Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Αγρονόμων και Τοπογράφων Μηχανικών – Μηχανικών Γεωπληροφορικής
Τομέας Τοπογραφίας
Εργαστήριο Φωτογραμμετρίας

Μοντελοποίηση και Ανάλυση Χωροχρονικών Επιδημιολογικών Δεδομένων με Τεχνικές Μηχανικής Μάθησης

Διπλωματική Εργασία

της

Ραφαέλας Ι. Ζουριδάκη

Επιβλέπων:

Νικόλαος Δουλάμης

Καθηγητής, ΕΜΠ

Αθήνα, Ιούλιος 2022

National Technical University of Athens
School of Rural, Surveying and Geoinformatics Engineering
Department of Topography
Laboratory of Photogrammetry

Modeling and Analysis of Spatiotemporal Epidemiological Data Using Machine Learning Techniques

Diploma Thesis

Rafaela I. Zouridaki

Supervisor:
Nikolaos Doulamis
Professor, NTUA

Athens, July 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Αγρονόμων και Τοπογράφων Μηχανικών – Μηχανικών Γεωπληροφορικής
Τομέας Τοπογραφίας
Εργαστήριο Φωτογραμμετρίας

Μοντελοποίηση και Ανάλυση Χωροχρονικών Επιδημιολογικών Δεδομένων με Χρήση Μηχανικής Μάθησης

Διπλωματική Εργασία

της

Ραφαέλας Ι. Ζουριδάκη

Επιβλέπων:
Νικόλαος Δουλάμης
Καθηγητής, ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19 Ιουλίου 2022.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
.....	.....	.....
Νικόλαος Δουλάμης	Κωνσταντίνος Καράντζαλος	Ανδρέας Γεωργόπουλος
Καθηγητής	Αναπληρωτής Καθηγητής	Καθηγητής

Αθήνα, Ιούλιος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Αγρονόμων και Τοπογράφων Μηχανικών – Μηχανικών Γεωπληροφορικής
Τομέας Τοπογραφίας
Εργαστήριο Φωτογραμμετρίας

(Υπογραφή)

. . . . . .

Ραφαέλα Ζουριδάκη

19 Ιουλίου 2022

Ευχαριστίες

Με την ολοκλήρωση των προπτυχιακών μου σπουδών νιώθω επιτακτική ανάγκη να ευχαριστήσω όλους όσοι στάθηκαν δίπλα μου και διαδραμάτισαν καθοριστικό ρόλο στην πορεία μου.

Πρωτίστως, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κύριο Νικόλαο Δουλάμη για την εμπιστοσύνη του και τη δυνατότητα την οποία μου παρείχε στο να γνωρίσω καλύτερα τον κόσμο της Μηχανικής Μάθησης και της Επιστήμης Δεδομένων. Στη συνέχεια, θα ήθελα να ευχαριστήσω τη Δρ. Μαρία Κασελίμη, για την ενδελεχή συμβολή της, την καθοδήγηση και τις γνώσεις τις οποίες μου προσέφερε κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Ακόμη, ένα μεγάλο ευχαριστώ για την πολύτιμη στήριξή του όλα αυτά τα χρόνια, θα ήθελα να πω στον κύριο Ανδρέα Γεωργόπουλο, σύμβουλο καθηγητή μου κατά τη φοίτησή μου στο ΕΜΠ. Ένα θερμό ευχαριστώ οφείλω και στον Ιωάννη Κόλια, υποψήφιο διδάκτορα της σχολής Μηχανολόγων Μηχανικών ΕΜΠ, για τις συμβουλές και τη στήριξή του κατά τη διάρκεια συγγραφής της παρούσας εργασίας.

Θα ήθελα να πω ένα μεγάλο ευχαριστώ, ακόμη, στους ανθρώπους με τους οποίους όλα αυτά τα χρόνια δημιουργήσαμε όμορφες και αξέχαστες στιγμές. Σε εκείνους τους ανθρώπους οι οποίοι στάθηκαν δίπλα μου και οι οποίοι με καθοδήγησαν όταν αυτό χρειαζόταν, στους φίλους μου.

Τέλος, θερμές ευχαριστίες δίνω στην οικογένειά μου, στους γονείς μου, Μαρία και Ιωάννη, για τη συμπαράσταση, τη στήριξη, την αγάπη και όλα όσα μου έχουν προσφέρει σε κάθε στάδιο της ζωής μου.

Αθήνα, Ιούλιος 2022
Ραφαέλα Ζουριδάκη

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως θέμα τη δημιουργία και ανάλυση μοντέλων χωροχρονικών επιδημιολογικών δεδομένων με χρήση τεχνικών μηχανικής μάθησης. Τα εν λόγω μοντέλα δημιουργούνται με αλγορίθμους μηχανικής μάθησης και προβλέπουν τα κρούσματα και τους θανάτους κορονοϊού για εννέα πόλεις. Ακόμη, μελετώνται εκείνες οι μεταβλητές οι οποίες διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη. Τέλος, επιλέγεται εκείνο το μοντέλο το οποίο έχει παρουσιάσει την καλύτερη προσαρμογή στα δεδομένα.

Τα τελευταία δύο χρόνια, ο πλανήτης έχει έρθει αντιμέτωπος με την πανδημία του κορονοϊού. Ένα μεγάλο μέρος της επιστημονικής κοινότητας έχει στρέψει το ενδιαφέρον της στη μελέτη, κατανόηση ακόμη και στην πρόβλεψη του φαινομένου του κορονοϊού. Συναντώνται αρκετά δημιουργηθέντα μοντέλα πρόβλεψης της διασποράς του κορονοϊού, αλλά και των θανάτων οι οποίοι αποδίδονται στον κορονοϊό. Επιπροσθέτως, γίνονται προσπάθειες προσδιορισμού των χαρακτηριστικών εκείνων τα οποία επηρεάζουν τη διάδοση και τους θανάτους. Έτσι, δημιουργείται η τρέχουσα εργασία, στην οποία παρουσιάζονται τα αποτελέσματα και η αξιολόγηση μοντέλων παλινδρόμησης και παρουσιάζονται εκείνα τα χαρακτηριστικά τα οποία φαίνεται να έχουν σημαντικό βάρος στις προβλέψεις.

Όπως ήδη αναφέρθηκε, χρησιμοποιήθηκε ένα σετ δεδομένων για εννέα διαφορετικές πρωτεύουσες. Τα δεδομένα αυτά, αρχικά, οπτικοποιήθηκαν με σκοπό να ελεγχθούν ως προς την πληρότητά τους, αλλά και με σκοπό την καλύτερη ερμηνεία και κατανόηση της διακύμανσης και της πιθανής σχέσης ανάμεσα στη διάδοση και στους θανάτους κορονοϊού. Στη συνέχεια, δημιουργήθηκαν μοντέλα με χρήση έξι διαφορετικών αλγορίθμων. Συγκεκριμένα, χρησιμοποιήθηκε το Multiple Linear Regression, το Support Vector Regression, το LASSO, το Gaussian Process Regression, το Random Forest Regression και το XGBoost Regression. Στη συνέχεια, τα μοντέλα αξιολογήθηκαν βάσει των αποδόσεών τους, τόσο με χρήση μετρικών αξιολόγησης μοντέλων παλινδρόμησης, όσο και με τη χρήση διαγραμμάτων. Τέλος, εντοπίζονται τα βάρη εκάστης ανεξάρτητης μεταβλητής στην πρόβλεψη. Για τα μοντέλα Random Forest Regression και XGBoost Regression χρησιμοποιείται έτοιμος τρόπος υπολογισμού τους μέσω χρήσης βιβλιοθήκης της Python, ενώ για τα υπόλοιπα μοντέλα υπολογίζονται τα βάρη των συντελεστών.

Από την παραπάνω ανάλυση, προκύπτει ότι τα καλύτερα αποτελέσματα, για την πλειονότητα των πόλεων, εμφανίζονται για το μοντέλο Random Forest Regression. Άρα, ως βέλτιστο μοντέλο επιλέγεται αυτό. Αρκετά ικανοποιητική απόδοση εμφάνισε και το μοντέλο του XGBoost Regressor. Επιπροσθέτως, οι μεταβλητές οι οποίες φαίνεται να επηρέασαν τις προβλέψεις σχετίζονται, κυρίως, με την κάλυψη των περιοχών σε βλάστηση, με ατμοσφαιρικούς ρύπους και με μετεωρολογικά φαινόμενα.

Εν κατακλείδι, ως τελικό πόρισμα για τη μοντελοποίηση και την ανάλυση των χωροχρονικών επιδημιολογικών δεδομένων με τεχνικές μηχανικής μάθησης είναι ότι καλύτερη απόδοση εμφανίζουν αλγόριθμοι οι οποίοι χρησιμοποιούν δένδρα αποφάσεων και τα χαρακτηριστικά τα οποία επηρεάζουν τη διάδοση και τους θανάτους είναι κατεξοχήν η μεταβλητή της κάλυψης σε βλάστηση και εν συνεχεία, οι ατμοσφαιρικοί ρύποι και τα μετεωρολογικά φαινόμενα.

Λέξεις Κλειδιά

Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Παλινδρόμηση, Χρονοσειρές, Μοντέλα Πρόβλεψης, Μετρικές Αξιολόγησης, Βαρύτητα Μεταβλητών, Κορονοϊό

Abstract

This thesis aims to create and analyze models of spatiotemporal epidemiological data using Machine Learning techniques. These models are created using ML algorithms, while focusing on predictions of cases and deaths of Covid-19, for nine cities. In addition, feature importance for each model is being extracted, in order to determine the possible factors that influence the Covid-19 spread. Subsequently, best fit model is the one that represents the best predictions for both propagation rates and mortality rates, based on model evaluation metrics.

Over the past two years, the pandemic of covid-19 has been decimating humanity. More and more scientists turn their interest to studying, understanding and predicting Covid-19 spread. In the existing literature, numerous prediction models for the spread and the mortality caused by covid-19 can be found. In addition, scientists show particular interest in appointing all possible features that affect the spread of covid. In this specific thesis, the results of the models' prediction and their evaluation are presented, as well as the feature importance of each variable.

In an endeavor to address the aforementioned issues, there is a vital need to understand the data set. Thus, firstly, a visualization is provided of the actual values of cases and deaths, for each city. Subsequently, six ML models were created. The algorithms used are Multiple Linear Regression, Support Vector Regression, LASSO, Gaussian Process Regression, Random Forest Regression, and XGBoost Regressor. Besides, regression evaluation metrics were utilized to evaluate the models' performance. Lastly, the score for all the input features was computed. Random Forest and XGBoost have built-in feature importance, therefore the scikit-learn package was used aiming to reckon that score. As for the rest of the models, coefficients were used as a crude type of feature importance.

The presented evidence conclude that the Random Forest was the model with the best performance. Furthermore, the key factors that influence the evolution trend of Covid-19 are related to urban vegetation, air pollutants such as SO_2, NO_2, O_3, PM_{xx} and meteorological conditions such as temperature, humidity, wind speed and precipitation.

Finally, algorithms based on decision trees tend to have better performance than simpler algorithms such as Linear Regression, LASSO, SVR and GPR. Ultimately there is a strong association between urban vegetation and the spread of the pandemic. Similar relationship can be found with respect to air pollutants and meteorological factors and the spread of Covid-19.

Key Words

Artificial Intelligence, Machine Learning, Regression, Timeseries, Forecasting Models, Evaluation Metrics, Feature Importance, Coronavirus

Πίνακας περιεχομένων

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Ακρωνύμια

AQI	Air Quality Index
EPA	Environmental Protection Agency
ESA	European Space Agency
EVS	Explained Variance Score
GPR	Gaussian Process Regression
KDE	Kernel Density Estimation
LASSO	Least Absolute Shrinkage and Selection Operator
LOOCV	Leave One Out Cross Validation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MdAE	Median Absolute Error
MSE	Mean Squared Error
NPL	Natural Language Processing
RMSE	Root Mean Squared Error
RSS	Residual Sum of Squares
SVM	Support Vector Machines
TSS	Total Sum of Squares
NO_2	Διοξείδιο του Αζώτου
SO_2	Διοξείδιο του Θείου
CO	Μονοξείδιο του Άνθρακα
O_3	Όζον
ΠΟΥ	Παγκόσμιος Οργανισμός Υγείας
PM	Σωματιδιακή Ύλη
ΤΝ	Τεχνητή Νοημοσύνη

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο

Σκοπός της παρούσας εργασίας είναι ο εντοπισμός εκείνου του μοντέλου μηχανικής μάθησης, το οποίο προβλέπει τη διάδοση και τους θανάτους του Covid-19, με την καλύτερη προσαρμογή, άρα με τα καλύτερα αποτελέσματα, για τα υπάρχοντα δεδομένα. Ακόμη, μελετώνται εκείνα τα χαρακτηριστικά τα οποία επηρεάζουν τα αποτελέσματα των προβλέψεων.

Συγκεκριμένα, αναπτύσσονται έξι διαφορετικά μοντέλα παλινδρόμησης για εννέα διαφορετικές πόλεις. Τα μοντέλα αυτά είναι το Multiple Linear Regression, το Support Vector Regression, το LASSO, το Gaussian Process Regression, το Random Forest Regression και το XGBoost Regression. Οι πόλεις οι οποίες μελετώνται είναι η Αθήνα, το Βερολίνο, οι Βρυξέλλες, η Λισαβόνα, το Λονδίνο, η Μαδρίτη, η Μόσχα, το Παρίσι και η Πράγα.

Κάθε μοντέλο αξιολογείται βάσει των μετρικών οι οποίες το περιγράφουν, αλλά και των δημιουργηθέντων διαγραμμάτων τα οποία αφορούν τις προβλέψεις. Πρόκειται για διαγράμματα στα οποία παρατίθενται οι τιμές των προβλέψεων με τις πραγματικές τιμές και διαγράμματα υπολοίπων.

Ακόμη, μέσα από τη δημιουργία διαγραμμάτων για τα βάρη εκάστης μεταβλητής, γίνεται εύκολα αντιληπτός ο ρόλος κάθε μεταβλητής στην πρόβλεψη. Η μελέτη αυτή, συντελεί στην καλύτερη κατανόηση του προβλήματος, όπως επίσης στον εντοπισμό των σημαινόντων στοιχείων της πρόβλεψης.

Τέλος, ερμηνεύονται τα αποτελέσματα για την απόδοση των μοντέλων και για τα βάρη των ανεξάρτητων μεταβλητών τα οποία εντοπίζονται. Επιδιώκεται να συγκριθούν τα αποτελέσματα της παρούσας εργασίας, με εκείνα από αντίστοιχες ερευνητικές εργασίες.

1.2 Δομή

Η εν λόγω διπλωματική εργασία διαρθρώνεται συνολικά σε πέντε κεφάλαια, μαζί με την εισαγωγή, και έχει την ακόλουθη δομή:

Στο Κεφάλαιο 2, κρίνεται σκόπιμο να γίνει αναφορά στη μετάδοση του Covid-19, στους παράγοντες οι οποίοι εικάζεται ότι επιδρούν. Συγκεκριμένα, αναλύονται οι ατμοσφαιρικοί παράγοντες και τα μετεωρολογικά φαινόμενα. Τέλος, κρίνεται σκόπιμο να αναφερθούν και να σχολιαστούν σχετικές εργασίες από ερευνητές. Πρόκειται για εργασίες, οι οποίες έχουν πραγματοποιηθεί και έχουν αναρτηθεί και οι οποίες χρησιμοποιούν χρονοσειρές και μοντέλα πρόβλεψης μηχανικής μάθησης για την πανδημία.

Στο Κεφάλαιο 3 παρουσιάζεται το θεωρητικό υπόβαθρο της εργασίας. Αναφέρονται οι βασικές αρχές της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των χρονοσειρών. Μεγαλύτερη εστίαση γίνεται στο κομμάτι της μηχανικής μάθησης το οποίο σχετίζεται με τους αλγορίθμους αλλά και με τις μετρικές οι οποίες χρησιμοποιούνται για την αξιολόγηση των αποτελεσμάτων.

Στο Κεφάλαιο 4 γίνεται η ανάλυση των πειραματικών αποτελεσμάτων. Στην αρχή, παρατίθενται διαγράμματα για τα αρχικά δεδομένα, για καλύτερη κατανόηση και εμπέδωση του προβλήματος. Στη συνέχεια, το ενδιαφέρον επικεντρώνεται στην ανάλυση και αξιολόγηση των μοντέλων πρόβλεψης, με τη βοήθεια πινάκων των μετρικών και διαγραμμάτων των τελικών αποτελεσμάτων.

Τέλος, στο Κεφάλαιο 5 γίνεται μία σύνοψη της ανάλυσης που πραγματοποιήθηκε στην παρούσα εργασία. Η σύνοψη αυτή επιτυγχάνεται μέσω συμπερασμάτων τα οποία προκύπτουν. Ακόμη, παρουσιάζονται ορισμένες ιδέες οι οποίες θα μπορούσαν να εφαρμοσθούν σε μελλοντική ερευνητική εργασία.

Κεφάλαιο 2

Μετάδοση Covid-19 και Παράγοντες Επίδρασης

2.1 Εισαγωγή

Στο δεύτερο κεφάλαιο, παρατίθενται πληροφορίες σχετικά με τα δεδομένα τα οποία χρησιμοποιήθηκαν για την τρέχουσα μοντελοποίηση και ανάλυση. Αναλύονται ορισμένα στοιχεία για την ατμοσφαιρική ρύπανση και για τους ρυπαντές. Ακόμη, κρίνεται αρκετά σημαντικό να παρουσιαστούν και να σχολιαστούν ορισμένα από τα αποτελέσματα προγενέστερων εργασιών οι οποίες σχετίζονται με μοντέλα πρόβλεψης για τη διασπορά και τους θανάτους του κορονοϊού, αλλά και για τον εντοπισμό των μεταβλητών εκείνων οι οποίες φαίνεται να επηρεάσουν σε αυτά τα φαινόμενα.

2.2 Δεδομένα

Τα δεδομένα τα οποία χρησιμοποιήθηκαν στην παρούσα εργασία, προέρχονται από τον ιστότοπο OurWorldInData. Στον εν λόγω ιστότοπο συναντώνται ερευνητικές εργασίες και δεδομένα, που αφορούν μείζονος σημασίας παγκόσμια προβλήματα, όπως είναι η φτώχεια, οι ασθένειες, η πείνα, η κλιματική αλλαγή, οι πόλεμοι, οι υπαρξιακοί κίνδυνοι και η ανισότητα (Wikipedia, 2022).

Σύμφωνα και με προηγούμενες έρευνες, είναι γεγονός ότι μοτίβα πανδημιών και μοτίβα μετάδοσης τους είναι συνυφασμένα με φυσικά και ανθρώπινα γεωγραφικά χαρακτηριστικά. Δηλαδή, η πρόοδος της εξάπλωσης μίας πανδημίας εξαρτάται από κοινωνικούς, οικονομικούς και περιβαλλοντικούς παράγοντες. Συνεπώς, για την τρέχουσα ποσοτική ανάλυση χρησιμοποιήθηκαν δεδομένα, τα οποία αφορούν ατμοσφαιρικούς ρύπους, όπως είναι για παράδειγμα η σωματιδιακή ύλη $PM_{2.5}$ και $PM_{2.5}$, το όζον, το NO_2 και το SO_2. Ακόμη, χρησιμοποιήθηκαν δεδομένα τα οποία αφορούν μετεωρολογικά φαινόμενα, όπως είναι η υγρασία, ο άνεμος, η πίεση, η θερμοκρασία. Επίσης, χρησιμοποιήθηκαν δεδομένα τα οποία αφορούν πληθυσμιακά στοιχεία, όπως είναι για παράδειγμα ο μέσος όρος της ηλικίας για κάθε πόλη, ηλικιακές ομάδες άνω των 65 και 70 ετών, στοιχεία τα οποία σχετίζονται με την υγεία όπως άνθρωποι οι οποίοι αντιμετωπίζουν διαβητολογικά και καρδιαγγειακά προβλήματα και κοινωνικο-οικονομικά στοιχεία, όπως εισοδηματικά κριτήρια. Τέλος, χρησιμοποιήθηκαν δεδομένα τα οποία αφορούν το πράσινο των πόλεων, όπως είναι η κάλυψη σε βλάστηση και η πυκνότητα των δέντρων.

Συγκεκριμένα, στον παρακάτω Πίνακα παρουσιάζονται τα ονόματα των μεταβλητών οι οποίες χρησιμοποιήθηκαν για την ανάλυση και η περιγραφή τους.

Ανεξάρτητες Μεταβλητές	**Περιγραφή**
Wind Speed	Μέση ημερήσια ταχύτητα ανέμου
Wind Gust	Μέση ημερήσια ριπή ανέμου
Temperature	Μέση ημερήσια θερμοκρασία
SO_2	Μέση ημερήσια τιμή διοξειδίου του θείου
Pressure	Μέση ημερήσια ατμοσφαιρική πίεση

$PM_{2.5}$	Μέση ημερήσια συγκέντρωση σωματιδιακής ύλης με διάμετρο μικρότερη των 2.5μm
PM_{10}	Μέση ημερήσια συγκέντρωση σωματιδιακής ύλης με διάμετρο μικρότερη των 10μm
O_3	Μέση ημερήσια τιμή διοξειδίου του όζοντος
NO_2	Μέση ημερήσια τιμή διοξειδίου του αζώτου
Humidity	Μέση ημερήσια υγρασία
Dew	Μέση ημερήσια τιμή σημείου δρόσου
Median Age	Μέση τιμή πληθυσμού ανά πόλη
Aged 65 Older	Πληθυσμός άνω των 65 ετών
Aged 70 Older	Πληθυσμός άνω των 70 ετών
GDP per capita	Κατά Κεφαλήν Εισόδημα
Cardiovascular Death Rate	Καρδιαγγειακός Κίνδυνος ανά Χώρα
Diabetes Prevalence	Ποσοστιαίος αριθμός διαβητικών στη χώρα
Female Smokers	Ποσοστιαίος δείκτης γυναικών καπνιστών ανά πόλη
Male Smokers	Ποσοστιαίος δείκτης ανδρών καπνιστών ανά πόλη
Tree Density	Πυκνότητα Δέντρων
LCV	Κάλυψη Εδάφους από Βλάστηση

Πίνακας 1 *Μεταβλητές των μοντέλων, Ιδία Επεξεργασία*

Αξίζει να σημειωθεί ότι η χρονική κλίμακα η οποία χρησιμοποιείται για τις παραπάνω μεταβλητές είναι το διάστημα μίας ημέρας. Αυτό, εύλογα δημιουργεί ορισμένες «ιδιαίτερες» συνθήκες. Για παράδειγμα, οι αισθητήρες των ατμοσφαιρικών ρύπων δεν λαμβάνουν μία μοναδική ημερήσια τιμή, αλλά πολλές τιμές μέσα στην ημέρα. Έτσι, επιλέγεται να χρησιμοποιηθεί η αντιπροσωπευτικότερη τιμή των μετρήσεων, ο μέσος όρος τους. Το κόστος, λοιπόν, το οποίο υπάρχει σε αυτές τις περιπτώσεις άπτεται του γεγονότος ότι χάνεται πληροφορία.

2.3 Ατμοσφαιρική Ρύπανση

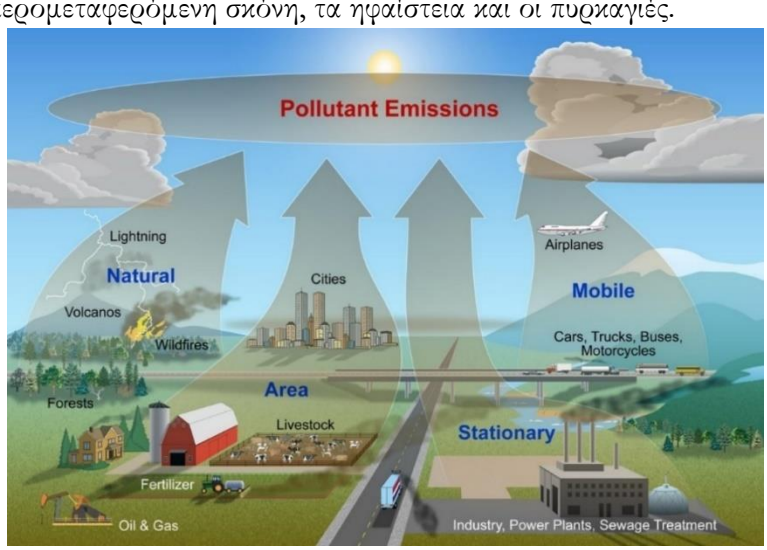
Κάθε είδος ρύπανσης του αέρα, η οποία προέρχεται είτε από φυσικά είτε από ανθρωπογενή αίτια, χαρακτηρίζεται ως ατμοσφαιρική ρύπανση. Πρόκειται για μία κατάσταση ανισορροπίας των στοιχείων της ατμόσφαιρας. Οι παράγοντες οι οποίοι την προκαλούν ονομάζονται ατμοσφαιρικοί ρυπαντές (Ευρωπαϊκός Οργανισμός Περιβάλλοντος, 2020). Οι ατμοσφαιρικοί ρυπαντές στις πόλεις αποτελούν κυρίως μείγμα πολλών διαφορετικών ρυπαντών, από τους οποίους μερικοί είναι ορατοί, όπως είναι η σκόνη και η αιθάλη, ενώ πολλοί είναι αόρατοι, όπως είναι τα πολύ μικρά σωματίδια ή τα αέρια.

Η ρύπανση διακρίνεται ως πρωτογενής ή δευτερογενής ανάλογα με τον τρόπο παραγωγής των ρυπαντών. Συγκεκριμένα, οι πρωτογενείς ρυπαντές είναι ουσίες οι οποίες παράγονται απευθείας από μια φυσική ή ανθρωπογενή διαδικασία, όπως η στάχτη από μια έκρηξη ηφαιστείου ή το μονοξείδιο του άνθρακα από τις εξατμίσεις των οχημάτων, αντίστοιχα. Από την άλλη πλευρά, οι δευτερογενείς ρυπαντές δεν απελευθερώνονται, αλλά δημιουργούνται στον αέρα όταν οι πρωτογενείς ρυπαντές αντιδρούν ή αλληλοεπιδρούν (ESA, 2013).

Αξίζει να σημειωθεί ότι σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ), για την ατμοσφαιρική ρύπανση ορίζεται ότι «περιορίζεται σε καταστάσεις στις οποίες η εξωτερική περιβαλλοντική ατμόσφαιρα περιέχει υλικά σε συγκέντρωση η οποία είναι επιβλαβής για τους ανθρώπους και το γύρω περιβάλλον τους» (ESA, 2013).

Επιπροσθέτως, σύμφωνα με τον ΠΟΥ, η ατμοσφαιρική ρύπανση είναι ο σημαντικότερος περιβαλλοντικός κίνδυνος για την υγεία στην Ευρωπαϊκή Ένωση (ΕΕ). Εκτιμάται ότι στην ΕΕ, προκαλεί κάθε χρόνο περίπου 400000 πρόωρους θανάτους, ενώ το εξωτερικό κόστος το οποίο συνεπάγεται σε ό,τι αφορά τις επιπτώσεις στην υγεία ανέρχεται σε εκατοντάδες δισεκατομμύρια ευρώ. Ιδιαίτερα εκτεθειμένοι είναι όσοι άνθρωποι ζουν σε αστικές περιοχές.

Στο παρακάτω Σχήμα απεικονίζονται οι τέσσερις κύριοι τύποι πηγών της ατμοσφαιρικής ρύπανσης. Πρόκειται για τις κινητές πηγές, όπως είναι τα μέσα μεταφοράς. Ακόμη, πρόκειται για σταθερές πηγές όπως είναι οι διάφορες βιομηχανικές εγκαταστάσεις και τα εργοστάσια. Εν συνεχεία, πηγές ατμοσφαιρικής ρύπανσης αποτελούν διάφορες εκτάσεις όπως είναι οι γεωργικές εκτάσεις, οι πόλεις, αλλά και τα τζάκια και τέλος είναι οι φυσικές πηγές όπως είναι η αερομεταφερόμενη σκόνη, τα ηφαίστεια και οι πυρκαγιές.



Σχήμα 1 *Πηγές ατμοσφαιρικής ρύπανσης, (National Park Service, 2018)*

2.3.1 Ρυπαντές

Οι κύριοι ρυπαντές οι οποίοι συναντώνται στην ατμόσφαιρα είναι το όζον (O_3), η Σωματιδιακή Ύλη/Particulate Matter (PM), το διοξείδιο του αζώτου (NO_2) καθώς επίσης το διοξείδιο του θείου (SO_2) και το μονοξείδιο του άνθρακα (CO). Στις επόμενες σειρές πρόκειται να ακολουθήσει μία σύντομη ανάλυση των προαναφερθέντων ρυπαντών.

2.3.1.1 Όζον

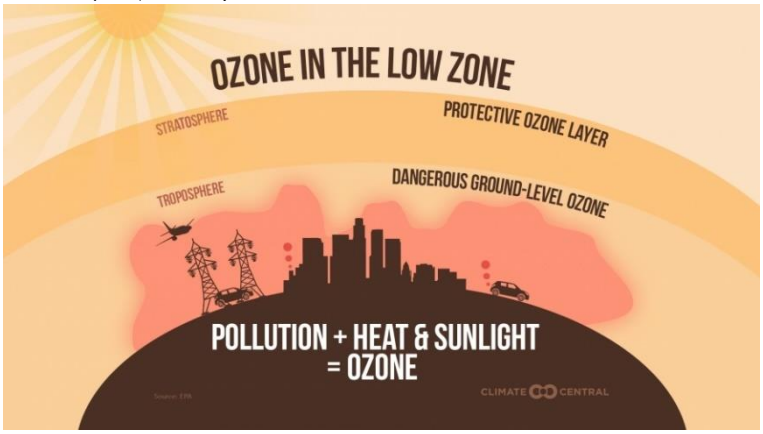
Το όζον, O_3, είναι ένα άχρωμο αέριο το οποίο σχηματίζεται μέσω χημικών αντιδράσεων μεταξύ δραστικών οργανικών αερίων και οξειδίων του αζώτου παρουσία ηλιακών ακτινών (ESA, 2013).

Το όζον σε επίπεδο εδάφους και έως 10 χιλιόμετρα πάνω από την επιφάνεια της Γης, τροποσφαιρικό όζον, αποτελεί έναν από τους δευτερογενείς ερεθιστικούς ρυπαντές οι οποίοι υπάρχουν στην αιθαλομίχλη η οποία δημιουργείται στις αστικές περιοχές. Δύναται να επιδεινώσει υπάρχουσες αναπνευστικές νόσους και να προκαλέσει ερεθισμό στο λαιμό, πονοκεφάλους και πόνο στο στήθος, αλλά και να επηρεάσει σημαντικά τη χλωρίδα των περιοχών.

Η συγκέντρωση όζοντος σε ανθυγιεινά επίπεδα συνήθως παρατηρείται κατά τις θερμές καλοκαιρινές μέρες σε αστικά περιβάλλοντα, παρά ταύτα είναι δυνατόν να εμφανίσει υψηλά επίπεδα ακόμη και σε πιο κρύους μήνες. Επιπροσθέτως, το όζον μπορεί να μεταφερθεί σε μακρινές αποστάσεις μέσω του αέρα. Αυτό έχει σαν αποτέλεσμα την παρατήρηση υψηλών επιπέδων όζοντος ακόμη και στις αγροτικές περιοχές. Τέλος, να σημειωθεί ότι το τροποσφαιρικό όζον είναι το κύριο συστατικό της αιθαλομίχλης (EPA, 2022).

Υπάρχει επίσης και το στρώμα του όζοντος στην στρατόσφαιρα σε ύψος 12-40 χιλιομέτρων επάνω από την επιφάνεια της Γης. Σε αντίθεση με το τροποσφαιρικό όζον, αυτό το στρώμα όζοντος λειτουργεί με έναν εξαιρετικά ευεργετικό ρόλο επειδή απορροφά την επικίνδυνη υπεριώδη ακτινοβολία (UV-B) η οποία εκπέμπεται από το Ήλιο, αποτρέποντάς την με αυτόν τον τρόπο στο να φτάσει στο έδαφος (UNESCO International Science, 1997).

Στο Σχήμα η οποία ακολουθεί απεικονίζεται το όζον της στρατόσφαιρας, αλλά και το τροποσφαιρικό όζον. Ακόμη, παρουσιάζονται οπτικά οι τρόποι με τους οποίους δημιουργείται ο εν λόγω ρυπαντής.



Σχήμα 2 *Όζον στρατόσφαιρας και τροπόσφαιρας, (Climate Central, 2019)*

2.3.1.2 *Σωματιδιακή Ύλη*

Η σωματιδιακή ύλη, PM, ή αλλιώς τα αιωρούμενα σωματίδια είναι μικρού μεγέθους στερεά, ξηρή σκόνη, ή υγρά αιωρήματα, σταγονίδια, τα οποία βρίσκονται διασκορπισμένα στην ατμόσφαιρα. Εμφανίζουν μεγάλη ποικιλία σε Σχήμα και μέγεθος, σε φυσικές ιδιότητες και σε χημική σύσταση (ThermiAir, 2019).

Η σωματιδιακή ύλη μπορεί να περιλαμβάνει ζωντανούς οργανισμούς όπως είναι οι ιοί, τα βακτήρια και η μούχλα. Η επίδρασή τους είναι αρνητική τόσο ως προς το περιβάλλον όσο και ως προς την ανθρώπινη υγεία. Τα σωματίδια τα οποία έχουν διάμετρο μικρότερη από 10 μικρόμετρα σχηματίζουν τα λεγόμενα αερολύματα (Σχολή Μηχανολόγων Μηχανικών - Πανεπιστήμιο Θεσσαλίας, 2020).

$PM_{2.5}$: Η λεπτή σωματιδιακή ύλη αποτελείται από εισπνεύσιμα σωματίδια ρυπαντών με διάμετρο μικρότερη από 2.5 μικρόμετρα. Συνεπώς, το $PM_{2.5}$ καλύπτει όλα τα σωματίδια μεταξύ 0 και 2.5 μικρομέτρων. Τα εν λόγω αερολύματα θεωρούνται αρκετά επικίνδυνα για τον άνθρωπο, καθώς είναι εφικτή η διείσδυση και εναπόθεση τους στο ανθρώπινο σώμα. Συγκεκριμένα, μπορούν να εισέλθουν στους πνεύμονες και στην κυκλοφορία του αίματος προκαλώντας με αυτόν τον τρόπο σοβαρά προβλήματα υγείας. Έτσι, λοιπόν, οι πιο βαριές επιπτώσεις τους παρουσιάζονται στους πνεύμονες και στην καρδιά. Η έκθεση σε μεγάλη ποσότητα των συγκεκριμένων ρυπαντών δύναται να προκαλέσει βήχα ή δυσκολία στην αναπνοή, επιδείνωση του άσθματος και ανάπτυξη χρόνιας αναπνευστικής νόσου.

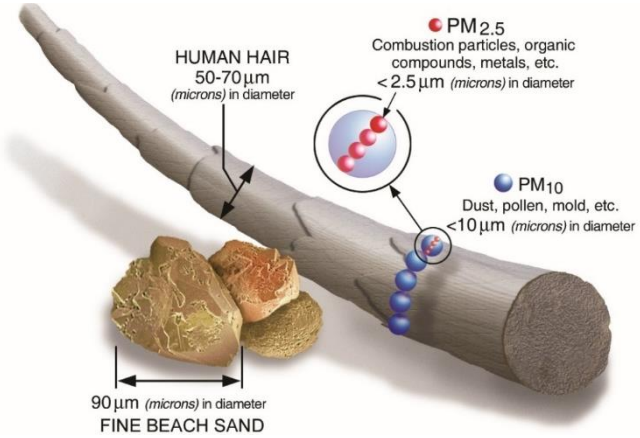
Τα αιωρούμενα σωματίδια τα οποία έχουν διάμετρο μικρότερη από 1 μικρόμετρο ονομάζονται λεπτομερή σωματίδια (fine mode particles) και χαρακτηρίζονται από μεγάλο χρόνο παραμονής στην ατμόσφαιρα.

Τα αιωρούμενα σωματίδια τα οποία έχουν διάμετρο μεγαλύτερη από 2.5 μικρόμετρα ονομάζονται χονδροειδή σωματίδια (coarse mode particles) είναι κυρίως πρωτογενή αλλά μπορεί να είναι και ανθρωπογενή ή να προέρχονται από φυσικά αίτια (Σχολή Μηχανολόγων Μηχανικών - Πανεπιστήμιο Θεσσαλίας, 2020).

PM_{10}: Η σωματιδιακή ύλη αποτελείται από εισπνεύσιμα σωματίδια ρυπαντών με διάμετρο μικρότερη από 10 μικρόμετρα. Συνεπώς, το PM_{10} καλύπτει όλα τα σωματίδια μεταξύ 0 και 10 μικρομέτρων. Τα σωματίδια τα οποία είναι μεγαλύτερα από 2.5 μικρόμετρα, και μικρότερα από 10 μικρόμετρα, είναι ικανά να διασχίσουν το λάρυγγα και τους πνεύμονες και μπορούν να επικαθίσουν στους αεραγωγούς προκαλώντας προβλήματα υγείας.

Η έκθεση σε μεγάλη ποσότητα των συγκεκριμένων ρυπαντών δύναται να οδηγήσει σε ερεθισμό των ματιών και του λαιμού, βήχα ή δυσκολία στην αναπνοή και επιδείνωση του άσθματος. Η συχνότερη και υπερβολική έκθεση μπορεί να έχει σοβαρότερες επιπτώσεις στην υγεία. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας και με τα εναρμονισμένα με την ΕΕ νομοθετημένα όρια για τα αιωρούμενα σωματίδια PM_{10} η μέση ημερήσια τιμή δεν θα πρέπει να υπερβαίνει την οριακή τιμή των 50 μικρογραμμαρίων ανά κυβικό μέτρο αέρα περισσότερες από 35 φορές το έτος (ThermiAir, 2019).

Τα σωματίδια τα οποία εμφανίζουν διάμετρο μεγαλύτερη των 10 μικρομέτρων συγκρατούνται από τα τριχίδια της μύτης, άρα δεν εισέρχονται στην αναπνευστική οδό. Αυτά τα σωματίδια προκαλούν ερεθισμό της μύτης και των ματιών, χωρίς όμως να προκαλούν μεγαλύτερες βλάβες στους οργανισμούς (ESA, 2013).



Σχήμα 3 *Σύγκριση μεγεθών PM, (EPA, 2021)*

2.3.1.3 *Διοξείδιο του αζώτου*

Το διοξείδιο του αζώτου, NO_2, αποτελεί έναν από τους πιο σημαντικούς αέριους ρύπους και καθίσταται τοξικό με την εισπνοή. Λόγω της δριμύτητας της ένωσης, είναι δυνατή η ανίχνευση ακόμη και χαμηλών συγκεντρώσεων μέσω της όσφρησης.

Παρά ταύτα, η αναπνοή σε περιβάλλον με υψηλά επίπεδα διοξειδίου του αζώτου αυξάνει τον κίνδυνο αναπνευστικών προβλημάτων και μειώνει τη λειτουργία των πνευμόνων. Συχνά παρουσιάζεται βήχας και δυσκολία στην αναπνοή. Ακόμη, ενδέχεται να εμφανιστούν και σοβαρότερα προβλήματα υγείας μέσω της παρατεταμένης έκθεσης, όπως για παράδειγμα αναπνευστικές λοιμώξεις, προβλήματα στο αίμα, στο ήπαρ και στον σπλήνα (Επιπτώσεις στην Υγεία, 2013).

Μερικές από τις πηγές διοξειδίου του αζώτου είναι οι μηχανές εσωτερικής καύσης, οι θερμοηλεκτρικοί σταθμοί, τα εργοστάσια χαρτοπολτού, δηλαδή οι βιομηχανικές διεργασίες και οι μεταφορές.

Μία μεγάλη θερμοκρασία καύσης αντιστοιχεί σε μεγάλη ποσότητα σχηματιζόμενων οξειδίων αζώτου. Μικρότερη θερμοκρασία οδηγεί σε μικρότερη ποσότητα σχηματιζόμενων οξειδίων αζώτου, όμως ο βαθμός απόδοσης της μηχανής μειώνεται αισθητά. Επιπροσθέτως, μπορεί να παραχθεί φυσικά κατά τη διάρκεια ηλεκτρικών καταιγίδων.

Αξίζει να σημειωθεί ότι το διοξείδιο του αζώτου διαδραματίζει σημαντικό ρόλο στην ατμοσφαιρική χημεία, με χαρακτηριστικά παραδείγματα το σχηματισμό του τροποσφαιρικού όζοντος και τη δημιουργία αιωρούμενων σωματιδίων (Wikipedia, 2021).

2.3.1.4 Διοξείδιο του θείου

Ο πρώτος ρυπαντής ο οποίος ιστορικά ανιχνεύθηκε στην ατμόσφαιρα ήταν μία ομάδα από ενώσεις οξειδίων του θείου, SO_2. Ως εκ τούτου, έχουν ληφθεί σημαντικά μέτρα ελέγχου των εκπομπών του ιδιαιτέρως στην Ευρώπη και στη Βόρεια Αμερική. Το διοξείδιο του θείου είναι ένα αέριο άχρωμο, όμως έχει μία ιδιαίτερη χαρακτηριστική οσμή. Τόσο το SO_2 όσο και τα προϊόντα των αντιδράσεων τα οποία ελευθερώνονται στην ατμόσφαιρα, σουφλίδια, αποτελούν σημαντικούς ρυπαντές της ατμόσφαιρας (Φωτιάδη, 2015).

Η έκθεση στο διοξείδιο του θείου μπορεί να προκαλέσει ερεθισμό στο λαιμό και τα μάτια. Επιπροσθέτως, μία μακροχρόνια έκθεση στο διοξείδιο του θείου μπορεί να προκαλέσει αναπνευστικά προβλήματα, να τροποποιήσει τον αμυντικό μηχανισμό των πνευμόνων και να επιδεινώσει τυχόν υπάρχουσες πνευμονολογικές και καρδιοαγγειακές παθήσεις, όπως το άσθμα, η χρόνια βρογχίτιδα και το εμφύσημα. Κατά κύριο λόγο, οι πιο ευπαθείς ομάδες είναι τα μικρά παιδιά και οι ηλικιωμένοι, αλλά και όλοι όσοι πάσχουν από καρδιοαγγειακές και πνευμονολογικές παθήσεις (Επιπτώσεις στην Υγεία, 2013).

Οι πηγές του είναι καύσεις στερεών ή υγρών καυσίμων τα οποία περιέχουν θείο. Οι φυσικές πηγές παραγωγής διοξειδίου του θείου στην ατμόσφαιρα είναι τα ηφαίστεια, οι πυρκαγιές και η σκόνη από απογυμνωμένο έδαφος. Όμως, υπάρχουν και οι ανθρωπογενείς πηγές διοξειδίου του θείου στην ατμόσφαιρα. Πρόκειται για διάφορες βιομηχανικές δραστηριότητες, χυτήρια μεταλλεύματος, οχήματα, αγροτικές δραστηριότητες, διύλιση πετρελαίου και γενικότερα βιομηχανικές κατεργασίες θειούχων ενώσεων (Μουστρής, 2015). Συνοπτικά, δηλαδή, πρόκειται για τη χημική βιομηχανία, την καύση και την επεξεργασία ορυκτών καυσίμων. Αυτό συμβαίνει καθώς οι γαιάνθρακες και το πετρέλαιο περιέχουν ενώσεις του θείου, οπότε εκλύεται διοξείδιο του θείου μέσω της καύσης τους.

Το διοξείδιο του θείου είναι μία ένωση διαλυτή στο νερό. Έτσι, από μόνο του όταν εισέρχεται σε έναν οργανισμό, απορροφάται από τα υγρά του ανώτερου αναπνευστικού συστήματος και έτσι ένα πολύ μικρό ποσοστό είναι ικανό να φτάσει στο κατώτερο αναπνευστικό σύστημα. Όμως, σε συνδυασμό με άλλους ρύπους με τους οποίους συνυπάρχει, όπως τον καπνό και τα αιωρούμενα σωματίδια, εισέρχεται στον οργανισμό και μεταφέρεται στους πνεύμονες μέσω της αναπνευστικής οδού.

Ακόμη, το διοξείδιο του θείου σε συνδυασμό με οξείδια του αζώτου αποτελούν πρόδρομο για τη δημιουργία της όξινης βροχής, η οποία δύναται να καταστεί επιζήμια τόσο για τα φυσικά οικοσυστήματα όσο και για τις ανθρωπογενείς κατασκευές, λχ. Αρχαιολογικοί χώροι. Τέλος, το διοξείδιο του θείου ευθύνεται και για τη δημιουργία σωματιδίων $PM_{2.5}$.

2.3.1.5 Μονοξείδιο του άνθρακα

Το μονοξείδιο του άνθρακα, CO, σχηματίζεται μέσω ατελούς καύσης οργανικής ύλης, δηλαδή καύσιμης ύλης η οποία περιέχει άνθρακα, από ηφαίστεια καθώς και μέσω φωτοχημικών αντιδράσεων οι οποίες λαμβάνουν χώρα στο τροποσφαιρικό τμήμα της ατμόσφαιρας.

Πρόκειται για ένα αέριο, ή υγρό, άχρωμο και πρακτικά άοσμο και άγευστο το οποίο όταν εισπνέεται σε υψηλά επίπεδα δύναται να γίνει αρκετά επιβλαβές για την υγεία των οργανισμών (EPA, 2010). Ενδεικτικά αναφέρεται ότι είναι ικανό να προκαλέσει από πονοκέφαλο, ζάλη, εμετό και συμπτώματα γρίπης, έως και θάνατο.

Στον εξωτερικό περιβάλλοντα χώρο οι κύριες πηγές παραγωγής μονοξειδίου του άνθρακα είναι τα αυτοκίνητα, τα φορτηγά και γενικότερα τα οχήματα και οι μηχανές τα οποία καίνε ορυκτά καύσιμα.

Το πιο χαρακτηριστικό πρόβλημα υγείας το οποίο είναι συνυφασμένο με τα ανησυχητικά επίπεδα μονοξειδίου του άνθρακα είναι η υποξία (hypoxia), δηλαδή η περιορισμένη διαθεσιμότητα οξυγόνου, η οποία δημιουργείται λόγω των υψηλών επιπέδων ανθρακοξυαιμοσφαιρίνης στο αίμα. Μέσω της εισπνοής αέρα με υψηλή περιεκτικότητα σε CO μειώνεται το οξυγόνο το οποίο μεταφέρεται μέσω του αίματος, σε διάφορα σημαντικά όργανα όπως η καρδιά και ο εγκέφαλος (Σιμιτσής, 2013). Άλλες πιθανές επιπτώσεις στην υγεία από το μονοξείδιο του άνθρακα μπορούν να εντοπισθούν στο κεντρικό νευρικό σύστημα, καθώς και στο αναπνευστικό σύστημα.

Σε υψηλά επίπεδα συγκέντρωσης μονοξειδίου του άνθρακα, τα οποία είναι πιο πιθανό να συμβούν σε εσωτερικούς και εσώκλειστους χώρους, προκαλείται ζάλη, σύγχυση, απώλεια αισθήσεων, ακόμη και θάνατος. Υψηλά επίπεδα συγκέντρωσης μονοξειδίου του άνθρακα δεν είναι πιθανό να συμβούν σε εξωτερικούς χώρους. Εν τούτοις, όταν υπάρχει αύξηση του επιπέδου συγκέντρωσης μονοξειδίου του άνθρακα σε εξωτερικό χώρο, τότε δύναται να δημιουργηθεί ανησυχία για ορισμένους ανθρώπους οι οποίοι ανήκουν σε ευπαθείς ομάδες, κυρίως άνθρωποι με καρδιαγγειακά προβλήματα. Η εν λόγω ανησυχία οφείλεται στο γεγονός ότι οι συγκεκριμένοι άνθρωποι έχουν ήδη μειωμένη ικανότητα στην πρόσληψη αίματος με επαρκή ποσότητα οξυγόνου στην καρδιά τους, όταν και όποτε απαιτείται η λήψη περισσότερης ποσότητας οξυγόνου από την καρδιά (EPA, 2021).

2.4 Προγενέστερες Εργασίες

Όπως έχει ήδη αναφερθεί, η πανδημία του κορονοϊού έχει κεντρίσει το ενδιαφέρον αρκετών επιστημόνων. Αρκετοί ερευνητές έχουν δημιουργήσει μοντέλα τα οποία προβλέπουν τη διάδοση του κορονοϊού και μέσω των οποίων εξετάζουν τις μεταβλητές εκείνες οι οποίες φαίνεται να ασκούν επιρροή στη διάδοση, αλλά και στις τιμές των θανάτων του κορονοϊού.

Στη συνέχεια, παρατίθενται τα αποτελέσματα ορισμένων εργασιών, οι οποίες εμφανίζουν παρόμοια μεθοδολογία με την τρέχουσα διπλωματική εργασία.

Σε μία παρεμφερή μελέτη των Α. Τέμενος, Ι. Τζώρτζης, Μ. Κασελίμη, Ι. Ράλλης, Α. Δουλάμης και Ν. Δουλάμης, μελετώνται τα αποτελέσματα διαφορετικών μοντέλων με χρήση ενός χωροχρονικού σετ δεδομένων, για οκτώ ευρωπαϊκές πόλεις. Επιπλέον, ερευνώνται οι μεταβλητές εκείνες οι οποίες εμφανίζουν σημαντικό στατιστικό βάρος στις προβλέψεις. Εκείνο το μοντέλο το οποίο εμφάνισε τα καλύτερα αποτελέσματα ήταν εκείνο του Random Forest. Τέλος, προέκυψε ότι οι τέσσερις σημαντικότεροι παράγοντες οι οποίοι φαίνεται να σχετίζονται με τη διάδοση του κορονοϊού για κάθε πόλη είναι η θερμοκρασία, η κινητικότητα, οι δραστηριότητες εμπορίου και η αναψυχή και τέλος η υπάρχουσα αστική βλάστηση. Ακόμη, παρατηρείται ότι παράγοντες του κλίματος και ατμοσφαιρικοί παράγοντες, επηρεάζουν τη

συμπεριφορά των μοντέλων (A. Temenos, I. Tzortzis, M. Kaselimi, I. Rallis, A. Doulamis, N. Doulamis, 2022).

Στο άρθρο των Ammar H. Elsheikh et al., αναπτύσσεται ένα LSTM μοντέλο βαθιάς μηχανικής μάθησης. Το μοντέλο αυτό προβλέπει τα κρούσματα, τις αναρρώσεις και τους θανάτους και εφαρμόζεται για έξι διαφορετικές χώρες, οι οποίες εφαρμόζουν διαφορετική πολιτική, με διαφορετικές κλιματολογικές και κοινωνικο-οικονομικές συνθήκες. Τέλος, προκύπτει ότι η απόδοσή του είναι καλύτερη, συγκριτικά με το ARIMA και το NARANN (Ammar H. Elsheikh et al., 2020).

Αρκετό ενδιαφέρον παρουσιάζει η εργασία των N. Ayoobi et all. Πρόκειται για μία εργασία στην οποία πραγματοποιήθηκε πρόβλεψη κρουσμάτων και θανάτων κορονοϊού με τεχνητές βαθιάς μηχανικής μάθησης. Συγκεκριμένα, χρησιμοποιήθηκαν μοντέλα όπως το LSTM, Bi-LSTM, Conv-LSTM, Bi-Conv-LSTM, GRU και Bi-GRU. Τα μοντέλα αυτά εφαρμόστηκαν για προβλέψεις μίας ημέρας, τριών και επτά ημερών. Οι συγγραφείς, ισχυρίζονται ότι μοντέλα όπως Bi-GRU και Bi-Conv-LSTM, δεν είχαν χρησιμοποιηθεί για προβλέψεις κορονοϊού. Επίσης, ισχυρίζονται ότι δεν εντοπίστηκε έρευνα η οποία να κάνει προβλέψεις για κάθε τρεις ή επτά ημέρες. Να σημειωθεί ότι ο σκοπός τους ήταν να δουν εάν η μείωση της υπολογιστικής ισχύς επιφέρει αποδεκτά τελικά αποτελέσματα (Nooshin Ayoobi et al., 2021).

Ακόμη, ενδιαφέρον παρουσιάζει η έρευνα των Ι. Κάβουρα, Μ. Κασελίμη, Ε. Πρωτοπαπαδάκης, Ν. Μπάκαλος, Ν. Δουλάμης και Α. Δουλάμης, η οποία επικεντρώνεται στη δημιουργία δύο μοντέλων βαθιάς μηχανικής μάθησης. Στο πρώτο μοντέλο επιδιώκεται να γίνει πρόβλεψη κρουσμάτων και θανάτων κορονοϊού και το δεύτερο μοντέλο προβλέπει τις εισαγωγές στο νοσοκομείο και τις διασωληνώσεις σε ΜΕΘ, λόγω κορονοϊού. Οι μέθοδοι οι οποίες χρησιμοποιήθηκαν είναι: Conv1D-LSTM, GRU, LSTM και SimpleRNN. Για το πρώτο μοντέλο φαίνεται ότι υπερέχει η μέθοδος SimpleRNN, ενώ για το δεύτερο ικανοποιητικά αποτελέσματα βγάζει η LSTM (Ioannis Kavouras et al., 2022).

Στη μελέτη των Lung-Chang Chien και Lung-Wen Chen επιλέχθηκαν 50 πόλεις των ΗΠΑ και 7 μεταβλητές, όπως η μέγιστη, η ελάχιστη και η μέση θερμοκρασία, η μέγιστη, ελάχιστη και μέση σχετική υγρασία και ο υετός. Αρχικά, μελετήθηκε η γραμμική συσχέτιση ανάμεσα στους μετεωρολογικούς παράγοντες και στον κορονοϊό και εντοπίστηκε ότι υψηλές θερμοκρασίες μειώνουν σημαντικά τη μετάδοση κορονοϊού, ενώ υψηλή σχετική υγρασία αυξάνει τον κίνδυνο. Επιπροσθέτως, παρατηρήθηκε ότι μία μη γραμμική συσχέτιση είναι πιθανό να περιγράφει καλύτερα την επίδραση των μετεωρολογικών παραγόντων στον κορονοϊό. Τέλος, στο τρίτο δημιουργηθέν μοντέλο, παρατηρήθηκε ότι τα υψηλότερα μεγέθη θερμοκρασιών δεν έχουν πλήρη αρνητική συσχέτιση με τον κορονοϊό (Lung-Chang Chien, Lung-Wen Chen, 2020).

Στην εργασία των Yongfa You και Shufen Pan, πραγματοποιήθηκε μελέτη επάνω στο πώς επηρεάζουν τη διάδοση του κορονοϊού τέσσερις μεταβλητές, αστική βλάστηση, θερμοκρασία αέρα, πυκνότητα πληθυσμού και αρχικό στάδιο εμφάνισης της λοίμωξης, για ορισμένες πόλεις των ΗΠΑ. Στην εργασία αυτή χρησιμοποιήθηκε η πολυμεταβλητή ποσοτική στατιστική τεχνική PAM. Σύμφωνα με τους Y. You και S. Pan «*η αστική βλάστηση μπορεί να μειώσει τη διάδοση του κορονοϊού. Κάθε 1% αύξηση στο ποσοστό της αστικής βλάστησης οδηγεί σε 2.6% μείωση των συνολικών κρουσμάτων κορονοϊού.*» Ακόμη, προέκυψε ότι η πληθυσμιακή πυκνότητα είναι έντονα συσχετισμένη με τα συνολικά κρούσματα κορονοϊού, καθώς υψηλές πληθυσμιακές πυκνότητες, οδηγούν σε υψηλό αριθμό κρουσμάτων. Επιπλέον παρατηρήθηκε ότι η σύντομη εφαρμογή περιοριστικών μέτρων επιδρά θετικά στον

περιορισμό της διάδοσης του κορονοϊού. Τέλος, παρατηρήθηκε ότι η ατμοσφαιρική θερμοκρασία εμφάνισε αρνητική συσχέτιση με τα συνολικά κρούσματα κορονοϊού, χωρίς όμως η συσχέτιση αυτή να είναι στατιστικά αξιόλογη (Yongfa You, Shufen Pan, 2020).

Τέλος, μία μελέτη των Ι. Κάβουρα, Ε. Πρωτοπαπαδάκη, Μ. Κασελίμη, Ν. Δουλάμη, φανέρωσε ότι υπάρχει σχέση ανάμεσα στον κορονοϊό και στην αλλαγή της ποιότητας του αέρα. Χρησιμοποιήθηκε ο δείκτης για την ατμοσφαιρική ποιότητα αέρα, Air Quality Index (AQI) και επίσης μελετήθηκε η διακύμανση των τιμών για τους ατμοσφαιρικούς ρυπαντές, σε διάφορες πόλεις. Οι μεταβολές στις τιμές των ρυπαντών προήλθαν λόγω των περιοριστικών μέτρων. Για να επιτευχθεί η εργασία αυτή, χρησιμοποιήθηκαν μοντέλα πρόβλεψης μηχανικής μάθησης, όπως LASSO, LR, RF Regression, Ridge, DNN και KNN (I. Kavouras et al., 2021).

Κεφάλαιο 3

Θεωρητικό Υπόβαθρο

3.1 Εισαγωγή

Το Κεφάλαιο 3 σχετίζεται με το πεδίο της Εφαρμοσμένης Επιστήμης Υπολογιστών και συγκεκριμένα με τον τομέα της Τεχνητής Νοημοσύνης. Πρόκειται να παρατεθούν πληροφορίες για την Τεχνητή Νοημοσύνη (ΤΝ), για τη Μηχανική Μάθηση και για τις Χρονοσειρές. Αναλυτικότερα, γίνεται μία προσπάθεια για καλύτερη επεξήγηση των «πτυχών» της μηχανικής μάθησης, όπως είναι οι κατηγορίες της, οι αλγόριθμοι οι οποίοι συναντώνται, αλλά κι οι μετρικές αξιολόγησης των μοντέλων μηχανικής μάθησης. Τέλος, αναλύεται το θεωρητικό υπόβαθρο το οποίο απαιτείται για την ανάπτυξη μοντέλου επιβλεπόμενης μηχανικής μάθησης με τεχνικές παλινδρόμησης για την πρόβλεψη των κρουσμάτων και των θανάτων Covid-19 με χρήση χρονοσειρών, σε συνδυασμό με τον τρόπο υλοποίησης του.

3.2 Τεχνητή Νοημοσύνη

Ως τεχνητή νοημοσύνη νοείται ο τομέας της επιστήμης των υπολογιστών στον οποίο επιδιώκεται ο σχεδιασμός υπολογιστικών συστημάτων τα οποία είναι ικανά να αναπαραγάγουν τις γνωστικές λειτουργίες ενός ανθρώπου, όπως μάθηση, σχεδιασμός και δημιουργικότητα. Μέσω της τεχνητής νοημοσύνης, οι μηχανές μαθαίνουν να επιλύουν προβλήματα και να επιδιώκουν την επίτευξη ενός στόχου, χάρη στην ικανότητά τους για προσαρμογή της συμπεριφορά τους λαμβάνοντας υπόψιν συνέπειες προηγούμενων δράσεων και δρώντας με αυτονομία. Τα δεδομένα τα οποία χρησιμοποιεί, επεξεργάζεται και ανταποκρίνεται μία μηχανή μπορεί να είναι ήδη έτοιμα, ή να τα συλλέγει από διάφορους αισθητήρες, όπως για παράδειγμα είναι η κάμερα και το μικρόφωνο (Ευρωπαϊκό Κοινοβούλιο, 2020).

Η τεχνητή νοημοσύνη έχει διάφορους υπό-κλάδους, η μελέτη των οποίων επιτρέπει την ευκολότερη κατανόηση στον τρόπο με τον οποίο γίνεται η εφαρμογή των κλάδων αυτών, επάνω σε διάφορα πεδία της βιομηχανίας. Οι κλάδοι αυτοί είναι: η Μηχανική Μάθηση, η Βαθιά Μάθηση, τα Νευρωνικά Δίκτυα, η Όραση Υπολογιστών, το Natural Language Processing (NLP), καθώς και το Cognitive Computing. Εκτενέστερη ανάλυση για τη Μηχανική Μάθηση πρόκειται να ακολουθήσει στη **3.3** συνεπώς στην τρέχουσα ενότητα πραγματοποιείται μία σύντομη αναφορά για τους υπόλοιπους κλάδους της Τεχνητής Νοημοσύνης.

Τα Νευρωνικά Δίκτυα λειτουργούν με παρόμοιες αρχές όπως οι ανθρώπινοι νευρώνες. Πρόκειται για μία διαδοχή αλγορίθμων οι οποίοι σκιαγραφούν τη σχέση η οποία υπάρχει ανάμεσα σε διάφορες υποκείμενες μεταβλητές και επεξεργάζονται τα δεδομένα όπως θα τα επεξεργαζόταν ένας ανθρώπινος εγκέφαλος.

Οι αλγόριθμοι της Όρασης Υπολογιστών επιδιώκουν να κατανοήσουν μία εικόνα. Έτσι, η εικόνα «σπάει» σε κομμάτια και γίνεται η μελέτη των διάφορων αντικειμένων, βάσει των τιμών των pixel. Η μηχανή, λοιπόν, ταξινομεί και μαθαίνει από ένα σύνολο εικόνων, ώστε να «εκμαιεύσει» μία καλύτερη τελική απόφαση βασισμένη σε προηγούμενες παρατηρήσεις.

Στη συνέχεια, το NLP είναι μία επιστήμη κατά την οποία μία μηχανή διαβάζει, καταλαβαίνει και ερμηνεύει μία γλώσσα. Άπαξ και η μηχανή κατανοήσει αυτό το οποίο ο χρήστης έχει την πρόθεση να επικοινωνήσει, τότε εκείνη αντιδρά με ανάλογο τρόπο.

Τέλος, οι αλγόριθμοι του cognitive computing επιδιώκουν να μιμηθούν το ανθρώπινο μυαλό, αναλύοντας κείμενο, ομιλία, Σχήμα και αντικείμενα με τρόπος με τους οποίους θα ακολουθούσε ένας άνθρωπος και επιδιώκει έτσι να δώσει το επιθυμητό αποτέλεσμα (Great Learning Team, 2022).

3.2.1 Εξηγήσιμη Τεχνητή Νοημοσύνη

Είναι γεγονός ότι η χρήση της Τεχνητής Νοημοσύνης έχει διαδραματίσει σημαντικό ρόλο στις επιστημονικές και τεχνολογικές εφαρμογές. Με το πέρασμα του καιρού οι ρίζες της Τεχνητής Νοημοσύνης γίνονται ολοένα και πιο δυνατές, επιτρέποντάς της να διεισδύσει ακόμη περισσότερο σε επιστημονικά και τεχνολογικά ζητήματα. Όμως, ένα βασικό ερώτημα το οποίο προκύπτει από όλα αυτά είναι το ένα πρέπει να εμπιστεύονται οι άνθρωποι έναν υπολογιστή.

Οι διάφορες εφαρμογές τεχνητής νοημοσύνης αποτελούν ένα όπλο. Ο λόγος για τον οποίο συμβαίνει αυτό, είναι ότι εάν δεν υπάρχει διαφάνεια, μπορεί κάποιος να δημιουργήσει εφαρμογές οι οποίες να διαιωνίζουν διάφορες κοινωνικά ζητήματα, όπως είναι προκαταλήψεις για τη φυλή, για το φύλο, για την ηλικία, για τη θρησκεία, για το σεξουαλικό προσανατολισμό. Για να μπορέσει να εξαλειφθεί αυτό το πρόβλημα, χρειάζεται οι εφαρμογές της τεχνητής νοημοσύνης, αφενός να σχεδιάζονται βάσει των αρχών που αναφέρθηκαν σε ακριβώς προηγούμενη ενότητα, αφετέρου να παρουσιάζουν διαφάνεια. Δηλαδή, για να μπορέσουν να είναι αξιόπιστες, δεν θα πρέπει ο τρόπος λειτουργίας τους να είναι ένα «μαύρο κουτί». Αντιθέτως, μέσω της διαφάνειας, οι μηχανικοί, οι χρήστες και οι διάφορες αρχές θα είναι σε θέση να ερμηνεύσουν, να καταλάβουν τον τρόπο με τον οποίο λειτουργεί ένα σύστημα και να το εμπιστευτούν (Σταμάτης, 2021).

Η Εξηγήσιμη Τεχνητή Νοημοσύνη – Explainable Artificial Intelligence (XAI), είναι το σύνολο των διαδικασιών και των μεθόδων, οι οποίες επιτρέπουν στους ανθρώπους να καταλάβουν και να εμπιστευτούν τα αποτελέσματα των εφαρμογών μηχανικής μάθησης. Συνεπώς, μέσω της εξηγήσιμης ΤΝ γίνονται κατανοητοί οι αλγόριθμοι μηχανικής μάθησης, η βαθιά μάθηση και τα νευρωνικά δίκτυα (Watson, 2021).

3.2.2 Θεμελιώδεις Αρχές Τεχνητής Νοημοσύνης

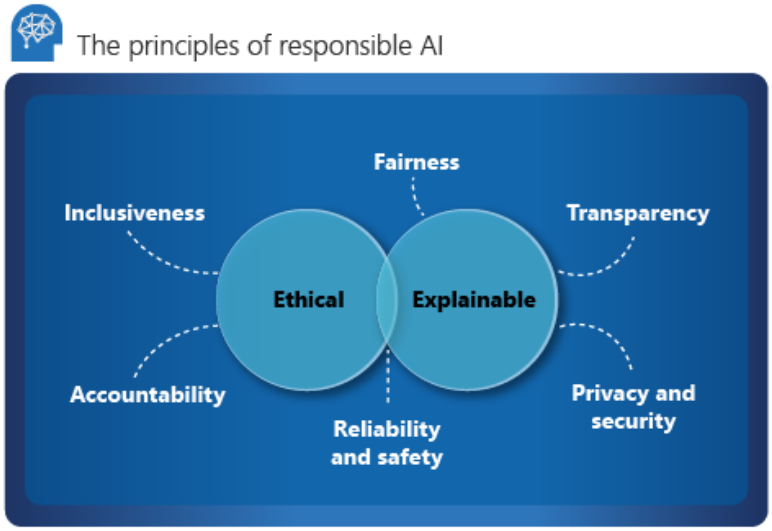
Σε κάθε περίπτωση, κατά την ανάπτυξη μίας εφαρμογής μηχανικής μάθησης χρειάζεται να λαμβάνονται υπόψιν ορισμένες θεμελιώδεις αρχές. Σύμφωνα με τη Microsoft, υπάρχουν έξι θεμελιώδεις αρχές για υπεύθυνα και σοβαρά συστήματα τεχνητής νοημοσύνης. Ένα υπεύθυνο και σοβαρό σύστημα τεχνητής νοημοσύνης, σημαίνει ότι έχει γίνει σχεδιασμός, ανάπτυξη και εφαρμογή της τεχνητής νοημοσύνης με καλή διάθεση, έχοντας ως στόχο την ενδυνάμωση των εργαζομένων και των επιχειρήσεων, με ταυτόχρονη δίκαιη επίδραση στα μέλη της κοινωνίας, ώστε οι εταιρείες να δημιουργούν ασφάλεια μέσω της εφαρμογής των συστημάτων της τεχνητής νοημοσύνης (Accenture, 2022).

Η πρώτη αρχή είναι η αρχή της αμεροληψίας (Fairness Principle). Μέσω της αρχής αυτής, μελετάται η υπευθυνότητα του συστήματος στην επιβεβαίωση ότι οι αλγόριθμοι και τα δεδομένα είναι όσο το δυνατόν αμερόληπτα και όσο το δυνατόν αντιπροσωπευτικά (Rao, 2021). Αυτό το οποίο επιδιώκεται είναι ο έλεγχος ότι οι αποφάσεις των συστημάτων δεν κάνουν διακρίσεις βάσει φύλου, φυλής, σεξουαλικού προσανατολισμού και θρησκείας των ανθρώπων.

Η δεύτερη είναι η αρχή της αξιοπιστίας και της ασφάλειας (Reliability and Safety Principle). Ένα σύστημα ΤΝ για να μπορέσει να είναι έμπιστο χρειάζεται να είναι πρωτίστως αξιόπιστο αλλά και ασφαλές. Έτσι, το σύστημα αυτό επιβάλλεται να λειτουργεί ακριβώς όπως

είχε σχεδιαστεί να λειτουργήσει. Με το πέρας του χρόνου, η απόδοση ενός συστήματος ενδέχεται να ελαττωθεί. Ως εκ τούτου, είναι αναγκαία μία ισχυρή διαδικασία εντοπισμού και ελέγχου του μοντέλου, ώστε να αξιολογείται η απόδοσή του και σε περιπτώσεις όπου κρίνεται αναγκαίο να γίνεται η επανεκπαίδευσή του (Microsoft, 2021).

Η τρίτη είναι η αρχή της ιδιωτικότητας και της ασφάλειας (Privacy and Security Principle). Μία βάση δεδομένων οφείλει να προστατεύει τα δεδομένα της, με τέτοιον τρόπο ώστε να μην προσβάλλεται η ιδιωτικότητα των ατόμων. Ένας τρόπος προστασίας είναι η προσθήκη θορύβου στα δεδομένα και η εφαρμογή τυχαιοποίησής τους, με σκοπό τη δυσκολότερη ανάκτηση πληροφοριών από τρίτους.



Σχήμα 4 *Αρχές υπεύθυνου συστήματος ΤΝ, (Microsoft, 2021)*

Η τέταρτη είναι η αρχή της «ευρύτητας» (Inclusiveness Principle). Διάφορες τεχνικές όπως speech-to-text, text-to-speech και οπτική αναγνώριση, χρησιμοποιούνται ώστε να μπορούν όλοι οι άνθρωποι να βρίσκονται σε θέση να χρησιμοποιούν τα συστήματα αυτά. Με αυτόν τον τρόπο δεν αποκλείονται άνθρωποι οι οποίοι ανήκουν σε μειονότητες, λόγω κάποιου πιθανού προβλήματός τους, όπως για παράδειγμα πρόβλημα ακοής.

Η πέμπτη είναι η αρχή της διαφάνειας (Transparency Principle). Μέσω της διαφάνειας γίνονται κατανοητά τόσο τα δεδομένα όσο και οι αλγόριθμοι οι οποίοι χρησιμοποιούνται για να εκπαιδεύσουν το μοντέλο, αλλά και το τελικό μοντέλο το οποίο προκύπτει. Με αυτόν τον τρόπο, χτίζεται εμπιστοσύνη.

Τέλος, η έκτη και τελευταία αρχή είναι η αρχή της ανάληψης ευθυνών (Accountability Principle). Οι σχεδιαστές των συστημάτων ΤΝ χρειάζεται να είναι υπεύθυνοι για τις αποφάσεις τους και τις πράξεις τους, ειδικότερα όσον αφορά αυτόματα/ανεξάρτητα συστήματα.

3.3 Μηχανική Μάθηση

Όπως αναφέρθηκε στην προηγούμενη υποενότητα, η Μηχανική Μάθηση αποτελεί ένα υπό-πεδίο της Τεχνητής Νοημοσύνης και αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην Τεχνητή Νοημοσύνη.

Η ιστορία της Μηχανικής Μάθησης ξεκινάει αρκετά χρόνια πριν, όταν το 1959 ο Arthur Samuel, Αμερικανός καθηγητής του πανεπιστημίου Stanford και πρωτοπόρος στα πεδία Τεχνητής Νοημοσύνης και Παιγνίων Υπολογιστών, εφηύρε τον όρο, παρουσιάζοντας ένα πρόγραμμα υπολογιστή για το παιχνίδι της Ντάμας, το οποίο και θεωρείται ως το πρώτο

πρόγραμμα «αυτομάθησης» υπολογιστή (Ανδρικάκης, 2017). Σύμφωνα με τον A. Samuel «*Οι υπολογιστές αποκτούν την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί*».

Συνεπώς, στο πεδίο της Μηχανικής Μάθησης μελετώνται και κατασκευάζονται αλγόριθμοι οι οποίοι αντλούν πληροφορίες από, πειραματικά, δεδομένα και βάσει των συγκεκριμένων δεδομένων ως αποτέλεσμα των αλγορίθμων είτε θα κάνουν προβλέψεις είτε θα ανασύρουν αποφάσεις (Χρυσοπούλου, 2019). Η δημιουργία μοντέλων πρόβλεψης με χρήση μηχανικής μάθησης ασχολείται με την ελαχιστοποίηση σφαλμάτων ενός μοντέλου και με τη λογική της δημιουργίας μοντέλων τα οποία θα προβλέπουν με ακρίβεια μία μεταβλητή. Έτσι, στην εφαρμοσμένη μηχανική μάθηση, δανείζονται και χρησιμοποιούνται αλγόριθμοι οι οποίοι προέρχονται από άλλα πεδία, όπως η στατιστική (Brownlee, 2016).

Σε αυτό το σημείο, κρίνεται να τονιστεί η αναγκαιότητα ύπαρξης δεδομένων. Ένα πρόγραμμα αποκτάει εμπειρία και μαθαίνει από δεδομένα. Θα μπορούσε κάποιος να πει με βεβαιότητα ότι τα δεδομένα αποτελούν το καύσιμο της Μηχανικής Μάθησης, συνεπώς χωρίς δεδομένα δεν υφίσταται Μηχανική Μάθηση. Ένα μοντέλο Μηχανικής Μάθησης μαθαίνει, λοιπόν, μέσω των ιστορικών δεδομένων τα οποία εισάγονται σε αυτό και εν συνεχεία, μέσω των αλγορίθμων πρόβλεψης γίνεται η πρόβλεψη ενός προϊόντος για ένα καινούργιο σύνολο δεδομένων. Η ακρίβεια των μοντέλων εξαρτάται από την ποιότητα και την ποσότητα των δεδομένων εισόδου. Συνήθως, ένα μεγαλύτερο πλήθος δεδομένων συντελεί στη δημιουργία ενός ισχυρότερου μοντέλου το οποίο προβλέπει το τελικό προϊόν αρκετά αποτελεσματικά (Great Learning Team, 2022). Να σημειωθεί, όμως, ότι μεγάλο πλήθος δεδομένων δεν αποτελεί πάντοτε πανάκεια, οπότε χρειάζεται να ληφθούν υπόψιν αρκετοί ακόμη παράγοντες οι οποίοι επηρεάζουν την ποιότητα των αποτελεσμάτων ενός μοντέλου. Άλλωστε, η χρήση σημαντικού όγκου δεδομένων απαιτεί και την αντίστοιχη υπολογιστική ισχύ, η οποία ενδέχεται να μην είναι διαθέσιμη. Εκτός από την ισχύ, απαιτείται και ο κατάλληλος χρόνος, τόσο για την προ-επεξεργασία των δεδομένων από το χρήστη, όσο και για την επεξεργασία των δεδομένων από τη μηχανή.

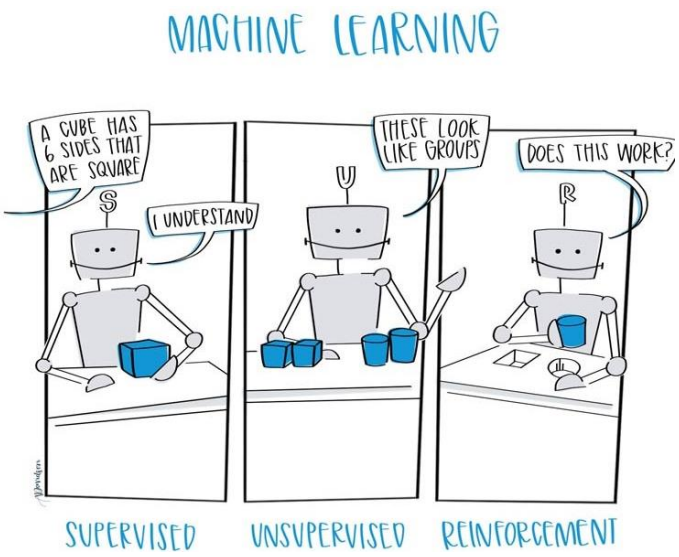
Ορισμένα στοιχεία τα οποία χρειάζεται να ληφθούν επιπροσθέτως υπόψιν για την ακρίβεια ενός μοντέλου, είναι οι εκάστοτε αλγόριθμοι οι οποίοι επιλέγονται να χρησιμοποιηθούν, αλλά και οι μετρικές αξιολόγησης οι οποίες συνοδεύουν τα τελικά αποτελέσματα. Λόγος γι' αυτά τα δύο στοιχεία πρόκειται να γίνει εκτενέστερα σε ακόλουθες υποενότητες.

3.3.1 Τύποι Μηχανικής Μάθησης

Μία εργασία Μηχανικής Μάθησης δύναται να ταξινομηθεί, ανάλογα με το εάν τα αρχικά δεδομένα έχουν ετικέτα, ή όχι, σε μία από τις ακόλουθες τρεις κατηγορίες[1]: Επιβλεπόμενη Μάθηση (Supervised Learning), Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning) και Ενισχυτική Μάθηση (Reinforcement Learning).

Στις επόμενες γραμμές ακολουθούν περισσότερες λεπτομέρειες για έκαστη κατηγορία με σκοπό την αρτιότερη επεξήγηση και κατανόηση.

[1] Να σημειωθεί σε αυτό το σημείο ότι συναντώνται επιπρόσθετοι τύποι Μηχανικής Μάθησης, λόγου χάριν Semi-supervised Learning, Self-supervised Learning, Active Learning κ.λπ., εν τούτοις στο πλαίσιο της τρέχουσας διπλωματικής εργασίας επιλέγεται να γίνει ανάλυση των τριών ευρέως χρησιμοποιούμενων και γνωστότερων τύπων.



Σχήμα 5 *Τύποι Μηχανικής Μάθησης, (sketchalytics, 2019)*

3.3.1.1 *Επιβλεπόμενη Μάθηση – Supervised Learning*

Ένας από τους βασικότερους τύπους Μηχανικής Μάθησης είναι η επιβλεπόμενη μηχανική μάθηση και γι' αυτό το λόγο, έχει συχνή εφαρμογή. Ο εν λόγω τύπος μάθησης, μπορεί να χρησιμοποιηθεί κατά κύριο λόγο σε προβλήματα πρόβλεψης (prediction), ταξινόμησης (classification) και διερμηνείας (interpretation). Στην επιβλεπόμενη μάθηση, ένας αλγόριθμος εκπαιδεύεται σε δεδομένα τα οποία έχουν ετικέτα. Συγκεκριμένα, παρατηρούνται δύο τύποι μεταβλητών. Τα labels, δηλαδή η μεταβλητή η οποία προσδοκάται να προβλεφθεί και τα features, δηλαδή οι μεταβλητές οι οποίες βοηθούν στην πρόβλεψη της μεταβλητής με ετικέτα.

Η εκπαίδευση του αλγορίθμου γίνεται με ένα μικρό εκπαιδευτικό σύνολο δεδομένων (training set), τα οποία αποτελούν μέρος του μεγαλύτερου συνόλου δεδομένων. Έτσι, το μοντέλο είναι σε θέση να εντοπίσει το υποβόσκων μοτίβο ανάμεσα στα δεδομένα, καθώς γνωρίζει τόσο τις μεταβλητές features, όσο και τη μεταβλητή label η οποία συσχετίζεται με τα features (Great Learning Team, 2022).

Επιστρέφοντας στο σύνολο των δεδομένων εκπαίδευσης, το οποίο αποτελείται από δεδομένα εισόδου (inputs) και δεδομένα εξόδου (outputs), εκείνο έχει παρόμοια χαρακτηριστικά με το τελικό σύνολο δεδομένων και παρέχει στον αλγόριθμο όλες τις παραμέτρους οι οποίες απαιτούνται για την επίλυση του προβλήματος. Άπαξ και ο αλγόριθμος εντοπίσει τη σχέση/μοτίβο μεταξύ των μεταβλητών, πλέον γνωρίζει τον τρόπο με τον οποίο λειτουργούν τα δεδομένα και ποια είναι η σχέση ανάμεσα στα δεδομένα εισόδου και εξόδου. Η λύση αναπτύσσεται, καθώς γίνεται χρήση του τελικού συνόλου δεδομένων, από το οποίο ο αλγόριθμος μαθαίνει με τον ίδιο τρόπο όπως και στην περίπτωση του εκπαιδευτικού συνόλου δεδομένων. Έτσι, το μοντέλο έχει δυνατότητα να βελτιώνεται ακόμη και μετά την ανάπτυξή του, με την πάροδο του χρόνου, καθώς βρίσκεται σε θέση να ανακαλύψει νέα μοτίβα και σχέσεις τα οποία προκύπτουν από τα νέα δεδομένα (Ελευθερίου, 2021).

3.3.1.2 *Μη επιβλεπόμενη Μάθηση – Unsupervised Learning*

Στην περίπτωση της μη επιβλεπόμενης μάθησης τα δεδομένα δεν έχουν ετικέτα. Ο εν λόγω τύπος μάθησης χρησιμοποιείται κατά κύριο λόγο σε προβλήματα ομαδοποίησης (clustering), σε προβλήματα ανάλυσης συσχετισμών (association analysis), αλλά και σε προβλήματα μείωσης διαστάσεων (dimensionality reduction) (Κοτρωνάκη, 2021). Σε αυτήν την

περίπτωση, δεν απαιτείται ανθρώπινη εργασία, ώστε να γνωστοποιήσει το σύνολο δεδομένων στη μηχανή.

Οι αλγόριθμοι βρίσκονται σε θέση να ανακαλύψουν μόνοι τους, με αφηρημένο τρόπο και χωρίς ανθρώπινη παρέμβαση, κρυφές δομές, όπως ομαδοποιήσεις ή μοτίβα, δεδομένων. Οι αλγόριθμοι της μη επιβλεπόμενης μάθησης έχουν την ικανότητα να προσαρμόζονται στα δεδομένα, γεγονός το οποίο επιφέρει ως αποτέλεσμα την δυναμική αλλαγή των κρυφών δομών και γεγονός το οποίο συντελεί σε ικανότητα για περισσότερη ανάπτυξη (Ελευθερίου, 2021).

3.3.1.3 *Ενισχυτική Μάθηση – Reinforcement Learning*

Η περίπτωση της ενισχυτικής μάθησης είναι εμπνευσμένη από τον τρόπο με τον οποίο οι άνθρωποι μαθαίνουν από δεδομένα. Ο αλγόριθμος της ενισχυτικής μάθησης βελτιώνεται και μαθαίνει από νέες συνθήκες μέσω της δοκιμής και του σφάλματος. Εν αντιθέσει με τις δύο προηγούμενες τεχνικές μάθησης, οι οποίες είχαν σαν στόχο την επισήμανση και την ομαδοποίηση δεδομένων, αντίστοιχα, η ενισχυτική μάθηση επιδιώκει την επίτευξη ενός επιθυμητού στόχου. Συγκεκριμένα, ένα ευνοϊκό αποτέλεσμα ενισχύεται, διαφορετικά ένα μη ευνοϊκό αποτέλεσμα αποθαρρύνεται. Ο διερμηνέας έπειτα από έκαστη επανάληψη του αλγορίθμου, αποφασίζει εάν το αποτέλεσμα είναι ευνοϊκό ή όχι. Στην περίπτωση, λοιπόν, κατά την οποία το αποτέλεσμα είναι ευνοϊκό, τότε ο διερμηνέας ανταμείβει τον αλγόριθμο. Διαφορετικά, ο αλγόριθμος «αναγκάζεται» να επαναλάβει τα βήματά του μέχρι να βγάλει ένα ευνοϊκό αποτέλεσμα. Όπως εύκολα μπορεί να γίνει αντιληπτό, στόχος του αλγορίθμου είναι να μεγιστοποιήσει τα ευνοϊκά αποτελέσματα, άρα και τη συνολική ανταμοιβή. Επιπροσθέτως, αξίζει να σημειωθεί ότι αρκετές φορές δεν υπάρχει απόλυτη λύση. Στις περιπτώσεις αυτές, χρησιμοποιείται μία βαθμολογία η οποία αφορά την αποτελεσματικότητα και η οποία εκφράζεται σε μία ποσοστιαία τιμή. Συνεπώς, μεγαλύτερη τιμή της ποσοστιαίας τιμής, συνεπάγεται και μεγαλύτερη ανταμοιβή (Ελευθερίου, 2021).

Τέλος, κρίνεται σκόπιμο να τονισθεί ότι εκτός από την «κλασική» ενισχυτική μάθηση, συναντάται και η βαθιά ενισχυτική μάθηση η οποία συνδυάζει στοιχεία της ενισχυτικής μάθησης και της βαθιάς μάθησης και στην οποία οι είσοδοι μεταβάλλονται ενδελεχώς (Κοτρωνάκη, 2021).

3.3.2 Αλγόριθμοι Μηχανικής Μάθησης

Όπως αναφέρθηκε και στην 3.3.1, διαχωρισμός της μηχανικής μάθησης γίνεται βάσει του τύπου της εξόδου τον οποίο φέρει ένα μοντέλο, στις πέντε ακόλουθες κατηγορίες: Ταξινόμηση (Classification), Παλινδρόμηση (Regression), Ομαδοποίηση (Clustering), Συσχέτιση (Association) και Μείωση Διαστάσεων (Dimensionality Reduction).

Ένας όρος ο οποίος χρειάζεται να «συστηθεί» σε αυτό το σημείο είναι η εξόρυξη δεδομένων (data mining). Η εξόρυξη δεδομένων είναι η διαδικασία κατά την οποία εντοπίζονται ανωμαλίες, μοτίβα και συσχετίσεις ανάμεσα σε ένα πλήθος σετ δεδομένων/βάσεις δεδομένων, με σκοπό την πρόβλεψη αποτελεσμάτων.

Στην περίπτωση της επιβλεπόμενης μάθησης, συναντώνται δύο τύποι προβλημάτων κατά τη διαδικασία εξόρυξης δεδομένων. Πρόκειται για τα προβλήματα ταξινόμησης και για τα προβλήματα παλινδρόμησης.

Στην περίπτωση της μη επιβλεπόμενης μάθησης, συναντώνται τα προβλήματα παλινδρόμησης, τα προβλήματα συσχέτισης και τα προβλήματα μείωσης διαστάσεων, όπως αναφέρθηκε και σε προηγούμενο εδάφιο (3.3.1). Στη συνέχεια, ακολουθεί ανάλυση εκάστης κατηγορίας καθώς επίσης και ανάλυση των αλγορίθμων οι οποίοι συναντώνται για κάθε κατηγορία.

Αναλυτικότερα:

3.3.2.1 Ταξινόμηση – Classification

Στα προβλήματα ταξινόμησης σκοπός είναι ο καθορισμός της κατηγορίας ενός αντικειμένου.

Συγκεκριμένα, χρησιμοποιείται ένας αλγόριθμος μέσω του οποίου γίνεται εκχώρηση δεδομένων ελέγχου σε καθορισμένες κλάσεις. Ο αλγόριθμος αυτός, από το σύνολο δεδομένων εκπαιδεύεται να αναγνωρίζει συγκεκριμένες οντότητες και επιδιώκει να εξάγει συμπεράσματα αναφορικά με τον τρόπο με τον οποίο οι οντότητες αυτές πρέπει να επισημανθούν ή να οριστούν, ώστε να ενταχθούν σε μία συγκεκριμένη κατηγορία (IBM Cloud Education, 2020).

Εκείνοι οι αλγόριθμοι οι οποίοι χρησιμοποιούνται κατά κύριο λόγο για τα προβλήματα ταξινόμησης είναι: Linear Classifiers, Support Vector Machines (SVM), Decision Trees, K-Nearest Neighbor και Random Forest.

3.3.2.2 Παλινδρόμηση – Regression

Στα προβλήματα παλινδρόμησης σκοπός είναι η πρόβλεψη μελλοντικών τιμών μέσω ενός μοντέλου το οποίο εκπαιδεύεται με ιστορικά δεδομένα. Συγκεκριμένα, η παλινδρόμηση χρησιμοποιείται με σκοπό την κατανόηση της συσχέτισης, των πιθανών σχέσεων, ανάμεσα στις εξαρτημένες και τις ανεξάρτητες μεταβλητές (IBM Cloud Education, 2020).

Εκείνοι οι αλγόριθμοι οι οποίοι χρησιμοποιούνται κατά κύριο λόγο για τα προβλήματα παλινδρόμησης είναι: Linear Regression, Logistic Regression, Polynomial Regression, Support Vector Regression, Lasso Regression, Gaussian Process Regression, Random Forest Regression και XGBoost Regression.

3.3.2.3 Ομαδοποίηση – Clustering

Στα προβλήματα ομαδοποίησης σκοπός είναι η ομαδοποίηση μεταβλητών με όμοια χαρακτηριστικά, σε ομάδες οι οποίες δεν είναι γνωστές εκ των προτέρων. Ευρέως χρησιμοποιούμενος αλγόριθμος σε προβλήματα ομαδοποίησης είναι ο K-means.

3.3.2.4 Συσχέτιση – Association

Στα προβλήματα συσχέτισης σκοπός είναι η εύρεση σχέσεων ανάμεσα στις μεταβλητές ενός συνόλου δεδομένων. Σε αυτήν την περίπτωση, ο αλγόριθμος εκείνος ο οποίος ξεχωρίζει και εφαρμόζεται συχνότερα είναι ο Apriori (Κοτρωνάκη, 2021).

3.3.2.5 Μείωση Διαστάσεων – Dimensionality Reduction

Τα προβλήματα μείωσης διαστάσεων συναντώνται όταν ο αλγόριθμος των διαστάσεων ή των χαρακτηριστικών ενός συνόλου δεδομένων είναι αρκετά υψηλός. Όπως έχει ήδη αναφερθεί, ένα μεγάλο πλήθος δεδομένων συνήθως δύναται να συντελέσει σε αύξηση ακρίβειας των αποτελεσμάτων, εν τούτοις μπορεί να παρατηρηθεί υπερβολική προσαρμογή του μοντέλου στα δεδομένα (overfitting), καθώς επίσης να δημιουργεί προβλήματα ως προς την οπτικοποίηση του συνόλου δεδομένων. Συνεπώς, στόχος της μείωσης διαστάσεων είναι η μείωση του αριθμού των δεδομένων εισόδου σε τέτοιο μέγεθος το οποίο να είναι διαχειρίσιμο από το μοντέλο. Εκείνοι οι αλγόριθμοι οι οποίοι χρησιμοποιούνται κατά κύριο λόγο για τα προβλήματα παλινδρόμησης είναι: Principal Component Analysis, Singular Value Decomposition και Autoencoders (Κοτρωνάκη, 2021).

Για την παρούσα εργασία, όπου συναντάται ένα πρόβλημα παλινδρόμησης, οι αλγόριθμοι οι οποίοι χρησιμοποιήθηκαν για την αναζήτηση τόσο του αριθμού των κρουσμάτων όσο και του αριθμού των θανάτων είναι ορισμένοι από όσους αναφέρθηκαν στο εδάφιο της Παλινδρόμησης. Δηλαδή, πρόκειται για τους αλγορίθμους Plain Linear Regression, Support Vector Regression, Lasso Regression, Gaussian Process Regression, Random Forest Regression και XGBoost Regression. Ως εκ τούτου, κρίνεται αναγκαίο να

πραγματοποιηθεί περαιτέρω ανάλυση των συγκεκριμένων αλγορίθμων, με σκοπό την καλύτερη ερμηνεία των τελικών αποτελεσμάτων στο Κεφάλαιο 4.

3.3.2.6 *Linear Regression*

Πρώτη μέθοδος η οποία αναλύεται είναι η μέθοδος της γραμμικής παλινδρόμησης, Linear Regression. Η γραμμική παλινδρόμηση είναι από τους πιο γνωστούς, απλούς και κατανοητούς αλγορίθμους τόσο στον κλάδο της στατιστικής όσο και στον κλάδο της μηχανικής μάθησης.

Συναντώνται δύο περιπτώσεις γραμμικής παλινδρόμησης. Στην περίπτωση της απλής γραμμικής παλινδρόμησης (simple linear regression), υπάρχει μία μόνο ανεξάρτητη μεταβλητή X (input/independent variable) και η εξαρτημένη Y (dependent variable) και αναζητείται η σχέση η οποία συναντάται ανάμεσα στις δύο αυτές μεταβλητές. Όταν συναντάται μία μόνο είσοδος, τότε μέσω στατιστικών μεγεθών εκτιμώνται οι διάφοροι συντελεστές. Τέτοια στατιστικά μεγέθη είναι η μέση τιμή, η τυπική απόκλιση, οι συσχετίσεις και οι συμμεταβλητότητες (Brownlee, 2016). Η σχέση ανάμεσα στις δύο μεταβλητές X και Y, ερμηνεύεται με ένα απλό γραμμικό μοντέλο, όπως περιγράφεται στην εξίσωση (1.1):

$$Y_i = b_0 + b_1 X_i + \varepsilon_i, \quad i = 1,2,\dots,n \qquad (1.1)$$

Όπου b_0, b_1 δύο άγνωστες σταθερές και $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_v$ ανεξάρτητες τυχαίες μεταβλητές, οι οποίες ονομάζονται σφάλματα των μετρήσεων και ακολουθούν κανονική κατανομή $N(0,\sigma^2)$, με άγνωστο σ^2.

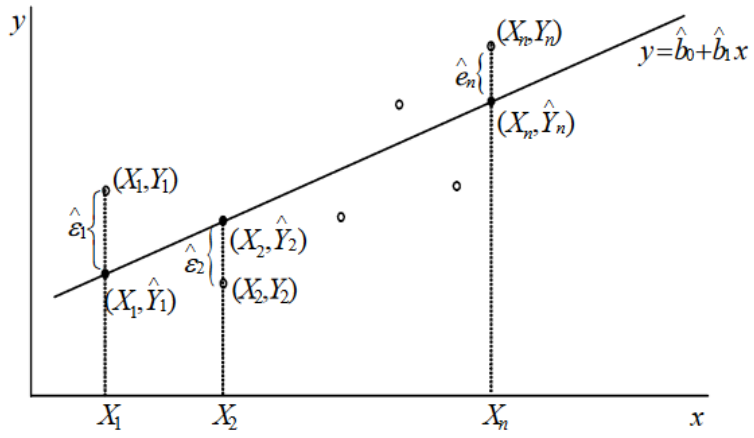
Βάσει των (X_i,Y_i) εκτιμώνται οι καλύτερες τιμές των παραμέτρων b_0, b_1 και σ^2. Έτσι, αφού εκτιμηθούν οι τιμές αυτές, τότε υπολογίζονται οι προβλέψεις των Y_i (Y predicted) ως:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i \qquad (1.2)$$

Η διαφορά ανάμεσα στην πραγματική τιμή και στην προβλεπόμενη τιμή ονομάζεται υπόλοιπο (residuals) και υπολογίζεται όπως παρουσιάζεται στην εξίσωση (1.3):

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i \qquad (1.3)$$

Μέσα από την ανάλυση απλής γραμμικής παλινδρόμησης προκύπτει η εκτιμημένη ευθεία γραμμικής παλινδρόμησης "$y = \hat{b}_0 + \hat{b}_1 x$", καθώς επίσης και τα υπόλοιπα $\hat{\varepsilon}_i$. Το Σχήμα το οποίο ακολουθεί, περιγράφει πλήρως ένα αποτέλεσμα μοντέλου απλής γραμμικής παλινδρόμησης.



Σχήμα 6 *Μοντέλο απλής γραμμικής παλινδρόμησης,* (Boutsikas, Ενότητα 5: Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression), 2004)

Στη δεύτερη περίπτωση, δηλαδή στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης (multiple linear regression), συναντώνται πολλαπλές ανεξάρτητες μεταβλητές X_1, X_2,…, X_{p-1}, από τις οποίες εξαρτάται η μεταβλητή Y. Στην περίπτωση αυτή, το μοντέλο το οποίο εφαρμόζεται είναι:

$$Y_i = b_0 + b_1 X_{i,1} + b_2 X_{i,2} + \cdots + b_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1,2,\dots,n \qquad (1.4)$$

Όπου τα $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_\nu$ ακολουθούν κανονική κατανομή N(0,1).

Το μοντέλο της εξίσωσης (1.4) μπορεί να γραφτεί σε μορφή πινάκων ως εξής:

$$\boldsymbol{Y} = \boldsymbol{Xb} + \varepsilon \qquad (1.5)$$

Όπου:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1,p-1} \\ 1 & X_{21} & & X_{2,p-1} \\ \vdots & & & \\ 1 & X_{n1} & \dots & X_{n,p-1} \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{p-1} \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Σε αυτήν την περίπτωση, τα διάφορα σημεία δεν «προσεγγίζουν» μία ευθεία, αλλά ένα υπερεπίπεδο με διάσταση p (Boutsikas, 2004).

3.3.2.7 Support Vector Regression

Η δεύτερη μέθοδος η οποία αναλύεται, είναι η Support Vector Regression. Λειτουργεί με παρόμοιο τρόπο όπως η μέθοδος Support Vector Machine σε προβλήματα ταξινόμησης.

Μία πιθανή ύπαρξη μη γραμμικότητας στα δεδομένα μπορεί να εντοπιστεί από τον αλγόριθμο. Με αυτόν τον τρόπο, είναι εφικτή η καλύτερη πρόβλεψη του μοντέλου. Ως υπερεπίπεδο ορίζεται εκείνη η ευθεία η οποία προσαρμόζεται στα δεδομένα. Τα σημεία εκείνα τα οποία βρίσκονται κοντά στο υπερεπίπεδο ονομάζονται "Support Vectors". Για να μπορέσει να εντοπιστεί ένα υπερεπίπεδο, μπορεί να χρησιμοποιηθεί ένα σετ μαθηματικών συναρτήσεων, γνωστό και ως kernels (Raj, 2020).

Στην περίπτωση του SVR, το ενδιαφέρον εστιάζεται στο σφάλμα και όχι τόσο στην πρόβλεψη. Συγκεκριμένα, επιδιώκεται το σφάλμα να είναι μικρότερο από ένα κατώφλι. Δηλαδή, η ελάχιστη απόσταση ανάμεσα στα σημεία εκπαίδευσης και στο υπερεπίπεδο μεγιστοποιείται από τον αλγόριθμο SVR (Singh, 2019).

Συναντώνται διάφορα είδη kernels τα γραμμικά, τα μη γραμμικά, τα πολυωνυμικά, τα Radial Basis Function (RBF) και τα σιγμοειδή. Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν γραμμικά kernels.

Τέλος, να σημειωθεί ότι αναμένεται οι προβλέψεις των SVR μοντέλων να είναι καλύτερες από εκείνες των μοντέλων απλής γραμμικής παλινδρόμησης. Αυτό συμβαίνει καθώς οι προβλεπόμενες τιμές των πρώτων βρίσκονται πιο κοντά στις πραγματικές τιμές (Singh, 2019).

3.3.2.8 LASSO Regression

Ακολουθεί η μέθοδος LASSO Regression. Τα αρχικά του προκύπτουν από το **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. Αποτελεί μία τεχνική κανονικοποίησης και πρόκειται για ένα μοντέλο το οποίο συρρικνώνει τις τιμές και. Γενικά, αυτό το οποίο ισχύει είναι ότι εάν ένα μοντέλο παλινδρόμησης χρησιμοποιεί την τεχνική κανονικοποίησης L1, τότε πρόκειται για παλινδρόμηση LASSO. Εάν χρησιμοποιεί την L2 τεχνική κανονικοποίησης, τότε πρόκειται για παλινδρόμηση Ridge (DataCamp Team, 2022).

Στη L1 κανονικοποίηση, οι συντελεστές οι οποίοι υπολογίζονται από το γραμμικό μοντέλο, συσσωρεύονται γύρω από ένα κεντρικό σημείο, όπως είναι ο μέσος. Έτσι, προστίθεται μία «ποινή» ίση με το απόλυτο μέγεθος του συντελεστή, η οποία ονομάζεται «α» ή λ. Συνεπώς, είναι αρκετά πιθανή η ύπαρξη λιγότερων συντελεστών, καθώς ορισμένοι από τους συντελεστές γίνονται ίσοι με το μηδέν. Οι τιμές συντελεστών οι οποίες βρίσκονται πιο κοντά στο μηδέν είναι περισσότερο επιρρεπείς σε «ποινές» (Great Learning Team, 2021).

Η μαθηματική εξίσωση για το LASSO, περιγράφεται παρακάτω. Πρόκειται για το τετραγωνικό άθροισμα των υπολοίπων με προσθήκη του λ επί το σύνολο της απόλυτης τιμής του μεγέθους των συντελεστών.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \qquad (1.6)$$

Όπου, λ φανερώνει το ποσό της συρρίκνωσης.

Όταν το λ=0, τότε λαμβάνονται υπόψιν όλες οι μεταβλητές και πρόκειται για ένα μοντέλο αντίστοιχο με εκείνο της γραμμικής παλινδρόμησης. Σε αυτήν την περίπτωση, το μοντέλο δημιουργείται μόνο με το τετράγωνο του αθροίσματος των υπολοίπων. Όταν το λ μεγαλώνει, λαμβάνονται υπόψιν λιγότερες μεταβλητές και όταν γίνεται άπειρο, τότε δεν λαμβάνονται υπόψιν οι μεταβλητές. Τέλος, μία αύξηση στο λ συντελεί σε αύξηση μεροληψίας και σε μείωση διασποράς.

3.3.2.9 Gaussian Process Regression

Όπως αναφέρει η Hilarie Sit στο άρθρο της, *«To Gaussian Process Regression είναι μία μη παραμετρική, Μπεϊζιανή προσέγγιση στην παλινδρόμηση»*. Ακόμη, θεωρείται ότι λειτουργεί καλά σε μικρά σετ δεδομένων (Sit, 2019).

Εφόσον πρόκειται για ένα μη παραμετρικό μοντέλο, ο υπολογισμός της κατανομής πιθανοτήτων υπολογίζεται για όλες τις συναρτήσεις οι οποίες εικάζεται ότι προσαρμόζονται στα δεδομένα. Δεν γίνεται, δηλαδή, υπολογισμός της κατανομής πιθανότητας για παραμέτρους μίας συγκεκριμένης συνάρτησης.

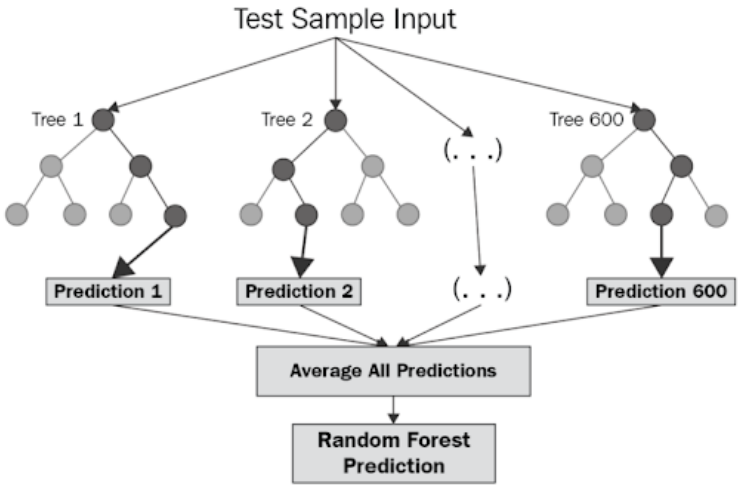
Πρόκειται για ένα μοντέλο το οποίο επεκτείνει με απλοϊκό τρόπο το μοντέλο της γραμμικής παλινδρόμησης. Ακόμη, μπορεί να παρέχει εκτιμητές αβεβαιότητας μαζί με τις προβλεπόμενες τιμές.

3.3.2.10 Random Forest Regression

Ο αλγόριθμος της επιβλεπόμενης μάθησης Random Forest Regression χρησιμοποιεί την τεχνική "ensemble learning". Δηλαδή, πρόκειται για μία τεχνική μέσω της οποίας συνδυάζονται από πολλαπλούς αλγορίθμους μηχανικής μάθησης οι προβλέψεις. Αυτό, όπως εύλογα μπορεί να γίνει αντιληπτό, οδηγεί σε καλύτερα και ακριβέστερα μοντέλα πρόβλεψης (Bakshi, 2020).

Το μοντέλο Random Forest αποτελείται από ένα συνονθύλευμα δένδρων αποφάσεων. Πρόκειται για τη μέθοδο "bagging" και όχι "boosting", του ensemble learning. Εξ ορισμού, η μέθοδος bagging λαμβάνει χώρα όταν κάθε μοντέλο τρέχει ανεξάρτητα από τα υπόλοιπα και στη συνέχεια, αθροίζει τα τελικά αποτελέσματα τους. Συνεπώς, τα δέντρα στο Random Forest, δεν έχουν καμία αλληλεπίδραση μεταξύ τους όσο υλοποιούνται, αλλά ενώνονται μεταξύ τους στο τέλος, ώστε να δημιουργηθεί μία πιο ακριβής και σταθερή πρόβλεψη. Σε προβλήματα παλινδρόμησης, το τελικό αποτέλεσμα προκύπτει ως ο μέσος όρος των μεμονωμένων δένδρων (Chakure, 2022).

Στο παρακάτω Σχήμα γίνεται αντιληπτός ο τρόπος λειτουργίας του Random Forest.



Σχήμα 7 *Δομή Random Forest, (Chakure, 2022)*

Θετικά του αλγορίθμου είναι ότι μπορεί να χρησιμοποιηθεί τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης. Επιπλέον, πρόκειται για έναν εύχρηστο αλγόριθμο ο οποίος εξαγάγει αρκετά ικανοποιητικά αποτελέσματα. Τέλος, οι πιθανότητες να υπερ-προσαρμοστεί το μοντέλο στα πραγματικά δεδομένα ελαττώνονται, όμως δεν μηδενίζονται. Από την άλλη, αρνητικά του είναι ότι ενδέχεται να εμφανίσει αργή απόδοση, λόγω μεγάλου πλήθους δένδρων. Επίσης, σε περιπτώσεις κατά τις οποίες συναντώνται σετ δεδομένων με πολύ θορυβώδεις διαδικασίες σε προβλήματα ταξινόμησης ή παλινδρόμησης, υπάρχει περίπτωση να γίνει υπερ-προσαρμογή (Donges, 2022).

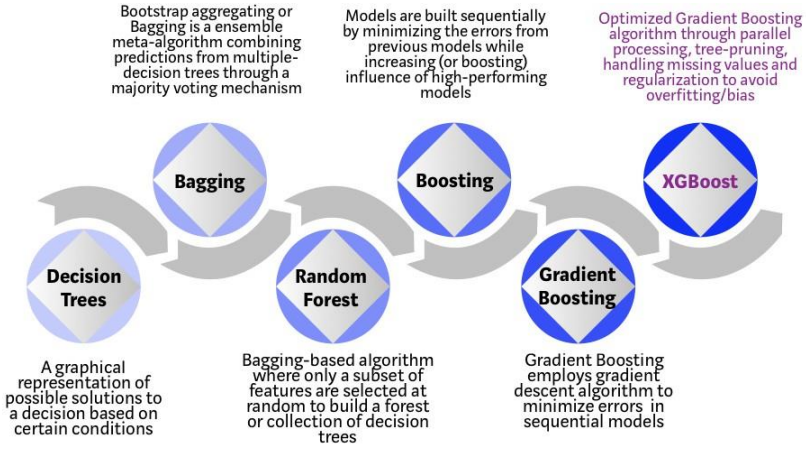
3.3.2.11 XGBoost Regression

Ο αλγόριθμος XGBoost χρησιμοποιεί την τεχνική "ensemble learning". Εν αντιθέσει με τον αλγόριθμο RF, η τεχνική η οποία χρησιμοποιείται είναι η τεχνική "gradient boosting". Είναι γεγονός ότι σε περιπτώσεις κατά τις οποίες συναντώνται μικρά ή μεσαία δεδομένα, οι αλγόριθμοι οι οποίοι έχουν βάσει στα δένδρα αποφάσεων φαίνεται να αποδίδουν καλύτερα (Morde, 2019). Η ιστορία του αλγορίθμου χρονολογεί λίγα χρόνια. Συγκεκριμένα, αναπτύχθηκε σε ένα ερευνητικό project από το πανεπιστήμιο της Washington, το 2016.

Η τεχνική του XGBoost στηρίζεται, λοιπόν, στη μέθοδο Gradient Boosting και συγκεκριμένα είναι η μέθοδος "Extreme Gradient Boosting". Μία ακόμη διαφοροποίηση του από το RF είναι ότι τα δέντρα τρέχουν παράλληλα, άρα απαιτείται λιγότερος χρόνος για την περαίωση μίας εργασίας. Πρόκειται για μία μέθοδο η οποία συνδυάζει τεχνικές βελτιστοποίησης λογισμικών και εξοπλισμού και παρέχει ικανοποιητικά αποτελέσματα σε σύντομο χρονικό διάστημα (Morde, 2019).

Αξίζει να σημειωθεί ότι ο εν λόγω αλγόριθμος μπορεί να χρησιμοποιηθεί σε προβλήματα ταξινόμησης και παλινδρόμησης. Δεν κάνει διακρίσεις στα λειτουργικά συστήματα, μπορεί να «τρέχει» σε Mac, Linux και σε Windows.

Στο παρακάτω Σχήμα παρουσιάζονται οι μέθοδοι για το ensemble learning, μαζί με μία σύντομη περιγραφή για καλύτερη κατανόησή τους.



Σχήμα 8 *Μέθοδοι ensemble learning, (Morde, 2019)*

3.3.3 Μετρικές Αξιολόγησης Μοντέλων Μηχανικής Μάθησης

Η ανάλυση και η πρόβλεψη μελλοντικών τιμών με χρήση χρονοσειρών και με εφαρμογές μηχανικής μάθησης, είναι αντικείμενο των μοντέλων παλινδρόμησης. Τα μοντέλα παλινδρόμησης ανήκουν στην επιβλεπόμενη μηχανική μάθηση και έχουν ως τελικό αποτέλεσμα ένα συνεχές προϊόν. Γι' αυτό το λόγο οι μετρικές οι οποίες συναντώνται, βασίζονται κυρίως στον υπολογισμό κάποιου είδους απόστασης ανάμεσα στην προβλεπόμενη και στην αληθή τιμή.

Εν αντιθέσει, τα μοντέλα ταξινόμησης, τα οποία ανήκουν στην επιβλεπόμενη μηχανική μάθηση, έχουν ως τελικό αποτέλεσμα ένα δυαδικό προϊόν, ένα διακριτό προϊόν. Γι' αυτό το λόγο οι μετρικές οι οποίες συναντώνται, βασίζονται κυρίως στη σύγκριση διακριτών κλάσεων. Οι μετρικές της ταξινόμησης, αποτιμούν την απόδοση ενός μοντέλου και πληροφορούν πόσο καλή ή κακή είναι η ταξινόμηση, όμως κάθε μία από αυτές την αξιολογεί με διαφορετικό τρόπο (Bajaj, 2022).

Οι μετρικές αξιολόγησης, λοιπόν, σε κάθε περίπτωση χρησιμοποιούνται με σκοπό την αξιολόγηση της απόδοσης και της αποτελεσματικότητας των μοντέλων μηχανικής μάθησης. Με βάση τα παραπάνω, γίνεται αντιληπτό ότι συναντώνται αρκετοί τύποι μετρικών αξιολόγησης οι οποίες χρησιμοποιούνται στη μηχανική μάθηση ανάλογα με το μοντέλο το οποίο χρησιμοποιείται, αλλά και με τα αποτελέσματα τα οποία προκύπτουν. Ακριβώς το ίδιο ισχύει και για την εκτίμηση της απόδοσης ενός μοντέλου πρόβλεψης με χρήση χρονοσειρών (Lendave, 2021).

Το ερώτημα το οποίο χρήζει να απαντηθεί στην υποενότητα αυτή δεν είναι κάποιο άλλο, πέρα από το ποια είναι η σπουδαιότητα της αξιολόγησης ενός μοντέλου μηχανικής μάθησης. Υπάρχουν αρκετά στάδια για την επίλυση των προβλημάτων μηχανικής μάθησης, όπως είναι η συλλογή των δεδομένων, το ίδιο το πρόβλημα, η προ-επεξεργασία των δεδομένων, οι κατάλληλοι μετασχηματισμοί, η εκπαίδευση του μοντέλου και τέλος η αξιολόγησή του.

Παρόλο που συναντώνται διάφορα στάδια, το στάδιο της αξιολόγησης ενός μοντέλου μηχανικής μάθησης είναι το πιο κρίσιμο, καθώς δίνει πληροφορίες για την ακρίβεια του μοντέλου πρόβλεψης. Άλλωστε, η απόδοση και η χρήση ενός μοντέλου μηχανικής μάθησης στηρίζεται στην ακρίβεια την οποία επιτυγχάνει. Συνεπώς, μετά το πέρας της εκπαίδευσης ενός μοντέλου μηχανικής μάθησης, δημιουργείται η αμφισβήτηση για την εφαρμοσιμότητα του μοντέλου για το προς επίλυση πρόβλημα και δημιουργούνται ερωτήσεις όπως «Είναι το μοντέλο αυτό κατάλληλο για το συγκεκριμένο πρόβλημα;», «Πόσο ακριβές είναι το μοντέλο αυτό;», «πώς μπορεί να κριθεί ότι το μοντέλο αυτό είναι το καταλληλότερο ως προς την προσαρμογή στο τρέχον πρόβλημα;». Τα ερωτήματα αυτά, λοιπόν, λύνονται μέσω της

τεχνικής της Αξιολόγησης Μοντέλου Μηχανικής Μάθησης, η οποία περιγράφει την απόδοση ενός μοντέλου και καθιστά σαφές εάν το σχεδιασμένο μοντέλο είναι, ή όχι, κατάλληλο για το προς μελέτη πρόβλημα. Μέσω της τεχνικής αυτής, ξεχωρίζει εκείνος ο αλγόριθμος ο οποίος ταιριάζει καλύτερα στο δοθέν σετ δεδομένων και στην πρόβλεψη των τελικών προϊόντων. Έτσι, από όλους τους αλγορίθμους οι οποίοι χρησιμοποιούνται, επιλέγεται εκείνος ο οποίος παρέχει περισσότερη ακρίβεια για τα δεδομένα εισόδου και θεωρείται ως ο καλύτερος για το μοντέλο μιας και προβλέπει αρτιότερα το τελικό αποτέλεσμα.

Η ακρίβεια είναι ένας πολύ καθοριστικός παράγοντας στην επίλυση προβλημάτων μηχανικής μάθησης. Εάν η ακρίβεια είναι υψηλή, τότε οι προβλέψεις του μοντέλου με βάση τα δοθέντα δεδομένα συναντώνται στο μέγιστο δυνατό βαθμό.

Για να μπορέσει να πραγματοποιηθεί η αξιολόγηση μοντέλου, το αρχικό σετ δεδομένων χωρίζεται σε δύο τύπους. Στα δεδομένα εκπαίδευσης και στα δεδομένα ελέγχου. Όπως έχει ήδη αναφερθεί σε προηγούμενο κεφάλαιο, το μοντέλο «χτίζεται» με τα δεδομένα εκπαίδευσης, training dataset, και έπειτα, η αξιολόγηση του πραγματοποιείται μέσω των δεδομένων ελέγχου, test dataset, τα οποία αποτελούνται από άγνωστα, προς το μοντέλο, δείγματα δεδομένων τα οποία δεν χρησιμοποιήθηκαν στη διαδικασία εκπαίδευσης.

Υπάρχουν δύο μέθοδοι αξιολόγησης της απόδοσης ενός μοντέλου. Η πρώτη είναι η μέθοδος "Holdout", η οποία χρησιμοποιεί δύο τύπους δεδομένων για εκπαίδευση και έλεγχο. Τα δεδομένα ελέγχου χρησιμοποιούνται για τον υπολογισμό της απόδοσης του μοντέλου το οποίο έχει εκπαιδευτεί μέσω των δεδομένων εκπαίδευσης. Η εν λόγω μέθοδος χρησιμοποιείται για να ελεγχθεί πόσο καλά το μοντέλο μηχανικής μάθησης έχει αναπτυχθεί μέσα από τη χρήση διάφορων αλγορίθμων για άγνωστα δείγματα δεδομένων. Πρόκειται για μία απλή, ευέλικτη και γρήγορη μέθοδο. Η δεύτερη μέθοδος είναι η "Cross Validation". Ως Cross Validation ορίζεται η διαδικασία με την οποία ολόκληρο το σετ δεδομένων χωρίζεται δείγματα δεδομένων και η αξιολόγηση του μοντέλου μηχανικής μάθησης προκύπτει μέσα από τη χρήση των συμπληρωματικών δειγμάτων δεδομένων, με σκοπό την γνωστοποίηση της ακρίβειας του μοντέλου. Ο υπολογισμός της εν λόγω μεθόδου δύναται να υλοποιηθεί με τρεις διαφορετικούς τρόπους.

Ο πρώτος τρόπος είναι η μέθοδος "Validation", κατά την οποία γίνεται μοίρασμα του σετ δεδομένων σε 50% δεδομένα εκπαίδευσης και 50% δεδομένα ελέγχου. Ένα πολύ σημαντικό μειονέκτημα της μεθόδου είναι ότι από το 50% των δεδομένων τα οποία πρόκειται να χρησιμοποιηθούν για έλεγχο, ενδέχεται να περιλαμβάνει σημαντικές πληροφορίες οι οποίες πιθανότατα θα χαθούν κατά την εκπαίδευση του μοντέλου. Συνεπώς, στον τρόπο αυτό παρατηρείται υψηλό συστηματικό σφάλμα.

Ο δεύτερος τρόπος είναι η "Leave one out cross validation (LOOCV)". Στη συγκεκριμένη μέθοδο εκπαιδεύονται όλα τα σετ δεδομένων στο μοντέλο, εκτός από ένα (data-point) το οποίο χρησιμοποιείται για τον έλεγχο. Αυτή η μέθοδος παρουσιάζει μικρότερο συστηματικό σφάλμα, όμως υπάρχει πιθανότητα να αποτύχει καθώς το ένα σημείο το οποίο έχει μείνει εκτός, ενδέχεται να είναι outlier σημείο στα δοθέντα δεδομένα, οπότε σε αυτήν την περίπτωση δεν είναι εφικτή η παρουσίαση καλύτερων αποτελεσμάτων με αξιόλογη ακρίβεια.

Τρίτος και τελευταίος τρόπος είναι η K-Fold Cross Validation. Πρόκειται για μία δημοφιλή μέθοδο η οποία χρησιμοποιείται για αξιολόγηση ενός μοντέλου μηχανικής μάθησης. Λειτουργεί χωρίζοντας τα δεδομένα σε k σημεία. Γίνεται η εκπαίδευση των k επιμέρους σημείων των δεδομένων στο μοντέλο και αφήνεται εκτός ένα (k-1) υποσύνολο το οποίο χρησιμοποιείται για την αξιολόγηση του εκπαιδευμένου μοντέλου. Αυτή η μέθοδος

έχει ως αποτελέσματα υψηλή ακρίβεια και δημιουργεί δεδομένα με λιγότερα συστηματικά σφάλματα (Garlapati, 2021).

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, χρησιμοποιήθηκε η μέθοδος "Holdout", χωρίζοντας τα δεδομένα σε δεδομένα ελέγχου με ποσοστό 20% και σε δεδομένα εκπαίδευσης με ποσοστό 80%.

Επιπροσθέτως αντικείμενο ενασχόλησης είναι οι χρονοσειρές, άρα το ενδιαφέρον επικεντρώνεται κατά κύριο λόγο στα μοντέλα παλινδρόμησης. Εν τούτοις, κρίνεται σκόπιμη μία σύντομη και περιεκτική αναφορά στις μετρικές αξιολόγησης μοντέλων ταξινόμησης για λόγους πληρότητας και αρτιότητας της θεωρίας και της ανάλυσης.

3.3.3.1 Μετρικές Αξιολόγησης Μοντέλων Παλινδρόμησης

Ο τρόπος υπολογισμού, για έκαστη μετρική, παρουσιάζεται στις ακόλουθες γραμμές.

Η πρώτη μετρική της εξίσωσης (1.7), μετρική του μέσου τετραγωνικού σφάλματος, υπολογίζεται ως το άθροισμα του τετραγώνου των διαφορών των προβλεπόμενων με των παρατηρούμενων τιμών, διαιρεμένο με το συνολικό αριθμό των παρατηρήσεων.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_{pred}^{(i)} - y_{obs}^{(i)}\right)^2 \qquad (1.7)$$

Όπου $y_{obs}^{(i)}$ είναι η *i* παρατηρούμενη τιμή, $y_{pred}^{(i)}$ είναι η αντίστοιχή της προβλεπόμενη τιμή και n είναι ο συνολικός αριθμός των παρατηρήσεων.

Αφού ορίστηκε το μέσο τετραγωνικό σφάλμα, το μέσο τετραγωνικό σφάλμα ρίζας – RMSE υπολογίζεται αρκετά εύκολα, στην εξίσωση (1.8). Ακόμη, μπορεί να οριστεί το ποσοστιαίο μέσο τετραγωνικό σφάλμα ρίζας – RMSPE, εξίσωση (1.9).

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{pred}^{(i)} - y_{obs}^{(i)}\right)^2} \qquad (1.8)$$

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{pred}^{(i)} - y_{obs}^{(i)}}{y_{obs}^{(i)}}\right)^2} * 100\ \% \qquad (1.9)$$

Εν συνεχεία, παρατίθεται για αρχή ο περιγραφικός τρόπος υπολογισμού του συντελεστή προσδιορισμού R^2.

$$R^2 = \frac{Διασπορά\ Μοντέλου}{Συνολική\ Διασπορά} \qquad (1.10)$$

Ο συντελεστής αυτός, αντιπροσωπεύει το τμήμα της διασποράς της εξαρτημένης μεταβλητής το οποίο εξηγείται από τις ανεξάρτητες μεταβλητές του μοντέλου. Υπολογίζει τη δύναμη της συσχέτισης ανάμεσα στο αναπτυχθέν μοντέλο, τις προβλέψεις του, και στην εξαρτημένη μεταβλητή.

Αναλυτικότερα, αρχικά υπολογίζεται η ολική διασπορά τετραγωνικού σφάλματος. Όταν είναι γνωστές οι τιμές των ανεξάρτητων μεταβλητών, είναι εύκολο να υπολογισθεί το σφάλμα της παλινδρόμησης. Όπως έχει αναφερθεί και σε προηγούμενη υποενότητα, ως υπόλοιπο ορίζεται η διαφορά ανάμεσα στην πραγματική και την προβλεπόμενη τιμή. Οπότε, καθορίζεται ότι το "Residual Sum of Squares" (RSS) ισούται με:

$$RSS = \sum_{i=1}^{n} (y_{obs}^{(i)} - y_{pred}^{(i)})^2 = \sum_{i=1}^{n} e_i^2 \quad (1.11)$$

Όπου $y_{obs}^{(i)}$ είναι η *i* παρατηρούμενη τιμή, $y_{pred}^{(i)}$ είναι η αντίστοιχή της προβλεπόμενη τιμή και e_i είναι το υπόλοιπο.

Στην περίπτωση κατά την οποία δεν είναι γνωστές οι τιμές των ανεξάρτητων μεταβλητών, οι μόνες γνωστές τιμές είναι των παρατηρούμενων μεταβλητών. Οπότε, μπορεί να υπολογισθεί ο μέσος όρος τους. Οπότε, υπολογίζεται η ολική διασπορά του συνόλου δεδομένων ανάμεσα στη μέση τιμή των μεταβλητών και στην τιμή της εκάστοτε μεταβλητής. Αυτό ορίζεται ως "Total Sum of Squares" (TSS).

$$TSS = \sum_{i=1}^{n} \left(y_{obs}^{(i)} - \bar{y_{obs}}\right)^2 \quad (1.12)$$

Επομένως, ο υπολογισμός του συντελεστή προσδιορισμού δύναται να πραγματοποιηθεί με τη χρήση των RSS και TSS, καθώς επιδιώκεται να γνωστοποιηθεί το ποσοστό της συνολικής διασποράς των μετρήσεων, το οποίο περιγράφεται από τις ανεξάρτητες μεταβλητές. Έτσι, εάν είναι γνωστό το ποσοστό της συνολικής διασποράς των μετρήσεων, το οποίο δεν περιγράφεται από κάποια γραμμή παλινδρόμησης, τότε προκύπτει το εξής:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (1.13)$$

Συνοψίζοντας, ο λόγος του RSS ως προς το TSS δηλώνει το συνολικό σφάλμα το οποίο παραμένει στο μοντέλο παλινδρόμησης. Αφαιρώντας το λόγο αυτό από τη μονάδα, προκύπτει το σφάλμα το οποίο έχει αφαιρεθεί χρησιμοποιώντας ανάλυση παλινδρόμησης, δηλαδή το R^2. Εάν η τιμή του R^2 είναι υψηλή, έστω 1, τότε το μοντέλο αντιπροσωπεύει τη μεταβλητότητα της εξαρτημένης μεταβλητής. Αντιθέτως, εάν η τιμή του R^2 είναι χαμηλή, τότε το μοντέλο δεν αντιπροσωπεύει τη μεταβλητότητα της εξαρτημένης μεταβλητής και δεν είναι πολύ καλύτερο από τη χρήση της μέσης τιμής. Επίσης, δηλώνει πως δεν χρησιμοποιούνται πληροφορίες από τις άλλες μεταβλητές. Μία αρνητική τιμή R^2 δηλώνει ότι το αποτέλεσμα είναι χειρότερο από εκείνο της μέσης τιμής. Το R^2 υπολογίζει τα σκόρπια σημεία από τη γραμμή της παλινδρόμησης. Σε εκείνη, δηλαδή, την περίπτωση κατά την οποία $R^2 = 1$, όλα τα σημεία πέφτουν πάνω στη γραμμή, καθώς οι προβλεπόμενες τιμές είναι ίδιες με τις πραγματικές τιμές. Ωστόσο, αξίζει να σημειωθεί ότι δεν θα ήταν σωστό κάποιος να επαναπαυτεί πλήρως στην τιμή του R^2, καθώς ένα καλό μοντέλο μπορεί να έχει χαμηλή τιμή, ενώ ένα μοντέλο με συστηματικά σφάλματα ενδέχεται να εμφανίζει μεγαλύτερη τιμή. Γι' αυτό το λόγο συνίσταται η οπτική απεικόνιση των υπολοίπων σε διάγραμμα διασποράς. (Yashwanth, 2020).

Σε αυτό το σημείο, κρίνεται αναγκαία η αναφορά μίας μετρικής η οποία συγχέεται αρκετά με την R^2. Πρόκειται για την "Explained Variance Score" (EVS). Η διαφορά τους έγκειται στο μέσο του σφάλματος. Δηλαδή, η EVS λαμβάνει υπόψιν τη διασπορά των υπολοίπων (residuals).

Αναλυτικότερα, ο τρόπος υπολογισμού της έχει ως εξής:

$$Explained\ Variance\ Score = 1 - \left[\frac{Variance\left(y_{pred}^{(i)} - y_{obs}^{(i)}\right)}{Variance\left(y_{obs}^{(i)}\right)}\right] \quad (1.14)$$

Όπου,

$$Variance\left(y_{pred}^{(i)} - y_{obs}^{(i)}\right) = \frac{1}{n}\sum_{i=1}^{0}\left[(y_{pred}^{(i)} - y_{obs}^{(i)})^2 - Mean\ Error\right] \quad (1.15)$$

Εύκολα μπορεί να γίνει αντιληπτό ότι η διαφορά ανάμεσα στις εξισώσεις (1.13) και (1.14) έγκειται στην αφαίρεση του μέσου σφάλματος. Οπότε, κατά τη σύγκριση της R^2 με την EVS αυτό το οποίο προκύπτει είναι το Mean Error. Συνεπώς, στην περίπτωση κατά την οποία R^2 = EVS, τότε αυτό συνεπάγεται ότι το Mean Error είναι ίσο με το μηδέν. Το Mean Error αντιπροσωπεύει την τάση της εκτιμήτριας, δηλαδή της μεροληπτικής ή αμερόληπτης εκτίμησης. Από τα παραπάνω προκύπτει ότι θα ήταν καλό να υπολογιζόταν και να λαμβάνονταν υπόψιν το μέσο του σφάλματος, ώστε η εκτιμήτρια να είναι αμερόληπτη (Yahya, 2020).

Ακολουθεί ο τύπος υπολογισμού του μέσου απόλυτου σφάλματος (MAE). Πρόκειται για ένα μέσο όρο της απόλυτης διαφοράς ανάμεσα στην πραγματική τιμή και στην τιμή η οποία προβλέπεται από το μοντέλο. Συνηθίζεται να μην προτιμάται σε περιπτώσεις όπου παρατηρούνται σημαντικές αποκλίνουσες τιμές.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{pred}^{(i)} - y_{obs}^{(i)}\right| \quad (1.16)$$

Όπως και στην περίπτωση του μέσου τετραγωνικού σφάλματος, αφού ορίστηκε το μέσο απόλυτο σφάλμα, για να καταστεί ευκολότερα αντιληπτό, το ποσοστιαίο μέσο απόλυτο σφάλμα ρίζας (MAPE) υπολογίζεται ως εξής:

$$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_{pred}^{(i)} - y_{obs}^{(i)}}{y_{obs}^{(i)}}\right|\right) * 100\% \quad (1.17)$$

Αν και το RMSE είναι μία από τις πιο συνήθεις μετρικές, η ερμηνεία του δεν είναι αρκετά εύκολη. Έτσι, ένας εναλλακτικός τρόπος είναι η παρατήρηση της κατανομής των ποσοστιαίων απόλυτων σφαλμάτων. Ως μετρική Max Error ορίζεται το χείριστο πιθανό σφάλμα ανάμεσα στην προβλεπόμενη και την αληθή τιμή.

Τις διάφορες μετρικές, οι οποίες συναντώνται στη βιβλιογραφία, μπορεί κάποιος να τις κατηγοριοποιήσει σε τέσσερις ομάδες, όπως προτάθηκε από τους Hyndman και Koehler (Hyndman, 2006).

Στην *πρώτη* κατηγορία ανήκουν οι μετρικές οι οποίες εξαρτώνται από την κλίμακα. Αναλυτικότερα, η κλίμακα των μετρήσεων αυτών εξαρτάται από την κλίμακα των δεδομένων. Δύναται να είναι χρήσιμες για τη σύγκριση διαφορετικών μεθόδων οι οποίες εφαρμόζονται στο ίδιο σετ δεδομένων. Οι μετρικές οι οποίες αποτελούν την εν λόγω κατηγορία είναι: Mean Squared Error, Root Mean Squared Error, Mean Absolute Error και Median Absolute Error. Από αυτές τις μετρικές, για αξιολόγηση προβλέψεων προτιμάται περισσότερο η χρήση του rMSE από το MSE, καθώς βρίσκεται στην ίδια κλίμακα με τα δεδομένα. Βέβαια, πρόκειται για δύο από τις δημοφιλέστερες μετρικές λόγω της σύνδεσής τους με τα στατιστικά μοντέλα. Εν τούτοις, παρουσιάζουν αξιοσημείωτη ευαισθησία στα πιο ιδιάζοντα σημεία/outliers, συγκριτικά με το MAE και το MdAE.

Στη *δεύτερη* κατηγορία ανήκουν οι μετρικές οι οποίες στηρίζονται σε ποσοστιαίες μετρήσεις σφαλμάτων. Τα εν λόγω σφάλματα έχουν το πλεονέκτημα να είναι ανεξάρτητα από την κλίμακα των δεδομένων και γι' αυτόν το λόγο χρησιμοποιούνται συχνά σε σύγκριση απόδοσης προβλέψεων ανάμεσα σε διαφορετικά σετ δεδομένων. Ο υπολογισμός ενός τέτοιου σφάλματος δίνεται από τον παρακάτω τύπο:

$$p_t = \frac{100e_t}{Y_t} \quad (1.18)$$

Όπου Y_t η παρατήρηση στη χρονική στιγμή t, F_t η πρόβλεψη της Y_t και όπου $e_t = Y_t - F_t$ το σφάλμα της πρόβλεψης.

Οι μετρικές οι οποίες αποτελούν τη δεύτερη κατηγορία είναι: Mean Absolute Percentage Error, Median Absolute Percentage Error, Root Mean Squared Percentage Error και Root Median Squared Percentage Error. Ένα μειονέκτημα των συγκεκριμένων μετρήσεων είναι η πιθανότητά τους να εμφανίσουν ακαθόριστη ή άπειρη τιμή, όταν $Y_t = 0$, καθώς επίσης να εμφανίσουν σημαντική ασσυμετρία όταν οι τιμές της Y_t βρίσκονται κοντά στο μηδέν.

Στην *τρίτη* κατηγορία, ανήκουν οι μετρητικές στις οποίες κάθε σφάλμα διαιρείται με ένα σφάλμα το οποίο προέκυψε μέσω της χρήσης μίας διαφορετικής μεθόδου πρόβλεψης. Έστω ότι το r_t δηλώνει το σχετικό σφάλμα και για το οποίο ισχύει η ακόλουθη εξίσωση:

$$r_t = \frac{e_t}{e_t^*} \quad (1.19)$$

Όπου e_t^* είναι το σφάλμα πρόβλεψης το οποίο αποκτήθηκε από μία μέθοδο – ορόσημο.

Οι μετρικές οι οποίες ανήκουν σε αυτήν την κατηγορία είναι η Mean Relative Absolute Error, η Median Relative Absolute Error και η Geometric Mean Relative Absolute Error. Οι μετρικές αυτές επιδιώκουν να αφαιρέσουν την κλίμακα των δεδομένων μέσα από τη σύγκριση των προβλέψεων με προβλέψεις – ορόσημα, συνήθως της naive μεθόδου. Εν τούτοις, εμφανίζουν προβλήματα. Συγκεκριμένα, παρατηρείται στατιστική κατανομή με απροσδιόριστο μέσο και άπειρη διασπορά.

Στην *τέταρτη* και τελευταία κατηγορία ανήκει η μετρική η οποία έχει "scaled error". Πρόκειται για τη Mean Absolute Scaled Error (MASE).

Ένα "scaled error" ορίζεται ως εξής:

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}\left|y_{obs}^{i} - y_{obs}^{i-1}\right|} \quad (1.20)$$

Οπότε προκύπτει ότι:

$$MASE = mean(q_t) \quad (1.21)$$

Μερικά από τα πλεονεκτήματα του MASE είναι ότι δεν δίνει ποτέ απροσδιόριστες ή άπειρες τιμές και είναι μία αρκετά καλή επιλογή για διακοπτόμενες σειρές. Τέλος, μπορεί να χρησιμοποιηθεί σε ενιαίες σειρές, όπως επίσης σαν εργαλείο για την αξιολόγηση πολλαπλών σειρών.

Σε κάθε περίπτωση, χρειάζεται να λαμβάνεται υπόψιν ότι η επιλογή και χρήση των κατάλληλων και ορθών μετρικών αξιολόγησης είναι αρκετά κρίσιμη. Μία λάθος μετρική αξιολόγησης αφενός θα επηρεάσει τη βελτιστοποίηση του μοντέλου και αφετέρου θα διαστρεβλώσει την αξιολόγηση των αλγορίθμων (Rink, 2021). Αδιαμφισβήτητα, ανάμεσα σε τόσες μετρικές, δεν υπάρχει κάποια η οποία να είναι καλύτερη. Κάθε στατιστική μέτρηση «συμπυκνώνει» ένα μεγάλο πλήθος δεδομένων σε μία μοναδική τιμή. Με αυτόν τον τρόπο, παρέχει μόνο μία προβολή των σφαλμάτων του μοντέλου, η οποία δίνει έμφαση σε μία συγκεκριμένη πλευρά των χαρακτηριστικών των σφαλμάτων από την απόδοση του μοντέλου (Chai and Draxler, 2014).

3.3.3.2 Μετρικές Αξιολόγησης Μοντέλων Ταξινόμησης

Μερικές από τις πιο γνωστές μετρικές αξιολόγησης, οι οποίες εφαρμόζονται σε αλγορίθμους ταξινόμησης είναι οι εξής: Classification Accuracy, Confusion Matrix, Logarithmic Loss, Area under Curve (AUC) και F-Measure.

Η πρώτη μετρική, δηλαδή το Classification Accuracy, είναι παρεμφερής με τον όρο «Ακρίβεια»/Accuracy. Πρόκειται για το λόγο των σωστών προβλέψεων προς το συνολικό αριθμό προβλέψεων του μοντέλου με βάση τα δοθέντα δεδομένα. Στην περίπτωση στην οποία τα δοθέντα δεδομένα έχουν τον ίδιο τύπο δεδομένων που σχετίζεται με το δοθέν πρόβλημα, τότε προκύπτει καλύτερη ακρίβεια. Εάν η ακρίβεια είναι υψηλή, τότε το μοντέλο είναι πιο ακριβές και δύναται να χρησιμοποιηθεί. Εάν η ακρίβεια είναι χαμηλή, τότε αυτό υποδηλώνει ότι τα δείγματα δεδομένων δεν έχουν ταξινομηθεί σωστά ώστε να ταιριάζουν στο δοσμένο πρόβλημα.

Στην περίπτωση του Confusion Matrix, δηλαδή ενός Πίνακας NxN ο οποίος χρησιμοποιείται για την αξιολόγηση ενός μοντέλου ταξινόμησης, όπου N είναι ο αριθμός των προβλεπόμενων κλάσεων, λειτουργεί επάνω σε ένα σετ δεδομένων ελέγχου του οποίου οι πραγματικές τιμές είναι γνωστές. Μέσω του εν λόγω Πίνακας, προκύπτουν τα νούμερα των σωστών και λαθεμένων προβλέψεων, καθώς επίσης χρησιμοποιείται ώστε να υπολογισθεί η ακρίβεια και ορθότητα του μοντέλου. Αποτελείται από τιμές όπως True Positive, False Positive, True Negative, και False Negative, οι οποίες χρησιμοποιούνται και για τον υπολογισμό της ακρίβειας, της ορθότητας της ανάκλησης, της ευαισθησίας και της καμπύλης AUC. Όλες αυτές οι μετρήσεις φανερώνουν την απόδοση ενός μοντέλου. Η ακρίβεια υπολογίζεται από το μέσο όρο των τιμών True Positive και True Negative του συνολικού δείγματος τιμών. Πρόκειται για μία μέτρηση η οποία γνωστοποιεί το συνολικό αριθμό των ορθών προβλέψεων οι οποίες πραγματοποιήθηκαν από το μοντέλο.

Εν συνεχεία, η Area under the ROC curve είναι μία από τις πιο χρησιμοποιούμενες μετρικές αξιολόγησης. Οι δείκτες True Positive και False Positive έχουν εύρος τιμών από το μηδέν έως το ένα. Οι εν λόγω δείκτες υπολογίζονται με χρήση διαφορετικών κατωφλίων. Το "Area under curve" είναι η γραφική παράσταση ανάμεσα στο δείκτη False Positive και το δείκτη True Positive.

Περνώντας στο "Logarithmic Loss", ή αλλιώς γνωστό ως Log Loss. Όπως ήδη αναφέρθηκε, η AUC-ROC καθορίζει την απόδοση του μοντέλου χρησιμοποιώντας προβλεπόμενες πιθανότητες, αλλά δεν λαμβάνει υπόψη την πιθανότητα του μοντέλου να προβλέψει την υψηλότερη πιθανότητα των δειγμάτων να είναι περισσότερο θετικά. Πρόκειται για μία τεχνική η οποία χρησιμοποιείται σε ταξινόμηση πολλών τάξεων και υπολογίζεται ως η αρνητική μέση τιμή του λογαρίθμου των σωστών προβλεπόμενων πιθανοτήτων για κάθε περίπτωση.

Τέλος, F-Measure ή αλλιώς F1-Score, είναι η καλύτερη μέτρηση της ακρίβειας του ελέγχου του ανεπτυγμένου μοντέλου. Όσο πιο ψηλή είναι η τιμή του, τόσο καλύτερη απόδοση έχει το μοντέλο (Garlapati, 2021).

3.3.4 Εφαρμογές Μηχανικής Μάθησης

Η μηχανική μάθηση ανήκει σε ένα πεδίο ευρέως αναπτυσσόμενο. Ειδικότερα τα τελευταία χρόνια εμφανίζεται μία στροφή της επιστημονικής κοινότητας, αλλά και της βιομηχανίας προς τη χρήση μεθόδων μηχανικής μάθησης για διάφορα προβλήματα. Μπορεί να χρησιμοποιεί σε ποίκιλλα πεδία, διαφορετικά μεταξύ τους.

Για παράδειγμα, στον κλάδο της ιατρικής, μπορεί να χρησιμοποιηθεί για ιατρικές προγνώσεις. Στον κλάδο του εμπορίου και στον τομέα επιχειρήσεων, μπορεί να χρησιμοποιηθεί για να βελτιωθεί η εξυπηρέτηση των καταναλωτών με χρήση chatbot. Εφαρμογές μηχανικής μάθησης συναντώνται επίσης και σε ζητήματα τα οποία σχετίζονται με τις γλώσσες. Μία εφαρμογή, δηλαδή, είναι σε θέση να αναγνωρίσει την ομιλία. Άπαξ και την αναγνωρίσει, μπορεί να την καταγράψει, να τη μεταφράσει, αλλά μπορεί και να απαντήσει ή

να εκτελέσει εντολές. Επιπροσθέτως, τον τελευταίο καιρό παρατηρείται η ενασχόληση των αυτοκινητοβιομηχανιών με την αυτόνομη οδήγηση. Έτσι, οι αυτοκινητοβιομηχανίες συνδυάζουν μεθόδους τεχνητής νοημοσύνης με σκοπό να δημιουργήσουν αυτοκίνητα τα οποία θα μπορούν να «διαβάζουν» το δρόμο. Διάφορες εφαρμογές της όρασης υπολογιστών, επίσης, χρησιμοποιούν τεχνικές μηχανικής μάθησης για να πετύχουν το σκοπό τους. Επιπλέον, μέσω της μηχανικής μάθησης, μπορεί να γίνει η αναγνώριση συναισθημάτων από κείμενα, με χαρακτηριστικό παράδειγμα η ταξινόμηση των σχολίων με βάση την εμπειρία του πελάτη σε θετικά και σε αρνητικά. Τέλος, η μηχανική μάθηση μπορεί να χρησιμοποιηθεί και σε μοντέλα πρόβλεψης τιμών διαφόρων χαρακτηριστικών.

3.4 Χρονοσειρές

Ως χρονοσειρά (Time Series), ή αλλιώς χρονολογική σειρά, ορίζεται ένα σύνολο ιστορικών παρατηρήσεων ενός μεγέθους, με σταθερό χρονικό βήμα, δηλαδή με σταθερό χρόνο δειγματοληψίας (Sampling Time). Μέσω μίας χρονοσειράς εκφράζεται η εξέλιξη μίας μεταβλητής σε βάθος χρόνου, το πώς δηλαδή μεταβάλλεται η τιμή μίας μεταβλητής στις διάφορες χρονικές στιγμές[2]. Δεδομένα χρονοσειρών μπορούν να λειτουργήσουν ως εργαλείο σε προβλήματα πρόβλεψης τιμών, καθώς επίσης και για ανάλυση και κατανόηση ενός φαινομένου.

Οι περισσότερες μέθοδοι πρόβλεψης χρησιμοποιούν ιστορικά δεδομένα τα οποία είναι προϊόντα μίας χρονοσειράς. Χαρακτηριστικά μοντέλα πρόβλεψης με χρήση χρονοσειρών είναι οι ημερήσιες τιμές μία μετοχής, οι ημερήσιες αφίξεις τουριστών σε ένα αεροδρόμιο, οι εβδομαδιαίες πωλήσεις ενός προϊόντος σε ένα χρονικό διάστημα, ο αριθμός ενοικιαζόμενων αυτοκινήτων σε μία χρονική περίοδο, οι μετεωρολογικές χρονοσειρές και φυσικά, οι χρονοσειρές οι οποίες αφορούν την εξέλιξη πανδημιών, όπως στην περίπτωση της πανδημίας του Covid-19. Δεν συναντάται περιορισμός ως προς το είδος των μεγεθών τα οποία συναντώνται στις χρονοσειρές. Για παράδειγμα, μπορούν να συναντηθούν τόσο διακριτικά μεγέθη σε διακριτό ή συνεχή χρόνο όσο και συνεχή μεγέθη σε διακριτό ή συνεχή χρόνο (Λοΐζου, 2020).

Σε περιπτώσεις πρόβλεψης με χρήση χρονοσειρών, λαμβάνεται υπόψιν ο χρονικός ορίζοντας της πρόβλεψης. Συναντώνται τρεις περιπτώσεις ανάλογα με τα χρονικά διαστήματα της πρόβλεψης. Αρχικά, συναντάται η βραχυπρόθεσμη πρόβλεψη (Short-term forecasting), στη συνέχεια συναντάται η μεσοπρόθεσμη πρόβλεψη (Medium-term forecasting) και τέλος, συναντάται η μακροπρόθεσμη πρόβλεψη (Long-term forecasting). Όπως τα ονόματά τους φανερώνουν, στην πρώτη περίπτωση ο χρονικός ορίζοντας είναι μικρός, δηλαδή μικρότερος από ένα έτος, στη δεύτερη περίπτωση ο χρονικός ορίζοντας είναι μεγαλύτερος από ένα έτος και στην τελευταία περίπτωση, ο ορίζοντας πρόβλεψης είναι μεγαλύτερος των τριών ετών (Μπίνου, 2016).

Οι προβλέψεις με χρήση χρονοσειρών αποτελούν ένα σημαντικό κομμάτι της μηχανικής μάθησης. Όπως αναφέρθηκε και σε προηγούμενη παράγραφο, αρκετά προβλήματα πρόβλεψης χρησιμοποιούν πληροφορία χρόνου. Εν τούτοις, εάν και αυτή η χρονική πληροφορία αποτελεί ένα επιπρόσθετο στοιχείο, καθιστά πιο ιδιαίτερη τη διαχείριση των μοντέλων πρόβλεψης, συγκριτικά με άλλα μοντέλα πρόβλεψης τα οποία δεν χρησιμοποιούν χρονοσειρές (Flovik, 2018).

[2] Οι χρονικές στιγμές μπορεί να είναι ώρες, ημέρες, εβδομάδες, μήνες, χρόνια κ.ο.κ.

3.4.1 Είδη Χρονοσειρών

Οι χρονοσειρές χωρίζονται σε δύο κατηγορίες ανάλογα με το πλήθος των μεταβλητών που περιέχουν. Συγκεκριμένα, συναντώνται οι Μονοδιάστατες Χρονοσειρές (Univariate Time Series) και οι Πολυδιάστατες Χρονοσειρές (Multivariate Time Series).

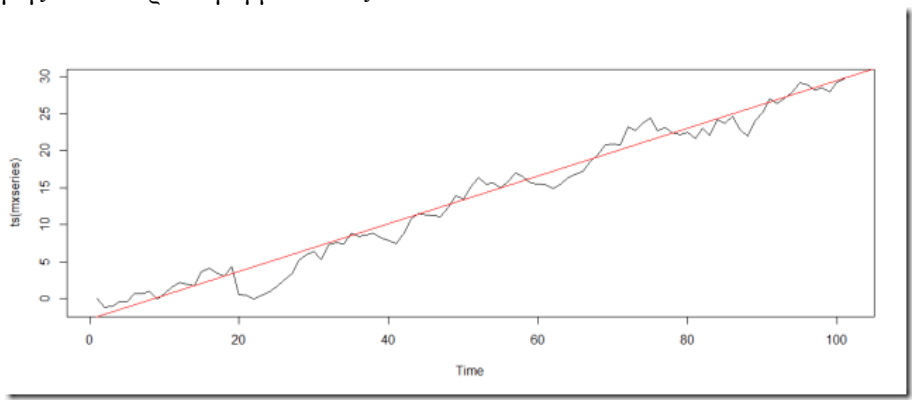
Στις μονοδιάστατες χρονοσειρές, όπως μαρτυρά και το όνομά τους, συναντάται μία μόνο μεταβλητή. Δηλαδή, συλλέγονται δεδομένα για τη διακύμανση μίας μόνο μεταβλητής στο χρόνο.

Στις πολυδιάστατες χρονοσειρές, συναντώνται περισσότερες από μία μεταβλητές. Δηλαδή, σε αυτήν την περίπτωση, συλλέγονται και χρησιμοποιούνται δεδομένα από διάφορες μεταβλητές. Παρουσιάζει αρκετό ενδιαφέρον η αναζήτηση κάποιας πιθανής εξάρτησης/σχέσης ανάμεσα στις μεταβλητές των πολυδιάστατων χρονοσειρών. Μέσω της αναζήτησης αυτής, εάν μία μεταβλητή επηρεάζει την πρόγνωση των μελλοντικών τιμών τότε συναντάται ομογένεια, διαφορετικά εάν δεν επηρεάζει την πρόγνωση συναντάται ετερογένεια (Λοΐζου, 2020).

3.4.2 Στασιμότητα Χρονοσειρών

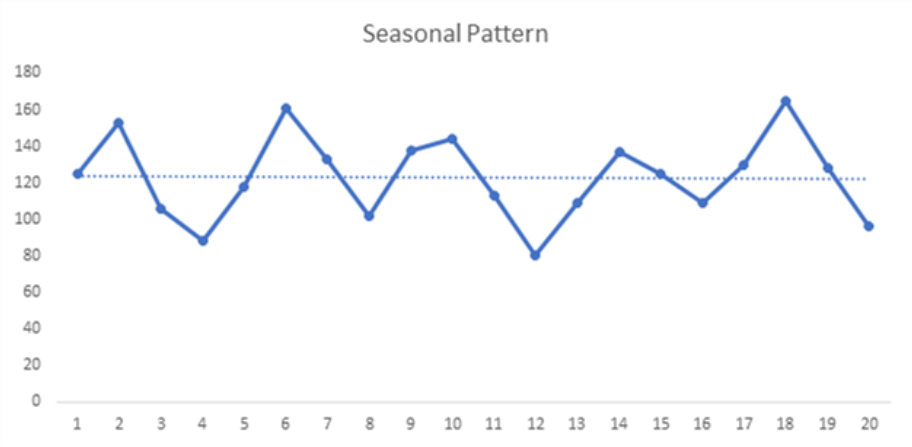
Οι χρονοσειρές διακρίνονται σε δύο κατηγορίες όσον αφορά τη στασιμότητά τους. Δηλαδή, μελετάται η περίπτωση κατά την οποία τα διάφορα στατιστικά μέτρα των χρονοσειρών μένουν αναλλοίωτα ή όχι. Στην περίπτωση κατά την οποία τα στατιστικά μέτρα, όπως είναι επί παραδείγματι η μέση τιμή και η διασπορά, παραμένουν αναλλοίωτα, τότε η χρονοσειρά θεωρείται Στάσιμη (Stationary Time Series). Εάν τα στατιστικά μέτρα μεταβάλλονται, τότε η χρονοσειρά θεωρείται Μη Στάσιμη (Non-stationary Time Series). Σε αυτήν την περίπτωση συναντώνται οι έννοιες της τάσης, της εποχικότητας, της κυκλικότητας και των ιδιαζόντων σημείων (Λοΐζου, 2020).

Ως τάση (trend) ορίζεται η κίνηση η οποία παρουσιάζει μία χρονοσειρά σε μία χρονική περίοδο. Η τάση μπορεί να είναι ανοδική, όταν η εμφανίζονται μεγαλύτερες τιμές της μεταβλητής στο βάθος χρόνου, καθοδική, όταν εμφανίζονται μικρότερες τιμές της μεταβλητής στο βάθος χρόνου, σύνθετη, όταν η τάση εμφανίζει και άνοδο και κάθοδο και τέλος, μπορεί να είναι ανύπαρκτη στην περίπτωση κατά την οποία η κίνηση των τιμών της μεταβλητής είναι παράλληλη με τον άξονα x.



Σχήμα 9 *Διάγραμμα Τάσης, (Buchta, 2015)*

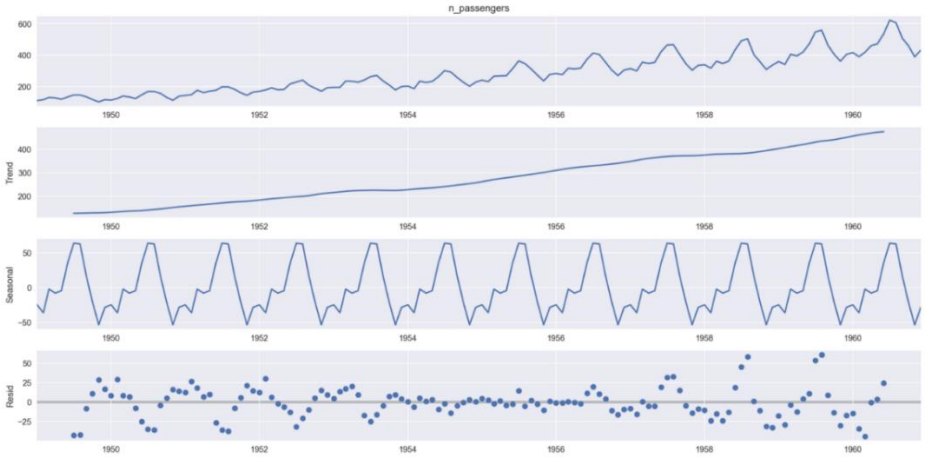
Ως εποχικότητα (seasonality) ορίζεται η ιδιότητα συστηματικής επανάληψης των τιμών μίας χρονοσειράς σε συγκεκριμένο χρονικό διάστημα. Πρόκειται για μία περιοδική διακύμανση, η οποία έχει ένα σταθερό μήκος.



Σχήμα 10 *Διάγραμμα Εποχικότητας, (Ranjan, 2020)*

Ως κυκλικότητα (cyclic) ορίζεται η ιδιότητα των τιμών μίας χρονοσειράς να εμφανίζει διακυμάνσεις οι οποίες επαναλαμβάνονται γύρω από τη γραμμή τάσης.

Τέλος, ως ιδιάζοντα σημεία (outliers) ορίζονται ως εκείνα τα σημεία τα οποία εμφανίζουν «ακραίες» τιμές σε μία γραφική παράσταση. Στις περισσότερες περιπτώσεις δεν αποτελούν πραγματικές μετρήσεις των μελετώμενων τιμών και οφείλονται είτε σε κάποια βλάβη του συστήματος, είτε σε κάποια εξωτερική παρεμβολή, όμως υπάρχει και η περίπτωση να αποτελούν ιδιαίτερες παρατηρήσεις από μη αναμενόμενα γεγονότα.



Σχήμα 11 *Αποσύνθεση Χρονοσειράς, (Lewinson, 2022)*

Στην παραπάνω Σχήμα παρουσιάζεται η αποσύνθεση μίας χρονοσειράς. Η χρονοσειρά απεικονίζεται στην πρώτη σειρά, στη συνέχεια παρουσιάζεται η γραφική παράσταση της τάσης της, της εποχικότητάς της και τέλος, παρουσιάζεται η γραφική παράσταση των ιδιαζόντων σημείων.

3.5 Υλοποίηση Μοντέλου Μηχανικής Μάθησης

Για να υλοποιηθεί ένα μοντέλο μηχανικής μάθησης χρειάζεται να ληφθούν υπόψιν ορισμένες παράμετροι, με σκοπό τον αρτιότερο, πληρέστερο και αποδοτικότερο σχεδιασμό του. Στη συνέχεια της τρέχουσας υποενότητας πρόκειται να αναλυθούν τα απαιτούμενα βήματα τα οποία οφείλει να ακολουθήσει κάποιος ο οποίος σκοπεύει στο σχεδιασμό ενός μοντέλου μηχανικής μάθησης.

3.5.1 Προσδιορισμός Προβλήματος

Αρχικά, χρειάζεται να γίνει σαφής προσδιορισμός του προβλήματος. Πρόκειται για ένα από τα βασικότερα θεμέλια για το «χτίσιμο» ενός μοντέλου μηχανικής μάθησης. Αυτό συμβαίνει,

διότι κατ' αυτόν τον τρόπο, γίνεται πλήρως κατανοητό το είδος του προβλήματος και έτσι εντοπίζονται τα κατάλληλα μέσα τα οποία επιβάλλεται να χρησιμοποιηθούν. Στην περίπτωση της παρούσας διπλωματικής εργασίας, το πρόβλημα είναι η ανάπτυξη ενός μοντέλου μηχανικής μάθησης το οποίο θα είναι ικανό να προβλέπει κρούσματα και θανάτους της Covid-19 με χρήση χρονοσειρών.

3.5.2 Είδος Δεδομένων Μοντέλου

Εφόσον καθορίστηκε το πρόβλημα, χρειάζεται να μελετηθεί το είδος των δεδομένων τα οποία θα εισαχθούν στο μοντέλο. Οπότε, σε αυτό το στάδιο, επιλέγονται δεδομένα από κατάλληλες πηγές τα οποία θα χρησιμοποιηθούν από τον αλγόριθμο του μοντέλου με τελικό στόχο την εξαγωγή αποτελεσμάτων. Όπως ήδη αναφέρθηκε, εδώ χρησιμοποιήθηκαν δεδομένα χρονοσειρών, τα οποία αντλήθηκαν από τον ιστότοπο "*OurWorldInData*". Περισσότερα για τα δεδομένα τα οποία χρησιμοποιήθηκαν παρουσιάζονται στο Κεφάλαιο 2.

3.5.3 Προ-επεξεργασία Δεδομένων

Στη συνέχεια, χρειάζεται να γίνει προ επεξεργασία των δεδομένων. Πρόκειται για μία διαδικασία μετασχηματισμού ανεπεξέργαστων δεδομένων, σε ένα κατανοητό φορμάτ. Χρειάζεται να γίνει έλεγχος της ποιότητας των δεδομένων προτού εφαρμοστούν οι αλγόριθμοι μηχανικής μάθησης. Η ποιότητα των δεδομένων μπορεί να σχετίζεται με την ακρίβεια, την πληρότητα και την αξιοπιστία (Anunaya, 2021). Συναντώνται τρία κυριότερα βήματα κατά τη διαδικασία της προ επεξεργασίας δεδομένων. Πρόκειται για τον καθαρισμό δεδομένων (data cleansing), το μετασχηματισμό των δεδομένων (data transformation) και τη μείωση των δεδομένων (data reduction).

Ως καθαρισμός δεδομένων ορίζεται εκείνη η τεχνική κατά την οποία αντιμετωπίζονται οι χαμένες/κενές τιμές, ο θόρυβος και άλλα προβλήματα των δεδομένων. Αρκετές φορές, συναντώνται δεδομένα τα οποία εμφανίζουν ακραίες τιμές και σφάλματα καθιστώντας έτσι την πληροφορία τους άχρηστη. Αυτά τα δεδομένα έχουν θόρυβο. Οι λαθεμένες τιμές, εάν δεν πρόκειται απλώς για ακραίες τιμές, δύναται να προκύψουν είτε από τεχνικές αστοχίες, όπως είναι η κακή λειτουργία αισθητήρων και συσκευών που καταγράφουν τα δεδομένα, προβλήματα κατά τη μετάδοση των δεδομένων, είτε από σφάλματα λογισμικού και χειριστών οι οποίοι τα καταχωρούν.

Σε περιπτώσεις προ επεξεργασίας για δεδομένα με θόρυβο χρησιμοποιούνται τρεις τεχνικές, η μέθοδος Binning, κατά την οποία όλα τα δεδομένα χωρίζονται σε τμήματα ίσου μεγέθους και στη συνέχεια εφαρμόζονται διάφορες μέθοδοι. Η διαχείριση κάθε τμήματος γίνεται ξεχωριστά (Jain, 2021). Στη συνέχεια, οι τιμές έκαστου τμήματος μπορούν να αντικατασταθούν είτε από τη μέση (mean) τιμή του τμήματος, είτε από την τιμή των ορίων του τμήματος, είτε από τη διάμεση (median) τιμή του τμήματος (Anunaya, 2021). Στην παλινδρόμηση (regression) επιδιώκεται να γίνει ομαλοποίηση των δεδομένων μέσω της προσαρμογής τους σε μία συνάρτηση παλινδρόμησης. Τέλος, στην ομαδοποίηση (clustering) δεδομένα με παρόμοια χαρακτηριστικά ομαδοποιούνται σε συστάδες, οπότε υπάρχει πιθανότητα με αυτόν τον τρόπο τα ιδιάζοντα σημεία να βρεθούν εκτός των συστάδων (Jain, 2021).

Ως μετασχηματισμός των δεδομένων ορίζεται η διαδικασία κατά την οποία μεταβάλλεται η δομή των δεδομένων. Συναντώνται διάφορες μέθοδοι μετασχηματισμού, οι οποίες θα αναλυθούν στη συνέχεια.

Πρώτη μέθοδος είναι η μέθοδος της κανονικοποίησης (normalization). Ως κανονικοποίηση ορίζεται η διαδικασία κατά την οποία οι τιμές των δεδομένων ανάγονται σε

ένα εύρος τιμών, για παράδειγμα από 0.0 έως 1.0 ή από -1.0 έως 1.0. Στην τρέχουσα εργασία, επειδή πρόκειται για ένα μοντέλο μηχανικής μάθησης έγινε κανονικοποίηση των τιμών σε εύρος (0,1).

Δεύτερη μέθοδος η οποία αναλύεται είναι η μέθοδος της διακριτικοποίησης (discretization). Η μέθοδος της διακριτικοποίησης εφαρμόζεται για να γίνει ο διαχωρισμός των συνεχών δεδομένων σε διαστήματα. Όπως εύκολα μπορεί να αντιληφθεί κάποιος, η διακριτικοποίηση μειώνει το μέγεθος των δεδομένων.

Τέλος, η μέθοδος της συνάθροισης (aggregation) είναι μία μέθοδος κατά την οποία τα δεδομένα αποθηκεύονται και παρουσιάζονται περιληπτικά. Το σετ δεδομένων το οποίο προέρχεται από διάφορες πηγές, ενοποιείται με την περιγραφή της ανάλυσης δεδομένων. Πρόκειται για ένα σημαντικό βήμα, καθώς η ποιότητα και η ποσότητα των δεδομένων καθορίζουν την ακρίβεια τους. Καλύτερα αποτελέσματα προκύπτουν με δεδομένα καλής ποιότητας και ικανοποιητικής ποσότητας (Anunaya, 2021).

Το τρίτο και τελευταίο βήμα το οποίο αναλύεται, κατά τη διαδικασία της επεξεργασίας δεδομένων, είναι η μείωση των δεδομένων. Πρόκειται για ένα βήμα το οποίο συντελεί στην μείωση του όγκου των δεδομένων. Ως αποτέλεσμα, η ανάλυση των δεδομένων γίνεται με ευκολότερο τρόπο, καθώς δεν υπάρχει αξιόλογη καθυστέρηση για την ανάλυση, ενώ τα αποτελέσματά της οφείλουν να μην επηρεάζονται σημαντικά από τη μείωση των δεδομένων. Ένα ακόμη θετικό στοιχείο της μείωσης των δεδομένων είναι η απελευθέρωση αποθηκευτικού χώρου. Οι μέθοδοι οι οποίες χρησιμοποιούνται για τη μείωση των δεδομένων είναι η μείωση της διάστασης (dimensionality reduction), η μείωση των πολλών αριθμών (numerosity reduction) και η συμπίεση των δεδομένων (data compression).

Κατά τη μείωση της διάστασης δεδομένων, γίνεται ελάττωση τυχαίων μεταβλητών, μέσω συνδυασμού και συγχώνευσης τιμών των δεδομένων, χωρίς όμως να χάσουν τα αρχικά τους δεδομένα. Αυτό έχει ως σκοπό τον περιορισμό της διάστασης του σετ δεδομένων. Πρόκειται για μία αρκετά χρήσιμη διαδικασία, καθώς σε διάφορες εφαρμογές το μέγεθος των δεδομένων είναι τεράστιο. Όπως και στην περίπτωση της μείωσης των δεδομένων, έτσι κι εδώ, η διεργασία αυτή συντελεί στο να απελευθερωθεί αποθηκευτικός χώρος και απαιτείται λιγότερος χρόνος για τον υπολογισμό. Κατά τη μέθοδο της μείωσης των πολλών αριθμών, γίνεται μικρότερη παρουσίαση δεδομένων μέσω ελάττωση του όγκου. Σε αυτό το στάδιο, δεν παρατηρείται απώλεια δεδομένων.

Τέλος, η συμπιεσμένη μορφή των δεδομένων, ορίζεται ως συμπίεση δεδομένων. Πρόκειται για μία διαδικασία κατά την οποία δύναται να υπάρξουν απώλειες δεδομένων, όπως υπάρχει η περίπτωση να μην υπάρξει απώλεια. Σε περιπτώσεις κατά τις οποίες συναντάται μείωση της πληροφορίας, η πληροφορία αυτή η οποία χάνεται είναι αχρείαστη (Anunaya, 2021).

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, πραγματοποιήθηκε κανονικοποίηση των δεδομένων. Τα δεδομένα αφορούν διάφορες κατηγορίες και έκαστη κατηγορία έχει διαφορετικό εύρος τιμών. Συνεπώς, για να μπορέσει το μοντέλο να διαχειριστεί τα εισαχθέντα δεδομένα, να τα επεξεργαστεί και τέλος να βγάλει ένα αποτέλεσμα, κρίθηκε σκόπιμο να γίνει η κανονικοποίησή τους.

3.5.4 Προσδιορισμός Δεδομένων Εκπαίδευσης

Σε ένα μοντέλο μηχανικής μάθησης από το σύνολο των δεδομένων τα οποία εισάγονται στο μοντέλο, χρειάζεται να γίνει ο διαχωρισμός τους σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου.

Τα δεδομένα εκπαίδευσης είναι εκείνα τα δεδομένα τα οποία χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου. Πρόκειται για ένα ποσοστό του αρχικού συνόλου δεδομένων. Το ποσοστό αυτό επιλέγεται ανάλογα με το πλήθος των συνολικών δεδομένων. Σε περιπτώσεις στις οποίες συναντάται μικρό πλήθος δεδομένων, τα δεδομένα εκπαίδευσης συνηθίζεται να παίρνουν ποσοστό 80%, ενώ για μεγαλύτερο πλήθος αρχικών δεδομένων μπορούν να πάρουν ποσοστό 70% ή 67%.

Τα δεδομένα ελέγχου, αποτελούν το υπόλοιπο ποσοστό. Πρόκειται για δεδομένα τα οποία χρησιμοποιούνται για να αξιολογήσουν την απόδοση του μοντέλου. Αντίστοιχα, λαμβάνουν τιμές 20%, σε περιπτώσεις στις οποίες το πλήθος των αρχικών δεδομένων είναι μικρό, ενώ μπορούν για μεγαλύτερα μεγέθη συνολικών δεδομένων να πάρουν τιμές 30% ή 33%.

Ένα μεγαλύτερο πλήθος δεδομένων ελέγχου δύναται να εξασφαλίσει περισσότερη ακρίβεια στον υπολογισμό της απόδοσης ενός μοντέλου μηχανικής μάθησης. Ακόμη, η διαδικασία εκπαίδευσης γίνεται σε λιγότερο χρόνο, καθώς χρησιμοποιείται μικρότερος αριθμός δεδομένων εκπαίδευσης και τα αποτελέσματα τα οποία προκύπτουν είναι πιο αξιόπιστα (Malato, 2020).

Όπως αναφέρεται και στην ενότητα 4.2, για τα δεδομένα τα οποία χρησιμοποιήθηκαν στο μελετώμενο πρόβλημα, το ποσοστό των δεδομένων εκπαίδευσης είναι ίσο με 80%, ενώ το ποσοστό των δεδομένων ελέγχου είναι ίσο με 20%, για κάθε μελετώμενη πόλη.

3.5.5 Επιλογή Αλγορίθμου

Στην υποενότητα 3.3.2 παρουσιάστηκαν ορισμένοι αλγόριθμοι οι οποίοι χρησιμοποιούνται σε προβλήματα μηχανικής μάθησης. Αυτό το οποίο αξίζει να σημειωθεί είναι ότι ένας αλγόριθμος μπορεί να εμφανίσει καλή απόδοση σε ορισμένα προβλήματα, όμως σε κάποια άλλα προβλήματα η απόδοσή του ενδέχεται να είναι μειωμένη.

Έτσι, στην εργασία αυτή, δημιουργούνται μοντέλα από τους προαναφερθέντες αλγορίθμους και επιλέγεται εκείνος ο αλγόριθμος ο οποίος φαίνεται να «προσαρμόζεται» καλύτερα στα δεδομένα και στο ίδιο το πρόβλημα της πρόβλεψης, τόσο των κρουσμάτων όσο και των θανάτων της Covid-19, για τις εννέα διαφορετικές μελετώμενες πόλεις της τρέχουσας εργασίας.

3.5.6 Αξιολόγηση Αποδοτικότητας

Τέλος, ένα ακόμη σημαντικό βήμα κατά τη δημιουργία μοντέλων μηχανικής μάθησης είναι η αξιολόγηση του μοντέλου. Δηλαδή, η αξιολόγηση των αποτελεσμάτων τα οποία προκύπτουν από το εκάστοτε μοντέλο. Όπως αναλύθηκε και στην 3.3.3, συναντώνται διάφορες μετρικές για την αξιολόγηση των μοντέλων. Οι μετρικές αυτές επιλέγονται βάσει του είδος των προβλημάτων, προβλήματα ταξινόμησης – προβλήματα παλινδρόμησης.

Καθώς η πρόβλεψη τιμών για τους θανάτους και τα κρούσματα της Covid-19 ανήκει στα προβλήματα παλινδρόμησης, οι μετρικές οι οποίες χρησιμοποιούνται για την αξιολόγηση των αναπτυχθέντων μοντέλων είναι μετρικές οι οποίες συναντώνται στην κατηγορία των μετρικών παλινδρόμησης. Συγκεκριμένα, για τους σκοπούς της ανάλυσης των χωροχρονικών επιδημιολογικών δεδομένων της παρούσας εργασίας οι μετρικές οι οποίες υπολογίσθηκαν και χρησιμοποιήθηκαν για την ανάλυση είναι οι R^2, η VAR, το MAE, το MAPE, το MSE και το rMSE.

3.6 Python και Μηχανική Μάθηση

Τα μοντέλα μηχανικής μάθησης στην τρέχουσα εργασία, επιλέχθηκε να δημιουργηθούν σε γλώσσα προγραμματισμού Python μέσω του Google Colaboratory (Colab). Πρόκειται για μία γλώσσα προγραμματισμού με πλούσιες βιβλιοθήκες, εύκολη στην κατανόηση και με ένα σημαντικό παγκόσμιο δίκτυο για βοήθεια.

3.6.1 Google Colaboratory

Το Google Colaboratory (Colab) είναι ένα προϊόν από το Google Research. Το Colab επιτρέπει τη δημιουργία κώδικα σε Python και πλούσιου κειμένου σε ένα αρχείο, μαζί με την ύπαρξη εικόνων και LaTeX. Τα αρχεία τα οποία δημιουργούνται αποθηκεύονται στο Google Drive.

Οι περισσότερες εργασίες οι οποίες λαμβάνουν χώρα στο Colab αφορούν προβλήματα μηχανικής μάθησης, ανάλυσης δεδομένων αλλά χρησιμοποιείται και για εκπαιδευτικούς σκοπούς.

Πρόκειται για ένα διαδραστικό περιβάλλον το οποίο δεν απαιτεί κάποια εγκατάσταση, όπως το Jypyter. Είναι εύκολο και εύχρηστο για τη δημιουργία και την εκτέλεση κώδικα. Τέλος, παρέχει δωρεάν πρόσβαση σε υπολογιστικούς πόρους, συμπεριλαμβανομένων των GPUs (Chng, 2022).

3.6.2 Βιβλιοθήκες

Όπως ήδη αναφέρθηκε, η Python διαθέτει πλούσιες βιβλιοθήκες οι οποίες δύναται να χρησιμοποιηθούν για διάφορες μελέτες, συμπεριλαμβανομένων των προβλημάτων μοντελοποίησης και ανάλυσης χωροχρονικών δεδομένων. Στην παρούσα εργασία, χρησιμοποιήθηκαν ορισμένες από τις πιο βασικές βιβλιοθήκες για χρήση μοντέλων μηχανικής μάθησης, για στατιστική ανάλυση, για υπολογισμούς και για οπτικοποιήσεις. Πρόκειται για τις Scikit-learn, NumPy, Pandas, Matplotlib, Seaborn και Yellowbrick.

Η Scikit-learn είναι πιθανότατα η πιο χρήσιμη βιβλιοθήκη μηχανικής μάθησης, ανοιχτού κώδικα, για προγραμματισμό σε Python. Στη Scikit-learn συναντάται ένα εύρος αλγορίθμων ταξινόμησης, παλινδρόμησης, ομαδοποίησης και «παρέχει» τα απαιτούμενα εργαλεία σε προβλήματα μείωσης διαστάσεων. Εμφανίζει διαλειτουργικότητα με βιβλιοθήκες όπως η NumPy και η SciPy. Χρησιμοποιείται, λοιπόν, για να δημιουργεί μοντέλα μηχανικής μάθησης και όχι για να διαβάζει ή/και να επεξεργάζεται τα δεδομένα.

Η NumPy, Numerical Python, είναι μία ευρέως χρησιμοποιούμενη βιβλιοθήκη, ενός πακέτου συναρτήσεων, για επιστημονικούς υπολογισμούς στη Python. Μέσω της χρήσης της εν λόγω βιβλιοθήκης, διευκολύνονται οι υπολογισμοί διανυσμάτων και πινάκων, αλλά και το διάβασμα και η εγγραφή συνόλων δεδομένων. Πολλές από τις βιβλιοθήκες επιστημονικών υπολογισμών της Python, όπως είναι η Scikit-learn και η Pandas, έχουν ως βάση τη NumPy.

Η Pandas είναι ένα πακέτο ανοιχτού κώδικα το οποίο χρησιμοποιείται στην επιστήμη και ανάλυση δεδομένων και σε προβλήματα μηχανικής μάθησης. Ορισμένες από τις εργασίες στις οποίες χρησιμοποιείται είναι η κανονικοποίηση των δεδομένων (data normalization), η στατιστική ανάλυση, η οπτικοποίηση δεδομένων και τέλος η φόρτωση και η αποθήκευση δεδομένων.

Η Matplotlib είναι βιβλιοθήκη η οποία χρησιμοποιείται για τη δημιουργία δισδιάστατων γραφημάτων και για την οπτικοποίηση δεδομένων. Τα γραφήματα μπορούν να δημιουργηθούν άμεσα, εύκολα και με λίγες γραμμές κώδικα. Αξίζει να σημειωθεί ότι τα

γραφήματά της θυμίζουν γραφήματα τα οποία προκύπτουν από τη MATLAB, χωρίς όμως να υπάρχει κάποια άμεση εξάρτηση μεταξύ τους.

Η Seaborn είναι μία βιβλιοθήκη η οποία χρησιμοποιείται για τη δημιουργία δισδιάστατων γραφημάτων. Έχει ως βάση τη Matplotlib. Τα γραφήματά της είναι περισσότερο «αισθητικά όμορφα» από εκείνα της Matplotlib.

Τέλος, η Yellowbrick είναι μία βιβλιοθήκη οπτικοποίησης και διαγνωστικών εργαλείων, επιτρέποντας τη γρηγορότερη επιλογή του βέλτιστου μοντέλου. Συνδυάζει τη δύναμη της Scikit-learn και την ευελιξία της Matplotlib για τη δημιουργία εύληπτων γραφημάτων (Dar, 2018).

Κεφάλαιο 4

Πειραματικά Αποτελέσματα

4.1 Εισαγωγή

Στο τρέχον κεφάλαιο πρόκειται να παρουσιαστεί μία σειρά αποτελεσμάτων τα οποία προέκυψαν μέσω της επεξεργασίας των αρχικών δεδομένων. Συγκεκριμένα, κρίνεται σκόπιμο να αναφερθούν οι ποίκιλες μετρικές αξιολόγησης για μοντέλα παλινδρόμησης, Regression Related Evaluated Metrics, οι οποίες χρησιμοποιήθηκαν στην ανάλυση και την εκτίμηση των τελικών αποτελεσμάτων και συμπερασμάτων. Εν συνεχεία, θίγεται η σπουδαιότητα της απεικόνισης/οπτικοποίησης δεδομένων κατά τη διαδικασία της ανάλυσης και επεξεργασίας τους. Τέλος, παρουσιάζονται οπτικοποιημένα τόσο τα αρχικά δεδομένα, συνοδευόμενα από μία στατιστική ανάλυση, όσο και τα τελικά αποτελέσματα πρόβλεψης για τα πιθανά κρούσματα και τους πιθανούς θανάτους της νόσου Covid-19 και με τεχνικές Επιβλεπόμενης Μηχανικής Μάθησης (Supervised Machine Learning).

4.2 Μετρικές Αξιολόγησης

Όπως αναφέρθηκε και στην **3.3.3** η απόδοση ενός μοντέλου μηχανικής μάθησης αξιολογείται μέσω κατάλληλων μετρικών. Από την αξιολόγηση του μοντέλου, κρίνεται εάν εκείνο μπορεί να χρησιμοποιηθεί, ως βέλτιστο, για το εκάστοτε μελετώμενο πρόβλημα. Στη 4.4 παρουσιάζονται τα αποτελέσματα έπειτα από την εφαρμογή των μοντέλων πρόβλεψης και αξιολογούνται βάσει των παραπάνω μετρητικών.

Σε κάθε περίπτωση, αυτό το οποίο αναμένεται από ένα μοντέλο είναι να παρουσιάζει την καλύτερη δυνατή ακρίβεια. Έτσι, το μοντέλο θα μπορέσει να χρησιμοποιηθεί αποτελεσματικά και να εξάγει ικανοποιητικά αποτελέσματα.

Το πρόβλημα της μοντελοποίησης και της ανάλυσης χρονοσειρών υπάγεται στα προβλήματα της παλινδρόμησης. Συνεπώς, στην περίπτωση της παρούσας διπλωματικής, οι μετρικές οι οποίες χρησιμοποιούνται για την αξιολόγηση των αναπτυχθέντων μοντέλων ανήκουν στην κατηγορία των μετρικών αξιολόγησης μοντέλων παλινδρόμησης.

Μετρικές οι οποίες εφαρμόζονται σε αλγορίθμους παλινδρόμησης είναι το Mean Squared Error (MSE), η Root Mean Squared Error (RMSE), η R^2, το Mean Absolute Error (MAE), το Mean Absolute Percentage Error (MAPE), το Median Absolute Error (MdAE), το Max Error και η Explained Variance Score. Από τις προαναφερθείσες μετρικές, επιλέχθηκε να χρησιμοποιηθούν για την αξιολόγηση των μοντέλων οι εξής: MSE, RMSE, R^2, MAE, MAPE και Explained Variance Score.

Το MAE, Mean Absolute Error, είναι το μέσο απόλυτο, μη αρνητικό, σφάλμα ανάμεσα στην πραγματική και τη προβλεπόμενη τιμή. Δύο θετικά του MAE είναι ότι αφενός γίνεται κατανοητό από τους τελικούς χρήστες ενός συστήματος και αφετέρου, η τιμή του δίνεται σε «μονάδες» της μελετώμενης μεταβλητής. Από την άλλη πλευρά, δύο αρνητικά του στοιχεία είναι ότι εμφανίζει ευαισθησία στα ιδιάζοντα σημεία και δευτερευόντως, τα αποτελέσματα του δεν μπορούν να χρησιμοποιηθούν για τη σύγκριση διαφορετικών μοντέλων (Allwright, What is a good MAE score?, 2021).

Το MAPE, Mean Absolute Percentage Error, είναι μία μετρική αξιολόγησης η οποία χρησιμοποιείται στα μοντέλα παλινδρόμησης. Επιστρέφει το απόλυτο μέσο σφάλμα ως

ποσοστό. Όσο πιο μικρή η τιμή του, τόσο πιο ακριβές είναι και το μοντέλο. Η διαφορά ανάμεσα στο MAE και το MAPE έγκειται στο ότι το MAE είναι μία απόλυτη τιμή, ενώ το MAPE είναι ποσοστό. Δύο από τα θετικά του MAPE είναι ότι μπορεί να γίνει εύκολα κατανοητό τόσο από τους προγραμματιστές όσο και από τους τελικούς χρήστες και χάρη στο MAPE είναι εφικτή η σύγκριση της ακρίβειας δύο διαφορετικών μοντέλων. Τέλος, το ένα από τα δύο αρνητικά του είναι ότι, σε περιπτώσεις στις οποίες συναντώνται τιμές κοντά ή ίσες με το μηδέν, τότε δεν ενδείκνυται η χρήση του, καθώς θα εμφανίσει μεγάλα νούμερα, εφόσον πραγματοποιείται διαίρεση με το μηδέν. Το δεύτερο αρνητικό του είναι ότι εμφανίζει ευαισθησία στα ιδιάζοντα σημεία (Allwright, 2021).

Το MSE, Mean Squared Error, είναι μία μετρική αξιολόγησης μοντέλων παλινδρόμησης. Όσο η τιμή του βρίσκεται πιο κοντά στο 0, τόσο καλύτερη είναι η απόδοση και η ακρίβεια του μοντέλου. Σε αντίθεση με τις δύο προηγούμενες μετρικές, το MSE δεν επιστρέφει τιμή στην ίδια «κλίμακα» της προβλεπόμενης μεταβλητής. Τα αρνητικά της εν λόγω μετρικές υπερέχουν έναντι του ενός θετικού της. Το θετικό της είναι ότι πρόκειται για μία εύκολα εφαρμόσιμη και κατανοήσιμη μετρική. Τα αρνητικά της είναι ότι το σφάλμα δεν δίνεται στις ίδιες μονάδες τις μελετώμενης μεταβλητής, είναι δυσνόητη η ερμηνεία της και είναι ευαίσθητη σε ιδιάζοντα σημεία (Allwright, 2022). Μία ισχυρή μετρική, χρειάζεται να επηρεάζεται λιγότερο από τα ιδιάζοντα σημεία. Το MSE πιθανότατα είναι λιγότερο ισχυρό από το MAE, καθώς μέσω του τετραγώνου των σφαλμάτων δίνεται μεγαλύτερη σπουδαιότητα στα ιδιάζοντα σημεία.

Το RMSE, Root Mean Squared Error, είναι ακόμη μία μετρική αξιολόγησης για μοντέλα παλινδρόμησης. Χρησιμοποιείται για να φανεί πόσο καλά το μοντέλο προσαρμόζεται σε άγνωστα δεδομένα. Η τιμή του RMSE είναι στην ίδια μονάδα με εκείνη της προβλεπόμενης μεταβλητής. Ως εκ τούτου, δεν υπάρχει κάποια καθορισμένη «καλή τιμή». Μία καλή τιμή RMSE εξαρτάται από την κλίμακα των προβλεπόμενων μεταβλητών. Χρειάζεται προσοχή, καθώς πρόκειται για μία μετρική ευαίσθητη στα ιδιάζοντα σημεία και δεν είναι εφικτή η σύγκριση των τιμών της σε διαφορετικούς κλάδους και σετ δεδομένων.

Η EVS ορίζεται το μέτρο εκείνο το οποίο φανερώνει πόσο απέχουν οι πραγματικές τιμές από το μέσο των προβλεπόμενων τιμών. Δηλαδή, τη διαφορά τους από τη μέση προβλεπόμενη τιμή. Συχνά, συγχέεται με τη R^2. Εν τούτοις, η EVS λαμβάνει υπόψιν τη διασπορά των υπολοίπων.

Η R^2 είναι μία μετρική αξιολόγησης της απόδοσης μοντέλων παλινδρόμησης. Εν αντιθέσει με τις προαναφερθείσες μετρικές, δεν περιγράφει πόσο ακριβείς είναι οι μετρήσεις, αλλά πρόκειται για ένα μέτρο καταλληλόλητας. Η R^2 είναι ότι μπορεί να πάρει από αρνητικές τιμές έως +1. Οι αρνητικές τιμές περιγράφουν μοντέλα τα οποία «υπολειτουργούν». Δηλαδή, περιγράφουν περιπτώσεις κατά τις οποίες οι ανεξάρτητες μεταβλητές δεν ερμηνεύουν τη μεταβλητότητα και συνεισφέρουν αρνητικά στο μοντέλο. Μία τιμή ίση με το 1, σημαίνει ότι η μεταβλητότητα ερμηνεύεται από τις ανεξάρτητες μεταβλητές. Μία τιμή ίση με το 0, σημαίνει ότι οι ανεξάρτητες μεταβλητές δεν ερμηνεύουν καμία μεταβλητότητα (Allwright, 2022). Γενικότερα, αυτό το οποίο ισχύει είναι ότι μία χαμηλή τιμή της R^2, συνήθως, φανερώνει ένα χαμηλό επίπεδο συσχέτισης, το οποίο υποδηλώνει ότι το μοντέλο παλινδρόμησης δεν είναι πλήρως σωστό. Τέλος, αξίζει να σημειωθεί ότι η τιμή της εξαρτάται από το είδος του προβλήματος. Για παράδειγμα, εάν τα δεδομένα έχουν κανονική κατανομή και το πρόβλημα είναι γραμμικό, τότε προκύπτει μία «καλή» τιμή για το R^2. Αντιθέτως, εάν τα δεδομένα εμφανίσουν μεγάλη διασπορά και το πρόβλημα είναι μη γραμμικό, τότε η R^2 θα έχει μικρότερη τιμή (Biswas, 2020).

Αξίζει να τονιστεί ότι η επιλογή και χρήση των κατάλληλων και ορθών μετρικών αξιολόγησης είναι αρκετά σημαντική. Ανάμεσα στις μετρικές οι οποίες εφαρμόστηκαν στο πλαίσιο της παρούσας διπλωματικής εργασίας, δεν υπάρχει κάποια η οποία να είναι καλύτερη από κάποια άλλη για την αξιολόγηση ενός μοντέλου παλινδρόμησης.

4.3 Οπτικοποίηση Δεδομένων και Αποτελεσμάτων

Ως οπτικοποίηση δεδομένων ορίζεται η γραφική αναπαράσταση της πληροφορίας και των δεδομένων. Μέσω διαγραμμάτων, γραφημάτων και χαρτών, μπορεί κάποιος να περιεργαστεί και να κατανοήσει ευκολότερα τάσεις και μοτίβα δεδομένων, καθώς επίσης, μπορεί να διακρίνει και τα outliers – ιδιάζοντα σημεία. Η παρουσίαση γραφημάτων δύναται με εύληπτο τρόπο να ενημερώσει πλήρως τους ενδιαφερόμενους αναγνώστες, εάν συγκρίνονταν με την παρουσίαση δεδομένων μέσα από πίνακες.

Ορισμένοι από τους πιο συχνούς τύπους αναπαράστασης δεδομένων είναι τα διαγράμματα – Charts, τα γραφήματα – Graphs, οι χάρτες – Maps, τα πληροφοριακά γραφήματα – Infographics, οι πίνακες – Tables και τα ταμπλό απεικόνισης – Dashboards (Tableau).

Η οπτικοποίηση δεδομένων δεν πρόκειται για μία απλή και ανώδυνη διαδικασία. Απαιτείται ιδιαίτερη προσοχή, ώστε η περιγραφή των δεδομένων να γίνει με τον αρτιότερο δυνατό τρόπο, χωρίς να αποκρυφθούν σημαντικά στοιχεία. Μερικά από τα ερωτήματα τα οποία χρειάζεται να λάβει υπόψιν κάποιος προτού ξεκινήσει τη διαδικασία της οπτικοποίησης σχετίζονται με την ειδικότητα του κοινού, με τις πιθανές ερωτήσεις γύρω από το εξεταζόμενο ζήτημα και με τις απαντήσεις οι οποίες προκύπτουν μέσω της ενασχόλησης με το εν λόγω ζήτημα.

Υπάρχουν τρεις διαφορετικοί τύποι για οπτικοποίηση οι οποίοι εφαρμόζονται για την ανάλυση ενός σετ δεδομένων. Ήτοι: μίας μεταβλητής, δύο μεταβλητών και πολλών μεταβλητών.

Στην περίπτωση της μίας μεταβλητής, univariate distribution, μελετάται πώς τα δεδομένα είναι διανεμημένα ή πώς βρίσκεται η μεταβολή της κατανομής για μία μονή μεταβλητή εντός του συνόλου δεδομένων.

Στην περίπτωση των δύο μεταβλητών, γίνεται προσπάθεια κατανόησης του μοτίβου ή κάποιας υποβόσκουσας δομής ανάμεσα σε δύο συνεχείς μεταβλητές. Σε αυτήν την περίπτωση συναντάται και η γραφική παράσταση των residuals – Residual Plot ανάμεσα σε δύο μεταβλητές καθώς επίσης και η γραφική παράσταση παλινδρόμησης – Regression Plot.

Στην περίπτωση των πολλών μεταβλητών, multivariate distribution, μελετάται η σχέση η οποία υπάρχει είτε ανάμεσα σε αρκετές συνεχείς μεταβλητές, είτε ανάμεσα σε μεταβλητές με διαφορετικούς τύπους (Pattabiraman).

Συναντώνται αρκετά μέσα με τα οποία μπορεί κάποιος να οπτικοποιήσει δεδομένα. Συγκεκριμένα, για την οπτικοποίηση μπορεί να χρησιμοποιηθεί κάποιο λογισμικό, λόγου χάρη Power BI, Tableau, είτε να χρησιμοποιηθεί κώδικας.

Στο πλαίσιο της εργασίας αυτής, επιλέχθηκε να χρησιμοποιηθούν βιβλιοθήκες από τη γλώσσα προγραμματισμού Python. Η Python προσφέρει μεγάλο εύρος εργαλείων και βιβλιοθηκών για οπτικοποίηση δεδομένων. Οι βιβλιοθήκες μπορούν να βοηθήσουν στη δημιουργία διαδραστικών και εξατομικευμένων διαγραμμάτων, γραφημάτων και σχεδίων με ευνόητες εντολές και λειτουργίες. Μερικές από τις ευρέως χρησιμοποιούμενες βιβλιοθήκες της Python είναι η Matplotlib, η Seaborn, η Potly, η Folium, η Ggplot, η Bokeh, η Pygal, η Geoplotlib, η Glearn, η Missingno και η Altair.

Στην παρούσα εργασία για την οπτικοποίηση των αρχικών δεδομένων, αλλά και των τελικών αποτελεσμάτων, όπως έχει αναφερθεί και στο Κεφάλαιο 3, χρησιμοποιήθηκε η Matplotlib, η Seaborn και η Yellowbrick Οι περισσότερες βιβλιοθήκες οπτικοποίησης έχουν αναπτυχθεί επάνω στη Matplotlib.

Αναλυτικότερα, η Matplotlib είναι η περισσότερο δημοφιλής βιβλιοθήκη για οπτικοποίηση στην Python. Πρόκειται για μία βιβλιοθήκη η οποία επιτρέπει τη δημιουργία στατικών, δυναμικών και διαδραστικών οπτικοποιήσεων. Χρησιμοποιεί αντικειμενοστραφή API για να ενσωματώσει γραφήματα σε εφαρμογές με χρήση Python. Η Seaborn είναι μία βιβλιοθήκη η οποία έχει βάσεις στη Matplotlib. Προσφέρει περιβάλλον υψηλού επιπέδου για το σχεδιασμό ελκυστικών και κατατοπιστικών στατιστικών γραφημάτων.

4.3.1 Ανάλυση και Οπτικοποίηση Αρχικών Δεδομένων

Κρίνεται αναγκαίο να γίνει οπτικοποίηση των αρχικών δεδομένων, ώστε να εντοπιστούν πιθανά μοτίβα συσχέτισης ανάμεσα στα δεδομένα, καθώς επίσης και να γίνει κατανόηση των τάσεών τους. Ακόμη μέσω της οπτικοποίησης των αρχικών δεδομένων δύναται να πραγματοποιηθεί ένας οπτικός έλεγχος των τιμών τους για εντοπισμό πιθανών λαθών, αλλά και ιδιαζόντων σημείων - outliers.

Στο πλαίσιο της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκαν δεδομένα για εννέα πόλεις, από εννέα διαφορετικές χώρες. Συγκεκριμένα, οι πόλεις αυτές είναι: Η Αθήνα, η Μόσχα, η Λισαβόνα, η Πράγα, η Μαδρίτη, το Παρίσι, το Λονδίνο, οι Βρυξέλες και το Βερολίνο.

Στον παρακάτω Πίνακας παρουσιάζεται αναλυτικότερα η χρονική περίοδος μελέτης για κάθε πόλη.

Πόλη	**Περίοδος Μελέτης**
Αθήνα	06/03/2020 – 29/12/2020
Μόσχα	11/03/2020 – 29/12/2020
Λισαβόνα	09/03/2020 – 26/12/2020
Πράγα	15/03/2020 – 29/12/2020
Μαδρίτη	06/03/2020 – 29/12/2020
Παρίσι	03/04/2020 – 29/12/2020
Λονδίνο	25/02/2020 – 29/12/2020
Βρυξέλες	22/03/2020 – 02/12/2020
Βερολίνο	25/02/2020 – 29/12/2020

Πίνακας 2 *Πόλεις και περίοδος μελέτης, Ιδία Επεξεργασία*

Τα δεδομένα τα οποία χρησιμοποιήθηκαν, όπως έχει ήδη αναφερθεί στο Κεφάλαιο 2, αφορούν διάφορες παραμέτρους της ατμόσφαιρας, όπως είναι οι ρυπαντές, η πίεση, ο άνεμος, η υγρασία και η θερμοκρασία, διάφορα πληθυσμιακά στοιχεία, όπως η μέση ηλικία, ο ηλικιωμένος πληθυσμός, άνδρες και γυναίκες καπνιστές, διαβητικοί και δείκτης καρδιαγγειακού κινδύνου, καθώς επίσης χρησιμοποιήθηκαν στοιχεία για τις μετακινήσεις με επαγγελματικά αυτοκίνητα, για το πράσινο κάθε περιοχής, όπως επίσης χρησιμοποιήθηκαν στοιχεία για το ακαθάριστο εγχώριο προϊόν κάθε πόλης.

Στις ακόλουθες υποενότητες, ακολουθεί μία σύντομη ανάλυση από την οπτικοποίηση μονοδιάστατων αλλά και πολυδιάστατων μεταβλητών, ανά πόλη.

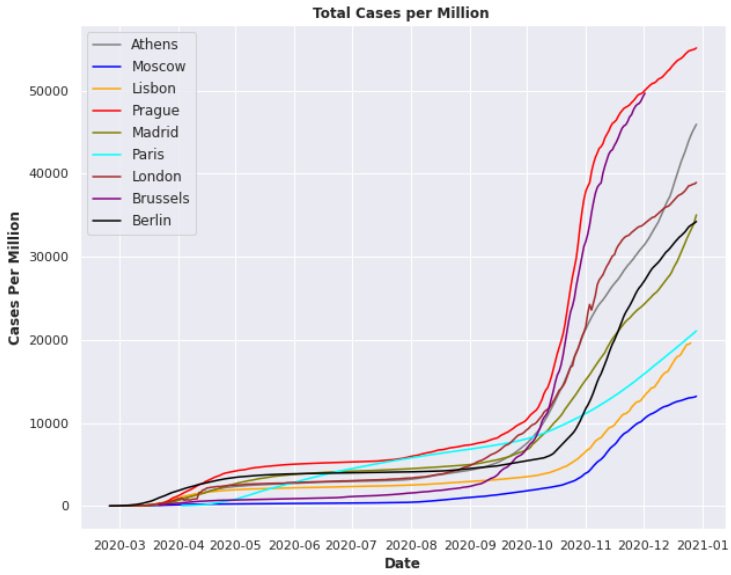
4.3.2 Ανάλυση και Οπτικοποίηση Αρχικών Δεδομένων – Περίπτωση Κρουσμάτων

Σε αρχικό στάδιο, γίνεται μία σύντομη παρουσίαση τόσο των συχνοτήτων των κρουσμάτων όσο και της ροής των κρουσμάτων, για έκαστη πόλη, στο χρονικό ορίζοντα για τον οποίο πραγματοποιείται η μελέτη.

Μέσω μίας ευθύγραμμης γραφικής παράστασης (Line Plot) απεικονίζεται η σχέση ανάμεσα σε δύο μεταβλητές από τις οποίες συχνά η μία είναι ο χρόνος και αποτυπώνεται στον άξονα x, καθώς είναι περισσότερο αντιληπτό στο ανθρώπινο μάτι η σύνδεση μεμονωμένων σημείων με μία γραμμή και με αυτόν τον τρόπο μπορεί να παρατηρηθεί μία τάση ή μία εποχικότητα.

Έτσι, χρησιμοποιείται κι εδώ μία τέτοιου είδους γραφική παράσταση, απεικονίζοντας τη σχέση η οποία επικρατεί ανάμεσα στο χρόνο και στο συνολικό αριθμό των κρουσμάτων. Σε αυτήν την περίπτωση, πρόκειται για ένα ζήτημα δύο μεταβλητών· του χρόνου και του συνολικού αριθμού κρουσμάτων.

Στην περίπτωση των συχνοτήτων, όπως είναι ήδη γνωστό από τη στατιστική, η οπτικοποίηση πραγματοποιείται με τη βοήθεια διαγραμμάτων κατανομών, στα οποία συμπεριλαμβάνονται και τα ιστογράμματα, για τη μονοδιάστατη μεταβλητή "total_cases_per_million", δηλαδή για το συνολικό αριθμό κρουσμάτων, ανά εκατομμύριο.

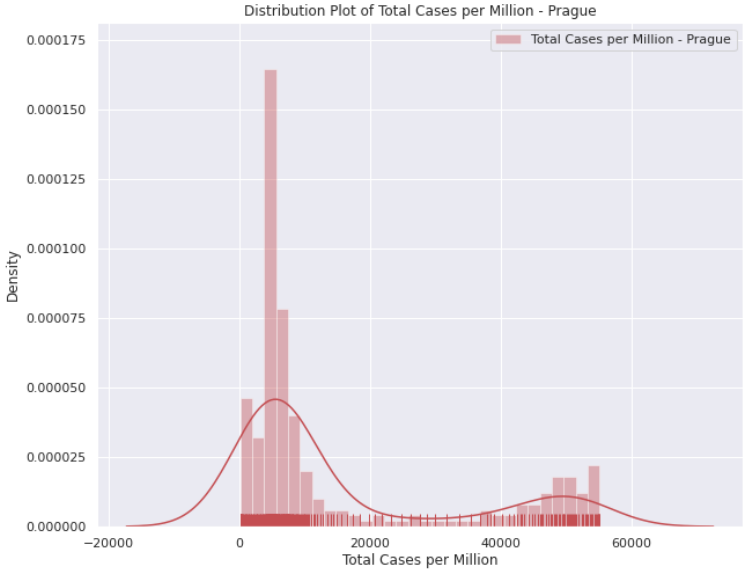


Σχήμα 12 *Συνολικός αριθμός κρουσμάτων ανά εκατομμύριο και ανά πόλη, Ιδία επεξεργασία*

Στο παραπάνω γράφημα παρουσιάζεται η σχέση η οποία υπάρχει ανάμεσα στις τιμές των κρουσμάτων για όλες τις μελετώμενες πόλεις και στη χρονική περίοδο μελέτης.

Αναλύοντας την τάση των κρουσμάτων, αυτό το οποίο προκύπτει είναι μία έντονη αύξηση του συνολικού αριθμού των τιμών των κρουσμάτων, ανά εκατομμύριο, από την περίοδο του Οκτωβρίου του 2020 και για τις εννέα πόλεις. Μελετώντας τις κλίσεις των γραφικών παραστάσεων εκάστης πόλης, παρατηρείται ότι έως και τον Οκτώβρη του 2020, ο αριθμός των κρουσμάτων φαίνεται να παρουσιάζει μικρές αυξήσεις, αλλά δίχως έντονες εξάρσεις. Εν τούτοις, από τα μέσα Οκτώβρη και έπειτα, παρατηρείται μία έντονη εκτίναξη του αριθμού των κρουσμάτων για όλες τις πόλεις. Συγκεκριμένα, εντονότερη αύξηση φαίνεται να παρουσιάζει τόσο η Πράγα όσο και οι Βρυξέλες, ενώ πιο ήπια αυξητική τάση φαίνεται να παρουσιάζει η Μόσχα και το Παρίσι.

Σύμφωνα με την απεικόνιση η οποία προηγήθηκε, **Σχήμα 12**, κρίνεται σκόπιμο να παρατεθούν ενδεικτικά δύο διαγράμματα κατανομών για την πόλη της Πράγας λόγω της έντονης αυξητικής τάσης των κρουσμάτων, από την περίοδο του Οκτωβρίου κι έπειτα και λόγω των μεγαλύτερων τιμών κρουσμάτων ανά εκατομμύριο, συγκριτικά με τις υπόλοιπες οκτώ μελετώμενες πόλεις. Το δεύτερο διάγραμμα αφορά την πόλη της Μόσχας στην οποία παρατηρείται ήπια αυξητική μεταβολή του αριθμού των κρουσμάτων την περίοδο του Οκτωβρίου, καθώς επίσης παρατηρούνται οι χαμηλότερες τιμές κρουσμάτων.



Σχήμα 13 *Διάγραμμα κατανομών συνολικών κρουσμάτων – Πράγα, Ιδία επεξεργασία*

Στο **Σχήμα 13** παρουσιάζεται το διάγραμμα κατανομών για τις τιμές των συνολικών κρουσμάτων για την πόλη της Πράγας. Πρόκειται για ένα διάγραμμα το οποίο αποτελείται από ένα ιστόγραμμα συχνοτήτων, από τη γραφική παράσταση Kernel Density Estimation (KDE) και από τη γραφική παράσταση rug. Και τα τρία αποτελούν διαφορετικούς τρόπους απεικόνισης της κατανομής των δεδομένων.

Εν συντομία, σε ένα ιστόγραμμα συχνοτήτων απεικονίζεται πόσο συχνά κάθε διαφορετική μεταβλητή από ένα σύνολο δεδομένων εμφανίζεται.

Μέσω του KDE μπορεί να υπολογισθεί η συνάρτηση πυκνότητας πιθανότητας για το εκάστοτε πεπερασμένο σετ δεδομένων, με μη παραμετρικό τρόπο. Στην περίπτωση αυτή, δεν θεωρείται ότι υπάρχει κάποια υποκείμενη κατανομή στα δεδομένα. Στη βιβλιοθήκη της Seaborn, η οποία και χρησιμοποιήθηκε για την απεικόνιση των αρχικών δεδομένων, υπάρχει ως προεπιλογή το Gaussian Kernel.

Η γραφική παράσταση rug χρησιμοποιείται για την οπτικοποίηση κατανομής των δεδομένων. Συγκεκριμένα, γίνεται απεικόνιση της οριακής κατανομής μέσω μικρών ευθύγραμμων τμημάτων τόσο στον άξονα x όσο και στον άξονα y. Ακόμη, μέσω της συνάρτησης αυτής επιδιώκεται η συμπλήρωση άλλων γραφημάτων, όπως το ιστόγραμμα συχνοτήτων και το KDE, μέσω της τοποθέτησης των μεμονωμένων παρατηρήσεων με ένα διακριτικό τρόπο. Η εν λόγω γραφική παράσταση είναι ανάλογη ενός ιστογράμματος με ραβδώσεις μηδενικού πάχους ή ανάλογη με ένα μονοδιάστατο διάγραμμα διασποράς.

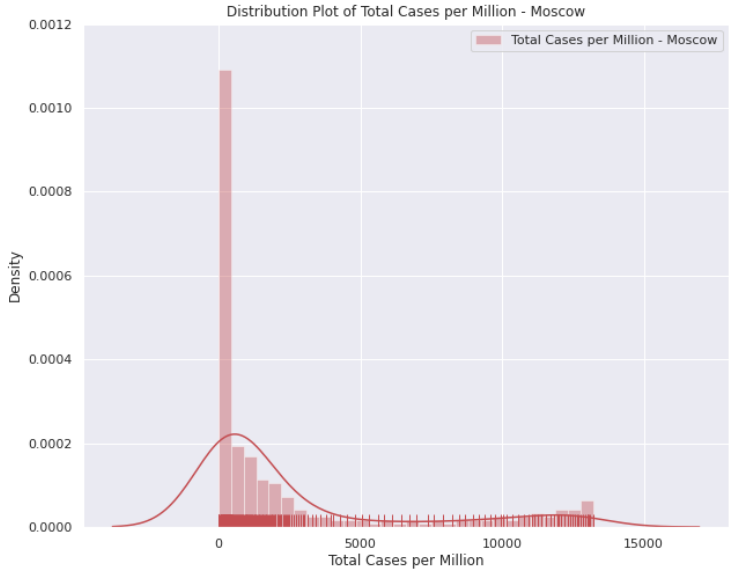
Παρατηρώντας το συγκεκριμένο διάγραμμα, προκύπτει ότι συναντάται μία σημαντική συγκέντρωση παρατηρήσεων για ένα χαρακτηριστικό εύρος τιμών. Πρόκειται για το εύρος από 0 έως περίπου 10000 κρουσμάτων. Ακόμη, παρατηρείται μία συγκέντρωση παρατηρήσεων για ένα μεγαλύτερο εύρος τιμών το οποίο βρίσκεται γύρω στο 50000. Για το

εύρος από περίπου 18000 έως 43000 δεν παρατηρείται ιδιαίτερη συγκέντρωση παρατηρήσεων.

Τα παραπάνω προκύπτουν αφενός από το ιστόγραμμα συχνοτήτων το οποίο παρουσιάζει μέγιστη τιμή περίπου για τα 8000 συνολικά κρούσματα, αφετέρου από το KDE το οποίο φαίνεται ότι γύρω στο 8000 αποκτά μία μέγιστη τιμή, της τάξεως 0.000045, και στη συνέχεια αρχίζει να φθίνει, καθώς και από τα ευθύγραμμα τμήματα του rug plot είναι πιο πυκνά, από το 0 έως περίπου το 10000. Ακόμη, υψηλές τιμές του ιστογράμματος εμφανίζονται μετά από τα περίπου 43000 συνολικά κρούσματα, οι οποίες βέβαια δεν ξεπερνάνε τις τιμές για το εύρος 0 έως 10000.

Μέσω σύγκρισης ανάμεσα στο **Σχήμα 12** και **Σχήμα 13**, επιβεβαιώνεται ότι από το Μάρτιο του 2020 έως και τον Οκτώβρη του 2020 οι τιμές των συνολικών κρουσμάτων στην Πράγα κυμαίνονταν σε ένα σχετικά σταθερό επίπεδο, λιγότερο του 10000. Όμως, από τον Οκτώβρη και μετά, για ένα πολύ μικρό χρονικό διάστημα παρατηρείται έντονη εκτίναξη του αριθμού των κρουσμάτων η οποία φαίνεται να «ελέγχεται» μετά από τις 43000.

Στο παρακάτω διάγραμμα, **Σχήμα 14**, παρουσιάζεται το διάγραμμα κατανομών για τις τιμές των συνολικών κρουσμάτων για τη Μόσχα. Πρόκειται για ένα διάγραμμα το οποίο αποτελείται από ένα ιστόγραμμα συχνοτήτων, από τη γραφική παράσταση Kernel Density Estimation (KDE) και από τη γραφική παράσταση rug.



Σχήμα 14 *Διάγραμμα κατανομών συνολικών κρουσμάτων – Μόσχα, Ιδία επεξεργασία*

Παρατηρώντας το γράφημα, αυτό το οποίο σημειώνεται είναι μία συγκέντρωση παρατηρήσεων για ένα εύρος τιμών από 0 έως περίπου 2000. Στη συνέχεια, παρατηρείται αραίωση των παρατηρήσεων από περίπου 2500 έως και 12000. Από τις 12500 παρατηρείται αύξηση της συγκέντρωσης των κρουσμάτων, αλλά όχι στον ίδιο βαθμό της αρχικής συγκέντρωσης.

Τα παραπάνω προκύπτουν αφενός από το ιστόγραμμα συχνοτήτων το οποίο παρουσιάζει μέγιστη τιμή περίπου για τα 1000 συνολικά κρούσματα, αφετέρου από το KDE το οποίο φαίνεται ότι γύρω στο 1000 αποκτά μία μέγιστη τιμή, της τάξεως 0.0002. Η τιμή αυτή είναι μεγαλύτερη από τη μέγιστη τιμή η οποία συναντάται στην πόλη της Πράγας με περισσότερα κρούσματα, και στη συνέχεια αρχίζει να φθίνει. Ακόμη, τα ευθύγραμμα τμήματα του rug plot είναι πιο πυκνά, από το 0 έως περίπου το 2500. Ακόμη, υψηλές τιμές του ιστογράμματος

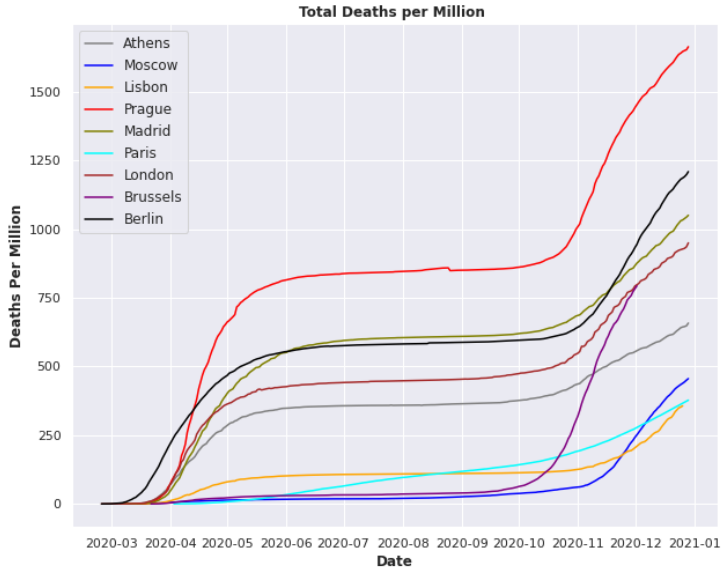
εμφανίζονται μετά από τα περίπου 12000 συνολικά κρούσματα, οι οποίες βέβαια δεν ξεπερνάνε τις τιμές για το εύρος 0 έως 0.

Μέσω σύγκρισης ανάμεσα στο **Σχήμα 12** και **Σχήμα 14**, επιβεβαιώνεται ότι από το Μάρτιο του 2020 έως και τον Οκτώβρη του 2020 οι τιμές των συνολικών κρουσμάτων στην Μόσχα κυμαίνονταν σε ένα σχετικά σταθερό επίπεδο, λιγότερο των 5000. Όμως, από το Νοέμβρη και μετά, για ένα πολύ μικρό χρονικό διάστημα παρατηρείται έντονη αύξηση του αριθμού των κρουσμάτων η οποία φαίνεται να «ελέγχεται» μετά από τις 12000.

4.3.3 Ανάλυση και Οπτικοποίηση Αρχικών Δεδομένων – Περίπτωση Θανάτων

Ομοίως και στην περίπτωση των θανάτων, γίνεται μία σύντομη παρουσίαση τόσο των συχνοτήτων των θανάτων όσο και της ροής των θανάτων, στο χρονικό ορίζοντα για τον οποίο πραγματοποιείται η μελέτη.

Οι μέθοδοι οι οποίες χρησιμοποιήθηκαν είναι ίδιες με εκείνες οι οποίες αναφέρθηκαν στο εδάφιο **4.3.2**. Δηλαδή, χρησιμοποιήθηκε η Line Plot, το KDE και η γραφική παράσταση rug. Η μόνη διαφορά έγκειται στη μονοδιάστατη μεταβλητή "total_deaths_per_million", δηλαδή για το συνολικό αριθμό θανάτων, ανά εκατομμύριο, η οποία χρησιμοποιήθηκε για τη δημιουργία των γραφημάτων.



Σχήμα 15 *Συνολικός αριθμός θανάτων ανά εκατομμύριο και ανά πόλη, Ιδία επεξεργασία*

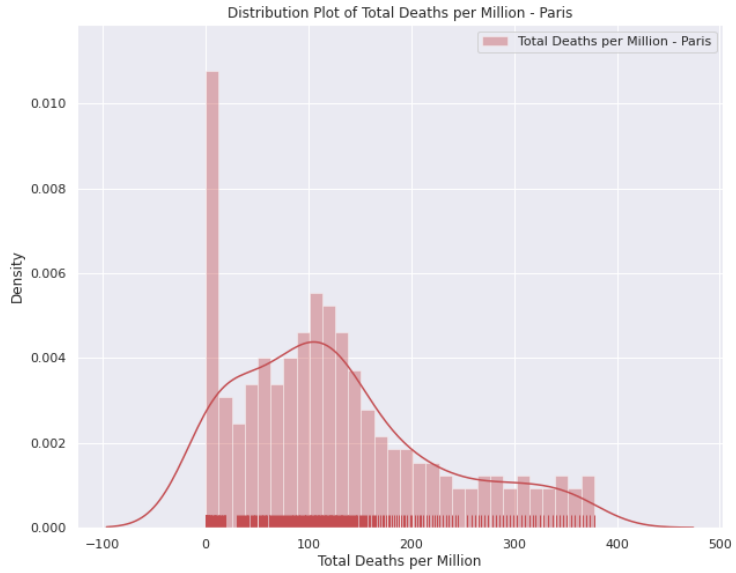
Στο **Σχήμα 15** παρουσιάζεται η σχέση η οποία υπάρχει ανάμεσα στις τιμές των θανάτων για όλες τις μελετώμενες πόλεις και στη χρονική περίοδο μελέτης.

Αναλύοντας την τάση των θανάτων, αυτό το οποίο προκύπτει είναι μία έντονη αύξηση των τιμών των θανάτων, ανά εκατομμύριο, στην περίοδο του Απριλίου του 2020 για πέντε πόλεις, Πράγα, Βερολίνο, Αθήνα, Λονδίνο και Μαδρίτη. Αυτή αύξηση συνοδεύεται για τους επόμενους έξι μήνες, δηλαδή έως και τις αρχές Νοέμβρη, από μια αρκετά ήπια αυξητική μεταβολή. Για τη Μόσχα, τη Λισαβόνα, τις Βρυξέλλες και το Παρίσι, από τον Απρίλιο του 2020 παρατηρείται μία σχετικά ήπια αυξητική τάση, έως και το Νοέμβριο του 2020 όπου και για τις εννέα πόλεις παρατηρείται ξανά μία έντονη αύξηση των συνολικών θανάτων, σε μικρό χρονικό διάστημα.

Παρατηρώντας τις κλίσεις των γραφικών παραστάσεων, προκύπτει ότι για την περίοδο του Απριλίου εντονότερη αυξητική τάση του αριθμού των θανάτων εμφανίζει η Πράγα και για την περίοδο του Νοεμβρίου έντονη αυξητική τάση του αριθμού των θανάτων εμφανίζει τόσο η Πράγα όσο και οι Βρυξέλλες. Αντιθέτως, μικρότερη αυξητική τάση για την περίοδο του

Απριλίου φαίνεται να εμφανίζει η Μόσχα και για την περίοδο του Νοεμβρίου ηπιότερη αυξητική τάση εμφανίζει τόσο η Αθήνα όσο και το Παρίσι. Τέλος, παρατηρείται ότι από τα μέσα Μαΐου έως και τα μέσα Οκτωβρίου, για όλες τις πόλεις η αύξηση του συνολικού αριθμού θανάτων είναι ήπια, καθώς δεν εμφανίζει έντονες μεταβολές.

Σύμφωνα με την απεικόνιση η οποία προηγήθηκε, **Σχήμα 15**, κρίνεται σκόπιμο να παρατεθούν ενδεικτικά δύο διαγράμματα κατανομών για την πόλη της Πράγας με τις έντονες μεταβολές και τις μεγαλύτερες τιμές θανάτων ανά εκατομμύριο, συγκριτικά με τις υπόλοιπες οκτώ μελετώμενες πόλεις, και για την πόλη του Παρισιού στην οποία εμφανίζονται ήπιες μεταβολές και σχετικά χαμηλές τιμές συνολικών θανάτων.



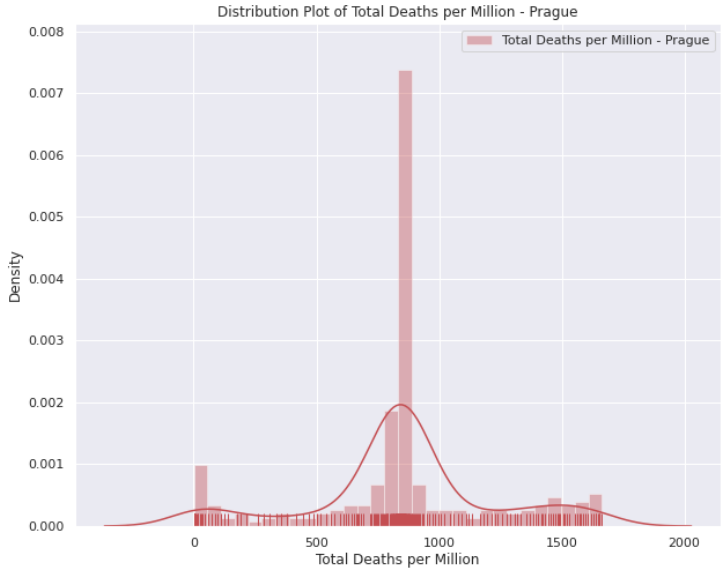
Σχήμα 16 *Διάγραμμα κατανομών συνολικών θανάτων – Παρίσι, Ιδία επεξεργασία*

Στο **Σχήμα 16** παρουσιάζεται ένα διάγραμμα κατανομών για τις τιμές των συνολικών θανάτων για την πόλη του Παρισιού. Το εικονιζόμενο διάγραμμα αποτελείται από ένα ιστόγραμμα συχνοτήτων, από τη γραφική παράσταση Kernel Density Estimation (KDE) και από τη γραφική παράσταση rug.

Αυτό το οποίο παρατηρείται στο **Σχήμα 16** είναι ότι υπάρχει μεγάλη συγκέντρωση παρατηρήσεων για συγκεκριμένο εύρος τιμών. Αυτό προκύπτει αφενός από τις τιμές του ιστογράμματος συχνοτήτων, αφετέρου από το KDE το οποίο φαίνεται ότι γύρω στο 120 αποκτά μία μέγιστη τιμή, της τάξεως 0.004, και στη συνέχεια αρχίζει να φθίνει. Επίσης, τα ευθύγραμμα τμήματα του rug plot είναι πιο πυκνά, από το 0 έως περίπου το 160, με ένα μικρό κενό γύρω στο 20.

Εάν συγκριθούν το **Σχήμα 15** και το **Σχήμα 16**, μπορεί εύκολα να παρατηρηθεί ότι πράγματι για μία μεγάλη περίοδο, από το Μάρτιο του 2020 έως και το Νοέμβριο του 2020, οι τιμές των συνολικών θανάτων στο Παρίσι ήταν σε παρόμοια και χαμηλά επίπεδα, μικρότερα των 200 θανάτων ανά εκατομμύριο. Για την ίδια περίοδο οι θάνατοι ανά εκατομμύριο κυμαίνονταν γύρω στους 100-150.

Στο παρακάτω, **Σχήμα 17**, παρουσιάζεται ένα διάγραμμα κατανομών για τις τιμές των συνολικών θανάτων για την πόλη της Πράγας.



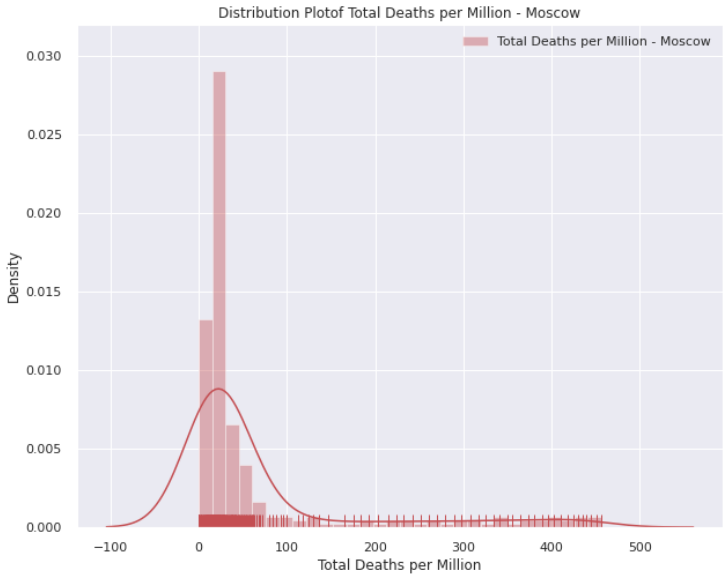
Σχήμα 17 *Διάγραμμα κατανομών συνολικών θανάτων – Πράγα, Ιδία επεξεργασία*

Παρατηρώντας το παραπάνω γράφημα, προκύπτει ότι υπάρχει μεγάλη συγκέντρωση παρατηρήσεων σε ένα συγκεκριμένο, διαφορετικό και μικρότερο εύρος, συγκριτικά με εκείνο της πόλης του Παρισιού. Αυτό προκύπτει αφενός από το ιστόγραμμα συχνοτήτων το οποίο εμφανίζει μέγιστες τιμές για το εύρος από 750 έως 900, αφετέρου από το KDE το οποίο φαίνεται ότι γύρω στο 900 αποκτά μία μέγιστη τιμή της τάξεως 0.002 και στη συνέχεια αρχίζει να φθίνει. Τέλος, από τα ευθύγραμμα τμήματα του rug plot τα οποία είναι πιο πυκνά, περίπου από το 750 έως περίπου το 900. Να σημειωθεί ότι οι τιμές του KDE είναι μικρότερες για την πόλη της Πράγας, συγκριτικά με εκείνες του Παρισιού.

Μέσω σύγκρισης από το **Σχήμα 15** και το **Σχήμα 17**, μπορεί εύκολα να παρατηρηθεί ότι πράγματι για μία μεγάλη περίοδο, από το Μάη του 2020 έως και το Οκτώβρη του 2020, οι τιμές των συνολικών θανάτων στην Πράγα ήταν σε σταθερά επίπεδα γύρω στο 800-850. Συνεπώς, αρκετές από τις παρατηρήσεις θα βρίσκονται σε αυτό το εύρος τιμών.

Ιδιαίτερο ενδιαφέρον θα παρουσίαζε και η οπτικοποίηση του αντίστοιχου γραφήματος για την πόλη της Μόσχας, στην περίπτωση των συνολικών θανάτων. Μέσω του συγκεκριμένου γραφήματος καθίσταται δυνατή η σύγκριση των δύο μελετώμενων περιπτώσεων, κρουσμάτων/θανάτων. Με αυτόν τον τρόπο μπορεί να μελετηθεί η κατανομή και η τάση των θανάτων στην πόλη της Μόσχας και να εντοπιστεί κάποια πιθανή συσχέτιση με την κατανομή και την τάση των κρουσμάτων.

Από το διάγραμμα στο **Σχήμα 15** γίνεται αντιληπτό ότι η Μόσχα εμφανίζει τις μικρότερες τιμές συνολικών θανάτων, συγκριτικά με τις υπόλοιπες πόλεις, από τον Απρίλιο του 2020 έως τα τέλη του Νοεμβρίου 2020. Από τα τέλη Νοεμβρίου, όμως, και έπειτα, εμφανίζει μία αυξητική μεταβολή. Αυτή η τάση, αποτυπώνεται και στο **Σχήμα 18**, στην οποία διακρίνεται ότι για ένα εύρος συνολικών θανάτων 0 έως περίπου 80, παρατηρείται σημαντική συγκέντρωση. Ο KDE λαμβάνει μέγιστη τιμή, δηλαδή περίπου 0.008, κοντά στην περιοχή του 40. Το ιστόγραμμα συχνοτήτων εμφανίζει μέγιστη τιμή στην ίδια περιοχή και η γραφική παράσταση rug παρουσιάζει αρκετές συγκεντρωμένες τιμές στο εύρος 0 έως 80.



Σχήμα 18 *Διάγραμμα κατανομών συνολικών θανάτων – Μόσχα, Ιδία επεξεργασία*

Λαμβάνοντας υπόψιν και το **Σχήμα 12**, η αύξηση των κρουσμάτων στις αρχές του Νοεμβρίου 2020, για την πόλη της Μόσχας, φαίνεται να συνοδεύεται και από αύξηση των θανάτων στα τέλη Νοεμβρίου 2020.

4.3.4 Συμπεράσματα

Από την παραπάνω μελέτη της οπτικοποίησης των αρχικών δεδομένων, προκύπτουν ορισμένα αξιόλογα συμπεράσματα, τα οποία λειτουργούν ως αρωγοί στην αποδοτικότερη μοντελοποίηση η οποία ακολουθεί.

Παρατηρείται ότι οι πόλεις, εμφανίζουν διαφορετική κατανομή τόσο των κρουσμάτων, όσο και των θανάτων, για την ίδια χρονική περίοδο μελέτης. Αυτό, όσον αφορά το κομμάτι της μηχανικής μάθησης, αυξάνει τη δυσκολία στο μοντέλο να διαχειριστεί αποτελεσματικά τα συνολικά δεδομένα, καθώς καλείται να συνδυάσει και να χρησιμοποιήσει δεδομένα από διαφορετικές κατανομές, με διαφορετικά βάρη. Γι' αυτό το λόγο επιλέχθηκε να δημιουργηθούν μοντέλα πρόβλεψης κρουσμάτων και μοντέλα πρόβλεψης θανάτων για κάθε πόλη, ξεχωριστά.

Ακόμη, παρατηρείται ότι το χρονικό διάστημα από τον Απρίλιο του 2020 έως τον Οκτώβρη του 2020 τα κρούσματα παραμένουν σε «ελεγχόμενα» επίπεδα. Εν τούτοις, από τον Οκτώβρη και μετά, σημειώνεται μία σημαντική εκτίναξη των κρουσμάτων. Αντίστοιχα, από τον Απρίλιο του 2020 έως και τα μέσα Οκτωβρίου, οι συνολικοί θάνατοι ανά εκατομμύριο εμφανίζουν μία αρκετά ήπια αυξητική τάση, η οποία όμως εκτινάσσεται από τέλη Οκτώβρη/αρχές Νοέμβρη. Πιθανότατα αυτό το γεγονός θα μπορούσε να συσχετιστεί με την ραγδαία αύξηση των κρουσμάτων, την ίδια χρονική περίοδο. Συμπεραίνεται, λοιπόν, ότι την περίοδο του Οκτωβρίου και Νοεμβρίου 2020, εμφανίζεται ένα έντονο δεύτερο κύμα.

Τέλος, μπορεί εύκολα να γίνει αντιληπτό ότι όσο μεγαλύτερο εύρος τιμών υπάρχει, τόσο μικρότερες θα είναι οι τιμές του διαγράμματος του KDE. Αυτό συμβαίνει καθώς σε μικρότερο εύρος τιμών, οι συγκεντρώσεις «περιορίζονται», ενώ για μεγαλύτερο εύρος τιμών, οι συγκεντρώσεις μπορούν να καταλάβουν περισσότερο χώρο.

4.4 Μοντέλα Μηχανικής Μάθησης

Σε αυτήν την υποενότητα αναφέρονται τα μοντέλα και οι αλγόριθμοι μηχανικής μάθησης οι οποίοι χρησιμοποιήθηκαν για την πρόβλεψη τόσο των κρουσμάτων όσο και των θανάτων της

Covid-19. Στη συνέχεια, παρουσιάζονται τα αποτελέσματα κάθε μεθόδου αφενός για τις προβλεπόμενες τιμές των κρουσμάτων και των θανάτων, αφετέρου για τις τιμές των σφαλμάτων τα οποία συνοδεύουν κάθε μοντέλο πρόβλεψης. Πρόκειται για το κρισιμότερο στάδιο της παρούσας μελέτης, καθώς αξιολογείται κάθε μέθοδος και συγκρίνονται όλες οι μέθοδοι μεταξύ τους με τελικό στόχο την επιλογή του καταλληλότερου αλγορίθμου πρόβλεψης για το χρησιμοποιηθέν σύνολο δεδομένων.

Στην παρούσα εργασία, σύμφωνα και με την ενότητα 3.3.2 από το Κεφάλαιο 3, χρησιμοποιήθηκαν τα κάτωθι μοντέλα: Multiple Linear Regression, Support Vector Regression, LASSO Regression, Gaussian Process Regression, Random Forest Model και XGBoost.

Υπενθυμίζεται ότι η γραμμική παλινδρόμηση – Linear Regression, είναι μία από τις πιο απλές τεχνικές μοντελοποίησης για τεχνικές επιβλεπόμενης μάθησης. Τα μοντέλα της γραμμικής παλινδρόμησης χρησιμοποιούνται για να παρουσιάσουν ή να προβλέψουν τη σχέση ανάμεσα σε δύο ή περισσότερες μεταβλητές.

Στη συνέχεια, το Support Vector Regression παρέχει τη δυνατότητα καθορισμού του μέγιστου επιτρεπόμενου σφάλματος το οποίο είναι αποδεκτό από το μοντέλο και στη συνέχεια, εντοπίζει μία κατάλληλη γραμμή η οποία εφαρμόζεται στα δεδομένα (Sharp, 2020).

Όσο αναφορά το LASSO Regression, η λέξη LASSO προέρχεται από τα αρχικά "**L**east **A**bsolute **S**hrinkage and **S**election **O**perator". Πρόκειται για ένα μοντέλο το οποίο χρησιμοποιείται επάνω σε μεθόδους παλινδρόμησης για πιο ακριβή πρόβλεψη. Μία στατιστική φόρμουλα για τεχνικές κανονικοποίησης. Το μοντέλο αυτό συρρικνώνει τις τιμές των δεδομένων σε ένα κεντρικό σημείο (Great Learning Team, 2021).

Το Gaussian Process Regression (GPR) είναι μία μη παραμετρική, Μπεϊζιανή προσέγγιση στην παλινδρόμηση. Εμφανίζει αρκετά πλεονεκτήματα και δύναται να λειτουργήσει καλά σε μικρά σετ δεδομένων καθώς επίσης έχει τη δυνατότητα να παρέχει μετρήσεις αβεβαιότητας στις προβλέψεις (Sit, 2019).

Το Random Forest Regression είναι ένας αλγόριθμος επιβλεπόμενης μηχανικής μάθησης, ο οποίος στηρίζεται στα δένδρα απόφασης (Decision Trees) και χρησιμοποιεί την τεχνική ensemble learning για παλινδρόμηση. Το ensemble learning αποτελεί ένα τομέα της μηχανικής μάθησης. Ο τομέας αυτός, στην περίπτωση της παλινδρόμησης, ασχολείται με το συνδυασμό προβλέψεων πολλαπλών μοντέλων παλινδρόμησης με σκοπό την αύξηση της απόδοσης και της ακρίβειας των προβλέψεων από εκείνες οι οποίες προέρχονται από ένα μόνο μοντέλο (Bakshi, 2020).

Τέλος, η τεχνική XGBoost, eXtreme Gradient Boosting, είναι μία αποτελεσματική ανοιχτού κώδικα εφαρμογή της τεχνικής Gradient Boosting. Η τεχνική Boosting λειτουργεί παρόμοια με την Random Forest. Η διαφορά τους έγκειται στο γεγονός ότι τα δέντρα δημιουργούνται διαδοχικά, συνεπώς, δεν είναι ασυσχέτιστα μεταξύ τους, όπως είναι στην Random Forest (Βαρελάς, 2019).

Στις ακόλουθες σελίδες παρουσιάζονται τα αποτελέσματα για τις προβλέψεις, αλλά και για τα σφάλματα τα οποία τις συνοδεύουν, για κάθε μοντέλο επιβλεπόμενης μηχανικής μάθησης, από τα προαναφερθέντα, για τα κρούσματα και για τους θανάτους Covid-19. Να τονισθεί ότι η ανάλυση για όλα τα μοντέλα μηχανικής μάθησης πραγματοποιήθηκε και για τις εννέα πόλεις. Τα αποτελέσματα τα οποία παρουσιάζονται και αναλύονται στις επόμενες σελίδες, για εξοικονόμηση χώρου, αφορούν ένα μέρος από τις εννέα αυτές πόλεις. Συγκεκριμένα, επιλέχθηκαν πέντε πόλεις οι οποίες φάνηκε να έχουν καλή απόδοση στα

μοντέλα πρόβλεψης τα οποία δημιουργήθηκαν. Πρόκειται για την Αθήνα, τη Μαδρίτη, τη Μόσχα, το Παρίσι και την Πράγα. Τα αποτελέσματα των υπόλοιπων τεσσάρων πόλεων, βρίσκονται στο παράρτημα.

4.4.1 Multiple Linear Regression

Το πρώτο μοντέλο το οποίο εφαρμόστηκε στα υπάρχοντα δεδομένα για τις εννέα διαφορετικές πόλεις, είναι το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης. Στο σημείο αυτό, παρατίθενται τα αποτελέσματα του αλγορίθμου για την πρόβλεψη των κρουσμάτων και των θανάτων της Covid-19. Επιπροσθέτως, παρατίθεται η αξιολόγηση της μεθόδου, η οποία προκύπτει κυρίως μέσω των τιμών των σφαλμάτων, αλλά και των γραφημάτων[3].

4.4.1.1 Πρόβλεψη Κρουσμάτων

Το μοντέλο πολλαπλής γραμμικής παλινδρόμησης το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των κρουσμάτων, είναι εκείνο για την πόλη του Παρισιού. Στις επόμενες σελίδες, ακολουθούν οι μετρικές αξιολόγησης του μοντέλου σε Πίνακα, αλλά και τα πειραματικά αποτελέσματα σε γραφήματα.

Μετρική Αξιολόγησης	**Παρίσι**
RMSE	1898.53 (cases per million)
R^2	0.902
EVS	0.906
MAE	1474.40 (cases per million)
MAPE	3.05%
MSE	3520145.14

Πίνακας 3 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LR, Παρίσι, Ιδία Επεξεργασία*

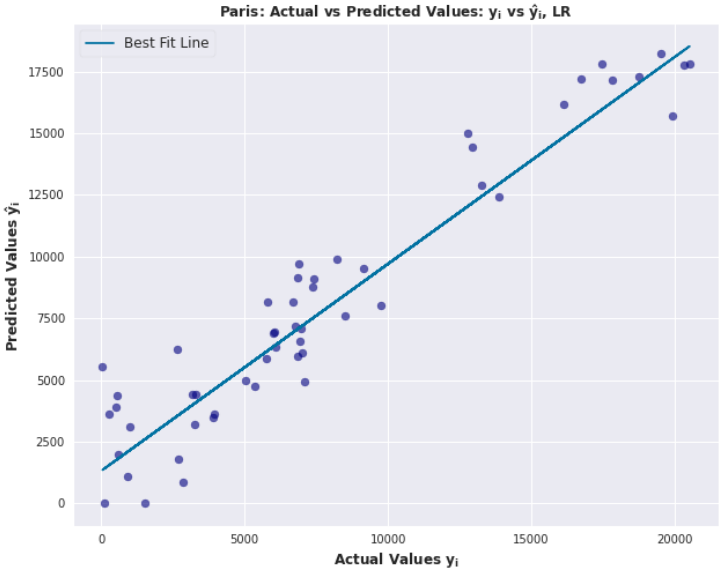
Βάσει όσων αναφέρθηκαν στην υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Παρισιού, εμφανίζουν ικανοποιητικές τιμές. Η R^2 και EVS έχουν υψηλές τιμές. Πρόκειται για τιμές οι οποίες βρίσκονται πολύ κοντά στο 0.90, πράγμα το οποίο υποδηλώνει πολύ υψηλή συσχέτιση. Στη συνέχεια, το MAE και το RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, βάσει του εύρους των τιμών των πραγματικών θανάτων (Βλέπε **Σχήμα 20**). Ακόμη το ποσοστό από το MAPE είναι χαμηλό και μικρότερο από 10%, άρα, λαμβάνοντας υπόψιν τη βιβλιογραφία (Allwright, 2021), θεωρείται πολύ καλό. Παρατηρείται, όμως, ότι το MSE εμφανίζει μία «μεγάλη» τιμή.

Όπως ήδη έχει αναφερθεί, δεν είναι εύκολη η ερμηνεία του συγκεκριμένου σφάλματος. Επίσης, χρειάζεται να ληφθεί υπόψιν, ότι το σφάλμα αυτό επηρεάζεται αρκετά από τυχόν ιδιάζοντα σημεία. Τέλος, όπως έχει αναφερθεί, δεν υπάρχει κάποια «σωστή» τιμή για το MSE. Ο κύριος σκοπός χρήσης του είναι η επιλογή μίας πρόβλεψης ενός μοντέλου, έναντι κάποιας άλλης.

Μία ορθή ερμηνεία μετρικών, απαιτεί την προσεκτική παρατήρηση του εύρους τιμών του μεγέθους το οποίο μελετάται. Στην περίπτωση αυτή, τα σφάλματα RMSE και MAE, εμφανίζουν φυσιολογικές τιμές, εάν κάποιος αναλογιστεί τις τιμές των κρουσμάτων ανά εκατομμύριο.

[3] Προτού ξεκινήσει η αξιολόγηση μοντέλου, αυτό το οποίο αξίζει να σημειωθεί είναι ότι για να γίνει η αξιολόγηση ενός μοντέλου πρόβλεψης με δεδομένα χρονοσειρών απαιτείται ιδιαίτερη προσοχή. Σίγουρα οι μετρικές αξιολόγησης βοηθούν στην αξιολόγηση, όμως δεν θα ήταν εντελώς σωστό κάποιος να επαναπαυτεί στις τιμές των μετρικών αυτών.

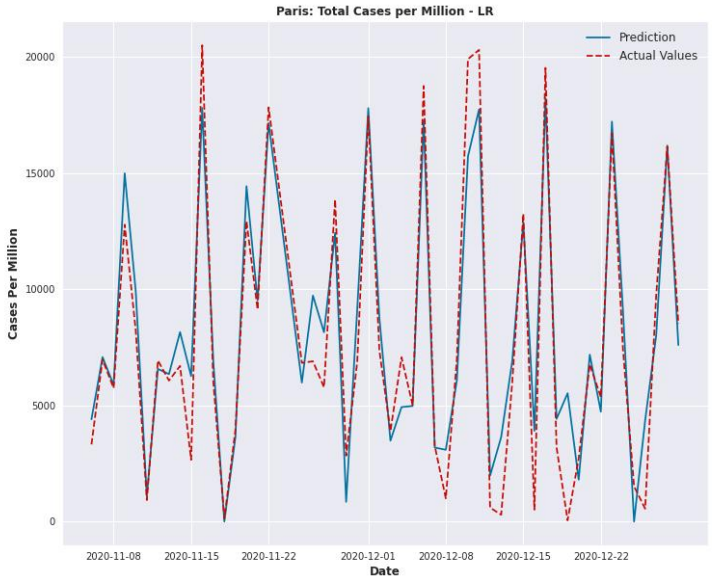
Για να είναι δυνατή, όμως, η άρτια ερμηνεία των παραπάνω μετρικών, χρειάζεται να παρατεθούν ορισμένα γραφήματα. Τα γραφήματα τα οποία επιλέγεται να παρουσιαστούν είναι το διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές, το διάγραμμα στο οποίο παρουσιάζονται ως προς το χρόνο, τόσο οι πραγματικές τιμές όσο και οι προβλεπόμενες και τέλος, το διάγραμμα των υπολοίπων ως προς train set και test set.



Σχήμα 19 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LR, Παρίσι, Ιδία Επεξεργασία*

Το διάγραμμα διασποράς αποτελεί έναν από τους πιο χρήσιμους τρόπους απεικόνισης της σχέσης η οποία επικρατεί ανάμεσα σε δύο, ποσοτικές, μεταβλητές. Οι τιμές της μίας μεταβλητής, στην περίπτωση αυτή της πραγματικής τιμής των κρουσμάτων ανά εκατομμύριο, εμφανίζονται στον άξονα x, ενώ οι τιμές της άλλης μεταβλητής, δηλαδή στην περίπτωση αυτή της προβλεπόμενης τιμής των κρουσμάτων ανά εκατομμύριο, εμφανίζονται στον άξονα y. Σε ένα διάγραμμα διασποράς μελετώνται κατά κύριο λόγο τέσσερις παράμετροι. Πρόκειται για τη μορφή, εάν είναι γραμμική ή όχι, για τη διεύθυνση, εάν είναι θετική ή αρνητική, για την ισχύ ανάμεσα στις δύο μεταβλητές, εάν είναι ισχυρή, ήπια ή αδύναμη. Τέλος, εξετάζεται εάν στο διάγραμμα εμφανίζονται ιδιάζοντα σημεία.

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη του Παρισιού, **Σχήμα 19**, παρατηρείται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή αφενός η γραμμικότητα, αφετέρου η υψηλή τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Πληροφορίες για την ισχύ ανάμεσα στις δύο μεταβλητές, παρέχονται από την κλίση της ευθείας των ελαχίστων τετραγώνων του διαγράμματος. Συγκεκριμένα, όσο η εν λόγω ευθεία βρίσκεται κοντά στις 45^o, τόσο περισσότερο δυνατή είναι η ισχύς. Συνεπώς, μικρότερες τιμές για την κλίση της ευθείας, συνεπάγονται και ηπιότερη ισχύ. Στην περίπτωση, όμως, του Παρισιού, θεωρείται ότι συναντάται μέτρια προς δυνατή ισχύς. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο ισχύει είναι ότι τα διαγράμματα διασποράς συχνά εμφανίζουν μία διάταξη, ένα μοτίβο. Ως ιδιάζον σημείο χαρακτηρίζεται εκείνο το οποίο δεν ταιριάζει στη διάταξη αυτή. Από το παραπάνω διάγραμμα εντοπίζονται πολύ λίγα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.



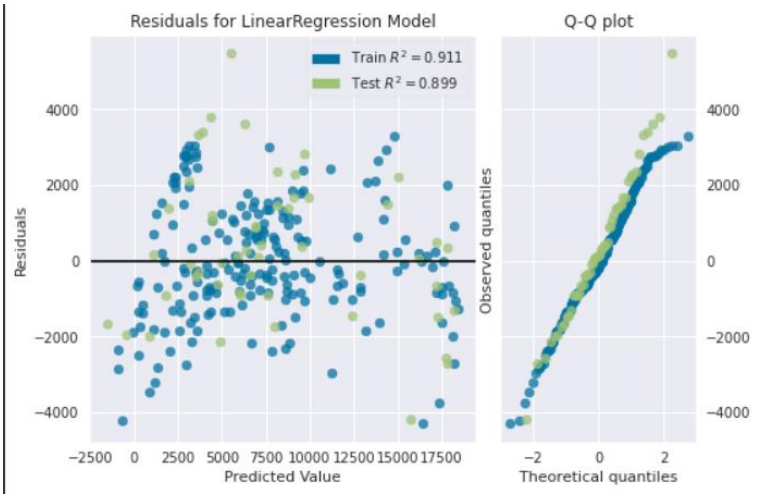
Σχήμα 20 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LR, Παρίσι, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων από το μοντέλο της γραμμικής παλινδρόμησης, αλλά κρίνεται σκόπιμο να απεικονισθούν και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο συγκεκριμένος αλγόριθμος έχει αποδώσει ικανοποιητικά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως για παράδειγμα γύρω στις 25/12/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα, γεγονός το οποίο μπορεί να επιβεβαιωθεί και από τις τιμές των μετρικών, **Πίνακας 8**.

Όπως έχει αναφερθεί, ως υπόλοιπο ορίζεται η διαφορά η οποία συναντάται ανάμεσα στην παρατηρούμενη τιμή και στην προβλεπόμενη τιμή μίας μεταβλητής, δηλαδή στο σφάλμα της πρόβλεψης.

Στο διάγραμμα των υπολοίπων (residuals) απεικονίζονται οι διαφορές ανάμεσα στις πραγματικές και στις προβλεπόμενες τιμές. Πρόκειται για ένα διάγραμμα διασποράς, το οποίο απεικονίζει τα υπόλοιπα στον κατακόρυφο άξονα και την εξαρτημένη μεταβλητή στον οριζόντιο άξονα. Με αυτόν τον τρόπο καθίσταται εύκολος ο εντοπισμός περιοχών οι οποίες έχουν τάση για λιγότερα ή περισσότερα σφάλματα (The scikit-yb developers, 2022). Ένα διάγραμμα υπολοίπων βοηθάει στο να καθοριστεί εάν ένα γραμμικό μοντέλο είναι κατάλληλο για την μοντελοποίηση των εκάστοτε χρησιμοποιούμενων δεδομένων. Ακόμη, ένα γράφημα υπολοίπων φανερώνει πώς τα δεδομένα παρεκκλίνουν από το μοντέλο. Σε περιπτώσεις στις οποίες τα υπόλοιπα είναι διασκορπισμένα με τυχαίο τρόπο γύρω από τη γραμμή y=0, αυτό σημαίνει ότι το γραμμικό μοντέλο προσεγγίζει τα δεδομένα καλά. Συνεπώς, μπορεί να χρησιμοποιηθεί το γραμμικό μοντέλο σε αυτό το πρόβλημα. Διαφορετικά, σε περιπτώσεις κατά τις οποίες τα υπόλοιπα εμφανίζουν ένα κυρτό μοτίβο, τότε αυτό υποδηλώνει ότι το γραμμικό μοντέλο αποτυπώνει μία τάση ορισμένων δεδομένων καλύτερα από κάποια άλλα. Ως εκ τούτου, συμπεραίνεται ότι είναι προτιμότερο να μην χρησιμοποιηθεί γραμμικό μοντέλο, αλλά ένα μη γραμμικό (Jiwon Park, 2021).

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 21 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LR, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα, **Σχήμα 21**, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Συνεπώς, ένα γραμμικό μοντέλο δύναται να περιγράψει το πρόβλημα αυτό. Επιπλέον, έχουν τυπωθεί και οι τιμές για το R^2, τόσο για το train set όσο και για το test set. Για το test set είχε γίνει ο υπολογισμός του R^2 μέσω της Scikit-learn. Να σχολιαστεί επίσης, ότι η τιμή του R^2 για το train set είναι αρκετά ικανοποιητική και λίγο υψηλότερη από εκείνη του test set.

Εκτός από το διάγραμμα των υπολοίπων, παρουσιάζεται και το διάγραμμα Q-Q. Πρόκειται για ένα διάγραμμα στον οριζόντιο άξονα του οποίου απεικονίζονται θεωρητικά ποσοστά, τα οποία είναι γνωστά ως τυπική κανονική μεταβλητή, μία μεταβλητή με μέσο μ=0 και τυπική απόκλιση σ=1. Στον κατακόρυφό του άξονα απεικονίζονται με τη σειρά οι τιμές της τυχαίας μεταβλητής της οποίας αναζητείται η κατανομή.

Στο διάγραμμα αυτό καθορίζεται εάν δύο σετ δεδομένων έρχονται από πληθυσμούς με κοινή κατανομή. Το διάγραμμα δημιουργείται μέσω της απεικόνισης των οντοτήτων του ενός σετ δεδομένων, επάνω σε εκείνες του άλλου σετ. Εάν και οι δύο οντότητες προέρχονται από την ίδια κατανομή, τότε τα σημεία σχηματίζουν μία γραμμή η οποία είναι σχεδόν ευθεία (Ford, 2015). Ακόμη, από την παρατήρηση μίας γραφικής παράστασης Q-Q μπορούν να αντληθούν επιπλέον πληροφορίες, οι οποίες αφορούν την ασυμμετρία μίας κατανομής και την κύρτωση της.

Από το γράφημα Q-Q, **Σχήμα 21**, φαίνεται ότι τα σημεία εμφανίζουν μία σχεδόν ευθεία γραμμή και διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Στη συνέχεια, ακολουθεί η ανάλυση για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο καλή, συγκριτικά με εκείνη της πόλης του Παρισιού.

Όπως και στην περίπτωση του Παρισιού, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

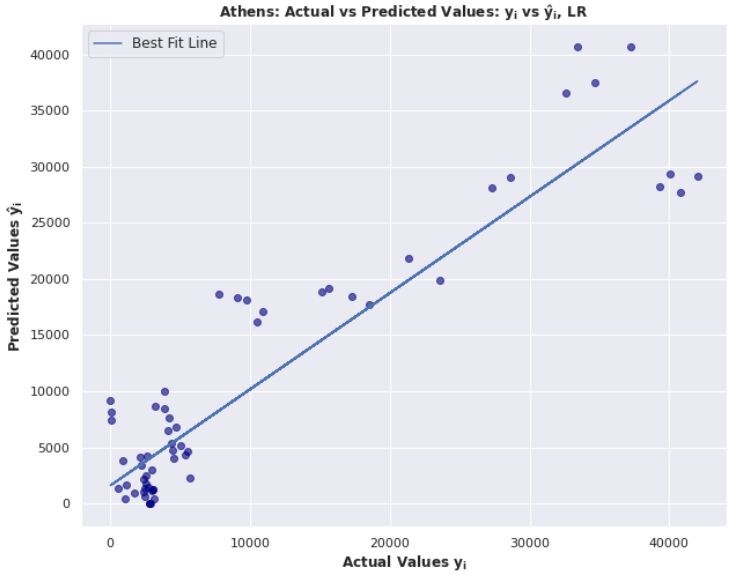
Μετρική Αξιολόγησης	**Αθήνα**
RMSE	5030.65 (cases per million)
R^2	0.845
EVS	0.845
MAE	3637.27 (cases per million)

MAPE	17.74%
MSE	25307459.09

Πίνακας 4 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LR, Αθήνα, Ιδία Επεξεργασία*

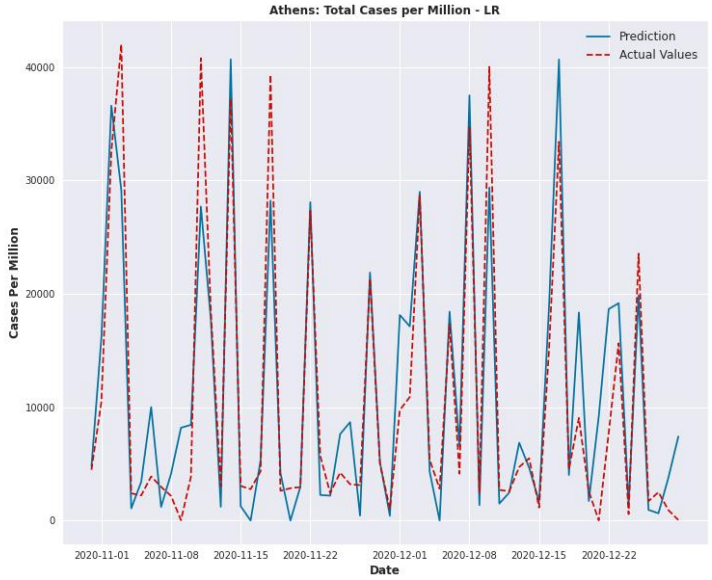
Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν σχετικά υψηλές τιμές. Έχουν τιμές κοντά στο 0.85, και ως εκ τούτου θεωρείται ότι έχουν ισχυρό μέγεθος συσχέτισης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 23**). Ακόμη, το ποσοστό από το MAPE είναι σχετικά χαμηλό, βρίσκεται κοντά στο 20%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται καλό.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Αθήνας.



Σχήμα 22 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LR, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο, για την πόλη της Αθήνας, **Σχήμα 22**, παρατηρείται ύπαρξη γραμμικότητας. Παρατηρείται, δηλαδή, ότι τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη». Όλα αυτά γίνονται αντιληπτά και μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, τα οποία απέχουν αρκετά από την ευθεία ελαχίστων τετραγώνων καθώς και από συγκεντρώσεις άλλων σημείων.



Σχήμα 23 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LR, Αθήνα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και απεικονίζονται και οι πραγματικές τιμές των κρουσμάτων, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για τον αλγόριθμο γραμμικής παλινδρόμησης, για την πόλη της Αθήνας, είναι ότι εντοπίζει την τάση των πραγματικών τιμών, εν τούτοις αδυνατεί να προβλέψει την ακριβή τιμή τόσο των μεγίστων, όσο και των ελαχίστων. Ακόμη, υπάρχουν σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως, όπως είναι για παράδειγμα η περίοδος από 2/11/2020 έως 8/11/2020. Τα παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, **Πίνακας 9**. Παρατηρείται ότι οι τιμές των κρουσμάτων φτάνουν στις 30000 και 40000, ανά εκατομμύριο, συνεπώς, τα σφάλματα της τάξεως των 3500 και 5000 χιλιάδων, θεωρούνται αποδεκτά.

Στο **Σχήμα 24** παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Σχήμα 24 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LR, Αθήνα, Ιδία Επεξεργασία*

Από το **Σχήμα 24** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, φαίνεται ότι δημιουργείται ένα ευθύγραμμο μοτίβο το οποίο «τέμνει» την ευθεία y=0, δεν είναι παράλληλο σε εκείνη. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο **Σχήμα 22** αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι υπάρχουν ορισμένα

ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο σχεδόν επικαλυπτόμενες ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Στη συνέχεια, παρατίθενται τα αποτελέσματα για την πόλη της Μαδρίτης.

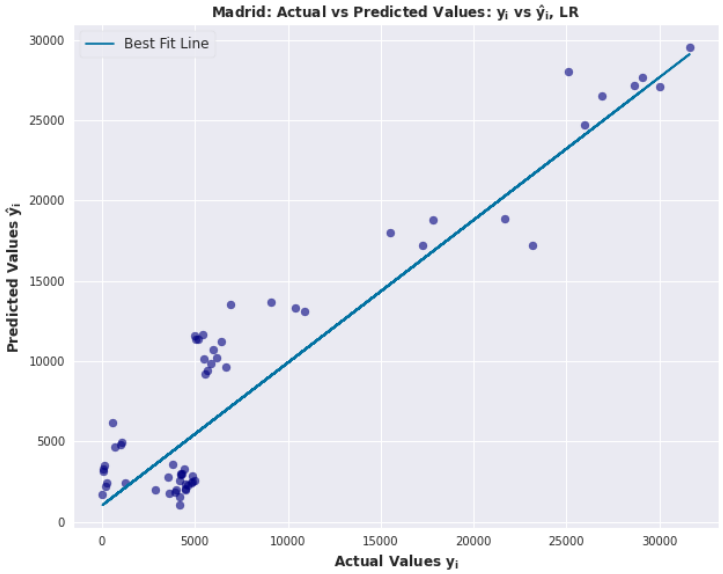
Ομοίως, παρουσιάζεται αρχικά ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	**Μαδρίτη**
RMSE	3308.69 (cases per million)
R^2	0.856
EVS	0.868
MAE	2859.63 (cases per million)
MAPE	4.33%
MSE	10947427.27

Πίνακας 5 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LR, Μαδρίτη, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν τη Μαδρίτη, κυμαίνονται σε ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν υψηλές τιμές. Έχουν τιμή μεγαλύτερη από 0.80, και ως εκ τούτου έχουν ισχυρό μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 26**). Ακόμη, το ποσοστό από το MAPE είναι σχετικά χαμηλό, βρίσκεται κοντά στο 4%, άρα, θεωρείται αρκετά καλό. Άλλωστε, όπως έχει αναφερθεί, όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Να σημειωθεί ότι το MSE εμφανίζει σημαντικά μεγάλη τιμή, όπως και το MSE για την πόλη της Αθήνας. Πιθανότατα αυτό να συμβαίνει λόγω ύπαρξης ιδιαζόντων σημείων. Όπως ήδη αναφέρθηκε και στην περίπτωση της πόλης του Παρισιού, πρόκειται για μία μετρική η οποία δεν είναι αρκετά εύκολο να ερμηνευτεί.

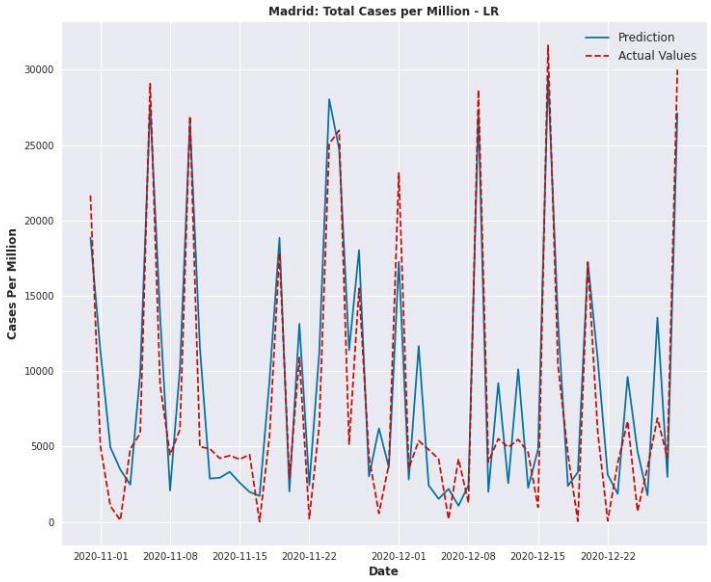
Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.



Σχήμα 25 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LR, Μαδρίτη, Ιδία Επεξεργασία*

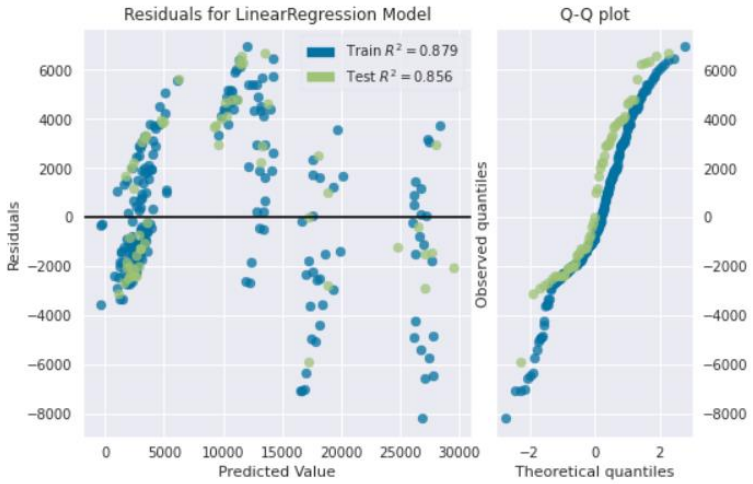
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, παρατηρείται ύπαρξη ήπιας γραμμικότητας. Παρατηρείται ότι, όπως και στην περίπτωση της Αθήνας, τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και φαίνεται να δημιουργούνται ορισμένες χαρακτηριστικές συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται λίγα ιδιάζοντα σημεία, τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία. Σε κάθε περίπτωση, χρειάζεται να ληφθεί υπόψιν ότι τα σημεία του διαγράμματος δεν έχουν ένα συγκεκριμένο, ξεκάθαρο, γραμμικό μοτίβο.

Στο διάγραμμα το οποίο ακολουθεί, **Σχήμα 26**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι καταγεγραμμένες τιμές των κρουσμάτων, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Μαδρίτης και για τον αλγόριθμο γραμμικής παλινδρόμησης, είναι ότι ο αλγόριθμος εντοπίζει σχετικά καλά την τάση των πραγματικών τιμών. Αδυνατεί να προβλέψει επιτυχώς την ακριβή τιμή τόσο για αρκετά μέγιστα, όσο και για αρκετά ελάχιστα. Ακόμη, υπάρχουν σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως, όπως το διάστημα ανάμεσα στις 8/12/2020 και στις 15/12/2020. Τα παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, όπως λ.χ. το MAE, **Πίνακας 10**. Εδώ, οι τιμές των MAE και RMSE είναι μικρότερες από εκείνες της Αθήνας, αλλά και τα συνολικά κρούσματα εμφανίζουν μέγιστες τιμές γύρω στο 25000 με 30000.



Σχήμα 26 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LR, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 27 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LR, Μαδρίτη, Ιδία Επεξεργασία*

Από το **Σχήμα 27** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα, αλλά όχι πλήρως «γραμμικά» γύρω από την ευθεία y=0. Επίσης, φαίνεται ότι δημιουργείται ένα ευθύγραμμο μοτίβο το οποίο «τέμνει» την ευθεία y=0, δεν είναι παράλληλο σε εκείνη. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο **Σχήμα 25** και λαμβάνοντας υπόψιν την υψηλή τιμή του MSE, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι πράγματι υπάρχουν ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες δεν επικαλύπτονται πλήρως. Οι εν λόγω γραμμές, όμως, ακολουθούν ίδια τάση. Δηλαδή, η γραμμή για τα δεδομένα ελέγχου φαίνεται να ακολουθεί την τάση της γραμμής των δεδομένων εκπαίδευσης, αλλά να βρίσκεται μερικές μονάδες υψηλότερα. Θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Ακολουθεί η ανάλυση για το μοντέλο πρόβλεψης κρουσμάτων για την πόλη της Μόσχας.

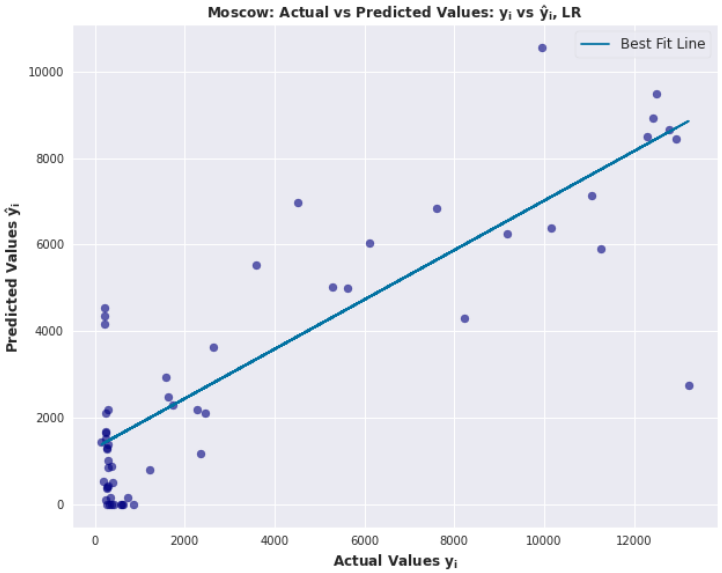
Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	**Μόσχα**
RMSE	2535.89 (cases per million)
R^2	0.686
EVS	0.693
MAE	1752.48 (cases per million)
MAPE	2.63%
MSE	6210695.47

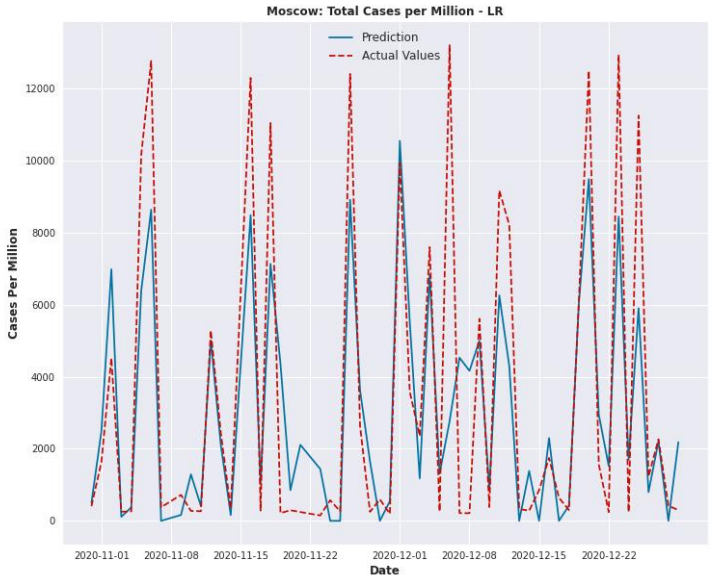
Πίνακας 6 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LR, Μόσχα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε σχετικά ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν μεσαία τιμή. Έχουν τιμή μεγαλύτερη από 0.60, και ως εκ τούτου έχουν μέτριο μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως το MAE και RMSE, εμφανίζουν μικρές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 29**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό, βρίσκεται κοντά στο 3%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο.

Από το διάγραμμα διασποράς το οποίο ακολουθεί, για την πόλη της Μόσχας, **Σχήμα 28**, ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μόσχας, δεν παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας. Παρατηρείται ότι, όπως και στην περίπτωση των προαναφερθέντων πόλεων, εκτός από το Παρίσι, τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη». Στην περίπτωση αυτή, δημιουργείται μία έντονη συστάδα για ένα μικρό εύρος τιμών. Αυτό το οποίο γίνεται αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων είναι ότι τα σημεία δεν βρίσκονται κοντά της γραμμικά. Με αυτόν τον τρόπο μπορεί να δικαιολογηθεί η σχετικά χαμηλή τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι αρκετά ήπια. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται μερικά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 28 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LR, Μόσχα, Ιδία Επεξεργασία*



Σχήμα 29 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LR, Μόσχα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών σε αρκετά σημεία. Εν τούτοις, σε κάποια άλλα φαίνεται να την «υπερεκτιμάει» όπως, για παράδειγμα, στα τέλη του Νοέμβρη (22/11/2020). Ακόμη, αδυνατεί να προβλέψει πλήρως την ακριβή τιμή τόσο για τα μέγιστα, όσο και για τα ελάχιστα. Έτσι, υπάρχουν σημεία τα οποία δεν έχουν προσαρμοστεί πλήρως. Τα παραπάνω, μπορούν να εντοπισθούν και να επιβεβαιωθούν και από τις τιμές των μετρικών, όπως λ.χ. το MAE και το RMSE, **Πίνακας 11**.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Σχήμα 30 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LR, Μόσχα, Ιδία Επεξεργασία*

Από το **Σχήμα 30** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, φαίνεται ότι δημιουργείται ένα ευθύγραμμο μοτίβο το οποίο «τέμνει» την ευθεία y=0, το οποίο, δηλαδή, δεν είναι παράλληλο σε εκείνη. Τα περισσότερα σημεία φαίνεται να βρίσκονται κυρίως σε αυτό το ευθύγραμμο τμήμα. Συνεπώς, παρατηρώντας ταυτόχρονα το **Σχήμα 28** και το **Σχήμα 30**, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι παρουσιάζονται ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των υπόλοιπων σημείων. Ενδιαφέρον παρουσιάζει το γεγονός ότι η απόδοση R^2 για τα δεδομένα εκπαίδευσης δεν είναι τόσο καλή όσο η απόδοση R^2 των δεδομένων ελέγχου.

Στη βιβλιογραφία συναντώνται τρεις κύριοι λόγοι για τους οποίους η ακρίβεια των δεδομένων εκπαίδευσης είναι μεγαλύτερη από εκείνη των δεδομένων ελέγχου. Πρωτίστως, υποβόσκει κάποια διαφορετική κατανομή ανάμεσα στα δεδομένα. Χρειάζεται να ληφθεί υπόψιν ότι στην τρέχουσα διπλωματική χρησιμοποιούνται δεδομένα χρονοσειρών. Δευτερευόντως, μπορεί να συμβεί σε περιπτώσεις στις οποίες συναντάται ένα μικρό σετ δεδομένων ελέγχου και το μοντέλο φαίνεται να έχει καλή απόδοση. Κατά τρίτον, είναι πιθανό να συνέβη ανεξήγητη, υψηλού βαθμού κανονικοποίηση (Krishna, 2020).

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία των δεδομένων εκπαίδευσης και τα σημεία των δεδομένων ελέγχου εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται σε ένα τμήμα τους. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Πέμπτη και τελευταία πόλη ανάλυσης για τα μοντέλα πρόβλεψης κρουσμάτων με γραμμική παλινδρόμηση, είναι η Πράγα.

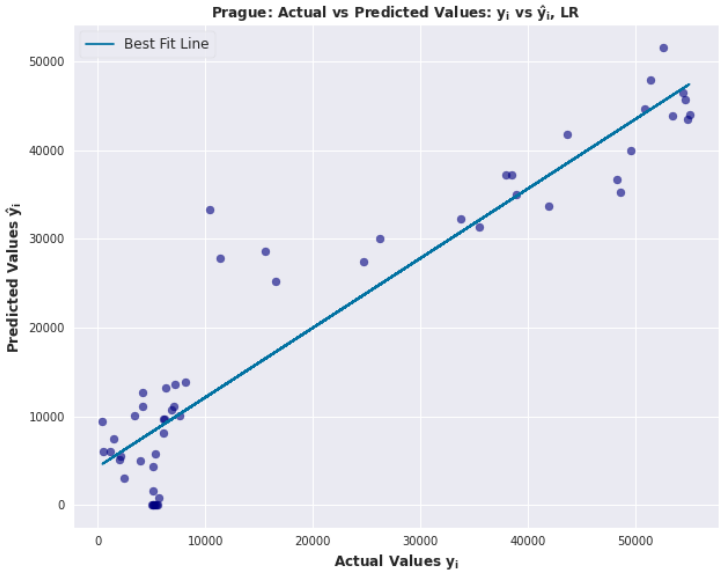
Όπως και στις προηγούμενες τέσσερις αναλύσεις, παρουσιάζεται αρχικά ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	**Πράγα**
RMSE	7586.98 (cases per million)
R^2	0.866
EVS	0.866
MAE	212.17 (cases per million)
MAPE	1.37%
MSE	52443580.36

Πίνακας 7 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LR, Πράγα, Ιδία Επεξεργασία*

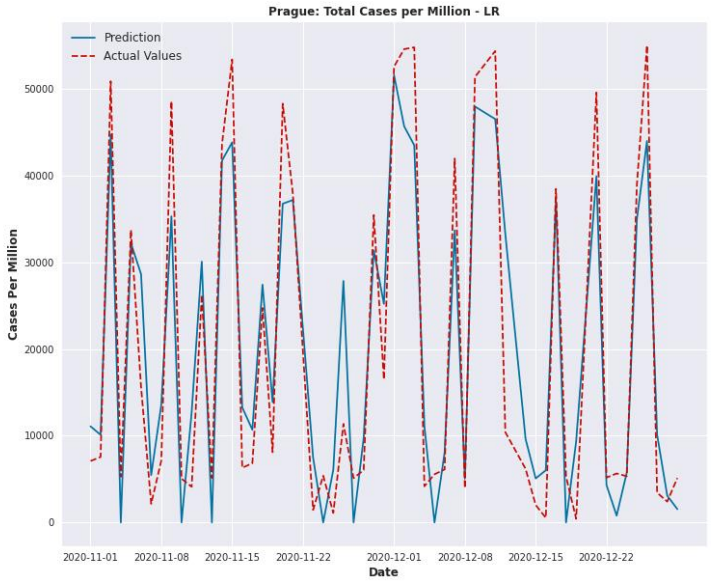
Οι μετρικές αξιολόγησης της Πράγας, κυμαίνονται σε ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν σχετικά υψηλή τιμή. Έχουν τιμή μεγαλύτερη από 0.80, και ως εκ τούτου έχουν σχετικά υψηλό μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές, λαμβάνοντας υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Πράγας (Βλέπε **Σχήμα 32**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό, βρίσκεται κοντά στο 2%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Εκείνη η μετρική η οποία φαίνεται να εμφανίζει σημαντικά μεγαλύτερη τιμή, συγκριτικά με τις τιμές των υπόλοιπων προηγηθέντων τεσσάρων μετρικών, είναι το MSE. Όπως ήδη αιτιολογήθηκε και στην περίπτωση της Μαδρίτης, η υψηλή τιμή στο MSE πιθανότατα να οφείλεται σε ύπαρξη ιδιαζόντων σημείων. Επιπροσθέτως, πιθανότατα υποδηλώνει στατιστικά ότι υπάρχει σημαντική μεροληψία στην εκτίμηση και στην διασπορά της εκτιμήτριας.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Πράγας.



Σχήμα 31 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LR, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Πράγας, παρατηρείται μία σχετική ύπαρξη γραμμικότητας. Παρατηρείται, επίσης, ότι τα σημεία του διαγράμματος σχηματίζουν μία ήπια ευθεία γραμμή σε ορισμένα σημεία, κοντά στην ευθεία των ελαχίστων τετραγώνων. Στην περίπτωση αυτή, φαίνεται να δημιουργούνται δύο χαρακτηριστικές συστάδες. Μία για μικρότερα μεγέθη και μία για μεγαλύτερα, η οποία χαρακτηρίζεται και από μία πιο ευδιάκριτη γραμμικότητα σημείων. Σίγουρα, η απεικόνιση αυτή δεν μπορεί να συγκριθεί με εκείνη για την πόλη του Παρισιού, **Σχήμα 19**. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται ορισμένα ιδιάζοντα σημεία, τα οποία βρίσκονται μακριά από τις συγκεντρώσεις των υπόλοιπων σημείων και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 32 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LR, Πράγα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών σε αρκετά σημεία. Εν τούτοις, σε κάποια άλλα φαίνεται να την «υπερεκτιμάει» όπως, για παράδειγμα, στα τέλη του Νοέμβρη (24/11/2020), γεγονός το οποίο φανερώνει ότι δεν μπορεί να κάνει πολύ καλή προσαρμογή στο μοντέλο. Ακόμη, αδυνατεί να προβλέψει πλήρως την ακριβή τιμή τόσο για τα μέγιστα, όσο και για τα ελάχιστα, στην πλειονότητα των περιπτώσεων. Έτσι, παρατηρούνται σημεία τα οποία δεν έχουν προσαρμοστεί πλήρως στην τάση και στις τιμές των πραγματικών δεδομένων. Τα παραπάνω, επιβεβαιώνονται και από τις τιμές των μετρικών, όπως λ.χ. το MAE και το RMSE, **Πίνακας 7**. Να σημειωθεί, όμως, ότι η αντιστοιχία κρουσμάτων και μετρικών θα μπορούσε να θεωρηθεί ότι εμφανίζει μία «αρμονία», καθώς για μέγιστα 50000 κρουσμάτων εντοπίζονται σφάλματα της τάξεων των 7500 κρουσμάτων.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Έπειτα, παρουσιάζεται το διάγραμμα Q-Q.

Σχήμα 33 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LR, Πράγα, Ιδία Επεξεργασία*

Από το **Σχήμα 33** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όπως και στις προηγούμενες περιπτώσεις, φαίνεται ότι δημιουργείται ένα ευθύγραμμο μοτίβο το οποίο «τέμνει» την ευθεία y=0, δηλαδή, δεν είναι παράλληλο σε εκείνη. Τα περισσότερα σημεία φαίνεται να βρίσκονται κυρίως σε αυτό το ευθύγραμμο τμήμα. Επίσης, αρκετά σημεία φαίνεται να είναι συγκεντρωμένα στις αρνητικές τιμές, κάτω από την ευθεία y=0. Συνεπώς, παρατηρώντας ταυτόχρονα το **Σχήμα 31** και το **Σχήμα 33**, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι παρουσιάζονται αρκετά ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από την ευθεία y=0 και τη συγκέντρωση των υπόλοιπων σημείων γύρω από αυτή. Όπως γίνεται αντιληπτό, και στην περίπτωση της πόλης της Πράγας για την πρόβλεψη των κρουσμάτων, η απόδοση του R^2 των δεδομένων ελέγχου είναι καλύτερη από εκείνη των δεδομένων εκπαίδευσης.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία από τα δεδομένα εκπαίδευσης και από τα δεδομένα ελέγχου εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται σε ένα μεγάλο μέρος τους. Συνεπώς, τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τα αποτελέσματα των μοντέλων για την πρόβλεψη κρουσμάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Α.

4.4.1.2 *Πρόβλεψη Θανάτων*

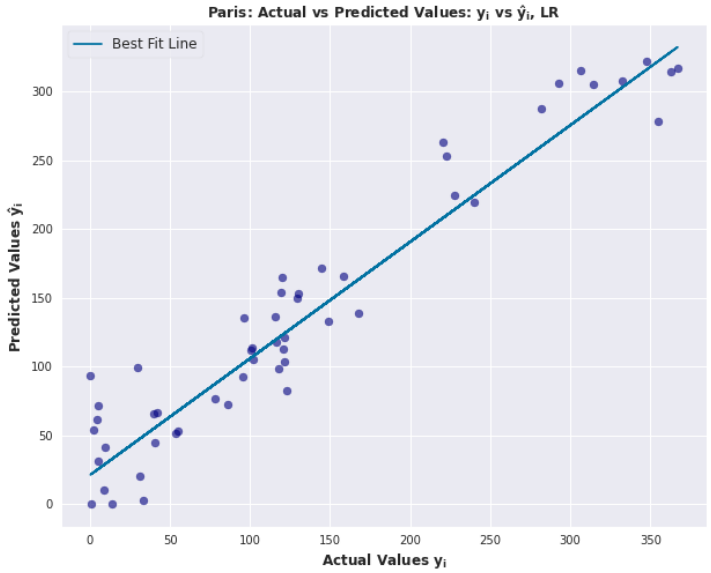
Το μοντέλο πολλαπλής γραμμικής παλινδρόμησης το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των θανάτων, είναι εκείνο για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	**Παρίσι**
RMSE	32.77 (deaths per million)
R^2	0.912
EVS	0.916
MAE	25.13 (deaths per million)
MAPE	6.30%
MSE	1045.64

Πίνακας 8 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LR, Παρίσι, Ιδία Επεξεργασία*

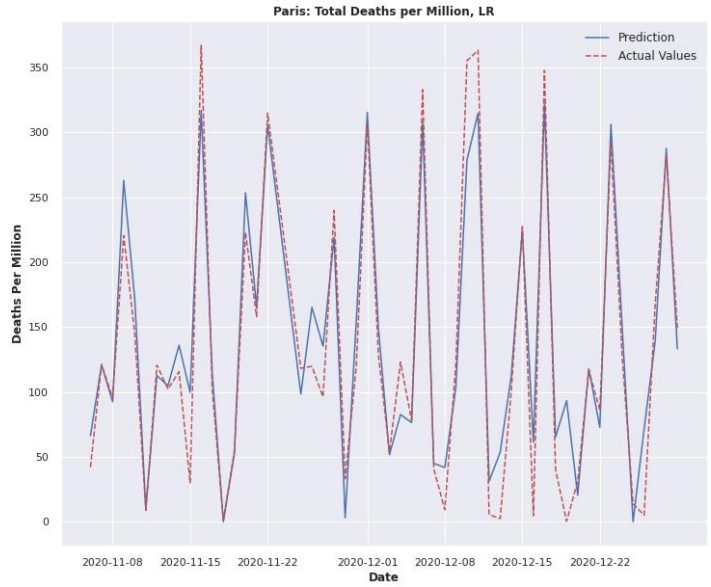
Βάσει όσων αναφέρθηκαν στην υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Παρισιού, είναι ικανοποιητικές. Η τιμή της R^2 και της EVS

ξεπερνάει το 0.90, άρα η συσχέτιση είναι πολύ υψηλή. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 35**). Ακόμη το ποσοστό από το MAPE είναι χαμηλό και μικρότερο από 10%, άρα θεωρείται πολύ καλό.



Σχήμα 34 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LR, Παρίσι, Ιδία Επεξεργασία*

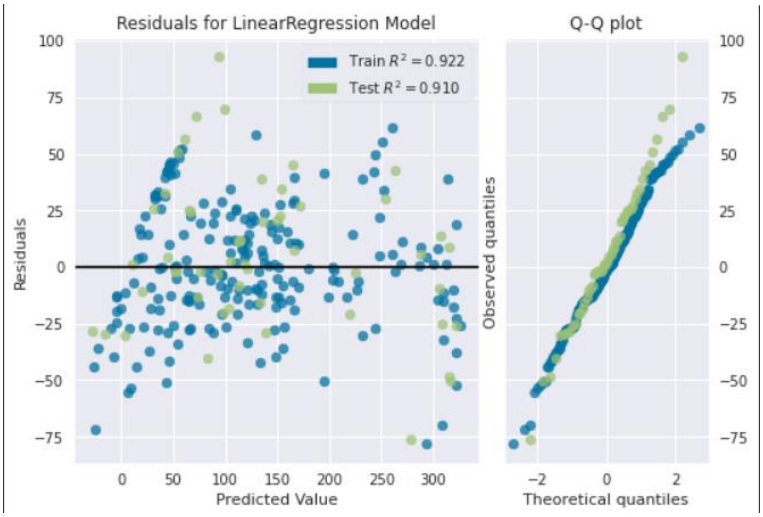
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη του Παρισιού, **Σχήμα 34**, παρατηρείται ότι υπάρχει γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή αφενός η γραμμικότητα, αφετέρου η υψηλή τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Στην περίπτωση του Παρισιού, θεωρείται ότι συναντάται μέτρια ισχύς, λόγω της κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται πολύ λίγα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.



Σχήμα 35 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LR, Παρίσι, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα, **Σχήμα 35**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων, αλλά κρίνεται σκόπιμο να απεικονισθούν και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος της γραμμικής παλινδρόμησης έχει αποδώσει ικανοποιητικά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα, γεγονός το οποίο μπορεί να επιβεβαιωθεί και από τις μετρικές, **Πίνακας 8**.

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 36 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LR, Παρίσι, Ιδία Επεξεργασία*

Από το **Σχήμα 36** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Συνεπώς, ένα γραμμικό μοντέλο δύναται να περιγράψει το πρόβλημα αυτό. Επιπλέον, έχουν τυπωθεί και οι τιμές για το R^2, τόσο για το train set όσο και για το test set. Για το test set είχε γίνει ο υπολογισμός του R^2 μέσω της Scikit-learn. Να σχολιαστεί επίσης, ότι η τιμή του R^2 για το train set είναι αρκετά ικανοποιητική και λίγο υψηλότερη από εκείνη του test set.

Εκτός από το διάγραμμα των υπολοίπων, παρουσιάζεται και το διάγραμμα Q-Q.

Από το διάγραμμα Q-Q, **Σχήμα 36**, φαίνεται ότι τα σημεία εμφανίζουν μία σχεδόν ευθεία γραμμή και διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Στη συνέχεια, ακολουθεί η ανάλυση για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο καλή, συγκριτικά με την πόλη του Παρισιού.

Όπως και στην περίπτωση του Παρισιού, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

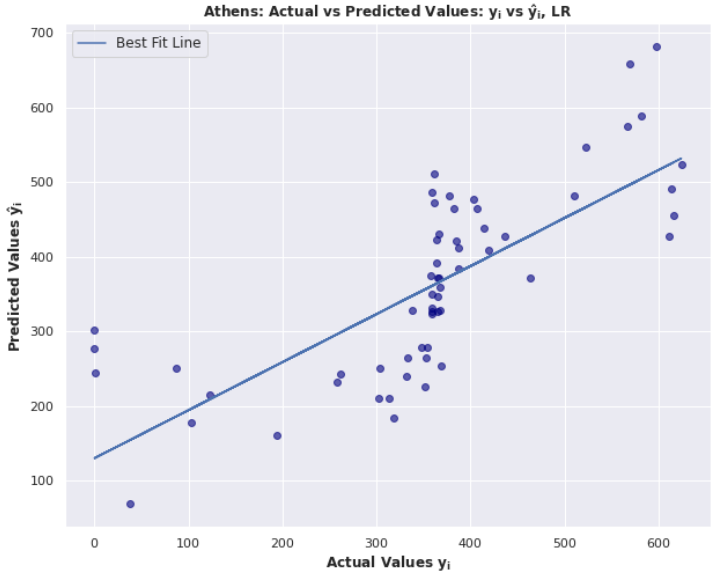
Μετρική Αξιολόγησης	**Αθήνα**
RMSE	98.11 (deaths per million)
R^2	0.550
EVS	0.552
MAE	73.35 (deaths per million)

MAPE	51.78%
MSE	9625.68

Πίνακας 9 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LR, Αθήνα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS δεν εμφανίζουν αρκετά υψηλές τιμές. Βρίσκονται στη μέση, με τιμές κοντά στο 0.50, και ως εκ τούτου χαρακτηρίζονται «μέτριες» (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων (Βλέπε **Σχήμα 38**). Ακόμη, το ποσοστό από το MAPE είναι σχετικά υψηλό, βρίσκεται κοντά στο 50%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται μέτριο.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Αθήνας.

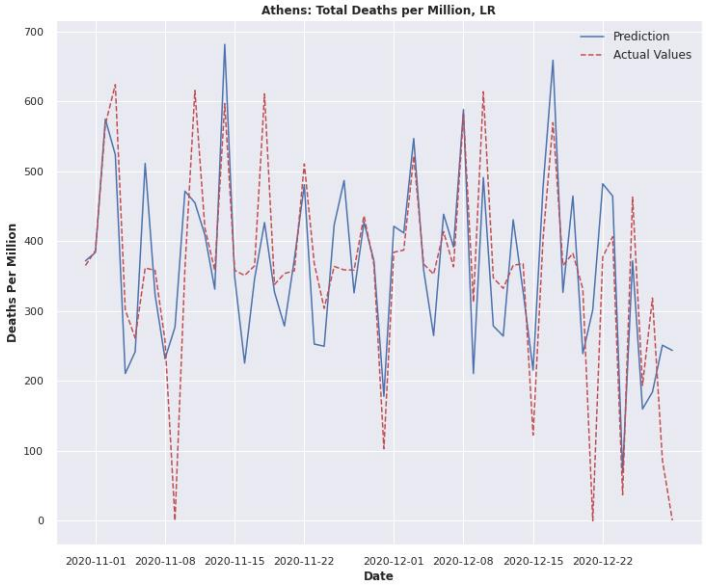


Σχήμα 37 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LR, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Αθήνας, **Σχήμα 37**, δεν παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας. Παρατηρείται, δηλαδή, ότι τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί τρεις συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Ακόμη, με αυτόν τον τρόπο μπορεί να δικαιολογηθεί η μέτρια τιμή του R^2, η οποία συναντάται στο μοντέλο αυτό. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι σχετικά ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, λαμβάνοντας, όμως, υπόψιν ότι τα σημεία του διαγράμματος δεν έχουν ένα συγκεκριμένο γραμμικό μοτίβο.

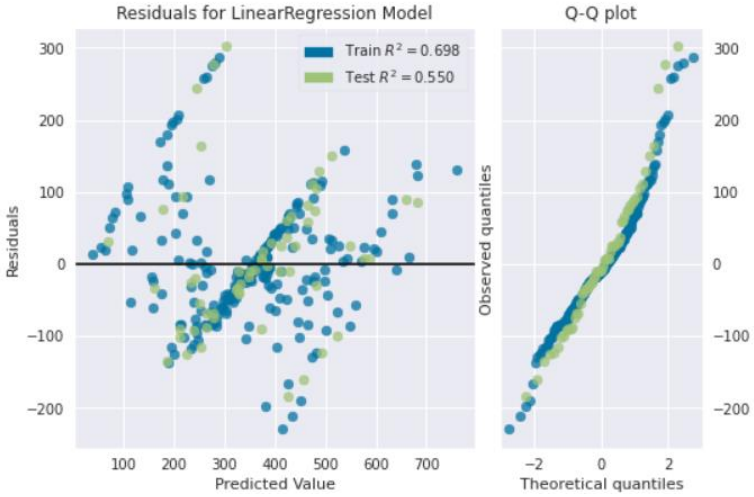
Στο παρακάτω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων, αλλά κρίνεται σκόπιμο να απεικονισθούν και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για τον αλγόριθμο γραμμικής παλινδρόμησης, για την πόλη της Αθήνας, είναι ότι εντοπίζει την τάση των πραγματικών τιμών, εν τούτοις αδυνατεί να προβλέψει πλήρως τόσο τα μέγιστα, όσο και τα

ελάχιστα. Ακόμη, υπάρχουν σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως. Τα παραπάνω, επιβεβαιώνονται από τις μετρικές, **Πίνακας 9**.



Σχήμα 38 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LR, Αθήνα, Ιδία Επεξεργασία*

Στο **Σχήμα 39** παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 39 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LR, Αθήνα, Ιδία Επεξεργασία*

Από το **Σχήμα 39** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, φαίνεται ότι δημιουργείται ένα ευθύγραμμο μοτίβο το οποίο «τέμνει» την ευθεία y=0, δεν είναι παράλληλο σε εκείνη. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο **Σχήμα 37** αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο επικαλυπτόμενες ευθείες γραμμές, άρα ταυτίζονται, και διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

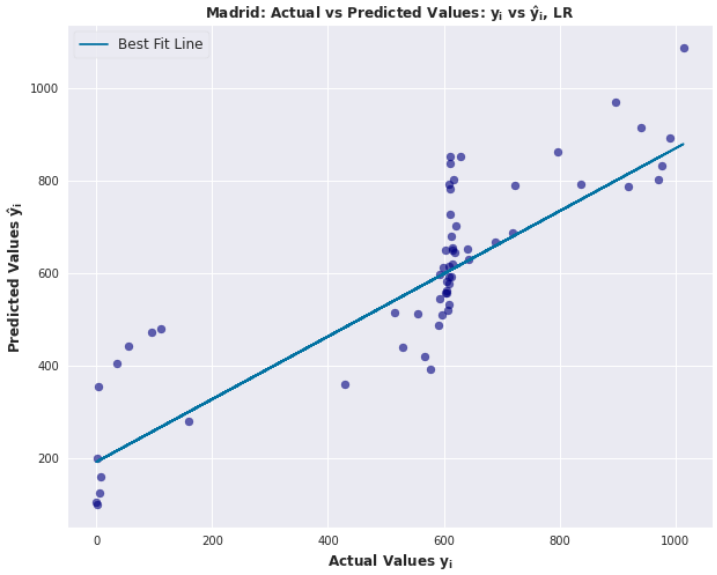
Μετά την Αθήνα, σειρά για ανάλυση έχει το μοντέλο πρόβλεψης κρουσμάτων για την πόλη της Μαδρίτης. Ομοίως, παρουσιάζεται αρχικά ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	Μαδρίτη
RMSE	147.51 (deaths per million)
R^2	0.703
EVS	0.733
MAE	107.50 (deaths per million)
MAPE	20.33%
MSE	21759.83

Πίνακας 10 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LR, Μαδρίτη, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν τη Μαδρίτη, κυμαίνονται σε πιο ικανοποιητικά και αποδεκτά πλαίσια, συγκριτικά με την Αθήνα. Η R^2 και EVS εμφανίζουν υψηλές τιμές. Έχουν τιμή μεγαλύτερη από 0.70, και ως εκ τούτου έχουν ισχυρό μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 41**). Ακόμη, το ποσοστό από το MAPE είναι σχετικά χαμηλό, βρίσκεται κοντά στο 20%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται σχετικά καλό. Άλλωστε, όπως έχει αναφερθεί, όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Να σημειωθεί ότι το MSE εμφανίζει σημαντικά μεγαλύτερη τιμή συγκριτικά με εκείνες των προηγούμενων πόλεων. Πιθανότατα αυτό συμβαίνει λόγω ύπαρξης ιδιαζόντων σημείων.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.



Σχήμα 40 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LR, Μαδρίτη, Ιδία Επεξεργασία*

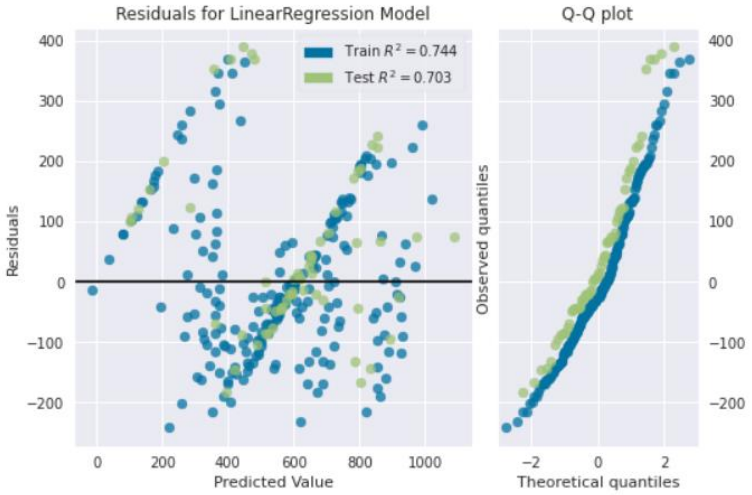
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, **Σχήμα 40**, δεν παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας. Παρατηρείται ότι, όπως και στην περίπτωση της Αθήνας, τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί τρεις χαρακτηριστικές συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων

τετραγώνων. Ακόμη, με αυτόν τον τρόπο μπορεί να δικαιολογηθεί η σχετικά καλή τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία. Σε κάθε περίπτωση, χρειάζεται να ληφθεί υπόψιν ότι τα σημεία του διαγράμματος δεν έχουν ένα συγκεκριμένο γραμμικό μοτίβο.

Στο διάγραμμα το οποίο ακολουθεί, **Σχήμα 41**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Μαδρίτης και για τον αλγόριθμο γραμμικής παλινδρόμησης, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών κατά ένα μεγάλο ποσοστό, εν τούτοις αδυνατεί να προβλέψει πλήρως την ακριβή τιμή τόσο για τα μέγιστα, όσο και για τα ελάχιστα. Ακόμη, υπάρχουν σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως. Τα παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, όπως λ.χ. το MAE, **Πίνακας 10**.

Σχήμα 41 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LR, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 42 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LR, Μαδρίτη, Ιδία Επεξεργασία*

Από το **Σχήμα 42** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, φαίνεται ότι δημιουργείται ένα ευθύγραμμο μοτίβο το οποίο «τέμνει» την ευθεία y=0, δεν είναι παράλληλο σε εκείνη. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο **Σχήμα 40** και λαμβάνοντας υπόψιν την υψηλή τιμή του MSE, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι πράγματι υπάρχουν ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες δεν επικαλύπτονται πλήρως. Η γραμμή για τα δεδομένα ελέγχου φαίνεται να ακολουθεί την τάση της γραμμής των δεδομένων εκπαίδευσης, αλλά να βρίσκεται μερικές μονάδες υψηλότερα.

Ακολουθεί η ανάλυση για το μοντέλο πρόβλεψης θανάτων για την πόλη της Μόσχας.

Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	**Μόσχα**
RMSE	81.61 (deaths per million)
R^2	0.603
EVS	0.610
MAE	54.55 (deaths per million)
MAPE	1.75%
MSE	6463.59

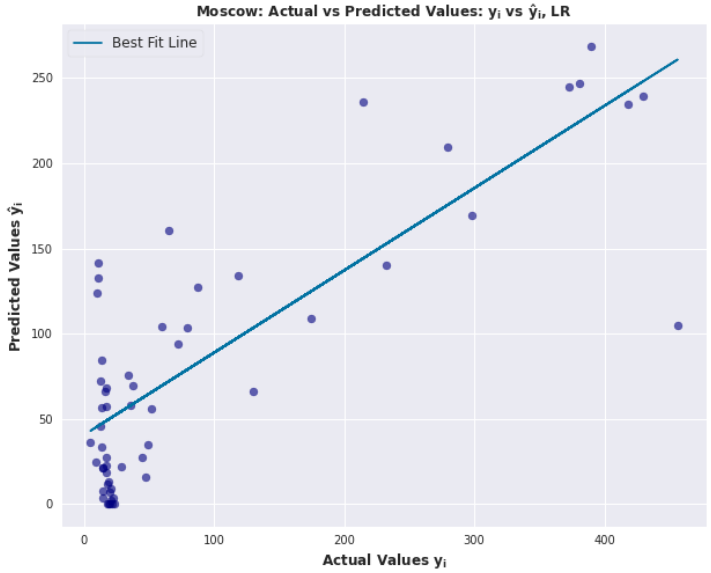
Πίνακας 11 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LR, Μόσχα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν μεσαία τιμή. Έχουν τιμή μεγαλύτερη από 0.60, και ως εκ τούτου έχουν μέτριο μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 44**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό, βρίσκεται κοντά στο 2%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο.

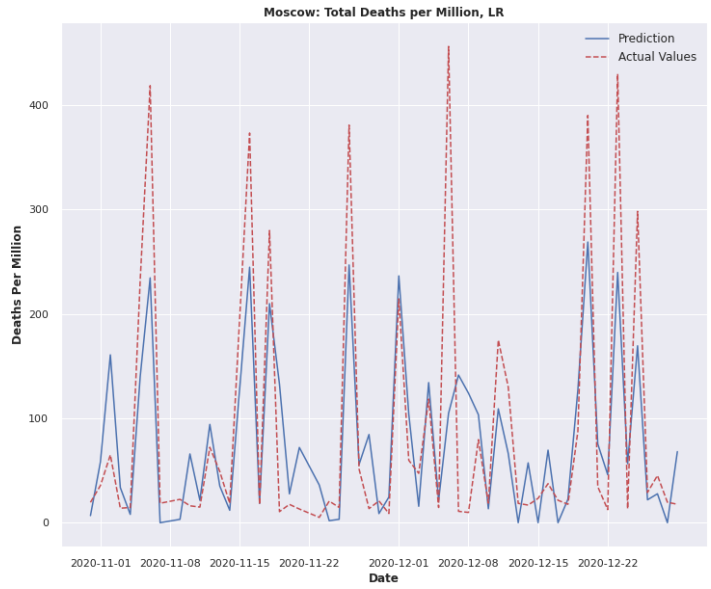
Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Μόσχας.

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μόσχας, δεν παρατηρείται ύπαρξη γραμμικότητας. Παρατηρείται ότι, όπως και στην περίπτωση των τριών προαναφερθέντων πόλεων, τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη». Στην περίπτωση αυτή, δημιουργείται μία έντονη συστάδα για ένα μικρό εύρος τιμών. Αυτό το οποίο γίνεται αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων είναι ότι τα σημεία δεν βρίσκονται κοντά της γραμμικά. Με αυτόν τον τρόπο μπορεί να δικαιολογηθεί η σχετικά χαμηλή τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, από το παραπάνω

διάγραμμα παρατηρούνται μερικά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 43 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LR, Μόσχα, Ιδία Επεξεργασία*



Σχήμα 44 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LR, Μόσχα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών σε αρκετά σημεία. Εν τούτοις, σε κάποια άλλα φαίνεται να την «υπερεκτιμάει» όπως, για παράδειγμα, στα μέσα του Δεκέμβρη (15/12/2020). Ακόμη, αδυνατεί να προβλέψει πλήρως την ακριβή τιμή τόσο για τα μέγιστα, όσο και για τα ελάχιστα. Έτσι, υπάρχουν σημεία τα οποία δεν έχουν προσαρμοστεί πλήρως. Τα παραπάνω, επιβεβαιώνονται και από τις τιμές των μετρικών, όπως λ.χ. το MAE και το RMSE, **Πίνακας 11**.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Σχήμα 45 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LR, Μόσχα, Ιδία Επεξεργασία*

Από το **Σχήμα 45** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, φαίνεται ότι δημιουργείται ένα ευθύγραμμο μοτίβο το οποίο «τέμνει» την ευθεία y=0, το οποίο, δηλαδή, δεν είναι παράλληλο σε εκείνη. Τα περισσότερα σημεία φαίνεται να βρίσκονται κυρίως σε αυτό το ευθύγραμμο τμήμα. Συνεπώς, παρατηρώντας ταυτόχρονα το **Σχήμα 43** και το **Σχήμα 45**, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι παρουσιάζονται ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των υπόλοιπων σημείων. Στο μοντέλο πρόβλεψης θανάτων για την πόλη της Μόσχας, φαίνεται ότι η απόδοση R^2 έχει υπολογισθεί να είναι μεγαλύτερη για τα δεδομένα ελέγχου, από ότι η R^2 για τα δεδομένα εκπαίδευσης.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερό τους τμήμα. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Έσχατη πόλη ανάλυσης, είναι η Πράγα.

Όπως και στις προηγούμενες τέσσερις αναλύσεις, παρουσιάζεται αρχικά ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

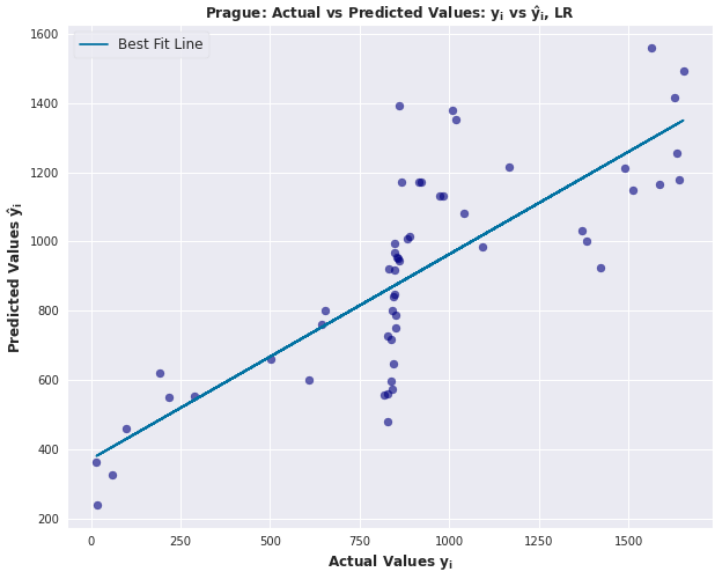
Μετρική Αξιολόγησης	**Πράγα**
RMSE	252.64 (deaths per million)
R^2	0.627
EVS	0.627
MAE	212.17 (deaths per million)
MAPE	1.03%
MSE	63828.73

Πίνακας 12 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LR, Πράγα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης της Πράγας, κυμαίνονται σε ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν μεσαία τιμή. Έχουν τιμή μεγαλύτερη από 0.60, και ως εκ τούτου έχουν μέτριο μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Πράγας (Βλέπε **Σχήμα 47**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά

χαμηλό, βρίσκεται κοντά στο 1%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Εκείνη η μετρική η οποία φαίνεται να εμφανίζει σημαντικά μεγαλύτερη τιμή, συγκριτικά με τις τιμές των υπόλοιπων προηγηθέντων τεσσάρων μετρικών, είναι το MSE. Όπως ήδη αιτιολογήθηκε και στην περίπτωση της Μαδρίτης, η υψηλή τιμή στο MSE πιθανότατα να οφείλεται σε ύπαρξη ιδιαζόντων σημείων. Επιπροσθέτως, πιθανότατα υποδηλώνει στατιστικά ότι υπάρχει σημαντική μεροληψία στην εκτίμηση και στην διασπορά της εκτιμήτριας.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Πράγας.

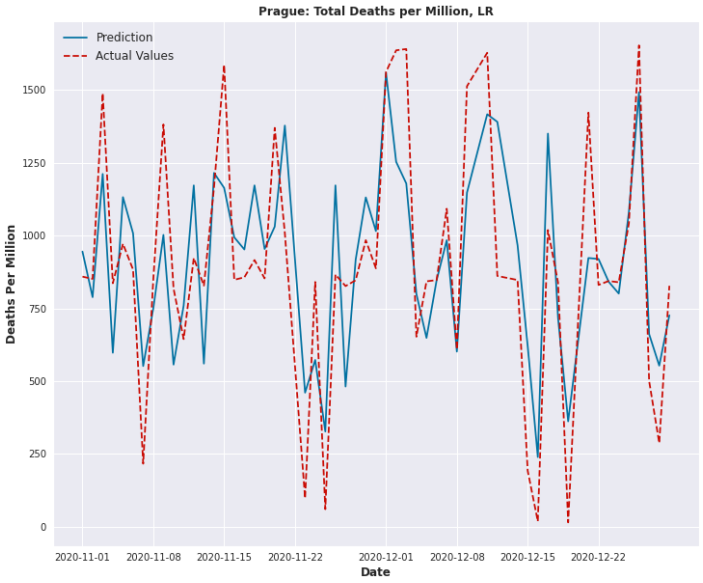


Σχήμα 46 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LR, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Πράγας, παρατηρείται μία ήπια ύπαρξη γραμμικότητας. Παρατηρείται ότι, όπως και στην περίπτωση των τεσσάρων προαναφερθέντων πόλεων, τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη». Στην περίπτωση αυτή, δημιουργείται μία χαρακτηριστική συστάδα. Παρατηρώντας την κατανομή των σημείων και τη σχέση που εμφανίζουν με την ευθεία ελαχίστων τετραγώνων, εδώ φαίνεται τα σημεία να βρίσκονται πιο «αρμονικά» κοντά στη γραμμή. Σίγουρα, η απεικόνιση αυτή δεν μπορεί να συγκριθεί με εκείνη για την πόλη του Παρισιού, **Σχήμα 34**. Με αυτόν τον τρόπο μπορεί να δικαιολογηθεί η σχετικά χαμηλή τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται ορισμένα ιδιάζοντα σημεία, τα οποία βρίσκονται αρκετά μακριά από τις συγκεντρώσεις των υπόλοιπων σημείων και από την ευθεία ελαχίστων τετραγώνων.

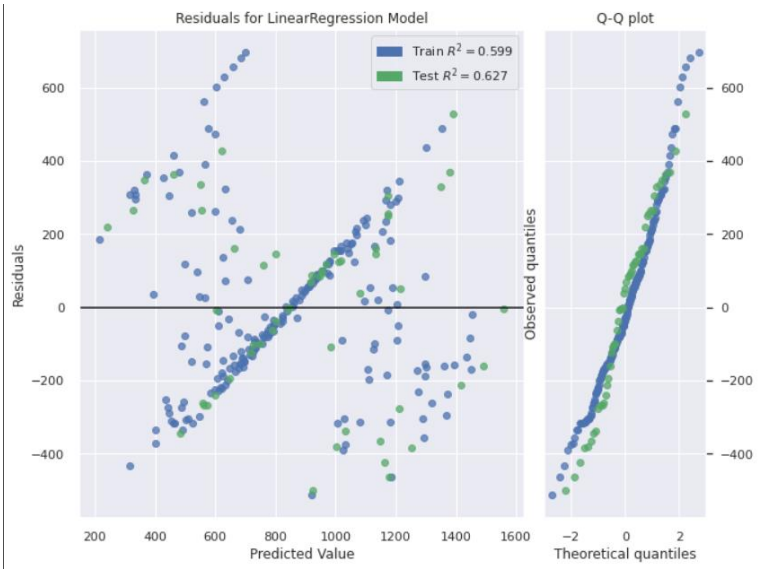
Στο παρακάτω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών σε αρκετά σημεία. Εν τούτοις, σε κάποια άλλα φαίνεται να την «υπερεκτιμάει» όπως, για παράδειγμα, στα μέσα του Δεκέμβρη (15/12/2020). Ακόμη, αδυνατεί να προβλέψει πλήρως την ακριβή τιμή τόσο για τα μέγιστα, όσο και για τα ελάχιστα. Έτσι, υπάρχουν σημεία τα

οποία δεν έχουν προσαρμοστεί πλήρως. Τα παραπάνω, επιβεβαιώνονται και από τις τιμές των μετρικών, όπως λ.χ. το MAE και το RMSE, **Πίνακας 12**.



Σχήμα 47 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LR, Πράγα, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 48 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LR, Πράγα, Ιδία Επεξεργασία*

Από το **Σχήμα 48** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όπως και στις προηγούμενες περιπτώσεις, φαίνεται ότι δημιουργείται ένα ευθύγραμμο μοτίβο το οποίο «τέμνει» την ευθεία y=0, δηλαδή, δεν είναι παράλληλο σε εκείνη. Τα περισσότερα σημεία φαίνεται να βρίσκονται κυρίως σε αυτό το ευθύγραμμο τμήμα. Συνεπώς, παρατηρώντας ταυτόχρονα το **Σχήμα 46** και το **Σχήμα 48**, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι παρουσιάζονται αρκετά ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των υπόλοιπων σημείων. Επίσης, παρατηρείται ότι το R^2 για τα δεδομένα εκπαίδευσης είναι μικρότερο από εκείνο των δεδομένων ελέγχου.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία από τα δεδομένα εκπαίδευσης και από τα δεδομένα ελέγχου εμφανίζουν δύο ευθείες γραμμές, οι οποίες όμως δεν επικαλύπτονται πλήρως. Μπορεί να θεωρηθεί ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή, καθώς η διαφοροποίησή τους είναι ελάχιστη.

Τα αποτελέσματα των μοντέλων πρόβλεψης θανάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Α.

4.4.2 SVR

Το δεύτερο μοντέλο το οποίο εφαρμόστηκε στα υπάρχοντα δεδομένα για τις εννέα διαφορετικές πόλεις, είναι το μοντέλο του Support Vector Regression. Στις επόμενες σελίδες ακολουθούν τα αποτελέσματα του αλγορίθμου για την Αθήνα, τη Μαδρίτη, τη Μόσχα, το Παρίσι και την Πράγα.

4.4.2.1 Πρόβλεψη Κρουσμάτων

Το μοντέλο SVR το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των κρουσμάτων, είναι εκείνο για την πόλη της Πράγας. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

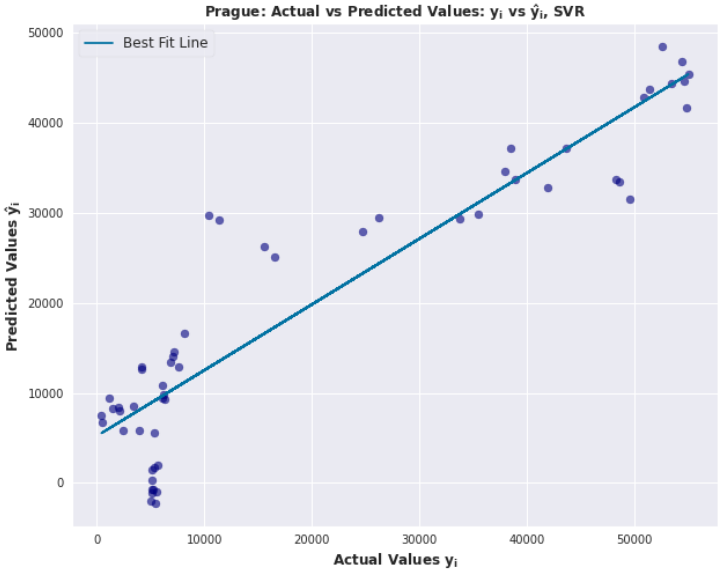
Μετρική Αξιολόγησης	Πράγα
RMSE	8114.60 (cases per million)
R^2	0.832
EVS	0.833
MAE	7061.33 (cases per million)
MAPE	1.45%
MSE	65846665.28

Πίνακας 13 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, SVR, Πράγα, Ιδία Επεξεργασία*

Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Πράγας, είναι ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.80, άρα η συσχέτιση είναι υψηλή. Δηλαδή, σημαίνει ότι 80% της μεταβολής της τελικής μεταβλητής μπορεί να εξηγηθεί από τις μεταβλητές εισόδου. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές 50000 κρούσματα (Βλέπε). Ακόμη το ποσοστό από το MAPE είναι χαμηλό και μικρότερο από 10%, άρα θεωρείται πολύ καλό. Τέλος, όπως έχει αναφερθεί, δεν υπάρχει κάποια «σωστή» τιμή για το MSE. Ο κύριος σκοπός χρήσης του είναι η επιλογή μίας πρόβλεψης ενός μοντέλου, έναντι κάποιας άλλης.

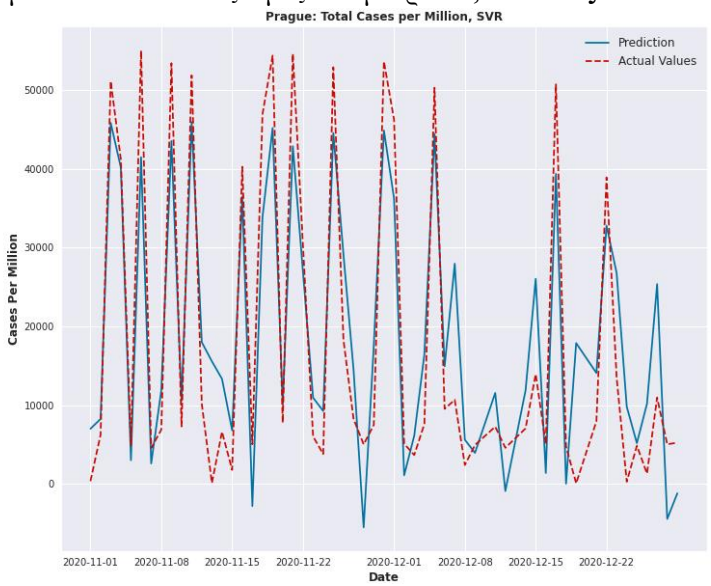
Στη συνέχεια, ακολουθεί το διάγραμμα διασποράς για την πόλη της Πράγας.

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Πράγας, **Σχήμα 49**, εμφανίζεται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν και να ακολουθούν και να μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Στην περίπτωση της Πράγας, θεωρείται ότι συναντάται μέτρια ισχύς, λόγω της κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζεται μία συστάδα σημείων η οποία απέχει από τα υπόλοιπα σημεία, όπως επίσης εντοπίζονται μεμονωμένα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.



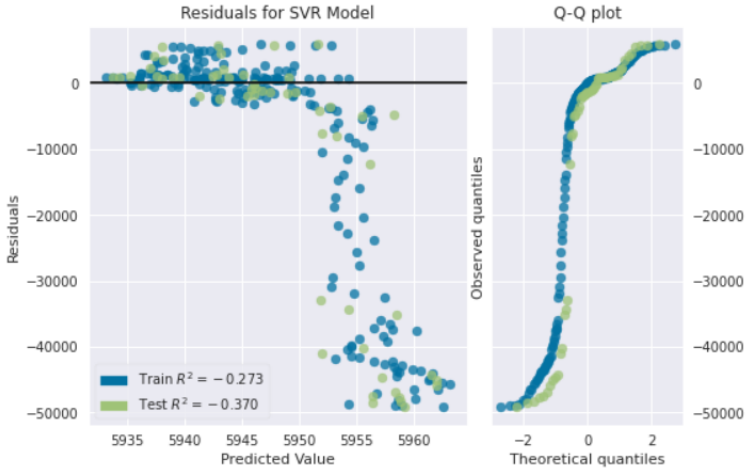
Σχήμα 49 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, SVR, Πράγα, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα, **Σχήμα 50**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων, και απεικονίζονται και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος SVR έχει αποδώσει ικανοποιητικά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως το διάστημα 7/12/2020 – 15/12/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα, γεγονός το οποίο μπορεί να επιβεβαιωθεί και από τις τιμές των μετρικών, **Πίνακας 13**.



Σχήμα 50 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, SVR, Πράγα, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 51 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, SVR, Πράγα, Ιδία Επεξεργασία*

Από το **Σχήμα 51** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται επίσης, ότι αρκετά σημεία υπολοίπων βρίσκονται σε σημαντική απόσταση από την ευθεία y=0. Γεγονός το οποίο φανερώνει ότι συναντώνται αρκετά πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού. Επιπλέον, έχουν τυπωθεί και οι τιμές για το R^2, τόσο για τα δεδομένα εκπαίδευσης, όσο και για τα δεδομένα ελέγχου. Για το test set είχε προηγηθεί ο υπολογισμός του R^2 και μέσω της βιβλιοθήκης της Scikit-learn.

Από το διάγραμμα Q-Q, **Σχήμα 51**, φαίνεται ότι τα σημεία των δύο σετ δεδομένων εμφανίζουν μία σχεδόν ευθεία γραμμή και διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Στη συνέχεια, ακολουθεί η ανάλυση για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη της Πράγας.

Όπως στην ανάλυση για την πόλη της Πράγας, έτσι ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

Μετρική Αξιολόγησης	**Αθήνα**
RMSE	4919.14 (cases per million)
R^2	0.844
EVS	0.845
MAE	3591.22 (cases per million)
MAPE	14.66%
MSE	24197925.86

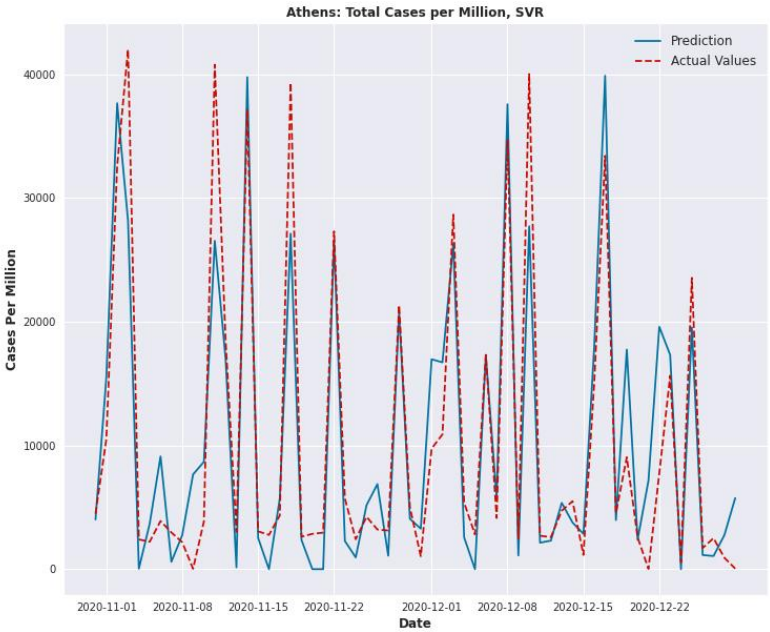
Πίνακας 14 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, SVR, Αθήνα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν σχετικά υψηλές τιμές. Βρίσκονται κοντά στο 0.85 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν υψηλή συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο βάσει του μεγέθους των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 53**). Ακόμη, το ποσοστό από το MAPE είναι σχετικά χαμηλό, βρίσκεται κοντά στο 15%, άρα θεωρείται καλό.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Αθήνας.

Σχήμα 52 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, SVR, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, **Σχήμα 52**, δεν παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας. Παρατηρείται, δηλαδή, ότι τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι σχετικά ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, λαμβάνοντας, όμως, υπόψιν ότι τα σημεία του διαγράμματος δεν έχουν ένα συγκεκριμένο γραμμικό μοτίβο.

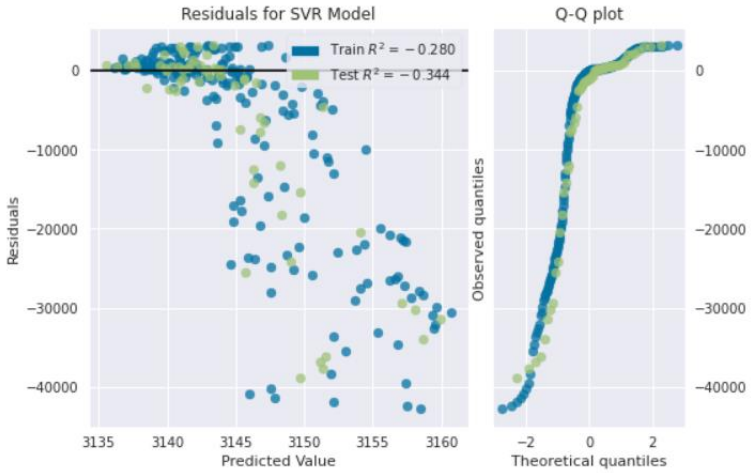


Σχήμα 53 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, SVR, Αθήνα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων, μαζί με τις πραγματικές τιμές των κρουσμάτων, για την αντίστοιχη χρονική περίοδο μελέτης.

Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για τον αλγόριθμο SVR, για την πόλη της Αθήνας, είναι ότι εντοπίζει την τάση των πραγματικών τιμών, εν τούτοις αδυνατεί να προβλέψει πλήρως τόσο όλα τα μέγιστα, όσο και όλα τα ελάχιστα. Ακόμη, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα. Τα παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, **Πίνακας 14**.

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 54 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, SVR, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όμως, παρατηρείται ότι αρκετά υπόλοιπα απέχουν σημαντικά από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο **Σχήμα 52**, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο επικαλυπτόμενες ευθείες γραμμές, άρα ταυτίζονται, και διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τρίτη πόλη ανάλυσης αποτελεί η πόλη της Μαδρίτης. Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης κρουσμάτων και για την πόλη της Μαδρίτης.

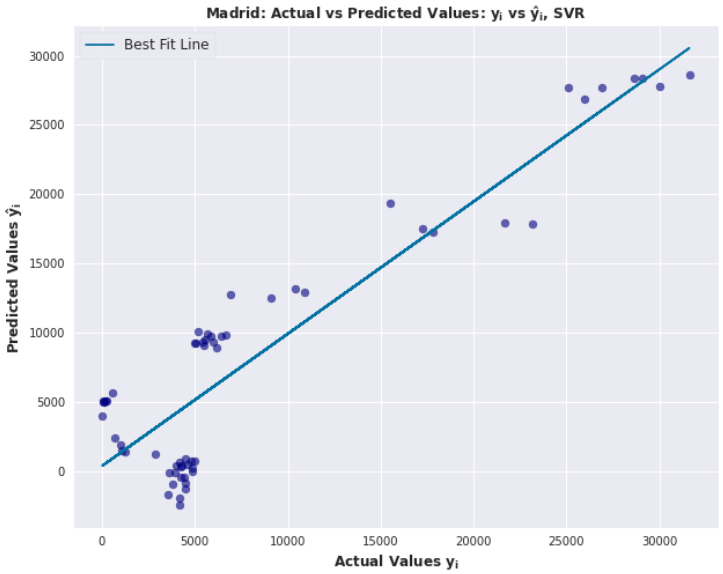
Μετρική Αξιολόγησης	**Μαδρίτη**
RMSE	3908.51 (cases per million)
R^2	0.798
EVS	0.798
MAE	3562.35 (cases per million)
MAPE	7.34%
MSE	15276436.84

Πίνακας 15 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, SVR, Μαδρίτη, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν τη Μαδρίτη, κυμαίνονται σε σχετικά ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν μεσαίες τιμές. Έχουν τιμή μεγαλύτερη από 0.70, και ως εκ τούτου θεωρείται ότι εμφανίζεται ισχυρό μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν

χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 56**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, βρίσκεται κοντά στο 8%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται σχετικά καλό. Άλλωστε, όπως έχει αναφερθεί, όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Να σημειωθεί ότι το MSE εμφανίζει σημαντικά μεγαλύτερη τιμή συγκριτικά με εκείνες των προηγούμενων πόλεων. Γεγονός το οποίο υποδηλώνει ότι το μοντέλο αυτό δεν είναι τόσο καλό όσο τα προηγούμενα.

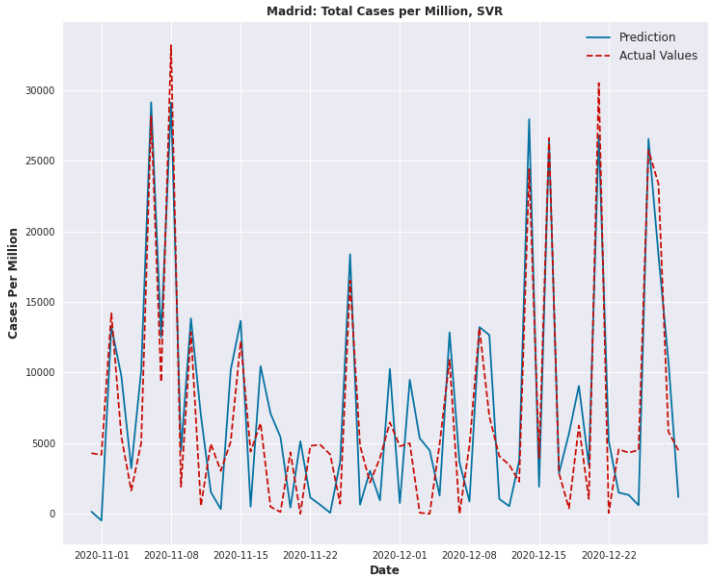
Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.



Σχήμα 55 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, SVR, Μαδρίτη, Ιδία Επεξεργασία*

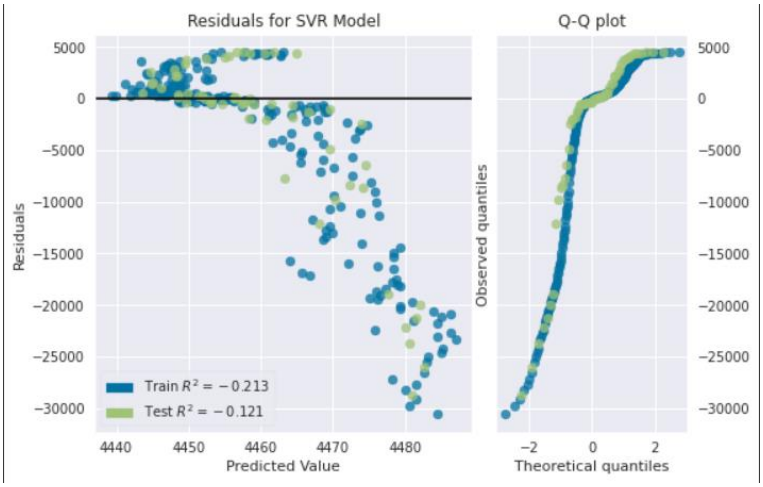
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, **Σχήμα 55**, δεν παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας. Παρατηρείται ότι, όπως και στην περίπτωση της Αθήνας, τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία «νοητή» ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί ορισμένες χαρακτηριστικές συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Ακόμη, με αυτόν τον τρόπο μπορεί να δικαιολογηθεί η τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται ορισμένα ιδιάζοντα σημεία, τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία. Σε κάθε περίπτωση, χρειάζεται να ληφθεί υπόψιν ότι τα σημεία του διαγράμματος δεν ακολουθούν ένα συγκεκριμένο γραμμικό μοτίβο.

Στο διάγραμμα το οποίο ακολουθεί, **Σχήμα 56**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Μαδρίτης και για το SVR, είναι ότι ο αλγόριθμος δεν εντοπίζει την τάση των πραγματικών τιμών με ιδιαίτερη επιτυχία, καθώς επίσης αδυνατεί να προβλέψει πλήρως την ακριβή τιμή τόσο για αρκετά μέγιστα, όσο και για αρκετά ελάχιστα. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές. Τα παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, όπως λ.χ. το MAE, **Πίνακας 15**.



Σχήμα 56 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, SVR, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 57 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, SVR, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα, **Σχήμα 57**, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, παρατηρείται ότι αρκετά από τα υπόλοιπα εμφανίζουν μεγάλες αρνητικές τιμές, μακριά από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές και λαμβάνοντας υπόψιν την υψηλή τιμή του MSE, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερο τμήμα τους. Έτσι, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Ακολουθεί η ανάλυση για το μοντέλο πρόβλεψης κρουσμάτων για την πόλη της Μόσχας.

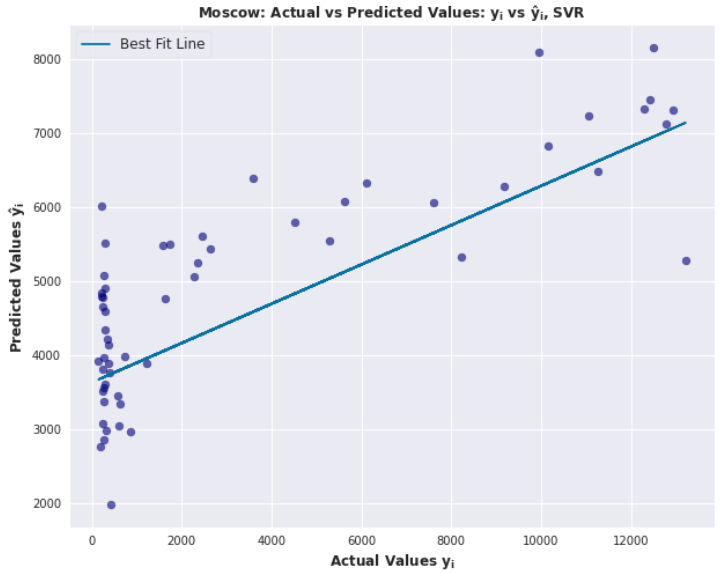
Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	Μόσχα
RMSE	3701.01 (cases per million)
R^2	0.307
EVS	0.424
MAE	3433.17 (cases per million)
MAPE	7.25%
MSE	13697448.27

Πίνακας 16 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, SVR, Μόσχα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν αρκετά χαμηλή τιμή. Έχουν τιμή μεγαλύτερη από 0.30 και μικρότερη από 0.50, και ως εκ τούτου οι μεταβλητές έχουν αδύναμο μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 59**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, βρίσκεται κοντά στο 7%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Το MSE, όπως στην περίπτωση της Αθήνας, της Μαδρίτης και της Πράγας, εμφανίζει 8ψήφιο αριθμό.

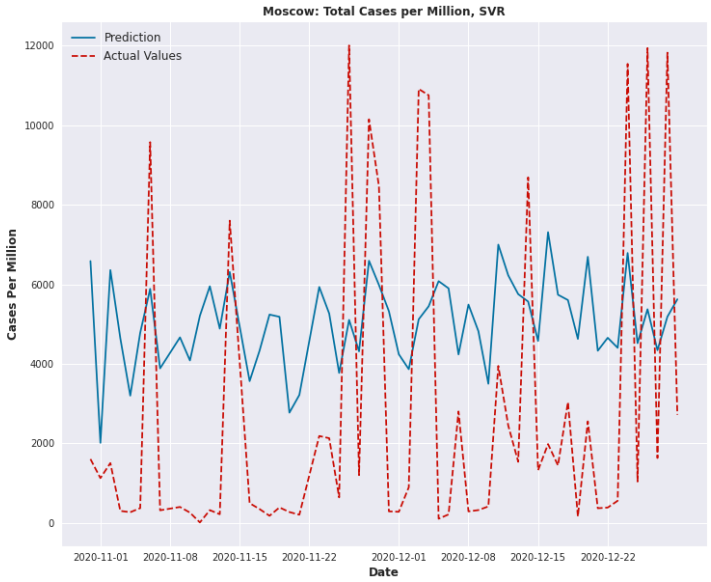
Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Μόσχας.



Σχήμα 58 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, SVR, Μόσχα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μόσχας, παρατηρείται ύπαρξη ήπιας γραμμικότητας. Αυτό το οποίο γίνεται αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων είναι ότι τα σημεία δεν βρίσκονται κοντά της γραμμικά για τις μικρότερες τιμές των πραγματικών μεταβλητών. Με αυτόν τον τρόπο μπορεί να δικαιολογηθεί η σχετικά χαμηλή τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, από ένα σημείο και μετά, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται

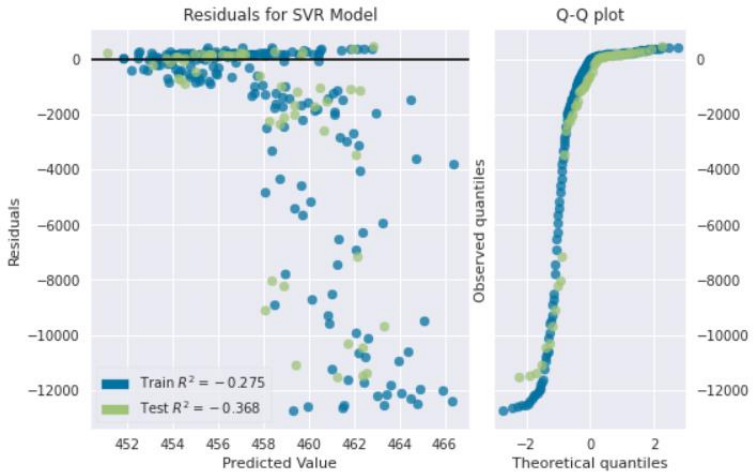
μερικά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 59 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, SVR, Μόσχα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι δεν έχει γίνει καθόλου καλή προσαρμογή του αλγορίθμου στις πραγματικές τιμές των κρουσμάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει κατά κάποιον τρόπο, την τάση των κρουσμάτων, όμως, αδυνατεί πλήρως να υπολογίσει τις πραγματικές τιμές στην πλειονότητα της χρονικής μελέτης. Σε αυτό το διάγραμμα, για πρώτη φορά στην έως τώρα ανάλυση, υπάρχουν πάρα πολλά σημεία τα οποία δεν έχουν προσαρμοστεί πλήρως. Τα παραπάνω, επιβεβαιώνονται και από τις ιδιαίτερες τιμές των μετρικών, όπως λ.χ. το MAE και το RMSE, **Πίνακας 16**.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 60 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, SVR, Μόσχα, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όμως, για μία ακόμη φορά για τον αλγόριθμο SVR, συναντώνται αρκετά σημεία σε μεγάλες αρνητικές τιμές, μακριά από την ευθεία y=0. Αυτό το

οποίο μπορεί να γίνει αντιληπτό, είναι ότι παρουσιάζονται ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των υπόλοιπων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερό τους τμήμα. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Για να ολοκληρωθεί η ανάλυση των μοντέλων SVR για πρόβλεψη κρουσμάτων στις μελετώμενες πόλεις, χρειάζεται να παρατεθούν τα αποτελέσματα τα οποία προκύπτουν για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	**Παρίσι**
RMSE	2930.74 (cases per million)
R^2	0.760
EVS	0.792
MAE	2433.17 (cases per million)
MAPE	4.12%
MSE	8589214.66

Πίνακας 17 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, SVR, Παρίσι, Ιδία Επεξεργασία*

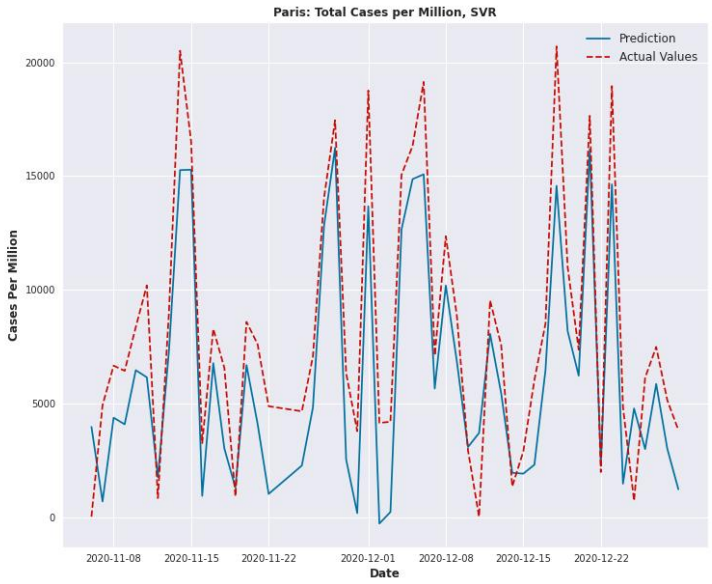
Βάσει όσων αναφέρθηκαν στην υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Παρισιού, είναι ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.70, άρα η συσχέτιση των μεταβλητών θεωρείται υψηλή. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 62**). Ακόμη το ποσοστό από το MAPE είναι χαμηλό και μικρότερο από 10%, άρα θεωρείται πολύ καλό. Το MSE εν αντιθέσει με την περίπτωση της Αθήνας, της Μαδρίτης, της Μόσχας και της Πράγας, εμφανίζει 7ψήφιο αριθμό. Άρα αριθμό μικρότερο από εκείνο των υπόλοιπων πόλεων.

Σχήμα 61 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, SVR, Παρίσι, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη του Παρισιού, παρατηρείται ότι υπάρχει μία σχετική γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του

διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται ήπια ισχύς, λόγω της μικρής κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.

Στο παρακάτω διάγραμμα, **Σχήμα 62**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος SVR έχει αποδώσει σχετικά καλά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα, γεγονός το οποίο μπορεί να επιβεβαιωθεί και από τις τιμές των μετρικών, **Πίνακας 8**. Υπάρχουν, τέλος, αρκετά σημεία τα οποία δεν έχει καταφέρει να προσαρμόσει στις πραγματικές τιμές των κρουσμάτων για την πόλη του Παρισιού.

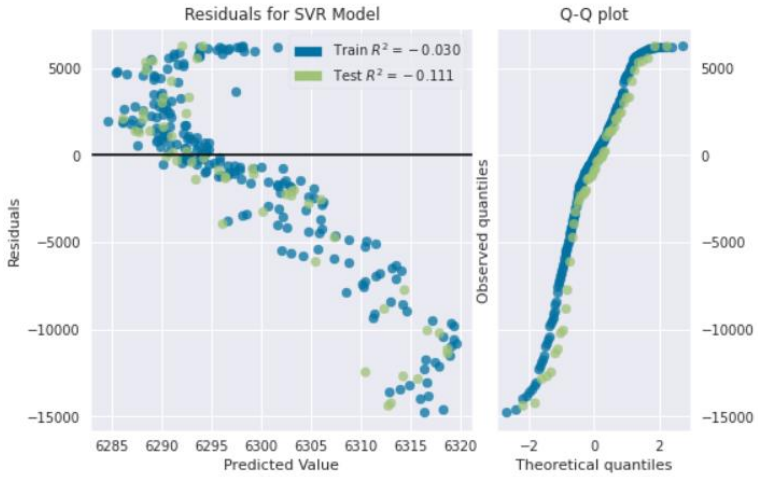


Σχήμα 62 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, SVR, Παρίσι, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όμως, τα σημεία αυτά και σε αυτήν την περίπτωση, εμφανίζουν σημαντικές αρνητικές τιμές, μακριά από τον οριζόντιο άξονα.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία εμφανίζουν δύο σχεδόν ταυτιζόμενες ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.



Σχήμα 63 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, SVR, Παρίσι, Ιδία Επεξεργασία*

Τα αποτελέσματα των μοντέλων πρόβλεψης κρουσμάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Β.

4.4.2.2 *Πρόβλεψη Θανάτων*

Το μοντέλο SVR το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των θανάτων, είναι εκείνο για την πόλη της Μαδρίτης. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

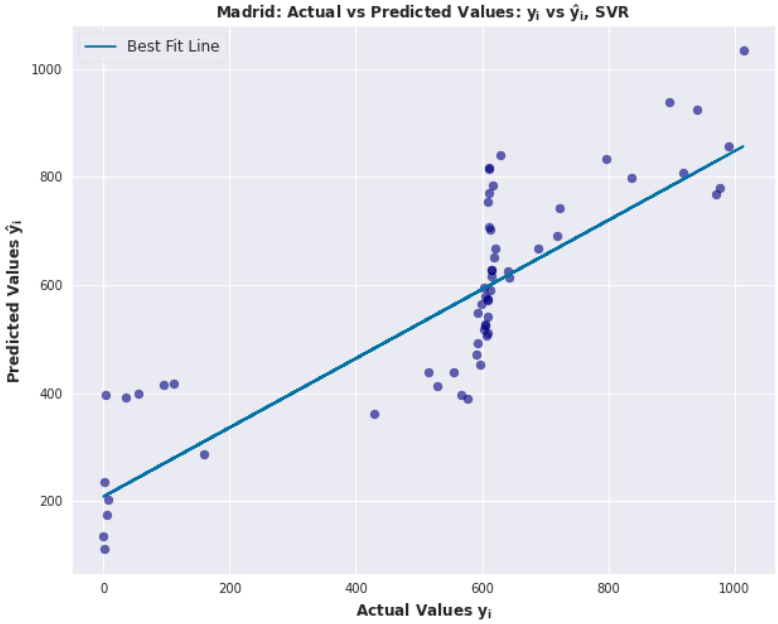
Μετρική Αξιολόγησης	**Μαδρίτη**
RMSE	147.30 (deaths per million)
R^2	0.703
EVS	0.713
MAE	113.24(deaths per million)
MAPE	25.00%
MSE	21696.47

Πίνακας 18 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, SVR, Μαδρίτη, Ιδία Επεξεργασία*

Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Πράγας, είναι ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.70, άρα η συσχέτιση των μεταβλητών είναι σχετικά υψηλή. Δηλαδή, σημαίνει ότι 70% της μεταβολής της τελικής μεταβλητής μπορεί να εξηγηθεί από τις μεταβλητές εισόδου. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές 1000 κρούσματα (Βλέπε **Σχήμα 65**). Ακόμη το ποσοστό από το MAPE είναι μεσαίο και ίσο με 25%, άρα θεωρείται μέτριο. Τέλος, όπως έχει αναφερθεί, δεν υπάρχει κάποια «σωστή» τιμή για το MSE. Ο κύριος σκοπός χρήσης του είναι η επιλογή μίας πρόβλεψης ενός μοντέλου, έναντι κάποιας άλλης. Η τιμή του είναι 5ψήφια.

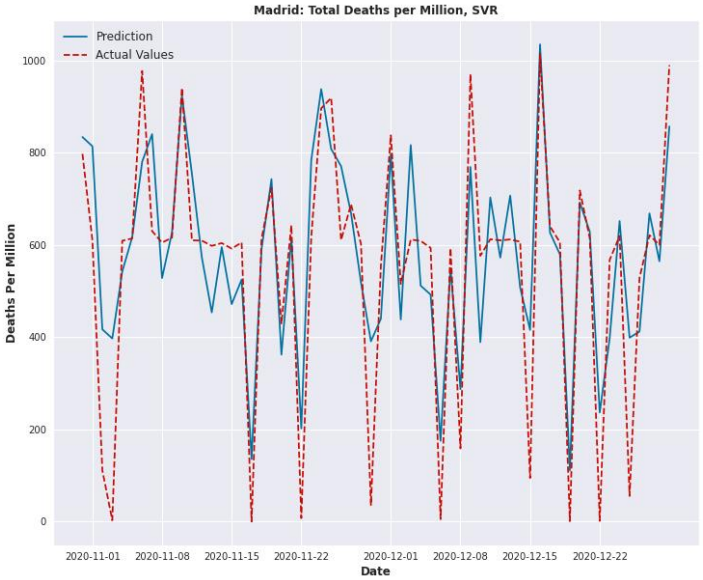
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, δεν εμφανίζεται ιδιαίτερη γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος δεν σχηματίζουν και δεν ακολουθούν, ούτε μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων μπορούν εύκολα να γίνουν όσα προαναφέρθηκαν αντιληπτά. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Στην περίπτωση της Μαδρίτης, θεωρείται ότι συναντάται ήπια ισχύς, λόγω της μικρής κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά

τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από μελέτη του παραπάνω διαγράμματος, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία απέχουν από τα υπόλοιπα και τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.



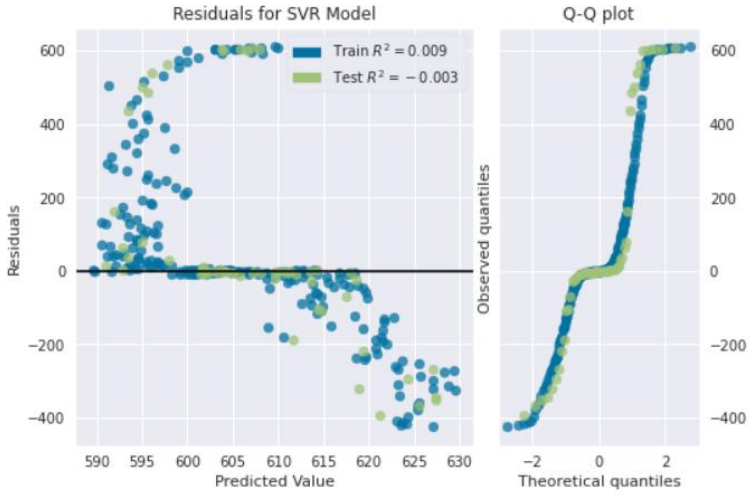
Σχήμα 64 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, SVR, Μαδρίτη, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα, **Σχήμα 65**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και απεικονίζονται και οι πραγματικές τιμές των θανάτων για την πόλη της Μαδρίτης και για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος SVR έχει αποδώσει μέτρια. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως το διάστημα 9/12/2020 – 15/12/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα, γεγονός το οποίο μπορεί να επιβεβαιωθεί και από τις τιμές των μετρικών, **Πίνακας 18**. Έτσι, αρκετά προβλεπόμενα σημεία φαίνεται ότι δεν προσαρμόζονται πλήρως στις πραγματικές τιμές των θανάτων.



Σχήμα 65 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, SVR, Μαδρίτη, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 66 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, SVR, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα, **Σχήμα 66** φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται επίσης, ότι αρκετά σημεία υπολοίπων βρίσκονται σε σημαντική απόσταση από την ευθεία y=0, τόσο σε θετικές τιμές, όσο και σε αρνητικές. Γεγονός το οποίο φανερώνει ότι συναντώνται αρκετά πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού. Επιπλέον, έχουν τυπωθεί και οι τιμές για το R^2, τόσο για τα δεδομένα εκπαίδευσης, όσο και για τα δεδομένα ελέγχου.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων εμφανίζουν μία σχεδόν ευθεία γραμμή και διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Στη συνέχεια, ακολουθεί η ανάλυση μοντέλων για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη της Μαδρίτης.

Ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

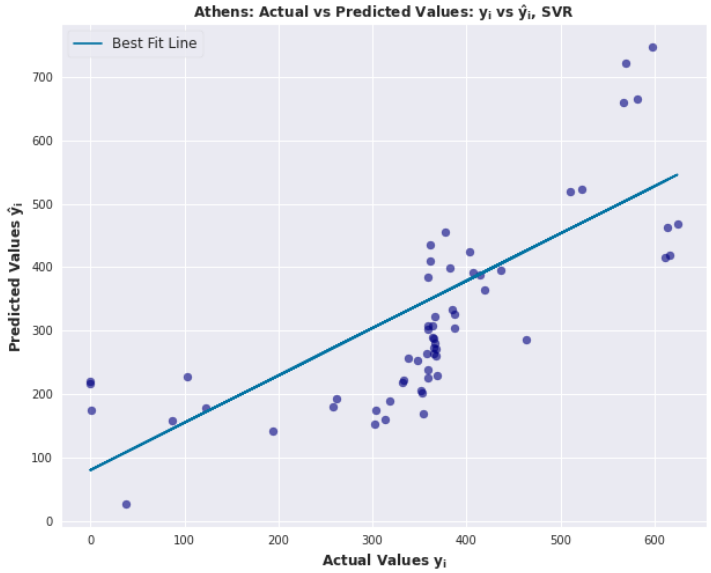
Μετρική Αξιολόγησης	**Αθήνα**
RMSE	111.04 (deaths per million)
R^2	0.424
EVS	0.510
MAE	96.64 (deaths per million)
MAPE	38.40%
MSE	12330.85

Πίνακας 19 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, SVR, Αθήνα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε χαμηλά πλαίσια. Η R^2 και EVS εμφανίζουν αρκετά χαμηλές τιμές. Βρίσκονται κοντά στο 0.50 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν αδύναμη συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο βάσει του μεγέθους των καταγεγραμμένων τιμών των

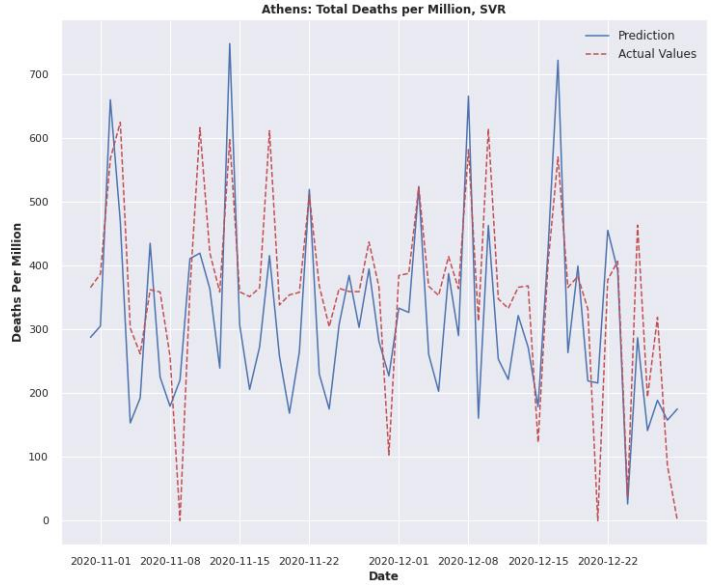
κρουσμάτων (Βλέπε **Σχήμα 68**). Ακόμη, το ποσοστό από το MAPE είναι σχετικά υψηλό, βρίσκεται κοντά στο 40%, άρα θεωρείται μέτριο.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Αθήνας.



Σχήμα 67 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, SVR, Αθήνα, Ιδία Επεξεργασία*

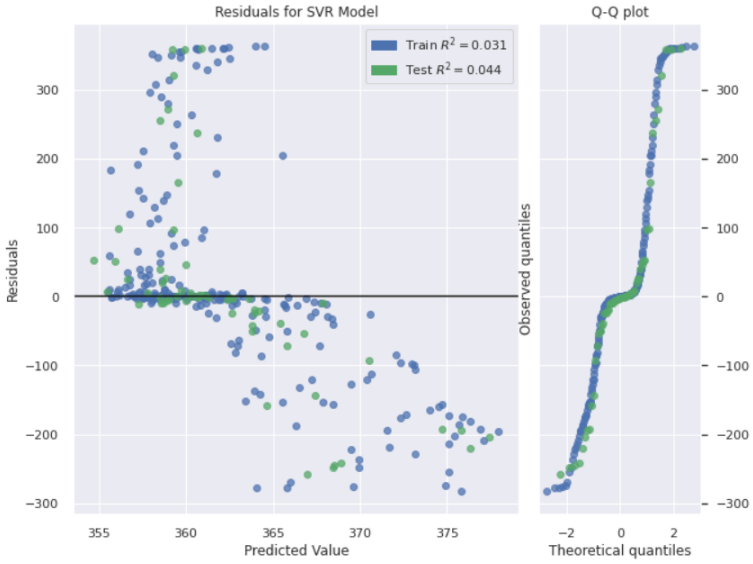
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, παρατηρείται σχετική ύπαρξη γραμμικότητας. Παρατηρείται, δηλαδή, ότι τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργούνται συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, καθώς παρατηρούνται σημεία τα οποία απέχουν τόσο από την ευθεία ελαχίστων τετραγώνων όσο και από τις συγκεντρώσεις των υπόλοιπων σημείων.



Σχήμα 68 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, SVR, Αθήνα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων, μαζί με τις πραγματικές τους τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για τον αλγόριθμο SVR, για την πόλη της Αθήνας, είναι ότι εντοπίζει την τάση των πραγματικών τιμών. Εν τούτοις αδυνατεί να προβλέψει πλήρως τις ακριβείς τιμές, τα περισσότερα μέγιστα, και τα περισσότερα ελάχιστα. Έτσι, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα. Σε ορισμένες περιπτώσεις φαίνεται να υπερεκτιμάει και σε άλλες φαίνεται να υποτιμάει τον προβλεπόμενο αριθμό κρουσμάτων. Τα παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, **Πίνακας 19**.

Στο ακόλουθο διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 69 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, SVR, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ακόμη, παρατηρείται ότι αρκετά υπόλοιπα απέχουν σημαντικά από την ευθεία y=0, τόσο ως προς τις θετικές τιμές όσο και προς τις αρνητικές. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές για την πόλη της Αθήνας, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο επικαλυπτόμενες ευθείες γραμμές, άρα ταυτίζονται. Ως εκ τούτου, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τρίτη πόλη ανάλυσης αποτελεί η πόλη της Μόσχας. Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

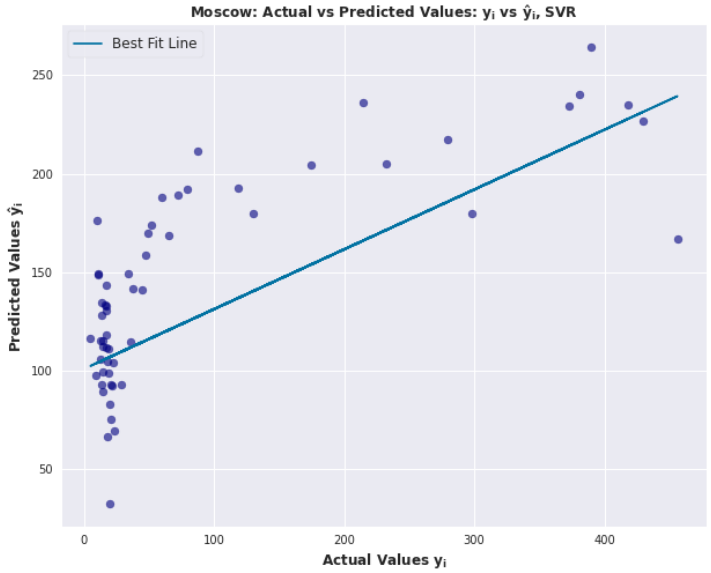
Μετρική Αξιολόγησης	**Μόσχα**
RMSE	110.36 (deaths per million)
R^2	0.252
EVS	0.442
MAE	100.93 (deaths per million)
MAPE	4.29%

MSE	12178.26

Πίνακας 20 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, SVR, Μόσχα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε χαμηλά πλαίσια. Η R^2 και EVS εμφανίζουν αρκετά χαμηλή τιμή. Η τιμή για το R^2 είναι μικρότερο από 0.30 και ως εκ τούτου οι μεταβλητές έχουν ανύπαρκτο ή πολύ αδύναμο μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 71**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, βρίσκεται κοντά στο 4%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Το MSE, όπως στην περίπτωση της Αθήνας και της Μαδρίτης, εμφανίζει 5ψήφιο αριθμό.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Μόσχας.

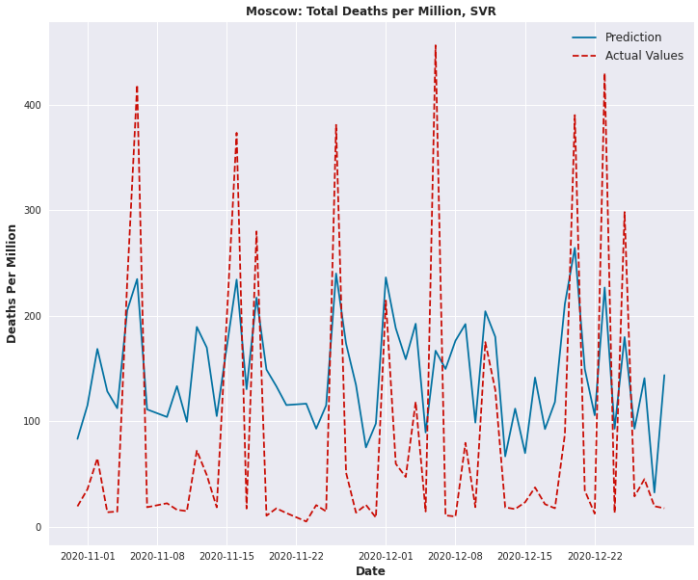


Σχήμα 70 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, SVR, Μόσχα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μόσχας, παρατηρείται ύπαρξη πολύ ήπιας γραμμικότητας. Αυτό το οποίο γίνεται αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων είναι ότι τα σημεία δεν βρίσκονται κοντά της με γραμμικό τρόπο. Έτσι, μπορεί να δικαιολογηθεί η σχετικά χαμηλή τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι αρκετά ήπια. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται μερικά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τις συγκεντρώσεις των υπόλοιπων σημείων και από την ευθεία ελαχίστων τετραγώνων.

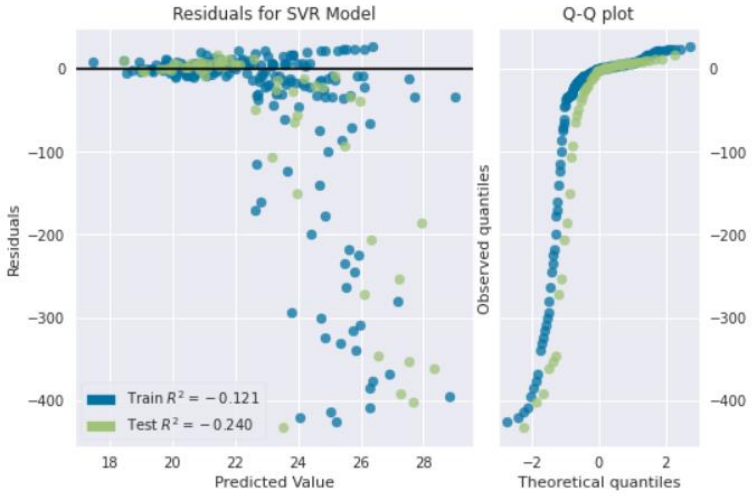
Στο παρακάτω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι δεν έχει γίνει καθόλου καλή προσαρμογή του αλγορίθμου στις πραγματικές τιμές των θανάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει κατά κάποιον τρόπο, την τάση των κρουσμάτων, όμως, αδυνατεί πλήρως να υπολογίσει τις πραγματικές τιμές στην πλειονότητα της χρονικής περιόδου μελέτης. Τα

παραπάνω, επιβεβαιώνονται και από τις ιδιαίτερες τιμές των μετρικών, όπως λ.χ. το R^2, το MAE και το RMSE, **Πίνακας 20**. Στις προβλεπόμενες τιμές, παρατηρείται μία μετατόπιση στις τιμές και την τάση, από εκείνη των πραγματικών τιμών. Όπως και στην περίπτωση των κρουσμάτων, για το μοντέλο SVR της Μόσχας, δεν έχει γίνει καθόλου καλή προσαρμογή.



Σχήμα 71 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, SVR, Μόσχα, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 72 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, SVR, Μόσχα, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όμως, για τον αλγόριθμο SVR συναντώνται ξανά αρκετά σημεία σε μεγάλες αρνητικές τιμές, μακριά από την ευθεία y=0. Αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι παρουσιάζονται ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των υπόλοιπων σημείων.

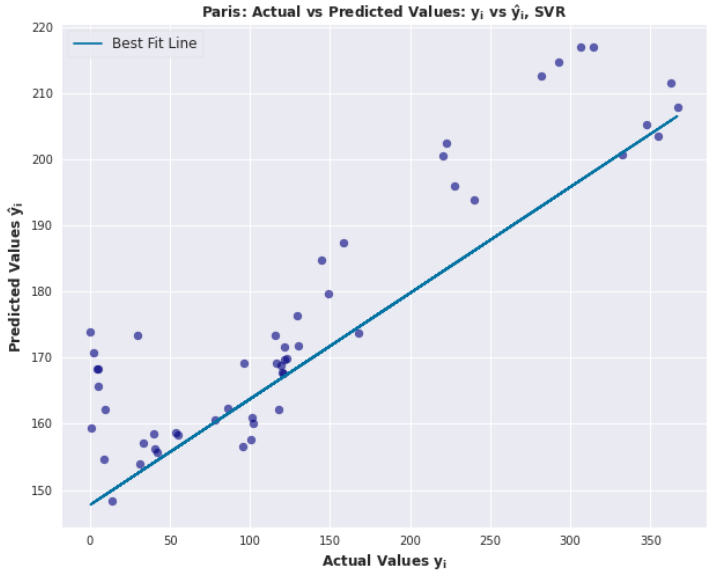
Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες δεν επικαλύπτονται πλήρως, όμως ακολουθούν την ίδια τάση. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη του Παρισιού.

Μετρική Αξιολόγησης	Παρίσι
RMSE	101.98 (deaths per million)
R^2	0.127
EVS	0.289
MAE	89.63 (deaths per million)
MAPE	16.14%
MSE	10400.79

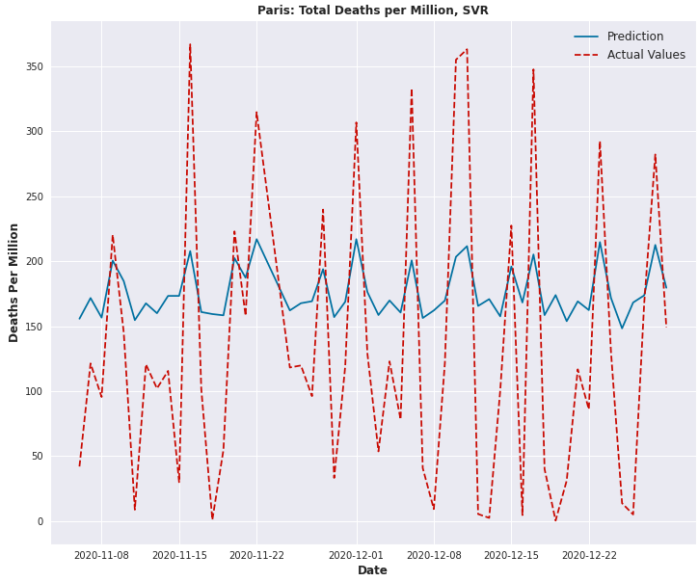
Πίνακας 21 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, SVR, Παρίσι, Ιδία Επεξεργασία*

Βάσει όσων αναφέρθηκαν στην υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Παρισιού, είναι δεν κυμαίνονται σε αρκετά ικανοποιητικά πλαίσια. Η τιμή της R^2 και της EVS είναι αρκετά χαμηλές. Η τιμή της R^2 βρίσκεται κοντά στο 0.10, άρα η συσχέτιση των μεταβλητών θεωρείται ανύπαρκτη ή πάρα πολύ χαμηλή. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε). Ακόμη, το ποσοστό από το MAPE είναι σχετικά χαμηλό και μικρότερο από 20%, άρα θεωρείται καλό. Το MSE, όπως και στην περίπτωση της Αθήνας, της Μαδρίτης και της Μόσχας, εμφανίζει το μικρότερο 5ψήφιο αριθμό.



Σχήμα 73 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, SVR, Παρίσι, Ιδία Επεξεργασία*

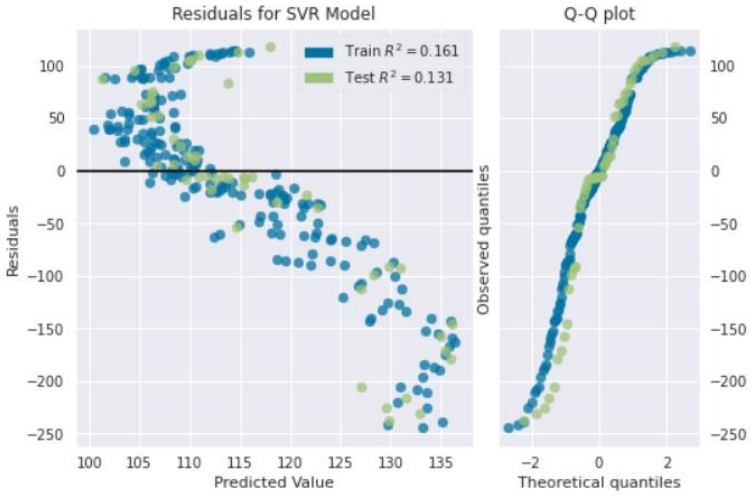
Από το παραπάνω διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη του Παρισιού, παρατηρείται ότι υπάρχει μία γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται ήπια ισχύς, λόγω της μικρής κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται λίγα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.



Σχήμα 74 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, SVR, Παρίσι, Ιδία Επεξεργασία*

Στο διάγραμμα, **Σχήμα 74**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος SVR δεν έχει αποδώσει καθόλου καλά. Δεν είναι ικανός να ακολουθήσει πιστά τη ροή και την τάση των πραγματικών τιμών. Ούτε μπορεί να προβλέψει με πλήρη επιτυχία όλα τα μέγιστα, αλλά και όλα τα ελάχιστα, γεγονός το οποίο μπορεί να επιβεβαιωθεί και από τις χαμηλές τιμές των μετρικών, **Πίνακας 21**. Άρα, αρκετά σημεία δεν έχουν προσαρμοστεί στις πραγματικές τιμές των θανάτων για την πόλη του Παρισιού.

Στη συνέχεια, παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 75 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, SVR, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι ένα πλήθος τους βρίσκεται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όμως, ένα σημαντικό πλήθος των σημείων, εμφανίζουν αρνητικές τιμές, μακριά από τον οριζόντιο άξονα.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία εμφανίζουν δύο σχεδόν ταυτιζόμενες ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

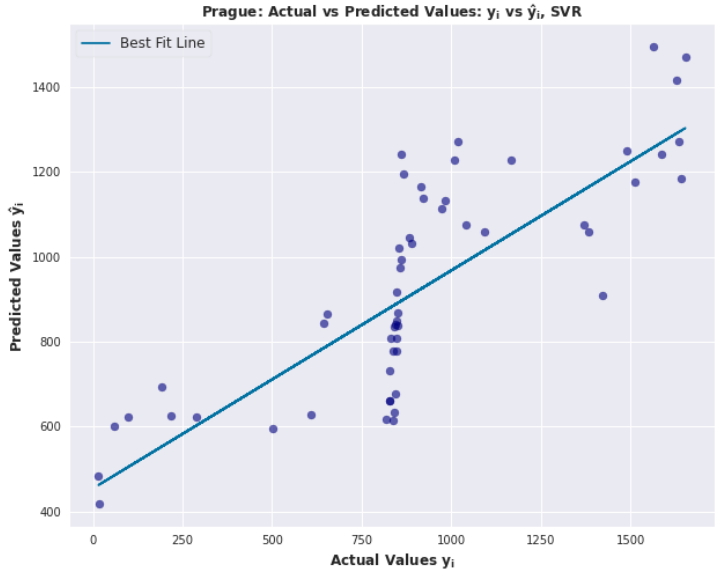
Για να ολοκληρωθεί η ανάλυση των μοντέλων SVR για πρόβλεψη θανάτων στις μελετώμενες πόλεις, χρειάζεται να παρατεθούν τα αποτελέσματα τα οποία προκύπτουν για την πόλη της Πράγας. Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης θανάτων για την πόλη της Πράγας.

Μετρική Αξιολόγησης	Πράγα
RMSE	256.13 (deaths per million)
R^2	0.616
EVS	0.622
MAE	206.43 (deaths per million)
MAPE	1.45%
MSE	65600.03

Πίνακας 22 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, SVR, Πράγα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την Πράγα, κυμαίνονται σε σχετικά ικανοποιητικά πλαίσια. Η R^2 και EVS εμφανίζουν μεσαίες τιμές. Έχουν τιμή μεγαλύτερη από 0.60, και ως εκ τούτου θεωρείται ότι εμφανίζεται μέτριο μέτρο επίδρασης, δηλαδή ότι η συσχέτιση των μεταβλητών είναι μέτρια (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Πράγας (Βλέπε **Σχήμα 76**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό, βρίσκεται κοντά στο 2%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται αρκετά καλό. Άλλωστε, όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Πρόκειται για το πιο χαμηλό MAPE το οποίο συναντάται στην περίπτωση των θανάτων με τον αλγόριθμο SVR. Να σημειωθεί ότι το MSE του εν λόγω μοντέλου, εμφανίζει μεγαλύτερη τιμή συγκριτικά με εκείνες των προηγούμενων πόλεων.

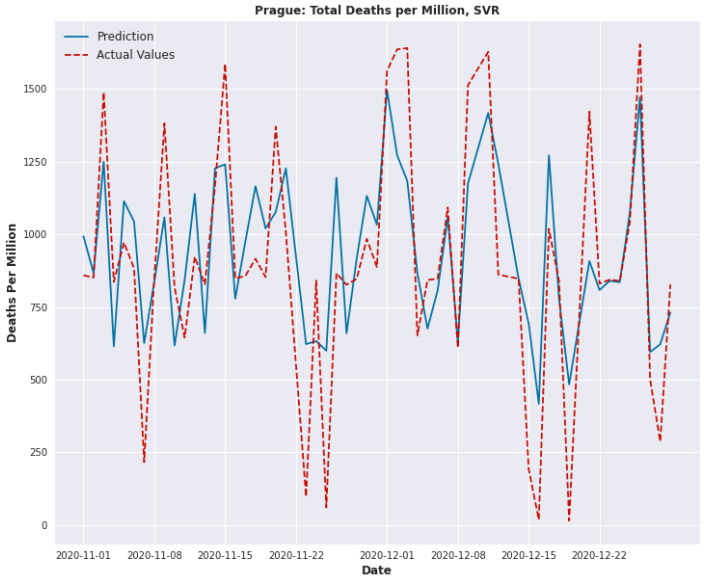
Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Πράγας.



Σχήμα 76 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, SVR, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Πράγας, παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας ανάμεσα στα σημεία. Παρατηρείται ότι η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί ορισμένες συστάδες. Όλα αυτά γίνονται αντιληπτά

μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται ορισμένα ιδιάζοντα σημεία, τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία. Σε κάθε περίπτωση, χρειάζεται να ληφθεί υπόψιν ότι τα σημεία του διαγράμματος δεν ακολουθούν ένα συγκεκριμένο γραμμικό μοτίβο.



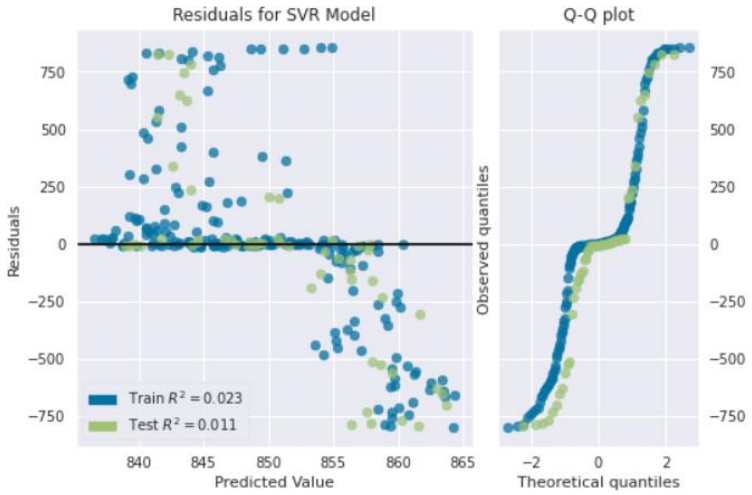
Σχήμα 77 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, SVR, Πράγα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα, **Σχήμα 77**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Πράγας και για το SVR, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών ικανοποιητικά, εν τούτοις αδυνατεί να προβλέψει πλήρως την ακριβή τιμή τόσο για αρκετά μέγιστα, όσο και για αρκετά ελάχιστα. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές. Χαρακτηριστικό είναι το παράδειγμα για τις ημερομηνίες 20/11/2020 έως και περίπου 25/11/2020. Τα παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, όπως λ.χ. το MAE, **Πίνακας 22**.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Από το διάγραμμα των υπολοίπων, **Σχήμα 78**, φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, παρατηρείται ότι αρκετά από τα υπόλοιπα εμφανίζουν μεγάλες αρνητικές τιμές και μεγάλες θετικές τιμές, μακριά από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές και λαμβάνοντας υπόψιν την υψηλή τιμή του MSE, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερο τμήμα τους. Έτσι, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.



Σχήμα 78 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, SVR, Πράγα, Ιδία Επεξεργασία*

Τα αποτελέσματα των μοντέλων πρόβλεψης θανάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Β.

4.4.3 LASSO

Τρίτο μοντέλο το οποίο εφαρμόστηκε στα υπάρχοντα δεδομένα για τις εννέα διαφορετικές πόλεις, είναι το LASSO. Στις επόμενες σελίδες ακολουθούν τα αποτελέσματα του αλγορίθμου για την Αθήνα, τη Μαδρίτη, τη Μόσχα, το Παρίσι και την Πράγα.

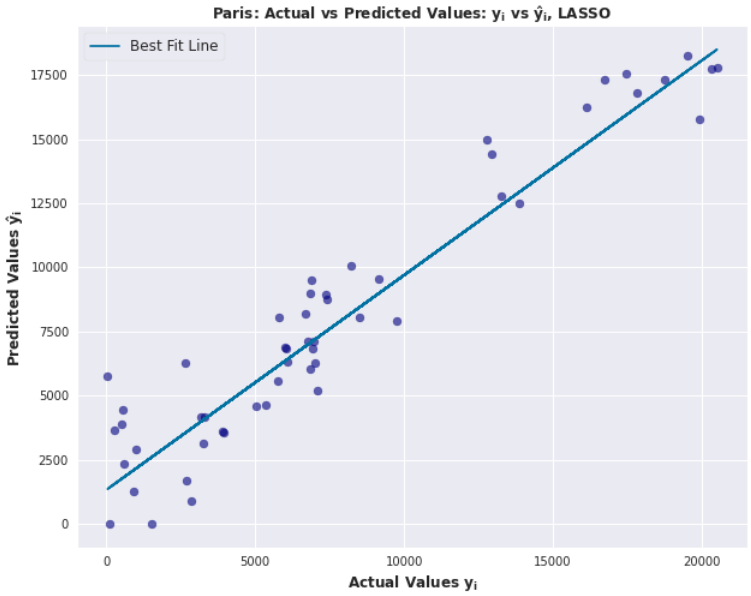
4.4.3.1 Πρόβλεψη Κρουσμάτων

Το LASSO μοντέλο το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των κρουσμάτων, είναι εκείνο για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	**Παρίσι**
RMSE	1874.56 (cases per million)
R^2	0.902
EVS	0.906
MAE	1434.74 (cases per million)
MAPE	3.00%
MSE	3499471.65

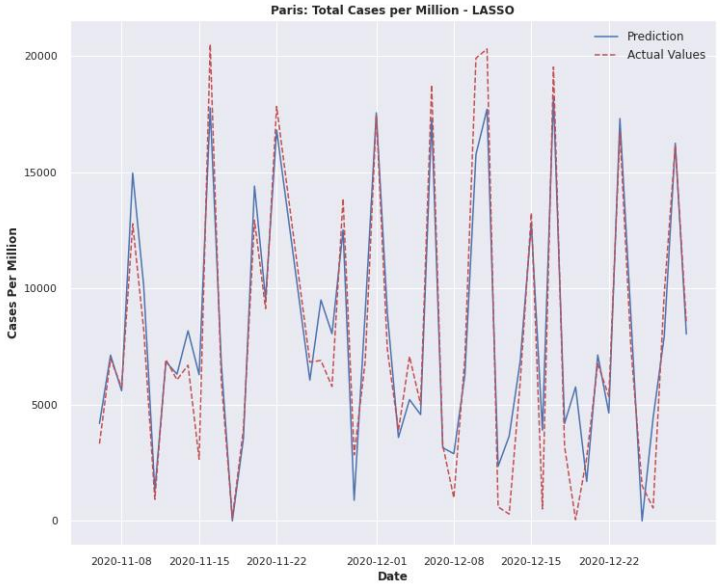
Πίνακας 23 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LASSO, Παρίσι, Ιδία Επεξεργασία*

Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Παρισιού, είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.90, άρα πρόκειται για υψηλή συσχέτιση μεταβλητών. Δηλαδή, σημαίνει ότι 90% της μεταβολής της τελικής μεταβλητής μπορεί να εξηγηθεί από τις μεταβλητές εισόδου. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές 20000 κρούσματα (Βλέπε **Σχήμα 80**). Ακόμη το ποσοστό από το MAPE είναι χαμηλό και μικρότερο από 10%, άρα θεωρείται πολύ καλό. Τέλος, όπως έχει αναφερθεί, δεν υπάρχει κάποια «σωστή» τιμή για το MSE. Ο κύριος σκοπός χρήσης του είναι η επιλογή μίας πρόβλεψης ενός μοντέλου, έναντι κάποιας άλλης. Στην περίπτωση αυτή, πρόκειται για έναν επταψήφιο αριθμό, ο οποίος, όπως θα αποδεχτεί στη συνέχεια, εμφανίζει τη μικρότερη τιμή για το Παρίσι.



Σχήμα 79 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LASSO, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη του Παρισιού, **Σχήμα 79**, εμφανίζεται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν, να ακολουθούν και να μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η γραμμικότητα αυτή. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Στην περίπτωση του Παρισιού, θεωρείται ότι συναντάται μέτρια ισχύς, λόγω της μεσαίας κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, εντοπίζονται ελάχιστα μεμονωμένα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.

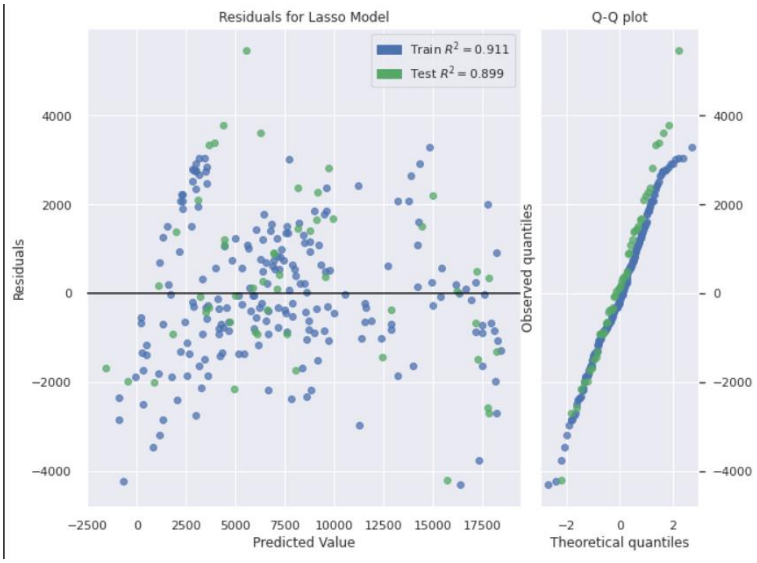


Σχήμα 80 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LASSO, Παρίσι, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα, **Σχήμα 80**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και απεικονίζονται και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος LASSO έχει αποδώσει ικανοποιητικά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια

ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως το διάστημα 12/11/2020 – 15/11/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα. Σε γενικές γραμμές, όμως, πρόκειται για μία αξιόλογη προσαρμογή του μοντέλου στις πραγματικές τιμές.

Ακολουθεί το διάγραμμα των υπολοίπων, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 81 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται επίσης, ότι αρκετά σημεία υπολοίπων βρίσκονται σε σημαντική απόσταση από την ευθεία y=0. Γεγονός το οποίο φανερώνει ότι συναντώνται αρκετά πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού. Επιπλέον, έχουν τυπωθεί και οι τιμές για το R^2, τόσο για τα δεδομένα εκπαίδευσης, όσο και για τα δεδομένα ελέγχου. Για το test set είχε προηγηθεί ο υπολογισμός του R^2 και μέσω της βιβλιοθήκης της Scikit-learn.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων εμφανίζουν δύο ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Στη συνέχεια, ακολουθεί η ανάλυση για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη του Παρισιού.

Όπως στην ανάλυση η οποία προηγήθηκε, έτσι ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

Μετρική Αξιολόγησης	**Αθήνα**
RMSE	4995.18 (cases per million)
R^2	0.847
EVS	0.850
MAE	3587.22 (cases per million)
MAPE	16.91%

MSE	23814970.45

Πίνακας 24 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LASSO, Αθήνα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε πολύ καλά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν σχετικά υψηλές τιμές. Βρίσκονται κοντά στο 0.85 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν υψηλή συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, βάσει του μεγέθους των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 83**). Ακόμη, το ποσοστό από το MAPE είναι σχετικά χαμηλό, βρίσκεται κοντά στο 17%, άρα θεωρείται καλό. Το MSE στην περίπτωση αυτή, είναι ένας 8ψήφιος αριθμός, άρα αριθμός μεγαλύτερος από εκείνο του μοντέλου του Παρισιού.

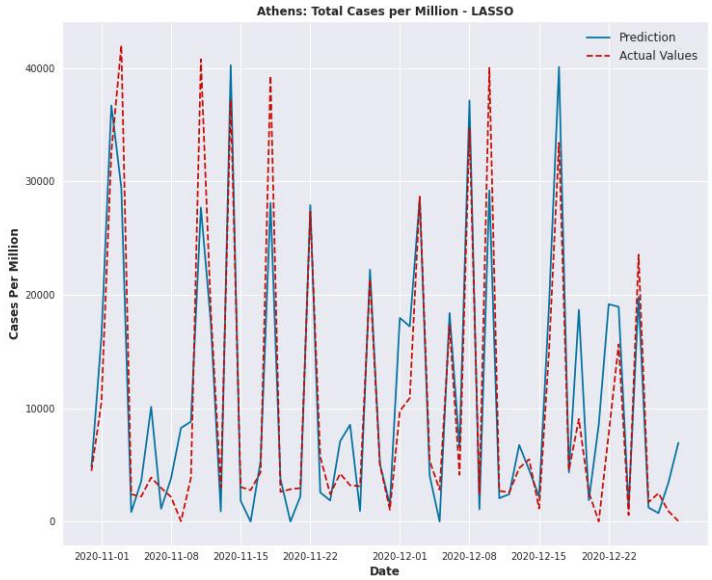
Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Αθήνας.

Σχήμα 82 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LASSO, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, δεν παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας. Παρατηρείται, δηλαδή, ότι από τα σημεία του διαγράμματος φαίνεται να μπορεί να περάσει μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι σχετικά ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, λαμβάνοντας, όμως, υπόψιν ότι τα σημεία του διαγράμματος δεν έχουν ένα συγκεκριμένο γραμμικό μοτίβο.

Στο παραπάνω διάγραμμα, **Σχήμα 83**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων, μαζί με τις πραγματικές τιμές των κρουσμάτων, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για τον αλγόριθμο SVR, για την πόλη της Αθήνας, είναι ότι εντοπίζει την τάση των πραγματικών τιμών, εν τούτοις αδυνατεί να προβλέψει πλήρως τόσο όλα τα μέγιστα, όσο και όλα τα ελάχιστα. Ακόμη, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα, όπως για παράδειγμα το διάστημα ανάμεσα στις 3/11/2020 έως 9/11/2020. Τα

παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, **Πίνακας 14**. Η προσαρμογή του μοντέλου στα πραγματικά δεδομένα δεν είναι άρτια, είναι απλώς ικανοποιητική.



Σχήμα 83 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LASSO, Αθήνα, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Σχήμα 84 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όμως, παρατηρείται ότι ένα πλήθος υπολοίπων απέχει σημαντικά από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και στις προβλεπόμενες τιμές, γίνεται αντιληπτό ότι συναντώνται ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο σχεδόν επικαλυπτόμενες ευθείες γραμμές και διαφοροποιούνται σε ένα μικρό εύρος στις ουρές τους. Θεωρείται, λοιπόν, ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

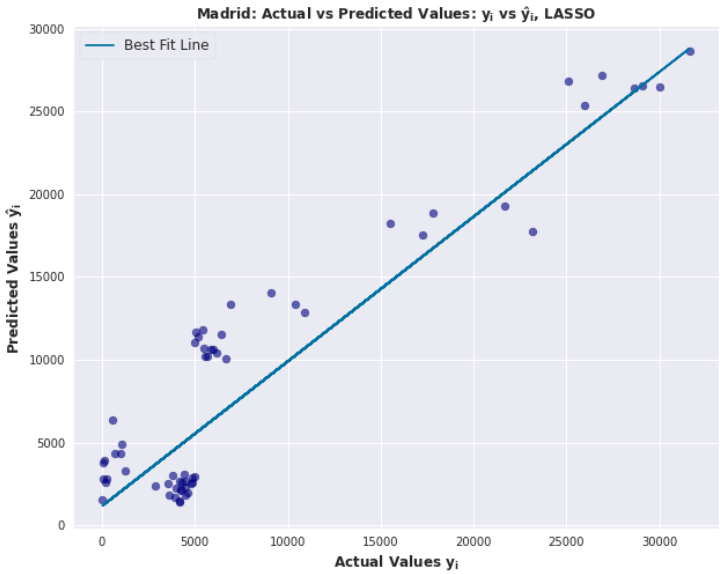
Τρίτη πόλη ανάλυσης αποτελεί η πόλη της Μαδρίτης. Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης κρουσμάτων.

Μετρική Αξιολόγησης	Μαδρίτη
RMSE	3413.64 (cases per million)
R^2	0.846
EVS	0.860
MAE	2987.26 (cases per million)
MAPE	4.32%
MSE	11652951.78

Πίνακας 25 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν τη Μαδρίτη, κυμαίνονται σε σχετικά ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν μεσαίες τιμές. Έχουν τιμή μεγαλύτερη από 0.80, και ως εκ τούτου θεωρείται ότι εμφανίζεται ισχυρό μέτρο επίδρασης των ανεξάρτητων μετρικών στην εξαρτημένη μεταβλητή (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 86**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, βρίσκεται κοντά στο 4%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται σχετικά καλό, αφού όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.

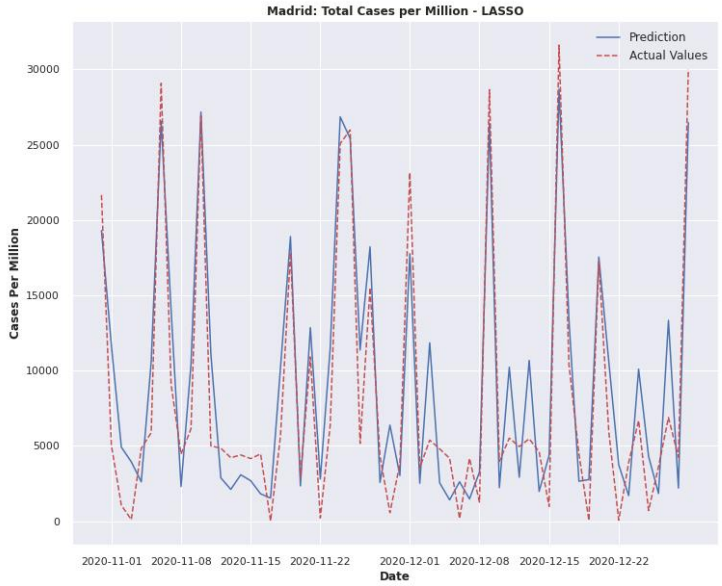


Σχήμα 85 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, παρατηρείται ότι δύναται η προσαρμογή ευθείας γραμμής στα δεδομένα. Παρουσιάζεται, όμως, ότι τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν μία «νοητή» ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί ορισμένες χαρακτηριστικές συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω

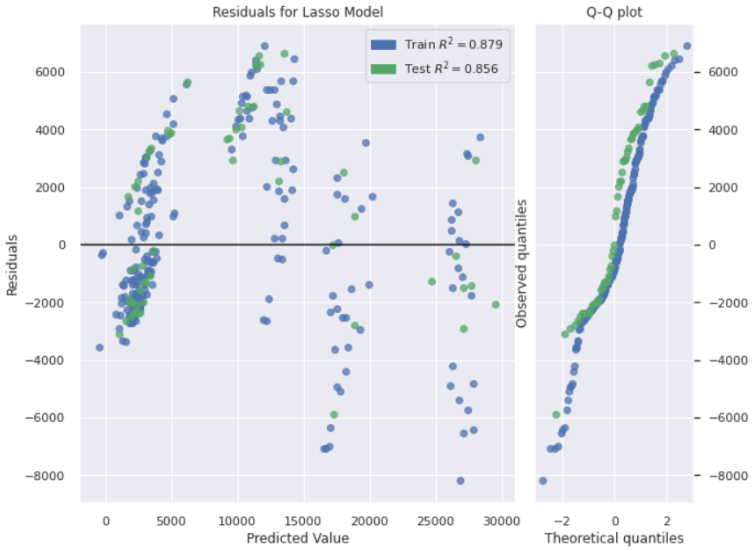
διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται ορισμένα ιδιάζοντα σημεία. Σε κάθε περίπτωση, χρειάζεται να ληφθεί υπόψιν ότι τα σημεία του διαγράμματος ακολουθούν ένα πιο «αφηρημένο» γραμμικό μοτίβο.

Στο διάγραμμα το οποίο ακολουθεί, **Σχήμα 86**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Μαδρίτης και για το LASSO, είναι ότι ο αλγόριθμος δεν εντοπίζει την τάση των πραγματικών τιμών με ιδιαίτερη επιτυχία, καθώς επίσης αδυνατεί να προβλέψει πλήρως την ακριβή τιμή κατά κύριο λόγο για μεσαίες τιμές, όπως για παράδειγμα η περίοδος 10/12/2020 με 12/12/2020. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές. Τα παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, όπως λ.χ. το MAE, **Πίνακας 25**.



Σχήμα 86 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 87 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων για το LASSO μοντέλο της Μαδρίτης, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0, αλλά σαν συστάδες. Επίσης, παρατηρείται ότι αρκετά από τα υπόλοιπα εμφανίζουν μεγάλες αρνητικές τιμές, μακριά από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές και λαμβάνοντας υπόψιν την υψηλή τιμή του MSE, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερο τμήμα τους. Έτσι, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Ακολουθεί η ανάλυση για το μοντέλο πρόβλεψης κρουσμάτων για την πόλη της Μόσχας. Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

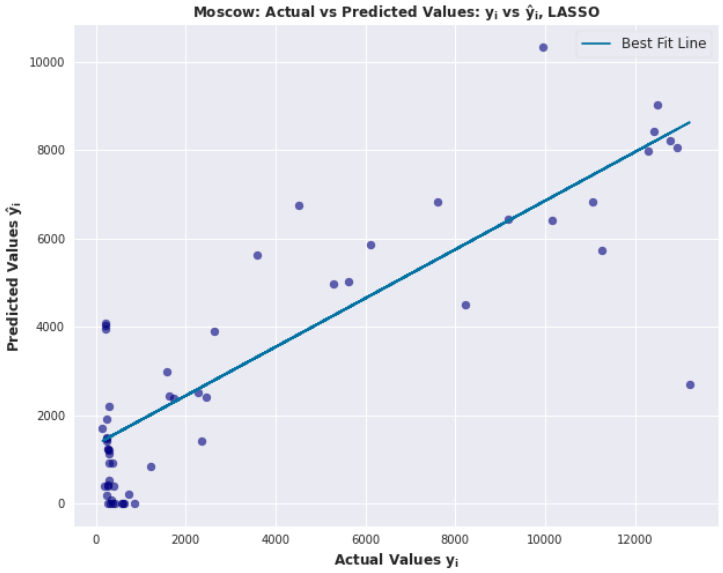
Μετρική Αξιολόγησης	**Μόσχα**
RMSE	3413.64 (cases per million)
R^2	0.846
EVS	0.860
MAE	χ (cases per million)
MAPE	2.51%
MSE	6410728.03

Πίνακας 26 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LASSO, Μόσχα, Ιδία Επεξεργασία*

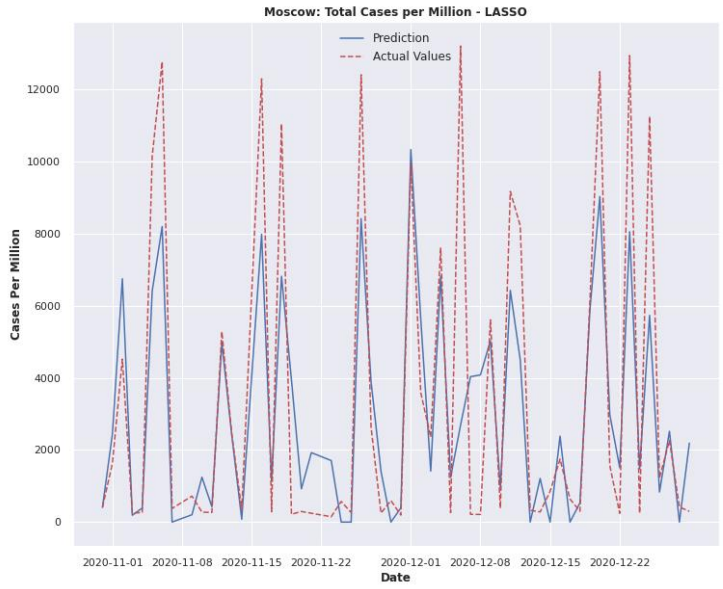
Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν αρκετά σχετικά υψηλή τιμή. Έχουν τιμή κοντά στο 0.85, και ως εκ τούτου οι ανεξάρτητες μεταβλητές έχουν ισχυρό μέτρο επίδρασης στην εξαρτημένη μεταβλητή (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 89**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, βρίσκεται κοντά στο 3%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Το MSE, όπως στην περίπτωση της Αθήνας και της Μαδρίτης, εμφανίζει 8ψήφιο αριθμό.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Μόσχας.

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μόσχας, παρατηρείται ύπαρξη γραμμικότητας. Αυτό το οποίο γίνεται αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων είναι ότι τα σημεία δεν βρίσκονται κοντά της γραμμικά για τις μικρότερες τιμές των πραγματικών μεταβλητών. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, κυρίως από ένα σημείο και μετά, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται μερικά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 88 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LASSO, Μόσχα, Ιδία Επεξεργασία*



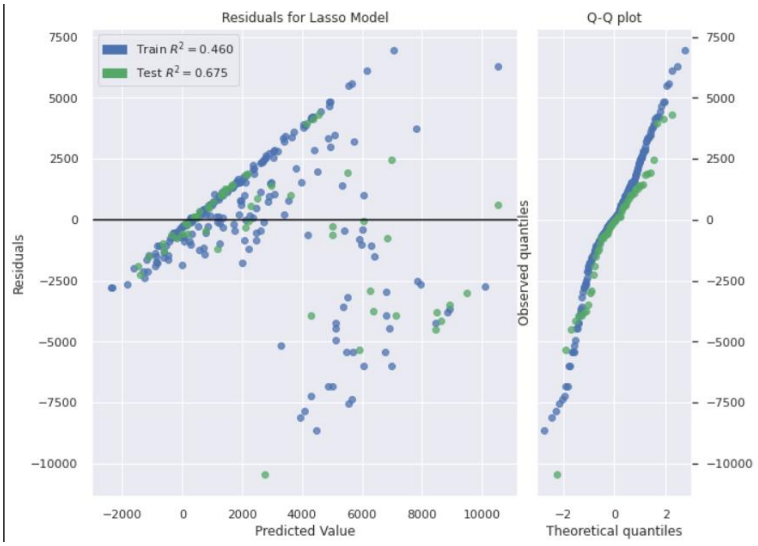
Σχήμα 89 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LASSO, Μόσχα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι δεν έχει γίνει πολύ ικανοποιητική προσαρμογή του αλγορίθμου στις πραγματικές τιμές των κρουσμάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει κατά κάποιον τρόπο, την τάση των κρουσμάτων, όμως, αδυνατεί πλήρως να υπολογίσει τις πραγματικές τιμές για τα μέγιστα, κυρίως. Συνεπώς, η προσαρμογή του μοντέλου στις πραγματικές τιμές δεν είναι αρκετά επιθυμητή.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Από το παρακάτω διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται ύπαρξη υπολοίπων σε ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Επίσης, εμφανίζονται σημεία με

σημαντικά μεγάλες αρνητικές τιμές. Τα σημεία αυτά πιθανότατα να μπορούσαν να χαρακτηριστούν ως ιδιάζοντα σημεία.



Σχήμα 90 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Μόσχα, Ιδία Επεξεργασία*

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία για τα δεδομένα ελέγχου και για τα δεδομένα εκπαίδευσης εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερό τους τμήμα. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Για να ολοκληρωθεί η ανάλυση των δημιουργηθέντων LASSO μοντέλων για πρόβλεψη κρουσμάτων στις μελετώμενες πόλεις, χρειάζεται να παρατεθούν τα αποτελέσματα τα οποία προκύπτουν για την Πράγα. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

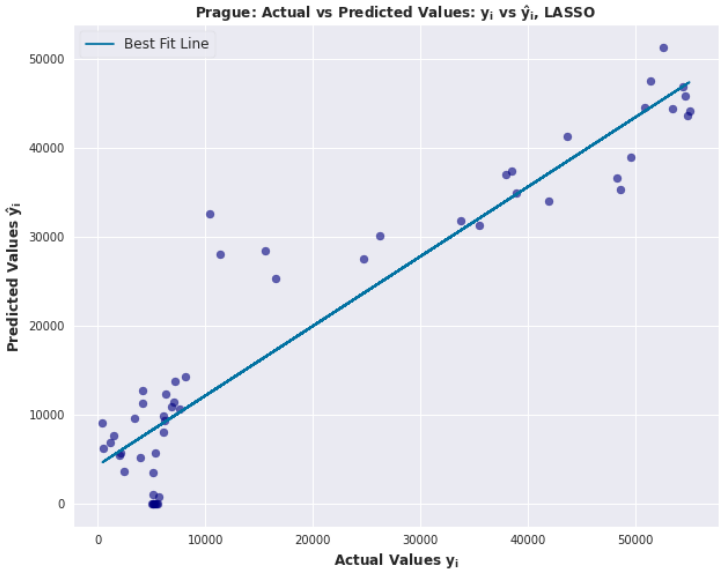
Μετρική Αξιολόγησης	**Πράγα**
RMSE	7565.90 (cases per million)
R^2	0.866
EVS	0.866
MAE	6323.24 (cases per million)
MAPE	1.39%
MSE	52507828.42

Πίνακας 27 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LASSO, Πράγα, Ιδία Επεξεργασία*

Βάσει όσων αναφέρθηκαν στην υποενότητα **4.2**, οι παραπάνω μετρικές αξιολόγησης είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.80, άρα η συσχέτιση των μεταβλητών θεωρείται υψηλή. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 92**). Ακόμη το ποσοστό από το MAPE είναι χαμηλό και μικρότερο από 10%, άρα θεωρείται πολύ καλό. Το MSE και για την πόλη της Πράγας, εμφανίζει το μεγαλύτερο 8ψήφιο αριθμό.

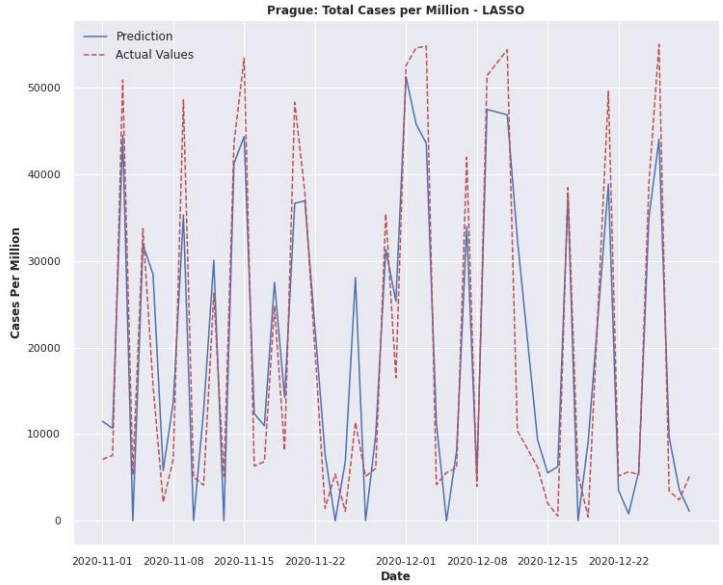
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Πράγας, **Σχήμα 91**, παρατηρείται ότι υπάρχει μία σχετική γραμμικότητα. Αυτό σημαίνει ότι στα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει

θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται ήπια ισχύς, λόγω της μικρής κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.



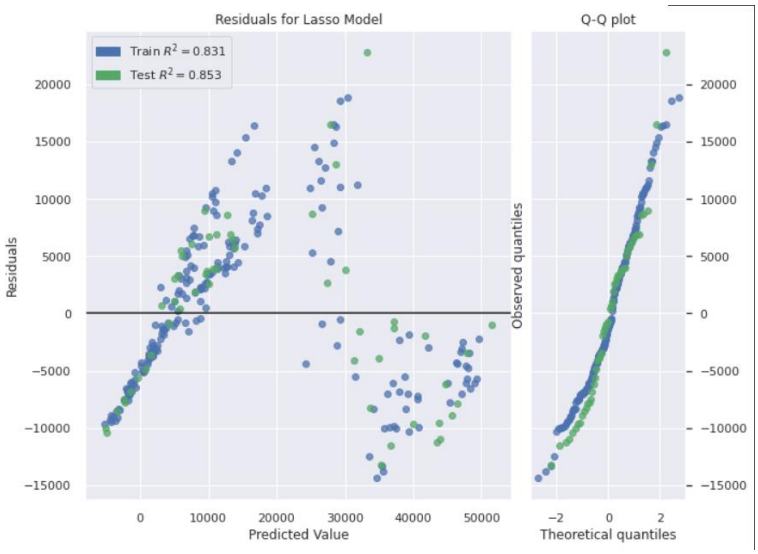
Σχήμα 91 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LASSO, Πράγα, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα, **Σχήμα 92**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος έχει αποδώσει σχετικά καλά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως για παράδειγμα 22/11/2020 – 25/11/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα. Υπάρχουν, λοιπόν, αρκετά σημεία τα οποία δεν έχει καταφέρει να προσαρμόσει στις πραγματικές τιμές των καταγεγραμμένων κρουσμάτων.



Σχήμα 92 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LASSO, Πράγα, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 93 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται ότι για σχετικά μικρές τιμές, δημιουργείται ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Αυτό το ευθύγραμμο τμήμα και η έντονη γραμμική συγκέντρωση των σημείων, μπορεί να ερμηνευτεί και από τη συστάδα η οποία παρατηρείται στο διάγραμμα από το **Σχήμα 91**. Ακόμη, συναντώνται σημεία τα οποία εμφανίζουν σημαντική απόσταση, μακριά από τον οριζόντιο άξονα.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία εμφανίζουν δύο σχεδόν ταυτιζόμενες ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος, κυρίως στην αριστερή ουρά. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τα αποτελέσματα των μοντέλων πρόβλεψης κρουσμάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Γ.

4.4.3.2 *Πρόβλεψη Θανάτων*

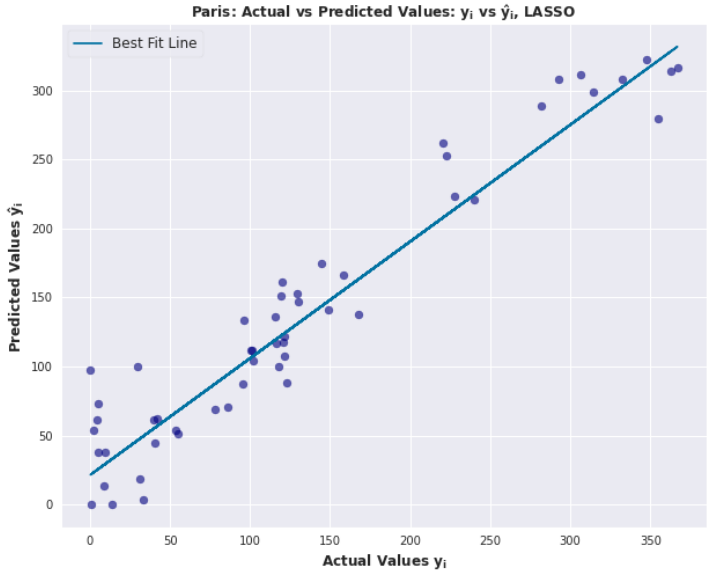
Το LASSO μοντέλο το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των θανάτων, είναι εκείνο για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	**Παρίσι**
RMSE	32.38 (deaths per million)
R^2	0.913
EVS	0.916
MAE	24.45 (deaths per million)
MAPE	6.20%
MSE	1040.20

Πίνακας 28 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LASSO, Παρίσι, Ιδία Επεξεργασία*

Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Πράγας, είναι ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.90, άρα η συσχέτιση των ανεξάρτητων μεταβλητών, με την εξαρτημένη μεταβλητή είναι αρκετά υψηλή.

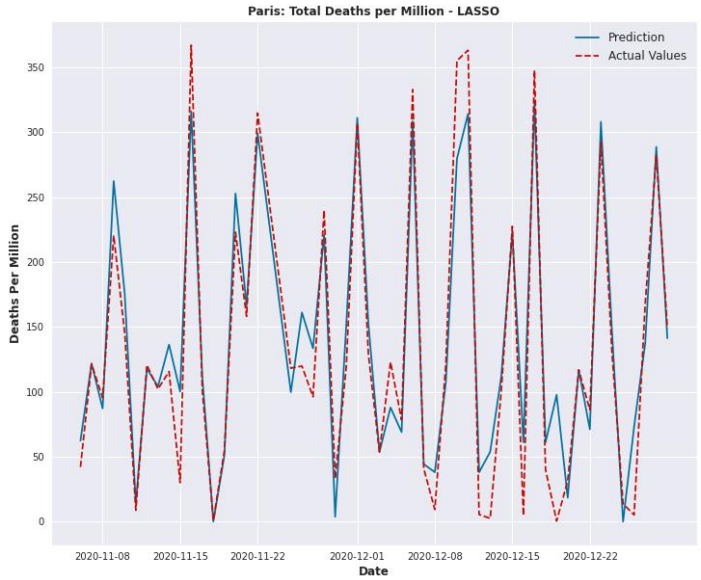
Δηλαδή, σημαίνει ότι 90% της μεταβολής της εξαρτώμενης μεταβλητής μπορεί να εξηγηθεί από τις ανεξάρτητες μεταβλητές. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές 350 κρούσματα (Βλέπε **Σχήμα 95Σχήμα 65**). Ακόμη το ποσοστό από το MAPE είναι μικρότερο από 10%, άρα θεωρείται πολύ καλό. Τέλος, όπως έχει αναφερθεί, δεν υπάρχει κάποια «σωστή» τιμή για το MSE. Ο κύριος σκοπός χρήσης του είναι η επιλογή ενός μοντέλου, έναντι κάποιου άλλου. Η τιμή του, στην περίπτωση αυτή, είναι 4ψήφια.



Σχήμα 94 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LASSO, Παρίσι, Ιδία Επεξεργασία*

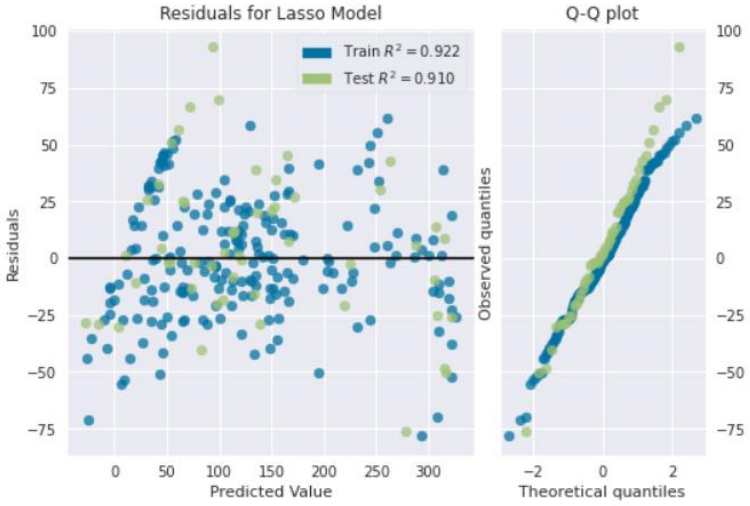
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για το Παρίσι, εμφανίζεται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος σχηματίζουν και ακολουθούν, όπως επίσης μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων μπορούν εύκολα να γίνουν όσα προαναφέρθηκαν αντιληπτά. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται μέτρια ισχύς, λόγω της μεσαίας κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από μελέτη του παραπάνω διαγράμματος, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία απέχουν από τα υπόλοιπα και τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.

Στο παρακάτω διάγραμμα, **Σχήμα 95**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και απεικονίζονται και οι πραγματικές τιμές των θανάτων για την πόλη του Παρισιού και για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος έχει αποδώσει καλά, όχι όμως πλήρως ικανοποιητικά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως το διάστημα 23/12/2020 – 25/12/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα. Έτσι, ορισμένα προβλεπόμενα σημεία φαίνεται ότι δεν προσαρμόζονται πλήρως στις πραγματικές τιμές των θανάτων.



Σχήμα 95 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LASSO, Παρίσι, Ιδία Επεξεργασία*

Στο διάγραμμα των υπολοίπων, παρουσιάζονται τα υπόλοιπα για το συγκεκριμένο μοντέλο και απεικονίζονται τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 96 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων, φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ένα γραμμικό μοντέλο μπορεί να περιγράψει συνεπώς τα δεδομένα αυτά. Επίσης, συναντώνται ορισμένα σημεία σε μακρινή απόσταση από τον οριζόντιο άξονα, γεγονός το οποίο φανερώνει ότι συναντώνται αρκετά πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων εμφανίζουν μία σχεδόν ευθεία γραμμή και διαφοροποιούνται σε ένα μικρό εύρος, στην δεξιά ουρά. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Στη συνέχεια, ακολουθεί η ανάλυση μοντέλων για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη της Μαδρίτης.

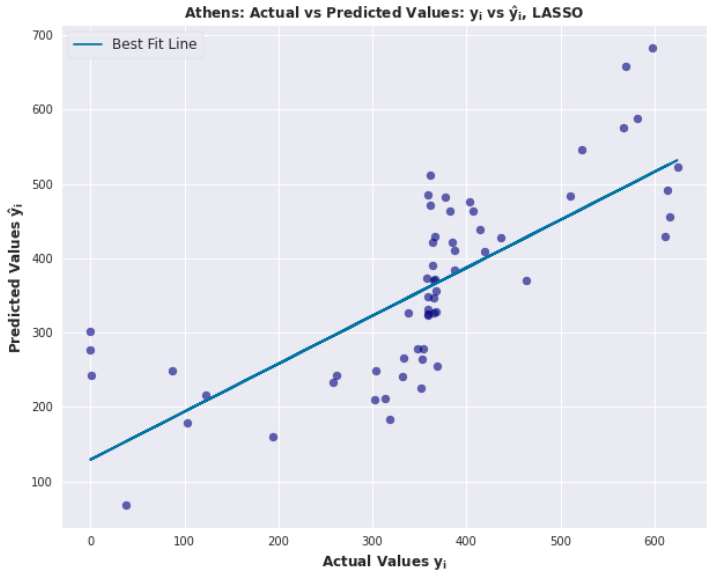
Ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

Μετρική Αξιολόγησης	Αθήνα
RMSE	97.72 (deaths per million)
R^2	0.554
EVS	0.555
MAE	72.95 (deaths per million)
MAPE	51.72%
MSE	9549.92

Πίνακας 29 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LASSO, Αθήνα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε σχετικά χαμηλά πλαίσια. Η R^2 και EVS εμφανίζουν χαμηλές τιμές. Βρίσκονται κοντά στο 0.50 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν αδύναμη συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο βάσει του μεγέθους των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 98**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά υψηλό, ξεπερνάει το 50%, άρα δεν θεωρείται καλό.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Αθήνας.



Σχήμα 97 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LASSO, Αθήνα, Ιδία Επεξεργασία*

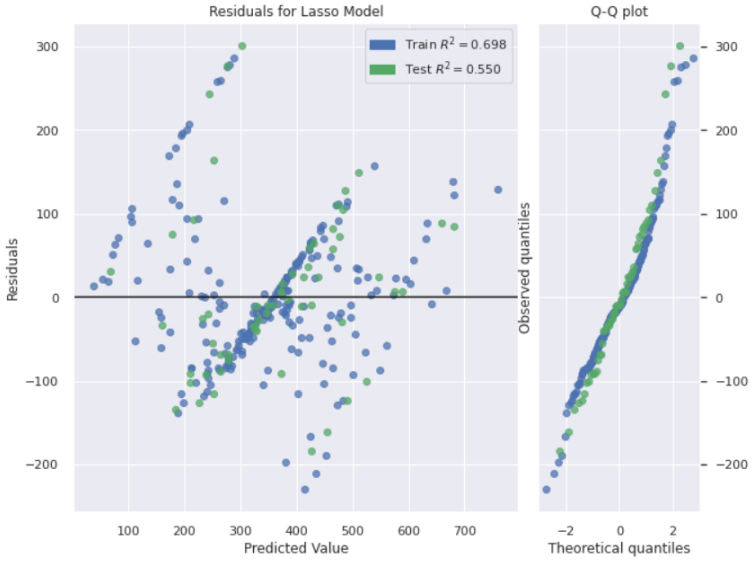
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, παρατηρείται σχετική ύπαρξη γραμμικότητας. Παρατηρείται, δηλαδή, ότι τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν ξεκάθαρα μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργούνται συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, καθώς παρατηρούνται σημεία τα οποία απέχουν τόσο από την ευθεία ελαχίστων τετραγώνων όσο και από τις συγκεντρώσεις των υπόλοιπων σημείων.

Στο παρακάτω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων, μαζί με τις πραγματικές τους τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές

γραμμές, εκείνο το οποίο παρατηρείται, για την πόλη της Αθήνας, είναι ότι το μοντέλο εντοπίζει την τάση των πραγματικών τιμών. Εν τούτοις, αδυνατεί να προβλέψει πλήρως τις ακριβείς τιμές, για τα περισσότερα μέγιστα και για τα περισσότερα ελάχιστα. Έτσι, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα. Σε ορισμένες περιπτώσεις φαίνεται να υπερεκτιμάει και σε άλλες φαίνεται να υποτιμάει τον προβλεπόμενο αριθμό κρουσμάτων. Τα παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, **Πίνακας 29**.

Σχήμα 98 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LASSO, Αθήνα, Ιδία Επεξεργασία*

Στο διάγραμμα από το **Σχήμα 99** παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 99 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LASSO, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ακόμη, παρατηρείται ότι δημιουργείται ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Επιπλέον, εμφανίζονται ορισμένα σημεία μακριά από τη συγκέντρωση των υπολοίπων και μακριά από τον άξονα y. Πρόκειται για πιθανά ιδιάζοντα σημεία.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο επικαλυπτόμενες ευθείες γραμμές, άρα ταυτίζονται, εκτός από τις ουρές τους. Ως εκ τούτου, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

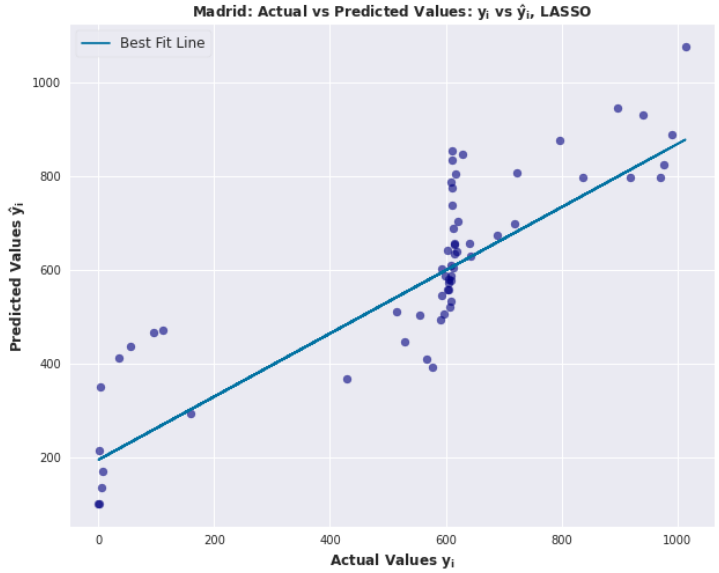
Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης θανάτων για την πόλη της Μαδρίτης.

Μετρική Αξιολόγησης	Μαδρίτη
RMSE	147.32 (deaths per million)
R^2	0.703
EVS	0.734
MAE	107.47 (deaths per million)
MAPE	19.83%
MSE	21702.64

Πίνακας 30 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την Μαδρίτη, κυμαίνονται σε σχετικά ικανοποιητικά πλαίσια. Η R^2 και EVS εμφανίζουν μεσαίες τιμές. Έχουν τιμή μεγαλύτερη από 0.70, και ως εκ τούτου θεωρείται ότι εμφανίζεται μέτριο μέτρο επίδρασης, δηλαδή ότι η συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη είναι μέτρια (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 101**). Ακόμη, το ποσοστό από το MAPE είναι μεσαίο, βρίσκεται κοντά στο 20%, άρα, θεωρείται καλό. Να σημειωθεί ότι το MSE του εν λόγω μοντέλου, είναι 5ψήφιος αριθμός.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.

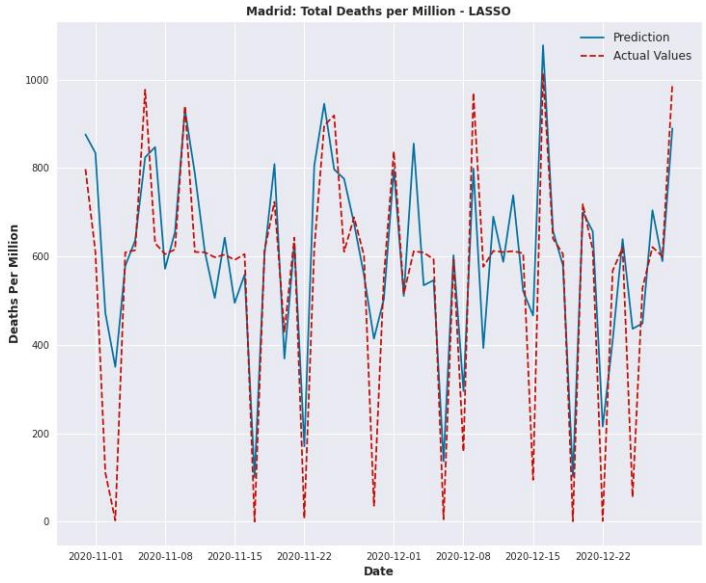


Σχήμα 100 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μαδρίτης , παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας ανάμεσα στα σημεία. Παρατηρείται ότι η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί ορισμένες συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη.

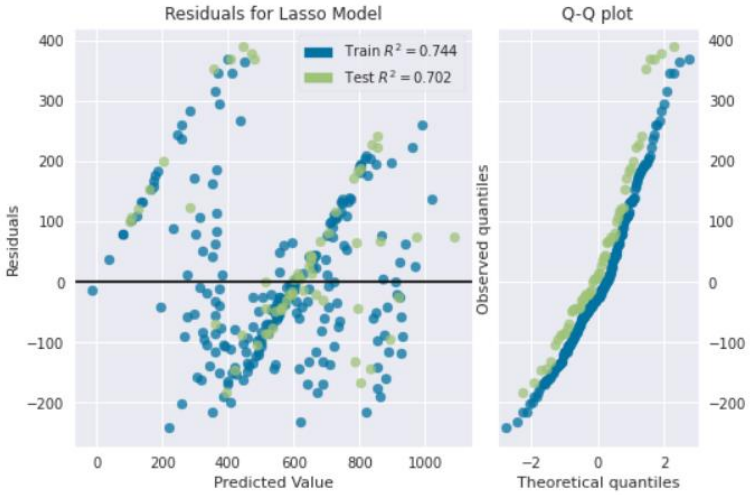
Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται ορισμένα ιδιάζοντα σημεία, τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία.

Στο παρακάτω διάγραμμα, **Σχήμα 101**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Πράγας και για το μοντέλο LASSO, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών σε ένα μέτριο επίπεδο, εν τούτοις αδυνατεί να προβλέψει πλήρως την ακριβή τιμή τόσο για αρκετά μέγιστα, όσο και για αρκετά ελάχιστα. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές. Χαρακτηριστικό είναι το παράδειγμα για τις ημερομηνίες 10/11/2020 έως και περίπου 15/11/2020. Τα παραπάνω, επιβεβαιώνονται από τις σχετικά χαμηλές τιμές των μετρικών, όπως λ.χ. το MAE, **Πίνακας 30**.



Σχήμα 101 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 102 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, παρατηρείται ότι δημιουργείται ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές, αυτό το οποίο μπορεί να γίνει θεωρηθεί είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα υπόλοιπα για τα δεδομένα εκπαίδευσης και για τα δεδομένα ελέγχου εμφανίζουν δύο ευθείες γραμμές, οι οποίες δεν επικαλύπτονται, όμως ακολουθούν το ίδιο μοτίβο, την ίδια τάση. Έτσι, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τέταρτη πόλη ανάλυσης αποτελεί η πόλη της Μόσχας. Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

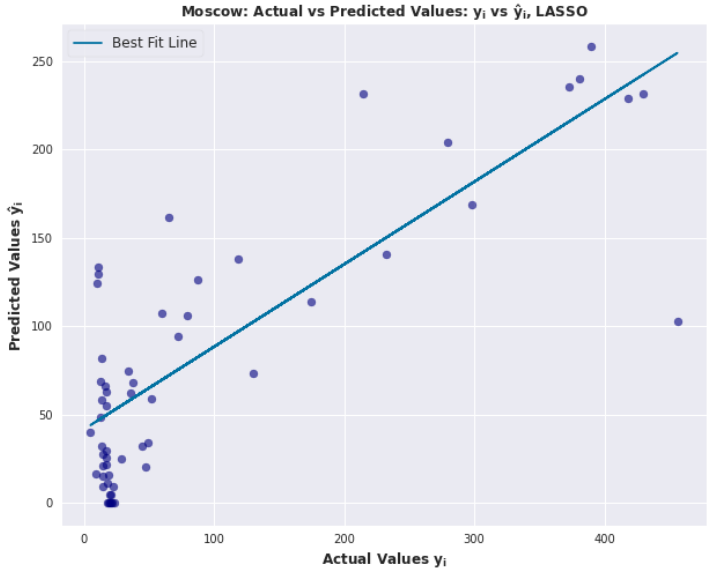
Μετρική Αξιολόγησης	**Μόσχα**
RMSE	82.69 (deaths per million)
R^2	0.593
EVS	0.600
MAE	54.86 (deaths per million)
MAPE	1.72%
MSE	6636.53

Πίνακας 31 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LASSO, Μόσχα, Ιδία Επεξεργασία*

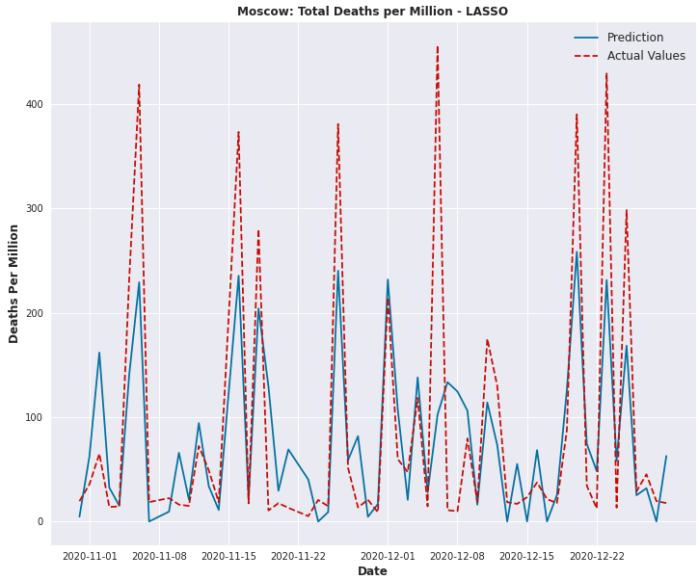
Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε μεσαία επίπεδα. Η R^2 και EVS εμφανίζουν αρκετά χαμηλή τιμή. Η τιμή για το R^2 βρίσκεται κοντά στο 0.60 και ως εκ τούτου οι ανεξάρτητες μεταβλητές έχουν μέτριο μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 104**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, βρίσκεται κοντά στο 2% και θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Το MSE, όπως στην περίπτωση της Αθήνας και του Παρισιού, εμφανίζει 4ψήφιο αριθμό.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Μόσχας.

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μόσχας, δεν παρατηρείται ιδιαίτερη γραμμικότητας. Αυτό το οποίο γίνεται αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων είναι ότι τα σημεία δεν βρίσκονται κοντά της με γραμμικό τρόπο. Έτσι, μπορεί να δικαιολογηθεί η «μεσαία» τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι αρκετά ήπια. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται μερικά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τις συγκεντρώσεις των υπόλοιπων σημείων και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 103 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LASSO, Μόσχα, Ιδία Επεξεργασία*



Σχήμα 104 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LASSO, Μόσχα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι έχει γίνει μία μέτρια προσαρμογή του αλγορίθμου στις πραγματικές τιμές των θανάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει κατά κάποιον τρόπο, την τάση των κρουσμάτων, όμως, αδυνατεί πλήρως να υπολογίσει τις πραγματικές τιμές στην πλειονότητα της χρονικής περιόδου μελέτης.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Από το διάγραμμα των υπολοίπων, **Σχήμα 105**, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όμως, παρατηρείται ότι η μεγαλύτερη συγκέντρωση δημιουργεί ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Ακόμη, παρατηρούνται ορισμένα σημεία με αρκετά μεγάλες τιμές, κυρίως αρνητικές, μακριά από την ευθεία y=0. Έτσι, ίσως πρόκειται για ιδιάζοντα σημεία, καθώς πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των υπόλοιπων σημείων.

Τέλος, από το διάγραμμα Q-Q, **Σχήμα 105**, παρατηρείται ότι τα σημεία των υπολοίπων για τα δύο σετ εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερό τους. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Σχήμα 105 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LASSO, Μόσχα, Ιδία Επεξεργασία*

Στο τελικό μοντέλο ανάλυσης για πρόβλεψη θανάτων με τον αλγόριθμο LASSO, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Πράγας.

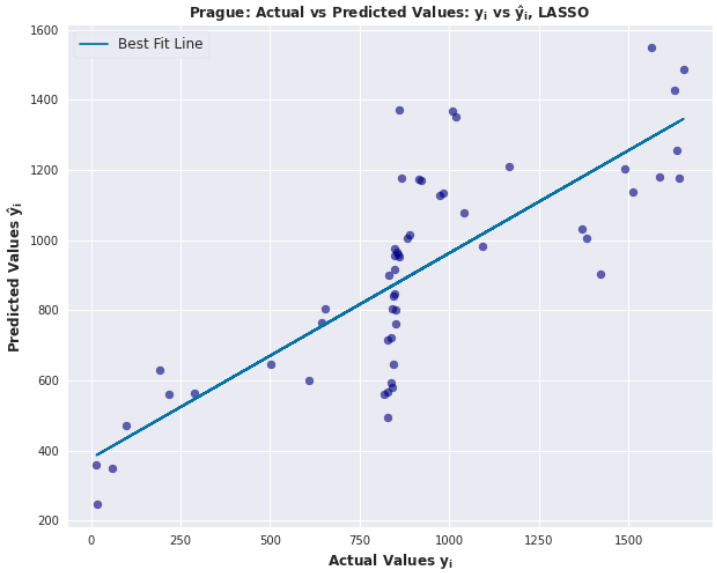
Μετρική Αξιολόγησης	**Πράγα**
RMSE	252.52 (deaths per million)
R^2	0.627
EVS	0.628
MAE	211.63 (deaths per million)
MAPE	1.04%
MSE	63765.56

Πίνακας 32 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LASSO, Πράγα, Ιδία Επεξεργασία*

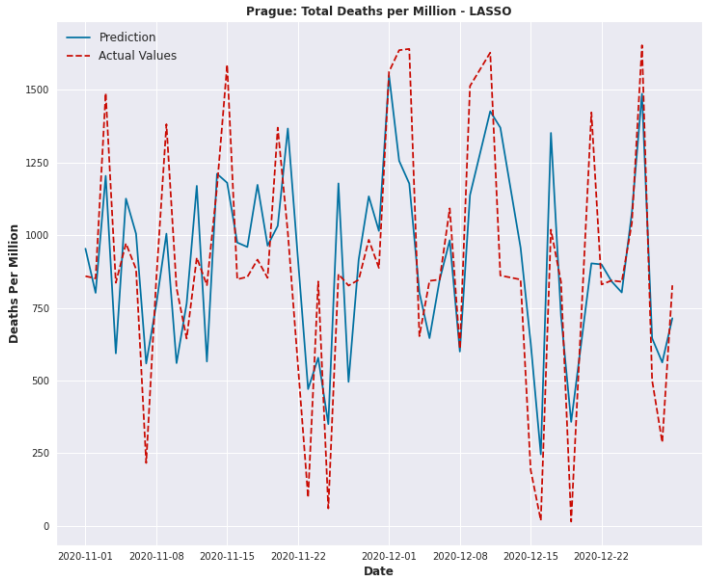
Βάσει όσων αναφέρθηκαν στην υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Πράγας, κυμαίνονται σε μέτρια πλαίσια. Η τιμή της R^2 και της EVS είναι σχετικά καλές. Η τιμή της R^2 βρίσκεται κοντά στο 0.63, άρα η συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη θεωρείται μέτρια. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 107**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό και μικρότερο από 2%, άρα θεωρείται πολύ καλό. Το MSE εμφανίζει το μεγαλύτερο 5ψήφιο αριθμό.

Από το διάγραμμα διασποράς, το οποίο ακολουθεί, **Σχήμα 106**, ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Πράγας, παρατηρείται ότι υπάρχει γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται αρκετά ήπια ισχύς, λόγω της μικρής κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται λίγα σημεία τα οποία

θα μπορούσαν να χαρακτηρισθούν έτσι, καθώς απέχουν αρκετά από την ευθεία ελαχίστων τετραγώνων και από συγκεντρώσεις άλλων σημείων.



Σχήμα 106 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LASSO, Πράγα, Ιδία Επεξεργασία*



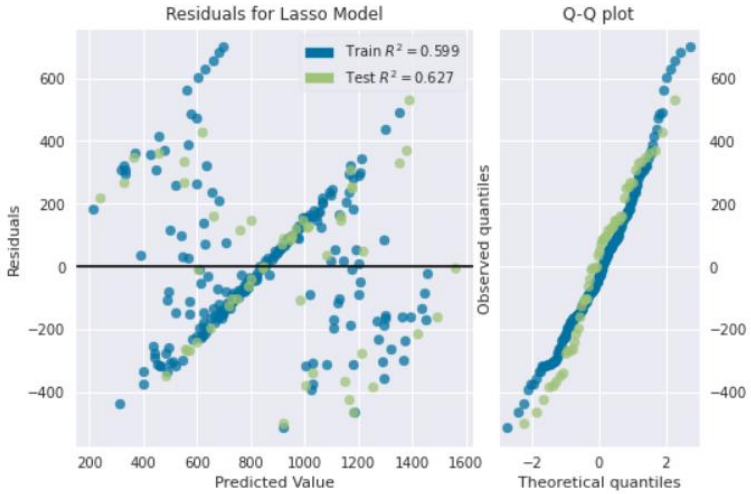
Σχήμα 107 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LASSO, Πράγα, Ιδία Επεξεργασία*

Στο διάγραμμα των προβλεπόμενων τιμών, **Σχήμα 107**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι η απόδοση του αλγορίθμου δεν είναι πολύ ικανοποιητική. Δεν είναι ικανός να ακολουθήσει πιστά τη ροή και την τάση των πραγματικών τιμών. Ούτε μπορεί να προβλέψει με πλήρη επιτυχία όλα τα μέγιστα, αλλά και όλα τα ελάχιστα. Άρα, αρκετά σημεία δεν έχουν προσαρμοστεί στις πραγματικές τιμές των θανάτων.

Στη συνέχεια, παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.

Από το διάγραμμα των υπολοίπων φαίνεται ότι ένα πλήθος τους βρίσκεται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ακόμη μία φορά παρατηρείται η συγκέντρωση αρκετών σημείων σε ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Τέλος,

εμφανίζονται σημεία τα οποία απέχουν από τον εν λόγω άξονα και τα οποία θα μπορούσε κάποιος να τα χαρακτηρίσει ως ιδιάζοντα σημεία.



Σχήμα 108 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, , LASSO, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων, εμφανίζουν δύο σχεδόν ταυτιζόμενες ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος, στην ουρά τους. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τα αποτελέσματα των μοντέλων πρόβλεψης θανάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Γ.

4.4.4 GPR

Τέταρτο μοντέλο το οποίο εφαρμόστηκε στα υπάρχοντα δεδομένα για τις εννέα διαφορετικές πόλεις, είναι το GPR. Στις επόμενες σελίδες ακολουθούν τα αποτελέσματα του αλγορίθμου για την Αθήνα, τη Μαδρίτη, τη Μόσχα, το Παρίσι και την Πράγα.

4.4.4.1 Πρόβλεψη Κρουσμάτων

Εκείνο το μοντέλο το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των κρουσμάτων με το GPR, είναι εκείνο για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	**Παρίσι**
RMSE	1894.76 (cases per million)
R^2	0.900
EVS	0.903
MAE	1469.57 (cases per million)
MAPE	3.05%
MSE	3590121.18

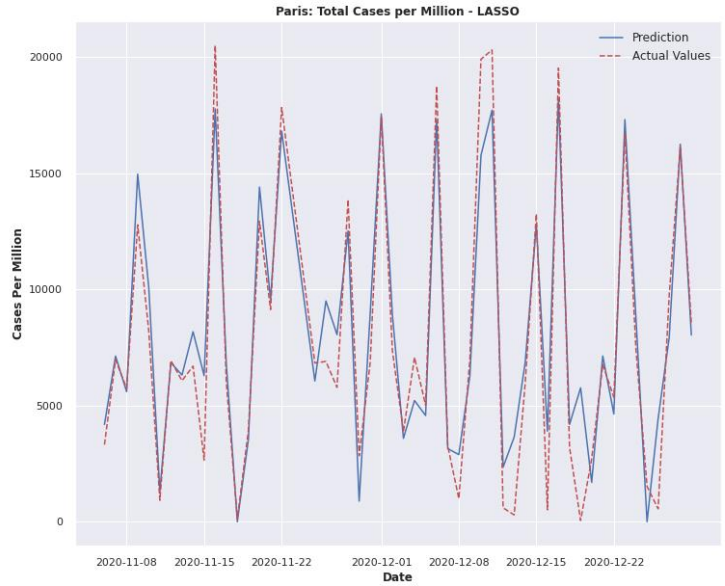
Πίνακας 33 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, GPR, Παρίσι, Ιδία Επεξεργασία*

Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Παρισιού, είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS είναι ίσες με 0.90, άρα πρόκειται για υψηλή συσχέτιση μεταβλητών. Δηλαδή, σημαίνει ότι 90% της μεταβολής της εξαρτημένης μεταβλητής μπορεί να εξηγηθεί από τις ανεξάρτητες μεταβλητές. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές τα περίπου 20000 κρούσματα (Βλέπε **Σχήμα**

110). Ακόμη το ποσοστό από το MAPE είναι χαμηλό και μικρότερο από 10%, άρα θεωρείται πολύ καλό. Τέλος, όπως έχει αναφερθεί και σε προηγούμενα, δεν υπάρχει κάποια «σωστή» τιμή για το MSE. Ο κύριος σκοπός χρήσης του είναι η επιλογή μίας πρόβλεψης ενός μοντέλου, έναντι κάποιας άλλης. Στην περίπτωση αυτή, πρόκειται για έναν επταψήφιο αριθμό, ο οποίος, όπως θα αποδεχτεί στη συνέχεια, εμφανίζει τη μικρότερη τιμή για το Παρίσι.

Σχήμα 109 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, GPR, Παρίσι, Ιδία Επεξεργασία*

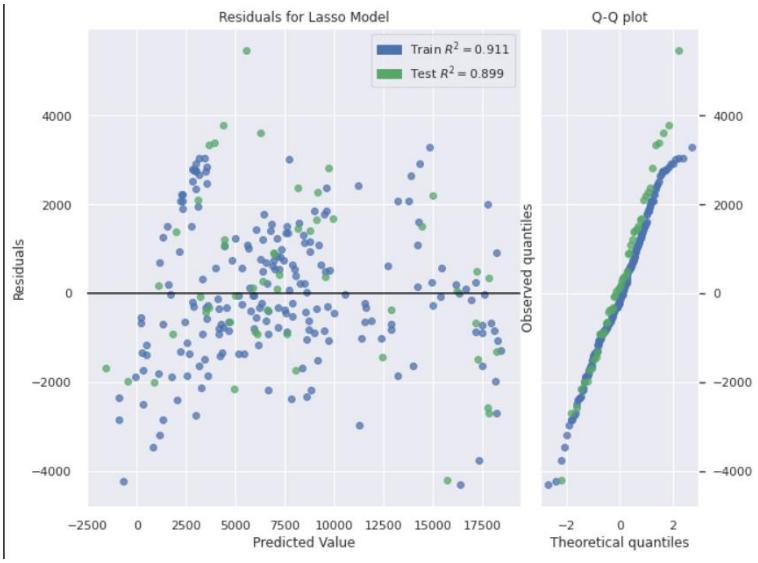
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη του Παρισιού, **Σχήμα 109**, εμφανίζεται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν, να ακολουθούν και να μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η γραμμικότητα αυτή. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Στην περίπτωση του Παρισιού, θεωρείται ότι συναντάται μέτρια ισχύς, λόγω της μεσαίας κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, εντοπίζονται ελάχιστα μεμονωμένα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.



Σχήμα 110 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, GPR, Παρίσι, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα, **Σχήμα 110**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και απεικονίζονται και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος GPR έχει αποδώσει ικανοποιητικά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως το διάστημα 10/11/2020 – 15/11/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα. Σε γενικές γραμμές, όμως, πρόκειται για μία αξιόλογη προσαρμογή του μοντέλου στις πραγματικές τιμές.

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 111 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται επίσης, ότι ορισμένα σημεία υπολοίπων βρίσκονται σε σημαντική απόσταση από την ευθεία y=0. Γεγονός το οποίο φανερώνει ότι συναντώνται αρκετά πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων εμφανίζουν δύο ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος, των δεξιών ουρών τους. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Στη συνέχεια, ακολουθεί η ανάλυση για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη του Παρισιού.

Όπως στην ανάλυση η οποία προηγήθηκε, έτσι ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

Μετρική Αξιολόγησης	**Αθήνα**
RMSE	5016.70 (cases per million)
R^2	0.845
EVS	0.849
MAE	3588.66 (cases per million)

MAPE	17.49%
MSE	24071027.62

Πίνακας 34 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, GPR, Αθήνα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε πολύ καλά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν υψηλές τιμές. Βρίσκονται κοντά στο 0.85 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν υψηλή συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, βάσει του μεγέθους των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 113**). Ακόμη, το ποσοστό από το MAPE είναι σχετικά χαμηλό, βρίσκεται κοντά στο 17%, άρα θεωρείται καλό. Το MSE στην περίπτωση αυτή, είναι ένας 8ψήφιος αριθμός, άρα αριθμός μεγαλύτερος από εκείνο του μοντέλου του Παρισιού.

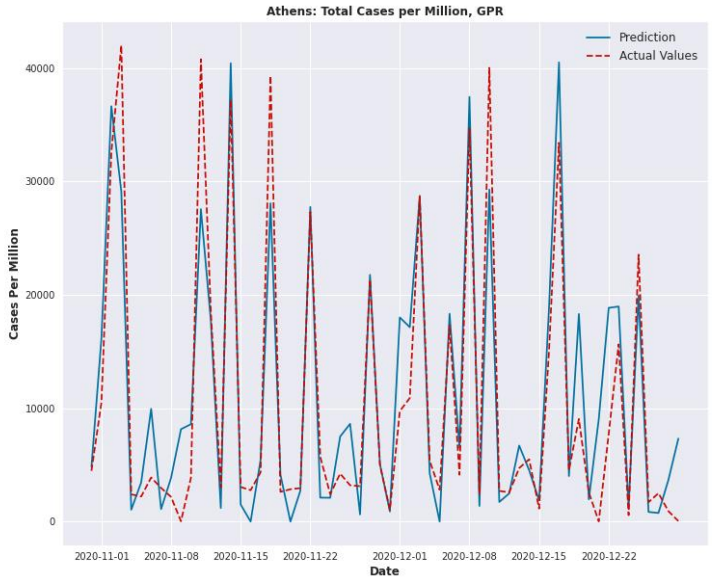
Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Αθήνας.

Σχήμα 112 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, GPR, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, δεν παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας. Παρατηρείται, δηλαδή, ότι από τα σημεία του διαγράμματος φαίνεται να μπορεί να περάσει μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, λαμβάνοντας, όμως, υπόψιν ότι τα σημεία του διαγράμματος δεν έχουν ένα συγκεκριμένο γραμμικό μοτίβο.

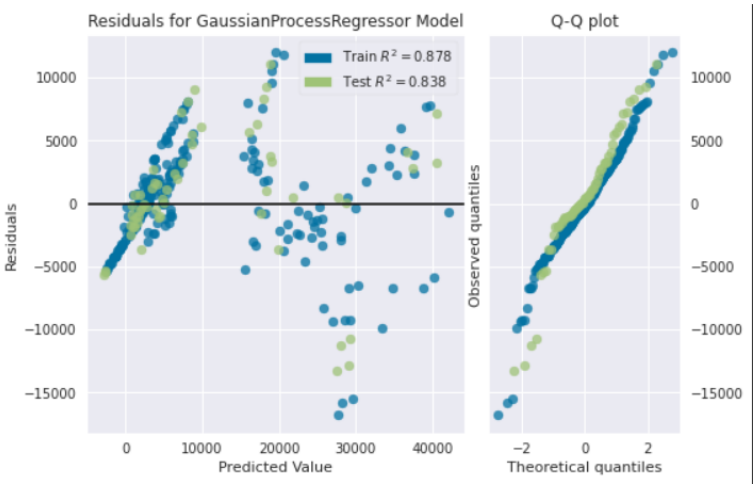
Ακολουθεί διάγραμμα στο οποίο απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων, μαζί με τις πραγματικές τιμές των κρουσμάτων, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για τον αλγόριθμο SVR, για την πόλη της Αθήνας, είναι ότι εντοπίζει την τάση των πραγματικών τιμών, εν τούτοις αδυνατεί να προβλέψει πλήρως τόσο όλα τα μέγιστα, όσο και όλα τα ελάχιστα. Ακόμη, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα, όπως για παράδειγμα το διάστημα ανάμεσα στις 3/11/2020 έως 9/11/2020. Τα

παραπάνω, επιβεβαιώνονται από τις τιμές των μετρικών, **Πίνακας 34**. Η προσαρμογή του μοντέλου στα πραγματικά δεδομένα δεν είναι άρτια, είναι απλώς ικανοποιητική.



Σχήμα 113 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, GPR, Αθήνα, Ιδία Επεξεργασία*

Παρατίθεται το διάγραμμα υπολοίπων μαζί με το διάγραμμα Q-Q.



Σχήμα 114 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, για σχετικά μικρές τιμές, παρατηρείται ότι υπάρχει συγκέντρωση σημείων σε ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Επιπροσθέτως, παρατηρείται ότι ένα μέρος υπολοίπων απέχει σημαντικά από την ευθεία y=0. Έτσι, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και στις προβλεπόμενες τιμές, γίνεται αντιληπτό ότι συναντώνται πιθανά ιδιάζοντα σημεία. Πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των παρατηρούμενων σημείων.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο σχεδόν επικαλυπτόμενες ευθείες γραμμές και διαφοροποιούνται σε ένα μικρό εύρος στις ουρές τους. Θεωρείται, λοιπόν, ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

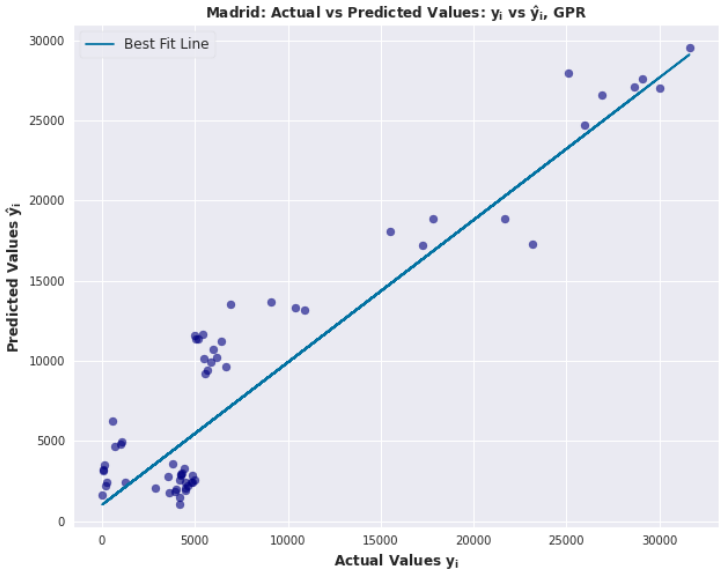
Τρίτη πόλη ανάλυσης αποτελεί η πόλη της Μαδρίτης. Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης κρουσμάτων.

Μετρική Αξιολόγησης	Μαδρίτη
RMSE	3316.62 (cases per million)
R^2	0.855
EVS	0.867
MAE	2867.93 (cases per million)
MAPE	4.27%
MSE	10999964.19

Πίνακας 35 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, GPR, Μαδρίτη, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν τη Μαδρίτη, κυμαίνονται σε σχετικά ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν σχετικά υψηλές τιμές. Έχουν τιμή μεγαλύτερη από 0.80, και ως εκ τούτου θεωρείται ότι εμφανίζεται ισχυρό μέτρο επίδρασης των ανεξάρτητων μετρικών στην εξαρτημένη μεταβλητή (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 116**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, βρίσκεται κοντά στο 4%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται σχετικά καλό, αφού όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Το MSE σε αυτήν την περίπτωση είναι ένας 8ψήφιος αριθμός, μικρότερος από εκείνο του μοντέλου πρόβλεψης για την πόλη της Αθήνας.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.

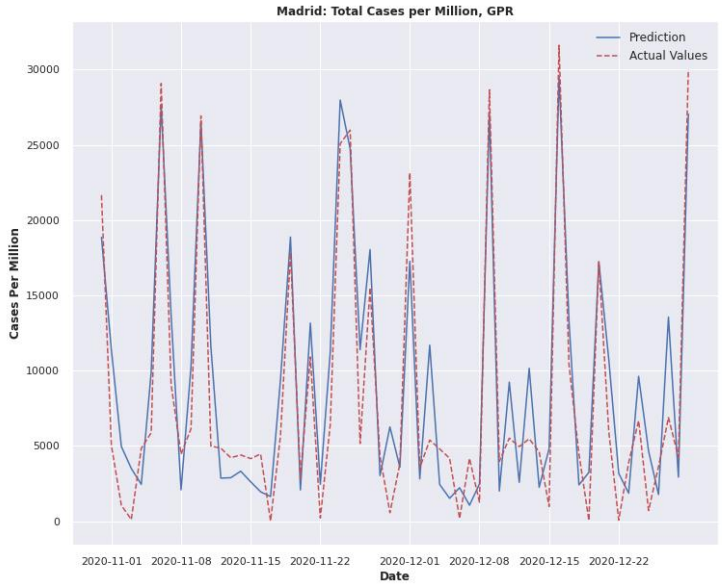


Σχήμα 115 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, GPR, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, παρατηρείται ότι υπάρχει μία σχετική γραμμικότητα. Δηλαδή, δύναται η προσαρμογή ευθείας γραμμής στα δεδομένα. Φαίνεται, όμως, ότι τα σημεία του διαγράμματος δεν σχηματίζουν μία «νοητή» ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί ορισμένες χαρακτηριστικές συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια.

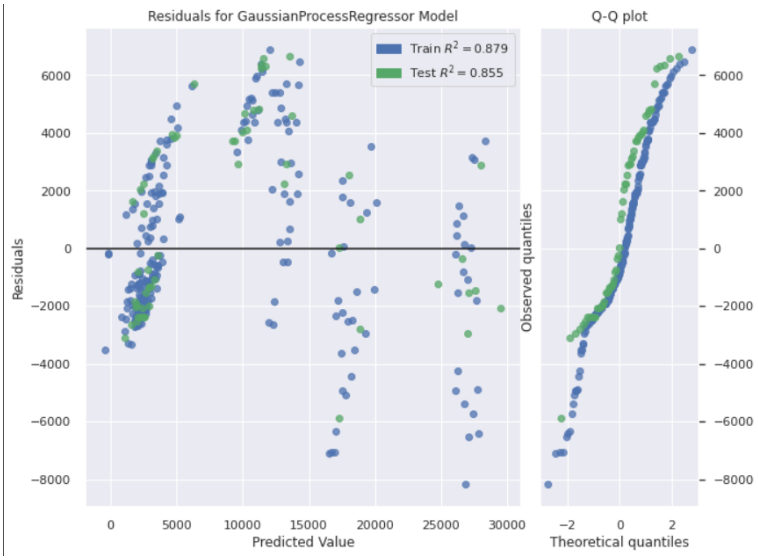
Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται ορισμένα ιδιάζοντα σημεία. Σε κάθε περίπτωση, χρειάζεται να ληφθεί υπόψιν ότι τα σημεία του διαγράμματος ακολουθούν ένα πιο «αφηρημένο» γραμμικό μοτίβο.

Στο διάγραμμα το οποίο ακολουθεί, **Σχήμα 116**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Μαδρίτης και για το GPR, είναι ότι ο αλγόριθμος δεν εντοπίζει την τάση των πραγματικών τιμών με ιδιαίτερη επιτυχία, καθώς επίσης αδυνατεί να προβλέψει πλήρως την ακριβή τιμή κατά κύριο λόγο για μεσαίες τιμές, όπως για παράδειγμα η περίοδος 10/12/2020 με 12/12/2020. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές. Άρα, ο αλγόριθμος για την πόλη αυτή, δεν φαίνεται να έχει κάνει πολύ καλή προσαρμογή.



Σχήμα 116 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, GPR, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 117 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων για το GPR μοντέλο της Μαδρίτης, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Η κατανομή τους αυτή, φαίνεται να είναι σαν συστάδες και φαίνεται να δημιουργεί ευθύγραμμα τμήματα τα οποία τέμνουν τον οριζόντιο άξονα. Επίσης, παρατηρείται ότι αρκετά από τα υπόλοιπα εμφανίζουν μεγάλες αρνητικές τιμές, μακριά από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές και λαμβάνοντας υπόψιν την υψηλή τιμή του MSE, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι υπάρχουν ορισμένα πιθανά ιδιάζοντα σημεία.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερο τμήμα τους. Έτσι, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Ακολουθεί η ανάλυση για το μοντέλο πρόβλεψης κρουσμάτων για την πόλη της Μόσχας. Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

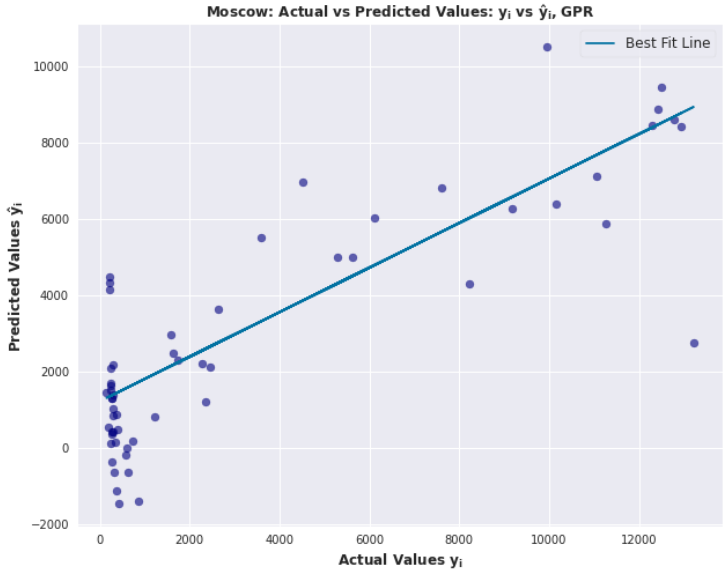
Μετρική Αξιολόγησης	**Μόσχα**
RMSE	2536.99 (cases per million)
R^2	0.674
EVS	0.684
MAE	1752.10 (cases per million)
MAPE	2.62%
MSE	6436341.54

Πίνακας 36 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, GPR, Μόσχα, Ιδία Επεξεργασία*

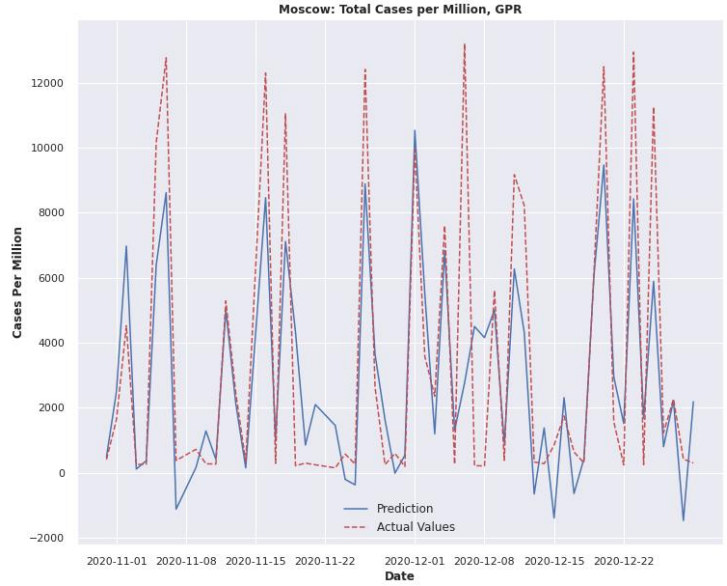
Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε μεσαία επίπεδα. Η R^2 και EVS εμφανίζουν σχετικά χαμηλή τιμή. Έχουν τιμή κοντά στο 0.68, και ως εκ τούτου οι ανεξάρτητες μεταβλητές έχουν μέτριο μέτρο επίδρασης στην εξαρτημένη μεταβλητή (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 119**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, βρίσκεται κοντά στο 3%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Το MSE εμφανίζει 7ψήφιο αριθμό, ο οποίος όμως είναι μεγαλύτερος από εκείνο του Παρισιού.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Μόσχας.

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μόσχας, παρατηρείται ύπαρξη γραμμικότητας. Αυτό το οποίο γίνεται αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων είναι ότι τα σημεία δεν βρίσκονται κοντά της γραμμικά για τις μικρότερες τιμές των πραγματικών μεταβλητών. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, κυρίως από ένα σημείο και μετά, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη πολύ μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι αρκετά ήπια. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται μερικά πιθανά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία και από την ευθεία ελαχίστων τετραγώνων.



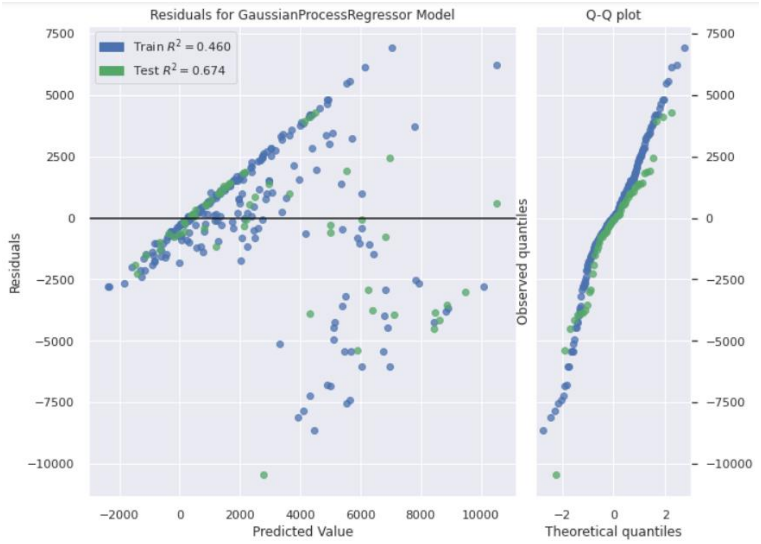
Σχήμα 118 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, GPR, Μόσχα, Ιδία Επεξεργασία*



Σχήμα 119 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, GPR, Μόσχα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι δεν έχει γίνει πολύ ικανοποιητική προσαρμογή του αλγορίθμου στις πραγματικές τιμές των κρουσμάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει κατά κάποιον τρόπο, την τάση των κρουσμάτων, όμως, αδυνατεί πλήρως να υπολογίσει τις πραγματικές τιμές για τα μέγιστα, για τα ελάχιστα, αλλά και για μικρές μεταβολές. Χαρακτηριστικά παραδείγματα είναι η περίοδος 17/11/2020 με 23/11/2020 και η περίοδος 10/12/2020 με 17/12/2020. Συνεπώς, η προσαρμογή του μοντέλου στις πραγματικές τιμές δεν είναι αρκετά επιθυμητή.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 120 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Μόσχα, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται ύπαρξη υπολοίπων σε ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Επίσης, εμφανίζονται σημεία με σημαντικά μεγάλες αρνητικές τιμές. Τα σημεία αυτά πιθανότατα να μπορούσαν να χαρακτηριστούν ως ιδιάζοντα σημεία.

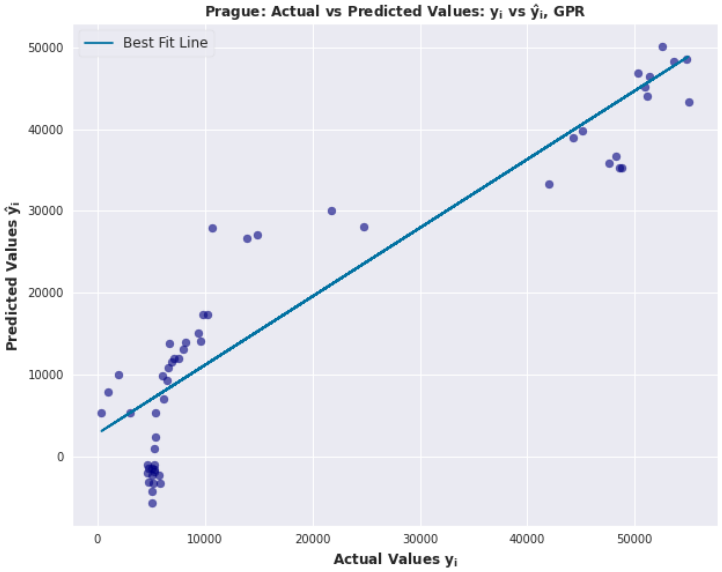
Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία για τα δεδομένα ελέγχου και για τα δεδομένα εκπαίδευσης εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερό τους τμήμα. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Για να ολοκληρωθεί η ανάλυση των δημιουργηθέντων GPR μοντέλων για πρόβλεψη κρουσμάτων στις μελετώμενες πόλεις, χρειάζεται να παρατεθούν τα αποτελέσματα τα οποία προκύπτουν για την πόλη της Πράγας. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	**Πράγα**
RMSE	7631.77 (cases per million)
R^2	0.851
EVS	0.852
MAE	6502.48 (cases per million)
MAPE	1.43%
MSE	58243983.35

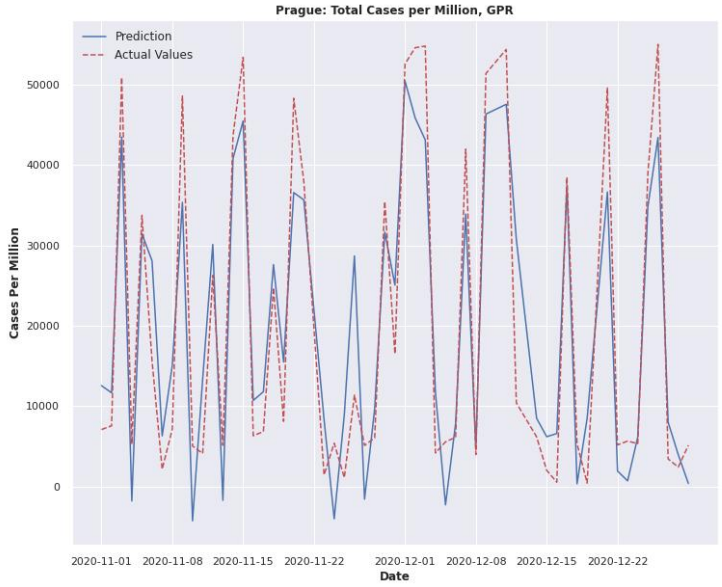
Πίνακας 37 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, GPR, Πράγα, Ιδία Επεξεργασία*

Βάσει όσων αναφέρθηκαν στην υποενότητα **4.2**, οι παραπάνω μετρικές αξιολόγησης είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.80, άρα η συσχέτιση των μεταβλητών θεωρείται υψηλή. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 122**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό και μικρότερο από 10%, άρα θεωρείται πολύ καλό. Το MSE και για την πόλη της Πράγας, εμφανίζει το μεγαλύτερο 8ψήφιο αριθμό.



Σχήμα 121 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, GPR, Πράγα, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Πράγας, παρατηρείται ότι υπάρχει μία σχετική γραμμικότητα. Αυτό σημαίνει ότι στα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή, αλλά εμφανίζεται και η δημιουργία συστάδων. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται ήπια ισχύς, λόγω της μικρής κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.

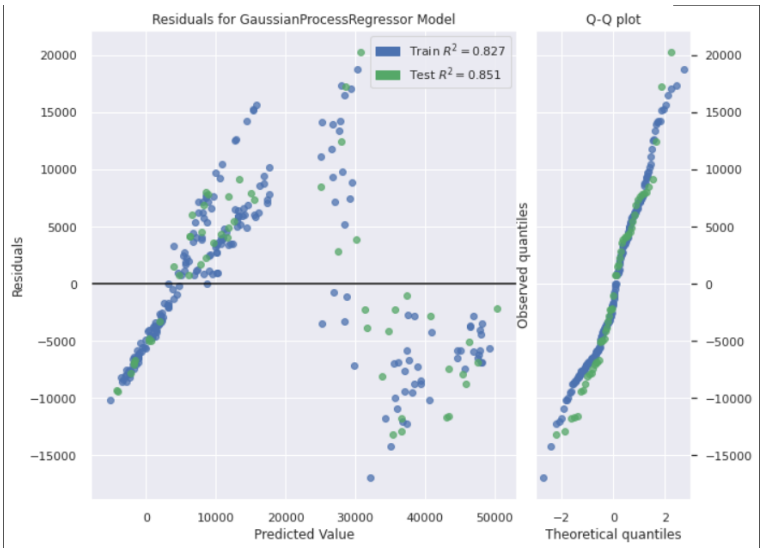


Σχήμα 122 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, GPR, Πράγα, Ιδία Επεξεργασία*

Στο διάγραμμα, **Σχήμα 122**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος έχει αποδώσει σχετικά καλά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως για παράδειγμα 22/11/2020 – 28/11/2020.

Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα. Υπάρχουν, λοιπόν, αρκετά σημεία τα οποία δεν έχει καταφέρει το μοντέλο να προσαρμόσει στις πραγματικές τιμές των καταγεγραμμένων κρουσμάτων.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 123 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται ότι για σχετικά μικρές τιμές, δημιουργείται ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Αυτό το ευθύγραμμο τμήμα και η έντονη γραμμική συγκέντρωση των σημείων, μπορεί να ερμηνευτεί και από τη συστάδα η οποία παρατηρείται στο διάγραμμα από το **Σχήμα 121**. Ακόμη, συναντώνται σημεία τα οποία εμφανίζουν σημαντική απόσταση, μακριά από τον οριζόντιο άξονα.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία εμφανίζουν δύο σχεδόν ταυτιζόμενες ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος, κυρίως στην αριστερή ουρά. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τα αποτελέσματα των μοντέλων πρόβλεψης κρουσμάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Δ.

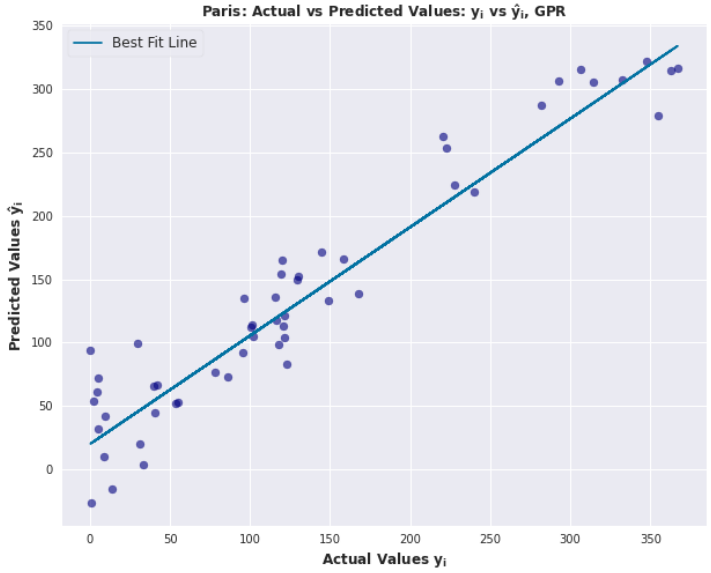
4.4.4.2 *Πρόβλεψη Θανάτων*

Το GPR μοντέλο το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των θανάτων, είναι εκείνο για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	**Παρίσι**
RMSE	32.75 (deaths per million)
R^2	0.910
EVS	0.912
MAE	25.11 (deaths per million)
MAPE	6.30%
MSE	1072.82

Πίνακας 38 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, GPR, Παρίσι, Ιδία Επεξεργασία*

Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Πράγας, είναι ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.90, άρα η συσχέτιση των ανεξάρτητων μεταβλητών, με την εξαρτημένη μεταβλητή είναι αρκετά υψηλή. Δηλαδή, σημαίνει ότι 91% της μεταβολής της εξαρτώμενης μεταβλητής μπορεί να εξηγηθεί από τις ανεξάρτητες μεταβλητές. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές 350 κρούσματα (Βλέπε **Σχήμα 125**). Ακόμη το ποσοστό από το MAPE είναι μικρότερο από 10%, άρα θεωρείται πολύ καλό. Τέλος, όπως έχει αναφερθεί, δεν υπάρχει κάποια «σωστή» τιμή για το MSE. Ο κύριος σκοπός χρήσης του είναι η επιλογή ενός μοντέλου, έναντι κάποιου άλλου. Η τιμή του, στην περίπτωση αυτή, είναι 4ψήφια και είναι η χαμηλότερη τιμή για MSE στα μοντέλα πρόβλεψης θανάτων με GPR.

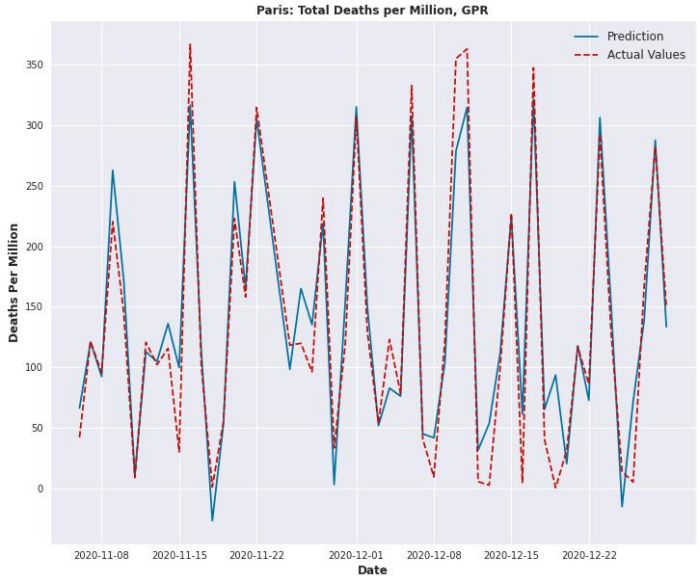


Σχήμα 124 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, GPR, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για το Παρίσι, εμφανίζεται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος σχηματίζουν και ακολουθούν, όπως επίσης μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων μπορούν εύκολα να γίνουν όσα προαναφέρθηκαν αντιληπτά. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται μέτρια ισχύς, λόγω της μεσαίας κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από μελέτη του παραπάνω διαγράμματος, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία απέχουν από τα υπόλοιπα και τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.

Στο παρακάτω διάγραμμα, **Σχήμα 125**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και απεικονίζονται και οι πραγματικές τιμές των θανάτων για την πόλη του Παρισιού και για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος έχει αποδώσει καλά, όχι όμως πλήρως ικανοποιητικά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως το διάστημα 23/11/2020 – 25/11/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και

κάποια ελάχιστα. Έτσι, ορισμένα προβλεπόμενα σημεία φαίνεται ότι δεν προσαρμόζονται πλήρως στις πραγματικές τιμές των θανάτων.

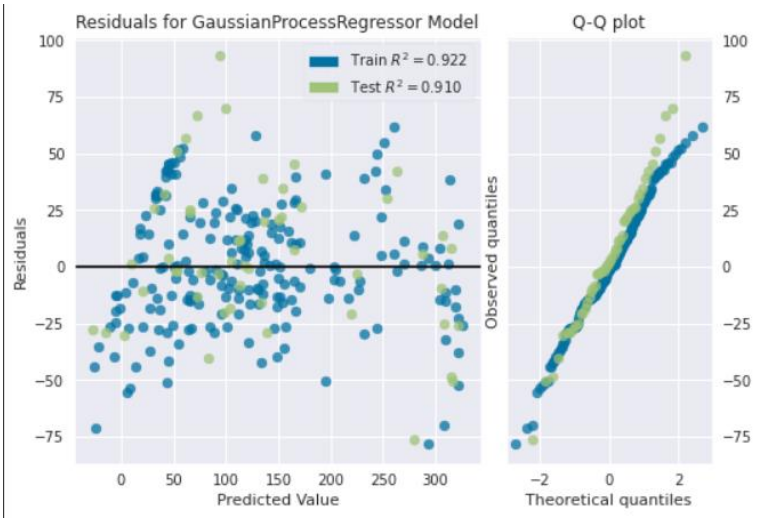


Σχήμα 125 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, GPR, Παρίσι, Ιδία Επεξεργασία*

Ακολουθεί το διάγραμμα των υπολοίπων, στο οποίο παρουσιάζονται τα υπόλοιπα για το συγκεκριμένο μοντέλο και απεικονίζονται τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.

Από το εν λόγω διάγραμμα, **Σχήμα 126**, φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Συνεπώς, ένα γραμμικό μοντέλο μπορεί να περιγράψει τα δεδομένα αυτά. Επίσης, παρατηρούνται ορισμένα σημεία σε μακρινή απόσταση από τον οριζόντιο άξονα, γεγονός το οποίο φανερώνει ότι συναντώνται αρκετά πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων εμφανίζουν μία σχεδόν ευθεία γραμμή και διαφοροποιούνται σε ένα μικρό εύρος, στην δεξιά ουρά. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.



Σχήμα 126 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Παρίσι, Ιδία Επεξεργασία*

Στη συνέχεια, ακολουθεί η ανάλυση μοντέλων για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη της Μαδρίτης.

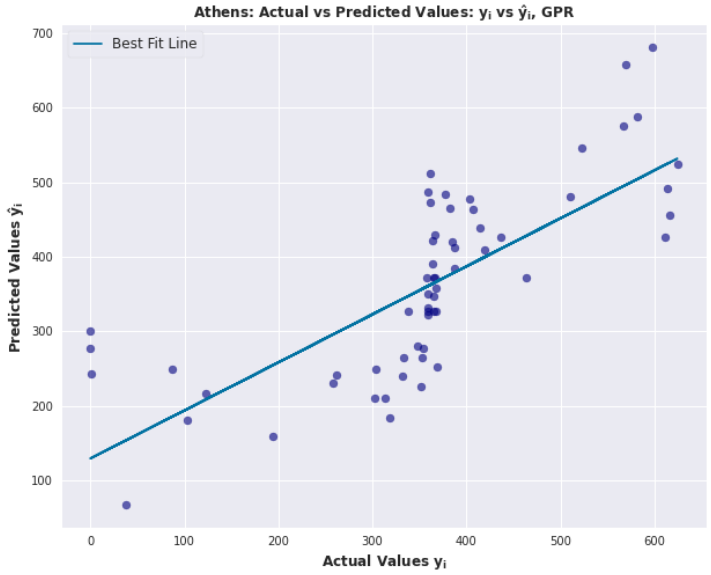
Ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

Μετρική Αξιολόγησης	Αθήνα
RMSE	97.87 (deaths per million)
R^2	0.552
EVS	0.554
MAE	73.11 (deaths per million)
MAPE	51.62%
MSE	9577.65

Πίνακας 39 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, GPR, Αθήνα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε σχετικά χαμηλά πλαίσια. Η R^2 και EVS εμφανίζουν χαμηλές τιμές. Είναι ίσες με 0.55 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν αδύναμη συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο βάσει του μεγέθους των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 128**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά υψηλό, ξεπερνάει το 50%, άρα δεν θεωρείται καλό. Το MSE, είναι ένας 4ψήφιος αριθμός, μεγαλύτερος από το αντίστοιχο MSE του Παρισιού.

Στη συνέχεια, βρίσκεται το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και στις προβλεπόμενες τιμές για την πόλη της Αθήνας και το μοντέλο GPR.

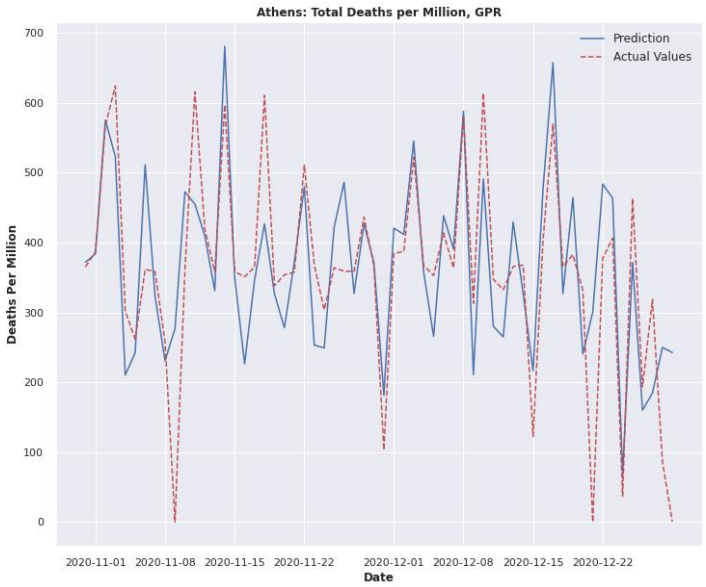


Σχήμα 127 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, GPR, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, παρατηρείται σχετική ύπαρξη γραμμικότητας. Παρατηρείται, δηλαδή, ότι τα σημεία του διαγράμματος δεν φαίνεται να σχηματίζουν ξεκάθαρα μία ευθεία γραμμή, αλλά η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργούνται συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι

συναντώνται μερικά ιδιάζοντα σημεία, καθώς παρατηρούνται σημεία τα οποία απέχουν τόσο από την ευθεία ελαχίστων τετραγώνων όσο και από τις συγκεντρώσεις των υπόλοιπων σημείων.

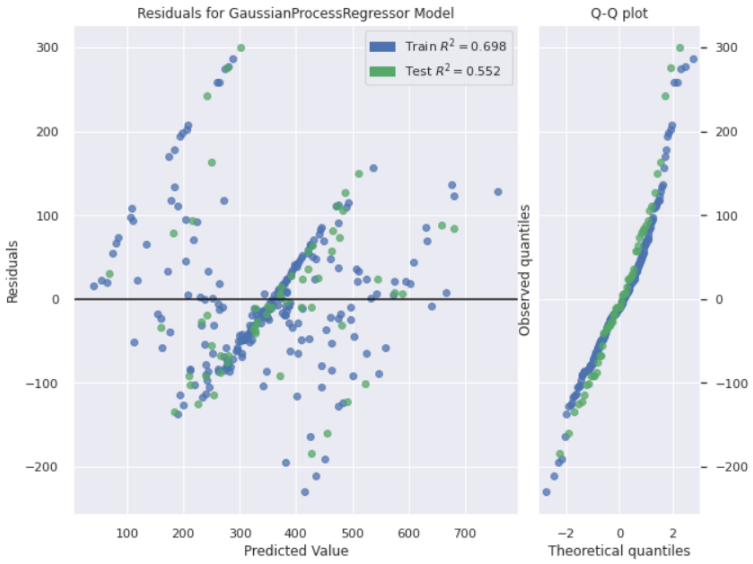
Στο παρακάτω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων, μαζί με τις πραγματικές τους τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, για την πόλη της Αθήνας, είναι ότι το GPR μοντέλο εντοπίζει την τάση των πραγματικών τιμών. Εν τούτοις, αδυνατεί να προβλέψει πλήρως τις ακριβείς τιμές, για τα περισσότερα μέγιστα και για τα περισσότερα ελάχιστα. Έτσι, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα. Σε ορισμένες περιπτώσεις φαίνεται να υπερεκτιμάει και σε άλλες φαίνεται να υποτιμάει τον προβλεπόμενο αριθμό κρουσμάτων. Τα παραπάνω, επιβεβαιώνονται από τις σχετικά χαμηλές τιμές των μετρικών, **Πίνακας 39**. Συμπεραίνεται, ότι η προσαρμογή του μοντέλου πρόβλεψης θανάτων, δεν είναι αρκετά ικανοποιητική για την περίπτωση του GPR και της πόλης της Αθήνας.



Σχήμα 128 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, GPR, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων, **Σχήμα 129**, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ακόμη, παρατηρείται ότι δημιουργείται ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Το ευθύγραμμο αυτό τμήμα, πρόκειται για τη συστάδα η οποία εντοπίζεται από το Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών, **Σχήμα 127**. Επιπλέον, εμφανίζονται ορισμένα σημεία μακριά από τη συγκέντρωση των υπολοίπων και μακριά από τον άξονα y. Πρόκειται για πιθανά ιδιάζοντα σημεία.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο επικαλυπτόμενες ευθείες γραμμές, άρα ταυτίζονται, εκτός από τις ουρές τους. Ως εκ τούτου, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.



Σχήμα 129 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, GPR, Αθήνα, Ιδία Επεξεργασία*

Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης θανάτων για την πόλη της Μαδρίτης.

Μετρική Αξιολόγησης	**Μαδρίτη**
RMSE	147.81 (deaths per million)
R^2	0.701
EVS	0.732
MAE	107.60 (deaths per million)
MAPE	19.87%
MSE	21848.34

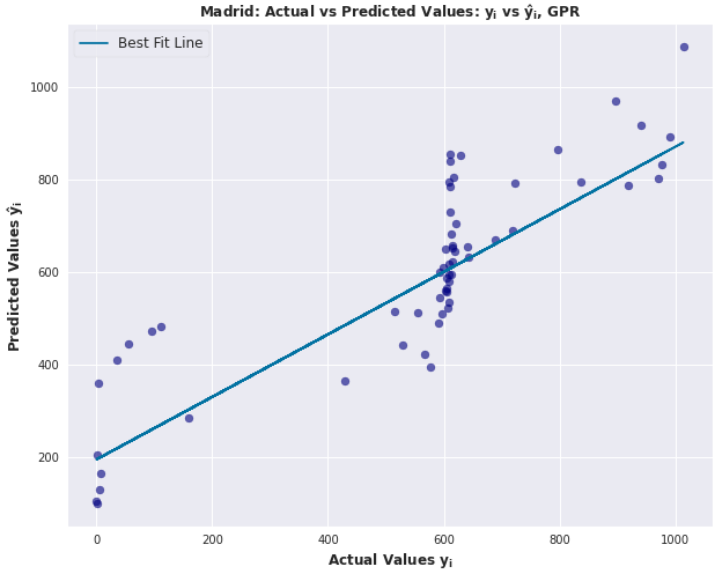
Πίνακας 40 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, GPR, Μαδρίτη, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την Μαδρίτη, κυμαίνονται σε σχετικά ικανοποιητικά πλαίσια. Η R^2 και EVS εμφανίζουν μεσαίες τιμές. Έχουν τιμή μεγαλύτερη από 0.70, και ως εκ τούτου θεωρείται ότι εμφανίζεται μέτριο μέτρο επίδρασης, δηλαδή ότι η συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη είναι μέτρια (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 131**). Ακόμη, το ποσοστό από το MAPE βρίσκεται κοντά στο 20%, άρα, θεωρείται καλό. Να σημειωθεί ότι το MSE του εν λόγω μοντέλου, είναι 5ψήφιος αριθμός.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.

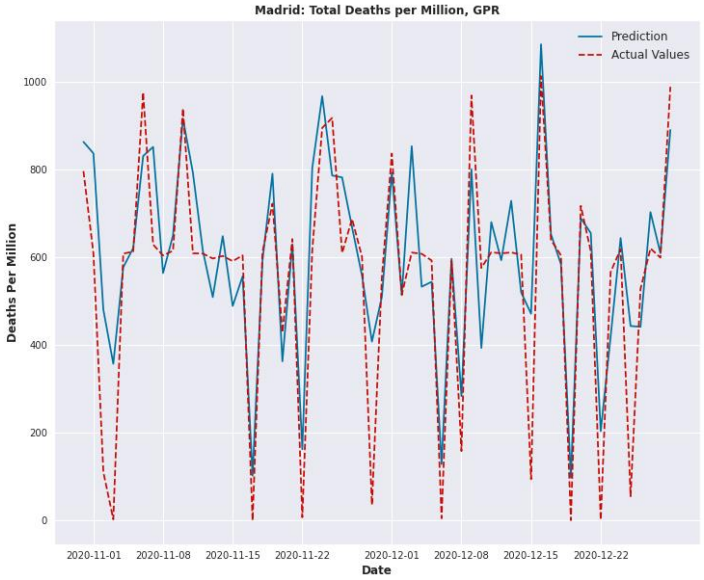
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, παρατηρείται ύπαρξη ιδιαίτερης γραμμικότητας ανάμεσα στα σημεία. Παρατηρείται ότι η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργεί ορισμένες συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ήπια. Τέλος, μελετώντας το παραπάνω διάγραμμα θα

μπορούσε κάποιος να θεωρήσει ότι συναντώνται ορισμένα ιδιάζοντα σημεία, τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία.



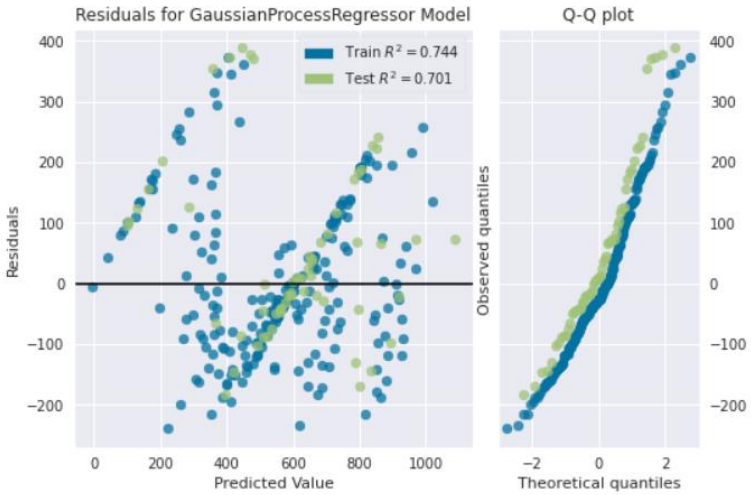
Σχήμα 130 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, GPR, Μαδρίτη, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα, **Σχήμα 131**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Πράγας και για το μοντέλο GPR, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών σε ένα μέτριο επίπεδο, εν τούτοις αδυνατεί να προβλέψει πλήρως την ακριβή τιμή τόσο για αρκετά μέγιστα, όσο και για αρκετά ελάχιστα, αλλά και για μεσαίες τιμές. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές των θανάτων. Χαρακτηριστικό είναι το παράδειγμα για τις ημερομηνίες 10/11/2020 έως και περίπου 15/11/2020. Τα παραπάνω, επιβεβαιώνονται από τις σχετικά χαμηλές τιμές των μετρικών, όπως λ.χ. το MAE, **Πίνακας 40**.



Σχήμα 131 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, GPR, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 132 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, GPR, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, παρατηρείται ότι δημιουργείται ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές, αυτό το οποίο μπορεί να γίνει θεωρηθεί είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα υπόλοιπα για τα δεδομένα εκπαίδευσης και για τα δεδομένα ελέγχου εμφανίζουν δύο ευθείες γραμμές, οι οποίες δεν επικαλύπτονται, όμως ακολουθούν το ίδιο μοτίβο, την ίδια τάση. Έτσι, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

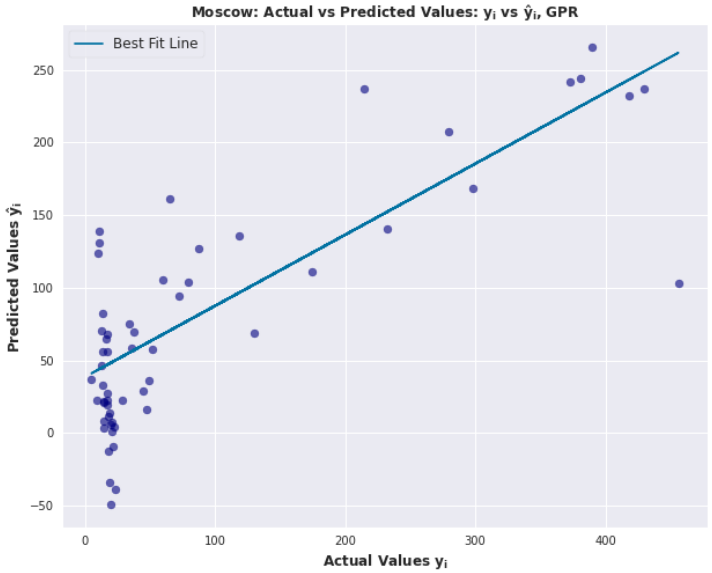
Τέταρτη πόλη ανάλυσης αποτελεί η πόλη της Μόσχας. Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	**Μόσχα**
RMSE	81.96 (deaths per million)
R^2	0.588
EVS	0.599
MAE	54.66 (deaths per million)
MAPE	1.73%
MSE	6716.63

Πίνακας 41 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, GPR, Μόσχα, Ιδία Επεξεργασία*

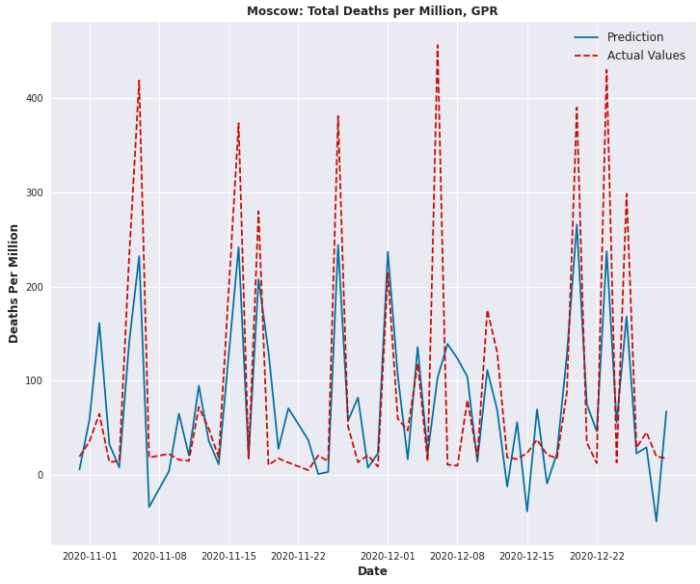
Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε μεσαία επίπεδα. Η R^2 και EVS εμφανίζουν αρκετά χαμηλή τιμή. Η τιμή για το R^2 βρίσκεται κοντά στο 0.60 και ως εκ τούτου οι ανεξάρτητες μεταβλητές έχουν μέτριο μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 134**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, βρίσκεται κοντά στο 2% και θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Το MSE, όπως στην περίπτωση της Αθήνας και του Παρισιού, εμφανίζει 4ψήφιο αριθμό.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Μόσχας.



Σχήμα 133 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, GPR, Μόσχα, Ιδία Επεξεργασία*

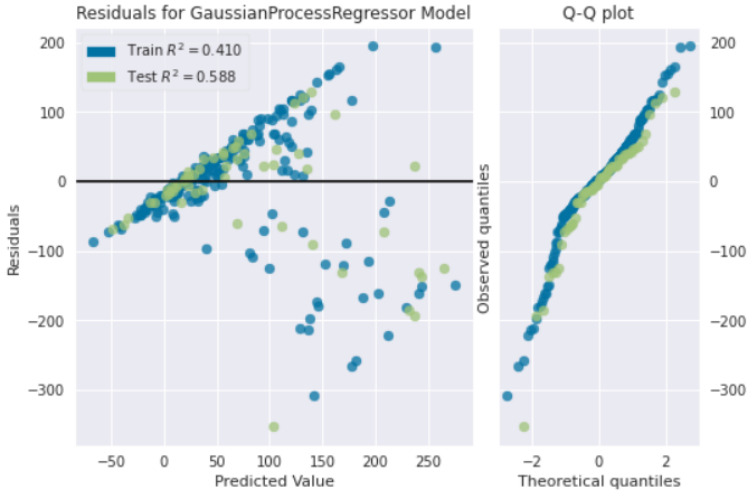
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μόσχας, παρατηρείται ιδιαίτερη γραμμικότητα. Ιδιαίτερη, καθώς για μικρές τιμές φαίνεται να δημιουργείται μία συστάδα η οποία δεν εμφανίζει έντονη γραμμικότητα. Έτσι, μπορεί να δικαιολογηθεί η «μεσαία» τιμή του R^2. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μικρή κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι αρκετά ήπια. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται μερικά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τις συγκεντρώσεις των υπόλοιπων σημείων και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 134 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, GPR, Μόσχα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι έχει γίνει μία μέτρια προσαρμογή του αλγορίθμου στις πραγματικές τιμές των θανάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει κατά κάποιον τρόπο, την τάση των κρουσμάτων, όμως, αδυνατεί πλήρως να υπολογίσει τις πραγματικές τιμές στην πλειονότητα της χρονικής περιόδου μελέτης. Συνεπώς, η προσαρμογή του αλγορίθμου σε αυτήν την περίπτωση δεν φαίνεται να είναι πλήρως ικανοποιητική.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 135 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, GPR, Μόσχα, Ιδία Επεξεργασία*

Από το διάγραμμα υπολοίπων, **Σχήμα 135**, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Όμως, παρατηρείται ότι η μεγαλύτερη συγκέντρωση δημιουργεί ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Πρόκειται για σημεία τα οποία εντοπίστηκαν στο διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών, ως συστάδα. Ακόμη, παρατηρούνται ορισμένα σημεία με μεγάλες τιμές, κυρίως αρνητικές, μακριά από την ευθεία y=0. Ίσως να πρόκειται για ιδιάζοντα σημεία, καθώς πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των υπόλοιπων σημείων.

Τέλος, από το διάγραμμα Q-Q, **Σχήμα 135**, παρατηρείται ότι τα σημεία των υπολοίπων για τα δύο σετ εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερό τους. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

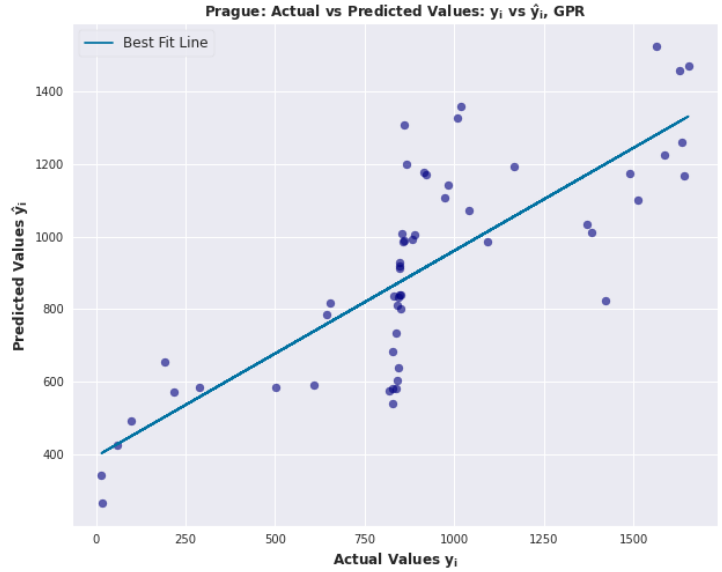
Στο τελικό μοντέλο ανάλυσης για πρόβλεψη θανάτων με τον αλγόριθμο GPR, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Πράγας.

Μετρική Αξιολόγησης	**Πράγα**
RMSE	255.29 (deaths per million)
R^2	0.619
EVS	0.619
MAE	210.58 (deaths per million)
MAPE	1.06%

MSE	65170.55

Πίνακας 42 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, GPR, Πράγα, Ιδία Επεξεργασία*

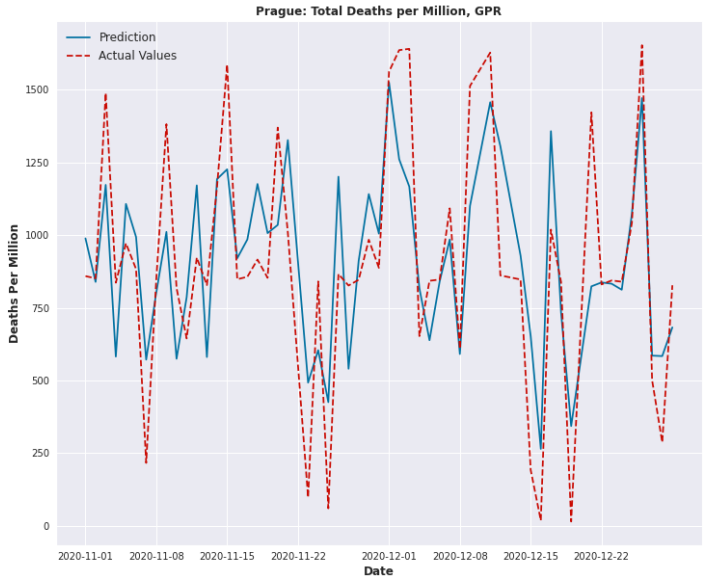
Βάσει όσων αναφέρθηκαν και στην υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Πράγας, κυμαίνονται σε μέτρια πλαίσια. Η τιμή της R^2 και της EVS είναι σχετικά καλές. Η τιμή της R^2 βρίσκεται κοντά στο 0.62, άρα η συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη θεωρείται μέτρια. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 137**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό και μικρότερο από 2%, άρα θεωρείται πολύ καλό. Το MSE εμφανίζει το μεγαλύτερο 5ψήφιο αριθμό, από τις 5 μελετώμενες πόλεις.



Σχήμα 136 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, GPR, Πράγα, Ιδία Επεξεργασία*

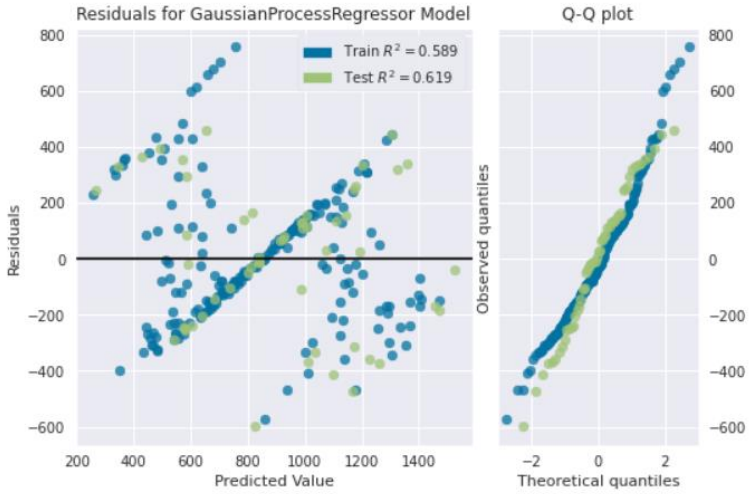
Από το διάγραμμα διασποράς, **Σχήμα 136**, ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Πράγας, παρατηρείται ότι υπάρχει γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται αρκετά ήπια ισχύς, λόγω της μικρής κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται λίγα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν έτσι, καθώς απέχουν αρκετά από την ευθεία ελαχίστων τετραγώνων και από συγκεντρώσεις άλλων σημείων.

Στο διάγραμμα των προβλεπόμενων τιμών, **Σχήμα 137**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι η απόδοση του αλγορίθμου δεν είναι πολύ ικανοποιητική. Δεν είναι ικανός να ακολουθήσει πιστά τη ροή και την τάση των πραγματικών τιμών. Ούτε μπορεί να προβλέψει με πλήρη επιτυχία όλα τα μέγιστα, αλλά και όλα τα ελάχιστα. Άρα, αρκετά σημεία δεν έχουν προσαρμοστεί στις πραγματικές τιμές των θανάτων.



Σχήμα 137 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, GPR, Πράγα, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζονται τα υπόλοιπα για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 138 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, , GPR, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι ένα πλήθος τους βρίσκεται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ακόμη μία φορά παρατηρείται η συγκέντρωση αρκετών σημείων σε ένα ευθύγραμμο τμήμα το οποίο τέμνει τον οριζόντιο άξονα. Τέλος, εμφανίζονται σημεία τα οποία απέχουν από τον εν λόγω άξονα και τα οποία θα μπορούσε κάποιος να τα χαρακτηρίσει ως ιδιάζοντα σημεία.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων, εμφανίζουν δύο σχεδόν ταυτιζόμενες ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τα αποτελέσματα των μοντέλων πρόβλεψης θανάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Δ.

4.4.5 RF Regression

Πέμπτο μοντέλο το οποίο εφαρμόστηκε στα υπάρχοντα δεδομένα για τις εννέα διαφορετικές πόλεις, είναι το Random Forrest Regression. Στις επόμενες σελίδες ακολουθούν τα

αποτελέσματα του αλγορίθμου για την Αθήνα, τη Μαδρίτη, τη Μόσχα, το Παρίσι και την Πράγα.

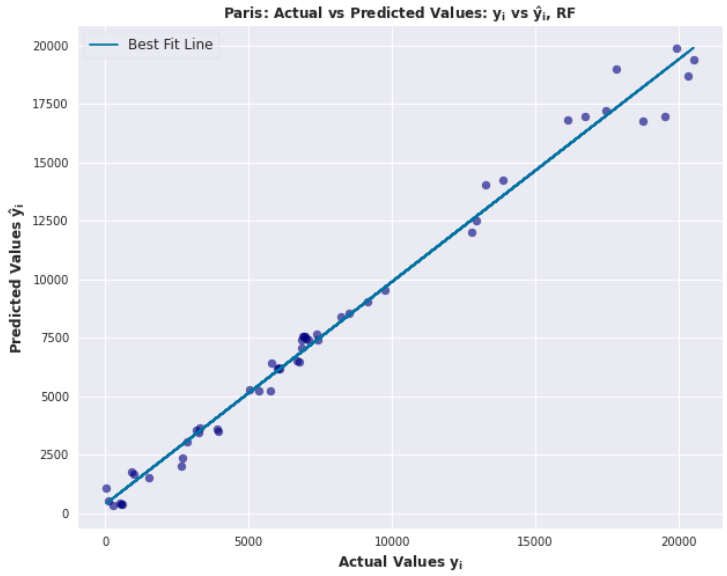
4.4.5.1 Πρόβλεψη Κρουσμάτων

Εκείνο το μοντέλο το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των κρουσμάτων με το RF, είναι εκείνο για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	Παρίσι
RMSE	714.92 (cases per million)
R^2	0.986
EVS	0.986
MAE	475.57 (cases per million)
MAPE	0.70%
MSE	0.00017

Πίνακας 43 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, RF, Παρίσι, Ιδία Επεξεργασία*

Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Παρισιού, είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS είναι αρκετά υψηλές και ίσες με 0.98, άρα πρόκειται για υψηλή συσχέτιση μεταβλητών. Δηλαδή, σημαίνει ότι 98% της μεταβολής της εξαρτημένης μεταβλητής μπορεί να εξηγηθεί από τις ανεξάρτητες μεταβλητές. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές τα περίπου 20000 κρούσματα (Βλέπε **Σχήμα 140**). Ακόμη το ποσοστό από το MAPE είναι πάρα πολύ χαμηλό και μικρότερο από 10%, άρα θεωρείται πολύ καλό. Τέλος, όπως έχει αναφερθεί και σε προηγούμενα, δεν υπάρχει κάποια «σωστή» τιμή για το MSE. Πρόκειται για το μικρότερο MSE το οποίο έχει συναντηθεί έως τώρα, αλλά και το μικρότερο MSE το οποίο συναντάται στην ανάλυση μοντέλων πρόβλεψης κρουσμάτων του RF.

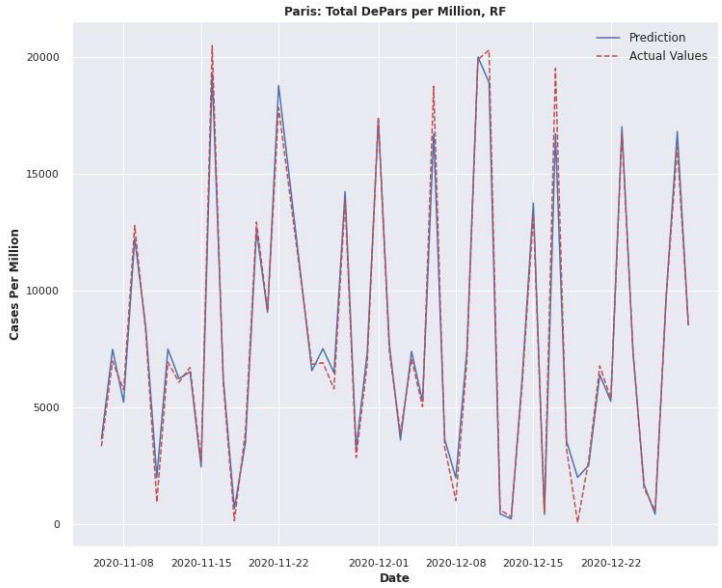


Σχήμα 139 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, RF, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη του Παρισιού, **Σχήμα 139**, εμφανίζεται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν, να ακολουθούν και να μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή.

Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η γραμμικότητα αυτή. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Στην περίπτωση του Παρισιού, θεωρείται ότι συναντάται σχετικά υψηλή ισχύς, λόγω της κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, εντοπίζονται ελάχιστα μεμονωμένα σημεία στα οποία θα μπορούσε να αποδοθεί ο χαρακτηρισμός αυτός.

Στο παρακάτω διάγραμμα, **Σχήμα 140**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και απεικονίζονται και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος RF έχει αποδώσει αρκετά ικανοποιητικά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες μικρές εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως το διάστημα 24/11/2020 – 25/11/2020. Εμφανίζονται σημεία τα οποία δεν έχουν προσαρμοστεί πλήρως στις πραγματικές τιμές των κρουσμάτων, όμως οι διαφορές τους από τις πραγματικές τιμές είναι πολύ μικρές και πρόκειται για πάρα πολύ μικρά χρονικά διαστήματα. Σε γενικές γραμμές, όμως, πρόκειται για μία αξιόλογη προσαρμογή του μοντέλου στις πραγματικές τιμές.

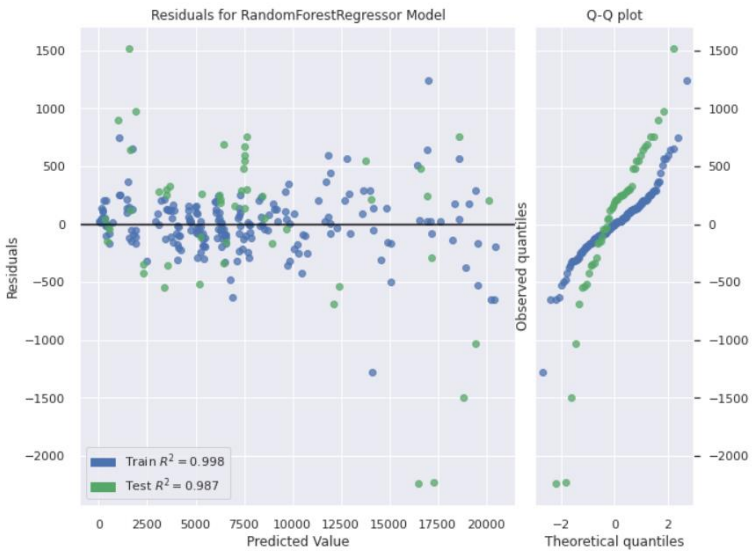


Σχήμα 140 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, RF, Παρίσι, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται επίσης, ότι ορισμένα σημεία υπολοίπων βρίσκονται σε σημαντική απόσταση από την ευθεία y=0. Γεγονός το οποίο φανερώνει ότι συναντώνται αρκετά πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων εμφανίζουν δύο ευθείες γραμμές, οι οποίες τέμνονται και δεν επικαλύπτονται. Θεωρείται, λοιπόν, ότι τα υπόλοιπα των δύο σετ δεδομένων δεν ακολουθούν την ίδια κατανομή.



Σχήμα 141 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Παρίσι, Ιδία Επεξεργασία*

Στη συνέχεια, ακολουθεί η ανάλυση για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη του Παρισιού.

Όπως στην ανάλυση η οποία προηγήθηκε, έτσι ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

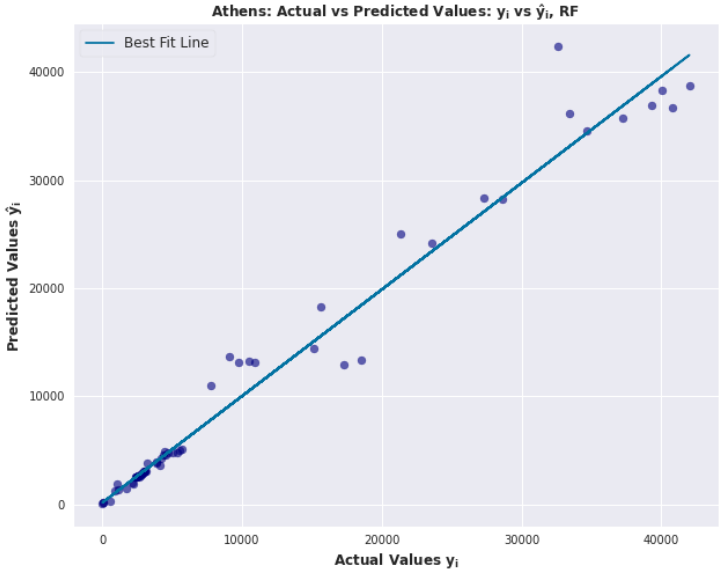
Μετρική Αξιολόγησης	**Αθήνα**
RMSE	2005.45 (cases per million)
R^2	0.974
EVS	0.974
MAE	1059.31 (cases per million)
MAPE	0.36%
MSE	0.00145

Πίνακας 44 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, RF, Αθήνα, Ιδία Επεξεργασία*

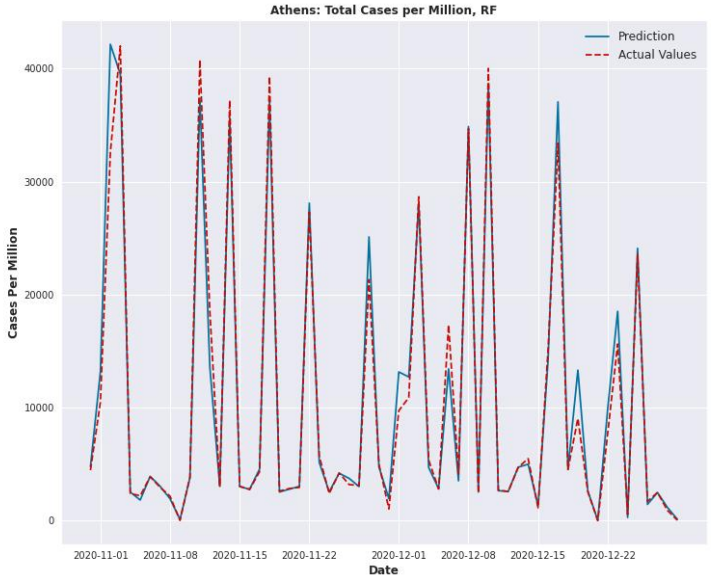
Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε πολύ καλά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν υψηλές τιμές. Είναι ίσες με 0.97 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν πολύ υψηλή συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, βάσει του μεγέθους των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 143**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό, βρίσκεται κοντά στο 0.40%, άρα θεωρείται πολύ καλό. Το MSE και σε αυτήν την περίπτωση εμφανίζει χαμηλή τιμή.

Από το παρακάτω διάγραμμα διασποράς, **Σχήμα 142**, ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, **Σχήμα 142**, παρατηρείται ύπαρξη γραμμικότητας. Παρατηρείται, δηλαδή, ότι από τα σημεία του διαγράμματος φαίνεται να μπορεί να περάσει μία ευθεία γραμμή. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων,

συμπεραίνεται ότι η ισχύς είναι ισχυρή. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά πιθανά ιδιάζοντα σημεία.



Σχήμα 142 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, RF, Αθήνα, Ιδία Επεξεργασία*



Σχήμα 143 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, RF, Αθήνα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων, μαζί με τις πραγματικές τιμές των κρουσμάτων, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για τον αλγόριθμο RF, για την πόλη της Αθήνας, είναι ότι εντοπίζει την τάση των πραγματικών τιμών, εν τούτοις αδυνατεί να προβλέψει πλήρως ορισμένα μέγιστα αλλά και ορισμένα ελάχιστα. Βέβαια, η αστοχία αυτή φαίνεται να εμφανίζει πολύ μικρές τιμές. Ακόμη, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα, όπως για παράδειγμα το διάστημα ανάμεσα στη 01/12/2020 έως 02/12/2020. Η προσαρμογή του μοντέλου στα πραγματικά δεδομένα είναι αρκετά ικανοποιητική.

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Σχήμα 144 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επιπροσθέτως, παρατηρείται ότι ένα σημείο απέχει σημαντικά από την ευθεία y=0, ως προς το θετικό άξονα. Έτσι, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και στις προβλεπόμενες τιμές, γίνεται αντιληπτό ότι πρόκειται για κάποιο πιθανό ιδιάζον σημείο.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο σχεδόν επικαλυπτόμενες ευθείες γραμμές και διαφοροποιούνται σε ένα μικρό εύρος στις ουρές τους. Θεωρείται, λοιπόν, ότι τα δύο σετ δεδομένων για τα υπόλοιπα ακολουθούν την ίδια κατανομή.

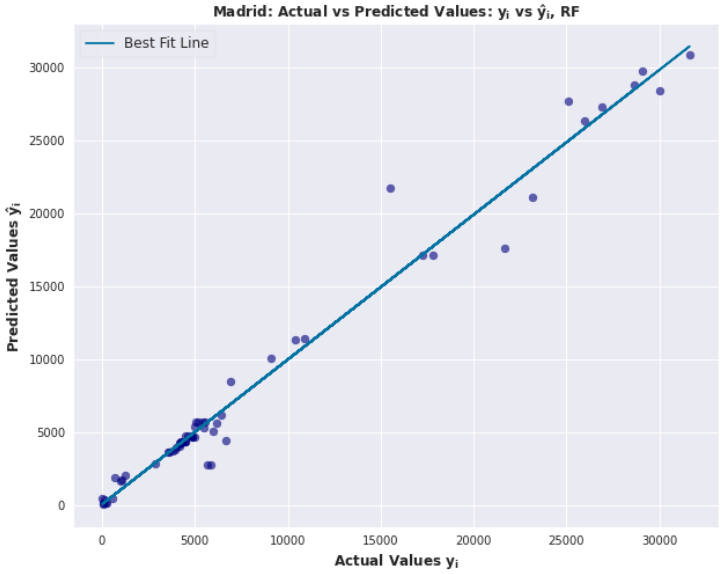
Τρίτη πόλη ανάλυσης αποτελεί η πόλη της Μαδρίτης. Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης κρουσμάτων και παρατίθενται τα γραφήματα των αποτελεσμάτων της πρόβλεψης.

Μετρική Αξιολόγησης	**Μαδρίτη**
RMSE	1318.10 (cases per million)
R^2	0.977
EVS	0.977
MAE	684.19 (cases per million)
MAPE	0.52%
MSE	0.00057

Πίνακας 45 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, RF, Μαδρίτη, Ιδία Επεξεργασία*

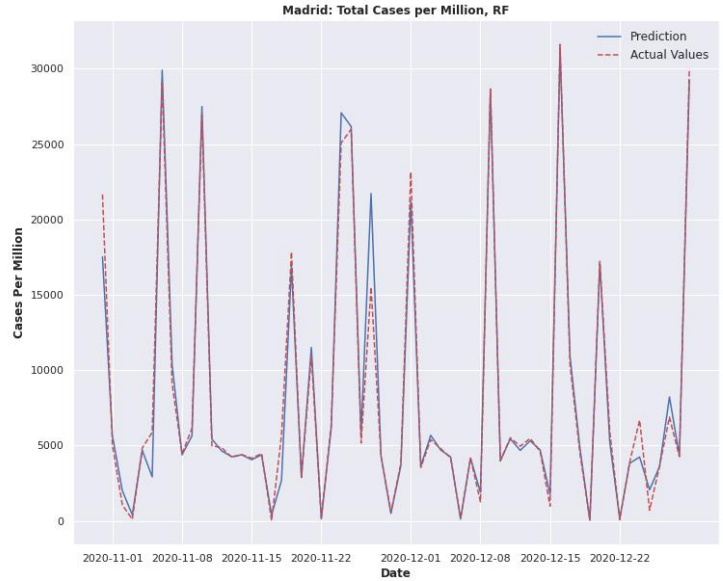
Οι μετρικές αξιολόγησης οι οποίες περιγράφουν τη Μαδρίτη, κυμαίνονται σε σχετικά ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν αρκετά υψηλές τιμές. Έχουν τιμή μεγαλύτερη από 0.90, και ως εκ τούτου θεωρείται ότι εμφανίζεται ισχυρό μέτρο επίδρασης των ανεξάρτητων μετρικών στην εξαρτημένη μεταβλητή (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 146**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό, μικρότερο από 1%, άρα, σύμφωνα με τη βιβλιογραφία, θεωρείται πολύ καλό, αφού όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Το MSE σε αυτήν την περίπτωση είναι ο δεύτερος μικρότερος αριθμός MSE ο οποίος καταγράφεται για τα μοντέλα πρόβλεψης κρουσμάτων και για το RF.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.



Σχήμα 145 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, RF, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, παρατηρείται ότι υπάρχει γραμμικότητα. Δηλαδή, δύναται η προσαρμογή ευθείας γραμμής στα δεδομένα. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι υψηλή. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται λίγα ιδιάζοντα σημεία.

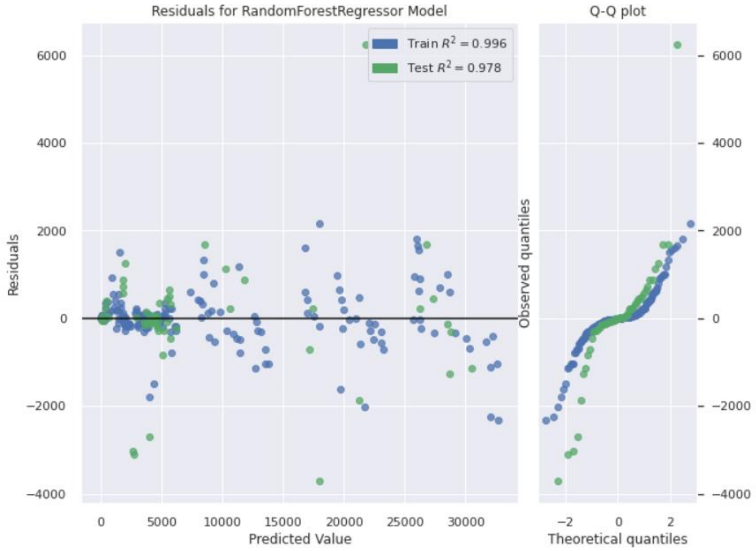


Σχήμα 146 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, RF, Μαδρίτη, Ιδία Επεξεργασία*

Στο διάγραμμα το οποίο ακολουθεί, **Σχήμα 146**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Μαδρίτης και για το RF, είναι ότι ο αλγόριθμος εντοπίζει την τάση και τη ροή των πραγματικών τιμών με ιδιαίτερη επιτυχία. Συναντώνται βέβαια περιπτώσεις κατά τις οποίες αδυνατεί να προβλέψει πλήρως την ακριβή τιμή για ορισμένα μέγιστα και για ορισμένες

μεσαίες τιμές, όμως πρόκειται για ένα πολύ μικρό χρονικό διάστημα και για μικρή απόκλιση. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές. Σε γενικές γραμμές, ο αλγόριθμος για την πόλη αυτή, φαίνεται να έχει κάνει καλή προσαρμογή.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 147 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων για το RF μοντέλο της Μαδρίτης, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, παρατηρείται ότι ορισμένα υπόλοιπα εμφανίζουν μεγάλες αρνητικές τιμές, μακριά από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές, αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι συναντώνται ορισμένα πιθανά ιδιάζοντα σημεία.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερο τμήμα τους. Έτσι, θεωρείται ότι τα δύο σετ δεδομένων για τα υπόλοιπα ακολουθούν την ίδια κατανομή.

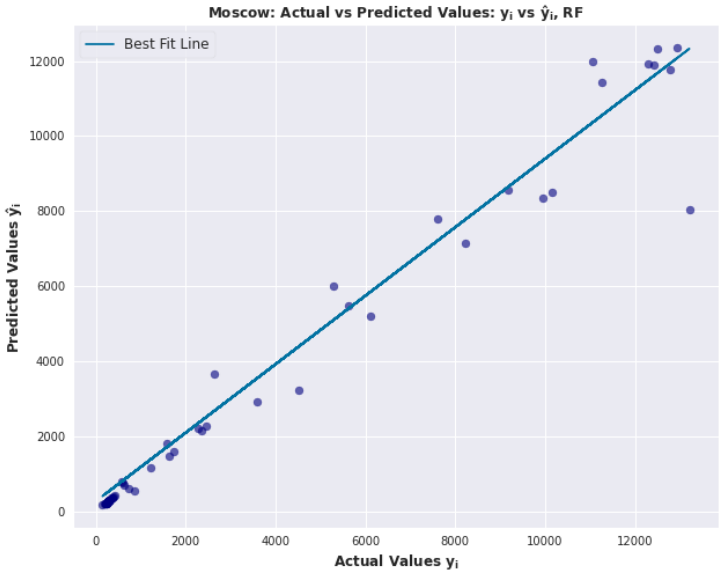
Ακολουθεί η ανάλυση για το μοντέλο πρόβλεψης κρουσμάτων για την πόλη της Μόσχας. Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	**Μόσχα**
RMSE	721.37 (cases per million)
R^2	0.974
EVS	0.976
MAE	343.61 (cases per million)
MAPE	0.13%
MSE	0.00017

Πίνακας 46 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, RF, Μόσχα, Ιδία Επεξεργασία*

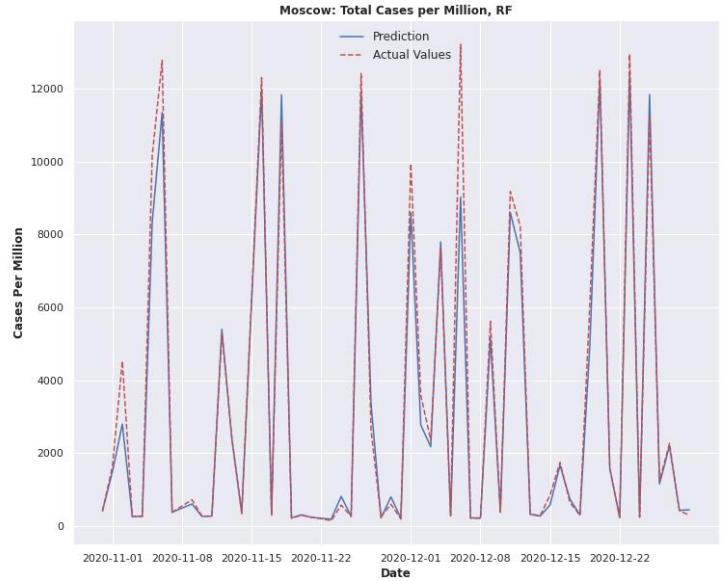
Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε υψηλά επίπεδα. Η R^2 και EVS εμφανίζουν μεγάλη τιμή. Έχουν τιμή κοντά στο 0.97, και ως εκ τούτου οι ανεξάρτητες μεταβλητές έχουν υψηλό μέτρο επίδρασης στην εξαρτημένη μεταβλητή (Moore, D. S., Notz,

W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 149**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, μικρότερο από 1%, άρα θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο. Το MSE εμφανίζει το μικρότερο αριθμό και είναι ίσο με το MSE για την πόλη του Παρισιού.



Σχήμα 148 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, RF, Μόσχα, Ιδία Επεξεργασία*

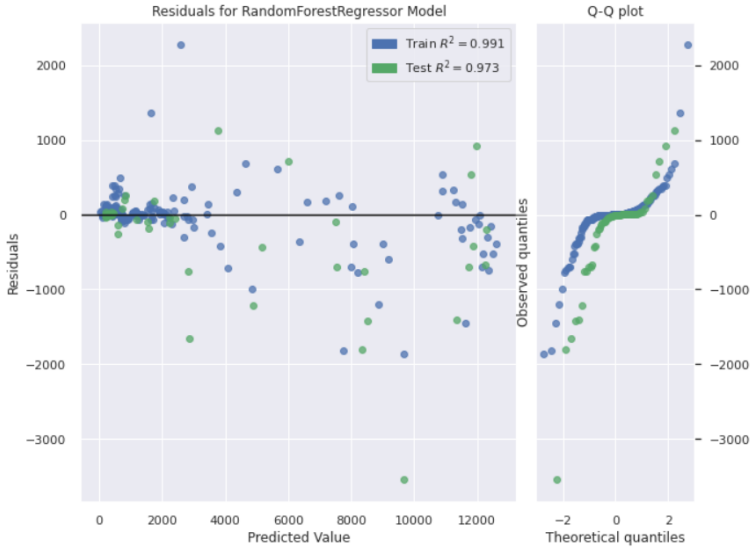
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μόσχας, παρατηρείται ύπαρξη γραμμικότητας. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι έντονη. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται λίγα πιθανά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 149 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, RF, Μόσχα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι έχει γίνει αρκετά ικανοποιητική προσαρμογή του αλγορίθμου στις πραγματικές τιμές των κρουσμάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει την τάση και τη ροή των κρουσμάτων, όμως και εδώ συναντώνται περιπτώσεις κατά τις οποίες αδυνατεί πλήρως να προβλέψει την ακριβή τιμή. Για μία ακόμη φορά, η διαφορά ανάμεσα στην προβλεπόμενη και την πραγματική τιμή, όπως αποτυπώνεται και στο διάγραμμα, φαίνεται να είναι μικρή και για ένα αρκετά σύντομο χρονικό διάστημα. Συνεπώς, η προσαρμογή του μοντέλου στις πραγματικές τιμές είναι αρκετά ικανοποιητική.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.



Σχήμα 150 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Μόσχα, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, εμφανίζονται λίγα σημεία με σημαντικά μεγάλες αρνητικές και θετικές τιμές. Τα σημεία αυτά πιθανότατα να μπορούσαν να χαρακτηριστούν ως ιδιάζοντα.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία για τα δεδομένα ελέγχου και για τα δεδομένα εκπαίδευσης, για τα υπόλοιπα, εμφανίζουν δύο ευθείες γραμμές, οι οποίες έχουν ίδια μορφή, αλλά δεν επικαλύπτονται σε όλο τους το τμήμα. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

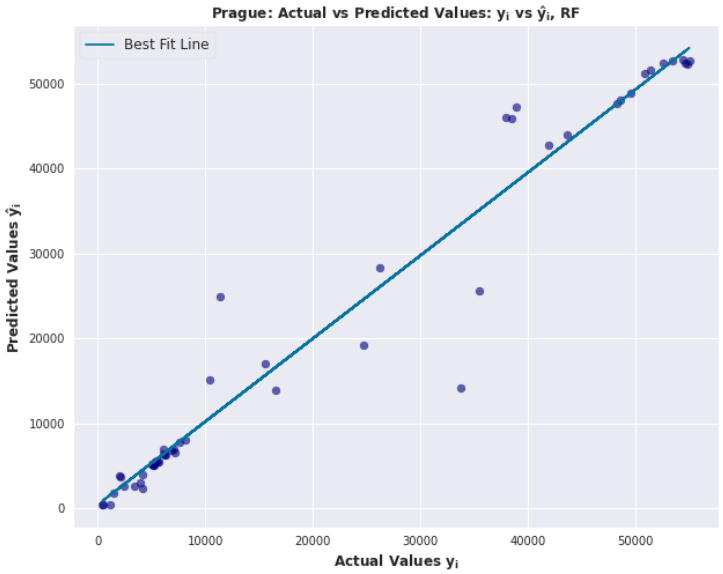
Για να ολοκληρωθεί η ανάλυση των δημιουργηθέντων RF μοντέλων για πρόβλεψη κρουσμάτων στις μελετώμενες πόλεις, χρειάζεται να παρατεθούν τα αποτελέσματα τα οποία προκύπτουν για την πόλη της Πράγας. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	**Πράγα**
RMSE	4117.15 (cases per million)
R^2	0.957
EVS	0.957
MAE	1906.26 (cases per million)
MAPE	0.15%

MSE	0.00558

Πίνακας 47 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, RF, Πράγα, Ιδία Επεξεργασία*

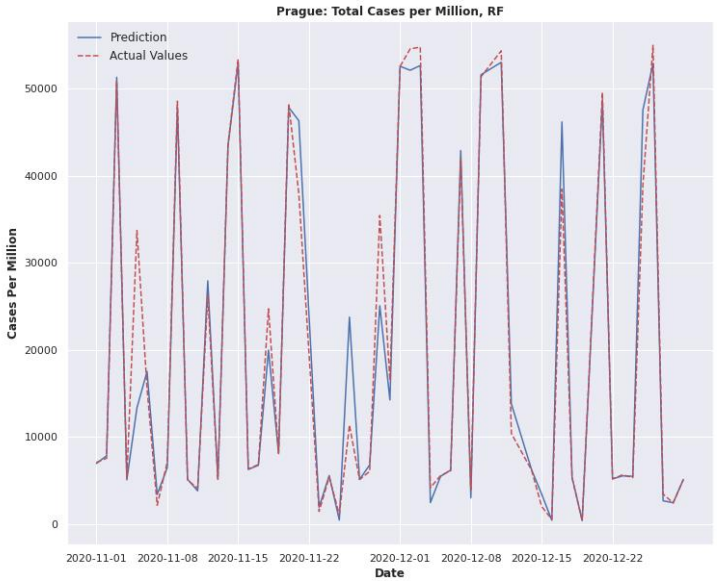
Βάσει όσων αναφέρθηκαν στην υποενότητα **4.2**, οι παραπάνω μετρικές αξιολόγησης είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.90, άρα η συσχέτιση των μεταβλητών θεωρείται υψηλή. Δηλαδή, το 96% των ανεξάρτητων μεταβλητών μπορεί να επηρεάσει την εξαρτημένη μεταβλητή. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 152**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό και μικρότερο από 1%, άρα θεωρείται πολύ καλό.



Σχήμα 151 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, RF, Πράγα, Ιδία Επεξεργασία*

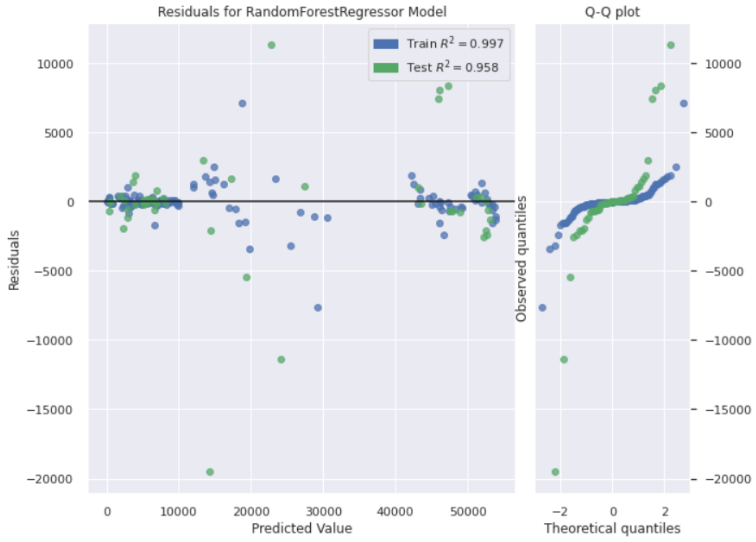
Από το παραπάνω διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Πράγας, συναντάται γραμμικότητα. Αυτό σημαίνει ότι στα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή, αλλά εμφανίζεται και η δημιουργία συστάδων. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται έντονη ισχύς, λόγω της μεγάλης κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.

Στο διάγραμμα, **Σχήμα 152**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος έχει αποδώσει αρκετά καλά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως για παράδειγμα 24/11/2020 έως 28/11/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα. Υπάρχουν, λοιπόν, ορισμένα σημεία τα οποία δεν έχει καταφέρει το μοντέλο να προσαρμόσει στις πραγματικές τιμές των καταγεγραμμένων κρουσμάτων. Η προσαρμογή του αλγορίθμου είναι ικανοποιητική, όμως εμφανίζει ορισμένες αστοχίες.



Σχήμα 152 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, RF, Πράγα, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 153 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ακόμη, συναντώνται σημεία τα οποία εμφανίζουν σημαντική απόσταση, μακριά από τον οριζόντιο άξονα. Πιθανότατα είναι ιδιάζοντα σημεία.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές οι οποίες δεν ταυτίζονται πλήρως, όμως ακολουθούν την ίδια σχεδόν μορφή. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τα αποτελέσματα των μοντέλων πρόβλεψης κρουσμάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Ε.

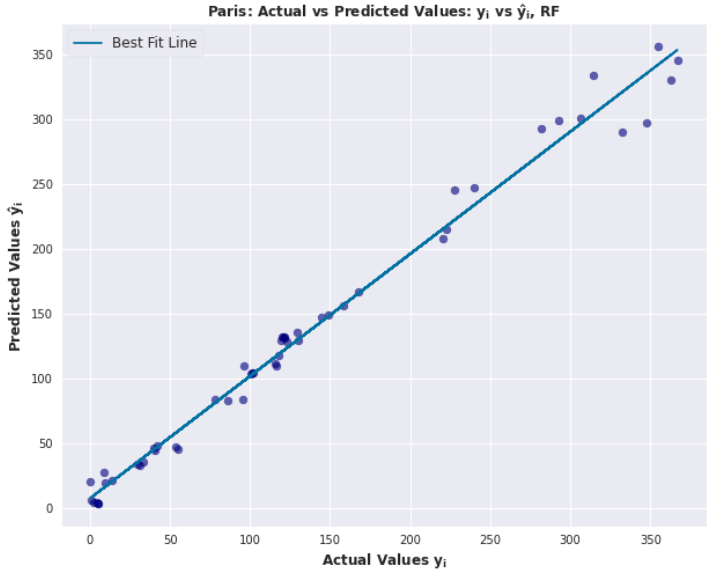
4.4.5.2 *Πρόβλεψη Θανάτων*

Το RF μοντέλο το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των θανάτων, είναι εκείνο για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	Παρίσι
RMSE	13.20 (deaths per million)
R^2	0.985
EVS	0.985
MAE	9.20 (deaths per million)
MAPE	1.50%
MSE	0.00006

Πίνακας 48 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, RF, Παρίσι, Ιδία Επεξεργασία*

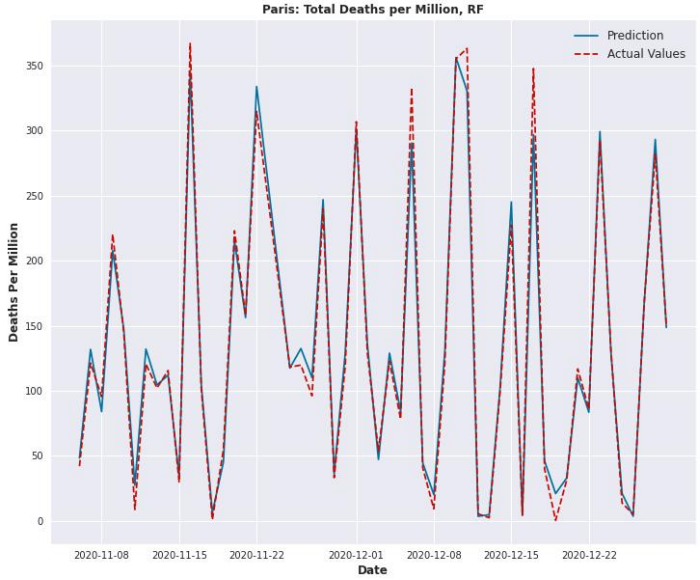
Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Πράγας, είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.90, άρα η συσχέτιση των ανεξάρτητων μεταβλητών, με την εξαρτημένη μεταβλητή είναι αρκετά υψηλή. Δηλαδή, σημαίνει ότι 98% της μεταβολής της εξαρτώμενης μεταβλητής μπορεί να εξηγηθεί από τις ανεξάρτητες μεταβλητές. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές 350 κρούσματα (Βλέπε **Σχήμα 155**). Ακόμη το ποσοστό από το MAPE είναι μικρότερο από 10%, άρα θεωρείται πολύ καλό. Τέλος, η τιμή του MSE είναι η χαμηλότερη τιμή η οποία συναντάται στα μοντέλα πρόβλεψης θανάτων με RF.



Σχήμα 154 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, RF, Παρίσι, Ιδία Επεξεργασία*

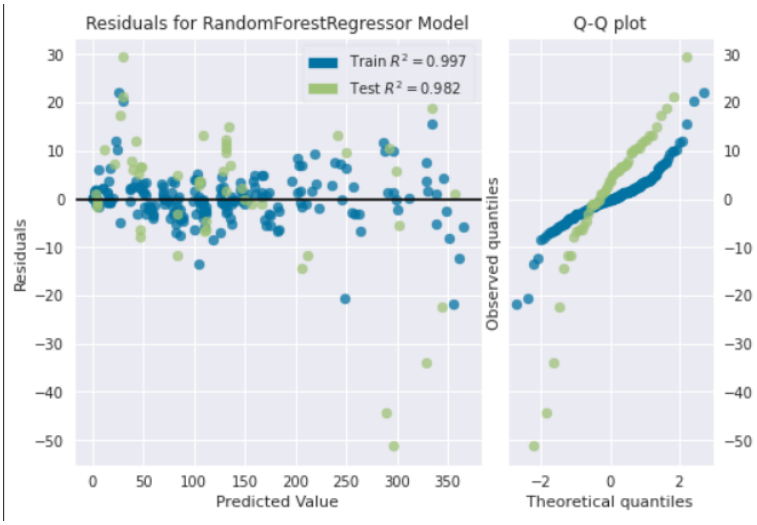
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για το Παρίσι, εμφανίζεται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος σχηματίζουν και ακολουθούν, όπως επίσης μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων μπορούν εύκολα να γίνουν όσα προαναφέρθηκαν αντιληπτά. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται έντονη ισχύς, λόγω της μεγάλης κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από μελέτη του παραπάνω διαγράμματος, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία απέχουν από τα υπόλοιπα και τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.

Στο παρακάτω διάγραμμα, **Σχήμα 155**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και απεικονίζονται και οι πραγματικές τιμές των θανάτων για την πόλη του Παρισιού και για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος έχει αποδώσει αρκετά καλά, αν και εμφανίζει ορισμένες αστοχίες. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως το διάστημα 24/11/2020 έως 25/11/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα. Έτσι, ορισμένα προβλεπόμενα σημεία φαίνεται ότι δεν προσαρμόζονται πλήρως στις πραγματικές τιμές των θανάτων. Εν τούτοις, πρόκειται για μία πολύ ικανοποιητική προσαρμογή του αλγορίθμου στα δεδομένα του Παρισιού.



Σχήμα 155 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, RF, Παρίσι, Ιδία Επεξεργασία*

Ακολουθεί το διάγραμμα των υπολοίπων, στο οποίο παρουσιάζονται τα υπόλοιπα για το συγκεκριμένο μοντέλο και απεικονίζονται τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα ελέγχου. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 156 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων, φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, παρατηρούνται ορισμένα σημεία σε μακρινή απόσταση από τον οριζόντιο άξονα, γεγονός το οποίο φανερώνει ότι συναντώνται αρκετά πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων εμφανίζουν δύο ευθείες γραμμές οι οποίες δεν ταυτίζονται πλήρως, αλλά τέμνονται σε ένα σημείο. Επίσης, αυτές οι γραμμές, δεν εμφανίζουν την ίδια τάση και ροή. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων δεν ακολουθούν την ίδια κατανομή.

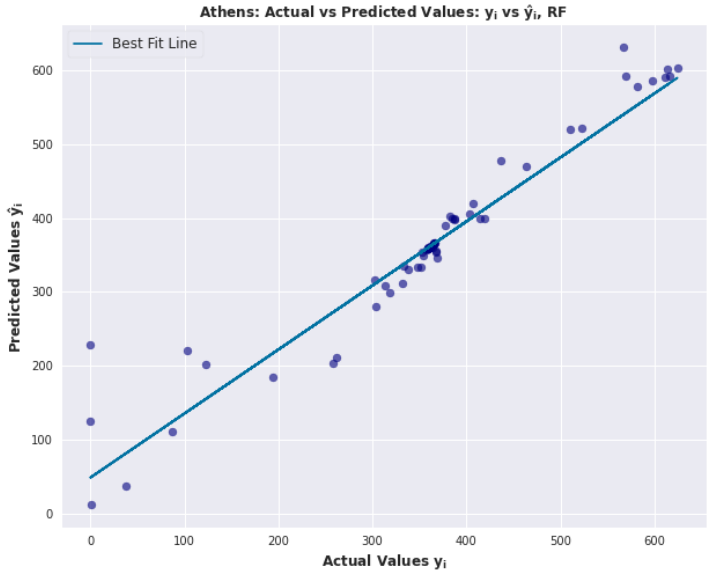
Στη συνέχεια, ακολουθεί η ανάλυση μοντέλων για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη της Μαδρίτης.

Ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

Μετρική Αξιολόγησης	**Αθήνα**
RMSE	42.94 (deaths per million)
R^2	0.914
EVS	0.916
MAE	19.87 (deaths per million)
MAPE	27.11%
MSE	0.00067

Πίνακας 49 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, RF, Αθήνα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε καλά πλαίσια. Η R^2 και EVS εμφανίζουν υψηλές τιμές. Είναι ίσες με 0.91 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν υψηλή συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο βάσει του μεγέθους των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 158Σχήμα 128**). Ακόμη, το ποσοστό από το MAPE είναι σχετικά υψηλό και ίσο με 27%, άρα θεωρείται μέτριο.



Σχήμα 157 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, RF, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, **Σχήμα 157**, παρατηρείται ύπαρξη γραμμικότητας. Παρατηρείται, δηλαδή, ότι τα σημεία του διαγράμματος σχηματίζουν μία ευθεία γραμμή, αλλά, επίσης, η κατανομή τους στο χώρο είναι πιο

«αφηρημένη» και δημιουργούνται συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι μέτρια προς έντονη. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, καθώς παρατηρούνται σημεία τα οποία απέχουν τόσο από την ευθεία ελαχίστων τετραγώνων όσο και από τις συγκεντρώσεις των υπόλοιπων σημείων.

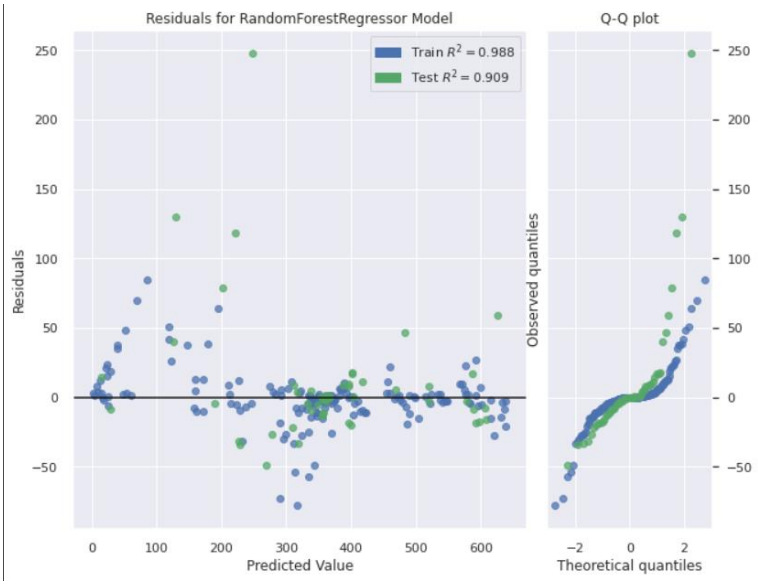
Σχήμα 158 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, RF, Αθήνα, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα, **Σχήμα 158**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων, μαζί με τις πραγματικές τους τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, για την πόλη της Αθήνας, είναι ότι το RF μοντέλο εντοπίζει την τάση των πραγματικών τιμών. Εν τούτοις, αδυνατεί να προβλέψει πλήρως τις ακριβείς τιμές, για ορισμένα μέγιστα και για ορισμένα ελάχιστα. Έτσι, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα. Σε ορισμένες περιπτώσεις φαίνεται να υπερεκτιμάει, έστω και για μικρή διαφορά τιμών, και σε άλλες φαίνεται να υποτιμάει, εμφανίζεται μεγαλύτερη διαφορά τιμών, τον προβλεπόμενο αριθμό κρουσμάτων. Συμπεραίνεται, ότι η προσαρμογή του μοντέλου πρόβλεψης θανάτων, δεν είναι αρκετά ικανοποιητική για την περίπτωση του RF και της πόλης της Αθήνας.

Στο διάγραμμα των υπολοίπων το οποίο ακολουθεί, παρουσιάζεται η κατανομή των υπολοίπων. Εκείνα, υπολογίζονται για το συγκεκριμένο μοντέλο και απεικονίζονται και για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Από το διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επιπλέον, εμφανίζονται ορισμένα σημεία μακριά από τη συγκέντρωση των υπολοίπων και μακριά από τον άξονα y. Πρόκειται για πιθανά ιδιάζοντα σημεία.

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα σημεία εμφανίζουν δύο ευθείες γραμμές οι οποίες επικαλύπτονται στο μεγαλύτερό τους τμήμα. Ως εκ τούτου, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.



Σχήμα 159 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, RF, Αθήνα, Ιδία Επεξεργασία*

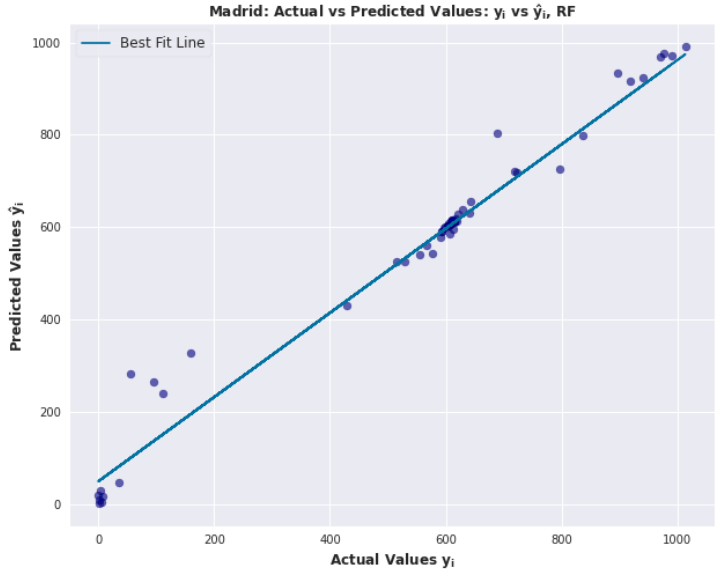
Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης θανάτων για την πόλη της Μαδρίτης.

Μετρική Αξιολόγησης	**Μαδρίτη**
RMSE	55.22 (deaths per million)
R^2	0.958
EVS	0.960
MAE	25.82 (deaths per million)
MAPE	5.37%
MSE	0.00110

Πίνακας 50 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, RF, Μαδρίτη, Ιδία Επεξεργασία*

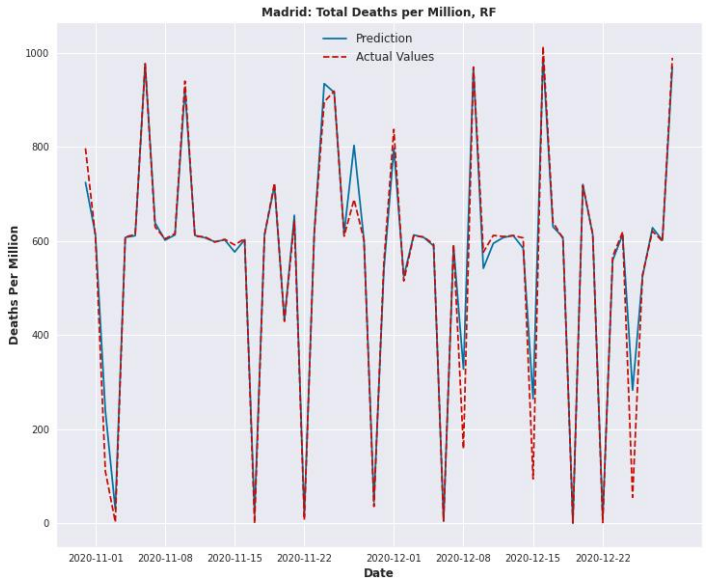
Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την Μαδρίτη, κυμαίνονται σε ικανοποιητικά πλαίσια. Η R^2 και EVS εμφανίζουν υψηλές τιμές. Έχουν τιμή μεγαλύτερη από 0.90, και ως εκ τούτου θεωρείται ότι εμφανίζεται υψηλό μέτρο επίδρασης, δηλαδή ότι η συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη είναι υψηλή (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 161**). Ακόμη, το ποσοστό από το MAPE βρίσκεται κοντά στο 5%, άρα, θεωρείται αρκετά καλό. Να σημειωθεί ότι το MSE του εν λόγω μοντέλου, είναι ο μεγαλύτερος αριθμός ο οποίος συναντάται για τα μοντέλα πρόβλεψης θανάτων με χρήση RF.

Από το διάγραμμα διασποράς, **Σχήμα 160**, ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, παρατηρείται γραμμικότητα ανάμεσα στα σημεία. Αυτό μπορεί να γίνει αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι μέτρια προς έντονη. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται ορισμένα ιδιάζοντα σημεία, τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 160 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, RF, Μαδρίτη, Ιδία Επεξεργασία*

Στο **Σχήμα 161** απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Πράγας και για το μοντέλο RF, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών σε ένα καλό επίπεδο, εν τούτοις αδυνατεί να προβλέψει πλήρως την ακριβή τιμή για ορισμένα μέγιστα και για ορισμένα ελάχιστα. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές των θανάτων. Χαρακτηριστικό είναι το παράδειγμα για τις ημερομηνίες 24/11/2020 έως και περίπου 25/11/2020, κατά το οποίο υπερεκτιμάει τις τιμές των θανάτων.



Σχήμα 161 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, RF, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές, αυτό το οποίο

μπορεί να γίνει θεωρηθεί είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία. Τα σημεία αυτά, όπως φαίνεται, βρίσκονται μακριά από τον οριζόντιο άξονα και μακριά από συγκεντρώσεις των υπόλοιπων σημείων.

Σχήμα 162 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, RF, Μαδρίτη, Ιδία Επεξεργασία*

Τέλος, από το διάγραμμα Q-Q παρατηρείται ότι τα υπόλοιπα για τα δεδομένα εκπαίδευσης και για τα δεδομένα ελέγχου εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερό τους τμήμα. Έτσι, θεωρείται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τέταρτη πόλη ανάλυσης αποτελεί η πόλη της Μόσχας. Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

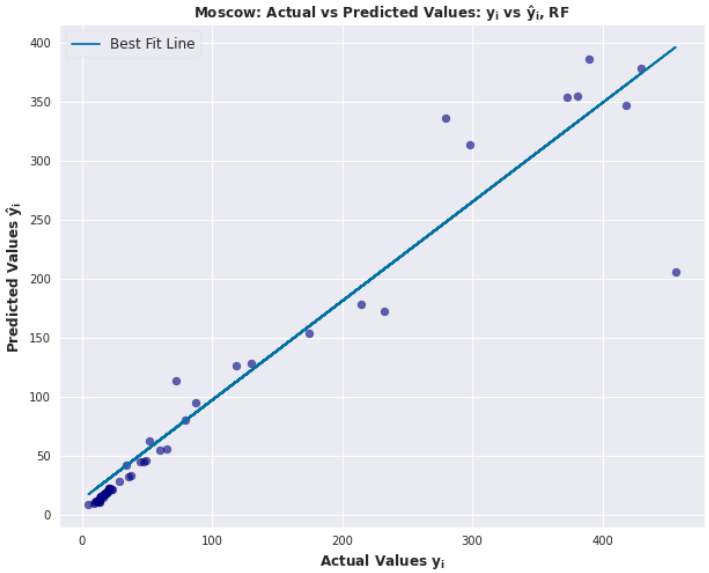
Μετρική Αξιολόγησης	**Μόσχα**
RMSE	35.02 (deaths per million)
R^2	0.925
EVS	0.928
MAE	13.03 (deaths per million)
MAPE	0.11%
MSE	0.00044

Πίνακας 51 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, RF, Μόσχα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε ικανοποιητικά επίπεδα. Η R^2 και EVS εμφανίζουν υψηλή τιμή. Η τιμή για το R^2 βρίσκεται κοντά στο 0.92 και ως εκ τούτου οι ανεξάρτητες μεταβλητές έχουν υψηλό μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 164**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό και μικρότερο από 1% και θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο.

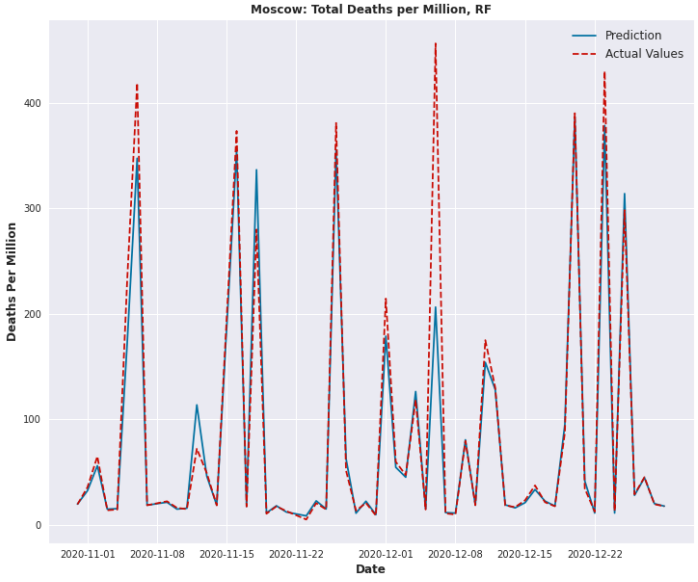
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μόσχας, παρατηρείται γραμμικότητα. Γραμμικότητα εμφανίζεται καθώς είναι εφικτή η προσαρμογή της ευθείας ελαχίστων τετραγώνων στα σημεία. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την

αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι αρκετά μέτρια προς έντονη. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται ορισμένα ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τις συγκεντρώσεις των υπόλοιπων σημείων και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 163 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, RF, Μόσχα, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα, **Σχήμα 164**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι έχει γίνει μία καλή προσαρμογή του αλγορίθμου στις πραγματικές τιμές των θανάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει την τάση και τη ροή των κρουσμάτων, όμως, αδυνατεί να υπολογίσει τις πραγματικές τιμές, κατά κύριο λόγο, σε ορισμένα μέγιστα. Η διαφοροποίηση αυτή των τιμών, έχει να κάνει με ένα πολύ μικρό χρονικό διάστημα, όμως στην περίπτωση της 06/12/2020, συναντάται μία σημαντική απόκλιση ανάμεσα σε προβλεπόμενη και πραγματική τιμή. Σε γενικά πλαίσια, η προσαρμογή του αλγορίθμου σε αυτήν την περίπτωση είναι καλή.

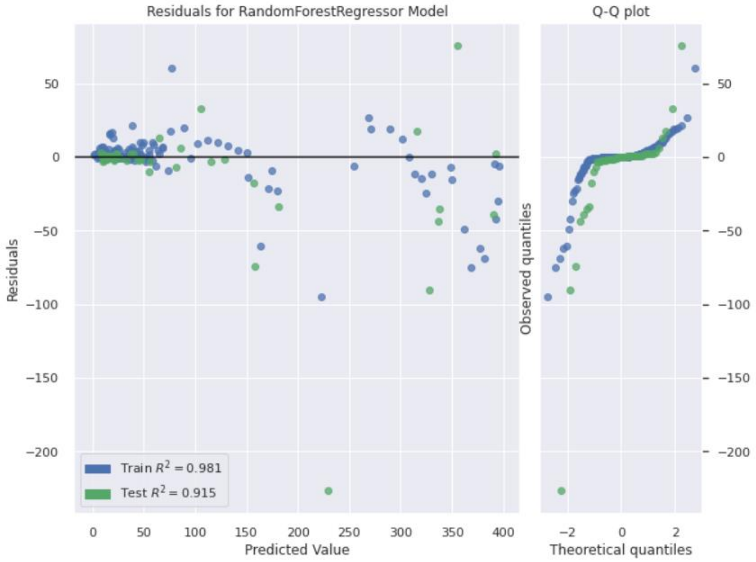


Σχήμα 164 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, RF, Μόσχα, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου και απεικονίζονται τόσο για το train set όσο και για το test set. Ακόμη, παρουσιάζεται το διάγραμμα Q-Q.

Από το διάγραμμα υπολοίπων, **Σχήμα 165**, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ακόμη, παρατηρούνται ορισμένα σημεία με μεγάλες τιμές, κυρίως αρνητικές, μακριά από την ευθεία y=0. Ίσως να πρόκειται για ιδιάζοντα σημεία, καθώς πρόκειται για σημεία τα οποία βρίσκονται μακριά από το μοτίβο και τη συγκέντρωση των υπόλοιπων σημείων.

Τέλος, από το διάγραμμα Q-Q, παρατηρείται ότι τα σημεία των υπολοίπων για τα δύο σετ εμφανίζουν δύο ευθείες γραμμές, οι οποίες επικαλύπτονται στο μεγαλύτερό τους μέρος. Συμπεραίνεται ότι τα δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.



Σχήμα 165 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, RF, Μόσχα, Ιδία Επεξεργασία*

Στο τελικό μοντέλο ανάλυσης για πρόβλεψη θανάτων με τον αλγόριθμο RF, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Πράγας.

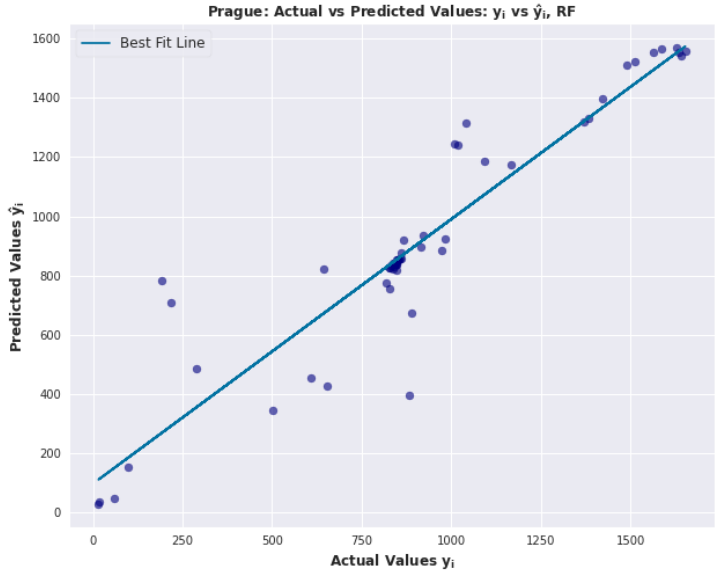
Μετρική Αξιολόγησης	**Πράγα**
RMSE	156.68 (deaths per million)
R^2	0.856
EVS	0.857
MAE	83.24 (deaths per million)
MAPE	0.22%
MSE	0.00886

Πίνακας 52 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, RF, Πράγα, Ιδία Επεξεργασία*

Βάσει όσων αναφέρθηκαν και στην υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Πράγας, κυμαίνονται σε καλά πλαίσια. Η τιμή της R^2 και της EVS είναι υψηλές, όχι όμως τόσο υψηλές όσο εκείνες των προηγούμενων μοντέλων. Η τιμή της R^2 βρίσκεται κοντά στο 0.85, άρα η συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη θεωρείται υψηλή. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 167**).

Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό και μικρότερο από 1%, άρα θεωρείται πολύ καλό.

Από το διάγραμμα διασποράς το οποίο ακολουθεί, **Σχήμα 166**, ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Πράγας, παρατηρείται ότι υπάρχει γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται μέτρια προς έντονη ισχύς, λόγω της σχετικά μεγάλης κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται λίγα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν έτσι, καθώς απέχουν αρκετά από την ευθεία ελαχίστων τετραγώνων και από συγκεντρώσεις άλλων σημείων.



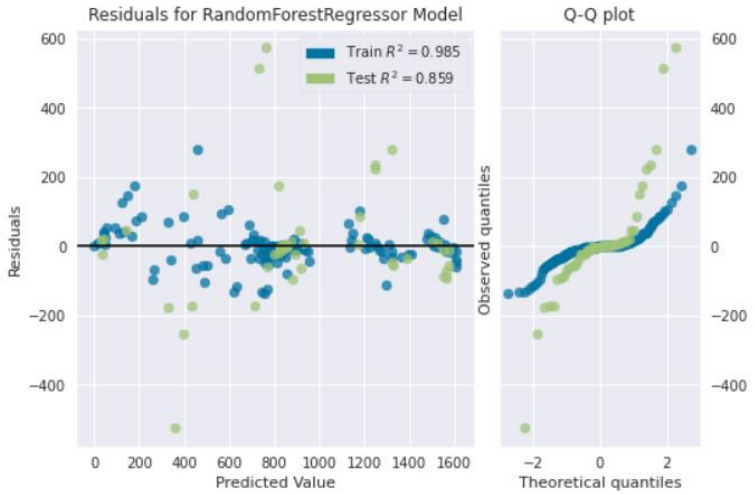
Σχήμα 166 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, RF, Πράγα, Ιδία Επεξεργασία*

Σχήμα 167 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, RF, Πράγα, Ιδία Επεξεργασία*

Στο διάγραμμα των προβλεπόμενων τιμών, **Σχήμα 167**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο.

Σε γενικές γραμμές παρατηρείται ότι η απόδοση του αλγορίθμου είναι ικανοποιητική, όμως εμφανίζει αρκετές αστοχίες. Δηλαδή, δεν είναι ικανός να ακολουθήσει πιστά την τάση των πραγματικών τιμών. Ούτε μπορεί να προβλέψει με πλήρη επιτυχία ορισμένα από τα μέγιστα, αλλά και ορισμένα από τα ελάχιστα. Άρα, όπως παρατηρείται και από το διάγραμμα, αρκετά σημεία δεν έχουν προσαρμοστεί στις πραγματικές τιμές των θανάτων.

Στη συνέχεια, παρουσιάζονται τα υπόλοιπα για το train set αλλά και για το test set. Ακόμη, παρουσιάζεται και το διάγραμμα Q-Q.



Σχήμα 168 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, RF, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι ένα πλήθος τους βρίσκεται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Τέλος, εμφανίζονται σημεία τα οποία απέχουν από τον οριζόντιο άξονα και τα οποία θα μπορούσε κάποιος να τα χαρακτηρίσει ως ιδιάζοντα.

Από το διάγραμμα Q-Q, φαίνεται ότι τα σημεία των δύο σετ δεδομένων, εμφανίζουν δύο σχεδόν ταυτιζόμενες ευθείες γραμμές, οι οποίες διαφοροποιούνται σε ένα μικρό εύρος. Θεωρείται, λοιπόν, ότι τα σημεία των δύο σετ δεδομένων ακολουθούν την ίδια κατανομή.

Τα αποτελέσματα των μοντέλων πρόβλεψης θανάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα Ε.

4.4.6 XGBoost Regression

Έκτο και τελευταίο μοντέλο το οποίο εφαρμόστηκε στα υπάρχοντα δεδομένα για τις εννέα διαφορετικές πόλεις, είναι το XGBoost Regression. Στις επόμενες σελίδες ακολουθούν τα αποτελέσματα του αλγορίθμου για την Αθήνα, τη Μαδρίτη, τη Μόσχα, το Παρίσι και την Πράγα.

4.4.6.1 Πρόβλεψη Κρουσμάτων

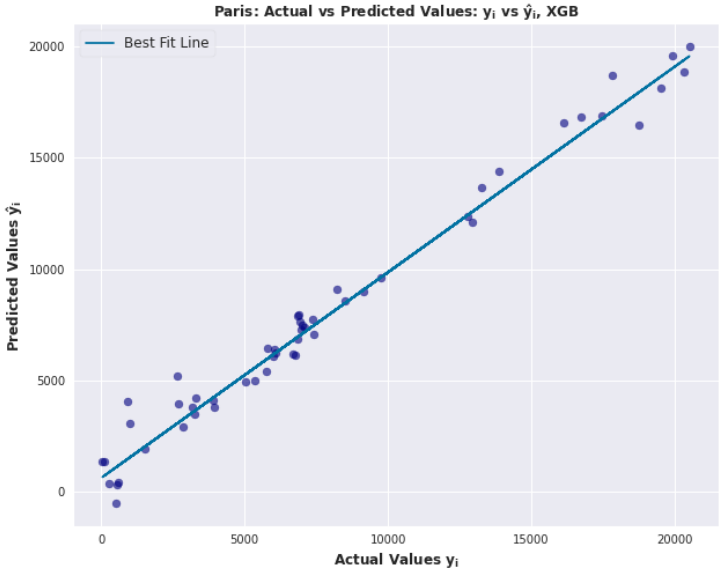
Εκείνο το μοντέλο το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των κρουσμάτων με το XGBoost, είναι εκείνο για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	Παρίσι
RMSE	932.58 (cases per million)
R^2	0.976
EVS	0.977
MAE	667.36 (cases per million)

MAPE	0.91%
MSE	0.00029

Πίνακας 53 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, XGBoost, Παρίσι, Ιδία Επεξεργασία*

Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Παρισιού, είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS είναι αρκετά υψηλές και ίσες με 0.97, άρα πρόκειται για υψηλή συσχέτιση μεταβλητών. Δηλαδή, σημαίνει ότι 97% της μεταβολής της εξαρτημένης μεταβλητής μπορεί να εξηγηθεί από τις ανεξάρτητες μεταβλητές. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές τα περίπου 20000 κρούσματα (Βλέπε **Σχήμα 170**). Ακόμη το ποσοστό από το MAPE είναι πάρα πολύ χαμηλό και μικρότερο από 1%, άρα θεωρείται πολύ καλό. Τέλος, όπως θα παρουσιαστεί και στη συνέχεια, πρόκειται για το μικρότερο MSE το οποίο συναντάται στην ανάλυση μοντέλων πρόβλεψης κρουσμάτων του XGBoost.

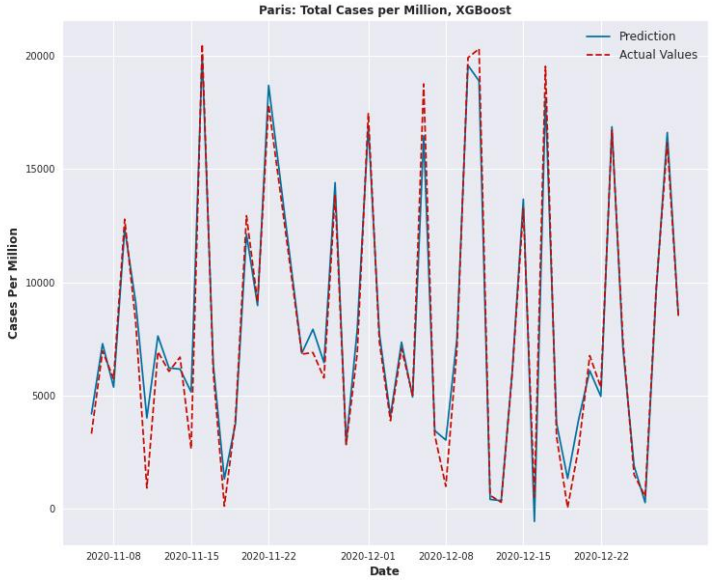


Σχήμα 169 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, XGBoost, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη του Παρισιού, **Σχήμα 169**, εμφανίζεται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν, να ακολουθούν και να μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η γραμμικότητα αυτή. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Στην περίπτωση του Παρισιού, θεωρείται ότι συναντάται έντονη ισχύς, λόγω της μεγάλης κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, εντοπίζονται ελάχιστα μεμονωμένα σημεία στα οποία θα μπορούσε να αποδοθεί ο χαρακτηρισμός αυτός.

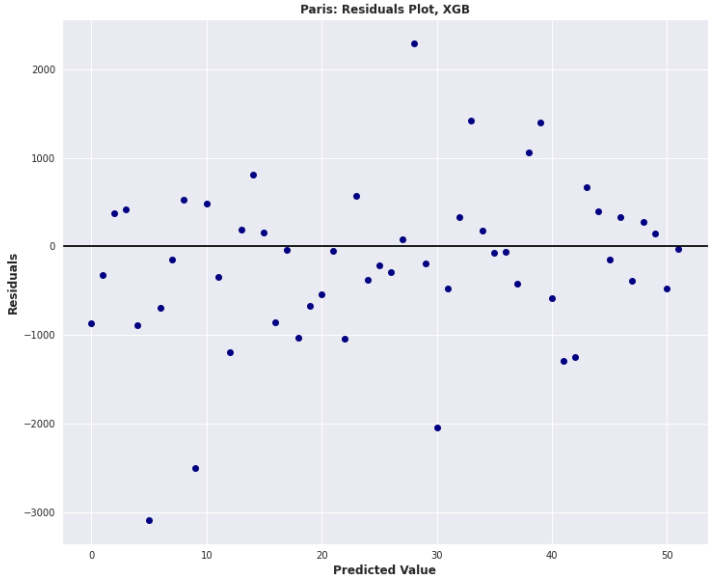
Στο παρακάτω διάγραμμα, **Σχήμα 170**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και απεικονίζονται και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος XGBoost έχει αποδώσει αρκετά ικανοποιητικά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες μικρές εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως το διάστημα 09/11/2020 – 15/11/2020. Εμφανίζονται σημεία τα οποία δεν έχουν

προσαρμοστεί πλήρως στις πραγματικές τιμές των κρουσμάτων, όμως οι διαφορές τους από τις πραγματικές τιμές είναι πολύ μικρές. Σε γενικές γραμμές, όμως, πρόκειται για μία αξιόλογη προσαρμογή του μοντέλου στις πραγματικές τιμές.



Σχήμα 170 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, XGBoost, Παρίσι, Ιδία Επεξεργασία*

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα, τα οποία υπολογίζονται για το συγκεκριμένο μοντέλο. Για το XGBoost Regression και για την απεικόνιση των υπολοίπων, όπως γίνεται αντιληπτό, δεν χρησιμοποιήθηκε η βιβλιοθήκη της Yellowbrick. Έτσι, ο υπολογισμός των υπολοίπων προέκυψε από τη διαφορά των πραγματικών τιμών του συνόλου ελέγχου με τις προβλεπόμενες τιμές.



Σχήμα 171 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Παρατηρείται επίσης, ότι ορισμένα σημεία υπολοίπων βρίσκονται σε σημαντική απόσταση από την ευθεία y=0. Γεγονός το οποίο φανερώνει ότι πρόκειται για πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού.

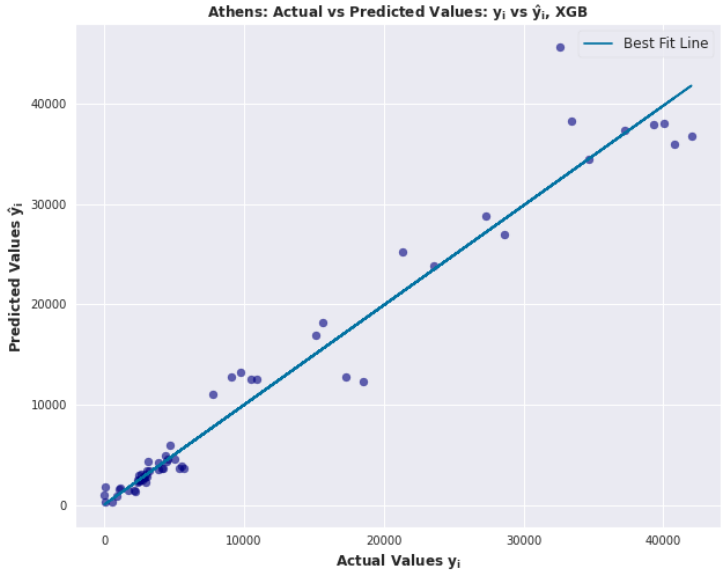
Στη συνέχεια, ακολουθεί η ανάλυση για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη του Παρισιού.

Όπως στην ανάλυση η οποία προηγήθηκε, έτσι ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

Μετρική Αξιολόγησης	Αθήνα
RMSE	2600.44 (cases per million)
R^2	0.957
EVS	0.957
MAE	1480.47 (cases per million)
MAPE	2.35%
MSE	0.00222

Πίνακας 54 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, XGBoost, Αθήνα, Ιδία Επεξεργασία*

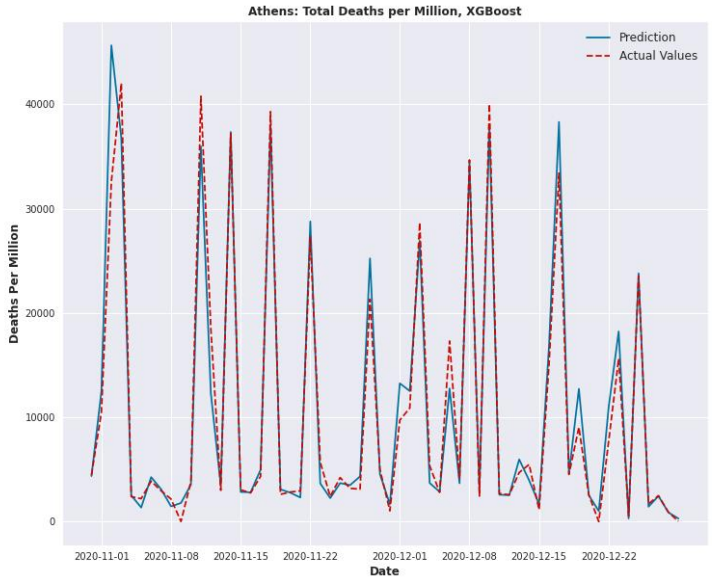
Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε πολύ καλά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν υψηλές τιμές. Είναι ίσες με 0.95 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν πολύ υψηλή συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, βάσει του μεγέθους των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 173**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό, μικρότερο του 3%, άρα θεωρείται πολύ καλό. Το MSE και σε αυτήν την περίπτωση εμφανίζει χαμηλή τιμή.



Σχήμα 172 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, XGBoost, Αθήνα, Ιδία Επεξεργασία*

Στο **Σχήμα 172**, παρατίθεται το διάγραμμα διασποράς για την πόλη της Αθήνας. Από το παρακάτω διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, παρατηρείται ύπαρξη γραμμικότητας. Παρατηρείται, δηλαδή, ότι από τα σημεία του διαγράμματος φαίνεται να μπορεί να περάσει μία ευθεία γραμμή. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι ισχυρή.

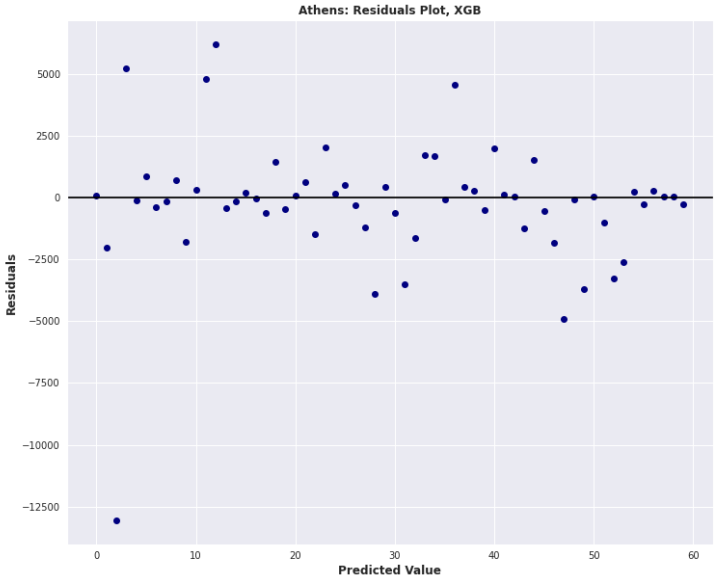
Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά πιθανά ιδιάζοντα σημεία.



Σχήμα 173 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, XGBoost, Αθήνα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων, μαζί με τις πραγματικές τιμές των κρουσμάτων, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για τον αλγόριθμο XGBoost, για την πόλη της Αθήνας, είναι ότι εντοπίζει την τάση των πραγματικών τιμών, εν τούτοις, αδυνατεί να προβλέψει πλήρως ορισμένα μέγιστα αλλά και ορισμένα ελάχιστα. Βέβαια, η αστοχία αυτή φαίνεται να εμφανίζει πολύ μικρές τιμές ως διαφορά. Ακόμη, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα, όπως για παράδειγμα το διάστημα ανάμεσα στη 09/11/2020 έως 10/11/2020. Σε γενικές γραμμές, η προσαρμογή του μοντέλου στα πραγματικά δεδομένα είναι αρκετά ικανοποιητική.

Στο παρακάτω διάγραμμα παρουσιάζονται τα υπόλοιπα. Σε κάθε περίπτωση από εδώ και στο εξής, ο υπολογισμός των υπολοίπων προκύπτει από τη διαφορά των πραγματικών τιμών του συνόλου ελέγχου με τις προβλεπόμενες τιμές.



Σχήμα 174 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επιπροσθέτως, παρατηρείται ότι ένα σημείο απέχει σημαντικά από την ευθεία y=0, ως προς τα αρνητικά. Πρόκειται πιθανότατα για ένα ιδιάζον σημείο.

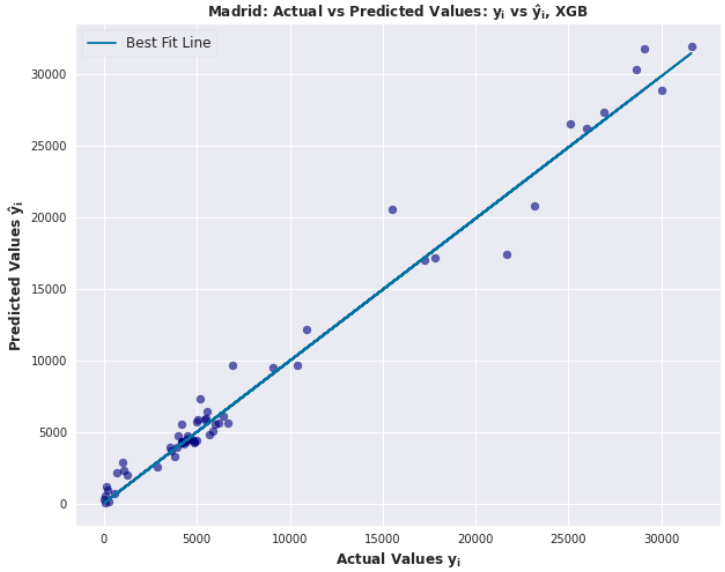
Τρίτη πόλη ανάλυσης αποτελεί η πόλη της Μαδρίτης. Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης κρουσμάτων, καθώς επίσης παρουσιάζονται και τα γραφήματα των αποτελεσμάτων.

Μετρική Αξιολόγησης	**Μαδρίτη**
RMSE	1279.86 (cases per million)
R^2	0.978
EVS	0.9780
MAE	834.35 (cases per million)
MAPE	0.71%
MSE	0.00054

Πίνακας 55 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν τη Μαδρίτη, κυμαίνονται σε σχετικά ικανοποιητικά και αποδεκτά πλαίσια. Η R^2 και EVS εμφανίζουν αρκετά υψηλές τιμές. Έχουν τιμή μεγαλύτερη από 0.90, και ως εκ τούτου θεωρείται ότι εμφανίζεται ισχυρό μέτρο επίδρασης των ανεξάρτητων μετρικών στην εξαρτημένη μεταβλητή (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 176**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό, μικρότερο από 1%, άρα θεωρείται πολύ καλό, αφού όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο.

Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.

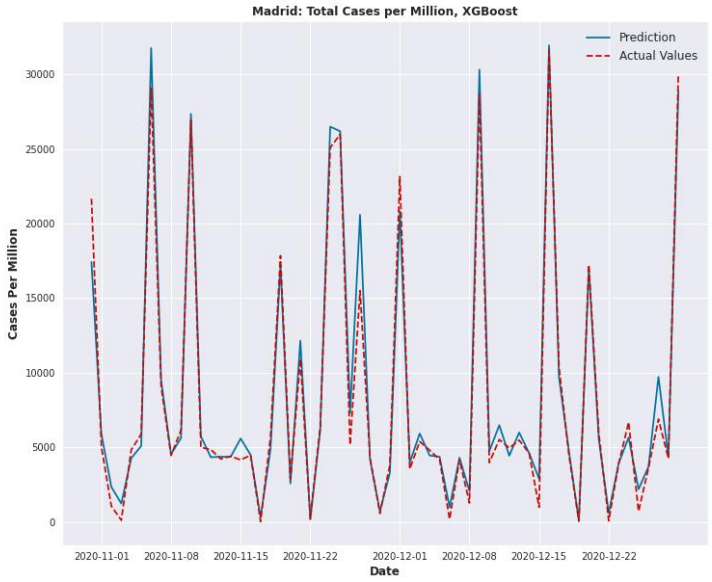


Σχήμα 175 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, παρατηρείται ότι υπάρχει γραμμικότητα. Δηλαδή, δύναται η προσαρμογή ευθείας γραμμής στα δεδομένα. Αυτό γίνεται αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον,

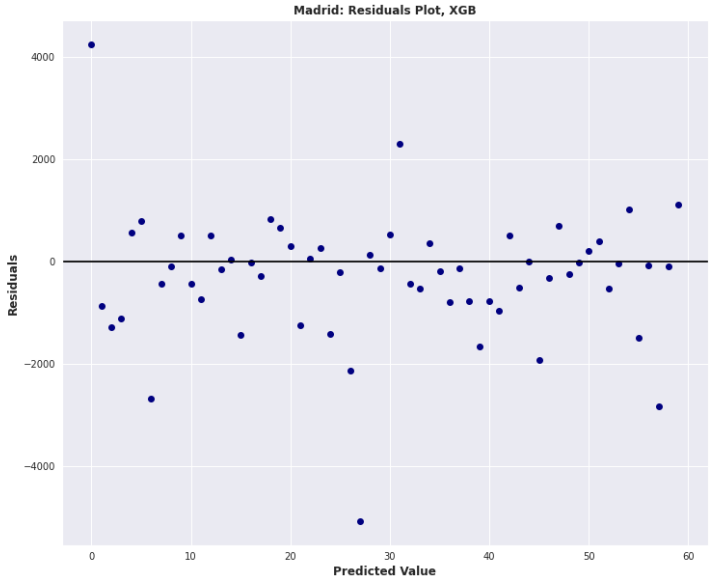
παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι έντονη. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται λίγα ιδιάζοντα σημεία.

Στο διάγραμμα το οποίο ακολουθεί, **Σχήμα 176**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Μαδρίτης και για το XGBoost, είναι ότι ο αλγόριθμος εντοπίζει την τάση και τη ροή των πραγματικών τιμών με ιδιαίτερη επιτυχία. Συναντώνται βέβαια περιπτώσεις κατά τις οποίες αδυνατεί να προβλέψει πλήρως την ακριβή τιμή για ορισμένα μέγιστα και για ορισμένες μεσαίες τιμές, όμως πρόκειται για ένα πολύ μικρό χρονικό διάστημα και για μικρή απόκλιση. Χαρακτηριστικό παράδειγμα είναι εκείνο από τις 9/12/2020 έως τις 15/12/2020. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές. Σε γενικές γραμμές, ο αλγόριθμος για την πόλη αυτή, φαίνεται να έχει κάνει καλή προσαρμογή.



Σχήμα 176 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του μοντέλου.



Σχήμα 177 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

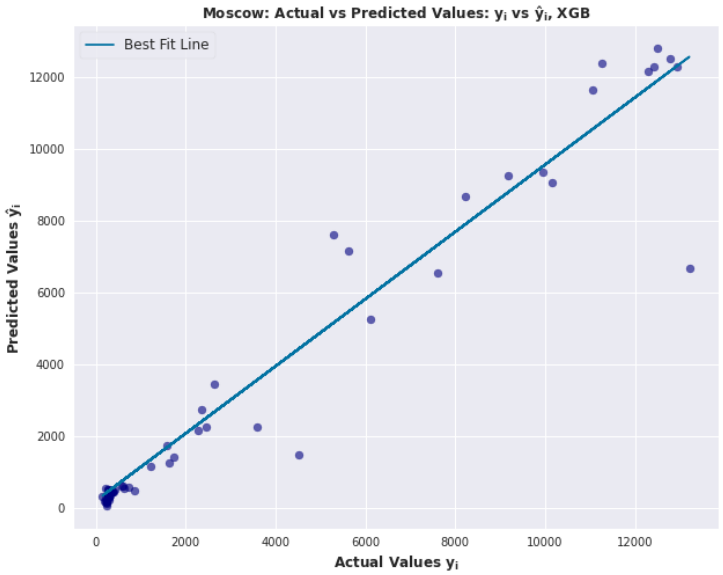
Από το διάγραμμα των υπολοίπων για το XGBoost μοντέλο της Μαδρίτης, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, παρατηρείται ότι ορισμένα υπόλοιπα εμφανίζουν μεγάλες τιμές, τόσο στον αρνητικό όσο και στο θετικό άξονα, μακριά από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές αυτό το οποίο μπορεί να γίνει αντιληπτό, είναι ότι συναντώνται ορισμένα πιθανά ιδιάζοντα σημεία.

Ακολουθεί η ανάλυση για το μοντέλο πρόβλεψης κρουσμάτων για την πόλη της Μόσχας. Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	**Μόσχα**
RMSE	1098.71 (cases per million)
R^2	0.939
EVS	0.940
MAE	484.69 (cases per million)
MAPE	0.26%
MSE	0.00040

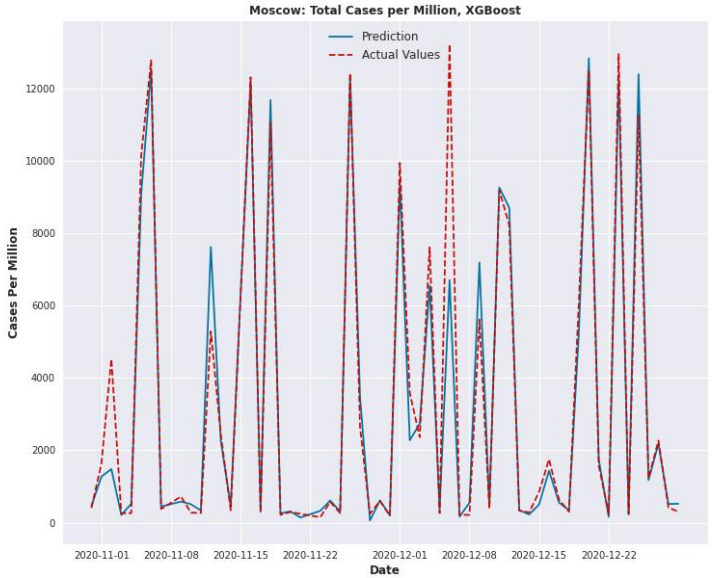
Πίνακας 56 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, XGBoost, Μόσχα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε υψηλά επίπεδα. Η R^2 και EVS εμφανίζουν μεγάλη τιμή. Έχουν τιμή κοντά στο 0.93, και ως εκ τούτου οι ανεξάρτητες μεταβλητές έχουν υψηλό μέτρο επίδρασης στην εξαρτημένη μεταβλητή (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των κρουσμάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 179Σχήμα 149**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό, μικρότερο από 1%, άρα θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο.



Σχήμα 178 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, XGBoost, Μόσχα, Ιδία Επεξεργασία*

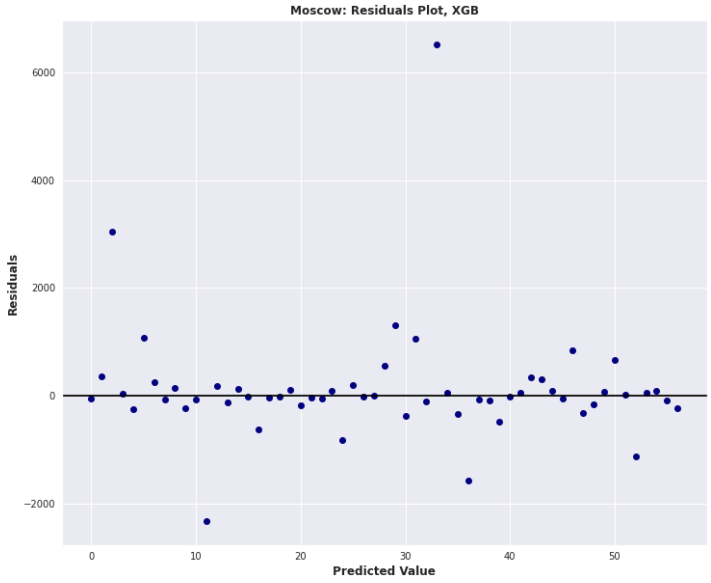
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Μόσχας, παρατηρείται ύπαρξη γραμμικότητας. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι έντονη. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται ορισμένα πιθανά ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία και από την ευθεία ελαχίστων τετραγώνων.



Σχήμα 179 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, XGBoost, Μόσχα, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι έχει γίνει αρκετά ικανοποιητική προσαρμογή του αλγορίθμου στις πραγματικές τιμές των κρουσμάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει την τάση και τη ροή των κρουσμάτων, όμως και εδώ συναντώνται περιπτώσεις κατά τις οποίες αδυνατεί πλήρως να προβλέψει την ακριβή τιμή. Για μία ακόμη φορά, η διαφορά ανάμεσα στην προβλεπόμενη και την πραγματική τιμή, όπως αποτυπώνεται και στο διάγραμμα, φαίνεται να είναι μικρή και για ένα αρκετά σύντομο χρονικό διάστημα. Συγκριτικά με τα προηγούμενα αναλυθέντα μοντέλα, αυτό το μοντέλο φαίνεται να μην έχει προσαρμοστεί πλήρως στα δεδομένα. Σε γενικές γραμμές, όμως, η προσαρμογή του μοντέλου στις πραγματικές τιμές είναι αρκετά ικανοποιητική.

Κλείνοντας την ανάλυση για το μοντέλο πρόβλεψης κρουσμάτων για την πόλη της Μόσχας και με τον αλγόριθμο XGBoost Regression, ακολουθεί το διάγραμμα των υπολοίπων του μοντέλου. Από το διάγραμμα των υπολοίπων, **Σχήμα 180**, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, εμφανίζονται λίγα σημεία με σημαντικά αρνητικές και θετικές τιμές. Τα σημεία αυτά πιθανότατα να μπορούσαν να χαρακτηριστούν ως ιδιάζοντα.



Σχήμα 180 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Μόσχα, Ιδία Επεξεργασία*

Για να ολοκληρωθεί η ανάλυση των δημιουργηθέντων XGBoost μοντέλων για πρόβλεψη κρουσμάτων στις μελετώμενες πόλεις, χρειάζεται να παρατεθούν τα αποτελέσματα τα οποία προκύπτουν για την πόλη της Πράγας. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

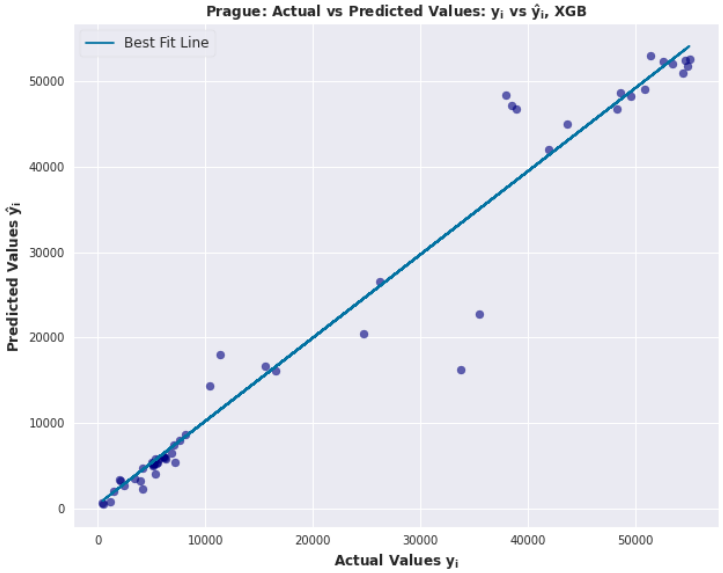
Μετρική Αξιολόγησης	**Πράγα**
RMSE	3934.90(cases per million)
R^2	0.961
EVS	0.961
MAE	1976.69 (cases per million)
MAPE	0.14%
MSE	0.00509

Πίνακας 57 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, XGBoost, Πράγα, Ιδία Επεξεργασία*

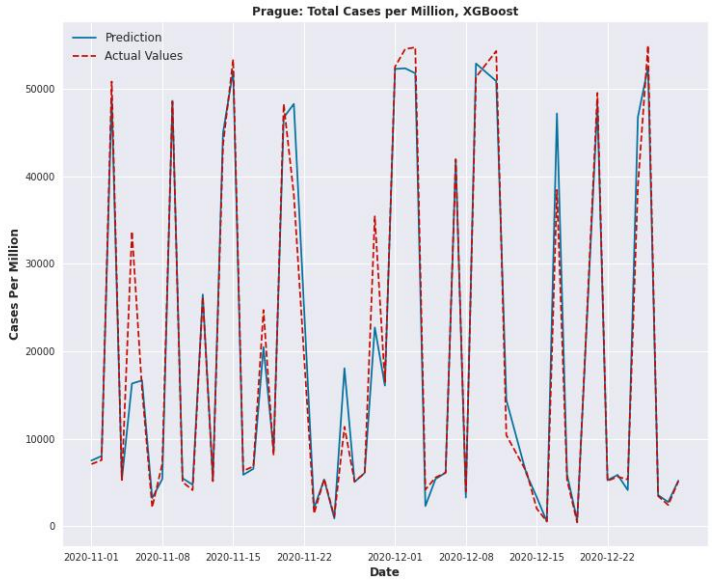
Βάσει όσων αναφέρθηκαν στην υποενότητα **4.2**, οι παραπάνω μετρικές αξιολόγησης είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.90, άρα η συσχέτιση των μεταβλητών θεωρείται υψηλή. Δηλαδή, το 96% των ανεξάρτητων μεταβλητών μπορεί να επηρεάσει την εξαρτημένη μεταβλητή. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τα κρούσματα ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 182**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό και μικρότερο από 1%, άρα θεωρείται πολύ καλό. Να σημειωθεί ότι για την πόλη της Πράγας συναντάται το μεγαλύτερο MSE σφάλμα, ανάμεσα στις 5 πόλεις οι οποίες αναλύονται.

Από το παρακάτω διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Πράγας, συναντάται γραμμικότητα. Αυτό σημαίνει ότι στα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή, αλλά εμφανίζεται και η δημιουργία συστάδων. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται έντονη ισχύς, λόγω της μεγάλης κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα

σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.



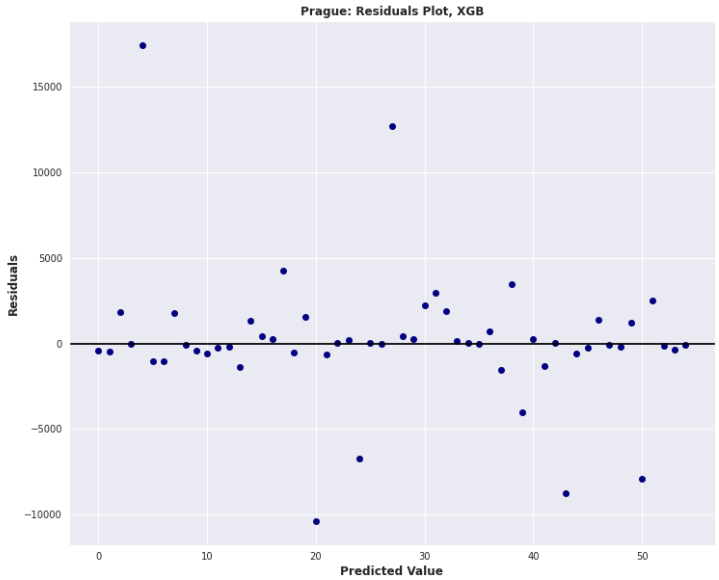
Σχήμα 181 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, XGBoost, Πράγα, Ιδία Επεξεργασία*



Σχήμα 182 *Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, XGBoost, Πράγα, Ιδία Επεξεργασία*

Στο **Σχήμα 182**, απεικονίζονται οι τιμές για την πρόβλεψη των κρουσμάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος έχει αποδώσει αρκετά καλά. Φαίνεται να ακολουθεί τη ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για ορισμένα μικρά χρονικά διαστήματα, όπως για παράδειγμα 04/11/2020 έως 06/11/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία, κατά κύριο λόγο, κάποια μέγιστα, αλλά και κάποια ελάχιστα. Υπάρχουν, λοιπόν, ορισμένα σημεία τα οποία δεν έχει καταφέρει το μοντέλο να προσαρμόσει στις πραγματικές τιμές των καταγεγραμμένων κρουσμάτων. Σημειώνεται, λοιπόν, ότι η προσαρμογή του αλγορίθμου είναι ικανοποιητική, όμως εμφανίζει ορισμένες αστοχίες.

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων.



Σχήμα 183 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ακόμη, συναντώνται σημεία τα οποία εμφανίζουν σημαντική απόσταση, μακριά από τον οριζόντιο άξονα. Πιθανότατα είναι ιδιάζοντα σημεία.

Τα αποτελέσματα των μοντέλων πρόβλεψης κρουσμάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα ΣΤ.

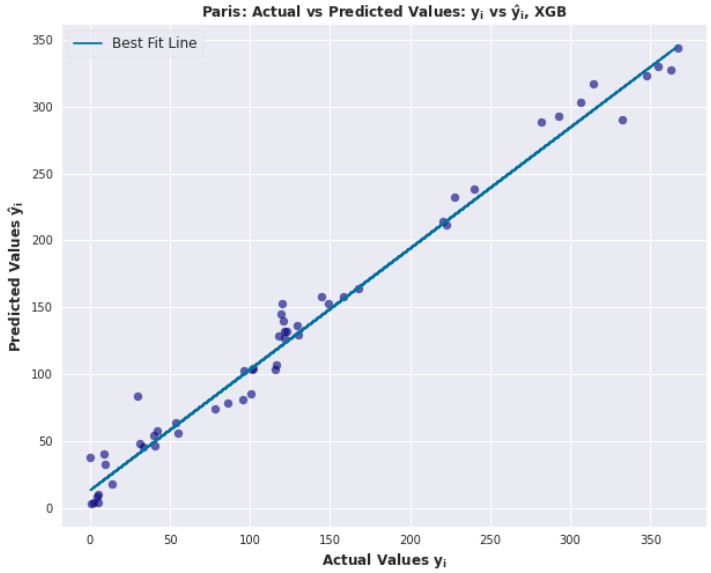
4.4.6.2 *Πρόβλεψη Θανάτων*

Το XGBoost μοντέλο το οποίο φαίνεται να ξεχωρίζει για την περίπτωση πρόβλεψης των θανάτων, είναι εκείνο για την πόλη του Παρισιού. Στη συνέχεια, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα.

Μετρική Αξιολόγησης	**Παρίσι**
RMSE	17.22 (deaths per million)
R^2	0.975
EVS	0.976
MAE	12.27 (deaths per million)
MAPE	2.14%
MSE	0.00011

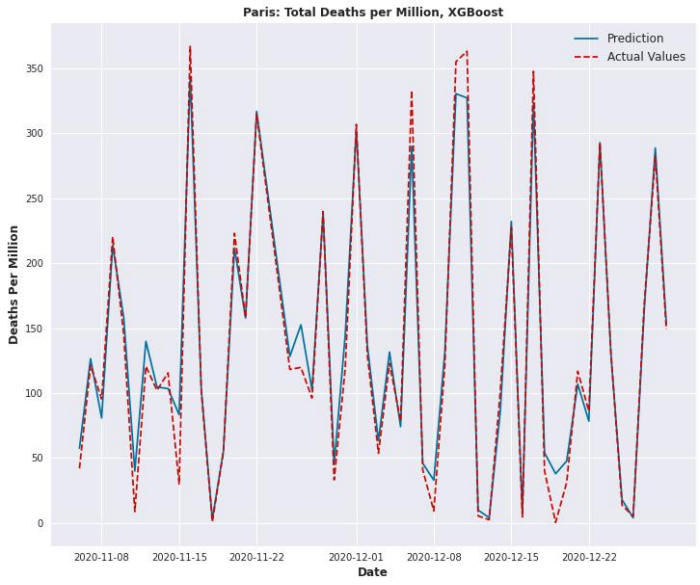
Πίνακας 58 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, XGBoost, Παρίσι, Ιδία Επεξεργασία*

Σύμφωνα με υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη του Πράγας, είναι αρκετά ικανοποιητικές. Η τιμή της R^2 και της EVS ξεπερνάει το 0.90, άρα η συσχέτιση των ανεξάρτητων μεταβλητών, με την εξαρτημένη μεταβλητή είναι αρκετά υψηλή. Δηλαδή, σημαίνει ότι 97% της μεταβολής της εξαρτώμενης μεταβλητής μπορεί να εξηγηθεί από τις ανεξάρτητες μεταβλητές. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών, μέγιστες τιμές 350 κρούσματα (Βλέπε **Σχήμα 185**). Ακόμη το ποσοστό από το MAPE είναι μικρότερο από 10%, άρα θεωρείται πολύ καλό. Τέλος, η τιμή του MSE είναι η χαμηλότερη τιμή η οποία συναντάται στα μοντέλα πρόβλεψης θανάτων με XGBoost.



Σχήμα 184 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, XGBoost, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για το Παρίσι, εμφανίζεται γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος σχηματίζουν και ακολουθούν, όπως επίσης μπορούν να προσαρμοσθούν πάνω σε μία ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων μπορούν εύκολα να γίνουν όσα προαναφέρθηκαν αντιληπτά. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται υψηλή ισχύς, λόγω της μεγάλης κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από μελέτη του παραπάνω διαγράμματος, είναι ότι εντοπίζονται ορισμένα σημεία τα οποία απέχουν από τα υπόλοιπα και τα οποία θα μπορούσαν να χαρακτηρισθούν ως ιδιάζοντα.

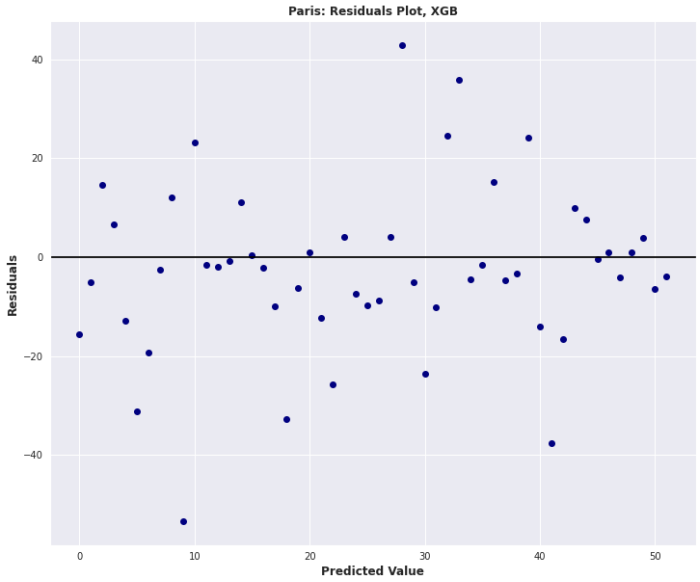


Σχήμα 185 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, XGBoost, Παρίσι, Ιδία Επεξεργασία*

Στο παραπάνω διάγραμμα, **Σχήμα 185**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και απεικονίζονται και οι πραγματικές τιμές των θανάτων για την πόλη του Παρισιού και για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι ο αλγόριθμος έχει αποδώσει αρκετά καλά, αν και εμφανίζει ορισμένες αστοχίες. Φαίνεται να ακολουθεί τη

ροή και τη τάση των πραγματικών τιμών, εμφανίζοντας βέβαια ορισμένες εξαιρέσεις για μερικά μικρά χρονικά διαστήματα, όπως το διάστημα 10/11/2020 έως 15/11/2020. Επιπροσθέτως, παρατηρείται ότι δεν μπορεί να προβλέψει με πλήρη επιτυχία κάποια μέγιστα, αλλά και κάποια ελάχιστα. Έτσι, ορισμένα προβλεπόμενα σημεία φαίνεται ότι δεν προσαρμόζονται πλήρως στις πραγματικές τιμές των θανάτων. Εν τούτοις, πρόκειται για μία πολύ ικανοποιητική προσαρμογή του αλγορίθμου στα δεδομένα του Παρισιού.

Ακολουθεί το διάγραμμα των υπολοίπων για το Παρίσι.



Σχήμα 186 *Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Παρίσι, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων, φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επίσης, παρατηρούνται ορισμένα σημεία σε μακρινή απόσταση από τον οριζόντιο άξονα, γεγονός το οποίο φανερώνει ότι συναντώνται αρκετά πιθανά ιδιάζοντα σημεία στην περίπτωση του μοντέλου αυτού.

Στη συνέχεια, ακολουθεί η ανάλυση μοντέλων για τις υπόλοιπες τέσσερις πόλεις. Η απόδοση του αλγορίθμου για τις εν λόγω πόλεις δεν είναι το ίδιο ικανοποιητική, συγκριτικά με την πόλη της Μαδρίτης.

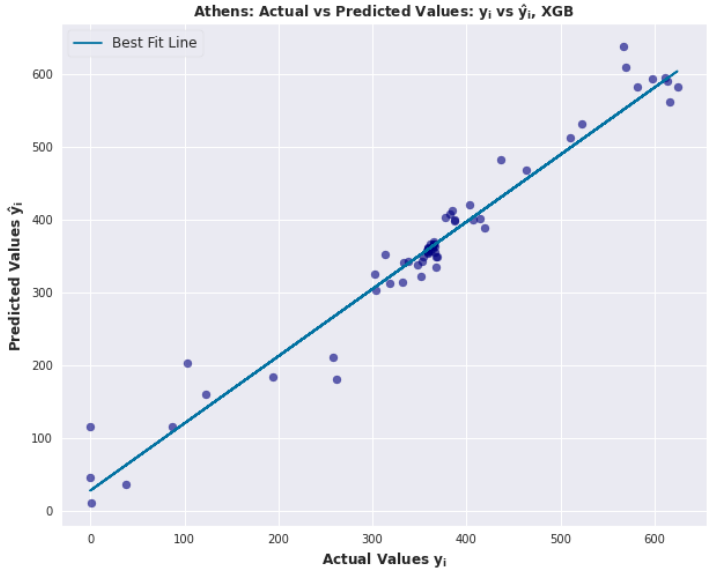
Ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Αθήνας.

Μετρική Αξιολόγησης	**Αθήνα**
RMSE	31.82 (deaths per million)
R^2	0.953
EVS	0.953
MAE	20.56 (deaths per million)
MAPE	11.42%
MSE	0.00037

Πίνακας 59 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, XGBoost, Αθήνα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Αθήνας, κυμαίνονται σε καλά πλαίσια. Η R^2 και EVS εμφανίζουν υψηλές τιμές. Είναι ίσες με 0.95 και ως εκ τούτου θεωρείται ότι οι μεταβλητές έχουν υψηλή συσχέτιση (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά

εκατομμύριο βάσει του μεγέθους των καταγεγραμμένων τιμών των κρουσμάτων (Βλέπε **Σχήμα 188**). Ακόμη, το ποσοστό από το MAPE είναι ίσο με 11%, άρα θεωρείται καλό. Να σημειωθεί ότι το MSE του εν λόγω μοντέλου, είναι ο δεύτερος μεγαλύτερος αριθμός ο οποίος συναντάται για τα μοντέλα πρόβλεψης θανάτων με χρήση XGBoost Regression.



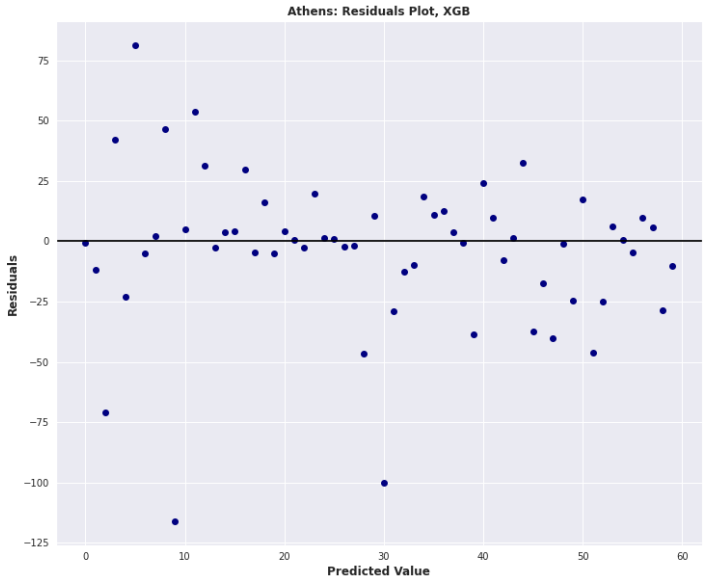
Σχήμα 187 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, XGBoost, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των κρουσμάτων ανά εκατομμύριο για την πόλη της Αθήνας, **Σχήμα 187**, παρατηρείται ύπαρξη γραμμικότητας. Παρατηρείται, δηλαδή, ότι τα σημεία του διαγράμματος σχηματίζουν μία ευθεία γραμμή, αλλά, επίσης, η κατανομή τους στο χώρο είναι πιο «αφηρημένη» και δημιουργούνται συστάδες. Όλα αυτά γίνονται αντιληπτά μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι μέτρια προς έντονη. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται μερικά ιδιάζοντα σημεία, καθώς παρατηρούνται σημεία τα οποία απέχουν τόσο από την ευθεία ελαχίστων τετραγώνων όσο και από τις συγκεντρώσεις των υπόλοιπων σημείων.

Στο παρακάτω διάγραμμα, **Σχήμα 188**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων, μαζί με τις πραγματικές τους τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, για την πόλη της Αθήνας, είναι ότι το XGBoost μοντέλο εντοπίζει την τάση των πραγματικών τιμών. Εν τούτοις, αδυνατεί να προβλέψει πλήρως τις ακριβείς τιμές, για ορισμένα μέγιστα και για ορισμένα ελάχιστα. Έτσι, υπάρχουν σημεία πρόβλεψης τα οποία δεν μπορεί να προσαρμόσει πλήρως στα πραγματικά δεδομένα. Σε ορισμένες περιπτώσεις φαίνεται να υπερεκτιμάει, έστω και για μικρή διαφορά τιμών, και σε άλλες φαίνεται να υποτιμάει, εμφανίζεται μεγαλύτερη διαφορά τιμών, τον προβλεπόμενο αριθμό κρουσμάτων. Συμπεραίνεται, ότι η προσαρμογή του μοντέλου πρόβλεψης θανάτων, δεν είναι άρτια, είναι, όμως, ικανοποιητική για την περίπτωση του XGBoost και της πόλης της Αθήνας.

Σχήμα 188 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, XGBoost, Αθήνα, Ιδία Επεξεργασία*

Στο διάγραμμα το οποίο ακολουθεί, παρουσιάζεται η κατανομή των υπολοίπων.



Σχήμα 189 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, XGBoost, Αθήνα, Ιδία Επεξεργασία*

Από το διάγραμμα φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επιπλέον, εμφανίζονται ορισμένα σημεία μακριά από τη συγκέντρωση των υπολοίπων και μακριά από τον άξονα y. Πρόκειται για πιθανά ιδιάζοντα σημεία.

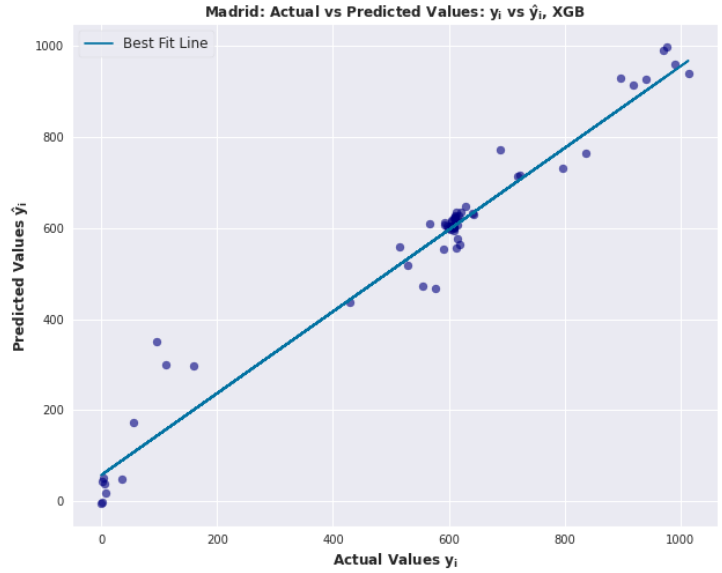
Στη συνέχεια, παρατίθεται ο Πίνακας των μετρικών για το μοντέλο πρόβλεψης θανάτων για την πόλη της Μαδρίτης.

Μετρική Αξιολόγησης	**Μαδρίτη**
RMSE	57.90 (deaths per million)
R^2	0.954
EVS	0.955
MAE	34.00 (deaths per million)
MAPE	1.75%
MSE	0.00121

Πίνακας 60 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης οι οποίες περιγράφουν την Μαδρίτη, κυμαίνονται σε ικανοποιητικά πλαίσια. Η R^2 και EVS εμφανίζουν υψηλές τιμές. Έχουν τιμή μεγαλύτερη από 0.90, και ως εκ τούτου θεωρείται ότι εμφανίζεται υψηλό μέτρο επίδρασης, δηλαδή ότι η συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη είναι υψηλή (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως είναι το MAE και RMSE, εμφανίζουν χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μαδρίτης (Βλέπε **Σχήμα 191**). Ακόμη, το ποσοστό από το MAPE βρίσκεται κοντά στο 2%, άρα, θεωρείται αρκετά καλό.

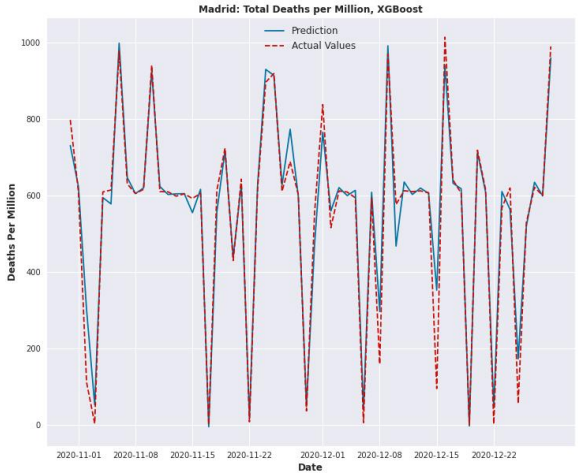
Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς για την πόλη της Μαδρίτης.



Σχήμα 190 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

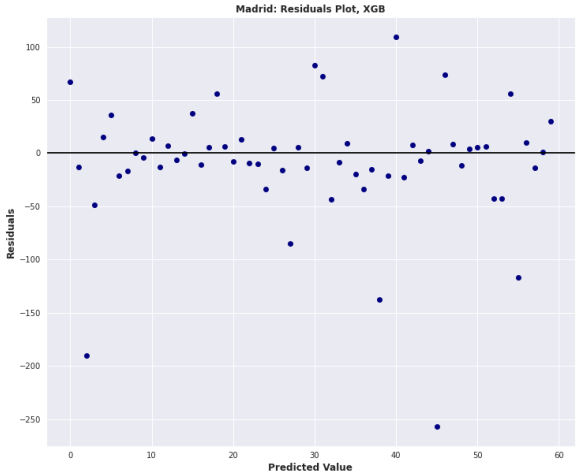
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μαδρίτης, παρατηρείται γραμμικότητα ανάμεσα στα σημεία. Αυτό μπορεί να γίνει αντιληπτό μέσω της παρεμβολής της ευθείας ελαχίστων τετραγώνων. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι μέτρια προς έντονη. Τέλος, μελετώντας το παραπάνω διάγραμμα θα μπορούσε κάποιος να θεωρήσει ότι συναντώνται ορισμένα ιδιάζοντα σημεία, τα οποία βρίσκονται αρκετά μακριά από τα υπόλοιπα σημεία και από την ευθεία ελαχίστων τετραγώνων.

Στο επόμενο διάγραμμα, **Σχήμα 191**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές τους ανά εκατομμύριο, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται για την πόλη της Πράγας και για το μοντέλο XGBoost, είναι ότι ο αλγόριθμος εντοπίζει την τάση των πραγματικών τιμών σε ένα καλό επίπεδο, εν τούτοις αδυνατεί να προβλέψει πλήρως την ακριβή τιμή για ορισμένα μέγιστα και για ορισμένα ελάχιστα. Έτσι, συναντώνται σημεία τα οποία δεν μπορεί να προσαρμόσει πλήρως στις πραγματικές τιμές των θανάτων. Χαρακτηριστικό είναι το παράδειγμα για τις ημερομηνίες 10/12/2020 έως και περίπου 15/12/2020, κατά το οποίο αστοχεί να προσαρμοστεί στις πραγματικές τιμές θανάτων.



Σχήμα 191 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

Στη συνέχεια, παρουσιάζεται το διάγραμμα των υπολοίπων του συγκεκριμένου μοντέλου.



Σχήμα 192 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι εκείνα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Επιπλέον, με μία παράλληλη δεύτερη ματιά στο διάγραμμα διασποράς ανάμεσα στις πραγματικές και τις προβλεπόμενες τιμές, αυτό το οποίο μπορεί να γίνει θεωρηθεί είναι ότι υπάρχουν ορισμένα ιδιάζοντα σημεία. Τα σημεία αυτά, όπως φαίνεται, βρίσκονται μακριά από τον οριζόντιο άξονα και μακριά από συγκεντρώσεις των υπόλοιπων σημείων.

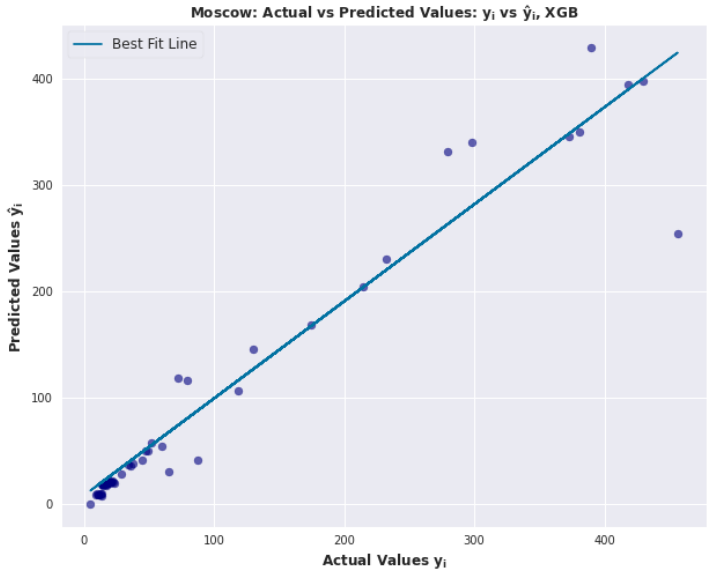
Τέταρτη πόλη ανάλυσης αποτελεί η πόλη της Μόσχας. Αρχικά, παρουσιάζεται ο Πίνακας με τις τιμές των μετρικών και εν συνεχεία, παρουσιάζονται τα δημιουργηθέντα γραφήματα.

Μετρική Αξιολόγησης	**Μόσχα**
RMSE	31.85 (deaths per million)
R^2	0.938
EVS	0.939
MAE	13.23 (deaths per million)
MAPE	0.17%
MSE	0.00037

Πίνακας 61 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, XGBoost, Μόσχα, Ιδία Επεξεργασία*

Οι μετρικές αξιολόγησης της Μόσχας, κυμαίνονται σε ικανοποιητικά επίπεδα. Η R^2 και EVS εμφανίζουν υψηλή τιμή. Η τιμή για το R^2 βρίσκεται κοντά στο 0.93 και ως εκ τούτου οι ανεξάρτητες μεταβλητές έχουν υψηλό μέτρο επίδρασης (Moore, D. S., Notz, W. I, & Flinger, M. A., 2013). Οι μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, όπως το MAE και RMSE, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας πάντοτε υπόψιν το μέγεθος των καταγεγραμμένων τιμών των θανάτων για την πόλη της Μόσχας (Βλέπε **Σχήμα 194**). Ακόμη, το ποσοστό από το MAPE είναι χαμηλό και μικρότερο από 1% και θεωρείται πολύ καλό, καθώς όσο πιο χαμηλές τιμές έχει το MAPE, τόσο περισσότερο ακριβές είναι το μοντέλο.

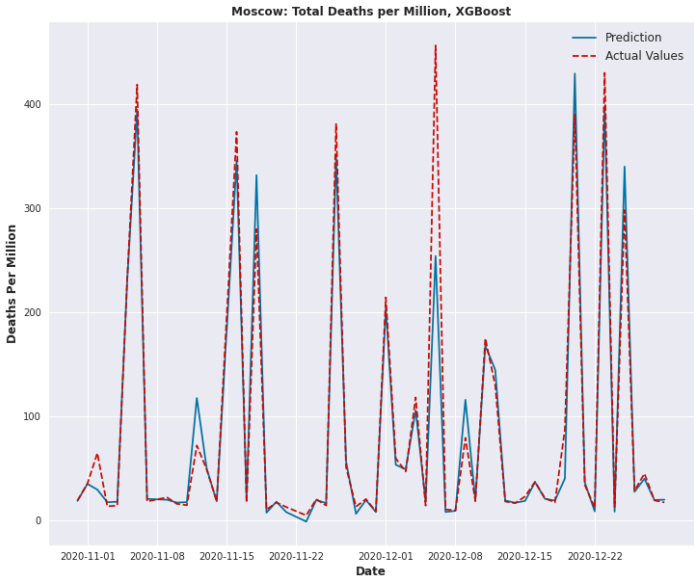
Στη συνέχεια, παρατίθεται το διάγραμμα διασποράς, όπως αυτό προκύπτει για τα δεδομένα της Μόσχας.



Σχήμα 193 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, XGBoost, Μόσχα, Ιδία Επεξεργασία*

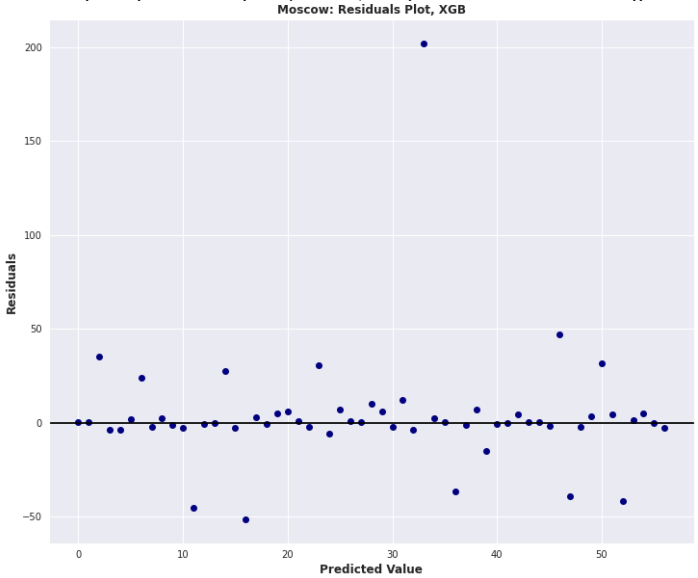
Από το διάγραμμα διασποράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Μόσχας, παρατηρείται γραμμικότητα. Γραμμικότητα εμφανίζεται καθώς είναι εφικτή η προσαρμογή της ευθείας ελαχίστων τετραγώνων στα σημεία. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Λαμβάνοντας υπόψιν τη σχετικά μεγάλη κλίση της ευθείας ελαχίστων τετραγώνων, συμπεραίνεται ότι η ισχύς είναι αρκετά μέτρια προς έντονη. Τέλος, από το παραπάνω διάγραμμα παρατηρούνται ορισμένα ιδιάζοντα σημεία τα οποία βρίσκονται αρκετά μακριά από τις συγκεντρώσεις των υπόλοιπων σημείων και από την ευθεία ελαχίστων τετραγώνων.

Στο παρακάτω διάγραμμα, **Σχήμα 194**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων και οι πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο μελέτης. Σε γενικές γραμμές, εκείνο το οποίο παρατηρείται, είναι ότι έχει γίνει μία καλή προσαρμογή του αλγορίθμου στις πραγματικές τιμές των θανάτων της Μόσχας. Φαίνεται ότι ο αλγόριθμος μπορεί να προβλέψει την τάση και τη ροή των κρουσμάτων, όμως, αδυνατεί να υπολογίσει τις πραγματικές τιμές, κατά κύριο λόγο, σε ορισμένα μέγιστα. Η διαφοροποίηση αυτή των τιμών, έχει να κάνει με ένα πολύ μικρό χρονικό διάστημα, όμως στην περίπτωση της 06/12/2020, συναντάται μία σημαντική απόκλιση ανάμεσα σε προβλεπόμενη και πραγματική τιμή. Σε γενικά πλαίσια, η προσαρμογή του αλγορίθμου σε αυτήν την περίπτωση είναι καλή.



Σχήμα 194 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, XGBoost, Μόσχα, Ιδία Επεξεργασία*

Από το διάγραμμα υπολοίπων, **Σχήμα 195**, φαίνεται ότι τα υπόλοιπα βρίσκονται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Ακόμη, παρατηρείται ένα σημείο με μεγάλη θετική τιμή, μακριά από την ευθεία y=0. Ίσως να πρόκειται για ιδιάζον σημείο, καθώς βρίσκεται μακριά από το μοτίβο και τη συγκέντρωση των υπόλοιπων σημείων.



Σχήμα 195 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, XGBoost, Μόσχα, Ιδία Επεξεργασία*

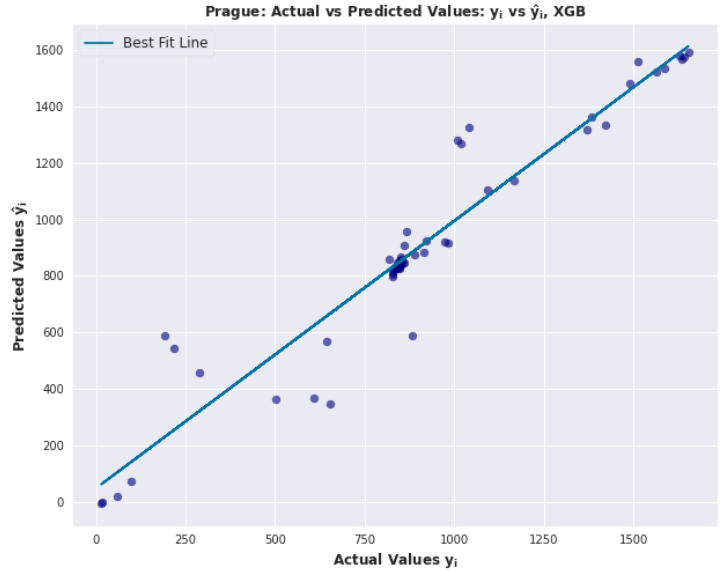
Στο τελικό μοντέλο ανάλυσης για πρόβλεψη θανάτων με τον αλγόριθμο XGBoost, ακολουθούν τα πειραματικά αποτελέσματα σε γραφήματα, αλλά και οι μετρικές αξιολόγησης σε Πίνακα, για την πόλη της Πράγας.

Μετρική Αξιολόγησης	**Πράγα**
RMSE	123.08 (deaths per million)
R^2	0.911
EVS	0.911
MAE	73.14 (deaths per million)
MAPE	0.19%
MSE	0.00547

Πίνακας 62 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, XGBoost, Πράγα, Ιδία Επεξεργασία*

Βάσει όσων αναφέρθηκαν και στην υποενότητα **4.2**, οι μετρικές αξιολόγησης οι οποίες περιγράφουν την πόλη της Πράγας, κυμαίνονται σε καλά πλαίσια. Η τιμή της R^2 και της EVS είναι υψηλές, όχι όμως τόσο υψηλές όσο εκείνες των προηγούμενων μοντέλων. Η τιμή της R^2 βρίσκεται κοντά στο 0.91, άρα η συσχέτιση των ανεξάρτητων μεταβλητών με την εξαρτημένη θεωρείται υψηλή. Το MAE και RMSE, μετρικές οι οποίες έχουν την ίδια μονάδα με την προβλεπόμενη μεταβλητή, εμφανίζουν σχετικά χαμηλές τιμές για τους θανάτους ανά εκατομμύριο, λαμβάνοντας υπόψιν το εύρος των πραγματικών τιμών (Βλέπε **Σχήμα 197**). Ακόμη, το ποσοστό από το MAPE είναι αρκετά χαμηλό και μικρότερο από 1%, άρα θεωρείται πολύ καλό. Στην περίπτωση αυτή, συναντάται το μεγαλύτερο MSE σφάλμα για τα μοντέλα πρόβλεψης θανάτων με XGBoost.

Από το διάγραμμα διασποράς το οποίο ακολουθεί, **Σχήμα 196**, ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές των θανάτων ανά εκατομμύριο για την πόλη της Πράγας, παρατηρείται ότι υπάρχει γραμμικότητα. Αυτό σημαίνει ότι τα σημεία του διαγράμματος φαίνεται να σχηματίζουν μία νοητή ευθεία γραμμή. Παρεμβάλλοντας την ευθεία ελαχίστων τετραγώνων γίνεται αντιληπτή η εν λόγω γραμμικότητα. Επιπλέον, παρατηρείται ότι υπάρχει θετική συσχέτιση, καθώς με την αύξηση της μίας μεταβλητής, αυξάνεται και η άλλη. Ακόμη, θεωρείται ότι συναντάται μέτρια προς έντονη ισχύς, λόγω της σχετικά μεγάλης κλίσης της ευθείας ελαχίστων τετραγώνων. Τέλος, όσον αφορά τα ιδιάζοντα σημεία, αυτό το οποίο προκύπτει από το παραπάνω διάγραμμα, είναι ότι εντοπίζονται λίγα σημεία τα οποία θα μπορούσαν να χαρακτηρισθούν έτσι, καθώς απέχουν αρκετά από την ευθεία ελαχίστων τετραγώνων και από συγκεντρώσεις άλλων σημείων.

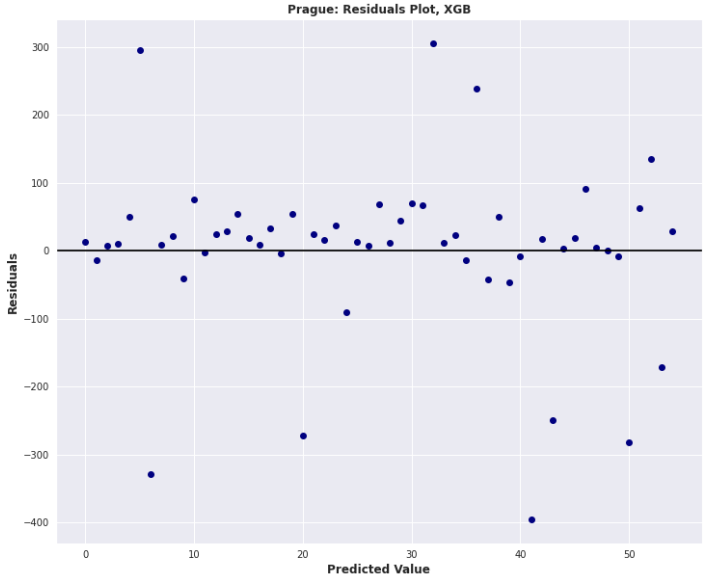


Σχήμα 196 *Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, XGBoost, Πράγα, Ιδία Επεξεργασία*

Στο διάγραμμα των προβλεπόμενων τιμών, **Σχήμα 197**, απεικονίζονται οι τιμές για την πρόβλεψη των θανάτων μαζί με τις πραγματικές τιμές, για την αντίστοιχη χρονική περίοδο. Σε γενικές γραμμές παρατηρείται ότι η απόδοση του αλγορίθμου είναι ικανοποιητική, όμως εμφανίζει αρκετές αστοχίες. Δηλαδή, δεν είναι ικανός να ακολουθήσει πιστά την τάση των πραγματικών τιμών. Ούτε μπορεί να προβλέψει με πλήρη επιτυχία ορισμένα από τα μέγιστα, αλλά και ορισμένα από τα ελάχιστα. Άρα, όπως παρατηρείται και από το διάγραμμα, αρκετά σημεία δεν έχουν προσαρμοστεί στις πραγματικές τιμές των θανάτων.

Σχήμα 197 *Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, XGBoost, Πράγα, Ιδία Επεξεργασία*

Ακολουθεί το διάγραμμα με την κατανομή των υπολοίπων.



Σχήμα 198 *Υπόλοιπα μοντέλου πρόβλεψης θανάτων, XGBoost, Πράγα, Ιδία Επεξεργασία*

Από το διάγραμμα των υπολοίπων φαίνεται ότι ένα πλήθος τους βρίσκεται διάσπαρτα κατανεμημένα γύρω από την ευθεία y=0. Τέλος, εμφανίζονται σημεία τα οποία απέχουν από τον οριζόντιο άξονα, τόσο προς τα αρνητικά όσο και προς τα θετικά, και τα οποία θα μπορούσε κάποιος να τα χαρακτηρίσει ως ιδιάζοντα.

Τα αποτελέσματα των μοντέλων πρόβλεψης θανάτων για το Βερολίνο, τις Βρυξέλλες, τη Λισαβόνα και το Λονδίνο βρίσκονται στο Παράρτημα ΣΤ.

4.4.7 Feature Importance

Στην τρέχουσα υποενότητα, παρατίθενται τα διαγράμματα για την επιλογή μεταβλητών. Συγκεκριμένα, παρουσιάζονται διαγράμματα με το Feature Importance κάθε πόλης και κάθε αλγορίθμου. Σε κάθε διάγραμμα, φαίνεται ο ρόλος τον οποίο διαδραματίζει έκαστη μεταβλητή στην πρόβλεψη, για το εκάστοτε μοντέλο και για τα εκάστοτε δεδομένα.

Πρόκειται για μία διαδικασία η οποία ανήκει στο τομέα της Εξηγήσιμης Τεχνητής Νοημοσύνης, καθώς μελετάται ο βαθμός με τον οποίο επηρεάζει κάθε μεταβλητή, η οποία χρησιμοποιήθηκε στην παρούσα ανάλυση, το αποτέλεσμα της πρόβλεψης.

Αναλυτικότερα, τα διαγράμματα τα οποία παρατίθενται στην παρούσα υποενότητα είναι το διάγραμμα του Feature Importance. Πρόκειται για διαγράμματα στα απεικονίζονται όλες οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν σε έκαστο μοντέλο, συνοδευόμενες από το επίπεδο της σχετικής σημαντικότητάς/βαρύτητάς τους κατά τη διαδικασία της πρόβλεψης. Ένα τέτοιο διάγραμμα, παρέχει καλύτερη αντίληψη των δεδομένων, καλύτερη κατανόηση του μοντέλου και επίσης, μπορεί να βοηθήσει στη μείωση του αριθμού των εισαχθέντων χαρακτηριστικών.

Συναντώνται διάφοροι τρόποι υπολογισμού και δημιουργίας ενός διαγράμματος σημαντικότητας των ανεξάρτητων μεταβλητών. Ο τρόπος ο οποίος επιλέχθηκε για να γίνει η απεικόνιση των διαγραμμάτων Feature Importance, για τρία από τα πέντε μοντέλα, είναι η χρήση συνόλου συντελεστών (coefficients).

Οι συντελεστές περιγράφουν τη μαθηματική σχέση η οποία υπάρχει ανάμεσα σε κάθε ανεξάρτητη μεταβλητή και στην εξαρτημένη. Στην περίπτωση του Feature Importance, ένας συντελεστής παλινδρόμησης φανερώνει μία αρνητική ή θετική συσχέτιση ανάμεσα σε κάθε ανεξάρτητη μεταβλητή και στην εξαρτώμενη. Ένας θετικός συντελεστής δείχνει ότι όσο η τιμή της ανεξάρτητης μεταβλητής αυξάνεται, τόσο το μέσο της εξαρτώμενης μεταβλητής τείνει να αυξάνεται. Αντιθέτως, όταν ένας συντελεστής μίας ανεξάρτητης μεταβλητής είναι αρνητικός, αυτό υποδηλώνει ότι η ανεξάρτητη μεταβλητή έχει αρνητική επίδραση στην εξαρτώμενη μεταβλητή. Δηλαδή, μία αύξηση της ανεξάρτητης μεταβλητής οδηγεί σε μείωση της εξαρτώμενης και αντίστροφα. Συμπερασματικά, η τιμή του συντελεστή υποδηλώνει πόσο ο μέσος της εξαρτημένης μεταβλητής μεταβάλλεται, μέσω της αλλαγής κατά μία μονάδα μίας, απομονωμένης, ανεξάρτητης μεταβλητής, κρατώντας τις υπόλοιπες μεταβλητές του μοντέλου σταθερές (Frost, 2017).

Για τα μοντέλα του Random Forrest Regression και XGBoost Regression, χρησιμοποιήθηκε διαφορετικός τρόπος υπολογισμού. Συγκεκριμένα, η βαρύτητα μεταβλητών υπολογίζεται ως τη συνολική μείωση στην ανομοιογένεια των κόμβων, έχοντας ως βάρος την πιθανότητα να βρεθεί ο κόμβος αυτός. Η πιθανότητα εύρεσης του συγκεκριμένου κόμβου μπορεί να υπολογισθεί από το πλήθος των δειγμάτων τα οποία φτάνουν τον εν λόγω κόμβο, διαιρεμένο του συνολικού πλήθους δειγμάτων. Όσο πιο υψηλή είναι η τιμή, τόσο πιο σημαντική είναι και η μεταβλητή (Ronaghan, 2018).

Στην περίπτωση του RF και του XGBoost, η βαρύτητα εκάστης μεταβλητής υπολογίζεται με διαφορετικό τρόπο, από ότι στα προηγούμενα τρία μοντέλα. Να σημειωθεί ότι σε αυτό το μοντέλο, δεν μπορεί να σχολιαστεί εάν η επιρροή μίας μεταβλητής έχει θετική ή αρνητική επίδραση. Αυτό, όμως, το οποίο γίνεται αντιληπτό είναι το πόσο συμβάλλει κάθε ανεξάρτητη μεταβλητή στην πρόβλεψη της εξαρτημένης μεταβλητής.

Στα γραφήματα τα οποία ακολουθούν, στον άξονα y βρίσκονται οι ονομασίες των μεταβλητών οι οποίες χρησιμοποιήθηκαν στην πρόβλεψη. Στον άξονα x εμφανίζονται τα βάρη.

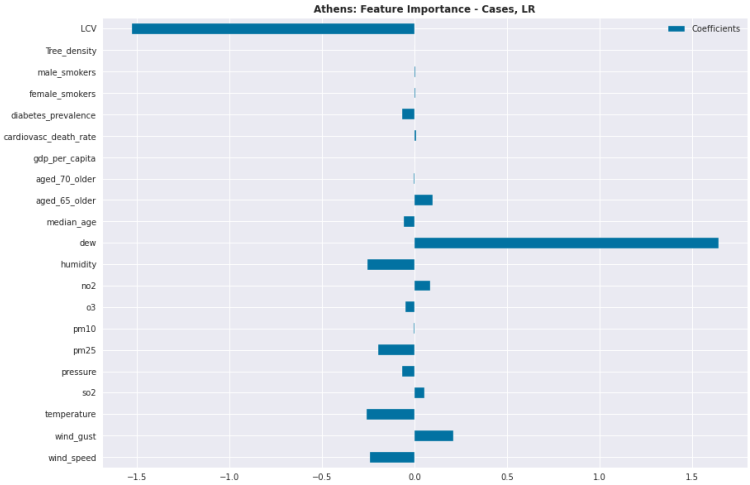
4.4.7.1 Αθήνα

Ακολουθώντας αλφαβητική σειρά, πρώτη πόλη για την οποία θα παρουσιαστούν τα διαγράμματα για τη σημαντικότητα των μεταβλητών, είναι η Αθήνα. Πρόκειται να παρατεθούν τα διαγράμματα τα οποία έχουν προκύψει αρχικά για τις προβλέψεις των κρουσμάτων και στη συνέχεια για τις προβλέψεις των θανάτων, για κάθε αλγόριθμο.

i. Πρόβλεψη Κρουσμάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.



Σχήμα 199 *Feature importance για πρόβλεψη κρουσμάτων, LR, Αθήνα, Ιδία Επεξεργασία*

Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή "dew". Πρόκειται για το σημείο δρόσου. Πρόκειται για μία τιμή μεγαλύτερη από 1.5. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV (Land Cover in Vegetation) και είναι περίπου ίσος με -1.5. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή είναι η ανεξάρτητη μεταβλητή του σημείου δρόσου.

Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης η θερμοκρασία (temperature), η ταχύτητα του ανέμου (wind speed), τα $PM_{2.5}$ και η υγρασία. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η μεταβλητή των ριπών ανέμου (wind gust). Δηλαδή, με την αύξηση των ριπών ανέμου, αναμένεται η αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων (tree density) και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

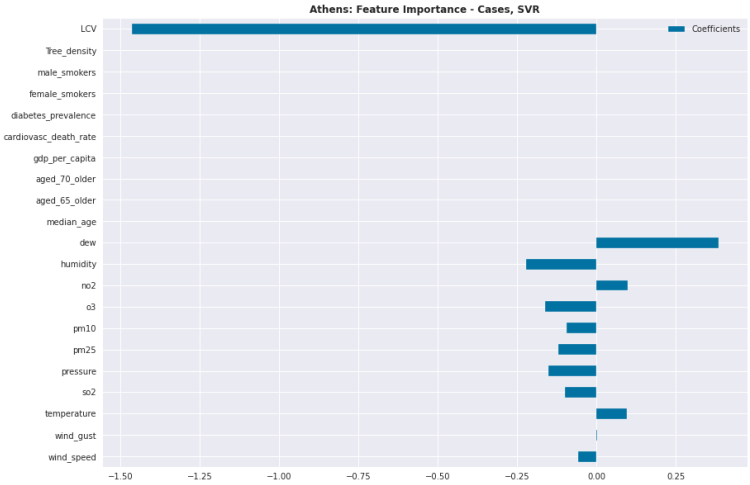
- SVR

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στη συνέχεια.

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή του σημείου δρόσου (dew). Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης η υγρασία, το όζον (O_3), τα PM_{10} και $PM_{2.5}$, η πίεση (pressure), το διοξείδιο του θείου (SO_2) και η ταχύτητα του ανέμου. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η

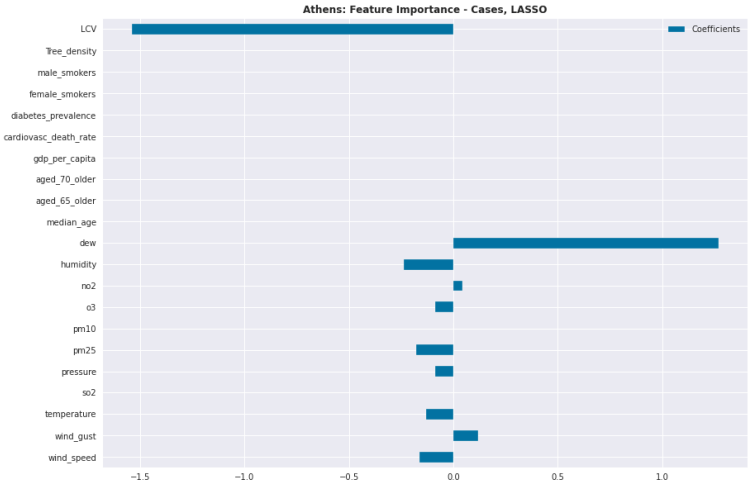
μεταβλητή του διοξειδίου του αζώτου (NO_2) και η θερμοκρασία. Δηλαδή, με την αύξηση της θερμοκρασίας και του NO_2, αναμένεται η αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων (tree density), άντρες και γυναίκες καπνιστές (male and female smokers), προβλήματα διαβήτη (diabetes prevalence) και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.



Σχήμα 200 *Feature importance για πρόβλεψη κρουσμάτων, SVR, Αθήνα, Ιδία Επεξεργασία*

- LASSO

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.



Σχήμα 201 *Feature importance για πρόβλεψη κρουσμάτων, LASSO, Αθήνα, Ιδία Επεξεργασία*

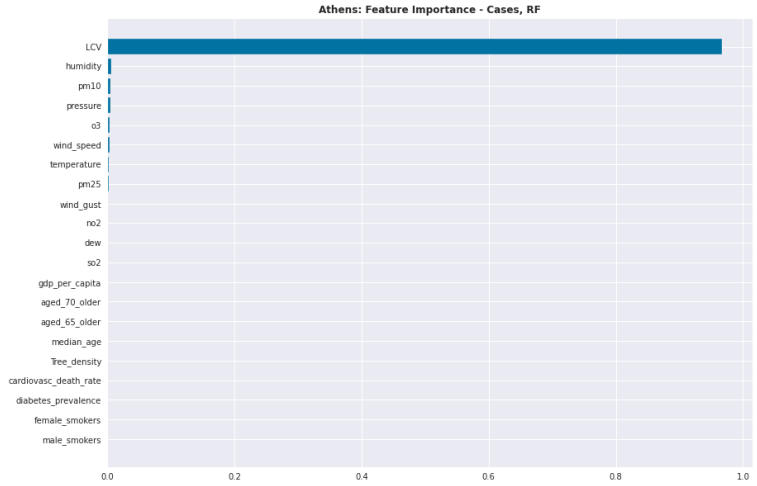
Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή του σημείου δρόσου (dew). Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης η υγρασία, το όζον, τα $PM_{2.5}$, η πίεση, η θερμοκρασία και η ταχύτητα του ανέμου. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η μεταβλητή της ταχύτητας του ανέμου και του διοξειδίου του αζώτου. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η

αύξηση των τιμών των κρουσμάτων και με τη μείωσή τους, αναμένεται και μείωση των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων (tree density), PM_{10}, άντρες και γυναίκες καπνιστές (male and female smokers), SO_2, προβλήματα διαβήτη (diabetes prevalence) και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- RF Regression

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.

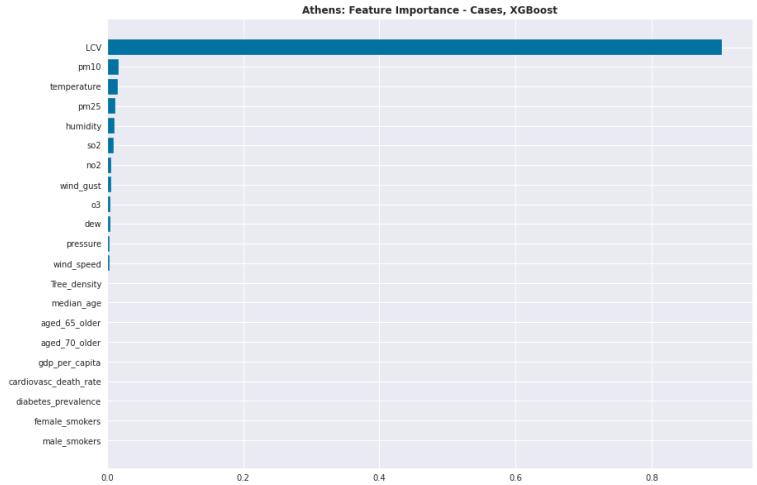


Σχήμα 202 *Feature importance για πρόβλεψη κρουσμάτων, RF, Αθήνα, Ιδία Επεξεργασία*

Με μία μελέτη του παραπάνω γραφήματος, παρατηρείται ότι μεγαλύτερη επιρροή, με τιμή πολύ κοντά στο 1, περίπου 0.96, φαίνεται να παρουσιάζει η μεταβλητή του LCV. Πρόκειται, λοιπόν, για την πιο σημαντική μεταβλητή για την πρόβλεψη του αριθμού των κρουσμάτων. Ακολουθεί με πολύ μικρότερη τιμή η υγρασία, τα PM10, η πίεση, το όζον, η ταχύτητα του ανέμου, η θερμοκρασία και τα $PM_{2.5}$. Οι τιμές αυτών των μεταβλητών δεν ξεπερνάνε το 0.1. Άρα, για το παραπάνω μοντέλο, θεωρείται ότι κατά κύριο λόγο μεγαλύτερη επιρροή στην πρόβλεψη ασκεί η μεταβλητή LCV.

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.



Σχήμα 203 *Feature importance για πρόβλεψη κρουσμάτων, XGBoost, Αθήνα, Ιδία Επεξεργασία*

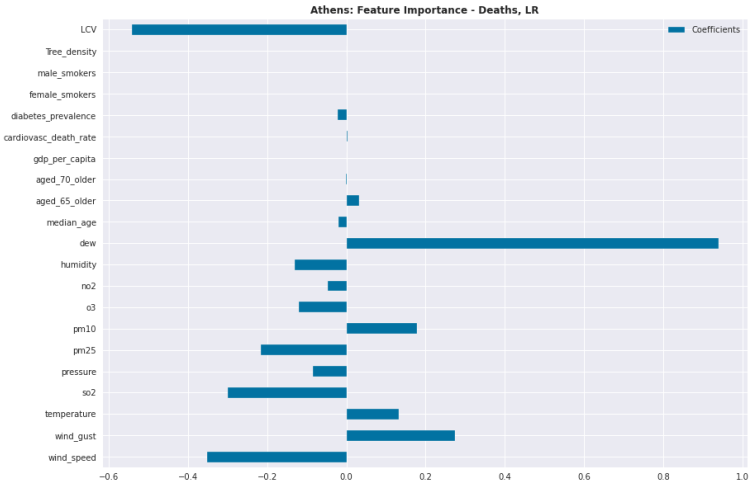
Όπως και στην περίπτωση του RF, έτσι κι εδώ, παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV, με τιμή μεγαλύτερη του 0.9. Ακολουθούν τα

PM10, η θερμοκρασία, τα PM2.5, η υγρασία, το SO2, το NO2, οι ριπές του ανέμου, το O3, η μεταβλητή του σημείου δρόσου η πίεση και η ταχύτητα του ανέμου. Όμως, οι τιμές αυτών των μεταβλητών δεν ξεπερνάνε το 0.1. Άρα, για το παραπάνω μοντέλο, θεωρείται ότι κατά κύριο λόγο μεγαλύτερη επιρροή στην πρόβλεψη ασκεί η μεταβλητή LCV.

ii. Πρόβλεψη Θανάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.



Σχήμα 204 *Feature importance για πρόβλεψη θανάτων, LR, Αθήνα, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα, εμφανίζονται οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν για την πρόβλεψη, μαζί με την τιμή των συντελεστών τους.

Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή του σημείου δρόσου. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή, καθώς εμφανίζει υψηλότερη τιμή, είναι η ανεξάρτητη μεταβλητή του σημείου δρόσου.

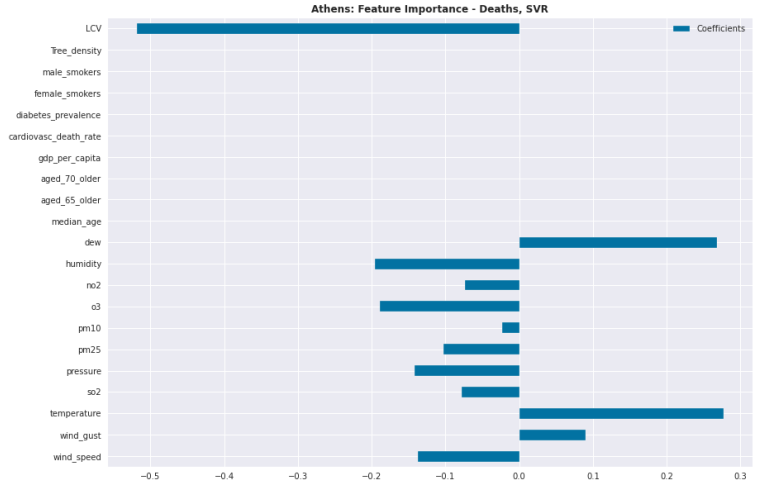
Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης η ταχύτητα ανέμου, το SO_2, τα $PM_{2.5}$, το O_3, η υγρασία και το NO_2. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των θανάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η μεταβλητή των ριπών ανέμου, της θερμοκρασίας και των PM_{10}. Δηλαδή, με την αύξηση των εν λόγω μεταβλητών, αναμένεται η αύξηση των τιμών των θανάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- SVR

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στη συνέχεια.

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή της θερμοκρασίας. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές

τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.

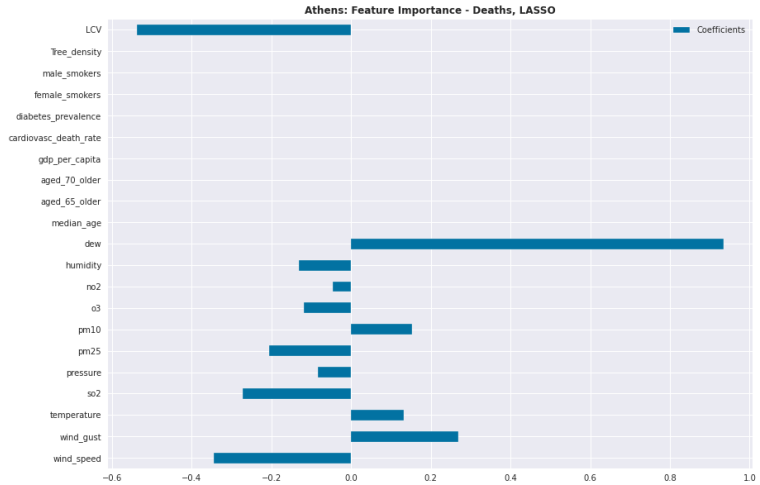


Σχήμα 205 *Feature importance για πρόβλεψη θανάτων, SVR, Αθήνα, Ιδία Επεξεργασία*

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης η υγρασία, το όζον, η πίεση η ταχύτητα ανέμου, τα $PM_{2.5}$, το διοξείδιο του θείου (SO_2) και τα PM_{10}. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η μεταβλητή του σημείου δρόσου και οι ριπές του ανέμου. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων (tree density), άντρες και γυναίκες καπνιστές (male and female smokers), προβλήματα διαβήτη (diabetes prevalence) και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- <u>LASSO</u>

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.



Σχήμα 206 *Feature importance για πρόβλεψη θανάτων, LASSO, Αθήνα, Ιδία Επεξεργασία*

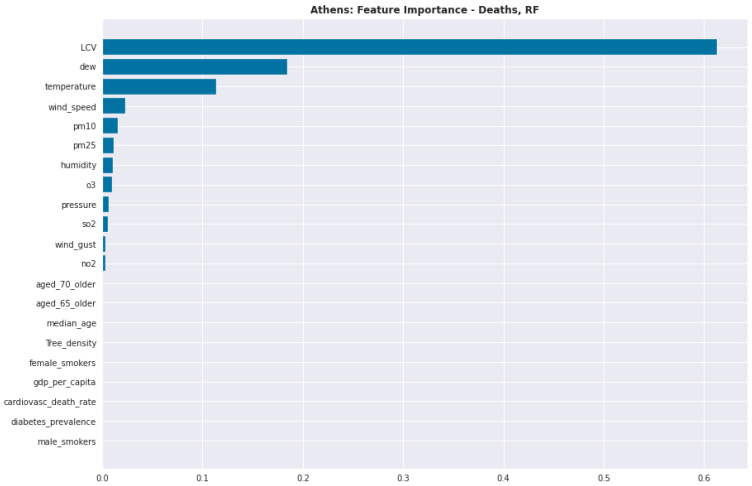
Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή του σημείου δρόσου (dew). Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των

κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το dew.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης η ταχύτητα του ανέμου, το SO_2, τα $PM_{2.5}$, η υγρασία και το όζον. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζουν οι ρίποι του ανέμου, τα PM_{10} και η θερμοκρασία. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η αύξηση των τιμών των θανάτων και με τη μείωσή τους, αναμένεται και μείωση των θανάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων (tree density), άντρες και γυναίκες καπνιστές (male and female smokers), προβλήματα διαβήτη (diabetes prevalence) και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- RF Regression

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.



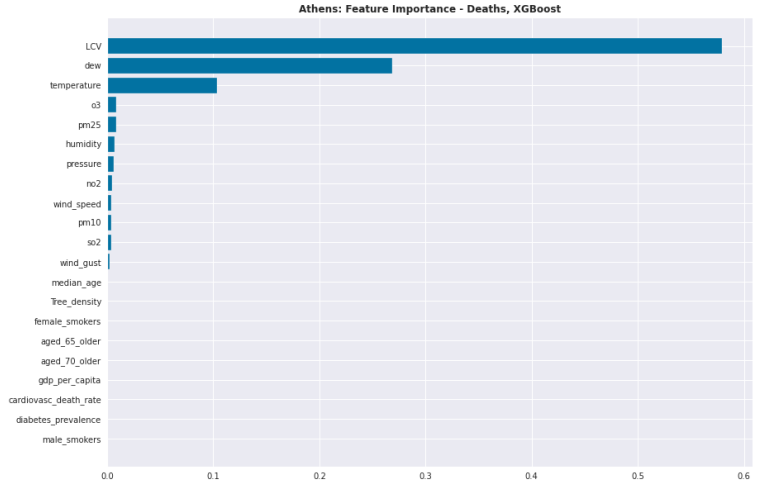
Σχήμα 207 *Feature importance για πρόβλεψη θανάτων, RF, Αθήνα, Ιδία Επεξεργασία*

Παρατηρείται ότι τη μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί το σημείο δρόσου, η θερμοκρασία, η ταχύτητα του ανέμου, τα PM_{10} και $PM_{2.5}$, η υγρασία, το όζον, η πίεση, το SO_2, οι ριπές του ανέμου και το NO_2. Οι υπόλοιπες μεταβλητές φαίνεται ότι έχουν ελάχιστη, έως καθόλου, βαρύτητα.

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV, με τιμή κοντά στο 0.6. Ακολουθεί το σημείο δρόσου, το οποίο εμφανίζει τη δεύτερη μεγαλύτερη τιμή κοντά στο 0.3, η θερμοκρασία, το όζον, τα $PM_{2,5}$, η υγρασία, η πίεση, το NO_2, η ταχύτητα του ανέμου, τα PM_{10}, το SO_2 και οι ριπές του ανέμου. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση.



Σχήμα 208 *Feature importance για πρόβλεψη θανάτων, XGBoost, Αθήνα, Ιδία Επεξεργασία*

Συμπεράσματα:

Έπειτα από την ανάλυση των παραπάνω διαγραμμάτων, προκύπτουν ορισμένα συμπεράσματα.

Όσον αφορά τα τρία πρώτα μοντέλα πρόβλεψης κρουσμάτων, οι μεταβλητές οι οποίες κυριαρχούν είναι η LCV και η dew. Για τον αλγόριθμο RF και XGBoost, δεν συναντάται έντονη επιρροή της dew στο τελικό αποτέλεσμα. Πρώτη θέση εξακολουθεί να έχει η LCV. Εν συνεχεία, μικρότερες τιμές επίδρασης παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία, πίεση και άνεμος. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα, με οικονομικούς παράγοντες, με ηλικιακούς παράγοντες και με το πράσινο της πόλης, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

Όσον αφορά τα μοντέλα πρόβλεψης θανάτων, προκύπτει ότι για την πόλη της Αθήνας, οι μεταβλητές οι οποίες φαίνεται να ασκούν σημαντική επιρροή είναι δύο. Η LCV και η dew. Σε τέσσερα από τα πέντε μοντέλα τα οποία αναλύθηκαν, κυριαρχούν οι δύο αυτές μεταβλητές για τις πρώτες δύο θέσεις. Λαμβάνοντας υπόψιν τους συντελεστές από τα τρία πρώτα μοντέλα, η LCV φαίνεται να ασκεί αρνητική επίδραση, ενώ η dew ασκεί θετική επίδραση στην πρόβλεψη. Στη συνέχεια, μικρότερες τιμές επίδρασης παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία και άνεμος. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα, με οικονομικούς παράγοντες, με ηλικιακούς παράγοντες και με το πράσινο, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

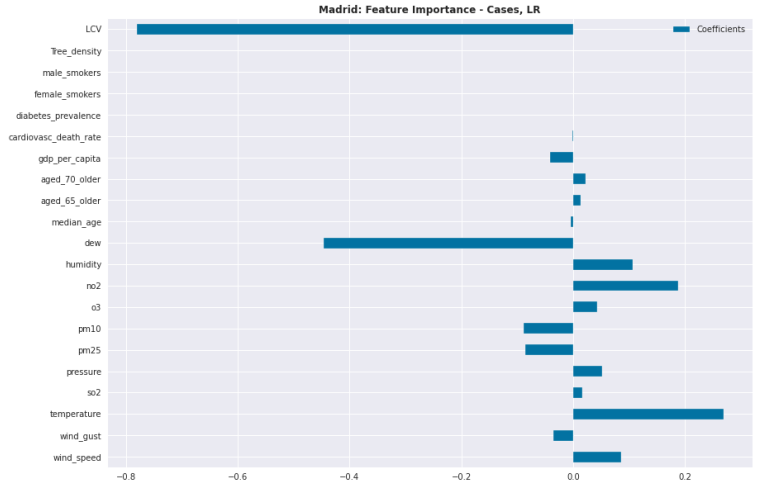
4.4.7.2 Μαδρίτη

Στη συνέχεια, δεύτερη πόλη για την οποία θα παρουσιαστούν τα διαγράμματα για τη σημαντικότητα των μεταβλητών, είναι η Μαδρίτη. Πρόκειται να παρατεθούν τα διαγράμματα τα οποία έχουν προκύψει αρχικά για τις προβλέψεις των κρουσμάτων και στη συνέχεια για τις προβλέψεις των θανάτων, για κάθε αλγόριθμο.

i. Πρόβλεψη Κρουσμάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.



Σχήμα 209 *Feature importance για πρόβλεψη κρουσμάτων, LR, Μαδρίτη, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα, εμφανίζονται οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν για την πρόβλεψη, μαζί με την τιμή των συντελεστών τους.

Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή της θερμοκρασίας. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή είναι η ανεξάρτητη μεταβλητή LCV.

Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης η μεταβλητή dew, τα $PM_{2.5}$ και PM_{10}, η ταχύτητα του ανέμου και η μεταβλητή gdp per capita. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει το NO_2, η υγρασία, οι ρίποι ανέμου, η πίεση, το O_3 και ο πληθυσμός με ηλικία μεγαλύτερη από 65 και 75 έτη. Δηλαδή, με την αύξηση των ριπών ανέμου, αναμένεται η αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων (tree density), οι άνδρες και οι γυναίκες καπνιστές φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

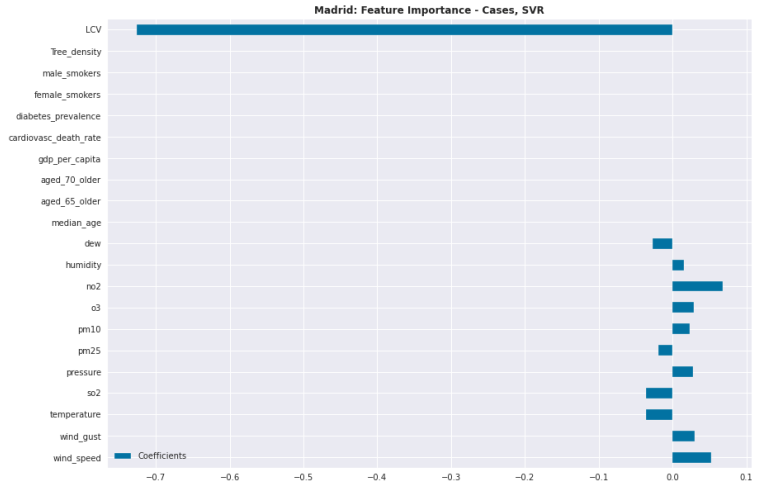
- <u>SVR</u>

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στο Σχήμα το οποίο ακολουθεί.

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή του NO_2. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, σημαντικά μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης η θερμοκρασία, το SO_2, το dew και τα $PM_{2.5}$. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η μεταβλητή της ταχύτητας του ανέμου, των ριπών ανέμου, της πίεσης, του O_3 και των PM_{10}. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές,

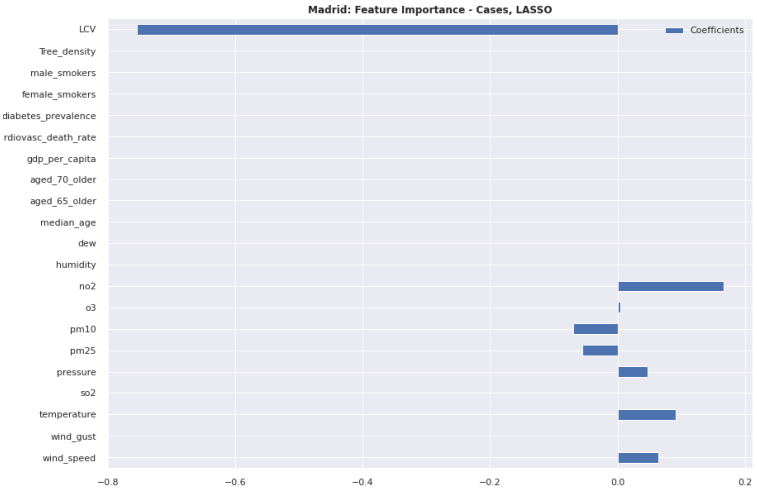
προβλήματα διαβήτη (diabetes prevalence), ηλικιακά χαρακτηριστικά και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.



Σχήμα 210 *Feature importance για πρόβλεψη κρουσμάτων, SVR, Μαδρίτη, Ιδία Επεξεργασία*

- LASSO

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.



Σχήμα 211 *Feature importance για πρόβλεψη κρουσμάτων, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

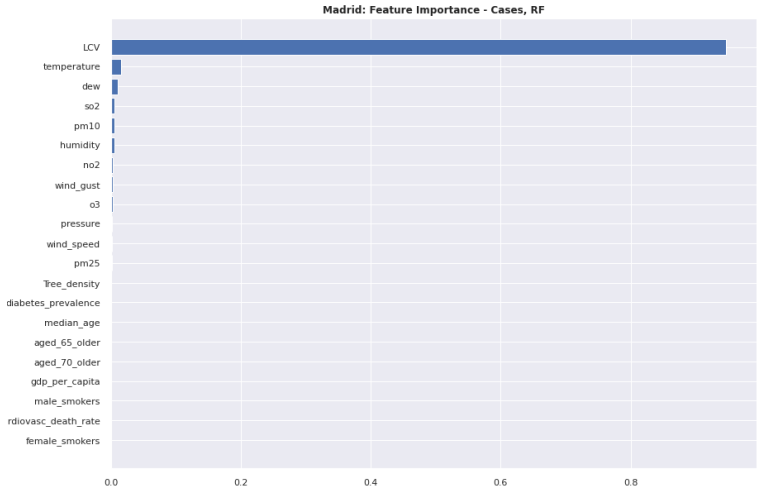
Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή του NO_2. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, αρκετά μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζουν τα $PM_{2.5}$ και PM_{10}. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η μεταβλητή της θερμοκρασίας, της ταχύτητας του ανέμου και της πίεσης. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η αύξηση των τιμών των κρουσμάτων και με τη μείωσή τους, αναμένεται και μείωση των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, SO_2, υγρασία, dew, προβλήματα διαβήτη (diabetes

prevalence) και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- RF Regression

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.



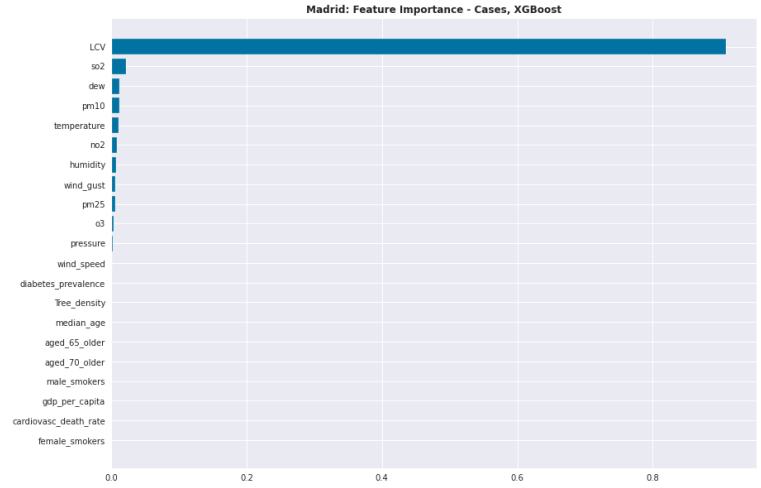
Σχήμα 212 *Feature importance για πρόβλεψη κρουσμάτων, RF, Μαδρίτη, Ιδία Επεξεργασία*

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το dew, το SO_2, τα PM_{10}, η υγρασία, το NO_2, οι ριπές του ανέμου και το όζον. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση στην πρόβλεψη των κρουσμάτων.

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.

Από το παρακάτω διάγραμμα παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί το SO_2, το dew, τα PM_{10}, η θερμοκρασία, το NO_2, η υγρασία, οι ριπές του ανέμου, τα $PM_{2.5}$, το O_3 και η πίεση. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση στην πρόβλεψη των κρουσμάτων.

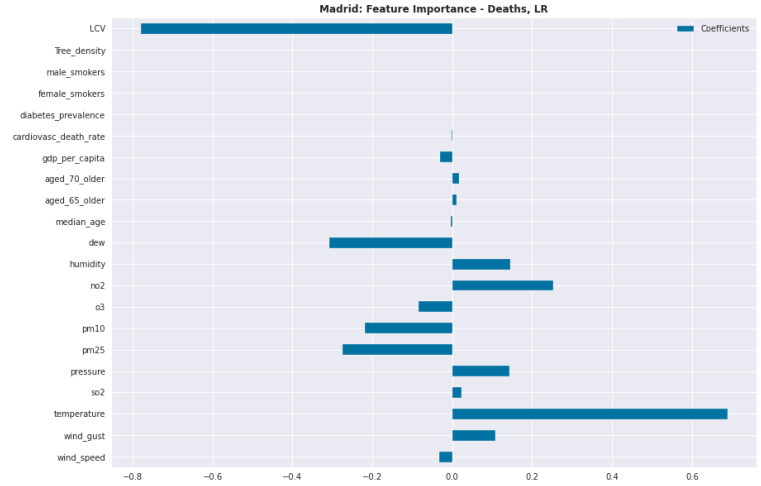


Σχήμα 213 *Feature importance για πρόβλεψη κρουσμάτων, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

ii. Πρόβλεψη Θανάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.



Σχήμα 214 *Feature importance για πρόβλεψη θανάτων, LR, Μαδρίτη, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα, εμφανίζονται οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν για την πρόβλεψη, μαζί με την τιμή των συντελεστών τους.

Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η θερμοκρασία Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή, καθώς εμφανίζει υψηλότερη τιμή, είναι η ανεξάρτητη μεταβλητή LCV.

Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης το dew, τα $PM_{2.5}$, τα PM_{10}, το O_3, οι ρίποι ανέμου και η μεταβλητή gdp per capita. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των θανάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει το NO_2, η υγρασία, η πίεση η ταχύτητα του ανέμου και το SO_2. Δηλαδή, με την αύξηση των εν λόγω μεταβλητών, αναμένεται η αύξηση των τιμών των θανάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άνδρες και γυναίκες καπνιστές, ηλικιακές ομάδες άνω των 65 και 70 ετών, φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

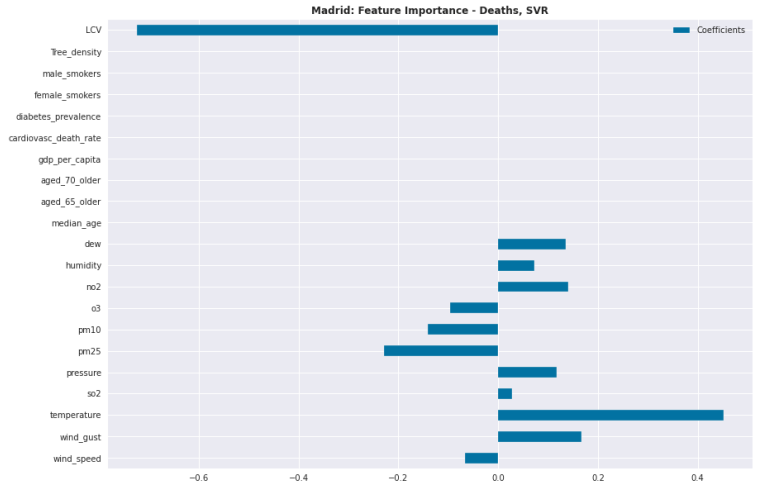
- <u>SVR</u>

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στη συνέχεια.

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή της θερμοκρασίας. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζουν επίσης τα $PM_{2.5}$, το O_3 και το NO_2. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζουν οι ριπές του ανέμου, το dew, το NO_2, η πίεση, η υγρασία και το SO_2. Δηλαδή, με την αύξηση

των μεταβλητών αυτών, αναμένεται αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, προβλήματα διαβήτη, ηλικιακές ομάδες άνω των 65 και 70 ετών και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

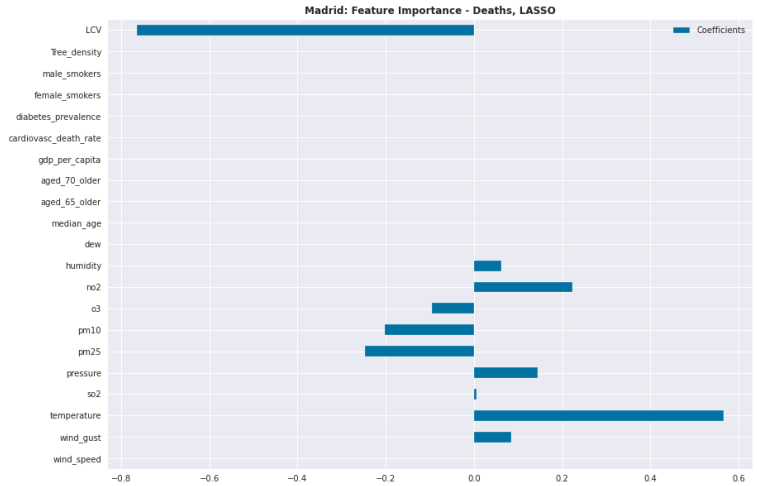


Σχήμα 215 *Feature importance για πρόβλεψη θανάτων, SVR, Μαδρίτη, Ιδία Επεξεργασία*

- LASSO

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.

Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή της θερμοκρασίας. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει ο δείκτης LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.



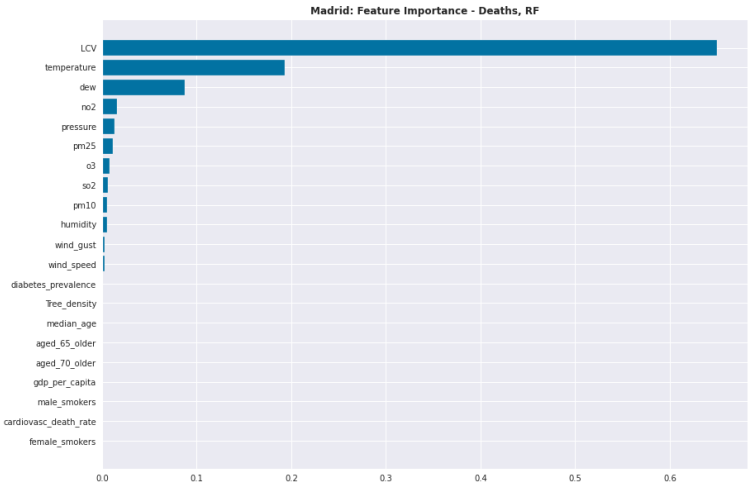
Σχήμα 216 *Feature importance για πρόβλεψη θανάτων, LASSO, Μαδρίτη, Ιδία Επεξεργασία*

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζουν επίσης τα $PM_{2.5}$, τα PM_{10} και το O_3. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει το NO_2, η πίεση, οι ρίποι του ανέμου και η υγρασία. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η αύξηση των τιμών των θανάτων και με τη μείωσή τους, αναμένεται και μείωση των θανάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες

καπνιστές, προβλήματα διαβήτη και το gdp per capita φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- RF Regression

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.

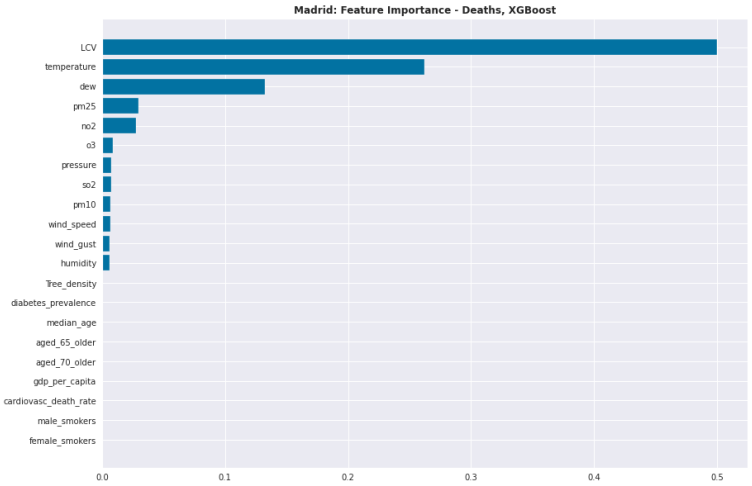


Σχήμα 217 *Feature importance για πρόβλεψη θανάτων, RF, Μαδρίτη, Ιδία Επεξεργασία*

Παρατηρείται ότι τη μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το σημείο δρόσου, το NO_2, η πίεση, $PM_{2.5}$, το O_3, το SO_2, τα PM_{10}, η υγρασία, οι ριπές του ανέμου και η ταχύτητα του ανέμου. Οι υπόλοιπες μεταβλητές φαίνεται ότι έχουν ελάχιστη, έως καθόλου, βαρύτητα.

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.



Σχήμα 218 *Feature importance για πρόβλεψη θανάτων, XGBoost, Μαδρίτη, Ιδία Επεξεργασία*

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το σημείο δρόσου, τα $PM_{2.5}$, το NO_2, το O_3, η πίεση, το SO_2, τα PM_{10}, η ταχύτητα του ανέμου και οι ριπές του ανέμου και η υγρασία. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση.

Συμπεράσματα:

Έπειτα από την ανάλυση των παραπάνω διαγραμμάτων, προκύπτουν ορισμένα συμπεράσματα.

Όσον αφορά τα τρία πρώτα μοντέλα πρόβλεψης κρουσμάτων, οι μεταβλητές οι οποίες κυριαρχούν είναι η LCV και το NO_2. Για τον αλγόριθμο RF, συναντάται πάλι έντονη επιρροή της LCV, όμως ως δεύτερη σημαντικότερη μεταβλητή είναι εκείνη της θερμοκρασίας. Για το XGBoost, πρώτη θέση εξακολουθεί να έχει η LCV, όμως δεύτερη έχει το NO_2. Εν συνεχεία, μικρότερες τιμές επίδρασης παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία, πίεση και άνεμος. Ακόμη, μεταβλητές οι οποίες σχετίζονται με οικονομικούς και ηλικιακούς παράγοντες, φαίνεται να ασκούν μία πολύ μικρή επιρροή. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα καπνιστών και με το πράσινο της πόλης, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

Όσον αφορά τα μοντέλα πρόβλεψης θανάτων, προκύπτει ότι για την πόλη της Μαδρίτης, οι μεταβλητές οι οποίες φαίνεται να ασκούν σημαντική επιρροή είναι δύο. Η LCV και η θερμοκρασία. Σε όλα μοντέλα τα οποία αναλύθηκαν, κυριαρχούν οι δύο αυτές μεταβλητές για τις πρώτες δύο θέσεις. Λαμβάνοντας υπόψιν τους συντελεστές από τα τρία πρώτα μοντέλα, η LCV φαίνεται να ασκεί αρνητική επίδραση, ενώ η θερμοκρασία ασκεί θετική επίδραση στην πρόβλεψη. Στη συνέχεια, μικρότερες τιμές επίδρασης παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία και άνεμος. Ακόμη, μεταβλητές οι οποίες σχετίζονται με οικονομικούς και ηλικιακούς παράγοντες, φαίνεται να ασκούν μία πολύ μικρή επιρροή. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα καπνιστών και με το πράσινο, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

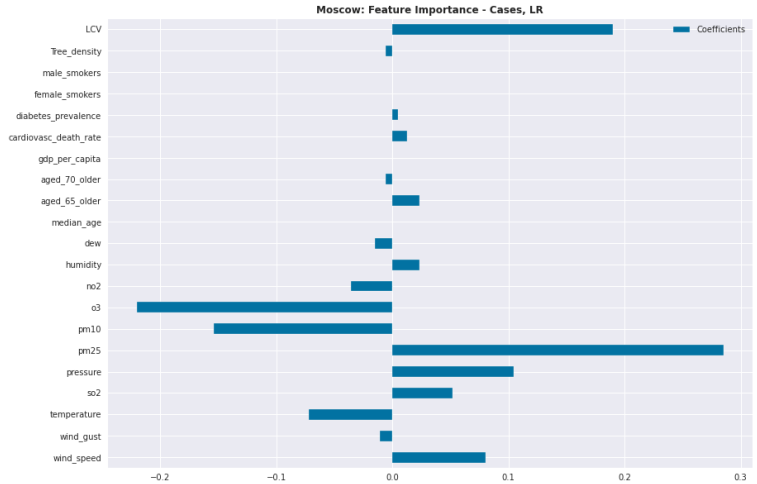
4.4.7.3 Μόσχα

Στη συνέχεια, τρίτη πόλη για την οποία θα παρουσιαστούν τα διαγράμματα για τη σημαντικότητα των μεταβλητών, είναι η Μόσχα. Πρόκειται να παρατεθούν τα διαγράμματα τα οποία έχουν προκύψει αρχικά για τις προβλέψεις των κρουσμάτων και στη συνέχεια για τις προβλέψεις των θανάτων, για κάθε αλγόριθμο.

iii. Πρόβλεψη Κρουσμάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.



Σχήμα 219 *Feature importance για πρόβλεψη κρουσμάτων, LR, Μόσχα, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα, εμφανίζονται οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν για την πρόβλεψη, μαζί με την τιμή των συντελεστών τους.

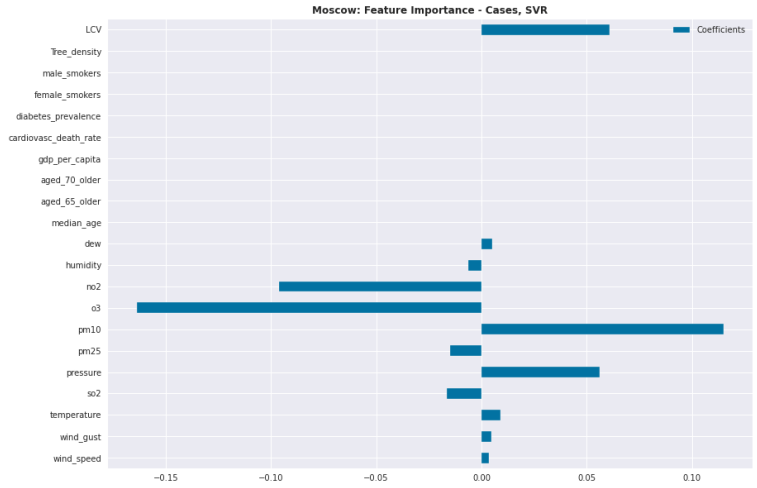
Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή της σωματιδιακής ύλης με διάμετρο μικρότερη από 2.5. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται

η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το όζον. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή της εν λόγω μεταβλητής, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή είναι τα $PM_{2.5}$.

Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζουν επίσης τα PM_{10}, η θερμοκρασία, το NO_2, το dew και οι ριπές ανέμου. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει το LCV, η πίεση, η ταχύτητα του ανέμου, το SO_2, η υγρασία, ο πληθυσμός μεγαλύτερος των 65 ετών και η μεταβλητή του δείκτη θνησιμότητας από καρδιαγγειακή ασθένεια (cardiovasc death rate). Δηλαδή, με την αύξηση των παραπάνω μεταβλητών, αναμένεται η αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων (tree density), οι άνδρες και οι γυναίκες καπνιστές φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- <u>SVR</u>

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στο Σχήμα το οποίο ακολουθεί.



Σχήμα 220 *Feature importance για πρόβλεψη κρουσμάτων, SVR, Μόσχα, Ιδία Επεξεργασία*

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή των PM_{10}. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, σημαντικά μεγαλύτερη αρνητική τιμή παρουσιάζει το όζον. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη του όζοντος.

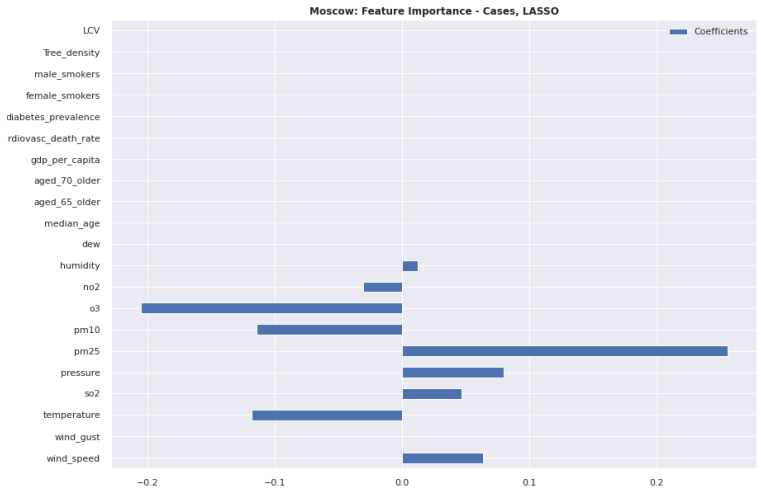
Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης το NO_2, τα $PM_{2.5}$, το SO_2 και η υγρασία. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει το LCV, η πίεση, η θερμοκρασία, οι ριπές ανέμου, το dew και η ταχύτητα ανέμου. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, προβλήματα διαβήτη (diabetes prevalence), ηλικιακά χαρακτηριστικά και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- <u>LASSO</u>

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.

Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή των $PM_{2.5}$. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, αρκετά μεγαλύτερη αρνητική τιμή παρουσιάζει το όζον, O_3. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για τα $PM_{2.5}$.

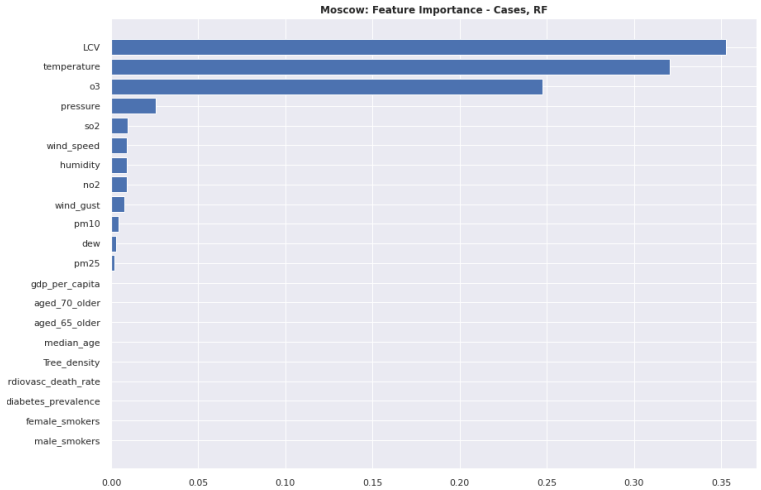
Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει η θερμοκρασία, τα PM_{10} και το NO_2. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η μεταβλητή της πίεσης, της ταχύτητας του ανέμου, του SO_2 και της υγρασίας. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η αύξηση των τιμών των κρουσμάτων και με τη μείωσή τους, αναμένεται και μείωση των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, dew, προβλήματα διαβήτη (diabetes prevalence) και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.



Σχήμα 221 *Feature importance για πρόβλεψη κρουσμάτων, LASSO, Μόσχα, Ιδία Επεξεργασία*

- <u>RF Regression</u>

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.

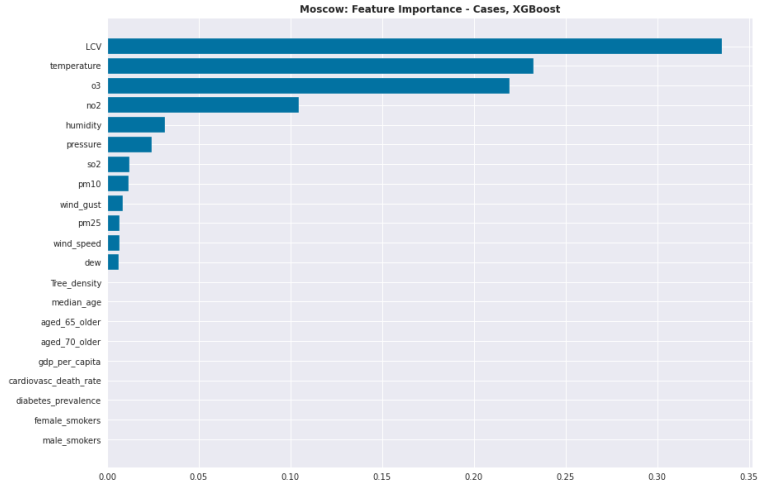


Σχήμα 222 *Feature importance για πρόβλεψη κρουσμάτων, RF, Μόσχα, Ιδία Επεξεργασία*

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το όζον, η πίεση, το SO_2, η ταχύτητα του ανέμου, η υγρασία, το NO_2, οι ριπές του ανέμου, τα PM_{10}, το dew και τα $PM_{2.5}$. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση στην πρόβλεψη των κρουσμάτων.

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.



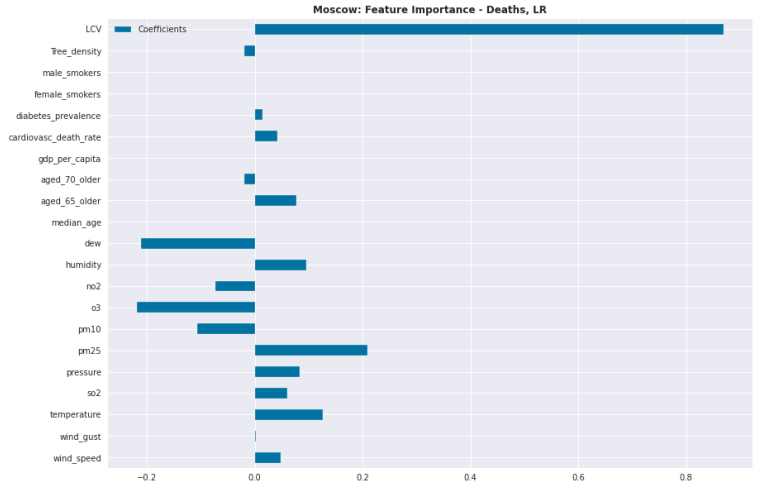
Σχήμα 223 *Feature importance για πρόβλεψη κρουσμάτων, XGBoost, Μόσχα, Ιδία Επεξεργασία*

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το όζον, το NO_2, η υγρασία, η πίεση, το SO_2, τα PM_{10}, οι ριπές του ανέμου, τα $PM_{2.5}$, η ταχύτητα του ανέμου και το dew. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση στην πρόβλεψη των κρουσμάτων.

iv. Πρόβλεψη Θανάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.



Σχήμα 224 *Feature importance για πρόβλεψη θανάτων, LR, Μόσχα, Ιδία Επεξεργασία*

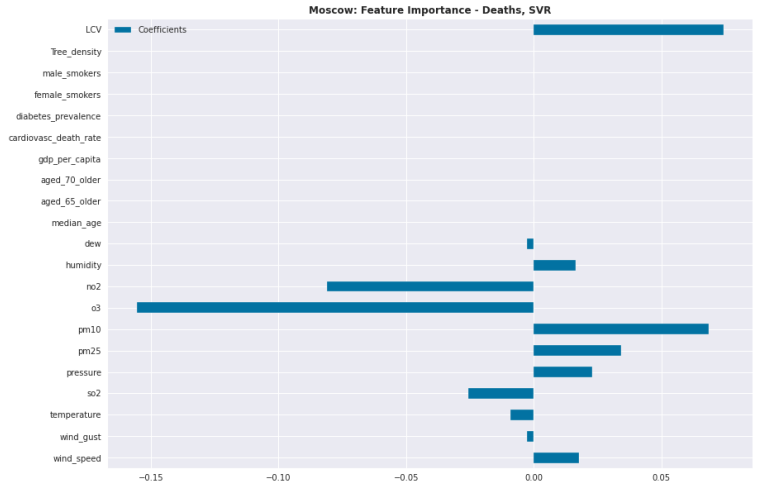
Από το παραπάνω διάγραμμα, εμφανίζονται οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν για την πρόβλεψη, μαζί με την τιμή των συντελεστών τους.

Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει το LCV. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το όζον. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή, καθώς εμφανίζει υψηλότερη τιμή, είναι η ανεξάρτητη μεταβλητή LCV.

Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης το dew, τα PM_{10}, το NO_2, η πυκνότητα δέντρων και η ηλικιακή ομάδα μεγαλύτερη των 70 ετών. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των θανάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζουν τα $PM_{2.5}$, η θερμοκρασία, η υγρασία, η πίεση, η ηλικιακή ομάδα μεγαλύτερη των 65 ετών, το SO_2, η ταχύτητα του ανέμου και ο δείκτης θνησιμότητας από καρδιαγγειακή ασθένεια. Δηλαδή, με την αύξηση των εν λόγω μεταβλητών, αναμένεται η αύξηση των τιμών των θανάτων. Τέλος, μεταβλητές όπως άνδρες και γυναίκες καπνιστές, μέση ηλικία, φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- SVR

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στη συνέχεια.



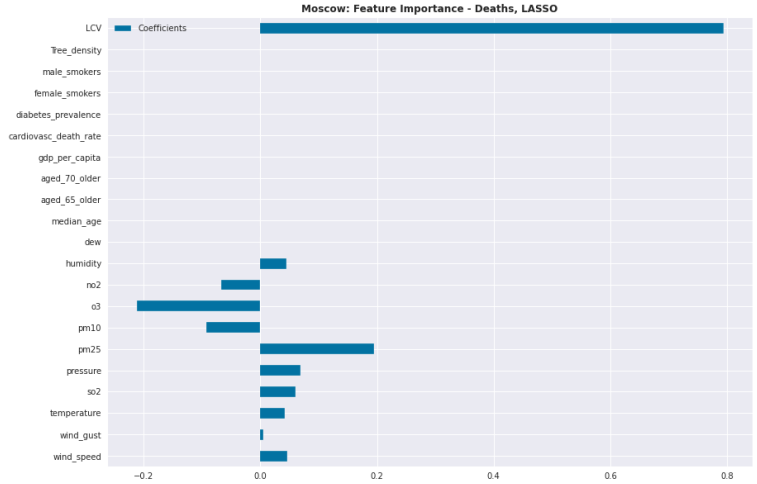
Σχήμα 225 *Feature importance για πρόβλεψη θανάτων, SVR, Μόσχα, Ιδία Επεξεργασία*

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή LCV. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το όζον. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το όζον.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζουν επίσης το NO_2, το SO_2, η θερμοκρασία και οι ριπές του ανέμου. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζουν τα PM_{10}, $PM_{2.5}$, η υγρασία και η ταχύτητα του ανέμου. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, προβλήματα διαβήτη, ηλικιακές ομάδες άνω των 65 και 70 ετών και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- LASSO

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.



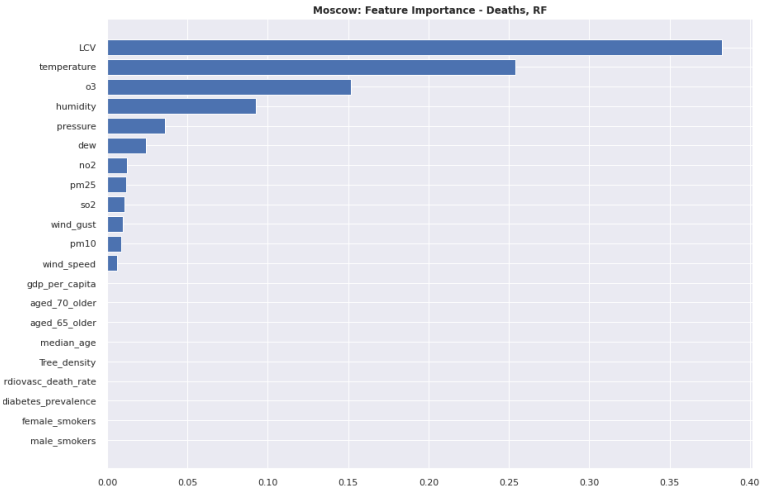
Σχήμα 226 *Feature importance για πρόβλεψη θανάτων, LASSO, Μόσχα, Ιδία Επεξεργασία*

Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή LCV. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το όζον. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζουν τα PM_{10} και το NO_2. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζουν τα $PM_{2.5}$, η πίεση, το SO_2, η ταχύτητα του ανέμου, η θερμοκρασία και η υγρασία. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η αύξηση των τιμών των θανάτων και με τη μείωσή τους, αναμένεται και μείωση των θανάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, προβλήματα διαβήτη και το gdp per capita φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- RF Regression

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.

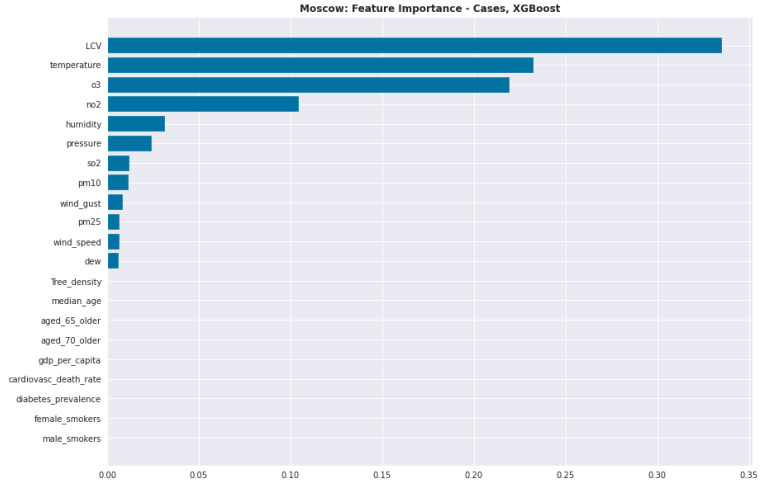


Σχήμα 227 *Feature importance για πρόβλεψη θανάτων, RF, Μόσχα, Ιδία Επεξεργασία*

Παρατηρείται ότι τη μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το O_3, η υγρασία, η πίεση, το dew, το NO_2, οι ριπές του ανέμου, τα PM_{10} και η ταχύτητα του ανέμου. Οι υπόλοιπες μεταβλητές φαίνεται ότι έχουν ελάχιστη, έως καθόλου, βαρύτητα.

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.



Σχήμα 228 *Feature importance για πρόβλεψη θανάτων, XGBoost, Μόσχα, Ιδία Επεξεργασία*

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το O_3, το NO_2, η υγρασία, η πίεση, το SO_2, τα PM_{10}, οι ριπές ανέμου, τα $PM_{2.5}$, η ταχύτητα ανέμου και το dew. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση.

Συμπεράσματα:

Έπειτα από την ανάλυση των παραπάνω διαγραμμάτων, προκύπτουν ορισμένα συμπεράσματα.

Όσον αφορά τα τρία πρώτα μοντέλα πρόβλεψης κρουσμάτων, οι μεταβλητές οι οποίες κυριαρχούν είναι τα PM_{10} και το O_3. Για τον αλγόριθμο RF και για το XGBoost συναντάται έντονη επιρροή της LCV, όμως ως δεύτερη σημαντικότερη μεταβλητή είναι εκείνη της θερμοκρασίας και το όζον καταλαμβάνει την τρίτη θέση. Εν συνεχεία, μικρότερες τιμές επίδρασης παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία, πίεση και άνεμος. Ακόμη, μεταβλητές οι οποίες σχετίζονται με οικονομικούς και ηλικιακούς παράγοντες, φαίνεται να ασκούν μία πολύ μικρή επιρροή. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα καπνιστών και με το πράσινο της πόλης, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

Όσον αφορά τα μοντέλα πρόβλεψης θανάτων, προκύπτει ότι για την πόλη της Μόσχας, οι μεταβλητές οι οποίες φαίνεται να ασκούν σημαντική επιρροή είναι δύο. Η LCV και το όζον. Σε όλα μοντέλα τα οποία αναλύθηκαν, κυριαρχούν οι δύο αυτές μεταβλητές για τις πρώτες δύο θέσεις. Εδώ, συγκριτικά με τις προηγούμενες δύο πόλεις, το LCV φαίνεται να έχει θετική επίδραση, λαμβάνοντας υπόψιν τους συντελεστές από τα τρία πρώτα μοντέλα, ενώ το όζον ασκεί αρνητική επίδραση στην πρόβλεψη. Όσον αφορά τα μοντέλα RF και XGBoost, η μεταβλητή εκείνη η οποία φαίνεται να ασκεί μεγαλύτερη επιρροή είναι η LCV, ακολουθεί η θερμοκρασία και τρίτη θέση καταλαμβάνει το όζον. Στη συνέχεια, μικρότερες τιμές επίδρασης παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία και άνεμος. Ακόμη, μεταβλητές οι οποίες σχετίζονται με οικονομικούς και

ηλικιακούς παράγοντες, φαίνεται να ασκούν μία πολύ μικρή επιρροή. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα καπνιστών και με το πράσινο, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

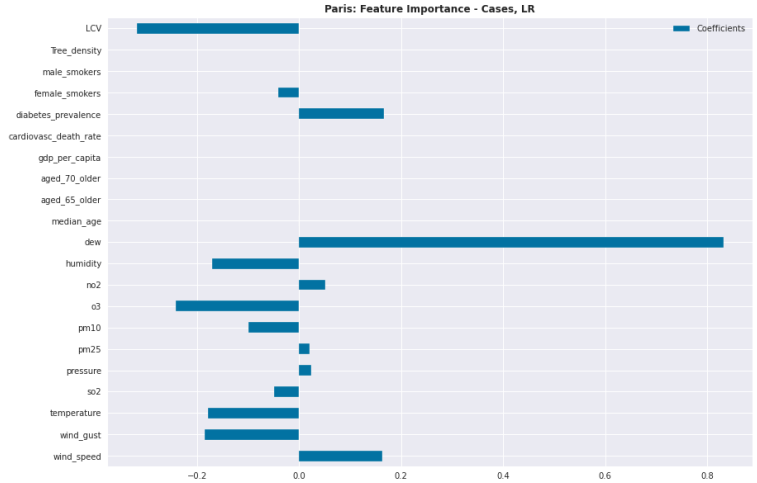
4.4.7.4 *Παρίσι*

Ακολουθεί η τέταρτη πόλη για την οποία θα παρουσιαστούν τα διαγράμματα για τη σημαντικότητα των μεταβλητών. Πρόκειται για την πόλη του Παρισιού. Παρατίθενται τα διαγράμματα τα οποία έχουν προκύψει αρχικά για τις προβλέψεις των κρουσμάτων και στη συνέχεια για τις προβλέψεις των θανάτων, για κάθε αλγόριθμο.

i. Πρόβλεψη Κρουσμάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.



Σχήμα 229 *Feature importance για πρόβλεψη κρουσμάτων, LR, Παρίσι, Ιδία Επεξεργασία*

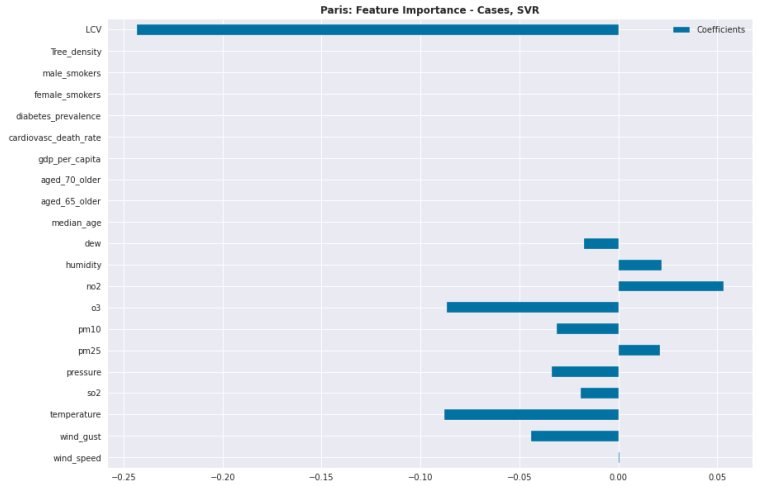
Από το παραπάνω διάγραμμα, εμφανίζονται οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν για την πρόβλεψη, μαζί με την τιμή των συντελεστών τους.

Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή dew. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή της εν λόγω μεταβλητής, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή είναι το dew.

Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης το O_3, οι ριπές ανέμου, η θερμοκρασία, η υγρασία, τα PM_{10} και οι γυναίκες καπνιστές. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η ταχύτητα του ανέμου, ο δείκτης επιπολασμού διαβήτη, το NO_2 και η πίεση με τα $PM_{2.5}$. Δηλαδή, με την αύξηση των παραπάνω μεταβλητών, αναμένεται η αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, οι άνδρες καπνιστές, ηλικιακές ομάδες μεγαλύτερες των 65 και 70 ετών, φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- SVR

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στο Σχήμα το οποίο ακολουθεί.



Σχήμα 230 *Feature importance για πρόβλεψη κρουσμάτων, SVR, Παρίσι, Ιδία Επεξεργασία*

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή του NO_2. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, σημαντικά μεγαλύτερη αρνητική τιμή παρουσιάζει το LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη του LCV.

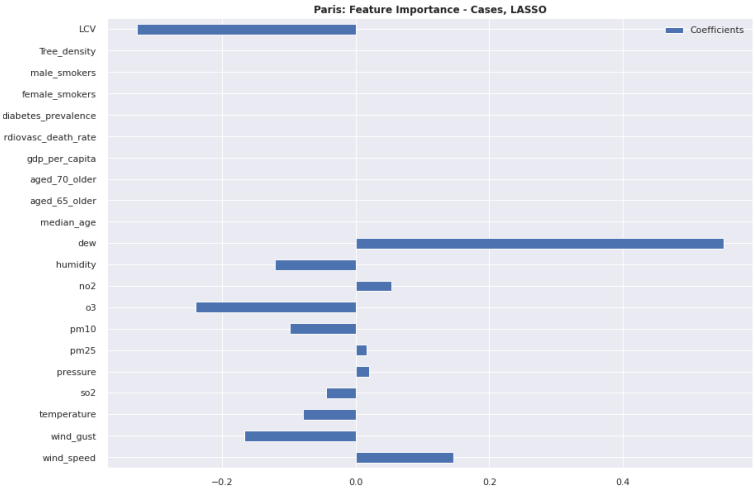
Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης το O_3, η θερμοκρασία, οι ριπές του ανέμου, η πίεση, τα PM_{10} και το dew. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει το NO_2, η υγρασία και τα $PM_{2.5}$. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, προβλήματα διαβήτη, ηλικιακές ομάδες άνω των 65 και 70 ετών και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- <u>LASSO</u>

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.

Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή dew. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, αρκετά μεγαλύτερη αρνητική τιμή παρουσιάζει το LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το dew.

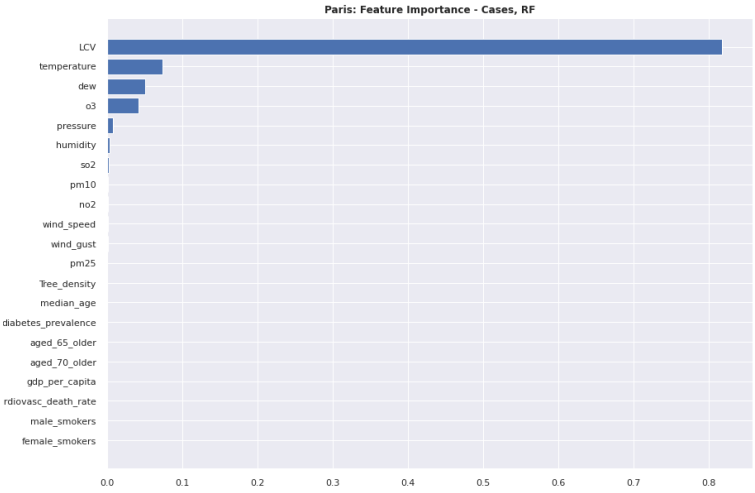
Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει το όζον, οι ριπές ανέμου, η υγρασία, τα PM_{10}, η θερμοκρασία και το SO_2. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η μεταβλητή της ταχύτητας του ανέμου και του NO_2. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η αύξηση των τιμών των κρουσμάτων και με τη μείωσή τους, αναμένεται και μείωση των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, ηλικιακές ομάδες άνω των 65 και 70 ετών, προβλήματα διαβήτη (diabetes prevalence) και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.



Σχήμα 231 *Feature importance για πρόβλεψη κρουσμάτων, LASSO, Παρίσι, Ιδία Επεξεργασία*

- RF Regression

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.



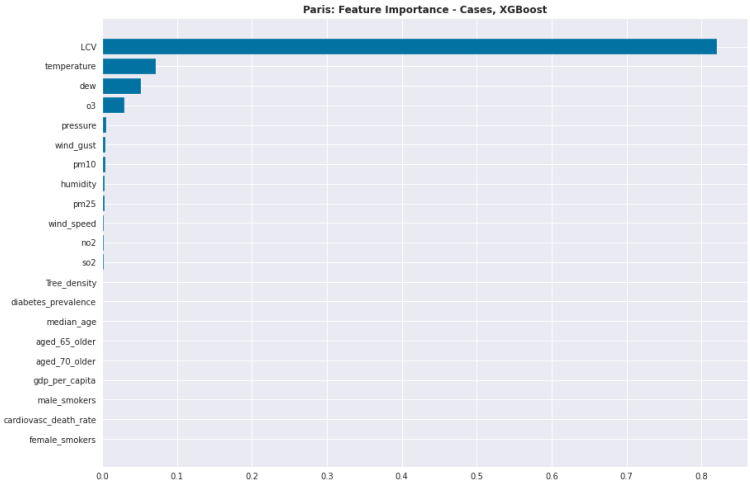
Σχήμα 232 *Feature importance για πρόβλεψη κρουσμάτων, RF, Παρίσι, Ιδία Επεξεργασία*

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το dew, το όζον, η πίεση, η υγρασία και το SO_2. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση στην πρόβλεψη των κρουσμάτων.

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.

Από το διάγραμμα του παρακάτω Σχήμα τος, παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV, με τιμή μεγαλύτερη από 0.8. Ακολουθεί η θερμοκρασία, το dew, το όζον, η πίεση, οι ριπές ανέμου, τα PM10, η υγρασία, τα PM2.5, η ταχύτητα του ανέμου, το NO2 και το SO2, των οποίων οι τιμές δεν ξεπερνάνε το 0.1. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση στην πρόβλεψη των κρουσμάτων. Άρα, και σε αυτήν την περίπτωση, κατά κύριο λόγο μεγαλύτερη επιρροή στην πρόβλεψη ασκεί η μεταβλητή LCV.



Σχήμα 233 *Feature importance για πρόβλεψη κρουσμάτων, XGBoost, Παρίσι, Ιδία Επεξεργασία*

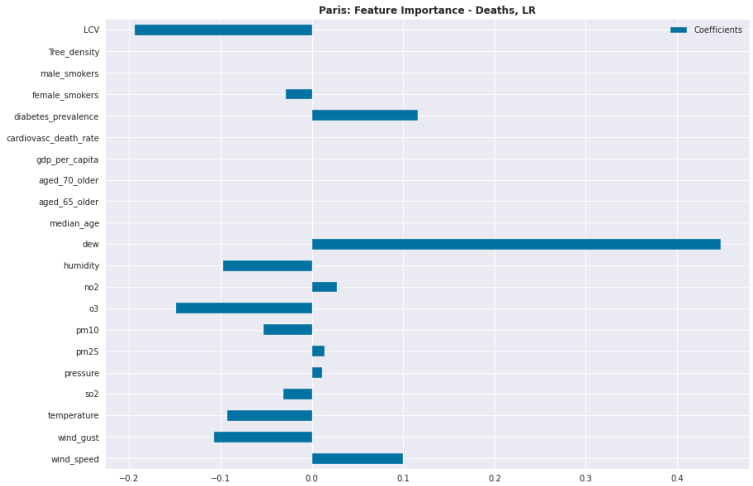
ii. Πρόβλεψη Θανάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.

Στο παρακάτω διάγραμμα, εμφανίζονται οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν για την πρόβλεψη, μαζί με την τιμή των συντελεστών τους.

Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει το dew. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή, καθώς εμφανίζει υψηλότερη τιμή, είναι η ανεξάρτητη μεταβλητή dew.



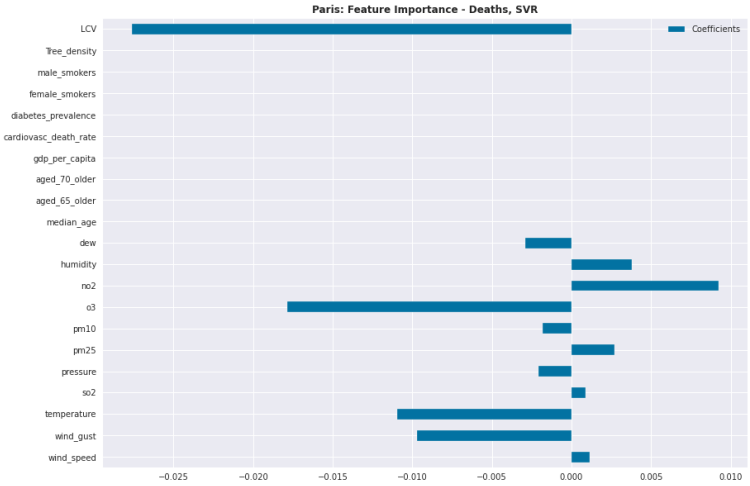
Σχήμα 234 *Feature importance για πρόβλεψη θανάτων με LR, Παρίσι, Ιδία Επεξεργασία*

Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης το O_3, οι ριπές ανέμου, η υγρασία, η θερμοκρασία, τα PM_{10}, το SO_2 και οι γυναίκες καπνιστές. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των θανάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει ο δείκτης επιπολασμού διαβήτη, η ταχύτητα του ανέμου, το NO_2, τα $PM_{2.5}$ και η πίεση. Δηλαδή, με την αύξηση των εν λόγω μεταβλητών, αναμένεται η αύξηση των τιμών των θανάτων. Τέλος,

μεταβλητές όπως άνδρες καπνιστές, μέση ηλικία, ηλικιακές ομάδες μεγαλύτερες των 65 και 70 ετών, φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- SVR

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στη συνέχεια.



Σχήμα 235 *Feature importance για πρόβλεψη θανάτων, SVR, Παρίσι, Ιδία Επεξεργασία*

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή NO_2. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζουν επίσης το O_3, η θερμοκρασία, οι ριπές ανέμου, το dew, η πίεση και τα PM_{10}. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η υγρασία, τα $PM_{2.5}$, η ταχύτητα ανέμου και το SO_2. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, ο προβλήματα διαβήτη, οι ηλικιακές ομάδες άνω των 65 και 70, ο δείκτης θνησιμότητας από καρδιαγγειακή ασθένεια ετών και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

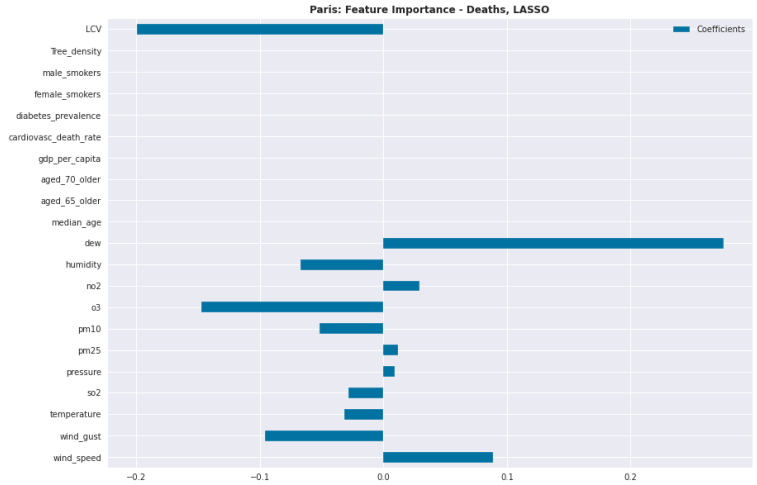
- LASSO

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.

Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή dew. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το dew.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει το O_3, οι ριπές ανέμου, η υγρασία, τα PM_{10}, η θερμοκρασία και το SO_2. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η ταχύτητα ανέμου, το NO_2, τα $PM_{2.5}$ και η πίεση. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η αύξηση των τιμών των θανάτων και με τη

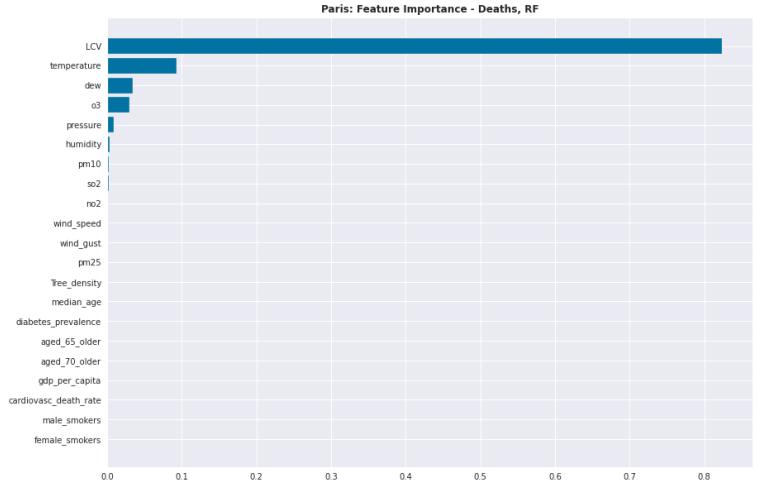
μείωσή τους, αναμένεται και μείωση των θανάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, ηλικιακές ομάδες άνω των 65 και 70 ετών, προβλήματα διαβήτη και το gdp per capita φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.



Σχήμα 236 *Feature importance για πρόβλεψη θανάτων, LASSO, Παρίσι, Ιδία Επεξεργασία*

- RF Regression

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.



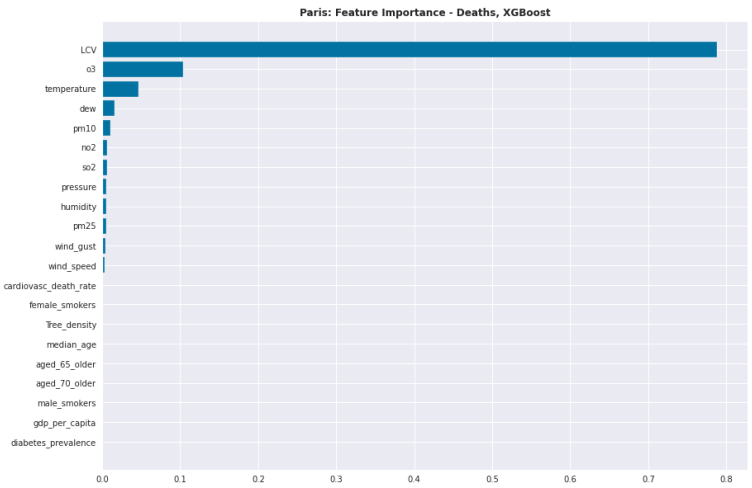
Σχήμα 237 *Feature importance για πρόβλεψη θανάτων, RF, Παρίσι, Ιδία Επεξεργασία*

Παρατηρείται ότι τη μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το dew, το O_3, η πίεση, η υγρασία, τα PM_{10} και το SO_2. Οι υπόλοιπες μεταβλητές φαίνεται ότι έχουν ελάχιστη, έως καθόλου, βαρύτητα.

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το O_3, το dew, τα PM_{10}, το NO_2, το SO_2, η πίεση, η υγρασία, τα $PM_{2.5}$, οι ριπές ανέμου και η ταχύτητα ανέμου. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση.



Σχήμα 238 *Feature importance για πρόβλεψη θανάτων, XGBoost, Παρίσι, Ιδία Επεξεργασία*

Συμπεράσματα:

Έπειτα από την ανάλυση των παραπάνω διαγραμμάτων, προκύπτουν ορισμένα συμπεράσματα.

Όσον αφορά τα τρίτα πρώτα μοντέλα πρόβλεψης κρουσμάτων, οι μεταβλητές οι οποίες κυριαρχούν είναι το LCV, το dew και το NO_2. Για τον αλγόριθμο RF και για το XGBoost συναντάται έντονη επιρροή της LCV, όμως ως δεύτερη σημαντικότερη μεταβλητή είναι εκείνη της θερμοκρασίας και το dew καταλαμβάνει την τρίτη θέση. Εν συνεχεία, μικρότερες τιμές επίδρασης παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία, πίεση και άνεμος. Ακόμη, μεταβλητές οι οποίες σχετίζονται με οικονομικούς και ηλικιακούς παράγοντες, φαίνεται να ασκούν μία πολύ μικρή επιρροή. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα καπνιστών και με το πράσινο της πόλης, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

Όσον αφορά τα μοντέλα πρόβλεψης θανάτων, προκύπτει ότι για την πόλη του Παρισιού, οι μεταβλητές οι οποίες φαίνεται να ασκούν σημαντική επιρροή είναι τρεις. Η LCV, το dew και το όζον. Σε όλα μοντέλα τα οποία αναλύθηκαν, κυριαρχούν οι τρεις αυτές μεταβλητές για τις πρώτες δύο θέσεις. Εδώ, το LCV και το O_3 φαίνεται να έχουν αρνητική επίδραση, λαμβάνοντας υπόψιν τους συντελεστές από τα τρία πρώτα μοντέλα, ενώ το dew ασκεί θετική επίδραση στην πρόβλεψη. Όσον αφορά το μοντέλο RF, η μεταβλητή εκείνη η οποία φαίνεται να ασκεί μεγαλύτερη επιρροή είναι η LCV, ακολουθεί η θερμοκρασία και τρίτη θέση καταλαμβάνει το dew. Για το XGBoost πρώτη θέση παρουσιάζεται για τη LCV, δεύτερη για το όζον και τρίτη για τη θερμοκρασία. Στη συνέχεια, μικρότερες τιμές επίδρασης παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία και άνεμος. Ακόμη, μεταβλητές οι οποίες σχετίζονται με οικονομικούς και ηλικιακούς παράγοντες, φαίνεται να ασκούν μία πολύ μικρή επιρροή. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα καπνιστών και με το πράσινο, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

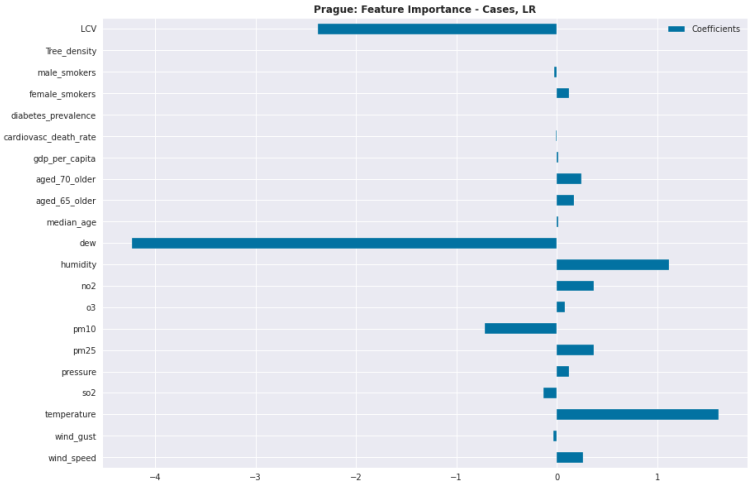
4.4.7.5 Πράγα

Ακολουθεί η τέταρτη πόλη για την οποία θα παρουσιαστούν τα διαγράμματα για τη σημαντικότητα των μεταβλητών. Πρόκειται για την πόλη του Παρισιού. Παρατίθενται τα διαγράμματα τα οποία έχουν προκύψει αρχικά για τις προβλέψεις των κρουσμάτων και στη συνέχεια για τις προβλέψεις των θανάτων, για κάθε αλγόριθμο.

iii. Πρόβλεψη Κρουσμάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.



Σχήμα 239 *Feature importance για πρόβλεψη κρουσμάτων, LR, Πράγα, Ιδία Επεξεργασία*

Από το παραπάνω διάγραμμα, εμφανίζονται οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν για την πρόβλεψη, μαζί με την τιμή των συντελεστών τους.

Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή της θερμοκρασίας. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το dew. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή της εν λόγω μεταβλητής, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή είναι το dew.

Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης το LCV, τα PM_{10} και το SO_2. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η υγρασία, το NO_2, τα $PM_{2.5}$, η ταχύτητα του ανέμου και το O_3. Δηλαδή, με την αύξηση των παραπάνω μεταβλητών, αναμένεται η αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, ηλικιακές ομάδες μεγαλύτερες των 65 και 70 ετών, φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

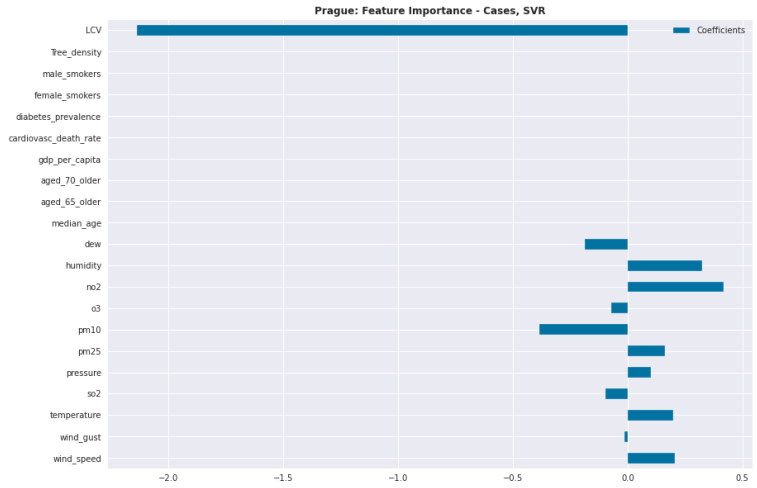
- SVR

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στο Σχήμα το οποίο ακολουθεί.

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή του NO_2. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, σημαντικά μεγαλύτερη αρνητική τιμή παρουσιάζει το LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη του LCV.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζουν επίσης τα PM_{10}, το dew και το O_3. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η υγρασία, η ταχύτητα ανέμου, η θερμοκρασία τα $PM_{2.5}$ και η πίεση. Τέλος, μεταβλητές όπως

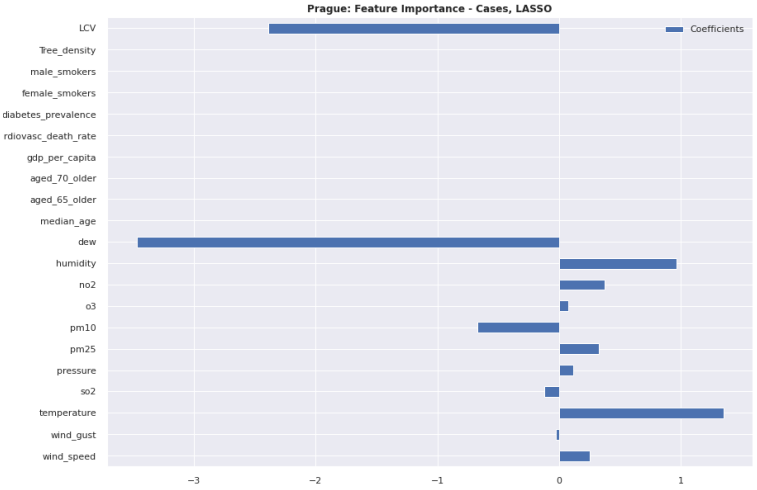
η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, προβλήματα διαβήτη, ηλικιακές ομάδες άνω των 65 και 70 ετών και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.



Σχήμα 240 *Feature importance για πρόβλεψη κρουσμάτων, SVR, Πράγα, Ιδία Επεξεργασία*

- LASSO

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.



Σχήμα 241 *Feature importance για πρόβλεψη κρουσμάτων, LASSO, Πράγα, Ιδία Επεξεργασία*

Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή της θερμοκρασίας. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των κρουσμάτων. Από την άλλη, αρκετά μεγαλύτερη αρνητική τιμή παρουσιάζει το dew. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του εν λόγω δείκτη, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το dew.

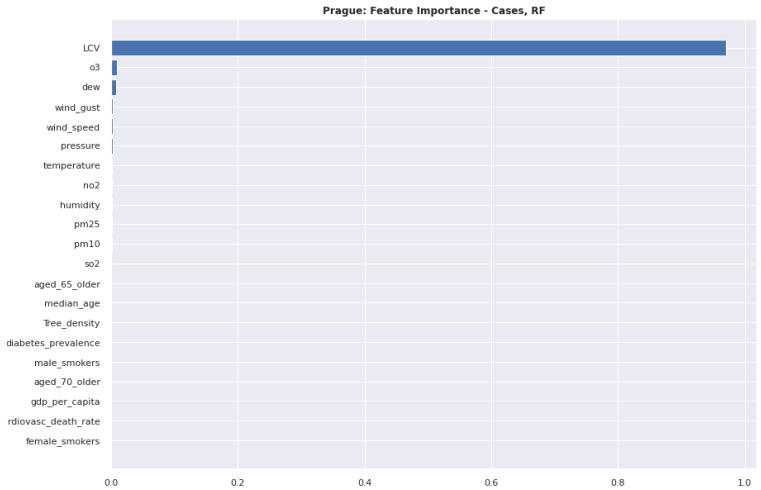
Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει το LCV, τα PM_{10}, και το SO_2. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η μεταβλητή της υγρασίας, του NO_2, των $PM_{2.5}$, της ταχύτητας του ανέμου, της πίεσης και του O_3. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η αύξηση των τιμών των κρουσμάτων και με τη μείωσή τους, αναμένεται και μείωση των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, ηλικιακές ομάδες

άνω των 65 και 70 ετών, προβλήματα διαβήτη (diabetes prevalence) και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- RF Regression

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.

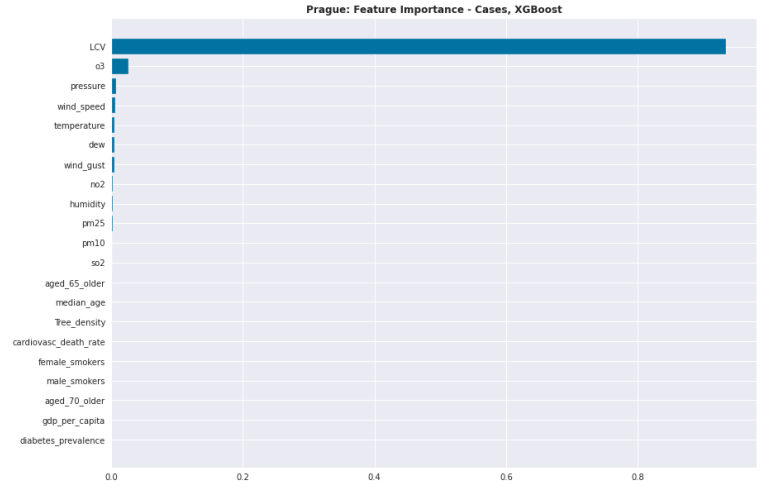
Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί το O_3, το dew, οι ριπές του ανέμου, η ταχύτητα του ανέμου και η πίεση. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση στην πρόβλεψη των κρουσμάτων.



Σχήμα 242 *Feature importance για πρόβλεψη κρουσμάτων, RF, Πράγα, Ιδία Επεξεργασία*

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.



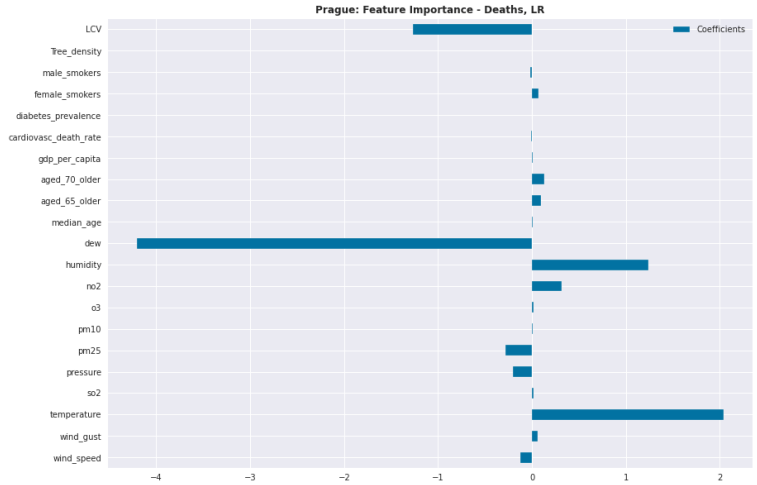
Σχήμα 243 *Feature importance για πρόβλεψη κρουσμάτων, XGBoost, Πράγα, Ιδία Επεξεργασία*

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί το O_3, η πίεση, η ταχύτητα του ανέμου, η θερμοκρασία, το dew, οι ριπές ανέμου, το NO_2, η υγρασία και τα $PM_{2.5}$. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση στην πρόβλεψη των κρουσμάτων.

iv. Πρόβλεψη Θανάτων

- Multiple Linear Regression

Το αποτέλεσμα το οποίο προκύπτει για τον αλγόριθμο της γραμμικής παλινδρόμησης παρουσιάζεται στη συνέχεια.



Σχήμα 244 *Feature importance για πρόβλεψη θανάτων με LR, Πράγα, Ιδία Επεξεργασία*

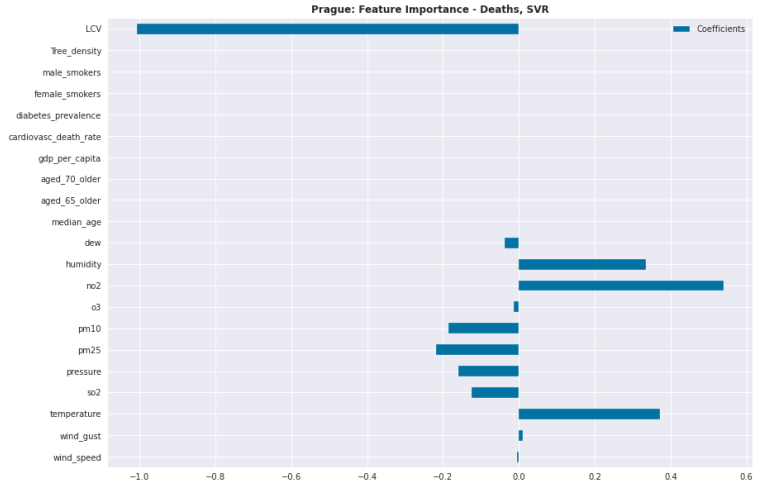
Από το παραπάνω διάγραμμα, εμφανίζονται οι ανεξάρτητες μεταβλητές οι οποίες χρησιμοποιήθηκαν για την πρόβλεψη, μαζί με την τιμή των συντελεστών τους.

Παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η θερμοκρασία. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το dew. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές τις δύο μέγιστες τιμές, εκείνη η οποία φαίνεται να ασκεί περισσότερη επιρροή, καθώς εμφανίζει υψηλότερη τιμή, είναι η ανεξάρτητη μεταβλητή dew.

Αρνητική επίδραση, με μία σχετικά έντονη επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει επίσης το LCV, τα $PM_{2.5}$ και η πίεση. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των θανάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η υγρασία, το NO_2, οι ηλικιακές ομάδες άνω των 70 και 65 ετών και οι γυναίκες καπνιστές. Δηλαδή, με την αύξηση των εν λόγω μεταβλητών, αναμένεται η αύξηση των τιμών των θανάτων. Τέλος, μεταβλητές όπως άνδρες καπνιστές, μέση ηλικία, φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- <u>SVR</u>

Το αποτέλεσμα το οποίο προκύπτει για το SVR αλγόριθμο παρουσιάζεται στη συνέχεια.



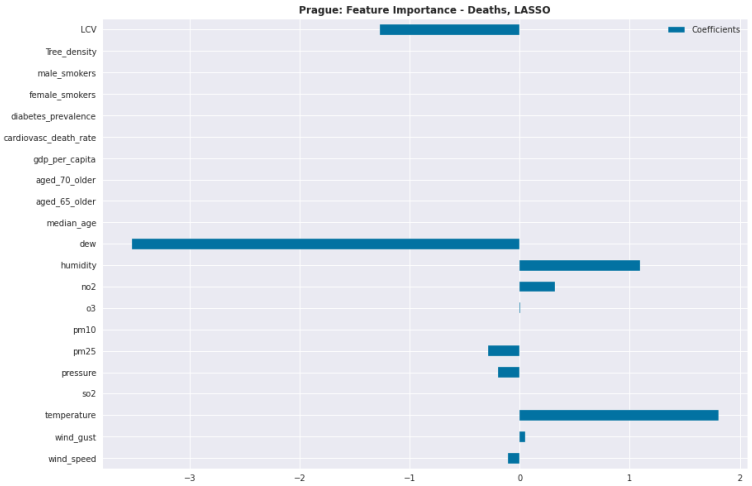
Σχήμα 245 *Feature importance για πρόβλεψη θανάτων, SVR, Πράγα, Ιδία Επεξεργασία*

Στην περίπτωση του SVR, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η μεταβλητή NO_2. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το LCV. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του, τόσο αυξάνεται η τιμή των θανάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το LCV.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζουν τα $PM_{2.5}$, τα PM_{10}, η πίεση, το SO_2, το dew και το O_3. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η θερμοκρασία και η υγρασία. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται αύξηση των τιμών των κρουσμάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, ο προβλήματα διαβήτη, οι ηλικιακές ομάδες άνω των 65 και 70, ο δείκτης θνησιμότητας από καρδιαγγειακή ασθένεια ετών και το ΑΕΠ (gdp per capita) φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- LASSO

Το αποτέλεσμα το οποίο προκύπτει για το LASSO αλγόριθμο παρουσιάζεται στη συνέχεια.



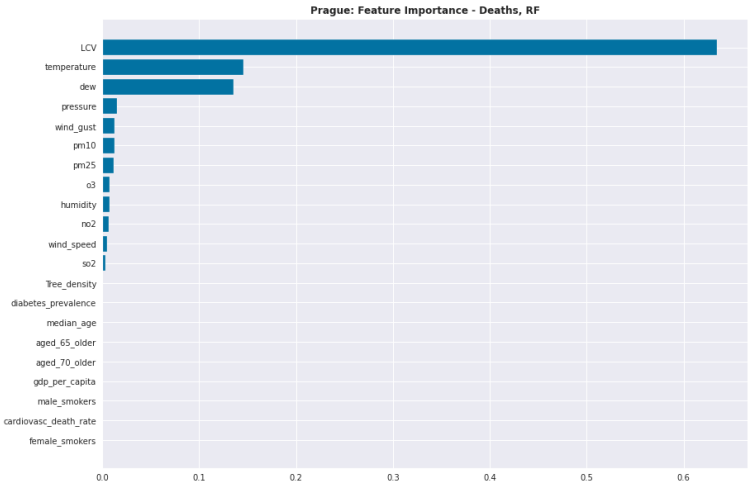
Σχήμα 246 *Feature importance για πρόβλεψη θανάτων, LASSO, Πράγα, Ιδία Επεξεργασία*

Στην περίπτωση του LASSO, παρατηρείται ότι μεγαλύτερη θετική επιρροή φαίνεται να παρουσιάζει η θερμοκρασία. Αυτό, λοιπόν, σημαίνει ότι όσο αυξάνεται η τιμή αυτή, τόσο τείνει να αυξάνεται και η τιμή της εξαρτημένης μεταβλητής, δηλαδή των θανάτων. Από την άλλη, μεγαλύτερη αρνητική τιμή παρουσιάζει το dew. Αυτό δηλώνει ότι, όσο μειώνεται η τιμή του, τόσο αυξάνεται η τιμή των κρουσμάτων. Ανάμεσα σε αυτές τις δύο μεταβλητές, εκείνη η οποία εμφανίζει την μεγαλύτερη, κατ' απόλυτο, τιμή είναι εκείνη για το dew.

Αξιόλογη αρνητική επίδραση, για την προβλεπόμενη μεταβλητή, παρουσιάζει το LCV, τα $PM_{2.5}$, η πίεση και η ταχύτητα του ανέμου. Δηλαδή, μία αύξηση των τιμών αυτών, οδηγεί σε μείωση των κρουσμάτων και vice versa. Αντιθέτως, αξιόλογη θετική επίδραση φαίνεται να εμφανίζει η υγρασία και το NO_2. Δηλαδή, με την αύξηση των μεταβλητών αυτών, αναμένεται η αύξηση των τιμών των θανάτων και με τη μείωσή τους, αναμένεται και μείωση των θανάτων. Τέλος, μεταβλητές όπως η πυκνότητα των δέντρων, άντρες και γυναίκες καπνιστές, ηλικιακές ομάδες άνω των 65 και 70 ετών, προβλήματα διαβήτη και το gdp per capita φαίνεται να ασκούν μηδαμινή, έως καθόλου, επιρροή.

- RF Regression

Το αποτέλεσμα το οποίο προκύπτει για το RF Regression παρουσιάζεται στη συνέχεια.

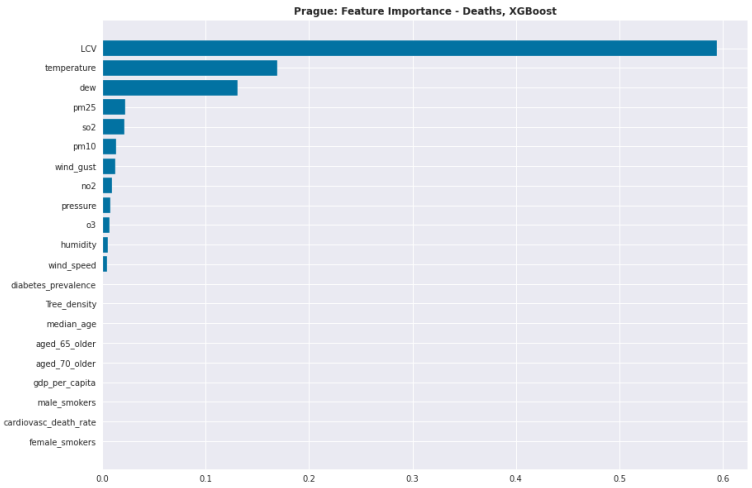


Σχήμα 247 *Feature importance για πρόβλεψη θανάτων, RF, Πράγα, Ιδία Επεξεργασία*

Παρατηρείται ότι τη μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το dew, η πίεση, οι ριπές ανέμου, τα PM_{10}, τα $PM_{2.5}$, το O_3, η υγρασία, το NO_2, η ταχύτητα του ανέμου και το SO_2. Οι υπόλοιπες μεταβλητές φαίνεται ότι έχουν ελάχιστη, έως καθόλου, βαρύτητα.

- XGBoost Regression

Το αποτέλεσμα το οποίο προκύπτει για το XGBoost Regression παρουσιάζεται στη συνέχεια.



Σχήμα 248 *Feature importance για πρόβλεψη θανάτων, XGBoost, Πράγα, Ιδία Επεξεργασία*

Παρατηρείται ότι μεγαλύτερη επιρροή φαίνεται να παρουσιάζει η μεταβλητή του LCV. Ακολουθεί η θερμοκρασία, το dew, τα $PM_{2.5}$, το SO_2, τα PM_{10}, οι ριπές ανέμου, το NO_2, η πίεση, το O_3, η υγρασία και η ταχύτητα του ανέμου. Οι υπόλοιπες μεταβλητές δεν εμφανίζουν κάποια αξιόλογη τιμή, συνεπώς θεωρείται ότι ασκούν μηδαμινή, έως καθόλου, επίδραση.

Συμπεράσματα:

Έπειτα από την ανάλυση των παραπάνω διαγραμμάτων, προκύπτουν ορισμένα συμπεράσματα.

Όσον αφορά τα τρία πρώτα μοντέλα πρόβλεψης κρουσμάτων, οι μεταβλητές οι οποίες κυριαρχούν είναι το dew, το LCV, η θερμοκρασία και το NO_2. Για τον αλγόριθμο RF και συναντάται έντονη επιρροή της LCV, ως δεύτερη σημαντικότερη μεταβλητή είναι το O_3 και το dew καταλαμβάνει την τρίτη θέση. Για το XGBoost, πρώτη μεταβλητή στην ιεράρχηση είναι το LCV, δεύτερη το O_3 και τρίτη η πίεση. Εν συνεχεία, μικρότερες τιμές επίδρασης

παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία, πίεση και άνεμος. Ακόμη, μεταβλητές οι οποίες σχετίζονται με οικονομικούς και ηλικιακούς παράγοντες, φαίνεται να ασκούν μία πολύ μικρή επιρροή. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα καπνιστών και με το πράσινο της πόλης, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

Όσον αφορά τα μοντέλα πρόβλεψης θανάτων, προκύπτει ότι για την πόλη της Πράγας, οι μεταβλητές οι οποίες φαίνεται να ασκούν σημαντική για τα τρία πρώτα μοντέλα είναι η θερμοκρασία, το dew, το LCV και το NO_2. Όσον αφορά τα μοντέλα RF και XGBoost, η μεταβλητή εκείνη η οποία φαίνεται να ασκεί μεγαλύτερη επιρροή είναι η LCV, ακολουθεί η θερμοκρασία και τρίτη θέση καταλαμβάνει το dew. Στη συνέχεια, μικρότερες τιμές επίδρασης παρουσιάζουν μεταβλητές οι οποίες σχετίζονται κυρίως με την ατμόσφαιρα, όπως ρύποι, υγρασία και άνεμος. Ακόμη, μεταβλητές οι οποίες σχετίζονται με οικονομικούς και ηλικιακούς παράγοντες, φαίνεται να ασκούν μία πολύ μικρή επιρροή. Τέλος, παρατηρείται ότι μεταβλητές οι οποίες σχετίζονται με φύλα καπνιστών και με το πράσινο, δεν διαδραματίζουν σημαντικό ρόλο στην πρόβλεψη.

4.4.8 Γενικά Συμπεράσματα

Από την παραπάνω ανάλυση τόσο για τις επιδόσεις των μοντέλων, όσο και για τη βαρύτητα εκάστης ανεξάρτητης μεταβλητής στις προβλέψεις δύναται να διεξαχθούν ορισμένα συμπεράσματα.

Το γενικό συμπέρασμα από τα παραπάνω είναι ότι τα αποτελέσματα για κάθε πόλη και για κάθε μοντέλο δεν ήταν το ίδιο ικανοποιητικά. Ορισμένες πόλεις φάνηκε να έχουν καλύτερες αποδόσεις από κάποιες άλλες. Το ίδιο ισχύει και για ορισμένα μοντέλα.

Αναλυτικότερα, από τις πέντε πόλεις, τα μοντέλα εκείνης η οποία ξεχώρισε για την απόδοσή της, για τις έξι μεθόδους και για τα δύο είδη προβλέψεων, ήταν εκείνα για την πόλη του Παρισιού. Αντιθέτως, τα μοντέλα για την πόλη της Μόσχας φαίνεται να ξεχώρισαν για την αρνητική τους προσαρμογή στα πραγματικά δεδομένα, άρα και για τη χαμηλή τους απόδοση και στις περιπτώσεις πρόβλεψης κρουσμάτων και στις περιπτώσεις πρόβλεψης θανάτων Covid-19. Σε αυτό το σημείο υπενθυμίζεται ότι δεδομένα για το Παρίσι χρησιμοποιούνται από τις 03/04/2020 έως τις 29/12/2020, ενώ για τη Μόσχα τα δεδομένα ξεκινάνε περίπου ένα μήνα πριν, 11/03/2020 έως τις 29/12/2020. Τα δεδομένα εκπαίδευσης επιλέχθηκε, λόγω του μικρού σχετικά όγκου δεδομένων, να αποτελούν το 80%. Συνεπώς, για τις δύο αυτές περιπτώσεις, αν και η Μόσχα έχει μεγαλύτερο αριθμό δεδομένων εκπαίδευσης, από ότι το Παρίσι, δεν φαίνεται να αποδίδει εξίσου ικανοποιητικά.

Όσον αφορά τα μοντέλα τα οποία χρησιμοποιήθηκαν, δύο από τα έξι εμφάνισαν περισσότερο αξιόλογα αποτελέσματα, για την πλειονότητα των μελετώμενων πόλεων και για τα δύο είδη προβλέψεων. Πρόκειται για το RF Regression και το XGBoost Regression. Στα εν λόγω μοντέλα εμφανίστηκαν υψηλές τιμές R^2 και EVS, χαμηλές δε τιμές για MAPE, MSE, RMSE και MAE. Απεναντίας, το μοντέλο εκείνο στο οποίο εμφανίστηκαν οι χαμηλότερες αποδόσεις για έκαστη πόλη, ήταν εκείνο του SVR. Σε αυτό το μοντέλο ορισμένες από τις μετρικές ήταν πάρα πολύ χαμηλές. Για να μην είναι πλήρως αυθαίρετη η σύγκριση μοντέλων, θεωρήθηκε ως μοντέλο βάσης (Baseline Model) το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης. Το συγκεκριμένο μοντέλο, δεν εμφάνισε άρτια προσαρμογή στις πραγματικές τιμές, όμως, τα αποτελέσματά του ήταν αρκετά αξιοπρεπή. Ως εκ τούτου, επιλέχθηκε να είναι το μοντέλο βάσης, σύμφωνα με το οποίο θα αξιολογηθούν τα υπόλοιπα δημιουργηθέντα μοντέλα.

Τέλος, σχετικά με τη βαρύτητα των μεταβλητών στις προβλέψεις, εκείνες οι οποίες φαίνεται να ξεχωρίζουν είναι η LCV και οι μεταβλητές οι οποίες σχετίζονται με μετεωρολογικά φαινόμενα και με ατμοσφαιρική ρύπανση. Παρατηρήθηκε ότι δεν εμφανίζουν οι ίδιες μεταβλητές, την ίδια βαρύτητα σε κάθε πόλη και σε κάθε μοντέλο πρόβλεψης. Διαφοροποιούνται, τόσο ανάλογα με την πόλη, όσο ανάλογα και με το εκάστοτε μέγεθος πρόβλεψης. Σε γενικές γραμμές, όμως, φαίνεται να κυριαρχεί η μεταβλητή LCV, για όλες τις πόλεις, για όλα τα μοντέλα και στα δύο είδη προβλέψεων. Όμως, οι υπόλοιπες μεταβλητές διαφοροποιούνται. Άρα, συμπεραίνεται ότι κάθε πόλη, με κάθε μοντέλο και για κάθε πρόβλεψη, φαίνεται να εμφανίζει τις δικές της βαρύτητες στις μεταβλητές. Να σημειωθεί ότι, όπως ήδη αναφέρθηκε στο Κεφάλαιο 2, μία προγενέστερη μελέτη δείχνει ότι η ύπαρξη βλάστησης έχει αρνητική επίδραση στο φαινόμενο διάδοσης του κορονοϊού.

Κεφάλαιο 5

Επίλογος

5.1 Εισαγωγή

Κλείνοντας το κεφάλαιο αυτό, κρίνεται σκόπιμο να παρουσιαστούν τα τελικά συμπεράσματα τα οποία προέκυψαν από τη μοντελοποίηση και την ανάλυση των χωροχρονικών επιδημιολογικών δεδομένων, για τις μελετώμενες πόλεις. Στη συνέχεια, ακολουθούν ορισμένες προτάσεις οι οποίες θα μπορούσαν να εφαρμοσθούν σε κάποια μελλοντική εργασία και έρευνα.

5.2 Πορίσματα

Καταρχάς, ορισμένα από τα δεδομένα τα οποία χρησιμοποιήθηκαν για την παρούσα ανάλυση είναι δυναμικά. Πρόκειται για μεταβλητές οι οποίες μεταβάλλονται με το χρόνο. Τέτοιες μεταβλητές σχετίζονται με τη ρύπανση του αέρα και με τα μετεωρολογικά φαινόμενα. Δηλαδή, περιγράφουν συνθήκες διαφορετικών περιοχών για μία συγκεκριμένη χρονική περίοδο. Εξ ορισμού, η ανάλυση και η μοντελοποίηση των εν λόγω δεδομένων είναι ένα σύνθετο ζήτημα. Σε κάθε περίπτωση, χρειάζεται να λαμβάνεται υπόψιν το είδος της εφαρμογής, ώστε να επιλεχθεί και η κατάλληλη συχνότητα μετρήσεων. Όπως γίνεται αντιληπτό, με βάση τα παραπάνω δεδομένα, η επιλογή αυτή επιφέρει και ένα κόστος, ακριβώς λόγω των παραδοχών. Εδώ, πληροφορίες οι οποίες σχετίζονται με ατμοσφαιρικούς ρύπους και με μετεωρολογικά φαινόμενα «έχασαν» πληροφορία, καθώς το διάστημα ανάλυσης ορίστηκε ως μία ημέρα. Άλλωστε, θα ήταν αδύνατο να θεωρηθεί ότι ο χρόνος είναι στατικός σε αυτές τις περιπτώσεις.

Εν συνεχεία, όπως παρουσιάστηκε και στην 4.4.8, δεν ήταν η απόδοση όλων των μοντέλων το ίδιο ικανοποιητική. Ξεχώρισαν κατά κύριο λόγο δύο μέθοδοι πρόβλεψης. Η πρώτη ήταν το RF Regression και η δεύτερη το XGBoost Regression.

Όπως παρουσιάστηκε και στην ανάλυση, τα μοντέλα αυτά εμφανίζουν χαμηλές τιμές σε μετρικές αξιολόγησης οι οποίες αφορούν τις ίδιες τις προβλέψεις, όπως είναι το σφάλμα MAE και RMSE. Ο χαρακτηρισμός «χαμηλές τιμές», έχει ως σημείο αναφοράς τις πραγματικές τιμές των προβλεπόμενων μεγεθών. Ακόμη, εμφανίζουν χαμηλές τιμές για το σφάλμα του MAPE. Αξίζει να επισημανθεί μία ακόμη φορά ότι οι χαμηλές τιμές MAPE δηλώνουν ότι ένα μοντέλο εμφανίζει μεγαλύτερη ακρίβεια. Επίσης, εμφανίζουν υψηλές τιμές για το R^2 και για το EVS. Οι υψηλές τιμές, σε αυτήν την περίπτωση, φανερώνουν τη συσχέτιση ανάμεσα στις ανεξάρτητες και στην εκάστοτε εξαρτημένη μεταβλητή. Ένα επιπρόσθετο γεγονός το οποίο φανερώνει ότι τα εν λόγω μοντέλα έχουν κάνει καλή προσαρμογή στα δύο μελετώμενα προβλήματα είναι οι τιμές για το MSE. Πρόκειται για αρκετά χαμηλές τιμές, εάν συγκριθούν με τις αντίστοιχες τιμές των υπόλοιπων μοντέλων.

Θα αποτελούσε σοβαρή παράλειψη να μην τονισθεί και η αξία των διαγραμμάτων, στην αξιολόγηση των μοντέλων πρόβλεψης. Από το διάγραμμα των προβλεπόμενων τιμών και των πραγματικών τιμών, γίνεται αντιληπτή η αξιόλογη προσαρμογή των εν λόγω μοντέλων, καθώς οι τιμές κατά κύριο λόγο επικαλύπτονται και διαφοροποιούνται ελάχιστα για πολύ μικρά χρονικά τμήματα. Ακόμη, από το διάγραμμα των υπολοίπων γίνονται αντιληπτές οι μικρές διαφορές ανάμεσα στις δύο αυτές τιμές.

Από αυτές τις δύο μεθόδους, επιλέγεται ως βέλτιστη, τόσο για την πρόβλεψη κρουσμάτων όσο και για την πρόβλεψη θανάτων, η μέθοδος RF Regression. Η επιλογή αυτή έγινε αφενός βάσει του μεγέθους των μετρικών, για όλες τις πόλεις και για τα δύο αναπτυχθέντα μοντέλα πρόβλεψης, κρουσμάτων, θανάτων, αφετέρου βάσει των «ιστορικών διαγραμμάτων», στα οποία παρουσιάστηκαν οι τιμές πρόβλεψης και οι πραγματικές τιμές για ένα συγκεκριμένο χρονικό διάστημα. Στο RF Regression, η πλειονότητα των μοντέλων εκάστης πόλης, φαινόταν να έκανε αρκετά καλή προσαρμογή στα πραγματικά δεδομένα.

Ένα σημαντικό κομμάτι της τρέχουσας εργασίας ήταν ο εντοπισμός του βάρους εκάστης μεταβλητής στην πρόβλεψη. Πρόκειται για ένα ζήτημα το οποίο άπτεται στο κομμάτι της εξηγήσιμης τεχνητής νοημοσύνης. Επίσης, ο εντοπισμός του βάρους, φαίνεται να αποτελεί απάντηση στο ερώτημα «Ποια χαρακτηριστικά, από το σετ δεδομένων, πρόκειται να λειτουργήσουν καλά για το μοντέλο μηχανικής μάθησης;». Αυτό το οποίο προέκυψε από την ανάλυση, είναι ότι παράγοντες όπως ατμοσφαιρικοί ρυπαντές καθώς επίσης και μετεωρολογικοί παράγοντες, εμφάνισαν σημαντικό βάρος στις προβλέψεις. Ακόμη, σημαντικό βάρος εμφάνισε και η μεταβλητή LCV, η οποία σχετίζεται με τη βλάστηση. Από την άλλη, μεταβλητές οι οποίες σχετίζονται με ζητήματα υγείας, όπως καρδιαγγειακά ζητήματα, διαβητολογικά, μεταβλητές όπως οι ηλικιακές ομάδες άνω των 65 ετών και οι καπνιστές, φαίνεται να άσκησαν μία επιρροή, όμως όπου εμφανιζόταν, η επιρροή αυτή ήταν ανεπαίσθητη.

Συγκεκριμένα, εκείνη η μεταβλητή η οποία εμφάνισε τη μεγαλύτερη βαρύτητα, για την πλειονότητα των μοντέλων, είναι η μεταβλητή της Land Cover in Vegetation, δηλαδή της βλάστησης. Σύμφωνα και με προηγούμενες έρευνες και εργασίες, όπως αναλύθηκαν στο Κεφάλαιο 2, η αστική βλάστηση εμφανίζει σημαντικό ρόλο στη διάδοση του κορονοϊού. Αν και δεν έχει αποδοθεί κάποια ακριβής ερμηνεία από τις προηγούμενες έρευνες, αυτό το οποίο εικάζεται στην παρούσα διπλωματική, έχει να κάνει με το γεγονός ότι η βλάστηση μπορεί να επιδράσει στη μείωση της διάδοσης και των θανάτων κορονοϊού, με τρεις εκδοχές. Πρωτίστως, η ύπαρξη πρασίνου στις πόλεις βοηθάει στη βελτίωση του μικροκλίματος, άρα και της ατμόσφαιρας. Αυτό έχει ως αποτέλεσμα, την ελάττωση των ατμοσφαιρικών ρύπων. Στη συνέχεια, εάν σε μία αστική περιοχή η χρήση γης δεν είναι βλάστηση, τότε η χρήση η οποία θα κυριαρχήσει είναι αρκετά πιθανό να περιλαμβάνει ανθρώπινες δραστηριότητες. Ως εκ τούτου, οι συγκεντρώσεις των ανθρώπων σε μία περιοχή θα αυξηθούν, άρα είναι αρκετά πιθανό να αυξηθεί και η διάδοση του κορονοϊού. Τέλος, σε αστικές περιοχές με βλάστηση, ο κόσμος μπορεί να κρατήσει τις κατάλληλες αποστάσεις, όπως αυτές ορίζονται κάθε φορά από τον αρμόδιο φορέα. Συνεπώς, γίνεται φανερό ότι χρειάζεται να λαμβάνονται υπόψιν όλες οι παράμετροι στον αστικό σχεδιασμό. Ένας καλύτερος αστικός σχεδιασμός, συντελεί και σε καλύτερη διαχείριση κοινωνικών «κρίσεων».

Τέλος, χρειάζεται να ληφθεί υπόψιν ότι είναι αρκετά πιθανό να εμφανίζονται στατιστικά λάθη, κατά τη διαδικασία υπολογισμού του Feature Importance. Αυτό, μπορεί να συμβαίνει όταν το μοντέλο εμφανίζει μεροληψία για ορισμένες μεταβλητές, έναντι κάποιων άλλων. Για παράδειγμα, η μεταβλητή LCV πιθανότατα εμφανίζει σημαντικό βάρος στις προβλέψεις διότι το μοντέλο τείνει να έχει μία προκατάληψη ως προς εκείνη τη μεταβλητή. Παρατηρώντας το dataset, αυτό το οποίο προκύπτει είναι ότι η τιμή του LCV μεταβάλλεται ανά μήνα, ενώ οι τιμές άλλων μεταβλητών οι οποίες δεν φάνηκε να εμφανίζουν στατιστική σημασία, είχαν σταθερή τιμή για όλη τη χρονική περίοδο μελέτης. Συνεπώς, τα αποτελέσματα μίας ερευνητής εργασίας χρειάζεται να τεθούν σε έλεγχο και σε περαιτέρω ανάλυση.

5.3 Μελλοντικές Κατευθύνσεις

Εκτός από την παράθεση συμπερασμάτων, θεωρείται ορθό να παρουσιαστούν ορισμένες προτάσεις οι οποίες σχετίζονται με πιθανή μελλοντική εργασία. Εκείνες προκύπτουν έπειτα από μία ενδελεχή ενασχόληση τόσο με τη μοντελοποίηση όσο και με την ανάλυση των χωροχρονικών επιδημιολογικών δεδομένων.

Αρχικά, θα μπορούσε να γίνει χρήση όγκου δεδομένων για μεγαλύτερο χρονικό διάστημα. Στην παρούσα εργασία, η περίοδος μελέτης αφορούσε τους πρώτους μήνες εμφάνισης της πανδημίας. Θεωρείται, ότι θα εμφάνιζε αρκετό ενδιαφέρον να χρησιμοποιούταν δεδομένα για ακόμη μεγαλύτερο χρονικό διάστημα. Άλλωστε, δύο χρόνια μετά, οι ημερήσιες καταγραφές των κρουσμάτων και των θανάτων εξακολουθούν να υπάρχουν στα δελτία ειδήσεων. Συνεπώς, πρόκειται για μία βάση δεδομένων η οποία αρχίζει αρκετά να εμπλουτίζεται. Η εκπαίδευση και ο έλεγχος του προτεινόμενου μοντέλου θα γίνει με μεγαλύτερο πλήθος δεδομένων. Συνεπώς, είναι πολύ πιθανό να βελτιωθεί και η ακρίβεια προσαρμογής των μοντέλων στις πραγματικές τιμές των κρουσμάτων και των θανάτων.

Άπαξ και χρησιμοποιηθεί μεγαλύτερος όγκος δεδομένων, θα μπορούσε να γίνει εφαρμογή διαφορετικών μοντέλων πρόβλεψης. Θα παρουσίαζε αρκετό ενδιαφέρον η χρήση στατιστικών μοντέλων ανάλυσης, όπως είναι το ARIMA, καθώς επίσης και χρήση Τεχνητών Νευρωνικών Δικτύων, όπως το LSTM. Τα αποτελέσματα των μεθόδων αυτών θα μπορούσαν να συγκριθούν λαμβάνοντας ως baseline model το μοντέλο του Random Forest το οποίο δημιουργήθηκε στην παρούσα εργασία. Πρόκειται για μεθόδους οι οποίες δεν χρησιμοποιήθηκαν στην παρούσα ανάλυση και θα μπορούσαν τα αποτελέσματά τους να συγκριθούν με εκείνα των χρησιμοποιηθέντων αλγορίθμων Μηχανικής Μάθησης.

Επιπροσθέτως, αφού τονίστηκε η κρισιμότητα επιλογής της κατάλληλης χρονικής κλίμακας, θα μπορούσε να γίνει μία μελέτη με μεγαλύτερο χρονικό βαθμό λεπτομέρειας. Αυτό σημαίνει ότι το διάστημα μελέτης δεν θα αποτελούν πλέον οι 24ώρες, αλλά ένα μικρότερο χρονικό τμήμα. Πιθανότατα, με αυτόν τον τρόπο δεν θα χάνεται σημαντική πληροφορία στις τιμές οι οποίες αφορούν φαινόμενα ατμοσφαιρικής ρύπανσης, μετεωρολογικά φαινόμενα και φαινόμενα μετακίνησης και συγκέντρωσης πληθυσμού.

Ακόμη, θα παρουσίαζε ενδιαφέρον η εκπαίδευση ενός μοντέλου με ένα συγκεκριμένο σετ δεδομένων και ο έλεγχος της απόδοσής του με χρήση «άγνωστων» σε αυτό δεδομένων. Με αυτόν τον τρόπο, θα ελεγχόταν και θα εξαγόταν συμπεράσματα σχετικά με τόσο καλά μπορεί ένα μοντέλο μηχανικής μάθησης να προσαρμοστεί και να προβλέψει τιμές, όταν του δίνονται εντελώς «διαφορετικά» δεδομένα, από εκείνα στα οποία έχει «προσαρμοστεί».

Τέλος, εφόσον προηγήθηκε η ανάλυση του βάρους εκάστης μεταβλητής, κρίνεται σκόπιμο να τονισθεί ότι θα μπορούσε να γίνει μία τροποποίηση των χρησιμοποιηθέντων μεταβλητών. Δηλαδή, μία προσθήκη μεταβλητών, όπως για παράδειγμα είναι το πλήθος του εμβολιασμένου πληθυσμού, θα εμφάνιζε αρκετό ενδιαφέρον στην ερμηνεία και την κατανόηση της πανδημίας. Εκτός από την προσθήκη, μία δοκιμή να αφαιρεθούν μεταβλητές με μικρό βάρος ή μεταβλητές οι οποίες εμφάνισαν σημαντικό βάρος στις προβλέψεις, θα έδειχνε εάν το μοντέλο οδηγείται σε καλύτερες, ή όχι, προβλέψεις. Για παράδειγμα, θα μπορούσε να αφαιρεθεί η μεταβλητή LCV, η οποία εμφανίζει σημαντικό βάρος στις προβλέψεις για την πλειονότητα των δημιουργηθέντων μοντέλων. Με αυτόν τον τρόπο θα εντοπιζόταν πώς δρουν οι υπόλοιπες μεταβλητές στην πρόβλεψη και θα ελεγχόταν εάν η αφαίρεση της εν λόγω μεταβλητής θα είχε επιρροή στις σχέσεις των εναπομεινάντων μεταβλητών.

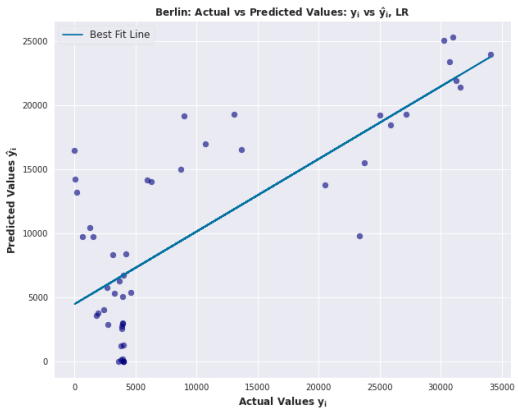
Παράρτημα Α

Στο εν λόγω παράρτημα παρουσιάζονται τα αποτελέσματα για το μοντέλο πρόβλεψης Multiple Linear Regression.

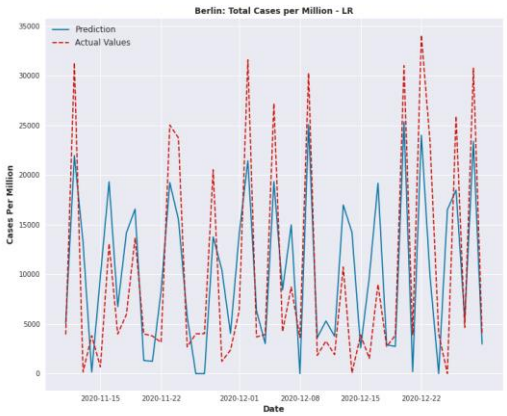
Πρόβλεψη Κρουσμάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	6957.88 (cases per million)
R^2	0.595
EVS	0.596
MAE	5799.41 (cases per million)
MAPE	38.24%
MSE	46466532.40

Πίνακας 63 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, Linear Regression, Βερολίνο, Ιδία Επεξεργασία*

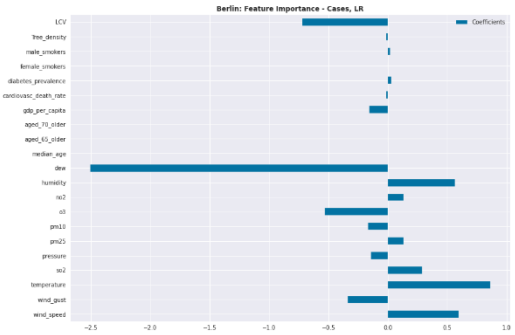


Σχήμα 249 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, Linear Regression, Βερολίνο



Σχήμα 250 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, Linear Regression, Βερολίνο

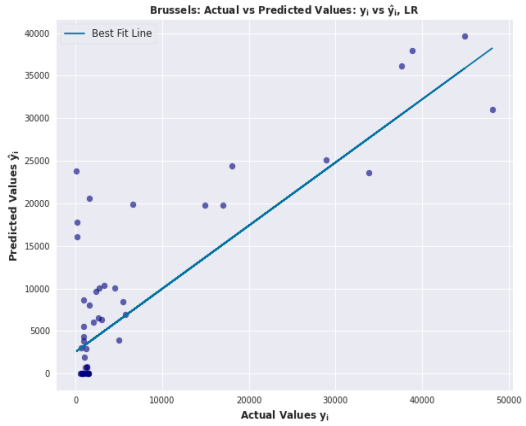
Σχήμα 251 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, Linear Regression, Βερολίνο



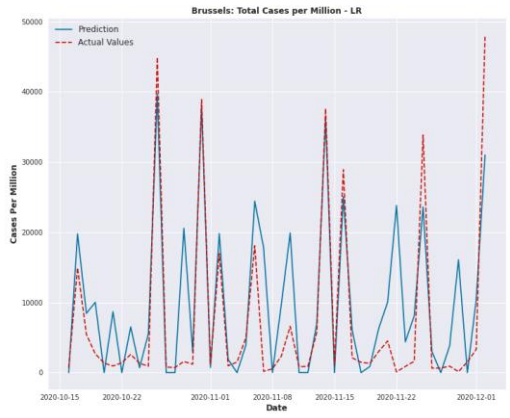
Σχήμα 252 Feature importance για πρόβλεψη κρουσμάτων, Linear Regression, Βερολίνο

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	8027.35 (cases per million)
R^2	0.664
EVS	0.704
MAE	6075.32 (cases per million)
MAPE	12.14%
MSE	54016503.52

Πίνακας 64 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, Linear Regression, Βρυξέλλες, Ιδία Επεξεργασία*

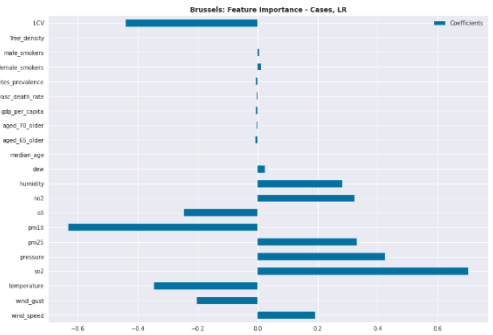


Σχήμα 253 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, Linear Regression, Βρυξέλλες



Σχήμα 254 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, Linear Regression, Βρυξέλλες

Σχήμα 255 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, Linear Regression, Βρυξέλλες

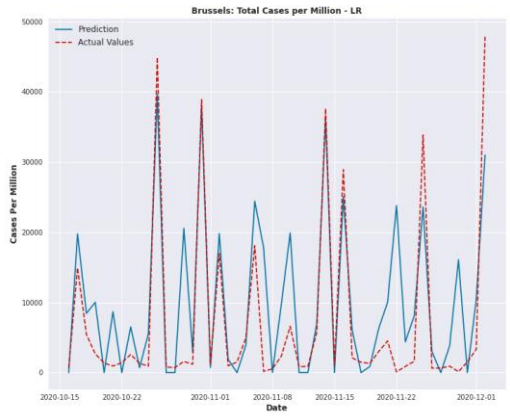


Σχήμα 256 Feature importance για πρόβλεψη κρουσμάτων, Linear Regression, Βρυξέλλες

Μετρική Αξιολόγησης	Λισαβόνα
RMSE	3004.91 (cases per million)
R^2	0.513
EVS	0.527
MAE	2248.08 (cases per million)
MAPE	21.14%
MSE	8701702.47

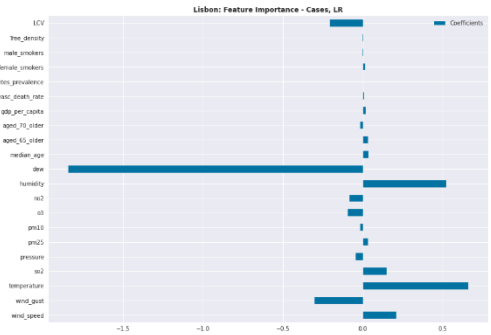
Πίνακας 65 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, Linear Regression, Λισαβόνα, Ιδία Επεξεργασία*

Σχήμα 257 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, Linear Regression, Λισαβόνα



Σχήμα 258 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, Linear Regression, Λισαβόνα

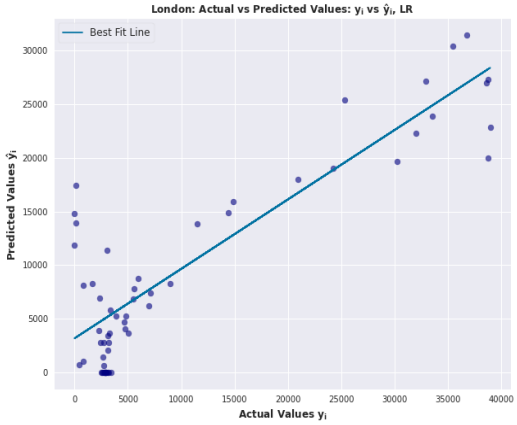
Σχήμα 259 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, Linear Regression, Λισαβόνα



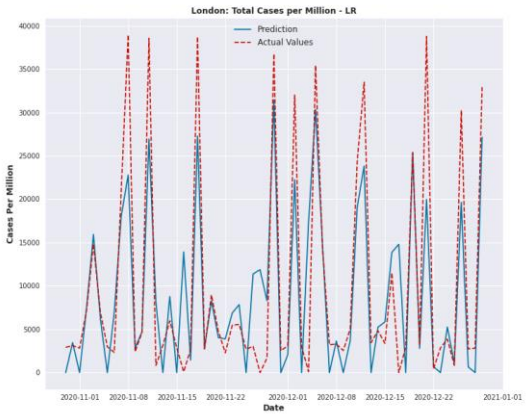
Σχήμα 260 Feature importance για πρόβλεψη κρουσμάτων, Linear Regression, Λισαβόνα

Μετρική Αξιολόγησης	Λονδίνο
RMSE	6913.75 (cases per million)
R^2	0.737
EVS	0.742
MAE	5024.07 (cases per million)
MAPE	776.17%
MSE	41071126.66

Πίνακας 66 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, Linear Regression, Λονδίνο, Ιδία Επεξεργασία*

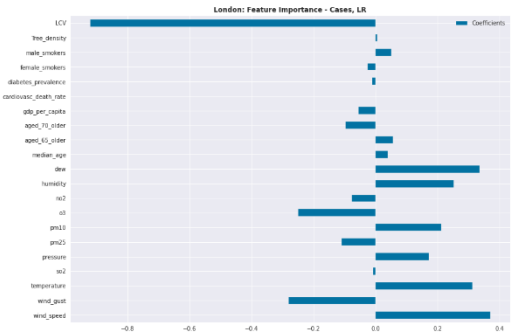


Σχήμα 261 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, Linear Regression, Λονδίνο



Σχήμα 262 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, Linear Regression, Λονδίνο

Σχήμα 263 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, Linear Regression, Λονδίνο

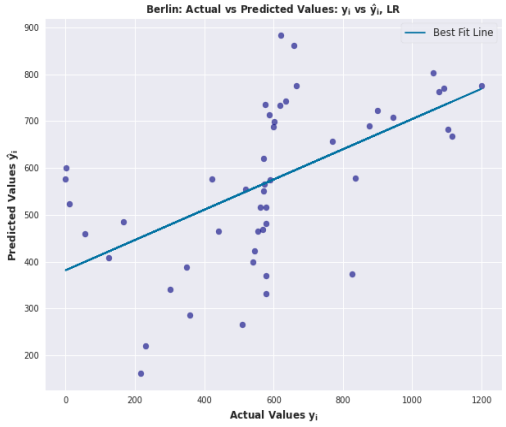


Σχήμα 264 Feature importance για πρόβλεψη κρουσμάτων, Linear Regression, Λονδίνο

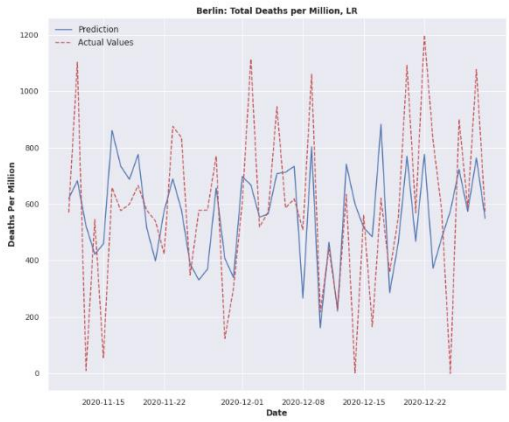
Πρόβλεψη Θανάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	247.34 (deaths per million)
R^2	0.306
EVS	0.310
MAE	192.46 (deaths per million)
MAPE	48.09%
MSE	61178.09

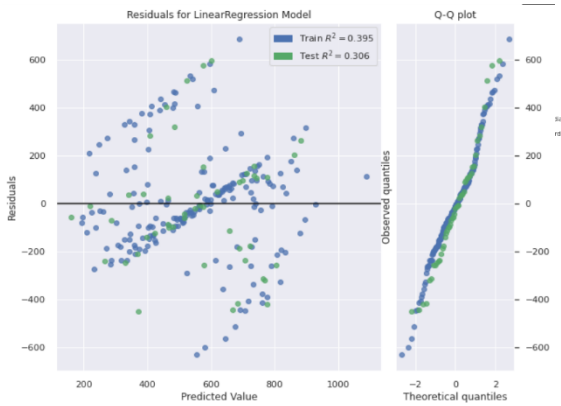
Πίνακας 67 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, Linear Regression, Βερολίνο, Ιδία Επεξεργασία*



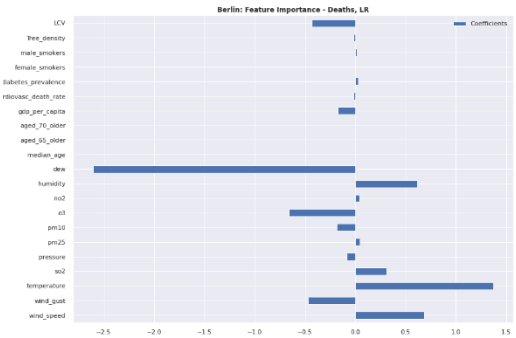
Σχήμα 265 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, Linear Regression, Βερολίνο, Ιδία Επεξεργασία



Σχήμα 266 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, Linear Regression, Βερολίνο, Ιδία Επεξεργασία



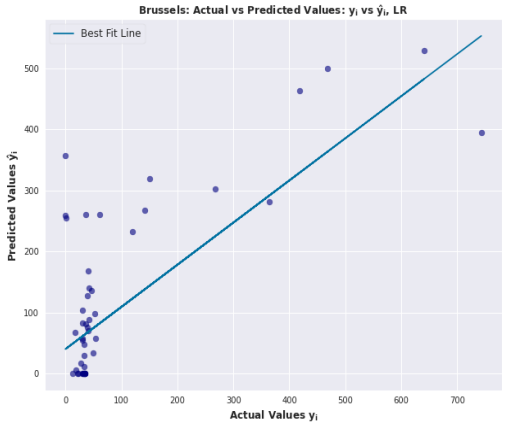
Σχήμα 267 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, Linear Regression, Βερολίνο, Ιδία Επεξεργασία



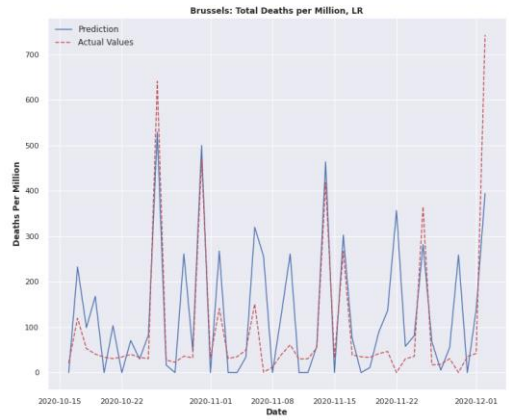
Σχήμα 268 Feature importance για πρόβλεψη θανάτων, Linear Regression, Βερολίνο, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	122.36 (deaths per million)
R^2	0.490
EVS	0.538
MAE	90.26 (deaths per million)
MAPE	107.00%
MSE	13071.26

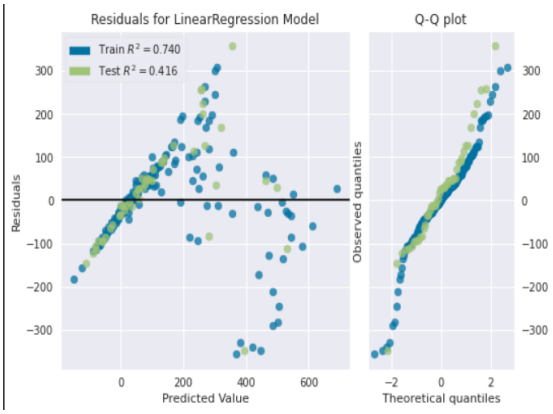
Πίνακας 68 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, Linear Regression, Βρυξέλλες, Ιδία Επεξεργασία*



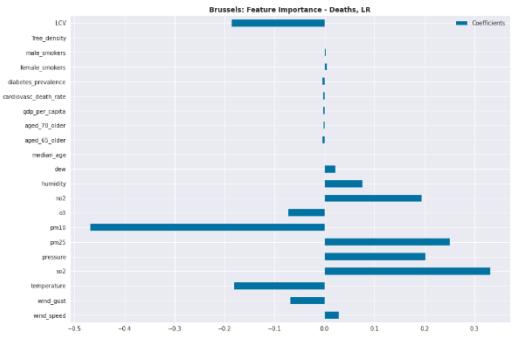
Σχήμα 269 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, Linear Regression, Βρυξέλλες, Ιδία Επεξεργασία



Σχήμα 270 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, Linear Regression, Βρυξέλλες, Ιδία Επεξεργασία



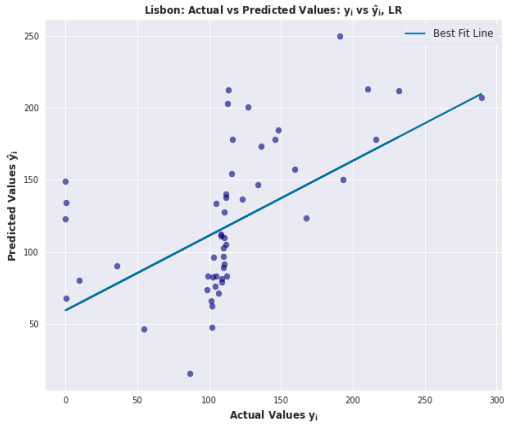
Σχήμα 271 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, Linear Regression, Βρυξέλλες, Ιδία Επεξεργασία



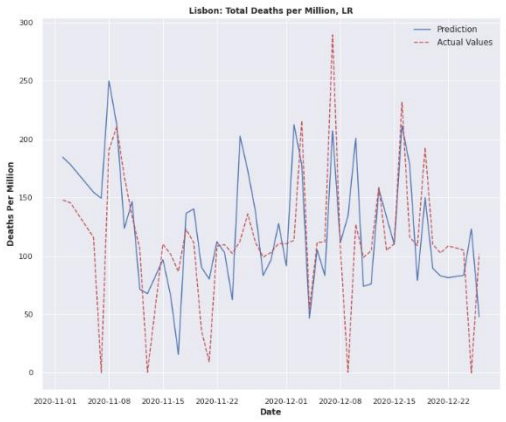
Σχήμα 272 Feature importance για πρόβλεψη θανάτων, Linear Regression, Βρυξέλλες, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Λισαβόνα
RMSE	51.89 (deaths per million)
R^2	0.139
EVS	0.172
MAE	39.33 (deaths per million)
MAPE	235.98%
MSE	2692.53

Πίνακας 69 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, Linear Regression, Λισαβόνα, Ιδία Επεξεργασία*

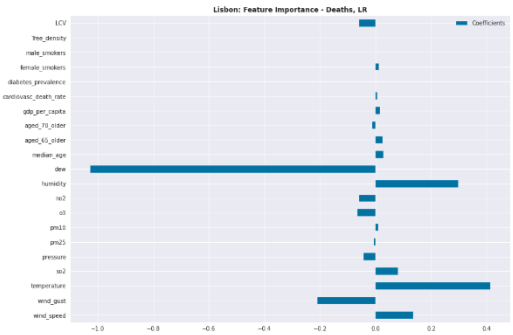


Σχήμα 273 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, Linear Regression, Λισαβόνα, Ιδία Επεξεργασία



Σχήμα 274 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, Linear Regression, Λισαβόνα, Ιδία Επεξεργασία

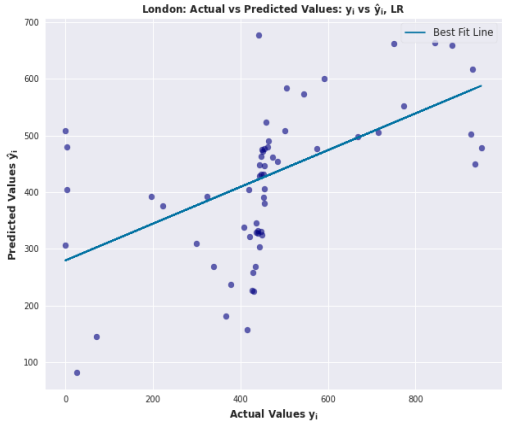
Σχήμα 275 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, Linear Regression, Λισαβόνα, Ιδία Επεξεργασία



Σχήμα 276 Feature importance για πρόβλεψη θανάτων, Linear Regression, Λισαβόνα, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Λονδίνο
RMSE	191.24 (deaths per million)
R^2	0.248
EVS	0.289
MAE	137.01 (deaths per million)
MAPE	254.11%
MSE	3653.77

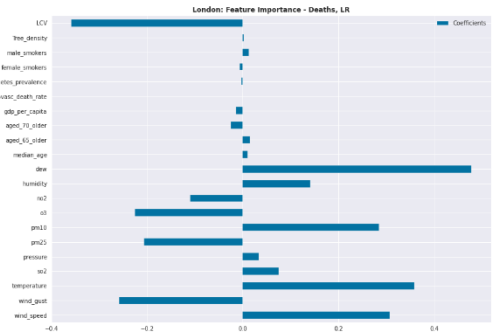
Πίνακας 70 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, Linear Regression, Λονδίνο, Ιδία Επεξεργασία*



Σχήμα 277 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, Linear Regression, Λονδίνο, Ιδία Επεξεργασία

Σχήμα 278 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, Linear Regression, Λονδίνο, Ιδία Επεξεργασία

Σχήμα 279 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, Linear Regression, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 280 Feature importance για πρόβλεψη θανάτων, Linear Regression, Λονδίνο, Ιδία Επεξεργασία

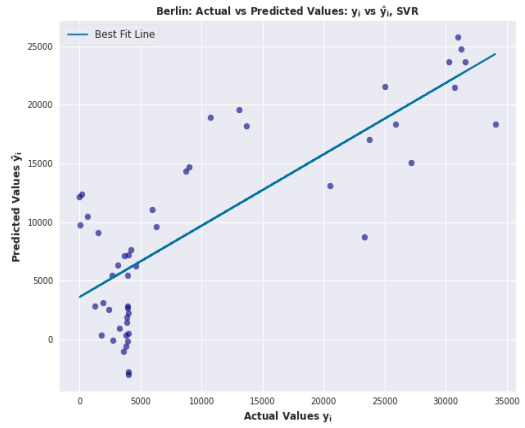
Παράρτημα Β

Στο συγκεκριμένο παράρτημα παρουσιάζονται τα αποτελέσματα για τις τέσσερις πόλεις και για το SVR μοντέλο.

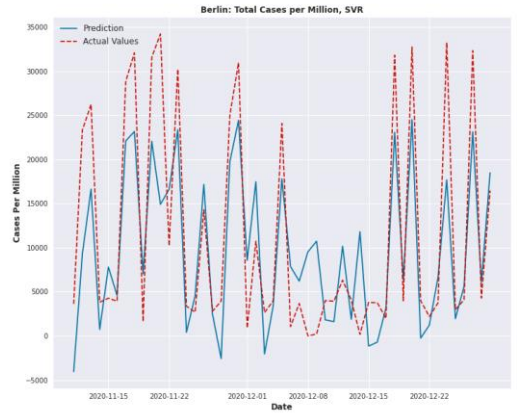
Πρόβλεψη Κρουσμάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	6517.84 (cases per million)
R^2	0.630
EVS	0.636
MAE	5394.76 (cases per million)
MAPE	28.39%
MSE	42482300.16

Πίνακας 71 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, SVR, Βερολίνο, Ιδία Επεξεργασία*

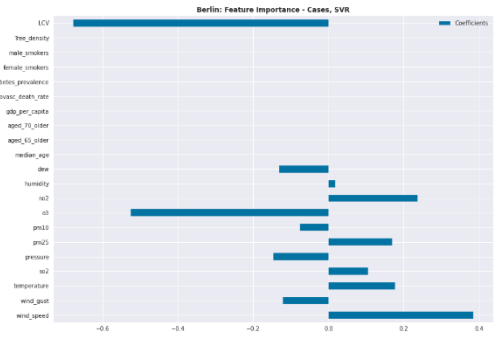


Σχήμα 281 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, SVR, Βερολίνο



Σχήμα 282 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, SVR, Βερολίνο

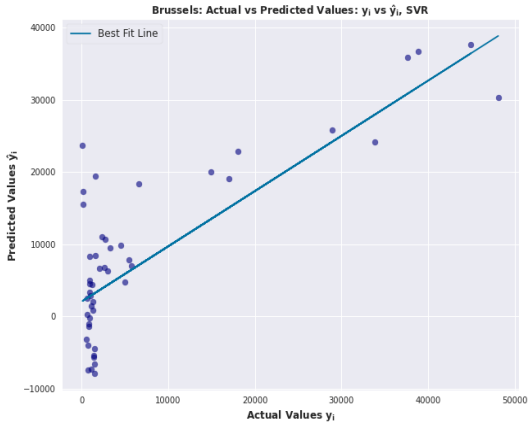
Σχήμα 283 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, SVR, Βερολίνο



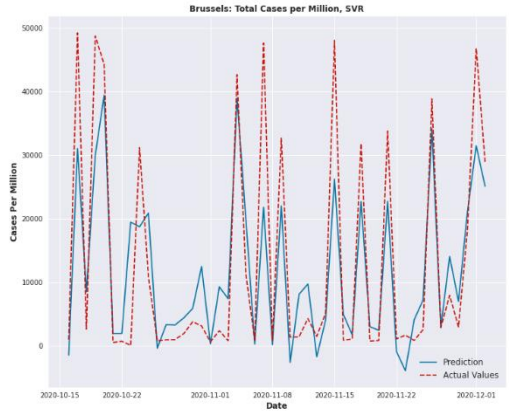
Σχήμα 284 Feature importance για πρόβλεψη κρουσμάτων, SVR, Βερολίνο

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	8027.35 (cases per million)
R^2	0.664
EVS	0.704
MAE	6075.32 (cases per million)
MAPE	12.14%
MSE	54016503.52

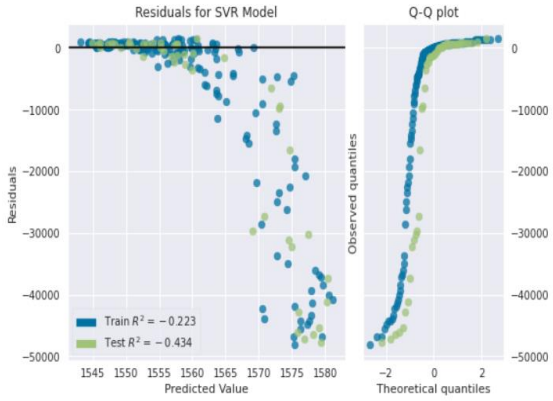
Πίνακας 72 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, SVR, Βρυξέλλες, Ιδία Επεξεργασία*



Σχήμα 285 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, SVR, Βρυξέλλες



Σχήμα 286 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, SVR, Βρυξέλλες

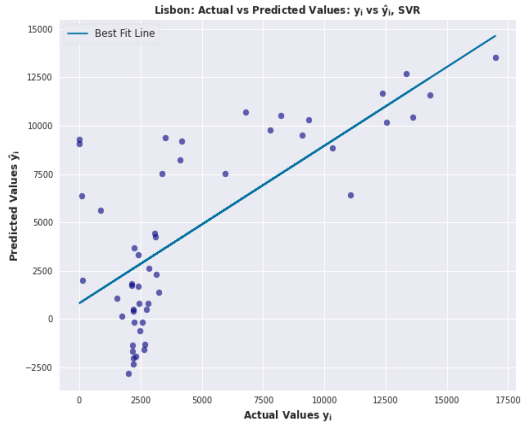


Σχήμα 287 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, SVR, Βρυξέλλες

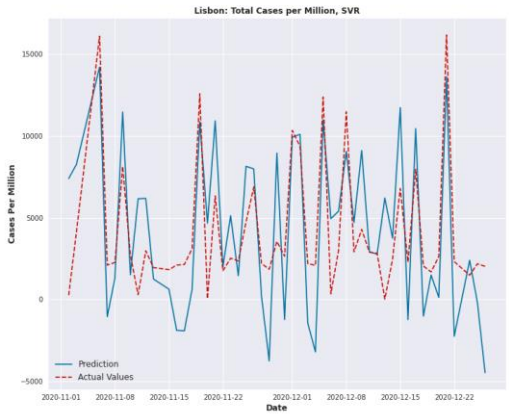
Σχήμα 288 Feature importance για πρόβλεψη κρουσμάτων, SVR, Βρυξέλλες

Μετρική Αξιολόγησης	Λισαβόνα
RMSE	3004.91 (cases per million)
R^2	0.513
EVS	0.527
MAE	2248.08 (cases per million)
MAPE	21.14%
MSE	8701702.47

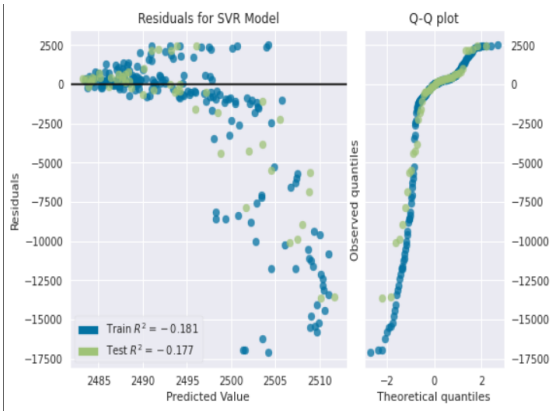
Πίνακας 73 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, SVR, Λισαβόνα, Ιδία Επεξεργασία*



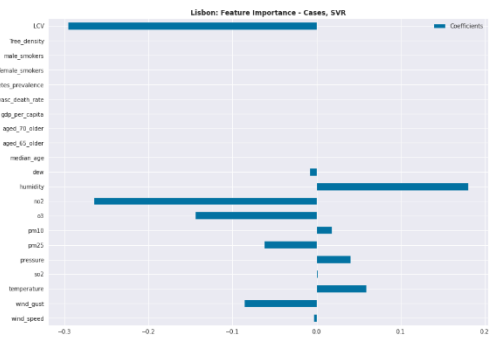
Σχήμα 289 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, SVR, Λισαβόνα



Σχήμα 290 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, SVR, Λισαβόνα



Σχήμα 291 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, SVR, Λισαβόνα

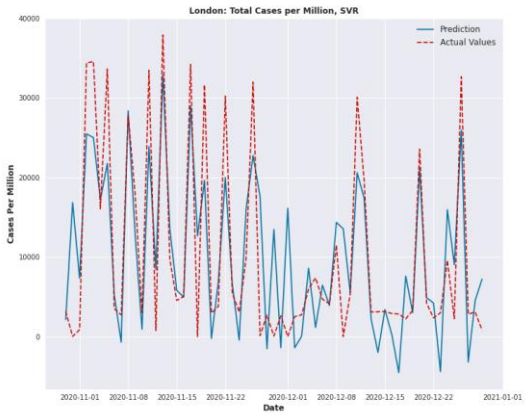


Σχήμα 292 Feature importance για πρόβλεψη κρουσμάτων, SVR, Λισαβόνα

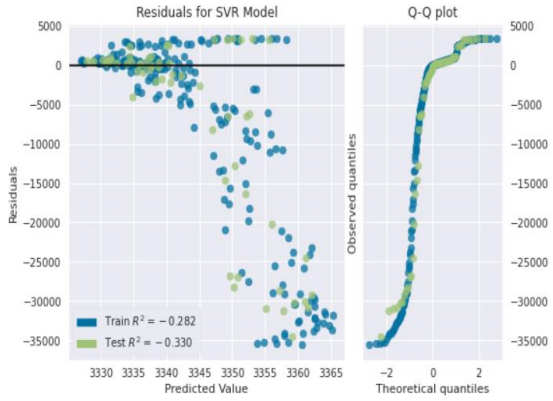
Μετρική Αξιολόγησης	Λονδίνο
RMSE	6913.75 (cases per million)
R^2	0.737
EVS	0.742
MAE	5024.07 (cases per million)
MAPE	776.17%
MSE	41071126.66

Πίνακας 74 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, SVR, Λονδίνο, Ιδία Επεξεργασία*

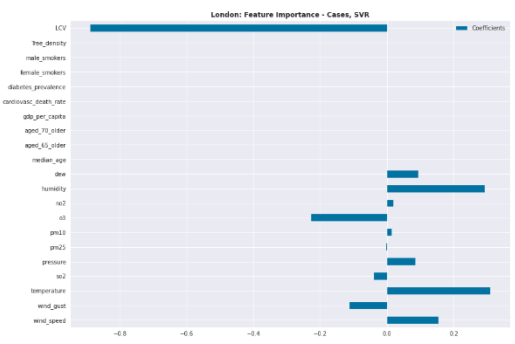
Σχήμα 293 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, SVR, Λονδίνο



Σχήμα 294 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, SVR, Λονδίνο



Σχήμα 295 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, SVR, Λονδίνο

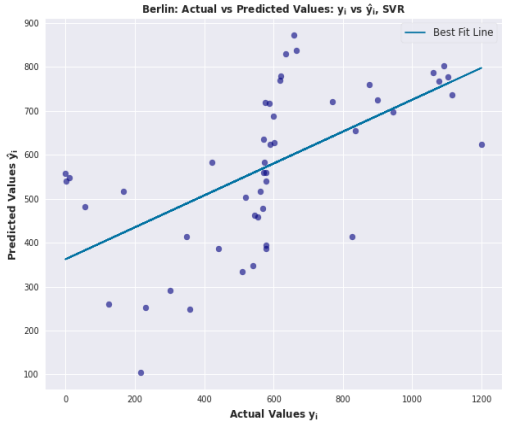


Σχήμα 296 Feature importance για πρόβλεψη κρουσμάτων, SVR, Λονδίνο

Πρόβλεψη Θανάτων

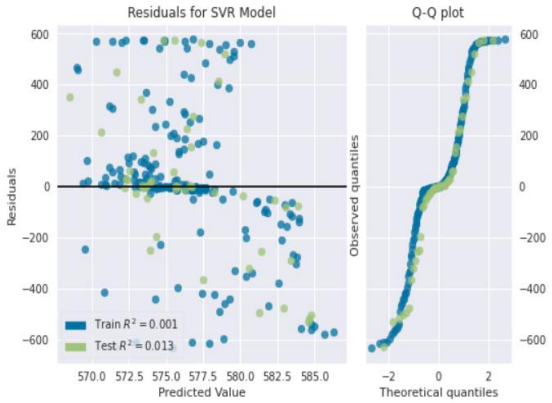
Μετρική Αξιολόγησης	Βερολίνο
RMSE	238.63 (deaths per million)
R^2	0.354
EVS	0.356
MAE	182.36 (deaths per million)
MAPE	46.23%
MSE	56945.81

Πίνακας 75 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, SVR, Βερολίνο, Ιδία Επεξεργασία*

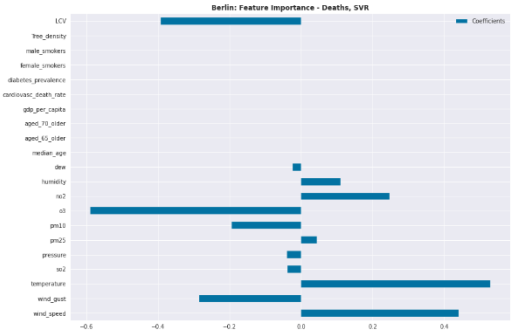


Σχήμα 297 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, SVR,, Ιδία Επεξεργασία

Σχήμα 298 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, SVR, Βερολίνο, Ιδία Επεξεργασία



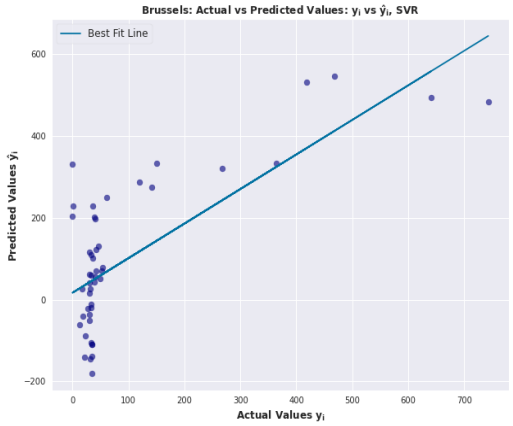
Σχήμα 299 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, SVR, Βερολίνο, Ιδία Επεξεργασία



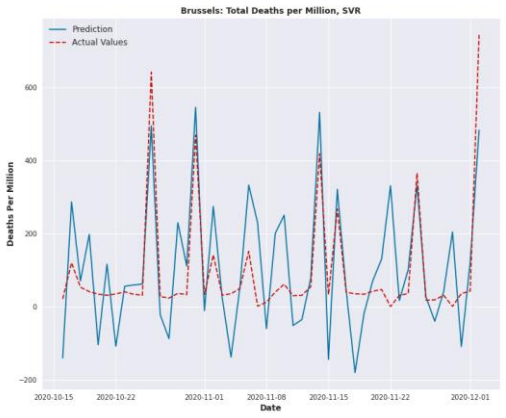
Σχήμα 300 Feature importance για πρόβλεψη θανάτων, SVR, Βερολίνο, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	127.90 (deaths per million)
R^2	0.362
EVS	0.368
MAE	102.02 (deaths per million)
MAPE	96.87%
MSE	16358.58

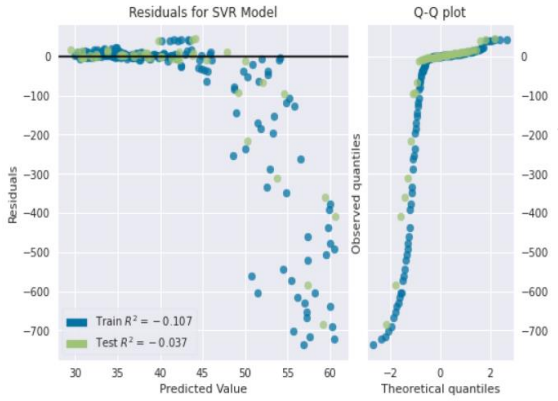
Πίνακας 76 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, SVR, Βρυξέλλες, Ιδία Επεξεργασία*



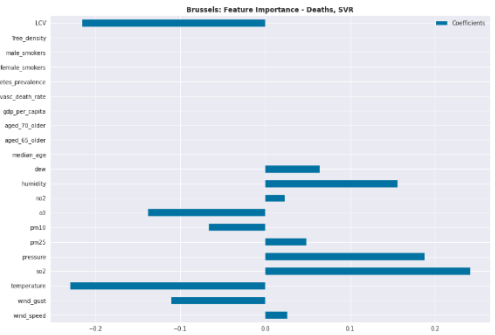
Σχήμα 301 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, SVR, Βρυξέλλες, Ιδία Επεξεργασία



Σχήμα 302 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, SVR, Βρυξέλλες, Ιδία Επεξεργασία



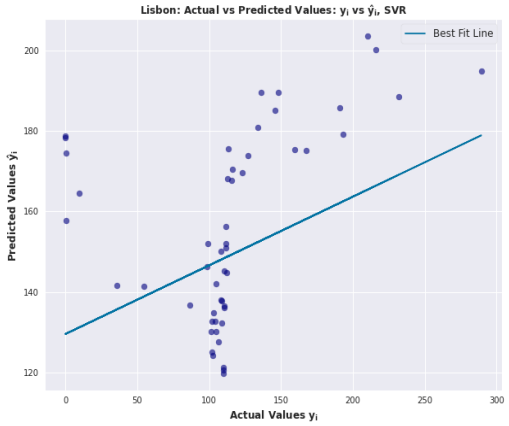
Σχήμα 303 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, SVR, Βρυξέλλες, Ιδία Επεξεργασία



Σχήμα 304 Feature importance για πρόβλεψη θανάτων, SVR, Βρυξέλλες, Ιδία Επεξεργασία

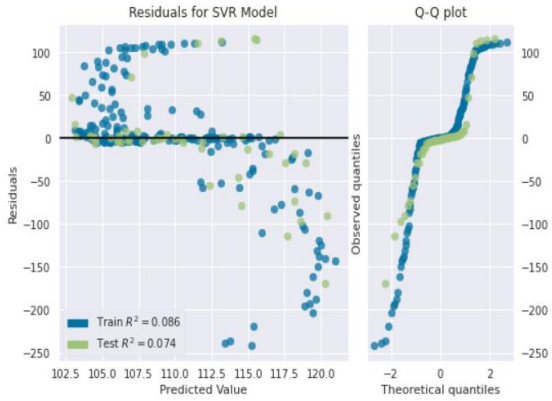
Μετρική Αξιολόγησης	Λισαβόνα
RMSE	66.63 (deaths per million)
R^2	-0.420
EVS	0.164
MAE	49.74 (deaths per million)
MAPE	313.41%
MSE	4439.67

Πίνακας 77 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, SVR, Λισαβόνα, Ιδία Επεξεργασία*

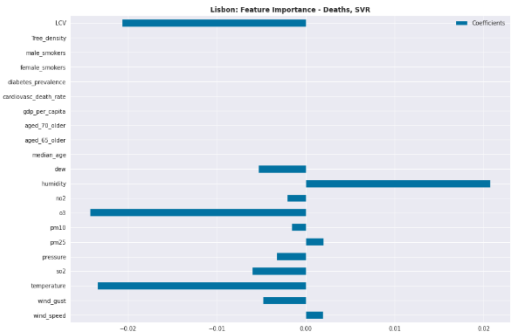


Σχήμα 305 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, SVR, Λισαβόνα, Ιδία Επεξεργασία

Σχήμα 306 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, SVR, Λισαβόνα, Ιδία Επεξεργασία



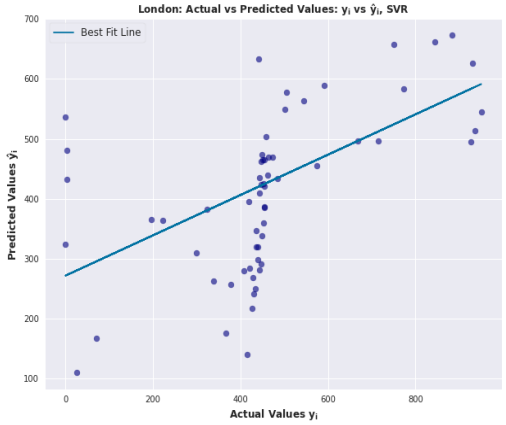
Σχήμα 307 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, SVR, Λισαβόνα, Ιδία Επεξεργασία



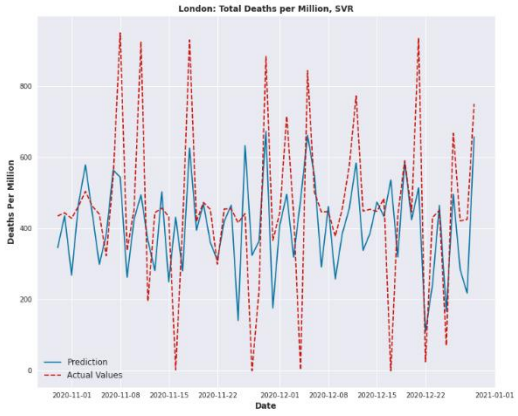
Σχήμα 308 Feature importance για πρόβλεψη θανάτων, SVR, Λισαβόνα, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Λονδίνο
RMSE	190.23 (deaths per million)
R^2	0.256
EVS	0.305
MAE	140.19 (deaths per million)
MAPE	268.86%
MSE	36187.62

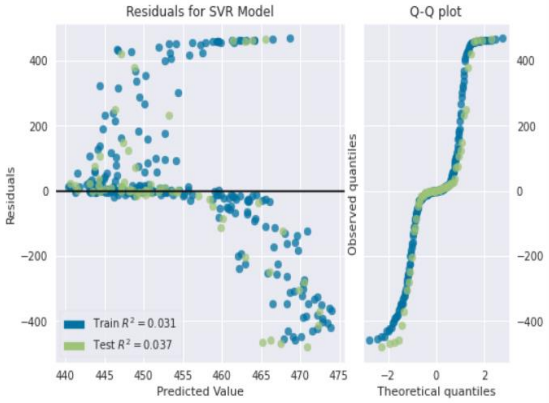
Πίνακας 78 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, SVR, Λονδίνο, Ιδία Επεξεργασία*



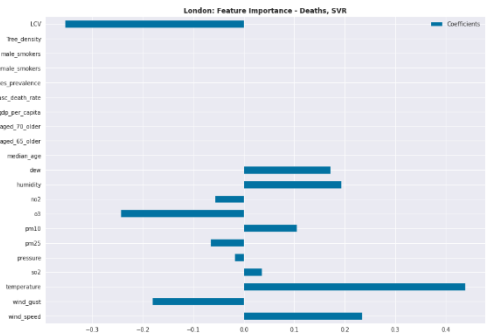
Σχήμα 309 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, SVR, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 310 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, SVR, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 311 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, SVR, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 312 Feature importance για πρόβλεψη θανάτων, SVR, Λονδίνο, Ιδία Επεξεργασία

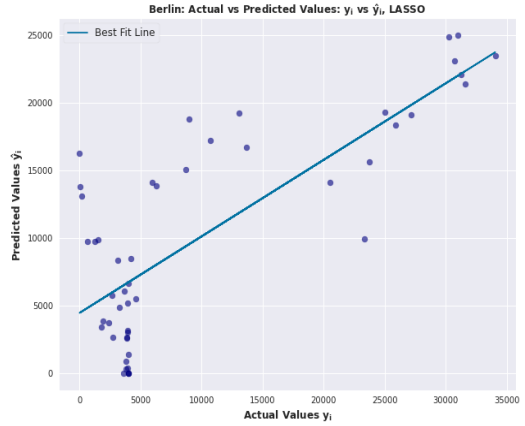
Παράρτημα Γ

Στο συγκεκριμένο παράρτημα παρουσιάζονται τα αποτελέσματα για το μοντέλο LASSO και για τις τέσσερις πόλεις.

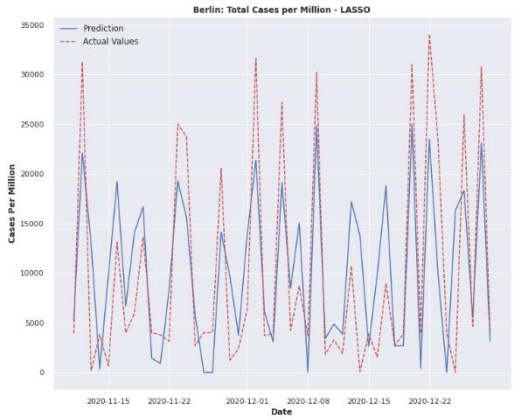
Πρόβλεψη Κρουσμάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	6893.01 (cases per million)
R^2	0.601
EVS	0.602
MAE	5737.80 (cases per million)
MAPE	37.65%
MSE	45800573.66

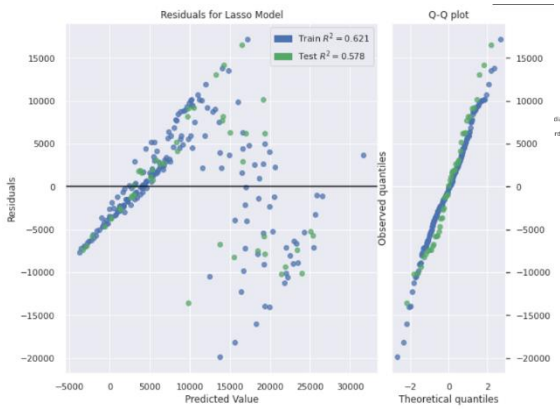
Πίνακας 79 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LASSO, Βερολίνο, Ιδία Επεξεργασία*



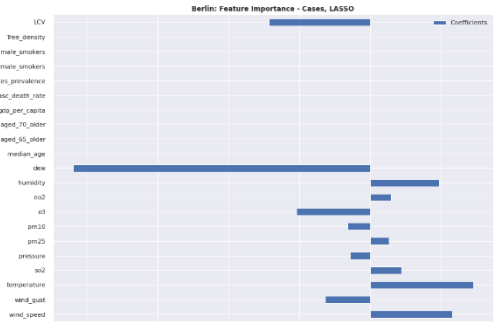
Σχήμα 313 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LASSO, Βερολίνο



Σχήμα 314 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LASSO, Βερολίνο



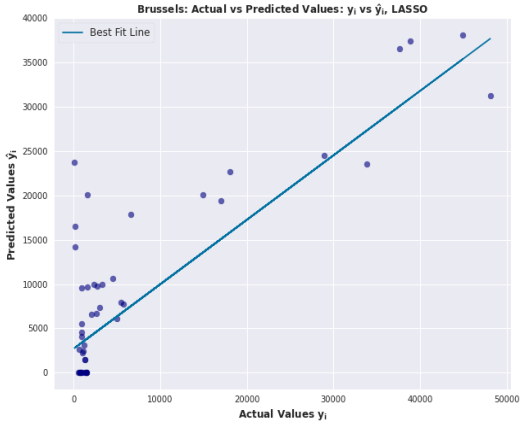
Σχήμα 315 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Βερολίνο



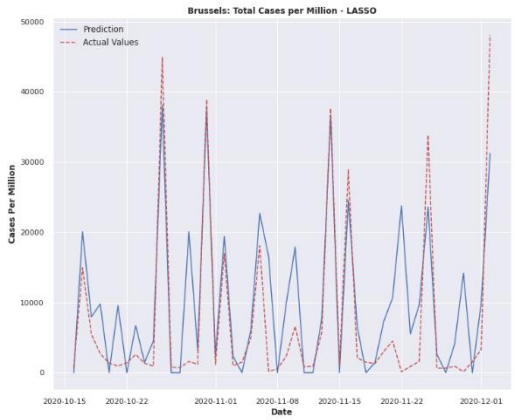
Σχήμα 316 Feature importance για πρόβλεψη κρουσμάτων, LASSO, Βερολίνο

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	7692.73 (cases per million)
R^2	0.679
EVS	0.720
MAE	5818.54 (cases per million)
MAPE	11.51%
MSE	51589418.49

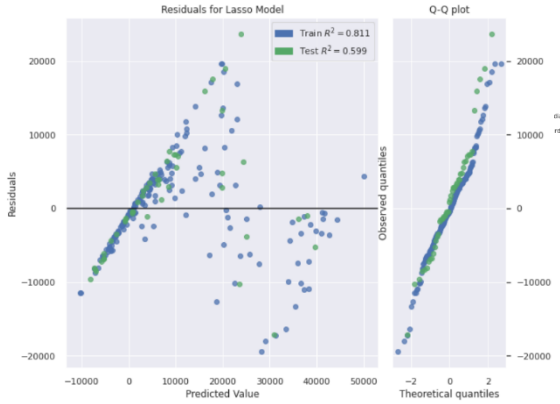
Πίνακας 80 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LASSO, Βρυξέλλες, Ιδία Επεξεργασία*



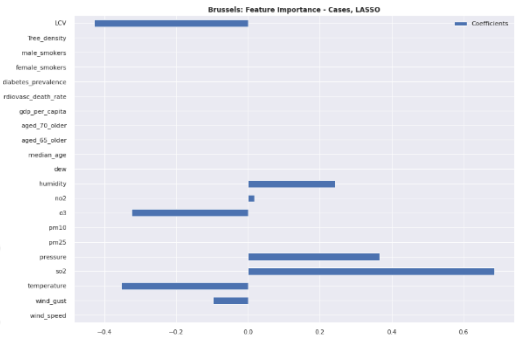
Σχήμα 317 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LASSO, Βρυξέλλες



Σχήμα 318 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LASSO, Βρυξέλλες



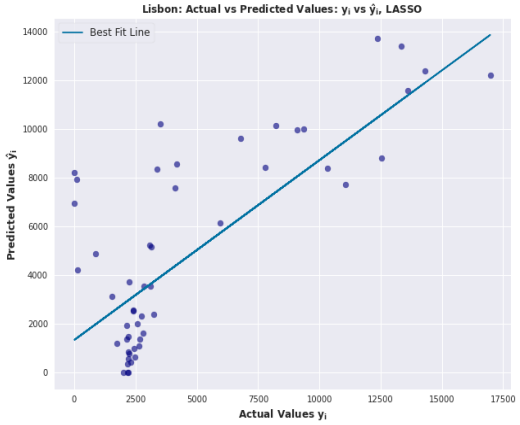
Σχήμα 319 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Βρυξέλλες



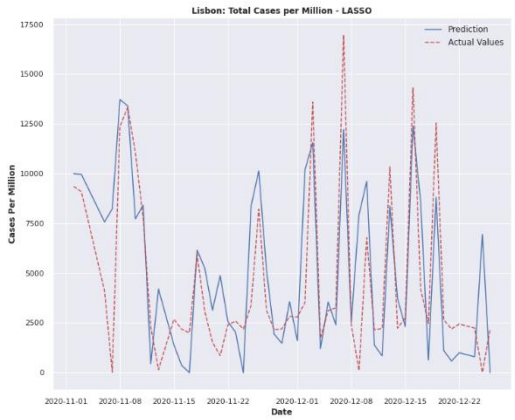
Σχήμα 320 Feature importance για πρόβλεψη κρουσμάτων, LASSO, Βρυξέλλες

Μετρική Αξιολόγησης	Λισαβόνα
RMSE	2984.73 (cases per million)
R^2	0.518
EVS	0.530
MAE	2238.81 (cases per million)
MAPE	21.46%
MSE	8609023.94

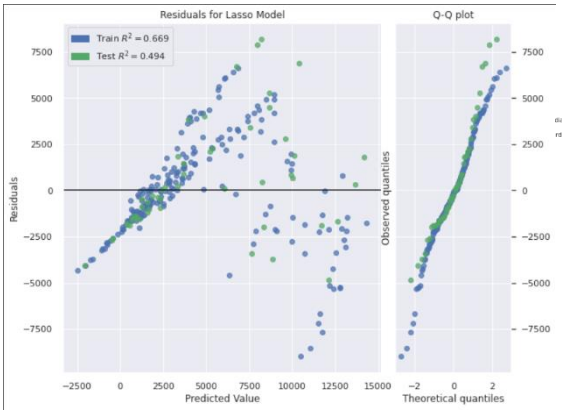
Πίνακας 81 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LASSO, Λισαβόνα, Ιδία Επεξεργασία*



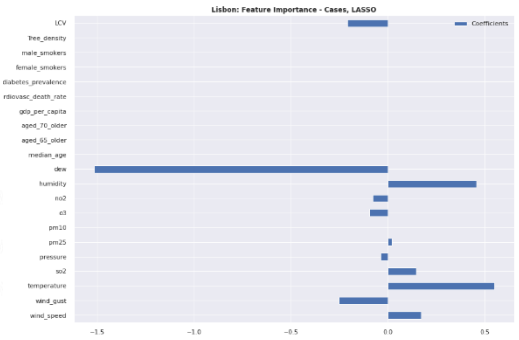
Σχήμα 321 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LASSO, Λισαβόνα



Σχήμα 322 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LASSO, Λισαβόνα



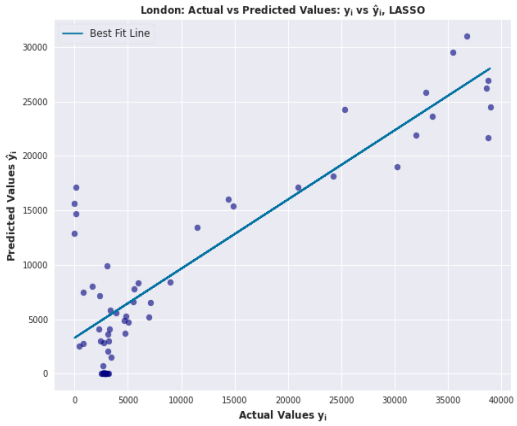
Σχήμα 323 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Λισαβόνα



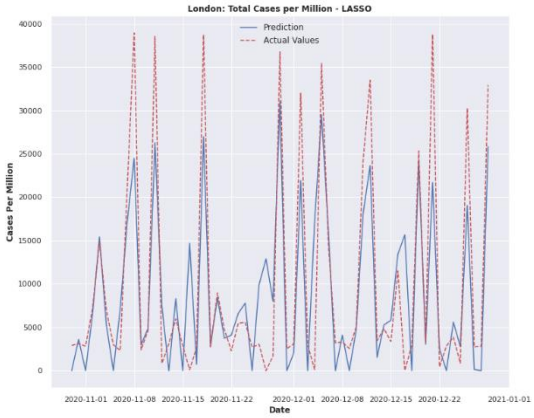
Σχήμα 324 Feature importance για πρόβλεψη κρουσμάτων, LASSO, Λισαβόνα

Μετρική Αξιολόγησης	Λονδίνο
RMSE	6810.85 (cases per million)
R^2	0.735
EVS	0.741
MAE	5028.78 (cases per million)
MAPE	840.62%
MSE	41302588.12

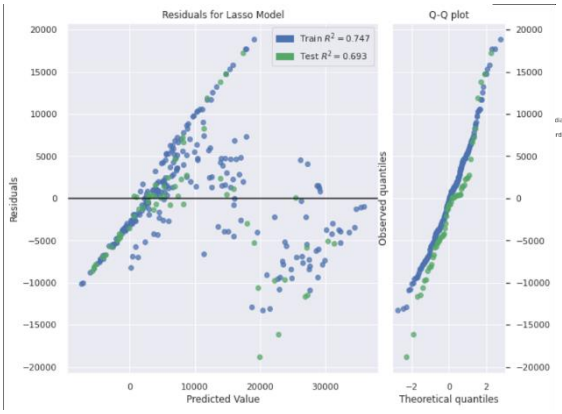
Πίνακας 82 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, LASSO, Λονδίνο, Ιδία Επεξεργασία*



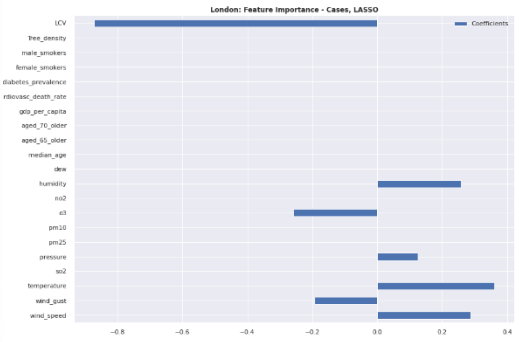
Σχήμα 325 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, LASSO, Λονδίνο



Σχήμα 326 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, LASSO, Λονδίνο



Σχήμα 327 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, LASSO, Λονδίνο

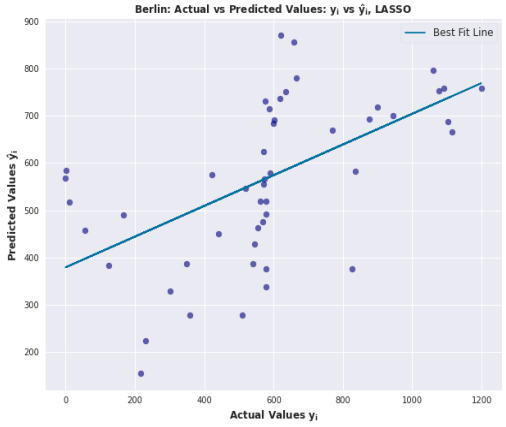


Σχήμα 328 Feature importance για πρόβλεψη κρουσμάτων, LASSO, Λονδίνο

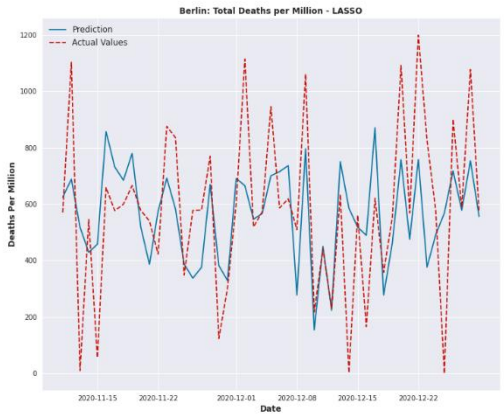
Πρόβλεψη Θανάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	245.92 (deaths per million)
R^2	0.314
EVS	0.318
MAE	190.52 (deaths per million)
MAPE	47.39%
MSE	60474.23

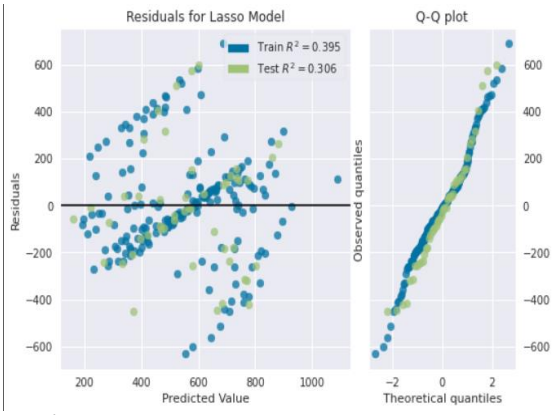
Πίνακας 83 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LASSO, Βερολίνο, Ιδία Επεξεργασία*



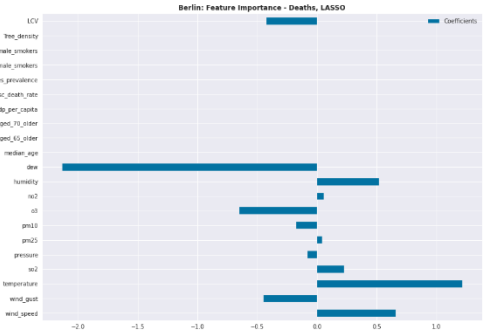
Σχήμα 329 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LASSO, Βερολίνο, Ιδία Επεξεργασία



Σχήμα 330 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LASSO, Βερολίνο, Ιδία Επεξεργασία



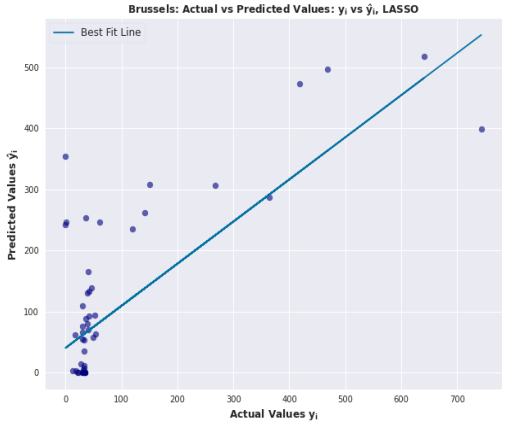
Σχήμα 331 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LASSO, Βερολίνο, Ιδία Επεξεργασία



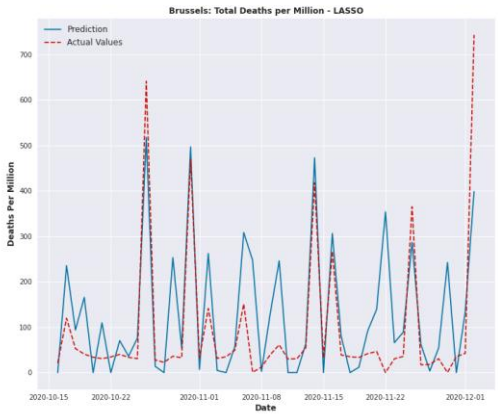
Σχήμα 332 Feature importance για πρόβλεψη θανάτων, LASSO, Βερολίνο, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	119.00 (deaths per million)
R^2	0.513
EVS	0.561
MAE	87.67 (deaths per million)
MAPE	104.84%
MSE	12485.20

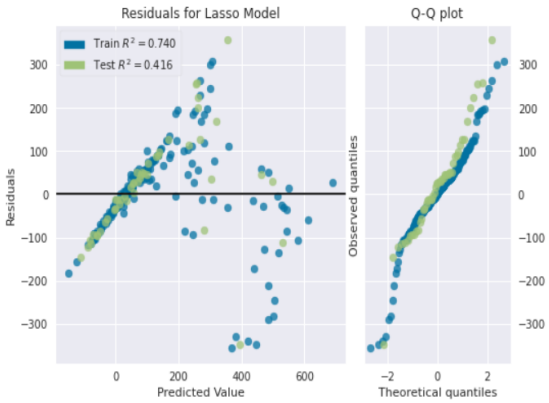
Πίνακας 84 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LASSO, Βρυξέλλες, Ιδία Επεξεργασία*



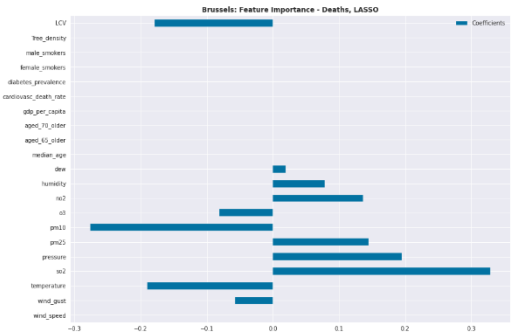
Σχήμα 333 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LASSO, Βρυξέλλες, Ιδία Επεξεργασία



Σχήμα 334 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LASSO, Βρυξέλλες, Ιδία Επεξεργασία



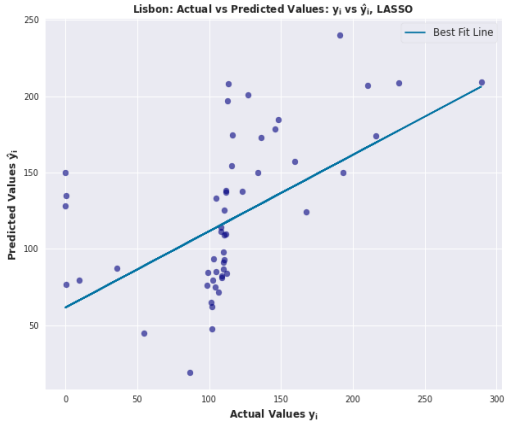
Σχήμα 335 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LASSO, Βρυξέλλες, Ιδία Επεξεργασία



Σχήμα 336 Feature importance για πρόβλεψη θανάτων, LASSO, Βρυξέλλες, Ιδία Επεξεργασία

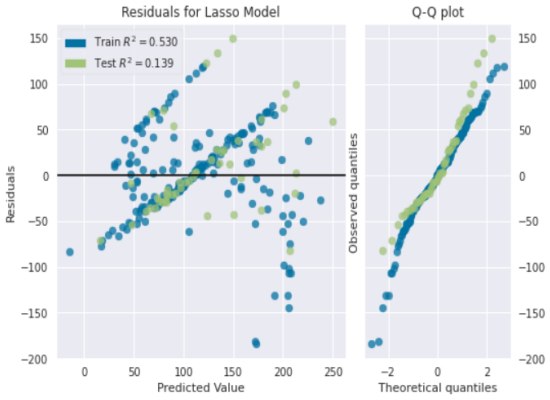
Μετρική Αξιολόγησης	Λισαβόνα
RMSE	51.64 (deaths per million)
R^2	0.147
EVS	0.178
MAE	39.26 (deaths per million)
MAPE	241.13%
MSE	2667.14

Πίνακας 85 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LASSO, Λισαβόνα, Ιδία Επεξεργασία*

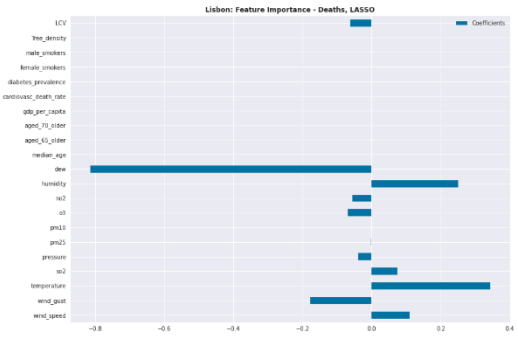


Σχήμα 337 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LASSO, Λισαβόνα, Ιδία Επεξεργασία

Σχήμα 338 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LASSO, Λισαβόνα, Ιδία Επεξεργασία



Σχήμα 339 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LASSO, Λισαβόνα, Ιδία Επεξεργασία

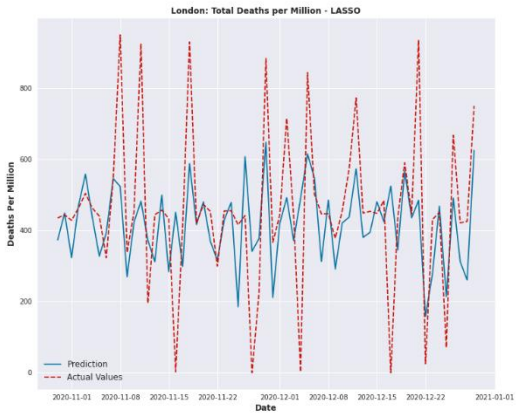


Σχήμα 340 Feature importance για πρόβλεψη θανάτων, LASSO, Λισαβόνα, Ιδία Επεξεργασία

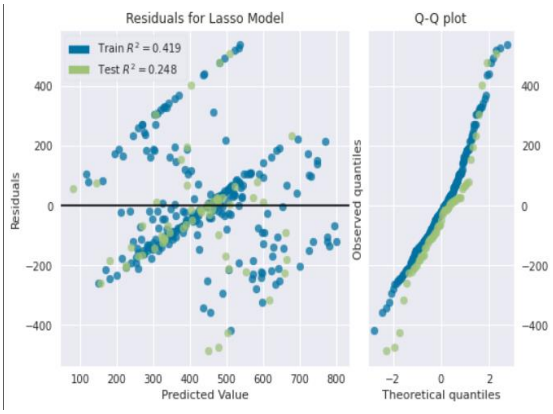
Μετρική Αξιολόγησης	Λονδίνο
RMSE	190.60 (deaths per million)
R^2	0.283
EVS	0.285
MAE	136.14 (deaths per million)
MAPE	276.51%
MSE	36329.25

Πίνακας 86 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, LASSO, Λονδίνο, Ιδία Επεξεργασία*

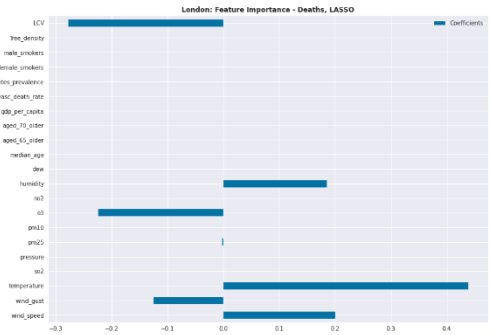
Σχήμα 341 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, LASSO, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 342 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, LASSO, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 343 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, LASSO, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 344 Feature importance για πρόβλεψη θανάτων, LASSO, Λονδίνο, Ιδία Επεξεργασία

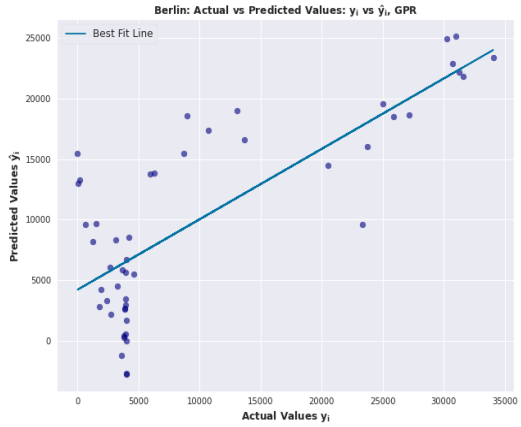
Παράρτημα Δ

Στο συγκεκριμένο παράρτημα παρουσιάζονται τα αποτελέσματα για τον αλγόριθμο GPR.

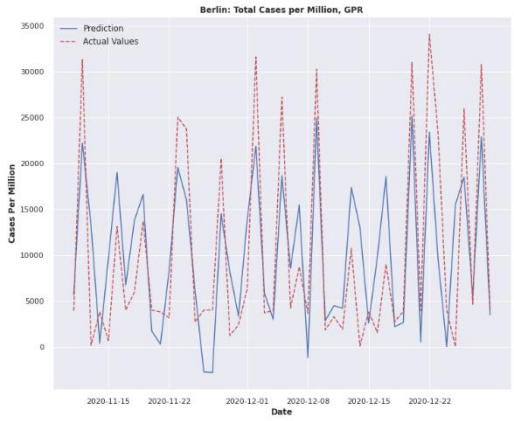
Πρόβλεψη Κρουσμάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	6745.41 (cases per million)
R^2	0.604
EVS	0.604
MAE	5618.86 (cases per million)
MAPE	35.86%
MSE	45500605.16

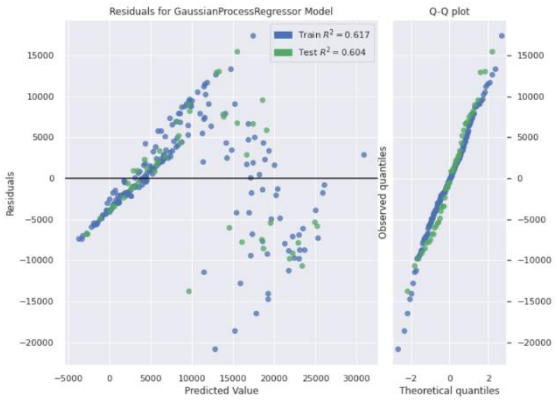
Πίνακας 87 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, GPR, Βερολίνο, Ιδία Επεξεργασία*



Σχήμα 345 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, GPR, Βερολίνο



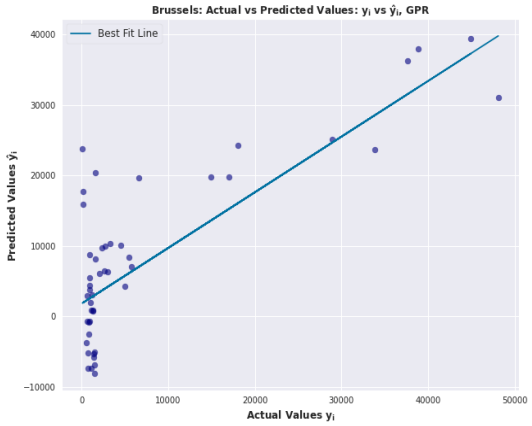
Σχήμα 346 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, GPR, Βερολίνο



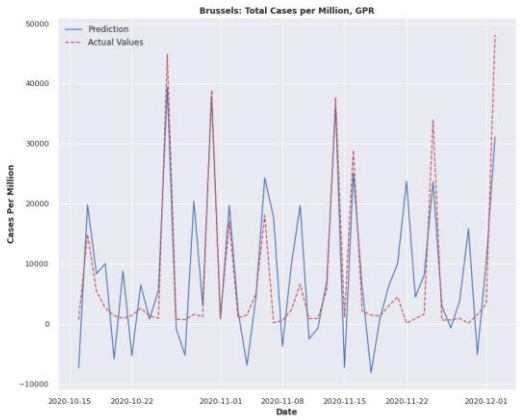
Σχήμα 347 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Βερολίνο

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	7980.87 (cases per million)
R^2	0.604
EVS	0.614
MAE	6034.90 (cases per million)
MAPE	12.06%
MSE	63694297.31

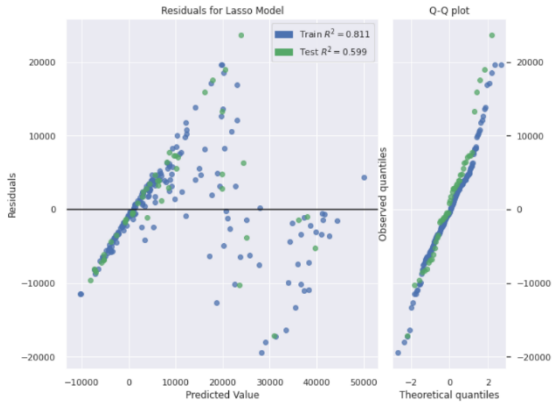
Πίνακας 88 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, GPR, Βρυξέλλες, Ιδία Επεξεργασία*



Σχήμα 348 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, GPR, Βρυξέλλες



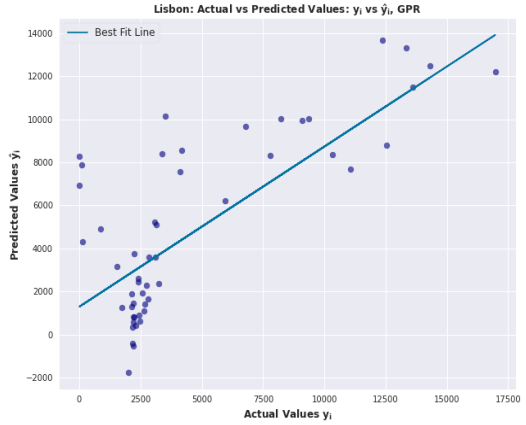
Σχήμα 349 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, GPR, Βρυξέλλες



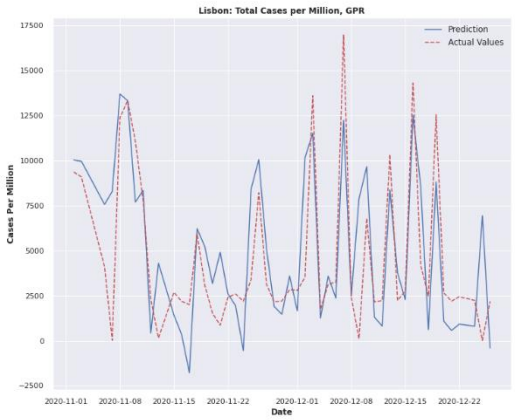
Σχήμα 350 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Βρυξέλλες

Μετρική Αξιολόγησης	Λισαβόνα
RMSE	2988.80 (cases per million)
R^2	0.500
EVS	0.509
MAE	2245.51 (cases per million)
MAPE	21.53%
MSE	8932942.62

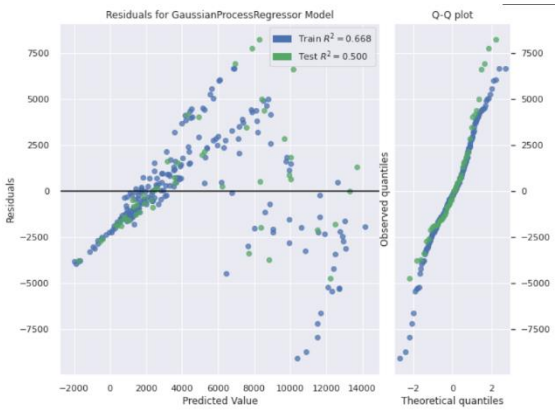
Πίνακας 89 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, GPR, Λισαβόνα, Ιδία Επεξεργασία*



Σχήμα 351 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, GPR, Λισαβόνα



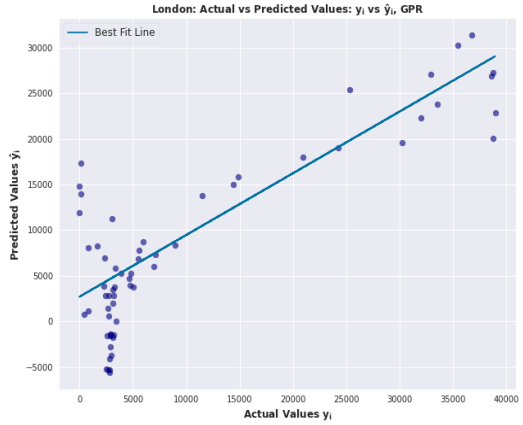
Σχήμα 352 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, GPR, Λισαβόνα



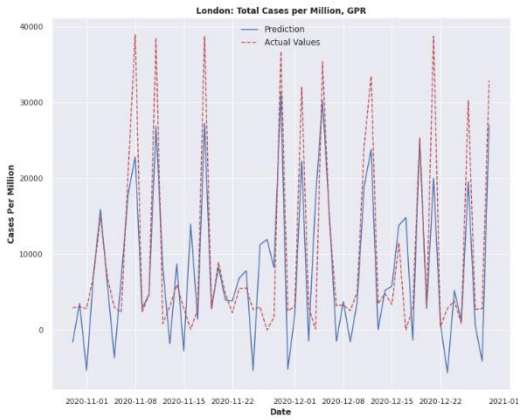
Σχήμα 353 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Λισαβόνα

Μετρική Αξιολόγησης	Λονδίνο
RMSE	6905.75 (cases per million)
R^2	0.694
EVS	0.711
MAE	5016.33 (cases per million)
MAPE	779.86%
MSE	47689433.96

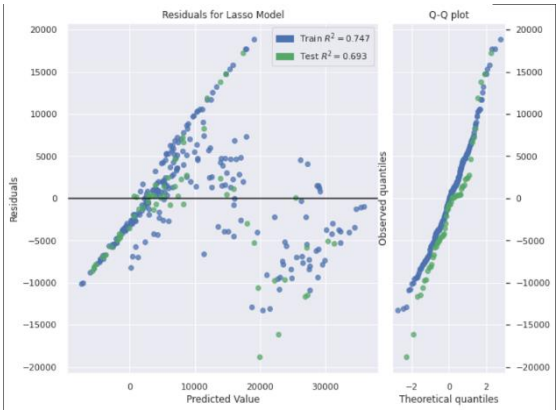
Πίνακας 90 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, GPR, Λονδίνο, Ιδία Επεξεργασία*



Σχήμα 354 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, GPR, Λονδίνο



Σχήμα 355 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, GPR, Λονδίνο

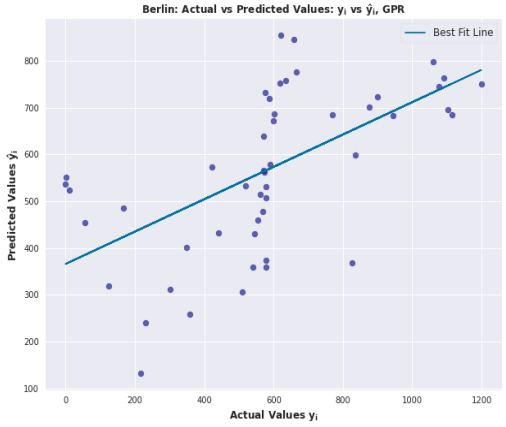


Σχήμα 356 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, GPR, Λονδίνο

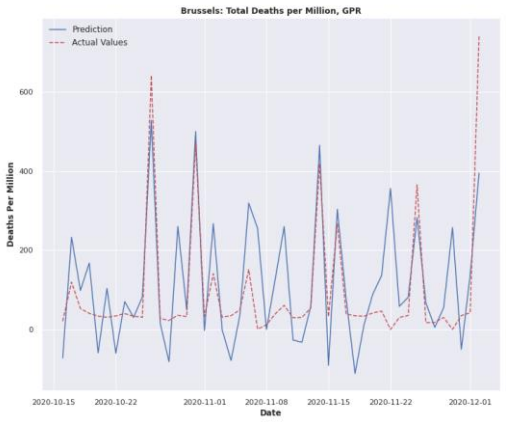
Πρόβλεψη Θανάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	240.35 (deaths per million)
R^2	0.345
EVS	0.350
MAE	186.38 (deaths per million)
MAPE	44.87%
MSE	57766.07

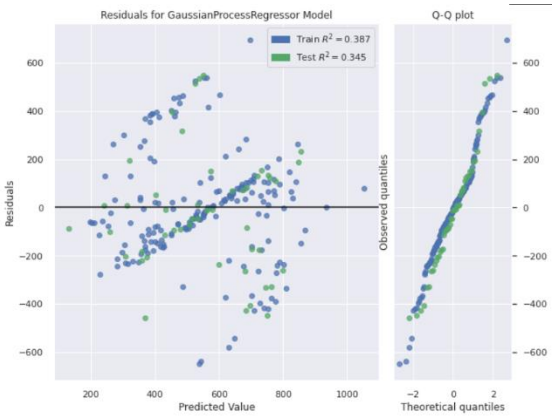
Πίνακας 91 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, GPR, Βερολίνο, Ιδία Επεξεργασία*



Σχήμα 357 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, GPR, Βερολίνο, Ιδία Επεξεργασία



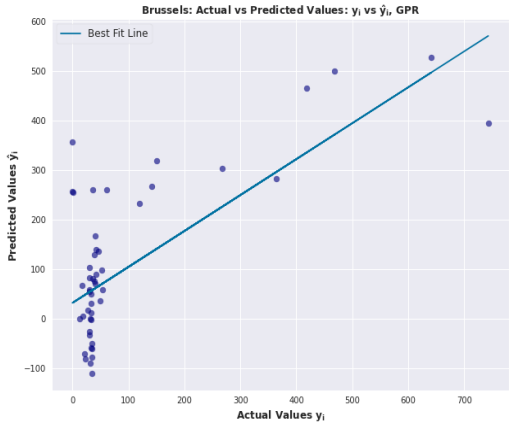
Σχήμα 358 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, GPR, Βερολίνο, Ιδία Επεξεργασία



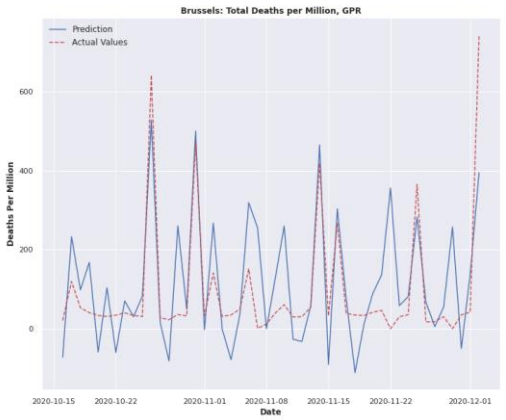
Σχήμα 359 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, GPR, Βερολίνο, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	122.06 (deaths per million)
R^2	0.419
EVS	0.437
MAE	90.03 (deaths per million)
MAPE	106.75%
MSE	14899.00

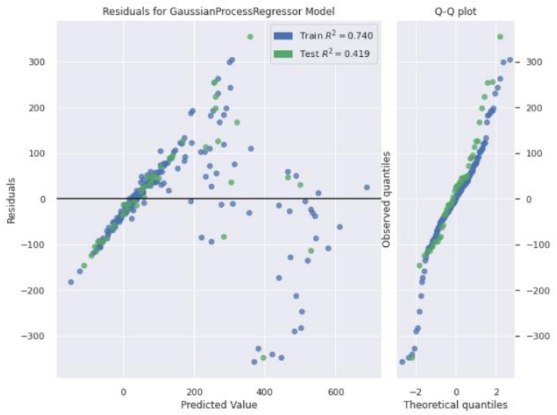
Πίνακας 92 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, GPR, Βρυξέλλες, Ιδία Επεξεργασία*



Σχήμα 360 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, GPR, Βρυξέλλες, Ιδία Επεξεργασία



Σχήμα 361 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, GPR, Βρυξέλλες, Ιδία Επεξεργασία



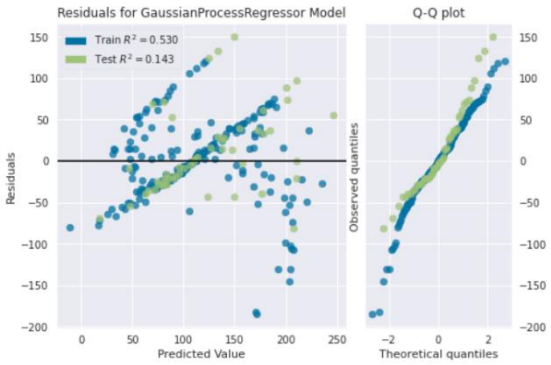
Σχήμα 362 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, GPR, Βρυξέλλες, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Λισαβόνα
RMSE	51.97 (deaths per million)
R^2	0.143
EVS	0.176
MAE	39.24 (deaths per million)
MAPE	237.59%
MSE	2678.16

Πίνακας 93 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, GPR, Λισαβόνα, Ιδία Επεξεργασία*

Σχήμα 363 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, GPR, Λισαβόνα, Ιδία Επεξεργασία

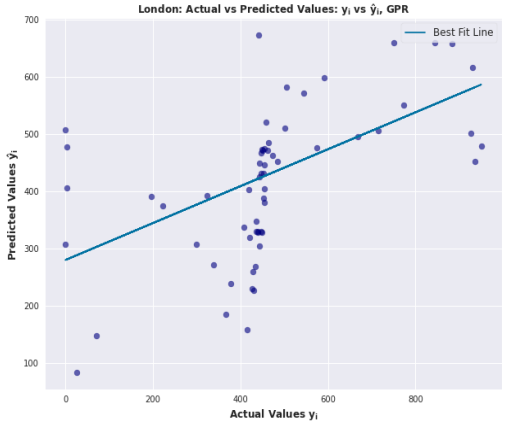
Σχήμα 364 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, GPR, Λισαβόνα, Ιδία Επεξεργασία



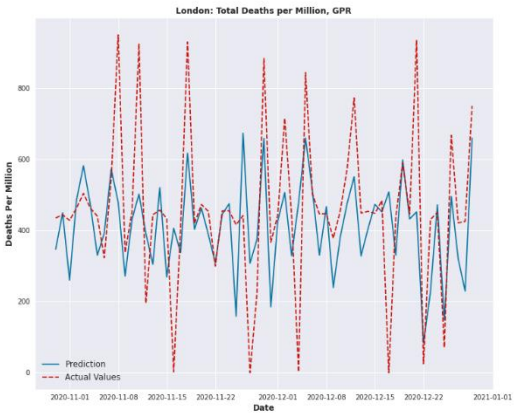
Σχήμα 365 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, GPR, Λισαβόνα, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Λονδίνο
RMSE	190.85 (deaths per million)
R^2	0.251
EVS	0.292
MAE	136.68 (deaths per million)
MAPE	255.07%
MSE	36425.55

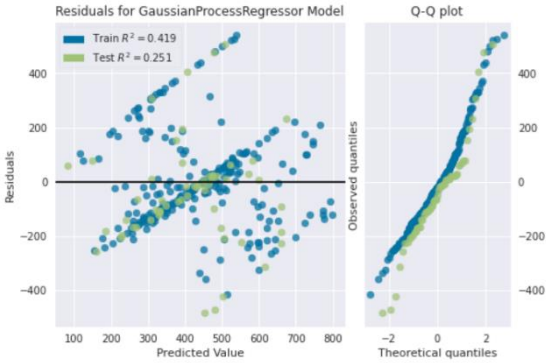
Πίνακας 94 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, GPR, Λονδίνο, Ιδία Επεξεργασία*



Σχήμα 366 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, GPR, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 367 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, GPR, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 368 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, GPR, Λονδίνο, Ιδία Επεξεργασία

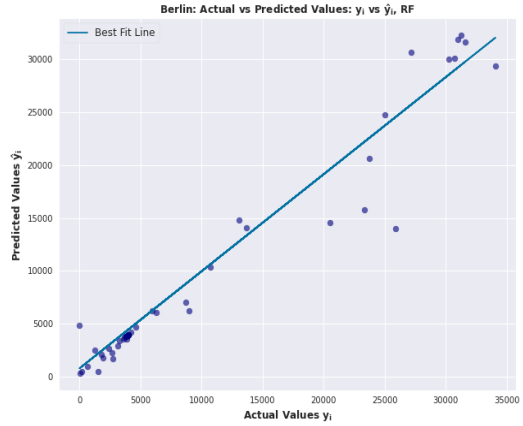
Παράρτημα Ε

Στο τρέχον παράρτημα παρουσιάζονται τα αποτελέσματα του μοντέλου Random Forest Regression.

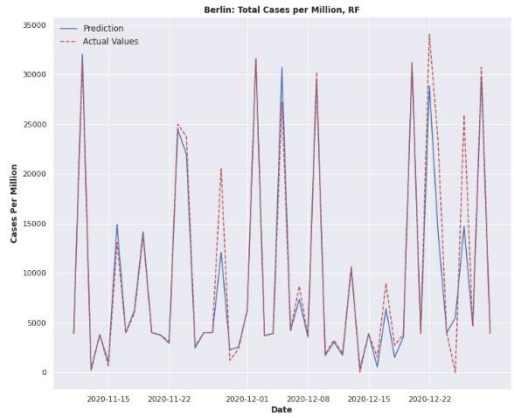
Πρόβλεψη Κρουσμάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	2271.09 (cases per million)
R^2	0.955
EVS	0.957
MAE	1095.12 (cases per million)
MAPE	7.04%
MSE	0.00170

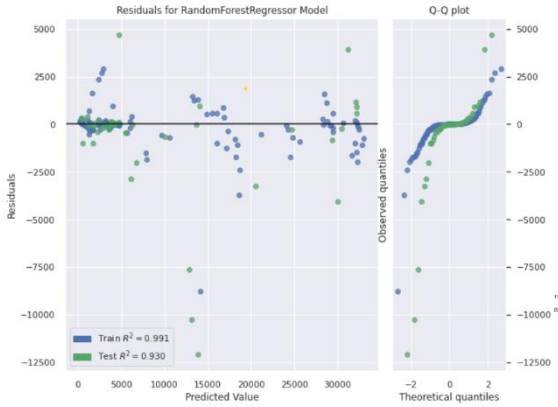
Πίνακας 95 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, RF, Βερολίνο, Ιδία Επεξεργασία*



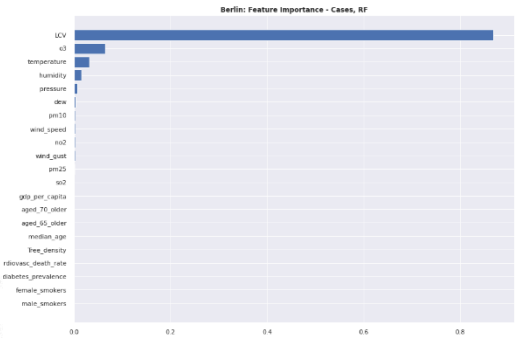
Σχήμα 369 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, RF, Βερολίνο



Σχήμα 370 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, RF, Βερολίνο



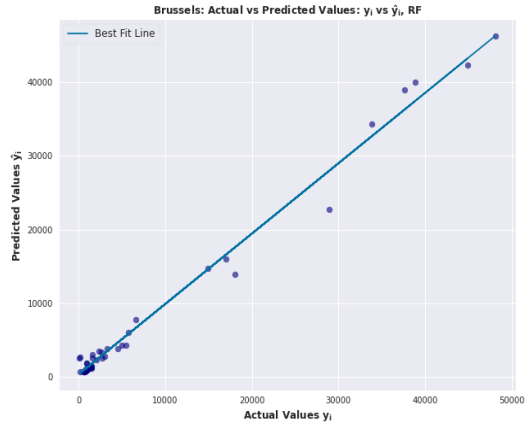
Σχήμα 371 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Βερολίνο



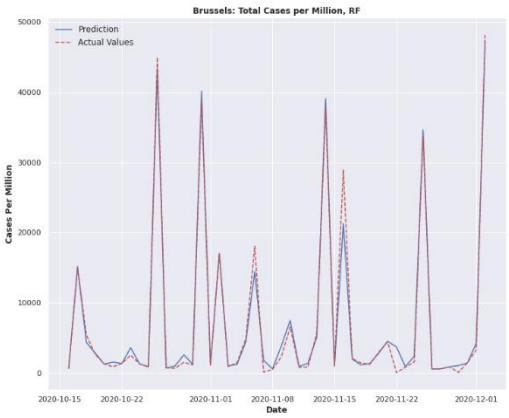
Σχήμα 372 Feature importance για πρόβλεψη κρουσμάτων, RF, Βερολίνο

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	1791.58 (cases per million)
R^2	0.980
EVS	0.980
MAE	829.49 (cases per million)
MAPE	1.59%
MSE	0.00106

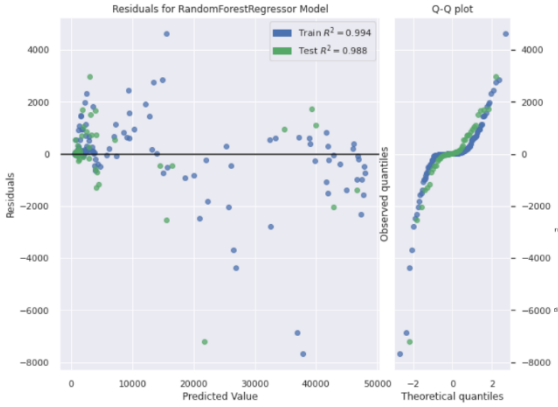
Πίνακας 96 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, RF, Βρυξέλλες, Ιδία Επεξεργασία*



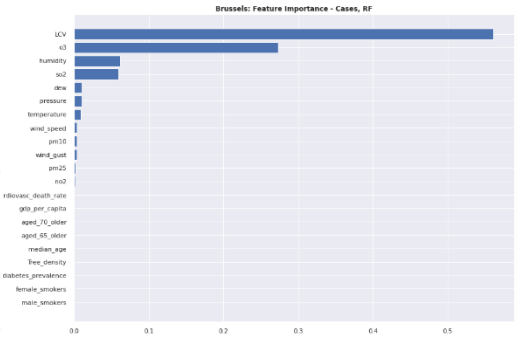
Σχήμα 373 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, RF, Βρυξέλλες



Σχήμα 374 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, RF, Βρυξέλλες



Σχήμα 375 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Βρυξέλλες

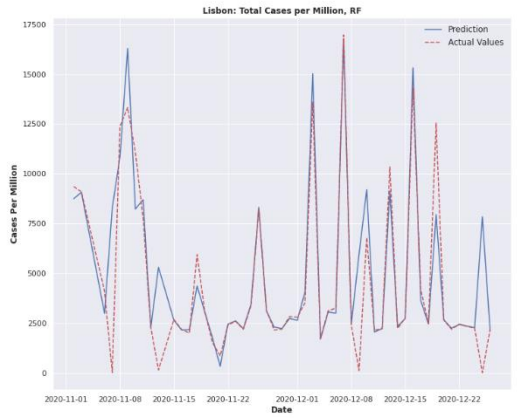


Σχήμα 376 Feature importance για πρόβλεψη κρουσμάτων, RF, Βρυξέλλες

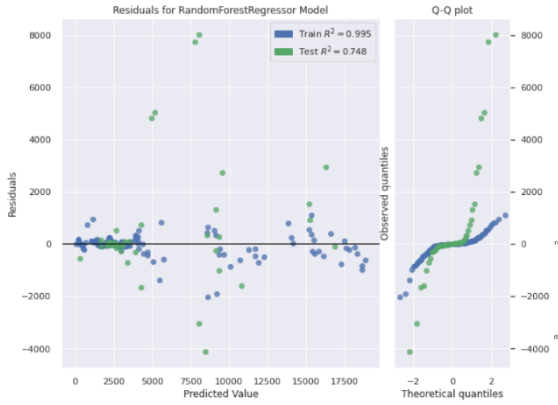
Μετρική Αξιολόγησης	Λισαβόνα
RMSE	2170.70 (cases per million)
R^2	0.736
EVS	0.748
MAE	1030.82 (cases per million)
MAPE	22.10%
MSE	0.00155

Πίνακας 97 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, RF, Λισαβόνα, Ιδία Επεξεργασία*

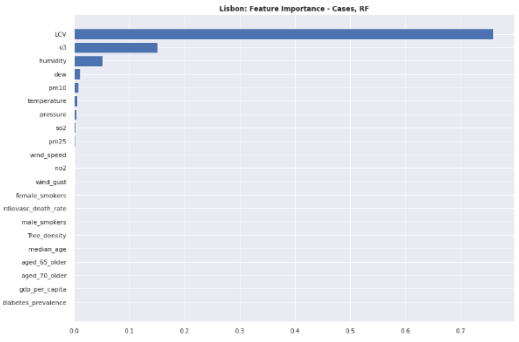
Σχήμα 377 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, RF, Λισαβόνα



Σχήμα 378 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, RF, Λισαβόνα



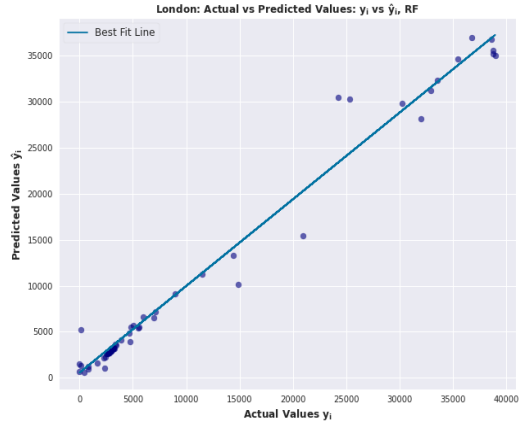
Σχήμα 379 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Λισαβόνα



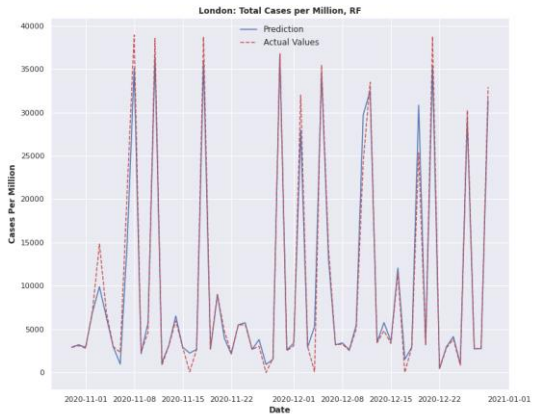
Σχήμα 380 Feature importance για πρόβλεψη κρουσμάτων, RF, Λισαβόνα

Μετρική Αξιολόγησης	Λονδίνο
RMSE	2023.58 (cases per million)
R^2	0.974
EVS	0.974
MAE	1050.72 (cases per million)
MAPE	38.36%
MSE	0.00135

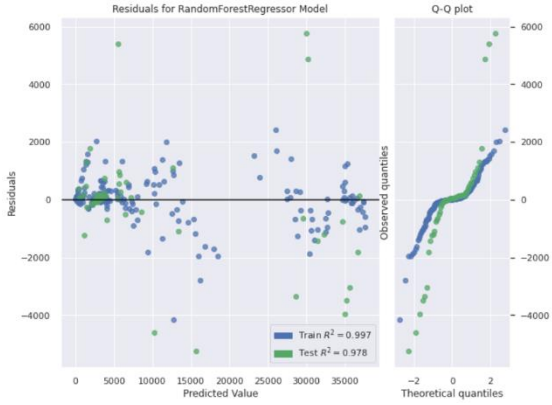
Πίνακας 98 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, RF, Λονδίνο, Ιδία Επεξεργασία*



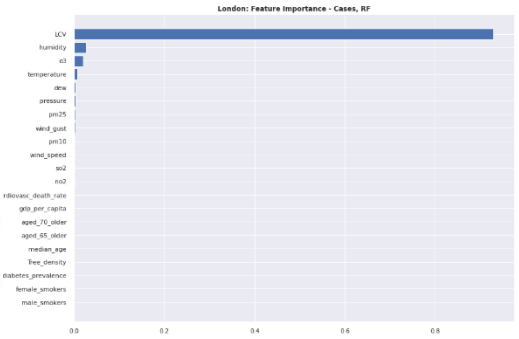
Σχήμα 381 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, RF, Λονδίνο



Σχήμα 382 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, RF, Λονδίνο



Σχήμα 383 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, RF, Λονδίνο

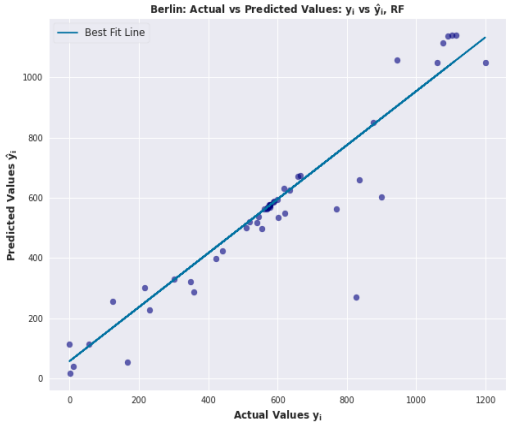


Σχήμα 384 Feature importance για πρόβλεψη κρουσμάτων, RF, Λονδίνο

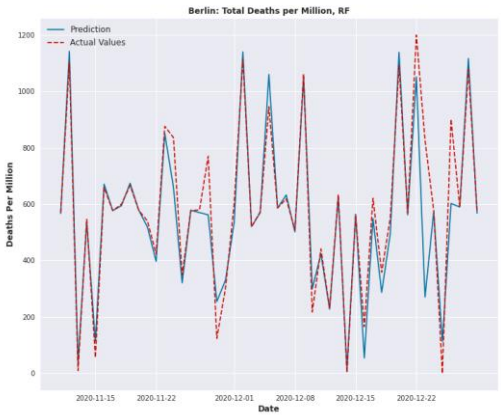
Πρόβλεψη Θανάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	122.13 (deaths per million)
R^2	0831
EVS	0.839
MAE	62.07 (deaths per million)
MAPE	11.74%
MSE	0.00538

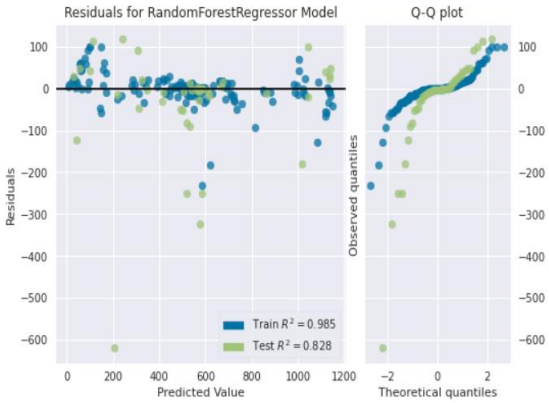
Πίνακας 99 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, RF, Βερολίνο, Ιδία Επεξεργασία*



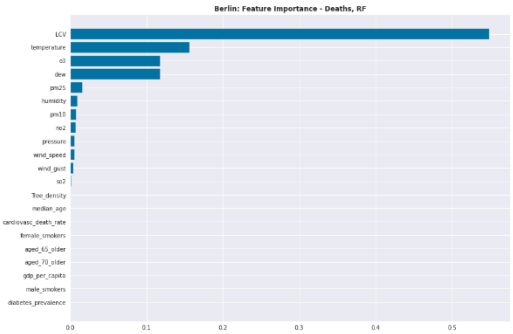
Σχήμα 385 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, RF, Βερολίνο, Ιδία Επεξεργασία



Σχήμα 386 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, RF, Βερολίνο, Ιδία Επεξεργασία



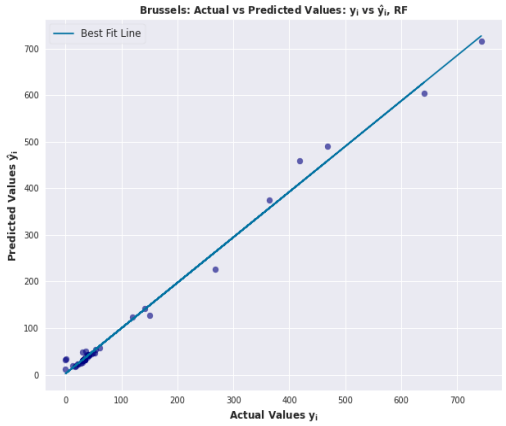
Σχήμα 387 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, RF, Βερολίνο, Ιδία Επεξεργασία



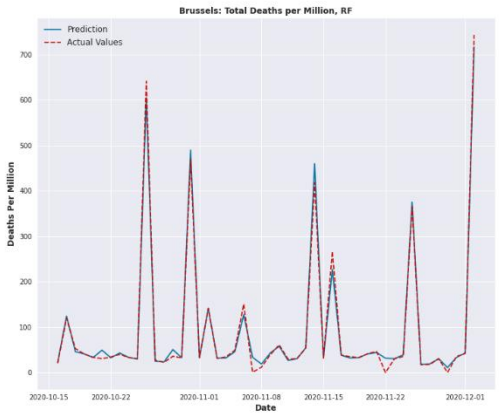
Σχήμα 388 Feature importance για πρόβλεψη θανάτων, RF, Βερολίνο, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	14.12 (deaths per million)
R^2	0.992
EVS	0.992
MAE	7.72 (deaths per million)
MAPE	3.99%
MSE	0.00007

Πίνακας 100 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, RF, Βρυξέλλες, Ιδία Επεξεργασία*

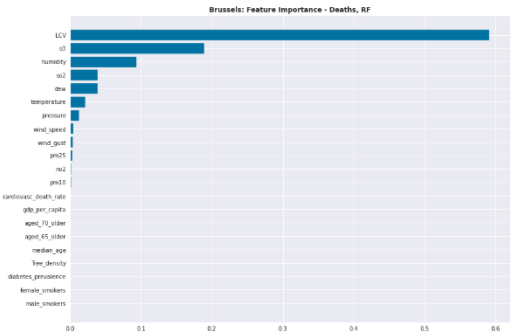


Σχήμα 389 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, RF, Βρυξέλλες, Ιδία Επεξεργασία



Σχήμα 390 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, RF, Βρυξέλλες, Ιδία Επεξεργασία

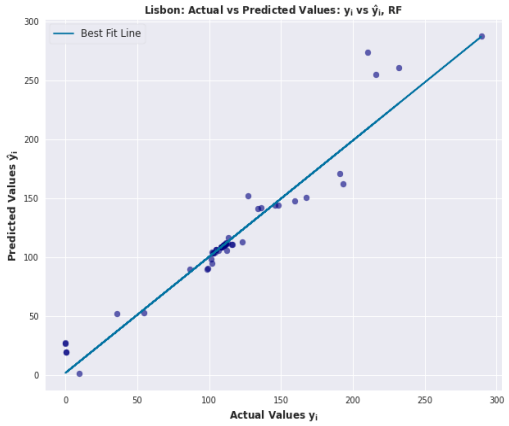
Σχήμα 391 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, RF, Βρυξέλλες, Ιδία Επεξεργασία



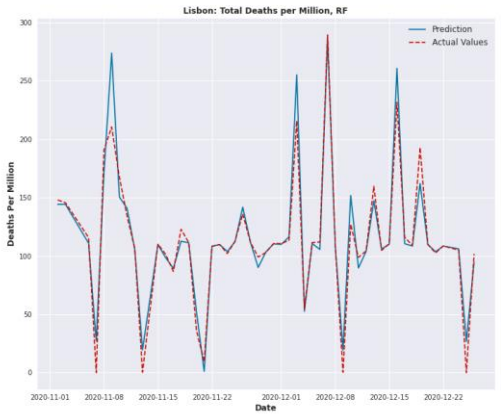
Σχήμα 392 Feature importance για πρόβλεψη θανάτων, RF, Βρυξέλλες, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Λισαβόνα
RMSE	13.67 (deaths per million)
R^2	0.940
EVS	0.941
MAE	7.35 (deaths per million)
MAPE	18.15%
MSE	0.00007

Πίνακας 101 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, RF, Λισαβόνα, Ιδία Επεξεργασία*

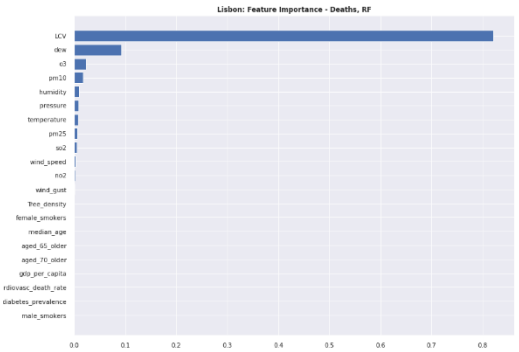


Σχήμα 393 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, RF, Λισαβόνα, Ιδία Επεξεργασία



Σχήμα 394 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, RF, Λισαβόνα, Ιδία Επεξεργασία

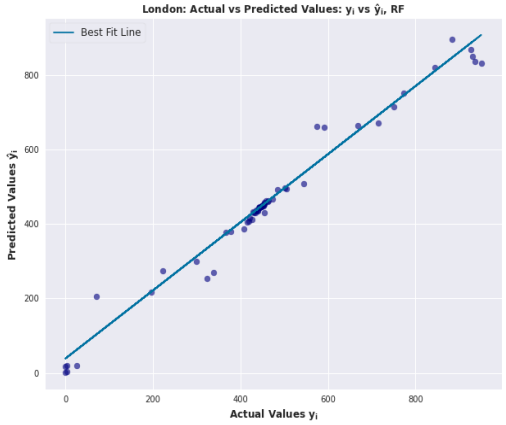
Σχήμα 395 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, RF, Λισαβόνα, Ιδία Επεξεργασία



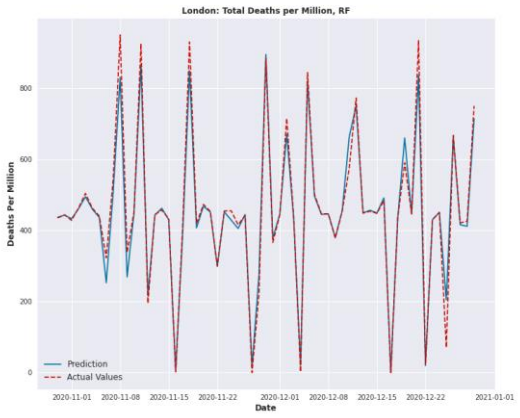
Σχήμα 396 Feature importance για πρόβλεψη θανάτων, RF, Λισαβόνα, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Λονδίνο
RMSE	38.27 (deaths per million)
R^2	0.970
EVS	0.970
MAE	21.18 (deaths per million)
MAPE	11.94%
MSE	0.00053

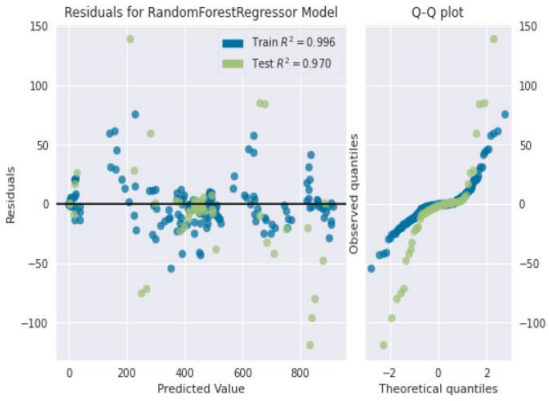
Πίνακας 102 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, RF, Λονδίνο, Ιδία Επεξεργασία*



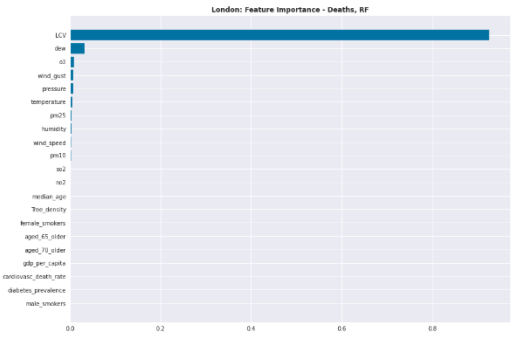
Σχήμα 397 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, RF, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 398 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, RF, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 399 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, RF, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 400 Feature importance για πρόβλεψη θανάτων, RF, Λονδίνο, Ιδία Επεξεργασία

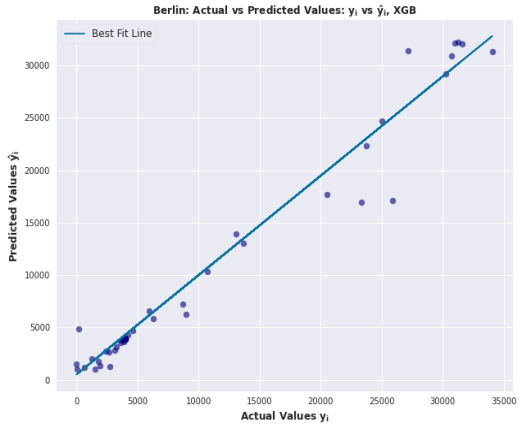
Παράρτημα ΣΤ

Τέλος, παρουσιάζονται τα αποτελέσματα για το XGBoost Regressor.

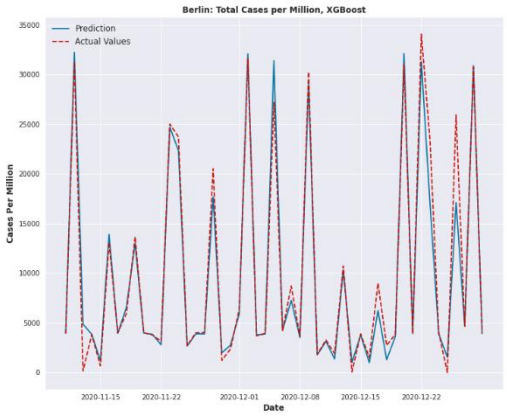
Πρόβλεψη Κρουσμάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	2009.74 (cases per million)
R^2	0.965
EVS	0.966
MAE	1035.96 (cases per million)
MAPE	3.86%
MSE	0.00133

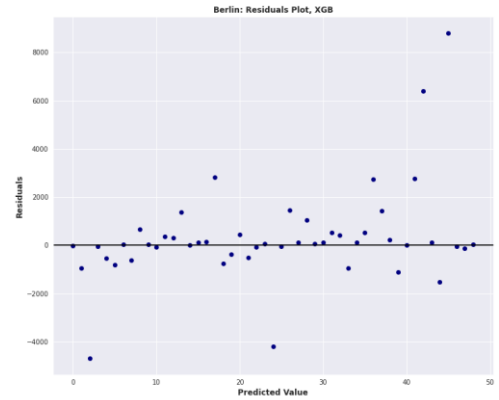
Πίνακας 103 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, XGBoost, Βερολίνο, Ιδία Επεξεργασία*



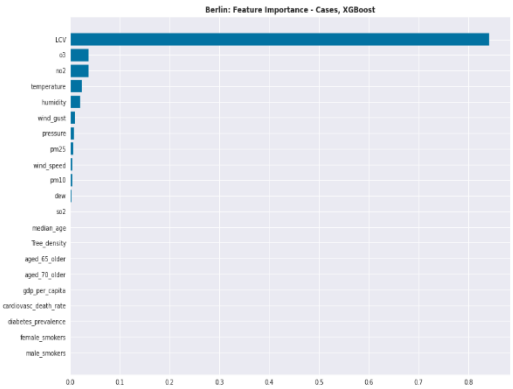
Σχήμα 401 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, XGBoost, Βερολίνο



Σχήμα 402 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, XGBoost, Βερολίνο



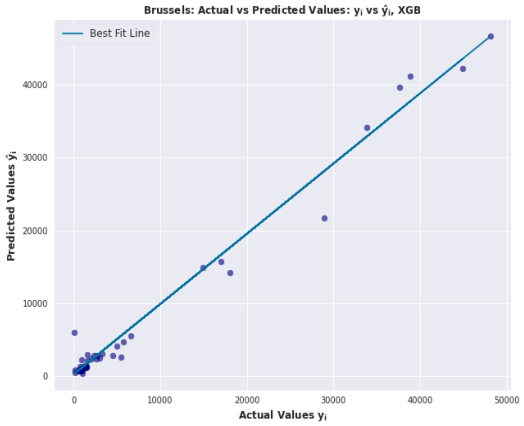
Σχήμα 403 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Βερολίνο



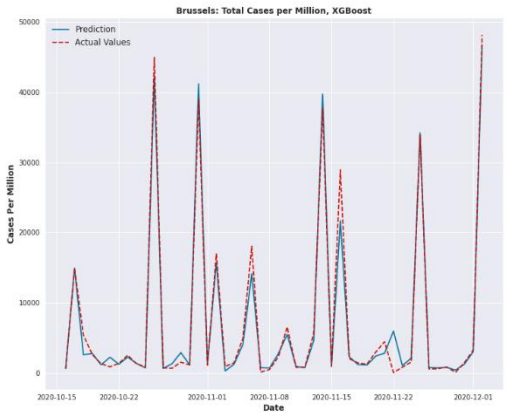
Σχήμα 404 Feature importance για πρόβλεψη κρουσμάτων, XGBoost, Βερολίνο

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	1720.83 (cases per million)
R^2	0.982
EVS	0.982
MAE	925.77 (cases per million)
MAPE	1.51%
MSE	0.00097

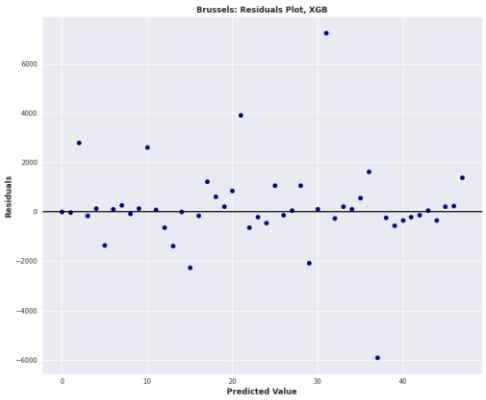
Πίνακας 104 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, XGBoost, Βρυξέλλες, Ιδία Επεξεργασία*



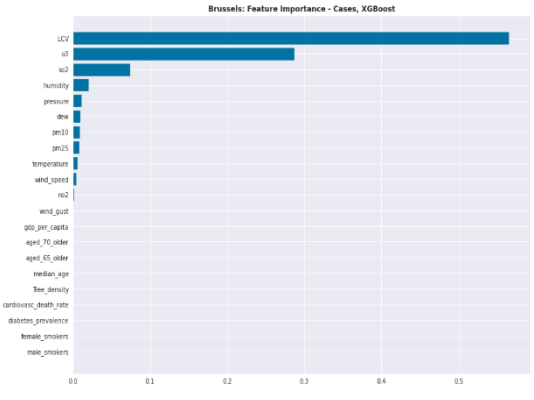
Σχήμα 405 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, XGBoost, Βρυξέλλες



Σχήμα 406 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, XGBoost, Βρυξέλλες



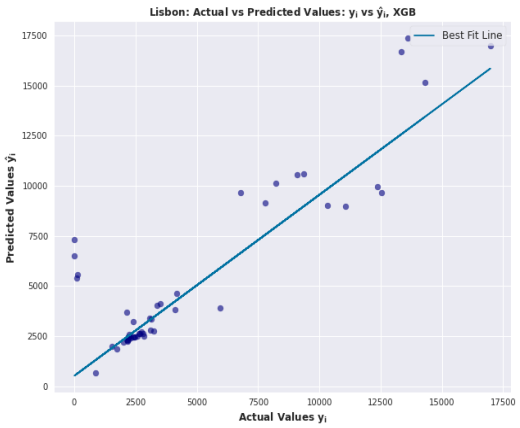
Σχήμα 407 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Βρυξέλλες



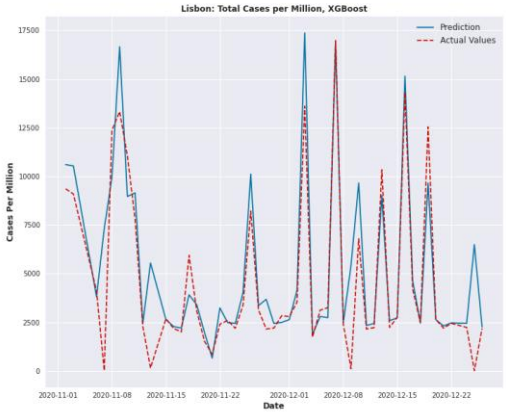
Σχήμα 408 Feature importance για πρόβλεψη κρουσμάτων, XGBoost, Βρυξέλλες

Μετρική Αξιολόγησης	Λισαβόνα
RMSE	2105.29 (cases per million)
R^2	0.752
EVS	0.779
MAE	1204.59 (cases per million)
MAPE	19.10%
MSE	0.00146

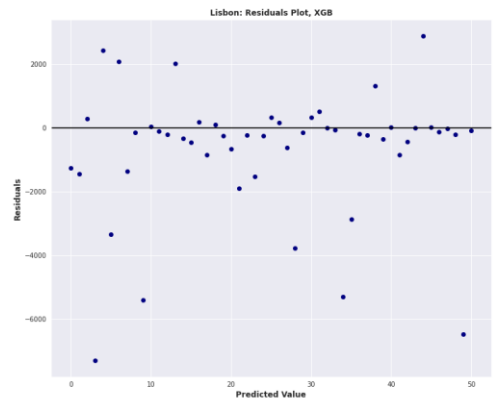
Πίνακας 105 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, XGBoost, Λισαβόνα, Ιδία Επεξεργασία*



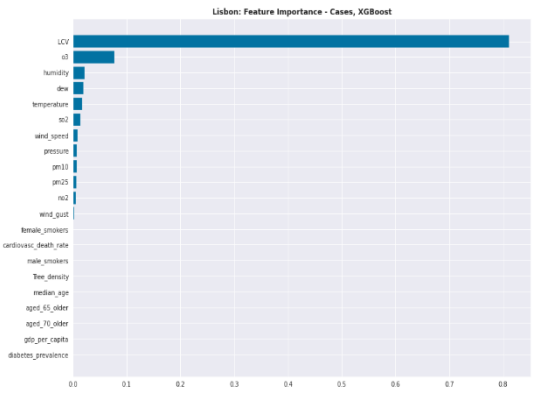
Σχήμα 409 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, XGBoost, Λισαβόνα



Σχήμα 410 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, XGBoost, Λισαβόνα



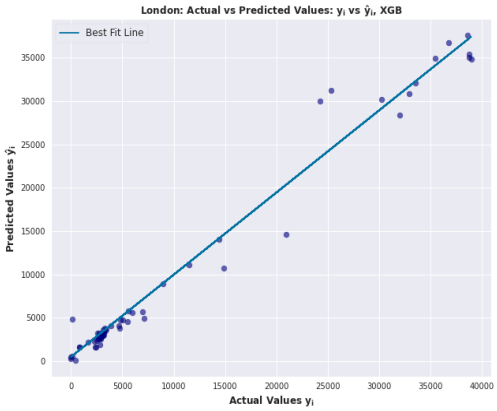
Σχήμα 411 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Λισαβόνα



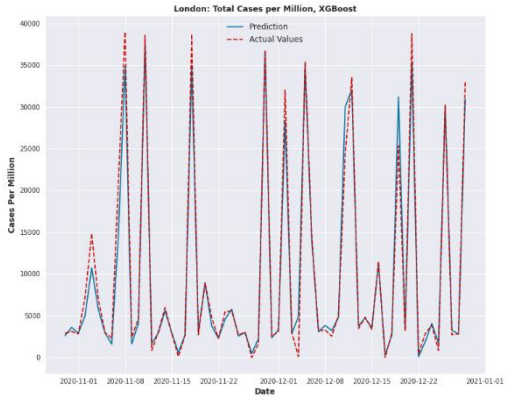
Σχήμα 412 Feature importance για πρόβλεψη κρουσμάτων, XGBoost, Λισαβόνα

Μετρική Αξιολόγησης	Λονδίνο
RMSE	1918.14 (cases per million)
R^2	0.976
EVS	0.977
MAE	1077.07 (cases per million)
MAPE	31.84%
MSE	0.00121

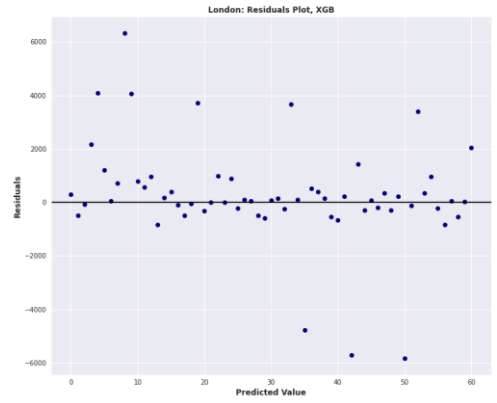
Πίνακας 106 *Μετρικές Αξιολόγησης, Πρόβλεψη Κρουσμάτων, XGBoost, Λονδίνο, Ιδία Επεξεργασία*



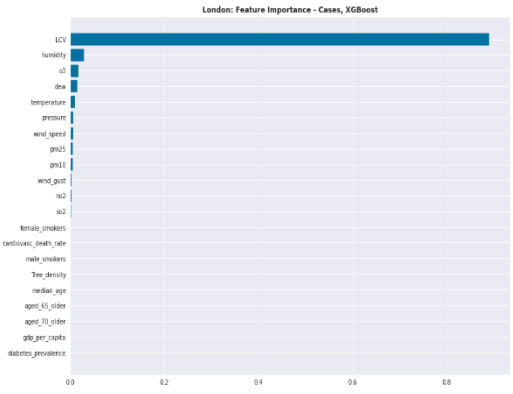
Σχήμα 413 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών κρουσμάτων, XGBoost, Λονδίνο



Σχήμα 414 Πρόβλεψη: Συνολικά κρούσματα στο εκατομμύριο, XGBoost, Λονδίνο



Σχήμα 415 Υπόλοιπα μοντέλου πρόβλεψης κρουσμάτων, XGBoost, Λονδίνο

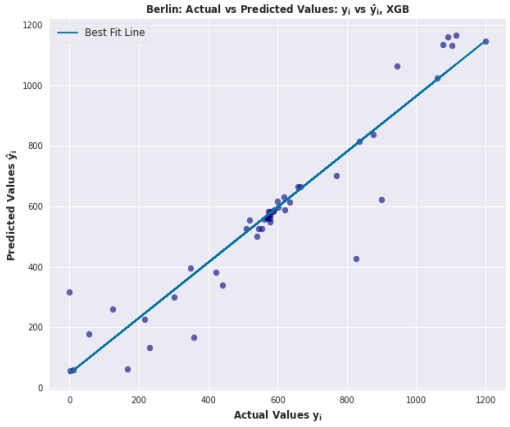


Σχήμα 416 Feature importance για πρόβλεψη κρουσμάτων, XGBoost, Λονδίνο

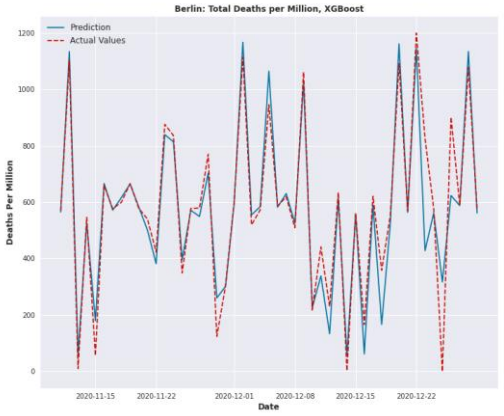
Πρόβλεψη Θανάτων

Μετρική Αξιολόγησης	Βερολίνο
RMSE	99.96 (deaths per million)
R^2	0.887
EVS	0.888
MAE	57.70 (deaths per million)
MAPE	23.63%
MSE	0.00361

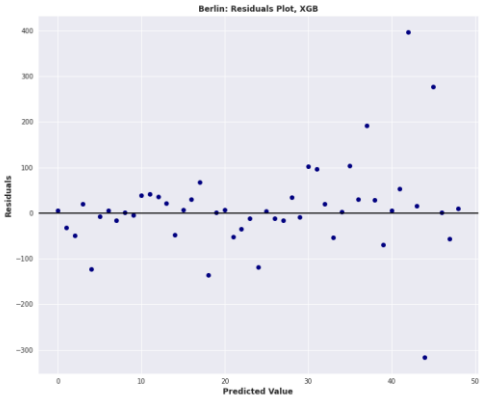
Πίνακας 107 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, XGBoost, Βερολίνο, Ιδία Επεξεργασία*



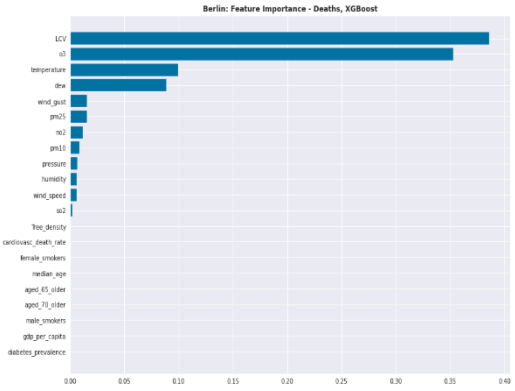
Σχήμα 417 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, XGBoost, Βερολίνο, Ιδία Επεξεργασία



Σχήμα 418 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, XGBoost, Βερολίνο, Ιδία Επεξεργασία



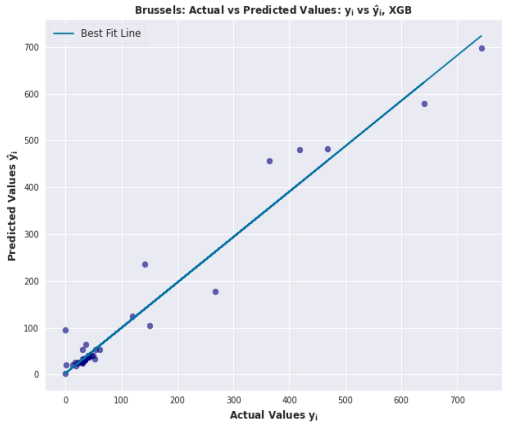
Σχήμα 419 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, XGBoost, Βερολίνο, Ιδία Επεξεργασία



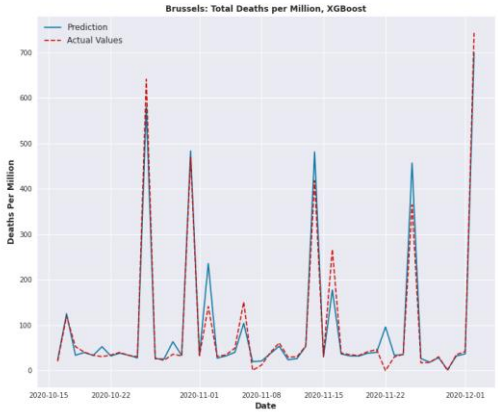
Σχήμα 420 Feature importance για πρόβλεψη θανάτων, XGBoost, Βερολίνο, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Βρυξέλλες
RMSE	32.04 (deaths per million)
R^2	0.960
EVS	0.960
MAE	16.68 (deaths per million)
MAPE	22.28%
MSE	0.00037

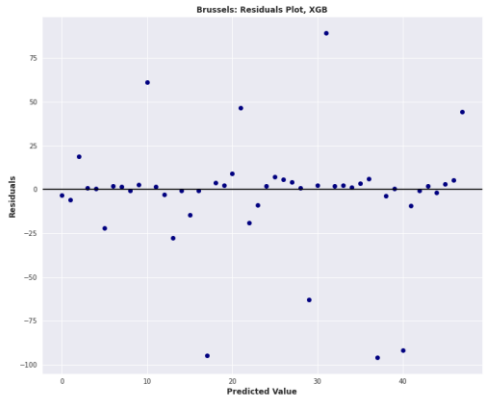
Πίνακας 108 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, XGBoost, Βρυξέλλες, Ιδία Επεξεργασία*



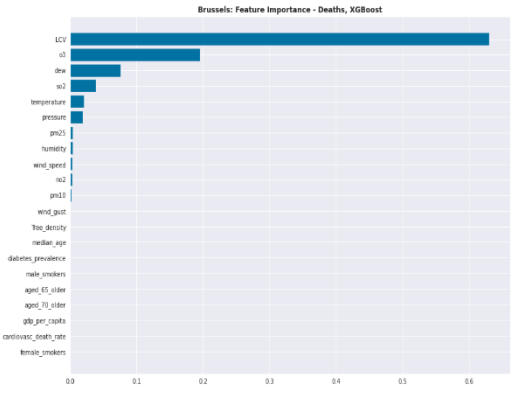
Σχήμα 421 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, XGBoost, Βρυξέλλες, Ιδία Επεξεργασία



Σχήμα 422 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, XGBoost, Βρυξέλλες, Ιδία Επεξεργασία



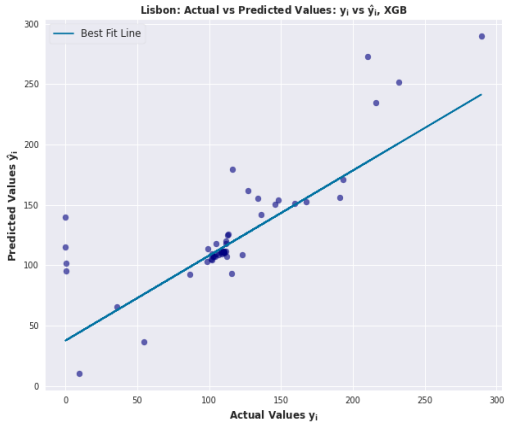
Σχήμα 423 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, XGBoost, Βρυξέλλες, Ιδία Επεξεργασία



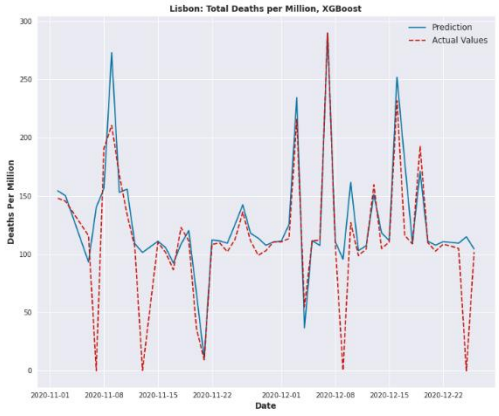
Σχήμα 424 Feature importance για πρόβλεψη θανάτων, XGBoost, Βρυξέλλες, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Λισαβόνα
RMSE	36.40 (deaths per million)
R^2	0.576
EVS	0.638
MAE	19.37 (deaths per million)
MAPE	220.87%
MSE	0.00048

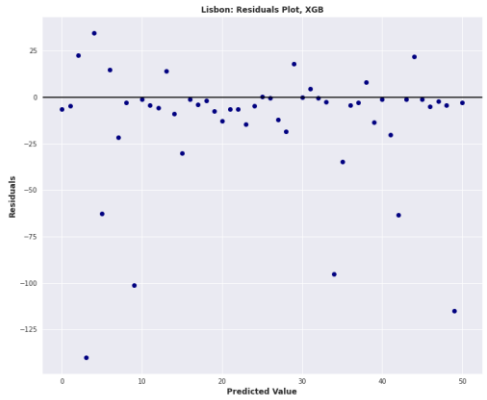
Πίνακας 109 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, XGBoost, Λισαβόνα, Ιδία Επεξεργασία*



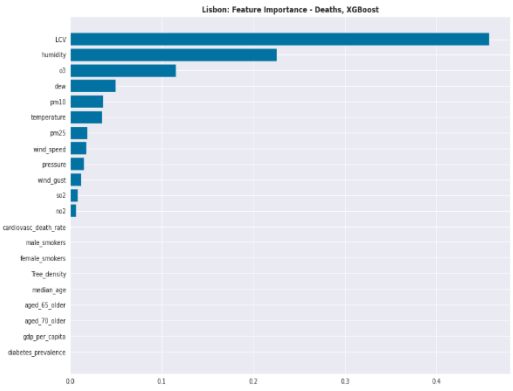
Σχήμα 425 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, XGBoost, Λισαβόνα, Ιδία Επεξεργασία



Σχήμα 426 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, XGBoost, Λισαβόνα, Ιδία Επεξεργασία



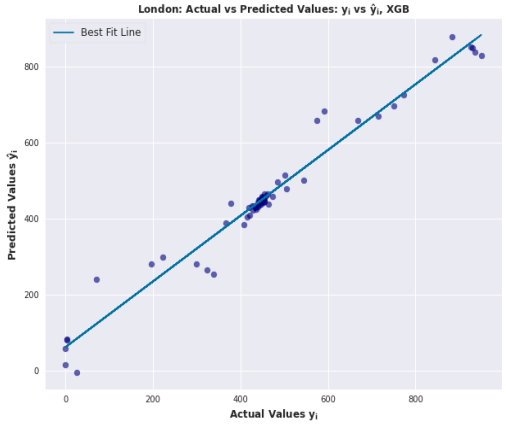
Σχήμα 427 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, XGBoost, Λισαβόνα, Ιδία Επεξεργασία



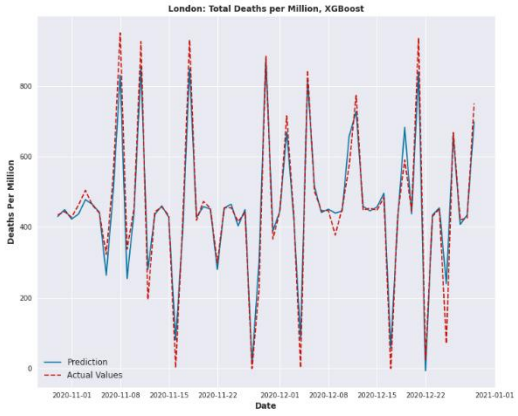
Σχήμα 428 Feature importance για πρόβλεψη θανάτων, XGBoost, Λισαβόνα, Ιδία Επεξεργασία

Μετρική Αξιολόγησης	Λονδίνο
RMSE	47.72 (deaths per million)
R^2	0.953
EVS	0.953
MAE	31.48 (deaths per million)
MAPE	18.54%
MSE	0.00082

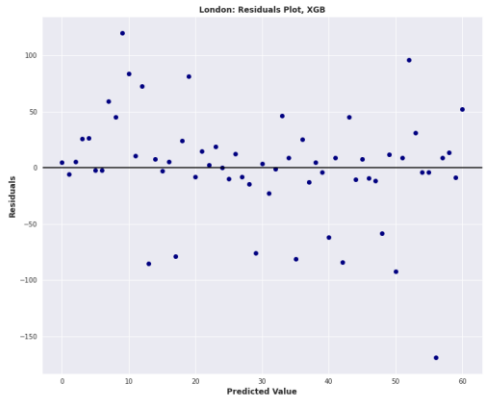
Πίνακας 110 *Μετρικές Αξιολόγησης, Πρόβλεψη Θανάτων, XGBoost, Λονδίνο, Ιδία Επεξεργασία*



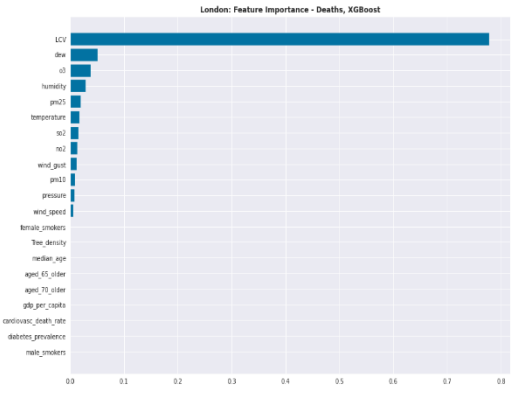
Σχήμα 429 Διάγραμμα διασποράς πραγματικών και προβλεπόμενων τιμών θανάτων, XGBoost, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 430 Πρόβλεψη: Συνολικοί θάνατοι στο εκατομμύριο, XGBoost, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 431 Υπόλοιπα μοντέλου πρόβλεψης θανάτων, XGBoost, Λονδίνο, Ιδία Επεξεργασία



Σχήμα 432 Feature importance για πρόβλεψη θανάτων, XGBoost, Λονδίνο, Ιδία Επεξεργασία

Αναφορές

A. Temenos, I. Tzortzis, M. Kaselimi, I. Rallis, A. Doulamis, N. Doulamis. (2022, Μάιος 31). Novel Insights in Spatial Epidemiology Utilizing Explainable. *MDPI*, σ. 20.

Allwright, S. (2021, Οκτώβριος 26). *What is a good MAE score?* Ανάκτηση από stephenallwright: https://stephenallwright.com/good-mae-score/

Allwright, S. (2021, Οκτώβριος 27). *What is a good MAPE score?* Ανάκτηση από stephenallwright: https://stephenallwright.com/good-mape-score/

Allwright, S. (2022, Απρίλιος 7). *What is a good MSE score? Mean Squared Error explained!* Ανάκτηση από stephenallwright: https://stephenallwright.com/good-mse-score/

Allwright, S. (2022, Απρίλιος 16). *What is a good R2 (R-Squared) score and how do I interpret it?* Ανάκτηση από stephenallwright: https://stephenallwright.com/good-r2-score/

Amadebai, E. (χ.χ.). *10 of the Best Data Visualization Libraries in Python.* Ανάκτηση από analyticsfordecisions: https://www.analyticsfordecisions.com/best-data-visualization-libraries-in-python/

Ammar H. Elsheikh et al. (2020, Ιούνιος 9). Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia. *Elsevier*, σ. 11.

Anunaya, S. (2021, Αύγουστος 10). *Data Preprocessing in Data Mining -A Hands On Guide.* Ανάκτηση από analyticsvidhya: https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/

Bajaj, A. (2022, Μάρτιος 18). *Performance Metrics in Machine Learning.* Ανάκτηση από neptune: https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

Bakshi, C. (2020, Ιούνιος 9). *Random Forest Regression.* Ανάκτηση από levelup.gitconnected: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Biswas, S. K. (2020, Ιούλιος 30). *What is the Acceptable MSE value and Coefficient of determination(R2)?* Ανάκτηση από researchgate: https://www.researchgate.net/post/What_is_the_Acceptable_MSE_value_and_Coefficient_of_determinationR2

Boutsikas, M. (2004). *Ενότητα 5: Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression).* Ανάκτηση από unipi: http://www.unipi.gr/faculty/mbouts/statprog/SPSS_lesson9-10.pdf

Boutsikas, M. (2004). *Ενότητα 6: Πολλαπλή Γραμμική Παλινδρόμηση (Multiple Linear Regression).* Ανάκτηση από unipi: https://www.unipi.gr/faculty/mbouts/statprog/SPSS_lesson11.pdf

Brownlee, J. (2016, Μάρτιος 25). *Linear Regression for Machine Learning.* Ανάκτηση από machinelearningmastery: https://machinelearningmastery.com/linear-regression-for-machine-learning/

Buchta, H. (2015, Οκτώβριος 19). *Trend in times series analysis.* Ανάκτηση από oraylis: https://www.oraylis.de/blog/2015/trend-in-times-series-analysis

Chai and Draxler. (2014, Ιούνιος 30). *Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature.* Ανάκτηση από Researchgate: https://www.researchgate.net/profile/Tianfeng-Chai/publication/272024186_Root_mean_square_error_RMSE_or_mean_absolute_error_MAE-_Arguments_against_avoiding_RMSE_in_the_literature/links/54e3776f0cf2b2314f5d2f3c/Root-mean-square-error-RMSE-or-mean-absolute-

Chakure, A. (2022, Μάρτιος 7). *Implementing Random Forest Regression in Python: An Introduction.* Ανάκτηση από builtin: https://builtin.com/data-science/random-forest-python

Chng, Z. M. (2022, Απρίλιος 28). *Google Colab for Machine Learning Projects.* Ανάκτηση από machinelearningmastery: https://machinelearningmastery.com/google-colab-for-machine-learning-projects/

Dar, P. (2018, Μάιος 27). *Yellowbrick – A set of Visualization Tools to Accelerate your Model Selection Process.* Ανάκτηση από analyticsvidhya: https://www.analyticsvidhya.com/blog/2018/05/yellowbrick-a-set-of-visualization-tools-to-accelerate-your-model-selection-process/

DataCamp Team. (2022, Μάρτιος 25). *Lasso and Ridge Regression Tutorial*. Ανάκτηση από datacamp: https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression

Donges, N. (2022, Ιουλίου 6). *Random Forest Algorithm: A Complete Guide*. Ανάκτηση από builtin: https://builtin.com/data-science/random-forest-algorithm

EPA. (2010, Ιούλιος). *Quantitative Risk and Exposure Assessment for Carbon Monoxide - Amended*. Ανάκτηση από epa.gov: https://www.epa.gov/sites/default/files/2020-07/documents/co-rea-amended-july2010.pdf

EPA. (2021, Ιούνιος 7). *Basic Information about Carbon Monoxide (CO) Outdoor Air Pollution* . Ανάκτηση από epa.gov: https://www.epa.gov/co-pollution/basic-information-about-carbon-monoxide-co-outdoor-air-pollution

EPA. (2022, Ιουνίου 14). *Ground-level Ozone Basics* . Ανάκτηση από epa.gov: https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics

ESA. (2013, Ιούνιος 3). *Ατμοσφαιρική ρύπανση*. Ανάκτηση από esa: https://www.esa.int/SPECIALS/Eduspace_Global_GR/SEMQ3T4SZLG_0.html

Flovik, V. (2018, Ιούνιος 7). *How (not) to use Machine Learning for time series forecasting: Avoiding the pitfalls*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/how-not-to-use-machine-learning-for-time-series-forecasting-avoiding-the-pitfalls-19f9d7adf424

Ford, C. (2015, Αύγουστος 26). *Understanding Q-Q Plots*. Ανάκτηση από library.virginia: https://data.library.virginia.edu/understanding-q-q-plots/

Frost, J. (2017, Απρίλιος 12). *How to Interpret P-values and Coefficients in Regression Analysis*. Ανάκτηση από statisticsbyjim: https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

Garlapati, H. V. (2021, Απρίλιος 9). *Machine Learning Model Evaluation*. Ανάκτηση από knowledgehut: https://www.knowledgehut.com/blog/data-science/machine-learning-model-evaluation

Great Learning Team. (2021, Δεκέμβριος 26). *A Complete understanding of LASSO Regression*. Ανάκτηση από mygreatlearning: https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/

Great Learning Team. (2022, Ιανουάριος 15). *Machine Learning Tutorial For Complete Beginners | Learn Machine Learning with Python*. Ανάκτηση από mygreatlearning: https://www.mygreatlearning.com/blog/machine-learning-tutorial/

Great Learning Team. (2022, Ιανουάριος 19). *What is Artificial Intelligence? How does AI work, Types, Trends and Future of it?* Ανάκτηση από mygreatlearning: https://www.mygreatlearning.com/blog/what-is-artificial-intelligence/#introduction-to-artificial-intelligence

Hyndman, R. J. (2006, Ιούνιος 4). *Another Look at Forecast-Accuracy Metrics*. Ανάκτηση από robjhyndman: https://robjhyndman.com/papers/foresight.pdf

I. Kavouras et al. (2021, Ιούνιος). Machine Learning Tools to Assess the Impact of COVID-19 Civil Measures in Atmospheric Pollution. *PETRA 2021*, σσ. 396–403.

IBM Cloud Education. (2020, Αύγουστος 19). *Supervised Learning*. Ανάκτηση από ibm: https://www.ibm.com/cloud/learn/supervised-learning

Ioannis Kavouras et al. (2022, Μάρτιος 23). COVID-19 Spatio-Temporal Evolution Using Deep Learning at a European Level. *MDPI*, σ. 25.

Jain, D. (2021, Ιούνιος 29). *Data Preprocessing in Data Mining*. Ανάκτηση από geeksforgeeks: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Jiwon Park. (2021, Ιούνιος 10). *How to Interpret a Residual Plot*. Ανάκτηση από study: https://study.com/skill/learn/how-to-interpret-a-residual-plot-explanation.html

Krishna, H. (2020, Σεπτέμβριος 25). *Test accuracy higher than the training accuracy*. Ανάκτηση από kaggle: https://www.kaggle.com/questions-and-answers/186681

Lendave, V. (2021, Νοέμβριος 1). *A Guide to Different Evaluation Metrics for Time Series Forecasting Models*. Ανάκτηση από analyticsindiamag: https://analyticsindiamag.com/a-guide-to-different-evaluation-metrics-for-time-series-forecasting-models/

Lewinson, E. (2022, Μάρτιος 22). *Time Series DIY: Seasonal Decomposition*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/time-series-diy-seasonal-decomposition-f0b469afed44

Lung-Chang Chien, Lung-Wen Chen. (2020, Ιούλιος 4). *Meteorological impacts on the incidence of COVID-19 in the U.S*. Ανάκτηση από Springer Link: https://link.springer.com/article/10.1007/s00477-020-01835-8

Malato, G. (2020, Μάιος 6). *Why training set should always be smaller than test set*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/why-training-set-should-always-be-smaller-than-test-set-61f087ed203c

Microsoft. (2021, Μάιος 18). *Responsible and trusted AI.* Ανάκτηση από microsoft: https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/trusted-ai

Minnesota Pollution Control Agency. (2020, Νοέμβριος 20). *Sources of air pollution that most impact health*. Ανάκτηση από pca.state.mn.us: https://www.pca.state.mn.us/air/sources-air-pollution-most-impact-health

Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). *The basic practice of statistics (6th ed.)*. New York: NY: W. H. Freeman and Company.

Morde, V. (2019, Απρίλιος 8). *XGBoost Algorithm: Long May She Reign!* Ανάκτηση από towardsdatascience: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

National Park Service. (2018, Ιανουάριος 17). *Where Does Air Pollution Come From?* Ανάκτηση από nps.gov: https://www.nps.gov/subjects/air/sources.htm

Nooshin Ayoobi et al. (2021, Μάρτιος 22). Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods. *ELSEVIER*, σ. 15.

Pattabiraman, M. G. (χ.χ.). *Data Visualization using Python.* Ανάκτηση από mygreatlearning: https://www.mygreatlearning.com/academy/learn-for-free/courses/data-visualization-using-python?utm_source_page=public_certificate_view&utm_source_cta=enrol_for_free

Raj, A. (2020, Οκτώβριος 3). *Unlocking the True Power of Support Vector Regression*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0

Ranjan, V. (2020, Ιούλιος 12). *What is a Time Series?* Ανάκτηση από medium: https://medium.com/analytics-vidhya/what-is-a-time-series-fab8ebc4451b

Rao, D. (2021, Μάιος 21). *Fairness in AI systems – Everything you need know!* . Ανάκτηση από persistent: https://www.persistent.com/blogs/fairness-in-ai-systems/

Rink, K. (2021, Οκτώβριος 21). *Time Series Forecast Error Metrics You Should Know.* Ανάκτηση από towardsdatascience: https://towardsdatascience.com/time-series-forecast-error-metrics-you-should-know-cc88b8c67f27

Ronaghan, S. (2018, Μάιος 11). *The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

Sharp, T. (2020, Μάρτιος 3). *An Introduction to Support Vector Regression (SVR).* Ανάκτηση από towardsdatascience: https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

Singh, A. (2019, Ιούνιος 23). *Support Vector Regression*. Ανάκτηση από medium: https://medium.com/@singhakshay.etw69/support-vector-regression-4216416f94c0

Sit, H. (2019, Ιούνιος 19). *Quick Start to Gaussian Process Regression.* Ανάκτηση από towardsdatascience: https://towardsdatascience.com/quick-start-to-gaussian-process-regression-36d838810319

sketchalytics. (2019, Αύγουστος 28). *The 3 Types of Machine Learning.* Ανάκτηση από ceralytics: https://www.ceralytics.com/3-types-of-machine-learning/

Tableau. (χ.χ.). *What Is Data Visualization? Definition, Examples, And Learning Resources.* Ανάκτηση από tableau: https://www.tableau.com/learn/articles/data-visualization

The scikit-yb developers. (2022, Μάρτιος 19). *Residuals Plot.* Ανάκτηση από .scikit-yb.org: https://www.scikit-yb.org/en/latest/api/regressor/residuals.html

ThermiAir. (2019, Μάρτιος). *Ατμοσφαιρική Ρύπανση*. Ανάκτηση από thermiair: http://www.thermiair.gr/project/air-quality/

UNESCO International Science. (1997). The chemistry of atmospheric policy Vol. XXII No. 2. *Technology & Environmental Education Newsletter*.

Watson. (2021, Μάρτιος 26). *Explainable AI*. Ανάκτηση από IBM: https://www.ibm.com/watson/explainable-ai

Wikipedia. (2021, Αυγούστου 13). *Διοξείδιο του αζώτου*. Ανάκτηση από wikipedia: https://el.wikipedia.org/wiki/%CE%94%CE%B9%CE%BF%CE%BE%CE%B5%CE%AF%CE%B4%CE%B9%CE%BF_%CF%84%CE%BF%CF%85_%CE%B1%CE%B6%CF%8E%CF%84%CE%BF%CF%85

Wikipedia. (2022, Ιανουάριος). *Our World in Data*. Ανάκτηση από wikipedia: https://en.wikipedia.org/wiki/Our_World_in_Data

Yahya. (2020, Φεβρουάριος 14). *Python sci-kit learn (metrics): difference between r2_score and explained_variance_score?* Ανάκτηση από stackoverflow: https://stackoverflow.com/questions/24378176/python-sci-kit-learn-metrics-difference-between-r2-score-and-explained-varian

Yashwanth, N. (2020, Οκτώβριος 7). *Evaluation metrics & Model Selection in Linear Regression.* Ανάκτηση από towardsdatascience: https://towardsdatascience.com/evaluation-metrics-model-selection-in-linear-regression-73c7573208be

Yongfa You, Shufen Pan. (2020, Σεπτέμβριος 4). *Urban Vegetation Slows Down the Spread of Coronavirus Disease in the United States.* Ανάκτηση από Advancing Earth and Space Science: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020GL089286

Accenture. (2022, Απρίλιος 6). *Responsible AI: Scale AI with confidence.* Ανάκτηση από accenture: https://www.accenture.com/us-en/services/applied-intelligence/ai-ethics-governance

Ανδρικάκης, Α. (2017). *Μηχανική Μάθηση σε Ανομοιογενή Δεδομένα - (Machine Learning in Imbalanced Data Sets).* Πειραιάς: Πανεπιστήμιο Πειραιώς.

Βαρελάς, Ν. Δ. (2019, Μάρτιος). *Μέθοδοι μελέτης του ρυθμού απώλειας πελατών και της αξίας συνολικού χρόνου ζωής πελάτη.* Ανάκτηση από dione.lib.unipi: https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/12122/Varelas_MES17016.pdf?sequence=2&isAllowed=y

Ελευθερίου, Ε. (2021). *Ανάπτυξη μοντέλου πρόβλεψης εξασθένησης σήματος λόγω βροχής σε δορυφορικές ζεύξεις με τεχνικές παλινδρόμησης μηχανικής μάθησης.* Εθνικό Μετσόβλιο Πολυτεχνείο.

Επιπτώσεις στην Υγεία. (2013). Ανάκτηση από airquality.dli.mlsi.gov: https://www.airquality.dli.mlsi.gov.cy/el/health-effects

Ευρωπαϊκό Κοινοβούλιο. (2020, Σεπτέμβριος 9). *Τι είναι η τεχνητή νοημοσύνη και πώς χρησιμοποιείται;* . Ανάκτηση από europarl.europa.eu: https://www.europarl.europa.eu/news/el/headlines/society/20200827STO85804/ti-einai-i-techniti-noimosuni-kai-pos-chrisimopoieitai

Ευρωπαϊκός Οργανισμός Περιβάλλοντος. (2020, Νοέμβριος 23). *Ατμοσφαιρική ρύπανση* . Ανάκτηση από eea.europa.eu: https://www.eea.europa.eu/el/themes/air/intro

Κοτρωνάκη, Ι. (2021). *Αναγνώριση ρωγμών με χρήση τεχνολογιών μη επιβλεπόμενης βαθιάς μηχανικής μάθησης.* Εθνικό Μετσόβιο Πολυτεχνείο.

Λοΐζου, Ε. (2020). *Ανάλυση Χρονοσειρών με εφαρμογή σε βιοϊατρικά δεδομένα.* Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο. Ανάκτηση από https://dspace.lib.ntua.gr/xmlui/bitstream/handle/123456789/52798/thesis3_rev11%209-2%20FINAL_rev-2.pdf?sequence=1&isAllowed=y

Μουστρής, Δ. (2015, Οκτώβριος 9). *Τεχνολογία Περιβαλλοντικών Μετρήσεων.* Ανάκτηση από eclass.teipir: http://eclass.teipir.gr/openeclass/modules/units/?course=MECH111&id=655

Μπίνου, Γ. (2016). *Ανάλυση και Πρόβλεψη Χρονοσειρών: Μέθοδοι και Εφαρμογές.* Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο.

Σιμιτσής, Δ. (2013). *ΜΟΝΟΞΕΙΔΙΟ ΤΟΥ ΑΝΘΡΑΚΑ - «Ο ΣΙΩΠΗΛΟΣ ΔΟΛΟΦΟΝΟΣ».* Ανάκτηση από eaps: https://www.eaps.gr/wp-content/uploads/2013/11/CO-simitsis.pdf

Σταμάτης, Δ. (2021, Νοέμβριος 30). *Εξηγήσιμη Τεχνητή Νοημοσύνη.* Ανάκτηση από people.iee.ihu: https://people.iee.ihu.gr/~demos/Downloads/AI_ST_10_XAI.pdf

Σχολή Μηχανολόγων Μηχανικών - Πανεπιστήμιο Θεσσαλίας. (2020). *Σωματιδιακοί Ρύποι - Αιωρούμενα Σωματίδια.* Ανάκτηση από mie.uth: http://www.mie.uth.gr/ekp_yliko/3_particulates.pdf

Φωτιάδη, Α. (2015, Φεβρουάριος 19). *Ατμοσφαιρική Ρύπανση - Ατμοσφαιρικοί Ρύποι*. Ανάκτηση από eclass.upatras: https://eclass.upatras.gr/modules/document/file.php/ENV113/5%CE%B7%20%CE%94%CE%B9%CE%AC%CE%BB%CE%B5%CE%BE%CE%B7.pdf

Χρυσοπούλου, Ζ. (2019). *Τεχνικές Μηχανικής Μάθησης για Ανίχνευση Ρητορικής Μίσους*. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.