



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ
ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές Μηχανικής Μάθησης για την πρόβλεψη και βελτιστοποίηση κατανάλωσης καυσίμου στην ναυτιλία

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Παναγιώτας-Ελευθερίας Α. Κουφού

Επιβλέπων : Στέφανος Κόλλιας,
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ
ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές Μηχανικής Μάθησης για την πρόβλεψη και βελτιστοποίηση κατανάλωσης καυσίμου στην ναυτιλία

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Παναγιώτας- Ελευθερίας Α. Κουφού

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12^η Οκτωβρίου 2023.

.....
Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

.....
Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

.....
Αθανάσιος Βουλόδημος
Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

.....

Παναγιώτα – Ελευθερία Α. Κουφού

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

Copyright © Παναγιώτα – Ελευθερία Α. Κουφού, 2023

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

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 δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Κρίσιμο ζήτημα, για τον τομέα της ναυτιλίας αποτελεί ο έλεγχος της κατανάλωσης καυσίμου, για λόγους οικονομικούς και περιβαλλοντικούς. Στην παρούσα εργασία, το πρόβλημα αυτό αντιμετωπίζεται από την σκοπιά της συνεισφοράς που μπορεί να έχει η μηχανική μάθηση στην πρόβλεψη της κατανάλωσης καυσίμου, βάση άλλων χαρακτηριστικών του πλοίου και των καιρικών συνθηκών, την παρούσα χρονική στιγμή ή σε βάθος μικρού χρονικού ορίζοντα και στον εντοπισμό μοτίβων που συσχετίζουν την κατανάλωση καυσίμου με τα λοιπά χαρακτηριστικά. Στο θεωρητικό μέρος, παρουσιάστηκαν οι βασικές αρχές και βήματα λειτουργίας ενός πλήθους αλγορίθμων επιβλεπόμενης, συλλογικής, ομοσπονδιακής μάθησης και αναδρομικών νευρωνικών δικτύων. Ακόμα, συζητήθηκαν οι κυριότερες προσεγγίσεις στην προεπεξεργασία των δεδομένων και οι δημοφιλέστερες μετρικές αξιολόγησης μοντέλων επιβλεπόμενης μάθησης. Στο πρακτικό μέρος, παρουσιάστηκαν τα αποτελέσματα από την εφαρμογή αυτών των αλγορίθμων και μεθόδων στα πραγματικά δεδομένα πλοίων, για την επιλογή διαφορετικών υπερπαραμέτρων, σχολιάστηκαν και αξιολογήθηκαν ως προς την ικανότητα τους να παράγουν γενικευμένες προβλέψεις. Τέλος, έγινε αναφορά στην συνεισφορά της εργασίας μας και σε μελλοντικές προεκτάσεις που θα μπορούσε να έχει.

Λέξεις-Κλειδιά : μηχανική μάθηση, ναυτιλία, κατανάλωση καυσίμου, επιβλεπόμενη μάθηση, νευρωνικά δίκτυα, συλλογική μάθηση, συστήματα μακράς-βραχείας μνήμης, ομοσπονδιακή μάθηση, προεπεξεργασία.

Abstract

A critical issue for the shipping sector is the control of fuel consumption, for economic and environmental reasons. In this thesis, this problem is addressed from the point of view of the contribution that machine learning can have in predicting fuel consumption, based on other attributes of the ship and weather conditions, at the present time or in a short time and in detecting patterns, which relate fuel consumption to the other attributes. In the theoretical part, the basic principles and operating steps of a number of supervised, ensemble, federated learning and recurrent neural network algorithms were presented. Also, the main approaches to data preprocessing and the most popular metric evaluations of supervised learning models were discussed. In the practical part, the results from the application of these algorithms and methods to real ship data, for the selection of different hyperparameters, were presented, commented and evaluated for their ability to produce generalized predictions. Finally, reference was made to the contribution of our work and future implications it could have.

Keywords: machine learning, shipping, fuel consumption, supervised learning, neural networks, ensemble learning, long short-term memory (LSTM) systems, federated learning, preprocessing.

Ευχαριστίες

Θα ήθελα αρχικά να ευχαριστήσω τον επιβλέποντα Καθηγητή Ε.Μ.Π. κ. Στέφανο Κόλλια, για την καθοδήγηση του. Ακόμα, θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή Ε.Κ.ΠΑ. κ. Παναγιώτη Τρακάδα, τον Επίκουρο Καθηγητή Ε.Κ.ΠΑ. κ. Γιώργο Αλεξανδρίδη και τον Υποψήφιο Διδάκτορα κ. Τάσο Γιαννόπουλο για την πολύτιμη στήριξη και βοήθεια τους. Φυσικά, ευχαριστώ τα λοιπά μέλη της εξεταστικής επιτροπής, τον κ. Γεώργιο Στάμου, Καθηγητή Ε.Μ.Π. και τον κ. Αθανάσιο Βουλόδημο, Επίκουρο Καθηγητή Ε.Μ.Π., για τον χρόνο και την αξιολόγησή τους. Επίσης, θα ήθελα να ευχαριστήσω τον Δημήτρη, τον Στέλιο και την Ηλέκτρα για τις ιδέες που μου έδωσαν όταν τις χρειαζόμουν. Ξεχωριστή θέση στις ευχαριστίες μου καταλαμβάνουν οι φίλες και φίλοι μου, Ελένη, Χαρά, Άννα, Πέγκυ, Νεκταρία, Μάρω, Λευτέρης, Ρωμυλία, Ανδρέας, Δημήτρης καθώς και όλα τα υπόλοιπα πολύ ξεχωριστά άτομα που γνώρισα στο Πολυτεχνείο, που με αγκάλιασαν και με στήριξαν. Τέλος, δεν θα μπορούσαν να λείπουν από τον αυτόν τον κατάλογο ευχαριστιών οι γονείς και η αδερφή μου, καθώς και όλη η υπόλοιπη οικογένειά μου, χωρίς των οποίων την βοήθεια δεν θα μπορούσα να ξεκινήσω το παραμικρό.

Παναγιώτα Κουφού, Σεπτέμβρης 2023

Περιεχόμενα

1. Εισαγωγή.....	12
2. Μηχανική Μάθηση	14
2.1 Ορισμός και Κατηγορίες Μηχανικής Μάθησης	14
2.2 Επιβλεπόμενη Μάθηση (Supervised Learning)	15
2.2.1 Γραμμική Παλινδρόμηση (Linear Regression)	16
2.2.2 Δένδρα Απόφασης (Decision Trees) σε προβλήματα Παλινδρόμησης	17
2.2.3 Παλινδρόμηση Διανουσμάτων Υποστήριξης (Support Vector Regression – SVR)	19
2.2.4 Τυχαίο Δάσος (Random Forest) σε προβλήματα Παλινδρόμησης.....	20
2.2.5 Νευρώνας Πολλαπλών Στρωμάτων (Multiple Layer Perceptron – MLP) ...	21
2.3 Συλλογική Μάθηση (Ensemble Learning)	24
2.3.1 Παλινδρόμηση Ψηφίσματος (Voting Regression)	24
2.3.2 Σακούλιασμα (Bagging) σε προβλήματα Παλινδρόμησης	25
2.3.3 Ενδυνάμωση (Boosting) σε προβλήματα Παλινδρόμησης	27
2.3.3.1 Τεχνική AdaBoost σε προβλήματα Παλινδρόμησης	27
2.3.3.2 Διαβαθμισμένη Ενδυνάμωση (Gradient Boosting) σε προβλήματα Παλινδρόμησης.....	28
2.3.4 Στοίβαξη (Stacking) σε προβλήματα Παλινδρόμησης.....	30
2.4 Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks, RNN).....	30
2.4.1 Συστήματα μακράς βραχυπρόθεσμης μνήμης (LSTM).....	31
2.5 Τεχνικές Προεπεξεργασίας (Preprocessing)	33
2.5.1 Διαχείριση απουσιαζουσών τιμών (missing values)	34
2.5.2 Διαχείριση των έκτοπων τιμών (outliers)	35
2.5.3 Μετασχηματισμός των δεδομένων	35
2.5.3.1 Κανονικοποίηση Ελαχίστου-Μεγίστου (Min-Max Scaling).....	36
2.6 Μετρικές της Επιβλεπόμενης Μάθησης.....	36
2.6.1 Μετρικές για προβλήματα παλινδρόμησης	37
2.7 Ομοσπονδιακή Μάθηση (Federated Learning)	38
2.7.1 Αλγόριθμος Συμψηφισμού Μέσου Όρου (FedAvg).....	43
2.8 Άλλες Μέθοδοι.....	43
2.8.1 Διασταυρωμένη Επικύρωση	44
2.8.2 Αναζήτηση Πλέγματος	44
2.8.3 Έλεγχος t του Student (Student’s t-test)	45

3.	Περιγραφή και Προετοιμασία Συνόλου Δεδομένων	46
3.1	Περιγραφή του Συνόλου Δεδομένων.....	46
3.2	Προεπεξεργασία Δεδομένων	49
4.	Πειραματικά Αποτελέσματα Μηχανικής Μάθησης	53
4.1	Μοντέλα Επιβλεπόμενης Μάθησης	53
4.1.1	Αποτελέσματα Γραμμικής Παλινδρόμησης.....	54
4.1.2	Αποτελέσματα Δέντρων Απόφασης	55
4.1.3	Αποτελέσματα Παλινδρόμησης Διανυσμάτων Υποστήριξης.....	56
4.1.4	Αποτελέσματα Τυχαίου Δάσους.....	56
4.1.5	Αποτελέσματα Νευρωνικού Δικτύου Πολλαπλών Στρωμάτων.....	58
4.1.6	Σύγκριση Αποτελεσμάτων Επιβλεπόμενης Μάθησης.....	59
4.2	Μοντέλα Συλλογικής Μάθησης	60
4.2.1	Αποτελέσματα Παλινδρόμησης Ψηφίσματος	61
4.2.2	Αποτελέσματα Σακουλιάσματος.....	61
4.2.3	Αποτελέσματα AdaBoost Τεχνικής	63
4.2.4	Αποτελέσματα Διαβαθμισμένης Ενδυνάμωσης.....	64
4.2.5	Αποτελέσματα Στοίβαξης	65
4.2.6	Σύγκριση Αποτελεσμάτων	65
4.3	Αποτελέσματα Αναδρομικών Νευρωνικών Δικτύων	67
4.4	Αποτελέσματα Ομοσπονδιακής Μάθησης.....	71
5.	Συμπεράσματα.....	74
5.1	Συνεισφορά Εργασίας	74
5.2	Μελλοντικές Προεκτάσεις	74
	Αναφορές.....	76
	Συνομογραφίες.....	80

Πίνακας Σχημάτων

Εικόνα 1: Παράδειγμα λειτουργίας της γραμμικής παλινδρόμησης. Με κόκκινο χρώμα, παρουσιάζονται οι πραγματικές τιμές της εξόδου του δείγματος και με μπλε χρώμα η γραμμή που προκύπτει από το μοντέλο γραμμικής παλινδρόμησης. Με πράσινο χρώμα, παρουσιάζονται οι γραμμές που χρησιμοποιούνται για τον υπολογισμό του σφάλματος. Η θέση και η κλίση της ευθείας μπορεί να τροποποιηθεί ώστε να επιτευχθεί ελαχιστοποίηση του σφάλματος. (Πηγή : [6]).	17
Εικόνα 2: Παράδειγμα της δημιουργίας ενός δένδρου απόφασης. (Πηγή: [7]).	18
Εικόνα 3: Παράδειγμα της διαφορετικής αντιμετώπισης της κατηγοριοποίησης δειγμάτων από ένα γραμμικό διαχωριστή και έναν διαχωριστή διανυσμάτων υποστήριξης. Βλέπουμε ότι η κοντινότερη στα τετράγωνα πράσινη γραμμή μπορεί να ελαχιστοποιεί το σφάλμα αλλά είναι επιρρεπής σε λανθασμένες ταξινομήσεις. Η πράσινη γραμμή του δεξιού σχήματος είναι το βέλτιστο υπερεπίπεδο. (Πηγή: [8]).	19
Εικόνα 4: Συνάρτηση πυρήνα RBF. (Πηγή: [9]).	20
Εικόνα 5 : Μοντέλο ενός τυχαίου δάσους. Οι προβλέψεις των N δέντρων απόφασης συμψηφίζονται ώστε να προκύψει το τελικό αποτέλεσμα. (Πηγή: [11]).	21
Εικόνα 6: Μοντέλο ενός νευρώνα Perceptron. (Πηγή: [13]).	22
Εικόνα 7: Ένας πολυστρωματικός νευρώνας (MLP). (Πηγή: [14]).	23
Εικόνα 8: Η δομή ενός μοντέλου Παλινδρόμησης Ψηφίσματος.	25
Εικόνα 9: Παράδειγμα του τρόπου δημιουργίας υποσυνόλων στην τεχνική του σακουλιάσματος.	26
Εικόνα 10: Τρόπος λειτουργίας της τεχνικής Σακουλιάσματος.	26
Εικόνα 11: Παράδειγμα της λειτουργίας της τεχνικής AdaBoost σε πρόβλημα κατηγοριοποίησης. (Πηγή: [18]).	28
Εικόνα 12: Τρόπος λειτουργίας ενός διαβαθμισμένου ενισχυτή. (Πηγή: [20])	29
Εικόνα 13: Δομή ενός μοντέλου στοίβαξης.	30
Εικόνα 14: Δομή ενός αναδρομικού δικτύου και το ανάπτυγμα που προκύπτει από το ξετύλιγμά του. (Πηγή: [22])	31
Εικόνα 15: Δομή ενός LSTM κελιού. (Πηγή: [23]).	32
Εικόνα 19: Γενική δομή ενός μοντέλου Ομοσπονδιακής Μάθησης.	40
Εικόνα 20: Κατανομές των χαρακτηριστικών του συνόλου δεδομένων μετά την προεπεξεργασία.	51
Εικόνα 21: Απόκλιση μεταξύ των προβλεπόμενων (predictions) και των πραγματικών (actual) τιμών. Η μπλε ευθεία είναι η ευθεία που παράγεται από την γραμμική παλινδρόμηση και θα ταυτιζόταν με τις κόκκινες κουκίδες σε περίπτωση που η πρόβλεψη ήταν πλήρως ακριβής.	54
Εικόνα 22: ΜΤΣ (MSE) ανάλογα με το πλήθος των δέντρων (Number of Trees) σε ένα μοντέλο τυχαίου δάσους (Random Forest). Το καλύτερο ΜΤΣ επιτυγχάνεται για 200 δέντρα.	57
Εικόνα 23: Αξιολόγηση των μοντέλων που χρησιμοποιήσαμε με βάση το ΜΤΣ (MSE). Τα μοντέλα που συγκρίνουμε είναι η γραμμική παλινδρόμηση (LR), το δέντρο απόφασης (Decision Tree), η παλινδρόμηση διανυσμάτων υποστήριξης (SVR Model),	

το τυχαίο δάσος (Random Forest) και το νευρωνικό δίκτυο πολλαπλών στρωμάτων (MLP).....	60
Εικόνα 24: Το ΜΤΣ (MSE) του μοντέλου σακουλιάσματος (Bagging) ως προς το ποσοστό επί της πληθικότητας του συνόλου εκπαίδευσης, των δειγμάτων που επιλέγονται για την εκπαίδευση του μοντέλου βάσης (Bootstrapping).....	62
Εικόνα 25: Το ΜΤΣ (MSE) του μοντέλου σακουλιάσματος (Bagging) ως προς τον αριθμό των μοντέλων βάσης (number of estimators).	62
Εικόνα 26: Το ΜΤΣ (MSE) του μοντέλου AdaBoost ως προς τον ρυθμό μάθησης (learning rate).	63
Εικόνα 27: : Το ΜΤΣ (MSE) του μοντέλου AdaBoost ως προς τον αριθμό των μοντέλων βάσης (number of estimators). Παρατηρούμε ότι δεν διαχειριζόμαστε καλύτερα τα σφάλματα για μεγαλύτερο πλήθος μοντέλων βάσης.....	63
Εικόνα 28: Το ΜΤΣ (MSE) του μοντέλου διαβαθμισμένης ενδυνάμωσης (GDBoost) ως προς τον ρυθμό μάθησης (learning rate).	64
Εικόνα 29: Το ΜΤΣ (MSE) του μοντέλου διαβαθμισμένης ενδυνάμωσης (GDBoost) ως προς τον αριθμό των μοντέλων βάσης (number of estimators).	65
Εικόνα 30: Το ΜΤΣ (MSE) ανά μοντέλο, όπου συγκρίνονται τα μοντέλα γραμμικής παλινδρόμησης (LR), δέντρου απόφασης (Decision Tree), παλινδρόμησης διανυσμάτων υποστήριξης (SVR), τυχαίου δάσους (Random Forest), νευρωνικού δικτύου πολλαπλών στρωμάτων (MLP), παλινδρόμησης ψηφίσματος (Voting Regressor), σακουλιάσματος (Bagging), AdaBoost , διαβαθμισμένης ενίσχυσης (GDBoost) και στοίβαξης (Stacking).....	66
Εικόνα 31: Μείωση της απώλειας (εδώ ΜΤΣ) κατά την διάρκεια της εκπαίδευσης (training loss). Ο οριζόντιος άξονας αντιστοιχεί στον αριθμό των εποχών.	69
Εικόνα 32: Παρουσιάζονται οι πραγματικές και οι προβλεπόμενες τιμές του ΜΕΡ σε χρόνο (time step) ανά 15 λεπτά. Η κόκκινη γραμμή αντιστοιχεί στις προβλεπόμενες τιμές και η μπλε στις πραγματικές.	70
Εικόνα 33: Απώλεια εκπαίδευσης (training loss) ως προς το πλήθος των εποχών. ...	71
Εικόνα 35: Το ΜΤΣ (MSE) που προκύπτει αν η εκπαίδευση του μοντέλου για τα πλοία 1 και 2 γίνει ξεχωριστά (MLP for 1,2) ή με ομοσπονδιακή μάθηση (FL for 1,2).	73
Εικόνα 36: Το ΜΤΣ (MSE) βάση αν το μοντέλο εκπαίδευσης που χρησιμοποιούμε ακολουθεί κεντρική (Centralized) ή ομοσπονδιακή (Federated) προσέγγιση (approach).....	73

Πίνακας Πινάκων

Πίνακας 1: Χαρακτηριστικά του αρχικού συνόλου δεδομένων.	48
Πίνακας 2: Χαρακτηριστικά του συνόλου δεδομένων, μετά την προεπεξεργασία. ...	50

1. Εισαγωγή

Ο τομέας της ναυτιλίας συγκεντρώνει τεράστιο οικονομικό, και όχι μόνο ενδιαφέρον, καθώς αποτελεί την εμπορική αρτηρία του πλανήτη. Συγκεκριμένα, 90% του παγκόσμιου εμπορίου διεξάγεται δια θαλάσσης [1]. Ταυτόχρονα, η διακίνηση εμπορευμάτων με την χρήση πλοίων ευθύνεται για το 3% των παγκόσμιων εκπομπών σε αέρια του θερμοκηπίου [2]. Επομένως, η μελέτη της κατανάλωσης καυσίμου στην κίνηση των πλοίων αποτελεί φλέγον ζήτημα καθώς σχετίζεται με τεράστια οικονομικά συμφέροντα, αλλά και με τις προσπάθειες περιορισμού του οικολογικού αποτυπώματος της ανθρωπότητας.

Πλήθος λύσεων έχουν προταθεί για τον έλεγχο της κατανάλωσης καυσίμου, ο οποίος μέχρι τώρα βασιζόταν αποκλειστικά στην εμπειρία των χειριστών του εκάστοτε πλοίου. Μία εξ αυτών των προσεγγίσεων στηρίζεται στην χρήση αλγορίθμων μηχανικής μάθησης, οι οποίοι αφενός μπορούν να προβλέψουν την κατανάλωση καυσίμου την τρέχουσα στιγμή αλλά και σε βάθος χρόνου και αφετέρου να εντοπίσουν δαπανηρές συμπεριφορές και συνθήκες. Τα αποτελέσματα αυτά μπορούν να χρησιμοποιηθούν άμεσα, για την πραγματοποίηση διορθώσεων που θα μειώσουν την κατανάλωση καυσίμου σε σύντομο χρονικό ορίζοντα, αλλά και στον σχεδιασμό μελλοντικών, πιο οικονομικών και φιλικών προς το περιβάλλον, ταξιδιών.

Παράλληλα, η επιθυμία για αποτελεσματικότερη αντιμετώπιση του προβλήματος της κατανάλωσης καυσίμου στην ναυτιλία, δημιουργεί την ανάγκη για διαμοιρασμό πληροφοριών μεταξύ των ναυτιλιακών εταιρειών. Ωστόσο, τα δεδομένα που απαιτείται να ανταλλαχθούν, υπόκεινται σε αυστηρά πρωτόκολλα ασφαλείας, με σκοπό την προστασία των εταιρικών συμφερόντων. Επομένως, αυτή την στιγμή, κρίσιμο πεδίο έρευνας αποτελούν οι προσπάθειες για την ανάπτυξη μεθόδων, οι οποίες θα επιτρέπουν την αποδοτική ανταλλαγή πληροφοριών μεταξύ των εταιρειών, τηρώντας παράλληλα όλες τις απαραίτητες προδιαγραφές προκειμένου να διασφαλιστεί η εμπιστευτικότητα των εταιρικών δεδομένων. Μία από αυτές της μεθόδους είναι και η ομοσπονδιακή μάθηση.

Ο στόχος της παρούσας εργασίας είναι η πρόβλεψη της ισχύος του κύριου κινητήρα, μέγεθος το οποίο είναι άμεσα συσχετισμένο με την κατανάλωση καυσίμου ενός πλοίου, με βάση άλλα χαρακτηριστικά του πλοίου και τις επικρατούσες καιρικές συνθήκες. Προκειμένου να πραγματοποιηθεί η πρόβλεψη αυτή θα μελετηθεί, θα περιγραφεί και θα εφαρμοστεί μια ευρεία γκάμα μεθόδων και μεθοδολογιών μηχανικής μάθησης. Ειδικότερα, θα παρουσιαστούν και θα δοκιμαστούν στην πράξη τεχνικές επιβλεπόμενης και συλλογικής μάθησης, αναδρομικών νευρωνικών δικτύων και ομοσπονδιακής μάθησης. Στα πλαίσια της διερεύνησης των μεθόδων μηχανικής μάθησης θα συζητηθεί και η σημασία των τεχνικών προεπεξεργασίας καθώς και μετρικές που χρησιμοποιούνται για την αξιολόγηση μοντέλων μηχανικής

μάθησης. Στο πρακτικό μέρος θα παρουσιάσουμε τα αποτελέσματα που προέκυψαν από την εφαρμογή αυτού του εκτενούς εύρους μεθόδων στα πραγματικά δεδομένα πλοίων. Τέλος, θα συζητήσουμε την συνεισφορά της εργασίας μας και πιθανές μελλοντικές προεκτάσεις.

2. Μηχανική Μάθηση

2.1 Ορισμός και Κατηγορίες Μηχανικής Μάθησης

Η Μηχανική Μάθηση (Machine Learning - ML) είναι ένας κλάδος της Τεχνητής Νοημοσύνης (Artificial Intelligence - AI), ο οποίος ασχολείται με την εκπαίδευση μοντέλων προκειμένου αυτές να δώσουν λύσεις σε περίπλοκα προβλήματα. Ο δημοφιλέστερος ορισμός της Μηχανικής Μάθησης έχει διατυπωθεί από τον Tom M. Mitchell, σύμφωνα με τον οποίον: «ένα υπολογιστικό πρόγραμμα, λέμε ότι μαθαίνει από μια εμπειρία E , ως προς μία κλάση εργασιών T και σύμφωνα με ένα μέτρο επίδοσης P , αν η επίδοση του στις εργασίες του T , όπως μετριέται από το P , αυξάνεται με την εμπειρία E » [3].

Ο όρος Μηχανική Μάθηση εισήχθη από τον Arthur Samuel, το 1952 [4], και έκτοτε ο κλάδος έχει γνωρίσει τεράστια άνθηση, με τις τεχνικές του να αποτελούν νευραλγικό κομμάτι πλήθους τεχνολογιών και με προοπτικές για μεγάλη εξέλιξη στο μέλλον. Γενικότερα, οι μέθοδοι της μηχανικής μάθησης μπορούν να χωριστούν σε τρεις κατηγορίες, ανάλογα με τον τύπο των δεδομένων που τροφοδοτούν το μοντέλο [5] :

- **Επιβλεπόμενη Μάθηση (Supervised Learning):** Στην επιβλεπόμενη μάθηση, το μοντέλο τροφοδοτείται με ζεύγη εισόδου - εξόδου και μαθαίνει, ανάλογα με την είσοδο, ποια είναι η επιθυμητή έξοδος, ώστε στη συνέχεια να εφαρμοστεί σε δεδομένα που δεν έχει ξανασυναντήσει και των οποίων την επιθυμητή έξοδο δεν γνωρίζει. Δηλαδή, μετά το πέρας της εκπαίδευσης, το μοντέλο θα πρέπει να είναι σε θέση να *γενικεύσει* τα συμπεράσματα που έβγαλε για την σχέση εισόδου – εξόδου, από τα ζεύγη εκπαίδευσης, και η γενίκευση του αυτή να οδηγήσει σε υψηλής ποιότητας προβλέψεις για άγνωστα, ως τώρα, δεδομένα [5].
- **Μη-Επιβλεπόμενη Μάθηση (Unsupervised Learning):** Στην μη επιβλεπόμενη μάθηση, τα δεδομένα που χρησιμοποιούνται, δεν συνοδεύονται από ετικέτες, που να μαρτυρούν την επιτυχία ή μη των προβλέψεων του μοντέλου. Η μη-επιβλεπόμενη μάθηση, επομένως, χρησιμοποιείται για την εξαγωγή μοτίβων και κρυμμένων χαρακτηριστικών από το σύνολο δεδομένων [5].
- **Ενισχυτική Μάθηση (Reinforcement Learning):** Στην ενισχυτική μάθηση, στόχος είναι η εκπαίδευση ενός δράστη (agent), ο οποίος παίρνει κάποιες αποφάσεις και ανάλογα με την επιτυχία τους ως προς έναν τελικό στόχο, τιμωρείται ή επιβραβεύεται [5]. Όταν ολοκληρωθεί η εκπαίδευση, ο δράστης θα πρέπει να είναι σε θέση, ανάλογα με την κατάσταση του κόσμου, να

πάρει αποφάσεις οι οποίες θα έχουν ως αποτέλεσμα την επίτευξη του στόχου, με τον βέλτιστο δυνατό τρόπο.

2.2 Επιβλεπόμενη Μάθηση (Supervised Learning)

Τα προβλήματα, τα οποία επιλύονται με αλγόριθμους επιβλεπόμενης μάθησης, μπορούν, με την σειρά τους, να χωριστούν σε δύο κατηγορίες, ανάλογα με το είδος των εξόδων που πρέπει να προβλέπουν:

- **Προβλήματα Κατηγοριοποίησης (Classification):** Στα προβλήματα κατηγοριοποίησης, η έξοδος αποτελείται από διακριτές κατηγορίες-κλάσεις και οι είσοδοι θα πρέπει, με την χρήση του μοντέλου, να τοποθετηθούν στην σωστή κατηγορία-κλάση [5]. Για παράδειγμα, ένα πρόβλημα κατηγοριοποίησης θα μπορούσε να είναι να αποφασιστεί, κατά πόσο το ζώο που παρουσιάζεται σε μια φωτογραφία είναι γάτα ή σκύλος. Στην προκειμένη περίπτωση, η είσοδος είναι φωτογραφίες ζώων και η διακριτή έξοδος είναι η κατηγορία γάτα ή σκύλος.
- **Προβλήματα Παλινδρόμησης (Regression):** Στα προβλήματα παλινδρόμησης, η έξοδος που πρέπει να προβλεφθεί είναι μια πραγματική τιμή [5]. Για παράδειγμα, ένα πρόβλημα παλινδρόμησης θα μπορούσε να είναι, με βάση κάποια στοιχεία για τον καιρό, να προβλεφθεί η θερμοκρασία για τις επόμενες ώρες. Προφανώς, η θερμοκρασία είναι μια συνεχής και όχι διακριτή τιμή, και για αυτό, αυτό το πρόβλημα θεωρείται πρόβλημα παλινδρόμησης.

Κάποιες ακόμα κεντρικές έννοιες της επιβλεπόμενης μάθησης που θα πρέπει να αποσαφηνιστούν, αποτελούν το *σύνολο εκπαίδευσης* (training set) και το *σύνολο ελέγχου* (test set). Και τα δύο σύνολα έχουν γνωστές εξόδους για κάθε είσοδο. Θα μπορούσαμε να φανταστούμε τα σύνολα αυτά, ως πίνακες. Κάθε γραμμή του πίνακα, αποτελεί μια είσοδο και ονομάζεται *δείγμα* (sample). Τα δείγματα με την σειρά τους μπορούν να οπτικοποιηθούν ως διανύσματα, που κάθε θέση τους αντιστοιχεί σε κάποιο *χαρακτηριστικό* (attribute), με βάση την τιμή του οποίου θα γίνει η εκπαίδευση. Κάθε δείγμα του συνόλου εκπαίδευσης και ελέγχου, όπως προείπαμε συνοδεύεται από την πραγματική του έξοδο, η οποία ονομάζεται και *ετικέτα* (label). Το σύνολο εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του μοντέλου. Κάθε δείγμα του τροφοδοτείται στο μοντέλο, το οποίο επεξεργάζεται τα χαρακτηριστικά του δείγματος και βάση αυτών κάνει μια πρόβλεψη για την έξοδο που θα έπρεπε να έχει αυτό το δείγμα. Στην συνέχεια, η πρόβλεψη αυτή συγκρίνεται με την πραγματική τιμή – ετικέτα του συγκεκριμένου δείγματος και ανάλογα με το πόσο καλή ήταν η πρόβλεψη τροποποιούνται οι παράμετροι του μοντέλου, ώστε να βελτιωθεί η ακρίβεια των μελλοντικών προβλέψεων. Μετά την ολοκλήρωση της εκπαίδευσης, χρησιμοποιείται το σύνολο ελέγχου για να ελεγχθεί

κατά πόσο ήταν επιτυχημένη η εκπαίδευση. Το σύνολο ελέγχου τροφοδοτείται στο σύστημα και οι προβλέψεις του συστήματος συγκρίνονται με τις ετικέτες του συνόλου ελέγχου. Η σύγκριση αυτή μπορεί να ποσοτικοποιηθεί με κάποιες μετρικές και να χρησιμοποιηθεί ως μέτρο της επιτυχίας της εκπαίδευσης.

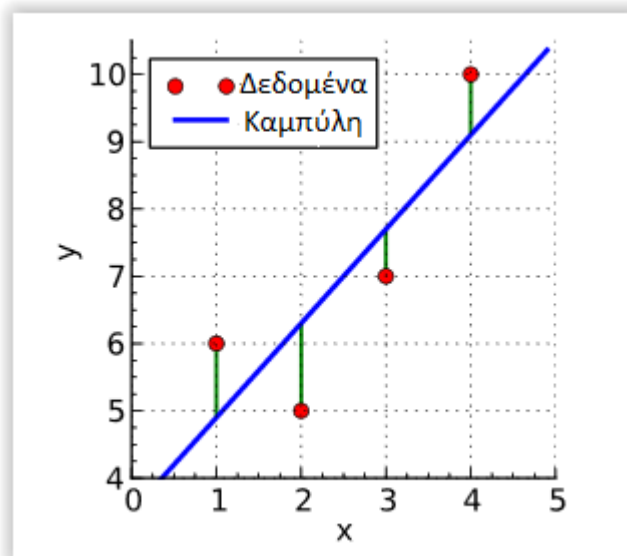
Στη συνέχεια, θα περιγράψουμε τις τεχνικές επιβλεπόμενης μάθησης που χρησιμοποιήσαμε για τους σκοπούς αυτής εδώ της εργασίας. Να σημειωθεί ακόμα ότι το πρόβλημα που καλεστήκαμε να αντιμετωπίσουμε ήταν πρόβλημα παλινδρόμησης (εφόσον στόχος μας ήταν η πρόβλεψη με βάση κάποια χαρακτηριστικά μιας πραγματικής τιμής) και ως εκ τούτου, οι τεχνικές που θα περιγράψουμε παρακάτω είναι προσαρμοσμένες ώστε να εφαρμόζονται σε προβλήματα παλινδρόμησης.

2.2.1 Γραμμική Παλινδρόμηση (Linear Regression)

Στόχος της *γραμμικής παλινδρόμησης* είναι να παράγει μια ευθεία (ή για περισσότερα χαρακτηριστικά – διαστάσεις κάποιο επίπεδο ή υπερεπίπεδο) η οποία θα αντιστοιχίζει τις τιμές των χαρακτηριστικών του δείγματος στην αντίστοιχη πραγματική τιμή – ετικέτα. Προφανώς, η σχέση ανάμεσα στα χαρακτηριστικά και στην πραγματική έξοδο του δείγματος μπορεί να είναι μη γραμμική και έτσι η *γραμμική παλινδρόμηση* αναζητά εκείνη την ευθεία η οποία να ελαχιστοποιεί το σφάλμα ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές για την έξοδο του δείγματος. Ο τρόπος που υπολογίζεται το σφάλμα μπορεί να διαφέρει από περίπτωση σε περίπτωση. Μια συνήθη μετρική, η οποία χρησιμοποιήθηκε και στην δική μας περίπτωση ως συνάρτηση σφάλματος είναι το Υπολειπόμενο Άθροισμα Τετραγώνων (RSS- Residual Sum of Squares) το οποίο υπολογίζεται από τον ακόλουθο τύπο:

$$RSS = \sum_i^n (y_i - f(x_i))^2$$

όπου y_i είναι η τιμή του i -οστού δείγματος και $f(x_i)$ η πρόβλεψη του μοντέλου. Συνήθως η τεχνική *γραμμικής παλινδρόμησης* δεν πετυχαίνει πολύ καλά αποτελέσματα ωστόσο είναι μια απλή και εύκολη τεχνική και μας δίνει ένα κάτω όριο για την ακρίβεια των προβλέψεων μας [5].

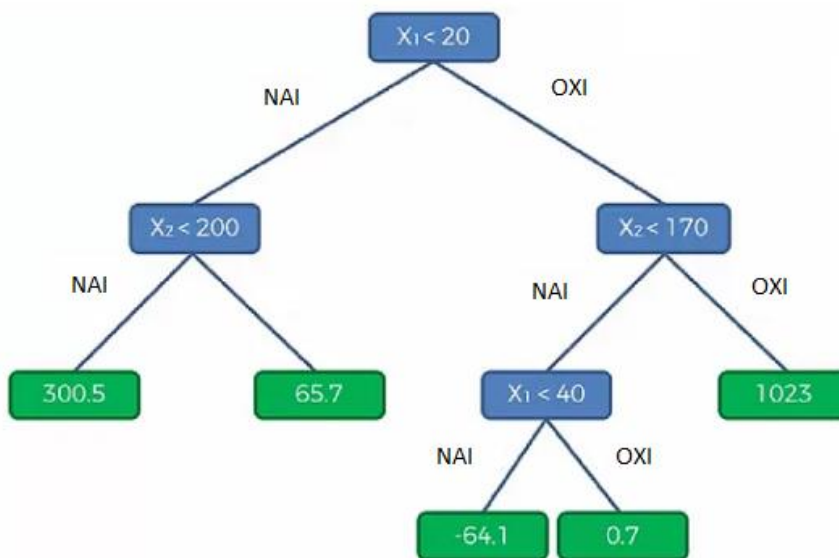
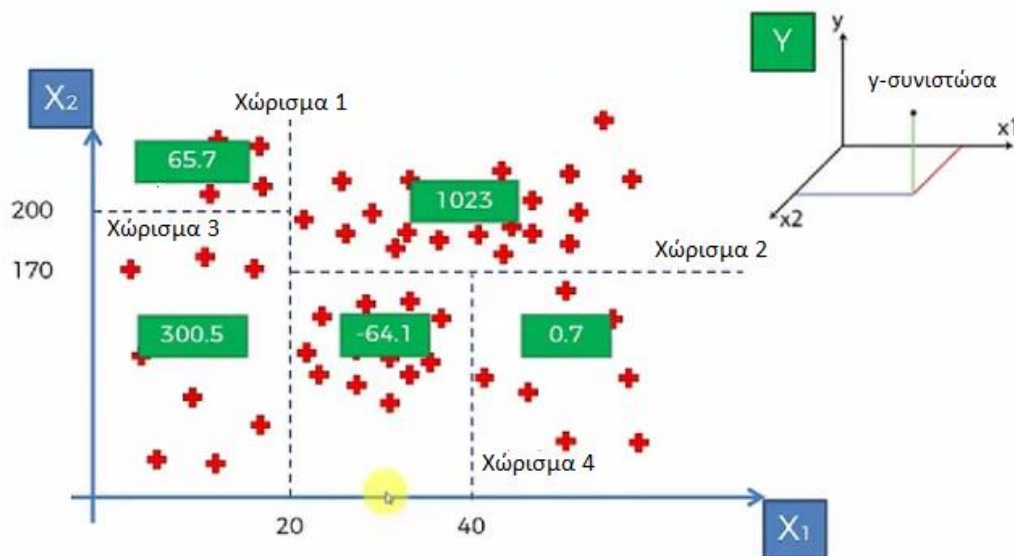


Εικόνα 1: Παράδειγμα λειτουργίας της γραμμικής παλινδρόμησης. Με κόκκινο χρώμα, παρουσιάζονται οι πραγματικές τιμές της εξόδου του δείγματος και με μπλε χρώμα η γραμμή που προκύπτει από το μοντέλο γραμμικής παλινδρόμησης. Με πράσινο χρώμα, παρουσιάζονται οι γραμμές που χρησιμοποιούνται για τον υπολογισμό του σφάλματος. Η θέση και η κλίση της ευθείας μπορεί να τροποποιηθεί ώστε να επιτευχθεί ελαχιστοποίηση του σφάλματος. (Πηγή : [6]).

2.2.2 Δένδρα Απόφασης (Decision Trees) σε προβλήματα Παλινδρόμησης

Τα δένδρα απόφασης κατασκευάζουν μια δεντρική δομή, προκειμένου, ανάλογα με τις τιμές των χαρακτηριστικών κάθε δείγματος, να επιστρέψουν μια προβλεπόμενη τιμή για την έξοδο. Κάθε εσωτερικός κόμβος του δέντρου αντιστοιχεί σε ένα κατώφλι για κάποιο επιλεγθέν χαρακτηριστικό, που εάν αυτό το χαρακτηριστικό το ικανοποιεί, ακολουθείται το μονοπάτι που ξεκινά από αυτόν τον κόμβο. Κάθε φύλλο του δέντρου αντιστοιχεί σε κάποια πρόβλεψη για την έξοδο (ή τις εξόδους) του συγκεκριμένου συνόλου δεδομένων. Η επιλογή του χαρακτηριστικού που ελέγχεται σε κάθε κόμβο, αλλά και του κατωφλίου για αυτό το χαρακτηριστικό, γίνεται μέσα από μια διαδικασία βελτιστοποίησης όπου το κατώφλι επιλέγεται έτσι ώστε να ελαχιστοποιείται κάποια συνάρτηση σφάλματος με βάση την απόσταση ανάμεσα στην προβλεπόμενη (μέση) τιμή των δειγμάτων που καταλήγουν σε αυτή την πλευρά του δέντρου και στην πραγματική τους τιμή. Αφού προκύψει κάποιο κατώφλι για κάθε χαρακτηριστικό, εκείνο το χαρακτηριστικό που δίνει τελικά την μικρότερη τιμή για την συνάρτηση σφάλματος επιλέγεται για να αντιπροσωπεύσει τον κόμβο. Τα φύλλα του δέντρου αποτελούν τις τελικές προβλέψεις για την έξοδο και η τιμή τους υπολογίζεται ως η μέση τιμή των δειγμάτων που κατηγοριοποιούνται σε εκείνο το φύλλο με βάση το δέντρο αποφάσεων. Η διαδικασία αυτή επαναλαμβάνεται αναδρομικά σε κάθε βήμα έως ότου χτιστεί το δέντρο αποφάσεων, ενώ ολοκληρώνεται όταν φτάσουμε σε κάποιο ορισμένο μέγιστο βάθος, ή πετύχουμε «καθαρότητα» των κόμβων (όλες οι πραγματικές τιμές

ταυτίζονται με την προβλεπόμενη) ή δεν μπορεί να γίνει περαιτέρω διαχωρισμός των δειγμάτων με βάση τις ρυθμίσεις του δέντρου [5].



Εικόνα 2: Παράδειγμα της δημιουργίας ενός δένδρου απόφασης. (Πηγή: [7]).

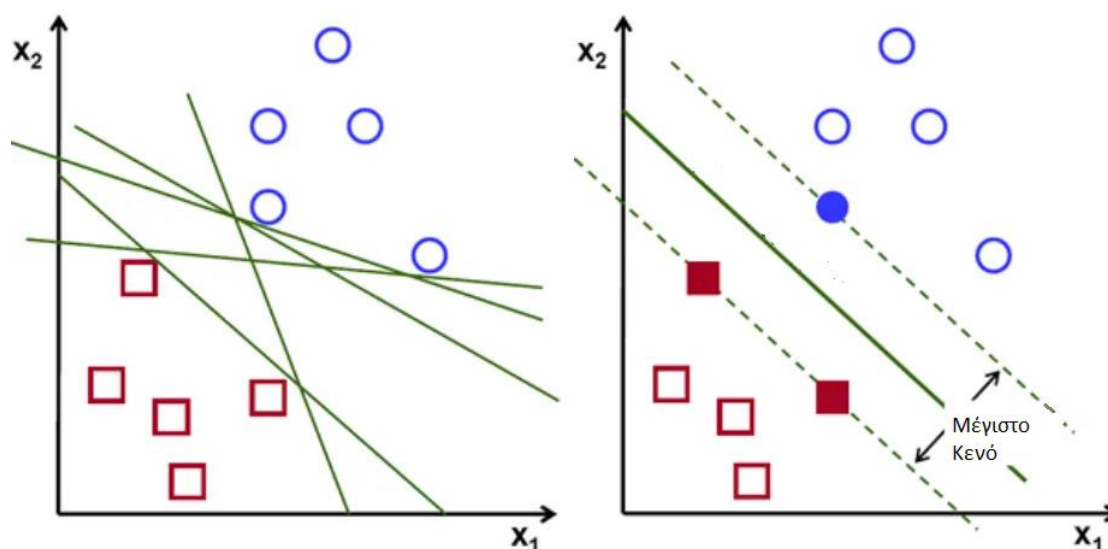
Στο παράδειγμα της Εικόνας 2, τα X_1 και X_2 αποτελούν τα χαρακτηριστικά του συνόλου δεδομένων. Αρχικά, ο χώρος θα πρέπει να χωριστεί σε δύο τμήματα, όπου κάθε τμήμα λαμβάνει ως ετικέτα την μέση τιμή των δειγμάτων που καταλήγουν σε αυτό. Δοκιμάζονται όλα τα πιθανά χωρίσματα για τα χαρακτηριστικά X_1 και X_2 . Το split (χώρισμα) 1 είναι εκείνο που ελαχιστοποιεί το σφάλμα ανάμεσα στις πραγματικές τιμές και στην πρόβλεψη των δειγμάτων. Έτσι το χωρίσμα 1 καθορίζει το περιεχόμενο του πρώτου κόμβου. Στην συνέχεια η διαδικασία επαναλαμβάνεται για τις υποπεριοχές που δημιούργησε το χωρίσμα αυτό. Βλέπουμε δηλαδή ότι το δέντρο αποφάσεων στην ουσία κατανέμει δείγματα σε υποπεριοχές του χώρου

χαρακτηριστικών, οι οποίες χαρακτηρίζονται από μια μέση τιμή – πρόβλεψη. Προφανώς, αν ο χώρος χαρακτηριστικών έχει περισσότερα από δύο χαρακτηριστικά, τότε τα χωρίσματα μας δεν θα αντιστοιχούν σε ευθείες αλλά σε υπερεπίπεδα.

Ένα θεμελιώδες πρόβλημα που αντιμετωπίζουν τα μοντέλα των δέντρων απόφασης είναι η τάση τους να υπερπροσαρμόζονται στα δεδομένα εκπαίδευσης, κατασκευάζοντας μεγάλα και πολύπλοκα δέντρα των οποίων οι κανόνες αποτυπώνουν επακριβώς την εικόνα του συνόλου εκπαίδευσης. Όπως θα εξηγηθεί αργότερα, το πρόβλημα αυτό μπορεί να διορθωθεί και με την χρήση τεχνικών συλλογικής μάθησης [5].

2.2.3 Παλινδρόμηση Διανυσμάτων Υποστήριξης (Support Vector Regression – SVR)

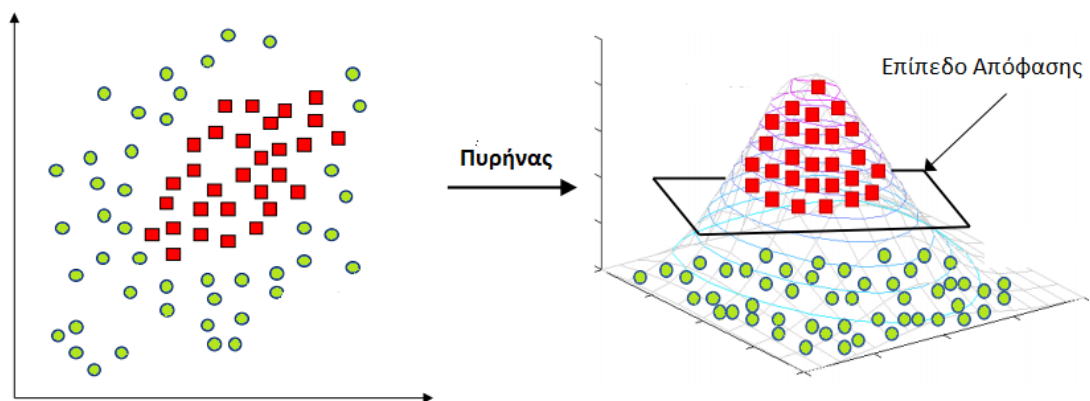
Η τεχνική των διανυσμάτων υποστήριξης έχει χρησιμοποιηθεί ευρέως σε προβλήματα κατηγοριοποίησης και έχει επεκταθεί σε προβλήματα παλινδρόμησης. Σε προβλήματα κατηγοριοποίησης, παρατηρείται ότι η εύρεση κάποιου γραμμικού συνόρου απόφασης μπορεί να πετυχαίνει ελαχιστοποίηση του σφάλματος για τα δεδομένα του συνόλου εκπαίδευσης, αλλά να αποτυγχάνει στην βέλτιστη ταξινόμηση άγνωστων δεδομένων. Η τεχνική των διανυσμάτων υποστήριξης κατορθώνει να το διορθώσει αυτό, αναζητώντας εκείνο το κατώφλι το οποίο θα μεγιστοποιεί το περιθώριο ανάμεσα στα κοντινότερα στην ευθεία διαχωρισμού δείγματα κάθε κατηγορίας.



Εικόνα 3: Παράδειγμα της διαφορετικής αντιμετώπισης της κατηγοριοποίησης δειγμάτων από ένα γραμμικό διαχωριστή και έναν διαχωριστή διανυσμάτων υποστήριξης. Βλέπουμε ότι η κοντινότερη στα τετράγωνα πράσινη γραμμή μπορεί να ελαχιστοποιεί το σφάλμα αλλά είναι επιρρεπής σε λανθασμένες ταξινομήσεις. Η πράσινη γραμμή του δεξιού σχήματος είναι το βέλτιστο υπερεπίπεδο. (Πηγή: [8]).

Η θέση του κατωφλίου υπολογίζεται με την επίλυση μιας εξίσωσης και τα κοντινότερα δείγματα στην περιοχή διαχωρισμού των κλάσεων είναι τα μόνα που

παίζουν ρόλο στον καθορισμό της θέσης του κατωφλίου. Μάλιστα, τα δείγματα αυτά ονομάζονται *διανύσματα υποστήριξης* και από αυτά προκύπτει το όνομα της μεθόδου. Στην περίπτωση που τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα τότε υπάρχει η δυνατότητα να χρησιμοποιηθεί το *τέχνασμα του πυρήνα* (kernel trick), όπου με την χρήση κάποιων *συναρτήσεων πυρήνα* (kernel functions) μετασχηματίζονται τα δείγματα σε έναν νέο χώρο καταστάσεων, στον οποίον, αν υπάρχουν αρκετά πολλά διαστάσεις, τα δείγματα, σχεδόν με βεβαιότητα, θα είναι γραμμικά διαχωρίσιμα. Το υπερεπίπεδο διαχωρισμού που δημιουργείται στον υψηλότερο χώρο διαστάσεων μπορεί να μετασχηματίζεται, στον χαμηλότερο χώρο διαστάσεων, σε μια πολύπλοκη καμπύλη διαχωρισμού. Οι συναρτήσεις πυρήνα διευκολύνουν τους υπολογισμούς, διότι αντί να μεταφέρουν τα δείγματα στο νέο χώρο υψηλότερων διαστάσεων, προσομοιώνουν αυτήν την διαδικασία, υπολογίζοντας τους συσχετισμούς ανάμεσα σε ζεύγη δειγμάτων [5]. Μια συνάρτηση πυρήνα που χρησιμοποιείται ευρέως είναι η *ακτινική συνάρτηση βάσης* (Radial Basis Function - RBF), η οποία μεταφέρει τα δείγματα σε υψηλότερο χώρο διαστάσεων, μεγεθύνοντας, κατά κάποιο τρόπο, τον κέντρο, με την χρήση γκαουσιανής συνάρτησης, όπως φαίνεται στην Εικόνα 4.



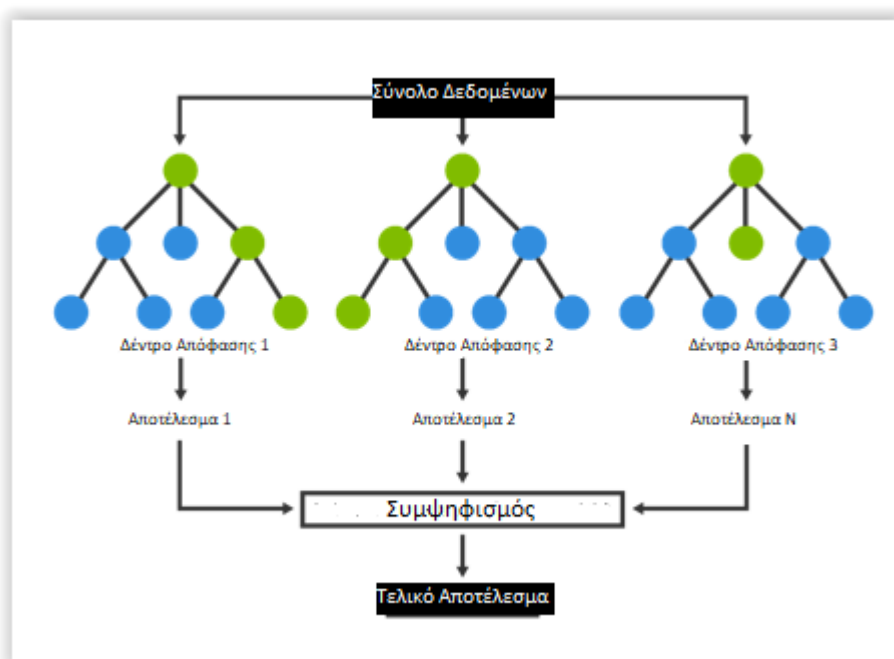
Εικόνα 4: Συνάρτηση πυρήνα RBF. (Πηγή: [9]).

Με παρόμοιο σκεπτικό, λειτουργούν τα διανύσματα υποστήριξης και για προβλήματα παλινδρόμησης. Αντί να γίνει προσπάθεια να βρεθεί ευθεία που να ταιριάζει καλύτερα στα δεδομένα εκπαίδευσης, που ωστόσο μπορεί να μην πετυχαίνει εξίσου καλά αποτελέσματα σε άγνωστα δείγματα, βρίσκεται η ευθεία η οποία μεγιστοποιεί το περιθώριο ανάμεσα στα κοντινότερα στην ευθεία δεδομένα και ελαχιστοποιεί το σφάλμα. Αν δεν είναι εφικτό η ευθεία αυτή να βρεθεί στον τρέχοντα χώρο διαστάσεων, μπορούν να χρησιμοποιηθούν οι συναρτήσεις πυρήνα για να γίνει μετάβαση σε χώρο περισσότερων διαστάσεων [10].

2.2.4 Τυχαίο Δάσος (Random Forest) σε προβλήματα Παλινδρόμησης

Τα μοντέλα τυχαίου δάσους αναπτύχθηκαν προκειμένου να διορθώσουν το πρόβλημα υπερπροσαρμογής των δέντρων απόφασης. Στηρίζονται σε τεχνικές

συλλογικής μάθησης και συγκεκριμένα σε μια εκδοχή της τεχνικής του σακουλιάσματος, η οποία θα περιγραφεί διεξοδικότερα στην συνέχεια. Η τεχνική τυχαίου δάσους κατασκευάζει δέντρα αποφάσεων με μία τυχαιότητα ως προς το ποια δείγματα χρησιμοποιούνται για να κατασκευαστεί κάθε δέντρο αλλά και ποια χαρακτηριστικά μπορούν να επιλεχθούν σε κάθε βήμα κατασκευής του δέντρου, ενώ, ακόμα και η επιλογή της τιμής κατωφλίου μπορεί να γίνει με εν μέρει τυχαίο τρόπο. Από τα δέντρα αποφάσεων παίρνουμε προβλέψεις και η μέση τιμή αυτών των προβλέψεων μας δίνει την τελική πρόβλεψη του μοντέλου. Σημαντική υπερπαραμέτρος του τυχαίου δάσους είναι ο αριθμός των δέντρων απόφασης που θα χρησιμοποιηθούν, όπου με τον όρο *υπερπαραμέτρος* αναφερόμαστε σε μια παράμετρο της οποίας η τιμή επιλέγεται πριν την εκπαίδευση και χρησιμοποιείται για να ελέγξει την διαδικασία της εκπαίδευσης, ενώ οι τιμές των απλών παραμέτρων προκύπτουν από την εκπαίδευση. Η κατασκευή τυχαίων δασών δίνει καλά αποτελέσματα και μπορεί να γίνει αποδοτικά [5]. Στην Εικόνα 5, βλέπουμε ένα τυχαίο δάσος, το οποίο αποτελείται από N δέντρα απόφασης και τα αποτελέσματά τους συμψηφίζονται ώστε να προκύψει το τελικό αποτέλεσμα.



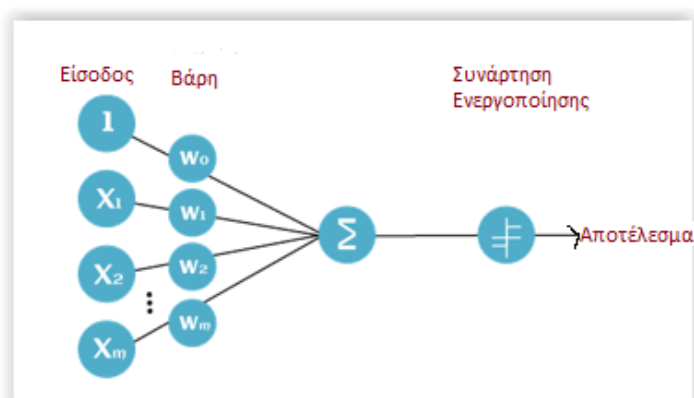
Εικόνα 5 : Μοντέλο ενός τυχαίου δάσους. Οι προβλέψεις των N δέντρων απόφασης συμψηφίζονται ώστε να προκύψει το τελικό αποτέλεσμα. (Πηγή: [11]).

2.2.5 Νευρώνας Πολλαπλών Στρωμάτων (Multiple Layer Perceptron – MLP)

Ένας πολυστρωματικός νευρώνας, ή σωστότερα ένα δίκτυο πολυστρωματικών νευρώνων αποτελεί μια ειδική κατηγορία ενός τεχνητού νευρωνικού δικτύου

(Artificial Neural Network – ANN) πρόσθιας τροφοδότησης ή απλώς νευρωνικού δικτύου (η πρόσθια τροφοδότηση συνήθως εννοείται, αν δεν αναφέρεται κάτι διαφορετικό) [12]. Τα νευρωνικά δίκτυα ονομάζονται έτσι, επειδή οι δομή τους προσομοιώνει την δομή των νευρώνων του εγκεφάλου [5]. Τα δίκτυα πρόσθιας τροφοδότησης έχουν συνδέσεις μόνο προς μια κατεύθυνση – δηλαδή, σχηματίζουν έναν κατευθυνόμενο ακυκλικό γράφο με καθορισμένους κόμβους εισόδου και εξόδου [5]. Τα δίκτυα πολυστρωματικών νευρώνων είναι πλήρως διασυνδεδεμένα, δηλαδή κάθε νευρώνας ενός επιπέδου συνδέεται με κάθε άλλο νευρώνα του επόμενου επιπέδου [5].

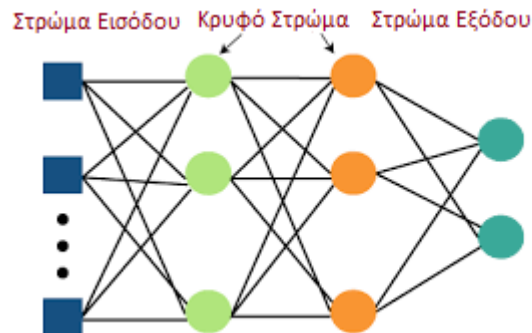
Ένα δίκτυο πολυστρωματικού νευρώνα βασίζεται στην δομική μονάδα του νευρώνα perceptron η οποία παρουσιάζεται στην Εικόνα 6. Ένας νευρώνας perceptron εμφανίζει στην έξοδό του το άθροισμα το σταθμισμένων χαρακτηριστικών της εισόδου συν κάποια ακόμα παράμετρο (bias) ακολουθούμενο από μια συνάρτηση ενεργοποίησης. Οι στάθμες με τις οποίες πολλαπλασιάζονται τα επιμέρους χαρακτηριστικά κάθε δείγματος πριν αθροιστούν, ονομάζονται *βάρη* (weights). Η *συνάρτηση ενεργοποίησης* (activation function) παίρνει το άθροισμα αυτό σαν είσοδο και παράγει την έξοδο του νευρώνα. Ένας πολυστρωματικός νευρώνας αποτελείται από πολλούς τέτοιους νευρώνες οι οποίοι οργανώνονται σε διαδοχικά επίπεδα. Πάντα υπάρχει ένα στρώμα εισόδου και ένα εξόδου ωστόσο μπορούν να υπάρχουν ένα ή περισσότερα κρυφά στρώματα. Ένα δίκτυο με ένα μόνο κρυφό στρώμα ονομάζεται δίκτυο *ρηχής μάθησης* (swallow learning) ενώ, ένα δίκτυο με δύο ή περισσότερα κρυφά στρώματα ονομάζεται δίκτυο *βαθιάς μάθησης* (deep learning) [5].



Εικόνα 6: Μοντέλο ενός νευρώνα Perceptron. (Πηγή: [13]).

Στόχος της διαδικασίας μάθησης σε ένα νευρωνικό δίκτυο είναι ο υπολογισμός των βαρών των ακμών που συνδέουν τους επιμέρους νευρώνες του δικτύου. Τα βάρη σε κάθε νευρώνα υπολογίζονται από ένα αλγόριθμο που ονομάζεται *κατάβαση κλίσης* (gradient descent), ο οποίος ανανεώνει τα βάρη κάθε νευρώνα προκειμένου να πετύχει τη βελτιστοποίηση μιας συνάρτησης σφάλματος, χρησιμοποιώντας μια στρατηγική *οπισθοδιάδοσης* (backpropagation) για να υπολογίσει αποτελεσματικά

την ανανέωση των βαρών. Η ανανέωση των βαρών γίνεται βάση ενός *ρυθμού μάθησης* (learning rate), ο οποίος καθορίζει πόσο πολύ θα μεταβάλλονται τα βάρη σε κάθε ανανέωση. Συγκεκριμένα η διαδικασία έχει ως εξής: Κάθε δείγμα εισάγεται στο νευρωνικό δίκτυο μέσω του στρώματος εισόδου, το οποίο έχει τόσους νευρώνες όσα και τα χαρακτηριστικά κάθε δείγματος. Στην συνέχεια, τα δεδομένα περνούν μια επεξεργασία μέσω των κρυφών στρωμάτων, με την χρήση των βαρών και των συναρτήσεων ενεργοποίησης, για να προκύψει τελικά κάποια πρόβλεψη στο επίπεδο εξόδου, το οποίο έχει τόσους νευρώνες όσες και οι δυνατές κατηγορίες στα προβλήματα κατηγοριοποίησης ή μόνο έναν νευρώνα αν πρόκειται για πρόβλημα παλινδρόμησης. Με βάση το σφάλμα σε σχέση με την πραγματική τιμή, τα βάρη αναπροσαρμόζονται μέσω της κατάβασης κλίσης και με την χρήση της οπισθοδιάδοσης και η διαδικασία επαναλαμβάνεται για όλα τα δείγματα, πιθανώς για περισσότερους από έναν κύκλους εκπαίδευσης. Μετά την ολοκλήρωση της εκπαίδευσης τα βάρη έχουν πάρει την τελική τους τιμή και το δίκτυο μπορεί να χρησιμοποιηθεί για την πρόβλεψη της εξόδου νέων δεδομένων.



Εικόνα 7: Ένας πολυστρωματικός νευρώνας (MLP). (Πηγή: [14]).

Τέλος, παρουσιάζονται επιγραμματικά κάποιες συναρτήσεις ενεργοποίησης που χρησιμοποιούνται ευρέως στα νευρωνικά δίκτυα: η *σιγμοειδής*, η *ReLU* και η *υπερβολική εφαπτομένη* [5].

- Η σιγμοειδής συνάρτηση:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- Η ανορθωμένη γραμμική μονάδα:

$$ReLU(x) = \max(0, x)$$

- Η υπερβολική εφαπτομένη:

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$

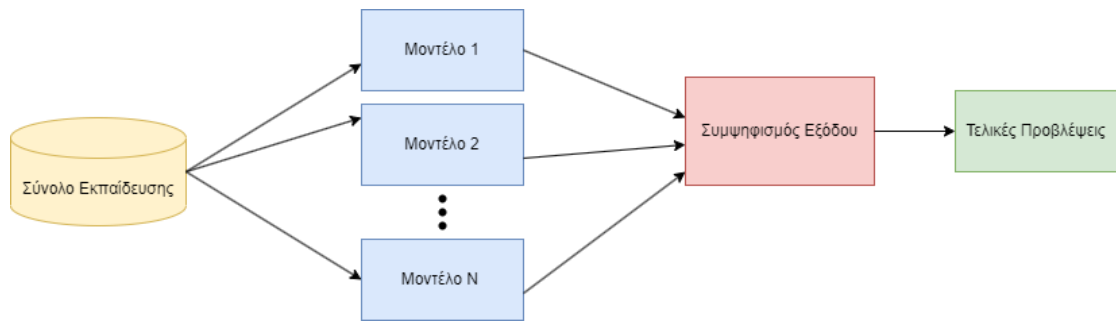
2.3 Συλλογική Μάθηση (Ensemble Learning)

Οι τεχνικές *συλλογικής μάθησης* (ensemble learning) είναι υποκατηγορία των τεχνικών επιβλεπόμενης μάθησης, ωστόσο, λόγω του εύρους των μεθόδων τους παρουσιάζονται ξεχωριστά. Η κεντρική φιλοσοφία της συλλογικής μάθησης έχει να κάνει με την δημιουργία μιας συλλογής επιμέρους μοντέλων από τα οποία προκύπτουν προβλέψεις και οι προβλέψεις αυτές συνδυάζονται, με διάφορες μεθόδους, προκειμένου να προκύψει κάποιο τελικό αποτέλεσμα [5]. Τα επιμέρους μοντέλα ονομάζονται *βασικά μοντέλα* (base models) και ο συνδυασμός τους *συλλογικό μοντέλο* (ensemble model). Ο λόγος που χρησιμοποιούνται οι τεχνικές συλλογικής μάθησης είναι ότι επιτυγχάνουν μείωση της μεροληψίας αλλά και της διακύμανσης. Με τον όρο *μεροληψία* (bias) περιγράφονται οι απλουστευτικές υποθέσεις που κάνει κάποιο μοντέλο για την μορφή της εξόδου προκειμένου να απλοποιήσει τους υπολογισμούς, ενώ με τον όρο *διακύμανση* (variance) ποσοτικοποιείται η επιρροή του συνόλου εκπαίδευσης στην μορφή της εξόδου [15]. Με την συλλογική μάθηση μειώνεται η μεροληψία καθώς ο συνδυασμός περισσότερων βασικών μοντέλων εισάγει μια επιπλέον εκφραστικότητα στην μορφή των εξόδων που μπορούν να παραχθούν (για παράδειγμα, το μοντέλο της γραμμικής παλινδρόμησης είναι πολύ πιο μεροληπτικό σε σχέση με τον συνδυασμό περισσότερων μοντέλων γραμμικής παλινδρόμησης). Επιπλέον, με την συλλογική μάθηση μειώνεται η διακύμανση επειδή υπάρχει μεγαλύτερη πιθανότητα να γίνουν περισσότερες σωστές ταξινομήσεις εξαιτίας του μεγαλύτερου αριθμού μοντέλων [5].

Υπάρχουν τέσσερις διαφορετικές προσεγγίσεις συλλογικής μάθησης: ο Συμψηφισμός (Averaging), το Σακούλιασμα (Bagging), η Στοίβαξη (Stacking) και η Ενδυνάμωση (Boosting). Στην συνέχεια, θα παρουσιαστούν οι μέθοδοι που χρησιμοποιήθηκαν από κάθε κατηγορία.

2.3.1 Παλινδρόμηση Ψηφίσματος (Voting Regression)

Το μοντέλο της παλινδρόμησης ψηφίσματος ανήκει στην κατηγορία των μοντέλων συμψηφισμού και στην πραγματικότητα αποτελεί την απλούστερη τεχνική συλλογικής μάθησης. Χρησιμοποιεί πλήρως εκπαιδευμένα βασικά μοντέλα, στο σύνολο του συνόλου εκπαίδευσης και βγάζει τον μέσο όρο των προβλέψεών τους. Στην Εικόνα 8 παρατηρούμε την δομή ενός τέτοιου μοντέλου. Χρησιμοποιούνται N μοντέλα βάσης, τα οποία εκπαιδεύονται ανεξάρτητα σε όλο το σύνολο εκπαίδευσης και κατόπιν τα αποτελέσματά τους συμψηφίζονται ώστε να προκύψει η τελική πρόβλεψη.



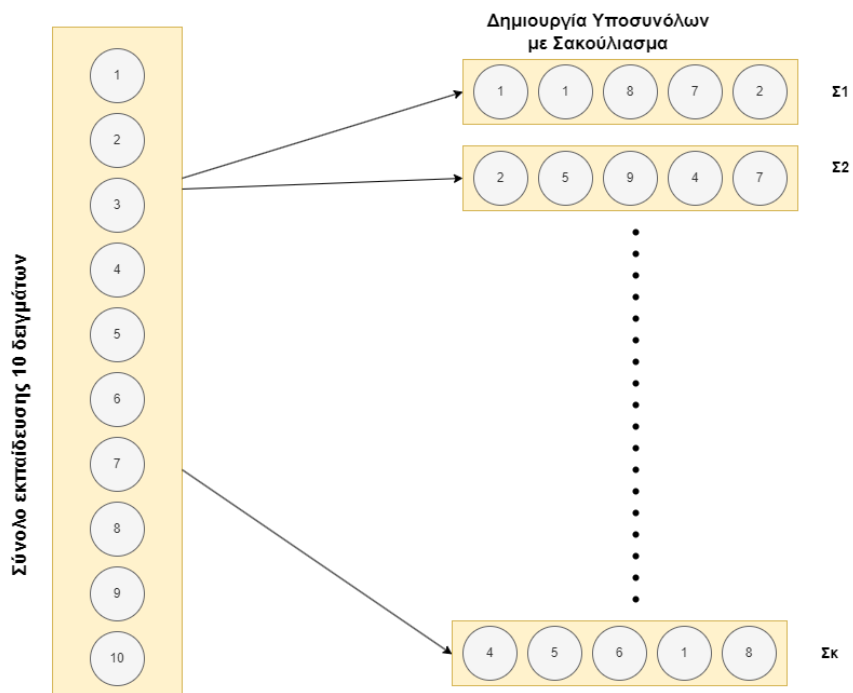
Εικόνα 8: Η δομή ενός μοντέλου Παλινδρόμησης Ψηφίσματος.

2.3.2 Σακούλιασμα (Bagging) σε προβλήματα Παλινδρόμησης

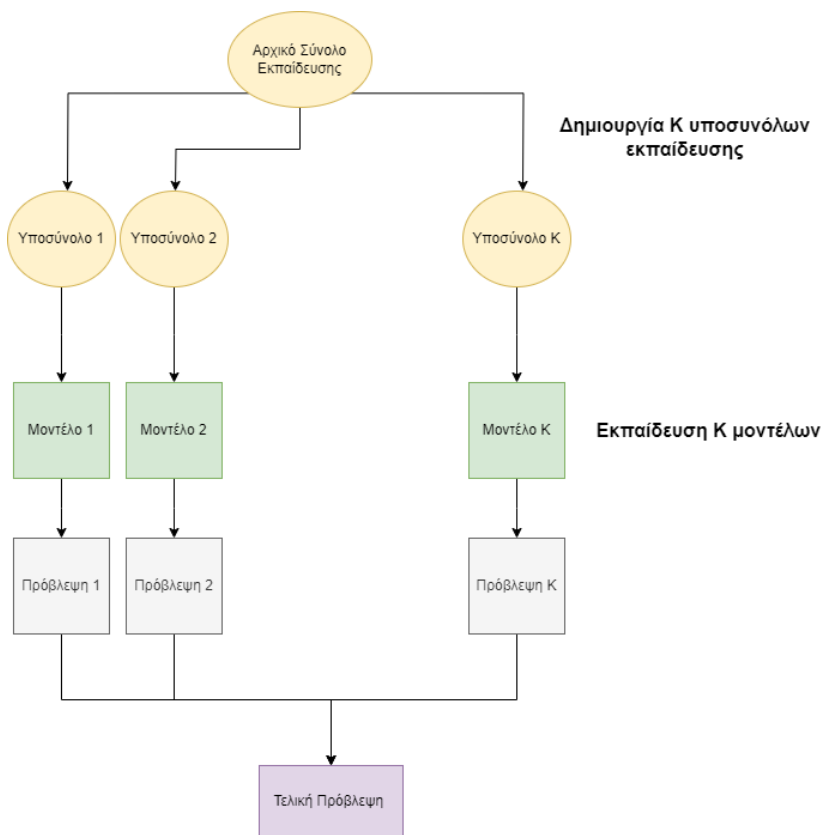
Κατά την τεχνική σακουλιάσματος δημιουργούνται K διακριτά υποσύνολα εκπαίδευσης, τα οποία κατόπιν χρησιμοποιούνται για την εκπαίδευση των K βασικών μοντέλων του συλλογικού μοντέλου. Η διαδικασία που ακολουθείται για να δημιουργηθούν υποσύνολα του συνόλου εκπαίδευσης είναι η εξής: Από το σύνολο εκπαίδευσης, επιλέγουμε τυχαία τον επιθυμητό αριθμό δειγμάτων (έστω N) με αντικατάσταση. Αυτό σημαίνει ότι αν ένα δείγμα επιλεγεί για κάποιο υποσύνολο του συνόλου εκπαίδευσης, μπορεί να επανεπιλεγεί είτε για το ίδιο είτε για κάποιο άλλο υποσύνολο. Η τεχνική αυτή, για την δημιουργία υποσυνόλων, μπορεί, παραδόξως, να επιτύχει πολύ καλά αποτελέσματα στην εξαγωγή στατιστικών χαρακτηριστικών του συνόλου, από κάποιο υποσύνολο αυτού, και για αυτό θα μπορούσε να χρησιμοποιηθεί και σε τεχνικές μηχανικής μάθησης όπου το σύνολο εκπαίδευσης δεν είναι επαρκώς μεγάλο ώστε να προκύψουν ικανοποιητικές προβλέψεις [5].

Μετά την δημιουργία των K υποσυνόλων εκπαίδευσης, καθένα από αυτά χρησιμοποιείται για την ανεξάρτητη εκπαίδευση ενός από τα K βασικά μοντέλα. Τα βασικά μοντέλα που χρησιμοποιούνται συνήθως στην τεχνική σακουλιάσματος είναι δέντρα απόφασης, χωρίς ωστόσο αυτό να σημαίνει ότι δεν μπορούν να χρησιμοποιηθούν και άλλα μοντέλα. Κάθε βασικό μοντέλο παράγει μια πρόβλεψη και ο μέσος όρος των προβλέψεων αυτών μας δίνει την τελική πρόβλεψη του συλλογικού μοντέλου παλινδρόμησης. Στην Εικόνα 10 παρουσιάζεται σχηματικά αυτή η διαδικασία [5].

Η τεχνική του σακουλιάσματος πετυχαίνει την μείωση της διακύμανσης, οπότε συχνά επιλέγεται σε περιπτώσεις που παρατηρείται υπερπροσαρμογή στο σύνολο εκπαίδευσης [5].



Εικόνα 9: Παράδειγμα του τρόπου δημιουργίας υποσυνόλων στην τεχνική του σακουλιάσματος.



Εικόνα 10: Τρόπος λειτουργίας της τεχνικής Σακουλιάσματος.

2.3.3 Ενδυνάμωση (Boosting) σε προβλήματα Παλινδρόμησης

Η καινοτομία που εισαγάγουν οι τεχνικές ενδυνάμωσης αφορά στη χρήση ενός σταθμισμένου συνόλου εκπαίδευσης [5]. Συγκεκριμένα κάθε δείγμα λαμβάνει ένα διαφορετικό βάρος ανάλογα με την δυσκολία του εκάστοτε μοντέλου να δώσει σωστή πρόβλεψη. Τα δείγματα που δίνουν μεγαλύτερο σφάλμα κατά την πρόβλεψή τους από κάποιο βασικό μοντέλο, στο επόμενο βασικό μοντέλο αποκτούν μεγαλύτερο βάρος. Η τεχνική της ενδυνάμωσης συνεχίζει να εισαγάγει μοντέλα μέχρι να γίνει σωστή πρόβλεψη για όλα τα δείγματα. Υπό την έννοια αυτή, οι τεχνικές ενδυνάμωσης είναι άπληστες, καθώς άπαξ και εισάγουν κάποιο βασικό μοντέλο αυτό σίγουρα θα συμπεριληφθεί στο συλλογικό μοντέλο, ενώ η όλη διαδικασία είναι ακολουθιακή. Όταν ολοκληρωθεί η εισαγωγή βασικών μοντέλων, η τελική πρόβλεψη προκύπτει από ένα σταθμισμένο μέσο όρο των βασικών μοντέλων, όπου μοντέλα με ποιοτικότερες προβλέψεις αποκτούν μεγαλύτερο βάρος [5]. Αποδεικνύεται ότι ακόμα και για πολύ απλά μοντέλα (weak estimators), η τεχνική ενίσχυσης μπορεί να δώσει εξαιρετικά υψηλή ακρίβεια στο σύνολο εκπαίδευσης, για έναν ικανό αριθμό μοντέλων [5]. Παρακάτω θα περιγράψουμε δύο τεχνικές ενίσχυσης: την τεχνική AdaBoost και την τεχνική Διαβαθμισμένης Ενδυνάμωσης (Gradient Boosting).

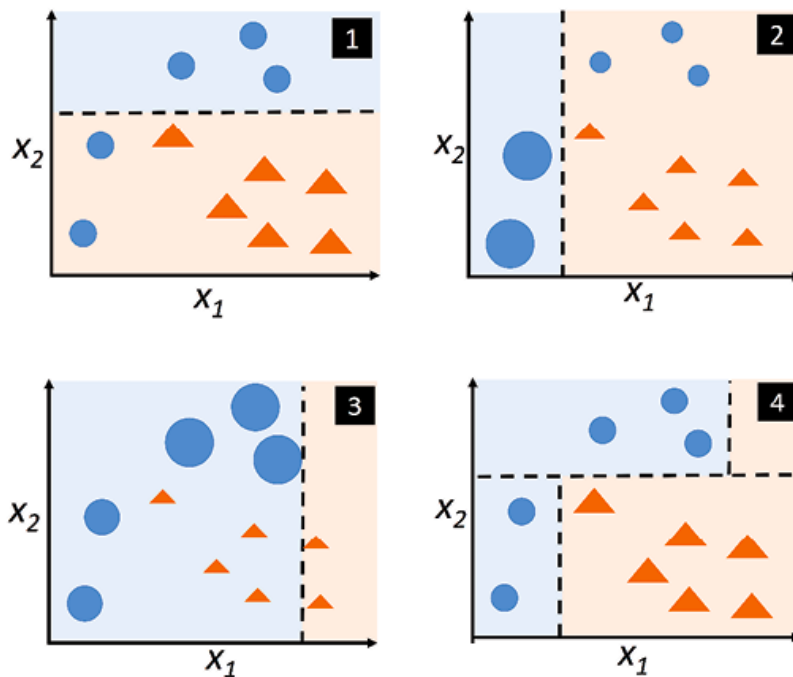
2.3.3.1 Τεχνική AdaBoost σε προβλήματα Παλινδρόμησης

Η AdaBoost είναι από τις πρώτες τεχνικές ενίσχυσης που εφαρμόστηκαν στην πράξη. Οι τεχνικές ενίσχυσης, όπως εξηγήθηκε, χρησιμοποιούν πολύ απλά βασικά μοντέλα και λειτουργούν σειριακά αξιοποιώντας τις αστοχίες κάθε επιπέδου προκειμένου να βελτιώσουν την απόδοση του επόμενου επιπέδου. Συγκεκριμένα, ο αλγόριθμος AdaBoost Αλγόριθμος ακολουθεί τα παρακάτω βήματα [16], [17]:

1. Κάθε δείγμα του συνόλου εκπαίδευσης αντιστοιχίζεται σε ένα βάρος. Τα βάρη αρχικοποιούνται όλα στην ίδια τιμή.
2. Κατασκευάζεται ένα απλό δέντρο απόφασης ενός επιπέδου με μόνο ρίζα και φύλλα (decision stump), με βάση το οποίο γίνεται πρόβλεψη για την έξοδο κάθε δείγματος.
3. Αφού ολοκληρωθεί η εκπαίδευση, υπολογίζεται η συνάρτηση σφάλματος, η οποία χρησιμοποιείται έτσι ώστε να ανανεωθούν κατάλληλα τα βάρη των δειγμάτων.
4. Ανανεώνονται τα βάρη των δειγμάτων. Τα δείγματα με τις χειρότερες προβλέψεις αποκτούν μεγαλύτερο βάρος σε σχέση με τα δείγματα με τις καλύτερες προβλέψεις.
5. Τα δείγματα τοποθετούνται σε κουβάδες (buckets) όπου το εύρος των κουβάδων καθορίζεται από το βάρος των δειγμάτων. Δείγματα με μεγαλύτερα βάρη τοποθετούνται σε κουβάδες μεγαλύτερου εύρους.

6. Δημιουργείται νέο σύνολο εκπαίδευσης, με βάση τους κουβάδες που έχουν δημιουργηθεί στο προηγούμενο βήμα. Αρχικά, επιλέγεται κάποια τυχαία τιμή από το 0 μέχρι το 1. Ο κουβάς μέσα στον οποίο θα πέσει αυτή η τιμή καθορίζει και το αντίστοιχο δείγμα που θα επιλεγεί να συμπεριληφθεί στο νέο σύνολο εκπαίδευσης. Προφανώς τα δείγματα σε κουβάδες μεγαλύτερου εύρους επιλέγονται συχνότερα.
7. Τα βήματα 2 – 6 επαναλαμβάνονται μέχρι να ισχύσει κάποια συνθήκη τερματισμού.
8. Για να προκύψει η πρόβλεψη του συλλογικού μοντέλου τα αποτελέσματα των μοντέλων βάσης συμψηφίζονται, με τα μοντέλα που μας έδωσαν τις καλύτερες προβλέψεις αποκτούν μεγαλύτερα βάρη στην λήψη της τελικής απόφασης.

Η Εικόνα 11 παρουσιάζει ένα παράδειγμα εφαρμογής της AdaBoost τεχνικής σε πρόβλημα κατηγοριοποίησης. Με την χρήση τεσσάρων μόνο απλών δέντρων απόφασης και την αναπροσαρμογή των βαρών των λανθασμένα ταξινομημένων δειγμάτων επιτυγχάνεται απόλυτα ακριβής ταξινόμηση. Η τεχνική AdaBoost λειτουργεί με τον ίδιο τρόπο και για προβλήματα παλινδρόμησης.



Εικόνα 11: Παράδειγμα της λειτουργίας της τεχνικής AdaBoost σε πρόβλημα κατηγοριοποίησης. (Πηγή: [18]).

2.3.3.2 Διαβαθμισμένη Ενδυνάμωση (Gradient Boosting) σε προβλήματα Παλινδρόμησης

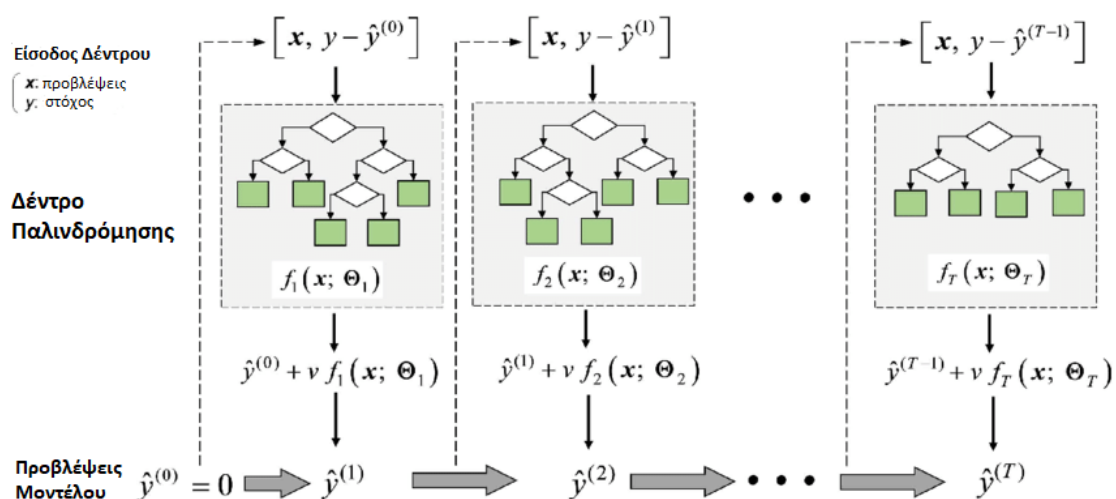
Η μέθοδος διαβαθμισμένης ενδυνάμωσης ανήκει και αυτή στις τεχνικές ενδυνάμωσης. Περιλαμβάνει πολλά βασικά μοντέλα, συνήθως μικρά δέντρα απόφασης, τα οποία διατάσσονται σειριακά, με καθένα από αυτά να τροφοδοτεί

την εκπαίδευση του επόμενου του και με τα αποτελέσματα τους να συμψηφίζονται, ώστε να προκύψει το τελικό αποτέλεσμα. Η βασική ιδέα της διαβαθμισμένης ενδυνάμωσης είναι ότι κάθε μοντέλο ζητάει από το επόμενο μοντέλο να προβλέψει τις αποκλίσεις ανάμεσα στις πραγματικές τιμές του και τις προβλέψεις του, όπου :

$$\text{απόκλιση} = \text{πραγματική τιμή} - \text{προβλεπόμενη τιμή}$$

Η τελική πρόβλεψη προκύπτει αθροίζοντας τις προβλέψεις των επιμέρους μοντέλων. Προκειμένου να αποφευχθεί η υπερπροσαρμογή οι προβλέψεις κάθε επιπέδου πολλαπλασιάζονται και με έναν ρυθμό μάθησης, πριν αθροιστούν στο τελικό αποτέλεσμα [5]. Συγκεκριμένα ο αλγόριθμος έχει ως εξής [19]:

1. Το μοντέλο αρχικοποιείται με μια πρώτη πρόβλεψη, η οποία συνήθως είναι ο μέσος όρος των εξόδων για το σύνολο εκπαίδευσης.
2. Δημιουργείται ένα πρώτο δέντρο απόφασης (συνήθως, μπορεί να χρησιμοποιηθεί και άλλο βασικό μοντέλο). Αυτό το δέντρο απόφασης, έχει ως σύνολο εκπαίδευσης τα δείγματα του αρχικού συνόλου εκπαίδευσης, ωστόσο του ζητείται να προβλέψει την απόκλιση ανάμεσα στην έξοδο κάθε δείγματος και στην πρόβλεψη που έγινε για αυτό το δείγμα.
3. Οι προβλέψεις του μοντέλου αυτού, πολλαπλασιάζονται με έναν ρυθμό μάθησης και προστίθενται στις αρχικές προβλέψεις. Το άθροισμα αυτό αποτελεί την τελική πρόβλεψη του νέου μοντέλου.
4. Τα βήματα 2 και 3 επαναλαμβάνονται μέχρι να ισχύσει κάποια συνθήκη τερματισμού. Κάθε νέο μοντέλο προσπαθεί να προβλέψει την απόκλιση της πρόβλεψης του προηγούμενου μοντέλου από την πραγματική έξοδο και η πρόβλεψη του, πολλαπλασιασμένη με τον ρυθμό μάθησης, προστίθεται στο άθροισμα των προβλέψεων των προηγούμενων μοντέλων.
5. Η τελική πρόβλεψη του συλλογικού μοντέλου είναι το άθροισμα που προκύπτει από τις προβλέψεις των βασικών μοντέλων.

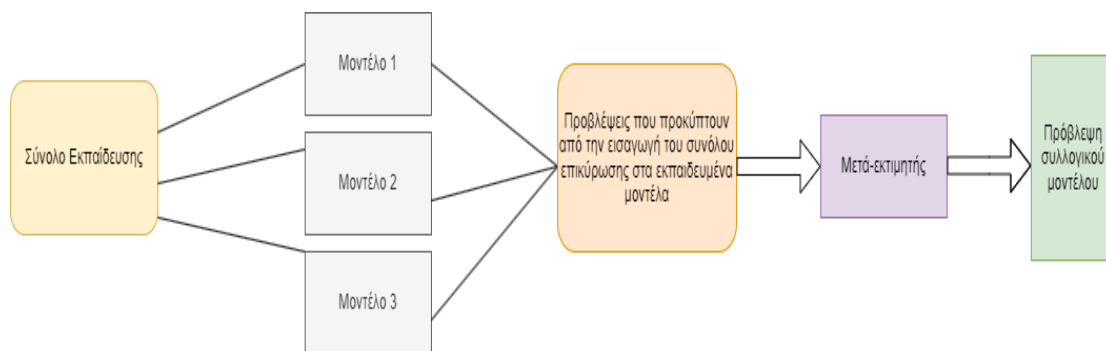


Εικόνα 12: Τρόπος λειτουργίας ενός διαβαθμισμένου ενισχυτή. (Πηγή: [20])

Στην Εικόνα 12 παρουσιάζεται, πολύ αναλυτικά, η δομή ενός διαβαθμισμένου ενισχυτή.

2.3.4 Στοίβαξη (Stacking) σε προβλήματα Παλινδρόμησης

Η μέθοδος στοίβαξης, διαφοροποιείται από τις τεχνικές σακουλιάσματος και ενίσχυσης, καθώς χρησιμοποιεί διαφορετικά μεταξύ τους μοντέλα, με υψηλότερες επιδόσεις πρόβλεψης από τα απλά δέντρα απόφασης [5]. Η ιδέα της είναι απλή: χρησιμοποιεί ένα σύνολο από k μοντέλα, τα οποία εκπαιδεύονται ανεξάρτητα μεταξύ τους και οι προβλέψεις τους τροφοδοτούνται σε ένα *μετά-εκτιμητή* (meta-estimator), ο οποίος κάνει και την τελική πρόβλεψη. Η επιλογή των βασικών μοντέλων όσο και του μετά-εκτιμητή αποτελεί μια σημαντική υπερπαράμετρο του συλλογικού μοντέλου. Τα μοντέλα εκπαιδεύονται σε διαφορετικά υποσύνολα του συνόλου εκπαίδευσης, για να εξασφαλιστεί όσο το δυνατόν μεγαλύτερη διαφοροποίηση μεταξύ τους. Αφού ολοκληρωθεί η εκπαίδευση, στα βασικά μοντέλα τροφοδοτούνται άγνωστα μέχρι τότε δεδομένα και οι προβλέψεις χρησιμεύουν ως σύνολο εκπαίδευσης του μετά-εκτιμητή, ο οποίος οφείλει να μάθει πώς να τις συνδυάζει αποτελεσματικά.



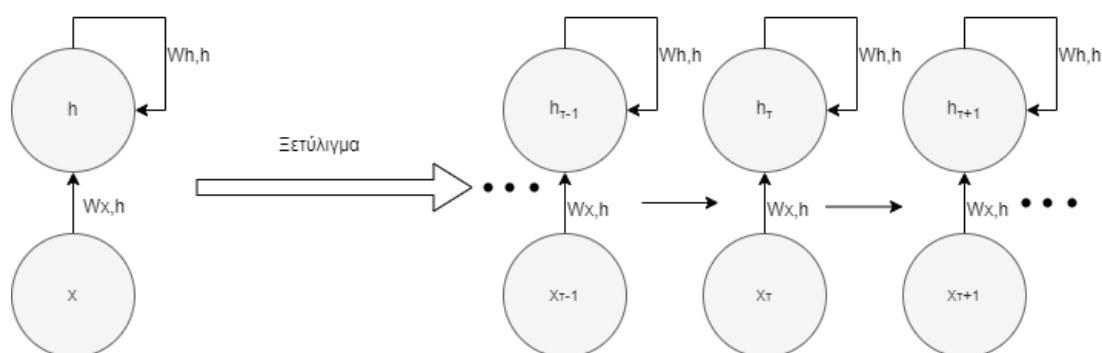
Εικόνα 13: Δομή ενός μοντέλου στοίβαξης.

Προκειμένου να εξασφαλιστεί η αξιοπιστία του ελέγχου, το σύνολο ελέγχου δεν πρέπει να τροφοδοτηθεί στα βασικά μοντέλα και να χρησιμοποιηθεί για την εκπαίδευση του μετά-μοντέλου. Για αυτό, εξαρχής γίνεται ο χωρισμός ενός επιπλέον συνόλου, το οποίο ονομάζεται *σύνολο επικύρωσης* (validation set). Η τεχνική της στοίβαξης συνήθως κατορθώνει να μειώσει την μεροληψία και ευνοεί την παράλληλη ανάπτυξη των μοντέλων [21], [5]. Στην Εικόνα 13 παρουσιάζεται η δομή ενός μοντέλου στοίβαξης.

2.4 Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks, RNN)

Τα αναδρομικά νευρωνικά δίκτυα εντάσσονται στην κατηγορία της επιβλεπόμενης μάθησης και αποτελούν μια κατηγορία νευρωνικών δικτύων που χρησιμοποιούνται για να αποτυπώσουν σχέσεις χρονικής αλληλουχίας και εξάρτησης που ενδέχεται να υπάρχουν μεταξύ των δεδομένων. Διαφοροποιούνται από τα απλά, νευρωνικά δίκτυα πρόσθιας τροφοδότησης, επειδή, σε αντίθεση με αυτά, δέχονται

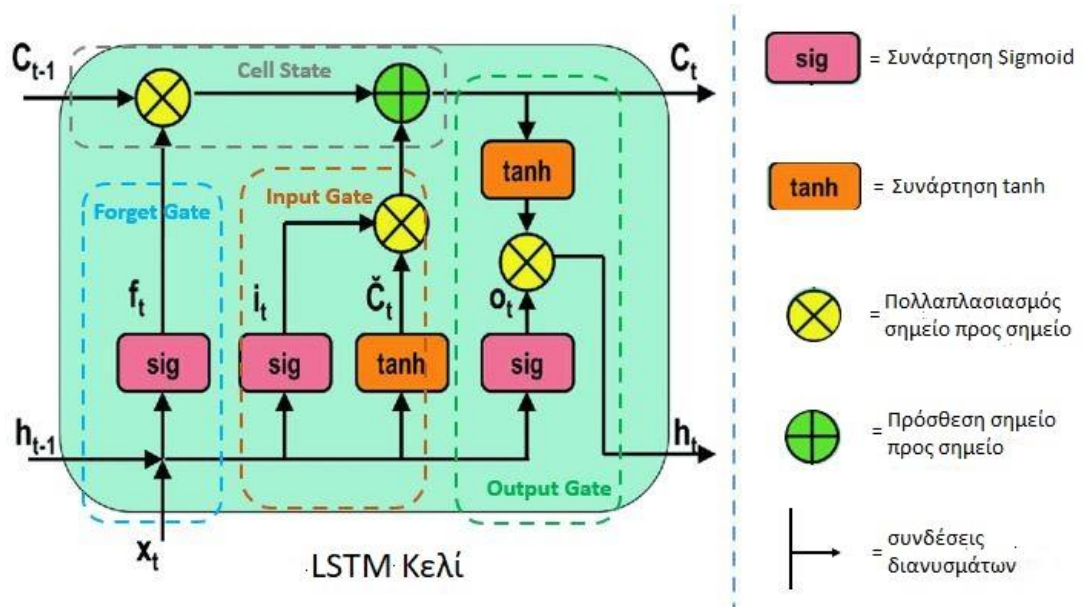
ανατροφοδότηση από την έξοδο τους. Η ιδιότητα τους αυτή, αποτυπώνεται με την παρουσία κύκλων στον γράφο του δικτύου. Στην Εικόνα 14 βλέπουμε την δομή ενός απλού αναδρομικού νευρωνικού δικτύου. Τα αναδρομικά νευρωνικά δίκτυα έχουν αποκτήσει μεγάλη δημοφιλία και βρίσκουν εφαρμογή σε πλήθος πεδίων, όπως η πρόβλεψη χρονοσειράς (time series forecasting) και η επεξεργασία φυσικής γλώσσας (natural language processing). Ωστόσο, τα απλά, αναδρομικά νευρωνικά δίκτυα αντιμετωπίζουν περιορισμούς εξαιτίας του προβλήματος της εξαφανιζόμενης/εκρηγνυώμενης κλίσης (vanishing/exploding gradient) που εμφανίζεται κατά την διαδικασία της οπισθοδιάδοσης και τα εμποδίζει από το να αποτυπώσουν χρονικές συσχετίσεις σε βάθος μεγαλύτερου χρόνου. Προκειμένου να αντιμετωπιστεί το πρόβλημα αυτό, έχουν προταθεί διάφορες αρχιτεκτονικές. Μια από τις πιο δημοφιλείς, που θα αξιοποιηθεί και σε αυτήν την εργασία είναι η αρχιτεκτονική της μακράς βραχυπρόθεσμης μνήμης (long-short term memory - LSTM) [22].



Εικόνα 14: Δομή ενός αναδρομικού δικτύου και το ανάπτυγμα που προκύπτει από το ξετύλιγμά του. (Πηγή: [22])

2.4.1 Συστήματα μακράς βραχυπρόθεσμης μνήμης (LSTM)

Τα συστήματα μακράς βραχυπρόθεσμης μνήμης (LSTM), όπως και άλλα αναδρομικά νευρωνικά δίκτυα, κατορθώνουν να εξαγάγουν τις χρονικές αλληλουχίες και συσχετίσεις που εμφανίζονται μεταξύ των προβλεπόμενων εξόδων. Επιπλέον τα LSTM, σε αντίθεση με άλλα επαναληπτικά μοντέλα, μπορούν να εξαγάγουν χρονικές αλληλουχίες σε πιο απομακρυσμένα χρονικά διαστήματα και να κρατούν μόνο το χρήσιμο μέρος αυτής της πληροφορίας κάθε φορά. Ο τρόπος που το κατορθώνουν αυτό θα αναλυθεί παρακάτω. Στην Εικόνα 15 βλέπουμε την δομή ενός LSTM κυττάρου. Κεντρικό στοιχείο αποτελεί η γραμμή της μακροχρόνιας μνήμης-πληροφορίας η οποία διατρέχει ολόκληρο το μοντέλο και στο σχήμα σημειώνεται ως Cell State. Κατόπιν, σε κάθε LSTM υπάρχουν 3 πύλες (gates), οι οποίες καθορίζουν το ποσοστό και το περιεχόμενο της πληροφορίας που θα μεταφερθεί στο επόμενο κελί.



Εικόνα 15: Δομή ενός LSTM κελιού. (Πηγή: [23]).

Η πρώτη πύλη ονομάζεται *forget gate*. Έχοντας ως είσοδο την βραχυπρόθεσμη μνήμη από τον προηγούμενο κελί (h_{t-1}) και την είσοδο του μοντέλου την χρονική στιγμή t (x_t), αποφασίζει ποιο κομμάτι της μακροπρόθεσμης μνήμης θα ξεχαστεί. Η συνάρτηση που περιγράφει το forget gate είναι η ακόλουθη:

$$f_t = sig(W_f * [h_{t-1}, x_t] + b_f)$$

όπου W_f και b_f είναι τα εκπαιδευσιμα βάρη και πολώσεις μεταξύ της εισόδου και του forget gate, ενώ, η συνάρτηση *sig* (σιγμοειδής) εξάγει ένα αποτέλεσμα μεταξύ 0-1. Για αποτέλεσμα 0, δεν θα περάσει καθόλου από την προγενέστερη πληροφορία, ενώ για αποτέλεσμα 1 η πληροφορία θα περάσει αυτούσια.

Η επόμενη πύλη ονομάζεται *input gate*. Η πύλη αυτή έχει δύο αρμοδιότητες. Αρχικά αποφασίζει ποιο κομμάτι της πληροφορίας θα ανανεωθεί, ενώ αποφασίζει παράλληλα και μια νέα υποψήφια κατάσταση κελιού (cell state). Ύστερα το input gate συνδυάζει τις δύο αυτές αποφάσεις, ώστε να καταλήξει σε μια ανανέωση της κατάστασης κελιού. Αυτό περιγράφεται από τις ακόλουθες εξισώσεις:

$$i_t = sig(W_i * [h_{t-1}, x_t] + b_i)$$

$$\tilde{c}_t = tanh(W_c * [h_{t-1}, x_t] + b_c)$$

Το i_t αποφασίζει ποιο κομμάτι της κατάστασης κελιού θα ανανεωθεί και για αυτό χρησιμοποιεί την σιγμοειδή συνάρτηση. Στην προκειμένη περίπτωση, αποτέλεσμα 0 σημαίνει ότι αυτή η πληροφορία δεν θα ανανεωθεί, ενώ διαφορετικά παίρνουμε ένα ποσοστό για το κατά πόσο η νέα προτεινόμενη κατάσταση κελιού θα επηρεάσει την τελική κατάσταση κελιού. Και πάλι βλέπουμε ότι το i_t λαμβάνει υπόψη του την είσοδο την τρέχουσα χρονική στιγμή και την είσοδο από το προηγούμενο κελί, ενώ

και πάλι υπάρχουν κάποια εκπαιδευσιμα βάρη τα οποία επηρεάζουν τα αποτελέσματα. Επιπλέον, το \tilde{C}_t είναι τη νέα προτεινόμενη κατάσταση κελιού. Η υπερβολική εφαιπτομένη χρησιμοποιείται για να κανονικοποιήσει την κατάσταση κελιού στο διάστημα $[-1,1]$. Και πάλι, η προτεινόμενη κατάσταση κελιού προκύπτει από την προηγούμενη είσοδο, την τρέχουσα είσοδο και κάποια εκπαιδευσιμα βάρη. Αφού προκύψουν τα i_t και \tilde{C}_t εμπλουτίζονται με τις νέες πληροφορίες ξ κατάσταση κελιού (μακροπρόθεσμη μνήμη). Η τελική κατάσταση κελιού (C_t) που προκύπτει περιγράφεται από την εξίσωση:

$$C_t = C_{t-1} * f_t + i_t * \tilde{C}_t$$

Τέλος, έχουμε την 3^η πύλη η οποία ονομάζεται *output gate*. Η πύλη αυτή φροντίζει να ανανεώσει την βραχυπρόθεσμη μνήμη η οποία θα λειτουργήσει ως είσοδος για το επόμενο LSTM κελί. Η βραχυπρόθεσμη μνήμη θα επηρεαστεί τελικά από την προηγούμενη είσοδο, την τρέχουσα είσοδο αλλά και την νέα κατάσταση κελιού. Αυτό παριστάνεται με τις ακόλουθες εξισώσεις:

$$o_t = sig(W_o * [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * tanh(C_t)$$

Κάθε μονάδα LSTM αποτελείται από κελιά LSTM, όπως αυτά που περιγράψαμε παραπάνω, τόσα όσο και το εύρος του χρονικού παραθύρου που εξετάζει το μοντέλο. Η αρχιτεκτονική που παρουσιάστηκε παραπάνω, έχει επιτρέψει στα LSTM μοντέλα να ανιχνεύουν μακροχρόνιες συσχετίσεις και έτσι να διαθέτουν πλήθος εφαρμογών, με πολύ καλά αποτελέσματα σε πολλούς τομείς της μηχανικής μάθησης [24].

2.5 Τεχνικές Προεπεξεργασίας (Preprocessing)

Μέχρι τώρα, μελετήθηκαν μοντέλα επιβλεπόμενης και μη μάθησης τα οποία λαμβάνουν σύνολα δεδομένων (χαρακτηρισμένα ή μη) και εκπαιδούνται ώστε να είναι σε θέση να εξάγουν προβλέψεις ή γενικότερα να παράγουν πληροφορία σχετικά με τα δεδομένα αυτά. Ωστόσο, υπάρχει ένα κρίσιμο στάδιο, που επηρεάζει καθοριστικά την αποτελεσματικότητα της εκπαίδευσης και βρίσκεται πριν την τροφοδότηση των δεδομένων στο μοντέλο. Πρόκειται για το στάδιο της προεπεξεργασίας των δεδομένων.

Τα δεδομένα τα οποία έχει συνήθως στην διάθεση του ο προγραμματιστής του μοντέλου βρίσκονται σε εντελώς ανεπεξέργαστη, πρωταρχική μορφή. Η πρωταρχική μορφή των δεδομένων μπορεί να περιλαμβάνει άγνωστες ή περιττές τιμές, επαναλήψεις και να μαστίζεται από πλήθος άλλων προβλημάτων και δυσμορφιών οι οποίες δυσχεραίνουν την εκπαίδευση του μοντέλου που πρόκειται να χρησιμοποιηθεί. Η ευαισθησία στην έλλειψη προεπεξεργασίας διαφέρει από

μοντέλο σε μοντέλο [25]. Ωστόσο, στην γενική περίπτωση φαίνεται ότι, η προεπεξεργασία των δεδομένων εισόδου ενός μοντέλου βελτιώνει την ποιότητα των αποτελεσμάτων του, μειώνοντας ταυτόχρονα τον χρόνο και την πολυπλοκότητα της εκπαίδευσης του και διευκολύνει την ικανότητα ερμηνευσιμότητας του μοντέλου, δηλαδή την ικανότητα του ανθρώπου να κατανοήσει πως προέκυψαν τα αποτελέσματα της εκπαίδευσης. Η βελτίωση της ποιότητας των αποτελεσμάτων της εκπαίδευσης έχει να κάνει και με την αποφυγή της υπερπροσαρμογής, η οποία μπορεί να οφείλεται στην ύπαρξη χαμηλής ποιότητας δειγμάτων-πληροφοριών στο σύνολο εκπαίδευσης, τα οποία δυσχεραίνουν την προσπάθεια του μοντέλου να γενικεύσει τα αποτελέσματά του [26].

Υπάρχει πληθώρα τεχνικών προεπεξεργασίας που μπορούν να χρησιμοποιηθούν προκειμένου να βελτιώσουν την ποιότητα των δεδομένων που θα τροφοδοτηθούν τελικά στο μοντέλο. Οι τεχνικές αυτές διακρίνονται σε τεχνικές πρόσθεσης στοιχείων, αφαίρεσης στοιχείων ή μετασχηματισμού των υπάρχοντων δεδομένων [25]. Παρακάτω, θα παρουσιαστούν συνοπτικά διάφορες τεχνικές προεπεξεργασίας και θα συζητηθεί το τι μπορεί να επιτευχθεί με την χρήση αυτών. Να σημειωθεί τέλος, ότι η βέλτιστη τεχνική ή ο συνδυασμός τεχνικών προεπεξεργασίας εξαρτάται τόσο από τα δεδομένα εκπαίδευσης όσο και από το μοντέλο εκπαίδευσης, ενώ σε ορισμένες περιπτώσεις η προεπεξεργασία είναι περιττή ή και επιβλαβής για την διαδικασία της μάθησης [25].

2.5.1 Διαχείριση απουσιαζουσών τιμών (missing values)

Είναι συχνό φαινόμενο στα δεδομένα μας να υπάρχουν δείγματα των οποίων ορισμένα χαρακτηριστικά είναι άγνωστα ή ανύπαρκτα. Αυτές οι απύουσες τιμές των χαρακτηριστικών μπορεί να δημιουργήσουν προβλήματα στην εκπαίδευση του μοντέλου και να επηρεάσουν την ποιότητα των τελικών αποτελεσμάτων. Βεβαίως, υπάρχουν και μοντέλα που μπορούν να διαχειριστούν τις άγνωστες τιμές [25]. Ωστόσο, στην γενική περίπτωση η παρουσία των απουσιαζουσών τιμών αντιμετωπίζεται είτε με την διαγραφή τους είτε με την τεχνική της *αντικατάστασης*. Κατά την διαγραφή, αφαιρούνται απλώς από το σύνολο δεδομένων τα δείγματα εκείνα τα οποία περιέχουν άγνωστες τιμές χαρακτηριστικών. Η διαδικασία αυτή μπορεί να δημιουργήσει πρόβλημα στην περίπτωση που το σύνολο δεδομένων δεν είναι μεγάλο ή υπάρχουν πολλές άγνωστες τιμές. Οπότε εναλλακτικά, μπορεί να χρησιμοποιηθεί η τεχνική της αντικατάστασης των απουσιαζουσών τιμών, κατά την οποία, η άγνωστη τιμή αντικαθίσταται από μια άλλη τιμή η οποία προκύπτει από τις τιμές των υπόλοιπων δειγμάτων για το ίδιο χαρακτηριστικό. Για παράδειγμα, η απουσιάζουσα τιμή θα μπορούσε να προκύψει ως ο μέσος όρος των γνωστών τιμών των υπόλοιπων δειγμάτων για το συγκεκριμένο χαρακτηριστικό, ή να προκύψει από ορισμένα μόνο «κοντινά» δείγματα ή ακόμα και να αποδοθεί τυχαία [25], [27].

2.5.2 Διαχείριση των έκτοπων τιμών (outliers)

Ως έκτοπες ορίζονται οι τιμές οι οποίες απέχουν σημαντικά από την πλειοψηφία των υπόλοιπων τιμών του ίδιου χαρακτηριστικού για το σύνολο δεδομένων [25]. Οι τιμές αυτές μπορούν να επηρεάσουν σημαντικά τις μετρικές της απόδοσης του μοντέλου που εξετάζουμε αλλά και να είναι επιζήμιες και για την εκπαίδευση γενικότερα [27]. Οι έκτοπες τιμές προκύπτουν συχνά από σφάλματα ή από ακραίες συνθήκες οι οποίες μπορεί να μην αφορούν την περίπτωση που εξετάζει η εκπαίδευση και επομένως, μπορούν να έχουν παρόμοια αντιμετώπιση με τις απουσιάζουσες τιμές. Ωστόσο, η αντικατάσταση ή διαγραφή τους δεν μπορεί να γίνει εύκολα γιατί συχνά είναι δείγματα κάποιων καταστάσεων, οι οποίες έχουν ενδιαφέρον να μελετηθούν [25]. Οι έκτοπες τιμές μπορούν να αντιμετωπιστούν αποτελεσματικά με κάποιες από τις τεχνικές μετασχηματισμού που θα δούμε αργότερα (χρήση «κουβάδων»), ενώ άλλες κάνουν την παρουσία των ακραίων τιμών ακόμα πιο ισχυρή (κανονικοποίηση ελαχίστου-μεγίστου).

2.5.3 Μετασχηματισμός των δεδομένων

Άλλη μια συχνή κατηγορία τεχνικών προεπεξεργασίας που εφαρμόζεται, πέρα από την πρόσθεση και την αφαίρεση στοιχείων, είναι και ο μετασχηματισμός τους. Μετασχηματισμός των δεδομένων μπορεί να συμβεί με διάφορους τρόπους και να έχει πολύ διαφορετικούς σκοπούς κατά περίπτωση. Κάποιοι από τους πιο ευρέως χρησιμοποιούμενους μετασχηματισμούς είναι η μείωση της διαστατικότητας του συνόλου δεδομένων και η κανονικοποίηση του. Η μείωση της διαστατικότητας έχει να κάνει με τον μετασχηματισμό του συνόλου δεδομένων έτσι ώστε, το νέο σύνολο δεδομένων που θα προκύψει να έχει δείγματα με μικρότερο αριθμό διαστάσεων-χαρακτηριστικών σε σχέση με το αρχικό. Η μείωση της διαστατικότητας μπορεί να είναι καθοριστική για την εξοικονόμηση χρόνου και πόρων κατά την εκπαίδευση του μοντέλου. Σημαντικός είναι και ο ρόλος της κανονικοποίησης των τιμών του συνόλου δεδομένων για την ομαλή λειτουργία του μοντέλου. Η κανονικοποίηση μάλιστα, είναι μια απαραίτητη διαδικασία στην περίπτωση που τα χαρακτηριστικά του συνόλου δεδομένων έχουν διαφορετικό εύρος τιμών. Σε αυτήν την περίπτωση, χαρακτηριστικά με μεγάλο εύρος τιμών μπορεί να επηρεάζουν δυσανάλογα, σε σχέση με την πραγματική τους σημασία, τους υπολογισμούς του μοντέλου, έχοντας αρνητικό αντίκτυπο στην απόδοση και στην δυνατότητα γενίκευσης του μοντέλου μετά την εκπαίδευση. Έτσι, κρίνεται απαραίτητη η κανονικοποίηση των τιμών των χαρακτηριστικών στην ίδια κλίμακα, συνήθως μεταξύ $[0,1]$ [27]. Κάποιες από τις πιο γνωστές τεχνικές κανονικοποίησης είναι η κανονικοποίηση ελαχίστου-μεγίστου και η κανονικοποίηση γύρω από τη μέση τιμή.

Άλλες γνωστές τεχνικές μετασχηματισμού του συνόλου δεδομένων είναι η χρήση «κουβάδων» και η ομαδοποίηση των δειγμάτων εισόδου [27]. Κατά την χρήση

«κουβάδων», τα δεδομένα μας χωρίζονται σε «κουβάδες», όπου ο καθένας διαθέτει μια συγκεκριμένη ετικέτα που ανταποκρίνεται σε ορισμένα χαρακτηριστικά των δειγμάτων που περιέχει. Στην συνέχεια, η τιμή των χαρακτηριστικών αυτών για κάθε δείγμα, αντικαθίσταται από την ετικέτα του «κουβά» στον οποίον ανήκει το δείγμα. Ένα χαρακτηριστικό παράδειγμα αυτής της τεχνικής είναι, για ένα σύνολο δεδομένων του οποίου τα δείγματα αντιστοιχούν σε ανθρώπους, τα άτομα να τοποθετούνται σε «κουβάδες» 0-18, 19-30, 31-50, 51+ βάση της ηλικίας τους. Έτσι, η χρήση «κουβάδων» μειώνει την επίδραση των ακραίων τιμών και απλοποιεί τους υπολογισμούς του μοντέλου, ενώ βελτιώνει την απόδοση του και το καθιστά καλύτερα ερμηνεύσιμο [25] , [27].

Παρόμοια τεχνική με την χρήση «κουβάδων» είναι και η ομαδοποίηση των δειγμάτων σε ομοειδή, η οποία μπορεί να χρησιμοποιήσει μη-επιβλεπόμενη μάθηση για να επιτύχει την ομαδοποίηση, ενώ κατά την χρήση «κουβάδων» η ομαδοποίηση των δειγμάτων γίνεται, συνήθως, βάσει του εύρους τιμών κάποιου χαρακτηριστικού. Πετυχαίνει και αυτή καλύτερη απόδοση, υψηλότερη ερμηνευσιμότητα και απλούστευση της εκπαίδευσης [27]. Στην συνέχεια, θα αναλυθεί ελαφρώς διεξοδικότερα η κανονικοποίηση ελαχίστου-μεγίστου η οποία χρησιμοποιήθηκε εκτεταμένα στην παρούσα εργασία.

2.5.3.1 Κανονικοποίηση Ελαχίστου-Μεγίστου (Min-Max Scaling)

Η κανονικοποίηση ελαχίστου-μεγίστου αποτελεί μια από τις πλέον συνηθισμένες τεχνικές κανονικοποίησης. Ονομάζεται έτσι επειδή χρησιμοποιεί την μέγιστη και την ελάχιστη τιμή ενός χαρακτηριστικού προκειμένου να πετύχει την κανονικοποίηση του στην κλίμακα [0,1]. Ο τύπος που χρησιμοποιείται για την κανονικοποίηση ελαχίστου-μεγίστου είναι ο ακόλουθος:

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

όπου, X_{sc} αντιστοιχεί στην κανονικοποιημένη τιμή του χαρακτηριστικού-δείγματος, το X στην αρχική τιμή, το X_{max} στην μέγιστη τιμή των δειγμάτων και το X_{min} στην ελάχιστη [27].

2.6 Μετρικές της Επιβλεπόμενης Μάθησης

Κατά την επιβλεπόμενη μάθηση, η επιτυχία της εκπαίδευσης του μοντέλου αξιολογείται με βάση την απόσταση μεταξύ των προβλέψεων του μοντέλου για τις εξόδους του συνόλου ελέγχου και των πραγματικών εξόδων αυτού. Προκειμένου να αποτυπωθεί αυτή η απόσταση χρησιμοποιούνται πολλές μετρικές, οι οποίες διαφοροποιούνται ως προς τα προβλήματα των οποίων την αξιολόγηση εξυπηρετούν και ως προς την συνιστώσα της εκπαίδευσης που τελικά απεικονίζουν.

Συγκεκριμένα, τα προβλήματα κατηγοριοποίησης χρειάζονται άλλες μετρικές για την αξιολόγηση τους σε σχέση με τα προβλήματα παλινδρόμησης, δεδομένου ότι στην πρώτη περίπτωση η έξοδος του προβλήματος είναι διακριτή (κάποια κλάση) και στην δεύτερη συνεχής. Στην συνέχεια, θα συζητηθούν κάποιες ευρέως διαδεδομένες μετρικές για προβλήματα παλινδρόμησης.

2.6.1 Μετρικές για προβλήματα παλινδρόμησης

Η φιλοσοφία των μετρικών για την αξιολόγηση των προβλημάτων παλινδρόμησης είναι η μέτρηση του σφάλματος ανάμεσα στην προβλεπόμενη και την κανονική τιμή για κάθε δείγμα του συνόλου ελέγχου. Υπάρχουν διάφορες μαθηματικά ορισμένες μετρικές οι οποίες μπορούν να χρησιμοποιηθούν προκειμένου να αποτιμηθεί το σφάλμα αυτό. Παρακάτω θα συζητηθούν τέσσερις μόνο από αυτές, οι οποίες αξιοποιήθηκαν στα πλαίσια της παρούσας εργασίας :

➤ Μέσο Τετραγωνικό Σφάλμα – ΜΤΣ (Mean Square Error) :

Το ΜΤΣ είναι ένας τρόπος να μετρήσουμε την απόκλιση ανάμεσα στις προβλεπόμενες τιμές του μοντέλου που εκπαιδεύτηκε και τις πραγματικές τιμές του συνόλου ελέγχου. Συγκεκριμένα, για κάθε δείγμα του συνόλου ελέγχου υπολογίζεται η διαφορά μεταξύ της προβλεπόμενης και της πραγματικής τιμής εξόδου, υψώνεται στο τετράγωνο και το αποτέλεσμα αυτό συμψηφίζεται με τα αποτελέσματα για όλα τα υπόλοιπα δείγματα του συνόλου ελέγχου, ώστε να προκύψει ο μέσος όρος τους. Ο ορισμός του ΜΤΣ είναι ο ακόλουθος [28]:

$$MT\sigma = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

όπου, n είναι ο αριθμός των δειγμάτων του συνόλου ελέγχου, Y_i η πραγματική τιμή του i -οστού δείγματος και \hat{Y}_i η προβλεπόμενη τιμή.

➤ Μέσο Απόλυτο Σφάλμα – ΜΑΣ (Mean Absolute Error) :

Το ΜΑΣ είναι πολύ παρόμοια μετρική με το ΜΤΣ, όμως αντί να μετράει το τετράγωνο της διαφοράς ανάμεσα στην πραγματική και προβλεπόμενη τιμή εξόδου του δείγματος, μετράει την απόλυτη τιμή αυτής της διαφοράς. Είναι προφανές ότι, για αυτόν τον λόγο, το ΜΤΣ τιμωρεί περισσότερο μεγάλα σφάλματα σε σχέση με το ΜΑΣ. Ο ορισμός του ΜΑΣ είναι ο ακόλουθος [28]:

$$MA\sigma = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

- Μέσο Απόλυτο Ποσοστιαίο Σφάλμα – ΜΑΠΣ (Mean Absolute Percentage Error) :

Το ΜΑΠΣ με την σειρά του είναι πολύ παρόμοιο με το ΜΑΣ, όμως διαιρεί την απόλυτη τιμή που υπολογίζεται για κάθε δείγμα με την μέγιστη πραγματική τιμή εξόδου του συνόλου δεδομένων. Αυτό γίνεται προκειμένου το ΜΑΠΣ να απεικονίζει το σχετικό σφάλμα χωρίς να επηρεάζεται από την κλίμακα στην οποία βρίσκονται οι έξοδοι του συνόλου ελέγχου. Ο ορισμός του ΜΑΠΣ είναι ο ακόλουθος [28] :

$$ΜΑΠΣ = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i - \hat{Y}_i|}{\max(\varepsilon, |Y_i|)}$$

όπου το $\varepsilon > 0$ είναι ένας πολύ μικρός θετικός αριθμός που χρησιμοποιείται για την πληρότητα του μαθηματικού ορισμού στην περίπτωση που $Y_i = 0$.

- Goodness of fit:

Η μετρική goodness of fit χρησιμοποιείται για να ελέγξει την ομοιότητα μεταξύ δύο καμπυλών, στην περίπτωση των προβλημάτων παλινδρόμησης, της καμπύλης που ορίζεται από τις πραγματικές εξόδους του συνόλου δεδομένων και της καμπύλης που ορίζεται από τις προβλεπόμενες τιμές του μοντέλου. Ισχύει ότι:

$$goodness\ of\ fit = 1 - \frac{|\overrightarrow{x_{ref}} - \vec{x}|}{|\overrightarrow{x_{ref}} - mean(\overrightarrow{x_{ref}})|}$$

όπου, $\overrightarrow{x_{ref}}$ είναι το διάνυσμα των πραγματικών τιμών, $mean(\overrightarrow{x_{ref}})$ η μέση τους τιμή και \vec{x} οι προβλεπόμενες τιμές. Για $goodness\ of\ fit = 1$ έχουμε ακριβή ταύτιση των δύο καμπυλών, για $goodness\ of\ fit = 0$ δεν έχουμε καλύτερη ταύτιση από ότι με μια ευθεία γραμμή και για πολύ μεγάλες αρνητικές τιμές έχουμε πολύ χαμηλή ομοιότητα των δύο καμπυλών [29].

Η μετρική που θα μας απασχολήσει κυρίως σε αυτήν την εργασία είναι το ΜΤΣ, που αποτελεί την πλέον συνηθισμένη μετρική για προβλήματα παλινδρόμησης, εξαιτίας του εύκολου υπολογισμού του και της άμεσης συσχέτισης του με τον μαθηματικό ορισμό της έννοιας της απόστασης.

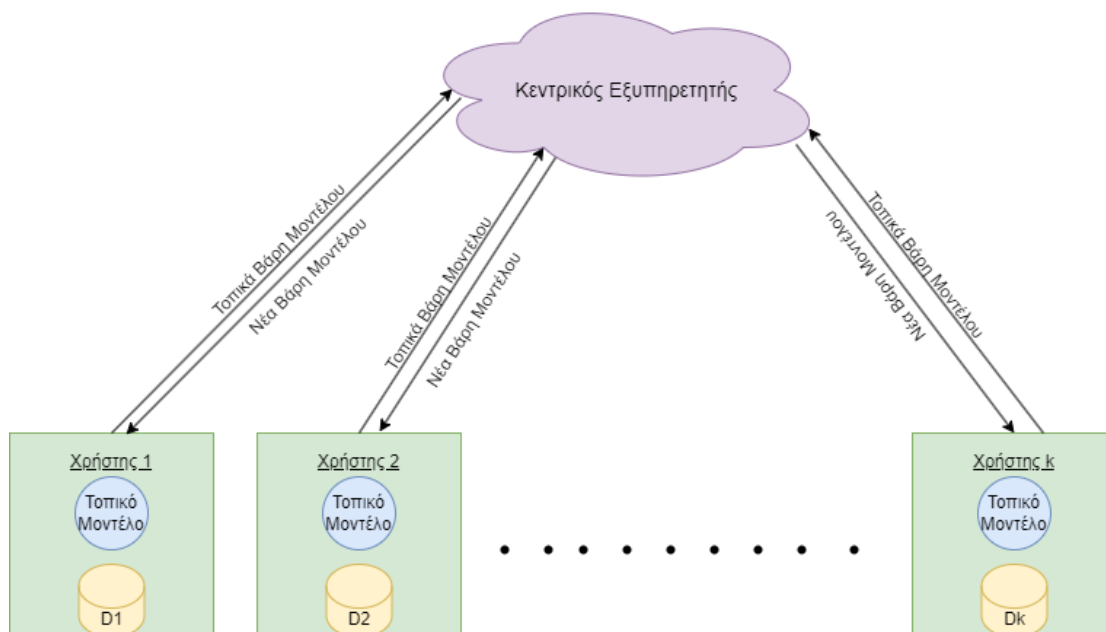
2.7 Ομοσπονδιακή Μάθηση (Federated Learning)

Ο όρος Ομοσπονδιακή Μάθηση (Federated Learning – FL) προτάθηκε πρώτη φορά το 2016 για να περιγράψει μοντέλα τα οποία ακολουθούν μια κατανεμημένη προσέγγιση στη μηχανική μάθηση, με έμφαση στην προστασία της ιδιωτικότητας των δεδομένων του χρήστη [30], [31], [1]. Η τεράστια κινητοποίηση για την προστασία των προσωπικών δεδομένων χρηστών, που προέκυψε μετά το σκάνδαλο της Cambridge Analytica και η συνεπακόλουθη θέσπιση αυστηρού νομοθετικού πλαισίου γύρω από το ζήτημα της προστασίας προσωπικών δεδομένων, με νόμους

όπως ο GDPR (General Data Protection Regulations) της Ευρωπαϊκής Ένωσης και οι νόμοι προστασίας των πολιτών της Κίνας, δημιουργούν την ανάγκη για την εύρεση νέων προσεγγίσεων στην μηχανική μάθηση, οι οποίες θα συμμορφώνονται με το πλαίσιο των κανονισμών που έχουν τεθεί [30], [31].

Στο πεδίο της μηχανικής μάθησης, έχει τεράστια σημασία η παρουσία μεγάλου όγκου δεδομένων προκειμένου να προκύψουν ποιοτικά και ακριβή αποτελέσματα. Ταυτόχρονα, την παρούσα χρονική στιγμή, πράγματι ο αριθμός των δεδομένων που διατίθενται είναι τεράστιος. Από την άλλη, η χρήση αυτών των δεδομένων για τους σκοπούς της μηχανικής μάθησης αντιμετωπίζει περιορισμούς οι οποίοι μεταξύ άλλων σχετίζονται με την κατανομή των δεδομένων αυτών σε νησίδες (data islands) [31]. Συγκεκριμένα, ως νησίδες δεδομένων μπορούν να νοηθούν αισθητήρες, προσωπικές συσκευές χρηστών (κινητά τηλέφωνα, υπολογιστές), ενώ με την εξέλιξη του Διαδικτύου των Πραγμάτων (Internet of Things) το φαινόμενο γίνεται εντονότερο. Μία προφανής λύση, ώστε ο τεράστιος αυτός όγκος δεδομένων να αξιοποιηθεί για την εκπαίδευση μοντέλων μηχανικής μάθησης και την παραγωγή πολύτιμης γνώσης, είναι να γίνει μια συγκεντρωτική διαχείριση των δεδομένων αυτών. Η λύση αυτή έχει ήδη χρησιμοποιηθεί ευρέως και προβλέπει την συλλογή των δεδομένων σε ένα κεντρικό εξυπηρετητή, την οργάνωση, επεξεργασία και αξιοποίηση τους [31]. Ωστόσο, οι νέοι αυστηροί κανονισμοί προστασίας προσωπικών δεδομένων θέτουν ισχυρούς φραγμούς στην συλλογή όγκου ευαίσθητων προσωπικών δεδομένων σε κάποιον κεντρικό εξυπηρετητή. Παράλληλα, η εφαρμογή τεχνικών μηχανικής μάθησης σε τοπικό επίπεδο (σε επίπεδο συσκευής χρήστη) δεν μπορεί να είναι αποτελεσματική εξαιτίας του μικρού πλήθους των δεδομένων που διατίθενται προκειμένου να γίνει η εκπαίδευση [31]. Η Ομοσπονδιακή Μάθηση έχει προταθεί ως λύση στο πρόβλημα αυτό.

Η φιλοσοφία της Ομοσπονδιακής Μάθησης ταιριάζει απόλυτα με την δομή και τις ιδιαιτερότητες της κατανομής των δεδομένων αυτή την στιγμή. Ο τρόπος λειτουργίας της έχει ως εξής: έστω ότι υπάρχουν k χρήστες οι οποίοι επιθυμούν να συμμετάσχουν στην διαδικασία της μάθησης. Καθένας από αυτούς έχει πρόσβαση μόνο στο δικό του σύνολο δεδομένων έστω D_i , όπου i ο εκάστοτε χρήστης. Ταυτόχρονα, υπάρχει και ένας κεντρικός εξυπηρετητής, ο οποίος αποφασίζει για το μοντέλο εκπαίδευσης που θα χρησιμοποιηθεί από τους επιμέρους χρήστες και «συντονίζει» την εκπαίδευση. Σύμφωνα με τα μοντέλα ομοσπονδιακής μάθησης, ο εκάστοτε χρήστης θα χρησιμοποιήσει το μοντέλο μηχανικής μάθησης, το οποίο έχει οριστεί από τον κεντρικό εξυπηρετητή, θα το εκπαιδεύσει τοπικά, πάνω στο δικό του σύνολο δεδομένων D_i , και θα αποστείλει της παραμέτρους του μοντέλου (και μόνο αυτές) που θα προκύψουν από την εκπαίδευση στον κεντρικό εξυπηρετητή. Ο κεντρικός εξυπηρετητής θα συλλέξει τις παραμέτρους που έλαβε από τους χρήστες, θα τις *συμψηφίσει* (aggregation) βάση κάποιας μαθηματικής φόρμουλας και θα αποστείλει στους χρήστες τις νέες παραμέτρους, ώστε αυτοί να ενημερώσουν τα τοπικά μοντέλα. Η διαδικασία επαναλαμβάνεται για όσους γύρους χρειαστεί, μέχρι να υπάρξει σύγκλιση [31].



Εικόνα 16: Γενική δομή ενός μοντέλου Ομοσπονδιακής Μάθησης.

Η Ομοσπονδιακή Μάθηση, με τον τρόπο που περιγράφηκε παραπάνω, επιτρέπει στους χρήστες να συνεργαστούν ώστε να μεγιστοποιήσουν τα οφέλη τους από την εφαρμογή των τεχνικών μηχανικής μάθησης, ενώ παράλληλα προστατεύουν τα ευαίσθητα προσωπικά τους δεδομένα. Επομένως τα κύρια πλεονεκτήματα της Ομοσπονδιακής Μάθησης έναντι των διάφορων μοντέλων κεντρικής μάθησης που είχαν χρησιμοποιηθεί μέχρι τώρα είναι δύο [31] :

1. Η ιδιωτικότητα. Αυτή επιτυγχάνεται καθώς ο κεντρικός εξυπηρετητής δεν έχει πρόσβαση απευθείας στα δεδομένα των χρηστών, αλλά μόνο στις παραμέτρους-βάρη εκπαίδευσης που μοιράζονται οι χρήστες. Πρόσβαση στα δεδομένα του έχει μόνο ο εκάστοτε χρήστης
2. Μείωση του υπολογιστικού κόστους και των πόρων της κεντρικής εκπαίδευσης. Αυτό επιτυγχάνεται καθώς η εκτέλεση των υπολογισμών διαμοιράζεται στους τοπικούς χρήστες, πράγμα που είναι πλέον εφικτό εξαιτίας της αύξησης της υπολογιστικής ισχύος των συσκευών των χρηστών.

Παράλληλα όμως, η Ομοσπονδιακή Μάθηση έχει να αντιμετωπίσει και κάποιες σημαντικές προκλήσεις [32], [31]:

1. Απαιτεί δυνατότητα γρήγορης επικοινωνίας σε πραγματικό χρόνο προκειμένου να μπορεί να λειτουργήσει αποτελεσματικά. Η χαμηλή ταχύτητα επικοινωνίας ωστόσο μπορεί να οφείλεται όχι μόνο στις δυνατότητες της σύνδεσης αλλά και του χρήστη. Κάποιες λύσεις που έχουν προταθεί για αυτό το πρόβλημα είναι η δυνατότητα ασύγχρονης επικοινωνίας ανάμεσα στον κεντρικό εξυπηρετητή και τους τοπικούς χρήστες. Επίσης, μια άλλη πρόταση είναι ότι οι αλγόριθμοι ομοσπονδιακής μάθησης δεν απαιτούν την συμμετοχή όλων των τοπικών χρηστών κάθε φορά για να λειτουργήσουν. Επομένως, θα μπορούσε ο κεντρικός

εξυπηρετητής να επιλέγει δειγματοληπτικά τους χρήστες με τους οποίους θα συνεργαστεί σε κάθε γύρο, εξασφαλίζοντας ωστόσο ότι θα συμμετάσχουν σε κάποιο ικανοποιητικό βαθμό όλοι οι χρήστες. Τέλος, αντίστοιχα, τα μοντέλα ομοσπονδιακής μάθησης, οφείλουν να δείχνουν ανοχή στα λάθη και αδυναμίες που είναι πολύ πιθανό να προκύψουν από τα τοπικά μοντέλα, σε τέτοιο βαθμό όμως που να μην επηρεάζεται σημαντικά η ποιότητα και η αξιοπιστία της εκπαίδευσης.

2. Παραμένει ο κίνδυνος διαρροής δεδομένων. Συγκεκριμένα, έρευνες έχουν δείξει ότι τα δεδομένα των χρηστών θα μπορούσαν να ανακτηθούν από τα βάρη-κλίσεις που ανταλλάσσονται ανάμεσα σε κεντρικό και τοπικό επίπεδο κατά τη διάρκεια της εκπαίδευσης. Μια λύση που συχνά εφαρμόζεται είναι να πραγματοποιείται *ομομορφική κρυπτογράφηση* (homomorphic encryption) στις παραμέτρους που ανταλλάσσονται ή το πρωτόκολλο της *διαφορικής ιδιωτικότητας* (differential privacy), το οποίο προβλέπει την προσθήκη θορύβου κατά την επικοινωνία των επιμέρους επιπέδων.
3. Η αποτελεσματικότητα και η αξιοπιστία της εκπαίδευσης εξαρτάται από την αξιοπιστία των χρηστών τοπικού επιπέδου. Υπάρχει ο κίνδυνος οι πληροφορίες που αποστέλλονται από τοπικούς χρήστες να είναι εσκεμμένα ή μη, χαμηλής ποιότητας ή παραπλανητικές. Για αυτό, έχουν προταθεί πολλοί αλγόριθμοι στους οποίους ο κεντρικός εξυπηρετητής επιλέγει βάση κάποιων κριτηρίων αξιοπιστίας και ποιότητας τους χρήστες με τους οποίους θα συνεργαστεί.
4. Υπάρχει σοβαρό πρόβλημα στην διαχείριση δεδομένων που δεν ικανοποιούν τις συνθήκες iid (ανεξαρτησίας και ομοιόμορφης κατανομή μεταξύ τους). Τα δεδομένα των επιμέρους χρηστών μπορεί να μην είναι ανεξάρτητα μεταξύ τους ή να μην έχουν κοινά χαρακτηριστικά (είτε εκ των πραγμάτων είτε λόγω διαφορετικής μορφοποίησης) ή ακόμα να είναι άνισα κατανεμημένα (χρήστες με πολύ περισσότερα χαρακτηριστικά από άλλους). Το γεγονός αυτό δυσχεραίνει σοβαρά την εκπαίδευση και υπονομεύει την ποιότητα των αποτελεσμάτων της. Να σημειωθεί εδώ ωστόσο, ότι στον τομέα της ναυτιλίας το πρόβλημα αυτό δεν είναι τόσο έντονο καθώς τα μοντέλα των αισθητήρων που χρησιμοποιούνται, είναι ως επί το πλείστον κοινά και ακολουθούν παρόμοια κωδικοποίηση [1].
5. Υπάρχει θέμα με την ερμηνευσιμότητα των αποτελεσμάτων που προκύπτουν από την εκπαίδευση με την χρήση ομοσπονδιακής μάθησης και πολλές έρευνες επικεντρώνονται στην κατασκευή ερμηνεύσιμων αλγορίθμων ομοσπονδιακής μάθησης.

Από τα παραπάνω, προκύπτει το συμπέρασμα ότι η Ομοσπονδιακή Μάθηση αποτελεί μια πολλά υποσχόμενη τεχνική, η οποία προτάσσει κατανεμημένες προσεγγίσεις προκειμένου να συμμορφωθεί με την τήρηση των κανονισμών για την προστασία των προσωπικών δεδομένων και να αντιμετωπίσει άλλες προκλήσεις. Παράλληλα, υπάρχει ένα ευρύ πεδίο έρευνας που αναζητά λύσεις στις αδυναμίες

της Ομοσπονδιακής Μάθησης, το οποίο αναμένεται να επεκταθεί καθώς υπάρχει πλήθος ανοικτών ζητημάτων.

Η Ομοσπονδιακή Μάθηση ανάλογα με τα μοτίβα κατανομής του χώρου δειγμάτων και του χώρου χαρακτηριστικών μπορεί να χωριστεί σε τρεις κατηγορίες, οι οποίες έχουν και διαφορετικές αλγοριθμικές προσεγγίσεις [31], [30]:

- Οριζόντια Ομοσπονδιακή Μάθηση:
Σε αυτήν την περίπτωση, οι τοπικοί χρήστες έχουν μικρή επικάλυψη μεταξύ τους, ωστόσο τα επιμέρους χαρακτηριστικά τους επικαλύπτονται σε μεγάλο βαθμό. Δηλαδή, έχουμε διαφορετικούς χρήστες με ανεξάρτητα, δικά τους σύνολα δεδομένων ο καθένας, αλλά με κοινά χαρακτηριστικά για τα δείγματα των συνόλων δεδομένων τους. Πρόκειται για την κλασικότερη δομή ομοσπονδιακής μάθησης και επιτρέπει την επέκταση του συνόλου δεδομένων που θα χρησιμοποιήσει το κεντρικό μοντέλο, εφόσον τα δείγματα χρηστών με κοινά χαρακτηριστικά μπορούν να ιδωθούν ως ένα ενιαίο σύνολο δεδομένων.
- Κατακόρυφη Ομοσπονδιακή Μάθηση:
Στην κατακόρυφη Ομοσπονδιακή Μάθηση, οι χρήστες έχουν μεγάλο βαθμό επικάλυψης, ενώ τα χαρακτηριστικά τους μικρό. Αυτή μορφολογία του χώρου δειγμάτων και δεδομένων μπορεί να αξιοποιηθεί για να αυξηθεί η διαστατικότητα του χώρου χαρακτηριστικών των κοινών χρηστών και να εμπλουτιστούν τα σύνολα δεδομένων τους με νέες πληροφορίες. Ένα παράδειγμα κατακόρυφης ομοσπονδιακής μάθησης είναι να υπάρχουν, για τον ίδιο χρήστη, ένα σύνολο δεδομένων με ταινίες που του αρέσουν και ένα άλλο με βιβλία και αυτά να χρησιμοποιηθούν για να γίνονται προτάσεις μουσικής στον χρήστη.
- Μεταφορά Ομοσπονδιακής Μάθησης:
Στην περίπτωση αυτή δεν υπάρχει επικάλυψη ούτε μεταξύ των χρηστών ούτε μεταξύ των χαρακτηριστικών των δειγμάτων τους. Οπότε, αυτό το σχήμα θα μπορούσε να αξιοποιηθεί στα πλαίσια της μεταφοράς μάθησης. Η μεταφορά μάθησης χρησιμοποιείται στην περίπτωση που ο χρήστης δεν διαθέτει επαρκή ποσότητα δεδομένων για την εκπαίδευση. Έτσι, κάποιος άλλος χρήστης με πλούσιο όγκο δεδομένων μπορεί να μοιραστεί τα μοντέλα που έχει αναπτύξει για την επίλυση παρόμοιων προβλημάτων με τον χρήστη που δεν διαθέτει επαρκή δεδομένα.

Στην παρούσα εργασία, χρησιμοποιήθηκε Ομοσπονδιακή Μάθηση για την ανταλλαγή γνώσης στα πλαίσια του τομέα της ναυτιλίας. Ακολουθήθηκε μια οριζόντια προσέγγιση ομοσπονδιακής μάθησης και χρησιμοποιήθηκε η απλούστερη εκδοχή του αλγορίθμου συμψηφισμού μέσου όρου (Federated Averaging – FedAvg), του οποίου τα βήματα θα περιγραφούν στην συνέχεια. Να τονιστεί, επίσης, ότι υπάρχουν πολλοί ακόμα αλγόριθμοι που ακολουθούν την προσέγγιση της ομοσπονδιακής μάθησης, ενσωματώνουν βελτιώσεις που αναφέρθηκαν παραπάνω, ενώ συνεχώς σχεδιάζονται νέοι.

2.7.1 Αλγόριθμος Συμψηφισμού Μέσου Όρου (FedAvg)

Ο αλγόριθμος FedAvg ακολουθεί την γενική δομή του οριζόντιου μοντέλου ομοσπονδιακής μάθησης, με την ιδιαιτερότητα ότι ο συμψηφισμός των τοπικών παραμέτρων στον κεντρικό εξυπηρετητή γίνεται χρησιμοποιώντας μια συνάρτηση μέσου όρου. Η δομή του μοντέλου είναι η εξής: υπάρχουν k τοπικοί χρήστες με το δικό τους σύνολο δεδομένων, έστω D_i , και το δικό τους τοπικό μοντέλο και ένας κεντρικός εξυπηρετητής, ο οποίος είναι υπεύθυνος για τα κεντρικά βάρη-παραμέτρους. Τα βήματα του αλγορίθμου παρουσιάζονται παρακάτω [33]:

1. Ο κεντρικός εξυπηρετητής αρχικοποιεί τα κεντρικά βάρη και τα γνωστοποιεί στους χρήστες, ώστε να αποτελέσουν τα αρχικά βάρη για τα τοπικά μοντέλα.
2. Οι χρήστες εκπαιδεύουν τα τοπικά μοντέλα πάνω στο δικό τους σύνολο δεδομένων και εξάγουν επικαιροποιημένα τοπικά βάρη, τα οποία αποστέλλουν στον κεντρικό εξυπηρετητή.
3. Ο κεντρικός εξυπηρετητής συμψηφίζει τα βάρη που παρέλαβε σύμφωνα με την ακόλουθη σχέση, ώστε να προκύψουν τα νέα κεντρικά βάρη :

$$f(w) = \sum_{i=1}^k \frac{n_k}{n} F_k(w)$$

όπου, $F_k(w)$ είναι τα τοπικά βάρη του χρήστη k , τα οποία συμμετέχουν στο άθροισμα των βαρών σταθμισμένα κατά την συνεισφορά του χρήστη στο τελικό ενιαίο σύνολο δεδομένων, δηλαδή κατά ένα παράγοντα $\frac{n_k}{n}$, όπου το n_k αντιστοιχεί στην πληθικότητα του συνόλου δεδομένων του χρήστη k και το n στο άθροισμα των k αυτών πληθικοτήτων.

4. Ο κεντρικός εξυπηρετητής αποστέλλει τις νέες κεντρικές παραμέτρους στους τοπικούς χρήστες και η διαδικασία επαναλαμβάνεται για έναν ορισμένο αριθμό εποχών (γύρων) ή μέχρι να υπάρξει σύγκλιση.

2.8 Άλλες Μέθοδοι

Παρακάτω θα περιγράψουμε κάποιες μεθοδολογίες οι οποίες μας επιτρέπουν να μελετήσουμε με πιο συστηματικό τρόπο την εκπαίδευση των μοντέλων που εξετάζουμε, καθώς αποτελούν εργαλεία για τον έλεγχο της αξιοπιστίας των αποτελεσμάτων και των ισχυρισμών μας, την μελέτη της υπερπροσαρμογής και την εύρεση βέλτιστων υπερπαραμέτρων.

2.8.1 Διασταυρωμένη Επικύρωση

Η *διασταυρωμένη επικύρωση* (cross validation) είναι μια ευρέως χρησιμοποιούμενη μέθοδος της μηχανικής μάθησης που χρησιμοποιείται για την αντικειμενικότερη αξιολόγηση της αποτελεσματικότητας ενός μοντέλου. Η διασταυρωμένη επικύρωση χωρίζει το σύνολο εκπαίδευσης σε k υποσύνολα, για αυτό ονομάζεται και συχνά k -πλη διασταυρωμένη επικύρωση. Στην συνέχεια επιλέγεται τυχαία ένα εκ των υποσυνόλων για να αποτελέσει το σύνολο επικύρωσης και τα υπόλοιπα απαρτίζουν το σύνολο ελέγχου. Η διαδικασία επαναλαμβάνεται k φορές, με το μοντέλο να επανεκπαιδεύεται, χρησιμοποιώντας κάθε φορά διαφορετικό σύνολο επικύρωσης. Οι μετρικές αξιολόγησης του μοντέλου για κάθε γύρο συνοψίζονται, έτσι ώστε να παράγουν μια αντικειμενικότερη μετρική για την απόδοση του μοντέλου. Με αυτόν τον τρόπο, τα δεδομένα εκπαίδευσης χρησιμοποιούνται στο έπακρο και επιτυγχάνεται η διερεύνηση των δυνατοτήτων γενίκευσης ενός μοντέλου, μειώνοντας σημαντικά τον κίνδυνο να προκύψουν παραπλανητικά συμπεράσματα εξαιτίας της τυχαιότητας κατά την επιλογή συνόλου εκπαίδευσης και περιορίζοντας τον κίνδυνο μη εντοπισμού της υπερπροσαρμογής. Σημαντικός παράμετρος είναι η τιμή k , για την οποία συχνά χρησιμοποιούνται οι τιμές 5 ή 10 [5].

2.8.2 Αναζήτηση Πλέγματος

Η *αναζήτηση πλέγματος* (Grid Search) αποτελεί μια μέθοδο εξαντλητικού ελέγχου διαφορετικών συνδυασμών υπερπαραμέτρων ενός μοντέλου, ώστε να βρεθεί ο καλύτερος εξ αυτών, βάση κάποιας μετρικής [34]. Η επιλογή των κατάλληλων υπερπαραμέτρων παίζει σημαντικό ρόλο στην τελική επιτυχία του μοντέλου και δεν προκύπτει από την εκπαίδευση του μοντέλου, αλλά μέσω δοκιμών. Η αναζήτηση πλέγματος μας παρέχει ένα συστηματικό τρόπο να δοκιμάσουμε διαφορετικούς συνδυασμούς υπερπαραμέτρων. Για κάθε μια από τις υπερπαραμέτρους που θα εξεταστούν δίνονται κάποιες συγκεκριμένες τιμές, που είναι επιθυμητό να ελεγχθούν και η συνάρτηση αναζήτησης πλέγματος δημιουργεί ένα «πλέγμα» όλων των δυνατών συνδυασμών και εκπαιδεύει το μοντέλο σε καθένα από αυτούς [35]. Με αυτό τον τρόπο, μπορεί να αξιολογήσει όλους τους συνδυασμούς και να εντοπίσει τον καλύτερο. Ο έλεγχος της επίδοσης του μοντέλου για κάθε διαφορετικό συνδυασμό έχει σημασία, διότι μπορεί να μας δώσει διαφορετικά και πιο αξιόπιστα αποτελέσματα από τον σειριακό έλεγχο κάθε υπερπαραμέτρου. Ωστόσο, να σημειώσουμε ότι και η αναζήτηση πλέγματος δεν γίνεται να είναι εξαντλητική. Συνήθως, εξετάζονται, για μια υπερπαραμέτρο, τιμές σε μια μεγάλη περιοχή αρχικά, η οποία στενεύει προκειμένου να εντοπιστεί μια μικρή περιοχή τιμών, η οποία για την δοθείσα υπερπαραμέτρο δίνει τα καλύτερα αποτελέσματα, σε συνδυασμό πάντα με τις τιμές των υπόλοιπων υπερπαραμέτρων. Συχνά, η αναζήτηση πλέγματος συνδυάζεται με διασταυρωμένη επικύρωση, ώστε να εξασφαλιστεί η αξιοπιστία των αποτελεσμάτων της [34].

2.8.3 Έλεγχος t του Student (Student's t-test)

Ο έλεγχος t είναι ένας στατιστικός έλεγχος, ο οποίος προτάθηκε το 1908 από τον William Gosset, σε δημοσίευσή του, υπό το ψευδώνυμο Student [36]. Ο έλεγχος αυτός χρησιμοποιείται για να διαπιστωθεί κατά πόσο η διαφορά δύο κατανομών είναι πραγματική ή οφείλεται σε στατιστική τυχαιότητα και βρίσκει συχνά εφαρμογή στην αξιολόγηση του καλύτερου μοντέλου μηχανικής μάθησης. Χρησιμοποιείται, συγκεκριμένα, για να ελέγξει κατά πόσο οι διαφορετικές επιδόσεις δύο μοντέλων είναι πραγματικές ή προκύπτουν από στατιστική τυχαιότητα κατά την δημιουργία του συνόλου εκπαίδευσης και κατά την εκπαίδευση του μοντέλου [37], [38]. Ο έλεγχος αυτός στηρίζεται σε μια βασική, *μηδενική υπόθεση* (null hypothesis), σύμφωνα με την οποία θεωρούμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στις επιδόσεις των δύο εξεταζόμενων μοντέλων [38], [37]. Η διαδικασία που ακολουθείται κατόπιν είναι η εξής: οι επιδόσεις των δύο μοντέλων υπολογίζονται για k διαφορετικά σύνολα εκπαίδευσης με την χρήση της διασταυρωμένης επικύρωσης και στη συνέχεια υπολογίζεται μια τιμή t , σύμφωνα με τον ακόλουθο τύπο και θεωρώντας ότι οι επιδόσεις των μοντέλων έχουν προκύψει ανεξάρτητα, ακολουθούν περίπου κανονική κατανομή και ισχύει το Null Hypothesis:

$$t = \frac{\bar{d}}{\bar{\sigma}}$$

Όπου,

$$\bar{d} = \frac{1}{k} \sum_{i=1}^k d_i$$

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^k (d_i - \bar{d})^2}{k - 1}$$

Με d_i να είναι η διαφορά μεταξύ των i -οστων επιδόσεων των εξεταζόμενων μοντέλων και $k - 1$ οι βαθμοί ελευθερίας, με k τον αριθμό των επαναλήψεων της διασταυρωμένης επικύρωσης [38], [39].

Αφού υπολογιστεί η τιμή του t , σε συνδυασμό με το πλήθος των βαθμών ελευθερίας, χρησιμοποιείται για να προκύψει, από μια δοσμένη συνάρτηση πυκνότητας-πιθανότητας για το t , η πιθανότητα P_{value} , να ισχύει η μηδενική υπόθεση. Η πιθανότητα αυτή συγκρίνεται με ένα επιλεγμένο επίπεδο σημαντικότητας α , συνήθως $\alpha = 0.05$. Αν η πιθανότητα είναι μικρότερη από το

επίπεδο σημαντικότητας τότε η μηδενική υπόθεση απορρίπτεται και τα δύο μοντέλα έχουν στατιστικά σημαντικά διαφορετικές επιδόσεις [38], [37].

Ωστόσο, η μέθοδος δεν επιτρέπει ασφαλείς υποθέσεις κατά την σύγκριση των επιδόσεων δύο μοντέλων μηχανικής μάθησης, για επιδόσεις που έχουν προκύψει από την διασταυρωμένη επικύρωση στο ίδιο σύνολο δεδομένων, καθώς τόσο οι επιδόσεις των δύο διαφορετικών μοντέλων δεν είναι ανεξάρτητες αφού προκύπτουν από την εκπαίδευση στο ίδιο σύνολο δεδομένων, ενώ και οι επιδόσεις για το ίδιο μοντέλο δεν είναι ανεξάρτητες μεταξύ τους αφού προκύπτουν από εκπαίδευση σε επικαλυπτόμενα υποσύνολα του συνόλου εκπαίδευσης. Έχουν προταθεί διάφορες μέθοδοι, για την διόρθωση του ελέγχου, όπως η εφαρμογή του *t*-ελέγχου με διασταυρωμένη επικύρωση [37]. Στην παρούσα εργασία, εφαρμόστηκε η πρόταση των Nadeau και Bengio, οι οποίοι προτείνουν την χρήση μιας διορθωμένης τιμής διακύμανσης, κατά τον υπολογισμό της τιμής *t*, ώστε να συμπεριληφθεί η επίδραση της εξάρτησης. Η διορθωμένη τιμή είναι [39]:

$$\bar{\sigma}_{mod}^2 = \left(\frac{1}{k} + \frac{n1}{n2}\right) \bar{\sigma}^2$$

Όπου *n1*, πληθικότητα του συνόλου εκπαίδευσης και *n2*, πληθικότητα του συνόλου ελέγχου.

Μετά τον υπολογισμό της τιμής *t*, με βάση την νέα διακύμανση η διαδικασία εύρεσης της πιθανότητας *P_{value}* και ο έλεγχος για την αποδοχή της μηδενικής υπόθεσης, ακολουθείται όπως περιεγράφηκε παραπάνω.

3. Περιγραφή και Προετοιμασία Συνόλου Δεδομένων

Το πρώτο βήμα πριν την εφαρμογή τεχνικών Μηχανικής Μάθησης είναι η προσεκτική μελέτη και κατανόηση του συνόλου δεδομένου που θα χρησιμοποιηθεί, το οποίο καθορίζει την μορφή και την στόχευση της εκπαίδευσης και την προεπεξεργασία του. Στην παρούσα ενότητα παρουσιάζονται τα βήματα αυτά.

3.1 Περιγραφή του Συνόλου Δεδομένων

Τα σύνολα δεδομένων που είχαμε στην διάθεση μας περιείχαν πληροφορίες για τα χαρακτηριστικά δύο δίδυμων φορτηγών πλοίων. Συγκεκριμένα, κάθε σύνολο δεδομένων περιείχε δείγματα δέκα χαρακτηριστικών τα οποία συλλέγονταν σε χρονικό διαστήματα του ενός (1) λεπτού, κατά την περίοδο ενός (1) έτους (από 01/07/2021 έως 30/06/2022) [40]. Κατά την διαδικασία αυτή καταγράφηκαν πολλά ταξίδια του ίδιου πλοίου σε διαφορετικές περιβαλλοντικές και καιρικές συνθήκες, καθώς και σε διαφορετικές συνθήκες φορτίου, με αποτέλεσμα την δημιουργία συνόλων δεδομένων 514.525 και 525.005 δειγμάτων, για το πρώτο και δεύτερο πλοίο, αντίστοιχα. Τα χαρακτηριστικά τα οποία συλλέχθηκαν είναι τα ακόλουθα [40]:

- Ταχύτητα ως προς το έδαφος (Speed Over Ground – SOG) :

Η ταχύτητα ως προς το έδαφος είναι η ταχύτητα του πλοίου ως προς κάποιο σταθερό σημείο, όπως το έδαφος ή κάποια σηματοδούρα. Μετριέται σε κόμβους (knots) , όπου ο κόμβος είναι μονάδα μέτρησης της ταχύτητας στην ναυτιλία και ένας κόμβος αντιστοιχεί σε ένα ναυτικό μίλι ανά ώρα.

- Ταχύτητα ως προς το νερό (Speed Through Water – STW) :
Η ταχύτητα ως προς το νερό είναι η σχετική ταχύτητα του πλοίου ως προς την ταχύτητα του νερού. Μετριέται επίσης σε κόμβους. Για παράδειγμα, ένα πλοίο με μηδενική ταχύτητα, σε νερό που έχει ταχύτητα 2 κόμβους θα έχει STW=0 knots αλλά SOG=2 knots.
- Κατεύθυνση (Heading) :
Είναι η κατεύθυνση προς την οποία κινείται το πλοίο. Μετριέται σε μοίρες, ξεκινώντας από 0° που αντιστοιχούν στον Βορρά και φτάνοντας, ακολουθώντας την φορά των δεικτών του ρολογιού, μέχρι τις 359°.
- Συνεχής Ταχύτητα του Ανέμου (Continuous Wind Speed – CWS) :
Πρόκειται για το μέτρο της πραγματική ταχύτητα του ανέμου μετρημένο σε $\frac{m}{sec}$ (μέτρα ανά δευτερόλεπτο).
- Διακριτή Ταχύτητα του Ανέμου (Discretized Wind Speed – DWS) :
Η διακριτή ταχύτητα του ανέμου προκύπτει άμεσα από την συνεχή ταχύτητα του ανέμου, μέσω του συσχετισμού της κλίμακας Μποφόρ (bft). Σύμφωνα με την κλίμακα Μποφόρ, κάθε συνεχή ταχύτητα του ανέμου, μετρημένη, στην προκειμένη περίπτωση σε $\frac{m}{sec}$, αντιστοιχίζεται σε μία κλίμακα bft που αριθμείται στο διάστημα 0-12, με το 0 να αντιστοιχεί στην νηνεμία και το 12 σε κατάσταση τυφώνα. Για παράδειγμα, για CWS=9.2 $\frac{m}{sec}$ έχουμε DWS=5bft.
- Κατεύθυνση του ανέμου (Wind Direction – WD) :
Πρόκειται για την κατεύθυνση του ανέμου ως προς το πλοίο. Μετριέται σε μοίρες, με 0° να αντιστοιχούν σε κατεύθυνση ανέμου προς την πλώρη.
- Βύθισμα Μπροστά (Draft Forward – DF) :
Το βύθισμα μπροστά αναφέρεται στο βύθισμα της πλώρης, ως προς το επίπεδο του νερού. Μετριέται σε μέτρα (m), ως προς την κάθετο της πλώρης βάση προκαθορισμένων επιπέδων βάθους.
- Βύθισμα Πίσω (Draft Aft – DA) :
Το βύθισμα πίσω αναφέρεται στο βύθισμα της πρύμνης, ως προς το επίπεδο του νερού. Μετριέται και αυτό σε μέτρα (m), ως προς την κάθετο της πρύμνης βάση προκαθορισμένων επιπέδων βάθους.
- Ισορροπία (Trim) :
Το Trim υπολογίζεται βάση της ακόλουθης σχέσης:

$$Trim = DF - DA \quad , (1)$$

Μετριέται σε μέτρα (m) ως προς ένδειξη-γραμμή που υπάρχει στο κέντρο του πλοίου και η διατήρηση ενός συγκεκριμένου εύρους τιμών του είναι καθοριστική για την ισορροπία του πλοίου. Επηρεάζεται, σε σημαντικό

βαθμό, από την κατανομή του φορτίου στο πλοίο, οπότε εμπεριέχει χρήσιμες πληροφορίες για αυτήν την παράμετρο [41] .

- **Ισχύς Κύριου Κινητήρα (Main Engine Power – MEP) :**
 Η ισχύς του κύριου κινητήρα είναι η συνολική ισχύς που παράγεται από τον κύριο κινητήρα (ή κινητήρες) προκειμένου να κινηθεί η προπέλα του πλοίου και μετρείται σε kW. Η ισχύς που παράγεται, προφανώς αντικατοπτρίζει τις ανάγκες σε ενέργεια του κινητήρα και επομένως είναι ανάλογη της κατανάλωσης καυσίμου.

Συνοπτικά, τα χαρακτηριστικά αυτά παρουσιάζονται στον Πίνακα 1.

Χαρακτηριστικό	Αγγλική Ονομασία	Αγγλική Συνομογραφία
Ταχύτητα ως προς το έδαφος	Speed Over Ground	SOG
Ταχύτητα ως προς το νερό	Speed Through Water	STW
Κατεύθυνση	Heading	-
Συνεχής Ταχύτητα Ανέμου	Continuous Wind Speed	CWS
Διακριτή Ταχύτητα Ανέμου	Discretized Wind Speed	DWS
Κατεύθυνση Ανέμου	Wind Direction	WD
Βύθισμα Μπροστά	Draft Forward	DF
Βύθισμα Πίσω	Draft Aft	DA
Ισορροπία	Trim	-
Ισχύς Κύριου Κινητήρα	Main Engine Power	MEP

Πίνακας 1: Χαρακτηριστικά του αρχικού συνόλου δεδομένων.

Στόχος της παρούσας εργασίας είναι η εφαρμογή τεχνικών μηχανικής μάθησης για την πρόβλεψη της κατανάλωσης καυσίμου. Όπως αναφέρθηκε, η κατανάλωση καυσίμου είναι ανάλογη του χαρακτηριστικού MEP. Άρα, η πρόβλεψη της κατανάλωσης καυσίμου θα μπορούσε να προκύψει άμεσα από την πρόβλεψη του MEP. Για αυτόν τον λόγο, στόχος της εκπαίδευσης που θα ακολουθήσει αποτελεί η πρόβλεψη της τιμής MEP, η οποία θα αποτελέσει και την έξοδο για τα προβλήματα επιβλεπόμενης μάθησης. Επιπλέον, το MEP (και η κατανάλωση καυσίμου) εξαρτώνται από την αντίσταση του πλοίου, η οποία επηρεάζεται από παράγοντες όπως η κατασκευή και συντήρηση του πλοίου, το φορτίο και οι καιρικές συνθήκες και από την ταχύτητα που καλείται να αναπτύξει το πλοίο [42]. Τα χαρακτηριστικά που έχουν μετρηθεί παρέχουν πληροφορίες για τις καιρικές συνθήκες, την ταχύτητα του πλοίου και την κατανομή του φορτίου του και μπορούν να μας δώσουν μια αρκετά πλήρη εικόνα για την τρέχουσα κατανάλωση καυσίμου. Ακόμα, εφόσον τα δεδομένα αφορούν ένα πλοίο και μάλιστα για χρονικό διάστημα ενός χρόνου, η παράμετροι της κατασκευής και της συντήρησης του πλοίου μπορούν να θεωρηθούν σταθερές και αποτυπώνονται εμμέσως στην μέτρηση του MEP, οπότε κατά μια έννοια λαμβάνονται υπόψιν στα πλαίσια της εκπαίδευσης.

3.2 Προεπεξεργασία Δεδομένων

Στόχος είναι να εφαρμοστούν διάφορες τεχνικές μηχανικής μάθησης στα πλαίσια της προσπάθειας πρόβλεψης της τιμής MEP , βάσει των χαρακτηριστικών που διαθέτουμε. Όμως το σύνολο δεδομένων που διαθέτουμε βρίσκεται σε πρωταρχική μορφή και, όπως εξηγήσαμε και στην Ενότητα 2.6, θα πρέπει να επεξεργαστούμε τα δεδομένα μας με διάφορες τεχνικές προεπεξεργασίας προκειμένου να εξάγουμε ποιοτικές προβλέψεις.

Το πρώτο βήμα στην διαδικασία της προεπεξεργασίας είναι η διαχείριση των απουσιαζουσών τιμών. Να σημειώσουμε εδώ, ότι δείγματα με τιμές που απουσιάζουν αντιστοιχούν περίπου σε ένα ποσοστό 8% του συνόλου δεδομένων, το οποίο, δεδομένου και του μεγέθους του συνόλου δεδομένων είναι αρκετά μικρό. Εξαιτίας της συχνής δειγματοληψίας (ανά 1 λεπτό) και δεδομένου ότι τα χαρακτηριστικά του πλοίου και του καιρού αλλάζουν με αργό ρυθμό, κρίνουμε ότι μεταξύ διαδοχικών δειγμάτων δεν θα υπάρχουν μεγάλες αποκλίσεις. Οπότε, επιλέγουμε να εφαρμόσουμε μια μέθοδο απόδοσης για να διαχειριστούμε τις άγνωστες τιμές, κατά την οποία αποδίδουμε στην άγνωστη τιμή του χαρακτηριστικού i τον μέσο όρο της πρώτης γνωστής προηγούμενης και της επόμενης τιμής του δείγματος που μας ενδιαφέρει για το χαρακτηριστικό που εξετάζουμε.

Στο επόμενο βήμα, πρέπει να διαχειριστούμε τις έκτοπες τιμές. Συγκεκριμένα, τα δείγματα εκείνα κατά τα οποία το πλοίο είναι στατικό (συνήθως επειδή βρίσκεται σε λιμάνι) δεν αφορούν την πρόβλεψη της κατανάλωσης καυσίμου και απορρίπτονται απευθείας από το σύνολο δεδομένων. Τα δείγματα αυτά προσδιορίζονται από την τιμή SOG , με την συνθήκη $SOG < 0.2 \text{ knots}$, για να θεωρηθεί το πλοίο στατικό. Μετά από την εφαρμογή ενός φίλτρου κινούμενου μέσου το οποίο χρησιμοποιείται για να επιτευχθεί εξομάλυνση στις καμπύλες των χαρακτηριστικών και να αντιμετωπιστεί η παραμόρφωση η οποία προκύπτει από την διαδικασία της δειγματοληψίας, προχωρήσαμε σε έναν επιπλέον έλεγχο για ακραίες τιμές. Συγκεκριμένα, μετά από μια επιθεώρηση των, εξομαλυμένων πλέον, καμπυλών των χαρακτηριστικών καταλήξαμε να απορρίψουμε τις ακραίες τιμές, χρησιμοποιώντας της συνθήκη:

$$|x| > \mu + 3\sigma$$

όπου, x είναι η εξεταζόμενη τιμή, μ η μέση τιμή της κατανομής του χαρακτηριστικού που ελέγχουμε και σ η τυπική απόκλιση της κατανομής.

Στην συνέχεια, προχωρήσαμε σε υποδειγματοληψία του συνόλου δεδομένων μας για να επιταχύνουμε την διαδικασία της εκπαίδευσης και να περιορίσουμε την περιττή πληροφορία. Επειδή, όπως εξηγήσαμε οι μεταβολές στην κατάσταση του

πλοίου και του καιρού είναι αργές, επιλέξαμε να κρατήσουμε δείγματα ανά 15 λεπτά, υποπολλαπλασιάζοντας με αυτόν τον τρόπο την πληθικότητα του συνόλου δεδομένων κατά ένα παράγοντα 15. Κατόπιν, για να μειώσουμε επιπλέον την διαστατικότητα του συνόλου και τον πλεονασμό πληροφορίας, εξετάσαμε αν υπάρχουν περιττά ή αλληλεξαρτώμενα χαρακτηριστικά και τα απορρίψαμε ή κρατήσαμε μόνο ένα, το οποίο να συμπυκνώνει ικανοποιητικά την κοινή πληροφορία, αντίστοιχα. Συγκεκριμένα, η κατεύθυνση του πλοίου δεν σχετίζεται με την κατανάλωση καυσίμου και επομένως, το χαρακτηριστικό αυτό απορρίφθηκε. Επιπλέον, είναι προφανές ότι οι τιμές CWS και DWS είναι εξαρτώμενες μεταξύ τους και προκύπτει άμεσα η μια από την άλλη μέσω της κλίμακας Μποφόρ, οπότε επιλέξαμε να κρατήσουμε μόνο το χαρακτηριστικό DWS. Παρόμοια, τα χαρακτηριστικά DF, DA και Trim συσχετίζονται μέσω της σχέσης (1) και επιλέξαμε να κρατήσουμε μόνο το Trim, το οποίο θεωρήσαμε ότι συμπυκνώνει καλύτερα όλη την πληροφορία. Τέλος, υπήρξε η σκέψη να απορριφθεί κάποιο από τα χαρακτηριστικά SOG ή STW, ωστόσο το STW μας παρέχει επιπλέον πληροφορίες για την ταχύτητα των ρευμάτων, ενώ μετά από έναν σύντομο έλεγχο της μεταξύ τους συσχέτισης προέκυψε ότι υπάρχει υψηλός βαθμός ανεξαρτησίας μεταξύ των δύο αυτών χαρακτηριστικών. Συγκεκριμένα, προέκυψε συντελεστής συσχέτισης Pearson, $R = 0.23$, ο οποίος υποδηλώνει μεγάλο βαθμό ανεξαρτησίας μεταξύ των δύο αυτών χαρακτηριστικών [40].

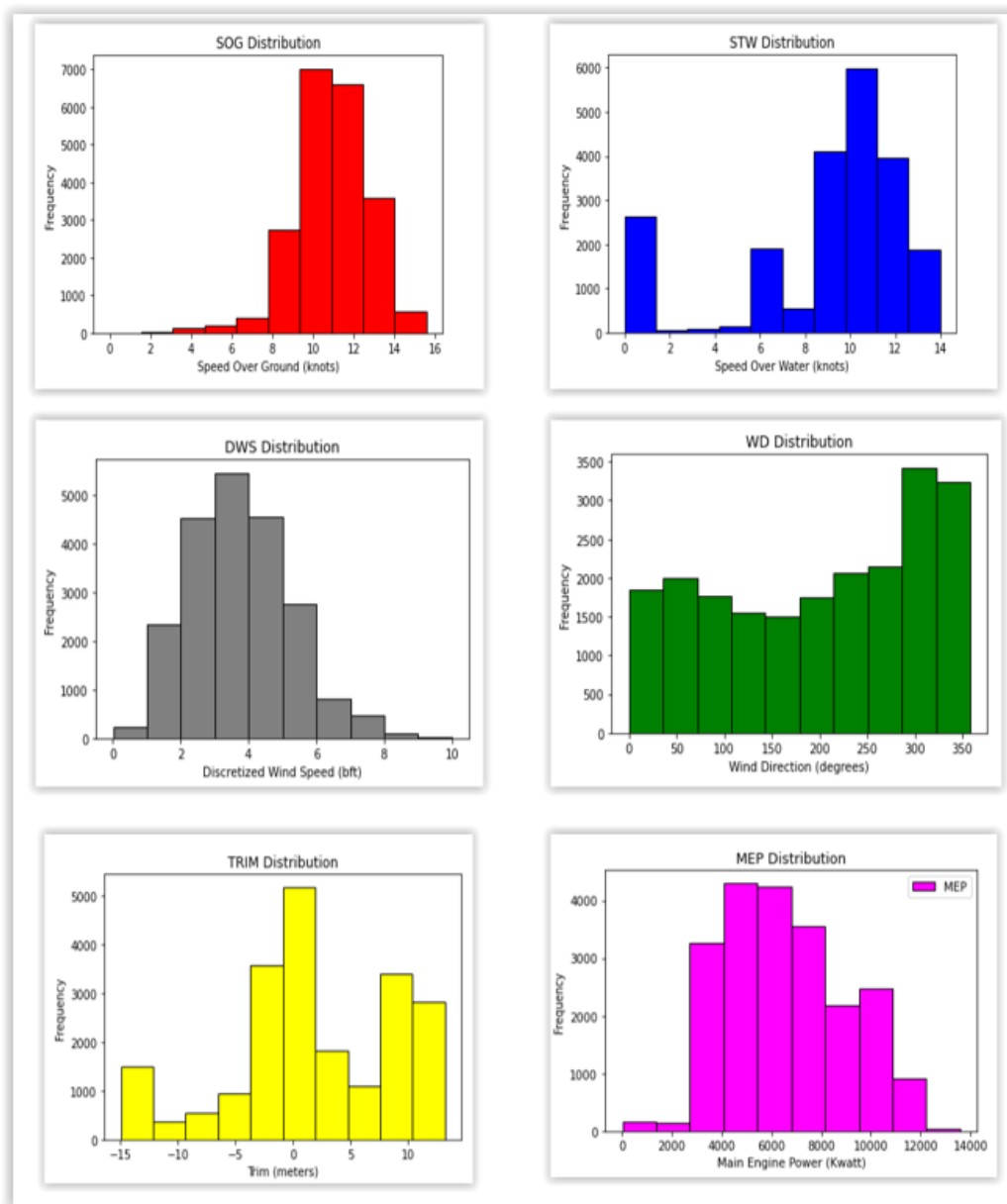
Στον πίνακα 2, παρουσιάζονται τα χαρακτηριστικά του τελικού συνόλου δεδομένων, που προέκυψε μετά την προεπεξεργασία.

Χαρακτηριστικό	Αγγλική Ονομασία	Αγγλική Συνομογραφία
Ταχύτητα ως προς το έδαφος	Speed Over Ground	SOG
Ταχύτητα ως προς το νερό	Speed Through Water	STW
Διακριτή Ταχύτητα Ανέμου	Discretized Wind Speed	DWS
Κατεύθυνση Ανέμου	Wind Direction	WD
Ισορροπία	Trim	-
Ισχύς Κύριου Κινητήρα	Main Engine Power	MEP

Πίνακας 2: Χαρακτηριστικά του συνόλου δεδομένων, μετά την προεπεξεργασία.

Μετά από αυτά τα βήματα της προεπεξεργασίας οι κατανομές των χαρακτηριστικών που απέμειναν παρουσιάζονται στην Εικόνα 17.

Τέλος, πάντοτε πριν προχωρήσουμε στην τροφοδότηση του συνόλου δεδομένων στο εκάστοτε μοντέλο εφαρμόζαμε την τεχνική κανονικοποίησης ελαχίστου-μεγίστου.



Εικόνα 17: Κατανομές των χαρακτηριστικών του συνόλου δεδομένων μετά την προεπεξεργασία.

Από την εποπτεία των κατανομών της Εικόνας 17 προκύπτουν κάποιες χρήσιμες παρατηρήσεις, οι οποίες αξίζει να σημειωθούν. Αρχικά, παρατηρούμε ότι πράγματι, καταφέραμε να εξαλείψουμε τις πολύ χαμηλές τιμές SOG, ωστόσο υπάρχουν ακόμα πολύ χαμηλές τιμές STW. Αυτό μπορεί να συμβεί σε περιπτώσεις που το πλοίο βρίσκεται εν πλω και κινείται με την ταχύτητα του νερού, επειδή, για παράδειγμα, έχει ρίξει άγκυρα, έχει χαμηλώσει την ταχύτητά του ή αντιμετωπίζει ισχυρά ρεύματα. Η παρατήρηση αυτή συνηγορεί επιπλέον στο γεγονός ότι τα χαρακτηριστικά SOG και STW περιέχουν διαφορετική πληροφορία. Κατά τα άλλα, για τα χαρακτηριστικά DWS και MEP παρατηρούμε μια σχεδόν ομαλή κατανομή με μηδενισμό των ακραίων τιμών, ενώ για το χαρακτηριστικό WD η κατανομή του είναι σχεδόν σταθερή, με περισσότερες τιμές για κατεύθυνση ανέμου σχεδόν προς την πλώρη. Επιπλέον, παρατηρούμε για το Trim μεγάλο μέρος των τιμών του να είναι συγκεντρωμένες γύρω από το 0, πράγμα που αντικατοπτρίζει την προσπάθεια να κατανεμηθεί το φορτίο ώστε να υπάρξει ισορροπημένη βύθιση του μπροστινού και πίσω μέρους. Ωστόσο, έχουμε και αρκετά δείγματα συγκεντρωμένα σε ακραίες τιμές, μέχρι 15m, ενώ τα περισσότερα από αυτά αντιστοιχούν σε θετικό Trim. Τα 15m είναι προφανώς το όριο που έχει οριστεί ότι δεν πρέπει να υπερβαίνει το Trim προκειμένου να εξασφαλιστεί η ευστάθεια του πλοίου. Βλέπουμε, επομένως ότι διαθέτουμε και ικανοποιητικό αριθμό δειγμάτων για αντίξοες συνθήκες αέρα (άνεμος στην πλώρη) και για άνιση κατανομή του Trim, οπότε μπορούν να μελετηθούν και αυτές οι δύσκολες περιπτώσεις.

4. Πειραματικά Αποτελέσματα Μηχανικής Μάθησης

Στο μεγαλύτερο μέρος της εργασίας χρησιμοποιήσαμε το σύνολο δεδομένων μόνο του πρώτου πλοίου για να τροφοδοτήσουμε την εκπαίδευση των μοντέλων μηχανικής μάθησης. Εκπαιδεύσαμε μοντέλα επιβλεπόμενης, συλλογικής μάθησης, αναδρομικά νευρωνικά δίκτυα, καθώς και μοντέλα ομοσπονδιακής μάθησης επί των δεδομένων και των δύο πλοίων, προκειμένου να εξετάσουμε την αποτελεσματικότητά τους στην εξαγωγή συμπερασμάτων για την παραγωγή MEP, στα συγκεκριμένα πλοία, αλλά και να μελετήσουμε πως θα μπορούσαν να συνεισφέρουν στην προσπάθεια πρόβλεψης της κατανάλωσης καυσίμου στον τομέα της ναυτιλίας γενικότερα.

4.1 Μοντέλα Επιβλεπόμενης Μάθησης

Στα πλαίσια της επιβλεπόμενης μάθησης, στόχος μας είναι η πρόβλεψη της τιμής του χαρακτηριστικού MEP για το πρώτο πλοίο, βάση των τιμών των χαρακτηριστικών SOG, STW, DWS, WD και Trim. Πρόκειται για πρόβλημα παλινδρόμησης, δεδομένου ότι η τιμή που καλούμαστε να προβλέψουμε είναι συνεχής. Όπως αναφέρθηκε στο θεωρητικό μέρος, για να εκπαιδευτεί ένα μοντέλο επιβλεπόμενης μάθησης πρέπει να τροφοδοτηθεί με ένα χαρακτηρισμένο σύνολο εκπαίδευσης, ενώ οι επιδόσεις του μοντέλου ελέγχονται βάση ενός συνόλου ελέγχου. Εμείς διαθέτουμε το προεπεξεργασμένο σύνολο δεδομένων του πλοίου 1, με τις ετικέτες των δειγμάτων να είναι οι τιμές του MEP. Χωρίζουμε το σύνολο δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου, με το σύνολο ελέγχου να αντιστοιχεί στο 10% του μεγέθους του αρχικού συνόλου δεδομένων. Τα σύνολα εκπαίδευσης και ελέγχου είναι μη επικαλυπτόμενα, ενώ τα δείγματα που τοποθετούνται στο σύνολο ελέγχου επιλέγονται τυχαία. Είναι σημαντικό να μην παρεισφρήσει πληροφορία από το σύνολο εκπαίδευσης στο σύνολο ελέγχου, ώστε να διατηρηθεί η αντικειμενικότητα της τελικής αξιολόγησης, αφού το μοντέλο δεν πρέπει να εκπαιδευτεί πάνω σε πληροφορίες του συνόλου ελέγχου γιατί τότε μπορεί να επιτύχει καλύτερες επιδόσεις, οι οποίες όμως δεν θα είναι αξιόπιστες, αφού το σύνολο ελέγχου δεν είναι πλέον άγνωστο. Για αυτόν τον λόγο, η κανονικοποίηση μεγίστου-ελαχίστου γίνεται βάση των τιμών μόνο του συνόλου εκπαίδευσης και κατόπιν, ο μετασχηματισμός που προκύπτει εφαρμόζεται και στο σύνολο ελέγχου. Μετά και την κανονικοποίηση, το σύνολο εκπαίδευσης τροφοδοτείται στα μοντέλα επιβλεπόμενης μάθησης, τα οποία εκπαιδεύονται ώστε να προβλέπουν στην έξοδο τους μια τιμή MEP, βάσει των τιμών των πέντε χαρακτηριστικών που λαμβάνουν. Μόλις ολοκληρωθεί η εκπαίδευση, η απόδοση των μοντέλων αξιολογείται με την βοήθεια του συνόλου ελέγχου. Τα μοντέλα

επιβλεπόμενης μάθησης που θα χρησιμοποιηθούν είναι τα εξής: γραμμική παλινδρόμηση, δέντρο αποφάσεων, παλινδρόμηση διανυσμάτων υποστήριξης, τυχαίο δάσος και νευρωνικό δίκτυο πολλαπλών στρωμάτων.

4.1.1 Αποτελέσματα Γραμμικής Παλινδρόμησης

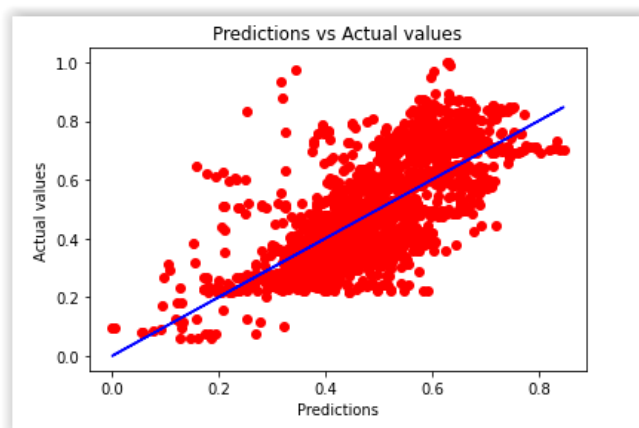
Μετά την ολοκλήρωση της εκπαίδευσης του μοντέλου γραμμικής παλινδρόμησης τροφοδοτήσαμε στο μοντέλο το σύνολο ελέγχου, ώστε να αξιολογήσουμε την απόδοση του βάσει των σφαλμάτων μεταξύ των προβλεπόμενων και των πραγματικών εξόδων του συνόλου ελέγχου. Οι μετρικές που χρησιμοποιήσαμε είναι το ΜΤΣ, ΜΑΣ και το ΜΑΠΣ. Τα αποτελέσματα που προέκυψαν παρουσιάζονται παρακάτω:

$$ΜΤΣ = 0.01541$$

$$ΜΑΣ = 0.09646$$

$$ΜΑΠΣ = 0.22555$$

Οι διαφορετικές τιμές των τριών μετρικών οφείλονται στις διαφορετικές μαθηματικές σχέσεις υπολογισμού τους. Χρησιμοποιήσαμε διαφορετικές μετρικές για να διερευνήσουμε τους διαφορετικούς τρόπους αξιολόγησης ενός μοντέλου παλινδρόμησης. Ωστόσο, στο εξής θα χρησιμοποιούμε ως μετρική μόνο το ΜΤΣ, το οποίο είναι πλήρως ικανό να αποτυπώσει την απόδοση ενός μοντέλου.



Εικόνα 18: Απόκλιση μεταξύ των προβλεπόμενων (*predictions*) και των πραγματικών (*actual*) τιμών. Η μπλε ευθεία είναι η ευθεία που παράγεται από την γραμμική παλινδρόμηση και θα ταυτιζόταν με τις κόκκινες κουκίδες σε περίπτωση που η πρόβλεψη ήταν πλήρως ακριβής.

Σε κάθε περίπτωση, παρατηρούμε ότι τα αποτελέσματα της εκπαίδευσης με γραμμική παλινδρόμηση είναι ήδη ικανοποιητικά, παρότι η γραμμική παλινδρόμηση θα αποτελέσει το απλούστερο μοντέλο μας. Για να έχουμε και μια εικόνα των αποτελεσμάτων μας δημιουργήσαμε ένα γράφημα των προβλεπόμενων προς των πραγματικών εξόδων του μοντέλου, που παρουσιάζεται στην Εικόνα 18.

Επιπλέον, για να βεβαιωθούμε για την αξιοπιστία των μετρήσεων μας και να ελέγξουμε το ενδεχόμενο της υπερπροσαρμογής, χρησιμοποιήσαμε την μέθοδο της 10-πλης διασταυρωμένης επικύρωσης. Κατασκευάσαμε δική μας συνάρτηση, η οποία αναλαμβάνει τον χωρισμό του συνόλου δεδομένων σε 10 τυχαία υποσύνολα, χωρίς επικαλύψεις. Κατόπιν, σε κάθε έναν από τους 10 γύρους που έπονται, επιλέγει ένα από τα δέκα υποσύνολα, διαφορετικό κάθε φορά, για σύνολο ελέγχου και κρατάει τα υπόλοιπα υποσύνολα ως σύνολο εκπαίδευσης. Στην συνέχεια, τα σύνολα εκπαίδευσης και ελέγχου που προκύπτουν κανονικοποιούνται και χρησιμοποιούνται για την εκπαίδευση και τον έλεγχο του τρέχοντος μοντέλου. Η συνάρτηση που υλοποιεί την διασταυρωμένη επικύρωση μας επιστρέφει μια λίστα με τις 10 μετρήσεις (εδώ MTΣ) που προκύπτουν για την επιτυχία των προβλέψεων, για το σύνολο εκπαίδευσης και το σύνολο ελέγχου. Στην προκειμένη περίπτωση, το εξεταζόμενο μοντέλο είναι γραμμική παλινδρόμηση και παρακάτω παρουσιάζονται τα MTΣ που προκύπτουν για κάθε σύνολο εκπαίδευσης και ελέγχου μετά την διασταυρωμένη επικύρωση.

$$MTΣ_{\text{Συνόλου Ελέγχου}} = \begin{bmatrix} 0.01499, 0.01539, 0.01538, 0.01482, 0.01453, \\ 0.01589, 0.01624, 0.01713, 0.01508, 0.01540 \end{bmatrix}$$

$$MTΣ_{\text{Συνόλου Εκπαίδευσης}} = \begin{bmatrix} 0.01550, 0.01546, 0.01546, 0.01552, 0.01588, \\ 0.01540, 0.01536, 0.01527, 0.01549, 0.01545 \end{bmatrix}$$

Παρατηρούμε ότι το αποτέλεσμα που πήραμε για το MTΣ επιβεβαιώνεται με την χρήση της διασταυρωμένης επικύρωσης. Επιπλέον, δεν υπάρχουν ενδείξεις για υπερπροσαρμογή του μοντέλου, δεδομένου ότι η επίδοση για την σύνολο ελέγχου δεν διαφέρει σημαντικά από την επίδοση για το σύνολο εκπαίδευσης. Μάλιστα, παρατηρούμε ότι, ορισμένες φορές, οι προβλέψεις για το σύνολο ελέγχου είναι καλύτερες από αυτές για το σύνολο εκπαίδευσης. Αυτό μπορεί να οφείλεται στον τυχαίο διαχωρισμό του συνόλου εκπαίδευσης και του συνόλου ελέγχου, εξαιτίας του οποίου μπορεί να κατέληξαν απλούστερα δείγματα στο σύνολο ελέγχου.

Να σημειώσουμε, ότι στο εξής η διασταυρωμένη επικύρωση χρησιμοποιήθηκε για όλα τα μοντέλα επιβλεπόμενης και συλλογικής μάθησης, συνήθως για τις βέλτιστες υπερπαραμέτρους, ωστόσο τα αποτελέσματα της παρουσιάζονται μόνο στην περίπτωση που παρουσιάζουν κάποιο επιπλέον ενδιαφέρον σε σχέση με την απλή μέτρηση του MTΣ, που δίνεται για όλα αυτά τα μοντέλα.

4.1.2 Αποτελέσματα Δέντρων Απόφασης

Η τιμή του MTΣ που προέκυψε μετά την τροφοδότηση του συνόλου ελέγχου στο δένδρο απόφασης είναι η ακόλουθη:

$$MTΣ = 0.000589$$

Ενδιαφέρον παρουσιάζουν, στην προκειμένη περίπτωση, και τα αποτελέσματα της διασταυρωμένης επικύρωσης:

$$MT\Sigma_{\text{Συνόλου Ελέγχου}} = \begin{bmatrix} 0.00064, 0.00065, 0.00069, 0.00049, 0.00065, \\ 0.00119, 0.00069, 0.00084, 0.00080, 0.00028 \end{bmatrix}$$

$$MT\Sigma_{\text{Συνόλου Εκπαίδευσης}} = \begin{bmatrix} 0.0, 0.0, 0.0, 0.0, 0.0, \\ 0.0, 0.0, 0.0, 0.0, 0.0 \end{bmatrix}$$

Με βάση τα αποτελέσματα των δέντρων απόφασης στο σύνολο εκπαίδευσης, παρατηρούμε ότι υπάρχει το ενδεχόμενο της υπερπροσαρμογής. Ωστόσο, το ενδεχόμενο αυτό απομακρύνεται εξαιτίας των επίσης πολύ χαμηλών σφαλμάτων του συνόλου ελέγχου. Συμπεραίνουμε, λοιπόν ότι δεν έχει γίνει υπερπροσαρμογή αλλά παρατηρούμε την ισχυρή ικανότητα των δέντρων απόφασης να αποτυπώσουν μοτίβα του συνόλου εκπαίδευσης. Επιπλέον, αναμένουμε ότι για τεχνικές που στηρίζονται στην λογική των δέντρων απόφασης αλλά περιορίζουν το ενδεχόμενο υπερπροσαρμογής, όπως τα τυχαία δάση, θα έχουμε πολύ καλά αποτελέσματα για το σύνολο δεδομένων μας και μάλιστα ίσως με ακόμα υψηλότερες δυνατότητες γενίκευσης.

4.1.3 Αποτελέσματα Παλινδρόμησης Διανυσμάτων Υποστήριξης

Η τιμή του $MT\Sigma$ που πήραμε για το μοντέλο διανυσμάτων υποστήριξης με ακτινική συνάρτηση βάσης στον πυρήνα είναι:

$$MT\Sigma = 0.004631$$

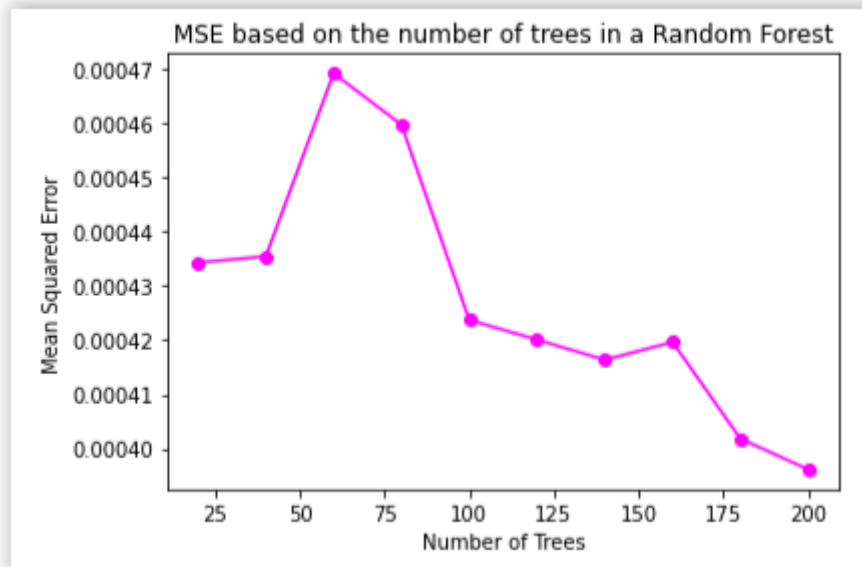
Η τιμή αυτή επαληθεύεται και από τα αποτελέσματα της διασταυρωμένης επικύρωσης.

Προκειμένου να βελτιώσουμε περαιτέρω την επίδοση του μοντέλου μας δοκιμάσαμε διαφορετικές συναρτήσεις πυρήνα. Συγκεκριμένα, δοκιμάσαμε γραμμικό πυρήνα, πολυωνυμικό τρίτου βαθμού και σιγμοειδές. Το χαμηλότερο σφάλμα προέκυψε ξεκάθαρα για την ακτινική συνάρτηση βάσης. Κάποια δυναμική βελτίωσης είχε ο πολυωνυμικός πυρήνας, για μεγαλύτερους βαθμούς, ωστόσο, δεν προχωρήσαμε σε περαιτέρω βελτιστοποίηση επειδή η εκπαίδευση με πυρήνα 3^{ου} βαθμού ήδη καθυστερούσε πολύ.

4.1.4 Αποτελέσματα Τυχαίου Δάσους

Στο μοντέλο τυχαίου δάσους, όλα τα δείγματα του συνόλου εκπαίδευσης χρησιμοποιούνται για την κατασκευή κάθε δέντρου απόφασης του τυχαίου δάσους, καθώς και όλα τα χαρακτηριστικά. Επιπλέον, ο αριθμός των δέντρων απόφασης που χρησιμοποιήσαμε αρχικά είναι 100. Ωστόσο, επειδή ο αριθμός των δέντρων απόφασης είναι μια σημαντική υπερπαραμέτρος του μοντέλου τυχαίου δάσους, όπως αναφέραμε και στην θεωρία αποφασίσαμε να ακολουθήσουμε μια διαδικασία

ρύθμισης, όπου δοκιμάσαμε διαφορετικούς αριθμούς δέντρων απόφασης και καταλήξαμε σε αυτόν που μας έδινε το ελάχιστο ΜΤΣ. Η διαδικασία αυτή παρουσιάζεται στην Εικόνα 19 για πλήθος δέντρων απόφασης από 20-200 με βήμα 20. Όπως βλέπουμε βέλτιστο ΜΤΣ επιτυγχάνεται για αριθμό δέντρων 200. Μεγαλύτεροι αριθμοί από το 200 ελέγχθηκαν, ωστόσο το 200 αποτελεί ελάχιστο, ενώ, όσο αυξάνεται ο αριθμός των δέντρων αυξάνεται και ο χρόνος εκπαίδευσης.



Εικόνα 19: ΜΤΣ (MSE) ανάλογα με το πλήθος των δέντρων (Number of Trees) σε ένα μοντέλο τυχαίου δάσους (Random Forest). Το καλύτερο ΜΤΣ επιτυγχάνεται για 200 δέντρα.

Το ΜΤΣ που επιτυγχάνεται για 200 δέντρα απόφασης είναι:

$$ΜΤΣ = 0.000396$$

Εδώ παραθέτουμε και τα αποτελέσματα της διασταυρωμένης επικύρωσης για τυχαίο δάσος με 200 δέντρα, ώστε να ελέγξουμε αν ο μεγάλος αριθμός δέντρων δεν δημιουργεί πρόβλημα υπερπροσαρμογής.

$$ΜΤΣ_{\text{Συνόλου Ελέγχου}} = \begin{bmatrix} 0.00025, 0.00031, 0.00049, 0.00050, 0.00065, \\ 0.00034, 0.00024, 0.00024, 0.00045, 0.00031 \end{bmatrix}$$

$$ΜΤΣ_{\text{Συνόλου Εκπαίδευσης}} = \begin{bmatrix} 0.00005, 0.00005, 0.00005, 0.00005, 0.00005, \\ 0.00005, 0.00005, 0.00005, 0.00005, 0.00005 \end{bmatrix}$$

Από την παρατήρηση των ΜΤΣ για το σύνολο εκπαίδευσης και το σύνολο ελέγχου, όπως αυτά προκύπτουν από την εφαρμογή της διασταυρωμένης επικύρωσης, παρατηρούμε ότι δεν φαίνεται να έχει συμβεί υπερπροσαρμογή του μοντέλου μας, καθώς το ΜΤΣ για το σύνολο εκπαίδευσης είναι μεν πολύ μικρό αλλά όχι μηδενικό ενώ το σφάλμα του συνόλου ελέγχου είναι αποδεκτά μεγαλύτερο από αυτού του συνόλου εκπαίδευσης. Ακόμα να σημειωθεί ότι ο μέσος όρος των ΜΤΣ που προκύπτουν από την διασταυρωμένη επικύρωση είναι:

$$ΜΤΣ = 0.00036$$

Επομένως, το ΜΤΣ που υπολογίσαμε για την πρόβλεψη από ένα μόνο σύνολο ελέγχου είναι πράγματι αντιπροσωπευτικό.

Επιπλέον, για την εύρεση του βέλτιστου αριθμού δένδρων δοκιμάσαμε και την αναζήτηση πλέγματος, αρχικά για μεγάλη περιοχή [10,310] με βήμα 50 και πήραμε βέλτιστα αποτελέσματα για μοντέλο με 110 δέντρα, οπότε δοκιμάσαμε περιοχή [10,210] με βήμα 30, πήραμε βέλτιστα αποτελέσματα για μοντέλο με 130 δέντρα, δοκιμάσαμε περιοχή [80,250] με βήμα 20, πήραμε για 200 δέντρα κ.ο.κ. Μετά από κάποιους ακόμα ελέγχους καταλήξαμε ότι η περιοχή γύρω από τα 200 δέντρα είναι πραγματικά η βέλτιστη επιλογή για τον αριθμό δέντρων του μοντέλου, με μικρές και εν πολλοίς τυχαίες αποκλίσεις. Αυτό βεβαίως ήταν αναμενόμενο, αφού στο συγκεκριμένο μοντέλο εξετάζουμε, ως σημαντική υπερπαραμέτρο, μόνο τον αριθμό των χρησιμοποιούμενων δέντρων, οπότε η αναζήτηση πλέγματος ακολουθεί την ίδια διαδικασία με την απλή δοκιμή και αποτύπωση σε διάγραμμα, διαφορετικών παραμέτρων.

4.1.5 Αποτελέσματα Νευρωνικού Δικτύου Πολλαπλών Στρωμάτων

Για να βελτιστοποιήσουμε την επίδοση του μοντέλου νευρωνικού δικτύου πολλαπλών στρωμάτων πειραματιστήκαμε με τις τιμές των τριών σημαντικότερων υπερπαραμέτρων: του αριθμού των κρυφών στρωμάτων, του αριθμού των νευρώνων σε κάθε κρυφό στρώμα και του ρυθμού μάθησης. Πρέπει να τονιστεί εδώ ότι δεν υπάρχει κάποιος κανόνας για την δομή του νευρωνικού δικτύου που πρέπει να κατασκευάσουμε. Ο αριθμός των κρυφών στρωμάτων και των νευρώνων σε αυτά προκύπτει από δοκιμές, είτε χρησιμοποιώντας «καλές» δομές, οι οποίες έχουν δοκιμαστεί με επιτυχία σε παρόμοια προβλήματα. Αφού πειραματιστήκαμε με διαφορετικές δομές για το νευρωνικό μας δίκτυο, με την χρήση της αναζήτησης πλέγματος, η οποία μας επέτρεψε να ελέγξουμε έναν ευρύ συνδυασμό παραμέτρων και λάβαμε υπόψιν και τον χρόνο εκπαίδευσης, ο οποίος αυξάνεται με την αύξηση του αριθμού των επιπέδων και νευρώνων, καταλήξαμε σε ένα νευρωνικό δίκτυο με δύο κρυφά στρώματα με 180 και 100 νευρώνες σε κάθε στρώμα, αντίστοιχα. Ο αριθμός των νευρώνων σε κάθε στρώμα προέκυψε από την σύγκλιση των αποτελεσμάτων της αναζήτησης πλέγματος σε μια μικρή περιοχή με τις καλύτερες επιδόσεις. Ο συγκεκριμένος αριθμός των νευρώνων για τον οποίον παίρνουμε τα καλύτερα αποτελέσματα ποικίλλει, σε διαφορετικούς κύκλους εκπαίδευσης, ενδεχομένως εξαιτίας της διαφορετικής αρχικοποίησης των παραμέτρων – βαρών, ωστόσο οι διαφορές αυτές είναι ελάχιστες. Επομένως, ως αριθμός νευρώνων σε κάθε στρώμα επιλέχθηκε μια συχνά εμφανιζόμενη τιμή σε μια μικρή περιοχή βέλτιστων αποτελεσμάτων. Επιπλέον, ο αριθμός των στρωμάτων που επιλέχθηκε, ακολούθησε αντίστοιχα την λογική της αναζήτησης πλέγματος, για την επιλογή του, ωστόσο, δεν επιλέχθηκε εντέλει ο βέλτιστος αριθμός στρωμάτων. Φαίνεται, ότι η αύξηση του αριθμού στρωμάτων, με την επιλογή του κατάλληλου αριθμού νευρώνων για το νέο στρώμα (πχ. για τρίτο στρώμα, 80 νευρώνες) βελτιώνει πράγματι το ΜΤΣ, ωστόσο σε πολύ μικρό βαθμό, της τάξης του 0.0001 για το τρίτο

στρώμα, ενώ παράλληλα αυξάνει σημαντικά τον χρόνο εκπαίδευσης και την πολυπλοκότητα του νευρωνικού δικτύου. Περαιτέρω αύξηση του αριθμού των στρωμάτων οδηγεί σε επιπλέον βελτίωση του ΜΤΣ (πχ. βελτίωση κατά 0.0016 για 6 στρώματα), αυξάνοντας ωστόσο σημαντικά την πολυπλοκότητα και τον χρόνο εκπαίδευσης του δικτύου. Ακόμα, η περαιτέρω αύξηση του αριθμού των στρωμάτων (πάνω από τα 6 στρώματα) δεν οδηγεί σε παραπάνω βελτίωση, καθώς δεν υπάρχει επαρκής αριθμός δεδομένων για την εκπαίδευση τόσων πολλών παραμέτρων ή εναλλακτικά, όλη η πληροφορία που περιέχουν τα δεδομένα εκπαίδευσης έχει ήδη αποτυπωθεί σε λιγότερες παραμέτρους. Επιπλέον, από την αναζήτηση πλέγματος προέκυψε ως βέλτιστη τιμή του ρυθμού μάθησης 0.01 για δοκιμή όλων των τιμών από 0.00001 έως 0.1 στην λογαριθμική κλίμακα.

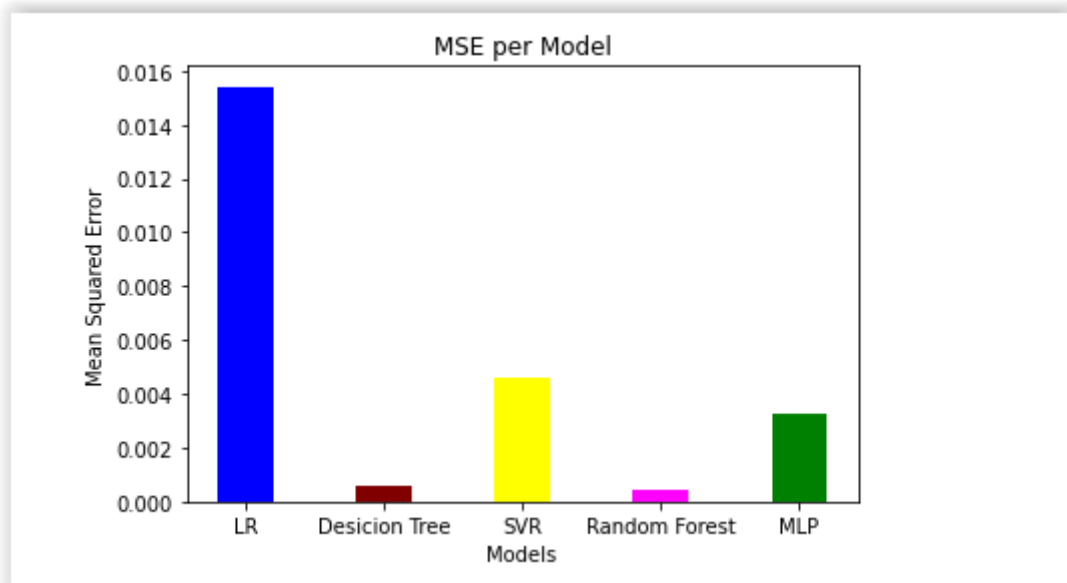
Η τιμή του ΜΤΣ που επιτυγχάνεται είναι:

$$MTS = 0.0031$$

Επιβεβαιώσαμε τα αποτελέσματα μας και αποκλείσαμε τον κίνδυνο υπερπροσαρμογής του μοντέλου με την χρήση της διασταυρωμένης επικύρωσης.

4.1.6 Σύγκριση Αποτελεσμάτων Επιβλεπόμενης Μάθησης

Στην Εικόνα 20 παρουσιάζονται συνοπτικά τα ΜΤΣ που πετυχαίνουν τα βέλτιστα μοντέλα επιβλεπόμενης μάθησης που κατασκευάσαμε. Παρατηρούμε ότι την χειρότερη απόδοση έχει η γραμμική παλινδρόμηση, πράγμα αναμενόμενο καθώς είναι ένα πολύ απλό μοντέλο που συνήθως δεν καταφέρνει να συλλάβει και να αποδώσει όλη την πολυπλοκότητα του προβλήματος. Τα καλύτερα αποτελέσματα, ωστόσο, τα δίνει το τυχαίο δάσος, πράγμα που αποτελεί έκπληξη δεδομένου ότι αναμέναμε ότι θα τα έδινε το νευρωνικό δίκτυο, εξαιτίας της ικανότητας του να αποτυπώνει πολυπλοκότερους συσχετισμούς.



Εικόνα 20: Αξιολόγηση των μοντέλων που χρησιμοποιήσαμε με βάση το ΜΤΣ (MSE). Τα μοντέλα που συγκρίνουμε είναι η γραμμική παλινδρόμηση (LR), το δέντρο απόφασης (Decision Tree), η παλινδρόμηση διανυσμάτων υποστήριξης (SVR Model), το τυχαίο δάσος (Random Forest) και το νευρωνικό δίκτυο πολλαπλών στρωμάτων (MLP).

Μια πιθανή εξήγηση για αυτό θα ήταν η υπερπροσαρμογή του νευρωνικού δικτύου. Ωστόσο, κάτι τέτοιο δεν φαίνεται να συμβαίνει, δεδομένου ότι, μετά από τον έλεγχο των αποτελεσμάτων μας με διασταυρωμένη επικύρωση, προκύπτει ότι το τυχαίο δάσος έχει καλύτερες επιδόσεις και στο σύνολο εκπαίδευσης σε σχέση με το νευρωνικό δίκτυο και επιπλέον, για το νευρωνικό δίκτυο υπάρχει μικρή μόνο απόκλιση στο ΜΤΣ μεταξύ του συνόλου εκπαίδευσης και του συνόλου ελέγχου. Μια άλλη εξήγηση θα μπορούσε να οφείλεται στην επιλογή μη βέλτιστης δομής για το νευρωνικό δίκτυο. Όμως, πρέπει να τονιστεί ότι δεν υπάρχει κάποιος κανόνας που να εγγυάται ποιο είναι το καλύτερο μοντέλο για κάθε πρόβλημα. Το «καλύτερο» μοντέλο είναι εντελώς σχετικό και εξαρτάται εν πολλοίς από τα χαρακτηριστικά του προβλήματος, για αυτό και για κάθε πρόβλημα δοκιμάζονται πολλές τεχνικές, ώστε να προκύψει η καλύτερη. Μπορεί, επομένως πράγματι, το βελτιστοποιημένο μοντέλο τυχαίου δάσους να είναι το καταλληλότερο για την πρόβλεψη των εξόδων αυτού του προβλήματος. Σε κάθε περίπτωση, οι αλγόριθμοι επιβλεπόμενης μάθησης που εφαρμόσαμε πετυχαίνουν εξαιρετικά καλές προβλέψεις για το πρόβλημά μας.

4.2 Μοντέλα Συλλογικής Μάθησης

Παρά τα πολύ καλά αποτελέσματα που παίρνουμε ήδη από την εφαρμογή των απλών αλγορίθμων επιβλεπόμενης μάθησης, αποφασίσαμε να χρησιμοποιήσουμε και μοντέλα συλλογικής μάθησης, σε μια προσπάθεια να βελτιώσουμε περαιτέρω την ακρίβεια των προβλέψεών μας και να διερευνήσουμε και άλλες τεχνικές στο πεδίο της επιβλεπόμενης μάθησης. Τα μοντέλα που χρησιμοποιήσαμε είναι η

παλινδρόμηση ψηφίσματος, το σακούλιασμα, οι τεχνικές ενδυνάμωσης AdaBoost, διαβαθμισμένη ενδυνάμωση και στοίβαξη.

Σε κάθε ένα από αυτά τα μοντέλα, χρησιμοποιήσαμε την τεχνική της διασταυρωμένης επικύρωσης, για να ελέγξουμε την αξιοπιστία των αποτελεσμάτων μας, σε σχέση με το ΜΤΣ, και για να αποκλείσουμε το ενδεχόμενο της υπερπροσαρμογής. Πράγματι, επιβεβαιώσαμε τα αποτελέσματα μας και πουθενά δεν ανέκυψε πρόβλημα υπερπροσαρμογής.

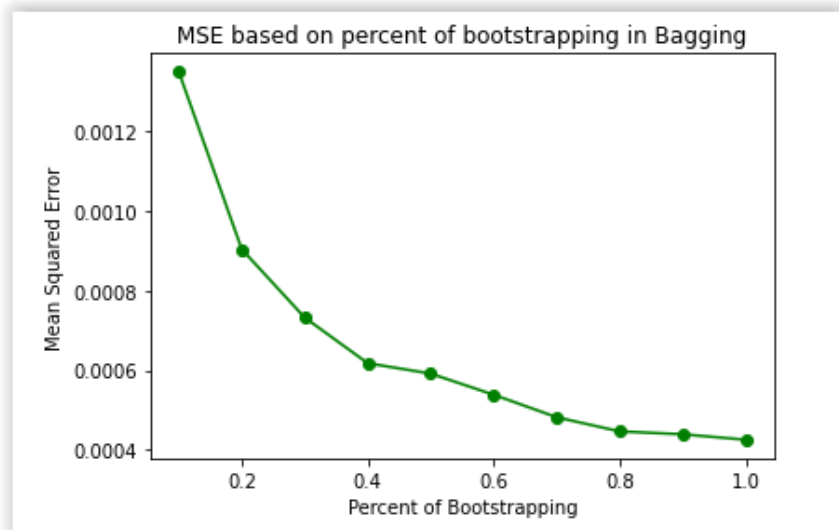
4.2.1 Αποτελέσματα Παλινδρόμησης Ψηφίσματος

Τα μοντέλα βάσης που επιλέξαμε να χρησιμοποιήσουμε είναι τα τρία μοντέλα επιβλεπόμενης μάθησης που μας έδωσαν τα καλύτερα αποτελέσματα, δηλαδή το δέντρο απόφασης, το τυχαίο δάσος και το νευρωνικό δίκτυο. Τα μοντέλα αυτά εκπαιδεύονται πάνω στο σύνολο εκπαίδευσης και η τελική πρόβλεψη του μοντέλου παλινδρόμησης ψηφίσματος προκύπτει από τον μέσο όρο των επιμέρους προβλέψεων. Το ΜΤΣ που πήραμε μετά την εκπαίδευση του μοντέλου παλινδρόμησης ψηφίσματος είναι:

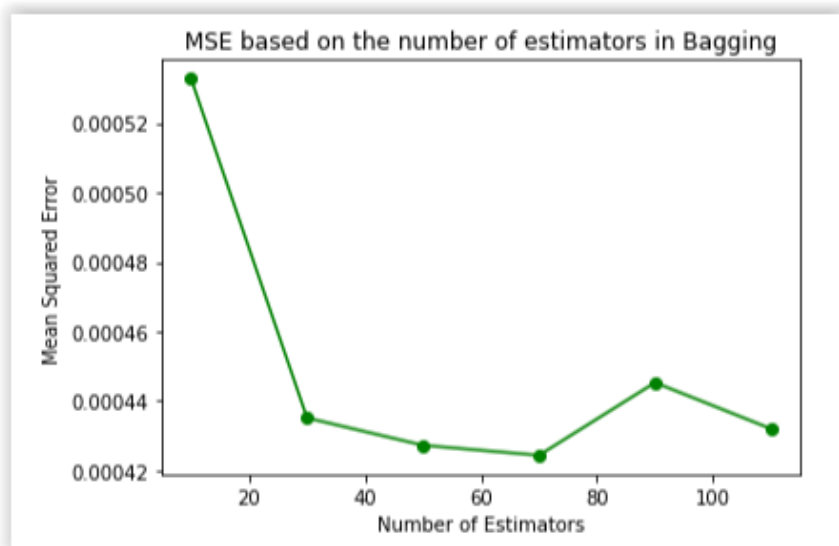
$$MT\sigma = 0.000386$$

4.2.2 Αποτελέσματα Σακουλιάσματος

Το βασικό μοντέλο που χρησιμοποιήσαμε είναι τα δένδρα απόφασης, το πλήθος των δειγμάτων που χρησιμοποιούνται για την εκπαίδευση του βασικού μοντέλου είναι ίσο με τον αριθμό των δειγμάτων του συνόλου εκπαίδευσης, ωστόσο, τα δείγματα επιλέγονται με αντικατάσταση, δηλαδή ένα δείγμα μπορεί να υπάρχει περισσότερες από μια φορές στο τελικό σύνολο εκπαίδευσης που θα τροφοδοτηθεί στο κάθε μοντέλο βάσης, ενώ άλλη μια σημαντική υπερπαραμέτρος είναι και ο αριθμός των μοντέλων βάσης που θα χρησιμοποιηθεί. Ως μοντέλο βάσης επιλέγουμε να κρατήσουμε το δέντρο απόφασης, το οποίο, όπως είδαμε και παραπάνω ανταποκρίνεται καλά στο πρόβλημα και εκπαιδεύεται γρήγορα. Στην συνέχεια ελέγχουμε κατά πόσο η εκπαίδευση του βασικού μοντέλου με αριθμό δειγμάτων ίσο με το πλήθος του συνόλου εκπαίδευσης είναι μια καλή επιλογή. Παρατηρούμε ότι, όταν επιλέγεται όλο το πλήθος των δειγμάτων του συνόλου εκπαίδευσης τότε επιτυγχάνονται οι χαμηλότερες τιμές ΜΤΣ. Ακόμη, θα ακολουθήσουμε μια διαδικασία ρύθμισης για να βρούμε τον βέλτιστο αριθμό δέντρων απόφασης για το συγκεκριμένο πρόβλημα. Βάση των αποτελεσμάτων που παρουσιάζονται στο διάγραμμα της Εικόνας 22, επιλέγουμε να χρησιμοποιήσουμε 70 δένδρα απόφασης.



Εικόνα 21: Το ΜΤΣ (MSE) του μοντέλου σακουλιάσματος (Bagging) ως προς το ποσοστό επί της πληθικότητας του συνόλου εκπαίδευσης, των δειγμάτων που επιλέγονται για την εκπαίδευση του μοντέλου βάσης (Bootstrapping).



Εικόνα 22: Το ΜΤΣ (MSE) του μοντέλου σακουλιάσματος (Bagging) ως προς τον αριθμό των μοντέλων βάσης (number of estimators).

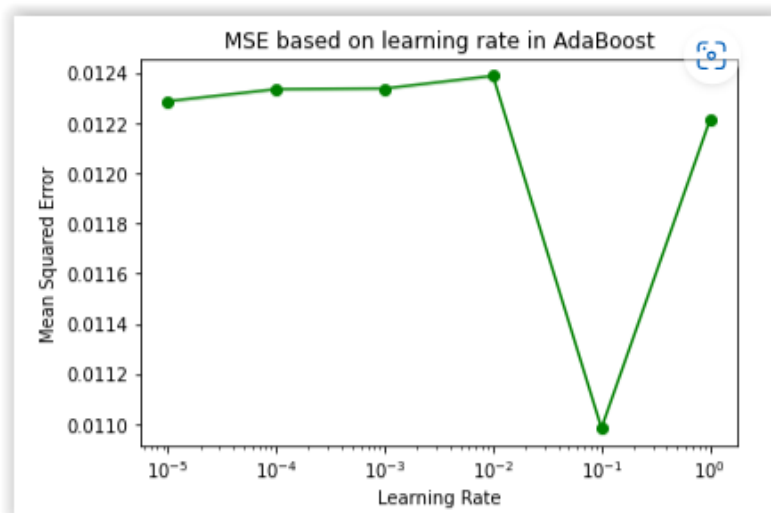
Το ΜΤΣ που προκύπτει τελικά είναι:

$$ΜΤΣ = 0.000414$$

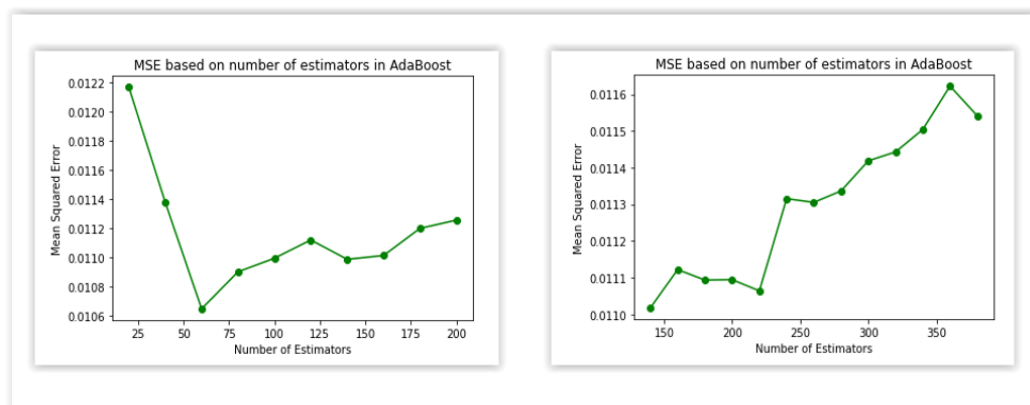
Να σημειώσουμε όμως ότι επειδή η τεχνική σακουλιάσματος εμπεριέχει μεγάλο βαθμό τυχαιότητας, τα αποτελέσματα σε κάθε διαφορετική επανάληψη του μοντέλου μπορεί να είναι ελαφρώς διαφορετικά.

4.2.3 Αποτελέσματα AdaBoost Τεχνικής

Η τεχνική AdaBoost διαθέτει πλήθος υπερπαραμέτρων , όπως το μοντέλο βάσης που θα χρησιμοποιηθεί , τον ρυθμό μάθησης με βάση τον οποίο ανανεώνονται τα βάρη και, πολύ σημαντικό, τον αριθμό των μοντέλων βάσης που θα χρησιμοποιηθούν. Το μοντέλο βάσης που χρησιμοποιεί η συγκεκριμένη τεχνική AdaBoost είναι δένδρο απόφασης με μέγιστο βάθος 3. Τα δένδρα απόφασης, όπως έχει αποδειχθεί, είναι μια καλή επιλογή μοντέλου βάσης για το πρόβλημα μας, ενώ αργότερα μπορεί να δοκιμαστεί και η χρήση απλών δένδρων απόφασης.



Εικόνα 23: Το ΜΤΣ (MSE) του μοντέλου AdaBoost ως προς τον ρυθμό μάθησης (learning rate).



Εικόνα 24: : Το ΜΤΣ (MSE) του μοντέλου AdaBoost ως προς τον αριθμό των μοντέλων βάσης (number of estimators). Παρατηρούμε ότι δεν διαχειριζόμαστε καλύτερα τα σφάλματα για μεγαλύτερο πλήθος μοντέλων βάσης.

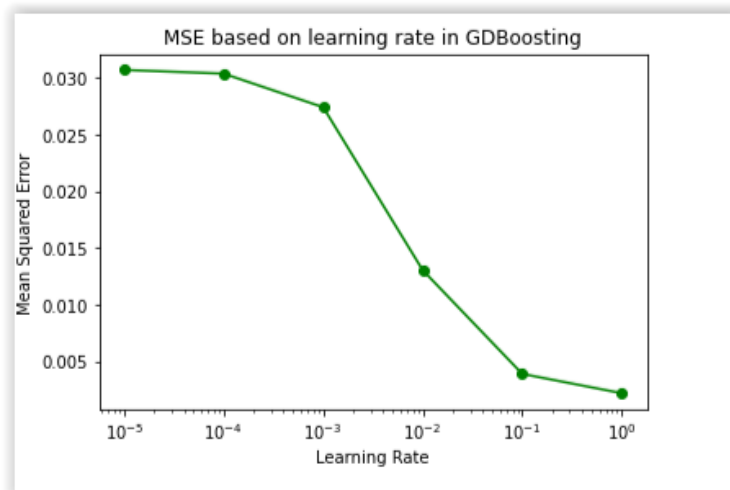
Κατά τα άλλα, κάναμε προσπάθεια να βρούμε τις βέλτιστες τιμές για τις υπερπαραμέτρους του ρυθμού μάθησης και του αριθμού των μοντέλων βάσης και οι προσπάθειες μας αποτυπώνονται στα διαγράμματα των Εικόνων 23 και 24. Βλέπουμε ότι, το καλύτερο ΜΤΣ προκύπτει για ρυθμό μάθησης 0.1 και για 60 δέντρα απόφασης. Το ΜΤΣ που προκύπτει είναι:

$$MΤΣ = 0.001659$$

Επιπλέον, για τις βέλτιστες υπερπαραμέτρους δοκιμάσαμε να χρησιμοποιήσουμε κολοβά δέντρα απόφασης, ωστόσο, πήραμε χειρότερα αποτελέσματα.

4.2.4 Αποτελέσματα Διαβαθμισμένης Ενδυνάμωσης

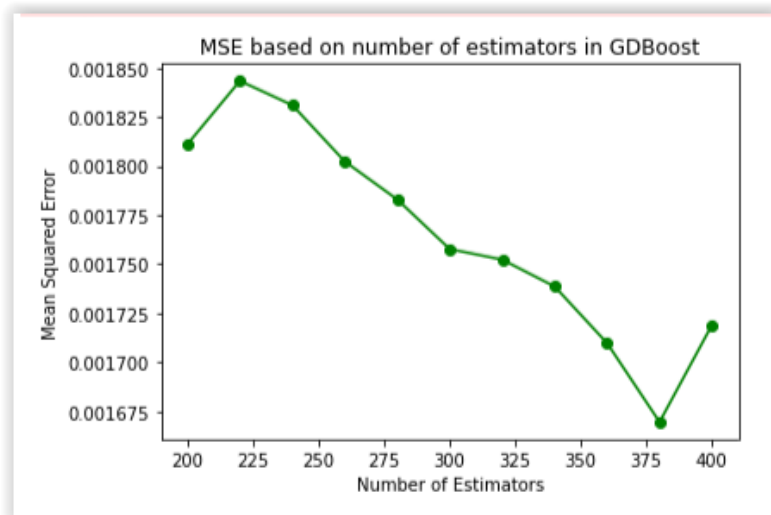
Το μοντέλο διαβαθμισμένης ενδυνάμωσης διαθέτει, όπως και το AdaBoost, πλήθος υπερπαραμέτρων, ωστόσο εμείς ασχοληθήκαμε με την βελτιστοποίηση του ρυθμού μάθησης που χρησιμοποιείται ως συντελεστής των προβλέψεων και του αριθμού των μοντέλων βάσης. Τα μοντέλα βάσης που χρησιμοποιούνται είναι δέντρα απόφασης μέγιστου βάθους 3 και όπως είδαμε δεν υπάρχει κάποια σημαντική βελτίωση που μπορούμε να επιτύχουμε αλλάζοντας τα μοντέλα βάσης. Στα διαγράμματα των Εικόνων 25 και 26 βλέπουμε πως επηρεάζεται το ΜΤΣ ως προς τον ρυθμό μάθησης και τον αριθμό των μοντέλων βάσης αντίστοιχα.



Εικόνα 25: Το ΜΤΣ (MSE) του μοντέλου διαβαθμισμένης ενδυνάμωσης (GDBosting) ως προς τον ρυθμό μάθησης (learning rate).

Ο αριθμός δέντρων απόφασης για το αρχικό μοντέλο είναι 100 και δοκιμάσαμε μόνο μεγαλύτερες τιμές. Καταλήγουμε να επιλέξουμε ρυθμό μάθησης 1 και 380 δέντρα απόφασης. Το ΜΤΣ που προκύπτει για αυτά είναι:

$$MT\Sigma = 0.001658$$



Εικόνα 26: Το ΜΤΣ (MSE) του μοντέλου διαβαθμισμένης ενδυνάμωσης (GDBoost) ως προς τον αριθμό των μοντέλων βάσης (number of estimators).

4.2.5 Αποτελέσματα Στοίβαξης

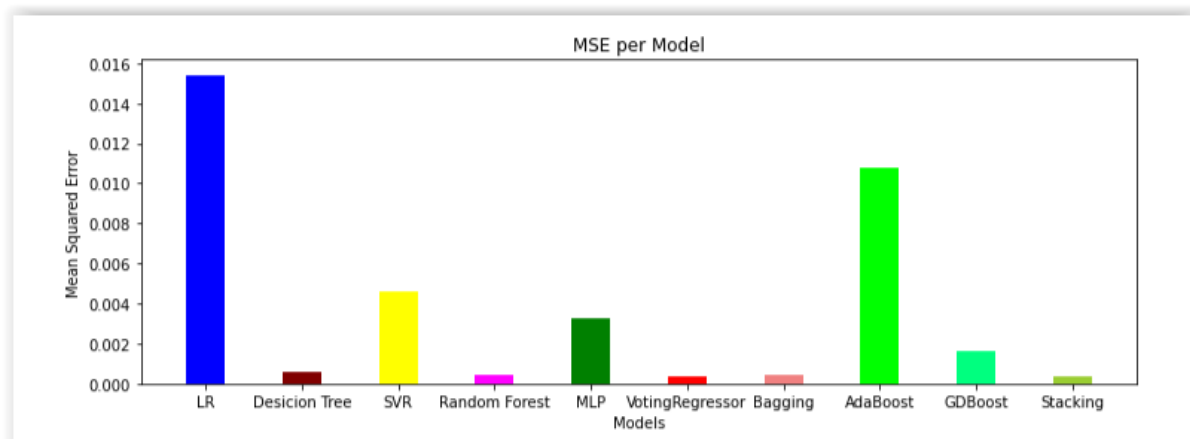
Ως βασικά μοντέλα στοίβαξης επιλέχθηκαν τα τρία καλύτερα μοντέλα της επιβλεπόμενης μάθησης, δηλαδή το δέντρο απόφασης, το τυχαίο δάσος και το νευρωνικό δίκτυο. Για τον μετά-εκτιμητή δοκιμάστηκαν διάφορες επιλογές, συγκεκριμένα γραμμική παλινδρόμηση, διανύσματα υποστήριξης και δέντρα απόφασης. Ωστόσο, το καλύτερο ΜΤΣ το πήραμε για τον την παλινδρόμηση Ridge. Πρόκειται για ένα είδος γραμμικής παλινδρόμησης που χρησιμοποιεί κάποια παράμετρο κανονικοποίησης στην συνάρτηση σφάλματος για να πετύχει καλύτερα αποτελέσματα. Συγκεκριμένα το ΜΤΣ που πήραμε για αυτό το μοντέλο είναι:

$$ΜΤΣ = 0.00039$$

Ο τρόπος με τον οποίο δημιουργήθηκε το σύνολο επικύρωσης, από το οποίο προέκυψαν οι προβλέψεις που τροφοδοτήθηκαν στον μετά-εκτιμητή ακολουθεί την τεχνική της διασταυρωμένης επικύρωσης.

4.2.6 Σύγκριση Αποτελεσμάτων

Τα καλύτερα αποτελέσματα τα παίρνουμε κατά σειρά για παλινδρόμηση ψηφίσματος, στοίβαξη, τυχαίο δάσος και σακούλιασμα με ελάχιστη απόκλιση.



Εικόνα 27: Το ΜΤΣ (MSE) ανά μοντέλο, όπου συγκρίνονται τα μοντέλα γραμμικής παλινδρόμησης (LR), δέντρου απόφασης (Decision Tree), παλινδρόμησης διανυσμάτων υποστήριξης (SVR), τυχαίου δάσους (Random Forest), νευρωνικού δικτύου πολλαπλών στρωμάτων (MLP), παλινδρόμησης ψηφίσματος (Voting Regressor), σακουλιάσματος (Bagging), AdaBoost, διαβαθμισμένης ενίσχυσης (GDBoost) και στοίβαξης (Stacking).

Αξίζει να σημειώσουμε εδώ ότι στην πραγματικότητα και το τυχαίο δάσος είναι μια τεχνική συλλογικής μάθησης και συγκεκριμένα μια παραλλαγή της τεχνικής σακουλιάσματος. Επομένως, είναι αναμενόμενο οι δύο αυτές τεχνικές να μας δίνουν καλά αποτελέσματα. Επιπλέον, οι τεχνικές παλινδρόμησης ψηφίσματος και στοίβαξης χρησιμοποιούν το μοντέλο του τυχαίου δάσους και προχωρούν σε μια διόρθωση των αποτελεσμάτων του, μέσω των προβλέψεων κάποιων επίσης πολύ ικανών μοντέλων. Φαίνεται μάλιστα ότι πετυχαίνουμε ένα κάτω όριο για το ελάχιστο ΜΤΣ, καθώς η εφαρμογή πιο σύνθετων τεχνικών δεν φαίνεται να πετυχαίνει σημαντικά καλύτερα αποτελέσματα σε σχέση με αυτά του τυχαίου δάσους. Το γεγονός αυτό ωστόσο, δεν μας προβληματίζει γιατί η ακρίβεια των προβλέψεων μας είναι πάρα πολύ καλή. Όμως, παρατηρούμε ότι η απόδοση των τεχνικών ενίσχυσης δεν είναι πολύ καλή για το πρόβλημά μας. Όπως είπαμε, εξετάσαμε το ενδεχόμενο της υπερπροσαρμογής, ωστόσο η απόκλιση του ΜΤΣ μεταξύ συνόλου εκπαίδευσης και συνόλου ελέγχου δεν είναι μεγάλη. Οπότε, συμπεραίνουμε ότι κάποιες παράμετροι του προβλήματος ή του συνόλου δεδομένων δεν ευνοούν την εκπαίδευση αποτελεσματικών μοντέλων ενίσχυσης. Εξάλλου, η απόδοση κάθε τεχνικής μπορεί να διαφέρει σημαντικά, ανάλογα με τις ιδιαιτερότητες του κάθε προβλήματος.

Ακόμα, για να εξετάσουμε ότι η κατάταξη των καλύτερων μοντέλων δεν οφείλεται σε στατιστική τυχαιότητα, εφαρμόσαμε τον t -έλεγχο σε μοντέλα με κοντινές επιδόσεις, ώστε να ελέγξουμε κατά πόσο έχουν στατιστικά σημαντικές διαφορές. Συγκεκριμένα, συγκρίναμε το μοντέλο τυχαίου δάσους με τα άλλα τρία καλύτερα μοντέλα, την παλινδρόμηση ψηφίσματος, την στοίβαξη και το σακούλιασμα αλλά και με το δέντρο απόφασης. Από τον t -έλεγχο αυτών των ζευγών, μόνο η P_{value} για τον συνδυασμό τυχαίου δάσους και στοίβαξης ξεπερνούσε το επίπεδο σημαντικότητας $\alpha = 0.05$ και επομένως τα μοντέλα αυτά κρίνονται ότι δεν έχουν σημαντικά στατιστικές διαφορές και επομένως, μπορεί να χρησιμοποιείται είτε το ένα είτε το άλλο χωρίς να κρίνεται κάποιο ως καλύτερο. Επίσης, το νευρωνικό δίκτυο πολλαπλών στρωμάτων συγκρίθηκε με τα κοντινά σε επιδόσεις μοντέλα,

παλινδρόμησης διανυσμάτων υποστήριξης και διαβαθμισμένης ενίσχυσης και συμπεράναμε ότι τα μοντέλα αυτά έχουν στατιστικά σημαντικές διαφορές επιδόσεων. Τέλος, παρότι δεν είναι εμφανές από το διάγραμμα προέκυψε ότι γραμμική παλινδρόμηση και η ενίσχυση AdaBoost δεν έχουν στατιστικά σημαντικές διαφορές και έτσι κατατάσσονται και τα δύο στα χειρότερα μοντέλα για το πρόβλημα μας.

4.3 Αποτελέσματα Αναδρομικών Νευρωνικών Δικτύων

Η επιτυχία των μέχρι τώρα προβλέψεων μας αφορούσε στο να δεχόμαστε σαν είσοδο τα χαρακτηριστικά του πλοίου και του καιρού που διαθέτουμε και να προβλέπουμε την τιμή του χαρακτηριστικού MEP για την τρέχουσα χρονική στιγμή. Ωστόσο, μεγάλη αξία θα είχε να μπορεί να γίνει πρόβλεψη της MEP και επομένως, της κατανάλωσης καυσίμου, σε βάθος κάποιου χρονικού ορίζοντα, ώστε να υπάρχει το περιθώριο πιθανών διορθώσεων ή αλλαγών. Επιπλέον, το σύνολο δεδομένων μας έχει την μορφή χρονοσειράς, εφόσον τα δείγματα για το ίδιο πλοίο καταγράφονται σε διάστημα 15 λεπτών και άρα ευνοεί την ανάπτυξη αναδρομικών μοντέλων όπως είναι τα LSTM μοντέλα.

Κατασκευάσαμε, λοιπόν, ένα μοντέλο LSTM για να προβλέψουμε την τιμή του χαρακτηριστικού MEP την χρονική στιγμή $t + 1$, λαμβάνοντας υπόψιν μας τις τιμές των πέντε χαρακτηριστικών εισόδου για τις χρονικές στιγμές $[t - 3, t]$ αλλά και τις τιμές του MEP για τις χρονικές στιγμές $[t - 3, t]$. Δεδομένου λοιπόν, ότι έχουμε μετρήσεις ανά 15 λεπτά, αυτό που κατορθώνουμε είναι, την τρέχουσα χρονική στιγμή και λαμβάνοντας υπόψιν μας προβλέψεις της τελευταίας μιας ώρας, να προβλέψουμε την τιμή MEP για 15 λεπτά αργότερα. Η δυνατότητα πρόβλεψης σε χρονικό ορίζοντα 15 λεπτών είναι ικανοποιητική, δεδομένου ότι στο διάστημα αυτό μπορούν να γίνουν αλλαγές οι οποίες να οδηγήσουν σε βελτιωμένη τιμή MEP.

Το πρώτο βήμα είναι να φορτώσουμε τα δεδομένα μας. Στα δεδομένα εισόδου προσθέσαμε και την στήλη MEP, την οποία όπως εξηγήσαμε θα χρησιμοποιήσουμε για να εξάγουμε προβλέψεις για μελλοντικές τιμές της MEP.

Σημαντικό κομμάτι της διαδικασίας αποτελεί η σωστή μορφοποίηση των δεδομένων προκειμένου να εισαχθούν στο LSTM μοντέλο. Δεδομένου ότι, για να εξάγουμε την τιμή MEP για την χρονική στιγμή $t + 1$ θα χρησιμοποιήσουμε τις τιμές των χαρακτηριστικών εισόδων και του MEP για τις χρονικές στιγμές από $[t - 3, t]$, θα πρέπει να δημιουργήσουμε τις στήλες αυτές, ως δεδομένα εισόδου. Επομένως πλέον, στα δεδομένα εισόδου μας θα έχουμε μια διάταξη με $4 \times 6 = 24$ στήλες, εφόσον σε κάθε χαρακτηριστικό θα έχουμε στήλες με τις τιμές του για $t - 3, t - 2, t - 1, t$, όπου κάθε μια από αυτές τις στήλες θα αποτελεί μετατόπιση κατά ένα της προηγούμενης. Επιπλέον η πρώτη τιμή MEP, για την οποία έχουμε επαρκή χαρακτηριστικά προκειμένου να την προβλέψουμε είναι η MEP(4) και για αυτό οι γραμμές των δεδομένων μας μειώνονται κατά τέσσερα. Προφανώς, ο αριθμός των

γραμμών μειώνεται με τον ίδιο τρόπο σε όλες τις στήλες αφού οι τελευταίες τέσσερις τιμές της στήλης $t - 3$ δεν θα είναι χρήσιμες για την πρόβλεψη, η πρώτη και οι τελευταίες τρεις της $t - 2$ κλπ. Με τον ίδιο τρόπο, μετατοπίζοντας κατά ένα την στήλη MEP παίρνουμε τις τιμές-στόχους. Πλέον, έχουμε όλα τα απαραίτητα δεδομένα που χρειάζονται να τροφοδοτηθούν στο LSTM μοντέλο για την πρόβλεψη της τιμής MEP μετά από 15 λεπτά.

Προτού προχωρήσουμε στην κατασκευή του μοντέλου, πρέπει να δημιουργήσουμε το σύνολο εκπαίδευσης και το σύνολο ελέγχου και να κανονικοποιήσουμε τα δεδομένα μας. Κρατάμε ως σύνολο εκπαίδευσης τα τελευταία 96 δείγματα του συνόλου δεδομένων, δηλαδή το τελευταίο 24ωρο. Τα δείγματα αυτά δεν θα συμπεριληφθούν στο σύνολο εκπαίδευσης, για να μην επηρεαστεί η αμεροληψία της διαδικασίας ελέγχου. Στην συνέχεια, προχωράμε σε κανονικοποίηση ελαχίστου-μεγίστου, προσέχοντας, όπως προείπαμε, να μην παρεισφρήσει πληροφορία του συνόλου ελέγχου στην κανονικοποίηση του συνόλου εκπαίδευσης.

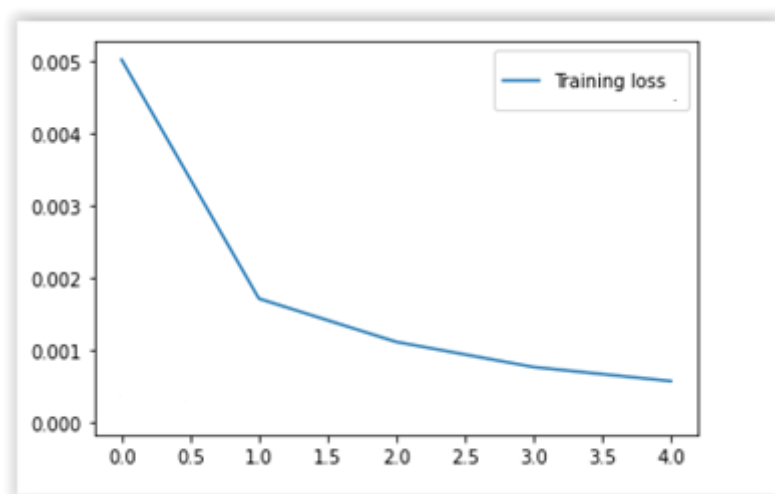
Το σχήμα των δεδομένων που απαιτεί το LSTM μοντέλο είναι [πλήθος δειγμάτων, πλήθος χρονικών βημάτων, πλήθος χαρακτηριστικών], οπότε στην δική μας περίπτωση είναι [πλήθος συνόλου δεδομένων - 4, 4, 6]. Αφού φέραμε τα δεδομένα στην κατάλληλη μορφή, πλέον χτίζουμε το μοντέλο μας. Για να εντοπίσουμε την βέλτιστη δομή μοντέλου χρησιμοποιήσαμε αναζήτηση πλέγματος, ώστε να καταλήξουμε, στον βέλτιστο ρυθμό μάθησης, πλήθος νευρώνων και κρυφών στρωμάτων και στο βέλτιστο ρυθμό dropout. Συγκεκριμένα :

- Αναζητήσαμε τον βέλτιστο αριθμό κρυφών στρωμάτων και νευρώνων και καταλήξαμε ότι η βέλτιστη επιλογή είναι ένα κρυφό στρώμα LSTM, 64 μονάδων, με την κάθε μονάδα περιλαμβάνει 4 LSTM κελιά (δεδομένου ότι έχουμε 4 χρονικά βήματα).
- Ως βέλτιστο ρυθμό μάθησης προσδιορίσαμε το 10^{-5} .
- Εξετάσαμε ως επιλογή την χρήση ενός Dropout στρώματος, το οποίο χρησιμοποιείται για την αποφυγή της υπερπροσαρμογής. Τελικά το ποσοστό dropout για το οποίο πήραμε τα βέλτιστα αποτελέσματα ήταν το 0% και επομένως το στρώμα Dropout δεν είναι απαραίτητο.

Επομένως, το μοντέλο που επιλέξαμε να χρησιμοποιήσουμε και το οποίο έχει πυρήνα του τις μονάδες LSTM τεσσάρων κυττάρων έχει, συνολικά την εξής μορφή: 1 κρυφό στρώμα LSTM με 64 μονάδες και το τελικό στρώμα, το οποίο προφανώς θα περιέχει μόνο έναν νευρώνα αφού το πρόβλημα μας είναι πρόβλημα παλινδρόμησης και προβλέπουμε μια και μόνη τιμή.

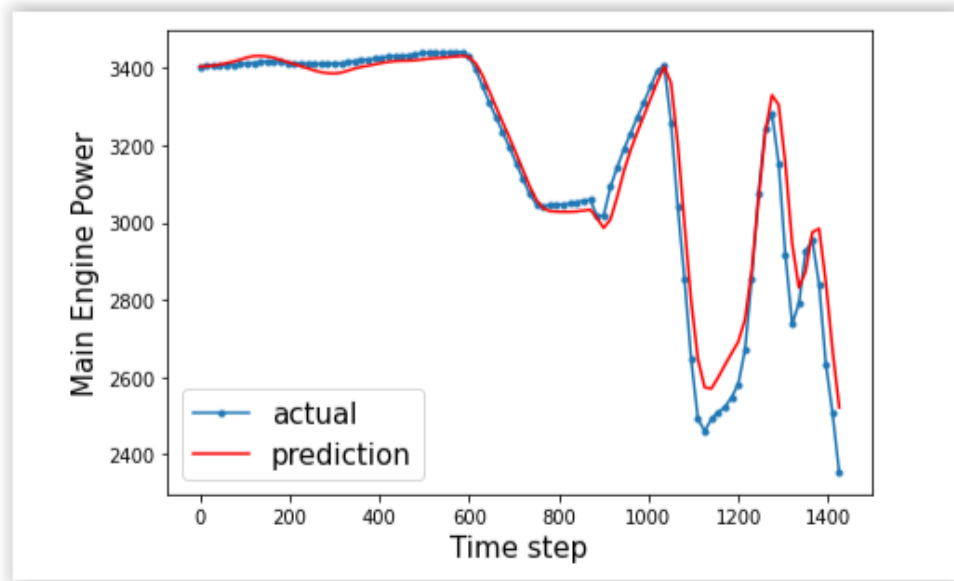
Η εκπαίδευση πραγματοποιήθηκε σε πέντε εποχές και με μέγεθος δέσμης (batch size) ίσο με 32. Το μέγεθος δέσμης καθορίζει πόσο συχνά θα γίνονται ανανεώσεις των βαρών από τον αλγόριθμο κατάβασης κλίσης. Αν είναι 32 τότε ανά 32 δείγματα θα ανανεώνονται τα βάρη του μοντέλου. Η εποχή, από την άλλη, εκφράζει ένα πλήρες πέρασμα όλων των δειγμάτων του συνόλου εκπαίδευσης. Κατά την προετοιμασία του μοντέλου, ορίστηκαν άλλες δύο παράμετροι, οι οποίες

επηρεάζουν την διαδικασία της εκπαίδευσης, ο βελτιστοποιητής και η συνάρτηση σφάλματος. Οι παράμετροι αυτοί καθορίζουν τον τρόπο με τον οποίο θα γίνει η ανανέωση των βαρών. Συγκεκριμένα, ο αλγόριθμος ο οποίος θα χρησιμοποιηθεί για την ανανέωση των βαρών είναι ο Adam, ο οποίος αποτελεί παραλλαγή του αλγορίθμου κατάβασης κλίσης, ενώ η μετρική, βάσει της οποίας υπολογίζεται η απόκλιση μεταξύ των πραγματικών και των προβλεπόμενων τιμών, κατά την εκπαίδευση, είναι το ΜΤΣ. Η απόκλιση αυτή, κατά την εκπαίδευση, διορθώνεται με την βοήθεια της τεχνικής της οπισθοδιάδοσης. Στην Εικόνα 28 παρουσιάζεται ένα διάγραμμα ενδεικτικό της πορείας της εκπαίδευσης. Από το διάγραμμα παρατηρούμε ότι η διαδικασία της εκπαίδευσης κυλάει ομαλά, δηλαδή όντως το μοντέλο εκπαιδεύεται με το πέρασμα των εποχών.



Εικόνα 28: Μείωση της απώλειας (εδώ ΜΤΣ) κατά την διάρκεια της εκπαίδευσης (training loss). Ο οριζόντιος άξονας αντιστοιχεί στον αριθμό των εποχών.

Μετά την ολοκλήρωση της εκπαίδευσης, θα πρέπει να αποτιμήσουμε την επιτυχία της, στην πρόβλεψη άγνωστων δεδομένων. Για αυτό τον σκοπό, σχεδιάζουμε της τελευταίες 96 πραγματικές τιμές του χαρακτηριστικού ΜΕΡ, σε αντιδιαστολή με τις 96 προβλεπόμενες τιμές, όπως αυτές προκύπτουν ως αποτέλεσμα του μοντέλου για το σύνολο ελέγχου, αφού εφαρμόσουμε τον αντίστροφο μετασχηματισμό της κανονικοποίησης ελαχίστου-μεγίστου. Για να αξιολογήσουμε την ομοιότητα των δύο καμπυλών που προκύπτουν χρησιμοποιούμε την μετρική goodness of fit. Για το μοντέλο μας πετυχαίνουμε $goodness\ of\ fit = 77\%$ και οι καμπύλες που προκύπτουν παρουσιάζονται στην Εικόνα 29.

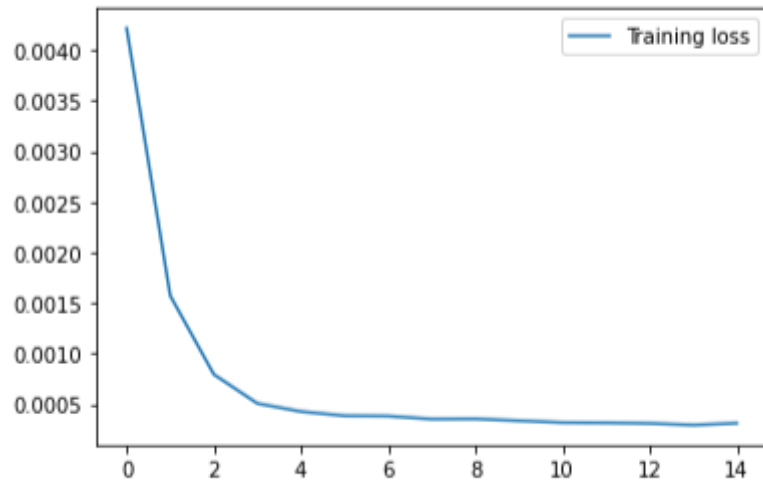


Εικόνα 29: Παρουσιάζονται οι πραγματικές και οι προβλεπόμενες τιμές του MEP σε χρόνο (time step) ανά 15 λεπτά. Η κόκκινη γραμμή αντιστοιχεί στις προβλεπόμενες τιμές και η μπλε στις πραγματικές.

Τα αποτελέσματα μας είναι πολύ ικανοποιητικά, ωστόσο έγιναν κάποιες προσπάθειες να επιτευχθούν καλύτερα αποτελέσματα, μέσω της ρύθμισης και της διόρθωσης των υπερπαραμέτρων της εκπαίδευσης του LSTM μοντέλου. Δοκιμάσαμε τα εξής :

- Αύξηση του αριθμού των εποχών :
Οι προβλέψεις μας σε αυτήν την περίπτωση θα μπορούσαν να παρουσιάσουν κάποια βελτίωση επειδή γίνεται καλύτερη προσαρμογή του μοντέλου στο σύνολο δεδομένων. Βεβαίως, αύξηση των εποχών ενέχει τον κίνδυνο της υπερπροσαρμογής. Ωστόσο παρατηρούμε ότι η αύξηση του πλήθους των εποχών δεν βελτιώνει τα αποτελέσματα μας και αυτό δικαιολογείται από το διάγραμμα της Εικόνας 30. Πράγματι, παρατηρούμε ότι το μοντέλο μας έχει προσαρμοστεί σχεδόν πλήρως στο σύνολο εκπαίδευσης, ήδη από τις πέντε εποχές και η αύξηση του πλήθους των εποχών οδηγεί σε μια απειροελάχιστη επιπλέον προσαρμογή, αυξάνοντας παράλληλα τον χρόνο εκπαίδευσης και τις πιθανότητες να σημειωθεί υπερπροσαρμογή. Έτσι, παραμένουμε στις πέντε εποχές.
- Αύξηση του χρονικού παραθύρου :
Θεωρητικά, μεγαλύτερο χρονικό παράθυρο θα μας έδινε καλύτερα αποτελέσματα, επειδή θα επέτρεπε την ανίχνευση συσχετίσεων σε μεγαλύτερο βάθος χρόνου. Δοκιμάσαμε παράθυρα εύρους 6 και 9 και πήραμε $goodness\ of\ fit = 79\%$ και $goodness\ of\ fit = 80\%$ αντίστοιχα. Ωστόσο αυτά τα αποτελέσματα επιτεύχθηκαν μόνο μετά την αύξηση των εποχών σε 10, διότι διαφορετικά τα σύνολα δεδομένων με μεγαλύτερα χρονικά παράθυρα είχαν χειρότερα αποτελέσματα. Αυτό μπορεί να οφειλόταν στο γεγονός ότι σε πέντε μόνο εποχές δεν ήταν εφικτό να αποτυπωθούν ικανοποιητικά οι πολυπλοκότερες συσχετίσεις. Ωστόσο,

παρατηρούμε ότι η βελτίωση με την αύξηση του χρονικού παραθύρου δεν είναι σημαντική και γενικότερα φαίνεται ότι το αρχικό μας μοντέλο έχει αποτυπώσει ήδη σε μεγάλο βαθμό τις συσχετίσεις που μπορούν να προκύψουν από το σύνολο δεδομένων.



Εικόνα 30: Απώλεια εκπαίδευσης (training loss) ως προς το πλήθος των εποχών.

Τέλος, δοκιμάσαμε το μοντέλο μας και στην πρόβλεψη επόμενων χρονικών στιγμών. Αναμέναμε ότι οι προβλέψεις μας θα ήταν χειρότερες σε βάθος χρόνου και πράγματι για την χρονική στιγμή $t+2$ πήραμε $goodness\ of\ fit = 54\%$, ενώ για την χρονική στιγμή $t+5$ πήραμε $goodness\ of\ fit = 9\%$ μόλις.

4.4 Αποτελέσματα Ομοσπονδιακής Μάθησης

Για να δοκιμάσουμε την ομοσπονδιακή μάθηση στην πράξη, χρησιμοποιήσαμε τα δεδομένα και των δύο πλοίων που είχαμε στην διάθεση μας και χτίσαμε μια αρχιτεκτονική, η οποία προσομοιώνει την δομή και τα βήματα του FedAvg αλγορίθμου. Τα σύνολα δεδομένων των δύο πλοίων είναι κοινά μορφοποιημένα και προεπεξεργασμένα, με τα δείγματα τους να έχουν τα ίδια χαρακτηριστικά, με την ίδια σειρά και έξοδο την τιμή του χαρακτηριστικού MEP. Αρχικά, χωρίζουμε κάθε ένα από τα σύνολα δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου (10% του συνολικού) και κατόπιν τα ζεύγη συνόλου εκπαίδευσης και συνόλου ελέγχου υφίστανται ξεχωριστή κανονικοποίηση ελαχίστου-μεγίστου το καθένα, προσέχοντας πάντα να μην υπάρξει διαρροή πληροφορίας από το σύνολο ελέγχου στο σύνολο εκπαίδευσης. Στην συνέχεια, χτίζουμε το μοντέλο που θα χρησιμοποιηθεί στην εκπαίδευση, τόσο από τον κεντρικό εξυπηρετητή όσο και από τους τοπικούς χρήστες. Αποφασίζουμε ότι το μοντέλο που θα χρησιμοποιήσουμε θα είναι ένα

νευρωνικό δίκτυο πολλαπλών στρωμάτων και συγκεκριμένα θα αποτελείται από δύο κρυφά στρώματα, με 180,100 νευρώνες αντίστοιχα και στρώμα εξόδου του ενός νευρώνα. Το δίκτυο αυτό έχει την ίδια δομή με αυτό που χρησιμοποιήσαμε στην επιβλεπόμενη μάθηση και το επιλέξαμε ακριβώς εξαιτίας των καλών του αποτελεσμάτων. Στην συνέχεια κατασκευάζουμε τρεις συναρτήσεις για να υλοποιήσουμε τον αλγόριθμο FedAvg :

- Η πρώτη υπολογίζει έναν παράγοντα διαβάθμισης, ο οποίος αποτυπώνει την συνεισφορά του εκάστοτε τοπικού χρήστη στην εκπαίδευση. Ο παράγοντας διαβάθμισης προκύπτει ως εξής :

$$\pi\delta = \frac{n_i}{\sum_{i=1}^2 n_i}$$

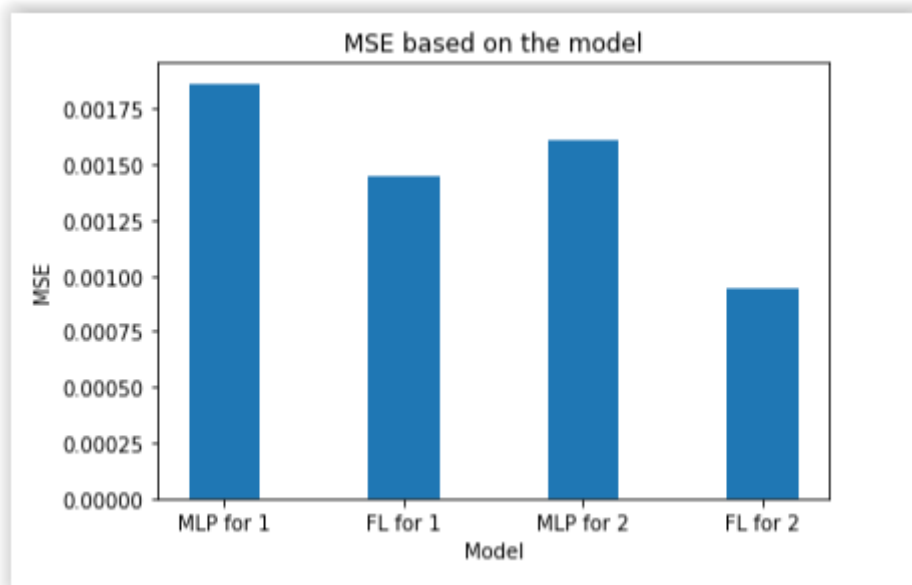
όπου, n_i είναι ο αριθμός των δειγμάτων του συνόλου εκπαίδευσης του τοπικού χρήστη i .

- Η δεύτερη εφαρμόζει τον παράγοντα διαβάθμισης του εκάστοτε χρήστη στα βάρη που προκύπτουν από την εκπαίδευση του τοπικού μοντέλου του χρήστη. Συγκεκριμένα πολλαπλασιάζει τον παράγοντα διαβάθμισης με κάθε βάρος.
- Η τρίτη συνάρτηση υλοποιεί τον συμψηφισμό των βαρών από τον κεντρικό εξυπηρετητή. Αφού ολοκληρωθεί η τοπική εκπαίδευση και των δύο μοντέλων η τρίτη συνάρτηση αθροίζει τα αντίστοιχα σταθμισμένα βάρη.

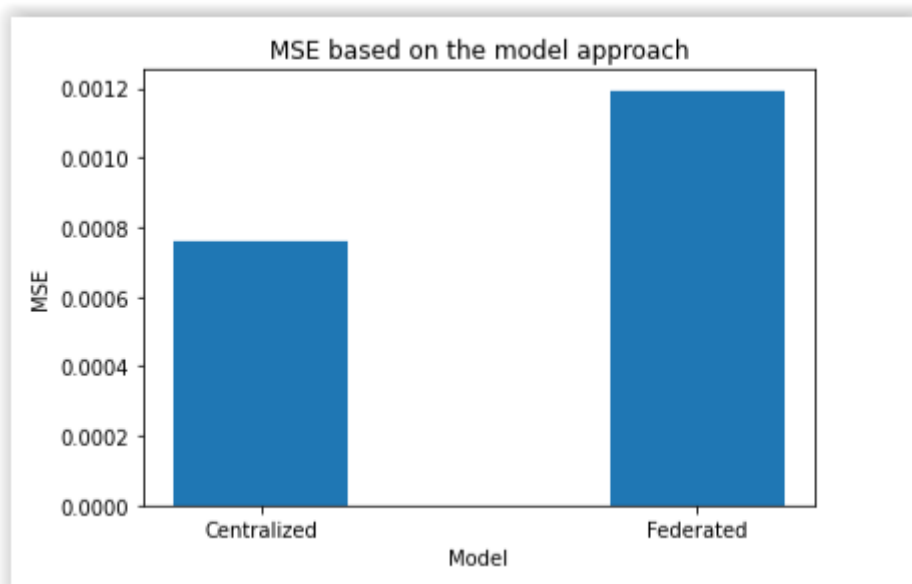
Εφόσον υλοποιήσαμε αυτές τις τρεις συναρτήσεις, ακολουθούμε τα βήματα του FedAvg αλγορίθμου. Αρχικά, ο κεντρικός εξυπηρετητής αρχικοποιεί τα βάρη του νευρωνικού δικτύου που ορίσαμε, για να υλοποιήσει την εκπαίδευση. Τα βάρη αυτά διαμοιράζονται στους τοπικούς χρήστες, οι οποίοι εκπαιδεύουν το νευρωνικό δίκτυο με τα κεντρικά βάρη που έλαβαν, πάνω στο δικό τους σύνολο εκπαίδευσης για δύο εποχές. Στην συνέχεια, εφαρμόζονται επί των νέων τοπικών βαρών οι τρεις συναρτήσεις που περιγράψαμε παραπάνω και από την τρίτη συνάρτηση προκύπτουν τα νέα κεντρικά βάρη, τα οποία θα διαμοιραστούν εκ νέου στους τοπικούς χρήστες. Η διαδικασία αυτή επαναλαμβάνεται για 40 γύρους.

Η επιλογή του αριθμού των τοπικών όσο και των κεντρικών γύρων εκπαίδευσης προέκυψε μετά από δοκιμές, σε μια προσπάθεια βελτιστοποίησης των αποτελεσμάτων της αρχιτεκτονικής. Στην Εικόνα 31 συγκρίνουμε τα αποτελέσματα που προκύπτουν αν η εκπαίδευση γίνει ξεχωριστά για τα αντίστοιχα σύνολα δεδομένων ή με διαμοιρασμό πληροφοριών μέσω της αρχιτεκτονικής που ακολουθεί τις αρχές της ομοσπονδιακής μάθησης.

Παρατηρούμε ότι η απόδοση της εκπαίδευσης των μοντέλων αυξάνεται όταν εφαρμόζονται οι αρχές της ομοσπονδιακής μάθησης και έτσι επιβεβαιώνουμε στην πράξη ότι η ανταλλαγή πληροφοριών μεταξύ διαφορετικών συνόλων δεδομένων ευνοεί την αποτελεσματικότερη εκπαίδευσή τους.



Εικόνα 31: Το ΜΤΣ (MSE) που προκύπτει αν η εκπαίδευση του μοντέλου για τα πλοία 1 και 2 γίνει ξεχωριστά (MLP for 1,2) ή με ομοσπονδιακή μάθηση (FL for 1,2).



Εικόνα 32: Το ΜΤΣ (MSE) βάση αν το μοντέλο εκπαίδευσης που χρησιμοποιούμε ακολουθεί κεντρική (Centralized) ή ομοσπονδιακή (Federated) προσέγγιση (approach).

Στην Εικόνα 32 παρουσιάζεται η απόδοση του FedAvg αλγορίθμου σε σύγκριση με ένα μοντέλο που συλλέγει τα δεδομένα και τον δύο πλοίων στον κεντρικό εξυπηρετητή και εφαρμόζει κεντρική εκπαίδευσης. Τα αποτελέσματα είναι χειρότερα για τον FedAvg αλγόριθμο, ωστόσο όχι σημαντικά χειρότερα, ώστε να μην θεωρούνται σημαντικά τα πλεονεκτήματα που εξασφαλίζουμε με την χρήση αλγορίθμων ομοσπονδιακής μάθησης, με κυριότερα από αυτά την ιδιωτικότητα και το διαμοιρασμό του κόστους εκπαίδευσης στους τοπικούς χρήστες.

5. Συμπεράσματα

5.1 Συνεισφορά Εργασίας

Στην παρούσα εργασία προχωρήσαμε σε μια ανασκόπηση των δημοφιλέστερων σύγχρονων τεχνικών μηχανικής μάθησης, οι οποίες εφαρμόστηκαν πάνω στα πραγματικά δεδομένα δύο πλοίων, ώστε να γίνει πρόβλεψη και εξαγωγή συμπερασμάτων για την ισχύ του κύριου κινητήρα και κατ' επέκταση, για την κατανάλωση καυσίμου. Καταφέραμε να επιτύχουμε ακριβείς προβλέψεις και να παρατηρούσαμε ενδιαφέροντα μοτίβα για την ισχύ κύριου κινητήρα, ενώ ταυτόχρονα εξετάσαμε, θεωρητικά και στην πράξη, μια ευρεία γκάμα αλγόριθμων επιβλεπόμενης, συλλογικής μάθησης καθώς και μοντέλα αναδρομικών νευρωνικών δικτύων και ομοσπονδιακής μάθησης. Έτσι, η εργασία αυτή αποτελεί μια σύνοψη των βασικών αρχών αυτών των μοντέλων, ενώ μελετάει την βελτιστοποίηση των υπερπαραμέτρων τους και την εφαρμογή τους στην αντιμετώπιση προβλημάτων του καίριου πεδίου της ναυτιλίας. Από την εφαρμογή των εξεταζόμενων αλγόριθμων στα πραγματικά δεδομένα των πλοίων, διαπιστώσαμε μια υπεροχή των μοντέλων τυχαίου δάσους στην πρόβλεψη της ισχύος κύριου κινητήρα, η οποία έγινε εντονότερη όταν αυτά συνδυάστηκαν, στα πλαίσια συλλογικής μάθησης, με άλλα ισχυρά μοντέλα σε δομή ψηφίσματος ή στοίβαξης. Τα αποτελέσματα αυτά καταδεικνύουν την ανάγκη διεξοδικής μελέτης και ελέγχου των διαθέσιμων τεχνικών στο εκάστοτε πρόβλημα, ώστε να εντοπιστεί αυτή που απεικονίζει πληρέστερα τις, εν γένει πολύπλοκες, συσχετίσεις που προκύπτουν εντός του συνόλου δεδομένων, αλλά και την ανάγκη βελτιστοποίησης αυτών των τεχνικών μέσω της δοκιμής διαφορετικών υπερπαραμέτρων. Τέλος, κατασκευάσαμε σύστημα μακράς βραχυχρόνιας μνήμης και το χρησιμοποιήσαμε αποτελεσματικά για την διεξαγωγή προβλέψεων σε βάθος χρόνου, ενώ κάναμε μια πρώτη επιτυχημένη απόπειρα να συνδυάσουμε, στα πλαίσια ομοσπονδιακής μάθησης, τα αποτελέσματα από την εκπαίδευση επιμέρους πλοίων, ώστε να επωφεληθούμε πλήρως από τις δυνατότητες διαμοιρασμού γνώσης, τηρώντας παράλληλα τα πρωτόκολλα προστασίας προσωπικών δεδομένων.

5.2 Μελλοντικές Προεκτάσεις

Η μελέτη που διεξήχθη στα πλαίσια της παρούσας εργασίας θα μπορούσε να επεκταθεί προς διάφορες κατευθύνσεις στο μέλλον. Συγκεκριμένα:

- Στο πεδίο της επιβλεπόμενης μάθησης, να χρησιμοποιηθούν πιο προηγμένες τεχνικές για τον ταυτόχρονο συντονισμό των υπερπαραμέτρων και να μελετηθούν κάποιοι ακόμα αλγόριθμοι συλλογικής μάθησης.

- Στο πεδίο των συστημάτων μακράς βραχυχρόνιας μνήμης να γίνει προσπάθεια βελτιώσεων, ώστε να επεκταθεί ο χρονικός ορίζοντας προβλέψεων.
- Στο πεδίο της ομοσπονδιακής μάθησης εντοπίζονται τα περισσότερα περιθώρια επέκτασης της παρούσας εργασίας. Στο μέλλον, μπορούν να γίνουν δοκιμές για την εφαρμογή του τρέχοντος αλγορίθμου ομοσπονδιακής μάθησης σε μεγάλης κλίμακας προβλήματα, όπου τα συνεργαζόμενα πλοία είναι πολύ περισσότερα, αλλά να δοκιμαστούν και διάφορες άλλες παραλλαγές αλγορίθμων ομοσπονδιακής μάθησης, οι οποίες μπορούν να διορθώσουν της αδυναμίες του απλού μοντέλου που κατασκευάστηκε εδώ.
- Όλες οι παραπάνω προσεγγίσεις μπορούν, τέλος, να εφαρμοστούν για την αντιμετώπιση άλλων προβλημάτων του τομέα της ναυτιλίας, όπως η προληπτική συντήρηση (predictive maintenance).

Αναφορές

- [1] Z. Zhang, C. Guan, H. Chen, X. Yang, W. Gong και A. Yang, «Adaptive Privacy-Preserving Federated Learning for Fault Diagnosis in Internet of Ships,» *IEEE Internet of Things Journal*, 2022.
- [2] A. King, «Emissions-free sailing is full steam ahead for ocean-going shipping,» *Horizon: The EU research and innovation magazine*, 06 09 2022.
- [3] T. Mitchell, *Machine Learning*, McGraw Hill, 1997.
- [4] «pandio.com,» 21 01 2021. [Ηλεκτρονικό]. Available: <https://pandio.com/when-was-machine-learning-invented/>. [Πρόσβαση 16 05 2023].
- [5] S. Russell και P. Norvig., *Artificial Intelligence: A Modern Approach*, 3η επιμ., Prentice Hall, 2010.
- [6] Wikipedia, «Wikipedia,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Linear_regression. [Πρόσβαση 18 05 2023].
- [7] M. Soni, «Medium,» 26 09 2020. [Ηλεκτρονικό]. Available: <https://maniksonituts.medium.com/what-is-decision-tree-regression-dcd0ea40a323>. [Πρόσβαση 18 05 2023].
- [8] D. A. Marvasti, «Towards Data Science,» 10 05 2020. [Ηλεκτρονικό]. Available: <https://towardsdatascience.com/support-vector-classifiers-and-logistic-regression-similarity-97ff06aa6ec3>. [Πρόσβαση 22 05 2023].
- [9] S. Saxena, «Analytics Vidhya,» 08 03 2021. [Ηλεκτρονικό]. Available: <https://www.analyticsvidhya.com/blog/2021/03/beginners-guide-to-support-vector-machine-svm/>. [Πρόσβαση 22 05 2023].
- [10] A. Sethi, «Analytics Vidhya,» 27 03 2020. [Ηλεκτρονικό]. Available: <https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/>. [Πρόσβαση 22 05 2023].
- [11] TIBCO, «TIBCO,» [Ηλεκτρονικό]. Available: <https://www.tibco.com/reference-center/what-is-a-random-forest>. [Πρόσβαση 23 05 2023].
- [12] Wikipedia, «Wikipedia,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Multilayer_perceptron#Terminology. [Πρόσβαση 23 05 2023].
- [13] Javatpoint, «Javatpoint,» [Ηλεκτρονικό]. Available:

- <https://www.javatpoint.com/perceptron-in-machine-learning>. [Πρόσβαση 23 05 2023].
- [14] Javatpoint, «Javatpoint,» [Ηλεκτρονικό]. Available: <https://www.javatpoint.com/multi-layer-perceptron-in-tensorflow>. [Πρόσβαση 23 05 2023].
- [15] J. Brownlee, «Machine Learning Mastery,» 18 03 2016. [Ηλεκτρονικό]. Available: <https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/>. [Πρόσβαση 23 05 2023].
- [16] R. E. Schapire, «Explaining AdaBoost,» σε *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, Berlin, Heidelberg, Springer Berlin Heidelberg, 2013, pp. 37-52.
- [17] C. Shetty, «Towards Data Science,» 30 12 2020. [Ηλεκτρονικό]. Available: <https://towardsdatascience.com/multiple-model-creation-using-adaboost-technique-2119b7aacf25>. [Πρόσβαση 25 05 2023].
- [18] F. Liu, «Linked In,» 25 03 2016. [Ηλεκτρονικό]. Available: <https://www.linkedin.com/pulse/what-adaboost-all-fuyang-liu>. [Πρόσβαση 24 05 2023].
- [19] T. Masui, «Towards Data Science,» 20 01 2022. [Ηλεκτρονικό]. Available: <https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502>. [Πρόσβαση 25 05 2023].
- [20] W. Mao-Xin, H. Duruo, W. Gang και L. Dian-Qing, «SS-XGBoost: A Machine Learning Framework for Predicting Newmark Sliding Displacements of Slopes,» *Journal of Geotechnical and Geoenvironmental Engineering*, 01 09 2020.
- [21] D. H. Wolpert, «Stacked generalization,» *Neural Networks*, pp. 241-259, 1992.
- [22] I. J. Goodfellow, Y. Bengio και A. Courville, *Deep Learning*, Cambridge, MA, USA: MIT Press, Deep Learning.
- [23] G. Singhal, «PLURALSIGHT,» 09 09 2020. [Ηλεκτρονικό]. Available: <https://www.pluralsight.com/guides/introduction-to-lstm-units-in-rnn>. [Πρόσβαση 2023 5 25].
- [24] C. Olah, «colah's blog,» 27 08 2015. [Ηλεκτρονικό]. Available: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>. [Πρόσβαση 2023 5 25].
- [25] M. Kuhn και K. Johnson, *Applied Predictive Modeling*, 2013.
- [26] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak και F. Herrera, «A survey on

data preprocessing for data stream mining: Current status and future directions,» *Neurocomputing*, pp. 39-57, 2017.

- [27] S. Maddali, «Towards Data Science,» 12 05 2022. [Ηλεκτρονικό]. Available: <https://towardsdatascience.com/what-are-the-most-important-preprocessing-steps-in-machine-learning-and-data-science-a7606d18f32a>. [Πρόσβαση 21 07 2023].
- [28] scikit-learn, «scikit-learn,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-error. [Πρόσβαση 07 23 2023].
- [29] MathWorks, «MathWorks,» [Ηλεκτρονικό]. Available: <https://www.mathworks.com/help/ident/ref/goodnessoffit.html>. [Πρόσβαση 28 07 2023].
- [30] IBM, «research.ibm.com,» [Ηλεκτρονικό]. Available: <https://research.ibm.com/blog/what-is-federated-learning>. [Πρόσβαση 23 07 2023].
- [31] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li και Y. Gao, «A survey on federated learning,» *Knowledge-Based Systems*, 15 03 2021.
- [32] S. Wang και Y. Zhang, «Multi-Level Federated Network Based on Interpretable Indicators for Ship Rolling Bearing Fault Diagnosis,» *Journal of Marine Science and Engineering*, 2022.
- [33] S. Tijani, «Towards Data Science,» 10 04 2020. [Ηλεκτρονικό]. Available: <https://towardsdatascience.com/federated-learning-a-step-by-step-implementation-in-tensorflow-aac568283399>. [Πρόσβαση 24 07 2023].
- [34] scikit-learn, «scikit-learn,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/grid_search.html. [Πρόσβαση 12 09 2023].
- [35] scikit-learn, «scikit-learn,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. [Πρόσβαση 12 09 2023].
- [36] Student, «The Probable Error of a Mean,» *Biometrika*, pp. 1-25, 1908.
- [37] mlxtend, «mlxtend,» Github, [Ηλεκτρονικό]. Available: https://rasbt.github.io/mlxtend/user_guide/evaluate/paired_ttest_kfold_cv/. [Πρόσβαση 20 09 2023].
- [38] J. Kiani, «Medium,» 26 09 2019. [Ηλεκτρονικό]. Available: <https://medium.com/analytics-vidhya/using-the-corrected-paired-students-t-test-for-comparing-the-performance-of-machine-learning-dc6529eaa97f>. [Πρόσβαση

20 09 2023].

- [39] C. Nadeau και Y. Bengio, «Inference for the Generalization Error,» *Machine Learning*, pp. 239-281, 01 09 2003.
- [40] A. Giannopoulos, P. Gkonis, P. Bithas, N. Nomikos, G. Ntroulias και P. Trakadas, «Federated Learning for Maritime Environments:Use Cases, Experimental Results, and Open Issues,» *TechRxiv*, 2023.
- [41] MariEMS, «Trim optimisation, Hull and propeller condition».
- [42] United States Naval Academy, «RESISTANCE AND POWERING OF SHIPS,» σε *EN400: Principles of Ship Performance*.

Συντομογραφίες

LSTM	Long Short-Term Memory
PCA	Principal Component Analysis
EM	Expectation-Maximization
FedAvg	Federated Averaging
SOG	Speed Over Ground
STW	Speed Through Water
DWS	Discretized Wind Speed
WD	Wind Direction
DF	Draft Forward
DA	Draft Aft
MEP	Main Engine Power
ΜΤΣ	Μέσο Τετραγωνικό Σφάλμα
ΜΑΣ	Μέσο Απόλυτο Σφάλμα
ΜΑΠΣ	Μέσο Απόλυτο Ποσοστιαίο Σφάλμα