Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Κυβερνοασφάλεια και Επιστήμη Δεδομένων»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Αναμενόμενος Χρόνος Άφιξης και Πρόβλεψη Τοποθεσίας σε ένα Αστικό Σενάριο Κινητικότητας** **Expected Time of Arrival and Location Prediction in an Urban Mobility Scenario**
Ονοματεπώνυμο Φοιτητή	**ΚΑΤΣΑΔΟΥΡΟΥ ΗΛΙΑΝΑ**
Πατρώνυμο	**ΓΕΩΡΓΙΟΣ**
Αριθμός Μητρώου	**ΜΠΚΕΔ21021**
Επιβλέπων	**ΘΕΟΔΩΡΙΔΗΣ ΙΩΑΝΝΗΣ, ΚΑΘΗΓΗΤΗΣ**

Ημερομηνία Παράδοσης **ΣΕΠΤΕΜΒΡΙΟΣ 2024**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Γιάννης Θεοδωρίδης **Καθηγητής**	**Νίκος Πελέκης** **Αναπλ. Καθηγητής**	**Άγγελος Πικράκης** **Επίκ. Καθηγητής**
Όνομα Επώνυμο Βαθμίδα	Όνομα Επώνυμο Βαθμίδα	Όνομα Επώνυμο Βαθμίδα

Ευχαριστίες

Πρωτίστως, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στον επιβλέπων καθηγητή μου, κύριο Ιωάννη Θεοδωρίδη, για την ευκαιρία να συνεργαστώ μαζί του και τον Γεώργιο Θεοδωρόπουλο για τη συνεχή υποστήριξη κατά τη συγγραφή αυτής της διατριβής. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, και κυρίως τον πατέρα μου, για την στήριξη που μου προσφέρουν σε κάθε νέο μου βήμα.

Περίληψη

Η Μηχανική Μάθηση είναι ένας τομέας της πληροφορικής που αναπτύσσεται ραγδαία, και μπορεί να προσφέρει λύσεις σε σύνθετα και απαιτητικά προβλήματα. Αυτή η διπλωματική εργασία στοχεύει στην ανάπτυξη και σύγκριση συστημάτων για την πρόβλεψη των στάσεων αποβίβασης επιβατών, με χρήση δεδομένων δημόσιων μεταφορών από την πόλη της Ρίγας, χρησιμοποιώντας μοντέλα μηχανικής μάθησης. Επιπρόσθετα, μελετά εις βάθος αλγορίθμους όπως οι Decision Tree, k-Nearest Neighbors, Random Forest, Bagging, Gradient Boosting,Kernel Ridge Regression, Νευρωνικά Δίκτυα, XGBoost και LightGBM, αναλύοντας τη θεωρητική βάση, τους τρόπους υλοποίησης, και τις μετρικές απόδοσής τους. Στη συνέχεια οι αλγόριθμοι αυτοί εφαρμόζονται στα δεδομένα που μελετάμε, και η ακρίβεια πρόβλεψης που προσφέρουν αξιολογείται και αιτιολογείται. Αυτή η έρευνα υπογραμμίζει τη σημασία της επιλογής του κατάλληλου μοντέλου βάσει των χαρακτηριστικών των δεδομένων και των απαιτήσεων του προβλήματος, παρέχοντας πρακτικές γνώσεις για την εφαρμογή τεχνικών μηχανικής μάθησης στην πρόβλεψη των στάσεων αποβίβασης των μέσων δημόσιας συγκοινωνίας.

Abstract

Machine Learning is a rapidly developing field in computer science that can offer solutions to complex and demanding problems. This thesis aims to develop and compare systems for predicting passenger disembarkation stops using public transportation data from the city of Riga, utilizing machine learning models. Additionally, it thoroughly examines algorithms such as Decision Tree, k-Nearest Neighbors, Random Forest, Bagging, Gradient Boosting, Kernel Ridge Regression, Neural Networks, XGBoost, and LightGBM, analyzing their theoretical foundations, implementation methods, and performance metrics. These algorithms are then applied to the collected transportation data, and the accuracy of their predictions is evaluated and justified. This research highlights the importance of selecting the appropriate model based on the characteristics of the data and the problem's requirements, providing practical insights into the application of machine learning techniques for predicting public transportation disembarkation stops.

ΠΕΡΙΕΧΟΜΕΝΑ

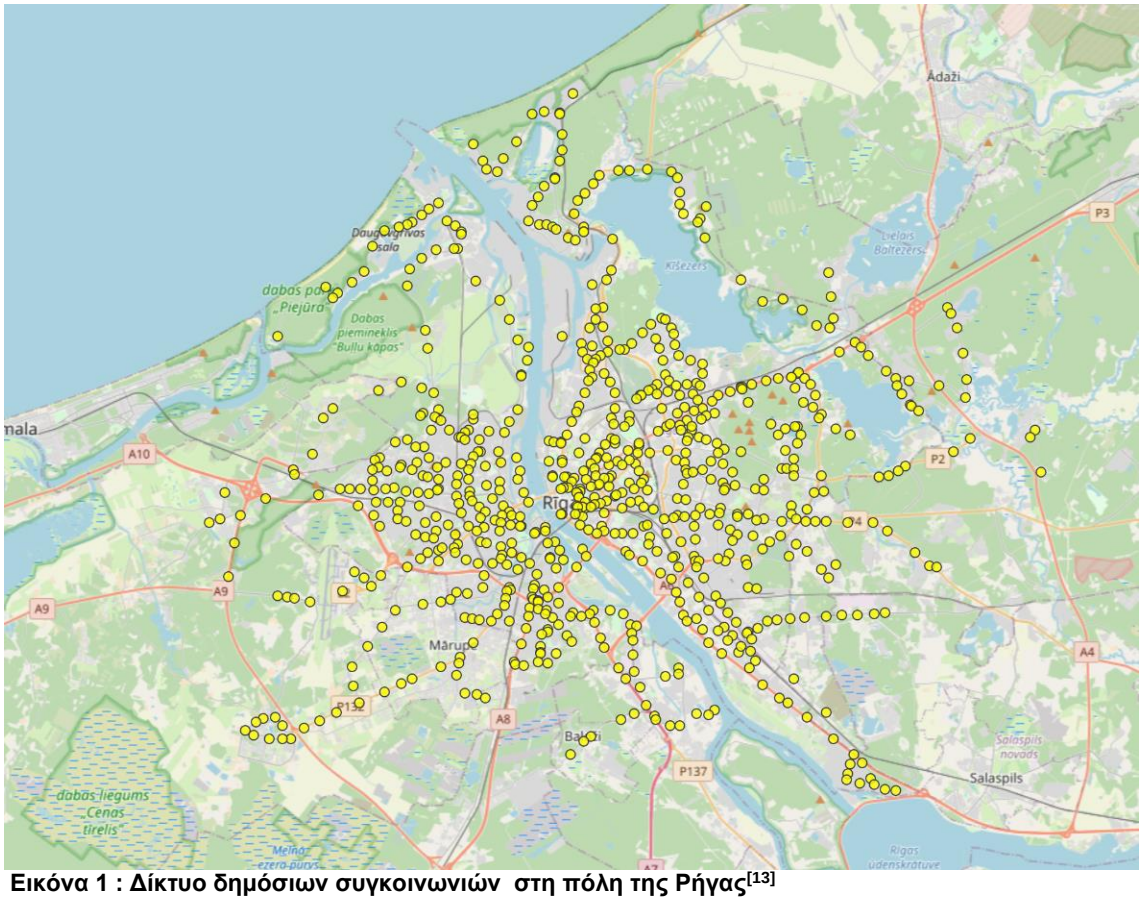
1. ΕΙΣΑΓΩΓΗ

1.1. Ορισμός του Προβλήματος

Στον τομέα των δημόσιων συγκοινωνιών, η ακριβής εκτίμηση των στάσεων επιβίβασης/αποβίβασης των επιβατών παρέχει ένα ισχυρό εργαλείο ώστε να βελτιστοποιούνται τα προσφερόμενα δρομολόγια και η συνολική εμπειρία των επιβατών.

Παρόλο που υπάρχουν πολλές προσπάθειες εκτίμησης, η πρόβλεψη των στάσεων επιβίβασης/αποβίβασης είναι μια πολύπλοκη και δύσκολη πρόκληση στον κλάδο της τεχνητής νοημοσύνης και απαιτεί συνεχή έρευνα προκειμένου να βελτιώνεται και να επεκτείνεται.

Στόχος αυτής της διπλωματικής εργασίας είναι η χρήση δεδομένων δημόσιων μεταφορών που παρέχονται από την πόλη της Ρίγας, προκειμένου να εκτιμηθούν με ακρίβεια οι στάσεις εξόδου συναρτήσει των στάσεων εισόδου για κάθε επικυρωμένο εισιτήριο το οποίο ανήκει σε ένα άτομο.



Εικόνα 1 : Δίκτυο δημόσιων συγκοινωνιών στη πόλη της Ρήγας[13]

Κάποιοι από τους λόγους για τους οποίους έχει αξία το συγκεκριμένο θέμα είναι οι εξής:

- Βελτίωση της Εμπειρίας των Επιβατών :

Η ακριβής πρόβλεψη των στάσεων εισόδου/εξόδου βοηθά στη βελτίωση της συνολικής εμπειρίας των επιβατών, καθώς μπορούν να προγραμματίζουν καλύτερα τα ταξίδια τους.

- Βελτίωση του Σχεδιασμού Μεταφορικών Συστημάτων :

Τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν για τον βελτιστοποιημένο σχεδιασμό των μεταφορικών συστημάτων, προσφέροντας αποτελεσματικότερες υπηρεσίες.

- Οικονομία Χρόνου και Κόστους :

Η ακριβής πρόβλεψη των σταθμών εισόδου/εξόδου μπορεί να συμβάλει στη μείωση του χρόνου και του κόστους των μετακινήσεων.

- Ενίσχυση του Σχεδιασμού Πόλης :

Οι πληροφορίες αυτές συμβάλλουν στην ενίσχυση του σχεδιασμού της πόλης, καθώς οι αρχές μπορούν να κατανοήσουν καλύτερα τις προτιμήσεις και τα πρότυπα μετακίνησης των πολιτών.

- Ασφάλεια και Αποτελεσματικότητα :

Η ακριβής εκτίμηση των σταθμών εισόδου/εξόδου συμβάλλει στη βελτίωση της ασφάλειας και της αποτελεσματικότητας των μεταφορικών υπηρεσιών.

- Εξατομίκευση και Βελτίωση των Υπηρεσιών :

Η κατανόηση των ζευγών αφετηρίας-προορισμού επιτρέπει την εξατομίκευση των υπηρεσιών, προσαρμόζοντας τις στις ανάγκες των πολιτών. Με αυτόν τον τρόπο, η χρήση αυτών των δεδομένων συνεισφέρει στη βελτίωση της κινητικότητας και της ποιότητας ζωής στην πόλη της Ρίγας.

Για την υλοποίηση του στόχου της εργασίας αυτής, θα συγκρίνουμε διάφορα συστήματα μηχανικής μάθησης χρησιμοποιώντας κατάλληλα επεξεργασμένα δεδομένα, με στόχο να αξιολογήσουμε την απόδοσή τους και να φτάσουμε στην δυνατόν καλύτερη πρόβλεψη των στάσεων εξόδου.

1.2. Σχετικές Εργασίες

Μερικά από τα έργα που έχουν ήδη υλοποιηθεί στον τομέα για την μελέτη του προβλήματος είναι τα εξής:

- Breiman, L. (2001). Random Forests. Machine Learning, 45. DOI: https://doi.org/10.1023/A:1010933404324.

  Το άρθρο περιγράφει τους αλγορίθμους Random Forests, οι οποίοι αποτελούν συνδυασμό προβλεπτικών μοντέλων δέντρων, όπου κάθε δέντρο βασίζεται σε έναν τυχαίο διανυσματικό συνδυασμό χαρακτηριστικών, τα οποία επιλέγονται ανεξάρτητα και με την ίδια κατανομή για όλα τα δέντρα στο δάσος. Το σφάλμα γενίκευσης των Random Forests συγκλίνει σε ένα όριο καθώς αυξάνεται ο αριθμός των δέντρων. Το σφάλμα αυτό εξαρτάται από τη δύναμη των επιμέρους δέντρων και τη συσχέτιση μεταξύ τους. Η τυχαία επιλογή χαρακτηριστικών για τη διάσπαση κάθε κόμβου αποδίδει χαμηλά ποσοστά σφάλματος, συγκρίσιμα με τον αλγόριθμο Adaboost, αλλά είναι πιο ανθεκτική στον θόρυβο. Επιπλέον, παρέχονται εσωτερικές εκτιμήσεις για την παρακολούθηση του σφάλματος, της δύναμης και της συσχέτισης, ενώ χρησιμοποιούνται και για τη μέτρηση της σημαντικότητας των χαρακτηριστικών. Αυτές οι ιδέες εφαρμόζονται επίσης και στην παλινδρόμηση.

- Hsu, Y., Chen, Y., & Perng, J. (2020). Estimation of the Number of Passengers in a Bus Using Deep Learning. Sensors, 20(8). DOI: https://doi.org/10.3390/s20082178

  Η παρούσα μελέτη προτείνει μια μέθοδο για την εκτίμηση του αριθμού των επιβατών σε ένα λεωφορείο. Η μέθοδος βασίζεται στο deep learning για την εκτίμηση της πληρότητας των επιβατών σε διάφορα σενάρια. Για να επιτευχθεί αυτό χρησιμοποιούνται δύο μέθοδοι deep learning: η πρώτη είναι ένας συνελικτικός

αυτόματος κωδικοποιητής, που χρησιμοποιείται κυρίως για την εξαγωγή χαρακτηριστικών από πλήθος επιβατών και για τον προσδιορισμό του αριθμού των ατόμων σε ένα πλήθος- η δεύτερη είναι η αρχιτεκτονική you only look once version 3, κυρίως για τον εντοπισμό της περιοχής στην οποία τα χαρακτηριστικά είναι πιο καθαρά σε ένα λεωφορείο. Τα αποτελέσματα που προκύπτουν από τις δύο μεθόδους αθροίζονται για τον υπολογισμό του τρέχοντος ποσοστού πληρότητας του λεωφορείου από επιβάτες. Για να αποδειχθεί η απόδοση του αλγορίθμου, πραγματοποιήθηκαν πειράματα για την εκτίμηση του αριθμού των επιβατών σε διαφορετικές ώρες και στάσεις λεωφορείων. Τα αποτελέσματα δείχνουν ότι το προτεινόμενο σύστημα αποδίδει καλύτερα από ορισμένες υπάρχουσες μεθόδους

.

- Liu, W., Tan, Q., & Liu, L. (2020). Destination Estimation for Bus Passengers Based on Data Fusion. Mathematical Problems in Engineering, 2020. DOI: https://doi.org/10.1155/2020/8305475.

  Η μελέτη αυτή αναλύει την εφαρμογή προηγμένων μεθόδων ανάλυσης δεδομένων για την εκτίμηση του προορισμού των επιβατών λεωφορείων. Οι συγγραφείς προτείνουν ένα αλγόριθμο εκτίμησης που αξιοποιεί ιστορικά δεδομένα επιβατών, περιλαμβάνοντας πληροφορίες όπως χρονικά σήματα εισόδου και εξόδου, σημεία στάσεων, και ιστορικά μοτίβα μετακίνησης. Το μοντέλο τους ενσωματώνει τεχνικές μηχανικής μάθησης, όπως αλγορίθμους ταξινόμησης και αλγορίθμους συστάσεων, για να προσφέρει δυναμικές εκτιμήσεις των τελικών προορισμών των επιβατών. Μέσω της χρήσης αυτών των δεδομένων, το μοντέλο βελτιώνει την ακρίβεια των προβλέψεων και παρέχει πολύτιμα δεδομένα για τη βελτίωση της σχεδίασης δρομολογίων και της διαχείρισης των δημόσιων συγκοινωνιών.

- Lu, X., & Wang, S. (2018). An Improved Taipei Bus Estimation-Time-of-Arrival (ETA) Model Based on Integrated Analysis on Historical and Real-time Bus Position. Proceedings of the 7th International Conference on Cartography and GIS (ICC & GIS). URL: https://iccgis2018.cartographygis.com/7ICCGIS_Proceedings/7_ICCGIS_2018%20(38).pdf.

  Στο άρθρο αυτό προτείνεται ένα βελτιωμένο μοντέλο εκτίμησης χρόνου άφιξης (ETA) για τα λεωφορεία της Ταϊπέι. Η μεθοδολογία βασίζεται στη συνδυαστική ανάλυση ιστορικών και πραγματικών δεδομένων θέσης των λεωφορείων. Εφαρμόζοντας τεχνικές πρόβλεψης μέσω αλγορίθμων μηχανικής μάθησης και βελτιστοποίησης, το μοντέλο εκτιμά με μεγαλύτερη ακρίβεια τον χρόνο άφιξης των λεωφορείων, ενσωματώνοντας παράγοντες όπως η κυκλοφοριακή ροή και οι χρονικές αποκλίσεις. Το άρθρο ακόμη εξετάζει τη χρήση GIS τεχνολογιών και την ακριβή ενσωμάτωση δεδομένων από διαφορετικές χρονικές περιόδους, με στόχο τη βελτίωση των αστικών συγκοινωνιών και της εμπειρίας των επιβατών.

- Tran, L., Mun, M., Lim, M., Yamato, J., Huh, N., & Shahabi, C. (2020). DeepTRANS: A Deep Learning System for Public Bus Travel Time Estimation using Traffic Forecasting. Proceedings of the VLDB Endowment, 13(12). DOI: https://doi.org/10.14778/3415478.3415518.

  Στο άρθρο αυτό παρουσιάζεται ένα σύστημα εκτίμησης χρόνου ταξιδιού λεωφορείων που βασίζεται σε deep learning. Το σύστημα DeepTRANS συνδυάζει μοντέλα

πρόβλεψης της κυκλοφορίας με δεδομένα θέσης λεωφορείων και χρησιμοποιεί νευρωνικά δίκτυα για την ακριβή εκτίμηση του χρόνου ταξιδιού σε πραγματικό χρόνο. Το μοντέλο ενσωματώνει ιστορικά δεδομένα κυκλοφορίας, μοτίβα καθυστερήσεων και άλλες χρονικές μεταβλητές, προσφέροντας βελτιωμένες προβλέψεις σε συνθήκες μεταβαλλόμενης κυκλοφορίας. Η μελέτη εξετάζει τη δομή του νευρωνικού δικτύου, τη διαδικασία εκπαίδευσης και την αξιολόγηση της απόδοσης του συστήματος, συμβάλλοντας στη βελτιστοποίηση της δημόσιας συγκοινωνίας μέσω καλύτερης διαχείρισης του χρόνου και της ροής των λεωφορείων.

2. ΔΕΔΟΜΕΝΑ

2.1. Περιγραφή Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν για την εκτίμηση των στάσεων εξόδων είναι δύο χρονικών περιόδων και συγκεκριμένα αυτά των επικυρωμένων εισιτηρίων για τις ημερομηνίες 11/09/2021 και 07/09/2021 στην πόλη της Ρήγας.

Τα δεδομένα αυτά δίνονται σε δύο αρχεία, μεγέθους 3.730.176 και 2.070.288 εγγραφών αντίστοιχα για τις δύο αυτές ημερομηνίας, και αποτελούνται το καθένα από τα εξής στοιχεία:

Ονομασία Στήλης	Επεξήγηση
GarNr	Αριθμός πινακίδας του εμπλεκόμενου οχήματος δημόσιας συγκοινωνίας.
direction	Κατεύθυνση της εκτελεσθείσας διαδρομής.
ValidTalonaId	Αριθμός ταυτοποίησης e-Validation
datetime	Χρονική σήμανση της επικύρωσης.
route	Όνομα ή αριθμός της ολοκληρωμένης διαδρομής
TripCompanyCode	Κωδικός της εταιρείας που εκτελεί την διαδρομή.
stop_name	Όνομα της στάσης επιβίβασης
stop_id	ID της στάσης επιβίβασης
geometry	Γεωγραφικές συντεταγμένες (γεωγραφικό πλάτος και μήκος) της στάσης.
exit_stop_id	ID της στάσης αποβίβασης.
exit_stop_name	Όνομα της στάσης αποβίβασης.
exit_geometry	Γεωγραφικές συντεταγμένες της στάσης αποβίβασης.

Πίνακας 1 : Πίνακας Περιεχομένων Δεδομένων [12]

2.2. Επεξεργασία Δεδομένων

Για να επιτευχθεί η πιο ακριβής πρόβλεψη της στάσης εξόδου και να εξασφαλιστεί η ποιότητα και η αξιοπιστία των παραγόμενων μοντέλων απαιτείται η κατάλληλη επεξεργασία των δεδομένων η οποία έγινε ως εξής:

- Αφαίρεση των εγγραφών οι οποίες δεν περιείχαν στοιχεία στην στήλη του stop_id ή/και του exit_stop_id για να διασφαλιστεί η μείωση του θορύβου και η βελτίωση ακρίβειας των μοντέλων, με την χρήση της βιβλιοθήκης pandas και συγκεκριμένα της συνάρτησης dropna().

- Αφαίρεση των στηλών που περιείχαν δεδομένα τα οποία δεν χρειάζονται για την διαδικασία εκπαίδευσης των μοντέλων για την αποφυγή του πιθανού overtrain και την αύξηση της αποδοτικότητας των μοντέλων.
- Χρήση κατηγοριών (labels): [Late, Early, Mid-Day, Late] για την ομαδοποίηση των ωρών σε δαφορετικές χρονικές περιόδους της ημέρας. Με αυτόν τον τρόπο μετατρέπονται οι ώρες σε κατηγοριοποιημένες μεταβλητές, οι οποίες μπορούν πιο εύκολα να χρησιμοποιηθούν στα μοντέλα μηχανικής μάθησης.

  Αυτή η κατηγοριοποίηση των ωρών βοηθά στην απλοποίηση της μεταβλητής datetime από μια συνεχή μεταβλητή σε μια διακριτή κατηγορική μεταβλητή, που μπορεί να χρησιμοποιηθεί πιο αποτελεσματικά. Με αυτόν τον τρόπο, η πληροφορία της ώρας της ημέρας διατηρείται χωρίς λάθη. Σε κατηγοριοποιημένες μεταβλητές, τα μοντέλα μπορούν να εκπαιδευτούν καλύτερα στις σχέσεις μεταξύ των χαρακτηριστικών και του στόχου (target variable) και να αποδώσουν καλύτερα στις προβλέψεις τους.
  Αρχικά, η στήλη datetime που περιέχει τις ημερομηνίες και ώρες, μετατρέπεται σε χρονικές περιόδους (bins) με τα εξής διαστήματα:
    - 00:00:00 έως 03:00:00
    - 03:00:00 έως 11:00:00
    - 11:00:00 έως 17:00:00
    - 17:00:00 έως 23:59:59

  Αυτές οι χρονικές περίοδοι αντιστοιχίζονται στις κατηγορίες:
    - Late (αργά το βράδυ): 00:00:00 έως 03:00:00
    - Early (πρωί): 03:00:00 έως 11:00:00
    - Mid-Day (μεσημέρι): 11:00:00 έως 17:00:00
    - Late (απόγευμα και βράδυ): 17:00:00 έως 23:59:59
- Αφαίρεση των πεντακοσίων σπανιότερων τιμών των στάσεων αποβίβασης για την αποφυγή προκατάληψης στα μοντέλα, και συνεπώς την αποφυγή παραμορφώσεων των προβλέψεων και την κατάληξη σε ανακριβή αποτελέσματα.
- Μετατροπή των δεδομένων σε ακέραιους αριθμούς για την διασφάλιση της συμβατότητας με τους αλγόριθμους που χρησιμοποιήθηκαν, καθώς οι περισσότεροι αλγόριθμοι μηχανικής μάθησης απαιτούν αριθμητική είσοδο.
- Διαχωρισμός του συνόλου δεδομένων σε υποσύνολα για την εκπαίδευση των μοντέλων.

  Τα x_train, x_test, y_train και y_test είναι οι μεταβλητές που προκύπτουν από τη διαδικασία διαχωρισμού του συνόλου δεδομένων σε υποσύνολα για την εκπαίδευση και την αξιολόγηση ενός μοντέλου μηχανικής μάθησης.

  Εξήγηση των Μεταβλητών:
    - x_train: Το σύνολο δεδομένων εκπαίδευσης (training set) που περιέχει τα χαρακτηριστικά (features) για την εκπαίδευση του μοντέλου.
    - y_train: Το σύνολο δεδομένων εκπαίδευσης που περιέχει τις ετικέτες ή τις τιμές στόχου (target values) που αντιστοιχούν στα χαρακτηριστικά του x_train.
    - x_test: Το σύνολο δεδομένων δοκιμής (test set) που περιέχει τα χαρακτηριστικά που θα χρησιμοποιηθούν για την αξιολόγηση του μοντέλου.
    - y_test: Το σύνολο δεδομένων δοκιμής που περιέχει τις ετικέτες ή τις τιμές στόχου που αντιστοιχούν στα χαρακτηριστικά του x_test.

  Ο διαχωρισμός του συνόλου δεδομένων σε υποσύνολα εκπαίδευσης και δοκιμής είναι μια κοινή πρακτική στη μηχανική μάθηση για τους εξής λόγους:

- Εκπαίδευση του Μοντέλου:

  Το x_train και το y_train χρησιμοποιούνται για την εκπαίδευση του μοντέλου. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο μαθαίνει τις σχέσεις μεταξύ των χαρακτηριστικών (x_train) και των τιμών στόχου (y_train).

- Αξιολόγηση του Μοντέλου:

  Το x_test και το y_test χρησιμοποιούνται για την αξιολόγηση του μοντέλου. Μετά την εκπαίδευση, το μοντέλο προβλέπει τις τιμές στόχου χρησιμοποιώντας τα δεδομένα δοκιμής (x_test). Οι προβλέψεις αυτές συγκρίνονται με τις πραγματικές τιμές (y_test) για να υπολογιστούν οι μετρικές απόδοσης όπως η ακρίβεια (accuracy), το μέσο απόλυτο σφάλμα (mean absolute error), το μέσο τετραγωνικό σφάλμα (mean squared error), κ.λπ.

Διαδικασία διαχωρισμού :

Η διαδικασία διαχωρισμού πραγματοποιείται με τη χρήση της συνάρτησης

train_test_split από τη βιβλιοθήκη sklearn:

Παράμετροι της train_test_split

- data: Το σύνολο δεδομένων που περιέχει τα χαρακτηριστικά.
- y: Το σύνολο δεδομένων που περιέχει τις τιμές στόχου.
- test_size=0.20: Ορίζει το ποσοστό των δεδομένων που θα χρησιμοποιηθεί για το σύνολο δοκιμής. Στην προκειμένη περίπτωση, το 20% των δεδομένων θα χρησιμοποιηθεί για το σύνολο δοκιμής και το 80% για το σύνολο εκπαίδευσης.
- random_state=0: Ορίζει τον αρχικό σπόρο (seed) για την τυχαιοποίηση, έτσι ώστε η διαδικασία να είναι αναπαραγώγιμη. Με την ίδια τιμή random_state, θα προκύψουν τα ίδια υποσύνολα κάθε φορά που εκτελείται ο κώδικας. Με αυτόν τον τρόπο, μπορούμε να αξιολογήσουμε την απόδοση του μοντέλου μας με δεδομένα που δεν έχουν χρησιμοποιηθεί στην εκπαίδευση, διασφαλίζοντας ότι το μοντέλο εκπαιδεύεται καλά σε νέα δεδομένα.

Με τη διαδικασία αυτή, δίνοντας ως input τα καθαρισμένα πλέον δεδομένα, εξασφαλίζουμε την απόδοση της καλύτερης δυνατής εκτίμησης των αλγορίθμων μηχανικής μάθησης που θα χρησιμοποιήσουμε στη συνέχεια.

3. ΑΛΓΟΡΙΘΜΟΙ

Οι αλγόριθμοι μηχανικής μάθησης επιτρέπουν τη δημιουργία μοντέλων που μπορούν να μάθουν από τα δεδομένα που τους δίνονται, να αναγνωρίζουν πρότυπα και να κάνουν προβλέψεις με ακρίβεια. Στο κεφάλαιο αυτό, αναλύονται διεξοδικά μια σειρά από δημοφιλείς αλγόριθμους μηχανικής μάθησης, οι οποίοι χρησιμοποιήθηκαν και για το πρόβλημα που εξετάζουμε, εστιάζοντας στη θεωρητική τους βάση, τον ψευδοκώδικα υλοποίησής τους, τις περιπτώσεις εφαρμογής τους, καθώς και παραδείγματα χρήσης τους. Οι αλγόριθμοι που θα εξετάσουμε περιλαμβάνουν τα Δέντρα Απόφασης, τους k-Κοντινότερους Γείτονες (k-NN), τα Τυχαία Δάση (Random Forest), την Τεχνική Bagging, την Ενίσχυση Κλίσης (Gradient Boosting), την Πυρηνική Ανάδρομη Παλινδρόμηση (Kernel Ridge Regression), τον XGBoost, τον LightGBM και τα Νευρωνικά Δίκτυα,. Με την κατανόηση αυτών των αλγορίθμων, δημιουργείται μια ολοκληρωμένη εικόνα των μεθόδων που χρησιμοποιούνται για την επίλυση πολύπλοκων προβλημάτων μηχανικής μάθησης και την ανάπτυξη ισχυρών και αξιόπιστων μοντέλων πρόβλεψης, και συνεπώς για τις τεχνικές που χρησιμοποιήθηκαν για την πρόβλεψη των στάσεων εξόδου που είναι το ζητούμενο αυτής της εργασίας.

3.1. Decision Tree (Δέντρο Απόφασης)

Εισαγωγή

Το Δέντρο Απόφασης (Decision Tree) είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Η κύρια ιδέα του αλγορίθμου είναι η δημιουργία ενός μοντέλου που αποτελείται από διαδοχικές αποφάσεις ή "κόμβους" που διαχωρίζουν τα δεδομένα με βάση τις τιμές των χαρακτηριστικών τους. Κάθε απόφαση σε κάθε κόμβο του δέντρου οδηγεί σε μία κατηγορία ή μία πρόβλεψη για την τιμή ενός στόχου, ανάλογα με το είδος του προβλήματος [21].

Θεωρητική Βάση

Το Δέντρο Απόφασης αποτελεί έναν από τους πιο διαδεδομένους αλγορίθμους στον τομέα της μηχανικής μάθησης λόγω της απλότητας της υλοποίησής του και της ευκολίας στην ερμηνεία των αποτελεσμάτων του. Το κύριο χαρακτηριστικό των Δέντρων Απόφασης είναι η ικανότητά τους να διαχωρίζουν τα δεδομένα σε ιεραρχικά επίπεδα, με κάθε κόμβο να αποτελεί ένα "τεστ" σε ένα χαρακτηριστικό, και κάθε κλάδος να αντιπροσωπεύει μία πιθανή εναλλακτική διαίρεση.

Η κατασκευή ενός Δέντρου Απόφασης γίνεται με βάση την επιλογή κατάλληλων χαρακτηριστικών για τη διαχωριστική διαδικασία και την εκτίμηση της καθαρότητας των κόμβων. Οι διαφορετικοί αλγόριθμοι καθαρότητας, όπως ο Gini impurity και η εντροπία (entropy), χρησιμοποιούνται για να αξιολογηθεί η επίδοση της κάθε διαχωριστικής στρατηγικής [4].

Ο αλγόριθμος ξεκινά με έναν κόμβο ρίζας και συνεχίζει την ανάπτυξη του δέντρου με τη διαίρεση των δεδομένων σε υποομάδες, καθορίζοντας δυναμικά τις συνθήκες τερματισμού όπως το μέγιστο βάθος.

Στάδια Εφαρμογής:

- Συγκέντρωση των δεδομένων που θα χρησιμοποιηθούν για την εκπαίδευση και την αξιολόγηση του μοντέλου.
- Διαχείριση των ελλιπών τιμών, των outliers και των ανωμαλιών στα δεδομένα.
- Διαχωρισμός του συνόλου δεδομένων σε εκπαιδευτικό σύνολο (training set) και σύνολο ελέγχου (test set).
- Καθορισμός των παραμέτρων του αλγορίθμου, όπως το μέγιστο βάθος του δέντρου, το ελάχιστο μέγεθος φύλλου, κλπ.
- Δοκιμή και ρύθμιση των υπερπαραμέτρων για βελτίωση της απόδοσης του μοντέλου.
- Επαναληπτική διαδικασία εκπαίδευσης και αξιολόγησης μέχρι να επιτευχθεί το επιθυμητό επίπεδο απόδοσης.

Οφέλη και Περιορισμοί

Οφέλη του αλγορίθμου Decision Tree :

- Το Δέντρο Απόφασης προσφέρει εύκολα ερμηνεύσιμες αποφάσεις και αποτελέσματα.
- Λόγω της διαφάνειάς τους, τα δέντρα απόφασης μπορούν να χρησιμοποιηθούν για την εξερεύνηση και την κατανόηση των σχέσεων στα δεδομένα. Οι πιο σημαντικές μεταβλητές και τα κρίσιμα σημεία διαχωρισμού μπορούν να αναγνωριστούν εύκολα [17].
- Τα δέντρα απόφασης μπορούν να χρησιμοποιηθούν για προβλήματα με αριθμητικά και κατηγορικά χαρακτηριστικά χωρίς την ανάγκη για κωδικοποίηση ή κανονικοποίηση των δεδομένων.

- Σε αντίθεση με άλλους αλγόριθμους, τα δέντρα απόφασης δεν απαιτούν πολλή προεπεξεργασία των δεδομένων, όπως κανονικοποίηση ή διαχείριση των μηδενικών τιμών.

Περιορισμοί του αλγορίθμου Decision Tree :

- Τα δέντρα απόφασης έχουν την τάση να υπερπροσαρμόζονται στα δεδομένα εκπαίδευσης, ειδικά αν το δέντρο είναι πολύ βαθύ. Αυτό σημαίνει ότι το μοντέλο μπορεί να έχει καλή απόδοση στα δεδομένα εκπαίδευσης αλλά να αποτυγχάνει στα δεδομένα ελέγχου ή σε νέα δεδομένα.
- Μικρές αλλαγές στα δεδομένα εκπαίδευσης μπορούν να οδηγήσουν σε μεγάλες αλλαγές στη δομή του δέντρου, προκαλώντας αστάθεια στα αποτελέσματα.
- Τα δέντρα απόφασης δεν είναι τόσο αποτελεσματικά για προβλήματα παλινδρόμησης όσο άλλοι αλγόριθμοι, όπως οι γραμμικά μοντέλα ή τα νευρωνικά δίκτυα.
- Η εκπαίδευση μεγάλων δέντρων απόφασης μπορεί να απαιτεί σημαντικό υπολογιστικό χρόνο και πόρους, ειδικά αν το σύνολο δεδομένων είναι μεγάλο.
- Τα δέντρα απόφασης μπορεί να δυσκολεύονται να εξάγουν σχέσεις που είναι πολυσύνθετες και μη γραμμικές.

3.2. K-Nearest Neighbors (k-NN)

Εισαγωγή

Ο αλγόριθμος k-Nearest Neighbors (k-NN) είναι ένας από τους απλούστερους και πιο ευέλικτους αλγορίθμους μηχανικής μάθησης και χρησιμοποιείται ευρέως τόσο για προβλήματα ταξινόμησης όσο και παλινδρόμησης [16].

Θεωρητική Βάση

Η βασική του k-NN αρχή είναι η ακόλουθη: όμοια παραδείγματα τείνουν να ανήκουν στην ίδια κλάση. Αυτό σημαίνει ότι ο k-NN κατασκευάζει το μοντέλο του χρησιμοποιώντας τα δεδομένα εκπαίδευσης και, κατά τη διάρκεια της πρόβλεψης, αποφασίζει την κλάση του νέου παραδείγματος βασιζόμενο στις κλάσεις των k πλησιέστερων γειτόνων του [6].

Στάδια Εφαρμογής :

- Αποθήκευση Δεδομένων Εκπαίδευσης : Ο αλγόριθμος απλά αποθηκεύει όλα τα δεδομένα εκπαίδευσης.
- Υπολογισμός Αποστάσεων : Κατά τη φάση της πρόβλεψης, για κάθε νέο παράδειγμα πρέπει να υπολογίστεί η απόστασή του από όλα τα παραδείγματα εκπαίδευσης. Οι αποστάσεις συνήθως υπολογίζονται με μετρικές όπως η ευκλείδεια απόσταση ή η απόσταση Manhattan [23].
- Επιλογή Κοντινότερων Γειτόνων : Επιλέγονται τα k πλησιέστερα παραδείγματα (γείτονες) με βάση τις μικρότερες αποστάσεις.
- Απόφαση Κλάσης ή Τιμής : Η τελική απόφαση για το νέο παράδειγμα γίνεται με βάση την πλειοψηφία των κλάσεων των κοντινότερων γειτόνων (στην περίπτωση ταξινόμησης) ή τον μέσο όρο των τιμών τους (στην περίπτωση παλινδρόμησης).

Οφέλη και Περιορισμοί

Οφέλη του αλγορίθμου k-NN :

- Απλός στην υλοποίηση και εύκολος στην κατανόηση, καθώς βασίζεται στην έννοια της απόστασης μεταξύ δειγμάτων.

- Δεν απαιτεί υπόθεση για την κατανομή των δεδομένων: Ο k-NN δεν κάνει υποθέσεις σχετικά με τη μορφή της κατανομής των δεδομένων και είναι εάν τα δεδομένα είναι γραμμικά διαχωρίσιμα ή όχι.
- Κατάλληλος για πολυκλασική ταξινόμηση, καθώς μπορεί να χειριστεί πολυκλασικές κατηγορίες χωρίς πρόβλημα.
- Σχετικά καλή απόδοση για μικρά σετ δεδομένων, ειδικά όταν ο αριθμός των γειτόνων (k) επιλέγεται κατάλληλα.

Περιορισμοί του αλγορίθμου k-NN :

- Αναγκαία επιλογή του k: Η σωστή επιλογή του αριθμού k είναι κρίσιμη. Ένα μικρό k μπορεί να οδηγήσει σε πολύ επηρεασμένη από θόρυβο ταξινόμηση, ενώ ένα μεγάλο k μπορεί να κάνει τον αλγόριθμο να χάσει τις λεπτομέρειες της δομής των δεδομένων [8].
- Ευαισθησία στις αρχικές, καθώς μικρές αλλαγές στα δεδομένα μπορεί να οδηγήσουν σε διαφορετικές ταξινομήσεις.

3.3. Random Forest

Εισαγωγή

Ο Random Forest είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Βασίζεται στην ιδέα της συλλογικής μάθησης (ensemble learning). Ο Random Forest συνδυάζει πολλαπλά δέντρα απόφασης για να βελτιώσει την ακρίβεια και να μειώσει την πιθανότητα υπερπροσαρμογής (overfitting).

Θεωρητική Βάση

Ο Random Forest αποτελείται από ένα σύνολο δέντρων απόφασης που λειτουργούν ως επιμέρους μοντέλα σε ένα σύνολο δεδομένων. Η βασική ιδέα είναι η χρήση της τυχαίας επιλογής υποσυνόλων δεδομένων και χαρακτηριστικών για την κατασκευή κάθε δέντρου [15].

Αυτή η διαδικασία αυξάνει την ποικιλία μεταξύ των δέντρων και ενισχύει τη συνολική ακρίβεια του μοντέλου [3].

Στάδια Εφαρμογής

- Δημιουργία πολλαπλών δέντρων απόφασης, χρησιμοποιώντας τυχαία υποσύνολα δεδομένων (bagging) και τυχαία υποσύνολα χαρακτηριστικών σε κάθε διακλάδωση του δέντρου.
- Συνδυασμός των αποτελεσμάτων των δέντρων, χρησιμοποιώντας πλειοψηφική ψήφο (majority vote) για ταξινόμηση ή μέσο όρο (average) για παλινδρόμηση.

Οφέλη και περιορισμοί

Οφέλη του αλγορίθμου Random Forest :

- Υψηλή ακρίβεια: Ο Random Forest συνδυάζει πολλαπλά δέντρα απόφασης για να βελτιώσει την ακρίβεια της πρόβλεψης.
- Αντοχή στην υπερπροσαρμογή (overfitting) λόγω του συνδυασμού πολλαπλών δέντρων.
- Δυνατότητα χειρισμού μεγάλων συνόλων δεδομένων
- Εκτίμηση της σημασίας των χαρακτηριστικών για την πρόβλεψη.

Περιορισμοί του αλγορίθμου Random Forest:

- Έλλειψη ερμηνευσιμότητας: Επειδή αποτελείται από πολλά δέντρα, μπορεί να είναι δύσκολο να ερμηνευθεί η τελική πρόβλεψη.
- Απαίτηση παραμέτρων: Ο αριθμός των δέντρων και ο αριθμός των χαρακτηριστικών που επιλέγονται τυχαία πρέπει να ρυθμιστούν κατάλληλα.

3.4. Bagging

Εισαγωγή

Ο αλγόριθμος Bagging (Bootstrap Aggregating) είναι μια τεχνική συλλογικής μάθησης που χρησιμοποιείται για τη βελτίωση της σταθερότητας και της ακρίβειας των μηχανών μάθησης.

Λειτουργεί με τη δημιουργία πολλαπλών εκδοχών ενός αλγορίθμου μάθησης, οι οποίες στη συνέχεια συνδυάζονται για να παράγουν μία τελική πρόβλεψη.

Ο Bagging είναι ιδιαίτερα αποτελεσματικός στη μείωση της διακύμανσης (variance) και της υπερπροσαρμογής (overfitting).

Θεωρητική Βάση

Ο Bagging χρησιμοποιεί την τεχνική της δειγματοληψίας με επανατοποθέτηση (bootstrap sampling) για να δημιουργήσει πολλά υποσύνολα των δεδομένων εκπαίδευσης. Για κάθε υποσύνολο, εκπαιδεύεται ένα μοντέλο (συνήθως ένα δέντρο απόφασης). Οι προβλέψεις των μοντέλων συνδυάζονται στη συνέχεια μέσω μιας διαδικασίας όπως η πλειοψηφική ψήφος (για ταξινόμηση) ή ο μέσος όρος (για παλινδρόμηση) [2].

Στάδια Εφαρμογής

- Δημιουργία πολλαπλών υποσυνόλων δεδομένων, χρησιμοποιώντας δειγματοληψία με επανατοποθέτηση από το αρχικό σύνολο δεδομένων.
- Εκπαίδευση μοντέλων σε κάθε υποσύνολο, χρησιμοποιώντας τον ίδιο αλγόριθμο μάθησης για κάθε υποσύνολο.
- Συνδυασμός των προβλέψεων των μοντέλων, χρησιμοποιώντας πλειοψηφική ψήφο ή μέσο όρο.

Οφέλη και Περιορισμοί

Οφέλη του αλγορίθμου Bagging:

- Μείωση της διακύμανσης (variance), συνδυάζοντας πολλά μοντέλα, γεγονός που βελτιώνει τη σταθερότητα και την ακρίβεια [5].
- Αντοχή στην υπερπροσαρμογή (overfitting), λόγω της χρήσης πολλαπλών μοντέλων.
- Ευελιξία: Ο Bagging μπορεί να χρησιμοποιηθεί με διάφορους αλγόριθμους μάθησης, αν και συνήθως εφαρμόζεται σε δέντρα απόφασης.
- Βελτίωση απόδοσης, ιδιαίτερα σε προβλήματα με υψηλή διακύμανση.

Περιορισμοί του αλγορίθμου Bagging:

- Απαίτηση παραμέτρων: Η επιλογή του αριθμού των μοντέλων και του μεγέθους των υποσυνόλων δεδομένων μπορεί να απαιτεί πειραματισμό.
- Έλλειψη ερμηνευσιμότητας: Όπως και με άλλες τεχνικές συλλογικής μάθησης, ο Bagging μπορεί να παράγει σύνθετα μοντέλα που είναι δύσκολο να ερμηνευτούν.

3.5. Gradient Boosting

Εισαγωγή

Ο αλγόριθμος Gradient Boosting είναι μια ισχυρή τεχνική συλλογικής μάθησης που χρησιμοποιείται για την αύξηση της απόδοσης των αλγορίθμων μάθησης. Βασίζεται στην ιδέα της σταδιακής ενίσχυσης αδύναμων μοντέλων (weak learners) για τη δημιουργία ενός ισχυρού μοντέλου (strong learner) [24]. Ο Gradient Boosting εφαρμόζεται ευρέως σε προβλήματα ταξινόμησης και παλινδρόμησης και έχει αποδειχθεί ιδιαίτερα αποτελεσματικός σε διάφορους τομείς, όπως η χρηματοοικονομική ανάλυση και η βιοπληροφορική.

Θεωρητική Βάση

Ο Gradient Boosting λειτουργεί με την εκπαίδευση μοντέλων σε διαδοχικές φάσεις, κάθε μία από τις οποίες προσπαθεί να διορθώσει τα λάθη του προηγούμενου μοντέλου. Αυτό επιτυγχάνεται μέσω της εκπαίδευσης ενός νέου μοντέλου για την πρόβλεψη των υπολειμμάτων (residuals) των προηγούμενων μοντέλων. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να επιτευχθεί ένας προκαθορισμένος αριθμός μοντέλων ή να βελτιωθεί η απόδοση κατά ένα συγκεκριμένο όριο [7].

Στάδια Εφαρμογής :

- Αρχικοποίηση του μοντέλου: Ένα αρχικό μοντέλο εκπαίδευσης, συνήθως είναι ένα απλό μοντέλο, όπως ο μέσος όρος των στόχων.
- Υπολογισμός των υπολειμμάτων: Τα υπολείμματα υπολογίζονται ως η διαφορά μεταξύ των πραγματικών τιμών και των προβλέψεων του τρέχοντος μοντέλου.
- Εκπαίδευση ενός νέου μοντέλου στα υπολείμματα: Ένα νέο μοντέλο εκπαιδεύεται για να προβλέψει τα υπολείμματα.
- Ενημέρωση του μοντέλου: Το νέο μοντέλο προστίθεται στο σύνολο των μοντέλων με έναν σταθερό συντελεστή μάθησης (learning rate).
- Επανάληψη: Η διαδικασία επαναλαμβάνεται για έναν προκαθορισμένο αριθμό επαναλήψεων ή μέχρι να επιτευχθεί η βέλτιστη απόδοση.

Οφέλη και περιορισμοί

Οφέλη του Gradient Boosting :

- Υψηλή απόδοση : Ο Gradient Boosting είναι γνωστός για την υψηλή του απόδοση σε προβλήματα τόσο ταξινόμησης όσο και παλινδρόμησης.
- Μείωση της μεροληψίας και της διακύμανσης : Συνδυάζοντας πολλά μοντέλα, ο Gradient Boosting μπορεί να μειώσει τόσο τη μεροληψία όσο και τη διακύμανση των προβλέψεων [11].
- Ευελιξία : Μπορεί να χρησιμοποιηθεί με διάφορους αλγόριθμους μάθησης και να προσαρμοστεί για διαφορετικά προβλήματα.
- Δυνατότητα χρήσης για ανάλυση σημαντικότητας χαρακτηριστικών : Ο Gradient Boosting μπορεί να χρησιμοποιηθεί για την εκτίμηση της σημαντικότητας των χαρακτηριστικών.

Περιορισμοί του αλγόριθμου Gradient Boosting:

- Απαίτηση παραμέτρων: Η ρύθμιση των υπερπαραμέτρων, όπως ο αριθμός των μοντέλων και ο συντελεστής μάθησης, μπορεί να απαιτεί πειραματισμό.
- Ευαισθησία σε θόρυβο: Ο Gradient Boosting μπορεί να υπερπροσαρμοστεί σε δεδομένα με πολύ θόρυβο αν δεν ρυθμιστούν σωστά οι υπερπαράμετροι.

3.6. Kernel Ridge Regression (KRR)

Εισαγωγή

Ο αλγόριθμος Kernel Ridge Regression (KRR) είναι μια επέκταση της κλασικής γραμμικής παλινδρόμησης που συνδυάζει τον Ridge Regression με τη χρήση πυρηνικών συναρτήσεων (kernels). Η χρήση των Kernels επιτρέπει την εφαρμογή της μεθόδου σε μη γραμμικά προβλήματα, προσφέροντας μεγάλη ευελιξία και αποτελεσματικότητα [22]. Ο KRR εφαρμόζεται ευρέως σε διάφορους τομείς, όπως η ανάλυση χρηματοοικονομικών δεδομένων, η αναγνώριση προτύπων και η βιοπληροφορική.

Θεωρητική Βάση

Ο Kernel Ridge Regression βασίζεται στην έννοια των πυρηνικών συναρτήσεων για να επεκτείνει την κλασική Ridge Regression σε μη γραμμικά προβλήματα. Η βασική ιδέα είναι να μετασχηματίσει τα δεδομένα εισόδου σε έναν υψηλότερης διάστασης χώρο χαρακτηριστικών, όπου είναι πιο πιθανό να είναι γραμμικά διαχωρίσιμα. Αυτό επιτυγχάνεται μέσω της χρήσης πυρηνικών συναρτήσεων, οι οποίες επιτρέπουν τον υπολογισμό του εσωτερικού γινομένου των μετασχηματισμένων δεδομένων χωρίς να απαιτείται η πραγματική μεταφορά τους στον υψηλότερης διάστασης χώρο (kernel trick) [18].

Η βασική εξίσωση για την Ridge Regression είναι η εξής:

$$\boldsymbol{W} = (\boldsymbol{X^T X} + \boldsymbol{\lambda I})^{-1} \boldsymbol{X^T y}$$

όπου X είναι ο πίνακας δεδομένων εισόδου, y είναι το διάνυσμα των στόχων, w είναι το διάνυσμα των βαρών, λ είναι η παράμετρος τακτικής και I είναι ο μοναδιαίος πίνακας.

Για τον Kernel Ridge Regression, χρησιμοποιείται η πυρηνική συνάρτηση K που αντικαθιστά το $X^T X$ με το K, όπου K είναι ο πίνακας πυρήνα που υπολογίζεται ως:

$$\boldsymbol{K_{ij}} = \boldsymbol{k}(\boldsymbol{x_i}, \boldsymbol{x_j})$$

Η τελική εξίσωση για την KRR γίνεται:

$$\boldsymbol{\alpha} = (\boldsymbol{K} + \boldsymbol{\lambda I})^{-1} \boldsymbol{y}$$

όπου α είναι το διάνυσμα των συντελεστών και K είναι ο πίνακας πυρήνα.

Οφέλη και Περιορισμοί

Οφέλη του Kernel Ridge Regression :

- Ευελιξία: Η χρήση πυρηνικών συναρτήσεων επιτρέπει την εφαρμογή του KRR σε μη γραμμικά προβλήματα.
- Καλή γενίκευση: Η τακτική (regularization) βοηθά στη μείωση της υπερπροσαρμογή και βελτιώνει τη γενίκευση του μοντέλου [19].
- Εύκολη εφαρμογή: Ο KRR είναι εύκολος στην εφαρμογή και μπορεί να χρησιμοποιηθεί με διάφορους πυρήνες.

Περιορισμοί του Kernel Ridge Regression

- Απαίτηση επιλογής πυρήνα: Η επιλογή της κατάλληλης πυρηνικής συνάρτησης και των παραμέτρων της μπορεί να είναι δύσκολη και απαιτεί πειραματισμό.

- Κλιμάκωση: Η απόδοση της KRR μπορεί να επηρεαστεί από την κλιμάκωση των δεδομένων εισόδου.

3.7. XGBoost (Extreme Gradient Boosting)

Εισαγωγή

Ο XGBoost (Extreme Gradient Boosting) είναι ένας από τους πιο δημοφιλείς και ισχυρούς αλγορίθμους μηχανικής μάθησης, ειδικά για δομημένα δεδομένα. Αναπτύχθηκε για να παρέχει υψηλή απόδοση, δυνατότητα παραλληλοποίησης και ευελιξία, επιτρέποντας την εφαρμογή του σε ένα ευρύ φάσμα προβλημάτων παλινδρόμησης και ταξινόμησης. Η ικανότητά του να διαχειρίζεται μεγάλες ποσότητες δεδομένων και να αποφεύγει την υπερπροσαρμογή τον καθιστά ένα πολύτιμο εργαλείο [9].

Θεωρητική Βάση

Ο XGBoost βασίζεται στην τεχνική του gradient boosting, η οποία συνδυάζει την απόδοση πολλαπλών αδύναμων αλγορίθμων (συνήθως δέντρων απόφασης) για να δημιουργήσει έναν ισχυρό τελικό αλγόριθμο. Κάθε νέο δέντρο στο σύνολο προστίθεται για να διορθώσει τα λάθη που έκαναν τα προηγούμενα δέντρα [20].

Η βασική ιδέα πίσω από το gradient boosting είναι να βελτιώνεται συνεχώς η απόδοση του μοντέλου προσθέτοντας νέα δέντρα απόφασης που διορθώνουν τα λάθη του προηγούμενου μοντέλου. Η διαδικασία αυτή επαναλαμβάνεται έως ότου επιτευχθεί η βέλτιστη απόδοση ή ικανοποιηθεί κάποιο κριτήριο διακοπής.

Απώλεια και Gradient

Για κάθε επανάληψη, προσθέτουμε ένα νέο δέντρο που εκπαιδεύεται για να προβλέψει το αρνητικό gradient της συνάρτησης απώλειας για τις τρέχουσες προβλέψεις του μοντέλου.

Η συνάρτηση απώλειας L για ένα δεδομένο δείγμα i με πρόβλεψη $\hat{y}_i^{(m-1)}$ είναι:

$$L\left(y_i, \hat{y}_i^{(m-1)}\right)$$

Το gradient της απώλειας είναι:

$$\boldsymbol{g_i^{(m)}} = \frac{\boldsymbol{dL}(y_i, \hat{y})}{\boldsymbol{d}\hat{y}}\Big|_{\hat{y}=\hat{y}_i^{(m-1)}}$$

Το νέο δέντρο εκπαιδεύεται για να προβλέψει το $g_i^{(m)}$

Συναρτήσεις Στόχου και Κανονικοποίηση

Ο XGBoost χρησιμοποιεί μια τεχνική κανονικοποίησης που ονομάζεται L2 κανονικοποίηση των βαρών των φύλλων, η οποία βοηθά στην αποφυγή υπερπροσαρμογής. Επίσης, χρησιμοποιεί παραμέτρους όπως το learning rate για να ελέγξει το μέγεθος των βημάτων βελτίωσης.

Οφέλη και περιορισμοί

Οφέλη του XGBoost :

- Υψηλή απόδοση : Ο XGBoost είναι γνωστός για την εξαιρετική του απόδοση σε διάφορες εφαρμογές μηχανικής μάθησης.
- Αποφυγή της υπερπροσαρμογής, χρησιμοποιώντας κανονικοποίηση και άλλες τεχνικές.

- Ευελιξία : Μπορεί να χρησιμοποιηθεί τόσο για ταξινόμηση όσο και για παλινδρόμηση, για διάφορες συναρτήσεις απώλειας.
- Παραλληλοποίηση : Εκμεταλλεύεται τις σύγχρονες υπολογιστικές αρχιτεκτονικές για γρήγορη εκπαίδευση.

Περιορισμοί του XGBoost :

- Σύνθετη ρύθμιση παραμέτρων: Απαιτείται προσεκτική ρύθμιση των υπερπαραμέτρων για την επίτευξη βέλτιστης απόδοση [20].
- Μνήμη: Μπορεί να απαιτεί μεγάλη ποσότητα μνήμης, ειδικά κατά την εκπαίδευση με μεγάλα σύνολα δεδομένων.

3.8. LightGBM (Light Gradient Machine)

Εισαγωγή

Ο LightGBM (Light Gradient Boosting Machine) είναι ένας ισχυρός και αποδοτικός αλγόριθμος μηχανικής μάθησης που αναπτύχθηκε από την Microsoft. Είναι ειδικά σχεδιασμένος για να παρέχει υψηλή απόδοση και αποτελεσματικότητα στην εκπαίδευση και την πρόβλεψη, κάνοντας χρήση τεχνικών βελτιστοποίησης μνήμης και ταχύτητας. Το LightGBM είναι ιδανικό για μεγάλα σύνολα δεδομένων και χρησιμοποιείται ευρέως σε προβλήματα ταξινόμησης και παλινδρόμησης.

Θεωρητική Βάση

Ο LightGBM βασίζεται στην τεχνική του gradient boosting, η οποία συνδυάζει πολλούς αδύναμους αλγορίθμους (συνήθως δέντρα απόφασης) για να δημιουργήσει έναν ισχυρό τελικό αλγόριθμο. Το LightGBM βελτιώνει την απόδοση και την αποτελεσματικότητα του gradient boosting μέσω διάφορων καινοτομιών, όπως η χρήση τεχνικών histogram-based και leaf-wise [14].

Histogram-based μέθοδος

Αντί να εξετάζει κάθε πιθανό split σημείο, η histogram-based μέθοδος του LightGBM ομαδοποιεί συνεχείς τιμές χαρακτηριστικών σε διακριτούς "κάδους" (bins). Αυτό μειώνει τον υπολογιστικό φόρτο και επιτρέπει γρηγορότερη εκπαίδευση, διατηρώντας παράλληλα την ακρίβεια του μοντέλου.

Leaf-wise στρατηγική

Σε αντίθεση με την παραδοσιακή level-wise στρατηγική, όπου τα δέντρα απόφασης αναπτύσσονται επίπεδο-επίπεδο, η leaf-wise στρατηγική του LightGBM αναπτύσσει τα φύλλα με τη μεγαλύτερη μείωση απώλειας πρώτα. Αυτό επιτρέπει τη δημιουργία βαθύτερων δέντρων και καλύτερης γενίκευσης, αλλά απαιτεί προσοχή για την αποφυγή υπερπροσαρμογής.

Οφέλη του LightGBM

Οφέλη του LightGBM :

- Υψηλή απόδοση: Ο LightGBM είναι γνωστό για την εξαιρετική του απόδοση και την ταχύτητα εκπαίδευσης, ειδικά σε μεγάλα σύνολα δεδομένων.
- Αποτελεσματικότητα: Χρησιμοποιεί histogram-based μεθόδους και leaf-wise στρατηγικές για να βελτιστοποιήσει την απόδοση και τη μνήμη [10].
- Αντιμετώπιση υπερπροσαρμογής: Παρέχει μηχανισμούς κανονικοποίησης και παραμέτρους για τον έλεγχο της υπερπροσαρμογής.
- Ευελιξία: Μπορεί να χρησιμοποιηθεί για ταξινόμηση, παλινδρόμηση, υποστηρίζοντας διάφορες συναρτήσεις απώλειας.

Περιορισμοί του LightGBM :

- Σύνθετη ρύθμιση παραμέτρων**:** Όπως και με το XGBoost, απαιτείται προσεκτική ρύθμιση των υπερπαραμέτρων για την επίτευξη βέλτιστης απόδοσης.
- Μνήμη**:** Μπορεί να απαιτεί μεγάλη ποσότητα μνήμης, ειδικά κατά την εκπαίδευση με μεγάλα σύνολα δεδομένων.
- Ευαισθησία σε μη ισορροπημένα δεδομένα**:** Ο LightGBM μπορεί να είναι ευαίσθητος σε μη ισορροπημένα σύνολα δεδομένων, απαιτώντας πρόσθετα βήματα προεπεξεργασίας.

3.9. Νευρωνικό Δίκτυο

Εισαγωγή

Τα νευρωνικά δίκτυα είναι από τα πιο ισχυρά εργαλεία στη μηχανική μάθηση και την τεχνητή νοημοσύνη. Εμπνευσμένα από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου, τα νευρωνικά δίκτυα αποτελούνται από νευρώνες οργανωμένους σε στρώματα, τα οποία συνεργάζονται για να μάθουν και να γενικεύσουν από τα δεδομένα. Τα νευρωνικά δίκτυα χρησιμοποιούνται ευρέως σε ποικίλες εφαρμογές, όπως η αναγνώριση εικόνας, η επεξεργασία φυσικής γλώσσας και τα αυτόνομα οχήματα.

Θεωρητική Βάση

Δομή Νευρωνικού Δικτύου [1] :

Ένα τυπικό νευρωνικό δίκτυο αποτελείται από τρία βασικά είδη στρωμάτων:

- Εισόδου (Input Layer) : Το στρώμα αυτό λαμβάνει τα αρχικά δεδομένα και τα περνάει στα επόμενα στρώματα.
- Κρυφά (Hidden Layers) : Αυτά τα στρώματα βρίσκονται μεταξύ του εισόδου και του εξόδου και είναι υπεύθυνα για την εξαγωγή και την αναγνώριση χαρακτηριστικών των δεδομένων. Κάθε κρυφό στρώμα αποτελείται από νευρώνες που εφαρμόζουν μη γραμμικές συναρτήσεις ενεργοποίησης.
- Εξόδου (Output Layer) : Το στρώμα αυτό παράγει την τελική πρόβλεψη του δικτύου.
  - *Νευρώνες και Συναρτήσεις Ενεργοποίησης*

Κάθε νευρώνας σε ένα στρώμα υπολογίζει μια γραμμική συνάρτηση των εισόδων του και στη συνέχεια εφαρμόζει μια μη γραμμική συνάρτηση ενεργοποίησης, όπως η sigmoid, η tanh ή η ReLU (Rectified Linear Unit) [8].

  - *Προώθηση (Forward Propagation)*

Η προώθηση περιλαμβάνει τον υπολογισμό των εξόδων των νευρώνων από το στρώμα εισόδου μέχρι το στρώμα εξόδου. Αυτό γίνεται με τον εξής τρόπο:

- Υπολογισμός της γραμμικής συνάρτησης $z = W \cdot x + b$, όπου W είναι τα βάρη, x οι εισροές και b οι προκαταλήψεις[25].
- Εφαρμογή της συνάρτησης ενεργοποίησης a = σ(z).
  - *Αντίστροφη Διάδοση (Backpropagation)*

Η αντίστροφη διάδοση είναι η διαδικασία εκπαίδευσης του νευρωνικού δικτύου. Περιλαμβάνει τον υπολογισμό των παραγώγων της συνάρτησης απώλειας σε σχέση με τα βάρη και τις προκαταλήψεις, και την ενημέρωσή τους χρησιμοποιώντας τον αλγόριθμο του gradient descent[26].

Οφέλη και Περιορισμοί

Οφέλη του Νευρωνικών Δικτύων :

- Ικανότητα Μάθησης από σύνθετα μοτίβα και σχέσεις από μεγάλα σύνολα δεδομένων.
- Ευελιξία**:** Μπορούν να εφαρμοστούν σε μια πληθώρα προβλημάτων, από ταξινόμηση και παλινδρόμηση μέχρι παραγωγή κειμένου και εικόνας.
- Αυτοβελτίωση: Μπορούν να βελτιώνουν την απόδοσή τους με την πάροδο του χρόνου καθώς εκτίθενται σε περισσότερα δεδομένα.
- Αυτοματισμός Χαρακτηριστικών: Μπορούν να εξάγουν αυτόματα σημαντικά χαρακτηριστικά από τα δεδομένα χωρίς την ανάγκη χειροκίνητης επεξεργασίας.

Περιορισμοί των Νευρωνικών Δικτύων :

- Είναι επιρρεπή σε υπερπροσαρμογή, ειδικά όταν το σύνολο εκπαίδευσης είναι μικρό.
- Μαύρο Κουτί: Η ερμηνευσιμότητα των αποτελεσμάτων μπορεί να είναι περιορισμένη, καθώς οι εσωτερικές διεργασίες ενός νευρωνικού δικτύου δεν είναι πάντα διαφανείς.
- Ανάγκη Μεγάλου Όγκου Δεδομένων για την αποτελεσματική εκπαίδευση και την αποφυγή υπερπροσαρμογής.

4. ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ

Στο κεφάλαιο αυτό, αναλύονται τα αποτελέσματα των παραπάνω αλγορίθμων μηχανικής μάθησης που εφαρμόστηκαν σε δύο διαφορετικά σύνολα δεδομένων όπως περιγράφονται στο Κεφάλαιο 2. Η ανάλυση των αποτελεσμάτων αυτών αποσκοπεί στην αξιολόγηση της απόδοσης των αλγορίθμων Decision Tree, k-Nearest Neighbors (k-NN), Random Forest, Bagging, Gradient Boosting, Kernel Ridge Regression, Neural Network, XGBoost και LightGBM, εξετάζοντας τη μεταβολή της ακρίβειας (accuracy) των προβλέψεων τους για το exit_stop_id συναρτήσει του stop_id.

Για την την ανάπτυξη των αλγορίθμων, χρησιμοποιήθηκαν διάφορες βιβλιοθήκες, με κύρια την scikit-learn. Η συγκεκριμένη βιβλιοθήκη αποτελεί βασικό εργαλείο για την εφαρμογή αλγορίθμων μηχανικής μάθησης και παρείχε πολλές από τις λειτουργίες που χρησιμοποιήθηκαν στην πειραματική μελέτη. Για την ταξινόμηση χρησιμοποιήθηκαν μοντέλα όπως το DecisionTreeClassifier, RandomForrestClassifier, GradientBoostingClassifier BaggingClassifier και το KNeighborsClassifier, τα οποία επέτρεψαν την εκπαίδευση των αντίστοιχων αλγορίθμων και πρόβλεψη των αποτελεσμάτων. Επίσης, για την παλινδρόμηση χρησιμοποιήθηκε το Kernel Ridge Regression μέσω της συνάρτησης KernelRidge, ενώ η κανονικοποίηση των δεδομένων έγινε με εργαλεία όπως το StandardScaler, και το LabelEncoder για την κωδικοποίηση κατηγοριών. Για την δημιουργία και την οπτικοποίηση των γραφημάτων χρησιμοποιήθηκε η βιβλιοθήκη MatPlotLib.

Εκτός από τις παραπάνω βασικές βιβλιοθήκες, χρησιμοποιήθηκαν και κάποιες πιο εξειδικευμένες. Συγκεκριμένα, η XGBoost εφαρμόζει τον αλγόριθμο Gradient Boosting, βελτιώνοντας την ακρίβεια μέσω του συνδυασμού πολλών μοντέλων. Αντίστοιχα η LightGBM προσφέρει μία πιο γρήγορη προσέγγιση στην ενίσχυση δέντρων απόφασης με στόχο την βελτιστοποίηση της ταχύτητας. Επιπλέον, η Keras, χρησιμοποιήθηκε για την υλοποίηση και εκπαίδευση του MLP νευρωνικού δυκτίου.

Η ακρίβεια (accuracy) είναι ένα βασικό μέτρο απόδοσης σε προβλήματα ταξινόμησης, το οποίο υπολογίζεται ως ο λόγος των σωστών προβλέψεων προς το συνολικό αριθμό των προβλέψεων. Συγκεκριμένα, εκφράζει το ποσοστό των στάσεων εξόδου που προβλέφθηκαν σωστά από τον αλγόριθμο. Η επιλογή της ακρίβειας ως μέτρο αξιολόγησης βασίζεται στην ικανότητά της να παρέχει μια άμεση και κατανοητή ένδειξη της αποτελεσματικότητας του μοντέλου.

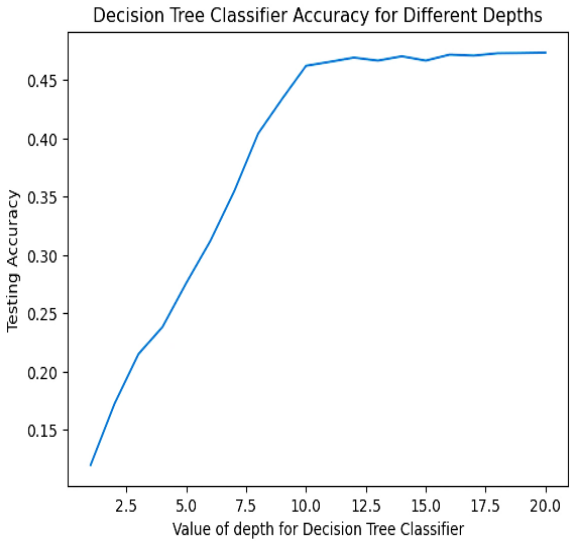
Μέσω αυτής της διαδικασίας, προσδιορίζονται οι αλγόριθμοι που αποδίδουν καλύτερα στα υπάρχοντα δεδομένα του προβλήματος που εξετάζουμε. Επιπλέον, παρατίθενται γραφήματα

που απεικονίζουν την πορεία της ακρίβειας των αλγορίθμων ανάλογα με την ημέρα που εξετάζεται και τα αίτια των διαφοροποιήσεων στις επιδόσεις τους. Τα συμπεράσματα βοηθούν στην κατανόηση της συμπεριφοράς των αλγορίθμων σε πραγματικά δεδομένα και προσφέρουν πολύτιμες πληροφορίες για μελλοντικές εφαρμογές και βελτιστοποιήσεις.

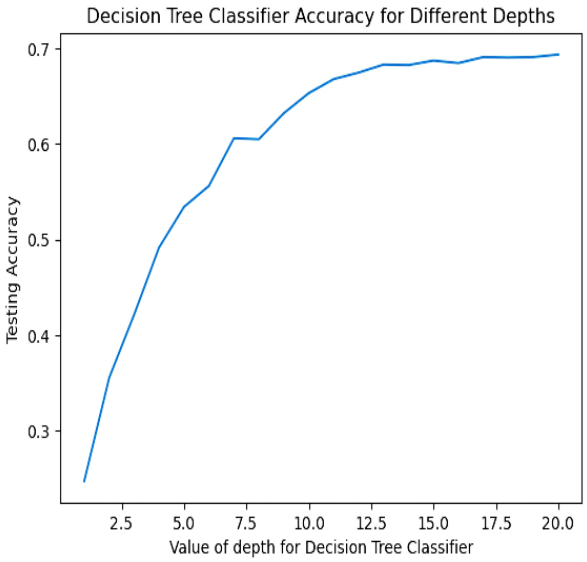
Για λόγους συντομίας, ως «Ημέρα 1» και «Ημέρα 2» περιγράφονται οι 07/09/2021 και 11/09/2021 αντίστοιχα.

4.1. Αποτελέσματα με χρήση Desicion Tree

Παρακάτω παρουσιάζεται η ακρίβεια πρόβλεψης του exit_stop_id,με χρήση του μοντέλου Decision Tree σε σχέση με το βάθος του δέντρου.



Εικ. 2: Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1



Εικ. 3: Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2

Βάθος Δέντρου	Ακρίβεια Ημέρα 1	Ακρίβεια Ημέρα 2
2.5	0.15	0.30
5.0	0.25	0.40
7.5	0.35	0.50
10.0	0.45	0.55
12.5	0.42	0.60
15.0	0.43	0.65
17.5	0.44	0.67
20.0	0.45	0.70

Πίνακας 2 : Πίνακας Τιμών Decision Tree

Είσοδος Αλγορίθμου

Τα δεδομένα εκπαίδευσης (x_train) που περιλαμβάνουν χαρακτηριστικά όπως η ώρα της ημέρας, η διαδρομή, η κατεύθυνση και το σημείο στάσης

Έξοδος Αλγορίθμου

Η προβλεπόμενη έξοδος του επιβάτη (exit_stop_id).

Ημέρα 1

- Για τα πολύ μικρά βάθη (από 2.5 έως περίπου 5), η ακρίβεια του μοντέλου αυξάνεται δραματικά, γεγονός που δείχνει ότι τα δέντρα με μικρό βάθος δεν έχουν την ικανότητα να συλλάβουν τις σύνθετες σχέσεις στα δεδομένα.
- Μετά το βάθος των 5 μονάδων, η ακρίβεια συνεχίζει να αυξάνεται αλλά με μικρότερο ρυθμό. Αυτό υποδηλώνει ότι το δέντρο αρχίζει να προσαρμόζεται καλύτερα στα δεδομένα, συλλαμβάνοντας περισσότερες λεπτομέρειες.
- Από το βάθος των 10 μονάδων και μετά, η αύξηση της ακρίβειας φαίνεται να σταθεροποιείται γύρω στο 0.45, με μικρές διακυμάνσεις.

Ημέρα 2

- Και εδώ, παρατηρείται μια σημαντική αύξηση της ακρίβειας στα μικρά βάθη (από 2.5 έως περίπου 7.5), που υποδηλώνει ότι τα πολύ ρηχά δέντρα δεν είναι επαρκώς ικανά να διαχωρίσουν τα δεδομένα.
- Μετά το βάθος των 7.5 μονάδων, η ακρίβεια συνεχίζει να αυξάνεται αλλά με μειωμένο ρυθμό. Αυτό το μοτίβο είναι παρόμοιο με το πρώτο διάγραμμα, δείχνοντας ότι το δέντρο αρχίζει να μαθαίνει περισσότερες λεπτομέρειες.
- Από το βάθος των 12 μονάδων και μετά, η ακρίβεια τείνει να σταθεροποιηθεί γύρω στο 0.7. Αυτή η τάση είναι πιο εμφανής από την πρώτη μέρα, υποδηλώνοντας ότι το μοντέλο έχει καλύτερη ικανότητα γενίκευσης στα δεδομένα της δεύτερης μέρας.

Σύγκριση των δύο ημερών

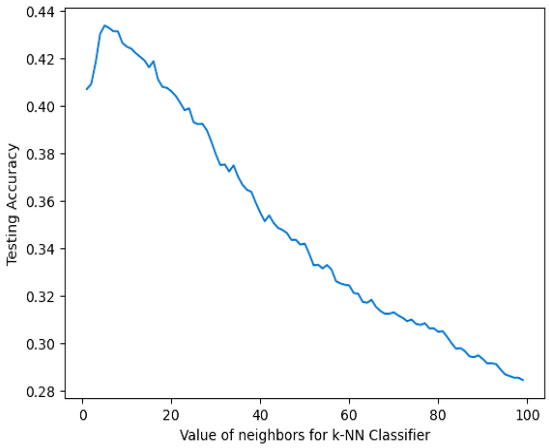
- Και στις δύο μέρες, παρατηρούμε μια παρόμοια τάση αύξησης της ακρίβειας με το βάθος του δέντρου, με την ακρίβεια να σταθεροποιείται σε μεγαλύτερα βάθη.
- Η ακρίβεια του μοντέλου είναι γενικά υψηλότερη στη δεύτερη μέρα σε σχέση με την πρώτη. Αυτό μπορεί να οφείλεται σε διάφορους παράγοντες, όπως η διαφορά στη σύνθεση των δεδομένων ή η διαφορετική κατανομή των δρομολογίων.
- Η πρώτη μέρα παρουσιάζει μια σταθεροποίηση της ακρίβειας γύρω στο 0.45, ενώ η δεύτερη μέρα γύρω στο 0.7. Αυτό υποδηλώνει ότι τα δεδομένα της δεύτερης μέρας ήταν πιθανώς πιο εύκολο να ταξινομηθούν ή περιείχαν λιγότερο θόρυβο.

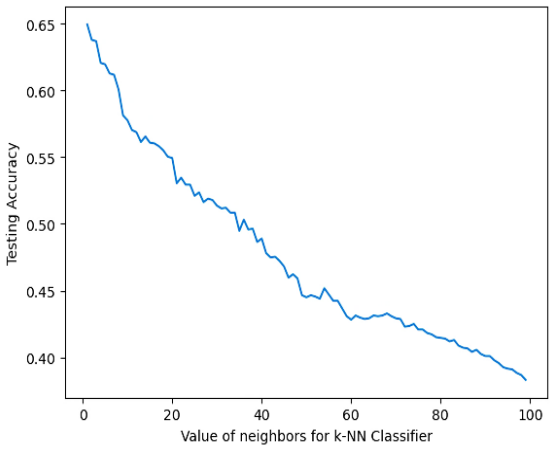
Συμπεράσματα

Η ανάλυση αυτή δείχνει ότι ο αλγόριθμος Decision Tree μπορεί να προσαρμοστεί καλά στα δεδομένα δρομολογίων, με την ακρίβεια να αυξάνεται όσο αυξάνεται το βάθος του δέντρου. Ωστόσο, υπάρχει ένα σημείο μετά το οποίο η περαιτέρω αύξηση του βάθους δεν προσφέρει σημαντική βελτίωση και μπορεί να οδηγήσει σε υπερεκπαίδευση. Η καλύτερη απόδοση του μοντέλου στη δεύτερη μέρα υποδηλώνει ότι η ποιότητα και η σύνθεση των δεδομένων παίζουν σημαντικό ρόλο στην αποτελεσματικότητα του αλγορίθμου.

4.2. Αποτελέσματα με χρήση K-Nearest Neighbors (k-NN)

Παρακάτω παρουσιάζεται η ακρίβεια πρόβλεψης του exit_stop_id,με χρήση του μοντέλου k-NN σε σχέση με τον αριθμό των γειτόνων.





Εικ.4:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1 **Εικ.5:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2**

Αριθμός Γειτόνων (k)	Ακρίβεια Ημέρα 1	Ακρίβεια Ημέρα 2
1	0.43	0.65
5	0.42	0.60
10	0.38	0.55
20	0.34	0.50
30	0.33	0.47
40	0.31	0.45
50	0.30	0.43
75	0.29	0.41
100	0.28	0.40

Πίνακας 3 : Πίνακας Τιμών k-NN

Είσοδος

Τα δεδομένα εκπαίδευσης με τα χαρακτηριστικά των διαδρομών και τις στάσεις. Αριθμός γειτόνων (n_neighbors = 20).

Έξοδος

Η προβλεπόμενη στάση εξόδου.

Ημέρα 1

- Στην αρχή, παρατηρείται μια αύξηση της ακρίβειας όσο αυξάνεται η τιμή του k από 1 έως περίπου 3. Αυτό υποδεικνύει ότι το μοντέλο βελτιώνεται καθώς λαμβάνει υπόψη περισσότερους γείτονες για την ταξινόμηση ενός σημείου δεδομένων.
- Η μέγιστη ακρίβεια φτάνει γύρω στο 0.42 για μικρές τιμές k (γύρω στο 3). Αυτό σημαίνει ότι το μοντέλο έχει καλύτερη απόδοση όταν λαμβάνει υπόψη λίγους γείτονες.
- Μετά από αυτή την κορυφή, η ακρίβεια αρχίζει να μειώνεται σταθερά καθώς αυξάνεται ο αριθμός των γειτόνων. Αυτή η μείωση υποδηλώνει ότι το μοντέλο γίνεται λιγότερο

ακριβές όταν λαμβάνει υπόψη πολλούς γείτονες, πιθανόν επειδή επηρεάζεται από την πολυπλοκότητα και την ποικιλία των δεδομένων.

- Για τιμές k από περίπου 50 και μετά, η ακρίβεια σταθεροποιείται γύρω στο 0.30. Αυτό δείχνει ότι το μοντέλο δεν καταφέρνει να ταξινομήσει σωστά με πολλούς γείτονες, πιθανόν επειδή υπεραπλουστεύει τα δεδομένα.

Ημέρα 2

- Αρχική Ακρίβεια: Η ακρίβεια ξεκινάει σε ένα υψηλότερο σημείο γύρω στο 0.65 για μικρές τιμές k, δείχνοντας ότι το μοντέλο έχει καλύτερη αρχική απόδοση.
- Μείωση Ακρίβειας: Παρόμοια με την πρώτη μέρα, η ακρίβεια μειώνεται καθώς αυξάνεται ο αριθμός των γειτόνων, αλλά σε μια πιο δραματική πτώση.
- Σταθεροποίηση**:** Μετά από τιμές k περίπου 50, η ακρίβεια σταθεροποιείται γύρω στο 0.40. Παρόλο που το μοτίβο είναι παρόμοιο με την πρώτη μέρα, η ακρίβεια είναι υψηλότερη σε όλες τις τιμές k.

Σύγκριση των δύο ημερών

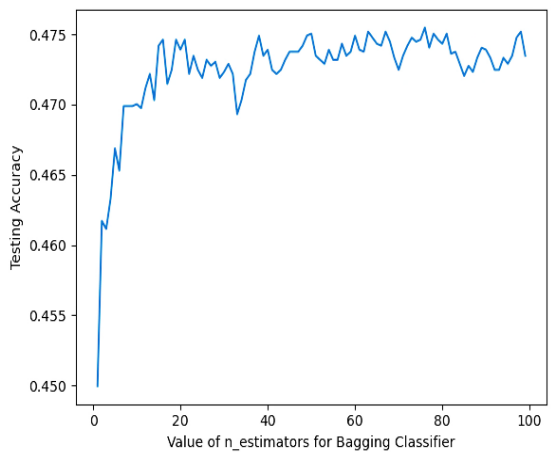
- Η ακρίβεια ξεκινάει από υψηλότερο σημείο την δεύτερη μέρα, υποδηλώνοντας καλύτερη αρχική απόδοση του μοντέλου k-NN με τα δεδομένα της δεύτερης μέρας.
- Και στις δύο μέρες, η ακρίβεια μειώνεται καθώς αυξάνεται ο αριθμός των γειτόνων, αλλά η μείωση είναι πιο έντονη την δεύτερη μέρα.
- Η ακρίβεια σταθεροποιείται σε χαμηλότερο επίπεδο την πρώτη μέρα (γύρω στο 0.30) σε σύγκριση με τη δεύτερη μέρα (γύρω στο 0.40).

Συμπεράσματα

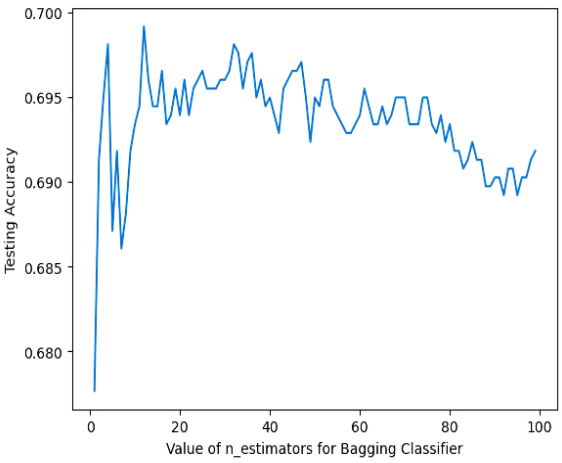
Η ανάλυση αυτή δείχνει ότι ο αλγόριθμος k-NN αποδίδει καλύτερα με λίγους γείτονες και η απόδοση του μειώνεται καθώς αυξάνεται ο αριθμός των γειτόνων. Η απόδοση του μοντέλου φαίνεται να εξαρτάται από την ποιότητα και τη σύνθεση των δεδομένων κάθε μέρας, με τη δεύτερη μέρα να δείχνει υψηλότερη ακρίβεια σε όλες τις τιμές k. Η σταθεροποίηση της ακρίβειας σε χαμηλότερο επίπεδο για μεγάλες τιμές k υποδηλώνει ότι η πολυπλοκότητα και η ποικιλία των δεδομένων επηρεάζει αρνητικά την απόδοση του μοντέλου όταν αυτό λαμβάνει υπόψη πολλούς γείτονες.

4.3. Αποτελέσματα με χρήση Bagging

Παρακάτω παρουσιάζεται η ακρίβεια πρόβλεψης του exit_stop_id,με χρήση του μοντέλου Bagging σε σχέση με τον αριθμό των εκτιμητών.



Εικ.6:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1



Εικ.7:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2

Αριθμός Εκτιμητών	Ακρίβεια Ημέρα 1	Ακρίβεια Ημέρα 2
20	0.470	0.695
40	0.474	0.695
60	0.472	0.692
80	0.470	0.690
100	0.470	0.693

Πίνακας 4 : Πίνακας Τιμών Bagging

Είσοδος

Τα ίδια χαρακτηριστικά όπως στους προηγούμενους αλγορίθμους. Ο αριθμός των δέντρων είναι n_estimators=200.

Έξοδος

Η πρόβλεψη στάσης εξόδου από το σύνολο των μοντέλων (σύνολο δέντρων).

Ημέρα 1

- Για τους μικρούς αριθμούς εκτιμητών (από 10 έως περίπου 20), η ακρίβεια του μοντέλου αυξάνεται σημαντικά, υποδεικνύοντας ότι μαθαίνει να προσαρμόζεται στα δεδομένα.
- Από τους 20 έως τους 40 εκτιμητές, η ακρίβεια συνεχίζει να αυξάνεται με διακυμάνσεις και με μικρότερο ρυθμό. Αυτό δείχνει ότι το μοντέλο αρχίζει να προσαρμόζεται καλύτερα στα δεδομένα, αξιοποιώντας περισσότερες πληροφορίες.
- Από τους 30 εκτιμητές και μετά, η αύξηση της ακρίβειας φαίνεται να σταθεροποιείται γύρω στο 0.475, με μικρές διακυμάνσεις. Αυτό υποδηλώνει ότι η περαιτέρω αύξηση του αριθμού των εκτιμητών δεν προσφέρει σημαντική βελτίωση στην ικανότητα του μοντέλου.

Ημέρα 2

- Και εδώ, παρατηρούμε μια σημαντική αύξηση της ακρίβειας στους πρώτους εκτιμητές (από 0 έως περίπου 20), με μεγάλη διακύμανση από τους 0 μέχρι 10.
- Μετά τους 20 εκτιμητές, η ακρίβεια συνεχίζει να αυξάνεται αλλά με μειωμένο ρυθμό, παρόμοιο με το πρώτο διάγραμμα, δείχνοντας ότι το μοντέλο αρχίζει να μαθαίνει περισσότερες λεπτομέρειες.
- Από τους 40 εκτιμητές και μετά, η ακρίβεια τείνει να σταθεροποιηθεί με πτωτική τάση γύρω στο 0.690, με μικρές διακυμάνσεις.

Σύγκριση των δύο ημερών

- Και στα δύο διαγράμματα παρατηρείται μια αρχική αύξηση στην ακρίβεια δοκιμής με την αύξηση του αριθμού των εκτιμητών. Αυτή η συμπεριφορά είναι συνήθης για τις μεθόδους συνόλου όπως το Bagging, όπου η αύξηση των εκτιμητών συνήθως βελτιώνει την απόδοση του μοντέλου μειώνοντας τη διασπορά.
- Η υψηλότερη ακρίβεια για την Ημέρα 1 επιτυγχάνεται γύρω στους 30 εκτιμητές, φτάνοντας περίπου στο 0.475 σε αντίθεση με την υψηλότερη ακρίβεια για την Ημέρα 2, που παρατηρείται λίγο νωρίτερα, γύρω στους 20 εκτιμητές, φτάνοντας κοντά στο 0.700.
- Και στις δύο μέρες παρατηρούνται διακυμάνσεις στην ακρίβεια μετά την επίτευξη του μέγιστου σημείου. Αυτές οι διακυμάνσεις υποδηλώνουν ότι η προσθήκη περισσότερων

εκτιμητών πέρα από ένα συγκεκριμένο σημείο δεν βελτιώνει σταθερά την απόδοση και μπορεί να εισάγει κάποια αστάθεια.

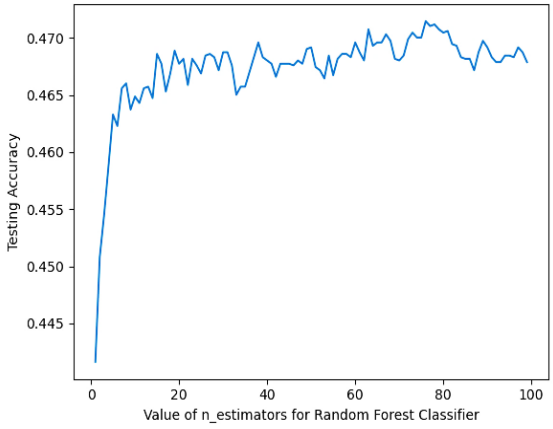
- Τα αποτελέσματα και των δύο ημερών υποδηλώνουν ότι ένας Bagging Classifier με περίπου 20-30 εκτιμητές είναι πιθανό να αποδώσει καλά, προσφέροντας μια ισορροπία μεταξύ προκατάληψης και διασποράς.

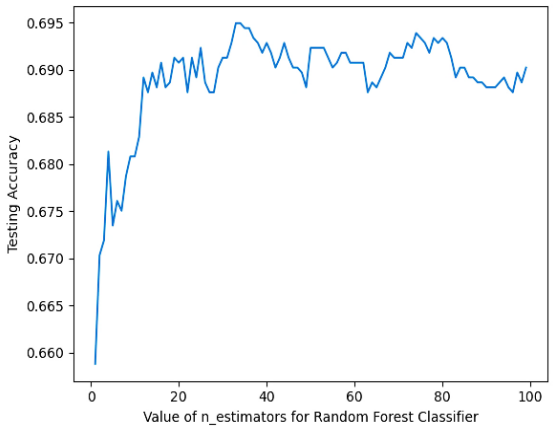
Συμπεράσματα

Η ανάλυση δείχνει ότι ο αλγόριθμος Bagging μπορεί να προσαρμοστεί καλά στα δεδομένα, με την ακρίβεια να αυξάνεται όσο αυξάνεται ο αριθμός των εκτιμητών. Ωστόσο, υπάρχει ένα σημείο μετά το οποίο η περαιτέρω αύξηση του αριθμού των εκτιμητών δεν προσφέρει σημαντική βελτίωση και μπορεί να οδηγήσει σε μείωση της ακρίβειας. Η καλύτερη απόδοση του μοντέλου στη δεύτερη μέρα υποδηλώνει ότι η ποιότητα και η σύνθεση των δεδομένων παίζουν σημαντικό ρόλο στην αποτελεσματικότητα του αλγορίθμου.

4.4. Αποτελέσματα με χρήση Random Forest

Παρακάτω παρουσιάζεται η ακρίβεια πρόβλεψης του exit_stop_id,με χρήση του μοντέλου Random Forest για διαφορετικές τιμές της παραμέτρου n_estimators, δηλαδή τον αριθμό των δέντρων στο δάσος.





Εικ.9:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1 Εικ.9: Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2

n_estimators (k)	Ακρίβεια Ημέρα 1	Ακρίβεια Ημέρα 2
5	0.450	0.65
20	0.465	0.60
40	0.467	0.55
60	0.468	0.50
80	0.469	0.47
100	0.470	0.45

Πίνακας 5 : Πίνακας Τιμών Random Forest

Είσοδος

Τα δεδομένα εκπαίδευσης και ο αριθμός των δέντρων (n_estimators=10).

Έξοδος

Η πρόβλεψη της στάσης εξόδου μέσω του συνδυασμού των προβλέψεων από τα πολλά δέντρα αποφάσεων.

Ημέρα 1

- Στις πρώτες τιμές των n_estimators παρατηρείται μια σταθερή άνοδος στην ακρίβεια, η οποία φτάνει το μέγιστο περίπου στους 20-25.
- Σταθεροποίηση**:** Μετά τις 25 εκτιμήσεις, η ακρίβεια σταθεροποιείται γύρω στο 0.465 με μικρές διακυμάνσεις.
- Αν και υπάρχουν κάποιες μικρές διακυμάνσεις, η γενική τάση παραμένει σταθερή χωρίς σημαντική βελτίωση.Αυτό υποδηλώνει ότι η προσθήκη περισσότερων δέντρων στο δάσος μετά από ένα σημείο δεν προσφέρει ουσιαστική βελτίωση στην απόδοση του μοντέλου, κάτι που είναι σύνηθες στο Random Forest.

Ημέρα 2

Στις πρώτες τιμές των n_estimators παρατηρείται μια σταθερή άνοδος στην ακρίβεια, η οποία φτάνει το μέγιστο περίπου στους 20-25, όπως και στην πρώτη ημέρα.

- Μετά τις 25 εκτιμήσεις, η ακρίβεια σταθεροποιείται γύρω στο 0.690 με μικρές διακυμάνσεις.
- Υπάρχουν μικρές διακυμάνσεις γύρω από αυτή την τιμή, αλλά χωρίς σημαντική βελτίωση με την προσθήκη περισσότερων δέντρων.

Σύγκριση των δύο ημερών

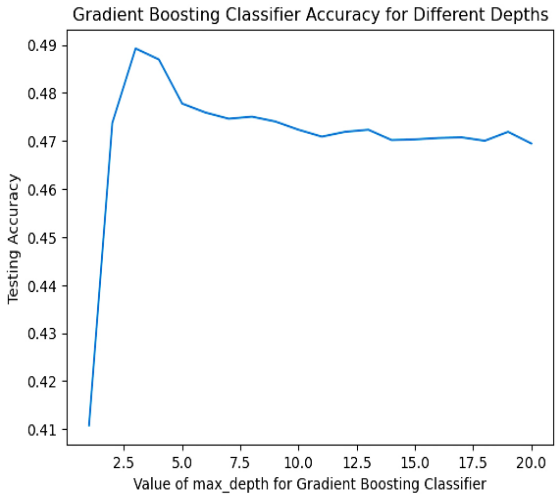
- Και για τις δύο ημέρες παρατηρείται ότι η ακρίβεια αυξάνεται γρήγορα, καθώς ο αριθμός των εκτιμητών αυξάνεται από το 0 έως περίπου το 20.
- Μετά από αυτό το σημείο, η αύξηση της ακρίβειας επιβραδύνεται και για τις δύο ημέρες και φαίνεται να σταθεροποιείται με μικρές διακυμάνσεις.
- Παρατηρείται ότι η συνολική ακρίβεια για τη δεύτερη ημέρα είναι σημαντικά υψηλότερη από την πρώτη ημέρα. Αυτό υποδηλώνει ότι τα δεδομένα της δεύτερης ημέρας είναι πιο συνεκτικά.

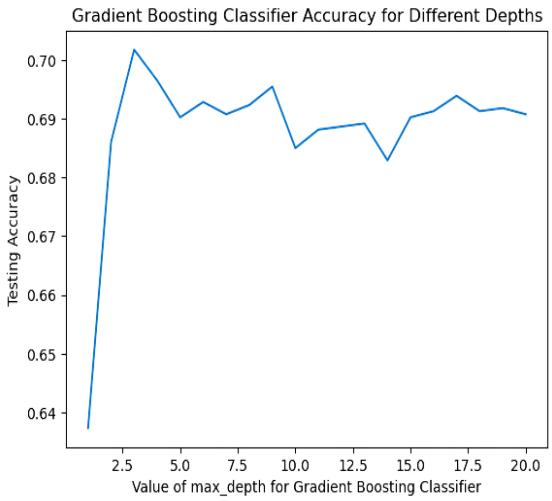
Συμπεράσματα

- Και στις δύο ημέρες, παρατηρούμε ότι η ακρίβεια του αλγορίθμου Random Forest αυξάνεται αρχικά με την αύξηση του αριθμού των δέντρων, αλλά στη συνέχεια σταθεροποιείται. Αυτό είναι χαρακτηριστικό των αλγορίθμων Random Forest, καθώς μετά από ένα σημείο οι προσθήκες νέων δέντρων δεν βελτιώνουν ουσιαστικά την απόδοση.
- Η σταθεροποίηση της ακρίβειας υποδηλώνει ότι το μοντέλο έχει μάθει τη δομή των δεδομένων και η προσθήκη περισσότερων δέντρων προσφέρει περιορισμένη βελτίωση.
- Παρατηρείται διαφορά στις τιμές της ακρίβειας μεταξύ των δύο ημερών (0.465 για την πρώτη ημέρα και 0.690 για τη δεύτερη). Αυτή η διαφορά μπορεί να οφείλεται στη διαφορά στα δεδομένα των δύο ημερών, όπως π.χ., διαφορετικά μοτίβα κίνησης.

4.5. Αποτελέσματα με χρήση Gradient Boosting

Παρακάτω παρουσιάζεται τη ακρίβεια πρόβλεψης του exit_stop_id,με χρήση του μοντέλου Gradient Boosting σε σχέση με τη μέγιστη τιμή του βάθους των δέντρων (max_depth).





Εικ.10:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1 Εικ.11: Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2

Βάθος Δέντρου	Ακρίβεια Ημέρα 1	Ακρίβεια Ημέρα 2
2.5	0.48	0.69
5.0	0.49	0.70
7.5	0.48	0.69
10.0	0.47	0.68
12.5	0.47	0.69
15.0	0.47	0.68

Πίνακας 6 : Πίνακας Τιμών Gradient Boosting

Είσοδος

Τα δεδομένα εκπαίδευσης και τα παραμετρικά χαρακτηριστικά όπως ο αριθμός των δέντρων (n_estimators=200) και το learning rate (0.1).

Έξοδος

Η πρόβλεψη της στάσης εξόδου, λαμβάνοντας υπόψη τα λάθη προηγούμενων δέντρων.

Ημέρα 1

- Στην αρχή, παρατηρείται μια σημαντική αύξηση της ακρίβειας καθώς το βάθος αυξάνεται από 1 έως περίπου 3.
- Η μέγιστη ακρίβεια φτάνει περίπου στο 0.49 για βάθη γύρω στο 3.5, υποδηλώνοντας ότι το μοντέλο έχει βέλτιστη απόδοση σε αυτό το βάθος.
- Μετά από αυτό το σημείο, η ακρίβεια παρουσιάζει μικρή πτώση και στη συνέχεια σταθεροποιείται γύρω στο 0.47-0.48 για βάθη από 7.5 και πάνω. Αυτό σημαίνει ότι το μοντέλο δεν βελτιώνεται από την αύξηση του βάθους πέρα από ένα σημείο.

Ημέρα 2

- Όπως και την πρώτη μέρα, η ακρίβεια αυξάνεται δραματικά στην αρχή και φτάνει στη μέγιστη τιμή περίπου στο βάθος 3.

- Η μέγιστη ακρίβεια για τη δεύτερη μέρα φτάνει περίπου στο 0.70 για βάθη γύρω στο 3, υποδεικνύοντας καλύτερη απόδοση σε σύγκριση με την πρώτη μέρα.
- Μετά από αυτό το σημείο, η ακρίβεια παρουσιάζει κάποιες διακυμάνσεις αλλά παραμένει σχετικά σταθερή γύρω στο 0.69 για μεγαλύτερα βάθη. Αυτό δείχνει ότι η αύξηση του βάθους πέρα από ένα σημείο δεν προσφέρει σημαντική βελτίωση.

Σύγκριση των δύο ημερών

- Η μέγιστη ακρίβεια τη δεύτερη μέρα (0.70) είναι σημαντικά υψηλότερη από την πρώτη μέρα (0.49). Αυτό υποδηλώνει ότι τα δεδομένα της δεύτερης μέρας ήταν πιο κατάλληλα για τον αλγόριθμο Gradient Boosting.
- Και στις δύο μέρες, η ακρίβεια τείνει να σταθεροποιηθεί μετά από ένα συγκεκριμένο βάθος.
- Η δεύτερη μέρα δείχνει περισσότερες διακυμάνσεις στην ακρίβεια καθώς αυξάνεται το βάθος, ενώ η πρώτη μέρα έχει πιο ομαλή καμπύλη μετά τη μέγιστη ακρίβεια.

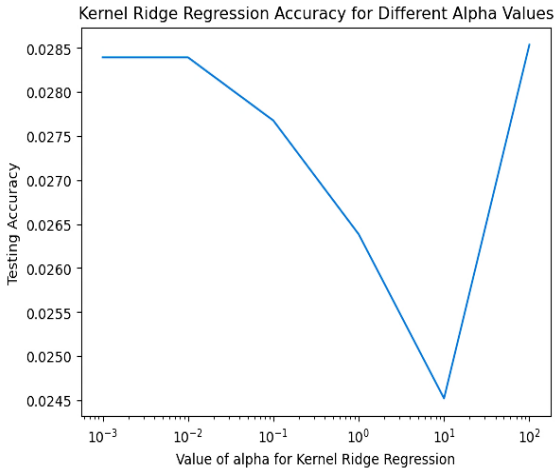
Συμπεράσματα

Η ανάλυση των αποτελεσμάτων δείχνει ότι ο αλγόριθμος Gradient Boosting αποδίδει καλύτερα για μικρά βάθη δέντρων (περίπου στο 3), με την απόδοση να μειώνεται ελαφρώς και να σταθεροποιείται καθώς αυξάνεται το βάθος. Η απόδοση του μοντέλου είναι σαφώς καλύτερη τη δεύτερη μέρα, υποδηλώνοντας ότι η ποιότητα και η σύνθεση των δεδομένων είναι καταλληλότερες για τον αλγόριθμό αυτό.

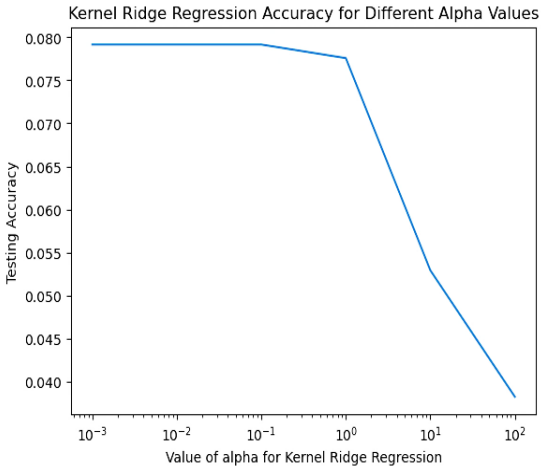
4.6. Αποτελέσματα με χρήση Kernel Ridge Regression

Παρακάτω παρουσιάζεται η ακρίβεια πρόβλεψης του exit_stop_id,με χρήση του μοντέλου Kernel Ridge Regression ανάλογα με τις τιμές των παραμέτρων alpha και gamma.

Alpha



Εικ.12:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1



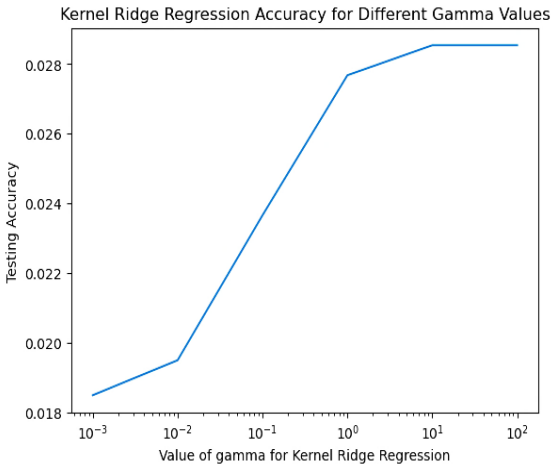
Εικ.13:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2

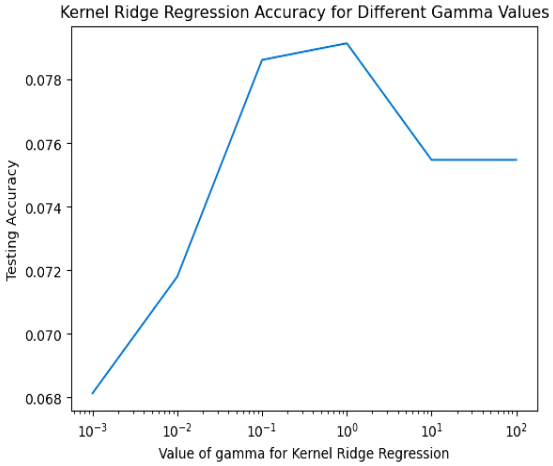
Τιμή Alpha	Ακρίβεια Ημέρα 1	Ακρίβεια Ημέρα 2
0.001	0.0285	0.0800
0.01	0.0280	0.0780
0.1	0.0275	0.0760
1	0.0265	0.0650

10	0.0255	0.0550
100	0.0245	0.0400

Πίνακας 7 : Πίνακας τιμών Kernel Ridge Regression

Gamma





Εικ.14:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1 **Εικ.15:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2**

Τιμή Alpha	Ακρίβεια Ημέρα 1	Ακρίβεια Ημέρα 2
0.001	0.0180	0.0680
0.01	0.0200	0.0720
0.1	0.0220	0.0740
1	0.0240	0.0780
10	0.0260	0.0760
100	0.0280	0.0760

Πίνακας 8 : Πίνακας Τιμών Kernel Ridge Regression

Είσοδος

Τα δεδομένα εκπαίδευσης και οι παράμετροι όπως alpha και gamma για τον υπολογισμό της παλινδρόμησης.

Έξοδος

Οι προβλέψεις της στάσης εξόδου.

Ημέρα 1

- **Alpha** :
    - Για τις πολύ μικρές τιμές alpha (10^{-3} έως 10^{-2}), η ακρίβεια παραμένει σταθερή γύρω στο 0.0285.
    - Με την αύξηση του alpha (10^{-1} έως 10^{1}), η ακρίβεια μειώνεται σταδιακά, φτάνοντας σε ένα ελάχιστο σημείο γύρω στο 0.0245.
    - Στη συνέχεια, με την περαιτέρω αύξηση του alpha, η ακρίβεια αυξάνεται ξανά απότομα, δείχνοντας ότι μεγάλες τιμές alpha μπορούν να βοηθήσουν στη βελτίωση της ακρίβειας.

- **Gamma** :
    - Για τις πολύ μικρές τιμές gamma (10^{-3} έως 10^{-2}), η ακρίβεια είναι χαμηλή, γύρω στο 0.018.
    - Με την αύξηση του gamma (10^{-1}έως 10^{2}), η ακρίβεια αυξάνεται σταθερά, φτάνοντας το μέγιστο σημείο γύρω στο 0.028.
    - Για ακόμα μεγαλύτερες τιμές gamma η ακρίβεια παραμένει σταθερή γύρω στο 0.028.

Ημέρα 2

- **Alpha** :
    - Για τις πολύ μικρές τιμές alpha (10^{-3} έως 10^{-2}), η ακρίβεια παραμένει σταθερή γύρω στο 0.080.
    - Με την αύξηση του alpha (10^{-1} έως 10^{1}), η ακρίβεια μειώνεται σταδιακά, φτάνοντας σε ένα ελάχιστο σημείο γύρω στο 0.040.
    - Η πτώση της ακρίβειας υποδηλώνει ότι η αύξηση του alpha πέρα από ένα συγκεκριμένο σημείο μπορεί να είναι επιζήμια για την απόδοση του μοντέλου.
- **Gamma** :
    - Για τις πολύ μικρές τιμές gamma (10^{-3} έως 10^{-2}), η ακρίβεια είναι χαμηλή, γύρω στο 0.068.
    - Με την αύξηση του gamma (10^{-1}έως 10^{1}), η ακρίβεια αυξάνεται σταθερά, φτάνοντας το μέγιστο σημείο γύρω στο 0.078.
    - Για πολύ μεγάλες τιμές gamma (10^{2}), η ακρίβεια παραμένει σταθερή.

Σύγκριση των δύο ημερών

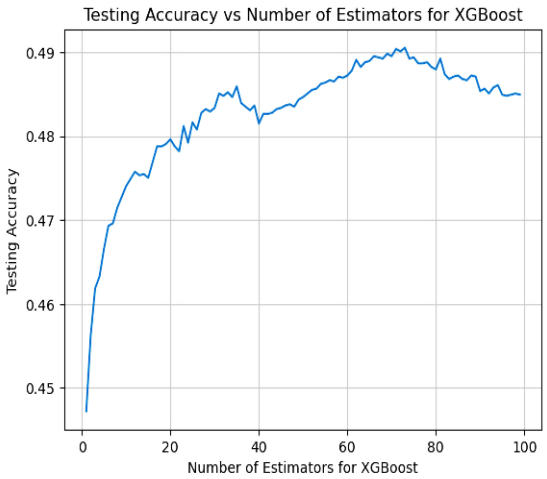
- **Alpha** :
    - Ημέρα 1: Η ακρίβεια μειώνεται σημαντικά με την αύξηση του alpha και αυξάνεται ξανά απότομα στις πολύ μεγάλες τιμές.
    - Ημέρα 2: Η ακρίβεια μειώνεται σταθερά με την αύξηση του alpha, χωρίς την απότομη αύξηση που παρατηρήθηκε στην Ημέρα 1.
- **Gamma** :
    - Και στις δύο μέρες, η αύξηση του gamma οδηγεί σε βελτίωση της ακρίβειας, με την ακρίβεια να φτάνει το μέγιστο σημείο σε υψηλότερες τιμές gamma.
    - Ημέρα 1: Η ακρίβεια φτάνει το μέγιστο σημείο γύρω στο 0.028 για μεγάλες τιμές gamma.
    - Ημέρα 2: Η ακρίβεια φτάνει το μέγιστο σημείο γύρω στο 0.078 και μειώνεται ελαφρώς για πολύ μεγάλες τιμές gamma.

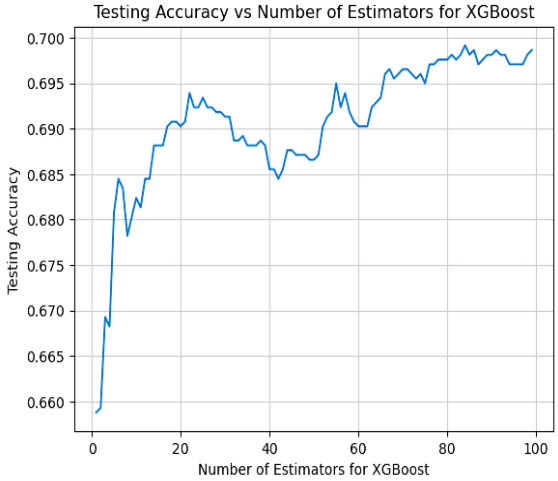
Συμπεράσματα

Η ανάλυση των αποτελεσμάτων δείχνει ότι η απόδοση του αλγορίθμου Kernel Ridge Regression επηρεάζεται σημαντικά από τις παραμέτρους alpha και gamma. Η βέλτιστη τιμή για την alpha μπορεί να ποικίλει σημαντικά ανάλογα με τα δεδομένα, ενώ η αύξηση του gamma γενικά βελτιώνει την ακρίβεια μέχρι ένα συγκεκριμένο σημείο. Για τη βέλτιστη απόδοση, είναι σημαντικό να γίνεται προσεκτική ρύθμιση των παραμέτρων μέσω διαδικασιών όπως cross-validation.

4.7. Αποτελέσματα με χρήση XGBoost

Παρακάτω παρουσιάζεται τη ακρίβεια πρόβλεψης του exit_stop_id,με χρήση του μοντέλου XGBoost σε σχέση με τον αριθμό των εκτιμητών.





Εικ.16:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1 **Εικ.17:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2**

Αριθμός Εκτιμητών	Ακρίβεια Ημέρα 1	Ακρίβεια Ημέρα 2
5	0.460	0.670
10	0.470	0.680
20	0.475	0.685
40	0.485	0.690
60	0.490	0.695
80	0.490	0.700

Πίνακας 9 : Πίνακας Τιμών XGBoost

Είσοδος

Δεδομένα εκπαίδευσης, αριθμός δέντρων (n_estimators=65).

Έξοδος

Προβλέψεις της στάσης εξόδου που βασίζονται σε ενισχυμένα δέντρα αποφάσεων.

Ημέρα 1

- Στο πρώτο γράφημα, βλέπουμε ότι η ακρίβεια της δοκιμής ξεκινά από περίπου 0.45 και αυξάνεται σταθερά καθώς αυξάνεται ο αριθμός των εκτιμητών. Αυτή η αρχική φάση δείχνει γρήγορη βελτίωση στην απόδοση.
- Η ακρίβεια φτάνει στο μέγιστο περίπου στο 0.49 όταν ο αριθμός των εκτιμητών είναι γύρω στους 70-80. Μετά από αυτό το μέγιστο, η ακρίβεια παρουσιάζει μικρές διακυμάνσεις αλλά γενικά σταθεροποιείται.
- Υπάρχουν μικρές διακυμάνσεις γύρω από την κορυφαία τιμή, αλλά η συνολική τάση δεν δείχνει σημαντικές βελτιώσεις πέραν των 70-80 εκτιμητών.

Ημέρα 2

- Στο δεύτερο γράφημα, η ακρίβεια της δοκιμής ξεκινά από περίπου 0.66 και δείχνει ένα παρόμοιο μοτίβο γρήγορης αύξησης καθώς αυξάνεται ο αριθμός των εκτιμητών.
- Η ακρίβεια φτάνει στο μέγιστο περίπου στο 0.70 γύρω στους 80 εκτιμητές. Μετά από αυτό το σημείο, η ακρίβεια σταθεροποιείται με μικρές διακυμάνσεις.
- Όπως και στο πρώτο γράφημα, υπάρχουν μικρές διακυμάνσεις γύρω από την κορυφαία τιμή, υποδεικνύοντας ότι η προσθήκη περισσότερων εκτιμητών πέρα από τους 80 δεν βελτιώνει σημαντικά την απόδοση.

Σύγκριση των δύο ημερών

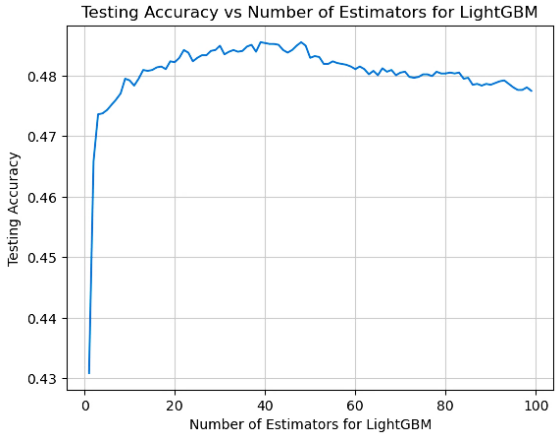
- Και στα δύο διαγράμματα παρατηρείται μια αρχική αύξηση στην ακρίβεια δοκιμής με την αύξηση του αριθμού των εκτιμητών.
- Η υψηλότερη ακρίβεια για την Ημέρα 1 επιτυγχάνεται γύρω στους 75 εκτιμητές, φτάνοντας σε αντίθεση με την υψηλότερη ακρίβεια για την Ημέρα 2, που παρατηρείται γύρω στους 80.
- Και στις δύο μέρες παρατηρούνται διακυμάνσεις στην ακρίβεια με μεγαλύτερη μεταβλητότητα την ημέρα δύο.
- Η απόδοση του αλγορίθμου σταθεροποιείται σχετικά μετά από έναν συγκεκριμένο αριθμό επαναλήψεων, δείχνοντας ότι έχει μάθει επαρκώς από τα δεδομένα της εκάστοτε ημέρας. Ωστόσο, η συνολική ακρίβεια της δεύτερης ημέρας είναι υψηλότερη σε σύγκριση με την πρώτη.

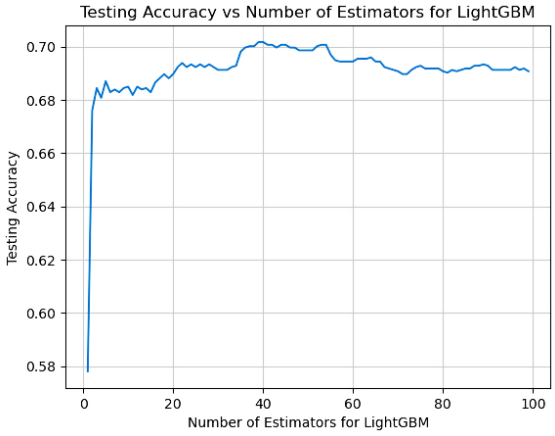
Συμπεράσματα

- Και τα δύο γραφήματα δείχνουν ένα συνεπές μοτίβο όπου η ακρίβεια αυξάνεται αρχικά με τον αριθμό των εκτιμητών, φτάνει σε μια κορυφή και στη συνέχεια σταθεροποιείται με μικρές διακυμάνσεις. Αυτό δείχνει ότι η απόδοση του XGBoost βελτιώνεται με περισσότερους εκτιμητές μέχρι ένα σημείο, μετά το οποίο οι οριακές βελτιώσεις είναι ελάχιστες.
- Για και τις δύο ημέρες, ο βέλτιστος αριθμός εκτιμητών φαίνεται να είναι γύρω στους 70-80. Πέρα από αυτό το εύρος, η προσθήκη περισσότερων εκτιμητών δεν βελτιώνει ουσιαστικά την ακρίβεια.
- Οι κορυφαίες τιμές ακρίβειας διαφέρουν μεταξύ των δύο ημερών (0.49 έναντι 0.70), κάτι που μπορεί να οφείλεται σε διαφορετικά σύνολα δεδομένων, διαφορετικά βήματα προεπεξεργασίας ή άλλους υποκείμενους παράγοντες. Αυτό τονίζει τη σημασία της ειδικής ρύθμισης για τον αριθμό των εκτιμητών στο XGBoost ανάλογα με το εκάστοτε σύνολο δεδομένων.

4.8. Αποτελέσματα με χρήση LightGBM

Παρακάτω παρουσιάζεται τη ακρίβεια πρόβλεψης του exit_stop_id,με χρήση του μοντέλου LightGBM για τις διαφορετικές τιμές του παραμέτρου n_estimators, δηλαδή τον αριθμό των εκτιμητών





Εικ.18:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1 **Εικ.19:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2**

n_estimators	Ακρίβεια Ημέρα 1	Ακρίβεια Ημέρα 2
20	0.480	0.690
40	0.480	0.680
60	0.475	0.675
80	0.470	0.680
100	0.468	0.680

Πίνακας 10 : Πίνακας Τιμών LightGBM

Είσοδος

Χαρακτηριστικά διαδρομών και στάσεων με αριθμό δέντρων (n_estimators=50).

Έξοδος

Προβλέψεις της στάσης εξόδου με χρήση μικρών φύλλων στα δέντρα.

Ημέρα 1

- Στις πρώτες τιμές των n_estimators παρατηρείται μια σταθερή άνοδος στην ακρίβεια, η οποία φτάνει το μέγιστο περίπου για τους 15-20.
- Μετά τις 20 εκτιμήσεις, η ακρίβεια σταθεροποιείται γύρω στο 0.48 με μικρές διακυμάνσεις.
- Αν και υπάρχουν κάποιες μικρές διακυμάνσεις, η γενική τάση παραμένει σταθερή χωρίς σημαντική βελτίωση. Αυτό υποδηλώνει ότι η προσθήκη περισσότερων εκτιμητών μετά από ένα σημείο δεν προσφέρει ουσιαστική βελτίωση στην απόδοση του μοντέλου, κάτι που είναι σύνηθες στους αλγορίθμους του τύπου Gradient Boosting όπως το LightGBM.

Ημέρα 2

- Στις πρώτες τιμές των n_estimators παρατηρείται μια σταθερή άνοδος στην ακρίβεια, η οποία φτάνει το μέγιστο περίπου για τους 15-20, όπως και στην πρώτη ημέρα.
- Μετά τις 20 εκτιμήσεις, η ακρίβεια σταθεροποιείται γύρω στο 0.69 με μικρές διακυμάνσεις.
- Υπάρχουν μικρές διακυμάνσεις γύρω από αυτή την τιμή, αλλά χωρίς σημαντική βελτίωση με την προσθήκη περισσότερων εκτιμητών.

Σύγκριση των δύο ημερών

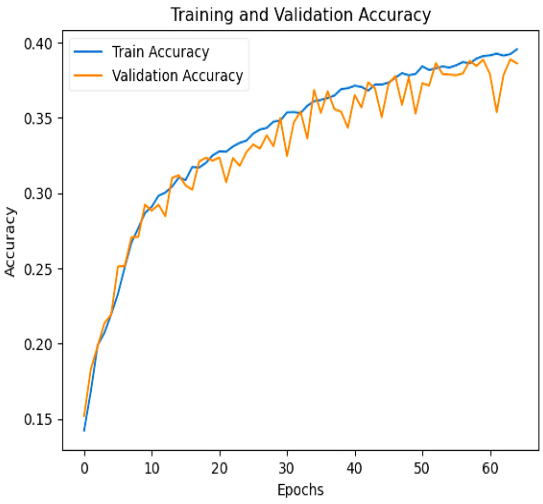
- Και για τις δύο ημέρες, η ακρίβεια αυξάνεται σταθερά με την αύξηση του αριθμού των επαναλήψεων (iterations). Ωστόσο, η συνολική ακρίβεια για τη δεύτερη ημέρα είναι υψηλότερη από την πρώτη ημέρα, γεγονός που να υποδηλώνει πως περισσότερες πληροφορίες που μπορούν να αξιοποιηθούν από τον αλγόριθμο.
- Η βελτίωση στην ακρίβεια φαίνεται να σταθεροποιείται μετά από έναν ορισμένο αριθμό επαναλήψεων, και για τις δύο ημέρες υποδεικνύοντας ότι ο αλγόριθμος έχει μάθει όσο το δυνατόν περισσότερα από τα δεδομένα που του δόθηκαν.

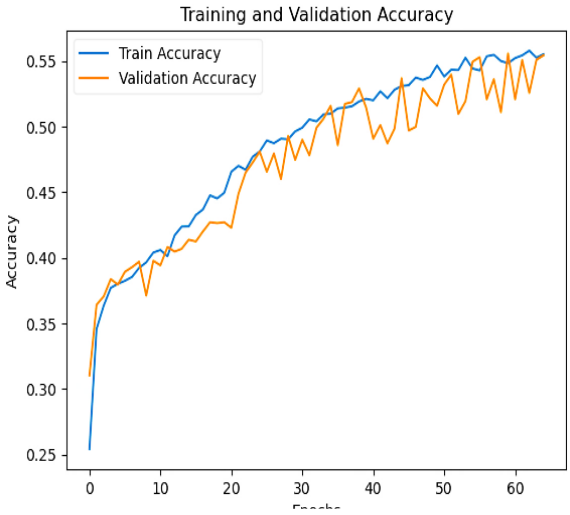
Συμπεράσματα

- Και στις δύο ημέρες, παρατηρούμε ότι η ακρίβεια του αλγορίθμου LightGBM αυξάνεται αρχικά με την αύξηση του αριθμού των εκτιμητών, αλλά στη συνέχεια σταθεροποιείται. Αυτό είναι χαρακτηριστικό των αλγορίθμων Gradient Boosting, καθώς μετά από ένα σημείο οι προσθήκες νέων εκτιμητών δεν βελτιώνουν ουσιαστικά την απόδοση.
- Η σταθεροποίηση της ακρίβειας υποδηλώνει ότι το μοντέλο έχει μάθει τη δομή των δεδομένων και η προσθήκη περισσότερων εκτιμητών προσφέρει περιορισμένη βελτίωση.
- Παρατηρείται διαφορά στις τιμές της ακρίβειας μεταξύ των δύο ημερών (0.48 για την πρώτη ημέρα και 0.69 για τη δεύτερη). Αυτή η διαφορά οφείλεται στη διαφορά των δεδομένα των δύο ημερών, όπως π.χ., διαφορετικά μοτίβα κίνησης ή ανωμαλίες στα δεδομένα.

4.9. Αποτελέσματα με χρήση Νευρωνικού Δικτύου Multilayer Perceptron (MLP) τεσσάρων επιπέδων

Παρακάτω παρουσιάζεται η ακρίβεια πρόβλεψης του exit_stop_id, κατά τη διάρκεια της εκπαίδευσης και της επικύρωσης με χρήση νευρωνικού δικτύου MLP τεσσάρων επιπέδων για τους διάφορους αριθμούς εποχών (epochs).





Εικ.20:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 1 **Εικ.21:Διάγραμμα Μεταβολής Ακρίβειας Ημέρα 2**

Εποχές	Ακρίβεια Εκπαίδευσης Ημέρα 1	Ακρίβεια Επικύρωσης Ημέρα 1	Ακρίβεια Εκπαίδευσης Ημέρα 2	Ακρίβεια Επικύρωσης Ημέρα 2
20	0.250	0.310	0.420	0.430
40	0.380	0.370	0.510	0.500
60	0.400	0.390	0.540	0.530

Πίνακας 11 : Πίνακας Τιμών Νευρωνικού Δικτύου

Είσοδος

Κλιμακωμένα χαρακτηριστικά με χρήση του StandardScaler και κατηγοριοποιημένη έξοδος (one-hot encoding) για το exit_stop_id.

Έξοδος

Οι πιθανότητες για κάθε στάση εξόδου.

Ημέρα 1

- Στην αρχή της εκπαίδευσης, η ακρίβεια αυξάνεται ραγδαία τόσο για το σύνολο εκπαίδευσης όσο και για το σύνολο επικύρωσης.
- Μετά από περίπου 20 εποχές, η αύξηση της ακρίβειας γίνεται πιο αργή και σταδιακά σταθεροποιείται.
- Η μέγιστη ακρίβεια που επιτυγχάνεται είναι περίπου 0.40 για το σύνολο εκπαίδευσης και το σύνολο επικύρωσης.
- Αυτό σημαίνει πως το νευρωνικό δίκτυο μαθαίνει από τα χαρακτηριστικά των δεδομένων της πρώτης ημέρας μη μπορώντας παρόλα αυτά να ξεπεράσει το ποσοστό ακρίβειας του 40% για τις 60 εποχές εκπαίδευσής του.

Ημέρα 2

- Η ακρίβεια αυξάνεται γρήγορα στις πρώτες εποχές και στη συνέχεια σταθεροποιείται.
- Η μέγιστη ακρίβεια που επιτυγχάνεται είναι περίπου 0.55 για το σύνολο εκπαίδευσης και το σύνολο επικύρωσης, η οποία είναι σημαντικά υψηλότερη σε σύγκριση με την πρώτη ημέρα.
- Αυτό υποδηλώνει ότι τα δεδομένα της δεύτερης ημέρας ενδέχεται να είναι πιο εύκολα προβλέψιμα και ότι το νευρωνικό δίκτυο μπόρεσε να μάθει καλύτερα από αυτά.

Σύγκριση των δύο ημερών

- Το νευρωνικό δίκτυο επιτυγχάνει καλύτερη απόδοση τη δεύτερη ημέρα σε σχέση με την πρώτη.
- Η ακρίβεια του μοντέλου είναι μεγαλύτερη τη δεύτερη ημέρα, φτάνοντας το 0.55 σε σύγκριση με το 0.40 της πρώτης ημέρας.
- Η σταθεροποίηση της ακρίβειας συμβαίνει περίπου στον ίδιο αριθμό εποχών και για τις δύο ημέρες, υποδηλώνοντας παρόμοια διαδικασία εκμάθησης.

Συμπεράσματα

- Η ακρίβεια τόσο στην εκπαίδευση όσο και στην επικύρωση αυξάνεται αυξάνεται συνεχώς με την αύξηση των εποχών (epochs).
- Στη συνέχεια, η ακρίβεια φαίνεται να σταθεροποιείται, υποδηλώνοντας ότι το δίκτυο έχει φτάσει κοντά στη μέγιστη απόδοσή του με τα δεδομένα εκπαίδευσης. Η διαφορά μεταξύ

της ακρίβειας εκπαίδευσης και της ακρίβειας επικύρωσης είναι μικρή, υποδηλώνοντας ότι το δίκτυο δεν υπερπροσαρμόζεται στα δεδομένα εκπαίδευσης και ότι το μοντέλο επιτυγχάνει καλή γενίκευση.

- Η σύγκλιση των καμπυλών εκπαίδευσης και επικύρωσης υποδηλώνει καλή γενίκευση του μοντέλου, και μικρές διακυμάνσεις στην ακρίβεια επικύρωσης είναι φυσιολογικές και οφείλονται στην στοχαστική φύση του αλγορίθμου εκπαίδευσης.
- Το νευρωνικό δίκτυο έχει μάθει τη δομή των δεδομένων εκπαίδευσης και η προσθήκη περισσότερων εποχών μετά από ένα σημείο δεν προσφέρει ουσιαστική βελτίωση στην απόδοση του μοντέλου.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Με βάση την ανάλυση των αποτελεσμάτων για τους αλγορίθμους Decision Tree, k-NN, Random Forest, Bagging, Gradient Boosting, Kernel Ridge Regression, Neural Network, XGBoost και LightGBM, μπορούμε να εξάγουμε σημαντικά συμπεράσματα για την εκτίμηση της στάσης εξόδου με χρήση των δεδομένων που δόθηκαν. Οι διάφορες μέθοδοι παρουσίασαν διαφορετικές επιδόσεις ανάλογα με τις παραμέτρους και τις συνθήκες εκπαίδευσης.

Οι παραδοσιακοί αλγόριθμοι όπως τα Decision Trees και k-NN έδειξαν ικανοποιητικά αποτελέσματα, αλλά οι σύγχρονες τεχνικές όπως τα Random Forest, Bagging και Gradient Boosting απέδειξαν ότι η συνδυαστική προσέγγιση (ensemble methods) προσφέρει σημαντικές βελτιώσεις στην ακρίβεια και την γενίκευση. Οι Kernel Ridge Regression και Neural Networks παρείχαν πιο ευέλικτες προσεγγίσεις με αξιοσημείωτη προσαρμοστικότητα στα δεδομένα, ενώ οι πιο πρόσφατοι αλγόριθμοι XGBoost και LightGBM ξεχώρισαν για την αποδοτικότητα και την ταχύτητά τους στην εκπαίδευση μεγάλων συνόλων δεδομένων.

Η συγκριτική μελέτη επιβεβαιώνει ότι δεν υπάρχει ένας "καλύτερος" αλγόριθμος που να ταιριάζει σε όλα τα προβλήματα. Αντιθέτως, η επιλογή του κατάλληλου αλγορίθμου εξαρτάται από τη φύση των δεδομένων, την απαίτηση για ακρίβεια, την υπολογιστική πολυπλοκότητα και τους περιορισμούς χρόνου. Συνολικά, η εφαρμογή και ο συνδυασμός διαφορετικών τεχνικών μπορεί να οδηγήσει στη βέλτιστη λύση, προσαρμοσμένη στις συγκεκριμένες ανάγκες του εκάστοτε προβλήματος.

Η παρούσα μελέτη παρέχει μια ολοκληρωμένη κατανόηση της συμπεριφοράς και των χαρακτηριστικών κάθε αλγορίθμου σε δεδομένα δρομολογίων, προσφέροντας χρήσιμες κατευθυντήριες γραμμές για μελλοντικές εφαρμογές και έρευνες στον τομέα της μηχανικής μάθησης.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

[1] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

[2] Breiman, L. (1996). *Bagging predictors. Machine learning, 24(2), (pp. 123-140).*

[3] Breiman, L. (2001). *Random forests. Machine learning, 45(1), (pp. 5-32).*

[4] Breiman, L. F. (1984). *Classification and Regression Trees.* Wadsworth.

[5] Bühlmann, P. &. (2002). *Analyzing bagging. The Annals of Statistics, 30(4), (pp. 927-961).*

[6] Cover, T. M. (1967). *Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), (pp. 21-27).*

[7] Friedman, J. H. (2001). *Greedy function approximation: a gradient boosting machine. Annals of statistics, (pp. 1189-1232).*

[8] Goodfellow, I. B. (2016). *Deep Learning.* MIT Press.

[9] Guestrin, T. C. (2016). *XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794) https://doi.org/10.1145/2939672.2939785.* New York, NY, USA: Association for Computing Machinery.

[10] Guolin Ke, Q. M. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree.* NIPS.

[11] Hastie, T. T. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. .* Springer.

[12] https://data.gov.lv/dati/lv/dataset/e-talonu-validaciju-dati-rigas-satiksme-sabiedriskajos-transportlidzeklos. (n.d.).

[13] https://saraksti.lv/. (n.d.).

[14] Ke, G. M. (2017). *LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (pp. 3146-3154).*

[15] Liaw, A. &. Wiener, M. (2002). *Classification and regression by random Forest (pp. 18-22).* R news, 2(3).

[16] Peterson, L. E. (2009). *K-nearest neighbor.* Scholarpedia, 4(2), 1883.

[17] Quinlan, J. R. (1986). *Induction of decision trees. Machine learning, 1(1), (pp. 81-106).*

[18] Saunders, C. G. (1998). *Ridge regression learning algorithm in dual variables. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML) (pp. 515-521).*

[19] Shawe-Taylor, J. &. (2004). *Methods for Pattern Analysis.* Cambridge University Press.

[20] Tianqi Chen, C. G. (2016). *XGBoost: A Scalable Tree Boosting System.* KDD.

[21] Kira, K., & Rendell, L. A. (1992). *A Practical Approach to Feature Selection. In Proceedings of the 9th International Conference on Machine Learning* (pp. 249-256).

[22] Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press.

[23] Aha, D. W., Kibler, D., & Albert, M. K. (1991). *Instance-Based Learning Algorithms. Machine Learning*, 6(1). (pp. 37-66).

[24] Friedman, J. H. (2002). *Stochastic Gradient Boosting. Computational Statistics & Data Analysis,* 38(4). (pp. 367-378).

[25] Haykin, S. (2009). *Neural Networks and Learning Machines (3rd ed.).* Pearson Education.

[26] Rojas, R. (1996). *Neural Networks: A Systematic Introduction*. Springer.