ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ανάπτυξη Πλατφόρμας για την Συλλογή και την Περιγραφική Ανάλυση Δεδομένων στα πλαίσιο μιας έξυπνης πόλης και πρόβλεψη χρονοσειρών με στατιστικά μοντέλα και μοντέλα μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ξενία Δημητρίου

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάπτυξη Πλατφόρμας για την Συλλογή και την Περιγραφική Ανάλυση Δεδομένων στα πλαίσιο μιας έξυπνης πόλης και πρόβλεψη χρονοσειρών με στατιστικά μοντέλα και μοντέλα μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ξενίας Δημητρίου

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή τον Φεβρουάριο του 2022.

.....	.....	.....
Δημήτριος Ασκούνης Καθηγητής Ε.Μ.Π.	Ιωάννης Ψαρράς Καθηγητής Ε.Μ.Π.	Χρυσόστομος Δούκας Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2022

.....

ΞΕΝΙΑΣ ΔΗΜΗΤΡΙΟΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Οι ολοένα και περισσότερες αναδυόμενες τεχνολογίες που ξεπροβάλλουν τα τελευταία χρόνια έχουν δημιουργήσει νέα όπλα στη φαρέτρα μας, με τα οποία μπορούμε να προσεγγίσουμε υπάρχοντα προβλήματα με έναν πιο σύγχρονο και φρέσκο τρόπο. Μάλιστα, επειδή οι νέες τεχνολογίες εκτείνονται σε ένα μεγάλο φάσμα της επιστήμης των υπολογιστών, ο συνδυασμός αυτών μπορεί να οδηγήσει σε συστήματα που παλαιότερα θα φάνταζαν σενάρια επιστημονικής φαντασίας. Στο πλαίσιο αυτό, είναι ευκαιρία να συνθέσουμε τις νέες τεχνολογίες για να επιλύσουμε προβλήματα της ανθρωπότητας, να δημιουργήσουμε ίσως νέες ανάγκες και εν τέλη να βελτιώσουμε την ποιότητα ζωής των ανθρώπων.

Υπό αυτό το πρίσμα, η συγκεκριμένη διπλωματική στοχεύει στη δημιουργία ενός εργαλείου που επικεντρώνεται στην ενίσχυση των διαδικασιών λήψης αποφάσεων από τις υπεύθυνες αρχές μιας πόλης. Αποτελεί ένα επικουρικό εργαλείου που θα στελεχώσει το στόλο των εφαρμογών που θα χρησιμοποιηθούν κατά την ψηφιακή μεταπήδηση των έξυπνων πλέον πόλεων. Είναι ένα εργαλείο που αρχικά συλλέγει δεδομένα από τους διάφορους αισθητήρες που υπάρχουν και θα υπάρξουν μέσα στην πόλη και αναλαμβάνει την αποθήκευσή τους σε κάποιο σύστημα αποθήκευσης δεδομένων. Ύστερα μέσα από αυτό μπορεί ο χρήστης του συστήματος να προβεί στην Περιγραφική και Διαγνωστική Ανάλυση των δεδομένων μέσα από μια μεγάλη ποικιλία διαγραμμάτων που στόχο έχουν την ανάδειξη "κρυμμένων" πληροφοριών, βοηθώντας τη δουλεία των αναλυτών. Τέλος, διαθέτει 9 αλγορίθμους πρόβλεψης χρονοσειρών και παρέχει στο χρήστη προβλέψεις για συγκεκριμένο χρονικό ορίζοντα. Οι προβλέψεις αυτές προέρχονται από στατιστικά μοντέλα και μοντέλα μηχανικής μάθησης που έχουν εκπαιδευτεί με στόχο την πιο ακριβή πρόβλεψη. Δίνει επίσης τη δυνατότητα στον χρήστη να πειραματιστεί με τις παραμέτρους των μοντέλων και να λάβει χρήσιμες πληροφορίες για τις μεταβολές των αλλαγών του.

Λέξεις Κλειδιά

Περιγραφική Ανάλυση, Διαγνωστική Ανάλυση, Προγνωστική Ανάλυση, Ανάλυση Δεδομένων, Έξυπνες Πόλεις, Στατιστικά Μοντέλα Πρόβλεψης, Μοντέλα Μηχανικής Μάθησης, Πρόβλεψη Χρονοσειρών

Abstract

The new technologies that have emerged in the last decade, have created new paths to the solution of many of humanity's problems in a fresh and more modern way. On top of that, considering that these new technologies spread in an extended spectrum of computer science, it’s getting more and more feasible and vital to combine them in order to create systems that possibly would have been claimed as scientific fiction some years ago. Thus, we have the opportunity to combine these new technologies in order to improve the quality of human life.

Under these circumstances, the goal of this diploma thesis is to create a tool-system that aims to help the authorities of a city to make better decisions for everyday problems. It will be a tool which will be a part of the suite of new systems that will take part in the new era of digital transformation in smart cities. To kick off with the details of the system, it initially collects the data coming from many different sensors across the city and saves them in a file storage system. After that, it continues with Descriptive Analytics and Diagnostic Analytics with this data with the use of an abundance of diagrams that aim to reveal insights from the initial data and help analysts’ work. Finally it takes it further with Predictive Analytics upon the time series and provides nine predictive algorithms in order to predict the value of the time series within a particular future time window. These algorithms consist of statistical algorithms and machine learning algorithms that have been trained to predict the most feasible accurate value for the future. The system also provides the user with an interface that he can change the parameters of these algorithms and test the behavior of the models.

Keywords:

Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Data Analysis, Smart Cities, Statistical Prediction Models, Machine Learning Models, Time Series Prediction

Ευχαριστίες

Μέσα σε αυτό το μακρύ ταξίδι των φοιτητικών ετών μου, το οποίο κλείνει με την παράδοση της παρούσας διπλωματικής, υπήρχαν αρκετοί άνθρωποι που στάθηκαν δίπλα μου και με βοήθησαν να διατηρήσω τους στόχους μου και να εξελιχθώ. Θα ήθελα να ευχαριστήσω καθέναν από αυτούς ξεχωριστά, γιατί ένα ταξίδι αποτελείται από πολλές μικρές κουκίδες που ενώνονται για να φτάσεις στο τέλος αυτού. Οι κουκίδες αυτές είναι κυριώς άνθρωποι και αν δεν υπήρχαν αυτοί, σίγουρα το ταξίδι θα ήταν διαφορετικό.

Πιο συγκεκριμένα θα ήθελα να ευχαριστήσω τους κολλητούς μου φίλους που βρίσκονταν πάντα εκεί και καταλάβαιναν την πίεση και τον χρόνο που μια τέτοια σχολή χρειάζεται να αφιερώσεις. Επίσης την οικογένεια μου που με στήριξαν σε κάθε φάση των φοιτητικών μου χρόνων, με στόχο το καλύτερο για μένα.

Ιδιαίτερες ευχαριστίες όμως θα ήθελα να δώσω και στους τόσους καθηγητές που μου δίδαξαν όχι μόνο ένα μάθημα, αλλά το τρόπο για να γίνεις σωστός άνθρωπος και επιστήμονας. Ευχαριστίες στον κ. Ασκούνη που επέβλεψε τη συγκεκριμένη διπλωματική εργασία και στους κ. Καψάλη και κ. Αλεξάκη που ήταν συνοδοιπόροι μου στο δρόμο της παρούσας διπλωματικής και με βοήθησαν με όλες τις δυσκολίες που βρέθηκαν.

Ξενίας Δημήτριος
Φεβρουάριος 2022

Πίνακας περιεχομένων

Κατάλογος εικόνων

1 Εισαγωγή

1.1 Αφόρμηση

Την τελευταία δεκαετία έχουν εξελιχθεί πάρα πολύ οι διαθέσιμες τεχνολογίες και έχουν ξεπροβάλλει νεες, οι οποίες πλέον μπορούν να υλοποιήσουν σενάρια που παλαιότερα θα φάνταζαν ότι βγήκαν από ταινία επιστημονικής φαντασίας. Η εξέλιξη μάλιστα αυτή παρατηρείται σε ένα τεράστιο φάσμα τεχνολογιών και συνεπώς είμαστε πλέον στην εποχή που μπορούμε να αξιοποιήσουμε τις νέες ανακαλύψεις και να τις συνδυάσουμε για να επιλύσουμε υπαρκτά προβλήματα της ανθρωπότητας, να βελτιώσουμε καθημερινές διαδικασίες, αλλά και να δημιουργήσουμε νέες ανάγκες σε αυτή. Έτσι μπορούμε πλέον να προσεγγίσουμε προβλήματα με νέες μεθόδους επίλυσης και νέα όπλα στη φαρέτρα μας. Και αυτά τα όπλα που αυτή τη στιγμή εξελίσσονται και δείχνουν τη δύναμή τους είναι η Τεχνητή Νοημοσύνη και τα εργαλεία επεξεργασίας δεδομένων Μεγάλης Κλίμακας.

Αφού η επιστήμη φτιάχνεται από τους ανθρώπους για τους ανθρώπους, είναι ανάγκη να χρησιμοποιήσουμε όλα αυτά τα σύγχρονα τεχνολογικά όπλα για λύσουμε προβλήματα που αφορούν πάρα πολύ κόσμο. Ένα από αυτά είναι τα προβλήματα που συναντάμε μέσα στην πόλη που ζούμε και συμβιώνουμε με τους συμπολίτες μας. Συνήθως όμως τέτοια προβλήματα λύνονται και πρέπει να λύνονται από τις τοπικές αρχές και τα τοπικά συμβούλια της κάθε πόλης. Συνεπώς είναι ανάγκη να δημιουργήσουμε εργαλεία και να τα δώσουμε στα χέρια των υπεύθυνων αρχών και τα οποία θα μπορούν να χρησιμοποιηθούν με στόχο τη βελτίωση ζωής στην πόλη ή και την προτροπή έκτακτων συμβάντων.

Τα εργαλεία αυτά μπορεί να είναι είτε εργαλεία που θα δρουν επικουρικά στη λήψη αποφάσεων και στο σχεδιασμό πλάνων δράσης της πολιτείας, είτε και συστήματα που θα έχουν

πιο άμεσο ρόλο. Ο κοινός παρανομαστής όμως των εργαλείων αυτών πρέπει να είναι η χρήση και η αξιοποίηση των σύγχρονων τεχνολογιών, ο οποίες είναι σίγουρο ότι θα δώσουν μια νέα διάσταση στη λήψη αποφάσεων από τις υπεύθυνες αρχές.

1.2 Αντικείμενο Διπλωματικής

Πάνω στο πλαίσιο που αναλύθηκε παραπάνω, σκοπός της παρούσας διπλωματικής είναι η υλοποίηση ενός εργαλείου που απευθύνεται στις υπεύθυνες αρχές μιας πόλης στο πλαίσιο του εκσυγχρονισμού των διαδικασιών λήψης αποφάσεων. Είναι ένα εργαλείο που μπορεί να χρησιμοποιηθεί επικουρικά στη διαδικασία αυτή, αλλά και να παρέχει πληροφορίες χρήσιμες προς τις υπεύθυνες αρχές και οργανισμούς για τη χάραξη σχεδίων δράσης. Συγκεκριμένα είναι ένα εργαλείο που είναι υπεύθυνο αρχικά για τη συλλογή δεδομένων από αισθητήρες που είναι κατανεμημένοι μέσα στον ιστό της πόλης και την αποθήκευσή τους σε ένα σύστημα αρχείων. Ύστερα το σύστημα διαθέτει μια γραφική διεπαφή προς τον χρήστη στην οποία αυτός μπορεί να δει τα δεδομένα που έχουν συλλεχθεί μέσα από μια μεγάλη ποικιλία διαγραμμάτων, που σκοπός τους είναι να αναδείξουν και πιο "κρυμμένες" πληροφορίες των δεδομένων. Τέλος, δίνει τη δυνατότητα στο χρήστη της εφαρμογής να χρησιμοποιήσει 9 διαφορετικούς αλγορίθμους πρόβλεψης, ώστε να λάβει πρόβλεψη για την τιμή των χρονοσειρών του συστήματος μέσα σε ένα συγκεκριμένο ορίζοντα πρόβλεψης. Μάλιστα μπορεί μέσα από το σύστημα να επιλέξει δικές τους παραμέτρους για τα μοντέλα πρόβλεψης και να πειραματιστεί με αυτά, δίχως τη γνώση προγραμματιστικών εργαλείων και γλωσσών προγραμματισμού, μέσα από ένα γραφικό περιβάλλον. Τα μοντέλα αυτά είναι στατιστικά μοντέλα και μοντέλα μηχανικής μάθησης και είναι εκπαιδευμένα να προβλέπουν την τιμή της χρονοσειράς με όσο το δυνατόν πιο ακριβή τρόπο.

Αποτελεί δηλαδή ένα σύστημα που αναλαμβάνει όλη τη διαδικασία ανάλυσης δεδομένων στα πλαίσια μιας έξυπνης πόλης, από την αρχή, που περιλαμβάνει την πρόσληψη δεδομένων από τους αισθητήρες, μέχρι το τέλος, που είναι η παροχή τιμών πρόβλεψης σε χρονοσειρές των δεδομένων προς τον χρήστη του συστήματος. Ο χρήστης μπορεί μέσα από αυτές να προβεί σε κάποια δράση με σκοπό την βελτίωση της ζωής των πολιτών στην πόλη, την αποφυγή έκτακτων γεγονότων μέσα στην πόλη, ακόμα και την μείωση της κυκλοφοριακής συμφόρησης.

1.3 Οργάνωση Κειμένου

1.2.1 Βιβλιογραφία και Επιστημονικό Υπόβαθρο

Αναφέρεται αρχικά η ανάγκη για τη συλλογή και την ανάλυση δεδομένων στο σύγχρονο ψηφιακό κόσμο μέσα σε μια πόλη, μέσα από αναφορές και έρευνες που έχουν διεξαχθεί σε έγκυρες δημοσιεύσεις. Στην πορεία παρατίθεται το επιστημονικό υπόβαθρο και αναλύονται βασικές έννοιες που χρησιμοποιούνται στα στάδια της Περιγραφικής Ανάλυσης, της Διαγνωστικής Ανάλυσης και της Προγνωστικής Ανάλυσης

1.2.2 Εργαλεία

Αναλύονται τα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν καθ' όλη τη διάρκεια ανάπτυξης του συστήματος.

1.2.3 Παρουσίαση Δεδομένων

Παρουσιάζονται τα δεδομένα που χρησιμοποιούνται από το σύστημα και τα οποία προέρχονται από την πόλη Aarhus της Δανίας. Επίσης παρέχεται και το σχήμα των δεδομένων αυτών αλλά και E-R διάγραμμα, το οποίο απεικονίζει την σύνδεση των δεδομένων.

1.2.4 Αρχιτεκτονική Συστήματος

Παρουσιάζεται η Αρχιτεκτονική του Συστήματος και τα εμπλεκόμενα υπο συστήματα, αλλά και το πως αυτά επικοινωνούν μεταξύ τους. Επίσης παρέχονται πληροφορίες για τον τρόπο δημιουργίας κάθε υπο συστήματος, αλλά και το ρόλο του στο συνολικό project.

1.2.5 Ανάλυση Δεδομένων

Αναλύονται τα δεδομένα σύμφωνα με τα βήματα της Περιγραφικής και της Διαγνωστικής Ανάλυσης και παρουσιάζονται τρόποι και τεχνικές κατά τα βήματα αυτά στα υπάρχοντα δεδομένα. Επίσης περιγράφεται αναλυτικά και το βήμα της Προγνωστικής Ανάλυσης, με τους διάφορους αλγορίθμους, και πώς αυτοί χρησιμοποιούν τα δεδομένα για να εκπαιδευτούν και να προβλέψουν.

1.2.6 Οδηγός Εργαλείου

Παρουσιάζεται η Γραφική Διεπαφή Χρήστη με το σύστημα μέσα από στιγμιότυπα του Web App. Επίσης αναλύονται οι περιπτώσεις χρήσης του εργαλείου αυτού.

1.2.7 Αποτελέσματα και Συμπεράσματα

Παρουσιάζονται τα αποτελέσματα των αλγορίθμων πρόβλεψης των χρονοσειρών, καθώς και τα συμπεράσματα ως προς τις ανάγκες ενός τέτοιου συστήματος.

1.2.8 Επόμενα βήματα

Παρουσιάζονται τα επόμενα βήματα που απαιτούνται προκειμένου το αντικείμενο της διπλωματικής να περάσει από το πλαίσιο μιας διπλωματικής και το στάδιο της πειραματικής ανάλυσης, σε ένα πραγματικό σύστημα που καταφέρνει να φέρει εις πέρας τις προσδοκίες από ένα τέτοιο σύστημα.

2

Βιβλιογραφία και Επιστημονικό Υπόβαθρο

2.1 Οι ευκαιρίες και η ανάγκη για την Συλλογή και την Ανάλυση Δεδομένων στα πλαίσια μιας έξυπνης πόλης

Οι ολοένα και περισσότερες αναδυόμενες τεχνολογίες τα τελευταία χρόνια τόσο στα πρωτόκολλα επικοινωνιών μέσω Διαδικτύου (5G) όσο και στα υπολογιστικά συστήματα, έχουν καταστήσει την επικοινωνία μεταξύ διαφορετικών συσκευών πιο εύκολη από ποτέ. Σύμφωνα με έρευνες πάνω από 50 δισεκατομμύρια συσκευές βρίσκονται συνδεδεμένες στο Διαδίκτυο και οι προβλέψεις για το μέλλον δείχνουν τον πολλαπλασιασμό του αριθμού αυτού. Όλες αυτές οι συσκευές στεγάζονται κάτω από τον γνωστό πλέον όρο “Διαδίκτυο των Πραγμάτων” (IoT). Το διαδίκτυο αυτό είναι ένας συνδυασμός από ενσωματωμένα συστήματα που περιλαμβάνουν ενσύρματες και ασύρματες επικοινωνίες, αισθητήρες, και φυσικά αντικείμενα που συνδέονται στο Internet και επικοινωνούν μεταξύ τους [1]. Κάθε τέτοιο σύστημα μπορεί να συλλέγει δεδομένα και να τα διακινεί μέσω του Διαδικτύου. Σκοπός της ύπαρξης ενός τέτοιου δικτύου είναι προφανώς η απλούστευση και ο εμπλουτισμός των ανθρώπινων διαδικασιών και εμπειριών, κάτι που άλλωστε πρέπει να πρεσβεύει η επιστήμη των υπολογιστών [2]. Μέσα σε

αυτό το πλαίσιο, είναι απαραίτητη η δημιουργία συστημάτων που μπορούν να έχουν πρόσβαση σε αυτά τα ακατέργαστα δεδομένα που διακινούνται μέσα από το διαδίκτυο, να τα συλλέγουν και να τα αναλύουν με στόχο την απόκτηση γνώσης, την οποία θα αξιοποιούν για την εκάστοτε κάθε φορά λειτουργία που θέλουν να βελτιώσουν ή δημιουργήσουν. Γίνεται έτσι αντιληπτό ότι η ολοένα και αυξανόμενη ένταξη νέων συσκευών στο Διαδίκτυο των Πραγμάτων και η συλλογή των δεδομένων τους από ένα κεντρικό σύστημα μπορεί να εκτοξεύσει και τις δυνατότητες της Επιστήμης των Δεδομένων, καθώς η τελευταία είναι άρρηκτα συνδεδεμένη με τα δεδομένα, που είναι άλλωστε η πηγή όλων των αναλύσεων. Πόσο μάλλον όταν πλέον δεν μιλάμε απλα για δεδομένα, αλλά για "Μεγάλα" Δεδομένα.

Αυτό ισχύει και στο πλαίσιο μιας έξυπνης πόλης, στην οποία εντάσσονται όλο και περισσότερες συσκευές και αισθητήρες κάτω από την ομπρέλα του Διαδικτύου των Πραγμάτων. Οι πόλεις ανέκαθεν έψαχναν νέους τρόπους για να ενισχύσουν την ποιότητα ζωής και να κάνουν τις ήδη υπάρχουσες υπηρεσίες αποδοτικότερες. Τα τελευταία χρόνια η έννοια των "έξυπνων" πόλεων παίζει ένα σημαντικό ρόλο στην βιομηχανία [3]. Με τον ολοένα και αυξανόμενο αριθμό του πληθυσμού στις αστικές πόλεις, αυτές αναζητούν τρόπους να λύσουν τα γενόμενα προβλήματα της αστικοποίησης. Συνεπώς, η ύπαρξη και η ανάπτυξη του Διαδικτύου των Πραγμάτων γεννά πολλές ευκαιρίες στους διοικούντες των Δήμων για μια ψηφιακή μεταπήδηση, με την οποία θα μπορούν να έχουν άμεσες αναφορές για διάφορα δεδομένα μέσα στην πόλη τους, μέσα από όλα τα "ζωντανά" δεδομένα που συγκεντρώνουν τα πληροφοριακά συστήματα που χρησιμοποιούν. Μέσα από αυτά τα δεδομένα θα μπορούν σε άμεσο χρόνο να αναγνωρίζουν κάποια πιθανή κρίση και να δρουν ανάλογα για την αντιμετώπιση της κατάστασης μέσω ενός σχεδίου δράσης. Έτσι με τη χρήση του Διαδικτύου των Πραγμάτων σε μια πόλη, την μετατρέπουμε σε μια "έξυπνη" πόλη, όπου πολλοί αισθητήρες, κτίρια και δομές επικοινωνούν μεταξύ τους, καθιστώντας πιο εύκολη την παρακολούθηση όλων αυτών των δεδομένων. Προβλήματα που μπορούν να λυθούν έτσι είναι η βελτίωση του κυκλοφοριακού προβλήματος, η αντιμετώπιση φυσικών καταστροφών, η διαχείριση υδάτων, η διαχείριση ενέργειας, η διαχείριση των ρυπαντών και προφανώς η βελτίωση της ποιότητας ζωής των πολιτών [4]. Μάλιστα σύμφωνα με τους συγγραφείς του "Vision of a smart city" [5] μακροπρόθεσμο όραμα είναι η δημιουργία έξυπνων πόλεων που θα έχουν συστήματα και δομές που θα μπορούν να παρακολουθούν από μόνες τους τα "Μεγάλα" Δεδομένα που καταφθάνουν και να δρουν αυτοβούλως εάν αυτό χρειαστεί.

2.2 Τα βήματα της Ανάλυσης των Δεδομένων

Η Επιστήμη των Δεδομένων είναι ένας συνδυασμός από διαφορετικούς επιστημονικούς τομείς που σχετίζονται με την εξόρυξη δεδομένων, την μηχανική μάθηση και άλλες τεχνικές για να βρούμε μοτίβα και να εξάγουμε γνώση από τα δεδομένα [4]. Η πλήρης διαδικασία που ακολουθούμε κατά την ανάλυση δεδομένων αποτελείται από 4 βήματα με τη σειρά: (i) περιγραφική ανάλυση των δεδομένων (τι συμβαίνει), (ii) διαγνωστική ανάλυση δεδομένων (γιατί συμβαίνει), (iii) προγνωστική ανάλυση δεδομένων (τι πρόκειται να συμβεί) και (iv) εντεταλμένη ανάλυση δεδομένων (τι ενέργεια πρέπει να ληφθεί). [6] Η διαδικασία αυτή ή ακόμα και μέρος της ακολουθείται τόσο από επιχειρήσεις για την βελτίωση των επιχειρηματικών τους διαδικασιών και την απόκτηση γνώσης (γνωστό και ως Business Intelligence (BI)), όσο και κατά τη διαδικασία ερευνών πάνω σε μια μεγάλη ποικιλία ερευνητικών θεμάτων, που βασίζονται στην Επιστήμη των Δεδομένων.

2.2.1 Περιγραφική Ανάλυση Δεδομένων

Αποτελεί το πρώτο βήμα και το πιο σημαντικό κατά τη διαδικασία της ανάλυσης δεδομένων καθώς μέσα από αυτό μπορούμε να διαπιστώσουμε το τί συμβαίνει γύρω μας. Η περιγραφική ανάλυση συμπεριλαμβάνει την άμεση παρακολούθηση της συμπεριφοράς ενός στόχου σε φυσικό πλαίσιο, ώστε να συγκεντρώσουμε πληροφορίες για τον στόχο αυτό [7]. Είναι ουσιαστικά η μετατροπή των ακατέργαστων δεδομένων σε μορφή τέτοια που να καθιστά εύκολη την κατανόηση και την ερμηνεία τους, ώστε να έχουμε καλύτερα οργανωμένη την πληροφορία για να την μελετήσουμε. Μας βοηθάει στην περιγραφή, την προβολή των δεδομένων και την περίληψη των δεδομένων με ένα δομημένο τρόπο που καθιστά πιο εύκολη την εξαγωγή “κρυμμένης” πληροφορίας. Μπορούμε να δούμε την κατανομή των δεδομένων, την ανίχνευση “ψευδών” δεδομένων και να παρατηρήσουμε ομοιότητες ή και σχέσεις μεταξύ των δεδομένων μας. Όλα αυτά τα αποτελέσματα θα μας βοηθήσουν για το επόμενο στάδιο της Διαγνωστικής Ανάλυσης και επιτυγχάνονται με τον υπολογισμό ή και την οπτικοποίηση διαφόρων μετρήσεων ενός μεταβαλλόμενου συνήθως στο χρόνο μεγέθους.

2.2.1.1 Σημαντικά Μεγέθη κατα την Περιγραφική Ανάλυση Δεδομένων

Πέρα από το βασικό μέγεθος που αναλύουμε, όπως είπαμε είναι σημαντικό να εξάγουμε και άλλες πληροφορίες για το μέγεθος αυτό, μέσα από τον υπολογισμό διάφορων στατιστικών δεικτών αλλά και χαρακτηριστικών του. Τέτοιοι δείκτες και χαρακτηριστικά είναι [9] [10]:

- *Μέτρηση συχνότητας*

  Στην Περιγραφική Ανάλυση είναι πολύ σημαντικό να γνωρίζουμε πόσο συχνά πραγματοποιείται ένα γεγονός, καθώς μας βοηθάει να καταλάβουμε πώς κινείται το μέγεθος στο χρόνο. Άρα είναι πολύ σημαντικό να οπτικοποιήσουμε και να μετρήσουμε τη συχνότητα.

- *Μέτρηση Κεντρικής Τάσης και Quartile Definition*

  Άλλο ένα σημαντικό μέγεθος είναι η μέτρηση της μέσης τιμής, του μέσου όρου, του Q1 και Q3, της ελάχιστης αλλά και της μέγιστης τιμής. Οι τιμές αυτές υπολογίζονται ως εξής. Έστω ότι έχουμε ένα ταξινομημένο σύνολο δεδομένων τιμών:

  { **59, 60, 65, 65, 68, 69, 70, 72, 75**, 75, **76, 77, 81, 82, 84, 87, 90, 95, 98**}

  - Ο Μέσος Όρος υπολογίζεται με τον γνωστό τύπο:

$$average(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  Ωστόσο κάποιες φορές μεγαλύτερη σημασία έχουν τα υπόλοιπα μεγέθη μέτρησης του συνόλου.

  - Η Μέση Τιμή είναι η τιμή της μέτρησης που χωρίζει το ανώτερο σύνολο τιμών σε 2 υποσύνολα ίδιου μεγέθους (εδώ το 75)
  - Q1 είναι η τιμή που χωρίζει το πρώτο μισό του συνόλου σε αντίστοιχα δυο νέα υποσύνολα ίδιου μεγέθους (εδώ το 68)

❖ Q3 είναι η τιμή που χωρίζει το δεύτερο μισό του συνόλου σε αντίστοιχα δυο νέα υποσύνολα ίδιου μεγέθους (εδώ το **84**)

- *Μέτρηση Τυπικής Απόκλισης και Διακύμανσης*

Η Τυπική Απόκλιση Εκφράζει το βαθμό κατά τον οποίο οι παρατηρήσεις είναι διεσπαρμένες γύρω από τη μέση τιμή και υπολογίζεται από τον τύπο:

$$s = \sqrt{\frac{(x_i - \bar{x})^2}{n-1}}$$

Εικόνα 2.1 Τύπος Τυπικής Απόκλισης

Συνδυάζοντας και τις ανώτερες τιμές, εαν το Q1 έχει απόλυτη διαφορά με τη μέση τιμή μεγαλύτερη από την απόλυτη διαφορά του Q2 με τη μέση τιμή, τότε καταλαβαίνουμε ότι το μέγεθος παρατήρησης έχει μεγαλύτερη διασπορά ως προς τις μικρότερες τιμές αντί για τις μεγαλύτερες.

Αντίστοιχα η διακύμανση ορίζεται ως το τετράγωνο της τυπικής απόκλισης.

- *Ποιοτικά Χαρακτηριστικά Χρονοσειρών [10]*

Όταν το μέγεθος που μελετάμε είναι μια χρονοσειρά, μπορουμε με τη μέθοδο της αποσύνθεσης να απομονώσουμε τις τέσσερις βασικές συνιστώσες της: τάση, κύκλος, εποχικότητα και τυχαιότητα. Οι συνιστώσες αυτές είναι πολύ σημαντικές γιατί μπορούμε να αντλήσουμε κρυμμένες πληροφορίες μέσα από τα δεδομένα μας και να μας βοηθήσει στην κατανόηση και την ερμηνεία των γεγονότων. Η **τάση** αντιπροσωπεύει τη γενική εικόνα της χρονοσειράς και θα μπορούσε να οριστεί σαν μια "μακροπρόθεσμη" μεταβολή του μέσου επιπέδου των τιμών της χρονοσειράς. Η **κυκλικότητα** αντιπροσωπεύει μια "κυματοειδή" μεταβολή που οφείλεται σε ειδικές εξωγενείς συνθήκες και εμφανίζεται κατα περιόδους. Η **εποχικότητα** ορίζεται σαν μια περιοδική διακύμανση που έχει σταθερό και μικρότερο του έτους μήκος. **Ασυνέχειες** ονομάζονται οι

απομονωμένες παρατηρήσεις που εμφανίζονται στο γράφημα κάποιας χρονοσειράς ως απότομες αλλαγές στο πρότυπο συμπεριφοράς της.

- *Στασιμότητα Χρονοσειράς [11] [12]*

  Επειδή κάποια μοντέλα πρόβλεψης χρονοσειρών που θα δούμε στην πορεία της ανάλυσης παίρνουν ως δεδομένο ότι η χρονοσειρά είναι στάσιμη, ένα σημαντικό χαρακτηριστικό ανάλυσης είναι η στασιμότητα της χρονοσειράς. Με λίγα λόγια μια χρονοσειρά είναι στάσιμη όταν δεν παρουσιάζει ούτε τάση ούτε εποχικότητα. Μαθηματικά περιγράφεται ως ότι έχει σταθερό μέσο όρο και σταθερή διακύμανση. Οι μέθοδοι για να αποφανθούμε αν μια χρονοσειρά είναι στάσιμη είναι: μέσω της μελέτης του κινητού μέσου όρου, αλλά συνήθως για πιο έγκυρα αποτελέσματα χρησιμοποιούμε το στατιστικό έλεγχο υποθέσεων ADF Test, το οποίο ουσιαστικά υλοποιεί ένα κλασσικό τεστ μηδενικής υπόθεσης και επιστρέφει μια τιμή **p**. Εάν αυτή η τιμή είναι **μικρότερη από 0.05** τότε αποκλείουμε τη μηδενική υπόθεση και θεωρούμε ότι η χρονοσειρά είναι **στάσιμη**. Σε αντίθετη περίπτωση η χρονοσειρά είναι **μη-στάσιμη**. Εάν δεν είναι στάσιμη μια χρονοσειρά και θέλουμε να χρησιμοποιήσουμε κάποιο μοντέλο πρόβλεψης που απαιτεί αυτή την υπόθεση, τότε μπορούμε να την κάνουμε στάσιμη είτε μέσω της αποσύνθεσης, είτε μέσω της τεχνικής "differencing", δηλαδή της διαφοράς μιας τιμής με την επόμενή της.

2.2.1.2 Οπτικοποίηση Δεδομένων και των χαρακτηριστικών τους

Πέρα από τον υπολογισμό των παραπάνω μεγεθών, είναι απαραίτητη και η οπτικοποίησης τους για την καλύτερη κατανόηση τόσο του μεγέθους παρατήρησης, όσο και της εξαγόμενης πληροφορίας. Αρχικά η γραφική αναπαράσταση των δεδομένων αποτελεί ένα πολύ σημαντικό εργαλείο για την ανάλυση δεδομένων και χρονοσειρών. Η αναπαράσταση έγκειται σε δισδιάστατη γραφική απεικόνιση των πραγματικών τιμών των διαθέσιμων δεδομένων ως προς το χρόνο. Από την αναπαράσταση των δεδομένων καθίστανται εμφανή τα ποιοτικά χαρακτηριστικά της χρονοσειράς (τάση, εποχικότητα, κύκλος, τυχαιότητα, ασυνέχειες) και βοηθούν τον αναλυτή να επιλέξει μεταξύ των εναλλακτικών μεθοδολογιών και

εργαλείων. Επιπλέον η γραφική απεικόνιση των δεδομένων ενδέχεται να αποκαλύψει και ακραίες ή και εσφαλμένες τιμές ακόμα και κρυφά μοτίβα με το μάτι. [8] [10]. Πέρα όμως από την οπτικοποίηση των παραπάνω ακατέργαστων δεδομένων, σημαντικό είναι να οπτικοποιούμε και τις μετρήσεις για τα μεγέθη αυτά που ήδη προαναφέραμε. Χρήσιμα διαγράμματα είναι:

- *Ιστόγραμμα Συχνότητας (ή Distribution Histogram)*

  Είναι η γραφική απεικόνιση στατιστικών συχνοτήτων περιοχών τιμών ενός μεγέθους. Σχηματίζεται από παρακείμενα ορθογώνια. Η επιφάνεια κάθε ορθογωνίου είναι μέτρο συχνότητας εμφάνισης της συγκεκριμένης τιμής περιοχών, ενώ το ύψος ισούται με το λόγο της συχνότητας προς το εύρος των τιμών που αντιπροσωπεύει το ορθογώνιο.

- *Box Plots*

  Διάγραμμα που απεικονίζει τα εξής 5 μεγέθη: ελάχιστη τιμή, Q1, μέση τιμή, Q3, μέγιστη τιμή, με τρόπο που ένας αναλυτής μπορεί πολύ εύκολα να αναγνωρίσει τις τιμές και να τις χρησιμοποιήσει για την περαιτέρω ανάλυση του μεγέθους παρατήρησης.

2.2.2 Διαγνωστική Ανάλυση Δεδομένων

Στο βήμα της Διαγνωστικής Ανάλυσης Δεδομένων προσπαθούμε να καταλάβουμε γιατί συμβαίνουν τα γεγονότα που παρατηρούμε και να εξάγουμε την “κρυμμένη” πληροφορία των δεδομένων μας και ακόμα και να γεννήσουμε νέα πληροφορία. Για παράδειγμα, έστω ότι έχουμε μια χρονοσειρά που περιγράφει τις πωλήσεις παγωτού στην Ελλάδα. Έχοντας κάνει την περιγραφική ανάλυση και έχοντας υπολογίσει σημαντικούς δείκτες της χρονοσειράς και τα ποιοτικά της χαρακτηριστικά, μπορούμε να εξηγήσουμε ότι η περιοδικότητα που παρατηρείται στις πωλήσεις (πιο πολλές το καλοκαίρι και λιγότερες το χειμώνα) οφείλεται στο ότι το καλοκαίρι κάνει πολλή ζέστη και ο κόσμος αγοράζει περισσότερα παγωτά για να δροσιστεί. Ο συλλογισμός αυτός προκύπτει από το γεγονός ότι στην περιγραφική ανάλυση έχουμε εντοπίσει

την περιοδικότητα και ουσιαστικά κατά την διαγνωστική ανάλυση προσπαθούμε να εξηγήσουμε γιατί συμβαίνει αυτό.

Επίσης το βήμα αυτό, ελέγχουμε τις μάλλον "ψεύτικες" τιμές της χρονοσειράς και προσπαθούμε να εξηγήσουμε εάν αυτή η τιμή είναι λογική ή όντως λανθασμένη. Για παράδειγμα, εάν η τιμή μιας χρονοσειράς παίρνει την τιμή 0, είναι μια τιμή υποψήφια ως λανθασμένη και χρήζει ανάγκη περαιτέρω εξέτασης για την εξακρίβωση της εγκυρότητας της τιμής.

Τέλος μια πολύ σημαντική διαδικασία στο βήμα αυτό είναι η μελέτη των συσχετίσεων μεταξύ διαφορετικών μεγεθών παρακολούθησης. Για παράδειγμα, είναι πολύ σημαντικό να ξέρουμε εάν δύο διαφορετικά μεγέθη συσχετίζονται μεταξύ τους. Βασικός δείκτης σε αυτή τη διαδικασία είναι η συνδιακύμανση που ορίζεται ώς:

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

2.2 Τύπος Συνδιακύμανσης

Βασική προϋπόθεση είναι οι δύο αυτές μεταβλητές να είναι διακριτές και τυχαίες. Αντί για την συνδιακύμανση μπορούμε να ελέγξουμε και την συσχέτιση (συνήθως κατά Pearson) ή αλλιώς συντελεστή γραμμικής συσχέτισης των δύο τυχαίων μεταβλητών, που ορίζεται ως:

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

2.3 Τύπος Συντελεστή Γραμμικής Συσχέτισης

Τα μεγέθη αυτά μπορούν να δείξουν τις συσχετίσεις πολλών μεταβλητών μεταξύ τους και να χρησιμοποιήσουμε την πληροφορία αυτή για τα μοντέλα πρόβλεψής μας.

2.2.3 Προγνωστική Ανάλυση Δεδομένων

Το στάδιο της Προγνωστικής Ανάλυσης Δεδομένων περιλαμβάνει πρωτίστως την κατά το δυνατόν ακριβέστερη εκτίμηση της ζητούμενης μεταβλητής στο μέλλον. Η παραγωγή προβλέψεων και ισχυρισμών επιτυγχάνεται με την αξιοποίηση της διαθέσιμης γνώσης και εμπειρίας και αγορά συνηθέστερα, μελλοντικά γεγονότα ή καταστάσεις που δεν έχουν ακόμα

παρατηρηθεί. Ανάλογα με το σκοπό χρήσης των παραγόμενων προβλέψεων, αλλά και τα διαθέσιμα μέσα, η διαδικασία παραγωγής προβλέψεων γίνεται με διαφορετικό τρόπο. [10]

2.2.3.1 Κατηγορίες Προβλέψων

Οι προβλέψεις μπορούν να χωριστούν σε 3 μεγάλες κατηγορίες και μετα σε πολλές υποκατηγορίες. Σύμφωνα με τη βιβλιογραφία [10] οι 3 μεγάλες κατηγορίες είναι η **στατιστική πρόβλεψη** που βασίζεται σε στατιστικές μεθόδους προβλέψεων, η **κριτική πρόβλεψη** που απαιτεί την συμβολή ειδικών και εμπειρογνώμων, και η **πρόβλεψη στόχου ή προϋπολογισμού**, που αφορά τον επιθυμητό στόχο. Εν τέλει, η **τελική πρόβλεψη** προκύπτει από τον γραμμικό συνδυασμό των τριών αυτών κατηγοριών. Στα πλαίσια της διπλωματικής θα επικεντρωθούμε στην στατιστική πρόβλεψη, καθώς μόνο αυτή την πρόβλεψη θα χρησιμοποιήσουμε.

2.2.3.2 Μοντέλα Στατιστικών Προβλέψεων

Οι στατιστικές προβλέψεις αναφέρονται στην εφαρμογή στατιστικών μοντέλων χρονοσειρών ή αιτιοκρατικών μοντέλων επι μιας σειρά δεδομένων με σκοπό την αυτοματοποιημένη και συστηματική παραγωγή προβλέψεων. Οι στατιστικές προβλέψεις είναι άμεσα εφαρμόσιμες και αποδεκτά ακριβείς, αν συνδυαστούν με ένα διάστημα εμπιστοσύνης. Οι στατιστικές μέθοδοι πρόβλεψης μπορούν να χρησιμοποιηθούν από τους διοικούντες οργανισμών μέσω εξειδικευμένων πληροφοριακών συστημάτων προβλέψεων, χωρίς να υπάρχει απαίτηση σε τεχνικές και στατιστικές γνώσεις και μπορούν να παράγουν ένα αποτέλεσμα αρκετά κοντά σε ακρίβεια με την μελλοντική τιμή, πράγμα που καθιστά τους οργανισμούς σε θέση να προβούν σε γρήγορες ενέργειες για την διαχείριση των επερχόμενων καταστάσεων. Πέρα από τα πολλά και σημαντικά πλεονεκτήματα των στατιστικών προβλέψεων, υπάρχουν ορισμένα προβλήματα στην εφαρμογή αυτών. Συγκεκριμένα, η στατιστικές μέθοδοι προβλέψεων προϋποθέτουν ότι το πρότυπο συμπεριφοράς της εκάστοτε χρονοσειράς θα συνεχιστεί στο μέλλον, γεγονός που δεν συμβαίνει πάντα. Επίσης, οι στατιστικές μέθοδοι δε λαμβάνουν υπόψη ειδικά γεγονότα και ενέργειες που ενδέχεται να πραγματοποιηθούν στο μέλλον. Τα μοντέλα στατιστικών προβλέψεων που χρησιμοποιούνται στην υπάρχουσα διπλωματική και αναλύονται σε αυτό το κομμάτι είναι τα εξής:

2.2.3.2.1 Απλή Εκθετική Εξομάλυνση (SES)

Το συγκεκριμένο μοντέλο σταθερού επιπέδου, που αναφέρεται επίσης και ως απλή εκθετική εξομάλυνση, περιγράφεται από τις εξισώσεις:

$$e_t = Y_t - F_t$$

$$S_t = S_{t-1} + a \cdot e_t$$

$$F_{t+1} = S_t$$

όπου *e* δηλώνει σφάλμα (απόκλιση πραγματικής τιμής από πρόβλεψη), *S* δηλώνει το επίπεδο, *F* την πρόβλεψη και *t* τη χρονική περίοδο. Η παράμετρος *a* αποτελεί το συντελεστή εξομάλυνσης της μεθόδου και μπορεί να λάβει τιμές στο διάστημα [0,1].

Ο βέλτιστος συντελεστής εξομάλυνσης καθορίζεται από παράγοντες της χρονοσειράς, αλλά τα υπολογιστικά συστήματα μπορούν να τον υπολογίσουν σχετικά εύκολα.

2.2.3.2.2 Πολλαπλή Γραμμική Παλινδρόμηση (MLR)

Είναι επέκταση της απλής γραμμικής παλινδρόμησης. Η απλή γραμμική παλινδρόμηση υποθέτει την ύπαρξη σχέσης ανάμεσα στη μεταβλητή πρόβλεψης και σε μια άλλη μεταβλητή (ανεξάρτητη μεταβλητή). Επίσης υποθέτει ότι η σχέση αυτή είναι γραμμική. Στην πολλαπλή γραμμική παλινδρόμηση υποθέτουμε ότι η μεταβλητή πρόβλεψης εξαρτάται από περισσότερες μεταβλητές μέσω μια γραμμικής σχέσης που ορίζεται ως:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_k \cdot X_k + e$$

όπου η *Y* εκφράζει την εξαρτημένη μεταβλητή, ενώ οι μεταβλητές X_1, X_2, έως X_κ εκφράζουν τις ανεξάρτητες μεταβλητές. Οι συντελεστές b_k είναι οι σταθερές παράμετροι. Τέλος, το *e* δηλώνει τον τυχαίο παράγοντα.

2.2.3.2.4 ARIMA

Τα ολοκληρωμένα αυτοπαλινδρομικά μοντέλα κινητών μέσων όρων (ARIMA) είναι στοχαστικά μαθηματικά μοντέλα με τα οποία προσπαθούμε να περιγράψουμε τη διαχρονική εξέλιξη κάποιου φυσικού μεγέθους. Δεδομένου ότι για την πλειοψηφία των φυσικών μεγεθών είναι αδύνατη η πλήρης γνώση και καταγραφή όλων των παραγόντων που επηρεάζουν την εξέλιξή τους στο χρόνο, είναι πολύ δύσκολη η διαχρονική περιγραφή περιγραφή του μεγέθους

από ένα ντετερμινιστικό μοντέλο Τα στοχαστικά αυτά μοντέλα εμπεριέχουν τον τυχαίο παράγοντα, τις τιμές του μεγέθους και ίσως κάποιους άλλους στοχαστικούς παράγοντες. Το μοντέλο που προκύπτει είναι ουσιαστικά ένας γραμμικός συνδυασμός των παραπάνω ποσοτήτων. Βασικό στοιχείο του ARIMA που το κάνει να διαφέρει απο το ARMA είναι ότι το μοντέλο ARMA χρησιμοποιείται σε στάσιμες χρονοσειρές. Αντιθέτως το ARIMA ανταπεξέρχεται σε μη-στασιμες διαδικασίες. το οποίο το καθιστά ένα από τα πιο χρησιμοποιημένα μοντέλα πρόβλεψης. Η γενική μορφή του υποδείγματος ARIMA (p,d,q) περιγράφεται από τη παρακάτω σχέση:

$$y'_t = a_1 y'_{t-1} + a_2 y'_{t-2} + \cdots + a_p y'_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_p \varepsilon_{t-p}$$

όπου, p οι παράμετροι της αυτοπαλίνδρομης διαδικασίας, d ο αριθμός των διαφορών προκειμένου η χρονοσειρά να γίνει στάσιμη, q οι παράμετροι της διαδικασίας του κινητού μέσου, $\alpha_1, \ldots, a_p$ οι παράμετροι του υποδείγματος για το AR, $\beta_1, \ldots, \beta_p$ οι παράμετροι του υποδείγματος για το MA με:

$$y'_t = y_t - y_{t-1}$$

2.2.3.2.5 Αυτοπαλινδρομικά διανυσματικά μοντέλα (VAR)

Τα αυτο παλινδρομικά διανυσματικά μοντέλα χρησιμοποιούνται για την πρόβλεψη μιας χρονοσειράς, όταν αυτή εξαρτάται από παραπάνω από μια χρονοσειρές. Η διαφορά με τα αυτοπαλινδρομικά μοντέλα κινητών μέσων όρων (ARIMA) είναι ότι τα τελευταία είναι μονο-κατευθυντικά ενώ τα Var δι-κατευθυντικά. Μαθηματικά αυτό σημαίνει ότι αν πάρουμε ένα απλό μοντέλο AR(p), έχει συνάρτηση πρόβλεψης:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \epsilon_t$$

Εικόνα 2.4 Μαθηματική Σχέση AR(p)

όπου β οι παράμετροι υποδείγματος του AR και ε ο τυχαίος παράγοντας.

Αντίθετα το VAR μοντέλο είναι ένας γραμμικός συνδυασμός των προηγούμενων τιμών τιμών των πολλαπλών μεταβλητών. Μαθηματικά για 2 μεταβλητές που έχουν συσχέτιση μεταξύ τους είναι το εξής:

$$Y_{1,t} = \alpha_1 + \beta_{11,1} Y_{1,t-1} + \beta_{12,1} Y_{2,t-1} + \epsilon_{1,t}$$
$$Y_{2,t} = \alpha_2 + \beta_{21,1} Y_{1,t-1} + \beta_{22,1} Y_{2,t-1} + \epsilon_{2,t}$$

2.5 Μαθηματικές Σχέσεις VAR

Βλέπουμε λοιπόν ότι χρησιμοποιούμε τις δυνατότητες των αυτοπαλινδρομικών μοντέλων και στο μοντέλο μας βάζουμε και περισσότερη γνώση μέσα από την ένταξη περισσότερων μεταβλητών σε αυτό.

2.2.3.3 Μοντέλα Μηχανικής Μάθησης και Βαθειών Νευρωνικών Δικτύων

Ο Arthur Samuel ορίζει ως μηχανική μάθηση «Το πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί». [13] Η μηχανική μάθηση επικεντρώνεται στην ανάπτυξη αλγορίθμων που μπορούν να έχουν πρόσβαση σε δεδομένα και να τα χρησιμοποιούν για να μάθουν και να βελτιώνονται, με στόχο τη καλύτερη δυνατή πρόβλεψη στο μέλλον, μέσα από την αποκτηθείσα εμπειρία. Η μηχανική μάθηση χωρίζεται σε τρεις κύριες κατηγορίες:

- Επιβλεπόμενη Μάθηση (Supervised Learning).

Ο αλγόριθμος εκπαιδεύεται σε ένα σύνολο παραδειγμάτων με ζευγάρια εισόδων και επιθυμητών εξόδων. Μερικοί από τους αλγορίθμους που χρησιμοποιούνται είναι:

- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM)
- Λογιστική Παλινδρόμηση (Logistic Regression)
- Nearest Neighbor
- Naive Bayes
- Δέντρα Αποφάσεων (Decision Trees)
- Νευρωνικά Δίκτυα (Neural Networks)

- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning).

Ο αλγόριθμος εκπαιδεύεται χρησιμοποιώντας πληροφορίες που δεν είναι ούτε ταξινομημένες αλλά ούτε επισημασμένες και επιτρέπουν το αλγόριθμο να ενεργεί χωρίς καθοδήγηση. Αλγόριθμοι για αυτή τη κατηγορία είναι:

• Συσταδοποίηση k-means

• Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)

• Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA)

❖ Ενισχυτική Μάθηση (Reinforcement Learning).

Ο αλγόριθμος μαθαίνει πως να συμπεριφέρεται σε ένα περιβάλλον εκτελώντας ενέργειες και βλέποντας αποτελέσματα.

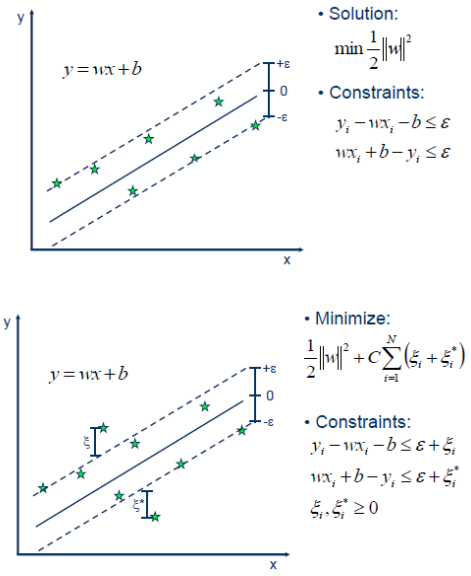
Για την πρόβλεψη χρονοσειρών θα χρησιμοποιήσουμε μόνο την πρώτη κατηγορία, την επιβλεπόμενη μάθηση και μάλιστα με εφαρμογή σε πρόβλημα πρόβλεψης χρονοσειράς, που είναι ένα λεγόμενο regression πρόβλημα. Θα χρησιμοποιήσουμε τους εξής αλγορίθμους που σύμφωνα με αυτή τη πηγή [15] είναι από τους αποδοτικότερους στην πρόβλεψη χρονοσειρών.

2.2.3.3.1 Random Forest

Το Random Forest χρησιμοποιεί στη βάση του Decision Trees, που είναι μια πολύ δημοφιλής κατηγορία στην μηχανική μάθηση. Σε προβλήματα regression, ο Random Forest επιστρέφει ένα μέσο όρο των προβλέψεων των κάθε ξεχωριστών δεντρών, το οποίο αντιμετωπίζει τον πρόβλημα της μεροληψίας των decision trees, όταν αυτά έχουν μεγάλο βάθος. Συνεπώς το Random Forest συνδυάζει την απλούστευση των Decision Trees μαζί με την ευελιξία, οδηγώντας σε καλύτερη ακρίβεια των προβλέψεων.

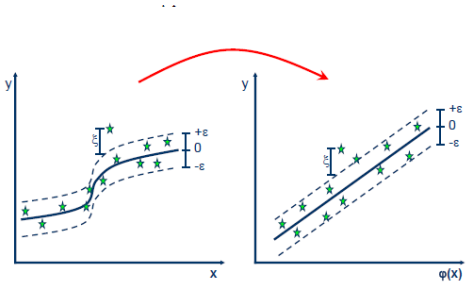
2.2.3.3.2 Support Vector Regression (SVR)

Το SVR είναι μια εξέλιξη του μοντέλου SVM. Το SVM χρησιμοποιείται ευρέως για προβλήματα κατηγοριοποίησης. Σε προβλήματα regression ωστόσο το SVR ακολουθεί την ίδια νοοτροπία της ελαχιστοποίησης του σφάλματος μέσω του ορισμού ενός hyperplane που προσπαθεί να μεγιστοποιήσει τα περιθώρια από τα δεδομένα. Μαθηματικά αυτό εκφράζεται ως εξής:



2.6 Διαγράμματα και Μαθηματικά μοντέλα γραμμικού SVR

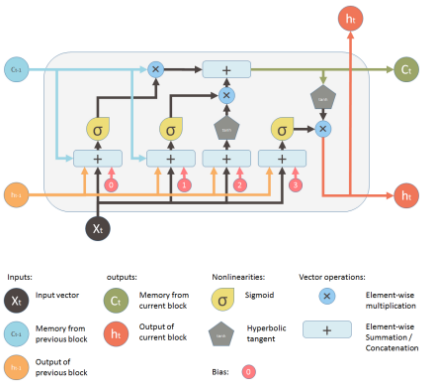
Προσπαθούμε δηλαδή να βρούμε το hyperplane, ώστε τα σημεία της χρονοσειράς να βρίσκονται μέσα στο χώρο που ορίζουν τα όρια που έχουμε θέση σε απόσταση ε από το hyperplane. Το hyperplane μπορεί να είναι ή γραμμικό όπως πάνω, ή και μη γραμμικό όπως παρακάτω:



2.7 Διαγράμματα Μη Γραμμικού SVR

2.2.3.3.3 LSTM

Αποτελεί προέκταση του RNN, ωστόσο λύνει το πρόβλημα κατα το οποίο σε ένα deep RNN δεν μπορεί να κρατηθεί πληροφορία σχετικά με μακροχρόνιες εξαρτήσεις, λόγω προβλημάτων στην ανατροφοδότησή τους. Τα LSTM αποτελούνται από 3 πύλες όπως φαίνεται παρακάτω:

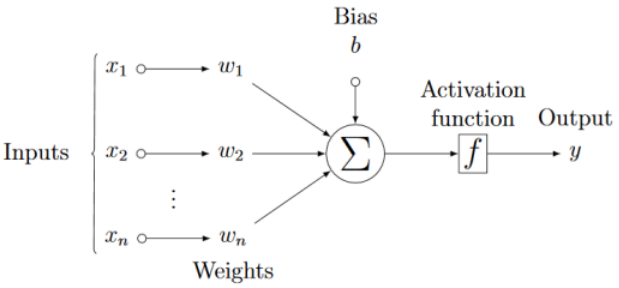


2.8 Εσωτερική Αρχιτεκτονική κυττάρου LSTM

Κάθε LSTM κύτταρο έχει εσωτερικά ένα cell στα οποία διατηρεί την κατάστασή του.[8] Και οι τρεις πύλες του σχήματος δέχονται την ίδια είσοδο από το εξωτερικό του LSTM κυττάρου και παράγουν έξοδό στο διάστημα 0 έως 1, με βάρη που βάρη που μαθαίνουν. Η πρώτη πύλη (input gate) καθορίζει το συντελεστή με τον οποίο θα πολλαπλασιαστεί η εξωτερική είσοδος του LSTM κυττάρου. Η δεύτερη πύλη (forget gate) καθορίζει το συντελεστή με τον οποίο θα πολλαπλασιαστεί η εσωτερική ανάδραση του LSTM κυττάρου. Η επόμενη κατάσταση καθορίζεται από την τρέχουσα κατάσταση και από την ανάδραση. Η Τρίτη πύλη (output gate) καθορίζει το συντελεστή με τον οποίο η εσωτερική κατάσταση του κυττάρου θα περάσει στην έξοδο. Αυτή επομένως μπορεί να απενεργοποιήσει συνολικά το LSTM cell. [18]

2.2.3.3.4 Deep Multiple-Layer Perceptron (MLP)

Η βασική μονάδα υπολογισμού σε ένα MLP είναι ο νευρώνας (Neuron), συχνά επονομαζόμενος και ως κόμβος. [14] Λαμβάνει εισόδους από άλλους κόμβους ή από μια εξωτερική πηγή και υπολογίζει μια έξοδο. Κάθε είσοδος πολλαπλασιάζεται με το αντίστοιχο βάρος (Weight) και υπολογίζεται τα ολικό άθροισμα των γινομένων. Ο κόμβος εφαρμόζει μια συνάρτηση ενεργοποίησης σε αυτό το άθροισμα και υπολογίζεται η έξοδος του νευρώνα. Παρακάτω φαίνεται η αναπαράσταση ενός νευρώνα με τρεις εισόδους καθώς και η εξίσωση της εξόδου.



Εικόνα 2.9: Αναπαράσταση ενός νευρώνα με τρεις εισόδους σε MLP

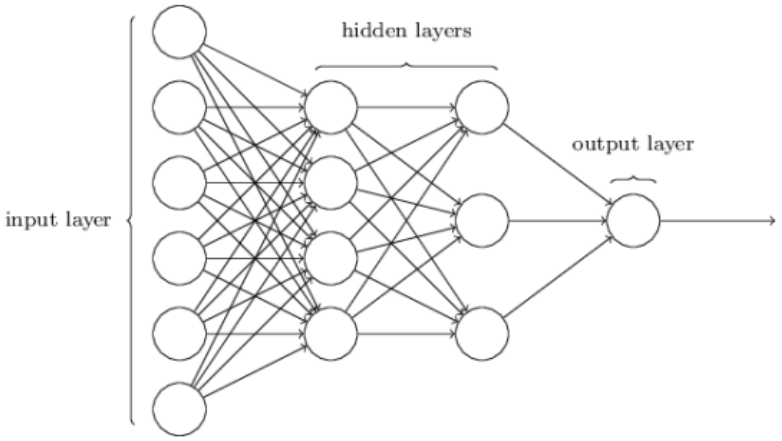
$$y = f(\sum_{i=1}^{n} w_i * x_i + b_i)$$

Εικόνα 2.10: Μαθηματική Σχέση νευρώνα MLP

Το παραπάνω δίκτυο παίρνει ως εισόδους τα X_1, X_2 έως X_n που έχουν για βάρη τα W_1, W_2 έως W_n αντίστοιχα. Επιπλέον υπάρχει ακόμα μια είσοδος 1 με βάρος b η οποία ονομάζεται πόλωση (bias). Η συνάρτηση f είναι μη γραμμική και ονομάζεται συνάρτηση ενεργοποίησης (activation function). Ο σκοπός της συνάρτησης ενεργοποίησης είναι να εισάγει μη γραμμικότητα στην έξοδο ενός νευρώνα. [14]. Αυτό είναι σημαντικό καθώς σχεδόν όλα τα πραγματικά δεδομένα είναι μη γραμμικά. Παραδείγματα τέτοιων συναρτήσεων είναι η σιγμοειδής συνάρτηση, η υπερβολική εφαπτομένη και η ReLU την οποία και θα χρησιμοποιήσουμε στις εφαρμογές αυτής τη εργασίας. Η ReLu αποδίδει συχνά καλύτερα από άλλες συναρτήσεις ενεργοποίησης για κρυφά επίπεδα. Ο βασικός λόγος της αυξημένης

απόδοσης οφείλεται στο γεγονός ότι η ReLU είναι μια γραμμική συνάρτηση μη κορεσμού. Ο κορεσμός είναι το μεγαλύτερο πρόβλημα των δύο προηγούμενων σιγμοειδών συναρτήσεων. Σε αντίθεση λοιπόν με την logistic ή tanh, η ReLU δεν κορέζεται στο -1, 0 ή 1. Οι πιο πρόσφατες έρευνες αναφέρουν ότι τα κρυμμένα επίπεδα του MLP πρέπει να χρησιμοποιούν την ενεργοποίηση του ReLu.

Η βασικότερη μορφή ενός νευρωνικού δικτύου είναι τα Πολυεπίπεδα Perceptrons (Multilayer Perceptrons - MLP). Οι νευρώνες στα MLP είναι οργανωμένοι σε επίπεδα (layers) και δεν υπάρχουν συνδέσεις μεταξύ νευρώνων του ίδιου επιπέδου. Το πρώτο από αυτά τα επίπεδα ονομάζεται επίπεδο εισόδου (input layer) και χρησιμοποιείται για την εισαγωγή των δεδομένων. Τα στοιχεία αυτού του επιπέδου δεν αποτελούν νευρώνες καθώς δεν εκτελούν κάποιον υπολογισμό. Ακολουθούν ένα ή περισσότερα κρυφά επίπεδα (hidden layers) και τέλος υπάρχει το επίπεδο εξόδου (output layer). Στην παρακάτω εικόνα φαίνεται ένα παράδειγμα ενός MLP με 6 εισοδους, 2 κρυφά επίπεδα με 4 και 3 νευρώνες αντίστοιχα για το κάθε επίπεδο και ένα επίπεδο εξόδου.



Εικόνα 2.11: MLP με 6 εισόδους, 2 κρυφά επίπεδα με 4 και 3 νευρώνες αντίστοιχα για το κάθε επίπεδο και ένα επίπεδο εξόδου

Ένα MLP νευρωνικό δίκτυο έχει τις εξής ιδιότητες:

- Feedforward αποκλειστικά, καμία ανάδραση.
- Πλήρως συνδεδεμένα επίπεδα, δηλαδή κάθε νευρώνας ενός επιπέδου συνδέεται με όλους τους νευρώνες του επόμενου. Οι νευρώνες του ίδιου επιπέδου δε συνδέονται μεταξύ τους.

- Μη-γραμμική συνάρτηση ενεργοποίησης (activation function).

Χαρακτηριστικά:

- Ένα Multilayer Perceptron δίκτυο, είναι σε θέση να υλοποιήσει γραμμικές και μη-γραμμικές συναρτήσεις.
- Σύμφωνα με το [17], τα Multilayer Feedforward Networks μπορούν να προσεγγίσουν οποιαδήποτε συνάρτηση με οσηδήποτε ακρίβεια.
- Εκπαίδευση με back-propagation

2.2.3.4 Αξιολόγηση Μοντέλων Πρόβλεψης

Όλοι οι παραπάνω αλγόριθμοι εκπαιδεύονται με βάση μια είσοδο από δεδομένα και σκοπός είναι να μπορούν να προβλέψουν την τιμή των δεδομένων. Στη περίπτωση που τα δεδομένα αυτά είναι χρονοσειρές, στόχος είναι η πρόβλεψη ενός ορίζοντα πρόβλεψης. Τα μοντέλα αυτά όμως χρειάζονται και μια αξιολόγηση για να γνωρίζουμε πόσο ακριβή είναι στις προβλέψεις τους. Για την αξιολόγησή τους χρησιμοποιούνται διάφοροι δείκτες μέτρησης του σφάλματος των μοντέλων. Στα ακόλουθα, y είναι το διάνυσμα της πραγματικής τιμής και f το διάνυσμα της τιμής πρόβλεψης. Το μέγεθος των διανυσμάτων στην περίπτωση των χρονοσειρών είναι ίσο με τον ορίζοντα πρόβλεψης n. Οι συναρτήσεις κόστους είναι οι ακόλουθες: [10]

- Μέσο απόλυτο σφάλμα (MAE)

Εκφράζει ένα μέτρο της ακρίβειας της πρόβλεψης έναντι των πραγματικών τιμών διατηρώντας τις μονάδες μέτρησης της αρχικής χρονοσειράς και έχει τύπο:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |Y_i - F_i|$$

- Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (MAPE)

Ορισμένες φορές είναι πιο χρήσιμος ο υπολογισμός των σφαλμάτων πρόβλεψης σε καθαρά ποσοστιαία μορφή. Αυτό, για παράδειγμα, θα ήταν χρήσιμο, όταν θέλουμε να συγκρίνουμε την ακρίβεια μιας μεθόδου πρόβλεψης που έχει εφαρμοστεί σε παραπάνω από μια χρονοσειρές.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{Yi - Fi}{Yi}\right| \cdot 100\ (\%)$$

- Μέσο Τετραγωνικό Σφάλμα (MSE)

  Όπως και το μέσο απόλυτο σφάλμα, είναι ένα μέτρο ακρίβειας της πρόβλεψης, το οποίο όμως δίνει μεγαλύτερο βάρος στα μεγάλα σφάλματα και μικρότερα βάρος στα μικρότερα σφάλματα.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} |Yi - Fi|^2$$

3 Εργαλεία

Στο συγκεκριμένο κεφάλαιο παρουσιάζονται τα κυριότερα εργαλεία που χρησιμοποιήθηκαν κατά τη διάρκεια υλοποίησης του συστήματος τόσο για το back-end όσο και για το front-end. Ακολουθούν:

3.1 Python Django and Django REST Framework

Για την υλοποίηση του back-end της εφαρμογής που παρουσιάζεται στην παρούσα διπλωματική, χρησιμοποιείται το Django Framework [21]. Σκοπός του, όπως και όλων φυσικά των frameworks είναι η αυτοματοποίηση διαδικασιών που συνδέονται με την ανάπτυξη λογισμικού στο κάθε τομέα ενδιαφέροντος. Έτσι προσφέρεται μεγαλύτερη κλιμακωσιμότητα του λογισμικού, μειώνεται ο χρόνος υλοποίησης βασικών διαδικασιών, προσφέρεται μεγαλύτερη σταθερότητα στο σύστημα και περισσότερη ασφάλεια, ενώ τέλος προσφέρεται καλύτερη ενσωμάτωση με τα υπόλοιπα συστήματα.

Έτσι λοιπόν και το Django είναι ένα πρωτοποριακό framework ανοιχτού κώδικα βασισμένο στη γλώσσα προγραμματισμού Python και ακολουθεί την αρχιτεκτονική model view controller (MVC). Είναι κατάλληλο για την ανάπτυξη back-end servers και μπορεί να διαχειριστεί πολλά και πλούσια δεδομένα με αρκετά μεγάλη ταχύτητα. Επίσης διαθέτει και ένα προαιρετικό admin panel για την CRUD (Create Read Update Delete) λειτουργίες. Έχει χρησιμοποιηθεί σε projects όπως το Instagram, το Pinterest και η Coursera.

Για την δημιουργία ενός RESTful API όμως με τη χρήση Django για το back-end, χρησιμοποιείται ένα άλλο framework επιπλέον, το Django REST Framework [22], το οποίο

επεκτείνει τις δυνατότητες του Django ώστε να ανταπεξέλθει στη δημιουργία ενός API back-end server.

3.2 ReactJS

Η ReactJS [23] είναι μια ανοικτού κώδικα βιβλιοθήκη της Javascript για front-end εφαρμογές, με τη οποία επιτυγχάνεται η δημιουργία user interfaces βασισμένο σε UI components. Δημιουργήθηκε και συντηρείται από την Meta (παλία Facebook) και από μια μεγάλη κοινότητα προγραμματιστών. Μπορεί να χρησιμοποιηθεί για την δημιουργία single-page web applications, μέχρι και σε πιο σύνθετες εφαρμογές με τη χρήση επιπλέον βιβλιοθηκών για το routing.

Στην βάση της η React δημιουργεί ένα Virtual DOM στην μνήμη της και σε αυτό κάνει όλες τις απαραίτητες αλλαγές. Στο τέλος συγκρίνει το Virtual DOM με το προηγουμενο DOM του περιηγητή και κάνει όσες αλλαγές χρειάζεται για να ανανεώσει το UI. Βασικό στοιχείο της React είναι τα Components, τα οποία ουσιαστικά είναι αυτά που επιστρέφουν HTML στο τέλος προς τον browser.

3.1 Material UI

Το Material UI [24] είναι μια βιβλιοθήκη που επιτρέπει να εισάγουμε διαφορετικά Components για να δημιουργήσουμε το user interface. Για παράδειγμα μας δίνει τη δυνατότητα να έχουμε έτοιμες φόρμες εισαγωγής δεδομένων όσων αφορά το styling και τα βασικά elements αυτών, γλιτώνοντας φυσικά χρόνο στους προγραμματιστές. Το Material UI είναι εμπνευσμένο από τις σχεδιαστικές αρχές της Google στη κατασκευή user interfaces.

3.3 amCharts 4

Το amCharts [25] είναι μια βιβλιοθήκη της Javascript με την οποία επιτυγχάνεται η απεικόνιση διαδραστικών γραφημάτων στο front-end της εφαρμογής.

3.4 Mapbox

Το Mapbox [26] είναι μια βιβλιοθήκη της Javascript για την χρήση διαδραστικών χαρτών στο front-end της εφαρμογής

3.5 OpenAPI Specification

Το OpenAPI Specification (OAS) [27] ορίζει έναν standard τρόπο διεπαφής των RESTful APIs, το οποίο επιτρέπει τόσο τους ανθρώπους όσο και τις μηχανές να εξερευνούν και να καταλαβαίνουν τις δυνατότητες του συστήματος χωρίς να έχουν πρόσβαση στον κώδικα ή το documentation. Αν έχει οριστεί σωστά, ο consumer μπορεί να καταλάβει και να αλληλεπιδράσει με μια απομακρυσμένη υπηρεσία με την ελάχιστη γνώση της υλοποίησης του.

3.6 Keras & Tensorflow

Για την δημιουργία της αρχιτεκτονικής των μοντέλων μηχανικής μάθησης που χρησιμοποιούνται στη παρούσα διπλωματική, βασικό εργαλείο είναι η πλατφόρμα Tensorflow μαζί με το API της Tensorflow, το Keras. Συγκεκριμένα, το Tensorflow [29] είναι μια ανοικτού κώδικα πλατφόρμα μηχανικής μάθησης που επιβλέπει και διαχειρίζεται από την αρχή μέχρι το τέλος τους πόρους για να μπορούν να εκτελούνται οι αλγόριθμοι μηχανικής μάθησης. Συγκεκριμένα, βασικός ρόλος του tensorflow είναι η αποδοτική εκτέλεση low-level εντολών μεταξύ τένσορων σε CPU, GPU ή TPU. Επίσης, αναλαμβάνει τον υπολογισμό παραγώγων διάφορων ολοκληρώσιμων εκφράσεων. Τέλος, είναι υπεύθυνο για τον κλιμακώσιμο υπολογισμό σε πολλές συσκευές, όπως clusters από εκατοντάδες GPUs και επίσης υλοποιεί την εξαγωγή γράφων σε εξωτερικα συστήματα, όπως servers, browsers κτλπ.

Το Keras [30] από την άλλη είναι ένα high-level API του Tensorflow, το οποίο είναι μια πάρα πολύ διαδεδομένη και χρησιμοποιημένη διεπαφή για την λύση προβλημάτων μηχανικής μάθησης, με στόχο το σύγχρονο deep learning. Διαθέτει τις απαραίτητες αφαιρέσεις και διάφορα building blocks για την ανάπτυξη λύσεων μηχανικής μάθησης και την άμεση διάθεσή τους σε παραγωγικά περιβάλλοντα με πολύ γρήγορη ταχύτητα. Βασικά εργαλεία του Keras που χρησιμοποιούνται στη διπλωματική είναι το LSTM, το Dense και το Sequential, τα οποία συνεργάζονται για την κατασκευή της αρχιτεκτονικής του μοντέλου και την εκπαίδευση αυτού.

3.7 Scikit-learn

Το Scikit-learn [31] είναι η πιο διαδεδομένη βιβλιοθήκη της Python για επίλυση προβλημάτων με μηχανική μάθηση. Περιέχει πολλά αποδοτικά εργαλεία για μηχανική μάθηση και στατιστικά μοντέλα σε προβλήματα που αφορούν την κατηγοριοποίηση, παλινδρομήσεις,

ομαδοποίηση και μείωση διαστάσεων. Βασικοί αλγόριθμοι που χρησιμοποιούνται στη διπλωματική αυτή είναι ο random forest, support-vector regression (SVR), linear model καθώς και διάφοροι scalers. Τέλος, από τη βιβλιοθήκη αυτή χρησιμοποιούνται και συναρτήσεις υπολογισμού κόστους οπως το mean absolute error κτλπ.

3.8 Statsmodels

Το Statsmodels [32] είναι μια βιβλιοθήκη της Python που παρέχει συναρτήσεις για τον υπολογισμό πολλών στατιστικών μοντέλων αλλά και statistical tests Στην παρούσα διπλωματική χρησιμοποιούνται συναρτήσεις όπως adfuller, SimpleExpSmoothing, ARIMA,VAR για τις στατιστικές μας προβλέψεις.

3.9 Prophet

Το Prophet [33] είναι μια διαδικασία πρόβλεψης, υλοποιημένη σε R και Python. Είναι γρήγορο και παρέχει αυτοματοποιημένες προβλέψεις για χρονοσειρές σε μοντέλα στα οποία η μη- γραμμική τάση συνδυάζεται με εποχικότητα. Είναι ένα εργαλείο που έχει αναπτυχθεί από την Meta (πρώην Facebook).

4 Παρουσίαση Δεδομένων

Όπως αναφέραμε και στο δεύτερο κεφάλαιο, για να χαρακτηρίσουμε μια πόλη ως "έξυπνη" πόλη, χρειάζονται δεδομένα από πολλές περιοχές ενδιαφέροντος, ώστε να μπορέσουν αυτά να συνδυαστούν για να προσφέρουν μια ολοκληρωμένη πληροφορία. Ωστόσο, η συλλογή πολλών και διαφορετικών δεδομένων μέσα στην πόλη καθίσταται δύσκολη σήμερα, λόγω της έλλειψης ανοιχτών δεδομένων που βρίσκονται στον παγκόσμιο ιστό. Βέβαια, ήδη πολλές πόλεις και δήμοι κάνουν προσπάθειες για την παροχή τέτοιων δεδομένων στο κοινό στα πλαίσια της ψηφιακής μεταπήδησης. Επομένως φαντάζει πολύ κοντινό τέτοιες εφαρμογές που θα αποσκοπούν στον εκσυγχρονισμό των πόλεων να αρχίζουν να ανθίζουν στο βραχυπρόθεσμο μέλλον και να εφαρμόζονται με αποτελεσματικότητα.

Προκειμένου να εφαρμόσουμε τα παραπάνω στάδια ανάλυσης και πρόβλεψης δεδομένων στα πλαίσια μιας έξυπνης πόλη αλλά και για την ανάπτυξη του εργαλείου, χρειαζόμαστε κάποια πραγματικά δεδομένα που προέρχονται από πολλές πηγές μέσα από την πόλη. Τα δεδομένα που θα χρησιμοποιηθούν επομένως στην παρούσα διπλωματική για την υλοποίηση των απαιτούμενων διαδικασιών προέρχονται από την δεύτερη μεγαλύτερη πόλη της Δανίας, το Aarhus. Τα δεδομένα αυτά διατίθενται ως ανοιχτά δεδομένα (open data) από τον Δήμο Aarhus μέσα από την πλατφόρμα Open Data DK (ODDK) [20] και δίνει την δυνατότητα σε εταιρείες και προγραμματιστές να αναπτύξουν τις καινοτόμες ιδέες τους για την βελτίωση της ποιότητας ζωής των πολιτών αλλά και τον εξορθολογισμό των αστικών λειτουργιών.

Ακολουθεί η παρουσίαση των δεδομένων και του σχήματος αυτών:

4.1 Μετεωρολογικά Δεδομένα για την πόλη Aarhus

Τα μετεωρολογικά δεδομένα είναι πάρα πολύ χρήσιμα για μια πόλη. Προφανώς δεν θα χρησιμοποιηθούν από το σύστημά μας για κάποια πρόβλεψη καιρού, μιας και μια άλλη επιστήμη ασχολείται με πλήρη ακρίβεια και επιτυχία για πολλά χρόνια. Ωστόσο υπάρχουν πολλοί τρόποι με τους οποίους μπορούμε να αξιοποιήσουμε τα δεδομένα καιρού και οι οποίοι παρουσιάζονται στην επόμενη ενότητα.

Dew Point Dataset (4,310 records)			
Περιγραφή	Περιέχει τις μετρήσεις του Dew point καθημερινά με μετρήσεις ανά μισή περίπου ώρα.		
Χρονική Διάρκεια	Αυγ – Σεπτ 2014		
Πεδία Δεδομένων (Σχήμα)	**Όνομα Πεδίου**	**Τύπος**	**Περιγραφή**
	Datetime	*DateTime*	Ημερομηνία και ώρα εγγραφής
	Value	*Number*	Τιμή (σε βαθμούς Κελσίου)

Humidity Dataset (4,310 records)	
Περιγραφή	Περιέχει τις μετρήσεις της Υγρασίας καθημερινά με μετρήσεις ανά μισή περίπου ώρα.

Χρονική Διάρκεια	Αυγ – Σεπτ 2014

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	Τύπος	Περιγραφή
Datetime	*DateTime*	Ημερομηνία και ώρα εγγραφής
Value	*Number*	Τιμή (επί τις %)

Pressure Dataset (4,310 records)

Περιγραφή	Περιέχει τις μετρήσεις της Πίεσης καθημερινά με μετρήσεις ανά μισή περίπου ώρα.
Χρονική Διάρκεια	Αυγ – Σεπτ 2014

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	Τύπος	Περιγραφή
Datetime	*DateTime*	Ημερομηνία και ώρα εγγραφής
Value	*Number*	Τιμή (σε mBar)

Temperature Dataset (4,310 records)

Περιγραφή	Περιέχει τις μετρήσεις της Θερμοκρασίας καθημερινά με μετρήσεις ανά μισή περίπου ώρα.
Χρονική Διάρκεια	Αυγ – Σεπτ 2014

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	Τύπος	Περιγραφή
Datetime	*DateTime*	Ημερομηνία και ώρα εγγραφής
Value	*Number*	Τιμή (σε βαθμούς Κελσίου)

Wind Direction Dataset (4,310 records)

Περιγραφή	Περιέχει τις μετρήσεις της Κατεύθυνσης τους Αέρα καθημερινά με μετρήσεις ανά μισή περίπου ώρα.
Χρονική Διάρκεια	Αυγ – Σεπτ 2014

Πεδία Δεδομένων

Όνομα Πεδίου	Τύπος	Περιγραφή

(Σχήμα)	**Datetime**	*DateTime*	Ημερομηνία και ώρα εγγραφής
	Value	*Number*	Τιμή (σε μοίρες)

Wind Speed Dataset (4,310 records)	
Περιγραφή	Περιέχει τις μετρήσεις της Ταχύτητας του Αέρα καθημερινά με μετρήσεις ανά μισή περίπου ώρα.
Χρονική Διάρκεια	Αυγ – Σεπτ 2014

	Όνομα Πεδίου	**Τύπος**	**Περιγραφή**
Πεδία Δεδομένων (Σχήμα)	**Datetime**	*DateTime*	Ημερομηνία και ώρα εγγραφής
	Value	*Number*	Τιμή (σε kph)

4.2 Δεδομένα των χώρων στάθμευσης του Aarhus

Στο συγκεκριμένο πεδίο παρουσιάζονται δεδομένα από 8 χώρους στάθμευσης που βρίσκονται μέσα στην πόλη του Aarhus. Ένα σημαντικό πρόβλημα που αντιμετωπίζουν οι πολίτες που μετακινούνται μέσα στις σύγχρονες πόλεις με τα αυτοκίνητά τους είναι το θέμα της στάθμευσης. Ο δήμος οφείλει να βρίσκει λύσεις σε αυτό το πρόβλημα των πολιτών και σίγουρα δεδομένα σχετικά με τους χώρους στάθμευσης είναι πολύ σημαντικά για την κατανόηση του προβλήματος.

Aarhus_Parking_Address Dataset (8 records)	
Περιγραφή	Περίεχει πληροφορίες για 8 χώρους στάθμευσης που υπάρχουν στην πόλη.
Χρονική Διάρκεια	-

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	Τύπος	Περιγραφή
garagecode	*String*	Το κωδικός-όνομα του χώρου στάθμευσης
city	*String*	Η πόλη στην οποία ανήκει ο χώρος στάθμευσης
postalcode	*Number*	Ο ταχυδρομικός κωδικός
street	*String*	Η οδός του χώρου

housenumber	*Number*	Το νούμερο διεύθυνσης του χώρου
latitude	*Number*	Γεωγραφικό πλάτος της διεύθυνσης του χώρου
longtitude	*Number*	Γεωγραφικό μήκος της διεύθυνσης του χώρου

Aarhus_Parking_info (55,2655 records)

Περιγραφή	Περιέχει την πληρότητα των άνωθεν χώρων στάθμευσης. Οι εγγραφές αφορούν μετρήσεις καθημερινές και ανά μισή ώρα.
Χρονική Διάρκεια	22 Μαΐου – 4 Νοεμβρίου 2014

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	**Τύπος**	**Περιγραφή**
vehiclecount	*Number*	Ο αριθμός των σταθμευμένων οχημάτων
updatetime	*DateTime*	Η ημερομηνία και ώρα μέτρησης της εγγραφής
totalspaces	*Number*	Ο συνολικό αριθμός χωρητικότητας του χώρου στάθμευσης
garagecode*	*String*	Το κωδικός-όνομα του χώρου στάθμευσης

*Πεδία με κοινό όνομα αναφέρονται στην ίδια εγγραφή και είναι ουσιαστικά το κλειδί σύνδεσης των δεδομένων.

4.3 Δεδομένα Κίνησης στους Δρόμους

Ένα από τα βασικότερα προβλήματα που καλείται ένας Δήμος να αντιμετωπίσει είναι η και κυκλοφοριακή συμφόρηση. Είναι μια πρόκληση που εάν αντιμετωπιστεί θα μπορέσει και να βελτιώσει τη μετακίνηση μέσα στη πόλη αλλά και να περιορίσει τους ρύπους από τα τόσα πολλά αυτοκίνητα που κυκλοφορούν. Για τον σκοπό αυτό χρησιμοποιούνται δεδομένα από 449 διαφορετικούς αισθητήρες σε διαφορετικά σημεία παρατήρησης μέσα στην πόλη του Aarhus.

Traffic_Aarhus Dataset (7,184,000 records)

Περιγραφή	Περιέχει διάφορες μετρήσεις σχετικά με την κίνηση στους δρόμους. Οι μετρήσεις αυτές γίνονται μεταξύ δυο σημείων παρατήρησης στα οποία υπάρχουν αισθητήρες και ουσιαστικά μετράνε την κίνηση σε ένα κομμάτι δρόμου που ορίζεται από τα δυο σημεία παρατήρησης κάθε φορά. Οι μετρήσεις αυτές γίνονται καθημερινώς και ανά 5 λεπτά.
Χρονική Διάρκεια	Αυγ – Οκτ 2014

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	Τύπος	Περιγραφή
timestamp	*DateTime*	Ημερομηνία και ώρα μέτρησης
avgMeasuredTime	*Number*	Χρονική διάρκεια που κράτησε η μέτρηση από το ένα σημείο στο άλλο

avgSpeed	*Number*	Μέση τιμή ταχύτητας των αυτοκινήτων που καταγράφηκαν
vehicleCount	*Number*	Συνολικός αριθμός αυτοκινήτων που καταγράφηκαν μεταξύ των 2 σημείων
Report_ID	*Number*	Το id του συγκεκριμένου τμήματος δρόμου που ορίζεται από τα δυο σημεία παρατήρησης

Traffic_Aarhus_Meta Dataset (450 records)

Περιγραφή	Περιέχει πληροφορίες για τα τμήματα δρόμου στα οποία διαθέτουμε τις άνωθεν πληροφορίες.
Χρονική Διάρκεια	-

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	Τύπος	Περιγραφή
point_1_street	*String*	Η οδός του πρώτου σημείου
point_1_city	*String*	Η πόλη που βρίσκεται το πρώτο σημείο
point_1_street_number	*Number*	Ο αριθμός διεύθυνσης του πρώτου σημείου
point_1_lat	*Number*	Το γεωγραφικό πλάτος του πρώτου σημείου

point_1_lng	*Number*	Το γεωγραφικό μήκος του πρώτου σημείου
point_1_postal_code	*Number*	Ο Ταχυδρομικό Κωδικός του πρώτου σημείου
point_1_country	*String*	Η χώρα που βρίσκεται το πρώτο σημείο
point_2_street	*String*	Η οδός του δεύτερου σημείου
point_2_city	*String*	Η πόλη που βρίσκεται το δεύτερο σημείο
point_2_street_number	*Number*	Ο αριθμός διεύθυνσης του δεύτερου σημείου
point_2_lat	*Number*	Το γεωγραφικό πλάτος του δεύτερου σημείου
point_2_lng	*Number*	Το γεωγραφικό μήκος του δεύτερου σημείου
point_1_postal_code	*Number*	Ο Ταχυδρομικό Κωδικός του δεύτερου σημείου
point_1_country	*String*	Η χώρα που βρίσκεται το δεύτερο σημείο
road_type	*String*	Το είδος του δρόμου που χαρακτηρίζεται από τα 2 σημεία
Report_ID*	*Number*	Το id του συγκεκριμένου τμήματος δρόμου που ορίζεται από τα δυο σημεία παρατήρησης

distance_in_meters	*Number*	Η απόσταση μεταξύ των 2 σημείων

*Πεδία με κοινό όνομα αναφέρονται στην ίδια εγγραφή και είναι ουσιαστικά το κλειδί σύνδεσης των δεδομένων.

4.4 Δεδομένα αισθητήρων στο κτίριο DOKK1

Μία "έξυπνη" πόλη πρέπει να διαθέτει και "έξυπνα" δημόσια κτίρια. Τα κτίρια αυτά πρέπει να διαθέτουν αισθητήρες σε διάφορα δωμάτια και ανάλογα με τις μετρήσεις να αυτοματοποιούν κάποιες διαδικασίες. Στο Aarhus ένα τέτοιο κτίριο είναι το DOKK1.

Το DOKK1 είναι ένα δημόσιο κτίριο που λειτουργεί ως κέντρο πολιτισμού και δημόσια βιβλιοθήκη. Πλήθος κόσμου το επισκέπτονται καθημερινά και σίγουρα η προσπάθεια για τον εκσυγχρονισμό του θα είναι ένα θετικό στοιχείο για τους πολίτες.

DOKK1_Sensors Dataset (412,498 records)

Περιγραφή	Περιέχει μετρήσεις από 37 αισθητήρες που βρίσκονται μέσα στο κτίριο DOKK1 για διάφορα μεγέθη ο οποίες καταγράφονται καθημερινά ανά μισή περίπου ώρα
Χρονική Διάρκεια	4 Μαρτ -19 Σεπτ 2020

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	Τύπος	Περιγραφή
date	*DateTime*	Ημερομηνία και ώρα μέτρησης
sensor	*Number*	Το id του αισθητήρα

temperature	*Number*	Θερμοκρασία (σε Κελσίου)
humidity	*Number*	Υγρασία
co2	*Number*	Διοξείδιο του Άνθρακα
voc	*Number*	Οργανικά στοιχεία
light_level	*Number*	Επίπεδα φωτισμού
sound	*Number*	Επίπεδα ήχου
occupancy	*Number*	0: Ο αισθητήρας δεν ανίχνευσε κίνηση 1: Ανιχνεύτηκε κίνηση

DOKK1_Sensors_meta Dataset (37 records)

Περιγραφή	Περιέχει πληροφορίες για τους 37 αισθητήρες που υπάρχουν μέσα στο κτίριο DOKK1
Χρονική Διάρκεια	-

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	**Τύπος**	**Περιγραφή**
sensor*	*Number*	Το id του αισθητήρα
location	*String*	Defualt: DOKK1

room	*String*	Το δωμάτιο του DOKK1 που βρίσκεται ο αισθητήρας
latitude	*Number*	Γεωγραφικό πλάτος που βρίσκεται ο αισθητήρας
longtitude	*Number*	Γεωγραφικό μήκος που βρίσκεται ο αισθητήρας

*Πεδία με κοινό όνομα αναφέρονται στην ίδια εγγραφή και είναι ουσιαστικά το κλειδί σύνδεσης των δεδομένων.

4.5 Δεδομένα για τα πολιτιστικά δρώμενα της πόλης

Τα πολιτιστικά δρώμενα που πραγματοποιούνται είναι μέρος της ζωής των κατοίκων μιας πόλης. Τέτοια δρώμενα προκαλούν μεγάλη συνάθροιση των πολιτών και η γνώση μιας τέτοιας πληροφορίας είναι σημαντική για την διαχείριση πιθανών κρίσεων αλλά και προβλημάτων που ένα τέτοιο γεγονός μπορεί να προκαλέσει.

Cultural_Events Dataset (100 records)			
Περιγραφή	Περιέχει πολιτιστικές εκδηλώσεις που πραγματοποιούνται στο Aarhus		
Χρονική Διάρκεια	5 Μαΐου 2014 – 25 Ιαν 2015		
Πεδία Δεδομένων (Σχήμα)	**Όνομα Πεδίου**	**Τύπος**	**Περιγραφή**
	title	*String*	Τίτλος της εκδήλωσης
	price	*String*	Εύρος τιμής εισιτηρίων

DateTime	*Datetime*	Ημερομηνία και ώρα της εκδήλωσης
city	*String*	Η πόλη που διοργανώνεται η εκδήλωση
latitude	*Number*	Γεωγραφικό πλάτος του χώρου που πραγματοποιείται η εκδήλωση
longitude	*Number*	Γεωγραφικό μήκος του χώρου που πραγματοποιείται η εκδήλωση
type_of_event	*String*	Ο τύπος της εκδήλωσης
genre	*String*	Πιο συγκεκριμένο είδος εκδήλωσης

Library_Events Dataset (1,549 records)			
Περιγραφή	Περιέχει πολιτιστικές εκδηλώσεις που πραγματοποιούνται από τις βιβλιοθήκες του Aarhus		
Χρονική Διάρκεια	10 Οκτ 2013 – 6 Ιουν 2015		
Πεδία Δεδομένων (Σχήμα)	**Όνομα Πεδίου**	**Τύπος**	**Περιγραφή**
	title	*String*	Τίτλος της εκδήλωσης

price	*String*	Εύρος τιμής εισιτηρίων
start_time	*Datetime*	Ημερομηνία και ώρα έναρξης της εκδήλωσης
end_time	*Datetime*	Ημερομηνία και ώρα λήξης της εκδήλωσης
city	*String*	Η πόλη που διοργανώνεται η εκδήλωση
latitude	*Number*	Γεωγραφικό πλάτος του χώρου που πραγματοποιείται η εκδήλωση
longtitude	*Number*	Γεωγραφικό μήκος του χώρου που πραγματοποιείται η εκδήλωση
library	*String*	Η βιβλιοθήκη που διοργανώνει την εκδήλωση

4.6 Δημογραφικά Χαρακτηριστικά

Τα δημογραφικά χαρακτηριστικά των πολιτών μιας πόλης είναι σίγουρα σημαντικά για την κατανόηση της σύνθεσης των πολιτών και την καλύτερη κατανόηση των συμπεριφορών και αναγκών τους.

Demographics Dataset (28 records)	
Περιγραφή	Περιέχει τα δημογραφικά χαρακτηριστικά των πολιτών του Aarhus

Χρονική Διάρκεια	Ιανουάριος 2013

Πεδία Δεδομένων (Σχήμα)

Όνομα Πεδίου	Τύπος	Περιγραφή
local_community	*String*	Η περιοχή του Aarhus
women	*Number*	Το πλήθος των γυναικών
men	*Number*	Το πλήθος των αντρών
total	*Number*	Το συνολικό πλήθος
0-2 yr	*Number*	Πλήθος 0-2 χρονών
3-5 yr	*Number*	Πλήθος 3-5 χρονών
6-15 yr	*Number*	Πλήθος 6-15 χρονών
16-19 yr	*Number*	Πλήθος 16-19 χρονών
20-24 yr	*Number*	Πλήθος 20-24 χρονών
25-64 yr	*Number*	Πλήθος 25-64 χρονών
65 yr -	*Number*	Πλήθος 65- χρονών

4.7 Ατμοσφαιρική Ρύπανση

Ένα εξίσου σημαντικό πρόβλημα είναι η ρύπανση της ατμόσφαιρας, το οποίο είναι ένα φαινόμενο που παρατηρείται ως επί το πλείστον πάνω από μεγάλα αστικά κέντρα. Εν έτη 2022 μάλιστα είναι πολλές οι ενέργειες για την προστασία του περιβάλλοντος και την υιοθέτηση νέων τρόπων λειτουργίας των μεγάλων αστικών πόλεων με γνώμονα την μείωση της ρύπανσης

του περιβάλλοντος. Έτσι είναι σημαντικό να γνωρίζουμε τους δείκτες από διάφορους παράγοντες που ρυπαίνουν την ατμόσφαιρα και να δράσουμε για την προστασία του περιβάλλοντος αλλά και της υγείας των πολιτών μας, αφού είναι άρρηκτα συνδεδεμένα αυτά τα δύο.

Στο συγκεκριμένο πεδίο έχουμε δεδομένα για διάφορους ρυπαντές του ατμοσφαιρικού αέρα. Τέτοιο ρυπαντές είναι το όζον, η σωματιδιακή ύλη που βρίσκεται στον αέρα, το μονοξείδιο του άνθρακα κ.ά. Τα μεγέθη αυτά μετριούνται σε κλίμα Air Quality Index και η ερμηνεία των τιμών φαίνεται στον παρακάτω πίνακα:

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
When the AQI is in this range:	*..air quality conditions are:*	*...as symbolized by this color:*
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Εικονα 4.1 Πίνακας Αντιστοιχίας AQI

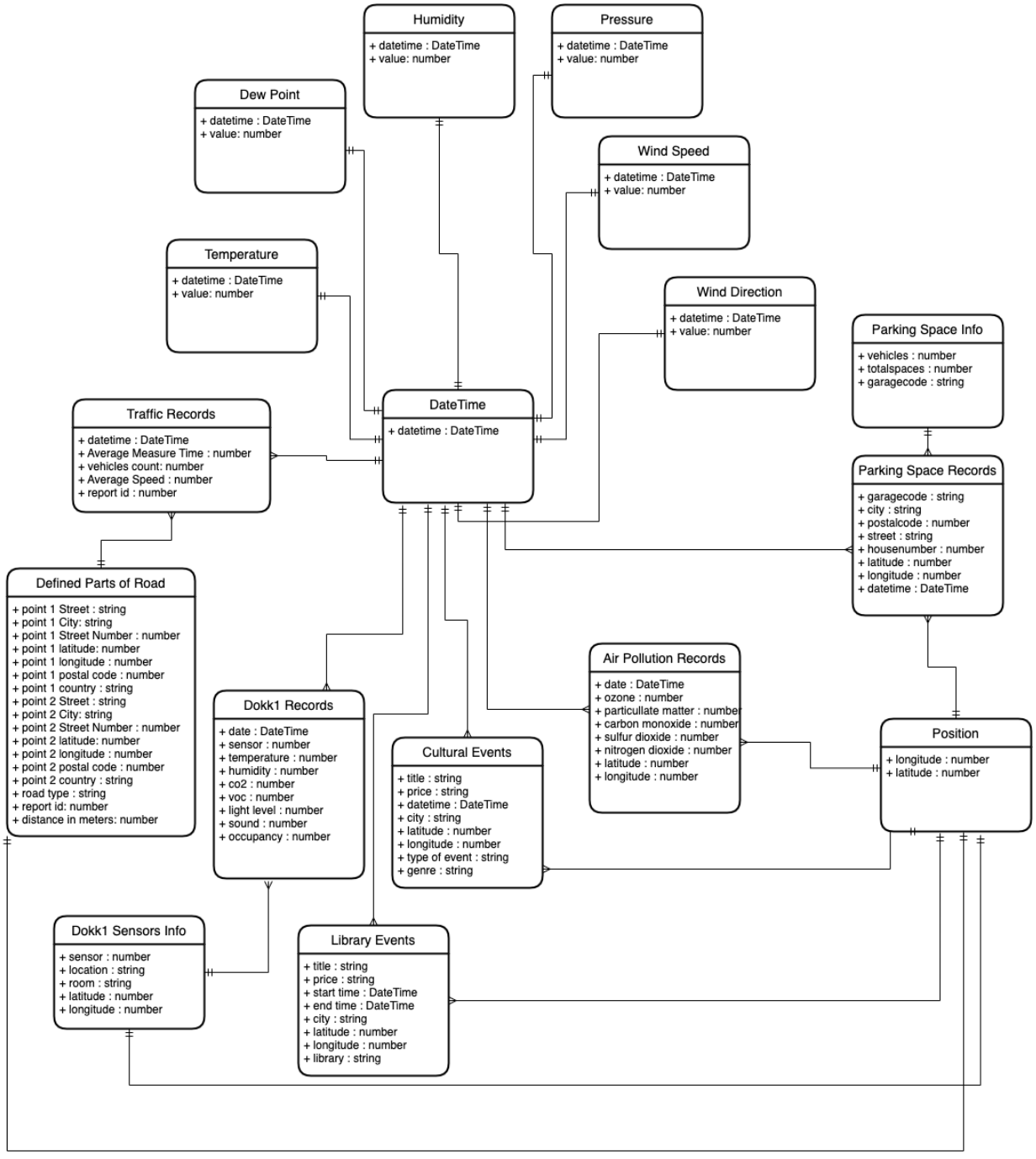
Αυτό που καταλαβαίνουμε είναι ότι όσο υψηλότερα είναι τα νούμερα των μετρήσεων τόσο πιο επικίνδυνα είναι για τη υγεία μας αλλά και για το περιβάλλον.

Air Pollution (7,900,604 records)	
Περιγραφή	Περιέχει μετρήσεις που αφορούν διάφορους παράγοντες που ρυπαίνουν την ατμόσφαιρα. Οι μετρήσεις προέρχονται από 449 διαφορετικά σημεία παρατήρησης στα οποία τοποθετήθηκαν ειδικοί αισθητήρες. Τα σημεία αυτά συμπίπτουν με τα σημεία παρατήρησης της κυκλοφορίας που αναφέραμε παραπάνω. Οι εγγραφές για κάθε αισθητήρα είναι καθημερινές και ανά 5 λεπτά καταγεγραμμένες.
Χρονική Διάρκεια	Αυγ – Οκτ 2014

	Όνομα Πεδίου	Τύπος	Περιγραφή
Πεδία Δεδομένων (Σχήμα)	**timestamp**	*DateTime*	Η ημερομηνία και ώρα μέτρησης
	ozone	*Number*	Το όζον (σε AQI*)
	particullate_matter	*Number*	Η σωματιδιακή ύλη (σε AQI*)
	carbon_monoxide	*Number*	Μονοξείδιο του Άνθρακα (σε AQI*)
	sulfur_dioxide	*Number*	Διοξείδιο του θείου (σε AQI*)
	nitrogen_dioxide	*Number*	Διοξείδιο του αζώτου (σε AQI*)
	latitude	*Number*	Γεωγραφικό πλάτος του αισθητήρα
	longitude	*Number*	Γεωγραφικό μήκος της διεύθυνσης του αισθητήρα

4.8 Logical E-R Data Modeling

Στην συνέχεια παρουσιάζεται σε ένα Logical E-R Data Modeling οι οντότητες των δεδομένων με τα γνωρίσματά τους, αλλά και τη σύνδεση αυτών μεταξύ τους.



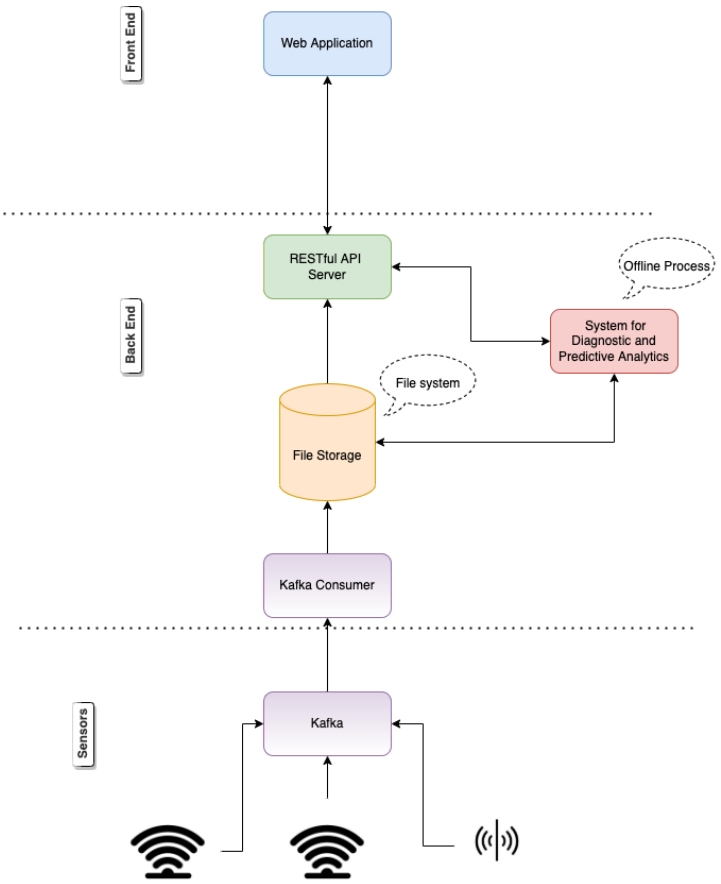
Εικόνα 4.2 Διάγραμμα οντοτήτων και σχέσεων δεδομένων

5 *Αρχιτεκτονική Συστήματος*

Το σύστημα μας αποτελείται από δύο βασικά μέρη και κάποια δευτέροντα. Όλα αυτά συνδυάζονται για να πετύχουμε να έχουμε ένα σύστημα που θα μπορεί να αποθηκεύει τα δεδομένα που λαμβάνει από διάφορους αισθητήρες και πηγές μέσα στην πόλη του Aarhus, να τα διαχειρίζεται και να τα προβάλλει στον χρήστη του συστήματος. Αποτελεί δηλαδή ένα ενιαίο κεντρικό πληροφοριακό σύστημα που λαμβάνει όλα τα δεδομένα της πόλης, τα αποθηκεύει, τα εξερευνά, δημιουργεί κάποια στατιστικά στοιχεία, αναζητά κρυφές πληροφορίες και εν τέλη όλα αυτά τα δίνει στον χρήστη της εφαρμογής για να μπορεί να τα αξιοποιήσει με τον δικό τρόπο. Επομένως είναι ένα εργαλείο που βοηθάει στην περιγραφική ανάλυση, τη διαγνωστική ανάλυση και την προγνωστική ανάλυση των εισερχόμενων δεδομένων και όλα αυτά κεντρικά, και πολύ προσιτά προς τον χρήστη της εφαρμογής μέσω της Γραφικής Διεπαφής Χρήστη του συστήματος. Μπορεί να γλιτώσει πολύ χρόνο σε έναν αναλυτή καθώς τα δεδομένα ανανεώνονται αυτόματα κάθε συγκεκριμένη χρονική στιγμή και μαζί με αυτά ανανεώνονται όλα τα διαγράμματα.

Τέλος, δίνει τη δυνατότητα στον χρήστη της εφαρμογής να προχωρήσει στην προγνωστική ανάλυση βασικών χρονοσειρών με βάση πολλά διαθέσιμα στατιστικά μοντέλα και μοντέλα μηχανικής μάθησης. Ειδικά στα μοντέλα μηχανικής μάθησης μπορεί να αλλάζει γραφικά τις παραμέτρους της αρχιτεκτονικής τους και να βλέπει κάθε φορά τη νέα πρόβλεψη, κάτι το οποίο δίνει νέες δυνατότητες σε αναλυτές που δεν έχουν γνώση από προγραμματισμό να έχουν πρόσβαση σε τέτοια μοντέλα πρόβλεψης.

Όπως αναφέραμε, το σύστημα αποτελείται από δύο βασικά μέρη, το Web App που είναι η Γραφική Διεπαφή Χρήστη με το σύστημά μας και έναν RESTful API Server, ο οποίος αναλαμβάνει να παρέχει τα δεδομένα με βάση διάφορα φίλτρα προς του καταναλωτές του. Επίσης, υπάρχει και ένα core σύστημα που δέχεται τα νέα δεδομένα και είναι υπεύθυνο για την διαγνωστική τους ανάλυση και προσπαθεί να προσαρμόσει τα μοντέλα μηχανικής μάθησης έτσι ώστε να πετυχαίνουμε καλύτερη πρόβλεψη κάθε φορά για τις χρονοσειρές μας. Έτσι πέρα από τη δυνατότητα που παρέχει το σύστημα στον χρήστη να αλλάζει παραμέτρους των μοντέλων, αυτό προσπαθεί κάθε φορά να υπολογίζει την καλύτερη αρχιτεκτονική των μοντέλων και να εμφανίζει την πρόβλεψή τους. Τέλος τα δεδομένα από τους διάφορους αισθητήρες μπορούν να λαμβάνονται μέσα από ένα Kafka, το οποίο αποτελεί ένα κατανεμημένο publish-subscribe messaging system. Έτσι αν οι αισθητήρες κάνουν publish τα δεδομένα εκεί και εμείς έχουμε το σύστημά μας να κάνει consume τα δεδομένα αυτά, μπορούμε ύστερα να τα αποθηκεύσουμε στη βάση δεδομένων μας ή σε κάποιο file storage system. Η αρχιτεκτονική του συστήματος γραφικά απεικονίζεται στην παρακάτω εικόνα:



Εικόνα 5.1 Αρχιτεκτονική Συστήματος

5.1 Η Γραφική Διεπαφή Χρήστη μέσω Web App

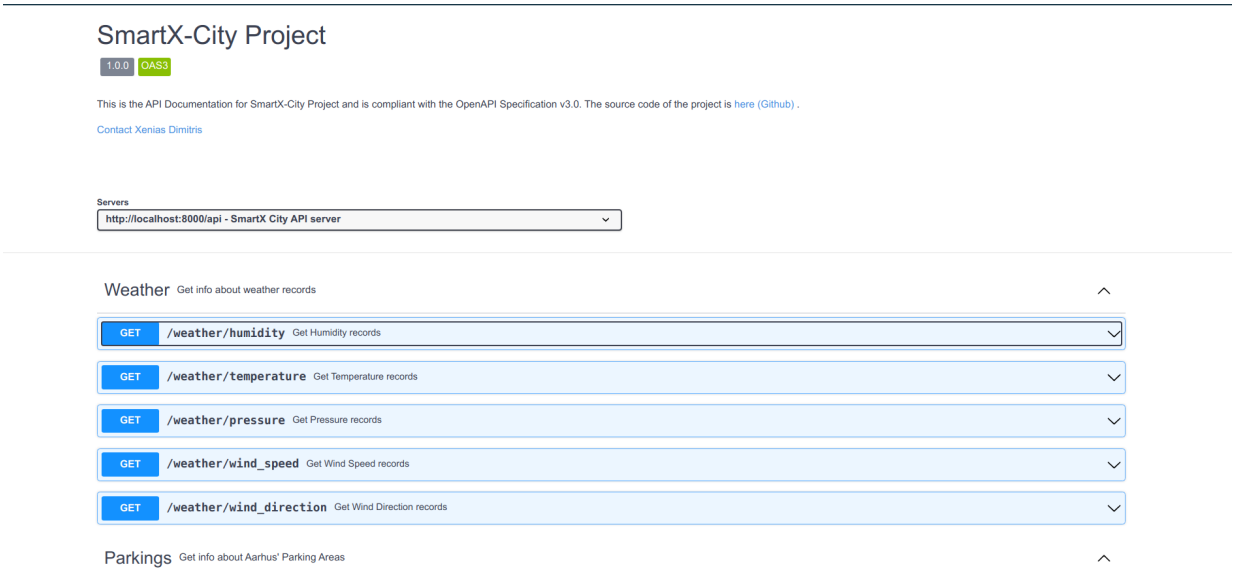
Όπως αναφέρθηκε, σκοπός του συστήματος είναι να δίνει τη δυνατότητα ενός χρήστη να δει τα δεδομένα τα οποία είναι αποθηκευμένα στο σύστημα και τα οποία προέρχονται από διάφορες πηγές και αισθητήρες μέσα από την πόλη Aarhus. Αυτό επιτυγχάνεται μέσω ενός GUI (Γραφική Διεπαφή Χρήστη). Τα δεδομένα αυτά πρέπει να παρουσιάζονται με γραφικό τρόπο στο UI, το οποίο πρέπει να είναι χρηστικό, εύκολο και "όμορφο". Έτσι θα μπορεί ο χρήστης, που πιθανότητα να είναι κάποιος αναλυτής, να μπορεί να δει δεδομένα με βάση δικά του καταχωρημένα φίλτρα και να μπορεί να προβεί σε κάποια περαιτέρω ανάλυση αυτών. Επειδή αυτό είναι και το πρώτο και πιο βασικό στάδιο της ανάλυσης για έναν αναλυτή, πρέπει η εφαρμογή να είναι προσβάσιμη εύκολα, και όλες της οι λειτουργίες να είναι κατανοητές από τον χρήστη. Η υλοποίηση του user interface με το σύστημά μας γίνεται μέσα από ένα web app. Το web app είναι γραμμένο σε ReactJS [23] και χρησιμοποιεί τις βιβλιοθήκες amCharts [25] και Mapbox [26] για την απεικόνιση δεδομένων σε γραφήματα και απεικόνισης δεδομένων σε χάρτη της πόλης. Παράλληλα για το καλύτερο User Experience (UX) χρησιμοποιείται η βιβλιοθήκη Material UI [24], που δημιουργεί ένα οικείο γραφικό περιβάλλον για τον χρήστη, αφού βασίζεται σε βασικά components που χρησιμοποιούν οι εφαρμογές της Google.

5.2 RESTful API Server

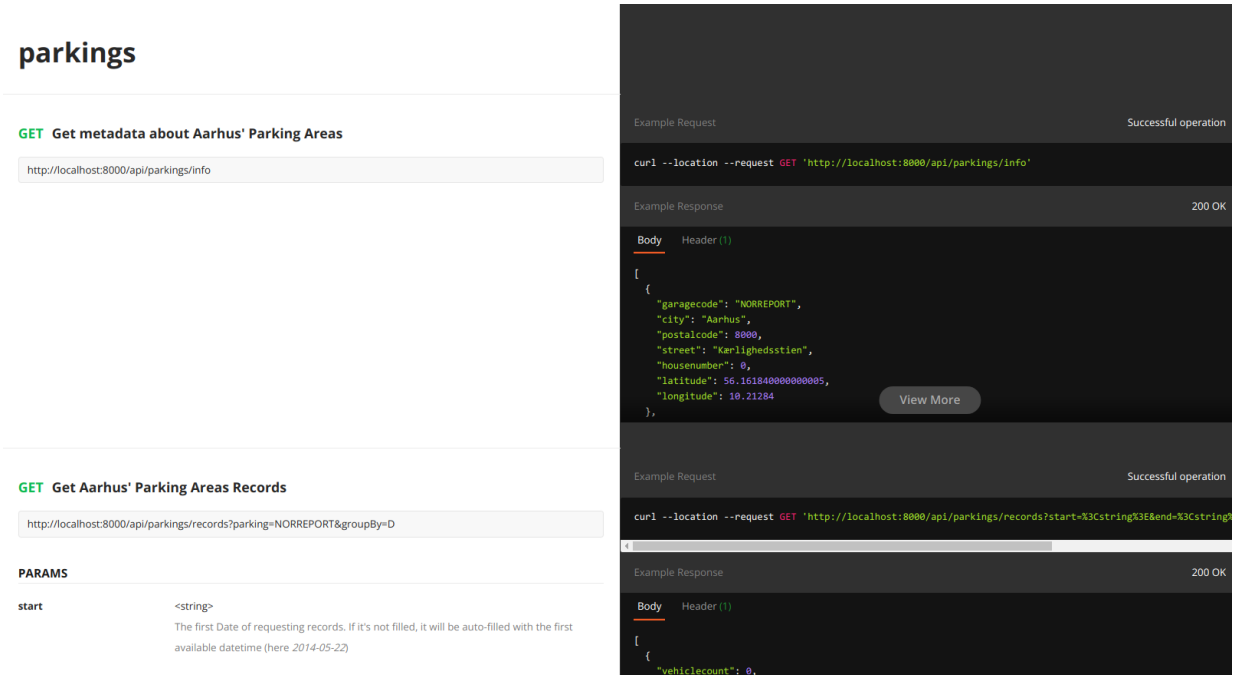
Το web app λοιπόν αποτελεί μια ξεχωριστή και αυτόνομη εφαρμογή που μπορεί να απεικονίζει τα δεδομένα του συστήματός μας. Συγκεκριμένα λαμβάνει τα απαραίτητα δεδομένα μέσα από API Calls προς τον back-end server μας, ο οποίος είναι υλοποιημένος με το Django Framework [21] και αποτελεί ουσιαστικά ένα RESTful API. Μια τέτοια αρχιτεκτονική που διαχωρίζει το front-end από το back-end είναι πολύ χρήσιμη στις μέρες μας. Συγκεκριμένα, όσο αποκεντρικοποιούμε λειτουργίες από ένα σύστημα, αυξάνουμε τις δυνατότητές του και το κάνουμε κλιμακώσιμο. Έτσι, το back-end ως REStful API μπορεί να συντηρηθεί και να υλοποιηθεί ανεξάρτητα από το front-end κομμάτι του. Το μόνο που μας ενδιαφέρει από πλευράς back-end είναι να ορίσουμε τις προδιαγραφές του API και να αναπτύξουμε κώδικα για να παρέχει δεδομένα με βάση τα calls σε συγκεκριμένα endpoints του. Αυτά τα endpoints μπορούν να τα καλέσουν όχι μόνο ένα web-app, αλλά χιλιάδες

consumers, κάτι που κάνει το σύστημά μας προσεγγίσιμο σε περισσότερους consumers και μπορούμε να παρέχουμε δεδομένα και σε άλλες υπηρεσίες.

Έτσι λοιπόν έχουμε το web app το οποίο κάνει API calls στον API server του συστήματος με βάση κάποια φίλτρα και το σύστημά μας μετά αναλαμβάνει να πάει στα δεδομένα τα οποία είναι αποθηκευμένα σε μία βάση δεδομένων (εδώ σε τοπικά αρχεία στο file system) και να αντλήσει τα δεδομένα με βάση τα φίλτρα του api call. Στην συνέχεια, στέλνει τα δεδομένα αυτά σε μια συγκεκριμένη μορφή (εδώ json format) και ο consumer τα διαχειρίζεται όπως αυτός θέλει με προβολή σε διαγράμματα και χάρτες. Γίνεται κατανοητό όμως ότι για να γνωρίζουν οι consumers πως να κάνουν τα api calls στο σύστημά μας, αυτό πρέπει να παρέχει ένα documentation. Το documentation αυτό μπορεί να γραφεί στο χέρι, αλλά ένας έξυπνος τρόπος είναι η δημιουργία ενός OpenAPI Specification [27]. Με τον τρόπο αυτό γράφουμε μια φορά τις προδιαγραφές του συστήματος σε μια συγκεκριμένη μορφή που ορίζει το πρωτόκολλο αυτό και μπορούμε με τη μορφή αυτή να εξάγουμε το documentation, ή να το φορτώσουμε στο Postman για την άμεση χρήση του, ή να χρησιμοποιήσουμε και πιο γραφικούς τρόπους μέσω Swagger. Στη συνέχεια παρουσιάζεται το API documentation που έχει γραφτεί μέσω του OpenAPI Specification 3.0 μέσω Postman και μέσω Swagger.



Εικόνα 5.2 API Specification από Swagger



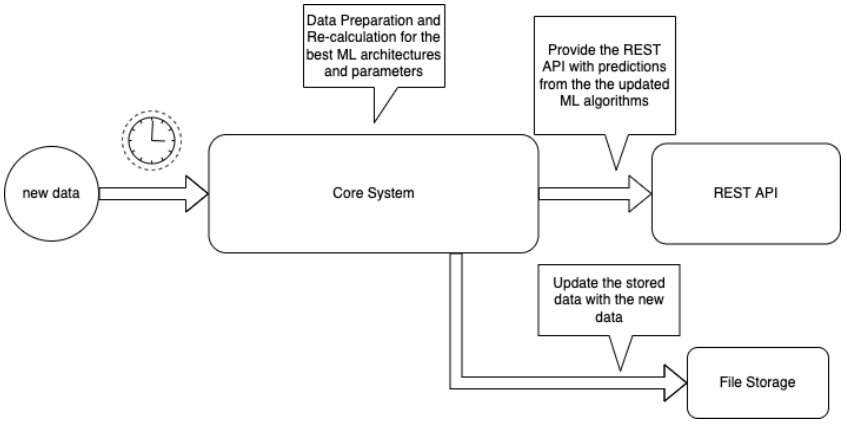
Εικόνα 5.3 API Specification από Postman

Ολόκληρο το API Specification μπορεί να διερευνηθεί ΕΔΩ

5.3 Core System

Όπως αναφέραμε η περιγραφική ανάλυση των δεδομένων γίνεται μέσα από τη Γραφική Διεπαφή Χρήστη του Web App από τον ίδιο το χρήστη. Η προετοιμασία των δεδομένων αλλά και Διαγνωστική Ανάλυση και η Προγνωστική Ανάλυση του συστήματος γίνεται μέσω ενός core συστήματος σε γλώσσα Python. Το σύστημα αυτό ενημερώνεται ανά χρονικά διαστήματα με τα νέα δεδομένα που εμπλουτίζουν την ιστορία των χρονοσειρών. Αφού εισέλθουν νέα δεδομένα, το σύστημα ξεκινάει μια διαδικασία προετοιμασίας δεδομένων για να ενταχθούν τα νέα δεδομένα, όπως πρέπει, στην ιστορία των χρονοσειρών και να αποθηκευτούν τα δεδομένα. Έχοντας πλέον ενημερωμένα και επικαιροποιημένα τα δεδομένα, το REST API μπορεί να προσφέρει τα νέα δεδομένα στο χρήστη .Στη συνέχεια το core system επαναλαμβάνει αυτοματοποιημένες διαδικασίες για την διαγνωστική ανάλυση των δεδομένων με βάση τη καλύτερη ιστορία που διαθέτει κάθε φορά και προσπαθεί να βρει τις καλύτερες παραμέτρους

για τα μοντέλα μηχανικής μάθησης ώστε να επιτυγχάνει κάθε φορά καλύτερη ακρίβεια στις προβλέψεις του. Έτσι όταν ο χρήστης διαλέξει μια πρόβλεψη για μια χρονοσειρά μέσω του γραφικού περιβάλλοντος, το αίτημα πάει στον API server και αυτός στη συνέχεια απευθύνεται στο core system που συνδέεται με τον server για να πάρει την πρόβλεψη της χρονοσειράς. Από τη στιγμή που το core system ενημερώνεται από τα νέα δεδομένα, η πρόβλεψη κάθε φορά είναι επικαιροποιημένη, έχοντας λάβει υπόψη τα νέα δεδομένα. Επειδή στο σύστημα αυτό γίνεται ο πιο μεγάλος όγκος της επεξεργασίας δεδομένων, το ονομάζουμε “core system” και ασχολείται καθαρά με τη διαχείριση και την ανάλυση δεδομένων, αποτελώντας μια αυτοτελή μονάδα στο συνολικό project.



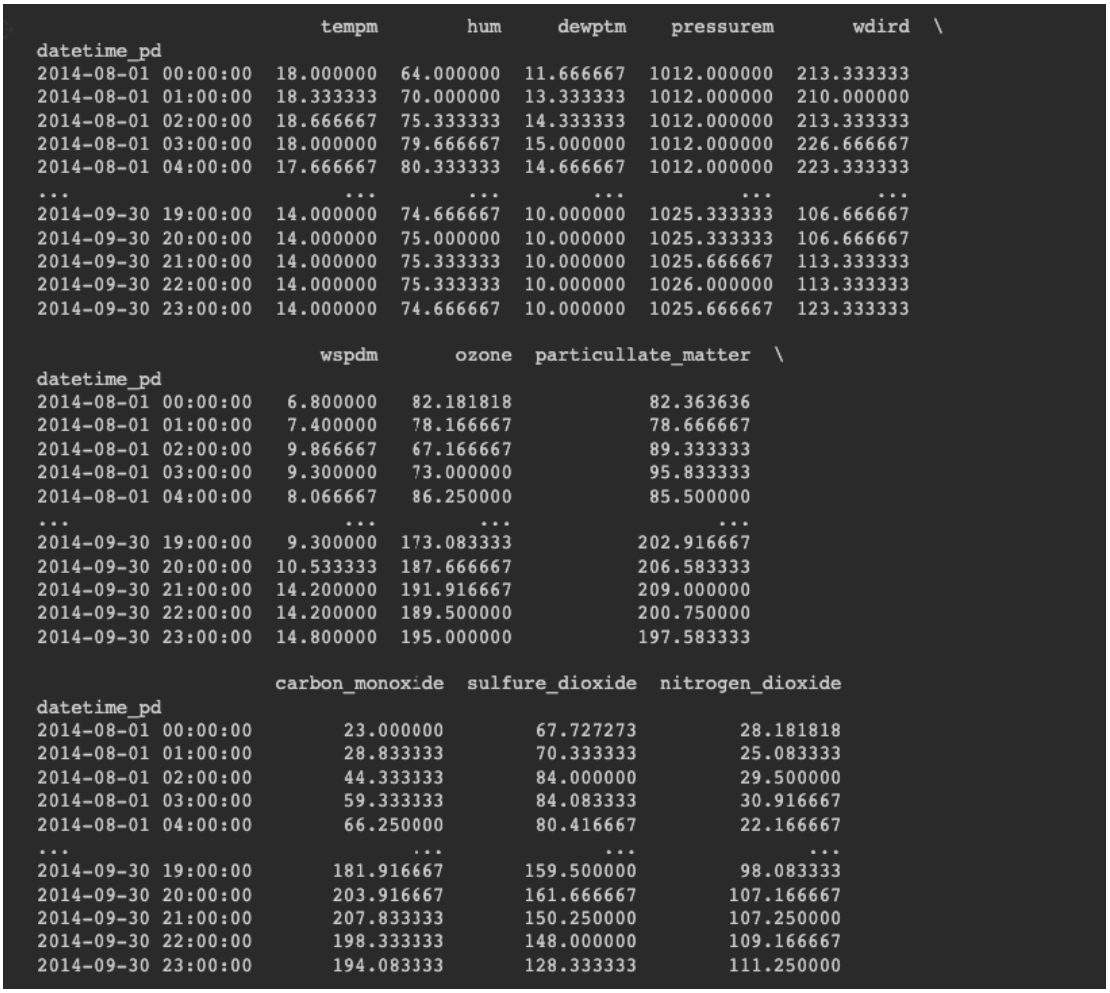
Εικόνα 5.4 core Functionality Component Diagram

6 *Ανάλυση Δεδομένων*

Σε κάθε project που είναι data oriented, τα δεδομένα είναι προφανώς το πιο σημαντικό και δύσκολο κομμάτι, και σε αυτά πρέπει να δοθεί το μεγαλύτερο βάρος δουλείας. Η διαδικασία μάλιστα αυτή για την επιτυχημένη διαχείριση των δεδομένων και αξιοποίηση αυτών για τον εκάστοτε σκοπό του συστήματος, είναι μια διαδικασία που αποτελείται από πολλά βήματα τα οποία μάλιστα είναι σειριακά και πρέπει κάθε φορά να εκτελούνται με προσοχή ώστε να είμαστε σίγουροι ότι το κομμάτι των δεδομένων και της ανάλυσης αυτών είναι σωστά εκτελεσμένα. Σε αντίθετη περίπτωση έχουμε αποτύχει στο πιο σημαντικό κομμάτι του συστήματος και άρα το σύστημά μας δεν έχει κάποια αξία να προσφέρει.

Έχοντας αναλύσει λοιπόν τα δεδομένα, και τη σύνδεσή τους, που χρησιμοποιούνται στο σύστημα αλλα και την αρχιτεκτονική του συστήματος, στο κεφάλαιο αυτό παρουσιάζεται η διαδικασία της ανάλυσης δεδομένων που υλοποιείται κυρίως στο core σύστημα που αναφέραμε. Στο κεφάλαιο 5, στην υποενότητα 5.3, αναφέρθηκε ότι όταν νέα ακατέργαστα δεδομένα εισέρχονται στο σύστημα, αυτά περνάνε από ένα core system για να γίνει η προετοιμασία τους ώστε να ενταχθούν στα κύρια δεδομένα που χρησιμοποιούνται από το σύστημα. Ύστερα το σύστημα χρησιμοποιεί τα κύρια δεδομένα για να κάνει την ανάλυσή τους. Στο κεφάλαιο αυτό αναλύονται τα βήματα της Διαγνωστικής Ανάλυσης και της Προγνωστικής Ανάλυσης που εκτελεί το core system, ώστε να παρέχει κάθε φορά επικαιροποιημένα μοντέλα πρόβλεψης και προβλέψεις που είναι ακριβείς προς τον χρήστη.

Τα κύρια δεδομένα του συστήματος έχουν την ακόλουθη μορφή:



datetime_pd	tempm	hum	dewptm	pressurem	wdird
2014-08-01 00:00:00	18.000000	64.000000	11.666667	1012.000000	213.333333
2014-08-01 01:00:00	18.333333	70.000000	13.333333	1012.000000	210.000000
2014-08-01 02:00:00	18.666667	75.333333	14.333333	1012.000000	213.333333
2014-08-01 03:00:00	18.000000	79.666667	15.000000	1012.000000	226.666667
2014-08-01 04:00:00	17.666667	80.333333	14.666667	1012.000000	223.333333
...	...	...	...	...	...
2014-09-30 19:00:00	14.000000	74.666667	10.000000	1025.333333	106.666667
2014-09-30 20:00:00	14.000000	75.000000	10.000000	1025.333333	106.666667
2014-09-30 21:00:00	14.000000	75.333333	10.000000	1025.666667	113.333333
2014-09-30 22:00:00	14.000000	75.333333	10.000000	1026.000000	113.333333
2014-09-30 23:00:00	14.000000	74.666667	10.000000	1025.666667	123.333333

datetime_pd	wspdm	ozone	particullate_matter
2014-08-01 00:00:00	6.800000	82.181818	82.363636
2014-08-01 01:00:00	7.400000	78.166667	78.666667
2014-08-01 02:00:00	9.866667	67.166667	89.333333
2014-08-01 03:00:00	9.300000	73.000000	95.833333
2014-08-01 04:00:00	8.066667	86.250000	85.500000
...	...	...	...
2014-09-30 19:00:00	9.300000	173.083333	202.916667
2014-09-30 20:00:00	10.533333	187.666667	206.583333
2014-09-30 21:00:00	14.200000	191.916667	209.000000
2014-09-30 22:00:00	14.200000	189.500000	200.750000
2014-09-30 23:00:00	14.800000	195.000000	197.583333

datetime_pd	carbon_monoxide	sulfure_dioxide	nitrogen_dioxide
2014-08-01 00:00:00	23.000000	67.727273	28.181818
2014-08-01 01:00:00	28.833333	70.333333	25.083333
2014-08-01 02:00:00	44.333333	84.000000	29.500000
2014-08-01 03:00:00	59.333333	84.083333	30.916667
2014-08-01 04:00:00	66.250000	80.416667	22.166667
...	...	...	...
2014-09-30 19:00:00	181.916667	159.500000	98.083333
2014-09-30 20:00:00	203.916667	161.666667	107.166667
2014-09-30 21:00:00	207.833333	150.250000	107.250000
2014-09-30 22:00:00	198.333333	148.000000	109.166667
2014-09-30 23:00:00	194.083333	128.333333	111.250000

Εικόνα 6.1 Η μορφή των Δεδομένων στο σύστημα

Συνεπώς όταν έχουμε αυτό το στιγμιότυπο των κύριων Δεδομένων και έρθουν κάποια νέα ακατέργαστα δεδομένα, αυτά τα ακατέργαστα δεδομένα θα αφορούν την αμέσως επόμενη χρονική στιγμή της τελευταίας χρονικά καταγραφής που υπάρχει στα κύρια Δεδομένα.

6.1 Προετοιμασία και Προεπεξεργασία Δεδομένων

Το πρώτο και πάρα πολύ σημαντικό βήμα που το core system αναλαμβάνει να πράξει μόλις λάβει νέα δεδομένα από ένα Kafka στην προκειμένη περίπτωση είναι η προεπεξεργασία των δεδομένων αυτών και η προετοιμασία τους για την ένταξή τους στα κύρια δεδομένα του file storage, τα οποία και το σύστημα προσφέρει στο χρήστη. Τα υπο-βήματα που ακολουθούνται κατά την προετοιμασία και προεπεξεργασία δεδομένων είναι τα ακόλουθα.

6.1.1 Ανίχνευση Χαμένων Μετρήσεων

Πολλές φορές αισθητήρες παρέχουν δεδομένα που ως τιμή μέτρησης δεν έχουν κάποιον αριθμό (NaN from Not A Number). Για παράδειγμα μπορεί να είναι κενό το πεδίο της τιμής, ή κάποιος περίεργος χαρακτήρας, αποτέλεσμα κάποιας δυσλειτουργίας του αισθητήρα. Επίσης, μπορεί για κάποια χρονική στιγμή να μην λάβουμε καν μέτρηση με το αναμενόμενο timestamp. Έτσι το core σύστημα ανιχνεύει τις τιμές αυτές ή χαμένες μετρήσεις και πρέπει να τις αντικαταστήσει με κάποιον τρόπο. Ο τρόπος που χρησιμοποιούμε εδώ είναι να αντιγράψουμε την τιμή της αμέσως προηγούμενης διαθέσιμης μέτρησης, καθώς θεωρούμε ότι δεν έχουμε τεράστια μεταβολές στα μεγέθη που μετράμε μέσα σε ένα τόσο μικρό χρονικό διάστημα.

6.1.2 Ανίχνευση μη Έγκυρων Μετρήσεων

Το ερώτημα εδώ είναι τί θεωρείται μη έγκυρη μέτρηση και πως εμείς μπορούμε να το ξέρουμε με σιγουριά. Γενικά δύο είναι οι περιπτώσεις μια τιμή να μην είναι έγκυρη. Η πρώτη περίπτωση είναι όταν βλέπουμε τιμή σε μια χρονοσειρά η οποία να είναι 0. Γενικά αυτό μπορεί να είναι μια σωστή μέτρηση καθώς το 0 για παράδειγμα μπορεί να είναι η τιμή της θερμοκρασίας, αλλά πρέπει να γίνει μια μελέτη για το αν αυτή η τιμή είναι στο πεδίο τιμών της χρονοσειράς και αν είναι λογικό να είναι 0 σύμφωνα και με τα υπόλοιπα εξαρτώμενα μεγέθη. Σε αντίθετη περίπτωση που δεν είναι λογικό, τότε την αντικαθιστούμε με την τιμή της αμέσως προηγούμενης διαθέσιμης τιμής που έχουμε, για τον ίδιο λόγο με προηγουμένως.

Η δεύτερη περίπτωση που μπορεί μια τιμή να μην είναι έγκυρη, είναι αν αυτή αποκλίνει πάρα πολύ από την μέση τιμή. Μαθηματικά όμως υπάρχει η μέθοδος IQR Method (identify outliers) που μας υποδεικνύει αν μια τιμή είναι μη έγκυρη για τον λόγο αυτό. Συγκεκριμένα υπολογίζει το IQR ως το Q3-Q1 [2.2.1.1] και θεωρεί ως μη έκγυρη τιμή οποιαδήποτε τιμή είναι εκτός ενός πεδίου τιμών που ορίζεται ως:

$$[Q1 - 1.5 * IQR, Q + 1.5 * QR]$$

Τα μεγέθη αυτά υπολογίζονται εύκολα καθώς μέσω της εντολής *.describe()* της *pandas* λαμβάνουμε τα quartiles εύκολα.

	tempm	hum	dewptm	pressurem	wdird	\
count	1464.000000	1464.000000	1464.000000	1464.000000	1464.000000	
mean	15.587318	70.082081	10.351890	1013.842099	184.881603	
std	3.577939	14.046408	2.795900	7.872343	79.927421	
min	2.666667	23.333333	0.000000	992.333333	0.000000	
25%	13.000000	61.333333	8.666667	1008.666667	116.666667	
50%	15.333333	73.000000	10.666667	1013.000000	200.000000	
75%	18.000000	81.000000	12.000000	1019.666667	246.666667	
max	27.000000	92.333333	19.000000	1030.000000	360.000000	

	wspdm	ozone	particullate_matter	carbon_monoxide	\
count	1464.000000	1464.000000	1464.000000	1464.000000	
mean	12.683094	153.355999	93.812691	113.660861	
std	6.806240	39.829207	48.423531	47.414795	
min	0.000000	24.416667	21.583333	21.500000	
25%	7.433333	129.250000	55.166667	76.062500	
50%	11.766667	160.041667	85.875000	115.041667	
75%	17.266667	185.020833	128.354167	147.791667	
max	34.566667	210.250000	209.000000	210.000000	

Εικόνα 6.2 Quartiles από pandas

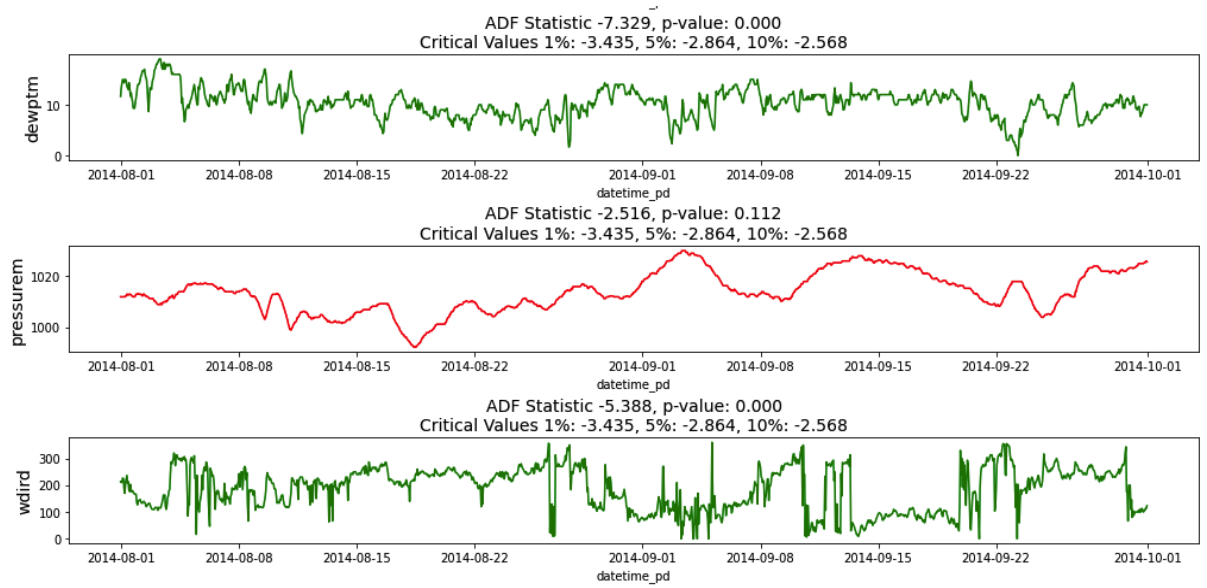
6.1.3 Περιγραφική Ανάλυση

Παρότι η περιγραφική ανάλυση γίνεται από τον χρήστη μέσα από την οπτικοποίηση των δεδομένων στο γραφικό περιβάλλον της εφαρμογής, το core system πρέπει να υπολογίσει κάποια πολύ βασικά ποιοτικά χαρακτηριστικά των χρονοσειρών. Τα χαρακτηριστικά αυτά απαντούν στο "τί γίνεται" στα δεδομένα μας, και για αυτό είναι και μέρος της προγνωστικής ανάλυσης όπως περιγράφηκε στην υποενότητα 2.2.1.1. Επομένως βασικά χαρακτηριστικά που βρίσκει το core system είναι τα ακόλουθα.

6.1.3.1 Στασιμότητα Χρονοσειρών

Στην υποενότητα 2.2.1.1 περιγράφτηκε πότε μια χρονοσειρά είναι στάσιμη. Επίσης αναφέρθηκε ότι κάποια στατιστικά μοντέλα απαιτούν οι χρονοσειρές που εισάγονται σε αυτά να είναι στάσιμες. Έτσι κάθε φορά το σύστημα πρέπει αν κάποια χρονοσειρά δεν είναι στάσιμη και θέλουμε να χρησιμοποιήσουμε κάποιο μοντέλο πρόβλεψης για αυτήν, να την μετατρέψει σε στάσιμη.

Η βιβλιοθήκη *Statsmodels* παρέχει μια συνάρτηση *adfuller* για το τεστ μηδενικής υπόθεσης που θα μας βοηθήσει για μελέτη της στασιμότητας. Με κόκκινο χρώμα εμφανίζονται

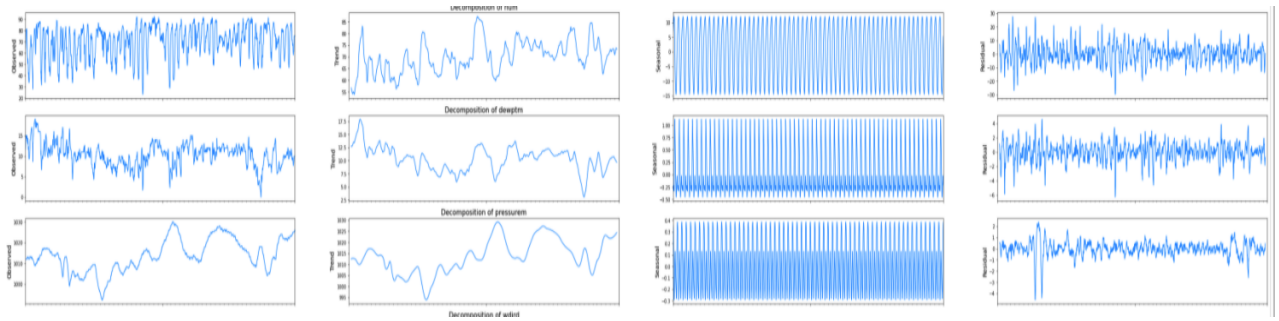


οι χρονοσειρές που δεν είναι στάσιμες, εμώ με πράσινο αυτές που είναι στάσιμες. Στη συνέχεια φαίνεται ένα screenshot από την ανάλυση στασιμότητας στα δεδομένα μας.

Εικόνα 6.3 ADFuller Test για στασιμότητα Χρονοσειρών

6.1.3.2 Ποιοτικά Χαρακτηριστικά Χρονοσειρών

Ως ποιοτικά χαρακτηριστικά ορίζονται η τάση, η εποχικότητα και η κυκλικότητα της χρονοσειράς. Οι νέες χρονοσειρές που προκύπτουν από την απο-εποχικοποίηση των χρονοσειρών μπορεί να φανούν χρήσιμες κατά το επόμενο στάδιο της διαγνωστικής ανάλυσης και για αυτό υπολογίζονται από το core system .Στη συνέχεια φαίνεται ένα screenshot από την απο-εποχικοποίηση των δεδομένων μας.



Εικόνα 6.4 Απο-εποχικοποίηση Χρονοσειρών

Συγκεκριμένα αριστερά βλέπουμε τις κανονικές χρονοσειρές, δίπλα τους την τάση της χρονοσειράς, πιο δίπλα την εποχικότητα και στο τέλος βλέπουμε την αρχική χρονοσειρά, αν αφαιρέσουμε από αυτήν τις συνισταμένες της τάσης και της εποχικότητας.

6.1.3.3 Feature Engineering

Ένας πολύ μεγάλο μέρος της κοινότητας που ασχολείται με την Ανάλυση Δεδομένων έχει τονίσει την σημασία του Feature Engineering για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης και ασχολείται με αυτό. Ως Feature Engineering ή σε ελεύθερη μετάφραση Μηχανική των Χαρακτηριστικών ορίζεται ως η διαδικασία με την οποία από τα ακατέργαστα δεδομένα που έχουμε, προσπαθούμε να εξάγουμε πληροφορία που θα την χρησιμοποιήσουμε ως νέο χαρακτηριστικό (feature) των δεδομένων, με στόχο να βελτιώσουμε την ποιότητα των αποτελεσμάτων από τους αλγορίθμους μηχανικής μάθησης, σε αντίθεση με το εάν κρατούσαμε τα αρχικά δεδομένα. Καταλαβαίνουμε ότι η διαδικασία αυτή είναι αρκετά εμπειρική σε ένα βαθμό και δεν είναι σίγουρο ότι πάντα θα βρούμε features που θα βοηθήσουν τον αλγόριθμο. Θέλει συνεχή αναζήτηση και έλεγχο της απόδοσης των αλγορίθμων και ξανά πάλι από την αρχή.

Στη δική μας περίπτωση το μόνο που μπορούμε να κάνουμε είναι να πειραματιστούμε με τα timestamps. Συγκεκριμένα, μπορούμε να δημιουργήσουμε επιπλέον χαρακτηριστικά στις χρονοσειρές μας, με την εισαγωγή των χαρακτηριστικών: year, month, day, day of the year, week of the year και quarter. Έτσι μπορεί η μηχανική μάθηση να εξάγει καλύτερα αποτελέσματα, έχοντας πιο πολλή πληροφορία για το timestamp έναντι απλά ενός datetime χαρακτηριστικού, το οποίο μάλιστα δεν λαμβάνει καν υπόψην. Για παράδειγμα:

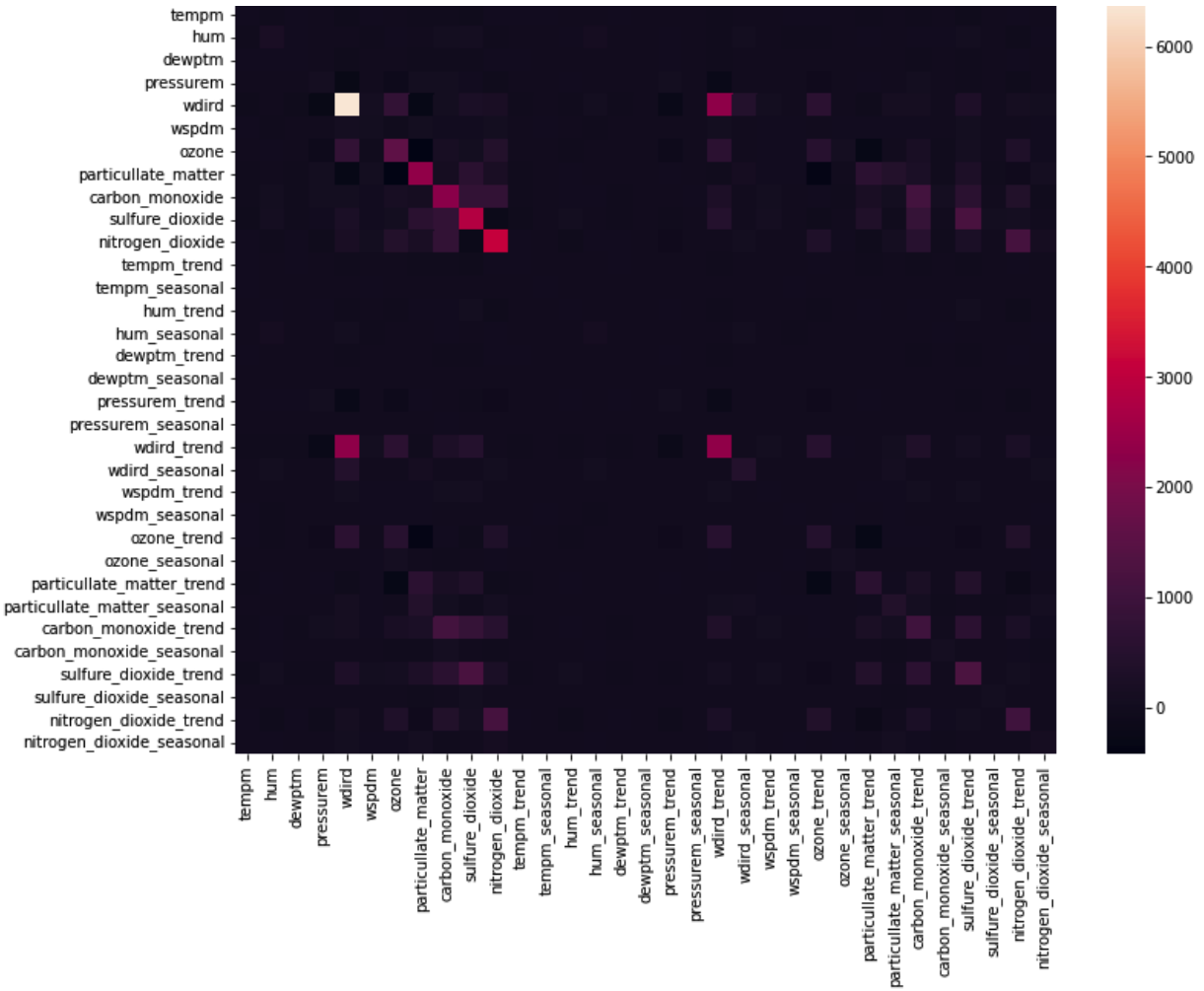
datetime_pd	year	month	day	day_of_year	week_of_year	quarter
2014-08-01 00:00:00	2014	8	1	213	31	3
2014-08-01 01:00:00	2014	8	1	213	31	3
2014-08-01 02:00:00	2014	8	1	213	31	3
2014-08-01 03:00:00	2014	8	1	213	31	3
2014-08-01 04:00:00	2014	8	1	213	31	3

Εικόνα 6.5 Feature Engineering

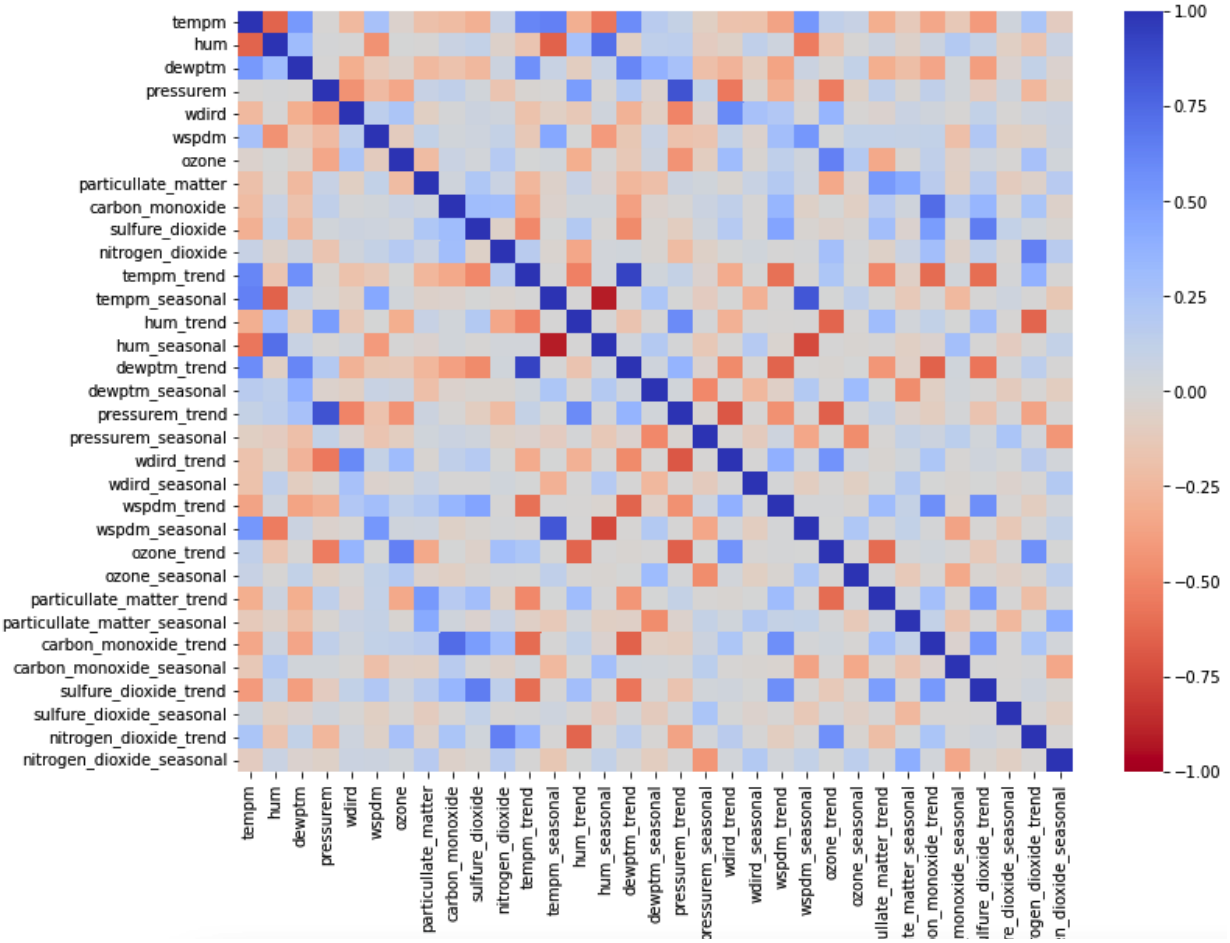
6.2. Διαγνωστική Ανάλυση

Όπως αναφέρθηκε και στην υποενότητα 2.2.2 στο σημείο αυτό το core system προσπαθεί να εξηγήσει "γιατί" οι χρονοσειρές εξελίσσονται με τον τρόπο που εξελίσσονται. Ο άνθρωπος μπορεί να χρησιμοποιήσει και το ένστικτό του, κάποια ειδικά γεγονότα που συνέβησαν, ακόμα και μια γρήγορη ανάλυση των χρονοσειρών γραφικά, για να δει την τάση και την εποχικότητα, ώστε να συνδυάσει τις τιμές με τα ποιοτικά χαρακτηριστικά της χρονοσειράς. Τέλος ένας άνθρωπος θα έλεγχε και πως συνδέονται οι χρονοσειρές μεταξύ τους, δηλαδή κατά πόσο η μεταβολή της μιας χρονοσειράς εξαρτάται από την άλλη.

Συνεπώς και το σύστημα πρέπει και προσομοιώνει τη διαδικασία που θα έκανε ένας άνθρωπος με το μάτι, αλλά το σύστημα το αποτυπώνει σε μαθηματικά και υπολογισμούς. Έτσι Υπολογίζει τον πίνακα συνδιακύμανσης και του συντελεστή γραμμικής συσχέτισης (2.2.2) και τα αποτελέσματα φαίνονται παρακάτω.



Εικόνα 6.6 Covariance Matrix



Εικόνα 6.7 Correlation matrix

Βλέπουμε αρχικά ότι υπολογίζουμε τις τιμές αυτές και για τα ποιοτικά χαρακτηριστικά κάθε χρονοσειράς. Ουσιαστικά με αυτόν τον τρόπο προσπαθούμε να βρούμε κάποια κρυφή εξάρτηση των χρονοσειρών. Για παράδειγμα, θα μπορούσε η τάση του μονοξειδίου το άνθρακα να εξαρτάται από την εποχικότητα της θερμοκρασία. Έτσι με το τέχνασμα αυτό μπορούμε να αντλήσουμε επιπλέον πληροφορία που μπορεί να φανεί χρήσιμη σε μοντέλα που θα λαμβανουν υποψη τις εξαρτήσεις αυτές.

Επίσης ως αναφορά την ερμηνεία των διαγραμμάτων στην εικονα 6.4 φαίνεται η τιμή της συνδιακύμανσης, και λόγω διαφοράς κλίμακας των δεδομένων δεν φαίνεται ξεκάθαρα το ζητούμενο μας. Ωστόσο στο δεύτερο διάγραμμα φαίνεται ξεκάθαρα πλέον μέσω του συντελεστή γραμμικής συσχέτισης το πόσο εξαρτάται η μία χρονοσειρά από την άλλη. Συγκεκριμένα, τιμές κοντά στα άκρα 1 και -1 δηλώνουν απόλυτη ή πάρα πολύ μεγάλη συσχέτιση, ενώ τιμές κοντά στο 0 δηλώνουν ελάχιστη συσχέτιση. Συνεπώς, τα έντονα χρώματα στον πίνακα συσχέτισης δηλώνουν απευθείας ότι υπάρχει κάποια συσχέτιση που πρέπει να ληφθεί υπόψη.

6.3 Προγνωστική Ανάλυση

Όλη αυτή η προηγούμενη ανάλυση έχει ως στόχο την καλύτερη πρόγνωση του συστήματός μας για τις μελλοντικές τιμές των χρονοσειρών μας. Όπως αναφέρθηκε και στην αρχιτεκτονική του συστήματος, το σύστημα παρέχει προβλέψεις με χρονικό ορίζοντα την επόμενη ώρα. Στα στατιστικά μοντέλα που αναλύθηκαν στο Κεφάλαιο 2 δεν υπάρχουν και πολλές παράμετροι, επομένως δεν δίνεται η δυνατότητα του χρήστη να πειραματιστεί με τις παραμέτρους τους. Αντιθέτως στα μοντέλα μηχανικής μάθησης , εκ των οποίων δύο είναι και μοντέλα βαθιάς μηχανικής μάθησης, υπάρχουν πολλές δυνατότητες παραμετροποίησης στην αρχιτεκτονικής και σε βασικά μεγέθη των μονάδων της αρχιτεκτονικής. Έτσι το σύστημά μας αναλύει σε συγκεκριμένα χρονικά διαστήματα τα νέα δεδομένα και προσπαθεί να βρει την αρχιτεκτονική και τις παραμέτρους που βελτιστοποιούν την ακρίβεια των προβλέψεων στο σύνολο των δεδομένων εκπαίδευσης. Παράλληλα, όμως, προσφέρει και τη δυνατότητα ο χρήστης να πειραματιστεί με την αρχιτεκτονική και τις παραμέτρους για σκοπούς έρευνας και προβάλλει κάθε φορά τις νέες προβλέψεις.

Στη συγκεκριμένη υποενότητα παρουσιάζεται η προγνωστική ανάλυση που κάνει το core system για να προβλέψει την τιμή του μονοξειδίου του άνθρακα σε χρονικό ορίζοντα μιας ώρας. Ωστόσο η ίδια διαδικασία γίνεται και για τις υπόλοιπες χρονοσειρές, όπως επίσης και μπορεί να επεκταθεί ο αλγόριθμος ώστε να προβλέπει μεγαλύτερο χρονικό ορίζοντα με μεγαλύτερη πιθανότητα σφάλματος. Ακολουθεί ο τρόπος που το σύστημα "χτίζει" τα μοντέλα πρόβλεψης. Η θεωρία αναλύεται στην υπο ενότητα 2.2.3.2, επομένως εδώ απλά παρουσιάζεται συνοπτικά πως εφαρμόζεται η θεωρία στις χρονοσειρές που έχουμε.

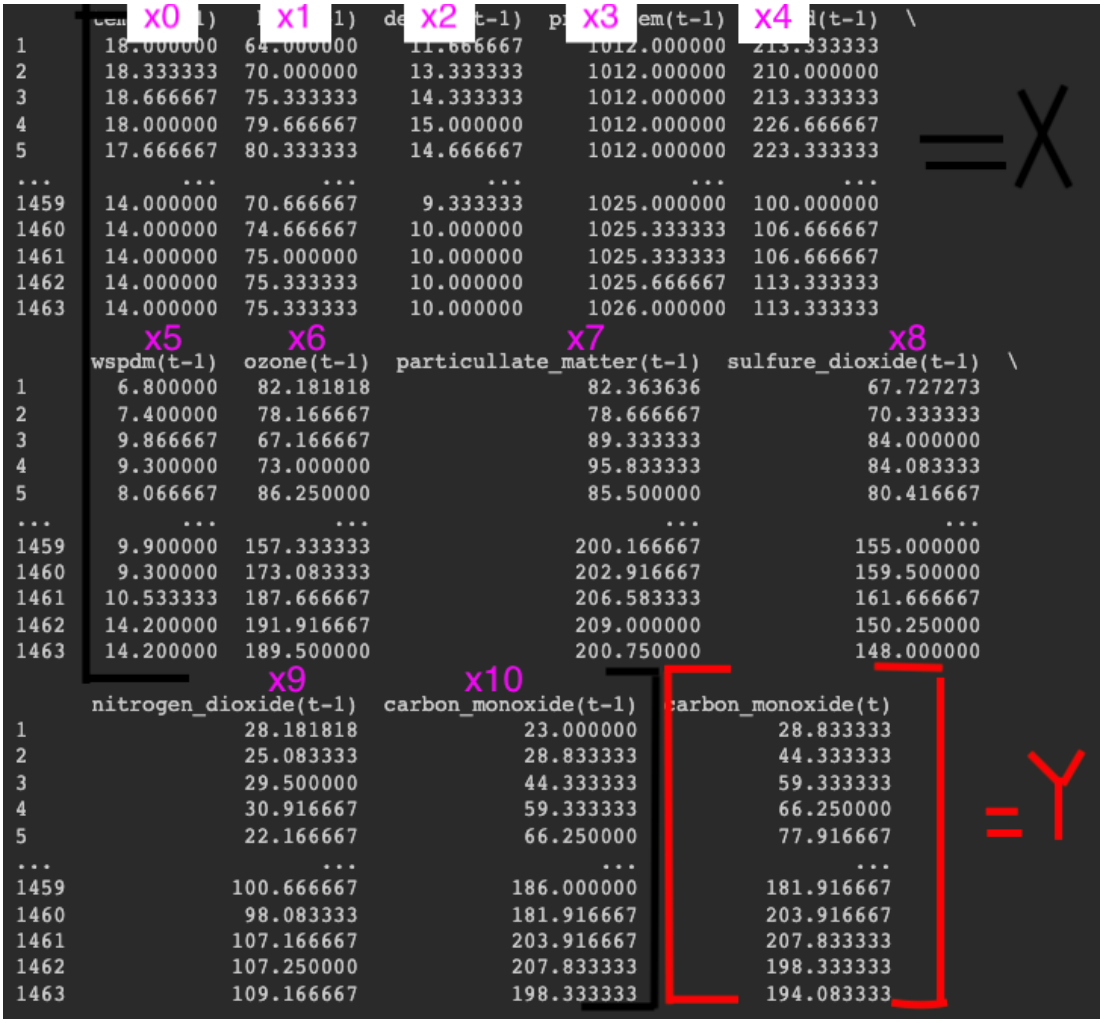
6.3.1 Απλή Εκθετική Εξομάλυνση (SES)

Μέσω της Βιβλιοθήκης Statsmodels παρέχεται η συνάρτηση SimpleExpSmoothing, η οποία δέχεται τα ιστορικά δεδομένα και προβλέπει την αμέσως επόμενη τιμή σε χρονικό ορίζοντα μίας ώρας.

6.3.2 Πολλαπλή Γραμμική Παλινδρόμηση (MLR)

Εδώ όπως αναφέρθηκε και στο θεωρητικό υπόβαθρο υπολογίζεται η πρόβλεψη μιας χρονοσειράς, δεδομένου ότι η τυχαία μεταβλητή της εξαρτάται από περισσότερες μεταβλητές

μέσω μια γραμμικής σχέσης. Έτσι το σύστημα μετατρέπει τα δεδομένα στη μορφή που φαίνεται παρακάτω:



	x0	x1	x2	x3	x4
1	18.000000	64.000000	11.666667	1012.000000	213.333333
2	18.333333	70.000000	13.333333	1012.000000	210.000000
3	18.666667	75.333333	14.333333	1012.000000	213.333333
4	18.000000	79.666667	15.000000	1012.000000	226.666667
5	17.666667	80.333333	14.666667	1012.000000	223.333333
...	...	...	...	...	...
1459	14.000000	70.666667	9.333333	1025.000000	100.000000
1460	14.000000	74.666667	10.000000	1025.333333	106.666667
1461	14.000000	75.000000	10.000000	1025.333333	106.666667
1462	14.000000	75.333333	10.000000	1025.666667	113.333333
1463	14.000000	75.333333	10.000000	1026.000000	113.333333

	x5	x6	x7	x8
	wspdm(t-1)	ozone(t-1)	particullate_matter(t-1)	sulfure_dioxide(t-1)
1	6.800000	82.181818	82.363636	67.727273
2	7.400000	78.166667	78.666667	70.333333
3	9.866667	67.166667	89.333333	84.000000
4	9.300000	73.000000	95.833333	84.083333
5	8.066667	86.250000	85.500000	80.416667
...	...	...	...	...
1459	9.900000	157.333333	200.166667	155.000000
1460	9.300000	173.083333	202.916667	159.500000
1461	10.533333	187.666667	206.583333	161.666667
1462	14.200000	191.916667	209.000000	150.250000
1463	14.200000	189.500000	200.750000	148.000000

	x9	x10	
	nitrogen_dioxide(t-1)	carbon_monoxide(t-1)	carbon_monoxide(t)
1	28.181818	23.000000	28.833333
2	25.083333	28.833333	44.333333
3	29.500000	44.333333	59.333333
4	30.916667	59.333333	66.250000
5	22.166667	66.250000	77.916667
...	...	...	...
1459	100.666667	186.000000	181.916667
1460	98.083333	181.916667	203.916667
1461	107.166667	203.916667	207.833333
1462	107.250000	207.833333	198.333333
1463	109.166667	198.333333	194.083333

Εικόνα 6.8 Τροποποιημένα δεδομένα

Αυτό γίνεται διότι από τη μαθηματική έκφραση το Y είναι η τιμή το carbon_monoxide για την στιγμή t, και οι μεταβλητές X_i είναι οι τιμές των μεταβλητών που θα συμπεριλάβουμε στο MLR για την χρονική στιγμή στιγμή t-1. Όπως είναι αναμενόμενο καμία τιμή της χρονικής περιόδου t των X_i δεν είναι γνωστή επομένως δεν μπορούμε να εκφράσουμε τη μαθηματική σχέση συναρτήσει αυτών των τιμών. Άλλωστε είναι και πιο λογικό να εκφράσουμε την τιμή της ζητούμενης μεταβλητής Y συναρτήσει των X_i της προηγούμενης χρονικής στιγμής. Έτσι λοιπόν δίνουμε στη συνάρτηση linear_model της sklearn τον πίνακα X και Y που έχουμε κάθε φορά, που ουσιαστικά με βάση την παραπάνω εικόνα ισχύει ότι data = [X | Y]. Η συνάρτηση εκπαίδευσης υπολογίζει τις παραμέτρους της γραμμικής σχέσης και έτσι όταν δώσουμε στο σύστημα ως είσοδο το X_i της t χρονικής στιγμής, αυτό υπολογίζει το Y της t+1 χρονικής

περιόδου. Επίσης κάθε νέα γραμμή του [X|Y] αναφέρεται στην επόμενη πραγματική χρονική στιγμή και αφού τα δεδομένα είναι πολλά, έχουμε πολλές τέτοιες στιγμές σειριακές.

Τέλος το σύστημα θα διαλέξει όσες μεταβλητές X_i θεωρεί ότι έχουν μεγάλη εξάρτηση με τη Y, αλλά είναι παράλληλα και ανεξάρτητες μεταξύ τους μεταβλητές, όπως ορίζει η θεωρία. Εδώ είναι που χρησιμοποιούνται τα στοιχεία που έχουν αναλυθεί παραπάνω.

6.3.3 Prophet

Το Prophet ασχολείται μόνο με τη ζητούμενη χρονοσειρά πρόβλεψης και λαμβάνει όλα τα ιστορικά δεδομένα και με βάση τον εσωτερικό της σχεδιασμό προχωράει στην πρόβλεψη της επόμενης χρονικής στιγμής.

6.3.4 ARIMA

Το ARIMA όπως έχει αναφερθεί προϋποθέτει η χρονοσειρά να είναι μη-στάσιμη. Έτσι κάθε φορά και ανάλογα με τη χρονοσειρά που προβλέπεται, χρησιμοποιείται η ανάλυση της στασιμότητας της χρονοσειράς, και εάν αυτή δεν είναι στάσιμη τότε μπορεί να χρησιμοποιηθεί το μοντέλο ARIMA. Σε αντίθετη περίπτωση δεν είναι εφικτή η χρήση του ARIMA μοντέλου, αλλά μπορεί να χρησιμοποιηθεί κάποιο ARMA μοντέλο. Εδώ η χρονοσειρά του μονοξειδίου του άνθρακα είναι μη στασιμη και μπορεί να προχωρήσει το σύστημα στην εκπαίδευση του μοντέλου. Σημειώνεται ότι στα στατιστικά μοντέλα, εκπαίδευση μοντέλου καλείται η διαδικασία εύρεσης των καλύτερων παραμέτρων της μαθηματικής σχέσης του μοντέλου. Και αφού μιλάμε για αυτο παλινδρομικό μοντέλο, δεν χρειάζεται κάποια άλλη χρονοσειρά παρά μόνο η ζητούμενη της πρόβλεψης. Έτσι η συνάρτηση ARIMA της Statsmodel βιβλιοθήκης δέχεται τη χρονοσειρά και προβλέπει την τιμή για χρονικό ορίζοντα μιας ώρας.

6.3.5 VAR

Όπως αναλύθηκε και στα ανώτερα κεφάλαια με τα VAR έχουμε περισσότερη γνώση γιατί εμπλουτίζουμε το μοντέλο όχι μόνο με την αυτοπαλινδρόμηση της ζητούμενης μεταβλητής, αλλά και με την αυτοπαλινδρόμηση των συσχετισμένων μεταβλητών με αυτήν. Έτσι το σύστημα υπολογίζει τις πιο συσχετισμένες, με τη ζητούμενη μεταβλητή, μεταβλητές και στη συνέχεια εκπαιδεύει το μοντέλο. Εκπαιδεύεται με τον X πίνακα των δεδομένων στην αρχικής τους μορφή (Εικόνα 6.1) και υπολογίζει τις τιμές όλων των μεταβλητών για την επόμενη χρονική στιγμή.

6.3.6 Random Forest

Εδώ θα χρησιμοποιήσουμε πάλι τα δεδομένα με τη μορφή της Εικόνας 6.8 που είδαμε παραπάνω. Και αυτό διότι πάλι η τιμή του μονοξειδίου του άνθρακα θέλουμε να θέσουμε ότι εξαρτάται από τις τιμές όλων των υπόλοιπων μεγεθών, την προηγούμενη χρονική στιγμή. Έτσι λοιπόν δημιουργούμε μέσω του Random Forest πολλαπλά δέντρα αποφάσεων που εξαρτάται από την παράμετρο nestimators που δέχεται η συνάρτηση RandomForestRegressor της sklearn. Τα δέντρα αυτά χτίζονται με βάση τις εγγραφές των δεδομένων εκπαίδευσης ως προς τις τιμές της ζητούμενης μεταβλητής. Με βάση τους estimators, όταν μια νέα σειρά δεδομένων έρθει στο σύστημα, τότε υπολογίζεται η τιμή όλων των estimators και βγαίνει ένας μέσος όρος των αποτελεσμάτων ως πρόβλεψη της ζητούμενης ποσότητας. Εδώ δίνεται στον χρήστη η ευκαιρία μέσω του Web App να πειραματιστεί με τον αριθμό των estimators ώστε να δει πως μεταβάλλεται η πρόβλεψη με βάση τις αλλαγμένες παραμέτρους.

6.3.7 SVR

Ο SVR δέχεται πάλι ως δεδομένα εκπαίδευσης τα πίνακα X της Εικόνας 6.8 και ως δεδομένα επικύρωσης εκπαίδευσης τις τιμές του Y. Αφού εκπαιδευτεί, τότε με βάση μια νέα είσοδο δεδομένων X_t προβλέπει την τιμή της Y_{t+1}. Για τη διαδικασία αυτή χρησιμοποιείται η συνάρτηση SVR της sklearn. Εδώ ο χρήστης μπορεί να επιλέξει τις παραμέτρους του C και του gamma, για να δει πως μεταβάλλεται η πρόβλεψη με βάση τις παραμέτρους αυτές.

6.3.8 LSTM

Για την ανάπτυξη του μοντέλου LSTM, χρησιμοποιείται το API του Tensorflow, το Keras. Με βάση το Keras θα χτιστεί η αρχιτεκτονική του LSTM, ωστόσο πριν από αυτό χρειάζεται μια μικρή προεπεξεργασία. Συγκεκριμένα πρέπει να χρησιμοποιηθεί ένας MinMaxScaler από τη βιβλιοθήκη sklearn, ώστε να μετατρέψει τα δεδομένα σε κλίμακα από 0 εώς 1. Αυτό συμβαίνει γιατί το μοντέλο LSTM είναι επιρρεπές σε μεγάλες μεταβολές τιμών και έτσι είναι ανάγκη να κάνουμε τα δεδομενα σε μια κλίμακα στην οποία μαθηματικά είναι κοντά. Επίσης τα δεδομένα πάλι θα χρησιμοποιηθούν με τη μορφή της Εικόνας 6.8.

Στη συνέχεια πρέπει να χωρίσουμε τα δεδομένα σε train και test. Συγκεκριμένα, το σύνολο των train δεδομένων χρησιμοποιείται για να βελτιστοποιήσει τις παραμέτρους τους συστήματος μέσα από την εκπαίδευση του μοντέλου. Παράλληλα χρησιμοποιείται και ένα

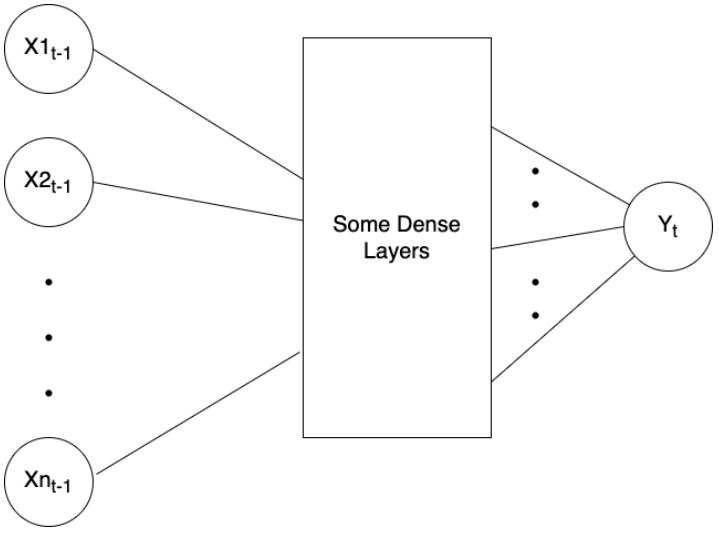
validation σύνολο δεδομένων το οποίο χρησιμοποιείται για την αποτίμηση του μοντέλου σε κάθε τέλος μιας εποχής. Την αποτίμηση αυτή την χρησιμοποιούμε για να καταλάβουμε εάν έχουμε κάποια εξέλιξη στην εκπαίδευση με το πέρας των εποχών ή όχι. Τέλος το test σύνολο δεδομένων αποτελεί ουσιαστικά και την πρόβλεψη της τελευταίας χρονικής στιγμής που βρίσκεται στα κύρια δεδομένα, καθώς έτσι μπορούμε να υπολογίσουμε και το τελικό accuracy του μοντέλου στα δεδομένα μας για χρονικό ορίζοντα μιας ώρας.

Τονίζεται ότι ο πίνακας X αποτελείται από τις μεταβλητές που έχουμε βρει κατά τη διαγνωστική ανάλυση ότι σχετίζονται περισσότερο με την ζητούμενη μεταβλητή πρόβλεψης. Αφού λοιπόν γίνεται η εκπαίδευση του μοντέλου, τότε με βάση την είσοδο των τιμών των μεταβλητών X για τη χρονική στιγμή t, αυτό υπολογίζει τη τιμή της Y τη χρονική στιγμή t+1. Το τελευταίο βήμα είναι να γίνει η αντίστροφη διαδικασία που έγινε στην αρχή με το scaling των δεδομένων, ώστε εν τέλη να λάβουμε την πρόβλεψη σε πραγματική τιμή.

Εδώ ο χρήστης μπορεί να πειραματιστεί με την αρχιτεκτονική του προβλήματος και συγκεκριμένα να ορίσει τα units του lstm, και τον αριθμό των ενδιάμεσων dense layers αλλά και τον αριθμό των κόμβων κάθε layer.

6.3.9 MLP

Το MLP χρησιμοποιεί ουσιαστικά Dense Layers από το API του Keras και έτσι χτίζεται η αρχιτεκτονική του. Τα δεδομένα πάλι εισαγωγής στο dense layer είναι οι μεταβλητές X_n που έχουμε βρει κατά τη διαγνωστική ανάλυση ότι σχετίζονται περισσότερο με τη ζητούμενη μεταβλητή πρόβλεψης και το αποτέλεσμα του τελικού κόμβου είναι η τιμή της ζητούμενης μεταβλητής. Συγκεκριμένα το μοντέλο έχει μια αρχιτεκτονική που μοιάζει κάπως έτσι:



Εικόνα 6.9 MLP Architecture

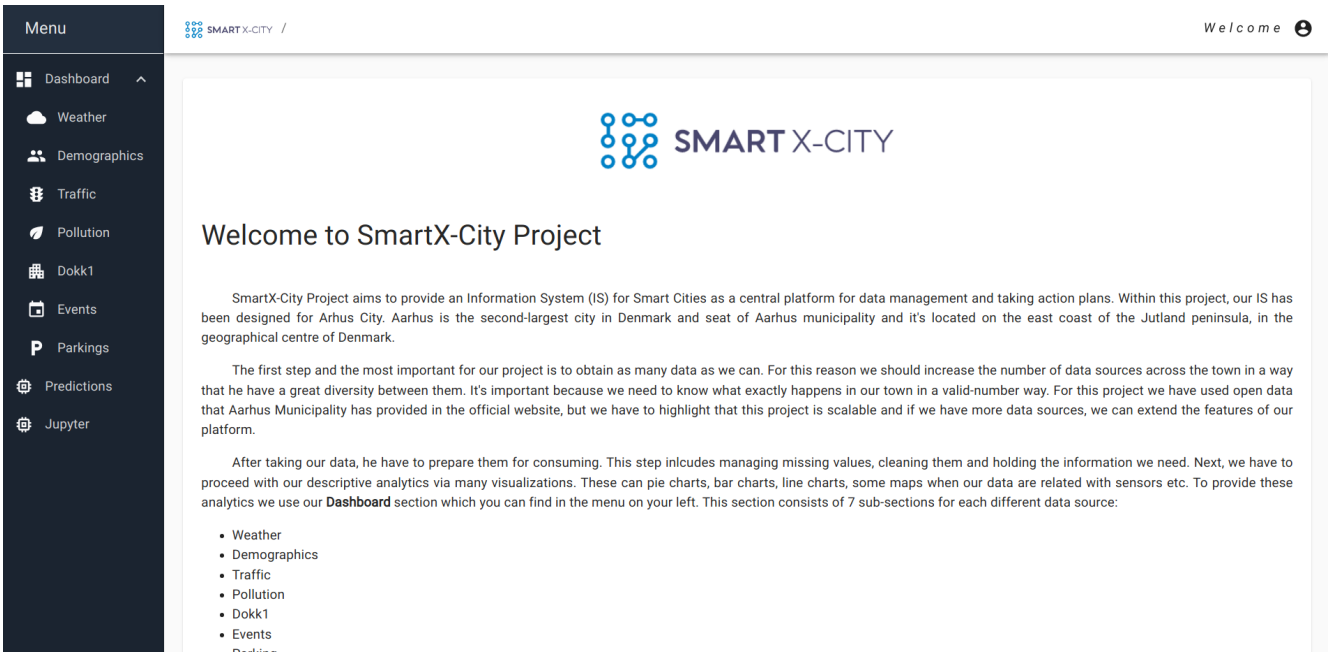
Βλέπουμε και μέσα από την εικόνα ότι πάλι τα δεδομένα είναι στη μορφή της Εικόνας 6.8 για τον ίδιο λόγο που έχει αναφερθεί και παραπάνω. Και εδώ ο αλγόριθμος εκπαιδεύεται με βάση ένα σύνολο δεδομένων εκπαίδευσης και αξιολογείται με βάση το σύνολο δεδομένων test. Εδώ το σύνολο test είναι τα δεδομένα της τελευταίας χρονικής στιγμής που βρίσκονται στα κύρια δεδομένα, και από το οποίο λαμβάνουμε τελικά και την ακρίβεια του μοντέλου.

Εδώ ο χρήστης του Web App μπορεί να πειραματιστεί με την αρχιτεκτονική όσων αφορά τον κουτί της εικόνας 6.9. Το κουτί αυτό μπορεί να αντικατασταθεί με κάποια dense layers που περιέχουν όσους κόμβους ο χρήστης επιθυμεί για να αξιολογήσει τη συμπεριφορά του μοντέλου συναρτήσει των κρυφών επιπέδων.

7 *Οδηγός Εργαλείου*

Στα προηγούμενα κεφάλαια αναλύθηκαν τόσο τα δεδομένα, αλλά και οι λειτουργίες που εκτελεί το core system για την ανάλυση αυτών και την υλοποίηση μοντέλων πρόβλεψης των χρονοσειρών. Το REST Api system λειτουργεί με έναν κλασικό τρόπο λειτουργίας ενός REST Api server και επομένως δεν χρήζει παραπάνω ανάλυσης πέρα της αναφοράς στο Κεφάλαιο 5. Το μόνο που μένει λοιπόν είναι να παρουσιαστεί σε αυτό το κεφάλαιο η Γραφική Διεπαφή Χρήστη με το σύστημα μέσω του Web App. Μάλιστα από τη στιγμή που είναι η μόνη διεπαφή του χρήστη με το σύστημά μας, χρήζει περαιτέρω ανάλυσης και παρουσίασης των δυνατοτήτων του για να μπορεί κάθε ένας να το χρησιμοποιήσει διαβάζοντας το κεφάλαιο αυτό. Παρατίθενται επίσης και διαγράμματα χρήσης και ακολουθιών που αποτελεί υλικό για την ανάλυση των Λειτουργικών Απαιτήσεων των εμπλεκόμενων μερών του συστήματος αυτού. Η παρουσίαση του εργαλείου περιλαμβάνει τόσο τα διαγράμματα αυτά όσο και στιγμιότυπα από το Web App.

7.1 Αρχική οθόνη και μενού



Εικόνα 7.1 Αρχική οθόνη

Βλέπουμε ότι η αρχική οθόνη περιλαμβάνει μια περιγραφή της εφαρμογής και αριστερά βρίσκεται το μενού περιήγησης του χρήστη. Μέσω αυτού μπορεί να χρησιμοποιήσει όλες τις λειτουργίες του εργαλείου. Το μενού χωρίζεται σε 3 ενότητες:

- Dashboard

  Περιλαμβάνει υπο ενότητες για κάθε διαφορετικό δεδομένο του συστήματος.

- Predictions

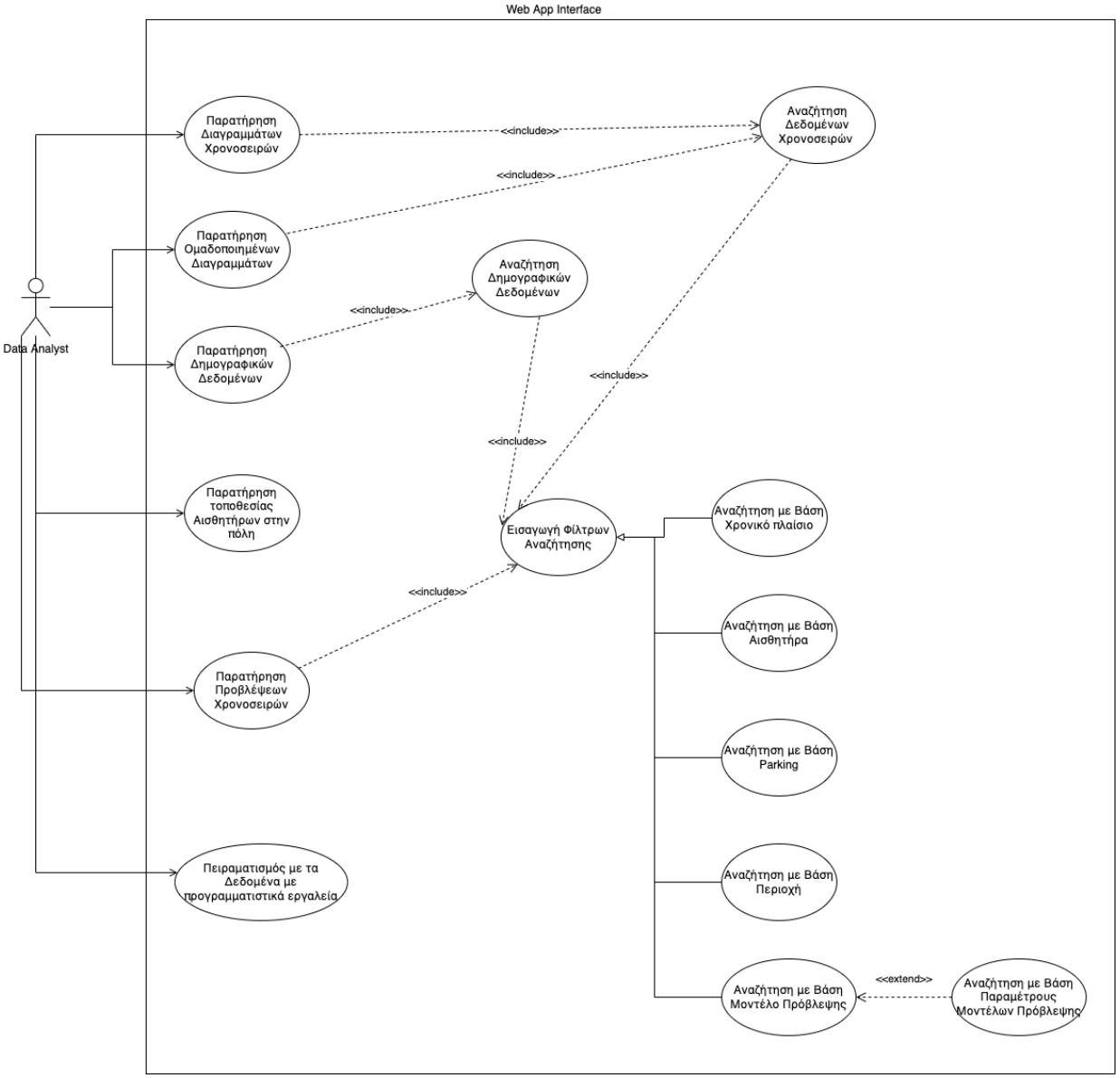
  Περιλαμβάνει την διεπαφή με τα μοντέλα πρόβλεψης των χρονοσειρών.

- Jupyter

  Περιλαμβάνει τη διεπαφή με το προγραμματιστικό εργαλείο, ώστε ο χρήστης να μπορεί να πειραματιστεί με τα δεδομένα και τα μοντέλα πρόβλεψης.

7.2 Περιπτώσεις Χρήσης Εφαρμογής

Στο ακόλουθο Use Case UML Diagram περιγράφονται οι περιπτώσεις χρήσης του εργαλείου μας με τους χρήστες αυτού, οι οποίες πραγματοποιούνται μέσα από την επιλογή στο μενού της εφαρμογής και αναφέρεται ουσιαστικά στις δυνατότητες που δίνει το εργαλείο αυτό στον χρήστη του.



Εικόνα 7.1 Use Case UML Diagram

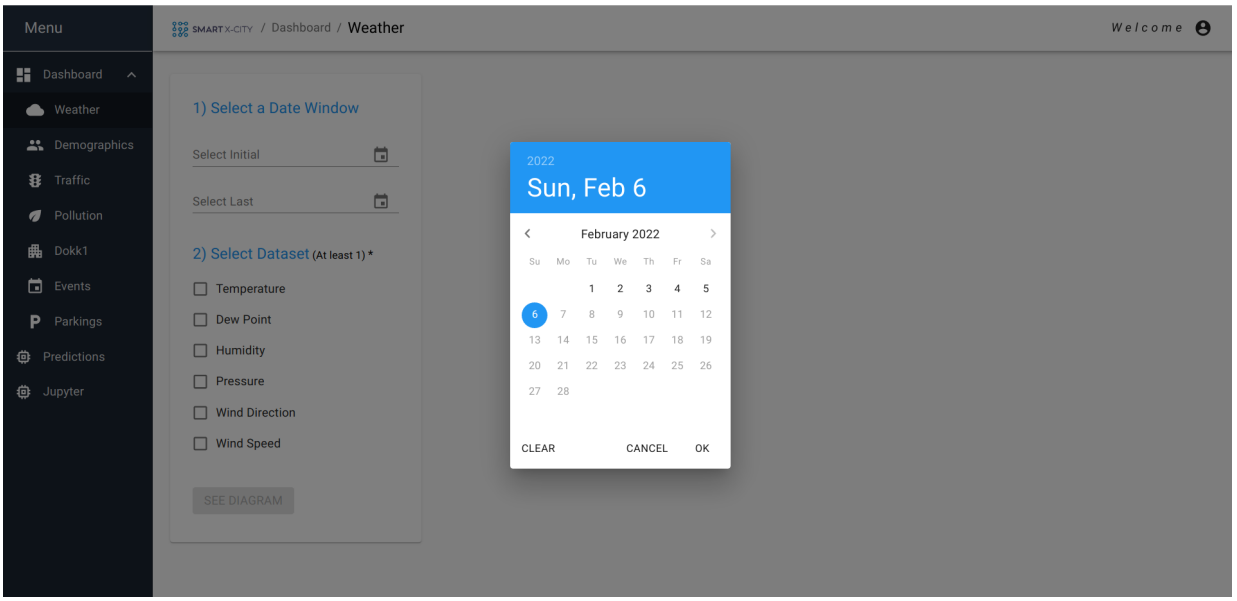
Ακολουθεί η ανάλυση των περιπτώσεων χρήσης μαζί με στιγμιότυπα οθονών.

7.3 Παρατήρηση Διαγραμμάτων Χρονοσειρών

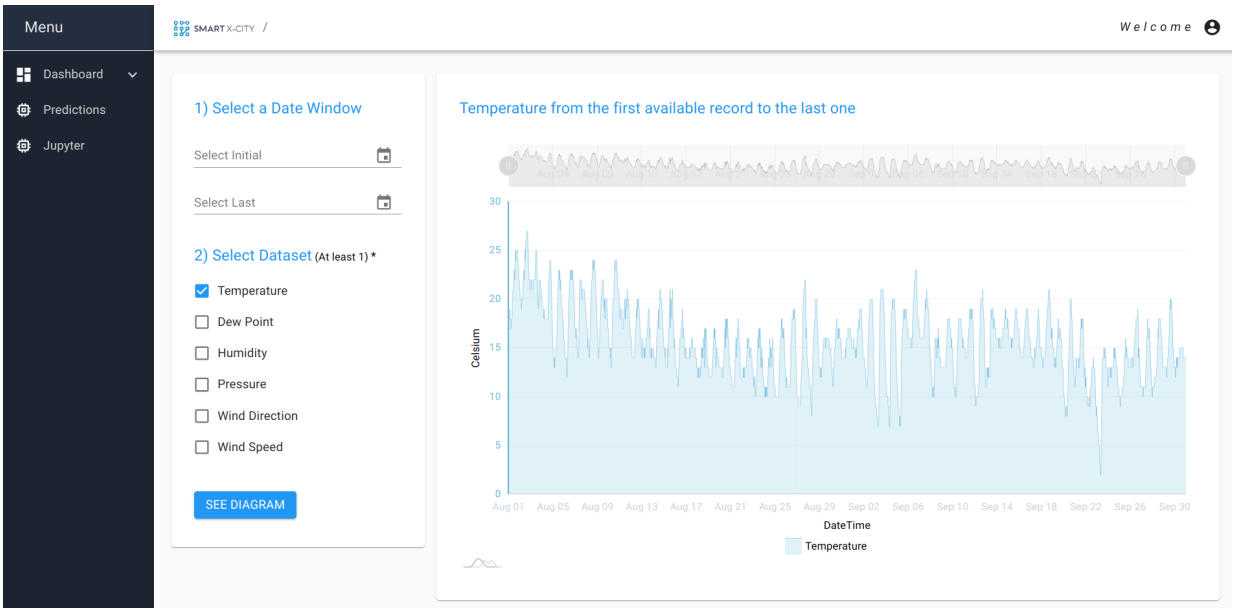
Το Web App αρχικά αποτελεί κατα κύριο λόγο ένα παρατηρητήριο των χρονοσειρών των διάφορων δεδομένων του συστήματος. Έτσι η παρατήρηση των διαγραμμάτων των χρονοσειρών, που αποτελούν την πλειοψηφία των δεδομένων είναι μια σημαντική λειτουργική απαίτηση του συστήματος που καλύπτεται. Παρέχεται μέσα από το μενού του Web app, στην υποενότητα του Dashboard και για κάθε δεδομένο που αποτελεί χρονοσειρά. Για την προβολή των δεδομένων απαιτείται και η συμπλήρωση των κατάλληλων φίλτρων αναζήτησης που κάθε φορά είναι διαφορετικό. Συγκεκριμένα:

- Μετεωρολογικά Δεδομένα

  Εδώ τα διαθέσιμα φίλτρα είναι το είδος των μετεωρολογικών δεδομένων και το χρονικό διάστημα για το οποίο ο χρήστης θέλει να δει τα δεδομένα.



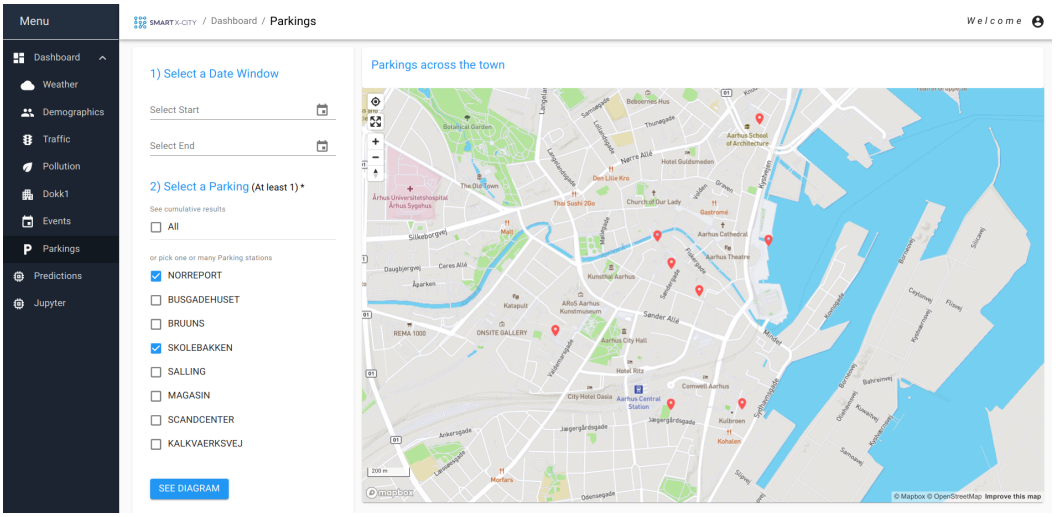
Εικόνα 7.3 Εισαγωγή Φίλτρου Χρονικού Πλαισίου



Εικόνα 7.3 Χρονοσειρά Μετεωρολογικών Δεδομένων

- Δεδομένα των χώρων στάθμευσης

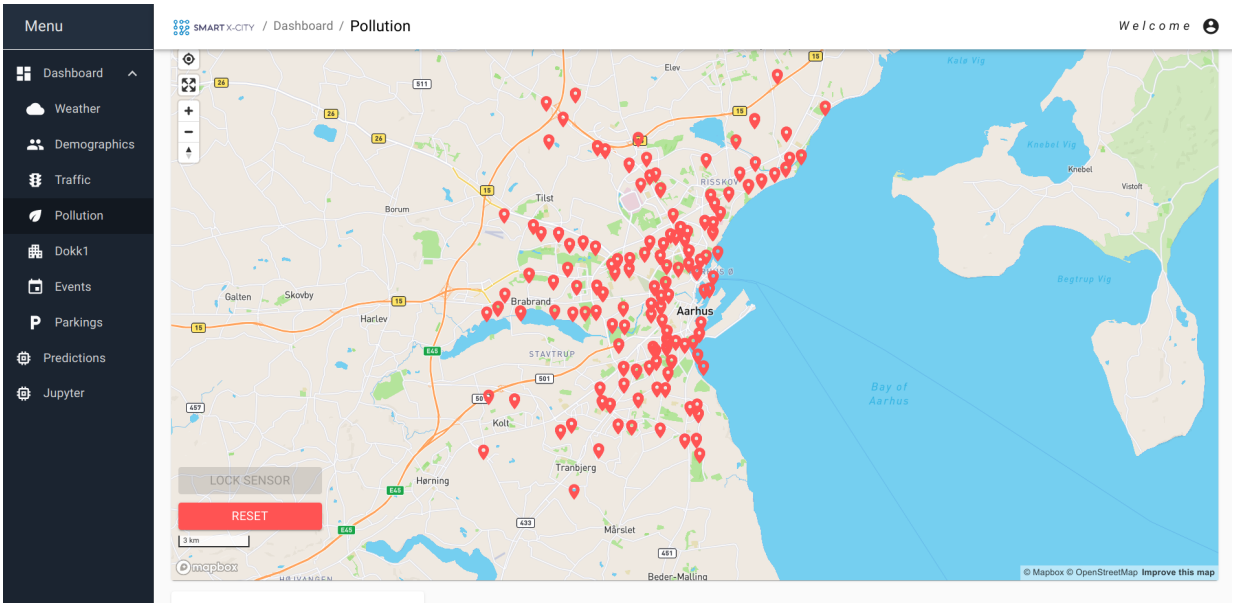
  Εδώ ο χρήστης επιλέγει το διαθέσιμο Parking ή και περισσότερα για τα οποία θέλει να αντλήσει τα δεδομένα και το χρονικό πλαίσιο και λαμβάνει τις κατειλημμένες θέσεις και τις κενές θέσεις.



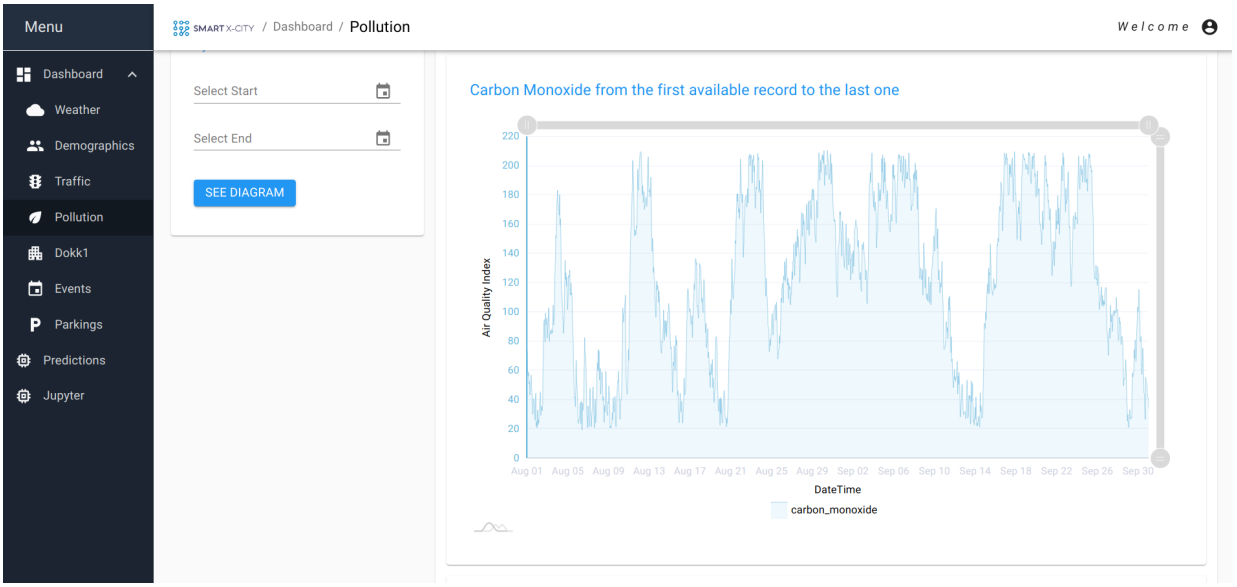
Εικόνα 7.4 Επιλογή Φίλτρων Για το Parking

- Ατμοσφαιρική Ρύπανση

  Εδώ ο χρήστης επιλέγει τον αισθητήρα μέσω του χάρτη που θέλει να δει τις μετρήσεις του και το χρονικό πλαίσιο και λαμβάνει τις χρονοσειρές των ρυπαντών.



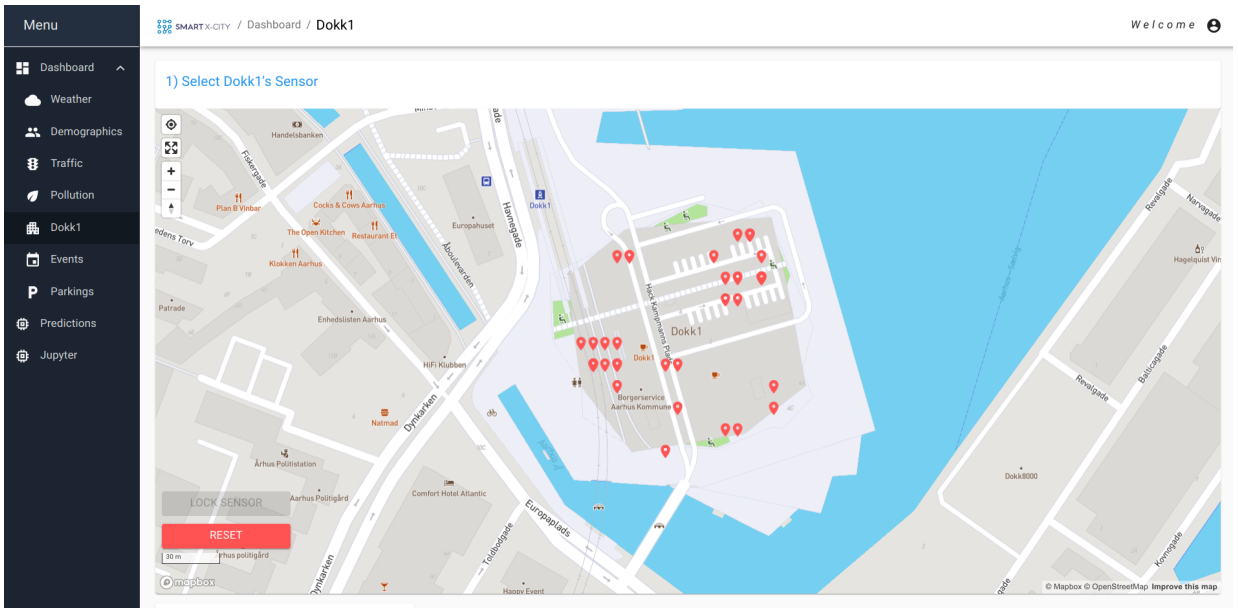
Εικόνα 7.5 Επιλογή Φίλτρων Για το Pollution



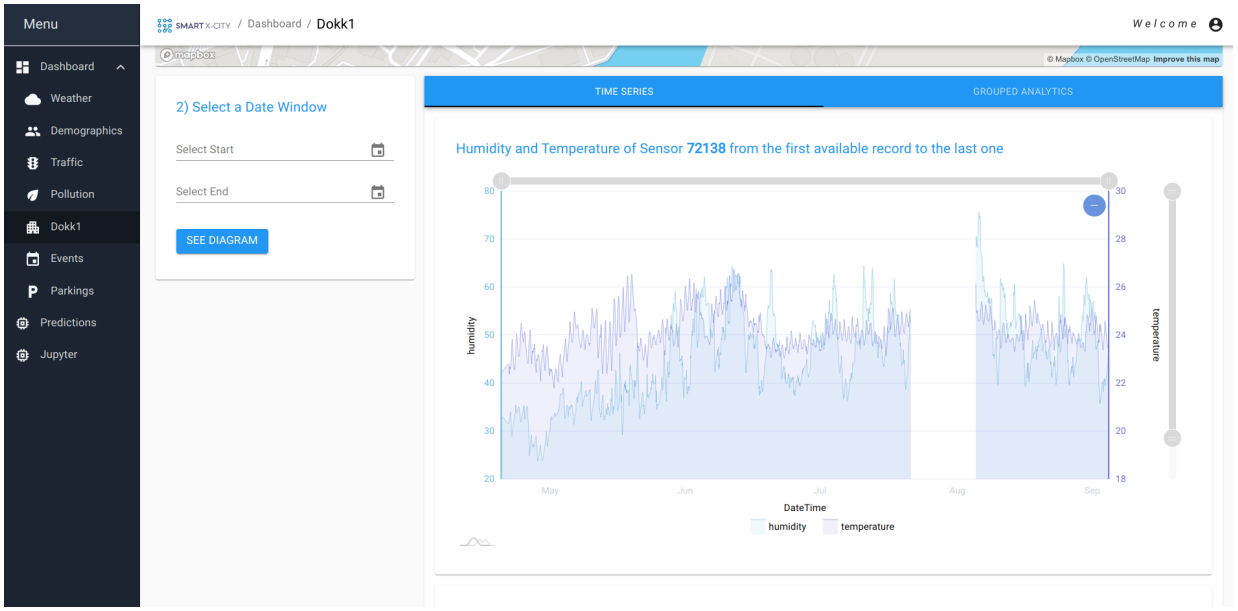
Εικόνα 7.6 Χρονοσειρά για τους ρυπαντές

- Δεδομένα Αισθητήρων στο κτίριο DOKK1

  Εδώ ο χρήστης επιλέγει τον αισθητήρα μέσω του χάρτη που θέλει να δει τις μετρήσεις του και το χρονικό πλαίσιο και λαμβάνει τις χρονοσειρές των διαθέσιμων μεγεθών.



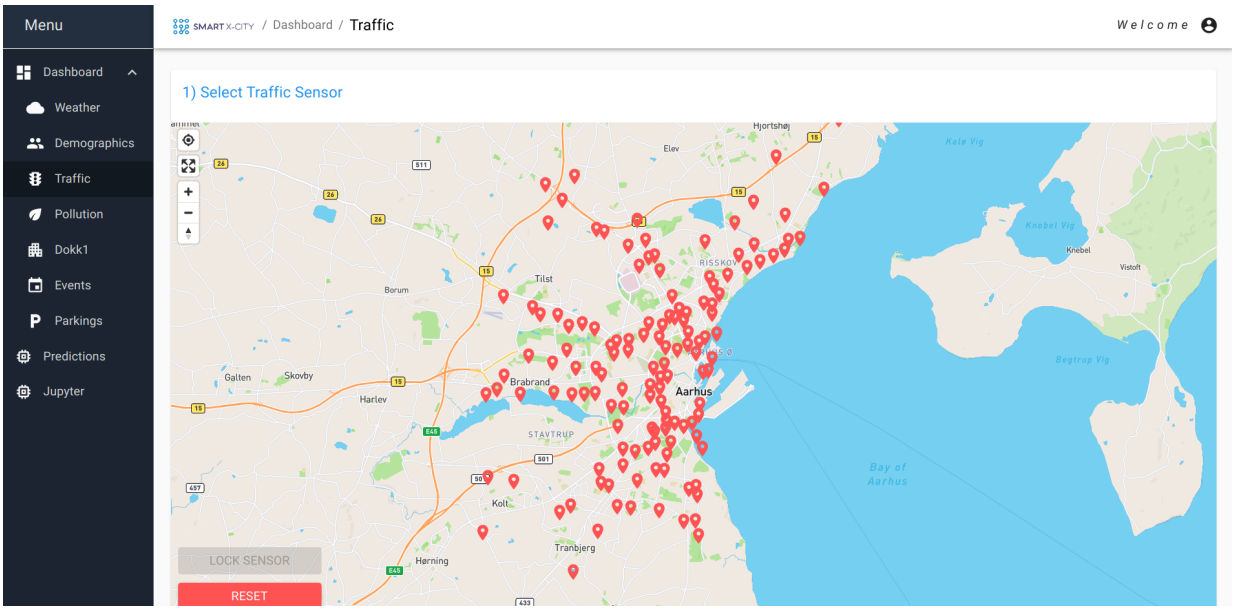
Εικόνα 7.7 Επιλογή Φίλτρων Για τον αισθητήρα μέτρησης στο Dokk1



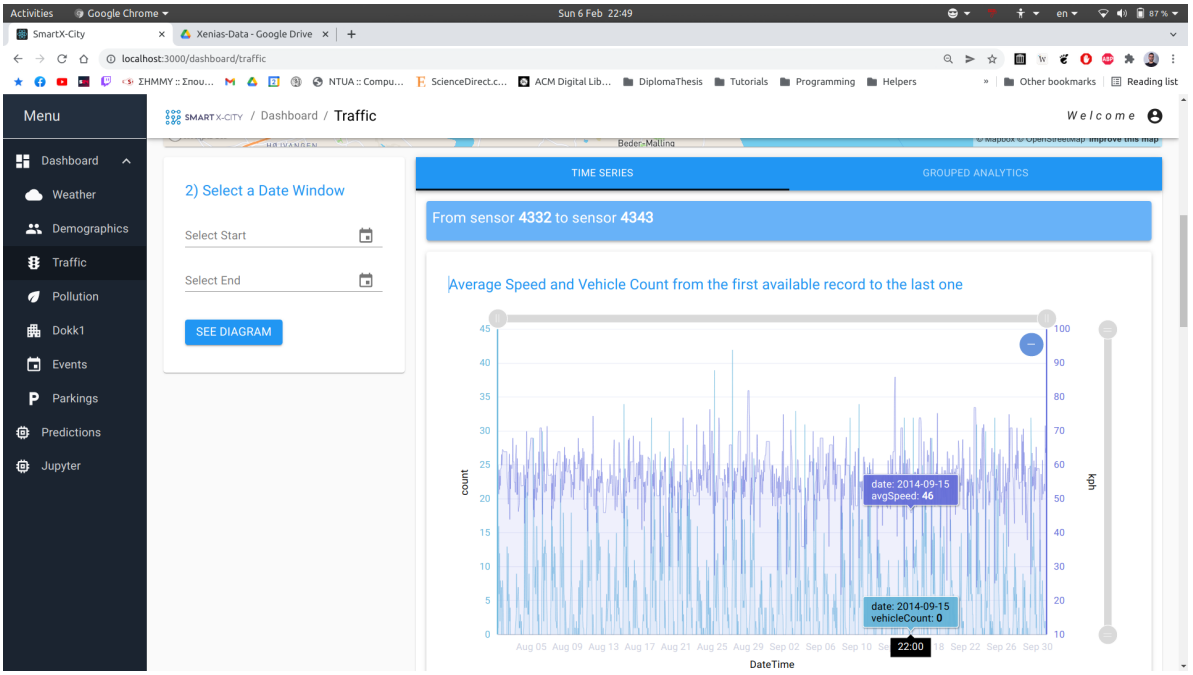
Εικόνα 7.8 Χρονοσειρά για τα διαθέσιμα μεγέθη στο Dokk1

- Δεδομένα κίνησης στους δρόμους

  Εδώ ο χρήστης επιλέγει στο χάρτη έναν αισθητήρα και λαμβάνει για το χρονικό πλαίσιο που εισήγαγε ο χρήστης τη κίνηση που υπάρχει μεταξύ του επιλεγμένου αισθητήρα και των γειτονικών του.



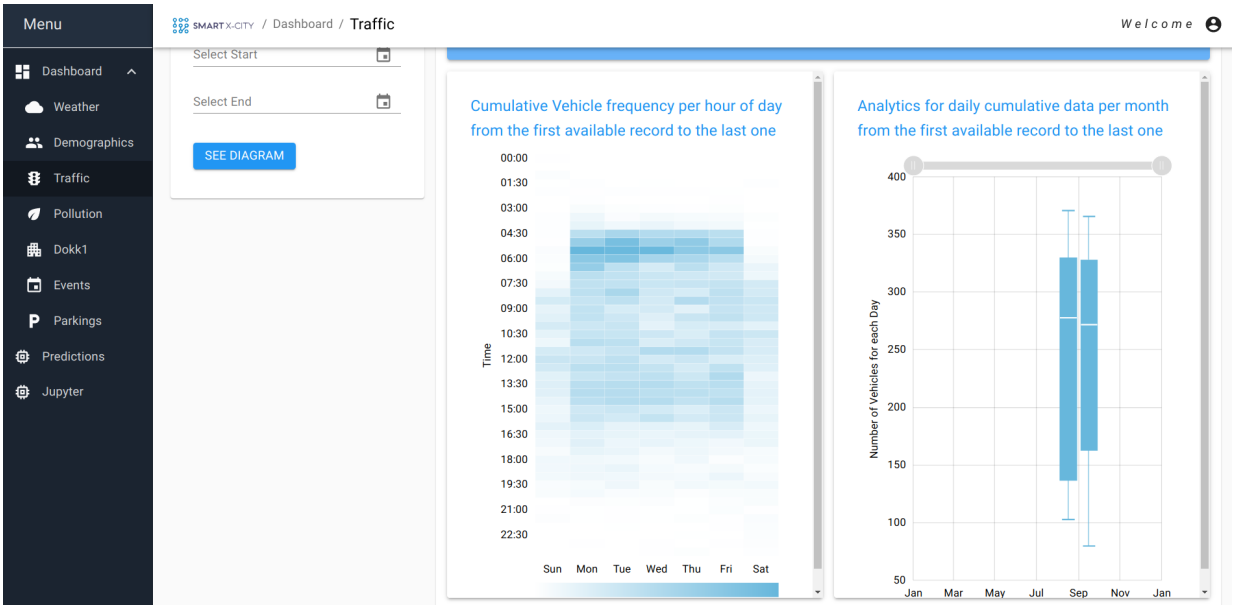
Εικόνα 7.9 Επιλογή Φίλτρων Για τον αισθητήρα μέτρησης της Κίνησης



Εικόνα 7.10 Χρονοσειρά για την κίνηση μεταξύ του επιλεγμένου αισθητήρα και ενός γειτονικού

7.4 Παρατήρηση Ομαδοποιημένων Διαγραμμάτων

Όλα τα άνωθεν διαγράμματα προσφέρονται για τον σκοπό της Περιγραφικής Ανάλυσης των Δεδομένων του συστήματος. Ωστόσο το εργαλείο πηγαίνει ένα βήμα παραπέρα και προσφέρει πιο ομαδοποιημένα διαγράμματα των δεδομένων, που θα βοηθήσουν τον χρήστη να καταλάβει καλύτερα το "τί γίνεται" και προφανώς πιο εύκολα να προχωρήσει στη Διαγνωστική Ανάλυση, δηλαδή στο "γιατί γίνεται¨ αυτό. Για τις παραπάνω χρονοσειρές τα ομαδοποιημένα διαγράμματα είναι ίδια και επομένως για συντομία παρατίθεται ένα στιγμιότυπο διαγράμματος για την κίνηση που καταγράφεται μεταξύ δύο γειτονικών αισθητήρων.



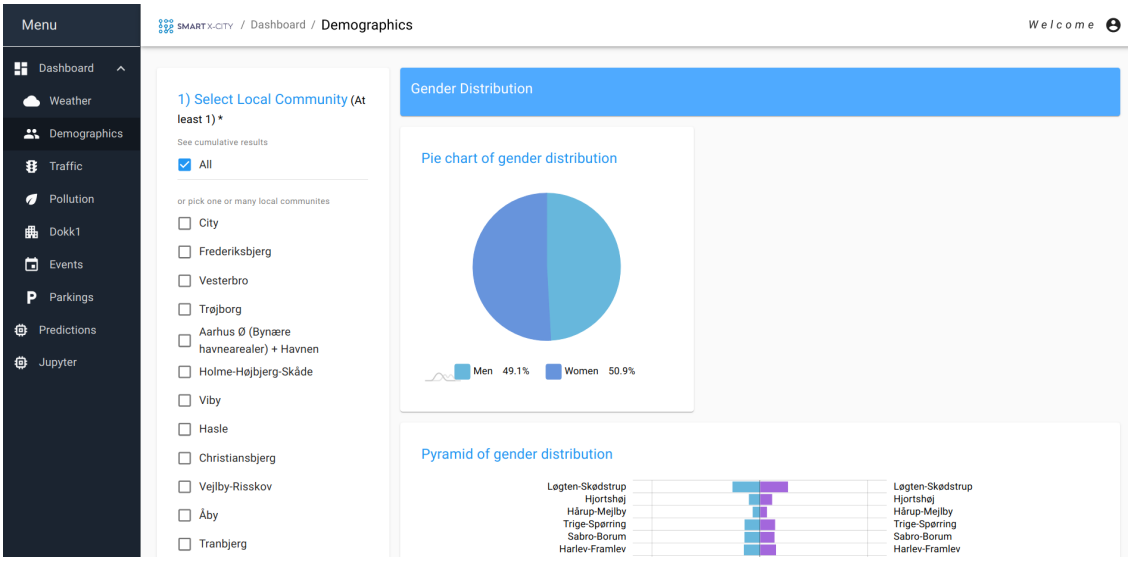
Εικόνα 7.11 Ομαδοποιημένα διαγράμματα Χρονοσειρών

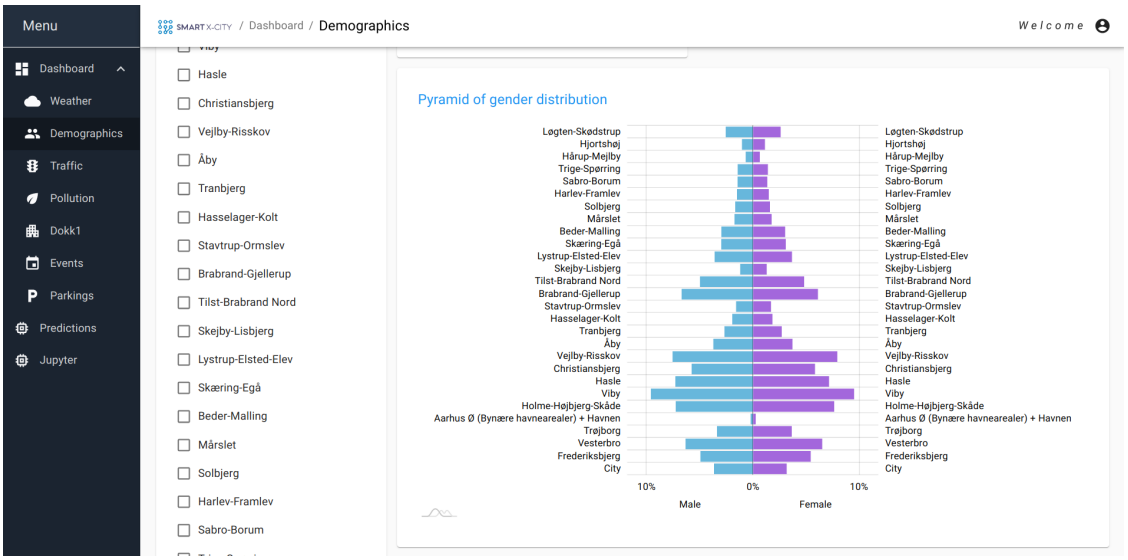
Συγκεκριμένα στο αριστερό διάγραμμα παρουσιάζονται τα δεδομένα ανά μισάωρο σε ένα διάγραμμα συχνότητας. Συγκεκριμένα, είναι ένας πίνακας με γραμμές τις ημέρες της εβδομάδας και στήλες τις ώρες της ημέρας. Έτσι ένα τετραγωνάκι αποτελεί ένα μέσο όρο του αριθμού των αυτοκινήτων που υπήρξε εκείνο το μισάωρο, εκείνη την ημέρα, και αυτό απεικονίζεται σε μια κλίμακα χρώματος, στην οποία όσο πιο έντονο είναι το χρώμα τόσο μεγαλύτερη κίνηση υπήρχε.

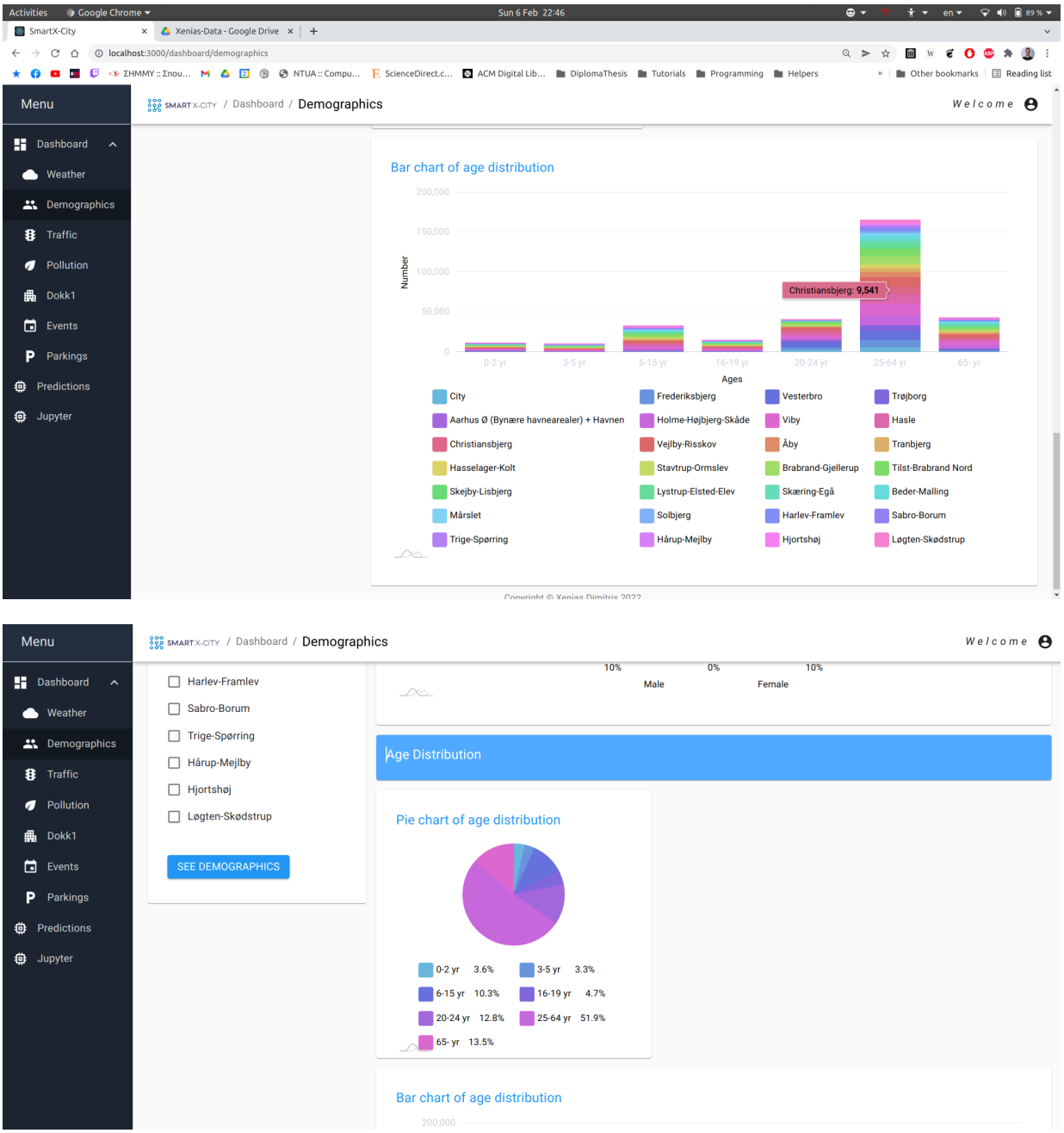
Στο δεξιά διάγραμμα παρακολουθούμε τα Quartiles όπως αναλύθηκαν στη θεωρία μαζί με τον μέσο όρο τιμών, την ελάχιστη τιμή και τη μέγιστη τιμή. Όλα αυτά τα quartiles είναι ομαδοποιημένα ανά μήνα και βλέπουμε κάπως την εποχικότητα εάν υπάρχει τον δεδομένων.

7.5 Παρατήρηση Δημογραφικών Δεδομένων

Την σημασία των δημογραφικών δεδομένων την αναφέραμε στο Κεφάλαιο 4.6. Επομένως εδώ ο χρήστης μπορεί να διαλέξει το δήμο που επιθυμεί ως φίλτρο και να λάβει διαγράμματα με την κατανομή των πολιτών σε ηλικία και φύλλο.





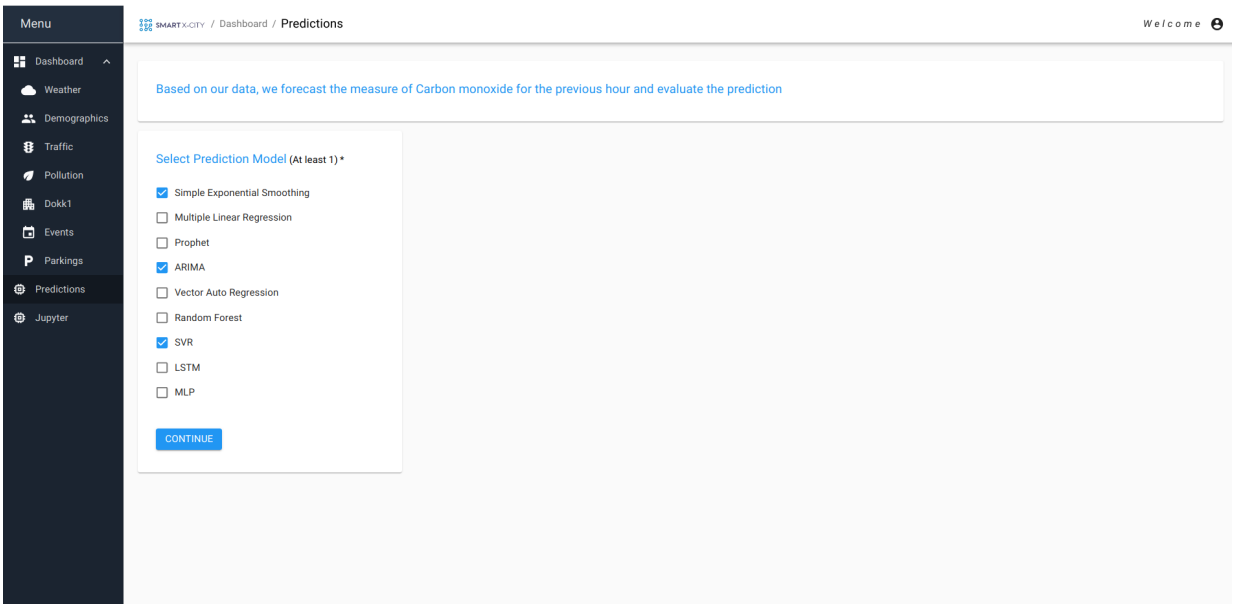


Εικόνα 7.12 Διαγράμματα Δημογραφικών Χαρακτηριστικών

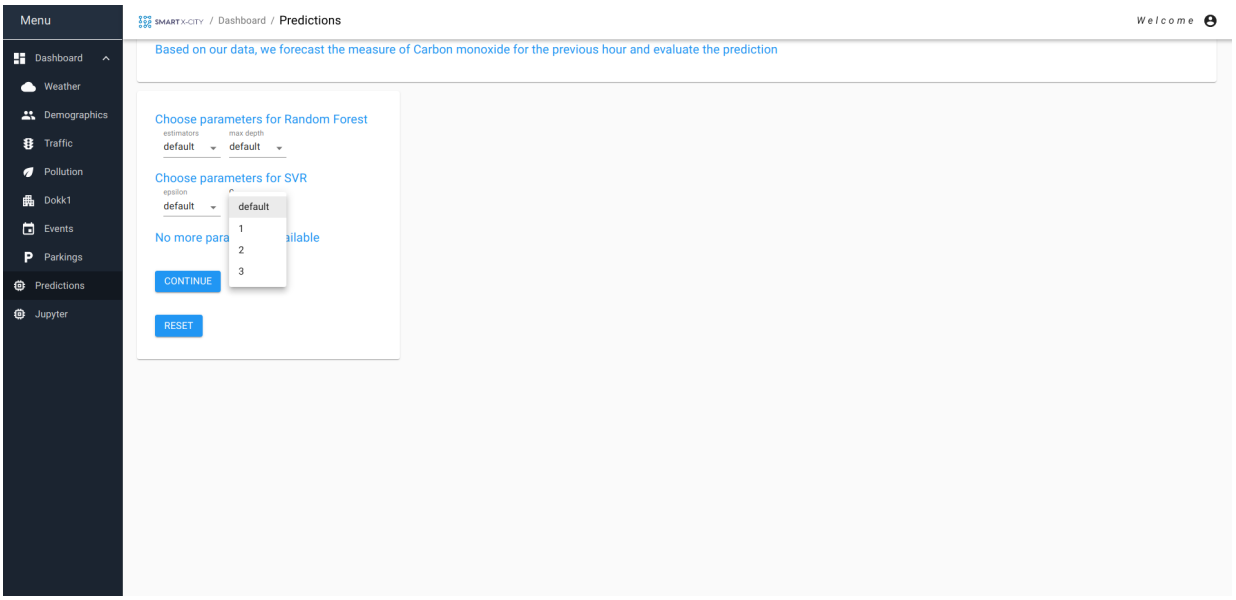
7.6 Παρατήρηση Προβλέψεων Χρονοσειρών

Στη Ενότητα αυτή ο χρήστης μπορεί να δει τις προβλέψεις της χρονοσειράς του μονοξειδίου του άνθρακα με βάση τα στατιστικά μοντέλα και τα μοντέλα μηχανικής μάθησης, που το core system εκπαιδεύει, για την επόμενη ώρα. Για διδακτικούς σκοπούς θεωρούμε την τελευταία διαθέσιμη μέτρηση ως μέτρηση για να τσεκάρουμε τις προβλέψεις του συστήματος, και επομένως στο τμήμα αυτό παρουσιάζεται και μια αξιολόγηση των επιλεγμένων μοντέλων ως προς την πρόβλεψή τους. Ο χρήστης αρχικά επιλέγει για ποια διαθέσιμα μοντέλα θέλει να λάβει πρόβλεψη. Στη συνέχεια, του ζητάται να

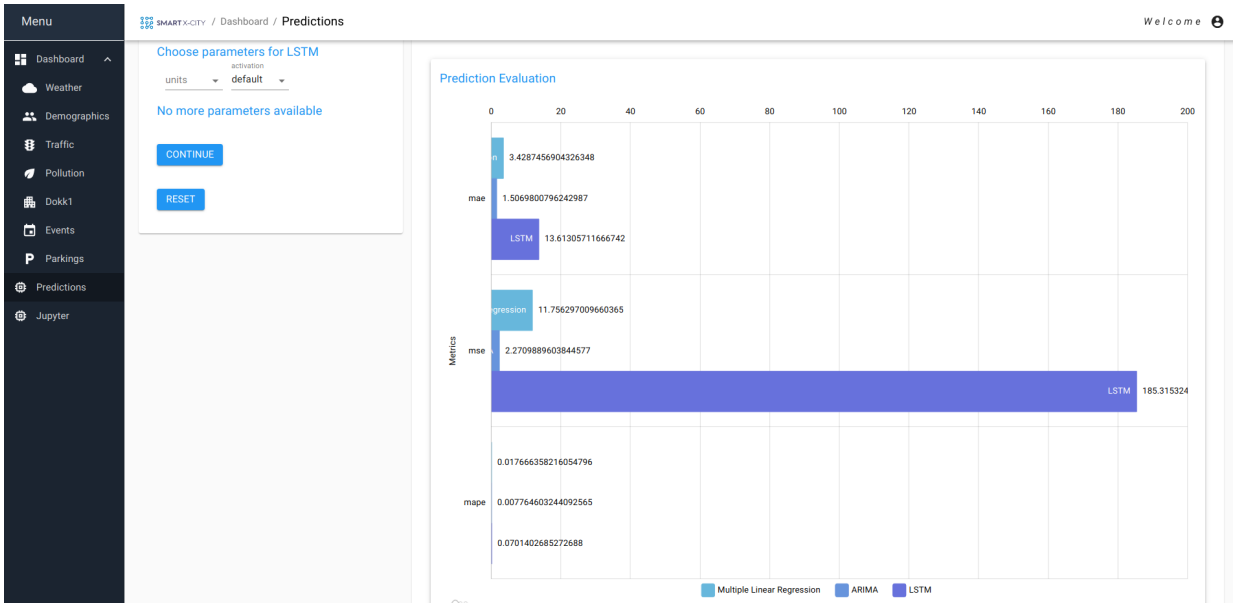
θέσει τις προσωπικές του επιλογές για τις παραμέτρους των μοντέλων, εάν επιθυμεί να πειραματιστεί με αυτές, και φυσικά εάν διατίθεται ως επιλογή για τα επιλεγμένα μοντέλα. Εάν δεν τεθούν συγκεκριμένες παράμετροι από τον χρήστη μέσω τη φόρμας, τότε το σύστημα επιστρέφει την πρόβλεψη για τις μέχρι τότε βέλτιστες αρχιτεκτονικές των μοντέλων που έχουν υπολογιστεί από το core system. Έτσι μόλις υπολογιστούν οι προβλέψεις, εμφανίζονται στο χρήστη. Εδώ εμφανίζεται μάλιστα και η αξιολόγηση τους, για διδακτικούς λόγους.



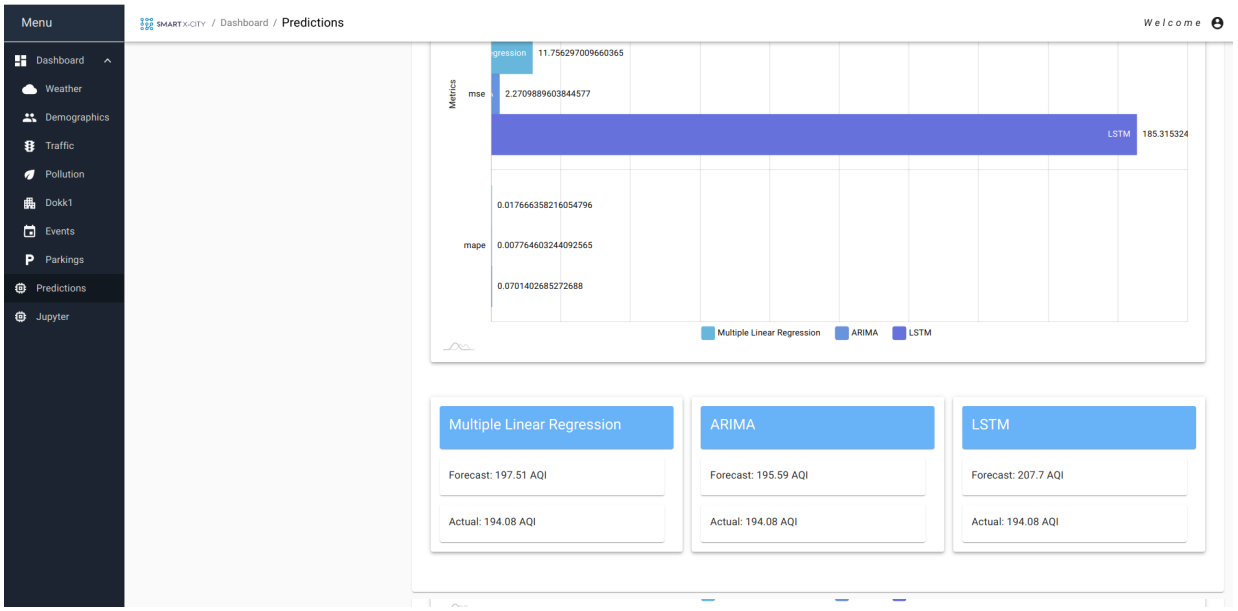
Εικόνα 7.13 Επιλογή Διαθέσιμων μοντέλων



Εικόνα 7.14 Επιλογή προσωπικών παραμέτρων στα μοντέλα Μηχανικής Μάθησης



Εικόνα 7.15 Διάγραμμα των μετρικών αξιολόγησης μοντέλων



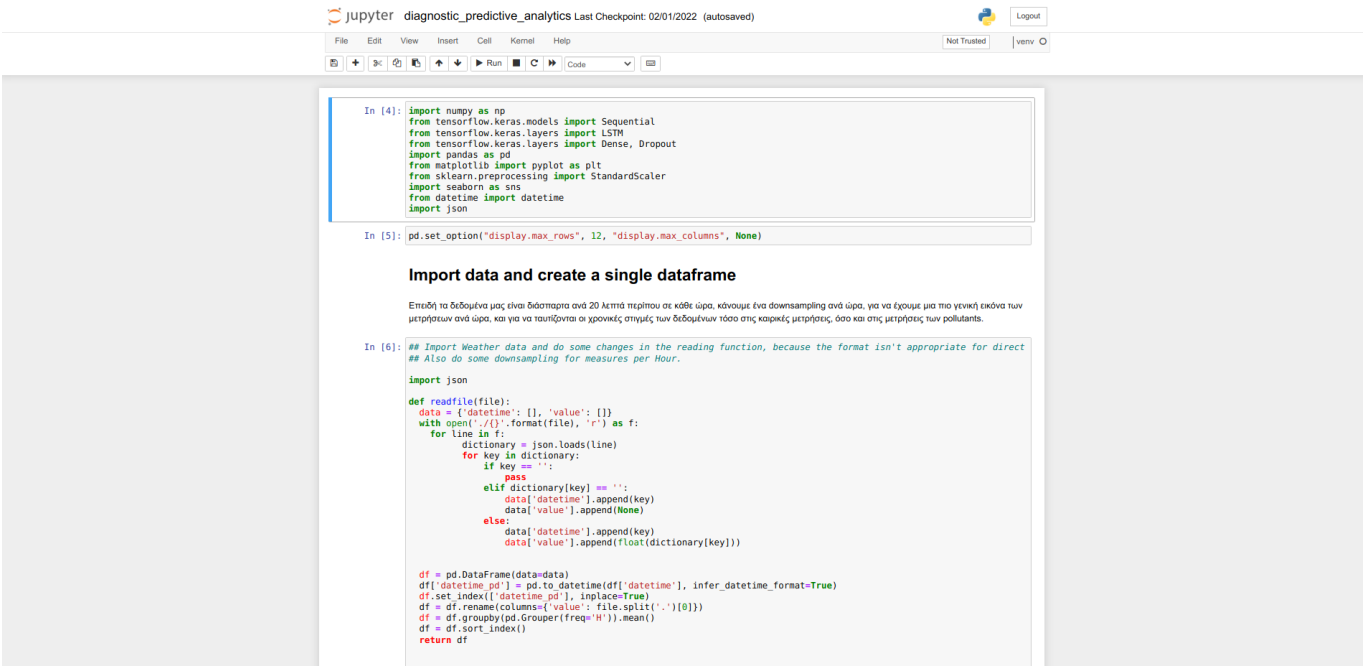
Εικόνα 7.16 Πραγματική και Προβλεπόμενη Τιμή

7.7 Παρατήρηση τοποθεσίας Αισθητήρων στην πόλη

Μέσα από τις εικόνες 7.5, 7.7, 7.9, ο χρήστης μπορεί να δει την τοποθεσία των αισθητήρων ρύπανσης, του Dokk1 και της κίνησης μέσα στην πόλη και να λάβει πληροφορίες τους.

7.8 Πειραματισμός με τα Δεδομένα μέσω προγραμματιστικής διεπαφής

Τελευταία περίπτωση χρήσης είναι ο χρήστης να πάει στην τρίτη επιλογή του μενού “Jupyter”, μέσω της οποίας συνδέεται σε μια προγραμματιστική διεπαφή του core συστήματος μέσω του jupyter notebook. Συγκεκριμένα βρίσκεται ενσωματωμένο στο Web App μια διεπαφή με έναν jupyter notebook server, ο οποίος δίνει τη δυνατότητα στον χρήστη να πειραματιστεί με τα δεδομένα μέσω προγραμματιστικής γλώσσας python και αποτελεί ουσιαστικά ένα “playground” για τα δεδομένα.



Εικόνα 7.17 Jupyter Notebook Server

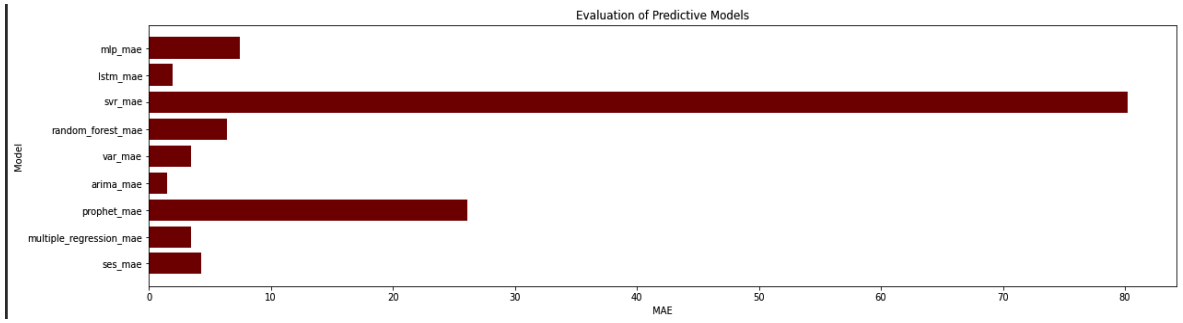
8 Αποτελέσματα και Συμπεράσματα

Όπως αναφέρθηκε και στην υπο ενότητα 2.2.3.4, η αξιολόγηση των μοντέλων πρόβλεψης επιτυγχάνεται μέσα από κάποιες μετρικές που εκφράζουν το πόσο ακριβή είναι τα αποτελέσματα των μοντέλων στις προβλέψεις τους. Έτσι ο χρήστης του συστήματος μπορεί να έχει ένα μέτρο σύγκρισης, αλλά και ένα μέτρο εμπιστοσύνης προς το κάθε μοντέλο και να πράξει αναλόγως. Επίσης στο Κεφάλαιο 6 αναφέραμε τον τρόπο με τον οποίο το core system αναλύει κάθε φορά τα δεδομένα που διαθέτει και εκπαιδεύει αλγορίθμους στατιστικής και μηχανικής μάθησης, ενώ στο Κεφάλαιο 7 είδαμε τον τρόπο που διαθέτει τα αποτελέσματα των αλγορίθμων αυτών στον χρήστη, μαζί με τις μετρικές αξιολόγησης και ακρίβειας. Ας μελετήσουμε όμως στο Κεφάλαιο αυτό τα αποτελέσματα των αλγορίθμων.

Αρχικά είναι γνωστό ότι όσα περισσότερα δεδομένα διαθέτει ένα σύστημα, τόσο μεγαλύτερο εύρος πληροφοριών έχει και συνεπώς μπορεί καλύτερα να εκπαιδευτεί και να βελτιωθεί στην πορεία. Τα αποτελέσματα επομένως της ακρίβειας των αλγορίθμων πρόβλεψης εξαρτώνται από το πλήθος αλλά και την ποιότητα και ποικιλομορφία των δεδομένων. Επομένως εδώ μελετάμε τα αποτελέσματα των αλγορίθμων με βάση τα δεδομένα που διαθέτουμε για το πειραματικό αυτό στάδιο. Στην πορεία και σε ένα πραγματικό σύστημα είναι αρκετά πιθανό να δούμε καλύτερα νούμερα ακρίβειας στις προβλέψεις και ίσως κάποιοι αλγόριθμοι να είναι πιο αποτελεσματικοί στην πορεία. Για αυτό το λόγο προσφέρεται και μια τόση μεγάλη γκάμα αλγορίθμων από το σύστημά μας, διότι έτσι έχουμε μια γενικευμένη

άποψη για τα αποτελέσματα καθενός από αυτούς, και στο μέλλον μπορούμε να λαμβάνουμε διαφορετικό κάθε φορά ως καλύτερο.

Στην προκειμένη περίπτωση λοιπόν για τα δεδομένα που έχουμε στο σύστημα, αξιολογούμε την ικανότητα των 9 αλγορίθμων πρόβλεψης στο να προβλέπουν την τιμή του μονοξειδίου του άνθρακα στην επόμενη χρονική στιγμή. Τα αποτελέσματα ακρίβειας των προβλέψεων φαίνονται στον κατώτερο πίνακα.



Εικόνα 8.1 Αξιολόγηση στα Μοντέλα Πρόβλεψης

Όπως ήταν αναμενόμενο τα στατιστικά μοντέλα είναι αρκετά ακριβή στις προβλεψεις τους, διότι η χρονοσειρά είναι διαθέσιμη για ένα μικρό χρονικό διάστημα και επομένως αυτοπαλινδρομικά και παλινδρομικά μοντέλα μπορούν να προσαρμοστούν και να προβλέψουν κάτι αρκετά ακριβές για την επόμενη χρονική στιγμή. Ο ARIMA μάλιστα πετυχαίνει το καλύτερο σκορ ακρίβειας στις προβλέψεις του (MAE), επιβεβαιώνοντας όσους χρησιμοποιούν τον αλγόριθμο αυτό για τέτοιες προβλέψεις και καθιστώντας τον ως έναν από τους πιο χρήσιμους αλγορίθμους πρόβλεψης χρονοσειρών.

Αντίθετα το SVR φαίνεται να πετυχαίνει τις χειρότερες τιμές πρόβλεψης, κάτι το οποίο σημαίνει ότι είναι πολύ δύσκολο να βρεθεί ένα hyperplane για τον αλγόριθμο σε μικρό χρόνο, και ειδικά για προβλήματα χρονοσειρών που τα δεδομένα ολοένα και αυξάνονται, ο SVR φαίνεται μη αποτελεσματικός σε προβλέψεις εντός εύλογου χρονικού διαστήματος.

Οι υπόλοιποι αλγόριθμοι βρίσκονται σχετικά κοντά σε ακρίβεια πρόβλεψης. Ωστόσο πρέπει να τονιστεί ότι οι αλγόριθμοι lstm και mlp, και κυρίως ο πρώτος, ανταγωνίζονται πολύ τα στατιστικά μοντέλα, παρότι λειτουργούν με άλλη φιλοσοφία και είναι αλγόριθμοι μηχανικής μάθησης. Εκτιμάται ότι με περισσότερα δεδομένα στο σύστημα, οι αλγόριθμοι αυτοί θα ξεπερνούν σε ακρίβεια και τον ARIMA, λόγω της ικανότητάς τους να δρουν σε μη στάσιμες

χρονοσειρές και να αντλούν περισσότερη πληροφορία, που οι ARIMA χάνουν κατά τη μετατροπή των χρονοσειρών σε στάσιμες.

Επίσης άλλο ένα πολύ σημαντικό αποτέλεσμα της διπλωματικής, είναι η ανάγκη για την ύπαρξη πολλών παράλληλων συστημάτων στην αρχιτεκτονική τέτοιων project. Πολλές φορές η εκπαίδευση αλγορίθμων μπορεί να είναι αρκετά χρονοβόρα και ο χρήστης δεν μπορεί να περιμένει για αυτά. Χρειάζονται λοιπόν συστήματα που θα είναι αποκλειστικά υπεύθυνα για την εκπαίδευση αλγορίθμων και θα αφιερώνουν όλες τις πηγές τους σε αυτή τη διαδικασία. Παράλληλα αυτά τα συστήματα ασύγχρονα πρέπει να ενημερώνουν τον server που παρέχει τα αποτελέσματα στον χρήστη, ώστε ο χρήστης να μην περιμένει σχεδόν καθόλου για την εκπαίδευση. Μάλιστα σε πραγματικό χρόνο είναι καλύτερο να δίνουμε προβλέψεις με βάση τα μη ανανεωμένα μοντέλα μέχρι να ανανεωθούν αυτά, παρά να περιμένουμε να ανανεώνονται κάθε φορά. Έτσι διατηρούμε την ταχύτητα του συστήματος και την αμεσότητά του, χωρίς να έχουμε κάποιο ιδιαίτερα μεγάλο κόστος στην πρόβλεψη, μιας και μεγάλες μεταβολές δεν συναντάμε εύκολα σε μικρά χρονικά διαστήματα.

9 Επόμενα Βήματα

Η παραπάνω διαδικασία και το σύστημα υλοποίησης αναπτύχθηκε στα πλαίσια της διπλωματικής εξέτασης και επομένως σίγουρα υπάρχουν πράγματα που πρέπει να γίνουν για να μπορέσει να σταθεί ως ένα πραγματικό σύστημα που θα προσφέρει κάποια αξία.

Βασικό στοιχείο όπως αναφέρθηκε και στα συμπεράσματα είναι η υποδομή του συστήματος. Έτσι είναι βασικό να μελετηθεί εκτενέστερα ο τρόπος άντλησης των δεδομένων μέσω κάποιου Kafka συστήματος, αλλά και ο αριθμός των παράλληλων συστημάτων επεξεργασίας των δεδομένων. Τα συστήματα αυτά μάλιστα επειδή θα διαχειρίζονται δεδομένα Μεγάλης Κλίμακας, θα πρέπει να συνδέονται με κατανεμημένα συστήματα αρχείων. Έτσι έαν για παράδειγμα κάποιο σύστημα είναι υπεύθυνο για την εκπαίδευση του LSTM αλγορίθμου, πρέπει να μελετηθεί πως θα στηθεί ο αλγόριθμος ώστε να εκπαιδεύεται ο αλγόριθμος με δεδομένα που βρίσκονται σε κάποιο HDFS. Επίσης η ανάγκη για περισσότερα παράλληλα συστήματα γεννά και την ανάγκη για την υλοποίηση πρωτοκόλλου επικοινωνίας μεταξύ των συστημάτων ώστε να μπορούν να συντονίζονται και εν τέλη να καταλήγουν όλα τα δεδομένα που επιθυμούμε στον χρήστη. Αυτό μπορεί να συνεπάγεται και στην δημιουργία περισσοτέρων RESTful API servers.

Επίσης είναι πολύ σημαντικό να μελετηθούν και οι ανάγκες των πόλεων. Αφού πρόκειται για ένα σύστημα που απευθύνεται σε αυτές, πρέπει να εκτελεστούν έρευνες για τις ανάγκες αυτών από ένα τέτοιο σύστημα. Έτσι θα εμπλουτιστούν τα χαρακτηριστικά και οι

παροχές της εφαρμογής προς τις υπεύθυνες υπηρεσίες του Δήμου και θα είναι ένα ουσιαστικό σύστημα που θα λύνει υπαρκτά προβλήματα και θα βελτιώνει τις διαδικασίες που χρησιμοποιούνται, ενσωματώνοντας νέες τεχνολογίες σε αυτές. Επίσης σημαντικό είναι και έρευνες σχετικά με τον τρόπο που πρέπει να σχεδιαστεί το Web App, ώστε να κάνει τους χρήστες αυτού να βρίσκουν αυτό που θέλουν άμεσα και γρήγορα, με την μέγιστη καλύτερη εμπειρία χρήστη.

Τέλος σημαντικό είναι να βρεθούν οι πηγές δεδομένων σε πραγματικό κόσμο και να υλοποιηθεί ένα πιλοτικό πρόγραμμα πάνω σε αυτά για να εξεταστεί η αποδοτικότητα του συστήματος. Ένα επίσης ενδιαφέρον χαρακτηριστικό θα ήταν η χρησιμοποίηση των δεδομένων για την πρόβλεψη της κίνησης στους δρόμους και την ανάπτυξη στην συνέχεια ενός συστήματος δυναμικού ελέγχου των φαναριών.

10 References

[1] The internet of things: a survey by L. Atzori, A. Iera and G. Morabito

[2] The computer for the 21st century by M. Weiser

[3] Towards a smart city based on cloud of things, a survey on the smart city vision and paradigms by R. Petrolo, V. Loscrì and N. Mitton

[4] Machine learning for internet of things data analysis: a survey by Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi,Amit P. Sheth

[5] The vision of a smart city by B. Bowerman, J. Braverman, J. Taylor, H. Todosow and U. Von Wimmersperg

[6] The Value of Descriptive Analytics: Evidence from Online Retailers by Ron Berman, Ayelet Israeli

[7] "Research Trends in Descriptive Analysis" by Kimberly N. Sloman

[8] Why is Data Visualization Important? What is Important in Data Visualization?" by Antony Unwin

[9] An Overview of Descriptive Analysis by Ayush Singh Rawat

[10] Επιχειρησιακές Προβλέψεις από Φώτιο Πετρόπουλο και Βασίλειο Ασημακόπουλος

[11] "Descriptive statistics in Time Series Modelling" by Snigdha Cheekoty

[12] Ανάλυση Δεδομένων από Δημήτρη Κουγιουμτζή

[13]Too Big to Ignore: The Business Case for Big Data by Phil Simon

[14]https://medium.com/shallow-thoughts-about-deep-learning/how-would-we-find-a-better-activation-function-than-relu-4409df217a5c

[15] https://nix-united.com/blog/find-out-how-to-use-machine-learning-for-time-series-forecasting/

[16] http://neuralnetworksanddeeplearning.com/chap1.html

[17] "Multilayer feedforward networks are universal approximators" by K. Hornik, M. Stinchcombe, and H. White

[18] "Deep Learning" by I. Goodfellow, Y. Bengio, and A. Courville

[19] "Multilayer feedforward networks are universal approximators," by K. Hornik, M. Stinchcombe, and H. White,

Appendix

[20] Open Data DK (ODDK) https://www.opendata.dk/search?q=tags:iot%20organization:city-of-aarhus

[21] Django Framework https://www.djangoproject.com/

[22] Django REST Framework https://www.django-rest-framework.org/

[23] ReactJS https://reactjs.org/

[24] Material UI https://mui.com/

[25] amCharts 4 https://www.amcharts.com/docs/v4/

[26] Mapbox https://www.mapbox.com/

[27] OpenAPI Specification https://swagger.io/specification/#:~:text=The%20OpenAPI%20Specification%20(OAS)%20defines,or%20through%20network%20traffic%20inspection.

[29] Tensorflow https://www.tensorflow.org/

[30] Keras https://keras.io/about/

[31] Scikit-learn https://scikit-learn.org/stable/index.html

[32] Statsmodels https://www.statsmodels.org/stable/index.html

[33] Prophet https://facebook.github.io/prophet/