ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	Μηχανική Μάθηση στην Πρόβλεψη της τιμής ενοικίασης Airbnb στο Άμστερνταμ Machine Learning Prediction of Amsterdam Airbnb Prices
Ονοματεπώνυμο Φοιτητή	Γεώργιος Σερβετάς
Πατρώνυμο	Δημήτριος
Αριθμός Μητρώου	ΜΠΠΛ/19049
Επιβλέπων	Διονύσιος Σωτηρόπουλος, Επίκουρος Καθηγητής

Ημερομηνία Παράδοσης Οκτώβριος 2023

Τριμελής Εξεταστική Επιτροπή

Διονύσιος Σωτηρόπουλος (Επίκ. Καθηγητής)	Γεώργιος Τσιχριντζής (Καθηγητής)	Ευάγγελος Σακκόπουλος (Αναπλ. Καθηγητής)

Περίληψη (Abstract)

Σε αυτή τη διατριβή, διερευνάται η χρήση και σύγκριση μοντέλων μηχανικής μάθησης για την πρόβλεψη της τιμής μιας καταχώρισης Airbnb. Πιο συγκεκριμένα, συλλέχτηκε ένα σύνολο δεδομένων από καταχωρίσεις της Airbnb στο Amsterdam με πλήθος χαρακτηριστικών, όπως ο τύπος της ιδιοκτησίας, ο αριθμός των υπνοδωματίων, η γειτονιά κτλ. Στην συνέχεια εφαρμόστηκε καθαρισμός και προεπεξεργασία των δεδομένων για να τροφοδοτηθούν στα μοντέλα μηχανικής εκμάθησης. Πολλά μοντέλα μηχανικής μάθησης, συμπεριλαμβανομένου του δέντρου παλινδρόμησης, του τυχαίου δάσους και της μηχανής διανυσματικής υποστήριξης εκπαιδεύτηκαν και δοκιμάστηκαν στο σύνολο δεδομένων. Τα αποτελέσματα των διαφορετικών μοντέλων συγκρίθηκαν ως προς την ακρίβεια πρόβλεψής αξιοποιώντας ως μετρική την MSE και την R2. Τα αποτελέσματα έδειξαν ότι το μοντέλο XGBoost είχε την υψηλότερη προγνωστική ακρίβεια. Ακόμα, εξετάστηκε η σημαντικότητα των χαρακτηριστικών του συνόλου δεδομένων ως προς την συμβολή τους στην πραγματοποίηση ακριβέστερων προβλέψεων. Εδώ το σημαντικότερο χαρακτηριστικό αποδεικνύετε πως είναι το πλήθος των ατόμων που μπορούν να διαμείνουν σε ένα κατάλυμα. Συνολικά, τα αποτελέσματα αυτής της μελέτης καταδεικνύουν την αποτελεσματικότητα των μοντέλων μηχανικής εκμάθησης στην πρόβλεψη της τιμής μιας καταχώρισης Airbnb και υπογραμμίζουν τη σημασία της εξέτασης τόσο της ακρίβειας πρόβλεψης όσο και της επιρροής της από τα σημαντικότερα χαρακτηριστικά.

In this thesis, the use and comparison of machine learning models to predict the price of an Airbnb listing is explored. More specifically, a dataset was collected from Airbnb listings in Amsterdam with a number of features such as property type, number of bedrooms, neighborhood, etc. The data was then cleaned and pre-processed to feed to the machine learning models. Several machine learning models including regression trees, random forest, and support vector machine were trained and tested on the dataset. The results of the different models were compared on the prediction accuracy by using MSE and R2 as metrics. The results showed that the XGBoost model had the highest predictive accuracy. However, the importance of features in terms of their contribution to making more accurate predictions was examined. Here the most important feature is proved that is the number of people who can stay in a property. Overall, the results of this study demonstrate the effectiveness of machine learning models in predicting the price of an Airbnb listing and highlights the importance of considering both the prediction accuracy and the influence of the most important features.

Ευχαριστίες (Acknowledgments)

Στην οικογένεια μου και στον καθηγητή μου Διονύσιο Σωτηρόπουλο

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

1. Εισαγωγή (Introduction)

Για τον άνθρωπο, η σημασία της παροχής ενός καλού τόπου διαβίωσης δεν μπορεί να υποεκτιμηθεί. Είτε πρόκειται για προσωρινή κατοικία για διακοπές, είτε για μόνιμη κατοικία, το περιβάλλον στο οποίο ζει έχει βαθιές επιπτώσεις στην υγεία, την ευημερία και τη συνολική ποιότητα ζωής του. Ένας άνετος, ασφαλής και λειτουργικός χώρος διαβίωσης είναι απαραίτητος για τη σωματική και ψυχική υγεία και μπορεί ακόμη και να επηρεάσει την κοινωνική και συναισθηματική ανάπτυξη του. Επομένως, η επένδυση σε ένα καλό μέρος διαβίωσης δεν είναι σημαντική μόνο για τη δική του προσωπική ικανοποίηση αλλά και για τη μακροπρόθεσμη βιωσιμότητα των κοινοτήτων αλλά και του πλανήτη γενικότερα [1]. Στον σημερινό κόσμο που η καθημερινότητα του μέσου ανθρώπου κινείται σε γοργούς ρυθμούς, η σημασία της εύρεσης ενός καλού τόπου διαβίωσης είναι ακόμα πιο σημαντική. Με την αυξανόμενη αστικοποίηση και την άνοδο της τεχνολογίας, πολλοί άνθρωποι βρίσκονται σε μικρά διαμερίσματα ή στενούς χώρους διαβίωσης. Αυτό μπορεί να οδηγήσει σε συναισθήματα απομόνωσης και αποσύνδεσης από τον φυσικό κόσμο, κάτι που μπορεί να έχει αρνητικό αντίκτυπο στην ψυχική και συναισθηματική του υγεία. Από την άλλη, ένα καλό μέρος διαβίωσης, παρέχει μια αίσθηση άνεσης και ασφάλειας, που επιτρέπει στον άνθρωπο, να ηρεμήσει και να χαλαρώσει μετά από μια κουραστική μέρα και ακόμα και να επανασυνδεθεί με τον εαυτό του. Για πολλούς ανθρώπους, ένας καλός χώρος διαβίωσης είναι επίσης πηγή υπερηφάνειας και αντανάκλαση του προσωπικού τους στυλ και ταυτότητας. Είναι ένας χώρος που μπορούν να επιμεληθούν και να διακοσμήσουν σύμφωνα με τις προτιμήσεις τους και που αναπαριστά την μοναδική προσωπικότητα και τα ενδιαφέροντά τους. Επίσης, ένα καλό μέρος διαβίωσης μπορεί να είναι πηγή υποστήριξης σε περιόδους άγχους ή αβεβαιότητας, παρέχοντας μια αίσθηση σταθερότητας [2].

Κατ' επέκταση, η ποιότητα ενός χώρου διαβίωσης επηρεάζει άμεσα την καθημερινή ζωή και τη ρουτίνα του κάθε ανθρώπου. Ένα καλά συντηρημένο και λειτουργικό σπίτι μπορεί να κάνει τις καθημερινές εργασίες και δραστηριότητες πιο αποτελεσματικές και ευχάριστες. Αντίθετα, ένας κακώς σχεδιασμένος ή μη προσεγμένος χώρος διαβίωσης μπορεί να οδηγήσει σε απογοήτευση και δυσφορία, παρεμποδίζοντας τη συνολική ποιότητα ζωής του εκάστοτε κατοίκου. Είναι λοιπόν, ζωτικής σημασίας για τους ανθρώπους να επενδύσουν στον καλύτερο δυνατό χώρο διαβίωσης που μπορούν να υποστηρίξουν οικονομικά (είτε για προσωρινή διαμονή είτε σε μόνιμη κατοικία) για να εξασφαλίσουν τη σωματική, ψυχική και συναισθηματική τους ευεξία [3].

Σε αυτήν την απαραίτητη ανάγκη του ανθρώπου έρχεται να συνεισφέρει η μηχανική μάθηση. Η μηχανική μάθηση είναι ένα ισχυρό εργαλείο που χρησιμοποιείται σε διάφορους κλάδους, για τη βελτιστοποίηση των διαδικασιών και την πραγματοποίηση προβλέψεων. Ένας τομέας όπου η μηχανική μάθηση εφαρμόζεται σφόδρα είναι στη σφαίρα των ακινήτων, και πιο συγκεκριμένα στην πρόβλεψη των τιμών των κατοικιών, είτε αγοράς είτε ενοικίασης. Αυτή η συνεισφορά είναι ιδιαίτερα σημαντική, καθώς η ακριβής πρόβλεψη των τιμών των κατοικιών μπορεί να βοηθήσει τους αγοραστές, τους πωλητές, τους ενοικιαστές, τους οικοδεσπότες και τους επενδυτές να λαμβάνουν τεκμηριωμένες αποφάσεις [4].

Τα τελευταία χρόνια, η Airbnb έχει αναδειχθεί ως κορυφαία εταιρία στην ενοικίαση ακινήτων, διαμερισμάτων κτλ. Η Airbnb επιτρέπει στους ιδιώτες να νοικιάζουν τα σπίτια ή τα διαμερίσματά τους σε ταξιδιώτες, παρέχοντας μια μοναδική και συχνά πιο προσιτή εναλλακτική λύση απέναντι στα παραδοσιακά ξενοδοχεία. Ως αποτέλεσμα, η Airbnb διαθέτει πληθώρα δεδομένων για τις τιμές διαφόρων ακινήτων σε όλο τον κόσμο. Αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν για την εκπαίδευση αλγορίθμων μηχανικής μάθησης για την πρόβλεψη των τιμών των σπιτιών ή των διαμερισμάτων. Υπάρχουν πολλές διαφορετικές προσεγγίσεις για τη χρήση της μηχανικής εκμάθησης για την πρόβλεψη των τιμών των κατοικιών [5].

Μια κοινή προσέγγιση είναι η χρήση ενός εποπτευόμενου αλγόριθμου εκμάθησης (Supervised Learning), όπου ο αλγόριθμος εκπαιδεύεται σε ένα σύνολο δεδομένων με γνωστές τιμές ενοικίασης κατοικιών και των σχετικών χαρακτηριστικών τους. Όπως η τοποθεσία, το μέγεθος και ο αριθμός των υπνοδωματίων κτλ. Ο αλγόριθμος μπορεί στη συνέχεια να χρησιμοποιηθεί για να κάνει προβλέψεις σε νέα δεδομένα και να προσφέρει εκτιμήσεις στους πελάτες της πλατφόρμας. Μια άλλη προσέγγιση είναι η χρήση της μάθησης χωρίς επίβλεψη (Unsupervised Learning), όπου στον αλγόριθμο δεν δίνονται προεπισημασμένα δεδομένα και πρέπει, μόνος του, να ανακαλύψει μοτίβα και σχέσεις μεταξύ των δεδομένων. Αυτή η προσέγγιση μπορεί να είναι χρήσιμη για τον εντοπισμό τάσεων και ανωμαλιών στις τιμές των κατοικιών και μπορεί να βοηθήσει στον εντοπισμό πιθανών παραγόντων που μπορεί να επηρεάσουν τις τιμές.

Εκτός από την Airbnb, άλλες εταιρείες χρησιμοποιούν επίσης μηχανική μάθηση για την πρόβλεψη των τιμών των κατοικιών. Για παράδειγμα, ορισμένοι ιστότοποι ακινήτων χρησιμοποιούν μηχανική μάθηση για να παρέχουν στους χρήστες εκτιμήσεις της αξίας ενός σπιτιού με βάση την τοποθεσία, το μέγεθός του και άλλους παράγοντες. Αυτό μπορεί να βοηθήσει τους πιθανούς αγοραστές και πωλητές να αποκτήσουν μια καλύτερη ιδέα για το τι αξίζει ένα ακίνητο και τι μπορούν να αναμένουν να λάβουν γι' αυτό [6].

Συνολικά, η χρήση της μηχανικής μάθησης στην πρόβλεψη των τιμών των κατοικιών είναι μια σημαντική εξέλιξη στον κλάδο των ακινήτων. Παρέχοντας πιο ακριβείς και ενημερωμένες πληροφορίες σχετικά με τις αξίες των ακινήτων, η μηχανική μάθηση μπορεί να βοηθήσει τους αγοραστές, τους πωλητές και τους επενδυτές να λάβουν πιο ενημερωμένες αποφάσεις και ενδεχομένως να εξοικονομήσουν χρήματα. Καθώς οι αλγόριθμοι μηχανικής εκμάθησης συνεχίζουν να βελτιώνονται και περισσότερα δεδομένα γίνονται διαθέσιμα, μπορούμε να περιμένουμε να δούμε ακόμη μεγαλύτερες προόδους σε αυτόν τον τομέα στο επερχόμενο μέλλον.

1.1. Στόχος Διπλωματικής (Thesis Objective)

Ο στόχος αυτής της διπλωματικής είναι η δημιουργία ενός μοντέλου μηχανικής μάθησης που να μπορεί να προβλέψει με ακρίβεια τις τιμές ενοικίασης κατοικιών της πλατφόρμας Airbnb στο Άμστερνταμ. Αυτό το μοντέλο θα βασίζεται σε ένα μεγάλο σύνολο ιστορικών δεδομένων που έχουν συλλεχθεί από την τον επίσημο ιστότοπο της Airbnb. Το πρώτο βήμα για τη δημιουργία αυτού του μοντέλου θα είναι η συγκέντρωση και ο καθαρισμός του συνόλου δεδομένων. Αυτό θα περιλαμβάνει τον εντοπισμό σχετικών μεταβλητών, όπως η τοποθεσία, το μέγεθος του καταλύματος κτλ., καθώς και η αφαίρεση τυχόν άσχετων ή ελλιπών δεδομένων. Στη συνέχεια, το καθαρισμένο σύνολο δεδομένων θα χωριστεί σε ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου. Το σύνολο εκπαίδευσης θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου και το σύνολο ελέγχου για την αξιολόγηση της απόδοσής του.

Το μοντέλο μηχανικής μάθησης θα αναπτυχθεί χρησιμοποιώντας έναν από του κυρίαρχους αλγορίθμους, όπως ένα δέντρο παλινδρόμησης, ένα τυχαίο δάσος παλινδρόμησης, μία μηχανή διανυσματικής υποστήριξης κτλ. Στην συνέχεια το μοντέλο εκπαιδευτεί και θα βελτιστοποιηθεί μέσω μιας διαδικασίας τελειοποίησης. Αυτό θα περιλαμβάνει την προσαρμογή των παραμέτρων του μοντέλου στα δεδομένα εκπαίδευσης για την αύξηση της απόδοσής του. Στόχος αυτής της διαδικασίας είναι να επιτευχθεί η υψηλότερη δυνατή ακρίβεια στην πρόβλεψη των τιμών των κατοικιών.

Τέλος, το εκπαιδευμένο και τελειοποιημένο μοντέλο θα εξεταστεί σε πραγματικό σενάριο. Αυτό θα περιλαμβάνει τη χρήση του μοντέλου για την πραγματοποίηση προβλέψεων σε νέα σύνολο δεδομένα ενοικίασης κατοικιών και τη σύγκριση των προβλέψεων του μοντέλου με τις

πραγματικές τιμές ενοικίασης. Η απόδοση του μοντέλου θα αξιολογηθεί με βάση την ακρίβεια και την ικανότητά του να παρέχει χρήσιμες και αξιόπιστες προβλέψεις.

Συνολικά, ο στόχος αυτής της διατριβής είναι να εξεταστεί πλήθος διαφορετικών αλγορίθμων που αποτελούν κορυφαίες επιλογές στην επίλυση προβλημάτων παλινδρόμησης στην μηχανική μάθηση. Καθώς και η σύγκριση αυτών μεταξύ τους για την ανεύρεση του αποδοτικότερου στην πρόβλεψη τιμών ενοικίασης ακινήτων στο Άμστερνταμ. Τέλος, μέσω της παραπάνω διαδικασίας θα αξιολογηθούν και τα χαρακτηριστικά των ακινήτων ως προς την σημαντικότητα τους στην επιρροή της τελικής τιμής ενοικίασης του εκάστοτε καταλύματος.

1.2. Ερευνητικά Ερωτήματα (Research Questions)

Κατ' επέκταση, για την επίτευξη του παραπάνω στόχου η ανάπτυξη της συγκεκριμένης διπλωματικής θα εστίαση στην απάντηση των παρακάτω ερευνητικών ερωτημάτων:

1. Ποια είναι τα χαρακτηριστικά που επηρεάζουν περισσότερο την τιμή ενοικίασης ενός ακινήτου στο Άμστερνταμ;
2. Ποιο μοντέλο είναι καταλληλότερο για τον προσδιορισμό της ενοικίασης του καταλύματος;

2. Υπάρχουσα Βιβλιογραφία (Literature Review)

Η θεματική που θίγεται σε αυτήν την εργασία, αφορά υποκατηγορία της επιστήμης των υπολογιστών που ονομάζεται μηχανική μάθηση. Η μηχανική μάθηση, στηρίζετε σε μοντέλα της στατιστικής, κλάδου της επιστήμης των μαθηματικών και χρησιμοποιείται για την μελέτη και υλοποίηση αλγορίθμων που έχουν την δυνατότητα να «μαθαίνουν» από τα δεδομένα, να μοντελοποιούν την φύση τους και με βάση αυτήν να κάνουν τις εκάστοτε προβλέψεις. Το πρόβλημα που μελετάται στην παρούσα εργασία, αποτελεί ένα πρόβλημα παλινδρόμησης επιβλεπόμενης μάθησης. Η μοντελοποίηση της σχέσης μεταξύ της εξαρτώμενης μεταβλητής, στην προκειμένη περίπτωση της τιμής ενοικίασης των καταλυμάτων, και των ανεξάρτητων μεταβλητών, που αποτελούν τα χαρακτηριστικά του κάθε καταλύματος, συνιστά τον απώτερο σκοπό της. Εδώ και αρκετά χρόνια, πραγματοποιούνται προσπάθειες και γίνονται μελέτες σχετικά με το παραπάνω πρόβλημα επιστρατεύοντας τεχνικές μηχανικής μάθησης για την επίλυση τού. Πολλοί μελετητές και επιστήμονες, των οποίων το έργο υπάρχει δημοσιευμένο έχουν προσεγγίσει το παραπάνω πρόβλημα. Οι μελέτες στις οποίες στηρίχθηκε η παρούσα έρευνα παρουσιάζονται στη συνέχεια.

Τα ξενοδοχεία κυριαρχούσαν εδώ και χρόνια στον κλάδο της φιλοξενίας μέχρι που εμφανίστηκαν επιχειρήσεις κοινής χρήσης, όπως η Airbnb. Η Airbnb, που ιδρύθηκε το 2008, παρουσίασε ένα νέο επιχειρηματικό μοντέλο οικονομίας διαμοιρασμού που έφερε επανάσταση στον κλάδο της φιλοξενίας και συνδέει τους ιδιοκτήτες ελεύθερων καταλυμάτων με ταξιδιώτες που αναζητούν προσωρινή διαμονή [7]. Από το 2019, υπάρχουν πάνω από 6 εκατομμύρια καταχωρίσεις στον ιστότοπο του Airbnb σε περίπου 220 χώρες και περιοχές, πραγματοποιώντας κατά μέσο όρο πάνω από 1 εκατομμύρια διαμονές ανά βραδιά [8].

Ο [9] εξηγεί ότι τα εργαλεία δυναμικής τιμολόγησης καταλυμάτων του Airbnb βασίζονται σε τεχνητή νοημοσύνη. Αναφέρει μάλιστα, πώς το αρχικό εργαλείο που κυκλοφόρησε το 2015 βασιζόταν σε τεχνικές παλινδρόμησης και χρησιμοποιούσε ως είσοδο τις ανέσεις (amenities) ενός καταλύματος και πληροφορίες σχετικά με τα γειτονικά ακίνητα. Πλέον, έχει αντικατασταθεί με εργαλείο της εταιρίας που βασίζεται στην ενισχυτική μάθηση (Reinforcement Learning). Παρόλα αυτά έχει ιδιαίτερο ενδιαφέρον ότι και η ίδια εταιρία που διαθέτει τον πλήρη όγκο δεδομένων στην αρχή επέλεξε να δημιουργεί προτάσεις τιμών χρησιμοποιώντας παραδοσιακές τεχνικές μηχανικής μάθησης.

Το 2019 οι [10], με τη χρήση των δεδομένων που παρέχονται από την πλατφόρμα Inside Airbnb, μελετούν την πρόβλεψη ενοικίασης των καταλυμάτων Airbnb στην Μελβούρνη της Αυστραλίας με την χρήση μοντέλων μηχανικής μάθησης. Η έρευνά τους περιλαμβάνει την σύγκριση διάφορων μοντέλων πρόβλεψης τιμής, όπως τα νευρωνικά δίκτυα, καθώς και παραδοσιακών μεθόδων μηχανικής μάθησης, όπως είναι τεχνικές παλινδρόμησης, τα τυχαία δάση (Random Forest) και η ενισχυτική κλίση (Gradient Boosting). Υπολογίζοντας το Μέσο Τετραγωνικό Σφάλμα (RMS) και τον συντελεστή προσδιορισμού R^2, αξιολογούν τα παραπάνω μοντέλα και διαπιστώνουν ότι καλύτερη απόδοση έχει η μέθοδος παλινδρόμησης με ενισχυτική κλίση, ενώ αμέσως μετά έρχεται η μέθοδος των τυχαίων δασών, η οποία ενδεχομένως να είχε βελτιωμένη απόδοση με αυστηρότερη επιλογή χαρακτηριστικών.

Ακόμα, για να μπορέσουν να καθορίσουν με ακρίβεια την τιμή ενοικίασης ενός καταλύματος για έναν host, πολλοί ερευνητές χρησιμοποιούν διάφορες μεθόδους με τρία κύρια στοιχεία:

i) Ένα δυαδικό μοντέλο ταξινόμησης προβλέπει την πιθανότητα κράτησης κάθε διανυκτέρευσης

ii) Ένα μοντέλο παλινδρόμησης προβλέπει το ιδανικό κόστος για κάθε διανυκτέρευση

iii) Εξατομικευμένο συλλογισμό πάνω στην πρόβλεψη του μοντέλου παλινδρόμησης για την παροχή των τελευταίων προτάσεων τιμής ανάλογα με τους στόχους του host [11]

Την παραπάνω προσέγγιση επιβεβαιώνει προγενέστερη έρευνα των [12], που χρησιμοποιούν μοντέλα παλινδρόμησης για την μεγιστοποίηση των εσόδων ενός host. Πιο συγκεκριμένα, εφαρμόζουν τον αλγόριθμο Gradient Boosting Machine (GBM) για να προβλέψουν την πιθανότητα κράτησης ενός καταλύματος. Στη συνέχεια, δημιούργησαν ένα μοντέλο παλινδρόμησης σχετικό με την πρόβλεψη τιμής για το εκάστοτε βράδυ και τέλος, προσάρμοζαν τις προτάσεις του μοντέλου παλινδρόμησης στους τους προσωπικούς στόχους που εκάστοτε host.

Οι [13] σύγκριναν μοντέλα όπως τυχαία δάση (Random Forest), μετεξέλιξη της ενισχυτικής κλίσης (XGBoost) και νευρωνικά δίκτια (Neural Networks) στην πρόβλεψη τιμών Airbnb πάνω σε δεδομένα της Νέας Υόρκης και του Παρισιού. Αξίζει να σημειωθεί ότι στο στάδιο της προεπεξεργασίας, κατάργησαν πολλά χαρακτηριστικά για να μειώσουν το θόρυβο και να δώσουν έμφαση σε χαρακτηριστικά όπως country_code και τον αριθμό υπνοδωματίων, τα οποία και θεώρησαν ως πιο ουσιώδη. Ακόμα, δεδομένα κειμένου, όπως η περιγραφή του καταλύματος, θεωρούνται επίσης χρήσιμα χαρακτηριστικά. Παρόλα αυτά όμως τέτοια χαρακτηριστικά απαιτούν ιδιαίτερη προεπεξεργασία καθώς είναι σε αδόμητη μορφή. Τέλος, μετά την εκτέλεση εκτεταμένης προεπεξεργασίας και καθαρισμού δεδομένων, το μοντέλο XGBoost πέτυχε την καλύτερη απόδοση. Ως μετρικές αξιολόγησης των μοντέλων χρησιμοποιήθηκαν οι: R^2 και το μέσο τετραγωνικό σφάλμα (MSE).

Οι ερευνητές στο Πανεπιστημίου του Στάνφορντ πραγματοποιούν παρόμοια ερεύνα, συγκρίνουν επίσης πολλαπλές μεθόδους μηχανικής μάθησης που περιλαμβάνουν: Γραμμική Παλινδρόμηση, μοντέλα που βασίζονται σε Δέντρα Παλινδρόμησης, Διανυσματικές Μηχανές Υποστήριξης για Παλινδρόμηση (SVR) και K-means (KMC). Οι κύριες συνεισφορές αυτής της εργασίας είναι ότι χρησιμοποιούν τεχνικές επιλογής χαρακτηριστικών (feature selection), οι οποίες και δίνουν τα 22 καλύτερα χαρακτηριστικά για την πρόβλεψη της τιμής ενοικίασης. Επίσης, πρόσθεσαν ανάλυση συναισθήματος για να εξετάσουν τις κριτικές των πελατών [14]. Ωστόσο, αυτή η έρευνα δεν προσέφερε τελικά ένα ολοκληρωμένο μοντέλο πρόβλεψης τιμών. Η Laura Lewis στην δουλειά της το 2019 μελετά επίσης τον καθορισμό τιμών σε καταχωρίσεις της Airbnb στο Λονδίνο κυρίως μέσο της μεθόδου XGBoost. Στην προσέγγιση της δίνει ιδιαίτερη έμφαση στις εργασίες προεπεξεργασίας, καθαρισμού και ανάλυση των δεδομένων [15].

Όπως γίνετε αντιληπτό, πολλές μέθοδοι μηχανικής μάθησης έχουν εφαρμοστεί για την πρόβλεψη της αξίας ενός καταλύματος. Παρόλα αυτά, το σημαντικότερο στοιχείο στην επίτευξη ανταγωνιστικών αποτελεσμάτων αποτελεί η σωστή επιλογή χαρακτηριστικών [16].

Κατ’ επέκταση, πολλές έρευνες εργάστηκαν σχετικά με το ποια χαρακτηριστικά επηρεάζουν την τιμή ενοικίασης του καταλύματος. Ορισμένες από αυτές, επιλέγουν να χωρίζουν τα χαρακτηρίστηκα ενός καταλύματος σε δυο κατηγορίες. Στα host-controlled χαρακτηριστικά, που είναι χαρακτηριστικά που παρέχονται από τον host είτε αφορούν πληροφορίες σχετικά με τον ίδιο τον host. Και στα out_of_host_controlled χαρακτηριστικά, που αφορούν πληροφορίες που δεν μπορεί να επηρεάσει ο host. Σύμφωνα με τον Brando MaNeil, τόσο τα host_controlled όσο και τα out_of_host_controlled χαρακτηριστικά είναι εξίσου σημαντικά για τον καθορισμό της τιμής ενοικίασης ενός καταλύματος στην Airbnb. Διαπίστωσε επίσης, ότι όταν συνδυάζονται χαρακτηριστικά host_controlled και χαρακτηριστικά out_of_host_controlled προβλέπεται με μεγαλύτερη ακρίβεια η τιμή ενοικίασης του εκάστοτε καταλύματος. Ωστόσο, τα host_controlled χαρακτηριστικά εξακολουθούν να έχουν μεγαλύτερη σημασία για την τιμή καταχώρισης από ό,τι τα out_of_host_controlled χαρακτηριστικά [17].

Προηγούμενη έρευνα έδειξε ότι η ενοικίαση ενός ολόκληρου σπιτιού θα έχει υψηλότερα μέσα έσοδα σε σχέση με την ενοικίαση του κάθε δωματίου ξεχωριστά. Επίσης, πιο επαγγελματίες hosts

αποκτούν κατά μέσο όρο περισσότερες κριτικές και τείνουν να έχουν υψηλότερα μηνιαία έσοδα. Στην Airbnb, αυτό το χαρακτηριστικό εμφανίζεται ως "super-host" [18]. Ακόμα, οι [19] επιβεβαίωσαν ότι χαρακτηριστικά που σχετίζονται με τοποθεσίες, τις υπηρεσίες καθώς και τις κριτικές πελατών θα μπορούσαν να επηρεάσουν τις τιμές των ενοικιαζόμενων καταλυμάτων. Προσδιόρισαν τα χαρακτηριστικά του host (host_controlled) ως σημαντικούς καθοριστικούς παράγοντες της τιμής.

3. Θεωρητικό Υπόβαθρο

3.1. Τεχνητή Νοημοσύνη και Μηχανική Μάθηση

Η Τεχνητή Νοημοσύνη (Τ.Ν.) είναι ο επιστημονικός τομέας που εστιάζει στην προσομοίωση της ανθρώπινης νοημοσύνης σε μηχανές και υπολογιστές με σκοπό να μιμηθούν τις γνωστικές λειτουργίες του ανθρώπινου μυαλού, όπως η μάθηση, η επίλυση προβλημάτων και η λήψη αποφάσεων [20]. Σκοπός της είναι η δημιουργία αλγορίθμων και μοντέλων, μέσω των οποίων οι υπολογιστές ανεξάρτητα θα αποφασίζουν για τον ακριβή τρόπο εκτέλεσης μια συγκεκριμένης διεργασίας μαθαίνοντας από παραδείγματα, σε αντίθεση με την προσέγγιση στην οποία ο υπολογιστής έπρεπε να προγραμματιστεί επακριβώς ανά περίπτωση. Επομένως, ένα μοντέλο Τεχνητής Νοημοσύνης θα μπορεί αργότερα (αφότου έχει εκπαιδευτεί) να επεκτείνει τον συλλογισμό του σε νέα (παρεμφερή) δεδομένα. Αυτό πετυχαίνετε, καθώς δημιουργεί εσωτερικά μια γενικευμένη μεθοδολογία που εξυπηρετεί όλες τις δυνατές καταστάσεις του προβλήματος που προσπαθεί να επιλύσει κατά των καλύτερο δυνατό τρόπο, όπως συνηθίζουν να κάνουν και οι άνθρωποι.

Η Μηχανική Μάθηση είναι ένα υποσύνολο της Τεχνητής Νοημοσύνης που περιλαμβάνει τη χρήση αλγορίθμων και μεθοδολογιών παρμένων από την στατιστική για να επιτρέψει σε έναν αλγόριθμο να βελτιώσει την απόδοσή του σε μια συγκεκριμένη εργασία μέσω της εμπειρίας. Μία από τις χαρακτηριστικές εφαρμογές της μηχανικής μάθησης είναι ο τομέας της ανάλυσης δεδομένων. Χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης, οι επιστήμονες δεδομένων είναι σε θέση να εξάγουν πολύτιμες γνώσεις και μοτίβα από μεγάλα και πολύπλοκα σύνολα δεδομένων. Πιο συγκεκριμένα, τα μοντέλα μηχανικής μάθησης, εξάγοντας μοτίβα από τα δεδομένα προσδιορίζουν σχέσεις μεταξύ τους και κατ' επέκταση αφότου επαναληφθεί αυτή η διαδικασία σε πολλά ιστορικά δεδομένα, μπορεί μετά να χρησιμοποιηθεί σε νέα δεδομένα και αξιοποιώντας αυτές τις σχέσεις σαν κανόνες να εκτιμήσει χαρακτηρίστηκα που επιθυμούν οι επιστήμονες δεδομένων [21].

Οι περισσότεροι άνθρωποι συνδέουν την Τεχνητή Νοημοσύνη και την Μηχανική Μάθηση με τα ρομπότ, καθώς [22] είναι πιο εξοικειωμένοι με αυτές τις τεχνολογίες μέσω ταινιών, σειρών και βιντεοπαιχνιδιών κτλ. Ωστόσο, η Τεχνητή Νοημοσύνη και η Μηχανική Μάθηση αναφέρονται σε κάθε είδους μηχανή ικανή να εκτελέσει πολλαπλές και σύνθετες υπολογιστικές εργασίες καθώς κάθε πληροφορία και εντολή που λαμβάνει το μοντέλο μεταφράζεται σε αριθμούς και μαθηματικές πράξεις. Ακόμα, δεδομένου ότι οι υπολογιστές βελτιστοποιούν συνεχώς την υπολογιστική τους ισχύ, ο χρόνος εκτέλεσης περίπλοκων μαθηματικών εξισώσεων και πράξεων έχει μειωθεί σημαντικά. Συνεπώς, η δυνατότητα ενός μοντέλου στο να «σκέφτεται» και να «αποφασίζει» αποτελεσματικά (με ακρίβεια και σε πραγματικό χρόνο), εξαρτάται από την ικανότητά του να εκτελεί γρήγορες μαθηματικές πράξεις. Έτσι, η Τεχνητή Νοημοσύνη και η Μηχανική Μάθηση έχουν γίνει ένας από τους πιο κυρίαρχους τομείς στην πρόσφατη έρευνα. Η ανάπτυξη της Τεχνητής Νοημοσύνης οδήγησε στην εφεύρεση νέων τεχνολογιών όσο και στην βελτιώσει των υπαρχόντων. Κατ' επέκταση, βιομηχανίες, εταιρείες, κατασκευαστές και ερευνητικά κέντρα χρησιμοποιούν την Τεχνητή Νοημοσύνη για να υποστηρίξουν ή να εκτελέσεων εξολοκλήρου λειτουργικές διαδικασίες.

Χάρη στην Τεχνητή Νοημοσύνη έχουν επωφεληθεί πολλά πεδία της καθημερινότητας του ανθρώπου. Η αυτόνομη οδήγηση τώρα δεν είναι μόνο ένα σενάριο γραμμένο για βιντεοπαιχνίδια, σειρές και ταινίες επιστημονικής φαντασίας. Σήμερα, πολλά αυτοκίνητα έχουν την ικανότητα να κατανοούν το περιβάλλον τους μέσω καμερών και αισθητήρων στο όχημα, και έτσι μπορούν να εκτελούν τις κατάλληλες εργασίες. Επιπλέον, εργαλεία μετάφρασης κειμένου και πλέον και ομιλίας έχουν βοηθήσει στην επικοινωνία ανθρώπων διαφορετικών εθνικοτήτων χωρίς την υπάρξει κοινής γλώσσας. Τόσο οι μικρές όσο και οι μεγάλες εταιρείες διαθέτουν τμήματα Analytics και Τ.Ν., ώστε να μπορούν να διεξάγουν στατιστικά αποτελέσματα σχετικά με τους παράγοντες που επηρεάζουν την αγορά και τα προϊόντα τους και ως εκ τούτου να βρίσκουν επιχειρηματικές λύσεις σχετικά με τα έσοδά τους.

Ένα ακόμα παράδειγμα αποτελεί ο τρόπος με τον οποίο η Airbnb χρησιμοποιεί μηχανική μάθηση και τεχνητή νοημοσύνη για να βελτιώσει τη στρατηγική τιμολόγησης των καταλυμάτων της. Αυτό το εργαλείο χρησιμοποιεί αλγόριθμους για την ανάλυση δεδομένων σχετικά με παράγοντες όπως η τοπική ζήτηση, ο ανταγωνισμός και οι αξιολογήσεις χρηστών για αυτόματη προσαρμογή των τιμών για κάθε καταχώριση σε πραγματικό χρόνο. Αυτό επιτρέπει στους οικοδεσπότες να ορίσουν μια βασική τιμή για την καταχώρισή τους και να αφήσουν τον αλγόριθμο να κάνει τα υπόλοιπα, βοηθώντας τους να μεγιστοποιήσουν τα κέρδη τους και να παραμείνουν ανταγωνιστικοί στην αγορά [6].

Τέλος, ένα ακόμα σημείο που αξίζει να αναφερθεί είναι ότι παρατηρώντας την σημαντική αύξηση τις χρήσεις τεχνητής νοημοσύνης σε κλάδους όπως στην υγειονομική περίθαλψη για τη βελτίωση της διάγνωσης και της θεραπείας, στην εκπαίδευση για την εξατομίκευση της μάθησης και στον στρατό για την ενίσχυση της αποτελεσματικότητας των οπλικών συστημάτων και γενικότερα η αυξανόμενη εξάρτηση από την τεχνητή νοημοσύνη έχει εγείρει ανησυχίες για πιθανές αρνητικές συνέπειες, όπως η απώλεια θέσεων εργασίας και η μεροληψία στη λήψη αποφάσεων.

3.2. Προεπεξεργασία (Preprocessing)

3.2.1. Κωδικοποίηση (One hot encoding)

Η one hot encoding αντιμετωπίζει την πρόκληση της αναπαράστασης κατηγορικών χαρακτηριστικών. Οι παραδοσιακές αναπαραστάσεις εκχωρούν αριθμητικές τιμές σε κάθε κατηγορία, που συνήθως κυμαίνονται από 0 έως $k-1$ για ένα χαρακτηριστικό με k κατηγορίες. Ωστόσο, αυτή η αναπαράσταση αποτυγχάνει να συλλάβει τις εγγενείς σχέσεις και ομοιότητες μεταξύ διαφορετικών κατηγοριών. Τα μοντέλα που βασίζονται στον υπολογισμό των αποστάσεων μεταξύ των χαρακτηριστικών ή εντός κάθε χαρακτηριστικού, μπορεί να ερμηνεύουν αυτές τις αριθμητικές αναπαραστάσεις με τρόπο που εισάγει ακούσιες προκαταλήψεις ή παρανοήσεις.

Η θεωρητική βάση της one hot encoding έχει τις ρίζες της στην ανάγκη δημιουργίας μιας κατηγορικής αναπαράστασης που εξαλείφει τυχόν ψευδείς υποθέσεις σχετικά με τις σχέσεις μεταξύ των κατηγοριών. Μετατρέποντας ένα κατηγορικό χαρακτηριστικό με k κατηγορίες σε k νέα δυαδικά χαρακτηριστικά, καθένα από τα οποία αντιστοιχεί σε μια συγκεκριμένη κατηγορία, μια one hot κωδικοποίηση διασφαλίζει ότι όλες οι κατηγορίες αντιμετωπίζονται ως ξεχωριστές οντότητες χωρίς καμία σιωπηρή σειρά ή ιεραρχία. Κάθε νέο χαρακτηριστικό παίρνει μια τιμή 0 ή 1, υποδεικνύοντας εάν μια συγκεκριμένη εγγραφή ανήκει στην αντίστοιχη κατηγορία ή όχι.

Επιπλέον, η one hot encoding επιτρέπει στα μοντέλα να εκτελούν υπολογισμούς και συγκρίσεις με ουσιαστικό και αμερόληπτο τρόπο. Δεδομένου ότι κάθε κατηγορία αντιπροσωπεύεται από ένα ξεχωριστό χαρακτηριστικό, το μοντέλο μπορεί να αξιολογήσει την παρουσία ή την απουσία μιας συγκεκριμένης κατηγορίας χωρίς υποθέσεις σχετικά με τη σχέση της με άλλες κατηγορίες. Αυτή η προσέγγιση διασφαλίζει ότι οι υπολογισμοί απόστασης ή οι αξιολογήσεις ομοιότητας βασίζονται αποκλειστικά στην παρουσία ή την απουσία μιας κατηγορίας και όχι σε αυθαίρετες αριθμητικές εκχωρήσεις.

Μια one hot encoding ενισχύει επίσης την ερμηνευτικότητα των αποτελεσμάτων του μοντέλου. Με τη μετατροπή των κατηγορικών χαρακτηριστικών σε δυαδικά χαρακτηριστικά, ο αντίκτυπος κάθε κατηγορίας στις προβλέψεις του μοντέλου γίνεται πιο διαφανής. Οι αναλυτές και τα ενδιαφερόμενα μέρη μπορούν εύκολα να ερμηνεύσουν τις τιμές των συντελεστών που σχετίζονται με κάθε κατηγορία, αποκτώντας πληροφορίες για το ποιες κατηγορίες έχουν σημαντική επιρροή στην παραγωγή του μοντέλου. Αυτή η ερμηνευσιμότητα διευκολύνει τη λήψη αποφάσεων και παρέχει μια βαθύτερη κατανόηση των παραγόντων που οδηγούν στις προβλέψεις του μοντέλου.

Είναι σημαντικό να σημειωθεί ότι η one hot encoding μπορεί να δημιουργήσει ορισμένες προκλήσεις, ιδιαίτερα όταν ασχολούμαστε με κατηγορηματικά χαρακτηριστικά που έχουν μεγάλο αριθμό κατηγοριών. Καθώς, ο αριθμός των κατηγοριών αυξάνεται, ο αριθμός των νέων χαρακτηριστικών που δημιουργούνται μέσω της one hot encoding αυξάνεται επίσης, οδηγώντας ενδεχομένως σε έναν χώρο χαρακτηριστικών υψηλότερων διαστάσεων. Αυτή η επέκταση μπορεί να αυξήσει την υπολογιστική πολυπλοκότητα και τις απαιτήσεις μνήμης του μοντέλου, ειδικά για μεγάλα σύνολα δεδομένων. Σε τέτοιες περιπτώσεις, τεχνικές μείωσης διαστάσεων ή εναλλακτικές μέθοδοι κωδικοποίησης μπορούν να διερευνηθούν για τον μετριασμό αυτών των προκλήσεων.

Συμπερασματικά, η one hot encoding είναι μια ισχυρή τεχνική για την αναπαράσταση κατηγορικών χαρακτηριστικών σε μοντέλα μηχανικής μάθησης. Με τη μετατροπή των κατηγορικών χαρακτηριστικών σε δυαδικά χαρακτηριστικά, εξαλείφει τυχόν υποθέσεις ή προκαταλήψεις σχετικά με τις σχέσεις μεταξύ των κατηγοριών. Αυτή η προσέγγιση διασφαλίζει ότι όλες οι κατηγορίες αντιμετωπίζονται ως ίσες και ανεξάρτητες, επιτρέποντας ακριβείς υπολογισμούς, αμερόληπτες συγκρίσεις και βελτιωμένη ερμηνεία των αποτελεσμάτων του μοντέλου. Ενώ η one hot encoding μπορεί να εισάγει προκλήσεις όσον αφορά την υπολογιστική πολυπλοκότητα για μεγάλους χώρους κατηγοριών χαρακτηριστικών, τα οφέλη της όσον αφορά τη διαφάνεια και τους ουσιαστικούς υπολογισμούς την καθιστούν πολύτιμο εργαλείο για αποτελεσματική μοντελοποίηση και λήψη αποφάσεων [23].

Στην παρακάτω εικόνα δίνεται οπτικά μια αναπαράσταση της συγκεκριμένης μεθόδου:

Type
AA
AB
CD
AA

Onehot Encoding →

Type	AA_Onehot	AB_Onehot	CD_Onehot
AA	1	0	0
AB	0	1	0
CD	0	0	1
AA	1	0	0

Figure 1: One-Hot Encoding

3.2.2. Ευρεση Παράτυπων Σημείων (Outlier Detection - Interquartile (IQR) method)

Η μέθοδος ενδοτεταρτομοριακό εύρους (IQR) προσφέρει μια απλή και διαισθητική προσέγγιση για τον εντοπισμό παράτυπων σημείων σε ένα σύνολο δεδομένων. Τα παράτυπα σημεία είναι σημεία δεδομένων που αποκλίνουν σημαντικά από τη συνολική κατανομή των δεδομένων και η παρουσία τους μπορεί να έχει αρνητικό αντίκτυπο στην απόδοση και την ακρίβεια των μοντέλων μηχανικής μάθησης.

Η θεωρητική βάση της μεθόδου του ενδοτεταρτομοριακό εύρους βασίζεται στην έννοια των τεταρτημορίων και στην κατανομή των δεδομένων μέσα σε αυτά. Τα τεταρτημόρια διαιρούν ένα σύνολο δεδομένων σε τέσσερα ίσα μέρη, παρέχοντας πληροφορίες για τη κατανομή και τη διακύμανση των δεδομένων. Το πρώτο τεταρτημόριο, $Q1$, αντιπροσωπεύει το $25ο$ ποσοστιαίο σημείο, υποδεικνύοντας ότι το 25% των δεδομένων βρίσκεται κάτω από αυτήν την τιμή. Το δεύτερο τεταρτημόριο, $Q2$, αντιστοιχεί στη διάμεσο του συνόλου, με το 50% των δεδομένων να πέφτει κάτω από αυτό το σημείο. Τέλος, το τρίτο τεταρτημόριο, το $Q3$, αντιπροσωπεύει το $75ο$ ποσοστιαίο σημείο, κάτω από το οποίο βρίσκεται το 75% των δεδομένων.

Για την εφαρμογή της μεθόδου του ενδοτεταρτομοριακού εύρους, ακολουθούνται τα παρακάτω βήματα:

1. Υπολογισμός του πρώτου τεταρτημορίου, $Q1$.
2. Υπολογισμός του τρίτου τεταρτημορίου, $Q3$.
3. Προσδιορισμός του ενδοτεταρτομοριακού εύρους, IQR, το οποίο αντιπροσωπεύει την απόσταση μεταξύ $Q1$ και $Q3$ και υπολογίζεται ως η απόλυτη διαφορά αυτών: $IQR = |Q3 - Q1|$.
4. Καθορισμός του αποδεκτού εύρους δεδομένων ορίζοντας το κάτω όριο ως $Q1 - 1{,}5 * IQR$ και το άνω όριο ως $Q3 + 1{,}5 * IQR$.
5. Κάθε δεδομένο που βρίσκεται εκτός αυτού του εύρους θεωρείται παράτυπο και θα πρέπει να επισημανθεί για περαιτέρω ανάλυση ή ενδεχομένως να αφαιρεθεί από το σύνολο δεδομένων.

Η μέθοδος ενδοτεταρτομοριακού εύρους για την ανίχνευση παράτυπων σημείων επιλέγεται πολλους λόγους. Πρώτον, παρέχει ένα ισχυρό μέτρο της κατανομής των δεδομένων, λαμβάνοντας υπόψη το κεντρικό 50% του συνόλου δεδομένων, ενώ αγνοεί τις ακραίες τιμές. Αυτό το χαρακτηριστικό το καθιστά λιγότερο ευαίσθητο στην παρουσία ακραίων τιμών σε σύγκριση με άλλα στατιστικά μέτρα, όπως ο μέσος όρος και η τυπική απόκλιση. Επιπλέον, είναι εύκολα ερμηνεύσιμη και ευρέως κατανοητή. Τα τεταρτημόρια και το IQR παρέχουν ουσιαστικές και διαισθητικές πληροφορίες σχετικά με τη κατανομή των δεδομένων και το εύρος εντός του οποίου βρίσκεται η πλειονότητα των δεδομένων. Προσδιορίζοντας ακραίες τιμές ως σημεία δεδομένων που εμπίπτουν εκτός αυτού του εύρους, κατεπέκταση, η μέθοδος προσφέρει ένα σαφές και απλό κριτήριο για τον εντοπισμό ανωμαλιών. Επίσης την ευελιξία στο χειρισμό των ακραίων τιμών. Ενώ η μέθοδος προσδιορίζει παράτυπες τιμές, δεν καθορίζει εγγενώς τον τρόπο χειρισμού τους. Οι ερευνητές μπορούν να επιλέξουν να αφαιρέσουν τα παράτυπα σημεία από το σύνολο δεδομένων εάν κριθούν λανθασμένα ή άσχετα με την ανάλυση. Εναλλακτικά, τα παράτυπα σημεία μπορούν να διατηρηθούν εάν αντιπροσωπεύουν σπάνια ή σημαντικά γεγονότα που δεν πρέπει να αγνοούνται από το μοντέλο.

Είναι σημαντικό να σημειωθεί ότι η μέθοδος ενδοτεταρτομοριακού εύρους έχει τους περιορισμούς της. Υποθέτει ότι τα δεδομένα ακολουθούν μια συμμετρική κατανομή και δεν έχει καλή απόδοση σε λοξές ή πολυτροπικές κατανομές. Σε τέτοιες περιπτώσεις, εναλλακτικές μέθοδοι ανίχνευσης παράτυπων σημείων ή μετασχηματισμοί μπορεί να είναι πιο κατάλληλοι. Επιπλέον, η μέθοδος IQR βασίζεται στην επιλογή του πολλαπλασιαστή 1,5 ως κατωφλίου για τον προσδιορισμό των ακραίων τιμών. Ενώ αυτή η τιμή χρησιμοποιείται συνήθως, μπορεί να προσαρμοστεί με βάση τις συγκεκριμένες απαιτήσεις και τα χαρακτηριστικά του συνόλου δεδομένων.

Συμπερασματικά, η μέθοδος ενδοτεταρτομοριακού εύρους παρέχει μια πρακτική και διαισθητική προσέγγιση για τον εντοπισμό ακραίων τιμών σε ένα σύνολο δεδομένων. Με τη χρήση των τεταρτημορίων και του ενδοτεταρτομοριακού εύρους, προσφέρει ένα ισχυρό μέτρο της εξάπλωσης των δεδομένων, ενώ είναι λιγότερο ευαίσθητο σε ακραίες τιμές. Η ερμηνευτικότητα της μεθόδου, η ανοχή σε ακραίες τιμές και η δυνατότητα ενσωμάτωσης με άλλες τεχνικές την καθιστούν πολύτιμο εργαλείο στο στάδιο της προεπεξεργασίας των δεδομένων. Οι ερευνητές μπορούν να αξιοποιήσουν τη μέθοδο ενδοτεταρτομοριακού εύρους για να ανιχνεύουν και να χειρίζονται αποτελεσματικά τα παράτυπα σημεία, βελτιώνοντας την ποιότητα και την αξιοπιστία των αναλύσεών τους και των μοντέλων μηχανικής μάθησης [25].

Η έννοια των τεταρτημορίων και του IQR μπορεί να παρουσιαστεί καλύτερα από το παρακάτω θηκόγραμμα που αναπαριστά την κατανομή ενός συνόλου δεδομένων αλλά και τα ποσοστιαία σημεία:

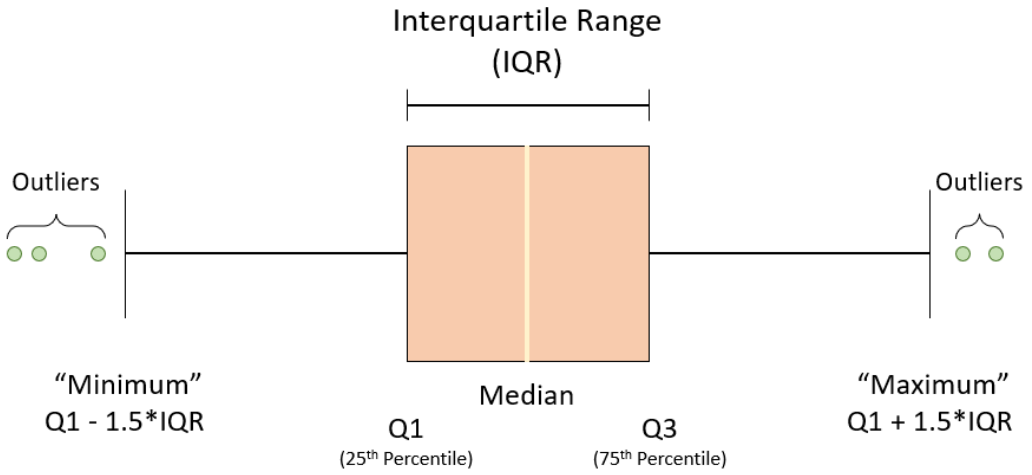


Figure 2: Interquartile Range (IQR) to Detect Outliers

3.2.3. Κανονικοποίηση (Scaling – MinMax Scaler)

Η κλιμάκωση, γνωστή και ως κανονικοποίηση δεδομένων, είναι μια θεμελιώδης τεχνική που χρησιμοποιείται στο στάδιο προεπεξεργασίας των δεδομένων στην μηχανική μάθηση. Σκοπός της είναι να μετασχηματίσει την κλίμακα των δεδομένων, διασφαλίζοντας ότι όλα τα χαρακτηριστικά βρίσκονται σε σταθερό εύρος διατηρώντας παράλληλα την αρχική τους κατανομή. Το σκεπτικό πίσω από την κλιμάκωση είναι να εξαλειφθούν οι προκαταλήψεις που μπορεί να προκύψουν κατά την εκτέλεση υπολογισμών ή συγκρίσεων μεταξύ χαρακτηριστικών με ανόμοιες κλίμακες. Φέρνοντας τα δεδομένα σε μια κοινή κλίμακα, τα μοντέλα μπορούν να αξιολογήσουν με ακρίβεια τη σημασία κάθε χαρακτηριστικού και να λάβουν τεκμηριωμένες αποφάσεις με βάση τη σχετική συνεισφορά τους.

Τα οφέλη της κλιμάκωσης είναι πολλά. Πρώτον, η κλιμάκωση βοηθά στην αποτροπή της κυριαρχίας χαρακτηριστικών με μεγάλα εύρη τιμών στη διαδικασία εκμάθησης. Εάν ορισμένα χαρακτηριστικά έχουν τιμές που εκτείνονται σε πολύ μεγαλύτερο εύρος από άλλα, τα μοντέλα μπορεί να αποδώσουν αδικαιολόγητα μεγάλη σημασία σε αυτά τα χαρακτηριστικά, επισκιάζοντας τις συνεισφορές άλλων. Η κλιμάκωση μετριάζει αυτό το πρόβλημα τοποθετώντας όλα τα χαρακτηριστικά σε συγκρίσιμη κλίμακα, επιτρέποντας στα μοντέλα να ζυγίζουν κάθε χαρακτηριστικό κατάλληλα με βάση την εγγενή του σημασία. Αυτό εξασφαλίζει μια δίκαιη και ισορροπημένη εξέταση όλων των χαρακτηριστικών κατά τη διάρκεια της εκπαιδευτικής διαδικασίας.

Δεύτερον, η κλιμάκωση διευκολύνει τη σύγκλιση και τη σταθερότητα πολλών αλγορίθμων βελτιστοποίησης που χρησιμοποιούνται στη μηχανική εκμάθηση. Αλγόριθμοι όπως η gradient descent, που στοχεύουν στην ελαχιστοποίηση του λάθους μεταξύ προβλέψεων και πραγματικών τιμών, βασίζονται σε αποτελεσματικές ενημερώσεις των παραμέτρων του μοντέλου. Όταν τα χαρακτηριστικά βρίσκονται σε διαφορετικές κλίμακες, η διαδικασία βελτιστοποίησης μπορεί να

παρεμποδιστεί, οδηγώντας σε αργή σύγκλιση ή ακόμα και αποτυχία σύγκλισης. Η κλιμάκωση των δεδομένων βοηθά στην άμβλυνση αυτών των προβλημάτων παρέχοντας ένα πιο ευνοϊκό τοπίο για βελτιστοποίηση, βελτιώνοντας την αποδοτικότητα και την αποτελεσματικότητα της μαθησιακής διαδικασίας.

Επιπλέον, η κλιμάκωση ενισχύει την ερμηνευτικότητα και την οπτική αναπαράσταση των δεδομένων. Όταν τα χαρακτηριστικά είναι στην ίδια κλίμακα, γίνεται πιο εύκολο να συγκριθούν και να κατανοηθούν οι σχέσεις τους οπτικά. Αυτό βοηθά στη διερευνητική ανάλυση δεδομένων, στην ερμηνεία του μοντέλου και στη λήψη αποφάσεων με βάση τις γνώσεις που αποκτήθηκε από τα δεδομένα. Η κλιμάκωση διασφαλίζει επίσης ότι το μέγεθος των συντελεστών ή βαρών που αποδίδονται σε κάθε χαρακτηριστικό είναι σημαντικό και ερμηνεύσιμο, επιτρέποντας στους ερευνητές να εξάγουν ουσιαστικά συμπεράσματα και να λαμβάνουν τεκμηριωμένες αποφάσεις με βάση τα αποτελέσματα του μοντέλου [26] [27].

Η θεωρητική βάση της κλιμάκωσης βρίσκεται στην έννοια της κανονικοποίησης χαρακτηριστικών. Η πιο συχνά χρησιμοποιούμενη τεχνική κλιμάκωσης είναι η ελάχιστη-μέγιστη κλίμακα ή η κανονικοποίηση στο εύρος $[0,1]$. Για κάθε χαρακτηριστικό, προσδιορίζονται η ελάχιστη και μέγιστη τιμή και, στη συνέχεια, εφαρμόζεται ένας μετασχηματισμός χρησιμοποιώντας τον ακόλουθο τύπο:

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

Αφαιρώντας την ελάχιστη τιμή από κάθε σημείο δεδομένων και διαιρώντας με το εύρος μεταξύ της ελάχιστης και της μέγιστης τιμής, τα δεδομένα μεταφέρονται στο επιθυμητό εύρος $[0,1]$. Είναι σημαντικό να σημειωθεί ότι αυτός ο μετασχηματισμός θα πρέπει να εφαρμόζεται ανεξάρτητα σε κάθε χαρακτηριστικό για να επιτευχθεί το επιθυμητό αποτέλεσμα κλιμάκωσης. Αυτή η διαδικασία κανονικοποίησης διασφαλίζει ότι κάθε χαρακτηριστικό έχει μια συνεπή κλίμακα, επιτρέποντας δίκαιες συγκρίσεις και υπολογισμούς μεταξύ διαφορετικών χαρακτηριστικών.

Άλλες τεχνικές κλιμάκωσης, όπως η τυποποίηση (επίσης γνωστή ως κανονικοποίηση βαθμολογίας z), επανακλιμακώνουν τα δεδομένα ώστε να έχουν μέσο όρο 0 και τυπική απόκλιση 1. Η τυποποίηση είναι ιδιαίτερα χρήσιμη όταν τα δεδομένα παρουσιάζουν κατανομή Gauss ή όταν απαιτείται από συγκεκριμένους αλγόριθμους που προϋποθέτουν τυποποιημένα χαρακτηριστικά.

Συμπερασματικά, η κλιμάκωση παίζει ζωτικό ρόλο στην προεπεξεργασία δεδομένων και στη μηχανική μάθηση. Φέρνοντας τα χαρακτηριστικά σε μια κοινή κλίμακα, εξαλείφει τις προκαταλήψεις και επιτρέπει δίκαιες αξιολογήσεις της σημασίας των χαρακτηριστικών. Τα οφέλη της κλιμάκωσης περιλαμβάνουν βελτιωμένη απόδοση μοντέλου, βελτιωμένη σύγκλιση αλγορίθμων βελτιστοποίησης, αυξημένη ερμηνευτικότητα των αποτελεσμάτων και καλύτερη οπτικοποίηση των σχέσεων δεδομένων. Οι ερευνητές και οι επαγγελματίες θα πρέπει να εξετάσουν προσεκτικά την κλιμάκωση ως ένα ουσιαστικό βήμα στη γραμμή προεπεξεργασίας των δεδομένων τους για να εξασφαλίσουν ακριβή και αξιόπιστα αποτελέσματα μοντελοποίησης.

3.3. Μοντέλα πρόβλεψης (Predictive Models)

Σε αυτό το κεφάλαιο θα περιγραφτεί η λειτουργικότητα των μοντέλων μηχανικής μάθησης που θα αξιοποιηθούν για την μοντελοποίηση του συγκεκριμένου προβλήματος.

3.3.1. Δεντρα Παλινδρόμησης (Regression Tree)

Ένα δέντρο παλινδρόμησης είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί για την πρόβλεψη αριθμητικών τιμών σε μια ποικιλία εφαρμογών. Ανήκει στην οικογένεια των δέντρων αποφάσεων και έχει σχεδιαστεί ειδικά για εργασίες παλινδρόμησης, όπου ο στόχος είναι να εκτιμηθεί μια μεταβλητή συνεχούς εξόδου με βάση ένα σύνολο χαρακτηριστικών εισόδου.

Στον πυρήνα του, ένα δέντρο παλινδρόμησης ακολουθεί μια διαδοχική προσέγγιση βασισμένη σε κανόνες, οργανωμένη σε δομή δέντρου. Το δέντρο αποτελείται από τρεις κύριους τύπους κόμβων: τον κόμβο ρίζας, τους εσωτερικούς κόμβους και τους κόμβους φύλλων. Ο ριζικός κόμβος αντιπροσωπεύει την αρχική κατάσταση του αλγορίθμου, που περιλαμβάνει ολόκληρο το σύνολο δεδομένων. Οι εσωτερικοί κόμβοι αντιστοιχούν σε συγκεκριμένα χαρακτηριστικά του συνόλου δεδομένων και περιέχουν κανόνες απόφασης, ενώ οι κόμβοι φύλλων αντιπροσωπεύουν τις τελικές προβλέψεις.

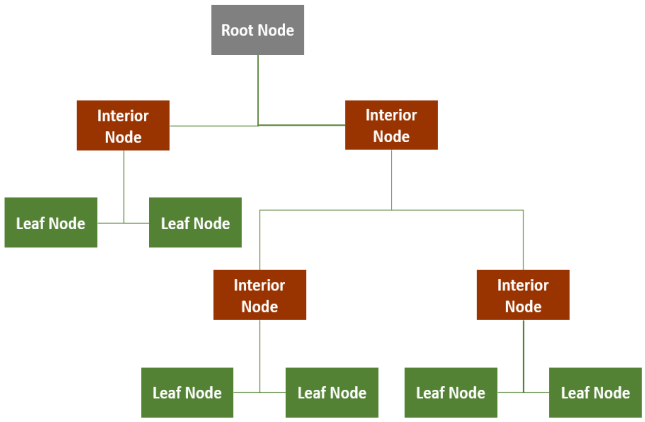


Figure 3: Decision Tree Architecture

Το στάδιο εκπαίδευσης ενός δέντρου παλινδρόμησης περιλαμβάνει τη διαίρεση του πολυδιάστατου χώρου εισόδου σε υποσύνολα με ιεραρχικό τρόπο. Αυτή η διαδικασία καθοδηγείται από μια σειρά ερωτήσεων που τίθενται στα δεδομένα, με κάθε ερώτηση να στοχεύει στη λήψη μιας δυαδικής απόφασης με βάση ένα συγκεκριμένο χαρακτηριστικό. Οι απαντήσεις σε αυτές τις ερωτήσεις, που συνήθως αντιπροσωπεύονται ως "Ναι" ή "Όχι", καθοδηγούν τα δεδομένα στη δομή του δέντρου μέχρι να φτάσουν σε έναν κόμβο φύλλου. Κάθε κόμβος φύλλου περιέχει μια προκαθορισμένη αριθμητική τιμή, η οποία επιλέγεται κατά τη διαδικασία εκπαίδευσης.

Για την κατασκευή ενός δέντρου παλινδρόμησης, ο αλγόριθμος επαναλαμβάνει τα χαρακτηριστικά του συνόλου δεδομένων. Για κάθε χαρακτηριστικό, αξιολογεί διαφορετικά σημεία διαχωρισμού για να καθορίσει το καταλληλότερο σημείο τομής. Η επιλογή του σημείου διαίρεσης βασίζεται σε μια μέτρηση διαχωρισμού, όπως η ελαχιστοποίηση του σφάλματος ή της απόστασης μεταξύ των προβλεπόμενων και των πραγματικών τιμών σε κάθε υποσύνολο. Επιλέγεται το χαρακτηριστικό που επιτυγχάνει τον καλύτερο διαχωρισμό και τα δεδομένα χωρίζονται ανάλογα σε δύο διακριτές ομάδες.

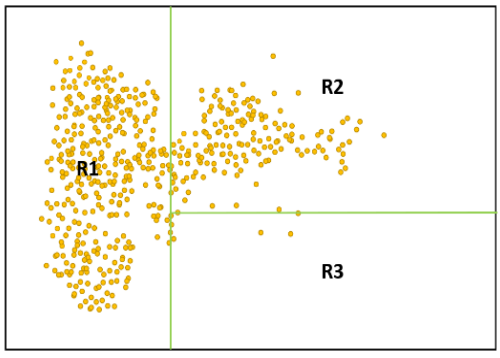


Figure 4: Decision tree Regressor - Inference

Με τη διαίρεση των δεδομένων σε υποσύνολα, ο αλγόριθμος προχωρά στον υπολογισμό της μέσης τιμής της μεταβλητής στόχου σε κάθε υποσύνολο. Αυτή η μέση τιμή χρησιμεύει ως η προβλεπόμενη αριθμητική τιμή για τυχόν

μελλοντικά δεδομένα που εμπίπτουν στο ίδιο υποσύνολο. Ο στόχος είναι να βρεθεί το χαρακτηριστικό που μεγιστοποιεί την ανομοιότητα μεταξύ των υποσυνόλων ως προς τις μέσες τιμές τους, δημιουργώντας έτσι διακριτές περιοχές για πρόβλεψη.

Με την επανάληψη αυτής της διαδικασίας επαναληπτικά, το δέντρο παλινδρόμησης προσθέτει περισσότερους κανόνες και αυξάνει το βάθος του, διαιρώντας περαιτέρω το σύνολο δεδομένων σε μικρότερες υποπεριοχές. Αυτή η επαναληπτική κατάτμηση και ανάθεση πρόβλεψης επιτρέπει στο δέντρο να καταγράφει σύνθετες σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Το βάθος του δέντρου καθορίζει τον αριθμό των ερωτήσεων που απαιτούνται για την επίτευξη μιας πρόβλεψης και μπορεί να ελεγχθεί για να αποφευχθεί η υπερβολική προσαρμογή θέτοντας ένα κριτήριο διακοπής.

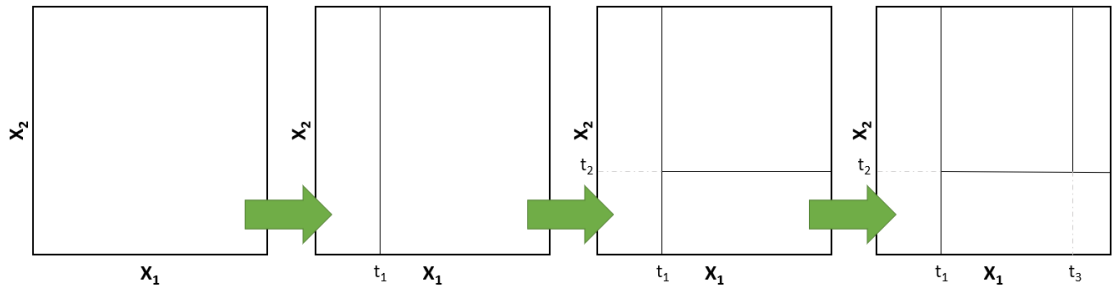


Figure 5: Decision Tree Regressor - Splitting the Data Process

Κατά τη φάση της πρόβλεψης, ένα νέο σημείο δεδομένων τροφοδοτείται στο εκπαιδευμένο δέντρο παλινδρόμησης. Ξεκινά από τον ριζικό κόμβο και ακολουθεί τους κανόνες απόφασης σε κάθε εσωτερικό κόμβο, με βάση τις τιμές των χαρακτηριστικών. Η διαδρομή μέσα από το δέντρο οδηγεί τα δεδομένα σε έναν συγκεκριμένο κόμβο φύλλου, όπου η προκαθορισμένη αριθμητική τιμή που έχει εκχωρηθεί σε αυτόν τον κόμβο φύλλου γίνεται η τελική πρόβλεψη για την είσοδο.

Τα δέντρα παλινδρόμησης προσφέρουν πολλά πλεονεκτήματα στη μηχανική μάθηση. Είναι εύκολο να κατανοηθούν και να ερμηνευτούν, παρέχοντας σαφείς γνώσεις σχετικά με τη διαδικασία λήψης αποφάσεων. Επιπλέον, μπορούν να χειριστούν τόσο κατηγορίες όσο και αριθμητικά χαρακτηριστικά χωρίς να απαιτείται εκτεταμένη προεπεξεργασία. Επιπλέον, τα δέντρα παλινδρόμησης μπορούν να συλλάβουν μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου, καθιστώντας τα κατάλληλα για ένα ευρύ φάσμα εφαρμογών.

Ωστόσο, είναι σημαντικό να σημειωθεί ότι τα δέντρα παλινδρόμησης είναι επιρρεπή σε υπερβολική προσαρμογή, ειδικά όταν το δέντρο γίνεται πολύ βαθύ και πολύπλοκο. Η υπερπροσαρμογή συμβαίνει όταν το μοντέλο ταιριάζει υπερβολικά στα δεδομένα εκπαίδευσης, με αποτέλεσμα κακή γενίκευση σε μη ορατά δεδομένα. Τεχνικές όπως το κλάδεμα, το οποίο περιλαμβάνει την αφαίρεση ή την κατάρρευση ορισμένων κόμβων στο δέντρο ή τη χρήση μεθόδων συνόλου όπως τα τυχαία δάση, μπορούν να βοηθήσουν στον μετριασμό της υπερβολικής προσαρμογής και στη βελτίωση της προγνωστικής απόδοσης των δέντρων παλινδρόμησης.

Συνοπτικά, ένα δέντρο παλινδρόμησης είναι ένας ευέλικτος αλγόριθμος μηχανικής μάθησης για την πρόβλεψη αριθμητικών τιμών. Διαχωρίζοντας αναδρομικά το σύνολο δεδομένων με βάση τις τιμές χαρακτηριστικών και εκχωρώντας μέσες προβλέψεις σε κάθε υποσύνολο, το δέντρο παλινδρόμησης καταγράφει σχέσεις και μοτίβα στα δεδομένα. Η διαδοχική δομή του που βασίζεται σε κανόνες παρέχει μια ερμηνεύσιμη και διαισθητική προσέγγιση για την πραγματοποίηση προβλέψεων. Ωστόσο, θα πρέπει να δοθεί ιδιαίτερη προσοχή στον έλεγχο του βάθους και της πολυπλοκότητας του δέντρου για να αποφευχθεί η υπερβολική προσαρμογή και να ενισχυθούν οι δυνατότητες γενίκευσης [28] [29].

3.3.2. Τυχαία Δάση (Random Forest)

Ο αλγόριθμος τυχαίων δασών είναι μια ευέλικτη και ισχυρή μέθοδος εκμάθησης συνόλου που συνδυάζει πολλαπλά δέντρα αποφάσεων για να κάνει ακριβείς προβλέψεις. Προτάθηκε για πρώτη φορά από τον Tin Kam Ho το 1995 [30] και έκτοτε έχει γίνει ένας από τους πιο ευρέως χρησιμοποιούμενους αλγόριθμους μηχανικής μάθησης.

Στον πυρήνα του, ένα τυχαίο δάσος αποτελείται από ένα σύνολο δέντρων απόφασης. Κάθε δέντρο στο δάσος χτίζεται ανεξάρτητα, χρησιμοποιώντας ένα υποσύνολο των δεδομένων εκπαίδευσης και ένα υποσύνολο των διαθέσιμων χαρακτηριστικών. Η τελική πρόβλεψη του τυχαίου δάσους προκύπτει από τη συγκέντρωση των προβλέψεων όλων των μεμονωμένων δέντρων.

Η κύρια ιδέα πίσω από τον αλγόριθμο τυχαίων δασών είναι ότι ένα διαφορετικό σύνολο δέντρων αποφάσεων, που συνεργάζονται ως επιτροπή, μπορούν να παράγουν ακριβέστερες προβλέψεις από οποιοδήποτε μεμονωμένο δέντρο. Συνδυάζοντας τις προβλέψεις πολλών δέντρων, το τυχαίο δάσος μπορεί να χειριστεί αποτελεσματικά πολύπλοκα μοτίβα και θόρυβο στα δεδομένα, οδηγώντας σε βελτιωμένη απόδοση γενίκευσης.

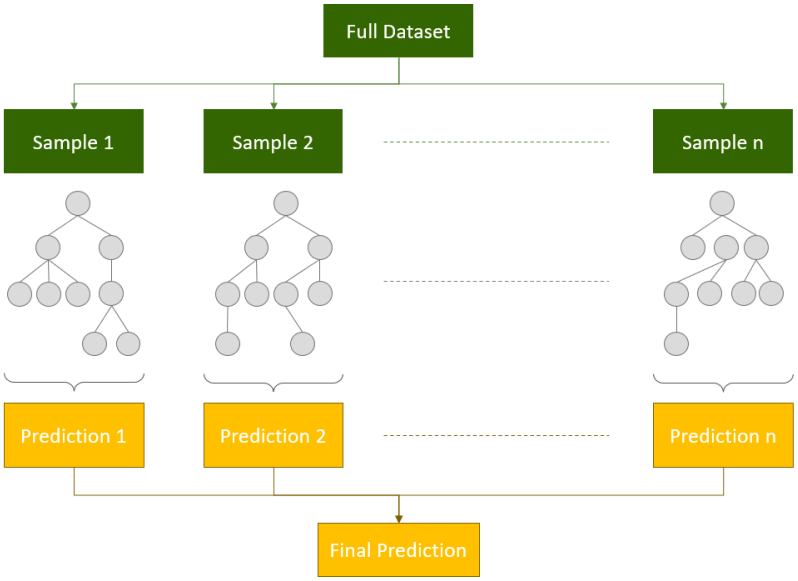


Figure 6: Random Forest Architecture

Για να δημιουργηθεί ένα τυχαίο δάσος, ο αλγόριθμος περνά από διάφορα βήματα:

1. **Δειγματοληψία δεδομένων:** Ένα τυχαίο υποσύνολο των δεδομένων εκπαίδευσης επιλέγεται για κάθε δέντρο στο δάσος. Αυτή η δειγματοληψία πραγματοποιείται με αντικατάσταση, μια τεχνική γνωστή ως bagging ή bootstrap aggregating. Η δειγματοληψία με αντικατάσταση σημαίνει ότι κάθε υποσύνολο μπορεί να περιέχει διπλότυπα στιγμιότυπα και ορισμένες παρουσίες μπορεί να παραμείνουν εντελώς εκτός. Αυτή η διαδικασία δειγματοληψίας εισάγει μεταβλητότητα στα δεδομένα εκπαίδευσης για κάθε δέντρο, καθιστώντας τα ανεξάρτητα το ένα από το άλλο.
2. **Δειγματοληψία χαρακτηριστικών:** Εκτός από τη δειγματοληψία των δεδομένων, ο αλγόριθμος τυχαίων δασών επιλέγει επίσης τυχαία ένα υποσύνολο χαρακτηριστικών για κάθε δέντρο. Αυτή η επιλογή εκτελείται συνήθως χωρίς αντικατάσταση, όπου κάθε δέντρο έχει πρόσβαση μόνο σε ένα κλάσμα των συνολικών χαρακτηριστικών. Ο

αριθμός των χαρακτηριστικών που λαμβάνονται υπόψη για κάθε δέντρο καθορίζεται από μια παράμετρο ή ευρετική που ορίζει ο χρήστης. Αυτή η τεχνική είναι γνωστή ως χαρακτηριστικές σακούλες ή μέθοδος τυχαίου υποχώρου. Χρησιμοποιώντας διαφορετικά υποσύνολα χαρακτηριστικών, το τυχαίο δάσος προωθεί την ποικιλομορφία μεταξύ των δέντρων, επιτρέποντάς τους να συλλάβουν διαφορετικές πτυχές των δεδομένων.

3. **Κατασκευή δέντρων:** Με τα δεδομένα και τα χαρακτηριστικά του δείγματος, κάθε δέντρο στο τυχαίο δάσος κατασκευάζεται ανεξάρτητα. Η διαδικασία κατασκευής ακολουθεί τον τυπικό αλγόριθμο δέντρου αποφάσεων, όπως ο αλγόριθμος ID3, C4.5 ή CART. Σε κάθε κόμβο του δέντρου, ο αλγόριθμος αναζητά το καλύτερο χαρακτηριστικό και το καλύτερο σημείο διαχωρισμού για να διαιρέσει τα δεδομένα με βάση ένα συγκεκριμένο κριτήριο, χρησιμοποιώντας συχνά μετρήσεις όπως κέρδος πληροφοριών ή ακαθαρσία Gini. Η διαδικασία διαχωρισμού συνεχίζεται αναδρομικά μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής, όπως η επίτευξη ενός μέγιστου βάθους ή ενός ελάχιστου αριθμού δειγμάτων σε έναν κόμβο φύλλου.
4. **Συνάθροιση δέντρων:** Μόλις χτιστούν όλα τα δέντρα, το τυχαίο δάσος συνδυάζει τις προβλέψεις τους για να αποκτήσει την τελική πρόβλεψη. Για εργασίες ταξινόμησης, η πιο κοινή μέθοδος συγκέντρωσης είναι η πλειοψηφία, όπου η κλάση που λαμβάνει τις περισσότερες ψήφους από τα μεμονωμένα δέντρα επιλέγεται ως η προβλεπόμενη τάξη. Για εργασίες παλινδρόμησης, οι προβλέψεις των μεμονωμένων δέντρων συνήθως υπολογίζονται κατά μέσο όρο για να ληφθεί μια συνεχής τιμή. Η τελική πρόβλεψη αντιπροσωπεύει τη συναίνεση μεταξύ των δέντρων, αξιοποιώντας τη συλλογική τους γνώση και μειώνοντας τον αντίκτυπο των μεροληψιών ή σφαλμάτων κάθε μεμονωμένου δέντρου.

Τα τυχαία δάση προσφέρουν πολλά πλεονεκτήματα σε σχέση με τα δέντρα μεμονωμένων αποφάσεων και άλλους αλγόριθμους μηχανικής μάθησης:

1. **Ανθεκτικότητα στην υπερπροσαρμογή**: Τα τυχαία δάση είναι λιγότερο επιρρεπή σε υπερπροσαρμογή σε σύγκριση με μεμονωμένα δέντρα απόφασης. Η προσέγγιση του συνόλου μειώνει τη διακύμανση του μοντέλου υπολογίζοντας τον μέσο όρο των προβλέψεων πολλών δέντρων, γεγονός που βοηθά στην εξομάλυνση του θορύβου και των ακραίων τιμών που υπάρχουν στα δεδομένα. Αυτό οδηγεί σε βελτιωμένη απόδοση γενίκευσης σε αόρατα δεδομένα.
2. **Χειρισμός δεδομένων υψηλών διαστάσεων**: Τα τυχαία δάση μπορούν να χειριστούν αποτελεσματικά σύνολα δεδομένων με μεγάλο αριθμό χαρακτηριστικών. Επιλέγοντας τυχαία υποσύνολα χαρακτηριστικών για κάθε δέντρο, ο αλγόριθμος εστιάζει σε διαφορετικά υποσύνολα μεταβλητών και μειώνει τον κίνδυνο συμπερίληψης άσχετων ή περιττών χαρακτηριστικών στο μοντέλο. Αυτό μπορεί να βοηθήσει στην αποφυγή υπερβολικής προσαρμογής και στη βελτίωση της αποτελεσματικότητας του αλγορίθμου.
3. **Εκτίμηση σημαντικότητας χαρακτηριστικών**: Τα τυχαία δάση παρέχουν ένα μέτρο της σημασίας των χαρακτηριστικών, υποδεικνύοντας τη σχετική σημασία κάθε χαρακτηριστικού για την πραγματοποίηση προβλέψεων. Αυτές οι πληροφορίες μπορεί να είναι πολύτιμες για την επιλογή χαρακτηριστικών, προσδιορίζοντας τις πιο σχετικές δυνατότητες για την εργασία. Η σημασία του χαρακτηριστικού συνήθως υπολογίζεται αναλύοντας τη μέση μείωση της ακρίβειας πρόβλεψης όταν ένα συγκεκριμένο χαρακτηριστικό μετατίθεται τυχαία.
4. **Σύλληψη μη γραμμική σχέσεων**: Τα τυχαία δάση μπορούν να καταγράψουν μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου.

Συνδυάζοντας πολλαπλά δέντρα αποφάσεων, το καθένα ικανό να συλλαμβάνει διαφορετικές πτυχές των δεδομένων, ο αλγόριθμος τυχαίων δασών μπορεί να μοντελοποιήσει πολύπλοκες αλληλεπιδράσεις και εξαρτήσεις μεταξύ των μεταβλητών. Αυτό το καθιστά κατάλληλο για εργασίες όπου η υποκείμενη σχέση είναι μη γραμμική ή παρουσιάζει αλληλεπιδράσεις μεταξύ χαρακτηριστικών.

5. **Outlier και Noise Handling**: Τα τυχαία δάση είναι ανθεκτικά σε ακραίες τιμές και θόρυβο στα δεδομένα. Δεδομένου ότι κάθε δέντρο είναι χτισμένο σε διαφορετικό υποσύνολο δεδομένων, ο αντίκτυπος μεμονωμένων θορυβωδών ή ακραίων περιπτώσεων μειώνεται. Οι ακραίες τιμές είναι πιθανό να απομονώνονται σε διαφορετικά υποσύνολα και να έχουν περιορισμένη επίδραση στην τελική πρόβλεψη λόγω του βήματος συνάθροισης.
6. **Παραλληλισμός**: Τα στάδια κατασκευής και πρόβλεψης των τυχαίων δασών μπορούν να παραλληλιστούν, καθιστώντας τα κατάλληλα για κατανεμημένα υπολογιστικά περιβάλλοντα. Κάθε δέντρο στο δάσος μπορεί να κατασκευαστεί ανεξάρτητα, επιτρέποντας την αποτελεσματική χρήση των υπολογιστικών πόρων.

Είναι σημαντικό να σημειωθεί ότι η απόδοση ενός τυχαίου δάσους μπορεί να επηρεαστεί από διάφορες παραμέτρους, όπως ο αριθμός των δέντρων στο σύνολο, το βάθος κάθε δέντρου, ο αριθμός των χαρακτηριστικών που λαμβάνονται υπόψη σε κάθε διάσπαση και το συγκεκριμένο κριτήριο διαχωρισμού που χρησιμοποιείται. Αυτές οι παράμετροι μπορούν να βελτιστοποιηθούν χρησιμοποιώντας τεχνικές όπως η διασταυρούμενη επικύρωση ή η αναζήτηση πλέγματος για τη βελτιστοποίηση της απόδοσης του μοντέλου.

Τα τυχαία δάση έχουν κερδίσει δημοτικότητα σε διάφορους τομείς και έχουν εφαρμοστεί με επιτυχία σε ένα ευρύ φάσμα προβλημάτων, όπως, ενδεικτικά, η ταξινόμηση εικόνων, η ανάλυση κειμένου, ο εντοπισμός απάτης, η βιοπληροφορική και τα συστήματα συστάσεων. Κατατάσσονται σταθερά μεταξύ των κορυφαίων αλγορίθμων μηχανικής μάθησης και θεωρούνται ευρέως ως ένα αξιόπιστο και ισχυρό εργαλείο για προγνωστική μοντελοποίηση.

Συνοπτικά, ο αλγόριθμος τυχαίων δασών είναι μια μέθοδος εκμάθησης συνόλου που συνδυάζει πολλαπλά δέντρα αποφάσεων για να κάνει ακριβείς προβλέψεις. Χρησιμοποιώντας δεδομένα και δειγματοληψία χαρακτηριστικών, δημιουργεί ένα ποικίλο σύνολο δέντρων που συνεργάζονται ως επιτροπή για να παράγουν ισχυρές προβλέψεις. Τα τυχαία δάση προσφέρουν βελτιωμένη απόδοση γενίκευσης, χειρίζονται δεδομένα υψηλών διαστάσεων, εκτιμούν τη σημασία των χαρακτηριστικών, καταγράφουν μη γραμμικές σχέσεις, χειρίζονται ακραίες τιμές και θόρυβο και μπορούν να παραλληλιστούν για αποτελεσματικούς υπολογισμούς. Έχουν αποδειχθεί ένα ευέλικτο και ισχυρό εργαλείο στον τομέα της μηχανικής μάθησης, επιτυγχάνοντας ανταγωνιστικά αποτελέσματα σε διάφορους τομείς [31].

3.3.3. Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine - SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) (Support Vector Machines - SVM) προτάθηκαν από τον Vladimir Vapnik ως γραμμικοί ταξινομητές, το 1963, αλλά απέκτησαν δημοσιότητα μετά το 1992, όταν ενισχύθηκαν με το κόλπο του πυρήνα (kernel trick), που επέτρεψε τη χρήση τους και σε μη γραμμικώς διαχωρίσιμα προβλήματα. Είναι μέθοδος για προβλήματα ταξινόμησης αλλά έχουν επεκταθεί και για παρεμβολή (Support Vector Regression - SVR). Στη δεκαετία του 2000, οι Μηχανές Διανυσμάτων Υποστήριξης εδραιώθηκαν ως μια από τις πιο διαδεδομένες μεθόδους (γραμμικής και μη) ταξινόμησης και παρεμβολής, αποτελώντας, μέχρι την έλευση των Deep-NN, τη βέλτιστη επιλογή για εφαρμογές όπως [32]:

- Αναγνώριση γραφής (handwriting recognition)

- Ταξινόμηση κειμένων (text categorization)
- Ταξινόμηση δεδομένων έκφρασης γονιδίων (gene expression data)
- Και γενικά για εφαρμογές με δεδομένα πολλών διαστάσεων

Η μέθοδος SVM επιδιώκει να βρει το σύνορο που απέχει όσο το δυνατόν περισσότερο από τα παραδείγματα των κλάσεων που διαχωρίζει. Υπάρχουν άπειρα επίπεδα διαχωρισμού, αλλά τα SVM ψάχνουν εκείνο που δεινή το μεγαλύτερο περιθώριο σφαλμάτων. Το σύνορο αυτό μπορεί να είναι:

- Γραμμή για δισδιάστατους χώρους
- Επίπεδο για τρισδιάστατους χώρους
- Υπερεπίπεδο για πολυδιάστατους χώρους

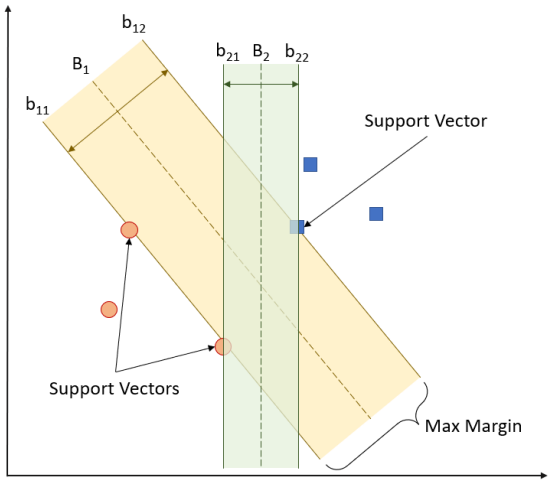


Figure 7: SVM Maximum Margin Hypersurface

Το σύνορο αυτό ονομάζεται υπερεπιφάνεια μεγίστου περιθωρίου (maximum margin hypersurface) και σε γραμμικώς διαχωρίσιμα προβλήματα ορίζεται από έναν πεπερασμένο αριθμό παραδειγμάτων του συνόλου εκπαίδευσης που ονομάζονται διανύσματα υποστήριξης (support vectors), εξου και η ονομασία του συγκεκριμένου αλγορίθμου. Για την εύρεση τους χρησιμοποιούνται αλγόριθμοι βελτιστοποίησης προβλημάτων τετραγωνικού προγραμματισμού (constrained quadratic optimization).

Επιπλέον, μέσω των συναρτήσεων πυρήνα (kernel functions), μπορούν να μετασχηματίσουν τον αρχικό χώρο του συνόλου δεδομένων έτσι ώστε μη-γραμμικώς διαχωρίσιμα προβλήματα να μετατραπούν σε γραμμικώς διαχωρίσιμα και τελικά να λυθούν με την ίδια μεθοδολογία. Επομένως, το κύριο πλεονέκτημα των SVM είναι ότι όχι μόνο δημιουργούν ένα διαχωριστικό σύνορο μεταξύ των επιμέρους κλάσεων (όπως άλλωστε κάνουν οι περισσότερες μέθοδοι ταξινόμησης), αλλά φροντίζουν αυτό το σύνορο να απέχει όσο το δυνατό περισσότερο από τα παραδείγματα των κλάσεων που διαχωρίζουν. Τα σύνορα απόφασης με μεγάλα περιθώρια έχουν κατά βάση μεγαλύτερη ανοχή σε φαινόμενα υπερπροσαρμογής. Ακόμη κι αν τα διανύσματα υποστήριξης είναι παραδείγματα με λίγο θόρυβο. Αν το περιθώριο είναι μικρό, τότε μια μικρή διαταραχή (π.χ. εξαιτίας θορύβου στα δεδομένα εκπαίδευσης) μπορεί να προκαλέσει σημαντική πτώση στην απόδοση του ταξινομητή.

Γραμμικώς Διαχωρίσιμα Προβλήματα

Έστω ένα σύνολο από πολυδιάστατα σημεία $x = \{x_1, x_2, \dots, x_n\}$ τα οποία ανήκουν στην κλάση $(+, τεράγωνα)$ ή στην κλάση $(-, κύκλοι)$, τις οποίες έστω ότι κωδικοποιούμε με 1 και −1. Έστω ότι το πρόβλημα είναι γραμμικώς διαχωρίσιμο σε δύο διαστάσεις (**Figure 16**). Το σύνορο θα έχει τη μορφή $w_1x_1 + w_2x_2 + b = 0$ ή $w \cdot x + b = 0$ με διανυσματική γραφή για τα βάρη και το σημείο. Εφαρμόζοντας τη σχέση αυτή για δύο σημεία α και β που βρίσκονται πάνω στο σύνορο και αφαιρώντας τις δύο σχέσεις προκύπτει: $w \cdot (x_\alpha - x_\beta) = 0$ που σημαίνει ότι το διάνυσμα των βαρών είναι κάθετο στο ζητούμενο σύνορο. Σε γραμμικώς διαχωρίσιμα προβλήματα, μπορούμε να ορίσουμε δύο επιπλέον σύνορα εκατέρωθεν του $w \cdot x + b = 0$ που επίσης χωρίζουν τις δύο κλάσεις και απέχουν όσο γίνεται περισσότερο μεταξύ τους. Η περιοχή που οριοθετούν αυτές οι δύο ευθείες (υπερεπιφάνειες, στη γενική περίπτωση) είναι το ζητούμενο μέγιστο περιθώριο (margin) και η ευθεία $w \cdot x + b = 0$ (γενικά υπερεπιφάνεια) βρίσκεται στη μέση αυτού του περιθωρίου. Η κλάση y ενός άγνωστου σημείου (δεδομένου) θα είναι:

$$y = \begin{cases} 1, & \alpha\nu\ w \cdot z + b > 0 \\ -1, & \alpha\nu\ w \cdot z + b < 0 \end{cases}$$

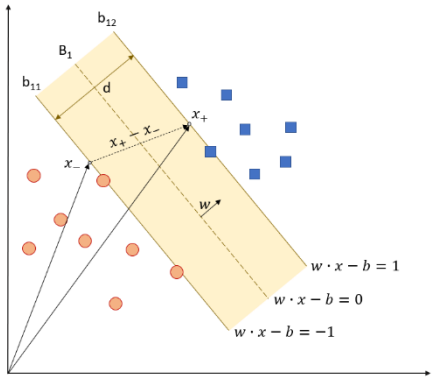


Figure 8: Linear SVM

Έστω x_+ και x_- είναι σημεία στα πάνω και κάτω όρια του περιθωρίου αντίστοιχα (Figure 8). Καθώς το διάνυσμα w είναι κάθετο στο ζητούμενο σύνορο, διαιρώντας το με το μήκος του $\|w\|$ προκύπτει ένα μοναδιαίο διάνυσμα που πολλαπλασιαζόμενο (εσωτερικό γινόμενο) με το διάνυσμα της διαφοράς $(x_+ - x_-)$, δίνει το ζητούμενο d: $(x_+ - x_-)\frac{w}{\|w\|} = d$ (Σε αυτή τη σχέση, για το $x_+ \cdot w$, η σχέση υπολογισμού του y δίνει $1 - b$ ενώ για το $x_- \cdot w$ δίνει $1 + b$. Αντικαθιστώντας στην προηγούμενη σχέση προκύπτει: $d = \frac{2}{\|w\|}$). Άρα η μεγιστοποίηση του d ισούται με την ελαχιστοποίηση του $\|w\|$.

Άρα η διαδικασία εκπαίδευσης της SVM μπορεί να οριστεί μαθηματικά ως η ελαχιστοποίηση του $\|w\|$ που μπορεί να γραφεί και ως: $\min_w \frac{\|w\|^2}{w}$.

Η σχέση αυτή. μαζί με τους περιορισμούς που ορίζει η έκφραση για τα y, αποτελεί τη μαθηματική διατύπωση της διαδικασίας εκπαίδευσης του SVM. Η επίλυση του προβλήματος γίνεται με τους πολλαπλασιαστές Lagrange (Lagrange multipliers) και προκύπτει ότι τα w και b που ορίζουν το σύνορο απόφασης εξαρτώνται μόνο από διανύσματα (σημεία) που βρίσκονται πάνω στις ευθείες b_1 και b_2, και από εσωτερικά γινόμενα αυτών. Τα διανύσματα αυτά ονομάζονται διανύσματα υποστήριξης (support vectors).

Soft Margin

Η κλασσική μέθοδος SVM αναζητά εκείνο το περιθώριο που διαχωρίζει όλα τα δεδομένα της μιας κλάσης από τα δεδομένα της άλλης. Σε προβλήματα του πραγματικού κόσμου όμως, μπορεί να μην είναι όλες οι παρατηρήσεις γραμμικά διαχωρίσιμες ή κάποια δεδομένα να έχουν λάθος ετικέτα ή να είναι πολύ σπάνια. Για την αντιμετώπιση του προβλήματος του απόλυτου γραμμικού

διαχωρισμού, οι Corinna Cortes and Vapnik πρότειναν το 1995 τη χρήση ενός χαλαρού περιθωρίου (soft margin) που επιτρέπει μερικά παραδείγματα να αγνοηθούν ή να τοποθετηθούν στη λάθος πλευρά ώστε να είναι εφικτή η εφαρμογή των SVM και σε αυτές τις περιπτώσεις. Αυτό το «λάθος» συχνά καταλήγει σε καλύτερα μοντέλα. Η διαδικασία υπολογισμού των w και b είναι πάλι παρόμοια με πριν, με τη διαφορά ότι η συνάρτηση απώλειας (loss ή cost function) της διαδικασίας βελτιστοποίησης του περιθωρίου περιέχει έναν επιπλέον όρο που ελέγχει τα σφάλματα του ταξινομητή. Αυτό το περιθώριο καθορίζεται από την παράμετρο C της συνάρτησης κόστους που ελέγχει την επίδραση κάθε διανύσματος υποστήριξης [33].

Kernel Trick

Το 1992 η ιδέα των SVM επεκτάθηκε και σε μη γραμμικούς ταξινομητές, εφαρμόζοντας κάποιο μετασχηματισμό $\Phi(x)$ στα δεδομένα έτσι ώστε: Το ζητούμενο μη γραμμικό σύνορο στον αρχικό χώρο, να καταστεί γραμμικό στο μετασχηματισμένο. Έτσι πλέον ήταν δυνατό να χρησιμοποιηθεί η προηγούμενη μεθοδολογία. Στο Figure 9 απεικονίζεται η μετατροπή μη γραμμικού προβλήματος ταξινόμησης (αριστερά) σε γραμμικό (δεξιά), με πολυωνυμικό μετασχηματισμό 2ου βαθμού. Όπως και στα γραμμικά SVM έτσι και τώρα, η υπερεπιφάνεια σύνορο εξαρτάται από διανύσματα υποστήριξης, ορισμένα όμως στο μετασχηματισμένο χώρο. Οι εξισώσεις για τον υπολογισμό του w και του b εμπεριέχουν και πάλι εσωτερικά γινόμενα μεταξύ ζευγαριών διανυσμάτων στον μετασχηματισμένο χώρο, δηλαδή εκφράσεις της μορφής $\Phi(u) \cdot \Phi(v)$. Οι υπολογισμοί όμως είναι περισσότερο πολύπλοκοι καθώς πέρα από τους απαιτούμενους μετασχηματισμούς, συνήθως απαιτείται και η μετάβαση σε περισσότερες διαστάσεις. Επιπρόσθετα δεν είναι ξεκάθαρο τι είδους μετασχηματισμός απαιτείται κάθε φορά. Τα παραπάνω προβλήματα αντιμετωπίζονται με το κόλπο του πυρήνα (kernel trick).

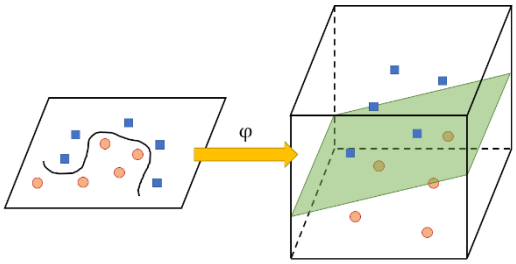


Figure 9: SVM Kernel Trick

Ο υπολογισμός του συνόρου των κλάσεων εμπεριέχει εσωτερικά γινόμενα με τα διανύσματα υποστήριξης. Το ίδιο ισχύει και στο μετασχηματισμένο χώρο, μόνο που η ανάλυση δείχνει ότι το εσωτερικό γινόμενο $\Phi(u) \cdot \Phi(v)$ των μετασχηματισμένων διανυσμάτων μπορεί να εκφραστεί ως συνάρτηση του εσωτερικού γινομένου των ίδιων διανυσμάτων στον αρχικό χώρο, του $u \cdot v$. Δηλαδή, είναι: $\Phi(u) \cdot \Phi(v) = f\,(u \cdot v)$. Η συνάρτηση f ονομάζεται συνάρτηση πυρήνα (kernel function) και συμβολίζεται με $K(u, v)$. Δηλαδή, μια συνάρτηση πυρήνα για δύο διανύσματα του αρχικού χώρου, είναι ισοδύναμη με το εσωτερικό γινόμενο των διανυσμάτων στο μετασχηματισμένο χώρο. Δηλαδή, αντί να προσδιοριστεί η συνάρτηση μετασχηματισμού Φ, τώρα αρκεί να προσδιοριστεί η συνάρτηση πυρήνα K. Έτσι, ο υπολογισμός των εσωτερικών γινομένων γίνεται στον αρχικό χώρο που είναι λιγότερων διαστάσεων και απαιτούνται λιγότεροι υπολογισμοί και κατ' επέκταση, μας απαλλάσσει από την ανάγκη προσδιορισμού της συνάρτησης Φ του μετασχηματισμού. Παραδείγματα τέτοιων συναρτήσεων είναι [34] [35]:

- **Dot**, εσωτερικού γινομένου ή ταυτοτική: $K(u, v) = u\,v$. Αυτός ο πυρήνας ισοδυναμεί με τη χρήση ενός SVM χωρίς πυρήνα
- **Polynomial**, συναρτήσεις βαθμού d: $K(u, v) = (u \cdot v + 1)^d$, όπου d ο βαθμός του πολυωνύμου
- **Radial**, συναρτήσεις ακτινικής βάσης - radial basis functions: $K(u, v) = e^{-g \cdot \|u-v\|^2}$, όπου g μια σταθερά, εξαρτώμενη από το πρόβλημα

- **Σιγμοειδείς**, χρησιμοποιούνται και στα Νευρωνικά Δίκτυα, όπως η υπερβολική εφαπτομένη $K(x, x') = tanh(ax^T x' + b)$.

Support Vector Regression (SVR)

Εδώ, στόχος είναι να βρεθεί μια συνάρτηση που προσεγγίζει τα σημεία εκπαίδευσης ελαχιστοποιώντας το σφάλμα πρόβλεψης. Η συνάρτηση απώλειας στην SVR δίνει έμφαση στο να μειώσει τους συντελεστές της ζητούμενης υπερεπιφάνειας μέσω $L2\ norm$, κάτι που περιορίζει την υπερπροσαρμογή. Ακόμα, προστίθενται περιορισμοί που ορίζουν ένα αποδεκτό όριο σφάλματος ε, εκατέρωθεν της ζητούμενης συνάρτησης παρεμβολής. Για τα σημεία που βρίσκονται εντός αυτού του ορίου, αλλά δεν είναι «πάνω» στη συνάρτηση, δεν επιβάλλεται καμία ποινή. Για το λόγο αυτό, η περιοχή εντός του ορίου σφάλματος ε, ονομάζεται και ε-σωλήνας (ε-tube). Με άλλα λόγια, «αγνοούμε» τα σημεία όπου η προβλεπόμενη τιμή y_i της συνάρτησης παρεμβολής για αυτά απέχει λιγότερο από ε από την πραγματική τους τιμή. Έτσι, αποφεύγουμε να κάνουμε υπερπροσαρμογή στα δεδομένα. Προφανώς, ένα τέτοιο μοντέλο, δεν είναι δυνατό να καλύπτει όλα τα δεδομένα, οπότε κάποια θα βρίσκονται εκτός των περιθωρίων (margins) που ορίζει το ε (λέγεται και μέθοδος ε-SVR). Σε αντιστοιχία με το χαλαρό περιθώριο (soft margin) στην ταξινόμηση, μπορεί να οριστεί ένα επιπλέον περιθώριο ξ (ανοχή) που καθορίζει πόσο ανεκτικοί είμαστε στην ύπαρξη σημείων εκτός του ε-σωλήνα που ορίζεται από το όριο σφάλματος ε. Έτσι, το ξ καθιστά δυνατή την εύρεση μιας λύσης στο πρόβλημα βελτιστοποίησης που, σε διαφορετική περίπτωση, μπορεί να μην υπήρχε. Δηλαδή, πλέον δεχόμαστε «σφάλματα» εκτός του ε-σωλήνα αλλά κερδίζουμε όσον αφορά την εύρεση μιας καλής λύσης (και ας μην είναι τέλεια). Στη μη γραμμική παρεμβολή μπορεί επίσης να εφαρμοστεί το κόλπο του πυρήνα (kernel trick) και να μετασχηματιστούν τα δεδομένα σε ανώτερη διάσταση, ώστε να καταστεί εφικτή η γραμμική προσαρμογή τους. Συνοψίζοντας, η μέθοδος SVR μας επιτρέπει να καθορίσουμε πόσο ανεκτικοί είμαστε σε σφάλματα, τόσο μέσω ενός αποδεκτού ορίου σφάλματος (περιθώριο ε) όσο και διαμέσου μιας ανοχής ξ στο κατά πόσο επιτρέπονται σημεία εκτός των κύριων ορίων σφάλματος.

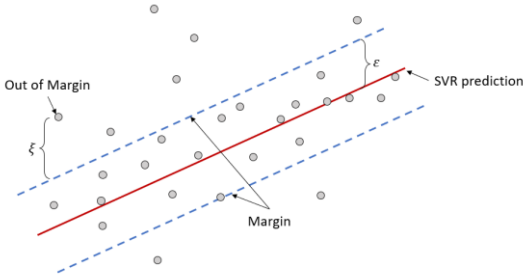


Figure 10: SVR

3.3.4. XGBoost

Ένα μοντέλο πρόβλεψης που προκύπτει από Μηχανική Μάθηση δεν είναι πάντα τέλειο. Η απόδοση του κυμαίνεται ανάλογα με τον αριθμό και την ποιότητα των δεδομένων και την καταλληλότητα του αλγορίθμου που χρησιμοποιήθηκε. Μια προφανής πρόταση σχετικά με την βελτίωση των προβλέψεων πάνω σε ένα συγκεκριμένο σύνολο δεδομένων είναι ο συνδυασμός των αποφάσεων πολλών διαφορετικών μοντέλων πρόβλεψης (ensemble techniques). Χαρακτηριστική μέθοδο για την πραγματοποίηση της παραπάνω πρότασης είναι αυτή της ενθυλάκωσης (Bagging) και ενίσχυσης (Boosting) που εφαρμόζει τον ίδιο αλγόριθμο σε διαφορετικά υποσύνολα δεδομένων. Για να μπορέσει να αποφέρει βελτιωμένα αποτελέσματα η παραπάνω μέθοδος, απαιτείται να εφαρμοστεί σε αλγορίθμους μάθησης οι οποίοι είναι ασταθείς (δηλ. έχουν μεγάλη διακύμανση (variance)). Αυτό σημαίνει, πως μικρές αλλαγές στα δεδομένα εκπαίδευσης προκαλούν αλλαγές και στο μοντέλο με αποτέλεσμα αυτό να βγάζει άλλες αποφάσεις. Ιδάλος, η εκπαίδευση του κάθε αλγορίθμου σε διαφορετικό υποσύνολο, ενδεχομένως να μην προσέφερε και διαφορετικά μοντέλα.

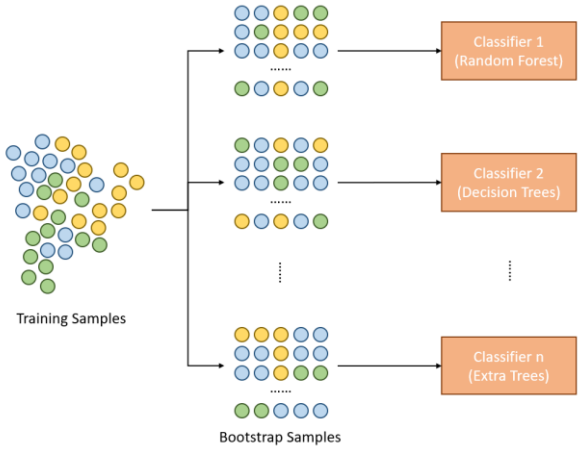


Figure 11: Bagging

Το Bagging αρχικά δημιουργεί (τυχαία) πολλά διαφορετικά σύνολα δεδομένων από το αρχικό μέσω "Δειγματοληψίας με Επανατοποθέτηση". Το κάθε νέο σύνολο δεδομένων έχει ίδιο αριθμό δεδομένων με το αρχικό, αλλά κάποια δεδομένα έχουν επαναληφθεί ενώ κάποια δεν έχουν συμπεριληφθεί καθόλου. Στη συνέχεια εφαρμόζει τον ασταθή αλγόριθμο μάθησης σε όλα τα νέα σύνολα δεδομένων και παράγει αντίστοιχα μοντέλα πρόβλεψης. Για τη διαδικασία πρόβλεψης λαμβάνονται υπόψιν οι αποφάσεις όλων των μοντέλων και η τελική τιμή είναι είτε η κατηγορία που συγκεντρώνει τις περισσότερες αποφάσεις μοντέλων (voting) είτε ο μέσος όρος των αριθμητικών προβλέψεων των διαφορετικών μοντέλων στην περίπτωση παλινδρόμησης.

Επέκταση της ενθυλάκωσης (Bagging), αποτελεί η ενίσχυση (Boosting). Είναι μια επαναληπτική διαδικασία που μετατρέπει όπως και παραπάνω, μετατρέπει ασταθείς αλγορίθμους σε ισχυρούς ελαττώνοντας το bias και το variance. Στηρίζεται στην ανάθεση βαρών (θετικών αριθμών) στα δεδομένα, έτσι ώστε ο αλγόριθμος μάθησης να επικεντρωθεί σε δεδομένα που συνήθως ταξινομούνται λάθος. Η πιθανότητα επιλογής ενός δεδομένου κατά τη δειγματοληψία είναι ανάλογη του βάρους του. Έτσι δεδομένα με μεγαλύτερο βάρος εμφανίζονται περισσότερες φορές ενώ δεδομένα με μικρότερο βάρος μπορεί να μην εμφανιστούν καθόλου. Ακόμα, τα βάρη χρησιμοποιούνται και για τον τρόπο υπολογισμού της απόδοσης ενός αλγορίθμου. Χωρίς βάρη, το ποσοστό λάθους είναι ο αριθμός των δεδομένων ελέγχου που

ταξινομούνται λάθος προς το συνολικό αριθμό των δεδομένων ελέγχου. Με βάρη, είναι το άθροισμα των βαρών των δεδομένων ελέγχου που ταξινομούνται λάθος προς το συνολικό άθροισμα των βαρών του συνόλου των δεδομένων ελέγχου (σταθμισμένο σφάλμα). Έτσι δεδομένα που έχουν μεγάλο βάρος, είναι και πιο πιθανό να επιλεχθούν στο υποσύνολο εκπαίδευσης και κατ' επέκταση να τα μάθει ο αλγόριθμος. Αλλά προκαλούν και μεγαλύτερο σφάλμα όταν δεν προβλεφθούν σωστά.

Η ενίσχυση (Boosting) είναι μια επαναληπτική διαδικασία όπου, τα διαφορετικά μοντέλα κατασκευάζονται το ένα μετά το άλλο και η απόδοση του προηγούμενου μοντέλου επηρεάζει την κατασκευή του επόμενου. Συγκεκριμένα προσπαθεί να κατασκευάσει το επόμενο μοντέλο έτσι ώστε να μην πραγματοποιεί τα ίδια λάθη με αυτά που έκανε το προηγούμενο (χρήση βαρών).

Διαδικασία Boosting:

1. Το πρώτο μοντέλο παράγεται από το αρχικό σύνολο δεδομένων. Θέτονται σε όλα τα δεδομένα του αρχικού συνόλου N ίσα βάρη $(1/N)$.
2. Έπειτα τα δεδομένα ταξινομούνται από το μοντέλο. Αν το σταθμισμένο σφάλμα e είναι πάνω από 50% $(e > 0.5)$, το μοντέλο απορρίπτεται, τα βάρη τίθενται στην τιμή $1/N$ και η διαδικασία επιστρέφει στο προηγούμενο βήμα. Ακόμα, αν η απόφαση του μοντέλου για κάποιο δεδομένο είναι λάθος τότε το βάρος του αυξάνεται ενώ αν είναι σωστή μειώνεται.
3. Η διαδικασία επαναλαμβάνεται από το βήμα 2 για τη μάθηση του επόμενου μοντέλου έως ότου επιτευχθεί ένα επιθυμητό όριο σφάλματος ή για προκαθορισμένο πλήθος κύκλων ενίσχυσης.

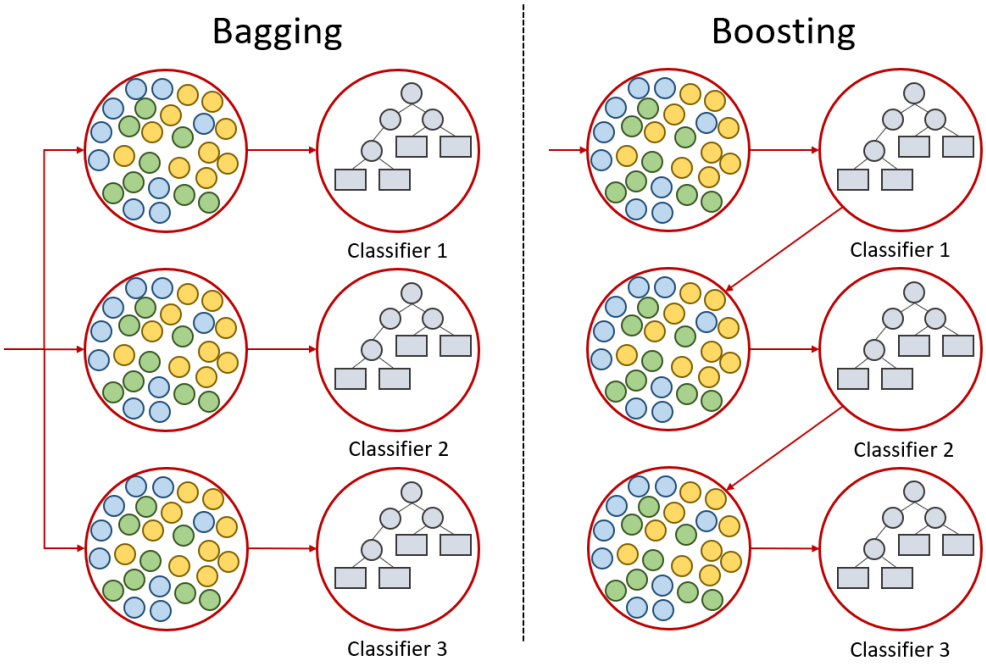


Figure 12: Bagging vs. Boosting

Αναφορικά με το κομμάτι ενημέρωσης των βαρών, πραγματοποιείται η παρακάτω διαδικασία. Για τα δεδομένα που ταξινομούνται λάθος το βάρος παραμένει όσο ήταν αρχικά, ενώ για αυτά που ταξινομούνται σωστά το βάρος μειώνεται αντιστρόφως ανάλογα με το ποσοστό λαθών e του ταξινομητή στα δεδομένα:

$$weight = weight \frac{e}{1-e}$$

Το σταθμισμένο σφάλμα ισούται με το άθροισμα των βαρών των δεδομένων που ο ταξινομητής ταξινομεί λάθος δια N, δηλαδή:

$$e_i = \frac{1}{N}\sum_{j=1}^{N} w_j$$

Στη συνέχεια τα βάρη κανονικοποιούνται έτσι ώστε το άθροισμα τους να παραμείνει όσο και πριν. Κάθε βάρος διαιρείται με το άθροισμα των νέων βαρών και πολλαπλασιάζεται με το άθροισμα των παλιών. Έτσι αυτόματα, κατά την διαδικασία δημιουργίας του συνόλου δεδομένων στο οποίο θα εκπαιδευτεί το επόμενο μοντέλο, αυξάνεται το βάρος των δεδομένων που ταξινομούνται λάθος και μειώνεται αυτών που ταξινομούνται σωστά [36].

Ακόμα, στο στάδιο της πρόβλεψης άγνωστων δεδομένων, συνδυάζονται οι αποφάσεις όλων των μοντέλων μέσω ψηφοφορίας με βάρη. Το βάρος της απόφασης κάθε μοντέλου είναι αντίστοιχο του ποσοστού λαθών e στα δεδομένα από τα οποία εκπαιδεύτηκε:

$$weight = -log\frac{e}{1-e}$$

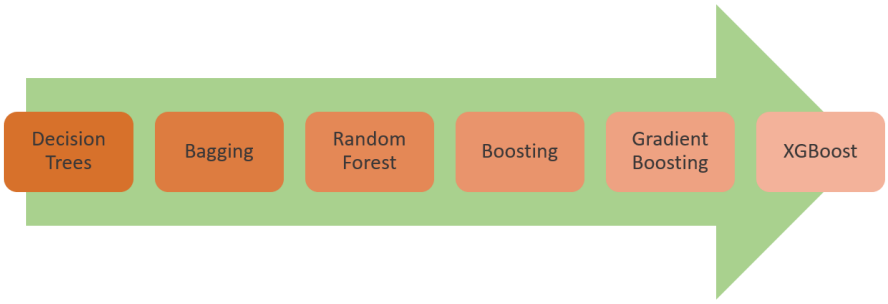


Figure 13: Evolution to XGBoost

Το boosting είναι ένας μετα-αλγόριθμος που συνδυάζει μοντέλα μηχανικής μάθησης με σκοπό κυρίως τη μείωση της μεροληψίας αλλά και της διακύμανσης στην επιβλεπόμενη μάθηση μέσω μιας διαδικασία ελαχιστοποίησης μιας κυρτής συνάρτησης κόστους. Η τεχνική του boosting προτάθηκε στο πλαίσιο του αλγορίθμου Adaboost (adaptive boosting) ο οποίος χρησιμοποιεί δένδρα ταξινόμησης σαν βασικό αλγόριθμο. Νεότεροι αλγόριθμοι βασισμένοι στο boosting πετυχαίνουν καλύτερα αποτελέσματα, όπως οι XGBoost, LPBoost, Totalboost, BrownBoost, LogitBoost, MadaBoost και άλλοι. Η διαφορά μεταξύ τους είναι κυρίως στον τρόπο υπολογισμού των βαρών στα δεδομένα .

Ο XGboost (Extreme Gradient boosting) προτάθηκε το 2016 από τους Tianqi Chen και Carlos Guestrin [37]. Είναι ένας αλγόριθμος ενίσχυσης ενός δένδρου απόφασης (ταξινόμησης/παρεμβολής) μέσω της επικλινούς ενίσχυσης (gradient boosting). Η επικλινής ενίσχυση είναι μια τεχνική ενίσχυσης που αντί να παράγει μοντέλα μεταβάλλοντας τα βάρη του συνόλου εκπαίδευσης, ενισχύει ένα αδύναμο μοντέλο μέσω μιας διαδικασίας βελτιστοποίησης επικλινούς καθόδου (gradient descent optimization procedure) ελαχιστοποιώντας μια κατάλληλη συνάρτηση κόστους. Πρόκειται για έναν εξαιρετικά δυνατό αλγόριθμο μηχανικής μάθησης που έχει αποκτήσει μεγάλη δημοσιότητα καθώς χρησιμοποιήθηκε από πολλές ερευνητικές ομάδες που κατάφεραν διακρίθηκαν σε διαγωνισμούς μηχανικής μάθησης.

3.4. Μεθοδολογίες Εκπαίδευσης (Training Techniques)

3.4.1. Διασταυρωμένη Επικύρωση k-Συνόλων (k-Fold Cross Validation)

Οι μέθοδοι επαναδειγματοληψίας έχουν αποκτήσει σημαντική θέση στις σύγχρονες στατιστικές και έχουν γίνει αναπόσπαστο εργαλείο για την ανάλυση δεδομένων. Αυτές οι μέθοδοι περιλαμβάνουν τη δημιουργία πολλαπλών δειγμάτων από ένα σύνολο δεδομένων εκπαίδευσης, επιτρέποντας στους ερευνητές να εξάγουν πολύτιμες γνώσεις και να βελτιώσουν την κατανόησή τους για τα στατιστικά μοντέλα. Με την εφαρμογή τεχνικών επαναδειγματοληψίας, όπως η Cross Validation, μπορεί κανείς να αξιολογήσει την ευρωστία, τη διακύμανση και την απόδοση ενός μοντέλου μέσω πολλαπλών επαναλήψεων.

Μια θεμελιώδης εφαρμογή των μεθόδων επαναδειγματοληψίας έγκειται στην εκτίμηση του βέλτιστου συνδυασμού υπερπαραμέτρων για μοντέλα μηχανικής μάθησης. Οι υπερπαράμετροι είναι σημαντικές ρυθμίσεις που επηρεάζουν άμεσα την απόδοση ενός μοντέλου και συνήθως επιλέγονται από τον ερευνητή. Ενώ οι ειδικοί στη μηχανική μάθηση μπορεί να έχουν μια γενική ιδέα για τις κατάλληλες τιμές υπερπαραμέτρων, η εύρεση των ακριβών τιμών που αποδίδουν τη βέλτιστη απόδοση είναι συχνά δύσκολη. Κατά συνέπεια, ένα μέσο για την αξιολόγηση της απόδοσης διαφορετικών συνδυασμών υπερπαραμέτρων, χωρίς την χρήση ενός ξεχωριστού συνόλου ελέγχου, καθίσταται απαραίτητο. Αυτή η προσέγγιση μετριάζει τον κίνδυνο επιλογής υπερπαραμέτρων που έχουν καλή απόδοση μόνο σε ένα συγκεκριμένο σύνολο δεδομένων αλλά δεν μπορούν να γενικευτούν σε άλλες περιπτώσεις [38] [39].

Μεταξύ των διαφόρων μεθόδων επαναδειγματοληψίας, η Cross Validation ξεχωρίζει ως κυρίαρχη και ευρέως υιοθετημένη τεχνική για την αξιολόγηση της απόδοσης του μοντέλου. Η μέθοδος Cross Validation περιλαμβάνει τη διαίρεση του συνόλου εκπαίδευσης σε k ίσα υποσύνολα, που συνήθως αναφέρονται ως "k-folds". Κατά τη διάρκεια κάθε επανάληψης Cross Validation, ένα υποσύνολο ορίζεται ως το δοκιμαστικό σύνολο, ενώ τα υπόλοιπα k-1 υποσύνολα χρησιμοποιούνται για την εκπαίδευση του μοντέλου. Αυτή η διαδικασία επαναλαμβάνεται k φορές, με κάθε υποσύνολο να χρησιμεύει ως δοκιμαστικό σύνολο μία φορά. Χρησιμοποιώντας όλες τις παρατηρήσεις στο σετ εκπαίδευσης για εκπαίδευση και αξιολόγηση μοντέλων, το Cross

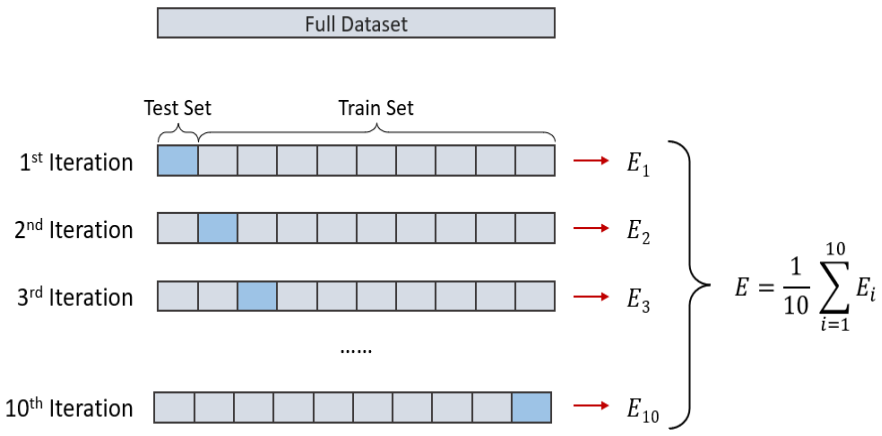


Figure 14: 10-Fold Cross Validation Scheme

Validation παρέχει μια ολοκληρωμένη αξιολόγηση της συνολικής απόδοσης του μοντέλου σε όλο το σύνολο.

Για τον υπολογισμό της συνολικής απόδοσης του μοντέλου χρησιμοποιώντας Cross Validation, χρησιμοποιείται ένας συγκεκριμένος τύπος. Ο τύπος περιλαμβάνει τη λήψη του μέσου όρου της απόδοσης που επιτυγχάνεται σε κάθε υποσύνολο:

$$CV_k = \frac{1}{k}\sum_{i=1}^{k}(Επίδοση\ στο\ i^{th}\ υποσύνολο)$$

Εδώ, το CV_k αντιπροσωπεύει τη μέτρηση απόδοσης Cross Validation, το k υποδηλώνει τον αριθμό των “folds” ή υποσυνόλων και το i αντιπροσωπεύει τον δείκτη κάθε υποσυνόλου. Η απόδοση που επιτυγχάνεται σε κάθε υποσύνολο αξιολογείται με βάση το επιλεγμένο κριτήριο αξιολόγησης, το οποίο ποικίλλει ανάλογα με το συγκεκριμένο πρόβλημα και τον στόχο του μοντέλου.

Το πλεονέκτημα του Cross Validation έγκειται στην ικανότητά του να παρέχει μια πιο αξιόπιστη εκτίμηση της απόδοσης ενός μοντέλου σε σύγκριση με τη χρήση ενός διαχωρισμού εκπαίδευσης-ελέγχου. Διαχωρίζοντας επανειλημμένα το σύνολο δεδομένων σε υποσύνολα εκπαίδευσης και ελέγχου, η Cross Validation καταγράφει τη μεταβλητότητα στην απόδοση του μοντέλου που προκύπτει από διαφορετικούς διαχωρισμούς εκπαίδευσης-ελέγχων. Αυτή η προσέγγιση προσφέρει πολύτιμες γνώσεις σχετικά με τις δυνατότητες γενίκευσης του μοντέλου και βοηθά τους ερευνητές να λαμβάνουν τεκμηριωμένες αποφάσεις σχετικά με την επιλογή υπερπαραμέτρων.

Επιπλέον, η Cross Validation δίνει τη δυνατότητα στους ερευνητές να εντοπίσουν πιθανά ζητήματα όπως η υπερπροσαρμογή ή η υποσυναρμολόγηση. Η υπερπροσαρμογή συμβαίνει όταν ένα μοντέλο αποδίδει εξαιρετικά καλά στα δεδομένα εκπαίδευσης αλλά αποτυγχάνει να γενικεύσει σε δεδομένα πέρα από αυτά. Χρησιμοποιώντας Cross Validation, οι ερευνητές μπορούν να αξιολογήσουν εάν η απόδοση του μοντέλου παραμένει σταθερή σε διαφορετικά υποσύνολα, παρέχοντας μια πιο ολοκληρωμένη αξιολόγηση των δυνατοτήτων γενίκευσής του. Ομοίως, η υποπροσαρμογή, η οποία υποδεικνύει ότι το μοντέλο δεν καταγράφει τα υποκείμενα μοτίβα στα δεδομένα, μπορεί επίσης να ανιχνευθεί μέσω Cross Validation.

Συνοπτικά, οι μέθοδοι επαναδειγματοληψίας, με την Cross Validation να είναι μια εξέχουσα προσέγγιση, έχουν φέρει επανάσταση στη στατιστική ανάλυση και την αξιολόγηση μοντέλων μηχανικής μάθησης. Δημιουργώντας επανειλημμένα δείγματα από ένα σύνολο δεδομένων εκπαίδευσης και χρησιμοποιώντας τα ληφθέντα υποσύνολα για εκπαίδευση και αξιολόγηση μοντέλων, οι ερευνητές μπορούν να εκτιμήσουν την απόδοση των μοντέλων και να βελτιστοποιήσουν τις ρυθμίσεις υπερπαραμέτρων. Η Cross Validation, ειδικότερα, διαιρεί το σετ εκπαίδευσης σε k-folds, επιτρέποντας ολοκληρωμένη αξιολόγηση της απόδοσης του μοντέλου και βοηθώντας στην επιλογή των βέλτιστων υπερπαραμέτρων. Με την ευρεία εφαρμογή και την ευρωστία του, το Cross Validation έχει γίνει ένα απαραίτητο εργαλείο για ερευνητές σε διάφορους τομείς που επιδιώκουν να βελτιώσουν την κατανόησή τους στα στατιστικά μοντέλα και να εξασφαλίσουν αξιόπιστη απόδοση σε αόρατα δεδομένα.

3.4.2. Αναζήτηση Πλέγματος (Grip Search)

Η αναζήτηση πλέγματος (Γνωστή και ως full factorial design [40]) είναι μια ισχυρή τεχνική που χρησιμοποιείται για τον συντονισμό της διαδικασίας προσδιορισμού των βέλτιστων υπερπαραμετρων ενός μοντέλου μηχανικής εκμάθησης. Λειτουργεί με εξαντλητική αναζήτηση συγκεκριμένων τιμών εντός ενός προκαθορισμένου εύρους για κάθε υπερπαράμετρο του μοντέλου. Η βασική αρχή είναι σχετικά απλή. Ο ερευνητής καθορίζει το εύρος τιμών για κάθε

υπερπαράμετρο που επιθυμεί να εξερευνήσει, δημιουργώντας ένα πλέγμα που περιλαμβάνει όλους τους πιθανούς συνδυασμούς αυτών των τιμών. Στη συνέχεια, η απόδοση του μοντέλου αξιολογείται για κάθε συνδυασμό μέσω μιας σειράς δοκιμών, στην οποία, συνήθως χρησιμοποιείται η διασταυρούμενη επικύρωση. Η αναζήτηση πλέγματος αυτοματοποιεί και εξερευνά συστηματικά το υποσύνολο του χώρου παραμέτρων που ορίζεται από τον ερευνητή.

Τα οφέλη της αναζήτησης πλέγματος είναι πολύπλευρα. Πρώτον, παρέχει ένα αποτελεσματικό και ολοκληρωμένο μέσο για τον προσδιορισμό των βέλτιστων τιμών παραμέτρων για ένα μοντέλο μηχανικής μάθησης. Αξιολογώντας εξαντλητικά όλους τους πιθανούς συνδυασμούς εντός του καθορισμένου χώρου παραμέτρων, η αναζήτηση πλέγματος προσφέρει την πιο ενδελεχή εξερεύνηση του τοπίου απόδοσης του μοντέλου. Αυτή η προσέγγιση διασφαλίζει ότι οι υπερπαράμετροι του μοντέλου ρυθμίζονται με ακρίβεια για να αποδίδουν τα καλύτερα δυνατά αποτελέσματα.

Η αναζήτηση πλέγματος είναι ιδιαίτερα πλεονεκτική όταν έχουμε να κάνουμε με πολύπλοκα μοντέλα που διαθέτουν πολλαπλές υπερπαραμέτρους. Τα μοντέλα με πολλές υπερπαραμέτρους μπορούν να έχουν τεράστιο χώρο παραμέτρων, καθιστώντας δύσκολη τη μη αυτόματη αναγνώριση των καταλληλότερων συνδυασμών. Η αναζήτηση πλέγματος απλοποιεί αυτή τη διαδικασία δοκιμάζοντας συστηματικά κάθε συνδυασμό μέσα στο πλέγμα, αφαιρώντας το βάρος της χειροκίνητης εξερεύνησης. Εξαλείφει την ανάγκη για διαίσθηση ή δοκιμή και σφάλμα, παρέχοντας μια πιο συστηματική και αξιόπιστη προσέγγιση στον συντονισμό υπερπαραμέτρων.

Επιπλέον, η αναζήτηση πλέγματος προσφέρει διαφάνεια και επαναληψιμότητα στη διαδικασία επιλογής μοντέλου. Ορίζοντας ρητά το εύρος παραμέτρων και αξιολογώντας συστηματικά κάθε συνδυασμό, η αναζήτηση πλέγματος διασφαλίζει ότι η διαδικασία είναι διαφανής και αναπαραγόμενη. Οι ερευνητές μπορούν να τεκμηριώσουν και να αναπαράγουν με ακρίβεια ολόκληρο τον χώρο αναζήτησης και τα αποτελέσματα αξιολόγησης, διευκολύνοντας την κατανόηση και την επικύρωση των επιλεγμένων τιμών υπερπαραμέτρων.

Η θεωρητική βάση της αναζήτησης πλέγματος βρίσκεται στην ιδέα της εξερεύνησης ενός προκαθορισμένου χώρου παραμέτρων για τη βελτιστοποίηση της απόδοσης του μοντέλου. Αξιολογώντας την απόδοση του μοντέλου για κάθε συνδυασμό μέσα στο πλέγμα, η αναζήτηση πλέγματος επιχειρεί να εντοπίσει τις υπερπαραμέτρους που αποδίδουν τα καλύτερα αποτελέσματα. Αυτή η προσέγγιση ευθυγραμμίζεται με την έννοια της επιλογής μοντέλου, όπου ο στόχος είναι να βρεθούν οι υπερπαράμετροι που βελτιστοποιούν την απόδοση του μοντέλου σε μια δεδομένη εργασία ή σύνολο δεδομένων.

Η χρήση της διασταυρούμενης επικύρωσης εντός της αναζήτησης πλέγματος ενισχύει την ευρωστία της διαδικασίας επιλογής παραμέτρων. Η διασταυρούμενη επικύρωση διαιρεί το σύνολο δεδομένων σε πολλαπλά υποσύνολα ή “folds”, επιτρέποντας στο μοντέλο να εκπαιδευτεί και να αξιολογηθεί σε διαφορετικούς συνδυασμούς δεδομένων. Με την ενσωμάτωση της διασταυρούμενης επικύρωσης στην αναζήτηση πλέγματος, η απόδοση κάθε συνδυασμού παραμέτρων αξιολογείται σε διάφορα υποσύνολα δεδομένων, παρέχοντας μια πιο αξιόπιστη εκτίμηση των δυνατοτήτων γενίκευσης του μοντέλου.

Παρά τα πλεονεκτήματά της, η αναζήτηση πλέγματος έχει περιορισμούς. Ένα βασικό μειονέκτημα είναι το υπολογιστικό του κόστος, ιδιαίτερα όταν έχουμε να κάνουμε με μεγάλους χώρους παραμέτρων και υπολογιστικά απαιτητικά μοντέλα. Καθώς το πλέγμα επεκτείνεται εκθετικά με τον αριθμό των παραμέτρων και τις πιθανές τιμές τους, η διαδικασία αναζήτησης μπορεί να γίνει χρονοβόρα και εξαιρετικά απαιτητική σε πόρους. Αυτό το ζήτημα γίνεται πιο έντονο όταν αντιμετωπίζουμε πολύπλοκα μοντέλα που απαιτούν εκτενή εκπαίδευση και αξιολόγηση.

Για να μετριάσουν τον υπολογιστικό φόρτο, οι ερευνητές χρησιμοποιούν συχνά τεχνικές βελτιστοποίησης, όπως η τυχαιοποιημένη αναζήτηση ή η Bayesian βελτιστοποίηση, που

παρέχουν πιο αποτελεσματική εξερεύνηση του χώρου παραμέτρων. Αυτές οι μέθοδοι δειγματίζουν συνδυασμούς παραμέτρων που βασίζονται σε στατιστικές αρχές ή χρησιμοποιούν ευρετικές μεθόδους για να καθοδηγήσουν τη διαδικασία αναζήτησης. Αν και αυτές οι εναλλακτικές λύσεις προσφέρουν ταχύτερη σύγκλιση σε βέλτιστες λύσεις, ενδέχεται να μην εγγυώνται την εξαντλητική εξερεύνηση ολόκληρου του χώρου παραμέτρων, χάνοντας πιθανώς ελπιδοφόρες δοκιμές.

Συμπερασματικά, η αναζήτηση πλέγματος είναι μια πολύτιμη τεχνική για τον συντονισμό υπερπαραμέτρων σε μοντέλα μηχανικής μάθησης. Με τη συστηματική διερεύνηση ενός προκαθορισμένου χώρου παραμέτρων και την αξιολόγηση της απόδοσης του μοντέλου για κάθε συνδυασμό, η αναζήτηση πλέγματος προσδιορίζει τις βέλτιστες τιμές υπερπαραμέτρων. Προσφέρει διαφάνεια, επαναληψιμότητα και ολοκληρωμένη διαδικασία αναζήτησης, δίνοντας τη δυνατότητα στους ερευνητές να τελειοποιήσουν τα μοντέλα τους και να επιτύχουν τη βέλτιστη απόδοση. Ωστόσο, οι ερευνητές θα πρέπει να προσέχουν το υπολογιστικό κόστος που σχετίζεται με την αναζήτηση πλέγματος, ειδικά για πολύπλοκα μοντέλα και μεγάλους χώρους παραμέτρων [41].

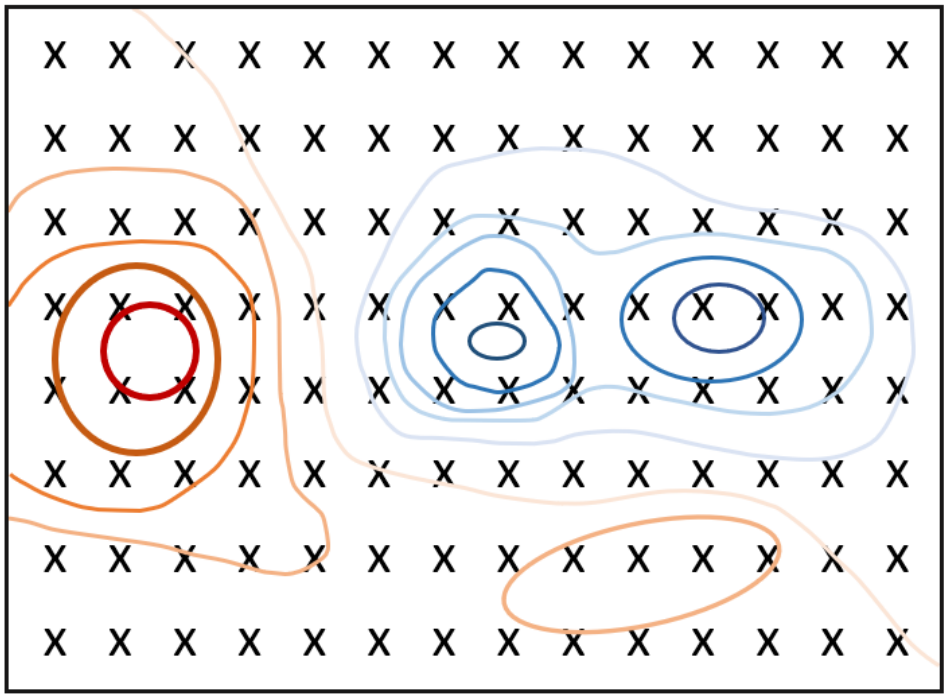


Figure 15: Grid Search for the parameters X_1 and X_2

4. Μεθοδολογία (Methodology - Algorithm Description)

4.1. Περιγραφή Δεδομένων (Dataset Description)

Το σύνολο δεδομένων που αξιοποιήθηκε στην συγκεκριμένη διπλωματική αποτελείτε από 18783 κατοικίες του Άμστερνταμ με 74 χαρακτηρίστηκα για την κάθε μία. Κατ' επέκταση οι αρχικές διαστάσεις του συνόλου δεδομένων είναι 18783 επί 74. Ακολουθεί πίνακας που περιγράφει σύντομα το καθένα από αυτά χαρακτηριστικά:

Table 1: Dataset Features Description

A/A	Χαρακτηριστικό	Περιγραφή
01	id	Το id της κατοικίας
02	listing_url	Η σελίδα από όπου έγινε το scrape των χαρακτηριστικών
03	scrape_id	Το id της scraping συνεδρίας
04	last_scraped	Πότε έγινε το τελευταίο scrape
05	name	Ο τίτλος της κατοικίας
06	description	Η περιγραφή της κατοικίας ως ελεύθερο κείμενο
07	neighborhood_overview	Περιγραφή της γειτονιάς της κατοικίας
08	picture_url	URL για εικόνες της κατοικίας
09	host_id	Το id του διαχειριστή της κατοικίας
10	host_url	Το URL του συγκεκριμένου host
11	host_name	Το όνομα του host
12	host_since	Από πότε υπάρχει αυτός ο host.
13	host_location	Σε ποια περιοχή διαμένει ο host
14	host_about	Περιγραφή του host
15	host_response_time	Ο μέσος χρόνος ανταπόκρισης του host
16	host_response_rate	Το rating του host ανάλογα με τον μέσο χρόνο ανταπόκρισης του
17	host_acceptance_rate	Ποσοστό των κρατήσεων που αναλαμβάνει ο host
18	host_is_superhost	Αν ο host είναι superhost
19	host_thumbnail_url	Η διεύθυνση URL της εικόνας thumbnail host
20	host_picture_url	URL για την εικόνα προφίλ του host
21	host_neighbourhood	Η γειτονία του host
22	host_listings_count	Το σύνολο των διαθέσιμων ακινήτων που διαχειρίζεται ο host
23	host_total_listings_count	Το σύνολο των ακινήτων που διαχειρίζεται ο host

24	host_verifications	Τα verifications που έχει ο host
25	host_has_profile_pic	Αν ο host έχει φωτογραφία προφίλ
26	host_identity_verified	Αν έχει διασταυρωθεί η ταυτότητα του
27	neighbourhood	Η πόλη στην οποία βρίσκετε το ακίνητο
28	neighbourhood_cleansed	Η γειτονιά στην οποία βρίσκετε το ακίνητο
29	neighbourhood_group_cleansed	Η γειτονιά στην οποία βρίσκετε το ακίνητο
30	latitude	Το γεωγραφικό πλάτος του ακινήτου
31	longitude	Το γεωγραφικό μήκος του ακινήτου
32	property_type	Το είδος της κατοικίας
33	room_type	Το είδος της κατοικίας
34	accommodates	Το πλήθος τον ανθρώπων που μπορεί να φιλοξενήσει
35	bathrooms	Τα μπάνια
36	bathrooms_text	Περιγραφή των μπάνιων
37	bedrooms	Το πλήθος των μπάνιων
38	beds	Το πλήθος των κρεβατιών
39	amenities	Οι ανέσεις που παρέχονται
40	price	Η τιμή ενοικιάσεις
41	minimum_nights	Η ελάχιστες μέρες διαμονής
42	maximum_nights	Η μέγιστες μερες διαμονής
43	minimum_minimum_nights	Το κάτω φράγμα της κατανομής των ελάχιστων διανυκτερεύσεων
44	maximum_minimum_nights	Το κάτω φράγμα της κατανομής των μέγιστων διανυκτερεύσεων
45	minimum_maximum_nights	Το άνω φράγμα της κατανομής των ελάχιστων διανυκτερεύσεων
46	maximum_maximum_nights	Το άνω φράγμα της κατανομής των μέγιστων διανυκτερεύσεων
47	minimum_nights_avg_ntm	Η μέση τιμή της κατανομής των ελάχιστων διανυκτερεύσεων
48	maximum_nights_avg_ntm	Η μέση τιμή της κατανομής των μέγιστων διανυκτερεύσεων
49	calendar_updated	Αν έχει ενημερωθεί η διαθεσιμότητα
50	has_availability	Αν είναι διαθέσιμο
51	availability_30	Οι ημέρες διαθεσιμότητας από τις επόμενες 30.
52	availability_60	Οι ημέρες διαθεσιμότητας από τις επόμενες 60.

53	availability_90	Οι ημέρες διαθεσιμότητας από τις επόμενες 90.
54	availability_365	Οι ημέρες διαθεσιμότητας μέσα στον επερχόμενο χρόνο
55	calendar_last_scraped	Πότε έγινε τελευταία φορά scrape η συγκεκριμένη κατοικία
56	number_of_reviews	Το πλήθος των αξιολογήσεων
57	number_of_reviews_ltm	Το πλήθος των αξιολογήσεων τους τελευταίους 12 μήνες (last twelve months)
58	number_of_reviews_l30d	Το πλήθος των αξιολογήσεων τις τελευταίες 130 ημέρες
59	first_review	Πότε έγινε η πρώτη αξιολόγηση
60	last_review	Πότε έγινε η τελευταία αξιολόγηση
61	review_scores_rating	Η συνολική βαθμολογία των αξιολογήσεων.
62	review_scores_accuracy	Η ακρίβεια των αξιολογήσεων.
63	review_scores_cleanliness	Η βαθμολογία σχετικά με την καθαριότητα
64	review_scores_checkin	Η βαθμολογία σχετικά με το Check-in
65	review_scores_communication	Η βαθμολογία σχετικά με την επικοινωνία
66	review_scores_location	Η βαθμολογία σχετικά με την τοποθεσία
67	review_scores_value	Η βαθμολογία σχετικά με την αξία του καταλύματος
68	license	Η άδεια της κατοικίας
69	instant_bookable	Αν είναι άμεσα διαθέσιμο
70	calculated_host_listings_count	Το πλήθος των κατοικιών που διαχειρίζεται ο συγκεκριμένος host
71	calculated_host_listings_count_entire_homes	Το πλήθος των κατοικιών τύπου «ολόκληρο οίκημα» διαχειρίζεται ο συγκεκριμένος host
72	calculated_host_listings_count_private_rooms	Το πλήθος των κατοικιών τύπου «Διαμέρισμα» διαχειρίζεται ο συγκεκριμένος host
73	calculated_host_listings_count_shared_rooms	Το πλήθος των κατοικιών τύπου «Κοινόχρηστο διαμέρισμα» διαχειρίζεται ο συγκεκριμένος host
74	reviews_per_month	Ο μέσος όρος αξιολογήσεων ανά μήνα

4.2. Καθαρισμός Δεδομένων (Data Cleaning

Καθώς όλο το σύνολο των δεδομένων έχει προκύψει μέσα από scraping, είναι αναμενόμενο πολλά από τα χαρακτηριστικά να περιέχουν τιμές σε μη αξιοποιήσιμη μορφή. Ο λόγος που συμβαίνει αυτό είναι ότι ο αλγόριθμος που συλλέγει τα δεδομένα από το διαδίκτυο, αποθηκεύει τυφλά οποιαδήποτε πληροφορία βρει από τις αντίστοιχες σελίδες, χωρίς να πραγματοποιεί καμία μορφή προεπεξεργασίας πάνω σε αυτά. Έτσι λανθασμένης μορφής πληροφορία, λάθος πληροφορία ή ακόμα και ελλιπής πληροφορία μεταφέρετε αυτούσια στο σύνολο δεδομένων. Κατ' επέκταση μεγάλο κομμάτι της συγκεκριμένης προσέγγισης έχει αφιερωθεί στο στάδιο του καθαρισμού των δεδομένων και στην συνέχεια στην προεπεξεργασία.

4.2.1. Απομάκρυνση Εκλιπόντων Τιμών (Cleaning NaN Values)

Αρχικά εξετάστηκε σε ποια από τα χαρακτηριστικά του συνόλου δεδομένων λείπουν παραπάνω από τις μισές τιμές. Ο λόγος που επιλέγουμε το 50% έγκειται στον τρόπο που θα αντιμετωπιστούν οι περιπτώσεις τον εκλιπόντων τιμών. Πιο συγκεκριμένα, στα χαρακτηριστικά που λείπουν τιμές θα εφαρμοστούν προσεγγίσεις (ανά περίπτωση) για να συμπληρωθούν τα κενά. Παρόλα αυτά όλες αυτές οι προσεγγίσεις συσχετίζονται άμεσα με την φυσική σημασία του χαρακτηριστικού, καθώς και με το πλήθος των τιμών που υπάρχουν. Δηλαδή, αν υπάρχουν μερικές εκλιπούσες τιμές (δηλαδή σε μικρό ποσοστό), τότε επιλέγετε η εκπαίδευση ενός μοντέλου μηχανικής μάθησης που θα αξιοποιήσει το υπόλοιπο σύνολο δεδομένων ώστε να συμπληρώσει αυτές τις τιμές όσο πιο ρεαλιστικά γίνετε. Ακόμα, απλούστερες μέθοδοι έχουν προταθεί όπως η συμπλήρωση όλων των εκλιπόντων τιμών με την μέση τιμή των υπολοίπων, την επικρατούσα τιμή κτλ. Παρόλα αυτά σε όλες τις περιπτώσεις συμπληρώνετε συνθετικά η χαμένη πληροφορία. Κατ' επέκταση αν το ποσοστό των εκλιπόντων τιμών ξεπερνά το 50% θεωρείτε ότι συμπληρώνοντας τιμές, το συγκεκριμένο χαρακτηριστικό θα αποτελείτε κυρίως από συνθετικές τιμές πάρα πραγματικές. Συνεπώς η πληροφορία που θα προσφέρει θα είναι πολύ μικρή. Εδώ όμως αξίζει να τονιστεί πως μεγάλο ρόλο παίζει και η σημαντικότητα του χαρακτηριστικού. Αν δηλαδή μιλάμε για ένα χαρακτηριστικό που όταν εμφανίζει μια συγκεκριμένη τιμή παρατηρείτε μεγάλη αλλαγή στο χαρακτηριστικό πρόβλεψής ενώ σε όλες τις υπόλοιπες περιπτώσεις το επηρεάζει σε αμελητέο βαθμό, τότε συστήνετε η διατήρηση του και η συμπλήρωση των εκλιπόντων τιμών με μια από τις προαναφερόμενες μεθόδους.

Κατ' επέκταση στο συγκεκριμένο σύνολο δεδομένων, παρατηρήθηκε πως τα παρακάτω χαρακτηριστικά εμφανίζουν ποσοστό εκλιπόντων τιμών πάνω από το 50%:

Table 2: Features with over 50% Nan Values

#	Χαρακτηριστικό	Ποσοστό εκλιπόντων τιμών	Συσχέτιση Spearman με την τιμή του ακινήτου
15	host_response_time	67%	4.8%
29	neighbourhood_group_cleansed	100%	-
35	bathrooms	100%	-
49	calendar_updated	100%	-
68	license	100%	-

Συνεπώς μπορούμε άφοβα να απομακρύνουμε τα εν λόγο χαρακτηριστικά.

4.2.2. Απομακρηνση Ασυσχέτιστων Χαρακτηρηστηκών (Cleaning Non-Relevant Features)

Ακόμα εκτός των παραπάνω χαρακτηριστικών, απομακρύνθηκαν επιπλέον χαρακτηρίστηκα των οποίων είτε η τιμές ήταν σε μη αξιοποιήσιμη μορφή, είτε δεν παρουσίαζαν συσχέτιση με το χαρακτηριστικό πρόβλεψης, είτε αποτελούνταν από μονάχα μια τιμή και συνεπώς δεν προσέδιδαν επιπλέον πληροφορία. Παρακάτω παρουσιάζονται όλα τα επιπλέον χαρακτηριστικά που απομακρύνθηκαν, καθώς και η αιτία που οδήγησε σε αυτό:

Table 3: Removed Features

#	Χαρακτηριστικό	Λόγος Απομάκρυνσης
02	listing_url	Διεύθυνση URL, Δεν επηρεάζει την τιμή ενοικίασης του καταλύματος
03	scrape_id	Δεν περιέχει χρήσιμη πληροφορία σχετικά με την τιμή ενοικίασης του καταλύματος
04	last_scraped	
05	name	
06	description	Μη δομημένο κείμενο
07	neighborhood_overview	
08	picture_url	Διεύθυνση URL, Δεν επηρεάζει την τιμή ενοικίασης του καταλύματος
10	host_url	Διεύθυνση URL, Δεν επηρεάζει την τιμή ενοικίασης του καταλύματος
13	host_location	Δεν περιέχει χρήσιμη πληροφορία σχετικά με την τιμή ενοικίασης του καταλύματος
14	host_about	Μη δομημένο κείμενο
19	host_thumbnail_url	Διεύθυνση URL, Δεν επηρεάζει την τιμή ενοικίασης του καταλύματος
20	host_picture_url	Διεύθυνση URL, Δεν επηρεάζει την τιμή ενοικίασης του καταλύματος
21	host_neighbourhood	Δεν περιέχει χρήσιμη πληροφορία σχετικά με την τιμή ενοικίασης του καταλύματος
22	host_listings_count	
23	host_total_listings_count	
24	host_verifications	
25	host_has_profile_pic	
27	neighbourhood	Μη δομημένο κείμενο και πολλά NaNs
30	latitude	Καλύπτεται από το χαρακτηριστικό neighbourhood_cleansed
31	longitude	
32	property_type	Ισχυρή συσχέτιση με το room_type και κρατάμε μόνο το room_type

47	minimum_nights_avg_ntm	Δεν θα υπάρχει σε μελλοντικές καταχωρίσεις. Άρα δεν θέλουμε το μοντέλο να μάθει να έχει αυτά χαρακτηριστικά στην είσοδο
48	maximum_nights_avg_ntm	
43	minimum_minimum_nights	
44	maximum_minimum_nights	
45	minimum_maximum_nights	
46	maximum_maximum_nights	
50	has_availability	Για όλα τα καταλλήματα παίρνει την ίδια τιμή
51	availability_30	Δεν περιέχει χρήσιμη πληροφορία σχετικά με την τιμή ενοικίασης του καταλύματος
52	availability_60	
53	availability_90	
54	availability_365	
55	calendar_last_scraped	
56	number_of_reviews	Δεν θα υπάρχει σε μελλοντικές καταχωρίσεις. Άρα δεν θέλουμε το μοντέλο να μάθει να έχει αυτά χαρακτηριστικά στην είσοδο
57	number_of_reviews_ltm	
58	number_of_reviews_l30d	
59	first_review	
60	last_review	
62	review_scores_accuracy	
63	review_scores_cleanliness	
64	review_scores_checkin	
65	review_scores_communication	
66	review_scores_location	
67	review_scores_value	
70	calculated_host_listings_count	Δεν περιέχει χρήσιμη πληροφορία σχετικά με την τιμή ενοικίασης του καταλύματος
71	calculated_host_listings_count_entire_homes	
72	calculated_host_listings_count_private_rooms	
73	calculated_host_listings_count_shared_rooms	
74	reviews_per_month	Δεν θα υπάρχει σε μελλοντικές καταχωρίσεις. Άρα δεν θέλουμε το μοντέλο να μάθει να έχει αυτά χαρακτηριστικά στην είσοδο

Σε αυτό το σημείο, ολοκληρώνεται η διαδικασία καθαρισμού των δεδομένων και ξεκινάει το κομμάτι της προεπεξεργασίας.

4.3. Προεπεξεργασία (Preprocessing)

Στο κομμάτι αυτό της προεπεξεργασίας χρειάστηκε να εφαρμοστούν διαφορετικές μέθοδοι ανά χαρακτηριστικό, καθώς το κάθε ένα είχε διαφορετική μορφή τιμών. Αρχικά, θα δοθεί έμφαση στην δημιουργία αντιπροσωπευτικής αναπαράστασης του κάθε χαρακτηριστικού σε αριθμητικές τιμές.

Όπως έχει προαναφερθεί, το σύνολο δεδομένων έχει συλλεχτεί από το διαδίκτυο και συνεπώς οι τιμές των χαρακτηριστικών δεν είναι σε κατάλληλη μορφή για να τροφοδοτηθούν στα μοντέλα μηχανικής μάθησης. Αφότου έχει ολοκληρωθεί το στάδιο της απομάκρυνσης χαρακτηριστικών που δεν θα χρειαστούν στην εκπαίδευση των μοντέλων, πρέπει τώρα όλα τα εναπομείναντα χαρακτηριστικά να έρθουν στην κατάλληλη μορφή. Αυτό σημαίνει πως όλα τα δεδομένα πρέπει να είναι αριθμοί που να μην περιέχουν σύμβολα, χαρακτήρες και επιπρόσθετα οι αριθμοί αυτή να μεταφέρουν όσο πιο πιστά γίνετε την φυσική σημασία του κάθε χαρακτηριστικού. Πρέπει δηλαδή, για το κάθε χαρακτηριστικό, να δημιουργηθεί μια αντιπροσωπευτική αναπαράσταση. Ακόμα, παράλληλα με αυτήν την διαδικασία θα εφαρμόζετε και συμπλήρωση των εκλιπόντων τιμών όπου αυτές εμφανίζονται, με σεβασμό πάντα στην φυσιολογία του εκάστοτε χαρακτηριστικού.

4.3.1. Κωδικοποίηση (Encoding)

Σε αυτό το σημείο, καθώς το κάθε χαρακτηριστικό του συνόλου δεδομένων απαιτεί διαφορετική προσέγγιση θα γινεί ανάλυση της προσεξεργασίας στο κάθε ένα ξεχωριστά:

host_since

Εδώ επιλέχτηκε να διατηρηθεί μόνο το έτος από το οποίο ξεκίνησε ο κάθε host και όχι ολόκληρη η ακριβής ημερομηνία. Αυτή η επιλογή έγινε διότι δεν θα μπορούσε να αποθηκευτεί εύκολα η πλήρης ημερομηνία σε μορφή που να είναι επαρκώς αξιοποιήσιμη από το εκάστοτε μοντέλο. Αλλά επίσης δεν αφαιρέθηκε τελείως καθώς υπάρχει περίπτωση όσο πιο παλιός να είναι ένας host τόσο καλύτερα να μπορεί να προωθήσει το κατάλυμα του και συνεπώς να πετύχει καλύτερες και πιο ανταγωνιστικές τιμές. Παρόλα αυτά, δεν επαρκεί αυτή η διαισθητική εξήγηση. Αντίθετα, θα αποφασίσει μελλοντικά το κάθε μοντέλο αν η πληροφορία αυτή του είναι χρήσιμη ή όχι.

Συνεπώς η προεπεξεργασία σε αυτό το χαρακτηριστικό είχε το παρακάτω αποτέλεσμα:

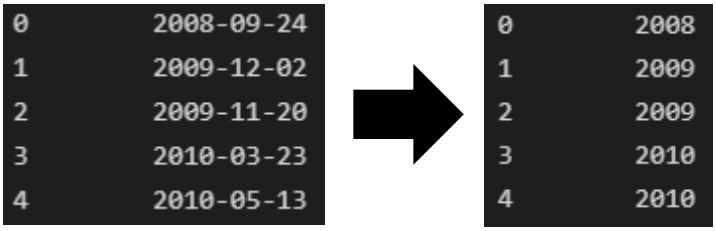


Figure 16: host_since preprocessing

host_response_rate και host_acceptance_rate

Εδώ επιλέχτηκε να μετατραπούν τα ποσοστά επί της εκατό σε δεκαδικά μεταξύ του 0 και 1 καθώς και η αντικατάσταση όλων των εκλιπόντων τιμών με την μέση τιμή των υπολοίπων. Ο λόγος που επιλέχτηκε η μέση τιμή είναι ότι αφού αυτά τα δυο χαρακτηριστικά αναφέρονται σε rating θέλουμε όλες όσες λείπουν να συμπληρωθούν με μια τιμή όσο το δυνατόν πιο ουδέτερη. Διότι, οι πολύ χαμηλές προσδίδουν αρνητική χροιά ενώ η πολύ ψηλές θετική.

Συνεπώς η προεπεξεργασία σε αυτά τα χαρακτηριστικών είχε το παρακάτω αποτέλεσμα:

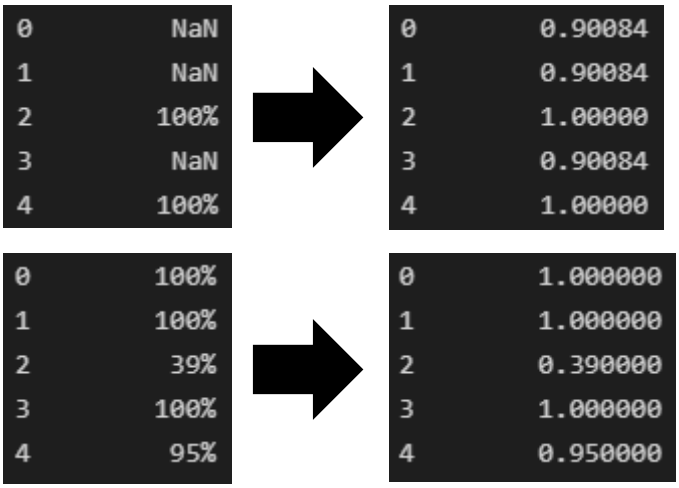


Figure 17: host_response_rate and host_acceptance_rate preprocessing

host_is_superhost, host_identity_verified και instant_bookable

εδώ επιλέχτηκε μια μορφή encoding. Καθώς και τα τρία αυτά χαρακτηριστικά παίρνουν τις τιμές ‘t’ και ‘f’ για True και False αντίστοιχα επιλέχτηκε να αντικατασταθούνε 1 και 0. Όλες οι ‘t’ τιμές με 1 και όλες οι ‘f’ με 0.

Συνεπώς η προεπεξεργασία σε αυτά τα χαρακτηριστικών είχε το παρακάτω αποτέλεσμα:

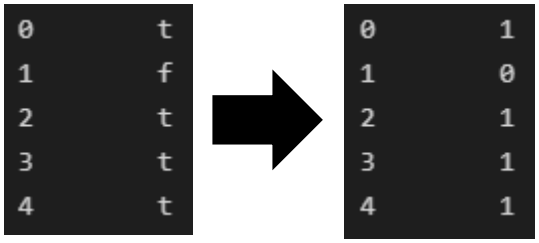


Figure 18: host_is_superhost, host_identity_verified and instant_bookable preprocessing

neighbourhood_cleansed

Εδώ επιλέχτηκε να πραγματοποιηθεί κατηγοριοποίηση των γειτονιών ανάλογα με την γεωγραφική τους τοποθεσία και στην συνέχεια η εφαρμογή one hot encoding. Η κατηγοριοποίηση έγινε βάση των επίσημων γειτονιών του Άμστερνταμ, όπως φαίνεται και στην παρακάτω εικόνα:

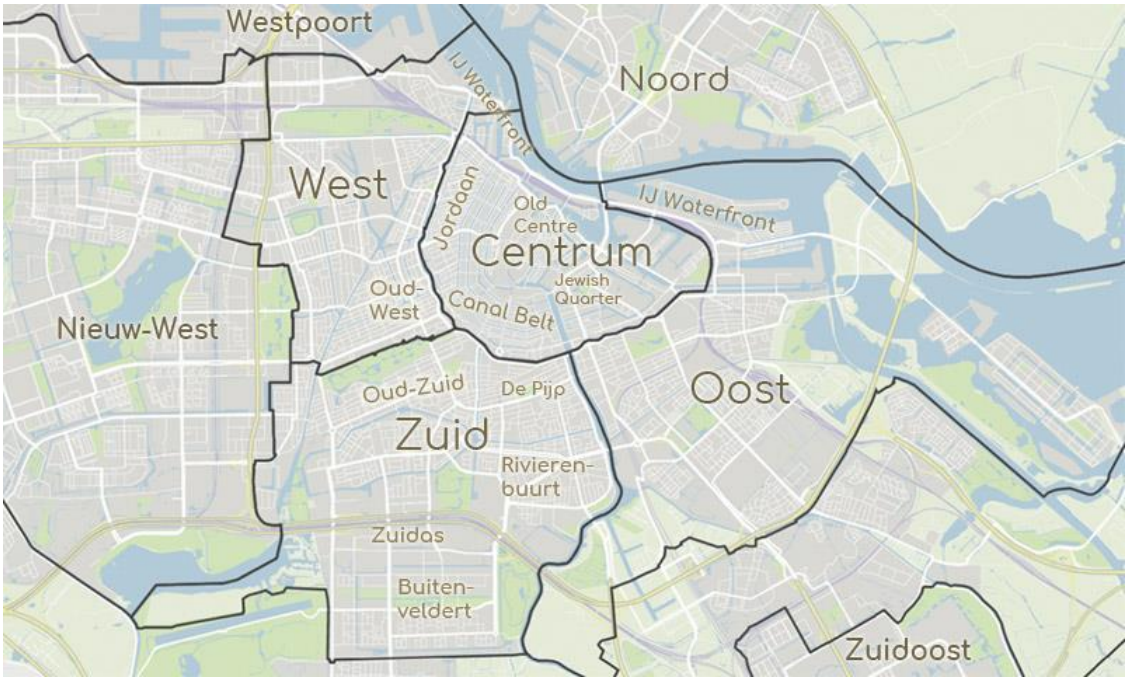
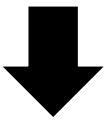


Figure 19: Amsterdam Neighborhoods layout

Με αυτόν τον τρόπο μειώνετε η πολυπλοκότητα αυτού του χαρακτηριστικού και αποκτά μια πιο ξεκάθαρη αναπαράσταση. Παρόλα αυτά όπως αναφέρθηκε κα πιο πάνω δεν αρκεί αυτό, αλλά πρέπει να εφαρμοστεί και one-hot encoding.

Συνεπώς, εφαρμόζοντας τις παραπάνω διαδικασίες το χαρακτηριστικό neighbourhood_cleansed παίρνει την εξής μορφή:

18716	Bos en Lommer
18717	Westerpark
18718	De Pijp - Rivierenbuurt
18719	Watergraafsmeer
18720	De Pijp - Rivierenbuurt



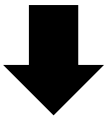
	Centrum	Zuid	West	Oost	Noord	Nieuw-West	Zuidoost
18716	0	0	1	0	0	0	0
18717	0	0	1	0	0	0	0
18718	0	1	0	0	0	0	0
18719	0	0	0	1	0	0	0
18720	0	1	0	0	0	0	0

Figure 20: neighbourhood_cleansed preprocessing

room_type

Εδώ επιλέχτηκε one-hot encoding για τις 4 διακριτές κατηγορίες αυτού του χαρακτηριστικού. Συνεπώς η προεπεξεργασία σε αυτά τα χαρακτηριστικών είχε το παρακάτω αποτέλεσμα:

0	Private room
1	Private room
2	Entire home/apt
3	Private room
4	Private room



	room_type_Entire home/apt	room_type_Hotel room	room_type_Private room	room_type_Shared room
0	0	0	1	0
1	0	0	1	0
2	1	0	0	0
3	0	0	1	0
4	0	0	1	0

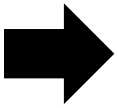
Figure 21: room_type preprocessing

bathrooms_text

Εδώ επιλέχτηκε μια παραλλαγή του one hot encoding που να αναπαριστά όσο καλυτέρα γίνεται το χαρακτηριστικό. Τα μπάνια χωρίστηκαν σε τρεις κατηγορίες baths, shared baths και private baths και ανάλογα με το πλήθος εμφανίσεων τους, συμπληρώθηκε η αντίστοιχη τιμή. Αυτή η προσέγγιση επιλέχτηκε καθώς από την μια υπάρχουν τρεις διακριτές κατηγορίες οπότε έπρεπε να δημιουργηθούν τρεις στήλες, μια για την κάθε μια. Αλλά από την άλλη υπήρχαν διαφοροποιήσεις εντός της κάθε κατηγορίας και κατ' επέκταση επιλέχθηκε να σημειώνεται η συχνότητα αντί για 0 και 1 που συμβαίνει στο κλασικό one hot encoding.

Συνεπώς η προεπεξεργασία σε αυτά τα χαρακτηριστικών είχε το παρακάτω αποτέλεσμα:

0	1.5 shared baths
1	1 private bath
2	1 bath
3	1.5 baths
4	1 shared bath



	baths	shared_baths	private_baths
0	0.0	1.5	0.0
1	0.0	0.0	1.0
2	1.0	0.0	0.0
3	1.5	0.0	0.0
4	0.0	1.0	0.0

Figure 22: bathrooms_text preprocessing

bedrooms και beds

Εδώ, καθώς οι τιμές των δυο αυτών χαρακτηριστικών είναι ήδη αριθμοί που εκφράζουν την συχνότητα εμφάνισης του κάθε χαρακτηριστικού, χρειάστηκε μόνο να διαχειριστούμε τις εκλιπούσες τιμές. Σε αυτήν την περίπτωση επιλέχτηκε να συμπληρωθούν όλες με την επικρατούσα τιμή, καθώς σε όλες τις υπάρχουσες περιπτώσεις η συχνότητα διέτρεχε ακέραιες τιμές και συνεπώς αν συμπληρωνόταν με την μέση τιμή, θα έπαιρνε δεκαδικές και δεν θα εξέφραζε την φυσική σημασία του χαρακτηριστικού

Συνεπώς η προεπεξεργασία σε αυτά τα χαρακτηριστικών είχε το παρακάτω αποτέλεσμα:

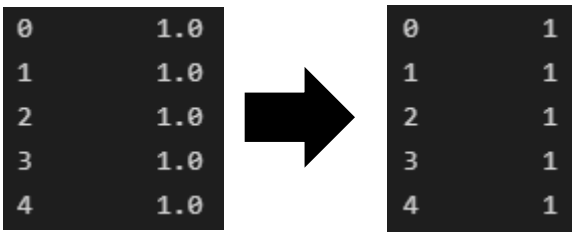
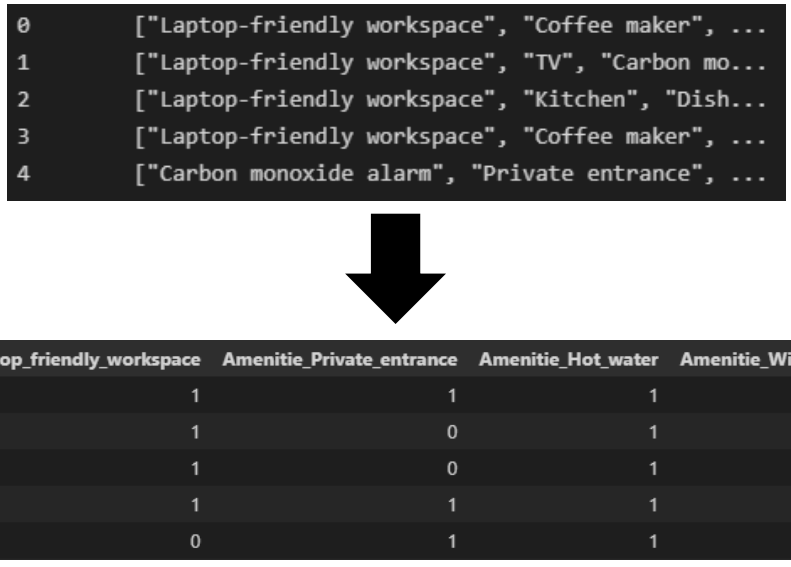


Figure 23: bedrooms and beds preprocessing

amenities

Εδώ επιλέχτηκε μια παραλλαγή του one hot encoding. Καθώς η τιμές του χαρακτηριστικού αναφέρονται σε επιπρόσθετες ανέσεις που διαθέτη το κάθε κατάλυμα, πολλοί hosts έχουν προσθέσει δικές τους επιλογές, μερικές φορές με κώδικές ονομασίες και συχνά σε αδόμητο κείμενο. Κατ' επέκταση για την αξιοποίησή αυτού του χαρακτηριστικού χρειάζεται μια καλή κωδικοποίηση. Συγκεκριμένα δημιουργήθηκαν γενικότερες κατηγορίες ανέσεων και στην συνέχεια ανάλογα με το πλήθος ανέσεων που κατείχε το κάθε κατάλυμα στην συγκεκριμένη κατηγορία, συμπληρώθηκε η αντίστοιχη συχνότητα. Συνολικά βρέθηκαν 232 διαφορετικές ανέσεις που συμπλήρωναν οι hosts, οι οποίες ομαδοποιήθηκαν σημασιολογικά σε 25 κατηγορίες.

Συνεπώς η προεπεξεργασία σε αυτά τα χαρακτηριστικών είχε το παρακάτω αποτέλεσμα:



	Amenitie_Laptop_friendly_workspace	Amenitie_Private_entrance	Amenitie_Hot_water	Amenitie_Wifi	Amenitie_TV
0	1	1	1	1	0
1	1	0	1	1	1
2	1	0	1	1	1
3	1	1	1	1	1
4	0	1	1	1	0

Figure 24: amenities preprocessing

price

Εδώ χρειάστηκε να καθαριστούν τα δεδομένα, καθώς οι τιμές του χαρακτηριστικού περιείχαν το σύμβολο του δολαρίου. Εκτός αυτού, δεν χρειάστηκε κάποια περεταίρω προεπεξεργασία.

Συνεπώς η προεπεξεργασία σε αυτά τα χαρακτηριστικών είχε το παρακάτω αποτέλεσμα:

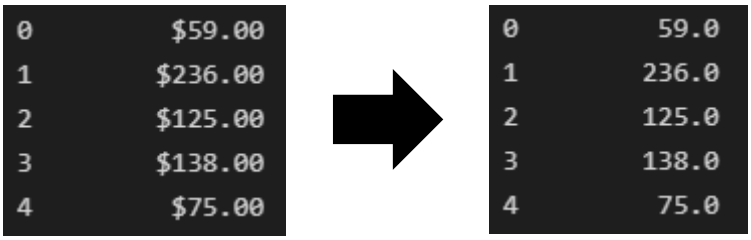


Figure 25: price preprocessing

4.3.2. Διαχωρισμός σε σύνολο Εκπαίδευσης και Ελέγχου (Train – Test Split)

Σε αυτό το σημείο, κρίνετε αναγκαίο να διαχωριστούν τα δεδομένα σε δυο υποσύνολα, το σύνολο εκπαίδευσης και το σύνολο ελέγχου. Αυτά τα δυο σύνολα είναι απαραίτητα για την εκπαίδευση μοντέλων μηχανικής μάθησης. Ο λόγος που γίνετε αυτός ο διαχωρισμός είναι για να μπορεί να είναι μελλοντικά δυνατή η αξιολόγηση των μοντέλων που εκπαιδεύτηκαν στο σύνολο εκπαίδευσης. Πιο συγκεκριμένα, τα μοντέλα θα αξιοποιήσουν το σύνολο εκπαιδεύσης για να μοντελοποιήσουν το πρόβλημα και στην συνέχεια για να αξιολογηθεί η επίδοση τους θα χρειαστεί ένα σύνολο του οποίου τα δεδομένα να είναι καινούρια για το μοντέλο και συνεπώς να υπολογιστεί σωστά η κάθε μετρική αξιολόγησης. Διαφορετικά, αν αυτό συνέβαινε στο ίδιο σύνολο, θα είναι σαν να «κλέβει» το μοντέλο, καθώς τα δεδομένα στα οποία θα αξιολογείτε θα είναι ακριβώς τα ίδια βάση των οποίων εκπαιδεύτηκε συνεπώς η επίδοση του δεν θα είναι απαραίτητα η ίδια σε δεδομένα που δεν έχει ξανά δει.

Επιπλέον, για τον συγκεκριμένο διαχωρισμό θα ακολουθηθεί η παρακάτω αναλογία. Το 70% του αρχικού συνόλου θα αξιοποιηθεί ως σύνολο εκπαίδευσης. Ενώ το υπόλοιπο 30% θα χρησιμοποιηθεί ως σύνολο ελέγχου.

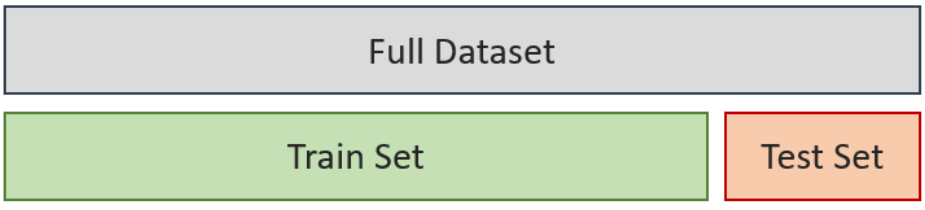


Figure 26: Train - Test Split

Ακόμα σημειώνεται πως σε όλα τα επόμενα στάδια προεπεξεργασίας, όλες οι παράμετροι των μεθόδων υπολογίζονται βάση του συνόλου εκπαίδευσης και μόνο η εφαρμογή των μεθοδολογιών πραγματοποιείται και στα δυο σύνολα. Ο λόγος που πραγματοποιείτε αυτό είναι για να αποφευχθεί κάθε είδος μεροληψίας και για να προσομοιωθούν στο σύνολο ελέγχου όσο καλύτερα γίνετε η συνθήκες πραγματικής πρόβλεψης. Αναλυτικότερα, όταν θα χρειαστεί να πραγματοποιήσει προβλέψεις το εκπαιδευμένο μοντέλο σε νέα δεδομένα, κάθε φορά θα

εφαρμόζεται η ίδια διαδικασία προεπεξεργασίας, έτσι ώστε τα δεδομένα να δίνονται στο μοντέλο σε παρόμοια μορφή με αυτήν της εκπαίδευσης.

Για αυτόν τον λόγω σε όλα τα παρακάτω βήματα προεπεξεργασίας, παρότι δεν θα αναφέρεται, οι παράμετροι των μεθοδολογιών θα υπολογίζονται μόνο στο σύνολο εκπαίδευσης και τελικά θα εφαρμόζονται και στα δύο.

4.3.3. Διαχείρηση Παράτυπων Σημείων (Handling Outliers)

Σε αυτό το σημείο, καθώς έχουν αναπαρασταθεί όλα τα δεδομένα με αριθμητικές τιμές και καθώς δεν έχει ακόμα πραγματοποιηθεί κανονικόποιηση των δεδομένων είναι το κατάλληλο σημείο για να εξεταστεί αν υπάρχουν παράτυπα σημεία ως προς το χαρακτηριστικό πρόβλεψης. Αρχικά, σχεδιάζεται το ιστόγραμμα της κατανομής του χαρακτηριστικού πρόβλεψης:

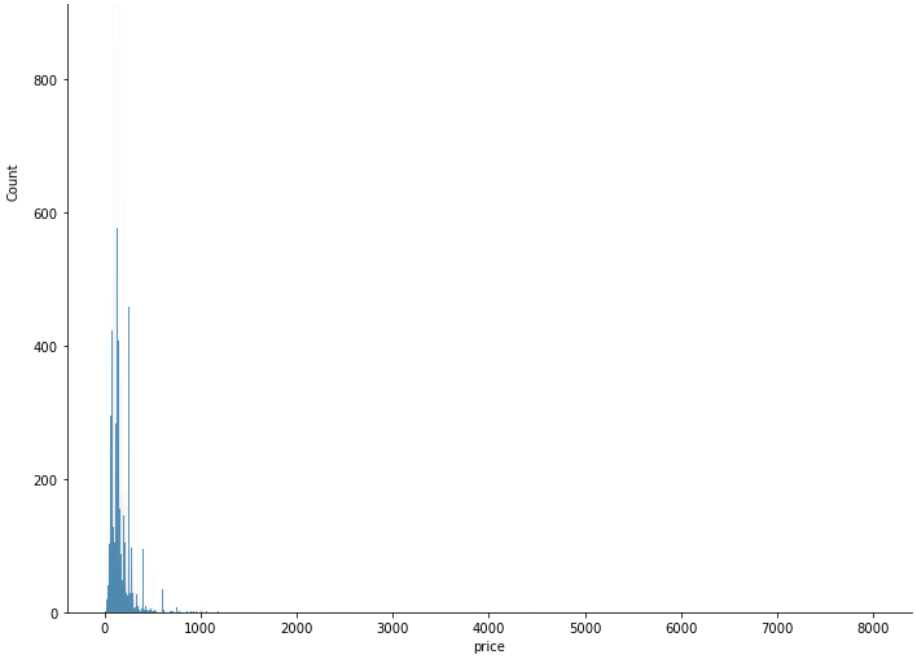


Figure 27: Price Histogram Before Outlier Removal

Όπως φαίνεται από το παραπάνω γράφημα, το μεγαλύτερο μέρος των δεδομένων συσσωρεύεται κάτω αριστερά. Παρόλα αυτά στον οριζόντιο άξονα, η συνάρτηση σχεδιασμού του γραφήματος επιλέγει να αποτυπώσει τιμές μέχρι το 8000. Αυτό σημαίνει πως ακόμα και να μην φαίνονται εδώ με το μάτι, υπάρχουν τιμές που πλησιάζουν κοντά σε αυτό το νούμερο. Παρόλα αυτά είναι πολύ λίγες σε αριθμό, για αυτό και δεν γίνονται αντιληπτές στο γράφημα. Κατ' επέκταση εφαρμόζοντας την μέθοδο ανεύρεσης παράτυπων σημείων που βασίζεται στο ενδοτεταρτομοροιακό εύρος αναγνωρίζονται ακριβώς ποιες είναι αυτές οι τιμές. Ποιο συγκεκριμένα υπολογίστηκαν τα: $Q1 = 95$, $Q3 = 180$, $IQR = 85$ και κατ' επέκταση, το αποδεκτό εύρος τιμών είναι το $[-32.5, 307.5]$. Συνεπώς, συνολικά αναγνωρίζονται 1025 παράτυπά σημεία. Τέλος σε αυτήν την διπλωματική επιλέχτηκαν να απομακρυνθούν τα συγκεκριμένα σημεία και άρα το γράφημα του ιστογράμματος του χαρακτηριστικού προβλέψεις μετατρέπεται όπως φαίνεται παρακάτω:

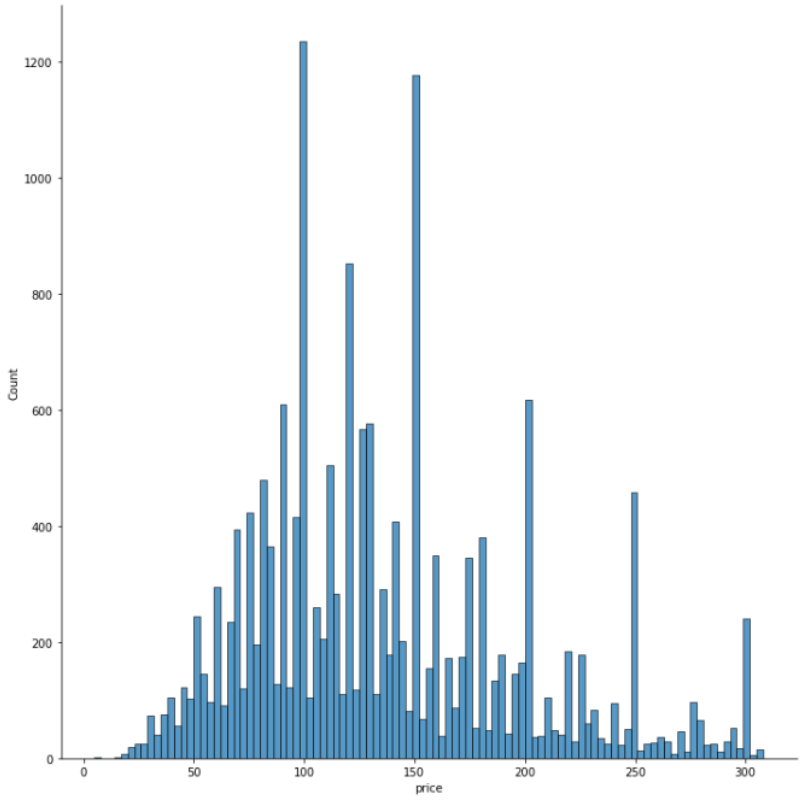


Figure 28: Price Histogram After Outlier Removal

Όπως φαίνεται από το νέο γράφημα, τώρα που έχουν αφαιρεθεί όλα τα παράτυπα σημεία, η κατανομή φαίνεται πολύ πιο ξεκάθαρα, χωρίς να αφήνει κενές περιοχές.

4.3.4. Κανονικοποίηση (Scaling)

Ως τελευταίο στάδιο της προεπεξεργασίας των δεδομένων εφαρμόστηκε κανονικοποίηση. Όλα τα δεδομένα κανονικοποιήθηκαν ανεξάρτητα στο $[0,1]$, όπως περιγράφεται αναλυτικά στο κεφάλαιο 3.2.3. Κατ' επέκταση, το σύνολο δεδομένων πήρε την παρακάτω μορφή:

	host_response_rate	host_acceptance_rate	host_is_superhost	host_identity_verified	accommodates	bedrooms
0	0.90084	1.00	1.0	1.0	0.066667	0.0
1	0.90084	1.00	0.0	1.0	0.066667	0.0
2	1.00000	0.39	1.0	0.0	0.133333	0.0
3	0.90084	1.00	1.0	1.0	0.066667	0.0
4	1.00000	0.95	1.0	1.0	0.066667	0.0

Figure 29: Dataset Final Overview

4.4. Εκπαίδευση Μοντέλων Μηχανικής Μάθησης (ML Models Training)

Σε όλα τα μοντέλα μηχανικής μάθησης που ακολουθούν εφαρμόστηκε η παρακάτω μεθοδολογία εκπαίδευσης. Για το κάθε μοντέλο, κατασκευάστηκε ένα "χαλαρό" πλέγμα συνδυασμού υπερπαραμέτρων πάνω στον χώρο υπερπαραμέτρων του συγκεκριμένου μοντέλου και στην συνέχεια εφαρμόστηκε Αναζήτηση Πλέγματος με Διασταυρωμένη επικύρωση ώστε να αξιολογηθεί ο εκάστοτε συνδυασμός υπερπαραμέτρων. Στην συνέχεια, αφότου εντοπιζόταν η περιοχή στην οποία το μοντέλο συμπεριφερόταν βελτηστα οσο αναφορά την πρόβλεψη του εξαρτημένου χαρακτηριστικού, επαναλαμβανόταν η παραπάνω διαδικασία αλλά με αυστηρότερο πλέγμα μόνο γύρο από την περιοχή της βέλτιστης συμπεριφοράς. Η παραπάνω διαδικασία ολοκληρωνόταν όταν το εκάστοτε μοντέλο κατέληγε στον καλύτερο συνδυασμό υπερπαραμέτρων. Στην συνέχεια, παρουσιάζονται όλες οι υπερπαράμετροι που δοκιμάστηκαν στο κάθε μοντέλο μηχανικής μάθησης που εκπαιδεύτηκε.

4.4.1. Δεντρα Παλινδρόμησης (Regression Tree)

Για το Δέντρο Παλινδρόμησης δοκιμάστηκαν οι παρακάτω υπερπαράμετροι:

Table 4: Υπερπαράμετροι Δέντρου Παλινδρόμησης

ΥΠΕΡΠΑΡΑΜΕΤΡΟΙ	ΤΙΜΕΣ
max_depth	$5, 6, 7, \dots, 50$
min_samples_leaf	$1, 2, 3, 4, 5$
min_samples_split	$1, 2, 3, 4, 5$

Ως καλύτερος συνδυασμός υπερπαραμέτρων βρέθηκε ο: max_depth:5, min_samples_leaf:2, min_samples_split:2. Με $R^2 = 0.364$

4.4.2. Τυχαία Δάση (Random Forest)

Για το Τυχαίο Δάσος Παλινδρόμησης δοκιμάστηκαν οι παρακάτω υπερπαράμετροι:

Table 5: Υπερπαράμετροι Τυχαίου Δάσους Παλινδρόμησης

ΥΠΕΡΠΑΡΑΜΕΤΡΟΙ	ΤΙΜΕΣ
max_depth	$5, 6, 7, \dots, 50$
min_samples_leaf	$1, 2, 3, 4, 5$
min_samples_split	$1, 2, 3, 4, 5$
n_estimators	$50, 100, 150, 200, \dots, 1000$

Ως καλύτερος συνδυασμός υπερπαραμέτρων βρέθηκε ο: max_depth:7, min_samples_leaf:3, min_samples_split:2, n_estimators:100. Με $R^2 = 0.405$

4.4.3. Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machine – SVM)

Για τις Μηχανές Διανυσματικής Υποστήριξης σε Παλινδρόμηση, δοκιμάστηκαν τρεις βασικές κατηγορίες μοντέλων. Τα μοντέλα που στηρίζονται σε γραμμικούς, πολυωνυμικούς και gaussian πυρήνες. Αναλόγως τον πυρήνα δοκιμάστηκαν οι αντίστοιχες υπερπαραμετροι:

Γραμμικός Πυρήνας:

Table 6: Υπερπαράμετροι Μηχανών Διανυσματικής Υποστήριξης με Γραμμικό Πυρήνα

ΥΠΕΡΠΑΡΑΜΕΤΡΟΙ	ΤΙΜΕΣ
C	$2^{-5}, 2^{-4}, 2^{-3}, \dots, 2^{7}$

Ως καλύτερος συνδυασμός υπερπαραμέτρων βρέθηκε ο: C:0.25. Με $R^2 = 0.392$

Πολυωνυμικός Πυρήνας:

Table 7: Υπερπαράμετροι Μηχανών Διανυσματικής Υποστήριξης με Πολυωνυμικό Πυρήνα

ΥΠΕΡΠΑΡΑΜΕΤΡΟΙ	ΤΙΜΕΣ
degree	$2, 3, 4, \dots, 10$
C	$2^{-5}, 2^{-4}, 2^{-3}, \dots, 2^{7}$

Ως καλύτερος συνδυασμός υπερπαραμέτρων βρέθηκε ο: degree: 2, C: 0.0625. Με $R^2 = 0.401$

Ακτινικός Πυρήνας:

Table 8: Υπερπαράμετροι Μηχανών Διανυσματικής Υποστήριξης με Ακτινικό Πυρήνα

ΥΠΕΡΠΑΡΑΜΕΤΡΟΙ	ΤΙΜΕΣ
gamma	$2^{-10}, 2^{-9}, 2^{-8}, \dots, 2^{4}$
C	$2^{-5}, 2^{-4}, 2^{-3}, \dots, 2^{7}$

Ως καλύτερος συνδυασμός υπερπαραμέτρων βρέθηκε ο: gamma:0.0625, C:0.125. Με $R^2 = 0.401$

4.4.4. XGBoost

Για τον αλγόριθμο XGBoost, δοκιμάστηκαν οι παρακάτω υπερπαράμετροι:

Table 9: Υπερπαράμετροι XGBoost

ΥΠΕΡΠΑΡΑΜΕΤΡΟΙ	ΤΙΜΕΣ
max_depth	$5, 6, 7, \dots, 50$
learning_rate	$0.01, 0.02, 0.03, \dots, 0.1$
n_estimators	$50, 100, 150, 200, \dots, 1000$
colsample_bytree	$0.1, 0.2, 0.3, \dots, 0.9$

Ως καλύτερος συνδυασμός υπερπαραμέτρων βρέθηκε ο: max_depth:8, learning_rate:0.01, n_estimators:400, colsample_bytree:0.2. Με $R^2 = 0.433$.

5. Results and Discussion

Συνοψίζοντας τις επιδόσεις όλων των μοντέλων μηχανικής μάθησης που επιλέχθηκαν για την επίλυση του συγκεκριμένου προβλήματος, προκύπτουν τα ακόλουθα αποτελέσματα:

Table 10: Αξιολόγηση Μοντέλων

Αλγόριθμος Μηχανικής Μάθησης	Mean Squerred Error στο Train Set	Mean Squerred Error στο Test Set	R^2-Score στο Train Set	R^2-Score στο Test Set
Δέντρο Παλινδρόμησης	36.036	36.744	0.387	0.364
Τυχαίο Δάσος Παλινδρόμησης	33.584	35.417	0.466	0.405
Μηχανή Διανυσματικής Υποστήριξης με Γραμμικό Πυρήνα	35.081	35.620	0.396	0.392
Μηχανή Διανυσματικής Υποστήριξης με Πολυωνυμικό Πυρήνα	34.355	35.235	0.427	0.401
Μηχανή Διανυσματικής Υποστήριξης με Ακτινικό Πυρήνα	34.441	35.300	0.426	0.401
XGBoost	**31.059**	**34.779**	**0.547**	**0.433**

Όπως φαίνεται από τον πίνακα αξιολόγησης των μοντέλων μηχανικής μάθησης, η κορυφαία απόδοση στο σύνολο ελέγχου φαίνεται από τον αλγόριθμο XGBoost (R^2: 0,433). Ωστόσο, ο αλγόριθμος XGBoost είναι επίσης αυτός που δείχνει τη μεγαλύτερη απόκλιση μεταξύ της απόδοσης στο σύνολο εκπαίδευσης και του συνόλου ελέγχου. Ωστόσο, αυτή η τάση του XGBoost να υπερεκπαιδεύεται δικαιολογείται από την κατασκευαστική του λογική, αφού όπως γνωρίζουμε χτίζεται διαδοχικά εστιάζοντας σε δεδομένα που δυσκολεύεται να προβλέψει επαρκώς. Αυτή λοιπόν η έμφαση στα 'δύσκολα' δεδομένα είναι και το στοιχείο που του επιτρέπει τελικά να πετύχει τις καλύτερες επιδόσεις.

Γενικά, παρατηρούμε ότι με τα διαθέσιμα δεδομένα, όλα τα μοντέλα επιτυγχάνουν καλές προβλέψεις και οι αποκλίσεις μεταξύ πραγματικών τιμών και προβλέψεων κυμαίνονται στο $\pm 30 - 35$ ευρώ, γεγονός που, δεδομένης της φύσης του προβλήματος, υποδηλώνει σχετικά χαμηλή απόκλιση.

Ακόμα εξετάζοντας ποια από τα χαρακτηρίστηκα των δεδομένων αξιοποιήθηκαν περισσότερο από τα μοντέλα δεντρικού χαρακτήρα. Προκύπτει πως για την εκτίμηση της τιμής ενοικίασης ενός καταλύματος, κυρίαρχο ρόλο (με σειρά σημαντικότητας) παίζουν τα παρακάτω χαρακτηριστικά:

Table 11: Σημαντικότερα Χαρακτηρίστηκα

A/A	Χαρακτηριστικό	Σημαντικότητα	Περιγραφή
1	accommodates	43%	Πλήθος ατόμων που μπορούν να νοικιάσουν τον χώρο για μια βραδιά
2	room_type_Entire home/apt	23%	Αν το κατάλυμα αποτελεί ολόκληρο σπίτι ή διαμέρισμα
3	bedrooms	8%	Το πλήθος των υπνοδωματίων

4	Centrum	3%	Αν το κατάλυμα βρίσκετε στο Κέντρο του Άμστερνταμ
5	baths	3%	Το πλήθος των μπάνιών
6	Amenitie_Kitchen_appliances	2%	Τι ποσοστό συσκευών κουζίνας διαθέτει το εν λόγο κατάλυμα
7	Amenitie_Safety	1%	Τι ποσοστό ασφάλειας κατέχει το εν λόγο κατάλυμα
8	Amenitie_Bathroom_appliances	1%	Τι ποσοστό συσκευών μπάνιου διαθέτει το εν λόγο κατάλυμα
9	Amenitie_Bathroom_staff	1%	Τι ποσοστό αντικειμένων μπάνιου διαθέτει το εν λόγο κατάλυμα
10	Amenitie_Luxurious	1%	Τι ποσοστό πολυτελούς εξοπλισμού κατέχει το εν λόγο κατάλυμα

Όπως βλέπουμε από τον πίνακα, τα χαρακτηριστικά που παίζουν τον πιο ουσιαστικό ρόλο είναι όλα εκείνα που επηρεάζουν τον αριθμό των ατόμων που μπορούν να μείνουν στο κατάλυμα (accommodates, bedrooms, baths). Το μέγεθος και η χωρητικότητα του ενοικιαζόμενου ακινήτου έχουν σημαντικό αντίκτυπο στην τιμή ενοικίασης, καθώς τα μεγαλύτερα ακίνητα που μπορούν να φιλοξενήσουν περισσότερα άτομα έχουν γενικά υψηλότερες τιμές.

Επιπλέον, αν το ενοικιαζόμενο κατάλυμα είναι ένα ολόκληρο σπίτι ή ένα ολόκληρο διαμέρισμα (room_type_Entire home/apt) έχει σημαντική επίδραση στην τιμή ενοικίασης. Αυτό το χαρακτηριστικό αντικατοπτρίζει το επίπεδο ιδιωτικότητας και αποκλειστικότητας που παρέχει το ενοικιαζόμενο ακίνητο, με ολόκληρα σπίτια ή διαμερίσματα να έχουν συνήθως υψηλότερες τιμές σε σύγκριση με κοινόχρηστα ή ιδιωτικά δωμάτια.

Το γεωγραφικό χαρακτηριστικό του εάν το κατάλυμα βρίσκεται στο κέντρο του Άμστερνταμ (Centrum) παίζει επίσης σημαντικό ρόλο στον καθορισμό της τιμής ενοικίασης. Τα ακίνητα που βρίσκονται σε προνομιακές τοποθεσίες, όπως τα κέντρα των πόλεων, τείνουν να είναι πιο επιθυμητά λόγω της εγγύτητάς τους σε ανέσεις, αξιοθέατα και βολικές επιλογές μεταφοράς. Ως εκ τούτου, συχνά έρχονται με μια κορυφαία τιμή.

Επιπλέον, η παρουσία συγκεκριμένων ανέσεων, όπως οι συσκευές κουζίνας (Amenitie_Kitchen_appliances) και τα χαρακτηριστικά ασφαλείας (Amenitie_Safety), μπορεί επίσης να επηρεάσει την τιμή ενοικίασης. Τα ακίνητα που είναι εξοπλισμένα με σύγχρονες συσκευές κουζίνας ή ενισχυμένα μέτρα ασφαλείας τείνουν να προσελκύουν πιο απαιτητικούς επισκέπτες και ενδέχεται να έχουν υψηλότερες τιμές ενοικίασης..

6. Conclusions

Η προσπάθεια ακριβής πρόβλεψης της τιμής ενοικίασης ενός καταλύματος αποδεικνύετε να είναι ένα πολυδιάστατο πρόβλημα με πολλές δυσκολίες. Παρόλα αυτά η ανάγκη ανεύρεσης μεθοδολογιών που θα επιλύουν βέλτιστα το παραπάνω πρόβλημα είναι αναγκαία, καθώς επηρεάζει σημαντικά την καθημερινή ποιότητα ζωής του μέσου ανθρώπου. Στην συγκεκριμένη διπλωματική, αποδεικνύουμε πως οι σύγχρονοι αλγόριθμοι μηχανικής μάθησης μπορούν και πετυχαίνουν κορυφαίες επιδόσεις. Έτσι δημιουργώντας ένα μοντέλο πρόβλεψης της τιμής ενοικίασης, κατασκευάζεται και ένας γενικός κανόνας αντιστοίχισης των χαρακτηριστικών ενός καταλύματος στις τιμές ενοικιάσεις. Και κατ' επέκταση, χρησιμοποιώντας αυτόν τον κανόνα μπορούμε και απαντάμε σε ένα από τα ερωτήματα της συγκεκριμένης διπλωματικής που είναι η εύρεση των κορυφαίων χαρακτηριστικών που επηρεάζουν την τιμή ενός καταλύματος, βλέποντας το ποσοστό επιρροής του εκάστοτε χαρακτηριστικού στον καθορισμό της τιμής. Τέλος, αναφορικά με το ερώτημα του ποιος αλγόριθμος μηχανικής μάθησης αποτελεί τον επικρατέστερο, κρίνοντας τις επιδώσεις τους. Η απάντηση είναι, ο XGBoost. Όλα τα μοντέλα φαίνεται να πετυχαίνουν εξαιρετικά ανταγωνιστικές επιδώσεις και με μικρή απόκλιση των προβλέψεων από τις πραγματικές τιμές. Αλλά το χαρακτηριστικό του XGBoost με το οποίο εστιάζει σε δύσκολα δεδομένα είναι αυτό που του επιτρέπει να ξεπεράσει όλα τα υπόλοιπα μοντέλα μηχανικής μάθησης.

7. Future Work

Στόχος της συγκεκριμένης διπλωματικής αποτέλεσε οι διερεύνηση βασικών αλγορίθμων μηχανικής μάθησης στην πρόβλεψη της αξίας ενοικίασης καταλυμάτων που παρέχονται μέσω της Airbnb, όπως και πραγματοποιήθηκε. Παρόλα αυτά, κατά την διάρκεια προεπεξεργασίας καθώς και κατά την επιλογή των αλγορίθμων που εκπαιδεύτηκαν σε προηγούμενες ενότητες, παρατηρήθηκαν τα παρακάτω:

1. Δεδομένα που περιείχαν αδόμητο κείμενο δεν μπόρεσαν να προεπεξεργαστούν κατάλληλος για να τροφοδοτηθούν στους επιλεγμένους αλγορίθμους και προτιμήθηκε να παραληφθούν. Ως μελλοντικό βήμα, θα παρουσίαζε ιδιαίτερο ενδιαφέρων η χρήση ενός γλωσσικού νευρωσικού δικτιού για την εξαγωγή χρήσιμων χαρακτηριστικών και αναπαράσταση του κάθε κειμένου σε έναν χώρο Ν διαστάσεων. Με αυτόν τον τρόπο, κάθε κείμενο θα αποκτούσε μια αναπαράσταση σταθερής διάστασης και θα μπορούσε να κωδικοποιηθεί καταλλήλως για να εισαχθεί ως επιπλέον χαρακτηριστικό στα μοντέλα μηχανικής μάθησης.

2. Ένας ακόμα παράγοντας που θα προσέφερε στην καλύτερη εκτίμηση των τιμών ενοικίασης είναι η αξιοποίηση των εικόνων που παρουσιάζονται στην σελίδα του εκάστοτε καταλύματος. Παρόλα αυτά, όπως και στη περίπτωση 1, αυτό συνεπάγεται την χρήση βαθιών συνελλικτικών νευρωνικών δικτύων (CNNs) για την εξαγωγή κωδικοποιήσιμων χαρακτηριστικών σταθερής διάστασης για κάθε εικόνα.

3. Ακόμα, για την εισαγωγή μεγαλύτερης μεταβλητότητας στο σύνολο δεδομένων και κατ' επέκταση την εκπαίδευση ενός μοντέλου που θα αναπτύξει έναν ακόμα ισχυρότερο και γενικό κανόνα για την τιμολόγηση των καταλυμάτων. Θα ήταν χρήση η εισαγωγή δεδομένων και από άλλες πόλεις, εκτός του Άμστερνταμ.

Τέλος, όπως γίνετε φανερό, ο τομέας της προβλέψεις τιμών είτε ενοικίασης, είτε πώλησης ακινήτων αποτελεί ένα αντικείμενο που χρήζει περαιτέρω ερεύνα και είναι πλούσιο σε αποτελέσματα αξιοποιώντας μεθόδους μηχανικής μάθησης.

8. Bibliograph

[1] A. Singh, L. Daniel, E. Baker and R. Bentley, "Housing Disadvantage and Poor Mental Health: A Systematic Review," *American Journal of Preventive Medicine,* vol. 57, no. 2, pp. 262-272, 2019.

[2] R. Cooper, C. Boyko and R. Codinhoto, "The effect of the physical environment on mental wellbeing," in *Mental capital and wellbeing*, 2010, pp. 967-1006.

[3] R. Bentley, E. Baker, K. Mason, S. V. Subramanian and A. M. Kavanagh, "Association Between Housing Affordability and Mental Health: A Longitudinal Analysis of a Nationally Representative Household Survey in Australia," *American Journal of Epidemiology,* vol. 174, no. 7, p. 753–760, 2011.

[4] S. B. Jha, V. Pandey, R. K. Jha and R. F. Babiceanu, "Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study," 2020.

[5] D. Guttentag, "Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector," *Current Issues in Tourism,* pp. 1192-1217, 2015.

[6] D. Islam, B. Li, K. S. Islam, R. Ahasan, R. Mia and E. Haque, "Airbnb rental price modeling based on Latent Dirichlet Allocation and MESF-XGBoost composite model," *Machine Learning with Applications,* vol. 7, 2022.

[7] R. R. R. Botsman, ”Product service systems,” in What’s Mine, New York: HarperCollins, 2010, pp. 106 - 108.

[8] "Fast facts," Dec 2022. [Online]. Available: https://news.airbnb.com/about-us/.

[9] D. Hill, "How Much Is Your Spare Room Worth?," *IEEE Spectrum,* vol. 52, no. 9, pp. 32-58, September 2015.

[10] T. Cai, K. Han and H. Wu, "Melbourne airbnb price prediction," 2019.

[11] A. Sihabuddin, "An Extreme Learning Machine Model Approach on Airbnb Base Price Prediction," *International Journal of Advanced Computer Science and Applications,* vol. 11, no. 11, January 2020.

[12] P. Ye, J. Qian, J. Chen, C. Wu, Y. Zhou, S. D. Mars, F. Yang and L. Zhang, "Customized Regression Model for Airbnb Dynamic Pricing," *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* pp. 932-940, July 2018.

[13] X. Z. Y. Z. Yuanhang Luo, "Predicting Airbnb Listing Price Across Different Cities," 2019.

[14] R. Pouya, N. Liubov and R. Hoormazd, "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis," *International Cross-Domain Conference for Machine Learning and Knowledge Extraction,* pp. 173-184, 2021.

[15] L. Lewis, "Predicting airbnb prices with machine learning and deep learning," May 2019.

[16] S. M. a. N. J. T. Mohd, "Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia," *International Journal of Recent Technology and Engineering (IJRTE),* vol. 8, no. 11, pp. 542-546, September 2019.

[17] B. McNeil, "Price prediction in the sharing economy: A case study," 2020.

[18] R. Deboosere, D. J. Kerrigan, D. Wachsmuth and A. M. Elgeneidy, "Location, location and professionalization: a multilevel hedonic analysis of airbnb listing prices and revenue," *Regional Studies, Regional Science,,* vol. 6, no. 1, p. :143–156, 2019.

[19] D. Wang and J. L. Nicolau, "Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com," *International Journal of Hospitality Management,* vol. 62, pp. 120-131, 2017.

[20] "What is artificial intelligence (AI)? definition, benefits and use cases.," [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence.

[21] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain and A. J. Aljaaf, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," in *Supervised and Unsupervised Learning for Data Science*, 2019.

[22] M. Zimmer, P. Viappiani and P. Weng, "Teacher-Student Framework: a Reinforcement Learning Approach," 2014.

[23] DanB, "Using Categorical Data with One Hot Encoding," 2017. [Online]. Available: https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/notebook.

[24] K. Singh and S. Upadhyaya, "Outlier Detection: Applications And Techniques," *International Journal of Computer Science Issues,* vol. 9, no. 1, 2012.

[25] H. P. Vinutha, B. Poornima and B. M. Sagar, "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset," in *Information and Decision Sciences*, 2018, p. 511–518.

[26] Y. Verma, "Why Data Scaling is important in Machine Learning & How to effectively do it," 2021. [Online]. Available: https://analyticsindiamag.com/why-data-scaling-is-important-in-machine-learning-how-to-effectively-do-it/.

[27] J. Hale, "Scale, Standardize, or Normalize with Scikit-Learn," Medium, 2019. [Online]. Available: https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02.

[28] J. R. Quinlan, "Induction of decision trees," *Machine Learning,* vol. 1, no. 1, p. 81–106, 1986.

[29] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification and regression trees, Monterey: Brooks/Cole Publishing, 1984.

[30] T. K. Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition,* vol. 1, pp. 278-282, 1995.

[31] S. Polamuri, "HOW THE RANDOM FOREST ALGORITHM WORKS IN MACHINE LEARNING," 2017. [Online]. Available: https://dataaspirant.com/random-forest-algorithm-machine-learing/.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, p. 273–297, 1995.

[33] D. R. Chen, Q. Wu, Y. Ying and D. X. Zhou, "Support Vector Machine Soft Margin Classifiers: Error Analysis," *Machine Learning Research,* vol. 5, pp. 1143-1175, 2004.

[34] E. Kim, "Everything You Wanted to Know about the," 2017.

[35] M. Hofmann, "Support Vector Machines - Kernels and the," 2006.

[36] "An End-to-End Guide to Understand the Math behind XGBoost," 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/.

[37] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* p. 785–794, 2016.

[38] D. Berrar, "Cross-validation," 2018.

[39] A. C. S. Arlot, "A survey of cross-validation procedures for model selection," *2010,* vol. 4, pp. 40 - 79, 2010.

[40] D. C. Montgomery, Design and Analysis of Experiments, John Wiley & Sons, Inc, 2013.

[41] M. Feurer and F. Hutter, "Hyperparameter optimization," *Automated machine learning,* pp. 3-33, 2019.

[42] R. A. Dubin, "Predicting house prices using multiple listings data," *The Journal of Real Estate Finance and Economics,* vol. 17, pp. 35-59, 1998.

[43] Y. Ma, Z. Zhang, A. Ihler and B. Pan, "Estimating warehouse rental price using machine learning techniques," *International Journal of Computers, Communications & Control,* vol. 13, 2018.

[44] J. Gu, M. Zhu and L. Jiang, "Housing price forecasting based on genetic algorithm and support vector machine," vol. 38, pp. 3383-3386, 2011.

[45] Y. Li, Q. Pan, T. Yang and L. Guo, "Reasonable price recommendation on airbnb using multi-scale clustering," *35th Chinese Control Conference (CCC). IEEE,* p. 7038–7041, 2016.

[46] Z. Zhang, R. J. Chen, L. D. Han and L. Yang, "Key factors affecting the price of airbnb listings: A geographically weighted approach," vol. 9, p. 1635, 2017.

[47] A. Varma, A. Sarma, S. Doshi and R. Nair, "House price prediction using machine learning and neural networks," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT).,* p. 1936–1939, 2018.

[48] P. R. Kalehbasti, L. Nikolenko and H. Rezaei, "“Airbnb price prediction using machine learning and sentiment analysis," 2019.

[49] Y. Luo, X. Zhou and Y. Zhou, "Predicting airbnb listing price across different cities," 2019.