ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΕΝΟΣ ΔΙΚΤΥΟΥ ΤΗΛΕΘΕΡΜΑΝΣΗΣ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Διπλωματική Εργασία

του

Αναστάσιου Κουρίδη

Επιβλέποντες: Δρ. Μιχαήλ Βασιλακόπουλος

Δρ. Δημήτριος Μπαργιώτας

Βόλος, Ιούνιος 2018

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

DATA ANALYSIS OF A DISTRICT HEATING NETWORK USING MACHINE LEARNING TECHNIQUES

Diploma Thesis

by

Anastasios Kouridis

Supervisors: Dr. Michael Vassilakopoulos

Dr. Dimitrios Bargiotas

Volos, June 2018

Διπλωματική Εργασία για την απόκτηση του Διπλώματος του Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών, του Πανεπιστημίου Θεσσαλίας, στα πλαίσια του Προγράμματος Προπτυχιακών Σπουδών του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Θεσσαλίας.

……

Κουρίδης Αναστάσιος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Πανεπιστημίου Θεσσαλίας

Στην οικογένεια & στους φίλους μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους επιβλέποντες καθηγητές Δρ. Μιχαήλ Βασιλακόπουλο και Δρ. Δημήτριο Μπαργιώτα για τις συμβουλές τους και την καθοδήγηση τους σε αυτή την εργασία.

Ιδιαίτερη αναφορά θα ήθελα να κάνω στον καθηγητή Δρ. Ηλία Χούστη για την εμπιστοσύνη που μου έδειξε κατά τη διάρκεια αυτής της εργασίας και την συνεχή καθοδήγηση του. Μου έδινε πολύτιμες συμβουλές σε κάθε απορία μου και ήταν πάντα διαθέσιμος όποτε τον χρειάστηκα. Είμαι πολύ τυχερός που ήμουν μαθητής του και τον είχα επιβλέποντα καθηγητή στην διπλωματική μου.

Επίσης θέλω να ευχαριστήσω την Δ.Ε.ΤΗ.Π. και συγκεκριμένα τον κ. Κωνσταντίνο Λάζογλου και τον κ. Ηλία Λευκόπουλο για την παροχή των δεδομένων λειτουργίας του δικτύου τηλεθέρμανσης και για την βοήθεια τους πάνω στα δεδομένα.

Τέλος, το μεγαλύτερο ευχαριστώ ανήκει δικαιωματικά στην οικογένεια μου και τους φίλους μου, οι οποίοι με στηρίζουν σε κάθε μου βήμα. Στους γονείς μου, Κώστα και Βάσω, που χάρις στις θυσίες τους όλα αυτά τα χρόνια κατάφερα να πραγματοποιήσω τους στόχους και τα όνειρα μου και γενικότερα, τα πάντα στη ζωή μου.

Σας ευχαριστώ όλους!!!

Κουρίδης Αναστάσιος

Βόλος, Ιούνιος 2018

Περιλήψη

Η αύξηση του πληθυσμού στις πόλεις μεγαλώνει την ενεργειακή ζήτηση η οποία επηρεάζει τις εκπομπές διοξειδίου του άνθρακα στο περιβάλλον. Μεγάλο μέρος της ενέργειας δαπανάται για την θέρμανση εσωτερικών χώρων και του νερού. Τα συστήματα τηλεθέρμανσης βελτιστοποιούν την παραγωγή ενέργειας επαναχρησιμοποιώντας την περίσσια ενέργεια από τις μονάδες συμπαραγωγής θερμότητας και ηλεκτρισμού. Η ακριβής πρόβλεψη των ενεργειακών αναγκών ενός δικτύου τηλεθέρμανσης για μία πόλη έχει όφελος για τους κατοίκους της αλλά και για την εταιρία διαχείρισης του δικτύου. Ο τομέας της επιστήμης δεδομένων με χρήση τεχνικών μηχανικής μάθησης μπορεί να βοηθήσει στην πρόβλεψη της ενεργειακής ζήτησης του δικτύου. Σε αυτή την έρευνα χρησιμοποιήθηκαν δεδομένα λειτουργίας από το δίκτυο τηλεθέρμανσης στην Πτολεμαΐδα το οποίο διαχειρίζεται η Δ.Ε.ΤΗ.Π., καθώς και ιστορικά δεδομένα διάφορων καιρικών παραγόντων από το OpenWeatherMap. Κατασκευάστηκαν μοντέλα με χρήση των τεχνικών γραμμικής παλινδρόμησης, δέντρων αποφάσεων, μηχανών διανυσματικής υποστήριξης και νευρωνικά δίκτυα. Επίσης έγινε μείωση διαστάσεων με ανάλυση κύριων συνιστωσών. Τα δεδομένα χωρίστηκαν σε δύο σύνολα το πρώτο για τους μήνες του χειμώνα και το δεύτερο για τους μήνες του φθινοπώρου και της άνοιξης. Για τα στάδια της ανάλυσης χρησιμοποιήθηκε η γλώσσα προγραμματισμού R και το περιβάλλον ανάπτυξης RStudio. Τα μοντέλα για την περίοδο του χειμώνα είχαν ακρίβεια πρόβλεψης 87.75% ενώ για την περίοδο του φθινοπώρου και άνοιξης 83.82%. Υπήρχε αδυναμία πρόβλεψης για τις ώρες όπου η ζήτηση είχε την μέγιστη τιμή της λόγω της επιρροής του ανθρώπινου παράγοντα, όμως ποτέ η πρόβλεψη δεν ήταν χαμηλότερη από την πραγματική. Ανακαλύφθηκαν επίσης συνήθειες των κατοίκων της πόλης εξετάζοντας την ενεργειακή ζήτηση του δικτύου τηλεθέρμανσης. Τέλος μέσω αυτής της μελέτης αναδείχθηκε η ανάγκη της ανάλυσης δεδομένων σε δίκτυα θερμικής ενέργειας.

Abstract

The growing population in cities increases the energy demand and affects the environment by increasing carbon emissions. Most of the energy consumed in cities is used for residential and commercial heating requirements such as space heating and water heating. District heating systems optimize the energy production by reusing waste energy with combined heat and power plants. An accurate forecast of the energy demand of a district heating network for a city has benefits for both residents and the network's management company. The field of data science with use of machine learning techniques can help predict energy demand on district heating networks. In this thesis, the operational data collected form district heating network in Ptolemaida which is operated by D.E.TI.P., as well as historical data for various weather factors from OpenWeatherMap.org. Prediction models were created using techniques such as linear regression, decision trees, support vector machines and neural networks. Also principal components analysis was used for dimensionality reduction. Data were split into two sets, the first one for the winter months and the second for the autumn and spring months. For the analysis process, programming language R and integrated development environment (IDE) RStudio were used. The proposed models achieved accuracies of 87.75% for winter season and 83.82% for autumn and spring season. There was an inability to predict the hours when demand had its maximum value due to the influence of the human factor, but the predicted value was never lower than the actual. The habits of city residents have also been discovered by looking at the energy demand of the district heating network. Finally, this thesis emerged the need of data analysis in thermal energy networks.

Πίνακας Περιεχομένων

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

ΚΕΦΑΛΑΙΟ 1

1. ΕΙΣΑΓΩΓΗ

1.1. Περιγραφή του προβλήματος

Η ψήφιση όλο και χαμηλότερων ορίων στις εκπομπές διοξειδίου του άνθρακα (CO_2) από τα κράτη μέλη της Ευρωπαϊκής Ένωσης αλλά και από τις υπόλοιπες χώρες του κόσμου, έχει οδηγήσει τις κυβερνήσεις στην παραγωγή ενέργειας από ανανεώσιμες πηγές [1]. Ο λόγος αυτός οδήγησε την Ευρωπαϊκή Ένωση να θέσει τρεις στόχους για το κλίμα και την ενέργεια μέχρι το έτος 2020, ώστε να αποκτήσει μια ενεργειακά αποδοτική και χαμηλών εκπομπών διοξειδίου του άνθρακα οικονομία. Αυτοί οι στόχοι είναι [1].

- μείωση των εκπομπών αερίων θερμοκηπίου της ΕΕ κατά 20% σε σύγκριση με τα επίπεδα εκπομπών του 1990
- αύξηση του μεριδίου της ανανεώσιμης ενέργειας στο 20% της συνολικής κατανάλωσης ενέργειας στην ΕΕ
- βελτίωση της ενεργειακής απόδοσης της ΕΕ κατά 20%

Όμως εκτός από τους στόχους έχουν οριστεί και πρόστιμα για τις χώρες αυτές που δεν βρίσκονται μέσα στα επιτρεπτά όρια εκπομπών διοξειδίου του άνθρακα [2].

Τα κτήρια αποτελούν το 40% της συνολικής κατανάλωσης ενέργειας και το 36% των συνολικών εκπομπών διοξειδίου του άνθρακα [3]. Το μεγαλύτερο μέρος της ενέργειας χρησιμοποιείται για την θέρμανση του χώρου ενώ μικρότερα ποσοστά καταλαμβάνουν η θέρμανση του νερού καθώς και άλλες εργασίες όπως το μαγείρεμα. Ένας από τους πιο αποδοτικούς και οικονομικούς τρόπους θέρμανσης είναι τα δίκτυα τηλεθέρμανσης (district heating networks) [4]. Σε συνδυασμό με τις μονάδες συμπαραγωγής θερμότητας και ηλεκτρισμού (combined heat and power), τα δίκτυα τηλεθέρμανσης χρησιμοποιούν την περισσευούμενη ενέργεια από την διαδικασία παραγωγής ηλεκτρισμού έτσι ώστε να ζεσταίνουν νερό το οποίο κυκλοφορεί στο δίκτυο τηλεθέρμανσης μιας πόλης και θερμαίνει τους χώρους και το νερό των κτηρίων. Με τον τρόπο αυτό μειώνονται η εκπομπές διοξειδίου του άνθρακα καθώς δεν χρειάζεται επιπλέον παραγωγή ενέργειας

για θέρμανση. Υπολογίζεται σε παγκόσμια κλίμακα ότι η τηλεθέρμανση μειώνει της εκπομπές CO_2 κατά 3-4% [5].

Όμως για να υπάρξει αυτό το όφελος από την χρήση των δικτύων τηλεθέρμανσης πρέπει οι μονάδες συμπαραγωγής θερμότητας και ηλεκτρισμού να μην χρησιμοποιούν ρυπογόνες πηγές ενέργειας. Για παράδειγμα μία μονάδα συμπαραγωγής με πηγή ενέργειας τον λιγνίτη θα έχει αυξημένες εκπομπές διοξειδίου του άνθρακα. Άρα η βιωσιμότητα της τηλεθέρμανσης εξαρτάται από την πηγή ενέργειας, όπως επίσης και οι περιοχές στις οποίες μπορούν να αναπτυχθούν τέτοια δίκτυα επειδή η θέρμανση του νερού γίνεται μέσα στης μονάδες συμπαραγωγής και η μεταφορά του σε μακρινές αποστάσεις έχει αποτέλεσμα την πτώση της θερμοκρασίας του, οπότε αχρηστεύετε.

1.2. Κίνητρα

Η ύπαρξη ενός δικτύου τηλεθέρμανσης για κάθε πόλη απαιτεί πολλές μονάδες συμπαραγωγής. Στα πλαίσια μιας χώρα είναι αδύνατον να συμβεί αυτό καθώς είναι ασύμφορο να υπάρχουν τόσες πολλές μονάδες. Μέχρι τώρα οι περισσότερες μονάδες συμπαραγωγής θερμότητας και ηλεκτρισμού χρησιμοποιούν μη ανανεώσιμες πηγές ενέργειας. Οπότε η ύπαρξη τους σε μεγάλες πόλεις θα δημιουργήσει προβλήματα. Όμως ένα δίκτυο τηλεθέρμανσης σε μεγάλη πόλη θα μείωνε κατά πολύ τις εκπομπές CO_2 και τα έξοδα θέρμανσης για κάθε κτήριο. Η χρήση ανανεώσιμων πηγών ενέργειας είναι μία λύση που θα βοηθήσει στην παρουσία της τηλεθέρμανσης σε μεγάλες πόλεις. Στο [6] αναπτύσσεται η ιδέα της 4^{ης} γενιάς τηλεθέρμανσης με χρήση των ευφυών θερμικών δικτύων [7]. Τα ευφυή θερμικά δίκτυα βασίζονται στην ιδέα των ευφυών δικτυών ενέργειας μόνο που στην περίπτωση αυτή η ενέργεια είναι θερμική και όχι ηλεκτρική. Πλέον δεν θα είναι ένας ο πάροχος και πολύ οι καταναλωτές αλλά θα υπάρχει η ιδιότητα του '' prosumer'', δηλαδή ένας καταναλωτής θα μπορεί να παρέχει στο δίκτυο την θερμική ενέργεια που παράγει. Με τον τρόπο αυτό το βάρος παραγωγής θερμικής ενέργειας μοιράζεται. Βέβαια πιο κατάλληλοι για τον ρόλο του '' prosumer'' δεν είναι οι οικίες αλλά μικρές βιομηχανίες ή οποιαδήποτε επιχείρηση μπορεί να διαθέτει έναν μεγάλο λέβητα. Όπως είναι λογικό ο αριθμός αυτών των επιχειρήσεων είναι αυξημένος σε μεγάλες πόλεις. Οπότε μέσω της 4^{ης} γενιάς είναι δυνατόν να αναπτυχθούν δίκτυα τηλεθέρμανσης σε μεγάλες πόλεις.

Πριν γίνει η μετάβαση στην επόμενη γενιά πολλά από τα ήδη υπάρχοντα δίκτυα τηλεθέρμανσης προμηθεύονται ζεστό νερό από μονάδες συμπαραγωγής οι οποίες χρησιμοποιούν πηγές ενέργειας με μεγάλες εκπομπές διοξειδίου του άνθρακα. Η μείωση των εκπομπών από την Ε.Ε αναγκάζει τέτοιου είδους μονάδες να υπολειτουργούν και μετά από καιρό να τερματίζουν την λειτουργία τους.

Για να συνεχίσουν τα δίκτυα τηλεθέρμανσης να υπάρχουν πρέπει να γίνεται σωστή πρόβλεψη των αναγκών τους σε θερμικό φορτίο. Μέχρι τώρα η πρόβλεψη γινόταν με βάση την εμπειρία των υπεύθυνων του δικτύου οι οποίοι αποφάσιζαν ποιες μπορεί να είναι οι ανάγκες των καταναλωτών σε ένα διάστημα για παράδειγμα τριών ημερών. Η τακτική αυτή μπορεί να οδηγήσει σε μεγάλες αποκλίσεις από την πραγματική ζήτηση ενέργειας του δικτύου με αποτέλεσμα είτε το δίκτυο να υπολειτουργεί είτε να έχει περισσότερη ενέργεια από όση χρειάζεται. Η τελευταία περίπτωση μεταφράζεται σε επιβάρυνση της μονάδας συμπαραγωγής θερμότητας και ηλεκτρισμού αλλά και σε επιπλέον έξοδα της εταιρίας που διαχειρίζεται το δίκτυο τηλεθέρμανσης.

Επίσης για την δημιουργία ενός ευφυούς δικτύου είναι αναγκαίο να υπάρχει μία σωστή εκτίμηση της ζήτησης του δικτύου έτσι ώστε να διαμορφώνονται οι τιμές πώλησης και αγοράς. Σε ένα τέτοιο δίκτυο η ύπαρξη αρκετών παραγωγών θερμικής ενέργειας δημιουργεί ένα πρόβλημα πολλών μεταβλητών το οποίο είναι αδύνατον να λυθεί μόνο με την εμπειρία κάποιων ανθρώπων.

Άρα συνοψίζοντας μία σωστή εκτίμηση των ημερήσιων ενεργειακών αναγκών ενός δικτύου τηλεθέρμανσης μπορεί να φέρει καλύτερη αξιοποίηση των ενεργειακών πόρων από της μονάδες παραγωγής ενέργειας, εξοικονόμηση χρημάτων στις εταιρίες διαχείρισης των δικτύων, μεταφορά στην 4^{η} γενιά τηλεθέρμανσης με χρήση ευφυών θερμικών δικτύων και δημιουργία δικτύων τηλεθέρμανσης σε μεγάλες πόλεις με παροχή ενέργειας από ανανεώσιμες πηγές.

1.3. Προσέγγιση προβλήματος

Σκοπός της εργασίας είναι να δώσει λύση στο πρόβλημα αναλύοντας δεδομένα λειτουργίας ενός δικτύου τηλεθέρμανσης με χρήση τεχνικών μηχανικής μάθησης

(machine learning). Η ανάλυση δεδομένων (data analysis) είναι μία διαδικασία από την οποία μπορούν να αναγνωριστούν πρότυπα (pattern recognition) και να δημιουργηθούν μοντέλα με στόχο την εξαγωγή πληροφορίας. Τα μοντέλα είναι ικανά να περιγράψουν την δομή και την συμπεριφορά των χαρακτηριστικών ενός συνόλου δεδομένων. Έτσι χρησιμοποιώντας ιστορικά δεδομένα περιοδικών καταστάσεων μπορεί να εξηγηθεί το παρελθόν, με στόχο να παρθούν αποφάσεις για το μέλλον και να ελεγχθούν υποθέσεις. Με πιο απλά λόγια η ανάλυση δεδομένων και η εξαγωγή συμπερασμάτων είναι σαν έναν άνθρωπο με πολύ καλή μνήμη και τεράστια εμπειρία. Ο παραλληλισμός αυτός έγινε επειδή στην ανάλυση δεδομένων, όπως μαρτυρά και το όνομα της, δεν χρησιμοποιούνται αυστηρά μαθηματικά μοντέλα φτιαγμένα αποκλειστικά για κάποιο φαινόμενο. Αντιθέτως τα μοντέλα είναι πολύ απλά, αλλά οι παράμετροι τους προσδιορίζονται μέσω των δεδομένων έτσι ώστε να προσεγγίζουν ικανοποιητικά το φαινόμενο που πρέπει να εξηγηθεί.

Κάποιες από τις τεχνικές είναι η γραμμική παλινδρόμηση (linear regression), τα δέντρα αποφάσεων (decision trees), οι μηχανές διανυσματικής υποστήριξης (support vector machines) και τα νευρωνικά δίκτυα (neural networks), οι οποίες θα εξηγηθούν και θα χρησιμοποιηθούν για την πρόβλεψη της ενεργειακής ζήτησης ενός δικτύου τηλεθέρμανσης.

1.4. Προκλήσεις και εμπόδια

Η πρόβλεψη ενός φαινομένου εξαρτάται από πολλούς παράγοντες που επηρεάζουν την ποιότητα των μοντέλων και κατ' επέκταση την ακρίβεια των εκτιμήσεων. Ένας από τους πιο σημαντικούς είναι το ίδιο το φαινόμενο. Στις περισσότερες περιπτώσεις τα φαινόμενα στον πραγματικό κόσμο δεν μπορούν να περιγραφούν από μία απλή ευθεία. Επίσης τα διαθέσιμα δεδομένα πρέπει να περιέχουν χρήσιμη πληροφορία και αρκετή έτσι ώστε να βρεθούν τα πραγματικά χαρακτηριστικά εκείνα επηρεάζουν την συμπεριφορά του φαινομένου.

Στο συγκεκριμένο πρόβλημα προσπαθούμε να προβλέψουμε την ζήτηση ενός δικτύου τηλεθέρμανσης εξάγοντας μία μόνο τιμή ανά ώρα για όλο το δίκτυο της πόλης. Τα διαθέσιμα δεδομένα είναι η τιμή αυτή από προηγούμενα έτη λειτουργίας, διάφοροι

καιρικοί παράγοντες και μία χρονοσφραγίδα για κάθε εγγραφή του συνόλου. Όμως η μία τιμή για την ζήτηση του δικτύου συμψηφίζει όλες τις τιμές από κάθε κτήριο της πόλης στις οποίες υπάρχει ο ανθρώπινος παράγοντας. Μία πόλη φυσικά δεν αποτελείται μόνο από ένα είδος κτηρίων. Υπάρχουν μονοκατοικίες, πολυκατοικίες, εμπορικά καταστήματα, γραφεία, χώροι εστίασης, καφετέριες και άλλα. Οπότε είναι λογικό η χρήση της τηλεθέρμανσης σε κάθε τέτοιο χώρο να έχει διαφορετικό μοτίβο. Η διαφορετικότητα συναντάται στις ώρες και μέρες λειτουργίας αλλά και στις κρίσιμες ώρες, δηλαδή όταν η ζήτηση για κάθε κτήριο έχει την μέγιστη τιμή. Επίσης υπάρχουν μέρες όπως οι Κυριακές και οι αργίες στις οποίες το μοτίβο χρήσης δεν ταιριάζει με τις υπόλοιπες. Για παράδειγμα οι άνθρωποι ξυπνάνε πιο αργά, τα καταστήματα είναι κλειστά, στις καφετέριες υπάρχει περισσότερος κόσμος.

Αυτή είναι και η μεγαλύτερη πρόκληση σε αυτή την έρευνα. Να προσδιοριστούν με ακρίβεια οι τιμές ζήτησης στις κρίσιμες ώρες μέσω μίας μόνο τιμής για όλο το δίκτυο, με τέτοιο τρόπο ώστε οι προβλέψεις να μην αναγκάζουν το δίκτυο να υπολειτουργεί αλλά και μην διαθέτει περισσότερο θερμικό φορτίο από όσο χρειάζεται.

1.5. Στόχοι

Ο κύριος στόχος της έρευνας αναφέρθηκε παραπάνω. Όμως τις περισσότερες φορές δεν γίνεται να λύσεις ένα τέτοιο πρόβλημα μόνο με μία προσπάθεια. Οπότε πέρα από το κύριο πρόβλημα η έρευνα προσπαθεί να αναδείξει την σημαντικότητα των δεδομένων και τα οφέλη της ανάλυσης και εξόρυξης πληροφορίας σε χώρους και τομείς που μέχρι τώρα δεν αξιοποιούνταν σε μεγαλύτερο ποσοστό.

ΚΕΦΑΛΑΙΟ 2

2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ ΚΑΙ ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Σε αυτό το κεφάλαιο γίνεται αναφορά στα δεδομένα που χρησιμοποιήθηκαν, από που προήλθαν και τι χαρακτηριστικά περιέχουν. Επίσης υπάρχει μία βιβλιογραφική ανασκόπηση σε παρόμοιες μελέτες προηγούμενων ετών και τέλος γίνεται περιγραφή των συστημάτων τηλεθέρμανσης.

2.1. Δεδομένα

Το δίκτυο τηλεθέρμανσης που επιλέχθηκε είναι στην πόλης της Πτολεμαΐδας στην Δυτική Μακεδονία. Αρμόδια για την διαχείριση του δικτύου είναι η Δημοτική Επιχείρηση Τηλεθέρμανσης Πτολεμαΐδας (Δ.Ε.ΤΗ.Π.) [8]. Η μονάδα παραγωγής ενέργειας του δικτύου είναι οι μονάδες συμπαραγωγής θερμότητας και ηλεκτρισμού του ατμοηλεκτρικού σταθμού (ΑΗΣ) Καρδιάς. Επίσης η επιχείρηση διαθέτει ένα λεβητοστάσιο για την κάλυψη των αναγκών του δικτύου στις ώρες αιχμής. Το καύσιμο του λέβητα είναι πετρέλαιο. Τα δεδομένα που χρησιμοποιήθηκαν στην ανάλυση προήλθαν από δύο πήγες. Η μία είναι η Δ.Ε.ΤΗ.Π. και συγκεκριμένα από το σύστημα SCADA που διαθέτει για την αυτοματοποιημένη λειτουργία και επίβλεψη του δικτύου της ενώ η δεύτερη πηγή είναι η ιστοσελίδα OpenWeatherMap [9] η οποία παρέχει ιστορικά δεδομένα καιρού για διάφορες τοποθεσίες ανά τον κόσμο.

Τα δεδομένα είναι από το έτος λειτουργίας 2012-2013 έως το έτος 2017-2018 και αφορούν τους μήνες Οκτώβριο μέχρι Απρίλιο. Από τα δεδομένα λείπουν οι μήνες Μάρτιος και Απρίλιος του 2018 εξαιτίας της χρονικής περιόδου που συλλέχθηκαν. Με βάση αυτές τις ημερομηνίες αγοράστηκαν τα δεδομένα καιρού από το OpenWeatherMap. Οι πληροφορίες που παρέχει το OpenWeatherMap είναι ωριαίες μετρήσεις για στοιχεία όπως ατμοσφαιρική πίεση, υγρασία, ταχύτητα ανέμου, κατεύθυνση ανέμου και νέφωση. Ενώ η Δ.Ε.ΤΗ.Π. μέσω του SCADA και των αισθητήρων της παρέχει μέση ωριαία τιμή εξωτερικής θερμοκρασίας, βαθμοημέρες, μέση ωριαία ζήτηση ενέργειας του δικτύου της πόλης και μέση ωριαία παροχή ενέργειας από τους

ατμοηλεκτρικούς σταθμούς (ΑΗΣ). Για κάθε μία εγγραφή του συνόλου των δεδομένων υπάρχει και η αντίστοιχη χρονοσφραγίδα (timestamp). Στον πίνακα 1 φαίνονται αναλυτικά τα στοιχεία των δύο πηγών.

Πίνακας 1. Στοιχεία πηγών δεδομένων

Στοιχείο	**Αναγνωριστικό (ID)**	**Μονάδα μέτρησης**	**Πηγή**
Χρονοσφραγίδα	datetime		Δ.Ε.ΤΗ.Π.
Εξ. Θερμοκρασία	deg	oC	Δ.Ε.ΤΗ.Π.
Βαθμοημέρες	degday	Καθαρός αριθμός	Δ.Ε.ΤΗ.Π.
Ζήτηση δικτύου	dhn	MWh	Δ.Ε.ΤΗ.Π.
Παροχή ΑΗΣ	chp	MWh	Δ.Ε.ΤΗ.Π.
Ατμ. Πίεση	pressure	hPa	OpenWeatherMap
Υγρασία	humidity	%	OpenWeatherMap
Ταχ. Ανέμου	wind_speed	m/s	OpenWeatherMap
Κατ. Ανέμου	wind_deg	Μοίρες (o)	OpenWeatherMap
Νέφωση	clouds_all	%	OpenWeatherMap

2.2. Βιβλιογραφική ανασκόπηση

Στο [10] γίνεται πρόβλεψη των ενεργειακών αναγκών διάφορων κτηρίων με χρήση τεχνητών νευρωνικών δικτύων (artificial neural networks - ANNs). Τα δεδομένα αφορούν χαρακτηριστικά των κτηρίων και τροφοδοτούνται σε ένα νευρωνικό δίκτυο που χρησιμοποιεί τον αλγόριθμο ανάστροφης διάδοσης σφάλματος (backpropagation) για την ελαχιστοποίηση των σφαλμάτων. Το δίκτυο πετυχαίνει ικανοποιητικά αποτελέσματα με ποσοστό επιτυχής πρόβλεψης 94.8% - 98.5%. Στο [11] επισημαίνεται η ανάγκη να υπάρχουν τα κατάλληλα δεδομένα έτσι ώστε να είναι ακριβής η πρόβλεψη του θερμικού φορτίου σε δίκτυο τηλεθέρμανσης. Για αυτό αναπτύσσεται ένας πολυεπίπεδος αισθητήρας δύο σταδίων με σκοπό την αυτόματη δημιουργία μοντέλων πρόβλεψης. Έτσι θα είναι δυνατή η βελτίωση του μοντέλου όσο δημιουργούνται νέα δεδομένα. Στο [12] προτείνεται ένα ανατροφοδοτούμενο νευρωνικό δίκτυο (recurrent neural network) το

οποίο είναι ικανό να μοντελοποιήσει τις περιόδους όπου η ζήτηση του δικτύου δεν έχει σταθερή μεταβολή. Σε αντίθεση με τα υπόλοιπα νευρωνικά δίκτυα που μέχρι εκείνη την στιγμή αδυνατούν να διαχειριστούν τέτοιες περιόδους. Στο [13] χρησιμοποιείται ένα νευρωνικό δίκτυο βασισμένο στο μοντέλο του Elman (ENN) για να βρεθούν οι παράγοντες με την μεγαλύτερη επιρροή στην ενεργειακή ζήτηση ενός δικτύου τηλεθέρμανσης. Εξετάζονται χρονικοί παράγοντες καθώς και καιρικοί όπως εξωτερική θερμοκρασία, ηλιακή ακτινοβολία και ταχύτητα ανέμου. Τα αποτελέσματα έδειξαν πως η εισαγωγή της ταχύτητας του ανέμου βελτιώνει την ακρίβεια πρόβλεψης του μοντέλου όμως η ταυτόχρονη παρουσία της ηλιακής ακτινοβολίας και ταχύτητας ανέμου στο σύνολο δεδομένων έχουν αρνητική επιρροή στην απόδοση του νευρωνικού δικτύου. Στο [14] χρησιμοποιούνται δεδομένα λειτουργίας από το δίκτυο τηλεθέρμανσης της Sueso και αναπτύσσονται μοντέλα με τις μεθόδους των μερικών ελαχίστων τετραγώνων, μηχανών διανυσματικής υποστήριξης για παλινδρόμηση και τεχνητών νευρωνικών δικτύων. Η περίοδος πρόβλεψης ήταν μία εβδομάδα του Φεβρουαρίου και τα δεδομένα που χρησιμοποιήθηκαν ήταν προηγούμενες τιμές κατανάλωσης και η θερμοκρασία του χώρου. Την καλύτερη συνολική επίδοση με τα μικρότερα σφάλματα πέτυχαν οι μηχανές διανυσματικής υποστήριξης με τα ποσοστά λάθους να κυμαίνονται από 3.98% έως 8.81%.

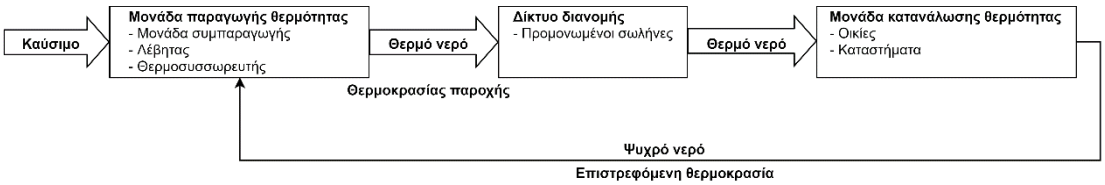
Στο [15] αναπτύχθηκε ένας online αλγόριθμος μηχανικής μάθησης για την πρόβλεψη του θερμικού φορτίου ενός δικτύου τηλεθέρμανσης. Το δίκτυο που χρησιμοποιήθηκε ήταν της πόλης Karlshamn στην Σουηδία. Η επιλογή του αλγορίθμου έγινε λόγω των πλεονεκτημάτων του. Έχουν μεγάλη υπολογιστική απόδοση και μπορούν να χειριστούν τυχόν αλλαγές στην συμπεριφορά της μεταβλητής πρόβλεψης. Το βασικό μοντέλο που χρησιμοποιήθηκε είναι το Fast Incremental Model Trees με Drift Detection (FIMT-DD). Το μέσο απόλυτο ποσοστιαίο σφάλμα του αλγορίθμου ήταν 4.77%. Στο [16] χρησιμοποιούνται 4 τεχνικές επιβλεπόμενης μάθησης (supervised learning), μηχανές διανυσματικής υποστήριξης για παλινδρόμηση (SVR), δέντρα αποφάσεων (decision trees), νευρωνικά δίκτυα πρόσθιας τροφοδότησης (feed forward neural networks - FFNN) και πολλαπλή γραμμική παλινδρόμηση (MLR). Τα δεδομένα συλλέχθηκαν από υποσταθμούς του δικτύου τηλεθέρμανσης της πόλης Skelleftea στην Σουηδία. Έγιναν προβλέψεις για μελλοντικά ωριαία διαστήματα έως 24 ωρών. Τα αποτελέσματα έδειξαν

ότι τα SVR πέτυχαν το μικρότερο μέσο απόλυτο ποσοστιαίο σφάλμα με τιμή 5.6%. Για το ίδιο δίκτυο τηλεθέρμανσης στο [17] κατασκευάστηκε ένα δίκτυο Bayes. Το τελικό μοντέλο είχε ακρίβεια πρόβλεψης της ζήτησης του δικτύου 76.74%. Στο [18] με χρήση δεδομένων από την εταιρία Eidsiva Bioenergi AS η οποία διαχειρίζεται το δίκτυο της Hamar στην Νορβηγία εξετάζονται 3 μοντέλα μηχανικής μάθησης. Αυτά είναι με την μέθοδο των μηχανών διανυσματικής υποστήριξης για παλινδρόμηση, μερικών ελαχίστων τετραγώνων (PLS) και τυχαία δάση (random forests). Τα δεδομένα αντιστοιχούν σε μετρήσεις κτηρίων από διάφορες περιοχές για μια περίοδο 29 εβδομάδων. Ως μετρικές αξιολόγησης χρησιμοποιούνται το μέσο απόλυτο σφάλμα (MAE) και το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE). Η τιμή του MAPE για τα SVR είναι 5.54% ενώ για τα PLS είναι 8.99%.

Στο [19] αναφέρονται τα οφέλη της ανάλυσης δεδομένων στην ανάπτυξη ευφυών ενεργειακών δικτύων. Η εγκατάσταση έξυπνων αισθητήρων παρέχει ένα μεγάλο όγκο δεδομένων για διαφορετικά χρονικά διαστήματα από τα οποία για να εξαχθεί πληροφορία, όπως να κατηγοριοποιηθούν τα δεδομένα, να γίνει πρόβλεψη της ζήτησης, να βελτιστοποιηθεί η παραγωγή ενέργειας, να εξαχθούν πολιτικές τιμολόγησης και να παρακολουθείτε το δίκτυο για τυχόν βλάβες, απαιτείται η χρήση τεχνικών ανάλυσης δεδομένων. Στο [20] χρησιμοποιείται συνδυασμός online αλγορίθμων μηχανικής μάθησης (ensembles of online machine learning algorithms) για την πρόβλεψη της ενεργειακής ζήτησης σε δίκτυο τηλεθέρμανσης. Τα δεδομένα αναφέρονται στους μήνες Ιανουάριο, Φεβρουάριο και Μάρτιο του 2016 για το δίκτυο της Rottne στην Σουηδία. Οι αποδόσεις των αλγορίθμων για αυτή την περίοδο είναι 73.24 για το MAE και 11.7% για το MAPE.

2.3. Συστήματα τηλεθέρμανσης

Τα συστήματα τηλεθέρμανσης έχουν δείξει ότι μπορούν να μειώσουν τις εκπομπές διοξειδίου του άνθρακα CO_2 και να αξιοποιήσουν στο έπακρο την διαθέσιμη ενέργεια [5, 21]. Η παραγωγή και η διανομή της θερμότητας από μία κεντρική μονάδα και όχι από πολλές, για παράδειγμα κάθε κτήριο μίας πόλης, έχει οικολογικά και οικονομικά οφέλη [21]. Τα συστήματα τηλεθέρμανσης αποτελούνται από 3 μέρη όπως φαίνεται στο Σχήμα 1.



Σχήμα 1. Διάγραμμα τηλεθέρμανσης [21]

2.3.1. Μονάδα παραγωγής θερμότητας

Η μονάδα παραγωγής θερμότητας αποτελεί την κύρια πηγή ενέργειας. Η παραγωγή μπορεί να γίνεται από έναν λέβητα ή έναν θερμοσυσσωρευτή ή από μία μονάδα συμπαραγωγής θερμότητας και ηλεκτρισμού (combined heat and power - CHP). Η πηγή ενέργειας της μονάδας μπορεί να είναι ορυκτά καύσιμα όπως λιγνίτης και φυσικό αέριο, γεωθερμική ενέργεια, βιομάζα ή κάποιος συνδυασμός των προηγούμενων.

2.3.2. Δίκτυο διανομής

Το δίκτυο διανομής αποτελείται από μονωμένους και προμονωμένους σωλήνες οι οποίοι μεταφέρουν ζεστό νερό και τροφοδοτούν κατοικίες, καταστήματα και οποιοδήποτε άλλο κτήριο που είναι συνδεδεμένο στο δικτύου. Οι σωλήνες επιστροφής μεταφέρουν το κρύο νερό, που προκύπτει από την θέρμανση των εσωτερικών χώρων, από τις μονάδες κατανάλωσης στην μονάδα παραγωγής.

2.3.3. Μονάδα κατανάλωσης θερμότητας

Η μονάδα κατανάλωσης θερμότητας είναι κάθε ένα κτήριο όπου απαιτείται θέρμανση. Κάθε κτίριο διαθέτει υποσταθμούς μέσω των οποίων γίνεται η διανομή του ζεστού νερού σε διαφορετικούς χρήστες του κτηρίου. Οι υποσταθμοί είναι εξοπλισμένοι με εναλλάκτες θερμότητας, οι οποίοι μεταφέρουν θερμότητα από τους προμονωμένους σωλήνες στο εσωτερικό δίκτυο θέρμανσης του κτηρίου. Οι εναλλάκτες θερμότητας χρησιμοποιούν το ζεστό νερό για να θερμάνουν τα θερμαντικά σώματα και το νερό των κτηρίων. Το κρύο νερό από το κτήριο επιστρέφεται στη συνέχεια στον σταθμό

τηλεθέρμανσης για να αναθερμανθεί. Το σύνολο του δικτύου σχηματίζει έναν κλειστό βρόγχο.

ΚΕΦΑΛΑΙΟ 3

3. ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Σε αυτό το κεφάλαιο περιγράφεται η έννοια της μηχανικής μάθησης και τεχνικών που χρησιμοποιούνται σε αυτήν. Αναλύεται επίσης η θεωρία πίσω από τις τεχνικές με τις οποίες θα κατασκευαστούν τα μοντέλα πρόβλεψης και αναφέρονται μετρικές με τις οποίες αξιολογείται η απόδοση τους.

3.1. Μηχανική μάθηση

Η μηχανική μάθηση (machine learning) ανήκει στο πεδίο της επιστήμης των υπολογιστών (computer science) και αναφέρεται στην ικανότητα των υπολογιστών συστημάτων να ‘’μαθαίνουν’’ χωρίς να έχουν προγραμματιστεί με ρητές εντολές. Μέσω μαθηματικών μοντέλων και τεχνικές στατιστικής μπορούν να ερμηνεύσουν φαινόμενα χρησιμοποιώντας δεδομένα που έχουν συλλεχθεί σχετικά με αυτά. Έτσι τα συστήματα αυτά είναι σε θέση να προβλέψουν μελλοντικές καταστάσεις με μικρό ποσοστό λάθους. Η συνεχής τροφοδοσία τους με νέα δεδομένα τα βοηθάει να βελτιώνουν τις προβλέψεις τους. Για τον λόγο αυτό χαρακτηρίζονται ως ‘’έξυπνα’’ και πολλές φορές η έννοια της μηχανικής μάθησης συνδέεται με αυτή της τεχνητής νοημοσύνης (artificial intelligence).

Η μηχανική μάθηση μπορεί να χωριστεί σε δύο μεγάλες κατηγορίες, την επιβλεπόμενη μάθηση (supervised learning) και την μη επιβλεπόμενη μάθηση (unsupervised learning). Στην πρώτη περίπτωση τα δεδομένα αποτελούνται από ζεύγη εισόδου-εξόδου. Κάθε παρατήρηση (observation) είναι ένα διάνυσμα από ανεξάρτητες μεταβλητές (είσοδος) και μία εξαρτημένη (έξοδος). Έτσι οι μέθοδοι επιβλεπόμενης μάθησης έχουν στην διάθεση τους τις τιμές εξόδου για κάθε παρατήρηση, με τις οποίες μπορούν να συγκρίνουν τα αποτελέσματα των μοντέλων πρόβλεψης που κατασκεύασαν και να τα βελτιώσουν. Στην περίπτωση της μη επιβλεπόμενης μάθησης δεν υπάρχει η τιμή της εξαρτημένης μεταβλητής. Στην κατηγορία αυτή οι μέθοδοι που χρησιμοποιούνται δεν σημαίνει ότι προσπαθούν να μαντέψουν την τιμή της εξόδου αλλά έχουν το χαρακτηριστικό της εξόρυξης πληροφορίας (data mining). Ανακαλύπτουν

μοτίβα μεταξύ των δεδομένων, όπως για παράδειγμα κοινά χαρακτηριστικά που τα κατατάσσουν σε ομάδες, και τα μετατρέπουν σε πιο ωφέλιμες μορφές για επόμενα στάδια της ανάλυσης, π.χ. μείωση διαστάσεων.

Για την δημιουργία ενός αξιόπιστου μοντέλου πρόβλεψης πρέπει να γίνει σωστή διαχείριση των διαθέσιμων δεδομένων. Πολλές φορές ο όγκος των δεδομένων δεν είναι αρκετός ή μπορεί να μην συλλεχθούν νέα δεδομένα για αρκετό καιρό έτσι ώστε να ελεγχθεί η απόδοση του μοντέλου πρόβλεψης. Για τους λόγους αυτούς υπάρχουν τεχνικές μέσω των οποίων παράγονται μοντέλα αξιόπιστα τα οποία έχουν υψηλή απόδοση ακόμα και σε νέα άγνωστα δεδομένα εισόδου.

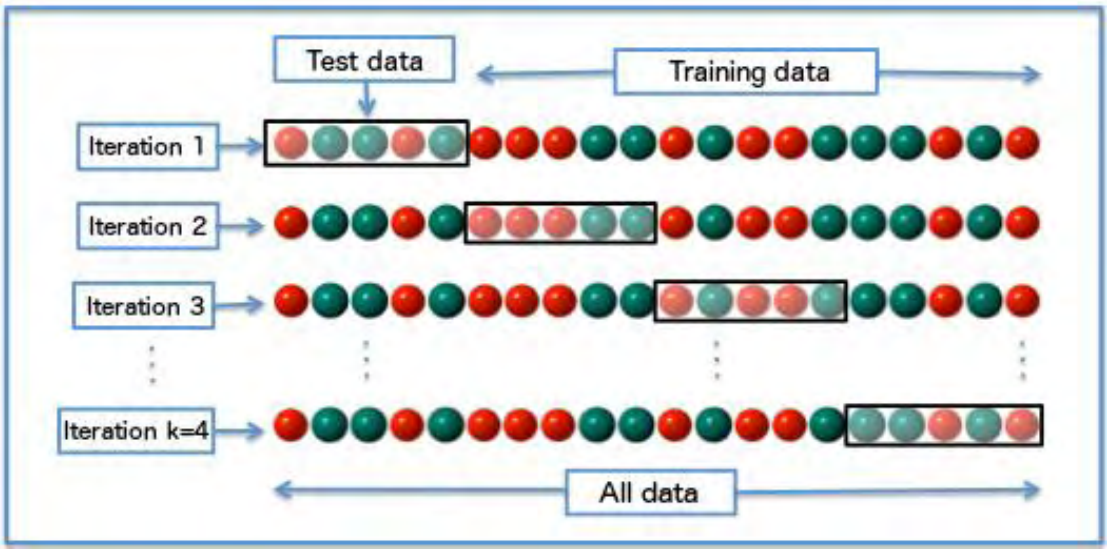
3.2. Σύνολο εκπαίδευσης και ελέγχου

Μία καλή πρακτική πριν την κατασκευή ενός μοντέλου είναι η διάσπαση του συνόλου των δεδομένων σε σύνολο εκπαίδευσης και ελέγχου (train and test set). Μία καλή αναλογία διάσπασης είναι 75/25 με το σύνολο εκπαίδευσης να είναι το 75% και το σύνολο ελέγχου το 25%. Φυσικά χρησιμοποιούνται και άλλες αναλογίες όμως πάντα το σύνολο εκπαίδευσης είναι αρκετά μεγαλύτερο από εκείνο του ελέγχου. Το σύνολο εκπαίδευσης χρησιμοποιείται για να δημιουργηθεί, ‘’εκπαιδευτεί’’ το μοντέλο πρόβλεψης ενώ το σύνολο ελέγχου για να εξεταστεί η απόδοση του μοντέλου σε άγνωστα δεδομένα. Έτσι έχοντας ως γνώμονα την απόδοση σε γνωστά και κυρίως σε άγνωστα δεδομένα μπορούμε να βελτιώσουμε το μοντέλο πρόβλεψης.

3.3. Cross-Validation

Λόγω του όγκου των δεομένων οι περισσότερες τεχνικές μηχανικής μάθησης δεν μπορούν να υπολογίσουν την ακριβή λύση των παραμέτρων ενός μοντέλου. Έτσι ξεκινώντας από τυχαίες τιμές για τους συντελεστές των παραμέτρων καταλήγουν μέσω ενός επαναληπτικού αλγορίθμου σε μια προσεγγιστική λύση. Όμως η λύση αυτή πολλές φορές εξαρτάται από το πως θα μοιραστούν τα δεδομένα σε σύνολο εκπαίδευσης και σύνολο ελέγχου με αποτέλεσμα να επηρεάζεται η απόδοση του μοντέλου. Το πρόβλημα αυτό μπορεί να λυθεί με την μέθοδο του cross-validation. Συγκεκριμένα θα αναφερθούμε στο k-fold cross-validation.

Το αρχικό σύνολο εκπαίδευσης σπάει σε *k* ίσα κομμάτια. Από τα *k* κομμάτια κάθε φορά το ένα από αυτά χρησιμοποιείται για να ελεγχθεί η απόδοση του μοντέλου που κατασκευάστηκε από τα υπόλοιπα *k-1* κομμάτια. Η διαδικασία αυτή επαναλαμβάνεται *k* φορές με το κομμάτι ελέγχου να είναι διαφορετικό σε κάθε επανάληψη, όπως φαίνεται στο Σχήμα 2. Στο τέλος τα μοντέλα συμψηφίζονται και δημιουργείται ένα τελικό μοντέλο. Μια τιμή που χρησιμοποιείται αρκετά συχνά για το *k* είναι το 10. Φυσικά μπορούν να οριστούν και άλλες τιμές.

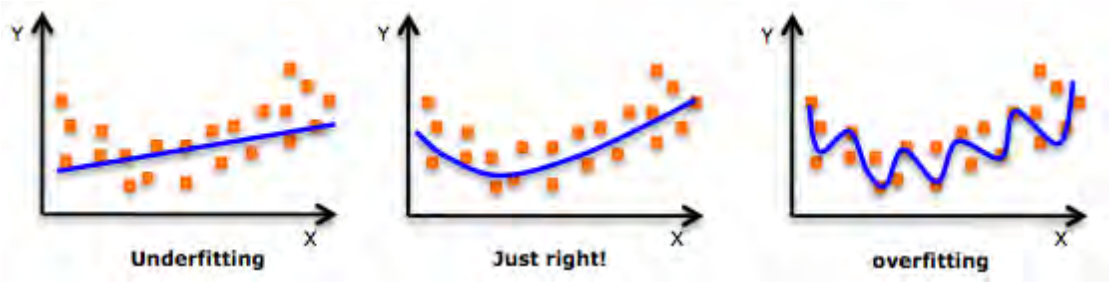


Σχήμα 2. Διαδικασία Cross-Validation [22]

3.4. Υπερπροσαρμογή και υποπροσαρμογή

Προηγουμένως έγινε αναφορά στο πόσο σημαντικό είναι ένα μοντέλο πρόβλεψης να αποδίδει εξίσου καλά σε άγνωστα δεδομένα. Πολλές φορές τα μοντέλα που δημιουργούνται είναι υπερπροσαρμοσμένα (overfitting) στα δεδομένα εκπαίδευσης. Αυτό σημαίνει ότι αποδίδουν σχεδόν βέλτιστα στις παρατηρήσεις από τις οποίες κατασκευάστηκαν αλλά η απόδοση τους πέφτει κατακόρυφα σε άγνωστες παρατηρήσεις. Αυτό είναι το φαινόμενο της υπερπροσαρμογής (overfitting). Όταν το μοντέλο δεν έχει ικανοποιητική απόδοση ούτε στα δεδομένα εκπαίδευσης τότε έχουμε το φαινόμενο της υποπροσαρμογής (underfitting).

Εφόσον ξεπεραστεί το φαινόμενο της υποπροσαρμογής, δηλαδή φτιαχτεί ένα μοντέλο με καλή απόδοση στα δεδομένα εκπαίδευσης, τότε το πιο σημαντικό είναι να μην υπάρχει το φαινόμενο της υπερπροσαρμογής. Αυτό συμβαίνει όταν το μοντέλο εκτός από την χρήσιμη πληροφορία ''εξηγεί'' και τον θόρυβο που υπάρχει ανάμεσα στα δεδομένα, Σχήμα 3.



Σχήμα 3. Παράδειγμα υποπροσαρμογής και υπερπροσαρμογής [23]

3.5. Μετρικές αξιολόγησης μοντέλων

Η απόδοση ενός μοντέλου πρόβλεψης είναι ένας από τους σημαντικότερους παράγοντες για την αξιολόγηση του. Στην επιβλεπόμενη μάθηση η μέτρηση της απόδοσης γίνεται συγκρίνοντας τις πραγματικές τιμές της εξαρτημένης μεταβλητής με εκείνες που παρήχθησαν από το μοντέλο πρόβλεψης. Η απόδοση του μοντέλου κρίνεται από το πόσο μικρό είναι το σφάλμα του, δηλαδή τις διαφορές μεταξύ των πραγματικών τιμών εξόδου με αυτές του μοντέλου. Υπάρχουν διάφορες μετρικές, κάποιες από αυτές είναι το μέσο τετραγωνικό σφάλμα (MSE), η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), το μέσο απόλυτο σφάλμα (MAE) και το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE). Στα πρώτα τρία η μονάδα μέτρησης είναι ίδια με εκείνη της εξαρτημένης μεταβλητής ενώ στο τελευταίο το σφάλμα μετριέται σε ποσοστό επί τις εκατό. Οι συναρτήσεις αυτές ονομάζονται επίσης συναρτήσεις κόστους. Εκτός από την αξιολόγηση ενός μοντέλου πρόβλεψης χρησιμοποιούνται και στην βελτιστοποίηση του. Οι μέθοδοι μηχανικής μάθησης που χρησιμοποιούν προσεγγιστικές λύσεις πετυχαίνουν καλύτερες αποδόσεις στα μοντέλα πρόβλεψης προσπαθώντας σε κάθε επανάληψή τους να ελαχιστοποιήσουν αυτές τις συναρτήσεις. Η επιλογή της μετρικής εξαρτάται από το φαινόμενο που πρέπει να ερμηνευθεί καθώς και από τις προτιμήσεις του χρήστη.

Στις παρακάτω εκφράσεις, όπου n είναι το πλήθος των παρατηρήσεων, y_i η πραγματική τιμή της εξαρτημένης μεταβλητής και $\hat{y}_i$ η τιμή της εξαρτημένης μεταβλητής που υπολογίστηκε από το μοντέλο πρόβλεψης.

3.5.1. Μέσο τετραγωνικό σφάλμα

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

3.5.2. Ρίζα μέσου τετραγωνικού σφάλματος

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

3.5.3. Μέσο απόλυτο σφάλμα

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

3.5.4. Μέσο απόλυτο ποσοστιαίο σφάλμα

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

3.6. Γραμμική Παλινδρόμηση

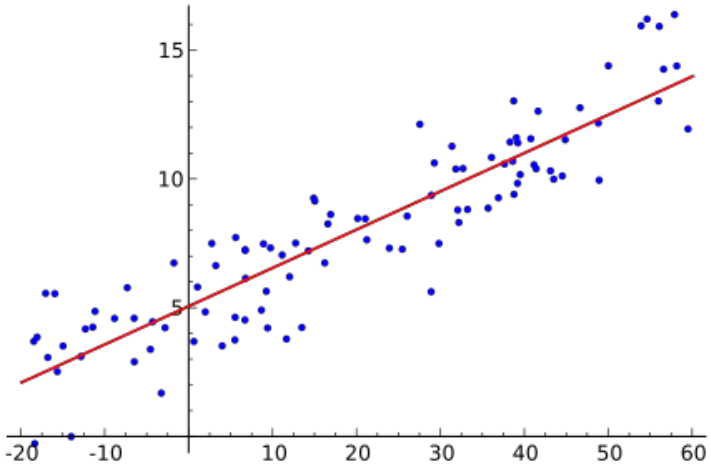
Η γραμμική παλινδρόμηση είναι ένας γραμμικός τρόπος μοντελοποίησης της σχέσης μεταξύ μιας συνεχής εξαρτημένης μεταβλητής και ενός συνόλου ανεξάρτητων μεταβλητών. Όταν υπάρχει μία ανεξάρτητη μεταβλητή η τεχνική ονομάζεται απλή γραμμική παλινδρόμηση ενώ όταν οι ανεξάρτητες μεταβλητές είναι παραπάνω από μία ονομάζεται πολλαπλή γραμμική παλινδρόμηση.

Στην περίπτωση της απλής γραμμικής παλινδρόμησης υπάρχει μία εξαρτημένη μεταβλητή (*y*) και μία ανεξάρτητη (*x*), οπότε κάθε παρατήρηση μπορεί να αναπαρασταθεί από ένα σημείο στον δισδιάστατο χώρο με τον συμβολισμό (y_i, x_i). Σκοπός της μεθόδου είναι να περιγράψει το σύνολο των δεδομένων μέσω μιας ευθείας,

Σχήμα 4. Υποθέτει δηλαδή πως υπάρχει μια γραμμική σχέση μεταξύ της εξαρτημένης και της ανεξάρτητης μεταβλητής. Η ευθεία που περιγράφει πως το *y* σχετίζεται με το *x* είναι:

$$y = \beta_0 + \beta_1 x$$

Στην παραπάνω εξίσωση το β_0 είναι το σημείο που η ευθεία τέμνει τον κάθετο άξονα και το β_1 είναι η κλίση της ευθείας. Εάν η τιμή του β_1 είναι πολύ κοντά στο μηδέν υποδεικνύει χαμηλή έως καθόλου γραμμική σχέση ενώ αν είναι μία μεγάλη θετική ή αρνητική τιμή τότε υπάρχει μεγάλη γραμμική εξάρτηση. Οι δύο αυτές σταθερές ονομάζονται συντελεστές ή παράμετροι της ευθείας παλινδρόμησης. Για να υπολογίσουμε την ευθεία θα δημιουργήσουμε ένα πιθανοτικό μοντέλο μέσω του οποίου θα αξιολογήσουμε πόσο ‘’καλά’’ περιγράφει τα δεδομένα.



Σχήμα 4. Απλή γραμμική παλινδρόμηση [24]

3.6.1. Πιθανοτικό μοντέλο για γραμμικά εξαρτώμενα δεδομένα

Έστω ότι έχουμε τα ζεύγη δεδομένων $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, υποθέτουμε ότι κάθε y_i αποτελεί συνάρτηση του x_i και υπολογίζεται από την ευθεία $y = \beta_0 + \beta_1 x$ στην οποία προσθέτουμε θόρυβο για να αναπαραστήσουμε την απόκλιση του σημείου από την ευθεία. Οπότε το πιθανοτικό μοντέλο για την αναπαράσταση των δεδομένων είναι

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Η μοντελοποίηση του θορύβου γίνεται μέσω μιας γκαουσιανής κατανομής με μέση τιμή μηδέν και σταθερή διασπορά, $\varepsilon \sim N(0, \sigma^2)$.

Για τον υπολογισμό των παραμέτρων της ευθείας θα χρησιμοποιηθεί η μέθοδος των ελαχίστων τετραγώνων (least squares method). Υποθέτοντας ότι τα δεδομένα δημιουργήθηκαν μέσω του παραπάνω πιθανοτικού μοντέλου, η ευθεία που τα περιγράφει καλύτερα θα είναι εκείνη με το μικρότερο άθροισμα τετραγώνων των σφαλμάτων. Η μαθηματική μορφή των παραπάνω είναι

$$\min_{\beta_0,\beta_1} : \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Αυτό είναι γνωστό και σαν πρόβλημα ελαχίστων τετραγώνων γραμμικής παλινδρόμησης. Έχοντας ένα σύνολο δεδομένων η λύση είναι

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

όπου $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ είναι η μέση τιμή των σημείων y και x αντίστοιχα.

3.6.2. Πολλαπλή γραμμική παλινδρόμηση

Στην προηγούμενη ενότητα ορίστηκε το πρόβλημα της απλής γραμμικής παλινδρόμησης και έγινε αναφορά στην διαδικασία εύρεσης των παραμέτρων της ευθείας. Επειδή η μέθοδος που θα χρησιμοποιηθεί στα επόμενα κεφάλαια θα είναι η πολλαπλή γραμμική παλινδρόμηση επιλέχθηκε να αναλυθούν μόνο σε αυτή την ενότητα οι μετρικές αξιολόγησης του μοντέλου και οι τεχνικές εκτίμησης της σημαντικότητας των παραμέτρων της ευθείας.

Αντίθετα με την απλή γραμμική παλινδρόμηση όπου υπήρχε μία ανεξάρτητη μεταβλητή (x) στην περίπτωση της πολλαπλής υπάρχει ένα διάνυσμα $(x_1, \dots, x_p)$ για κάθε εγγραφή *i* στο σύνολο των δεδομένων. Οπότε για κάθε εγγραφή υπάρχουν *p* διαφορετικές ανεξάρτητες μεταβλητές ή μεταβλητές πρόβλεψης ή χαρακτηριστικά. Η λέξη χαρακτηριστικά χρησιμοποιείται συχνά για να αναφερθεί κανείς στις ανεξάρτητες μεταβλητές επειδή πολλές φορές περιέχουν πληροφορία για τα χαρακτηριστικά της

εξαρτημένης μεταβλητής. Για παράδειγμα αν η εξαρτημένη μεταβλητή αναφέρετε σε ανθρώπους τότε οι ανεξάρτητες μπορεί να είναι το ύψος, το βάρος και το άνοιγμα των χεριών τους ή μπορεί να είναι μοντέλα αυτοκινήτων και τα χαρακτηριστικά τους να είναι τα κυβικά, η τελική ταχύτητα και το βάρος τους.

Στην περίπτωση της πολλαπλής παλινδρόμησης η γραμμική συνάρτηση που θα χρησιμοποιηθεί δεν θα είναι ευθεία, αν για παράδειγμα το *p=2* τότε θα ήταν ένα επίπεδο. Η γενική μορφή της συνάρτησης μέσω της οποίας θα προσπαθήσουμε να προβλέψουμε το *y* είναι:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

3.6.3. Πιθανοτικό μοντέλο

Για το ορισμό του πιθανοτικού μοντέλου θα χρησιμοποιηθούν πίνακες και διανύσματα για μια πιο συμπαγής αναπαράσταση του. Τα δεδομένα εισόδου $(x_1, \dots, x_p)$ θα έχουν την μορφή πίνακα (X). Ο πίνακας X έχει διαστάσεις *n x p* όπου κάθε γραμμή είναι μία εγγραφή και κάθε στήλη αντιστοιχεί σε ένα χαρακτηριστικό. Καθώς κάθε αποτέλεσμα y_i είναι αριθμός όλα τα αποτελέσματα μαζί θα αποτελούν ένα στηλοδιάνυσμα μεγέθους *n*. Έτσι το πιθανοτικό μοντέλο μπορεί να εκφραστεί ως

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon$$

όπου το β είναι ένα διάνυσμα p θέσεων με τους συντελεστές του μοντέλου και το ε είναι ένα διάνυσμα μεγέθους n όπου κάθε στοιχείο του ε_i αναπαριστά τον θόρυβο μέσω μιας γκαουσιανής κατανομής με μηδενική μέση τιμή και σταθερή διασπορά. Παρατηρήστε πως στην πολλαπλή παλινδρόμηση δεν υπάρχει ο σταθερός όρος β_0 όπως στην απλή. Αυτό δεν σημαίνει πως έχει απαλειφθεί, απλώς τις περισσότερες φορές προστίθεται μια στήλη από άσσους στον πίνακα X για την αναπαράσταση του.

Έτσι καταλήγουμε στο παρακάτω πρόβλημα βελτιστοποίησης

$$\min_{\beta} : \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

όπου το X_i αναφέρεται στην i-οστη γραμμή του πίνακα X και σκοπός είναι να βρεθούν οι τιμές του β που ελαχιστοποιούν το παραπάνω άθροισμα.

Με χρήση γραμμικής άλγεβρας το πρόβλημα εύρεσης βέλτιστης λύσης καταλήγει στην μορφή

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

3.6.4. Υπολογιστική πολυπλοκότητα της πολλαπλής γραμμικής παλινδρόμησης

Η πολλαπλή γραμμική παλινδρόμηση με χρήση της μεθόδου των ελαχίστων τετραγώνων σε N παρατηρήσεις και p συντελεστές έχει υπολογιστική πολυπλοκότητα:

- $O(p^2N)$ για τον πολλαπλασιασμό X^TX
- $O(pN)$ για τον πολλαπλασιασμό X^Ty
- $O(p^3)$ για τον υπολογισμό της LU παραγοντοποίησης για το γινόμενο $(X^T X)^{-1} X^T y$

Υποθέτουμε πως το N > p αλλιώς ο πίνακας X^TX δεν θα ήταν μοναδικός και δεν θα οριζόταν ο αντίστροφος του. Άρα η συνολική πολυπλοκότητα είναι η μεγαλύτερη από τις τρεις, δηλαδή $O(p^2N)$.

3.6.5. Μετρικές αξιολόγησης

Αφού υπολογιστεί το μοντέλο της παλινδρόμησης και βρεθούν οι παράμετροι των γραμμικών εξισώσεων στην συνέχει πρέπει να αξιολογηθεί πόσο καλά εξηγεί το μοντέλο τα πραγματικά δεδομένα και κατά πόσο κάθε μία ανεξάρτητη μεταβλητή συνεισφέρει στην εξήγηση αυτή. Για τον λόγο αυτό αρχικά θα ορίσουμε κάποιες ποσότητες σχετικά με την διασπορά των παρατηρήσεων γύρω από το μοντέλο παλινδρόμησης και την μέση τιμή της εξαρτημένης μεταβλητής.

Γενικά για την i-οστή παρατήρηση, Σχήμα 5, του συνόλου των δεδομένων ισχύει

$$y_i - \bar{y} = (\hat{y_i} - \bar{y}) + (y_i - \hat{y_i})$$

y_i i-οστή πραγματική τιμή εξαρτημένης μεταβλητής

$\bar{y}$ μέση τιμή εξαρτημένης μεταβλητής

$\hat{y_i}$ i-οστή προβλεπόμενη τιμή εξαρτημένης μεταβλητής από το μοντέλο παλινδρόμησης

Σχήμα 5. Σφάλματα i-οστης παρατήρησης

Εάν για κάθε μια από τις παραπάνω τρεις διαφορές πάρουμε το άθροισμα των τετραγώνων τους για κάθε i τότε προκύπτουν οι ποσότητες

Ολικό άθροισμα τετραγώνων (Total sum of squares):

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Ολικό άθροισμα παλινδρόμησης (Regression sum of squares):

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Ολικό άθροισμα τετραγώνων των σφαλμάτων (Error sum of squares):

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

Δηλαδή, *SST = SSR + SSE*

Το SST μετράει τη συνολική διασπορά των παρατηρήσεων y_i από την μέση τιμή $\bar{y}$, δηλαδή πόσο μακριά βρίσκονται. Το SSR εκφράζει το μέρος της διασποράς που οφείλεται στο μοντέλο παλινδρόμησης ενώ το SSE είναι η διασπορά λόγω άλλων παραγόντων που δεν μπορούν να εξηγηθούν από το μοντέλο.

Από τις παραπάνω ποσότητες προκύπτει ο συντελεστής προσδιορισμού (coefficient of determination) R^2 ο οποίος εκφράζει το κομμάτι της συνολικής διασποράς της εξαρτημένης μεταβλητής που μπορεί να εξηγηθεί από τις ανεξάρτητες μεταβλητές.

$$R^2 = \frac{SSR}{SST} \ , \quad 0 \leq \ R^2 \leq 1$$

Ο συντελεστής προσδιορισμού αποτελεί μια καλή μετρική για μοντέλα απλής γραμμικής παλινδρόμησης. Στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης δεν βοηθάει στην εκτίμηση της ποιότητας του μοντέλου διότι όσο περισσότερες είναι οι ανεξάρτητες μεταβλητές τόσο μεγαλύτερη είναι η τιμή του. Αυτό οφείλεται στο ότι μαζί με την ουσιαστική πληροφορία το μοντέλο ‘’εξηγεί’’ και θόρυβο με αποτέλεσμα να αυξάνεται η τιμή του συντελεστή. Για τον λόγο αυτό ορίστηκε ο διορθωμένος συντελεστής προσδιορισμού (adjusted coefficient of determination) $\bar{R}^2$.

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$

όπου *n* το σύνολο των παρατηρήσεων και *p* το πλήθος των ανεξάρτητων μεταβλητών.

Ο διορθωμένος συντελεστής προσδιορισμού είναι ένα καλύτερο μέτρο σύγκρισης μεταξύ μοντέλων πολλαπλής παλινδρόμησης με διαφορετικό πλήθος ανεξάρτητων μεταβλητών και έχει μικρότερη τιμή από τον συντελεστή προσδιορισμού.

Μία ακόμα μετρική για την διασπορά των σφαλμάτων του μοντέλου είναι το τυπικό σφάλμα της εκτίμησης (standard error of the estimate).

$$s = \sqrt{\frac{SSE}{k}}$$

k = n – p – 1 για πολλαπλή παλινδρόμηση

k = n – 2 για απλή παλινδρόμηση

Πέραν όμως των μετρικών αξιολόγησης του μοντέλου συνολικά, υπάρχουν τρόποι για την εκτίμηση κάθε μίας ανεξάρτητης μεταβλητής και του αντίστοιχου συντελεστή της. Έτσι συμπαιρένουμε πότε ένα χαρακτηριστικό (ανεξάρητητη μεταβλητή)

βοηθάει στην παραγωγή ενός μοντέλου που περιγράφει με περισσότερη ακρίβεια τα δεδομένα και πότε όχι.

Ένας από τους πιο εύκολους τρόπους αλλά παράλληλα αρκέτα χρήσιμος είναι να εξετάσουμε την τιμή της παραμέτρου. Εάν η τιμή είναι πολύ κοντά στο μηδέν τότε η συγκεκριμένη μεταβλητή δεν συμβάλει αρκετά στο συνολικό μοντέλο. Εάν όμως είναι αρκετά μεγαλύτερη ή μικρότερη από το μηδέν τότε υπάρχει άλλος τρόπος να διαπιστώσουμε την σημαντικότητα της.

Μέσω του στατιστικού ελέγχου υποθέσεων (Hypothesis Testing) μπορούμε να ελέγξουμε για κάθε παράμετρο μεταβλητής την σημαντικότητα της στο συνολικό μοντέλο. Αρχικά διατυπώνουμε την μηδενική υπόθεση H_0 στην οποία υποστηρίζουμε πως η τιμή της παραμέτρου είναι μηδέν $\beta_i = 0$, δηλαδή η αντίστοιχη ανεξάρτητη μεταβλητή δεν προσφέρει πληροφορία στο μοντέλο. Ελέγχουμε με ποια πιθανότητα είναι δυνατό να συμβεί αυτό. Για το σκοπό αυτό, ορίζουμε ένα ποσοστό για παράδειγμα 5% ή 1% το οποίο δείχνει πόσο σίγουροι είμαστε για το αποτέλεσμα. Αν η πιθανότητα ικανοποίησης της μηδενικής υπόθεσης είναι κάτω από αυτό το ποσοστό, τότε μπορούμε να πούμε ότι ισχύει η εναλλακτική υπόθεση H_a ($\beta_i \neq 0$). Άρα η ανεξάρτητη μεταβλητή είναι σημαντική για το μοντέλο και προσφέρει στην ‘’εξήγηση’’ των δεδομένων.

3.7. Δέντρα αποφάσεων

Τα δέντρα αποφάσεων (decision trees) είναι προγνωστικά μοντέλα τα οποία χρησιμοποιούνται για την αναπαράσταση ταξινομητών (classifiers) καθώς και μοντέλων παλινδρόμησης (regression models). Σε αντίθεση με τις περισσότερες μεθόδους πρόβλεψης οι οποίες βασίζονται στην στατιστική τα δέντρα αποφάσεων έχουν έναν πιο ανθρώπινο τρόπο στην λειτουργίας τους. Η τελική απόφαση προέρχεται μέσω επιλογών ανάμεσα σε δύο οι περισσότερα σύνολα το οποίο μοιάζει με τον τρόπο που λαμβάνουν αποφάσεις οι άνθρωποι.

Όταν ένα δέντρο απόφασης χρησιμοποιείται για διαδικασίες κατάταξης (classification), είναι καταλληλότερο να αναφέρεται ως ένα δέντρο ταξινόμησης. Όταν χρησιμοποιείται για διαδικασίες παλινδρόμησης, είναι πιο κατάλληλο να αναφέρεται ως δέντρο παλινδρόμησης. Ο διαχωρισμός αυτός γίνεται με βάση την εξαρτημένη

μεταβλητή. Όταν παίρνει διακριτές τιμές τότε έχουμε ένα δέντρο ταξινόμησης ενώ όταν είναι συνεχής ένα δέντρο παλινδρόμησης.

Τα δέντρα αποφάσεων διαιρούν το σύνολο δεδομένων σε μικρότερα κομμάτια και αυτά τα κομμάτια σε ακόμα μικρότερα φτιάχνοντας έτσι το αντίστοιχο δέντρο απόφασης. Το τελικό αποτέλεσμα είναι ένα δέντρο με κόμβους απόφασης και τους τελικούς κόμβους που ονομάζονται φύλλα. Οι ενδιάμεσοι κόμβοι ή κόμβοι απόφασης έχουν δύο ή περισσότερες διακλαδώσεις με την κάθε μία να αντιστοιχεί σε ένα διαφορετικό διάστημα τιμών ή στις διακριτές τιμές ενός χαρακτηριστικού, δηλαδή μιας ανεξάρτητης μεταβλητής. Τα φύλλα του δέντρου αποτελούν την απόφαση για την εξαρτημένη μεταβλητή, δηλαδή την τιμή της. Η εξαρτημένη μεταβλητή αντιστοιχεί στον υψηλότερο κόμβο του δέντρου ο οποίος ονομάζεται ρίζα. Στα δέντρα αποφάσεων μπορούν να υπάρξουν ανεξάρτητες μεταβλητές είτε με συνεχής είτε με διακριτές τιμές.

3.7.1. Αλγόριθμοι δέντρων αποφάσεων

Υπάρχουν δύο κύριες ομάδες στις οποίες χωρίζονται οι αλγόριθμοι των δέντρων αποφάσεων, οι από πάνω προς τα κάτω (top-down) και οι από κάτω προς τα πάνω (bottom-up). Το μεγαλύτερο μέρος των αλγορίθμων ανήκει στην πρώτη κατηγορία. Οι πιο γνωστοί από αυτούς είναι ο ID3 [25] (Quinlan, 1986), C4.5 [26] (Quinlan, 1993), CART [27] (Breiman, 1984). Αυτοί με την σειρά τους μπορούν να χωριστούν σε εκείνους που αναπτύσσουν το δέντρο και στην συνέχεια το ‘’κλαδεύουν’’ (C4.5, CART) και σε αυτούς που εκτελούν μόνο την φάση της ανάπτυξης.

Οι αλγόριθμοι αυτοί χρησιμοποιούν μία αναδρομική από πάνω προς τα κάτω άπληστη αναζήτηση για να καταλήξουν στο τελικό δέντρο. Στο εκάστοτε σύνολο δεδομένων ο αλγόριθμος ψάχνει το χαρακτηριστικό εκείνο (ανεξάρτητη μεταβλητή) το οποίο δίνει την καλύτερη διάσπαση. Η επιλογή του κατάλληλου χαρακτηριστικού γίνεται με βάση κάποιες μετρικές. Το δέντρο σταματά να αναπτύσσεται όταν ικανοποιηθούν οι συνθήκες διακοπής.

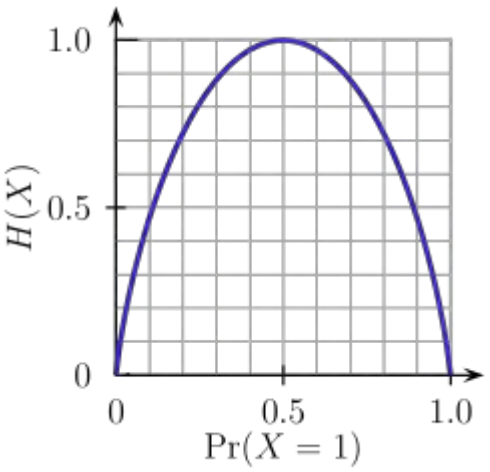
3.7.2. Συνθήκες διακοπής

- Για όλες τις παρατηρήσεις που βρίσκονται στον συγκεκριμένο κόμβο η εξαρτημένη μεταβλητή έχει την ίδια κλάση. Η συνθήκη αυτή χρησιμοποιείται μόνο σε δέντρα ταξινόμησης.
- Το δέντρο έχει φτάσει στο μέγιστο επιτρεπτό βάθος. Το βάθος ορίζεται από τον χρήστη.
- Ο αριθμός των παρατηρήσεων στον κόμβο είναι μικρότερος από τον ελάχιστο επιτρεπτό. Πάλι το όριο καθορίζεται από τον χρήστη.
- Οι μετρικές για την βέλτιστη διάσπαση δεν είναι μεγαλύτερες από κάποιο όριο.

3.7.3. Μετρικές βέλτιστης διάσπασης

Η εντροπία και το κέρδος πληροφορίας (information gain) χρησιμοποιούνται για την εύρεση της καταλληλότερης παραμέτρου διαχωρισμού του συνόλου των δεδομένων σε κάθε κόμβο του δέντρου.

Η εντροπία χρησιμοποιείται για τον υπολογισμό της ομοιογένειας σε ένα σύνολο. Αν το σύνολο είναι απόλυτα ομοιογενές ή αν είναι ισόποσα μοιρασμένο τότε η εντροπία είναι μηδέν, Σχήμα 6.



Σχήμα 6. Εντροπία [28]

Η εντροπία ενός συνόλου *S* υπολογίζεται από τον τύπο:

$$E(S) = \sum_{i=1}^{c} p_i log_2 p_i$$

όπου *S* η ανεξάρτητη μεταβλητή, *c* το πλήθος των κλάσεων της *S* και p_i η πιθανότητα της κλάσης *i* στο σύνολο *S*.

Για την εύρεση του καταλληλότερου χαρακτηριστικού διάσπασης πρέπει αρχικά να υπολογιστεί η εντροπία της εξαρτημένης μεταβλητής χρησιμοποιώντας τον παραπάνω τύπο. Στην συνέχεια υπολογίζεται κατά πόσο θα μειωθεί η εντροπία εάν επιλεγεί μία παράμετρος *X*.

$$E(T, X) = \sum_{c \in X} P(c)E(c)$$

όπου *T* η εξαρτημένη μεταβλητή, *X* η παράμετρος (ανεξάρτητη μεταβλητή) , *c* το πλήθος των κλάσεων της *X*.

Τέλος η παράμετρος που επιλέγεται είναι αυτή που δίνει το μεγαλύτερο κέρδος πληροφορίας (information gain).

$$IG(T, X) = E(T) - E(T, X)$$

Η μετρική αυτή δείχνει ότι επιλέγονται ως παράμετροι διάσπασης εκείνες που μοιράζουν πιο ομοιόμορφα το σύνολο δεδομένων.

Ο δείκτης Gini είναι ένα μέτρο για το πόσο συχνά ένα τυχαία επιλεγμένο στοιχείο ενός συνόλου είναι εσφαλμένα ταξινομημένο εφόσον έχει ταξινομηθεί τυχαία. Ο δείκτης Gini σε ένα σύνολο με *J* κλάσεις υπολογίζεται από τον τύπο:

$$I_G(p) = 1 - \sum_{i=1}^{J} p_i^2$$

όπου p_i η πιθανότητα της κλάσης *i*.

3.7.4. Αλγόριθμος CART

Ο αλγόριθμος CART (Classification And Regression Tree) είναι ένας από τους πιο γνωστούς τρόπους κατασκευής δέντρων αποφάσεων. Όπως δείχνει και το όνομα του

μπορεί να χρησιμοποιηθεί σε δέντρα ταξινόμησης αλλά και σε δέντρα παλινδρόμησης. Ο αλγόριθμος κατασκευάζει δυαδικά δέντρα δηλαδή κάθε εσωτερικός κόμβος περιέχει μόνο δύο διακλαδώσεις. Για τον χειρισμό μεταβλητών με διακριτές τιμές ο αλγόριθμος σε κάθε κόμβο μοιράζει τα δεδομένα έτσι ώστε να μην υπάρχουν παρατηρήσεις της ίδιας κλάσης και στα δύο κλαδιά του δέντρου. Για μεταβλητές με συνεχή δεδομένα ο CART χωρίζει το σύνολο των παρατηρήσεων σε δύο διαστήματα $(-\infty,\alpha]$ και $(\alpha, +\infty)$. Το ένα διάστημα βρίσκεται στο αριστερό κλαδί και το άλλο στο δεξί. Μία ακόμη διαφορά μεταξύ διακριτών και συνεχών μεταβλητών είναι ότι οι πρώτες συναντιούνται μόνο μία φορά σε κάποιο επίπεδο του δέντρου ως οι καταλληλότερες ενώ οι δεύτερες μπορούν να επιλεγούν περισσότερες από μία φορές.

Η εύρεση της καταλληλότερης παραμέτρου στον CART γίνεται με την χρήση του δείκτη Gini

$$I_G(X) = 1 - \sum_{c \in X} P(c)^2$$

όπου *X* η ανεξάρτητη μεταβλητή και *P(c)* η πιθανότητα της κλάσης *c* της μεταβλητής *X*.

Στην περίπτωση που ο CART κατασκευάζει δέντρο παλινδρόμησης τα φύλλα του δέντρου δεν θα αντιστοιχούν σε κάποια κλάση της εξαρτημένης μεταβλητής αλλά στην μέση τιμή των παρατηρήσεων που υπάρχουν στον κόμβο αυτό. Η επιλογή της μέσης τιμής σαν τιμή της εξαρτημένης μεταβλητής χρησιμοποιείται αρκετά συχνά, μπορούν όμως να χρησιμοποιηθούν και άλλες ποσότητες.

3.7.5. Βελτιστοποιήσεις

Το δέντρο που θα αναπτυχθεί σε επόμενο κεφάλαιο είναι ένα δέντρο παλινδρόμησης. Για τον λόγο αυτό οι βελτιστοποιήσεις που αναφέρονται εδώ επικεντρώνονται στο πως επηρεάζουν τέτοιου είδους δέντρα. Βέβαια οι ίδιες βελτιστοποιήσεις μπορούν να εφαρμοστούν και σε δέντρα ταξινόμησης.

Στα δέντρα αποφάσεων η βελτιστοποίηση αναφέρεται ως κλάδεμα (pruning) του δέντρου. Με τον όρο βελτιστοποίηση ή κλάδεμα εννοούμε την μείωση του μεγέθους (πολυπλοκότητας) του μοντέλου αλλά και το κόστος χρήσης του, όταν δηλαδή θα χρησιμοποιηθεί το μοντέλο για νέα δεδομένα. Ένα δέντρο με μεγάλο αριθμό τερματικών

κόμβων (φύλλα) έχει αυξημένη υπολογιστική πολυπλοκότητα τόσο στην κατασκευή του όσο και στην χρήση του. Κάποιες από τις συνθήκες διακοπής που αναφέρθηκαν προηγουμένως μπορούν να χρησιμοποιηθούν για το κλάδεμα του δέντρου όπως το μέγιστο επιτρεπόμενο βάθος ή ο ελάχιστος αριθμός παρατηρήσεων σε ένα κόμβο. Η τελευταία συνθήκη χρησιμοποιείται από τον αλγόριθμο CART.

Ένας ακόμη λόγος για να βελτιστοποιηθεί ένα δέντρο είναι η αποφυγή της υπερπροσαρμογής (overfitting). Στην περίπτωση των μοντέλων παλινδρόμησης αυτό συμβαίνει όταν το μοντέλο παρουσιάζει αρκετά μικρό σφάλμα στα δεδομένα εκπαίδευσης αλλά πολύ μεγαλύτερο σε άλλα δεδομένα που δεν χρησιμοποιήθηκαν για την κατασκευή του μοντέλου. Στα δέντρα αποφάσεων το πρόβλημα αυτό λύνεται μέσω της παραμέτρου πολυπλοκότητας (complexity parameter) *cp*. Εάν κάποια υποψήφια διακλάδωση δεν αυξάνει την συνολική τιμή του συντελεστή προσδιορισμού R^2 τουλάχιστον κατά *cp* τότε αυτή δεν πραγματοποιείται. Οπότε το συγκεκριμένο κλαδί σταματάει να αναπτύσσεται και έτσι μειώνεται η πολυπλοκότητα του δέντρου. Για την εύρεση της καταλληλότερης τιμής του *cp* χρησιμοποιείται η μέθοδος του cross-validation. Με τον τρόπο αυτό επιλέγεται η τιμή που παράγει το δέντρο με το χαμηλότερο σφάλμα στο οποίο όμως δεν υπάρχει το πρόβλημα της υπερπροσαρμογής.

3.8. Μηχανές Διανυσματικής Υποστήριξης

Μία ακόμη μέθοδος της μηχανικής μάθησης είναι οι μηχανές διανυσματικής υποστήριξης (support vector machines) ή όπως είναι γνωστά με το ακρωνύμιο SVM. Η μέθοδος των SVM παρουσιάστηκε από τους Vladimir N. Vapnik και Alexey Ya. Chervonenkis το 1963 ενώ η τελική τους έκδοση που περιελάβανε την βελτιστοποίηση των χαλαρών περιθωρίων και την χρήση των συναρτήσεων πυρήνα δημοσιεύθηκε το 1995 [29]. Αρχικά τα SVM χρησιμοποιούνταν σε προβλήματα ταξινόμησης όμως το 1995 ο Vapnik πρότεινε έναν τρόπο έτσι ώστε να εφαρμοστούν και σε προβλήματα παλινδρόμησης. Θα αναφερθούμε αρχικά στα SVM για ταξινόμηση στον δισδιάστατο χώρο ενώ στην συνέχεια θα γίνει περιγραφή της μεθόδου παλινδρόμησης και πως αυτές οι μέθοδοι μπορούν να επεκταθούν σε μη γραμμικά προβλήματα με την χρήση συναρτήσεων πυρήνα.

3.8.1. Ταξινόμηση

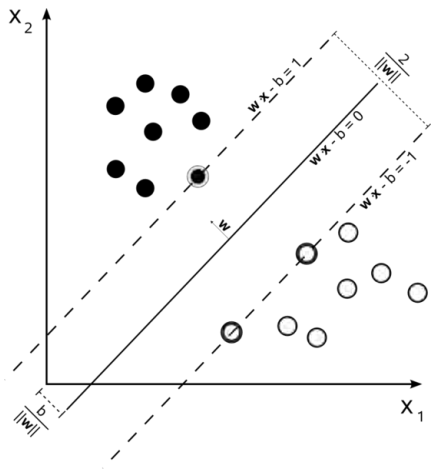
Σκοπός των SVM είναι να ορίσουν το υπερεπίπεδο, που μεγιστοποιεί το περιθώριο μεταξύ των δύο κλάσεων, από το οποίο θα ταξινομηθούν τα δεδομένα [30]. Τα διανύσματα που χρησιμοποιούνται για να οριστεί το υπερεπίπεδο ονομάζονται διανύσματα υποστήριξης (support vectors). Η εξίσωση της γραμμής που θα χωρίζει τις δύο κλάσεις είναι $wx + b = 0$. Θέλουμε να βρούμε την γραμμή με το μεγαλύτερο περιθώριο έτσι ώστε η απόσταση των δύο κλάσεων να είναι η μεγαλύτερη δυνατή. Χρησιμοποιώντας δύο διανύσματα υποστήριξης $wx + b = \pm 1$, θέλουμε η απόσταση μεταξύ τους να μεγιστοποιείται.

Θεωρούμε το σημείο x_0 στο υπερεπίπεδο $wx + b = 1$ οπότε έχουμε την εξίσωση $wx_0 + b = 1$. Για να βρούμε την απόσταση μεταξύ των δύο υπερεπιπέδων πρέπει να υπολογίσουμε την κάθετη απόσταση του x_0 από το $wx + b = -1$, την οποία ονομάζουμε r. Το $\frac{w}{\|w\|}$ είναι το μοναδιαίο διάνυσμα του υπερεπιπέδου $wx + b = 1$. Οπότε έχουμε

$$w\left(x_0 + r\frac{w}{\|w\|}\right) + b = 1$$

$$wx + b = -1$$

Από τις παραπάνω εξισώσεις βρίσκουμε ότι η απόσταση είναι $r = \frac{2}{\|w\|}$, Σχήμα 7.



Σχήμα 7. SVM για ταξινόμηση [31]

Επιστρέφοντας λοιπόν στον αρχικό σκοπό, την μεγιστοποίηση της απόστασης r, ισοδύναμα μπορούμε να ελαχιστοποιήσουμε το $\frac{\|w\|}{2}$. Επομένως έχουμε το πρόβλημα

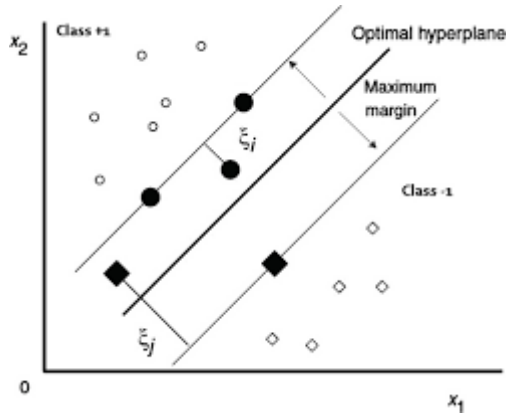
$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \tau.\omega \ \ y_i(wx_i + b) \geq 1, \forall x_i$$

Ο περιορισμός σημαίνει ότι

$$(wx_i + b) \geq 1, \alpha\nu\ y_i = 1, \qquad (wx_i + b) \leq -1, \alpha\nu\ y_i = -1$$

Άρα τα δεδομένα της κλάσης 1 πρέπει να είναι στην δεξιά πλευρά του υπερεπιπέδου $wx + b = 0$ ενώ της κλάσης 2 στην αριστερή πλευρά.

Μέχρι τώρα θεωρούσαμε πως τα δεδομένα ήταν πλήρως διαχωρίσιμα χωρίς κάποιο από αυτά να ταξινομείται σε λάθος κλάση. Επειδή όμως στον πραγματικό κόσμο αυτό είναι πολύ σπάνιο έως απίθανο να συμβεί πρέπει να βρεθεί ένας τρόπος έτσι ώστε να εισάγουμε στο αρχικό μοντέλο τα λάθη ταξινόμησης. Την λύση σε αυτό δίνουν τα χαλαρά περιθώρια. Το όνομα τους δείχνει πως πλέον επιτρέπεται κάποια από τα δεδομένα να έχουν ταξινομηθεί σε διαφορετική κλάση, υπάρχει όμως μία ποινή για αυτό. Πλέον σκοπός του μοντέλου είναι να μεγιστοποιήσει το περιθώριο μεταξύ των κλάσεων ενώ παράλληλα να ελαχιστοποιήσει το πλήθος των παρατηρήσεων που δεν ταξινομήθηκαν στην σωστή κλάση. Για να γίνει αυτό εισάγουμε την μεταβλητή ξ. Ο αλγόριθμος προσπαθεί να κρατήσει το ξ σε όσο πιο χαμηλή τιμή γίνεται ενώ παράλληλα μεγιστοποιεί το περιθώριο. Το ξ μετράει το άθροισμα των αποστάσεων, των παρατηρήσεων που έχουν ταξινομηθεί λάθος και εκείνων που βρίσκονται εντός του περιθωρίου, από το αντίστοιχο διάνυσμα υποστήριξης, Σχήμα 8.



Σχήμα 8. SVM για ταξινόμηση με χαλαρά περιθώρια [31]

Οπότε η εξίσωση παίρνει της εξής μορφή

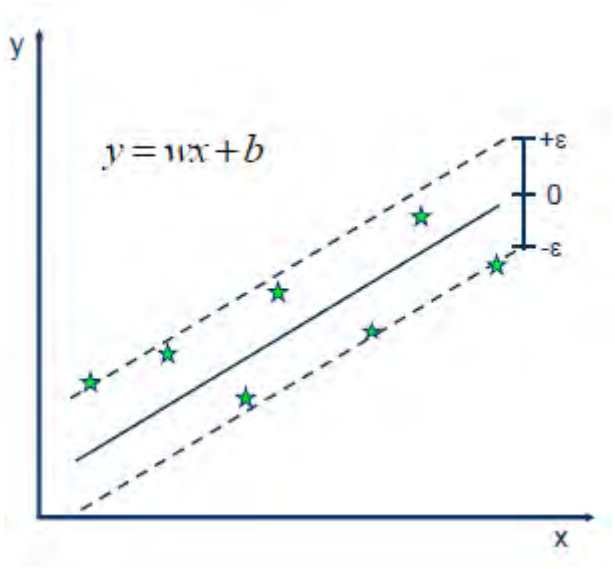
$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \quad \tau.\omega \; y_i(wx_i + b) \geq 1 - \xi_i, \forall x_i \; \kappa\alpha\iota \; \xi_i \geq 0$$

Η σταθερά *C* χρησιμοποιείται για την ισοστάθμιση (trade-off) μεταξύ του περιθωρίου και του σφάλματος των λανθασμένα ταξινομημένων παρατηρήσεων. Ανάλογα με την τιμή του δίνεται ‘’βάρος’’ στο περιθώριο ή στο σφάλμα. Αν το *C* είναι μηδέν τότε αγνοούμε τα χαλαρά περιθώρια.

3.8.2. Παλινδρόμηση

Έχουμε αναφέρει και σε προηγούμενο κεφάλαιο ότι σκοπός της παλινδρόμησης στον δισδιάστατο χώρο είναι να ορίσει μια ευθεία η οποία θα ελαχιστοποιεί τα σφάλματα, δηλαδή τις διαφορές μεταξύ των πραγματικών σημείων και των αντίστοιχων σημείων της ευθείας. Ένα ισχυρό κίνητρο για την χρήση των SVM στην παλινδρόμηση είναι ότι μπορούν πολύ ευκολά να επεκταθούν από την γραμμική στην μη γραμμική με την χρήση των συναρτήσεων πυρήνα.

Όπως στην ταξινόμηση έτσι και στην παλινδρόμηση η βασική ιδέα είναι ίδια, έχοντας ένα υπερεπίπεδο θέλουμε να μεγιστοποιήσουμε τις αποστάσεις των παρατηρήσεων από αυτό. Αρχικά να αναφέρουμε ότι στην περίπτωση της παλινδρόμησης το αποτέλεσμα είναι πραγματικός αριθμός οπότε είναι πολύ δύσκολο να γίνουν οι υπολογισμοί με το χέρι. Τα διανύσματα υποστήριξης έχουν τον ρόλο ενός περιθωρίου ανοχής (epsilon). Κάθε διάνυσμα απέχει από το κέντρο απόσταση *ε*. Τα δύο διανύσματα ορίζουν το μοντέλο παλινδρόμησης και το κεντρικό υπερεπίπεδο μοντελοποιεί τις παρατηρήσεις. Η βέλτιστη τοποθέτηση των τριών διανυσμάτων είναι εκείνη στην οποία όλες οι παρατηρήσεις βρίσκονται πιο κοντά στα εξωτερικά διανύσματα, δηλαδή οι αποστάσεις τους από το κεντρικό υπερεπίπεδο μεγιστοποιούνται, Σχήμα 9.

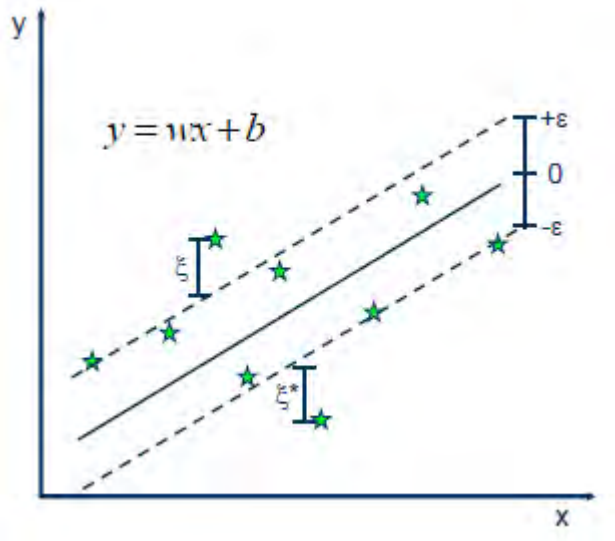


Σχήμα 9. SVM για παλινδρόμηση [31]

Η τελική μορφή του μοντέλου δεν διαφέρει πολύ από αυτή του μοντέλου ταξινόμησης

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \tau.\omega \quad \begin{cases} y_i - wx_i - b \leq \varepsilon \\ wx_i + b - y_i \leq \varepsilon \end{cases}$$

Όπως είναι φυσικό βέβαια, σε πραγματικά προβλήματα δεν γίνεται να βρίσκονται όλες οι παρατηρήσεις εντός ενός διαστήματος με πλάτος *2ε* για αυτό όπως και πριν έτσι και τώρα υπάρχουν τα χαλαρά περιθώρια. Στην περίπτωση της παλινδρόμησης ορίζουμε για κάθε παρατήρηση το ζευγάρι μεταβλητών ξ_i και ξ_i'. Οι μεταβλητές αυτές δείχνουν πόσο απέχει η παρατήρηση από τα διανύσματα υποστήριξης. Εάν το σημείο βρίσκεται εντός των δύο διανυσμάτων τότε οι μεταβλητές έχουν τιμή μηδέν. Εάν βρίσκεται από την πάνω πλευρά των διανυσμάτων τότε η ξ_i είναι ίση με την απόσταση από το πάνω διάνυσμα ενώ αν βρίσκεται από την κάτω πλευρά τότε η ξ_i' είναι ίση με τη απόσταση από το κάτω διάνυσμα, Σχήμα 10.



Σχήμα 10. SVM για παλινδρόμηση με χαλαρά περιθώρια [31]

Άρα η εξίσωση του μοντέλου είναι

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i}^{N}(\xi_i + \xi_i') \quad \tau.\omega \quad \begin{cases} y_i - wx_i - b \le \varepsilon + \xi_i, \forall x_i \\ wx_i + b - y_i \le \varepsilon + \xi_i', \forall x_i \end{cases} \kappa\alpha\iota\ \xi_i, \xi_i' \ge 0$$

3.8.3. Τρικ του πυρήνα

Προηγουμένως αναφέραμε ότι ένα πλεονέκτημα των SVM στην παλινδρόμηση είναι ότι μπορούν πολύ εύκολα να χρησιμοποιηθούν και για μη γραμμικά μοντέλα. Βέβαια η μέθοδος των συναρτήσεων πυρήνα μπορεί να εφαρμοστεί είτε σε μοντέλα παλινδρόμησης είτε σε μοντέλα ταξινόμησης. Με τις συναρτήσεις πυρήνα μπορούμε να μετατρέψουμε τα δεδομένα σε έναν χώρο μεγαλύτερης διάστασης από αυτόν που βρίσκονται. Οι συναρτήσεις πυρήνα έχουν την εξής μορφή [32]:

$$K(\bar{x}, \bar{y}) = \varphi(\bar{x}) \cdot \varphi(\bar{y})\ \mu\varepsilon\ \bar{x}, \bar{y}\ \in\ R^n$$

$$\varphi\colon R^n \to R^m\ \mu\varepsilon\ m \ge n$$

Οι πιο γνωστοί είναι:

- Γραμμικός (Linear)

$$K(\bar{x}, \bar{y}) = \bar{x} \cdot \bar{y} + \gamma$$

η παράμετρος *γ* καθορίζεται από τον χρήστη

- Πολυωνυμικός (Polynomial)

$$K(\bar{x}, \bar{y}) = (\bar{x} \cdot \bar{y} + 1)^p$$

η παράμετρος *p* καθορίζεται από τον χρήστη

- Ακτινωτός (Radial Basis Function (RBF))

$$K(\bar{x}, \bar{y}) = e^{-\frac{\|\bar{x}-\bar{y}\|}{2\sigma^2}}$$

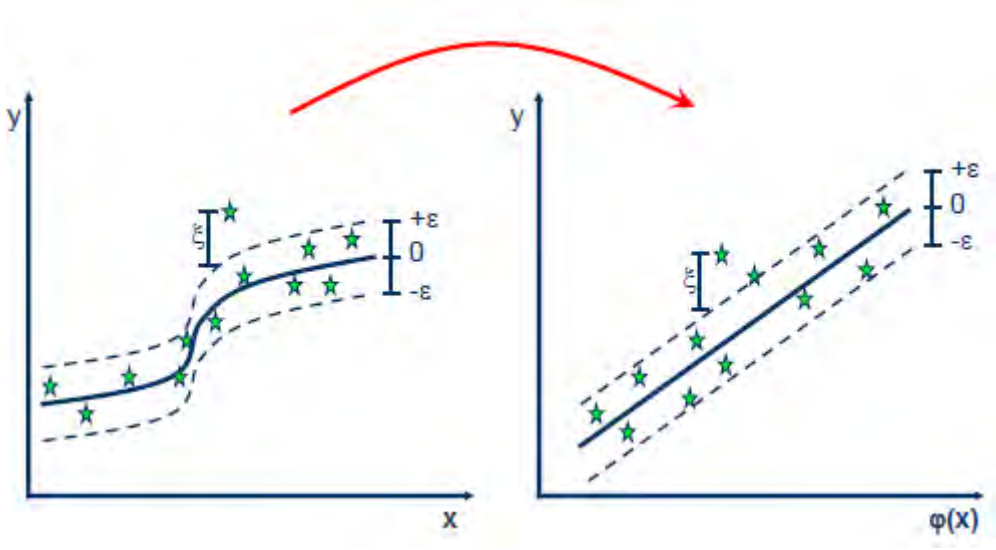
η παράμετρος *σ* καθορίζεται από τον χρήστη

Για να εφαρμόσουμε τις συναρτήσεις πυρήνα πρέπει να μετατρέψουμε το μοντέλο στην δυική του μορφή [33]:

$$y = \sum_{i=1}^{N} (a_i - a_i^*) \cdot \langle \varphi(x_i), \varphi(x) \rangle + b$$

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

Με την παραπάνω εξίσωση και την κατάλληλη συνάρτηση πυρήνα μπορούμε να φτιάξουμε ένα μη γραμμικό μοντέλο παλινδρόμησης, μεταφέροντας τα δεδομένα σε ένα χώρο μεγαλύτερης διάστασης στον οποίο είναι γραμμικά διαχωρίσιμα, Σχήμα 11.



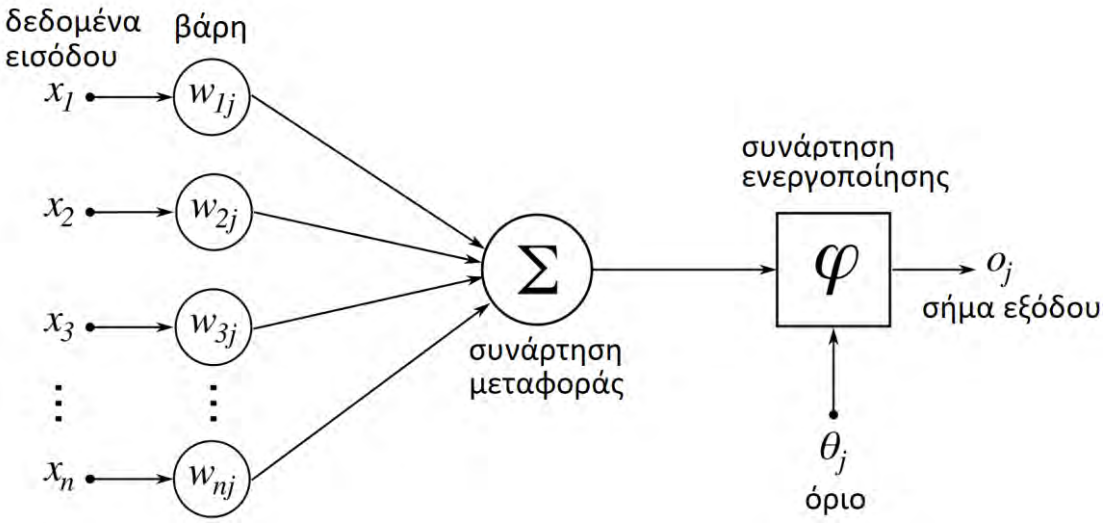
Σχήμα 11. Τρικ του πυρήνα [31]

3.9. Τεχνητά Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα (artificial neural networks - ANN) είναι συστήματα που βασίζονται στα βιολογικά νευρωνικά δίκτυα του εγκεφάλου [34]. Ο ανθρώπινος εγκέφαλος έχει περίπου 100 εκατομμύρια νευρώνες οι οποίοι επικοινωνούν μεταξύ τους μέσω συνδέσεων που ονομάζονται συνάψεις από τις οποίες περνούν ηλεκτροχημικά σήματα. Αν το άθροισμα αυτό των σημάτων υπερβαίνει κάποιο όριο τότε στέλνεται μία απόκριση μέσω του νευροάξονα (axon). Τα ANNs προσπαθούν να αναπαράγουν τον τρόπο με τον οποίο η πληροφορία μεταδίδεται μέσω των βιολογικών νευρώνων, επεξεργάζεται και στην συνέχεια παράγεται το τελικό αποτέλεσμα. Βέβαια η πολυπλοκότητα αυτών των δύο δικτύων δεν μπορεί να συγκριθεί καθώς ο ανθρώπινος εγκέφαλος έχει εκατομμύρια νευρώνες, και φυσικά πολύ μεγαλύτερη πολυπλοκότητα, σε σχέση με ένα τεχνητό νευρωνικό δίκτυο το οποίο έχει πολύ λιγότερους νευρώνες. Να αναφέρουμε επίσης ότι τα τεχνητά νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν για ταξινόμηση αλλά και για παλινδρόμηση.

Στο τεχνητό μοντέλο οι νευρώνες ονομάζονται κόμβοι (nodes). Η δομή κάθε κόμβου φαίνεται στο Σχήμα 12. Στην είσοδο υπάρχουν δεδομένα που δεν έχουν υποστεί ακόμα κάποια επεξεργασία (raw data) ή είναι τιμές εξόδου από άλλους νευρώνες. Η συνάρτηση μεταφοράς αθροίζει όλες τις εισόδους και αν η συνολική τιμή ξεπερνά κάποιο

όριο τότε η συνάρτηση ενεργοποίησης παράγει ένα σήμα εξόδου. Στην συνέχεια το σήμα εξόδου αποτελεί την έξοδο του νευρωνικού δικτύου ή είναι είσοδος σε επόμενο νευρώνα, ανάλογα με την αρχιτεκτονική του δικτύου. Για την δημιουργία ενός νευρωνικού δικτύου χρησιμοποιούνται πολύ τέτοιοι νευρώνες με την κατάλληλη τοπολογία.



Σχήμα 12. Δομή κόμβου νευρωνικού δικτύου

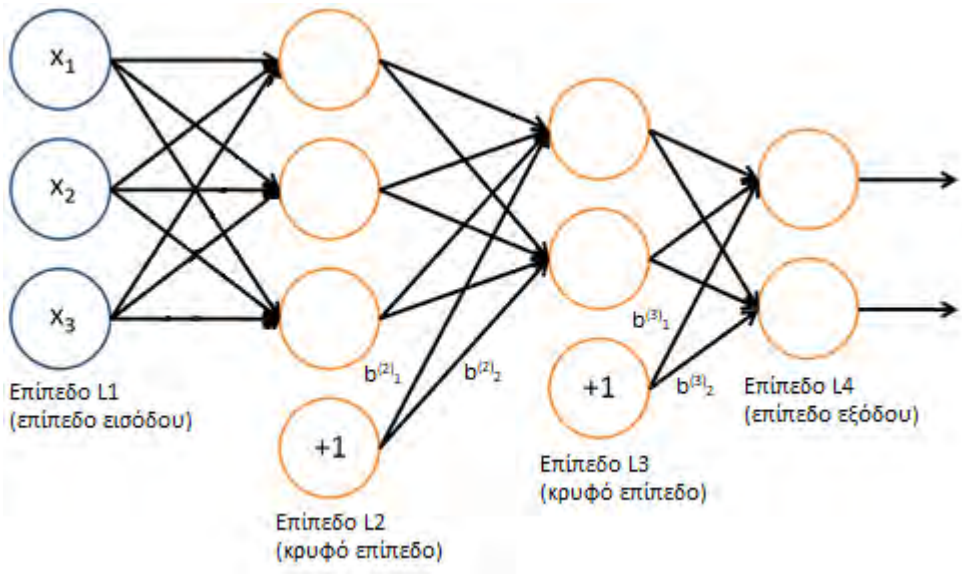
3.9.1. Νευρωνικά δίκτυα με απλή τροφοδότηση

Υπάρχουν πολλές και διαφορετικές κατηγορίες νευρωνικών δικτύων, κάποια αποδίδουν καλύτερα σε συγκεκριμένους τομείς, αλλά υπάρχουν και εκείνα που είναι γενικού σκοπού τα οποία όμως αποδίδουν εξίσου καλά. Μία τέτοια κατηγορία είναι τα νευρωνικά δίκτυα με απλή τροφοδότηση (feed forward neural networks) και το πιο γνωστό μοντέλο από αυτά είναι του πολυεπίπεδου αισθητήρα (multilayer perceptron). Το δίκτυο στο Σχήμα 13 είναι ένα τέτοιο δίκτυο.

Ο πολυεπίπεδος perceptron αποτελείται από ένα επίπεδο (layer) εισόδου, ένα επίπεδο εξόδου και ένα ή περισσότερα κρυφά επίπεδα. Το επίπεδο εισόδου διαβάζει τα δεδομένα που του παρέχει ο χρήστης, στα κρυφά επίπεδα συμβαίνει το μεγαλύτερο μέρος της ‘’εκμάθησης’’ και στο επίπεδο εξόδου παρουσιάζεται η τελική τιμή της εξαρτημένης μεταβλητής. Στο επίπεδο εισόδου κάθε νευρώνας αντιστοιχεί σε ένα διάνυσμα εισόδου δηλαδή μια ανεξάρτητη μεταβλητή ή αλλιώς ένα χαρακτηριστικό. Οι νευρώνες ενός επιπέδου επικοινωνούν μόνο με νευρώνες επιπέδων που βρίσκονται

δίπλα τους. Για παράδειγμα, το εσωτερικό επίπεδο *i* επικοινωνεί μόνο με το επίπεδο *i+1* στο οποίο στέλνει πληροφορία και με το επίπεδο *i-1* από το οποίο δέχεται πληροφορία. Κάθε νευρώνας επικοινωνεί είτε με όλους τους νευρώνες του επόμενου επιπέδου είτε με μερικούς από αυτούς. Κάθε τέτοια σύνδεση δεν μεταφέρει απλά την τιμή του προηγούμενου νευρώνα αλλά πολλαπλασιάζεται και με ένα βάρος $w_{ij}^{(l)}$. Το *i* αντιστοιχεί στην πηγή δηλαδή τον νευρώνα από όπου ξεκινάει η σύνδεση, το *j* τον νευρώνα που καταλήγει και το *l* στο επίπεδο.

Παρατηρήστε στο νευρωνικό δίκτυο του Σχήματος 13 τα βάρη $b_i^{(l)}$ $i = 1,2$ $l = 2,3$. Το συγκεκριμένο βάρος ονομάζεται πόλωση (bias) και υπάρχει σε κάθε κρυφό επίπεδο. Η πόλωση έχει τιμή εισόδου *x* ίση με την μονάδα και κάποιο βάρος *w*. Άρα το *x* παραλείπεται και έτσι η πόλωση λειτουργεί σαν ένα βάρος το οποίο συμμετέχει στην συνάρτηση μεταφοράς η οποία αθροίζει όλους τους όρους του νευρώνα. Ο ρόλος της πόλωσης αλλάζει ανάλογα με την συνάρτηση ενεργοποίησης. Για παράδειγμα αν η συνάρτηση ενεργοποίησης είναι η γραμμική, η πόλωση έχει τον ρόλο του σταθερού όρου στο γραμμικό μοντέλο παλινδρόμησης. Αν η συνάρτηση ενεργοποίησης είναι η σιγμοειδής τότε η πόλωση αλλάζει το όριο ταξινόμησης μεταξύ των κλάσεων.



Σχήμα 13. Νευρωνικό δίκτυο με απλή τροφοδότηση

3.9.2. Συναρτήσεις ενεργοποίησης

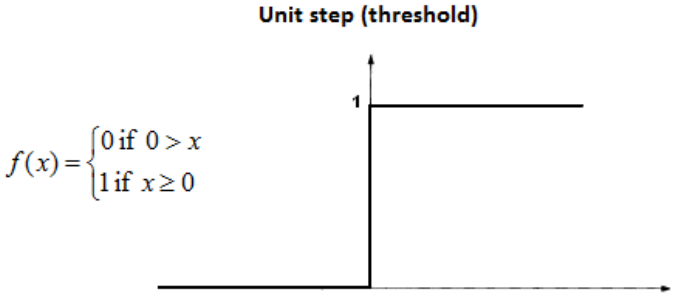
Ένα από τα πλεονεκτήματα των νευρωνικών δικτύων είναι να προσεγγίζουν οποιαδήποτε συνάρτηση εφόσον υπάρχει ο απαραίτητος όγκος δεδομένων. Η ιδιότητα αυτή εξαρτάται άμεσα από την επιλογή της κατάλληλης συνάρτησης ενεργοποίησης. Οι συναρτήσεις ενεργοποίησης βοηθούν το νευρωνικό δίκτυο να μάθει από μη γραμμικά χαρακτηριστικά που βρίσκονται στα δεδομένα. Η είσοδος στην συνάρτηση ενεργοποίησης είναι η έξοδος του νευρώνα που ονομάζεται συνάρτηση μεταφοράς, δηλαδή το βεβαρημένο άθροισμα των εισόδων του νευρώνα. Η παρακάτω συνάρτηση περιγράφει την συνάρτηση ενεργοποίησης.

$$o_j = \phi(b_j + \sum_{i=1}^{p} w_i x_i)$$

Όπου o_j είναι η έξοδος της συνάρτησης ενεργοποίησης στον νευρώνα *j* σε κάποιο επίπεδο και με $\phi(\cdot)$ συμβολίζεται η συνάρτηση ενεργοποίησης. Αν η τιμή της εξόδου ξεπερνά κάποιο όριο τότε ο συγκεκριμένος νευρώνας ενεργοποιείται και η τιμή του προωθείται στους νευρώνες του επόμενου επιπέδου.

Βηματική συνάρτηση ή συνάρτηση ορίου διέγερσης

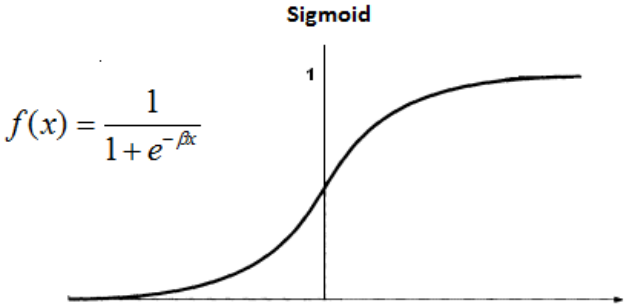
Αν η έξοδος της συνάρτησης μεταφοράς είναι μεγαλύτερη του ορίου (bias) τότε η έξοδος της συνάρτησης ενεργοποίησης είναι 1, διαφορετικά είναι 0, Σχήμα 14. Η συγκεκριμένη συνάρτηση δεν χρησιμοποιείται συχνά επειδή είναι απόλυτη στην τιμή εξόδου και δεν είναι παραγωγίσιμη σε όλο το πεδίο ορισμού της, το οποίο απαιτείται για την βελτιστοποίηση του δικτύου και θα εξηγηθεί στην συνέχεια.



Σχήμα 14. Βηματική συνάρτηση ενεργοποίησης

Σιγμοειδής συνάρτηση

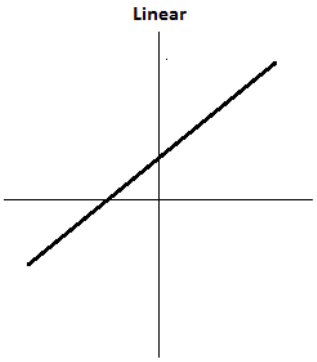
Πρόκειται για την περισσότερο χρησιμοποιούμενη συνάρτηση ενεργοποίησης στα πολυεπίπεδα νευρωνικά δίκτυα απλής τροφοδότησης. Μέσω αυτής τα νευρωνικά δίκτυα μπορούν να ‘’μάθουν’’ από μη γραμμικά χαρακτηριστικά. Επίσης είναι παραγωγίσιμη σε όλο το πεδίο ορισμού της το οποίο όπως προαναφέραμε είναι πολύ σημαντικό και τέλος οι τιμές εξόδου της βρίσκονται στο διάστημα τιμών [0, 1], Σχήμα 15.



Σχήμα 15. Σιγμοειδής συνάρτηση ενεργοποίησης

Γραμμική συνάρτηση

Όπως και στην γραμμική παλινδρόμηση, η γραμμική συνάρτηση ενεργοποίησης μετατρέπει το βεβαρημένο άθροισμα της συνάρτησης μεταφοράς μέσω μιας γραμμικής συνάρτησης, Σχήμα 16. Η συγκεκριμένη συνάρτηση χρησιμοποιείται ανάμεσα στο τελευταίο κρυφό επίπεδο και στο επίπεδο εξόδου του νευρωνικού δικτύου για προβλήματα παλινδρόμησης.



Σχήμα 16. Γραμμική συνάρτηση ενεργοποίησης

3.9.3. Αλγόριθμος ελάττωσης της παραγώγου

Ο αλγόριθμος ελάττωσης της παραγώγου (gradient descent) είναι ένας αλγόριθμος βελτιστοποίησης ο οποίος ελαχιστοποιεί μια συνάρτηση ως προς κάποιες παραμέτρους της [35]. Τις περισσότερες φορές οι συναρτήσεις αυτές είναι συναρτήσεις κόστους μέσω των οποίων αξιολογούνται μοντέλα πρόβλεψης. Για παράδειγμα, σε ένα μοντέλο απλής γραμμικής παλινδρόμησης μια συνάρτηση κόστους ή σφάλματος είναι

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

Για να ελαχιστοποιηθεί το σφάλμα πρέπει να βρεθούν οι παράμετροι της ευθείας παλινδρόμησης έτσι ώστε το μέσο τετραγωνικό σφάλμα (MSE) να έχει την ελάχιστη δυνατή τιμή. Αυτό μπορεί να γίνει με την μέθοδο των ελαχίστων τετραγώνων. Όταν όμως ο όγκος των δεδομένων είναι πολύ μεγάλος τότε το υπολογιστικό κόστος αυξάνεται και είναι ασύμφορο να χρησιμοποιηθεί αυτή η μέθοδος. Ο gradient descent είναι ένας προσεγγιστικός επαναληπτικός αλγόριθμος ο οποίος μπορεί να δώσει μια λύση πολύ κοντά σε αυτή των ελαχίστων τετραγώνων.

Έστω ότι έχουμε την συνάρτηση κόστους μέσου τετραγωνικού σφάλματος

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

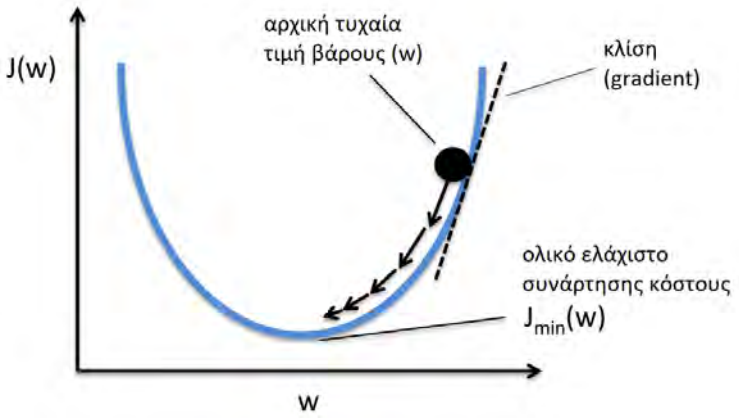
όπου $y^{(i)}$ είναι η πραγματική τιμή της παρατήρησης *i*, $h_\theta\left(x^{(i)}\right)$ είναι η τιμή εξόδου του μοντέλου που έχει επιλεχθεί για την παρατήρηση *i* και θ οι παράμετροι ή αλλιώς τα βάρη του μοντέλου. Σκοπός λοιπόν της gradient descent είναι να ελαχιστοποιήσει την συνάρτηση κόστους ως προς τα βάρη.

$$\min_\theta J(\theta)$$

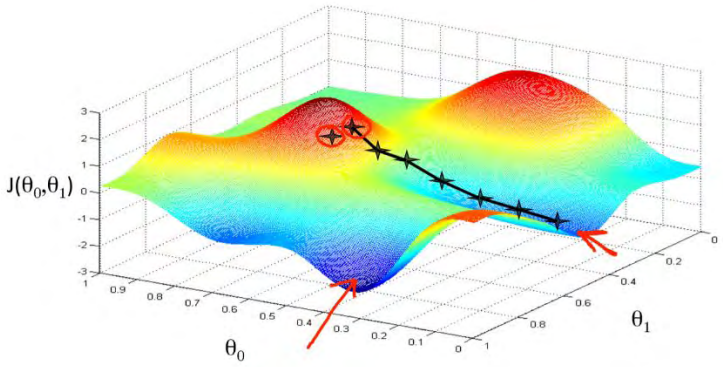
Αυτό το πετυχαίνει αν σε κάθε επανάληψη του αλγορίθμου οι παράμετροι ϑ ανανεώνονται σύμφωνα με την παρακάτω έκφραση έως ότου το σφάλμα γίνει μικρότερο από κάποιο όριο ή η διαφορά της μεταβολής του μεταξύ διαδοχικών επαναλήψεων γίνει μικρότερη από κάποια τιμή ή ο αλγόριθμος φτάσει τον μέγιστο αριθμό επαναλήψεων.

$$\theta_j := \theta_j - a\frac{\partial}{\partial\theta_j}J(\theta) \qquad j = 1, \dots, p$$

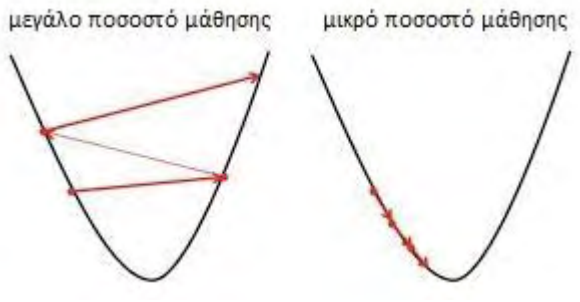
Όπου θ_j η j-οστή παράμετρος του μοντέλου, α το ποσοστό μάθησης (learning rate) το οποίο επιλέγετε από τον χρήστη και $\frac{\partial}{\partial\theta_j}J(\theta)$ η μερική παράγωγος ως προς θ_j της συνάρτησης κόστους $J(\theta)$. Η μερική παράγωγος χρησιμοποιείται γιατί σε μια συνάρτηση η παράγωγος αντιστοιχεί στην κλίση της. Οπότε υπολογίζοντας την μερική παράγωγο της συνάρτησης ως προς μια παράμετρο βρίσκουμε την κατεύθυνση που πρέπει να ακολουθήσουμε για να φτάσουμε σε κάποιο τοπικό ή στο ολικό ελάχιστο της συνάρτησης, Σχήμα 17, 18. Η τιμή του α αντιστοιχεί στο μέγεθος του βήματος για την εύρεση του ελαχίστου. Αν η τιμή του α είναι μικρή τότε θα χρειαστούν περισσότερες επαναλήψεις μέχρι να συγκλίνει ο αλγόριθμος ενώ αν είναι μεγάλη υπάρχει η πιθανότητα να υπερπηδήσει το σημείο όπου βρίσκεται το τοπικό ή ολικό ελάχιστο, Σχήμα 19.



Σχήμα 17. Εύρεση ελαχίστου σε δισδιάστατο χώρο



Σχήμα 18. Εύρεση ελαχίστου σε τρισδιάστατο χώρο



Σχήμα 19. Ποσοστό μάθησης

3.9.4. Αλγόριθμος ανάστροφης διάδοσης σφάλματος

Η απόδοση ενός νευρωνικού δικτύου που ανήκει στην κατηγορία της επιβλεπόμενης μάθησης (supervised learning) μπορεί να ελεγχθεί από το μέγεθος του σφάλματος. Στην περίπτωση της παλινδρόμησης, πόσο κοντά στην πραγματική τιμή είναι η έξοδος του δικτύου ενώ στην περίπτωση της ταξινόμησης πόσες παρατηρήσεις βρίσκονται σε λάθος κλάση. Με τον αλγόριθμο ανάστροφης διάδοσης σφάλματος (backpropagation) τα νευρωνικά δίκτυα μπορούν να βελτιώνουν το αρχικό σφάλμα της εκπαίδευσης με αποτέλεσμα να ‘’μαθαίνουν’’, δηλαδή να πλησιάζουν τις πραγματικές τιμές των δεδομένων [36].

Αρχικά στα βάρη κάθε νευρώνα ανατίθεται μια τυχαία τιμή. Έπειτα κάθε παρατήρηση του συνόλου των δεδομένων περνάει μέσα από το δίκτυο ξεκινώντας από το πρώτο επίπεδο και καταλήγοντας στο επίπεδο εξόδου, δίνοντας μία τιμή πρόβλεψης για την εξαρτημένη μεταβλητή. Αφού περάσουν όλες οι παρατηρήσεις τότε μπορεί να ελεγχθεί το σφάλμα, δηλαδή η διαφορά των τιμών εξόδου του νευρωνικού δικτύου από τις πραγματικές τιμές της εξαρτημένης μεταβλητής μέσω μιας συνάρτησης κόστους. Στο σημείο αυτό μέσω του αλγορίθμου ελάττωσης της παραγώγου (gradient descent) υπολογίζουμε τις μερικές παραγώγους των βαρών του τελικού κόμβου. Έτσι γνωρίζουμε για κάθε βάρος κατά πόσο επηρεάζει το σφάλμα. Στην συνέχεια η πληροφορία αυτή μεταφέρεται προς τα πίσω στο προηγούμενο επίπεδο όπου και εκεί υπολογίζονται οι μερικές παράγωγοι των βαρών, αυτή την φορά όμως χρησιμοποιώντας και την πληροφορία που πήραν από το επίπεδο που βρίσκεται μπροστά τους. Το ίδιο γίνεται και για τα επόμενα επίπεδα. Στο επίπεδο εισόδου δεν γίνονται αλλαγές διότι εκεί υπάρχουν

οι πραγματικές τιμές των ανεξάρτητων μεταβλητών. Τέλος τα βάρη ανανεώνονται σύμφωνα με την παρακάτω έκφραση

$$w = w - \eta \nabla_w J(w)$$

Η διαδικασία αυτή συνεχίζεται μέχρι μία από τις παρακάτω συνθήκες να είναι αληθής

- Το σφάλμα του δικτύου να είναι μικρότερο από κάποιο όριο που ορίζει ο χρήστης
- Η διαφορά των σφαλμάτων μεταξύ δύο επαναλήψεων να είναι μικρότερη από κάποια τιμή
- Ο αλγόριθμος να φτάσει τον μέγιστο αριθμό επαναλήψεων

3.9.5. Υπολογιστικό κόστος

Συγκριτικά με της παραδοσιακές μεθόδους μηχανικής μάθησης, όπως για παράδειγμα την γραμμική παλινδρόμηση, ένα νευρωνικό δίκτυο έχει μεγαλύτερο υπολογιστικό κόστος επειδή το πλήθος των παραμέτρων (βάρη) που πρέπει να προσδιορίσει είναι πολύ μεγαλύτερο. Έτσι εκτός από τα βάρη του δικτύου που επηρεάζουν την απόδοση του σχετικά με τα δεδομένα, υπάρχουν και άλλοι παράγοντες που καθορίζουν το υπολογιστικό κόστος του. Ένα νευρωνικό δίκτυο με μεγάλο χρόνο εξαγωγής αποτελεσμάτων θα ήταν πρακτικά άχρηστο σε εφαρμογές που απαιτούν γρήγορους υπολογισμούς. Δύο από τους πιο σημαντικούς παράγοντες είναι το βάθος του, δηλαδή ο αριθμός των επιπέδων του, και το πλάτος του, ο αριθμός των κόμβων σε κάθε επίπεδο. Το βάθος επηρεάζει κυρίως τον χρόνο εκπαίδευσης ενώ το πλάτος το μέγεθος του σφάλματος. Ένα νευρωνικό δίκτυο με μεγάλο πλάτος θα υπερπροσαρμόζει (overfitting) τα δεδομένα εκπαίδευσης έχοντας μικρό σφάλμα, αλλά θα έχει πολύ μεγαλύτερο σφάλμα σε νέα δεδομένα. Σε συνδυασμό με τεχνικές όπως τον διαχωρισμό του συνόλου των δεδομένων σε κομμάτι εκπαίδευσης και κομμάτι ελέγχου (train and test set) και με την μέθοδο του cross-validation μπορεί να βρεθεί το κατάλληλο βάθος και πλάτος του δικτύου.

3.10. Μείωση Διαστάσεων

Το πρόβλημα της ανάλυσης δεδομένων αποτελείται από κάποια στάδια. Ένα από αυτά, και αρκετά σημαντικό, είναι εκείνο της προεπεξεργασίας των δεδομένων. Για την δημιουργία ενός αξιόπιστου μοντέλου πρέπει τα δεδομένα που θα χρησιμοποιηθούν να είναι κατάλληλα και να περιέχουν όση περισσότερη πληροφορία γίνεται χωρίς να υπάρχουν χαρακτηριστικά που θα οδηγήσουν σε λάθος εκτιμήσεις. Επίσης ο όγκος των δεδομένων είναι ένα σοβαρό ζήτημα. Μπορεί πλέον αρκετά συστήματα να βασίζονται στο υπολογιστικό νέφος (cloud computing) [37] υπάρχουν όμως και εκείνα που είναι αναγκασμένα να κάνουν υπολογισμούς με περιορισμένη υπολογιστική μνήμη. Στο στάδιο της προεπεξεργασίας υπάρχουν μέθοδοι που ανήκουν στην κατηγορία της μη επιβλεπόμενης μάθησης (unsupervised learning) όπως η μείωση διαστάσεων (dimensionality reduction), οι οποίες βοηθούν στην ανακάλυψη απλούστερων και πιο συμπαγών αναπαραστάσεων των αρχικών δεδομένων, έτσι ώστε να παρέχουν χρήσιμο περιεχόμενο σε επόμενα στάδια της ανάλυσης ή να δώσουν στον χρήστη μια καλύτερη κατανόηση των δεδομένων. Μία τέτοια τεχνική μείωσης διαστάσεων είναι η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis) PCA.

Έστω ότι προσπαθούμε να εξηγήσουμε κάποιο άγνωστο φαινόμενο. Για να το καταφέρουμε αυτό συλλέγουμε δεδομένα από οτιδήποτε νομίζουμε πως σχετίζεται με αυτό. Αν γνωρίζαμε από πριν τι ακριβώς πρέπει να μετρήσουμε τότε θα μπορούσαμε να κατασκευάσουμε ένα απλό μοντέλο για να περιγράψουμε το φαινόμενο. Επειδή όμως αυτό δεν μπορεί να γίνει καταλήγουμε να συλλέγουμε αρκετά χαρακτηριστικά πολλά από τα οποία είναι άσχετα με το φαινόμενο και δεν προσφέρουν πληροφορία στην εξήγηση του. Όμως μέσω της PCA μπορούμε να αξιολογήσουμε τα χαρακτηριστικά αυτά και να δούμε πόση πληροφορία προσφέρει το καθένα. Έτσι θα τα μετατρέψουμε σε μια μικρότερη διάσταση από αυτήν που βρίσκονται κρατώντας μόνο την πληροφορία που είναι σχετική με το φαινόμενο.

3.10.1. Ανάλυση Κύριων Συνιστωσών

Ένας τρόπος για να μειωθεί η διάσταση των δεδομένων θα ήταν να επιλεγούν τα χρήσιμα χαρακτηριστικά και να απορριφθούν τα υπόλοιπα. Αυτή όμως δεν είναι μία

καλή τεχνική. Η ανάλυση κύριων συνιστωσών (PCA) είναι μία τεχνική που χρησιμοποιεί έναν ορθογώνιο μετασχηματισμό, δηλαδή προβάλει τα δεδομένα σε έναν νέο χώρο, για να μετατρέψει ένα σύνολο παρατηρήσεων με πιθανώς συσχετιζόμενες μεταβλητές (χαρακτηριστικά), σε έναν γραμμικό συνδυασμό νέων χαρακτηριστικών τα οποία είναι μία καλή αναπαράσταση των αρχικών δεδομένων. Τα χαρακτηριστικά ονομάζονται κύριες συνιστώσες (principal components) και το πλήθος τους είναι ίδιο με αυτό των αρχικών χαρακτηριστικών.

Η πρώτη κύρια συνιστώσα (first principal component) περιέχει όσο το δυνατόν περισσότερη πληροφορία από τα αρχικά δεδομένα. Σε όρους στατιστικής αυτό μεταφράζεται στην μεγαλύτερη δυνατή διακύμανση (variance). Η δεύτερη κύρια συνιστώσα έχει την δεύτερη μεγαλύτερη διακύμανση και είναι κάθετη στην πρώτη. Κάθε επόμενη συνιστώσα έχει την αμέσως επόμενη μεγαλύτερη διακύμανση και είναι κάθετη στις προηγούμενες. Η ιδιότητα αυτή των νέων χαρακτηριστικών δείχνει γιατί η ανάλυση κύριων συνιστωσών είναι κατάλληλη για μείωση διαστάσεων σε δεδομένα. Επίσης οι κύριες συνιστώσες έχουν την ιδιότητα να ανακατασκευάζουν τα αρχικά χαρακτηριστικά το οποίο είναι πολύ σημαντικό. Για παράδειγμα, η πληροφορία που θα εξαχθεί από μία ανάλυση δεδομένων στην οποία χρησιμοποιήθηκε η PCA πρέπει να ερμηνευθεί με βάση τις αρχικές μεταβλητές και όχι τις κύριες συνιστώσες. Οπότε είναι αναγκαίο να υπάρχει τρόπος να επιστρέψουμε στις αρχικές μεταβλητές χωρίς μεγάλες απώλειες. Άρα οι κύριες συνιστώσες πρέπει να έχουν το ελάχιστο δυνατό σφάλμα ανακατασκευής.

3.10.2. Υπολογισμός των κύριων συνιστωσών

Υπενθυμίζεται ότι η PCA προσπαθεί να βρει την κατεύθυνση στην οποία η προβολή των δεδομένων θα έχει την μεγαλύτερη διακύμανση (variance). Έχοντας έναν πίνακα X με n παρατηρήσεις και m μεταβλητές και την προβολή του Xw, η διακύμανση του είναι

$$\frac{1}{n-1}(Xw)^T Xw = w^T \left(\frac{1}{n-1} X^T X\right) w = w^T C w$$

όπου από κάθε στήλη του X έχει αφαιρεθεί η μέση τιμή της έτσι ώστε κάθε μεταβλητή να έχει την ίδια επιρροή στην εξήγηση της διακύμανσης. Διαφορετικά οι μεταβλητές με μεγαλύτερες τιμές θα επηρέαζαν περισσότερο την διακύμανση. C είναι ο πίνακας συνδιακύμανσης (covariance) του X. Ο πίνακας συνδιακύμανσης έχει μέγεθος $m \times m$ και κάθε στοιχείο του αντιστοιχεί στην συνδιακύμανση μεταξύ δύο μεταβλητών. Η συνδιακύμανση δύο μεταβλητών (χαρακτηριστικών) x_j και x_k είναι

$$\sigma_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

όπου $\bar{x}_j$ είναι η μέση τιμή του διανύσμτος $\bar{x}_j = \sum_{i=1}^{n} x_{ij}$

Μπορούμε επίσης να εκφράσουμε τον υπολογισμό του πίνακα συνδιακύμανσης σε μορφή πινάκων

$$C = \frac{1}{n-1}\left((X - \bar{x})^T(X - \bar{x})\right) = \frac{1}{n-1}X^T X$$

Η τελική ισότητα προκύπτει επειδή υποθέσαμε πως από τον πίνακα X έχουν αφαιρεθεί οι μέσες τιμές των διανυσμάτων.

Συνοψίζοντας πρέπει να υπολογιστεί το w έτσι ώστε η διακύμναση $w^T C w$ να είναι η μεγαλύτερη δυνατή. Άρα πρέπει να μεγιστοποιηθεί η συνάρτηση της διακύμανσης ως προς το w. Υπάρχει όμως ένας περιορισμός ως προς το μέτρο του $w, \|w\| = w^T w = 1$. Το διάνυσμα w πρέπει να έχει μέτρο ίσο με την μονάδα διαφορετικά η έκφραση της διακύμανσης θα μπορούσε να μεγιστοποιηθεί πολλαπλασιάζοντας το w με κάποιον αριθμό και έτσι η μέγιστη τιμή να είναι άπειρη. Οπότε η συνάρτηση της PCA είναι

$$\max_{w} \; w^T C w$$

$$subject\ to \;\; w^T w = 1$$

Άρα καταλήγουμε σε ένα πρόβλημα βελτιστοποίησης συνάρτησης με περιορισμό από άλλη συνάρτηση. Ένα τέτοιο πρόβλημα μπορεί να λυθεί με την βοήθεια του πολλαπλασιαστή Lagrange [38]. Έτσι καταλήγουμε στην ελαχιστοποίηση μιας συνάρτησης σφάλματος

$$L = w^T C w - \lambda(w^T w - 1)$$

Για την εύρεση του ελαχίστου μιας συνάρτησης πρέπει να βρεθεί το σημείο όπου μηδενίζεται η παράγωγος της. Για να λύσουμε ως προς w θέτουμε την μερική παράγωγο της L ως προς w ίση με μηδέν

$$\frac{\partial L}{\partial w} = 0 \Rightarrow Cw - \lambda w = 0 \Rightarrow Cw = \lambda w$$

Η τελευταία ισότητα είναι ένα πρόβλημα εύρεσης ιδιοτιμών και ιδιοδιανυσμάτων με το w να αντιστοιχεί στο ιδιοδιάνυσμα και το λ στην ιδιοτιμή η οποία είναι και η εξηγούμενη διακύμανση. Κάθε ιδιοδιάνυσμα δείχνει την κατεύθυνση της κύριας συνιστώσας ενώ η ιδιοτιμή του δείχνει το μέγεθος της διακύμανσης που αντιστοιχεί σε αυτήν συγκριτικά με τις υπόλοιπες.

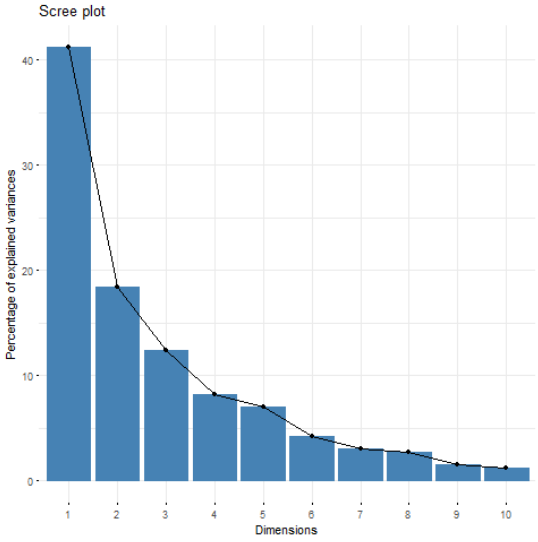
Έχοντας λοιπόν τα νέα χαρακτηριστικά για το σύνολο των δεδομένων μπορούμε να επιλέξουμε πόση από την πληροφορία θα κρατήσουμε προβάλλοντας τα δεδομένα σε έναν νέο χώρο μικρότερης διάστασης. Αυτό μπορεί να γίνει με έναν απλό πολλαπλασιασμό πινάκων

$$Y = XW$$

όπου W είναι ένα πίνακας με στήλες τα πρώτα d ιδιοδιανύσματα με τις μεγαλύτερες ιδιοτιμές και Y ο πίνακας με τα δεδομένα στον νέο πλέον *d*-διάστατο χώρο.

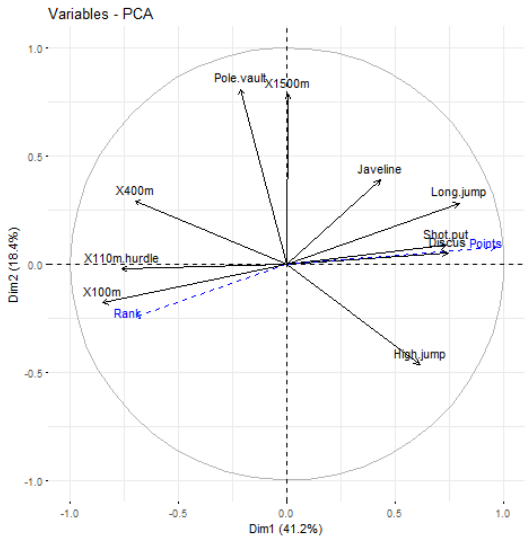
3.10.3. Γραφικές παραστάσεις

Μία γραφική παράσταση που βοηθάει στην επιλογή του νέου χώρου διαστάσεων είναι η scree plot, Σχήμα 20, στην οποία φαίνεται η διακύμανση κάθε κύριας συνιστώσας αλλά και η αθροιστική διακύμανση.



Σχήμα 20. Παράδειγμα scree plot

Μία ακόμη είναι η απεικόνιση της κατεύθυνσης και του μεγέθους των κύριων συνιστωσών στον δισδιάστατο χώρο όπου στον οριζόντιο άξονα βρίσκεται η πρώτη κύρια συνιστώσα και στον κατακόρυφο η δεύτερη, Σχήμα 21. Σε αυτή την γραφική παράσταση είναι ευκολότερο να ανακαλυφθούν τυχόν σχέσεις μεταξύ των κύριων συνιστωσών.



Σχήμα 21. Κατεύθυνση και μέγεθος κύριων συνιστωσών

ΚΕΦΑΛΑΙΟ 4

4. ΥΛΟΠΟΙΗΣΗ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στο προηγούμενο κεφάλαιο αναφερθήκαμε στο θεωρητικό κομμάτι και στα χαρακτηριστικά των μεθόδων που θα χρησιμοποιηθούν για την δημιουργία μοντέλων, μέσω των οποίων θα προβλεφθεί η ωριαία ζήτηση του θερμικού φορτίου του δικτύου της Δ.Ε.ΤΗ.Π. Σε αυτό το κεφάλαιο περιγράφουμε την διαδικασία επεξεργασίας των δεδομένων πριν την ανάλυση καθώς και τα αποτελέσματα των μοντέλων πρόβλεψης.

4.1. Λογισμικό

Για την ανάλυση χρησιμοποιήθηκε η γλώσσα προγραμματισμού R [39] και το περιβάλλον ανάπτυξης (IDE) RStudio [40]. Στην παρακάτω λίστα εμφανίζονται τα πιο σημαντικά πακέτα της R για αυτή την ανάλυση.

- stats (built-in)
- caret
- ggplot2
- FactoMineR
- factoextra
- rpart
- kernlab
- parallel
- doParallel
- nnet
- NeuralNetTools

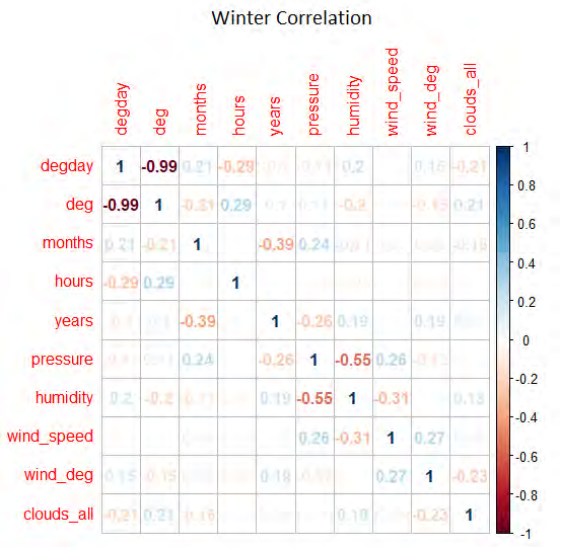
4.2. Προεπεξεργασία

Στο στάδιο της προεπεξεργασίας εξάγονται πληροφορίες σχετικά με κάθε μεταβλητή που θα συμμετέχει στην δημιουργία μοντέλων πρόβλεψης, όπως για παράδειγμα τα πεδία τιμών τους. Επίσης εξετάζεται η ύπαρξη συσχέτισης (correlation) μεταξύ δύο ή

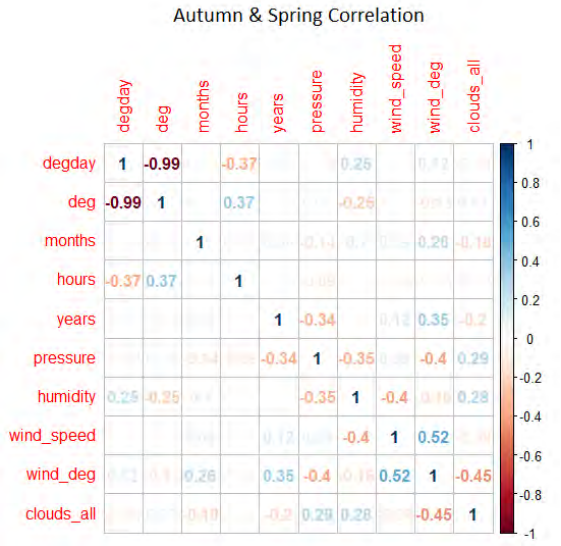
περισσότερων μεταβλητών. Πέρα από αυτά ένα μεγάλο κομμάτι της προεπεξεργασίας είναι ο ‘’καθαρισμός’’ των δεδομένων. Με τον όρο ‘’καθαρισμό’’ εννοείται η εύρεση εγγραφών με ελλιπή στοιχεία, λάθος μετρήσεις που οφείλονται σε σφάλματα αισθητήρων και η απαλοιφή τους. Σε ορισμένες περιπτώσεις στο στάδιο αυτό περιλαμβάνεται και η διαγραφή παρατηρήσεων (εγγραφών) οι μεταβλητές των οποίων έχουν ακραίες τιμές (outliers). Όμως σε αυτή την ανάλυση είναι σημαντικό να υπάρχουν ακραίες τιμές στο σύνολο των δεδομένων επειδή υπάρχει η ανάγκη τα μοντέλα πρόβλεψης που θα κατασκευαστούν να αποδίδουν καλά σε ακραίες τιμές.

Από το αρχικό σύνολο δεδομένων διαγράφηκαν όλες οι παρατηρήσεις που είχαν τουλάχιστον μία μεταβλητή με κενή τιμή. Σε συνεργασία με τους υπεύθυνους του δικτύου της Δ.Ε.ΤΗ.Π. ορίστηκαν κάποια όρια κάτω από τα οποία το δίκτυο δεν λειτουργεί. Αυτά είναι όταν η ζήτηση του δικτύου έχει τιμή μικρότερη από 20 MWh ($\mathrm{dhn} < 20$) ή η παροχή ενέργειας από τους ΑΗΣ είναι μικρότερη από 15 MWh ($chp < 15$). Η τελευταία συνθήκη δείχνει πως υπεύθυνοι για την μη κανονική λειτουργία του δικτύου είναι οι ΑΗΣ. Αν και η Δ.Ε.ΤΗ.Π. διαθέτει δικές της δεξαμενές νερού, όταν το δίκτυο λειτουργεί αποκλειστικά με αυτές δεν μπορεί να θεωρηθεί κανονική η λειτουργία του. Με βάση αυτά τα όρια διαγράφηκαν όσες παρατηρήσεις δεν τα ικανοποιούσαν. Επίσης από τα αρχεία καταγραφών των βλαβών βρέθηκε πως για τον μήνα Ιανουάριο του 2017 οι μετρήσεις ήταν εσφαλμένες οπότε διαγράφηκαν. Τέλος χρησιμοποιώντας την χρονοσφραγίδα των δεδομένων δημιουργήθηκαν δύο νέες μεταβλητές η *hours* και η *months*. Η πρώτη έχει σαν τιμές τις ώρες τις ημέρας και η δεύτερη τους μήνες. Έτσι προσφέρουμε παραπάνω πληροφορία στα δεδομένα η οποία θα βοηθήσει τα μοντέλα πρόβλεψης.

Επειδή υπάρχουν μεγάλες διακυμάνσεις στην θερμοκρασία οι οποίες επηρεάζουν με διαφορετικό τρόπο την ζήτηση του δικτύου, το σύνολο δεδομένων χωρίστηκε σε δύο κομμάτια με βάση την χρονική περίοδο στην οποία αναφέρονται. Με τον τρόπο αυτό πιστεύουμε πως οι προβλέψεις των μοντέλων θα είναι πιο εύστοχες. Το πρώτο κομμάτι είναι οι χειμερινοί μήνες Δεκέμβριος, Ιανουάριος και Φεβρουάριος ενώ το δεύτερο οι μήνες Οκτώβριος, Νοέμβριος, Μάρτιος και Απρίλιος [41].



Σχήμα 22. Συσχέτιση δεδομένων χειμώνα



Σχήμα 23. Συσχέτιση δεδομένων φθινοπώρου και άνοιξης

Για τον έλεγχο της συσχέτισης μεταξύ δύο ανεξάρτητων μεταβλητών χρησιμοποιήθηκε ο συντελεστής Pearson. Ο συντελεστής Pearson έχει τιμή +1 στην περίπτωση μίας τέλειας γραμμικής εξάρτησης και -1 στην περίπτωση τέλειας αντιστρόφως γραμμικής εξάρτησης. Στα Σχήματα 22 και 23 φαίνονται οι συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών για τα δύο σύνολα δεδομένων. Η συσχέτιση μεταξύ της εξωτερικής θερμοκρασίας (deg) και των βαθμοημερών (degday) είναι σχεδόν -1 το οποίο δηλώνει μία ισχυρή αντιστρόφως ανάλογη γραμμική εξάρτηση. Οι

βαθμοημέρες είναι ένας τρόπος αναπαράστασης της εξωτερικής θερμοκρασίας αντιστοιχίζοντας διαστήματα συνεχών τιμών σε διακριτές τιμές.

Το μείον (-) οφείλεται στο ότι όσο μεγαλύτερη είναι η τιμή της βαθμοημέρας τόσο μικρότερη είναι η τιμή της εξωτερικής θερμοκρασίας. Λόγω της μεγάλης εξάρτησης μεταξύ των δύο μεταβλητών και επειδή μας ενδιαφέρει να προβλέψουμε συνεχείς τιμές, η μεταβλητή των βαθμοημερών αφαιρέθηκε από τα σύνολα δεδομένων (Χειμώνας, Φθινόπωρο & Άνοιξη).

Τα βήματα που αναφέρονται στην συνέχεια ακολουθήθηκαν και για τα δύο σύνολα δεδομένων. Αρχικά τα δεδομένα χωρίστηκαν σε κομμάτι εκπαίδευσης και κομμάτι ελέγχου με την αναλογία 75/25. Έπειτα για την δημιουργία κάθε μοντέλου εφαρμόστηκε η τεχνική 10 fold cross-validation στο κομμάτι εκπαίδευσης (training set). Στα μοντέλα που κατασκευάστηκαν τροφοδοτήθηκαν τα δεδομένα ελέγχου (test set) και παρήχθησαν οι τιμές εξόδου, δηλαδή οι προβλέψεις για την ζήτηση ενέργειας του δικτύου τηλεθέρμανσης. Οι τιμές αυτές συγκρίθηκαν με τις πραγματικές και υπολογίστηκε η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) για τα δεδομένα εκπαίδευσης ως μέτρο σύγκρισης των αποτελεσμάτων. Επίσης για κάθε μοντέλο υπολογίστηκε ο συντελεστής προσδιορισμού (R^2) καθώς και το RMSE του συνόλου εκπαίδευσης. Οι παράμετροι που αφορούσαν την δομή των μοντέλων, για παράδειγμα αριθμός κόμβων σε κρυφό επίπεδο νευρωνικού δικτύου, υπολογίστηκαν κατά την διαδικασία του cross-validation. Η τιμή που επιλέχθηκε ήταν αυτή που έδινε το μικρότερο RMSE.

Τα μοντέλα που κατασκευάστηκαν είναι

- Πολλαπλή γραμμική παλινδρόμηση
- Δέντρα αποφάσεων
- Μηχανές διανυσματικής υποστήριξης για παλινδρόμηση
    - Γραμμικός πυρήνας
    - Πολυωνυμικός πυρήνας 2^{ου} βαθμού
    - Ακτινωτός πυρήνας
- Νευρωνικά δίκτυα

Στα παραπάνω μοντέλα εφαρμόστηκαν μόνο τα δεδομένα της Δ.Ε.ΤΗ.Π. Στο μοντέλο με την καλύτερη απόδοση χρησιμοποιήθηκαν τα δεδομένα και από τις δύο πηγές για να

εξεταστεί κατά πόσο θα βελτιωθεί το μοντέλο έχοντας περισσότερη πληροφορία. Τέλος με την ανάλυση κύριων συνιστωσών (PCA) μειώθηκε η διάσταση των δεδομένων των δύο πηγών. Με τα νέα δεδομένα δημιουργήθηκε ένα μοντέλο με την αποδοτικότερη μέθοδο και ελέγχθηκε η επιρροή της μείωσης των διαστάσεων στην ακρίβεια πρόβλεψης. Στην συνέχεια παρουσιάζονται τα αποτελέσματα κάθε μοντέλου.

4.3. Αποτελέσματα

Οι τιμές της ρίζας του μέσου τετραγωνικού σφάλματος (RMSE) και του συντελεστή προσδιορισμού (R^2) φαίνονται στους συγκεντρωτικούς πίνακες 2, 3, 4, 5, 6.

4.3.1. Πολλαπλή γραμμική παλινδρόμηση

Τα αποτελέσματα για το μοντέλο πολλαπλής γραμμικής παλινδρόμησης στο σύνολο δεδομένων για του χειμερινούς μήνες φαίνονται στο Σχήμα 24. Στην στήλη *Estimate* φαίνονται οι συντελεστές κάθε παραμέτρου. Καμία τιμή δεν είναι πολύ κοντά στο μηδέν οπότε αυτή είναι μία πρώτη ένδειξη ότι κάθε παράμετρος προσφέρει στην εξήγηση των δεδομένων. Παρατηρούμε ότι έχει δημιουργηθεί μία παράμετρος για κάθε ώρα και μήνα λόγω του τύπου των δύο μεταβλητών αλλά δεν υπάρχει παράμετρος για τον μήνα Ιανουάριο (*months1*) και την ώρα 00:00 (*hours0*). Η επιλογή της ώρα και του μήνα είναι τυχαία. Η απουσία τους δεν σημαίνει ότι δεν συμμετέχουν στο μοντέλο. Η παρουσία τους δηλώνεται με την απουσία όλων των άλλων μηνών και ωρών. Αυτό είναι ένα χαρακτηριστικό αυτού του τύπου των μεταβλητών. Η ύπαρξη τιμής για αυτές τις μεταβλητές απλά δηλώνει ότι αναφερόμαστε στον αντίστοιχο μήνα ή ώρα. Διαφορετικά έχουν τιμή μηδέν. Μία ακόμα σημαντική στήλη είναι η τελευταία ($\Pr(>|t|)$). Η στήλη αυτή αναφέρεται στον στατιστικό έλεγχο υποθέσεων (hypothesis testing) που γίνεται σε κάθε παράμετρο του μοντέλου για να διαπιστωθεί η σημαντικότητα του. Η υπόθεση ελέγχεται με ποσοστό 5% και εφόσον η τιμή είναι μικρότερη του 0.05 τότε δεχόμαστε την εναλλακτική υπόθεση η οποία είναι ότι η τιμή του συντελεστή της παραμέτρου είναι διάφορη του μηδενός οπότε η παράμετρος είναι σημαντική για το μοντέλο. Στην συγκεκριμένη περίπτωση κάθε παράμετρος είναι σημαντική για το μοντέλο.

Winter

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-47.366  -4.845   0.330   4.850  30.389

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.49623    0.52340  79.282  < 2e-16 ***
deg         -2.18787    0.02506 -87.306  < 2e-16 ***
months2     -1.74259    0.25394  -6.862 7.44e-12 ***
months12    -2.43350    0.25270  -9.630  < 2e-16 ***
hours1      -3.93436    0.70520  -5.579 2.52e-08 ***
hours2      -6.60282    0.71603  -9.221  < 2e-16 ***
hours3      -6.98189    0.69856  -9.995  < 2e-16 ***
hours4      -6.80458    0.70605  -9.638  < 2e-16 ***
hours5      -4.40176    0.69576  -6.327 2.68e-10 ***
hours6       7.28296    0.70231  10.370  < 2e-16 ***
hours7      36.08377    0.68706  52.519  < 2e-16 ***
hours8      47.29420    0.69545  68.005  < 2e-16 ***
hours9      45.90243    0.69978  65.595  < 2e-16 ***
hours10     36.09135    0.70437  51.239  < 2e-16 ***
hours11     29.94686    0.70887  42.246  < 2e-16 ***
hours12     30.38978    0.73031  41.612  < 2e-16 ***
hours13     37.17553    0.69910  53.176  < 2e-16 ***
hours14     40.66278    0.73432  55.374  < 2e-16 ***
hours15     37.62296    0.73502  51.186  < 2e-16 ***
hours16     36.43800    0.72865  50.008  < 2e-16 ***
hours17     40.87734    0.73570  55.563  < 2e-16 ***
hours18     45.51491    0.72312  62.942  < 2e-16 ***
hours19     48.48909    0.69455  69.813  < 2e-16 ***
hours20     45.18414    0.71502  63.193  < 2e-16 ***
hours21     36.55986    0.69411  52.672  < 2e-16 ***
hours22     21.77667    0.71742  30.354  < 2e-16 ***
hours23      8.80034    0.71135  12.371  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.046 on 6260 degrees of freedom
Multiple R-squared:  0.8554,    Adjusted R-squared:  0.8548
F-statistic:  1424 on 26 and 6260 DF,  p-value: < 2.2e-16
```

Σχήμα 24. Αποτελέσματα γραμμική παλινδρόμησης δεδομένων χειμώνα

Στο Σχήμα 25 φαίνονται τα αποτελέσματα του μοντέλου για τους φθινοπωρινούς και ανοιξιάτικους μήνες. Η παράμετρος της ώρας που λείπει είναι η ίδια, έχει αλλάξει όμως ο μήνας όπου σε αυτή την περίπτωση είναι ο Μάρτιος. Από την τελευταία στήλη παρατηρούμε ότι κάθε παράμετρος του μοντέλου είναι σημαντική στην εξήγηση της ενεργειακής ζήτησης του δικτύου της τηλεθέρμανσης.

Autumn & Spring

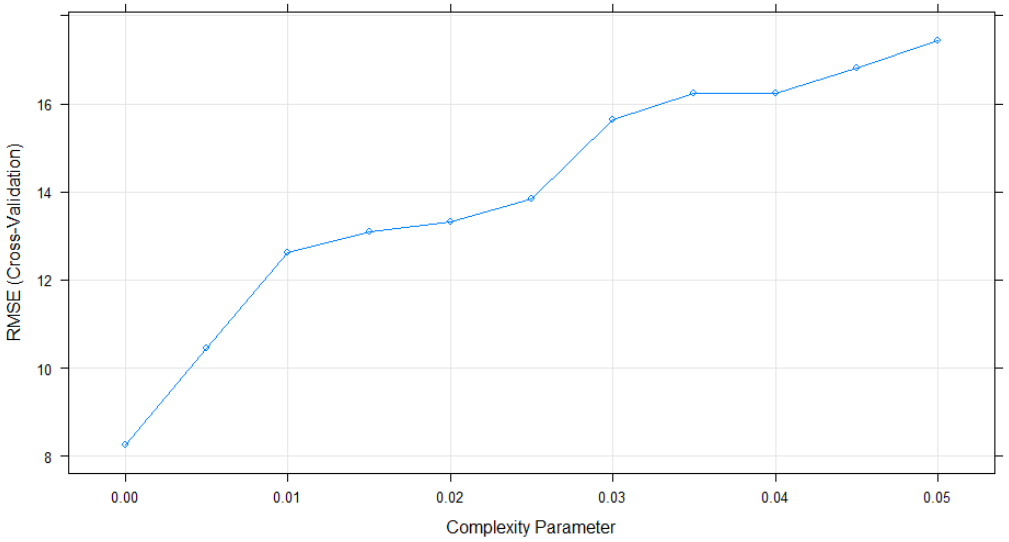
```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-38.932  -4.842  -0.093   4.928  35.920

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.07031    0.53583  84.114  < 2e-16 ***
deg         -2.48044    0.02706 -91.664  < 2e-16 ***
months4     -7.00502    0.29011 -24.146  < 2e-16 ***
months10    -9.42994    0.32758 -28.787  < 2e-16 ***
months11    -2.45350    0.23303 -10.529  < 2e-16 ***
hours1      -4.90672    0.78212  -6.274 3.75e-10 ***
hours2      -7.93920    0.83615  -9.495  < 2e-16 ***
hours3      -9.48739    0.82491 -11.501  < 2e-16 ***
hours4      -9.02415    0.83525 -10.804  < 2e-16 ***
hours5      -7.11953    0.78048  -9.122  < 2e-16 ***
hours6       2.99208    0.69922   4.279 1.90e-05 ***
hours7      24.10574    0.64633  37.296  < 2e-16 ***
hours8      31.87299    0.64832  49.162  < 2e-16 ***
hours9      29.72822    0.64896  45.809  < 2e-16 ***
hours10     26.10510    0.66488  39.263  < 2e-16 ***
hours11     24.02712    0.69372  34.635  < 2e-16 ***
hours12     24.40614    0.71054  34.349  < 2e-16 ***
hours13     29.16705    0.72332  40.324  < 2e-16 ***
hours14     32.47066    0.72085  45.045  < 2e-16 ***
hours15     30.93967    0.71627  43.196  < 2e-16 ***
hours16     30.01704    0.71487  41.990  < 2e-16 ***
hours17     32.22136    0.70112  45.957  < 2e-16 ***
hours18     35.61416    0.68848  51.729  < 2e-16 ***
hours19     38.25708    0.67763  56.457  < 2e-16 ***
hours20     36.76718    0.66141  55.589  < 2e-16 ***
hours21     28.48281    0.65178  43.700  < 2e-16 ***
hours22     17.73390    0.66140  26.813  < 2e-16 ***
hours23      8.12668    0.67223  12.089  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.914 on 6642 degrees of freedom
Multiple R-squared:  0.7591,    Adjusted R-squared:  0.7581
F-statistic: 775.1 on 27 and 6642 DF,  p-value: < 2.2e-16
```

Σχήμα 25. Αποτελέσματα γραμμική παλινδρόμησης δεδομένων φθινοπώρου άνοιξης

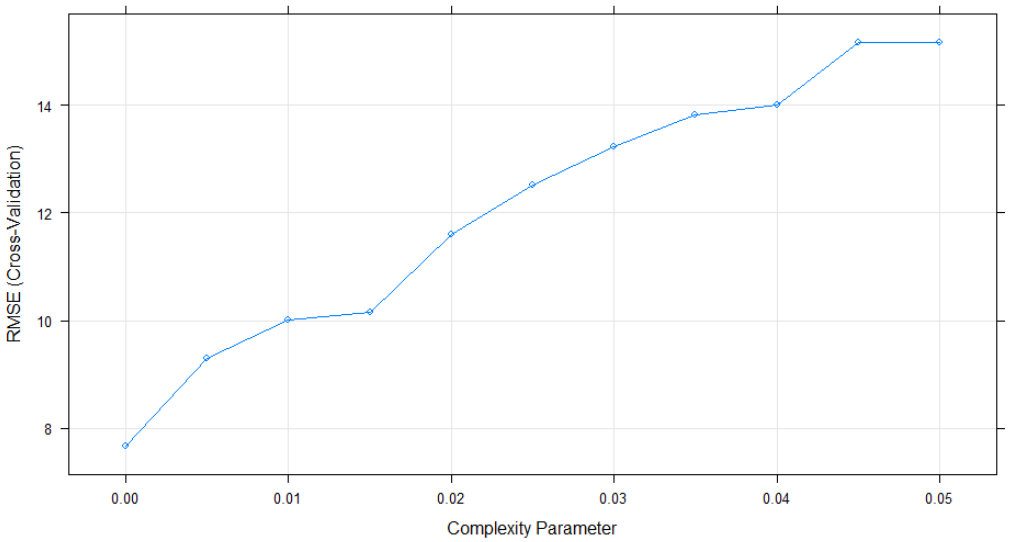
4.3.2. Δέντρα αποφάσεων

Στο συγκεκριμένο μοντέλο υπάρχει μία παράμετρος ρύθμισης (tuning parameter) η οποία αναφέρθηκε στο προηγούμενο κεφάλαιο, η παράμετρος πολυπλοκότητας (complexity parameter) cp. Υπενθυμίζουμε ότι η τιμή αυτής της παραμέτρου δηλώνει κατά πόσο πρέπει να βελτιώνεται το R^2 για να πραγματοποιείται η υποψήφια διακλάδωση σε κάποιο κόμβο του δέντρου. Αν η cp είναι μηδέν υπάρχει πιθανότητα να εμφανιστεί το φαινόμενο της υπερπροσαρμογής (overfitting). Στο Σχήμα 26 φαίνονται οι διαφορετικές τιμές για την cp και το RMSE του συνόλου εκπαίδευσης των χειμερινών μηνών ενώ στο Σχήμα 27 φαίνονται τα αντίστοιχα των υπόλοιπων μηνών. Στο τελικό

δέντρο η cp έχει τιμή 0 όμως το φαινόμενο της υπερπροσαρμογής δεν υπάρχει επειδή το δέντρο αποδίδει εξίσου καλά, με μικρό RMSE και στα σύνολα ελέγχου (test set) όπως φαίνεται στους πίνακες 2 και 4.



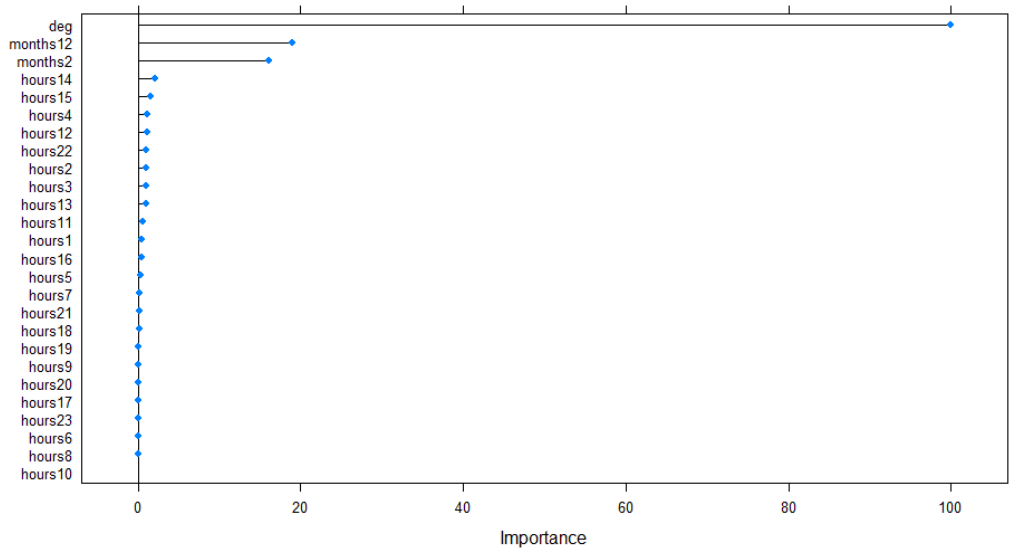
Σχήμα 26. Δέντρα αποφάσεων παράμετρος πολυπλοκότητας δεδομένων χειμώνα



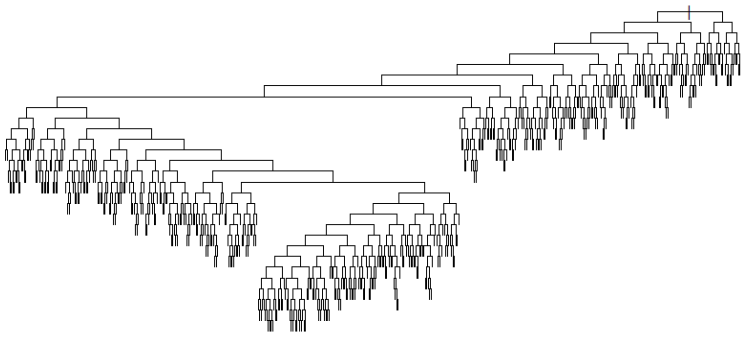
Σχήμα 27. Δέντρα αποφάσεων παράμετρος πολυπλοκότητας δεδομένων φθινοπώρου άνοιξης

Από το Σχήμα 28 βλέπουμε την σημαντικότητα κάθε μεταβλητής για το χειμερινό σύνολο δεδομένων. Με βάση αυτή την κατάταξη επιλέγονται οι μεταβλητές ως κριτήριο διάσπασης των κόμβων ξεκινώντας από την ρίζα του δέντρου μέχρι τα φύλλα. Στο Σχήμα 29 υπάρχει μία γραφική αναπαράσταση του ‘’χειμερινού’’ δέντρου. Αντίστοιχα στο Σχήμα

30 φαίνεται η σημαντικότητα των μεταβλητών του δεύτερου συνόλου δεδομένων ενώ η γραφική του αναπαράσταση βρίσκεται στο Σχήμα 31.



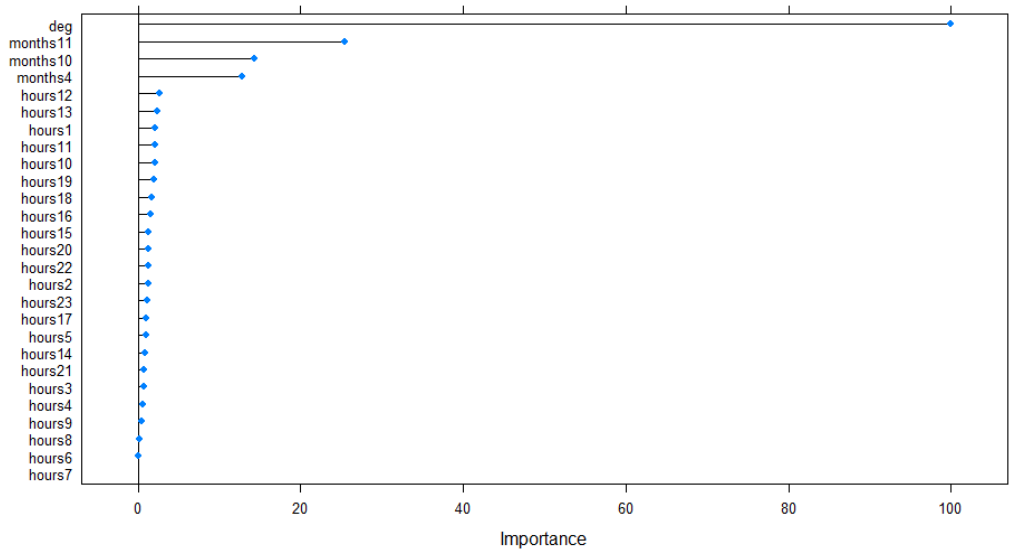
Σχήμα 28. Σημαντικότητα μεταβλητών δεδομένων χειμώνα



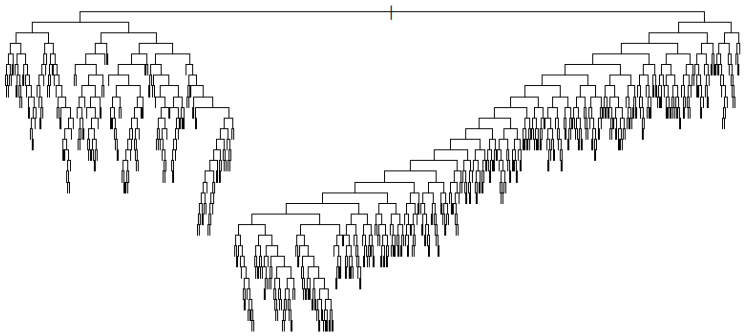
Σχήμα 29. Αναπαράσταση δέντρου απόφασης δεδομένων χειμώνα

Παρατηρούμε ότι και στα δύο σύνολα δεδομένων η πιο σημαντική μεταβλητή είναι της εξωτερικής θερμοκρασίας. Είναι λογικό να συμβαίνει διότι αυτή καθορίζει άμεσα την ζήτηση στο δίκτυο. Οι ώρες βρίσκονται στις τελευταίες θέσεις. Μπορεί η ζήτηση του δικτύου να διαφέρει μεταξύ πρωινών και βραδινών ωρών όμως δεν υπάρχει τόση μεγάλη διαφορά μεταξύ διπλανών ωρών. Για αυτό και η σημαντικότητα των ωρών

είναι χαμηλή ενώ αυτή των μηνών είναι πιο μεγάλη, επειδή η διάρκεια ενός μήνα είναι σαφώς μεγαλύτερη από την διάρκεια μίας ώρας.



Σχήμα 30. Σημαντικότητα μεταβλητών δεδομένων φθινοπώρου άνοιξης

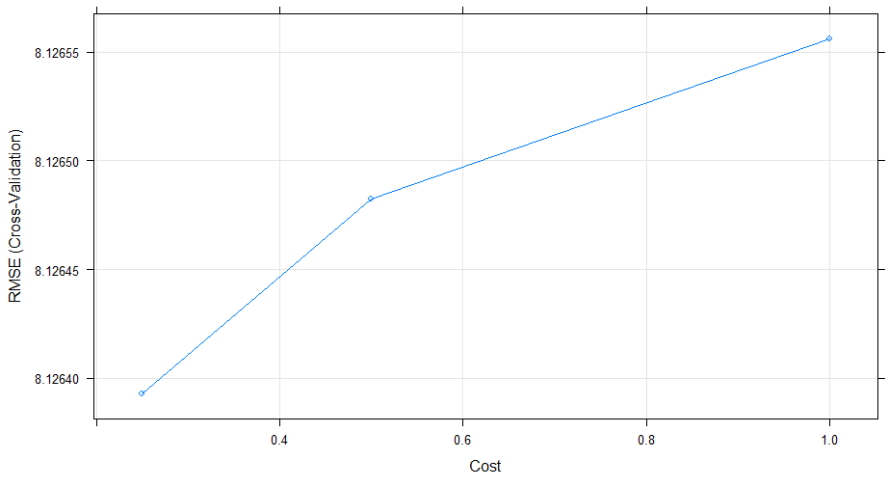


Σχήμα 31. Αναπαράσταση δέντρου απόφασης δεδομένων φθινοπώρου άνοιξης

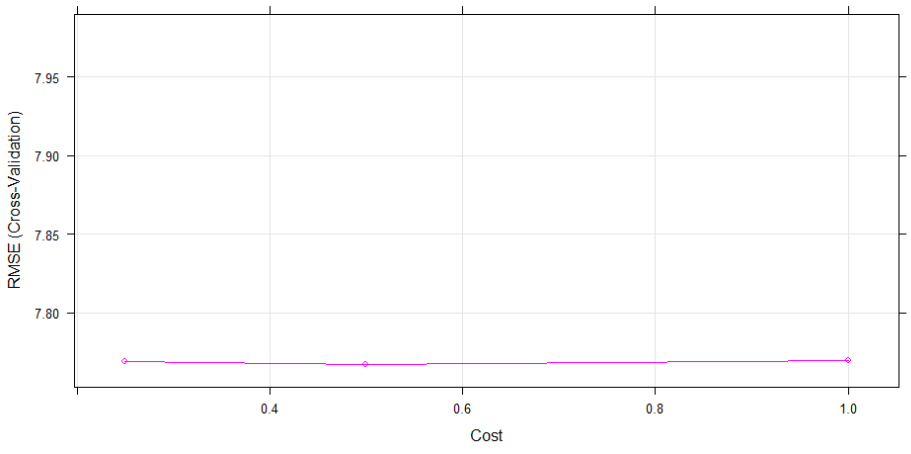
4.3.3. Μηχανές διανυσματικής υποστήριξης για παλινδρόμηση

Στα παρακάτω μοντέλα η παράμετρος ρύθμισης είναι το κόστος (*C*). Για τα μοντέλα παλινδρόμησης η τιμή του *C* καθορίζει πόσο μεγάλη θα είναι η ποινή για την συνάρτηση κόστους εξαιτίας των παρατηρήσεων που δεν βρίσκονται εντός του σωλήνα με πλάτος *2ε*. Το *C* επιλέχθηκε μέσω της διαδικασίας του cross-validation με κριτήρια το RMSE και την αποφυγή της υπερπροσαρμογής (overfitting). Στα Σχήματα 32, 33 και 34

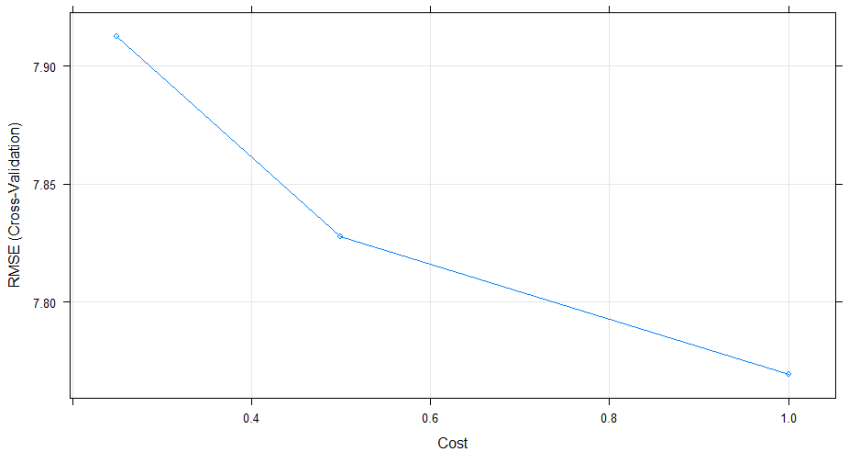
φαίνονται τα αποτελέσματα για τα τρία μοντέλα SVR του χειμερινού συνόλου δεδομένων. Παρατηρούμε πως για τα μοντέλα με τον γραμμικό και ακτινωτό πυρήνα υπάρχει αισθητή μεταβολή στο RMSE για διαφορετικές τιμές του *C*. Αντίθετα στο μοντέλο με τον πολυωνυμικό πυρήνα βαθμού 2 το RMSE είναι σχεδόν ίδιο για κάθε τιμή του *C*. Στα Σχήματα 35, 36, 37 βλέπουμε τα αποτελέσματα για το σύνολο δεδομένων με τους φθινοπωρινούς και ανοιξιάτικους μήνες. Για την παράμετρο ρύθμισης *C* στα SVR μοντέλα αυτού του συνόλου ισχύουν οι ίδιες παρατηρήσεις με αυτές των χειμερινών μοντέλων.



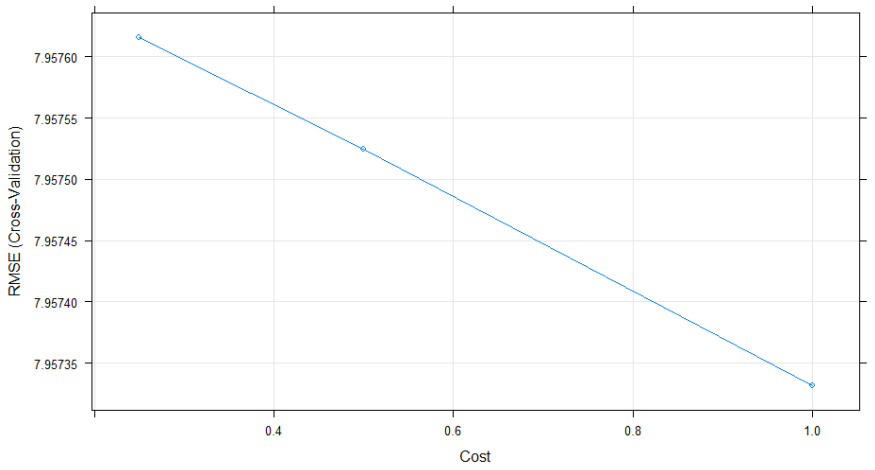
Σχήμα 32. Χειμώνας, παράμετρος κόστους για SVM με γραμμικό πυρήνα



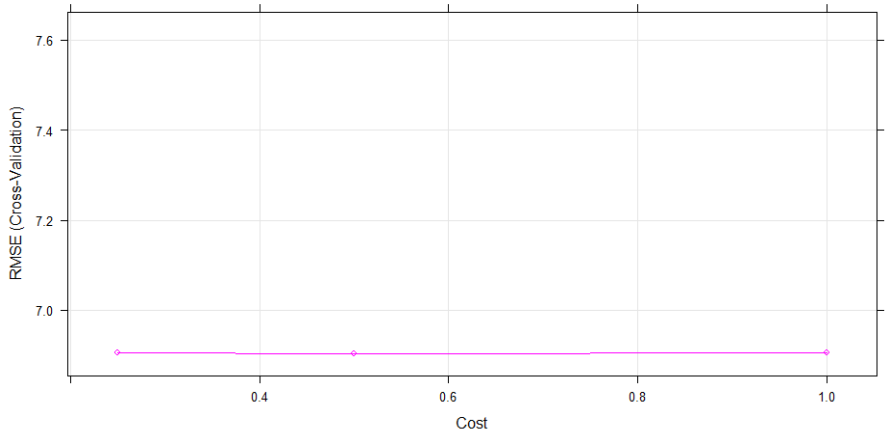
Σχήμα 33. Χειμώνας, παράμετρος κόστους για SVM με πολυωνυμικό πυρήνα



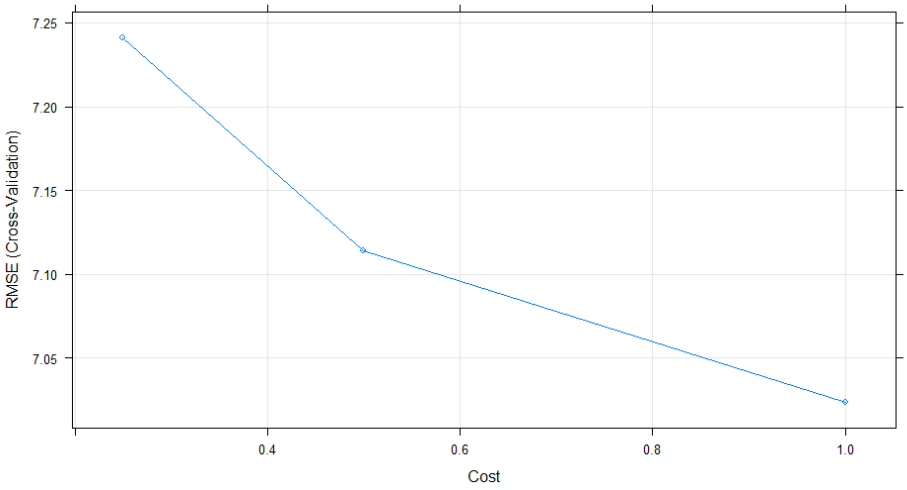
Σχήμα 34. Χειμώνας, παράμετρος κόστους για SVM με RBF πυρήνα



Σχήμα 35. Φθινόπωρο & Άνοιξη, παράμετρος κόστους για SVM με γραμμικό πυρήνα



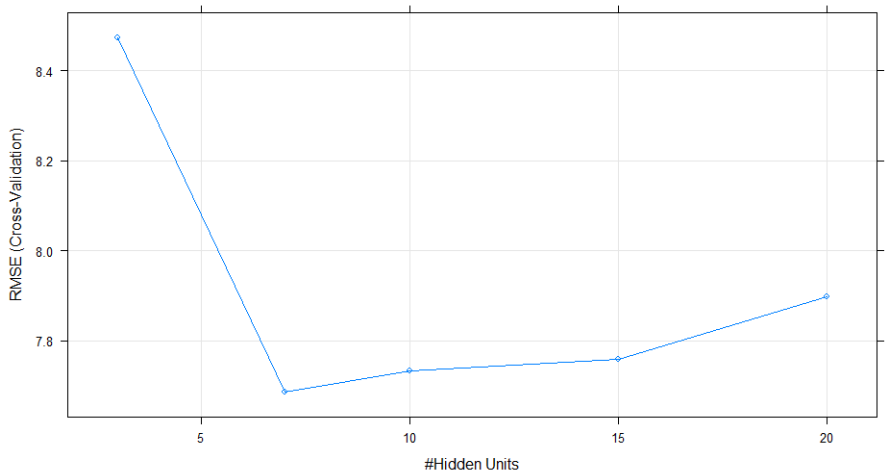
Σχήμα 36. Φθινόπωρο & Άνοιξη, παράμετρος κόστους για SVM με πολυωνιμικό πυρήνα



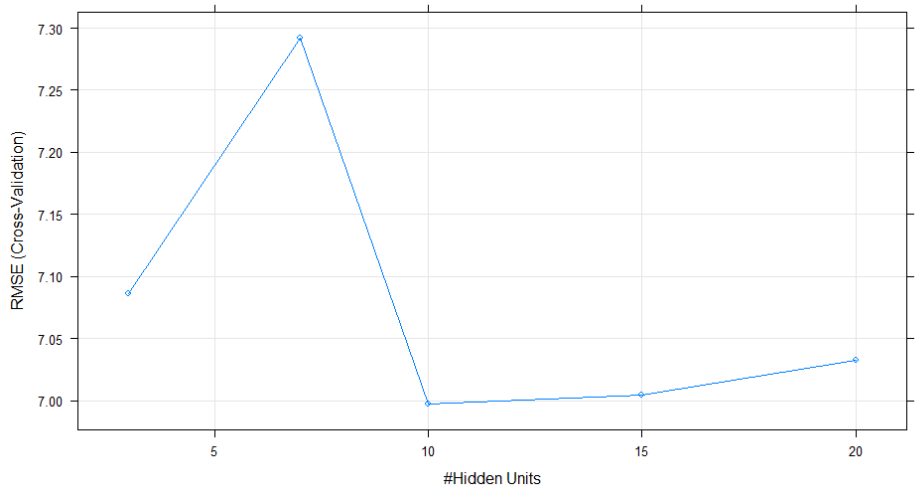
Σχήμα 37. Φθινόπωρο & Άνοιξη, παράμετρος κόστους για SVM με RBF πυρήνα

4.3.4. Νευρωνικά δίκτυα

Τα δύο νευρωνικά δίκτυα που κατασκευάστηκαν, ένα για κάθε σύνολο δεδομένων, είχαν την εξής ιδιαιτερότητα. Η παράμετρος ρύθμισης ήταν ο αριθμός των κόμβων στο κρυφό επίπεδο και ήταν διαφορετική για κάθε δίκτυο. Στο δίκτυο για το χειμερινό σύνολο δεδομένων ο αριθμός των κόμβων που επιλέχθηκε ήταν 7, Σχήμα 38. Ενώ στο δίκτυο για δεύτερο σύνολο δεδομένων ήταν 10, Σχήμα 39.

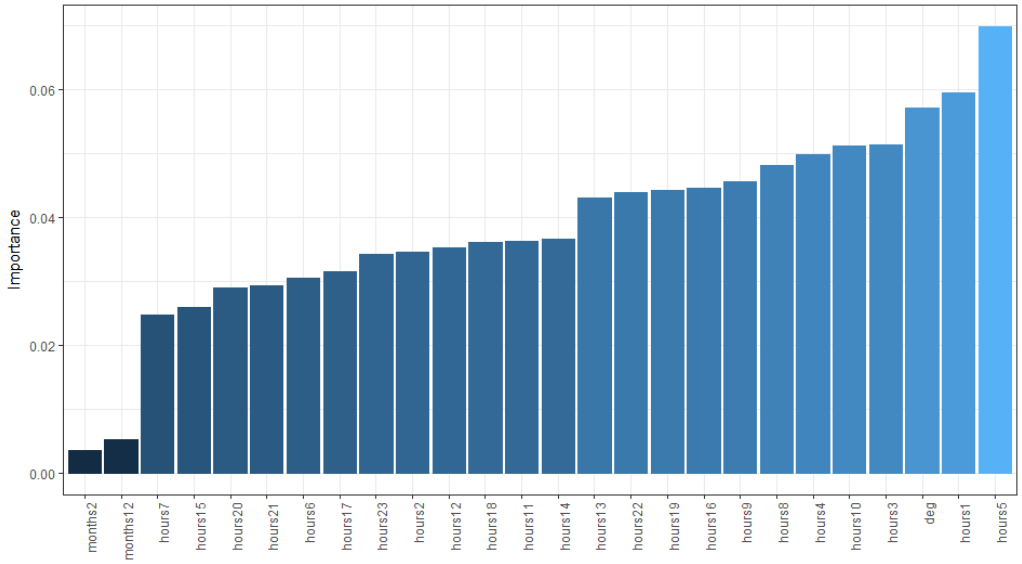


Σχήμα 38. Χειμώνας, αριθμός κόμβων κρυφού επιπέδου

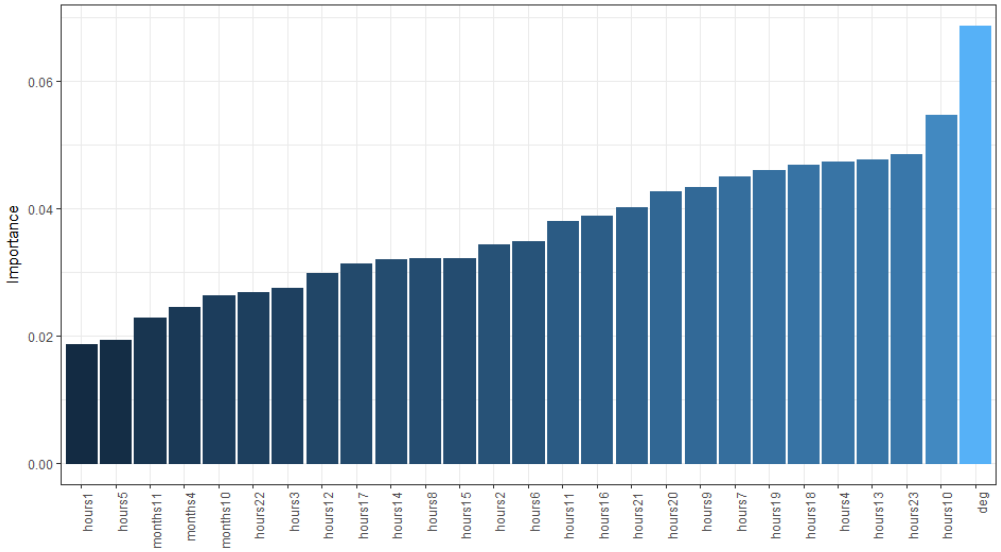


Σχήμα 39. Φθινόπωρο & Άνοιξη, αριθμός κόμβων κρυφού επιπέδου

Στα σχήματα 40 και 41 βλέπουμε για το χειμερινό και το φθινοπωρινό και ανοιξιάτικο σύνολο δεδομένων αντίστοιχα πόσο σημαντική ήταν κάθε μεταβλητή για το νευρωνικό δίκτυο, δηλαδή κατά πόσο βοήθησε στην πρόβλεψη της ζήτησης ενέργειας του δικτύου τηλεθέρμανσης.



Σχήμα 40. Χειμώνας, σημαντικότητα μεταβλητών για το νευρωνικό δίκτυο



Σχήμα 41. Φθινόπωρο & Άνοιξη, σημαντικότητα μεταβλητών για το νευρωνικό δίκτυο

Παρατηρούμε ότι η κατάταξη των μεταβλητών διαφέρει μεταξύ των δύο συνόλων. Ενώ στο σύνολο δεδομένων της φθινοπωρινής και ανοιξιάτικης περιόδου η μεταβλητή της εξωτερικής θερμοκρασίας είναι πρώτη στο σύνολο της χειμερινής περιόδου είναι τρίτη.

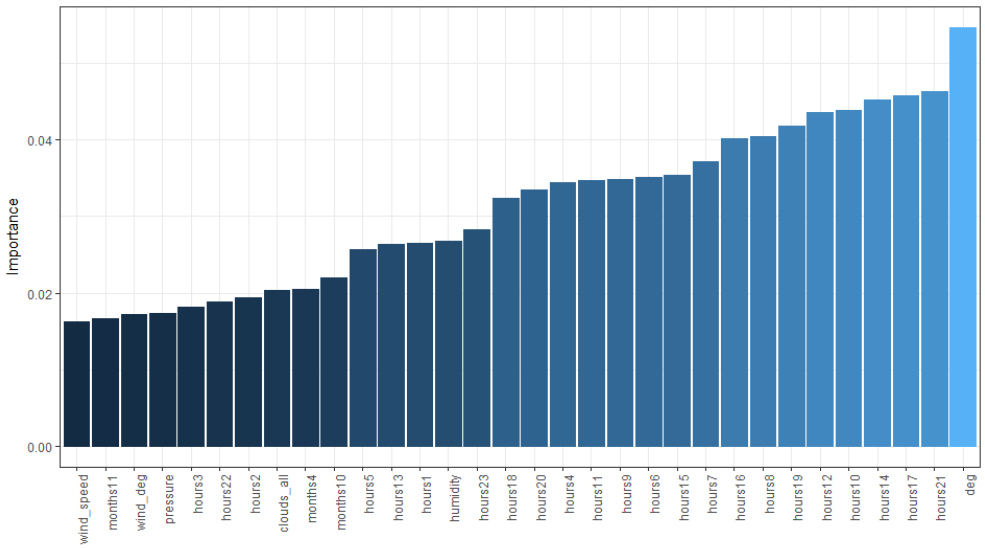
Συγκρίνοντας τις τιμές του RMSE για τα σύνολα εκπαίδευσης και ελέγχου και τις τιμές του R^2, οι οποίες δεν διαφέρουν σημαντικά μεταξύ τους για κάθε μοντέλο, καταλήγουμε στο συμπέρασμα ότι το πιο αποδοτικό για τους μήνες Δεκέμβριο, Ιανουάριο και Φεβρουάριο είναι το νευρωνικό δίκτυο με $RMSE = 7.9105$ στο σύνολο ελέγχου και $R^2 = 0.8677$. Για τους μήνες Οκτώβριο, Νοέμβριο, Μάρτιο και Απρίλιο το μοντέλο με τις καλύτερες προβλέψεις στο σύνολο εκπαίδευσης και ελέγχου είναι το SVR με πολυωνυμικό πυρήνα δεύτερου βαθμού. Έχει $RMSE = 6.5697$ στο σύνολο ελέγχου και $R^2 = 0.817$.

Έχοντας βρει τα καλύτερα μοντέλα χρησιμοποιώντας μόνο τα δεδομένα της Δ.Ε.ΤΗ.Π. θέλουμε να εξετάσουμε κατά πόσο θα βελτιωθεί η απόδοση τους εάν τα τροφοδοτήσουμε με περισσότερη πληροφορία σχετικά και με άλλους καιρικούς παράγοντες όπως ατμοσφαιρική πίεση και υγρασία. Θα προσθέσουμε στα δεδομένα της Δ.Ε.ΤΗ.Π. τα δεδομένα από το OpenWeatherMap και θα κατασκευάσουμε δύο νέα μοντέλα, ένα για κάθε σύνολο δεδομένων για να δούμε αν θα μειωθεί κι άλλο το RMSE.

Για το χειμερινό σύνολο δεδομένων θα κατασκευαστεί ένα νευρωνικό δίκτυο με 7 κόμβους στο κρυφό επίπεδο ενώ για το δεύτερο σύνολο ένα SVR με πολυωνυμικό πυρήνα δεύτερου βαθμού και ρυθμιστική παράμετρο κόστους ίση με 1. Όμως επειδή θέλουμε να ελέγξουμε πόσο επηρεάζει κάθε καιρικός παράγοντας το σύνολο δεδομένων για τους μήνες του φθινοπώρου και της άνοιξης, θα κατασκευάσουμε επίσης ένα νευρωνικό δίκτυο με 10 κόμβους στο κρυφό επίπεδο για τα δεδομένα αυτά.

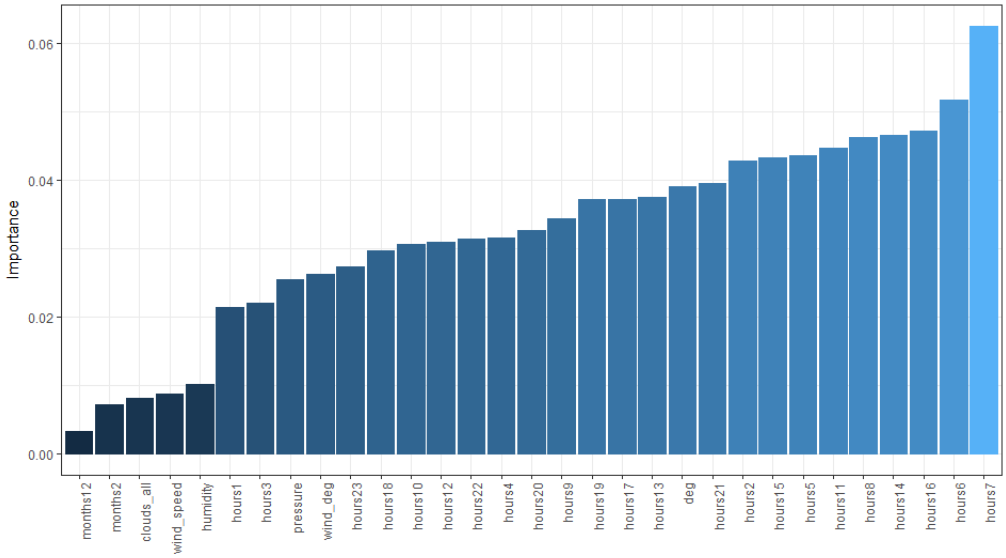
4.3.5. Μοντέλα πρόβλεψης για όλους τους καιρικούς παράγοντες

Για το μοντέλο SVR με πολυωνυμικό πυρήνα δεύτερου βαθμού στα δεδομένα των μηνών Οκτωβρίου, Νοεμβρίου, Μαρτίου και Απριλίου παρατηρούμε από τον πίνακα 6 ότι πετυχαίνει χαμηλότερο σφάλμα. Βέβαια η διαφορά από το προηγούμενο καλύτερο μοντέλο δεν είναι αρκετά σημαντική. Για το νευρωνικό δίκτυο του ίδιου συνόλου δεδομένων βλέπουμε ότι το σφάλμα του είναι σχεδόν ίδιο με εκείνο του άλλου νευρωνικού δικτύου με τα λιγότερα δεδομένα. Όμως σε σχέση με το μοντέλο SVR που κατασκευάστηκε από τα ίδια δεδομένα υστερεί σε απόδοση. Ο λόγος που κατασκευάστηκε είναι για να εξετάσουμε πόσο ωφέλιμοι είναι οι υπόλοιποι καιρικοί παράγοντες. Αυτό φαίνεται στο Σχήμα 42. Οι νέοι καιρικοί παράγοντες βρίσκονται αρκετά χαμηλά στην κατάταξη. Στο επόμενο κεφάλαιο θα αναλυθεί αν αξίζει να συμπεριλάβουμε αυτά τα δεδομένα με σκοπό να παράγουμε πιο αποδοτικά μοντέλα.



Σχήμα 42. Φθινόπωρο & Άνοιξη, σημαντικότητα μεταβλητών νευρωνικού δικτύου με χρήση όλων των καιρικών παραγόντων

Για το χειμερινό σύνολο δεδομένων βλέπουμε πως το νευρωνικό δίκτυο πετυχαίνει μικρότερο σφάλμα $RMSE = 7.4477$ σε σχέση με το προηγούμενο. Η διαφορά είναι σχεδόν μισή μονάδα το οποίο είναι μία αξιόλογη βελτίωση. Η σημαντικότητα των μεταβλητών φαίνεται στο Σχήμα 43.



Σχήμα 43. Χειμώνας, σημαντικότητα μεταβλητών νευρωνικού δικτύου με χρήση όλων των καιρικών παραγόντων

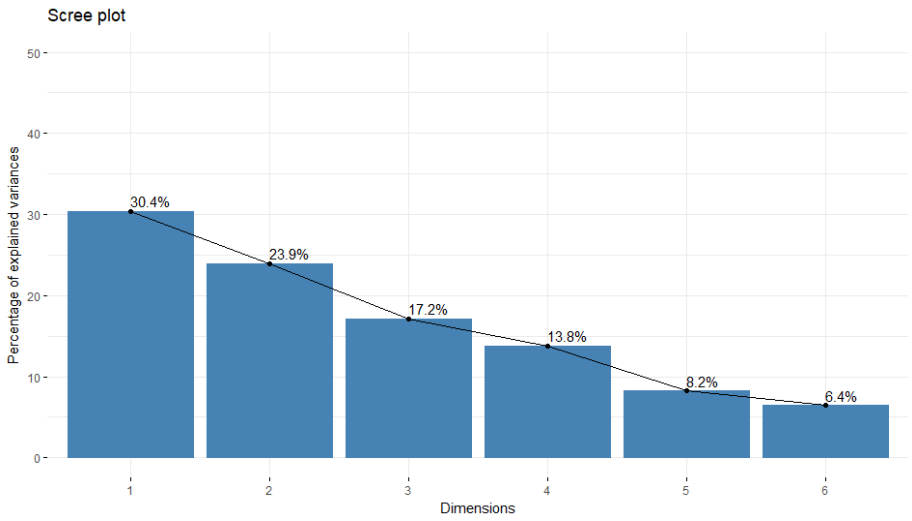
Όπως πριν έτσι και σε αυτό το νευρωνικό δίκτυο οι νέοι καιρικοί παράγοντες βρίσκονται σε χαμηλές θέσεις. Αυτό που αξίζει να αναφερθεί είναι η υποχώρηση της εξωτερικής θερμοκρασίας από την πρώτη στην εντέκατη θέση.

4.3.6. Μοντέλα πρόβλεψης με χρήση PCA

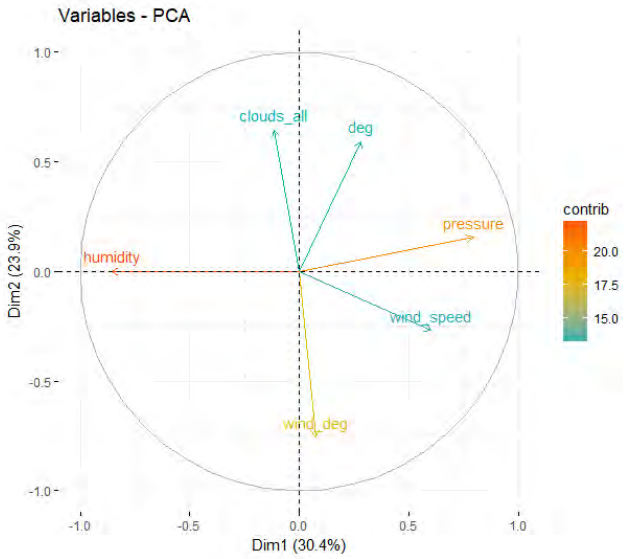
Από την προηγούμενη ενότητα είδαμε ότι η χρήση περισσότερων καιρικών παραγόντων πέραν της εξωτερικής θερμοκρασίας μπορεί να δώσει καλύτερα αποτελέσματα πρόβλεψης. Όμως ο όγκος των δεδομένων προς επεξεργασία αυξήθηκε αρκετά και αυτό επηρεάζει το υπολογιστικό κόστος. Θέλουμε λοιπόν να εξετάσουμε κατά πόσο η μείωση της διάστασης των δεδομένων επηρεάζει την απόδοση των μοντέλων. Μέχρι τώρα χρησιμοποιήθηκαν 6 καιρικοί παράγοντες, εξωτερική θερμοκρασία, ατμοσφαιρική πίεση, υγρασία, ταχύτητα ανέμου, κατεύθυνση ανέμου και νέφωση. Θα

χρησιμοποιήσουμε την μέθοδο της ανάλυσης κύριων συνιστωσών (PCA) για να μειώσουμε την διάσταση από 6 σε κάποια μικρότερη. Όμως θα κρατήσουμε εκτός τις μεταβλητές της ώρας και του μήνα, τις οποίες θα ενσωματώσουμε στην συνέχεια στις εναπομείναντες κύριες συνιστώσες. Η επιλογή αυτή έγινε για να μην χαθεί η πληροφορία από αυτές τις δύο μεταβλητές. Επειδή έχουν διακριτές τιμές ενώ οι καιρικοί παράγοντες συνεχείς, αν στην PCA χρησιμοποιούσαμε όλες τις μεταβλητές θα χάναμε την δυνατότητα να ξεχωρίζουμε τις ώρες και τους μήνες.

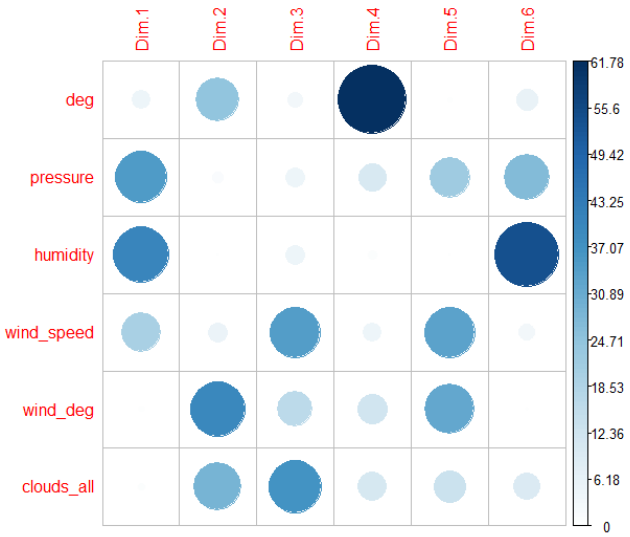
Για το χειμερινό σύνολο δεδομένων το ποσοστό της διακύμανσης (variance) που εξηγείται από κάθε κύρια συνιστώσα φαίνεται στο Σχήμα 44. Παρατηρούμε ότι στις πρώτες 4 συνιστώσες βρίσκεται το 85.31% της διακύμανσης των δεδομένων. Με άλλα λόγια στις πρώτες 4 συνιστώσες υπάρχει το 85.31% της πληροφορίας των αρχικών δεδομένων. Στο Σχήμα 45 βλέπουμε την διεύθυνση και το μέτρο κάθε καιρικού παράγοντα ως προς τις πρώτες δύο κύριες συνιστώσες. Ακόμη στο Σχήμα 46 φαίνεται η προσφορά κάθε μεταβλητής στις κύριες συνιστώσες.



Σχήμα 44. Χειμώνας, PCA scree plot



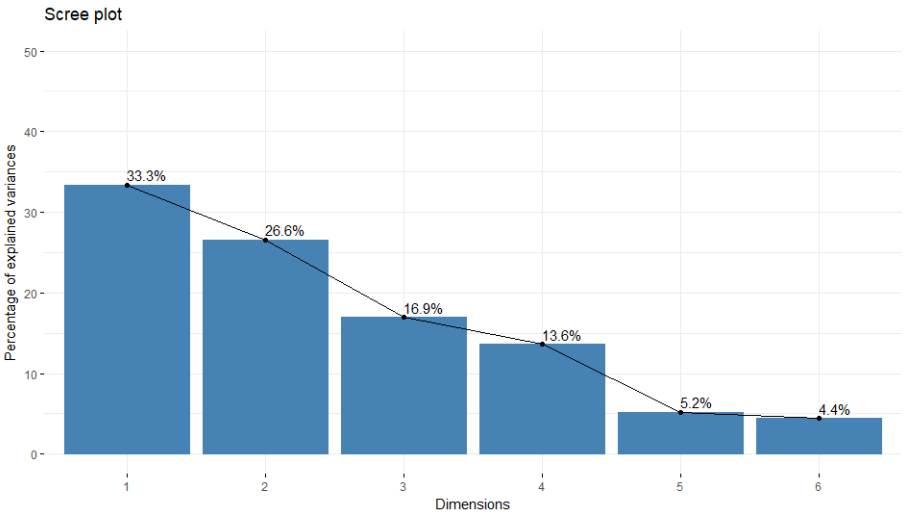
Σχήμα 45. Χειμώνας, μέγεθος και κατεύθυνση των μεταβλητών ως προς τις πρώτες δύο κύριες συνιστώσες



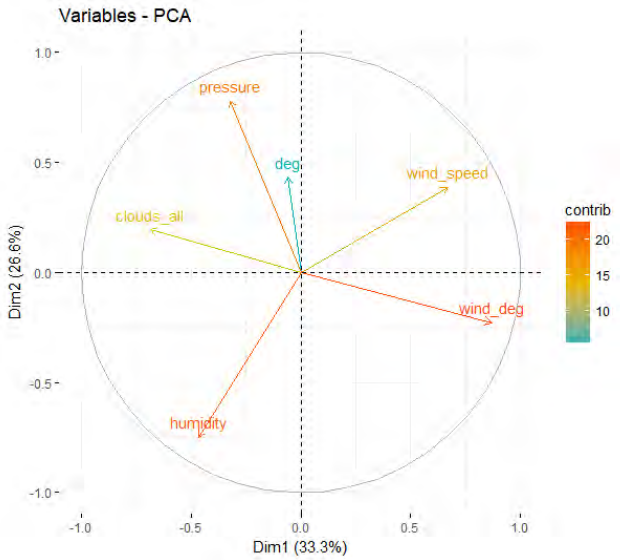
Σχήμα 46. Χειμώνας, προσφορά κάθε μεταβλητής στις κύριες συνιστώσες

Θα επιλεγούν οι πρώτες 4 κύριες συνιστώσες επειδή σε αυτές υπάρχει το 85.31% της πληροφορίας το οποίο είναι ένα μεγάλο ποσοστό. Επίσης στις πρώτες 4 κύριες συνιστώσες υπάρχει πληροφορία από κάθε καιρικό παράγοντα. Τέλος θα προκύψει ένα σύνολο δεδομένων 6 διαστάσεων, 4 κύριες συνιστώσες, ώρες και μήνες, το οποίο σε σχέση με το αρχικό είναι κατά δύο διαστάσεις μικρότερο.

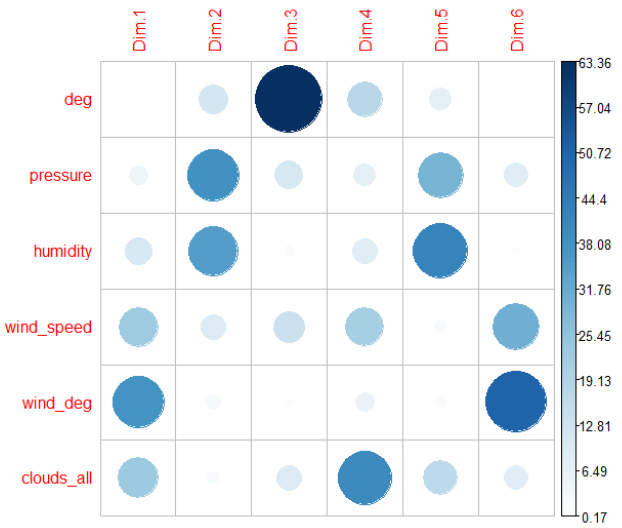
Για το σύνολο δεδομένων των υπόλοιπων μηνών οι αντίστοιχες γραφικές παραστάσεις υπάρχουν στα Σχήματα 47, 48 και 49. Επιλέχθηκαν οι πρώτες 4 κύριες συνιστώσες για τους ίδιους λόγους όμως σε αυτή την περίπτωση το ποσοστό της πληροφορίας είναι μεγαλύτερο 90.43%.



Σχήμα 47. Φθινόπωρο & Άνοιξη, PCA scree plot

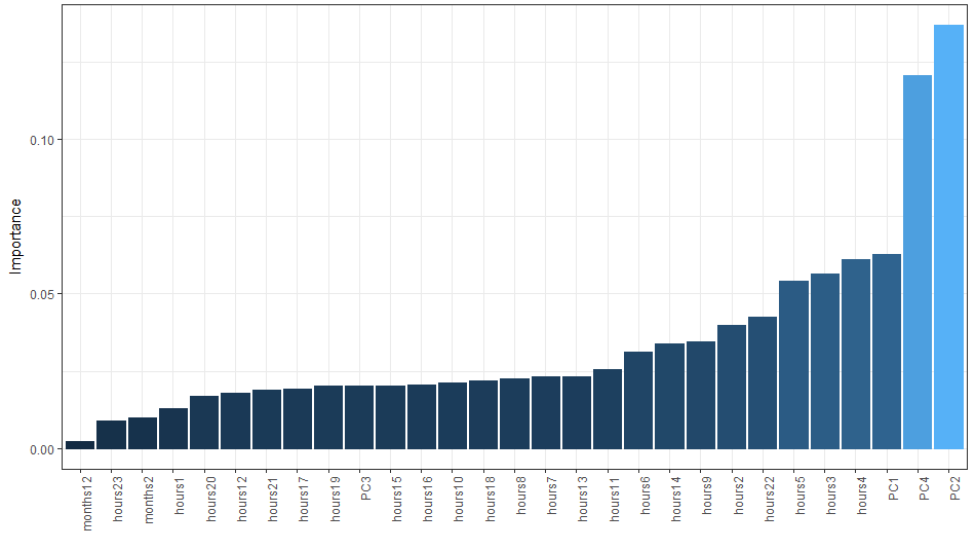


Σχήμα 48. Φθινόπωρο & Άνοιξη, μέγεθος και κατεύθυνση των μεταβλητών ως προς τις πρώτες δύο κύριες συνιστώσες

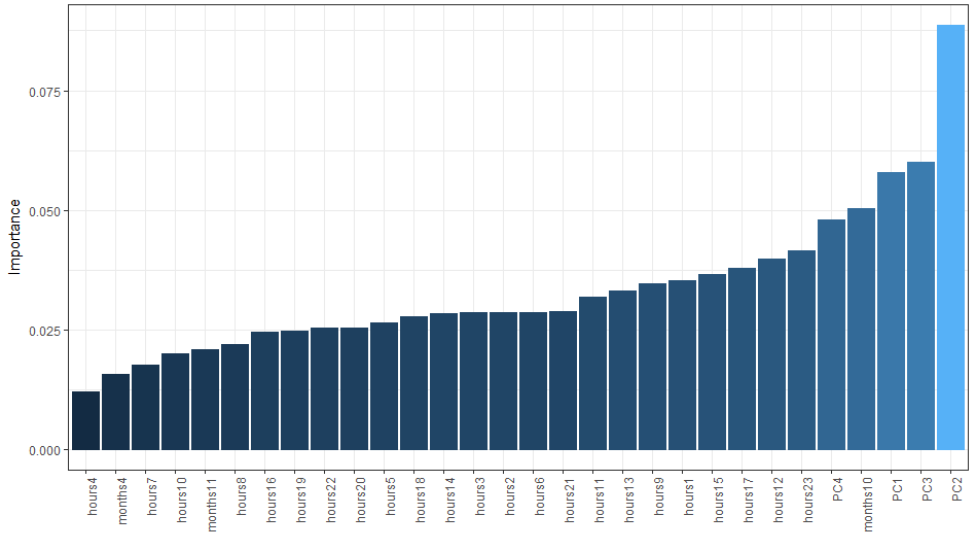


Σχήμα 49. Φθινόπωρο & Άνοιξη, προσφορά κάθε μεταβλητής στις κύριες συνιστώσες

Με τα νέα σύνολα δεδομένων κατασκευάστηκαν τρία μοντέλα πρόβλεψης ένα για τους χειμερινούς μήνες και δύο για τους υπόλοιπους. Για τους χειμερινούς φτιάχτηκε ένα νευρωνικό δίκτυο με 7 κόμβους στο κρυφό επίπεδο. Για το άλλο σύνολο κατασκευάστηκε ένα μοντέλο με την μέθοδο SVR με πολυωνυμικό πυρήνα δεύτερου βαθμού και ένα νευρωνικό δίκτυο με 10 κόμβους στο κρυφό επίπεδο. Τα σφάλματα των μοντέλων καθώς και το R^2 φαίνονται στους πίνακες 3 και 5. Βλέπουμε πως οι αποδόσεις τους είναι χαμηλότερες ακόμα και από εκείνες των πρώτων μοντέλων που χρησιμοποιούσαν μόνο τα δεδομένα της Δ.Ε.ΤΗ.Π. Άρα η επιλογή μίας ή δύο κύριων συνιστωσών για να είναι ο όγκος των νέων δεδομένων περίπου ίδιος με των πρώτων θα έδινε ακόμα χειρότερες αποδόσεις αφού θα χρησιμοποιούνταν ακόμη λιγότερη πληροφορία. Από τα νευρωνικά δίκτυα των δύο συνόλων βλέπουμε την σημαντικότητα των νέων μεταβλητών, Σχήμα 50, 51. Παρατηρούμε πως και στα δύο δίκτυα οι 3 από τις 4 κύριες συνιστώσες είναι οι πιο σημαντικές, δηλαδή αυτές που προσφέρουν περισσότερο στην πρόβλεψη της ζήτησης. Αυτό είναι αρκετά σημαντικό διότι δείχνει ότι με την χρήση της PCA καταφέραμε να συγκεντρώσουμε ωφέλιμη πληροφορία σε μικρότερο χώρο διαστάσεων από τον αρχικό.



Σχήμα 50. Χειμώνας, σημαντικότητα μεταβλητών νευρωνικού δικτύου με χρήση PCA



Σχήμα 51. Φθινόπωρο & Άνοιξη, σημαντικότητα μεταβλητών νευρωνικού δικτύου με χρήση PCA

Πίνακας 2. Χειμώνας, αποτελέσματα RMSE και R^2

Winter Dataset								
	Linear Regression		**Decision Trees**		**SVR Linear**		**SVR Polynomial**	
	Train	Test	Train	Test	Train	Test	Train	Test
RMSE	8.0636	8.2720	8.2571	8.4934	8.1263	8.3927	7.7674	8.0702
R^2	0.8543		0.8475		0.8523		0.8651	

Πίνακας 3. Χειμώνας, αποτελέσματα RMSE και R^2

Winter Dataset								
	SVR RBF		**NN**		**NN All Factors**		**NN PCA**	
	Train	Test	Train	Test	Train	Test	Train	Test
RMSE	7.7693	8.0515	7.6853	7.9105	7.4	7.4477	8.6591	8.4649
R^2	0.8645		0.8677		0.8775		0.831	

Πίνακας 4. Φθινόπωρο & Άνοιξη, αποτελέσματα RMSE και R2

Autumn & Spring Dataset								
	Linear Regression		**Decision Trees**		**SVR Linear**		**SVR Polynomial**	
	Train	Test	Train	Test	Train	Test	Train	Test
RMSE	7.9224	7.7535	7.6679	7.3443	7.9573	7.8109	6.9042	6.5697
R^2	0.7577		0.7744		0.7559		0.817	

Πίνακας 5. Φθινόπωρο & Άνοιξη, αποτελέσματα RMSE και R^2

Autumn & Spring Dataset								
	SVR RBF		**NN**		**NN All Factors**		**NN PCA**	
	Train	Test	Train	Test	Train	Test	Train	Test
RMSE	7.0234	6.6516	6.9973	6.7046	6.675	6.522	7.829	7.7362
R^2	0.8095		0.8105		0.8286		0.7642	

Πίνακας 6. Φθινόπωρο & Άνοιξη, αποτελέσματα RMSE και R^2

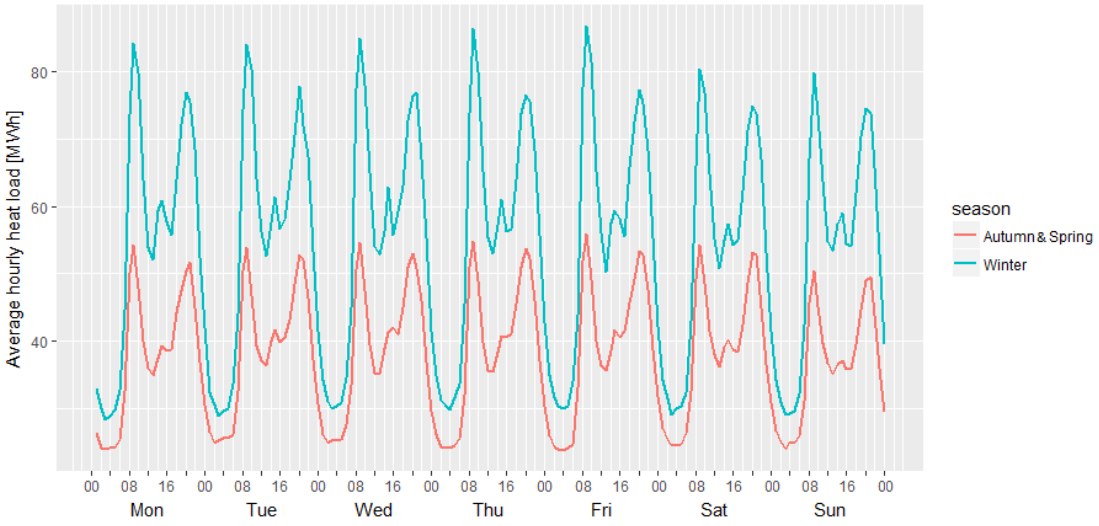
Autumn & Spring Dataset				
	SVR Polynomial All Factors		**SVR Polynomial PCA**	
	Train	Test	Train	Test
RMSE	6.4829	6.3819	8.2391	8.0039
R^2	0.8382		0.7380	

ΚΕΦΑΛΑΙΟ 5

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

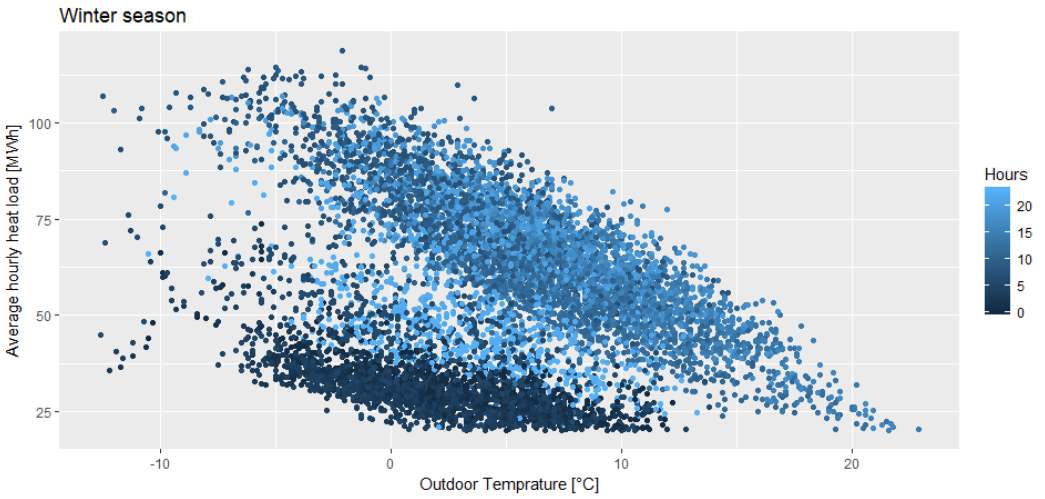
Στο κεφάλαιο αυτό παρουσιάζονται τα συμπεράσματα που προέκυψαν από την ανάλυση δεδομένων καθώς και από την δημιουργία των μοντέλων πρόβλεψης. Γίνονται συγκρίσεις μεταξύ των μοντέλων και ανακαλύπτεται πληροφορία για αυτά αλλά και για θέματα όπως την ενεργειακή απόδοση των κτηρίων και τις καθημερινές συνήθειες των κατοίκων.

Ξεκινώντας από τα δύο σύνολα δεδομένων πριν ακόμα κατασκευαστούν τα μοντέλα πρόβλεψης μπορούν να εξαχθούν συμπεράσματα σχετικά με την πολιτική λειτουργίας του δικτύου. Από τα Σχήματα 53 και 54 παρατηρούμε ότι ο μεγαλύτερος όγκος των δεδομένων συγκεντρώνεται σε δύο περιοχές. Αυτό οφείλεται στον τρόπο λειτουργίας του δικτύου ο οποίος ονομάζεται night setback control [41]. Το night setback control συμβαίνει όταν στους θερμοστάτες των κτηρίων ορίζεται μία χαμηλότερη εσωτερική θερμοκρασία για τις βραδινές ώρες με αποτέλεσμα η ζήτηση να είναι χαμηλότερη. Στο Σχήμα 52 φαίνεται η εβδομαδιαία μέση τιμή ζήτησης για τα δύο σύνολα δεδομένων. Η συγκεκριμένη μορφή της γραφικής παράστασης αντιστοιχεί σε δίκτυα με την λειτουργία night setback control. Παρατηρούμε πως υπάρχει κατακόρυφη πτώση στην ζήτηση για το διάστημα ωρών 23:00 – 06:00.

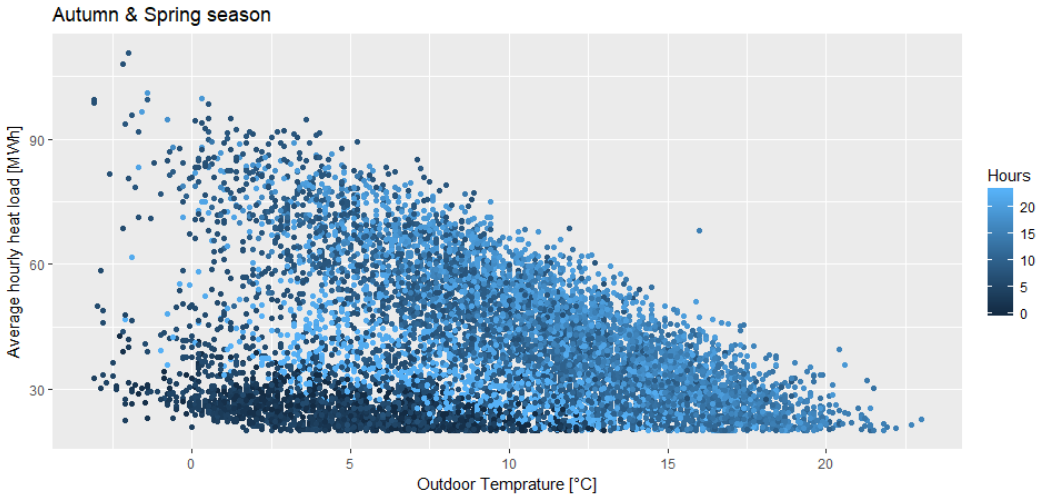


Σχήμα 52. Εβδομαδιαία μέση τιμή ζήτησης

Στα Σχήματα 53 και 54 οι εγγραφές που βρίσκονται στον κάτω μέρος των γραφικών παραστάσεων αντιστοιχούν σε βραδινές ώρες. Μεταξύ των δύο γραφικών παραστάσεων υπάρχει μία διαφορά. Για τους μήνες του φθινοπώρου και της άνοιξης ο άξονας της εξωτερικής θερμοκρασίας έχει μεγαλύτερες τιμές και δεν υπάρχει ξεκάθαρος διαχωρισμός μεταξύ των βραδινών και πρωινών ωρών. Ένας λόγος που συμβαίνει αυτό είναι ότι οι τιμές της ζήτησης των πρωινών ωρών βρίσκονται πιο κοντά σε αυτές των βραδινών στην περίπτωση του συνόλου δεδομένων για το φθινόπωρο και την άνοιξη. Όμως η μεγάλη διαφορά στο σύνολο δεδομένων του χειμώνα μπορεί να οδηγήσει σε συμπεράσματα σχετικά με τις θερμικές απώλειες των κτηρίων. Στην περίπτωση ενός κτηρίου με χαμηλές θερμικές απώλειες και με χρήση του night setback control η μείωση της εσωτερικής θερμοκρασίας δεν θα ήταν τόσο μεγάλη έτσι ώστε να υπάρξει σημαντική διαφορά στην ζήτηση μεταξύ βραδινών και πρωινών ωρών. Όμως σε αυτή την περίπτωση παρατηρούμε το αντίθετο. Άρα καταλήγουμε στο συμπέρασμα πως τα περισσότερα κτήρια της πόλης έχουν υψηλές θερμικές απώλειες. Στο πρώτο κεφάλαιο αναφερθήκαμε στην 4^{η} γενιά τηλεθέρμανσης. Μία σημαντική αλλαγή σε αυτή την γενιά είναι η κυκλοφορία του νερού στο δίκτυο σε χαμηλότερη θερμοκρασία. Αυτό για να συμβεί πρέπει τα κτήρια να έχουν ελάχιστες θερμικές απώλειες.

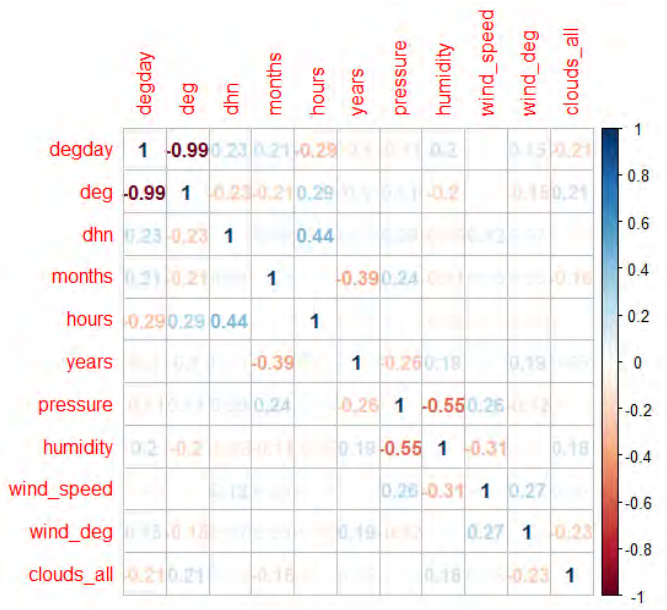


Σχήμα 53. Δεδομένα χειμώνα

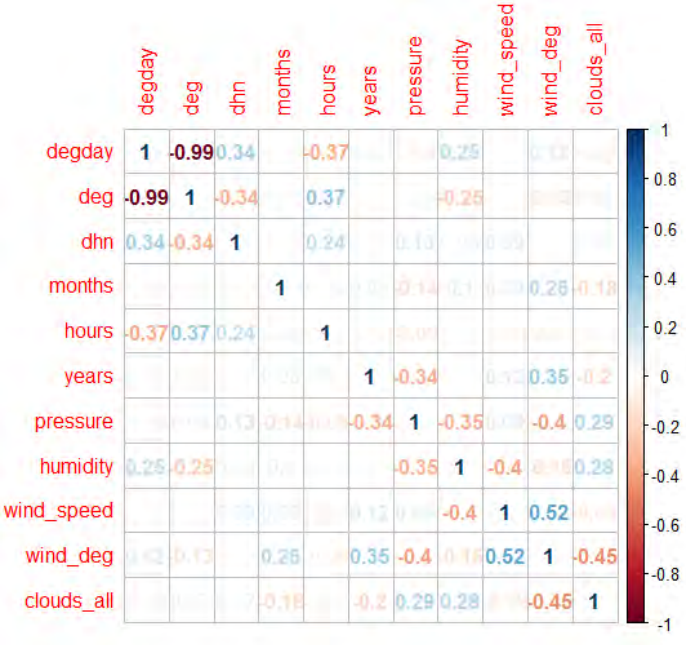


Σχήμα 54. Δεδομένα Φθινόπωρο & Άνοιξη

Εκτός από τα δεδομένα της Δ.Ε.ΤΗ.Π. και του OpenWeatherMap στην ανάλυση προστέθηκαν δύο ακόμη χαρακτηριστικά, η ώρα και ο μήνας. Μπορεί να υπήρχαν στο σύνολο δεδομένων με την μορφή της χρονοσφραγίδας αλλά δεν γινόταν να αξιοποιηθούν. Για να ελέγξουμε αν η επιλογή μας, να υπάρχουν αυτά τα χαρακτηριστικά στην ανάλυση, βοήθησε στην πρόβλεψη της ζήτησης και πρόσθεσε χρήσιμη πληροφορία θα υπολογίσουμε την συσχέτιση τους με τα υπόλοιπα χαρακτηριστικά και με την μεταβλητή πρόβλεψης. Χρησιμοποιώντας τον συντελεστή Pearson υπολογίσαμε την συσχέτιση, τα αποτελέσματα φαίνονται στα Σχήματα 55 και 56. Στο σύνολο δεδομένων για τους μήνες του χειμώνα η συσχέτιση μεταξύ της ώρας και της ζήτησης του δικτύου είναι 0.44 ενώ για το άλλο σύνολο είναι 0.24. Όπως είδαμε παραπάνω αυτό οφείλεται στον έντονο διαχωρισμό που υπάρχει στο σύνολο δεδομένων του χειμώνα εξαιτίας της ώρας. Τέλος και στα δύο σύνολα η τιμή συσχέτισης της ώρας (hours) με την μεταβλητή πρόβλεψης (dhn) είναι από τις μεγαλύτερες συγκριτικά με τις υπόλοιπες ανεξάρτητες μεταβλητές. Άρα η προσθήκη της στα μοντέλα πρόβλεψης είχε σημαντική βοήθεια.



Σχήμα 55. Χειμώνας, συσχέτιση μεταβλητών



Σχήμα 56. Φθινόπωρο & Άνοιξη, συσχέτιση μεταβλητών

Στο προηγούμενο κεφάλαιο τα αποτελέσματα παρουσιάστηκαν ανά μοντέλο πρόβλεψης. Σε αυτό το κεφάλαιο θα ομαδοποιήσουμε τις παρατηρήσεις ανά χρονική περίοδο έτσι ώστε να συγκρίνουμε τα μοντέλα μεταξύ τους έχοντας κοινό σύνολο δεδομένων. Θα ξεκινήσουμε με τους μήνες Δεκέμβριο, Ιανουάριο και Φεβρουάριο και

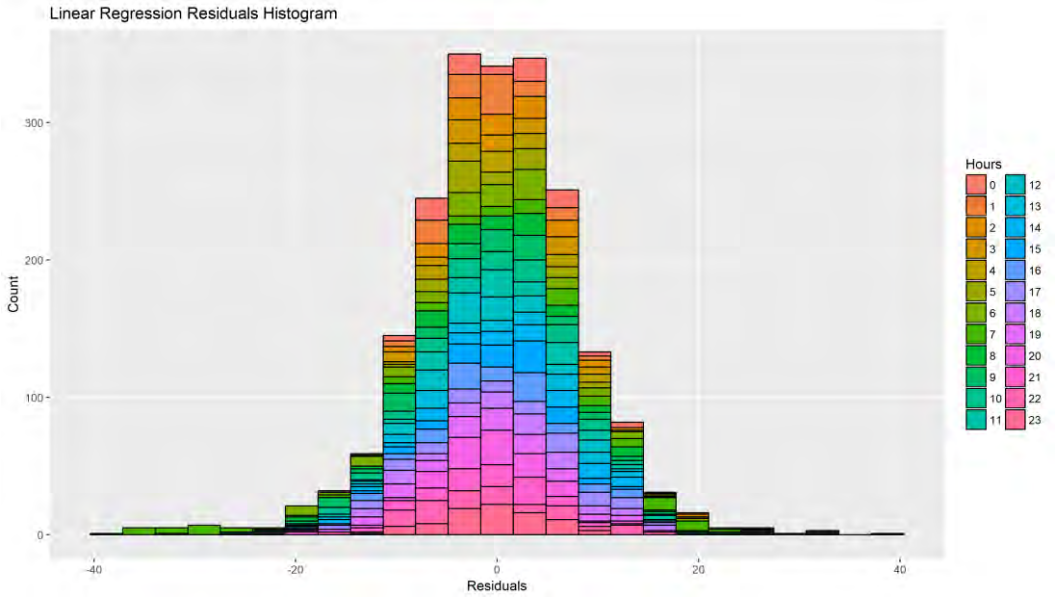
στην συνέχεια θα αναφερθούμε στο σύνολο δεδομένων για τους μήνες Οκτώβριο, Νοέμβριο, Μάρτιο και Απρίλιο. Για κάθε μοντέλο δημιουργήθηκε ένα ιστόγραμμα με το σφάλμα κάθε πρόβλεψης. Το σφάλμα συχνά αναφέρεται και ως υπόλοιπο αφού ορίζει την διαφορά της τιμής πρόβλεψης από την πραγματική. Για αυτό οι μπάρες του ιστογράμματος που βρίσκονται στο αρνητικό μέρος του οριζόντιου άξονα αντιστοιχούν στις περιπτώσεις όπου το μοντέλο προέβλεψε μία τιμή μεγαλύτερη από την πραγματική ενώ στο θετικό μέρος του άξονα συμβαίνει το αντίθετο. Δημιουργήθηκε επίσης μία γραφική παράσταση με ζεύγη δεδομένων. Κάθε ζεύγος περιέχει την τιμή πρόβλεψης και την πραγματική τιμή. Στον κατακόρυφο άξονα βρίσκονται οι τιμές των μοντέλων ενώ στον οριζόντιο οι πραγματικές τιμές. Υπάρχει επίσης η ευθεία $y = x$ η οποία δηλώνει μηδενικό σφάλμα και άλλες δύο ευθείες. Η πρώτη είναι η $y = x + rmse$ και η δεύτερη η $y = x - rmse$. Οι παρατηρήσεις εντός των δύο ευθειών έχουν σφάλμα μικρότερο ή ίσο με το συνολικό $RMSE$ του μοντέλου. Στο ιστόγραμμα και την γραφική παράσταση οι παρατηρήσεις ομαδοποιούνται με βάση την ώρα της ημέρας στην οποία αναφέρονται. Η ομαδοποίηση θα βοηθήσει στην εξαγωγή συμπερασμάτων σχετικά με την απόδοση του μοντέλου σε διαφορετικές ώρες.

5.1. Χειμώνας

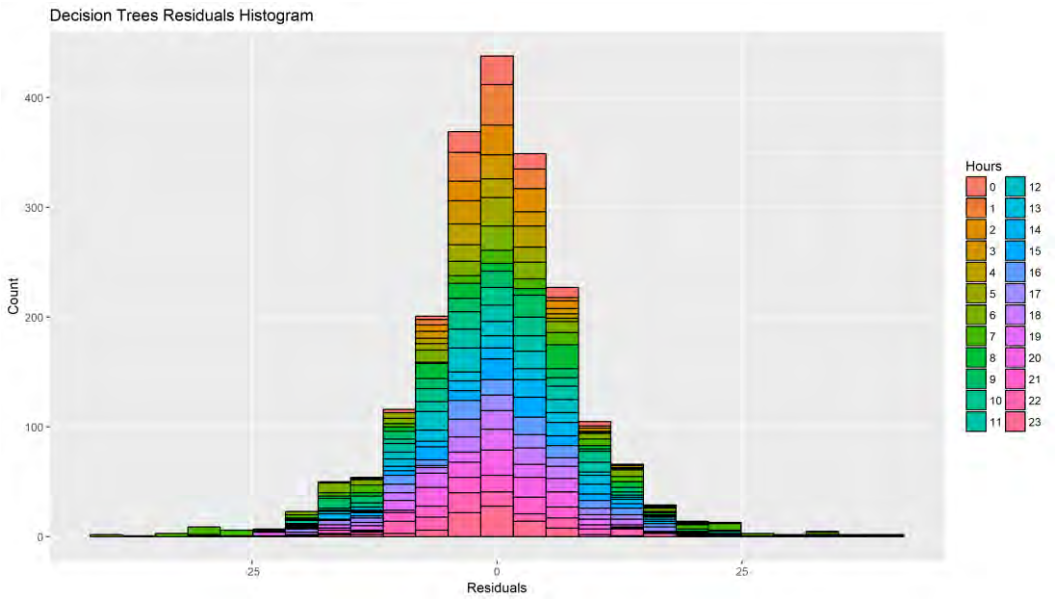
Θυμίζουμε για την περίοδο του χειμώνα κατασκευάστηκαν συνολικά 8 μοντέλα. Τα 6 χρησιμοποιώντας μόνο τα δεδομένα της Δ.Ε.ΤΗ.Π. και από την μέθοδο με το μικρότερο RMSE κατασκευάστηκαν άλλα δύο. Ένα χρησιμοποιώντας όλους τους καιρικούς παράγοντες και ένα με χρήση της ανάλυσης κύριων συνιστωσών (PCA). Στα Σχήματα 57 έως 64 υπάρχουν τα ιστόγραμμα για κάθε μοντέλο που κατασκευάστηκε. Ένα γενικό συμπέρασμα για όλα τα ιστογράμματα είναι πως ο κύριος όγκος των σφαλμάτων βρίσκεται κοντά στο μηδέν το οποίο δηλώνει πως τα μοντέλα έχουν ικανοποιητική απόδοση στο μεγαλύτερο μέρος της ημέρας. Όμως για το διάστημα ανάμεσα στις ώρες 06:00 και 08:00 τα σφάλματα απομακρύνονται αρκετά από το μηδέν. Επίσης στα περισσότερα μοντέλα για αυτές τις ώρες τα σφάλματα είναι περισσότερα στην αρνητική μεριά του άξονα σε σχέση με την θετική. Αυτό δηλώνει πως τα μοντέλα πρόβλεψης εκτίμησαν πως η ζήτηση του δικτύου θα είναι μεγαλύτερη από την πραγματική.

Είναι σημαντικό να υπάρχει μία ακριβής πρόβλεψη για τις ώρες αιχμής, δηλαδή περίπου από τις 06:00 μέχρι τις 08:00. Είναι το σημείο όπου η ζήτηση φτάνει στην μέγιστη τιμή της. Προτεραιότητα του δικτύου είναι να ικανοποιήσει την ζήτηση των κατοίκων της πόλης. Τα μοντέλα πρόβλεψης για τις ώρες αιχμής προβλέπουν μεγαλύτερη ζήτηση οπότε θα ικανοποιήσουν τις ανάγκες των πολιτών. Όμως οι μεγάλες αποκλίσεις από την πραγματική τιμή επηρεάζουν οικονομικά την εταιρία διαχείρισης του δικτύου.

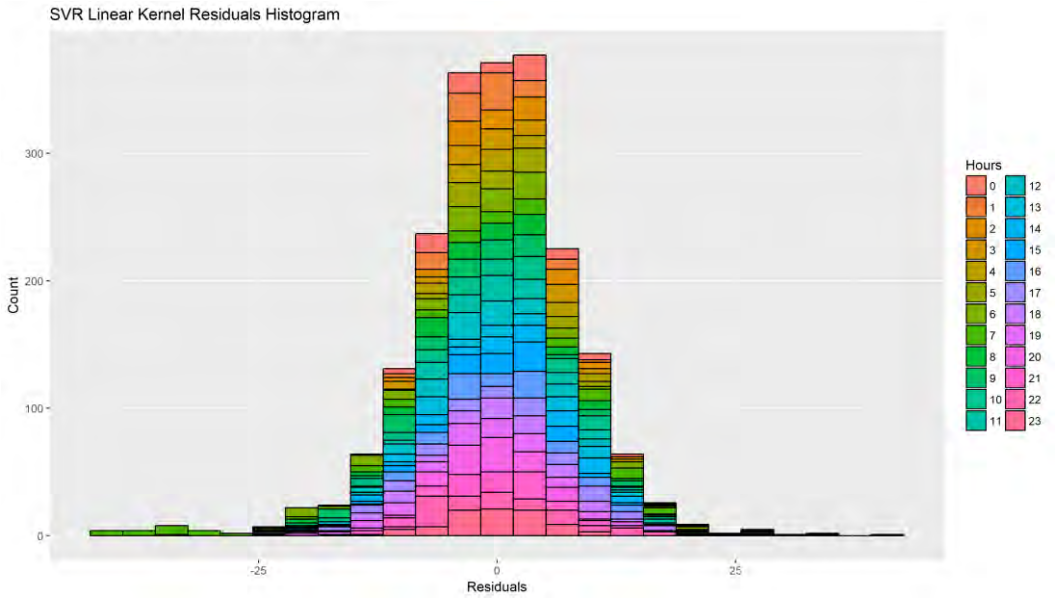
Συγκεκριμένα τα ιστογράμματα των μοντέλων πολλαπλής γραμμικής παλινδρόμησης, Σχήμα 57, μηχανών διανυσματικής υποστήριξης με γραμμικό πυρήνα, Σχήμα 59, και το νευρωνικό δίκτυο με τις κύριες συνιστώσες, Σχήμα 64, έχουν σφάλματα έως και 40MWh. Στα υπόλοιπα μοντέλα τα μέγιστα σφάλματα είναι κοντά στις 30MWh.



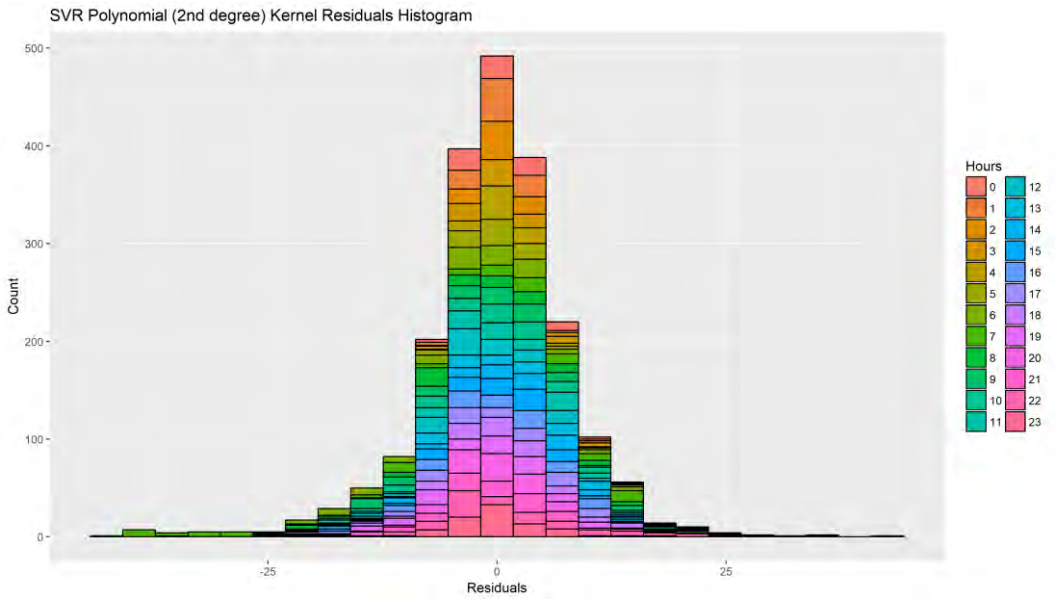
Σχήμα 57. Χειμώνας, ιστόγραμμα γραμμικής παλινδρόμησης



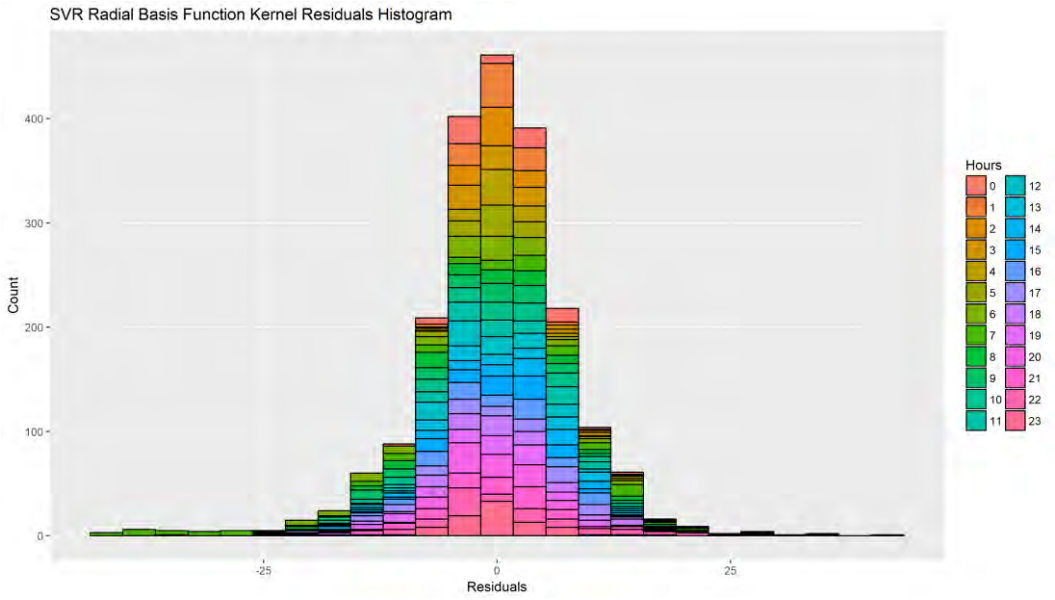
Σχήμα 58. Χειμώνας, ιστόγραμμα δέντρων απόφασης



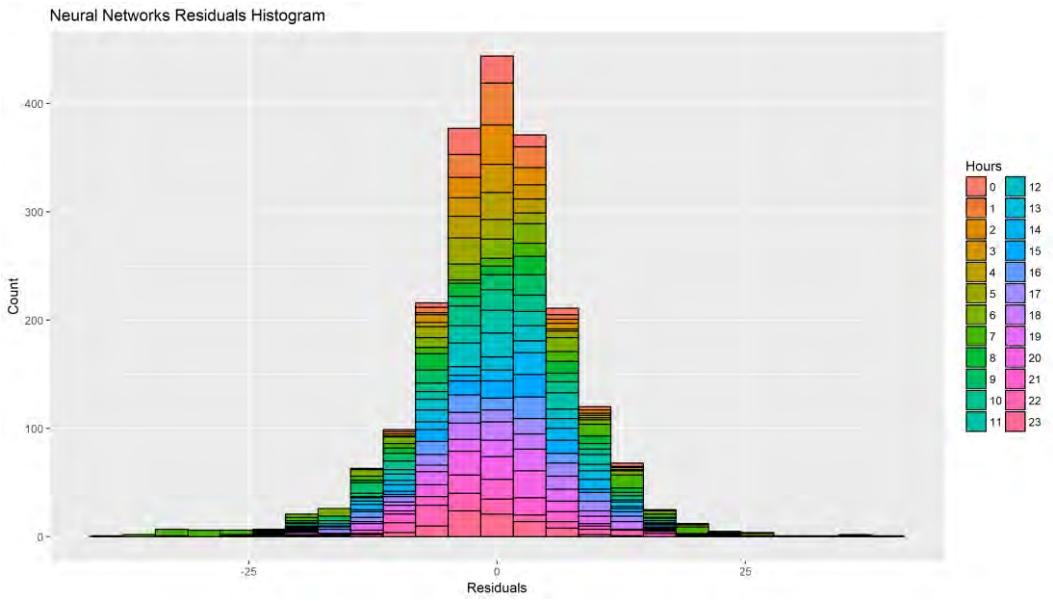
Σχήμα 59. Χειμώνας, ιστόγραμμα SVR με γραμμικό πυρήνα



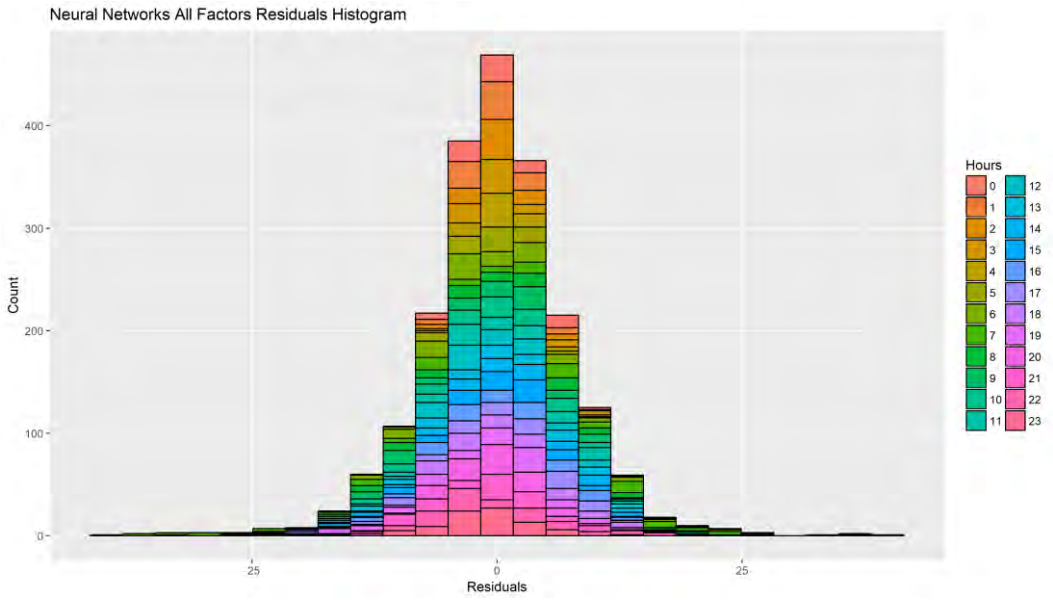
Σχήμα 60. Χειμώνας, ιστόγραμμα SVR με πολυωνυμικό πυρήνα



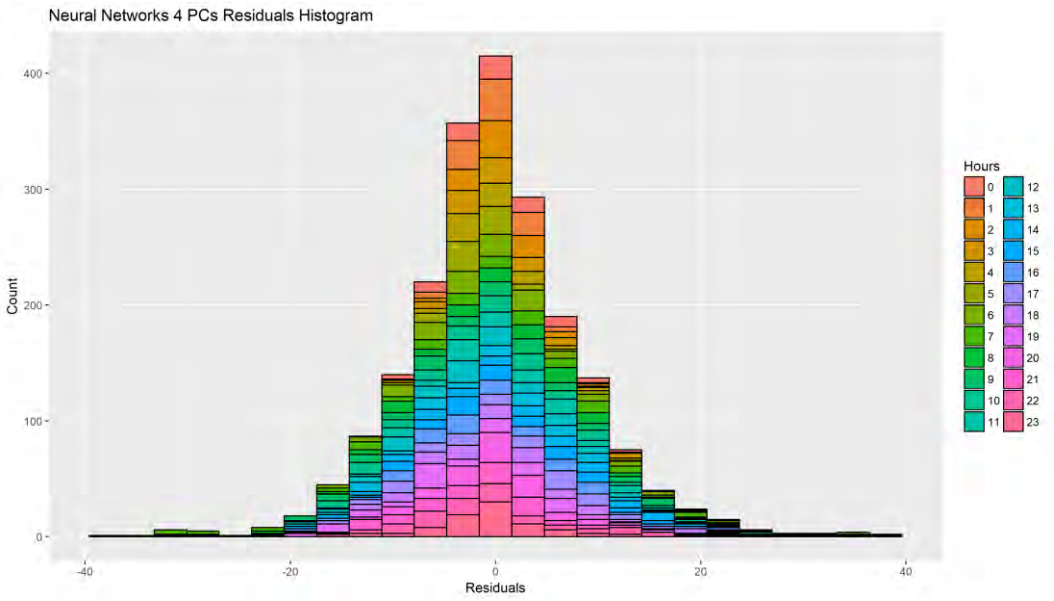
Σχήμα 61. Χειμώνας, ιστόγραμμα SVR με RBF πυρήνα



Σχήμα 62. Χειμώνας, ιστόγραμμα νευρωνικού δικτύου

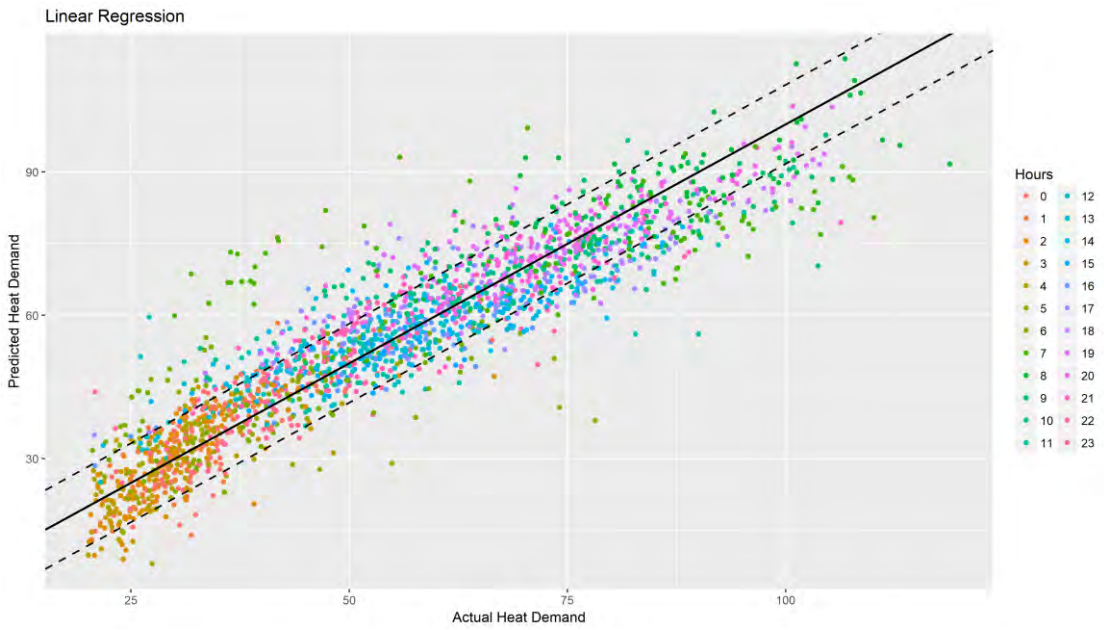


Σχήμα 63. Χειμώνας, ιστόγραμμα νευρωνικού δικτύου με όλους τους καιρικούς παράγοντες

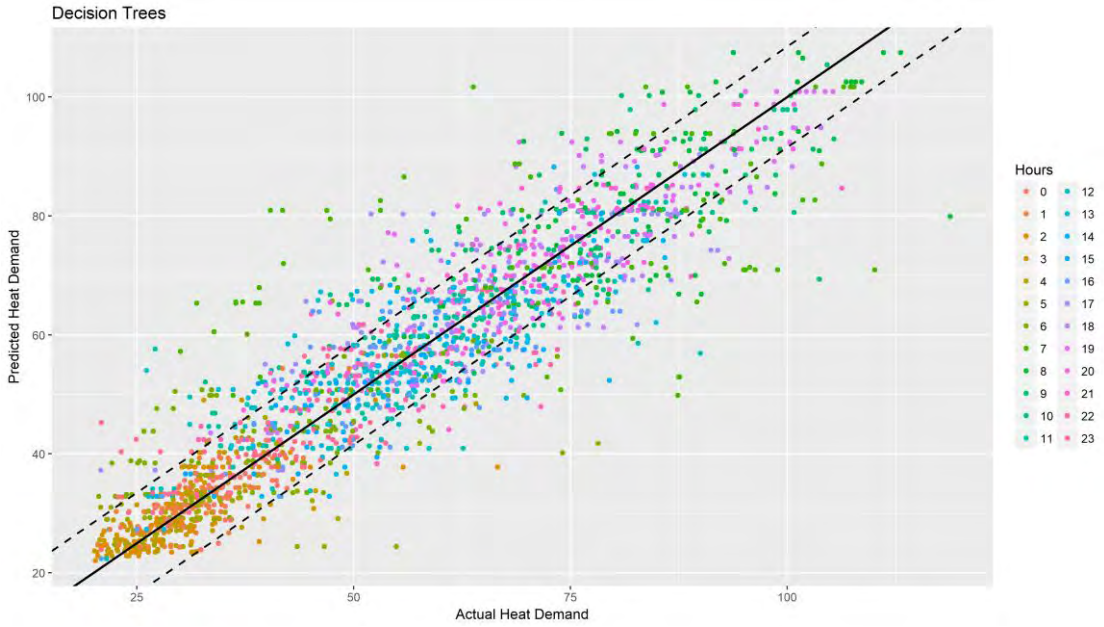


Σχήμα 64. Χειμώνας, ιστόγραμμα νευρωνικού δικτύου με χρήση PCA

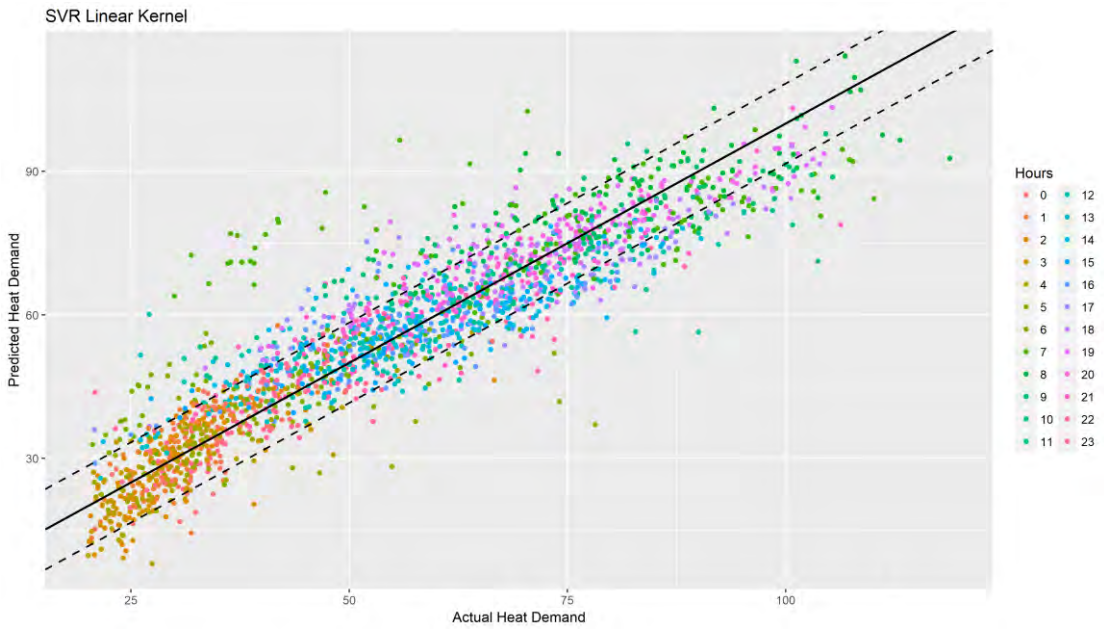
Από τα ιστογράμματα μπορούμε να καταλάβουμε μόνο το μέγεθος του σφάλματος και όχι σε ποια διαστήματα τιμών ζήτησης αναφέρονται. Με την γραφική παράσταση της τιμής πρόβλεψης ως προς την πραγματική τιμή θα εξάγουμε συμπεράσματα για την προέλευση των σφαλμάτων. Στα Σχήματα 65-72 παρουσιάζονται οι γραφικές παραστάσεις για όλα τα μοντέλα. Σε όλα τα μοντέλα οι εκτιμήσεις για τις ώρες 23:00 – 06:00 βρίσκονται εντός των δύο ευθειών και πολύ κοντά στην ευθεία μηδενικού σφάλματος. Επίσης εκτός από τις πρωινές ώρες αιχμής υπάρχει και δεύτερο διάστημα ωρών με μεγάλη ζήτηση. Αυτό είναι περίπου στις ώρες 18:00 με 20:00. Συγκεκριμένα τα δύο νευρωνικά δίκτυα, Σχήμα 70 και 71, εκτός από εκείνο με τις κύριες συνιστώσες, συγκεντρώνουν τις περισσότερες παρατηρήσεις εντός των δύο ευθειών. Αντίθετα στο νευρωνικό δίκτυο με τις κύριες συνιστώσες, Σχήμα 72, αρκετές παρατηρήσεις βρίσκονται μακριά από τις δύο ευθείες.



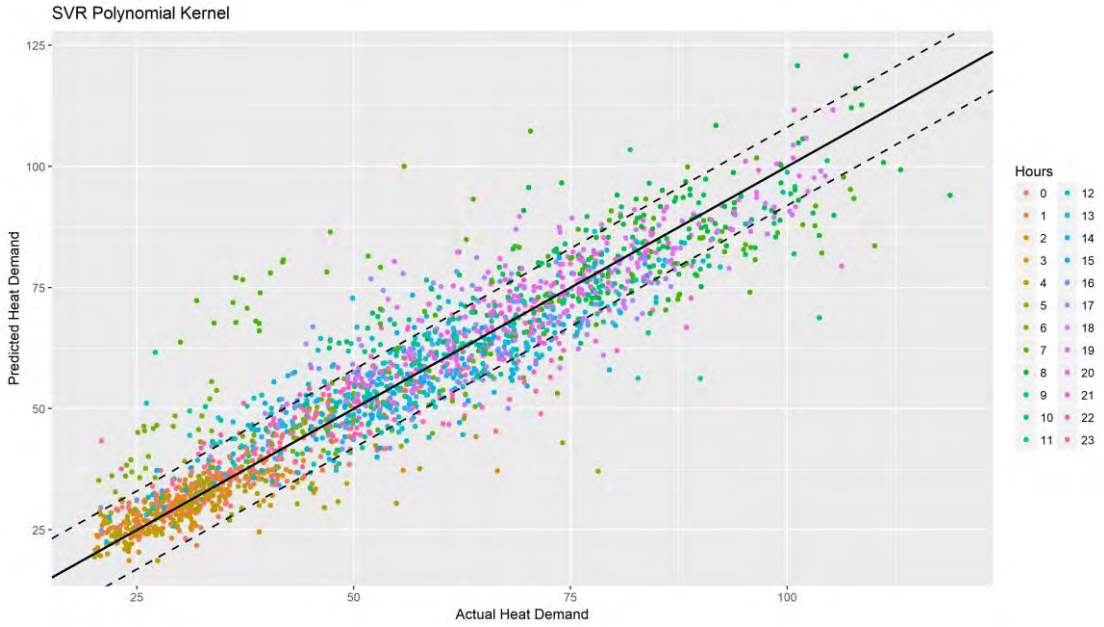
Σχήμα 65. Χειμώνας, γραμμική παλινδρόμηση



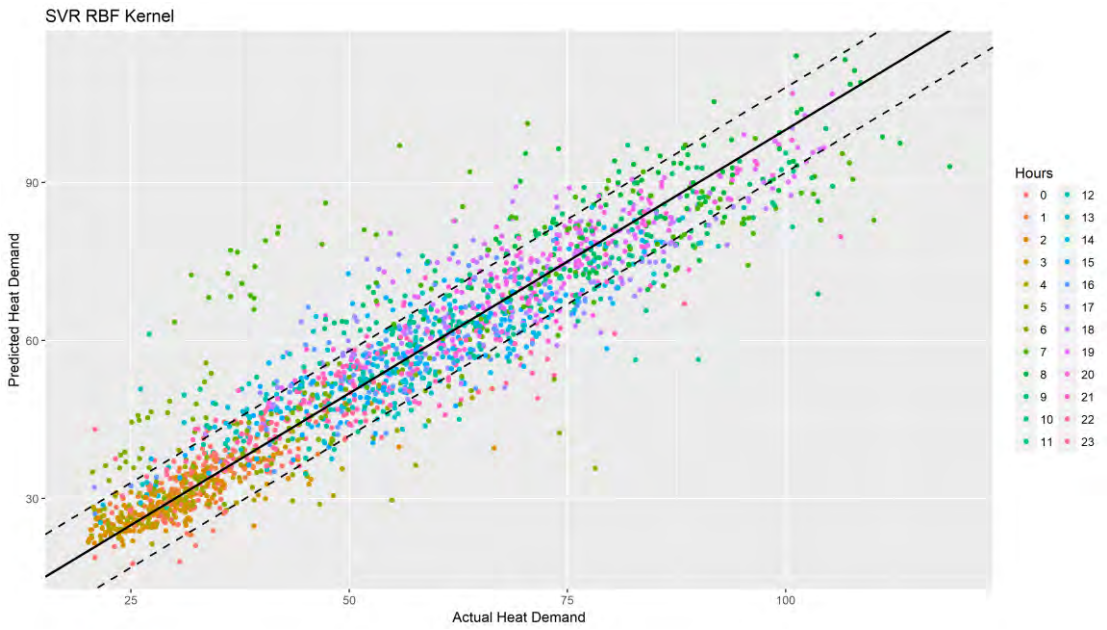
Σχήμα 66. Χειμώνας, δέντρα απόφασης



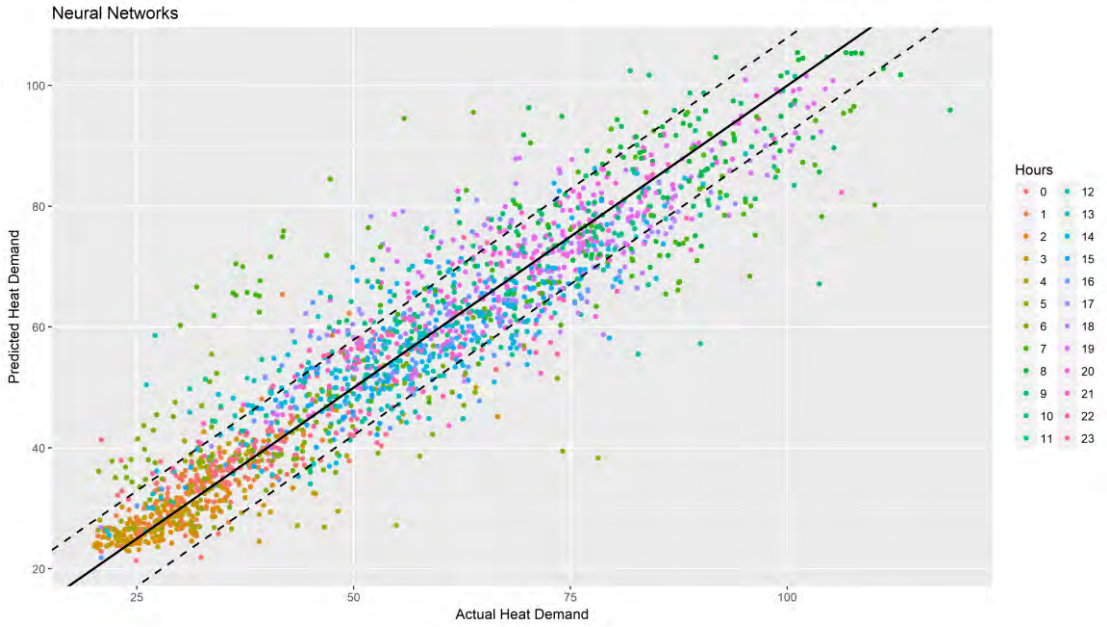
Σχήμα 67. Χειμώνας, SVR με γραμμικό πυρήνα



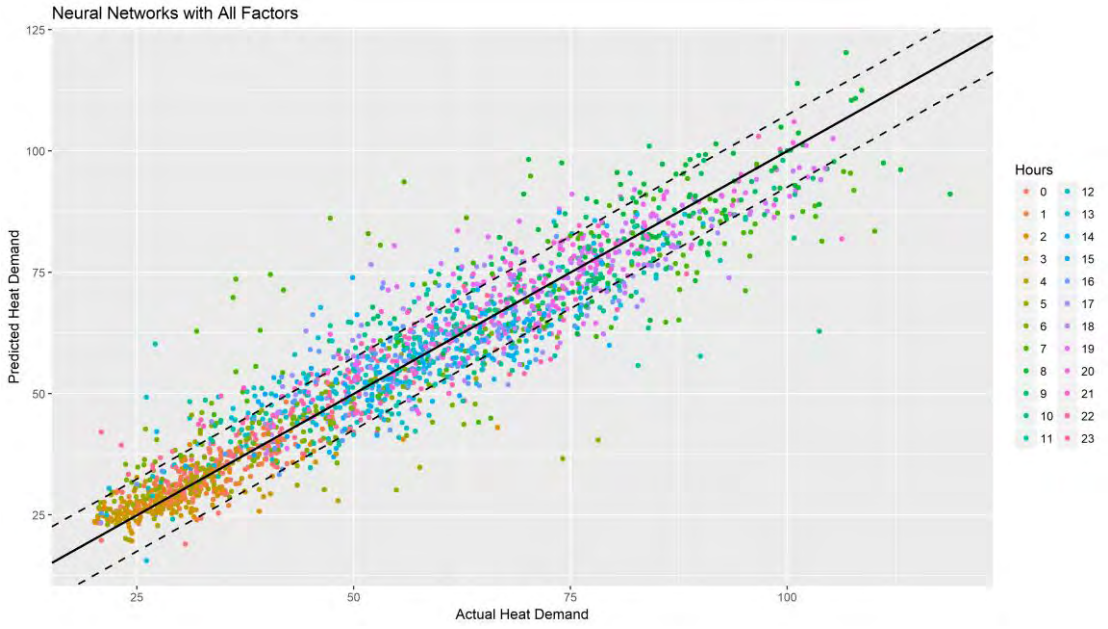
Σχήμα 68. Χειμώνας, SVR με πολυωνυμικό πυρήνα



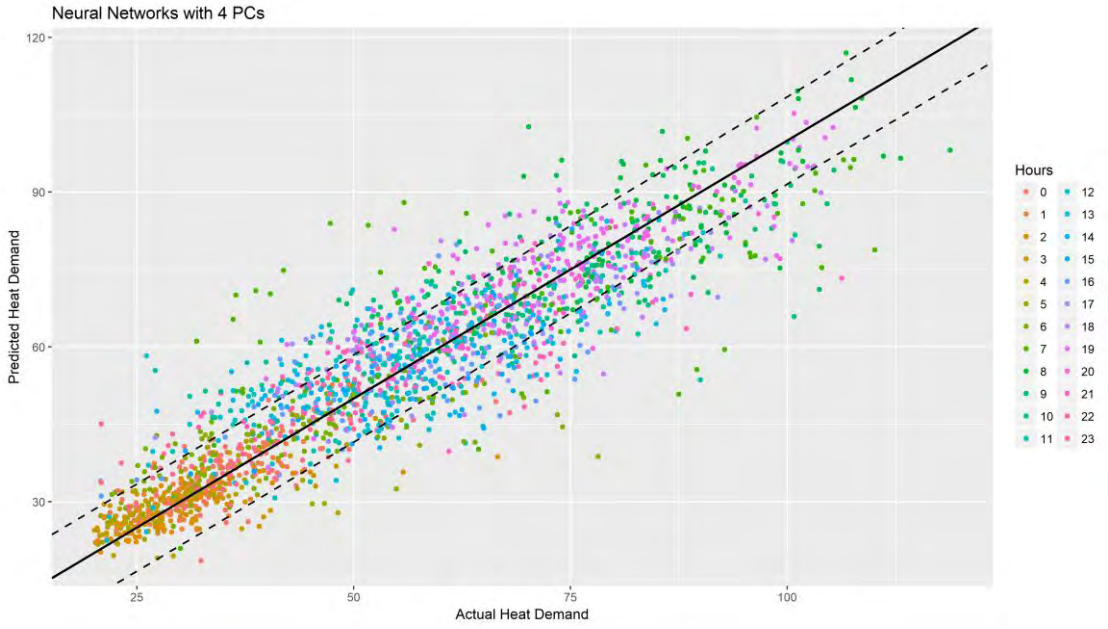
Σχήμα 69. Χειμώνας, SVR με RBF πυρήνα



Σχήμα 70. Χειμώνας, Νευρωνικό δίκτυο



Σχήμα 71. Χειμώνας, Νευρωνικό δίκτυο με όλους τους καιρικούς παράγοντες



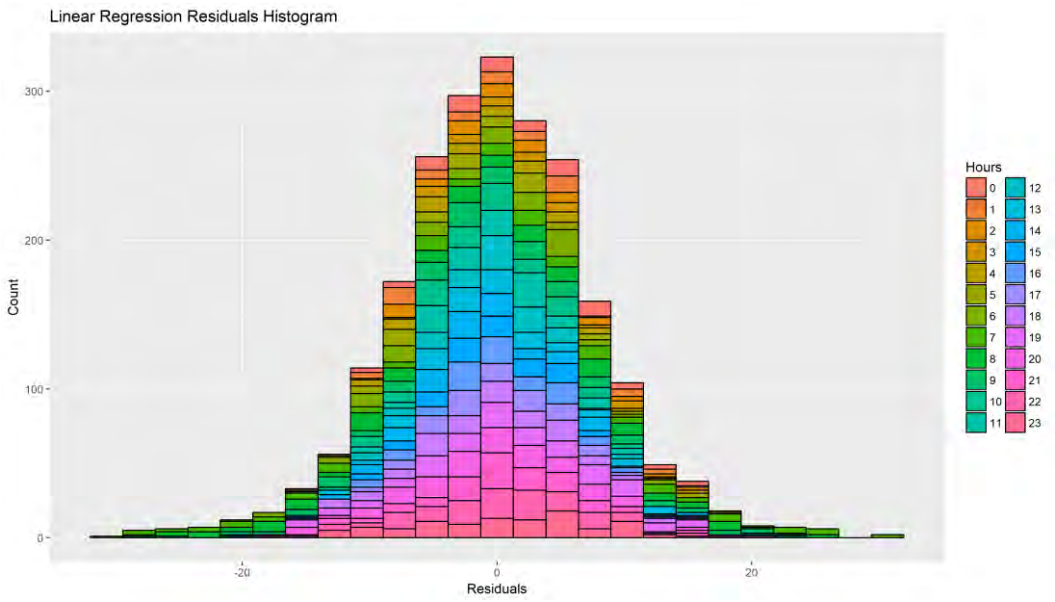
Σχήμα 72. Χειμώνας, Νευρωνικό δίκτυο με χρήση PCA

5.2. Φθινόπωρο & Άνοιξη

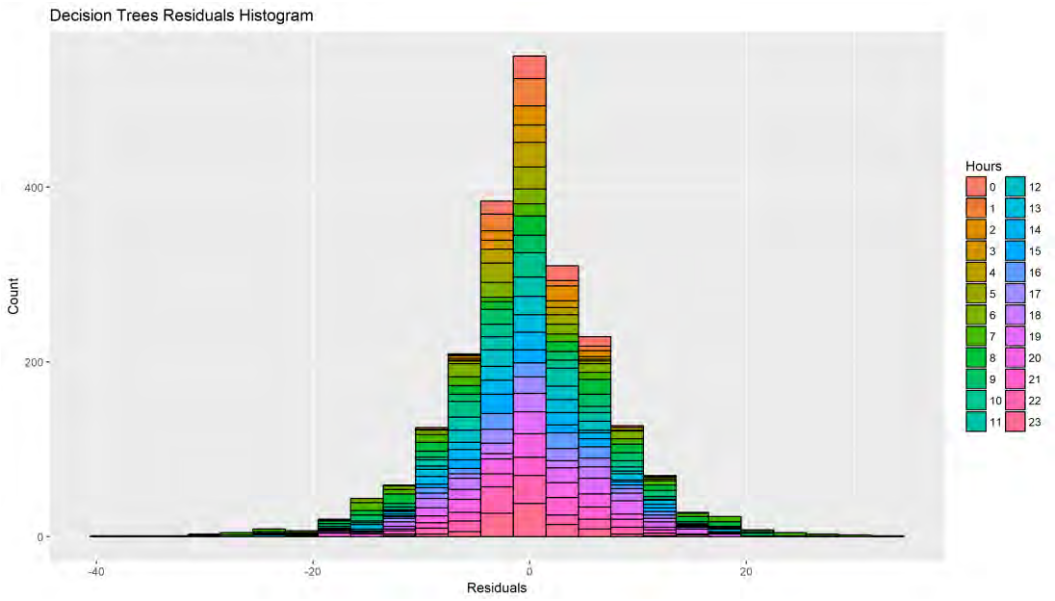
Για το σύνολο δεδομένων του φθινοπώρου και της άνοιξης κατασκευάστηκαν συνολικά 10 μοντέλα. Από τα βασικά 6 αυτό με το μικρότερο RMSE ήταν των μηχανών

διανυσματικής υποστήριξης με πολυωνυμικό πυρήνα. Οπότε με την μέθοδο αυτή κατασκευάστηκαν άλλα δύο μοντέλα ένα με όλους τους καιρικούς παράγοντες και ένα με ανεξάρτητες μεταβλητές τις πρώτες 4 κύριες συνιστώσες την ώρα και τον μήνα. Αντίστοιχα κατασκευάστηκαν δύο νευρωνικά δίκτυα με τα ίδια χαρακτηριστικά.

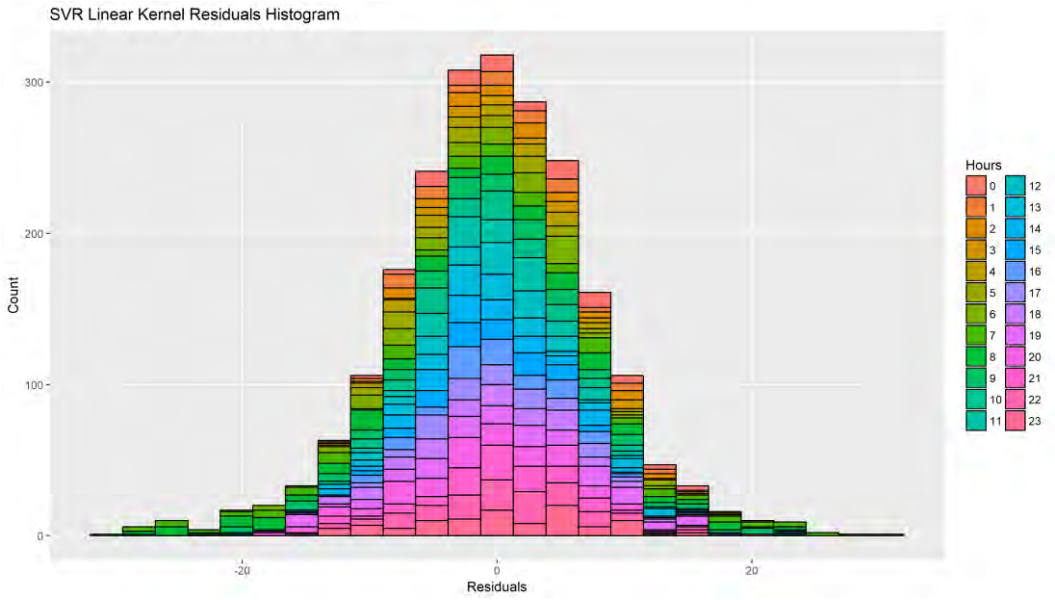
Τα ιστογράμματα των μοντέλων φαίνονται στα Σχήματα 73-82. Οι γενικές παρατηρήσεις σχετικά με το που βρίσκεται το μεγαλύτερο μέρος των σφαλμάτων είναι ίδιες με αυτές του προηγούμενου συνόλου δεδομένων. Τα μεγάλα σφάλματα αντιστοιχούν στις ώρες γύρω από τις 07:00. Αξίζει να αναφερθούμε στα ιστογράμματα των μοντέλων με χρήση της τεχνικής μηχανών διανυσματικής υποστήριξης με πολυωνυμικό πυρήνα δεύτερου βαθμού, Σχήμα 76. Το μοντέλο αυτό είχε το μικρότερο RMSE και η τεχνική αυτή επιλέχθηκε για να κατασκευαστεί το μοντέλο με όλους τους καιρικούς παράγοντες. Αν και συγκριτικά με τα άλλα μοντέλα έχουν τα περισσότερα σφάλματα κοντά στο 0, παράλληλα είναι εκείνα που έχουν και τα μεγαλύτερα σφάλματα. Συγκεκριμένα το μοντέλο με όλους τους καιρικούς παράγοντες, Σχήμα 81, έχει σφάλματα έως 45MWh μόνο στην αρνητική μεριά του άξονα ενώ αυτό με τα δεδομένα της Δ.Ε.ΤΗ.Π., Σχήμα 76, μέχρι 38MWh.



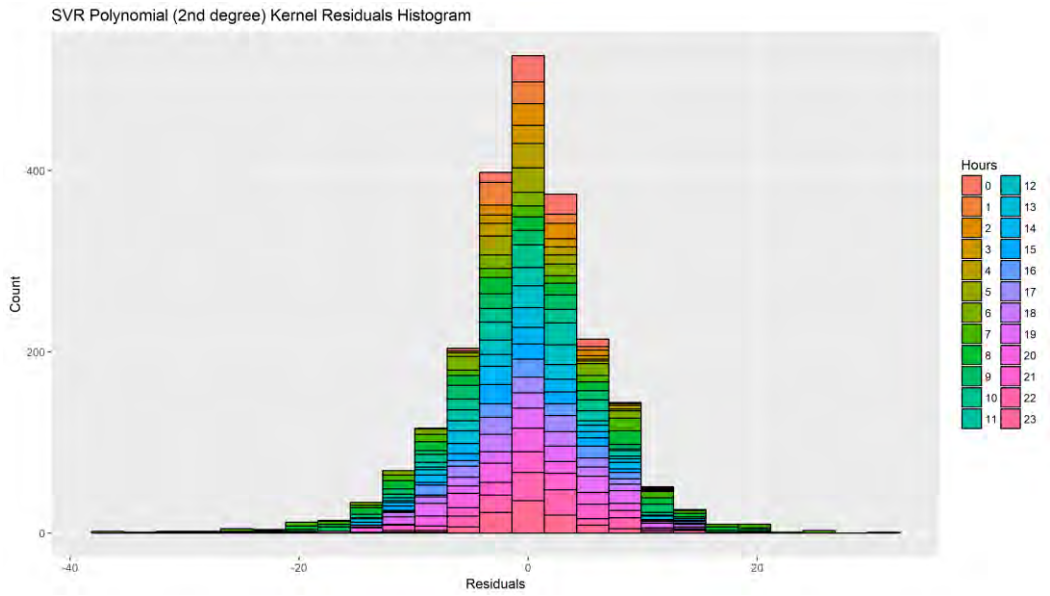
Σχήμα 73. Φθινόπωρο & Άνοιξη, ιστόγραμμα γραμμικής παλινδρόμησης



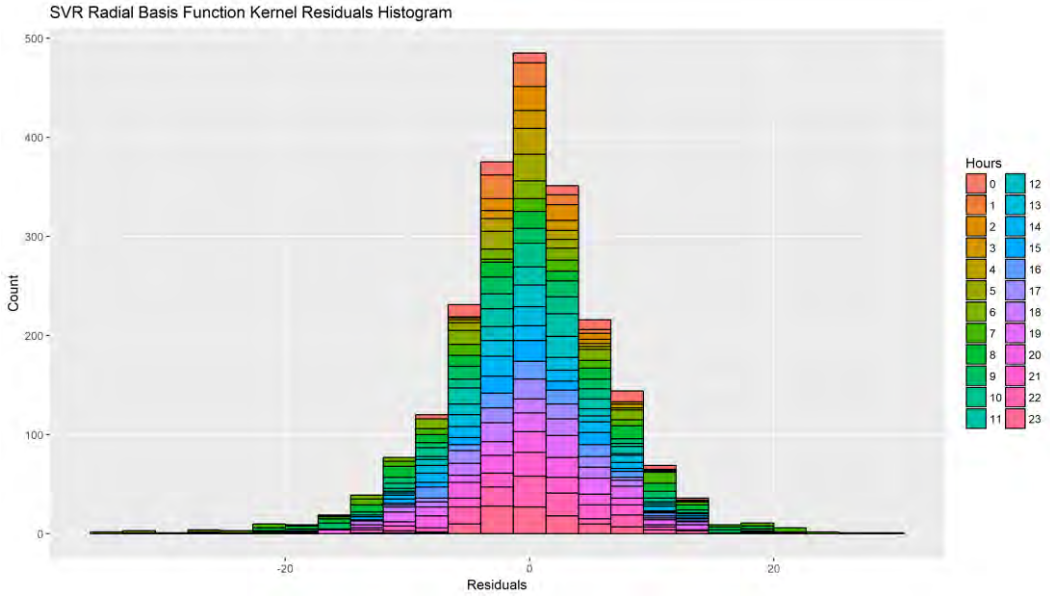
Σχήμα 74. Φθινόπωρο & Άνοιξη, ιστόγραμμα δέντρων απόφασης



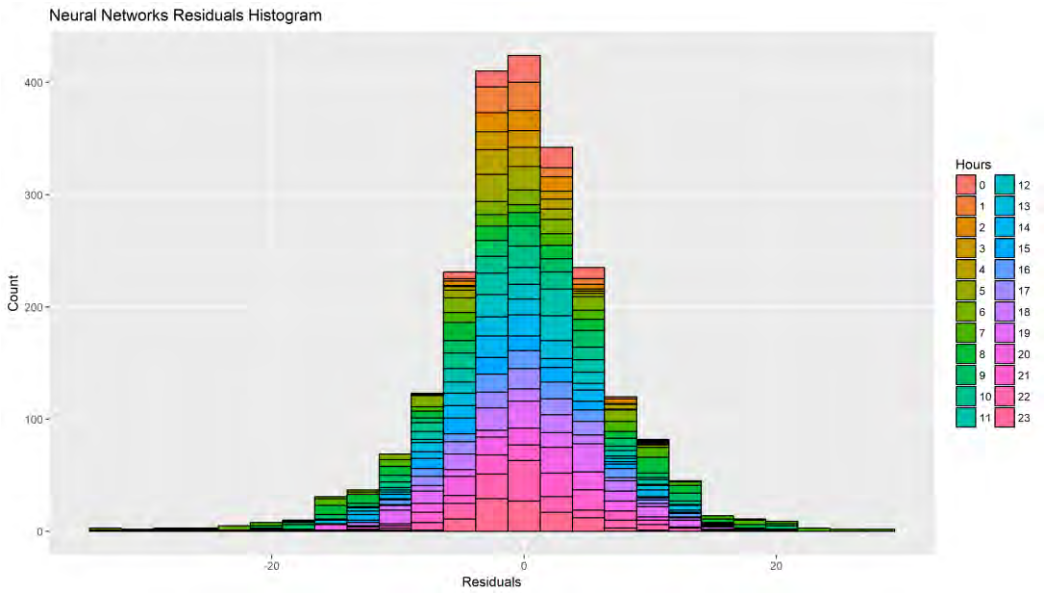
Σχήμα 75. Φθινόπωρο & Άνοιξη, ιστόγραμμα SVR με γραμμικό πυρήνα



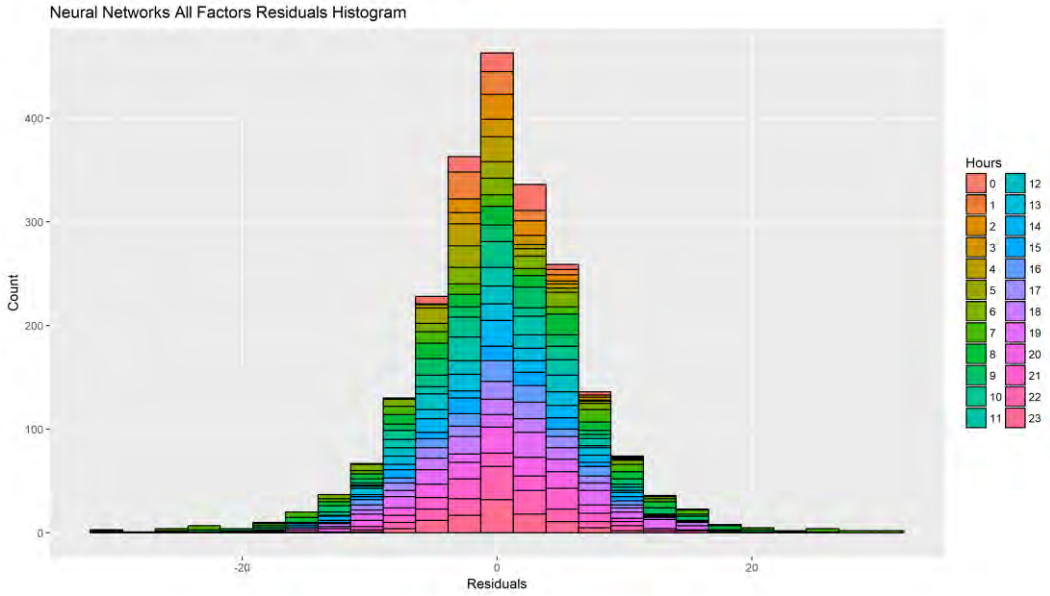
Σχήμα 76. Φθινόπωρο & Άνοιξη, ιστόγραμμα SVR με πολυωνυμικό πυρήνα



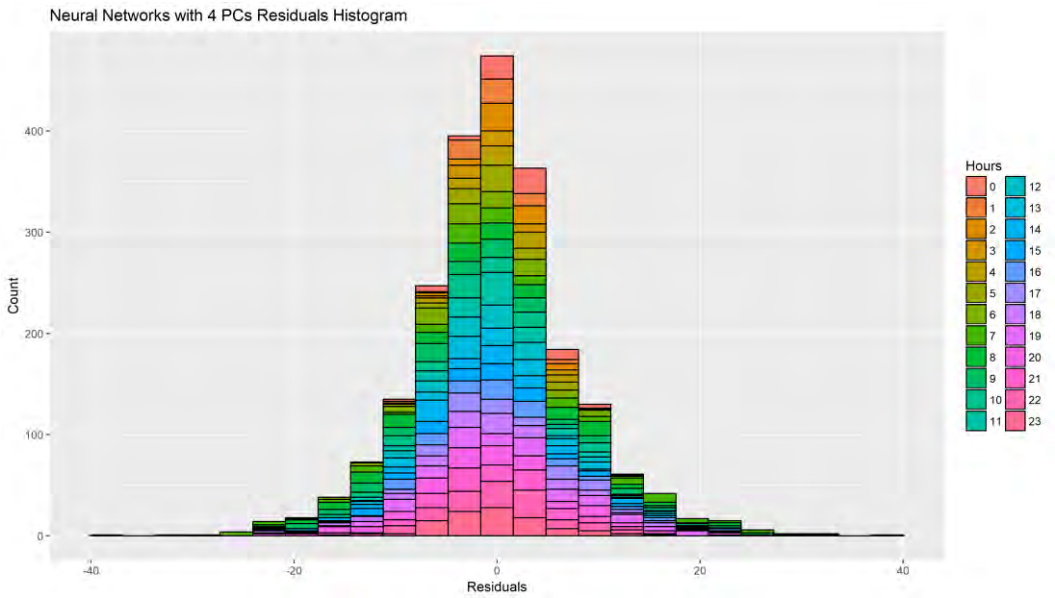
Σχήμα 77. Φθινόπωρο & Άνοιξη, ιστόγραμμα SVR με RBF πυρήνα



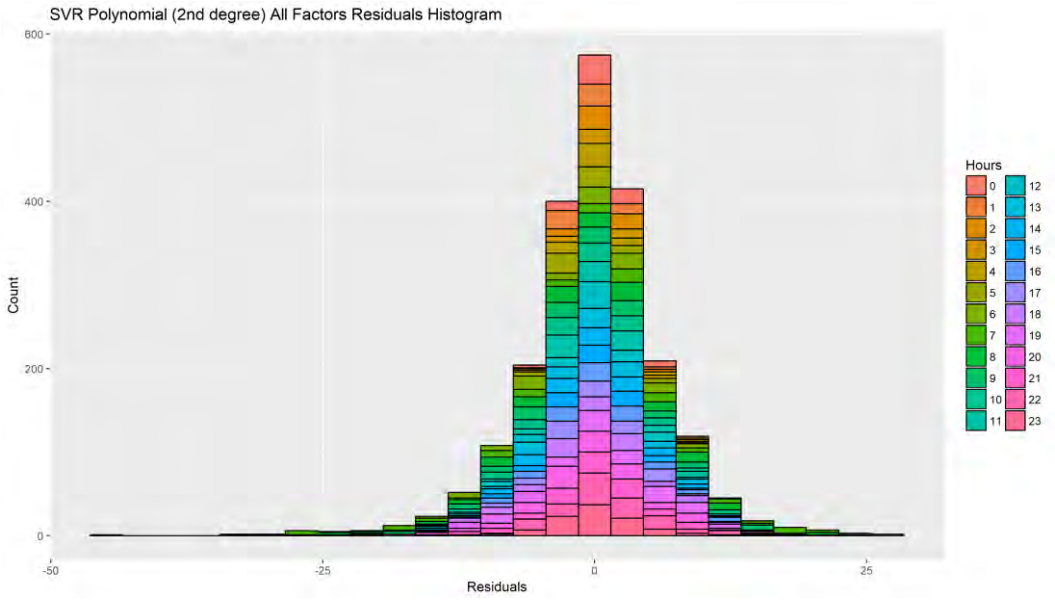
Σχήμα 78. Φθινόπωρο & Άνοιξη, ιστόγραμμα νευρωνικού δικτύου



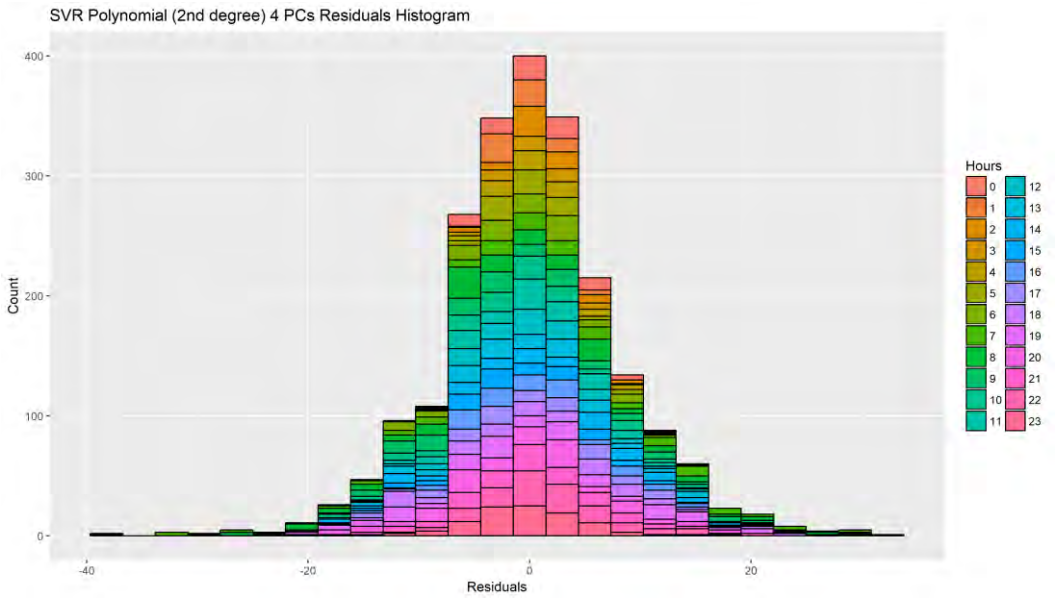
Σχήμα 79. Φθινόπωρο & Άνοιξη, ιστόγραμμα νευρωνικού δικτύου με όλους τους καιρικούς παράγοντες



Σχήμα 80. Φθινόπωρο & Άνοιξη, ιστόγραμμα νευρωνικού δικτύου με χρήση PCA

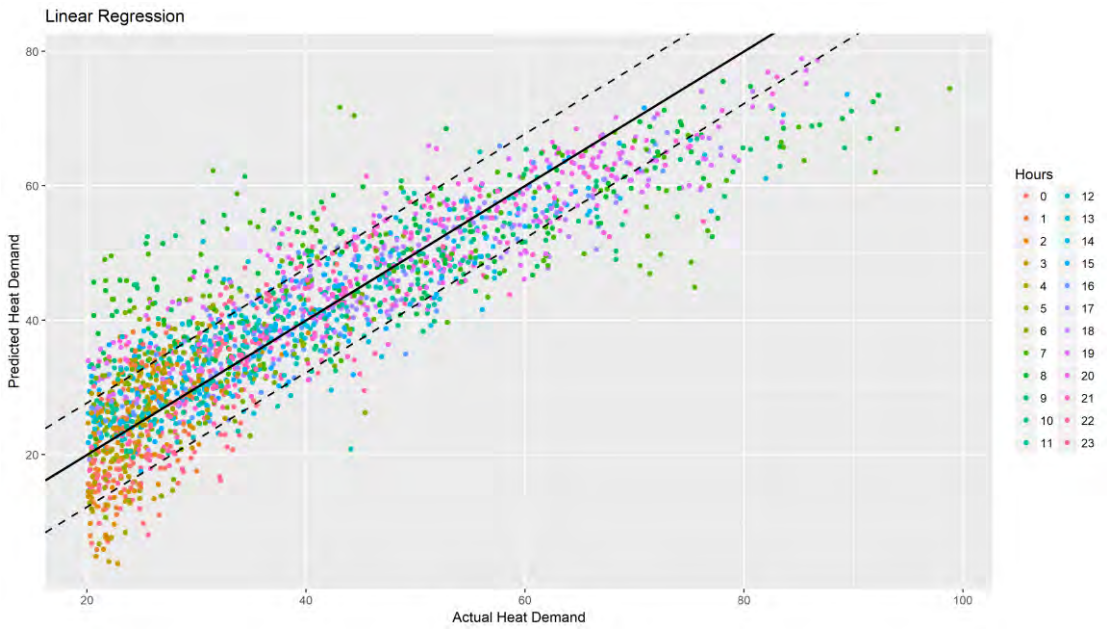


Σχήμα 81. Φθινόπωρο & Άνοιξη, ιστόγραμμα SVR με πολυωνυμικό πυρήνα και όλους τους καιρικούς παράγοντες

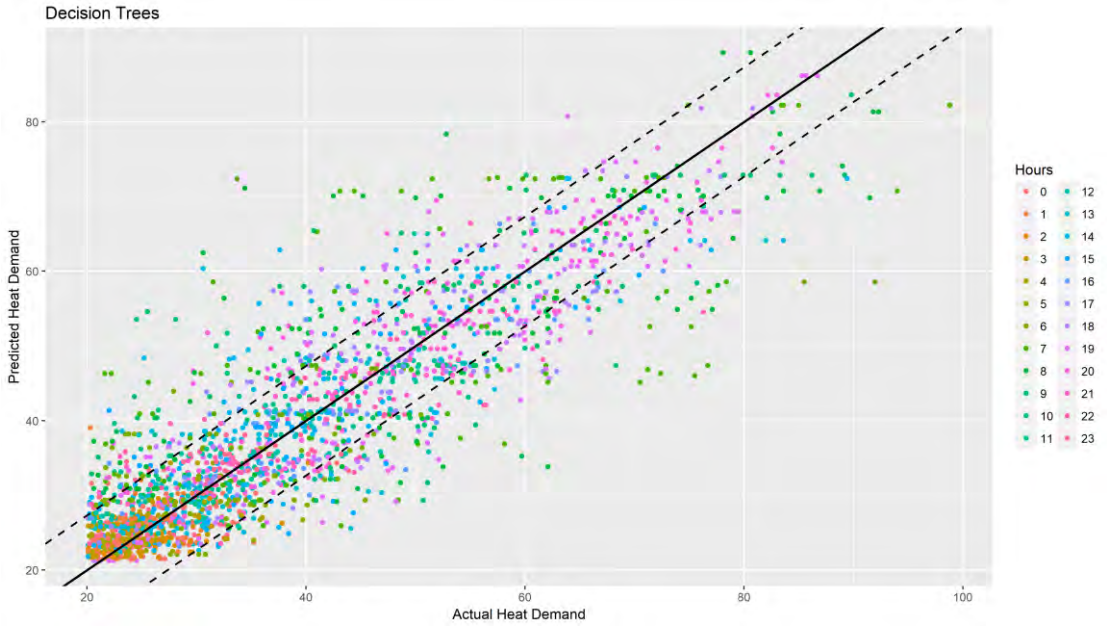


Σχήμα 82. Φθινόπωρο & Άνοιξη, ιστόγραμμα SVR με πολυωνυμικό πυρήνα και χρήση PCA

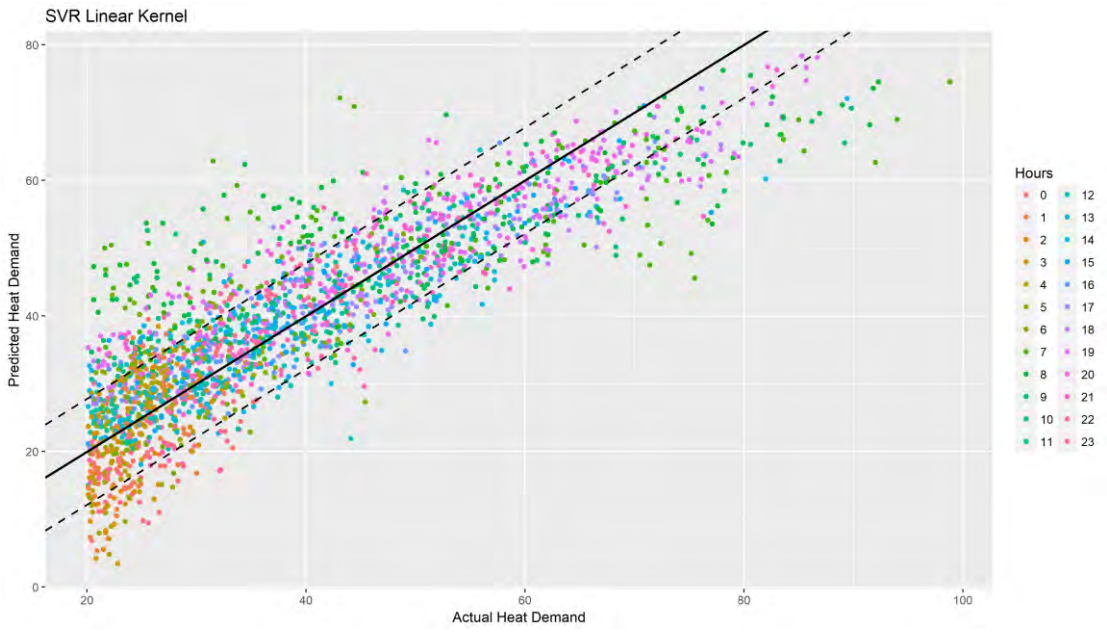
Για τις γραφικές παραστάσεις των μοντέλων ανάμεσα στις τιμές πρόβλεψης και τις πραγματικές τα συμπεράσματα είναι παρόμοια με αυτά του προηγούμενου συνόλου. Μία διαφορά ανάμεσα στα δύο σύνολα είναι οι χαμηλότερες τιμές ζήτησης και οι υψηλότερες τιμές εξωτερικής θερμοκρασίας στο σύνολο δεδομένων του φθινοπώρου και της άνοιξης, το οποίο είναι λογικό να συμβαίνει. Πέρα από αυτό, όπως πριν έτσι και σε αυτές τις γραφικές παραστάσεις, Σχήματα 83-92, διαπιστώνεται πως υπάρχει και δεύτερο διάστημα ωρών αιχμής, όχι τόσο έντονο με αυτό του χειμώνα, αλλά συγκριτικά με τις υπόλοιπες ώρες της ημέρας. Πάλι τα μεγάλα σφάλματα αντιστοιχούν στις ώρες αιχμής και συναντώνται σε όλο το πεδίο τιμών ζήτησης του δικτύου. Μια γενική παρατήρηση είναι πως τα μοντέλα δεν μπορούν να προβλέψουν με μικρή απόκλιση την τιμή ζήτησης στις ώρες αιχμής όμως τις περισσότερες φορές η πρόβλεψη είναι μεγαλύτερη το οποίο βοηθάει στην κανονική λειτουργία του δικτύου.



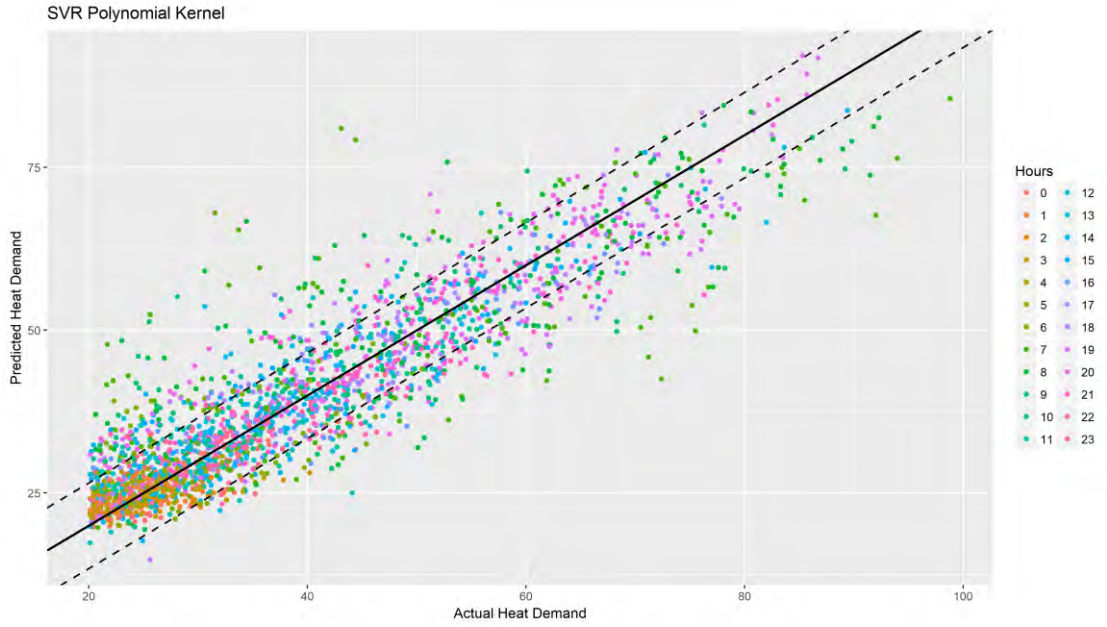
Σχήμα 83. Φθινόπωρο & Άνοιξη, γραμμική παλινδρόμηση



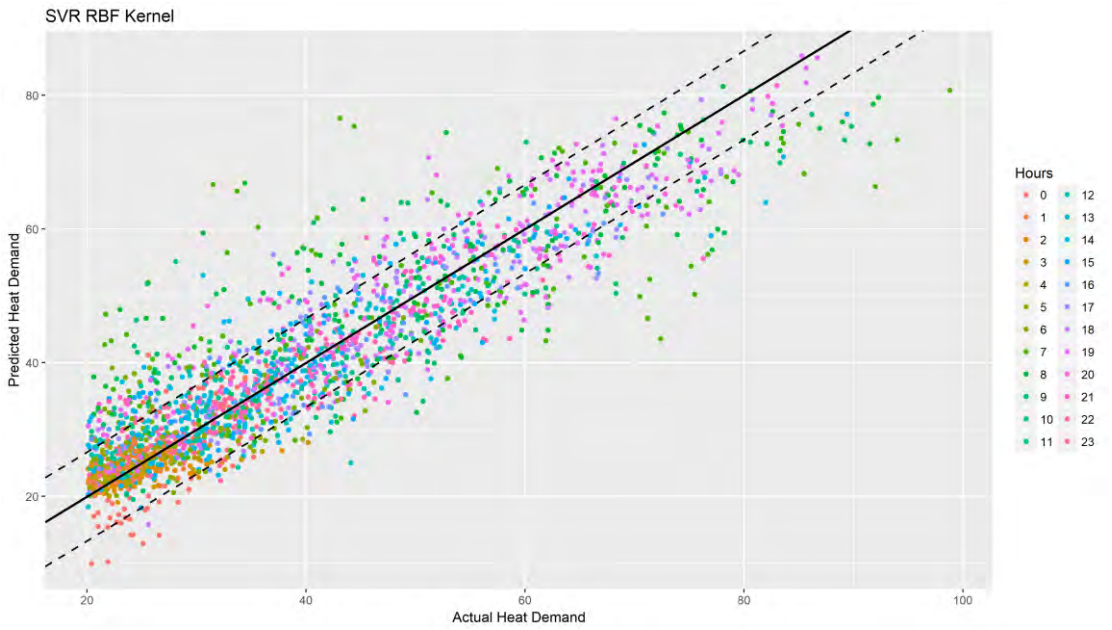
Σχήμα 84. Φθινόπωρο & Άνοιξη, δέντρα απόφασης



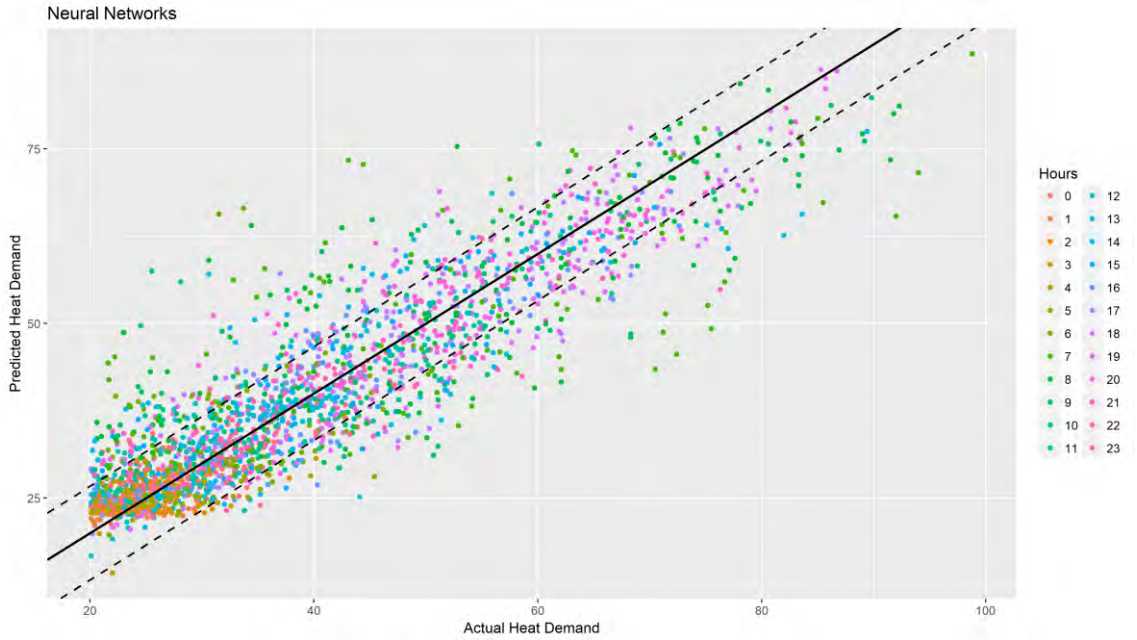
Σχήμα 85. Φθινόπωρο & Άνοιξη, SVR με γραμμικό πυρήνα



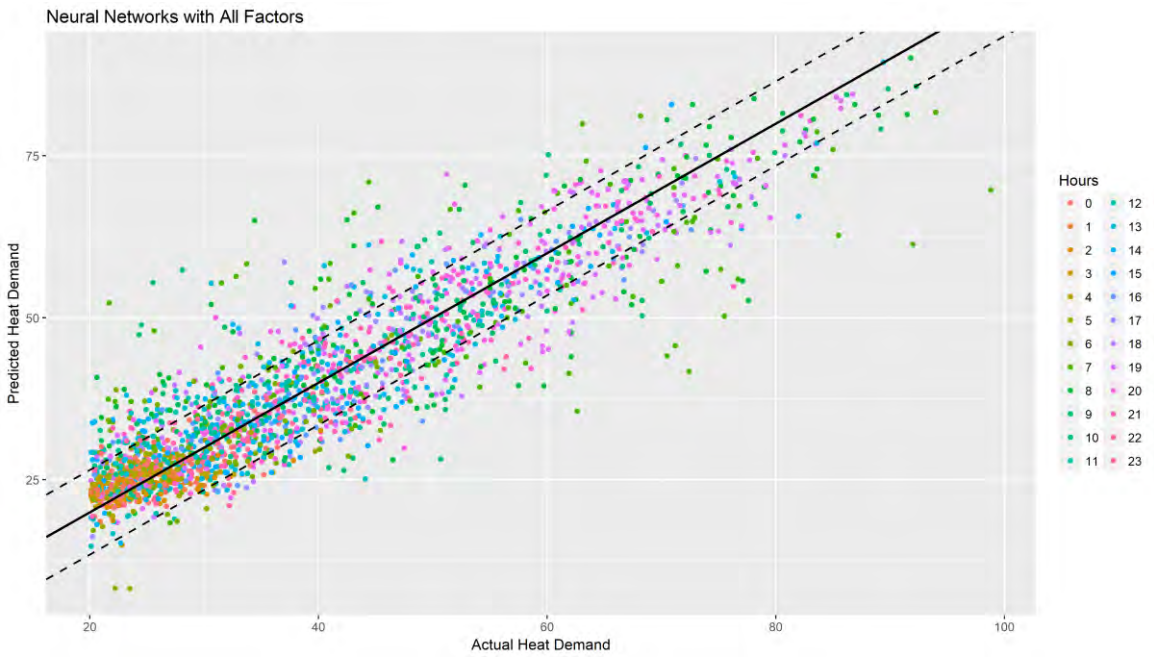
Σχήμα 86. Φθινόπωρο & Άνοιξη, SVR πολυωνυμικό πυρήνα



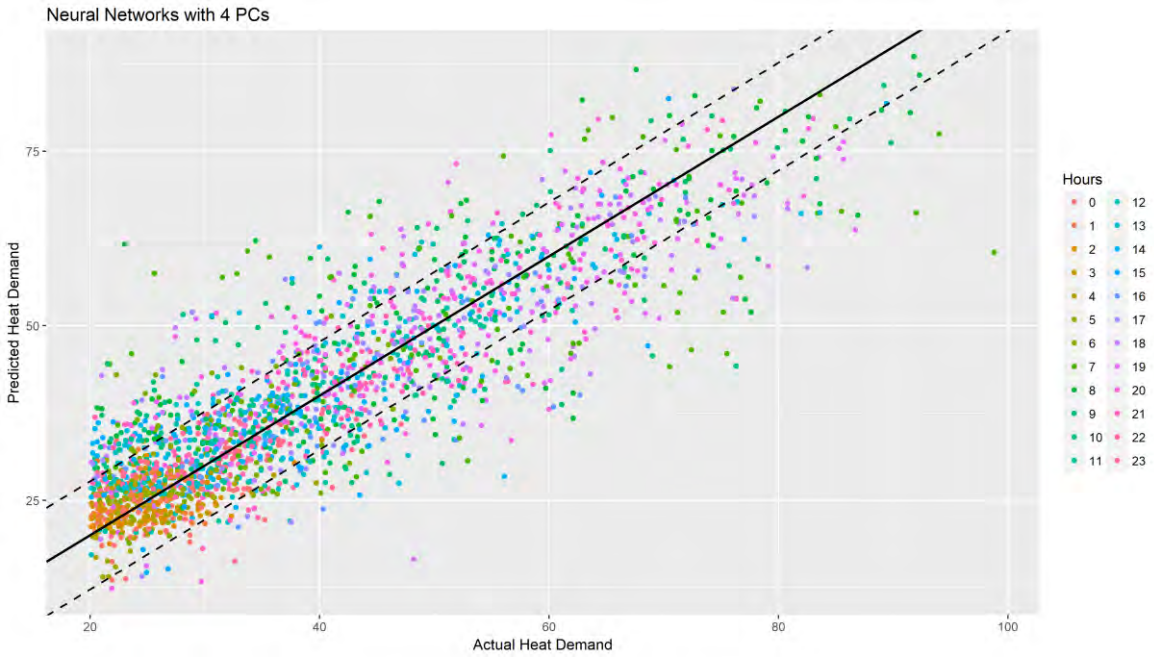
Σχήμα 87. Φθινόπωρο & Άνοιξη, SVR με RBF πυρήνα



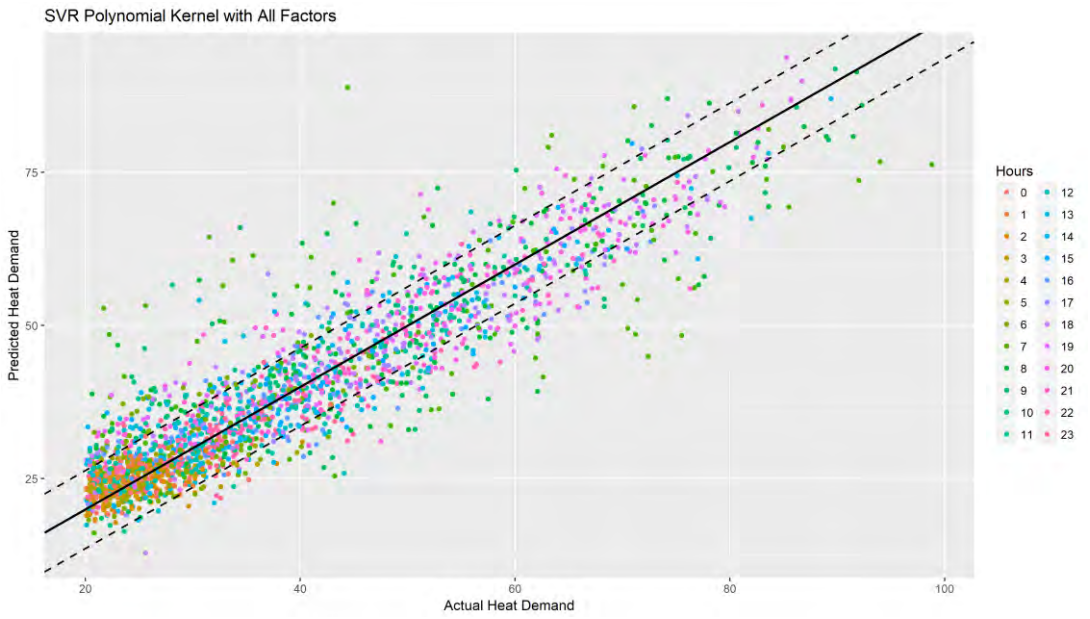
Σχήμα 88. Φθινόπωρο & Άνοιξη, Νευρωνικό δίκτυο



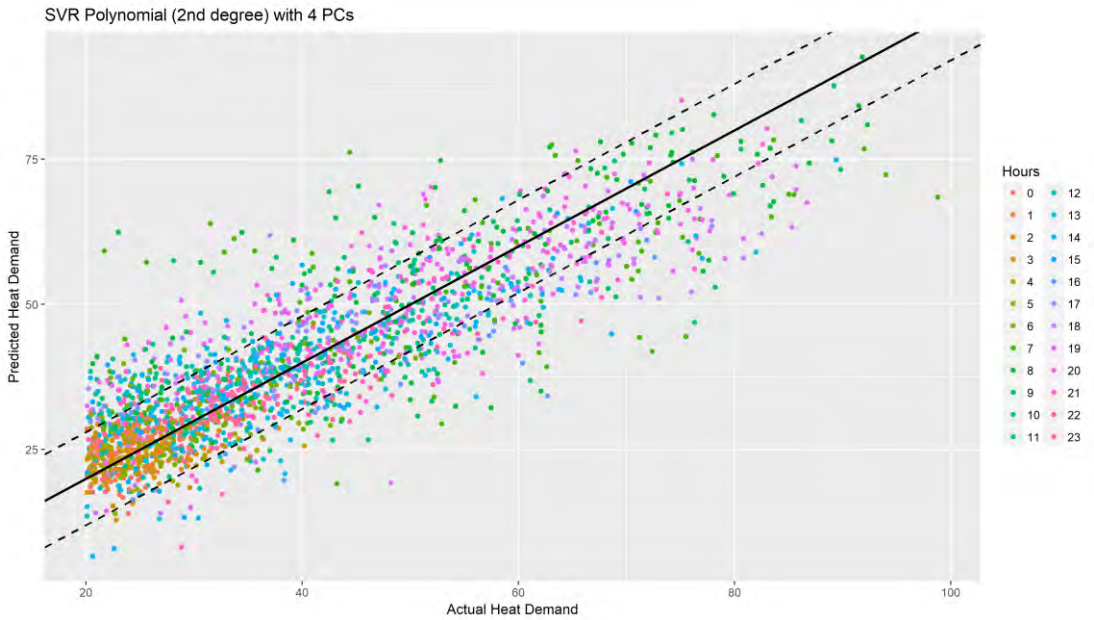
Σχήμα 89. Φθινόπωρο & Άνοιξη, Νευρωνικό δίκτυο με όλους τους καιρικούς παράγοντες



Σχήμα 90. Φθινόπωρο & Άνοιξη, Νευρωνικό δίκτυο με χρήση PCA



Σχήμα 91. Φθινόπωρο & Άνοιξη, SVR με πολυωνυμικό πυρήνα και όλους τους καιρικούς παράγοντες

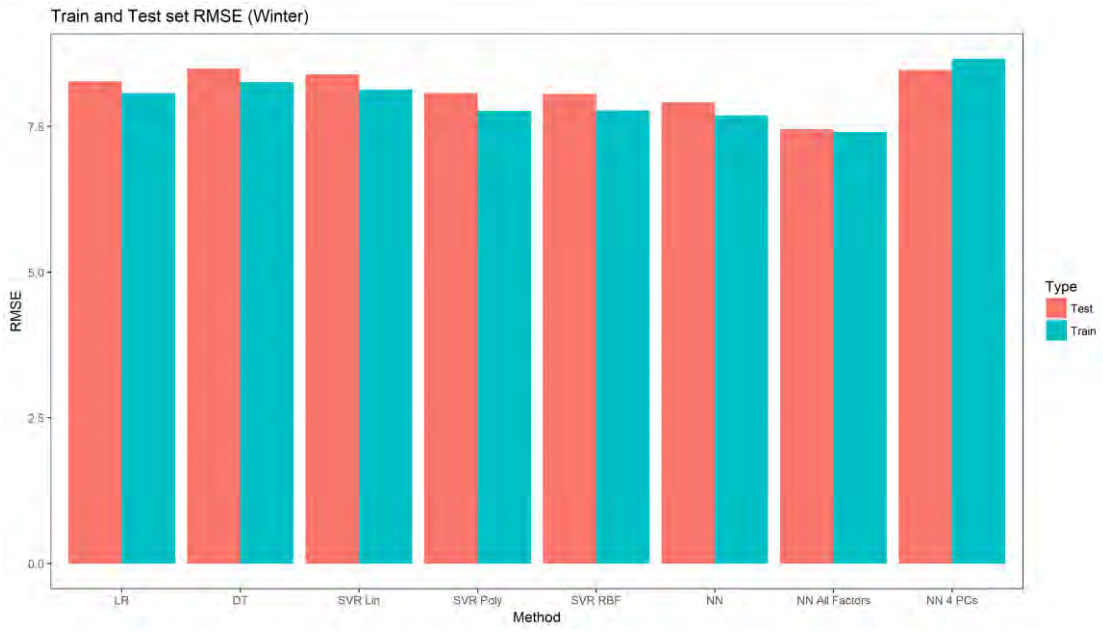


Σχήμα 92. Φθινόπωρο & Άνοιξη, SVR με πολυωνυμικό πυρήνα και χρήση PCA

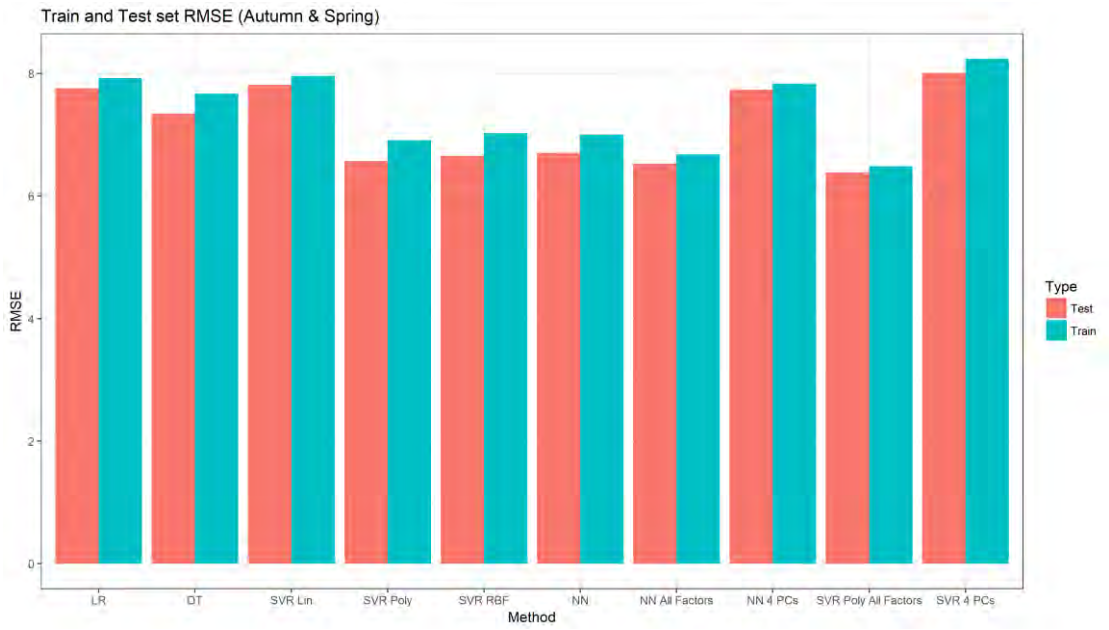
Τα μοντέλα των δύο συνόλων δεδομένων που κατασκευάστηκαν χρησιμοποιώντας κύριες συνιστώσες δεν κατάφεραν να συγκεντρώσουν την απαραίτητη πληροφορία σε μικρότερη διάσταση έτσι ώστε να έχουν παρόμοια απόδοση με τα υπόλοιπα μοντέλα. Αυτό φαίνεται και από τους πίνακες 3, 5, 6 παρατηρώντας τα RMSE

των μοντέλων ή από τα Σχήματα 93, 94 όπου υπάρχει γραφική σύγκριση των RMSE των μοντέλων για κάθε σύνολο δεδομένων. Επίσης για τα ίδια μοντέλα το R^2 είναι μικρότερο οπότε δεν κατάφεραν να εξηγήσουν το ίδιο ποσοστό πληροφορίας στα δεδομένα εκπαίδευσης συγκριτικά με τα άλλα. Παρόλα αυτά η διαφορά των αποδόσεων τους σε σχέση με τα υπόλοιπα δεν είναι αρκετά μεγάλη.

Για τα υπόλοιπα μοντέλα του χειμώνα το R^2 βρίσκεται μεταξύ 84% και 87%, πίνακες 2-6. Αυτό σημαίνει ότι μπορούν να εξηγήσουν ένα αρκετά μεγάλο ποσοστό των δεδομένων εκπαίδευσης με το ανεξήγητο ποσοστό να οφείλεται στις προβλέψεις για τις ώρες αιχμής. Αντίθετα στα μοντέλα του φθινοπώρου και της άνοιξης το R^2 κυμαίνεται από 75% μέχρι 83%. Θα περίμενε κάποιος το ποσοστό να ήταν μεγαλύτερο επειδή στις ώρες αιχμής οι διαφορές δεν είναι τόσο έντονες. Όμως όπως είδαμε στα Σχήματα 93, 94, στο σύνολο δεδομένων του χειμώνα η ομαδοποίηση λόγω της ώρας είναι πιο εμφανής σε αντίθεση με το άλλο σύνολο δεδομένων. Έχοντας εισάγει ως ανεξάρτητη μεταβλητή την ώρα η πληροφορία αυτή βοηθάει τα μοντέλα του χειμώνα να εξηγήσουν καλύτερα τα δεδομένα εκπαίδευσης.



Σχήμα 93. Χειμώνας, RMSE για σύνολο εκπαίδευσης και ελέγχου



Σχήμα 94. Φθινόπωρο & Άνοιξη, RMSE για σύνολο εκπαίδευσης και ελέγχου

Καταλήγουμε λοιπόν στο γενικό συμπέρασμα πως τα μοντέλα δεν μπορούν να προσεγγίσουν με ακρίβεια την τιμή ζήτησης του δικτύου στις ώρες αιχμής αλλά έχουν αρκετά καλή απόδοση στην διάρκεια της μέρας. Οι πρωινές ώρες αιχμής συμπίπτουν με τον τρόπο λειτουργίας του night setback control, δηλαδή η ώρα αιχμής είναι ίδια με την στιγμή αλλαγής των ορίων στο night setback control. Όμως δεν είναι αυτό υπεύθυνο για την απότομη αλλαγή στην ζήτηση. Ίσως χωρίς το πρόγραμμα να μην ήταν τόσο απότομη η αλλαγή αλλά πάλι θα υπήρχε και θα ήταν έντονη.

5.3. Ώρες αιχμής

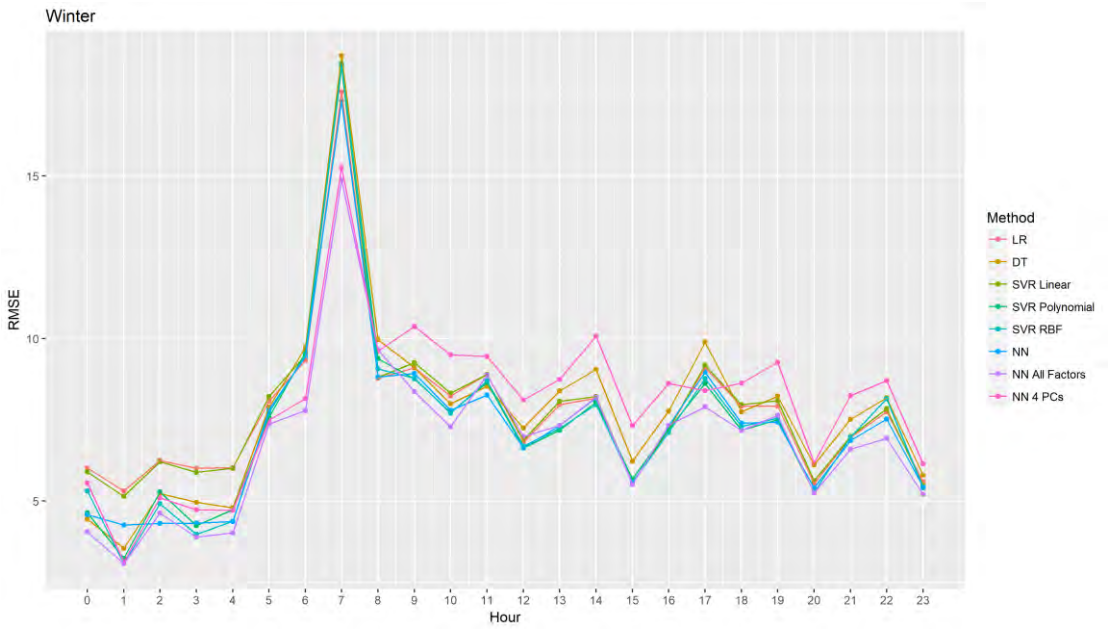
Από προηγούμενες παρατηρήσεις καταλήξαμε ότι υπάρχουν δύο διαστήματα ωρών, όπου η ζήτηση του δικτύου είναι πολύ μεγαλύτερη με εκείνη της υπόλοιπης ημέρας. Το πρώτο και πιο σημαντικό είναι τις ώρες 06:00 έως 08:00 το πρωί, με την κορύφωση να συμβαίνει στις 07:00 ενώ το δεύτερο είναι από τις 18:00 μέχρι τις 20:00.Οι λόγοι για τους οποίους συμβαίνει αυτό εξαρτώνται κυρίως από τον ανθρώπινο παράγοντα. Στην πρώτη περίπτωση εκείνες τις ώρες ξυπνάνε οι περισσότεροι άνθρωποι οπότε θα αυξήσουν την ζήτηση έτσι ώστε να θερμάνουν τις οικίες τους. Η έναρξη λειτουργίας των δημόσιων υπηρεσιών συμπίπτει με τις πρωινές ώρες αιχμής. Ομοίως για το δεύτερο διάστημα είναι

οι ώρες που αρκετοί επιστρέφουν στις οικίες τους αλλά και η χρονική στιγμή όπου η εξωτερική θερμοκρασία αρχίζει να μειώνεται. Οπότε η αδυναμία των μοντέλων πρόβλεψης να υπολογίσουν με ακρίβεια την ζήτηση του δικτύου στις ώρες αιχμής οφείλεται στην απουσία πληροφορίας σχετικά με τον ανθρώπινο παράγοντα.

5.4. Ωριαία πρόβλεψη

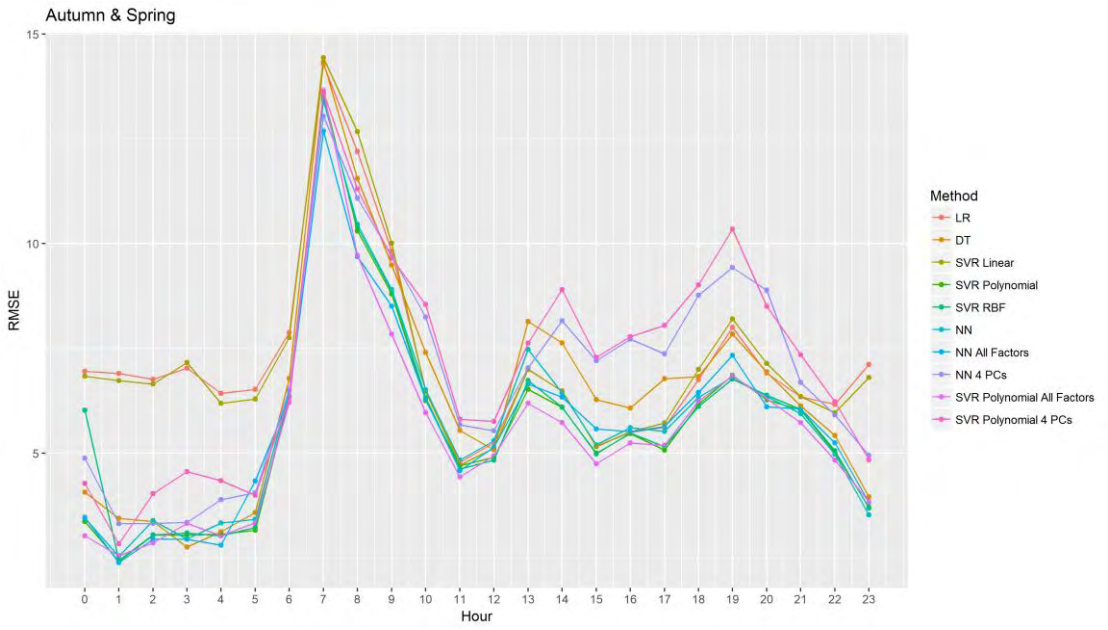
Θέλοντας να αξιολογήσουμε τα μοντέλα για κάθε ώρα της ημέρας ομαδοποιήσαμε τα σφάλματα με κριτήριο την ώρα της αντίστοιχης παρατήρησης. Στην συνέχεια υπολογίσαμε το RMSE για κάθε ώρα και για κάθε μοντέλο και στα δύο σύνολα ελέγχου (test set) των δεδομένων. Τα αποτελέσματα για το σύνολο δεδομένων του χειμώνα φαίνονται στο Σχήμα 95 και για τα δεδομένα του φθινοπώρου και της άνοιξης στο Σχήμα 96. Στις γραφικές παραστάσεις απεικονίζεται το RMSE οπότε μας ενδιαφέρει να παραμένει σε χαμηλές τιμές.

Στην περίπτωση του χειμώνα, ξεκινώντας από τις πρωινές ώρες αιχμής παρατηρούμε ότι κανένα μοντέλο δεν πετυχαίνει RMSE μικρότερο από 15. Μόνο τα δύο νευρωνικά δίκτυα, το ένα με όλους τους καιρικούς παράγοντες και το άλλο με τις κύριες συνιστώσες, βρίσκονται κοντά στο 15, όμως για τις υπόλοιπες ώρες το νευρωνικό δίκτυο με τις κύριες συνιστώσες έχει μία από τις χειρότερες αποδόσεις. Το νευρωνικό δίκτυο με τα δεδομένα της Δ.Ε.ΤΗ.Π. και το άλλο με όλα τα δεδομένα είναι αυτά που έχουν το μικρότερο RMSE για κάθε ώρα της ημέρας. Για τις υπόλοιπες μεθόδους η γραφική του RMSE έχει σχεδόν το ίδιο μοτίβο κατά την διάρκεια της ημέρας.



Σχήμα 95. Χειμώνας, ωριαία διακύμανση RMSE

Στην αντίστοιχη γραφική παράσταση για το σύνολο δεδομένων του φθινοπώρου και της άνοιξης παρατηρούμε και τις απογευματινές ώρες αιχμής, κάτι που δεν φαινόταν στο προηγούμενο σύνολο. Το μοντέλο των μηχανών διανυσματικής υποστήριξης με πολυωνυμικό πυρήνα για τα δεδομένα της Δ.Ε.ΤΗ.Π έχει το μικρότερο RMSE ενώ μετά από αυτό βρίσκεται το αντίστοιχο μοντέλο με όλους τους καιρικούς παράγοντες.



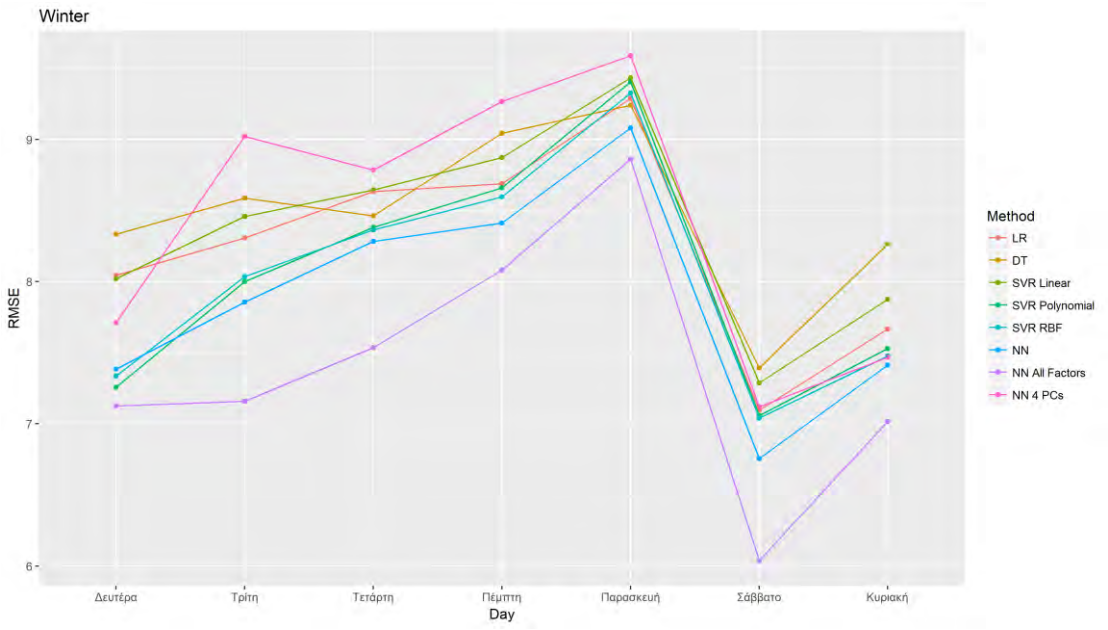
Σχήμα 96. Φθινόπωρο & Άνοιξη, ωριαία διακύμανση RMSE

Από αυτές τις δύο γραφικές παραστάσεις βλέπουμε την μεγάλη διαφορά στο RMSE για τις ώρες αιχμής. Οπότε η οποιαδήποτε βελτίωση για αυτές τις ώρες θα είχε ως αποτέλεσμα μια καλύτερη συνολική απόδοση.

5.5. Εβδομαδιαία πρόβλεψη

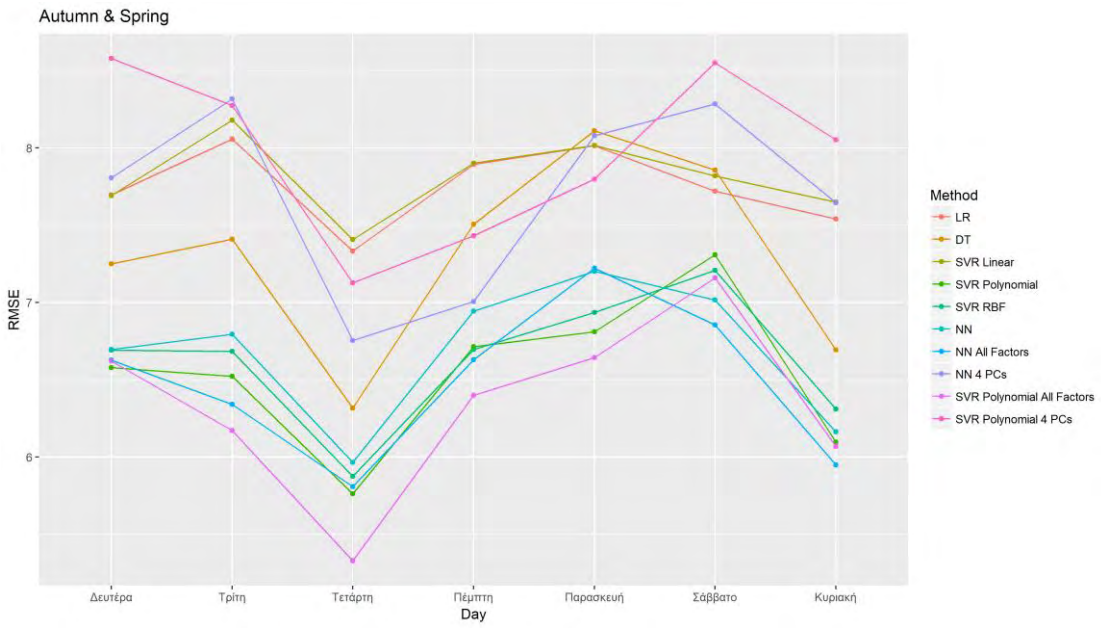
Οι συγκεκριμένες γραφικές παραστάσεις, Σχήματα 97, 98, είναι παρόμοιες με τις δύο προηγούμενες μόνο που εδώ τα σφάλματα ομαδοποιούνται ανάλογα με την ημέρα της εβδομάδας. Σκοπός είναι να εξετάσουμε πως συμπεριφέρονται τα μοντέλα για κάθε διαφορετική ημέρα χωρίς όμως να τους έχει δοθεί η συγκεκριμένη πληροφορία κατά της διαδικασία κατασκευής τους. Τα αποτελέσματα είναι αρκετά ενδιαφέροντα επειδή αποκαλύπτουν χαρακτηριστικά που οφείλονται στον ανθρώπινο παράγοντα. Μέχρι τώρα καταλήξαμε στο συμπέρασμα πως τα μοντέλα δεν μπορούν να προβλέψουν απότομες αυξητικές αλλαγές στην ζήτηση του δικτύου και έτσι έχουν ένα σχετικά σταθερό RMSE στην διάρκεια της ημέρας εκτός από τις ώρες αιχμής. Επίσης όπως είδαμε από τα ιστογράμματα και τις γραφικές παραστάσεις των τιμών πρόβλεψης ως προς τις πραγματικές τιμές, για τις ώρες που η ζήτηση έχει χαμηλές τιμές τα σφάλματα είναι σχεδόν μηδενικά. Άρα τα μοντέλα μπορούν να αναγνωρίσουν και να προβλέψουν με βεβαιότητα την χαμηλή τιμή της ζήτησης.

Για την γραφική παράσταση του χειμώνα, Σχήμα 97, παρατηρούμε αύξηση του RMSE κατά την διάρκεια της εβδομάδας και μία απότομη μείωση το Σάββατο. Όπως αναφέραμε προηγουμένως τα μοντέλα αναγνωρίζουν της χαμηλές τιμές ζήτησης. Άρα το Σάββατο η ζήτηση είναι μικρότερη σε σχέση με τις καθημερινές λόγω του χαμηλού RMSE. Αυτό είναι λογικό να συμβαίνει καθώς το Σάββατο υπάρχει μία ομάδα ανθρώπων που σε αντίθεση με τις καθημερινές ξυπνάει πιο αργά, αυτή η ομάδα είναι οι μαθητές. Έτσι στην πρωινή ώρα αιχμής το πλήθος των ατόμων είναι αρκετά λιγότερο. Μετά τις 09:00 η εξωτερική θερμοκρασία ανεβαίνει οπότε η ζήτηση δεν είναι μεγάλη και δεν δημιουργείται δεύτερη ώρα αιχμής. Το ίδιο συμβαίνει και την Κυριακή μόνο που σε αυτή την περίπτωση το RMSE είναι μεγαλύτερο από εκείνο του Σαββάτου αλλά όχι μεγαλύτερο των καθημερινών.



Σχήμα 97. Χειμώνας, εβδομαδιαία διακύμανση RMSE

Στο δεύτερο σύνολο δεδομένων λόγω πιο ήπιων καιρικών συνθηκών οι διαφορές δεν είναι τόσο μεγάλες μεταξύ των καθημερινών ημερών και του σαββατοκύριακου. Όμως την Τετάρτη παρατηρείται μία αισθητή μείωση στο RMSE. Για κάποιο λόγο ζήτηση του δικτύου πέφτει την Τετάρτη και έτσι τα μοντέλα έχουν καλύτερη απόδοση. Δεν μπορούμε να είμαστε απόλυτα σίγουροι αλλά ίσως η μείωση αυτή οφείλεται στο γεγονός πως κάθε Τετάρτη στην Πτολεμαΐδα πραγματοποιείται η λαϊκή αγορά. Αυτό αναγκάζει αρκετούς πολίτες να βρίσκονται σε εξωτερικό χώρο οπότε η ζήτηση του δικτύου πέφτει για κάποιες ώρες. Βέβαια δεν συμβαίνει το ίδιο για το σύνολο δεδομένων των χειμερινών μηνών καθώς οι καιρικές συνθήκες δεν το επιτρέπουν. Οι πολίτες χαμηλώνουν την ζήτηση τους μήνες του φθινοπώρου και της άνοιξης επειδή η διαφορά εσωτερικής και εξωτερικής θερμοκρασίας δεν είναι μεγάλη και έτσι θα μπορέσουν να θερμάνουν γρήγορα την οικία τους χωρίς μεγάλο οικονομικό κόστος. Το ίδιο δεν μπορεί να συμβεί για τους μήνες του χειμώνα οπότε και δεν παρατηρείται η αντίστοιχη μείωση την Τετάρτη.



Σχήμα 98. Φθινόπωρο & Άνοιξη, εβδομαδιαία διακύμανση RMSE

5.6. Τελικά συμπεράσματα

Σκοπός της ανάλυσης ήταν να εξετάσει την δυνατότητα ορισμένων τεχνικών μηχανικής μάθησης, να προβλέψουν την ενεργειακή ζήτηση ενός δικτύου τηλεθέρμανσης. Δεν υπάρχει λόγος να αναδειχθεί ένα μοντέλο ως το καλύτερο. Άλλωστε οι αποδόσεις τους ως προς το RMSE και το R^2, που φαίνονται στους πίνακες 2-6, βρίσκονται πολύ κοντά και σε όλα τα μοντέλα υπάρχουν οι ίδιες αδυναμίες. Η προσθήκη περισσότερων καιρικών παραγόντων βελτίωσε τις αποδόσεις αλλά η εισαγωγή τους εξαρτάται από τις ανάγκες κάθε συστήματος. Από την ανάλυση αναδείχθηκαν τα προβλήματα των μοντέλων όπως και η έλλειψη πληροφορίας για τον ανθρώπινο παράγοντα. Ένας από τους πιο σημαντικούς στόχους ήταν η πρόβλεψη στις ώρες αιχμής. Σε αυτές τις ώρες τα μοντέλα είχαν την χειρότερη απόδοση. Όμως τώρα γνωρίζουμε πως αρκεί μόνο η βελτίωση σε αυτές τις ώρες καθώς τα μοντέλα ανταποκρίνονται αρκετά καλά στις υπόλοιπες ώρες τις ημέρας. Σίγουρα αυτό που κατάφερε η ανάλυση ήταν να δείξει τα οφέλη που προκύπτουν από την αξιοποίηση των δεδομένων μέσα από τέτοιες τεχνικές.

5.7. Μελλοντικές προτάσεις

Ο ανθρώπινος παράγοντας επηρεάζει αρκετά την ζήτηση του δικτύου. Με την εισαγωγή μετρήσεων από διαφορετικούς τύπους κτηρίων όπως οικίες, καταστήματα, ιατρεία, καφετέριες και δημόσιες υπηρεσίες, τα νέα μοντέλα θα γνωρίζουν σε τι κτήριο αναφέρεται κάθε εγγραφή και έτσι θα έχουν πιο στοχευμένες προβλέψεις. Για τα μοντέλα παρατηρήσαμε πως ορισμένες τεχνικές έχουν καλύτερη απόδοση σε διαφορετικές ώρες και μέρες τις ημέρας. Με την κατασκευή ενός συνόλου μάθησης (ensemble learning) η τελική τιμή της εξαρτημένης μεταβλητής θα προέρχεται από ένα συνδυασμό πολλών τεχνικών. Για παράδειγμα για κάθε ώρα θα επιλέγεται η τεχνική με την καλύτερη πρόβλεψη. Τέλος ο διαχωρισμός της ημέρας σε λιγότερα διαστήματα με μεγαλύτερο εύρος ωρών ίσως βοηθήσει στις προβλέψεις των μοντέλων επειδή δεν θα υπάρχουν πολλές διαφορετικές τιμές για κάθε στιγμή της ημέρας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Απόφαση αριθ. 406/2009/ΕΚ του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου, της 23ης Απριλίου 2009, περί των προσπαθειών των κρατών μελών να μειώσουν τις οικείες εκπομπές αερίων θερμοκηπίου, ώστε να τηρηθούν οι δεσμεύσεις της Κοινότητας για μείωση των εκπομπών αυτών μέχρι το 2020 OJ L 140, 5.6.2009, p. 136–148 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

[2] EU Emissions Trading System (EU ETS) - Climate Action - European Commission, Climate Action - European Commission, 2018. Available at: https://ec.europa.eu/clima/policies/ets_en. Accessed June 28, 2018.

[3] EU Parliament, Directive 2010/31/eu of the European parliament and of the council of 19 may 2010 on the energy performance of buildings, 2010.

[4] Webb M, Smart 2020: Enabling The Low Carbon Economy in The Information Age, A report by the climate group on behalf of the Global eSustainability Initiative (GeSI). Global eSustainability Initiative (GeSI), Tech. Rep, 2008.

[5] Logstor, White Paper On District Heating And District Cooling Solutions In An Environmental Perspective., Tech. Rep, 2007.

[6] Lund H, Werner S, Wiltshire R et al, 4th Generation District Heating (4GDH), Energy, 2014; 68: 1-11.

[7] Stănişteanu C, Smart Thermal Grids – A Review, The Scientific Bulletin of Electrical Engineering Faculty, 2017.

[8] Δ.Ε.ΤΗ.Π. Tpt.gr. 2018. Available at: http://www.tpt.gr/. Accessed April 28, 2018.

[9] Current weather and forecast - OpenWeatherMap. Openweathermap.org. 2018. Available at: https://openweathermap.org/. Accessed April 30, 2018.

[10] Ekici B, Aksoy U, Prediction of building energy consumption by using artificial neural networks, Advances in Engineering Software, 2009; 40(5): 356-362.

[11] Grzenda M, Macukow B., Demand prediction with multistage neural processing, Advances in NaturaL Computation and Data Mining, 2006, 131-141.

[12] Sakawa M, Matsui T, Heat load prediction in district heating and cooling systems through recurrent neural networks, International Journal of Operational Research, 2015; 23(3): 284.

[13] Xie J, Li H, Ma Z et al., Analysis of Key Factors in Heat Demand Prediction with Neural Networks, Energy Procedia, 2017; 105: 2965-2970.

[14] Park T, Kim U, Kim L, Jo B, Yeo Y, Heat consumption forecasting using partial least squares, artificial neural network and support vector regression techniques in district heating systems, Korean Journal of Chemical Engineering. 2010; 27(4): 1063-1071.

[15] Provatas S, Lavesson N, Johansson C, An online machine learning algorithm for heat load forecasting in district heating systems, Proceedings of the 14th International Symposium on District Heating and Cooling, 2014.

[16] Idowu S, Saguna S, Ahlund C, Schelén O, Forecasting heat load for smart district heating systems: A machine learning approach, Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on. IEEE, 2014.

[17] Rohan N. A Bayesian approach for forecasting heat load in district heating systems, 2015.

[18] Dalipi F, Yildirim Yayilgan S, Gebremedhin A, Data-Driven Machine-Learning Model in District Heating System for Heat Load Prediction: A Comparison Study, Applied Computational Intelligence and Soft Computing, 2016; 2016: 1-11.

[19] Ma Z, Xie J, Li H et al., The Role of Data Analysis in the Development of Intelligent Energy Networks. IEEE Netw, 2017; 31(5): 88-95.

[20] Johansson C, Bergkvist M, Geysen D, Somer O, Lavesson N, Vanhoudt D, Operational Demand Forecasting In District Heating Systems Using Ensembles Of Online Machine Learning Algorithms, Energy Procedia, 2017; 116:208-216.

[21] Ibrahim Dinçer I, Zamfirescu C, District energy systems, In Sustainable Energy Systems and Applications, 2012.

[22] Cross-validation (statistics). Enwikipediaorg. 2018. Available at: https://en.wikipedia.org/wiki/Cross-validation_(statistics). Accessed June 20, 2018.

[23] Model Fit: Underfitting vs. Overfitting - Amazon Machine Learning. Docsawsamazoncom. 2018. Available at:

https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html. Accessed June 20, 2018.

[24] Linear regression. Enwikipediaorg. 2018. Available at: https://en.wikipedia.org/wiki/Linear_regression. Accessed June 20, 2018.

[25] Quinlan J., Induction of decision trees, Mach Learn, 1986; 1(1) :81-106.

[26] Quinian R., C4.5: Programs for Machine Learning, Morgan Kaufmann; 1993.

[27] Breiman L, Friedman J, Olshen R, Stone C., Classification and Regression Trees. Taylor & Francis; 1984.

[28] Entropy (information theory). Enwikipediaorg. 2018. Available at: https://en.wikipedia.org/wiki/Entropy_(information_theory). Accessed June 20, 2018.

[29] Cortes C, Vapnik V., Support-vector networks, Mach Learn. 1995; 20(3): 273-297.

[30] Lin C. J., A guide to support vector machines. Department of Computer Science, National Taiwan University. 2006.

[31] Support Vector Regression. Saedsayadcom. 2018. Available at: http://www.saedsayad.com/support_vector_machine_reg.htm. Accessed June 20, 2018.

[32] Giannouli D., Support Vector Machines in Classification and Regression Problems, 2014.

[33] Bauckhage C. The Dual Problem of L2 SVM Training. 2018.

[34] Haykin S. Neural Networks. Upper Sadle River, N.J.: Prentice Hall; 1999.

[35] Robbins H, Monro S, The Annals of Mathematical Statistics, 1951; 22(3): 400-407.

[36] Goodfellow I, Bengio Y, Courville. Deep Learning, MIT Press; 2016, p 196

[37] Mell P, Grance T, SP 800-145. The NIST Definition of Cloud Computing, National Institute of Standards & Technology, Gaithersburg, MD, 2011.

[38] Lagrange multipliers, examples. Khan Academy. Available at: https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/constrained-optimization/a/lagrange-multipliers-examples. Accessed May 14, 2018.

[39] R: The R Project for Statistical Computing. R-project.org. Available at: https://www.r-project.org/. Accessed May 3, 2018.

[40] RStudio. Available at: https://www.rstudio.com/. Accessed May 3, 2018.

[41] Gadd H, Werner S. Heat load patterns in district heating substations, Appl Energy, 2013; 108: 176-183.