ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΔΙΚΤΥΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΣΧΟΛΗ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΜΣ: ΠΛΗΡΟΦΟΡΙΚΗΣ & ΔΙΚΤΥΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Δεδομένων Προόδου Μαθητών για Προβλέψεις Απόδοσης

Περσεφόνη Πατσέα - 185

Επιβλέπων: Νικόλαος Γιαννακέας

Άρτα, Μάρτιος 2025

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΜΣ: ΠΛΗΡΟΦΟΡΙΚΗΣ & ΔΙΚΤΥΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Δεδομένων Προόδου Μαθητών για Προβλέψεις Απόδοσης

Περσεφόνη Πατσέα - 185

Επιβλέπων: Νικόλαος Γιαννακέας

Άρτα, Μάρτιος 2025

Analysis of Student Progress Data for Performance Predictions

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Άρτα, 28/03/2025

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1) Επιβλέπων καθηγητής
   Νικόλαος Γιαννακέας,
   Επίκουρος Καθηγητής

2) Χρήστος Γκόγκος,
   Καθηγητής

3) Ιωάννης Τσούλος,
   Καθηγητής

Ο Προϊστάμενος του Τμήματος

Αλέξανδρος Τζάλλας

Υπογραφή

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα διπλωματική εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Πατσέα Περσεφόνη

Υπογραφή

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Γιαννακέα Νικόλαο, για την καθοδήγηση που προσέφερε και το χρόνο που διέθεσε, δίνοντας κατ' αυτόν τον τρόπο χρήσιμες συμβουλές και οδηγίες, για την περάτωση της διπλωματικής εργασίας. Στο ίδιο πλαίσιο ευγνωμοσύνης, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου, που στάθηκαν σημαντικοί αρωγοί στην προσπάθεια μου, και για την προσφορά της αμέριστης υποστήριξής και καθοδήγησης τους, σε κάθε στάδιο της προσωπικής μου πορείας.

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στη χρήση αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψη της ακαδημαϊκής απόδοσης μαθητών, βασισμένη σε δύο σύνολα δεδομένων (student-mat.csv και student-por.csv). Η πρόβλεψη της μαθητικής απόδοσης αποτελεί σημαντικό ερευνητικό πεδίο, καθώς επιτρέπει την εφαρμογή εξατομικευμένων στρατηγικών μάθησης και έγκαιρων παρεμβάσεων για την ενίσχυση των μαθητών. Στη βιβλιογραφία, έχουν προταθεί διάφορες μέθοδοι πρόβλεψης, όπως Γραμμική Παλινδρόμηση, Decision Trees, Random Forest και Neural Networks, με έμφαση στη σημασία χαρακτηριστικών όπως οι βαθμολογίες, η συμμετοχή σε δραστηριότητες και οι κοινωνικοοικονομικοί παράγοντες.

Αρχικά, πραγματώθηκε προεπεξεργασία δεδομένων, όπου εφαρμόστηκαν τεχνικές διαχείρισης ελλιπών τιμών, κανονικοποίησης και μετατροπής κατηγορικών χαρακτηριστικών. Στη συνέχεια, αναπτύχθηκαν και αξιολογήθηκαν τα εξής δώδεκα μοντέλα ταξινόμησης: Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbors, Support Vector Machines (SVM - SVC), Naïve Bayes, MLP, Gradient Boosting, AdaBoost, XGBoost, Bagging Classifier και Extra Trees.

Για την αξιολόγηση των μοντέλων, χρησιμοποιήθηκαν μετρικές απόδοσης όπως Accuracy, Precision, Recall, F1-score, ενώ πραγματοποιήθηκε και σύγκριση Confusion Matrices, Heatmaps, Learning Curves και Bar Charts. Τα αποτελέσματα έδειξαν ότι τα Gradient Boosting, XGBoost και Random Forest υπερτερούν σε ακρίβεια και γενίκευση, με ακρίβεια κοντά στο 90%. Αντίθετα, το Naïve Bayes είχε τη χαμηλότερη απόδοση (~58%), υποδεικνύοντας περιορισμένη καταλληλότητα για την ανάλυση των συγκεκριμένων δεδομένων.

Επιπροσθέτως, τα αποτελέσματα κατέδειξαν ότι το dataset student-por.csv παρείχε ελαφρώς καλύτερη απόδοση των μοντέλων σε σχέση με το student-mat.csv, γεγονός που ενδέχεται να σχετίζεται με τη διαφορετική κατανομή των χαρακτηριστικών των μαθητών.

Τέλος, γίνεται νύξη προτάσεις για βελτίωση, όπως η βελτιστοποίηση υπερπαραμέτρων (Grid Search, Bayesian Optimization), η ενσωμάτωση επιπλέον χαρακτηριστικών και η διερεύνηση εναλλακτικών μεθόδων Μηχανικής Μάθησης, όπως Νευρωνικά Δίκτυα.

Λέξεις-κλειδιά: Μηχανική Μάθηση, Πρόβλεψη Απόδοσης Μαθητών, Ταξινόμηση, Ανάλυση Δεδομένων.

Abstract

This thesis focuses on the use of Machine Learning algorithms to predict students' academic performance, based on two datasets (student-mat.csv and student-por.csv). The prediction of student performance is an important research area, as it allows the implementation of personalized learning strategies and early interventions to support students. In the literature, several prediction methods have been proposed, such as Linear Regression, Decision Trees, Random Forest and Neural Networks, with a focus on the importance of characteristics such as grades, activity participation and socioeconomic factors.

Initially, data preprocessing was carried out, where techniques of missing value management, normalization and categorical attribute transformation were applied. Then, twelve classification models were developed and evaluated: random forest, decision tree, decision tree, logistic regression, K-Nearest Neighbors, Support Vector Machines (SVM - SVC), Naïve Bayes, MLP, Gradient Boosting, AdaBoost, XGBoost, Bagging Classifier and Extra Trees.

To evaluate the models, performance metrics such as Accuracy, Precision, Recall, F1-score were used, and Confusion Matrices, Heatmaps, Learning Curves and Bar Charts were also compared. The results showed that Gradient Boosting, XGBoost and Random Forest outperformed in terms of accuracy and generalization, with an accuracy close to 90%. In contrast, Naïve Bayes had the lowest performance (~58%), indicating limited suitability for the analysis of this data.

Additionally, the results showed that the student-por.csv dataset provided slightly better model performance than student-mat.csv, which may be related to the different distribution of student characteristics.

Finally, suggestions for improvement are hinted at, such as hyperparameter optimization (Grid Search, Bayesian Optimization), incorporating additional features and exploring alternative Machine Learning methods such as Neural Networks.

Keywords: Machine Learning, Student Performance Prediction, Classification, Data Analysis.

Πίνακας περιεχομένων

Λίστα Εικόνων

Λίστα Εξισώσεων

Πίνακας Συντομογραφιών

BN: Bayesian Networks

DT: Decision Tree

FDN: Feedward Dense Network

GB: Gradient Boosting

GNB: Gaussian Naïve Bayes

GRU: Gated Recurrent Unit

KNN: K-Nearest Neighbors

LDA: Linear Discriminant Analysis

LR: Logistic Regression

MLP: Multilayer Perceptron

NB: Naïve Bayes

NN: Neural Networks

RF: Random Forest

RFR: Random Forest Regression

RNN: Recurrent Neural Networks

SGD: Stochastic Gradient Descent

SVC: Support Vector Classifier

SVM: Support Vector Machine

Κεφάλαιο 1ο

1. Εισαγωγή

1.1. Ορισμός της Μηχανικής Μάθησης στην Εκπαίδευση

Το βασικότερο χαρακτηριστικό της ανθρώπινης φύσης, είναι η ικανότητα του ανθρώπου να μαθαίνει και να γίνεται καλύτερος στις καθημερινές του δραστηριότητες, με τη βοήθεια της εμπειρίας. Στην ουσία, όταν γεννιέται ο άνθρωπος, δε γνωρίζει τίποτα. Όμως με το πέρασμα του χρόνου, μέρα με τη μέρα, μαθαίνει ποικιλία πραγμάτων, είτε μέσω άλλων ανθρώπων, είτε μόνος του. Ακριβώς το ίδιο, συμβαίνει και με τους υπολογιστές, ειδάλλως μηχανές. Οι υπολογιστές λοιπόν, συγκεντρώνουν πληθώρα στοιχείων και δεδομένων, ούτως ώστε να είναι σε ετοιμότητα να εξάγουν από μόνοι τους συμπεράσματα [1].

Η Μηχανική Μάθηση (Machine Learning), απαρτίζει σημαντικό πεδίο στην επιστήμη των υπολογιστών, η οποία αναπτύχθηκε κατόπιν της μελέτης υπολογιστικής θεωρίας μάθησης και αναγνώρισης προτύπων στην Τεχνητή Νοημοσύνη – Artificial Intelligence, και ασχολείται με τη δημιουργία αλγορίθμων, που εκπαιδεύονται, χωρίς ωστόσο να έχουν επιδεχτεί προγραμματισμό συγκεκριμένων κανόνων. Πιο συγκεκριμένα, οι συγκεκριμένοι αλγόριθμοι, κάνουν χρήση δεδομένων, για να μπορούν να εφεύρουν σχέσεις και μοτίβα, προκειμένου να λάβουν αποφάσεις, ή να κάνουν προβλέψεις. Οι πρώτες αναφορές για τη Μηχανική Μάθηση, έκαναν την εμφάνισή τους τη δεκαετία του 1960. Η χρήση, όμως, των συγκεκριμένων τεχνικών, αυξήθηκε εκθετικά, κατόπιν της δεκαετίας του 1990. Απότοκο αυτού, ήταν η ανάπτυξη επιστημονικών κλάδων, όπως οι υπερυπολογιστές, η εξόρυξη δεδομένων και η ψηφιοποίηση αρχείων. Σ’ αυτό το σημείο, αξίζει να σημειωθεί, πως αλγόριθμος καλείται ένα σύνολο εντολών, όπου καθορίζει μία διεργασία που κατασκευάστηκε για την επίλυση προβλημάτων ή υπολογιστικών διεργασιών [2]. Ο Άρθουρ Σάμουελ, το 1959, διατύπωσε τον ορισμό της

Μηχανικής Μάθησης ως εξής: «Το πεδίο μελέτης που προσδίδει στους υπολογιστές την ικανότητα του να μαθαίνουν, χωρίς ωστόσο να έχουν ρητά προγραμματιστεί» [3]**.** Όλη αυτή η διαδικασία πραγματώνεται χωρίς καμία ανθρώπινη παρέμβαση. Πιο συγκεκριμένα, η εκπαίδευση των Μηχανικών Αλγορίθμων πραγματοποιείται από παραδείγματα και καταστάσεις, στα οποία αντιλαμβάνονται και εξετάζουν δεδομένα, και στοχεύουν στις μελλοντικές προβλέψεις [1].

Βασικός στόχος της Μηχανικής Μάθησης, είναι η διερεύνηση της μελέτης και κατασκευής των αλγορίθμων, όπου έχουν τη δυνατότητα να εκπαιδεύονται από τα δεδομένα, να μαθαίνουν, και να κάνουν σχετικά με αυτά προβλέψεις. Οι αλγόριθμοι αυτοί εκτελούνται ύστερα από την κατασκευή μοντέλων, και προέρχονται από πειραματικά δεδομένα, ώστε να πραγματοποιούν προβλέψεις που θα βασίζονται στα δεδομένα ή θα διεξάγουν αποφάσεις που θα αποδίδονται ως το αποτέλεσμα.

Η Μηχανική Μάθηση χρησιμοποιείται σε μία σωρεία υπολογιστικών εργασιών, που καθιστά με αυτόν τον τρόπο ανέφικτο το σχεδιασμό και τον προγραμματισμό αλγορίθμων.

Όσον αφορά την ανάλυση δεδομένων, το Machine Learning συγκροτεί τη μέθοδο, που ασχολείται με την εφεύρεση πολύπλοκων αλγορίθμων και μοντέλων, που έχουν σκοπό την πρόβλεψη. Τα δεδομένα αυτά είναι τόσο αναλυτικά, ώστε να μπορούν προσφέρουν σε αναλυτές, μηχανικούς, επιστήμονες δεδομένων και ερευνητές, τη δυνατότητα να εξάγουν αξιόπιστα αποτελέσματα και αποφάσεις, καθώς επίσης και να δώσουν έμφαση σε συσχετίσεις, με τη βοήθεια της μάθησης που προκύπτουν από δεδομένα τάσεων και ιστορικών σχέσεων [3]**.**

Μιας και σήμερα, ο όγκος δεδομένων έχει αυξηθεί εκθετικά, ήταν επιτακτική η ανάγκη να υπάρξουν συστήματα, τα οποία θα είναι ικανά να φέρουν εις πέρας την επεξεργασία των πολύπλοκων δεδομένων.

Τα δεδομένα αυτά, φερόμενα και ως Big Data, είναι πολύπλοκα και μεγάλα, και συχνά είναι διαχειρίσιμα από διάφορα μοντέλα Μηχανικής Μάθησης. Ενδεικτικά ένα τέτοιο μοντέλο Machine Learning, είναι το Deep Learning.

Όπως προαναφέρθηκε, η Μηχανική Μάθηση, συγκροτείται από μηχανές που τροφοδοτούνται με σωρεία δεδομένων, και έχουν στόχο την ανάλυση και εξαγωγή συμπερασμάτων. Κατόπιν, κρατούν όλα τα δεδομένα, ώστε να τα βελτιώσουν, προκειμένου κάθε φορά να παράγονται ακριβέστερα δεδομένα.

Οι περισσότερες εργασίες πλέον, μπορούν να αυτοματοποιηθούν μέσω των μηχανών, γεγονός που προτρέπει έναν όλο και αυξανόμενο αριθμό οργανισμών, να επικεντρωθούν γύρω από το Machine Learning, μαζί με ότι έπεται, με σκοπό ν αυτοματοποιήσουν στο έπακρο τις διεργασίες τους, με μεγαλύτερη ταχύτητα και ακρίβεια, απ' ότι συνέβαινε μέχρι τώρα, που τις υλοποιούσε ο άνθρωπος. Εξάλλου, η επιστήμη των δεδομένων, υιοθετείται μέρα με τη μέρα από τις εταιρείες, σα να είναι μονόδρομος, μιας και έχει εισέλθει σε τεράστιο βαθμό στην καθημερινή ζωή του ανθρώπου [1].

Σε ένα μαθησιακό σύστημα, είναι διαθέσιμα εκπαιδευτικά σήματα και ανατροφοδότηση, τα οποία αναλόγως με τη φύση αυτών Η ταξινόμηση αυτή βασίζεται στα είδη των αντιμετωπιζόμενων προβλημάτων [4] και διακρίνονται σε τρεις κατηγορίες: στην επιτηρούμενη, τη μη επιτηρούμενη και την ενισχυτική μάθηση. Αναλυτικότερα:

i) **Επιτηρούμενη Μάθηση – Supervised Learning, ειδάλλως μάθηση με επίβλεψη ή επιβλεπόμενη μάθηση:** Απαρτίζει ένα πρόγραμμα υπολογισμού το οποίο επιδέχεται τις παραδειγματικές εισόδους, αλλά και τα θεμιτά αποτελέσματα από έναν ειδικό, ώστε να επιτευχθεί ένας και μόνος στόχος: Την εκμάθηση ενός γενικού κανόνα, με σκοπό να αντιστοιχίζονται οι είσοδοι με τα κατάλληλα αποτελέσματα [1]. Αφορά τη διεξαγωγή συμπερασμάτων, με βάση παλιότερα συλλεχθέντα δεδομένα. Εκτενέστερα, φέρει αρκετές ομοιότητες με τον τρόπο που λειτουργεί και συμπεριφέρεται ο άνθρωπος, μιας και χρησιμοποιεί γνώσεις και εμπειρίες που έχει λάβει στο παρελθόν, ούτως ώστε να λάβει σημαντικές αποφάσεις για το παρόν του, και να δημιουργήσει καλύτερες εκβάσεις για το μέλλον του. Ένα χαρακτηριστικό παράδειγμα Επιβλέπουσας Μάθησης, είναι οι εξατομικευμένες προτάσεις

προϊόντων, που η Amazon συστήνει στους χρήστες της, βάσει των προϊόντων που έχουν αγοράσει ή απλά δει, στο παρελθόν [3]. Επίσης, είναι αξιοσημείωτο ότι, συναντάται όταν ο αλγόριθμος παράγει μία συνάρτηση, που αναπαριστά ένα σύνολο εκπαίδευσης, τις λεγόμενες δεδομένες εισόδους -input-, σε γνώριμες αναμενόμενες εξόδους -output-. Αυτού του είδους η μάθηση, στοχεύει στο να γενικεύει τη συγκεκριμένη συνάρτηση, ακόμη και για εισόδους που φέρουν άγνωστη έξοδο. Είναι συνηθισμένο, τα δεδομένα να φέρουν ετικέτες -labels-, που φανερώνουν τη σύνδεση με την έξοδο, οι οποίες έχουν τοποθετηθεί από κάποιον κώδικα, ή ακόμη και από τον ίδιον τον άνθρωπο. Σ' αυτό άλλωστε οφείλεται και η ονομασία της (επιβλεπόμενη). Μπορεί επίσης να χρησιμοποιηθεί και σε προβλήματα:

- ✓ **Διερμηνείας (Interpretation)**
- ✓ **Πρόγνωσης (Prediction)**
- ✓ **Ταξινόμησης (Classification)** [5] [2]

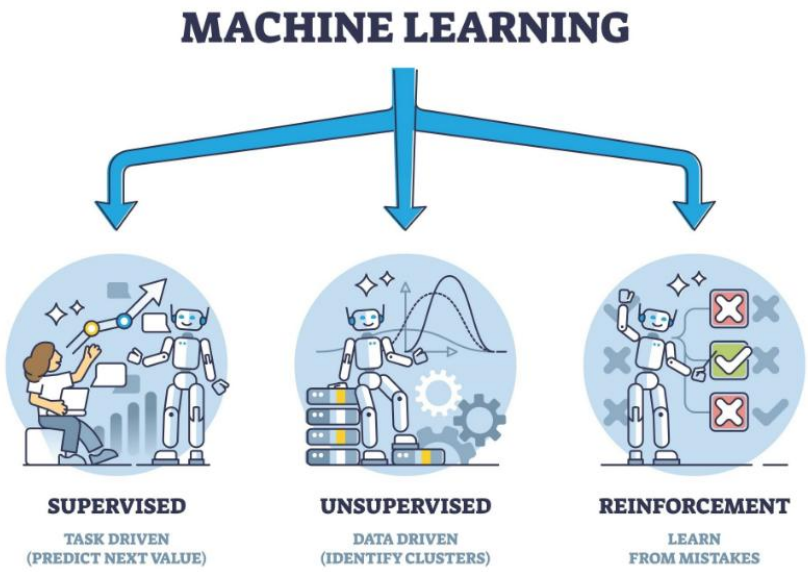
ii) **Μη Επιτηρούμενη Μάθηση – Unsupervised Learning, ειδάλλως μάθηση με χωρίς επίβλεψη ή μη επιβλεπόμενη μάθηση:** Στην προκειμένη περίπτωση, χωρίς να προσδίδεται στον αλγόριθμο μάθησης κάποια εμπειρία, οφείλει να εντοπίσει τη δομή που έχουν τα δεδομένα εισόδου. Το συγκεκριμένο είδος μάθησης μπορεί να είναι είτε αυτοσκοπός, είτε μέσο για ένα τέλος. Δηλαδή, μπορεί είτε να ανακαλύπτει σε δεδομένα κρυμμένα μοτίβα, είτε να είναι χαρακτηριστικό της μάθησης, αντίστοιχα [3]**.** Επίσης, σχετίζεται με τους αλγορίθμους που αποπειρώνται να ανακαλύψουν ποικίλα άγνωστα μοτίβα των δεδομένων τους, χωρίς να είναι ενήμερα για την παρουσία επισημασμένων δεδομένων. Η παρούσα μέθοδος χρησιμοποιείται για τον υπολογισμό πιθανών μεγεθών της αγοράς, ενός νέου προϊόντος, για το οποίο δεν είναι γνωστός μεγάλος αριθμός δεδομένων. Το αποτέλεσμα θα είναι η μηχανή να λειτουργήσει με τα δεδομένα που έχει, με σκοπό να προχωρήσει στην ομαδοποίηση τους σε συστάδες, τα λεγόμενα clusters, και να τα παρουσιάσει οπτικά -K-Means Clustering- [1]. Τέλος, παρατηρείται όταν ο αλγόριθμος δημιουργεί ένα μοντέλο, κάποιων συνόλων εισόδων -input-, με τη μορφή παρατήρησης, χωρίς ωστόσο να είναι γνωστές οι επιθυμητές έξοδοι.

Σ' αυτήν την περίπτωση, ο ίδιος ο υπολογιστής είναι υπεύθυνος να αναγνωρίσει τα προϋπάρχοντα μοτίβα. Προσπαθεί δηλαδή να αναπαράγει, το πως ο άνθρωπος χωρίζει τα πράγματα σε κατηγορίες. Συναντάται κυρίως σε προβλήματα:

- ✓ **Ομαδοποίησης (Clustering)**, όπου τα δεδομένα κατατάσσονται σε ομάδες -clusters-, βάσει των κοινών στοιχείων και πληροφοριών που διαθέτουν.
- ✓ **Ανάλυσης Συσχετισμών (Association Analysis)** [5] [2]

iii) **Ενισχυτική Μάθηση:** Συμβαίνει όταν στον υπολογιστή ένα πρόγραμμα αλληλεπιδρά με ένα δυναμικό περιβάλλον, όταν οφείλεται να υλοποιηθεί ένας ορισμένος στόχος, φερειπείν η οδήγηση ενός αυτοκινήτου. Αυτό συμβαίνει χωρίς την παρουσία κάπου ειδικού, που να υπενθυμίζει άμα πλησιάζει στο στόχο του ή όχι. Κάτι παρόμοιο συμβαίνει με τα παιχνίδια που παίζονται εναντίον αντιπάλου [3]**.** Επιπροσθέτως, Κατασκευάζεται ένα εικονικό περιβάλλον, το οποίο απαρτίζεται από συγκεκριμένους κανόνες, με αποτέλεσμα ο υπολογιστής να αλληλεπιδρά με το περιβάλλον αυτό, έως ότου να επιτευχθεί κάποιος στόχος, όπως για παράδειγμα να μεγιστοποιηθεί κάποιο σκορ. Η ενισχυτική μάθηση, χρησιμοποιείται ωστόσο για την εκμάθηση μερικών παιχνιδιών από τους υπολογιστές, τα οποία οι άνθρωποι δε μπορούν να ανταγωνιστούν. Σημειώνεται σε περιπτώσεις που ο αλγόριθμος ανακαλύπτει κάποιες ενέργειες στρατηγικής, μέσω άμεσων αλληλεπιδράσεων από το περιβάλλον. Χρησιμοποιείται σε προβλήματα:

- ✓ **Σχεδιασμού (Planning)**, στα οποί ενδεικτικά περιλαμβάνονται η βελτιστοποίηση εργασιών εργοστασιακών χώρων, και ο έλεγχος κίνησης ρομπότ [5] [2]



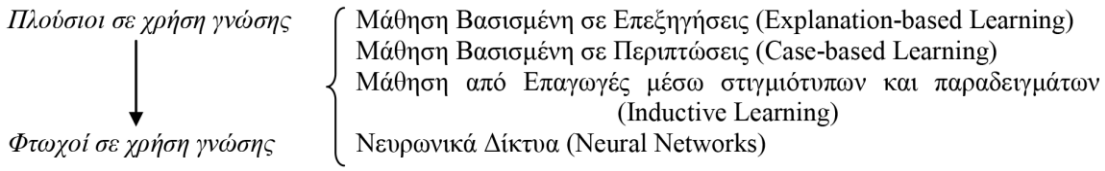
Εικόνα 1: Είδη Μάθησης ML

Η ταξινόμηση συχνά επικαλύπτεται, και είναι αρκετά ασαφής, συνεπώς είναι απαιτητικό μία ορισμένη μέθοδος να κατηγοριοποιηθεί σε μία υποκατηγορία. Τουτέστιν, όπως υποδηλώνει και η ονομασία της, η Ημί-Επιβλεπόμενη Μάθηση, είναι μερικώς Μη Επιβλεπόμενη και μερικώς Επιβλεπόμενη [4]. Γι' αυτό η Ημί-Επιβλεπόμενη Μάθηση, κατατάσσεται ανάμεσα στην Επιτηρούμενη και Μη Μάθηση, στην οποία δίνεται από το δάσκαλο, ένα εκπαιδευτικό σήμα που ωστόσο είναι ελλιπές. Το σήμα αυτό απαρτίζει ένα εκπαιδευτικό σύνολο, όπου αρκετά από τα από τα αποτελέσματα στόχους θα απουσιάζουν. Στη δεδομένη αρχή, τοποθετείται μία ειδική περίπτωση η λεγόμενη Μεταγωγή, στην οποία κατά τη διάρκεια του χρόνου εκμάθησης, απουσιάζει ένα τμήμα των στόχων, με το σύνολο καταστάσεων όμως του προβλήματος, να είναι γνωστά [3].

Για να επιλυθεί οποιοδήποτε πρόβλημα, στον τομέα της Μηχανικής Μάθησης, υφίσταται ο καταλληλότερος τρόπος μάθησης, και για τον αντίστοιχο τρόπο μάθησης, υπάρχει περισσότερους από ένας αλγορίθμους, οι οποίοι είναι κατάλληλοι, ώστε να μπορέσουν να χρησιμοποιηθούν.

Οι αλγόριθμοι Μηχανικής Μάθησης, διαχειρίζονται κατά αποκλειστικότητα τη γνώση, και την αναπαριστούν με διάφορους τρόπους, οι οποίοι είναι μαθηματικοποιημένοι, θεωρούμενοι από

τους συγκεκριμένους αλγορίθμους, κάθε φορά, κατάλληλοι ώστε να την εκφράσουν. Επιπροσθέτως, υπάρχουν συγκεκριμένοι αλγόριθμοι που λαμβάνουν σαν είσοδο αποκλειστικά και μόνο παρατηρήσεις, ενώ υπάρχουν άλλοι οι οποίοι επιδέχονται περισσότερες από μία προϋπάρχουσες γνώσεις. Η κάτωθι εικόνα απεικονίζει την προσπάθεια κατάταξης αλγορίθμων, με βασικό κριτήριο τον τρόπο μάθησης, να βασίζεται λιγότερο ή περισσότερο στην υπάρχουσα γνώση [5].



Εικόνα 2: Προσπάθεια Κατάταξης Αλγορίθμων

Στη Μηχανική Μάθηση, κατατάσσεται μία ακόμη κατηγορία εκμάθησης, η Μετά-Μάθηση (Meta Learning). Η εν λόγω διαδικασία, βοηθά τη μηχανή να μάθει να αναπτύσσει τις δικές της μεθόδους επαγωγής, στηριζόμενες σε προηγούμενη εμπειρία. Επιπροσθέτως, υπάρχει και η Αναπτυξιακή Μάθηση, η φερόμενη και ως Developmental Robotics, που προορίζεται για να υλοποιούν την εκμάθηση τα ρομπότ. Με αυτόν τον τρόπο, παράγεται η δική της ακολουθία από μαθησιακές καταστάσεις, με σκοπό το ρομπότ να κατέχει συσσωρευτικά πολυάριθμες δεξιότητες, χάρις στη βοήθεια αυτόνομης αυτό-εξερεύνησης, καθώς επίσης και κοινωνικής αλληλεπίδρασης, ανάμεσα σε (ανθρώπους) εκπαιδευτές, με τη χρήση μηχανισμών καθοδήγησης, λόγου χάρη μίμηση, ωρίμανση και ενεργητική μάθηση.

Μία ακόμη κατηγορία προβλημάτων Machine Learning, προκύπτει όταν θεωρηθεί επιθυμητό το αποτέλεσμα που εξάγει ένα σύστημα Μηχανικής Μάθησης.

i) **Ταξινόμηση:** Τα δεδομένα εισόδου κατατάσσονται σε τουλάχιστον δύο κλάσεις, με αποτέλεσμα η μηχανή οφείλει να δημιουργήσει ένα μοντέλο, που θα συσχετίζει τα δεδομένα σε τουλάχιστον μια κλάση. Η δεδομένη ταξινόμηση είναι γνωστή και ως Multi-Label Ταξινόμηση, η οποία ωστόσο περιλαμβάνει την Επιτηρούμενη Μάθηση. Τέτοιο παράδειγμα ταξινόμησης, είναι και τα «Φίλτρα Spam», όπου

είσοδοι θεωρούνται μηνύματα και emails, ενώ κλάσεις οι κατηγορίες: «spam» και «no spam».

ii) **Συσταδοποίηση:** Συμβαίνει, όταν ένα σύνολο από εισόδους, έγκειται να ομαδοποιηθεί. Εν αντιθέσει με την ταξινόμηση, στην προκειμένη περίπτωση, οι ομάδες είναι άγνωστες από την αρχή, κάνοντας την εν λόγω διάκριση Τυπική Εργασία μιας Μη Επιτηρούμενης Μάθησης.

iii) **Παλινδρόμηση:** Απαρτίζει ένα ακόμη πρόβλημα της Επιτηρούμενης Μάθησης, και τα αποτελέσματα που προκύπτουν δεν είναι διακριτά, αλλά συνεχή.

iv) **Μείωση Διαστασιμότητας (Dimensionality Reduction):** Σε αυτά του είδους τα προβλήματα, τα δεδομένα περνάνε από μία διαδικασία απλοποίησης και αντιστοίχισης, σε έναν άλλο χώρο, με λιγότερες ωστόσο διαστάσεις. Το “Topic Modeling” -Μοντέλο Θεμάτων-, συγκροτεί ένα σχετικό πρόβλημα, στο οποίο η μηχανή είναι υπεύθυνη για τον εντοπισμό εγγράφων παρόμοιων θεμάτων, σε σύγκριση με άλλα έγγραφα που είναι γραμμένα σε φυσική γλώσσα.

v) **Εκτίμηση Πυκνότητας:** Εντοπίζει και κατανέμει τα δεδομένα εισόδου στο χώρο.

Προσεγγίσεις:

i) **Εκμάθηση με Δέντρο Απόφασης:** Χρησιμοποιείται ένα δέντρο απόφασης, σαν προγνωστικό μοντέλο, όπου υλοποιεί αντιστοιχίσεις παρατηρήσεων σε συμπεράσματα. Παραδείγματος χάριν, η παρατήρηση μπορεί να σχετίζεται με ένα στοιχείο, ενώ το συμπέρασμα να αφορά την τιμή στόχο του συγκεκριμένου αντικειμένου.

ii) **Εκμάθηση με Κανόνες Συσχέτισης:** Συνιστά μία μέθοδο ανακάλυψης ανάμεσα στις μεταβλητές μεγάλων βάσεων δεδομένων.

iii) **Τεχνητά Νευρωνικά Δίκτυα:** Συγκροτεί έναν αλγόριθμο Μάθησης, εμπνευσμένος από τις πτυχές λειτουργίας και τη δομή των Βιολογικών Νευρωνικών Δικτύων. Η δομή αυτών των υπολογισμών, στηρίζεται σε ένα τμήμα που συνδέεται εσωτερικά με τεχνητούς νευρώνες, που επεξεργάζουν την πληροφορία και υλοποιούν υπολογισμούς, με το να επικοινωνούν μεταξύ τους. Τα εκσυγχρονισμένα Νευρωνικά

Δίκτυα κατατάσσονται στα είδη Μη Γραμμικής Στατιστικής Μοντελοποίησης Δεδομένων. Συχνά η χρήση τους επιτυγχάνεται ώστε να μοντελοποιηθούν σύνθετες σχέσεις ανάμεσα σε δεδομένα εξόδου και εισόδου, να εντοπιστούν στατιστικές δομές σε μία μη γνώριμη κοινή κατανομή πιθανοτήτων, ή να ανακαλυφθούν πρότυπα στα δεδομένα, ανάμεσα σε παρατηρούμενες μεταβλητές.

iv) **Βαθιά Μάθηση:** Η συγκεκριμένη ιδέα αναπτύχθηκε εξαιτίας στην ανάπτυξη των GPU, οι οποίοι επιλέγονταν για προσωπική χρήση, καθώς επίσης και στην πτώση τιμών των υλικών των τελευταίων χρόνων. Με τη συγκεκριμένη προσέγγιση επιτυγχάνεται η μοντελοποίηση του τρόπου με τον οποίο ένας ανθρώπινος εγκέφαλος επεξεργάζεται τον ήχο και το φως, μετατρέποντας τα σε ακοή και όραση, αντίστοιχα. Ενδεικτικά μερικές επιτυχημένες εφαρμογές που εντάσσονται στη Βαθιά Μάθηση, είναι η Αναγνώριση Ομιλίας και η Μηχανική Όραση.

v) **Επαγωγικός Λογικός Προγραμματισμός - ILP:** Απαρτίζει μία προσέγγιση που καθορίζει τη μάθηση, χρησιμοποιώντας τον εν λόγω προγραμματισμό, ως το βασικό τρόπο απαρίθμησης γνωστικού υποβάθρου, παραδειγμάτων εισόδου, αλλά και υποθέσεων. Το σύστημα ILP, κατασκευάζει υποτιθέμενα λογικά προγράμματα, που ενέχουν όλα τα θετικά, χωρίς όμως να εμπεριέχονται αρνητικά παραδείγματα. Σ' αυτό οφείλεται η κωδικοποίηση γνωστικού υποβάθρου, καθώς και ενός συνόλου παραδειγμάτων, τα οποία προβάλλονται σαν μία λογική βάση, γεμάτη από γεγονότα. Επιπροσθέτως, ο προγραμματισμός αυτός, είναι ένα παρόμοιο κομμάτι, στο οποίο σημαντικό ρόλο παίζουν κάθε είδος προγραμματιστικής γλώσσας, και όχι μόνο λογικού προγραμματισμού, με σκοπό να αναπαρασταθούν υποθέσεις, όπως είναι για παράδειγμα τα συναρτησιακά προγράμματα.

vi) **Μηχανές Διανυσμάτων Υποστήριξης:** Αποτελούν μία ομάδα μεθόδων Επιτηρούμενης Μάθησης, που στοχεύουν στην παλινδρόμηση και στην ταξινόμηση. Στη συγκεκριμένη περίπτωση, προσδίδεται ένα σύνολο από εκπαιδευτικά παραδείγματα, με κάθε φορά να σημειώνεται ποια από τις κατηγορίες αντιστοιχίζεται στο παράδειγμα. Μία τέτοια μηχανή, δημιουργεί ένα μοντέλο πρόβλεψης, το οποίο εξετάζει σε ποια κατηγορία συμπίπτει το καινούριο παράδειγμα.

vii) **Ομαδοποίηση:** Είναι η προσέγγιση στην οποία ένα σύνολο από παρατηρήσεις κατηγοριοποιείται σε σύνολα, ούτως ώστε οι παρατηρήσεις που εντάσσονται στην ίδια ομάδα -cluster-, να είναι όμοιες, όπως είναι προκαθορισμένες από κάποια κριτήρια, εν αντιθέσει με κάποιες παρατηρήσεις προερχόμενες από εντελώς διαφορετικά υποσύνολα, να χαρακτηρίζονται ως ανόμοιες. Τεχνικές κατηγοριοποίησης, που δεν είναι ίδιες μεταξύ τους, καταλήγουν σε υποθέσεις σχετιζόμενες με τη δομή των δεδομένων, οι οποίες ωστόσο είναι διαφορετικές, και συνήθως ορίζονται από κάποια μέσα ομοιότητας, και η αξιολόγηση τους επιτυγχάνεται, φερειπείν, ως προς την ομοιότητα ανάμεσα στα μέλη του ίδιου του cluster, τη λεγόμενη εσωτερική συνοχή, αλλά και το διαχωρισμό που πραγματώνεται μεταξύ των διαφορετικών ομάδων. Επιπλέον, υπάρχουν και άλλες μέθοδοι, που είναι βασισμένες στη συνεκτικότητα των γραφημάτων και στην εκτιμώμενη πυκνότητα. Τέλος, η ομαδοποίηση συγκροτεί μία μέθοδο Μη Επιτηρούμενης Μάθησης, και μία τεχνική που είναι ευρέως γνωστή στη Στατιστική Ανάλυση Δεδομένων.

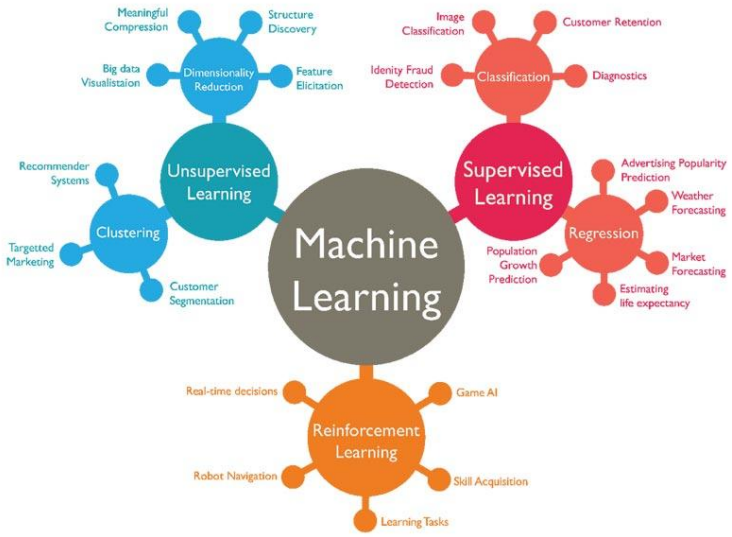
viii) **Δίκτυα Bayes** ή **Άκυκλο Γραφικό Μοντέλο** ή **Δίκτυο Εμπιστοσύνης:** Συνιστά ένα Πιθανοθεωρητικό Γραφικό Μοντέλο, απεικονιζόμενο από μία ομάδα τυχαίων μεταβλητών και τη μεταξύ τους ανεξαρτησία διαμέσου σε ένα Κατευθυνόμενο Άκυκλο Γράφο, η οποία διάμεσος όμως, δεν παύει να είναι υποθετική. Παραδείγματος χάριν, σε ένα τέτοιο δίκτυο αναπαρίστανται η πιθανοθεωρητική σχέση, ανάμεσα σε συμπτώματα και ασθένειες. Έχοντας σα δεδομένο τα συμπτώματα, το δίκτυο έχει τη δυνατότητα υπολογισμού πιθανοτήτων παρουσίας, σε διάφορες ασθένειες.

ix) **Ενισχυτική Μάθηση:** Αντικείμενο της είναι η ενασχόληση με το πως ένας πράκτορας (Υποκείμενο), οφείλει να δράσει σ' ένα περιβάλλον, ούτως ώστε κάποιος ενδεχόμενος ορισμός της μακροπρόθεσμης ανταμοιβής, να μεγιστοποιηθεί. Στόχος των αλγορίθμων Ενισχυτικής Μάθησης, είναι ο εντοπισμός μίας πολιτικής, η οποία θα αντιστοιχεί τις καταστάσεις του περιβάλλοντος, με τις αντίστοιχες ενέργειες που υλοποιεί το Υποκείμενο (πράκτορας), στις συγκεκριμένες καταστάσεις. Κύρια διαφορά της, εν συγκρίσει με τα υπόλοιπα είδη μάθησης, είναι ότι τα σωστά

ζευγάρια δεδομένων εισόδου και εξόδου, δεν έχουν ξαναπαρουσιαστεί, καθώς δεν έχουν επιδιορθωθεί οι βέλτιστες δυνατές ενέργειες.

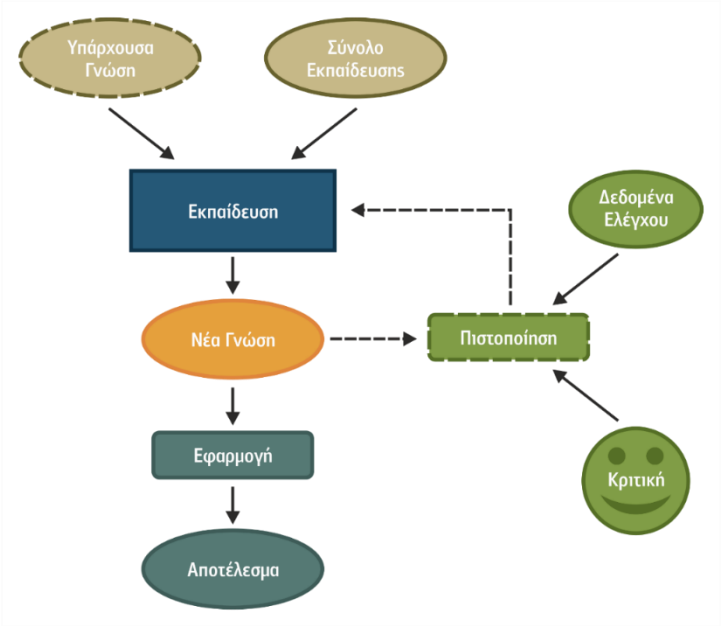
x) **Εκμάθηση με Μέτρο Ομοιότητας:** Στην εν λόγω κατηγορία, η Μηχανή Μάθησης αντλεί ζεύγη από παραδείγματα, τα οποία όμως χαρακτηρίζονται ως όμοια και ανόμοια. Σε αυτή την περίπτωση, η Μηχανή Μάθησης, κρίνεται υπεύθυνη για να εκπαιδευτεί σε μία συνάρτηση μετρικής απόστασης, ή μία συνάρτηση ομοιότητας, η οποία θα προβλέπει αν δύο νέα αντικείμενα που εισάγονται είναι όμοια. Η παρούσα τεχνική είναι γνωστή για τη χρήση της στα Συστήματα Σύστασης.

xi) **Γενετικοί Αλγόριθμοι - GA:** Συγκροτούν μία Αναζήτηση Εύρεσης, όπου αντιγράφει την προσέγγιση της Φυσικής Επιλογής, και οι μέθοδοι που χρησιμοποιούνται είναι για να δημιουργηθούν νέοι γονότυποι, προκειμένου να βρεθούν αποτελεσματικές λύσεις ενός συγκεκριμένου προβλήματος. Τέτοιες μέθοδοι είναι αυτές της διασταύρωσης και της μετάλλαξης. Οι πρώτοι Γενετικοί Αλγόριθμοι, έκαναν την εμφάνισή τους στη Μηχανική Μάθηση, περί τα 1980 και 1990. Τέλος, διάφορες τεχνικές Μηχανικής Μάθησης, χρησιμοποιήθηκαν ώστε να βελτιωθούν οι αποδόσεις των Εξελικτικών και Γενετικών Αλγορίθμων [3].



Εικόνα 3: Κατηγορίες Προβλημάτων ML

Στην παρακάτω εικόνα αναπαρίσταται η αποτύπωση του γενικού τρόπου όπου λειτουργούν οι αλγόριθμοι Μηχανικής Μάθησης. Η εκπαίδευση, απαρτίζει το βασικότερο στάδιο κάθε αλγορίθμου, κατά την οποία ο ίδιος ο αλγόριθμος σαν είσοδο χρησιμοποιεί ένα training set -σύνολο δεδομένων εκπαίδευσης-, προκειμένου να επιτευχθεί ο σκοπός του, που είναι η παραγωγή νέας γνώσης. Επιπρόσθετα, έχει τη δυνατότητα να χρησιμοποιήσει την υπάρχουσα γνώση είτε λιγότερο, είτε περισσότερο, είτε καθόλου.



Εικόνα 4: Φάσεις Μηχανικής Μάθησης

Όπως παρατηρείται, η εκπαίδευση συνοδεύεται από τη φάση της πιστοποίησης, δηλαδή της νέας γνώσης που παράγεται. Συχνά, η πιστοποίηση υλοποιείται αρχικά από τον αλγόριθμο τον ίσιο, με τη βοήθεια του recall -ανάκλησης-, του test data -ελέγχου δεδομένων-, και τέλος, μέσω της κριτικής που επιτυγχάνει ο ίδιος ο χρήστης, λαμβάνοντας υπόψιν τις γνώσεις του, για το πρόβλημα που έχει να λύσει ο αλγόριθμος. Συνοψίζοντας, η καινούρια γνώση που προσδίδεται, είναι σημαντική, προκειμένου να επιλυθούν τα πραγματικά προβλήματα [5].

Τα Μοντέλα Μηχανικής Μάθησης, είναι μαθηματικά πλαίσια ή αλγόριθμοι, τα οποία δίνουν τη δυνατότητα στους υπολογιστές να λάβουν προβλέψεις και αποφάσεις, βάσει δεδομένων. Τα εν λόγω μοντέλα εμβαθύνουν σε μοτίβα, εντός συνόλων δεδομένων, διευκολύνοντας τη βελτίωση της απόδοσης, χωρίς να επαναπρογραμματίζονται ρητά για κάθε εργασία.

Τα Μοντέλα ταξινομούνται σε 3 κατηγορίες, αναλόγως τα δεδομένα εισόδου και τα επιθυμητά αποτελέσματα: Επιβλεπόμενη Μάθηση, Μη Επιβλεπόμενη Μάθηση και Ενισχυτική Μάθηση. Εκμεταλλεύοντας τεράστιες ποσότητες από δεδομένα, τα Μοντέλα Μηχανικής Μάθησης, είναι ικανά να προχωρήσουν στον εντοπισμό τάσεων, ταξινόμηση πληροφοριών, αλλά και στη δημιουργία νέου περιεχομένου, κάνοντας έτσι αναγκαία τα εργαλεία σε κλάδους που ασχολούνται από την Υγειονομική Περίθαλψη έως τα Οικονομικά, και από την Τεχνητή Νοημοσύνη έως το Μάρκετινγκ.

Τουτέστιν, τα Μοντέλα Μηχανικής Μάθησης, συγκροτούν αλγορίθμους, που είναι υπεύθυνοι να παίρνουν αποφάσεις, ή να προβλέπουν, μέσα από δεδομένα που έχουν εκπαιδευτεί. Αυτό έχει σαν αποτέλεσμα, συν τω χρόνω, η βελτίωση απόδοσης, μέσω αναγνώρισης προτύπων.

Πλεονεκτήματα & Μειονεκτήματα Μοντέλων Μηχανικής Μάθησης:

Τα Μοντέλα Μηχανικής Μάθησης εμφανίζουν έναν αρκετά ικανοποιητικό αριθμό πλεονεκτημάτων. Ενδεικτικά, μερικά από αυτά, είναι η ικανότητα ταχείας ανάλυσης, τεραστίου όγκου, δεδομένων, αλλά και εντοπισμού προτύπων, τα οποία ωστόσο δεν καθίστανται ευκρινή στους ανθρώπινους αναλυτές. Επιπροσθέτως, ενισχύουν την ακρίβεια πρόβλεψης διαφόρων εφαρμογών, αυτοματοποιούν τις επαναλαμβανόμενες εργασίες βελτιώνουν, και τέλος βελτιώνουν τις διαδικασίες λήψης αποφάσεων.

Από την άλλη πλευρά, φέρουν έναν μεγάλο αριθμό μειονεκτημάτων. Αναλυτικότερα, είναι πολύ πιθανό να υπάρξουν λανθασμένα αποτελέσματα, εξαιτίας της μεροληψίας των δεδομένων εκπαίδευσης, της πολυπλοκότητας του ερμηνευτικού μοντέλου, καθώς και της απαίτησης σημαντικών υπολογιστικών πόρων. Επιπροσθέτως, η πιθανότητα έλλειψης διαφάνειας, είναι πολύ υψηλή, εξαιτίας της εξάρτησης από τη Μηχανική Μάθηση, κάνοντας έτσι απαιτητική την κατανόηση λήψης αποφάσεων, κάτι που προκαλεί ανησυχίες σε σημαντικούς τομείς (πρακτικές πρόσληψης και ποινική δικαιοσύνη).

Εν κατακλείδι, τα Μοντέλα Μηχανικής Μάθησης, προσφέρουν δραστικά μέσα ώστε να επιτευχθεί η αυτοματοποίηση και η ανάλυση δεδομένων, τα οποία όμως είναι συνοδευόμενα από προκλήσεις και συσχετίζονται με ηθικές επιπτώσεις, ερμηνευτικότητα και προκατάληψη.

Οφέλη των Μοντέλων Μηχανικής Μάθησης

Τα εν λόγω Μοντέλα, είναι ιδιαίτερα ωφέλιμα σε αρκετούς κλάδους, ενδυναμώνοντας έτσι τη διαδικασία λήψης αποφάσεων και την αποτελεσματικότητα. Ένα από τα σημαντικότερα προτερήματα είναι η ικανότητα ανάλυσης τεράστιων ποσοτήτων από δεδομένα, ταχύτατα και με ακρίβεια, φανερώνοντας ιδέες και μοτίβα, όπου θα ήταν δύσκολα ανιχνεύσιμα από τον άνθρωπο. Το συγκεκριμένο, καταλήγει σε αναβαθμισμένα στοιχεία αναλυτικής πρόβλεψης, διευκολύνοντας κατ' αυτόν τον τρόπο, τις επιχειρήσεις να κάνουν πιο εύκολη την πρόβλεψη των τάσεων και τη λήψη τεκμηριωμένων αποφάσεων. Πέρα απ' αυτό, τα Μοντέλα Μηχανικής Μάθησης, αυτοματοποιούν τις περιοδικές εργασίες, ελαττώνοντας έτσι την πιθανότητα ανθρώπινου λάθους, και συνεισφέροντας στους υπαλλήλους ωφέλιμο χρόνο, με σκοπό να εστιάσουν σε πρωτοβουλίες στρατηγικού χαρακτήρα. Συμπληρωματικά, εν καιρώ, αναβαθμίζονται και προσαρμόζονται, όσο παρουσιάζονται νέα δεδομένα, κατοχυρώνοντας έτσι πως οι προβλέψεις εξακολουθούν να είναι ακριβείς και επαρκείς. Γενικά, όταν τα Μοντέλα Μηχανικής Μάθησης ενσωματωθούν, μπορούν να οδηγήσουν σε καινοτομία, να ενισχύσουν την παραγωγικότητα και να προσφέρουν ανταγωνιστικότητα στα βασιζόμενα δεδομένα.

Συνοψίζοντας, τα Μοντέλα Μηχανικής Μάθησης οδηγούν στην καινοτομία και παρέχουν ανταγωνιστικό πλεονέκτημα. Αυτό επιτυγχάνεται με το να στηρίζουν την αποτελεσματικότητα, με την ανάλυση τεράστιων συνόλων δεδομένων με γρήγορη σχετικά ταχύτητα, την αναβάθμιση πρόβλεψης λεπτομερών στοιχείων, την αυτοματοποίηση εργασιών, ώστε να επιτευχθεί ο περιορισμός σφαλμάτων, και την προσαρμογή της διατήρησης της ακρίβειας, με την πάροδο του καιρού.

Τα Μοντέλα Μηχανικής Μάθησης, έρχονται αντιμέτωπα με πληθώρα προκλήσεων. Απότοκο αυτού είναι η αλλαγή στην αξιοπιστία και στην απόδοση. Η ποσότητα και η ποιότητα των δεδομένων, συγκροτούν τη σημαντικότερη πρόκληση. Τα Μοντέλα χρειάζονται οπωσδήποτε τεράστια και ποικίλα datasets, με σκοπό η μάθηση να είναι αποτελεσματική, μιας και μεροληπτικά δεδομένα ή δεδομένα κάκιστης ποιότητας, είναι πολύ πιθανό να καταλήξουν σε προβλέψεις που είναι μη έγκυρες. Επίσης, συναντάται εύκολα το φαινόμενο της υπερπροσαρμογής. Σε αυτή την περίπτωση, το Μοντέλο ναι μεν εκπαιδεύεται εξαιρετικά στα

εκπαιδευτικά δεδομένα, όμως αποτυγχάνει να ερμηνευτεί αποδοτικά σε καινούρια δεδομένα. Επιπλέον, ένα άλλο ζήτημα που απασχολούν τα Μοντέλα Μηχανικής Μάθησης, είναι η ερμηνευτικότητα. Η ερμηνευτικότητα απασχολεί ιδιαίτερα πολυπλοκότερα Μοντέλα, όπως είναι λόγου χάρη τα Βαθιά Νευρωνικά Δίκτυα, τα οποία παίζουν τον ρόλο «μαύρων κουτιών», κάνοντας έτσι ακατόρθωτο για το χρήστη να κατανοήσει πως να λαμβάνει αποφάσεις. Ακόμη, ένας παράγοντας που ενδεχομένως να είναι περιοριστικός, είναι οι υπολογιστικοί πόροι, μιας και η εκπαίδευση των εξελιγμένων μοντέλων, χρειάζεται αρκετό χρόνο και επεξεργαστική ισχύ. Τελευταίο αλλά εξίσου σημαντικό, είναι ο ηθικός παράγοντας, στον οποίο συμπεριλαμβάνονται οι ανησυχίες για την αλγοριθμική μεροληψία και το απόρρητο. Είναι αναγκαίο να αναφερθεί ότι εφαρμόζονται σπουδαίες προκλήσεις στη δημιουργία λύσεων Μηχανικής Μάθησης.

Ανακεφαλαιώνοντας, τα Μοντέλα Μηχανικής Μάθησης έρχονται αντιμέτωπα με ποικίλες προκλήσεις. Συνοπτικά, μερικές από αυτές είναι οι υψηλές υπολογιστικές απαιτήσεις, η έλλειψη ερμηνείας, η υπερπροσαρμογή, η ποσότητα και η ποιότητα δεδομένων, καθώς και οι ηθικές ανησυχίες, φερειπείν διάφορα ζητήματα απορρήτου και μεροληψία. Όλες αυτές οι προκλήσεις είναι ικανές ώστε να αναστείλουν την υπεύθυνη ανάπτυξης και την αποτελεσματικότητά τους [4].

1.2. Σκοπός Μελέτης

Η ακαδημαϊκή επίδοση επικεντρώνεται στο αποτέλεσμα που φέρει η διαδικασία της εκπαίδευσης, και ειδικότερα στο βαθμό που ένα εκπαιδευτικό ίδρυμα, ένας δάσκαλος ή ένας φοιτητής / μαθητής, έχει κατορθώσει σε κάποιον εκπαιδευτικό στόχο. Η επιρροή που έχουν οι σπουδαστές σε συγκεκριμένα μαθήματα, δύναται να αποδοθούν αν ληφθεί υπόψιν μία σωρεία από παραμέτρους, όπως είναι λόγου χάρη η συμμετοχή των φοιτητών σε ατομικές και ομαδικές δραστηριότητες και εργασίες, η βαθμολογία τους σε διαγωνίσματα και τεστ, καθώς επίσης και οι συχνές του συμμετοχές στο εκάστοτε μάθημα.

Συνεπώς, η συνολική επίδοση που παρουσιάζει ένας σπουδαστής φαίνεται στη συνολική βαθμολογία που του παραδίδεται σε κάθε μάθημα μεμονωμένα, αλλά και στο συνολικό μέσο όρο όλων των μαθημάτων [6].

Τα Μοντέλα πρόβλεψης, φερόμενα και ως υπολογιστικά ή μαθηματικά εργαλεία, είναι υπεύθυνα για την πρόβλεψη μελλοντικών καταστάσεων ή αποτελεσμάτων, και βασίζονται σε τάσεις και ιστορικά δεδομένα. Όσον αφορά στον κλάδο της Μηχανικής Μάθησης, τα μοντέλα Πρόβλεψης, συγκροτούνται από αλγορίθμους, οι οποίοι είναι εκπαιδευμένοι στην αναγνώριση μοτίβων στα δεδομένα, και στην αναγνώριση προβλέψεων σε δεδομένα που διαβάζουν για πρώτη φορά.

i) **Εισαγωγή Δεδομένων:** Τα Μοντέλα αντλούν τα δεδομένα εισόδου τους, είτε αυτά είναι μεταβλητές, είτε χαρακτηριστικά, από ένα σύνολο δεδομένων (dataset).
ii) **Εκπαίδευση:** Τα Μοντέλα υλοποιούν την εκπαίδευσή τους, σε ένα υποσύνολο από δεδομένα, με στόχο την αναγνώριση σχέσεων, ανάμεσα σε χαρακτηριστικά και στόχους. Με λίγα λόγια, τι μπορεί να προβλεφθεί και τι όχι.
iii) **Πρόβλεψη:** Μετά το πέρας της εκπαίδευσης, το Μοντέλο είναι ικανό να προχωρήσει στην πρόβλεψη νέων δεδομένων.
iv) **Αξιολόγηση:** Η απόδοση του κάθε Μοντέλου αποτιμάται από διάφορες μετρικές, παραδείγματος χάριν recall, precision, F1-Score, accuracy, και ούτω καθεξής.

Τα Μοντέλα Πρόβλεψης χωρίζονται σε τρεις διαφορετικές κατηγορίες: στα Απλά Μοντέλα, στα Σύνθετα Μοντέλα, αλλά και στα Μοντέλα Χρονοσειρών. Αναλυτικότερα:

i) **Απλά Μοντέλα:** Συνήθως είναι τα πρώτα μοντέλα που χρησιμοποιούνται για τη λύση ενός προβλήματος. Αυτό συμβαίνει διότι είναι ευκολότερα ως προς την κατανόηση, την ερμηνεία και την εφαρμογή διαφόρων προβλημάτων. Μερικά είδη Απλών Μοντέλων είναι η Γραμμική Παλινδρόμηση και το Logistic Regression. Πιο συγκεκριμένα:
    - ✓ **Linear Regression – Γραμμική Παλινδρόμηση:** Συνιστά το βασικότερο και το πιο κατανοητό μοντέλο, το οποίο χρησιμοποιείται για την πρόβλεψη συνεχώς τιμών. Παραδείγματος χάριν τέτοιες συνεχείς τιμές είναι οι τιμές των σπιτιών και τα έσοδα. Αυτό που χαρακτηρίζει τη Γραμμική

Παλινδρόμηση είναι το πόσο απλό, γρήγορο και εύκολα ερμηνεύσιμο είναι σα μοντέλο. Εντούτοις, ο σημαντικότερος περιορισμός του είναι η δυσκολία του στην απόδοση σε πολύπλοκα δεδομένα, τα οποία ωστόσο δεν είναι γραμμικά συσχετισμένα.

- ✓ **Logistic Regression:** Μοντέλο που συγκροτείται αποκλειστικά σε ταξινομήσεις -Binary Classification-, όπως λόγου χάρη ένας μαθητής θα επιτύχει ή όχι σε ένα μάθημα. Το Διακρίνεται από την ταχύτητα, την εύκολη ερμηνεία και την αποδοτικότητα του σε μικρά σύνολα δεδομένων. Όμως, είναι σημαντικό να τονιστεί ότι η απόδοση του είναι αρκετά χαμηλή σε μη γραμμικά και πολύπλοκα δεδομένα.

ii) **Σύνθετα Μοντέλα:** Προορίζονται για πολύπλοκες και δυσκολότερες προβλέψεις. Στη συγκεκριμένη κατηγορία, κάνουν την εμφάνισή τους πιο εξελιγμένοι αλγόριθμοι, οι οποίοι είναι ικανοί να αποκομίσουν περίπλοκες σχέσεις ανάμεσα στα δεδομένα. Στη συγκεκριμένη κατηγορία, μπορεί εύκολα να συναντηθούν τα Decision Tree, Random Forest, Gradient Boosting Models, Support Vector Machines, καθώς και Neural Networks. Ειδικότερα:

- ✓ **Decision Trees:** Απαρτίζουν τα μοντέλα, όπου παράγουν ένα δέντρο αποφάσεων. Στο συγκεκριμένο δέντρο, σε κάθε κόμβο του, υλοποιείται μία ερώτηση που σχετίζεται με τα δεδομένα, και με τη σειρά τους, τα δεδομένα αυτά διαχωρίζονται με βάση τις απαντήσεις που λαμβάνονται. Χαρακτηρίζονται από την εύκολη ερμηνεία τους, αλλά και από την επιρρέπεια στο overfitting. Με άλλα λόγια, overfitting είναι διαδικασία στην οποία το μοντέλο προχωρά στην υπερβολική εκμάθηση του συνόλου δεδομένων, με αποτέλεσμα να μη μπορεί να αποδώσει αποτελεσματικά σε νέα δεδομένα.
- ✓ **Random Forest:** Απαρτίζει ένα σύνολο που αποτελείται από πολλά Δέντρα Απόφασης, με την τελική πρόβλεψη να επιτυγχάνεται κατόπιν ψηφοφορίας των αποτελεσμάτων των δέντρων. Είναι γνωστό για τη βελτιωμένη γενίκευση και το μειωμένο overfitting του. Όμως, αξίζει να σημειωθεί πως, συγκριτικά με τα απλά Δέντρα Απόφασης, είναι το λιγότερο ανιχνεύσιμο.

- ✓ **Gradient Boosting Models:** Συγκροτούν εξελιγμένες εκδόσεις των Δέντρων Αποφάσεων, με στόχο την εκμάθηση λαθών από προηγούμενα μοντέλα, ώστε να βελτιωθεί η ακρίβειά τους. Είναι αξιοσημείωτο, πως χαρακτηρίζεται από υψηλή ακρίβεια σε αρκετά σύνολα δεδομένων, με περισσότερο απαιτούμενο χρόνο για την εκπαίδευση.
- ✓ **Support Vector Machines (SVMs):** Σ' αυτή την κατηγορία ανήκουν τα Μοντέλα τα οποία προσπαθούν να εντοπίσουν το καλύτερο επίπεδο ή τη γραμμή, και κατατάσσει τα δεδομένα σε κατηγορίες. Βασικό τους προτέρημα είναι η καλή τους απόδοση σε δεδομένα που φέρουν ξεκάθαρα όρια, ενώ δεν ευδοκιμούν σε σχετικά μεγάλα σύνολα δεδομένων.
- ✓ **Neural Networks:** Τα Νευρωνικά Δίκτυα, είναι Μοντέλα που φέρουν αρκετές ομοιότητες με τη δομή ενός ανθρώπινου εγκεφάλου. Ο λόγος ύπαρξής τους οφείλεται στη λύση σύνθετων προβλημάτων, όπως είναι η φυσική γλώσσα και οι εικόνες. Χαρακτηρίζονται κυρίως από την ικανότητα τους να διαχειρίζονται τεράστια πολυπλοκότητα και όγκο δεδομένων. Ωστόσο, τα Νευρωνικά Δίκτυα, απαιτούν περισσότερο χρόνο και υπολογιστική ισχύ, ώστε να εκπαιδευτούν.

iii) Μοντέλα Χρονοσειρών: Τα εν λόγω Μοντέλα, είναι προορισμένα ώστε να αναλύουν δεδομένα που σχετίζονται με το χρόνο. Είναι τα καταλληλότερα για προβλέψεις βάσει των χρονικών εξελίξεων. Ενδεικτικά παραδείγματα Μοντέλων Χρονοσειρών είναι το ARIMA και το LSTM. Για την ακρίβεια:

- ✓ **ARIMA:** Αποτελεί ένα στατιστικό Μοντέλο, το οποίο στηρίζεται σε κινούμενους μέσους όρους και αυτοπαλινδρόμηση, με σκοπό να προβλέπει δεδομένα με βάση τη χρονική εξάρτηση. Αξίζει να σημειωθεί πως ενδείκνυται για γραμμικές χρονοσειρές, παρόλο που δεν αποδίδει αποτελεσματικά σε χρονοσειρές που δεν είναι γραμμικές.
- ✓ **Long Short-Term Memory (LSTM):** Συγκροτεί έναν τύπο Νευρωνικού Δικτύου, σχεδιασμένο για να κατανοεί μακροχρόνιες εξαρτήσεις στα δεδομένα. Αξίζει να επισημανθεί ότι έχει τη δυνατότητα να συγκρατεί

σημαντικές πληροφορίες σε τεράστιες ακολουθίες δεδομένων, μολονότι ο απαιτούμενος χρόνος και η υπολογιστική ισχύ, είναι περισσότερα.

Εκείνο που έχει μεγαλύτερη βαρύτητα είναι ότι τα Μοντέλα Πρόβλεψης, μπορούν να εφαρμοστούν σε αρκετούς τομείς της καθημερινότητας. Πιο συγκεκριμένα, συναντώνται στην εκπαίδευση όπου προβλέπεται η απόδοση των μαθητών, εντοπίζονται οι μαθητές που υπόκεινται σε κίνδυνο αποτυχίας, εξατομικεύεται το εκπαιδευτικό περιεχόμενο, καθώς και σχεδιάζονται στρατηγικές βελτιώσεις. Έπειτα, ένας ακόμη κλάδος που χρησιμοποιούνται τα Μοντέλα Πρόβλεψης, είναι η Ιατρική. Για την ακρίβεια, χρησιμοποιούνται ούτως ώστε να προβλέπονται οι διαγνώσεις ασθενειών, να υπολογίζονται οι πιθανότητες επιπλοκών διάφορων χειρουργικών επεμβάσεων, να εξατομικεύονται θεραπείες, καθώς να παρακολουθούνται ασθενείς με τη βοήθεια φορετών συσκευών (wearables). Επιπροσθέτως, παρατηρούνται και στα Οικονομικά όπου προβλέπονται χρηματιστηριακές τιμές, αναλύονται πιστωτικοί κίνδυνοι, εκτιμώνται αποδόσεις επενδύσεων και παρακολουθούνται οικονομικές τάσεις σε πραγματικό χρόνο. Επιπλέον, η παρουσία των Μοντέλων Πρόβλεψης στην ασφάλεια είναι απαραίτητη μιας και μπορούν να προβλέψουν κυβερνοεπιθέσεις, να ανιχνεύσουν απάτες και να αξιολογήσουν τους κινδύνους σε πραγματικό χρόνο. Πέρα από αυτούς τους τομείς, μπορούν να συναντηθούν και σε άλλους τομείς, όπως είναι το Marketing, η Βιομηχανία, οι Μεταφορές, η Ενέργεια και η Λιανική Πώληση.

Η χρήση των Μοντέλων Πρόβλεψης είναι αρκετά σημαντική, μιας και είναι ικανά να αυξήσουν την αποτελεσματικότητα, να επιτρέψουν τη λήψη αποφάσεων βάσει δεδομένων, αλλά και να μειώσουν την αβεβαιότητα διαφόρων τομέων. Παραδείγματος χάριν, σε εκπαιδευτικά πλαίσια, έχουν τη δυνατότητα να προσφέρουν βοήθεια σε συστήματα και καθηγητές, με σκοπό την κατανόηση του πότε ένας μαθητής απαιτεί επιπλέον υποστήριξη.

Επίσης, ενδυναμώνουν την έγκαιρη παρέμβαση, αποτρέποντας έτσι αρνητικές επιπτώσεις, όπως λόγου χάρη η αποχώρηση από το εκπαιδευτικό σύστημα ή τη σχολική αποτυχία. Είναι

αξιοσημείωτο, όπως αναφέρθηκε και παραπάνω, σε άλλους κλάδους -Ιατρική-, σώζουν ζωές, μέσα από έγκαιρες διαγνώσεις και βελτιώσεων θεραπευτικών στρατηγικών.

Εμφανής είναι ακόμη και η σημασία τους σε καταστάσεις οικονομικών αποφάσεων, με το να ελαχιστοποιούν κινδύνους επενδύσεων, ή στη βιομηχανία, με το να βελτιστοποιούνται οι διαδικασίες παραγωγής και η μείωση του κόστους συντήρησης του εξοπλισμού. Τέλος, με τη συνεχόμενη ανάλυση των δεδομένων, προσφέρεται η δυνατότητα προσαρμογής σε περιβάλλοντα τα οποία είναι δυναμικά, κατοχυρώνοντας τη βέλτιστη δυνατή απόδοση.

Σκοπός Μελέτης

Η Μηχανική Μάθηση συγκροτεί ένα βασικό εργαλείο, και στοχεύουν στην ανάλυση δεδομένων, κυρίως στον εκπαιδευτικό κλάδο, για την αξιοποίηση και καλύτερη κατανόηση των πληροφοριών, τα οποία προσκομίζονται απ' αυτά. Η δεδομένη εργασία στοχεύει στα κάτωθι:

- **i)** Την εξέταση δυνατοτήτων Μηχανικής Μάθησης για την ανάλυση δεδομένων μαθητών, δίνοντας έμφαση στις προβλέψεις των αποδόσεων τους.
- **ii)** Την ανάπτυξη και αξιολόγηση Μοντέλων Πρόβλεψης, τα οποία εμπεριέχουν ακριβείς αξιολογήσεις των ακαδημαϊκών πορειών των φοιτητών, τα οποία είναι βασιζόμενα σε χαρακτηριστικά και ιστορικά δεδομένα, όπως είναι οι συμμετοχές τους στα μαθήματα, οι βαθμολογίες τους, η κοινωνικοοικονομική κατάσταση, καθώς και άλλοι παράγοντες.
- **iii)** Ο εντοπισμός κύριων παραγόντων που επιδρούν στην απόδοση των φοιτητών, περιλαμβάνοντας τόσο τις προσωπικές, όσο και τις ακαδημαϊκές πτυχές, φερειπείν οι συνήθειες μελέτης ή η πρόσβαση σε πόρους εκπαιδευτικού περιεχομένου.

Η εφαρμογή των συγκεκριμένων μοντέλων παρέχει σπουδαία υποστήριξη σε ιδρύματα και εκπαιδευτικούς, συνεισφέροντας:

- **i)** Στην αναβάθμιση των συνολικών εκπαιδευτικών εμπειριών, διευκολύνοντας την έγκαιρη παρέμβαση ώστε να αποφευχθούν οι ακαδημαϊκές αποτυχίες.

ii) Στο σχεδιασμό προσαρμοσμένων παρεμβάσεων, όπως είναι η προσαρμογή των εκπαιδευτικών διαδικασιών στις ανάγκες του εκάστοτε φοιτητή.

iii) Στη λήψη αποφάσεων, χάρις στην αναγνώριση φοιτητών οι οποίοι ωστόσο έχουν ανάγκη από άμεση βοήθεια.

Η παρούσα εργασία δεν επικεντρώνεται μόνο στην ακρίβεια των Μοντέλων Πρόβλεψης, καθώς επίσης και στο πόσο χρήσιμα είναι τα αποτελέσματα που προκύπτουν, προκειμένου να προσαρμόζονται σε νέες στρατηγικές υποστήριξης και διδασκαλίας. Ακόμη, τονίζεται η σημασία των δεδομένων που χρησιμοποιούνται, εξασφαλίζοντας κατ' αυτόν τον τρόπο πως η προσέγγιση πληροί τις προϋποθέσεις για ποικίλα εκπαιδευτικά περιβάλλοντα.

Κεφάλαιο 2ο

2. Literature Review

2.1. Μηχανική Μάθηση στην Εκπαίδευση

1) Εισαγωγή

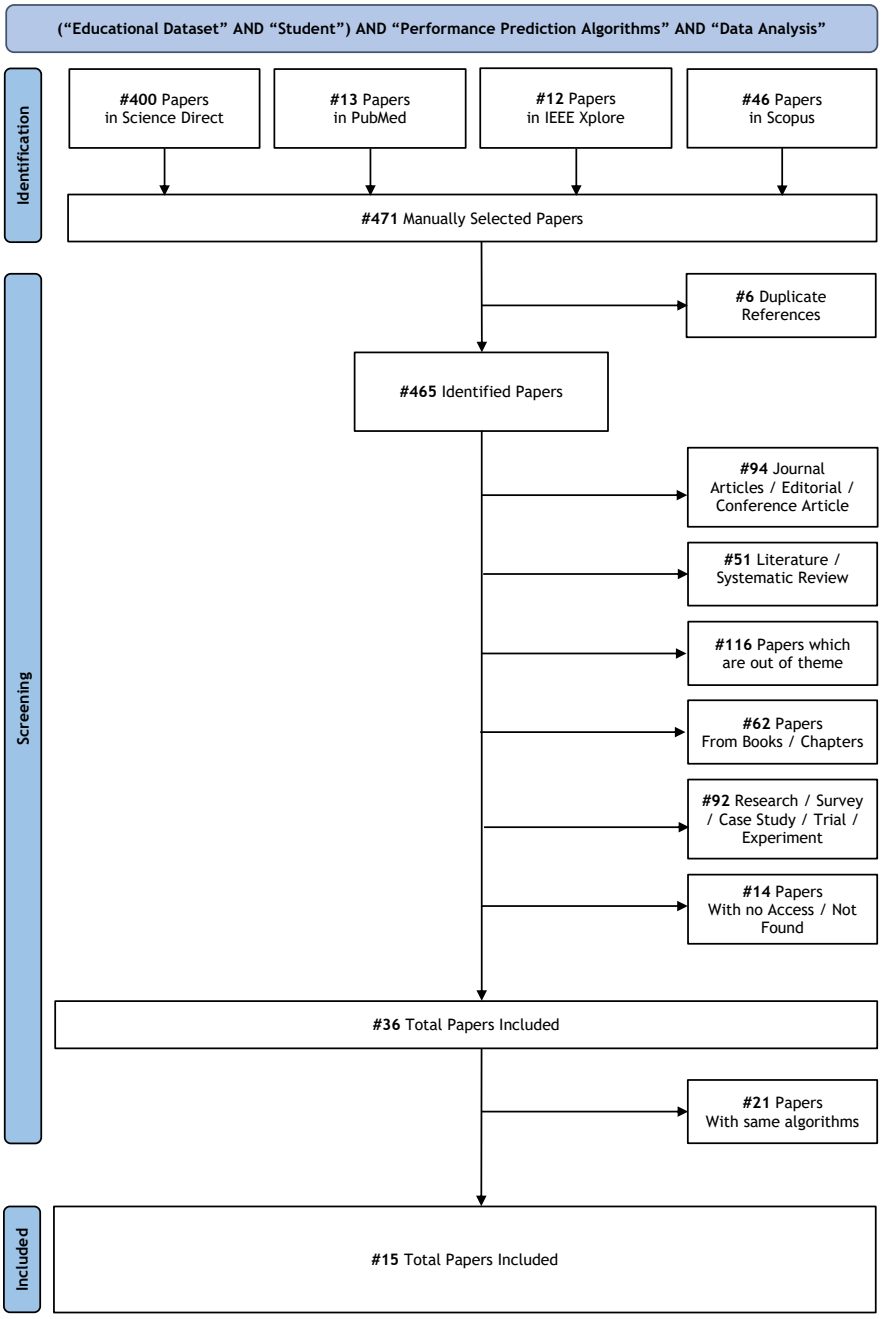
Στο πλαίσιο της διπλωματικής εργασίας, πραγματοποιήθηκε συστηματική ανασκόπηση της βιβλιογραφίας ακολουθώντας τη μεθοδολογία PRISMA, με στόχο την κατανόηση και αξιολόγηση αλγορίθμων Μηχανικής Μάθησης (ML) για την πρόβλεψη της απόδοσης φοιτητών. Συλλέχθηκαν συνολικά 471 papers από βάσεις δεδομένων όπως IEEE Xplore, PubMed, Science Direct και Scopus, εκ των οποίων επιλέχθηκαν τελικά 15 papers βάσει καθορισμένων κριτηρίων επιλογής.

2) Μεθοδολογία

Τα papers καλύπτουν χρονική περίοδο από το 2014 έως το 2024 και εστιάζουν σε τεχνικές ML, οι οποίες εφαρμόζονται σε εκπαιδευτικά δεδομένα. Τα μοντέλα που χρησιμοποιήθηκαν περιλαμβάνουν τόσο κλασικούς ταξινομητές όσο και υβριδικές προσεγγίσεις, με στόχο τη βελτίωση της ακρίβειας πρόβλεψης. Στις συχνότερα χρησιμοποιούμενες μεθόδους συγκαταλέγονται:

- ✓ **Random Forest - RF**
- ✓ **Support Vector Machine - SVM**
- ✓ **Logistic Regression - LR**
- ✓ **K-Nearest Neighbors – KNN**
- ✓ **Decision Tree - DT**
- ✓ **Naïve Bayes - NB**
- ✓ **Gradient Boosting - GB**
- ✓ **Ensemble Learning - Bagging, Stacking**
- ✓ **Neural Networks – NN**

Επιπλέον, σε ορισμένες περιπτώσεις εφαρμόστηκαν τεχνικές μείωσης διαστάσεων (PCA) και εξισορρόπησης δεδομένων (oversampling) για τη βελτίωση της απόδοσης των μοντέλων.



Εικόνα 5: PRISMA

Κάτωθι, παρατίθεται ο πίνακας της βιβλιογραφικής ανασκόπησης, με τις μελέτες, που έχουν ληφθεί υπόψιν.

AUTHOR	YEAR	METHOD	RESULT
Munde, et al. [7]	2024	Ανάλυση παραγόντων που επηρεάζουν την αγορά (βιωσιμότητα, δημογραφικά στοιχεία, κοκ). Χρησιμοποιήθηκαν 8 αλγόριθμοι Μηχανικής Μάθησης: NB, SVM, LDA, KNN, DT, RF, AdaBoost & Bagging Classifier. Οι αξιολογήσεις απόδοσης υπολογίστηκαν με μετρικές, όπως Confusion Matrix, Ακρίβεια, F1-Score, recall & precision.	Οι αλγόριθμοι RF, DT, LDA & Bagging Classifier, εμφάνισαν 100% ακρίβεια στα train set, σε αντίθεση με τα test set, όπου η ακρίβεια του RF έφτασε στα 53%. Ο RF αναδείχθηκε ως ο αποτελεσματικότερος αλγόριθμος, χάρις στο Ensemble Learning.
Kanchon, et al. [8]	2024	Παρουσίαση Μεθόδου ενίσχυσης εξατομικευμένης μάθησης, για την ανίχνευση διαφόρων στυλ μάθησης. Η μελέτη περιλαμβάνει τη δημιουργία ενός προσαρμοσμένου dataset από φοιτητές του Πανεπιστημίου του Μπαγκλαντές, αντλώντας δεδομένα από το Moodle LMS. Για την ταξινόμηση του dataset χρησιμοποιήθηκαν 5 αλγόριθμοι: DT, RF, SVM, XGBoost & LR.	Η τεχνική με την ανάπτυξη του αλγορίθμου XGBoost, επέφερε καλύτερα και πιο έγκυρα αποτελέσματα στις προβλέψεις των μετά-μαθητών, συγκριτικά με τους υπόλοιπους αλγορίθμους.
K, et al. [9]	2024	Χρήση υβριδικού μοντέλου που συνδυάζει τεχνικές Reinforcement Learning, για την πρόβλεψη απόδοσης μαθητών.	Λόγω του ότι τα συμβατικά μοντέλα ML παρουσιάζουν ελλείψεις στη διαφάνεια και στην ερμηνεία των αποφάσεων, προστέθηκε το

		Χρησιμοποιήθηκαν 6 μοντέλα πρόβλεψης Machine Learning: LR, MLP, NB, SVM, KNN & RF.	μοντέλο RLCHI, το οποίο απέδωσε σημαντικά καλύτερα, από ένα μοντέλο RL, που βασίζεται αποκλειστικά σε Μηχανική Μάθηση.
Al-Tameemi, et al. [10]	2024	Πρόβλεψη ακαδημαϊκής απόδοσης των μαθητών, χρησιμοποιώντας ένα υβριδικό συνδυασμό επιβλεπόμενων & μη, τεχνικών Μηχανικής Μάθησης, των Datasets NAU & OULAD. Χρησιμοποιήθηκε η μέθοδος PCA για τη μείωση της πολυπλοκότητας των δεδομένων, και τη βελτίωση της οπτικοποίησης. Οι 6 ταξινομητές επιβλεπόμενης μάθησης που χρησιμοποιήθηκαν είναι οι: KNN, SVM, NB, DT, RF & FDN. Ο κάθε ταξινομητής εκπαιδεύτηκε ξεχωριστά για κάθε cluster.	Το paper αναφέρει πως το μοντέλο επέφερε τη μεγαλύτερη ακρίβεια στην πρόβλεψη απόδοσης των μαθητών, χρησιμοποιώντας τα clustered datasets, είναι το FDN με 95,3%. Μετά ακολούθησε το DT με 93,7%, το RF με 93,4%. Κατόπιν, το SVM με 90,9%, το KNN με 88,7%, ενώ τέλος το NB, με τη μικρότερη ακρίβεια ύψους 83,4%.
Paneru, et al. [11]	2024	Τεχνικές Μηχανικής Μάθησης & Ανάπτυξη Εφαρμογών Πρόβλεψης, ώστε να ενισχυθεί η βιωσιμότητα κύκλων φόρτισης μπαταριών στη μετακίνηση ηλεκτρικών οχημάτων. Για το λόγο	Τα αποτελέσματα έδειξαν ότι ο αλγόριθμος RFR επέφερε καλύτερα αποτελέσματα, με τη μεγαλύτερη απόδοση και στα train και στα test set, εν αντιθέσει με τους αλγορίθμους CatBoost, SVM & GRU.

		αυτό, υλοποιήθηκαν 4 αλγόριθμοι: RFR, SVM, CatBoost & GRU.	
Verma, et al. [12]	2024	Ανάλυση Δεδομένων για την κατανόηση διαφόρων χαρακτηριστικών υβριδικών μαθησιακών περιβαλλόντων, που επηρεάζουν την ευημερία των φοιτητών. Για τη συγκεκριμένη υλοποίηση εφαρμόστηκαν 3 αλγόριθμοι Μηχανικής Μάθησης: RF, XGBoost & LR.	Ο αλγόριθμος RF, εμφάνισε ακρίβεια 88%, καθιστώντας τον ως τον αποτελεσματικότερο αλγόριθμο της υλοποίησης. Ο αλγόριθμος με τη χειρότερη απόδοση αποδείχθηκε ο LR με 79%, ενώ ο XGBoost παρουσίασε ακρίβεια 85%.
Daza, et al. [13]	2023	Χρήση μεθόδου εξισορρόπησης δεδομένων & τεχνικών Ensemble Learning, ώστε να εντοπιστούν και αξιολογηθούν τα επίπεδα άγχους. Η τεχνική που υλοποιήθηκε είναι η Ensemble Stacking, με τη βοήθεια 4 διαφορετικών διαμορφώσεων: Stacking 1A, 2A, 3A, 4A, με τη χρήση 6 αλγορίθμων: BN, KNN, DT, GB, LR & SVM. Για την αξιολόγηση των μοντέλων έγινε χρήση του cross-validation, ενώ για την εξισορρόπηση των δεδομένων το oversampling.	Όσων αφορά τις επιδόσεις των αλγορίθμων, χωρίς εξισορρόπηση δεδομένων, περισσότερη ακρίβεια παρουσίασε ο SVM με 96,49%. Ύστερα ακολούθησαν ο GB (89,47%), ο BN (87,72%), ο KNN (80,70%) και τέλος ο DT (73,68%).

Pratama, et al. [14]	2023	Σύγκριση 11 paper που σχετίζονται με την ψυχική υγεία, με τη βοήθεια αλγορίθμων Machine Learning. Οι αλγόριθμοι αυτοί είναι: SVM, KNN, NGBoost, RNN, SGD, DT, RF, XGBoost & LR, για την ανάλυση & σύγκριση των datasets.	Οι XGBoost & LGBM, είναι αποτελεσματικότεροι στην έγκαιρη ανίχνευση & ταξινόμηση καταστάσεων ψυχικής υγείας, εν συγκρίσει με τους υπόλοιπους αλγορίθμους. Ο XGBoost σημειώνει υψηλή ακρίβεια, παρόλο το μεγαλύτερο χρόνο εκτέλεσής του.
Althubiti, et al. [15]	2023	Ανάπτυξη Μοντέλου Πρόβλεψης ακαδημαϊκών σπουδών, ώστε να βοηθήσει τους φοιτητές να επιλέξουν κατεύθυνση βάσει της βαθμολογίας τους. Χρησιμοποιήθηκαν 5 αλγόριθμοι: GB, KNN, LR, NN & RF. Η απόδοση μοντέλου αξιολογήθηκε με τη χρήση accuracy, AUC-ROC & πίνακα σύγχυσης (Confusion Matrix).	Ο αλγόριθμος LR, παρείχε την καλύτερη απόδοση, όσον αφορά την ακρίβεια και την τιμή AUC-ROC. Επόμενος, ήταν ο αλγόριθμος GB, ενώ τελευταίος αναδείχθηκε ο KNN, μιας και εμφάνισε την πιο αδύναμη απόδοση (**LR>GB>NN>RF>KNN**).
Lotfy, et al. [16]	2023	Ανάπτυξη ενισχυμένου συστήματος αυτόματης βαθμολόγησης εκθέσεων, με τη χρήση Τεχνικών Μηχανικής Μάθησης. Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι οι κάτωθι: RF, Bagging, KNN, DT, Lasso & AdaBoost.	Μεγάλυτερη ακρίβεια, έναντι των υπόλοιπων 5 αλγορίθμων, εμφάνισε ο RF, ενώ ο DT, τη χειρότερη λόγω της αστάθειάς του, αλλά και την περιορισμένη ικανότητα του να διαχειρίζεται πολύπλοκα datasets.

Mustafa Yağcı [17]	2022	Μελετάται η εξόρυξη εκπαιδευτικών δεδομένων, για την πρόβλεψη βαθμών τελικών εξετάσεων φοιτητών. Χρησιμοποιήθηκαν 6 αλγόριθμοι Μηχανικής Μάθησης: RF, NN, SVM, LR, NB & KNN.	Οι αλγόριθμοι RF, NN & SVM, παρουσίασαν υψηλότερη ακρίβεια ταξινόμησης, ενώ ο KNN είχε τη χαμηλότερη. Τα αποτελέσματα έδειξαν ότι οι βαθμοί των φοιτητών στις ενδιάμεσες εξετάσεις είναι σημαντικοί προγνωστικοί παράγοντες για τους βαθμούς των τελικών εξετάσεων.
Verma, et al. [18]	2022	Χρήση τεχνικών εξόρυξης δεδομένων για την πρόβλεψη των ακαδημαϊκών επιδόσεων φοιτητών. Για τη συγκεκριμένη υλοποίηση χρησιμοποιήθηκαν 5 αλγόριθμοι Μηχανικής Μάθησης: DT (CART), NB, KNN, RF & SVM. Με σκοπό τη βελτίωση της ακρίβειας, συνδυάστηκαν 3 NB Classifiers με τη χρήση Bagging.	Το μοντέλο NB παρουσίασε την υψηλότερη ακρίβεια (89,61%), ενώ ακολούθησε το DT με 85,71%. Κατόπιν, ήταν τα KNN (84,42%), SVM (81,81%) και RF (79,22%). Ιδιαίτερη επιτυχία, ύψους 91%, παρουσίασε η βελτιωμένη έκδοση του Proposed Ensemble Model (Bagging).
Akbar, et al. [19]	2021	Χρησιμοποιείται ένα dataset πρόσληψης μιας πανεπιστημιούπολης. Στην υλοποίηση συγκρίθηκαν 8 αλγόριθμοι ταξινόμησης Επιβλεπόμενης Μάθησης: LR, SVC, KNN, GNB, DT, RF, GB & LDA. Οι μετρικές αξιολόγησης των επιδόσεων που χρησιμοποιήθηκαν ήταν τα:	Ο LDA υπερείχε έναντι των άλλων αλγορίθμων. Ο συντονισμένος LDA πέτυχε υψηλότερη μέση ακρίβεια, F1-Score, G-Mean & ανταγωνιστικό ROC AUC Score. Αξίζει να σημειωθεί πως οι διαφορές των τιμών στη σύγκριση των αλγορίθμων, δεν απείχαν πολύ μεταξύ τους.

		accuracy, F1-Score, G-Mean & ROC AUC Score, με πενταπλές διασταυρούμενες επικυρώσεις για την εξασφάλιση αξιόπιστων αποτελεσμάτων.	
Paul, et al. [20]	2020	Το paper ασχολείται με την πρόβλεψη καλοήθη ή κακοήθη καρκίνο του μαστού, χρησιμοποιώντας τεχνικές feature selection, καθώς και 6 μοντέλα Μηχανικής Μάθησης (SVM, NB, DT, KNN, LR & MLP).	Το SVM έδειξε να έχει την καλύτερη απόδοση, συγκριτικά με τα υπόλοιπα μοντέλα, ενώ τα LR & KNN, παρουσίασαν καλές επιδόσεις. Τέλος, τη χειρότερη απόδοση εμφάνισε το MLP, τόσο σα μεμονωμένη χρήση, όσο και συνδυαστικά με άλλες μεθόδους.
Mayilvaganan, et al. [21]	2015	Ανάλυση γνωστικών δεξιοτήτων των μαθητών, μέσω ερωτηματολογίων προσωπικότητας, λογικής & αριθμητικής ικανότητας, με τη χρήση τεχνικών εξόρυξης δεδομένων. Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι 3: NB Classification, FP-Growth Association και τέλος ο K-Means Clustering.	Ο NB απέδωσε εξαιρετικά στην κατηγοριοποίηση των μαθητών, ενώ οι FP-Growth & K-Means, επέφεραν σημαντικές συσχετίσεις & ομάδες δεξιοτήτων.

3) Αποτελέσματα

Τα papers ανέδειξαν σημαντικές διαφορές στην απόδοση των αλγορίθμων. Ενδεικτικά:

- ✓ **Munde et al. (2024):** Ο RF εμφάνισε 100% ακρίβεια στο train set, αλλά μόλις 53% στο test set, αναδεικνύοντας προβλήματα overfitting.
- ✓ **Kanchon et al. (2024):** Ο XGBoost αποδείχθηκε ο αποδοτικότερος αλγόριθμος, υπερέχοντας στην ταξινόμηση εξατομικευμένων στυλ μάθησης.
- ✓ **Al-Tameemi et al. (2024):** Το FDN πέτυχε ακρίβεια 95,3% σε clustered datasets, ακολουθούμενο από DT (93,7%) και RF (93,4%).
- ✓ **Daza et al. (2023):** Η μέθοδος Ensemble Stacking με oversampling βελτίωσε την ακρίβεια, με τον SVM να φτάνει το 96,49%.
- ✓ **Akbar et al. (2021):** Ο LDA αναδείχθηκε ως ο πιο αποδοτικός αλγόριθμος, επιτυγχάνοντας την υψηλότερη ακρίβεια και F1-score.

4) Συμπεράσματα

Η ανασκόπηση κατέδειξε τη σημασία της επιλογής του κατάλληλου αλγορίθμου ανάλογα με το είδος και τη δομή των δεδομένων. Τα Ensemble Methods και τα υβριδικά μοντέλα τείνουν να προσφέρουν υψηλότερη ακρίβεια, ενώ η χρήση τεχνικών προεπεξεργασίας όπως η PCA και το oversampling συμβάλλουν στη βελτίωση των αποτελεσμάτων. Η εφαρμογή διαφορετικών μετρικών αξιολόγησης (accuracy, precision, recall, F1-score) προσφέρει μια ολοκληρωμένη εικόνα της απόδοσης κάθε μοντέλου.

Η ανάλυση αυτή αποτελεί τη βάση για τη βελτίωση και τον εμπλουτισμό των μεθοδολογιών που θα εφαρμοστούν στην παρούσα διπλωματική εργασία, στοχεύοντας στη δημιουργία ενός ισχυρού και ακριβούς μοντέλου πρόβλεψης της ακαδημαϊκής απόδοσης των φοιτητών.

2.2. Παράγοντες που επηρεάζουν την Απόδοση των Μαθητών

Πολλοί μαθητές εμφανίζουν χαμηλή επίδοση στο σχολείο και τις περισσότερες φορές η αποτυχία τους αποδίδεται σε αδυναμία ή τεμπελιά να ασχοληθούν με τα μαθήματα. Όμως αυτή η χαμηλή σχολική επίδοση των μαθητών μπορεί να συμβαίνει εξαιτίας πολλών άλλων παραγόντων. Αναλυτικότερα:

Αίσθημα δειλίας ή ντροπής που προσδίδεται μέσω της έλλειψης φιλικών σχέσεων, της αδυναμίας συμμετοχής σε διάφορες δραστηριότητες και δει σχολικές, καθώς και την απομόνωση.

Έλλειψη σχολικής ωριμότητας που επιφέρει στο μαθητή δυσκολία στη γραφή και στην ανάγνωση, αλλά και στην παρακολούθηση των μαθημάτων του, αδυναμία στο να προσαρμόζεται σε κανόνες και στο περιβάλλον της τάξης, φτωχό λεξιλόγιο και τέλος δυσκολία στις συναναστροφές του. Η σχολική ωριμότητα αναφέρεται στο πως εξελίσσεται η παιδική νοημοσύνη, και διαφέρει από μαθητή σε μαθητή, τη βιολογική του ωρίμανση, την κοινωνική και συναισθηματική του εξέλιξη.

Τρόπος διδασκαλίας και συμπεριφοράς εκπαιδευτικών απέναντι στους μαθητές, επηρεάζει στο έπακρο τη σχολική επίδοσή τους. Κάθε είδους τιμωρίας ή επίπληξης, δημιουργούν προβλήματα στην ψυχοσύνθεση των μαθητών. Απότοκο αυτού η δημιουργία επιθετικών συμπεριφορών, προβλημάτων ένταξης στο σύνολο, μαθησιακών δυσκολιών, άγχος, συναισθημάτων απόρριψης και αποτυχίας, καθώς και ανασφάλειας. Συνοψίζοντας, όταν το σχολικό πλαίσιο δεν υποστηρίζει τους μαθητές, τότε είναι πιο εμφανείς οι παραπάνω συμπεριφορές.

Απουσία εσωτερικών κινήτρων η οποία δεν αποτελεί την αξία της γνώσης, αλλά και την εκπαιδευτική διαδικασία στη βάση.

Μαθησιακές δυσκολίες, είτε αυτές είναι ακουστικές, είτε οπτικές, που οδηγούν σε χαμηλή επίδοση στο σχολείο, μιας και υπάρχει αδυναμία παρακολούθησης του ρυθμού μάθησης της τάξης.

Ελλιπής φοίτηση με την παρουσία αδικαιολόγητων απουσιών είναι ένδειξη αρνητικής στάσης απέναντι στο σχολείο. Απόρροια αυτού μαθητές χαμηλής επίδοσης, σχολική στασιμότητα, ακόμη και εγκατάλειψη.

Οικογενειακός παράγοντας που είναι υψίστης σημασίας για τη χαμηλή σχολική επίδοση. Αυτό οφείλεται στην έλλειψη ενδιαφέροντος και υποστήριξης από την ίδια τους την οικογένεια, και εμφανίζεται μέσω της κούρασης των μαθητών, αδυναμία προσοχής στο μάθημα, δυσκολία σε οποιαδήποτε συζήτηση, κλείσιμο στον εαυτό τους και μελαγχολική έκφραση.

Ψυχική πίεση που ασκείται πολύ συχνά από καθηγητές και γονείς, ώστε οι μαθητές να επιτύχουν υψηλές επιδόσεις. Πολλές φορές, τα αποτελέσματα που επιτυγχάνονται από την υπερβολική απαίτηση είναι τα άκρως αντίθετα.

Χαμηλό αίσθημα αυτοεκτίμησης που επιδρά σημαντικά στην απόδοση των μαθητών. Ο μόνος τρόπος για να αλλάξει αυτό είναι να δημιουργηθεί ισχυρή αυτοεκτίμηση και εμπιστοσύνη από το μαθητή προς τις δυνάμεις του.

Είναι σημαντικό να αναφερθεί ότι μπορούν να αντιμετωπιστούν οι αιτίες που επιβραδύνουν τη σχολική ανάπτυξη των μαθητών. Πιο συγκεκριμένα:

- ✓ Όσον αφορά τη σχολική ανωριμότητα, οι ίδιοι οι γονείς οφείλουν να κατανοήσουν τα παιδιά τους, και να μη το αποδίδουν σε αδιαφορία ή τεμπελιά. Επίσης, μπορούν να παρακινούν τα παιδιά τους, και οι καταστάσεις που αντιμετωπίζουν να αξιολογηθούν από ειδικούς επιστήμονες, τα οποία θα οδηγήσουν στη βέλτιστη αντιμετώπιση του προβλήματος.
- ✓ Τα εσωτερικά κίνητρα που οφείλουν να αποκτήσουν οι μαθητές, ώστε να κατανοήσουν τις αξίες της γνώσης, διότι τα εσωτερικά κίνητρα έχουν πρωτεύων ρόλο στη διά βίου του μαθητή.
- ✓ Η σχολική επίδοση των μαθητών επηρεάζεται στο έπακρο από τις προσδοκίες, φιλοδοξίες, αξίες, συμπεριφορές και στάσεις των γονιών τους. Οι γονείς φροντίζουν τα παιδιά τους να γίνονται ανεξάρτητα, και τα παρακινούν με διάφορες εργασίες, με σκοπό να δημιουργηθούν κίνητρα επίδοσης. Τα κίνητρα επίδοσης είναι οι απαιτήσεις που φέρουν οι γονείς, ώστε να ανταποκρίνονται στις δυνατότητες των παιδιών τους.
- ✓ Για τα παιδιά που είναι συνεσταλμένα, πρέπει να παρέχεται ιδιαίτερη προσοχή, σωστή προσέγγιση και ενθάρρυνση για συμμετοχή σε διάφορες δραστηριότητες. Επιπλέον, είναι αρκετά ωφέλιμο να τονίζονται τα θετικά στοιχεία των μαθητών από τους εκπαιδευτές τους,

ώστε να επαινείται η οποιαδήποτε προσπάθεια που υλοποιείται ανεξαρτήτων του αποτελέσματος.

Εν τέλει, οι μαθητές καλό είναι να νιώθουν υποστήριξη από το ίδιο το σύστημα, δηλαδή το σχολείο, νιώθοντας σεβασμό και υποστήριξη ως προς την προσωπικότητας τους. Τέλος, οι μαθητές πρέπει να συμμετέχουν σε ποικίλες δραστηριότητες, προκειμένου να ανακαλύψουν τις ικανότητές τους [22].

Η απόδοση των μαθητών μπορεί να επηρεαστεί από πλήθος παραγόντων, και διαφέρουν ανάλογα το εκπαιδευτικό, οικονομικό και κοινωνικό υπόβαθρο. Αναλυτικότερα:

i) **Δέσμευση και Συμμετοχή:** Ο χρόνος που αφιερώνεται για μελέτη, η συμμετοχή σε μαθήματα, ακόμη και σε δραστηριότητες, συσχετίζονται άμεσα με την επίδοση των μαθητών.

ii) **Αλληλεπίδραση με το Εκπαιδευτικό Προσωπικό:** Η αποτελεσματικότητα της διδασκαλίας και η σχέση ανάμεσα σε εκπαιδευτή και εκπαιδευόμενο, επηρεάζουν εξίσου σημαντικά την ακαδημαϊκή πρόοδο.

iii) **Δημογραφικά Χαρακτηριστικά:** Μελέτες έχουν αποδείξει ότι το πολιτισμικό υπόβαθρο, η ηλικία και το φύλο των μαθητών, επηρεάζουν την απόδοση τους.

iv) **Κοινωνικοί και Ψυχολογικοί Παράγοντες:** Τα επίπεδα άγχους, η υποστήριξη, τα κίνητρα και η αυτοεκτίμηση, είναι και αυτοί σημαντικοί παράγοντες για την επίδοση των μαθητών.

v) **Κοινωνικοοικονομικοί Παράγοντες:** Η πρόσβαση σε εκπαιδευτικούς πόρους, το μορφωτικό επίπεδο που έχουν οι γονείς, αλλά και το οικογενειακό εισόδημα, απαρτίζουν σημαντικούς παράγοντες στην ακαδημαϊκή πρόοδο των μαθητών.

Προγνωστικά μοντέλα έχουν δημιουργηθεί χάρις την ανάλυση των παραπάνω παραγόντων, προκειμένου να εντοπίζονται μαθητές με χαμηλή απόδοση, και να δημιουργηθούν στρατηγικές παρέμβασης.

Επιπροσθέτως, για την αξιολόγηση των μαθητών, δημιουργούνται κρίσιμοι δείκτες, όπως είναι η παρακολούθηση, ο χρόνος συμμετοχής και οι βαθμολογίες. Πιο συγκεκριμένα:

i) **Παρακολούθηση:** Οι απουσίες από τις παρακολουθήσεις μαθημάτων μπορούν να οδηγήσουν σε χαμηλή απόδοση, αφού υποδεικνύονται κατ' αυτόν τον τρόπο διάφορα κοινωνικοοικονομικά προβλήματα. Συνοψίζοντας, η συχνή παρακολούθηση συνδέεται άρρηκτα με την ακαδημαϊκή επιτυχία.

ii) **Χρόνος Συμμετοχής:** Ο χρόνος που αφιερώνεται σε ομαδικές εργασίες, διαδικτυακές πλατφόρμες και άλλες δραστηριότητες, αποκαλύπτει το βαθμό που εμπλέκεται ένας μαθητής. Έχει αποδειχθεί ότι υψηλότερες επιδόσεις αποδίδονται σε μαθητές που αφιερώνουν χρόνο σε οργανωμένες δραστηριότητες.

iii) **Βαθμολογίες:** Συγκροτούν τον αποτελεσματικότερο και βασικότερο τρόπο μέτρησης της απόδοσης των μαθητών. Μέσα από τις αναλύσεις των βαθμολογιών, αναγνωρίζονται ποικίλα πρότυπα συμπεριφοράς, όπως είναι λόγου χάρη ο εντοπισμός των μαθητών που τείνουν να αποτύχουν, ή ο συσχετισμός ανάμεσα σε απόδοση και μαθήματα.

Τα προαναφερθέντα εμπεριέχουν σημαντικές πληροφορίες και δεδομένα στους εκπαιδευτικούς, ώστε να εφαρμοστούν εξατομικευμένες παρεμβάσεις. Επίσης, είναι απόλυτα πιθανό να βελτιωθεί η ακρίβεια των μοντέλων πρόβλεψης, μέσω της συσχέτισης των μετρήσεων που επιτυγχάνεται με δεδομένα [23].

Η φράση "Υπολογιστικά Μοντέλα που έχουν εφαρμοστεί", αναφέρεται σε τεχνικές, αλγορίθμους και μεθόδους που χρησιμοποιήθηκαν προκειμένου να αναλυθούν δεδομένα και να εξαχθούν χρήσιμα συμπεράσματα, όπως η κατηγοριοποίηση και η πρόβλεψη. Στον κλάδο της εκπαιδευτικής ανάλυσης, τα εν λόγω μοντέλα εφαρμόζονται για τον εντοπισμό παραγόντων που επηρεάζουν την απόδοση των μαθητών, και δημιουργούν προβλέψεις που σχετίζονται με τη μελλοντική τους επίδοση.

i) **Γραμμική Παλινδρόμηση - Linear Regression:** Συγκροτεί ένα υπολογιστικό μοντέλο που χρησιμοποιείται για τη μελέτη σχέσεων μεταξύ διαφορετικών μεταβλητών.

Η Μηχανική Μάθηση, συγκροτεί έναν κλάδο της Τεχνητής Νοημοσύνης, που επικεντρώνεται στην ανάπτυξη αλγορίθμων και στατιστικών μοντέλων τα οποία "μαθαίνουν" από δεδομένα και μπορούν να κάνουν προβλέψεις. Η γραμμική παλινδρόμηση ανήκει στην κατηγορία των εποπτευόμενων αλγορίθμων, καθώς βασίζεται σε δεδομένα με ετικέτες, δηλαδή σε δεδομένα των οποίων η επιθυμητή έξοδος είναι ήδη γνωστή. Ο αλγόριθμος προσαρμόζει τα δεδομένα σε μια βέλτιστη γραμμική συνάρτηση, όπου στη συνέχεια μπορεί να χρησιμοποιηθεί για προβλέψεις σε καινούρια σύνολα δεδομένων.

Κατηγορίες εποπτευόμενης μάθησης

Η εποπτευόμενη μάθηση περιλαμβάνει δύο βασικές κατηγορίες:

- ✓ **Ταξινόμηση - Classification**: Πρόβλεψη της κατηγορίας στην οποία ανήκει ένα δεδομένο.
- ✓ **Παλινδρόμηση - Regression**: Πρόβλεψη μιας συνεχούς μεταβλητής.

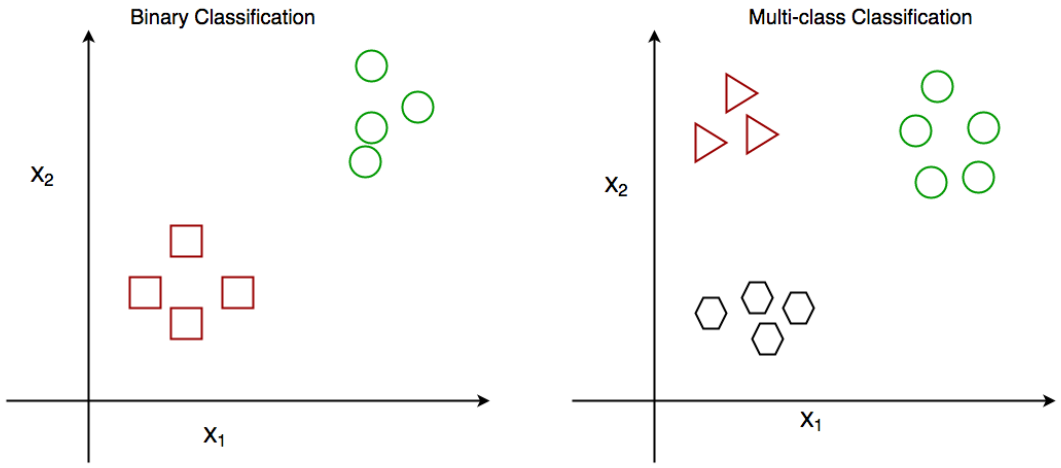
Η γραμμική παλινδρόμηση είναι ένας αλγόριθμος εποπτευόμενης μάθησης που υπολογίζει τη σχέση μεταξύ μιας εξαρτημένης μεταβλητής και ενός ή περισσότερων ανεξάρτητων μεταβλητών, προσαρμόζοντας μια γραμμική εξίσωση στα δεδομένα.

Ανάλογα με τον αριθμό των μεταβλητών που εμπλέκονται, διακρίνεται σε:

- ✓ **Απλή Γραμμική Παλινδρόμηση**: Όταν υπάρχει μία ανεξάρτητη μεταβλητή.
- ✓ **Πολλαπλή Γραμμική Παλινδρόμηση**: Όταν υπάρχουν περισσότερες ανεξάρτητες μεταβλητές.
- ✓ **Μονομεταβλητή Γραμμική Παλινδρόμηση**: Όταν υπάρχει μόνο μία εξαρτημένη μεταβλητή.
- ✓ **Πολυμεταβλητή Παλινδρόμηση**: Όταν υπάρχουν περισσότερες εξαρτημένες μεταβλητές.

Ένα από τα μεγαλύτερα πλεονεκτήματα της γραμμικής παλινδρόμησης είναι η ερμηνευσιμότητά της. Η μαθηματική της εξίσωση παρέχει σαφείς συντελεστές, οι οποίοι δείχνουν πώς κάθε ανεξάρτητη μεταβλητή επηρεάζει την εξαρτημένη, επιτρέποντας καλύτερη κατανόηση της σχέσης μεταξύ των δεδομένων.

Επιπλέον, η γραμμική παλινδρόμηση αποτελεί τη βάση για πιο εξελιγμένα υπολογιστικά μοντέλα. Τεχνικές όπως οι μηχανές υποστήριξης διανυσμάτων (SVMs) και οι αλγόριθμοι κανονικοποίησης βασίζονται σε αυτήν, ενώ παίζει επίσης σημαντικό ρόλο στη δοκιμή υποθέσεων, επιτρέποντας στους ερευνητές να επικυρώνουν θεωρίες και υποθέσεις σχετικά με τα δεδομένα τους [24].



Εικόνα 6: Δυαδική Ταξινόμηση VS Ταξινόμησης Πολλαπλών Κατηγοριών

ii) **Αλγόριθμοι Κατηγοριοποίησης - Classification Algorithms:** Οι αλγόριθμοι κατηγοριοποίησης χρησιμοποιούνται για την ομαδοποίηση δεδομένων σε συγκεκριμένες κατηγορίες, και βασίζονται σε διάφορα χαρακτηριστικά. Ορισμένοι από τους πιο διαδεδομένους αλγόριθμους είναι:

- ✓ **Decision Trees - Δέντρα Απόφασης**: Δημιουργούν ένα σύνολο κανόνων με βάση τα διαθέσιμα δεδομένα, βοηθώντας στη λήψη αποφάσεων.
- ✓ **Random Forest**: Συνδυάζει πολλά δέντρα απόφασης για πιο αξιόπιστα αποτελέσματα.

- ✓ **Logistic Regression**: Χρησιμοποιείται για προβλήματα δυαδικής ταξινόμησης.
- ✓ **Support Vector Machines - SVM**: Δημιουργούν ένα όριο που διαχωρίζει τις κατηγορίες των δεδομένων με τον καλύτερο δυνατό τρόπο.

Η εποπτευόμενη μάθηση -Supervised Learning- περιλαμβάνει ένα σύνολο δεδομένων όπου υπάρχει ήδη μια γνωστή έξοδος (Y) και χρησιμοποιείται ένας αλγόριθμος για να μάθει τη σχέση μεταξύ εισόδων (X) και εξόδων (Y). Στόχος είναι η κατασκευή ενός μοντέλου που, όταν του δοθούν νέα δεδομένα, θα μπορεί να κάνει προβλέψεις με ακρίβεια.

Τα προβλήματα εποπτευόμενης μάθησης χωρίζονται σε δύο βασικές κατηγορίες:

- ✓ **Παλινδρόμηση - Regression:** Χρησιμοποιείται όταν πρέπει να προβλεφθεί μια συνεχής αριθμητική τιμή.
- ✓ **Ταξινόμηση - Classification:** Εφαρμόζεται όταν τα δεδομένα είναι αναγκαίο να κατηγοριοποιηθούν.

Η ταξινόμηση είναι μια βασική διαδικασία στη μηχανική μάθηση, όπου τα δεδομένα ταξινομούνται σε προκαθορισμένες κατηγορίες με βάση συγκεκριμένα χαρακτηριστικά. Ο στόχος είναι η δημιουργία ενός μοντέλου που μπορεί να προβλέψει σωστά την κατηγορία ενός νέου δείγματος.

Υπάρχουν διάφοροι τύποι ταξινόμησης στη μηχανική μάθηση:

i) **Δυαδική Ταξινόμηση - Binary Classification:** Τα δεδομένα ταξινομούνται σε μία από δύο κατηγορίες.

ii) **Ταξινόμηση Πολλαπλών Κατηγοριών - Multi-Class Classification:** Τα δεδομένα ανήκουν σε περισσότερες από δύο κατηγορίες.

iii) **Ταξινόμηση Πολλαπλών Ετικετών - Multi-Label Classification:** Κάθε δείγμα μπορεί να ανήκει σε περισσότερες από μία κατηγορίες ταυτόχρονα.

iv) **Ανισορροπημένη Ταξινόμηση - Imbalanced Classification:** Όταν μία από τις κατηγορίες είναι πολύ σπάνια σε σχέση με τις άλλες.

v) **Αλγόριθμοι Ταξινόμησης:** Οι αλγόριθμοι ταξινόμησης χωρίζονται σε δύο βασικές κατηγορίες:

1) Γραμμικοί Ταξινομητές - Linear Classifiers: Αυτοί οι αλγόριθμοι δημιουργούν ένα γραμμικό όριο απόφασης ανάμεσα στις κατηγορίες. Είναι απλοί και αποδοτικοί υπολογιστικά. Παραδείγματα:

- **Logistic Regression**
- **Support Vector Machines (με γραμμικό πυρήνα)**
- **Perceptron μονής στρώσης**
- **Stochastic Gradient Descent (SGD) Classifier**

2) Μη Γραμμικοί Ταξινομητές (Non-Linear Classifiers): Δημιουργούν πιο πολύπλοκα όρια απόφασης, καταγράφοντας πιο σύνθετες σχέσεις στα δεδομένα. Παραδείγματα:

- **k-Nearest Neighbors (k-NN)**
- **Support Vector Machines (SVM) με μη γραμμικούς πυρήνες**
- **Naïve Bayes**
- **Decision Trees (Δέντρα Απόφασης)**

3) Ταξινομητές Μάθησης Συνόλου - Ensemble Learning Classifiers: Συνδυάζουν πολλά απλά μοντέλα για να βελτιώσουν την απόδοση:

- **Random Forest**
- **AdaBoost**
- **Bagging Classifier**

- **Voting Classifier**
- **Extra Trees Classifier**

4) Νευρωνικά Δίκτυα - Neural Networks: Τα Πολυεπίπεδα Τεχνητά Νευρωνικά Δίκτυα (Multilayer Artificial Neural Networks) χρησιμοποιούνται για πιο σύνθετες ταξινομήσεις, όπως η αναγνώριση εικόνων και η επεξεργασία φυσικής γλώσσας.

Οι προαναφερθέντες αλγόριθμοι είναι βασικά εργαλεία στη μηχανική μάθηση και χρησιμοποιούνται σε πολλούς τομείς, όπως η ανάλυση δεδομένων, η ιατρική διάγνωση και οι προτάσεις περιεχομένου [25].

iii) Αλγόριθμοι Συνόλου - Ensemble Methods: Συνδυάζουν πολλαπλά απλά μοντέλα για να δημιουργήσουν ένα πιο ισχυρό και ακριβές μοντέλο πρόβλεψης.

Το Ensemble Learning είναι μια μέθοδος μηχανικής μάθησης που βελτιώνει την απόδοση ενός μοντέλου συνδυάζοντας προβλέψεις από πολλά ανεξάρτητα μοντέλα. Η βασική ιδέα στηρίζεται στην αξιοποίηση των διαφορετικών δυνατοτήτων κάθε μοντέλου, ώστε να επιτευχθεί μεγαλύτερη ακρίβεια και γενίκευση.

Ενώ μεμονωμένα μοντέλα μπορεί να έχουν περιορισμούς ή σφάλματα, η συνδυαστική τους χρήση μπορεί να οδηγήσει σε πιο αξιόπιστα αποτελέσματα. Αυτή η προσέγγιση θεωρείται ως ένας τρόπος αντιμετώπισης των αδύναμων μοντέλων μάθησης, τα οποία, αν και πιο υπολογιστικά απαιτητικά, προσφέρουν υψηλότερη απόδοση σε σχέση με ένα μεμονωμένο μοντέλο που έχει εκπαιδευτεί εκτενώς.

Σε αυτές τις τεχνικές, πολλαπλά μοντέλα εκπαιδεύονται χρησιμοποιώντας το ίδιο σύνολο δεδομένων. Τα αποτελέσματά τους συνδυάζονται για να παράγουν μια συνολική απόφαση. Τα βασικά μοντέλα μπορεί να περιλαμβάνουν Δέντρα Αποφάσεων, Γραμμικά Μοντέλα, Υποστηρικτές Διανυσμάτων (SVMs), Νευρωνικά Δίκτυα ή άλλους αλγορίθμους πρόβλεψης.

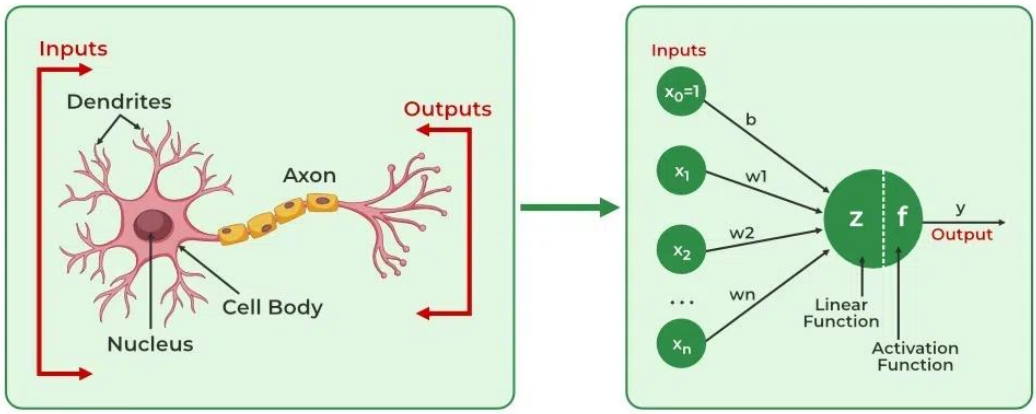
Δημοφιλείς Τεχνικές Ensemble Learning

- ✓ **Bagging**: Χρησιμοποιείται για την κατασκευή αλγορίθμων όπως το Random Forest.
- ✓ **Boosting**: Βελτιώνει αλγορίθμους όπως το Adaboost και το XGBoost.

Προηγμένες Τεχνικές Ensemble Learning

- ✓ **Gradient Boosting Machines - GBM**: Δημιουργεί μια ακολουθία δέντρων αποφάσεων, όπου κάθε νέο δέντρο προσπαθεί να διορθώσει τα σφάλματα των προηγούμενων, βελτιώνοντας την ακρίβεια της πρόβλεψης.
- ✓ **Extreme Gradient Boosting - XGBoost**: Χρησιμοποιεί μηχανισμούς όπως κλάδεμα δέντρων, τακτική -regularization- και παράλληλη επεξεργασία, καθιστώντας το ιδανικό για μεγάλες βάσεις δεδομένων.
- ✓ **CatBoost**: Σχεδιασμένο να διαχειρίζεται κατηγορηματικά χαρακτηριστικά χωρίς εκτεταμένη προεπεξεργασία, παρέχοντας γρήγορη εκπαίδευση και υψηλή ακρίβεια.
- ✓ **Stacking**: Συνδυάζει τις προβλέψεις πολλαπλών μοντέλων με έναν επιπλέον μετα-αλγόριθμο που μαθαίνει πώς να συνδυάζει τις επιμέρους προβλέψεις για μεγαλύτερη ακρίβεια.
- ✓ **Random Subspace Method**: Χρησιμοποιεί υποσύνολα των χαρακτηριστικών εισόδου, προσφέροντας μεγαλύτερη ποικιλομορφία και μειώνοντας το overfitting.
- ✓ **Τυχαίες Παραλλαγές Δασών**: Εισάγουν διαφοροποιήσεις στη δημιουργία δέντρων, στην επιλογή χαρακτηριστικών ή στη βελτιστοποίηση του μοντέλου, βελτιώνοντας τη συνολική απόδοση.

Η επιλογή της κατάλληλης τεχνικής εξαρτάται από τη φύση των δεδομένων, το πρόβλημα που πρέπει να λυθεί και τους διαθέσιμους υπολογιστικούς πόρους. Συχνά απαιτούνται δοκιμές και πειραματισμοί για την επίτευξη των βέλτιστων αποτελεσμάτων [26].



Εικόνα 7: Αναλογία μεταξύ Βιολογικού & Τεχνητού Νευρώνα

iv) Νευρωνικά Δίκτυα - Neural Networks: Βασικές Αρχές και Εφαρμογές

Τα νευρωνικά δίκτυα είναι μοντέλα μηχανικής μάθησης εμπνευσμένα από τη λειτουργία του ανθρώπινου εγκεφάλου. Χρησιμοποιούνται για την ανάλυση σύνθετων δεδομένων με πολλές μεταβλητές, αναγνωρίζοντας μοτίβα και λαμβάνοντας αποφάσεις μέσω της διαδικασίας εκμάθησης χαρακτηριστικών.

Αποτελούνται από διασυνδεδεμένους κόμβους -νευρώνες- που οργανώνονται σε επίπεδα και μαθαίνουν μέσω προσαρμογής των παραμέτρων τους. Τα Νευρωνικά Δίκτυα είναι η βάση της βαθιάς μάθησης -Deep Learning- και βρίσκουν εφαρμογή σε πολλούς τομείς, όπως η επεξεργασία φυσικής γλώσσας, η όραση υπολογιστών και τα αυτόνομα οχήματα.

Ένα Νευρωνικό Δίκτυο αποτελείται από τρία βασικά στοιχεία:

1) **Νευρώνες – Nodes:** Οι μονάδες επεξεργασίας δεδομένων, που λαμβάνουν εισόδους, τις επεξεργάζονται και παράγουν εξόδους μέσω συναρτήσεων ενεργοποίησης.

2) **Συνδέσεις**: Οι δεσμοί μεταξύ των νευρώνων, που καθορίζουν τη ροή των πληροφοριών μέσω βαρών και προκαταλήψεων -biases-.
3) **Στρώματα – Layers:** Τα επίπεδα στα οποία κατανέμονται οι νευρώνες:
    - **Είσοδος - Input Layer**: Λαμβάνει τα δεδομένα.
    - **Κρυφά επίπεδα - Hidden Layers**: Επεξεργάζονται τα δεδομένα και εξάγουν χαρακτηριστικά.
    - **Έξοδος - Output Layer**: Παράγει την τελική πρόβλεψη ή απόφαση του δικτύου.

Τα Νευρωνικά δίκτυα μαθαίνουν μέσω προσαρμογής των βαρών τους, ακολουθώντας μια επαναληπτική διαδικασία που περιλαμβάνει:

1) **Υπολογισμό εισόδου**: Τα δεδομένα τροφοδοτούνται στο δίκτυο.
2) **Πρόβλεψη εξόδου**: Το δίκτυο χρησιμοποιεί τις τρέχουσες παραμέτρους για να δημιουργήσει μια εκτίμηση.
3) **Βελτιστοποίηση – Backpropagation:** Το σφάλμα μεταξύ πρόβλεψης και πραγματικών δεδομένων υπολογίζεται και χρησιμοποιείται για τη ρύθμιση των βαρών, ώστε να βελτιώνεται η ακρίβεια.

Η διαδικασία αυτή πραγματοποιείται μέσω αλγορίθμων όπως ο Σταδιακός Καθοδικός Βαθμός (Gradient Descent), ο οποίος προσαρμόζει τα βάρη ώστε να μειώσει το συνολικό σφάλμα του δικτύου.

Εφαρμογές Νευρωνικών Δικτύων

Η ικανότητα των νευρωνικών δικτύων να αναγνωρίζουν περίπλοκα μοτίβα τα καθιστά ιδανικά για πολλές εφαρμογές:

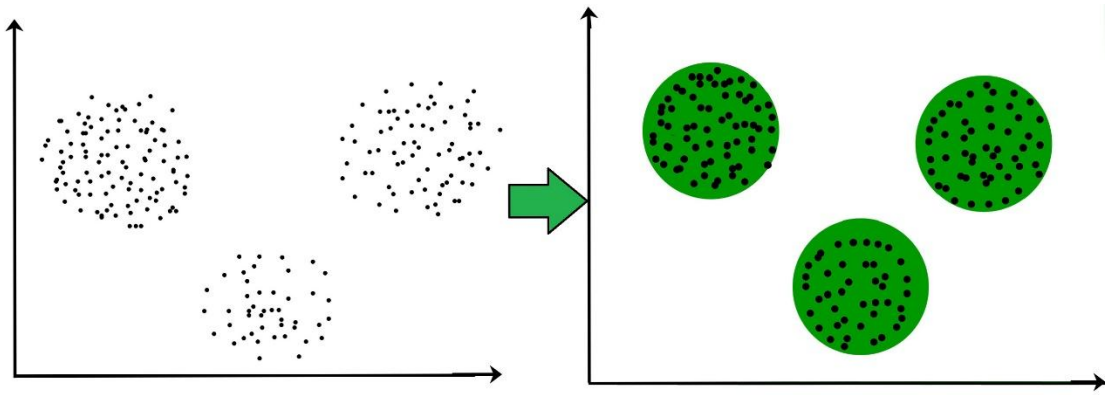
- **Επεξεργασία Φυσικής Γλώσσας – NLP:** Αυτόματη μετάφραση, chatbot, ανάλυση συναισθήματος.

- **Υπολογιστική Όραση**: Αναγνώριση προσώπων, διάγνωση ιατρικών εικόνων.
- **Αυτόνομα Οχήματα**: Ανάλυση περιβάλλοντος και λήψη αποφάσεων.
- **Συστήματα Σύστασης**: Προτάσεις περιεχομένου σε πλατφόρμες όπως Netflix, YouTube.

Τα νευρωνικά δίκτυα αποτελούν τον πυρήνα της σύγχρονης τεχνητής νοημοσύνης (AI), επιτρέποντας την ανάπτυξη μοντέλων που μπορούν να αναλύσουν μεγάλα σύνολα δεδομένων και να προσαρμοστούν σε δυναμικά περιβάλλοντα.

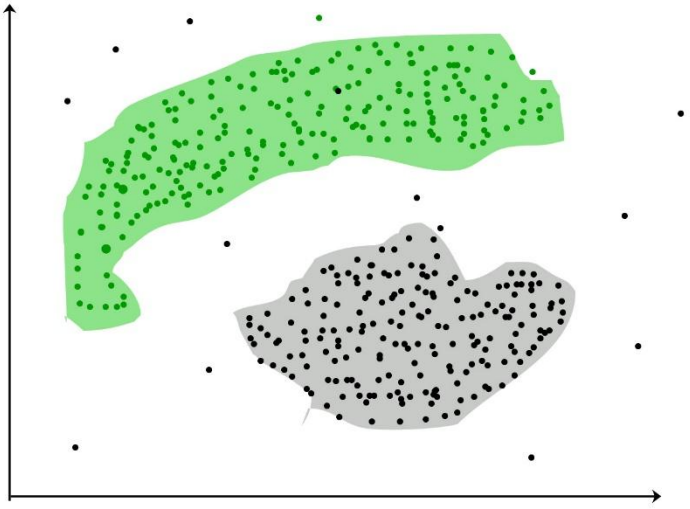
Με την εξέλιξη της υπολογιστικής ισχύος και την ανάπτυξη εξειδικευμένων μοντέλων όπως τα convolutional neural networks (CNNs) και τα transformers, η έρευνα συνεχώς προχωρά, διαμορφώνοντας το μέλλον της τεχνολογίας [27].

Στο διάγραμμα που ακολουθεί, αναπαρίστανται 3 κυκλικά σμήνη τα οποία σχηματίστηκαν βάσει της απόστασης.



Εικόνα 8: Κυκλικά Σμήνη

Στο παρακάτω γράφημα, απεικονίζονται τα συμπλέγματα που σχηματίζονται, τα οποία ωστόσο δεν έχουν κυκλικό σχήμα.



Εικόνα 9: Συμπλέγματα που σχηματίζονται

v) **Clustering:** Η ομαδοποίηση είναι μια τεχνική που χρησιμοποιείται για την ταξινόμηση δεδομένων σε ομάδες με βάση τις ομοιότητές τους. Είναι μια μορφή μη επιβλεπόμενης μάθησης -unsupervised learning-, όπου τα δεδομένα δεν έχουν προκαθορισμένες ετικέτες ή κατηγορίες. Σε αντίθεση με τις μεθόδους επιβλεπόμενης μάθησης, όπου έχουμε μια μεταβλητή-στόχο, η ομαδοποίηση στοχεύει στον εντοπισμό δομών και προτύπων μέσα στα δεδομένα.

Η βασική ιδέα πίσω από την ομαδοποίηση είναι η αναγνώριση ομάδων -clusters- που περιέχουν δεδομένα με κοινά χαρακτηριστικά. Αυτό γίνεται μέσω μέτρων απόστασης, όπως:

- ✓ **Ευκλείδεια απόσταση** - Euclidean distance
- ✓ **Απόσταση Manhattan** - Manhattan distance
- ✓ **Συσχέτιση συνημίτονου** - Cosine similarity

Τα σύνολα δεδομένων που αναλύονται μέσω clustering μπορεί να περιέχουν δεδομένα που δεν ακολουθούν αυστηρά κυκλικά ή γεωμετρικά σχήματα. Ορισμένοι αλγόριθμοι μπορούν να ανιχνεύσουν ομάδες με πιο σύνθετες και ακανόνιστες μορφές.

Υπάρχουν δύο βασικές προσεγγίσεις στην ομαδοποίηση:

i) **Σκληρή Ομαδοποίηση - Hard Clustering**: Κάθε δεδομένο ανήκει αποκλειστικά σε μία ομάδα.

ii) **Απαλή Ομαδοποίηση - Soft Clustering**: Κάθε δεδομένο μπορεί να έχει μια πιθανότητα να ανήκει σε διαφορετικές ομάδες.

Η ομαδοποίηση είναι μια σημαντική τεχνική στην ανάλυση δεδομένων, καθώς μας βοηθά να οργανώσουμε και να κατανοήσουμε μεγάλες ποσότητες μη δομημένων πληροφοριών. Οι πιο γνωστοί αλγόριθμοι για ομαδοποίηση περιλαμβάνουν:

i) **Ομαδοποίηση με βάση το Κέντρο Βάρους - Centroid-based clustering**: Ο αλγόριθμος χωρίζει τα δεδομένα σε Κ ομάδες, προσδιορίζοντας ένα κεντρικό σημείο για κάθε cluster και αναδιαμορφώνοντας τις ομάδες με βάση την εγγύτητα των δεδομένων σε αυτό.

ii) **Ομαδοποίηση με βάση την Πυκνότητα - Density-based clustering**: Χρησιμοποιεί περιοχές υψηλής συγκέντρωσης δεδομένων για τον καθορισμό των ομάδων, αντί να βασίζεται σε αυστηρά όρια ή προκαθορισμένα κέντρα.

iii) **Ιεραρχική Ομαδοποίηση - Hierarchical clustering:** Αντί να καθορίζει εξαρχής έναν αριθμό clusters, δημιουργεί μια ιεραρχική δομή όπου τα δεδομένα είτε συγχωνεύονται είτε χωρίζονται σταδιακά, σχηματίζοντας ένα δενδροειδές διάγραμμα (dendrogram).

iv) **Ομαδοποίηση με βάση την Κατανομή - Distribution-based clustering:** Υποθέτει ότι τα δεδομένα ακολουθούν συγκεκριμένες πιθανότητες κατανομής και τα ομαδοποιεί με βάση το πόσο ταιριάζουν σε αυτές τις κατανομές.

Χρήση της Ομαδοποίησης στην Εκπαίδευση

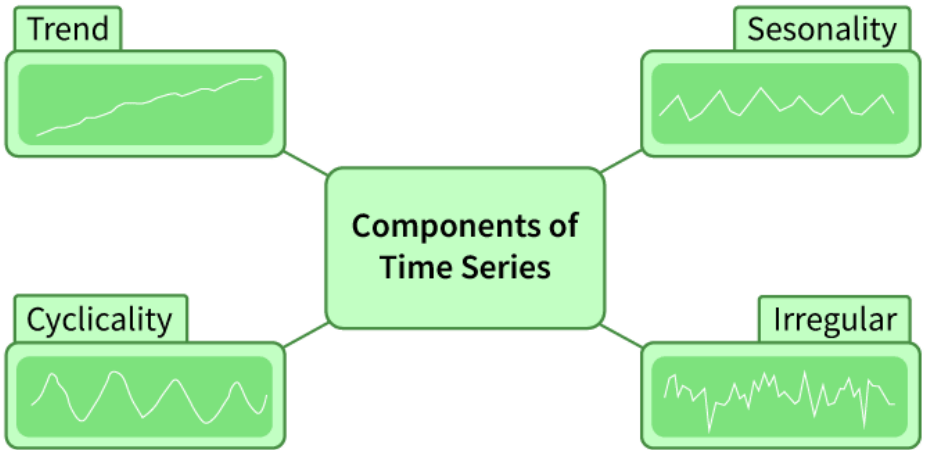
Στον τομέα της εκπαίδευσης, η ομαδοποίηση χρησιμοποιείται συχνά για την κατηγοριοποίηση μαθητών με βάση διαφορετικά χαρακτηριστικά, όπως:

- ✓ Πρότυπα μελέτης
- ✓ Επίπεδα απόδοσης
- ✓ Συμμετοχή σε διαδικτυακές πλατφόρμες μάθησης

Φερειπείν, ένας αλγόριθμος K-Means θα μπορούσε να χρησιμοποιηθεί για να διαχωρίσει μαθητές σε ομάδες υψηλής, μέσης και χαμηλής επίδοσης, επιτρέποντας στους εκπαιδευτικούς να προσαρμόσουν το διδακτικό υλικό ανάλογα με τις ανάγκες κάθε ομάδας.

Η κατανόηση και η σωστή εφαρμογή των τεχνικών ομαδοποίησης μπορεί να βοηθήσει σημαντικά στη λήψη αποφάσεων, είτε πρόκειται για εκπαιδευτικά περιβάλλοντα είτε για επιχειρηματικές και τεχνολογικές εφαρμογές [28].

Στοιχεία δεδομένων χρονοσειρών: Υπάρχουν τέσσερα κύρια στοιχεία μιας χρονοσειράς:



Εικόνα 10: Τέσσερα Κύρια Στοιχεία Χρονοσειράς

vi) **Time Series Analysis:** Η ανάλυση χρονοσειρών αφορά τη μελέτη δεδομένων που καταγράφονται διαδοχικά στον χρόνο, σε τακτά χρονικά διαστήματα. Αυτή η μέθοδος βοηθά στην κατανόηση της εξέλιξης μιας μεταβλητής με την πάροδο του χρόνου, όπως για παράδειγμα η ακαδημαϊκή πρόοδος ενός μαθητή κατά τη διάρκεια ενός εξαμήνου, οι αλλαγές στις τιμές των μετοχών ή οι διακυμάνσεις της θερμοκρασίας σε μία περιοχή.

Η απεικόνιση αυτών των δεδομένων γίνεται συνήθως με ένα διάγραμμα όπου ο χρόνος τοποθετείται στον οριζόντιο άξονα και οι τιμές της μεταβλητής στον κατακόρυφο άξονα. Με αυτόν τον τρόπο, μπορούν εύκολα να εντοπιστούν μοτίβα, τάσεις και αλλαγές που συμβαίνουν σε βάθος χρόνου.

Η ανάλυση χρονοσειρών χρησιμοποιείται ευρέως σε πολλούς τομείς, καθώς προσφέρει χρήσιμες πληροφορίες για τη λήψη αποφάσεων. Ορισμένα από τα βασικά της πλεονεκτήματα περιλαμβάνουν:

i) **Πρόβλεψη μελλοντικών τάσεων**: Επιτρέπει την εκτίμηση της μελλοντικής εξέλιξης μιας μεταβλητής, κάτι που είναι ιδιαίτερα χρήσιμο σε επιχειρήσεις για τη διαχείριση αποθεμάτων, τις επενδύσεις και τις στρατηγικές μάρκετινγκ.

ii) **Ανίχνευση προτύπων και αποκλίσεων:** Με τη μελέτη δεδομένων σε βάθος χρόνου, μπορούν να εντοπιστούν επαναλαμβανόμενα μοτίβα, αλλά και απροσδόκητες αλλαγές που μπορεί να οφείλονται σε συγκεκριμένα γεγονότα ή σφάλματα.

iii) **Διαχείριση κινδύνου**: Οι επιχειρήσεις και οι οργανισμοί μπορούν να αναγνωρίσουν πιθανούς κινδύνους και να λάβουν προληπτικά μέτρα για να τους μετριάσουν.

iv) **Στρατηγικός σχεδιασμός**: Η κατανόηση της εξέλιξης των δεδομένων με την πάροδο του χρόνου βοηθά στη λήψη στρατηγικών αποφάσεων σε τομείς όπως η οικονομία, η υγεία και η εκπαίδευση.

v) **Ανταγωνιστικό πλεονέκτημα**: Η ανάλυση χρονοσειρών επιτρέπει στις επιχειρήσεις να προσαρμόζονται γρήγορα στις αλλαγές, να βελτιώνουν τη διαχείριση των πόρων τους και να παραμένουν μπροστά από τον ανταγωνισμό [29].

Βασικά Στοιχεία Δεδομένων Χρονοσειρών

i) **Τάση**: Αναφέρεται στη γενική κατεύθυνση των δεδομένων με την πάροδο του χρόνου. Μπορεί να είναι ανοδική, καθοδική ή σταθερή.

ii) **Εποχικότητα**: Είναι τα επαναλαμβανόμενα μοτίβα που εμφανίζονται σε συγκεκριμένα χρονικά διαστήματα.

iii) **Κυκλικές διακυμάνσεις**: Αυτές αφορούν μεγαλύτερες χρονικές περιόδους και σχετίζονται με οικονομικούς ή επιχειρηματικούς κύκλους.

iv) **Τυχαίοι παράγοντες (θόρυβος)**: Πρόκειται για απρόβλεπτες διακυμάνσεις που δεν συνδέονται με κάποια τάση ή εποχικότητα. Μπορεί να οφείλονται σε εξωτερικούς παράγοντες, σφάλματα μέτρησης ή τυχαία γεγονότα.

Η ανάλυση χρονοσειρών είναι ένα ισχυρό εργαλείο που βοηθά στην κατανόηση της εξέλιξης δεδομένων στον χρόνο, επιτρέποντας τη λήψη τεκμηριωμένων αποφάσεων και την πρόβλεψη μελλοντικών αλλαγών [30].

2.3. Decision Tree

Ένα Δέντρο Απόφασης απαρτίζει μία δομή διαμερισμάτων αναδρομικού χαρακτήρα, τα οποία υποστηρίζουν αποφάσεις που χειρίζεται ένα μοντέλο αποφάσεων. Η δομή αυτή, παρομοιάζεται με ένα δέντρο, αλλά και με τις ενδεχόμενες συνέπειές τους, στα οποία περιλαμβάνονται τα αποτελέσματα τυχαίων γεγονότων, της χρησιμότητας και του κόστους των πόρων. Αυτό συγκροτεί έναν τρόπο εμφάνισης αλγορίθμου που εμπεριέχει αποκλειστικά υπό όρους εντολές ελέγχου [31].

Το Δέντρο Απόφασης απαρτίζει ένα ισχυρό και δημοφιλές εργαλείο, το οποίο χρησιμοποιείται σε αρκετούς τομείς, παραδείγματος χάριν Στατιστική, Εξόρυξη Δεδομένων και Μηχανική Μάθηση, προσφέροντας ένα διαισθητικό και σαφή τρόπο ώστε να λαμβάνονται αποφάσεις βάσει δεδομένων, μοντελοποιώντας κατ' αυτόν τον τρόπο σχέσεις μεταξύ διαφορετικών μεταβλητών [32]**.** Επί προσθέτως, αποτελεί έναν μη παραμετρικό αλγόριθμο μάθησης με επίβλεψη, ο οποίος είναι ευρέως γνωστός για τη χρήση του τόσο σε θέματα παλινδρόμησης, όσο και σε θέματα

ταξινόμησης. Κατέχει επίσης, μία ιεραρχική δενδρική δομή, που απαρτίζεται από έναν κόμβο, τη λεγόμενη ρίζα, κλαδιά, κόμβους φύλλων, αλλά και εσωτερικών κόμβων.

Όπως αναπαρίσταται και στο παρακάτω διάγραμμα, ένα Δέντρο Απόφασης φέρει την όψη ενός διαγράμματος ροής, και χρησιμοποιείται προκειμένου να λαμβάνονται αποφάσεις και προβλέψεις. Απαρτίζεται από κόμβους που αντιπροσωπεύουν ποικίλες αποφάσεις ή δοκιμές χαρακτηριστικών, κλάδους που αντιπροσωπεύουν το αποτέλεσμα των εν λόγω αποφάσεων, και κόμβους φύλλων που αντιπροσωπεύουν εν τέλει τις προβλέψεις και τα αποτελέσματα. Κάθε εσωτερικός κόμβος αντιστοιχεί σε μία δοκιμή, κάθε κλάδος στο αποτέλεσμα της δοκιμής, ενώ κάθε κόμβος φύλλου σε μία συνεχή τιμή ή ετικέτα κλάσης [32]**.** Επίσης, έχει ως αφετηρία τη ρίζα, η οποία ωστόσο δε διαθέτει άλλους εισερχόμενους κόμβους. Όμως, οι κόμβοι που εξέρχονται της ρίζας, τροφοδοτούν τους υπόλοιπους κόμβους που εισέρχονται, τους λεγόμενους κόμβους απόφασης. Βάσει των διαθέσιμων χαρακτηριστικών, και οι δυο τύποι κόμβων, έχουν τη δυνατότητα να πραγματώνουν αξιολογήσεις, με σκοπό να σχηματιστούν ομοιογενή υποσύνολα, και να συμβολίζονται με τερματικούς κόμβους ή κόμβους φύλλων. Όλοι οι κόμβοι φύλλων εκφράζουν όλα τα ενδεχόμενα αποτελέσματα εντός του συνόλου δεδομένων [33].

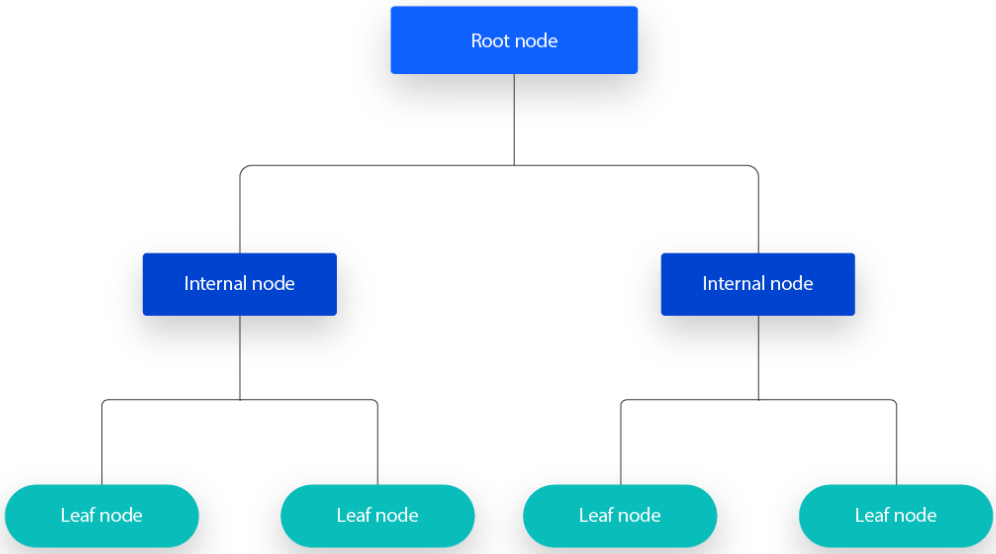
Δομή Δέντρου Απόφασης:

i) **Κόμβος Ρίζας:** Αποτελεί ολόκληρο το σύνολο δεδομένων και την αρχική απόφαση που οφείλει να ληφθεί

ii) **Εσωτερικοί Κόμβοι:** Συγκροτούν δοκιμές ή αποφάσεις σε χαρακτηριστικά. Κάθε εσωτερικός κόμβος περιλαμβάνει έναν ή περισσότερα κλαδιά.

iii) **Κλαδιά:** Απαρτίζουν το αποτέλεσμα μιας δοκιμής ή απόφασης, οδηγώντας σε έναν άλλο κόμβο.

iv) **Κόμβοι Φύλλων:** Αντιπροσωπεύουν την τελική πρόβλεψη ή απόφαση. Στους συγκεκριμένους κόμβους δεν πραγματώνονται περαιτέρω διασπάσεις [32]**.**



Εικόνα 11: Δομή Διαγράμματος Ροής

Ο παραπάνω τύπος δομής ενός διαγράμματος ροής, παράγει μία εύπεπτη αναπαράσταση λήψης αποφάσεων, διευκολύνοντας έτσι στις ποικίλες ομάδες ενός οργανισμού, να αντιληφθούν όσο το δυνατόν καλύτερα, για ποιο λόγο έλαβαν την απόφαση.

Η εκμάθηση ενός Δέντρου Απόφασης, απαιτεί μία στρατηγική τύπου «Διαίρει & Βασίλευε», υλοποιώντας μία άπληστη αναζήτηση, ώστε να εντοπιστούν τα βέλτιστα σημεία διαχωρισμού εντός ενός δέντρου. Η συγκεκριμένη διαδικασία διάσπασης επαναλαμβάνεται από πάνω προς τα κάτω, με τη βοήθεια ενός αναδρομικού τύπου, μέχρις ότου όλα τα έγγραφα, ή η πλειοψηφία τους, να αρχειοθετηθεί υπό συγκεκριμένες ετικέτες κλάσης. Η πολυπλοκότητα του δέντρου απόφασης, είναι πρωταρχικός παράγοντας, μιας και από αυτόν εξαρτάται η ταξινόμηση όλων των σημείων δεδομένων, ως ομοιογενή σύνολα. Αντιθέτως, τα μικρότερα δέντρα, είναι ευκολότερα να επιτύχουν αμιγείς κόμβους φύλλων, ειδάλλως σε μία μόνο κλάση να υπάρχουν αρκετά σημεία δεδομένων. Παρ' όλα αυτά, όταν ένα δέντρο απόφασης μεγαλώνει σε μέγεθος, γίνεται πιο δύσκολη η διατήρηση αυτής της καθαρότητας, με αποτέλεσμα να συμπίπτουν ελάχιστα δεδομένα σε ένα δεδομένο υποδέντρο. Το φαινόμενο αυτό, γνωστό και ως κατακερματισμός δεδομένων, ελλοχεύει τον κίνδυνο υπερπροσαρμογής. Απόρροια αυτού, τα δέντρα απόφασης να προτιμούν άλλα μικρά δέντρα, κάτι που συνάδει με την αρχή της φειδωλότητας του ξυραφιού του Όκαμ: «Οι οντότητες δε θα πρέπει να πολλαπλασιάζονται πέραν κάθε ανάγκης». Είναι αξιοσημείωτο πως τα

Δέντρα Απόφασης, οφείλουν να προσθέτουν πολυπλοκότητα, μόνο εφόσον είναι αναγκαίο, μιας και η πιο απλή εξήγηση συνήθως είναι και η βέλτιστη. Για να επιτευχθεί η μείωση της πολυπλοκότητας και η αποφυγή της υπερβολικής προσαρμογής, πραγματώνεται συνήθως το λεγόμενο «κλάδεμα», κατά το οποίο αφαιρούνται κλαδιά που διαχωρίζονται σε χαρακτηριστικά χαμηλής σημασίας. Επιπλέον, η προσαρμογή του μοντέλου αξιολογείται στη συνέχεια με τη βοήθεια της διασταυρούμενης επικύρωσης. Τέλος, ένας ακόμη τρόπος όπου τα δέντρα απόφασης διατηρούν την ακρίβεια τους, είναι ο σχηματισμός συνόλων, μέσα από αλγορίθμους τυχαίων δασών, κατά τον οποίο ο ταξινομητής προβλέπει πιο ακριβή αποτελέσματα, κυρίως όταν τα δέντρα που είναι μεμονωμένα, δε συσχετίζονται μεταξύ τους.

Τη δεκαετία του 1960, αναπτύχθηκε ο αλγόριθμος του Hunt, ο οποίος κάνει λόγο για τη μοντελοποίηση της ανθρώπινης μάθησης στην Ψυχολογία, και απαρτίζει τη βάση αρκετών δημοφιλών αλγορίθμων Δέντρων Απόφασης. Ενδεικτικά, μερικοί από αυτούς είναι οι κάτωθι:

i) **ID3 - Iterative Dichotomiser 3:** Ο αλγόριθμος αυτός, πιστώθηκε από τον Ross Quinlan, και αξιοποιεί το κέρδος και την εντροπία της πληροφορίας, ως μετρικές προκειμένου να αξιολογηθούν οι υποψήφιοι διαχωρισμοί.

ii) **C4.5:** Επέκταση του αλγορίθμου ID3, και αναπτύχθηκε επίσης από τον Ross Quinlan. Έχει τη δυνατότητα να αξιολογήσει σημεία διαχωρισμού εντός των δέντρων αποφάσεων, με τη χρήση αναλογιών κέρδους ή κέρδους πληροφορίας.

iii) **CART - Classification And Regression Trees – Δέντρα Ταξινόμησης και Παλινδρόμησης:** Ο δεδομένος αλγόριθμος εισήχθη από τον Leo Breiman, και συχνά κάνει χρήση της ακαθαρσίας Gini, ώστε να προσδιοριστεί το ιδανικό χαρακτηριστικό γι' αυτόν το διαχωρισμό. Η Gini προσμετρά πόσο συχνά, ένα εντελώς τυχαίο χαρακτηριστικό ταξινομείται λανθασμένα. Όσο χαμηλότερη είναι η τιμή Gini, τόσο ιδανικότερη είναι [33].

Όπως προαναφέρθηκε, ένα Δέντρο Απόφασης, έχει τη δομή ενός διαγράμματος ροής, όπου κάθε κόμβος εκπροσωπεί κάποια δοκιμή σ' ένα χαρακτηριστικό. Κάθε τομέας, με τη σειρά του, εκφράζει το αποτέλεσμα αυτής της δοκιμής, και κάθε κόμβος -φύλλο- συνιστά μία ετικέτα κλάσης, στην οποία η απόφαση παίρνεται κατόπιν υπολογισμού όλων αυτών των

χαρακτηριστικών. Σε κάθε διαδρομή, από τη ρίζα μέχρι το φύλλο, εφαρμόζονται οι κανόνες ταξινόμησης.

Στην ανάλυση των αποφάσεων, τα δέντρα απόφασης και τα στενά συνδεδεμένα διαγράμματα ροής, χρησιμοποιούνται σαν αναλυτικά και οπτικά εργαλεία που υποστηρίζουν αποφάσεις και κατ' αυτόν τον τρόπο υπολογίζονται όλες οι προσδοκώμενες τιμές, ή η επικείμενη χρησιμότητα, των εναλλακτικών, και δει, των ανταγωνιστικών.

Ένα Δέντρο Απόφασης συγκροτείται από κόμβους τριών τύπων:

i) **Κόμβοι Ευκαιρίας**, όπου εκπροσωπούνται από κύκλους.
ii) **Κόμβοι Απόφασης**, που εκφράζονται από τετράγωνα.
iii) **Τελικοί Κόμβοι**, οι οποίοι συνίστανται από τρίγωνα.

Τα Δέντρα Απόφασης χρησιμοποιούνται συχνά στη διαχείριση λειτουργιών και στην επιχειρησιακή έρευνα. Στην πραγματικότητα, οι αποφάσεις οφείλουν να υφίστανται ηλεκτρονική λήψη, χωρίς να ανακαλούνται από ελλιπή γνώση. Ένα Δέντρο Απόφασης πρέπει να παρομοιάζεται με ένα μοντέλο από πιθανότητες, σαν έναν αλγόριθμο μοντέλου επιλογής ή ένα μοντέλο καλύτερης επιλογής σε απευθείας σύνδεση. Μία ακόμη χρήση των εν λόγω δέντρων, είναι σα μέσο περιγραφής, ώστε να υπολογίζονται οι πιθανότητες υπό όρους [31].

Η διαδικασία δημιουργίας ενός Δέντρου Απόφασης, περιλαμβάνει τα εξής βήματα:

i) **Επιλογή Καλύτερου Χαρακτηριστικού:** Για το διαχωρισμό των δεδομένων, πρέπει να επιλεχθεί το καλύτερο χαρακτηριστικό, και αυτό συμβαίνει με το να χρησιμοποιηθεί μία μετρική, όπως είναι φερειπείν το κέρδος πληροφορίας, η Gini, ή η εντροπία.
ii) **Διαχωρισμός Συνόλου Δεδομένων:** Βάσει του επιλεγμένου χαρακτηριστικού, το σύνολο δεδομένων χωρίζεται σε διάφορα υποσύνολα.
iii) **Επανάληψη Διαδικασίας:** Η παραπάνω διαδικασία επαναλαμβάνεται αναδρομικά για κάθε υποσύνολο, δημιουργώντας έτσι ένα νέο φύλλο ή εσωτερικό κόμβο, μέχρις ότου να ικανοποιηθεί ένα κριτήριο διακοπής. Δηλαδή, όλα τα παραδείγματα σε έναν κόμβο να επιτυγχάνουν ένα προκαθορισμένο βάρος ή να ανήκουν στην ίδια κλάση.

Ενδεικτικά παρατίθενται μερικά από τα πλεονεκτήματα που φέρουν τα Δέντρα Απόφασης:

i) **Ερμηνευτικότητα και Απλότητα:** Τα Δέντρα Απόφασης είναι εύκολα στην ερμηνεία και την κατανόηση. Η οπτική αναπαράστασή τους, αντικατοπτρίζεται στενά στις ανθρώπινες διαδικασίες λήψης αποφάσεων.

ii) **Ευελιξία:** Μπορούν να χρησιμοποιηθούν τόσο για παλινδρόμηση, όσο και για εργασίες ταξινόμησης.

iii) **Κλιμάκωση Χαρακτηριστικών:** Τα Δέντρα Απόφασης, δεν απαιτούν κάποια κλιμάκωση ή κανονικοποίηση δεδομένων.

iv) **Μη-Γραμμικές Σχέσεις:** Προσφέρεται η δυνατότητα καταγραφής Μη-Γραμμικών Σχέσεων μεταξύ μεταβλητών-στόχων και χαρακτηριστικών.

Κάτωθι αναφέρονται κάποια από τα μειονεκτήματα που έχουν τα Δέντρα Απόφασης:

i) **Υπερπροσαρμογή:** Τα Δέντρα Απόφασης μπορούν με ευκολία να υπερπροσαρμόσουν τα εκπαιδευτικά δεδομένα, κυρίως αν είναι βαθιά συνδυαστικά με πολλούς κόμβους.

ii) **Αστάθεια:** Ορισμένες μεταβολές δεδομένων, ενδεχομένως να οδηγήσουν στη δημιουργία εντελώς διαφορετικών δέντρων.

iii) **Μεροληψία:** Τα χαρακτηριστικά που φέρουν περισσότερα επίπεδα, έχουν περισσότερες πιθανότητες να κυριαρχήσουν στη δομή του δέντρου.

Οι τεχνικές κλαδέματος, είναι ευρέως γνωστές, προκειμένου να ξεπεραστούν τα φαινόμενα της υπερπροσαρμογής. Το κλάδεμα έχει την ικανότητα μείωσης μεγέθους του δέντρου, με το να αφαιρούνται κόμβοι, παρέχοντας μικρή ισχύ στην ταξινόμηση περιπτώσεων. Οι τύποι κλαδέματος είναι δύο:

i) **Προ-Κλάδεμα – Pre-Pruning (Early Stopping):** Σταματά την ανάπτυξη του δέντρου, αμέσως μόλις εκπληρώσει κάποια ορισμένα κριτήρια, όπως είναι ενδεικτικά ο ελάχιστος αριθμός δειγμάτων ανά φύλλο και το μέγιστο βάθος.

ii) **Μετά-Κλάδεμα – Post-Pruning:** Αφαιρεί κλαδιά, από ένα πλήρως ανεπτυγμένο δέντρο, τα οποία ωστόσο δεν παρέχουν σημαντική ισχύ [32].

2.4. Random Forest

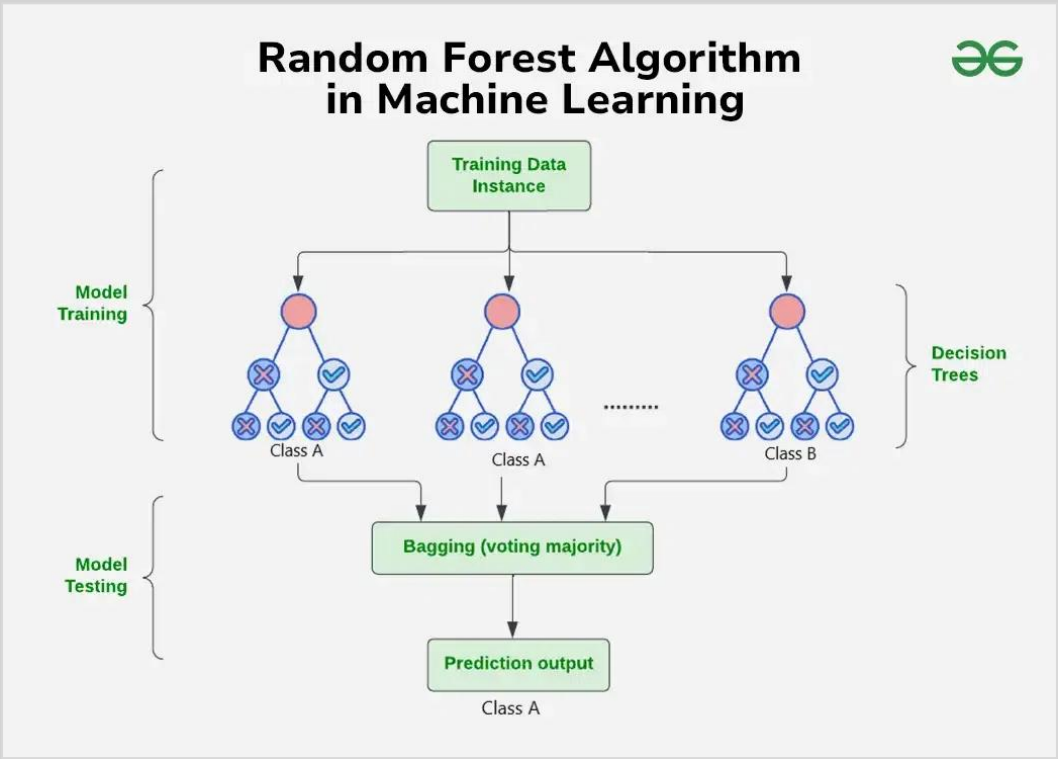
Όπως προαναφέρθηκε, η Μηχανική Μάθηση είναι ένα συνονθύλευμα της επιστήμης της στατιστικής και των υπολογιστών, σημείωσε σημαντική πρόοδο με έναν αλγόριθμο εν ονόματι Random Forest. Τα τυχαία Δέντρα Απόφασης ή τα Τυχαία Δάση, απαρτίζουν μία συνεργατική ομάδα Δέντρων Απόφασης, που συνεργάζονται ώστε να παράξουν μία ενιαία έξοδο. Έκανε την έναρξή του το 2001 μέσω του Leo Breimen [34], και εντέλει κατοχυρώθηκε από τους Leo Breiman και Adele Cutler, ο οποίος συνδυάζει την έξοδο πολλαπλών δέντρων απόφασης για να καταλήξει σε ένα ενιαίο αποτέλεσμα. Η ευκολία χρήσης και η ευελιξία του είναι οι λόγοι που έχουν τροφοδοτήσει την υιοθέτησή του, καθώς χειρίζεται τόσο προβλήματα ταξινόμησης όσο και παλινδρόμησης, και κάπως έτσι το Random Forest έγινε ακρογωνιαίος λίθος για τους λάτρεις της Μηχανικής Μάθησης [35].

Ο αλγόριθμος Random Forest απαρτίζει μία ισχυρή τεχνική εκμάθησης δέντρων στον τομέα της Μηχανικής Μάθησης. Λειτουργεί δημιουργώντας έναν αριθμό Δέντρων Απόφασης, σύμφωνα με τη φάση της εκπαίδευσης. Κάθε δέντρο κατασκευάζεται με τη χρήση ενός τυχαίου υποσυνόλου του συνόλου δεδομένων, ώστε να μετρηθεί το τυχαίο υποσύνολο χαρακτηριστικών. Η εν λόγω τυχαιότητα εισάγει τη μεταβλητότητα ανάμεσα στα μεμονωμένα δέντρα, βελτιώνοντας τη συνολική απόδοση πρόβλεψης και μειώνοντας κατ’ αυτόν τον τρόπο τον κίνδυνο υπερπροσαρμογής.

Κατά τη διαδικασία της πρόβλεψης, ο αλγόριθμος συγκεντρώνει τα αποτελέσματα όλων των δέντρων, είτε με το Μέσο Όρο -για εργασίες παλινδρόμησης-, είτε με ψηφοφορία -για εργασίες ταξινόμησης-. Η δεδομένη συνεργατική διαδικασία λήψης αποφάσεων , υποστηρίζεται από πολλαπλά δέντρα με τις γνώσεις τους, παρέχοντας ένα παράδειγμα ακριβώς και σταθερών αποτελεσμάτων. Επιπλέον, ο αλγόριθμος Random Forest, χρησιμοποιείται ευρέως για λειτουργίες παλινδρόμησης και ταξινόμησης, τα οποία φημίζονται για την ικανότητα τους στο χειρισμό πολύπλοκων δεδομένων, στη μείωση υπερπροσαρμογής και στην παροχή αξιόπιστων προβλέψεων σε ποικίλα περιβάλλοντα [34].

Συνοψίζοντας, Ο αλγόριθμος Random Forest αποτελεί επέκταση της μεθόδου bagging, καθώς χρησιμοποιεί τόσο το bagging όσο και την τυχαιότητα των χαρακτηριστικών για τη δημιουργία ενός ασυσχέτιστου δάσους Δέντρων Απόφασης. Η τυχαιότητα των χαρακτηριστικών, γνωστή και

ως feature bagging ή «The Random Subspace Method», δημιουργεί ένα τυχαίο υποσύνολο χαρακτηριστικών, το οποίο εξασφαλίζει χαμηλή συσχέτιση μεταξύ των δέντρων απόφασης. Αυτή είναι η βασικότερη διαφορά μεταξύ των Δέντρων Απόφασης και των Τυχαίων Δασών. Αν και τα Δέντρα Απόφασης εξετάζουν όλες τις πιθανές διαχωρίσεις χαρακτηριστικών, τα Τυχαία Δάση επιλέγουν μόνο ένα υποσύνολο αυτών των χαρακτηριστικών [35].



Εικόνα 12: Λειτουργία Random Forest

Λειτουργία Αλγορίθμου Random Forest:

Βήμα 1ο: Επιλογή τυχαίων K σημείων δεδομένων από το σύνολο εκπαίδευσης.

Βήμα 2ο: Δημιουργία Δέντρων Απόφασης που σχετίζονται με τα επιλεγμένα σημεία δεδομένων (υποσύνολα).

Βήμα 3ο: Επιλογή αριθμού N για τα Δέντρα Απόφασης που ο χρήστης επιθυμεί να κατασκευάσει.

Βήμα 4ο: Επανάληψη Βημάτων 1 & 2.

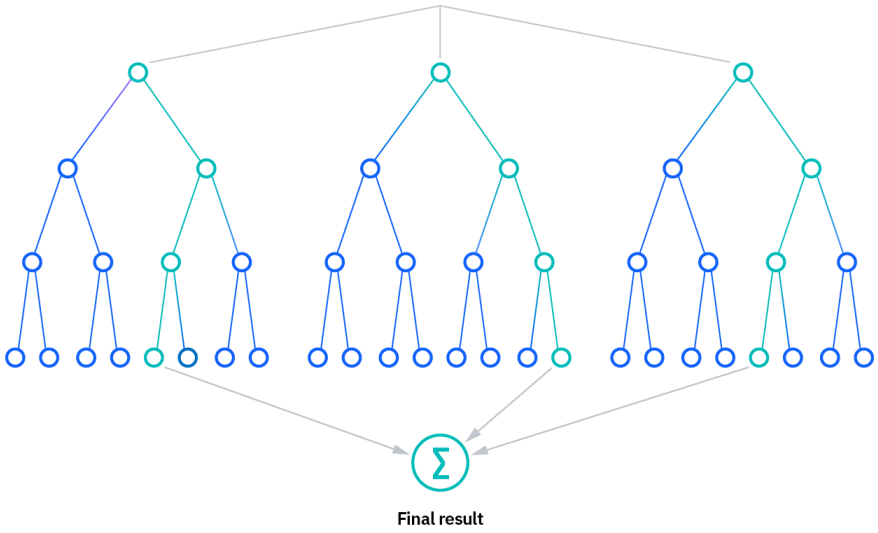
Βήμα 5ο: Για νέα σημεία δεδομένων, ο χρήστης βρίσκει τις προβλέψεις κάθε δέντρου απόφασης και αναθέτει τα νέα σημεία δεδομένων στην κατηγορία που κερδίζει στην πλειοψηφία των ψήφων.

Ο αλγόριθμος Random Forest, υλοποιείται σε διάφορα βήματα. Τα βήματα αυτά παρατίθενται παρακάτω:

- ✓ **Σύνολο Δέντρων Απόφασης:** Κατασκευάζεται ένας μεγάλος αριθμός από Δέντρα Απόφασης. Τα δέντρα αυτά είναι σαν μεμονωμένοι εμπειρογνώμονες, καθένα από τα οποία ειδικεύεται σε μια συγκεκριμένη πτυχή των δεδομένων. Είναι αξιοσημείωτο ότι λειτουργούν ανεξάρτητα, ελαχιστοποιώντας έτσι τον κίνδυνο, το μοντέλο να επηρεαστεί υπερβολικά από τις αποχρώσεις ενός μεμονωμένου δέντρου.
- ✓ **Τυχαία επιλογή χαρακτηριστικών:** Για να διασφαλιστεί ότι κάθε δέντρο απόφασης στο σύνολο του φέρνει μια μοναδική προοπτική, το Random Forest χρησιμοποιεί τυχαία επιλογή χαρακτηριστικών. Κατά τη διάρκεια της εκπαίδευσης κάθε δέντρου, επιλέγεται ένα τυχαίο υποσύνολο χαρακτηριστικών. Η τυχαιότητα αυτή διασφαλίζει ότι κάθε δέντρο επικεντρώνεται σε διαφορετικές πτυχές των δεδομένων, προωθώντας ένα ποικίλο σύνολο προβλέψεων στο σύνολο.
- ✓ **Συγκέντρωση Bootstrap ή Bagging:** Η τεχνική του bagging αποτελεί ακρογωνιαίο λίθο της στρατηγικής εκπαίδευσης του Random Forest, περιλαμβάνοντας τη δημιουργία πολλαπλών δειγμάτων bootstrap από το αρχικό σύνολο δεδομένων, επιτρέποντας έτσι τη δειγματοληψία περιπτώσεων με αντικατάσταση. Αυτό έχει ως αποτέλεσμα, διαφορετικά υποσύνολα δεδομένων για κάθε δέντρο απόφασης, εισάγοντας μεταβλητότητα στη διαδικασία εκπαίδευσης και καθιστώντας το μοντέλο πιο εύρωστο.
- ✓ **Λήψη αποφάσεων και ψηφοφορία:** Όταν πρόκειται να γίνουν προβλέψεις, κάθε δέντρο απόφασης στο Random Forest δίνει την ψήφο του. Για εργασίες ταξινόμησης, η τελική πρόβλεψη καθορίζεται από τον τρόπο (πιο συχνή πρόβλεψη) σε όλα τα δέντρα. Στις εργασίες παλινδρόμησης, λαμβάνεται ο Μέσος Όρος των προβλέψεων των μεμονωμένων δέντρων. Ο εν λόγω εσωτερικός μηχανισμός ψηφοφορίας εξασφαλίζει μια ισορροπημένη και συλλογική διαδικασία λήψης αποφάσεων [34].

Επιπλέον, είναι αξιοσημείωτο να αναφερθεί ότι οι αλγόριθμοι Τυχαίου Δάσους περιέχει τρεις κύριες υπερπαραμέτρους, οι οποίες πρέπει να οριστούν πριν από την εκπαίδευση. Αυτές περιλαμβάνουν το μέγεθος των κόμβων, τον αριθμό των δέντρων και τον αριθμό των δειγματοληπτικών χαρακτηριστικών. Από εκεί και πέρα, ο ταξινομητής Τυχαίου Δάσους μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων παλινδρόμησης ή ταξινόμησης.

Ο αλγόριθμος Random Forest αποτελείται από μια συλλογή δέντρων απόφασης και κάθε δέντρο στο σύνολο αποτελείται από ένα δείγμα δεδομένων που αντλείται από ένα σύνολο εκπαίδευσης με αντικατάσταση, το οποίο ονομάζεται δείγμα bootstrap. Από αυτό το δείγμα εκπαίδευσης, το ένα τρίτο αυτού του δείγματος τίθεται στην άκρη ως δεδομένα δοκιμής, γνωστό ως δείγμα out-of-bag (oob). Ένα άλλο παράδειγμα τυχαιότητας εισάγεται στη συνέχεια μέσω του feature bagging, προσθέτοντας μεγαλύτερη ποικιλομορφία στο σύνολο δεδομένων και μειώνοντας τη συσχέτιση μεταξύ των δέντρων απόφασης. Ανάλογα με τον τύπο του προβλήματος, ο προσδιορισμός της πρόβλεψης θα διαφέρει. Για ένα έργο παλινδρόμησης, τα μεμονωμένα δέντρα απόφασης θα υπολογίσουν το μέσο όρο, ενώ για ένα έργο ταξινόμησης, η ψήφος πλειοψηφίας -δηλαδή η πιο συχνή κατηγορική μεταβλητή- θα δώσει την προβλεπόμενη κλάση. Τέλος, το δείγμα oob χρησιμοποιείται στη συνέχεια για διασταυρούμενη επικύρωση, οριστικοποιώντας την εν λόγω πρόβλεψη.



Εικόνα 13: Λειτουργία Random Forest

Υπάρχουν ορισμένα σημαντικά πλεονεκτήματα και προκλήσεις που παρουσιάζει ο αλγόριθμος τυχαίου δάσους (Random Forest), όταν χρησιμοποιείται για προβλήματα ταξινόμησης ή παλινδρόμησης. Μερικά από αυτά περιλαμβάνουν:

Βασικά πλεονεκτήματα:

i) **Κίνδυνος Υπερπροσαρμογής:** Τα δέντρα αποφάσεων διατρέχουν τον κίνδυνο υπερπροσαρμογής, καθώς τείνουν να προσαρμόζουν στενά όλα τα δείγματα εντός των δεδομένων εκπαίδευσης. Ωστόσο, όταν υπάρχει ένας ισχυρός αριθμός δέντρων απόφασης σε ένα τυχαίο δάσος, ο ταξινομητής δεν θα υπερπροσαρμόσει το μοντέλο, καθώς ο μέσος όρος των ασυσχέτιστων δέντρων μειώνει τη συνολική διακύμανση και το σφάλμα πρόβλεψης.

ii) **Ευελιξία:** Δεδομένου ότι το τυχαίο δάσος μπορεί να χειριστεί τόσο τις εργασίες παλινδρόμησης όσο και τις εργασίες ταξινόμησης με υψηλό βαθμό ακρίβειας, είναι μια δημοφιλής μέθοδος μεταξύ των επιστημονικών δεδομένων. Η ταξινόμηση σε σάκους χαρακτηριστικών καθιστά επίσης τον ταξινομητή random forest ένα αποτελεσματικό εργαλείο για την εκτίμηση ελλιπών τιμών, καθώς διατηρεί την ακρίβεια όταν ένα μέρος των δεδομένων απουσιάζει.

iii) **Προσδιορισμός Σημασίας Χαρακτηριστικών:** Το τυχαίο δάσος καθιστά εύκολη την αξιολόγηση της σημασίας των μεταβλητών ή της συμβολής τους στο μοντέλο. Υπάρχουν μερικοί τρόποι για την αξιολόγηση της σημασίας των χαρακτηριστικών. Η σημασία Gini και η μέση μείωση της ακαθαρσίας (MDI) χρησιμοποιούνται συνήθως για να μετρήσουν πόσο μειώνεται η ακρίβεια του μοντέλου όταν αποκλείεται μια συγκεκριμένη μεταβλητή. Ωστόσο, η σημαντικότητα μετάθεσης, επίσης γνωστή ως μέση μείωση της ακρίβειας (MDA), είναι ένα άλλο μέτρο σημαντικότητας. Το MDA προσδιορίζει τη μέση μείωση της ακρίβειας με την τυχαία αντιμετάθεση των τιμών των χαρακτηριστικών σε δείγματα oob [35].

Μερικά από τα βασικά χαρακτηριστικά του Random Forest, αναφέρονται και αναλύονται παρακάτω. Ακριβέστερα:

i) **Υψηλή ακρίβεια πρόβλεψης:** Ο αλγόριθμος Random Forest, θα μπορούσε να χαρακτηριστεί ως μια ομάδα μάγων λήψης αποφάσεων. Κάθε μάγος, δηλαδή κάθε δέντρο απόφασης, εξετάζει ένα μέρος του προβλήματος και μαζί υφαίνουν τις γνώσεις τους σε ένα ισχυρό μωσαϊκό πρόβλεψης. Αυτή η ομαδική εργασία οδηγεί συχνά σε ένα πιο ακριβές μοντέλο από αυτό που θα μπορούσε να επιτύχει ένας μόνο μάγος.

ii) **Αντοχή στην υπερβολική προσαρμογή:** Ο δεδομένος αλγόριθμος, είναι σαν ένας ψύχραιμος μέντορας που καθοδηγεί τους μαθητευόμενούς του, και πιο συγκεκριμένα τα δέντρα αποφάσεων. Έτσι, αντί να αφήνει κάθε μαθητευόμενο να απομνημονεύει κάθε λεπτομέρεια της εκπαίδευσής του, ενθαρρύνει μια πιο ολοκληρωμένη κατανόηση. Αυτή η προσέγγιση βοηθά στην αποφυγή υπερβολικής εμπλοκής με τα δεδομένα εκπαίδευσης, γεγονός που καθιστά το μοντέλο λιγότερο επιρρεπές στην υπερπροσαρμογή.

iii) **Χειρισμός μεγάλων συνόλων δεδομένων:** Ο Random Forest το αντιμετωπίζει σαν έμπειρος εξερευνητής με μια ομάδα βοηθών -δέντρα απόφασης-. Κάθε βοηθός αναλαμβάνει ένα μέρος του συνόλου δεδομένων, διασφαλίζοντας ότι η εξερεύνηση δεν είναι μόνο ενδελεχής αλλά και εκπληκτικά γρήγορη.

iv) **Αξιολόγηση της σημασίας των μεταβλητών:** Ο Random Forest θα μπορούσε επίσης να χαρακτηριστεί εκτός των άλλων, και ως ένας ντετέκτιβ σε μια σκηνή εγκλήματος, ο οποίος υπολογίζει ποια στοιχεία -χαρακτηριστικά- έχουν τη μεγαλύτερη σημασία. Αξιολογεί τη σημασία κάθε στοιχείου για την επίλυση της υπόθεσης, βοηθώντας στην επικέντρωση των βασικών στοιχείων που οδηγούν στις προβλέψεις.

v) **Ενσωματωμένη διασταυρούμενη επικύρωση:** Καθώς κάθε δέντρο απόφασης εκπαιδεύεται από τον Random Forest, θέτει επίσης στην άκρη μια μυστική ομάδα περιπτώσεων (εκτός σακούλας) για δοκιμή. Αυτή η ενσωματωμένη επικύρωση διασφαλίζει ότι το μοντέλο δεν είναι απλώς άριστο στην εκπαίδευση, αλλά αποδίδει καλά και σε νέες προκλήσεις.

vi) **Χειρισμός ελλιπών τιμών:** Η ζωή είναι γεμάτη αβεβαιότητες, όπως ακριβώς και τα σύνολα δεδομένων με ελλείπουσες τιμές. Το Random Forest είναι ο φίλος που προσαρμόζεται στην κατάσταση, κάνοντας προβλέψεις χρησιμοποιώντας τις

διαθέσιμες πληροφορίες. Δεν αναστατώνεται από τα κομμάτια που λείπουν, αντιθέτως εστιάζει σε αυτά που μπορεί να πει με σιγουριά.

vii) **Παραλληλοποίηση για ταχύτητα:** Θα μπορούσε κάθε δέντρο απόφασης, να παρομοιαστεί ως ένας εργάτη που αντιμετωπίζει ταυτόχρονα ένα κομμάτι ενός παζλ. Αυτή η παράλληλη προσέγγιση αξιοποιεί τη δύναμη της σύγχρονης τεχνολογίας, καθιστώντας την όλη διαδικασία ταχύτερη και αποτελεσματικότερη για τον χειρισμό έργων μεγάλης κλίμακας [34].

Βασικές προκλήσεις:

i) **Χρονοβόρα Διαδικασία:** Η επεξεργασία των δεδομένων μπορεί να είναι αργή, καθώς υπολογίζονται δεδομένα για κάθε μεμονωμένο δέντρο απόφασης.

ii) **Περισσότεροι Πόροι:** Δεδομένου ότι τα τυχαία δάση επεξεργάζονται μεγαλύτερα σύνολα δεδομένων, απαιτούν περισσότερους πόρους ώστε να αποθηκευτούν αυτά τα δεδομένα.

iii) **Πολυπλοκότητα**: Η πρόβλεψη ενός μεμονωμένου δέντρου απόφασης είναι ευκολότερο να ερμηνευτεί, εν συγκρίσει με την πρόβλεψη ενός δάσους.

Ο αλγόριθμος του τυχαίου δάσους έχει εφαρμοστεί σε διάφορους κλάδους, επιτρέποντάς τους να λαμβάνουν καλύτερες επιχειρηματικές αποφάσεις. Ορισμένες περιπτώσεις χρήσης, περιλαμβάνουν:

i) **Χρηματοοικονομικά:** Είναι ένας αλγόριθμος που προτιμάται έναντι άλλων, καθώς μειώνει τον χρόνο που δαπανάται για τη διαχείριση δεδομένων και τις εργασίες προεπεξεργασίας. Μπορεί να χρησιμοποιηθεί για την αξιολόγηση πελατών με υψηλό πιστωτικό κίνδυνο, για την ανίχνευση απάτης και για προβλήματα τιμολόγησης δικαιωμάτων προαίρεσης.

ii) **Υγειονομική περίθαλψη:** Ο αλγόριθμος Random Forest έχει εφαρμογές στην υπολογιστική βιολογία, επιτρέποντας στους γιατρούς να αντιμετωπίσουν προβλήματα όπως η ταξινόμηση γονιδιακής έκφρασης, η ανακάλυψη βιοδεικτών και ο σχολιασμός

ακολουθιών. Ως αποτέλεσμα, οι γιατροί μπορούν να κάνουν εκτιμήσεις γύρω από την ανταπόκριση των φαρμάκων σε συγκεκριμένα φάρμακα.

iii) **Ηλεκτρονικό εμπόριο:** Μπορεί να χρησιμοποιηθεί για μηχανές συστάσεων για σκοπούς διασταυρούμενων πωλήσεων [35].

2.5. Logistic Regression

Η λογιστική παλινδρόμηση χρησιμοποιείται σε διάφορους τομείς, όπως η μηχανική μάθηση, στους περισσότερους ιατρικούς τομείς και στις κοινωνικές επιστήμες. Για παράδειγμα, το Trauma and Injury Severity Score (TRISS), το οποίο χρησιμοποιείται ευρέως για την πρόβλεψη της θνησιμότητας σε τραυματισμένους ασθενείς, αναπτύχθηκε αρχικά από τους Boyd et al. χρησιμοποιώντας λογιστική παλινδρόμηση [36].

Η Λογιστική Παλινδρόμηση είναι ένας αλγόριθμος μηχανικής μάθησης με επίβλεψη, που χρησιμοποιείται για εργασίες ταξινόμησης, όπου ο στόχος είναι να προβλεφθεί η πιθανότητα ότι μια περίπτωση ανήκει ή όχι σε μια δεδομένη κλάση. Ο Logistic Regression είναι ένας στατιστικός αλγόριθμος που αναλύει τη σχέση μεταξύ δύο παραγόντων δεδομένων [34].

Ο εν λόγω αλγόριθμος μπορεί να χρησιμοποιηθεί για την πρόβλεψη του κινδύνου εμφάνισης μιας συγκεκριμένης νόσου (π.χ. διαβήτης- στεφανιαία νόσος), με βάση τα παρατηρούμενα χαρακτηριστικά ενός ασθενή (ηλικία, φύλο, δείκτης μάζας σώματος, αποτελέσματα διαφόρων αιματολογικών εξετάσεων κ.λπ.).

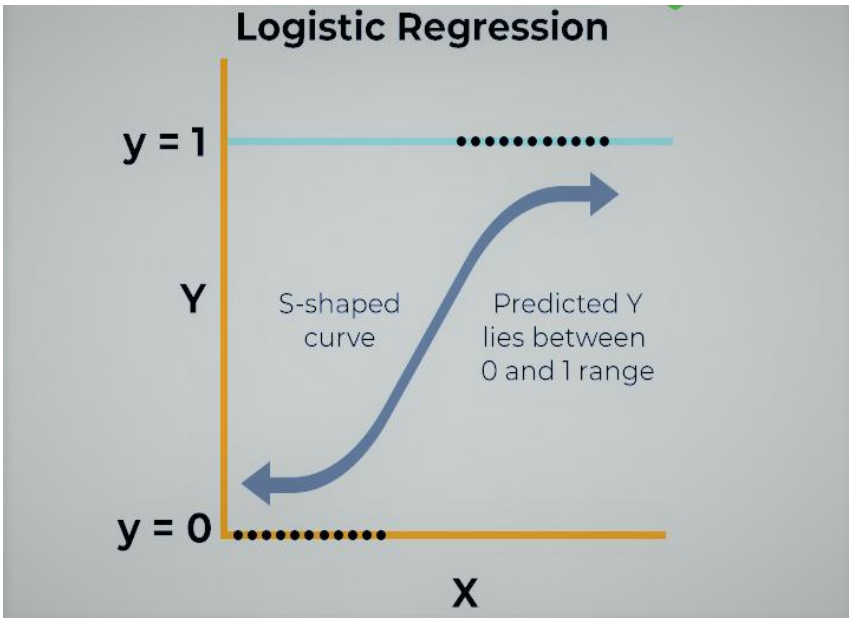
Η τεχνική αυτή μπορεί επίσης να χρησιμοποιηθεί στη Μηχανική Μάθηση, ιδίως για την πρόβλεψη της πιθανότητας αποτυχίας μιας δεδομένης διαδικασίας, ενός συστήματος ή ενός προϊόντος. Επιπροσθέτως, χρησιμοποιείται σε εφαρμογές μάρκετινγκ, όπως η πρόβλεψη της τάσης ενός πελάτη να αγοράσει ένα προϊόν ή να διακόψει μια συνδρομή. Στα οικονομικά, μπορεί να χρησιμοποιηθεί για την πρόβλεψη της πιθανότητας ένα άτομο να καταλήξει στο εργατικό δυναμικό, ενώ μια επιχειρηματική εφαρμογή θα ήταν η πρόβλεψη της πιθανότητας ένας ιδιοκτήτης σπιτιού να αθετήσει την υποθήκη του. Τα υπό συνθήκη τυχαία πεδία, μια επέκταση της λογιστικής παλινδρόμησης σε διαδοχικά δεδομένα, χρησιμοποιούνται στην επεξεργασία φυσικής γλώσσας. Οι σχεδιαστές καταστροφών και οι μηχανικοί βασίζονται σε αυτά τα μοντέλα

για να προβλέψουν τη λήψη αποφάσεων από τους ιδιοκτήτες σπιτιών ή τους ενοίκους κτιρίων σε εκκενώσεις μικρής και μεγάλης κλίμακας, όπως πυρκαγιές κτιρίων, πυρκαγιές και τυφώνες. Τα μοντέλα αυτά βοηθούν στην ανάπτυξη αξιόπιστων σχεδίων διαχείρισης καταστροφών και ασφαλέστερου σχεδιασμού για το δομημένο περιβάλλον.

Ειδικότερα, η λογιστική παλινδρόμηση είναι ένας αλγόριθμος μηχανικής μάθησης με επίβλεψη που χρησιμοποιείται ευρέως για δυαδικές εργασίες ταξινόμησης, όπως ο εντοπισμός του κατά πόσον ένα μήνυμα ηλεκτρονικού ταχυδρομείου είναι ανεπιθύμητο ή όχι και η διάγνωση ασθενειών με την αξιολόγηση της παρουσίας ή της απουσίας συγκεκριμένων καταστάσεων με βάση τα αποτελέσματα των εξετάσεων των ασθενών. Η προσέγγιση αυτή χρησιμοποιεί τη λογιστική (ή σιγμοειδή) συνάρτηση για να μετατρέψει έναν γραμμικό συνδυασμό χαρακτηριστικών εισόδου σε μια τιμή πιθανότητας που κυμαίνεται μεταξύ 0 και 1. Αυτή η πιθανότητα υποδεικνύει την πιθανότητα ότι μια δεδομένη είσοδος αντιστοιχεί σε μια από τις δύο προκαθορισμένες κατηγορίες. Ο βασικός μηχανισμός της λογιστικής παλινδρόμησης βασίζεται στην ικανότητα της λογιστικής συνάρτησης να μοντελοποιεί με ακρίβεια την πιθανότητα δυαδικών αποτελεσμάτων. Με τη χαρακτηριστική της καμπύλη σχήματος S, η λογιστική συνάρτηση αντιστοιχίζει αποτελεσματικά οποιονδήποτε αριθμό πραγματικής αξίας σε μια τιμή εντός του διαστήματος 0 έως 1. Αυτό το χαρακτηριστικό την καθιστά ιδιαίτερα κατάλληλη για εργασίες δυαδικής ταξινόμησης, όπως η ταξινόμηση των ηλεκτρονικών μηνυμάτων σε «spam» ή «μη spam». Υπολογίζοντας την πιθανότητα η εξαρτημένη μεταβλητή να κατηγοριοποιηθεί σε μια συγκεκριμένη ομάδα, η λογιστική παλινδρόμηση παρέχει ένα πιθανοτικό πλαίσιο που υποστηρίζει τη λήψη τεκμηριωμένων αποφάσεων.

Η λογιστική παλινδρόμηση χρησιμοποιείται για τη δυαδική ταξινόμηση και χρησιμοποιείται σιγμοειδή συνάρτηση, η οποία λαμβάνει είσοδο ως ανεξάρτητες μεταβλητές και παράγει μια τιμή πιθανότητας μεταξύ 0 και 1.

Για παράδειγμα, υπάρχουν δύο κλάσεις. Κλάση 0 και Κλάση 1. Εάν η τιμή της λογιστικής συνάρτησης για μια είσοδο είναι μεγαλύτερη από 0,5 (τιμή κατωφλίου) τότε ανήκει στην κλάση 1, διαφορετικά ανήκει στην κλάση 0. Αναφέρεται ως παλινδρόμηση επειδή είναι η επέκταση της γραμμικής παλινδρόμησης αλλά χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης.



Εικόνα 14: Λειτουργία Μοντέλου Logistic Regression

Βασικά σημεία:

i) Η λογιστική παλινδρόμηση προβλέπει την έξοδο μιας κατηγορικής εξαρτημένης μεταβλητής. Επομένως, το αποτέλεσμα πρέπει να είναι μια κατηγορική ή διακριτή τιμή.

ii) Μπορεί να είναι Ναι ή Όχι, 0 ή 1, αληθές ή Ψευδές, και ούτω καθεξής. Αλλά αντί να δίνει την ακριβή τιμή όπως 0 και 1, δίνει τις πιθανοτικές τιμές που βρίσκονται μεταξύ 0 και 1.

iii) Στη λογιστική παλινδρόμηση, αντί να προσαρμόζεται μια γραμμή παλινδρόμησης, προσαρμόζεται μια λογιστική συνάρτηση σχήματος «S», η οποία προβλέπει δύο μέγιστες τιμές (0 ή 1).

Με βάση τις κατηγορίες, η λογιστική παλινδρόμηση μπορεί να ταξινομηθεί σε τρεις κατηγορίες:

i) **Διωνυμική:** Στη διωνυμική λογιστική παλινδρόμηση, μπορούν να υπάρχουν μόνο δύο πιθανοί τύποι των εξαρτημένων μεταβλητών, όπως 0 ή 1, επιτυχία ή αποτυχία κ.λπ.

ii) **Πολυωνυμική:** Στην πολυωνυμική λογιστική παλινδρόμηση, μπορεί να υπάρχουν 3 ή περισσότεροι πιθανοί μη ταξινομημένοι τύποι της εξαρτημένης μεταβλητής, όπως «γάτα», «σκύλοι» ή «πρόβατα».

iii) **Ordinal:** Στην ordinal Logistic regression, μπορεί να υπάρχουν 3 ή περισσότεροι πιθανοί διατεταγμένοι τύποι εξαρτημένων μεταβλητών, όπως «χαμηλός», «μέτριος» ή «υψηλός».

Διερευνώνται οι παραδοχές της λογιστικής παλινδρόμησης, καθώς η κατανόηση αυτών των παραδοχών είναι σημαντική για τη διασφάλιση της κατάλληλης εφαρμογής του μοντέλου. Οι παραδοχές αυτές περιλαμβάνουν τα κάτωθι:

i) **Ανεξάρτητες Παρατηρήσεις:** Κάθε παρατήρηση είναι ανεξάρτητη από την άλλη. Αυτό σημαίνει ότι δεν υπάρχει συσχέτιση μεταξύ οποιωνδήποτε μεταβλητών εισόδου.

ii) **Δυαδικές Εξαρτημένες Μεταβλητές:** Θεωρείται ότι η εξαρτημένη μεταβλητή πρέπει να είναι δυαδική ή διχοτομική, δηλαδή μπορεί να λάβει μόνο δύο τιμές. Για περισσότερες από δύο κατηγορίες χρησιμοποιούνται συναρτήσεις SoftMax.

iii) **Γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών και των λογαριθμικών αποδόσεων:** Η σχέση μεταξύ των ανεξάρτητων μεταβλητών και των log odds της εξαρτημένης μεταβλητής πρέπει να είναι γραμμική.

iv) **Ακραίες Τιμές:** Δεν θα πρέπει να υπάρχουν ακραίες τιμές στο σύνολο δεδομένων.

v) **Μέγεθος Δείγματος:** Το μέγεθος του δείγματος είναι επαρκώς μεγάλο.

Παρακάτω ακολουθούν ορισμένοι συνήθεις όροι που εμπλέκονται στη λογιστική παλινδρόμηση:

i) **Ανεξάρτητες Μεταβλητές:** Τα χαρακτηριστικά εισόδου ή οι παράγοντες πρόβλεψης που εφαρμόζονται στις προβλέψεις της εξαρτημένης μεταβλητής.

ii) **Εξαρτημένη Μεταβλητή:** Η μεταβλητή-στόχος σε ένα μοντέλο λογιστικής παλινδρόμησης, η οποία πρέπει να προβλεφθεί.

iii) **Λογιστική Συνάρτηση:** Ο τύπος που χρησιμοποιείται για την αναπαράσταση του τρόπου με τον οποίο οι ανεξάρτητες και οι εξαρτημένες μεταβλητές σχετίζονται μεταξύ τους. Η λογιστική συνάρτηση μετατρέπει τις μεταβλητές εισόδου σε μια τιμή

πιθανότητας μεταξύ 0 και 1, η οποία αντιπροσωπεύει την πιθανότητα η εξαρτημένη μεταβλητή να είναι 1 ή 0.

iv) **Πιθανότητες:** Είναι ο λόγος του να συμβεί κάτι προς κάτι που δεν συμβαίνει. διαφέρει από την πιθανότητα, καθώς η πιθανότητα είναι ο λόγος του να συμβεί κάτι προς όλα όσα θα μπορούσαν ενδεχομένως να συμβούν.

v) **Λογαριθμικές Αποδόσεις:** Το log-odds, γνωστό και ως συνάρτηση logit, είναι ο φυσικός λογάριθμος των αποδόσεων. Στη λογιστική παλινδρόμηση, οι λογαριθμικές αποδόσεις της εξαρτημένης μεταβλητής μοντελοποιούνται ως γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών και της τομής.

vi) **Συντελεστής:** Οι εκτιμώμενες παράμετροι του μοντέλου λογιστικής παλινδρόμησης, δείχνουν πώς οι ανεξάρτητες και εξαρτημένες μεταβλητές σχετίζονται μεταξύ τους.

vii) **Διακοπή:** Ένας σταθερός όρος στο μοντέλο λογιστικής παλινδρόμησης, ο οποίος αντιπροσωπεύει τις λογαριθμικές πιθανότητες όταν όλες οι ανεξάρτητες μεταβλητές είναι ίσες με μηδέν.

viii) **Εκτίμηση Μέγιστης Πιθανοφάνειας:** Η μέθοδος που χρησιμοποιείται για την εκτίμηση των συντελεστών του μοντέλου λογιστικής παλινδρόμησης, η οποία μεγιστοποιεί την πιθανότητα παρατήρησης των δεδομένων με δεδομένο το μοντέλο.

Το μοντέλο λογιστικής παλινδρόμησης μετατρέπει την έξοδο συνεχών τιμών της γραμμικής συνάρτησης παλινδρόμησης σε έξοδο κατηγορικών τιμών χρησιμοποιώντας μια σιγμοειδή συνάρτηση, η οποία απεικονίζει οποιοδήποτε σύνολο ανεξάρτητων μεταβλητών εισόδου πραγματικής τιμής σε μια τιμή μεταξύ 0 και 1. Η συνάρτηση αυτή είναι γνωστή ως λογιστική συνάρτηση [37].

Πλεονεκτήματα Λογιστικής Παλινδρόμησης:

i) **Απλότητα:** Η λογιστική παλινδρόμηση είναι σχετικά απλή και εύκολη στην ερμηνεία σε σύγκριση με πιο σύνθετους αλγορίθμους όπως τα δέντρα αποφάσεων ή τα Νευρωνικά Δίκτυα. Παρέχει σαφή κατανόηση της σχέσης μεταξύ των μεταβλητών εισόδου και του προβλεπόμενου αποτελέσματος.

ii) **Ευελιξία:** Παρά το γεγονός ότι έχει σχεδιαστεί κυρίως για δυαδική ταξινόμηση, η λογιστική παλινδρόμηση μπορεί να επεκταθεί για να χειριστεί προβλήματα πολλαπλών κατηγοριών με τη χρήση τεχνικών όπως η παλινδρόμηση one-vs-rest ή softmax.

iii) **Αποτελεσματικότητα:** Η λογιστική παλινδρόμηση είναι γνωστή για την αποτελεσματικότητά της σε σενάρια όπου ο αριθμός των χαρακτηριστικών είναι σχετικά μικρός σε σύγκριση με το μέγεθος του δείγματος. Μπορεί να παρέχει αξιόπιστες προβλέψεις ακόμη και με περιορισμένα δεδομένα.

Περιορισμοί της λογιστικής παλινδρόμησης:

i) **Υπόθεση Γραμμικότητας:** Η λογιστική παλινδρόμηση υποθέτει γραμμική σχέση μεταξύ των μεταβλητών εισόδου και των λογαριθμικών πιθανοτήτων του αποτελέσματος. Εάν η πραγματική σχέση είναι μη γραμμική, η λογιστική παλινδρόμηση ενδέχεται να μην είναι σε θέση να συλλάβει αποτελεσματικά πολύπλοκα μοτίβα.

ii) **Περιορισμένες αλληλεπιδράσεις χαρακτηριστικών:** Η λογιστική παλινδρόμηση δεν μπορεί να συλλάβει τις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών, εκτός εάν περιλαμβάνονται ρητά στο μοντέλο. Αυτό μπορεί να περιορίσει την ικανότητά της να χειρίζεται πολύπλοκες σχέσεις μεταξύ μεταβλητών.

iii) **Ευαίσθητη στις ακραίες τιμές:** Οι ακραίες τιμές στο σύνολο δεδομένων μπορούν να ασκήσουν δυσανάλογο αντίκτυπο στο μοντέλο λογιστικής παλινδρόμησης, διαστρεβλώνοντας ενδεχομένως τα αποτελέσματα. Τεχνικές προεπεξεργασίας, όπως η αφαίρεση ακραίων τιμών ή η ισχυρή παλινδρόμηση, μπορούν να βοηθήσουν στον μετριασμό αυτού του ζητήματος.

Συμπερασματικά, ο αλγόριθμος λογιστικής παλινδρόμησης παρέχει μια ισχυρή και ερμηνεύσιμη προσέγγιση σε προβλήματα δυαδικής ταξινόμησης. Η ευκολία χρήσης, η ευελιξία και η αποτελεσματικότητά του τον καθιστούν απαραίτητο εργαλείο τόσο για τους επιστήμονες δεδομένων όσο και για τους επαγγελματίες. Ωστόσο, όπως κάθε αλγόριθμος, έχει τους

περιορισμούς και τις παραδοχές του που πρέπει να εξετάζονται προσεκτικά κατά την εφαρμογή του σε σενάρια του πραγματικού κόσμου [38].

2.6. K-NN

Στη στατιστική, ο αλγόριθμος k-κοντινότερων γειτόνων (k-NN) είναι μια μη παραμετρική μέθοδος μάθησης με επίβλεψη. Αναπτύχθηκε για πρώτη φορά από τους Evelyn Fix και Joseph Hodges το 1951, και αργότερα επεκτάθηκε από τον Thomas Cover. Τις περισσότερες φορές, χρησιμοποιείται για ταξινόμηση, ως ταξινομητής k-NN, η έξοδος του οποίου είναι η συμμετοχή σε μια κλάση. Ένα αντικείμενο ταξινομείται από μια πληθώρα ψήφων των γειτόνων του, με το αντικείμενο να κατατάσσεται στην κλάση που είναι πιο κοινή μεταξύ των k πλησιέστερων γειτόνων του (το k είναι ένας θετικός ακέραιος, συνήθως μικρός). Εάν k = 1, τότε το αντικείμενο απλώς κατατάσσεται στην κλάση αυτού του μοναδικού πλησιέστερου γείτονα.

Ο αλγόριθμος k-NN μπορεί επίσης να γενικευτεί για παλινδρόμηση. Στην παλινδρόμηση k-NN, επίσης γνωστή ως εξομάλυνση του πλησιέστερου γείτονα, η έξοδος είναι η τιμή της ιδιότητας για το αντικείμενο. Η τιμή αυτή είναι ο μέσος όρος των τιμών των k πλησιέστερων γειτόνων. Εάν k = 1, τότε η έξοδος ανατίθεται απλώς στην τιμή αυτού του μοναδικού πλησιέστερου γείτονα, γνωστή και ως παρεμβολή πλησιέστερου γείτονα.

Τόσο για την ταξινόμηση όσο και για την παλινδρόμηση, μια χρήσιμη τεχνική μπορεί να είναι η ανάθεση βαρών στις συνεισφορές των γειτόνων, έτσι ώστε οι πλησιέστεροι γείτονες να συνεισφέρουν περισσότερο στον μέσο όρο από ό,τι οι πιο απομακρυσμένοι. Για παράδειγμα, ένα κοινό σχήμα στάθμισης συνίσταται στο να δίνεται σε κάθε γείτονα ένα βάρος $\frac{1}{d}$, όπου d είναι η απόσταση από τον γείτονα.

Η είσοδος αποτελείται από τα k πλησιέστερα παραδείγματα εκπαίδευσης σε ένα σύνολο δεδομένων. Οι γείτονες λαμβάνονται από ένα σύνολο αντικειμένων για τα οποία είναι γνωστή η κλάση (για ταξινόμηση k-NN) ή η τιμή της ιδιότητας του αντικειμένου (για παλινδρόμηση k-NN). Αυτό μπορεί να θεωρηθεί ως το σύνολο εκπαίδευσης για τον αλγόριθμο, αν και δεν απαιτείται ρητό βήμα εκπαίδευσης.

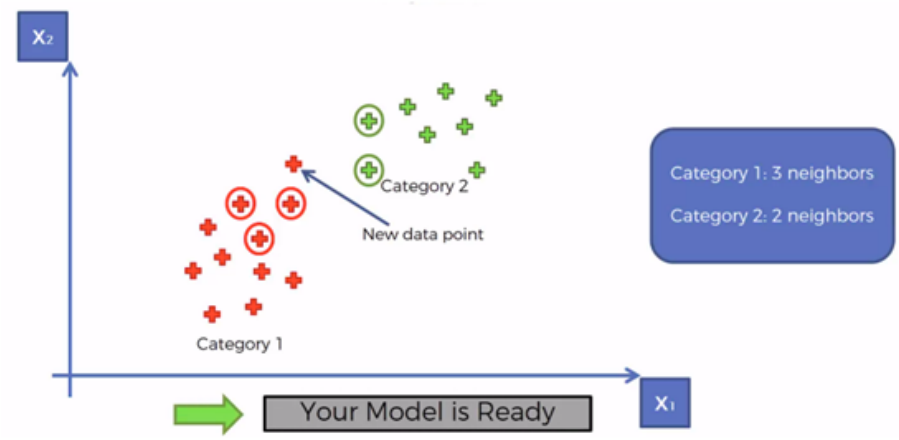
Μια ιδιαιτερότητα (μερικές φορές ακόμη και μειονέκτημα) του αλγορίθμου k-NN είναι η ευαισθησία του στην τοπική δομή των δεδομένων. Στην ταξινόμηση k-NN η συνάρτηση προσεγγίζεται μόνο τοπικά και όλος ο υπολογισμός αναβάλλεται μέχρι την αξιολόγηση της συνάρτησης. Δεδομένου ότι αυτός ο αλγόριθμος βασίζεται στην απόσταση, εάν τα χαρακτηριστικά αντιπροσωπεύουν διαφορετικές φυσικές μονάδες ή έρχονται σε πολύ διαφορετικές κλίμακες, τότε η κανονικοποίηση των δεδομένων εκπαίδευσης ως προς τα χαρακτηριστικά μπορεί να βελτιώσει σημαντικά την ακρίβειά του [39].

Η τεχνική K-NN (K-Nearest Neighbors), απαρτίζει έναν αλγόριθμο Μηχανικής Μάθησης, ο οποίος, όπως προαναφέρθηκε, χρησιμοποιείται για εργασίες τόσο ταξινόμησης, παλινδρόμησης, όσο και σε αρκετούς άλλους τομείς. Πιο συγκεκριμένα, στα χρηματοοικονομικά, ο K-NN χρησιμοποιείται ώστε να αξιολογεί πιστοληπτικές ικανότητες και να εντοπίζει απάτες με το να αναλύει πρότυπες συναλλαγές. Επίσης, επιτακτική είναι η ανάγκη στην υγειονομική περίθαλψη, με το να υλοποιεί διαγνώσεις νόσων ταξινομώντας δεδομένα ασθενών βάσει ιστορικών περιστατικών. Επιπλέον, εντοπίζει εφαρμογές σε συστήματα συστάσεων, στα οποία προτείνει υπηρεσίες ή προϊόντα, λαμβάνοντας υπόψιν συμπεριφορές και προτιμήσεις χρηστών. Τέλος, η υπολογιστική όραση και η αναγνώριση εικόνας αξιοποιούνται από τον K-NN, προκειμένου να ταξινομούνται αντικείμενα, ενώ επεξεργάζεται τη φυσική γλώσσα με την κατηγοριοποίηση κειμένου. Η αποτελεσματικότητά και η απλότητα του δεδομένου αλγορίθμου, τον κάνουν τη δημοφιλέστερη επιλογή σε ποικίλα πρακτικά σενάρια [40]. Λειτουργεί βάσει της αρχής πως παρόμοια σημεία δεδομένων έχουν περισσότερες πιθανότητες να συναντηθούν στον ίδιο χώρο χαρακτηριστικών. Ο δεδομένος αλγόριθμος μελετά τις ετικέτες γύρω από ένα σημείο δεδομένων στόχου, σε έναν επιλεγμένο αριθμό σημείων δεδομένων, ώστε να επιτευχθεί η πρόβλεψη που αφορά την κλάση που εμπίπτει με το σημείο των δεδομένων. Ο K-NN είναι πρωτίστως απλός αλγόριθμος, ωστόσο είναι και ισχυρός, γι’ αυτό το λόγο αποτελεί έναν από τους δημοφιλέστερους αλγορίθμους Μηχανικής Μάθησης.

Όταν πραγματοποιεί προβλέψεις, ο K-NN προσδιορίζει τα «K» κοντινότερα εκπαιδευτικά παραδείγματα, σε μια συγκεκριμένη είσοδο, βάσει μετρήσεων απόστασης, φερειπείν η Ευκλείδεια Απόσταση. Έπειτα, ο αλγόριθμος, στην ταξινόμηση, καταχωρεί την κοινή ή στην παλινδρόμηση, υπολογίζει το μέσο όρο των τιμών που φέρουν οι γεότονες ώστε να καθαριστεί η έξοδος. Η αποτελεσματικότητα και η απλότητα που φέρει, τον καθιστούν τη δημοφιλέστερη επιλογή

διάφορών εφαρμογές, αν και πρέπει να σημειωθεί πως ενδεχομένως να είναι μεγάλος υπολογιστικά σε τεράστια σύνολα δεδομένων.

Εν συντομία, ο K-NN συγκροτεί μία εποπτευόμενη μέθοδο Μηχανικής Μάθησης, όπου χρησιμοποιείται σε περιπτώσεις παλινδρόμησης και ταξινόμησης, χάρις στον προσδιορισμό των «K» πλησιέστερων σημείων δεδομένων στο χώρο των χαρακτηριστικών, καθώς και με τη χρήση ετικετών ή των αντίστοιχων τιμών τους, προκειμένου να επιτευχθούν προβλέψεις.



Εικόνα 15: Λειτουργία Μοντέλου k-NN

Κατά την εκτέλεση ενός K-NN αλγορίθμου, πραγματώνονται τρεις βασικές φάσεις. Αρχικά, πραγματώνεται διευθέτηση K επιλεγμένων αριθμών γειτόνων και κατόπιν υλοποιείται η εκτίμηση απόστασης ανάμεσα σε παρεχόμενα παραδείγματα – δοκιμών και παραδειγμάτων δεδομένων. Εν συνεχεία, υλοποιείται η κατηγοριοποίηση των αποστάσεων που έχουν υπολογιστεί, μέσω της αποδοχής ετικετών των K κορυφαίων καταχωρήσεων. Τέλος, πραγματοποιείται επιστροφή πρόβλεψης παραδείγματος δοκιμής.

Καταρχάς, στο πρώτο βήμα η μεταβλητή «K», διαλέγεται από τον αλγόριθμο και υποδεικνύει στον αλγόριθμο πόσα σημεία δεδομένων, δηλαδή πόσοι γείτονες, πρέπει να προσμετρηθούν στην απόδοση κρίσης για ομάδες που ανήκουν στο παράδειγμα στόχος. Εν συνεχεία του δευτέρου βήματος, η απόσταση ανάμεσα στο παράδειγμα στόχου και σε οποιοδήποτε άλλο του συνόλου δεδομένου, ελέγχεται από το μοντέλο. Έπειτα, η κάθε απόσταση προστίθεται σε μία λίστα και κατόπιν ταξινομείται, ενώ στο τέλος η ταξινομημένη λίστα περνάει από έναν έλεγχο, ενώ υλοποιείται επιστροφή ετικετών των κορυφαίων στοιχείων K. Για παράδειγμα, αν το K ορίζεται στο 5, το μοντέλο εκτελεί έλεγχο στο σημείο δεδομένων του στόχου, για ετικέτες των 5 κορυφαίων

κοντινότερων σημείων δεδομένων. Προκειμένου να αποδοθεί μία πρόβλεψη που σχετίζεται με το σημείο δεδομένων προορισμού, πρωταρχικός παράγοντας είναι αν η εργασία είναι έργο ταξινόμησης ή παλινδρόμησης. Αυτό συμβαίνει διότι, για μία εργασία ταξινόμησης χρησιμοποιείται η λειτουργία Κ κορυφαίων ετικετών, ενώ ο τρόπος του Μέσου Όρου των Κ κορυφαίων ετικετών, χρησιμοποιείται σε περιπτώσεις παλινδρόμησης. Τέλος, οι μαθηματικές πράξεις που θα χρησιμοποιηθούν στις εκτελέσεις του K-NN, εξαρτώνται άμεσα από τη μέτρηση απόστασης που έχει επιλεχθεί [41] [40].

Κατά τη χρήση του αλγορίθμου K-NN, επικρατεί η προϋπόθεση ότι μπορεί να επιλεχθεί μία λανθασμένη τιμή για το Κ. Αυτή η τιμή μπορεί να είναι είτε λάθος αριθμός γειτόνων, είτε ακατάλληλη για το Κ. Σε περίπτωση που συμβεί αυτό, κάθε πρόβλεψη που επιστρέφεται είναι εκτός λειτουργίας. Γι' αυτό το λόγο, είναι απαραίτητο κατά τη χρήση του αλγορίθμου, να προτιμάται η καταλληλότερη τιμή για το Κ. Δηλαδή, να επιλέγεται τιμή για το Κ, η οποία από τη μία να ελαχιστοποιεί τον αριθμό σφαλμάτων, ενώ από την άλλη να αυξάνει στο μέγιστο την ικανότητα του μοντέλου να προβλέπει σε δεδομένα που είναι αόρατα.

Όσο το Κ φέρει χαμηλές τιμές, τόσο ασταθής και αναξιόπιστες είναι οι προβλέψεις που παρέχει ο K-NN. Χαρακτηριστικό παράδειγμα αυτού, είναι η περίπτωση τω 7 γειτόνων, στο οποίο ο K-NN λειτουργεί με μία τιμή 2, όπου του ζητείται να κάνει μία πρόβλεψη λαμβάνοντας υπόψιν τους 2 πλησιέστερους γείτονες. Εφόσον η πλειοψηφία, πέντε στα εφτά, των γειτόνων ανήκουν στην ίδια κατηγορία, ας ονομαστεί Μπλε, και οι υπόλοιποι 2 πλησιέστεροι γείτονες ανήκουν στην Κόκκινη, η πρόβλεψη που θα υλοποιήσει το μοντέλο θα είναι η Κόκκινη κατηγορία, ενώ η καλύτερη πρόβλεψη θα ήταν η Μπλε. Αυτό συμβαίνει λόγω του ότι όσο αυξάνεται η ακτίνα , το μοντέλο εξετάζει μόνο τα σημεία που είναι κοντινότερα στο σημείο στόχος. Απότοκο αυτού η λανθασμένη ταξινόμηση και η μειωμένη ακρίβεια στην πρόβλεψη και απόδοση του αλγορίθμου [41].

Η απόδοση της ταξινόμησης του πλησιέστερου γείτονα Κ μπορεί συχνά να βελτιωθεί σημαντικά μέσω της εποπτευόμενης μετρικής μάθησης. Δημοφιλείς αλγόριθμοι είναι η ανάλυση συνιστωσών γειτονιάς και ο πλησιέστερος γείτονας μεγάλου περιθωρίου. Οι επιβλεπόμενοι αλγόριθμοι

μετρικής μάθησης χρησιμοποιούν τις πληροφορίες ετικέτας για να μάθουν μια νέα μετρική ή ψευδομετρική.

Όταν τα δεδομένα εισόδου σε έναν αλγόριθμο είναι πολύ μεγάλα για να επεξεργαστούν και υπάρχει η υποψία ότι είναι περιττά, όπως είναι παραδείγματος χάριν η ίδια μέτρηση σε πόδια και μέτρα, τότε τα δεδομένα εισόδου μετατρέπονται σε ένα σύνολο χαρακτηριστικών μειωμένης αναπαράστασης, που ονομάζεται επίσης διάνυσμα χαρακτηριστικών. Ο μετασχηματισμός των δεδομένων εισόδου σε σύνολο χαρακτηριστικών ονομάζεται εξαγωγή χαρακτηριστικών. Εάν τα χαρακτηριστικά που εξάγονται είναι προσεκτικά επιλεγμένα, αναμένεται ότι το σύνολο χαρακτηριστικών θα εξάγει τις σχετικές πληροφορίες από τα δεδομένα εισόδου προκειμένου να εκτελεστεί η επιθυμητή εργασία χρησιμοποιώντας αυτή τη μειωμένη αναπαράσταση αντί της εισόδου πλήρους μεγέθους. Η εξαγωγή χαρακτηριστικών πραγματοποιείται στα ακατέργαστα δεδομένα πριν από την εφαρμογή του αλγορίθμου k-NN στα μετασχηματισμένα δεδομένα στο χώρο των χαρακτηριστικών.

Παράδειγμα τυπικού υπολογιστικού αγωγού υπολογιστικής όρασης υπολογιστών για την αναγνώριση προσώπου με χρήση k-NN που περιλαμβάνει βήματα προεπεξεργασίας εξαγωγής χαρακτηριστικών και μείωσης διαστάσεων και συνήθως υλοποιείται με το OpenCV. Εκτενέστερα, αρχικά πραγματοποιείται ανίχνευση προσώπου Haar και ανάλυση εντοπισμού μέσης μετατόπισης, ενώ τέλος προβάλλεται PCA ή Fisher LDA στο χώρο χαρακτηριστικών, ακολουθούμενη από ταξινόμηση k-NN [40].

Παρακάτω αναλύονται τα προτερήματα και τα μειονεκτήματα που εμφανίζει ο δεδομένος αλγόριθμος.

Αναλυτικότερα, ο δεδομένος αλγόριθμος χρησιμοποιείται και για εργασίες ταξινόμησης, αλλά και για εργασίες παλινδρόμησης, εν αντιθέσει με διάφορους εποπτευόμενους αλγορίθμους μάθησης, μιας και είναι εξαιρετικά απλός και ακριβής ως προς τη χρήση του. Τέλος, είναι εύκολος ως προς την ερμηνεία, την κατανόηση και την εφαρμογή, καθώς μπορεί να χρησιμοποιηθεί σε ποικίλα προβλήματα, μιας και δεν βγάζει αυθαίρετα συμπεράσματα για τα δεδομένα.

Αντιθέτως, ο K-NN, είναι υπολογιστικά ακριβός και απαιτεί τεράστια μνήμη, αφότου αποθηκεύει την πλειοψηφία των δεδομένων του. Επίσης, τα τεράστια σύνολα δεδομένων, είναι ικανά να

κάνουν προβλέψεις υψηλής διάρκειας, αν και είναι αρκετά ευαίσθητος στην κλίμακα συνόλου δεδομένων, και άσχετα χαρακτηριστικά μπορούν να απορρίψουν εύκολα τον αλγόριθμο.

Εν κατακλείδι, ο αλγόριθμος K-NN (k-Nearest Neighbors – k-πλησιέστερων γειτόνων), συγκροτεί τον πιο απλό αλγόριθμο Μηχανικής Μάθησης. Είναι ναι μεν απλός, αλλά επίσης και ισχυρός, αφού προσδίδει υψηλή ακρίβεια σε σχεδόν όλα τα προβλήματα [41].

2.7. SVM – SVC

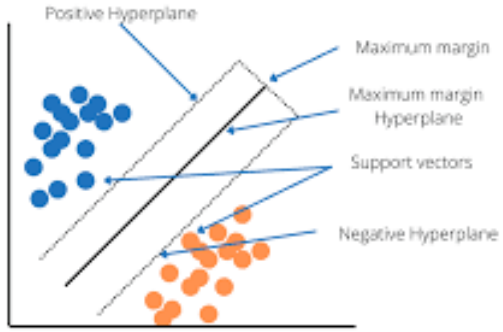
Οι Support Vector Machines συγκροτούν αλγορίθμους Επιβλεπόμενης Μάθησης και η χρήση τους περιορίζεται σε παλινδρόμηση και ταξινόμηση, με ιδιαίτερη έμφαση στην ταξινόμηση [42]. Στη μηχανική μάθηση, οι μηχανές διανυσμάτων υποστήριξης είναι εποπτευόμενα μοντέλα μέγιστου περιθωρίου με σχετικούς αλγορίθμους μάθησης που αναλύουν δεδομένα για ταξινόμηση και ανάλυση παλινδρόμησης. Οι SVM αναπτύχθηκαν στα εργαστήρια AT&T Bell Laboratories και αποτελούν ένα από τα πιο μελετημένα μοντέλα, καθώς βασίζονται σε πλαίσια στατιστικής μάθησης της θεωρίας VC που προτάθηκαν από τους Vapnik (1982, 1995) και Chervonenkis (1974).

Εκτός από την εκτέλεση γραμμικής ταξινόμησης, τα SVM μπορούν να εκτελέσουν αποτελεσματικά μη γραμμική ταξινόμηση χρησιμοποιώντας το τέχνασμα του πυρήνα, αναπαριστώντας τα δεδομένα μόνο μέσω ενός συνόλου συγκρίσεων ομοιότητας ανά ζεύγη μεταξύ των αρχικών σημείων δεδομένων με τη χρήση μιας συνάρτησης πυρήνα, η οποία τα μετατρέπει σε συντεταγμένες σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης. Έτσι, τα SVM χρησιμοποιούν το τέχνασμα του πυρήνα για να αντιστοιχίσουν σιωπηρά τις εισόδους τους σε χώρους χαρακτηριστικών υψηλών διαστάσεων, όπου μπορεί να πραγματοποιηθεί γραμμική ταξινόμηση. Τα SVM μπορούν επίσης να χρησιμοποιηθούν για εργασίες παλινδρόμησης, όπου ο στόχος γίνεται ευαίσθητος.

Ο αλγόριθμος ομαδοποίησης διανυσμάτων υποστήριξης, που δημιουργήθηκε από τους Hava Siegelmann και Vladimir Vapnik, εφαρμόζει τη στατιστική των διανυσμάτων υποστήριξης, που αναπτύχθηκε στον αλγόριθμο των μηχανών διανυσμάτων υποστήριξης, για την κατηγοριοποίηση μη επισημειωμένων δεδομένων. Αυτά τα σύνολα δεδομένων απαιτούν προσεγγίσεις μάθησης

χωρίς επίβλεψη, οι οποίες προσπαθούν να βρουν φυσική ομαδοποίηση των δεδομένων σε ομάδες και στη συνέχεια να αντιστοιχίσουν νέα δεδομένα σύμφωνα με αυτές τις ομάδες.

Η δημοτικότητα των SVMs οφείλεται πιθανότατα στην ευχέρεια θεωρητικής ανάλυσης και στην ευελιξία τους να εφαρμόζονται σε μια μεγάλη ποικιλία εργασιών, συμπεριλαμβανομένων των



Εικόνα 16: Λειτουργία Μοντέλου SVM - SVC

προβλημάτων δομημένης πρόβλεψης. Δεν είναι σαφές ότι τα SVM έχουν καλύτερη προβλεπτική απόδοση από άλλα γραμμικά μοντέλα, όπως η λογιστική παλινδρόμηση και η γραμμική παλινδρόμηση [43].

Οι μηχανές διανυσμάτων υποστήριξης (SVM) είναι ένα σύνολο μεθόδων μάθησης με επίβλεψη που χρησιμοποιούνται για ταξινόμηση, παλινδρόμηση και ανίχνευση ακραίων τιμών [44] και είναι ιδιαίτερα προσαρμόσιμες, καθιστώντας τες κατάλληλες για διάφορες εφαρμογές, όπως ταξινόμηση κειμένου, ταξινόμηση εικόνων, ανίχνευση ανεπιθύμητης αλληλογραφίας, αναγνώριση γραφής, ανάλυση γονιδιακής έκφρασης, ανίχνευση προσώπου και ανίχνευση ανωμαλιών.

Οι SVM είναι ιδιαίτερα αποτελεσματικές επειδή επικεντρώνονται στην εύρεση του μέγιστου διαχωριστικού υπερεπιπέδου μεταξύ των διαφορετικών κλάσεων στο χαρακτηριστικό-στόχο, καθιστώντας τες ανθεκτικές τόσο για δυαδική όσο και για ταξινόμηση πολλαπλών κλάσεων.

Ενώ μπορεί να εφαρμοστεί σε προβλήματα παλινδρόμησης, η SVM είναι καταλληλότερη για εργασίες ταξινόμησης. Ο πρωταρχικός στόχος του αλγορίθμου SVM είναι να προσδιορίσει το βέλτιστο υπερεπίπεδο σε έναν χώρο N διαστάσεων που μπορεί να διαχωρίσει αποτελεσματικά τα σημεία δεδομένων σε διαφορετικές κλάσεις στο χώρο χαρακτηριστικών. Ο αλγόριθμος

διασφαλίζει ότι μεγιστοποιείται το περιθώριο μεταξύ των πλησιέστερων σημείων διαφορετικών κλάσεων, γνωστών και ως διανύσματα υποστήριξης.

Η διάσταση του υπερεπιπέδου εξαρτάται από τον αριθμό των χαρακτηριστικών. Για παράδειγμα, εάν υπάρχουν δύο χαρακτηριστικά εισόδου, το υπερεπίπεδο είναι απλώς μια γραμμή, ενώ εάν υπάρχουν τρία χαρακτηριστικά εισόδου, το υπερεπίπεδο γίνεται ένα δισδιάστατο επίπεδο. Καθώς ο αριθμός των χαρακτηριστικών αυξάνεται πέραν των τριών, αυξάνεται και η πολυπλοκότητα της απεικόνισης του υπερεπιπέδου.

Θεωρούνται δύο ανεξάρτητες μεταβλητές, x_1 και x_2, και μία εξαρτημένη μεταβλητή που αναπαρίσταται είτε ως μπλε κύκλος είτε ως κόκκινος κύκλος.

- ✓ Σε αυτό το σενάριο, το υπερεπίπεδο είναι μια γραμμή επειδή υπάρχουν δύο χαρακτηριστικά, x_1 και x_2 αντίστοιχα.
- ✓ Υπάρχουν πολλαπλές γραμμές, ή υπερεπίπεδα που μπορούν να διαχωρίσουν τα σημεία δεδομένων.
- ✓ Η πρόκληση είναι να προσδιοριστεί το καλύτερο υπερεπίπεδο που μεγιστοποιεί το περιθώριο διαχωρισμού μεταξύ του κόκκινου και του μπλε κύκλου.

Πρόβλεψη εάν ο καρκίνος είναι καλοήθης ή κακοήθης. Η χρήση ιστορικών δεδομένων σχετικά με ασθενείς που έχουν διαγνωστεί με καρκίνο επιτρέπει στους γιατρούς να διακρίνουν τις κακοήθεις περιπτώσεις από τις καλοήθεις, καθώς τους δίνονται ανεξάρτητα χαρακτηριστικά.

Τα βήματα που ακολουθούνται είναι τα εξής:

i) Φόρτωση συνόλου δεδομένων για τον καρκίνο του μαστού από το sklearn.datasets
ii) Διαχωρισμός χαρακτηριστικών εισόδου και τις μεταβλητές-στόχους.
iii) Κατασκευή και εκπαίδευση ταξινομητών SVM χρησιμοποιώντας πυρήνα RBF.
iv) Σχεδιασμός διαγράμματος διασποράς χαρακτηριστικών εισόδου.
v) Σχεδιασμός ορίου απόφασης.
vi) Απεικόνιση ορίου απόφασης

Πλεονεκτήματα Μηχανής Διανυσμάτων Υποστήριξης (SVM)

i) **Απόδοση υψηλών διαστάσεων:** Η SVM υπερέχει σε χώρους υψηλών διαστάσεων, καθιστώντας την κατάλληλη για την ταξινόμηση εικόνων και την ανάλυση γονιδιακής έκφρασης.

ii) **Μη γραμμική ικανότητα:** Χρησιμοποιώντας συναρτήσεις πυρήνα όπως η RBF και το πολυώνυμο, το SVM χειρίζεται αποτελεσματικά μη γραμμικές σχέσεις.

iii) **Ανθεκτικότητα σε ακραίες τιμές:** Το χαρακτηριστικό μαλακό περιθώριο επιτρέπει στο SVM να αγνοεί τις ακραίες τιμές, ενισχύοντας την ανθεκτικότητα στην ανίχνευση ανεπιθύμητων μηνυμάτων και στην ανίχνευση ανωμαλιών.

iv) **Δυαδική και πολυταξική υποστήριξη:** Το SVM είναι αποτελεσματικό τόσο για δυαδική ταξινόμηση όσο και για ταξινόμηση πολλαπλών κατηγοριών, κατάλληλο για εφαρμογές στην ταξινόμηση κειμένου.

v) **Αποδοτικότητα μνήμης:** Ο SVM επικεντρώνεται στα διανύσματα υποστήριξης, γεγονός που τον καθιστά αποδοτικό στη μνήμη σε σύγκριση με άλλους αλγορίθμους.

Μειονεκτήματα Μηχανής Διανυσμάτων Υποστήριξης (SVM)

i) **Αργή εκπαίδευση:** Η SVM μπορεί να είναι αργή για μεγάλα σύνολα δεδομένων, επηρεάζοντας την απόδοση της SVM σε εργασίες εξόρυξης δεδομένων.

ii) **Δυσκολία συντονισμού παραμέτρων:** Η επιλογή του σωστού πυρήνα και η προσαρμογή παραμέτρων όπως το C απαιτεί προσεκτικό συντονισμό, επηρεάζοντας τους αλγορίθμους SVM.

iii) **Ευαισθησία στον θόρυβο:** Ο SVM δυσκολεύεται με θορυβώδη σύνολα δεδομένων και επικαλυπτόμενες κλάσεις, περιορίζοντας την αποτελεσματικότητα σε πραγματικές περιπτώσεις.

iv) **Περιορισμένη ερμηνευσιμότητα:** Η πολυπλοκότητα του υπερεπιπέδου σε υψηλότερες διαστάσεις καθιστά το SVM λιγότερο ερμηνεύσιμο από άλλα μοντέλα.

v) **Ευαισθησία στην κλιμάκωση χαρακτηριστικών:** Η σωστή κλιμάκωση των χαρακτηριστικών είναι απαραίτητη- διαφορετικά, τα μοντέλα SVM μπορεί να έχουν κακές επιδόσεις.

Συμπερασματικά, οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) είναι ισχυροί αλγόριθμοι στη μηχανική μάθηση, ιδανικοί τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Διακρίνονται στην εύρεση του βέλτιστου υπερεπιπέδου για το διαχωρισμό των δεδομένων, καθιστώντας τα κατάλληλα για εφαρμογές όπως η ταξινόμηση εικόνων και η ανίχνευση ανωμαλιών.

Η προσαρμοστικότητα του SVM μέσω των συναρτήσεων πυρήνα του επιτρέπει να χειρίζεται αποτελεσματικά τόσο γραμμικά όσο και μη γραμμικά δεδομένα. Ωστόσο, πρέπει να ληφθούν υπόψη προκλήσεις όπως ο συντονισμός των παραμέτρων και οι δυνητικά αργοί χρόνοι εκπαίδευσης σε μεγάλα σύνολα δεδομένων.

Η κατανόηση της SVM είναι ζωτικής σημασίας για τους επιστήμονες δεδομένων, καθώς ενισχύει την ακρίβεια πρόβλεψης και τη λήψη αποφάσεων σε διάφορους τομείς, συμπεριλαμβανομένης της εξόρυξης δεδομένων και της τεχνητής νοημοσύνης [45].

Η μέθοδος της ταξινόμησης διανυσμάτων υποστήριξης μπορεί να επεκταθεί για την επίλυση προβλημάτων παλινδρόμησης. Η μέθοδος αυτή ονομάζεται παλινδρόμηση διανυσμάτων υποστήριξης.

Το μοντέλο που παράγεται από την ταξινόμηση διανυσμάτων υποστήριξης, εξαρτάται μόνο από ένα υποσύνολο των δεδομένων εκπαίδευσης, επειδή η συνάρτηση κόστους για την κατασκευή του μοντέλου δεν ενδιαφέρεται για τα σημεία εκπαίδευσης που βρίσκονται εκτός του περιθωρίου. Κατ' αναλογία, το μοντέλο που παράγεται από τη διανυσματική παλινδρόμηση υποστήριξης εξαρτάται μόνο από ένα υποσύνολο των δεδομένων εκπαίδευσης, επειδή η συνάρτηση κόστους αγνοεί τα δείγματα των οποίων η πρόβλεψη βρίσκεται κοντά στο στόχο τους.

Επιπροσθέτως, υπάρχουν τρεις διαφορετικές υλοποιήσεις της διανυσματικής παλινδρόμησης υποστήριξης: SVR, NuSVR και LinearSVR. Η LinearSVR παρέχει μια ταχύτερη υλοποίηση από την SVR, αλλά λαμβάνει υπόψη μόνο τον γραμμικό πυρήνα, ενώ η NuSVR υλοποιεί μια ελαφρώς διαφορετική διατύπωση από την SVR και τη LinearSVR. Λόγω της υλοποίησής της στη liblinear, η LinearSVR κανονικοποιεί επίσης τη διατομή, εφόσον λαμβάνεται υπόψη. Το αποτέλεσμα αυτό μπορεί ωστόσο να μειωθεί με προσεκτική ρύθμιση της παραμέτρου intercept_scaling, η οποία επιτρέπει στον όρο intercept να έχει διαφορετική συμπεριφορά κανονικοποίησης σε σύγκριση με

τα άλλα χαρακτηριστικά. Συνεπώς, τα αποτελέσματα ταξινόμησης και η βαθμολογία μπορεί να διαφέρουν από τους άλλους δύο ταξινομητές [44].

Οι SVM μπορούν να χρησιμοποιηθούν για την επίλυση διαφόρων προβλημάτων του πραγματικού κόσμου:

- ✓ Ορισμένες μέθοδοι για ρηχή σημασιολογική ανάλυση βασίζονται σε μηχανές διανυσμάτων υποστήριξης.
- ✓ Η ταξινόμηση εικόνων μπορεί επίσης να πραγματοποιηθεί με τη χρήση SVM. Τα πειραματικά αποτελέσματα δείχνουν ότι οι SVMs επιτυγχάνουν σημαντικά υψηλότερη ακρίβεια αναζήτησης από τα παραδοσιακά σχήματα βελτίωσης ερωτημάτων μετά από μόλις τρεις έως τέσσερις γύρους ανατροφοδότησης συνάφειας. Αυτό ισχύει επίσης για τα συστήματα κατάτμησης εικόνων, συμπεριλαμβανομένων εκείνων που χρησιμοποιούν μια τροποποιημένη έκδοση SVM που χρησιμοποιεί την προνομιακή προσέγγιση όπως προτάθηκε από τον Vapnik.
- ✓ Ταξινόμηση δορυφορικών δεδομένων όπως τα δεδομένα SAR με χρήση SVM με επίβλεψη.
- ✓ Χειρόγραφοι χαρακτήρες μπορούν να αναγνωριστούν με τη χρήση SVM.

Ο αλγόριθμος SVM έχει εφαρμοστεί ευρέως στις βιολογικές και άλλες επιστήμες. Έχουν χρησιμοποιηθεί για την ταξινόμηση πρωτεϊνών με έως και 90% των ενώσεων να ταξινομούνται σωστά. Ως μηχανισμός για την ερμηνεία των μοντέλων SVM έχουν προταθεί δοκιμές μεταβολής με βάση τα βάρη SVM. Τα βάρη της μηχανής διανυσμάτων υποστήριξης έχουν επίσης χρησιμοποιηθεί στο παρελθόν για την ερμηνεία των μοντέλων SVM. Η μεταγενέστερη ερμηνεία των μοντέλων μηχανών διανυσμάτων υποστήριξης με σκοπό τον εντοπισμό των χαρακτηριστικών που χρησιμοποιούνται από το μοντέλο για την πραγματοποίηση προβλέψεων είναι ένας σχετικά νέος τομέας έρευνας με ιδιαίτερη σημασία στις βιολογικές επιστήμες [43].

2.8. Naïve Bayes

Στη στατιστική, οι ταξινομητές Naive Bayes είναι μια οικογένεια γραμμικών «πιθανοτικών ταξινομητών» που υποθέτει ότι τα χαρακτηριστικά είναι υπό όρους ανεξάρτητα, δεδομένης της κατηγορίας-στόχου [46], και βασίζονται στο θεώρημα του Bayes. Παρά την «αφελή» υπόθεση της ανεξαρτησίας των χαρακτηριστικών, αυτοί οι ταξινομητές χρησιμοποιούνται ευρέως για την απλότητα και την αποτελεσματικότητά τους στη μηχανική μάθηση [47]. Η ισχύς (αφέλεια) αυτής της παραδοχής δίνει στον ταξινομητή το όνομά του. Αυτοί οι ταξινομητές είναι από τα απλούστερα μοντέλα δικτύων Bayes.

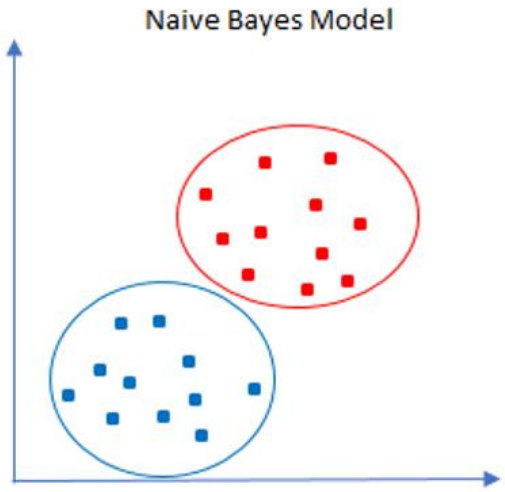
Οι ταξινομητές Naive Bayes είναι εξαιρετικά κλιμακούμενοι, απαιτώντας έναν αριθμό παραμέτρων γραμμικά προς τον αριθμό των μεταβλητών -χαρακτηριστικά/προγνωστικοί παράγοντες- σε ένα πρόβλημα μάθησης. Η εκπαίδευση μέγιστης πιθανοφάνειας μπορεί να γίνει με την αξιολόγηση μιας έκφρασης κλειστής μορφής, η οποία απαιτεί γραμμικό χρόνο, αντί για ακριβή επαναληπτική προσέγγιση όπως χρησιμοποιείται για πολλούς άλλους τύπους ταξινομητών.

Οι ταξινομητές Naive Bayes είναι μια συλλογή αλγορίθμων ταξινόμησης που βασίζονται στο θεώρημα του Bayes. Δεν πρόκειται για έναν ενιαίο αλγόριθμο αλλά για μια οικογένεια αλγορίθμων όπου όλοι τους μοιράζονται μια κοινή αρχή, δηλαδή ότι κάθε ζεύγος χαρακτηριστικών που ταξινομούνται είναι ανεξάρτητο μεταξύ τους, αρκεί να θεωρηθεί αρχικά ένα σύνολο δεδομένων.

Ένας από τους πιο απλούς και αποτελεσματικούς αλγορίθμους ταξινόμησης, ο ταξινομητής Naïve Bayes βοηθά στην ταχεία ανάπτυξη μοντέλων Μηχανικής Μάθησης με γρήγορες δυνατότητες πρόβλεψης.

Ο αλγόριθμος Naïve Bayes χρησιμοποιείται για προβλήματα ταξινόμησης. Χρησιμοποιείται επίσης, σε μεγάλο βαθμό στην ταξινόμηση κειμένου. Στις εργασίες ταξινόμησης κειμένου, τα δεδομένα περιέχουν υψηλή διάσταση, καθώς κάθε λέξη αντιπροσωπεύει ένα χαρακτηριστικό στα δεδομένα. Επιπλέον, είναι γνωστή για το φιλτράρισμα ανεπιθύμητης αλληλογραφίας, την ανίχνευση συναισθημάτων, την ταξινόμηση αξιολογήσεων κ.λπ. Το πλεονέκτημα της χρήσης του Naïve Bayes είναι η ταχύτητά του. Είναι γρήγορο και η πρόβλεψη είναι εύκολη με υψηλή διάσταση δεδομένων.

Αυτό το μοντέλο προβλέπει την πιθανότητα μια περίπτωση να ανήκει σε μια κλάση με ένα δεδομένο σύνολο τιμών χαρακτηριστικών. Είναι ένας πιθανοτικός ταξινομητής. Αυτό συμβαίνει επειδή υποθέτει ότι ένα χαρακτηριστικό στο μοντέλο είναι ανεξάρτητο από την ύπαρξη ενός άλλου χαρακτηριστικού. Με άλλα λόγια, κάθε χαρακτηριστικό συμβάλλει στις προβλέψεις χωρίς καμία σχέση μεταξύ τους. Στον πραγματικό κόσμο, αυτή η συνθήκη ικανοποιείται σπάνια. Τέλος,



Εικόνα 17: Μοντέλο Naive Bayes

είναι σημαντικό να αναφερθεί ότι γίνεται χρήση του θεωρήματος Bayes στον αλγόριθμο για την εκπαίδευση και την πρόβλεψη [47].

Στη βιβλιογραφία της στατιστικής, τα μοντέλα naive Bayes είναι γνωστά με διάφορα ονόματα, όπως simple Bayes και independence Bayes. Όλα αυτά τα ονόματα αναφέρονται στη χρήση του θεωρήματος Bayes στον κανόνα απόφασης του ταξινομητή, αλλά το naive Bayes δεν είναι απαραίτητα μια μέθοδος Bayes.

Ο αφελής Bayes είναι μια απλή τεχνική για την κατασκευή ταξινομητών: μοντέλα που αποδίδουν ετικέτες κλάσης σε περιπτώσεις προβλήματος, οι οποίες αναπαρίστανται ως διανύσματα τιμών χαρακτηριστικών, όπου οι ετικέτες κλάσης προέρχονται από κάποιο πεπερασμένο σύνολο. Δεν υπάρχει δηλαδή ένας μοναδικός αλγόριθμος για την εκπαίδευση τέτοιων ταξινομητών, αλλά μια οικογένεια αλγορίθμων που βασίζονται σε μια κοινή αρχή: όλοι οι ταξινομητές Naive Bayes υποθέτουν ότι η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι ανεξάρτητη από την τιμή οποιουδήποτε άλλου χαρακτηριστικού, δεδομένης της μεταβλητής κλάσης. Για παράδειγμα, ένα φρούτο μπορεί να θεωρηθεί μήλο εάν είναι κόκκινο, στρογγυλό και έχει διάμετρο περίπου 10 cm.

Ένας ταξινομητής naive Bayes θεωρεί ότι κάθε ένα από αυτά τα χαρακτηριστικά συμβάλλει ανεξάρτητα στην πιθανότητα ότι το φρούτο αυτό είναι μήλο, ανεξάρτητα από τις πιθανές συσχετίσεις μεταξύ των χαρακτηριστικών χρώμα, στρογγυλότητα και διάμετρος.

Σε πολλές πρακτικές εφαρμογές, η εκτίμηση των παραμέτρων για τα μοντέλα naive Bayes χρησιμοποιεί τη μέθοδο της μέγιστης πιθανοφάνειας- με άλλα λόγια, μπορεί κανείς να εργαστεί με το μοντέλο naive Bayes χωρίς να αποδεχθεί την πιθανότητα Bayes ή να χρησιμοποιήσει οποιεσδήποτε μεθόδους Bayes.

Παρά τον αφελή σχεδιασμό τους και τις προφανώς υπεραπλουστευμένες υποθέσεις τους, οι ταξινομητές naive Bayes έχουν λειτουργήσει αρκετά καλά σε πολλές σύνθετες καταστάσεις του πραγματικού κόσμου. Το 2004, μια ανάλυση του προβλήματος της ταξινόμησης Bayes έδειξε ότι υπάρχουν βάσιμοι θεωρητικοί λόγοι για τη φαινομενικά απίθανη αποτελεσματικότητα των ταξινομητών naive Bayes. Παρόλα αυτά, μια ολοκληρωμένη σύγκριση με άλλους αλγορίθμους ταξινόμησης το 2006 έδειξε ότι η ταξινόμηση Bayes υπερτερεί έναντι άλλων προσεγγίσεων, όπως τα ενισχυμένα δέντρα ή τα τυχαία δάση.

Ένα πλεονέκτημα του naive Bayes είναι ότι απαιτεί μόνο μια μικρή ποσότητα δεδομένων εκπαίδευσης για την εκτίμηση των παραμέτρων που είναι απαραίτητες για την ταξινόμηση [46].

Στα πλεονεκτήματα του ταξινομητή Naïve Bayes, κατατάσσεται ότι είναι εύκολος στην υλοποίηση και υπολογιστικά αποδοτικός, ενώ είναι αποτελεσματικός σε περιπτώσεις με μεγάλο αριθμό χαρακτηριστικών. Επιπροσθέτως, αποδίδει καλά ακόμη και με περιορισμένα δεδομένα εκπαίδευσης, καθώς επίσης και στην παρουσία κατηγορικών χαρακτηριστικών. Τέλος, για αριθμητικά χαρακτηριστικά τα δεδομένα υποτίθεται ότι προέρχονται από κανονικές κατανομές.

Αντιθέτως, στα μειονεκτήματα του ταξινομητή Naïve Bayes κατατάσσεται το γεγονός ότι υποθέτει ότι τα χαρακτηριστικά είναι ανεξάρτητα, κάτι που μπορεί να μην ισχύει πάντα στα δεδομένα του πραγματικού κόσμου, ενώ μπορεί να επηρεαστεί από άσχετα χαρακτηριστικά. Τελικά, έχει τη δυνατότητα να αναθέσει μηδενική πιθανότητα σε αθέατα γεγονότα, οδηγώντας σε κακή γενίκευση.

Εφαρμογές του ταξινομητή Naïve Bayes

i) **Φιλτράρισμα ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου:** Ταξινομεί τα μηνύματα ηλεκτρονικού ταχυδρομείου ως ανεπιθύμητα ή μη ανεπιθύμητα με βάση τα χαρακτηριστικά.

ii) **Ταξινόμηση κειμένου:** Χρησιμοποιείται στην ανάλυση συναισθήματος, στην κατηγοριοποίηση εγγράφων και στην ταξινόμηση θεμάτων.

iii) **Ιατρική διάγνωση:** Βοηθά στην πρόβλεψη της πιθανότητας μιας ασθένειας με βάση τα συμπτώματα.

iv) **Βαθμολόγηση πιστώσεων:** Αξιολογεί την πιστοληπτική ικανότητα ατόμων για την έγκριση δανείων.

v) **Πρόβλεψη καιρού:** Ταξινομεί τις καιρικές συνθήκες με βάση διάφορους παράγοντες [47].

2.9. MLP

Ένα Τεχνητό Νευρωνικό Δίκτυο (ANN) συγκροτεί ένα μοντέλο Μηχανικής Μάθησης εμπνευσμένο από τη δομή και τη λειτουργία του διασυνδεδεμένου δικτύου νευρώνων του ανθρώπινου εγκεφάλου. Αποτελείται από διασυνδεδεμένους κόμβους που ονομάζονται Τεχνητοί Νευρώνες και είναι οργανωμένοι σε επίπεδα. Η πληροφορία ρέει μέσω του δικτύου, με κάθε νευρώνα να επεξεργάζεται σήματα εισόδου και να παράγει ένα σήμα εξόδου που επηρεάζει άλλους νευρώνες του δικτύου.

Το Πολυεπίπεδο Perceptron (MLP) είναι ένας τύπος Τεχνητού Νευρωνικού Δικτύου που αποτελείται από πολλαπλά στρώματα νευρώνων. Οι νευρώνες στο MLP χρησιμοποιούν συνήθως μη γραμμικές συναρτήσεις ενεργοποίησης, επιτρέποντας στο δίκτυο να μαθαίνει σύνθετα μοτίβα στα δεδομένα. Τα MLP είναι σημαντικά στη Μηχανική Μάθηση διότι έχουν τη δυνατότητα εκμάθησης μη γραμμικών σχέσεων στα δεδομένα, καθιστώντας τα μοντέλα ισχυρά για εργασίες όπως η ταξινόμηση, η παλινδρόμηση και η αναγνώριση προτύπων [48].

Τα σύγχρονα Νευρωνικά δίκτυα εκπαιδεύονται με τη χρήση Οπισθοδιάδοσης και στην καθομιλουμένη αναφέρονται ως δίκτυα «βανίλιας». Τα MLP προέκυψαν από την προσπάθεια βελτίωσης των Perceptrons ενός στρώματος, τα οποία μπορούσαν να εφαρμοστούν μόνο σε γραμμικά διαχωρίσιμα δεδομένα. Ένα perceptron παραδοσιακά χρησιμοποιούσε μια συνάρτηση

βήματος Heaviside ως μη γραμμική συνάρτηση ενεργοποίησης. Ωστόσο, ο αλγόριθμος Οπισθοδιάδοσης απαιτεί από τα σύγχρονα MLP να χρησιμοποιούν συνεχείς συναρτήσεις ενεργοποίησης, όπως η σιγμοειδής ή η ReLU.

Τα Πολυστρωματικά Perceptrons αποτελούν τη βάση της βαθιάς μάθησης και είναι εφαρμόσιμα σε ένα τεράστιο σύνολο διαφορετικών τομέων.

Ένα πολυστρωματικό perceptron είναι ένας τύπος Νευρωνικού Δικτύου πρόωσης που απαρτίζεται από πλήρως συνδεδεμένους νευρώνες με ένα μη γραμμικό είδος συνάρτησης ενεργοποίησης, οργανωμένους σε στρώματα. Χρησιμοποιείται ευρέως για τη διάκριση δεδομένων που δεν είναι γραμμικά διαχωρίσιμα [49].

Τα MLP έχουν χρησιμοποιηθεί ευρέως σε διάφορους τομείς, όπως είναι μεταξύ άλλων η αναγνώριση εικόνας, η επεξεργασία φυσικής γλώσσας και η αναγνώριση ομιλίας. Η ευελιξία στην αρχιτεκτονική τους και η ικανότητά τους να προσεγγίζουν οποιαδήποτε συνάρτηση υπό ορισμένες συνθήκες τα καθιστούν θεμελιώδες δομικό στοιχείο στη βαθιά μάθηση και την έρευνα των Νευρωνικών Δικτύων. Κάτωθι παρατίθενται οι βασικές έννοιες τους.

Το στρώμα εισόδου αποτελείται από κόμβους ή νευρώνες που λαμβάνουν τα αρχικά δεδομένα εισόδου. Κάθε νευρώνας αντιπροσωπεύει ένα χαρακτηριστικό ή μια διάσταση των δεδομένων εισόδου. Ο αριθμός των νευρώνων στο στρώμα εισόδου καθορίζεται από τη διαστατικότητα των δεδομένων εισόδου.

Μεταξύ των στρωμάτων εισόδου και εξόδου, μπορεί να υπάρχουν ένα ή περισσότερα στρώματα νευρώνων. Κάθε νευρώνας σε ένα κρυφό στρώμα λαμβάνει εισόδους από όλους τους νευρώνες στο προηγούμενο στρώμα (είτε το στρώμα εισόδου είτε ένα άλλο κρυφό στρώμα) και παράγει μια έξοδο που μεταβιβάζεται στο επόμενο στρώμα. Ο αριθμός των κρυφών στρωμάτων και ο αριθμός των νευρώνων σε κάθε κρυφό στρώμα είναι υπερπαράμετροι που πρέπει να καθοριστούν κατά τη φάση σχεδιασμού του μοντέλου.

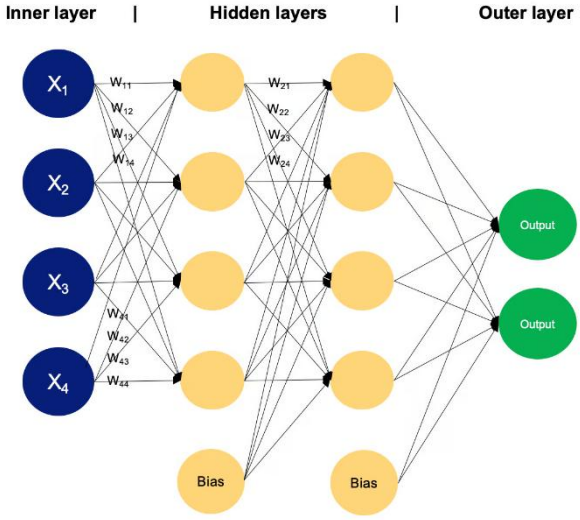
Αυτό το στρώμα αποτελείται από νευρώνες που παράγουν την τελική έξοδο του δικτύου. Ο αριθμός των νευρώνων στο στρώμα εξόδου εξαρτάται από τη φύση της εργασίας. Στη δυαδική ταξινόμηση, μπορεί να υπάρχουν είτε ένας είτε δύο νευρώνες ανάλογα με τη συνάρτηση ενεργοποίησης και να αντιπροσωπεύουν την πιθανότητα να ανήκει κανείς σε μία κλάση, ενώ σε εργασίες ταξινόμησης πολλαπλών κλάσεων, μπορεί να υπάρχουν πολλοί νευρώνες στο στρώμα εξόδου.

Οι νευρώνες σε γειτονικά στρώματα είναι πλήρως συνδεδεμένοι μεταξύ τους. Κάθε σύνδεση έχει ένα σχετικό βάρος, το οποίο καθορίζει την ισχύ της σύνδεσης. Αυτά τα βάρη μαθαίνονται κατά τη διάρκεια της διαδικασίας εκπαίδευσης.

Εκτός από τους νευρώνες εισόδου και τους κρυμμένους νευρώνες, κάθε στρώμα (εκτός από το στρώμα εισόδου) περιλαμβάνει συνήθως έναν νευρώνα προκατάληψης που παρέχει μια σταθερή είσοδο στους νευρώνες του επόμενου στρώματος. Οι νευρώνες προκατάληψης έχουν το δικό τους βάρος που σχετίζεται με κάθε σύνδεση, το οποίο επίσης μαθαίνεται κατά τη διάρκεια της εκπαίδευσης.

Ο νευρώνας προκατάληψης μετατοπίζει ουσιαστικά τη συνάρτηση ενεργοποίησης των νευρώνων στο επόμενο στρώμα, επιτρέποντας στο δίκτυο να μάθει μια μετατόπιση ή προκατάληψη στο όριο απόφασης. Με την προσαρμογή των βαρών που συνδέονται με τον νευρώνα προκατάληψης, το MLP μπορεί να μάθει να ελέγχει το όριο ενεργοποίησης και να προσαρμόζεται καλύτερα στα δεδομένα εκπαίδευσης.

Είναι αξιοσημείωτο να αναφερθεί ότι στο πλαίσιο των MLP, η προκατάληψη μπορεί να αναφέρεται σε δύο συναφείς αλλά διαφορετικές έννοιες: την προκατάληψη ως γενικό όρο στη μηχανική μάθηση και τον νευρώνα προκατάληψης. Στη Γενική Μηχανική Μάθηση, η προκατάληψη αναφέρεται στο σφάλμα που εισάγεται με την προσέγγιση ενός πραγματικού προβλήματος με ένα απλοποιημένο μοντέλο. Η μεροληψία μετρά πόσο καλά το μοντέλο μπορεί να συλλάβει τα υποκείμενα μοτίβα στα δεδομένα. Μια υψηλή μεροληψία υποδηλώνει ότι το



Εικόνα 18: Λειτουργία Μοντέλου MLP

μοντέλο είναι πολύ απλοποιημένο και μπορεί να μην προσαρμόζεται επαρκώς στα δεδομένα, ενώ μια χαμηλή μεροληψία υποδηλώνει ότι το μοντέλο καταγράφει καλά τα υποκείμενα πρότυπα.

Συνήθως, κάθε νευρώνας στα κρυφά στρώματα και στο στρώμα εξόδου εφαρμόζει μια συνάρτηση ενεργοποίησης στο σταθμισμένο άθροισμα των εισόδων του. Οι συνήθεις συναρτήσεις ενεργοποίησης περιλαμβάνουν τις sigmoid, tanh, ReLU (Rectified Linear Unit) και softmax. Αυτές οι συναρτήσεις εισάγουν μη γραμμικότητα στο δίκτυο, επιτρέποντάς του να μαθαίνει σύνθετα μοτίβα στα δεδομένα.

Τα MLP εκπαιδεύονται χρησιμοποιώντας τον αλγόριθμο Backpropagation, ο οποίος υπολογίζει τις κλίσεις μιας συνάρτησης απώλειας σε σχέση με τις παραμέτρους του μοντέλου και ενημερώνει τις παραμέτρους επαναληπτικά για την ελαχιστοποίηση της απώλειας [48].

2.10. Gradient Boosting

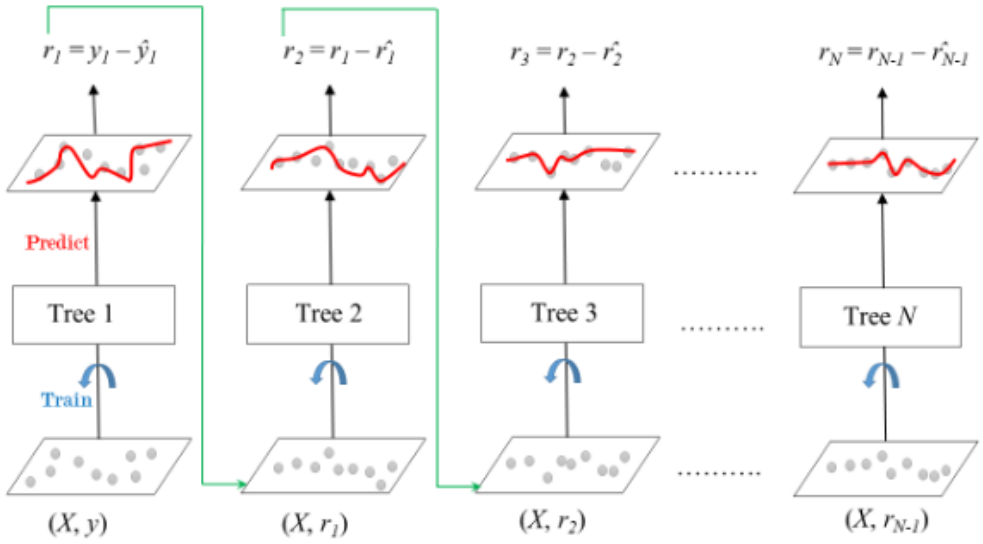
Το Gradient boosting είναι μια τεχνική συνόλου Μηχανικής Μάθησης που συνδυάζει διαδοχικά τις προβλέψεις πολλαπλών αδύναμων μαθητών, συνήθως δέντρων απόφασης. Στοχεύει στη βελτίωση της συνολικής απόδοσης πρόβλεψης βελτιστοποιώντας τα βάρη του μοντέλου με βάση τα σφάλματα των προηγούμενων επαναλήψεων, μειώνοντας σταδιακά τα σφάλματα πρόβλεψης και ενισχύοντας την ακρίβεια του μοντέλου. Αυτή η τεχνική χρησιμοποιείται συνηθέστερα για γραμμική παλινδρόμηση [50]. Το Gradient Boosting είναι ένας δημοφιλής αλγόριθμος ενίσχυσης στη Μηχανική Μάθηση που χρησιμοποιείται για εργασίες ταξινόμησης και παλινδρόμησης. Το Boosting είναι ένα είδος μεθόδου μάθησης συνόλου που εκπαιδεύει το μοντέλο διαδοχικά και κάθε νέο μοντέλο προσπαθεί να διορθώσει το προηγούμενο μοντέλο. Συνδυάζει διάφορους αδύναμους εκπαιδευόμενους σε ισχυρούς εκπαιδευόμενους. Υπάρχουν δύο πιο δημοφιλείς αλγόριθμοι ενίσχυσης, δηλαδή:

i) AdaBoost (αναπτύσσεται σε παρακάτω ενότητα)

ii) Gradient Boosting

Ο αλγόριθμος Gradient Boosting είναι ένας ισχυρός αλγόριθμος ενίσχυσης που συνδυάζει διάφορους αδύναμους εκπαιδευόμενους σε ισχυρούς εκπαιδευόμενους, στον οποίο κάθε νέο μοντέλο εκπαιδεύεται για να ελαχιστοποιήσει τη συνάρτηση απώλειας, όπως το μέσο τετραγωνικό σφάλμα ή η διασταυρούμενη εντροπία του προηγούμενου μοντέλου με τη χρήση κλίσης καθόδου. Σε κάθε επανάληψη, ο αλγόριθμος υπολογίζει την κλίση της συνάρτησης απώλειας σε σχέση με τις προβλέψεις του τρέχοντος συνόλου και στη συνέχεια εκπαιδεύει ένα νέο αδύναμο μοντέλο για να ελαχιστοποιήσει αυτή την κλίση. Οι προβλέψεις του νέου μοντέλου προστίθενται στη συνέχεια στο σύνολο και η διαδικασία επαναλαμβάνεται μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής.

Σε αντίθεση με το AdaBoost, τα βάρη των περιπτώσεων εκπαίδευσης δεν τροποποιούνται, αντίθετα, κάθε πρόβλεψη εκπαιδεύεται χρησιμοποιώντας τα εναπομείναντα σφάλματα του προηγούμενου ως ετικέτες. Υπάρχει μια τεχνική που ονομάζεται Gradient Boosted Trees της οποίας ο βασικός μαθητής είναι το CART (Classification and Regression Trees). Το παρακάτω διάγραμμα εξηγεί πώς εκπαιδεύονται τα Gradient Boosted Trees για προβλήματα παλινδρόμησης.



Εικόνα 19: Εκπαίδευση Gradient Boosted Trees

Το σύνολο συγκροτείται από M δέντρα. Το δέντρο1 εκπαιδεύεται χρησιμοποιώντας τον πίνακα χαρακτηριστικών X και τις ετικέτες y. Οι προβλέψεις με την ετικέτα y_1(hat) χρησιμοποιούνται για τον προσδιορισμό των υπολειμματικών σφαλμάτων r_1 του συνόλου εκπαίδευσης. Το δέντρο2 εκπαιδεύεται στη συνέχεια χρησιμοποιώντας τον πίνακα χαρακτηριστικών X και τα υπολειμματικά σφάλματα r_1 του δέντρου1 ως ετικέτες. Τα προβλεπόμενα αποτελέσματα r_1(hat) χρησιμοποιούνται στη συνέχεια για τον προσδιορισμό του υπολειμματικού r_2. Η διαδικασία επαναλαμβάνεται έως ότου εκπαιδευτούν και τα M δέντρα που αποτελούν το σύνολο. Υπάρχει μια σημαντική παράμετρος που χρησιμοποιείται σε αυτή την τεχνική και είναι γνωστή ως Shrinkage (συρρίκνωση). Η συρρίκνωση αναφέρεται στο γεγονός ότι η πρόβλεψη κάθε δέντρου στο σύνολο συρρικνώνεται αφού πολλαπλασιαστεί με το ρυθμό μάθησης (eta) που κυμαίνεται μεταξύ 0 και 1. Υπάρχει συμβιβασμός μεταξύ του eta και του αριθμού των εκτιμητών, ο μειούμενος ρυθμός μάθησης πρέπει να αντισταθμίζεται με αύξηση των εκτιμητών προκειμένου να επιτευχθεί ορισμένη απόδοση του μοντέλου. Εφόσον όλα τα δέντρα έχουν εκπαιδευτεί τώρα,

μπορούν να γίνουν προβλέψεις. Κάθε δέντρο προβλέπει μια ετικέτα και η τελική πρόβλεψη δίνεται από τον τύπο [51]:

$$\boldsymbol{y(pred) = y_1 + (eta \cdot r_1) + (eta \cdot r_2) + \cdots + (eta \cdot r_N)}$$ Εξίσωση. 1

Τα σφάλματα διαδραματίζουν σημαντικό ρόλο σε κάθε αλγόριθμο μηχανικής μάθησης. Υπάρχουν δύο κύριοι τύποι σφαλμάτων: σφάλμα προκατάληψης και σφάλμα διακύμανσης. Ο αλγόριθμος gradient boost βοηθά στην ελαχιστοποίηση σφαλμάτων προκατάληψης του μοντέλου. Η βασική ιδέα πίσω από αυτόν τον αλγόριθμο είναι η διαδοχική δημιουργία μοντέλων και αυτά τα επόμενα μοντέλα προσπαθούν να μειώσουν τα σφάλματα του προηγούμενου μοντέλου.

Πώς επιτυγχάνεται αυτό; Πώς μειώνεται το σφάλμα; Με την κατασκευή ενός νέου μοντέλου με βάση τα σφάλματα ή τα κατάλοιπα του προηγούμενου μοντέλου.

Όταν η στήλη-στόχος είναι συνεχής χρησιμοποιείται ένας Gradient Boosting Regressor, ενώ όταν πρόκειται για πρόβλημα ταξινόμησης, ένας Gradient Boosting Classifier. Η μόνη διαφορά μεταξύ των δύο είναι η «συνάρτηση απώλειας». Ο στόχος είναι η ελαχιστοποίηση αυτής της συνάρτησης απώλειας προσθέτοντας αδύναμους εκπαιδευόμενους χρησιμοποιώντας την κάθοδο κλίσης. Δεδομένου ότι βασίζεται στη συνάρτηση απωλειών, για προβλήματα παλινδρόμησης, υπάρχουν διαφορετικές συναρτήσεις απωλειών, όπως το μέσο τετραγωνικό σφάλμα (MSE), ενώ για ταξινόμηση, διαφορετικές συναρτήσεις, όπως η λογαριθμική πιθανοφάνεια (log-likelihood) [50].

Η ενίσχυση βαθμίδας είναι μια τεχνική Μηχανικής Μάθησης που βασίζεται στην ενίσχυση σε ένα λειτουργικό χώρο, όπου ο στόχος είναι τα ψευδό-κατάλοιπα έναντι των απλών κατάλοιπων όπως στην παραδοσιακή ενίσχυση. Προσδίδει ένα μοντέλο πρόβλεψης με τη μορφή ενός συνόλου αδύναμων μοντέλων πρόβλεψης, δηλαδή μοντέλων που κάνουν πολύ λίγες υποθέσεις για τα δεδομένα, τα οποία είναι συνήθως απλά δέντρα απόφασης. Όταν ένα δέντρο απόφασης είναι ο αδύναμος μαθητής, ο αλγόριθμος που προκύπτει ονομάζεται gradient-boosted trees- και συνήθως υπερτερεί του random forest. Όπως και με άλλες μεθόδους boosting, ένα μοντέλο gradient-boosted trees χτίζεται σταδιακά, αλλά γενικεύει τις άλλες μεθόδους επιτρέποντας τη βελτιστοποίηση μιας αυθαίρετης διαφοροποιήσιμης συνάρτησης απώλειας.

Η ενίσχυση βαθμίδας μπορεί να χρησιμοποιηθεί για την κατάταξη της σημασίας των χαρακτηριστικών, η οποία συνήθως βασίζεται στη συνένωση της συνάρτησης σημαντικότητας των μαθητών βάσης. Για παράδειγμα, εάν ένας αλγόριθμος δέντρων με ενίσχυση βαθμίδας έχει αναπτυχθεί χρησιμοποιώντας δέντρα απόφασης με βάση την εντροπία, ο αλγόριθμος συνόλου κατατάσσει τη σημασία των χαρακτηριστικών με βάση την εντροπία, επίσης με την επιφύλαξη ότι αυτή υπολογίζεται κατά μέσο όρο σε όλους τους μαθητές βάσης.

Ενώ η ενίσχυση μπορεί να αυξήσει την ακρίβεια ενός μαθητή βάσης, όπως ένα δέντρο απόφασης ή μια γραμμική παλινδρόμηση, θυσιάζει την καταληπτότητα και την ερμηνευσιμότητα. Λόγου χάρη, η παρακολούθηση της πορείας που ακολουθεί ένα δέντρο απόφασης για να λάβει την απόφασή του, είναι τετριμμένη και αυτονόητη, αλλά η παρακολούθηση των πορειών εκατοντάδων ή χιλιάδων δέντρων είναι πολύ πιο δύσκολη. Για να επιτευχθούν τόσο η απόδοση όσο και η ερμηνευσιμότητα, ορισμένες τεχνικές συμπίεσης μοντέλων επιτρέπουν τη μετατροπή ενός XGBoost σε ένα ενιαίο «αναγεννημένο» δέντρο απόφασης που προσεγγίζει την ίδια συνάρτηση απόφασης. Επιπλέον, η υλοποίησή του μπορεί να είναι πιο δύσκολη λόγω της υψηλότερης υπολογιστικής απαίτησης [52].

2.11. AdaBoost

Ο AdaBoost (Adaptive Boosting) αποτελεί έναν μετά-οριστή που ξεκινά με την προσαρμογή ενός ταξινομητή στο αρχικό σύνολο δεδομένων και στη συνέχεια προσαρμόζει πρόσθετα αντίγραφα του ταξινομητή στο ίδιο σύνολο δεδομένων, αλλά όπου τα βάρη των εσφαλμένα ταξινομημένων περιπτώσεων προσαρμόζονται ούτως ώστε οι επόμενοι ταξινομητές να εστιάζουν περισσότερο στις δύσκολες περιπτώσεις [53]. Είναι ένας μετά-αλγόριθμος στατιστικής ταξινόμησης που διατυπώθηκε από τους Yoav Freund και Robert Schapire το 1995, οι οποίοι κέρδισαν το βραβείο Gödel το 2003 για το έργο τους. Μπορεί να χρησιμοποιηθεί σε συνδυασμό με πολλούς τύπους αλγορίθμων μάθησης για τη βελτίωση της απόδοσης. Η έξοδος πολλαπλών ασθενών μαθητών συνδυάζεται σε ένα σταθμισμένο άθροισμα που αντιπροσωπεύει την τελική έξοδο του

ενισχυμένου ταξινομητή. Συνήθως, το AdaBoost παρουσιάζεται για δυαδική ταξινόμηση, αν και μπορεί να γενικευτεί σε πολλαπλές κλάσεις ή σε περιορισμένα διαστήματα πραγματικών τιμών.

Ο AdaBoost είναι προσαρμοστικός, με την έννοια ότι οι επόμενοι αδύναμοι μαθητές (μοντέλα) προσαρμόζονται υπέρ των περιπτώσεων που έχουν ταξινομηθεί εσφαλμένα από τα προηγούμενα μοντέλα. Σε ορισμένα προβλήματα, μπορεί να είναι λιγότερο επιρρεπής στην υπερπροσαρμογή από άλλους αλγορίθμους μάθησης. Οι επιμέρους εκπαιδευόμενοι μπορεί να είναι αδύναμοι, αλλά εφόσον η απόδοση του καθενός είναι ελαφρώς καλύτερη από την τυχαία εικασία, το τελικό μοντέλο μπορεί αποδεδειγμένα να συγκλίνει σε έναν ισχυρό εκπαιδευόμενο.

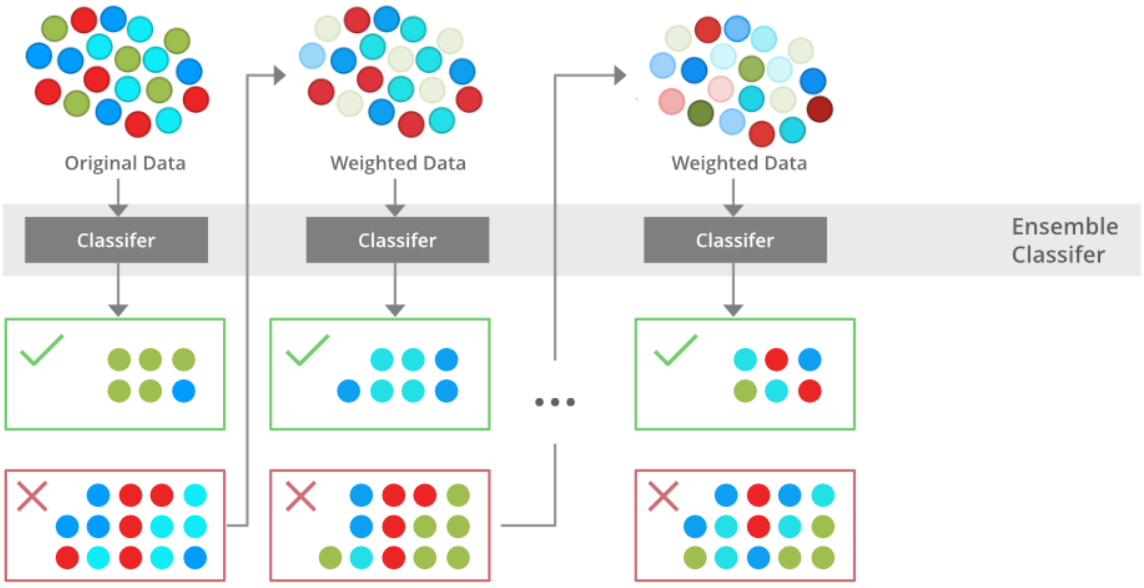
Παρόλο που ο AdaBoost χρησιμοποιείται συνήθως για να συνδυάσει αδύναμους εκπαιδευόμενους βάσης, έχει αποδειχθεί ότι μπορεί επίσης να συνδυάσει αποτελεσματικά ισχυρούς εκπαιδευόμενους βάσης, όπως βαθύτερα δέντρα αποφάσεων, παράγοντας για ένα ακόμη πιο ακριβές μοντέλο.

Κάθε αλγόριθμος μάθησης τείνει να ταιριάζει σε ορισμένους τύπους προβλημάτων καλύτερα από άλλους, και συνήθως έχει πολλές διαφορετικές παραμέτρους και διαμορφώσεις που πρέπει να ρυθμιστούν πριν επιτύχει τη βέλτιστη απόδοση σε ένα σύνολο δεδομένων. Ο AdaBoost, με τα δέντρα απόφασης ως αδύναμους μαθητές, αναφέρεται συχνά ως ο καλύτερος ταξινομητής από το κουτί. Όταν χρησιμοποιείται με τη μάθηση δέντρων απόφασης, οι πληροφορίες που συλλέγονται σε κάθε στάδιο του αλγορίθμου AdaBoost σχετικά με τη σχετική «σκληρότητα» κάθε δείγματος εκπαίδευσης τροφοδοτούνται στον αλγόριθμο αύξησης των δέντρων, έτσι ώστε τα μεταγενέστερα δέντρα να τείνουν να επικεντρώνονται σε παραδείγματα που είναι πιο δύσκολα στην ταξινόμηση [54].

Ο AdaBoost είναι ένα σύνολο μάθησης που χρησιμοποιείται στη Μηχανική Μάθηση για προβλήματα ταξινόμησης και παλινδρόμησης. Η βασική ιδέα πίσω από το AdaBoost είναι η επαναληπτική εκπαίδευση του αδύναμου ταξινομητή στο σύνολο δεδομένων εκπαίδευσης με κάθε διαδοχικό ταξινομητή να δίνει μεγαλύτερη βαρύτητα στα σημεία δεδομένων που έχουν ταξινομηθεί λάθος. Το τελικό μοντέλο AdaBoost αποφασίζεται συνδυάζοντας όλους τους αδύναμους ταξινομητές που χρησιμοποιήθηκαν για την εκπαίδευση με τη βαρύτητα που δίνεται στα μοντέλα ανάλογα με τις ακρίβειές τους. Το αδύναμο μοντέλο που έχει την υψηλότερη

ακρίβεια λαμβάνει την υψηλότερη βαρύτητα, ενώ το μοντέλο που έχει τη χαμηλότερη ακρίβεια λαμβάνει χαμηλότερη βαρύτητα.

Το παρακάτω διάγραμμα εξηγεί τον αλγόριθμο AdaBoost με πολύ απλό τρόπο. Η κατανόηση του επεξηγείται με τέσσερα απλά βήματα, κατόπιν μιας σταδιακής διαδικασίας:



Εικόνα 20: Λειτουργία AdaBoost

Βήμα 1: Αρχικό μοντέλο (B1)

- ✓ Το σύνολο δεδομένων αποτελείται από πολλαπλά σημεία δεδομένων (κόκκινοι, μπλε και πράσινοι κύκλοι).
- ✓ Σε κάθε σημείο δεδομένων αποδίδεται ίση βαρύτητα.
- ✓ Ο πρώτος ασθενής ταξινομητής προσπαθεί να δημιουργήσει ένα όριο απόφασης.
- ✓ 8 σημεία δεδομένων ταξινομούνται λανθασμένα.

Βήμα 2: Προσαρμογή των βαρών (B2)

- ✓ Στα εσφαλμένα ταξινομημένα σημεία από το B1 αποδίδονται υψηλότερα βάρη (εμφανίζονται ως πιο σκούρα σημεία στο επόμενο βήμα).

- ✓ Εκπαιδεύεται ένας νέος ταξινομητής με ένα βελτιωμένο όριο απόφασης που εστιάζει περισσότερο στα προηγουμένως λανθασμένα ταξινομημένα σημεία.
- ✓ Ορισμένα προηγουμένως κακώς ταξινομημένα σημεία ταξινομούνται τώρα σωστά.
- ✓ 6 σημεία δεδομένων ταξινομούνται εσφαλμένα.

Βήμα 3: Περαιτέρω προσαρμογή (B3)

- ✓ Τα πρόσφατα κακώς ταξινομημένα σημεία από το B2 λαμβάνουν υψηλότερα βάρη για να εξασφαλιστεί καλύτερη ταξινόμηση.
- ✓ Ο ταξινομητής προσαρμόζεται ξανά χρησιμοποιώντας ένα βελτιωμένο όριο απόφασης και 4 σημεία δεδομένων παραμένουν εσφαλμένα ταξινομημένα.

Βήμα 4: Τελικό ισχυρό μοντέλο (B4 - Ensemble Model)

- ✓ Ο τελικός ταξινομητής συνόλου συνδυάζει τους B1, B2 και B3 προκειμένου να λάβει τις δυνάμεις όλων των αδύναμων ταξινομητών.

Συγκεντρώνοντας πολλαπλά μοντέλα, το ensemble μοντέλο επιτυγχάνει υψηλότερη ακρίβεια από οποιοδήποτε μεμονωμένο αδύναμο μοντέλο [55].

2.12. XGBoost

Ο XGBoost (eXtreme Gradient Boosting) είναι μια βιβλιοθήκη λογισμικού ανοικτού κώδικα που παρέχει ένα πλαίσιο regularizing gradient boosting για τις γλώσσες C++, Java, Python, R, Julia, Perl, και Scala. Λειτουργεί σε Linux, Microsoft Windows, και macOS. Από την περιγραφή του έργου, στοχεύει στην παροχή μιας «κλιμακούμενης, φορητής και κατανεμημένης βιβλιοθήκης Gradient Boosting (GBM, GBRT, GBDT)». Τρέχει σε ένα μόνο μηχάνημα, καθώς και στα κατανεμημένα πλαίσια επεξεργασίας Apache Hadoop, Apache Spark, Apache Flink και Dask.

Ο XGBoost απέκτησε μεγάλη δημοτικότητα και προσοχή στα μέσα της δεκαετίας του 2010 ως ο αλγόριθμος επιλογής για πολλές νικήτριες ομάδες διαγωνισμών Μηχανικής Μάθησης [56]. Από την εισαγωγή του το 2014, ο XGBoost έχει γίνει ο αλγόριθμος Μηχανικής Μάθησης που επιλέγουν οι επιστήμονες δεδομένων και οι μηχανικοί Μηχανικής Μάθησης. Πρόκειται για μια βιβλιοθήκη ανοικτού κώδικα που μπορεί να εκπαιδεύσει και να δοκιμάσει μοντέλα σε μεγάλες

ποσότητες δεδομένων. Έχει χρησιμοποιηθεί σε πολλούς τομείς, από την πρόβλεψη των ποσοστών κλικ σε διαφημίσεις μέχρι την ταξινόμηση γεγονότων φυσικής υψηλής ενέργειας. Ο XGBoost είναι ιδιαίτερα δημοφιλές επειδή είναι τόσο γρήγορος, και αυτή η ταχύτητα δεν έχει κόστος στην ακρίβεια.

Ο XGBoost είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που μπορεί να βοηθήσει το χρήστη να κατανοήσει στο έπακρο τα δεδομένα, με σκοπό να λάβει καλύτερες αποφάσεις.

Ο XGBoost είναι μια εφαρμογή των δέντρων απόφασης με ενίσχυση κλίσης, gradient-boosting. Έχει χρησιμοποιηθεί από επιστήμονες δεδομένων και ερευνητές παγκοσμίως για τη βελτιστοποίηση των μοντέλων μηχανικής μάθησης.

Ο XGBoost έχει σχεδιαστεί για ταχύτητα, ευκολία χρήσης και απόδοση σε μεγάλα σύνολα δεδομένων. Δεν απαιτεί βελτιστοποίηση των παραμέτρων ή ρύθμιση, πράγμα που σημαίνει ότι μπορεί να χρησιμοποιηθεί αμέσως μετά την εγκατάσταση χωρίς περαιτέρω ρυθμίσεις.

Ο XGBoost είναι μια ευρέως διαδεδομένη υλοποίηση της ενίσχυσης βαθμίδας, gradient boosting. Παρακάτω αναλύονται ορισμένα χαρακτηριστικά του XGBoost που το καθιστούν τόσο ελκυστικό.

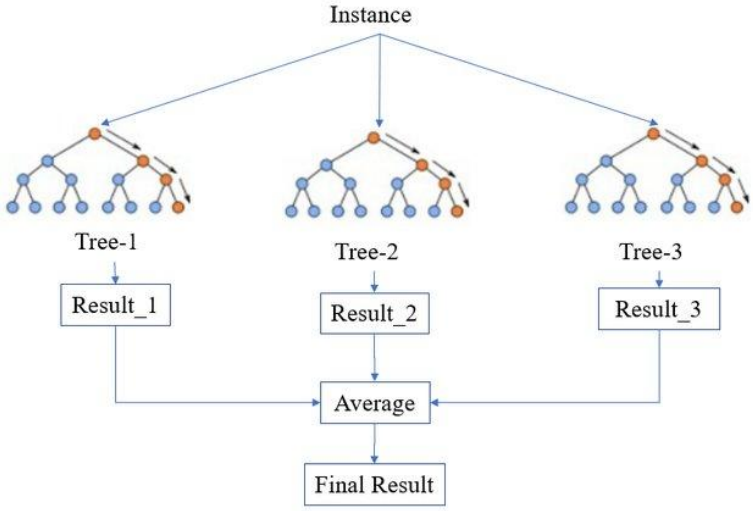
i) Ο XGBoost προσφέρει κανονικοποίηση, η οποία επιτρέπει τον έλεγχο υπερπροσαρμογής εισάγοντας ποινές $\frac{L_1}{L_2}$ στα βάρη και τις προκαταλήψεις κάθε δέντρου. Αυτό το χαρακτηριστικό δεν είναι διαθέσιμο σε πολλές άλλες υλοποιήσεις του gradient boosting.

ii) Ένα άλλο χαρακτηριστικό του XGBoost είναι η ικανότητά του να χειρίζεται αραιά σύνολα δεδομένων χρησιμοποιώντας τον αλγόριθμο σταθμισμένου σκίτσου κβαντιλίων. Αυτός ο αλγόριθμος επιτρέπει την αντιμετώπιση μη μηδενικών καταχωρήσεων στον πίνακα χαρακτηριστικών, διατηρώντας παράλληλα την ίδια υπολογιστική πολυπλοκότητα με άλλους αλγορίθμους όπως η στοχαστική κάθοδος κλίσης.

iii) Ο XGBoost διαθέτει επίσης μια δομή μπλοκ για παράλληλη μάθηση. Αυτό καθιστά εύκολη την κλιμάκωση σε πολυπύρηνες μηχανές ή συστάδες. Χρησιμοποιεί επίσης επίγνωση της

κρυφής μνήμης, η οποία συμβάλλει στη μείωση της χρήσης μνήμης κατά την εκπαίδευση μοντέλων με μεγάλα σύνολα δεδομένων.

iv) Τέλος, το XGBoost προσφέρει δυνατότητες υπολογισμού εκτός πυρήνα χρησιμοποιώντας δομές δεδομένων που βασίζονται σε δίσκο αντί για δομές στη μνήμη κατά τη φάση υπολογισμού [53].



Εικόνα 21: Λειτουργία Μοντέλου XGBoost

Ενώ τα κύρια χαρακτηριστικά του XGBoost που τον κάνουν να διαφέρει από άλλους αλγορίθμους gradient boosting περιλαμβάνουν:

- ✓ Έξυπνη τιμωρία των δέντρων
- ✓ Αναλογική συρρίκνωση των κόμβων φύλλων
- ✓ Newton Boosting
- ✓ Πρόσθετη παράμετρος τυχαιοποίησης
- ✓ Υλοποίηση σε μεμονωμένα, κατανεμημένα συστήματα και υπολογισμούς εκτός πυρήνα
- ✓ Αυτόματη επιλογή χαρακτηριστικών
- ✓ Θεωρητικά αιτιολογημένη σταθμισμένη σκιαγράφηση κβαντιλίων για αποδοτικό υπολογισμό
- ✓ Παράλληλη ενίσχυση δενδρικής δομής με αραιότητα
- ✓ Αποδοτική δομή μπλοκ με δυνατότητα προσωρινής αποθήκευσης για εκπαίδευση δέντρων απόφασης [56].

Ο XGBoost είναι ένας αλγόριθμος gradient boosting για μάθηση με επίβλεψη. Πρόκειται για μια εξαιρετικά αποδοτική και κλιμακούμενη υλοποίηση του αλγορίθμου boosting, με επιδόσεις συγκρίσιμες με αυτές άλλων σύγχρονων αλγορίθμων μηχανικής μάθησης στις περισσότερες περιπτώσεις.

Ο XGBoost χρησιμοποιείται για τους εξής δύο λόγους: ταχύτητα εκτέλεσης και απόδοση του μοντέλου.

Η ταχύτητα εκτέλεσης είναι ζωτικής σημασίας επειδή είναι απαραίτητη για την εργασία με μεγάλα σύνολα δεδομένων. Όταν χρησιμοποιείται ο XGBoost, δεν υπάρχουν περιορισμοί ως προς το μέγεθος του συνόλου δεδομένων, οπότε ο χρήστης μπορεί να εργαστεί με σύνολα δεδομένων που είναι μεγαλύτερα από ό,τι θα ήταν δυνατό με άλλους αλγορίθμους.

Η απόδοση του μοντέλου είναι επίσης απαραίτητη, επειδή επιτρέπει τη δημιουργία μοντέλων που μπορούν να έχουν καλύτερες επιδόσεις από άλλα μοντέλα. Ο XGBoost έχει συγκριθεί με διάφορους αλγορίθμους, όπως το random forest (RF), τα gradient boosting machines (GBM) και τα gradient boosting decision trees (GBDT). Αυτές οι συγκρίσεις δείχνουν ότι ο XGBoost υπερτερεί έναντι αυτών των άλλων αλγορίθμων στην ταχύτητα εκτέλεσης και στην απόδοση του μοντέλου.

Ο Gradient boosting είναι ένας αλγόριθμος ML που δημιουργεί μια σειρά μοντέλων και τα συνδυάζει για να δημιουργήσει ένα συνολικό μοντέλο που είναι ακριβέστερο από κάθε μεμονωμένο μοντέλο της ακολουθίας.

Υποστηρίζει τόσο προβλήματα πρόβλεψης μοντέλων παλινδρόμησης όσο και προβλήματα πρόβλεψης ταξινόμησης.

Για να προσθέσει νέα μοντέλα σε ένα υπάρχον, χρησιμοποιεί έναν αλγόριθμο βαθμωτής καθόδου που ονομάζεται gradient boosting.

Το gradient boosting υλοποιείται από τη βιβλιοθήκη XGBoost, επίσης γνωστή ως πολλαπλά προσθετικά δέντρα παλινδρόμησης, στοχαστικό gradient boosting ή gradient boosting machines.

Ο XGBoost είναι μια εξαιρετικά φορητή βιβλιοθήκη σε πλατφόρμες OS X, Windows και Linux. Χρησιμοποιείται επίσης στην παραγωγή από οργανισμούς σε διάφορους κάθετους κλάδους, συμπεριλαμβανομένων των χρηματοοικονομικών και του λιανικού εμπορίου.

Επίσης είναι ανοικτού κώδικα αλγόριθμος, επομένως είναι δωρεάν στη χρήση, και διαθέτει μια μεγάλη και αυξανόμενη κοινότητα επιστημονικών δεδομένων που συμβάλλουν ενεργά στην ανάπτυξή του. Η βιβλιοθήκη χτίστηκε από την αρχή για να είναι αποτελεσματική, ευέλικτη και φορητή.

Μπορεί κάποιος να χρησιμοποιήσει τον XGBoost για ταξινόμηση, παλινδρόμηση, κατάταξη, ακόμα και για προκλήσεις πρόβλεψης που καθορίζονται από τον χρήστη [57].

Ένα από τα βασικά χαρακτηριστικά του XGBoost είναι ο αποτελεσματικός χειρισμός των ελλιπών τιμών, ο οποίος του επιτρέπει να χειρίζεται δεδομένα πραγματικού κόσμου με ελλιπείς τιμές χωρίς να απαιτείται σημαντική προεπεξεργασία. Επιπλέον, ο XGBoost διαθέτει ενσωματωμένη υποστήριξη για παράλληλη επεξεργασία, καθιστώντας δυνατή την εκπαίδευση μοντέλων σε μεγάλα σύνολα δεδομένων σε εύλογο χρονικό διάστημα.

Ο XGBoost μπορεί να χρησιμοποιηθεί σε μια ποικιλία εφαρμογών, μεταξύ άλλων, συμπεριλαμβανομένων διαγωνισμών Kaggle, συστημάτων συστάσεων και πρόβλεψης του ποσοστού κλικ. Είναι επίσης εξαιρετικά προσαρμόσιμο και επιτρέπει τη λεπτομερή ρύθμιση διαφόρων παραμέτρων του μοντέλου για τη βελτιστοποίηση της απόδοσης.

Το XgBoost σημαίνει Extreme Gradient Boosting, το οποίο προτάθηκε από τους ερευνητές του Πανεπιστημίου της Ουάσινγκτον. Πρόκειται για μια βιβλιοθήκη γραμμένη σε C++ η οποία βελτιστοποιεί την εκπαίδευση για το Gradient Boosting.

Πλεονεκτήματα XGBoost:

i) **Απόδοση:** Ο XGBoost έχει ένα ισχυρό ιστορικό παραγωγής αποτελεσμάτων υψηλής ποιότητας σε διάφορες εργασίες μηχανικής μάθησης, ιδίως στους διαγωνισμούς Kaggle, όπου αποτελεί δημοφιλή επιλογή για τις νικητήριες λύσεις.
ii) **Επεκτασιμότητα:** Έχει σχεδιαστεί για αποτελεσματική και κλιμακούμενη εκπαίδευση μοντέλων μηχανικής μάθησης, καθιστώντας το κατάλληλο για μεγάλα σύνολα δεδομένων.
iii) **Προσαρμοστικότητα:** Διαθέτει ένα ευρύ φάσμα υπερπαραμέτρων που μπορούν να προσαρμοστούν για τη βελτιστοποίηση της απόδοσης, καθιστώντας το ιδιαίτερα προσαρμόσιμο.
iv) **Χειρισμός ελλιπών τιμών:** Διαθέτει ενσωματωμένη υποστήριξη για το χειρισμό ελλιπών τιμών, καθιστώντας εύκολη την εργασία με δεδομένα του πραγματικού κόσμου που συχνά έχουν ελλιπείς τιμές.
v) **Ερμηνευσιμότητα:** Εν αντιθέσει με ορισμένους αλγορίθμους μηχανικής μάθησης που μπορεί να είναι δύσκολο να ερμηνευθούν, ο XGBoost παρέχει δυνατότητες εισαγωγής χαρακτηριστικών, επιτρέποντας την καλύτερη κατανόηση των μεταβλητών που είναι πιο σημαντικές για την πραγματοποίηση προβλέψεων.

Μειονεκτήματα του XGBoost:

i) **Υπολογιστική πολυπλοκότητα:** Το XGBoost μπορεί να είναι υπολογιστικά εντατικό, ιδίως όταν εκπαιδεύονται μεγάλα μοντέλα, γεγονός που το καθιστά λιγότερο κατάλληλο για συστήματα περιορισμένων πόρων.
ii) **Υπερπροσαρμογή:** Σαν αλγόριθμος είναι επιρρεπής σε υπερπροσαρμογή, ειδικά όταν εκπαιδεύεται σε μικρά σύνολα δεδομένων ή όταν χρησιμοποιούνται πάρα πολλά δέντρα στο μοντέλο.
iii) **Ρύθμιση Υπερπαραμέτρων:** Ο XGBoost διαθέτει πολλές υπερπαραμέτρους που μπορούν να ρυθμιστούν, γεγονός που καθιστά σημαντική τη σωστή ρύθμιση των παραμέτρων για τη βελτιστοποίηση της απόδοσης. Ωστόσο, η εύρεση του βέλτιστου συνόλου παραμέτρων μπορεί να είναι χρονοβόρα και απαιτεί τεχνογνωσία.
iv) **Απαιτήσεις Μνήμης:** Είναι εντάσεως μνήμης, ιδίως όταν δουλεύει με μεγάλα σύνολα δεδομένων, γεγονός που το καθιστά λιγότερο κατάλληλο για συστήματα με περιορισμένους πόρους μνήμης.

Εν κατακλείδι, η XGBoost είναι μια βελτιστοποιημένη κατανεμημένη βιβλιοθήκη ενίσχυσης κλίσης που έχει σχεδιαστεί για αποτελεσματική και κλιμακούμενη εκπαίδευση μοντέλων μηχανικής μάθησης. Πρόκειται για μια μέθοδο μάθησης συνόλου που συνδυάζει τις προβλέψεις πολλαπλών αδύναμων μοντέλων για να παράγει μια ισχυρότερη πρόβλεψη. Το XGBoost σημαίνει «Extreme Gradient Boosting» και έχει γίνει ένας από τους πιο δημοφιλείς και ευρέως χρησιμοποιούμενους αλγορίθμους μηχανικής μάθησης λόγω της ικανότητάς του να χειρίζεται μεγάλα σύνολα δεδομένων και της ικανότητάς του να επιτυγχάνει κορυφαίες επιδόσεις σε πολλές εργασίες μηχανικής μάθησης, όπως η ταξινόμηση και η παλινδρόμηση [58].

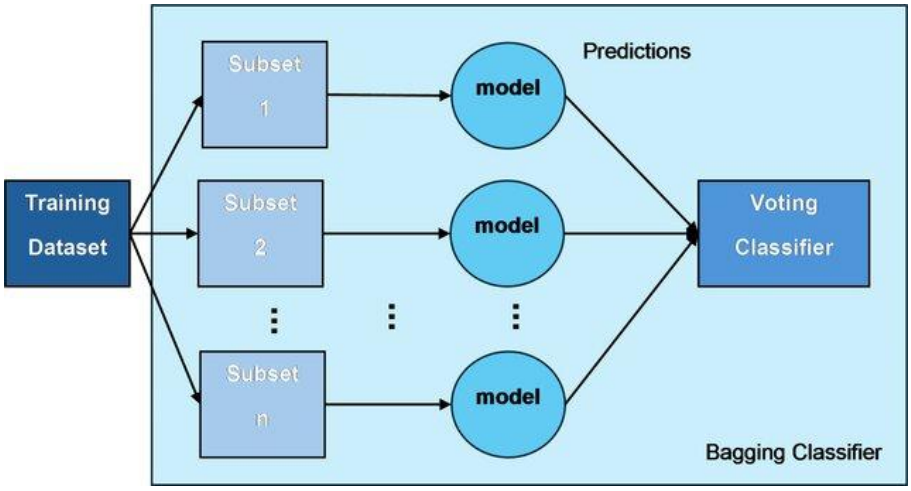
2.13. Bagging Classifier

Το Bagging (ή Bootstrap aggregating) είναι ένας τύπος μάθησης συνόλου, κατά τον οποίο πολλαπλά βασικά μοντέλα εκπαιδεύονται ανεξάρτητα και παράλληλα σε διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης. Κάθε υποσύνολο δημιουργείται με δειγματοληψία bootstrap, κατά την οποία τα σημεία δεδομένων επιλέγονται τυχαία με αντικατάσταση. Στην περίπτωση του ταξινομητή bagging, η τελική πρόβλεψη γίνεται με άθροιση των προβλέψεων του μοντέλου όλων των βάσεων με τη χρήση πλειοψηφικής ψηφοφορίας. Στα μοντέλα παλινδρόμησης, η τελική πρόβλεψη γίνεται με τη μέση τιμή των προβλέψεων του μοντέλου όλων των βάσεων και αυτό είναι γνωστό ως παλινδρόμηση bagging. Επιπρόσθετα, το bagging συμβάλλει στη βελτίωση της ακρίβειας και στη μείωση της υπερπροσαρμογής, ιδίως σε μοντέλα που έχουν υψηλή διακύμανση [59].

Ο ταξινομητής Bagging είναι ένας μετά-εκτιμητής συνόλου που προσαρμόζει βασικούς ταξινομητές ο καθένας σε τυχαία υποσύνολα του αρχικού συνόλου δεδομένων και στη συνέχεια αθροίζει τις μεμονωμένες προβλέψεις τους, είτε με ψηφοφορία είτε με μέσο όρο, για να σχηματίσει μια τελική πρόβλεψη. Ένας τέτοιος μετά-εκτιμητής μπορεί συνήθως να χρησιμοποιηθεί ως ένας τρόπος μείωσης της διακύμανσης ενός εκτιμητή μαύρου κουτιού, όπως για παράδειγμα ενός δέντρου απόφασης, εισάγοντας τυχαιοποίηση στη διαδικασία κατασκευής του και στη συνέχεια δημιουργώντας ένα σύνολο από αυτό.

Κάθε βασικός ταξινομητής εκπαιδεύεται παράλληλα με ένα σύνολο εκπαίδευσης το οποίο δημιουργείται με τυχαία άντληση, με αντικατάσταση, N παραδειγμάτων ή δεδομένων, από το

αρχικό σύνολο δεδομένων εκπαίδευσης, όπου N είναι το μέγεθος του αρχικού συνόλου εκπαίδευσης. Το σύνολο εκπαίδευσης για κάθε έναν από τους βασικούς ταξινομητές είναι ανεξάρτητο μεταξύ του. Πολλά από τα αρχικά δεδομένα μπορεί να επαναλαμβάνονται στο σύνολο εκπαίδευσης που προκύπτει, ενώ άλλα μπορεί να παραλείπονται.



Εικόνα 22: Δομή Μοντέλου Gradient Boosting

Το bagging μειώνει την υπερπροσαρμογή – διακύμανση, με μέσο όρο ή ψηφοφορία. Ωστόσο, αυτό οδηγεί σε αύξηση της προκατάληψης, η οποία αντισταθμίζεται όμως από τη μείωση της διακύμανσης.

Το Gradient Boosting είναι ένας δημοφιλής αλγόριθμος ενίσχυσης. Στο gradient boosting, κάθε προβλεπτής διορθώνει το σφάλμα του προκατόχου του. Σε αντίθεση με το AdaBoost, τα βάρη των περιπτώσεων εκπαίδευσης δεν τροποποιούνται, αντίθετα, κάθε προγνωστικός παράγοντας εκπαιδεύεται χρησιμοποιώντας τα εναπομείναντα σφάλματα του προκατόχου ως ετικέτες.

Υπάρχει μια τεχνική που ονομάζεται Gradient Boosted Trees της οποίας ο βασικός μαθητής είναι το CART (Classification and Regression Trees) [58].

Τα βασικά βήματα του τρόπου λειτουργίας ενός ταξινομητή bagging είναι τα εξής:

- ✓ **Δειγματοληψία Bootstrap:** Στη δειγματοληψία Bootstrap δειγματοληπτούνται τυχαία «n» υποσύνολα των αρχικών δεδομένων εκπαίδευσης με αντικατάσταση. Αυτό το βήμα διασφαλίζει ότι τα βασικά μοντέλα εκπαιδεύονται σε διαφορετικά υποσύνολα δεδομένων, καθώς ορισμένα δείγματα μπορεί να εμφανίζονται πολλές φορές στο νέο υποσύνολο, ενώ άλλα μπορεί να παραλείπονται. Μειώνει τους κινδύνους υπερπροσαρμογής και βελτιώνει την ακρίβεια του μοντέλου.
- ✓ **Εκπαίδευση Βασικού Μοντέλου:** Στο bagging, χρησιμοποιούνται πολλαπλά μοντέλα βάσης. Μετά τη δειγματοληψία bootstrap, κάθε βασικό μοντέλο εκπαιδεύεται ανεξάρτητα χρησιμοποιώντας έναν συγκεκριμένο αλγόριθμο μάθησης, όπως δέντρα απόφασης, μηχανές διανυσμάτων υποστήριξης ή Νευρωνικά Δίκτυα σε ένα διαφορετικό υποσύνολο δεδομένων bootstrap. Αυτά τα μοντέλα ονομάζονται συνήθως «αδύναμοι εκπαιδευόμενοι» επειδή μπορεί να μην είναι ιδιαίτερα ακριβή από μόνα τους. Δεδομένου ότι το βασικό μοντέλο εκπαιδεύεται ανεξάρτητα από διαφορετικά υποσύνολα δεδομένων. Για να γίνει το μοντέλο υπολογιστικά αποδοτικό και λιγότερο χρονοβόρο, τα βασικά μοντέλα μπορούν να εκπαιδευτούν παράλληλα.
- ✓ **Συγκέντρωση:** Αφού εκπαιδευτούν όλα τα μοντέλα βάσης, χρησιμοποιούνται για την πραγματοποίηση προβλέψεων στα αθέατα δεδομένα, δηλαδή στο υποσύνολο των δεδομένων στα οποία δεν έχει εκπαιδευτεί το συγκεκριμένο μοντέλο βάσης. Στον ταξινομητή συσσωμάτωσης, η προβλεπόμενη ετικέτα κλάσης για τη δεδομένη περίπτωση επιλέγεται με βάση την ψηφοφορία πλειοψηφίας. Η κλάση που έχει την πλειοψηφία των ψήφων είναι η πρόβλεψη του μοντέλου.
- ✓ **Αξιολόγηση εκτός σακούλας (OOB):** Ορισμένα δείγματα αποκλείονται από το υποσύνολο εκπαίδευσης συγκεκριμένων βασικών μοντέλων κατά τη διάρκεια της μεθόδου bootstrapping. Αυτά τα «out-of-bag» δείγματα μπορούν να χρησιμοποιηθούν για την εκτίμηση της απόδοσης του μοντέλου χωρίς την ανάγκη διασταυρούμενης επικύρωσης.
- ✓ **Τελική πρόβλεψη:** Μετά τη συγκέντρωση των προβλέψεων από όλα τα μοντέλα βάσης, το Bagging παράγει μια τελική πρόβλεψη για κάθε περίπτωση.

Πλεονεκτήματα του ταξινομητή bagging

i) **Βελτιωμένη απόδοση πρόβλεψης:** Ο ταξινομητής bagging συχνά υπερτερεί έναντι των μεμονωμένων ταξινομητών μειώνοντας την υπερπροσαρμογή και αυξάνοντας την ακρίβεια πρόβλεψης. Συνδυάζοντας πολλαπλά μοντέλα βάσης, μπορεί να γενικεύσει καλύτερα σε αθέατα δεδομένα.

ii) **Ανθεκτικότητα:** Η ταξινόμηση σε σάκους μειώνει τον αντίκτυπο των ακραίων τιμών και του θορύβου στα δεδομένα, συγκεντρώνοντας τις προβλέψεις από πολλαπλά μοντέλα. Αυτό ενισχύει τη συνολική σταθερότητα και ευρωστία του μοντέλου.

iii) **Μειωμένη Διακύμανση:** Δεδομένου ότι κάθε βασικό μοντέλο εκπαιδεύεται σε διαφορετικά υποσύνολα δεδομένων, η διακύμανση του συγκεντρωτικού μοντέλου μειώνεται σημαντικά σε σύγκριση με ένα μεμονωμένο μοντέλο.

iv) **Παραλληλοποίηση:** Η σακκοποίηση επιτρέπει την παράλληλη επεξεργασία, καθώς κάθε βασικό μοντέλο μπορεί να εκπαιδευτεί ανεξάρτητα. Αυτό το καθιστά υπολογιστικά αποδοτικό, ιδίως για μεγάλα σύνολα δεδομένων.

v) **Ευελιξία:** Ο ταξινομητής bagging είναι μια ευέλικτη τεχνική που μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα αλγορίθμων μηχανικής μάθησης, συμπεριλαμβανομένων των δέντρων απόφασης, των τυχαίων δασών και των μηχανών διανυσμάτων υποστήριξης.

Μειονεκτήματα του Bagging :

i) **Απώλεια Ερμηνευσιμότητας:** Το bagging περιλαμβάνει τη συγκέντρωση προβλέψεων από πολλαπλά μοντέλα, γεγονός που καθιστά δυσκολότερη την ερμηνεία των επιμέρους συνεισφορών κάθε βασικού μοντέλου. Αυτό μπορεί να περιορίσει τη δυνατότητα εξαγωγής ακριβών επιχειρηματικών συμπερασμάτων από το σύνολο.

ii) **Υπολογιστικά Δαπανηρό:** Καθώς αυξάνεται ο αριθμός των επαναλήψεων, δηλαδή τα δείγματα bootstrap, αυξάνεται και το υπολογιστικό κόστος του bagging. Αυτό το καθιστά λιγότερο κατάλληλο για εφαρμογές πραγματικού χρόνου όπου η αποδοτικότητα και η ταχύτητα είναι ζωτικής σημασίας. Για το χειρισμό μεγάλων συνόλων δεδομένων και πολυάριθμων επαναλήψεων απαιτείται συχνά αποτελεσματική παράλληλη επεξεργασία ή κατανεμημένα συστήματα.

iii) **Λιγότερο Ευέλικτη:** Το bagging είναι πιο αποτελεσματικό όταν εφαρμόζεται σε αλγορίθμους που είναι λιγότερο σταθεροί ή πιο επιρρεπείς στην υπερπροσαρμογή. Οι αλγόριθμοι που είναι ήδη σταθεροί ή παρουσιάζουν χαμηλή μεροληψία ενδέχεται να μην επωφεληθούν σημαντικά από το bagging, καθώς υπάρχει λιγότερη διακύμανση προς μείωση εντός του συνόλου.

Ο ταξινομητής bagging μπορεί να εφαρμοστεί σε διάφορες εργασίες του πραγματικού κόσμου:

i) **Ανίχνευση Απάτης:** Ο ταξινομητής ταξινόμηση σε σάκους μπορεί να χρησιμοποιηθεί για την ανίχνευση δόλιων συναλλαγών με τη συγκέντρωση προβλέψεων από πολλαπλά μοντέλα ανίχνευσης απάτης.
ii) **Φιλτράρισμα ανεπιθύμητης αλληλογραφίας:** Ο ταξινομητής bagging μπορεί να χρησιμοποιηθεί για το φιλτράρισμα μηνυμάτων spam με τη συγκέντρωση προβλέψεων από πολλαπλά φίλτρα spam που έχουν εκπαιδευτεί σε διαφορετικά υποσύνολα των μηνυμάτων spam.
iii) **Βαθμολόγηση πιστοληπτικής ικανότητας:** Ο ταξινομητής bagging μπορεί να χρησιμοποιηθεί για τη βελτίωση της ακρίβειας των μοντέλων πιστωτικής βαθμολόγησης συνδυάζοντας τις προβλέψεις πολλαπλών μοντέλων που έχουν εκπαιδευτεί σε διαφορετικά υποσύνολα των πιστωτικών δεδομένων.
iv) **Ταξινόμηση εικόνων:** Ο ταξινομητής bagging μπορεί να χρησιμοποιηθεί για τη βελτίωση της ακρίβειας των εργασιών ταξινόμησης εικόνων συνδυάζοντας τις προβλέψεις πολλαπλών ταξινομητών που εκπαιδεύονται σε διαφορετικά υποσύνολα των εικόνων εκπαίδευσης.
v) **Επεξεργασία φυσικής γλώσσας:** Σε εργασίες NLP, ο ταξινομητής bagging μπορεί να συνδυάσει προβλέψεις από πολλαπλά γλωσσικά μοντέλα για να επιτύχει καλύτερα αποτελέσματα ταξινόμησης κειμένου.

Συμπερασματικά, ο ταξινομητής bagging, ως τεχνική μάθησης συνόλου, προσφέρει μια ισχυρή λύση για τη βελτίωση της προβλεπτικής απόδοσης και της ευρωστίας του μοντέλου. Επιπροσθέτως, αποφεύγει την υπερβολική προσαρμογή, βελτιώνει τη γενίκευση και δίνει

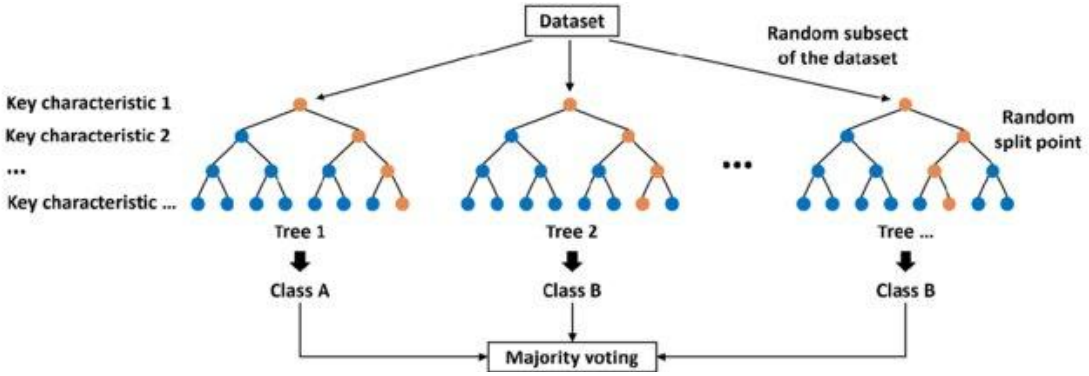
αξιόπιστες προβλέψεις για ένα ευρύ φάσμα εφαρμογών, χρησιμοποιώντας τη συλλογική σοφία πολλών βασικών μοντέλων [59].

2.14. Extra Trees

Τα Extra trees, συντομογραφία για τα εξαιρετικά τυχαία δέντρα, είναι μια μέθοδος Μηχανικής Μάθησης με επίβλεψη που κάνει χρήση δέντρων απόφασης και χρησιμοποιείται από το εργαλείο Train Using AutoML.

Ο αλγόριθμος extra trees, όπως και ο αλγόριθμος random forests, δημιουργεί πολλά δέντρα αποφάσεων, αλλά η δειγματοληψία για κάθε δέντρο είναι τυχαία, χωρίς αντικατάσταση. Αυτό δημιουργεί ένα σύνολο δεδομένων για κάθε δέντρο με μοναδικά δείγματα. Ένας συγκεκριμένος αριθμός χαρακτηριστικών, από το συνολικό σύνολο των χαρακτηριστικών, επιλέγεται επίσης τυχαία για κάθε δέντρο. Το πιο σημαντικό και μοναδικό χαρακτηριστικό των επιπλέον δέντρων είναι η τυχαία επιλογή μιας τιμής διαχωρισμού για ένα χαρακτηριστικό. Αντί του υπολογισμού μιας τοπικά βέλτιστης τιμής με τη χρήση του Gini ή της εντροπίας για τη διάσπαση των δεδομένων, ο αλγόριθμος επιλέγει τυχαία μια τιμή διάσπασης. Αυτό καθιστά τα δέντρα διαφοροποιημένα και ασυσχέτιστα [60].

Ο ταξινομητής εξαιρετικά τυχαίων δέντρων (Extra Trees Classifier) είναι ένας τύπος τεχνικής μάθησης συνόλου που συγκεντρώνει τα αποτελέσματα πολλαπλών απο-συσχετιζόμενων δέντρων απόφασης που συλλέγονται σε ένα «δάσος» για την εξαγωγή του αποτελέσματος ταξινόμησης. Εννοιολογικά, μοιάζει πολύ με έναν ταξινομητή τυχαίου δάσους και διαφέρει από αυτόν μόνο στον τρόπο κατασκευής των δέντρων απόφασης στο δάσος. Κάθε δέντρο απόφασης στο δάσος



Εικόνα 23: Δομή Μοντέλου Extra Trees

Extra Trees Forest κατασκευάζεται από το αρχικό δείγμα εκπαίδευσης. Στη συνέχεια, σε κάθε κόμβο δοκιμής, σε κάθε δέντρο παρέχεται ένα τυχαίο δείγμα k χαρακτηριστικών από το σύνολο χαρακτηριστικών από το οποίο κάθε δέντρο απόφασης πρέπει να επιλέξει το καλύτερο χαρακτηριστικό για το διαχωρισμό των δεδομένων με βάση κάποια μαθηματικά κριτήρια (συνήθως το δείκτη Gini). Αυτό το τυχαίο δείγμα χαρακτηριστικών οδηγεί στη δημιουργία πολλαπλών αποδιαταγμένων δέντρων απόφασης. Για να πραγματοποιηθεί η επιλογή χαρακτηριστικών με την παραπάνω δομή του δάσους, κατά τη διάρκεια της κατασκευής του δάσους, για κάθε χαρακτηριστικό υπολογίζεται η κανονικοποιημένη συνολική μείωση των μαθηματικών κριτηρίων που χρησιμοποιούνται στην απόφαση του χαρακτηριστικού του διαχωρισμού (αν ο δείκτης Gini χρησιμοποιείται στην κατασκευή του δάσους). Η τιμή αυτή ονομάζεται σημασία Gini του χαρακτηριστικού. Για να πραγματοποιηθεί η επιλογή χαρακτηριστικών, κάθε χαρακτηριστικό ταξινομείται σε φθίνουσα σειρά σύμφωνα με τη σημασία Gini του κάθε χαρακτηριστικού και ο χρήστης επιλέγει τα κορυφαία k χαρακτηριστικά σύμφωνα με την επιλογή του.

Ο ταξινομητής Extra Trees για την επιλογή χαρακτηριστικών προσφέρει πολλά πλεονεκτήματα:

i) **Ανθεκτικότητα στο θόρυβο και στα άσχετα χαρακτηριστικά:** Ο ταξινομητής Extra Trees χρησιμοποιεί πολλαπλά δέντρα αποφάσεων και επιλέγει χαρακτηριστικά με βάση τις βαθμολογίες σημαντικότητάς τους, καθιστώντας τον λιγότερο ευαίσθητο στο θόρυβο και τα άσχετα χαρακτηριστικά. Μπορεί να χειριστεί αποτελεσματικά σύνολα δεδομένων με μεγάλο αριθμό χαρακτηριστικών και θορυβώδη δεδομένα.
ii) **Υπολογιστική αποδοτικότητα:** Ο ταξινομητής Extra Trees Classifier κατασκευάζει δέντρα απόφασης παράλληλα, γεγονός που μπορεί να επιταχύνει σημαντικά τη διαδικασία εκπαίδευσης σε σύγκριση με άλλες τεχνικές επιλογής χαρακτηριστικών. Είναι ιδιαίτερα χρήσιμος για σύνολα δεδομένων υψηλής διάστασης όπου η αποδοτικότητα είναι ζωτικής σημασίας.
iii) **Μείωση Μεροληψίας:** Η τυχαία επιλογή υποσυνόλων και τα τυχαία σημεία διαχωρισμού στον Extra Trees Classifier συμβάλλουν στη μείωση της μεροληψίας που μπορεί να προκύψει από τη χρήση ενός μόνο δέντρου απόφασης. Με την εξέταση πολλαπλών

δέντρων απόφασης, παρέχει μια πιο ολοκληρωμένη αξιολόγηση της σημασίας των χαρακτηριστικών.

iv) **Κατάταξη Χαρακτηριστικών:** Το Extra Trees Classifier αποδίδει βαθμολογίες σημαντικότητας σε κάθε χαρακτηριστικό, επιτρέποντάς σας να τα κατατάξετε με βάση τη σχετική τους σημασία. Αυτή η κατάταξη μπορεί να παρέχει πληροφορίες σχετικά με τη συνάφεια και τη συμβολή κάθε χαρακτηριστικού στη μεταβλητή-στόχο.

v) **Χειρισμός Πολυσυμβατότητας:** Ο ταξινομητής Extra Trees μπορεί να χειριστεί αποτελεσματικά συσχετιζόμενα χαρακτηριστικά. Με την τυχαία επιλογή υποσυνόλων χαρακτηριστικών και τη χρήση τυχαίων διαχωρισμών, μειώνει τον αντίκτυπο της πολυσυγγραμμικότητας, σε αντίθεση με τις μεθόδους που βασίζονται σε ρητές συσχετίσεις χαρακτηριστικών.

vi) **Ευελιξία Επιλογής Χαρακτηριστικών:** Η διαδικασία επιλογής χαρακτηριστικών στον ταξινομητή Extra Trees Classifier βασίζεται στις εισαγωγές χαρακτηριστικών, επιτρέποντας την προσαρμογή κατωφλίου για τη συμπερίληψη χαρακτηριστικών σύμφωνα με τις ανάγκες του κάθε χρήστη. Μπορεί να επιλέξει να συμπεριλάβει μόνο τα πιο σημαντικά χαρακτηριστικά ή ένα μεγαλύτερο υποσύνολο, ανάλογα με την επιθυμητή ισορροπία μεταξύ της μείωσης των χαρακτηριστικών και της απόδοσης του μοντέλου.

vii) **Γενίκευση και Ερμηνευσιμότητα:** Με την επιλογή ενός υποσυνόλου σχετικών χαρακτηριστικών, ο ταξινομητής Extra Trees μπορεί να βελτιώσει τη γενίκευση του μοντέλου μειώνοντας την υπερβολική προσαρμογή. Επιπλέον, τα επιλεγμένα χαρακτηριστικά μπορούν να παρέχουν ερμηνεύσιμες πληροφορίες σχετικά με τους παράγοντες που οδηγούν τις προβλέψεις και επηρεάζουν τη μεταβλητή-στόχο.

Τα πλεονεκτήματα που αναφέρθηκαν παραπάνω, καθιστούν τον Extra Trees Classifier ένα πολύτιμο εργαλείο για την επιλογή χαρακτηριστικών, ιδίως όταν πρόκειται για σύνολα δεδομένων υψηλών διαστάσεων, θορυβώδη δεδομένα και καταστάσεις όπου η υπολογιστική αποτελεσματικότητα είναι απαραίτητη [61].

Κεφάλαιο 3ο

3. Μεθοδολογία

3.1. Περιγραφή Δεδομένων

Η παρούσα μελέτη βασίζεται σε δύο datasets, το **"student-por.csv"** και το **"student-mat.csv"**, τα οποία περιέχουν δεδομένα μαθητών από πορτογαλικά σχολεία. Τα συγκεκριμένα datasets είναι διαθέσιμα από το UCI Machine Learning Repository και χρησιμοποιούνται ευρέως για την ανάλυση εκπαιδευτικών δεδομένων. Το "student-mat.csv" περιέχει πληροφορίες για μαθητές που παρακολουθούν το μάθημα των Μαθηματικών, ενώ το "student-por.csv" αφορά μαθητές που παρακολουθούν το μάθημα της Πορτογαλικής Γλώσσας.

Κάθε dataset περιέχει **33 χαρακτηριστικά** που σχετίζονται με δημογραφικές πληροφορίες, οικογενειακό περιβάλλον, κοινωνικοοικονομικούς παράγοντες, καθώς και ακαδημαϊκά στοιχεία των μαθητών. Κάποια από τα βασικά χαρακτηριστικά είναι:

i) **Demographic Data - Δημογραφικά Δεδομένα**

- ✓ **sex:** Φύλο μαθητή (M/F)
- ✓ **age:** Ηλικία μαθητή
- ✓ **address:** Τύπος διαμονής (αστική ή αγροτική περιοχή)

ii) **Family & Social Data - Οικογενειακά & Κοινωνικά Δεδομένα**

- ✓ **famsize:** Μέγεθος οικογένειας (≤3 ή >3 μέλη)
- ✓ **Pstatus:** Κατάσταση διαμονής γονέων (μαζί ή χωριστά)
- ✓ **Medu, Fedu:** Εκπαιδευτικό επίπεδο μητέρας και πατέρα

- ✓ **famsup:** Παροχή οικογενειακής υποστήριξης στη μελέτη (ναι/όχι)
- ✓ **romantic:** Αν ο μαθητής έχει ρομαντική σχέση (ναι/όχι)

iii) **Academic Performance Data - Ακαδημαϊκή Απόδοση**

- ✓ **failures**: Αριθμός προηγούμενων αποτυχιών
- ✓ **studytime:** Χρόνος μελέτης ανά εβδομάδα
- ✓ **schoolsup:** Παροχή επιπλέον εκπαιδευτικής υποστήριξης (ναι/όχι)
- ✓ **G1, G2, G3:** Βαθμολογίες σε τρία διαφορετικά χρονικά σημεία

Υπάρχουν αρκετοί (382) μαθητές που ανήκουν και στα δύο σύνολα δεδομένων .

Αυτοί οι μαθητές μπορούν να εντοπιστούν με την αναζήτηση πανομοιότυπων χαρακτηριστικών που χαρακτηρίζουν κάθε μαθητή.

Ο στόχος της ανάλυσης είναι η πρόβλεψη της ακαδημαϊκής επίδοσης των μαθητών. Ο στόχος - target variable- ορίστηκε βάσει του τελικού βαθμού **G3**, ο οποίος αναπαριστά την τελική επίδοση των μαθητών.

Για τη μοντελοποίηση, η συνεχής μεταβλητή G3 κατατάχθηκε σε τρεις διακριτές κατηγορίες απόδοσης:

i) **Χαμηλή Απόδοση**: $G3 \leq 9$

ii) **Μέτρια Απόδοση**: $10 \leq G3 \leq 14$

iii) **Υψηλή Απόδοση**: $G3 \geq 15$

Η εν λόγω κατηγοριοποίηση επελέγη καθώς επιτρέπει την εύκολη ερμηνεία των προβλέψεων, βοηθώντας καθηγητές και εκπαιδευτικά ιδρύματα να εντοπίσουν μαθητές που χρειάζονται επιπλέον υποστήριξη ή μπορούν να επωφεληθούν από προχωρημένα προγράμματα.

Η ανάλυση πραγματοποιήθηκε μέσω 12 αλγορίθμων πρόβλεψης, και αξιολογήθηκε η ακρίβειά τους βάσει των precision, recall, F1-score και support. Ο συνολικός χρόνος εκτέλεσης καταγράφηκε, καθώς και ο χρόνος εκπαίδευσης και πρόβλεψης κάθε μοντέλου.

Η συλλογή δεδομένων αποτελεί το πρώτο και ίσως πιο κρίσιμο στάδιο της διαδικασίας ανάλυσης δεδομένων. Στην παρούσα εργασία, χρησιμοποιήθηκαν δεδομένα από εκπαιδευτικές πλατφόρμες και ανοιχτά datasets, με σκοπό την ανάλυση της προόδου των μαθητών και την πρόβλεψη της απόδοσής τους. Τα δεδομένα περιλαμβάνουν πληροφορίες όπως βαθμολογίες, χρόνο συμμετοχής, παρακολούθηση μαθημάτων, αλλά και δημογραφικά στοιχεία, που συλλέχθηκαν μέσω αξιόπιστων πηγών και πλατφορμών.

Για την επίτευξη των στόχων της μελέτης, χρησιμοποιήθηκε η κύρια πηγή δεδομένων **Open University Learning Analytics Dataset (OULAD)**, η οποία εμπεριέχει δεδομένα από μαθητές που συμμετείχαν σε προγράμματα εξ αποστάσεως εκπαίδευσης. Ενδεικτικά από τα δεδομένα που περιλαμβάνονται, είναι τα κάτωθι:

- ✓ Αρχεία συμμετοχής σε δραστηριότητες.
- ✓ Αξιολογήσεις, όπως είναι τα quizzes και τα assignments.
- ✓ Συνολική επίδοση στο μάθημα.

Τα εν λόγω δεδομένα είναι ανώνυμα και διατίθενται από το Open University για ερευνητικούς σκοπούς.

Η συλλογή των δεδομένων έγινε μέσω των εξής βημάτων:

- ✓ **Επιλογή πηγών:** Εντοπίστηκαν δημόσια διαθέσιμα datasets, από το OULAD, τα οποία είναι εύκολα προσβάσιμα και καλύπτουν τη θεματική της μελέτης.
- ✓ **Λήψη και αποθήκευση:** Τα datasets αντλήθηκαν από τις αντίστοιχες πλατφόρμες και αποθηκεύτηκαν τοπικά για περαιτέρω επεξεργασία.

- ✓ **Έλεγχος δεδομένων:** Έγινε ενδελεχής έλεγχος για τη διασφάλιση της πληρότητας και της ποιότητας των δεδομένων. Διερευνήθηκαν τυχόν ελλείψεις ή σφάλματα, προκειμένου να διασφαλιστεί ότι τα δεδομένα είναι κατάλληλα για ανάλυση.

Κατά τη διαδικασία συλλογής δεδομένων, παρατηρήθηκαν ορισμένοι περιορισμοί, της η ελλιπής κάλυψη συγκεκριμένων ομάδων πληθυσμού και η ανομοιογένεια στη μορφή των δεδομένων. Αυτά αντιμετωπίστηκαν μέσω προσεκτικής επιλογής των πηγών και της προεπεξεργασίας.

Τα δεδομένα που συλλέχθηκαν αποτελούν τη βάση για την προεπεξεργασία και την ανάλυση που περιγράφονται στα επόμενα κεφάλαια. Η ανάλυσή τους θα επιτρέπει τη δημιουργία μοντέλων πρόβλεψης που μπορούν να αξιολογήσουν την απόδοση των μαθητών και να παρέχουν χρήσιμες πληροφορίες για την εκπαιδευτική διαδικασία. Πρόσβαση στο διαθέσιμο dataset.

3.2. Προεπεξεργασία Δεδομένων

Η προεπεξεργασία δεδομένων αποτελεί βασικό στάδιο σε κάθε ανάλυση δεδομένων, καθώς διασφαλίζει την ποιότητα και την αξιοπιστία των δεδομένων που θα χρησιμοποιηθούν στα επόμενα βήματα. Η διαδικασία αυτή περιλαμβάνει τον καθαρισμό, τη μετατροπή και την προσαρμογή των δεδομένων, ώστε να είναι κατάλληλα για χρήση από αλγορίθμους μηχανικής μάθησης.

Η προεπεξεργασία των δεδομένων της μελέτης περιλάμβανε τα εξής βήματα:

i) **Αφαίρεση ελλιπών ή άχρηστων δεδομένων:** Αφού πρώτα εντοπίστηκαν στήλες ή γραμμές με υψηλό ποσοστό ελλιπών τιμών, όπως κενά δεδομένα ή μηδενικές τιμές, έπειτα οι συγκεκριμένες τιμές αφαιρέθηκαν ή αντικαταστάθηκαν με τη χρήση τεχνικών όπως ο υπολογισμός μέσου όρου ή διαμέσου, όπου ήταν απαραίτητο.

ii) **Αφαίρεση πλεονασμού:** Στήλες όπως "Unnamed: 0" ή "Unnamed: 3", αφαιρέθηκαν, καθώς δεν περιείχαν χρήσιμες πληροφορίες. Επιπλέον, τα δεδομένα καθαρίστηκαν από περιττές πληροφορίες που δεν σχετίζονται με την πρόβλεψη απόδοσης.

iii) **Κωδικοποίηση κατηγορικών μεταβλητών:** Οι κατηγορικές μεταβλητές, όπως η χώρα ή η περιγραφή, κωδικοποιήθηκαν σε αριθμητική μορφή, μέσω one-hot encoding ή label encoding. Με αυτόν τον τρόπο διευκολύνεται η επεξεργασία τους από τα μοντέλα μηχανικής μάθησης.

iv) **Κανονικοποίηση ή κανονικοποίηση δεδομένων:** Χρησιμοποιήθηκαν τεχνικές κανονικοποίησης -normalization- ή τυποποίησης -standardization- για τη μετατροπή αριθμητικών μεταβλητών σε κλίμακα, εφόσον είναι απαραίτητο. Αυτό διασφαλίζει ότι όλες οι μεταβλητές να έχουν παρόμοιο εύρος τιμών και έτσι δε δημιουργείται ανισορροπία στα μοντέλα.

v) **Αντιμετώπιση ανισορροπίας δεδομένων:** Σε περιπτώσεις που τα δεδομένα περιείχαν ανισόρροπη κατανομή (π.χ., υπερβολικά πολλές τιμές για μία κατηγορία), εφαρμόστηκαν τεχνικές όπως undersampling ή oversampling για τη βελτίωση της ισορροπίας.

Προκειμένου να εξεταστεί η ποιότητα και η κατανομή των δεδομένων, δημιουργήθηκαν γραφήματα και στατιστικές αναλύσεις. Παραδείγματος χάριν:

- ✓ Εμφάνιση διανομής βαθμολογιών ή κατηγοριών με χρήση bar plots και histograms.
- ✓ Εντοπισμός πιθανών outliers μέσω box plots.

Κατά την προεπεξεργασία των δεδομένων εντοπίστηκαν προβλήματα, όπως:

- ✓ **Δεδομένα εκτός εύρους – outliers:** Αντιμετωπίστηκαν μέσω αντικατάστασης ή αφαίρεσης.
- ✓ **Ελλιπείς τιμές:** Σε ορισμένες περιπτώσεις χρησιμοποιήθηκαν μέθοδοι όπως η μέση τιμή, προκειμένου να καλυφθούν.

Η προεπεξεργασία των δεδομένων είναι ένα κρίσιμο βήμα πριν από την εφαρμογή των αλγορίθμων μηχανικής μάθησης, καθώς διασφαλίζει ότι τα δεδομένα είναι καθαρά, συνεκτικά και έτοιμα για ανάλυση. Επιπλέον, διευκολύνει την καλύτερη απόδοση των αλγορίθμων μηχανικής μάθησης, καθώς μειώνει τις πιθανότητες λάθους ή παρερμηνείας.

Αρχικά, έγινε έλεγχος για ελλιπές ή μη έγκυρο περιεχόμενο στα δεδομένα. Ο εντοπισμός των ελλειπών τιμών πραγματοποιήθηκε με τη χρήση της μεθόδου .isnull().sum(), προκειμένου να καταμετρηθεί το ποσοστό των δεδομένων που έλειπαν σε κάθε στήλη.

- ✓ Για τις ποσοτικές μεταβλητές, αν υπήρχαν ελλιπές τιμές, εφαρμόστηκε μέση τιμή (mean imputation) ώστε να διατηρηθεί η συνοχή των δεδομένων.
- ✓ Για τις κατηγορικές μεταβλητές, οι ελλείψεις αντιμετωπίστηκαν είτε με τη συμπλήρωση της συχνότερης τιμής (mode imputation) είτε με την εισαγωγή μιας ξεχωριστής κατηγορίας "Unknown".
- ✓ Πραγματοποιήθηκε έλεγχος για ανωμαλίες στις τιμές, δηλαδή για τιμές εκτός εύρους στις βαθμολογίες G1-G2-G3, καθώς και απομακρύνθηκαν πιθανά ακραία σημεία (outliers).

Οι κατηγορικές μεταβλητές, όπως το φύλο, η οικογενειακή κατάσταση, η στήριξη στη μελέτη κ.λπ., δεν μπορούν να χρησιμοποιηθούν απευθείας από τους περισσότερους αλγορίθμους μηχανικής μάθησης. Για αυτόν τον λόγο, εφαρμόστηκε One-Hot Encoding, το οποίο δημιουργεί δυαδικές (0-1) στήλες για κάθε κατηγορία των αρχικών μεταβλητών.

Στην ουσία, στην One-Hot Encoding, οι αριθμητικές μεταβλητές διατηρούνται ως έχουν, ενώ οι κατηγορικές μετατρέπονται σε δυαδικές για την αποφυγή πολυσυγγραμμικότητας.

Μετά την επεξεργασία των μεταβλητών, τα δεδομένα χωρίστηκαν σε:

- ✓ **Χαρακτηριστικά (X)**: Όλες οι ανεξάρτητες μεταβλητές που χρησιμοποιήθηκαν για την πρόβλεψη.

- ✓ **Στόχο (y)**: Η κατηγοριοποιημένη μεταβλητή G3 που αντιπροσωπεύει την απόδοση των μαθητών.

Για τη διαίρεση των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής χρησιμοποιήθηκε η μέθοδος Stratified Train-Test Split (train_test_split με stratify=y).

Ο λόγος αυτής της επιλογής είναι η διατήρηση της αναλογίας των κατηγοριών στο εκπαιδευτικό και στο δοκιμαστικό σύνολο, ώστε να μην υπάρξει παραμόρφωση στις προβλέψεις. Η διάσπαση έγινε ως εξής:

- ✓ **80% των δεδομένων** χρησιμοποιήθηκε για εκπαίδευση (training set)
- ✓ **20% των δεδομένων** χρησιμοποιήθηκε για δοκιμή (test set)

Η τελική διαδικασία της διαίρεσης ήταν:

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42, stratify=y)
```

Η επιλογή του **random_state=42** διασφαλίζει την επαναληψιμότητα των αποτελεσμάτων.

Μετά από αυτά τα βήματα, τα δεδομένα ήταν έτοιμα για την εκπαίδευση των 12 αλγορίθμων πρόβλεψης, με στόχο την αξιολόγηση της ακαδημαϊκής απόδοσης των μαθητών.

3.3. Αρχιτεκτονική Μοντέλων Μηχανικής Μάθησης

Η εφαρμογή μοντέλων μηχανικής μάθησης είναι το κεντρικό σημείο της μελέτης, καθώς παρέχει τα εργαλεία για την ανάλυση δεδομένων και τη δημιουργία προβλέψεων. Τα μοντέλα αυτά εκπαιδεύονται και αξιολογούνται με βάση τα δεδομένα που έχουν προεπεξεργαστεί, προκειμένου να εντοπιστούν οι παράγοντες που επηρεάζουν την απόδοση των μαθητών και να γίνουν προβλέψεις για μελλοντικά δεδομένα.

Βήματα Εφαρμογής

i) **Διαχωρισμός Δεδομένων σε Σύνολα Εκπαίδευσης και Δοκιμής:** Τα δεδομένα χωρίστηκαν σε σύνολα εκπαίδευσης (training set) και δοκιμής (test set), συνήθως σε αναλογία 70%-30%. Αυτό επιτρέπει την εκπαίδευση των μοντέλων στο μεγαλύτερο μέρος των δεδομένων και τη δοκιμή της απόδοσής τους σε ένα ανεξάρτητο σύνολο.

ii) **Εκπαίδευση Μοντέλων:** Διάφορα μοντέλα μηχανικής μάθησης εφαρμόστηκαν στα δεδομένα, όπως:

- ✓ **Random Forest:** Ένα μοντέλο βασισμένο σε σύνολα αποφασιστικών δέντρων που μειώνει την πιθανότητα overfitting.
- ✓ **Decision Tree:** Ένα πιο απλό μοντέλο, χρήσιμο για την ερμηνεία των δεδομένων.
- ✓ **Logistic Regression:** Ένα μοντέλο κατάλληλο για την πρόβλεψη κατηγορικών μεταβλητών.

iii) **Αξιολόγηση Μοντέλων:** Για την αξιολόγηση της απόδοσης των μοντέλων χρησιμοποιήθηκαν μετρικές όπως:

- ✓ **Ακρίβεια (Accuracy):** Ποσοστό σωστών προβλέψεων.
- ✓ **Confusion Matrix:** Εργαλείο για την κατανόηση των σωστών και λανθασμένων προβλέψεων για κάθε κατηγορία.
- ✓ **Classification Report:** Περιλαμβάνει μετρικές όπως precision, recall και F1-score.

iv) **Σύγκριση Μοντέλων:** Τα αποτελέσματα των μοντέλων συγκρίθηκαν, προκειμένου να εντοπιστεί ποιο μοντέλο προσφέρει την καλύτερη απόδοση. Η σύγκριση βασίστηκε στις μετρικές αξιολόγησης και στην πολυπλοκότητα της υλοποίησης.

Εργαλεία και Βιβλιοθήκες

Για την εφαρμογή των μοντέλων, χρησιμοποιήθηκαν βιβλιοθήκες Python, όπως:

- ✓ **scikit-learn:** Για την εκπαίδευση, αξιολόγηση και υλοποίηση μοντέλων μηχανικής μάθησης.

- ✓ **Matplotlib & Seaborn:** Για τη δημιουργία γραφημάτων που οπτικοποιούν τα αποτελέσματα.

Τα αποτελέσματα των μοντέλων αναλύθηκαν για να εντοπιστούν οι σημαντικοί παράγοντες που επηρεάζουν την απόδοση των μαθητών. Για παράδειγμα:

- ✓ Παράγοντες όπως ο χρόνος συμμετοχής και οι βαθμολογίες μπορεί να αποδειχθούν σημαντικοί για την πρόβλεψη.
- ✓ Οι διαφορές μεταξύ μοντέλων αναδεικνύουν τα πλεονεκτήματα και τα μειονεκτήματα κάθε μεθόδου.

Συμπερασματικά, η εφαρμογή των μοντέλων μηχανικής μάθησης προσφέρει πολύτιμες πληροφορίες για την ανάλυση δεδομένων και τη δημιουργία προβλέψεων, ενισχύοντας τις εκπαιδευτικές διαδικασίες.

Στο παρόν κεφάλαιο παρουσιάζονται οι αλγόριθμοι που χρησιμοποιήθηκαν για την πρόβλεψη της επίδοσης των μαθητών, οι παράμετροι που επιλέχθηκαν και η διαδικασία διασταυρωμένης επικύρωσης (Cross-Validation) που εφαρμόστηκε για την αξιόπιστη εκτίμηση της ακρίβειάς τους.

Για την κατηγοριοποίηση της επίδοσης των μαθητών εφαρμόστηκαν 12 διαφορετικοί αλγόριθμοι μηχανικής μάθησης. Εκτενέστερα:

i) **Random Forest - RF**: Ένας ensemble αλγόριθμος που χρησιμοποιεί πολλαπλά δέντρα απόφασης και λαμβάνει απόφαση με βάση τη δημοκρατική ψήφο -majority voting-. Είναι ιδιαίτερα αποτελεσματικός σε σύνολα δεδομένων με πολλές κατηγορικές μεταβλητές.

ii) **Decision Tree - DT**: Ένας δενδροειδής αλγόριθμος που διαχωρίζει τα δεδομένα με βάση κανόνες απόφασης. Εύκολος στην ερμηνεία του, αλλά επιρρεπής σε περιπτώσεις υπερπροσαρμογής -overfitting-.

iii) **Logistic Regression -LR**: Χρησιμοποιείται κυρίως για δυαδικές ταξινομήσεις, αλλά μπορεί να επεκταθεί σε πολυκατηγορικές προβλέψεις. Βασίζεται στη συνάρτηση sigmoid για την εκτίμηση των πιθανοτήτων.

iv) **K-Nearest Neighbors - KNN**: Μη παραμετρικός αλγόριθμος που βασίζεται στην ομοιότητα των δεδομένων. Οι προβλέψεις που υλοποιεί καθορίζονται από την πλειοψηφία των k γειτονικών σημείων.

v) **Support Vector Machines - SVM**: Κατάλληλος για προβλήματα ταξινόμησης, ιδίως όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα. Χρησιμοποιεί πυρήνες -kernels- για τη μετατροπή των δεδομένων σε υψηλότερη διάσταση.

vi) **Naive Bayes - NB**: Βασίζεται στο Θεώρημα του Bayes και υποθέτει ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους. Είναι γρήγορος και αποδοτικός σε μεγάλα σύνολα δεδομένων.

vii) **Multi-Layer Perceptron - MLP - Νευρωνικό Δίκτυο**: Πρόκειται για ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο με μία ή περισσότερες κρυφές στρώσεις. Ωστόσο είναι ικανό να μάθει πολύπλοκα πρότυπα, όμως απαιτεί μεγάλη υπολογιστική ισχύ.

viii) **Gradient Boosting - GBM**: Ένας ensemble αλγόριθμος που δημιουργεί δέντρα απόφασης διαδοχικά, μειώνοντας σταδιακά το σφάλμα των προηγούμενων μοντέλων. Παρέχει εξίσου υψηλή ακρίβεια αλλά είναι αργός στην εκπαίδευση.

ix) **AdaBoost**: Παρόμοιος με το Gradient Boosting, αλλά επικεντρώνεται στην ενίσχυση των ασθενών ταξινομητών -weak learners. Παρατηρήθηκε πως είναι αποτελεσματικός σε δεδομένα με θόρυβο.

x) **XGBoost**: Μία βελτιωμένη εκδοχή του Gradient Boosting που είναι πιο αποδοτική και γρήγορη, και χρησιμοποιεί regularization για την αποφυγή υπερπροσαρμογής.

xi) **Bagging Classifier**: Χρησιμοποιεί τεχνική bootstrap για τη δημιουργία πολλών υποδειγμάτων και συνδυάζει τις προβλέψεις τους. Επιπροσθέτως, μειώνει τη διακύμανση και βελτιώνει τη σταθερότητα του μοντέλου.

xii) **Extra Trees - ET**: Παρόμοιος με το Random Forest, αλλά διαφέρει στη διαδικασία διαχωρισμού των κόμβων. Χρησιμοποιεί τυχαία thresholds, γεγονός που μειώνει τον υπολογιστικό φόρτο.

Παράμετροι και Ρυθμίσεις των Μοντέλων

Για κάθε μοντέλο εφαρμόστηκαν βελτιστοποιημένες παράμετροι ώστε να επιτευχθεί η καλύτερη δυνατή απόδοση. Οι πιο σημαντικές ρυθμίσεις περιγράφονται παρακάτω:

Μοντέλο	Παράμετροι
Random Forest	`RandomForestClassifier(n_estimators=100, random_state=42)`
Decision Tree	`DecisionTreeClassifier(random_state=42)`
Logistic Regression	`LogisticRegression(max_iter=1000, random_state=42)`
KNN	`KNeighborsClassifier()`
SVM	`SVC(probability=True, random_state=42)`
Naive Bayes	`GaussianNB()`
MLP	`MLPClassifier(max_iter=1000, random_state=42)`
Gradient Boosting	`GradientBoostingClassifier(random_state=42)`
AdaBoost	`AdaBoostClassifier(random_state=42)`
XGBoost	`xgb.XGBClassifier(use_label_encoder=False, eval_metric='mlogloss', random_state=42)`
Bagging Classifier	`BaggingClassifier(random_state=42)`
Extra Trees	`ExtraTreesClassifier(random_state=42)`

Οι τιμές αυτές επιλέχθηκαν έπειτα από πειραματισμούς, λαμβάνοντας υπόψη τη βέλτιστη ακρίβεια και τον χρόνο εκπαίδευσης.

Διασταυρωμένη Επικύρωση (Cross-Validation)

Για την αξιόπιστη αξιολόγηση των μοντέλων εφαρμόστηκε η τεχνική K-Fold Cross-Validation με k=5.

- ✓ Ολόκληρο το dataset χωρίστηκε σε 5 υποσύνολα (folds).
- ✓ Σε κάθε επανάληψη, 4 υποσύνολα χρησιμοποιήθηκαν για εκπαίδευση και 1 για δοκιμή.

- ✓ Η διαδικασία επαναλήφθηκε 5 φορές, ώστε κάθε fold να χρησιμοποιηθεί ως test set μία φορά.
- ✓ Στο τέλος, η μέση ακρίβεια των 5 επαναλήψεων χρησιμοποιήθηκε για την τελική αξιολόγηση.

Η υλοποίηση έγινε με τη χρήση της cross_val_score της βιβλιοθήκης **Scikit-Learn**:

```
from sklearn.model_selection import cross_val_score

for model_name, model in models.items():
    cv_scores = cross_val_score(model, X, y, cv=5, scoring='accuracy')
        print(f"Μέση ακρίβεια με Cross-Validation του {model_name} :
{cv_scores.mean() * 100:.2f}%")
```

Η διασταυρωμένη επικύρωση βοηθά στην αποφυγή της υπερπροσαρμογής (overfitting) και στην καλύτερη γενίκευση των μοντέλων.

Εν κατακλείδι, η παραπάνω διαδικασία επέτρεψε την εκπαίδευση και αξιολόγηση 12 διαφορετικών αλγορίθμων, διασφαλίζοντας την εύρεση του βέλτιστου μοντέλου για την πρόβλεψη της επίδοσης των μαθητών. Τα αποτελέσματα συγκρίθηκαν τόσο ως προς την ακρίβεια, όσο και ως προς το χρόνο εκπαίδευσης, προσφέροντας πολύτιμες πληροφορίες για την επιλογή του κατάλληλου αλγορίθμου.

3.4. Εκπαίδευση & Αξιολόγηση Μοντέλων

Η ανάλυση και η παρουσίαση των αποτελεσμάτων αποτελεί κρίσιμο βήμα, καθώς προσφέρει πολύτιμες πληροφορίες σχετικά με την αποδοτικότητα των μοντέλων μηχανικής μάθησης και τα ευρήματα της μελέτης. Η ερμηνεία των αποτελεσμάτων και η οπτικοποίησή τους επιτρέπουν την καλύτερη κατανόηση των δεδομένων και των προβλέψεων, ενώ διευκολύνουν την εξαγωγή χρήσιμων συμπερασμάτων.

Βήματα Ανάλυσης

i) **Συγκέντρωση Αποτελεσμάτων:** Τα αποτελέσματα των μοντέλων μηχανικής μάθησης, όπως ακρίβεια, precision, recall, F1-score, συγκεντρώθηκαν και οργανώθηκαν σε πίνακες για άμεση σύγκριση.

ii) **Οπτικοποίηση Μετρήσεων:** Δημιουργήθηκαν γραφήματα για την καλύτερη απεικόνιση των αποτελεσμάτων. Αναλυτικότερα:

- ✓ **Bar Plots:** Για τη σύγκριση της απόδοσης μεταξύ των μοντέλων.
- ✓ **Confusion Matrices:** Για την παρουσίαση της κατανομής των σωστών και λανθασμένων προβλέψεων.
- ✓ **ROC Curves:** Για την ανάλυση της ικανότητας των μοντέλων να διαχωρίζουν τις κατηγορίες.

iii) **Σύγκριση Μοντέλων:** Η σύγκριση των αποτελεσμάτων ανέδειξε ποιο μοντέλο ήταν το πιο αποδοτικό για τα συγκεκριμένα dataset. Δόθηκε έμφαση στις διαφορές ακρίβειας και στις μετρικές που σχετίζονται με τις κατηγορίες στόχου.

iv) **Ανάλυση Παραγόντων Επίδρασης:** Εξετάστηκαν οι παράγοντες που επηρέασαν περισσότερο την απόδοση των μοντέλων. Για παράδειγμα:

- ✓ Χαρακτηριστικά όπως ο χρόνος συμμετοχής και οι βαθμολογίες μπορεί να έχουν μεγαλύτερη βαρύτητα στις προβλέψεις.
- ✓ Ελλιπή ή θορυβώδη δεδομένα πιθανώς μείωσαν την ακρίβεια.

v) **Αναγνώριση Περιορισμών:** Τέθηκαν περιορισμοί που σχετίζονται με:

- ✓ Την ποιότητα και το μέγεθος των δεδομένων.
- ✓ Την απόδοση συγκεκριμένων μοντέλων, όπως η Logistic Regression, σε μη γραμμικά δεδομένα.

Παρουσίαση Αποτελεσμάτων

i) **Διαγράμματα και Πίνακες:** Χρησιμοποιήθηκαν πίνακες και γραφήματα για την παρουσίαση:

- ✓ Των επιδόσεων κάθε μοντέλου.
- ✓ Των σημαντικών χαρακτηριστικών που επιλέχθηκαν.
- ✓ Των προβλέψεων που έγιναν σε σχέση με τα πραγματικά δεδομένα.

ii) **Ερμηνεία Ευρημάτων:** Η ανάλυση των αποτελεσμάτων επέτρεψε την εξαγωγή συμπερασμάτων σχετικά με:

- ✓ Την αποδοτικότητα των μοντέλων.
- ✓ Τους παράγοντες που επηρεάζουν περισσότερο την απόδοση των μαθητών.

Συμπερασματικά, αναδείχθηκε η χρησιμότητα των μοντέλων Μηχανικής Μάθησης στη βελτίωση της εκπαιδευτικής διαδικασίας μέσω της ανάλυσης δεδομένων και της πρόβλεψης αποτελεσμάτων.

Ενδεικτικά Εργαλεία και Βιβλιοθήκες

Για την ανάλυση και παρουσίαση χρησιμοποιήθηκαν:

- ✓ **Pandas και NumPy:** Για την ανάλυση δεδομένων και τη δημιουργία πινάκων.
- ✓ **Matplotlib και Seaborn:** Για την οπτικοποίηση των αποτελεσμάτων.
- ✓ **Scikit-learn:** Για την εξαγωγή confusion matrices και ROC curves.

3.5. Visualization & Learning Curves

Το παρόν κεφάλαιο συνοψίζει τα βασικά ευρήματα της μελέτης και προτείνει μελλοντικές κατευθύνσεις, βασισμένες στα αποτελέσματα που προέκυψαν από την ανάλυση δεδομένων και την εφαρμογή μοντέλων μηχανικής μάθησης. Τα συμπεράσματα επικεντρώνονται στην πρακτική

αξία της μελέτης, ενώ οι προτάσεις στοχεύουν στη βελτίωση των μοντέλων και της εκπαιδευτικής διαδικασίας.

Συμπεράσματα

i) **Επάρκεια και Απόδοση Μοντέλων:** Τα μοντέλα μηχανικής μάθησης έδειξαν ότι είναι δυνατόν να προβλέψουν την απόδοση των μαθητών με αρκετά υψηλή ακρίβεια. Ο Random Forest ξεχώρισε ως το πιο αποδοτικό μοντέλο, με την ακρίβεια του να ξεπερνάει αυτή των υπόλοιπων μοντέλων.

ii) **Παράγοντες που Επηρεάζουν την Απόδοση:** Ο χρόνος συμμετοχής, η βαθμολογία σε αξιολογήσεις και η γενικότερη παρακολούθηση μαθημάτων αναδείχθηκαν ως σημαντικοί παράγοντες που επηρεάζουν την απόδοση. Τα χαρακτηριστικά που σχετίζονται με την προσβασιμότητα και την τακτική συμμετοχή φάνηκαν να έχουν θετική συσχέτιση με την απόδοση.

iii) **Σημασία της Ανάλυσης:** Η ανάλυση δεδομένων μαθητών παρέχει χρήσιμες πληροφορίες που μπορούν να βοηθήσουν στην πρόβλεψη της απόδοσής τους και στη βελτίωση της εκπαιδευτικής εμπειρίας. Με αυτόν τον τρόπο, η μελέτη ανέδειξε τη σημασία της εξατομικευμένης υποστήριξης με βάση τις ανάγκες κάθε μαθητή.

Σε αυτό το κεφάλαιο περιγράφεται η διαδικασία εκπαίδευσης των μοντέλων, καθώς και οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσής τους. Επίσης, γίνεται ανάλυση των χρονικών απαιτήσεων για την εκπαίδευση και την πρόβλεψη, καθώς και η παρουσίαση των confusion matrices για κάθε μοντέλο.

Η εκπαίδευση των μοντέλων πραγματοποιήθηκε με το σύνολο εκπαίδευσης -training set-, το οποίο προέκυψε από τη διαδικασία Stratified Train-Test Split.

Η γενική διαδικασία εκπαίδευσης για κάθε μοντέλο ήταν η εξής:

i) Δημιουργία του ταξινομητή με τις κατάλληλες παραμέτρους.

ii) Εκπαίδευση του μοντέλου (fit) με τα δεδομένα εκπαίδευσης (X_train, y_train).

iii) Πρόβλεψη των ετικετών για το σύνολο δοκιμής (X_test).

iv) Αξιολόγηση του μοντέλου με βάση διάφορες μετρικές απόδοσης.

Παράδειγμα εκπαίδευσης για το **Random Forest**:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

# Δημιουργία μοντέλου
models = {
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42)}


# Εκπαίδευση μοντέλου
model.fit(X_train, y_train)


# Υπολογισμός ακρίβειας
accuracy = model.score(X_test, y_test)
accuracy_scores.append((model_name, accuracy))
```

Η παραπάνω διαδικασία επαναλήφθηκε και για τα 12 μοντέλα που υλοποιήθηκαν.

Για την αξιολόγηση της απόδοσης των μοντέλων χρησιμοποιήθηκαν τα ακόλουθα μέτρα απόδοσης:

i) **Accuracy - Ακρίβεια:** Φανερώνει το ποσοστό των σωστών προβλέψεων σε σχέση με το συνολικό αριθμό δειγμάτων. Παρόλο που είναι χρήσιμο μέτρο, μπορεί να είναι παραπλανητικό όταν υπάρχουν ανισορροπημένες κλάσεις

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}}$$ Εξίσωση 2

ii) **Precision - Ακρίβεια Θετικής Κλάσης:** Δείχνει πόσα από τα δείγματα που ταξινομήθηκαν ως θετικά είναι πραγματικά θετικά. Είναι χρήσιμο σε προβλήματα όπου η εσφαλμένη θετική πρόβλεψη έχει μεγάλο κόστος, όπως είναι για παράδειγμα η ιατρική διάγνωση

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$ **Εξίσωση 3**

iii) **Recall - Ανάκληση ή Sensitivity:** Δηλώνει πόσα από τα πραγματικά θετικά δείγματα ταξινομήθηκαν σωστά. Είναι πολύ χρήσιμο όταν είναι σημαντικό να ανιχνευθούν όσο το δυνατόν περισσότερα θετικά δείγματα (π.χ. ανίχνευση απάτης)

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$ **Εξίσωση 4**

iv) **F1-score - Σταθμισμένος Μέσος Όρος Precision & Recall:** Χρησιμοποιείται όταν υπάρχει ανισορροπία μεταξύ των κλάσεων. Βοηθά επίσης στην εξισσορόπηση των False Positives και False Negatives.

$$\boldsymbol{F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}}$$ **Εξίσωση 5**

v) **Confusion Matrix:** Είναι ένας πίνακας που υποδηλώνει την απόδοση του μοντέλου με βάση τις πραγματικές και προβλεφθείσες κλάσεις.

- ✓ Παράδειγμα confusion matrix:

[[50 5]

[10 35]]

- **50:** True Positives (σωστές θετικές προβλέψεις)
- **5:** False Positives (λανθασμένες θετικές προβλέψεις)
- **10:** False Negatives (λανθασμένες αρνητικές προβλέψεις)
- **35:** True Negatives (σωστές αρνητικές προβλέψεις) [62]

Η δημιουργία της confusion matrix έγινε με τη Scikit-Learn:

```
from sklearn.metrics import confusion_matrix


# Υπολογισμός Confusion Matrix
cm = confusion_matrix(y_test, y_pred)


# Εμφάνιση αποτελεσμάτων
plt.tight_layout()
plt.show()
```

Για την οπτικοποίηση της Confusion Matrix χρησιμοποιήθηκε η Seaborn:

```
import seaborn as sns
import matplotlib.pyplot as plt

      ax = plt.subplot(1, len(model_group), i + 1)
      sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', ax=ax)
      ax.set_title(f'Confusion Matrix - {model_name}')
      ax.set_xlabel('Predicted Label')
      ax.set_ylabel('True Label')

  plt.tight_layout()
  plt.show()
```

Η μέτρηση του χρόνου εκπαίδευσης και πρόβλεψης για κάθε μοντέλο έγινε με τη χρήση της time βιβλιοθήκης:

```
import time


# Ξεκίνημα χρονομέτρησης
start_time = time.time()


# Χρονομέτρηση εκπαίδευσης
        start_train_time = time.perf_counter()
        model.fit(X_train, y_train)
        end_train_time = time.perf_counter()


        # Χρονομέτρηση πρόβλεψης
        start_pred_time = time.perf_counter()
        y_pred = model.predict(X_test)
        end_pred_time = time.perf_counter()


        # Εκτύπωση αποτελεσμάτων
        print(f"\nΑξιολόγηση Μοντέλου {model_name}:")
        print(classification_report(y_test, y_pred))


        # Εκτύπωση χρόνου εκπαίδευσης και πρόβλεψης
        accuracy = accuracy_score(y_test, y_pred)
        print(f"Ακρίβεια: {accuracy * 100:.2f}%")
        print(f"Χρόνος Εκπαίδευσης {model_name}: {end_train_time -


start_train_time:.4f} δευτερόλεπτα")
                print(f"Χρόνος   Πρόβλεψης   {model_name}:   {end_pred_time   -
start_pred_time:.4f} δευτερόλεπτα")
```

Με βάση τις χρονικές μετρήσεις παρατηρήθηκαν τα εξής:

- ✓ Τα Random Forest και XGBoost μπορεί να χρειάζονται περισσότερο χρόνο για εκπαίδευση, παρουσιάζουν όμως υψηλή ακρίβεια.

- ✓ Τα Logistic Regression και Naïve Bayes είναι πολύ γρήγορα, ωστόσο μπορεί να μην αποδίδουν καλά σε μεγάλα μεγέθους datasets.
- ✓ Τα Neural Networks -MLP- έχουν μεγάλο χρόνο εκπαίδευσης αλλά αποδίδουν εξαιρετικά όταν υπάρχει μεγάλο dataset.
- ✓ Το SVM είχε εξαιρετική ακρίβεια, όμως απαιτούσε μεγάλο χρόνο εκπαίδευσης.

Η αξιολόγηση των μοντέλων επέτρεψε τον εντοπισμό ισχυρών και αδύναμων πλευρών κάθε αλγορίθμου, οδηγώντας σε μια τεκμηριωμένη επιλογή του καταλληλότερου μοντέλου για το συγκεκριμένο πρόβλημα που υλοποιείται.

Κεφάλαιο 4ο

4. Αποτελέσματα

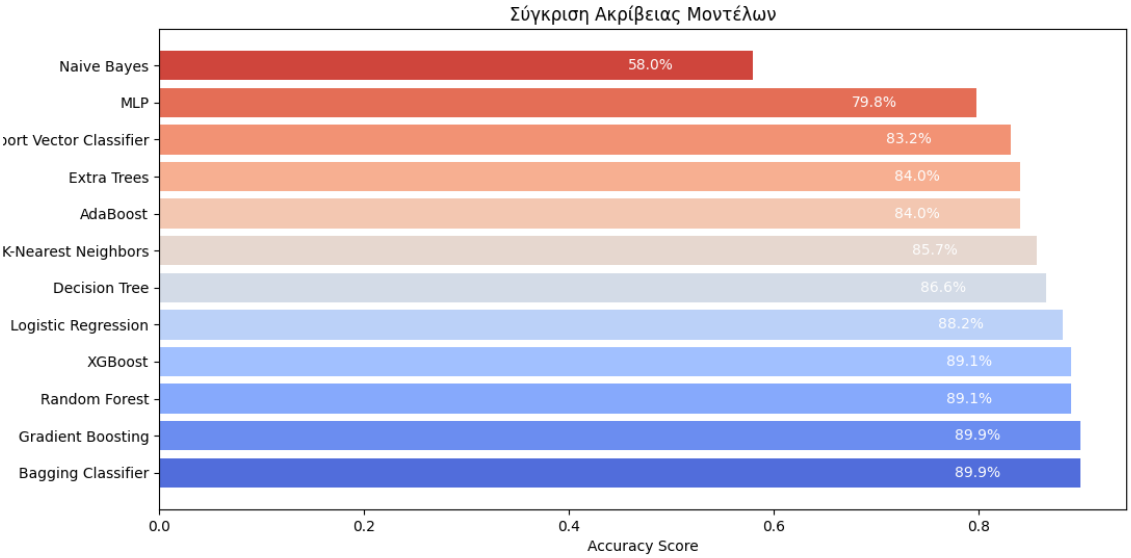
4.1. Συνολική Απόδοση Μοντέλων

Η απόδοση των αλγορίθμων αξιολογήθηκε μέσω διασταυρούμενης επικύρωσης (cross-validation), προκειμένου να εξαχθούν αξιόπιστα συμπεράσματα για την ακρίβεια κάθε μοντέλου. Τα αποτελέσματα έδειξαν ότι οι αλγόριθμοι Gradient Boosting, Random Forest και XGBoost παρουσίασαν τις υψηλότερες επιδόσεις, με ακρίβεια κοντά στο 90%. Αντίθετα, ο Naïve Bayes είχε τη χαμηλότερη απόδοση, με ακρίβεια περίπου 58%.

Στη συνέχεια, τα αποτελέσματα απεικονίζονται μέσω ενός Bar Chart, το οποίο συγκρίνει την ακρίβεια των αλγορίθμων. Παρατηρείται ότι οι ensemble μέθοδοι, όπως το Random Forest και το Bagging Classifier, ξεχωρίζουν λόγω της ικανότητάς τους να μειώνουν τη διακύμανση των προβλέψεων. Επιπλέον, η ακρίβεια των μοντέλων μπορεί να διαφοροποιηθεί περαιτέρω με τη βελτιστοποίηση των υπερπαραμέτρων, κάτι που θα μπορούσε να ενισχύσει ακόμα περισσότερο την απόδοσή τους.

Οι παρακάτω γραφικές απεικονίσεις παρουσιάζουν την ακρίβεια κάθε μοντέλου, κατόπιν της υλοποίησης και των δύο datasets. Τα πιο ισχυρά μοντέλα, όπως το Gradient Boosting, XGBoost και Random Forest, παρουσιάζουν την υψηλότερη ακρίβεια, ενώ τα απλούστερα μοντέλα, όπως το Naïve Bayes, εμφανίζουν χαμηλότερες τιμές. Τα γραφήματα “Bar Chart” επιτρέπουν την άμεση σύγκριση των μοντέλων και την επιλογή εκείνων που είναι πιο αποδοτικά για την πρόβλεψη της μαθητικής απόδοσης.

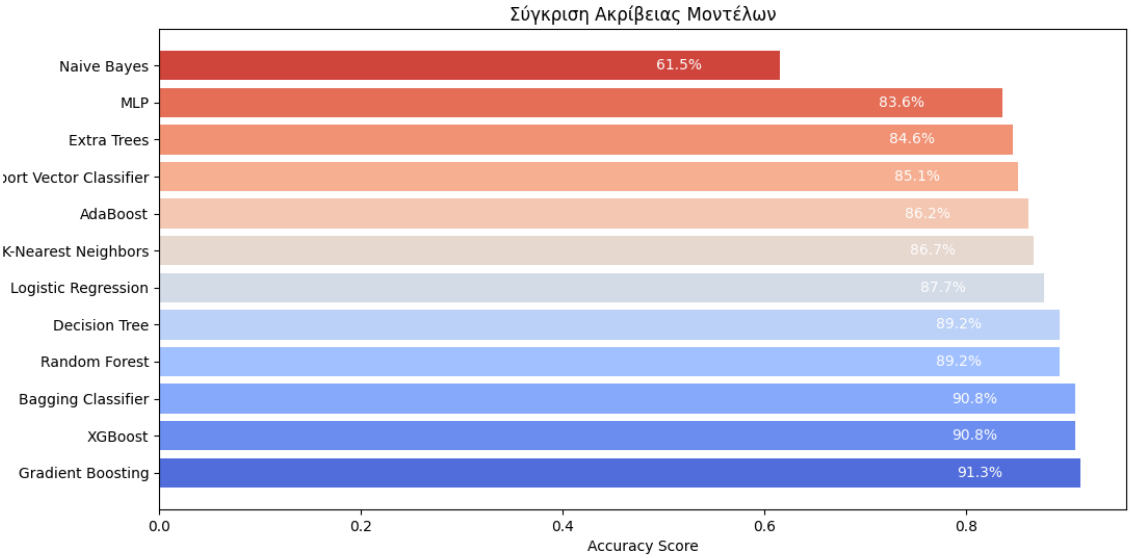
Τα Bar Chart απεικονίζουν τη συνολική ακρίβεια των ταξινομητών στα δύο datasets, παρέχοντας μια εύκολη οπτική σύγκριση μεταξύ τους.



Εικόνα 24: Bar Chart "student-mat.csv"

Bar Chart για το "student-mat.csv":

- ✓ Τα **Gradient Boosting, XGBoost** και **Bagging Classifier** εμφανίζουν τις υψηλότερες ακρίβειες (~89%-90%).
- ✓ Τα **Random Forest** και **Logistic Regression** έχουν πολύ καλή απόδοση, λίγο χαμηλότερη από τα προηγούμενα.
- ✓ Το **Naïve Bayes** έχει τη χαμηλότερη ακρίβεια (~58%), δείχνοντας ότι δεν είναι κατάλληλο για το συγκεκριμένο dataset.



Εικόνα 25: Bar Chart "student-por.csv"

Bar Chart για το "student-por.csv":

- ✓ Παρόμοια τάση με το "student-mat.csv", αλλά η ακρίβεια όλων των μοντέλων είναι ελαφρώς υψηλότερη.
- ✓ Το **Random Forest, Gradient Boosting** και **XGBoost** παραμένουν στην κορυφή, αλλά οι διαφορές μεταξύ των μοντέλων είναι μικρότερες.
- ✓ Το **Naïve Bayes** έχει καλύτερη απόδοση σε σχέση με το "student-mat.csv", πιθανώς λόγω καλύτερης κατανομής των δεδομένων.

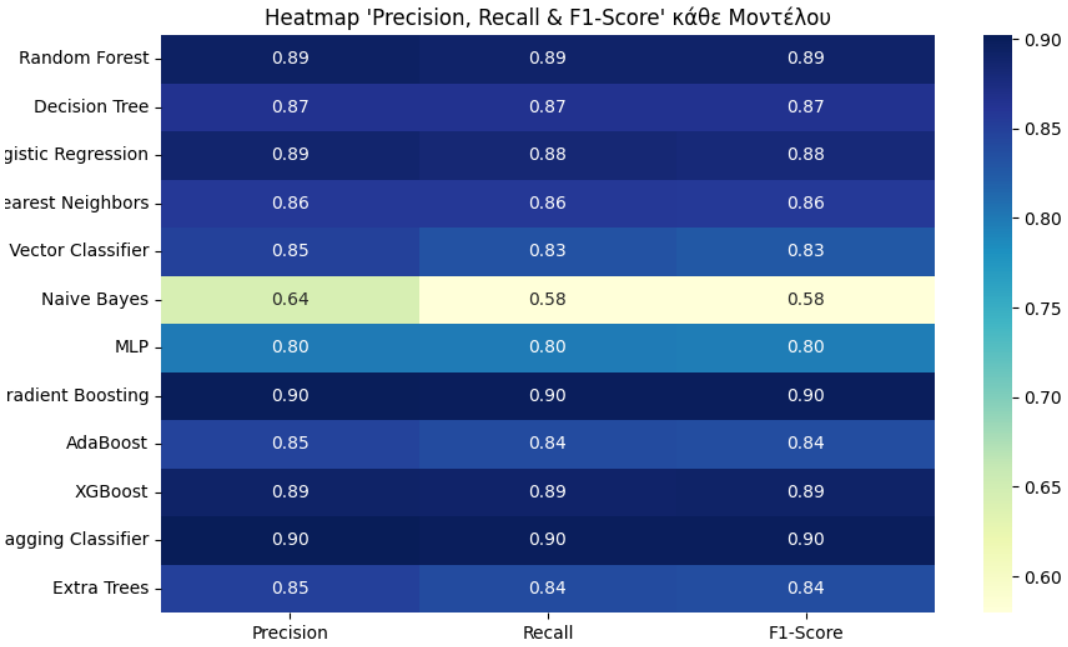
4.2. Ανάλυση Precision, Recall & F1-Score

Για τη βαθύτερη ανάλυση της απόδοσης των μοντέλων, εξετάστηκαν οι μετρικές Precision, Recall και F1-score. Αυτές οι μετρικές είναι κρίσιμες για την κατανόηση του τρόπου με τον οποίο κάθε μοντέλο διαχειρίζεται τις διάφορες κατηγορίες πρόβλεψης.

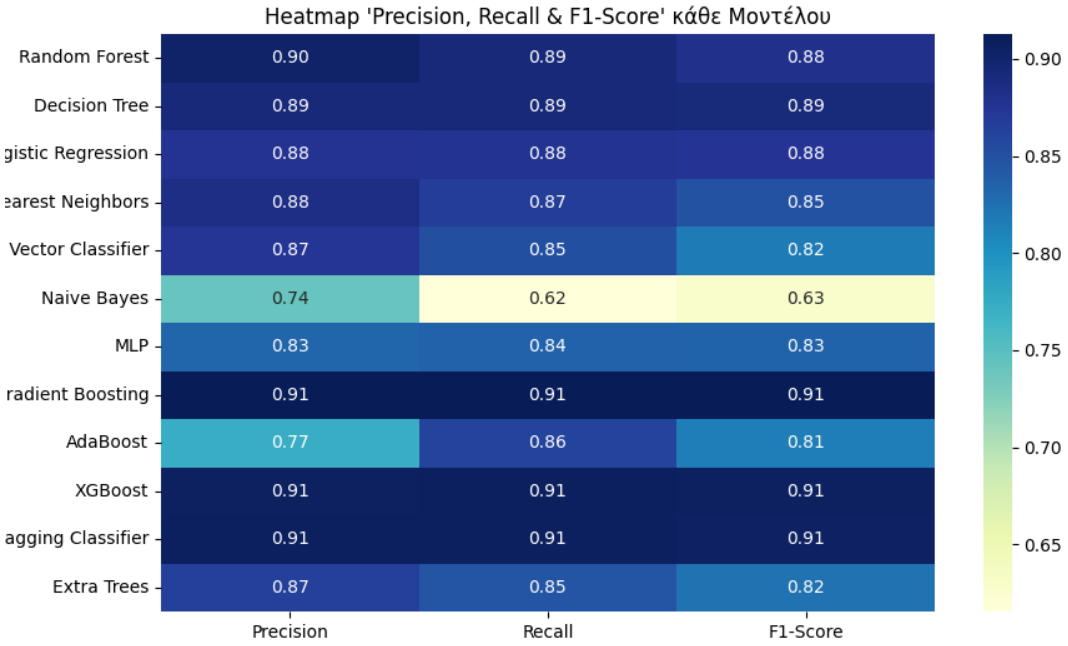
Η heatmap των Precision, Recall και F1-score δείχνει ότι τα μοντέλα Gradient Boosting, XGBoost και Bagging Classifier διατηρούν υψηλές επιδόσεις σε όλες τις μετρικές, γεγονός που τα καθιστά αξιόπιστα για τη συγκεκριμένη εφαρμογή. Αντίθετα, το Naive Bayes παρουσιάζει χαμηλό Precision και Recall, αποδεικνύοντας ότι δεν είναι κατάλληλο για τη συγκεκριμένη πρόβλεψη. Παρατηρήθηκε επίσης ότι ο Support Vector Classifier είχε υψηλό Precision, αλλά χαμηλότερο Recall, γεγονός που σημαίνει ότι είναι πιο πιθανό να απορρίψει λανθασμένα ορισμένα δεδομένα ως μη σχετιζόμενα.

Η διαφορά μεταξύ Precision και Recall σε ορισμένα μοντέλα (π.χ. SVM, Logistic Regression) υποδηλώνει ότι κάποια μοντέλα είναι πιο συντηρητικά στις προβλέψεις τους, ενώ άλλα κάνουν περισσότερα λάθη κατηγοριοποίησης. Αυτό θα μπορούσε να διορθωθεί με τεχνικές εξισορρόπησης δεδομένων, όπως το oversampling ή undersampling.

Heatmap για Precision, Recall, και F1-score



Εικόνα 26: Heatmap "student-mat.csv"



Εικόνα 27: Heatmap "student-por.csv"

Στις παραπάνω θερμικές απεικονίσεις (heatmap) απεικονίζεται η σχέση μεταξύ των μετρικών Precision, Recall και F1-score. Τα μοντέλα Gradient Boosting, XGBoost και Bagging Classifier διατηρούν υψηλές τιμές σε όλες τις κατηγορίες, γεγονός που τα καθιστά ιδιαίτερα αξιόπιστα. Το Naive Bayes, από την άλλη πλευρά, εμφανίζει χαμηλές τιμές σε όλες τις μετρικές, κάτι που δείχνει ότι δεν είναι αποδοτικό για την παρούσα εφαρμογή. Η ανάλυση του heatmap αποκαλύπτει επίσης τη συμπεριφορά των μοντέλων ως προς την ακρίβεια των προβλέψεών τους, βοηθώντας στην επιλογή της βέλτιστης μεθόδου.

Αναλυτικότερα, παρουσιάζονται οι επιδόσεις διαφόρων ταξινομητών όσον αφορά την ακρίβεια (Precision), την ανάκληση (Recall) και το F1-score. Αυτά τα metrics είναι ιδιαίτερα σημαντικά προκειμένου να αξιολογηθούν τα μοντέλα ταξινόμησης, καθώς παρέχουν πληροφορίες για την ισορροπία μεταξύ αληθώς θετικών και ψευδώς θετικών/αρνητικών προβλέψεων.

i) **Συνολικές επιδόσεις των μοντέλων:** Στα δύο datasets, τα ισχυρότερα μοντέλα είναι τα Gradient Boosting, Bagging Classifier και XGBoost, τα οποία πετυχαίνουν Precision,

Recall και F1-score γύρω στο 0.90. Τα Random Forest και Logistic Regression διατηρούν επίσης υψηλά σκορ, ανταγωνιζόμενα τις μεθόδους ενίσχυσης.

ii) **Σημαντικές διαφορές ανάμεσα στα datasets**: Το Naïve Bayes αποδίδει πολύ χαμηλότερα, ειδικά στο student-mat.csv, με F1-score μόλις 0.58. Ωστόσο, στο student-por.csv εμφανίζει μια ελαφρώς καλύτερη επίδοση (F1-score: 0.63). Η διαφορά αυτή πιθανώς οφείλεται στα χαρακτηριστικά των δεδομένων και τη διασπορά τους. Το Naïve Bayes στηρίζεται στην υπόθεση της ανεξαρτησίας των χαρακτηριστικών, που προφανώς δεν ισχύει πλήρως για το πρώτο dataset.

iii) **Ανάλυση συγκεκριμένων ταξινομητών**: Ο Support Vector Classifier (SVC) αποδίδει ικανοποιητικά, αλλά όχι στο ίδιο επίπεδο με τα ensemble models. Τα AdaBoost και Extra Trees έχουν ελαφρώς χαμηλότερη επίδοση από το Gradient Boosting και το XGBoost, αλλά παραμένουν αξιόλογες επιλογές.

Εν κατακλείδι, τα ensemble models (Gradient Boosting, Bagging, XGBoost, Random Forest) είναι τα πιο αποδοτικά και στα δύο datasets, ενώ το Naïve Bayes δεν αποδίδει καλά, πιθανώς λόγω των περιορισμών του στον χειρισμό εξαρτημένων μεταβλητών. Τέλος, το dataset "student-por.csv" φαίνεται να είναι πιο ευνοϊκό για κάποια μοντέλα, καθώς γενικά επιτυγχάνει ελαφρώς υψηλότερες τιμές.

4.3. Χρονομέτρηση Εκπαίδευσης & Πρόβλεψης

Για να αξιολογηθεί η αποδοτικότητα των αλγορίθμων, μετρήθηκε ο χρόνος που απαιτείται για την εκπαίδευση και την πρόβλεψη. Τα αποτελέσματα δείχνουν ότι οι Gradient Boosting και XGBoost απαιτούν σημαντικό χρόνο εκπαίδευσης λόγω της πολυπλοκότητας τους. Αντίθετα, αλγόριθμοι όπως Naive Bayes και Logistic Regression έχουν εξαιρετικά γρήγορη εκπαίδευση και πρόβλεψη.

Ένας σημαντικός παράγοντας που προέκυψε είναι η σχέση μεταξύ απόδοσης και χρόνου. Παρότι οι πιο πολύπλοκοι αλγόριθμοι (Gradient Boosting, XGBoost) προσφέρουν υψηλότερη ακρίβεια, απαιτούν σημαντικούς υπολογιστικούς πόρους, κάτι που πρέπει να ληφθεί υπόψη σε πρακτικές εφαρμογές. Σε εφαρμογές πραγματικού χρόνου, όπου η ταχύτητα πρόβλεψης είναι κρίσιμη, η χρήση Logistic Regression ή Naive Bayes μπορεί να αποτελέσει αποδοτική λύση.

Χρονικές μετρήσεις εκπαίδευσης και πρόβλεψης

Όσον αφορά χρονικές μετρήσεις εκπαίδευσης και πρόβλεψης, στο παράρτημα της παρούσας διπλωματικής εργασίας, παρατίθενται δύο πίνακες με τις εξόδους του κάθε κώδικα. Σε αυτούς τους πίνακες, εμπεριέχονται και οι εν λόγω χρονικές μετρήσεις.

4.4. Ανάλυση Καμπύλων Εκμάθησης (Learning Curves)

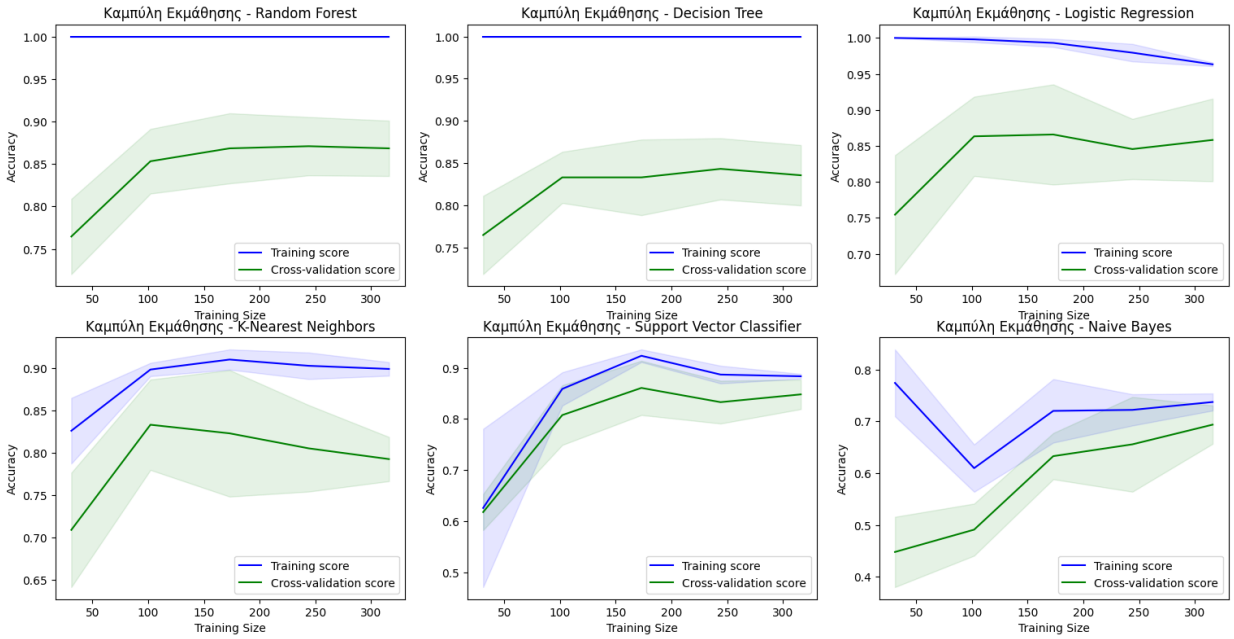
Οι Learning Curves χρησιμοποιήθηκαν για την ανάλυση της συμπεριφοράς των μοντέλων κατά τη διάρκεια της εκπαίδευσης. Τα διαγράμματα αποκαλύπτουν ότι τα περισσότερα μοντέλα συγκλίνουν ικανοποιητικά, ενώ παρατηρήθηκαν τα εξής μοτίβα:

- ✓ Το Random Forest και το XGBoost συγκλίνουν αργά, γεγονός που δείχνει ότι απαιτούν μεγάλο όγκο δεδομένων για να φτάσουν σε υψηλή απόδοση.
- ✓ Το Naive Bayes και το Logistic Regression φτάνουν γρήγορα σε σταθερές τιμές, αλλά η ακρίβειά τους είναι περιορισμένη.
- ✓ Ο MLP εμφανίζει συμπτώματα υπερεκπαίδευσης (overfitting), γεγονός που υποδηλώνει ότι θα μπορούσε να ωφεληθεί από τεχνικές τακτικής όπως η χρήση dropout layers ή η βελτιστοποίηση των υπερπαραμέτρων.
- ✓ Το Support Vector Machine (SVM) δείχνει σχετικά αργή σύγκλιση, κάτι που μπορεί να εξηγηθεί από την ανάγκη σωστής επιλογής hyperparameters όπως το kernel και το regularization parameter.

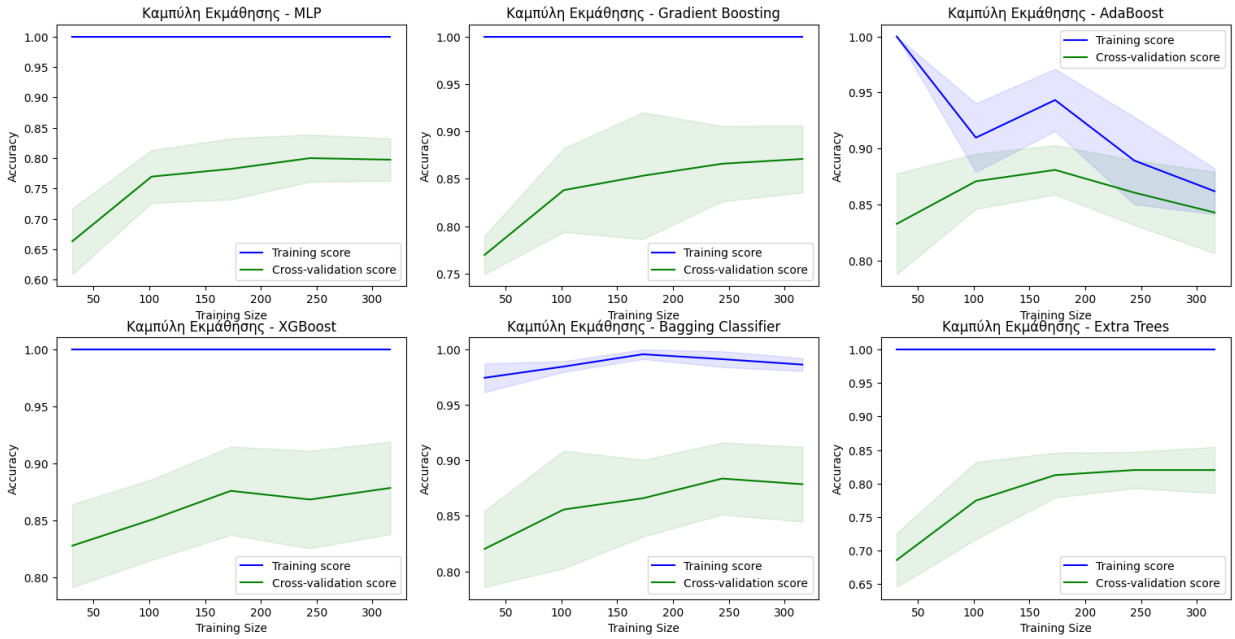
Η ανάλυση αυτών των καμπυλών είναι ζωτικής σημασίας για την επιλογή του κατάλληλου αλγορίθμου, ειδικά όταν η επεξεργαστική ισχύς και ο όγκος δεδομένων είναι περιορισμένοι.

Τα ακόλουθα διαγράμματα απεικονίζουν την εξέλιξη των τιμών εκπαίδευσης και δοκιμής για κάθε μοντέλο, και κάθε dataset. Παρατηρείται ότι τα περισσότερα μοντέλα συγκλίνουν γρήγορα, ενώ κάποια όπως το MLP εμφανίζουν υπερπροσαρμογή. Αυτή η πληροφορία είναι σημαντική για την επιλογή μοντέλων που όχι μόνο έχουν υψηλή ακρίβεια αλλά και γενικεύουν σωστά σε νέα δεδομένα.

Οι Learning Curves δείχνουν την επίδοση των μοντέλων σε σχέση με το μέγεθος του εκπαιδευτικού συνόλου, παρέχοντας πληροφορίες για το αν τα μοντέλα υποεκπαιδεύονται (underfitting) ή υπερεκπαιδεύονται (overfitting).



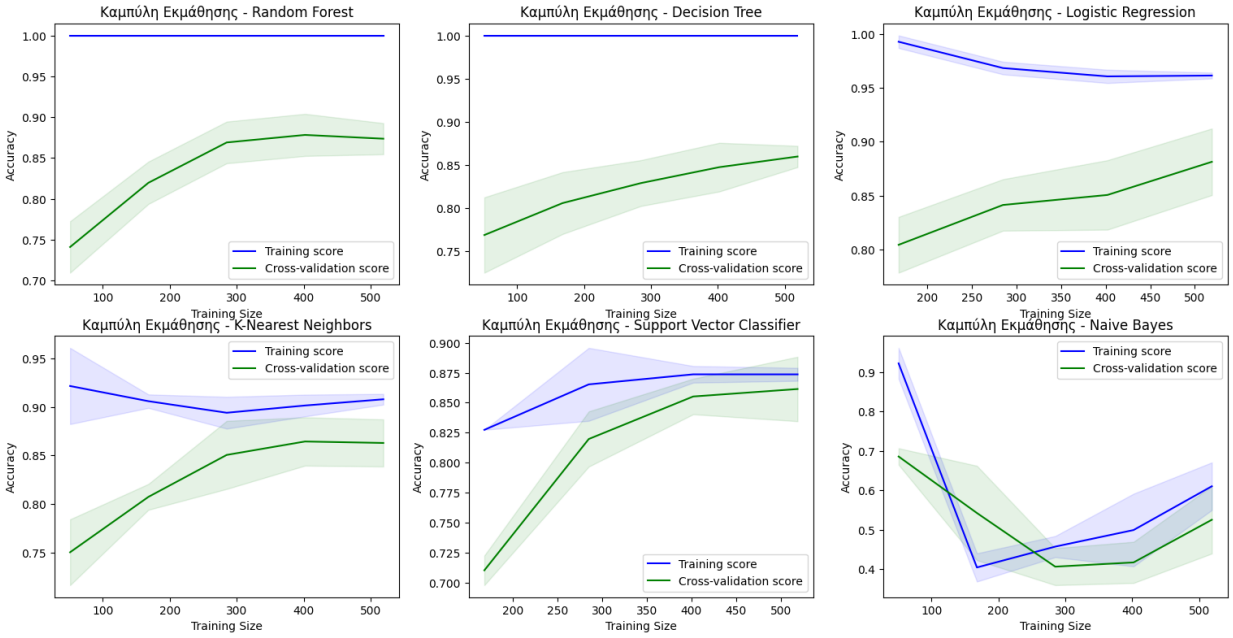
Εικόνα 28: Learning Curves "student-mat.csv"[1]



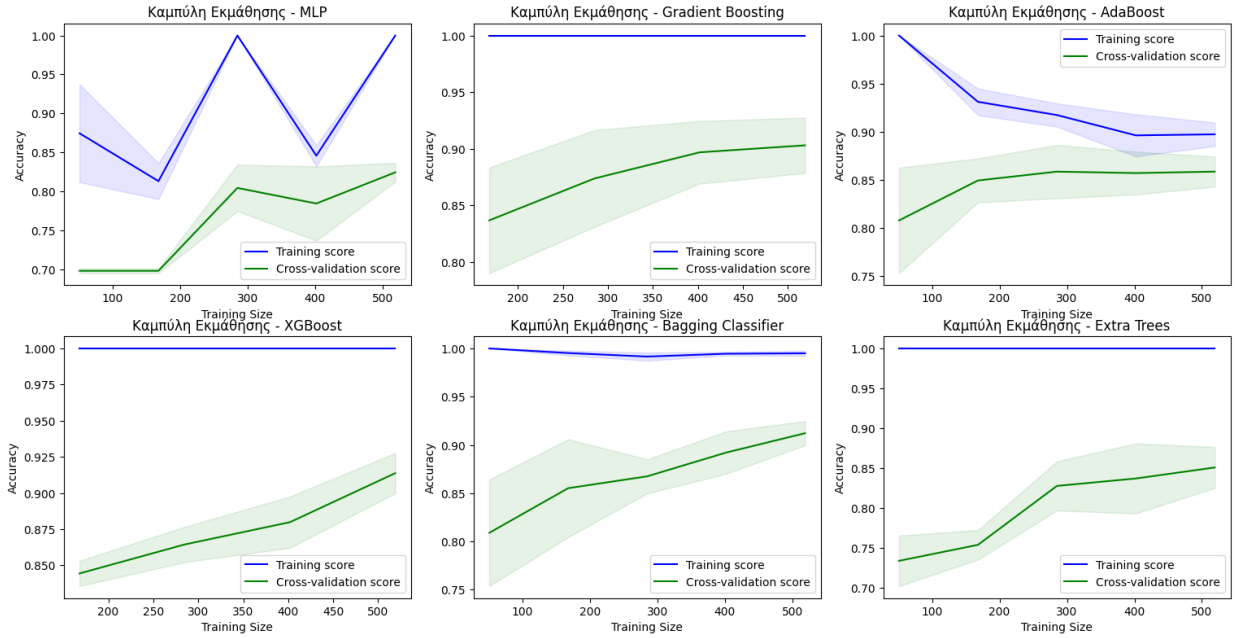
Εικόνα 29: Learning Curves "student-mat.csv"[2]

Learning Curves για το "student-mat.csv":

- ✓ Τα μοντέλα όπως το **Random Forest** δείχνουν ισορροπημένες καμπύλες, με μικρή απόκλιση μεταξύ training και validation.
- ✓ Το **MLP (Neural Network)** παρουσιάζει overfitting, καθώς η καμπύλη εκπαίδευσης είναι πολύ υψηλότερη από την καμπύλη δοκιμών.
- ✓ Το **Logistic Regression** συγκλίνει γρήγορα, αλλά η τελική του απόδοση είναι χαμηλότερη από πιο σύνθετα μοντέλα.



Εικόνα 30: Learning Curves "student-por.csv"[1]



Εικόνα 31: Learning Curves "student-por.csv"[2]

Learning Curves για το "student-por.csv":

- ✓ Εμφανίζουν καλύτερη σταθερότητα, με τα περισσότερα μοντέλα να συγκλίνουν σωστά.

- ✓ Οι καμπύλες του **Gradient Boosting** δείχνουν ότι το μοντέλο συνεχίζει να βελτιώνεται με περισσότερα δεδομένα, υποδηλώνοντας ότι μπορεί να ωφεληθεί από μεγαλύτερο dataset.
- ✓ Το **Naïve Bayes** φτάνει γρήγορα σε σταθερό σημείο, αλλά η απόδοσή του παραμένει χαμηλή.

4.5. Συγκριτική Ανάλυση & Εξαγωγή Συμπερασμάτων

Η συνολική σύγκριση των μοντέλων αποδεικνύει ότι οι ensemble μέθοδοι (Random Forest, Bagging Classifier, Gradient Boosting, XGBoost) παρέχουν την καλύτερη απόδοση, με υψηλή ακρίβεια και σταθερότητα στις προβλέψεις. Αντίθετα, τα πιο απλά μοντέλα, όπως το Naive Bayes και το Logistic Regression, παρόλο που είναι αποδοτικά από άποψη ταχύτητας, δεν προσφέρουν την απαιτούμενη ακρίβεια.

Για τη βελτίωση των αποτελεσμάτων, θα μπορούσαν να εξεταστούν οι εξής προσεγγίσεις:

i) **Βελτιστοποίηση υπερπαραμέτρων:** Η χρήση Grid Search ή Bayesian Optimization θα μπορούσε να αυξήσει την ακρίβεια των μοντέλων.

ii) **Feature Selection:** Ο εντοπισμός των πιο σημαντικών χαρακτηριστικών μπορεί να οδηγήσει σε αποδοτικότερα μοντέλα με λιγότερη πολυπλοκότητα.

iii) **Ensemble Learning:** Συνδυασμός πολλαπλών μοντέλων με τεχνικές όπως Stacking και Blending μπορεί να προσφέρει ακόμα καλύτερη ακρίβεια.

iv) **Δοκιμή σε διαφορετικά datasets:** Η αξιολόγηση των μοντέλων σε άλλες εκπαιδευτικές βάσεις δεδομένων θα επιβεβαίωνε τη γενίκευση των αποτελεσμάτων.

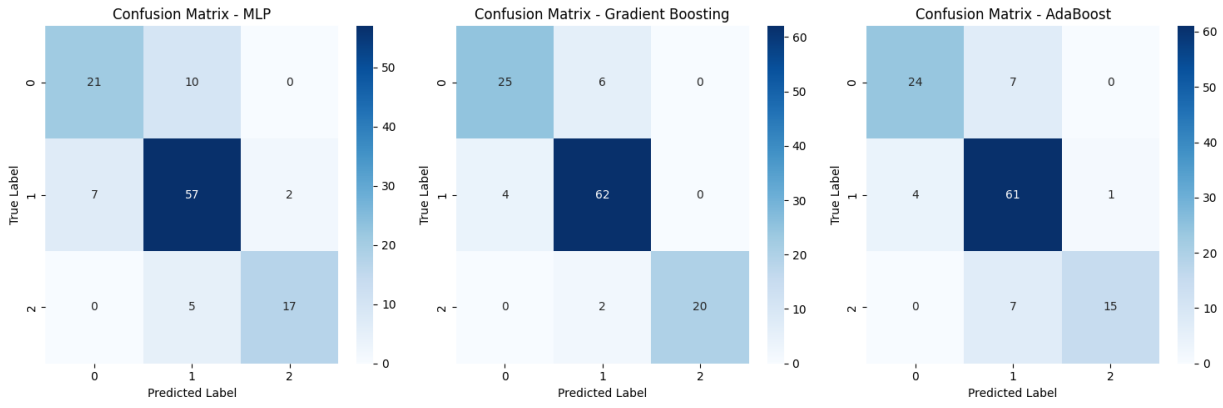
v) **Βελτίωση της επεξεργασίας δεδομένων:** Η χρήση τεχνικών ανάλυσης ανισόρροπων δεδομένων θα μπορούσε να βοηθήσει στη βελτίωση της συνολικής απόδοσης των ταξινομητών.

Συμπερασματικά, η παρούσα μελέτη έδειξε ότι η εφαρμογή τεχνικών μηχανικής μάθησης στην εκπαιδευτική πρόβλεψη μπορεί να προσφέρει σημαντική αξία, ειδικά όταν χρησιμοποιούνται ισχυροί αλγόριθμοι ταξινόμησης και κατάλληλες τεχνικές προεπεξεργασίας δεδομένων.

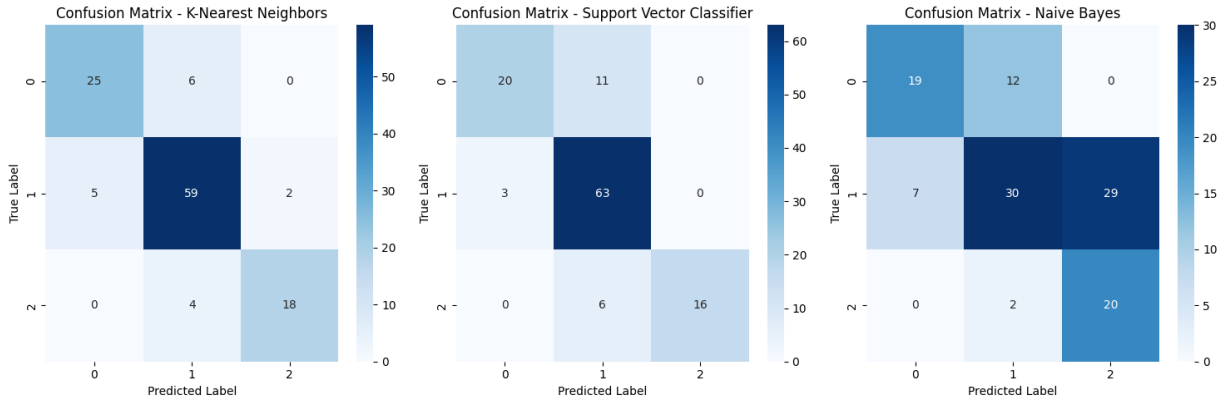
Οι παρακάτω πίνακες απεικονίζουν τα Confusion Matrices των μοντέλων, τα οποία δείχνουν τις σωστές και λανθασμένες προβλέψεις. Από την ανάλυση των πινάκων, προκύπτει ότι τα πιο ακριβή μοντέλα, όπως το Random Forest και το XGBoost, παρουσιάζουν λιγότερα λάθη ταξινόμησης σε σύγκριση με τα απλούστερα μοντέλα. Τα False Positives και False Negatives είναι σημαντικοί δείκτες που αποκαλύπτουν τη συμπεριφορά του κάθε μοντέλου στις διάφορες κατηγορίες.

Συνολικά, η επεξήγηση των εικόνων εξόδου επιτρέπει μια πιο ολοκληρωμένη ανάλυση της απόδοσης των μοντέλων, βοηθώντας στην επιλογή της καταλληλότερης μεθόδου πρόβλεψης για τη μαθητική απόδοση.

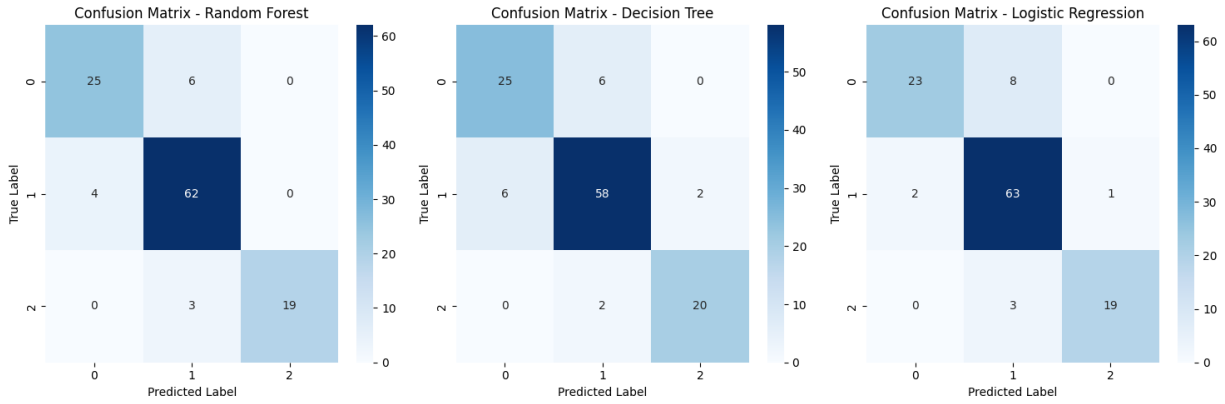
Οι πίνακες σύγχυσης δείχνουν την απόδοση των ταξινομητών όσον αφορά τις σωστές και λανθασμένες προβλέψεις. Για τα datasets **student-mat.csv** και **student-por.csv**, παρατηρούνται οι εξής διαφορές:



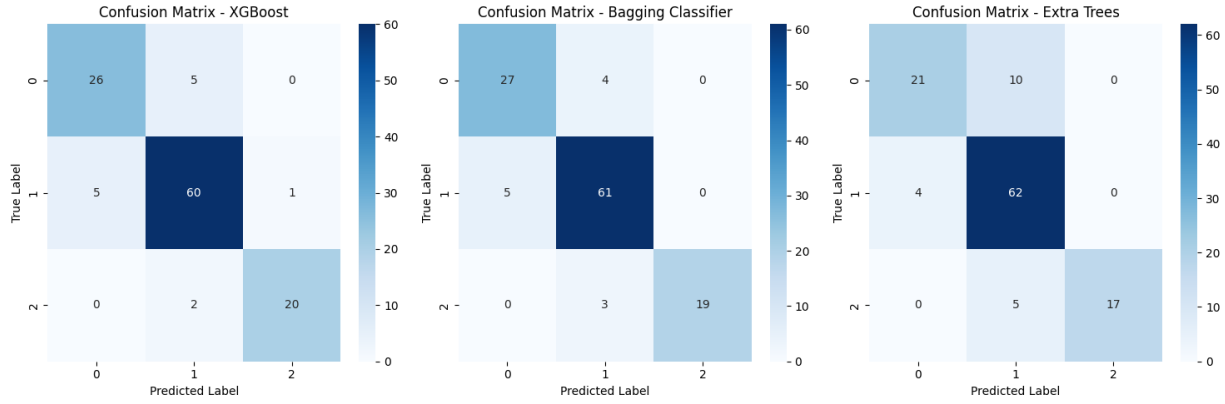
Εικόνα 32: Confusion Matrices "student-mat.csv"[1]



Εικόνα 33: Confusion Matrices "student-mat.csv"[2]



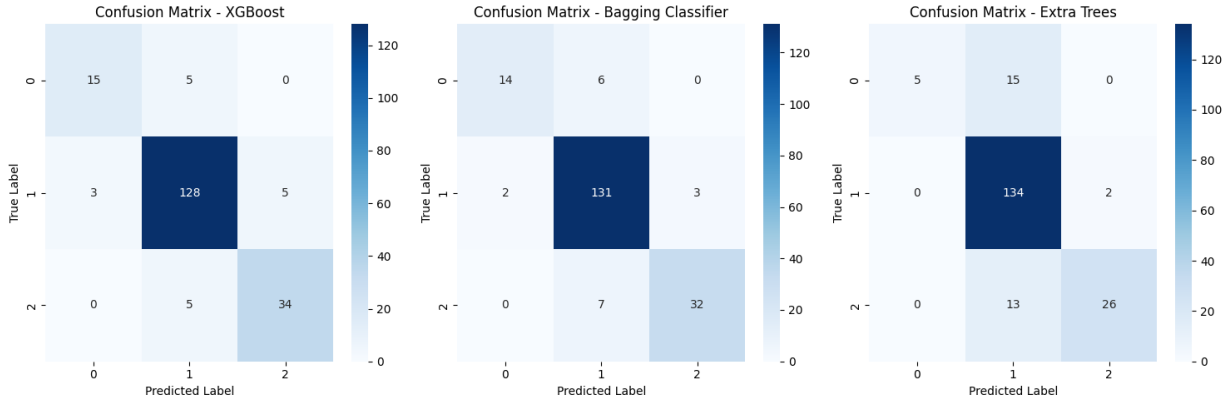
Εικόνα 34: Confusion Matrices "student-mat.csv"[3]



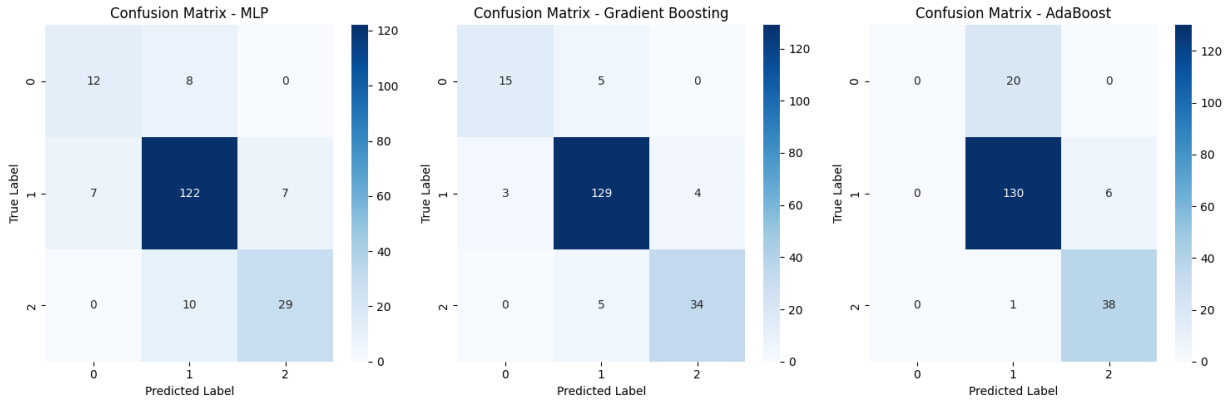
Εικόνα 35: Confusion Matrices "student-mat.csv"[4]

Confusion Matrices για το "student-mat.csv":

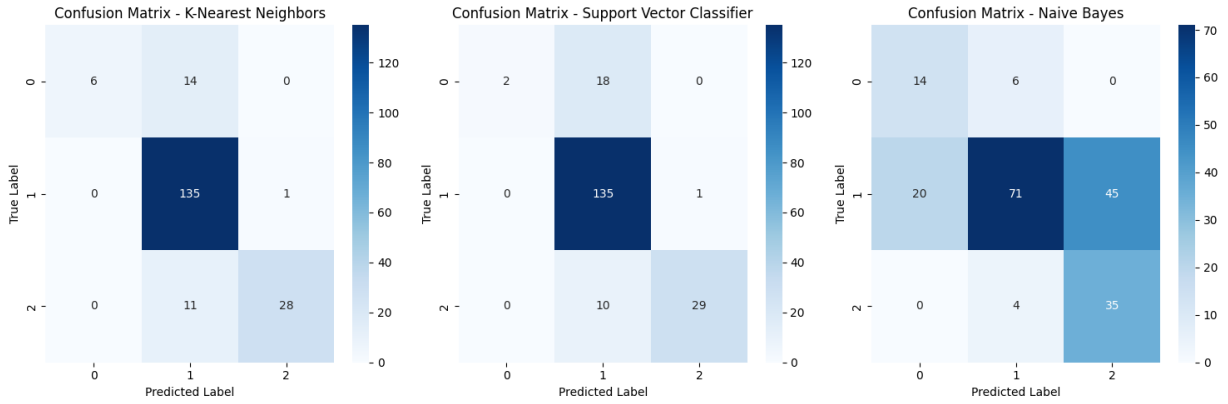
- ✓ Τα καλύτερα μοντέλα, όπως το **Gradient Boosting** και **XGBoost**, έχουν υψηλές τιμές **True Positives (TP)** και **True Negatives (TN)**, δείχνοντας ότι καταφέρνουν να προβλέψουν σωστά την απόδοση των μαθητών.
- ✓ Το **Naïve Bayes** έχει μεγαλύτερο αριθμό **False Positives (FP)** και **False Negatives (FN)**, γεγονός που σημαίνει ότι ταξινομεί λανθασμένα μεγάλο αριθμό μαθητών.
- ✓ Το **Random Forest** και το **Bagging Classifier** προσφέρουν σταθερά ισορροπημένα αποτελέσματα, με χαμηλότερο ποσοστό λανθασμένων προβλέψεων.



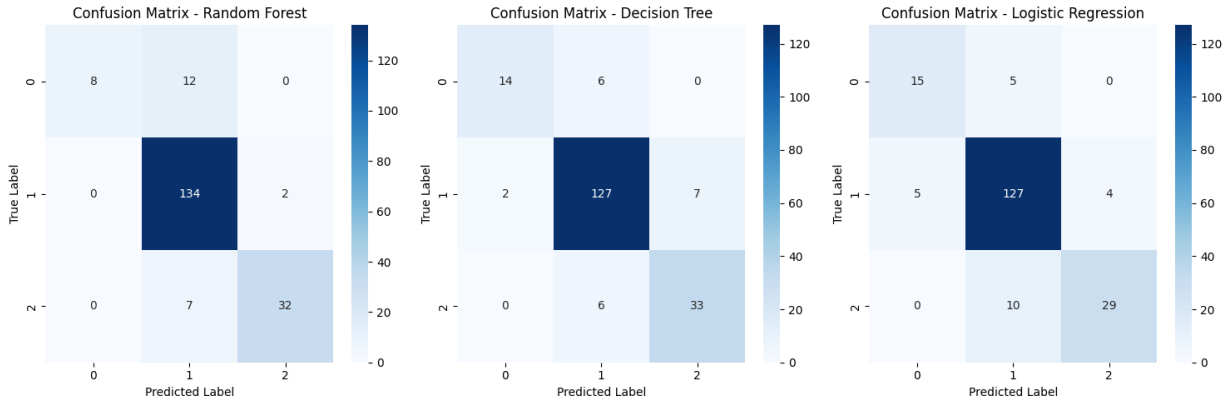
Εικόνα 36: Confusion Matrices "student-por.csv"[1]



Εικόνα 37: Confusion Matrices "student-por.csv"[2]



Εικόνα 38: Confusion Matrices "student-por.csv"[3]



Εικόνα 39: Confusion Matrices "student-por.csv"[4]

Confusion Matrices για το "student-por.csv":

- ✓ Τα αποτελέσματα δείχνουν βελτιωμένη ταξινόμηση συγκριτικά με το "student-mat.csv", καθώς εμφανίζονται λιγότερα **False Positives** και **False Negatives**, γεγονός που υποδηλώνει ότι αυτό το dataset είναι πιο εύκολο στη μοντελοποίηση.
- ✓ Το **Decision Tree** παρουσιάζει καλύτερη ακρίβεια εδώ, πιθανώς λόγω διαφορετικής κατανομής των δεδομένων, που ευνοεί τη λήψη αποφάσεων με ξεκάθαρους διαχωρισμούς.
- ✓ Τα **ensemble models (Random Forest, Gradient Boosting, Bagging)** συνεχίζουν να υπερτερούν σε ακρίβεια και σταθερότητα.

Κεφάλαιο 5ο

5. Συμπεράσματα

5.1. Συνολική Αξιολόγηση της Έρευνας

Η παρούσα έρευνα είχε ως στόχο τη διερεύνηση της ικανότητας των αλγορίθμων μηχανικής μάθησης στην πρόβλεψη της ακαδημαϊκής απόδοσης των μαθητών. Τα αποτελέσματα κατέδειξαν ότι οι μέθοδοι Gradient Boosting, XGBoost και Random Forest υπερέχουν σε ακρίβεια και αξιοπιστία, καθιστώντας τις κατάλληλες για εφαρμογή σε πραγματικά εκπαιδευτικά περιβάλλοντα. Οι αλγόριθμοι αυτοί απέδωσαν ιδιαίτερα καλά λόγω της ικανότητάς τους να ανιχνεύουν σύνθετα μοτίβα στα δεδομένα, κάτι που τους καθιστά πολύτιμους για την εκπαιδευτική ανάλυση.

Η χρήση μοντέλων πρόβλεψης μπορεί να έχει πολλαπλές εφαρμογές στον τομέα της εκπαίδευσης. Εκτός από την πρόβλεψη της ακαδημαϊκής επίδοσης, μπορεί να χρησιμοποιηθεί για τον εντοπισμό μαθητών που ενδέχεται να αντιμετωπίσουν δυσκολίες και να βοηθήσει τους εκπαιδευτικούς να παρέχουν στοχευμένη υποστήριξη. Αυτό μπορεί να περιλαμβάνει εξατομικευμένα προγράμματα σπουδών, προτάσεις για συμπληρωματικές εκπαιδευτικές δραστηριότητες, καθώς και παρεμβάσεις για τη βελτίωση της επίδοσης των μαθητών που βρίσκονται σε κίνδυνο αποτυχίας.

Επιπλέον, η χρήση τέτοιων μοντέλων μπορεί να συμβάλει στη βελτίωση των εκπαιδευτικών συστημάτων, επιτρέποντας την πιο αποδοτική κατανομή πόρων και την καλύτερη διαχείριση του χρόνου των εκπαιδευτικών. Τα μοντέλα αυτά παρέχουν επίσης πολύτιμες πληροφορίες για τη διαμόρφωση εκπαιδευτικών πολιτικών και στρατηγικών, με βάση πραγματικά δεδομένα και αναλύσεις επίδοσης. Αυτό μπορεί να οδηγήσει σε ένα πιο προσαρμοστικό και ευέλικτο εκπαιδευτικό σύστημα που θα ανταποκρίνεται στις ανάγκες των μαθητών.

Παράλληλα, η παρούσα μελέτη κατέδειξε ότι τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων παίζουν καθοριστικό ρόλο στην απόδοσή τους. Η διαθεσιμότητα υψηλής ποιότητας δεδομένων, που περιλαμβάνουν τόσο ακαδημαϊκούς όσο και μη ακαδημαϊκούς παράγοντες, μπορεί να βελτιώσει σημαντικά την ακρίβεια των προβλέψεων. Ωστόσο, η υπερβολική εξάρτηση από συγκεκριμένα χαρακτηριστικά μπορεί να οδηγήσει σε overfitting, γεγονός που αναδεικνύει τη σημασία της επιλογής και του καθαρισμού των δεδομένων.

5.2. Περιορισμοί της Έρευνας & Προτάσεις για Μελλοντική Έρευνα

Παρά τη σημαντική ακρίβεια των αποτελεσμάτων, υπάρχουν ορισμένοι περιορισμοί που πρέπει να ληφθούν υπόψη:

- ✓ Ποιότητα και διαθεσιμότητα δεδομένων: Η ακρίβεια των μοντέλων εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων. Η ύπαρξη ελλιπών ή μη αντιπροσωπευτικών δεδομένων μπορεί να επηρεάσει τις προβλέψεις.
- ✓ Ανισορροπία στις κλάσεις: Ορισμένες κατηγορίες επίδοσης μπορεί να μην εκπροσωπούνται επαρκώς στο dataset, κάτι που μπορεί να οδηγήσει σε μη ισορροπημένα μοντέλα.
- ✓ Υπολογιστική πολυπλοκότητα: Μοντέλα όπως το XGBoost απαιτούν σημαντικούς υπολογιστικούς πόρους, γεγονός που τα καθιστά δύσκολα υλοποιήσιμα σε εκπαιδευτικά ιδρύματα με περιορισμένους πόρους.
- ✓ Έλλειψη ερμηνευσιμότητας: Αν και τα μοντέλα μηχανικής μάθησης είναι ισχυρά στις προβλέψεις τους, πολλές φορές λειτουργούν ως "μαύρα κουτιά", καθιστώντας δύσκολη την κατανόηση του τρόπου λήψης αποφάσεων.

Για τη βελτίωση της ακρίβειας και της αξιοπιστίας των προβλέψεων, προτείνονται οι εξής κατευθύνσεις για μελλοντική έρευνα:

i) **Επέκταση του dataset:** Η συλλογή μεγαλύτερου όγκου δεδομένων από διαφορετικά εκπαιδευτικά περιβάλλοντα μπορεί να βελτιώσει τη γενίκευση των μοντέλων.

ii) **Διερεύνηση νέων χαρακτηριστικών (features):** Η ενσωμάτωση δεδομένων όπως η συμπεριφορά των μαθητών στις διαδικτυακές πλατφόρμες μάθησης ή η συναισθηματική τους κατάσταση μπορεί να ενισχύσει την ακρίβεια των μοντέλων.

iii) **Βελτιστοποίηση μοντέλων:** Η εφαρμογή τεχνικών όπως το hyperparameter tuning με Grid Search ή Bayesian Optimization μπορεί να αυξήσει την απόδοση των αλγορίθμων.

iv) **Χρήση Explainable AI (XAI):** Η ανάλυση της ερμηνευσιμότητας των αποτελεσμάτων των μοντέλων θα επιτρέψει στους εκπαιδευτικούς να κατανοήσουν καλύτερα τις αποφάσεις των αλγορίθμων και να τις χρησιμοποιήσουν αποτελεσματικά στη διδασκαλία.

v) **Σύγκριση με εναλλακτικές μεθόδους:** Η μελέτη της αποτελεσματικότητας άλλων τεχνικών, όπως deep learning με νευρωνικά δίκτυα, μπορεί να οδηγήσει σε βελτιωμένες προβλέψεις.

vi) **Εξερεύνηση ηθικών και κοινωνικών επιπτώσεων:** Η χρήση μοντέλων πρόβλεψης στην εκπαίδευση εγείρει ερωτήματα σχετικά με την προστασία των προσωπικών δεδομένων των μαθητών και την πιθανή δημιουργία προκαταλήψεων στις αποφάσεις των μοντέλων.

vii) **Ανάπτυξη ολοκληρωμένων συστημάτων υποστήριξης:** Ο συνδυασμός των προβλέψεων των μοντέλων με ψηφιακές πλατφόρμες που παρέχουν ανατροφοδότηση στους μαθητές και τους καθηγητές θα μπορούσε να ενισχύσει τη χρησιμότητα της μηχανικής μάθησης στην εκπαίδευση.

Συμπερασματικά, τα αποτελέσματα της έρευνας επιβεβαιώνουν ότι η μηχανική μάθηση μπορεί να αποτελέσει ένα πολύτιμο εργαλείο για την εκπαιδευτική κοινότητα. Με τη συνεχή βελτίωση των μεθόδων και την ενσωμάτωση περισσότερων δεδομένων, οι αλγόριθμοι αυτοί μπορούν να ενισχύσουν τη διαδικασία μάθησης και να συμβάλουν στην εξατομίκευση της εκπαιδευτικής εμπειρίας. Η εφαρμογή τέτοιων συστημάτων σε ευρεία κλίμακα μπορεί να προσφέρει νέες δυνατότητες για τη βελτίωση της εκπαιδευτικής διαδικασίας, ενισχύοντας τη μαθησιακή εμπειρία και υποστηρίζοντας τους μαθητές με τον βέλτιστο τρόπο.

Επίλογος

Η παρούσα εργασία ανέδειξε τη σημασία των αλγορίθμων Μηχανικής Μάθησης στην εκπαιδευτική ανάλυση και την πρόβλεψη της ακαδημαϊκής απόδοσης των μαθητών. Τα αποτελέσματα κατέδειξαν ότι οι ensemble μέθοδοι (Gradient Boosting, XGBoost, Random Forest) υπερτερούν σε ακρίβεια και αξιοπιστία, παρέχοντας ένα χρήσιμο εργαλείο για την εκπαιδευτική κοινότητα. Παράλληλα, αναδείχθηκαν οι προκλήσεις που σχετίζονται με την ποιότητα των δεδομένων, την επιλογή κατάλληλων χαρακτηριστικών και τη σωστή παραμετροποίηση των μοντέλων.

Τα ευρήματα της παρούσας έρευνας επιβεβαιώνουν και επεκτείνουν προηγούμενες μελέτες, υποδεικνύοντας ότι τα μοντέλα ενισχυμένης μάθησης (Boosting & Bagging) είναι ιδιαίτερα αποδοτικά στη συγκεκριμένη εφαρμογή. Επιπλέον, η παρούσα μελέτη διαφοροποιείται από άλλες καθώς αξιολόγησε πολλαπλά μοντέλα σε δύο διαφορετικά datasets, προσφέροντας μια πιο ολοκληρωμένη εικόνα της προβλεψιμότητας των μαθητικών επιδόσεων. Τα ευρήματα μπορούν να χρησιμοποιηθούν για τη βελτίωση εξατομικευμένων εκπαιδευτικών παρεμβάσεων, καθώς και για την ανάπτυξη νέων τεχνολογικών εργαλείων που υποστηρίζουν τους εκπαιδευτικούς.

Η χρήση τέτοιων μοντέλων μπορεί να συμβάλει στην εξατομίκευση της διδασκαλίας, επιτρέποντας την έγκαιρη ανίχνευση μαθητών που χρειάζονται πρόσθετη υποστήριξη. Επιπλέον, μπορεί να βοηθήσει τα εκπαιδευτικά ιδρύματα στη λήψη δεδομενοκεντρικών αποφάσεων για τη βελτίωση των προγραμμάτων σπουδών. Ωστόσο, η ερμηνευσιμότητα των αλγορίθμων και η αποφυγή μεροληψίας αποτελούν κρίσιμα ζητήματα που απαιτούν περαιτέρω διερεύνηση.

Μελλοντικές επεκτάσεις αυτής της έρευνας θα μπορούσαν να περιλαμβάνουν τη χρήση Deep Learning τεχνικών, όπως είναι για παράδειγμα τα Νευρωνικά Δίκτυα - MLP, LSTMs, τη συλλογή επιπρόσθετων δεδομένων από ετερογενείς πηγές και τη διερεύνηση της Explainable AI (XAI) για την καλύτερη κατανόηση των προβλέψεων των μοντέλων. Συνεπώς, η συνεχής εξέλιξη των μεθόδων Μηχανικής Μάθησης μπορεί να αποτελέσει βασικό μοχλό καινοτομίας στην εκπαίδευση, παρέχοντας βελτιωμένες στρατηγικές μάθησης και υποστήριξης μαθητών.

Παράρτημα

https://github.com/ppatsea/Performance-Predictions_Thesis

Το εν λόγω παράρτημα επιτρέπει την εύκολη αναπαραγωγή των αποτελεσμάτων και παρέχει μια ολοκληρωμένη εικόνα για το περιβάλλον υλοποίησης, τις απαιτήσεις και τις επιδόσεις των μοντέλων. Επομένως, στο συγκεκριμένο σημείο επεξηγείται η λειτουργία του κώδικα που χρησιμοποιήθηκε για την εκπαίδευση και αξιολόγηση μοντέλων πρόβλεψης φοιτητικής απόδοσης στα datasets "student-mat.csv" και "student-por.csv".

1) Φόρτωση και Καθαρισμός Δεδομένων: Ο κώδικας ξεκινάει με την εισαγωγή των απαραίτητων βιβλιοθηκών και την καταστολή των warnings. Τα δεδομένα φορτώνονται από το UCI Repository και προετοιμάζονται:

- ✓ **Αφαίρεση κενών τιμών:** Το dataset φιλτράρεται για να περιλαμβάνει μόνο δείγματα όπου υπάρχει διαθέσιμος τελικός βαθμός (G3).
- ✓ **Αφαίρεση στηλών χωρίς πληροφορία:** Στήλες του συνόλου δεδομένων που εμπεριέχουν αποκλειστικά κενές τιμές διαγράφονται.
- ✓ **Εμφάνιση πρώτων γραμμών:** Για να ελεγχθεί η δομή του dataset, κατά την έξοδο εμφανίζονται οι πρώτες εγγραφές.

```
✓ # Φόρτωση δεδομένων
✓ url = "https://raw.githubusercontent.com/ppatsea/Performance-
  Predictions_Thesis/refs/heads/main/dataset/student-mat.csv"
✓ data = pd.read_csv(url, sep=";")
✓
✓ # Εμφάνιση πρώτων γραμμών για να κατανοήσουμε τη δομή του dataset
✓ print(data.head())
✓
✓ # Καθαρισμός δεδομένων: αφαίρεση κενών τιμών
✓ data.dropna(subset=['G3'], inplace=True)  # Αφήνουμε μόνο τις γραμμές με
  τιμές για το 'G3' (τελικός βαθμός)
✓
✓ # Αφαίρεση κενών στηλών
✓ data.dropna(axis=1, how='all', inplace=True)
```

2) Μετατροπή και Διαχωρισμός Δεδομένων: Η στήλη G3 κατηγοριοποιείται σε τρεις κατηγορίες:

- **Χαμηλή επίδοση** ($G3 \leq 8$)
- **Μέτρια επίδοση** ($9 \leq G3 \leq 14$)
- **Υψηλή επίδοση** ($G3 \geq 15$)

Οι κατηγορικές μεταβλητές μετατρέπονται σε αριθμητικές μέσω One-Hot Encoding, ενώ τα δεδομένα διαχωρίζονται σε σύνολα εκπαίδευσης και δοκιμής με Stratified Train-Test Split.

```
# Χωρισμός χαρακτηριστικών (X) και στόχου (y)
X = data.drop(columns=['G3'])  # Αφαιρούμε τη στήλη 'G3' γιατί είναι ο στόχος

# Κατηγοριοποίηση της μεταβλητής στόχου (G3) σε 3 επίπεδα απόδοσης
def categorize_g3(score):
    if score <= 8:
        return 0  # Χαμηλή επίδοση
    elif 9 <= score <= 14:
        return 1  # Μέτρια επίδοση
    else:
        return 2  # Υψηλή επίδοση

y = data['G3'].apply(categorize_g3)

# Μετατροπή κατηγορικών μεταβλητών σε αριθμητικές
X = pd.get_dummies(X, drop_first=True)

# Stratified Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42, stratify=y)
```

3) Εκπαίδευση και Αξιολόγηση Μοντέλων: Στον κώδικα υλοποιούνται 12 αλγόριθμοι Μηχανικής Μάθησης και αξιολογεί τις επιδόσεις τους με Cross-Validation (5-fold). Κάθε μοντέλο εκπαιδεύεται και αξιολογείται με βάση την ακρίβεια.

```
# Δημιουργία λίστας μοντέλων προς αξιολόγηση
models = {
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42),
    'Decision Tree': DecisionTreeClassifier(random_state=42),
```

```
    'Logistic Regression': LogisticRegression(max_iter=1000, random_state=42),
    'K-Nearest Neighbors': KNeighborsClassifier(),
    'Support Vector Classifier': SVC(probability=True, random_state=42),
    'Naive Bayes': GaussianNB(),
    'MLP': MLPClassifier(max_iter=1000, random_state=42),
    'Gradient Boosting': GradientBoostingClassifier(random_state=42),
    'AdaBoost': AdaBoostClassifier(random_state=42),
                    'XGBoost':         xgb.XGBClassifier(use_label_encoder=False,
eval_metric='mlogloss', random_state=42),
    'Bagging Classifier': BaggingClassifier(random_state=42),
    'Extra Trees': ExtraTreesClassifier(random_state=42),
}

# Ξεκίνημα χρονομέτρησης
start_time = time.time()

# Cross-validation και αξιολόγηση μοντέλων
for model_name, model in models.items():
    cv_scores = cross_val_score(model, X, y, cv=5, scoring='accuracy')
        print(f"Μέση   ακρίβεια   με   Cross-Validation   του   {model_name}   :
{cv_scores.mean() * 100:.2f}%")
```

4) Οπτικοποίηση Αποτελεσμάτων: Η απόδοση των μοντέλων παρουσιάζεται μέσω Confusion Matrices, Bar Charts, Heatmaps και Learning Curves προκειμένου να κατανοηθούν καλύτερα τα αποτελέσματα.

```
# Σχεδίαση Bar Chart με την παλέτα coolwarm και εμφάνιση τιμών πάνω από τις
μπάρες
plt.figure(figsize=(12, 6))
colors = sns.color_palette("coolwarm", len(accuracy_df))  # χρήση της παλέτας
coolwarm
bars = plt.barh(accuracy_df['Model'], accuracy_df['Accuracy'], color=colors)
plt.xlabel("Accuracy Score")
plt.title("Σύγκριση Ακρίβειας Μοντέλων")

# Εμφάνιση τιμών πάνω από τις μπάρες
for bar, accuracy in zip(bars, accuracy_df['Accuracy']):
    plt.text(bar.get_width() - 0.1, bar.get_y() + bar.get_height() / 2,
             f"{accuracy * 100:.1f}%", va='center', ha='center', color='white')

plt.show()
```

5) Χρονικές Μετρήσεις Εκπαίδευσης και Πρόβλεψης: Ο συνολικός χρόνος εκτέλεσης του κάθε μοντέλου υπολογίζεται και εμφανίζεται, μαζί με τον χρόνο εκπαίδευσης και πρόβλεψης.

```
for i, model_name in enumerate(model_group):
        model = models[model_name]

        # Χρονομέτρηση εκπαίδευσης
        start_train_time = time.perf_counter()
        model.fit(X_train, y_train)
        end_train_time = time.perf_counter()

        # Χρονομέτρηση πρόβλεψης
        start_pred_time = time.perf_counter()
        y_pred = model.predict(X_test)
        end_pred_time = time.perf_counter()


# Εκτύπωση χρόνου εκπαίδευσης και πρόβλεψης
        accuracy = accuracy_score(y_test, y_pred)
        print(f"Ακρίβεια: {accuracy * 100:.2f}%")
              print(f"Χρόνος  Εκπαίδευσης  {model_name}:  {end_train_time  -
start_train_time:.4f} δευτερόλεπτα")
               print(f"Χρόνος  Πρόβλεψης  {model_name}:  {end_pred_time  -
start_pred_time:.4f} δευτερόλεπτα")


# Εκτύπωση συνολικού χρόνου εκτέλεσης
end_time = time.time()
execution_time = end_time - start_time
print()
print(f"Συνολικός Χρόνος Εκτέλεσης: {execution_time:.2f} δευτερόλεπτα")
```

Συμπερασματικά, και οι δύο κώδικες επιτρέπουν την ανάλυση δεδομένων φοιτητικής απόδοσης με τη χρήση 12 διαφορετικών αλγορίθμων, την αξιολόγηση των αποτελεσμάτων τους και την οπτικοποίηση της απόδοσής τους. Η χρήση μεθόδων όπως Stratified Train-Test Split, Cross-Validation, και Heatmaps διασφαλίζουν την ακριβή και αξιόπιστη αξιολόγηση των μοντέλων.

Λειτουργία του κώδικα που χρησιμοποιήθηκε για την εκπαίδευση και αξιολόγηση μοντέλων πρόβλεψης φοιτητικής απόδοσης στα datasets "student-mat.csv" και "student-por.csv".

1) Περιγραφή του Περιβάλλοντος Υλοποίησης: Η υλοποίηση πραγματοποιήθηκε σε VSCode, χρησιμοποιώντας τη γλώσσα προγραμματισμού Python σε εικονική μηχανή (VM).

Εργαλεία και Βιβλιοθήκες: Για την ανάπτυξη και αξιολόγηση των μοντέλων μηχανικής μάθησης χρησιμοποιήθηκαν οι ακόλουθες βιβλιοθήκες:

- ✓ **Pandas**: Διαχείριση και προεπεξεργασία δεδομένων.
- ✓ **NumPy**: Υπολογισμοί και διαχείριση πολυδιάστατων πινάκων.
- ✓ **Scikit-Learn**: Εκπαίδευση και αξιολόγηση μοντέλων μηχανικής μάθησης.
- ✓ **Matplotlib & Seaborn**: Οπτικοποίηση δεδομένων και αποτελεσμάτων.
- ✓ **XGBoost**: Gradient boosting αλγόριθμος για βελτιωμένες προβλέψεις.

Τεχνικές Απαιτήσεις:

- ✓ **RAM**: 128 GB
- ✓ **CPU**: AMD EPYC 7552 48-Core Processor, 2196 MHz, 48 Cores, 48 Logical Processor(s)
- ✓ **Λειτουργικό Σύστημα**: Microsoft Windows 10 Pro

2) Φόρτωση και Καθαρισμός Δεδομένων: Ο κώδικας ξεκινάει με την εισαγωγή των απαραίτητων βιβλιοθηκών και την καταστολή των warnings. Τα δεδομένα φορτώνονται από το UCI Repository και προετοιμάζονται:

- ✓ **Αφαίρεση κενών τιμών**: Το dataset φιλτράρεται για να περιλαμβάνει μόνο δείγματα όπου υπάρχει διαθέσιμος τελικός βαθμός (G3).
- ✓ **Αφαίρεση στηλών χωρίς πληροφορία**: Στήλες που περιέχουν αποκλειστικά κενές τιμές διαγράφονται.

- ✓ **Εμφάνιση πρώτων γραμμών**: Για να ελεγχθεί η δομή του dataset, εμφανίζονται οι πρώτες εγγραφές.

3) Εκτελέσιμες Οδηγίες (Reproducibility): Για να εκτελεστεί ο κώδικας από την αρχή, ακολουθούνται τα παρακάτω βήματα:

i) Λήψη των Datasets: Τα datasets "student-mat.csv" και "student-por.csv" είναι διαθέσιμα στο **UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/320/student+performance)**

ii) Εγκατάσταση Απαιτούμενων Βιβλιοθηκών: Για την εκτέλεση του κώδικα απαιτείται η εγκατάσταση των παρακάτω βιβλιοθηκών:

pip install pandas, numpy, scikit-learn, matplotlib, seaborn, xgboost

iii) Εκτέλεση του Κώδικα: Ο κώδικες μπορούν να εκτελεστούν σε **VSCode**, τρέχοντας το αρχείο student-mat.py και student-por.py, με τις εντολές:

python student-mat.py και **python student-por.py**, αντίστοιχα

iv) Συγκριτικός Πίνακας Αποτελεσμάτων: Ο παρακάτω πίνακας συγκρίνει τις επιδόσεις των αλγορίθμων που χρησιμοποιήθηκαν στην ανάλυση.

student-mat.py

Μοντέλο	**Accuracy (%)**	**Precision**	**Recall**	**F1-score**	**Χρόνος Εκπαίδευσης (sec)**	**Χρόνος Πρόβλεψης (sec)**

Random Forest	89.08	0.89	0.89	0.89	0.1385	0.0077
Decision Tree	86.55	0.87	0.87	0.87	0.0035	0.0013
Logistic Regression	88.24	0.89	0.88	0.88	0.2870	0.0014
K-Nearest Neighbors	85.71	0.86	0.86	0.86	0.0019	0.0121
Support Vector Classifier	83.19	0.85	0.83	0.83	0.0168	0.0040
Naïve Bayes	57.98	0.64	0.58	0.58	0.0022	0.0013
MLP	79.83	0.80	0.80	0.80	1.1801	0.0031
Gradient Boosting	89.92	0.90	0.90	0.90	0.3829	0.0025
AdaBoost	84.03	0.85	0.84	0.84	0.0767	0.0095
XGBoost	89.08	0.89	0.89	0.89	0.1327	0.0151
Bagging Classifier	89.92	0.90	0.90	0.90	0.0265	0.0028
Extra Trees	84.03	0.85	0.84	0.84	0.1152	0.0084

student-por.py

Μοντέλο	**Accuracy (%)**	**Precision**	**Recall**	**F1-score**	**Χρόνος Εκπαίδευσης (sec)**	**Χρόνος Πρόβλεψης (sec)**
Random Forest	89.23	0.90	0.89	0.88	0.1789	0.0106

Decision Tree	89.23	0.89	0.89	0.89	0.0048	0.0017
Logistic Regression	87.69	0.88	0.88	0.88	0.4318	0.0017
K-Nearest Neighbors	86.67	0.88	0.87	0.85	0.0026	0.0178
Support Vector Classifier	85.13	0.87	0.85	0.82	0.0386	0.0084
Naïve Bayes	61.54	0.74	0.62	0.63	0.0027	0.0016
MLP	83.59	0.83	0.84	0.83	2.7562	0.0875
Gradient Boosting	91.28	0.91	0.91	0.91	0.8144	0.0054
AdaBoost	86.15	0.77	0.86	0.81	0.0993	0.0120
XGBoost	90.77	0.91	0.91	0.91	1.0972	0.0172
Bagging Classifier	90.77	0.91	0.91	0.91	0.0339	0.0039
Extra Trees	84.62	0.87	0.85	0.82	0.1567	0.0120

Οι κλάσεις έχουν διαφορετικά πλήθη, γι' αυτό οι πίνακες συμπληρώθηκαν με τις τιμές του weighted avg. Αυτό συνέβη γιατί λαμβάνει υπόψη το πλήθος των παραδειγμάτων κάθε κλάσης (support), οπότε αντανακλά καλύτερα τη συνολική απόδοση του μοντέλου.

Βιβλιογραφία

[1] Big Blue Data Academy, «Machine Learning,» 2023. [Ηλεκτρονικό]. Available: https://bigblue.academy/gr/ti-einai-to-machine-learning.

[2] 2'science, «Τι είναι η μηχανική μάθηση (machine learning); – Μέρος Α: "Εισαγωγή",» 2021. [Ηλεκτρονικό]. Available: https://2science.gr/machine-learning-1/.

[3] Βικιπαίδεια, 2025. [Ηλεκτρονικό]. Available: https://el.wikipedia.org/wiki/%CE%9C%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE_%CE%BC%CE%AC%CE%B8%CE%B7%CF%83%CE%B7.

[4] Elements of AI, «Τα είδη της μηχανικής μάθησης,» [Ηλεκτρονικό]. Available: https://course.elementsofai.com/el/4/1.

[5] Κάλλιπος, «Μηχανική Μάθηση,» [Ηλεκτρονικό]. Available: http://repfiles.kallipos.gr/html_books/93/04a-main.html.

[6] ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ, «Μέθοδοι Ταξινόμησης και Παλινδρόμησης για την Πρόβλεψη και την Ανίχνευση Μοτίβων σε Δεδομένα Ακαδημαϊκών Επιδόσεων,» 2015. [Ηλεκτρονικό]. Available: https://ikee.lib.auth.gr/record/281642/files/GRI-2016-15980.pdf.

[7] J. K. Anjali Munde, «Predictive Modelling of Customer Sustainable Jewelry Purchases Using Machine Learning Algorithms,» 2024. [Ηλεκτρονικό]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924007427.

[18] M. S. K. F. N. R. T. R. K. Md. Kabin Hasan Kanchon, «Enhancing personalized learning: AI-driven identification of learning styles and content modification strategies,» 2024. [Ηλεκτρονικό]. Available: https://www.researchgate.net/publication/382087344_Enhancing_personalized_learning_AI-driven_identification_of_learning_styles_and_content_modification_strategies.

[19] S. P. K. K. Y. A. S. Vimarsha K, «Student Performance Prediction: A Co-Evolutionary Hybrid Intelligence model,» 2024. [Ηλεκτρονικό]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924007191.

[20] J. X. I. H. A. S. A. Ghaith Al-Tameemia, «A Hybrid Machine Learning Approach for Predicting Student Performance Using Multi-class Educational Datasets,» 2024. [Ηλεκτρονικό]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924013449.

[21] D. P. M. B. P. S. C. S. Biplov Paneru, «Advancing sustainable mobility: Dynamic predictive modeling of charging cycles in electric vehicles using machine learning techniques and predictive application development,» 2024. [Ηλεκτρονικό]. Available: https://www.sciencedirect.com/science/article/pii/S2772941924000863.

[22] Z. I. D. K. Chaman Verma, «An investigation of novel features for predicting student happiness in hybrid learning platforms – An exploration using experiments on trace data,» 2024. [Ηλεκτρονικό]. Available: https://www.sciencedirect.com/science/article/pii/S2667096824000089.

[23] A.-P. J. B. O. A. J. P. Alfredo Daza, «Stacking ensemble learning model for predict anxiety level in university students using balancing methods,» 2023. [Ηλεκτρονικό]. Available: https://www.sciencedirect.com/science/article/pii/S2352914823001867.

[24] K. I. R. S. S. Rizaldi Ardika Mahendra Pratamaa, «Technique of Mental Health Issues Classification based on Machine Learning: Systematic Literature Review,» 2023.

[14] [Ηλεκτρονικό]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923016770.

[15] T. M. A. M. O. A. Thamer Althubiti, «Developing A Predictive Model for Selecting Academic Track Via GPA by using Classification Algorithms: Saudi Universities as Case Study,» 2023. [Ηλεκτρονικό]. Available: https://thesai.org/Publications/ViewPaper?Volume=14&Issue=5&Code=IJACSA&SerialNo=39.

[16] A. S. M. E. A. A.-E. Nourmeen Lotfy, «An Enhanced Automatic Arabic Essay Scoring System Based on Machine Learning Algorithms,» 2023. [Ηλεκτρονικό]. Available: https://www.techscience.com/cmc/v77n1/54456.

[17] M. Yağcı, «Educational data mining: prediction of students' academic performance using machine learning algorithms,» 2022. [Ηλεκτρονικό]. Available: https://slejournal.springeropen.com/articles/10.1186/s40561-022-00192-z.

[18] R. K. Y. a. K. K. Swati Verma, «Prediction of Academic Performance of Engineering Students by Using Data Mining Techniques,» 2022. [Ηλεκτρονικό]. Available: https://www.ijiet.org/show-180-2305-1.html.

[19] M. A. A. B. R. V. Daud Muhajir, «Improving classification algorithm on education dataset using hyperparameter tuning,» 2022. [Ηλεκτρονικό]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921023954.

[20] I. R. P. S. S. A. M. S. J. J. N. Dhanya R, «F-test feature selection in Stacking ensemble model for breast cancer prediction,» 2020. [Ηλεκτρονικό]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920311467.

[D. K. M. Mayilvaganana, «Cognitive Skill Analysis for Students through Problem Solving
2 Based on Data Mining Techniques,» 2015. [Ηλεκτρονικό]. Available:
1 https://www.sciencedirect.com/science/article/pii/S1877050915004524.
]

[4ο Γυμνάσιο Ζωγράφου, «Ποιοί παράγοντες επηρεάζουν τη σχολική επίδοση των μαθητών;,»
2 2013. [Ηλεκτρονικό]. Available: https://blogs.sch.gr/4gymzogr/?p=395.
2
]

[S. V. Cristóbal Romero, «Educational Data Mining: A Review of the State of the Art,» 2010.
2 [Ηλεκτρονικό]. Available: https://ieeexplore.ieee.org/document/5524021.
3
]

[geeks for geeks, «Linear Regression in Machine learning,» 2025. [Ηλεκτρονικό]. Available:
2 https://www.geeksforgeeks.org/ml-linear-regression/.
4
]

[geeks for geeks, «Getting started with Classification,» 2025. [Ηλεκτρονικό]. Available:
2 https://www.geeksforgeeks.org/getting-started-with-classification/.
5
]

[geeks for geeks, «Ensemble Learning,» 2025. [Ηλεκτρονικό]. Available:
2 https://www.geeksforgeeks.org/a-comprehensive-guide-to-ensemble-learning/.
6
]

[geeks for geeks, «What is a Neural Network?,» 2025. [Ηλεκτρονικό]. Available:
2 https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/.

[27] geeks for geeks, «Clustering in Machine Learning,» 2025. [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/clustering-in-machine-learning/.

[28] geeks for geeks, «Time Series Analysis and Forecasting,» 2024. [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/time-series-analysis-and-forecasting/.

[29] S. O. Oppong, «Predicting Students' Performance Using Machine Learning Algorithms: A Review,» 2023. [Ηλεκτρονικό]. Available: https://www.researchgate.net/publication/372491497_Predicting_Students'_Performance_Using_Machine_Learning_Algorithms_A_Review.

[30] ΒΙΚΙΠΑΙΔΕΙΑ, «Decision Tree,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Decision_tree.

[31] geeks for geeks, «Decision Tree,» 2025. [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/decision-tree/.

[32] IBM, «What is a decision tree?,» [Ηλεκτρονικό]. Available: https://www.ibm.com/think/topics/decision-trees.

[geeks for geeks, «Random Forest Algorithm in Machine Learning,» 2025. [Ηλεκτρονικό].
3 Available: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/.
4
]

[IBM, «What is random forest?,» [Ηλεκτρονικό]. Available:
3 https://www.ibm.com/think/topics/random-forest.
5
]

[Wikipedia, «Logistic regression,» 2025. [Ηλεκτρονικό]. Available:
3 https://en.wikipedia.org/wiki/Logistic_regression.
6
]

[geeks for geeks, «Logistic Regression in Machine Learning,» 2025. [Ηλεκτρονικό]. Available:
3 https://www.geeksforgeeks.org/understanding-logistic-regression/.
7
]

[Medium, «Logistic Regression Algorithm,» 2023. [Ηλεκτρονικό]. Available:
3 https://medium.com/@francescofranco_39234/logistic-regression-algorithm-6451c7928375.
8
]

[Wikipedia, «k-nearest neighbors algorithm,» [Ηλεκτρονικό]. Available:
3 https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.
9
]

[easiio, «Algorithm: The Core of Innovation,» [Ηλεκτρονικό]. Available:
4 https://www.easiio.com/knn-algorithm/.

0
]

[UNITE.AI, «Τι είναι το KNN;,» 2020. [Ηλεκτρονικό]. Available:
4 https://www.unite.ai/el/what-is-k-nearest-neighbors/.
1
]

[tutor.edu.gr, «Support Vector Machines (SVMs),» 2018. [Ηλεκτρονικό]. Available:
4 https://tutor.edu.gr/data-mining/svm.
2
]

[Wikipedia, «Support vector machine,» [Ηλεκτρονικό]. Available:
4 https://en.wikipedia.org/wiki/Support_vector_machine.
3
]

[scikit learn, «Support Vector Machines,» [Ηλεκτρονικό]. Available: https://scikit-
4 learn.org/1.5/modules/svm.html.
4
]

[geeks for geeks, «Support Vector Machine (SVM) Algorithm,» 2025. [Ηλεκτρονικό].
4 Available: https://www.geeksforgeeks.org/support-vector-machine-algorithm/.
5
]

[Wikipedia, «Naive Bayes classifier,» [Ηλεκτρονικό]. Available:
4 https://en.wikipedia.org/wiki/Naive_Bayes_classifier.
6
]

[47] geeks for geeks, «Naive Bayes Classifiers,» 2025. [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/naive-bayes-classifiers/.

[48] datacamp, «Multilayer Perceptrons in Machine Learning: A Comprehensive Guide,» 2024. [Ηλεκτρονικό]. Available: https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning?dc_referrer=https%3A%2F%2Fwww.google.com%2F.

[49] Wikipedia, «Multilayer Perceptron,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Multilayer_perceptron.

[50] analytics vidhya, «Gradient Boosting Algorithm: A Complete Guide for Beginners,» 2024. [Ηλεκτρονικό]. Available: https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/.

[51] geeks for geeks, «Gradient Boosting in ML,» 2023. [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/ml-gradient-boosting/.

[52] Wikipedia, «Gradient Boosting,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Gradient_boosting.

[55] simplilearn, «What is XGBoost? An Introduction to XGBoost Algorithm in Machine Learning,» 2025. [Ηλεκτρονικό]. Available: https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article.

3
]

[Wikipedia, «AdaBoost,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/AdaBoost.
5
4
]

[geeks for geeks, «Implementing the AdaBoost Algorithm From Scratch,» 2025.
5 [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/implementing-the-adaboost-
5 algorithm-from-scratch/.
]

[wikipedia, «XGBoost,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/XGBoost.
5
6
]

[simplilearn, «What is XGBoost? An Introduction to XGBoost Algorithm in Machine
5 Learning,» 2025. [Ηλεκτρονικό]. Available: https://www.simplilearn.com/what-is-xgboost-
7 algorithm-in-machine-learning-article.
]

[geeks for geeks, «XGBoost,» 2025. [Ηλεκτρονικό]. Available:
5 https://www.geeksforgeeks.org/xgboost/.
8
]

[geeks for geeks, «ML | Bagging classifier,» 2025. [Ηλεκτρονικό]. Available:
5 https://www.geeksforgeeks.org/ml-bagging-classifier/.
9
]

[esri, «How Extra trees classification and regression algorithm works,» [Ηλεκτρονικό]. Available: https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-extra-tree-classification-and-regression-works.htm.
]

[geeks for geeks, «ML | Extra Tree Classifier for Feature Selection,» 2023. [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/.
]

[Π. Πατσέα, «Τμήμα Πληροφορικής & Τηλεπικοινωνιών,» 2022. [Ηλεκτρονικό]. Available: https://olympias.lib.uoi.gr/jspui/bitstream/123456789/37979/1/%CE%A0%CE%91%CE%A4%CE%A3%CE%95%CE%91%20%CE%A0%CE%95%CE%A1%CE%A3%CE%95%CE%A6%CE%9F%CE%9D%CE%97%20-%20%CE%A0%CE%9B%CE%97%CE%A1%CE%9F%CE%A6%CE%9F%CE%A1%CE%99%CE%9A%CE%97%CE%A3.pdf.
]

[easiio, «Machine Learning Models,» [Ηλεκτρονικό]. Available: https://www.easiio.com/el/easiio-machine-learning-models/.
]

[el eitca, «Τι είναι ο αλγόριθμος Gradient Boosting;,» 2023. [Ηλεκτρονικό]. Available: https://el.eitca.org/artificial-intelligence/eitc-ai-gcml-google-cloud-machine-learning/advancing-in-machine-learning/automl-vision-part-2/what-is-the-gradient-boosting-algorithm/.
]

[scikit learn, «AdaBoostClassifier,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.AdaBoostClassifier.html.
]

[next brain, «Extra-trees,» 2023. [Ηλεκτρονικό]. Available:
6 https://help.nextbrain.ai/en/article/extra-trees-1ka8ie5/.
6
]