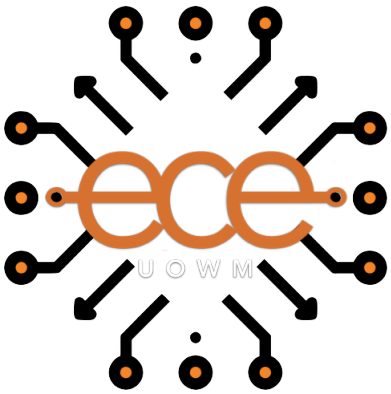


Πανεπιστήμιο Δυτικής Μακεδονίας

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Σύγκριση μεθόδων μηχανικής μάθησης σε δεδομένα που προέρχονται από αισθητήρες

Αθανάσιος Κούρας (ΑΜ: 698)

Επιβλέπων Καθηγητής: Νικόλαος Πλόσκας

Εργαστήριο Ευφυών Συστημάτων & Βελτιστοποίησης

5 Ιουλίου 2021

Περίληψη

Η συλλογή δεδομένων από αισθητήρες και η εξαγωγή χρήσιμων πληροφοριών από ακατέργαστα δεδομένα απαιτούν συγκεκριμένη διαχείριση στον τομέα της επιστήμης των υπολογιστών. Στην παρούσα διπλωματική, πραγματοποιείται η μοντελοποίηση με τη μέθοδο της παλινδρόμησης των δεδομένων θερμοκρασίας που προέρχονται από αισθητήρες τοποθετημένους σε οκτώ διαφορετικούς χώρους μίας κατοικίας στην Λευκόβρυση Κοζάνης, με σκοπό την ακριβή πρόβλεψη της θερμοκρασίας κάθε χώρου. Έπειτα από τη συλλογή των ακατέργαστων δεδομένων από το δίκτυο αισθητήρων που έχει τοποθετηθεί, ακολουθεί η επεξεργασία και η ανάλυσή τους. Χρησιμοποιούνται επτά αλγόριθμοι εποπτευόμενης μηχανικής μάθησης και ένα στατιστικό μοντέλο για την πρόβλεψη θερμοκρασίας σε κάθε χώρο ξεχωριστά. Στη συνέχεια, γίνεται σύγκριση της ακρίβειας κάθε μοντέλου, με στόχο την εξέταση της αποτελεσματικότητάς τους. Για περαιτέρω βελτίωση της ακρίβειας των μοντέλων, χρησιμοποιούνται επιπλέον ιστορικά δεδομένα θερμοκρασίας, πίεσης και υγρασίας. Τα αποτελέσματα του πειράματος δείχνουν ότι οι αλγόριθμοι των δέντρων αποφάσης και τυχαίων δασών επιτυγχάνουν την υψηλότερη ακρίβεια πρόβλεψης, ενώ τα συμπληρωματικά ιστορικά δεδομένα βελτιώνουν την απόδοση των αλγοριθμικών μοντέλων.

Λέξεις κλειδιά: μηχανική μάθηση, πρόβλεψη θερμοκρασίας, δεδομένα αισθητήρων, παλινδρόμηση, χρονοσειρές

Abstract

Data collection via sensors and the extraction of useful knowledge from raw data demand specific handling in computer science. This thesis conducts the modelling of temperature data derived from sensors placed in eight different rooms of a residence located in Lefkovrysi, Kozani, aiming to accurately predict the temperature in each room. The collection of the raw data from the sensor infrastructure is followed by data processing and analysis. Seven supervised learning algorithms along with a statistical model are utilized for the prediction of temperature in each room separately. Subsequently, a comparison between the level of accuracy of each model is drawn in order to examine the efficiency of the algorithms. For further improvement of the models' accuracy, additional historical data of temperature, atmospheric pressure and humidity were used. The experimental results show that the Decision Tree and Random Forest algorithms achieve the highest prediction accuracy, whilst the supplementary historical data enhance the models' performance.

Keywords: machine learning, temperature prediction, sensor data, regression, time series

Δήλωση Πνευματικών Δικαιωμάτων

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα Διπλωματική Εργασία με τίτλο ''Σύγκριση μεθόδων μηχανικής μάθησης σε δεδομένα που προέρχονται από αισθητήρες'' καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας και αναφέρονται ρητώς μέσα στο κείμενο που συνοδεύουν, και η οποία έχει εκπονηθεί στο Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας, υπό την επίβλεψη του μέλους του Τμήματος κ. Πλόσκα αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή / και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και μόνο.

Υπογραφή Φοιτητή

Ευχαριστίες

Για την εκπόνηση της παρούσας διπλωματικής θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντά μου κύριο Πλόσκα Νικόλαο, Επίκουρο Καθηγητή του Εργαστηρίου Ευφυών Συστημάτων και Βελτιστοποίησης του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας που με τις κατευθυντήριες οδηγίες του και τη συνεργασία μας συνέβαλλε στη διεκπεραίωση της παρούσας διατριβής. Επίσης θα ήθελα να ευχαριστήσω τον κύριο Γκάλφα Νικόλαο, μέλος Ε.ΔΙ.Π του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας για την παροχή των δεδομένων από τους αισθητήρες που εγκαταστάθηκαν με σκοπό τη λήψη αποτελεσμάτων. Ακόμη, θα ήθελα να ευχαριστήσω τη Διεύθυνση της Εθνικής Μετεωρολογικής Υπηρεσίας (Ε.Μ.Υ) για τη διάθεση δεδομένων θερμοκρασίας, πίεσης και υγρασίας για την πόλη της Κοζάνης την περίοδο 2019 έως 2020, που ήταν απαραίτητα για τη συγκριτική αξιόλογηση των αποτελεσμάτων. Τέλος, δε θα μπορούσα να παραλείψω την οικογένεια μου, ιδιαίτερα τους γονείς μου, για τη στήριξη που μου παρείχαν με το δικό τους τρόπο όλα αυτά τα χρόνια.

Περιεχόμενα

1 Εισαγωγή **14**

1.1 Ορισμός του προβλήματος 14

1.2 Κίνητρα και στόχοι υλοποίησης 15

1.3 Διάρθρωση κειμένου 15

2 Μηχανική Μάθηση **17**

2.1 Εισαγωγή στη μηχανική μάθηση 17

2.1.1 Εποπτευόμενη μάθηση 17

2.1.2 Μη εποπτευόμενη μάθηση 19

2.1.3 Ενισχυτική μάθηση 19

2.2 Ορισμοί και βασικές έννοιες 20

2.2.1 Προεπεξεργασία δεδομένων 20

2.2.2 Δεδομένα εκπαίδευσης και δεδομένα δοκιμών 20

2.2.3 Μηχανική χαρακτηριστικών 21

2.2.4 Διασταυρούμενη επικύρωση 21

2.2.5 Ρύθμιση υπερπαραμέτρων 22

2.2.6 Κλιμάκωση χαρακτηριστικών 22

2.2.7 Υποεκπαίδευση και υπερεκπαίδευση 23

2.3 Αλγόριθμοι μηχανικής μάθησης 24

2.3.1 Γραμμική παλινδρόμηση 24

2.3.2 Παλινδρόμηση Lasso 24

2.3.3 Παλινδρόμηση Ridge 25

2.3.4 Δέντρα απόφασης 25

2.3.5 Τυχαία Δάση 27

2.3.6 Μηχανές Διανυσμάτων Υποστήριξης 29

2.3.7 Multi-layer Perceptron 31

2.4 Μοντέλα Στατιστικής Ανάλυσης 33
2.4.1 Εισαγωγή 33
2.4.2 Μοντέλο ARIMA 33

3 Σχετική Βιβλιογραφία **35**
3.1 Εισαγωγή 35
3.2 Μελέτη για την πρόβλεψη του καιρού στο Tennessee. 35
3.3 Εφαρμογή αλγόριθμων μηχανικής μάθησης σε δεδομένα αισθητήρων για συστάσεις άρδευσης. 37
3.4 Μελέτη για την πρόβλεψη της αστικής θερμοκρασίας της Ζυρίχης χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης. 39
3.5 Σύγκριση αλγόριθμων μηχανικής μάθησης για την πρόβλεψη της εσωτερικής θερμοκρασίας σε έξυπνα σπίτια. 41
3.6 Πρόβλεψη της εσωτερικής θερμοκρασίας σε ένα IoT σενάριο. 44
3.7 Εφαρμογή αλγόριθμων μηχανικής μάθησης σε δεδομένα αισθητήρων για την πρόβλεψη του αριθμού ατόμων σε ένα χώρο. 46
3.8 Πρόβλεψη της μελλοντικής ωριαίας κατανάλωσης ενέργειας με τη χρήση αλγόριθμων μηχανικής μάθησης. 48
3.9 Πρόβλεψη της ποιότητας του αέρα με χρήση του IoT και μηχανικής μάθησης. 50
3.10 Σύγκριση τεχνικών μηχανικής μάθησης για την πρόβλεψη της ηλιακής ενέργειας με δεδομένα καιρού από αισθητήρες. 51
3.11 Παρακολούθηση του καιρού σε πραγματικό χρόνο και πρόβλεψη με τη χρήση αισθητήρων σε αστικά λεωφορεία και μηχανικής μάθησης. . 54
3.12 Πρόβλεψη του θερμικού φορτίου σε δίκτυα τηλεθέρμανσης χρησιμοποιώντας μηχανική εκμάθηση και συμβουλές από ειδικούς. 56

4 Υπολογιστικά Πειράματα **58**
4.1 Εισαγωγή 58
4.2 Περιγραφή συστήματος αισθητήρων 58
4.3 Συλλογή δεδομένων 59
4.4 Προεπεξεργασία δεδομένων 60
4.4.1 Δεδομένα που λείπουν 61

4.4.2 Μηχανική χαρακτηριστικών 63
4.4.3 Κλιμάκωση Χαρακτηριστικών 65
4.4.4 Διαχωρισμός δεδομένων σε σύνολα εκπαίδευσης και σύνολα δοκιμών 66
4.5 Ανάλυση δεδομένων 66
4.6 Μοντελοποίηση προβλήματος παλινδρόμησης 72
4.6.1 Περιγραφή 72
4.6.2 Μοντελοποίηση με το στατιστικό μοντέλο ARIMA 77
4.6.3 Διαγράμματα ισοτιμίας 80

5 Συμπεράσματα **99**

Κατάλογος σχημάτων

2.1 Είδη μηχανικής μάθησης. 18
2.2 Διασταυρούμενη επικύρωση 5 πτυχώσεων [Muller and Guido, 2016]. . 22
2.3 Παράδειγμα αλγόριθμου Δέντρων απόφασης [Drakos, 2019]. 26
2.4 Παράδειγμα αλγόριθμου Random Forest [Wood, 2021]. 28
2.5 Παράδειγμα γραμμικά διαχωρίσιμου SVM δύο διαστάσεων. Με διακεκομμένες γραμμές παρουσιάζονται κάποια από τα άπειρα υπερεπίπεδα ή όρια απόφασης που υπάρχουν [Han et al., 2012]. 30
2.6 Παράδειγμα μη διαχωρίσιμου γραμμικά SVM δύο διαστάσεων όπου το όριο απόφασης που προκύπτει είναι μη γραμμικό [Han et al., 2012]. 30
2.7 Παράδειγμα αλγόριθμου MLP με ένα κρυφό επίπεδο [Scikit-learn, 2021]. 32

3.1 (α) Το RMSE προσθέτοντας γειτονικές πόλεις, (β) Το RMSE όταν αυξηθούν τα δεδομένα στο σετ εκπαίδευσης, (γ) Το RMSE για όλα τα μοντέλα μηχανικής μάθησης [Jakaria et al., 2020]. 37
3.2 Παράδειγμα ενός δέντρου απόφασης το οποίο ήταν μέρος του GBRT μοντέλου [Goldstein et al., 2018]. 39
3.3 Ποσοστό επιτυχίας για όλα τα μοντέλα που χρησιμοποιήθηκαν [Goldstein et al., 2018]. 40
3.4 Σύκριση μεταξύ των προβλεπόμενων και των δεδομένων μετρήσεων για α) τα δεδομένα αισθητήρων αναφοράς (RMSE=1.43), β) τα δεδομένα μετεωρολογικών σταθμών (RMSE=1.35) και γ) τα δεδομένα ατομικών αισθητήρων (RMSE=1.71) [Zumwald et al., 2021]. 41
3.5 Μέση τιμή του R-coefficient των 20 καλύτερων αλγόριθμων παλινδρόμησης για ορίζοντα πρόβλεψης τριών ωρών [Alawadi et al., 2020]. . . 43

3.6 **(Αριστερά)** Η εξέλιξη των δύο χρονοσειρών, όπου παρουσιάζεται η εσωτερική (κόκκινο) και η εξωτερική (μπλε) θερμοκρασία. **(Δεξιά)** Η εξέλιξη των αλλαγών της εξωτερικής θερμοκρασίας ως συνάρτηση της εσωτερικής θερμοκρασίας [Monteiro et al., 2018]. 45
3.7 **(Αριστερά)** Η τιμή MSE που παρατηρήθηκε για τις 6 μεθόδους πρόβλεψης. **(Δεξιά)** Η εξέλιξη του MSE για τη γραμμική παλινδρόμηση, ως συνάρτηση των ημερών που χρησιμοποιήθηκαν για τα δεδομένα εκπαίδευσης [Monteiro et al., 2018]. 46
3.8 Αξιολόγηση των αλγόριθμων κατηγοριοποίησης και παλινδρόμησης [Moraru et al., 2010]. 48
3.9 Γραφήματα γραμμικής παλινδρόμησης για δύο συνεχόμενες ημέρες [Kumar et al., 2020]. 50
3.10 Μετρήσεις απόδοσης για το μοντέλο σύγκρισης και για το προτεινόμενο μοντέλο [Kumar et al., 2020]. 51
3.11 Μέθοδοι μηχανικής μάθησης [Carrera and Kim, 2020]. 53
3.12 Συγκριτικά στατιστικά στοιχεία των μοντέλων πρόβλεψης με βάση τα δεδομένα πρόβλεψης και δεδομένα παρατήρησης καιρού, χρησιμοποιώντας διασταυρούμενη επικύρωση 10 πτυχώσεων [Carrera and Kim, 2020]. 54
3.13 Αποτελέσματα πρόβλεψης θερμοκρασίας, υγρασίας και πίεσης στις 21 Μαρτίου 2020: Σύγκριση LSTM και MLP μοντέλου [Huang et al., 2020]. 55
3.14 Σύνολο χαρακτηριστικών με ώρα της ημέρας (HoD), ημέρα της εβδομάδας (DoW), ημέρα του χρόνου (DoY), προβλεπόμενη εξωτερική θερμοκρασία (T_{out}), Θερμικό φορτίο προηγούμενης μέρας (P_{t-24}) και προηγούμενης εβδομάδας (P_{t-168}), σήματα ελεγκτή προηγούμενης ημέρας (dT_{t-24}) και προηγούμενης εβδομάδας (dT_{t-168}) [Geysen et al., 2017]. 57
3.15 Σύγκριση της μέτρησης MAPE χωρίς επανεκπαίδευση (πάνω) και με επανεκπαίδευση (κάτω) [Geysen et al., 2017]. 57

4.1 Το μενού του αισθητήρα της εξωτερικής αυλής και οι πληροφορίες που καταγράφονται. 59
4.2 Το νέο DataFrame με τα νέα χαρακτηριστικά που δημιουργήθηκαν. . . 64

4.3 Το νέο DataFrame με τα νέα χαρακτηριστικά που δημιουργήθηκαν με τα ιστορικά δεδομένα της Ε.Μ.Υ. 65
4.4 Η οπτική απεικόνιση των δεδομένων θερμοκρασίας στον χώρο του μπάνιου και του γκαράζ. 68
4.5 Η οπτική απεικόνιση των δεδομένων θερμοκρασίας στον χώρο του γραφείου και του υπνοδωματίου. 68
4.6 Η οπτική απεικόνιση των δεδομένων θερμοκρασίας στον χώρο του 2ου και 3ου υπνοδωματίου. 69
4.7 Η οπτική απεικόνιση των δεδομένων θερμοκρασίας στον χώρο της κουζίνας και της εξωτερικής αυλής. 69
4.8 Η κατανομή των τιμών για τους χώρους του μπάνιου, γκαράζ, γραφείου και υπνοδωματίου. 70
4.9 Η κατανομή των τιμών για τους χώρους του δεύτερου και τρίτου υπνοδωματίου, της κουζίνας και της εξωτερικής αυλής. 71
4.10 Οπτικοποίηση του συνόλου δεδομένων και των προβλεπόμενων δεδομένων του μοντέλου ARIMA για τους πρώτους τέσσερις χώρους. . . . 79
4.11 Οπτικοποίηση του συνόλου δεδομένων και των προβλεπόμενων δεδομένων του μοντέλου ARIMA για τους υπόλοιπους τέσσερις χώρους. . 80
4.12 Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του μπάνιου. 83
4.13 Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του γκαράζ. 84
4.14 Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του γραφείου. 85
4.15 Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του υπνοδωματίου. 86
4.16 Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του δεύτερου υπνοδωματίου. 87
4.17 Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του τρίτου υπνοδωματίου. 88
4.18 Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο της κουζίνας. 89

4.19 Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο της αυλής. 90
4.20 Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του μπάνιου. 91
4.21 Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του γκαράζ. 92
4.22 Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του γραφείου. 93
4.23 Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του υπνοδωματίου. 94
4.24 Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του δεύτερου υπνοδωματίου. 95
4.25 Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του τρίτου υπνοδωματίου. 96
4.26 Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο της κουζίνας. 97
4.27 Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο της αυλής. 98

Κατάλογος πινάκων

4.1 Οι μετρήσεις του R^2 και RMSE χωρίς διασταυρούμενη επικύρωση. . . 73
4.2 Οι μετρήσεις του R^2 και RMSE με διασταυρούμενη επικύρωση. 74
4.3 Οι υπερπαράμετροι που επιλέχθηκαν για βελτίωση. 75
4.4 Οι μετρήσεις του R^2 και RMSE με ρύθμιση υπερπαραμέτρων. 76
4.5 Οι μετρήσεις του R^2 και RMSE με ρύθμιση υπερπαραμέτρων με τα δεδομένα της Ε.Μ.Υ. 77
4.6 Οι παράμετροι που επιλέχθηκαν και τα αποτελέσματα των μετρήσεων του μοντέλου ARIMA. 78

Κατάλογος απεικονίσεων

4.1 Αφαίρεση των γραμμάτων T και Z από τη μεταβλητή του χρόνου. . . 60
4.2 Εξαγωγή του μέσου όρου για κάθε ώρα ξεχωριστά. 61
4.3 Συμπλήρωση των ελλειπόντων τιμών που προέκυψαν στο DataFrame με τα δεδομένα της Ε.Μ.Υ με τη χρήση της μεθόδου KNNImputer. . . 63
4.4 Δημιουργία χαρακτηριστικών που σχετίζονται με την ημερομηνία και την ώρα. 63
4.5 Διαχωρισμός των χαρακτηριστικών σε ανεξάρτητες μεταβλητές εισόδου x και εξαρτημένης μεταβλητής εξόδου y. 65
4.6 Κλιμάκωση των χαρακτηριστικών με τη χρήση της μεθόδου StandardScaler. 65
4.7 Διαχωρισμός του συνόλου δεδομένων σε υποσύνολα εκπαίδευσης και δοκιμών. 66

Κεφάλαιο 1

Εισαγωγή

1.1 Ορισμός του προβλήματος

Η συλλογή δεδομένων ήταν ανέκαθεν αναγκαία στην καθημερινότητα και εξυπηρετούνταν σε βαθμό αντίστοιχο με τα μέσα της κάθε χρονικής περιόδου. Οι άνθρωποι συλλέγουν δεδομένα με σκοπό τη μετάφρασή τους σε χρήσιμες πληροφορίες όπως η αναγνώριση μοτίβων και η εξαγωγή συμπερασμάτων από αυτά. Στις μέρες μας, είναι απαραίτητη η συλλογή δεδομένων και πληροφοριών για διάφορους τομείς και επιστημονικούς κλάδους, όπως η γεωργία, η ιατρική, η μετεωρολογία, η εναέρια κυκλοφορία, κ.α. Ο όγκος των διαθέσιμων δεδομένων έχει αυξηθεί δραματικά σε σύγκριση με το παρελθόν, καθιστώντας αναγκαία την εφαρμογή ενός αυτοματοποιημένου τρόπου συλλογής και διαχείρισής τους. Το πρόβλημα της αυτοματοποιημένης συλλογής δεδομένων επιλύεται με τη χρήση αισθητήρων, συσκευών που τοποθετούνται σε οποιοδήποτε περιβάλλον με στόχο την ανίχνευση ενός φυσικού μεγέθους και την παραγωγή μίας μετρήσιμης εξόδου από αυτό. Τα δεδομένα αισθητήρων που συλλέγονται συνεχώς από πολλαπλούς κόμβους αισθητήρων, προωθούνται μέσω ασύρματων δικτύων σε μία κεντρική βάση για περεταίρω επεξεργασία [Wang and Liu, 2011]. Παραδείγματα μετρήσεων ενός αισθητήρα αποτελούν η μέτρηση τιμών στην ατμόσφαιρα όπως η υγρασία ή η θερμοκρασία, η μέτρηση της κατανάλωσης ενέργειας ή ακόμη και μετρήσεις τιμών στον ανθρώπινο οργανισμό. Καθώς ο όγκος δεδομένων που καταγράφονται από αισθητήρες ξεπερνά τις δυνατότητες επεξεργασίας από τον άνθρωπο, είναι αναγκαία η συλλογή και επεξεργασία τους από υπολογιστικά μηχανήματα.

1.2 Κίνητρα και στόχοι υλοποίησης

Στη σύγχρονη εποχή, η επιστήμη των υπολογιστών εξελίσσεται με ραγδαίους ρυθμούς και ο τομέας της μηχανικής μάθησης προσφέρει πολλές επιλογές στη μελέτη και ανάπτυξη μοντέλων πρόβλεψης δεδομένων που προέρχονται από αισθητήρες, ακόμη και για προσωπική χρήση. Η πρόβλεψη τιμών από μετρήσεις αισθητήρων μπορεί να είναι ιδιαίτερα επωφελής και υπάρχει δυνατότητα εξατομικευμένης εφαρμογής με τα κατάλληλα δεδομένα. Για παράδειγμα, είναι εφικτό να παρακολουθούμε την εναλλαγή των καιρικών φαινομένων με ακρίβεια για τον τόπο που βρισκόμαστε και να προγραμματίζουμε τις εργασίες και δραστηριότητές μας. Ακόμη, μπορούμε να ρυθμίζουμε τα συστήματα κατανάλωσης ενέργειας του σπιτιού μας, όπως τα συστήματα θέρμανσης, ψύξης και τους ηλιακούς θερμοσίφωνες. Με βάση αυτές τις πληροφορίες δίνεται η δυνατότητα να ζούμε σε ένα περιβάλλον που λειτουργεί σύμφωνα με τις ανάγκες αλλά και τις συνήθειές μας, εξοικονομώντας ταυτόχρονα πόρους ενέργειας και χρήματα. Στην παρούσα διπλωματική στόχος είναι η πρόβλεψη της θερμοκρασίας σε οκτώ διαφορετικούς χώρους ενός σπιτιού στην Λευκόβρυση Κοζάνης χρησιμοποιώντας δεδομένα θερμοκρασίας που προέρχονται από αισθητήρες τοποθετημένους σε κάθε έναν από αυτούς τους χώρους. Προκειμένου να γίνουν αυτές οι προβλέψεις θερμοκρασίας, χρησιμοποιήθηκαν αλγόριθμοι μηχανικής μάθησης, οι οποίοι έχουν τη δυνατότητα να παράγουν ακριβή και αξιόπιστα αποτελέσματα σχετικά με τις τιμές της θερμοκρασίας. Τα αποτελέσματα των προβλέψεων θα μπορούσαν να χρησιμοποιηθούν για τη χαρτογράφηση των αναγκών κατανάλωσης ενέργειας σε κάθε χώρο και τη ρύθμιση των πόρων που παρέχονται.

1.3 Διάρθρωση κειμένου

Η παρούσα διπλωματική εργασία αποτελείται από πέντε κεφάλαια. Στο δεύτερο κεφάλαιο αναλύονται οι βασικές έννοιες του τομέα της μηχανικής μάθησης καθώς και η ανάλυση των αλγόριθμων μηχανικής μάθησης που θα χρησιμοποιηθούν στο συγκεκριμένο πείραμα. Στη συνέχεια, στο τρίτο κεφάλαιο παρουσιάζονται 11 μελέτες οι οποίες βασίζονται στην ανάπτυξη και εφαρμογή αλγοριθμικών μοντέλων για την πρόβλεψη μετρήσεων σε δεδομένα που προέρχονται από αισθητήρες. Το τέταρτο κεφάλαιο περιέχει τη συνολική διαδικασία του πειράματος. Το πείραμα

ξεκινά από τη λήψη των δεδομένων από τους αισθητήρες, την επεξεργασία των δεδομένων και τη δημιουργία του συνόλου δεδομένων. Συνεχίζοντας, δημιουργούνται αλγοριθμικά μοντέλα μηχανικής μάθησης και εκπαιδεύονται στο σύνολο δεδομένων. Ακολουθούν οι τρόποι διαχείρισης του πειράματος, οι τεχνικές που εφαρμόστηκαν και παρουσιάζεται η ανάλυση και τα αποτελέσματα των αλγόριθμων που χρησιμοποιήθηκαν. Τέλος, στο πέμπτο κεφάλαιο παρουσιάζονται τα συμπεράσματα που προέκυψαν κατά τη διαδικασία του πειράματος και αφορούν την εκτίμηση της κάθε μεθόδου μεμονωμένα καθώς και τη σύγκρισή τους ως προς την αποδοτικότητά τους.

Κεφάλαιο 2

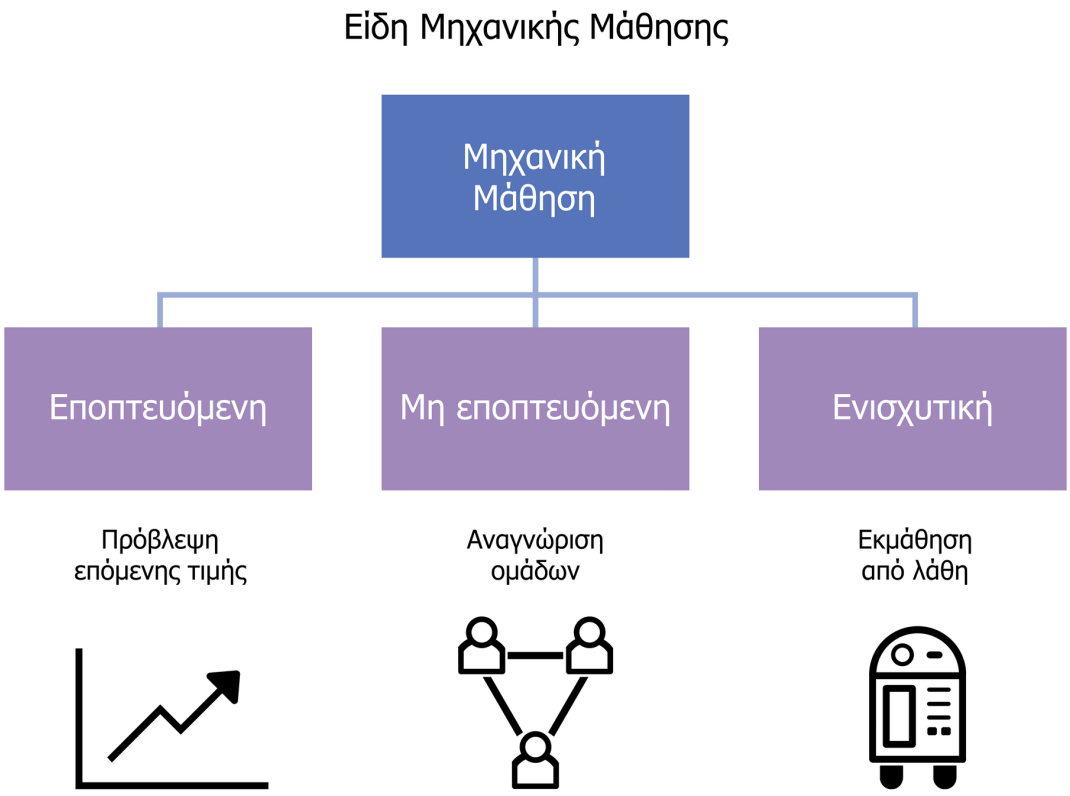
Μηχανική Μάθηση

2.1 Εισαγωγή στη μηχανική μάθηση

Η μηχανική μάθηση (machine learning) είναι ένα πεδίο της επιστήμης υπολογιστών, η οποία παρέχει στα συστήματα την ικανότητα να μαθαίνουν αυτοματοποιημένα και να βελτιώνονται με βάση την εμπειρία, χωρίς να έχουν προγραμματιστεί ρητά. Στο πεδίο της μηχανικής μάθησης συνδυάζονται η επιστήμη των υπολογιστών, η στατιστική και η τεχνητή νοημοσύνη, με σκοπό τη μελέτη, την κατασκευή και την ανάλυση των αλγόριθμων μηχανικής μάθησης. Η μηχανική μάθηση εστιάζει στην κατασκευή αλγόριθμων που έχουν πρόσβαση σε δεδομένα και τα χρησιμοποιούν για να μάθουν. Η διαδικασία εκμάθησης ξεκινά από ένα σύνολο παρατηρήσεων ή δεδομένων, όπως παραδείγματα ή οδηγίες, με σκοπό την εύρεση διάφορων μοτίβων στα δεδομένα και τη λήψη καλύτερων αποφάσεων στο μέλλον, με βάση τα παραδείγματα που παρέχουμε. Ο πρωταρχικός στόχος είναι, να επιτρέπεται στους υπολογιστές να μαθαίνουν αυτόματα χωρίς την ανθρώπινη παρέμβαση ή βοήθεια. Οι αλγόριθμοι μηχανικής μάθησης συχνά κατηγοριοποιούνται ανάλογα με το αν εκπαιδεύονται με ανθρώπινη επίβλεψη ή όχι. Στο Σχήμα 2.1 απεικονίζονται τα τρία βασικά είδη μηχανικής μάθησης τα οποία είναι: η εποπτευόμενη μάθηση (supervised learning), η μη εποπτευόμενη μάθηση (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning).

2.1.1 Εποπτευόμενη μάθηση

Στην εποπτευόμενη μάθηση, χρησιμοποιούμε τις μεταβλητές πρόβλεψης/χαρακτηριστικά με σκοπό να προβλέψουμε τη μεταβλητή στόχο. Είναι η μέθοδος κατά



Σχήμα 2.1: Είδη μηχανικής μάθησης.

την οποία μαθαίνουμε στη μηχανή χρησιμοποιώντας δεδομένα με ετικέτα, δηλαδή τα χαρακτηριστικά των δεδομένων εισόδου και εξόδου είναι επισημασμένα. Αρχικά, δημιουργούμε ένα μοντέλο μηχανικής μάθησης με βάση τα δεδομένα εισόδου και εξόδου τα οποία αποθηκεύουμε σε ένα σετ δεδομένων εκπαίδευσης και στη συνέχεια προσπαθούμε να κάνουμε προβλέψεις για καινούργια άγνωστα σε εμάς δεδομένα. Στην εποπτευόμενη μάθηση απαιτείται η ανθρώπινη παρέμβαση καθώς τα δεδομένα εισόδου, εξόδου όπως και οι αλγόριθμοι και τα σενάρια παρέχονται από τον άνθρωπο στο σύστημα. Οι δύο βασικοί τύποι εποπτευόμενης μάθησης είναι η κατηγοριοποίηση (classification) και η παλινδρόμηση (regression).

Κατηγοριοποίηση είναι η διαδικασία κατά την οποία προσπαθούμε να αναγνωρίσουμε και να ομαδοποιήσουμε τα χαρακτηριστικά των δεδομένων σε προκαθορισμένες υποκατηγορίες. Παραδείγματα κατηγοριοποίησης είναι η αναγνώριση ενός spam e-mail στα εισερχόμενά μας ή η αναγνώριση και ομαδοποίηση διαφόρων ειδών φυτών. Στην παλινδρόμηση, στόχος είναι η πρόβλεψη μίας συνεχούς μεταβλητής βασιζόμενη στα χαρακτηριστικά εισόδου-εξόδου. Ένα παράδειγμα παλινδρόμησης είναι η πρόβλεψη της τιμής διάφορων σπιτιών με βάση το μέγεθός τους, την περιοχή

και την τιμή τους. Κάποιοι από τους πιο ευρέως χρησιμοποιούμενους αλγόριθμους εποπτευόμενης μάθησης είναι οι εξής: αλγόριθμος γραμμικής παλινδρόμησης, αλγόριθμος λογιστικής παλινδρόμησης, τα δέντρα απόφασης, των μηχανών διανυσμάτων υποστήριξης, των τυχαίων δασών και των νευρωνικών δικτύων [Baştanlar and Özuysal, 2014].

2.1.2 Μη εποπτευόμενη μάθηση

Στη μη εποπτευόμενη μάθηση, παρέχουμε στον αλγόριθμο ένα σύνολο δεδομένων χωρίς ετικέτες και ζητάμε να εξαγάγει γνώσεις από αυτά τα δεδομένα. Αρχικά δημιουργούμε ένα μοντέλο μηχανικής μάθησης με βάση τα δεδομένα εισόδου χωρίς ετικέτα, και ο αλγόριθμος χωρίς καθοδήγηση αλλά και χωρίς επίγνωση των δεδομένων εξόδου, προσπαθεί να βγάλει συμπεράσματα για τα δεδομένα χωρίς ετικέτα. Αυτή η διαδικασία συνήθως γίνεται μέσω της εξερεύνησης και ομαδοποίησης παρόμοιων παραδειγμάτων/χαρακτηριστικών ή τη μείωση των διαστάσεων των δεδομένων με σκοπό να ταιριάζουν καλύτερα σε ένα μοντέλο πρόβλεψης. Μερικοί αλγόριθμοι μη εποπτευόμενης μάθησης είναι οι εξής: αλγόριθμος k-means ομαδοποίησης, της ανάλυσης κύριων συνιστωσών, της ιεραρχικής ομαδοποίησης και της πολλαπλής μάθησης [Baştanlar and Özuysal, 2014].

2.1.3 Ενισχυτική μάθηση

Το τρίτο είδος μηχανικής μάθησης είναι η ενισχυτική μάθηση, στο οποίο όπως και στη μη εποπτευόμενη δεν υπάρχουν δεδομένα με ετικέτα. Ο μαθησιακός αλγόριθμος εκπαιδεύεται στο να παίρνει μια σειρά από αποφάσεις στο περιβάλλον, μέσα από έναν μηχανισμό επιβράβευσης και τιμωρίας για κάθε πράξη. Για παράδειγμα, σε ένα περιβάλλον ενός αυτόματου αυτοκινήτου, ορίζουμε την αναγνώριση των ορίων του δρόμου και τα όρια ταχύτητας, και επιβραβεύουμε τον αλγόριθμο κάθε φορά που προχωρά εντός ορίων με αποδεκτή ταχύτητα, και τον τιμωρούμε κάθε φορά που ξεπερνά αυτά τα όρια. Ο αλγόριθμος δεν μπορεί να έχει επίγνωση όλων των πιθανών σεναρίων στο περιβάλλον, αλλά μπορεί να εκπαιδεύεται με βάση αυτά που συναντά και μαθαίνει με την πάροδο του χρόνου [Thrun, 1992].

2.2 Ορισμοί και βασικές έννοιες

2.2.1 Προεπεξεργασία δεδομένων

Η προεπεξεργασία δεδομένων (data preprocessing) αποτελεί ένα από τα πιο σημαντικά και κρίσιμα βήματα στην εφαρμογή των αλγορίθμων μηχανικής μάθησης. Είναι η διαδικασία κατά την οποία, προετοιμάζουμε και οργανώνουμε τα ακατέργαστα δεδομένα, έτσι ώστε να είναι κατάλληλα για ένα μοντέλο μηχανικής μάθησης. Αυτή η προετοιμασία συνήθως αφορά τον καθαρισμό των δεδομένων, όπως για παράδειγμα από κενά ή κατεστραμμένα σημεία δεδομένων, ή τη μετατροπή των δεδομένων σε μια πιο κατανοητή μορφή.

2.2.2 Δεδομένα εκπαίδευσης και δεδομένα δοκιμών

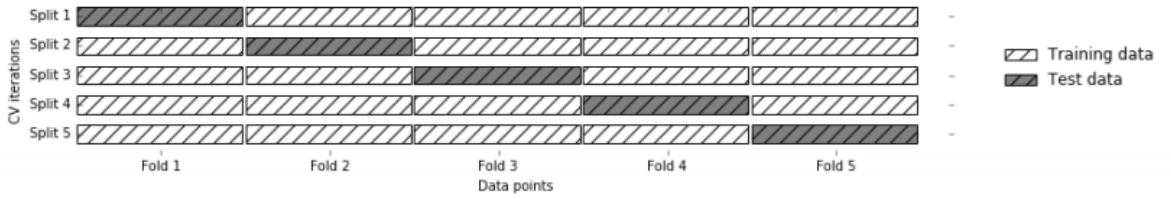
Για την αποτελεσματική εφαρμογή ενός αλγόριθμου σε ένα σύνολο δεδομένων, θα πρέπει να έχουμε διαχωρίσει αυτό το σύνολο σε δύο υποσύνολα, το σύνολο εκπαίδευσης (training set) και το σύνολο δοκιμών (test set). Τα δύο αυτά σύνολα είναι σημαντικό κομμάτι στη δημιουργία αλγοριθμικών μοντέλων, καθώς και στην αξιολόγηση αυτών των μοντέλων. Αρχικά θα πρέπει να διαχωρίσουμε αυτά τα δύο σύνολα έτσι ώστε να μην υπάρχουν τα ίδια δεδομένα και στο σύνολο εκπαίδευσης και στο σύνολο δοκιμών. Συνήθως χωρίζουμε τα δύο υποσύνολα δεδομένων (80% σύνολο εκπαίδευσης - 20% σύνολο δοκιμών), όπου το σύνολο εκπαίδευσης αποτελεί το πραγματικό σύνολο δεδομένων που χρησιμοποιούμε για να εκπαιδεύσουμε τον αλγόριθμο και το σύνολο δοκιμών αποτελεί το σύνολο δεδομένων που χρησιμοποιείται για την αξιολόγηση του αλγοριθμικού μοντέλου. Επίσης, θα πρέπει να φροντίσουμε να είναι τυχαία τα δεδομένα που βρίσκονται στο κάθε υποσύνολο, έτσι ώστε να έχουμε σε κάθε υποσύνολο μία γενική εικόνα για τα δεδομένα. Για παράδειγμα, δε θέλουμε στο σύνολο εκπαίδευσης να βρίσκονται δεδομένα που αφορούν μόνο τη χειμερινή περίοδο και στο σύνολο δοκιμών δεδομένα που αφορούν μόνο την καλοκαιρινή περίοδο, καθώς οι προβλέψεις μας δε θα ήταν ακριβείς. Όσο μεγαλύτερο είναι το μέγεθος του συνόλου εκπαίδευσης, τόσο περισσότερη γνώση θα λαμβάνει ο αλγόριθμος από τα δεδομένα και οι προβλέψεις θα είναι πιο ακριβείς.

2.2.3 Μηχανική χαρακτηριστικών

Η μηχανική χαρακτηριστικών (feature engineering) είναι η διεργασία κατά την οποία επεξεργαζόμαστε τα ακατέργαστα δεδομένα για να δημιουργήσουμε χαρακτηριστικά που αντιπροσωπεύουν καλύτερα το πρόβλημα που θέλουμε να αναλύσουμε με τα μοντέλα πρόβλεψης, με σκοπό τη βελτίωση της προγνωστικής δύναμης των αλγοριθμικών μοντέλων. Για παράδειγμα, σε ένα σύνολο δεδομένων με μια ημερομηνία για είσοδο και μια μεταβλητή όπως ο αριθμός ανθρώπων που χρησιμοποιούν το μετρό, η πρόβλεψη βασίζεται περισσότερο στην ημέρα και την ώρα, παρά σε ολόκληρη την ημερομηνία. Θα μπορούσαμε να χρησιμοποιήσουμε ως χαρακτηριστικά (features) τις ώρες, καθώς έχει περισσότερη κίνηση τις πρωινές ώρες ή την ημέρα, καθώς η κίνηση διαφέρει το Σάββατο και την Κυριακή από τις καθημερινές. Έτσι, μπορούμε να βγάλουμε πιο εύστοχα συμπεράσματα για τα δεδομένα και τις προβλέψεις των αλγοριθμικών μοντέλων [Dami and Esterabi, 2021].

2.2.4 Διασταυρούμενη επικύρωση

Η διασταυρούμενη επικύρωση (cross validation) είναι μια τεχνική για την αξιολόγηση μοντέλων μηχανικής μάθησης, κατά την οποία εκπαιδεύουμε τα μοντέλα σε υποσύνολα των διαθέσιμων δεδομένων εισόδου, χρησιμοποιώντας τα υπόλοιπα υποσύνολα δεδομένων για την αξιολόγησή των μοντέλων. Η πιο διαδεδομένη έκδοση διασταυρούμενης επικύρωσης είναι η k-fold cross validation ενώ υπάρχουν διάφορες παραλλαγές διασταυρούμενης επικύρωσης. Στην διασταυρούμενη επικύρωση k-fold, διαχωρίζουμε τα δεδομένα εισόδου σε υποσύνολα δεδομένων, τα οποία ονομάζουμε πτυχώσεις. Ο αριθμός των πτυχώσεων k ορίζεται από το χρήστη και συνήθως είναι 5 ή 10. Στη συνέχεια, χρησιμοποιούμε τη μία πτύχωση ως σετ δεδομένων δοκιμών και τις υπόλοιπες ως σετ δεδομένων εκπαίδευσης, αξιολογώντας συνεχώς την ακρίβεια των μοντέλων μηχανικής μάθησης όπως φαίνεται στην Εικόνα 2.2. Αυτή η διεργασία συνεχίζεται μέχρι να έχουν χρησιμοποιηθεί όλες οι πτυχώσεις ως δεδομένα δοκιμών. Στο τέλος της διαδικασίας, έχουμε συλλέξει τιμές ακρίβειας των μοντέλων για την αξιολόγηση, ανάλογα με τον αριθμό των πτυχώσεων [Muller and Guido, 2016].



Σχήμα 2.2: Διασταυρούμενη επικύρωση 5 πτυχώσεων [Muller and Guido, 2016].

2.2.5 Ρύθμιση υπερπαραμέτρων

Κατά τη δημιουργία ενός μοντέλου μηχανικής μάθησης, έχουμε πολλές επιλογές σχετικά με τη ρύθμιση των παραμέτρων ενός μοντέλου μηχανικής μάθησης. Οι παράμετροι που καθορίζουν την αρχιτεκτονική ενός μοντέλου μηχανικής μάθησης, ονομάζονται υπερπαράμετροι (hyperparameters). Συχνά, δε ξέρουμε ποια επιλογή στον σχεδιασμό της αρχιτεκτονικής του μοντέλου θα ήταν η βέλτιστη και συνεπώς θα θέλαμε να διερευνήσουμε μια σειρά δυνατοτήτων. Η ρύθμιση των υπερπαραμέτρων (hyperparameter tuning) είναι η διεργασία κατά την οποία ζητάμε από το σύστημα να εξερευνήσει το εύρος των τιμών των υπερπαραμέτρων που θα καθορίσουν την αρχιτεκτονική του μοντέλου, με σκοπό την εύρεση της βέλτιστης αρχιτεκτονικής του μοντέλου μηχανικής μάθησης. Έτσι, μέσα από τα κριτήρια αξιολόγησης που έχουμε ορίσει, μπορούμε να κρίνουμε την αποτελεσματικότητα και την ακρίβεια του μοντέλου μηχανικής μάθησης [Bardenet et al., 2013].

2.2.6 Κλιμάκωση χαρακτηριστικών

Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης δεν αποδίδουν καλά, όταν τα ανεξάρτητα χαρακτηριστικά των δεδομένων εισόδου είναι σε διαφορετικές κλίμακες. Για παράδειγμα, αν ένα χαρακτηριστικό εισόδου έχει ως τιμή τα 1000 μέτρα, ενώ ένα άλλο έχει ως τιμή τα 2 χιλιόμετρα, ο αλγόριθμος λαμβάνει τα 1000 μέτρα ως πολύ μεγαλύτερη τιμή από τα 2 χιλιόμετρα, με αποτέλεσμα να μας δώσει λάθος προβλέψεις. Για την αντιμετώπιση αυτού του προβλήματος, χρησιμοποιούμε την τεχνική κλιμάκωσης χαρακτηριστικών (feature scaling), έτσι ώστε να φέρουμε όλες τις τιμές στα ίδια μεγέθη. Οι δύο πιο σημαντικές τεχνικές κλιμάκωσης χαρακτηριστικών είναι η μέγιστη-ελάχιστη κανονικοποίηση (min-max normalization) και η τυποποιημένη κανονικοποίηση (standardization) [Muller, 2017].

Η μέγιστη-ελάχιστη κανονικοποίηση αναπροσαρμόζει σε μία νέα κλίμακα των

τιμών μεταξύ 0 και 1. Αυτό επιτυγχάνεται αφαιρώντας την ελάχιστη τιμή και στη συνέχεια διαιρώντας με τη μέγιστη τιμή μείον την ελάχιστη τιμή.

$$x' = \frac{x_i - min(x)}{max(x) - min(x)}$$

Η τεχνική της τυποποίησης (standardization) αναπροσαρμόζει σε μία νέα κλίμακα η οποία έχει κατανομή με μέση τιμή 0 και η διακύμανση ίση με 1. Αυτό επιτυγχάνεται αφαιρώντας τη μέση τιμή και στη συνέχεια διαιρώντας με τη διακύμανση, έτσι ώστε η νέα κατανομή να έχει διακύμανση μονάδας.

$$x' = \frac{x_i - mean(x)}{StandardDeviation}$$

2.2.7 Υποεκπαίδευση και υπερεκπαίδευση

Ένα συνηθισμένο πρόβλημα στη διαδικασία εκπαίδευσης των μοντέλων μηχανικής μάθησης, είναι το πόσο καλά τα μοντέλα ευθυγραμμίζονται με τα δεδομένα εκπαίδευσης. Στην περίπτωση της υποεκπαίδευσης (underfitting), το μοντέλο μηχανικής μάθησης δεν παράγει ακριβείς προβλέψεις και αυτό μπορεί να σημαίνει ότι είτε το μοντέλο μηχανικής μάθησης είναι πολύ απλό για να δημιουργήσει ένα σταθερό πρότυπο μάθησης ή ότι αποδίδει πολύ άσχημα με τα δεδομένα εκπαίδευσης. Η υποεκπαίδευση μπορεί να αποφευχθεί με την αύξηση ή ρύθμιση των παραμέτρων του μοντέλου μηχανικής μάθησης, με τη δημιουργία ενός πιο σύνθετου μοντέλου ή με την αύξηση των δεδομένων για την εκπαίδευση του μοντέλου.

Το αντίθετο συμβαίνει στην περίπτωση της υπερεκπαίδευσης (overfitting), όπου το μοντέλο μηχανικής μάθησης προσαρμόζεται τέλεια στα δεδομένα εκπαίδευσης. Παρά τα καλά αποτελέσματα που παράγει το μοντέλο μηχανικής μάθησης, δεν ξέρουμε πως μπορεί να ανταποκριθεί σε μελλοντικά παραδείγματα δεδομένων. Γενικά, η υπερεκπαίδευση συμβαίνει συνήθως σε μη γραμμικά ή μη παραμετρικά μοντέλα μηχανικής μάθησης. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι η διασταυρούμενη επικύρωση k-fold, καθώς χωρίζοντας τα δεδομένα σε υποσύνολα δεδομένων, η τυχόν υπερεκπαίδευση θα μειωθεί σημαντικά [Muller, 2017].

2.3 Αλγόριθμοι μηχανικής μάθησης

2.3.1 Γραμμική παλινδρόμηση

Η γραμμική παλινδρόμηση (linear regression) είναι ένας αλγόριθμος μηχανικής μάθησης βασισμένος στην εποπτευόμενη μάθηση. Εκτελεί μία διεργασία παλινδρόμησης, όπου η προβλεπόμενη έξοδος είναι συνεχής και έχει σταθερή κλίση. Χρησιμοποιείται για την πρόβλεψη συνεχών τιμών όπως είναι οι πωλήσεις μιας εταιρίας, οι θερμοκρασίες μιας περιοχής ή ο μισθός των εργαζομένων. Σκοπός είναι η έρευση μίας γραμμικής σχέσης μεταξύ των ανεξάρτητων μεταβλητών εισόδου x και της εξαρτώμενης μεταβλητής εξόδου y. Η γραμμική εξίσωση εκχωρεί έναν συντελεστή κλίμακας (coefficient) σε κάθε τιμή εισόδου που εκφράζεται με το κεφαλαίο γράμμα b, καθώς και ένα ακόμα συντελεστή που ονομάζεται συντελεστής αναχαίτισης (intercept coefficient) που δίνει στη γραμμή ένα βαθμό ελευθερίας προς τα πάνω ή προς τα κάτω σε μία γραφική παράσταση.

Έτσι, σε ένα απλό πρόβλημα παλινδρόμησης με ένα x και ένα y, το γραμμικό μοντέλο θα είναι της μορφής:

$$y = b_0 + b_1 * x$$

Όπου b_0 είναι ο συντελεστής αναχαίτησης (intercept coefficient), b_1 ο συντελεστής κλίμακας, x είναι η ανεξάρτητη μεταβλητή και y είναι η εξαρτημένη μεταβλητή.

Όταν οι ανεξάρτητες μεταβλητές x είναι παραπάνω από μία, τότε έχουμε πολλαπλή παλινδρόμηση (multiple linear regression) και το γραμμικό μοντέλο είναι της μορφής:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_n * x_n$$

Δύο σημαντικές τεχνικές γραμμικής παλινδρόμησης είναι οι εξής: γραμμική παλινδρόμηση lasso (lasso regression) και η γραμμική παλινδρόμηση ridge (ridge regression) [Muller and Guido, 2016].

2.3.2 Παλινδρόμηση Lasso

Στη γραμμική παλινδρόμηση lasso εκτελούμε κανονικοποίηση L1 (L1 regularization), η οποία προσθέτει στη συνάρτηση του γραμμικού μοντέλου μία ποινή ίση με την απόλυτη τιμή του μεγέθους των συντελεστών (coefficients). Αυτού του είδους η

κανονικοποίηση μπορεί να οδηγήσει στην αραίωση των συντελεστών καθώς μερικοί συντελεστές μπορεί να τείνουν προς το μηδέν και να αφαιρεθούν από το μοντέλο. Η συνάρτηση της παλινδρόμησης Lasso έχει την εξής μορφή:

$$\sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{M} b_j x_{ij}\right)^2 + \alpha \sum_{j=0}^{M} |x_j|$$

Σκοπός της γραμμικής παλινδρόμησης lasso είναι η αντιμετώπιση του προβλήματος υπερεκπαίδευσης, καθώς και στην εξάλειψη του μεγάλου αριθμού ανεξάρτητων μεταβλητών X που δεν είναι χρήσιμες για τις εκτιμήσεις [Tibshirani, 1996].

2.3.3 Παλινδρόμηση Ridge

Στη γραμμική παλινδρόμηση ridge εκτελούμε κανονικοποίηση L2 (L2 regularization), η οποία προσθέτει στη συνάρτηση του γραμμικού μοντέλου μία ποινή ίση με το τετράγωνο του μεγέθους των συντελεστών. Σε αντίθεση με την κανονικοποίηση L1, όλοι οι συντελεστές συρρικνώνονται από τον ίδιο παράγοντα και έτσι δεν εξαλείφεται κανένας. Η συνάρτηση της παλινδρόμησης Ridge δίνεται από τον εξής τύπο:

$$\sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{M} b_j x_{ij}\right)^2 + \alpha \sum_{j=0}^{M} x_j^2$$

Χρησιμοποιείται κυρίως σε περιπτώσεις όπου οι ανεξάρτητες μεταβλητές εισόδου x χαρακτηρίζονται από πολλαπλή συσχέτιση (multicollinearity). Έτσι αποφεύγεται το πρόβλημα της υπερεκπαίδευσης που μπορεί να είχαμε σε μία απλή γραμμική παλινδρόμηση [Hoerl, 2020].

2.3.4 Δέντρα απόφασης

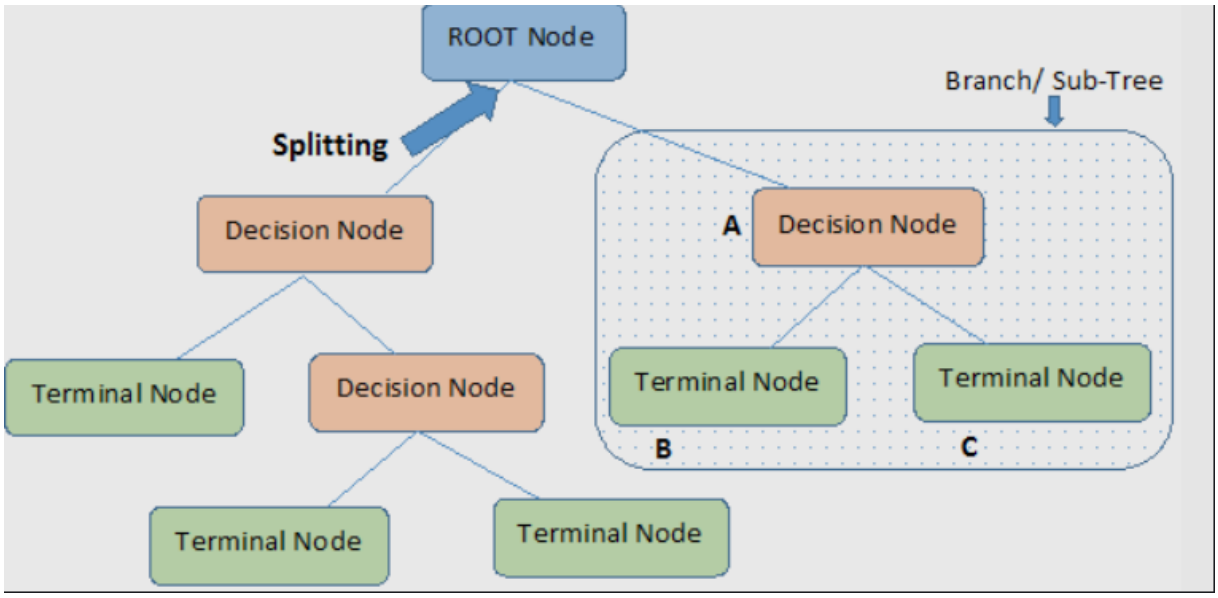
Τα Δέντρα απόφασης (Decision Trees), είναι ένας από τους πιο συχνά χρησιμοποιούμενους και πρακτικούς αλγόριθμους εποπτευόμενης μηχανικής μάθησης, ο οποίος μπορεί να χρησιμοποιηθεί για κατηγοριοποίηση αλλά και για παλινδρόμηση. Ο αλγόριθμος των δέντρων απόφασης χωρίζει το σύνολο δεδομένων σε υποσύνολα δεδομένων και στη συνέχεια μέσω ενός συνόλου ερωτήσεων δημιουργεί συσχετίσεις μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών. Ξεκινά από έναν αρχικό κόμβο (root node) και επεκτείνει προς τα κάτω δύο κλαδιά (branches) στα οποία

ανάλογα με την απάντηση, θα διαλέξει ποιο θα επεκτείνει στη συνέχεια, όπως φαίνεται στο Σχήμα 2.3. Στην περίπτωση της παλινδρόμησης, τα δέντρα απόφασης χρησιμοποιούν τη συνάρτηση του μέσου τετραγώνου σφάλματος (MSE) για να αποφασίσουν αν θα αναπτύξουν δύο ή περισσότερους κόμβους. Η συνάρτηση MSE είναι ο μέσος όρος του τετραγωνικού σφάλματος που προκύπτει μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών και εκφράζεται ως εξής:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - yi')^2$$

όπου n είναι ο συνολικός αριθμός των παρατηρήσεων, το y_i είναι οι πραγματικές τιμές και y_i' οι προβλεπόμενες τιμές των παρατηρήσεων.

Έτσι, σε κάθε βήμα ο αλγόριθμος παλινδρόμησης δέντρων απόφασης αξιολογεί πόσο καλός ήταν ο διαχωρισμός των δύο φύλλων (leafs) που δημιούργησε. Ο αλγόριθμος θα σταματήσει όταν υπάρχει μία κατάσταση τερματισμού που ορίζεται από τις υπερπαραμέτρους όπως το μέγιστο βάθος (max_depth). Επίσης, ο αλγόριθμος τερματίζει όταν υπάρχει μόνο ένα φύλλο κόμβου και δεν είναι εφικτός κάποιος άλλος διαχωρισμός [Drakos, 2019].



Σχήμα 2.3: Παράδειγμα αλγόριθμου Δέντρων απόφασης [Drakos, 2019].

Πλεονεκτήματα του αλγόριθμου [Rokach and Maimon, 2005]

- Μπορεί να χρησιμοποιηθεί σε προβλήματα κατηγοριοποίησης αλλά και παλινδρόμησης.

- Μπορεί να χειριστεί προβλήματα πολλαπλών εξόδων.
- Απαιτεί λιγότερη προετοιμασία των δεδομένων σε σχέση με τους υπόλοιπους αλγόριθμους.
- Ο τύπος των δεδομένων δεν είναι περιορισμός, καθώς μπορεί να χειριστεί τόσο αριθμητικά όσο και κατηγορικά δεδομένα (categorical data).
- Οι τυχόν μη γραμμικές σχέσεις μεταξύ των παραμέτρων δεν επηρεάζουν την επίδοσή του.

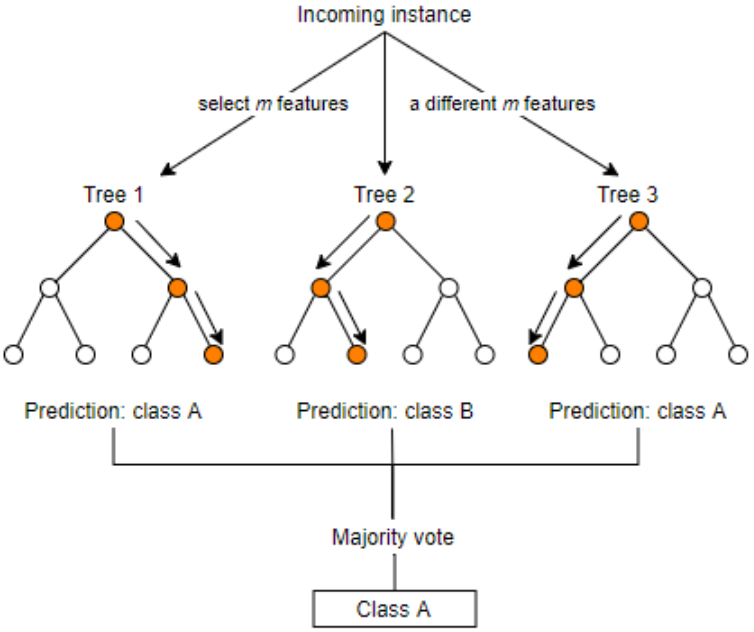
Μειονεκτήματα του αλγόριθμου [Rokach and Maimon, 2005]

- Μπορεί να δημιουργήσει υπερβολικά μεγάλα δέντρα τα οποία δεν κάνουν καλή γενίκευση (generalize) του συνόλου δεδομένων και συνεπώς να έχουμε το πρόβλημα της υπερεκπαίδευσης. Για να αποφύγουμε την υπερεκπαίδευση χρησιμοποιούμε μία τεχνική κλαδέματος (pruning), έτσι ώστε να μειώσουμε το μέγεθος του δέντρου χωρίς να μειώσουμε την ακρίβεια της πρόβλεψής του.
- Μπορεί να παρουσιάσει αστάθεια στην επίδοσή του, γιατί μικρές διακυμάνσεις στο σύνολο δεδομένων μπορούν να έχουν ως αποτέλεσμα τη δημιουργία ενός τελείως διαφορετικού δέντρου απόφασης.
- Δεν είναι δυνατή η επέκταση του αλγορίθμου.

2.3.5 Τυχαία Δάση

Τα τυχαία δάση (Random Forests) είναι ένας εύχρηστος αλγόριθμος εποπτευόμενης μάθησης που χρησιμοποιεί την bagging τεχνική (bootstrap aggregating), μία τεχνική συνολικής μάθησης (ensemble learning) για την διεργασία της παλινδρόμησης. Ένας αλγόριθμος τυχαίων δασών λειτουργεί κατασκευάζοντας πολλαπλά δέντρα απόφασης (Decision Trees) κατά τη διάρκεια εκπαίδευσης και παράγει ως έξοδο τον μέσο όρο των προβλέψεων των δέντρων συνολικά. Στο Σχήμα 2.4 παρουσιάζεται η αρχιτεκτονική του αλγόριθμου Random Forest. Για την υλοποίηση ενός μοντέλου Random Forest, επιλέγουμε το σύνολο των δέντρων (n estimators) που θέλουμε να δημιουργηθούν. Το κάθε δέντρο έχει ως σύνολο ένα τυχαίο δείγμα απο το γενικό σύνολο εκπαίδευσης (training set) και λειτουργεί παράλληλα και ανεξάρτητα

με τα υπόλοιπα δέντρα. Στη συνέχεια, τα δέντρα εκπαιδεύονται και το κάθε δέντρο δίνει ξεχωριστές προβλέψεις για το τυχαίο σύνολο δεδομένων που έχει εκπαιδευτεί. Τέλος, ως έξοδος δίνεται ο μέσος όρος των προβλέψεων από όλα τα δέντρα, όπου είναι και η τελική πρόβλεψη του μοντέλου.



Σχήμα 2.4: Παράδειγμα αλγόριθμου Random Forest [Wood, 2021].

Πλεονεκτήματα του αλγόριθμου [Cutler et al., 2012]

- Είναι πολύ σταθερός στις προβλέψεις. Ακόμα και αν ένα νέο υποσύνολο δεδομένων προστεθεί στο γενικό σύνολο δεδομένων, ο αλγόριθμος δε θα επηρεαστεί σε μεγάλο βαθμό καθώς μπορεί τα νέα δεδομένα να επηρεάσουν ένα δέντρο, αλλά είναι δύσκολο να επηρεάσουν όλα τα δέντρα.
- Είναι αποτελεσματικός όταν το σύνολο δεδομένων είναι υπερβολικά μεγάλο.
- Έχει ακριβείς προβλέψεις ακόμα και όταν υπάρχει μεγάλος αριθμός δεδομένων που λείπουν.
- Καθώς βασίζεται στα δέντρα απόφασης (Decision Trees) έχει όλα τα πλεονεκτήματα που προσφέρουν αυτά.

Μειονεκτήματα του αλγόριθμου [Cutler et al., 2012]

- Μπορεί να παρατηρηθεί υπερεκπαίδευση σε μερικά σύνολα δεδομένων.

- Απαιτεί περισσότερο χρόνο εκπαίδευσης σε σχέση με τα δέντρα απόφασης, καθώς δημιουργεί πολλά από αυτά.

2.3.6 Μηχανές Διανυσμάτων Υποστήριξης

Η μηχανή διανυσμάτων υποστήριξης (support vector machine) είναι ένας αλγόριθμος εποπτευόμενης μηχανικής μάθησης ο οποίος χρησιμοποιείται κυρίως σε εργασίες κατηγοριοποίησης αλλά και παλινδρόμησης. Σε μία εργασία κατηγοριοποίησης (Support Vector Classification - SVC) μεταξύ δύο κλάσεων, στόχος είναι να βρεθεί ένα υπερεπίπεδο (hyperplane) που θα δημιουργεί ένα όριο απόφασης (decision boundary) που θα διαχωρίζει τα δεδομένα των δύο κατηγοριών σε δύο ημιεπίπεδα διατηρώντας μία μέγιστη απόσταση (maximum margin) μεταξύ των κατηγοριών. Σκοπός είναι η εύρεση του υπερεπιπέδου που διαχωρίζει καλύτερα τα δεδομένα των δύο κατηγοριών. Τα σημεία δεδομένων (data points) που βρίσκονται στο όριο της μέγιστης απόστασης (maximum margin) του υπερπλάνου που διαχωρίζει τις δύο κατηγορίες δεδομένων, ονομάζονται διανύσματα υποστήριξης (support vectors).

Στην περίπτωση της παλινδρόμησης (Support Vector Regression - SVR), η μέθοδος της παλινδρόμησης χρησιμοποιεί τις ίδιες αρχές και χαρακτηριστικά με τη μέθοδο της κατηγοριοποίησης με διανύσματα υποστήριξης, με κάποιες διαφορές. Αρχικά, δημιουργείται ένα υπερπλάνο και στη συνέχεια δύο γραμμές ορίων (boundary lines). Η απόσταση του υπερπλάνου από κάθε γραμμή ορίου είναι ίση με ϵ.

Η εξίσωση του υπερπλάνου είναι:

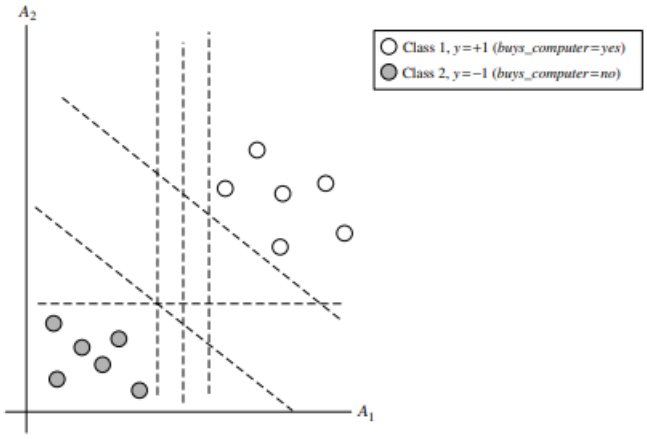
$$wx + b = 0$$

και η εξίσωση των γραμμών ορίου είναι:

$$wx + b = \pm\epsilon$$

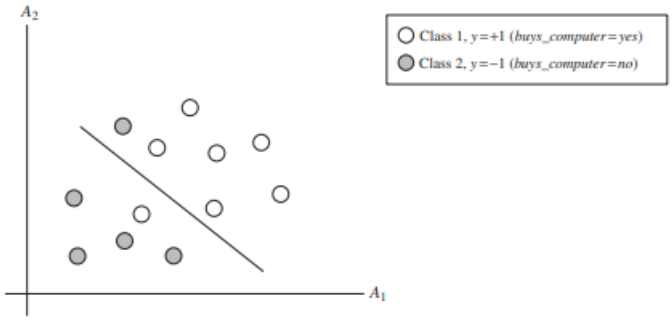
Έτσι, η εξίσωση που ικανοποιεί τον αλγόριθμο παλινδρόμησης με μηχανές διανυσμάτων υποστήριξης είναι η εξής:

$$\epsilon \leq y - wx - b \leq +\epsilon$$



Σχήμα 2.5: Παράδειγμα γραμμικά διαχωρίσιμου SVM δύο διαστάσεων. Με διακεκομμένες γραμμές παρουσιάζονται κάποια από τα άπειρα υπερεπίπεδα ή όρια απόφασης που υπάρχουν [Han et al., 2012].

Στο Σχήμα 2.5 φαίνονται οι SVM για ένα πρόβλημα κατηγοριοποίησης δύο κλάσεων που μπορεί να διαχωριστεί γραμμικά. Όταν μπορούμε να διαχωρίσουμε τα δεδομένα με μία ευθεία γραμμή, τότε το μοντέλο μηχανών διανυσμάτων υποστήριξης είναι γραμμικό (linear SVM). Αντιθέτως, όταν δεν είναι εφικτό να διαχωρίσουμε τα δεδομένα με μία ευθεία γραμμή, χρησιμοποιούμε το μη-γραμμικό μοντέλο (non-linear SVM) όπως φαίνεται στο Σχήμα 2.6. Σε αυτές τις περιπτώσεις χρησιμοποιούμε τις συναρτήσεις πυρήνα (kernel functions), οι οποίες μετατρέπουν τους μη-γραμμικούς χώρους σε γραμμικούς. Μετατρέπουν τα δεδομένα σε μία άλλη διάσταση έτσι ώστε τα δεδομένα να μπορούν να κατηγοριοποιηθούν. Οι πιο γνωστές συναρτήσεις πυρήνα είναι οι πολυωνυμικές, οι σιγμοειδείς και η radial basis function (RBF).



Σχήμα 2.6: Παράδειγμα μη διαχωρίσιμου γραμμικά SVM δύο διαστάσεων όπου το όριο απόφασης που προκύπτει είναι μη γραμμικό [Han et al., 2012].

Πλεονεκτήματα του αλγόριθμου [Pan et al., 2008]

- Είναι αποτελεσματικός σε χώρους πολλών διαστάσεων.

- Δυνατότητα ενημέρωσης και βελτίωσης του μοντέλου απόφασης.
- Διαφορετικές συναρτήσεις πυρήνα μπορούν να χρησιμοποιηθούν για διαφορετικές συναρτήσεις αποφάσεων.
- Δυνατότητα πρόσθεσης των συναρτήσεων πυρήνα για τη δημιουργία καλύτερων συναρτήσεων πυρήνα.

Μειονεκτήματα του αλγόριθμου [Pan et al., 2008]

- Χαμηλή απόδοση όταν ο αριθμός των χαρακτηριστικών είναι μεγαλύτερος από τον αριθμό των δεδομένων.
- Είναι υπολογιστικά ακριβός.

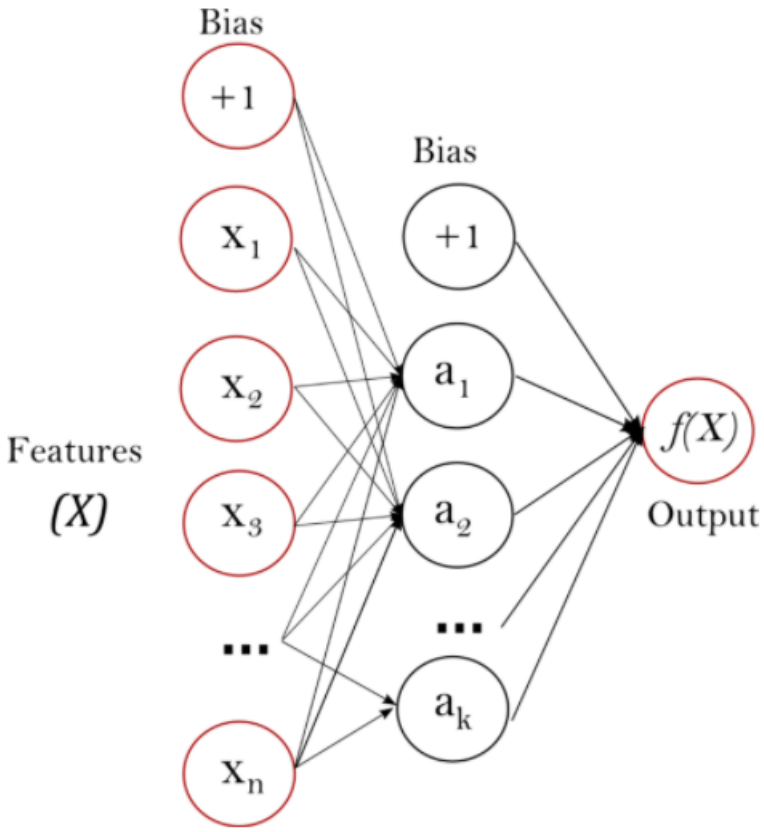
2.3.7 Multi-layer Perceptron

Ο Multi-layer Perceptron (MLP) είναι ένας αλγόριθμος εποπτευόμενης μάθησης ο οποίος μαθαίνει μία συνάρτηση $f(\cdot) : R^m \rightarrow R^o$ μέσω της εκπαίδευσης στο σύνολο δεδομένων, όπου το m είναι η διάσταση των δεδομένων εισόδου και το o είναι η διάσταση των δεδομένων εξόδου. Δίνοντας ένα σύνολο χαρακτηριστικών $X = x_1, x_2, ..., x_m$ και ένα στόχο y, μπορεί να μάθει μία μη-γραμμική προσεγγιστική συνάρτηση για προβλήματα παλινδρόμησης και κατηγοριοποίησης. Ανάμεσα στα δύο επίπεδα εισόδου και εξόδου του αλγορίθμου MLP, μπορούν αν υπάρξουν ένα ή περισσότερα μη-γραμμικά κρυφά επίπεδα (hidden layers).

Το επίπεδο εισόδου (στην αριστερή πλευρά του Σχήματος 2.7) αποτελείται από μία λίστα νευρώνων $X_i|x_1, x_2, ..., x_m$ τα οποία αντιπροσωπεύουν τα χαρακτηριστικά εισόδου. Κάθε νευρώνας στο κρυφό επίπεδο μετασχηματίζει την τιμή του προηγούμενου επιπέδου με ένα βάρος το οποίο προκύπτει από το γραμμικό άθροισμα $w_1x_1 + w_2x_2 + ... + w_mx_m$ ακολουθούμενο από μία συνάρτηση ενεργοποίησης $g(\cdot) : R \rightarrow R-$ όπως αυτό της υπερβολικής εφαπτομένης. Το επίπεδο εξόδου (στη δεξιά πλευρά του Σχήματος 2.7) λαμβάνει τις τιμές από το τελευταίο κρυφό επίπεδο και τις μετασχηματίζει στην κατάλληλη μορφή.

Πλεονεκτήματα του αλγόριθμου [Scikit-learn, 2021]

- Δυνατότητα να μάθει μη γραμμικά μοντέλα.



Σχήμα 2.7: Παράδειγμα αλγόριθμου MLP με ένα κρυφό επίπεδο [Scikit-learn, 2021].

- Δυνατότητα να μάθει μοντέλα σε πραγματικό χρόνο χρησιμοποιώντας μερική εφαρμογή (partial_fit).
- Μπορεί να εκτελεί πολλές διεργασίες παράλληλα χωρίς να επηρεάζει την απόδοση του συστήματος.

Μειονεκτήματα του αλγόριθμου [Scikit-learn, 2021]

- Είναι πολύ ευαίσθητος στην κλιμάκωση χαρακτηριστικών (feature scaling).
- Τα κρυφά επίπεδα έχουν μη κυρτή συνάρτηση απωλειών όπου μπορεί να υπάρχουν περισσότερα από ένα τοπικά ελάχιστα. Έτσι, η αρχικοποίηση διαφορετικών τυχαίων βαρών μπορεί να οδηγήσει σε διαφορετική ακρίβεια επικύρωσης.
- Χρειάζεται η ρύθμιση των υπερπαραμέτρων του MLP όπως είναι ο αριθμός των νευρώνων, των επιπέδων καθώς και οι επαναλήψεις.

2.4 Μοντέλα Στατιστικής Ανάλυσης

2.4.1 Εισαγωγή

Εκτός από τους αλγόριθμους μηχανικής μάθησης, για την ανάλυση δεδομένων συνήθως χρησιμοποιούμε μοντέλα στατιστικής. Στατιστικά μοντέλα είναι τα μαθηματικά μοντέλα που ενσωματώνουν ένα σύνολο στατιστικών υποθέσεων σχετικά με τη δημιουργία και την ανάλυση δειγμάτων δεδομένων. Ο σκοπός ενός μοντέλου στατιστικής είναι να παρατηρήσει διάφορα μοτίβα στο σύνολο των δεδομένων και να εξαγάγει συμπεράσματα τα οποία στη συνέχεια θα αξιολογηθούν. Χρησιμοποιούνται κύριως για πρόβλεψη μελλοντικών τιμών όπως η πρόβλεψη του καιρού ή για την πρόβλεψη των τιμών κάποιας μετοχής.

2.4.2 Μοντέλο ARIMA

Το μοντέλο ARIMA [Jain and Mallick, 2017] (AutoRegressive Integrated Moving Average) ανήκει στην κατηγορία των στατιστικών μοντέλων για την ανάλυση και πρόβλεψη δεδομένων χρονοσειρών. Είναι ένα αλγοριθμικό μοντέλο που χρησιμοποιεί την εξάρτηση μεταξύ παρατήρησης και υπολλειματικών σφαλμάτων (residual errors) από ένα μοντέλο κινητού μέσου όρου (moving average) που εφαρμόζεται σε παρατηρήσεις με καθυστέρηση (lagged values). Βασίζεται στην ιδέα ότι οι πληροφορίες των προηγούμενων παρατηρήσεων των δεδομένων χρονοσειρών μπορούν να χρησιμοποιηθούν για την πρόβλεψη των μελλοντικών παρατηρήσεων.

Ένα μοντέλο ARIMA χαρακτηρίζεται από τρεις όρους p, d, q όπου το p είναι η σειρά του όρου AutoRegressive (AR), το q είναι η σειρά του όρου Moving Average (MA) και το d είναι ο αριθμός των διαφορών (differencing) που απαιτούνται για να είναι οι χρονοσειρές σταθερές (stationary).

Όρος της διαφοράς

Ο όρος AR στο μοντέλο ARIMA σημαίνει ότι είναι ένα γραμμικό μοντέλο παλινδρόμησης (linear regression model). Καθώς τα γραμμικά μοντέλα παλινδρόμησης λειτουργούν καλύτερα όταν οι προγνωστικοί παράγοντες δε συσχετίζονται και είναι ανεξάρτητοι μεταξύ τους, είναι απαραίτητο να κάνουμε τα δεδομένα χρονοσειρών σταθερά. Μία σταθερή χρονοσειρά είναι αυτή της οποίας οι στατιστικές ιδιότητες όπως η μέση τιμή (mean), η διακύμανση (variance) και η αυτοσυσχέτιση

(autocorrelation) δεν εξαρτώνται από το χρόνο στον οποίο παρατηρείται η σειρά. Αυτό συνήθως γίνεται με την τεχνική της διαφοράς, δηλαδή αφαιρώντας την προηγούμενη τιμή από την τωρινή τιμή. Η τιμή του όρου d, είναι ο ελάχιστος αριθμός διαφοράς που απαιτείται για να γίνει η χρονοσειρά σταθερή.

AR Μοντέλο

Ένα AR μοντέλο χρονοσειρών, χρησιμοποιεί παρατηρήσεις από προηγούμενα χρονικά βήματα (lags) ως είσοδο σε μία εξίσωση παλινδρόμησης για να προβλέψει την τιμή y_t στο επόμενο βήμα. Η εξίσωση είναι η εξής:

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + ... + \beta_p y_{t-p} + \epsilon_1$$

όπου y_{t-1} είναι το πρώτο προηγούμενο βήμα (lag1) της σειράς, β_1 είναι ο συντελεστής (coefficient) που εκτιμά το μοντέλο και το α είναι ο όρος αναχαίτισης (intercept) που εκτιμάται επίσης από το μοντέλο.

Moving Average Μοντέλο

Ομοίως, ένα Moving Average (MA) μοντέλο είναι αυτό του οποίου η προβλεπόμενη τιμή y_t εξαρτάται μόνο από τα προηγούμενα σφάλματα πρόβλεψης (lagged forecast errors).

$$y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + ... + \phi_q \epsilon_{t-q}$$

όπου οι όροι σφάλματος ϵ είναι τα σφάλματα του AR μοντέλου των αντίστοιχων προηγούμενων βημάτων [Jain and Mallick, 2017].

Κεφάλαιο 3

Σχετική Βιβλιογραφία

3.1 Εισαγωγή

Στις μέρες μας, η χρήση αισθητήρων για τη συλλογή δεδομένων και στη συνέχεια η εφαρμογή της μηχανικής μάθησης για την πρόβλεψη διάφορων μετρήσεων, όπως η θερμοκρασία, η υγρασία, η κατανάλωση ενέργειας, αποτελεί σημαντικό και απαραίτητο κομμάτι για πολλούς τομείς, όπως η γεωργία, οι υπηρεσίες εναέριας κυκλοφορίας, το σύστημα υγείας, ο έλεγχος της ενέργειας και του περιβάλλοντος. Καθώς το κόστος των εξαρτημάτων αισθητήρων μειώνεται συνεχώς, ο αριθμός των δεδομένων που συλλέγονται για ανάλυση αυξάνεται ακόμη περισσότερο. Σύγχρονες μελέτες έχουν δείξει ότι η στατιστική ανάλυση και μοντελοποίηση με βάση τη μηχανική μάθηση, μπορεί να προσφέρει πολλές πληροφορίες και απαντήσεις στα ερωτήματα πρόβλεψης που δημιουργούνται. Παρακάτω, παρουσιάζονται μελέτες στις οποίες έχουν εφαρμοσθεί αλγόριθμοι μηχανικής μάθησης σε δεδομένα που προέρχονται από αισθητήρες για την ανάλυση και τη βελτίωση των προβλέψεων διάφορων μετρήσεων.

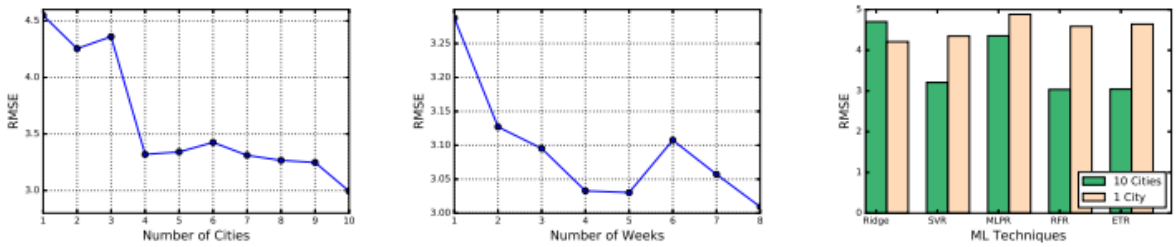
3.2 Μελέτη για την πρόβλεψη του καιρού στο Tennessee.

Οι Jakaria et al. [Jakaria et al., 2020] κάνουν ανάλυση και πρόγνωση του καιρού της πόλης Nashville του Tenessee χρησιμοποιώντας πέντε αλγόριθμους μηχανικής μάθησης. Οι μέθοδοι που χρησιμοποιούν έχουν ως είσοδο ιστορικά δεδομένα από 10 πόλεις και η μεταβλητή στόχος είναι η ωριαία θερμοκρασία της επόμενης μέρας στην πόλη Nashville. Αρχικά, συλλέγουν τα δεδομένα του καιρού, όπως θερμοκρασία, υγρασία, βροχόπτωση και ταχύτητα του ανέμου, για την πόλη Nashville αλλά

και για άλλες εννέα κοντινές πόλεις. Στη συνέχεια επεξεργάζονται τα ακατέργαστα δεδομένα, συνδυάζοντας όλα τα καιρικά δεδομένα των πόλεων σε μία συγκεκριμένη χρονική στιγμή και δημιουργώντας έτσι νέα σημεία δεδομένων. Αφού έχουν διασφαλίσει ότι κάθε γραμμή περιέχει δεδομένα από όλες τις πόλεις, αφαιρούν όσα χαρακτηριστικά έχουν κενά ή μη έγκυρα δεδομένα. Ακόμη, μετατρέπουν όσα χαρακτηριστικά έχουν κατηγορικά δεδομένα σε τιμές dummy μεταβλητών, δηλαδή σε τιμές 0 ή 1. Τέλος, στο σύνολο των δεδομένων που έχει προκύψει πραγματοποιούν κλιμάκωση των χαρακτηριστικών και χωρίζουν το σύνολο δεδομένων σε σετ εκπαίδευσης και σετ δοκιμών. Εφαρμόζοντας τους αλγόριθμους μηχανικής μάθησης στην αρχή μόνο για την πόλη Nashville και στη συνέχεια προσθέτοντας τις υπόλοιπες πόλεις, υπολογίζουν τη ρίζα του μέσου τετραγώνου σφάλματος (Root Mean Squared Error - RMSE) για να αξιολογήσουν τα αποτελέσματα.

Εφαρμόζοντας τους αλγόριθμους μηχανικής μάθησης, παρατηρούν ότι προσθέτοντας τα δεδομένα μίας ακόμη πόλης, η ακρίβεια του μοντέλου αυξάνεται. Προσθέτοντας όμως περισσότερες πόλεις, η ακρίβεια του μοντέλου μειώνεται ελαφρώς. Επίσης, παρατηρούν ότι όσο αυξάνονται οι εβδομάδες δεδομένων που παρέχουν στο σετ εκπαίδευσης, τα μοντέλα παρουσιάζουν μεγαλύτερη ακρίβεια στις προβλέψεις. Στο Σχήμα 3.1, φαίνεται η επίδραση που έχει ο αριθμός των πόλεων και ο αριθμός των εβδομάδων στο RMSE των μοντέλων, καθώς και το RMSE για τα πέντε μοντέλα μηχανικής μάθησης που χρησιμοποιήθηκαν. Η πρώτη τεχνική μηχανικής μάθησης που χρησιμοποίησαν είναι η παλινδρόμηση Ridge, όπου έδειξε σχετικά υψηλό RMSE πάνω από 4 και στην περίπτωση της εκπαίδευσης με δεδομένα μίας πόλης αλλά και με δεδομένα 10 πόλεων. Στη συνέχεια χρησιμοποίησαν την παλινδρόμηση με μηχανές διανυσμάτων υποστήριξης όπου έδειξε να έχει πολύ καλύτερη ακρίβεια, με RMSE περίπου 3. Αντίθετα, στην περίπτωση των δεδομένων από 10 πόλεις σε σχέση με τα δεδομένα μίας πόλης το RMSE ήταν περίπου 4.5. Έπειτα, πραγματοποίησαν παλινδρόμηση με MLP που αποτελείται από δύο κρυφά επίπεδα, με 100 βάρη στο πρώτο κρυφό επίπεδο και 50 βάρη στο δεύτερο κρυφό επίπεδο. Η παλινδρόμηση με MLP έδειξε υψηλό RMSE, περισσότερο από 4.0 και στις δύο περιπτώσεις δεδομένων.

Τέλος, χρησιμοποιήθηκαν οι αλγόριθμοι παλινδρόμησης Random Forest και Extra Tree Regressor όπου είχανε το χαμηλότερο RMSE περίπου 3 για την περίπτωση



Σχήμα 3.1: (α) Το RMSE προσθέτοντας γειτονικές πόλεις, (β) Το RMSE όταν αυξηθούν τα δεδομένα στο σετ εκπαίδευσης, (γ) Το RMSE για όλα τα μοντέλα μηχανικής μάθησης [Jakaria et al., 2020].

των δεδομένων 10 πόλεων και συνεπώς τη μεγαλύτερη ακρίβεια στις προβλέψεις. Αντιθέτως, είχανε αρκετά υψηλό RMSE, άνω του 4 στα δεδομένα μίας πόλης, γεγονός που υποδηλώνει την ανάγκη να μελετηθούν τα δεδομένα των γειτονικών πόλεων για την πρόβλεψη σε μία περιοχή.

3.3 Εφαρμογή αλγόριθμων μηχανικής μάθησης σε δεδομένα αισθητήρων για συστάσεις άρδευσης.

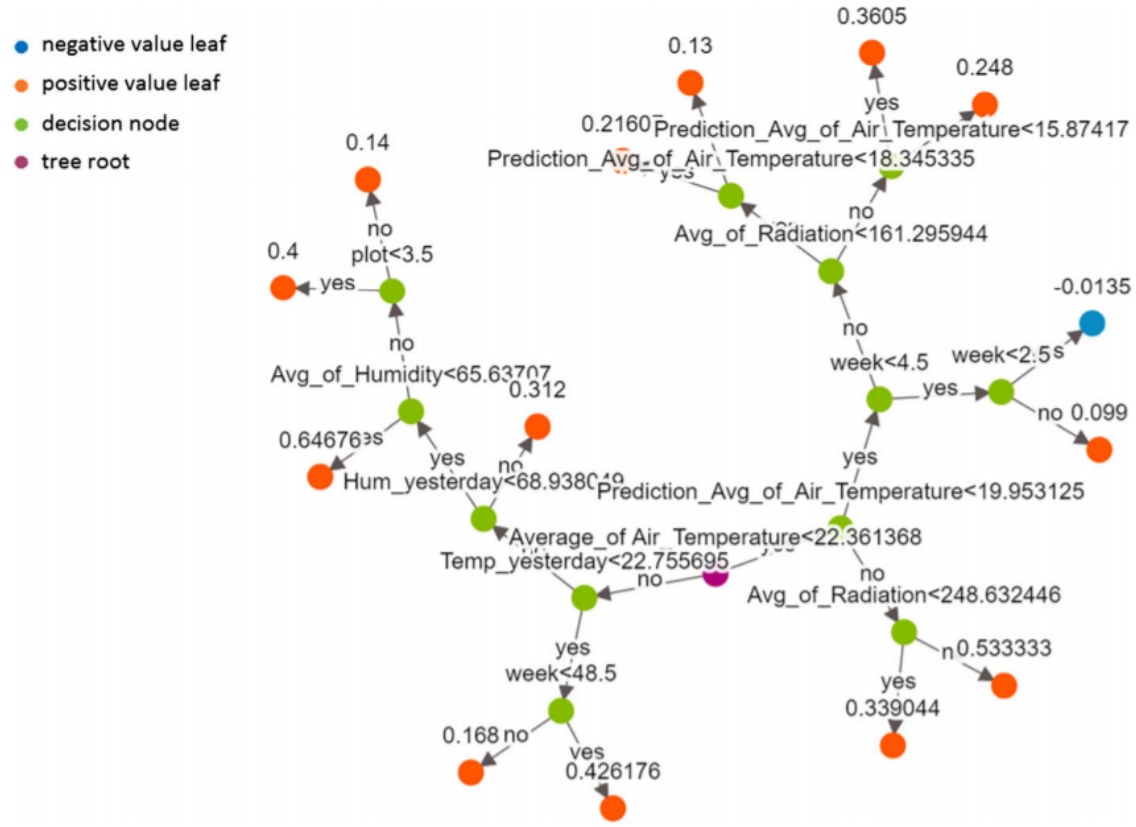
Οι Goldstein et al. [Goldstein et al., 2018] χρησιμοποιήσαν αλγόριθμους μηχανικής μάθησης για την πρόβλεψη και σύσταση των αποφάσεων άρδευσης ενός αγρονόμου με σκοπό την αυτοματοποίηση μελλοντικών αποφάσεων άρδευσης, αλλά και τη χρήση τους ως συμβουλευτικό μέσο για τη βελτιστοποίηση του σχεδιασμού άρδευσης στο μέλλον.

Αρχικά χρησιμοποιήθηκαν δεδομένα από διάφορους αισθητήρες που συλλέγουν την υγρασία εδάφους σε διαφορετικό εδαφικό βάθος, καθώς και δεδομένα από μετεωρολογικούς σταθμούς, όπως θερμοκρασία, υγρασία και ηλιακή ακτινοβολία. Για τον καθαρισμό και τις μετατροπές των δεδομένων που συλλέχθηκαν, οι ερευνητές συμβουλεύτηκαν τον αγρονόμο, ο οποίος έβαλε κάποια όρια στις τιμές των δεδομένων των αισθητήρων. Στην περίπτωση που οι τιμές ήταν λανθασμένες θα έπρεπε να φιλτραριστούν. Έτσι, ο αγρονόμος όρισε τις τιμές υγρασίας εδάφους μικρότερες της τιμής 17 ή μεγαλύτερες του 39% ως ασυνήθιστες και συνεπώς αφαιρέθηκαν από το σύνολο δεδομένων και αντικαταστάθηκαν από τη μέση τιμή των δύο γειτονικών αισθητήρων. Ακόμη, τα δεδομένα καιρικών φαινομένων από τον μετεωρολογικό σταθμό χρησιμοποιήθηκαν για την πρόβλεψη της μέσης εβδομαδιαίας θερμοκρασίας, υγρασίας και ηλιακής ακτινοβολίας την εβδομάδα που προετοιμάστηκε το πλάνο

άρδευσης.

Για την ανάπτυξη του μοντέλου πρόβλεψης και σύστασης άρδευσης επιλέχθηκαν οι εξής αλγόριθμοι μηχανικής μάθησης: γραμμική παλινδρόμηση, δέντρα παλινδρόμησης και ο αλγόριθμος δέντρων κατηγοριοποίησης Boosted Trees Classifier (BTC). Αρχικά, για την επιλογή των χαρακτηριστικών εκτιμήθηκε η συσχέτιση (correlation) μεταξύ των χαρακτηριστικών. Για παράδειγμα, τα δεδομένα αισθητήρων έδειξαν ότι έχουν χαμηλή συσχέτιση περίπου 0.1 σε σχέση με τα μετεωρολογικά δεδομένα. Έτσι, οι ερευνητές δημιούργησαν τρεις προαιρετικές κατηγορίες δεδομένων για να ομαδοποιήσουν τα δεδομένα και συνεπώς ορίστηκαν 8 υποσύνολα δεδομένων. Στη γραμμική παλινδρόμηση, το ποσοστό επιτυχίας και για τα 8 σύνολα δεδομένων ήταν παρόμοιο, με χαμηλότερο RMSE αυτό του δεύτερου υποσυνόλου δεδομένων. Το υψηλότερο συνολικά ποσοστό επιτυχίας είχε το έκτο υποσύνολο δεδομένων, αν και το 52.3% θεωρείται σχετικά χαμηλό. Για τη βελτιστοποίηση των αλγόριθμων γραμμικής παλινδρόμησης χρησιμοποιήθηκε η παλινδρόμηση Lasso η οποία τελικά δε βελτίωσε την επίδοση των αλγόριθμων όσον αφορά το RMSE ή το συνολικό ποσοστό επιτυχίας. Αντίστοιχα, αναπτύχθηκαν 8 μοντέλα gradient boost δέντρων απόφασης (GBRT) για τα 8 υποσύνολα δεδομένων όπου κάθε μοντέλο περιελάμβανε μία σειρά από δέντρα απόφασης. Στο Σχήμα 3.2 φαίνεται η αρχιτεκτονική ενός δέντρου απόφασης του GBRT μοντέλου. Συγκριτικά με το μοντέλο γραμμικής παλινδρόμησης, παρουσίασε πολύ καλύτερο RMSE ίσο με 0.11 στο πρώτο υποσύνολο δεδομένων, ενώ το καλύτερο συνολικό ποσοστό επιτυχίας παρουσιάστηκε στο πέμπτο υποσύνολο με τιμή 92%.

Τέλος, χρησιμοποιήθηκε η τεχνική των boosted δέντρων κατηγοριοποίησης (boosted trees classifier) για διαφορετικά σετ χαρακτηριστικών. Ορίστηκαν 5 κατηγορίες, ανάλογες με τις 5 κατηγορίες άρδευσης. Τα αποτελέσματα δείχνουν ότι το πρώτο υποσύνολο το οποίο είναι βασισμένο σε ολόκληρο το σύνολο χαρακτηριστικών, είχε την καλύτερη απόδοση από τα υπόλοιπα μοντέλα σε όλες τις μετρήσεις απόδοσης. Συνοψίζοντας, οι αλγόριθμοι δέντρων απόφασης φαίνεται ότι μπορούν να προβλέψουν με ακρίβεια την απόφαση ενός αγρονόμου να χρησιμοποιήσει την άρδευση ή όχι. Στο Σχήμα 3.3 απεικονίζεται το ποσοστό επιτυχίας όλων των μοντέλων που χρησιμοποιήθηκαν. Παρόλο που τα μοντέλα με δέντρα κατηγοριοποίησης έχουν μεγαλύτερη ακρίβεια, τα μοντέλα με δέντρα παλινδρόμησης συστήνονται για την πρόβλεψη της

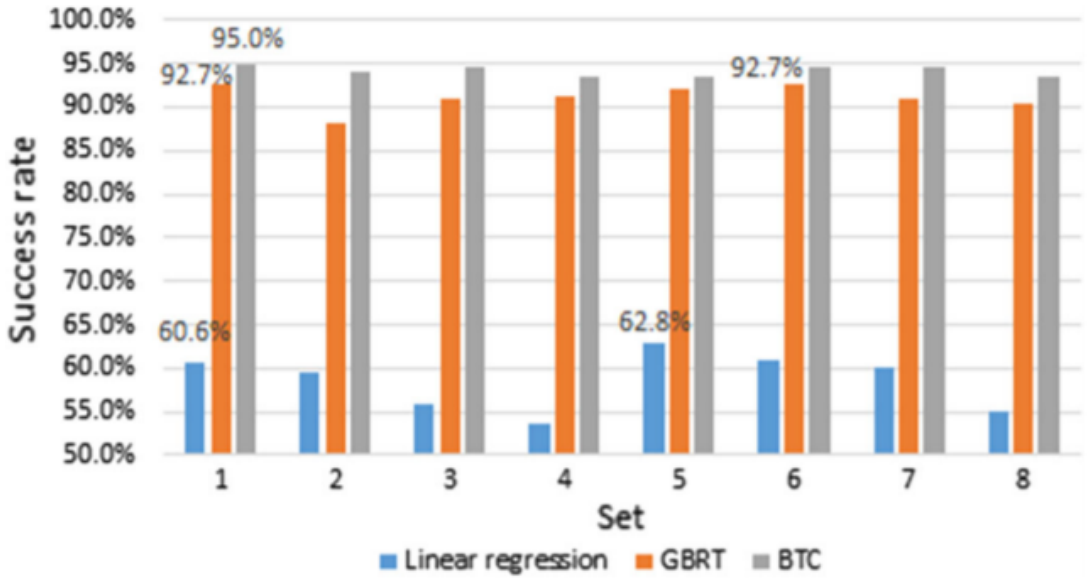


Σχήμα 3.2: Παράδειγμα ενός δέντρου απόφασης το οποίο ήταν μέρος του GBRT μοντέλου [Goldstein et al., 2018].

ποσότητας άρδευσης, καθώς μπορούν να προβλέψουν με ακριβείς τιμές.

3.4 Μελέτη για την πρόβλεψη της αστικής θερμοκρασίας της Ζυρίχης χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης.

Οι Zumwald et al. [Zumwald et al., 2021] εστιάζουν στο πως μπορούν να χρησιμοποιηθούν η μηχανική μάθηση και δεδομένα μετεωρολογικών σταθμών σε συνδυασμό με δεδομένα από ατομικούς αισθητήρες για τη δημιουργία και αξιολόγηση χαρτών υψηλής ανάλυσης της θερμοκρασίας του αέρα στην πόλη της Ζυρίχης. Για τη συλλογή των δεδομένων χρησιμοποίησαν 634 ατομικούς μετεωρολογικούς σταθμούς. Στα δεδομένα αυτά εφάρμοσαν στατιστικούς ελέγχους για να εξαλείψουν τυχόν λανθασμένες μετρήσεις. Αυτό είχε ως αποτέλεσμα στο να μειωθούν τα δεδομένα τους κατά 43%. Στη συνέχεια, χρησιμοποίησαν δεδομένα από 14 μετεωρολογικούς σταθμούς του Ομοσπονδιακού γραφείου Μετεωρολογίας και Κλιματολογίας της Ελβετίας. Τέλος, χρησιμοποιήθηκαν δεδομένα θερμοκρασίας από 43 αισθητήρες χαμη-



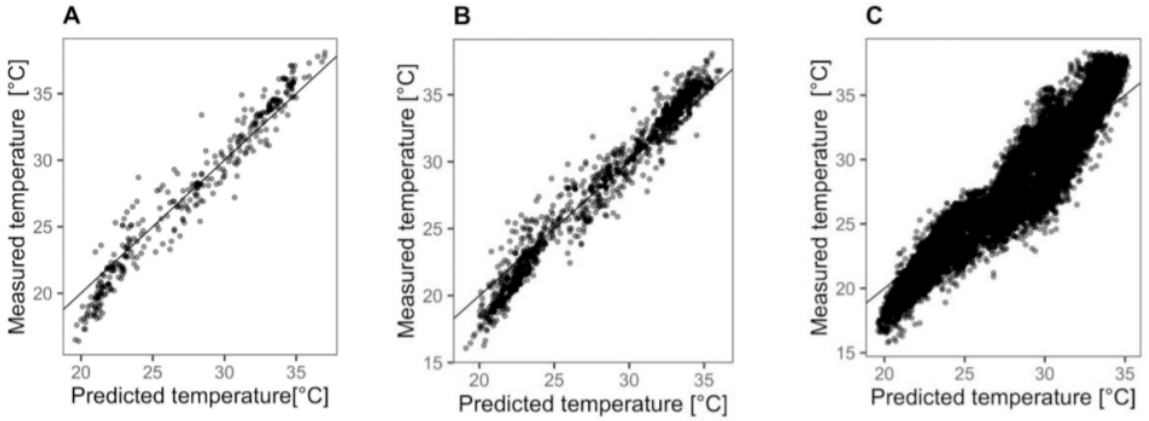
Σχήμα 3.3: Ποσοστό επιτυχίας για όλα τα μοντέλα που χρησιμοποιήθηκαν [Goldstein et al., 2018].

λού κόστους που χρησιμοποιούν το LoRaWAN δίκτυο. Το σύνολο δεδομένων εκπαίδευσης που δημιουργήθηκε αποτελούνταν από ωριαία δεδομένα θερμοκρασίας και από τους 691 αισθητήρες. Οι φυσικές αρχές που επηρεάζουν τη χωρική μεταβλητότητα της αστικής θερμοκρασίας, όπως η απόσταση μεταξύ των κτιρίων, το υλικό κατασκευής τους ή το υψόμετρο στο συγκεκριμένο σημείο της πόλης, έπρεπε να ληφθούν υπόψιν. Συνεπώς, σαν μεταβλητές εισόδου (predictor variables), ορίστηκαν το υψόμετρο και μία απλοποιημένη κατηγοριοποίηση 5 κλάσεων χρήσης της γης (κτίρια, σιδηρόδρομοι και δρόμοι, αστικοί χώροι πράσινου, επιφανειακά νερά και δάση). Επιπλέον, σαν συνεχής μεταβλητή εισόδου ορίστηκε ο όγκος του κτιρίου.

Για την πρόβλεψη της θερμοκρασίας σε αστικό περιβάλλον απαιτούνται μη γραμμικές προσεγγίσεις μοντελοποίησης. Χρησιμοποιήθηκε ο αλγόριθμος δάσους ποσοτικής παλινδρόμησης (quantile regression forest - QRF), μία επέκταση του αλγόριθμου τυχαίων δασών, ο οποίος χρησιμοποιεί τον αλγόριθμο τυχαίων δασών για να αντλήσει πληροφορίες από τους όρους ποσοτικά και μπορεί να χρησιμοποιηθεί για την πρόβλεψη αβεβαιότητας.

Η αξιολόγηση του μοντέλου έγινε με τη χρήση του RMSE ξεχωριστά για τα δεδομένα αισθητήρων αναφοράς, τα δεδομένα μετεωρολογικών σταθμών και τα δεδομένα ατομικών αισθητήρων, όπως φαίνεται στο Σχήμα 3.4. Το RMSE παρατηρήθηκε υψηλότερο κατά τις ώρες 03:00 έως 05:00 και χαμηλότερο τις ώρες 07:00

με 08:00. Οι προβλέψεις θερμοκρασίας είναι υπερεκτιμημένες τις βραδινές ώρες ξεκινώντας από τις 03:00 και υποτιμημένες ξεκινώντας από τις 07:00 το πρωί. Για την κατανόηση της συγκεκριμένης μοντελοποίησης και της ευαισθησίας του αριθμού αισθητήρων, το σύνολο δεδομένων χωρίστηκε σε 6 υποσύνολα με διαφορετικό αριθμό δεδομένων και αισθητήρων στο καθένα και έγινε σύγκριση της απόδοσης κάθε συνόλου με την απόδοση του γενικού συνόλου δεδομένων.



Σχήμα 3.4: Σύκριση μεταξύ των προβλεπόμενων και των δεδομένων μετρήσεων για α) τα δεδομένα αισθητήρων αναφοράς (RMSE=1.43), β) τα δεδομένα μετεωρολογικών σταθμών (RMSE=1.35) και γ) τα δεδομένα ατομικών αισθητήρων (RMSE=1.71) [Zumwald et al., 2021].

Η σημασία της κάθε μεταβλητής εισόδου (variable importance) στο μοντέλο, μας δίνει μία ένδειξη για το ποιες μεταβλητές ήταν οι πιο σημαντικές για την απόδοση του μοντέλου. Οι δύο πιο σημαντικές ήταν η θερμοκρασία του περιβάλλοντος αέρα στα 2 μέτρα κατά 33% και αυτή της θερμοκρασίας αέρα στα 5 εκατοστά από το έδαφος κατά 21%. Το καλύτερο RMSE παρουσίασε το γενικό σύνολο δεδομένων με τιμή 1.69, ενώ το πρώτο υποσύνολο παρουσίασε υψηλότερο RMSE ίσο με 1.83. Η μείωση της απόδοσης ήταν αισθητή στο δεύτερο υποσύνολο με το RMSE να φτάνει το 2.21. Συνεπώς, τα πειράματα επικύρωσης έδειξαν ότι με την αύξηση των δεδομένων, αυξήθηκε και η επίδοση, ενώ με την εξαίρεση κάποιων αισθητήρων, μειώθηκε αισθητά η απόδοση του μοντέλου.

3.5 Σύγκριση αλγόριθμων μηχανικής μάθησης για την πρόβλεψη της εσωτερικής θερμοκρασίας σε έξυπνα σπίτια.

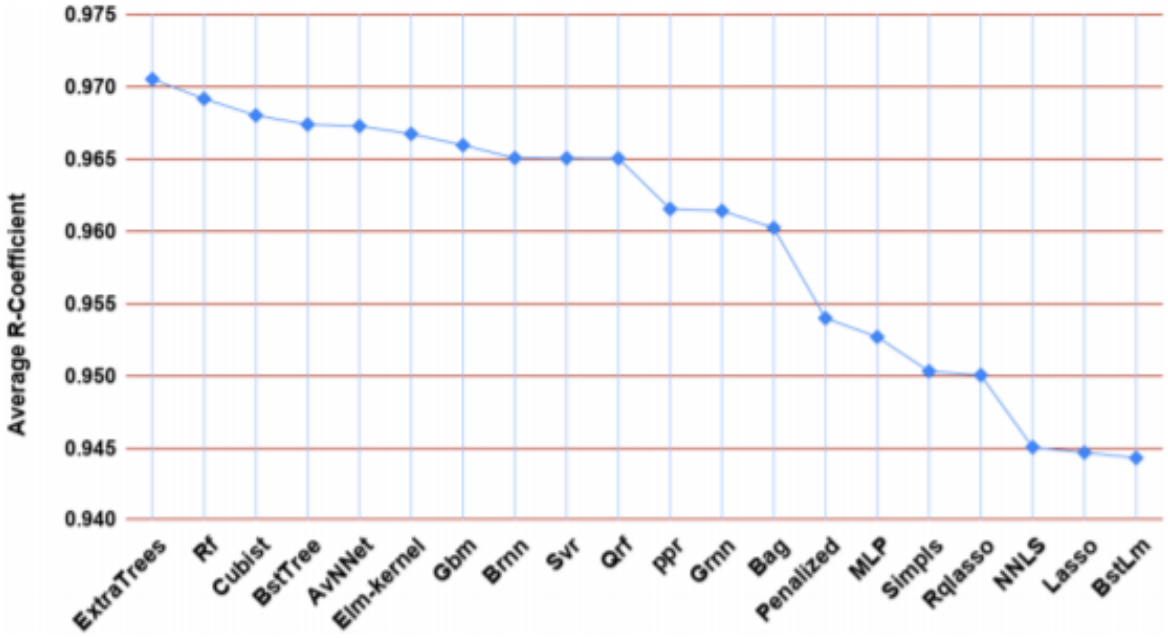
Στις μέρες μας, τα έξυπνα σπίτια αναλαμβάνουν αυτοματοποιημένα τον συντονισμό των συστημάτων θέρμανσης, εξαερισμού και ψύξης σύμφωνα με τις συνήθειες

των χρηστών με σκοπό τη βελτίωση της ικανοποίησης του χρήστη και την εξοικονόμησης ενέργειας. Οι Alawadi et al. [Alawadi et al., 2020] θέλοντας να πετύχουν μεγαλύτερη ακρίβεια στον αυτόματο συντονισμό των συστημάτων συγκρίνουν 36 αλγόριθμους μηχανικής μάθησης για την πρόβλεψη της εσωτερικής θερμοκρασίας ενός έξυπνου σπιτιού.

Τα πειράματα έγιναν σε ένα από τα 45 κτίρια του Πανεπιστημίου Santiago de Compostela (USC) στα οποία έχουν τοποθετηθεί αισθητήρες για την καταγραφή και διαχείριση των δεδομένων από συστήματα θέρμανσης, εξαερισμού και ψύξης. Το δίκτυο αισθητήρων στο συγκεκριμένο κτίριο, συλλέγει και αναφέρει 667 σήματα δεδομένων κάθε 10 δευτερόλεπτα. Ακόμη, χρησιμοποιήθηκαν δεδομένα τα οποία προέρχονται από τον πλησιέστερο μετεωρολογικό σταθμό. Τα χαρακτηριστικά του συνόλου δεδομένων που προέκυψε χωρίζονται σε δύο είδη, συνεχείς τιμές ή δυαδικές. Δεδομένα με συνεχείς τιμές είναι τα εξής: θερμοκρασία δαπέδου, θερμοκρασία του Air-Condition, υγρασία του Air-Condition και εσωτερική θερμοκρασία από τους αισθητήρες και θερμοκρασία, υγρασία, ακτινοβολία από τον μετεωρολογικό σταθμό. Τα δεδομένα με δυαδικές τιμές ήταν η κατάσταση του θερμαινόμενου δαπέδου (λειτουργία ON/OFF) και η κατάσταση του Air-Condition (λειτουργία ON/OFF).

Σε αυτή την έρευνα, έγινε σύγκριση συνολικά 36 αλγόριθμων μηχανικής μάθησης που ανήκουν σε 20 οικογένειες αλγορίθμων, με σκοπό να προσδιορίσουν ποιοι αλγόριθμοι είναι πιο ακριβείς στην πρόβλεψη της εσωτερικής θερμοκρασίας του κτιρίου που μελετήθηκε. Τα πειράματα για κάθε αλγόριθμο έγιναν με 10 επαναλήψεις, χρησιμοποιώντας τυχαία σημεία του συνόλου δεδομένων και χωρίζοντας το σύνολο σε 70% για εκπαίδευση, 15% για επικύρωση και 15% για δοκιμές. Επίσης, σε κάθε αλγόριθμο έγινε ρύθμιση των υπερπαραμέτρων και επιλέχθηκαν οι υπερπαράμετροι που είχαν την καλύτερη συνολική απόδοση. Για την αξιολόγηση των αλγόριθμων χρησιμοποιήθηκε η συσχέτιση Pearson (R-coefficient) και το RMSE. Οι αλγόριθμοι που χρησιμοποιήθηκαν ανήκουν στις εξής οικογένειες αλγορίθμων: Bayesian models, Bagging ensembles, Boosting ensembles, Gaussian processes, Generalized additive models, Generalized linear regression, Gradient boosting machines, Independent Component Analysis, Least absolute shrinkage, Least squares, Linear regression, Neural Networks, Partial least squares, Prototype models, Quantile regression, Random Forests, Regression Trees, Ridge regression, Support Vector Regression.

Στο Σχήμα 3.5 παρουσιάζονται οι 20 καλύτεροι αλγόριθμοι παλινδρόμησης με βάση τη μέση τιμή του R-coefficient. Ο αλγόριθμος που είχε τις καλύτερες συνολικές επιδόσεις ήταν ο αλγόριθμος ExtraTrees που ανήκει στην οικογένεια των Random Forests αλγόριθμων. Τα αποτελέσματα του RMSE ήταν 0.05807 και του R-coefficient ήταν 0.97052. Στη συνέχεια έγινε το τεστ Friedman για 3 διαφορετικούς ορίζοντες πρόβλεψης (για την επόμενη μία έως τρεις ώρες πρόβλεψης), ένα μη παραμετρικό στατιστικό τεστ για την κατάταξη των αποτελεσμάτων των αλγορίθμων ως προς το RMSE και R-coefficient.



Σχήμα 3.5: Μέση τιμή του R-coefficient των 20 καλύτερων αλγόριθμων παλινδρόμησης για ορίζοντα πρόβλεψης τριών ωρών [Alawadi et al., 2020].

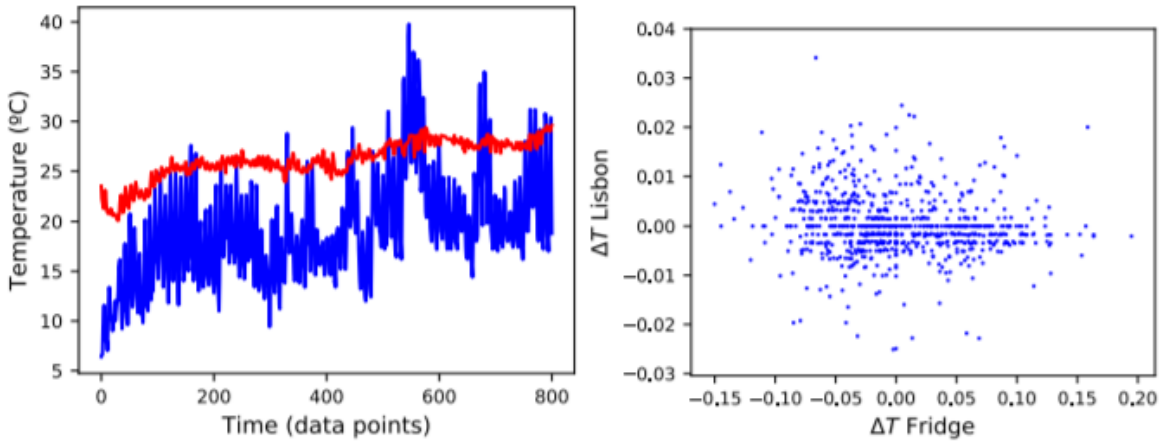
Τα αποτελέσματα αυτού του συγκριτικού πειράματος δείχνουν ότι ο αλγόριθμος ExtraTrees είχε τις καλύτερες επιδόσεις, καθώς φαίνεται να είναι λιγότερο ευαίσθητος στον θόρυβο των δεδομένων, ενώ τα μοντέλα νευρωνικών δικτύων δείχνουν πολύ πιο ευαίσθητα. Επίσης, η αύξηση του ορίζοντα πρόβλεψης έως και τρεις ώρες μετά δε φαίνεται να επηρεάζει την ακρίβεια της πρόβλεψης σχετικά με το RMSE και R-coefficient. Μερικοί επίσης καλοί αλγόριθμοι παλινδρόμησης είναι οι αλγόριθμοι: random forest, cubist, gradient boosting of regression trees (BstTree), average neural network committee (AvNNet) και kernel ELM (Elm-kernel).

3.6 Πρόβλεψη της εσωτερικής θερμοκρασίας σε ένα IoT σενάριο.

Οι Monteiro et al. [Monteiro et al., 2018] εξετάζουν πως μπορούν να χρησιμοποιηθούν οι αισθητήρες θερμοκρασίας στο εσωτερικό ενός σπιτιού και στη συνέχεια αλληλεπιδρώντας με τα υπάρχοντα μετεωρολογικά δεδομένα θερμοκρασίας να προβλέψουν την εσωτερική θερμοκρασία του σπιτιού, με τη βοήθεια αλγόριθμων μηχανικής μάθησης αλλά και στατιστικών μοντέλων. Το σύνολο δεδομένων αποτελείται από δύο υποσύνολα δεδομένων. Το πρώτο υποσύνολο περιλαμβάνει την εξέλιξη της εσωτερικής θερμοκρασίας που καταγράφεται από έναν εσωτερικό αισθητήρα ενός ψυγείου που βρίσκεται στο δωμάτιο του πανεπιστημίου UNINOVA της Λισαβόνας. Τα δεδομένα συλλέχθηκαν από το Μάρτιο μέχρι τον Ιούλιο του 2017, με 9 τιμές την ημέρα οι οποίες αφορούν τη μέση θερμοκρασία περίπου κάθε 9600 δευτερόλεπτα. Το δεύτερο υποσύνολο που συμπληρώνει το σύνολο δεδομένων, παρουσιάζει την αντίστοιχη εξωτερική θερμοκρασία στην πόλη της Λισαβόνας, όπως αυτή αποκτήθηκε από μετεωρολογικό ιστότοπο. Και οι δύο χρονοσειρές παρουσιάζονται στο Σχήμα 3.6. Στα αριστερά οι 801 παρατηρήσεις θερμοκρασιών μεταξύ $20.1°C$ έως $29.6°C$ για τον εσωτερικό χώρο, και μεταξύ $6.4°C$ και $39.8°C$ για τον εξωτερικό χώρο. Ακόμη, στα δεξιά του Σχήματος 3.6 φαίνεται ένα scatter plot που παρουσιάζει πως οι αλλαγές στη θερμοκρασία εσωτερικού χώρου ακολουθούν τη θερμοκρασία που παρατηρήθηκε στον εξωτερικό. Αυτές οι αλλαγές υπολογίζονται από τον εξής τύπο:

$$\Delta T_t = \log_{10} T_{t+1}/T_t$$

Οι τεχνικές που χρησιμοποιήθηκαν για την πρόβλεψη της θερμοκρασίας ήταν η γραμμική παλινδρόμηση και το στατιστικό μοντέλο ARIMA. Αυτές οι τεχνικές ανήκουν στην κατηγορία αφελούς μεθόδου πρόβλεψης (naive forecast method), όπως και τα μοντέλα που χρησιμοποιούν τον μέσο όρο της ημέρας ή την τελευταία καταγεγραμμένη τιμή για την πρόβλεψη και επικύρωση των τιμών θερμοκρασίας. Αυτές οι μέθοδοι πρόβλεψης επιλέχθηκαν γιατί είναι λιγότερο υπολογιστικά ακριβές, γεγονός που πρέπει να λαμβάνεται υπόψιν σε συσκευές IoT, σε σχέση με μεθόδους μηχανικής μάθησης, όπως οι μηχανές διανυσμάτων υποστήριξης ή τα νευρωνικά δίκτυα. Για τη γραμμική παλινδρόμηση οι ερευνητές εξέτασαν την πιθανότητα οι προηγούμενες και οι μελλοντικές τιμές θερμοκρασίας να σχετίζονται με μία γραμμική σχέση.

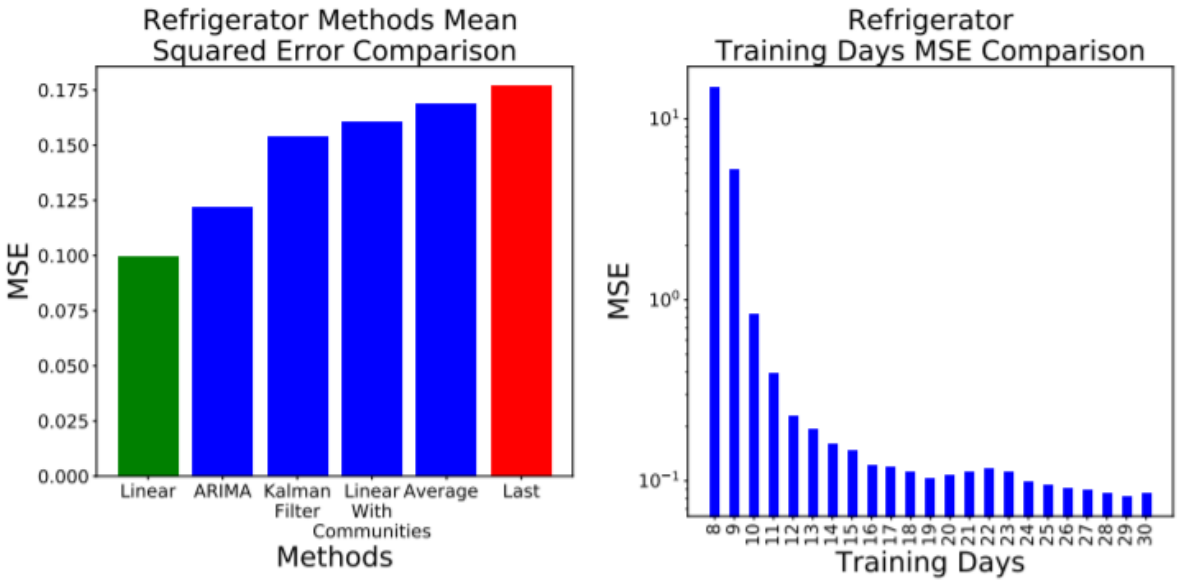


Σχήμα 3.6: **(Αριστερά)** Η εξέλιξη των δύο χρονοσειρών, όπου παρουσιάζεται η εσωτερική (κόκκινο) και η εξωτερική (μπλε) θερμοκρασία. **(Δεξιά)** Η εξέλιξη των αλλαγών της εξωτερικής θερμοκρασίας ως συνάρτηση της εσωτερικής θερμοκρασίας [Monteiro et al., 2018].

Καθώς καταγράφονται εννέα τιμές θερμοκρασίας την ημέρα, κατασκευάστηκε ένα μοντέλο όπου χρησιμοποιούσε τις οχτώ τελευταίες τιμές για την πρόβλεψη, με την τελευταία τιμή να είναι η τιμή στόχος (target value). Για τη βελτίωση του παραπάνω γραμμικού μοντέλου, έγινε φιλτράρισμα των δεδομένων μέσω των δικτύων. Δηλαδή, λαμβάνοντας υπόψη ότι κάποιες μέρες έχουν παρόμοια χαρακτηριστικά σχετικά με τη θερμοκρασία, δημιουργήθηκαν διάφορα προφίλ με βάση τη θερμοκρασία και στη συνέχεια ομαδοποιήθηκαν. Αυτό είχε ως αποτέλεσμα να υπάρχει μεγαλύτερη συσχέτιση στις τιμές της κάθε ομάδας και κάποιες μέρες να απορριφθούν από το σετ δεδομένων εκπαίδευσης καθώς δεν προσφέρουν γνώσεις στο μοντέλο, παρά μόνο θόρυβο.

Στη συνέχεια, χρησιμοποιήθηκε το μοντέλο στατιστικής ανάλυσης ARIMA αλλά και μία διαφορετική προσέγγιση για την πρόβλεψη τιμών σε δεδομένα χρονοσειρών, αυτή του φίλτρου Kalman. Το φίλτρο Kalman είναι ένας αλγόριθμος που χρησιμοποιεί μια σειρά μετρήσεων που παρατηρούνται με την πάροδο του χρόνου, που περιέχει στατιστικό θόρυβο και άλλες ανακρίβειες, και παράγει εκτιμήσεις άγνωστων μεταβλητών που τείνουν να είναι πιο ακριβείς από αυτές που βασίζονται σε μία μοναδική μέτρηση. Για την επικύρωση των μοντέλων χρησιμοποιήθηκε η συνάρτηση του MSE και τα αποτελέσματα παρουσιάζονται στο παρακάτω σχήμα.

Σύμφωνα με το Σχήμα 3.7, παρατηρήθηκε ότι την καλύτερη επίδοση είχε η γραμμική παλινδρόμηση με $MSE = 0.100$, ενώ τη χειρότερη η πρόβλεψη με βάση την προηγούμενη τιμή με $MSE = 0.175$. Επίσης, όπως φαίνεται στα δεξιά του Σχήματος



Σχήμα 3.7: **(Αριστερά)** Η τιμή MSE που παρατηρήθηκε για τις 6 μεθόδους πρόβλεψης. **(Δεξιά)** Η εξέλιξη του MSE για τη γραμμική παλινδρόμηση, ως συνάρτηση των ημερών που χρησιμοποιήθηκαν για τα δεδομένα εκπαίδευσης [Monteiro et al., 2018].

3.7, οι 30 ημέρες δεδομένων εκπαίδευσης είναι κάπως μεγάλος και συντηρητικός αριθμός, καθώς παρόμοιες επιδόσεις μπορούν να επιτευχθούν και στις 19 ημέρες δεδομένων εκπαίδευσης. Συνολικά, την καλύτερη απόδοση στις προβλέψεις αλλά και στις υπολογιστικές απαιτήσεις είχε η γραμμική παλινδρόμηση.

3.7 Εφαρμογή αλγόριθμων μηχανικής μάθησης σε δεδομένα αισθητήρων για την πρόβλεψη του αριθμού ατόμων σε ένα χώρο.

Οι Moraru et al. [Moraru et al., 2010] περιγράφουν ένα ολοκληρωμένο σύστημα ενός κόμβου αισθητήρα και τη χρήση αλγόριθμων μηχανικής μάθησης για την πρόβλεψη του αριθμού ατόμων που βρίσκονται σε ένα κλειστό χώρο. Ο κόμβος αισθητήρα καταγράφει και συλλέγει μετρήσεις δεδομένων θερμοκρασίας, υγρασίας, έντασης φωτός και πίεσης στο εργαστήριο των ερευνητών για ένα διάστημα 15 ημερών. Αυτές οι μετρήσεις εξαρτώνται άμεσα από την ανθρώπινη παρουσία στο εργαστήριο. Επιπλέον, συλλέξανε χειροκίνητα δεδομένα σχετικά με την ανθρώπινη παρουσία και άλλα γεγονότα που συμβαίνουν στο εργαστήριο. Το σύστημα κόμβου αισθητήρα λειτουργεί μαζεύοντας δεδομένα από τους αισθητήρες και στη συνέχεια μεταδίδει αυτά τα δεδομένα σε μία μηχανή αποθήκευσης. Στη συνέχεια, χρησι-

μοποιούνται μέθοδοι προεπεξεργασίας δεδομένων για την αυτόματη ενσωμάτωση των δεδομένων που συλλέγονται από αισθητήρες και δεδομένων που συλλέχθηκαν χειροκίνητα, για την εκπαίδευση σε αλγόριθμους μηχανικής μάθησης. Το σύνολο δεδομένων αποτελείται από τα εξής χαρακτηριστικά: θερμοκρασία, υγρασία, ένταση φωτός και πίεση. Επιπλέον, προστέθηκαν χειροκίνητα άλλα τρία χαρακτηριστικά: ο αριθμός ατόμων στο εργαστήριο, ο αριθμός των υπολογιστών που λειτουργούν και η θέση του παραθύρου. Το πρώτο στάδιο της προεπεξεργασίας δεδομένων ήταν η ευθυγράμμιση των δεδομένων αισθητήρων και των χειροκίνητα συλλεγόμενων δεδομένων σχετικά με τη χρονική στιμή (time stamp). Στο επόμενο βήμα, έγινε μείωση των δεδομένων με τη ρύθμιση του χρόνου συλλογής δεδομένων και αλλάζοντάς το από 10 δευτερόλεπτα σε ένα λεπτό. Ακόμη, αφαίρεσαν τα δεδομένα που συλλέγονταν την περίοδο της νύχτας (08:00 ΜΜ έως 06:00 ΠΜ) για να αποφύγουν πολλά χαρακτηριστικά με τιμή 0. Τέλος, το σύνολο δεδομένων που προέκυψε και χρησιμοποιήθηκε στη διαδικασία εκμάθησης περιείχε 16,578 μετρήσεις και 9 κατηγορίες χαρακτηριστικών.

Χρησιμοποιήθηκαν τεχνικές κατηγοριοποίησης καθώς και παλινδρόμησης. Προκειμένου να αξιολογηθεί η ακρίβεια της πρόβλεψης κάθε αλγόριθμου, πραγματοποιήθηκαν πειράματα σε δύο περιπτώσεις. Στην πρώτη περίπτωση το σύνολο δεδομένων αποτελείται από χαρακτηριστικά δεδομένων που προέρχονται μόνο από αισθητήρες, ενώ η δεύτερη περίπτωση προσθέτει στα δεδομένα αισθητήρα, δεδομένα που εισάγονται με μη αυτόματο τρόπο συλλογής. Οι μέθοδοι κατηγοριοποίησης που χρησιμοποιήθηκαν ήταν, ο αλγόριθμος J48 που ανήκει στην οικογένεια των δέντρων απόφασης και τα Bayesian δίκτυα. Για τη μέθοδο της παλινδρόμησης, χρησιμοποιήθηκε ο αλγόριθμος της γραμμικής παλινδρόμησης.

Στο Σχήμα 3.8 φαίνεται η απόδοση των τριών διαφορετικών μοντέλων μηχανικής μάθησης για τα δύο διαφορετικά σύνολα δεδομένων. Οι μετρήσεις αξιολόγησης που χρησιμοποιήθηκαν ήταν το μέσο απόλυτο σφάλμα (MAE), το RMSE και η ακρίβεια (ACC). Συνοψίζοντας, τις καλύτερες επιδόσεις είχαν τα Bayesian δίκτυα με το μικρότερο $RMSE = 0.24$ και $MAE = 0.1$ για το σύνολο δεδομένων με δεδομένα από αισθητήρα και δεδομένα που εισήχθησαν με μη αυτόματο τρόπο, καθώς και τα δέντρα απόφασης που φαίνεται ότι μπορούν να προβλέψουν με ακρίβεια τον αριθμό ατόμων με βάση τα δεδομένα αισθητήρων. Η βελτίωση των αλγόριθμων μπορεί να

Evaluation \ Algorithm		J48	BayesNet	Linear Regression
Simple Dataset	MAE	0.17	0.12	0.44
	RMSE	0.29	0.26	0.54
	ACC	73%	80%	—
Augmented Dataset	MAE	0.15	0.1	0.34
	RMSE	0.27	0.24	0.45
	ACC	78%	83%	—

MAE:mean absolute error; RMSE: root mean squared error;
ACC: accuracy

Σχήμα 3.8: Αξιολόγηση των αλγόριθμων κατηγοριοποίησης και παλινδρόμησης [Moraru et al., 2010].

επιτευχθεί με την επέκταση του συστήματος κόμβου αισθητήρων και τη συλλογή περισσότερων δεδομένων.

3.8 Πρόβλεψη της μελλοντικής ωριαίας κατανάλωσης ενέργειας με τη χρήση αλγόριθμων μηχανικής μάθησης.

Οι Edwards et al. [Edwards et al., 2012] εξετάζουν επτά τεχνικές μηχανικής μάθησης για την πρόβλεψη της ωριαίας κατανάλωσης ενέργειας σε τρεις διαφορετικές κατοικίες με τη χρήση δεδομένων από ένα σύνολο αισθητήρων. Στο κάθε ένα από αυτά τα σπίτια έχουν τοποθετηθεί 140 αισθητήρες που συλλέγουν δεδομένα ανά 15 λεπτά από την 1η Ιανουαρίου 2010 έως 31 Δεκεμβρίου 2010. Επίσης, κάθε σπίτι είναι εξοπλισμένο με αυτοματοποιημένα χειριστήρια που διαχειρίζονται το άνοιγμα/κλείσιμο της πόρτας ψυγείου, τη χρήση του φούρνου, του πληντυρίου και του ντους. Τα αυτοματοποιημένα χειριστήρια λειτουργούν με βάση ένα μοτίβο που είναι σύμφωνο με τα γενικά πρότυπα χρήσης ενέργειας σε σπίτια της Αμερικής. Αυτό το μοτίβο διασφαλίζει ότι το σύνολο δεδομένων είναι πιο σταθερό και δεν επηρεάζεται από παράγοντες διαφορετικής συμπεριφοράς σε κάθε σπίτι, κάνοντας πιο εύκολη τη σύγκριση των δεδομένων μεταξύ των σπιτιών. Κάθε σπίτι λειτουργεί σύμφωνα με αυτά τα πρότυπα, με εξαίρεση το 2ο και 3ο σπίτι που είναι εξοπλισμένα με πιο

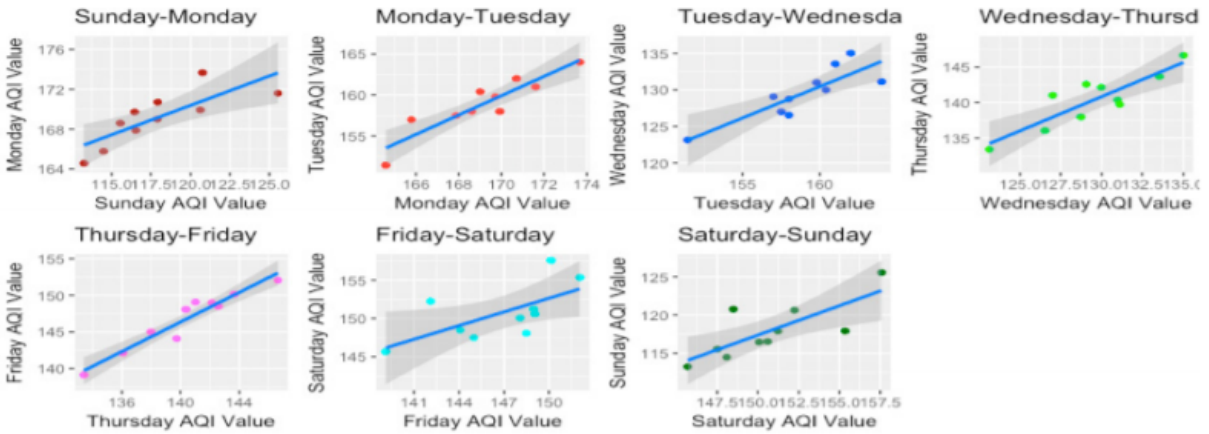
ενεργειακά αποδοτικές συσκευές, ενώ μόνο το 3ο σπίτι διαθέτει μία σειρά φωτοβολταϊκών για την παραγωγή ηλεκτρικής ενέργειας και ηλιακό θερμοσίφωνα.

Το αρχικό ενδιαφέρον των ερευνητών είναι ποιο μοντέλο μηχανικής μάθησης αποδίδει καλύτερα στην πρόβλεψη κατανάλωσης ενέργειας για την επόμενη ώρα. Δοκίμασαν κάθε τεχνική με ένα αριθμό διαφορετικών διαμορφώσεων, καθώς με τη χρήση K-Fold cross validation όπου το κάθε μοντέλο εκπαιδεύεται και δοκιμάζεται χρησιμοποιώντας 10 πτυχώσεις. Στη συνέχεια, παρουσιάζονται οι μετρήσεις απόδοσης που χρησιμοποίησαν για την αξιολόγηση του κάθε μοντέλου. Ο συντελεστής μεταβλητότητας (CV) που καθορίζει πόσο διαφέρει το συνολικό σφάλμα πρόβλεψης σε σχέση με το μέσο όρο του στόχου. Το μέσο συστηματικό σφάλμα (MBE) όπου καθορίζει πόσο πιθανό είναι το συγκεκριμένο μοντέλο να υπερεκτιμήσει ή υποτιμήσει την πραγματική κατανάλωση ενέργειας. Τέλος, χρησιμοποιήθηκε η μέτρηση του μέσου απόλυτου ποσοστού σφάλματος (MAPE), μία μέτρηση που χρησιμοποιείται συνήθως για την αξιολόγηση της ακρίβειας της παλινδρόμησης. Τα μοντέλα μηχανικής μάθησης που χρησιμοποιήθηκαν ήταν τα εξής: γραμμική παλινδρόμηση, Feedforward νευρωνικά δίκτυα (FFNN), παλινδρόμηση μηχανών υποστήριξης, least squares μηχανές υποστήριξης (LS-SVM), Hierarchical Mixture of Experts (HME) με γραμμική παλινδρόμηση, HME με Feedforward νευρωνικά δίκτυα (HME-FFNN) και Fuzzy C-Means με Feedforward νευρωνικά δίκτυα (FCM-FFNN). Τα αποτελέσματα έδειξαν πως το μοντέλο που αποδίδει καλύτερα στην πρόβλεψη της κατανάλωσης ενέργειας κάθε σπιτιού είναι το LS-SVM, ενώ φάνηκε ότι το απλό FFNN μοντέλο είχε χαμηλότερη απόδοση από τις νεότερες τεχνικές που δοκιμάστηκαν όπως HME-FFNN, LS-SVM και FCM-FFNN. Τέλος, τα αποτελέσματα δείχνουν πως ενώ η παλινδρόμηση μηχανών υποστήριξης και οι least squares μηχανές υποστήριξης έχουν παρόμοια αποτελέσματα όσον αφορά το CV και το MAPE, τα αποτελέσματα MBE που παρουσιάζει η μέθοδος SVR είναι πολύ χειρότερα, γεγονός που κάνει την LS-SVM την προτεινόμενη τεχνική.

3.9 Πρόβλεψη της ποιότητας του αέρα με χρήση του IoT και μηχανικής μάθησης.

Οι Kumar et al. [Kumar et al., 2020] εφαρμόζουν μία τεχνική μηχανικής μάθησης σε δεδομένα χρονοσειρών που προέρχονται από αισθητήρες και εξετάζουν την ακρίβεια πρόβλεψης της ποιότητας του αέρα σε ένα IoT περιβάλλον. Τα δεδομένα ποιότητας του αέρα μετρήθηκαν από αισθητήρες θερμοκρασίας και αέρα. Τα δεδομένα που συλλέχθηκαν κατά την περίοδο μεταξύ Ιουλίου και Σεπτεμβρίου 2018, αφορούν την περιοχή του Νέου Δελχί Ινδίας. Το σύνολο δεδομένων περιέχει τιμές του συστατικού ποιότητας του αέρα (AQI) και PPM τιμές συγκέντρωσης διάφορων ρυπογόνων αερίων, όπως μονοξείδιο του άνθρακα, διοξείδιο του άνθρακα, αμμωνία και ακετόνη. Τα δεδομένα συλλέχθηκαν από μία IoT διάταξη η οποία αποτελείται από ένα MQ-135 αισθητήρα αερίου, DHT-22 αισθητήρα υγρασίας και θερμοκρασίας και ένα μικροελεγκτή Arduino-Uno. Για την αντιμετώπιση των ελλειπών τιμών και στη συνέχεια την εφαρμογή της παλινδρόμησης σε αυτά τα δεδομένα, οι ερευνητές χρησιμοποίησαν τον αλγόριθμο expectation-maximization (EM) που λειτουργεί μεταξύ ελλειπόντων δεδομένων και διαθέσιμων δεδομένων.

Οι ερευνητές εξέτασαν την τάση (trend) της ποιότητας του αέρα για την περίοδο μίας εβδομάδας και παρατήρησαν ότι υπάρχει ένα μοτίβο στην εναλλαγή των τιμών κατά τη διάρκεια της εβδομάδας. Εφάρμοσαν γραμμική παλινδρόμηση για κάθε ζευγάρι συνεχόμενων ημερών για την πρόβλεψη της τιμής PPM χρησιμοποιώντας την τιμή της ποιότητας του αέρα της προηγούμενης ημέρας, όπως φαίνεται στο Σχήμα 3.9.



Σχήμα 3.9: Γραφήματα γραμμικής παλινδρόμησης για δύο συνεχόμενες ημέρες [Kumar et al., 2020].

Οι μετρήσεις απόδοσης που χρησιμοποιήθηκαν για την αξιολόγηση του μοντέλου ήταν το MAE, το MSE, το RMSE και το MAPE. Ως εργαλείο σύγκρισης χρησιμοποιήθηκε ένα μοντέλο για την πρόβλεψη της ποιότητας του αέρα στην Κίνα βασισμένο στην παλινδρόμηση με μηχανές διανυσμάτων υποστήριξης [Liu et al., 2017]. Ακόμη, εφαρμόζοντας τη γραμμική παλινδρόμηση σε δεδομένα διάφορων περιοχών του Νέου Δελχί, παρατηρήθηκε ότι βελτιώθηκε σημαντικά η μέτρηση του μέσου απόλυτου ποσοστού σφάλματος σε σχέση με το μοντέλο σύγκρισης [Liu et al., 2017].

Models	MAE	MSE	RMSE	MAPE
Benchmark Model	7.2675	108.5975	10.355	0.0881
Proposed Model	7.57	107.9533	10.29667	0.0769

Σχήμα 3.10: Μετρήσεις απόδοσης για το μοντέλο σύγκρισης και για το προτεινόμενο μοντέλο [Kumar et al., 2020].

Όπως φαίνεται στο Σχήμα 3.10, το προτεινόμενο μοντέλο παρουσιάζει καλύτερα αποτελέσματα στις μετρήσεις απόδοσης από το μοντέλο σύγκρισης. Συνοψίζοντας, η χρήση ιστορικών δεδομένων που συλλέγονται από αισθητήρες για την ακριβή πρόβλεψη της ποιότητας του αέρα μπορεί να βοηθήσει στη λήψη αποφάσεων και την οργάνωση των επόμενων βημάτων για την ελαχιστοποίηση της ρύπανσης του αέρα. Ωστόσο, μπορεί να υπάρξουν και άλλοι παράγοντες όπως η πυκνότητα της κυκλοφορίας, οι δραματικές αλλαγές του καιρού ή οι ιδιαιτερότητες μίας συγκεκριμένης περιοχής, που μπορεί να επηρεάσουν την ακρίβεια του μοντέλου και θα πρέπει να ληφθούν υπόψη για τη βελτίωση της ακρίβειας του αποτελέσματος.

3.10 Σύγκριση τεχνικών μηχανικής μάθησης για την πρόβλεψη της ηλιακής ενέργειας με δεδομένα καιρού από αισθητήρες.

Οι Carrera & Kim [Carrera and Kim, 2020] προτείνουν ένα πλαίσιο σύγκρισης διαφόρων τεχνικών μηχανικής μάθησης για την πρόβλεψη της ηλιακής ενέργειας. Η πρόβλεψη αφορά το εργοστάσιο ηλιακής ενέργειας Yeongam που βρίσκεται στην περιοχή South Jeolla στην Νότια Κορέα. Το πλαίσιο σύγκρισης τεχνικών μηχανι-

κής μάθησης αποτελείται από 5 μέρη: τη συλλογή δεδομένων, την προεπεξεργασία δεδομένων, τη διασταυρούμενη επικύρωση, τη διασταυρούμενη επικύρωση 10 πτυχώσεων και την επιλογή υποσυνόλου χαρακτηριστικών (subset selection). Το σύνολο δεδομένων δημιουργήθηκε από τέσσερα είδη δεδομένων. Τα δεδομένα παραγωγής ηλιακής ενέργειας παρέχονται από την ανοιχτή πύλη δεδομένων της Κορέας και αποτελούνται από ωριαία δεδομένα την περίοδο 1 Ιανουαρίου 2013 έως 31 Δεκεμβρίου 2015. Τα δεδομένα ηλιακής ανύψωσης λαμβάνονται από τη γεωγραφική θέση του σταθμού ηλεκτροπαραγωγής Yeongam και αντιπροσωπεύει τη θέση του ηλίου από την προοπτική του σταθμού σε συντεταγμένες γεωγραφικού πλάτους και μήκους. Στη συνέχεια, τα μετεωρολογικά δεδομένα που παρέχονται από τη Μετεωρολογική Διοίκηση της Κορέας (ΜΔΚ), είναι ο πραγματικός καιρός που μετράται. Τέλος, τα αρχεία πρόβλεψης που ανακοινώνονται από τη ΜΔΚ δίνονται σε περίοδο τριών ωρών από την έναρξη στις 02:00 το πρωί για κάθε ημέρα. Το σύστημα μετεωρολογικής παρατήρησης της ΜΔΚ αποτελείται από ένα πύργο ύψους 10 μέτρων. Στην κορυφή είναι τοποθετημένοι αισθητήρες μέτρησης της ταχύτητας και της κατεύθυνσης του αέρα, σε ύψος 1.5 μέτρο από το έδαφος είναι τοποθετημένοι οι αισθητήρες θερμοκρασίας και υγρασίας και σε ύψος περίπου 50 με 60 εκατοστά από το έδαφος είναι οι αισθητήρες πίεσης. Σχετικά με την προεπεξεργασία δεδομένων, εξαιρέθηκαν οι μετρήσεις χωρίς πληροφορίες για το ηλιακό φως (00:00–05:00 και 20:00–24:00). Στη συνέχεια, οι διαδοχικές μετρήσεις σε περιόδους τριών ωρών, συγχωνεύονται σε μία μέτρηση δημιουργώντας ένα σύνολο 5 μετρήσεων τη μέρα. Αυτή η διεργασία συγχώνευσης έγινε για να ταιριάζουν τα δεδομένα ηλιακής ενέργειας και τα μετεωρολογικά δεδομένα, τα οποία δίνονται ανά μία ώρα, με τα δεδομένα πρόβλεψης του καιρού τα οποία δίνονται ανά τρεις ώρες. Επιπλέον, τα δεδομένα ηλιακής ενέργειας τεμαχίστηκαν σε 7 χρονικές περιόδους έτσι ώστε τα μοντέλα μηχανικής μάθησης να μπορούν να προβλέπουν κάθε περίοδο μία προς μία. Τέλος, στις ποιοτικές μεταβλητές έγινε μετατροπή όλων των κατηγορικών μεταβλητών σε πολλαπλές δυαδικές μεταβλητές.

Στη συνέχεια της μελέτης, χρησιμοποιήθηκαν διάφοροι αλγόριθμοι μηχανικής μάθησης για πολλαπλές ανεξάρτητες μεταβλητές οι οποίοι αξιολογήθηκαν και κατηγοριοποιήθηκαν σε τρεις κατηγορίες. Όπως φαίνεται και στο Σχήμα 3.11, οι αλγόριθμοι αυτοί κατηγοριοποιούνται ως εξής: αλγόριθμοι μονής παλινδρόμησης, αλ-

γόριθμοι ensemble bagging και ensemble boosting. Οι μετρήσεις για την απόδοση των μοντέλων που χρησιμοποιήθηκαν ήταν αυτή του RMSE, του MAE και αυτή του συντελεστή προσδιορισμού R^2. Στις τρεις στατιστικές αναλύσεις που πραγματοποιήθηκαν, παρουσιάζονται οι στατιστικές πληροφορίες όλων των μοντέλων παλινδρόμησης σχετικά με το RMSE, MAE και R^2, καθώς και ο μέσος όρος και η τυπική απόκλιση (STD) για την κάθε μέτρηση. Στην πρώτη στατιστική ανάλυση με βάση τα δεδομένα παρατήρησης καιρού, καλύτερη απόδοση είχε ο αλγόριθμος Decision Tree με ($RMSE = 694.24$, $MAE = 478.77$ και $R^2 = 61.6\%$) ενώ ο αλγόριθμος Gradient Boosting είχε εξίσου καλή απόδοση.

Single Regression	Ensemble (Bagging)	Ensemble (Boosting)
Linear regression	Bagging	AdaBoost
Huber	Random forest	Gradient boosting
Ridge	Extra trees	CatBoost
Lasso		XGBoost
Elastic Net		
Decision tree		
k-NN		
SVR		

Σχήμα 3.11: Μέθοδοι μηχανικής μάθησης [Carrera and Kim, 2020].

Στη συνέχεια, έγινε στατιστική ανάλυση με βάση τα δεδομένα πρόβλεψης του καιρού, όπου την καλύτερη απόδοση είχε ο αλγόριθμος k-NN με μετρήσεις $RMSE = 542.09$, $MAE = 350.20$ και $R^2 = 76.6\%$. Για τα ensemble μοντέλα όπως και προηγουμένως, ο αλγόριθμος Gradient Boosting είχε καλά αποτελέσματα RMSE και R^2 ($RMSE = 531.85$ και $R^2 = 77.5\%$), ενώ ο αλγόριθμος Random Forest παρουσίασε τα καλύτερα αποτελέσματα σχετικά με τη μέτρηση MAE, περίπου ίσο με 340.

Τέλος, στο Σχήμα 3.12 φαίνεται η τρίτη στατιστική ανάλυση στην οποία χρησιμοποιήθηκαν τα δεδομένα πρόβλεψης καιρού και τα δεδομένα παρατηρήσεων του καιρού. Από τους αλγόριθμους μονής παλινδρόμησης ο αλγόριθμος k-NN είχε τις καλύτερες μετρήσεις απόδοσης ($RMSE = 547.79$, $MAE = 358.28$ και $R^2 = 76.0\%$). Στα ensemble μοντέλα, οι αλγόριθμοι bagging και Gradient boosting είχαν καλύτερες αποδόσεις από τα μοντέλα μονής παλινδρόμησης, με καλύτερη απόδοση αυτή του Gradient boosting ($RMSE = 517.56$, $R^2 = 78.6\%$).

Prediction Models		RMSE		MAE		R^2	
		Mean	STD	Mean	STD	Mean	STD
Single regression models	Linear regression	638.18	43.81	497.20	38.50	0.6741	0.0440
	Huber	668.09	34.30	497.38	25.83	0.6419	0.0494
	Ridge	638.06	43.62	496.58	38.22	0.6742	0.0440
	Lasso	638.36	43.61	495.98	38.02	0.6740	0.0437
	Elastic net	639.37	42.45	495.47	35.45	0.6727	0.0445
	Decision tree	558.35	49.92	366.54	36.39	0.7512	0.0354
	k-NN	547.79	56.24	358.28	31.88	0.7601	0.0395
	SVR	656.00	38.92	483.36	29.68	0.6551	0.0472
Ensemble models (bagging)	Bagging	530.90	42.84	338.03	27.46	0.7770	0.0295
	Random forest	525.88	46.17	339.20	29.82	0.7791	0.0324
	Extra trees	524.92	48.44	340.73	31.94	0.7805	0.0287
	AdaBoost	612.09	38.51	454.29	40.33	0.7088	0.0416
Ensemble models (boosting)	**Gradient boosting**	517.56	42.09	341.22	24.99	0.7864	0.0267
	CatBoost	536.23	52.59	364.20	32.35	0.7704	0.0353
	XGBoost	518.30	43.45	343.65	25.59	0.7850	0.0324

Σχήμα 3.12: Συγκριτικά στατιστικά στοιχεία των μοντέλων πρόβλεψης με βάση τα δεδομένα πρόβλεψης και δεδομένα παρατήρησης καιρού, χρησιμοποιώντας διασταυρούμενη επικύρωση 10 πτυχώσεων [Carrera and Kim, 2020].

3.11 Παρακολούθηση του καιρού σε πραγματικό χρόνο και πρόβλεψη με τη χρήση αισθητήρων σε αστικά λεωφορεία και μηχανικής μάθησης.

Οι Huang et al. [Huang et al., 2020] παρουσιάζουν ένα σύστημα παρακολούθησης καιρού σε πραγματικό χρόνο, καθώς και ένα σύστημα πρόβλεψης του καιρού βασιζόμενο στη διαχείριση των πληροφοριών που προέρχονται από τα αστικά λεωφορεία. Με τη χρήση της μηχανικής μάθησης ολοκληρώνεται η επικοινωνία και η ανάλυση των πληροφοριών μεταξύ των αστικών λεωφορείων, των στάσεων λεωφορείων και των αισθητήρων. Τα λεωφορεία και οι σταθμοί λεωφορείων έχουν εξοπλιστεί με Raspberry Pi 3 Model B+ και διάφορους αισθητήρες καιρού για τη μεταξύ τους ασύρματη επικοινωνία, έτσι ώστε να επιτευχθεί η μετάδοση και συλλογή δεδομένων. Τα δεδομένα συλλέγονται από 5 είδη αισθητήρων, τους αισθητήρες θερμοκρασίας/υγρασίας, υπεριώδεις αισθητήρες, αισθητήρες βροχόπτωσης, αισθητήρες συγκέντρωσης PM 2.5 αερίου και αισθητήρες πίεσης αέρα. Οι διαθέσιμοι αισθητήρες καταγράφουν τις μετρήσεις, την ώρα και την ταυτότητα (ID) του σταθμού λεωφορείου που στέλνονται, για διάφορες κατηγορίες δεδομένων και τις αποθηκεύουν σε μία βάση δεδομένων. Στη συνέχεια, εφαρμόζονται δύο αλγόριθμοι μηχανικής μάθησης για την πρόβλεψη των τιμών θερμοκρασίας, υγρασίας και πίεσης αέρα.

Για να εξεταστεί η απόδοση της πρόβλεψης, τα δεδομένα εισόδου χωρίζονται σε 4 κατηγορίες G1, G2, G3 και G4. Αυτό γίνεται για να εξεταστούν τα δεδομένα και η απόδοση της πρόβλεψης σε διάφορες χρονικές περιόδους. Για την κατηγορία G1, σκοπός είναι να εξερευνήσουν τα χαρακτηριστικά των δεδομένων σε σύντομες παρακείμενες χρονικές περιόδους (π.χ. 3 ώρες), ενώ για την κατηγορία G2 σε μέτριες χρονικές περιόδους (π.χ. 12 ώρες) και G3, G4 σε μεγάλες χρονικές περίοδους (π.χ. 24 και 48 ώρες), αντίστοιχα.

Τα μοντέλα μηχανικής μάθησης που χρησιμοποιήθηκαν είναι το νευρωνικό δίκτυο multilayer perceptron και το μοντέλο deep learning LSTM. Οι μετρήσεις απόδοσης που χρησιμοποιήθηκαν για την αξιολόγηση σύγκρισης των προβλεπόμενων καιρικών τιμών (θερμοκρασία, υγρασία, πίεση) και των πραγματικών δεδομένων καιρού είναι το RMSE, το MAE και το σφάλμα επί τις εκατό (percentage error). Για το MLP μοντέλο, τα καλύτερα συνολικά αποτελέσματα για τις τρεις προβλέψεις (θερμοκρασία, υγρασία, πίεση) παρουσιάστηκαν στην κατηγορία G1 εισόδου δεδομένων, ενώ η κατηγορία G2 δεδομένων εισόδου είχε καλύτερα αποτελέσματα για τα δεδομένα υγρασίας. Ομοίως, για το LSTM μοντέλο τα καλύτερα συνολικά αποτελέσματα παρουσιάστηκαν στην κατηγορία δεδομένων εισόδου G1. Στη συνέχεια, έγινε σύγκριση των δύο μοντέλων για την πρόβλεψη της 21ης Μαρτίου 2020, με δεδομένα εισόδου κατηγορίας G1 στο καθένα.

		Temperature	Humidity	Pressure
RMSE	LSTM	1.3219	2.8696	0.7676
	MLP	0.907	6.7972	1.0369
MAE	LSTM	1.0561	2.2483	0.6557
	MLP	0.7731	5.604	0.8433
Percentage Error	LSTM	4.15%	3.39%	0.07%
	MLP	3.17%	8.61%	0.08%

Σχήμα 3.13: Αποτελέσματα πρόβλεψης θερμοκρασίας, υγρασίας και πίεσης στις 21 Μαρτίου 2020: Σύγκριση LSTM και MLP μοντέλου [Huang et al., 2020].

Όπως φαίνεται και στο Σχήμα 3.13, και τα δύο μοντέλα είχανε αρκετά καλές προβλέψεις, με αυτή του MLP μοντέλου να είναι ελαφρώς καλύτερη στα δεδομένα θερμοκρασίας. Αντιθέτως στα δεδομένα υγρασίας και πίεσης αέρα, το μοντέλο LSTM έχει πολύ ανώτερες επιδόσεις στις προβλέψεις των δεδομένων. Αυτό μπορεί να οφείλεται στο βάθος του μοντέλου κατά την εκπαίδευση και στις παραμέτρους που χρησιμοποιούνται σε κάθε επίπεδο. Συνοψίζοντας, το σύστημα παρακολούθη-

σης καιρού σε πραγματικό χρόνο και η εφαρμογή αλγόριθμων μηχανικής μάθησης στα δεδομένα των αισθητήρων μπορούν να χρησιμοποιηθούν αποτελεσματικά για την πρόβλεψη των καιρικών δεδομένων.

3.12 Πρόβλεψη του θερμικού φορτίου σε δίκτυα τηλεθέρμανσης χρησιμοποιώντας μηχανική εκμάθηση και συμβουλές από ειδικούς.

Οι Geysen et al. [Geysen et al., 2017] χρησιμοποιούν τη μηχανική μάθηση και ένα σύστημα από συμβουλές ειδικών σαν είσοδο για την ακριβή πρόβλεψη του θερμικού φορτίου σε ένα δίκτυο τηλεθέρμανσης. Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν είναι η γραμμική παλινδρόμηση, τα τεχνητά νευρωνικά δίκτυα (ANN), οι μηχανές διανυσμάτων υποστήριξης και τα εξαιρετικά τυχαιοποιημένα δέντρα παλινδρόμησης (Extremely randomized tree regressors). Αντί να συγκριθεί η κάθε μέθοδος ξεχωριστά, ο σκοπός της μελέτης είναι να εφαρμόσει έναν αλγόριθμο ικανό να συνδυάσει πληροφορίες από N ειδικούς πρόβλεψης θερμικού φορτίου, με ένα τρόπο που να παρακολουθεί πάντα τον καλύτερο από αυτούς τους N ειδικούς.

Στο πείραμα χρησιμοποιήθηκαν τα δεδομένα από ένα δίκτυο τηλεθέρμανσης στο Rottne της Σουηδίας το οποίο αποτελείται από 200 κατοικίες. Κάθε κατοικία διαθέτει έναν υποσταθμό τηλεθέρμανσης που ελέγχεται από έναν ελεγκτή. Κατά τη διάρκεια του πειράματος προστέθηκε μία συσκευή που δίνει τη δυνατότητα απομακρυσμένης αλληλεπίδρασης και τροποποίησης του ελεγκτή μέσω ενός μηχανισμού παράκαμψης του αισθητήρα θερμοκρασίας εξωτερικού χώρου. Αυτό έδωσε τη δυνατότητα ελέγχου του υποσταθμού και της αποστολής εναλλακτικών σημάτων θερμοκρασίας στα οποία στη συνέχεια ο ελεγκτής θα ανταποκριθεί με βάση τις προεπιλεγμένες ρυθμίσεις. Επιπλέον αισθητήρες τοποθετήθηκαν για τη μέτρηση των δεδομένων θερμοκρασίας από την προμήθεια και επιστροφή στο σύστημα θέρμανσης, καθώς και για τη μέτρηση δεδομένων από το μετρητή θερμότητας. Το τελικό σύνολο δεδομένων που προέκυψε αποτελούνταν από τα δεδομένα του πειράματος σε πραγματικό χρόνο για το δίκτυο τηλεθέρμανσης, καθώς και από ιστορικά δεδομένα των σημάτων ελεγκτή, του θερμικού φορτίου και της πρόβλεψης καιρικών φαινομένων για το δίκτυο τηλεθέρμανσης. Τα ακατέργαστα δεδομένα καλύπτουν

μία περίοδο 27 μηνών, από το Νοέμβριο του 2014 έως το Φεβρουάριο του 2017 τα οποία χωρίστηκαν σε ένα ποσοστό 75% για εκπαίδευση και 25% για δοκιμές.

	Timing			Temp	Thermal load		Control signal	
	HoD	DoW	DoY	$\hat{T}_{out}$	P_{t-24}	P_{t-168}	dT_{t-24}	dT_{t-168}
Full set	√	√	√	√	√	√	√	√
Set-dT	√	√	√	√	√	√	×	×
Set-lags	√	√	√	√	×	×	×	×

Σχήμα 3.14: Σύνολο χαρακτηριστικών με ώρα της ημέρας (HoD), ημέρα της εβδομάδας (DoW), ημέρα του χρόνου (DoY), προβλεπόμενη εξωτερική θερμοκρασία (T_{out}), Θερμικό φορτίο προηγούμενης μέρας (P_{t-24}) και προηγούμενης εβδομάδας (P_{t-168}), σήματα ελεγκτή προηγούμενης ημέρας (dT_{t-24}) και προηγούμενης εβδομάδας (dT_{t-168}) [Geysen et al., 2017].

Στο Σχήμα 3.14 φαίνεται το σύνολο χαρακτηριστικών που χρησιμοποιήθηκαν. Το πρώτο σετ περιέχει όλα τα χαρακτηριστικά, ενώ στο δεύτερο δεν περιέχονται τα ιστορικά δεδομένα σημάτων του ελεγκτή. Τέλος, στο τρίτο σετ δεν περιέχονται τα ιστορικά δεδομένα σημάτων του ελεγκτή ούτε τα ιστορικά δεδομένα θερμικού φορτίου. Η μέτρηση της απόδοσης του κάθε μοντέλου έγινε με την τιμή του MAPE. Ακόμη χρησιμοποιήθηκε η διασταυρούμενη επικύρωση GridSearchCV της βιβλιοθήκης scikit-learn για την αξιολόγηση των μοντέλων πρόβλεψης.

	LR	ETR	SVM	ANN	Forecaster
No retrain	17.34 %	12.34 %	14.54 %	**11.92 %**	12.06 %
Daily retrain	17.27 %	12.42 %	14.72 %	**11.56 %**	11.95 %

Σχήμα 3.15: Σύγκριση της μέτρησης MAPE χωρίς επανεκπαίδευση (πάνω) και με επανεκπαίδευση (κάτω) [Geysen et al., 2017].

Τα αποτελέσματα όπως φαίνεται στο Σχήμα 3.15 έδειξαν πως η γραμμική παλινδρόμηση είχε τις χειρότερες επιδόσεις ($MAPE = 17.34\%$), ενώ τα νευρωνικά δίκτυα και τα εξαιρετικά τυχαιοποιημένα δέντρα παλινδρόμησης είχαν τις καλύτερες επιδόσεις ($MAPE = 11.92\%$ και $MAPE = 12.34\%$ αντίστοιχα). Ακόμη, έγινε ανάλυση του διαστήματος επανεκπαίδευσης και το συμπέρασμα ήταν πως η επανεκπαίδευση δεν αυξάνει την απόδοση πρόβλεψης.

Κεφάλαιο 4

Υπολογιστικά Πειράματα

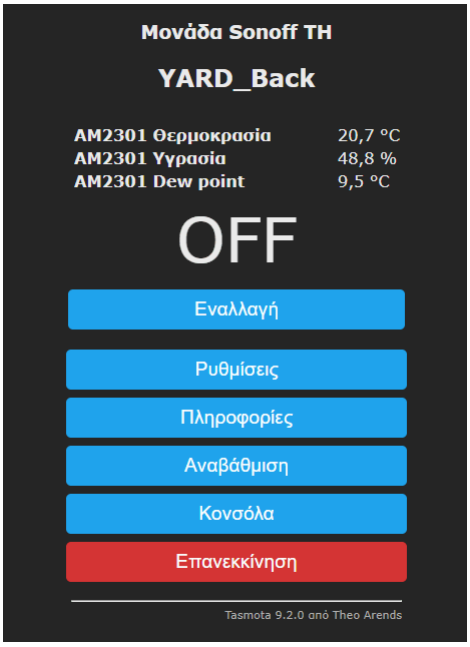
4.1 Εισαγωγή

Ο κύριος στόχος του πειράματος είναι η πρόβλεψη της θερμοκρασίας χρησιμοποιώντας ως είσοδο δεδομένα από αισθητήρες που έχουν τοποθετηθεί σε μία κατοικία στην Λευκόβρυση Κοζάνης. Οι προβλέψεις γίνονται σε 7 διαφορετικά σημεία του σπιτιού όπως τα υπνοδωμάτια, το μπάνιο, η κουζίνα, η αυλή, το γκαράζ και το χώρο του γραφείου που βρίσκεται στο Πανεπιστήμιο Δυτικής Μακεδονίας. Για την πρόβλεψη των τιμών θερμοκρασίας χρησιμοποιήθηκαν αλγόριθμοι μηχανικής μάθησης, όπως η γραμμική παλινδρόμηση, η παλινδρόμηση Ridge, η παλινδρόμηση Lasso, οι μηχανές διανυσμάτων υποστήριξης, τα δέντρα απόφασης, τα τυχαία δάση και το νευρωνικό δίκτυο MLP. Χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python (version 3.9) και η βιβλιοθήκη μηχανικής μάθησης scikit-learn. Ακόμη, χρησιμοποιήθηκαν οι βιβλιοθήκες NumPy, Pandas, Matplotlib καθώς και το Optuna, ένα πλαίσιο λογισμικού για τη βελτιστοποίηση των υπερπαραμέτρων των αλγόριθμων μηχανικής μάθησης.

4.2 Περιγραφή συστήματος αισθητήρων

Η αρχιτεκτονική του συστήματος συλλογής δεδομένων που χρησιμοποιήθηκε βασίζεται στα έξυπνα ρελέ Sonoff-Touch [Sonoff, 2020] που έχουν τοποθετηθεί στους επτά χώρους του σπιτιού και στο χώρο του γραφείου, τα οποία συνδέονται αυτόματα στο δίκτυο Wi-Fi του σπιτιού. Προσθέτοντας στα ρελέ το firmware Tasmota [Tasmota, 2016] γίνονται επεκτάσιμα και επιτρέπεται η σύνδεση διαφόρων ειδών αισθητήρων όπως αισθητήρες θερμοκρασίας, υγρασίας και πίεσης. Στο

συγκεκριμένο πείραμα χρησιμοποιήθηκαν οι μετρήσεις θερμοκρασίας κάθε χώρου, καθώς εκείνη την περίοδο είχε απενεργοποιηθεί η καταγραφή των μετρήσεων από τους αισθητήρες υγρασίας και πίεσης. Για τους εσωτερικούς χώρους του σπιτιού χρησιμοποιήθηκαν οι ψηφιακοί αισθητήρες αέρα Bosch BMP280, ενώ για τον εξωτερικό χώρο της αυλής χρησιμοποιήθηκε ο ψηφιακός αισθητήρας αέρα AM2301. Οι μετρήσεις των αισθητήρων συλλέγονται μέσω ενός διακομιστή MQTT Broker, ο οποίος χρησιμοποιώντας το πρωτόκολλο MQTT τις διαβιβάζει στο περιβάλλον του Openhab [openHAB, 2021], ένα ανοιχτού κώδικα λογισμικό για τον οικιακό αυτοματισμό. Μία από τις αυτοματοποιημένες διεργασίες του λογισμικού Openhab είναι η συλλογή των δεδομένων και στη συνέχεια η αποθήκευσή τους στη βάση δεδομένων στο InfluxDB [influxdata, 2021], ένα λογισμικό ανοιχτού πηγαίου κώδικα για την αποθήκευση και διαχείριση βάσεων δεδομένων χρονοσειρών. Η Εικόνα 4.1 δείχνει το μενού ενός αισθητήρα για ένα από τους οκτώ χώρους, με ταυτόχρονη ένδειξη των μετρήσεων.



Σχήμα 4.1: Το μενού του αισθητήρα της εξωτερικής αυλής και οι πληροφορίες που καταγράφονται.

4.3 Συλλογή δεδομένων

Στο συγκεκριμένο πείραμα χρησιμοποιήθηκαν δύο είδη δεδομένων. Αρχικά χρησιμοποιήθηκαν δεδομένα τα οποία προέρχονται από τους αισθητήρες σε κάθε διαφορετικό χώρο του σπιτιού τα οποία στέλνονται και αποθηκεύονται σε μία βάση

δεδομένων στο InfluxDB. Τα δεδομένα των αισθητήρων που αφορούν τη θερμοκρασία σε διαφορετικούς χώρους του σπιτιού και αποτελούνται από την καταγραφή του χρόνου και της τιμής θερμοκρασίας ανά διαστήματα 10 λεπτών. Το σύνολο των δεδομένων που προέρχονται από τους αισθητήρες καλύπτει μία περίοδο περίπου 18 μηνών από 21 Μαρτίου 2019 έως 2 Σεπτεμβρίου 2020. Επιπλέον, χρησιμοποιήθηκαν ιστορικά μετεωρολογικά δεδομένα (θερμοκρασία, πίεση και υγρασία) για την πόλη της Κοζάνης την ίδια χρονική περίοδο, τα οποία λάβαμε από την Εθνική Μετεωρολογική Υπηρεσία (Ε.Μ.Υ). Τέλος, με τη σύνδεση του InfluxDB με την Python, έγινε εξαγωγή των δεδομένων σε αρχεία csv, όπως και τα δεδομένα που είχαν ληφθεί από την Ε.Μ.Υ.

4.4 Προεπεξεργασία δεδομένων

Μετά τη συλλογή δεδομένων ήταν απαραίτητος ο έλεγχος και η προεπεξεργασία των ακατέργαστων δεδομένων πριν την εφαρμογή των μοντέλων μηχανικής μάθησης. Αρχικά, μία αλλαγή που έπρεπε να γίνει ήταν η μορφή της μεταβλητής του χρόνου των δεδομένων, καθώς ήταν σε μορφή UTC και περιείχε τα γράμματα T και Z ανάμεσα στην ημέρα και την ώρα. Παρακάτω φαίνεται ο Κώδικας 4.1 που υλοποιήθηκε για την αφαίρεση των γραμμάτων T και Z από τη μεταβλητή του χρόνου. Σκοπός ήταν η μετατροπή της μεταβλητής του χρόνου στην τοπική ώρα και η αφαίρεση των γραμμάτων T και Z έτσι ώστε να υπάρχει μία καλύτερη εικόνα των δεδομένων και να είναι ευκολότερα στη διαχείριση.

Κώδικας 4.1: Αφαίρεση των γραμμάτων T και Z από τη μεταβλητή του χρόνου.

```
import pandas as pd

#insert bathroom temperature data into dataframe
df = pd.read_csv('bath_temp.csv')

#remove letters T and Z from time column
df['time'] = df['time'].str.replace('T', '')
df['time'] = df['time'].str.replace('Z', '')
```

```
df.to_csv(r'D:\final_bath_temp.csv', index=False)
```

4.4.1 Δεδομένα που λείπουν

Έχοντας μία καλύτερη εικόνα των δεδομένων, παρατηρήθηκε ένα από τα πιο συνηθισμένα προβλήματα που προκύπτει σε δεδομένα που προέρχονται από αισθητήρες, η αποτυχία αποστολής δεδομένων σε σταθερό χρόνο. Ενώ οι αισθητήρες είχαν ρυθμιστεί έτσι ώστε να στέλνουν μετρήσεις δεδομένων ανά 10 λέπτα, υπήρχαν αρκετές μετρήσεις οι οποίες δεν είχαν καταγραφεί. Κάνοντας έλεγχο για την ύπαρξη κενών τιμών στο σύνολο δεδομένων, δε βρέθηκαν κενές τιμές δεδομένων αλλά υπήρξαν τιμές που δεν είχαν καταγραφεί. Παρατηρώντας ότι οι μετρήσεις θερμοκρασίας δε μεταβάλλονται σημαντικά σε διάστημα μίας ώρας, αποφασίστηκε να ομαδοποιηθούν τα δεδομένα κάθε ώρας και να εξαχθεί ο μέσος όρος των μετρήσεων θερμοκρασίας, πίεσης και υγρασίας για κάθε μία ώρα της ημέρας. Αυτή η διαδικασία έγινε ομαδοποιώντας τα δεδομένα που ανήκουν σε μία μοναδική ώρα της κάθε ημέρας και στη συνέχεια κρατώντας το μέσο όρο των τιμών τους, όπως φαίνεται στο τμήμα του Κώδικα 4.2. Η διαδικασία έγινε ξεχωριστά για τα δεδομένα θερμοκρασίας σε κάθε σημείο του σπιτιού. Στη συνέχεια, η μεταβλητή του χρόνου μετατράπηκε στην επιθυμητή μορφή datetime (Έτος/Μήνας/Ημέρα/Ωρα). Τέλος, αποθηκεύτηκε το νέο σύνολο δεδομένων σε ένα ξεχωριστό DataFrame για κάθε μέτρηση καταλήγοντας σε 8 αρχεία csv, ένα για κάθε μέρος του σπιτιού και το χώρο του γραφείου.

Κώδικας 4.2: Εξαγωγή του μέσου όρου για κάθε ώρα ξεχωριστά.

```
def get_hour(string):
    hr = datetime.fromisoformat(string).hour   #Gets hour from
       datestring
    return hr
def get_date(string):
    date = datetime.strptime(string, '%Y-%m-%d %H:%M:%S')   #
       Gets date from datestring
    return str(date.date())
def get_average(df):
```

```
    average = df['value'].sum() / len(df['value'])   #
       calculates the average of values
    return average

df = pd.read_csv('final_bath_temp.csv')
print(df.head())
hr = [get_hour(i) for i in df['time']]
date = [get_date(i) for i in df['time']]

df['hour'] = hr
df['date'] = date
df.drop('time', axis=1, inplace=True)
unique_dates = np.unique(df['date'])  # returns list
avg_list = []
u_list = []
h_list = []

for u in unique_dates:
    unique_df = df[df['date'] == u]  # returns df containing
       only unique dates
    unique_hr = np.unique(unique_df['hour'])  # returns list
       containing only unique hours
    for h in unique_hr:
        unique_hr_df = unique_df[unique_df['hour'] == h]
        average = get_average(unique_hr_df)
        u_list.append(u)
        h_list.append(h)
        avg_list.append(average)
```

Δεδομένα της Ε.Μ.Υ που λείπουν.

Έχοντας τα οκτώ DataFrame που περιείχαν τα δεδομένα θερμοκρασίας κάθε χώρου, έπρεπε να γίνει συνένωση του κάθε DataFrame με τα ιστορικά δεδομένα

θερμοκρασίας, πίεσης και υγρασίας που αποκτήθηκαν από την Ε.Μ.Υ. Η συνένωση των δεδομένων αισθητήρων με τα επιπρόσθετα ιστορικά δεδομένα έπρεπε να γίνει για τις δοκιμές που θα ακολουθήσουν ως τρόπος βελτίωσης των αλγορίθμων. Καθώς τα δεδομένα της Ε.Μ.Υ περιείχαν μετρήσεις οι οποίες λάμβαναν χώρα την ίδια χρονική περίοδο, δεν υπήρχαν καταγραφές δεδομένων ακριβώς τις ίδιες χρονικές στιγμές. Αυτό είχε ως αποτέλεσμα κατά τη συνένωση των δύο συνόλων δεδομένων, να δημιουργηθούν κενά σημεία δεδομένων στο νέο σύνολο. Για το λόγο αυτό χρησιμοποιήθηκε η μέθοδος KNNImputer η οποία παρέχεται από τη βιβλιοθήκη scikit-learn. Η μέθοδος αυτή συμπληρώνει τις τιμές που λείπουν με το μέσο όρο συγκεκριμένων γειτονικών τιμών της κάθε στήλης. Στον Κώδικα 4.3 φαίνεται η διαδικασία συμπλήρωσης των τιμών που έλειπαν μέσω της μεθόδου KNNImputer όπου η τιμή συμπλήρωσης αποτελείται απο το μέσο όρο των δύο γειτονικών τιμών.

Κώδικας 4.3: Συμπλήρωση των ελλειπόντων τιμών που προέκυψαν στο DataFrame με τα δεδομένα της Ε.Μ.Υ με τη χρήση της μεθόδου KNNImputer.

```
#fill the missing values
from sklearn.impute import KNNImputer
imputer = KNNImputer(n_neighbors=2, weights="uniform")
imputer.fit(merged_data.iloc[:, 0:].values)
merged_data.iloc[:, 0:] = imputer.transform(merged_data.iloc[:,
    0:].values)
```

4.4.2 Μηχανική χαρακτηριστικών

Η πρόβλεψη δεδομένων χρονοσειρών μπορεί να θεωρηθεί ως εποπτευόμενο μαθησιακό πρόβλημα. Προτού χρησιμοποιηθεί η μηχανική μάθηση, θα πρέπει το πρόβλημα πρόβλεψης δεδομένων χρονοσειρών να επαναδιαμορφωθεί ως πρόβλημα εποπτευόμενης μηχανικής μάθησης. Δηλαδή, να μετατραπεί από μία ακολουθία δεδομένων, σε ζεύγη ακολουθιών εισόδου και εξόδου. Συνεπώς, χρησιμοποιήθηκε μία τεχνική μηχανικής χαρακτηριστικών για τη δημιουργία χαρακτηριστικών που σχετίζονται με την ημερομηνία και την ώρα. Στον Κώδικα 4.4 παρουσιάζεται η δημιουργία τεσσάρων νέων χαρακτηριστικών (Έτος/Μήνας/Ημέρα/Ωρα) από τη μεταβλητή της χρονικής σήμανσης των αισθητήρων.

Κώδικας 4.4: Δημιουργία χαρακτηριστικών που σχετίζονται με την ημερομηνία και την ώρα.

```
# generate new features from data
df['year'] = df.Time.dt.year
df['month'] = df.Time.dt.month
df['day'] = df.Time.dt.day
df['hour'] = df.Time.dt.hour


print(df.head())
```

Έτσι, το νέο DataFrame που προκύπτει περιέχει τις ημερομηνίες, την ώρα και τη μέτρηση, καθώς και τις τιμές των νέων χαρακτηριστικών που δημιουργήθηκαν. Στην Εικόνα 4.2 παρουσιάζεται το νέο DataFrame με τα νέα χαρακτηριστικά που σχετίζονται με την ημερομηνία και την ώρα. Ακόμη, στην Εικόνα 4.3 απεικονίζεται το DataFrame με χαρακτηριστικά που σχετίζονται με την ημερομηνία και την ώρα, καθώς και τα ιστορικά δεδομένα θερμοκρασίας, πίεσης και υγρασίας από την Ε.Μ.Υ.

	Time	avg_value	year	month	day	hour
0	2019-03-21 14:00:00	22.400000	2019	3	21	14
1	2019-03-21 15:00:00	22.800000	2019	3	21	15
2	2019-03-21 16:00:00	23.000000	2019	3	21	16
3	2019-03-21 17:00:00	22.900000	2019	3	21	17
4	2019-03-21 18:00:00	22.800000	2019	3	21	18
...	...	...	...	..	..	..
7521	2020-09-02 18:00:00	27.816667	2020	9	2	18
7522	2020-09-02 19:00:00	27.750000	2020	9	2	19
7523	2020-09-02 20:00:00	27.750000	2020	9	2	20
7524	2020-09-02 21:00:00	27.760000	2020	9	2	21
7525	2020-09-02 22:00:00	27.900000	2020	9	2	22

[7526 rows x 6 columns]

Σχήμα 4.2: Το νέο DataFrame με τα νέα χαρακτηριστικά που δημιουργήθηκαν.

Στη συνέχεια αφαιρέθηκε η μεταβλητή του χρόνου και τοποθετήθηκαν ως μεταβλητές εισόδου x οι μεταβλητές των νέων χαρακτηριστικών που σχετίζονται με το χρόνο και ως μεταβλητή εξόδου y οι μετρήσεις θερμοκρασίας, πίεσης και υγρασίας των δεδομένων. Στον Κώδικα 4.5 φαίνεται η διαδικασία διαχωρισμού των χαρα-

	Time	avg_value	Temp	Humid	Pressure	year	month	day	hour
0	2019-03-21 14:00:00	22.400000	15.500000	66.366611	1013.950012	2019	3	21	14
1	2019-03-21 15:00:00	22.800000	13.600000	41.960093	1027.500000	2019	3	21	15
2	2019-03-21 16:00:00	23.000000	15.200000	63.586129	1013.000000	2019	3	21	16
3	2019-03-21 17:00:00	22.900000	19.000000	44.431891	1012.250000	2019	3	21	17
4	2019-03-21 18:00:00	22.800000	9.000000	60.478589	1029.000000	2019	3	21	18
...	...	...	...	...	...	...	..	..	..
7521	2020-09-02 18:00:00	27.816667	24.700001	44.773493	1012.500000	2020	9	2	18
7522	2020-09-02 19:00:00	27.750000	28.449999	24.378388	1010.100006	2020	9	2	19
7523	2020-09-02 20:00:00	27.750000	28.449999	24.378388	1010.100006	2020	9	2	20
7524	2020-09-02 21:00:00	27.760000	22.500000	52.297075	1015.000000	2020	9	2	21
7525	2020-09-02 22:00:00	27.900000	22.200000	44.994762	1008.699982	2020	9	2	22

[7526 rows x 9 columns]

Σχήμα 4.3: Το νέο DataFrame με τα νέα χαρακτηριστικά που δημιουργήθηκαν με τα ιστορικά δεδομένα της Ε.Μ.Υ.

κτηριστικών σε ανεξάρτητες μεταβλητές εισόδου x και της εξαρτημένης μεταβλητής εξόδου y.

Κώδικας 4.5: Διαχωρισμός των χαρακτηριστικών σε ανεξάρτητες μεταβλητές εισόδου x και εξαρτημένης μεταβλητής εξόδου y.

```
x = df.drop(['Time', 'bath_avg_value'], axis=1)   #store
   independent variables into "x" vector


y = df.bath_avg_value.values    #store dependent variable into
   "y" vector
```

4.4.3 Κλιμάκωση Χαρακτηριστικών

Για την καλύτερη εφαρμογή και στη συνέχεια απόδοση των αλγοριθμικών μοντέλων μηχανικής μάθησης, έπρεπε να γίνει κλιμάκωση των χαρακτηριστικών (feature scaling). Η μέθοδος κλιμάκωσης που επιλέχθηκε είναι η τυποποιημένη κανονικοποίηση και επιτυγχάνεται μέσω της βιβλιοθήκης Scikit-learn και της συνάρτησης StandardScaler όπως παρουσιάζεται στον Κώδικα 4.6.

Κώδικας 4.6: Κλιμάκωση των χαρακτηριστικών με τη χρήση της μεθόδου StandardScaler.

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

```
x_scaled = scaler.fit_transform(x)
```

4.4.4 Διαχωρισμός δεδομένων σε σύνολα εκπαίδευσης και σύνολα δοκιμών

Στη συνέχεια διαχωρίστηκε το σύνολο δεδομένων σε υποσύνολα εκπαίδευσης και δοκιμών για τη δημιουργία αλγοριθμικών μοντέλων πρόβλεψης. Δηλαδή, χωρίστηκαν οι ανεξάρτητες μεταβλητές εισόδου x σε υποσύνολα εκπαίδευσης και δοκιμών και αντίστοιχα χωρίστηκε και η εξαρτημένη μεταβλητή y. Το ποσοστό δεδομένων που χρησιμοποιήθηκαν για εκπαίδευση ήταν το 80% του γενικού συνόλου δεδομένων, ενώ το ποσοστό δεδομένων για δοκιμές ήταν 20%, όπως φαίνεται στον Κώδικα 4.7. Αυτό επιτεύχθη με τη χρήση της βιβλιοθήκης Scikit-learn όπου δημιούργησε τα νέα υποσύνολα εκπαίδευσης (x_train, y_train) και (x_test, y_test) για τις δοκιμές.

Κώδικας 4.7: Διαχωρισμός του συνόλου δεδομένων σε υποσύνολα εκπαίδευσης και δοκιμών.

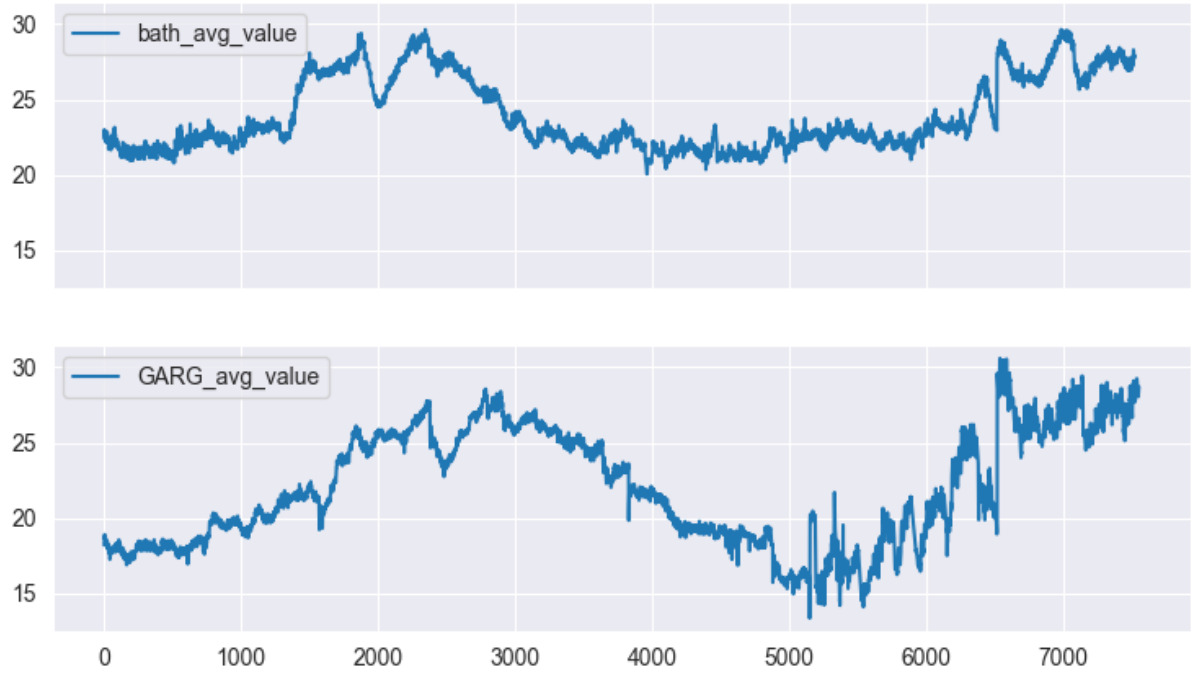
```
# split the dataset

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x_scaled, y
    , test_size=0.2, random_state=1)
```

4.5 Ανάλυση δεδομένων

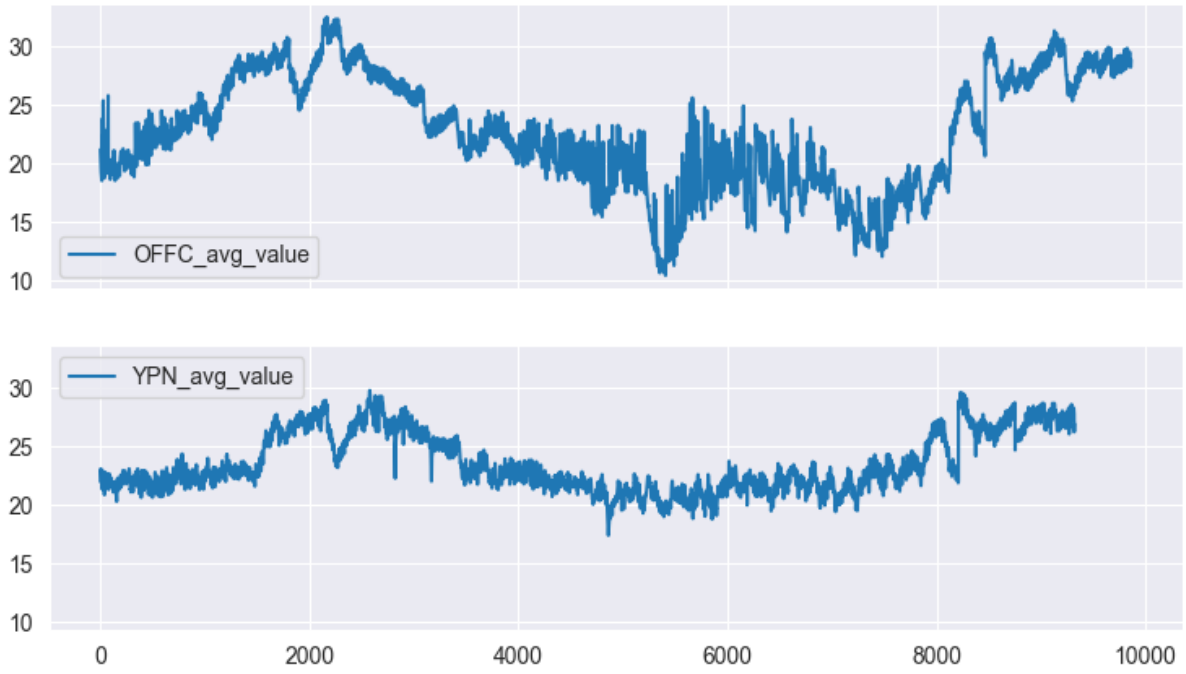
Αφού έγινε η συλλογή και επεξεργασία των δεδομένων, ήταν απαραίτητη η οπτικοποίηση των δεδομένων για να πραγματοποιηθεί μία ανάλυση μεταξύ των εξαρτημένων μεταβλητών εισόδου και της ανεξάρτητης μεταβλητής θερμοκρασίας. Χρησιμοποιήθηκε η βιβλιοθήκη matplotlib για να οπτικοποιηθεί η μεγάλη ποσότητα δεδομένων έτσι ώστε να παρουσιαστούν σε μία μορφή που να απεικονίζει διάφορα μοτίβα, τάσεις και συσχετίσεις που διαφορετικά δε θα ήταν εφικτό να παρατηρηθούν.

Αρχικά, οπτικοποιήθηκε το σύνολο δεδομένων για να ελεγχθεί εάν υπάρχουν κάποιες ασυνήθιστες μετρήσεις ή αν υπάρχει θόρυβος (white noise) στα δεδομένα χρονοσειρών, δηλαδή αν οι μεταβλητές είναι ανεξάρτητες και ταυτόσημα κατανεμημένες

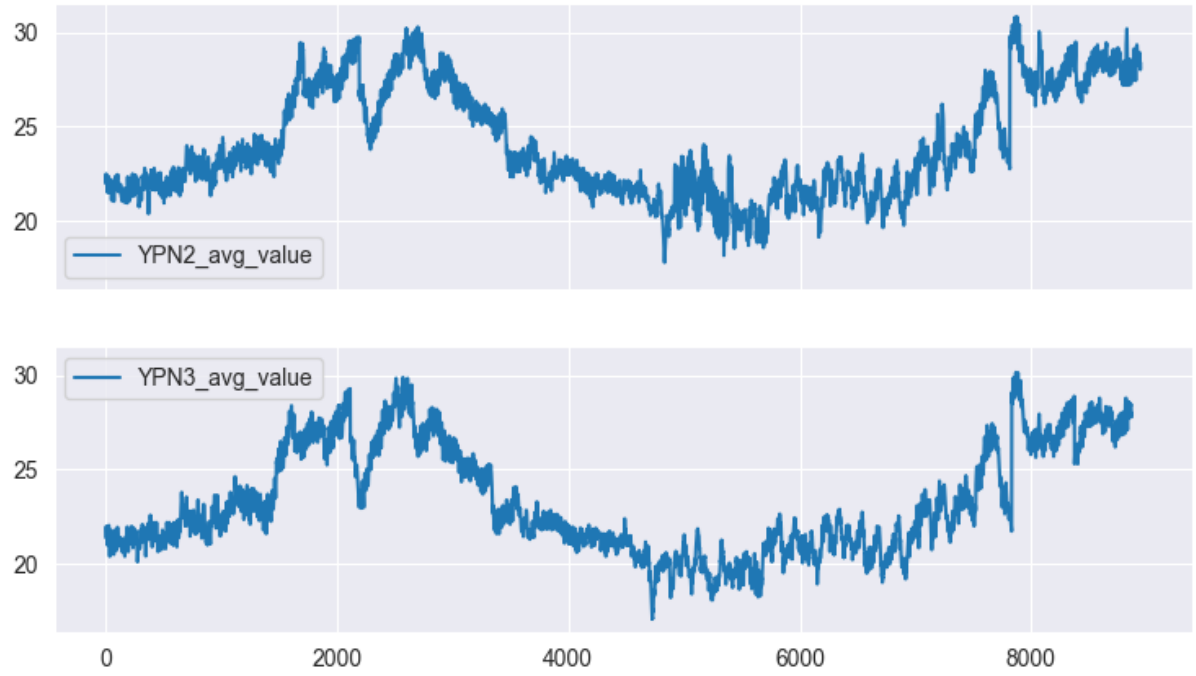
με μέση τιμή μηδέν. Αυτό σημαίνει ότι όλες οι τιμές έχουν την ίδια διακύμανση και μηδενική συσχέτιση με τις υπόλοιπες τιμές της χρονοσειράς και τέτοιες διακυμάνσεις δεδομένων δε θα μπορούσαν να ερμηνευτούν από τα μοντέλα παλινδρόμησης. Όπως φαίνεται στα δεδομένα κάθε χώρου δεν παρατηρούνται ασυνήθιστες τιμές ή κάποια σταθερή τάση μεταβολής των δεδομένων ανάλογα με το χρόνο. Ακόμη, στο Σχήμα 4.4 παρατηρείται ότι στα δεδομένα θερμοκρασίας του γκαράζ (GARG) η ελάχιστη τιμή είναι $13.4°$C καθώς δεν υπάρχει σύστημα θέρμανσης σε αυτό το χώρο. Το ίδιο παρατηρείται στο Σχήμα 4.5 στα δεδομένα θερμοκρασίας του γραφείου (OFFC) με ελάχιστη θερμοκρασία $10.4°$C όπως και στα δεδομένα για την εξωτερική αυλή (YARD) του Σχήματος 4.7 με ελάχιστη τιμή $-5.4°$C. Στα δεδομένα των υπόλοιπων χώρων παρατηρείται μία πιο ομαλή διαφοροποίηση των τιμών θερμοκρασίας όπως για παράδειγμα στο Σχήμα 4.6 όπου το εύρος των τιμών της θερμοκρασίας είναι μικρότερο, καθώς πρόκειται για θερμοκρασίες δύο παρόμοιων χώρων (Υπνοδωμάτιο 2, Υπνοδωμάτιο 3). Ακόμη, έγινε οπτικοποίηση της κατανομής των τιμών θερμοκρασίας για κάθε χώρο ξεχωριστά για να ελέγχθεί η ελάχιστη και η μέγιστη τιμή, καθώς και ποια τιμή συναντάται πιο συχνά σε κάθε χώρο. Όπως φαίνεται στο Σχήμα 4.8 οι θερμοκρασίες του μπάνιου και του υπνοδωματίου κυμαίνονται από $20°$C έως $30°$C ενώ του γκαράζ και του γραφείου από $10°$C έως $30°$C. Αντίστοιχα, στο Σχήμα 4.9 τα δύο υπνοδωμάτια και η κουζίνα έχουν παρόμοιο εύρος τιμών θερμοκρασίας από $17°$C έως $30°$C ενώ η εξωτερική αυλή έχει το μεγαλύτερο έυρος σε σχέση με τους υπόλοιπους χώρους, από $-5°$C έως $40°$C.



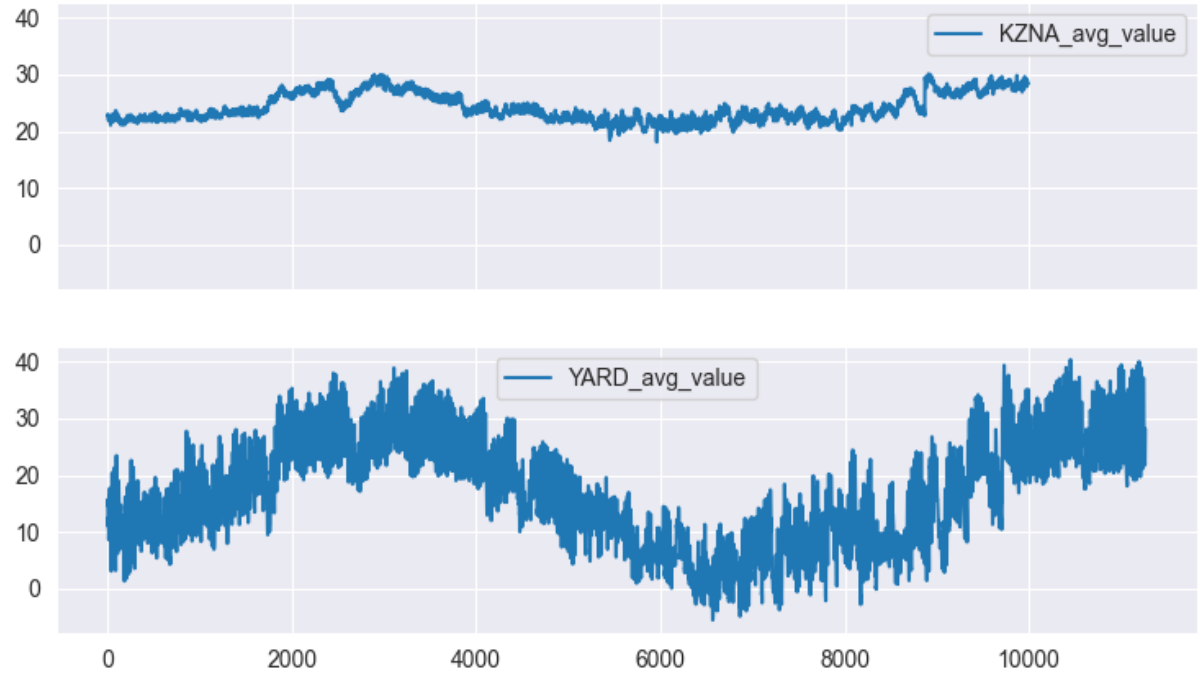
Σχήμα 4.4: Η οπτική απεικόνιση των δεδομένων θερμοκρασίας στον χώρο του μπάνιου και του γκαράζ.



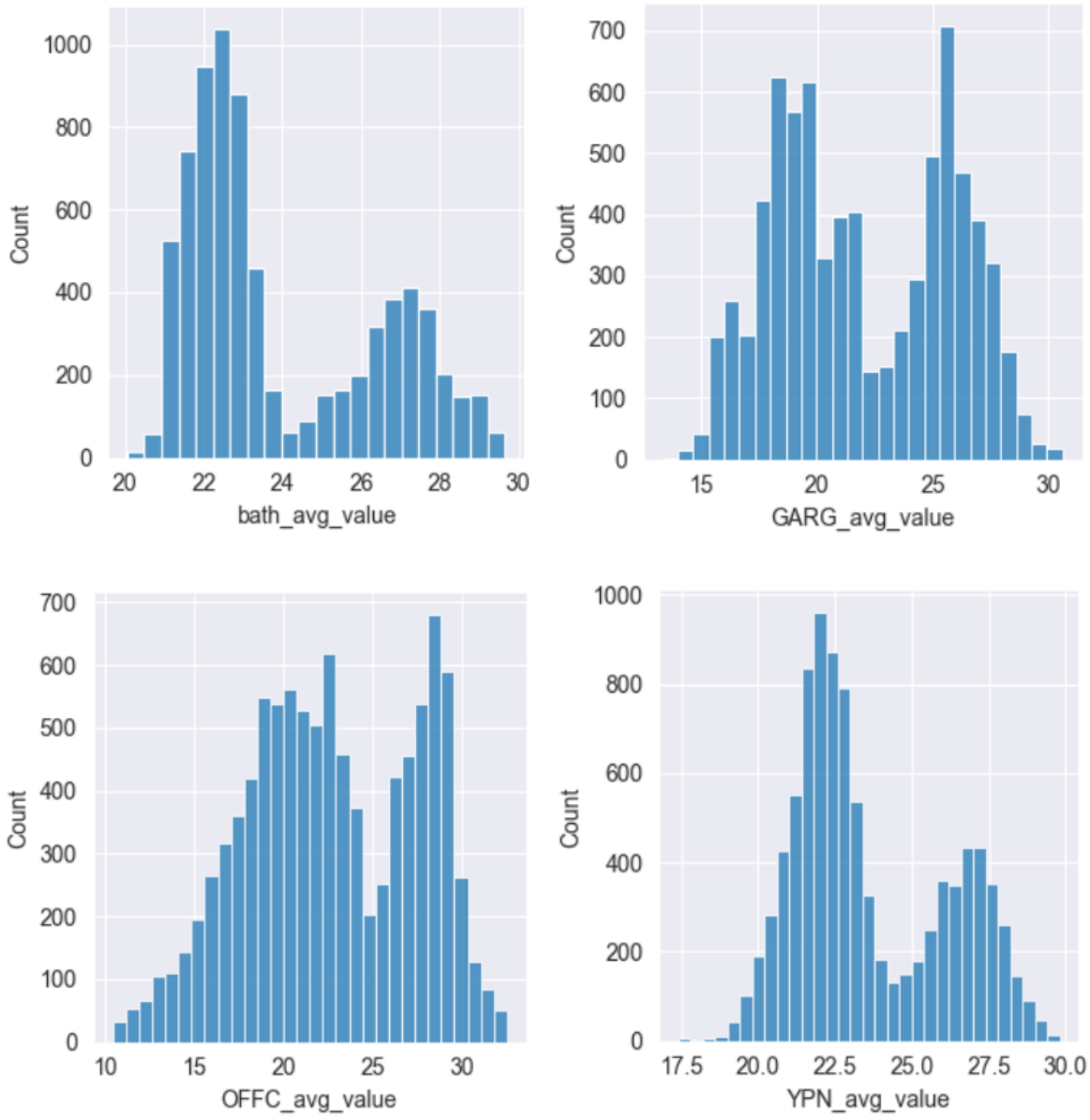
Σχήμα 4.5: Η οπτική απεικόνιση των δεδομένων θερμοκρασίας στον χώρο του γραφείου και του υπνοδωματίου.



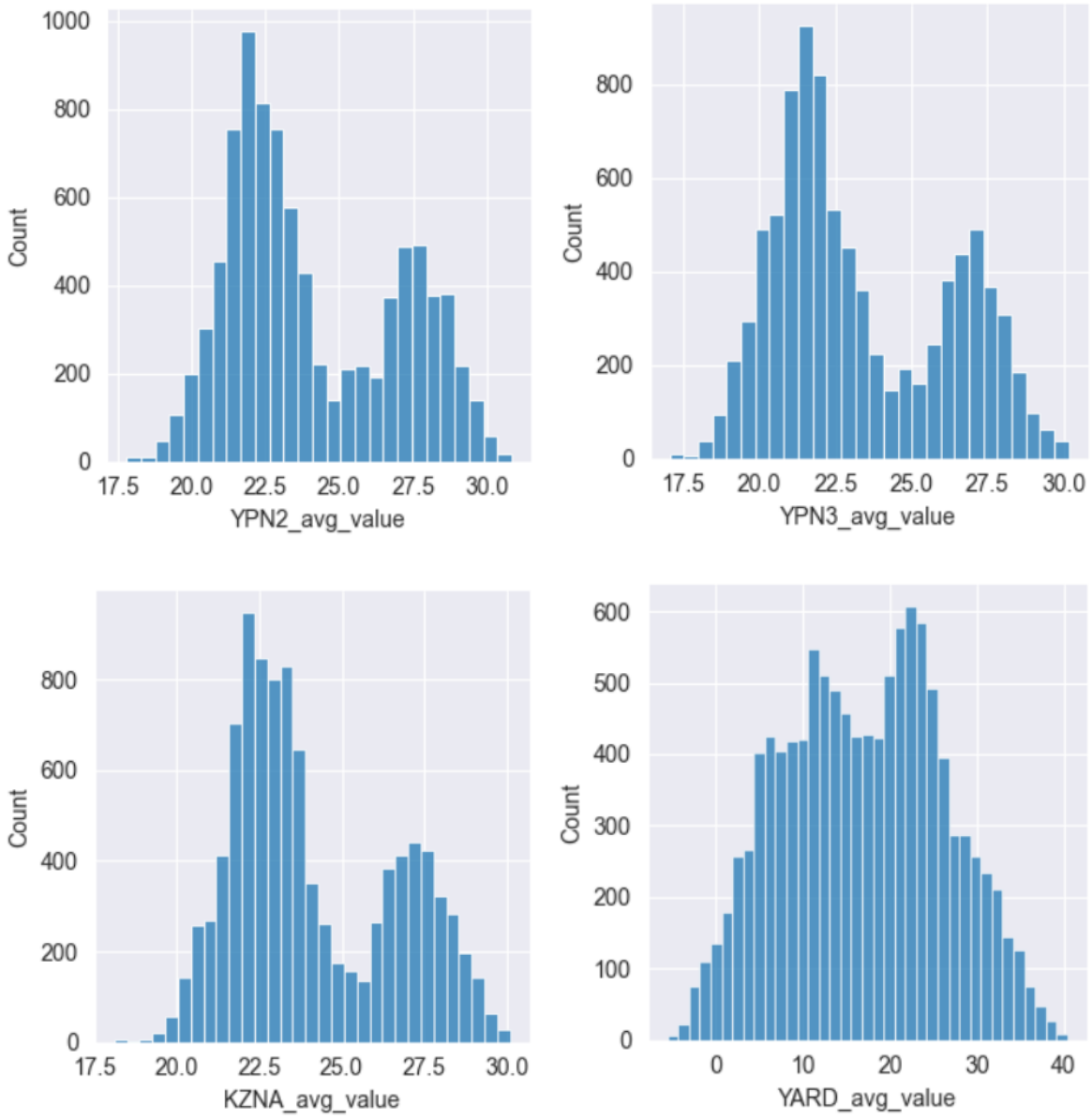
Σχήμα 4.6: Η οπτική απεικόνιση των δεδομένων θερμοκρασίας στον χώρο του 2ου και 3ου υπνοδωματίου.



Σχήμα 4.7: Η οπτική απεικόνιση των δεδομένων θερμοκρασίας στον χώρο της κουζίνας και της εξωτερικής αυλής.



Σχήμα 4.8: Η κατανομή των τιμών για τους χώρους του μπάνιου, γκαράζ, γραφείου και υπνοδωματίου.



Σχήμα 4.9: Η κατανομή των τιμών για τους χώρους του δεύτερου και τρίτου υπνοδωματίου, της κουζίνας και της εξωτερικής αυλής.

4.6 Μοντελοποίηση προβλήματος παλινδρόμησης

4.6.1 Περιγραφή

Η αντιμετώπιση του προβλήματος πρόβλεψης της θερμοκρασίας έγινε με τη χρήση αλγόριθμων παλινδρόμησης, με σκοπό να προβλέψουμε τις ακριβείς τιμές θερμοκρασίας σε κάθε χώρο. Χρησιμοποιήθηκαν οι εξής αλγόριθμοι μηχανικής μάθησης: γραμμική παλινδρόμηση, παλινδρόμηση Ridge, παλινδρόμηση Lasso, παλινδρόμηση με μηχανές διανυσμάτων υποστήριξης, δέντρα απόφασης, τυχαία δάση και MLP. Για την αξιολόγηση των μοντέλων παλινδρόμησης χρησιμοποιήθηκαν οι μετρήσεις του συντελεστή προσδιορισμού και της ρίζας του μέσου τετραγώνου σφάλματος. Στον Πίνακα 4.1 φαίνονται τα αποτελέσματα των αλγόριθμων μηχανικής μάθησης για τους οκτώ διαφορετικούς χώρους του σπιτιού χωρίς τη χρήση διασταυρούμενης επικύρωσης. Παρατηρείται ότι οι γραμμικοί αλγόριθμοι παλινδρόμησης όπως η γραμμική παλινδρόμηση, η παλινδρόμηση Ridge και Lasso δε γενικεύουν καλά το σύνολο δεδομένων και οι προβλέψεις τους είναι ανακριβείς. Αυτό ίσως οφείλεται στο γεγονός ότι τα νέα χαρακτηριστικά που δημιουργήθηκαν, δεν βοηθούν τα γραμμικά μοντέλα να αποκτήσουν ουσιαστική γνώση των τιμών θερμοκρασίας και συνεπώς να προβλέψουν με ακρίβεια τις μελλοντικές τιμές θερμοκρασίας. Ακόμη, το υψηλότερο RMSE παρατηρείται στο χώρο της εξωτερικής αυλής το οποίο είναι φυσιολογικό, καθώς αυτός ο χώρος είχε το μεγαλύτερο εύρος τιμών θερμοκρασίας. Τέλος, η μεγαλύτερη ακρίβεια επιτεύχθηκε με τους αλγόριθμους Random Forest και Decision Trees, δύο πολύ ισχυρούς αλγόριθμους που μπορούν να προβλέψουν με ακρίβεια την τιμή θερμοκρασίας και να γενικεύσουν το σύνολο δεδομένων με πιο αποδοτικό τρόπο.

Πίνακας 4.1: Οι μετρήσεις του R^2 και RMSE χωρίς διασταυρούμενη επικύρωση.

(Α) Μετρήσεις για τους πρώτους 4 χώρους του σπιτιού.

Αλγόριθμος	Μπάνιο		Γκαράζ		Γραφείο		Υπνοδωμάτιο	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Linear Reg	0.125	2.284	0.078	3.664	0.123	4.552	0.088	2.416
Ridge Reg	0.125	2.284	0.078	3.664	0.123	4.552	0.088	2.416
Lasso Reg	0.108	2.305	0.074	3.673	0.120	4.558	0.072	2.437
SVR	0.844	0.961	0.935	0.968	0.843	1.922	0.831	1.039
Decision Tree	0.995	0.171	0.993	0.297	0.990	0.476	0.989	0.270
Random Forest	0.996	0.148	0.994	0.285	0.992	0.433	0.991	0.235
MLP	0.956	0.508	0.963	0.729	0.910	1.455	0.914	0.738

(Β) Μετρήσεις για τους υπόλοιπους 4 χώρους του σπιτιού.

Αλγόριθμος	Υπνοδωμάτιο 2		Υπνοδωμάτιο 3		Κουζίνα		Εξωτερική Αυλή	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Linear Reg	0.090	2.728	0.058	2.829	0.078	2.363	0.105	8.701
Ridge Reg	0.090	2.728	0.058	2.829	0.078	2.363	0.105	8.701
Lasso Reg	0.084	2.738	0.045	2.849	0.067	2.377	0.106	8.699
SVR	0.863	1.057	0.854	1.110	0.824	1.032	0.877	3.218
Decision Tree	0.993	0.225	0.994	0.208	0.988	0.266	0.975	1.454
Random Forest	0.995	0.194	0.996	0.165	0.991	0.225	0.982	1.224
MLP	0.936	0.718	0.940	0.709	0.906	0.751	0.903	2.855

Στη συνέχεια, για τη βελτίωση των χαμηλών επιδόσεων των γραμμικών μοντέλων και την αποφυγή πιθανής υπερεκπαίδευσης στους υπόλοιπους αλγόριθμους, χρησιμοποιήθηκε η τεχνική της διασταυρούμενης επικύρωσης k-Fold με 10 πτυχώσεις. Τα αποτελέσματα που παρουσιάζονται στον Πίνακα 4.2 δείχνουν ότι η διασταυρούμενη επικύρωση δε βελτίωσε σημαντικά τα μοντέλα μηχανικής μάθησης σχετικά με το R^2 και RMSE. Υπήρξαν μικρές διαφοροποιήσεις στις μετρήσεις, ενώ τη μεγαλύτερη μεταβολή στην απόδοση είχε η παλινδρόμηση Lasso, η οποία φάνηκε πως δε γενίκευε καλά τα δεδομένα και η απόδοση του μοντέλου ήταν αρκετά πιο χαμηλή. Επίσης, οι αλγόριθμοι SVR και MLP παρουσίασαν μείωση στην ακρίβεια και στο RMSE σε όλους τους χώρους του σπιτιού. Η μείωση αυτή οφείλεται στο γεγονός πως αυτοί οι αλγόριθμοι παρά την ικανοποιητική απόδοσή τους, παρουσιάζουν αστάθεια εφαρμόζοντάς τους σε 10 πτυχώσεις του συνόλου δεδομένων. Αντιθέτως, οι αλγόριθμοι Decision Trees και Random Forest φάνηκε πως διατήρησαν τις επιδόσεις πρόβλεψης. Αυτό οφείλεται στο γεγονός ότι οι αλγόριθμοι Decision Trees και Random Forest είχαν την ίδια προβλεπτική ισχύ σε κάθε μία από τις 10 πτυχώσεις του συνόλου δεδομένων και έτσι διατηρήθηκε η υψηλή ακρίβεια και το RMSE στα ίδια επίπεδα.

Πίνακας 4.2: Οι μετρήσεις του R^2 και RMSE με διασταυρούμενη επικύρωση.

(Α) Μετρήσεις για τους πρώτους 4 χώρους του σπιτιού.

Αλγόριθμος	Μπάνιο		Γκαράζ		Γραφείο		Υπνοδωμάτιο	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Linear Reg	0.111	2.283	0.079	3.617	0.114	4.526	0.063	2.396
Ridge Reg	0.111	2.283	0.079	3.617	0.114	4.526	0.063	2.396
Lasso Reg	−0.002	2.425	−0.001	3.774	0.067	4.645	−0.002	2.479
SVR	0.837	0.975	0.926	1.022	0.833	1.958	0.824	1.036
Decision Tree	0.994	0.171	0.995	0.243	0.987	0.533	0.986	0.281
Random Forest	0.996	0.140	0.996	0.205	0.993	0.402	0.991	0.225
MLP	0.892	0.778	0.927	1.023	0.887	1.632	0.873	0.886

(Β) Μετρήσεις για τους υπόλοιπους 4 χώρους του σπιτιού.

Αλγόριθμος	Υπνοδωμάτιο 2		Υπνοδωμάτιο 3		Κουζίνα		Εξωτερική Αυλή	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Linear Reg	0.077	2.725	0.039	2.780	0.074	2.358	0.100	8.736
Ridge Reg	0.077	2.725	0.039	2.780	0.074	2.358	0.100	8.736
Lasso Reg	−0.001	2.840	−0.007	2.837	−0.003	2.457	0.075	8.856
SVR	0.851	1.094	0.855	1.077	0.826	1.022	0.875	3.252
Decision Tree	0.993	0.226	0.994	0.207	0.987	0.268	0.972	1.526
Random Forest	0.995	0.192	0.996	0.163	0.992	0.216	0.983	1.179
MLP	0.897	0.929	0.900	0.887	0.873	0.861	0.882	3.155

Καθώς η τεχνική της διασταυρούμενης επικύρωσης δε φάνηκε να επηρεάζει την απόδοση των μοντέλων σε σημαντικό βαθμό, εφαρμόστηκε η τεχνική της ρύθμισης

των υπερπαραμέτρων του κάθε μοντέλου μηχανικής μάθησης. Χρησιμοποιήθηκε το πλαίσιο λογισμικού Optuna για τη βελτίωση των υπερπαραμέτρων των αλγόριθμων μηχανικής μάθησης και οι υπερπαράμετροι που επιλέχθηκαν για να ρυθμιστούν, καθώς και το εύρος των τιμών κάθε υπερπαραμέτρου παρουσιάζονται στον Πίνακα 4.3. Μέσω του Optuna γίνονται 50 δοκιμές για κάθε αλγόριθμο, με διασταυρούμενη επικύρωση 10 πτυχώσεων σε κάθε δοκιμή και επιστρέφονται ως έξοδος οι μετρήσεις R^2 και RMSE, καθώς και οι τιμές των υπερπαραμέτρων που ορίστηκαν για βελτίωση.

Πίνακας 4.3: Οι υπερπαράμετροι που επιλέχθηκαν για βελτίωση.

Αλγόριθμος	Υπερπαράμετροι (Εύρος τιμών)	
Ridge Reg	alpha (0.0 - 2.0)	
Lasso Reg	alpha (0.0 - 2.0)	
SVR	C (0.0 - 10)	epsilon (0.0 - 2.0)
Decision Tree	max_depth (2 - 10)	min_samples_split (1 - 40)
Random Forest	n_estimators (10 - 50)	max_depth (2 - 10)
MLP	alpha (0.0 - 2.0)	max_iter (10 - 500)

Στον Πίνακα 4.4 παρουσιάζονται τα αποτελέσματα των μετρήσεων R^2 και RMSE για τους επτά χώρους του σπιτιού και το χώρο του γραφείου, με τη ρύθμιση των υπερπαραμέτρων για τη βελτίωση της απόδοσης κάθε μοντέλου. Τα αποτελέσματα της γραμμικής παλινδρόμησης παρέμειναν ίδια, καθώς δεν έγινε κάποια ρύθμιση υπερπαραμέτρων. Παρατηρήθηκε μία μικρή αύξηση της απόδοσης της παλινδρόμησης Ridge, ενώ μεγαλύτερη ήταν η βελτίωση στην παλινδρόμηση Lasso. Ακόμη, οι αλγόριθμοι Decision Tree και Random Forest διατήρησαν την ίδια απόδοση, ενώ σε μερικές περιπτώσεις μειώθηκε ελάχιστα. Αυτό οφείλεται στο γεγονός ότι το εύρος των υπερπαραμέτρων που επιλέχθηκε για αυτούς τους δύο αλγόριθμους ήταν μικρότερο από το απαιτούμενο, καταπιέζοντας έτσι τη μέγιστη προβλεπτική ισχύ των δύο αλγόριθμων. Τέλος, οι αλγόριθμοι SVR και MLP είχανε τη μεγαλύτερη βελτίωση στις μετρήσεις R^2 και RMSE δείχνοντας ότι η ρύθμιση των υπερπαραμέτρων είναι αρκετά σημαντικός παράγοντας για αυτούς τους αλγόριθμους μηχανικής μάθησης.

Στη συνέχεια, για την περαιτέρω βελτίωση των μετρήσεων αξιολόγησης των αλγοριθμικών μοντέλων, έγινε εισαγωγή των ιστορικών δεδομένων της Ε.Μ.Υ. που αφορούν τη θερμοκρασία, πίεση και υγρασία για την πόλη της Κοζάνης εκείνη

Πίνακας 4.4: Οι μετρήσεις του R^2 και RMSE με ρύθμιση υπερπαραμέτρων.

(Α) Μετρήσεις για τους πρώτους 4 χώρους του σπιτιού.

Αλγόριθμος	Μπάνιο		Γκαράζ		Γραφείο		Υπνοδωμάτιο	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Linear Reg	0.111	2.283	0.079	3.617	0.114	4.526	0.063	2.396
Ridge Reg	0.119	2.287	0.081	3.621	0.117	4.542	0.070	2.402
Lasso Reg	0.119	2.287	0.081	3.621	0.117	4.542	0.070	2.402
SVR	0.896	0.782	0.942	0.903	0.870	1.735	0.863	0.921
Decision Tree	0.989	0.250	0.991	0.339	0.972	0.796	0.963	0.478
Random Forest	0.991	0.220	0.993	0.302	0.980	0.674	0.970	0.426
MLP	0.956	0.547	0.960	0.753	0.928	1.281	0.923	0.702

(Β) Μετρήσεις για τους υπόλοιπους 4 χώρους του σπιτιού.

Αλγόριθμος	Υπνοδωμάτιο 2		Υπνοδωμάτιο 3		Κουζίνα		Εξωτερική Αυλή	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Linear Reg	0.077	2.725	0.039	2.780	0.074	2.358	0.100	8.736
Ridge Reg	0.080	2.731	0.042	2.794	0.075	2.359	0.098	8.746
Lasso Reg	0.080	2.731	0.042	2.794	0.075	2.359	0.098	8.745
SVR	0.892	0.935	0.886	0.961	0.861	0.912	0.896	2.962
Decision Tree	0.976	0.421	0.976	0.433	0.971	0.412	0.936	2.256
Random Forest	0.984	0.357	0.984	0.361	0.978	0.355	0.954	1.966
MLP	0.930	0.742	0.937	0.707	0.914	0.710	0.901	2.915

την περιόδο. Τα δεδομένα των τριών μετρήσεων χρησιμοποιήθηκαν ως δεδομένα εισόδου, ενώ ως δεδομένα εξόδου παρέμειναν τα δεδομένα θερμοκρασίας κάθε χώρου ξεχωριστά. Εφαρμόστηκε η τεχνική της ρύθμισης υπερπαραμέτρων μέσω του Optuna και κρατήθηκε η μέση τιμή των μετρήσεων μετά από διασταυρούμενη επικύρωση 10 πτυχώσεων. Τα αποτελέσματα του Πίνακα 4.5 δείχνουν ότι η ακρίβεια της γραμμικής παλινδρόμησης, της παλινδρόμησης Ridge και Lasso βελτιώθηκε σε μεγάλο βαθμό. Χαρακτηριστικό παράδειγμα είναι ότι αυτοί οι τρεις αλγόριθμοι, για το χώρο της εξωτερικής αυλής, είχαν $R^2 = 0.974\%$ αντί για $R^2 = 0.098\%$ που παρουσίασαν χωρίς τα δεδομένα της Ε.Μ.Υ, πλησιάζοντας τις επιδόσεις των υπόλοιπων αλγόριθμων. Αυτό συμβαίνει γιατί η θερμοκρασίες που παρουσιάζονται στο χώρο της αυλής είναι σχεδόν ίδιες με τις καταγεγραμμένες θερμοκρασίες της Ε.Μ.Υ που μετρήθηκαν εκείνη την περίοδο, καθώς πρόκειται για θερμοκρασίες εξωτερικού χώρου. Επίσης, σημαντικός παράγοντας σε αυτή την απόδοση είναι ο αριθμός των δεδομένων της εξωτερικής αυλής καθώς ήταν περισσότερα, περίπου 11.000 αντί για 8.000 που ήταν τα δεδομένα των υπόλοιπων χώρων. Στους υπόλοιπους χώρους του σπιτιού οι θερμοκρασίες διαφέρουν από αυτές των μετρήσεων της Ε.Μ.Υ, υπάρχουν και οι παράγοντες θέρμανσης και ψύξης του σπιτιού, γιαυτό δεν βελτιώθηκε η απόδοση των αλγόριθμων σε τόσο μεγάλο βαθμό. Ακόμη, βελτίωση της απόδο-

σης παρουσιάστηκε και στους αλγόριθμους SVR και MLP. Αντιθέτως, η απόδοση των αλγόριθμων Decision Tree και Random Forest μειώθηκε αισθητά, ιδιαίτερα στις μετρήσεις του RMSE. Αυτό οφείλεται στο γεγονός ότι το εύρος των τιμών των υπερπαραμέτρων αυτών των δύο αλγόριθμων ήταν μικρότερο από το απαιτούμενο για την επίτευξη του βέλτιστου αποτελέσματος. Έτσι, στις παραπάνω δοκιμές που δε ρυθμίστηκαν οι υπερπαράμετροι κάθε αλγόριθμου, οι δύο αλγόριθμοι Decision Tree και Random Forest είχαν μεγαλύτερη ακρίβεια στις προβλέψεις.

Πίνακας 4.5: Οι μετρήσεις του R^2 και RMSE με ρύθμιση υπερπαραμέτρων με τα δεδομένα της Ε.Μ.Υ.

(Α) Μετρήσεις για τους πρώτους 4 χώρους του σπιτιού.

Αλγόριθμος	Μπάνιο		Γκαράζ		Γραφείο		Υπνοδωμάτιο	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Linear Reg	0.676	1.376	0.773	1.792	0.804	2.127	0.691	1.374
Ridge Reg	0.678	1.380	0.774	1.792	0.806	2.128	0.702	1.359
Lasso Reg	0.678	1.380	0.774	1.792	0.806	2.128	0.702	1.359
SVR	0.915	0.706	0.957	0.778	0.921	1.356	0.901	0.779
Decision Tree	0.948	0.558	0.973	0.607	0.949	1.081	0.926	0.666
Random Forest	0.969	0.428	0.984	0.473	0.965	0.892	0.953	0.538
MLP	0.947	0.549	0.968	0.683	0.947	1.096	0.929	0.658

(Β) Μετρήσεις για τους υπόλοιπους 4 χώρους του σπιτιού.

Αλγόριθμος	Υπνοδωμάτιο 2		Υπνοδωμάτιο 3		Κουζίνα		Εξωτερική Αυλή	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Linear Reg	0.751	1.414	0.792	1.292	0.708	1.310	0.974	1.461
Ridge Reg	0.756	1.405	0.793	1.297	0.710	1.318	0.974	1.468
Lasso Reg	0.756	1.405	0.793	1.297	0.710	1.318	0.974	1.468
SVR	0.921	0.796	0.931	0.747	0.900	0.772	0.981	1.261
Decision Tree	0.944	0.668	0.937	0.713	0.942	0.579	0.978	1.349
Random Forest	0.963	0.544	0.963	0.547	0.962	0.472	0.983	1.170
MLP	0.944	0.689	0.950	0.638	0.933	0.635	0.981	1.247

4.6.2 Μοντελοποίηση με το στατιστικό μοντέλο ARIMA

Έχοντας τις μετρήσεις πρόβλεψης κάθε χώρου με διάφορες τεχνικές μοντελοποίησης, αποφασίστηκε η σύγκρισή τους με το στατιστικό μοντέλο ARIMA, ένα κλασικό τρόπο πρόβλεψης δεδομένων χρονοσειρών. Το σύνολο δεδομένων περιείχε τις μετρήσεις θερμοκρασίας κάθε χώρου ξεχωριστά και χωρίστηκε σε υποσύνολο εκπαίδευσης (80%) και υποσύνολο δοκιμών (20%). Αρχικά, έγινε ένας αυτοματοποιημένος έλεγχος για τις τιμές των παραμέτρων του μοντέλου ARIMA (p, d, q) μέσω της συνάρτησης auto_arima(). Η συνάρτηση αυτή κάνει δοκιμές του μοντέλου ARIMA με διαφορετικές παραμέτρους και χρησιμοποιεί ως μέτρο σύγκρισης

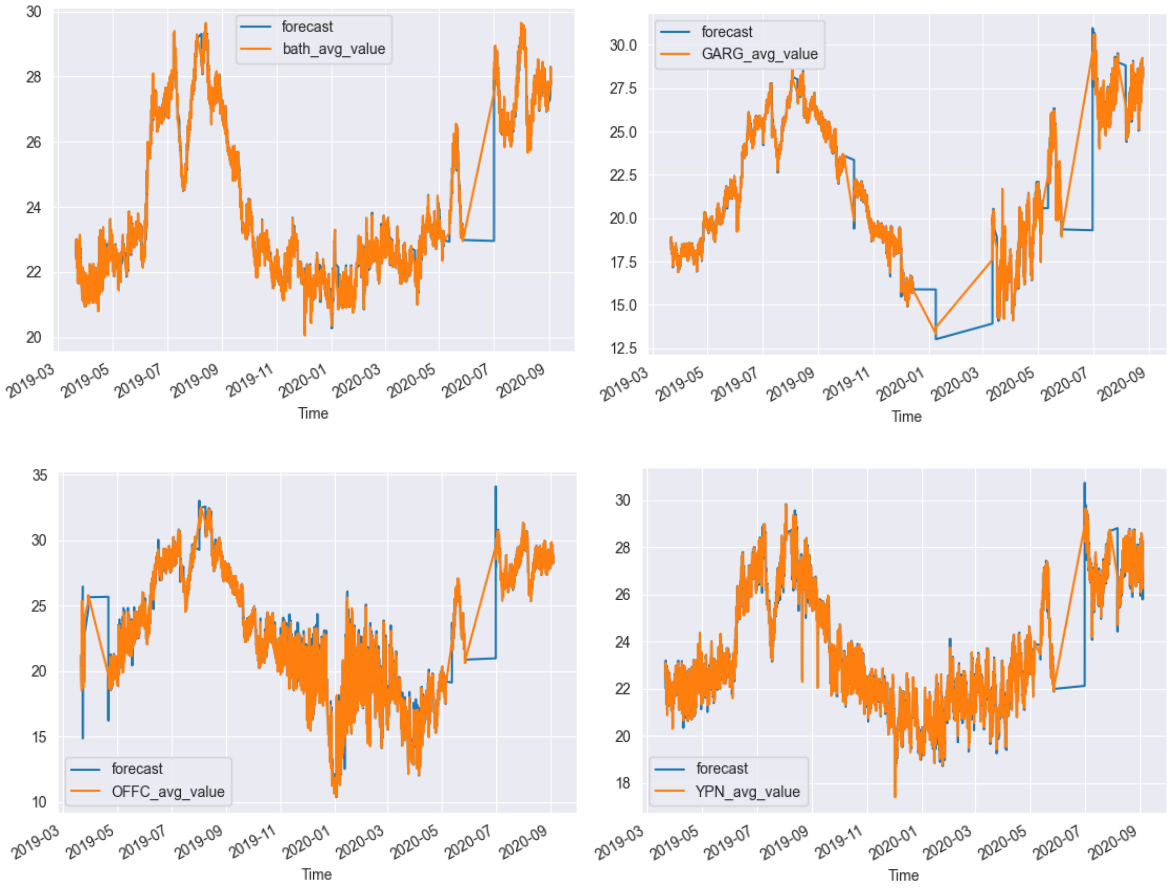
τη μέτρηση του κριτηρίου πληροφοριών Akaike (AIC). Η μέτρηση AIC είναι μία μαθηματική μέθοδος για την αξιολόγηση του πόσο καλά ένα μοντέλο ταιριάζει στα δεδομένα από τα οποία δημιουργήθηκε. Σκοπός της συνάρτησης auto_arima() είναι η εύρεση των παραμέτρων που επιστρέφουν τη χαμηλότερη μέτρηση AIC για κάθε σύνολο δεδομένων. Στη συνέχεια, εφαρμόζοντας το μοντέλο ARIMA με τις προτεινόμενες παραμέτρους, χρησιμοποιήθηκε η μέτρηση του RMSE για την αξιολόγηση της ακρίβειας κάθε μοντέλου. Στον Πίνακα 4.6 φαίνονται τα αποτελέσματα των μετρήσεων σε κάθε χώρο χρησιμοποιώντας το μοντέλο ARIMA, καθώς και οι βέλτιστοι παράμετροι που επιλέχθηκαν. Αναλύοντας τη στήλη των παραμέτρων του μοντέλου ARIMA, παρατηρείται ότι σε κάθε χώρο η τιμή του όρου της διαφοράς ισούται με 1 ($d = 1$), γεγονός που υποδηλώνει ότι όλα τα δεδομένα ήταν μη σταθερά και έπρεπε να γίνει έστω μία φορά η τεχνική της διαφοράς.

Πίνακας 4.6: Οι παράμετροι που επιλέχθηκαν και τα αποτελέσματα των μετρήσεων του μοντέλου ARIMA.

Χώρος Σπιτιού	(p, d, q)	RMSE	AIC
Μπάνιο	(2, 1, 3)	0.176	-6,940.3
Γκαράζ	(5, 1, 4)	0.371	154
Γραφείο	(2, 1, 5)	0.251	4,746.9
Υπνοδωμάτιο	(4, 1, 3)	0.242	-5,429.9
Υπνοδωμάτιο 2	(2, 1, 3)	0.214	-4,834.3
Υπνοδωμάτιο 3	(3, 1, 2)	0.223	-7,556.6
Κουζίνα	(3, 1, 2)	0.225	-3,321.5
Εξωτερική Αυλή	(3, 1, 3)	1.051	31,850

Τα αποτελέσματα των μετρήσεων φαίνονται ικανοποιητικά, καθώς οι τιμές του RMSE είναι πολύ χαμηλές και ξεπερνούν σε ακρίβεια τις τιμές των αλγόριθμων Decision Tree και Random Forest που είχαν τις καλύτερες επιδόσεις από τους αλγόριθμους μηχανικής μάθησης. Στη συνέχεια, έγινε η οπτικοποίηση των προβλεπόμενων τιμών θερμοκρασίας σε συνδυασμό με τις πραγματικές τιμές του συνόλου δεδομένων για να ελεγχθεί το πόσο καλά γενίκευσε το σύνολο δεδομένων το μοντέλο ARIMA. Στο Σχήμα 4.10 φαίνονται οι οπτικοποιήσεις για τους χώρους του μπάνιου, γκαράζ, γραφείου και του πρώτου υπνοδωματίου. Παρατηρείται ότι στο χώρο του μπάνιου που παρουσίασε το χαμηλότερο RMSE από όλους τους χώρους, οι τιμές των δεδομένων σχεδόν συμπίπτουν. Επίσης, τα δεδομένα που αφορούν το

χώρο του γκαράζ φαίνεται να διαφέρουν, κάτι το οποίο εξηγείται από το υψηλότερο RMSE που παρουσίασε αυτός ο χώρος σε σχέση με τους υπόλοιπους τρεις χώρους. Ακόμη, παρατηρείται ότι σε κάποιες χρονικές περιόδους, όπως για παράδειγμα την περίοδο Μαΐου-Ιουνίου 2020, το μοντέλο ARIMA δεν καταφέρνει να προβλέψει καλά τα δεδομένα. Αυτό μπορεί να οφείλεται στο γεγονός ότι εκείνη τη χρονική περίοδο υπάρχει σημαντική αύξηση στις τιμές της θερμοκρασίας σε σχέση με τις χαμηλότερες τιμές θερμοκρασίας των προηγούμενων μηνών.



Σχήμα 4.10: Οπτικοποίηση του συνόλου δεδομένων και των προβλεπόμενων δεδομένων του μοντέλου ARIMA για τους πρώτους τέσσερις χώρους.

Στο Σχήμα 4.11 φαίνονται οι οπτικοποιήσεις για τους χώρους του δεύτερου και τρίτου υπνοδωματίου, της κουζίνας και της εξωτερικής αυλής. Παρατηρείται ότι στα δύο υπνοδωμάτια όπου το RMSE ήταν το χαμηλότερο από τους τέσσερις αυτούς χώρους, τα δεδομένα συμπίπτουν με αυτά των πραγματικών μετρήσεων. Το χαμηλό RMSE και το γεγονός ότι συμπίπτουν οι προβλεπόμενες τιμές με τις πραγματικές τιμές οφείλεται στο ότι τα υπνοδωμάτια διατηρούν πιο σταθερές θερμοκρασίες από τους υπόλοιπους χώρους, και παρουσιάζουν το μικρότερο εύρος τιμών θερμοκρα-

σίας. Αντιθέτως, στο χώρο της αυλής παρατηρείται μεγάλη διακύμανση στις τιμές των δεδομένων, κάτι το οποίο είναι λογικό καθώς τα δεδομένα αυτού του χώρου είχαν τη μεγαλύτερη διακύμανση και παρουσίασαν το μεγαλύτερο RMSE. Κρίνοντας από τις οπτικοποιήσεις και των οκτώ χώρων, παρατηρείται ότι η περίοδος Μαΐου-Ιουλίου είχε τη μεγαλύτερη απότομη αλλαγή θερμοκρασίας κάτι το οποίο είναι φυσιολογικό καθώς βρίσκεται στους μήνες αλλαγής δύο εποχών.

Σχήμα 4.11: Οπτικοποίηση του συνόλου δεδομένων και των προβλεπόμενων δεδομένων του μοντέλου ARIMA για τους υπόλοιπους τέσσερις χώρους.

4.6.3 Διαγράμματα ισοτιμίας

Για την καλύτερη απεικόνιση και σύγκριση αποτελεσμάτων, παρουσιάζονται τα διαγράμματα ισοτιμίας κάθε χώρου με τις διαφορετικές τεχνικές μοντελοποίησης που χρησιμοποιήθηκαν. Ο άξονας x περιέχει τις πραγματικές μετρήσεις θερμοκρασίας που καταγράφηκαν από τους αισθητήρες κάθε χώρου, ενώ ο άξονας y περιέχει τις τιμές που προβλέφθηκαν από κάθε αλγόριθμο. Η ευθεία $y = x$ απεικονίζει τα σημεία που οι πραγματικές τιμές και οι προβλεπόμενες τιμές συμπίπτουν και όσο πιο

κοντά βρίσκονται τα σημεία προς την ευθεία, τόσο μεγαλύτερη είναι η ακρίβεια κάθε αλγοριθμικού μοντέλου. Ακόμη, στους άξονες x και y κάθε διαγράμματος ισοτιμίας, μπορεί να παρατηρηθεί η ελάχιστη και η μέγιστη θερμοκρασία που αναμένεται σε κάθε χώρο ξεχωριστά.

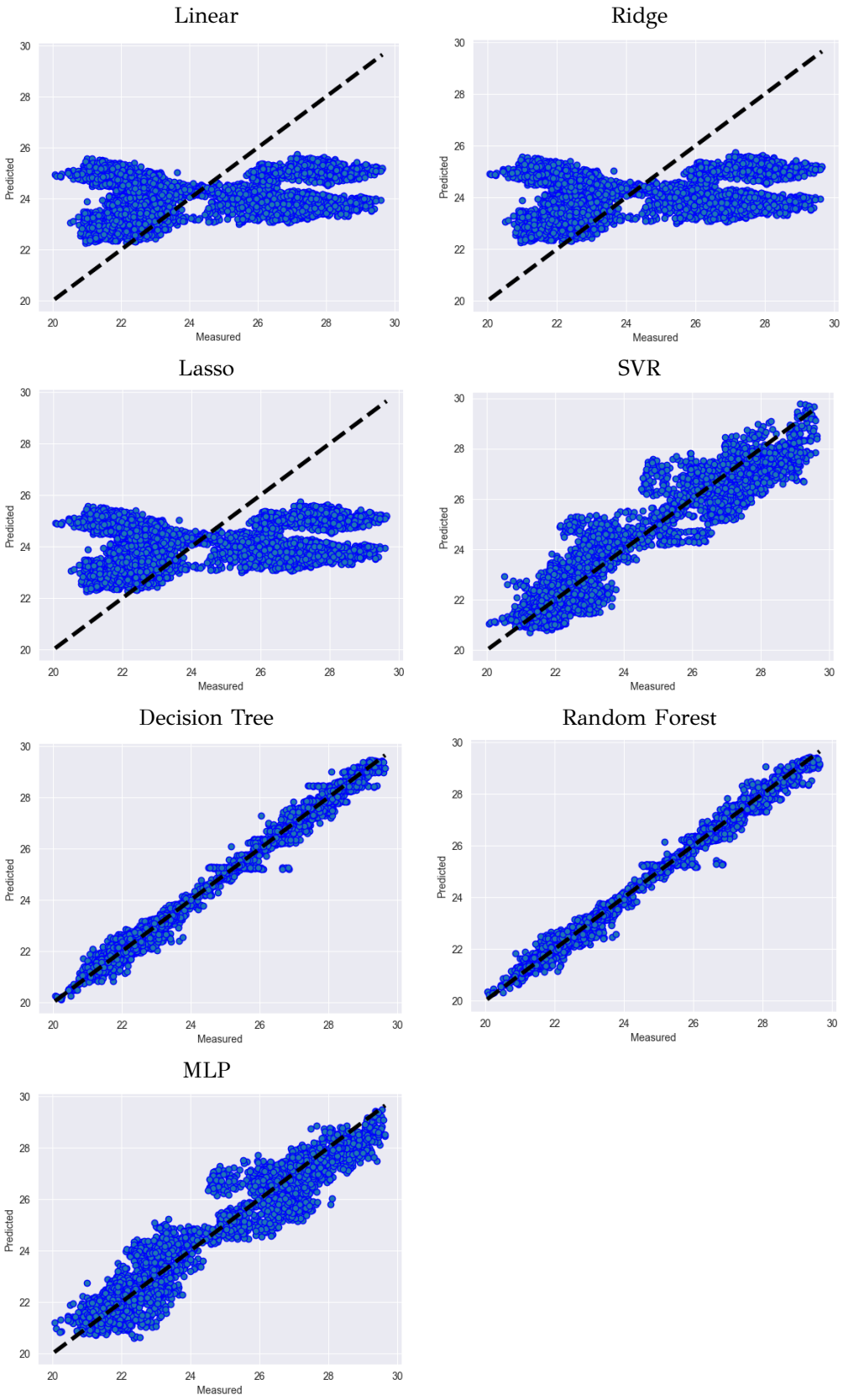
Διαγράμματα ισοτιμίας με τα δεδομένα του χρόνου ως είσοδο.

Παρακάτω παρουσιάζονται τα διαγράμματα ισοτιμίας (Σχήματα 4.12, 4.13, 4.14, 4.15, 4.16, 4.17, 4.18, 4.19) για κάθε χώρο ξεχωριστά, έπειτα από διασταυρούμενη επικύρωση 10 πτυχώσεων και ρύθμιση των υπερπαραμέτρων μέσω του Optuna. Είναι εμφανής η διαφορά των γραμμικών αλγόριθμων σε σχέση με τους υπόλοιπους σχετικά με το πόσο κοντά βρίσκονται οι τιμές θερμοκρασίας στη γραμμή $y = x$. Στους γραμμικούς αλγόριθμους οι προβλεπόμενες θερμοκρασίες σε σχέση με τις πραγματικές μετρήσεις θερμοκρασίας απέχουν αρκετά από την ευθεία $y = x$ δείχνοντας πως οι μετρήσεις είναι ανακριβείς. Επίσης, οι αλγόριθμοι SVR και MLP παρουσιάζουν ικανοποιητική ακρίβεια και μπορούν να θεωρηθούν ως αξιόπιστα μοντέλα πρόβλεψης, αν και υπάρχουν αρκετά σημεία δεδομένων που φαίνεται η απόκλιση της πρόβλεψης από τις πραγματικές τιμές θερμοκρασίας. Τέλος, φαίνεται η μεγάλη ακρίβεια που παρουσιάζουν οι αλγόριθμοι Decision Tree και Random Forest, καθώς οι προβλεπόμενες τιμές θερμοκρασίας έχουν σχεδόν τις ίδιες τιμές με αυτές που μετρήθηκαν από τους αισθητήρες. Αυτό έχει ως αποτέλεσμα όλα τα σημεία δεδομένων να εφάπτονται σχεδόν με την ευθεία $y = x$.

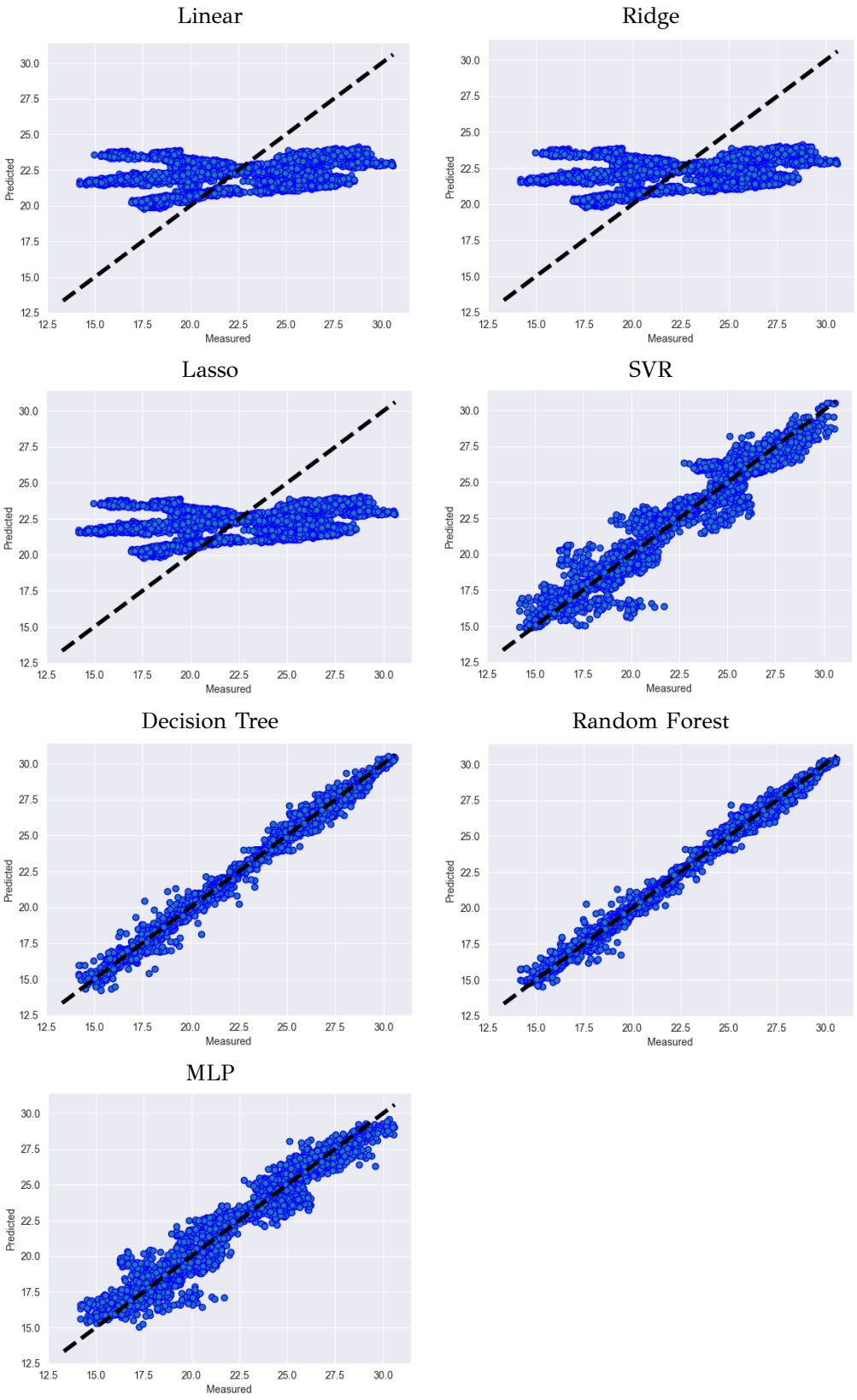
Διαγράμματα ισοτιμίας με τα δεδομένα της Ε.Μ.Υ. ως είσοδο.

Στη συνέχεια παρουσιάζονται τα διαγράμματα ισοτιμίας (Σχήματα 4.20, 4.21, 4.22, 4.23, 4.24, 4.25, 4.26, 4.27) έχοντας ως είσοδο τα δεδομένα της θερμοκρασίας, πίεσης και υγρασίας τα οποία ελήφθησαν από την Ε.Μ.Υ. με σκοπό τη βελτίωση της ακρίβειας κάθε αλγόριθμου. Προκειμένου να είναι ευδιάκριτος ο διαχωρισμός τους από τα υπόλοιπα διαγράμματα, απεικονίζονται με διαφορετικό χρώμα. Παρατηρείται ότι οι αλγόριθμοι γραμμικής παλινδρόμησης, παλινδρόμησης Ridge και παλινδρόμησεη Lasso βελτίωσαν τις επιδόσεις τους σε μεγάλο βαθμό και τα προβλεπόμενα δεδομένα πλησιάζουν αρκετά τις τιμές των πραγματικών δεδομένων που μετρήθηκαν. Χαρακτηριστικό είναι το διάγραμμα ισοτιμίας του Σχήματος 4.27 όπου

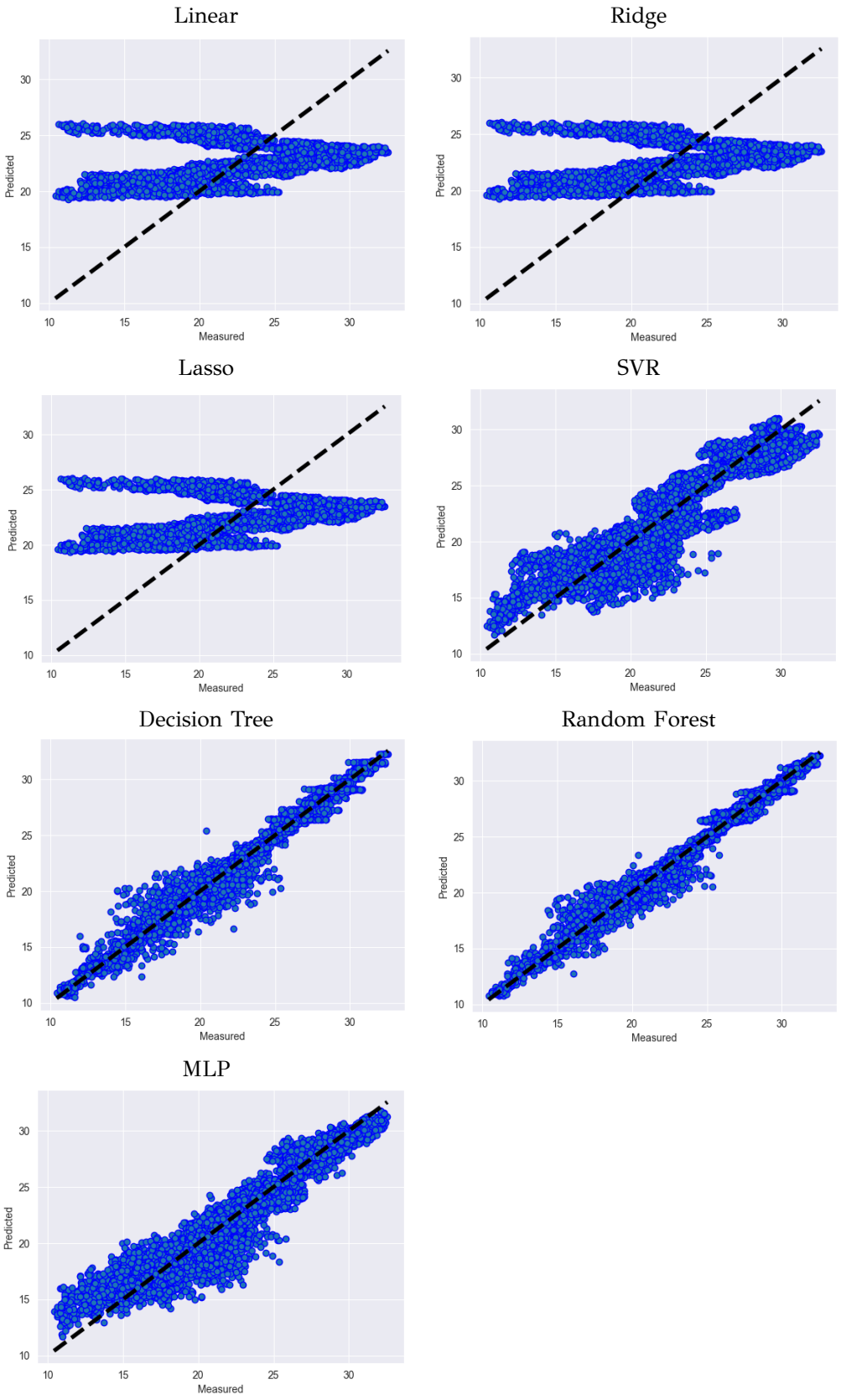
φαίνεται η βέλτιστη απόδοση των γραμμικών μοντέλων που επιτεύχθηκε σε σχέση με τους υπόλοιπους χώρους. Επίσης, φαίνεται πως παρά την υψηλή απόδοση και των επτά αλγόριθμων μηχανικής μάθησης, υπάρχουν αρκετά σημεία δεδομένων που έχουν αποτυχημένη πρόβλεψη σε σχέση με τις μετρήσεις. Αυτό οφείλεται στο γεγονός ότι ο χώρος της αυλής έχει τη μεγαλύτερη διακύμανση στις τιμές θερμοκρασίας που μετρήθηκαν και αυτό οδηγεί τους αλγόριθμους σε μερικές λάθος προβλέψεις. Ακόμη παρατηρείται η ελάχιστη αλλά σημαντική βελτίωση των αλγόριθμων SVR και MLP, καθώς και η ελάχιστη μείωση της ακρίβειας των αλγόριθμων Decision Tree και Random Forest. Οι τέσσερις αλγόριθμοι SVR, Decision Trees, Random Forest και MLP φαίνεται να έχουν παρόμοιες μετρήσεις ακρίβειας πρόβλεψης, καθιστώντας τις προβλέψεις τους ως ισχυρές και αξιόπιστες.



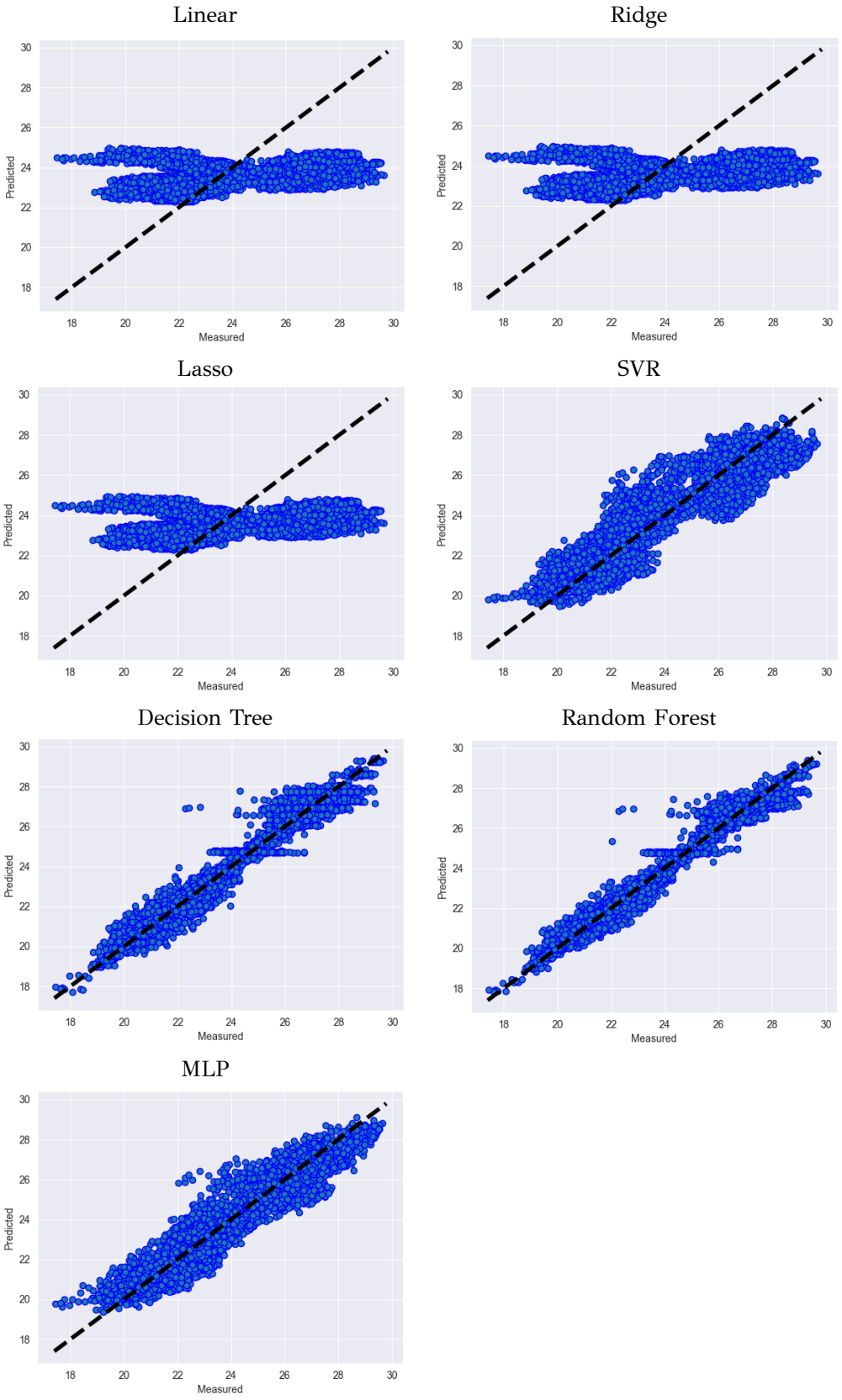
Σχήμα 4.12: Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του μπάνιου.



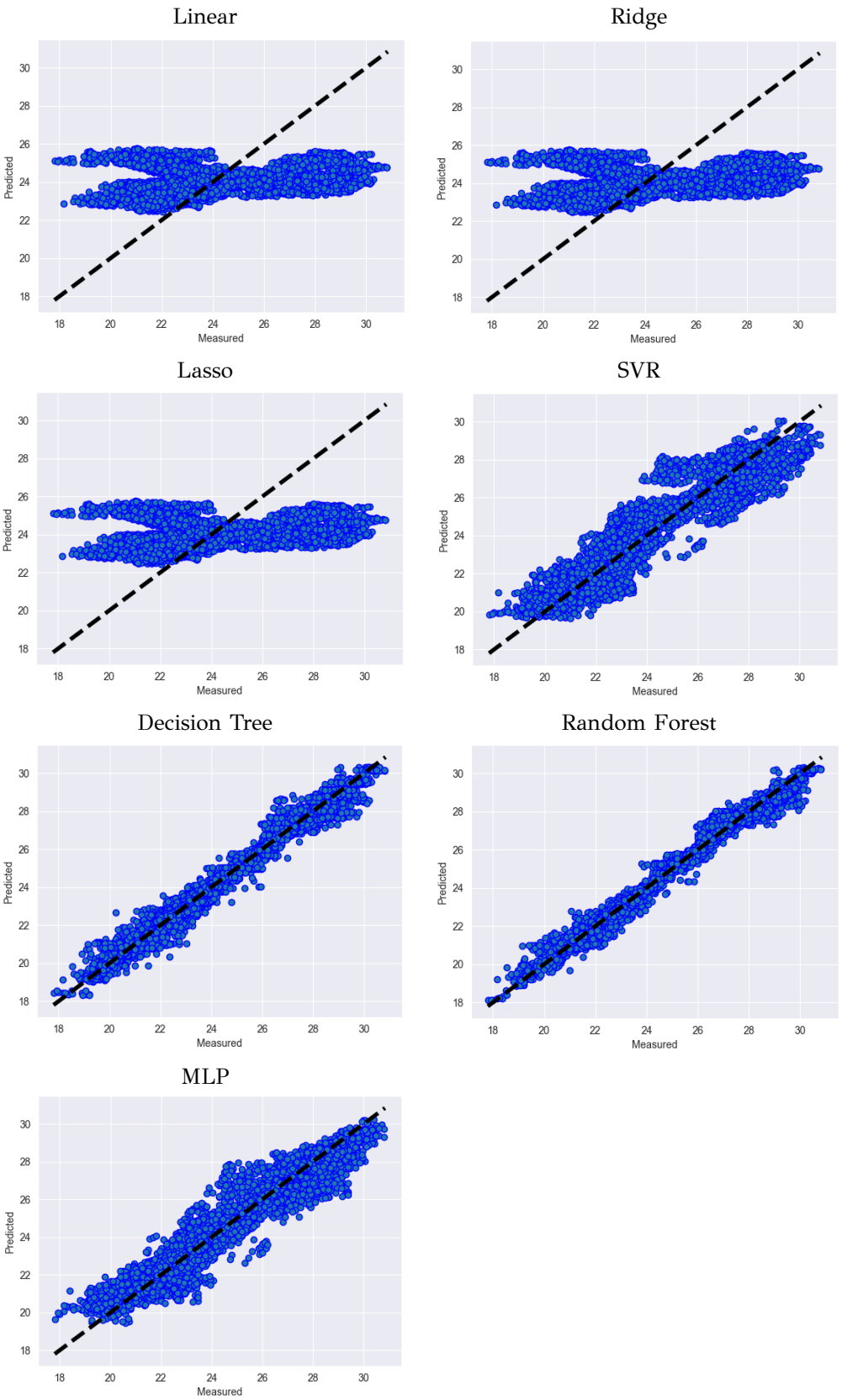
Σχήμα 4.13: Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του γκαράζ.



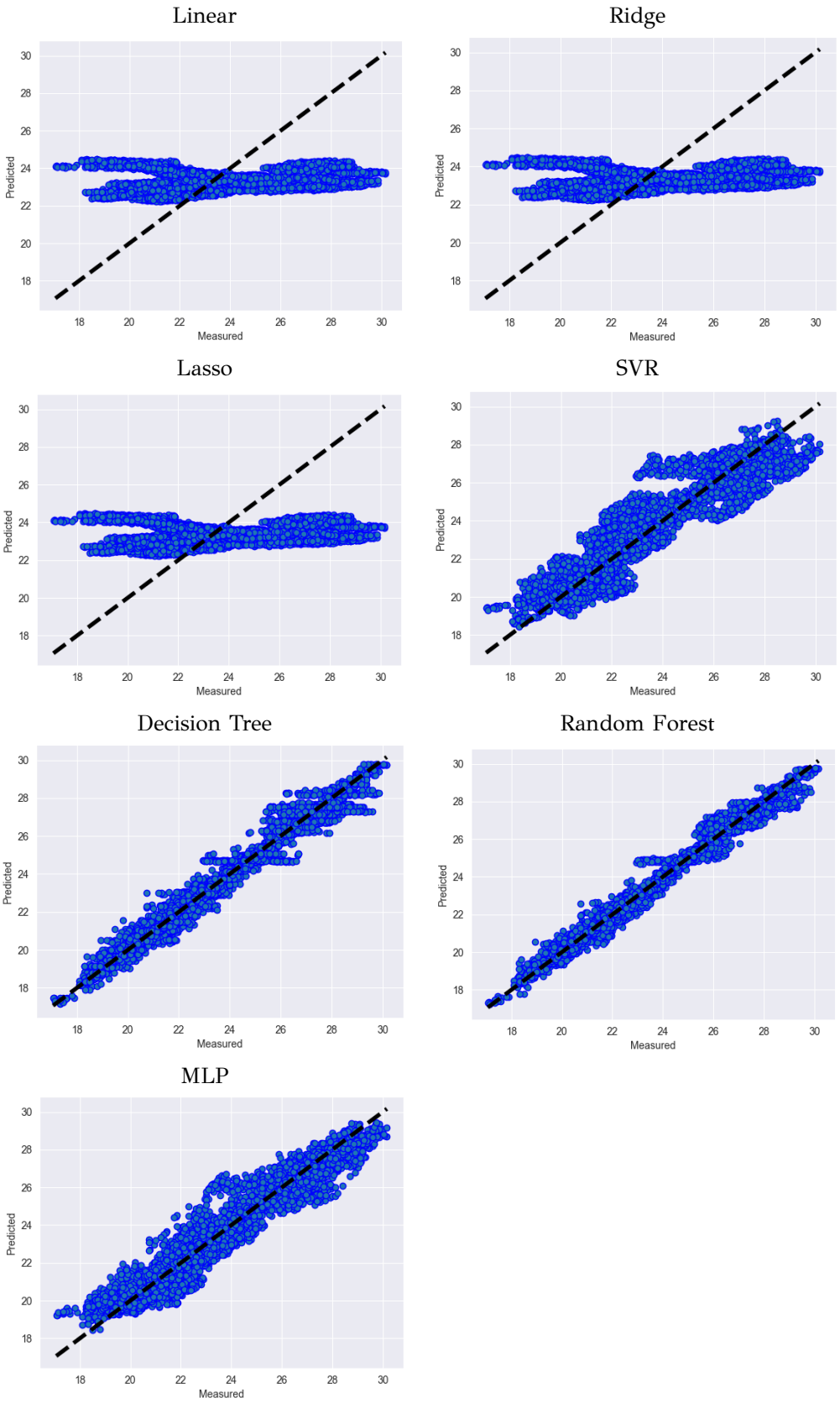
Σχήμα 4.14: Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του γραφείου.



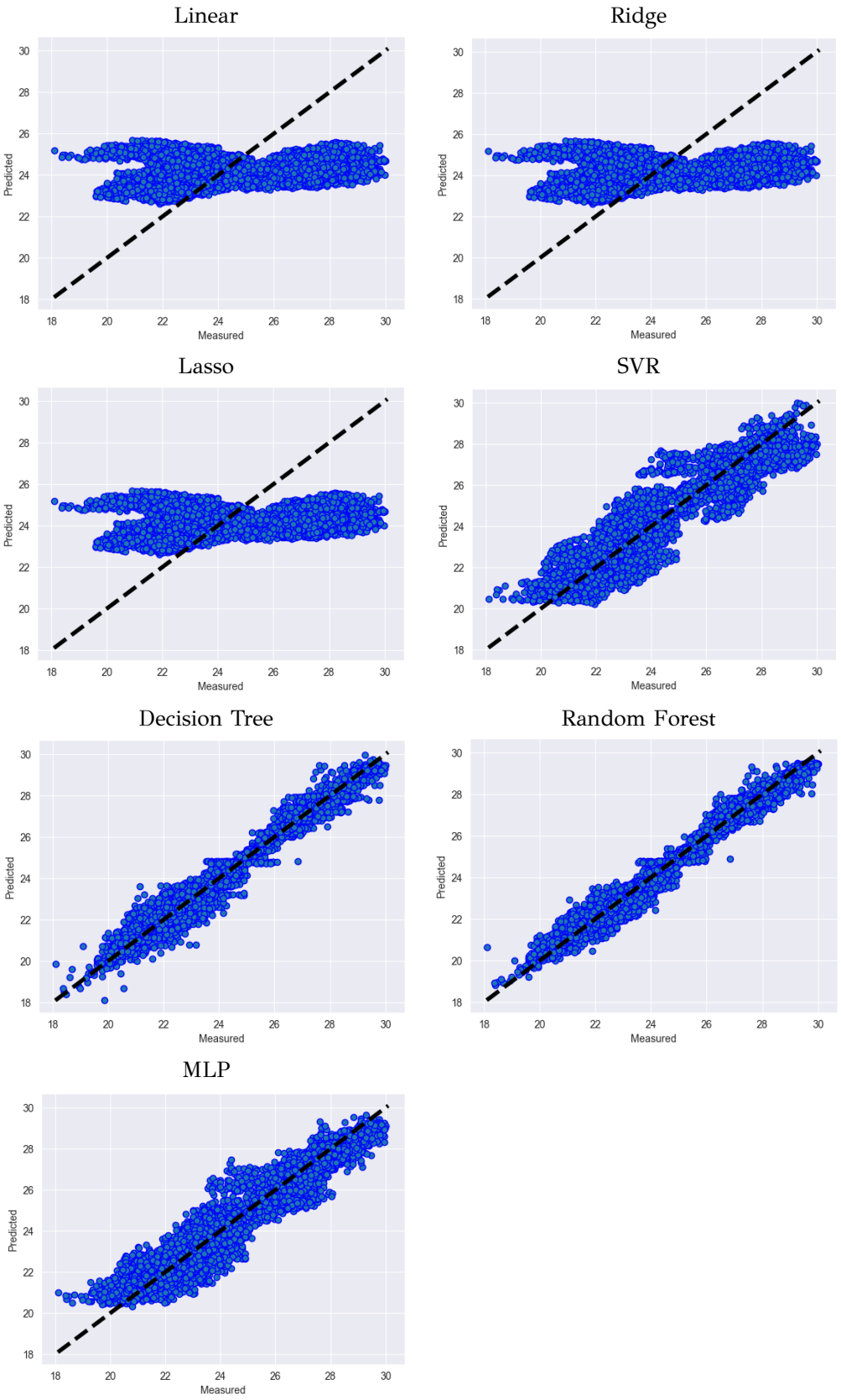
Σχήμα 4.15: Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του υπνοδωματίου.



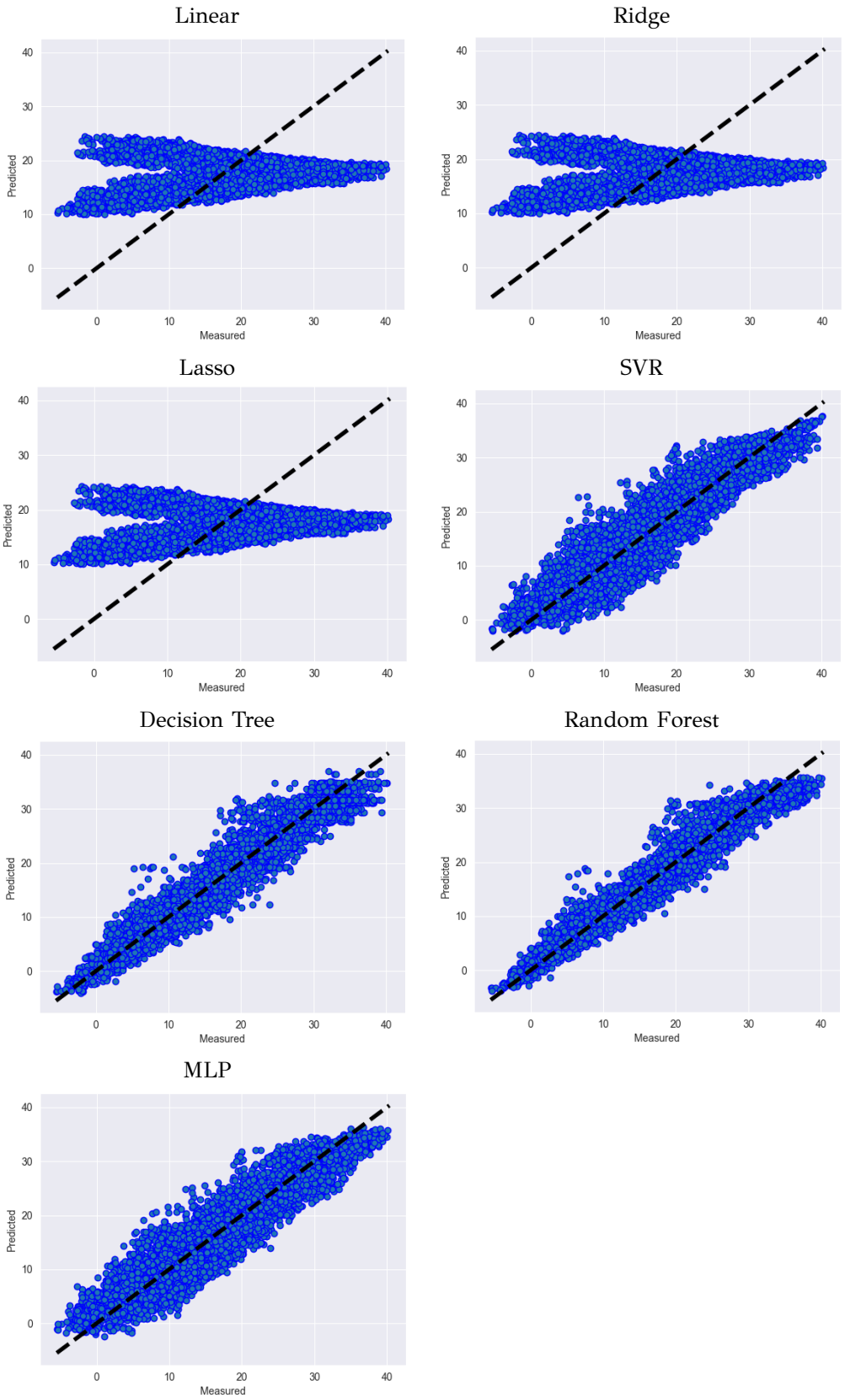
Σχήμα 4.16: Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του δεύτερου υπνοδωματίου.



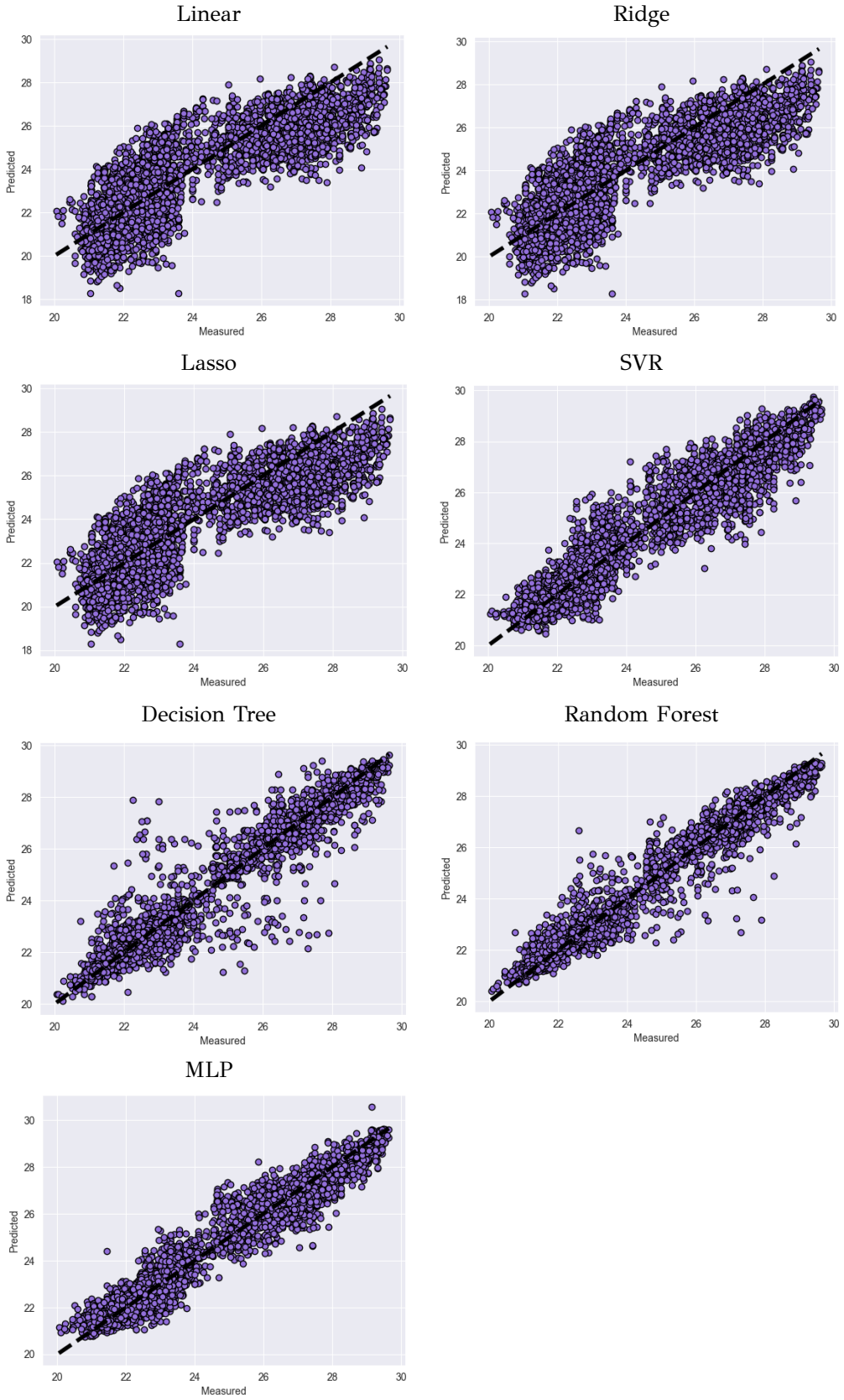
Σχήμα 4.17: Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο του τρίτου υπνοδωματίου.



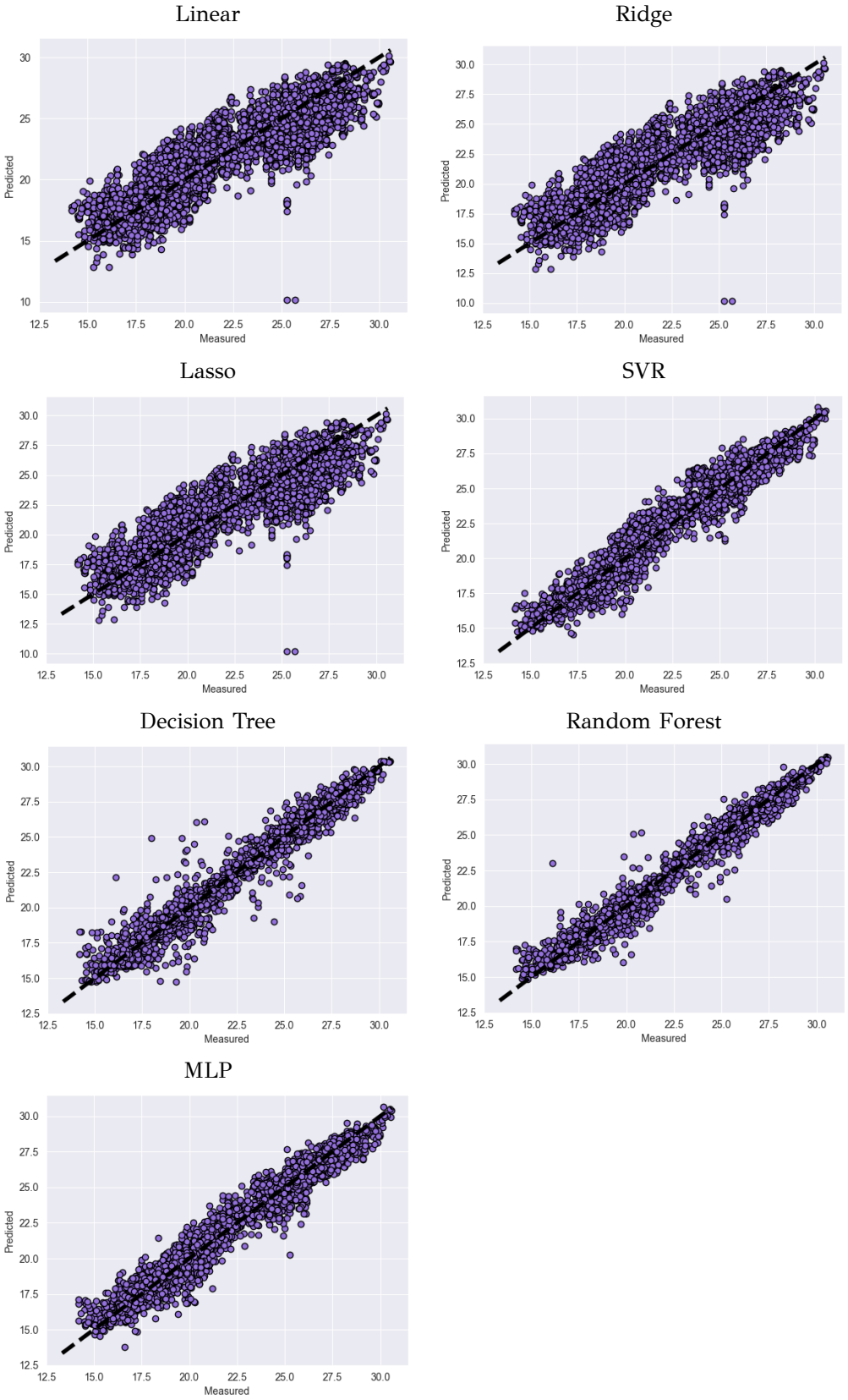
Σχήμα 4.18: Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο της κουζίνας.



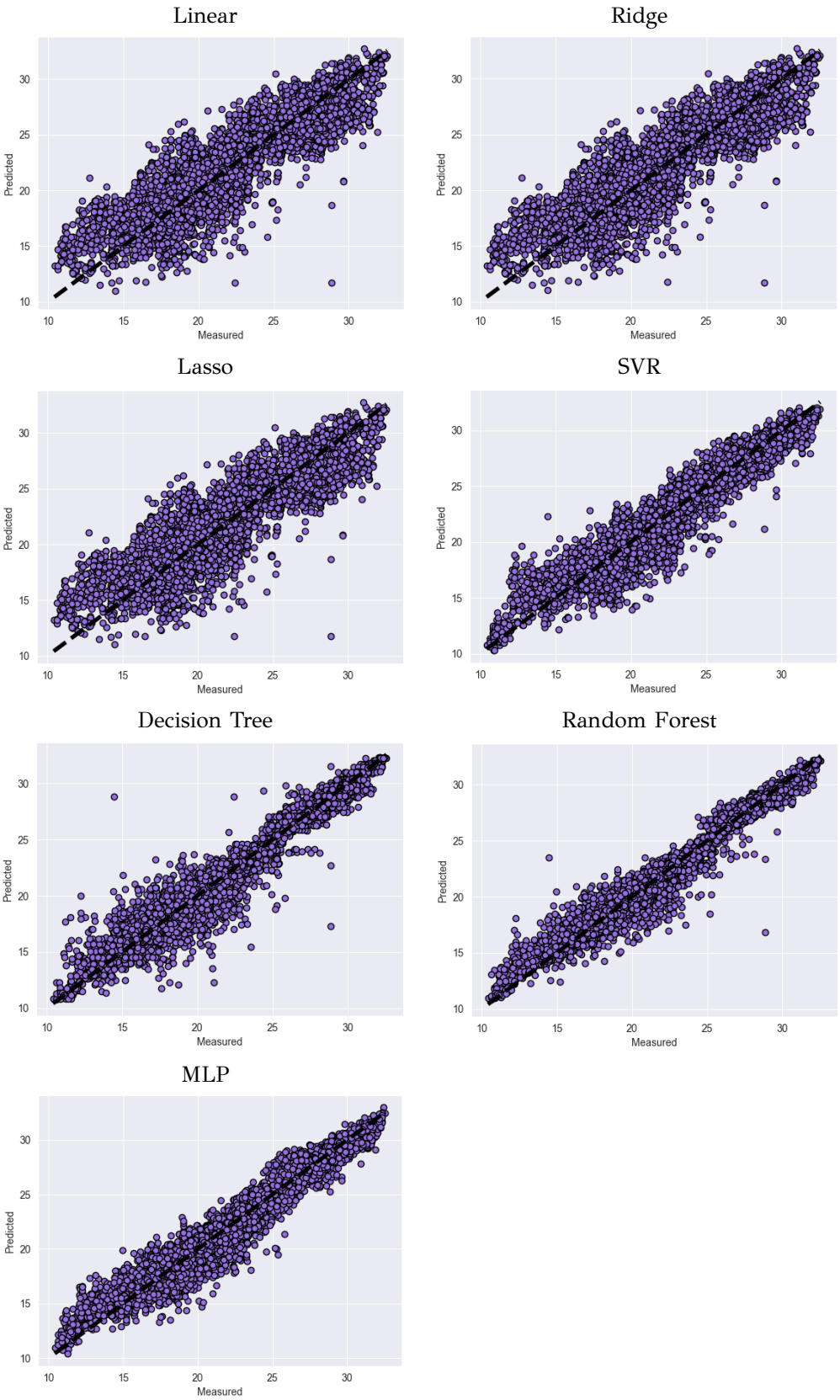
Σχήμα 4.19: Διαγράμματα ισοτιμίας με ρύθμιση υπερπαραμέτρων για το χώρο της αυλής.



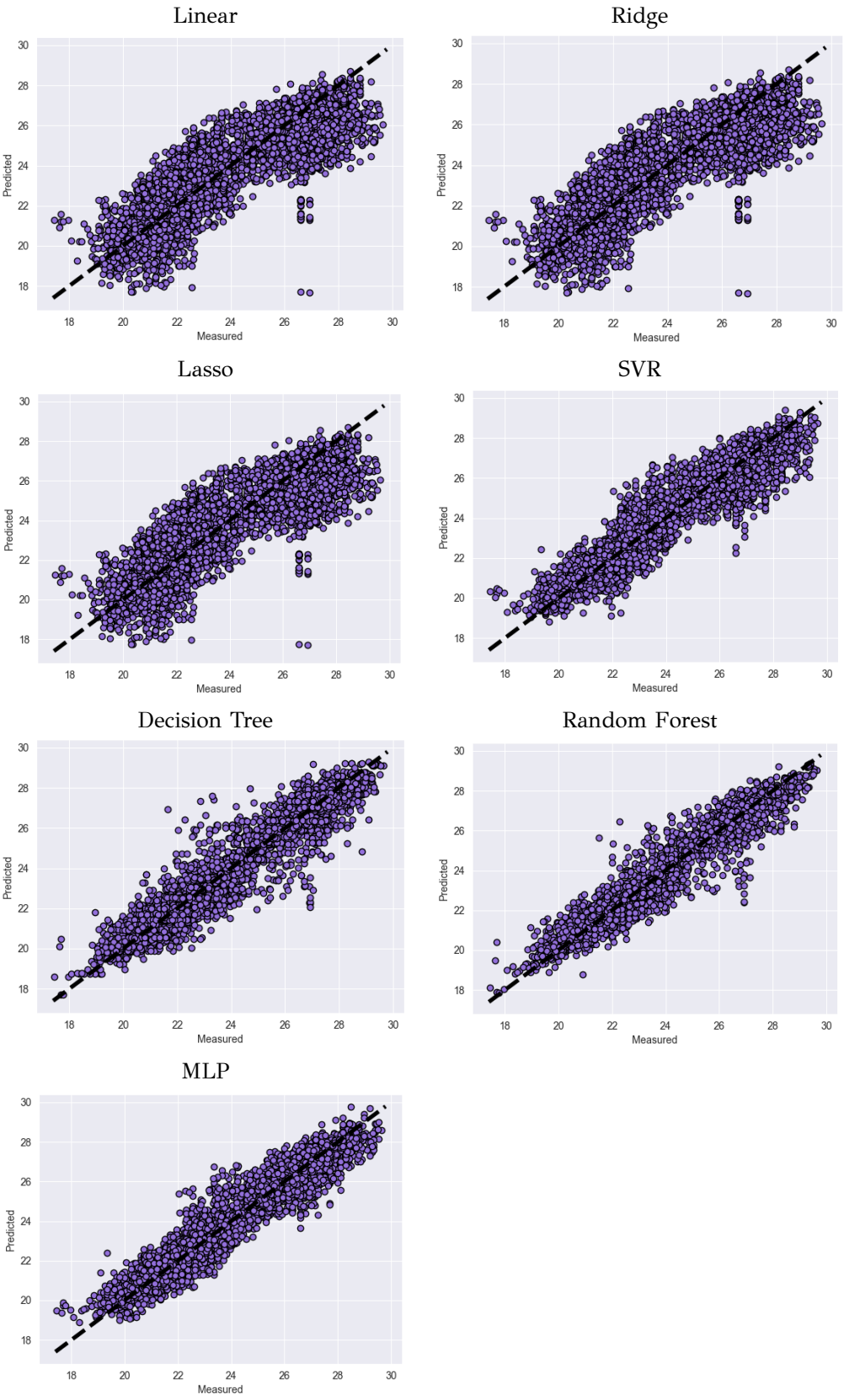
Σχήμα 4.20: Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του μπάνιου.



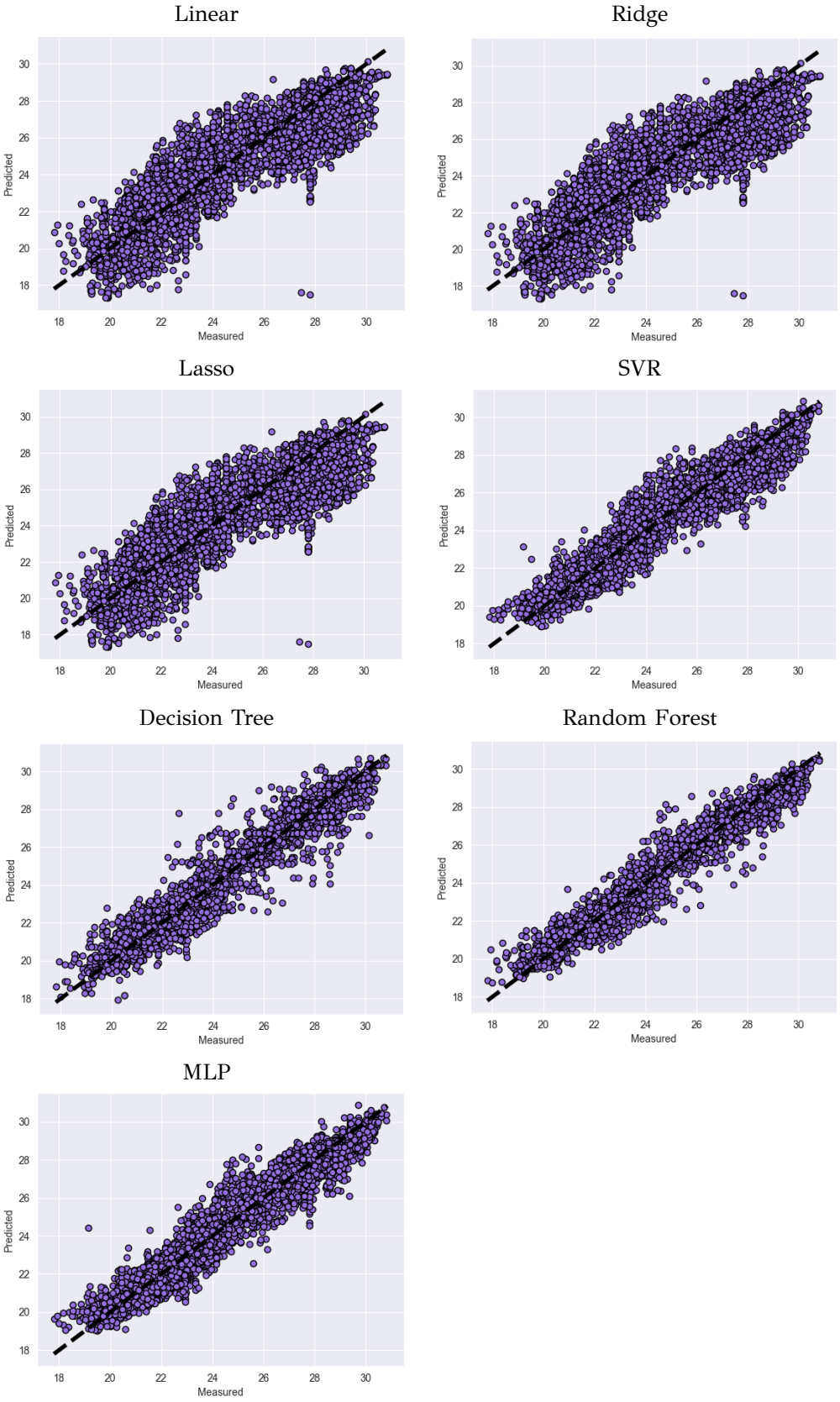
Σχήμα 4.21: Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του γκαράζ.



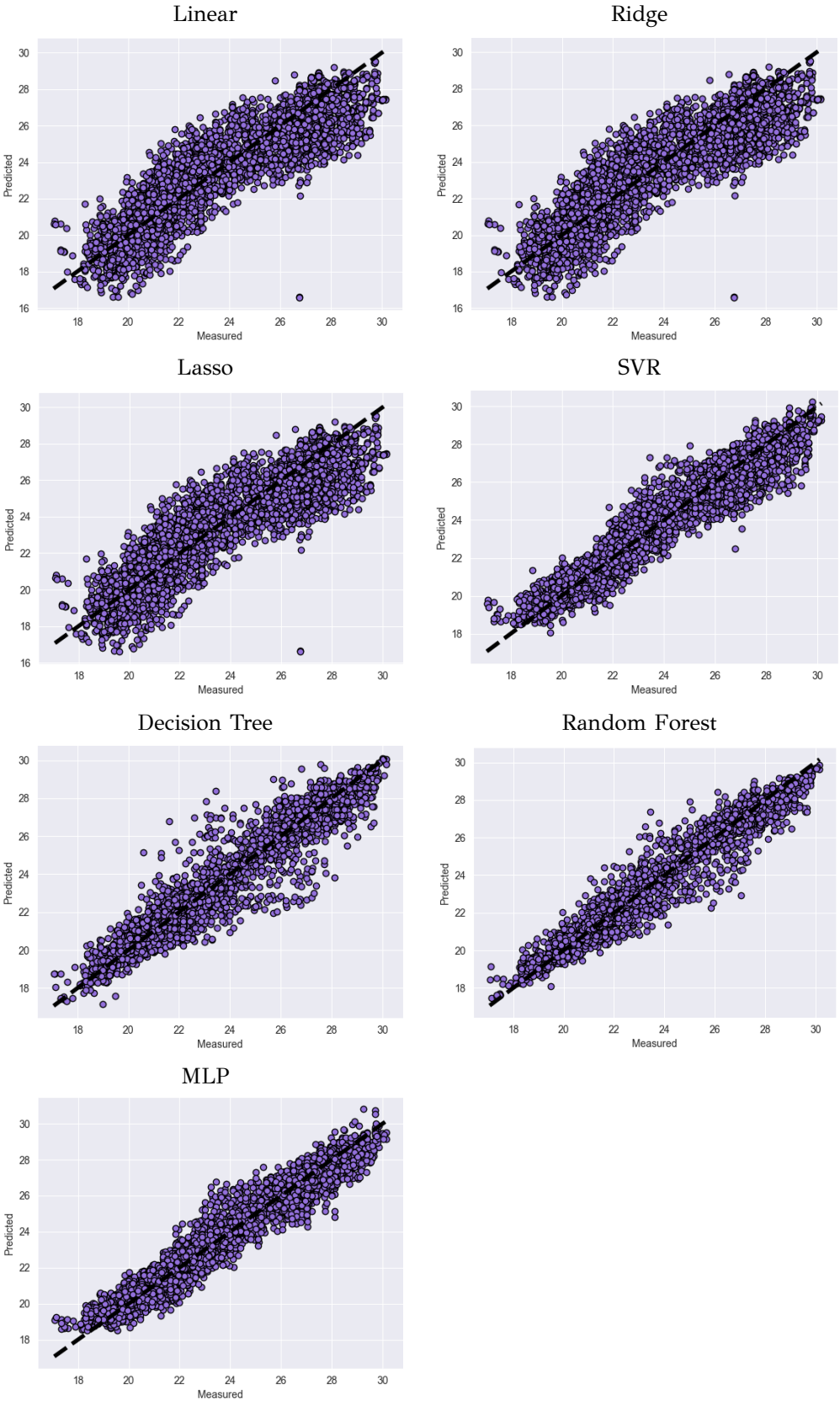
Σχήμα 4.22: Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του γραφείου.



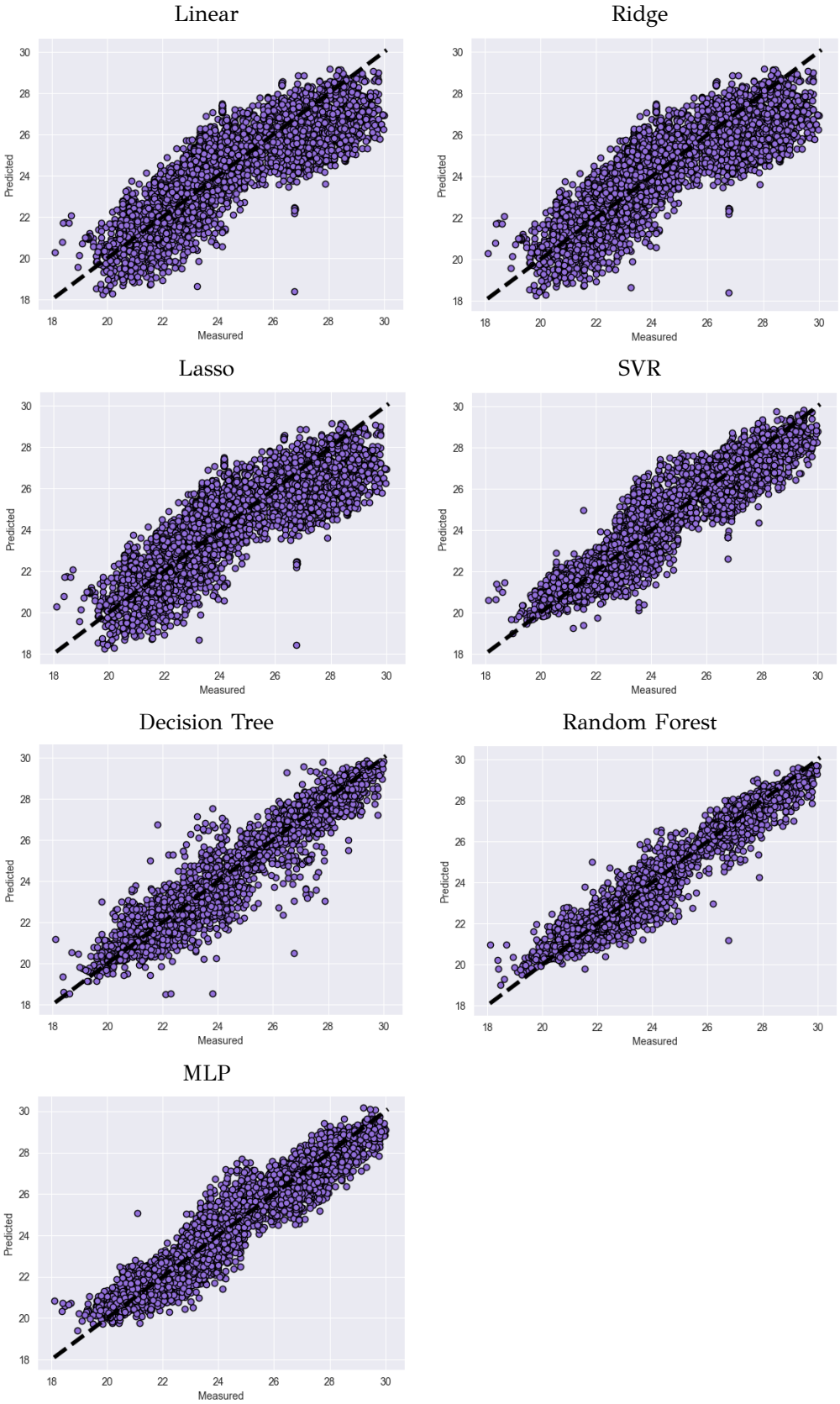
Σχήμα 4.23: Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του υπνοδωματίου.



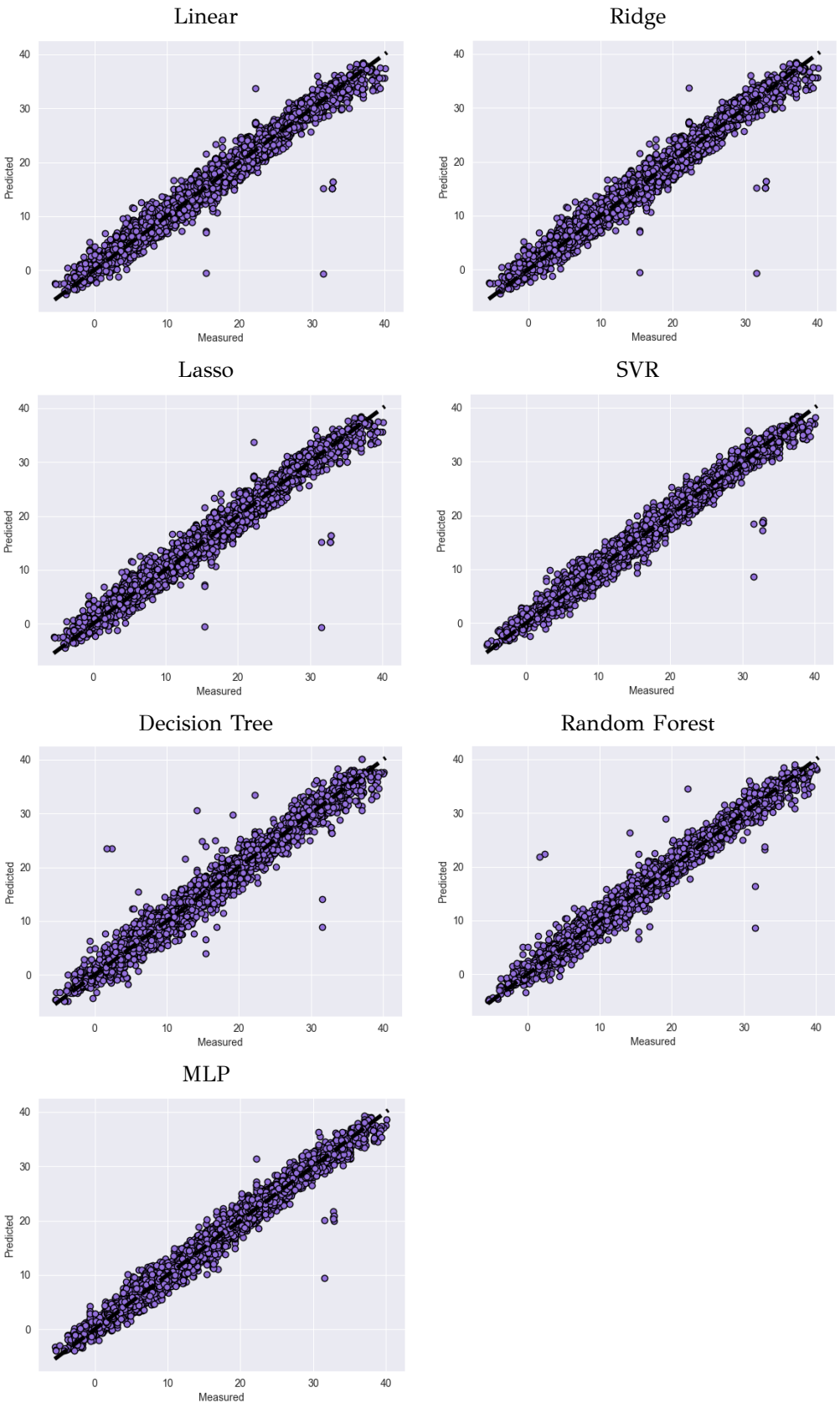
Σχήμα 4.24: Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του δεύτερου υπνοδωματίου.



Σχήμα 4.25: Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο του τρίτου υπνοδωματίου.



Σχήμα 4.26: Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο της κουζίνας.



Σχήμα 4.27: Διαγράμματα ισοτιμίας με είσοδο τα δεδομένα της Ε.Μ.Υ. και ρύθμιση υπερπαραμέτρων για τον χώρο της αυλής.

Κεφάλαιο 5

Συμπεράσματα

Στην παρούσα διπλωματική πραγματοποιήθηκε η μοντελοποίηση της θερμοκρασίας για επτά διαφορετικούς χώρους ενός σπιτιού στην πόλη της Κοζάνης και για ένα χώρο γραφείου στο Πανεπιστήμιο Δυτικής Μακεδονίας με τη χρήση αλγόριθμων μηχανικής μάθησης και ενός στατιστικού μοντέλου πρόβλεψης. Η προσέγγιση του προβλήματος πρόβλεψης της θερμοκρασίας έγινε με τη χρήση αλγόριθμων παλινδρόμησης, καθώς θέλαμε να προβλέψουμε την ακριβή ωριαία μέτρηση της θερμοκρασίας. Εφαρμόστηκε η τεχνική της μηχανικής χαρακτηριστικών για τη δημιουργία νέων χαρακτηριστικών βασιζόμενα στις χρονικές στιγμές των μετρήσεων, τα οποία θα λειτουργούσαν ως δεδομένα εισόδου για την πρόβλεψη της θερμοκρασίας. Αρχικά χρησιμοποιήθηκαν οι γραμμικοί αλγόριθμοι παλινδρόμησης, όπως η απλή γραμμική παλινδρόμηση, η παλινδρόμηση Ridge και η παλινδρόμηση Lasso. Τα αποτελέσματα αυτών των αλγόριθμων ήταν να μη γενικεύσουν καλά το σύνολο των δεδομένων και οι προβλέψεις τους να θεωρηθούν ανακριβείς. Η ακρίβεια των τριών μοντέλων της γραμμικής παλινδρόμησης ήταν περίπου στο 10%, γεγονός που δείχνει την χαμηλή προβλεπτική ισχύ αυτών των αλγοριθμικών μοντέλων στο συγκεκριμένο πρόβλημα. Στη συνέχεια, έγιναν δοκιμές με τους αλγόριθμους των δέντρων απόφασης, των τυχαίων δασών και της παλινδρόμησης με μηχανές διανυσμάτων υποστήριξης, καθώς και με το νευρωνικό δίκτυο MLP. Τα αποτελέσματα αυτών των αλγόριθμων ήταν ικανοποιητικά, πάνω από 82%, καθιστώντας αξιόπιστα αυτά τα μοντέλα μηχανικής μάθησης. Μάλιστα, οι τιμές ακρίβειας του αλγόριθμου τυχαίων δασών πλησίασαν το 99.5%, έχοντας τη μεγαλύτερη προβλεπτική ισχύ. Στην παρούσα διπλωματική πραγματοποιήθηκε η μοντελοποίηση της θερμοκρασίας για επτά διαφορετικούς χώρους ενός σπιτιού στην πόλη της Κοζάνης και για ένα χώρο

γραφείου στο Πανεπιστήμιο με τη χρήση αλγόριθμων μηχανικής μάθησης και ενός στατιστικού μοντέλου πρόβλεψης. Η προσέγγιση του προβλήματος πρόβλεψης της θερμοκρασίας έγινε με τη χρήση αλγόριθμων παλινδρόμησης, καθώς θέλαμε να προβλέψουμε την ακριβή ωριαία μέτρηση της θερμοκρασίας. Εφαρμόστηκε η τεχνική της μηχανικής χαρακτηριστικών για τη δημιουργία νέων χαρακτηριστικών βασιζόμενα στις χρονικές στιγμές των μετρήσεων, τα οποία θα λειτουργούσαν ως δεδομένα εισόδου για την πρόβλεψη της θερμοκρασίας. Αρχικά χρησιμοποιήθηκαν οι γραμμικοί αλγόριθμοι παλινδρόμησης, όπως η απλή γραμμική παλινδρόμηση, η παλινδρόμηση Ridge και η παλινδρόμηση Lasso. Τα αποτελέσματα αυτών των αλγόριθμων ήταν να μη γενικεύσουν καλά το σύνολο των δεδομένων και οι προβλέψεις τους να θεωρηθούν ανακριβείς. Η ακρίβεια των τριών μοντέλων της γραμμικής παλινδρόμησης ήταν περίπου στο 10%, γεγονός που δείχνει την χαμηλή προβλεπτική ισχύ αυτών των αλγοριθμικών μοντέλων στο συγκεκριμένο πρόβλημα. Στη συνέχεια, έγιναν δοκιμές με τους αλγόριθμους των δέντρων απόφασης, των τυχαίων δασών και της παλινδρόμησης με μηχανές διανυσμάτων υποστήριξης, καθώς και με το νευρωνικό δίκτυο MLP. Τα αποτελέσματα αυτών των αλγόριθμων ήταν ικανοποιητικά, πάνω από 82%, καθιστώντας αξιόπιστα αυτά τα μοντέλα μηχανικής μάθησης. Μάλιστα, οι τιμές ακρίβειας του αλγόριθμου τυχαίων δασών πλησίασαν το 99.5%, έχοντας τη μεγαλύτερη προβλεπτική ισχύ.

Στη συνέχεια, επιδιώκοντας τη βελτίωση των παραπάνω μοντέλων, εφαρμόσθηκε η τεχνική της διασταυρούμενης επικύρωσης με 10 πτυχώσεις. Τα αποτελέσματα ακρίβειας παρέμειναν περίπου τα ίδια σε όλους τους αλγόριθμους, έχοντας μία τάση να μειώνονται ελάχιστα. Εκτός από την περίπτωση της γραμμικής παλινδρόμησης Lasso, όπου η διασταυρούμενη επικύρωση είχε ως αποτέλεσμα να μειώσει αισθητά την ακρίβεια, φτάνοντας σε σημείο να είναι αρνητική, περίπου -0.001%. Συνεχίζοντας τις δοκιμές, χρησιμοποιήθηκε το πλαίσιο λογισμικού Optuna, για τη ρύθμιση των υπερπαραμέτρων του κάθε αλγόριθμου. Η ρύθμιση των υπερπαραμέτρων φάνηκε να βελτιώνει αισθητά την ακρίβεια κάθε μοντέλου, ειδικά των μοντέλων παλινδρόμησης με μηχανές διανυσμάτων υποστήριξης και του MLP. Η ακρίβεια αυτών των μοντέλων πλησίασε το ποσοστό του 96%, αποδεικνύοντας πόσο σημαντική είναι η σωστή ρύθμιση των υπερπαραμέτρων για αυτά τα δύο μοντέλα.

Επιπλέον, έγιναν προσπάθειες βελτίωσης των μοντέλων με την εισαγωγή των

ιστορικών δεδομένων της Ε.Μ.Υ. που αφορούν μετρήσεις θερμοκρασίας, πίεσης και υγρασίας για την πόλη της Κοζάνης την ίδια περίοδο, στο σύνολο δεδομένων. Πραγματοποιήθηκαν δοκιμές των μοντέλων με τις τεχνικές της διασταυρούμενης επικύρωσης και της ρύθμισης υπερπαραμέτρων, με τα δεδομένα της Ε.Μ.Υ. ως είσοδο και τις μετρήσεις θερμοκρασίας κάθε χώρου ως δεδομένα εξόδου. Τα αποτελέσματα ακρίβειας παρουσίασαν σημαντική βελτίωση, ειδικά στους αλγόριθμους γραμμικής παλινδρόμησης, Ridge και Lasso. Η ακρίβεια αυτών των μοντέλων ξεπέρασε το 67% και έφτασε το 97% έχοντας παρόμοια προβλεπτική ισχύ με τα πιο σύνθετα μοντέλα. Τα υπόλοιπα μοντέλα διατήρησαν την υψηλή ακρίβεια και αξιοπιστία ως προς την πρόβλεψη της θερμοκρασίας σε κάθε χώρο. Τέλος, πραγματοποιήθηκαν δοκιμές με το στατιστικό μοντέλο ARIMA έτσι ώστε να γίνει σύγκριση του RMSE με αυτό των υπόλοιπων αλγοριθμικών μοντέλων μηχανικής μάθησης. Τα αποτελέσματα έδειξαν πως το μοντέλο ARIMA παρουσιάζει χαμηλές τιμές RMSE, στα επίπεδα της απόδοσης των καλύτερων αλγοριθμικών μοντέλων, οπότε η προβλεπτική του ισχύς κρίθηκε ως αξιόπιστη.

Συνοψίζοντας, στο συγκεκριμένο πείραμα πραγματοποιήθηκε επιτυχώς η ακριβής πρόβλεψη της ωριαίας τιμής της θερμοκρασίας για κάθε ένα από τους οκτώ χώρους που είχαν τοποθετηθεί αισθητήρες μέτρησης θερμοκρασίας. Η ακρίβεια των μοντέλων θα μπορούσε να αυξηθεί περαιτέρω με την εισαγωγή περισσότερων δεδομένων όπως η ταχύτητα ή η κατεύθυνση του αέρα. Ακόμη, η ακραία αλλαγή των καιρικών συνθηκών αποτελεί σημαντικό παράγοντα για την προβλεπτική ισχύ κάθε μοντέλου, συνεπώς θα πρέπει να λαμβάνεται υπόψιν σε μελλοντικά αντίστοιχα πειράματα.

Βιβλιογραφία

[Alawadi et al., 2020] Alawadi, S., Mera, D., Fernández-Delgado, M., Alkhabbas, F., Olsson, C. M., and Davidsson, P. (2020). A comparison of machine learning algorithms for forecasting indoor temperature in smart buildings. *Energy Systems*, pages 1–17.

[Bardenet et al., 2013] Bardenet, R., Brendel, M., Kégl, B., and Sebag, M. (2013). Collaborative hyperparameter tuning. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 199–207, Atlanta, Georgia, USA. PMLR.

[Baştanlar and Özuysal, 2014] Baştanlar, Y. and Özuysal, M. (2014). *Introduction to Machine Learning*, pages 105–128. Humana Press, Totowa, NJ.

[Carrera and Kim, 2020] Carrera, B. and Kim, K. (2020). Comparison analysis of machine learning techniques for photovoltaic prediction using weather sensor data. *Sensors*, 20(11).

[Cutler et al., 2012] Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). *Random Forests*, pages 157–175. Springer US, Boston, MA.

[Dami and Esterabi, 2021] Dami, S. and Esterabi, M. (2021). Predicting stock returns of tehran exchange using lstm neural network and feature engineering technique. *Multimedia Tools and Applications*, pages 1–24.

[Drakos, 2019] Drakos, G. (2019). Decision tree regressor explained in depth. `https://gdcoder.com/decision-tree-regressor-explained-in-depth/`.

[Edwards et al., 2012] Edwards, R., New, J., and Parker, L. (2012). Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings - ENERG BLDG*, 49:591–603.

[Geysen et al., 2017] Geysen, D., Somer, O., Johansson, C., Brage, J., and Vanhoudt, D. (2017). Operational thermal load forecasting in district heating networks using machine learning and expert advice. *Energy and Buildings*, 162.

[Goldstein et al., 2018] Goldstein, A., Fink, L., Meitin, A., Bohadana, S., Lutenberg, O., and Ravid, G. (2018). Applying machine learning on sensor data for irrigation recommendations: Revealing the agronomist's tacit knowledge. *Precision Agriculture*, 47:1–24.

[Han et al., 2012] Han, J., Kamber, M., and Pei, J. (2012). *Data mining concepts and techniques, third edition*. Morgan Kaufmann Publishers, Waltham, Mass.

[Hoerl, 2020] Hoerl, R. (2020). Ridge regression: A historical context. *Technometrics*, 62:420–425.

[Huang et al., 2020] Huang, Z.-Q., Chen, Y.-C., and Wen, C.-Y. (2020). Real-time weather monitoring and prediction using city buses and machine learning. *Sensors*, 20(18).

[influxdata, 2021] influxdata (2021). Influxdb, the platform for building and operating time series applications. `https://www.influxdata.com/`.

[Jain and Mallick, 2017] Jain, G. and Mallick, B. (2017). A study of time series models arima and ets. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2898968/`.

[Jakaria et al., 2020] Jakaria, A., Hossain, M. M., and Rahman, M. (2020). Smart weather forecasting using machine learning: A case study in tennessee. *ArXiv*, abs/2008.10789.

[Kumar et al., 2020] Kumar, R., Kumar, P., and Kumar, Y. (2020). Time series data prediction using iot and machine learning technique. *Procedia Computer Science*, 167:373–381.

[Liu et al., 2017] Liu, B.-C., Binaykia, A., Chang, P.-C., Tiwari, M., and Tsao, C.-C. (2017). Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang. *PLoS ONE*, 12.

[Monteiro et al., 2018] Monteiro, P., Zanin, M., Menasalvas, E., Pimentao, J., and Sousa, P. (2018). Indoor temperature prediction in an iot scenario. *Sensors*, 18:3610.

[Moraru et al., 2010] Moraru, A., Pesko, M., Porcius, M., Fortuna, C., and Mladenić, D. (2010). Using machine learning on sensor data. *CIT*, 18.

[Muller, 2017] Muller, A. C. (2017). *Hands-On Machine Learning with Scikit-Learn and Tensorflow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, Inc.

[Muller and Guido, 2016] Muller, A. C. and Guido, S. (2016). *Introduction to Machine Learning with Python: a guide for Data Scientists*. O'Reilly Media, Inc.

[openHAB, 2021] openHAB (2021). openhab, empowering the smart home. `https://www.openhab.org/`.

[Pan et al., 2008] Pan, Y., Jiang, J., Wang, R., and Cao, H. (2008). Advantages of support vector machine in qspr studies for predicting auto-ignition temperatures of organic compounds. *Chemometrics and Intelligent Laboratory Systems*, 92(2):169–178.

[Rokach and Maimon, 2005] Rokach, L. and Maimon, O. (2005). *Decision Trees*, pages 165–192. Springer US, Boston, MA.

[Scikit-learn, 2021] Scikit-learn (2021). Neural network models (supervised). `https://scikit-learn.org/stable/modules/neural_networks_supervised.html#regression`.

[Sonoff, 2020] Sonoff (2020). Sonoff official homepage | smart home automation. `https://sonoff.tech/`.

[Tasmota, 2016] Tasmota (2016). Open source firmware for esp8266 devices. https://tasmota.github.io/docs/.

[Thrun, 1992] Thrun, S. B. (1992). Efficient exploration in reinforcement learning. Technical report.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[Wang and Liu, 2011] Wang, F. and Liu, J. (2011). Networked wireless sensor data collection: Issues, challenges, and approaches. *IEEE Communications Surveys Tutorials*, 13(4):673–687.

[Wood, 2021] Wood, T. (2021). Random forests's definition. https://deepai.org/machine-learning-glossary-and-terms/random-forest.

[Zumwald et al., 2021] Zumwald, M., Knüsel, B., Bresch, D. N., and Knutti, R. (2021). Mapping urban temperature using crowd-sensing data and machine learning. *Urban Climate*, 35:100739.