ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Πρόβλεψη κατανάλωσης φυσικού αερίου με τεχνολογίες επιστήμης δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΜΕΞΗ ΚΩΝΣΤΑΝΤΙΝΟΥ

Επιβλέπων : **Γεώργιος Στάμου**
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΈΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Πρόβλεψη κατανάλωσης φυσικού αερίου με τεχνολογίες επιστήμης δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΜΕΞΗ ΚΩΝΣΤΑΝΤΙΝΟΥ

Επιβλέπων : Γεώργιος Στάμου
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 24η Φεβρουαρίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γεώργιος Στάμου	Γεώργιος Γκούμας	Στέφανος Κόλλιας
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2021

(Υπογραφή)

.....

ΜΕΞΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

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

Περίληψη

Το θέμα της παρούσας διπλωματικής εργασίας είναι η πρόβλεψη της κατανάλωσης φυσικού αερίου με χρήση τεχνικών επιστήμης δεδομένων και μηχανικής μάθησης. Αρχικά έγινε μια εκτεταμένη καταγραφή και ανάλυση σχετικών εργασιών τεχνολογιών αιχμής πρόβλεψης κατανάλωσης φυσικού αερίου, καθώς και συλλογή και παρουσίαση διαθέσιμων συνόλων δεδομένων. Στη συνέχεια, έγινε μια διεξοδική μελέτη και ανάλυση της μεθοδολογίας που πρέπει να εφαρμοστεί προκειμένου να πραγματοποιηθούν σωστά οι σχετικές προβλέψεις και σχεδιάστηκαν και υλοποιήθηκαν μια σειρά μοντέλων μηχανικής μάθησης. Ιδιαίτερη έμφαση δόθηκε στην μελέτη των διάφορων καιρικών και ημερολογιακών παραγόντων που επηρεάζουν την κατανάλωση φυσικού αερίου και στην τελική επιλογή των παραγόντων εκείνων που θα αποτελέσουν τα χαρακτηριστικά των μοντέλων πρόβλεψης. Στα πλαίσια της διπλωματικής, πραγματοποιήθηκαν τρία πειράματα πρόβλεψης κατανάλωσης φυσικού αερίου, χρησιμοποιώντας ένα διαθέσιμο σύνολο δεδομένων. Χωρίζοντας με τρεις διαφορετικούς τρόπους το σύνολο δεδομένων σε σύνολα εκπαίδευσης (train set) και δοκιμής (test set), εξετάστηκε η απόδοση των μοντέλων στο πρόβλημα της ωριαίας και ημερήσιας πρόβλεψης κατανάλωσης φυσικού αερίου, καθώς και στον υπολογισμό της μέγιστης ωριαίας κατανάλωσης που αναμένεται κατά τη διάρκεια της ημέρας. Ακολούθησε εκτενής μελέτη των αποτελεσμάτων με χρήση κατάλληλων δεικτών αξιολόγησης, ενώ επίσης, έγινε συγκριτική αξιολόγηση της απόδοσης των μοντέλων που σχεδιάστηκαν στην παρούσα διπλωματική με την απόδοση που πέτυχαν οι ερευνητές σχετικής εργασίας που χρησιμοποίησε το ίδιο σύνολο δεδομένων.

Λέξεις Κλειδιά:

Επιστήμη Δεδομένων, Μηχανική Μάθηση, Τεχνητή Νοημοσύνη, Κατανάλωση Φυσικού Αερίου, Πρόβλεψη Κατανάλωσης

Abstract

The subject of this diploma thesis is forecasting natural gas consumption using data science and machine learning techniques. An extensive recording and analysis of related work on state-of-the-art gas consumption forecasting technologies was carried out, as well as the collection and presentation of available datasets. A thorough study and analysis of the methodology to be applied in order to make the relevant predictions was carried out and a series of machine learning models were designed and implemented. Particular emphasis was placed on the study of the various weather and calendar factors that affect gas consumption and the final selection of those factors that will be the features of the forecasting models. As part of the thesis, three gas consumption forecasting experiments were performed, using an available dataset. Dividing the dataset into training and test in three different ways, the performance of the models in the problem of hourly and daily gas consumption forecasting was examined, as well as in the prediction of the peak of the day hourly consumption. This was followed by an extensive study of the results using appropriate evaluation metrics, while also a comparative evaluation of the performance of the models designed in this thesis with the performance achieved by a relevant study using the same data set.

Keywords:

Data Science, Machine Learning, Artificial Intelligence, Natural Gas Consumption, Forecasting Consumption

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου Γεώργιο Στάμου για την ευκαιρία που μου δόθηκε να εργαστώ στο εξαιρετικά αυτό ενδιαφέρον θέμα της διπλωματικής εργασίας μου. Οφείλω ένα τεράστιο ευχαριστώ στον κύριο Δρ. Θεόδωρο Δαλαμάγκα, ο οποίος μου προσέφερε πραγματικά, κάθε δυνατή βοήθεια κατά την εκπόνηση της διπλωματικής μου εργασίας. Ακόμα, θέλω να ευχαριστήσω τον κύριο Χατζηλένα Χρήστο για τη βοήθειά του σε ορισμένα τεχνικά θέματα.

Τέλος, θέλω να ευχαριστήσω ιδιαίτερα τους φίλους μου και την οικογένειά μου που ήταν πάντα δίπλα μου.

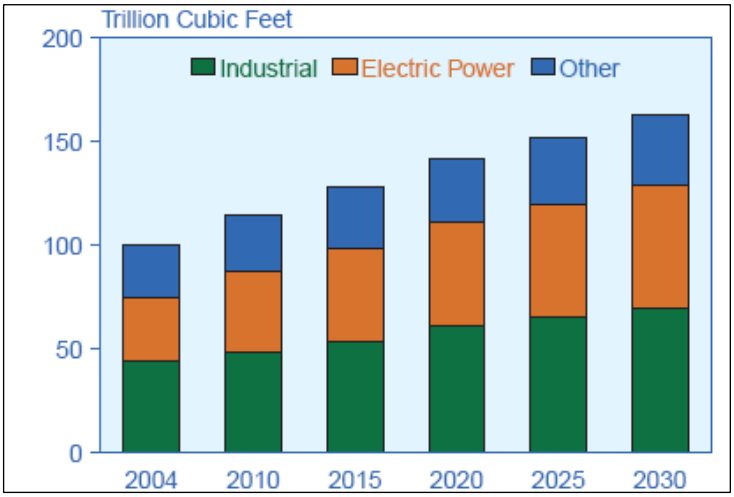
Αφιερώνω την εργασία αυτή στον αδελφό μου Θοδωρή.

Πίνακας περιεχομένων

1 *Εισαγωγή*

Το φυσικό αέριο είναι αέριο μείγμα κορεσμένων υδρογονανθράκων με μικρό αριθμό ατόμων άνθρακα. Είναι άχρωμο και άοσμο στην καθαρή του μορφή και εξάγεται από υπόγειες κοιλότητες. Όταν καίγεται παράγει μεγάλη ποσότητα ενέργειας, ενώ εξαιτίας των ιδιοτήτων του θεωρείται οικολογικό καύσιμο. Είναι μια από τις πιο καθαρές, ασφαλείς και χρήσιμες πηγές ενέργειας. Σε αντίθεση με άλλα ορυκτά καύσιμα, το φυσικό αέριο καίγεται καθαρά παράγοντας κυρίως διοξείδιο του άνθρακα, υδρατμούς και μικρά ποσά οξειδίων του αζώτου. Ως αποτέλεσμα των παραπάνω, η χρήση του φυσικού αερίου πρόκειται να αυξηθεί ακόμη περισσότερο στο μέλλον, καθώς είναι λιγότερο επιβλαβές για το περιβάλλον σε σχέση με άλλα καύσιμα.

Το φυσικό αέριο αποτελεί βασική πηγή παραγωγής ηλεκτρικής ενέργειας σε μονάδες συνδυασμένου κύκλου, ενώ χρησιμοποιείται και για την παραγωγή υδρογόνου. Συχνή είναι επίσης η χρήση του ως καύσιμο οχημάτων. Όλο και μεγαλύτερη αύξηση της χρήσης του παρατηρείται στον οικιακό και εμπορικό τομέα, για θέρμανση και κλιματισμό εσωτερικών χώρων, θέρμανση νερού, μαγειρική καθώς και διάφορες άλλες εξειδικευμένες εργασίες. Διαδεδομένη είναι η χρήση του και στον αγροτικό τομέα, με το φυσικό αέριο να χρησιμοποιείται για θέρμανση σύγχρονων θερμοκηπίων και παραγωγή διοξειδίου του άνθρακα. Οι τομείς με τη μεγαλύτερη κατανάλωση φυσικού αερίου παγκοσμίως είναι ο βιομηχανικός και η παραγωγή ηλεκτρικής ενέργειας.



Εικόνα 1: Παγκόσμια κατανάλωση φυσικού αερίου ανά τομέα
(Πηγή: Energy Information Administration)

Υπάρχουν πολλοί παράγοντες που επηρεάζουν την κατανάλωση του φυσικού αερίου. Ειδικότερα, στον οικιακό και εμπορικό τομέα, όπου η κυριότερη χρήση του φυσικού αερίου είναι η θέρμανση εσωτερικών χώρων, ο σημαντικότερος παράγοντας που επηρεάζει την κατανάλωση είναι οι καιρικές συνθήκες που επικρατούν και ιδιαίτερα η θερμοκρασία περιβάλλοντος. Άλλοι καιρικοί παράγοντες, όπως η υγρασία, η ηλιακή ακτινοβολία, η ένταση και η κατεύθυνση του ανέμου, είναι επίσης παράγοντες που επηρεάζουν την κατανάλωση και πρέπει να λαμβάνονται υπόψη. Σημαντικό ρόλο στην κατανάλωση φυσικού αερίου φαίνεται να έχουν και ημερολογιακοί παράγοντες, όπως η ημέρα του χρόνου και της εβδομάδας, ο μήνας, η εποχή του χρόνου, ακόμα και η ώρα της ημέρας. Το είδος της ημέρας, δηλαδή αν πρόκειται για επίσημη αργία, διακοπές, καθημερινή ή Σαββατοκύριακο είναι ακόμα ένας ημερολογιακός παράγοντας που φαίνεται να έχει επίδραση στην κατανάλωση. Οι ημερολογιακοί παράγοντες αυτοί είναι σημαντικοί, καθώς συνδέονται σε μεγάλο βαθμό με τα επίπεδα κατανάλωσης, καθώς, όπως είναι αναμενόμενο, τα επίπεδα και οι διακυμάνσεις της κατανάλωσης είναι σαφώς υψηλότερες κατά τους χειμερινούς μήνες, όταν υπάρχει ανάγκη θέρμανσης, ενώ τους καλοκαιρινούς μήνες η κατανάλωση είναι μικρή και χωρίς ιδιαίτερες διακυμάνσεις. Επιπρόσθετα, διάφοροι κοινωνικοοικονομικοί παράγοντες έχουν προταθεί ως παράγοντες που συνδέονται με την κατανάλωση ενέργειας και ειδικότερα του φυσικού αερίου. Ο πληθυσμός και η μέση ηλικία των κατοίκων μιας περιοχής, το ΑΕΠ, το μέσο εισόδημα και η τιμή του φυσικού αερίου και του πετρελαίου είναι οι βασικότεροι που έχουν προταθεί και χρησιμοποιηθεί στη βιβλιογραφία. Στον βιομηχανικό τομέα ωστόσο, η κατανάλωση παραμένει σε σχετικά σταθερά επίπεδα ανεξάρτητα από την εποχή του χρόνου. Στον τομέα αυτό, οι καιρικοί παράγοντες δεν είναι αυτοί που επηρεάζουν περισσότερο την κατανάλωση, ενώ άλλοι κοινωνικοοικονομικοί ή παράγοντες σχετικοί με την ενεργειακή πολιτική και τη νομοθεσία του κράτους φαίνεται να έχουν τη σημαντικότερη επίδραση.

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η πρόβλεψη της κατανάλωσης φυσικού αερίου με χρήση τεχνικών επιστήμης δεδομένων και μηχανικής μάθησης. Στα πλαίσια της εργασίας, έγινε μια διεξοδική μελέτη και ανάλυση της μεθοδολογίας που πρέπει να εφαρμοστεί προκειμένου να πραγματοποιηθούν σωστά οι σχετικές προβλέψεις και σχεδιάστηκαν και υλοποιήθηκαν μια σειρά μοντέλων μηχανικής μάθησης. Η συνεισφορά της διπλωματικής εργασίας συνοψίζεται ως εξής:

- Εκτεταμένη καταγραφή, ανάλυση και σύγκριση σχετικών εργασιών τεχνολογιών αιχμής πρόβλεψης κατανάλωσης φυσικού αερίου.
- Συλλογή, συγκριτική παρουσίαση και μελέτη διαθέσιμων συνόλων δεδομένων.
- Καθορισμός μεθοδολογίας δειγματοληψίας συνόλων εκπαίδευσης (train set) και δοκιμής (test set) και εκπαίδευσης μοντέλων μηχανικής μάθησης στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου.
- Εκτεταμένη αξιολόγηση ενός μεγάλου αριθμού μοντέλων μηχανικής μάθησης στο πρόβλημα της πρόβλεψης της ωριαίας και ημερήσιας κατανάλωσης φυσικού αερίου, καθώς και της μέγιστης ωριαίας κατανάλωσης της ημέρας.
- Τα πειραματικά αποτελέσματα έδειξαν ότι τα τρία μοντέλα που παρουσίασαν την υψηλότερη απόδοση σε όρους RMSE και MAE στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης φυσικού αερίου ήταν τα *XGBoost*, *SVR* και τα *τεχνητά νευρωνικά δίκτυα.*

Η οργάνωση του κειμένου της διπλωματικής εργασίας έχει πραγματοποιηθεί ως εξής. Στο *Κεφάλαιο 2*, παρουσιάζεται το θεωρητικό υπόβαθρο σχετικά με τις τεχνικές προβλέψεων και αναλύονται κι αξιολογούνται οι σχετικές εργασίες τεχνολογιών αιχμής πρόβλεψης της κατανάλωσης φυσικού αερίου, οι οποίες αποτέλεσαν σημείο εκκίνησης και αναφοράς για την παρούσα διπλωματική. Στο *Κεφάλαιο 3*, γίνεται παρουσίαση και μελέτη των σχετικών διαθέσιμων συνόλων δεδομένων και η επιλογή εκείνου που χρησιμοποιήθηκε ως case study στην ανάλυση που ακολούθησε. Στο *Κεφάλαιο 4*, παρουσιάζεται η διαδικασία της μελέτης του συνόλου δεδομένων που επιλέχθηκε, της προεπεξεργασίας των δεδομένων, καθώς και της ετοιμασίας των μοντέλων πρόβλεψης. Στο *Κεφάλαιο 5,* γίνεται η παρουσίαση, ο σχολιασμός και η αξιολόγηση των πειραματικών αποτελεσμάτων. Στο *Κεφάλαιο 6*, γίνεται η σύνοψη της εργασίας, παρουσιάζονται τα συμπεράσματα που προέκυψαν, καθώς και πιθανές μελλοντικές επεκτάσεις.

2 *Θεωρητικό Υπόβαθρο και Σχετικές Εργασίες*

2.1 Τεχνικές Προβλέψεων

2.1.1 Ορισμός και Διαδικασία Πρόβλεψης

Ως πρόβλεψη μπορεί να οριστεί η εκτίμηση αβέβαιων μελλοντικών γεγονότων. Οι προβλέψεις μπορούν να γίνουν βασισμένες στην εμπειρία και την παρατήρηση, σε στατιστικές μεθόδους καθώς και σε πολύπλοκα μαθηματικά μοντέλα. Χρησιμοποιούνται για τη βελτίωση της λήψης αποφάσεων και σχεδιασμού. [PA2013]

Η διαδικασία παραγωγής προβλέψεων είναι μια απαιτητική διαδικασία. Στην παράγραφο αυτή θα περιγραφούν εν συντομία τα πέντε βασικά βήματα που είναι απαραίτητα σε μια διαδικασία παραγωγής και αξιολόγησης προβλέψεων:

1. *Καθορισμός του προβλήματος.* Αποτελεί συχνά το δυσκολότερο μέρος της διαδικασίας. Στο βήμα αυτό γίνεται προσπάθεια σαφούς προσδιορισμού των μεγεθών που πρόκειται να προβλεφθούν, καθώς και ο τρόπος που αυτές θα χρησιμοποιηθούν.
2. *Συλλογή των δεδομένων.* Αρκετός χρόνος θα πρέπει να αναλωθεί στην ορθή συλλογή και συντήρηση των δεδομένων. Εκτός των μετρήσιμων αριθμητικών δεδομένων, σημαντική αποδεικνύεται και η χρήση εμπειρικών πληροφοριών όταν αυτές είναι διαθέσιμες. Η διαδικασία αυτή αποδεικνύεται συχνά χρονοβόρα.
3. *Προεπεξεργασία των δεδομένων.* Σκοπός του βήματος αυτού είναι η απόκτηση μιας ολοκληρωμένης αίσθησης των διαθέσιμων δεδομένων, ώστε να αναγνωριστούν πιθανά λανθάνοντα πρότυπα, σημαντικές τάσεις ή εποχικότητα και ασυνήθιστες

τιμές. Η διαδικασία αυτή θα οδηγήσει σε μια εξομαλυμένη σειρά δεδομένων, έτοιμη για την εφαρμογή μοντέλων πρόβλεψης.

4. *Επιλογή μεθόδων πρόβλεψης.* Στο βήμα αυτό επιτυγχάνεται η ορθή επιλογή κατάλληλων μοντέλων πρόβλεψης, αλλά και η επιλογή των παραμέτρων τους που θα οδηγήσουν στα πλέον ακριβή αποτελέσματα.
5. *Χρήση και αξιολόγηση των μοντέλων πρόβλεψης.* Στο τελευταίο στάδιο τα επιλεγμένα μοντέλα χρησιμοποιούνται ώστε να παραχθούν οι ζητούμενες προβλέψεις. Το κατά πόσο τα επιλεγμένα μοντέλα και προβλέψεις είναι ικανοποιητικές κρίνονται μόνο από το χρόνο, καθώς νέα δεδομένα γίνονται διαθέσιμα. Η αξιολόγηση και η μέτρηση της ακρίβειας των προβλέψεων επιτυγχάνεται με εξειδικευμένους στατιστικούς δείκτες.

2.1.2 Κατηγορίες Μεθόδων Πρόβλεψης

Οι μέθοδοι πρόβλεψης διακρίνονται σε τρεις μεγάλες κατηγορίες σύμφωνα με τη διαδικασία παραγωγής τους:

Ποσοτικές Μέθοδοι. Οι ποσοτικές μέθοδοι αναφέρονται στην εφαρμογή στατιστικών μοντέλων χρονοσειρών ή αιτιοκρατικών μοντέλων επί μιας σειράς δεδομένων με σκοπό την αυτοματοποιημένη και συστηματική παραγωγή προβλέψεων. Οι στατιστικές προβλέψεις είναι άμεσα εφαρμόσιμες και αποδεκτά ακριβείς, αν συνδυαστούν με κατάλληλα διαστήματα εμπιστοσύνης. Προϋποθέτουν ότι το πρότυπο (συμπεριφορά) της εκάστοτε χρονοσειράς θα συνεχιστεί στο μέλλον, γεγονός που ωστόσο δε συμβαίνει πάντα. Επίσης, οι μέθοδοι αυτές δε λαμβάνουν υπόψη ειδικά γεγονότα και ενέργειες που ενδέχεται να πραγματοποιηθούν στο άμεσο μέλλον. Ακόμα, βασική παραδοχή των αιτιοκρατικών μοντέλων είναι η ύπαρξη σταθερή συσχέτισης μεταξύ του προς πρόβλεψη μεγέθους και άλλων παραγόντων, χωρίς ωστόσο να είναι απαραίτητη η ύπαρξη χρονικής εξάρτησης. Τέλος, η συλλογή των δεδομένων μπορεί πολλές φορές να είναι μια δύσκολη και χρονοβόρα διαδικασία, καθώς συνήθως απαιτείται μεγάλο πλήθος ιστορικών δεδομένων προκειμένου να παραχθούν οι ζητούμενες προβλέψεις. Τέτοια μοντέλα είναι οι μέθοδοι εκθετικής εξομάλυνσης, τα μοντέλα παλινδρόμησης, τα μοντέλα ARIMA και τα τεχνητά νευρωνικά δίκτυα.

Κριτικές Μέθοδοι. Οι κριτικές μέθοδοι πρόβλεψης δεν έχουν τις ίδιες απαιτήσεις σε δεδομένα όπως οι στατιστικές μέθοδοι. Τα δεδομένα των κριτικών μεθόδων αποτελούν προϊόν διαίσθησης, κρίσης και συσσωρευμένης γνώσης από πλευράς εμπειρογνωμόνων. Οι μέθοδοι αυτές μπορούν να λάβουν υπόψη ειδικά γεγονότα και ενέργειες, ενώ ταυτόχρονα έχουν τη δυνατότητα να αντισταθμίζουν ανεπάρκειες και ελλείψεις σε ιστορικά δεδομένα. Είναι κατάλληλες όταν θίγονται ηθικά ζητήματα που υπερισχύουν των οικονομικών και τεχνολογικών παραγόντων. Οι μέθοδοι αυτές πρέπει να λειτουργούν συμπληρωματικά με τις

στατιστικές μεθόδους. Μεταξύ των πιο διαδεδομένων κριτικών μεθόδων συγκαταλέγονται η απλή κρίση, η μέθοδος Delphi και οι δομημένες αναλογίες.

Τεχνολογικές Μέθοδοι. Οι τεχνολογικές μέθοδοι πρόβλεψης αφορούν κυρίως μακροπρόθεσμα πλάνα τεχνολογικής, οικονομικής, κοινωνικής και πολιτικής φύσης. Διακρίνονται σε διερευνητικές και κανονιστικές. Οι πρώτες έχουν ως σημείο εκκίνησης το παρελθόν και το παρόν και στοχεύουν στη διερεύνηση όλων των πιθανών μελλοντικών περιπτώσεων. Οι κανονιστικές έχουν προκαθορισμένους στόχους και εξετάζουν τη δυνατότητα επίτευξής τους, σύμφωνα με τους υπάρχοντες περιορισμούς και διαθέσιμους πόρους [PA2013].

2.1.3 Ορίζοντας Πρόβλεψης

Οι μέθοδοι προβλέψεων σπάνια περιορίζονται στην πρόβλεψη μόνο της αμέσως επόμενης περιόδου της υπό μελέτη χρονοσειράς. Συνήθως, απαιτείται η παραγωγή προβλέψεων για αρκετές περιόδους στο μέλλον. Ο ορίζοντας πρόβλεψης είναι ο δείκτης που δείχνει για πόσες μελλοντικές περιόδους της ζητούμενης χρονοσειράς καλείται κανείς να δώσει εκτιμήσεις μέσω μεθοδολογιών πρόβλεψης. Ο ορισμός του ορίζοντα πρόβλεψης έχει σημαντική επίδραση στην επιλογή της καταλληλότερης τεχνικής προβλέψεων. Ως προς το χρονικό ορίζοντα οι προβλέψεις διακρίνονται σε:

Πολύ βραχυπρόθεσμη πρόβλεψη. Η τιμή του χρονικού ορίζοντα πρόβλεψης κυμαίνεται από μερικά λεπτά έως και μια ώρα. Οι προβλέψεις αυτού του είδους βρίσκουν χρησιμότητα σε real time εφαρμογές, όπως η συντήρηση και η διάγνωση σφαλμάτων ενός ηλεκτρικού δικτύου.

Βραχυπρόθεσμη πρόβλεψη. Η τιμή του χρονικού ορίζοντα πρόβλεψης κυμαίνεται από μια ώρα έως και μερικές εβδομάδες. Βρίσκει εφαρμογή στη λήψη επιχειρησιακών και λειτουργικών αποφάσεων.

Μεσοπρόθεσμη πρόβλεψη. Η τιμή του χρονικού ορίζοντα πρόβλεψης κυμαίνεται από μερικές εβδομάδες έως και 3 χρόνια. Αποτελεί τη συνηθέστερη κατηγορία πρόβλεψης και βρίσκει εφαρμογές στον οικονομικό σχεδιασμό επιχειρήσεων και στη λήψη τακτικών αποφάσεων, όπως ο προγραμματισμός συντήρησης, η αξιολόγηση επάρκειας και η διαχείριση περιορισμένων ενεργειακών μονάδων.

Μακροπρόθεσμη πρόβλεψη. Η τιμή του χρονικού ορίζοντα πρόβλεψης κυμαίνεται από 3 έως και 10 χρόνια, ενώ ενδεχομένως να φτάσει μέχρι και μερικές δεκαετίες στο μέλλον. Αναφέρεται στο μακροχρόνιο τεχνοοικονομικό σχεδιασμό των επιχειρήσεων και τη λήψη στρατηγικών αποφάσεων [PA2013].

2.1.4 Χρονοσειρές

Οι χρονοσειρές αποτελούν ένα σύνολο διαδοχικών παρατηρήσεων της τιμής κάποιου φυσικού ή άλλου μεγέθους και περιγράφουν την εξέλιξή του στο χρόνο. Οι παρατηρήσεις αυτές λαμβάνονται σε ισαπέχουσες χρονικές στιγμές ή περιόδους. Οι διαδοχικές αυτές παρατηρήσεις δεν είναι ανεξάρτητες μεταξύ τους και μπορούν να χρησιμοποιηθούν στην προσπάθεια πρόβλεψης μελλοντικών τιμών τους. Όταν οι διαδοχικές παρατηρήσεις είναι εξαρτημένες, οι μελλοντικές τιμές μπορούν να προσδιοριστούν ακριβώς από τις προηγούμενες. Τα μοντέλα που περιγράφουν αυτή την εξέλιξη ονομάζονται *ντετερμινιστικά.* Ωστόσο, κάτι τέτοιο δε συμβαίνει με τις πραγματικές χρονοσειρές, καθώς το μέλλον καθορίζεται μόνο μερικώς από το παρελθόν. Έτσι, οι χρονοσειρές θεωρείται ότι περιγράφονται από μοντέλα που περιέχουν τον τυχαίο παράγοντα, τα οποία ονομάζονται *στοχαστικά.* Η συστηματική μελέτη μιας χρονοσειράς ξεκινά με την επισκόπηση του γραφήματός της στο πεδίο του χρόνου. Με τον τρόπο αυτό προκύπτουν τα βασικά ποιοτικά της χρονοσειράς, τα οποία είναι η τάση, η εποχιακότητα, η κυκλικότητα, οι ασυνέχειες και τυχαιότητα.

Η *τάση* θα μπορούσε να οριστεί σα μια μακροπρόθεσμη μεταβολή του μέσου επιπέδου των τιμών της χρονοσειράς. Αντιπροσωπεύει τη γενική εικόνα της χρονοσειράς, ενώ για την εύρεσή της απαιτείται ικανός αριθμός παρατηρήσεων καθώς και σωστή εκτίμηση του κατάλληλου μήκους της περιόδου μέσα στο οποίο θα αναζητηθεί η ύπαρξη τάσης.

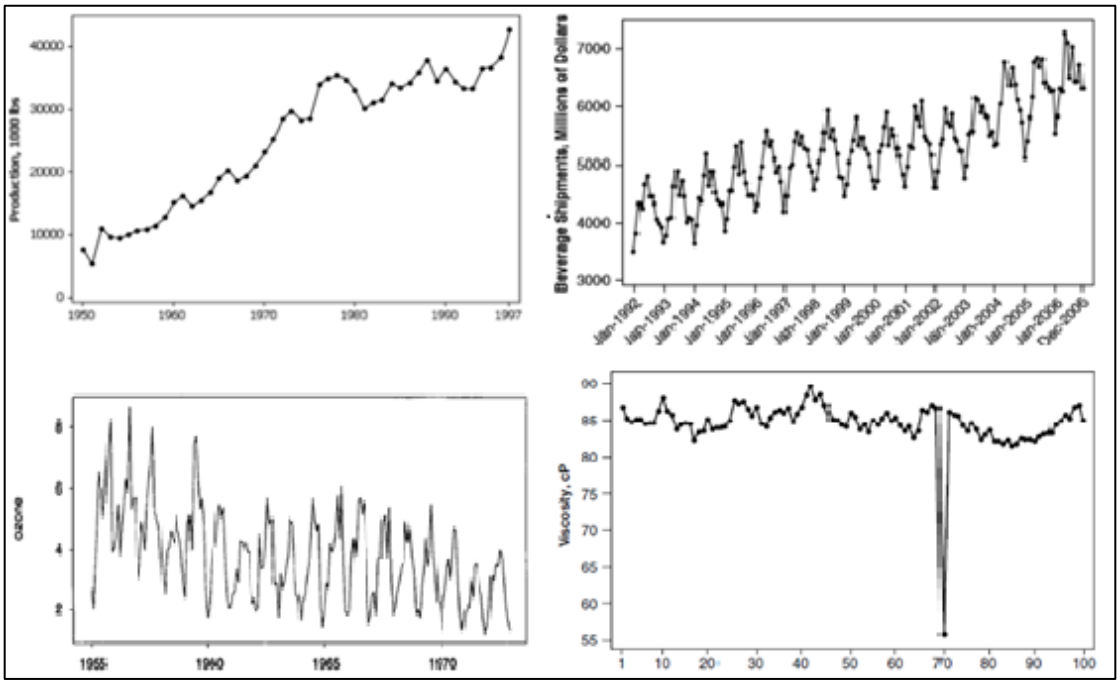
Η *κυκλικότητα* αντιπροσωπεύει μια κυματοειδή μεταβολή που οφείλεται σε ειδικές εξωγενείς συνθήκες και εμφανίζεται κατά περιόδους. Οι περίοδοι δεν είναι απαραίτητα σταθερές και το μήκος τους είναι κατά κανόνα μεγαλύτερο τους ενός έτους. Κυκλικότητα εμφανίζουν συνήθως οι χρονοσειρές των περισσότερων οικονομικών μεγεθών, όπως του ΑΕΠ και των τιμών των μετοχών.

Η *εποχιακότητα* ορίζεται σα μια περιοδική διακύμανση που έχει σταθερό και μικρότερο του έτους μήκος. Η διακύμανση αυτή είναι συνήθως κατανοητή και προβλέψιμη. Συναντάται κυρίως σε χρονοσειρές, οι τιμές των οποίων επηρεάζονται σημαντικά από την εποχή. Η μέτρηση και η απομόνωσή της είναι σχετικά εύκολη και γίνεται με διάφορες μεθόδους ώστε να προκύψουν τα λεγόμενα αποεποχικοποιημένα δεδομένα.

Ασυνέχειες ονομάζονται οι απομονωμένες παρατηρήσεις που εμφανίζονται στο γράφημα κάποιας χρονοσειράς ως απότομες αλλαγές στο πρότυπο συμπεριφοράς της και δε θα μπορούσαν να έχουν προβλεφθεί από την ιστορία της. Τέτοιες αλλαγές ενδέχεται να έχουν παροδικό ή μόνιμο χαρακτήρα. Στην πρώτη περίπτωση έχει επικρατήσει η ορολογία outliers και η επίδρασή τους έχει μικρή διάρκεια. Ένα outlier μπορεί να αντιπροσωπεύει μια ασυνήθιστη παρατήρηση που οφείλεται σε κάποιο εξαιρετικό και απρόβλεπτο γεγονός. Στην

περίπτωση που οι παρατηρούμενες ασυνέχειες έχουν μόνιμο χαρακτήρα ονομάζονται level-shifts, αφού εμφανίζονται ως απότομες αλλαγές στο μέσο επίπεδο των τιμών της χρονοσειράς.

Τέλος, η *τυχαιότητα* αντιπροσωπεύει τις μη κανονικές διακυμάνσεις που απομένουν όταν όλα τα υπόλοιπα συστατικά στοιχεία της χρονοσειράς έχουν απομονωθεί. Οι διακυμάνσεις αυτές μπορεί να αντιπροσωπεύουν μια εντελώς τυχαία μεταβλητή που εκφράζει τον τυχαία παράγοντα μιας στοχαστικής διαδικασίας. [PA2013]



Εικόνα 2: Παραδείγματα χρονοσειρών με
Τάση (πάνω αριστερά), Κυκλικότητα (πάνω δεξιά),
Εποχιακότητα (κάτω αριστερά), Ασυνέχεια (κάτω δεξιά)

2.1.5 Μηχανική Μάθηση

Η μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Οι αλγόριθμοι αυτοί βελτιώνουν τη συμπεριφορά τους σε κάποια εργασία χρησιμοποιώντας την εμπειρία τους. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα. Ο Άρθουρ Σάμουελ ορίζει τη μηχανική μάθηση ως "*Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί*".

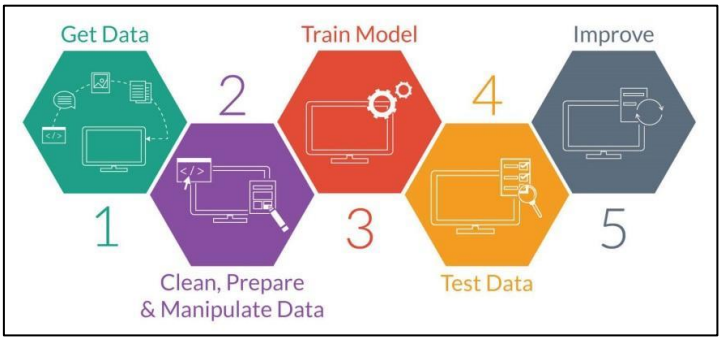
Ο τομέας της μηχανικής μάθησης αναπτύσσει τρεις τρόπους μάθησης, ανάλογους με τους τρόπους με τους οποίους μαθαίνει ο άνθρωπος:

Επιβλεπόμενη μάθηση (Supervised Learning). Η επιβλεπόμενη μάθηση είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα ταξινόμησης (classification), πρόγνωσης (prediction) και διερμηνείας (interpretation).

Μη επιβλεπόμενη μάθηση (Unsupervised Learning). Στην μη επιβλεπόμενη μάθηση ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα ανάλυσης συσχετισμών (association analysis) και ομαδοποίησης (clustering).

Ενισχυτική μάθηση (Reinforcement Learning). Στην ενισχυτική μάθηση ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα σχεδιασμού, όπως ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.

Για κάθε πρόβλημα προς επίλυση στο χώρο της μηχανικής μάθησης υπάρχει ένας κατάλληλος τρόπος μάθησης και για κάθε τρόπο μάθησης υπάρχει τουλάχιστον ένας κατάλληλος αλγόριθμος που μπορεί να χρησιμοποιηθεί. Ορισμένοι αλγόριθμοι δέχονται ως είσοδο μόνο παρατηρήσεις και άλλοι λαμβάνουν υπόψη τους λίγο ή περισσότερο την προϋπάρχουσα γνώση. Στην επόμενη εικόνα, αποτυπώνεται ο γενικός τρόπος λειτουργίας των αλγορίθμων μηχανικής μάθησης. Η σημαντικότερη φάση κάθε αλγόριθμου είναι η εκπαίδευση, όπου ο αλγόριθμος χρησιμοποιεί ως είσοδο ένα σύνολο δεδομένων εκπαίδευσης (training set) προς επίτευξη του σκοπού του, τη δημιουργία νέας γνώσης. Την εκπαίδευση ακολουθεί η φάση της πιστοποίησης της παραγόμενης νέας γνώσης με τη βοήθεια δεδομένων ελέγχου (test set) και στη συνέχεια, μέσω κριτικής που κάνει ο χρήστης βάσει των γνώσεων που διαθέτει για το πρόβλημα που επιχειρεί να λύσει ο αλγόριθμος. Τέλος, η νέα γνώση δίνεται προς χρήση σε εφαρμογές στις οποίες είναι απαραίτητη, για να λυθούν πραγματικά προβλήματα.



Εικόνα 3: Γενικός τρόπος λειτουργίας αλγορίθμων μηχανικής μάθησης

2.2 Τεχνολογίες Αιχμής Πρόβλεψης Κατανάλωσης ΦΑ

Η δημοσίευση εργασιών σχετικών με την πρόβλεψη κατανάλωσης φυσικού αερίου έχει ξεκινήσει ήδη από τα μισά του προηγούμενου αιώνα, ενώ την τελευταία δεκαετία οι σχετικές ερευνητικές δραστηριότητες παρουσιάζουν ραγδαία αύξηση. Οι εργασίες αυτές παρουσιάζουν διαφορές ως προς το χρονικό ορίζοντα πρόβλεψης, ο οποίος κυμαίνεται από μερικές ώρες και φτάνει έως προβλέψεις επόμενων δεκαετιών. Επίσης, η κλίμακα στην οποία γίνονται οι προβλέψεις μπορεί να είναι από μια μεμονωμένη κατοικία, μια πόλη, ένα νομό έως εθνικό ή και παγκόσμιο επίπεδο. Μπορεί επίσης να αφορούν τον οικιακό, τον εμπορικό ή το βιομηχανικό τομέα μιας περιοχής καθώς και συνδυασμό αυτών. Διαφορές παρουσιάζουν ακόμα οι μέθοδοι με τις οποίες οι εκάστοτε ερευνητές προσεγγίζουν το πρόβλημα. Ποσοτικές μέθοδοι, όπως η μέθοδος των χρονοσειρών, μέθοδοι παλινδρόμησης, ασαφής λογική, μοντέλα μηχανικής μάθησης και τεχνητά νευρωνικά δίκτυα είναι οι μέθοδοι που συναντώνται συχνότερα. Ένα επιπλέον σημείο διαφοροποίησης των εργασιών είναι τα χαρακτηριστικά, τα οποία επιλέγονται ως μεταβλητές εισόδου των μοντέλων που χρησιμοποίησαν. Τα χαρακτηριστικά με την μεγαλύτερη επίδραση στην κατανάλωση φυσικού αερίου είναι οι καιρικοί παράγοντες και ιδιαίτερα η εξωτερική θερμοκρασία. Ωστόσο, διάφοροι άλλοι κοινωνικοοικονομικοί παράγοντες, όπως το ΑΕΠ, η τιμή του φυσικού αερίου και του πετρελαίου, έχουν ληφθεί σα μεταβλητές εισόδου, κυρίως σε μελέτες μακροχρόνιων προβλέψεων

Κάθε διανομέας φυσικού αερίου είναι υποχρεωμένος να κάνει μια εκτίμηση της ποσότητας αερίου που απαιτείται για την επόμενη ημέρα ή χρονική περίοδο από τον προμηθευτή του. Στην περίπτωση που η πραγματική κατανάλωση ξεπεράσει κάποια επιτρεπόμενα όρια σε σχέση με την προβλεπόμενη τιμή, επιβάλλονται χρηματικά πρόστιμα στον εν λόγω διανομέα. Κατά συνέπεια, καθώς η εσφαλμένη εκτίμηση της κατανάλωσης συνεπάγεται υψηλά κόστη, είναι απαραίτητη η δημιουργία μοντέλων ικανών να προβλέψουν με ακρίβεια μελλοντικές καταναλώσεις.

Στο παρόν κεφάλαιο παρουσιάζονται σχετικές εργασίες στην υπάρχουσα βιβλιογραφία πρόβλεψης κατανάλωσης φυσικού αερίου με χρήση τεχνικών μηχανικής μάθησης, οι οποίες απετέλεσαν σημείο αφετηρίας και αναφοράς για την παρούσα διπλωματική εργασία.

2.2.1 Τεχνολογίες ANN για ημερήσια πρόβλεψη κατανάλωσης ΦΑ

Η παρούσα εργασία [BM1995] αποτελεί μια προσπάθεια χρήσης των *τεχνητών νευρωνικών δικτύων* (*Artificial Neural Networks* ή *ANN*) στο πρόβλημα της ημερήσιας πρόβλεψης κατανάλωσης φυσικού αερίου. Ο σημαντικότερος παράγοντας που επηρεάζει τα επίπεδα κατανάλωσης είναι η θερμοκρασία, καθώς η κύρια χρήση του φυσικού αερίου είναι η

θέρμανση εσωτερικών χώρων. Η ένταση του ανέμου αποτελεί έναν επίσης σημαντικό παράγοντα, καθώς η ανάγκη για θέρμανση αυξάνεται τις μέρες με αέρα. Έτσι, τα **χαρακτηριστικά** που χρησιμοποίησαν οι ερευνητές στα μοντέλα ήταν η *θερμοκρασία*, η *ταχύτητα ανέμου*, *προγενέστερες καταναλώσεις* και η *ημέρα της εβδομάδας*. Χρησιμοποιήθηκε επίσης ο *παράγοντας HDD* (heating degree day). Ο παράγοντας αυτός αποτελεί συνάρτηση της θερμοκρασίας και είναι ένα υπολογιζόμενο μέγεθος που χρησιμοποιείται για την ποσοτική εκτίμηση της απαιτούμενης ενέργειας για τη θέρμανση των εσωτερικών χώρων των κτιρίων. Η συσχέτισή του με την κατανάλωση φυσικού αερίου είναι γραμμική. Είναι σημαντικό να αναφερθεί ότι χρησιμοποιήθηκαν τόσο πραγματικές, όσο και προβλεπόμενες τιμές για τη θερμοκρασία, τον HDD και την ταχύτητα ανέμου, καθώς σε πραγματικές συνθήκες μόνο οι προβλεπόμενες τιμές των μεγεθών αυτών είναι διαθέσιμες. Το σύνολο δεδομένων που χρησιμοποιήθηκε αφορά μια πολιτεία των ΗΠΑ για το διάστημα 1990 έως 1995.

Τα **μοντέλα** που εξετάστηκαν, όπως αναφέρθηκε ήταν *τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης* (*ANN*). Όπως αναφέρουν οι ερευνητές, οι σημαντικότερες πηγές σφαλμάτων στις προβλέψεις κατανάλωσης φυσικού αερίου είναι τα σφάλματα στις προβλέψεις των καιρικών παραγόντων και τα σφάλματα που παράγουν τα διάφορα μαθηματικά μοντέλα στην πρόβλεψη της κατανάλωσης. Με τη χρήση των *ANN* γίνεται προσπάθεια ελαχιστοποίησης της δεύτερης κατηγορίας σφαλμάτων, μέσω της ικανότητας τους να μοντελοποιούν μη γραμμικές σχέσεις. Ως μέτρο ποιότητας των προβλέψεων χρησιμοποιήθηκε ο δείκτης *RMSE*, ενώ η απόδοση των νευρωνικών συγκρίθηκε με την απόδοση της μεθόδου *Linear Regression* (*LR*) που χρησιμοποιείτο για την παραγωγή αντίστοιχων προβλέψεων.

Το πρώτο νευρωνικό δίκτυο που σχεδιάστηκε, εκπαιδεύτηκε με δεδομένα της περιόδου 1992-1993. Οι προβλέψεις που παρήγαγε αφορούσαν την περίοδο 1993-1994. Η απόδοσή του ήταν σημαντικά καλύτερη από της μεθόδου *LR* που εκπαιδεύτηκε με τα ίδια δεδομένα. Συγκεκριμένα, το σφάλμα σε όρους *RMSE* ελαττώθηκε κατά 32% με τη χρήση του *ANN*. Οι ερευνητές παρατήρησαν ότι το νευρωνικό αυτό είχε την τάση να παράγει προβλέψεις χαμηλότερες από τις πραγματικές τιμές κατανάλωσης κατά τους Φθινοπωρινούς μήνες και υψηλότερες κατά την Άνοιξη. Κατά συνέπεια, θεώρησαν ότι η προσθήκη της *ημέρας του χρόνου* ως χαρακτηριστικό θα βελτιώσει την απόδοσή του. Επιπλέον, οι τιμές των καιρικών παραγόντων και της κατανάλωσης για την ίδια ημέρα της προηγούμενης εβδομάδας προστέθηκαν ως χαρακτηριστικά του νευρωνικού.

Το νέο νευρωνικό δίκτυο εκπαιδεύτηκε με δεδομένα για την περίοδο 1990-1993 και παρήγαγε προβλέψεις για το διάστημα 1 Ιουνίου 1993 έως 29 Μαρτίου 1994. Το σφάλμα του *ANN* ήταν μικρότερο από το αντίστοιχο της μεθόδου *LR* για κάθε μήνα που παρήχθησαν προβλέψεις, με εξαίρεση να αποτελούν ο Ιανουάριος και ο Φεβρουάριος του 1994. Κατά τους

μήνες αυτούς, το *ANN* δεν κατάφερε να προβλέψει με ακρίβεια την κατανάλωση φυσικού αερίου, παράγοντας μεγαλύτερο σφάλμα από της μεθόδου *LR*. Ο Ιανουάριος του 1994 παρουσίασε ιδιαίτερα έντονο κρύο, κάτι που δεν είχε παρατηρηθεί ξανά μέχρι τότε και δεν υπήρχε αντίστοιχη περίοδος στα δεδομένα εκπαίδευσης του μοντέλου. Κατά συνέπεια, το *ANN* προέβλεψε χαμηλότερες καταναλώσεις από τις πραγματικές. Αντίθετα, κατά τον επόμενο μήνα, το *ANN* προέβλεψε υψηλότερες καταναλώσεις από τις πραγματικές. Αυτό οφείλεται στη συμπεριφορά των καταναλωτών, οι οποίοι κατά το μήνα αυτό μείωσαν σημαντικά την κατανάλωση φυσικού αερίου λόγω των υψηλών λογαριασμών του προηγούμενου μήνα.

Στη συνέχεια, ένα νέο νευρωνικό δίκτυο εκπαιδεύτηκε με δεδομένα της περιόδου 8 Ιανουαρίου 1990 έως 31 Μαΐου 1994 και παρήγαγε προβλέψεις για την περίοδο 1994-1995. Το *ANN* αυτό χρησιμοποίησε αποκλειστικά προβλεπόμενες τιμές για τους καιρικούς παράγοντες και τις προγενέστερες καταναλώσεις. Αποτέλεσμα αυτού, ήταν η μειωμένη απόδοση του μοντέλου σε σχέση με τα προηγούμενα δύο, τα οποία χρησιμοποιούσαν και πραγματικές τιμές των παραπάνω μεταβλητών. Τέλος, το τελευταίο *ANN* κλήθηκε να παράγει προβλέψεις για την περίοδο 1994-1995 μιας άλλης περιοχής, προκειμένου να μελετηθεί η απόδοσή του μοντέλου σε διαφορετικές περιοχές. Αν και το σφάλμα αυξήθηκε, η απόδοση εξακολουθούσε να είναι καλύτερη των εμπειρικών μοντέλων.

Στα **συμπεράσματα** οι ερευνητές καταλήγουν πως η μέθοδος των *τεχνητών νευρωνικών δικτύων* παράγει ακριβέστερες προβλέψεις σε σχέση με γραμμικά μοντέλα όπως η *LR*, μειώνοντας το σφάλμα σε όρους *RMSE* στο μισό. Επόμενες εργασίες πρέπει να εστιαστούν στην καλύτερη μελέτη της συμπεριφοράς των καταναλωτών και την προσθήκη νέων χαρακτηριστικών ως μεταβλητές εισόδου.

2.2.2 Μακροπρόθεσμη πρόβλεψη οικιακής κατανάλωσης ΦΑ με τεχνολογίες ANN

Στην παρούσα εργασία [HH2013] επιχειρείται πρόβλεψη της κατανάλωσης φυσικού αερίου σε ετήσια βάση με χρήση *τεχνητών νευρωνικών δικτύων* (*ANN*) για τον οικιακό τομέα της επαρχίας Kerman στο νοτιοανατολικό Ιράν. Συγκεκριμένα, πραγματοποιήθηκαν προβλέψεις για κάθε μια από τις πόλεις της επαρχίας ξεχωριστά, ενώ στη συνέχεια η συνολική κατανάλωση της Kerman υπολογίστηκε από το άθροισμα των επιμέρους καταναλώσεων. Τα δεδομένα κατανάλωσης που χρησιμοποιήθηκαν, αφορούν τα έτη 2000 έως και 2008 και έγιναν διαθέσιμα από εταιρία φυσικού αερίου της περιοχής. Οι ερευνητές πραγματοποίησαν πρόβλεψη της κατανάλωσης για τα επόμενα 20 χρόνια, χρησιμοποιώντας ως **χαρακτηριστικά** την *ελάχιστη θερμοκρασία* κάθε έτους, τον *αριθμό του έτους* και το *ρυθμό αύξησης του πληθυσμού*.

Τα **μοντέλα** που χρησιμοποιήθηκαν ήταν τα νευρωνικά δίκτυα *πρόσθιας τροφοδότησης* (*feedforward ANN*) δύο κρυμμένων επιπέδων με 12 νευρώνες σε κάθε ένα από αυτά. Για τα κρυμμένα επίπεδα χρησιμοποιήθηκε η *tan-sigmoid function* ως συνάρτηση ενεργοποίησης, ενώ για το επίπεδο εξόδου χρησιμοποιήθηκε η *γραμμική συνάρτηση ενεργοποίησης*.

Καθώς επιχειρήθηκε πρόβλεψη καταναλώσεων για μελλοντικά έτη, έγινε η υπόθεση πως η ελάχιστη θερμοκρασία και ο ρυθμός αύξησης του πληθυσμού κάθε πόλης θα παραμείνουν στα ίδια επίπεδα για την επόμενη εικοσαετία. Όλες οι μεταβλητές κανονικοποιήθηκαν ώστε να ανήκουν στο διάστημα [0,1] χρησιμοποιώντας τη μέθοδο *min-max normalization*. Τα νευρωνικά δίκτυα εκπαιδεύτηκαν και αξιολογήθηκαν χρησιμοποιώντας δεδομένα κατανάλωσης της περιόδου 2000-2008 και στη συνέχεια παρήγαγαν προβλέψεις για τα έτη 2008-2028.

Τα νευρωνικά δίκτυα που σχεδιάστηκαν παρήγαγαν ιδιαίτερα ακριβείς προβλέψεις, το μέγιστο ποσοστιαίο σφάλμα των οποίων έφτασε μόλις το 3%. Το σφάλμα αυτό είναι πλήρως αποδεκτό και αποδεικνύει την ακρίβεια και την **απόδοση** της μεθόδου στη μακροπρόθεσμη πρόβλεψη κατανάλωσης.

Στα **συμπεράσματα**, οι ερευνητές καταλήγουν πως τα νευρωνικά δίκτυα αποτελούν ιδιαίτερα αξιόπιστη μέθοδο για την μακροπρόθεσμη πρόβλεψη κατανάλωσης φυσικού αερίου και άρα ένα εξαιρετικό εργαλείο λήψης αποφάσεων σχετικών με την ενεργειακή πολιτική. Τονίζουν επίσης, πως η ελάχιστη θερμοκρασία αποτελεί το παράγοντα με τη σημαντικότερη επίδραση στα επίπεδα της οικιακής κατανάλωσης φυσικού αερίου. Τέλος, αναφέρουν ότι τα αποτελέσματα της έρευνας προβλέπουν μια ραγδαία αύξηση της κατανάλωσης του φυσικού αερίου στις υπό μελέτη περιοχές κατά τα έτη 2008 έως 2013, κάτι που σταδιακά θα μειώνεται. Επίσης, φαίνεται πως η κατανάλωση θα παραμένει σταθερή μετά το 2028.

2.2.3 Η ηλιακή ακτινοβολία στα μοντέλα πρόβλεψης οικιακής κατανάλωσης ΦΑ

Η εργασία αυτή [SP2014] μελετά την επιρροή της ηλιακής ακτινοβολίας ως χαρακτηριστικό σε μοντέλα πρόβλεψης βραχυπρόθεσμης κατανάλωσης φυσικού αερίου στον οικιακό τομέα. Ως **χαρακτηριστικά** των μοντέλων που δοκιμάστηκαν χρησιμοποιήθηκαν οι *προγενέστερες καταναλώσεις*, η *μέση εξωτερική θερμοκρασία*, η *ηλιακή ακτινοβολία*, καθώς και *ημερολογιακοί παράγοντες*. Η μελέτη πραγματοποιήθηκε χρησιμοποιώντας δύο διαφορετικά σύνολα δεδομένων. Το πρώτο αφορά μια μεμονωμένη κατοικία που χρησιμοποιεί το φυσικό αέριο αποκλειστικά για θέρμανση και το δεύτερο περιέχει δεδομένα κατανάλωσης μιας τοπικής εταιρίας διανομής φυσικού αερίου, οι καταναλωτές της οποίας χρησιμοποιούν επίσης το φυσικό αέριο σχεδόν αποκλειστικά για θέρμανση. Συνεπώς, η κατανάλωση είναι ως επί το πλείστον εξαρτώμενη από τις καιρικές συνθήκες τις περιοχής. Τα

σύνολα δεδομένων αφορούν την περίοδο 2011 έως 2013, ενώ κανένα από τα δύο σύνολα δεδομένων δεν είναι δημοσίως διαθέσιμο.

Στα πλαίσια της έρευνας δοκιμάστηκαν *γραμμικά* και *μη γραμμικά* **μοντέλα**. Τα *γραμμικά μοντέλα* που δοκιμάστηκαν ήταν το *Autoregressive Model with Exogenous Inputs* (*ARX*) και *Stepwise Regression* (*SR*). Αντίστοιχα, τα *μη γραμμικά μοντέλα* ήταν τα *τεχνητά νευρωνικά δίκτυα* (*ANN*) και *Support Vector Regression* (*SVR*). Ως μέτρο αξιολόγησης της ακρίβειας των μοντέλων χρησιμοποιήθηκε ο δείκτης *MARNE*.

Προκειμένου να εξεταστεί κατά πόσο συνεισφέρει στη βελτίωση της απόδοσης των μοντέλων πρόβλεψης η προσθήκη της ηλιακής ακτινοβολίας ως χαρακτηριστικό, τα διάφορα μοντέλα, εκπαιδεύτηκαν τόσο λαμβάνοντας όσο και χωρίς να ληφθεί υπόψη η μεταβλητή αυτή. Οι προβλέψεις που έγιναν ήταν σε χρονικό ορίζοντα ημέρας. Για την εκπαίδευση των μοντέλων χρησιμοποιήθηκαν τα δεδομένα της περιόδου 2011-2012, ενώ η απόδοσή τους αξιολογήθηκε στην περίοδο 2012-2013. Οι μετρήσεις θερμοκρασίας αφορούσαν μέση ημερήσια θερμοκρασία, ενώ οι μετρήσεις ηλιακής ακτινοβολίας και κατανάλωσης φυσικού αερίου ήταν εκφρασμένες σε ημερήσια βάση. Οι ερευνητές χρησιμοποίησαν ως χαρακτηριστικά τόσο τις προγενέστερες τιμές θερμοκρασίας και ηλιακής ακτινοβολίας όπως αυτές έγιναν διαθέσιμες από τα ιστορικά στοιχεία, όσο και μελλοντικές τους τιμές, όπως αυτές προκύπτουν από μοντέλα πρόβλεψης μετεωρολογικών παραγόντων.

Τα γραμμικά μοντέλα είχαν **καλύτερη απόδοση** σε σχέση με τα μη γραμμικά. Ειδικότερα, η μέθοδος *ARX* παρουσίασε την υψηλότερη απόδοση στο σύνολο αξιολόγησης, ξεπερνώντας τις πιο σύνθετες μεθόδους των νευρωνικών δικτύων και SVR. Επίσης, η χρήση της ηλιακής ακτινοβολίας ως επιπλέον μεταβλητή εισόδου των μοντέλων βελτίωσε αισθητά την απόδοσή τους.

Στα **συμπεράσματα** της έρευνας, γίνεται σαφές ότι λαμβάνοντας υπόψη την ηλιακή ακτινοβολία όλα τα μοντέλα παράγουν ακριβέστερες προβλέψεις και γενικεύουν καλύτερα, κατά συνέπεια αυτή πρέπει να λαμβάνεται υπόψη σε επόμενες αντίστοιχες μελέτες. Σημαντικό είναι επίσης το γεγονός ότι στην εργασία αυτή τα, αν και απλούστερα, γραμμικά μοντέλα απέδωσαν καλύτερα από τα πιο περίπλοκα μη γραμμικά. Τέλος, οι ερευνητές τονίζουν τη σημασία της χρήσης μεταγενέστερων τιμών των καιρικών παραγόντων ως χαρακτηριστικά των μοντέλων, καθώς με τον τρόπο αυτό τα μοντέλα προσαρμόζονται καλύτερα σε συνθήκες ρεαλιστικών προβλέψεων.

2.2.4 Τεχνολογίες ANN για ωριαία πρόβλεψη κατανάλωσης ΦΑ

Στην παρούσα εργασία [Szo2015], ο ερευνητής επιχειρεί να κάνει πρόβλεψη ωριαίας κατανάλωσης φυσικού αερίου της πόλης Szczecin στην Πολωνία, με χρήση *τεχνητών νευρωνικών δικτύων* (*ANN*). Ως **χαρακτηριστικά** χρησιμοποιήθηκαν η *θερμοκρασία*, καθώς

και *ημερολογιακοί παράγοντες*, ενώ επιλέχθηκαν να αγνοηθούν διάφοροι κοινωνικοοικονομικοί παράγοντες, καθώς σύμφωνα με τον ερευνητή οι παράγοντες αυτοί δεν έχουν σημαντική επίδραση στη βραχυπρόθεσμη κατανάλωση φυσικού αερίου. Αναλυτικότερα, τα χαρακτηριστικά που χρησιμοποιήθηκαν ως μεταβλητές εισόδου των νευρωνικών δικτύων ήταν η *θερμοκρασία*, ο *μήνας*, η *μέρα του μήνα* καθώς και η *ώρα της ημέρας*. Το σύνολο δεδομένων που χρησιμοποιήθηκε δεν είναι δημοσίως διαθέσιμο και αφορά την περίοδο από 1 Ιανουαρίου 2009 έως και 31 Δεκεμβρίου 2011.

Τα **μοντέλα** που εξετάστηκαν ήταν τα *τεχνητά νευρωνικά δίκτυα* (*ANN*). Συγκεκριμένα, σχεδιάστηκαν και εκπαιδεύτηκαν *νευρωνικά δίκτυα πολλαπλών επιπέδων* (*Multilayer Perceptron* ή *MLP*) ενός κρυμμένου επιπέδου, με αριθμό νευρώνων που κυμάνθηκε μεταξύ 5 και 40. Ως *συνάρτηση ενεργοποίησης* (*activation function*) χρησιμοποιήθηκε η *υπερβολική εφαπτομένη* (*hyperbolic tangent function*) για το κρυμμένο επίπεδο και η *εκθετική συνάρτηση* (*exponential function*) για το επίπεδο εξόδου. Στη συνέχεια, τα νευρωνικά δίκτυα αυτά συγκρίθηκαν ως προς την ακρίβεια των προβλέψεών τους. Η απόδοση των νευρωνικών δικτύων αξιολογήθηκε με χρήση των δεικτών *R, RMSE* και *MAPE*.

Το παραπάνω σύνολο δεδομένων χωρίστηκε σε τρία επιμέρους ανεξάρτητα υποσύνολα, τα οποία περιείχαν διαφορετικό αριθμό καταγραφών. Τα δύο απ' αυτά χρησιμοποιήθηκαν για την εκπαίδευση των νευρωνικών και το τελευταίο για τον έλεγχο της απόδοσής τους. Με τον τρόπο αυτό, νευρωνικά δίκτυα ίδιας αρχιτεκτονικής εκπαιδεύτηκαν ανεξάρτητα, με διαφορετικό αριθμό δειγμάτων, με αποτέλεσμα να παρουσιάσουν διαφορετικό συντελεστή συσχέτισης *R*. Ο ερευνητής παρατήρησε πως όσο αυξανόταν ο αριθμός των νευρώνων του κρυμμένου επιπέδου, τόσο βελτιωνόταν η ακρίβεια των προβλέψεων και αυξανόταν ο συντελεστής συσχέτισης. Επιπλέον, παρατήρησε ότι σε νευρωνικά δίκτυα με μικρό αριθμό νευρώνων στο κρυμμένο επίπεδο, ο συντελεστής συσχέτισης ήταν μεγαλύτερος όταν αυτά εκπαιδεύονταν στο σύνολο με το μικρότερο αριθμό δειγμάτων. Αντίθετα, όσο αυξανόταν ο αριθμός των νευρώνων αυτών, ο συντελεστής συσχέτισης ήταν μεγαλύτερος στα νευρωνικά δίκτυα που εκπαιδεύτηκαν στο μεγαλύτερο σε πλήθος καταγραφών σύνολο.

Τα τρία νευρωνικά δίκτυα που παρουσίασαν την **καλύτερη απόδοση** ήταν το νευρωνικό με 25 νευρώνες στο κρυμμένο επίπεδο, το οποίο εκπαιδεύτηκε στο σύνολο με το μικρότερο αριθμό δειγμάτων, ενώ επίσης τα νευρωνικά με 25 και 36 νευρώνες στο κρυμμένο επίπεδο, τα οποία εκπαιδεύτηκαν στο σύνολο με το μεγαλύτερο αριθμό δειγμάτων. Ο ερευνητής σημειώνει ότι καλύτερη συμφωνία μεταξύ των προβλεπόμενων από τα νευρωνικά τιμών και των πραγματικών καταναλώσεων υπήρχε τους καλοκαιρινούς μήνες, όταν η κατανάλωση ήταν αισθητά μικρότερη. Αντίθετα, το σφάλμα σε όρους *RMSE* ήταν μεγαλύτερο τους πρώτους χειμερινούς μήνες, λόγω των αιφνίδιων αλλαγών στα επίπεδα

κατανάλωσης. Σημαντικό είναι επίσης να αναφερθεί ότι τα νευρωνικά που εκπαιδεύτηκαν με το σύνολο με τις περισσότερες καταγραφές είχαν καλύτερη απόδοση στις προβλέψεις.

Στα **συμπεράσματα,** ο ερευνητής καταλήγει σημειώνοντας πως η απόδοση των νευρωνικών δικτύων ως μοντέλα πρόβλεψης ξεπερνούν αυτή των απλούστερων μοντέλων, ενώ θεωρεί πως η χρήση υβριδικών μοντέλων και άλλων τεχνικών τεχνητής νοημοσύνης θα συνεισφέρει σε ακόμα ακριβέστερες προβλέψεις.

2.2.5 Πρόβλεψη κατανάλωσης ΦΑ με χρήση μοντέλων μηχανικής μάθησης

Στην εργασία αυτή [SO2018] μελετάται η απόδοση διαφόρων μοντέλων μηχανικής μάθησης στο πρόβλημα της πρόβλεψης της ημερήσιας κατανάλωσης φυσικού αερίου, χρησιμοποιώντας ως **χαρακτηριστικά** τη *μέση ημερήσια θερμοκρασία*, *ημερολογιακούς παράγοντες* και *προγενέστερες καταναλώσεις*. Τα μοντέλα που χρησιμοποιήθηκαν θεωρούνται σχετικά απλά, καθώς δε δοκιμάστηκαν υβριδικά ή μοντέλα ensemble learning. Ωστόσο, η ακρίβεια των προβλέψεων ήταν συγκρίσιμη με αυτή αντίστοιχων ερευνών. Το σύνολο δεδομένων που χρησιμοποιήθηκε δεν είναι δημοσίως διαθέσιμο και αφορά την περίοδο από 23 Φεβρουαρίου 2013 έως 30 Σεπτεμβρίου 2016 μιας κομητείας στην Κροατία.

Τα **μοντέλα** πρόβλεψης που δοκιμάστηκαν ήταν τα *Decision Trees, Support Vector Regression (SVR), Linear* και *Lasso Regression* καθώς και *τεχνητά νευρωνικά δίκτυα* (*ANN*). Ως μέτρο της ποιότητας πρόβλεψης των μοντέλων χρησιμοποιήθηκε ο δείκτης *RMSE*. Ένα μοντέλο θα θεωρηθεί ότι παράγει ικανοποιητικές προβλέψεις αν ο δείκτης αυτός είναι εντός του πεδίου $\pm 5\%$ της διαφοράς μεταξύ της μέγιστης και ελάχιστης κατανάλωσης.

Το σύνολο δεδομένων που χρησιμοποιήθηκε χωρίστηκε σε δύο επιμέρους σύνολα, τα fitting και test sets, τα οποία χρησιμοποιήθηκαν για την εκπαίδευση και την αξιολόγηση των μοντέλων αντίστοιχα. Το fitting set, περιείχε δεδομένα από τις 23 Φεβρουαρίου 2013 έως τις 30 Σεπτεμβρίου 2016 και χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων και για το στάδιο του cross validation με τον αλγόριθμο *5-fold cross validation*. Στο στάδιο αυτό υπολογίστηκαν και οι βέλτιστες τιμές των υπερπαραμέτρων των μοντέλων. Το test set, περιείχε δεδομένα από 1 Οκτωβρίου 2016 έως 30 Σεπτεμβρίου 2017 και χρησιμοποιήθηκε για την αξιολόγηση των μοντέλων.

Τα μοντέλα που παρουσίασαν την **καλύτερη απόδοση** ήταν τα *Lasso Regression*, *νευρωνικό δίκτυο πρόσθιας τροφοδότησης (feedforward ANN)* και *SVR*, με το *Lasso Regression*, παρά την απλότητά του, να ξεπερνά σε όρους *RMSE* όλα τα υπόλοιπα. Όλα τα μοντέλα παρήγαγαν προβλέψεις με αποκλίσεις εντός των αποδεκτών ορίων, με εξαίρεση να αποτελεί η μέθοδος *Decision Trees*, γεγονός που ήταν αναμενόμενο λόγω της απλότητας της μεθόδου.

Στα **συμπεράσματα**, οι ερευνητές καταλήγουν ότι πιθανός τρόπος βελτίωσης της απόδοσης των μοντέλων είναι η προσθήκη επιπλέον καιρικών, ακόμα και κοινωνικοοικονομικών παραγόντων ως χαρακτηριστικά των μοντέλων. Παράλληλα, η χρήση πιο σύνθετων μεθόδων, όπως τα υβριδικά ή τα μοντέλα ensemble learning, είναι επίσης πιθανό να συνεισφέρει στη βελτίωση της ακρίβειας των προβλέψεων.

2.2.6 Βραχυπρόθεσμη πρόβλεψη κατανάλωσης ΦΑ με χρήση Βαθιάς Μάθησης

Στην εργασία αυτή [MP2018] εξετάζεται η απόδοση των *Deep Neural Networks* (*DNN*) στο πρόβλημα της βραχυπρόθεσμης πρόβλεψης κατανάλωσης φυσικού αερίου. Ως **χαρακτηριστικά** των μοντέλων χρησιμοποιήθηκαν οι καιρικοί παράγοντες *σημείο δρόσου, HDD* (heating degree days) και *CDD* (cooling degree days) καθώς και οι ημερολογιακοί παράγοντες *ημέρα της εβδομάδας* και *ημέρα του χρόνου*. Ως μέτρο της ποιότητας των προβλέψεων χρησιμοποιήθηκε ο δείκτης *WMAPE*. Το σύνολο δεδομένων που χρησιμοποιήθηκε αφορά μηνιαίες καταναλώσεις 62 περιοχών των Η.Π.Α για περίοδο 10 ετών.

Τα **μοντέλα** πρόβλεψης που δοκιμάστηκαν ήταν τα *DNN, ANN* και *Linear Regression* (*LR*). Οι ερευνητές σημειώνουν ότι οι βασικές διαφορές μεταξύ των *ANN* και *DNN* είναι ο αριθμός των επιπέδων που περιέχουν, με τα *DNN* να περιέχουν μεγαλύτερο αριθμό, καθώς και ο αλγόριθμος εκπαίδευσης που χρησιμοποιούν.

Τα μοντέλα εκπαιδεύτηκαν με δεδομένα 10 ετών, ενώ παρήγαγαν προβλέψεις για ένα έτος για 62 περιοχές των Η.Π.Α. Το *ANN* που σχεδιάστηκε ήταν ένα νευρωνικό δίκτυο δύο κρυμμένων επιπέδων με 12 και 4 νευρώνες αντίστοιχα, ενώ η εκπαίδευσή του έγινε με χρήση του αλγορίθμου οπισθοδιάδοσης του σφάλματος (backpropagation). Στη συνέχεια σχεδιάστηκε ένα *DNN* ίδιας αρχιτεκτονικής με το παραπάνω *ANN*, στο οποίο όμως έγινε χρήση του αλγορίθμου contrastive divergence algorithm. Σκοπός της δημιουργίας του *DNN* αυτού, ήταν να διαπιστωθεί κατά πόσο η χρήση διαφορετικής μεθόδου εκπαίδευσης των νευρωνικών βελτιώνει την απόδοση των προβλέψεων. Τέλος, σχεδιάστηκε ένα μεγαλύτερο *DNN,* 4 κρυμμένων επιπέδων με 60 νευρώνες για τα τρία πρώτα και 12 νευρώνες για το τέταρτο κρυμμένο επίπεδο. Έτσι, ήταν δυνατόν να διαπιστωθεί κατά πόσο θα βελτιώσει την απόδοση των τελικών προβλέψεων η χρήση μιας περίπλοκης αρχιτεκτονικής ενός μεγαλύτερου νευρωνικού δικτύου. Τα νευρωνικά δίκτυα που σχεδιάστηκαν, καθώς και η μέθοδος *LR* συγκρίθηκαν ως προς την απόδοσή τους στο σύνολο αξιολόγησης.

Τα *DNN* παρουσίασαν την **καλύτερη απόδοση** ως προς το μέσο όρο του δείκτη *WMAPE* για τις 62 περιοχές, έχοντας μάλιστα τη μικρότερη τυπική απόκλιση σε σχέση με τις άλλες δύο μεθόδους. Το γεγονός αυτό αναδεικνύει την υπεροχή των πιο περίπλοκων μοντέλων, όπως τα *DNN* σε σχέση με απλούστερες μεθόδους. Σημαντικό είναι επίσης να

αναφερθεί ότι το *DNN* με την πιο σύνθετη αρχιτεκτονική ξεπέρασε το απλούστερο, παρουσιάζοντας μικρότερο μέσο σφάλμα καθώς και μικρότερη τυπική απόκλιση για τις 62 περιοχές. Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός ότι, σε ορισμένες από τις 62 περιοχές η απλή μέθοδος *LR* ξεπέρασε σε απόδοση όλα τα υπόλοιπα μοντέλα.

Στα **συμπεράσματα** οι ερευνητές επισημαίνουν την υπεροχή των *DNN* στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου σε σχέση με τις απλούστερες μεθόδους των *ANN* και *LR*, με το *DNN* μάλιστα με την πιο περίπλοκη αρχιτεκτονική να παρουσιάζει την καλύτερη απόδοση όλων. Ωστόσο, καταλήγουν σημειώνοντας ότι σε ορισμένες περιπτώσεις, ακόμα και ιδιαίτερα απλές μέθοδοι, όπως η *LR*, είναι δυνατόν να αποδώσουν καλύτερα από οποιαδήποτε άλλη πιο σύνθετη μέθοδο και επομένως πρέπει κι αυτές να λαμβάνονται υπόψη.

2.2.7 Σύγκριση μεθόδων πρόβλεψης οικιακής κατανάλωσης ΦΑ

Στην εργασία αυτή [HP2018] διάφορα μοντέλα πρόβλεψης δοκιμάζονται στο πρόβλημα της ωριαίας πρόβλεψης οικιακής κατανάλωσης φυσικού αερίου. Ως **χαρακτηριστικά** των μοντέλων χρησιμοποιήθηκε η *θερμοκρασία*, λαμβάνοντας υπόψη τόσο τις πραγματικές όσο και τις προβλεπόμενες από μετεωρολογικά μοντέλα τιμές της, καθώς και οι ημερολογιακοί παράγοντες, *ώρα της ημέρας, ώρα της εβδομάδας, ημέρα διακοπών, ημέρα μεταξύ διακοπών και Σαββατοκύριακου*. Το σύνολο δεδομένων που χρησιμοποιήθηκε αφορά την οικιακή κατανάλωση της πόλης Ljubljana στη Σλοβενία. Περιλαμβάνει δεδομένα κατανάλωσης 8 χειμερινών περιόδων, φτάνοντας τις 36.106 εγγραφές.

Τα **μοντέλα** πρόβλεψης που δοκιμάστηκαν ήταν τα *Linear Regression* (*LR*), *Kernel Machine* (*KM*), *Kernel Machine with Memory* (*KMM*) και *Recurrent Neural Network* (*RNN*) καθώς και κάποια *εμπειρικά μοντέλα* που αναπτύχθηκαν βασισμένα στην ανάλυση των δεδομένων. Οι δείκτες που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων ήταν οι *MAE* και *MAPE*.

Τα μοντέλα εκπαιδεύτηκαν με τα δεδομένα 7 περιόδων και παρήγαν προβλέψεις για την απομείνασα περίοδο. Αυτό υλοποιήθηκε για όλους τους συνδυασμούς των 8 περιόδων, εκτελώντας δηλαδή ένα κύκλο 8 ανεξάρτητων επαναλήψεων. Επιπλέον, η εκπαίδευση των μοντέλων έγινε δύο φορές, με χρήση των πραγματικών και των προβλεπόμενων τιμών της θερμοκρασίας, προκειμένου να αναδειχθεί η επίπτωση της χρήσης των δεύτερων στην απόδοση των μοντέλων. Οι προβλέψεις που παρήγαγαν τα μοντέλα ήταν βραχυπρόθεσμες, σε χρονικό ορίζοντα μέχρι 60 ωρών.

Τα μοντέλα με την **καλύτερη απόδοση** ήταν τα *RNN* και *LR*. Ειδικότερα, το μοντέλο *RNN*, παρουσίασε το μικρότερο σφάλμα, το οποίο μάλιστα παρέμενε σταθερό, ανεξάρτητα από τον χρονικό ορίζοντα πρόβλεψης. Αντίθετα, το σφάλμα της μεθόδου *LR* είχε αισθητή αύξηση όσο ο χρονικός ορίζοντας πρόβλεψης αυξανόταν. Τα *εμπειρικά μοντέλα* που

σχεδιάστηκαν από τους ερευνητές καθώς και η μέθοδος *KM* παρουσίασαν μέτρια απόδοση, ενώ η μέθοδος *KMM* ήταν η λιγότερο ακριβής. Το σφάλμα των παραπάνω μοντέλων αυξανόταν όσο μεγάλωνε ο χρονικός ορίζοντας των προβλέψεων, ενώ όταν απαιτούνταν μόνο ημερήσια πρόβλεψη, τα σφάλματα είχαν την ελάχιστη τιμή τους. Σημαντικό είναι επίσης να αναφερθεί, ότι όλα τα μοντέλα είχαν μέγιστο σφάλμα κατά την πρόβλεψη της κατανάλωσης «ειδικών ημερών», όπως οι διακοπές και οι αργίες. Εξαίρεση αποτέλεσε η *LR*, η οποία κατά τις περιόδους αυτές ξεπερνούσε σε απόδοση ακόμα και το *RNN*. Τα αποτελέσματα έδειξαν επίσης, ότι τόσο *εμπειρικά μοντέλα*, όσο και το *RNN* παρουσίασαν καλύτερη απόδοση όταν εκπαιδεύτηκαν με χρήση των προβλεπόμενων της θερμοκρασίας. Το γεγονός αυτό δεν ήταν αναμενόμενο, καθώς τα σφάλματα στις προβλέψεις της θερμοκρασίας δημιουργούν θόρυβο και κατά συνέπεια μείωση της ακρίβειας των τελικών προβλέψεων.

Στα **συμπεράσματα**, οι ερευνητές επισημαίνουν την υπεροχή των μη γραμμικών μοντέλων στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου. Τα μοντέλα αυτά μάλιστα, φαίνεται να αποδίδουν καλύτερα όταν η εκπαίδευσή τους γίνεται με χρήση των προβλεπόμενων τιμών της θερμοκρασίας αντί των πραγματικών. Εξαίρεση αποτελεί η μέθοδος *LR*, η οποία παρά την απλότητά της παρουσίασε ικανοποιητικά αποτελέσματα, γεγονός που ενδεχομένως οφείλεται στην ευκολία εκπαίδευσής της. Μειονέκτημα του μοντέλου *RNN*, αποτελεί η ανάγκη του για μεγάλο όγκο δεδομένων για την εκπαίδευσή του. Τέλος, τα εμπειρικά μοντέλα ήταν λιγότερο ακριβή, ωστόσο απαιτούνται ελάχιστοι πόροι για την εκπαίδευσή τους. Οι ερευνητές καταλήγουν πως η χρήση του μοντέλου *RNN*, σε συνδυασμό με τη μέθοδο *LR*, αποτελεί ένα ιδιαίτερο αξιόπιστο εργαλείο πρόβλεψης κατανάλωσης φυσικού αερίου.

2.2.8 Βραχυπρόθεσμη πρόβλεψη κατανάλωσης ΦΑ με χρήση υβριδικής μεθόδου

Στην εργασία αυτή [WL2018] προτείνεται ένα νέο υβριδικό μοντέλο για την βραχυπρόθεσμη πρόβλεψη κατανάλωσης φυσικού αερίου. Το μοντέλο αυτό ονομάζεται *FSA-LGA-SVR* και συνδυάζει τους αλγορίθμους *Factor Selection Algorithm* (*FSA*) και *Life Genetic Algorithm* (*LGA*) με τη μέθοδο πρόβλεψης *Support Vector Regression* (*SVR*). Ως **χαρακτηριστικά** των μοντέλων χρησιμοποιήθηκαν η ελάχιστη, μέση και μέγιστη τιμή της *ημερήσιας θερμοκρασίας, σημείου δρόσου, υγρασίας, ορατότητας, ατμοσφαιρικής πίεσης, ταχύτητας ανέμου,* καθώς και η *ημερήσια* τιμή του *υετού.* Το σύνολο δεδομένων που χρησιμοποιήθηκε αφορά τρεις πόλεις της Ελλάδας (Αθήνα, Θεσσαλονίκη και Λάρισα) και έγινε διαθέσιμο από το ΔΕΣΦΑ. Τα δεδομένα αφορούν καταναλώσεις των πόλεων για την περίοδο 1/1/2015 έως 31/12/2017.

Το **μοντέλα** που δοκιμάστηκαν ήταν τα *ANN, GA-SVR, LGA-SVR* και *FSA-LGA-SVR.* Ο αλγόριθμος *FSA* χρησιμοποιείται για την αυτόματη επιλογή των παραγόντων εκείνων που

επηρεάζουν σημαντικά την κατανάλωση φυσικού αερίου, μέσω της ανάλυσης δεικτών συσχέτισης (correlation coefficient analysis). Ο αλγόριθμος *LGA*, πρόκειται για μια τροποποιημένη από τους ερευνητές μορφή του *Genetic Algorithm* (*GA*) και χρησιμοποιήθηκε για την εύρεση των βέλτιστων τιμών των υπερπαραμέτρων του μοντέλου *SVR.*

Τα πρώτα δύο έτη του συνόλου δεδομένων (1/1/2015 έως 31/12/2016) χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων, ενώ το τελευταίο έτος (1/1/2017 έως 31/12/2017) για την αξιολόγησή τους. Οι ερευνητές επισημαίνουν ότι η κατανάλωση φυσικού αερίου παρουσιάζει έντονη εποχιακότητα και κατά συνέπεια οι παράγοντες που την επηρεάζουν δεν είναι οι ίδιοι σε κάθε περίοδο. Έτσι, το αρχικό σύνολο δεδομένων χωρίστηκε σε 12 επιμέρους υποσύνολα, ένα για κάθε μήνα του έτους. Στη συνέχεια, ο αλγόριθμος *FSA* εφαρμόστηκε σε κάθε ένα μήνα προκειμένου να εξάγει τους διαφορετικούς παράγοντες που επηρεάζουν περισσότερο την κατανάλωση σε κάθε έναν από αυτούς. Στη συνέχεια, το μοντέλο πρόβλεψης *SVR* εκπαιδεύτηκε και παρήγαγε προβλέψεις, χρησιμοποιώντας ως χαρακτηριστικά τους παράγοντες εκείνους που εξήχθησαν κατά την προηγούμενη διαδικασία. Στο στάδιο αυτό, οι βέλτιστες τιμές των υπερπαραμέτρων του *SVR* υπολογίστηκαν με τη βοήθεια των αλγορίθμων *GA* και *LGA.*

Οι ερευνητές αναφέρουν πως η χρήση του αλγορίθμου *LGA* έναντι του *GA* στο στάδιο υπολογισμού των υπερπαραμέτρων, βελτιώνει σημαντικά την απόδοση του *SVR.* Ακόμα, το μοντέλο *LGA-SVR* παράγει πιο ακριβείς προβλέψεις και από το *ANN.* Σημαντικό είναι επίσης να αναφερθεί, ότι η χρήση του αλγορίθμου *FSA,*όταν αυτή γίνεται στο σύνολο του έτους αυξάνει το σφάλμα των προβλέψεων. Αντίθετα, όταν εφαρμόζεται σε κάθε ένα μήνα ξεχωριστά, η τελική απόδοση του μοντέλου *FSA-LGA-SVR* βελτιώνεται αισθητά. Επομένως, το μοντέλο *FSA-LGA-SVR* είχε την **καλύτερη απόδοση** σε σχέση με τα υπόλοιπα μοντέλα που δοκιμάστηκαν.

Στα **συμπεράσματα** οι ερευνητές αναφέρουν ότι το προτεινόμενο υβριδικό μοντέλο *FSA-LGA-SVR* παράγει ιδιαίτερα ακριβείς προβλέψεις στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου. Τονίζουν ακόμα τη σημασία της ξεχωριστής επιλογής παραγόντων που επηρεάζουν την κατανάλωση για τις διαφορετικές περιόδους του χρόνου, καθώς αυτό αυξάνει σημαντικά την απόδοση των μοντέλων. Τέλος, επισημαίνουν ότι τα μεγαλύτερα σφάλματα παρήχθησαν κατά τους χειμερινούς μήνες. Προκειμένου να ελαχιστοποιηθούν τα σφάλματα αυτά, είναι σημαντικό να χρησιμοποιηθούν ως χαρακτηριστικά των μοντέλων παράγοντες επιπλέον των καιρικών.

2.2.9 Ημερήσια πρόβλεψη κατανάλωσης ΦΑ με εφαρμογή υβριδικής μεθόδου (Ι)

Στην εργασία αυτή [WL2019] σχεδιάστηκε ένα καινοτόμο υβριδικό μοντέλο με σκοπό την ημερήσια πρόβλεψη κατανάλωσης φυσικού αερίου. Το υβριδικό αυτό μοντέλο

ονομάζεται ISSA-LSTM και συνδυάζει τη μέθοδο *ISSA* και το τεχνητό νευρωνικό δίκτυο *Long Short-Term Memory* (*LSTM*). Ως **χαρακτηριστικά** των μοντέλων χρησιμοποιήθηκαν η ελάχιστη, μέση και μέγιστη τιμή της *ημερήσιας θερμοκρασίας, σημείου δρόσου, υγρασίας, ορατότητας, ατμοσφαιρικής πίεσης, ταχύτητας ανέμου,* καθώς και η *ημερήσια* τιμή του *υετού.* Επιπλέον, η *ημέρα της εβδομάδας,* ο *μήνας του έτος*, ο *αριθμός του έτους* καθώς και η *ημερήσια τιμή του φυσικού αερίου* λήφθηκαν υπόψη. Στα πλαίσια της εργασίας χρησιμοποιήθηκαν δεδομένα τεσσάρων πόλεων, οι οποίες ανήκουν σε τρεις κλιματικές ζώνες. Οι πόλεις αυτές είναι το Λονδίνο, η Μελβούρνη, το Χονγκ-Κονγκ και η Καρδίτσα και τα δεδομένα κατανάλωσης αφορούν την περίοδο 1/1/2015 έως 30/5/2018.

Τα **μοντέλα** που δοκιμάστηκαν ήταν τα *ISSA-LSTM, SSA-LSTAM, LSTM, linear regression* (*LR*), *Support Vector Regression* (*SVR*) και *Back-Propagation Neural Network* (*BPNN*). Το *ISSA* πρόκειται για μια τροποποιημένη μορφή της γνωστής μεθόδου *Singular Spectrum Analysis* (*SSA*), η οποία χρησιμοποιείται στην αποσύνθεση χρονοσειρών για την αφαίρεση του θορύβου. Η μέθοδος SSA, λόγω της ντετερμινιστικής φύσης της, αποτυγχάνει να δώσει ικανοποιητικά αποτελέσματα όταν η χρονοσειρά περιέχει υψηλά ποσοστά θορύβου. Αποτέλεσμα αυτού, είναι η δημιουργία αρνητικών τιμών κατανάλωσης φυσικού αερίου κατά την ανακατασκευή της χρονοσειράς. Το γεγονός αυτό μειώνει την απόδοση των μοντέλων πρόβλεψης και επομένως δημιουργείται η ανάγκη η δημιουργία μιας καλύτερης μεθόδου αποσύνθεσης και ανακατασκευής της χρονοσειράς, όπως η προτεινόμενη από τους ερευνητές μέθοδος ISSA. Οι δείκτες που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων ήταν οι R^2, *MAE, MAPE, RMSE* και *MARNE*.

Τα δεδομένα της περιόδου 1/1/2015 έως 1/1/2018 κάθε πόλης χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων, ενώ τα δεδομένα από 1/1/2018 έως 30/5/2018 χρησιμοποιήθηκαν για την αξιολόγησή τους. Τόσο η μέθοδος SSA, όσο και η ISSA χρησιμοποιήθηκαν στο στάδιο της προεπεξεργασίας για την αποσύνθεση και ανακατασκευή της χρονοσειράς των καταναλώσεων, με σκοπό την αφαίρεση του θορύβου από τα ιστορικά δεδομένα. Στη συνέχεια, η ανακατασκευασμένη αυτή χρονοσειρά χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου LSTM, παράγοντας αντίστοιχα τα υβριδικά μοντέλα SSA-LSTM και ISSA-LSTM. Επιπρόσθετα, εκπαιδεύτηκαν και δοκιμάστηκαν τα μοντέλα LR, SVR, LSTM και BPNN. Η παραγωγή προβλέψεων από κάθε μοντέλα επαναλήφθηκε 30 φορές προκειμένου να ληφθεί ο μέσος όρος των σφαλμάτων.

Τα αποτελέσματα της εργασίας ανέδειξαν το προτεινόμενο υβριδικό μοντέλο *ISSA-LSTM*, ως το μοντέλο με την **καλύτερη απόδοση**. Ειδικότερα, τα τρία βασισμένα στο *LSTM* μοντέλα, είχαν το μικρότερο σφάλμα προβλέψεων σε όλες τις πόλεις, ενώ η χρήση της μεθόδου ISSA έναντι της SSA οδήγησε σε περαιτέρω αύξηση της απόδοσης. Επιπλέον, η απόδοση του *ISSA-LSTM* παρέμεινε σε σταθερά επίπεδα σε όρους MARNE στις διαφορετικές

πόλεις, για τις οποίες παρήχθησαν προβλέψεις. Κατά συνέπεια, το μοντέλο αυτό μπορεί να χρησιμοποιηθεί για παραγωγή προβλέψεων κατανάλωσης περιοχών που ανήκουν σε διαφορετικές κλιματικές ζώνες. Ενδιαφέρον παρουσιάζει το γεγονός ότι η απόδοση της μεθόδου *LR* ήταν συγκρίσιμη, ακόμα και καλύτερη, με αυτή των πιο σύνθετων μοντέλων *SVR* και *BPNN*.

Στα **συμπεράσματα**, οι ερευνητές καταλήγουν επισημαίνοντας την υπεροχή της μεθόδου *ISSA* έναντι της *SSA* σε χρονοσειρές κατανάλωσης φυσικού αερίου. Ακόμα, ο συνδυασμός της μεθόδου αυτής με το μοντέλο *LSTM*, οδηγεί σε υψηλής ακρίβειας προβλέψεις, ενώ μάλιστα το υβριδικό μοντέλο *ISSA-LSTM* που προτάθηκε, είναι ικανό να παράγει ακριβείς προβλέψεις, ακόμα και για περιοχές που ανήκουν σε διαφορετικές κλιματικές ζώνες και άρα διαθέτουν αρκετά διαφορετικά επίπεδα κατανάλωσης φυσικού αερίου και καταναλωτικές συνήθειες.

2.2.10 Ημερήσια πρόβλεψη κατανάλωσης ΦΑ με εφαρμογή υβριδικής μεθόδου (II)

Στην εργασία αυτή [WD2019] προτείνεται ένα νέο υβριδικό μοντέλο για την ακριβή πρόβλεψη της κατανάλωσης φυσικού αερίου, συνδυάζοντας τον αλγόριθμο *Principal Component Correlation Analysis* (*PCCA*) και το τεχνητό νευρωνικό δίκτυο *Long Short-Term Memory* (*LSTM*). Ως **χαρακτηριστικά** των μοντέλων χρησιμοποιήθηκαν η ελάχιστη, μέση και μέγιστη τιμή της *ημερήσιας θερμοκρασίας, σημείου δρόσου, υγρασίας, ορατότητας, ατμοσφαιρικής πίεσης, ταχύτητας ανέμου,* καθώς και η *ημερήσια* τιμή του *υετού.* Τα σύνολα δεδομένων που χρησιμοποιήθηκαν αφορούσαν τις πόλεις Xi'an της Κίνας και Αθήνα της Ελλάδος. Τα δεδομένα της πόλης Xi'an έγιναν διαθέσιμα από εταιρία φυσικού αερίου της πόλης, ενώ τα δεδομένα της Αθήνας από το ΔΕΣΦΑ.

Τα **μοντέλα** που δοκιμάστηκαν ήταν τα *SVR, BPNN, LSTM, PCA-LSTM* και το προτεινόμενο *PCCA-LSTM*. Οι βέλτιστες τιμές των υπερπαραμέτρων των προηγούμενων μοντέλων υπολογίστηκαν με τη βοήθεια του αλγορίθμου *Genetic Algorithm* (*GA*). Η μέθοδος *Principal Component Analysis* (*PCA*), αποτελεί μια μαθηματική διαδικασία, η οποία μετατρέπει συσχετισμένες μεταβλητές σε έναν αριθμό ασυσχέτιστων μεταξύ τους μεταβλητών. Η μέθοδος *PCCA*, είναι μια τροποποιημένη από τους ερευνητές μορφή της *PCA*. Οι δυο αυτές μέθοδοι, χρησιμοποιήθηκαν στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου, προκειμένου να εξάγουν τα συστατικά εκείνα των παραγόντων που επηρεάζουν την κατανάλωση, μειώνοντας παράλληλα το μέγεθος του συνόλου δεδομένων. Οι δείκτες που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων ήταν οι *MAE, MAPE, RMSE* και *MARNE*.

Τα σύνολα δεδομένων που κάθε πόλης, χωρίστηκαν σε δυο επιμέρους υποσύνολα για την εκπαίδευση και την αξιολόγηση των μοντέλων αντίστοιχα. Για την πόλη Xi'an, το

σύνολο εκπαίδευσης περιείχε δεδομένα από 1/1/2015 έως 31/9/2017, ενώ το σύνολο αξιολόγησης από 1/9/2017 έως 30/9/2017. Αντίστοιχα για την Αθήνα, το σύνολο εκπαίδευσης περιείχε δεδομένα από 1/1/2015 έως 31/12/2017 και το σύνολο αξιολόγησης από 1/1/2018 έως 28/2/2018. Κατά το στάδιο της εκπαίδευσης, τόσο η μέθοδος *PCA*, όσο και η *PCCA* εφαρμόστηκαν κατά τη διαδικασία επιλογής των παραγόντων που επηρεάζουν την κατανάλωση. Κάθε μέθοδος εξήγαγε διαφορετικούς παράγοντες, οι οποίοι στη συνέχεια αποτέλεσαν τα χαρακτηριστικά του μοντέλου *LSTM*, δημιουργώντας τα υβριδικά μοντέλα *PCA-LSTM* και *PCCA-LSTM* αντίστοιχα.

Η χρήση των μεθόδων *PCA* και *PCCA* για την εξαγωγή των παραγόντων που επηρεάζουν την κατανάλωση και τη μείωση του μεγέθους του αρχικού συνόλου δεδομένων μειώνει σημαντικά τον απαιτούμενο υπολογιστικό χρόνο. Οι ερευνητές αναφέρουν ότι και στα δύο case studies, το προτεινόμενο μοντέλο *PCCA-LSTM* παρουσίασε την **καλύτερη απόδοση**. Το *BPNN* είχε το μεγαλύτερο σφάλμα προβλέψεων, ενώ όλα τα υπόλοιπα μοντέλα θεωρήθηκαν ιδιαίτερα ακριβή. Είναι σημαντικό να αναφερθεί ότι η χρήση του αλγορίθμου *PCA* για την εξαγωγή των παραμέτρων, αύξησε το σφάλμα των τελικών προβλέψεων. Αυτό συνέβη, καθώς ο αλγόριθμος αυτός παρέλειπε δευτερεύοντες παράγοντες που είχαν συσχέτιση με την κατανάλωση. Αντίθετα, η χρήση του τροποποιημένου αλγόριθμου *PCCA* βελτίωσε την απόδοση των προβλέψεων.

Στα **συμπεράσματα** οι ερευνητές επισημαίνουν την υπεροχή του αλγορίθμου *PCCA*, η χρήση του οποίου βελτίωνε την απόδοση των μοντέλων πρόβλεψης. Ακόμα, αναφέρουν ότι το τεχνητό νευρωνικό δίκτυο *LSTM* παρουσίασε καλύτερη απόδοση από τα μοντέλα *BPNN* και *SVR*, γεγονός που αναδεικνύει την υπεροχή του στο πρόβλημα της ημερήσιας πρόβλεψης κατανάλωσης φυσικού αερίου. Τέλος, καταλήγουν σημειώνοντας ότι το προτεινόμενο υβριδικό μοντέλο *PCCA-LSTM* αποτελεί ένα εύρωστο μοντέλο, το οποίο έχει χαμηλές υπολογιστικές απαιτήσεις, ενώ η ακρίβειά του στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου είναι υψηλή.

2.2.11 Μηνιαία πρόβλεψη κατανάλωσης ΦΑ με χρήση μοντέλων μηχ. μάθησης

Στην εργασία αυτή [BE2019] δοκιμάζονται τρεις μέθοδοι μηχανικής μάθησης για την μηνιαία πρόβλεψη της κατανάλωσης στην πόλη Istanbul της Τουρκίας. Πρόκειται για την πόλη με τη μεγαλύτερη κατανάλωση φυσικού αερίου στην Τουρκία. Τα **χαρακτηριστικά** που χρησιμοποιήθηκαν για την πρόβλεψη ήταν η *θερμοκρασία*, η *τιμή του φυσικού αερίου*, η *τιμή του πετρελαίου*, ο *πληθυσμός της περιοχής*, η *συναλλαγματική ισοτιμία*, οι *προγενέστερες καταναλώσεις*, το *ΑΕΠ* καθώς και *ημερολογιακοί παράγοντες*. Το σύνολο δεδομένων που χρησιμοποιήθηκε αφορά μηνιαίες καταναλώσεις για τα έτη 2004-2015 και έγινε διαθέσιμο στους ερευνητές από το μεγαλύτερο διανομέα φυσικού αερίου της Τουρκίας.

Τα **μοντέλα** πρόβλεψης που δοκιμάστηκαν ήταν η *Linear Regression* (*LR*), *Support Vector Regression* (*SRV*) και *τεχνητά νευρωνικά δίκτυα* (*ANN*). Ως μέτρο της ποιότητας πρόβλεψης των μοντέλων χρησιμοποιήθηκαν οι δείκτες *MAPE* και *MSE*.

Το σύνολο δεδομένων που χρησιμοποιήθηκε χωρίστηκε σε δύο επιμέρους σύνολα, τα οποία χρησιμοποιήθηκαν για την εκπαίδευση και την αξιολόγηση των μοντέλων αντίστοιχα. Το σύνολο που χρησιμοποιήθηκε για την αξιολόγηση, περιείχε δεδομένα από 1 Νοεμβρίου 2005 έως 30 Οκτωβρίου 2014, ενώ το σύνολο της αξιολόγησης περιείχε δεδομένα από 1 Νοεμβρίου 2014 έως 30 Οκτωβρίου 2015. Ο αλγόριθμος 7-*fold cross validation* εφαρμόστηκε για την εκπαίδευση των μοντέλων και την επιλογή των βέλτιστων υπερπαραμέτρων τους. Τα μοντέλα στη συνέχεια δοκιμάστηκαν στο σύνολο αξιολόγησης.

Η μέθοδος με την **καλύτερη απόδοση** αποδείχθηκε η *SVR* με *πολυωνιμική συνάρτηση πυρήνα τρίτου βαθμού* (*polynomial cubic kernel function*), η οποία ξεπέρασε σε ακρίβεια προβλέψεων τις άλλες δύο μεθόδους.

Στα **συμπεράσματα**, οι ερευνητές καταλήγουν ότι η κατανάλωση φυσικού αερίου παρουσιάζει έντονη εποχικότητα, παρουσιάζοντας σημαντική αύξηση τους χειμερινούς μήνες σε σχέση με τους καλοκαιρινούς. Τονίζουν επίσης, την έντονα αρνητική συσχέτιση που παρουσιάζει η θερμοκρασία με την κατανάλωση. Ειδικότερα, η κατανάλωση φαίνεται να μειώνεται γραμμικά με τη θερμοκρασία καθώς αυτή αυξάνεται έως τους 20^oC, ενώ από εκεί και πέρα η αύξηση της θερμοκρασίας δεν επηρεάζει περεταίρω την κατανάλωση. Αναφέρουν ακόμα, ότι το ΑΕΠ, καθώς και η συναλλαγματική ισοτιμία δεν επηρεάζουν την κατανάλωση. Αντίθετα, η αύξηση του πληθυσμού τα περιοχής αναμένεται να επιφέρει σημαντική αύξηση στα επίπεδα κατανάλωσης. Τέλος, τονίζουν ότι η μέθοδος *SVR* μπορεί να χρησιμοποιηθεί ως εργαλείο λήψης σημαντικών αποφάσεων σε μεγάλες αγορές φυσικού αερίου, παράγοντας αξιόπιστες προβλέψεις κατανάλωσης.

2.2.12 Πρόβλεψη οικιακής κατανάλωσης ΦΑ με χρήση μοντέλων μηχανικής μάθησης και καιρικών δεδομένων ως χαρακτηριστικά

Στην εργασία αυτή [KV2019], διάφορα μοντέλα μηχανική μάθησης δοκιμάζονται στο πρόβλημα της ωριαίας, ημερήσιας και εβδομαδιαίας πρόβλεψης της οικιακής κατανάλωσης φυσικού αερίου. Ως **χαρακτηριστικά** χρησιμοποιήθηκαν η *θερμοκρασία, ταχύτητα ανέμου* και *βροχόπτωση*, καθώς και οι ημερολογιακοί παράγοντες *ώρα της ημέρας, ημέρα της εβδομάδας* και *εποχή του χρόνου*. Το σύνολο δεδομένων που χρησιμοποιήθηκε είναι διαθέσιμο και περιγράφεται αναλυτικά στο *Κεφάλαιο 3.1* της παρούσας διπλωματικής. Περιέχει δεδομένα κατανάλωσης 52 κτηρίων μιας περιοχής στην Ολλανδία για την περίοδο 2/2017 έως 12/2017.

Τα **μοντέλα** που δοκιμάστηκαν ήταν τα *Linear Regression* (*LR*), *Deep Neural Network* (*DNN*), *Long Short-Term Memory* (*LSTM*), *Gated Recurrent Unit* (*GRU*), *Convolution Neural Network* (*CNN*) καθώς και το μοντέλο *CNN+RNN+DNN,* το οποίο αποτελείται από ένα επίπεδο *CNN,* το οποίο ακολουθούν ένα επίπεδο *LSTM* και ένα επίπεδο *DNN*. Ως μέτρο αξιολόγησης των μοντέλων χρησιμοποιήθηκαν οι δείκτες *MSE, RMSE, MAPE* και *SMAPE.*

Οι επιμέρους καταναλώσεις των 52 κτηρίων αθροίστηκαν, λαμβάνοντας έτσι τη συνολική μέση κατανάλωση της περιοχής. Οι πρώτοι 6 μήνες του συνόλου δεδομένων χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων, ενώ οι υπόλοιποι 3 για την αξιολόγησή τους. Ως *συνάρτηση κόστους* (*loss function*) στα τεχνητά νευρωνικά δίκτυα που σχεδιάστηκαν χρησιμοποιήθηκε η MSE, ενώ οι αλγόριθμοι βελτιστοποίησης Adam και Nadam χρησιμοποιήθηκαν για την ελαχιστοποίησή της. Προκειμένου να μελετηθεί η απόδοση των μοντέλων σε διαφορετικούς χρονικούς ορίζοντες πρόβλεψης, οι αρχικές ωριαίες προβλέψεις που παρήχθησαν, αθροίστηκαν κατάλληλα δίνοντας τις αντίστοιχες ημερήσιες και εβδομαδιαίες προβλέψεις. Στη συνέχεια τα μοντέλα συγκρίθηκαν ως προς την απόδοσή τους στους διάφορους χρονικούς ορίζοντες πρόβλεψης.

Το μοντέλο με την **καλύτερη απόδοση** σε όρους MAPE στις ωριαίες προβλέψεις ήταν το *DNN*, ενώ η μέθοδος *LR* ήταν αυτή με την καλύτερη απόδοση στις ημερήσιες και εβδομαδιαίες προβλέψεις. Ενδιαφέρον παρουσιάζει το γεγονός ότι το *DNN* είχε σημαντικά καλύτερη απόδοση από τα υπόλοιπα πιο σύνθετα νευρωνικά δίκτυα που σχεδιάστηκαν, γεγονός που οφείλεται ακριβώς στην απλότητά του. Ειδικότερα, σε όλους τους χρονικούς ορίζοντες πρόβλεψης το μοντέλο *LSTM* είχε τη χειρότερη απόδοση. Οι ερευνητές αναφέρουν ότι όλα τα μοντέλα έτειναν να παράγουν προβλέψεις χαμηλότερες από τις πραγματικές τιμές κατανάλωσης, ενώ τα σφάλματα των προβλέψεών τους αυξάνονταν κατά τους χειμερινούς μήνες.

Στα **συμπεράσματα** οι ερευνητές επισημαίνουν πως η προσθήκη επιπλέον παραγόντων, πέρα των καιρικών, ως χαρακτηριστικά των μοντέλων θα επιφέρει περεταίρω βελτίωση των προβλέψεων. Ένας τέτοιος παράγοντας που φαίνεται να έχει επίδραση στην κατανάλωση φυσικού αερίου είναι η κατανάλωση ηλεκτρικής ενέργειας. Ιδιαίτερα σημαντικό είναι η χρήση δεδομένων κατανάλωσης για χρονικό διάστημα μεγαλύτερο του ενός έτους, προκειμένου να εντοπιστούν οι διάφορες καταναλωτικές συνήθειες που δημιουργούνται. Τέλος, αναφέρουν ότι επόμενες εργασίες πρέπει να ασχοληθούν με το πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου και να εφαρμόσουν τις μεθόδους αυτές σε μεμονωμένα κτήρια και κατοικίες.

2.2.13 Επίπτωση μετάδοσης σφαλμάτων πρόβλεψης θερμοκρασίας στην πρόβλεψη οικιακής κατανάλωσης ΦΑ με χρήση μοντέλων μηχανικής μάθησης

Στην εργασία αυτή [FM2020] μελετάται το πρόβλημα της ημερήσιας πρόβλεψης οικιακής κατανάλωσης φυσικού αερίου σε εθνικό επίπεδο, χρησιμοποιώντας ως **χαρακτηριστικά** τις προγενέστερες *καταναλώσεις*, *ημερολογιακούς παράγοντες* και τη *θερμοκρασία*. Ακόμα, οι ερευνητές χρησιμοποίησαν ως επιπλέον χαρακτηριστικό τον *παράγοντα HDD* (heating degree days). Βασικός σκοπός της εργασίας ήταν η μελέτη της επίπτωσης της μετάδοσης των σφαλμάτων της πρόβλεψης της θερμοκρασίας στην πρόβλεψη της κατανάλωσης του φυσικού αερίου. Για το λόγο αυτό χρησιμοποιήθηκαν ως χαρακτηριστικά τόσο οι πραγματικές τιμές της θερμοκρασίας, όσο και οι προβλεπόμενες από κατάλληλα μετεωρολογικά μοντέλα, καθώς όπως είναι λογικό, σε εφαρμογές πρόβλεψης κατανάλωσης στον πραγματικό κόσμο, οι μελλοντικές τιμές της θερμοκρασίας δεν είναι διαθέσιμες και κατά συνέπεια θα πρέπει να χρησιμοποιηθούν προβλέψεις αυτών. Το σύνολο δεδομένων που χρησιμοποιήθηκε αφορά την Ιταλία για την περίοδο 2007 έως 2017.

Τα **μοντέλα** πρόβλεψης που χρησιμοποιήθηκαν ήταν τα *Ridge Regression, Gaussian Process (GP), k-Nearest Neighbor (KNN), Artificial Neural Network (ANN)* και *Torus model.* Ως μέτρο της ποιότητας πρόβλεψης των μοντέλων χρησιμοποιήθηκαν οι δείκτες RMSE και MAE.

Τα δεδομένα που αφορούσαν τα έτη 2007 έως 2014 χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων, ενώ τα έτη 2015 έως 2017 χρησιμοποιήθηκαν για την αξιολόγησή τους. Ο υπολογισμός των βέλτιστων υπερπαραμέτρων των μοντέλων έγινε με χρήση του αλγορίθμου *5-fold cross validation*. Η διαδικασία αυτή έγινε δύο φορές, μία με χρήση των πραγματικών τιμών της θερμοκρασίας και μια με τις προβλεπόμενες τιμές από μοντέλα πρόβλεψης μετεωρολογικών παραγόντων. Με τον τρόπο αυτό, μελετήθηκε όπως αναφέρθηκε η επίπτωση των σφαλμάτων στην πρόβλεψη της θερμοκρασίας στην πρόβλεψη της κατανάλωσης του φυσικού αερίου.

Το μοντέλο που παρουσίασε την **καλύτερη απόδοση** ως προς το δείκτη RMSE ήταν το *ANN*, ενώ με μικρή διαφορά ακολούθησε το *GP*. Η μέθοδος *KNN* αποδείχθηκε με διαφορά η χειρότερη όλων. Συγκρίνοντας τα μοντέλα ως προς το δείκτη MAE, η εικόνα παραμένει ίδια με το *GP* αυτή τη φορά να αποδίδει λίγο καλύτερα από το *ANN*. Αναλυτικότερα, οι ερευνητές επισημαίνουν ότι το μοντέλο *GP* φαίνεται να αποδίδει καλύτερα κατά τους θερινούς μήνες, όταν η κατανάλωση παραμένει σε χαμηλά επίπεδα και παρουσιάζει έντονη εβδομαδιαία περιοδικότητα. Αντίθετα, το *ANN* αποδεικνύεται καλύτερο κατά τους χειμερινούς μήνες, όταν η κατανάλωση είναι υψηλή και η θερμοκρασία αποτελεί το βασικότερο παράγοντα που επηρεάζει την κατανάλωση. Αυτό πιθανόν να οφείλεται στην ικανότητα των νευρωνικών

δικτύων να περιγράψουν τη μη γραμμική επίδραση της θερμοκρασίας στην κατανάλωση. Όταν χρησιμοποιήθηκαν οι προβλεπόμενες τιμές της θερμοκρασίας αντί των πραγματικών, τα σφάλματα που παρήγαγαν τα μοντέλα αυξήθηκαν. Συγκεκριμένα, οι ερευνητές αναφέρουν ότι τα σφάλματα στην πρόβλεψη της τιμής της θερμοκρασίας σε όρους MSE είναι υπεύθυνα για το 18% του αντίστοιχου σφάλματος στην πρόβλεψη της κατανάλωσης του φυσικού αερίου.

Στα **συμπεράσματα**, καταλήγουν σημειώνοντας τη σημασία της μετάδοσης του σφάλματος των προβλέψεων της θερμοκρασίας στην πρόβλεψη της κατανάλωσης φυσικού αερίου, γεγονός που συνεπάγεται τη χρήση όσο το δυνατόν ακριβέστερων μοντέλων πρόβλεψης των μετεωρολογικών παραγόντων. Τέλος, αναφέρουν πως μεταγενέστερες έρευνες πρέπει να επικεντρωθούν στη χρήση πιο σύνθετων μοντέλων νευρωνικών δικτύων ως μοντέλα πρόβλεψης, όπως τα RNN και LSTM καθώς και στην περεταίρω μελέτη της διάδοσης του σφάλματος των προβλεπόμενων τιμών της θερμοκρασίας στην πρόβλεψη της κατανάλωσης.

2.3 Αξιολόγηση Τεχνολογιών Πρόβλεψης Κατανάλωσης ΦΑ

Μελετώντας τις εργασίες που αναλύθηκαν στην προηγούμενη ενότητα, καθίσταται σαφές ότι έχουν γίνει πολλές προσπάθειες σχεδιασμού και ανάπτυξης μοντέλων πρόβλεψης της κατανάλωσης φυσικού αερίου. Ανάλογα με τα διαθέσιμα δεδομένα, η **κλίμακα** των προβλέψεων κυμάνθηκε από μια μεμονωμένη κατοικία έως ολόκληρη χώρα. Οι περισσότερες σχετικές εργασίες ωστόσο, αφορούσαν την πρόβλεψη κατανάλωσης πόλης. Ο **χρονικός ορίζοντας** των προβλέψεων αφορούσε ωριαίες έως ετήσιες προβλέψεις, με την πλειοψηφία των εργασιών να επιχειρούν ημερήσια πρόβλεψη κατανάλωσης. Ακόμα, διαφορετικές **μέθοδοι προβλέψεων** προτάθηκαν και δοκιμάστηκαν στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου. Στον επόμενο πίνακα, αποτυπώνονται οι *μέθοδοι πρόβλεψης* καθώς και ο *χρονικός ορίζοντας πρόβλεψης* των εργασιών της προηγούμενης ενότητας.

Δημοσίευση	Μέθοδος πρόβλεψης						Χρονικός ορίζοντας πρόβλεψης				
	ANN	SVR	Linear/ Lasso regression	Hybrid models	Μαθηματικά/ Εμπειρικά μοντέλα	Άλλα μοντέλα μηχανικής μάθησης	Ώρα	Ημέρα	Εβδομάδα	Μήνας	Έτος
[BM1995]	✔							✔			
[HH2013]	✔										✔
[SP2014]	✔	✔			✔	✔		✔			
[Szo2015]	✔		✔				✔				
[SO2018]	✔	✔						✔			
[MP2018]	✔		✔					✔			
[HP2018]	✔		✔		✔	✔	✔				
[WL2018]	✔	✔		✔				✔			
[WL2019]	✔	✔		✔				✔			
[WD2019]	✔	✔		✔				✔			
[BE2019]	✔	✔	✔							✔	
[KV2019]	✔		✔				✔	✔	✔		
[FM2020]	✔					✔		✔			

Πίνακας 1: Μέθοδοι και χρονικός ορίζοντας πρόβλεψης σχετικών εργασιών

2.3.1 Επισκόπηση μεθόδων πρόβλεψης

Όπως γίνεται προφανές, η πλέον χρησιμοποιούμενη μέθοδος πρόβλεψης στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου ήταν τα *τεχνητά νευρωνικά δίκτυα* και υβριδικές μέθοδοι βασισμένες σε αυτά. Η ικανότητα των νευρωνικών δικτύων να μοντελοποιούν μη γραμμικές συσχετίσεις μεταξύ μεταβλητών είναι ιδιαίτερα χρήσιμη στο παρόν πρόβλημα, λόγω της έντονης εποχικότητας της κατανάλωσης φυσικού αερίου και της μη γραμμικής εξάρτησής της από πλήθος παραγόντων. Βασικό μειονέκτημα της εν λόγω μεθόδου, αποτελεί η ανάγκη μεγάλου πλήθους ιστορικών δεδομένων για την εκπαίδευση των μοντέλων. Οι εργασίες που εφάρμοσαν με επιτυχία τα τεχνητά νευρωνικά δίκτυα, διέθεταν μεγάλο πλήθος ιστορικών δεδομένων, το οποίο συχνά κάλυπτε περίοδο πολλών ετών. Αντίθετα, εργασίες με ιστορικά δεδομένα μικρής χρονικής περιόδου δεν πέτυχαν εξίσου καλά αποτελέσματα [KV2019].

Μια εξίσου δημοφιλής μέθοδος πρόβλεψης στο παρόν πρόβλημα, ήταν η μέθοδος *Support Vector Regression* (*SVR*). Η εν λόγω μέθοδος, πολλές φορές συνδυάστηκε με άλλους αλγορίθμους με σκοπό τη δημιουργία ισχυρών μοντέλων πρόβλεψης. Αλγόριθμοι όπως ο *Genetic Algorithm* (*GA*), χρησιμοποιήθηκαν για την βελτιστοποίηση των υπερπαραμέτρων της *SVR* με σκοπό την περεταίρω αύξηση της απόδοσής της, ενώ τα υβριδικά μοντέλα που σχεδιάστηκαν βασισμένα στη *SVR* ξεπέρασαν σε απόδοση ακόμα σύνθετα νευρωνικά δίκτυα περίπλοκης αρχιτεκτονικής [WL2018].

Δεν ήταν λίγες οι εργασίες εκείνες που χρησιμοποίησαν τις γραμμικές μεθόδους παλινδρόμησης *Linear Regression* (*LR*), *Ridge Regression* και *Lasso Regression*. Είναι σημαντικό να αναφερθεί ότι οι εν λόγω μέθοδοι, αν και ιδιαίτερα απλές, παρήγαγαν γενικά αξιόπιστες προβλέψεις, ενώ πολλές φορές ξεπέρασαν σε απόδοση τις προηγούμενες σύνθετες, μη γραμμικές μεθόδους των νευρωνικών δικτύων και της SVR. Ενδιαφέρον παρουσιάζει η βελτίωση της απόδοσης των προβλέψεων των νευρωνικών δικτύων, όταν αυτά συνδυάζονται με την *LR* για την εκτίμηση της κατανάλωσης «ειδικών ημερών», όπως οι επίσημες αργίες και οι διακοπές [HP2018]. Επίσης, άλλα γραμμικά μαθηματικά μοντέλα, όπως το *Autoregressive Model with Exogenous Inputs* (*ARX*) και *Stepwise Regression* (*SR*) εφαρμόστηκαν από κάποιους ερευνητές [SP2014]. Τα μοντέλα αυτά αποδείχθηκαν εξίσου αποδοτικά, ξεπερνώντας κάποιες φορές σε ακρίβεια πιο σύνθετα, μη γραμμικά μοντέλα. Σημαντικό πλεονέκτημα των μεθόδων αυτών, είναι η μεγάλη ευκολία εκπαίδευσής τους και οι ελάχιστες υπολογιστικές τους απαιτήσεις.

Στα πλαίσια της σύγκρισης μεθόδων πρόβλεψης, διάφορα *στατιστικά* και *εμπειρικά μοντέλα πρόβλεψης* σχεδιάστηκαν από ερευνητές [HP2018]. Τα μοντέλα αυτά αποτελούν εμπειρικές μαθηματικές σχέσεις που προκύπτουν από την παρατήρηση και εμπειρία των εκάστοτε ερευνητών, σχετικά με τους παράγοντες που επηρεάζουν την κατανάλωση. Αν και οι μέθοδοι αυτές απαιτούν ελάχιστους πόρους για την εκπαίδευσή τους, η απόδοσή τους δεν είναι ικανοποιητική, καθώς ακόμα και τα απλούστερα γραμμικά μοντέλα, όπως η *LR* παράγουν σημαντικά καλύτερα αποτελέσματα.

Τέλος, λίγες εργασίες χρησιμοποίησαν άλλες μεθόδους μηχανικής μάθησης στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου, όπως τα *Decision Trees* [SO2018], το *k-Nearest Neighbor* [FM2020] και άλλα. Τα μοντέλα αυτά, ωστόσο, είχαν πάντα σημαντικά μειωμένη απόδοση σε σχέση με όλα τα υπόλοιπα που χρησιμοποιήθηκαν και κατά συνέπεια δε χρειάζεται να δοκιμάζονται σε επόμενες σχετικές εργασίες.

2.3.2 Επισκόπηση μεταβλητών εισόδου μοντέλων

Ένα ακόμα ιδιαίτερα σημαντικό στοιχείο των μελετών πρόβλεψης κατανάλωσης φυσικού αερίου, είναι η επιλογή των μεταβλητών που θα χρησιμοποιηθούν ως **χαρακτηριστικά** (features) των μοντέλων πρόβλεψης. Στον *Πίνακα 2*, παρουσιάζονται συγκεντρωτικά όλοι εκείνοι οι παράγοντες που χρησιμοποιήθηκαν ως μεταβλητές εισόδου των μοντέλων στις εργασίες που αναλύθηκαν στην προηγούμενη ενότητα. Είναι σαφές ότι οι παράγοντες που επηρεάζουν περισσότερο τα επίπεδα κατανάλωσης φυσικού αερίου είναι οι *καιρικοί παράγοντες* και ειδικότερα η *θερμοκρασία περιβάλλοντος*. Πολλοί ερευνητές χρησιμοποίησαν επίσης και τον παράγοντα *HDD* (*Heating Degree Days*), ο οποίος εξαρτάται από τη θερμοκρασία ενώ παρουσιάζει γραμμική συσχέτιση με την κατανάλωση φυσικού

αερίου [FM2020]. Η *ηλιακή ακτινοβολία*, αποδείχθηκε ως ένας ακόμα καιρικός παράγοντας που βελτιώνει αισθητά την απόδοση των μοντέλων πρόβλεψης, όταν λαμβάνεται υπόψη [SP2014]. Πολλές εργασίες χρησιμοποίησαν ως μεταβλητή εισόδου την *ταχύτητα του ανέμου,* ενώ δεν ήταν λίγες εκείνες, οι οποίες χρησιμοποίησαν πλήθος άλλων καιρικών παραγόντων, όπως η *υγρασία,* το *σημείο δρόσου,* η *τιμή του υετού* και άλλους. Η ακριβής επιλογή του βέλτιστου συνδυασμού παραπάνω καιρικών παραγόντων ως χαρακτηριστικά, αποτελεί μέρος του προβλήματος της πρόβλεψης κατανάλωσης φυσικού αερίου και εξαρτάται άμεσα από το σύνολο δεδομένων που εξετάζεται. Περιοχές με κλιματικές και γεωγραφικές διαφορές, παρουσιάζουν διαφορετικές συνήθειες κατανάλωσης, με αποτέλεσμα η κατανάλωση φυσικού αερίου να μην εξαρτάται από τους ίδιους καιρικούς παράγοντες [WL2019]. Αλγόριθμοι αυτοματοποιημένης επιλογής των χαρακτηριστικών αυτών χρησιμοποιήθηκαν με επιτυχία από κάποιους ερευνητές [WD2019].

Σε εφαρμογές πρόβλεψης κατανάλωσης φυσικού αερίου στον πραγματικό κόσμο, οι πραγματικές τιμές των παραπάνω καιρικών παραγόντων δεν θα είναι διαθέσιμες, παρά μόνο οι προβλεπόμενες τιμές τους από κατάλληλα μετεωρολογικά μοντέλα. Κάποιες από τις σχετικές εργασίες, χρησιμοποίησαν τόσο τις πραγματικές, όσο και τις προβλεπόμενες αυτές τιμές των καιρικών παραγόντων ως χαρακτηριστικά των μοντέλων, προκειμένου να μελετηθεί η επίπτωση τους στην τελική απόδοση των μοντέλων. Τα σφάλματα που προκύπτουν στις προβλεπόμενες τιμές των καιρικών παραγόντων, μεταδίνονται στις προβλέψεις της κατανάλωσης φυσικού αερίου, μειώνοντας σημαντικά την ακρίβειά τους. Είναι επιτακτική, επομένως, η χρήση όσο το δυνατόν ακριβέστερων τιμών για τις προβλεπόμενες τιμές των καιρικών παραγόντων. [FM2020]. Ενδιαφέρον παρουσιάζει το γεγονός ότι κάποια μοντέλα παρουσίασαν αύξηση της απόδοσής του όταν η εκπαίδευσή τους έγινε με τις προβλεπόμενες τιμές καιρικών παραγόντων αντί των πραγματικών [HP2018]. Το μη αναμενόμενο αυτό γεγονός αναφέρθηκε για τη μέθοδο *Recurrent Neural Network* (*RNN*).

Η κατανάλωση φυσικού αερίου, όπως έχει ήδη αναφερθεί, χαρακτηρίζεται από έντονη εποχιακότητα, καθώς οι καταναλωτικές συνήθειες διαφέρουν από εποχή σε εποχή. Συνεπώς, διάφοροι *ημερολογιακοί παράγοντες* έχουν προταθεί και χρησιμοποιηθεί με επιτυχία ως χαρακτηριστικά μοντέλων πρόβλεψης. Κατά τους χειμερινούς μήνες, όταν υπάρχει ανάγκη θέρμανσης, η κατανάλωση είναι υψηλή. Αντίθετα, το καλοκαίρι τα επίπεδα κατανάλωσης παραμένουν χαμηλά. Ακόμα, περιοδικότητα παρουσιάζουν τα επίπεδα της κατά τη διάρκεια της εβδομάδας, με τα Σαββατοκύριακα να χαρακτηρίζονται από χαμηλότερα επίπεδα κατανάλωσης σε σχέση με τις καθημερινές, ενώ η υψηλότερη ζήτηση φυσικού αερίου, φαίνεται να εμφανίζεται στη μέση της εβδομάδας [MP2018]. Επίσης, διακυμάνσεις στα επίπεδα της κατανάλωσης παρατηρούνται και εντός της ημέρας. Κατά τις εργάσιμες ώρες, όταν οι άνθρωποι απουσιάζουν από τις κατοικίες τους και οι θερμοστάτες είναι κλειστοί, η

κατανάλωση είναι μηδενική. Αντίθετα, τις απογευματινές ώρες, η εικόνα αυτή αλλάζει, καθώς υπάρχει ανάγκη χρήσης του φυσικού αερίου. Τέλος, οι «ειδικές ημέρες», όπως επίσημες αργίες και διακοπές, φαίνεται να έχουν επίσης επίδραση στα επίπεδα κατανάλωσης. Ειδικότερα, η κατανάλωση φυσικού αερίου φαίνεται να είναι υψηλότερη από τη συνηθισμένη την προηγούμενη ημέρα των διακοπών, ενώ κατά τη διάρκεια τέτοιων «ειδικών ημερών» το μέγιστο της κατανάλωσης εμφανίζεται κατά τις πρωινές ώρες [Szo2015].

Η τελευταία κατηγορία παραγόντων που χρησιμοποιήθηκαν από κάποιους ερευνητές ήταν διάφοροι *κοινωνικοοικονομικοί παράγοντες*, όπως το ΑΕΠ, ο ρυθμός αύξησης του πληθυσμού και η τιμή του φυσικού αερίου. Οι παράγοντες αυτοί χρησιμοποιήθηκαν σε μελέτες μακροχρόνιας πρόβλεψης [HH2013]. Κάποιοι ερευνητές, μάλιστα, ανέφεραν ότι οι παράγοντες αυτοί δεν παρουσιάζουν σημαντική επίδραση στη βραχυπρόθεσμη κατανάλωση φυσικού αερίου και δε θα πρέπει να λαμβάνονται υπόψη [Szo2015].

Χαρακτηριστικά	**Δημοσίευση**												
	[BM1995]	[HH2013]	[SP2014]	[Szo2015]	[SO2018]	[MP2018]	[HP2018]	[WL2018]	[WL2019]	[WD2019]	[BE2019]	[KV2019]	[FM2020]
Προγενέστερες καταναλώσεις	✔		✔		✔		✔				✔		✔
Θερμοκρασία	✔	✔	✔	✔	✔		✔	✔	✔	✔	✔	✔	✔
HDD	✔					✔							✔
Ταχύτητα ανέμου	✔							✔	✔	✔		✔	
Ηλιακή ακτινοβολία			✔										
Σημείο δρόσου						✔		✔	✔	✔			
Υγρασία								✔	✔	✔			
Ορατότητα								✔	✔	✔			
Ατμοσφαιρική πίεση								✔	✔	✔			
Τιμή υετού								✔	✔	✔		✔	
Ώρα της ημέρας				✔			✔					✔	
Ημέρα της εβδομάδας	✔		✔	✔	✔	✔	✔		✔			✔	✔
Ημέρα του μήνα				✔									
Ημέρα του έτους	✔					✔							
Μήνας				✔					✔		✔		
Εποχή											✔	✔	
Έτος		✔							✔				✔
Διακοπές			✔		✔		✔						
Τιμή φυσικού αερίου									✔		✔		
Τιμή πετρελαίου											✔		
Πληθυσμός		✔									✔		
ΑΕΠ											✔		
Συναλλαγματική ισοτιμία											✔		

Πίνακας 2: Μεταβλητές εισόδου των μοντέλων σχετικών εργασιών

3 *Συγκριτική Μελέτη Συνόλων Δεδομένων*

Σε αυτό το κεφάλαιο παρουσιάζονται τα διαθέσιμα σύνολα δεδομένων κατανάλωσης φυσικού αερίου. Για κάθε σύνολο δεδομένων (dataset) υπάρχει γενική περιγραφή του, καθώς και λεπτομερής πίνακας με τα μεταδεδομένα (metatada) που περιέχει. Τέλος, ακολουθεί ο σχολιασμός των πλεονεκτημάτων και μειονεκτημάτων του συνόλου δεδομένων ως προς το πρόβλημα που διαπραγματεύεται η παρούσα διπλωματική εργασία.

3.1 KB-74-OPSCHALER

Το σύνολο δεδομένων (https://github.com/deKeijzer/KB-74-OPSCHALER) αυτό περιέχει δεδομένα κατανάλωσης φυσικού αερίου και ηλεκτρικής ενέργειας 52 κτηρίων σε μια περιοχή της Ολλανδίας για περίοδο 9 μηνών σε ωριαία βάση. Συγκεκριμένα, τα δεδομένα αυτά αφορούν την περίοδο από 02-2017 έως 12-2017. Η ακριβής τοποθεσία των κτηρίων αυτών είναι άγνωστη, καθώς το σύνολο δεδομένων υπόκειται στους νόμους σχετικά με την προστασία των προσωπικών δεδομένων. Τα δεδομένα κατανάλωσης ηλεκτρικής ενέργειας έχουν δειγματολογεί σε βάση 10 δευτερολέπτων, ενώ η κατανάλωση φυσικού αερίου σε ωριαία βάση. Επιπρόσθετα με τις καταναλώσεις, είναι διαθέσιμες μετρήσεις καιρικών παραγόντων από κοντινό μετεωρολογικό σταθμό για τη δεδομένη χρονική περίοδο. Οι μετρήσεις αυτές έγιναν διαθέσιμες από το Royal Netherlands Meteorological Institute (KNMI) station στο Rotterdam. Το κτήριο που βρίσκεται σε μεγαλύτερη απόσταση από το μετεωρολογικό σταθμό, βρίσκεται περί τα 103 km βορειοανατολικά.

Ο πίνακας που ακολουθεί περιέχει όλους τους διαθέσιμους καιρικούς παράγοντες (πεδία του συνόλου δεδομένων) και τις μετρήσεις κατανάλωσης μαζί με σύντομη ερμηνεία τους.

Παράμετρος	Μονάδα	Περιγραφή	Sample rate
DD	deg	Διεύθυνση ανέμου	15 min
DR	s	Διάρκεια βροχόπτωσης	
FX	m/s	Μέγιστη ριπή ανέμου στα 10 m	
FF	m/s	Ταχύτητα ανέμου στα 10 m	
N	okta	Νεφοκάλυψη	
P	hPa	Ατμοσφαιρική πίεση	
Q	W/m^2	Ηλιακή ακτινοβολία	
RG	mm/h	Βροχόπτωση	
SQ	min	Διάρκεια ηλιοφάνειας	
T	^oC	Θερμοκρασία στο 1.5 m	
T10	^oC	Ελάχιστη θερμοκρασία στα 10 cm	
TD	^oC	Σημείο δρόσου	
U	-	Σχετική υγρασία στο 1.5 m	
VV	m	Ορατότητα	
Timestamp	-	Χρονοσήμανση στιγμής λήψης δεδομένων από μετρητή σε τοπική ώρα	10 s
ePower	kWh	Ηλεκτρική ενέργεια που παραδίδεται στον πελάτη	
gasMeter	m^3	Τελευταία ωριαία τιμή ποσότητας (temperature converted) φυσικού αερίου που παραδόθηκε στον πελάτη	1 h
gasPower	m^3/h	Διαφορά μεταξύ τρέχουσας και προηγούμενης τιμής της gasMeter	

Πίνακας 3

Το σύνολο δεδομένων αυτό θεωρήθηκε το πλέον κατάλληλο για την προκειμένη διπλωματική εργασία. Παρέχει μετρήσεις κατανάλωσης φυσικού αερίου οικιακών καταναλωτών σε ωριαία βάση, κάτι που είναι ιδιαίτερα δύσκολο να βρεθεί. Επιπλέον, παρέχει πλήθος καιρικών παραγόντων, οι οποίοι αποτελούν το βασικότερο παράγοντα που επηρεάζει την οικιακή ενεργειακή κατανάλωση και ιδιαίτερα την κατανάλωση φυσικού αερίου.

Μειονέκτημα αποτελεί η σχετικά μικρή χρονική περίοδος που αναφέρεται το σύνολο δεδομένων. Οι μετρήσεις αφορούν ένα μόνο έτος, γεγονός που δεν επιτρέπει τη μελέτη της κατανάλωσης κατά την πάροδο των ετών, ενώ επίσης είναι αδύνατη η μελέτη της επίδρασης

κοινωνικοοικονομικών παραγόντων, όπως το ΑΕΠ ή το μέσο εισόδημα, καθώς οι παράγοντες αυτοί υπολογίζονται σε ετήσια βάση.

3.2 The Almanac of Minutely Power Dataset (AMPds)

Το σύνολο δεδομένων AMPds (https://dataverse.harvard.edu) περιέχει μετρήσεις κατανάλωσης ενέργειας μιας κατοικίας στο Βανκούβερ του Καναδά. Σε αυτό έχουν καταγραφεί μετρήσεις κατανάλωσης ηλεκτρικής ενέργειας, νερού και φυσικού αερίου σε διαστήματα ενός λεπτού. Οι συνολικά 21 μετρητές ισχύος, 2 μετρητές νερού και 2 μετρητές φυσικού αερίου έχουν καταγράψει πάνω από ένα εκατομμύρια μετρήσεις σε διάστημα δύο ετών (4/2012 έως 3/2014). Επιπλέον, υπάρχουν ωριαία δεδομένα διαφόρων καιρικών παραγόντων όπως αυτά κατεγράφησαν από τον μετεωρολογικό σταθμό YVR του Καναδά. Τέλος, έχουν προστεθεί στοιχεία σχετικά με τις χρεώσεις των διάφορων παρόχων ενέργειας.

Στον παρακάτω πίνακα φαίνονται όλοι οι τύποι των αρχείων με μια σύντομη περιγραφή τους.

Αρχείο	**Περιγραφή**
Climate_HistoricalNormals.csv	Σύνοψη των μέσων κλιματικών συνθηκών που παρατηρήθηκαν κατά τα έτη 1981 έως 2010
Climate_HourlyWeather.csv	Ωριαίες μετρήσεις καιρικών παραγόντων όπως μετρήθηκαν από το μετεωρολογικό σταθμό YVR
Electricity_???.csv	Αρχείο κατανάλωσης ηλεκτρικής ενέργειας για κάθε ένα μετρητή και υπομετρητή της κατοικίας
Electricity_?.csv	Ένα αρχείο για κάθε στιγμιαία μέτρηση κάθε μετρητή και υπομετρητή ηλεκτρικής ενέργειας
Electricity_Billing.csv	Τιμές δεδομένων που συλλέγονται από κάθε δήλωση λογαριασμού ηλεκτρικής ενέργειας
Electricity_Monthly.csv	Μηνιαία κατανάλωση ενέργειας
Electricity_Statements.pdf	Αντίγραφα κάθε δήλωσης λογαριασμού ηλεκτρικής ενέργειας κατά την περίοδο συλλογής δεδομένων
Manual_*.csv	Εγχειρίδια για τις συσκευές που χρησιμοποιήθηκαν
Metering_*.csv	Τεχνικές προδιαγραφές και περιγραφή του εξοπλισμού μετρήσεων που χρησιμοποιήθηκε
NaturalGas_Billing.csv	Τιμές δεδομένων που συλλέγονται από κάθε δήλωση λογαριασμού φυσικού αερίου
NaturalGas_FRG.csv	Μετρήσεις κατανάλωσης φυσικού αερίου από υπομετρητή φούρνου
NaturalGas_HeatValues.csv	Ημερήσια κατανάλωση φυσικού αερίου για θέρμανση
NaturalGas_Monthly.csv	Μηνιαία κατανάλωση φυσικού αερίου
NaturalGas_Statements.pdf	Αντίγραφα κάθε δήλωσης λογαριασμού φυσικού αερίου κατά την περίοδο συλλογής δεδομένων

NaturalGas_WHG.csv	Μετρήσεις κατανάλωσης φυσικού αερίου για ολόκληρη την κατοικία
Water_Billing.csv	Τιμές δεδομένων που συλλέγονται από κάθε δήλωση λογαριασμού νερού
Water_DWW.csv	Μετρήσεις κατανάλωσης νερού από το πλυντήριο πιάτων
Water_HTW.csv	Μετρήσεις κατανάλωσης νερού από υπομετρητή μονάδας ζεστού νερού
Water_QualityReport_2012.pdf	Η ετήσια έκθεση ποιότητας νερού της πόλης για το 2012
Water_QualityReport_2013.pdf	Η ετήσια έκθεση ποιότητας νερού της πόλης για το 2013
Water_QualityReport_2014.pdf	Η ετήσια έκθεση ποιότητας νερού της πόλης για το 2014
Water_WHW.csv	Μετρήσεις κατανάλωσης νερού για ολόκληρη την κατοικία
Water_ZonesMap.pdf	Χάρτης της πόλης που απεικονίζει τις διάφορες ζώνες νερού. Η υπό εξέταση κατοικία βρίσκεται στη ζώνη 585

Πίνακας 4

Αναλυτικότερα, κάθε αρχείο κατανάλωσης φυσικού αερίου (NaturalGas_**.csv) περιλαμβάνει τα εξής πεδία:

Πεδίο	**Περιγραφή**	**Μονάδα**
0	Χρονοσήμανση στιγμής λήψης δεδομένων	s
1	Μέτρηση κατανάλωσης	dm^3
2	Μέσος ρυθμός κατανάλωσης	dm^3/h
3	Στιγμιαίος ρυθμός κατανάλωσης	dm^3/h

Πίνακας 5

Κάθε αρχείο κατανάλωσης ηλεκτρικής ενέργειας (Electricity_???.csv) περιλαμβάνει τα εξής πεδία:

Πεδίο	**Περιγραφή**	**Μονάδα**
0	Χρονοσήμανση στιγμής λήψης δεδομένων	s
1	Τάση (V)	V
2	Ρεύμα (I)	A
3	Συχνότητα (f)	Hz
4	Συντελεστής ισχύος μετατόπισης (DPF)	ratio
5	Πραγματικός συντελεστής ισχύος (APF)	ratio
6	Πραγματική ισχύς (P)	W
7	Πραγματική ενέργεια (Pt)	Wh
8	Άεργος ισχύς (Q)	VAR
9	Άεργος ενέργεια (Qt)	VARh
10	Φαινόμενη ισχύς (S)	VA
11	Φαινόμενη ενέργεια (St)	VAh

Πίνακας 6

Κάθε αρχείο κατανάλωσης νερού (Water_***.csv) περιλαμβάνει τα εξής πεδία:

Πεδίο	**Περιγραφή**	**Μονάδα**
0	Χρονοσήμανση στιγμής λήψης δεδομένων	s
1	Μέτρηση κατανάλωσης	L
2	Μέσος ρυθμός κατανάλωσης	L/h
3	Στιγμιαίος ρυθμός κατανάλωσης	L/h

Πίνακας 7

Σημειώνεται, ότι για τη χρονοσήμανση (timestamp), χρησιμοποιείται η Unix Timestamp.

Το εν λόγω σύνολο δεδομένων, αποτελεί το μοναδικό ελεύθερα διαθέσιμο που διαθέτει μετρήσεις κατανάλωσης, τόσο ηλεκτρικής ενέργειας, όσο νερού και φυσικού αερίου σε βάση λεπτού για μεγάλη χρονική περίοδο και μάλιστα σε επίπεδο νοικοκυριού. Πλεονέκτημα αποτελεί επίσης το γεγονός ότι μαζί με τις μετρήσεις κατανάλωσης παρέχονται και δεδομένα σχετικά με τις χρεώσεις λογαριασμών καθώς και διάφοροι μετεωρολογικοί παράγοντες που ενδεχομένως επηρεάζουν την κατανάλωση.

Στα αρνητικά εντάσσεται το γεγονός ότι τα δεδομένα που καταγράφονται αφορούν τις καταναλώσεις μιας μόνο κατοικίας, γεγονός που το καθιστά ακατάλληλο για μελέτες πρόβλεψης κατανάλωσης ενέργειας. Ωστόσο, αποτελεί αξιόλογο σύνολο δεδομένων για μελέτη σχετική με τον επιμερισμό της καταναλισκόμενης ενέργειας μεταξύ των συσκευών της κατοικίας.

3.3 *U.S Energy Information Administration*

Η μηνιαία κατανάλωση φυσικού αερίου των Η.Π.Α είναι διαθέσιμη από το site του σχετικού κυβερνητικού οργανισμού (https://www.eia.gov/naturalgas/data.php). Τα ιστορικά δεδομένα της συνολικής μηνιαίας κατανάλωσης οικιακού τομέα των Η.Π.Α είναι διαθέσιμα από το 1973, ενώ οι επιμέρους μηνιαίες καταναλώσεις κάθε πολιτείας ξεχωριστά γίνονται διαθέσιμες από το έτος 1989. Επιπρόσθετα, υπάρχουν αντίστοιχα δεδομένα σχετικά με την κατανάλωση φυσικού αερίου του βιομηχανικού και εμπορικού τομέα.

Η μορφή που παρουσιάζονται οι καταναλώσεις στο σύνολο δεδομένων αυτό παρουσιάζεται στον παρακάτω πίνακα. Η ίδια μορφή ισχύει τόσο στη συνολική κατανάλωση των Η.Π.Α, όσο και στις επιμέρους καταναλώσεις κάθε πολιτείας ξεχωριστά.

Μήνας **Έτος**	Οι μήνες σε αύξουσα σειρά
Αύξουσα σειρά των ετών που υπάρχουν καταγραφές	Καταναλώσεις φυσικού αερίου μετρημένες σε Million Cubic Feet

Πίνακας 8

Όπως τα περισσότερα διαθέσιμα σύνολα δεδομένων κατανάλωσης φυσικού αερίου, έτσι και το παρόν διαθέτει τα δεδομένα κατανάλωσης σε μεγάλη κλίμακα, καθώς και με μικρό ρυθμό δειγματοληψίας. Για την πρόβλεψη κατανάλωσης φυσικού αερίου οικιακών καταναλωτών είναι απαραίτητη η διάθεση καταναλώσεων σε όσο χαμηλότερο επίπεδο είναι δυνατό, ιδανικά σε επίπεδο νοικοκυριού. Παράλληλα, ο ρυθμός δειγματοληψίας πρέπει να είναι σε ωριαία ή ημερήσια βάση, ώστε να είναι δυνατό να απεικονιστούν οι διακυμάνσεις που παρουσιάζει κατανάλωση εντός της εβδομάδας ή ακόμα και της ημέρας. Από τα παραπάνω, γίνεται σαφές ότι ούτε το συγκεκριμένο σύνολο δεδομένων είναι κατάλληλο για το πρόβλημα που διαπραγματεύεται η εν λόγω εργασία.

3.4 *Electricity and Gas Consumption for LBNL Building 74*

Το σύνολο δεδομένων αυτό (https://openei.org/datasets/dataset/lbnl-building-74) περιέχει μετρήσεις κατανάλωσης ηλεκτρικής ενέργειας και φυσικού αερίου του κτηρίου 74 στο Εθνικό Εργαστήριο Λώρενς στο Μπέρκλεϋ (*Lawrence Berkeley National Laboratory*). Πρόκειται για ένα ομοσπονδιακό ερευνητικό κέντρο επιστήμης και τεχνολογίας των Η.Π.Α. Οι μετρήσεις κατανάλωσης ηλεκτρικής ενέργειας έχουν ληφθεί σε βάση 15 λεπτών, ενώ οι μετρήσεις κατανάλωσης φυσικού αερίου σε βάση 5 λεπτών. Οι μετρήσεις αυτές αφορούν το χρονικό διάστημα από 1/1/2014 έως 30/6/2015.

Η μορφή που δίνονται οι μετρήσεις τις ηλεκτρικής ενέργειας είναι η εξής:

Πεδίο	Περιγραφή	Μονάδα
0	Χρονοσήμανση στιγμής λήψης δεδομένων	s
1	Μέτρηση κατανάλωσης ηλεκτρικής ενέργειας	kWh

Πίνακας 9

Αντίστοιχα, η μορφή που δίνονται οι μετρήσεις του φυσικού αερίου είναι:

Πεδίο	Περιγραφή	Μονάδα
0	Χρονοσήμανση στιγμής λήψης δεδομένων	DD/MM/YY hh:mm
1	Ρυθμός κατανάλωσης φυσικού αερίου	therms/h
2	Μέτρηση κατανάλωσης φυσικού αερίου	therms

Πίνακας 10

Σημειώνεται ότι η μονάδα *therm* είναι μονάδα μέτρησης θερμικής ενέργειας και ισούται με 100000 Btu (*British thermal units*). Η μονάδα αυτή χρησιμοποιείται ευρέως για τη διατίμηση φυσικού αερίου.

Το σύνολο δεδομένων αυτό παρέχει μετρήσεις κατανάλωσης ενέργειας για ικανοποιητικό χρονικό διάστημα με πολύ καλό ρυθμό δειγματοληψίας. Όπως αναφέρθηκε ήδη, οι μετρήσεις αυτές αναφέρονται σε κάποιο κτήριο ενός εργαστηρίου. Είναι προφανές, ότι ένα τέτοιο κτήριο θα παρουσιάζει πολύ διαφορετικές συνήθειες κατανάλωσης ενέργειας από οικιακούς καταναλωτές. Το γεγονός αυτό, σε συνδυασμό με το ότι, όπως και σε προηγούμενο σύνολο δεδομένων, οι μετρήσεις είναι διαθέσιμες για ένα μεμονωμένο κτήριο καθιστά το σύνολο δεδομένων αυτό ακατάλληλο για μελέτες πρόβλεψης κατανάλωσης φυσικού αερίου οικιακών καταναλωτών.

3.5 City of Mesa Natural Gas Consumption

Το σύνολο δεδομένων (https://data.mesaaz.gov/Energy-Resources/Natural-Gas-Consumption/t9u6-caye) αυτό περιέχει μηνιαίες καταναλώσεις φυσικού αερίου της πόλης Μέσα (Mesa) στην κομητεία Μαρικόπα, στην πολιτεία Αριζόνα των Η.Π.Α, για την περίοδο από τον Ιανουάριου του 2015 έως και τον Αύγουστο του 2020. Η μονάδα μέτρησης της κατανάλωσης φυσικού αερίου στην οποία δίνονται οι μετρήσεις είναι τα *therms*. Πέρα από τις καταναλώσεις, το σύνολο δεδομένων παρέχει πληροφορίες σχετικά με το είδος των καταναλωτών. Συγκεκριμένα, για κάθε μήνα καταγράφεται το ποσό κατανάλωσης για κάθε κατηγορία καταναλωτών (οικιακοί καταναλωτές, εμπορικοί καταναλωτές, άλλες δημόσιες αρχές και υπηρεσίες) καθώς και το πλήθος των καταναλωτών που ανήκουν στην εν λόγω

κατηγορία. Η μορφή που έχει το σύνολο δεδομένων μαζί με τα μεταδεδομένα που περιέχει φαίνεται στον παρακάτω πίνακα:

Πεδίο	Περιγραφή
Month	Ο μήνας χρέωσης και καταγραφής της κατανάλωσης
Year	Το έτος που ανήκει ο παραπάνω μήνας
Consumer Type	Η κατηγορία καταναλωτών (*οικιακοί, εμπορικοί, άλλες δημόσιες αρχές και υπηρεσίες)*
Amount	Η ποσότητα κατανάλωσης φυσικού αερίου μετρημένη σε *therms*
Customer Count	Ο αριθμός καταναλωτών που ανήκει στην εν λόγω κατηγορία

Πίνακας 11

Όπως έχει ήδη αναφερθεί, για την πρόβλεψη κατανάλωσης φυσικού αερίου οικιακών καταναλωτών είναι απαραίτητη η διάθεση καταναλώσεων σε όσο χαμηλότερο επίπεδο είναι δυνατό, ιδανικά σε επίπεδο νοικοκυριού και μάλιστα με όσο το δυνατό μεγαλύτερο ρυθμό δειγματοληψίας. Αυτό είναι σημαντικό καθώς η οικιακή κατανάλωση παρουσιάζει σημαντικές διακυμάνσεις εντός της ημέρας. Από τα παραπάνω, γίνεται σαφές ότι το συγκεκριμένο σύνολο δεδομένων δεν είναι κατάλληλο για το πρόβλημα που διαπραγματεύεται η εν λόγω εργασία, καθώς δεν ικανοποιεί τις παραπάνω προϋποθέσεις, ενώ ακόμα οι καταγραφές που περιέχει δεν είναι αρκετές καθώς αφορούν μικρό χρονικό διάστημα.

3.6 Chicago's Energy Usage 2010

Στο σύνολο δεδομένων αυτό (https://data.cityofchicago.org/Environment-Sustainable-Development/Energy-Usage-2010/8yq3-m6wp) υπάρχουν καταγραφές σχετικά με την κατανάλωση ηλεκτρικής ενέργειας και φυσικού αερίου για τον οικιακό, εμπορικό και βιομηχανικό τομέα της πόλης του Σικάγο των Η.Π.Α για το έτος 2010. Τα δεδομένα ηλεκτρικής ενέργειας που έχουν καταγραφεί αποτελούν το 68% της συνολικής χρήσης στην πόλη, ενώ τα δεδομένα φυσικού αερίου αποτελούν το 81% της συνολικής κατανάλωσης φυσικού αερίου στο Σικάγο για το 2010. Η μορφή που έχει το σύνολο δεδομένων μαζί με τα μεταδεδομένα που περιέχει φαίνεται στον παρακάτω πίνακα:

Πεδίο	Περιγραφή
COMMUNITY AREA NAME	Όνομα κοινοτικής περιοχής
CENSUS BLOCK	Αριθμός γεωκωδικοποίησης που λαμβάνεται στους αλγόριθμους αντιστοίχισης διευθύνσεων
BUILDING TYPE	Ο τύπος του κτηρίου (*οικιακό, εμπορικό, βιομηχανικό*)

BUILDING_SUBTYPE	Υποκατηγορία κτηρίου (*μονοκατοικία, πολυκατοικία, δημοτικά*)
KWH “month” 2010	Κατανάλωση ηλεκτρικής ενέργειας σε κιλοβατώρες (kWh) για τον αναφερόμενο μήνα
TOTAL KWH	Συνολική κατανάλωση ηλεκτρικής ενέργειας (kWh) για το 2010 συνολικά
ELECTRICITY ACCOUNTS	Αριθμός λογαριασμών στον οποίο αναφέρονται οι καταναλώσεις ηλεκτρικής ενέργειας. Ένας λογαριασμός δεν αντιστοιχεί σε ένα κτίριο
ZERO KWH ACCOUNTS	Αριθμός λογαριασμών με μηδενική κατανάλωση σε kWh για το 2010
THERM “month” 2010	Κατανάλωση φυσικού αερίου σε *therms* για τον αναφερόμενο μήνα
TOTAL THERMS	Συνολική κατανάλωση φυσικού αερίου (*therms*) για το 2010 συνολικά
GAS ACCOUNTS	Αριθμός λογαριασμών στον οποίο αναφέρονται οι καταναλώσεις φυσικού αερίου. Ένας λογαριασμός δεν αντιστοιχεί σε ένα κτίριο
KWH TOTAL SQFT	Συνολικά τετραγωνικά πόδια (ft^2) που σχετίζονται με τη χρήση ηλεκτρικής ενέργειας το 2010
THERMS TOTAL SQFT	Συνολικά τετραγωνικά πόδια (ft^2) που σχετίζονται με τη χρήση φυσικού αερίου το 2010
KWH MEAN 2010	Μέση κατανάλωση ηλεκτρικής ενέργειας για το 2010
KWH STANDARD DEVIATION 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
KWH MINIMUM 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
KWH “1^{st}, 2^{nd},3^{rd}” QUARTILE 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
KWH MAXIMUM 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
KWH SQFT MEAN 2010	Μέση κατανάλωση ηλεκτρικής ενέργειας ανά τετραγωνικά πόδια (ft^2)
KWH SQFT STANDARD DEVIATION 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
KWH SQFT MINIMUM 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
KWH SQFT “1^{st}, 2^{nd},3^{rd}” QUARTILE 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
KWH SQFT MAXIMUM 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
THERM MEAN 2010	Μέση κατανάλωση φυσικού αερίου για το 2010
THERM STANDARD DEVIATION 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
THERM MINIMUM 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
THERM “1^{st}, 2^{nd},3^{rd}” QUARTILE 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
THERM MAXIMUM 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
THERMS SQFT MEAN 2010	Μέση κατανάλωση φυσικού αερίου ανά τετραγωνικά πόδια (ft^2)

THERMS SQFT STANDARD DEVIATION 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
THERMS SQFT MINIMUM 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
THERMS SQFT "1st, 2nd,3rd" QUARTILE 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
THERMS SQFT MAXIMUM 2010	(Δεν υπάρχει διαθέσιμη περιγραφή)
TOTAL POPULATION	Συνολικός πληθυσμός σύμφωνα με την απογραφή του 2010 (αναφέρεται σε κάθε CENSUS BLOCK)
TOTAL UNITS	Συνολικός αριθμός οικιστικών μονάδων σύμφωνα με την απογραφή του 2010
AVERAGE BUILDING AGE	Μέση ηλικία κτηρίων
AVERAGE HOUSESIZE	Μέσο μέγεθος νοικοκυριών
OCCUPIED UNITS	Αριθμός οικιστικών μονάδων που κατοικούνται
OCCUPIED UNITS PERCENTAGE	(Δεν υπάρχει διαθέσιμη περιγραφή)
RENTER-OCCUPIED HOUSING UNITS	Αριθμός οικιστικών μονάδων που ενοικιάζονται
RENTER-OCCUPIED HOUSING PERCENTAGE	Ποσοστό οικιστικών μονάδων που ενοικιάζονται
OCCUPIED HOUSING UNITS	Αριθμός οικιστικών μονάδων που κατοικούνται

Πίνακας 12

Το σύνολο δεδομένων αυτό είναι επίσης ακατάλληλο για την εν λόγω διπλωματική εργασία για τους λόγους που έχουν ήδη αναφερθεί. Ωστόσο, το γεγονός ότι παρέχει διάφορες πληροφορίες σχετικές με τα κτήρια για τα οποία δίνει τις καταναλώσεις ενέργειας, ίσως το καθιστά κατάλληλο για κάποια μελέτη σχετική με την ενεργειακή κατανάλωση και την κατανομή αυτής στα διάφορα είδη καταναλωτών. Επιπλέον μειονέκτημα στο σύνολο δεδομένων αυτό αποτελεί το γεγονός ότι έχει αρκετές κενές εγγραφές.

3.7 City of Gainesville Natural Gas Consumption

Το σύνολο δεδομένων αυτό (https://data.cityofgainesville.org/Utilities/GRU-Customer-Natural-Gas-Consumption/gxhb-disj) αφορά την πόλη Gainesville στην περιοχή Florida των Η.Π.Α. Περιέχει δεδομένα μηνιαίας κατανάλωσης φυσικού αερίου σε *therms* από τον Ιανουάριου του 2012 έως τον Αύγουστο του 2020. Η μορφή που έχει το σύνολο δεδομένων μαζί με τα μεταδεδομένα που περιέχει φαίνεται στον παρακάτω πίνακα:

Πεδίο	**Περιγραφή**
Service Address	Διεύθυνση καταναλωτών φυσικού αερίου
Service City	Πόλη καταναλωτών φυσικού αερίου
Month	Μήνας που εκδόθηκε ο λογαριασμός
Year	Έτος που εκδόθηκε ο λογαριασμός
Date	Ακριβής ημερομηνία έκδοσης λογαριασμού
Therm Consumption	Κατανάλωση φυσικού αερίου σε *therms*
Latitute	Γεωγραφικό πλάτος που βρίσκεται ο καταναλωτής
Longtitude	Γεωγραφικό μήκος που βρίσκεται ο καταναλωτής
Location	Συντεταγμένες τοποθεσίας καταναλωτή

Πίνακας 13

Το συγκεκριμένο σύνολο δεδομένων, αν και εκ πρώτης όψης φαίνεται ενδιαφέρον, παρουσιάζει σημαντικά μειονεκτήματα για την παρούσα εργασία. Το dataset περιέχει πάνω από 1.5 εκατομμύριο εγγραφές, οι οποίες αναφέρονται σε πολύ μεγάλο αριθμό κτηρίων, Service Addresses, όπως ονομάζεται το αντίστοιχο πεδίο. Τα κτήρια αυτά ωστόσο, δεν είναι μόνο οικιακοί, αλλά όλων των ειδών καταναλωτές, όπως επίσης εμπορικοί, βιομηχανικοί και δημόσιοι φορείς, ενώ παράλληλα δεν υπάρχει κάποιος τρόπος να ξεχωριστούν οι παραπάνω κατηγορίες καταναλωτών. Αποτέλεσμα αυτού είναι η κατανάλωση φυσικού αερίου να παρουσιάζει πολύ μεγάλο εύρος τιμών, ενώ υπάρχει περίπτωση να υπάρχουν αρκετές λανθασμένες εγγραφές (outliers). Ακόμα, πρόβλημα αποτελεί το γεγονός ότι, αν και υπάρχουν πολλά διαφορετικά κτήρια για τα οποία υπάρχουν μετρήσεις καταναλώσεων, σε ελάχιστα υπάρχουν μετρήσεις για όλους τους μήνες που περιλαμβάνονται. Συγκεκριμένα, το κτήριο με τις περισσότερες εγγραφές περιλαμβάνει καταγραφές για περί τις 120 μηνιαίες καταναλώσεις, ενώ υπάρχουν πολλά άλλα για τα οποία έχει γίνει καταγραφή κατανάλωσης μόνο για ένα μήνα. Τα παραπάνω καθιστούν και αυτό το σύνολο δεδομένων ακατάλληλο για την παρούσα εργασία.

3.8 Συγκεντρωτικός πίνακας παρουσίασης συνόλων δεδομένων κατανάλωσης φυσικού αερίου

KB-74-OPSCHALER		AMPds		City of Gainesville		Chicago Energy Usage 2010		LBNL Building 74		City of Mesa		U.S EIA	
Πεδίο	**Περιγραφή**	**Πεδίο**	**Περιγραφή**	**Πεδίο**	**Περιγραφή**	**Πεδίο**	**Περιγραφή**	**Πεδίο**	**Περιγραφή**	**Πεδίο**	**Περιγραφή**	**Πεδίο**	**Περιγραφή**
DD	Διεύθυνση ανέμου	unix_ts	Χρονοσήμανση στιγμής λήψης δεδομένων	Service Addre	Διεύθυνση καταναλωτή	CENSUS BLOCK	Αριθμός γεωκωδικοποίησης	Timestamp	Χρονοσήμανση στιγμής λήψης δεδομένων	Year	Έτος μέτρησης	Year	Έτος μέτρησης
DR	Διάρκεια βροχόπτωσης	counter	Μέτρηση κατανάλωσ φυσικού αερίου	Service City	Πόλη καταναλωτή	BUILDING TYPE	Τύπος κτηρίου	kWh Total Electricity	Μέτρηση κατανάλωσης ηλεκτρικής ενέργειας	Month	Μήνας μέτρησης	Month	Μήνας μέτρησης
FX	Μέγιστη ριπή ανέμου	avg_rate	Μέσος ρυθμός κατανάλωσης	Month	Μήνας μέτρησης	BUILDING_SUBTYPE	Υποκατηγορία κτηρίου	THERMS/hr	Ρυθμός κατανάλωσης φυσικού αερίου	Consumption	Μέτρηση κατανάλωσης	Consumption	Μέτρηση κατανάλωσης
FF	Ταχύτητα ανέμου	inst_rate	Στιγμιαίος ρυθμός κατανάλωσης	Year	Έτος μέτρησης	KWH “month” 2010	Μηνιαία κατανάλωση ηλεκτρικής ενέργειας	THERMS	Μέτρηση κατανάλωσης φυσικού αερίου	Amount	Ποσότητα κατανάλωσης		
N	Νεφοκάλυψη	Temp	Θερμοκρασία	Date	Ημερομηνία μέτρησης	TOTAL KWH	Ετήσια κατανάλωση ηλεκτρικής ενέργειας						
P	Ατμοσφαιρική πίεση	Dew Point	Σημείο δρόσου	Consumption	Μέτρηση κατανάλωσης	THERM “month” 2010	Μηνιαία κατανάλωση φυσικού αερίου						
Q	Ηλιακή ακτινοβολία	Rel Hum	Σχετική υγρασία	Latitude	Γεωγραφικό πλάτος	TOTAL THERMS	Ετήσια κατανάλωση φυσικού αερίου						
RG	Βροχόπτωση	Wind Dir	Κατεύθυνση ανέμου	Longtitude	Γεωγραφικό μήκος								
SQ	Διάρκεια ηλιοφάνειας	Wind Spd	Ταχύτητα ανέμου	Location	Συντεταγμένες τοποθεσίας								
T	Θερμοκρασία	Visibility	Ορατότητα										
T10	Ελάχιστη θερμοκρασία												
TD	Σημείο δρόσου												
U	Σχετική υγρασία												
VV	Ορατότητα												
Timestamp	Χρονοσήμανση στιγμής λήψης δεδομένων												
ePower	Ηλεκτρική ενέργεια που παραδίδεται στον πελάτη												
gasMeter	Τελευταία ωριαία τιμή ποσότητας φυσικού αερίου που παραδόθηκε στον πελάτη												
gasPower	Διαφορά μεταξύ τρέχουσας και προηγούμενης τιμής της gasMeter												

Πίνακας 14

3.9 Συμπεράσματα

Από την έρευνα για διαθέσιμα σύνολα δεδομένων (datasets) κατανάλωσης φυσικού αερίου και τη μελέτη της σχετικής βιβλιογραφίας, γίνεται σαφές ότι υπάρχει μια δυσκολία στη χρήση τέτοιων συνόλων δεδομένων, κυρίως λόγω της υπάρχουσας νομοθεσίας σχετικά με την προστασία προσωπικών δεδομένων, η οποία δεν επιτρέπει στους παρόχους φυσικού αερίου να δημοσιοποιούν τις καταναλώσεις των πελατών τους.

Αρκετά από τα δημοσίως διαθέσιμα σύνολα δεδομένων που υπάρχουν περιέχουν μικρό αριθμό καταγραφών, οι οποίες αφορούν συνήθως μικρό χρονικό διάστημα, ενώ ο ρυθμός δειγματοληψίας είναι επίσης μικρός, συνήθως σε μηνιαία ή ετήσια βάση.

Για την παρούσα εργασία, το σύνολο δεδομένων που κρίθηκε κατάλληλο είναι αυτό που παρουσιάστηκε στην *Υποενότητα 3.1* (KB-74-OPSCHALER). Συνοπτικά, στο σύνολο αυτό δίνονται μετρήσεις ωριαίας κατανάλωσης φυσικού αερίου 52 κτηρίων σε μια περιοχή της Ολλανδίας, για περίοδο 9 μηνών του έτους 2017. Επιπρόσθετα, γίνονται διαθέσιμες μετρήσεις ηλεκτρικής ενέργειας, καθώς και πλήθους καιρικών παραγόντων. Το σύνολο δεδομένων αυτό επιλέχθηκε λόγω του ρυθμού δειγματοληψίας των δεδομένων που περιέχει, ο οποίος ήταν ο μικρότερος όλων των διαθέσιμων συνόλων, ενώ παράλληλα έκανε διαθέσιμες μετρήσεις σημαντικού πλήθους παραγόντων, οι οποίοι θα χρησιμοποιηθούν ως υποψήφια χαρακτηριστικά (features) των μοντέλων πρόβλεψης που θα σχεδιαστούν.

4 Μεθοδολογία Πρόβλεψης Κατανάλωσης ΦΑ

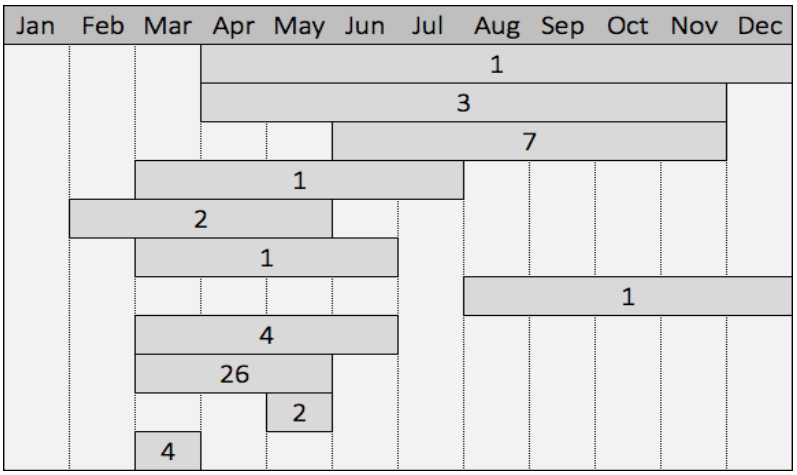
4.1 Σύνοψη

Στις ενότητες που ακολουθούν, αναλύεται η διαδικασία της προεπεξεργασίας των δεδομένων (preprocessing) και της επιλογής των χαρακτηριστικών των μοντέλων (feature engineering). Γίνεται η παρουσίαση του συνόλου δεδομένων, ενώ επίσης εξάγονται χρήσιμα συμπεράσματα, σχετικά με τα χαρακτηριστικά της κατανάλωσης φυσικού αερίου, όπως διάφορες τάσεις και καταναλωτικές συνήθειες που παρουσιάζονται μέσα στο έτος. Η διαδικασία της επιλογής των χαρακτηριστικών, συνίσταται στις μεθοδολογίες και παρατηρήσεις που οδήγησαν στην επιλογή των παραγόντων εκείνων, που θα χρησιμοποιηθούν τελικά ως μεταβλητές εισόδου των μοντέλων για την πρόβλεψη της κατανάλωσης του φυσικού αερίου. Στη συνέχεια, εξηγείται η μεθοδολογία που ακολουθήθηκε για την εκπαίδευση των μοντέλων και την ρύθμιση των υπερπαραμέτρων τους. Τέλος, παρουσιάζονται οι δείκτες που χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσης των μοντέλων πρόβλεψης, καθώς και τα μοντέλα πρόβλεψης που δοκιμάστηκαν.

4.2 Προεπεξεργασία Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε, περιείχε μετρήσεις ωριαίας κατανάλωσης φυσικού αερίου της περιόδου 02/2017 έως 12/2017. Οι μετρήσεις έγιναν από έξυπνους μετρητές που ήταν εγκατεστημένοι σε 52 κτήρια μιας περιοχής στην Ολλανδία. Η περίοδος καταγραφής δεν ήταν ίδια για όλα τα κτήρια, αλλά κυμαίνεται από 1 έως 9 μήνες. Στην παρακάτω εικόνα φαίνεται η κατανομή της περιόδου καταγραφής της κατανάλωσης των

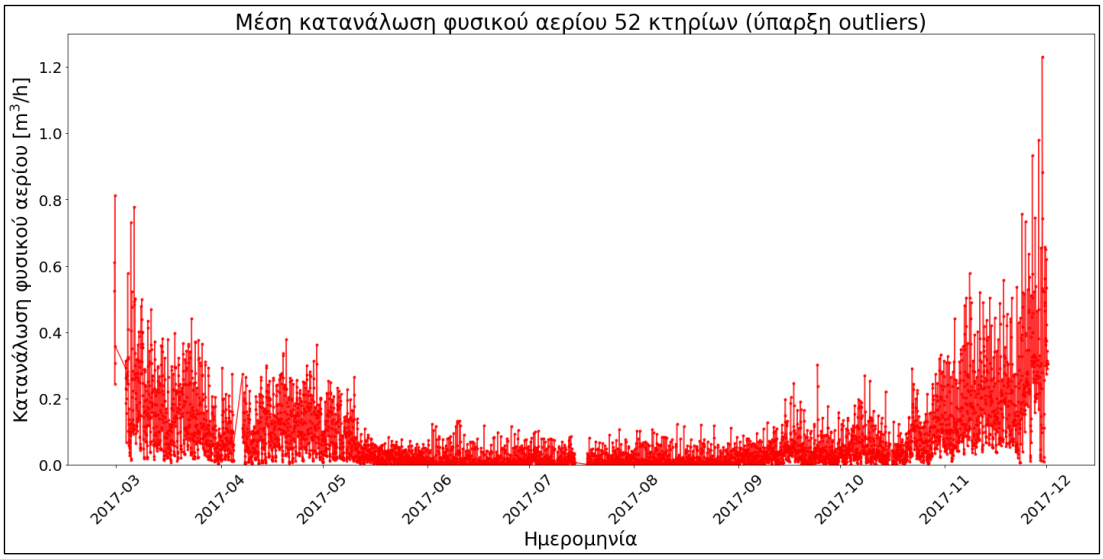
κτηρίων. Ο αριθμός στη κάθε γκρι περιοχή είναι το πλήθος των κτηρίων για τα οποία είναι διαθέσιμες οι μετρήσεις κατανάλωσης κατά τη διάρκεια αυτής της περιόδου.



Εικόνα 4: Κατανομή περιόδου καταγραφής δεδομένων για τα 52 κτήρια

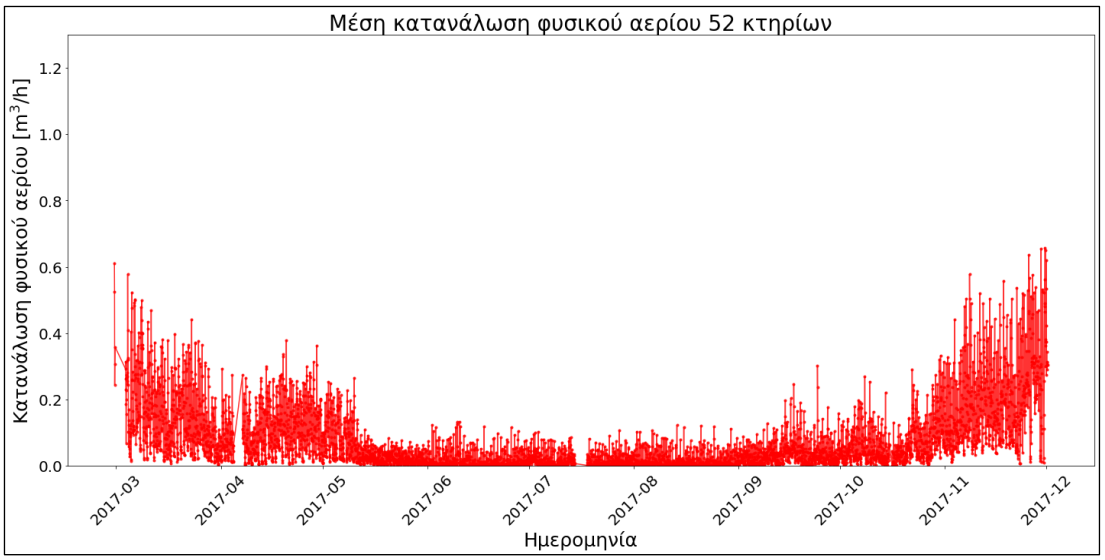
Οι μετρήσεις κατανάλωσης φυσικού αερίου ήταν σε ωραία βάση. Αντίθετα, οι μετρήσεις των διαφόρων καιρικών παραγόντων, καθώς και της κατανάλωσης ηλεκτρικής ενέργειας, όπως γίνονται διαθέσιμες από το σύνολο δεδομένων, ήταν σε διαφορετική χρονική βάση. Επομένως, έγινε ο υπολογισμός της μέσης ωριαίας τιμής καθενός από τους παράγοντες αυτούς, προκειμένου να υπάρχει κοινός ρυθμός δειγματοληψίας για όλους τους παράγοντες του συνόλου δεδομένων.

Υπολογίστηκε η μέση τιμή της ωριαίας κατανάλωσης, μαζί με τη τυπική απόκλιση της για το σύνολο των 52 κτηρίων. Έτσι, προέκυψε η συνολική μέση ωριαία κατανάλωση φυσικού αερίου των κτηρίων. Η μελέτη που ακολουθεί και οι προβλέψεις που θα παραχθούν θα αφορούν τη συνολική κατανάλωση φυσικού αερίου και για τα 52 αυτά κτήρια.



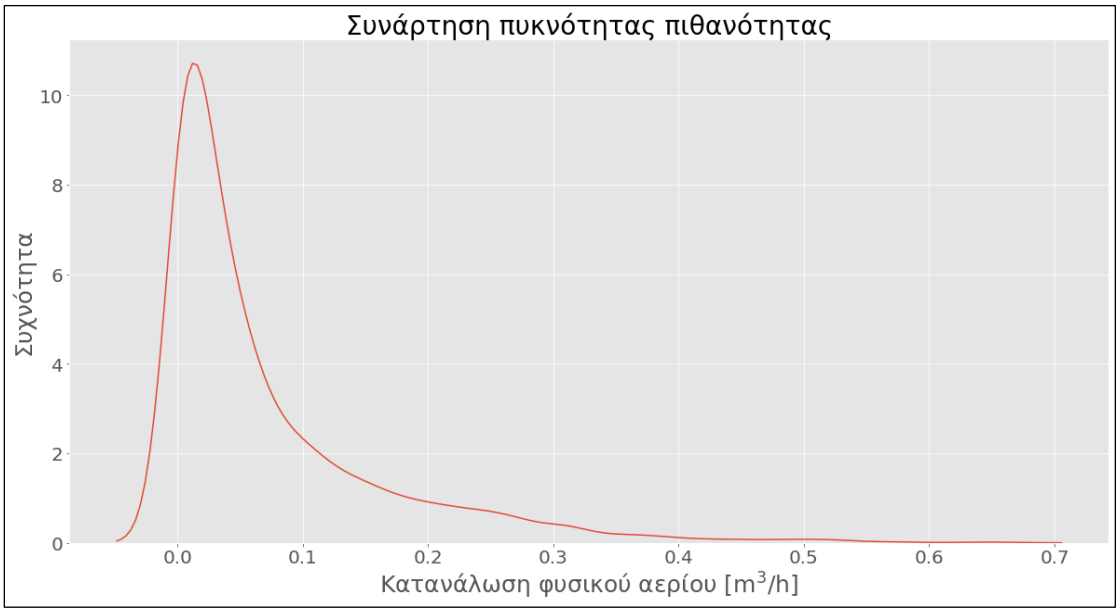
Εικόνα 5: Μέση ωριαία κατανάλωση φυσικού αερίου με ύπαρξη ακραίων τιμών

Παρατηρώντας την παραπάνω γραφική παράσταση, φαίνεται πως κάποιες τιμές κατανάλωσης είναι ιδιαίτερα υψηλές. Συγκεκριμένα, 11 μετρήσεις ωριαίας κατανάλωσης ξεπέρασαν τα $0.7\ m^3/h$. Οι μετρήσεις αυτές αντιμετωπίστηκαν ως outliers και διαγράφησαν από το σύνολο δεδομένων. Ακόμα, φαίνεται να υπάρχουν κάποιες κενές εγγραφές. Καθώς το πλήθος τους ήταν ιδιαίτερα μικρό κι αφού δεν υπάρχει επιστημονικός λόγος να γίνει κάτι διαφορετικό, επιλέχθηκε να διαγραφούν, ακολουθώντας την τακτική της σχετικής εργασίας [KV2019]. Στην *Εικόνα 6*, αποτυπώνεται η ωριαία κατανάλωση φυσικού αερίου του συνόλου των 52 κτηρίων κατά το υπό μελέτη χρονικό διάστημα. Η εικόνα αυτή παρουσιάζει την αναμενόμενη μορφή της καμπύλης κατανάλωσης του φυσικού αερίου. Κατά τους χειμερινούς μήνες η κατανάλωση είναι υψηλή, λόγω της ανάγκης θέρμανσης των κτηρίων, ενώ τους καλοκαιρινούς μήνες παραμένει σε χαμηλά επίπεδα.



Εικόνα 6: Μέση ωριαία κατανάλωση φυσικού αερίου

Ενδιαφέρον παρουσιάζει η συνάρτηση πυκνότητας πιθανότητας της κατανάλωσης φυσικού αερίου. Όπως φαίνεται στην *Εικόνα 7*, η κατανάλωση δεν ακολουθεί την κανονική κατανομή, αλλά μια ασύμμετρη κατανομή με θετική τιμή ασυμμετρίας. Παρουσιάζει, δηλαδή, μια «ουρά» που απλώνεται προς τα δεξιά. Η κορυφή της καμπύλης βρίσκεται σε μια σχετικά χαμηλή τιμή της κατανάλωσης, η οποία παραμένει σταθερή καθ' όλη τη διάρκεια του έτους και αφορά χρήσεις του φυσικού αερίου ανεξάρτητες από την ανάγκη θέρμανσης, όπως το μαγείρεμα και το ζέσταμα του νερού. Οι περισσότερες παρατηρήσεις βρίσκονται δεξιά της κορυφής αυτής και αφορούν την επιπλέον κατανάλωση φυσικού αερίου που απαιτείται λόγω της ανάγκης θέρμανσης κατά τους χειμερινούς μήνες.



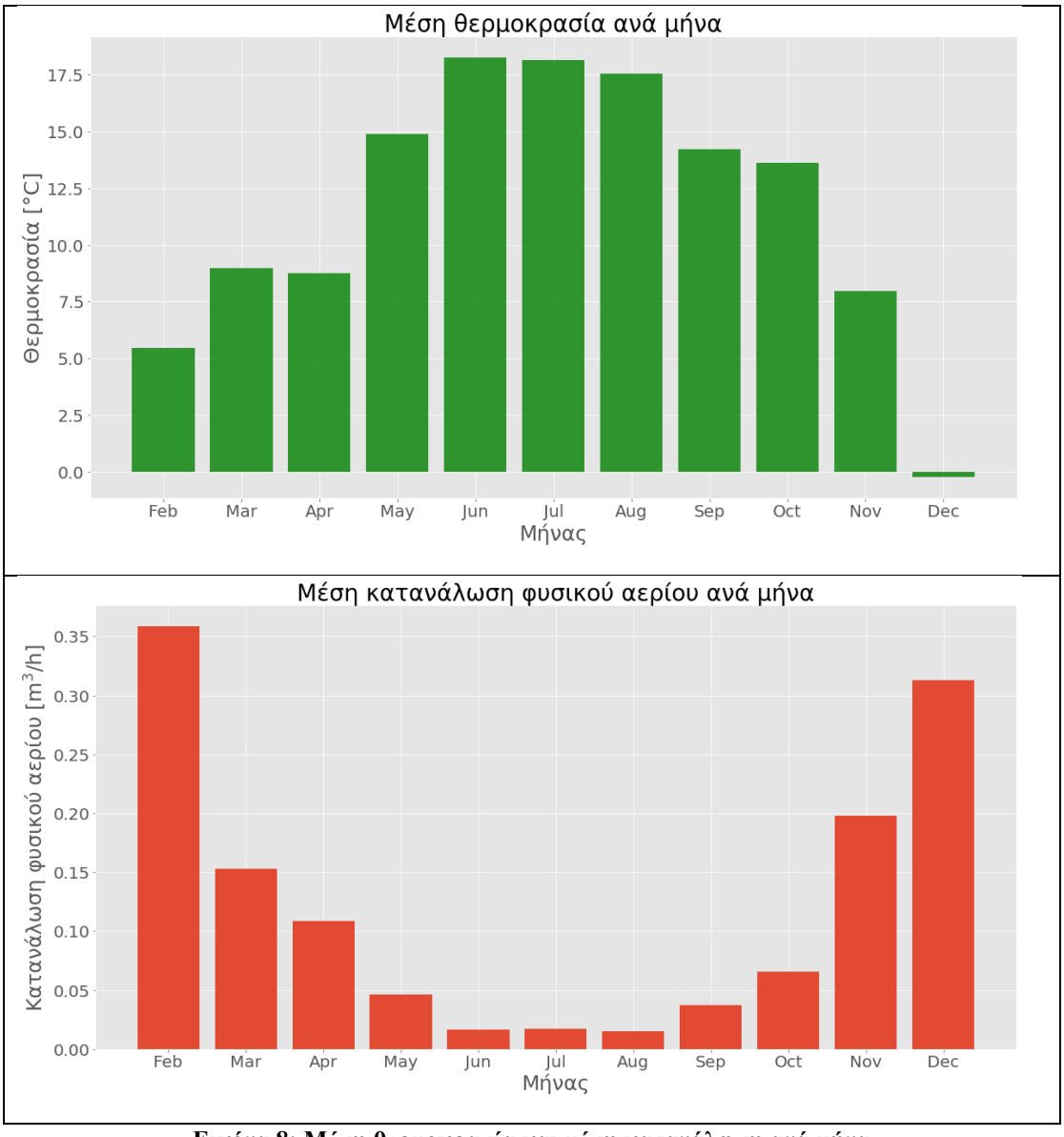
Εικόνα 7: Συνάρτηση πυκνότητας πιθανότητας κατανάλωσης φυσικού αερίου

Στον *Πίνακα 15* φαίνονται κάποια βασικά στατιστικά στοιχεία για τη κατανάλωση της περιοχής κατά το υπό μελέτη έτος.

Κατανάλωση φυσικού αερίου $[m^3/h]$	
Μέση τιμή	0.072582
Τυπική απόκλιση	0.093298
Ελάχιστη τιμή	0
25 εκατοστημόριο	0.010193
50 εκατοστημόριο	0.035658
75 εκατοστημόριο	0.098511
Μέγιστη τιμή	0.657043

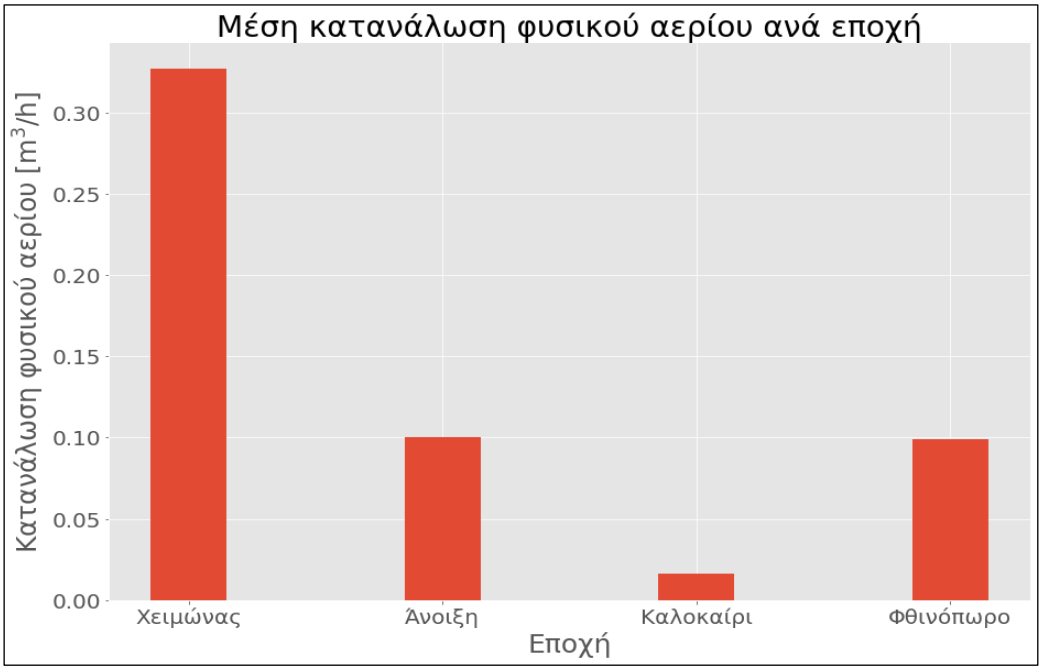
Πίνακας 15: Στατιστικά στοιχεία κατανάλωσης

Όπως έχει αναφερθεί, ο σημαντικότερος παράγοντας που επηρεάζει την κατανάλωση φυσικού αερίου είναι η θερμοκρασία. Στην *Εικόνα 8*, απεικονίζονται η μέση τιμή της θερμοκρασίας για κάθε μήνα του έτους 2017, καθώς και η αντίστοιχη μέση μηνιαία κατανάλωση φυσικού αερίου. Από την εικόνα αυτή, γίνεται προφανές ότι κατά τους καλοκαιρινούς μήνες, κατά τους οποίους η θερμοκρασία είναι υψηλή, η κατανάλωση βρίσκεται σε ιδιαίτερα χαμηλά επίπεδα. Αντίθετα, τους χειμερινούς μήνες, όταν η θερμοκρασία είναι χαμηλή, η κατανάλωση φυσικού αερίου φτάνει τα μέγιστα επίπεδά της, λόγω της ανάγκης θέρμανσης που δημιουργείται. Ακόμα, φαίνεται πως ανεξάρτητα από τη θερμοκρασία, ακόμα και κατά τους πλέον θερμούς μήνες, η κατανάλωση φυσικού αερίου δεν μηδενίζεται ποτέ, καθώς καλύπτει εργασίες ανεξάρτητες από την ανάγκη θέρμανσης, όπως αναφέρθηκε στην προηγούμενη παράγραφο.

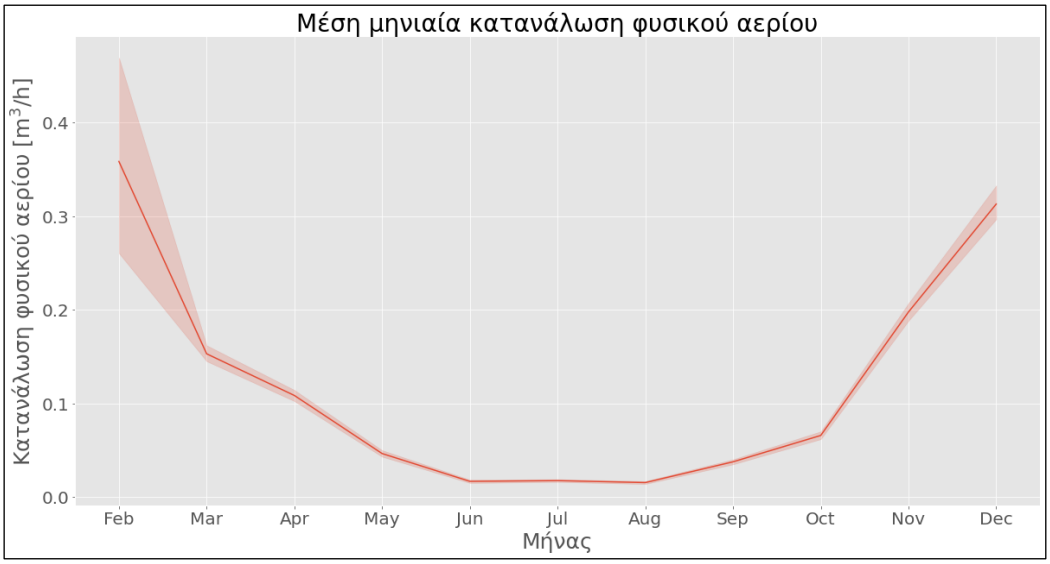


Εικόνα 8: Μέση θερμοκρασία και μέση κατανάλωση ανά μήνα

Τα επίπεδα κατανάλωσης σε κάθε εποχή του χρόνου, καθώς και η καμπύλη της μηνιαίας κατανάλωσης κατά το έτος 2017, απεικονίζονται στις δύο επόμενες εικόνες αντίστοιχα.



Εικόνα 9: Μέση κατανάλωση φυσικού αερίου ανά εποχή

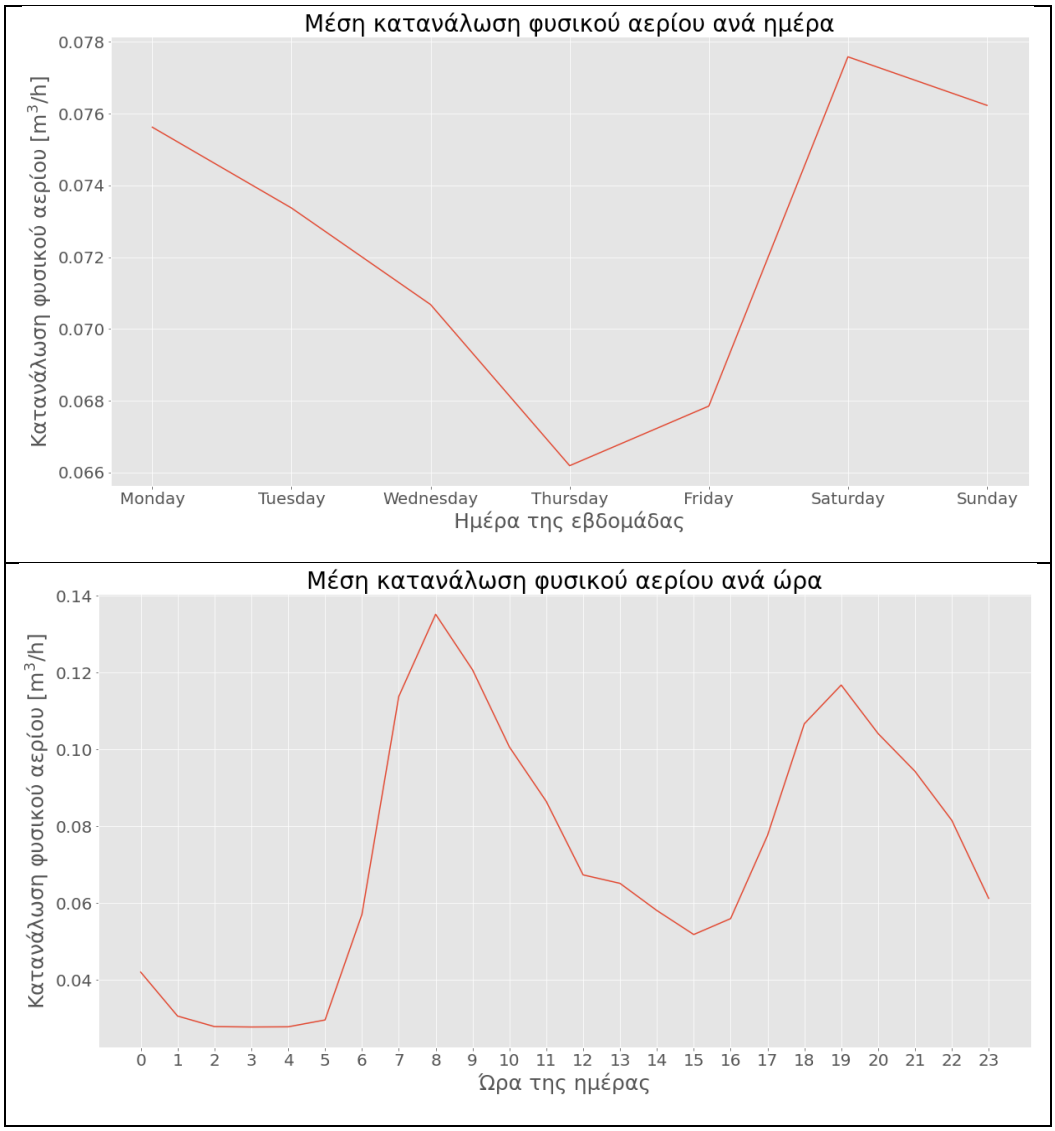


Εικόνα 10: Μηνιαία κατανάλωση φυσικού αερίου κατά το έτος 2017

Από τις *Εικόνες 9* και *10*, γίνεται εμφανής η έντονη εποχιακότητα που παρουσιάζει η κατανάλωση φυσικού αερίου. Κατά τους χειμερινούς μήνες, η κατανάλωση είναι ιδιαίτερα υψηλή και με σχετικά μεγάλη τυπική απόκλιση από τη μέση τιμή. Είναι η περίοδος που υπάρχει μεγάλη ανάγκη θέρμανσης των κατοικιών. Αντίθετα, τους υπόλοιπους μήνες, όταν οι θερμοκρασίες είναι πιο υψηλές και παύει να υπάρχει ανάγκη θέρμανσης, η κατανάλωση

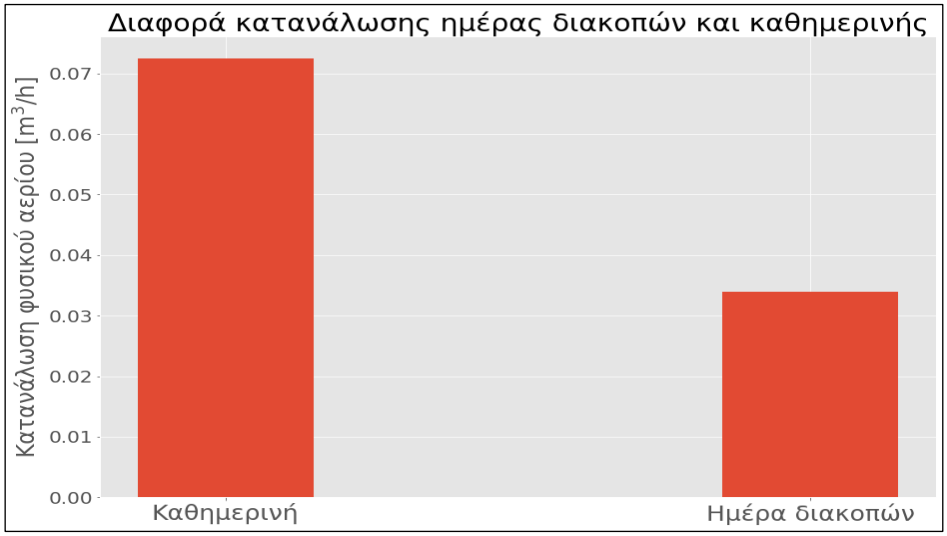
φαίνεται να παραμένει σε πολύ χαμηλά επίπεδα με σχεδόν μηδενική, μάλιστα, τυπική απόκλιση.

Διακυμάνσεις στα επίπεδα της κατανάλωσης παρουσιάζονται ακόμα και εντός της εβδομάδας, καθώς επίσης και κατά τη διάρκεια της ημέρας. Όπως φαίνεται στις παρακάτω εικόνες, η κατανάλωση φτάνει το μέγιστο επίπεδό της το Σαββατοκύριακο, ενώ στη συνέχεια παρατηρείται μια σταδιακή πτώση μέχρι την Πέμπτη. Τότε η κατανάλωση έχει την ελάχιστη τιμή της. Παράλληλα, κατά τη διάρκεια της ημέρας, η κατανάλωση παίρνει τις μέγιστες τιμές της τις πρωινές ώρες, καθώς και το απόγευμα, ενώ τις βραδινές ώρες είναι σχεδόν μηδενική.



Εικόνα 11: Μέση κατανάλωση φυσικού αερίου ανά ημέρα (πάνω) και ανά ώρα (κάτω)

Διαφορετικά φαίνεται να είναι επίσης, τα επίπεδα κατανάλωσης μεταξύ των ημερών διακοπών και των υπόλοιπων ημερών του χρόνου, όπως αναφέρεται και στη σχετική βιβλιογραφία. [FM2020]. Συγκεκριμένα, τις ημέρες των διακοπών η κατανάλωση φυσικού αερίου μειώνεται στο μισό σε σχέση με τις υπόλοιπες ημέρες του έτους.



Εικόνα 12: Διαφορά κατανάλωσης ημέρας διακοπών και υπόλοιπων ημερών του έτους

Όπως έχει ήδη αναφερθεί, η κατανάλωση φυσικού αερίου εξαρτάται σε μεγάλο βαθμό από διάφορους ημερολογιακούς παράγοντες. Οι παραπάνω παρατηρήσεις είναι πολύ σημαντικές για την επιλογή των χαρακτηριστικών των μοντέλων πρόβλεψης, καθώς αναδεικνύουν τις τάσεις και καταναλωτικές συμπεριφορές που διαμορφώνονται από τους παράγοντες αυτούς.

4.3 Επιλογή Χαρακτηριστικών

Ιδιαίτερα σημαντικό βήμα μιας εργασίας πρόβλεψης κατανάλωσης φυσικού αερίου, είναι η επιλογή των μεταβλητών εκείνων, που θα χρησιμοποιηθούν ως χαρακτηριστικά των μοντέλων πρόβλεψης. Αν και το σύνολο δεδομένων που χρησιμοποιήθηκε παρείχε μετρήσεις πλήθους καιρικών παραγόντων, η τυφλή χρήση όλων τους ως χαρακτηριστικά των μοντέλων, δε συνεπάγεται υψηλότερη απόδοση προβλέψεων. Αντίθετα, η χρήση περιττών χαρακτηριστικών κατά την εκπαίδευση των μοντέλων, αυξάνει τη διακύμανση τους (variance), γεγονός που συχνά οδηγεί στο φαινόμενο της υπερπροσαρμοστηκότητας (overfitting) [MC2009]. Έτσι, η ανάλυση συσχέτισης (correlation analysis) των παραγόντων και ο συνδυασμός τους για δημιουργία νέων με υψηλότερη συσχέτιση με την κατανάλωση είναι ουσιώδους σημασίας για την τελική επιλογή των χαρακτηριστικών των μοντέλων.

Το βασικότερο εργαλείο επιλογής χαρακτηριστικών των μοντέλων είναι η *ανάλυση συσχέτισης* (*correlation analysis*). Ο συντελεστής συσχέτισης είναι ένας δείκτης, ο οποίος

καθορίζει το βαθμό συσχέτισης ανάμεσα σε δύο μεταβλητές. Αποτελεί στατιστικό κριτήριο που χρησιμοποιείται για να διαπιστωθεί αν υπάρχει αλληλεξάρτηση μεταξύ των μεταβλητών.

Οι παράγοντες, οι οποίοι εξετάστηκαν και επιλέχθηκαν ως χαρακτηριστικά των μοντέλων στην παρούσα εργασία, ήταν η κατανάλωση ηλεκτρικής ενέργειας, διάφοροι καιρικοί παράγοντες, καθώς και ημερολογιακοί παράγοντες. Η κατανάλωση ηλεκτρικής ενέργειας και οι καιρικοί παράγοντες αποτελούν συνεχείς μεταβλητές. Μεταβλητές δηλαδή, οι οποίες μπορούν να πάρουν οποιαδήποτε τιμή σε ένα δεδομένο διάστημα. Οι παράγοντες αυτοί θα αποτελέσουν τα *αριθμητικά χαρακτηριστικά* (*numerical features*) των μοντέλων. Αντίστοιχα, οι ημερολογιακοί παράγοντες αποτελούν διακριτές μεταβλητές. Οι παράγοντες αυτοί, θα αποτελέσουν τα *κατηγορικά χαρακτηριστικά* (*categorical features*) των μοντέλων. Η διαδικασία επιλογής των χαρακτηριστικών, επιμερίστηκε σε δύο βήματα. Το πρώτο βήμα αποτελεί τη διαδικασία επιλογής των αριθμητικών χαρακτηριστικών που θα χρησιμοποιηθούν ως μεταβλητές εισόδου των μοντέλων, ενώ το δεύτερο βήμα τη διαδικασία επιλογής των αντίστοιχων κατηγορικών χαρακτηριστικών.

4.3.1 Μελέτη συσχετίσεων αριθμητικών χαρακτηριστικών

Για την επιλογή των χαρακτηριστικών αυτών, εξετάστηκε η συσχέτιση των διαφόρων αριθμητικών παραγόντων με την κατανάλωση φυσικού αερίου, με τη βοήθεια του συντελεστή συσχέτισης Pearson. Ο συντελεστής αυτός, αφορά δύο αριθμητικές μεταβλητές που αποτελούνται από συνεχή δεδομένα και προσπαθεί να εξετάσει κατά πόσο οι μεταβλητές αυτές σχετίζονται μεταξύ τους και ποια είναι η ένταση της σχέσης. Για δύο συνεχείς τυχαίες μεταβλητές ο *συντελεστής συσχέτισης Pearson* συμβολίζεται με r και ορίζεται από τη σχέση:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

όπου,

$$s_{xy} = Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \underline{x}) \cdot (y_i - \underline{y})}{n-1}$$

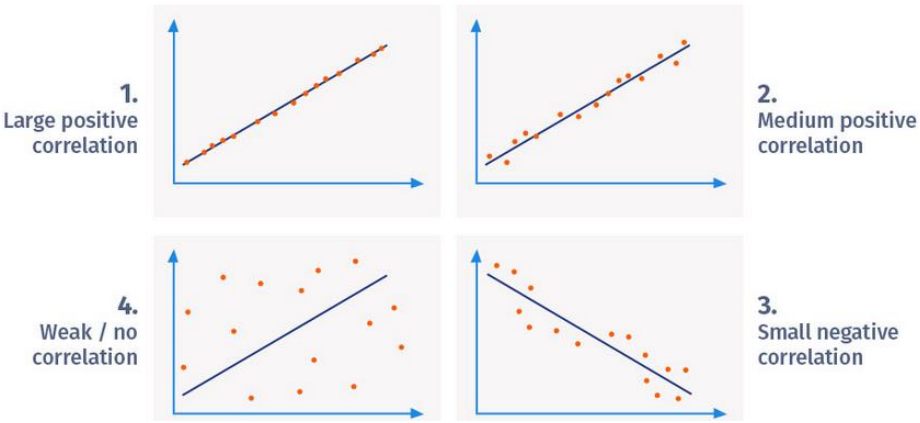
$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \underline{x}\right)^2} \text{ και } s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \underline{y}\right)^2}$$

Επομένως,

$$r == \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \underline{x}) \cdot (y_i - \underline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \underline{x}\right)^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \underline{y}\right)^2}}$$

Ο συντελεστής συσχέτισης Pearson δίνει ένα μέτρο του μεγέθους της γραμμικής συσχέτισης μεταξύ δύο μεταβλητών. Λαμβάνει τιμές στο κλειστό διάστημα $[-1,1]$.

- Όταν $r = 1$, υπάρχει τέλεια θετική συσχέτιση ανάμεσα στις δύο μεταβλητές.
- Όταν $0 < r < 1$, υπάρχει θετική συσχέτιση ανάμεσα στις δύο μεταβλητές, ενώ όσο πιο κοντά στη μονάδα βρίσκεται ο συντελεστής, τόσο πιο ισχυρή είναι η συσχέτιση.
- Όταν $r = 0$, δεν υφίσταται καμία γραμμική συσχέτιση. Αυτό δε σημαίνει, ωστόσο, ότι δεν υπάρχει άλλου είδους συσχέτιση μεταξύ των μεταβλητών.
- Όταν $-1 < r < 0$, υπάρχει αρνητική συσχέτιση μεταξύ των μεταβλητών.
- Όταν $r = -1$, υπάρχει τέλεια αρνητική συσχέτιση μεταξύ των μεταβλητών.



Εικόνα 13: Διαγράμματα απεικόνισης συσχετίσεων

Το σύνολο δεδομένων, περιείχε μετρήσεις της κατανάλωσης ηλεκτρικής ενέργειας καθώς και των καιρικών παραγόντων που καταγράφονται στον *Πίνακα 16*.

Περιγραφή	**Μονάδα**
Διεύθυνση ανέμου	deg
Διάρκεια βροχόπτωσης	s
Μέγιστη ριπή ανέμου στα 10 m	m/s
Ταχύτητα ανέμου στα 10 m	m/s
Νεφοκάλυψη	okta
Ατμοσφαιρική πίεση	hPa
Ηλιακή ακτινοβολία	W/m^2
Βροχόπτωση	mm/h
Διάρκεια ηλιοφάνειας	min
Θερμοκρασία στο 1.5 m	^oC
Ελάχιστη θερμοκρασία στα 10 cm	^oC
Σημείο δρόσου	^oC
Σχετική υγρασία στο 1.5 m	-
Ορατότητα	m

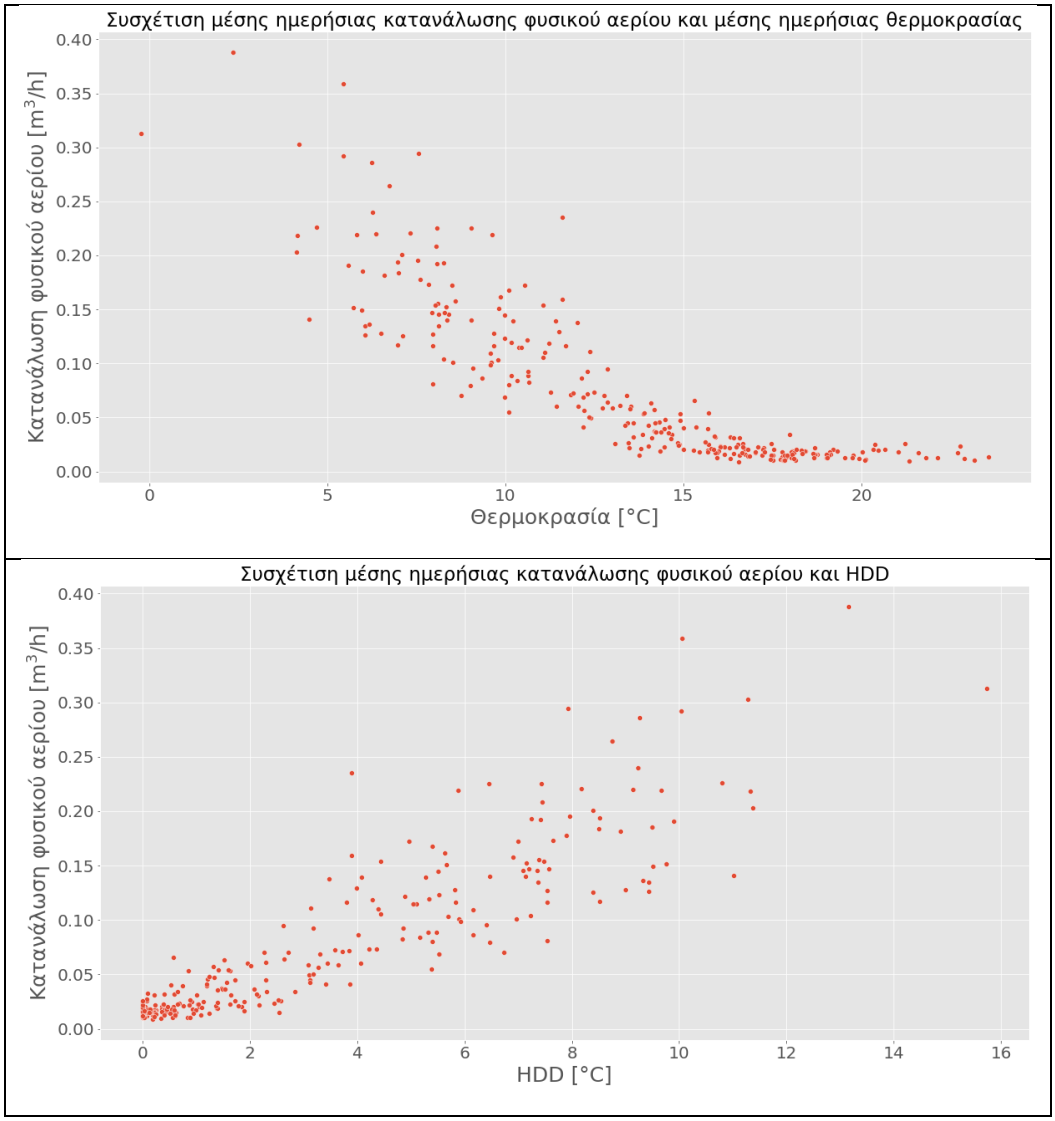
Πίνακας 16: Καιρικοί παράγοντες

Πέρα από τους διαθέσιμους από το σύνολο δεδομένων καιρικούς παράγοντες, σχηματίστηκε επίσης ο *παράγοντας HDD* και εξετάστηκε η συσχέτισή του με την κατανάλωση φυσικού αερίου. Ο παράγοντας *Heating Degree Days* (*HDD*), είναι ένα υπολογιζόμενο μέγεθος που χρησιμοποιείται για την ποσοτική εκτίμηση και υπολογισμό της απαιτούμενης ενέργειας για τη θέρμανση των εσωτερικών χώρων. Ο παράγοντας αυτός ορίζεται από τη σχέση:

$$HDD(T) = max(T_b - T, 0)$$

Η μεταβλητή T_b, ονομάζεται θερμοκρασία βάσης και εξαρτάται από τη γεωγραφική θέση και τις επικρατούσες στο υπό εξέταση σημείο κλιματολογικές συνθήκες. Στην παρούσα εργασία χρησιμοποιήθηκε η τιμή $T_b = 15.5^oC$, η οποία είναι η τιμή της θερμοκρασίας βάσης που χρησιμοποιείται για τις χώρες της Ευρωπαϊκής Ένωσης. Ο παράγοντας HDD παρουσιάζει καλύτερη γραμμική συσχέτιση με την κατανάλωση φυσικού αερίου σε σχέση με τη θερμοκρασία.

Όπως φαίνεται στην *Εικόνα 14*, η μέση ημερήσια κατανάλωση μειώνεται γραμμικά με την αύξηση της θερμοκρασίας. Αυτό, ωστόσο, συμβαίνει μέχρι η θερμοκρασία να φτάσει τους $15^oC - 16^oC$. Στην θερμοκρασία αυτή, παύει να υπάρχει ανάγκη για θέρμανση και οι θερμοστάτες κλείνουν. Κατά συνέπεια, η κατανάλωση σταματά να έχει γραμμική συσχέτιση με τη θερμοκρασία, καθώς φτάνει σε ένα ελάχιστο επίπεδο χωρίς περεταίρω μείωση με την αύξηση της θερμοκρασίας. Ο συντελεστής συσχέτισης Pearson μεταξύ της μέσης ημερήσιας κατανάλωσης φυσικού αερίου και της μέσης ημερήσιας τιμής της θερμοκρασίας είναι $r = -0.86$. Παρατηρώντας, λοιπόν την *Εικόνα 14*, γίνεται εμφανής η καλύτερη συσχέτιση του παράγοντα HDD με την κατανάλωση, η οποία αυξάνεται γραμμικά με την αύξηση της τιμής του. Ο συντελεστής συσχέτισης Pearson μεταξύ της μέσης ημερήσιας κατανάλωσης φυσικού αερίου και του HDD είναι $r = 0.90$.



Εικόνα 14: Συσχέτιση κατανάλωσης φυσικού αερίου και θερμοκρασίας (πάνω), Συσχέτιση κατανάλωσης φυσικού αερίου και HDD (κάτω)

Στον *Πίνακα 17* αποτυπώνεται η τιμή του συντελεστή συσχέτισης Pearson της μέσης ωριαίας κατανάλωσης φυσικού αερίου με κάθε έναν από τους καιρικούς παράγοντες, καθώς και με την κατανάλωση ηλεκτρικής ενέργειας. Ο παράγοντας HDD, η θερμοκρασία και το σημείο δρόσου αποτελούν τους τρεις παράγοντες με την υψηλότερη συσχέτιση με την κατανάλωση φυσικού αερίου, επιβεβαιώνοντας έτσι το γεγονός ότι η θερμοκρασία αποτελεί το βασικότερο παράγοντα που επηρεάζει την κατανάλωση φυσικού αερίου.

Ενδιαφέρον παρουσιάζει το γεγονός ότι η κατανάλωση φυσικού αερίου φαίνεται να έχει σημαντική γραμμική συσχέτιση με την κατανάλωση ηλεκτρικής ενέργειας. Η χρήση του ηλεκτρικού ρεύματος είναι άμεσα συνδεδεμένη με τη σύγχρονη ζωή και ενδεχομένως τα επίπεδα κατανάλωσής του υποδεικνύουν μια ανάλογη τάση γενικότερης ενεργειακής κατανάλωσης.

Η ηλιακή ακτινοβολία φαίνεται επίσης να σχετίζεται γραμμικά με την κατανάλωση φυσικού αερίου και άρα να αποτελεί παράγονται που επηρεάζει τα επίπεδά της, όπως έχει προταθεί κι από σχετική εργασία [SP2014].

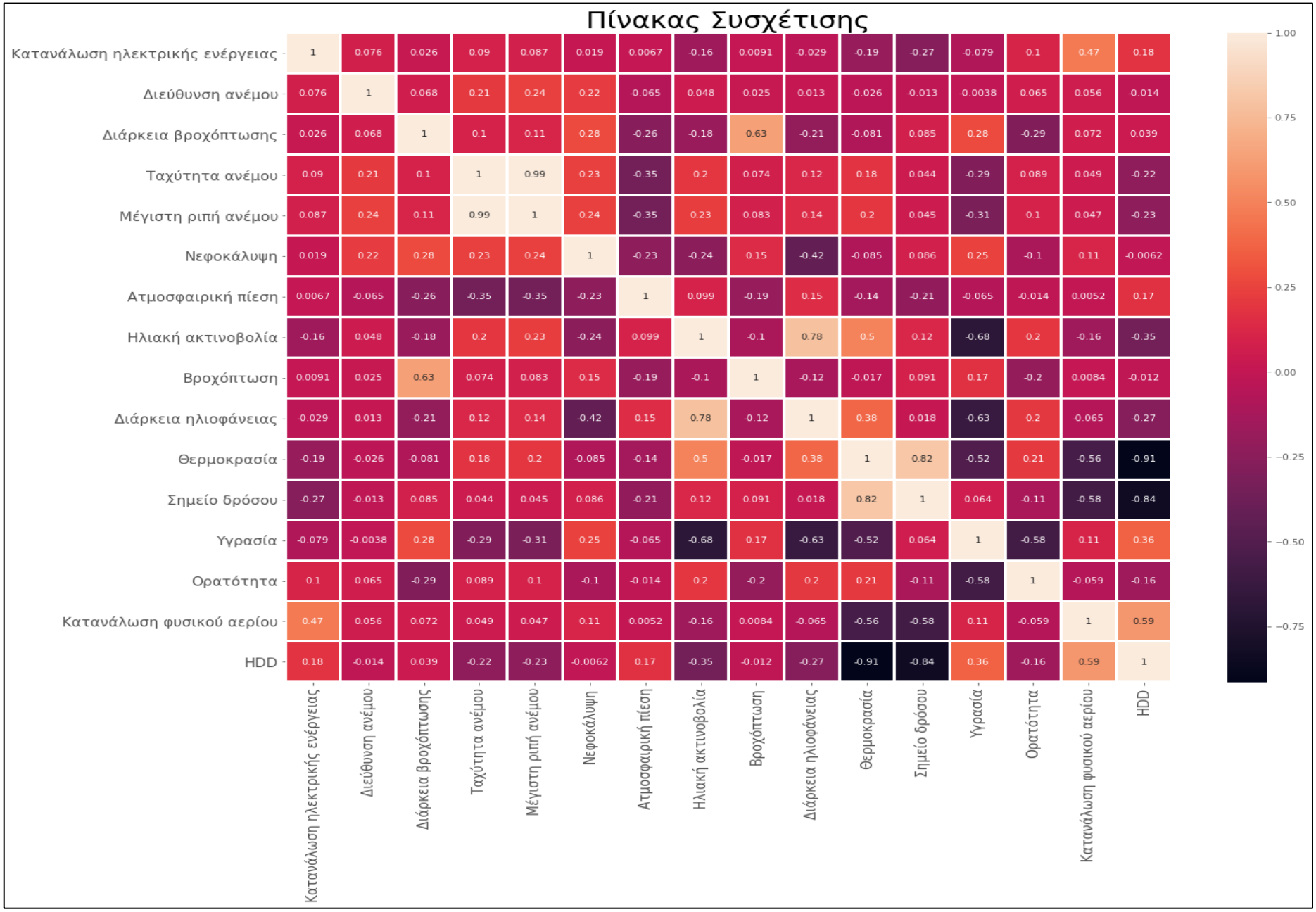
Τέλος, σύμφωνα με τις αντίστοιχες τιμές του συντελεστή συσχέτισης Pearson, οι υπόλοιποι καιρικοί παράγοντες δε φαίνεται να διαθέτουν γραμμική συσχέτιση με την κατανάλωση φυσικού αερίου.

Συντελεστής συσχέτισης με κατανάλωση φυσικού αερίου	
HDD	0.592
Κατανάλωση ηλεκτρικής ενέργειας	0.467
Σχετική υγρασία	0.110
Νεφοκάλυψη	0.110
Διάρκεια βροχόπτωσης	0.072
Διεύθυνση ανέμου	0.055
Ταχύτητα ανέμου	0.045
Μέγιστη ριπή ανέμου	0.047
Βροχόπτωση	0.008
Ατμοσφαιρική πίεση	0.005
Ορατότητα	-0.059
Διάρκεια ηλιοφάνειας	-0.065
Ηλιακή ακτινοβολία	-0.157
Σημείο δρόσου	-0.575
Θερμοκρασία	-0.557

Πίνακας 17: Συντελεστές συσχέτισης παραγόντων με μέση ημερήσια κατανάλωση φυσικού αερίου

Στην *Εικόνα 15*, αποτυπώνεται ο πίνακας συσχετίσεων των παραπάνω μεταβλητών. Σε αυτόν καταγράφονται τόσο οι συσχετίσεις των διάφορων παραγόντων με την κατανάλωση φυσικού αερίου, όσο και οι μεταξύ τους συσχετίσεις. Για την καλύτερη οπτική παρουσίαση των συσχετίσεων αυτών δημιουργήθηκε ο αντίστοιχος χάρτης θερμότητας (heat map).

Από τον πίνακα συσχετίσεων και τον αντίστοιχο χάρτη θερμότητας, φαίνεται πως κάποιοι από τους καιρικούς παράγοντες παρουσιάζουν ισχυρή γραμμική συσχέτιση μεταξύ τους. Αυτό είναι αναμενόμενο, καθώς το σημείο δρόσου και ο παράγοντας HDD εξαρτώνται άμεσα από την θερμοκρασία, η διάρκεια ηλιοφάνειας από τη νεφοκάλυψη, η ηλιακή ακτινοβολία από τη διάρκεια ηλιοφάνειας, ενώ υπάρχουν κι άλλες λιγότερο προφανείς συσχετίσεις.



Εικόνα 15: Χάρτης θερμότητας συσχετίσεων

Προκειμένου τα μοντέλα πρόβλεψης που θα σχεδιαστούν να λάβουν τη μεγαλύτερη δυνατή πληροφορία από τους διαθέσιμους καιρικούς παράγοντες, έγινε συνδυασμός κάποιων από αυτούς, με σκοπό τη δημιουργία νέων σύνθετων παραγόντων, οι οποίοι θα παρουσιάζουν αυξημένη συσχέτιση με την κατανάλωση φυσικού αερίου και θα οδηγήσουν τελικά σε μεγαλύτερη απόδοση προβλέψεων.

Οι νέοι σύνθετοι παράγοντες αυτοί, προέκυψαν προσθέτοντας, αφαιρώντας και πολλαπλασιάζοντας τις τιμές μέτρησης συσχετισμένων μεταξύ τους καιρικών παραγόντων. Στα πλαίσια αυτής της τεχνικής κατασκευής χαρακτηριστικών, σχηματίστηκαν τέσσερις νέοι σύνθετοι παράγοντες:

1. Ο πρώτος, αποτελεί συνδυασμό των ισχυρά συσχετισμένων μεταξύ τους θερμοκρασιακών παραγόντων *HDD*, *θερμοκρασία* και *σημείου δρόσου*. Ορίζεται από τη σχέση:

$$[Θερμοκρασία\ συνδυασμός] = [θερμοκρασία] + [σημείο\ δρόσου] - [HDD]$$

2. Ο δεύτερος, αποτελεί συνδυασμό των ισχυρά συσχετισμένων μεταξύ τους παραγόντων *διάρκεια ηλιοφάνειας* και *ηλιακής ακτινοβολίας*. Ορίζεται από τη σχέση:

$$[Ακτινοβολία\ συνδυασμός] = [διάρκεια\ ηλιοφάνειας] + [ηλιακή\ ακτινοβολία]$$

3. Ο τρίτος, αποτελεί συνδυασμό των ισχυρά συσχετισμένων μεταξύ τους παραγόντων *ταχύτητα ανέμου*, *μέγιστη ριπή ανέμου* και *διεύθυνση ανέμου*. Ορίζεται από τη σχέση:

$$[Άνεμος\ συνδυασμός] = [ταχύτητα\ ανέμου] + [μέγιστη\ ριπή\ ανέμου] + [διεύθυνση\ ανέμου]$$

4. Ο τέταρτος αποτελεί συνδυασμό των ισχυρά συσχετισμένων μεταξύ τους παραγόντων *διάρκεια βροχόπτωσης*, *ατμοσφαιρική πίεση*, *σχετική υγρασία*, *νεφοκάλυψη* και *βροχόπτωση*. Ορίζεται από τη σχέση:

$$[Βροχή\ συνδυασμός] = [διάρκεια\ βροχόπτωσης] + [ατμοσφαιρική\ πίεση] + [σχετική\ υγρασία] \cdot [νεφοκάλυψη] + [βροχόπτωση]$$

Στον *Πίνακα 18* καταγράφονται οι τιμές του συντελεστή συσχέτισης Pearson της μέσης ωριαίας κατανάλωσης φυσικού αερίου με τους τέσσερις νέους παράγοντες που σχηματίστηκαν:

Συντελεστής συσχέτισης με κατανάλωση φυσικού αερίου	
Βροχή συνδυασμός	0.133
Άνεμος συνδυασμός	0.058
Ακτινοβολία συνδυασμός	-0.156
Θερμοκρασία συνδυασμός	-0.602

Πίνακας 18: Συντελεστές συσχέτισης παραγόντων που σχηματίστηκαν με μέση ωριαία κατανάλωση φυσικού αερίου

4.3.2 Επιλογή αριθμητικών χαρακτηριστικών

Στον *Πίνακα 19*, καταγράφεται η τιμή του συντελεστή συσχέτισης Pearson τόσο των διαθέσιμων από το σύνολο δεδομένων καιρικών παραγόντων, όσο και των σύνθετων παραγόντων που σχηματίστηκαν. Επιπλέον, καταγράφεται η τιμή του συντελεστή συσχέτισης της κατανάλωσης ηλεκτρικής ενέργειας, καθώς και της τιμής της κατανάλωσης φυσικού αερίου της ακριβώς προηγούμενης ώρας.

Συντελεστής συσχέτισης με κατανάλωση φυσικού αερίου	
Κατανάλωσης ΦΑ προηγούμενης ώρας	0.837
HDD	0.592
Κατανάλωση ηλεκτρικής ενέργειας	0.467
Βροχή συνδυασμός	0.133
Σχετική υγρασία	0.110
Νεφοκάλυψη	0.110
Διάρκεια βροχόπτωσης	0.072
Άνεμος συνδυασμός	0.058
Διεύθυνση ανέμου	0.055
Ταχύτητα ανέμου	0.045
Μέγιστη ριπή ανέμου	0.047
Βροχόπτωση	0.008
Ατμοσφαιρική πίεση	0.005
Ορατότητα	-0.059
Διάρκεια ηλιοφάνειας	-0.065
Ακτινοβολία συνδυασμός	-0.156
Ηλιακή ακτινοβολία	-0.157
Σημείο δρόσου	-0.575
Θερμοκρασία	-0.557
Θερμοκρασία συνδυασμός	-0.602

Πίνακας 19: Συντελεστής συσχέτισης αριθμητικών παραγόντων με μέση ωριαία κατανάλωση φυσικού αερίου

Από τον *Πίνακα 19*, φαίνεται ότι οι νέοι σύνθετοι παράγοντες που σχηματίστηκαν, παρουσιάζουν καλύτερη συσχέτιση με την κατανάλωση φυσικού αερίου σε σχέση με τους επιμέρους παράγοντες που τους αποτελούν. Ο παράγοντας [$Άνεμος\_συνδυασμός$], εξακολουθεί, ωστόσο, να παρουσιάζει ιδιαίτερα χαμηλή συσχέτιση με την κατανάλωση φυσικού αερίου και επιλέχθηκε να μη χρησιμοποιηθεί ως χαρακτηριστικό των μοντέλων.

Στον *Πίνακα 20* καταγράφεται η τελική επιλογή των μεταβλητών που θα αποτελέσουν τα αριθμητικά χαρακτηριστικά των μοντέλων.

Αριθμητικά χαρακτηριστικά μοντέλων πρόβλεψης
Κατανάλωση φυσικού αερίου προηγούμενης ώρας
Κατανάλωση ηλεκτρικής ενέργειας
Θερμοκρασία συνδυασμός
Βροχή συνδυασμός
Ακτινοβολία συνδυασμός

Πίνακας 20: Αριθμητικά χαρακτηριστικά μοντέλων πρόβλεψης

4.3.3 Μελέτη συσχετίσεων κατηγορικών χαρακτηριστικών

Στην περίπτωση κατά την οποία πρέπει να χρησιμοποιηθεί μια κατηγορική μεταβλητή X με ν επίπεδα $X = 0,1,2,\dots,\nu$ σαν ανεξάρτητη μεταβλητή, δηλαδή σα χαρακτηριστικό ενός μοντέλου πρόβλεψης, τότε πρέπει αντί αυτής να εισαχθούν στο μοντέλο πρόβλεψης $\nu - 1$ το πλήθος δυαδικές μεταβλητές. Έτσι, οι παρακάτω ημερολογιακοί παράγοντες, τροποποιήθηκαν κατάλληλα, προκειμένου να χρησιμοποιηθούν ως χαρακτηριστικά των μοντέλων πρόβλεψης:

- *Ώρα της ημέρας*
- *Ημέρα της εβδομάδας*
- *Μήνας του έτους*
- *Εποχή του έτους*

Επιπλέον, οι δύο επόμενοι ημερολογιακοί παράγοντες εξετάστηκαν ως πιθανά χαρακτηριστικά των μοντέλων:

- *Σαββατοκύριακο*: Δυαδική μεταβλητή, η οποία λαμβάνει τιμή 1 όταν η προκείμενη μέρα είναι Σαββατοκύριακο, διαφορετικά λαμβάνει τιμή 0.
- *Διακοπές*: Δυαδική μεταβλητή, η οποία λαμβάνει τιμή 1 όταν η προκείμενη μέρα είναι ημέρα διακοπών, διαφορετικά λαμβάνει τιμή 0.

Έτσι, ο συνολικός αριθμός των δυαδικών μεταβλητών που προκύπτουν είναι 48 και αποτελείται από:

- 24 δυαδικές μεταβλητές που αντιστοιχούν σε κάθε ώρα της ημέρας
- 7 δυαδικές μεταβλητές που αντιστοιχούν σε κάθε ημέρα της εβδομάδας
- 11 δυαδικές μεταβλητές που αντιστοιχούν σε κάθε μήνα του χρόνου που περιλαμβάνεται στο σύνολο δεδομένων
- 4 δυαδικές μεταβλητές που αντιστοιχούν στις τέσσερις εποχές του έτους
- 1 δυαδική μεταβλητή για το Σαββατοκύριακο
- 1 δυαδική μεταβλητή για τις Διακοπές

Η επιλογή του κατάλληλου συνδυασμού τους έγινε μελετώντας τη συσχέτιση των παραγόντων αυτών με την ωριαία τιμή της κατανάλωσης φυσικού αερίου, με βοήθεια του συντελεστή συσχέτισης Point Biserial. Ο συντελεστής αυτός, φανερώνει το βαθμό συσχέτισης δύο μεταβλητών, η μια εκ των οποίων έχει δύο μόνο βαθμίδες (διχοτομημένη) και η άλλη είναι συνεχής και αριθμητική. Η ερμηνεία του είναι ίδια με την αντίστοιχη ερμηνεία του συντελεστή συσχέτισης Pearson.

Έστω Y μια συνεχής τυχαία μεταβλητή και X μια δυαδική τυχαία μεταβλητή, όπου $X \in (0,1)$ και έστω ότι υπάρχουν n διαθέσιμα ζευγάρια παρατηρήσεων $(Y_k, X_k), k = 1,2, \dots, n$. Ο *συντελεστής συσχέτισης Point Biserial*, ορίζεται από τη σχέση:

$$r_{pb} = \left(\frac{\underline{Y_1} - \underline{Y_0}}{S_Y}\right)\sqrt{\frac{np_0(1-p_0)}{n-1}}$$

όπου,

$$S_Y = \sqrt{\frac{\sum_{k=1}^{n}\left(Y_k - \underline{Y}\right)^2}{n-1}}, \quad \underline{Y} = \frac{\sum_{k=1}^{n} Y_k}{n}, \quad p_1 = \frac{\sum_{k=1}^{n} X_k}{n}, \quad p_0 = 1 - p_1$$

Με Y_k, X_k η k-οστή παρατήρηση του κάθε δείγματος, $\underline{Y}$ η μέση τιμή της συνεχούς μεταβλητής Y, n το πλήθος του δείγματος, p_1η μέση τιμή των παρατηρήσεων της X και p_0 η απόκλιση της μέσης τιμής των X από τη μονάδα.

Στον επόμενο πίνακα αποτυπώνεται η τιμή του συντελεστή συσχέτισης Point Biserial της μέσης ωριαίας κατανάλωσης φυσικού αερίου με κάθε έναν από τις δυαδικές μεταβλητές που προέκυψαν από τους ημερολογιακούς παράγοντες.

Συντελεστής συσχέτισης με κατανάλωση φυσικού αερίου			
Νοέμβριος	0.473	Τετάρτη	-0.008
Μάρτιος	0.298	Ώρα 12:00	-0.011
Άνοιξη	0.210	Διακοπές	-0.013
Φθινόπωρο	0.205	Ώρα 13:00	-0.017
Ώρα 08:00	0.140	Παρασκευή	-0.020
Απρίλιος	0.131	Ώρα 23:00	-0.025
Χειμώνας	0.124	Οκτώβριος	-0.026
Ώρα 09:00	0.108	Πέμπτη	-0.028
Ώρα 19:00	0.099	Ώρα 14:00	-0.032
Δεκέμβριος	0.097	Ώρα 06:00	-0.035
Ώρα 07:00	0.093	Ώρα 16:00	-0.037
Φεβρουάριος	0.077	Ώρα 15:00	-0.046
Ώρα 18:00	0.076	Ώρα 00:00	-0.068
Ώρα 20:00	0.070	Ώρα 01:00	-0.094
Ώρα 10:00	0.063	Ώρα 05:00	-0.096
Ώρα 21:00	0.049	Ώρα 02:00	-0.100
Ώρα 11:00	0.031	Ώρα 04:00	-0.100
Σαββατοκύριακο	0.030	Ώρα 03:00	-0.100
Σάββατο	0.022	Μάιος	-0.100
Ώρα 22:00	0.020	Σεπτέμβριος	-0.135
Κυριακή	0.016	Ιούλιος	-0.200
Δευτέρα	0.014	Ιούνιος	-0.215
Ώρα 17:00	0.011	Αύγουστος	-0.224
Τρίτη	0.004	Καλοκαίρι	-0.426

Πίνακας 21: Συντελεστής συσχέτισης ημερολογιακών παραγόντων με μέση ωριαία κατανάλωση φυσικού αερίου

Η εποχή του καλοκαιριού, καθώς και οι καλοκαιρινοί μήνες του έτους, είναι οι παράγοντες με την ισχυρότερη αρνητική συσχέτιση με την κατανάλωση φυσικού αερίου. Αυτό είναι αναμενόμενο, λόγω των γνωστών καταναλωτικών συνηθειών, σύμφωνα με τις οποίες, οι θερμοστάτες τότε παραμένουν κυρίως κλειστοί. Αντίθετα, τη μεγαλύτερη θετική συσχέτιση, παρουσιάζουν οι μήνες και οι εποχές του χρόνου με τα υψηλότερα επίπεδα κατανάλωσης φυσικού αερίου. Αν και ο Φεβρουάριος με το Δεκέμβρη είναι οι μήνες με τη μεγαλύτερη κατανάλωση, σύμφωνα με την ανάλυση της *Ενότητα 4.2*, στον παραπάνω πίνακα φαίνεται πως ο Νοέμβρης και ο Μάρτιος έχουν υψηλότερη συσχέτιση με αυτή. Αυτό

συμβαίνει, διότι το πλήθος εγγραφών κατά το Φεβρουάριο και Δεκέμβρη είναι μικρό, σε αντίθεση με τους επόμενους μήνες και κατά συνέπεια η συσχέτιση αυτή δε γίνεται εμφανής σε όρους συντελεστών συσχέτισης. Οι υπόλοιποι ημερολογιακοί παράγοντες, όπως οι ώρες της ημέρας και οι ημέρες της εβδομάδας, παρουσιάζουν μικρή τιμή συντελεστή συσχέτισης με την κατανάλωση. Ακόμα, ο συντελεστής συσχέτισης των διακοπών και Σαββατοκύριακων, αν και εκφράζει τη γενική κατεύθυνση με την έννοια του προσήμου, δε υπονοεί ισχυρή συσχέτιση. Αυτό οφείλεται στο σχετικά μικρό πλήθος των εγγραφών του συνόλου δεδομένων. Με την αύξηση των εγγραφών για περισσότερα του ενός έτη παρατηρήσεων, οι παραπάνω συσχετίσεις θα γίνουν πιο ενδεικτικές των πραγματικών καταναλωτικών τάσεων.

Προκειμένου να μοντελοποιηθεί καλύτερα η συσχέτιση των ημερολογιακών παραγόντων με την κατανάλωση φυσικού αερίου, όπως και στην περίπτωση των αριθμητικών χαρακτηριστικών, έγινε συνδυασμός κάποιων από τις παραπάνω δυαδικές μεταβλητές, με σκοπό τη δημιουργία νέων, οι οποίες θα παρουσιάζουν αυξημένη συσχέτιση με την κατανάλωση φυσικού αερίου και θα οδηγήσουν τελικά σε μεγαλύτερη απόδοση προβλέψεων. Δημιουργήθηκαν επομένως πέντε νέες μεταβλητές:

1. $Σαββατοκύριακα\ αιχμής = (Νοέμβριος + Μάρτιος) \cdot Σαββατοκύριακο$
2. $Μήνες\ αιχμής = Νοέμβριος + Δεκέμβριος + Φεβρουάριος$
3. $Μήνες\ ύφεσης = Ιούνιος + Ιούλιος + Αύγουστος + Σεπτέμβριος$
4. $Ώρες\ αιχμής = Ώρα\ 07{:}00 + Ώρα\ 08{:}00 + Ώρα\ 09{:}00 + Ώρα\ 11{:}00 + Ώρα\ 17{:}00 + Ώρα\ 18{:}00 + Ώρα\ 19{:}00 + Ώρα\ 20{:}00 + Ώρα\ 21{:}00 + Ώρα\ 22{:}00$
5. $Ώρες\ ύφεσης = Ώρα\ 00{:}00 + Ώρα\ 01{:}00 + Ώρα\ 02{:}00 + Ώρα\ 04{:}00 + Ώρα\ 05{:}00$

4.3.4 Επιλογή κατηγορικών χαρακτηριστικών

Οι τιμές του συντελεστή συσχέτισης της μέσης ωριαίας κατανάλωσης φυσικού αερίου με τους πέντε νέους παράγοντες που σχηματίστηκαν φαίνονται παρακάτω:

Συντελεστής συσχέτισης με κατανάλωση φυσικού αερίου	
Μήνες αιχμής	0.487
Σαββατοκύριακα αιχμής	0.323
Ώρες αιχμής	0.305
Ώρες ύφεσης	-0.257
Μήνες ύφεσης	-0.490

Πίνακας 22:Συντελεστές συσχέτισης νέων παραγόντων με μέση ωριαία κατανάλωση φυσικού αερίου

Η συσχέτιση των νέων αυτών μεταβλητών με την κατανάλωση φυσικού αερίου είναι ισχυρότερη από τις συσχετίσεις των επιμέρους μεταβλητών. Επιπλέον, με τη δημιουργία αυτών των σύνθετων μεταβλητών, επιτυγχάνεται η μοντελοποίηση των βασικότερων τάσεων της κατανάλωσης εντός του έτους, ενώ επίσης, μειώνεται σημαντικά το συνολικό πλήθος των χαρακτηριστικών που θα χρησιμοποιηθούν και άρα ο χρόνος εκπαίδευσης των μοντέλων. Ειδικότερα, οι μεταβλητές *Μήνες αιχμής* και *Μήνες ύφεσης*, συνδυάζουν σε μια μεταβλητή, τους μήνες με τη μεγαλύτερη και την ελάχιστη κατανάλωση αντίστοιχα. Με ανάλογο τρόπο, οι μεταβλητές *Ώρες αιχμής* και *Ώρες ύφεσης*, συνδυάζουν σε μια μεταβλητή τις ώρες της ημέρας που η κατανάλωση αναμένεται μέγιστη και σχεδόν μηδενική, αντίστοιχα. Τέλος, η μεταβλητή *Σαββατοκύριακα αιχμής*, μοντελοποιεί το γεγονός ότι η κατανάλωση τα Σαββατοκύριακα αναμένεται υψηλότερη των καθημερινών, ειδικά τους μήνες αιχμής Νοέμβριο και Μάρτιο. Στον *Πίνακα 23* καταγράφεται η τελική επιλογή των μεταβλητών που θα αποτελέσουν τα κατηγορικά χαρακτηριστικά των μοντέλων.

Κατηγορικά χαρακτηριστικά μοντέλων πρόβλεψης
Μήνες αιχμής
Σαββατοκύριακα αιχμής
Ώρες αιχμής
Άνοιξη
Φθινόπωρο
Χειμώνας
Ώρες ύφεσης
Καλοκαίρι
Μήνες ύφεσης

Πίνακας 23: Κατηγορικά χαρακτηριστικά μοντέλων πρόβλεψης

4.3.5 Σύνοψη αριθμητικών και κατηγορικών χαρακτηριστικών

Στον *Πίνακα 24,* παρουσιάζονται όλοι οι παράγοντες που επιλέχθηκαν να χρησιμοποιηθούν ως χαρακτηριστικά των μοντέλων πρόβλεψης. Επίσης, γίνεται μια σύντομη περιγραφή τους, καθώς και σημειώνεται το αν ήταν διαθέσιμοι από το αρχικό σύνολο δεδομένων ή σχηματίστηκαν στα πλαίσια της ανάλυσης που προηγήθηκε.

Χαρακτηριστικά μοντέλων πρόβλεψης	**Περιγραφή**	**Διαθέσιμο**	**Σχηματίστηκε**
Αριθμητικά χαρακτηριστικά			
Κατανάλωσης ΦΑ προηγούμενης ώρας	Η τιμή της κατανάλωσης φυσικού αερίου την προηγούμενη ώρα από την ώρα της ζητούμενης πρόβλεψης κατανάλωσης ΦΑ		✔
Κατανάλωση ηλεκτρικής ενέργειας	Η τιμή της κατανάλωσης της ηλεκτρικής ενέργειας την ώρα της ζητούμενης πρόβλεψης κατανάλωσης ΦΑ	✔	
Θερμοκρασία συνδυασμός	Η τιμή της σύνθετης μεταβλητής που σχηματίστηκε σύμφωνα με τη σχέση της *Υποενότητας 4.3.1* την ώρα της ζητούμενης πρόβλεψης κατανάλωσης ΦΑ		✔
Βροχή συνδυασμός	Η τιμή της σύνθετης μεταβλητής που σχηματίστηκε, σύμφωνα με τη σχέση της *Υποενότητας 4.3.1* την ώρα της ζητούμενης πρόβλεψης κατανάλωσης ΦΑ		✔
Ακτινοβολία συνδυασμός	Η τιμή της σύνθετης μεταβλητής που σχηματίστηκε, σύμφωνα με τη σχέση της *Υποενότητας 4.3.1* την ώρα της ζητούμενης πρόβλεψης κατανάλωσης ΦΑ		✔
Κατηγορικά χαρακτηριστικά			
Μήνες αιχμής	Δυαδική μεταβλητή που λαμβάνει τιμή 1, όταν η ζητούμενη ωριαία πρόβλεψη κατανάλωσης ανήκει στους μήνες Νοέμβριο, Δεκέμβριο ή Φεβρουάριο, όπως αναφέρεται στην *Υποενότητα 4.3.3*		✔
Σαββατοκύριακα αιχμής	Δυαδική μεταβλητή που λαμβάνει τιμή 1, όταν η ζητούμενη ωριαία πρόβλεψη κατανάλωσης ΦΑ αφορά Σαββατοκύριακο των μηνών Νοεμβρίου και Μαρτίου, όπως αναφέρεται στην *Υποενότητα 4.3.3*		✔
Ώρες αιχμής	Δυαδική μεταβλητή που λαμβάνει τιμή 1, όταν η ζητούμενη ωριαία πρόβλεψη κατανάλωσης ανήκει σε ώρες αιχμής, όπως αναφέρεται στην *Υποενότητα 4.3.3*		✔
Άνοιξη	Δυαδική μεταβλητή που λαμβάνει τιμή 1, όταν η ζητούμενη ωριαία πρόβλεψη κατανάλωσης ανήκει στους μήνες της Άνοιξης		✔
Φθινόπωρο	Δυαδική μεταβλητή που λαμβάνει τιμή 1, όταν η ζητούμενη ωριαία πρόβλεψη κατανάλωσης ανήκει στους μήνες του Φθινοπώρου		✔
Χειμώνας	Δυαδική μεταβλητή που λαμβάνει τιμή 1, όταν η ζητούμενη ωριαία πρόβλεψη κατανάλωσης ανήκει στους μήνες του Χειμώνα		✔

Καλοκαίρι	Δυαδική μεταβλητή που λαμβάνει τιμή 1, όταν η ζητούμενη ωριαία πρόβλεψη κατανάλωσης ανήκει στους μήνες του Καλοκαιριού		✔
Ώρες ύφεσης	Δυαδική μεταβλητή που λαμβάνει τιμή 1, όταν η ζητούμενη ωριαία πρόβλεψη κατανάλωσης ανήκει σε ώρες ύφεσης, όπως αναφέρεται στην *Υποενότητα 4.3.3*		✔
Μήνες ύφεσης	Δυαδική μεταβλητή που λαμβάνει τιμή 1, όταν η ζητούμενη ωριαία πρόβλεψη κατανάλωσης ανήκει στους μήνες Ιούνιο, Ιούλιο, Αύγουστο ή Σεπτέμβρη όπως αναφέρεται στην *Υποενότητα 4.3.3*		✔
Μεταβλητή στόχος			
Κατανάλωση φυσικού αερίου	Η τιμή πρόβλεψης των μοντέλων της ωριαίας κατανάλωσης φυσικού αερίου	✔	

Πίνακας 24: Χαρακτηριστικά μοντέλων πρόβλεψης

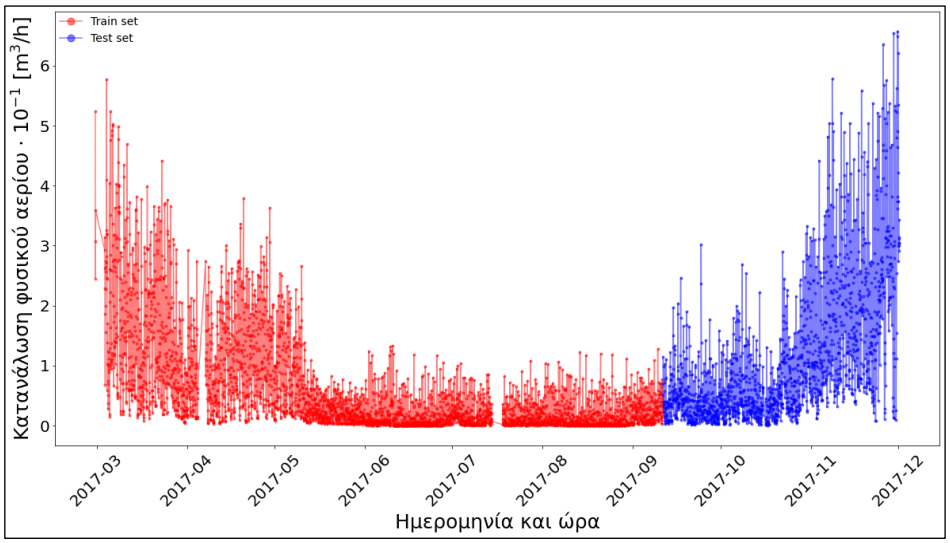
4.4 Εκπαίδευση και Ρύθμιση Υπερπαραμέτρων των Μοντέλων

Στο παρόν κεφάλαιο, παρουσιάζεται η μεθοδολογία που ακολουθήθηκε για την εκπαίδευση των μοντέλων πρόβλεψης που εξετάστηκαν, καθώς και για την ρύθμιση των υπερπαραμέτρων τους (hyperparameter tuning).

4.4.1 Διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο δοκιμής

Το σύνολο δεδομένων χωρίστηκε σε δύο επιμέρους σύνολα, τα οποία χρησιμοποιήθηκαν για την εκπαίδευση και την αξιολόγηση των μοντέλων, αντίστοιχα. Το *σύνολο εκπαίδευσης* (*train set*) περιείχε το 70% του συνόλου δεδομένων, ενώ το *σύνολο δοκιμής* (*test set*) το 30%. Στα πλαίσια της εργασίας, πραγματοποιήθηκαν τρεις διαφορετικοί τρόποι διαχωρισμού (split) του συνόλου δεδομένων, σύμφωνα με τις παραπάνω αναλογίες. Σε κάθε έναν, οι υπερπαράμετροι των μοντέλων πρόβλεψης ρυθμίστηκαν εκ νέου, αξιολογήθηκαν οι παραχθείσες προβλέψεις και σχολιάστηκαν τα αντίστοιχα συμπεράσματα.

Στον **πρώτο τρόπο**, το *σύνολο εκπαίδευσης* (*train set*) περιείχε το πρώτο 70% του συνόλου δεδομένων, ενώ το *σύνολο δοκιμής* (*test set*) το τελευταίο 30%. Αυτός είναι ο τρόπος διαχωρισμού των δεδομένων που ακολουθήθηκε από την εργασία [KV2019].

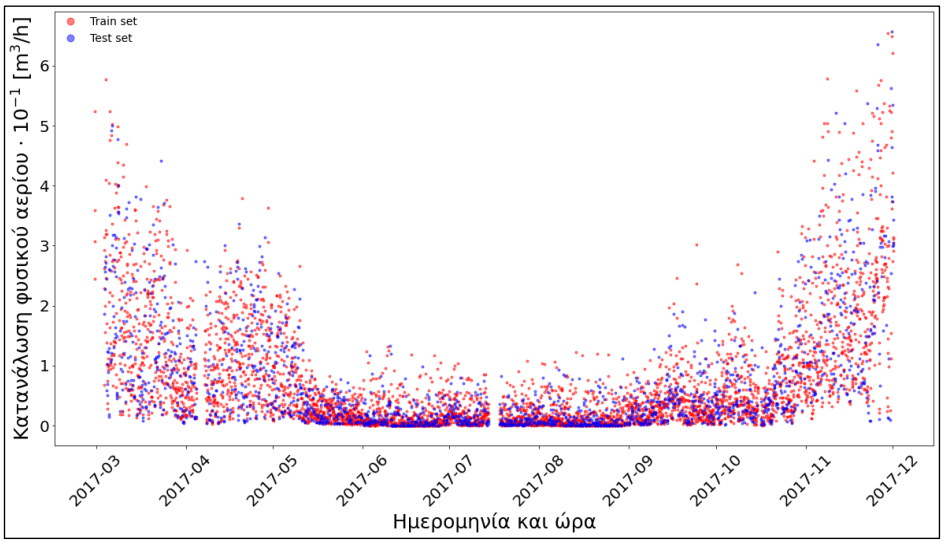


Εικόνα 16: Σύνολα εκπαίδευσης και δοκιμής μοντέλων (I)

Με τον τρόπο δειγματοληψίας αυτό, ωστόσο, τα μοντέλα πρόβλεψης εκπαιδεύονται με ένα μόνο μέρος του συνόλου δεδομένων, συγκεκριμένα με δεδομένα των σχετικά θερμότερων πρώτων μηνών του έτους, ενώ καλούνται να παράγουν προβλέψεις για μήνες που δεν

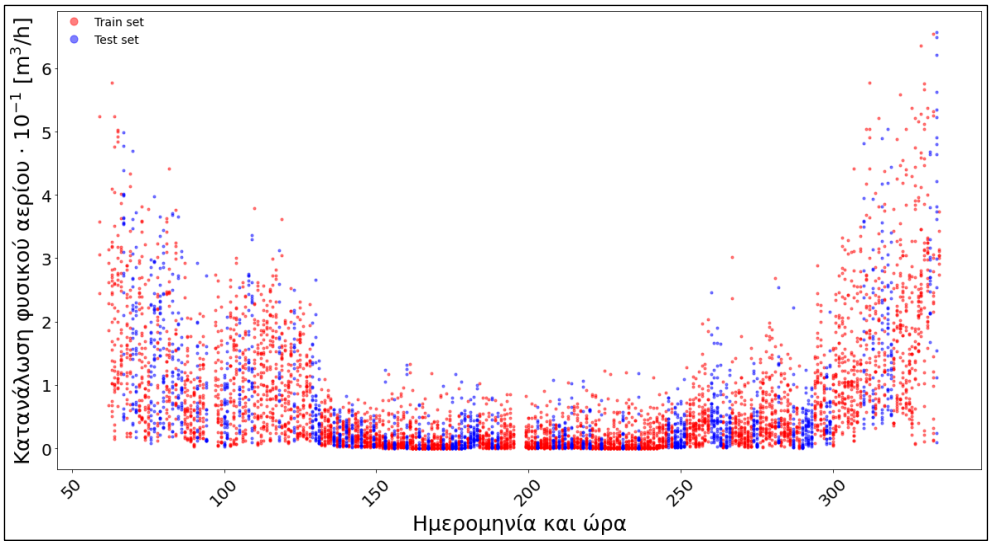
περιλαμβάνονται στο σύνολο εκπαίδευσής τους. Έτσι, τα μοντέλα πρόβλεψης αυτά θα χαρακτηρίζονται από ένα βαθμό μεροληψίας (bias).

Στον **δεύτερο τρόπο**, τα *σύνολα εκπαίδευσης* (*train set*) και *δοκιμής* (*test set*) περιλάμβαναν δεδομένα από όλο το έτος, στις αναλογίες που αναφέρθηκαν. Με αυτό τον τρόπο διαχωρισμού, παύει να υπάρχει μεροληψία (bias) κατά την εκπαίδευση των μοντέλων.



Εικόνα 17: Σύνολα εκπαίδευσης και δοκιμής μοντέλων (ΙΙ)

Στον **τρίτο τρόπο**, τα *σύνολα εκπαίδευσης* (*train set*) και *δοκιμής* (*test set*) περιλάμβαναν δεδομένα από όλο το έτος, στις αναλογίες που αναφέρθηκαν, αλλά ο διαχωρισμός έγινε σύμφωνα με την ημέρα της κατανάλωσης. Έτσι, εξασφαλίζεται ότι όλες οι ώρες μιας ημέρας θα ανήκουν ενιαία είτε στο σύνολο εκπαίδευσης είτε στο σύνολο δοκιμής. Ο τρόπος διαχωρισμού αυτός χρησιμοποιήθηκε για την παραγωγή ημερησίων προβλέψεων κατανάλωσης, αθροίζοντας τις επιμέρους ωριαίες προβλέψεις των μοντέλων.



Εικόνα 18: Σύνολα εκπαίδευσης και δοκιμής μοντέλων (ΙΙΙ)

Είναι σημαντικό τα σύνολα αυτά να κανονικοποιηθούν προτού χρησιμοποιηθούν για την εκπαίδευση των μοντέλων, όπως απαιτείται από τα περισσότερα μοντέλα μηχανικής μάθησης, προκειμένου να έχουν σωστή συμπεριφορά. *Κανονικοποίηση* (*Standarization*) ή *κλιμάκωση των χαρακτηριστικών εισόδου* (*Feature Scaling*), είναι μια διαδικασία μετασχηματισμού των δεδομένων, κατά την οποία οι τιμές των αριθμητικών χαρακτηριστικών αντικαθίστανται με άλλες, οι οποίες ανήκουν στην ίδια κλίμακα. Σημειώνεται ότι μόνο τα αριθμητικά χαρακτηριστικά απαιτείται να κανονικοποιηθούν και όχι και τα κατηγορικά. Υπάρχουν διάφορες μέθοδοι κανονικοποίησης. Στην παρούσα διπλωματική εργασία, η μέθοδος χρησιμοποιήθηκε η *μέθοδος της τυποποίησης* (*Standard Scaler*). Η μέθοδος αυτή κρατάει το μέσο όρο των υπό κανονικοποίηση αριθμητικών τιμών στο 0 και την τυπικής απόκλιση στο 1, έχοντας τη μορφή της κανονικής κατανομής. Έτσι, μετά την κανονικοποίηση αυτή, οι τιμές όλων των αριθμητικών χαρακτηριστικών των μοντέλων ανήκουν στο διάστημα $[-1,1]$. Υπολογίζεται από τον τύπο:

$$x_{std}^{i} = \frac{x^{i} - \mu_x}{\sigma_x}$$

Το x^i είναι η τιμή του συγκεκριμένο δείγματος, το μ_x ο δειγματικός μέσος και σ_x η τυπική απόκλιση.

4.4.2 Βέλτιστη ρύθμιση των υπερπαραμέτρων

Η σωστή λειτουργία πολλών από τους αλγορίθμους μηχανικής μάθησης, βασίζεται στη σωστή ρύθμιση των παραμέτρων τους. Οι *υπερπαράμετροι* (*Hyperparameters*), είναι οι μεταβλητές εκείνες που διέπουν τη διαδικασία εκπαίδευσης του μοντέλου. Η τιμή τους πρέπει να ρυθμιστεί από το χειριστή προτού αρχίσει η διαδικασία εκπαίδευσης, σε αντίθεση με τις απλές παραμέτρους του μοντέλου, η τιμή των οποίων υπολογίζεται αυτόματα από το ίδιο το μοντέλο κατά την εκπαίδευση. Η επιλογή των σωστών υπερπαραμέτρων είναι καθοριστικής σημασίας για την αποδοτική λειτουργία του εκάστοτε μοντέλου. Ωστόσο, δεν υπάρχουν συγκεκριμένοι κανόνες που να προσδιορίζουν πως θα γίνει η συγκεκριμένη επιλογή. Οι αλγόριθμοι που επιτελούν την ρύθμιση των υπερπαραμέτρων λειτουργούν σύμφωνα με τη μέθοδο δοκιμής-σφάλματος, προβαίνοντας σε συνεχόμενες προσαρμογές του μέχρις ότου φτάσουν στις βέλτιστες τιμές. Επομένως, σημαντικό βήμα αποτελεί η επιλογή των τιμών εκείνων που θα υποβληθούν σε αυτή τη διαδικασία βελτιστοποίησης. Υπάρχουν ελάχιστες καθολικές συμβουλές σχετικά με την επιλογή των τιμών αυτών, ενώ η τελική επιτυχία της διαδικασίας εξαρτάται σε μεγάλο βαθμό από την εμπειρία του ερευνητή. Η επιλογή τους πρέπει να γίνεται με σύνεση, καθώς κάθε παράμετρος που επιλέγεται να ρυθμιστεί μπορεί να αυξήσει εκθετικό τον απαιτούμενο αριθμό δοκιμών. Αφού επιλεχθούν οι παράμετροι προς ρύθμιση, εφαρμόζονται αλγόριθμοι, οι οποίοι προελαύνουν το χώρο αναζήτησης που

δημιουργείται, το μέγεθος του οποίου εξαρτάται από το πλήθος και το εύρος των υπερπαραμέτρων που έχουν επιλεχθεί να ρυθμιστούν. Οι δύο πλέον χρησιμοποιούμενοι από τους αλγορίθμους αυτούς είναι οι:

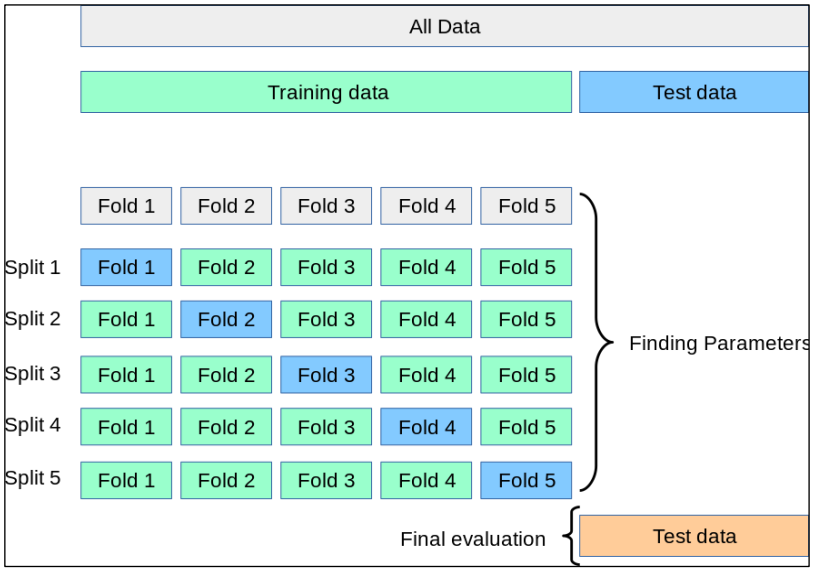
Αναζήτηση πλέγματος (Grid Search). Η αναζήτηση πλέγματος αποτελεί τον απλούστερο αλγόριθμο βελτιστοποίησης υπερπαραμέτρων. Ο αλγόριθμος αυτός εκτελεί μια εξαντλητική αναζήτηση στο προκαθορισμένο χώρο αναζήτησης που δημιουργείται. Ο χώρος αναζήτησης μπορεί να καταλήξει να αποτελεί ένα υπερεπίπεδο δεκάδων διαστάσεων, ανάλογα με το πλήθος των προς ρύθμιση παραμέτρων.

Τυχαία αναζήτηση (Randomized Search). Στην τυχαία αναζήτηση, ο χώρος αναζήτησης διασχίζεται τυχαία έως ότου ικανοποιηθεί κάποιο κριτήριο τερματισμού, όπως ο αριθμός των επαναλήψεων. Ο αλγόριθμος αυτός δεν εγγυάται την εύρεση της βέλτιστης λύσης, αλλά λειτουργεί ικανοποιητικά σε προβλήματα που το πλήθος των υπερπαραμέτρων είναι μικρό, ενώ επιλέγεται επίσης, όταν δεν είναι διαθέσιμη μεγάλη υπολογιστική ισχύς.

Η *υπερπροσαρμογή* (*overfitting*) αποτελεί ένας από τους κύριους λόγους μη καλής απόδοσης ενός μοντέλου μηχανικής μάθησης. Ο ορός *overfitting* χρησιμοποιείται στην επιβλεπόμενη μάθηση για να δηλώσει την κατάσταση κατά την οποία ένα μοντέλο έχει εκπαιδευτεί και εξειδικευτεί στο σύνολο εκπαίδευσης του προβλήματος με αποτέλεσμα να παρουσιάζει χαμηλή ακρίβεια στην πρόβλεψη στο σύνολο δοκιμής.

Η τεχνική *Cross-Validation* (*CV*), αποτελεί την πλέον ενδεικνυόμενη λύση για το παραπάνω πρόβλημα. Με την τεχνική αυτή, δεν απαιτείται πλέον η δέσμευση ενός μέρους του συνόλου εκπαίδευσης σα σύνολο αξιολόγησης, κατά συνέπεια τα μοντέλα εκπαιδεύονται τελικά με το μέγιστο δυνατό αριθμό δειγμάτων. Το σύνολο δοκιμής, ωστόσο, εξακολουθεί να υπάρχει για την τελική αξιολόγηση των μοντέλων. Η βασική πρακτική της εν λόγω τεχνικής ονομάζεται *k-fold Cross-Validation*. Σύμφωνα με αυτή, επιλέγεται ένας σταθερός αριθμός από *folds* (*πτυχές*), δηλαδή συνεχόμενες διαιρέσεις των δεδομένων. Τα δεδομένα διαχωρίζονται σε k προσεγγιστικά ίσα folds και κάθε ένα στη συνέχεια θα χρησιμοποιηθεί επαναληπτικά για την αξιολόγηση, ενώ τα υπόλοιπα για την εκπαίδευση των μοντέλων. Τα $(k-1)$ folds δηλαδή, χρησιμοποιούνται σα σύνολο εκπαίδευσης, ενώ το 1 fold λειτουργεί σα σύνολο αξιολόγησης, με την όλη διαδικασία να επαναλαμβάνεται k φορές. Έτσι, κάθε δείγμα του συνόλου δεδομένων χρησιμοποιείται ακριβώς μια φορά ως μέλος του συνόλου αξιολόγησης και $k-1$ φορές ως μέλος του συνόλου εκπαίδευσης. Τυπικές τιμές του k είναι της τάξεως του 5 έως 10. Η συνολική αξιολόγηση του μοντέλου προκύπτει από τη μέση τιμή των επιμέρους αξιολογήσεων που προέκυψαν κατά τις k επαναλήψεις. Η διαδικασία αυτή, μπορεί να επαναληφθεί για κάθε τιμή των υπερπαραμέτρων που δοκιμάζονται, ώστε να επιλεχθούν τελικά οι βέλτιστες. Είναι σαφές ότι η πρακτική αυτή έχει μεγάλο υπολογιστικό κόστος, καθώς απαιτούνται πολλοί κύκλοι εκπαίδευσης του μοντέλου Το βασικό της

πλεονέκτημα όμως, είναι ότι δε δεσμεύει μεγάλο μέρος των διαθέσιμων δειγμάτων για την αξιολόγηση, κάτι που είναι θεμελιώδους σημασίας όταν ο αριθμός δειγμάτων είναι περιορισμένος. Στην παρούσα εργασία, χρησιμοποιήθηκε η τεχνική *5-fold Cross-Validation*.



Εικόνα 19: Σχηματική απεικόνιση της τεχνικής 5-fold Cross-Validation

Στην παρούσα εργασία, τόσο η μέθοδος *Grid Search* όσο και η *Randomized Search* χρησιμοποιήθηκαν μέσω των σχετικών συναρτήσεων της βιβλιοθήκης Scikit-learn, για την βέλτιστη ρύθμιση των υπερπαραμέτρων των μοντέλων. Η δεύτερη μέθοδος χρησιμοποιήθηκε, λόγω του απαγορευτικά υψηλού υπολογιστικού χρόνου που απαιτούταν, σε περιπτώσεις μοντέλων με μεγάλο αριθμό υπερπαραμέτρων και πολλές υπό δοκιμή τιμές. Οι δύο αυτοί αλγόριθμοι, δέχονται σαν ορίσματα το μοντέλο πρόβλεψης, τα σύνολα τιμών των υπερπαραμέτρων που θα δοκιμαστούν, την τεχνική Cross-Validation που θα εφαρμοστεί και τον τρόπο με τον οποίο θα γίνει η αξιολόγηση (scoring) του μοντέλου. Στη συνέχεια, για κάθε δυνατό συνδυασμό των υπερπαραμέτρων του ορίσματος, εκτελείται η σχετική τεχνική Cross-Validation και προκύπτει η αξιολόγηση του εκάστοτε μοντέλου. Τέλος, ο αλγόριθμος επιλέγει το μοντέλο με εκείνες τις υπερπαραμέτρους που έδωσαν την υψηλότερη απόδοση στο σύνολο αξιολόγησης της τεχνικής Cross-Validation.

4.5 Δείκτες Αξιολόγησης Μοντέλων

Αφού ρυθμιστούν οι υπερπαράμετροι και κατασκευαστεί το βέλτιστο μοντέλο, η τελική αξιολόγηση της απόδοσής του γίνεται στο *σύνολο δοκιμής* (*test set*). Με αυτόν τον τρόπο, αξιολογείται η ικανότητά του να πραγματοποιεί ακριβείς προβλέψεις σε καινούρια δεδομένα. Η μέτρηση της ακρίβειας των μοντέλων επιτυγχάνεται με χρήση ειδικών στατιστικών δεικτών σφαλμάτων. Οι δείκτες σφαλμάτων που χρησιμοποιήθηκαν στην

παρούσα εργασία είναι οι R^2, MAE, MSE και $RMSE$. Στη συνέχεια παρουσιάζεται αναλυτικά κάθε ένας από αυτούς.

4.5.1 Συντελεστής προσδιορισμού R^2

Ο *συντελεστής προσδιορισμού* (*coefficient of determination*), συμβολίζεται με R^2 και μετρά πόση διακύμανση της εξαρτημένης μεταβλητής κατάφεραν να ερμηνεύσουν οι ανεξάρτητες μεταβλητές. Είναι το απλούστερο μέτρο της ικανότητας ενός συνόλου παραγόντων να ερμηνεύσουν ένα φαινόμενο. Αποτελεί, δηλαδή ένα μέτρο του πόσο καλά το μοντέλο έχει προσαρμοστεί στα δεδομένα με τα οποία εκπαιδεύτηκε και συνεπώς του πόσο αποτελεσματικά θα μπορεί να πραγματοποιήσει προβλέψεις σε νέα δεδομένα. Ο συντελεστής προσδιορισμού R^2 είναι ο λόγος της διακύμανσης των εκτιμημένων τιμών της εξαρτημένης μεταβλητής προς τη διακύμανση των πραγματικών τιμών της εξαρτημένης μεταβλητής και υπολογίζεται ως εξής:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\widehat{y_i} - \underline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \underline{y}\right)^2}$$

όπου n είναι ο αριθμός των παρατηρήσεων, y_i είναι οι πραγματικές τιμές της εξαρτημένης μεταβλητής Y, $\underline{y}$ είναι η μέση τιμή της μεταβλητής Y και $\widehat{y_i}$ είναι οι εκτιμημένες από το μοντέλο πρόβλεψης τιμές της Y.

Οι τιμές του συντελεστή προσδιορισμού R^2 κυμαίνονται από το 0 έως 1, ενώ όσο η τιμή του πλησιάζει προς το 1, τόσο καλύτερη προσαρμογή έχει το μοντέλο προς τη μεταβλητή στόχο που προβλέπει.

4.5.2 Μέσο Απόλυτο Σφάλμα (MAE)

Το *μέσο απόλυτο σφάλμα* (*mean absolute error*), συμβολίζεται *MAE* και αποτελεί μέτρο της αστοχίας της προβλεπόμενης τιμής ως προς την πραγματική, χωρίς ωστόσο να λαμβάνεται υπόψη η κατεύθυνση της πρόβλεψης. Διατηρεί τις μονάδες μέτρησης της αρχικής χρονοσειράς, ενώ όσο μεγαλύτερη είναι η τιμή του δείκτη, τόσο μικρότερη προκύπτει η ακρίβεια του μοντέλου. Υπολογίζεται σύμφωνα με τη σχέση:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

4.5.3 Μέσο Τετραγωνικό Σφάλμα (MSE)

Το *μέσο τετραγωνικό σφάλμα* (*mean square error*), συμβολίζεται *MSE* και αποτελεί ένα μέτρο της ακρίβειας των προβλέψεων, δίνοντας ωστόσο πολύ μεγαλύτερο βάρος στα μεγάλα σφάλματα και μικρότερο βάρος στα μικρά σφάλματα. Όμοια με τον δείκτη MAE, υψηλότερες τιμές του *MSE* συνεπάγονται μειωμένη ακρίβεια προβλέψεων. Υπολογίζεται σύμφωνα με τη σχέση:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

4.5.4 Ρίζα Μέσου Τετραγωνικού Σφάλματος (RMSE)

Η *ρίζα μέσου τετραγωνικού σφάλματος (root mean square error)*, συμβολίζεται *RMSE* και έχει τις ίδιες ιδιότητες που αναφέρθηκαν για το μέσο τετραγωνικό σφάλμα (MSE). Βασική διαφορά, αποτελεί το γεγονός ότι το *RMSE* εκφράζεται σε μονάδες της αρχικής χρονοσειράς. Συχνά προτιμάται έναντι του MSE, καθώς τα αποτελέσματά του εκφράζονται στην ίδια κλίμακα με τα δεδομένα πρόβλεψης. Ο δείκτης αυτός είναι πιο ευαίσθητος στα μεγαλύτερα κατ' απόλυτο τιμή σφάλματα. Υπολογίζεται σύμφωνα με τη σχέση:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

4.6 Μοντέλα Δοκιμής

Στην παρούσα ενότητα παρουσιάζονται τα μοντέλα μηχανικής μάθησης που δοκιμάστηκαν στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου.

4.6.1 Linear Regression

Η *γραμμική παλινδρόμηση* (*Linear Regression*) αποτελεί ένα τρόπο μοντελοποίησης της σχέσης μεταξύ μιας μεταβλητής εξόδου και μιας ή περισσοτέρων εισόδων. Η μεταβλητή εξόδου ονομάζεται εξαρτημένη μεταβλητή, ενώ οι μεταβλητές εισόδου ονομάζονται ανεξάρτητες μεταβλητές. Στο μοντέλο της γραμμικής παλινδρόμησης γίνεται η υπόθεση ότι η σχέση αυτή είναι γραμμική, γεγονός που στην πραγματικότητα δε συμβαίνει συχνά. Σημαντική προϋπόθεση για την παραγωγή του μοντέλου αυτού είναι η απουσία συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών. Συχνά, όταν οι ανεξάρτητες μεταβλητές είναι περισσότερες από μια, το μοντέλο ονομάζεται *πολλαπλή γραμμική παλινδρόμηση* (*multiple linear regression*). Το μοντέλο έχει την εξής μορφή:

$$Y = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_i X_{ip} + e_i, \text{ για κάθε δείγμα } i = 1,2,\dots,n$$

Στην παραπάνω σχέση, Y είναι η εξαρτημένη μεταβλητή, X_{ij} είναι το $i^{στο}$ δείγμα της j ανεξάρτητης μεταβλητής X, όπου $j = 1,2,\dots,p$. Ακόμα, b_i είναι οι παράμετροι του μοντέλου και e_i είναι η απόκλιση από τις πραγματικές τιμές. Οι σταθερές παράμετροι b_i του μοντέλου, υπολογίζονται με τη μέθοδο των ελαχίστων τετραγώνων. Στη μέθοδο των *κανονικών ελαχίστων τετραγώνων*, το πρόβλημα έγκειται στην ελαχιστοποίηση της συνάρτησης απωλειών:

$$L(X, y) = \sum_{i=1}^{n} (\hat{y_i} - y_i)^2 = \sum_{i=1}^{n} (b \cdot X_i - y_i)^2$$

Η γραμμική παλινδρόμηση, αν και αποτελεί μια απλή μέθοδο, μπορεί να προσφέρει μια βασική εκτίμηση της απόδοσης.

4.6.2 Ridge Regression

Συχνά στο γενικό γραμμικό μοντέλο είναι πιθανό μια ή περισσότερες ανεξάρτητες μεταβλητές να είναι γραμμικά συσχετισμένες. Το γεγονός αυτό οδηγεί σε αυξημένα τυπικά σφάλματα και συνεπώς δυσκολεύει την επίδραση της εκτίμησης κάθε ανεξάρτητης μεταβλητής στην εξαρτημένη μεταβλητή. Έτσι, δεν είναι δυνατόν να εντοπιστούν οι στατιστικά σημαντικές μεταβλητές. Σε τέτοιες περιπτώσεις, η ανάλυση παλινδρόμησης μπορεί να πραγματοποιηθεί αφού αφαιρεθεί μια μεταβλητή από το γραμμικά εξαρτημένο σύνολο. Με λίγα λόγια πραγματοποιείται συρρίκνωση (shrinkage) του μοντέλου. Η

παλινδρόμηση κορυφογραμμής (*Ridge Regression*), χρησιμοποιεί την ίδια λογική με τη μέθοδο των ελαχίστων τετραγώνων, εισάγει όμως και μια *παράμετρο ομαλοποίησης alpha* (*regularization parameter*). Η παράμετρος αυτή έχει θετική τιμή και ο ρόλος της είναι να μειώνει το πλάτος των παραμέτρων παλινδρόμησης b_i του γραμμικού μοντέλου και άρα την επίδραση των «άσχετων» χαρακτηριστικών, με αποτέλεσμα να μειώνει έτσι τη μεταβλητότητα (variance) των εκτιμήσεων και τον κίνδυνο του overfitting. Ο στόχος της μεθόδου είναι η ελαχιστοποίηση της συνάρτησης απωλειών:

$$L(X, y) = \sum_{i=1}^{n} (b \cdot X_i - y_i)^2 + alpha \cdot \sum_{j=1}^{p} {b_j}^2$$

Όταν η παράμετρος *alpha* είναι μηδέν, η *Ridge Regression* ισοδυναμεί με την απλή *Linear Regression*, ενώ όσο αυξάνεται η τιμή της παραμέτρου, τόσο περισσότερο μειώνονται τα πλάτη των παραμέτρων b_i. Η τιμή της παραμέτρου *alpha* πρέπει να επιλεχθεί προσεκτικά, καθώς υπερβολικά μεγάλες τιμές της, το μοντέλο χάνει την ικανότητα να πραγματοποιεί αξιόπιστες προβλέψεις και οδηγείται στο φαινόμενο του underfitting. Η παραπάνω τεχνική κανονικοποίησης που χρησιμοποιεί η *Ridge Regression* ονομάζεται *L2 κανονικοποίηση* (*L2 regularization*).

4.6.3 Support Vector Regression

Οι *μηχανές διανυσμάτων υποστήριξης* (*support vector machines ή SVM*), είναι ένας αλγόριθμος επιτηρούμενης μάθησης που αρχικά χρησιμοποιήθηκε για προβλήματα ταξινόμησης (classification). Η *παλινδρόμηση διανυσμάτων υποστήριξης* (*support vector regression* ή *SVR*), αποτελεί μια επέκταση του αλγορίθμου αυτού, για χρήση του σε προβλήματα παλινδρόμησης. Ένα κύριο χαρακτηριστικό του αλγορίθμου *SVR*, είναι ότι αντί της ελαχιστοποίησης του σφάλματος της εκπαίδευσης που παρατηρήθηκε, προσπαθεί να ελαχιστοποιήσει ένα γενικευμένο όριο λάθους έτσι ώστε να επιτευχθεί μία πιο γενικευμένη απόδοση. Αυτό το γενικευμένο όριο σφάλματος είναι ο συνδυασμός του σφάλματος εκπαίδευσης και ενός όρου κανονικοποίησης ο οποίος ελέγχει την πολυπλοκότητα του χώρου υπόθεσης. Δεδομένου, δηλαδή, ενός συνόλου δεδομένων εκπαίδευσης $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset X \times R$, είναι να βρεθεί μια συνάρτηση $f(x)$, η οποία να έχει το πολύ ε απόκλιση από τους παρατηρούμενους στόχους y_i για όλα τα δεδομένα εκπαίδευσης και την ίδια στιγμή να είναι όσο το δυνατόν πιο επίπεδη. Αναλυτικότερα, η είσοδος x αρχικά χαρτογραφείται επάνω σε ένα m-διάστατο χώρο χαρακτηριστικών, χρησιμοποιώντας μια σταθερή μη γραμμική χαρτογράφηση και στην συνέχεια ένα γραμμικό μοντέλο κατασκευάζεται σε αυτό το χώρο.

Το γραμμικό μοντέλο αυτό περιγράφεται ως εξής:

$$f(x) = \langle w, x \rangle + b$$

με $w \epsilon X$ και $b \epsilon R$, όπου $\langle \cdot, \cdot \rangle$ δηλώνει το εσωτερικό γινόμενο στο X. Η μείωση της πολυπλοκότητας του μοντέλου, πραγματοποιείται ελαχιστοποιώντας το $\|w\|^2 = \langle w, w \rangle$. Προκύπτει, λοιπόν, το εξής πρόβλημα ελαχιστοποίησης:

$$minimize \frac{1}{2} \|w\|^2$$

$$\text{Υπό τη συνθήκη: } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases}$$

Σύμφωνα με τις παραπάνω σχέσεις, τα ζεύγη (x_i, y_i) προσεγγίζονται με ακρίβεια ε. Επειδή όμως, υπάρχει χρήσιμη πληροφορία και εκτός του ε, χρησιμοποιούνται οι μη αρνητικές μεταβλητές ξ_i, ξ_i^* με $i = 1,2, \dots, n$, προκειμένου να μετρηθεί η απόκλιση των δειγμάτων εκπαίδευσης εκτός της ζώνης ε που δημιουργείται εκατέρωθεν των (x_i, y_i). Ειδικότερα, το ξ_i περιγράφει τα δείγματα πάνω από τη ζώνη και το ξ_i^* περιγράφει τα δείγματα κάτω από τη ζώνη. Έτσι το πρόβλημα ελαχιστοποίησης διαμορφώνεται τελικά ως εξής:

$$minimize \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

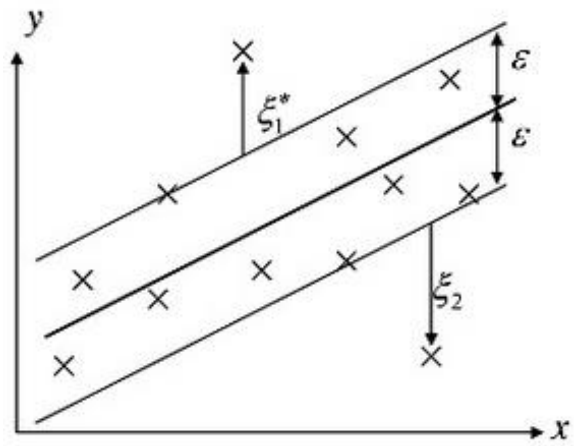
$$\text{Υπό τη συνθήκη: } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, i = 1,2, \dots, n \end{cases}$$

όπου $C > 0$ είναι μια σταθερά που καθορίζει τη σχέση μεταξύ της ομαλότητας της f και του ποσοστού μέχρι το οποίο είναι αποδεκτές αποκλίσεις μεγαλύτερες από ε. Το πρόβλημα αυτό μπορεί να επιλυθεί με χρήση τεχνικών των πολλαπλασιαστών Lagrange, σύμφωνα με τις οποίες προκύπτει ότι:

$$f(x) = \sum_{i=1}^{n} (a_i - a_i^*) K(x_i, x) + b$$

όπου, a_i και a_i^* είναι οι πολλαπλασιαστές Lagrange και $K(x_i, x)$ είναι η συνάρτηση πυρήνα. Η πιο δημοφιλής συνάρτηση πυρήνα είναι η radial-basis function (RBF) που είναι της μορφής:

$$K(x_i, x) = exp(-\gamma \|x_i - x\|^2)$$



Εικόνα 20: Σχηματική απεικόνιση των βασικότερων εννοιών του μοντέλου SVR

4.6.4 AdaBoost Regression

Ο αλγόριθμος *AdaBoost* ανήκει στην κατηγορία των μεθόδων ενίσχυσης. Με τον όρο ενίσχυση (boosting), εννοείται μια οικογένεια αλγορίθμων συνδυασμού μοντέλων, οι οποίες αποσκοπούν στη μετατροπή αδύναμων μοντέλων με χαμηλή απόδοση, σε ισχυρά μοντέλα πρόβλεψης. Οι αλγόριθμοι αυτοί βρίσκουν εφαρμογή τόσο σε προβλήματα ταξινόμησης, όσο και παλινδρόμησης. Η βασική ιδέα του αλγορίθμου *AdaBoost*, είναι η κατασκευή μιας ακολουθίας αδύναμων μοντέλων, με τροποποιημένες κάθε φορά εκδοχές των δεδομένων εκπαίδευσης. Έπειτα, οι προβλέψεις από όλα τα μοντέλα αυτά, συνδυάζονται με κάποια βαρύτητα για να δώσουν την τελική πρόβλεψη. Ο αλγόριθμος ξεκινά εκπαιδεύοντας ένα μοντέλο με το αρχικό σύνολο εκπαίδευσης. Στη συνέχεια εφαρμόζει συντελεστές βαρύτητας $w_1, w_2, \dots, w_n$ σε κάθε ένα από τα n δείγματα εκπαίδευσης και εκπαιδεύει εκ νέου το προηγούμενο μοντέλο με το νέο τροποποιημένο σύνολο εκπαίδευσης. Ο συντελεστής βαρύτητας του κάθε δείγματος επιλέγεται με τέτοιο τρόπο ώστε τα δείγματα που στο προηγούμενο βήμα δεν προβλέφθηκαν σωστά να έχουν αυξημένη βαρύτητα, ενώ για τα δείγματα που η πρόβλεψη έγινε σωστά ο συντελεστής βαρύτητάς τους μειώνεται. Η διαδικασία αυτή επαναλαμβάνεται και καθώς οι επαναλήψεις αυξάνονται, τα δείγματα εκείνα που είναι πιο δύσκολο να προβλεφθούν αποκτούν όλο και μεγαλύτερη επίδραση. Επομένως, κάθε διαδοχικό μοντέλο που εκπαιδεύεται αναγκάζεται να επικεντρωθεί σε εκείνα τα δείγματα που δεν προβλέφθηκαν σωστά από τα προηγούμενα μοντέλα. Έτσι, ο συνδυασμός αυτός των αδύναμων μοντέλων, οδηγεί τελικά σε ένα ισχυρό μοντέλο, το οποίο μπορεί να παράγει ακριβείς προβλέψεις.

4.6.5 XGBoost Regression

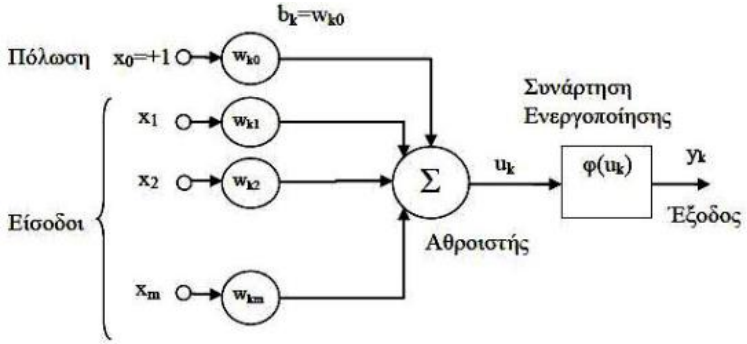
Η μέθοδος *XGBoost* είναι η συντομογραφία του Extreme Gradient Boosting και αποτελεί επίσης έναν αλγόριθμο ενίσχυσης. Η μέθοδος *Gradient Boosting,* αποτελεί μια ισχυρή τεχνική συνδυασμού αδύναμων μοντέλων, με στόχο τη δημιουργία ενός ισχυρότερου. Αρχικά, εκπαιδεύεται ένα μοντέλο σε ένα υποσύνολο των δεδομένων εκπαίδευσης. Χρησιμοποιώντας αυτό το μοντέλο, γίνονται προβλέψεις σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης και υπολογίζεται το σφάλμα του. Αυτές οι προβλέψεις είναι ανεπαρκείς και το αδύναμο μοντέλο θα πρέπει να ενισχυθεί σε μεταγενέστερες επαναλήψεις. Αυτός είναι ο λόγος για τον οποίον δημιουργείται ένα νέο μοντέλο, λαμβάνοντας υπόψη, τα σφάλματα που υπολογίστηκαν πριν, προκειμένου να διορθώσει τα λάθη του πρώτου. Τέλος, οι προβλέψεις του νέου μοντέλου συνδυάζονται με του προηγούμενου. Η μεγάλη διαφορά της μεθόδου αυτής από τη μέθοδο AdaBoost, έγκειται στο γεγονός, πως αντί να αλλάξει τη βαρύτητα των δειγμάτων που δεν προβλέφθηκαν σωστά, προσπαθεί να εκπαιδεύσει κάθε νέο μοντέλο αξιοποιώντας τα υπολειπόμενα σφάλματα του προκατόχου του. Η μέθοδος *XGBoost* είναι μια από τις ταχύτερες υλοποιήσεις του gradient boosting, ιδιαίτερα χρησιμοποιώντας ως τα αδύναμα μοντέλα είναι τα *δέντρα ταξινόμησης και παλινδρόμησης* (*CART* ή *Classification And Regression Trees*). Δεδομένου του συνόλου $F = \{f_1, f_2, \dots, f_m\}$ από m το πλήθος αδύναμων αλγορίθμων μάθησης, η τελικές προβλέψεις του μοντέλου *XGBoost* δίνονται από τη σχέση:

$$\widehat{y_i} = \sum_{t=1}^{m} f_t(x_i), f_t \in F$$

4.6.6 Artificial Neural Network

Κατ' αναλογία με ένα δίκτυο νευρώνων εγκεφάλου, ένα *τεχνητό νευρωνικό δίκτυο* (*Artificial Neural Network*) αποτελείται από ένα σύνολο τεχνητών νευρώνων που αλληλεπιδρούν, συνδεόμενοι μεταξύ τους με τις λεγόμενες συνάψεις (synapses). Ο βαθμός αλληλεπίδρασης είναι διαφορετικός για κάθε ζεύγος νευρώνων και καθορίζεται από τα λεγόμενα συναπτικά βάρη (synaptic weights). Συγκεκριμένα, καθώς το νευρωνικό δίκτυο αλληλεπιδρά με το περιβάλλον και μαθαίνει από αυτό, τα συναπτικά βάρη μεταβάλλονται συνεχώς, ενδυναμώνοντας ή αποδυναμώνοντας την ισχύ του κάθε δεσμού. Όλη η εμπειρική γνώση που αποκτά επομένως το νευρωνικό δίκτυο από το περιβάλλον κωδικοποιείται στα συναπτικά βάρη. Τα τελευταία χρόνια το ενδιαφέρον για τα νευρωνικά δίκτυα αυξάνεται συνεχώς καθώς εφαρμόζονται με μεγάλη επιτυχία σε ένα πολύ μεγάλο φάσμα τομέων της επιστήμης και της τεχνολογίας. Στην πραγματικότητα, τα νευρωνικά δίκτυα εισάγονται οπουδήποτε τίθεται θέμα πρόβλεψης, ταξινόμησης ή ελέγχου. Σε αναλογία με το βιολογικό

νευρώνα του εγκεφάλου, ο τεχνητός νευρώνας (artificial neuron) είναι η δομική μονάδα του τεχνητού νευρωνικού δικτύου. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη.



Εικόνα 21: Σχηματική αναπαράσταση μη γραμμικού νευρώνα

Στο νευρώνα αυτό, η πληροφορία ρέει πάντα προς μία κατεύθυνση, από αριστερά προς τα δεξιά, χωρίς να υπάρχει δηλαδή κανένας βρόχος ανάδρασης. Οι μαθηματικές εξισώσεις που περιγράφουν τη λειτουργία του νευρώνα είναι οι εξής:

$$u_k = \sum_{j=0}^{m} w_{kj} x_j$$

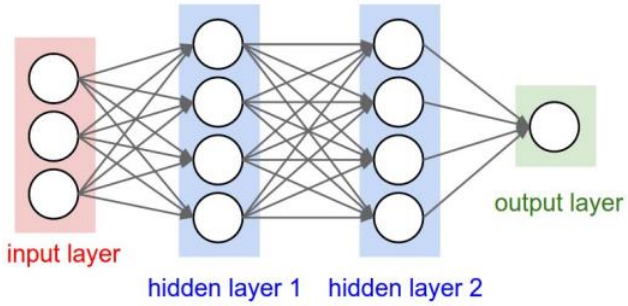
$$y_k = \varphi(u_k)$$

Σημαντικό ρόλο στην εκπαίδευση των τεχνητών νευρωνικών δικτύων έχουν οι συναρτήσεις ενεργοποίησης $\varphi(\cdot)$, οι οποίες θα πρέπει να είναι παραγωγίσιμες και άρα συνεχείς. Οι βασικότερες είναι:

- Η βηματική συνάρτηση (step function): $f(x) = \{0, x < 0\ 1, x \geq 0$
- Η γραμμική συνάρτηση: $f(x) = ax$
- Η υπερβολική εφαπτομένη: $f(x) = tanhx$
- Η λογιστική σιγμοειδής συνάρτηση: $f(x) = \frac{1}{1+e^{-x}}$
- Η διπολική σιγμοειδής συνάρτηση: $f(x) = \frac{1+e^{-x}}{1+e^{x}}$

Τα νευρωνικά δίκτυα χαρακτηρίζονται επίσης από την αρχιτεκτονική τους, η οποία καθορίζει τη διάταξη των συνδέσεων των νευρώνων καθώς και τον αριθμό και τον τύπο τους. Διακρίνονται σε δίκτυα με ανατροφοδότηση (recurrent networks), τα οποία περιέχουν

συνδέσεις με ανατροφοδότηση και σε δίκτυα προσοτροφοδότησης (feed forward networks), στα οποία η ροή των δεδομένων είναι αυστηρά από την είσοδο προς την έξοδο. Στην παρούσα εργασία σχεδιάστηκε τεχνητό νευρωνικό δίκτυο προσοτροφοδότησης. Η επεξεργασία των δεδομένων μπορεί να επεκτείνεται σε πολλές μονάδες, οι οποίες είναι οργανωμένες σε επίπεδα. Ως επίπεδο, ορίζεται το στρώμα στο οποίο γίνεται κάποιος υπολογισμός. Στη γενική περίπτωση ένα τεχνητό νευρωνικό δίκτυο περιέχει ένα ή περισσότερα κρυφά επίπεδα (hidden layers) των οποίων οι κόμβοι ονομάζονται κρυμμένοι νευρώνες.



Εικόνα 22: Τεχνητό νευρωνικό δίκτυο προσοτροφοδότησης δύο κρυμμένων επιπέδων

Η εκπαίδευση ενός νευρωνικού δικτύου γίνεται με χρήση παραδειγμάτων εκπαίδευσης και ενός αλγορίθμου εκπαίδευσης. Ο αλγόριθμος εκπαίδευσης, αποτελεί μια επαναληπτική διαδικασία, κατά την οποία αλλάζουν οι ελεύθεροι παράμετροι του δικτύου έτσι ώστε να μειωθεί το σφάλμα μεταξύ επιθυμητής και πραγματικής εξόδου του δικτύου. Ο αλγόριθμος εκπαίδευσης που χρησιμοποιείται συνηθέστερα σε πολυεπίπεδα νευρωνικά δίκτυα είναι ο αλγόριθμος *οπισθοδιάδοσης του σφάλματος* (*back propagation*). Στον αλγόριθμο αυτό, η μεταβολή των βαρών βασίζεται στον υπολογισμό της συνεισφοράς κάθε βάρους στο συνολικό σφάλμα. Τα τεχνητά νευρωνικά δίκτυα έχουν πετύχει αρκετά εντυπωσιακά αποτελέσματα, ενώ υπάρχουν πολλά μοντέλα δικτύων με διαφορετική φιλοσοφία και τρόπο λειτουργίας.

5 *Πειραματικά αποτελέσματα*

5.1 Γενικά

Στο κεφάλαιο αυτό περιγράφεται η υλοποίηση των μοντέλων πρόβλεψης, παρουσιάζονται τα αποτελέσματά εκτέλεσής τους όσον αφορά την ποιότητα της πρόβλεψης, καθώς και τα συμπεράσματα που προέκυψαν από την αξιολόγησή τους.

Όπως αναφέρθηκε στο *Κεφάλαιο 4*, πραγματοποιήθηκαν τρεις διαφορετικοί τρόποι διαχωρισμού (split) του συνόλου δεδομένων σε σύνολο εκπαίδευσης και δοκιμής, οι οποίοι αποτέλεσαν τρία ξεχωριστά πειράματα. Ο πρώτος τρόπος ήταν ο διαχωρισμός τους σύμφωνα με τη μεθοδολογία της εργασίας [KV2019]. Τα μοντέλα εκπαιδεύτηκαν με το πρώτο 70% του συνόλου δεδομένων και παρήγαγαν ωριαίες προβλέψεις κατανάλωσης φυσικού αερίου για το τελευταίο 30%. Στη συνέχεια, οι ωριαίες αυτές προβλέψεις αθροίστηκαν κατάλληλα, δίνοντας την ημερήσια πρόβλεψη κατανάλωσης. Επιπλέον, εξετάστηκε η απόδοση των μοντέλων στην εύρεση της μέγιστης ωριαίας κατανάλωσης (peak) που παρουσιάστηκε κατά τη διάρκεια της ημέρας.

Στο δεύτερο πείραμα, τα σύνολα εκπαίδευσης και δοκιμής περιείχαν δεδομένα από όλο το έτος. Έτσι, τα μοντέλα εκπαιδεύτηκαν με δεδομένα όλου του έτους και παρήγαγαν ωριαίες προβλέψεις κατανάλωσης.

Στο τρίτο πείραμα, τα σύνολα εκπαίδευσης και δοκιμής περιείχαν δεδομένα από όλο το έτος, η δειγματοληψία όμως έγινε σύμφωνα με την ημέρα αντί για την ώρα. Έτσι, τόσο το σύνολο εκπαίδευσης, όσο και το δοκιμής περιείχαν ωριαίες μετρήσεις ενιαίων ημερών, αντί για τυχαίες ωριαίες μετρήσεις από όλο το έτος. Με τον τρόπο αυτό, ήταν δυνατή η παραγωγή ωριαίων προβλέψεων κατανάλωσης, οι οποίες στη συνέχεια θα αθροιστούν και θα δώσουν την αντίστοιχη ημερήσια πρόβλεψη κατανάλωσης φυσικού αερίου. Επιπλέον, εξετάστηκε η

απόδοση των μοντέλων στον υπολογισμό εύρεση της μέγιστης ωριαίας κατανάλωσης που παρουσιάστηκε κατά τη διάρκεια της ημέρας.

Όλοι οι υπολογισμοί έγιναν με χρήση της γλώσσας προγραμματισμού Python και των βιβλιοθηκών Pandas, NumPy. Τα μοντέλα πρόβλεψης που δοκιμάστηκαν, υλοποιήθηκαν με τη βοήθεια των βιβλιοθηκών Scikit-Learn, xgboost, Keras και Tensorflow, ενώ η γραφική απεικόνιση των αποτελεσμάτων και τα διάφορα διαγράμματα έγιναν με χρήση των βιβλιοθηκών Matplotlib και Seaborn.

5.2 Παραμετροποίηση μοντέλων

Η εύρεση των βέλτιστων υπερπαραμέτρων των μοντέλων έγινε χρησιμοποιώντας τις τεχνικές Grid Search, Randomized Search και Cross-Validation, όπως περιγράφηκε στην *Ενότητα 4.4.2*. Η αξιολόγηση (scoring) των μοντέλων από τους αλγορίθμους και η επιλογή αυτών με τον καλύτερο συνδυασμό υπερπαραμέτρων, έγινε σύμφωνα με το μέσο τετραγωνικό σφάλμα (MSE).

Στις τρεις επόμενες υποενότητες, παρουσιάζονται οι τιμές των βέλτιστων υπερπαραμέτρων των μοντέλων, όπως προέκυψαν για κάθε ένα από τα τρία πειράματα που πραγματοποιήθηκαν.

5.2.1 Διαχωρισμός σύμφωνα με τη σχετική εργασία (Πείραμα 1)

Στον *Πίνακα 25*, παρουσιάζονται οι βέλτιστες υπερπαράμετροι των μοντέλων παλινδρόμησης που δοκιμάστηκαν, όπως προέκυψαν μετά τη διαδικασία τη ρύθμισής τους. Εξαιρείται η μέθοδος Linear regression, η οποία δε διαθέτει υπερπαραμέτρους προς ρύθμιση.

Μοντέλο	**Υπερπαράμετρος**	**Τιμή**
Ridge Regression	*alpha*	10
SVR	*kernel*	rbf
	C	3.39
	epsilon	0.13
	gamma	0.08
SVR	*kernel*	Linear
	C	0.02
	epsilon	0.02
	fit_intercept	True
AdaBoost	*learning_rate*	0.19
	n_estimators	50
	loss	exponential

XGBoost	*booster*	gbtree
	learning_rate	0.05
	gamma	0.18
	lambda	0.27
	colsample_bylevel	0.3
	colsample_bytree	0.9
	max_depth	31
	min_child_weight	9
	subsample	0.9
	n_estimators	155

Πίνακας 25: Υπερπαράμετροι μοντέλων παλινδρόμησης

Πέρα από τα παραπάνω μοντέλα παλινδρόμησης, σχεδιάστηκε και αξιολογήθηκε στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου ένα *προσοτροφοδοτούμενο τεχνητό νευρωνικό δίκτυο* (*ANN*), το οποίο εκπαιδεύτηκε με τον αλγόριθμο *Adam*, με εκπαίδευση ανά εποχή (epoch). Το ANN που σχεδιάστηκε περιείχε 2 κρυμμένα επίπεδα με 150 νευρώνες στο κάθε ένα. Η συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε σε όλα τα επίπεδα, εκτός του επιπέδου εξόδου ήταν η *Rectified Linear Unit* (*ReLU*), η οποία ορίζεται σύμφωνα με τη σχέση:

$$ReLU(z) = max(0, z)$$

Η αρχικοποίηση των βαρών κάθε επιπέδου έγινε με δειγματοληψία από την *περικομμένη κανονική κατανομή* (*truncated normal distribution*), η οποία έχει συνάρτηση πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής που ακολουθεί την κανονική κατανομή, αλλά οι τιμές της είναι φραγμένες. Η συνάρτηση κόστους ήταν κι εδώ το μέσο τετραγωνικό σφάλμα (MSE).

Συνοπτικά, οι βασικότεροι παράμετροι του νευρωνικού δικτύου που σχεδιάστηκε αποτυπώνονται στον επόμενο πίνακα:

Παράμετρος	**Τιμή**
Συνάρτηση ενεργοποίησης (activation function)	ReLU
Συνάρτηση απωλειών (loss function)	MSE
Αλγόριθμος μάθησης (optimizer)	Adam
Ρυθμός εκπαίδευσης (learning rate)	10^{-5}
Αριθμός εποχών (epochs)	500
Μέγεθος παρτίδας (batch size)	120
Αριθμός κρυμμένων επιπέδων (hidden layers)	2
Αριθμός νευρώνων κρυμμένων επιπέδων	150

Πίνακας 26: Παράμετροι νευρωνικού δικτύου

5.2.2 Διαχωρισμός με τυχαία δειγματοληψία ωριαίων καταναλώσεων (Πείραμα 2)

Στον *Πίνακα 27*, παρουσιάζονται οι βέλτιστες υπερπαράμετροι των μοντέλων παλινδρόμησης που δοκιμάστηκαν, όπως προέκυψαν μετά τη διαδικασία τη ρύθμισής τους.

Μοντέλο	**Υπερπαράμετρος**	**Τιμή**
Ridge Regression	*alpha*	25
SVR	*kernel*	rbf
	C	3.23
	epsilon	0.17
	gamma	0.07
AdaBoost	*learning_rate*	0.16
	n_estimators	66
	loss	exponential
XGBoost	*booster*	gbtree
	learning_rate	0.02
	gamma	0.94
	lambda	0.21
	colsample_bylevel	0.4
	colsample_bytree	0.9
	max_depth	36
	min_child_weight	11
	subsample	0.3
	n_estimators	295

Πίνακας 27: Υπερπαράμετροι μοντέλων παλινδρόμησης

Επιπλέον, ένα *τεχνητό νευρωνικό δίκτυο* (*ANN*) ίδιας αρχιτεκτονικής με αυτό της προηγούμενης υποενότητας εκπαιδεύτηκε και αξιολογήθηκε στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης. Συνοπτικά, οι βασικότεροι παράμετροι του νευρωνικού δικτύου που σχεδιάστηκε αποτυπώνονται στον επόμενο πίνακα:

Παράμετρος	**Τιμή**
Συνάρτηση ενεργοποίησης (activation function)	ReLU
Συνάρτηση απωλειών (loss function)	MSE
Αλγόριθμος μάθησης (optimizer)	Adam
Ρυθμός εκπαίδευσης (learning rate)	10^{-5}
Αριθμός εποχών (epochs)	500
Μέγεθος παρτίδας (batch size)	120
Αριθμός κρυμμένων επιπέδων (hidden layers)	2
Αριθμός νευρώνων κρυμμένων επιπέδων	150

Πίνακας 28: Παράμετροι νευρωνικού δικτύου

5.2.3 Διαχωρισμός με τυχαία δειγματοληψία ωριαίων καταναλώσεων ενιαίων ημερών (Πείραμα 3)

Στον *Πίνακα 29*, παρουσιάζονται οι βέλτιστες υπερπαράμετροι των μοντέλων παλινδρόμησης που δοκιμάστηκαν, όπως προέκυψαν μετά τη διαδικασία τη ρύθμισής τους.

Μοντέλο	**Υπερπαράμετρος**	**Τιμή**
Ridge Regression	*alpha*	25
SVR	*kernel*	rbf
	C	3.22
	epsilon	0.13
	gamma	0.04
SVR	*kernel*	Linear
	C	0.02
	epsilon	0.02
	fit_intercept	False
AdaBoost	*learning_rate*	0.31
	n_estimators	14
	loss	exponential

XGBoost	*booster*	gbtree
	learning_rate	0.02
	gamma	0.55
	lambda	0.01
	colsample_bylevel	0.3
	colsample_bytree	0.9
	max_depth	44
	min_child_weight	18
	subsample	0.7
	n_estimators	360

Πίνακας 29: Υπερπαράμετροι μοντέλων παλινδρόμησης

Επιπλέον, δύο *τεχνητά νευρωνικά δίκτυα* (*ANN*) σχεδιάστηκαν και αξιολογήθηκαν στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου. Ένα τεχνητό νευρωνικό ίδιας αρχιτεκτονικής με αυτό των προηγούμενων πειραμάτων, καθώς κι ένα νευρωνικό δίκτυο με ένα επιπλέον κρυμμένο επίπεδο 150 νευρώνων. Συνοπτικά, οι βασικότεροι παράμετροι των δύο νευρωνικών δικτύων που σχεδιάστηκαν αποτυπώνονται στον *Πίνακα 30*.

Παράμετρος	**Τιμή**
Συνάρτηση ενεργοποίησης (activation function)	ReLU
Συνάρτηση απωλειών (loss function)	MSE
Αλγόριθμος μάθησης (optimizer)	Adam
Ρυθμός εκπαίδευσης (learning rate)	10^{-5}
Αριθμός εποχών (epochs)	500
Μέγεθος παρτίδας (batch size)	120
Αριθμός κρυμμένων επιπέδων (hidden layers)	2 ή 3
Αριθμός νευρώνων κρυμμένων επιπέδων	150

Πίνακας 30: Παράμετροι των νευρωνικών δύο και τριών κρυμμένων επιπέδων

5.3 Αξιολόγηση αποτελεσμάτων

Σε κάθε ένα από τα τρία πειράματα που πραγματοποιήθηκαν, τα μοντέλα που σχεδιάστηκαν, εκπαιδεύτηκαν στη συνέχεια με τα δεδομένα του συνόλου εκπαίδευσης. Η ρύθμιση των υπερπαραμέτρων τους έγινε με χρήση της τεχνικής cross-validation, όπως έχει ήδη αναλυθεί. Στη συνέχεια, εξετάστηκε η απόδοσή τους στο σύνολο εκπαίδευσης (train set), καθώς και κατά τη διαδικασία του cross-validation, στο σύνολο αξιολόγησης (validation set). Η τελική αξιολόγηση και σύγκριση των μοντέλων έγινε σύμφωνα με την απόδοσή τους στο σύνολο δοκιμής (test set). Η αξιολόγηση των προβλέψεων έγινε χρησιμοποιώντας τους δείκτες που παρουσιάστηκαν στην *Ενότητα 4.2*.

5.3.1 Σύνοψη αποτελεσμάτων

Στην υποενότητα αυτή γίνεται η συνοπτική παρουσίαση των αποτελεσμάτων των τριών πειραμάτων που πραγματοποιήθηκαν. Σε κάθε μοντέλο καταγράφεται η απόδοσή του, σύμφωνα με την τιμή κάθε δείκτη αξιολόγησης, στο σύνολο εκπαίδευσης (train set), στο σύνολο αξιολόγησης (validation set), κατά τη διαδικασία του cross validation, καθώς και στο σύνολο δοκιμής (test set).

ΠΕΙΡΑΜΑ 1: ***Διαχωρισμός σύμφωνα με τη σχετική εργασία***

Στον *Πίνακα 31* παρουσιάζεται η απόδοση των μοντέλων σε ωριαία βάση προβλέψεων του πρώτου πειράματος που πραγματοποιήθηκε, χρησιμοποιώντας τη μεθοδολογία διαχωρισμού (split) των δεδομένων της εργασίας [KV2019]. Επίσης, στον *Πίνακα 33* παρουσιάζεται η απόδοση των μοντέλων στο πρόβλημα της ημερήσιας πρόβλεψης κατανάλωσης, ενώ στον *Πίνακα* 32 και *Πίνακα 34,* καταγράφονται οι αντίστοιχες αποδόσεις των μοντέλων της σχετικής εργασίας στο σύνολο δοκιμής (test set).

Στο πείραμα αυτό, τα μοντέλα που παρουσίασαν την υψηλότερη απόδοση στο σύνολο δοκιμής στην πρόβλεψη της ωριαίας κατανάλωσης ήταν τα *Linear Regression, Ridge Regression*, καθώς και η μέθοδος των *τεχνητών νευρωνικών δικτύων* (*ANN*). Παρατηρώντας τον *Πίνακα 31*, φαίνεται πως ενώ τα σύνθετα μοντέλα παλινδρόμησης SVR και XGBoost είχαν την καλύτερη απόδοση στο σύνολο αξιολόγησης, τα μοντέλα αυτά δε γενίκευσαν εξίσου καλά στο σύνολο δοκιμής. Η αισθητά αυτή μειωμένη απόδοση των μοντέλων στο σύνολο δοκιμής, συγκριτικά με την απόδοσή τους στο σύνολο αξιολόγησης, οφείλεται στο γεγονός της εισαγωγής μεροληψίας (bias), λόγω της μεθοδολογίας δειγματοληψίας των συνόλων εκπαίδευσης και δοκιμής που ακολούθησαν οι ερευνητές της εργασίας [KV2019].

	Linear Regression			Ridge Regression			SVR			LinearSVR			AdaBoost			XGBoost			ANN		
	Train	**Val**	**Test**	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**
R^2	0.8001	0.7952	0.6591	0.8000	0.7959	0.654	0.88	0.8506	0.6329	0.7999	0.7958	0.6270	0.8102	0.7873	0.6287	0.9538	0.8583	0.6330	0.8583	0.8186	0.6605
MAE	0.2331	0.2339	0.3969	0.2336	0.2336	0.3974	0.1729	0.1895	0.4043	0.2342	0.2348	0.4223	0.2411	0.2453	0.4275	0.1127	0.1806	0.4182	0.1820	0.1880	0.4117
MSE	0.1209	0.1226	0.4354	0.121	0.1221	0.442	0.0726	0.0899	0.4689	0.1210	0.1221	0.4765	0.1148	0.1270	0.4743	0.0280	0.0849	0.4688	0.0857	0.1026	0.4337
RMSE	0.3477	0.3495	0.6599	0.3478	0.3489	0.6648	0.2694	0.2989	0.6848	0.3479	0.3490	0.6903	0.3389	0.3561	0.6887	0.1672	0.2908	0.6847	0.2927	0.3204	0.6585

Πίνακας 31: Απόδοση ωριαίων προβλέψεων πρώτου πειράματος στα σύνολα εκπαίδευσης (train), αξιολόγησης (validation) και δοκιμής (test)

	Linear Regression	DNN	LSTM	GRU	CNN	CNN/RNN/DNN
	Test	**Test**	**Test**	**Test**	**Test**	**Test**
MSE	0.62	0.67	1	1.19	0.84	0.91
RMSE	0.7874	0.8185	1	1.09	0.9165	0.9539

Πίνακας 32: Απόδοση ωριαίων προβλέψεων μοντέλων σχετικής εργασίας στο σύνολο δοκιμής (test set)

	Linear Regression	Ridge Regression	SVR	Linear SVR	AdaBoost	XGBoost	ANN
	Test	Test	Test	Test	Test	Test	Test
R^2	0.9804	0.9700	0.8873	0.9112	0.9705	0.8784	0.9363
MAE	1.6239	2.053	4.4248	4.3108	1.895	4.8926	3.8061
MSE	6.2954	9.6314	36.1746	28.52	9.4845	39.0512	20.8818
RMSE	2.5091	3.1035	6.0145	5.3404	3.0797	6.2491	4.5697

Πίνακας 33: Απόδοση μοντέλων πρώτου πειράματος σε ημερήσια βάση προβλέψεων στο σύνολο δοκιμής (test set)

Στο πρόβλημα της ημερήσιας πρόβλεψης κατανάλωσης, οι μέθοδοι *Linear Regression, Ridge Regression, AdaBoost* καθώς και το *τεχνητό νευρωνικό δίκτυο* (*ANN*) παρουσίασαν την υψηλότερη απόδοση προβλέψεων. Η μέθοδος *Linear Regression*, μάλιστα, είχε την καλύτερη απόδοση όλων.

	Linear Regression	DNN	LSTM	GRU	CNN	CNN/RNN/DNN
	Test	Test	Test	Test	Test	Test
MSE	99	104	206	264	115	184
RMSE	9.9499	10.1980	14.3527	16.2480	10.7238	13.5647

Πίνακας 34: Απόδοση μοντέλων σχετικής εργασίας σε ημερήσια βάση προβλέψεων στο σύνολο δοκιμής (test set)

Στον *Πίνακα 35,* παρουσιάζεται η απόδοση των τριών καλύτερων μεθόδων του *Πειράματος 1* και της σχετικής εργασίας στο σύνολο δοκιμής (test set) στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης φυσικού αερίου.

	Διπλωματική εργασία			[KV2019]		
	Linear Regression	Ridge Regression	ANN	Linear Regression	DNN	CNN
	Test	Test	Test	Test	Test	Test
MSE	0.4354	0.442	0.4337	0.62	0.67	0.84
RMSE	0.6599	0.6648	0.6585	0.7874	0.8185	0.9165

Πίνακας 35: Σύγκριση καλύτερων μεθόδων πρόβλεψης *Πειράματος 1* και σχετικής εργασίας σε ωριαία βάση προβλέψεων στο σύνολο δοκιμής (test set)

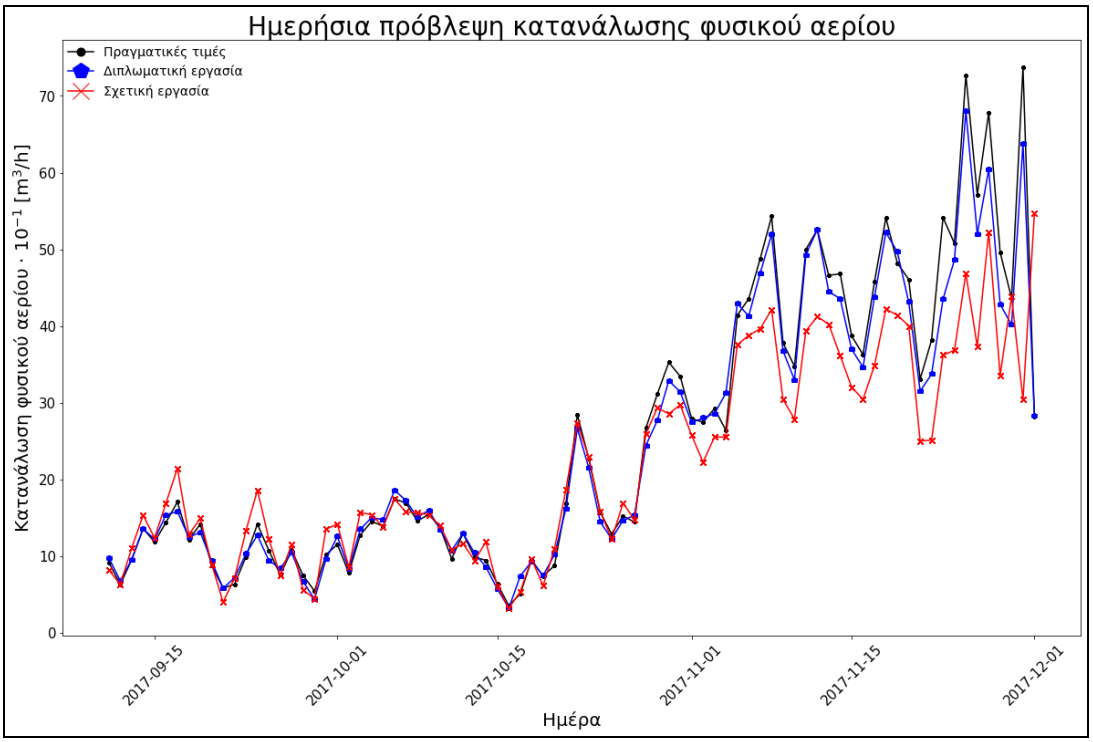
Αντίστοιχα, στον *Πίνακα 36,* παρουσιάζεται η απόδοση των τριών καλύτερων μεθόδων στο πρόβλημα της ημερήσιας πρόβλεψης κατανάλωσης φυσικού αερίου.

	Διπλωματική εργασία			**[KV2019]**		
	Linear Regression	**Ridge Regression**	**AdaBoost**	**Linear Regression**	**DNN**	**CNN**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**
MSE	6.2954	9.6314	9.4845	99	104	115
RMSE	2.5091	3.1035	3.0797	9.9499	10.1980	10.7238

Πίνακας 36: Σύγκριση καλύτερων μεθόδων πρόβλεψης *Πειράματος 1* και σχετικής εργασίας σε ημερήσια βάση προβλέψεων στο σύνολο δοκιμής (test set)

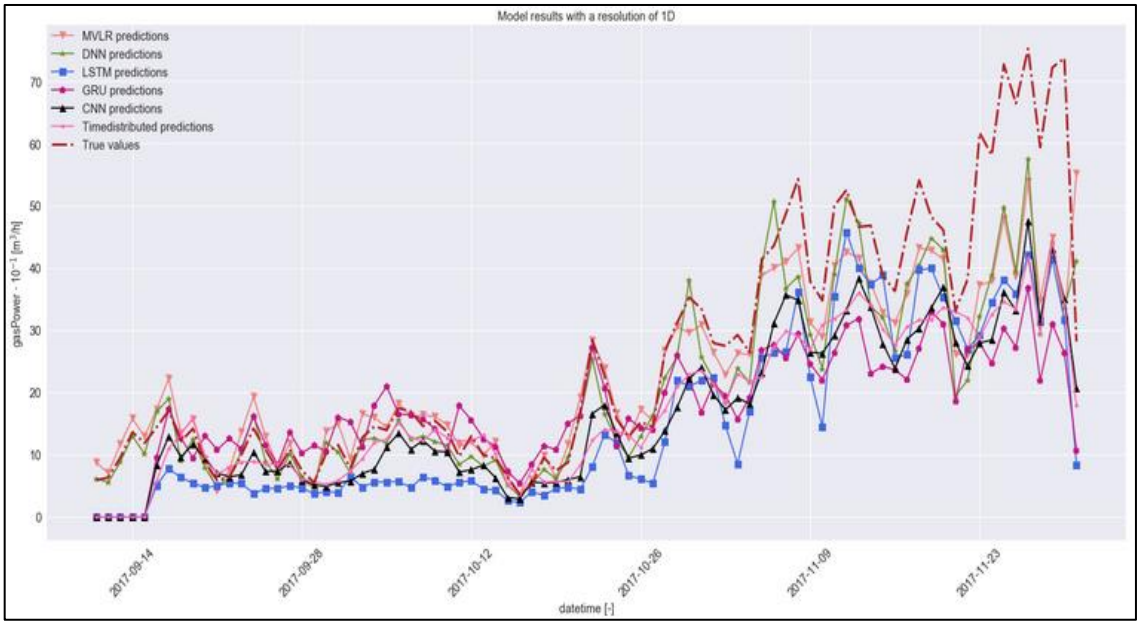
Συγκρίνοντας τους παραπάνω πίνακες, καθίσταται σαφές πως τα μοντέλα που υλοποιήθηκαν στην παρούσα διπλωματική εργασία ξεπέρασαν σημαντικά σε απόδοση αυτά της εργασίας [KV2019], τόσο στην παραγωγή ωριαίων, όσο και ημερησίων προβλέψεων κατανάλωσης.

Στα πλαίσια της διπλωματικής, παρήχθησαν προβλέψεις με τη μέθοδο Linear Regression, η οποία εκπαιδεύτηκε πρώτα χρησιμοποιώντας τα χαρακτηριστικά (features) της εργασίας [KV2019], επαληθεύοντας τα αποτελέσματα που παρουσιάζονται στη σχετική δημοσίευση. Σημειώνεται ότι η μέθοδος Linear Regression αναδείχθηκε ως η καλύτερη μέθοδος και των δύο εργασιών στην πρόβλεψη της ημερήσιας κατανάλωσης. Στην *Εικόνα 23,* παρουσιάζονται οι ημερήσιες προβλέψεις της μεθόδου Linear Regression των δύο εργασιών. Παρατηρώντας την εικόνα αυτή, γίνεται αισθητή η βελτίωση της απόδοσης της εν λόγω μεθόδου, όταν αυτή εκπαιδεύτηκε με τα χαρακτηριστικά που προέκυψαν από την ανάλυση της παρούσας διπλωματικής. Η βελτίωση αυτή γίνεται ιδιαίτερα εμφανής κατά τους χειμερινούς μήνες, όταν τα επίπεδα κατανάλωσης είναι υψηλότερα. Κατά τους μήνες αυτούς, τα σφάλματα της μεθόδου Linear Regression της εργασίας [KV2019] είναι αρκετά υψηλά, προβλέποντας μάλιστα συστηματικά αρκετά χαμηλότερες καταναλώσεις από τις πραγματικές. Αντίθετα, οι προβλέψεις που παρήχθησαν από την αντίστοιχη μέθοδο της παρούσας διπλωματικής, σημείωσαν σημαντικά υψηλότερη ακρίβεια τόσο κατά τη διάρκεια αυτών των περιόδων, όσο και των υπόλοιπων μηνών που περιλαμβάνονταν στο σύνολο δοκιμής.



Εικόνα 23: Σύγκριση Linear Regression των δύο εργασιών

Επιπλέον, στην *Εικόνα 24*, απεικονίζονται οι προβλέψεις σε ημερήσια βάση όλων των μοντέλων που σχεδιάστηκαν από τους ερευνητές στην εργασία [KV2019]. Όπως και στην περίπτωση της μεθόδου Linear Regression, κατά τους χειμερινούς μήνες του έτους τα σφάλματα όλων των μοντέλων της εργασίας [KV2019] είναι ιδιαίτερα υψηλά, προβλέποντας αρκετά χαμηλότερες καταναλώσεις από τις πραγματικές. Ακόμα, οι περισσότερες από τις μεθόδους πρόβλεψης της εργασίας αυτής, απέτυχαν να παράγουν ακριβείς προβλέψεις και κατά τη διάρκεια των υπόλοιπων μηνών του συνόλου δοκιμής.



Εικόνα 24: Ημερήσιες προβλέψεις μοντέλων εργασίας [KV2019]

Η διαφορά αυτή στην ακρίβεια των προβλέψεων των δύο εργασιών οφείλεται στην επιλογή των χαρακτηριστικών των μοντέλων, καθώς και στα μοντέλα πρόβλεψης που χρησιμοποιήθηκαν. Η εργασία [KV2019] περιορίστηκε στην επιλογή όσο το δυνατόν λιγότερων χαρακτηριστικών, παραλείποντας μια εκτενέστερη διαδικασία επιλογής και δημιουργίας χαρακτηριστικών. Αντίθετα, στην παρούσα διπλωματική, η επιλογή των χαρακτηριστικών αποτέλεσε μια ιδιαίτερα σημαντική διαδικασία, η οποία οδήγησε σε σαφώς αυξημένη απόδοση προβλέψεων. Επιπλέον, η επιλογή σύνθετων τεχνητών νευρωνικών δικτύων ως μοντέλα πρόβλεψης, όπως τα LSTM και CNN, δεν είχε τα αναμενόμενα αποτελέσματα. Αυτό οφείλεται στο γεγονός ότι σύνολο δεδομένων που χρησιμοποιήθηκε δεν περιέχει μεγάλο πλήθος εγγραφών, ενώ οι εγγραφές που περιλαμβάνει καλύπτουν μόνο ένα έτος με αποτέλεσμα τα σύνολα εκπαίδευσης και δοκιμής να χαρακτηρίζονται από διαφορετικές συνθήκες κατανάλωσης.

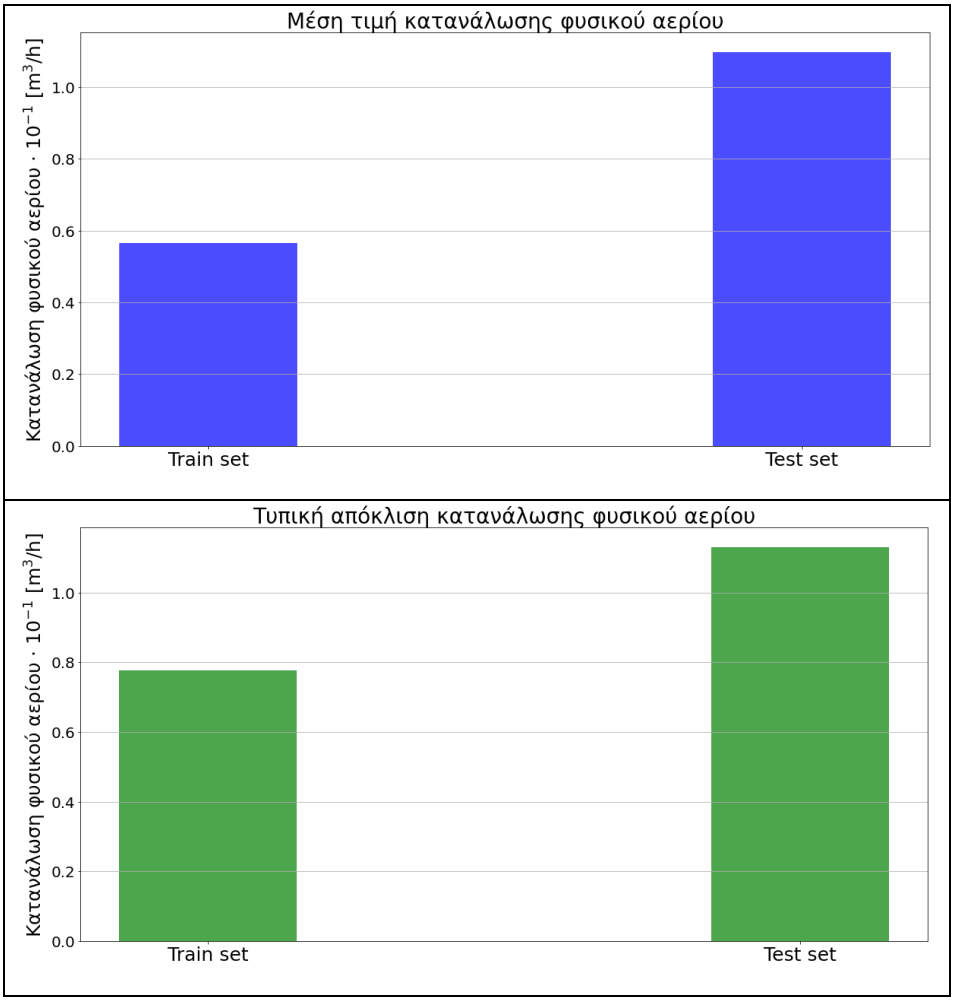
Στον *Πίνακα 37*, καταγράφεται η απόδοση των μοντέλων της διπλωματικής στο πρόβλημα του υπολογισμού της μέγιστης (peak) κατανάλωσης της ημέρας στο σύνολο δοκιμής.

	Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**
R^2	0.8614	0.8459	0.5446	0.7961	0.6551	0.4801	0.7708
MAE	0.4947	0.5224	0.9273	0.6246	0.6650	0.9568	0.7100
MSE	0.3911	0.4349	1.2852	0.5754	0.9733	1.1467	0.6467
RMSE	0.6254	0.6595	1.1337	0.7586	0.9866	1.2113	0.8042

Πίνακας 37: Απόδοση μοντέλων πρώτου πειράματος στον υπολογισμό της μέγιστης ωριαίας κατανάλωσης της ημέρας στο σύνολο δοκιμής (test set)

Οι γραμμικές μέθοδοι *Linear Regression* και *Ridge Regression* παρουσίασαν την υψηλότερη απόδοση και στο πρόβλημα του υπολογισμού της μέγιστης κατανάλωσης της ημέρας. Ακόμα, τα μοντέλο *LinearSVR*, δηλαδή η μέθοδος SVR με γραμμικό πυρήνα (kernel), καθώς και το *τεχνητό νευρωνικό δίκτυο* (*ANN*) είχαν επίσης σχετικά καλή απόδοση. Τα υπόλοιπα μοντέλα παρουσίασαν ιδιαίτερα μειωμένη απόδοση και υψηλά σφάλματα προβλέψεων.

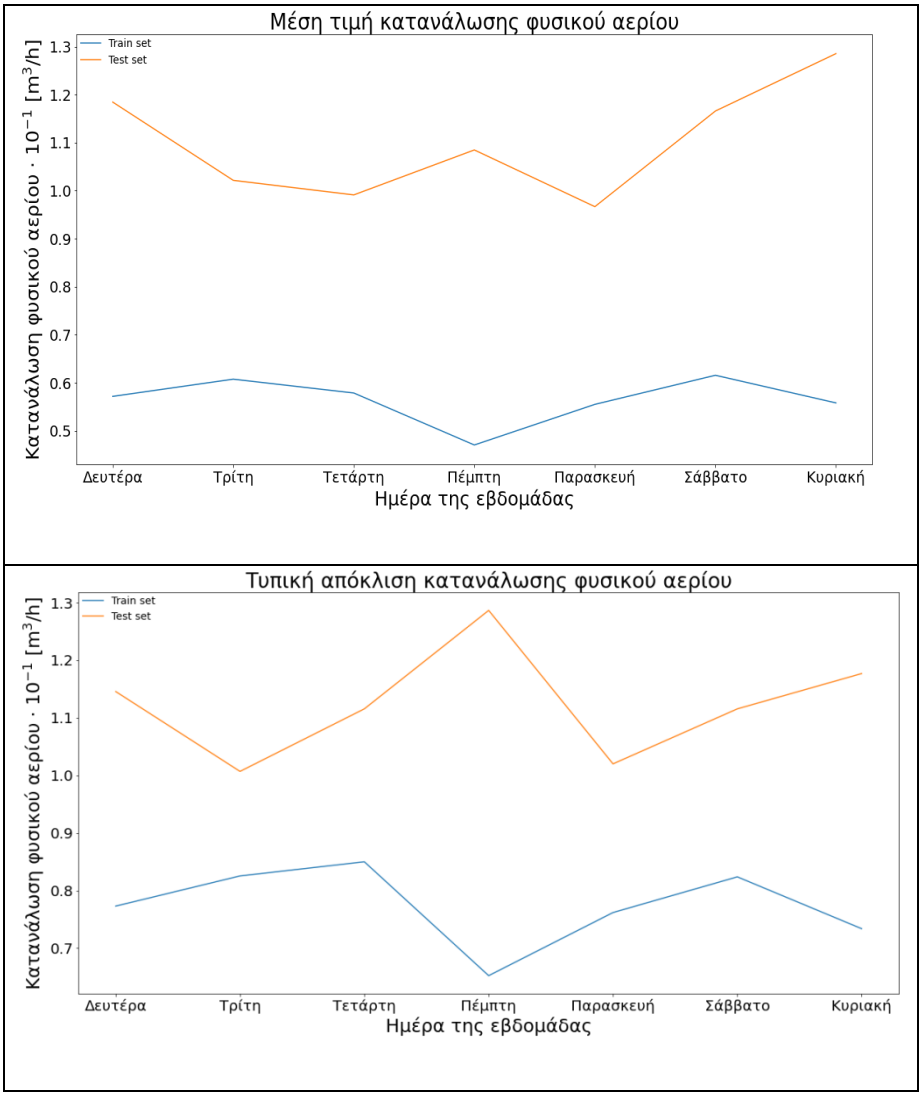
Είναι σημαντικό να αναφερθεί, ότι ο τρόπος διαχωρισμού του συνόλου των δεδομένων που ακολουθήθηκε από την εργασία [KV2019] εισάγει ένα βαθμό μεροληψίας (bias) στα μοντέλα πρόβλεψης. Αυτό συμβαίνει διότι, τα μοντέλα εκπαιδεύτηκαν με δεδομένα μόνο των πρώτων μηνών του έτους, ενώ κλήθηκαν να παράγουν προβλέψεις για τους επόμενους μήνες, οι οποίοι φυσικά χαρακτηρίζονται από διαφορετικές καιρικές και καταναλωτικές συνθήκες, καθώς και διαφορετικά επίπεδα κατανάλωσης. Τα παραπάνω γίνονται εμφανή υπολογίζοντας τη μέση τιμή και τυπική απόκλιση των συνόλων εκπαίδευσης (train set) και δοκιμής (test set) αντίστοιχα, όπως φαίνεται στην *Εικόνα 25.*



Εικόνα 25: Μέση τιμή και τυπική απόκλιση κατανάλωσης φυσικού αερίου στα σύνολα εκπαίδευσης και δοκιμής

Είναι σαφές ότι το σύνολο δοκιμής χαρακτηρίζεται από αρκετά υψηλότερη μέση τιμή κατανάλωσης φυσικού αερίου σε σχέση με το σύνολο εκπαίδευσης, καθώς επίσης και από μεγάλη τιμή τυπική απόκλισης, γεγονός που συνεπάγεται μεγαλύτερες διακυμάνσεις. Αυτό είναι αναμενόμενο, καθώς το σύνολο δοκιμής περιλαμβάνει χειμερινούς μήνες, κατά τους οποίους τα επίπεδα κατανάλωσης αναμένονται υψηλά. Στη συνέχεια παρουσιάζονται τα

διαγράμματα της μέσης τιμής και τυπικής απόκλισης της κατανάλωσης στα δύο σύνολα για κάθε ημέρα της εβδομάδας.



Εικόνα 26: Μέση τιμή και τυπική απόκλιση κατανάλωσης κάθε ημέρας τη εβδομάδας στα σύνολα εκπαίδευσης και δοκιμής

Το πρόβλημα της μεροληψίας (bias) που περιγράφηκε, επιλύεται αλλάζοντας των τρόπο διαχωρισμού των δεδομένων στα σύνολο εκπαίδευσης και δοκιμής, σύμφωνα με τους τρόπου που έχουν περιγραφεί στην *Υποενότητα 4.4.1*, όπως αποδεικνύεται από την ανάλυση που ακολουθεί.

<u>ΠΕΙΡΑΜΑ 2</u>: ***Διαχωρισμός με τυχαία δειγματοληψία ωριαίων καταναλώσεων***

Στο πείραμα αυτό, η δειγματοληψία των συνόλων εκπαίδευσης και δοκιμής ήταν τυχαία, περιλαμβάνοντας ωριαίες καταναλώσεις από όλο το σύνολο δεδομένων. Έτσι, τα μοντέλα εκπαιδεύτηκαν με δεδομένα όλου του έτους και παρήγαγαν ωριαίες προβλέψεις κατανάλωσης. Στον *Πίνακα 38,* παρουσιάζεται η απόδοση των μοντέλων που εξετάστηκαν.

Τα μοντέλα που παρουσίασαν την υψηλότερη απόδοση στο σύνολο δοκιμής (test set) ήταν τα *SVR, ANN* και *XGBoost*, με το τελευταίο να πετυχαίνει μάλιστα το μικρότερο σφάλμα σε όρους MAE και RMSE.

Στον *Πίνακα 39*, παρουσιάζεται η απόδοση των τριών καλύτερων μεθόδων του *Πειράματος 2* της παρούσας διπλωματικής, καθώς και των αντίστοιχων μοντέλων της εργασίας [KV2019] στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης φυσικού αερίου. Η σημαντική διαφορά του *Πειράματος 2* με το *Πείραμα 1* και τη σχετική εργασία [KV2019], έγκειται στη μεθοδολογία διαχωρισμού του συνόλου δεδομένων στα σύνολα εκπαίδευσης (train set) και δοκιμής (test set). Παρατηρώντας τον *Πίνακα 39*, γίνεται εμφανής η σημαντική βελτίωση της απόδοσης όλων των μοντέλων σε σχέση με τα μοντέλα της εργασίας [KV2019]. Το γεγονός αυτό οφείλεται ακριβώς, στην ορθότερη μεθοδολογία δειγματοληψίας των συνόλων εκπαίδευσης και δοκιμής. Τα μοντέλα παύουν να χαρακτηρίζονται από μεροληψία (bias) και προσαρμόζονται καλύτερα στα δεδομένα εκπαίδευσης, καθώς εκπαιδεύονται πλέον με δεδομένα όλων των μηνών του έτους. Αυτό γίνεται εμφανές με την αύξηση της τιμής του συντελεστή προσδιορισμού R^2 στο σύνολο δοκιμής, καθώς και με τη σημαντική μείωση των σφαλμάτων των προβλέψεων σε όρους MAE και RMSE σε σχέση με τις αντίστοιχες τιμές της εργασίας [KV2019].

	Linear Regression			Ridge Regression			SVR			AdaBoost			XGBoost			ANN		
	Train	**Val**	**Test**	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**
R^2	0.7695	0.7669	0.7309	0.7695	0.7672	0.7313	0.8340	0.7990	0.7592	0.7565	0.7310	0.7029	0.8658	0.8084	0.7865	0.8136	0.8101	0.7758
MAE	0.2830	0.2843	0.3028	0.2825	0.2835	0.3021	0.2272	0.2486	0.2605	0.3113	0.3172	0.3273	0.2098	0.2415	0.2497	0.2390	0.2298	0.2536
MSE	0.1985	0.2006	0.2358	0.1985	0.2003	0.2355	0.1430	0.1732	0.2111	0.2097	0.2296	0.2604	0.1156	0.1648	0.1871	0.1605	0.1459	0.1965
RMSE	0.4455	0.4466	0.4856	0.4456	0.4463	0.4852	0.3782	0.4144	0.4594	0.4579	0.4785	0.5103	0.3400	0.4044	0.4325	0.4007	0.3819	0.4432

Πίνακας 38: Απόδοση ωριαίων προβλέψεων δεύτερου πειράματος στα σύνολα εκπαίδευσης (train), αξιολόγησης (validation) και δοκιμής (test)

	Διπλωματική εργασία			[KV2019]		
	XGBoost	**ANN**	**SVR**	**Linear Regression**	**DNN**	**CNN**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**
MSE	0.1871	0.1965	0.2111	0.62	0.67	0.84
RMSE	0.4325	0.4432	0.4594	0.7874	0.8185	0.9165

Πίνακας 39: Σύγκριση καλύτερων μεθόδων *Πειράματος 2*, εκπαιδευμένων με δεδομένα όλου του έτους και σχετικής εργασίας, εκπαιδευμένων με δεδομένα μόνο των πρώτων μηνών του έτους, σε ωριαία βάση προβλέψεων

<u>ΠΕΙΡΑΜΑ 3</u>: ***Διαχωρισμός με τυχαία δειγματοληψία ωριαίων καταναλώσεων ενιαίων ημερών***

Στο πείραμα αυτό η δειγματοληψία των συνόλων εκπαίδευσης και δοκιμής ήταν τυχαία, περιλαμβάνοντας ωριαίες καταναλώσεις από όλο το σύνολο δεδομένων, αλλά ο διαχωρισμός έγινε σύμφωνα με την ημέρα της κατανάλωσης. Έτσι, εξασφαλίστηκε ότι όλες οι ώρες μιας ημέρας θα ανήκουν ενιαία είτε στο σύνολο εκπαίδευσης είτε στο σύνολο δοκιμής. Στον *Πίνακα 40* παρουσιάζεται η απόδοση των μοντέλων σε ωριαία βάση προβλέψεων του, ενώ στον *Πίνακα 41* παρουσιάζεται η απόδοση των μοντέλων στο πρόβλημα της ημερήσιας πρόβλεψης κατανάλωσης.

Τα μοντέλα που παρουσίασαν την υψηλότερη απόδοση στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης ήταν τα *SVR, ANN* και *XGBoost*, με το τελευταίο να πετυχαίνει μάλιστα το μικρότερο σφάλμα σε όρους MAE και RMSE.

Παρόμοια, ήταν η απόδοση των μοντέλων και στο πρόβλημα της ημερήσιας πρόβλεψης κατανάλωσης, όπου οι μέθοδοι *XGBoost* και *ANN* αποδείχθηκαν οι πλέον αποδοτικές, παράγοντας το μικρότερο σφάλμα πρόβλεψης σε όρους MAE και RMSE.

	Linear Regression			Ridge Regression			SVR			Linear SVR			AdaBoost			XGBoost			ANN (2 Hidden)			ANN (3 Hidden)		
	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test
R^2	0.7505	0.7446	0.7744	0.7504	0.7451	0.7743	0.7975	0.7738	0.8124	0.7502	0.7448	0.7746	0.7378	0.7092	0.7462	0.8933	0.7764	0.8313	0.7966	0.7768	0.8128	0.8094	0.7841	0.8181
MAE	0.2936	0.2954	0.2660	0.2931	0.2945	0.2861	0.2391	0.2534	0.2375	0.2931	0.2945	0.2857	0.3131	0.3149	0.3029	0.1855	0.2499	0.2284	0.2461	0.2522	0.2446	0.2370	0.2471	0.2388
MSE	0.2157	0.2191	0.1958	0.2157	0.2187	0.1958	0.1750	0.1949	0.1628	0.2159	0.2189	0.1956	0.2267	0.2486	0.2203	0.0923	0.1919	0.1464	0.1758	0.1893	0.1624	0.1647	0.1830	0.1579
RMSE	0.4644	0.4668	0.4425	0.4645	0.4664	0.4425	0.4183	0.4397	0.4035	0.4647	0.4666	0.4423	0.4761	0.4976	0.4693	0.3037	0.4370	0.3826	0.4193	0.4350	0.4030	0.4059	0.4278	0.3973

Πίνακας 40: Απόδοση ωριαίων προβλέψεων τρίτου πειράματος στα σύνολα εκπαίδευσης (train), αξιολόγησης (validation) και δοκιμής (test)

	Linear Regression	Ridge Regression	SVR	Linear SVR	AdaBoost	XGBoost	ANN 2 Hidden	ANN 3 Hidden
	Test	Test	Test	Test	Test	Test	Test	Test
R^2	0.9683	0.9682	0.9662	0.9687	0.9647	0.9726	0.9712	0.9725
MAE	1.9187	1.9265	1.7479	1.9141	1.9742	1.6127	1.7946	1.7170
MSE	8.2311	8.2478	8.7671	8.1279	9.1673	7.1144	7.4756	7.1280
RMSE	2.8690	2.8719	2.9609	2.8509	3.0278	2.6673	2.7342	2.6698

Πίνακας 41: Απόδοση μοντέλων τρίτου πειράματος σε ημερήσια βάση προβλέψεων στο σύνολο δοκιμής (test set)

Στον *Πίνακα 42,* παρουσιάζεται η απόδοση των τριών καλύτερων μεθόδων του *Πειράματος 3* και της εργασίας [KV2019] στο σύνολο δοκιμής (test set) στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης φυσικού αερίου.

	Διπλωματική εργασία			**[KV2019]**		
	XGBoost	**ANN 3 Hidden**	**SVR**	**Linear Regression**	**DNN**	**CNN**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**
MSE	0.1464	0.1579	0.1628	0.62	0.67	0.84
RMSE	0.3826	0.3973	0.4035	0.7874	0.8185	0.9165

Πίνακας 42: Σύγκριση καλύτερων μεθόδων *Πειράματος 3*, εκπαιδευμένων με δεδομένα ωριαίων καταναλώσεων ενιαίων ημερών όλου του έτους και σχετικής εργασίας, εκπαιδευμένων με δεδομένα μόνο των πρώτων μηνών του έτους, σε ωριαία βάση προβλέψεων

Αντίστοιχα, στον *Πίνακα 43,* παρουσιάζεται η απόδοση των τριών καλύτερων μεθόδων του *Πειράματος 3* και της σχετικής εργασίας στο σύνολο δοκιμής (test set) στο πρόβλημα της ημερήσιας πρόβλεψης κατανάλωσης φυσικού αερίου.

	Διπλωματική εργασία			**[KV2019]**		
	XGBoost	**ANN 2 Hidden**	**ANN 3 Hidden**	**Linear Regression**	**DNN**	**CNN**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**
MSE	7.1144	7.4756	7.1280	99	104	115
RMSE	2.6673	2.7342	2.6698	9.9499	10.1980	10.7238

Πίνακας 43: Σύγκριση καλύτερων μεθόδων *Πειράματος 3*, εκπαιδευμένων με δεδομένα ωριαίων καταναλώσεων ενιαίων ημερών όλου του έτους και σχετικής εργασίας, εκπαιδευμένων με δεδομένα μόνο των πρώτων μηνών του έτους, σε ημερήσια βάση προβλέψεων

Συγκρίνοντας τους παραπάνω πίνακες, καθίσταται σαφές πως τα μοντέλα που υλοποιήθηκαν στην παρούσα διπλωματική εργασία ξεπέρασαν σημαντικά σε απόδοση αυτά της εργασίας [KV2019], τόσο στην παραγωγή ωριαίων, όσο και ημερησίων προβλέψεων κατανάλωσης. Το γεγονός αυτό οφείλεται, όπως και στην περίπτωση του *Πειράματος 2*, στη μεθοδολογία δειγματοληψίας των συνόλων εκπαίδευσης και δοκιμής, καθώς πλέον τα μοντέλα παύουν να χαρακτηρίζονται από μεροληψία (bias) και προσαρμόζονται καλύτερα στα δεδομένα εκπαίδευσης, καθώς εκπαιδεύονται με δεδομένα όλων των μηνών του έτους. Η μεθοδολογία δειγματοληψίας που εφαρμόστηκε στο *Πείραμα 3*, αποτελεί την πλέον κατάλληλη για την παραγωγή ημερησίων προβλέψεων κατανάλωσης. Η βελτίωση της απόδοσης των προβλέψεων σε σχέση με την εργασία [KV2019] είναι εμφανής, καθώς τα σφάλματα των προβλέψεων σε όρους RMSE είναι σημαντικά μικρότερα.

Στον *Πίνακα 44*, καταγράφεται η απόδοση των μοντέλων στο πρόβλημα του υπολογισμού της μέγιστης (peak) κατανάλωσης της ημέρας.

	Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN 2 Hidden**	**ANN 3 Hidden**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**
R^2	0.8477	0.8471	0.7655	0.8482	0.7705	0.8646	0.8099	0.8039
MAE	0.4139	0.4145	0.5212	0.4126	0.4385	0.4177	0.4905	0.4817
MSE	0.3152	0.3163	0.4852	0.3141	0.4749	0.2801	0.3933	0.4058
RMSE	0.5615	0.5624	0.6966	0.5605	0.6891	0.5293	0.6271	0.6370

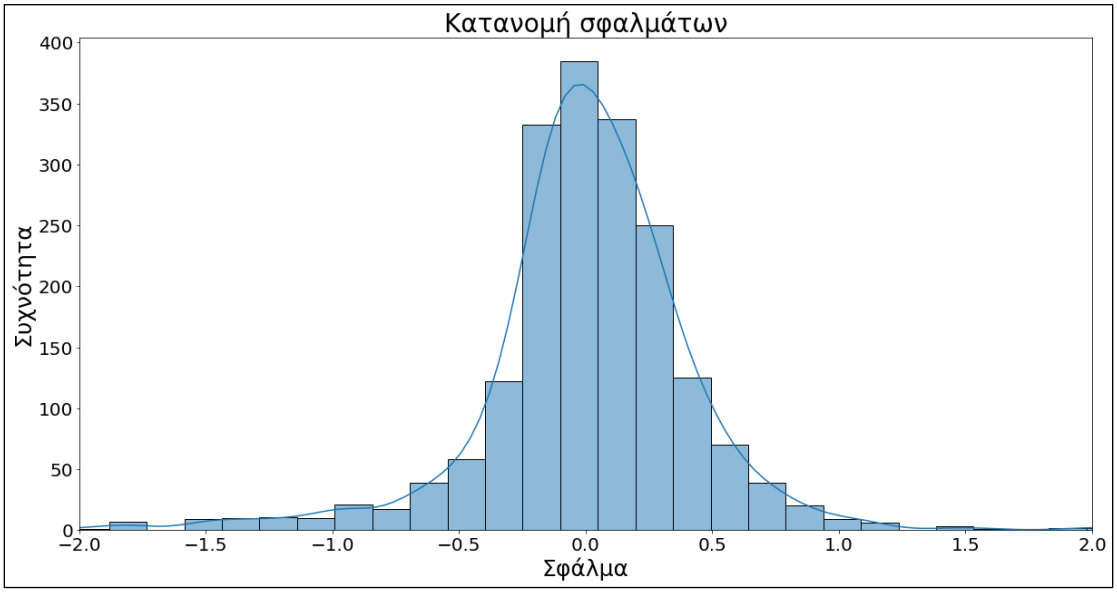
Πίνακας 44: Απόδοση μοντέλων τρίτου πειράματος στον υπολογισμό της μέγιστης ωριαίας κατανάλωσης της ημέρας στο σύνολο δοκιμής (test set)

Η μέθοδος *XGBoost* παρουσίασε την υψηλότερη απόδοση προβλέψεων, ενώ ενδιαφέρον αποτελεί το γεγονός, ότι στην περίπτωση του υπολογισμού της μέγιστης αναμενόμενης κατανάλωσης, οι γραμμικές μέθοδοι *Linear Regression* και *Ridge Regression* αποδείχθηκαν πιο αποδοτικές από τις υπόλοιπες πιο σύνθετες μεθόδους. Επίσης, η μέθοδος *LinearSVR*, δηλαδή το μοντέλο SVR με γραμμικό πυρήνα, παρουσίασε απόδοση αντίστοιχη με των δύο γραμμικών μεθόδων που αναφέρθηκαν.

5.3.2 Ποιοτική ανάλυση λειτουργίας μοντέλων

Είναι σημαντικό, προτού γίνει η αξιολόγηση και σύγκριση των μοντέλων, να έχει εξασφαλιστεί η σωστή λειτουργία τους, έτσι ώστε να αποφεύγονται προβλήματα όπως η υπερπροσαρμογή (overfitting) και η υπεραπλούστευση (underfitting). Ο καλύτερος τρόπος να γίνει αυτό, είναι μέσο της μελέτης των σφαλμάτων (residuals) που παράγει το κάθε μοντέλο. Με τον όρο σφάλμα (residual), εννοείται η διαφορά μεταξύ των προβλεπόμενων τιμών από το μοντέλο και των πραγματικών τιμών της μεταβλητής στόχου (target value). Ο σχεδιασμός του ιστογράμματος των σφαλμάτων αποτελεί βασικό τρόπο ανάλυσης των σφαλμάτων που παράγονται. Όταν το σφάλμα αυτό ακολουθεί κανονική κατανομή με μηδενική μέση τιμή, σημαίνει ότι τα σφάλματα παράγονται με τυχαίο τρόπο και άρα το μοντέλο δεν κάνει συστηματικά λάθος προβλέψεις για κάποιο συγκεκριμένο εύρος τιμών της μεταβλητής στόχου. Αντίθετα, αν η κατανομή των σφαλμάτων δεν είναι η προαναφερθείσα, υπάρχει κάποια συστηματική παραγωγή σφαλμάτων από το μοντέλο.

Στα μοντέλα που υλοποιήθηκαν έγινε η παραπάνω μελέτη και αποδείχθηκε ότι τα σφάλματα που παράγονται ακολουθούν πράγματι κανονική κατανομή με μηδενική μέση τιμή. Ενδεικτικά παρουσιάζεται η εν λόγω κατανομή για ένα από τα μοντέλα που υλοποιήθηκαν. Η παραπάνω εικόνα είναι παρόμοια για όλα τα μοντέλα και των τριών πειραμάτων.



Εικόνα 27: Κατανομή σφαλμάτων (residuals) μοντέλων

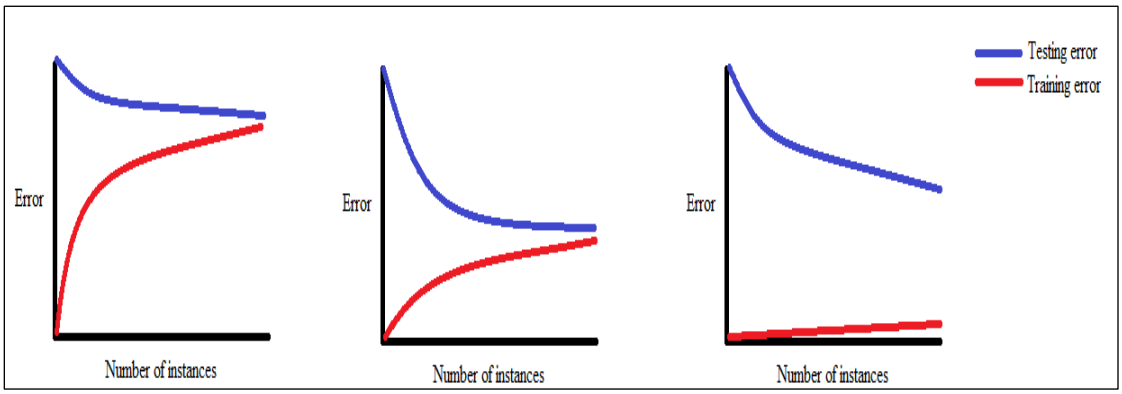
Στον *Πίνακα 45*, καταγράφεται η ακριβής μέση τιμή και τυπική απόκλιση των σφαλμάτων των μοντέλων, όπως προέκυψε για κάθε ένα από τα τρία πειράματα που πραγματοποιήθηκαν. Από τον πίνακα αυτό, επιβεβαιώνεται το γεγονός ότι η μέση τιμή των σφαλμάτων των μοντέλων σε όλα τα πειράματα είναι μηδενική. Ενδιαφέρον παρουσιάζει το γεγονός, ότι στο *Πείραμα 2* και *Πείραμα 3*, η μέση τιμή και η τυπική απόκλιση των σφαλμάτων των μοντέλων είναι σημαντικά μικρότερες σε σχέση με το *Πείραμα 1*. Το γεγονός αυτό οφείλεται στη μεθοδολογία δειγματοληψίας των συνόλων εκπαίδευσης και δοκιμής που εφαρμόστηκε σε κάθε πείραμα. Συγκεκριμένα, όπως έχει ήδη αναφερθεί, η μεθοδολογία που εφαρμόστηκε στο *Πείραμα 1*, εισάγει ένα βαθμό μεροληψίας (bias) στα μοντέλα, τα οποία εκπαιδεύονται με δεδομένα κατανάλωσης των πρώτων μόνο μηνών του έτους, ενώ παράγουν προβλέψεις για τους τελευταίου μήνες, οι οποίοι ωστόσο δεν περιλαμβάνονται καθόλου στο σύνολο εκπαίδευσης. Αντίθετα, η ορθότερη μεθοδολογία δειγματοληψίας των δεδομένων των δύο επόμενων πειραμάτων οδηγεί σε υψηλότερη ακρίβεια προβλέψεων και μείωση των σφαλμάτων, όπως υποδεικνύουν και οι τιμές του *Πίνακα 45*.

	Πείραμα 1	**Πείραμα 2**	**Πείραμα 3**
Μέση τιμή	-0.1286	0.0198	0.0083
Τυπική απόκλιση	0.6589	0.4685	0.4253

Πίνακας 45: Μέση τιμή και τυπική απόκλιση σφαλμάτων

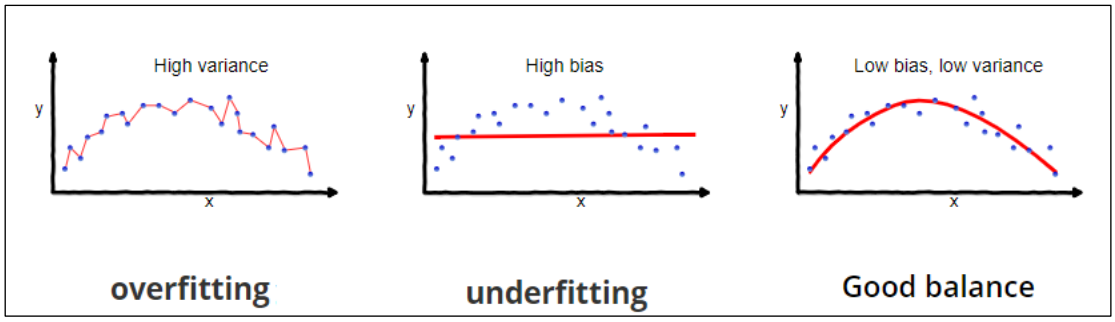
Επιπρόσθετα, η *καμπύλη μάθησης* (*learning curve*) αποτελεί ένα ακόμα εργαλείο διάγνωσης των μοντέλων μηχανικής μάθησης. Η καμπύλη μάθησης είναι ένα γράφημα, το οποίο συγκρίνει την απόδοση ενός μοντέλου κατά την εκπαίδευση και κατά τη δοκιμή, ενώ τα παραδείγματα εκπαίδευσης αυξάνονται. Μελετώντας τις καμπύλες αυτές, μπορούν να

διαγνωστούν προβλήματα κατά την εκπαίδευση του μοντέλου, όπως η υπερπροσαρμογή (overfitting) και η υπεραπλούστευση (underfitting).



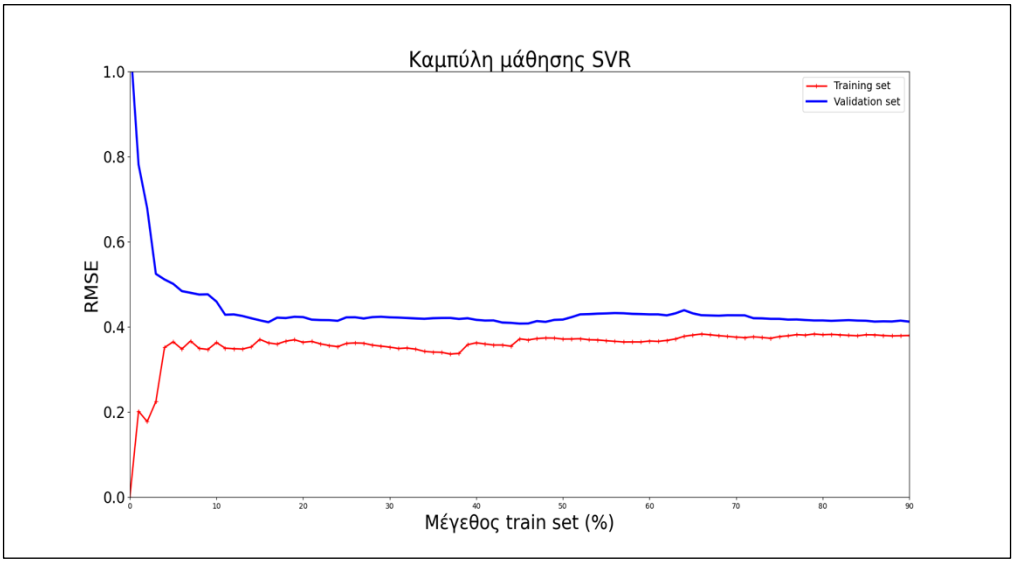
Εικόνα 28: Ενδεικτικές καμπύλες μάθηση (learning curves) με Υψηλή προτίμηση (αριστερά), Ιδεατή (κέντρο), Υψηλή διακύμανση (δεξιά)

Στην ιδεατή περίπτωση, η απόδοση των μοντέλων αυξάνεται όσο αυξάνεται το πλήθος των παραδειγμάτων εκπαίδευσης. Στην περίπτωση αυτή, τα σφάλματα στο σύνολο εκπαίδευσης (training error) και στο σύνολο αξιολόγησης (validation error), συγκλίνουν σε μια σχετικά χαμηλή τιμή. Όταν η τιμή σφάλματος, στην οποία τα δύο αυτά σύνολα συγκλίνουν παραμένει σε υψηλά επίπεδα, ανεξάρτητα από τον αριθμό των παραδειγμάτων εκπαίδευσης, τότε το μοντέλο παράγει σφάλματα με συστηματικό τρόπο και χαρακτηρίζεται από υψηλή προτίμηση (high bias). Αυτό είναι ενδεικτικό της ύπαρξης του φαινομένου underfitting. Όταν υπάρχει μεγάλη διαφορά μεταξύ των δύο αυτών σφαλμάτων, ενώ οι αντίστοιχες καμπύλες δε συγκλίνουν παρά την αύξηση των παραδειγμάτων εκπαίδευσης, το μοντέλο παρουσιάζει υψηλή διακύμανση (high variance) και για την βελτίωσή του απαιτείται προσθήκη επιπλέον δεδομένων εκπαίδευσης και χρήση λιγότερων ή απλούστερων χαρακτηριστικών. Κάτι τέτοιο είναι ενδεικτικό της ύπαρξης του φαινομένου overfitting. [AC1991], [Per2011]



Εικόνα 29: Ενδεικτικές περιπτώσεις overfitting, underfitting και καλής προσαρμογής

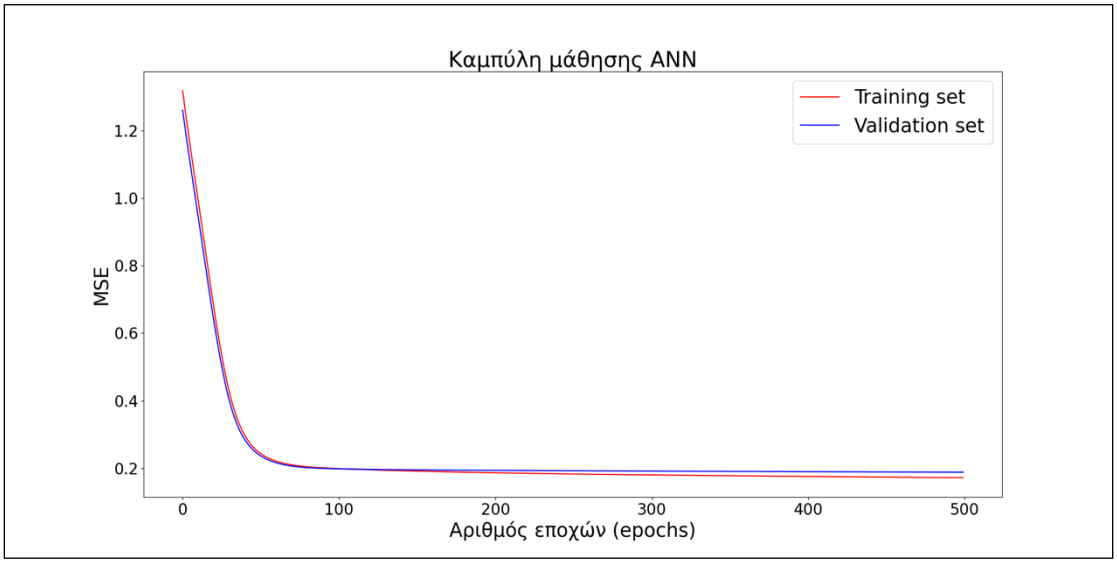
Στα πλαίσια της εργασίας, δημιουργήθηκαν οι καμπύλες μάθησης όλων των μοντέλων και των τριών πειραμάτων. Ενδεικτικά, στην *Εικόνα 30,* παρουσιάζεται η καμπύλη μάθησης της μεθόδου SVR του δεύτερου πειράματος. Η εικόνα αυτή είναι παρόμοια για όλα τα μοντέλα και των τριών πειραμάτων.



Εικόνα 30: Καμπύλη μάθησης

Οι καμπύλες μάθησης, πλησιάζουν αρκετά την εικόνα της ιδεατής περίπτωσης, όπως αυτή αναλύθηκε νωρίτερα. Τα σφάλματα των συνόλων εκπαίδευσης και αξιολόγησης, συγκλίνουν σε μια σχετικά μικρή τιμή όσο αυξάνεται το πλήθος των δεδομένων εκπαίδευσης, ενώ η προσθήκη επιπλέον δεδομένων στο σύνολο, είναι πιθανό να οδηγήσει σε ακόμα καλύτερη σύγκλιση και μείωση του σφάλματος των δύο συνόλων.

Τέλος, παρουσιάζεται η καμπύλη μάθησης των τεχνητών νευρωνικών δικτύων (ANN) που σχεδιάστηκαν. Στην περίπτωση των τεχνητών νευρωνικών δικτύων, η διάγνωση τυχόν προβλημάτων γίνεται μελετώντας τη καμπύλη της συνάρτησης απωλειών, στην προκειμένη περίπτωση του μέσου τετραγωνικού σφάλματος (MSE), όπως μεταβάλλεται σε κάθε εποχή (epoch) κατά τη διαδικασία της εκπαίδευσης. Ενδεικτικά, στην *Εικόνα 31,* παρουσιάζεται η καμπύλη μάθησης ενός από τα νευρωνικά δίκτυα που σχεδιάστηκαν. Η εικόνα αυτή είναι παρόμοια για όλα τα τεχνητά νευρωνικά δίκτυα και των τριών πειραμάτων.



Εικόνα 31: Καμπύλη μάθησης τεχνητού νευρωνικού δικτύου

Από την εικόνα της καμπύλης μάθησης, δε φαίνεται να παρουσιάζονται προβλήματα κατά την εκπαίδευση των νευρωνικών δικτύων. Η συνάρτηση απωλειών στο σύνολο αξιολόγησης μειώνεται αισθητά όσο αυξάνονται οι εποχές (epochs), ενώ φαίνεται να συγκλίνει σε μια σταθερή τιμή, πλησιάζοντας τη τιμή σύγκλισης του συνόλου εκπαίδευσης. Η προσθήκη επιπλέον παραδειγμάτων στο σύνολο δεδομένων αναμένεται να βελτιώσει την απόδοση του νευρωνικού δικτύου.

5.3.3 Ανάλυση συντελεστή προσδιορισμού R^2

Ο συντελεστής προσδιορισμού R^2, μπορεί να δώσει μια εκτίμηση της ικανότητας των μοντέλων να προσαρμοστούν στα δεδομένα εκπαίδευσης. Στην περίπτωση μιας «τέλειας» πρόβλεψης ο συντελεστής προσδιορισμού θα είχε τιμή $R^2 = 1$. Στον *Πίνακα 46*, παρουσιάζεται η τιμή του συντελεστή προσδιορισμού των ωριαίων προβλέψεων στο σύνολο δοκιμής (test set) κάθε μοντέλου και για τα τρία πειράματα που πραγματοποιήθηκαν.

Πείραμα 1							
Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN**	
Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**	
0.6591	0.6540	0.6329	0.6270	0.6287	0.6330	**0.6605**	
Πείραμα 2							
Linear Regression	**Ridge Regression**	**SVR**	**AdaBoost**		**XGBoost**	**ANN**	
Test	**Test**	**Test**	**Test**		**Test**	**Test**	
0.7309	0.7313	0.7592	0.7029		**0.7865**	0.7758	
Πείραμα 3							
Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN 2 Hidden**	**ANN 3 Hidden**
Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**
0.7744	0.7743	0.8124	0.7746	0.7462	**0.8313**	0.8128	0.8181

Πίνακας 46: Συντελεστής προσδιορισμού R^2 ωριαίων προβλέψεων στο σύνολο δοκιμής (test set)

Από τον παραπάνω πίνακα, γίνεται σαφής η βελτίωση της ικανότητας των μοντέλων να προσαρμοστούν στα δεδομένα εκπαίδευσης στα *Πειράματα 2* και *3*, γεγονός που οφείλεται στην ορθότερη μεθοδολογία δειγματοληψίας που εφαρμόστηκε. Ακόμα, τόσο στο *Πείραμα 2,* όσο και στο *Πείραμα 3* φαίνεται η υπεροχή της μεθόδου ***XGBoost*** στο πρόβλημα της πρόβλεψης ωριαίας κατανάλωσης φυσικού αερίου. Η μέθοδος των ***ANN***, καθώς και η ***SVR*** παρουσιάζουν επίσης υψηλές τιμές του συντελεστή προσδιορισμού και άρα ικανότητα προσαρμογής στα δεδομένα εκπαίδευσης. Εξαίρεση στα παραπάνω αποτελεί το *Πείραμα 1*, στο οποίο λόγω της μεθοδολογίας δειγματοληψίας που εφαρμόστηκε, οι γραμμικές μέθοδοι *Linear Regression* και *Ridge Regression* παρουσίασαν καλύτερη προσαρμογή στα δεδομένα σε σχέση με τις υπόλοιπες μεθόδους, χωρίς όμως να ξεπερνούν σε απόδοση τη μέθοδο των *ANN.* Σε κάθε περίπτωση, η μέθοδος *AdaBoost* αποδείχθηκε εκείνη με τη χειρότερη απόδοση.

5.3.4 Ανάλυση δεικτών MAE και RMSE

Πέρα από την ικανότητα προσαρμογής των μοντέλων στα δεδομένα εκπαίδευσης, σημαντικό βήμα για την αξιολόγηση της απόδοσής τους αποτελεί η ανάλυση των σφαλμάτων που αυτά παράγουν, μέσω των δεικτών αξιολόγησης MAE και RMSE. Οι δείκτες αυτοί, έχουν χρησιμοποιηθεί από πλήθος εργασιών πρόβλεψης κατανάλωσης ενέργειας για την αξιολόγηση των προβλέψεων των μοντέλων που υλοποιήθηκαν. Η σχετική βιβλιογραφία ωστόσο, φαίνεται να αναδεικνύει το δείκτη RMSE ως το πλέον κατάλληλο για το πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου. Ο δείκτης αυτός, σύμφωνα με την [CD2014], είναι καταλληλότερος για την αξιολόγηση των μοντέλων σε σχέση με τον MAE, όταν η κατανομή του σφάλματος των μοντέλων είναι κανονική. Ακόμα, μεγάλο πλεονέκτημα του δείκτη RMSE, ιδιαίτερα στη βραχυπρόθεσμη πρόβλεψης της κατανάλωσης, αποτελεί το γεγονός ότι ο δείκτης αυτός «τιμωρεί» περισσότερο τα μεγαλύτερα σφάλματα σε σχέση με τα πιο μικρά. Έτσι, τα σφάλματα σε ώρες ή μέρες αιχμής επιβαρύνονται περισσότερο. Αυτό είναι ιδιαίτερο σημαντικό, καθώς όταν η ζήτηση φυσικού αερίου είναι υψηλότερη, τότε η τιμή του φυσικού αερίου είναι επίσης πιο υψηλή, γεγονός που συνεπάγεται μεγάλο κόστος στην περίπτωση λανθασμένης εκτίμησης της [MP2018].

Στον *Πίνακα 47*, παρουσιάζονται οι τιμές των δύο αυτών δεικτών κάθε μοντέλου των τριών πειραμάτων στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης φυσικού αερίου στο σύνολο δοκιμής (test set).

	Πείραμα 1							
	Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN**	
	Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**	
MAE	**0.3969**	**0.3974**	0.4043	0.4223	0.4275	0.4182	**0.4117**	
RMSE	**0.6599**	**0.6648**	0.6848	0.6903	0.6887	0.6847	**0.6585**	
	Πείραμα 2							
	Linear Regression	**Ridge Regression**	**SVR**	**AdaBoost**		**XGBoost**	**ANN**	
	Test	**Test**	**Test**	**Test**		**Test**	**Test**	
MAE	0.3028	0.3021	**0.2605**	0.3273		**0.2497**	**0.2536**	
RMSE	0.4856	0.4852	**0.4594**	0.5103		**0.4325**	**0.4432**	
	Πείραμα 3							
	Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN 2 Hidden**	**ANN 3 Hidden**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**
MAE	0.2660	0.2861	0.2375	0.2857	0.3029	**0.2284**	**0.2446**	**0.2388**
RMSE	0.4425	0.4425	0.4035	0.4423	0.4693	**0.3826**	**0.4030**	**0.3973**

Πίνακας 47: Δείκτες MAE και RMSE των ωριαίων προβλέψεων στο σύνολο δοκιμής (test set)

Ο *Πίνακας 47* επιβεβαιώνει την υπεροχή των μεθόδων ***XGBoost*** και ***ANN*** στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης φυσικού αερίου. Το σφάλμα της μεθόδου *XGBoost* είναι το μικρότερο όλων των μεθόδων, τόσο σε όρους MAE όσο και σε όρους RMSE. Εξαίρεση αποτελεί το πρώτο πείραμα, στο οποίο η μέθοδος με το μικρότερο σφάλμα είναι το *ANN*, ενώ οι γραμμικές μέθοδοι *Linear Regression* και *Ridge Regression* αποδείχθηκαν πιο αποδοτικές από τη *XGBoost*. Υψηλή απόδοση στο παρόν πρόβλημα, φαίνεται να έχει επίσης η μέθοδος ***SVR***, όπως υποδεικνύουν οι δείκτες MAE και RMSE των *Πειραμάτων 2* και *3*. Τέλος, η μέθοδος *AdaBoost*, αποδείχθηκε η μέθοδος με τη χειρότερη απόδοση στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης φυσικού αερίου. Είναι σημαντικό να αναφερθεί η σημαντική αύξηση της απόδοσης των προβλέψεων όλων των μοντέλων των *Πειραμάτων 2* και *3*, σε σχέση με το *Πείραμα 1*. Το γεγονός αυτό γίνεται εμφανές από την αισθητή μείωση των σφαλμάτων των προβλέψεων σε όρους MAE και RMSE και οφείλεται στην ορθότερη μεθοδολογία δειγματοληψίας των δεδομένων στα σύνολα εκπαίδευσης και δοκιμής.

Στο *Πείραμα 1* και *2*, πραγματοποιήθηκε επίσης ημερήσια πρόβλεψη κατανάλωσης φυσικού αερίου. Η ημερήσια αυτή εκτίμηση της κατανάλωσης, προέκυψε με άθροιση των επιμέρους ωριαίων καταναλώσεων. Στον *Πίνακα 48*, παρουσιάζονται οι τιμές των δεικτών MAE και RMSE κάθε μοντέλου των δύο πειραμάτων στο σύνολο δοκιμής (test set).

	Πείραμα 1							
	Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN**	
	Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**	
MAE	**1.6239**	**2.053**	4.4248	4.3108	1.8950	4.8926	3.8061	
RMSE	**2.5091**	**3.1035**	6.0145	5.3404	3.0797	6.2491	4.5697	
	Πείραμα 3							
	Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN 2 Hidden**	**ANN 3 Hidden**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**
MAE	1.9187	1.9265	1.7479	1.9141	1.9742	**1.6127**	**1.7946**	**1.7170**
RMSE	2.8690	2.8719	2.9609	2.8509	3.0278	**2.6673**	**2.7342**	**2.6698**

Πίνακας 48: Δείκτες MAE και RMSE των ημερησίων προβλέψεων στο σύνολο δοκιμής (test set)

Από τον *Πίνακα 48*, αποδεικνύεται η υπεροχή των μεθόδων *XGBoost* και *ANN* και στο πρόβλημα της ημερήσιας πρόβλεψης κατανάλωσης στο *Πείραμα 3*. Οι μέθοδοι αυτοί παρουσιάζουν το μικρότερο σφάλμα σε όρους MAE και RMSE. Αντίστοιχα, στο *Πείραμα 1*, οι γραμμικές μέθοδοι *Linear Regression* και *Ridge Regression* ήταν αυτές με το μικρότερο σφάλμα προβλέψεων.

Τέλος, στο *Πείραμα 1* και *3* πραγματοποιήθηκε επίσης πρόβλεψη της μέγιστης ωριαίας κατανάλωσης της ημέρας. Οι αποδόσεις των μοντέλων σε όρους MAE και RMSE αποτυπώνονται στον *Πίνακα 49.*

	Πείραμα 1						
	Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**
MAE	**0.4947**	**0.5224**	0.9273	0.6246	0.6650	0.9568	0.7100
RMSE	**0.6254**	**0.6595**	1.1337	0.7586	0.9866	1.2113	0.8042

	Πείραμα 3							
	Linear Regression	**Ridge Regression**	**SVR**	**Linear SVR**	**AdaBoost**	**XGBoost**	**ANN 2 Hidden**	**ANN 3 Hidden**
	Test	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**	**Test**
MAE	**0.4139**	**0.4145**	0.5212	0.4126	0.4385	**0.4177**	0.4905	0.4817
RMSE	**0.5615**	**0.5624**	0.6966	0.5605	0.6891	**0.5293**	0.6271	0.6370

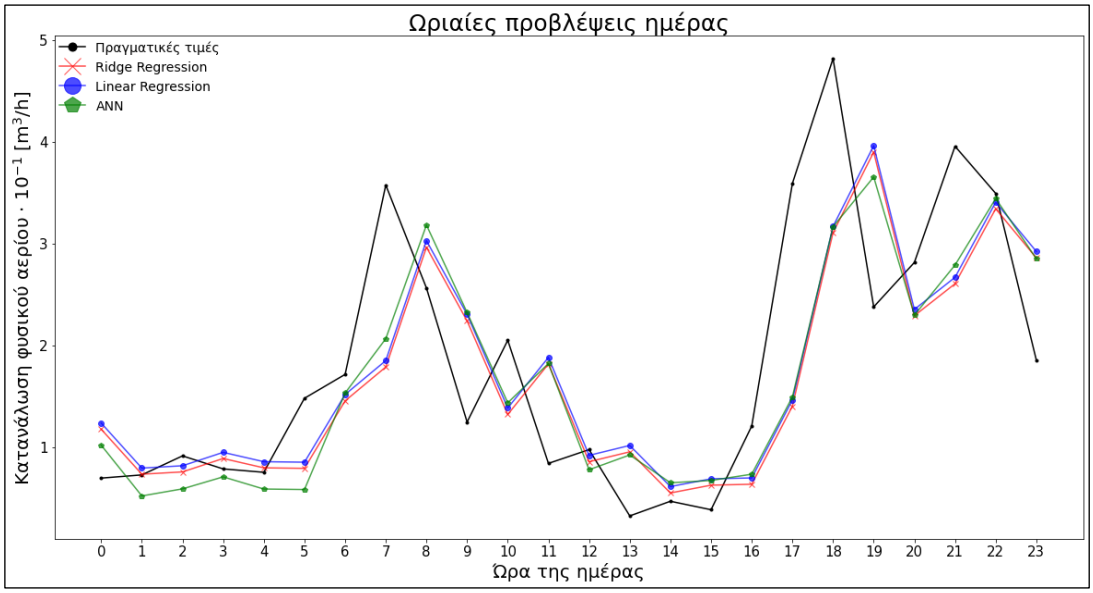
Πίνακας 49: Δείκτες MAE και RMSE προβλέψεων μέγιστης ωριαίας κατανάλωσης της ημέρας στο σύνολο δοκιμής (test set)

Όπως ήταν αναμενόμενο, λόγω των αντίστοιχων αποδόσεων στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης φυσικού αερίου, η μέθοδος ***XGBoost*** αποτέλεσε τη μέθοδο με το μικρότερο σφάλμα στο *Πείραμα 3*, ενώ οι γραμμικές μέθοδοι ***Linear*** *και* ***Ridge Regression*** τις αντίστοιχες μεθόδους του *Πειράματος 1*. Ενδιαφέρον αποτελεί το γεγονός, ότι οι δύο αυτές γραμμικές μέθοδοι παρουσίασαν επίσης υψηλή απόδοση και στο *Πείραμα 3*, στο πρόβλημα της πρόβλεψης της μέγιστης ωριαίας κατανάλωσης της ημέρας. Τέλος, όπως φαίνεται και από τον παραπάνω πίνακα, η μεθοδολογία δειγματοληψίας των συνόλων εκπαίδευσης και δοκιμής που εφαρμόστηκε στο *Πείραμα 3*, αύξησε την απόδοση των μοντέλων σε σχέση με το *Πείραμα 1,* μειώνοντας σημαντικά τα αντίστοιχα σφάλματα προβλέψεων.

5.3.5 Οπτική απεικόνιση απόδοσης προβλέψεων

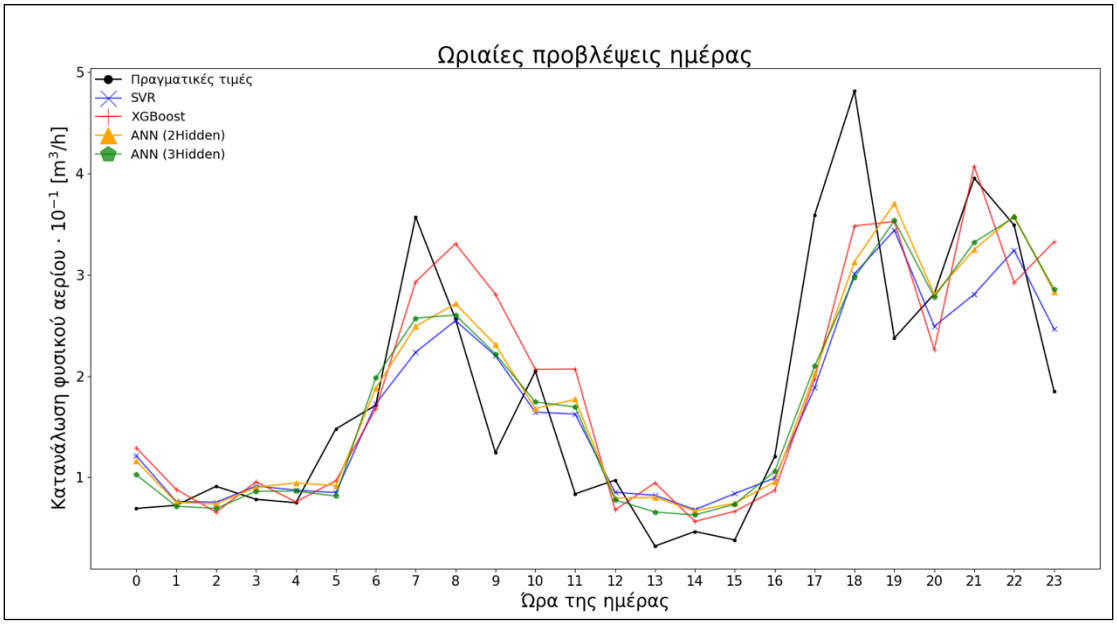
Για την καλύτερη μελέτη της απόδοσης των μοντέλων, θα παρουσιαστούν αναλυτικότερα οι προβλέψεις τους σε κάποιες ενδεικτικές περιπτώσεις του συνόλου δοκιμής. Με τον τρόπο αυτό, μπορεί να γίνει πιο εμφανής η συμπεριφορά των μοντέλων και ευκολότερη η σύγκριση μεταξύ τους. Συγκεκριμένα, για τα *Πειράματα 1* και 3, θα παρουσιαστούν οι ωριαίες προβλέψεις των μοντέλων με την υψηλότερη απόδοση κατά τη διάρκεια μιας ημέρας του Νοεμβρίου. Το μήνα αυτό, οι καταναλώσεις αναμένονται υψηλές.

Στο *Πείραμα 1*, τα μοντέλα με την υψηλότερη απόδοση στο πρόβλημα της ωριαίας πρόβλεψης κατανάλωσης φυσικού αερίου, ήταν τα γραμμικά μοντέλα ***Linear Regression*** και ***Ridge Regression***, καθώς και η μέθοδος των ***τεχνητών νευρωνικών δικτύων*** (***ANN***). Στην *Εικόνα 32*, αποτυπώνονται οι ωριαίες προβλέψεις των τριών αυτών μοντέλων κατά τη διάρκεια μιας ενδεικτικής ημέρας του έτους.



Εικόνα 32: Ωριαίες προβλέψεις αποδοτικότερων μοντέλων Πειράματος 1

Αντίστοιχα, στο *Πείραμα 3*, τα μοντέλα με την υψηλότερη απόδοση ήταν τα ***XGBoost***, ***SVR*** καθώς και τα ***τεχνητά νευρωνικά δίκτυα*** (***ANN***). Στην *Εικόνα 33*, αποτυπώνονται οι ωριαίες προβλέψεις των τριών αυτών μοντέλων κατά τη διάρκεια της ίδιας ημέρας που παρουσιάστηκε και στο *Πείραμα 1*.



Εικόνα 33: Ωριαίες προβλέψεις αποδοτικότερων μοντέλων Πειράματος 3

Από τις δύο προηγούμενες εικόνες, φαίνεται πως τα μοντέλα του *Πειράματος 3*, παρήγαγαν συστηματικότερα ακριβέστερες προβλέψεις. Ιδιαίτερα ακριβής μάλιστα, είναι η μέθοδος *XGBoost*, όπως προέκυψε και από την ανάλυση των *Υποενοτήτων 5.3.3- 5.3.4*. Είναι σημαντικό να αναφερθεί, ότι η υπό εξέταση ημέρα χαρακτηρίζεται από υψηλά επίπεδα κατανάλωσης με έντονες ωριαίες διακυμάνσεις της. Τα μοντέλα που σχεδιάστηκαν στο *Πείραμα 3*, έδειξαν μεγαλύτερη ικανότητα στην πρόβλεψη αυτών των διακυμάνσεων, γεγονός που οφείλεται στην ορθότερη μεθοδολογία που ακολουθήθηκε κατά την εκπαίδευσή τους. Τέλος, επιβεβαιώνεται το γεγονός ότι κανένα μοντέλο δεν παρήγαγε συστηματικά σφάλματα, όπως η σταθερή πρόβλεψη χαμηλότερων ή υψηλότερων καταναλώσεων από τις πραγματικές.

5.4 Συμπεράσματα

Από την παρουσίαση των αποτελεσμάτων και την ανάλυση που προηγήθηκε γίνεται σαφής η σημασία της σωστής επιλογής των χαρακτηριστικών των μοντέλων πρόβλεψης, καθώς και της μεθοδολογίας δειγματοληψίας των συνόλων εκπαίδευσης και δοκιμής των μοντέλων πρόβλεψης. Όπως αποδείχτηκε, η επιλογή μόνο λίγων βασικών χαρακτηριστικών, όπως έγινε στην εργασία [KV2019], χωρίς τη σχετική ανάλυση και μελέτη που προηγήθηκε στην παρούσα διπλωματική, οδηγεί σε μειωμένη απόδοση προβλέψεων, ακόμα κι αν τα χαρακτηριστικά αυτά διαθέτουν κάποια συσχέτιση με την κατανάλωση φυσικού αερίου. Επιπλέον, φάνηκε πως ανάλογα με το μέγεθος και το είδος του συνόλου δεδομένων, η επιλογή των μοντέλων πρόβλεψης που θα δοκιμαστούν στο πρόβλημα είναι θεμελιώδους σημασίας. Το συγκεκριμένο σύνολο δεδομένων δεν περιλαμβάνει σημαντικά μεγάλο πλήθος εγγραφών, ενώ καλύπτει μόλις ένα έτος. Έτσι, μέθοδοι που είναι ιδιαίτερα απαιτητικές ως προς το πλήθος των δεδομένων εκπαίδευσης, όπως τα σύνθετα τεχνητά νευρωνικά δίκτυα αποδείχθηκαν μη αποτελεσματικές. Αντίθετα, οι μέθοδοι παλινδρόμησης που δοκιμάστηκαν στην παρούσα διπλωματική αποδείχθηκαν πιο αποτελεσματικές και ακριβείς από τα τεχνητά νευρωνικά δίκτυα σύνθετης αρχιτεκτονικής της [KV2019]. Ειδικότερα, τα μοντέλα *XGBoost SVR,* καθώς και το *τεχνητό νευρωνικό δίκτυο 2 ή 3 κρυμμένων επιπέδων* παρουσίασαν την υψηλότερη απόδοση ωριαίων προβλέψεων κατανάλωσης φυσικού αερίου. Συνολικά, το μοντέλο παλινδρόμησης *XGBoost,* θα μπορούσε να θεωρηθεί πως παρήγαγε συστηματικά τις προβλέψεις με το μικρότερο σφάλμα σε όρους MAE και RMSE. Τέλος, θεμελιώδους σημασίας αποδείχθηκε πως είναι η μεθοδολογία δειγματοληψίας των δεδομένων των συνόλων εκπαίδευσης και δοκιμής των μοντέλων. Ο τρόπος διαχωρισμού των δύο συνόλων που εφαρμόστηκε στο *Πείραμα 1*, καθώς και στην εργασία [KV2019] εισήγαγε ένα βαθμό μεροληψίας (bias), καθώς τα μοντέλα εκπαιδεύτηκαν με δεδομένα μόνο των πρώτων μηνών του έτους, ενώ παρήγαγαν προβλέψεις για μήνες που δεν υπήρχαν στο σύνολο εκπαίδευσής

τους και χαρακτηρίζονταν από αρκετά διαφορετικές συνθήκες και επίπεδα κατανάλωσης. Αντίθετα, η μεθοδολογία δειγματοληψίας που εφαρμόστηκε στα *Πειράματα 2* και *3*, αύξησε σημαντικά την απόδοση των προβλέψεων, καθώς τα μοντέλα εκπαιδεύτηκαν με δεδομένα όλων των μηνών του έτους.

6 *Επίλογος*

6.1 Συμπεράσματα και Επεκτάσεις

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου με χρήση τεχνικών επιστήμης δεδομένων και μηχανικής μάθησης. Αρχικά έγινε μια εκτεταμένη καταγραφή και ανάλυση σχετικών εργασιών τεχνολογιών αιχμής πρόβλεψης κατανάλωσης φυσικού αερίου, καθώς και συλλογή, και παρουσίαση διαθέσιμων συνόλων δεδομένων.

Στη συνέχεια, καθορίστηκε και παρουσιάστηκε αναλυτικά η μεθοδολογία που απαιτείται για την εκπαίδευση μοντέλων μηχανικής μάθησης στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου. Ιδιαίτερη έμφαση δόθηκε στην μελέτη των διάφορων καιρικών και ημερολογιακών παραγόντων που επηρεάζουν την κατανάλωση φυσικού αερίου και στην τελική επιλογή των παραγόντων εκείνων που θα αποτελέσουν τα χαρακτηριστικά των μοντέλων πρόβλεψης.

Ακολούθως, παρουσιάστηκε η μεθοδολογία της βέλτιστης ρύθμισης των υπερπαραμέτρων των μοντέλων και έγινε εκτεταμένη αξιολόγηση ενός μεγάλου αριθμού μοντέλων μηχανικής μάθησης στο πρόβλημα. Στα πλαίσια της διπλωματικής, πραγματοποιήθηκαν τρία πειράματα πρόβλεψης κατανάλωσης φυσικού αερίου, χρησιμοποιώντας ένα διαθέσιμο σύνολο δεδομένων. Χωρίζοντας με τρεις διαφορετικούς τρόπους το σύνολο δεδομένων σε σύνολα εκπαίδευσης (train set) και δοκιμής (test set), εξετάστηκε η απόδοση των μοντέλων στο πρόβλημα της ωριαίας και ημερήσιας πρόβλεψης κατανάλωσης φυσικού αερίου, καθώς και στον υπολογισμό της μέγιστης ωριαίας κατανάλωσης που αναμένεται κατά τη διάρκεια της ημέρας. Ειδικότερα, αρχικά πραγματοποιήθηκαν ωριαίες προβλέψεις κατανάλωσης φυσικού αερίου, οι οποίες στη

συνέχεια, με κατάλληλη άθροιση έδωσαν την αντίστοιχη ημερήσια πρόβλεψη της κατανάλωσης.

Ακολούθησε εκτενής μελέτη των αποτελεσμάτων με χρήση κατάλληλων δεικτών αξιολόγησης, ενώ επίσης, έγινε συγκριτική αξιολόγηση της απόδοσης των μοντέλων που σχεδιάστηκαν στην παρούσα διπλωματική με την απόδοση που πέτυχαν οι ερευνητές σχετικής εργασίας που χρησιμοποίησε το ίδιο σύνολο δεδομένων. Ακόμα, αναλύθηκαν τα μειονεκτήματα της μεθοδολογίας που ακολούθησε η σχετική εργασία ως προς το διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής και προτάθηκαν δύο διαφορετικοί τρόποι μέσω των υπόλοιπων δύο πειραμάτων που πραγματοποιήθηκαν. Επιπλέον, έγινε ποιοτική ανάλυση της λειτουργίας των μοντέλων, προκειμένου να εξασφαλιστεί η σωστή λειτουργία τους κατά την παραγωγή των προβλέψεων. Η ανάλυση αυτή, αποτελεί σημαντικό εργαλείο διάγνωσης των μοντέλων, προκειμένου να γίνονται αντιληπτά προβλήματα όπως η υπερπροσαρμογή (overfitting) και η υπεραπλούστευση (underfitting).

Τα πειραματικά αποτελέσματα ανέδειξαν τις σύνθετες μεθόδους παλινδρόμησης XGBoost και SVR ως τις πλέον αποδοτικές στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου. Ειδικότερα, η μέθοδος XGBoost φάνηκε να παράγει συστηματικότερα τις προβλέψεις με το μικρότερο σφάλμα σε όρους MAE και RMSE. Η μέθοδος των τεχνητών νευρωνικών δικτύων, παρουσίασε επίσης υψηλή απόδοση στο πρόβλημα της πρόβλεψης κατανάλωσης φυσικού αερίου. Οι απλές γραμμικές μέθοδοι Ridge Regression και Linear Regression, απέδωσαν χειρότερα από τις προηγούμενες πιο σύνθετες μεθόδους, ωστόσο θα μπορούσαν να δώσουν μια γρήγορη εκτίμηση της κατανάλωσης και να αποτελέσουν μοντέλα αναφοράς. Επίσης, η ρύθμιση των υπερπαραμέτρων των μοντέλων αποδείχθηκε ιδιαίτερο σημαντικό βήμα για την παραγωγή αξιόπιστων προβλέψεων.

Στα πειράματα που πραγματοποιήθηκαν, ως χαρακτηριστικά των μοντέλων πρόβλεψης χρησιμοποιήθηκαν διάφοροι καιρικοί και ημερολογιακοί παράγοντες. Για τους καιρικούς παράγοντες αυτούς, λήφθηκαν υπόψη οι πραγματικές τιμές τους, όπως έγιναν διαθέσιμες από μετεωρολογικό σταθμό. Ωστόσο, κάτι τέτοιο δεν είναι δυνατό σε πραγματικές συνθήκες πρόβλεψης, καθώς μόνο οι προβλεπόμενες, από σχετικά μοντέλα, τιμές των παραγόντων αυτών θα είναι διαθέσιμες εκ των προτέρων. Μια προοπτική επέκταση της εργασίας, είναι η χρήση προβλεπόμενων τιμών των καιρικών παραγόντων ως χαρακτηριστικά των μοντέλων και η μελέτη της επίδρασης του σφάλματος πρόβλεψης των παραγόντων αυτών στις τελικές προβλέψεις της κατανάλωσης του φυσικού αερίου.

Στην παρούσα εργασία, όλα τα μοντέλα πρόβλεψης εκπαιδεύτηκαν χρησιμοποιώντας τα ίδια χαρακτηριστικά. Μια ακόμα επέκταση της εργασίας, είναι η ξεχωριστή επιλογή χαρακτηριστικών για κάθε ένα μοντέλο πρόβλεψης. Με τον τρόπο αυτό θα επιτευχθεί επιπλέον αύξηση της απόδοσης των μοντέλων, καθώς κάθε μοντέλο θα εκπαιδευτεί με το

βέλτιστο συνδυασμό χαρακτηριστικών, παράγοντας ακριβέστερες προβλέψεις. Η επιλογή των χαρακτηριστικών αυτών θα μπορούσε να γίνει, τόσο με τη μελέτη των συσχετίσεών τους με την κατανάλωση, όσο και αυτοματοποιημένα με χρήση σχετικών αλγορίθμων.

Η πρόβλεψη της κατανάλωσης φυσικού αερίου έγινε στο σύνολο της περιοχής. Ο λόγος είναι, ότι η περίοδος καταγραφής δεν ήταν ίδια για όλα τα κτήρια. Ακόμα, ο σχετικά μικρός αριθμός εγγραφών του συνόλου δεδομένων καθιστούσε αδύνατη την πρόβλεψη της κατανάλωσης για κάθε καταναλωτή ξεχωριστά. Ενδιαφέρουσα επέκταση της εργασίας, θα ήταν η παραγωγή επιπλέον δεδομένων κατανάλωσης για κάθε κτήριο. Αυτό προϋποθέτει τη στατιστική μελέτη της κατανάλωσης και την εύρεση της κατανομής πιθανότητας που προσομοιάζει καλύτερα την κατανάλωση φυσικού αερίου. Με τον τρόπο αυτό θα είναι διαθέσιμος μεγαλύτερος αριθμός δεδομένων και άρα δυνατή η πρόβλεψη των επιμέρους καταναλώσεων.

7 *Βιβλιογραφία*

[BM1995] R. H. Brown and I. Matin, "Development of artificial neural network models to predict daily gas consumption," Proceedings of IECON '95 - 21st Annual Conference on IEEE Industrial Electronics, Orlando, FL, USA, 1995, pp. 1389-1394 vol.2, doi: 10.1109/IECON.1995.484153.

[HH2013] Hosseinzadeh, Mostafa. (2013). Prediction of Long Term Residential Natural Gas Consumption Using ANN. Journal of Applied Mechanical Engineering. 02. 10.4172/2168-9873.1000120.

[SP2014] Soldo, Božidar & Potočnik, Primož & Šimunović, Goran & Šarić, Tomislav & Govekar, Edvard. (2014). Improving the residential natural gas consumption forecasting models by using solar radiation. Energy and Buildings. 69. 498–506. 10.1016/j.enbuild.2013.11.032.

[Szo2015] Szoplik, Jolanta. (2015). Forecasting of natural gas consumption with artificial neural networks. Energy. 85. 10.1016/j.energy.2015.03.084.

[SO2018] Z. Sičanica and Z. Oklopčić, "Countywide natural gas consumption forecast, a machine learning approach," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2018, pp. 1070-1073, doi: 10.23919/MIPRO.2018.8400195.

[MP2018] Merkel, Gregory & Povinelli, Richard & Brown, Ronald. (2018). Short-Term Load Forecasting of Natural Gas with Deep Neural Network Regression †. Energies. 11. 2008. 10.3390/en11082008.

[HP2018] Hribar, Rok & Potočnik, Primož & Šilc, Jurij & Papa, Gregor, 2019. "A comparison of models for forecasting the residential natural gas demand of an urban area," Energy, Elsevier, vol. 167(C), pages 511-522.

[WL2018] Wei, Nan & Li, Changjun & LI, CHAN & Xie, Hanyu & Du, Zhongwei & Zhang, Qiushi & Zeng, Fanhua. (2018). Short-Term Forecasting of Natural Gas Consumption Using Factor Selection Algorithm and Optimized Support Vector Regression. Journal of Energy Resources Technology. 141. 10.1115/1.4041413.

[WL2019] Wei, Nan & Li, Changjun & Peng, Xiaolong & Li, Yang & Zeng, Fanhua. (2019). Daily natural gas consumption forecasting via the application of a novel hybrid model. Applied Energy. 250. 358-368. 10.1016/j.apenergy.2019.05.023.

[WD2019] Wei, Nan & Li, Changjun & Duan, Jiehao & Liu, Jinyuan & Zeng, Fanhua. (2019). Daily Natural Gas Load Forecasting Based on a Hybrid Deep Learning Model. Energies. 12. 15. 10.3390/en12020218.

[BE2019] Beyca, Omer Faruk & Ervural, Beyzanur Cayir & Tatoglu, Ekrem & Ozuyar, Pinar Gokcin & Zaim, Selim, 2019. "Using machine learning tools for forecasting natural gas consumption in the province of Istanbul," Energy Economics, Elsevier, vol. 80(C), pages 937-949.

[KV2019] de Keijzer, Brian & de Visser, Pol & Garcia Romillo, Victor & Muñoz, Víctor & Boesten, Daan & Meezen, Megan & Salcedo Rahola, Tadeo Baldiri. (2019). Forecasting residential gas consumption with machine learning algorithms on weather data.

[FM2020] Marziali, Andrea & Fabbiani, Emanuele & Nicolao, Giuseppe. (2019). Forecasting residential gas demand: machine learning approaches and seasonal role of temperature forecasts.

[CD2014] Chai, Tianfeng & Draxler, R.R.. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. Geoscientific Model Development. 7. 1247-1250. 10.5194/gmd-7-1247-2014.

[SM2013] Sutskever, I. & Martens, J. & Dahl, G. & Hinton, G.. (2013). On the importance of initialization and momentum in deep learning. 30th International Conference on Machine Learning, ICML 2013. 1139-1147.

[MC2009] Munson, M. & Caruana, Rich. (2009). On Feature Selection, Bias-Variance, and Bagging. 144-159. 10.1007/978-3-642-04174-7_10.

[Per2011] Perlich, Claudia. (2011). Learning Curves in Machine Learning. 10.1007/978-0-387-30164-8_452.

[PA2013] Πετρόπουλος Φ. and Ασημακόπουλος Β. (2013) *Επιχειρησιακές Προβλέψεις*, Αθήνα: Εκδόσεις Συμμετρία

[Ger2019] Geron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly.