Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Ψηφιακός Πολιτισμός, Έξυπνες Πόλεις, IoT και Προηγμένες Ψηφιακές Τεχνολογίες»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	Νευρωνικά Δίκτυα και Μηχανική Μάθηση, Εφαρμογές στη Βιοϊατρική και Βιοπληροφορική Neural Networks and Machine Learning, Applications in Biomedicine and Bioinformatics
Ονοματεπώνυμο Φοιτητή	Γκουλούση Ζαχαρούλα
Πατρώνυμο	Ευάγγελος
Αριθμός Μητρώου	ΨΠΟΛ/ 2207
Επιβλέπων	Φιλιππάκης Μιχαήλ, Καθηγητής

Ημερομηνία Παράδοσης **Ιούλιος 2025**

Τριμελής Εξεταστική Επιτροπή

Μιχαήλ Φιλιππάκης Καθηγητής	Δημήτριος Βέργαδος Καθηγητής	Άγγελος Μιχάλας Διδάσκων ΠΜΣ

Περίληψη

Στην παρούσα διπλωματική εργασία διερευνάται η πρόβλεψη εμφάνισης καρδιακής προσβολής βάσει κάποιων συμπτωμάτων με την βοήθεια αλγόριθμων Μηχανικής Μάθησης για δυαδική ταξινόμηση, με σκοπό τη σύγκρισή των αποτελεσμάτων των αλγορίθμων αυτών. Η χρήση προγνωστικών μεθόδων με τη βοήθεια της εξόρυξης δεδομένων και ιδιαίτερα της Μηχανικής Μάθησης μπορεί να συμβάλει στην έγκαιρη διάγνωση, ώστε να αποφευχθούν οι σοβαρές επιπλοκές στην υγεία των ασθενών. Η υλοποίηση έγινε με τη βοήθεια της βιβλιοθήκης Scikit-learn της γλώσσας προγραμματισμού Python με ένα συγκεκριμένο σύνολο δεδομένων που διατίθεται δημόσια από το αποθετήριο της Kaggle. Το σύνολο δεδομένων αφορά παρατηρήσεις που συλλέχθηκαν από 918 ασθενείς, το οποίο περιλαμβάνει τα συχνότερα συμπτώματα της ασθένειας. Οι αλγόριθμοι μηχανικής μάθησης που εφαρμόστηκαν στο σύνολο δεδομένων είναι η Λογιστική Παλινδρόμηση (Logistic Regression), οι Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines) και ο Αφελής Μπεϋζιανός (Naive Bayes). Η σύγκριση των αποτελεσμάτων έγινε με διάφορες μετρικές απόδοσης που αφορούν προβλήματα ταξινόμησης. Τα συγκριτικά αποτελέσματα υποδεικνύουν ως καλύτερο αλγόριθμο για το συγκεκριμένο σύνολο δεδομένων τον Αφελή Μπεϋζιανό. Το σχετικό μοντέλο παρουσιάζει τις καλύτερες μετρικές απόδοσης, με το μοντέλο Μηχανών Υποστήριξης Διανυσμάτων να ακολουθεί.

Λέξεις – Κλειδιά

Μηχανική μάθηση, Αλγόριθμοι Ταξινόμησης, Λογιστική Παλινδρόμηση (LR), Μηχανές Υποστήριξης Διανυσμάτων (SVM), Αφελής Μπεϋζιανός (NB), Μετρικές Απόδοσης, Πίνακας Σύγχυσης, Καμπύλη ROC

Abstract

In this thesis, the problem of predicting the occurrence of heart attack based on symptoms using Machine Learning algorithms for binary classification is investigated in order to compare the results of these algorithms. The use of predictive methods with the help of data mining and especially Machine Learning can contribute to early diagnosis to avoid serious health complications in patients. The implementation was done with the help of the Scikit-learn library of the Python programming language with a specific dataset publicly available from the Kaggle repository. The dataset is observations collected from 918 patients, which includes the most common symptoms of the disease. The machine learning algorithms applied to the dataset are Logistic Regression, Support Vector Machines and Naive Bayes. The results were compared with various performance metrics related to classification problems. The comparative results indicate Naïve Bayes as the best algorithm for the given dataset. The associated model shows the best performance metrics, with Support Vector Machines following.

Keywords

Machine Learning, Classification Algorithms, Logistic Regression (LR), Support Vector Machines (SVM), Naïve Bayes (NB), Metrics, Confusion Matrix, ROC Curve

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1. Εισαγωγή

1.1. Εισαγωγή

Η παρούσα διπλωματική εργασία στοχεύει στην ανασκόπηση του πεδίου εφαρμογών της Μηχανικής Μάθησης, όπου θα αναδειχθεί πόσο ικανές είναι αυτές οι μέθοδοι να διαχειρίζονται δεδομένα από διάφορα πεδία της Βιοπληροφορικής αλλά και να κάνουν σωστή πρόβλεψη του αποτελέσματος, δεδομένων κάποιων συνθηκών. Στις μέρες μας, οι αλγόριθμοι τεχνητής νοημοσύνης έχουν ολοένα και μεγαλύτερη εφαρμογή στον τομέα της υγείας. Ταξινομούν τα δεδομένα των ασθενών για να προβλέψουν ποιοι θα αναπτύξουν ιατρικές παθήσεις όπως καρδιακές παθήσεις, ενώ παράλληλα χρησιμοποιούνται βοηθητικά και υποστηρικτικά στην ιατρική γνωμάτευση των ιατρών.

1.2. Αντικείμενο μελέτης

Στην παρούσα εργασία θα μελετήσουμε διάφορους αλγόριθμους μηχανικής μάθησης που έχουν εφαρμογή στον τομέα της Βιοϊατρικής και της Βιοπληροφορικής. Με τη χρήση της γλώσσας προγραμματισμού Python θα πραγματοποιηθεί συγκριτική μελέτη και αξιολόγηση διάφορων αλγορίθμων ταξινόμησης ως προς το σύνολο δεδομένων που έχουμε επιλέξει για το σκοπό αυτό. Το σύνολο δεδομένων αναφέρεται σε ένα σύνολο ατόμων που έχει επιλεγεί και έχουν επιλεχθεί και καταγραφεί σημαντικοί παράγοντες, όπου η μεταβολή των τιμών τους είναι κρίσιμη, καθώς επηρεάζουν την υγεία των ατόμων αυτών. Το σύνολο δεδομένων αποτελείται από 918 εγγραφές και 11 παράγοντες ή χαρακτηριστικά, με την 12η τελευταία στήλη να είναι το τελικό αποτέλεσμα που δείχνει την εμφάνιση καρδιακής προσβολής ή τη φυσιολογική λειτουργία της καρδιάς. Για να υπάρχει μεγαλύτερη ακρίβεια στην ορθότητα των δεδομένων έχει γίνει ο απαραίτητος έλεγχος, για τον καθαρισμό των δεδομένων από μηδενικές και κενές τιμές.

Με την έννοια ταξινόμηση δεδομένων εννοούμε μια τεχνική κατά την οποία παρόμοια σύνολα δεδομένων ταξινομούνται σε ένα ενιαίο σύνολο με βάση τα χαρακτηριστικά τους που τα προσδιορίζουν. Στην ταξινόμηση, ένας ταξινομητής ή ένα μοντέλο κατασκευάζεται για να προβλέψει τα χαρακτηριστικά της ετικέτας κλάσης. Κύριος στόχος της ταξινόμησης είναι η πρόβλεψη της τιμής του αποτελέσματος για κάθε είσοδο δεδομένων. Υπάρχουν πολλές τεχνικές για την επίλυση προβλημάτων ταξινόμησης όπως η Λογιστική Παλινδρόμηση (Logistic Regression), ο Αφελής Μπεϋζιανός (Naive Bayes), ο k - πλησιέστερος γείτονας (K-Nearest Neighbors – KNN), οι Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machine - SVM) και τα δέντρα απόφασης (Decision Trees). Στην παρούσα εργασία θα ασχοληθούμε και θα μελετήσουμε την απόδοση των αλγορίθμων της Λογιστικής Παλινδρόμησης, των Μηχανών Υποστήριξης Διανυσμάτων και του Αφελή Μπεϋζιανού.

2. Βιβλιογραφική επισκόπηση

2.1 Μηχανική μάθηση

Στην εποχή των Μεγάλων Δεδομένων (Big Data), ιδιαίτερη αξία έχει η εξαγωγή χρήσιμης πληροφορίας από όλον αυτό τον όγκο δεδομένων. Η Μηχανική Μάθηση παρέχει την τεχνική βάση για την εξόρυξη των δεδομένων και χρησιμοποιείται για την εξαγωγή πληροφορίας από ακατέργαστα δεδομένα σε βάσεις δεδομένων, μια πληροφορία που μπορεί να εκφραστεί σε κατανοητή μορφή και να χρησιμοποιηθεί για διάφορους σκοπούς. Η Μηχανική Μάθηση είναι ένα υποσύνολο της Τεχνητής Νοημοσύνης (ΤΝ), μιας ευρύτερης έννοιας της επιστήμης των υπολογιστών που αφορά τη δημιουργία έξυπνων μηχανών που μπορούν να προσομοιώσουν τις ανθρώπινες ικανότητες [1].

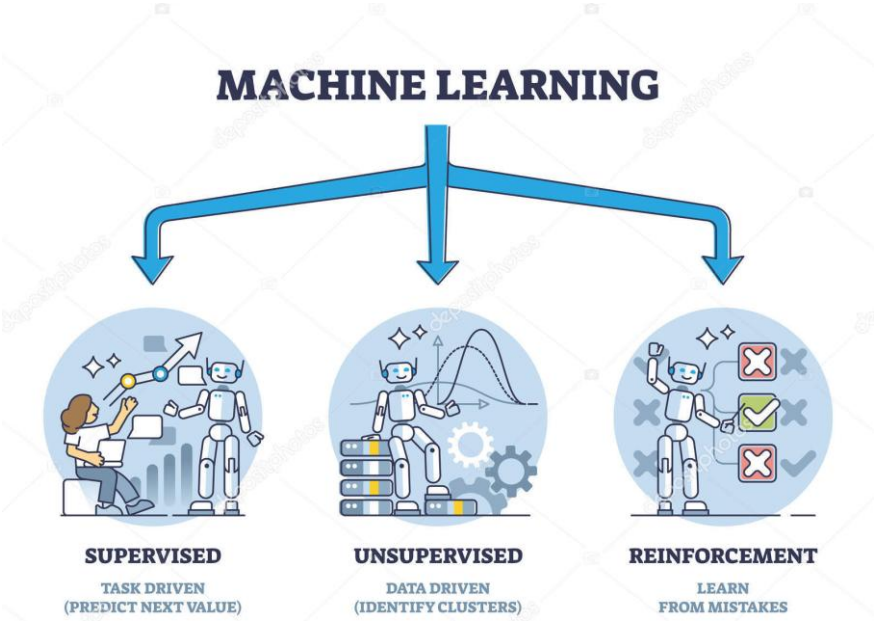
Έχουν αναπτυχθεί διάφοροι ορισμοί για τη Μηχανική Μάθηση. Μηχανική μάθηση είναι η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης (Carbonell, 1987) [26]. Μηχανική μάθηση είναι η διαδικασία κατά την οποία ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία σε σχέση με μια κατηγορία εργασιών και μια μετρική απόδοσης, αν η απόδοσή του σε εργασίες της, βελτιώνεται με την απόκτηση εμπειρίας (Mitchell, 1997) [1].

Εικόνα 2.1 Μηχανική Μάθηση [33]

Πηγή: https://analyzing-testing.netzsch.com/el/blog/2021/apo-ta-exupna-dedomena-sten-tekhnete-noemosune

2.1.1 Κατηγορίες Μηχανικής Μάθησης

Η Μηχανική Μάθηση χωρίζεται σε τρεις υποκατηγορίες μάθησης, όμοιες με τους τρόπους με τους οποίους μαθαίνει και εκπαιδεύεται και ο άνθρωπος, δηλαδή την επιβλεπόμενη μάθηση, την μη επιβλεπόμενη μάθηση και την ενισχυτική μάθηση. Οι αλγόριθμοι αποκτούν με αυτόν τον τρόπο την εμπειρία, δηλαδή εκπαιδεύονται, με βάση το σύνολο εκπαίδευσης που λαμβάνουν (train set). Η κατηγοριοποίηση γίνεται ανάλογα με το εάν η εκπαίδευση γίνεται με ανθρώπινη επίβλεψη – παρέμβαση ή όχι [2].



Εικόνα 2.2 Κατηγορίες Μηχανικής Μάθησης [35]

Πηγή: https://depositphotos.com/gr/vector/types-machine-learning-algorithms-classification-outline-diagram-labeled-educational-scheme-618897540.html

Οι κατηγορίες Μηχανικής Μάθησης είναι τρείς και παρουσιάζονται παρακάτω:

- Επιβλεπόμενη μάθηση (Supervised learning): Στην επιβλεπόμενη μάθηση, η διαδικασία εκμάθησης βασίζεται στην αντιστοίχιση ζευγών εισόδου και εξόδου. Το μοντέλο εκπαιδεύεται από ένα σύνολο δεδομένων εκπαίδευσης (train set) στο οποίο είναι γνωστά τα δεδομένα εισόδου και εξόδου. Στο τέλος της εκπαίδευσης, το μοντέλο είναι σε θέση να κάνει πρόβλεψη της εξόδου σε άγνωστα δεδομένα. Η επιβλεπόμενη μάθηση χρησιμοποιείται για την επίλυση

προβλημάτων ταξινόμησης (classification), πρόγνωσης (prediction) και διερμηνείας (interpretation).

- Μη επιβλεπόμενη μάθηση (Unsupervised learning): Στη μη επιβλεπόμενη μάθηση, το μοντέλο προσπαθεί να κάνει συσχετίσεις μεταξύ των δεδομένων εισόδου και εξόδου χωρίς κάποια καθοδήγηση. Η μη επιβλεπόμενη μάθηση χρησιμοποιείται σε προβλήματα ανάλυσης συσχετίσεων (association analysis) και ομαδοποίησης (clustering).

- Ενισχυτική μάθηση (Reinforcement learning): Η ενισχυτική μάθηση είναι μια εξελιγμένη μορφή μηχανικής μάθησης. Βασίζεται σε διάφορους τομείς όπως για παράδειγμα η θεωρία των παιγνίων αλλά και η συμπεριφορική ψυχολογία. Στην περίπτωση αυτή, το μοντέλο του αλγορίθμου λαμβάνει αποφάσεις με βάση τα δεδομένα εισόδου και ανάλογα με το αν οι αποφάσισες του ήταν σωστές ή όχι, επιβραβεύεται ή τιμωρείται. Η τακτική αυτή εξυπηρετεί ώστε το μοντέλο να μπαίνει στη διαδικασία να τροποποιεί τις αποφάσεις του, με αποτέλεσμα τη λήψη καλύτερων προβλέψεων συνολικά. Η ενισχυτική μάθηση χρησιμοποιείται σε προβλήματα σχεδιασμού (planning) και βελτιστοποίησης.

Η διαδικασία της Μηχανικής Μάθησης περιλαμβάνει αρκετά στάδια τα οποία περιγράφονται αναλυτικά στη συνέχεια [27]:

- Συλλογή δεδομένων: Αρχικά επιλέγουμε το σύνολο δεδομένων, το οποίο θα εξετάσουμε και πάνω στο οποίο θα πειραματιστούμε.

- Προ επεξεργασία δεδομένων (preprocessing data): Τα αρχικά μας δεδομένα μπορεί να είναι σε οποιαδήποτε μορφή. Για να μπορούμε να τα αξιοποιήσουμε και να τα επεξεργαστούμε, χρειάζεται να τα μετατρέψουμε σε κατάλληλη μορφή αναγνωρίσιμη από τους αλγορίθμους μηχανικής μάθησης. Γίνεται διαχωρισμός των χαρακτηριστικών που μας ενδιαφέρουν, κανονικοποίηση των δεδομένων και απαλοιφή κενών τιμών.

- Επιλογή αλγορίθμου: Όταν ολοκληρώσουμε με το αρχικό στάδιο της προ επεξεργασίας των δεδομένων, επιλέγουμε τον αλγόριθμο που θα χρησιμοποιήσουμε στην εκπαίδευση του μοντέλου μας. Η επιλογή του αλγορίθμου γίνεται με κριτήριο τι θέλουμε να κάνουμε με το σύνολο δεδομένων, τις απαιτήσεις σε χρόνο και την υπολογιστική ισχύ που μπορούμε να διαθέσουμε. Υπάρχει δυνατότητα να επιλέξουμε παραπάνω από ένα αλγόριθμο και να πραγματοποιήσουμε συνδυασμό αυτών.

- Εκπαίδευση: Στο στάδιο της εκπαίδευσης, το μοντέλο λαμβάνει το σύνολο δεδομένων που θα ορίσουμε για την διαδικασία της εκπαίδευσης. Πρόκειται για ένα τμήμα του συνόλου δεδομένων μέσα από το οποίο είναι γνωστές οι σχέσεις εισόδου και εξόδου. Στόχος είναι το μοντέλο να κάνει πρόβλεψη, η οποία δίνει την σωστή έξοδο για κάθε δεδομένο εισόδου.

- Έλεγχος: Στο στάδιο του ελέγχου, εισάγουμε ένα νέο για το μοντέλο μας σύνολο δεδομένων. Μπορούμε να αξιοποιήσουμε ένα υποσύνολο του αρχικού

μας συνόλου δεδομένων (test set), το οποίο είναι άγνωστο για το μοντέλο. Το μοντέλο καλείται να εκτελέσει τη διαδικασία μηχανικής μάθησης και να δώσει αποτελέσματα με βάση τα δεδομένα εισόδου που λαμβάνει. Στο τέλος συγκρίνουμε τα αποτελέσματα που έδωσε το μοντέλο με τα πραγματικά αποτελέσματα και αξιολογούμε την απόδοση του μοντέλου.

- Επικύρωση: Στο στάδιο της επικύρωσης, γίνεται μια σύνοψη και ανάλυση των αποτελεσμάτων. Έχουμε τη δυνατότητα να προβούμε σε παραμετροποιήσεις, να αλλάξουμε το ποσοστό διάτμησης των δεδομένων εκπαίδευσης και ελέγχου, ακόμη και να χρησιμοποιήσουμε διαφορετικά δεδομένα ή διαφορετικούς αλγορίθμους.

Στην εργασία αυτή θα εξετάσουμε την κατηγορία της επιβλεπόμενης μάθησης (supervised learning) και πώς αυτή εφαρμόζεται σε προβλήματα ταξινόμησης. Στην επιβλεπόμενη μάθηση παρέχεται σε ένα σύστημα Μηχανικής Μάθησης ένα σύνολο ακατέργαστων δεδομένων, το οποίο αποτελεί την τροφή για να εκπαιδευτεί ο αλγόριθμος [28]. Το σύνολο των δεδομένων περιέχει τα στοιχεία με τα χαρακτηριστικά τους και κάθε στοιχείο του συνόλου συνδέεται με ένα στόχο (target). Τέτοιου είδους δεδομένα, ονομάζονται δεδομένα με ετικέτα (labeled). Τα χαρακτηριστικά μπορούν να πάρουν αριθμητικές (numeric) ή μη αριθμητικές τιμές (nominal ή categorical). Όταν το χαρακτηριστικό εξόδου έχει αριθμητική τιμή τότε αναφερόμαστε σε παλινδρόμηση. Σε διαφορετική περίπτωση, που το χαρακτηριστικό έχει μη αριθμητική τιμή τότε αναφερόμαστε σε ταξινόμηση και οι διακριτές τιμές του χαρακτηριστικού ονομάζονται κλάσεις. Για παράδειγμα, σε μια έρευνα που αφορά την πρόβλεψη μιας ασθένειας ο στόχος είναι η απάντηση εάν ένας άνθρωπος έχει μια ασθένεια ή όχι. Για μια τέτοια έρευνα, οι ειδικοί συλλέγουν περιπτώσεις ασθενών, όπου ένας ασθενής έχει ή δεν έχει την ασθένεια με την καταγραφή μιας σειράς χαρακτηριστικών (features) που θα βοηθήσουν στην πρόβλεψη της εμφάνισής της. Για κάθε ασθενή δίνεται μια ετικέτα (label) όπου δηλώνεται εάν έχει την ασθένεια ή όχι. Για αυτά τα δεδομένα, μπορεί να γίνει η κατάλληλη μορφοποίηση, έτσι ώστε να είναι σε κατάλληλη μορφή κατανοητή και επεξεργάσιμη για τον υπολογιστή. Το σύνολο των παραδειγμάτων σε αυτή τη μορφή αποτελούν ένα σύνολο δεδομένων εκπαίδευσης (train set) μέσω του οποίου θα αποκτηθεί η εμπειρία και θα εκπαιδευτεί ο αλγόριθμος. Ο αλγόριθμος Μηχανικής Μάθησης προσδιορίζει πώς θα βγει το συμπέρασμα εάν ο ασθενής έχει την ασθένεια ή όχι με τη γενίκευση από τα δεδομένα που λαμβάνει. Γίνονται συνεχόμενες δοκιμές και επαναλήψεις όπου ο αλγόριθμος Μηχανικής Μάθησης καλείται να βρει μια όσο το δυνατόν απλούστερη μαθηματική συνάρτηση με βάση τα χαρακτηριστικά που του δίνονται από τα παραδείγματα, για να εξάγει το ορθό αποτέλεσμα για τον στόχο. Το απαιτούμενο επίπεδο ορθότητας ή ακρίβειας (accuracy) του αλγόριθμου εξαρτάται από την εκάστοτε εργασία και είναι το μέτρο απόδοσης του αλγόριθμου. Βέβαια, δεν πρέπει να παραλειφθεί το γεγονός ότι το ζητούμενο που κυρίως μας ενδιαφέρει είναι το πόσο καλά αποδίδει ο αλγόριθμος Μηχανικής Μάθησης σε δεδομένα που δεν γνωρίζει. Με αυτό τον τρόπο προσδιορίζεται πόσο αποτελεσματικά θα χρησιμοποιηθεί σε εφαρμογές με πραγματικά δεδομένα. Έτσι, η απόδοση του αλγόριθμου αποτιμάται με την χρήση ενός συνόλου δοκιμής (test set), το οποίο διαχωρίζεται από τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση του συστήματος Μηχανικής Μάθησης. Η

δυνατότητα της καλής απόδοσης ενός αλγόριθμου σε νέα δεδομένα, τα οποία το σύστημα δεν γνωρίζει ονομάζεται γενίκευση (generalization) [3].

Η επιβλεπόμενη μάθηση χωρίζεται σε δύο κατηγορίες την ταξινόμηση ή κατηγοριοποίηση (classification) και την παλινδρόμηση (regression). Στην πρώτη κατηγορία, αναμένεται η πρόβλεψη για την ταξινόμηση των δεδομένων σε διάφορες εκ των προτέρων γνωστές κατηγορίες και μάλιστα στην περίπτωση όπου οι κατηγορίες ταξινόμησης είναι δύο, έχουμε τη δυαδική ταξινόμηση (binary classification). Ενώ στην περίπτωση της παλινδρόμησης αναμένεται η πρόβλεψη μόνο μιας τιμής.

Οι αλγόριθμοι δυαδικής ταξινόμησης είναι [12]:

- Η Λογιστική Παλινδρόμηση (Logistic Regression)
- Οι Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines – SVM)
- Τα Δέντρα Απόφασης (Decision Trees) και τα τυχαία δάση (Random Forests)
- Ο απλοϊκός Μπεϋζιανός (Naïve Bayes)
- Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANN)
- Οι K-πλησιέστεροι γείτονες (K-Nearest Neighbors - KNN)

Το σύστημα Μηχανικής Μάθησης εκπαιδεύεται από τα δεδομένα που λαμβάνει και δημιουργεί μοντέλα πρόβλεψης με σκοπό όταν θα λάβει νέα δεδομένα που δεν έχει ξαναδεί, να προβλέψει το αποτέλεσμα για τα δεδομένα αυτά. Η ακρίβεια της προβλεπόμενης εξόδου εξαρτάται από την ποσότητα και την ποιότητα των δεδομένων, καθώς ο μεγάλος όγκος δεδομένων βοηθά στη δημιουργία ενός καλύτερου μοντέλου με μεγαλύτερη ακρίβεια πρόβλεψης.

2.1.2 Αλγόριθμοι Μηχανικής Μάθησης

2.1.2.1 Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση (Logistic Regression) είναι μια χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών [6]. Χρησιμοποιείται με σκοπό την εκχώρηση δεδομένων σε μία πραγματική μεταβλητή πρόβλεψης [7], όπως ισχύει και στην περίπτωση της κατηγοριοποίησης όταν είναι διακριτή, διαφορετικά καλείται παλινδρόμηση αν η μεταβλητή είναι συνεχής [8]. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και έπειτα καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Αποτέλεσμα της παλινδρόμησης όταν

χρησιμοποιείται ως τεχνική εξόρυξης δεδομένων, αποτελεί ένα μοντέλο που χρησιμοποιείται μετέπειτα για να προβλέψει τις τιμές της κατηγορίας για τα νέα δεδομένα.

Το Γενικευμένο Νευρωνικό Δίκτυο Παλινδρόμησης (Generalized Regression Neural Network - GRNN) μπορεί να εκτελέσει διεργασίες παλινδρόμησης για να κατασκευάσει ένα μοντέλο παλινδρόμησης. Τα μοντέλα παλινδρόμησης περιλαμβάνουν τις ακόλουθες μεταβλητές:

Α) Οι άγνωστες παράμετροι συσχέτισης που δηλώνονται ως (διάνυσμα).

Β) Οι ανεξάρτητες μεταβλητές (διάνυσμα).

Γ) Η εξαρτώμενη μεταβλητή Y.E (Y | X) = f (X , β)

Η Ανάλυση Παλινδρόμησης βοηθά να κατανοήσουμε την μεταβολή της εξαρτώμενης μεταβλητής Y όταν μεταβάλλεται μία από τις ανεξάρτητες μεταβλητές X, ενώ οι άλλες ανεξάρτητες μεταβλητές παραμένουν σταθερές. Συνήθως, επιδιώκεται να εξακριβωθεί η αιτιώδης επίδραση μιας μεταβλητής επάνω σε μια άλλη. Για τέτοια ζητήματα, συγκεντρώνονται τα δεδομένα που αφορούν τις μεταβλητές ενδιαφέροντος και υιοθετείται η παλινδρόμηση για να υπολογίσει την ποσοτική επίδραση των μεταβλητών επάνω στη μεταβλητή που επηρεάζουν. Πρόκειται δηλαδή για ένα στατιστικό μοντέλο που δίνει μία πιθανότητα μεταξύ του 0 και του 1, παίρνοντας ένα μοντέλο και χρησιμοποιώντας μία λογιστική συνάρτηση, μοντελοποιεί μία δυαδική μεταβλητή. Αξιολογείται επίσης η "στατιστική σημασία" των κατ' εκτίμηση συσχετίσεων, δηλαδή ο βαθμός εμπιστοσύνης (confidence) ότι η αληθινή συσχέτιση είναι κοντά στην εκτίμηση που πραγματοποιήθηκε. Η ανάλυση παλινδρόμησης για πρόβλεψη και πρόγνωση έχει ουσιαστική επικάλυψη με τον τομέα της Μηχανικής Μάθησης.

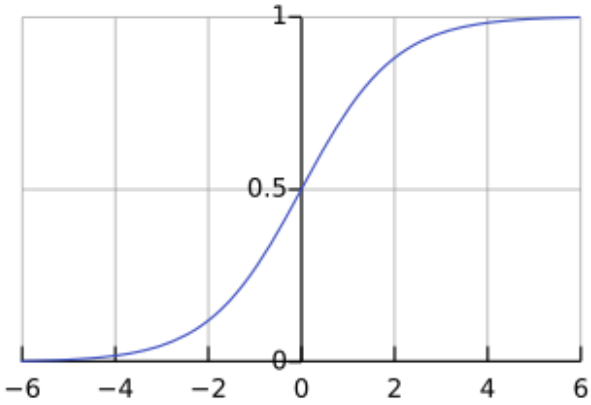
Στη Λογιστική Παλινδρόμηση το λογιστικό μοντέλο είναι ένα μη γραμμικό μοντέλο, τα σφάλματα, του οποίου δεν υπακούν στην κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή. Η Λογιστική Παλινδρόμηση χρησιμοποιείται όταν επιθυμούμε να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού, ή ενός συμβάντος. Είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης για την περίπτωση όπου η εξαρτημένη μεταβλητή Y είναι δίτιμη, δηλαδή παίρνει την τιμή 0 όταν το χαρακτηριστικό δεν υπάρχει ή την τιμή 1 σε διαφορετική περίπτωση.

Η λογιστική συνάρτηση είναι η σιγμοειδής συνάρτηση, μια μαθηματική συνάρτηση της οποίας η καμπύλη έχει τη μορφή S. Δέχεται ως παράμετρο μια λογιστική πραγματική τιμή z και εξάγει ως αποτέλεσμα μία πραγματική τιμή που ανήκει στο διάστημα [0,1] και υπολογίζεται από τον τύπο:

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

Στο μοντέλο της Λογιστικής Παλινδρόμησης, υπολογίζεται το σταθμισμένο άθροισμα: $z = w0x0 + w1x1 + w2x2 + \cdots + wnxn$ των χαρακτηριστικών της εισόδου, και στο z εφαρμόζεται η σιγμοειδής συνάρτηση, η οποία επιστρέφει τιμές στο εύρος [0,1], οπότε η πιθανότητα $p = \sigma(z)$, όπου z είναι το σταθμισμένο άθροισμα κάθε εισόδου [3]. Επομένως, για ένα στιγμιότυπο του συνόλου εκπαίδευσης x η πρόβλεψη της εξόδου $\hat{y}$ υπολογίζεται:

$$\hat{y} = \begin{cases} 0, \sigma(z) < 0.5 \\ 1, \sigma(z) \geq 0.5 \end{cases}$$



Εικόνα 2.3 Σιγμοειδής καμπύλη [39]

Πηγή:https://el.wikipedia.org/wiki/%CE%A3%CE%B9%CE%B3%CE%BC%CE%BF%CE%B5%CE%B9%CE%B4%CE%AE%CF%82_%CF%83%CF%85%CE%BD%CE%AC%CF%81%CF%84%CE%B7%CF%83%CE%B7

Με άλλα λόγια πρόκειται για ένα εποπτευόμενο αλγόριθμο Μηχανικής Μάθησης που χρησιμοποιείται για εργασίες ταξινόμησης, αναλύοντας τη σχέση μεταξύ δύο παραγόντων και προβλέπει την πιθανότητα ένα στιγμιότυπο να ανήκει σε μια κλάση ή όχι. Χρησιμοποιείται για δυαδική ταξινόμηση, η οποία λαμβάνει στην είσοδο ανεξάρτητες μεταβλητές και παράγει μια τιμή πιθανότητας μεταξύ 0 και 1 [11]. Για παράδειγμα, έχουμε δύο κλάσεις 0 και 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.

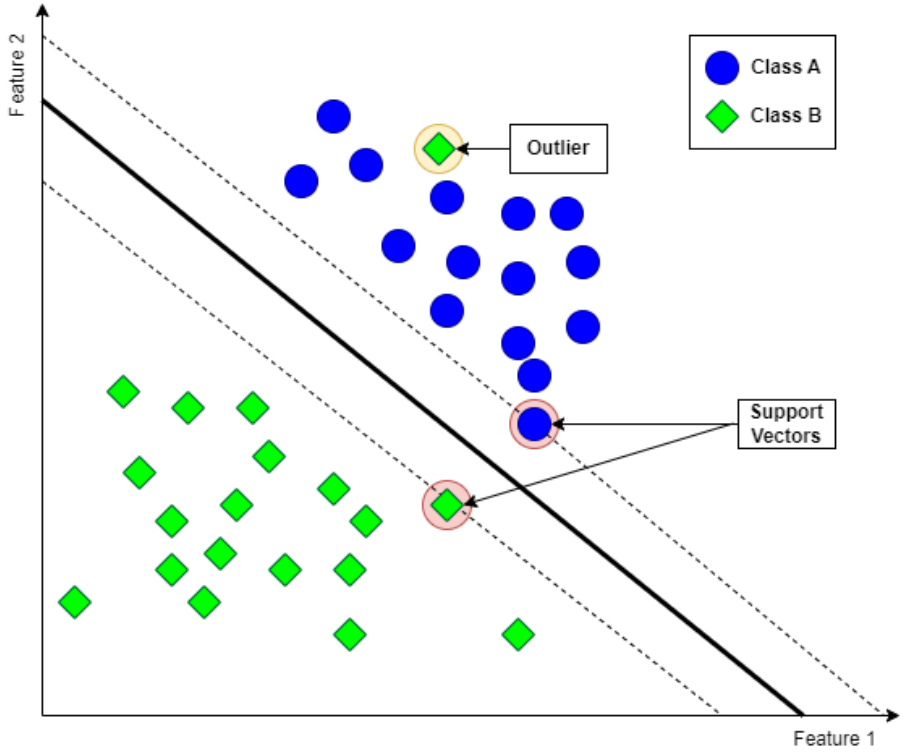
Το μοντέλο Λογιστικής Παλινδρόμησης μετατρέπει την έξοδο συνεχούς τιμής της συνάρτησης γραμμικής παλινδρόμησης σε έξοδο κατηγορικής τιμής χρησιμοποιώντας μια σιγμοειδή συνάρτηση, η οποία αντιστοιχίζει οποιοδήποτε σύνολο ανεξάρτητων μεταβλητών με πραγματική αξία που εισάγονται σε μια τιμή μεταξύ του 0 και του 1. Αυτή η συνάρτηση είναι γνωστή ως λογιστική συνάρτηση. Είναι σημαντικό

στο τέλος να μπορούμε να αξιολογήσουμε την απόδοση του μοντέλου ώστε να έχουμε σαφή εικόνα για την ακρίβεια για την επιτυχή ταξινόμηση των δεδομένων μας. Η αξιολόγηση του μοντέλου της λογιστικής παλινδρόμησης γίνεται με τη βοήθεια των ακόλουθων μετρικών απόδοσης [11], της ορθότητας (accuracy), της ακρίβειας (precision), της ευαισθησίας (sensitivity) και του F1-score.

Τα πλεονεκτήματα της Λογιστικής Παλινδρόμησης είναι η δυνατότητα ταξινόμησης σε περισσότερες κλάσεις, η καλή απόδοση σε μικρά σύνολα δεδομένων και σε δεδομένα που είναι γραμμικά διαχωρίσιμα και το γεγονός ότι είναι εύκολα υλοποιήσιμη και γρήγορη στην διαδικασία της εκπαίδευσης. Τα μειονεκτήματα της Λογιστικής Παλινδρόμησης είναι η ύπαρξη γραμμικότητας μεταξύ των χαρακτηριστικών εισόδου και εξόδου και το γεγονός ότι σε δύσκολα προβλήματα με πολλά χαρακτηριστικά δεν έχει καλή απόδοση.

2.1.2.2 Μηχανές Υποστήριξης Διανυσμάτων

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM) είναι μια αρκετά δημοφιλής τεχνική ταξινόμησης μηχανικής μάθησης. Είναι μια μέθοδος που δημιουργεί ένα ή περισσότερα υπερεπίπεδα σε έναν πολυδιάστατο χώρο, ώστε να ξεχωρίζει διαφορετικές κατηγορίες δεδομένων. Ο πολυδιάστατος αυτός χώρος θα αποτελεί την επιφάνεια απόφασης, με τέτοιο τρόπο ώστε να μεγιστοποιείται η απόσταση που διαχωρίζει τα πλησιέστερα δείγματα που ανήκουν σε διαφορετικές κλάσεις και έχουν ως στόχο την σχεδίαση ενός υπολογιστικά αποδοτικού τρόπου για την εκμάθηση κατάλληλων διαχωριστικών επιπέδων. Η επιλογή του υπέρ-επιπέδου γίνεται με τέτοιο τρόπο ώστε να απέχει όσο το δυνατόν περισσότερο από τα κοντινότερα θετικά και αρνητικά δείγματα (υπέρ-επίπεδο μεγίστου περιθωρίου) [14]. Οι Μηχανές Υποστήριξης Διανυσμάτων αν και έχουν καλή απόδοση, έχουν ένα σοβαρό μειονέκτημα το ότι είναι μια χρονοβόρα δοκιμή [16]. Διαισθητικά, ένας καλός διαχωρισμός επιτυγχάνεται από το υπέρ-επίπεδο που έχει τη μεγαλύτερη απόσταση από τα πλησιέστερα σημεία δεδομένων εκπαίδευσης οποιασδήποτε κατηγορίας (το λεγόμενο λειτουργικό περιθώριο), καθώς γενικά όσο μεγαλύτερο είναι το περιθώριο τόσο μικρότερο είναι το σφάλμα γενίκευσης του ταξινομητή.



Εικόνα 2.4 Διάνυσμα υποστήριξης [40]

Πηγή: https://towardsdatascience.com/support-vector-machine-with-scikit-learn-a-friendly-introduction-a2969f2ff00d/

Στην εικόνα παρουσιάζεται η γραφική αναπαράσταση της λειτουργίας ενός ταξινομητή μηχανών υποστήριξης διανυσμάτων για δύο κατηγορίες δεδομένων (Κατηγορία Α και Β), όπου στόχος είναι η εύρεση του υπερ-επιπέδου (μαύρη γραμμή) που διαχωρίζει τις κατηγορίες με το μεγαλύτερο δυνατό περιθώριο. Οι διακεκομμένες γραμμές δείχνουν τα όρια του περιθωρίου, ενώ τα σημεία που εφάπτονται σε αυτά ονομάζονται διανύσματα υποστήριξης και είναι κρίσιμα για τον προσδιορισμό της γραμμής απόφασης. Το outlier είναι ένα σημείο που βρίσκεται στη «λάθος» πλευρά, δηλαδή μέσα στην περιοχή της άλλης κατηγορίας και μπορεί να επηρεάσει αρνητικά την ακρίβεια του μοντέλου. Η εικόνα αποτυπώνει ξεκάθαρα τη βασική αρχή των Μηχανών Υποστήριξης Διανυσμάτων, το διαχωρισμό με το μέγιστο περιθώριο και βασισμένο μόνο στα πιο κοντινά σημεία των κατηγοριών.

Σε περιπτώσεις όπου τα δεδομένα δεν είναι απολύτως διαχωρίσιμα, ο ταξινομητής επιτρέπει σε μερικά δείγματα να βρίσκονται εντός ή εκτός του ορίου περιθωρίου, εισάγοντας μεταβλητές σφάλματος που τιμωρούνται μέσω μιας παραμέτρου ποινής C, η οποία ελέγχει την ισορροπία μεταξύ ακρίβειας και

ομαλοποίησης. Για να αντιμετωπιστεί η μη γραμμική διαχωρισιμότητα, τα δεδομένα μετασχηματίζονται σε έναν υψηλότερης διάστασης χώρο μέσω μιας συνάρτησης πυρήνα (kernel), όπως γραμμικό, πολυωνυμικό ή ακτινικής βάσης (RBF), ώστε να γίνουν γραμμικά διαχωρίσιμα, χωρίς όμως να υπολογίζεται ρητά ο μετασχηματισμός [14].

Στα πλεονεκτήματα των Μηχανών Υποστήριξης Διανυσμάτων σημειώνεται ότι έχουν χαμηλό υπολογιστικό κόστος, είναι εύκολα στην υλοποίηση και δεν επηρεάζονται από ακραίες τιμές, σε αντίθεση με την παλινδρόμηση. Ενώ, στα μειονεκτήματα των Μηχανών Υποστήριξης Διανυσμάτων σημειώνεται η δύσκολη επιλογή του κατάλληλου πυρήνα για την εφαρμογή μη γραμμικού διανύσματος, όταν το πρόβλημα δεν είναι γραμμικά διαχωρίσιμο.

2.1.2.3 Αφελής Μπεϋζιανός

Η επιβλεπόμενη μάθηση είναι μια υποκατηγορία της Μηχανικής Μάθησης όπου το μοντέλο του αλγορίθμου εκπαιδεύεται σε ένα συγκεκριμένο σύνολο δεδομένων (dataset). Το σύνολο δεδομένων χρησιμοποιείται για την εκπαίδευση των αλγορίθμων ώστε να ταξινομούν τα δεδομένα με όσο το δυνατόν μεγαλύτερη ακρίβεια. Τα δεδομένα εισόδου είναι γνωστά και έχουν αντιστοιχισθεί με την σωστό αποτέλεσμα εξόδου. Κάθε δηλαδή σημείο δεδομένων εισόδου σχετίζεται με την αντίστοιχη ετικέτα (label) εξόδου.

Στην επιβλεπόμενη μάθηση, το σύστημα προσπαθεί να “μάθει” μια συνάρτηση η οποία ονομάζεται συνάρτηση στόχος και χρησιμοποιείται για την πρόβλεψη της τιμής εξόδου, f(x), δεδομένης μιας τιμής εισόδου x. Το σύνολο παρατηρήσεων D αποτελείται από ζεύγη τιμών εισόδου - εξόδου. Πιο συγκεκριμένα:

$$D = \{xi, yi\}$$

όπου xi η τιμή της εισόδου και yi η τιμή της εξόδου. Στόχος του συστήματος είναι να χρησιμοποιήσει τα ζεύγη τιμών που δέχεται ως είσοδο, ώστε να εκπαιδευτεί και τελικά να κατασκευάσει μια συνάρτηση f τέτοια ώστε δεδομένης μιας νέας τιμής x που δεν ανήκει στο σύνολο D, η f(x) να προβλέπει την αντίστοιχη πραγματική τιμή εξόδου y με τη μέγιστη δυνατή ακρίβεια.

Ο Αφελής Μπεϋζιανός (Naïve Bayes) ανήκει στους αλγόριθμους της επιβλεπόμενης Μηχανικής Μάθησης και βασίζεται στην εφαρμογή του θεωρήματος του Bayes με την «αφελή» υπόθεση της υπό όρους ανεξαρτησίας μεταξύ του διανύσματος των χαρακτηριστικών x και της αντίστοιχης μεταβλητής κλάσης y. Το θεώρημα του Bayes δηλώνει την ακόλουθη σχέση, δεδομένης της μεταβλητής κλάσης ταξινόμησης y και το διάνυσμα με τις τιμές των *n* χαρακτηριστικών $x1$ έως xn:

$$P(y \mid x1, \dots, xn) = P(y)P(x1, \dots, xn \mid y)P(x1, \dots, xn)$$

Χρησιμοποιώντας την απλοϊκή υπόθεση της υπό όρους ανεξαρτησίας η οποία είναι:

$$P(xi|y, x1, \dots, xi\text{-}1, xi\text{+}1, \dots, xn) = P(xi|y)$$

για όλα τα i η παραπάνω σχέση απλοποιείται σε:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)}$$

Εικόνα 2.5 Τύπος αφελούς Μπεϋζιανού

Καθώς η $P(x1, \dots, xn)$ είναι σταθερή με δεδομένη την τιμή εισόδου x, εφαρμόζεται ο ακόλουθος κανόνας ταξινόμησης για την πρόβλεψη y^:

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

Εικόνα 2.6 Κανόνας ταξινόμησης πρόβλεψης

Στη συνέχεια, χρησιμοποιείται η μέγιστη εκ των υστέρων (Maximum A Posteriori MAP) εκτίμηση για την εκτίμηση της $P(y)$ που είναι η σχετική συχνότητα της κλάσης y στο σύνολο εκπαίδευσης και την εκτίμηση της $P(xi \mid y)$. Οι διαφορετικοί Μπεϋζιανοί ταξινομητές διαφέρουν κυρίως με βάση την υπόθεση σε σχέση με την κατανομή της $P(xi \mid y)$, δηλαδή εάν είναι Γκαουσιανή (Gaussian) ή πολυωνυμική. Παρά τις απλοποιημένες παραδοχές τους, οι Αφελείς Μπεϋζιανοί ταξινομητές λειτουργούν αρκετά καλά σε πολλές πραγματικές καταστάσεις, όπως στην ταξινόμηση εγγράφων και το φιλτράρισμα μηνυμάτων spam του ηλεκτρονικού ταχυδρομείου. Απαιτούν μια μικρή ποσότητα δεδομένων εκπαίδευσης για την εκτίμηση των απαραίτητων παραμέτρων [14].

Το θεώρημα Bayes είναι μια μέθοδος για τον υπολογισμό της υπό όρους πιθανότητας. Η παραδοσιακή μέθοδος υπολογισμού της πιθανότητας υπό όρους, δηλαδή η πιθανότητα να συμβεί ένα γεγονός δεδομένης της εμφάνισης ενός διαφορετικού γεγονότος, είναι η χρήση του τύπου πιθανοτήτων υπό όρους, υπολογίζοντας την κοινή πιθανότητα του πρώτου γεγονότος και του δεύτερου γεγονότος να συμβαίνουν ταυτόχρονα και στη συνέχεια διαιρώντας το με την πιθανότητα να συμβεί το δεύτερο γεγονός. Ωστόσο, η υπό όρους πιθανότητα μπορεί επίσης να υπολογιστεί με ελαφρώς διαφορετικό τρόπο χρησιμοποιώντας το θεώρημα Bayes [30]. Κατά τον υπολογισμό της υπό όρους πιθανότητας με το θεώρημα Bayes, γίνονται τα ακόλουθα βήματα:

- Προσδιορισμός της πιθανότητας η συνθήκη Β να είναι αληθής, υποθέτοντας ότι η συνθήκη Α είναι αληθής.
- Προσδιορισμός της πιθανότητας η συνθήκη Α να είναι αληθής.
- Πολλαπλασιασμός των δύο πιθανοτήτων Α και Β.
- Διαίρεση με την πιθανότητα να συμβεί η συνθήκη Β.

Αυτό σημαίνει ότι ο τύπος για το θεώρημα Bayes θα μπορούσε να εκφραστεί ως εξής:

$$P(A|B) = P(B|A)*P(A) / P(B)$$

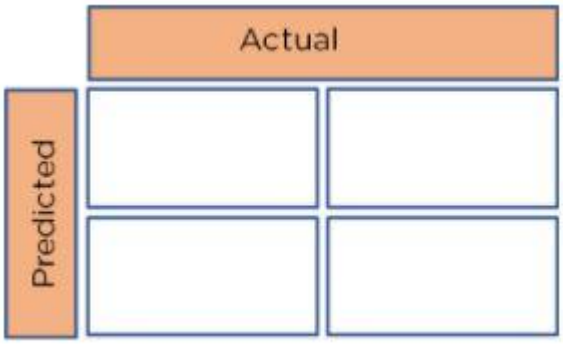
Ο υπολογισμός της υπό όρους πιθανότητας όπως αυτός είναι ιδιαίτερα χρήσιμος όταν η αντίστροφη υπό όρους πιθανότητα μπορεί να υπολογιστεί εύκολα ή όταν ο υπολογισμός της κοινής πιθανότητας θα ήταν πολύ δύσκολος.

Τα πλεονεκτήματα του Αφελή Μπεϋζιανού (Naïve Bayes) είναι η υψηλή ταχύτητα στην εκπαίδευση του μοντέλου, ότι είναι απλός και έχει καλή απόδοση σε σύνθετα προβλήματα και είναι καλή επιλογή όταν έχουμε κατηγορικά δεδομένα. Ενώ, τα μειονεκτήματα του Αφελή Μπεϋζιανού είναι ότι επεξεργάζεται μόνο κατηγορικά δεδομένα, δηλαδή δεδομένα τα οποία μπορεί να είναι αριθμητικά και χρειάζεται να μετατραπούν σε κατηγορικά, υπάρχει το φαινόμενο μηδενικής συχνότητας, στο οποίο η απουσία κατηγορικής τιμής θα δώσει μηδενική τιμή στο στάδιο ελέγχου και τέλος όσο μεγαλύτερη συσχέτιση υπάρχει μεταξύ των χαρακτηριστικών, τόσο χειρότερη απόδοση έχει ο αλγόριθμος.

2.1.3 Μετρικές αξιολόγησης

Ο πίνακας σύγχυσης ή αλλιώς confusion matrix βοηθά στην αξιολόγηση της απόδοσης του μοντέλου ταξινόμησης συγκρίνοντας τις προβλεπόμενες τιμές σε σχέση με τις πραγματικές τιμές για ένα σύνολο δεδομένων. Πρόκειται ουσιαστικά για μια μέθοδο οπτικοποίησης των αποτελεσμάτων ενός αλγορίθμου ταξινόμησης και έχει εφαρμογή σε διάφορους αλγορίθμους ταξινόμησης Μηχανικής Μάθησης όπως τον Αφελή Μπεϋζιανό,, τη Λογιστική Παλινδρόμηση, τα Δέντρα Αποφάσεων και άλλα [20].

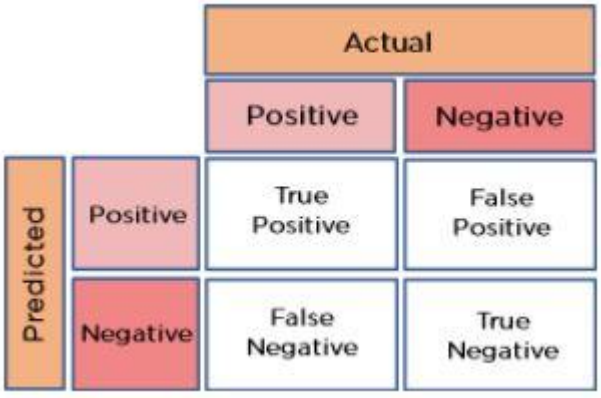
Ο πίνακας σύγχυσης αποτελείται από στήλες που αντιπροσωπεύουν τις προβλεπόμενες τιμές μιας δεδομένης κλάσης και από σειρές που αντιπροσωπεύουν τις πραγματικές τιμές μιας δεδομένης κλάσης ή το αντίστροφο.



Εικόνα 2.7 Βασική δομή πίνακα σύγχυσης [36]

Πηγή: https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning

Ο πίνακας σύγχυσης είναι ιδιαίτερα χρήσιμος, σε σχέση με άλλες μετρικές απόδοσης όπως η ακρίβεια, καθώς δημιουργεί μια πιο ολοκληρωμένη εικόνα του τρόπου απόδοσης ενός μοντέλου. Μπορούμε να λάβουμε τέσσερις διαφορετικούς συνδυασμούς από τις προβλεπόμενες και πραγματικές τιμές ενός ταξινομητή, όπως τα ψευδώς αρνητικά (FN), τα αληθινά αρνητικά (TN), τα ψευδώς θετικά (FP) και τα αληθινά θετικά (TP), όπως αυτά παρουσιάζονται και στην παρακάτω εικόνα:



Εικόνα 2.8 Τιμές πίνακα σύγχυσης [36]

Πηγή:https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning

Η αποτίμηση της ποιότητας ενός μοντέλου για δυαδική ταξινόμηση, δηλαδή ενός δυαδικού ταξινομητή γίνεται με βάση κάποια μέτρα απόδοσης. Γενικά, σε μια εργασία ταξινόμησης, ακολουθούνται οι εξής παραδοχές για την ονοματολογία και τους συμβολισμούς [17]. Οι κλάσεις είναι δύο: η θετική (Positive – P) και η αρνητική (Negative – N). Ο αριθμός των προβλέψεων που ανήκουν στη θετική κλάση, ονομάζονται αληθώς θετικά (True Positive - TP) και αναφέρεται στον αριθμό των σωστών προβλέψεων για την ασθένεια. Ο αριθμός των προβλέψεων που ανήκουν στην αρνητική κλάση, ονομάζονται αληθώς αρνητικά (True Negative - TN) και είναι ο αριθμός των σωστών προβλέψεων για μη ασθένεια. Ο αριθμός προβλέψεων που ταξινομήθηκαν στη θετική κλάση, ενώ στην πραγματικότητα ανήκουν στην αρνητική, ονομάζονται ψευδώς θετικά (False Positive - FP), ενώ ο αριθμός των προβλέψεων που ταξινομήθηκαν στην αρνητική κλάση ενώ στην πραγματικότητα ανήκουν στη θετική, ονομάζονται ψευδώς αρνητικά (False Negative - FN).

Ο πίνακας σύγχυσης αποτελεί ένα εξαιρετικά χρήσιμο εργαλείο για τον υπολογισμό και άλλων μετρήσεων που θα μας βοηθήσουν περαιτέρω στην αξιολόγηση των μοντέλων. Αυτές είναι η ορθότητα, η ακρίβεια, η ευαισθησία, το F1 score όπως παρουσιάζονται αναλυτικά στην συνέχεια [12], [18], [19].

Ορθότητα: Η ορθότητα δείχνει το ποσοστό των σωστά ταξινομημένων περιπτώσεων, δηλαδή κατά πόσο τα δεδομένα και τα συμπεράσματα που προκύπτουν από αυτά είναι αληθή, σωστά και χωρίς σφάλματα.

$$acc = \frac{TP + TN}{TP + FP + FN + TN}$$

Η συμβολή της ορθότητας είναι σημαντική κατά την εισαγωγή των δεδομένων, καθώς τα λάθη που μπορεί να γίνουν οδηγούν σε μείωση της ορθότητας. Τα σφάλματα, ελλιπείς ή ψευδείς τιμές επηρεάζουν την ορθότητα των συμπερασμάτων. Έτσι, αν τα δεδομένα ενός μοντέλου είναι λανθασμένα, τότε και τα αποτελέσματα που θα μας δώσει θα είναι εσφαλμένα. Επίσης, υπάρχει κίνδυνος σωστά στατιστικά αποτελέσματα να ερμηνευτούν λανθασμένα.

Ακρίβεια: Η ακρίβεια εστιάζει στην ακρίβεια των θετικών προβλέψεων, δηλαδή στο πόσο κοντά βρίσκονται οι μετρήσεις ή οι προβλέψεις σε μια πραγματική τιμή.

$$prec = \frac{TP}{TP + FP}$$

Ιδιαίτερη προσοχή χρειάζεται η αξιολόγηση των αποτελεσμάτων της ακρίβειας, καθώς μπορεί να είναι παραπλανητική σε ανισόρροπα σύνολα δεδομένων. Αν για παράδειγμα, το 90%των δεδομένων ανήκουν σε μια κατηγορία και το μοντέλο προβλέπει πάντα αυτή την κατηγορία, η ακρίβεια μπορεί να είναι στο 90% αλλά δεν θα είναι χρήσιμη πληροφορία. Συνεπώς υψηλή ακρίβεια δεν συνεπάγεται πάντα καλά αποτελέσματα.

Ευαισθησία: Η ευαισθησία είναι ένας πολύ σημαντικός δείκτης στην ανάλυση δεδομένων, κυρίως σε προβλήματα ταξινόμησης (classification) και ιδιαίτερα όταν υπάρχουν δύο κατηγορίες. Μετρά το ποσοστό των σωστά προβλεπόμενων θετικών περιπτώσεων μεταξύ όλων των πραγματικών θετικών περιπτώσεων, δηλαδή την ικανότητα ενός μοντέλου να εντοπίζει σωστά τα θετικά δείγματα.

$$sen = \frac{TP}{TP + FN}$$

Η συμβολή της ευαισθησίας είναι σημαντική όταν επιθυμούμε να εντοπίσουμε όλα τα θετικά αποτελέσματα. Για παράδειγμα σε ιατρικές διαγνώσεις, είναι προτιμότερο να εντοπιστεί ένα ψευδώς αρνητικό από το να μην εντοπιστεί ο ασθενής με την ασθένεια.

F1-Score: Το F1-score είναι η αρμονική μέση ακρίβεια και ευαισθησία. Είναι ένας πολύ χρήσιμος δείκτης αξιολόγησης στην ανάλυση δεδομένων, ιδιαίτερα όταν έχουμε ανισόρροπες (μη ισομερείς) κατηγορίες και θέλουμε να ισορροπήσουμε ανάμεσα σε ευαισθησία και ακρίβεια. Το F1-score χρησιμοποιείται όταν τα False Positive και False Negative είναι εξίσου σημαντικά, σε ανισόρροπα σύνολα δεδομένων, όταν η απλή ακρίβεια δεν δίνει πλήρη και ικανοποιητική εικόνα.

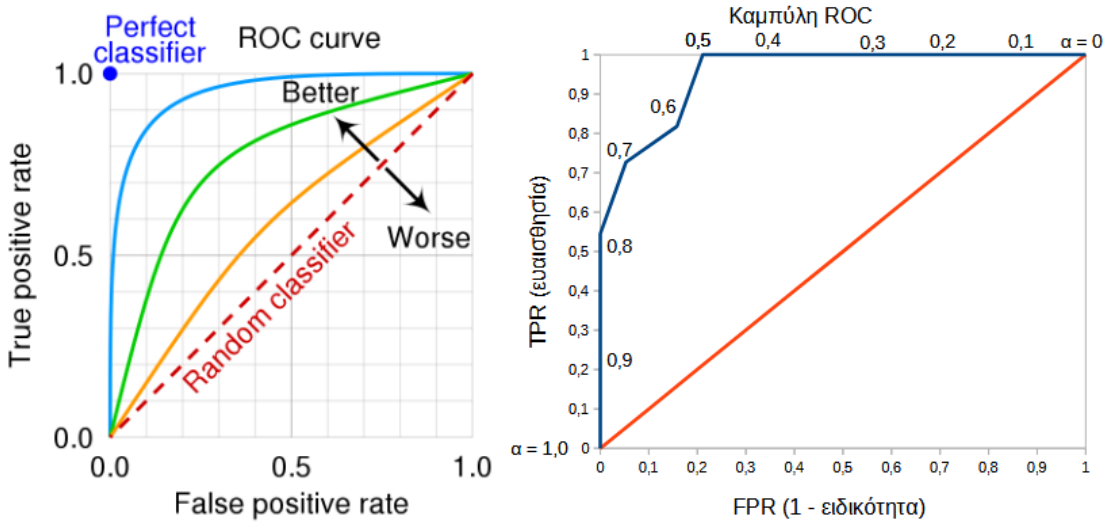
$$FI - score = 2 * \left(\frac{precision * recall}{precision + recall} \right)$$

Η καμπύλη χαρακτηριστικού δέκτη (Receiver Operating Characteristic Curve ή ROC) αποτελεί επίσης μια χρήσιμη τεχνική για την οργάνωση, την επιλογή και την απεικόνιση των ταξινομητών με βάση τη γραφική τους αναπαράσταση. Στην καμπύλη ROC απεικονίζεται ο πραγματικός θετικός ρυθμός έναντι του ψευδώς θετικού σε διάφορα κατώφλια. Το AUC-ROC μετρά την περιοχή κάτω από αυτήν την καμπύλη, παρέχοντας ένα συνολικό μέτρο της απόδοσης ενός μοντέλου σε διαφορετικά κατώφλια ταξινόμησης. Οι πιθανότητες κυμαίνονται μεταξύ 0 και 1, και εκφράζουν την πιθανότητα ενός συμβάντος ως αναλογία τόσο των εμφανίσεων όσο και των μη εμφανίσεων. Το 0 είναι το χαμηλότερο όριο και εκφράζει την μικρότερη πιθανότητα, ενώ το 1 είναι το υψηλότερο όριο και εκφράζει την μεγαλύτερη πιθανότητα [21].

Η τιμή του ορίου, ορίζεται από τον εκάστοτε ερευνητή σε οποιαδήποτε τιμή μεταξύ των τιμών 0 και 1, με τις διάφορες επιλογές να αντιστοιχούν σε διαφορετικούς πίνακες σύγχυσης, με διαφορετικές τιμές ευαισθησίας και ειδικότητας. Το σύνολο των ζευγών (ευαισθησία, 1 – ειδικότητα) είναι δυνατόν να τοποθετηθούν σε ένα διάγραμμα που θα διευκολύνει την απόφαση του ερευνητή. Το διάγραμμα αυτό είναι γνωστό ως καμπύλη λειτουργικού χαρακτηριστικού δέκτη - Receiver Operating Characteristic Curve ή πιο απλά καμπύλη ROC [30].

Η ονομασία καμπύλη ROC χρονολογείται από τον Δεύτερο Παγκόσμιο Πόλεμο και τους ηλεκτρολόγους μηχανικούς που υποστήριζαν τη λειτουργία των ραντάρ στα αεροπλάνα. Όταν η ενίσχυση του σήματος στο ραντάρ ήταν μηδενική, τότε δεν ανιχνευόταν κανένα εχθρικό αεροπλάνο. Όσο αυξανόταν η ενίσχυση του σήματος,

γινόταν ανίχνευση όλο και περισσότερων αντικειμένων, αυξανόταν και ο θόρυβος, με αποτέλεσμα να δημιουργούνται λανθασμένα θετικά αποτελέσματα (FP), δηλαδή να αναγνωρίζονται ως εχθρικά αεροπλάνα αντικείμενα που φαίνονταν σαν αεροπλάνα αλλά στην πραγματικότητα δεν ήταν, όπως σύννεφα βροχής ή κοπάδια πουλιών. Από κάποια ορισμένη τιμή, η περαιτέρω αύξηση της ενίσχυσης ήταν αντιπαραγωγική, καθώς ο θόρυβος (FP) άρχιζε να υπερτερεί των πραγματικών σημάτων (TP). Η καμπύλη ROC αντιπροσωπεύει ένα απλό στη δημιουργία διάγραμμα που στόχο έχει να προσφέρει βοήθεια στον χειριστή του ραντάρ να αποφασίσει ποιο επίπεδο ενίσχυσης προσφέρει ικανοποιητικά ποσοστά ευαισθησίας (True Positive Rate, TPR) και σχετικά μεγάλα ποσοστά ειδικότητας ή ισοδύναμα μικρή πιθανότητα εσφαλμένης ανίχνευσης (False Positive Rate FPR = 1 – ειδικότητα).



Εικόνα 2.9 Καμπύλη λειτουργικού χαρακτηριστικού δέκτη [35]

Πηγή:https://utopia.duth.gr/epdiaman/files/kedivim/section7/ROC%20Curve.pdf

Το εμβαδόν της περιοχής κάτω από την ROC καμπύλη ονομάζεται AUC (Area Under Curve) και είναι ένα βασικό μέτρο αξιολόγησης της μεθόδου ταξινόμησης. Αν η τιμή του AUC < 0,5 τότε η επίδοση του μοντέλου είναι χειρότερη από την τυχαία επιλογή και πρέπει να απορριφθεί εξ' ολοκλήρου. Όσο το AUC πλησιάζει πιο κοντά στην τιμή 1, τόσο καλύτερη είναι η διαδικασία ταξινόμησης στο σύνολό της. Επιπλέον, το AUC επιτρέπει τη σύγκριση δύο ή περισσότερων διαφορετικών μοντέλων που στοχεύουν στην ταξινόμηση των παρατηρήσεων [3].

Τα ζεύγη των τιμών (FPR, TPR) = (1 – ειδικότητα, ευαισθησία), αν αναπαρασταθούν σε διαγραμματική μορφή, συνιστούν την καμπύλη λειτουργικού χαρακτηριστικού δέκτη (καμπύλη ROC). Ο ερευνητής μπορεί να επιλέξει το όριο α με το οποίο επιτυγχάνεται αξιοπρεπής ευαισθησία και ειδικότητα στο μοντέλο πρόγνωσης

που έχει αναπτύξει. Η καμπύλη ROC, συνήθως απεικονίζεται μαζί με την ευθεία y = x, η οποία απευθύνεται στην τυχαία ταξινόμηση, όπου FPR = TPR. Ως εκ τούτου, επιθυμία είναι η επιλογή του ορίου α, να εξασφαλίζει ένα ζεύγος (FPR, TPR) όσο το δυνατόν πιο μακριά από την ευθεία y=x [35].

3. Πειραματική προσέγγιση

Το κεφάλαιο αυτό αφορά τη μεθοδολογία υλοποίησης του έργου. Αρχικά παρουσιάζεται το περιβάλλον υλοποίησης στην γλώσσα προγραμματισμού Python. Στη συνέχεια, παρουσιάζονται τα βασικά βήματα της υλοποίησης.

3.1 Περιγραφή βάσης δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε στην εργασία περιέχει παρατηρήσεις από 918 ασθενείς με 12 χαρακτηριστικά ανά στιγμιότυπο και μία μεταβλητή, που καλείται κλάση ταξινόμησης και που περιγράφει εάν ο ασθενής θα έχει καρδιακό πρόβλημα ή όχι. Η δομή των δεδομένων του συνόλου, καθώς και οι τιμές που παίρνουν τα στιγμιότυπα περιγράφεται αναλυτικά στον Πίνακα 1.1 Στο σύνολο των δεδομένων δεν υπάρχουν ελλιπείς τιμές, επομένως τα στοιχεία αφορούν ολόκληρο το σύνολο δεδομένων με τις 918 εγγραφές.

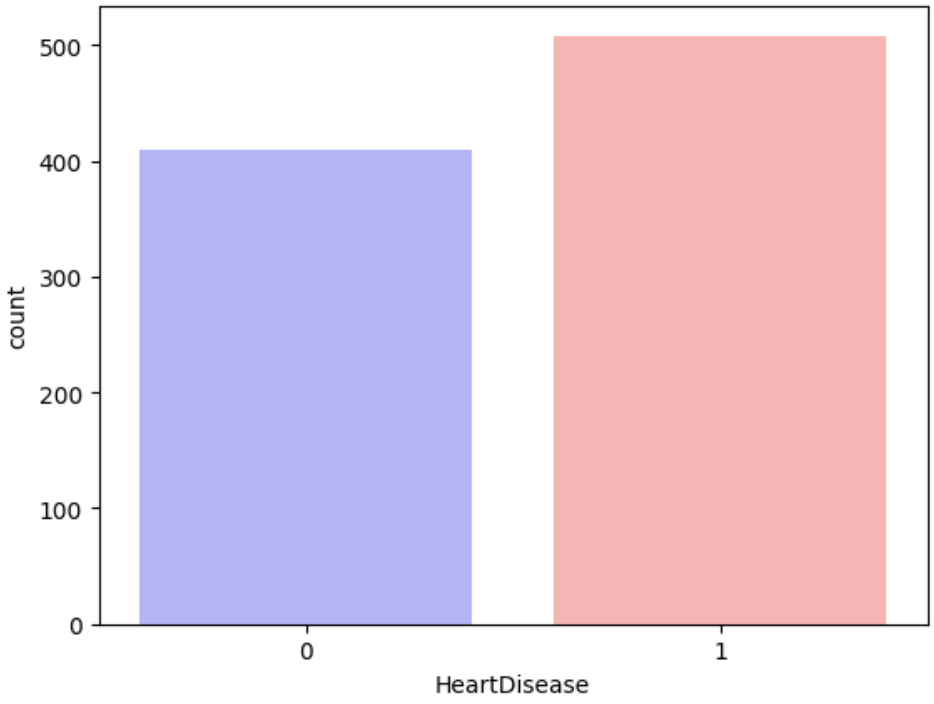
	ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ	**ΤΙΜΕΣ**
1	Age: Ηλικία	έτη
2	Sex: Φύλο	M, F
3	ChestPainType: τύπος πόνου στο θώρακα	[TA: Τυπικός στηθαγχικός πόνος, ATA: Άτυπος πόνος, NAP: Μη στηθαγχικός πόνος, ASY: Ασυμπτωματικός]
4	RestingBP: Αρτηριακή πίεση σε κατάσταση ηρεμίας	< 90: Χαμηλή πίεση, 90 – 120: φυσιολογική, 120 - 129: αυξημένη, 130 - 139: υπέρταση στάδιο 1, >=140: υπέρταση στάδιο 2, >=180: υπερτασική κρίση (επείγον)
5	Cholesterol: Χοληστερόλη [mm/dl]	< 200: φυσιολογικό, 200 - 239: οριακά υψηλό, >=240: υψηλό – αυξημένος καρδιαγγειακός

		κίνδυνος
6	FastingBS: Σάκχαρο στο αίμα μετά από τουλάχιστον 8 ώρες χωρίς φαγητό.	[1: > 120 mg/dl πιθανό πρόβλημα (υπεργλυκαιμία), 0: =< 120 mg/dl φυσιολογικό]
7	RestingECG: Αποτελέσματα ηλεκτροκαρδιογραφήματος σε ηρεμία	[0 - Normal: Φυσιολογικό καρδιογράφημα, 1-ST: Με ανωμαλία του κύματος ST-T (Ανωμαλία στα κύματα ST ή T – μπορεί να υποδηλώνει ισχαιμία ή υπερτροφία), 2: υπερτροφία της αριστερής κοιλίας με τα κριτήρια του Estes)
8	MaxHR: Μέγιστος καρδιακός ρυθμός	[Τιμή μεταξύ 60 and 202]
9	ExerciseAngina: Στηθάγχη που προκαλείται από άσκηση	[1: Ναι (πιθανότητα για ισχαιμία ή στεφανιαία νόσο), 0: Όχι]
10	Oldpeak ST: δείκτης των δοκιμασιών κοπώσεως πριν ή μετά την άσκηση	[0.0: Καμία κατάσπαση ST → φυσιολογικό, >0.0: Παρουσία κατάσπασης (ή ανάσπασης) ST → πιθανή ισχαιμία, >1.0: Αυξημένη πιθανότητα στεφανιαίας νόσου, >2.0: Ισχυρή ένδειξη ισχαιμίας (πολύ παθολογικό)]
11	ST_Slope: κλίση του ST τμήματος στο καρδιογράφημα κατά τη διάρκεια της κορυφαίας άσκησης, είναι σημαντικός δείκτης σε δοκιμασίες κοπώσεως	[Up: ανιούσα (φυσιολογική), Flat: επίπεδη (μπορεί να υποδηλώνει ισχαιμία), Down: κατιούσα (ανησυχητική)]
12	HeartDisease: έξοδος/ αποτέλεσμα	[1: Καρδιακή προσβολή, 0: Φυσιολογική λειτουργία]

Πίνακας 3.1 Περιγραφή Στοιχείων Δομής Δεδομένων

Η δομή των δεδομένων του συνόλου, καθώς και οι τιμές που παίρνουν τα στιγμιότυπα φαίνεται πιο καθαρά στον Πίνακα 1. Μέσα σε παρένθεση αναγράφονται οι ετικέτες των χαρακτηριστικών, όπως αναφέρονται στο αρχικό σύνολο.

Στο σύνολο που διατίθεται στο αποθετήριο, δεν υπάρχουν ελλιπείς τιμές, συνεπώς τα στοιχεία αφορούν 918 εγγραφές. Από αυτές τις εγγραφές, προκύπτει ότι 410 στιγμιότυπα και σε ποσοστό 44,66% ανήκουν στην κλάση Negative (0), ενώ 508 στιγμιότυπα και σε ποσοστό 55,34% ανήκουν στην κλάση Positive (1), όπως φαίνεται στις Εικόνες 3.1 και 3.2 παρακάτω.



Εικόνα 3.1 Διάγραμμα της κλάσης ταξινόμησης

HeartDisease	count
1	508
0	410

Εικόνα 3.2 Αρχικές τιμές της βάσης δεδομένων

3.2 Περιβάλλον Υλοποίησης

Για τη δημιουργία του κώδικα της εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, η οποία είναι η πλέον δημοφιλέστερη γλώσσα προγραμματισμού για προβλήματα ΒΜ και ανάλυσης δεδομένων. Η Python είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου και γενικού σκοπού, που δημιουργήθηκε από τον Guido van Rossum και παρουσιάστηκε το 1991. Η φιλοσοφία του σχεδιασμού της γλώσσας Python δίνει έμφαση στην αναγνωσιμότητα του κώδικα. Οι γλωσσικές δομές και η αντικειμενοστραφής προσέγγιση της στοχεύουν να βοηθήσουν τους προγραμματιστές να γράφουν σαφή και λογικό κώδικα για μικρά και μεγάλα πρότζεκτ. Η γλώσσα Python είναι ισχυρή και γρήγορη, τρέχει παντού, είναι προσιτή, εύχρηστη και ανοιχτού κώδικα. Η υλοποίηση της εργασίας έγινε στο περιβάλλον της ελεύθερης διανομής του Google Colab, το οποίο είναι βασισμένο στην Python 3.10 και είναι μία από τις δημοφιλέστερες πλατφόρμες για την επιστήμη των δεδομένων καθώς περιέχει πολλές βιβλιοθήκες ανοιχτού κώδικα διαφορετικών αλγορίθμων.

3.2.1 Βιβλιοθήκες Python

Οι σημαντικότερες βιβλιοθήκες της γλώσσας προγραμματισμού Python που χρησιμοποιήθηκαν στην εργασία είναι ελεύθερες και ανοιχτού κώδικα, υποστηρίζονται από μεγάλες κοινότητες προγραμματιστών και διαθέτουν πολύ καλή τεκμηρίωση. Οι βιβλιοθήκες που χρησιμοποιήθηκαν και περιλαμβάνονται στη διανομή του Google Colab περιγράφονται συνοπτικά στη συνέχεια.

SciPy: Η βασική βιβλιοθήκη για επιστημονικούς και τεχνικούς υπολογισμούς [13]. Η SciPy περιέχει λειτουργικές μονάδες για βελτιστοποίηση, γραμμική άλγεβρα, ολοκλήρωση, παρεμβολή, ειδικές λειτουργίες και άλλες εργασίες κοινές στην επιστήμη και τη μηχανική. Βασίζεται στο αντικείμενο των πινάκων NumPy, αποτελεί μέρος της στοίβας NumPy και το οικοσύστημά της περιλαμβάνει ένα σύνολο επιστημονικών βιβλιοθηκών. Τα βασικά πακέτα- βιβλιοθήκες του οικοσυστήματος της SciPy είναι τα παρακάτω:

NumPy: Η βασική βιβλιοθήκη για N - διάστατους πίνακες. Η NumPy προσθέτει υποστήριξη για μεγάλους, πολυδιάστατους πίνακες και μητρώα, μαζί με μια μεγάλη συλλογή μαθηματικών συναρτήσεων υψηλού επιπέδου για λειτουργίες γραμμικής άλγεβρας. Δημιουργήθηκε το 2005, με βάση τις πρώτες εργασίες των βιβλιοθηκών Numerical και Numarray.[22]

Pandas: Η βιβλιοθήκη για δομές δεδομένων και ανάλυση. Η βιβλιοθήκη Pandas είναι για χειρισμό και ανάλυση δεδομένων. Συγκεκριμένα, προσφέρει δομές δεδομένων και

λειτουργίες για χειρισμό αριθμητικών πινάκων και χρονοσειρών. Το όνομα προέρχεται από τον όρο "panel data", έναν όρο οικονομετρίας για σύνολα δεδομένων που περιλαμβάνει παρατηρήσεις σε πολλαπλές χρονικές περιόδους για τα ίδια αντικείμενα.[23]

Matplotlib: Η βιβλιοθήκη 2D σχεδίασης και η αριθμητική επέκταση μαθηματικών της NumPy. Παρέχει μια αντικειμενοστραφή διεπαφή προγραμματισμού (Application Programming Interface – API) για την ενσωμάτωση σχεδίων σε εφαρμογές με χρήση εργαλείων γενικής χρήσης γραφικών διεπαφών χρήστη (Graphical User Interface – GUI). Η Pyplot είναι μια ενότητα της matplotlib που παρέχει διεπαφή τύπου MATLAB13. Έχει σχεδιαστεί έτσι ώστε, να μπορεί να χρησιμοποιηθεί όπως το MATLAB [25], με τη δυνατότητα χρήσης της Python και το πλεονέκτημα του ότι είναι δωρεάν και ανοιχτού κώδικα.[24]

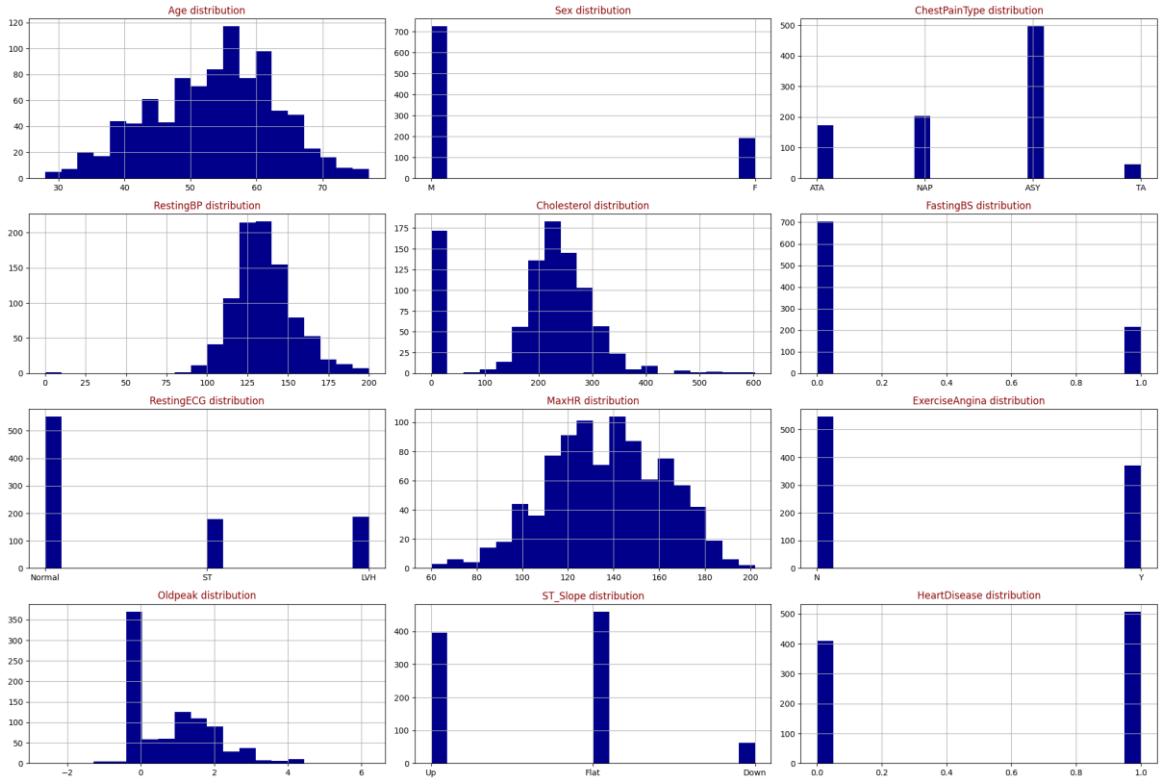
Scikit-learn: Η βιβλιοθήκη για εφαρμογές Μηχανικής Μάθησης. Διαθέτει διάφορους αλγόριθμους για ταξινόμηση, παλινδρόμηση, συσταδοποίηση, προεπεξεργασία δεδομένων (εξαγωγή χαρακτηριστικών και εξομάλυνση). Έχει σχεδιαστεί για να λειτουργεί με τις αριθμητικές και επιστημονικές βιβλιοθήκες NumPy και SciPy.[14]

3.3 Επεξεργασία και ανάλυση δεδομένων

Μετά την εξερεύνηση του περιεχομένου του συνόλου των δεδομένων, ακολουθεί η διερεύνηση των κατανομών διάφορων χαρακτηριστικών και η μεταξύ τους συσχέτιση, όπως περιγράφεται αναλυτικά παρακάτω.

3.3.1 Συσχέτιση χαρακτηριστικών με την κλάση ταξινόμησης

Η συσχέτιση κάθε δυαδικού χαρακτηριστικού από το σύνολο δεδομένων σε σχέση με την κλάση της ταξινόμησης φαίνεται στην Εικόνα 3.3 με την αναπαράσταση ιστογράμματος.



Εικόνα 3.3 Συσχέτιση Χαρακτηριστικών με την Κλάση Ταξινόμησης

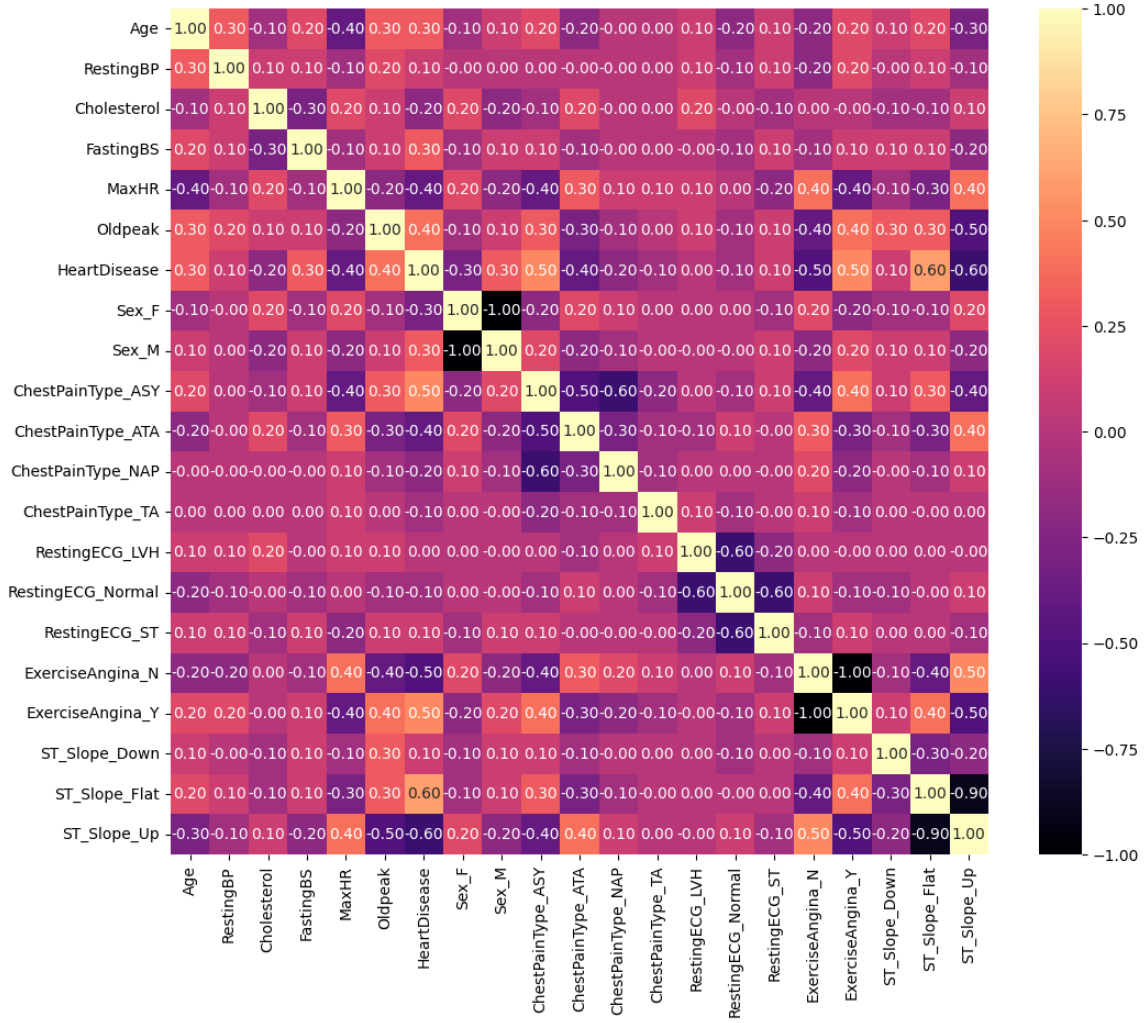
Τα ιστογράμματα στην παραπάνω Εικόνα 3.3 απεικονίζουν την κατανομή των μεταβλητών του συνόλου δεδομένων που σχετίζονται με την υγεία της καρδιάς. Στο πρώτο γράφημα Age distribution παρουσιάζονται οι ηλικίες, οι οποίες συγκεντρώνονται κυρίως στο εύρος μεταξύ 50 – 65 ετών, με μέγιστη τιμή τα 55 έτη. Η κατανομή είναι κανονική με μικρή ασυμμετρία. Στο δεύτερο γράφημα Sex distribution υπάρχει σαφής υπεροχή των ανδρών (M) σε σχέση με τις γυναίκες (F). Στο γράφημα ChestPainType distribution παρουσιάζεται ο πιο συχνός τύπος πόνου ο ASY (ασυμπτωματικός), που υποδεικνύει παρουσία ισχαιμίας. Ακολουθούν οι τύποι NAP, ATA, και λίγοι με TA. Το ιστόγραμμα RestingBP (Αρτηριακή πίεση ηρεμίας) δείχνει ότι πολλοί ασθενείς έχουν αρτηριακή πίεση σε ηρεμία στο εύρος τιμών μεταξύ 120 – 140mmHg. Το ιστόγραμμα Cholesterol distribution απεικονίζει τις τιμές της χοληστερίνης, οι οποίες επικεντρώνονται γύρω στα 200 – 250mg/dL. Βέβαια, υπάρχουν και ακραίες τιμές έως 600mg/dL, αλλά είναι πιο σπάνιες. Στη συνέχεια, το ιστόγραμμα FastingBS (Σάκχαρο νηστείας) έχει δυαδική μεταβλητή, καθώς οι περισσότεροι έχουν τιμή FastingBS = 0 (δηλαδή <120 mg/dL). Λιγότεροι έχουν τιμή FastingBS = 1 που υποδηλώνει υψηλό σάκχαρο νηστείας. Στο γράφημα RestingECG (ηλεκτροκαρδιογράφημα σε ηρεμία) η

πλειοψηφία έχει κανονικό καρδιογράφημα. Οι άλλες κατηγορίες (ST, LVH) εμφανίζονται σε μικρότερες αναλογίες. Το MaxHR (Μέγιστος καρδιακός ρυθμός) ιστόγραμμα, δείχνει υψηλή συγκέντρωση τιμών γύρω από τους 140 bpm. Έχει σχεδόν κανονική κατανομή με μέγιστες τιμές 190 – 200. Στο ιστόγραμμα ExerciseAngina η πλειοψηφία δεν εμφανίζει στηθάγχη κατά την διάρκεια άσκησης (N). Υπάρχει όμως σημαντικός αριθμός με τιμή (Y). Το ιστόγραμμα Oldpeak (κατάθλιψη ST) δείχνει ότι οι περισσότερες τιμές συγκεντρώνονται γύρω από το 0, δηλαδή δεν εμφανίζεται κατάθλιψη ST. Αρκετοί παρουσιάζουν ελαφρά κατάθλιψη, λίγοι όμως έχουν τιμή πάνω από 4. Στη συνέχεια, το ιστόγραμμα ST_Slope δείχνει ότι οι περισσότεροι έχουν τιμή Flat slope (επίπεδη), λιγότεροι έχουν Up (φυσιολογική) και πολύ λίγοι Down (ανησυχητική). Το τελευταίο γράφημα είναι αυτό του στόχου HeartDisease. Εδώ πολλοί ασθενείς έχουν καρδιοπάθεια (τιμή 1). Το σύνολο δεδομένων είναι ελαφρώς ανισομερές, με τους περισσότερους να παρουσιάζουν πρόβλημα υγείας.

Συγκεντρωτικά, μπορούμε να εξάγουμε τα ακόλουθα συμπεράσματα. Το σύνολο δεδομένων αφορά κυρίως μεσήλικες άνδρες. Οι περισσότεροι πάσχουν από ασυμπτωματικό πόνο στο στήθος και έχουν φυσιολογικά ή ελαφρώς αυξημένα επίπεδα πίεσης / χοληστερίνης. Η πλειοψηφία έχει ιστορικό καρδιοπάθειας, κάτι που υποδηλώνει ότι πρόκειται για κλινικά δεδομένα με σκοπό την ανάλυση πρόγνωσης ή πρόληψης.

3.3.2 Συσχετίσεις Χαρακτηριστικών

Για τη συσχέτιση των χαρακτηριστικών του συνόλου δεδομένων ανά δύο σε ένα διάγραμμα, χρησιμοποιούμε το διάγραμμα «θερμότητας» (heatmap) του πίνακα συσχέτισης με τιμές Pearson μεταξύ διαφόρων χαρακτηριστικών από το σύνολο δεδομένων, εφόσον όλες οι τιμές του συνόλου είναι αριθμητικές [37]. Πρώτα λοιπόν μετατρέπουμε όλες τις τιμές σε αριθμητικές και στη συνέχεια εξάγουμε το μητρώο ή αλλιώς πίνακα συσχέτισης για όλο το σύνολο δεδομένων σε μορφή heatmap όπως φαίνεται στην Εικόνα 3.4.



Εικόνα 3.4 Πίνακας Συσχέτισης Χαρακτηριστικών

Στον πίνακα συσχέτισης υπάρχουν οι συντελεστές συσχέτισης, οι οποίοι κυμαίνονται από το -1 έως το 1 και δείχνουν το βαθμό συσχέτισης μεταξύ δύο χαρακτηριστικών. Είναι σαφές ότι όσο πιο κοντά στο 1 βρίσκονται δύο συντελεστές τόσο πιο ισχυρή θετική συσχέτιση υπάρχει μεταξύ τους. Σε αντίθετη περίπτωση, όταν δηλαδή οι συντελεστές είναι κοντά στο -1, σημαίνει ότι μεταξύ των συντελεστών υπάρχει μικρή συσχέτιση ή διαφορετικά αρνητική συσχέτιση. Τέλος, όταν οι συντελεστές είναι κοντά στο 0, σημαίνει ότι δεν υπάρχει γραμμική συσχέτιση μεταξύ τους [12].

Η ανάλυση των συσχετίσεων δείχνει ότι η παρουσία καρδιοπάθειας έχει ισχυρή συσχέτιση με συγκεκριμένα χαρακτηριστικά. Πιο συγκεκριμένα, παρατηρείται θετική

συσχέτιση με τον ασυμπτωματικό πόνο στο στήθος (ChestPainType_ASY), την κατάθλιψη ST (Oldpeak), καθώς και με τη στενοκαρδία κατά την άσκηση (ExerciseAngina_Y), γεγονός που υποδηλώνει ότι τα άτομα με αυτά τα χαρακτηριστικά έχουν αυξημένες πιθανότητες εμφάνισης καρδιοπάθειας. Αντίθετα, παρατηρείται αρνητική συσχέτιση με την ανοδική κλίση ST (ST_Slope_Up), καθώς και με τύπους πόνου στο στήθος όπως οι NAP και ATA, που συνδέονται με χαμηλότερο κίνδυνο. Επιπλέον, η ηλικία σχετίζεται αρνητικά με τον μέγιστο καρδιακό ρυθμό, κάτι αναμενόμενο από φυσιολογικής πλευράς, ενώ η επίπεδη κλίση ST (ST_Slope_Flat) εμφανίζει την ισχυρότερη θετική συσχέτιση με την καρδιοπάθεια. Συνολικά, μεταβλητές όπως οι Oldpeak, ExerciseAngina, ST_Slope και ChestPainType_ASY φαίνεται να είναι οι πιο σημαντικές για την πρόβλεψη καρδιοπάθειας, ενώ χαρακτηριστικά όπως η χοληστερίνη (Cholesterol) και το σάκχαρο νηστείας (FastingBS) εμφανίζουν περιορισμένη προγνωστική αξία.

Πιο αναλυτικά, θετικά συσχετισμένα είναι το ChestPainType_ASY: +0.52 (ισχυρή θετική συσχέτιση), δηλαδή όσοι έχουν ασυμπτωματικό πόνο στο στήθος συχνά έχουν καρδιοπάθεια. Το χαρακτηριστικό Oldpeak: +0.40 υποδηλώνει αυξημένη κατάθλιψη ST που σχετίζεται με πιθανή καρδιοπάθεια και τέλος το χαρακτηριστικό ExerciseAngina_Y: +0.49 δείχνει στενοκαρδία με την άσκηση σχετίζεται ισχυρά με καρδιοπάθεια. Αρνητικά συσχετισμένα με την καρδιά είναι το ST_Slope_Up: -0.50, δηλαδή η κανονική ανοδική κλίση ST υποδεικνύει χαμηλότερη πιθανότητα καρδιοπάθειας. Το χαρακτηριστικό ChestPainType_NAP: -0.43 και ChestPainType_ATA: -0.38 συνδέονται με υγιέστερα άτομα. Άλλες αξιοσημείωτες συσχετίσεις είναι ο MaxHR με την Age: -0.40 (Φυσιολογικό: οι μεγαλύτεροι σε ηλικία έχουν χαμηλότερο μέγιστο καρδιακό ρυθμό). Επίσης η συσχέτιση των χαρακτηριστικών ST_Slope_Flat και HeartDisease: +0.60 (Ισχυρότερη συσχέτιση στο γράφημα — επίπεδη ST κλίση σηματοδοτεί υψηλότερο κίνδυνο).

3.3.3 Προετοιμασία δεδομένων

Το αρχικό σύνολο δεδομένων χωρίστηκε σε δύο επιμέρους μικρότερα σύνολα, το σύνολο δεδομένων εκπαίδευσης (train set), που αποτελείται από το 80% του αρχικού συνόλου και το σύνολο δεδομένων επαλήθευσης (test set), που αποτελείται από το υπόλοιπο 20% του αρχικού συνόλου δεδομένων. Η διαδικασία αυτή πραγματοποιήθηκε ξεχωριστά για όλους τους αλγορίθμους που χρησιμοποιήθηκαν. Επίσης, υπολογίστηκε ο πίνακας συσχετίσεων (confusion matrix) για κάθε έναν αλγόριθμο, καθώς και η καμπύλη ROC. Το σύνολο δεδομένων της εκπαίδευσης (train set) αποτελείται από το μεγαλύτερο μέρος των δεδομένων του αρχικού συνόλου δεδομένων. Ο καθαρισμός των δεδομένων είναι απαραίτητος πριν την υλοποίηση οποιουδήποτε αλγορίθμου, ώστε να αποφευχθεί τυχόν σύγχυση και λάθη που θα οδηγήσουν σε μη έγκυρη ταξινόμηση. Ο τρόπος με τον οποίο επιλέγουμε και

χωρίζουμε από το γενικό σύνολο δεδομένων μας το σύνολο δεδομένων εκπαίδευσης (train set), είναι ιδιαίτερα σημαντικός αφού χρειάζεται να περιλαμβάνει έναν ορισμένο όγκο πληροφορίας για την εκπαίδευση του μοντέλου, με σκοπό την εκμάθηση σωστής ταξινόμησης ή πρόβλεψης όσον το δυνατόν καλύτερου και ακριβέστερου αποτελέσματος. Προσοχή χρειάζεται ωστόσο, το μοντέλο να εκπαιδευτεί όσο χρειάζεται και όχι υπεραρκετά, καθώς θα έχουμε τα αντίθετα αποτελέσματα και όχι τα επιθυμητά. Όσο περισσότερο εκπαιδεύεται ένα μοντέλο υπάρχει ο κίνδυνος να γίνει παραπάνω εκπαίδευση από την επιθυμητή και συνεπώς τα αποτελέσματα της πρόβλεψης να μην είναι τα επιθυμητά. Αυτό συμβαίνει γιατί το μοντέλο παπαγαλίζει τα ήδη γνωστά δεδομένα σε βαθμό τέτοιο που όταν θα κληθεί να αντιμετωπίσει νέα δεδομένα να μην μπορεί να υπολογίσει το αποτέλεσμα.

Στο τέλος κάθε ανάλυσης αλγορίθμου πραγματοποιείται ο υπολογισμός των μετρικών στατιστικής ανάλυσης με σκοπό να γίνει η αξιολόγηση κάθε αλγορίθμου ξεχωριστά αλλά και η μεταξύ τους σύγκριση. Οι βασικές μετρικές για την στατιστική ανάλυση των δεδομένων που λαμβάνουμε είναι ο μέσος όρος, η τυπική απόκλιση, η παλινδρόμηση, ο έλεγχος υποθέσεων και ο προσδιορισμός του μεγέθους του δείγματος.

3.3.4 Μετρικές απόδοσης αλγορίθμων

Οι μετρικές απόδοσης για τη δυαδική ταξινόμηση είναι πολλές και η βιβλιοθήκη Scikit-Learn διαθέτει τις περισσότερες από αυτές που θα παρουσιαστούν στη συνέχεια. Η ορθότητα (accuracy) είναι το πιο συνηθισμένο μέτρο απόδοσης ενός μοντέλου. Η ορθότητα είναι η αναλογία των παραδειγμάτων για τα οποία έχει γίνει η σωστή πρόβλεψη επί του συνόλου των προβλέψεων και με βάση τις παραπάνω συμβάσεις και ισοδυναμεί με τον δείκτη λάθους (error rate) του μοντέλου [3]. Η ορθότητα δίνεται από τον τύπο:

$$acc = \frac{TP + TN}{TP + FP + FN + TN}$$

Η ορθότητα ως μέτρο απόδοσης είναι η αναλογία των παραδειγμάτων για τα οποία έγινε σωστή πρόβλεψη. Όμως, σε προβλήματα του πραγματικού κόσμου οι κλάσεις των δεδομένων στα σύνολα δεν είναι κατανεμημένες με ισορροπία. Για παράδειγμα, σε ένα σύνολο δεδομένων που αφορούν μια πολύ σπάνια περίπτωση ασθένειας μπορεί το 90% των παραδειγμάτων να ανήκουν στην αρνητική κλάση και μόνο το 10% στη θετική. Σε τέτοια περίπτωση, μπορεί να έχουμε ένα μοντέλο με υψηλή

ορθότητα, αλλά το μοντέλο δεν έχει προγνωστική δύναμη. Συνεπώς, η ορθότητα ως μετρική απόδοσης δεν αρκεί.

Η ακρίβεια (precision), είναι η αναλογία του κάθε παραδείγματος που προβλέπεται ως θετικό και είναι πραγματικά θετικό. Μπορεί να θεωρηθεί ως μέτρο θορύβου στις προβλέψεις, δηλαδή την πιθανότητα της ορθότητας για μια θετική πρόβλεψη. Ο υπολογισμός της ακρίβειας γίνεται από τον τύπο:

$$prec = \frac{TP}{TP + FP}$$

Η ανάκληση (recall), γνωστή και ως ευαισθησία (sensitivity) ή δείκτης αληθώς θετικών (True Positive Rate – TPR), είναι η αναλογία κάθε θετικού παραδείγματος που είναι πραγματικά θετικό. Η ανάκληση μετρά την ικανότητα του μοντέλου να αναγνωρίζει ένα παράδειγμα της θετικής κλάσης και υπολογίζεται από τον τύπο:

$$recall = TPR = \frac{TP}{TP + FN}$$

Συνοπτικά, η ακρίβεια είναι το κλάσμα των προβλέψεων που αναφέρθηκαν από το μοντέλο ως σωστές και η ανάκληση είναι το κλάσμα των πραγματικών παραδειγμάτων που έχουν προβλεφθεί. Σε πολλές περιπτώσεις, εάν θέλουμε να συνοψίσουμε την απόδοση του μοντέλου, επιδιώκοντας ένα είδος ισορροπίας μεταξύ της ακρίβειας και την ανάκλησης, χρησιμοποιούμε μια άλλη μετρική που συνδυάζει τις δύο μετρικές, τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης που ονομάζεται F1 αποτέλεσμα (F1-score) και υπολογίζεται από τον τύπο:

$$FI - score = 2 * \left(\frac{precision * recall}{precision + recall} \right)$$

Η προσδιοριστικότητα *(*specificity*)*, γνωστή και ως δείκτης αληθώς αρνητικών (True Negative Rate – TNR) μετρά τις ορθά αρνητικές προβλέψεις στο σύνολο των ορθών αρνητικών παραδειγμάτων και υπολογίζεται από τον τύπο:

$$specificity = TNR = \frac{TN}{TN + FP}$$

Η μετρική του δείκτη ψευδώς αρνητικών (False Positive Rate - *FPR*) αντιστοιχεί στην αναλογία των αρνητικών παραδειγμάτων που θεωρήθηκαν ως θετικά, σε σχέση με όλα τα αρνητικά παραδείγματα. Όσο μεγαλύτερος είναι ο *FPR*, πόσα περισσότερα αρνητικά παραδείγματα έχουν ταξινομηθεί λάθος. Ο *FPR* υπολογίζεται από τον τύπο:

$$FPR = \frac{FP}{FP + TN} = 1 - specificity = 1 - TNR$$

Όταν χρησιμοποιούνται ως μετρικές εκτίμησης απόδοσης του δυαδικού ταξινομητή η ακρίβεια και η ανάκληση, συνηθίζεται να απεικονίζονται γραφικά με την PR καμπύλη (PR curve) με την ακρίβεια να αντιστοιχεί στον άξονα των y και την ανάκληση στον άξονα των x.

Σε ένα πρόβλημα ταξινόμησης συνηθίζεται να δημιουργείται ο πίνακας σύγχυσης (confusion matrix), που συνήθως αναφέρεται ως πίνακας ταξινόμησης**.** Πρόκειται για ένα δισδιάστατο πίνακα, ο οποίος παρέχει τη δυνατότητα οπτικοποίησης της απόδοσης ενός μοντέλου.

Ένας εύκολος σχετικά γραφικός τρόπος και ο πιο συνηθισμένος για την ποιοτική εκτίμηση της απόδοσης ενός ταξινομητή είναι η λήψη της χαρακτηριστικής καμπύλης λειτουργίας (Receiving Operating Characteristic Curve – ROC Curve). Η ROC καμπύλη είναι η σχεδίαση της μετρικής TPR, δηλαδή της ανάκλησης, σε σχέση με την μετρική FPR και αναπαριστά τα TP και τα FP, δηλαδή την πιθανότητα στην οποία ένα παράδειγμα προβλέπεται να ανήκει σε μια κλάση. Η ROC καμπύλη χρησιμοποιείται ως μια γενική μετρική του μοντέλου. Όσο καλύτερο είναι ένα μοντέλο, τόσο υψηλότερη είναι η καμπύλη και συνεπώς τόσο μεγαλύτερη η επιφάνεια που υπάρχει κάτω από την καμπύλη. Στις βιβλιοθήκες μηχανικής μάθησης υπάρχει μια συνάρτηση για τον υπολογισμό της επιφάνειας κάτω από την ROC καμπύλη (Area Under the Curve – AUC) η οποία ονομάζεται ROC AUC ή απλά AUC. Ο τέλειος ταξινομητής έχει ROC AUC ίση με 1, ενώ ένας κακός ταξινομητής θα έχει 0.5, ίσως και λιγότερο. Συνεπώς, όσο πιο κοντά η ROC AUC στο 1, τόσο καλύτερος είναι ο ταξινομητής.

Ένας άλλος συντελεστής ποιοτικός συντελεστής που χρησιμοποιείται στην ταξινόμηση, είναι ο συντελεστής συσχέτισης του Matthews (Matthews Correlation Coefficient - MCC) [32]. Ο MCC χρησιμοποιείται στη Μηχανική Μάθηση ως μέτρο της ποιότητας της ταξινόμησης. Λαμβάνει υπόψη τα αληθώς και ψευδώς θετικά και αρνητικά και θεωρείται γενικά ως ισορροπημένο μέτρο που μπορεί να χρησιμοποιηθεί ακόμη και αν οι τάξεις έχουν πολύ διαφορετικά μεγέθη. Ο MCC είναι ουσιαστικά μια τιμή συντελεστή συσχέτισης μεταξύ -1 και +1. Ένας συντελεστής +1 αντιπροσωπεύει μια τέλεια πρόβλεψη, 0 μια μέση τυχαία πρόβλεψη και -1 μια αντίστροφη πρόβλεψη. Ο MCC υπολογίζεται από τον τύπο:

$$MCC = TP * TN - FP * FN \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

3.3.5. Τεχνικές Αποτίμησης

Η διαδικασία δημιουργίας ενός προγνωστικού μοντέλου για μια εργασία και ένα συγκεκριμένο σύνολο δεδομένων είναι η εκπαίδευση. Ο αλγόριθμος Μηχανικής Μάθησης με συνεχείς δοκιμές και επαναλήψεις καλείται να εξάγει το ορθό αποτέλεσμα για τον στόχο. Μετά από ορισμένες επαναλήψεις, το μοντέλο αποτιμάται με βάση τα

μέτρα απόδοσης. Εφόσον η απόδοση δεν είναι ικανοποιητική, αναζητούνται οι βέλτιστες παράμετροι και το μοντέλο εκπαιδεύεται ξανά. Όταν η απόδοση είναι ικανοποιητική, με βάση τις βέλτιστες παραμέτρους, το μοντέλο αποτιμάται τελικά σε ένα σύνολο δεδομένων που δεν έχει ξαναδεί. Η δυνατότητα του μοντέλου να αποδίδει καλά σε νέα δεδομένα ονομάζεται γενίκευση (generalization).
Η εκμάθηση των παραμέτρων μιας συνάρτησης πρόβλεψης και η δοκιμή της στα ίδια δεδομένα είναι μεθοδολογικό λάθος: ένα μοντέλο που θα επαναλάμβανε τις προβλέψεις στα ίδια δεδομένα θα είχε τέλεια απόδοση, αλλά δεν θα μπορούσε να κάνει σωστές προβλέψεις σε νέα δεδομένα. Αυτή η κατάσταση ονομάζεται υπερταίριασμα ή υπερ προσαρμογή (overfitting), όπου το μοντέλο διαχωρίζει μεν πολύ σωστά τα δεδομένα, αλλά με μεγάλη λεπτομέρεια και θα αποτύχει στην περίπτωση εισαγωγής νέων δεδομένων όπου θα δημιουργηθεί μεγάλο λάθος γενίκευσης. Για την αποφυγή της υπερ προσαρμογής, μια συνηθισμένη πρακτική κατά την εκτέλεση ενός εποπτευόμενου πειράματος Μηχανικής Μάθησης το μοντέλο να εκπαιδεύεται σε ένα μέρος του συνόλου δεδομένων και η απόδοσή του να μετράται, δηλαδή να επικυρώνεται κατά κάποιο τρόπο, στο υπόλοιπο του συνόλου δεδομένων. Αφού ρυθμιστούν οι παράμετροι του μοντέλου με βάση τα δεδομένα της εκπαίδευσης (train data), το μοντέλο αποτιμάται για τα δεδομένα της δοκιμής (test data).

Μια συνηθισμένη τεχνική επικύρωσης (validation) όταν διατίθεται ένα συγκεκριμένο σύνολο δεδομένων για πειραματισμό, είναι το σύνολο των δεδομένων να διαχωρίζεται (split) σε δύο υποσύνολα σε μια αναλογία π.χ. 80% - 20%, όπου το μοντέλο θα εκπαιδευτεί στο 80% των παραδειγμάτων και το υπόλοιπο 20% των παραδειγμάτων θα χρησιμοποιηθεί για να μετρηθεί η απόδοση του μοντέλου.

Στην παρούσα εργασία ως τεχνική επικύρωσης επιλέχθηκε η τεχνική διαχωρισμού του αρχικού συνόλου των δεδομένων σε ποσοστά 80% - 20%. Σημαντικός παράγοντας κατά τον διαχωρισμό των δεδομένων είναι να διασφαλίζεται ότι κάθε φορά θα λαμβάνεται το ίδιο υποσύνολο στιγμιότυπων των δεδομένων και για το στάδιο της εκπαίδευσης και για το στάδιο της δοκιμής αντίστοιχα. Η δυνατότητα αυτή παρέχεται από την βιβλιοθήκη scikit-learn, μέσω της παραμέτρου random state, η οποία συνήθως λαμβάνει τιμές από 0 έως 42.

3.3.6 Επιλογή αλγορίθμων μηχανικής μάθησης

Για τη δημιουργία του καταλληλότερου προγνωστικού μοντέλου με βάση το σύνολο των δεδομένων που περιέχει κατά βάση τα συμπτώματα της ασθένειας, επιλέχθηκαν οι πιο αντιπροσωπευτικοί από τους δημοφιλέστερους αλγόριθμους δυαδικής ταξινόμησης που διαθέτει η βιβλιοθήκη scikit-learn. Πιο συγκεκριμένα, επιλέχθηκαν για τις προβλέψεις τρεις συνολικά αλγόριθμοι:

1. Η Λογιστική Παλινδρόμηση (LR)
2. Οι Μηχανές Υποστήριξης Διανυσμάτων (SVM)
3. Ο Αφελής Μπεϋζιανός (NB)

Κάθε αλγόριθμος που διατίθεται από την scikit-learn διατίθεται ως κλάση της Python, και υπάρχουν εξ' ορισμού (default) τιμές για τις σχετικές υπέρ παραμέτρους του, αναφέρονται απλά ως παράμετροι, όπου μπορεί κανείς να τις δει συνοπτικά στην βιβλιογραφία του ιστοτόπου της βιβλιοθήκης.

Στην παρούσα εργασία, για την εκπαίδευση των μοντέλων για κάθε έναν από τους αλγόριθμους που επιλέχθηκαν ως αρχικές παράμετροι αφέθηκαν οι εξ ορισμού παράμετροι της βιβλιοθήκης scikit-learn, με το σκεπτικό ότι, αυτές οι παράμετροι μπορούν στη συνέχεια να αλλάξουν εφόσον για κάποιο μοντέλο που δεν αποδίδει καλά, μπορούν να αλλάξουν προκειμένου να βελτιστοποιηθεί. Αξίζει να σημειωθεί ότι, για πολλούς αλγόριθμους Μηχανικής Μάθησης προκειμένου να έχουν καλή απόδοση, είναι πολύ σημαντική η κλιμάκωση των χαρακτηριστικών (feature scaling). Ο πιο συνηθισμένος τρόπος για την κλιμάκωση των αριθμητικών τιμών στο ίδιο εύρος είναι η κανονικοποίηση (normalization), όπου οι αριθμητικές τιμές μετατρέπονται στο εύρος [0, 1]. Η βιβλιοθήκη scikit-learn παρέχει τη συνάρτηση MinMaxScaler για τον σκοπό αυτό. Ένας άλλος τρόπος είναι η προτυποποίση (standardization) κατά την οποία οι τιμές των χαρακτηριστικών κλιμακώνονται έτσι ώστε να έχουν τις ιδιότητες της πρότυπης κανονικής κατανομής με μέση τιμή μ=0 και τυπική απόκλιση σ=1. Η βιβλιοθήκη scikit-learn παρέχει τη συνάρτηση StandardScaler για τον σκοπό αυτό. Συνεπώς, η κλιμάκωση των δεδομένων κρίνεται απαραίτητη, θα εφαρμοστούν και οι δοκιμές γίνονται και με τις δύο τεχνικές κλιμάκωσης.

4. Αποτελέσματα

4.1 Υλοποίηση Λογιστικής Παλινδρόμησης

Αρχικά φτιάξαμε το μοντέλο της λογιστικής παλινδρόμησης το οποίο χρειάζεται να εκπαιδεύσουμε. Από τη διαδικασία εκπαίδευσης προκύπτει το ακόλουθο αποτέλεσμα:

```
Number of iterations: [7]
Accuracy of Logistic Regression model - test set: 59.2391304347826%

The Beta parameters: [[ 1.31236456e+00  1.38026389e+00 -2.38481887e+00  4.72906390e-02
  -4.24002758e+00  2.56413717e-01 -6.08186007e-02  6.31673219e-02
   1.27483716e-01 -7.59209743e-02 -4.38973066e-02 -3.34681914e-03
   1.52543152e-02 -2.59052251e-02  1.39845786e-02 -1.34846470e-01
   1.37195191e-01  1.48415340e-02  1.48950643e-01 -1.60458508e-01]]

The Intercept (bias): [1.54039342]

Model performance for test set
--------------------------------
R2 score: -67.49605534652261%
RMSE: 40.76086956521739%
```

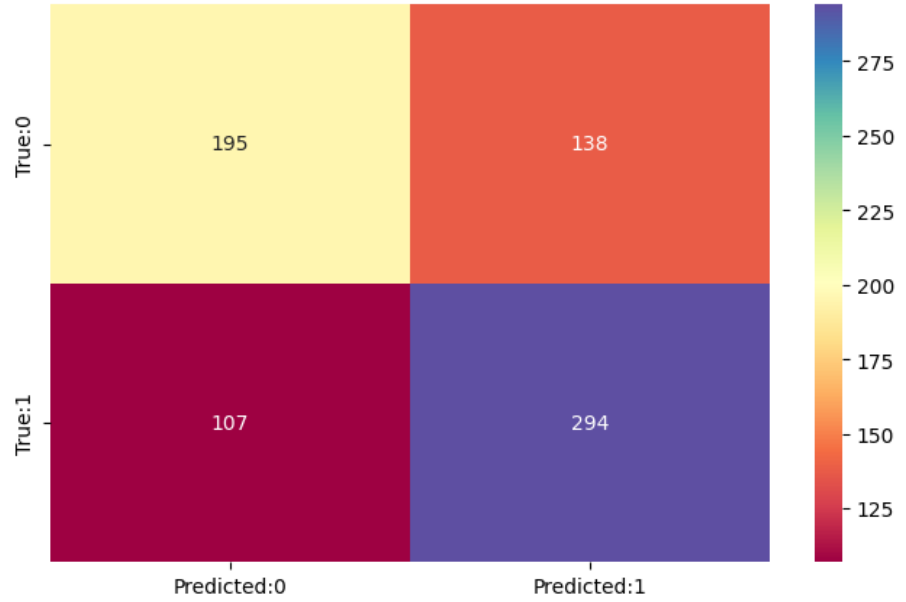
Εικόνα 4.1 Μοντέλο Λογιστικής Παλινδρόμησης

Όπως προκύπτει από την παραπάνω εικόνα, το μοντέλο ελέγχου πραγματοποίησε 7 επαναλήψεις εκπαίδευσης. Η ακρίβεια στο σύνολο δοκιμής (test set) είναι 59.2%, που σημαίνει ότι το μοντέλο ταξινόμησε σωστά περίπου το 59% των παραδειγμάτων στο test set. Οι Beta parameters είναι οι συντελεστές (βάρη) του μοντέλου για κάθε χαρακτηριστικό (feature). Είναι ένας πίνακας 2x10, που δείχνει τις τιμές των συντελεστών για κάθε μία από τις δύο κατηγορίες. Μεγάλοι θετικοί ή αρνητικοί συντελεστές δείχνουν ισχυρή συσχέτιση με την αντίστοιχη κατηγορία. Π.χ. το -4.24 στην πρώτη γραμμή δείχνει ισχυρή αρνητική επίδραση του αντίστοιχου χαρακτηριστικού στην πρόβλεψη αυτής της κατηγορίας. Η απόδοση του μοντέλου στο Test Set αξιολογείται επίσης από τους συντελεστές R^2 Score και RMSE. Το R^2 Score μας έδωσε αρνητική τιμή -67.50%, πράγμα που σημαίνει ότι το μοντέλο αποδίδει χειρότερα και από μία απλή πρόβλεψη με τον μέσο όρο των τιμών. Το R^2 χρησιμοποιείται κυρίως για παλινδρόμηση και η χρήση του σε λογιστική παλινδρόμηση είναι ασυνήθιστη. Η Ρίζα του Μέσου Τετραγωνικού Σφάλματος (RMSE) έδωσε τιμή 40.76 και δείχνει τη μέση απόσταση των προβλέψεων από τις πραγματικές τιμές. Όσο πιο μικρή τιμή έχει τόσο το καλύτερο.

Το μοντέλο δεν αποδίδει καλά και αυτό φαίνεται από τη χαμηλή ακρίβεια (~59%), το πολύ αρνητικό R^2 και το υψηλό RMSE. Πιθανές αιτίες μπορεί να είναι η

κακή ποιότητα των δεδομένων, να μην έχουν γίνει σωστά τα αρχικά βήματα της προ επεξεργασίας των δεδομένων ή η λογιστική παλινδρόμηση μπορεί να μην είναι το κατάλληλο μοντέλο για την επίλυση αυτού του προβλήματος.

Στη συνέχεια δημιουργήσαμε τον πίνακα σύγχυσης (confusion matrix) για το σύνολο δεδομένων του train set και πήραμε τα παρακάτω αποτελέσματα:



Εικόνα 4.2 Λογιστική Παλινδρόμηση - Πίνακας Σύγχυσης train set

Classification report of LR Confusion matrix - Train set :

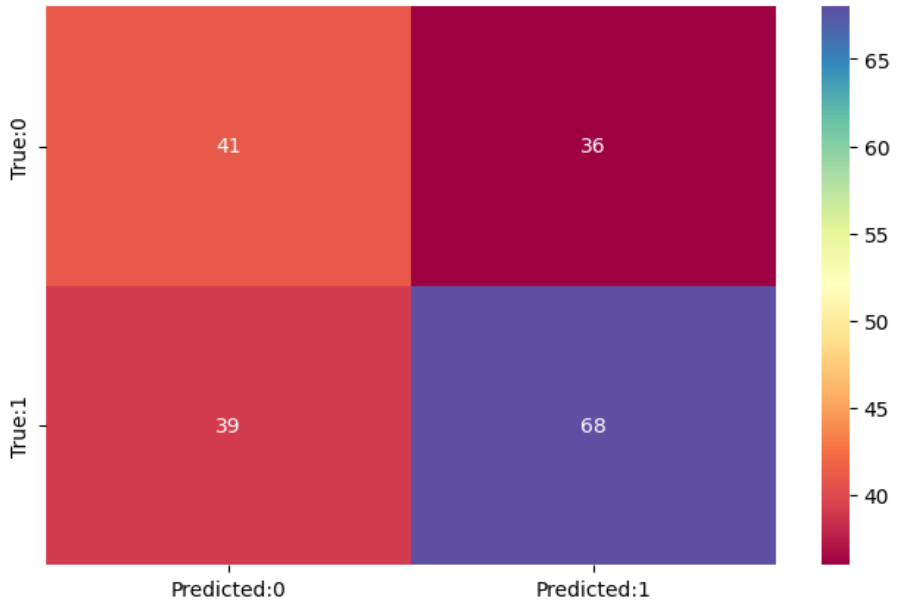
	precision	recall	f1-score	support
0	0.65	0.59	0.61	333
1	0.68	0.73	0.71	401
accuracy			0.67	734
macro avg	0.66	0.66	0.66	734
weighted avg	0.66	0.67	0.66	734

Εικόνα 4.3 Λογιστική Παλινδρόμηση - Report train set πίνακα σύγχυσης

Όπως έχει ήδη αναφερθεί και σε προηγούμενη ενότητα, ο πίνακας σύγχυσης (confusion matrix) είναι ένα σημαντικό εργαλείο που μας βοηθάει να εξάγουμε συμπεράσματα για την απόδοση ενός μοντέλου. Στον πίνακα σύγχυσης παρουσιάζονται οι τέσσερις διαφορετικοί συνδυασμοί των προβλεπόμενων και των

πραγματικών τιμών του ταξινομητή. Αυτοί οι συνδυασμοί είναι ο TP (αληθώς θετικά), ο TN (αληθώς αρνητικά), ο FP (ψευδώς θετικά) και ο FN (ψευδώς αρνητικά). Ο πίνακας σύγχυσης του train set δείχνει ότι 195 ασθενείς προβλέφθηκαν ότι θα πάσχουν από καρδιακές παθήσεις και ότι είναι πιθανό να πάθουν καρδιακή προσβολή, με την πρόβλεψη να είναι ΣΩΣΤΗ (TP). 294 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, η πρόβλεψη είναι ΣΩΣΤΗ (TN). 138 ασθενείς προβλέφθηκαν ότι θα υποφέρουν από καρδιακές παθήσεις και είναι πιθανό να πάθουν καρδιακή προσβολή, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FP). 107 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, όμως η πρόβλεψη είναι ΛΑΘΟΣ (FN). Η κατάσταση ταξινόμησης του μοντέλου δείχνει ότι το 68% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να πάθουν καρδιακή προσβολή (1) προβλέφθηκε σωστά και το 65% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να μην πάθουν καρδιακή προσβολή (0) προβλέφθηκε επίσης σωστά.

Σχεδιάσαμε αντίστοιχα και τον πίνακα σύγχυσης (confusion matrix) για τα δεδομένα του test set, όπως παρουσιάζεται παρακάτω:



Εικόνα 4.4 Λογιστική Παλινδρόμηση - Πίνακας Σύγχυσης test set

Classification report of LR Confusion matrix - Test set:

	precision	recall	f1-score	support
0	0.5125	0.5325	0.5223	77
1	0.6538	0.6355	0.6445	107
accuracy			0.5924	184
macro avg	0.5832	0.5840	0.5834	184
weighted avg	0.5947	0.5924	0.5934	184

Εικόνα 4.5 Λογιστική Παλινδρόμηση - Report test set πίνακα σύγχυσης

Ο πίνακας σύγχυσης του test set δείχνει ότι 41 ασθενείς προβλέφθηκαν ότι θα πάσχουν από καρδιακές παθήσεις και ότι είναι πιθανό να πάθουν καρδιακή προσβολή, η πρόβλεψη είναι ΣΩΣΤΗ (TP). 68 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, η πρόβλεψη είναι ΣΩΣΤΗ (TN). 36 ασθενείς προβλέφθηκαν ότι θα υποφέρουν από καρδιακές παθήσεις και είναι πιθανό να πάθουν έμφραγμα, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FP). 39 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FN). Η κατάσταση ταξινόμησης του μοντέλου δείχνει ότι το 65% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να πάθουν καρδιακή προσβολή (1) προβλέφθηκε σωστά και το 51% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να μην πάθουν καρδιακή προσβολή (0) προβλέφθηκε επίσης σωστά.

Για το train set μπορούμε επίσης να υπολογίσουμε τις μετρικές sensitivity, specificity, positive predictive value και negative predictive value όπως παρουσιάζεται παρακάτω:

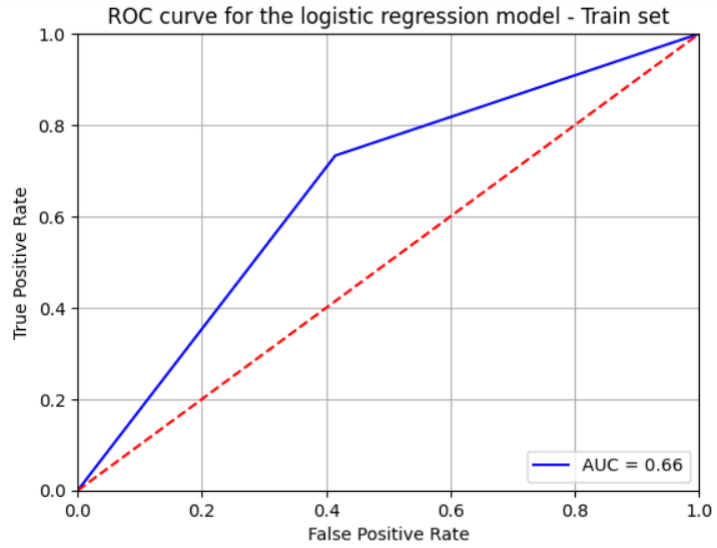
```
Sensitivity: 63.55140186915887

Specificity: 53.246753246753244

Positive Predictive Value: 65.38461538461539

Negative Predictive Value: 51.24999999999999
```

Εικόνα 4.6 Λογιστική Παλινδρόμηση – Μετρικές train set πίνακα σύγχυσης

Από τις παραπάνω μετρήσεις καταλαβαίνουμε ότι το μοντέλο λογιστικής παλινδρόμησης έχει μεγαλύτερο ποσοστό sensitivity από specificity. Αυτό σημαίνει ότι το μοντέλο είναι σε θέση να εντοπίσει το 63,5% των ασθενών που πάσχουν από καρδιακές παθήσεις και το 53,2% των ασθενών που δεν πάσχουν. Η υψηλή Θετική Προβλεπτική Τιμή (Positive Predictive Value) μπορεί να δώσει ικανοποιητικά αποτελέσματα 65,3% ότι ο ασθενής έχει καρδιακή νόσο σε περίπτωση θετικού

αποτελέσματος, ενώ η υψηλότερη Αρνητική Προβλεπτική Τιμή (Negative Predictive Value) μπορεί σίγουρα να πει το 51,2% της πιθανότητας ότι ο ασθενής δεν έχει τη νόσο.

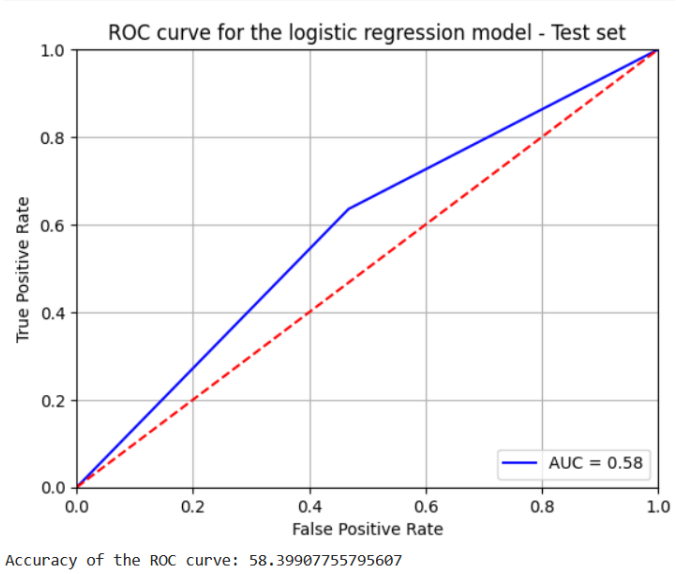
Έπειτα, έγινε αναπαράσταση της καμπύλης ROC για να απεικονιστεί η απόδοση του μοντέλου της λογιστικής παλινδρόμησης στο σύνολο δεδομένων train test και για το σύνολο δεδομένων test set αντίστοιχα.



Accuracy of the ROC curve: 65.93763339399248

Εικόνα 4.7 Λογιστική Παλινδρόμηση – Καμπύλη ROC για το train set

Από το παραπάνω γράφημα παίρνουμε τις ακόλουθες πληροφορίες, που αφορούν το σύνολο δεδομένων εκπαίδευσης του ταξινομητή της λογιστικής παλινδρόμησης. Η μπλε καμπύλη αναπαριστά την πραγματική καμπύλη ROC του μοντέλου. Δείχνει την απόδοση του σε διαφορετικά κατώφλια πιθανότητας. Κάθε σημείο πάνω στη γραμμή δείχνει τη σχέση μεταξύ των σωστών θετικών προβλέψεων (TPR) (ευαισθησία) και των λανθασμένων θετικών προβλέψεων (FPR). Όσο πιο πάνω και αριστερά είναι η καμπύλη τόσο καλύτερη απόδοση έχει. Η κόκκινη γραμμή αντιπροσωπεύει έναν τυχαίο ταξινομητή με Το AUC όπως έχει ήδη αναφερθεί, είναι το εμβαδόν που βρίσκεται κάτω από την καμπύλη ROC και μετράει την ικανότητα του μοντέλου να διακρίνει τις θετικές από τις αρνητικές κλάσεις. Η τιμή AUC=0.66 μας πληροφορεί πως το μοντέλο έχει διακριτική ικανότητα, η οποία δεν είναι ιδιαίτερα ισχυρή. Γενικά το AUC μετρά την απόσταση της μπλε από την κόκκινη γραμμή, δηλαδή πόσο πιο ψηλά βρίσκεται η μπλε από την κόκκινη γραμμή. Κυμαίνεται στο διάστημα από 0.5 (τυχαία πρόβλεψη) έως 1.0 (τέλεια πρόβλεψη). Επομένως, η τιμή του 0.66 υποδηλώνει ότι το μοντέλο είναι καλό αλλά χρειάζεται βελτίωση, καθώς μπορεί να κάνει διάκριση μεταξύ των κλάσεων αλλά υστερεί σε αξιοπιστία.

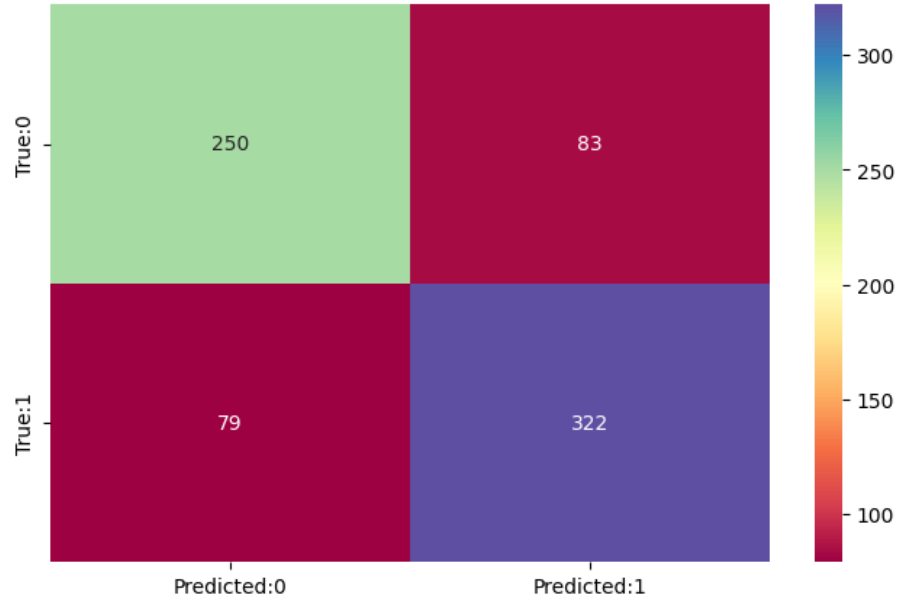


Εικόνα 4.8 Λογιστική Παλινδρόμηση – Καμπύλη ROC για το test set

Το παραπάνω γράφημα μας δίνει πληροφορίες, που αφορούν το σύνολο δεδομένων ελέγχου. Εδώ το AUC = 0.58 συνεπάγεται ότι το μοντέλο έχει μικρή διαχωριστική ικανότητα. Μερικά αίτια στα οποία θα μπορούσε να οφείλεται η χαμηλή απόδοση του μοντέλου είναι η χαμηλή ποιότητα των χαρακτηριστικών, η ανεπαρκής εκπαίδευση ή η ακαταλληλότητα του μοντέλου λογιστικής παλινδόμησης για την επίλυση του συγκεκριμένου προβλήματος.

4.2 Υλοποίηση Μηχανών Υποστήριξης Διανυσμάτων

Στη συνέχεια, δημιουργήσαμε το μοντέλο του SVM ταξινομητή. Η διαδικασία που ακολουθήσαμε είναι όμοια με αυτής της λογιστικής παλινδρόμησης. Δηλαδή, χρησιμοποιήσαμε το ίδιο σύνολο δεδομένων για το train set και το ίδιο σύνολο δεδομένων για το test set με την λογιστική παλινδρόμηση και δημιουργήσαμε τον πίνακα σύγχυσης και την καμπύλη ROC.



Εικόνα 4.9 Μηχανές Υποστήριξης Διανυσμάτων - Πίνακας Σύγχυσης train set

Classification report of SVM Confusion matrix - Train set:

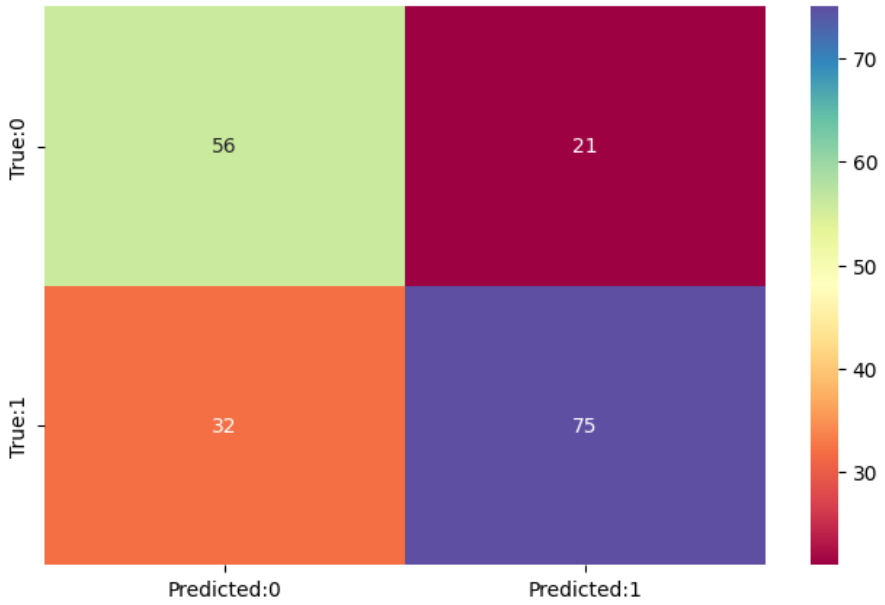
	precision	recall	f1-score	support
0	0.76	0.75	0.76	333
1	0.80	0.80	0.80	401
accuracy			0.78	734
macro avg	0.78	0.78	0.78	734
weighted avg	0.78	0.78	0.78	734

Εικόνα 4.10 Μηχανές Υποστήριξης Διανυσμάτων - Report train set πίνακα σύγχυσης

Ο πίνακας σύγχυσης του test set δείχνει ότι 250 ασθενείς προβλέφθηκαν ότι θα πάσχουν από καρδιακές παθήσεις και ότι είναι πιθανό να πάθουν καρδιακή προσβολή, η πρόβλεψη είναι ΣΩΣΤΗ (TP). 322 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, η πρόβλεψη είναι ΣΩΣΤΗ (TN). 83 ασθενείς προβλέφθηκαν ότι θα υποφέρουν από καρδιακές παθήσεις και είναι πιθανό να πάθουν έμφραγμα, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FP). 79 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FN). Η κατάσταση ταξινόμησης του μοντέλου δείχνει ότι το 80% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να πάθουν καρδιακή προσβολή (1)

προβλέφθηκε σωστά και το 76% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να μην πάθουν καρδιακή προσβολή (0) προβλέφθηκε επίσης σωστά.

Σχεδιάσαμε αντίστοιχα και τον πίνακα σύγχυσης (confusion matrix) για τα δεδομένα του test set, όπως παρουσιάζεται παρακάτω:



Εικόνα 4.11 Μηχανές Υποστήριξης Διανυσμάτων - Πίνακας Σύγχυσης test set

Classification report of SVM Confusion matrix - Test set:

	precision	recall	f1-score	support
0	0.64	0.73	0.68	77
1	0.78	0.70	0.74	107
accuracy			0.71	184
macro avg	0.71	0.71	0.71	184
weighted avg	0.72	0.71	0.71	184

Εικόνα 4.12 Μηχανές Υποστήριξης Διανυσμάτων - Report test set πίνακα σύγχυσης

Ο πίνακας σύγχυσης του test set δείχνει ότι 56 ασθενείς προβλέφθηκαν ότι θα πάσχουν από καρδιακές παθήσεις και ότι είναι πιθανό να πάθουν καρδιακή προσβολή, η πρόβλεψη είναι ΣΩΣΤΗ (TP). 75 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, η πρόβλεψη είναι ΣΩΣΤΗ (TN). 21 ασθενείς προβλέφθηκαν ότι θα

υποφέρουν από καρδιακές παθήσεις και είναι πιθανό να πάθουν έμφραγμα, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FP). 32 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FN). Η κατάσταση ταξινόμησης του μοντέλου δείχνει ότι το 78% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να πάθουν καρδιακή προσβολή (1) προβλέφθηκε σωστά και το 64% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να μην πάθουν καρδιακή προσβολή (0) προβλέφθηκε επίσης σωστά.

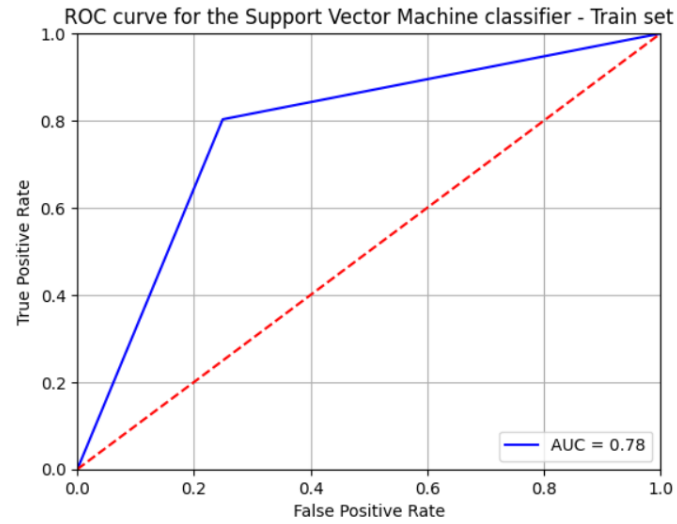
Για το train set μπορούμε επίσης να υπολογίσουμε τις μετρικές sensitivity, specificity, positive predictive value και negative predictive value όπως παρουσιάζεται παρακάτω:

```
Sensitivity: 70.09345794392523

Specificity: 72.72727272727273

Positive Predictive Value: 78.125

Negative Predictive Value: 63.63636363636363
```

Εικόνα 4.13 Μηχανές Υποστήριξης Διανυσμάτων – Μετρικές train set πίνακα σύγχυσης

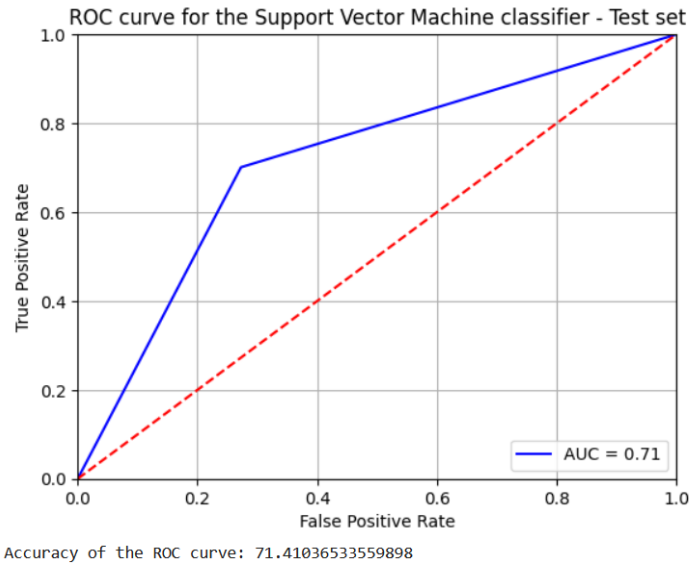
Από τις παραπάνω μετρήσεις καταλαβαίνουμε ότι το μοντέλο support vector machine έχει μεγαλύτερο ποσοστό specificity από sensitivity. Αυτό σημαίνει ότι το μοντέλο είναι σε θέση να αναγνωρίσει το 70% των ασθενών που πάσχουν από καρδιακές παθήσεις και το 72,7% των ασθενών που δεν πάσχουν. Η υψηλή Θετική Προβλεπτική Τιμή (Positive Predictive Value) μπορεί να δώσει ικανοποιητικά αποτελέσματα 78% ότι ο ασθενής έχει καρδιακή νόσο σε περίπτωση θετικού αποτελέσματος, ενώ η υψηλότερη Αρνητική Προβλεπτική Τιμή (Negative Predictive Value) μπορεί σίγουρα να πει το 63,6% της πιθανότητας ότι ο ασθενής δεν έχει τη νόσο.

Έπειτα, έγινε αναπαράσταση της καμπύλης ROC για να απεικονιστεί η απόδοση του μοντέλου του Support Vector Machine στο σύνολο δεδομένων train test και για το σύνολο δεδομένων test set αντίστοιχα.



Εικόνα 4.14 Μηχανές Υποστήριξης Διανυσμάτων – Καμπύλη ROC για το train set

Όμοια με το γράφημα της λογιστικής παλινδρόμησης, το παραπάνω γράφημα για το σύνολο δεδομένων εκπαίδευσης του ταξινομητή SVM έχουμε τη μπλε γραμμή για την καμπύλη ROC του μοντέλου και την κόκκινη γραμμή των τυχαίων προβλέψεων. Το αποτέλεσμα του AUC = 0.78. Αυτό σημαίνει ότι το μοντέλο έχει 78% πιθανότητα να ταξινομεί σωστά ένα θετικό δείγμα από ότι ένα αρνητικό. Η τιμή 0.78 υποδηλώνει καλή προβλεπτική ικανότητα. Σε αντίθεση, με την καμπύλη ROC της λογιστικής παλινδρόμησης, η καμπύλη αυτή παρουσιάζει απότομη άνοδο και δείχνει ότι το μονέλο έχει την ικανότητα να εντοπίζει πολλά αληθώς θετικά (TP) με χαμηλό false positive rate (FPR) – θετικό. Στη συνέχεια, παρατηρείται μια κάμψη, η οποία υποδηλώνει ότι όσο το κατώφλι μικραίνει τόσο αυξάνονται τα (FP) ψευδώς θετικά. Γενικά η πορεία πάνω από τη διαγώνιο δείχνει ότι το μοντέλο έχει καλύτερη απόδοση από την τύχη.

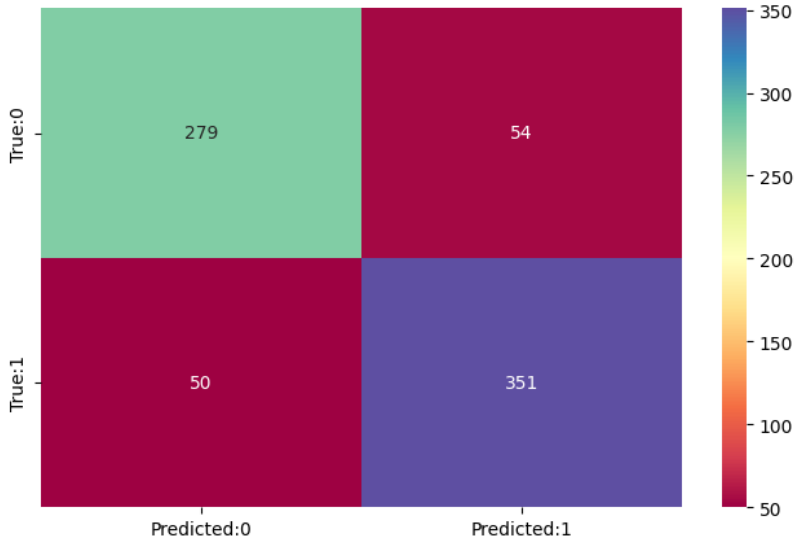


Εικόνα 4.15 Μηχανές Υποστήριξης Διανυσμάτων – Καμπύλη ROC για το test set

Το παραπάνω γράφημα μας δίνει πληροφορίες, που αφορούν το σύνολο δεδομένων ελέγχου. Εδώ το AUC = 0.71 σημαίνει ότι το μοντέλο έχει καλή απόδοση, καλύτερη από τυχαία απόδοση (0.5) και έχει ικανότητα διάκρισης μεταξύ των δύο τάξεων. Παρατηρούμε όμως ότι σε σχέση με το train set, υπάρχει μικρή μείωση του AUC (από 0.78 → 0.71) η οποία είναι αναμενόμενη και δείχνει ότι το μοντέλο δεν κάνει overfitting σε βαθμό ανησυχητικό. Η καμπύλη ROC έχει καλή καμπυλότητα, δηλαδή συγκεντρώνεται πάνω αριστερά – μια ένδειξη ότι το μοντέλο κάνει καλές προβλέψεις. Ο συνδιασμός με άλλους δείκτες όπως η ακρίβεια ή το F1 score θα δώσει καλύτερη εικόνα.

4.3 Υλοποίηση Αφελή Μπεϋζιανού

Τέλος, δημιουργήσαμε το μοντέλο του Αφελή Μπεϋζιανού. Όπως και στην περίπτωση των άλλων δύο ταξινομητών – της λογιστικής παλινδρόμησης και του Μηχανών Υποστήριξης Διανυσμάτων - δημιουργήσαμε τον πίνακα σύγχυσης και την καμπύλη ROC για το σύνολο δεδομένων εκπαίδευσης και ελέγχου αντίστοιχα.



Εικόνα 4.16 Αφελής Μπεϋζιανός - Πίνακας Σύγχυσης train set

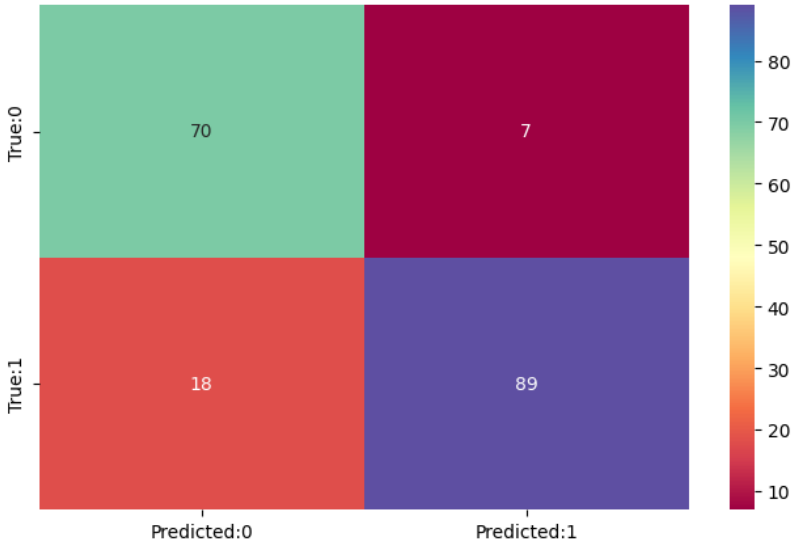
Classification report of NB Confusion matrix - Train set:

	precision	recall	f1-score	support
0	0.85	0.84	0.84	333
1	0.87	0.88	0.87	401
accuracy			0.86	734
macro avg	0.86	0.86	0.86	734
weighted avg	0.86	0.86	0.86	734

Εικόνα 4.17 Αφελής Μπεϋζιανός - Report train set πίνακα σύγχυσης

Ο πίνακας σύγχυσης του συνόλου εκπαίδευσης (test set) δείχνει ότι 279 ασθενείς προβλέφθηκαν ότι θα πάσχουν από καρδιακές παθήσεις και ότι είναι πιθανό να πάθουν καρδιακή προσβολή, η πρόβλεψη είναι ΣΩΣΤΗ (TP). 351 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, η πρόβλεψη είναι ΣΩΣΤΗ (TN). 54 ασθενείς προβλέφθηκαν ότι θα υποφέρουν από καρδιακές παθήσεις και είναι πιθανό να πάθουν έμφραγμα, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FP). 50 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FN). Η κατάσταση ταξινόμησης του μοντέλου δείχνει ότι το 87% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να πάθουν καρδιακή προσβολή (1) προβλέφθηκε σωστά και το 85% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να μην πάθουν καρδιακή προσβολή (0) προβλέφθηκε επίσης σωστά.

Σχεδιάσαμε αντίστοιχα και τον πίνακα σύγχυσης (confusion matrix) για τα δεδομένα του συνόλου δεδομένων ελέγχου (test set), όπως παρουσιάζεται παρακάτω:



Εικόνα 4.18 Αφελής Μπεϋζιανός - Πίνακας Σύγχυσης test set

Classification report of NB Confusion matrix - Test set:

	precision	recall	f1-score	support
0	0.80	0.91	0.85	77
1	0.93	0.83	0.88	107
accuracy			0.86	184
macro avg	0.86	0.87	0.86	184
weighted avg	0.87	0.86	0.86	184

Εικόνα 4.19 Αφελής Μπεϋζιανός - Report test set πίνακα σύγχυσης

Ο παραπάνω πίνακας σύγχυσης δείχνει ότι 70 ασθενείς προβλέφθηκαν ότι θα πάσχουν από καρδιακές παθήσεις και ότι είναι πιθανό να πάθουν καρδιακή προσβολή, η πρόβλεψη είναι ΣΩΣΤΗ (TP). 89 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, η πρόβλεψη είναι ΣΩΣΤΗ (TN). 7 ασθενείς προβλέφθηκαν ότι θα υποφέρουν από καρδιακές παθήσεις και είναι πιθανό να πάθουν έμφραγμα, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FP). 18 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FN). Η κατάσταση ταξινόμησης του μοντέλου δείχνει ότι το 93% των προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να πάθουν καρδιακή προσβολή (1) προβλέφθηκε σωστά και το 80% των

προβλέψεων που αφορούσε την πιθανότητα οι ασθενείς να μην πάθουν καρδιακή προσβολή (0) προβλέφθηκε επίσης σωστά.

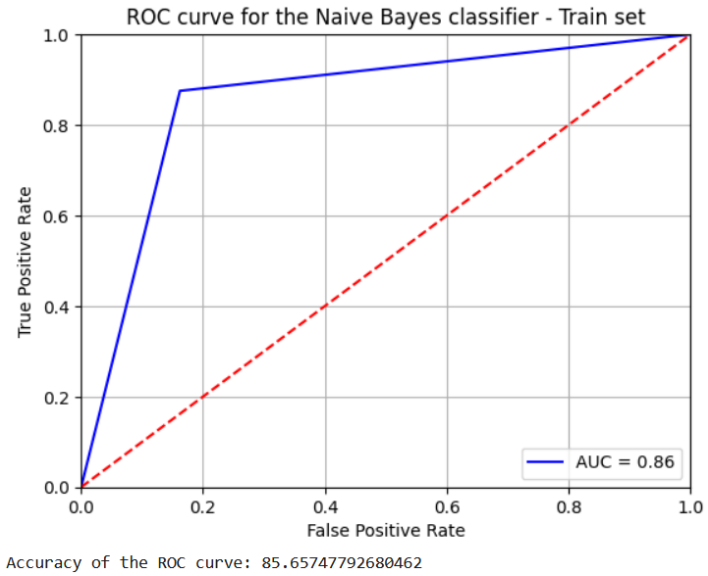
Για το σύνολο εκπαίδευσης (train set) μπορούμε επίσης να υπολογίσουμε τις μετρικές sensitivity, specificity, positive predictive value και negative predictive value όπως παρουσιάζεται παρακάτω:

```
Sensitivity: 83.17757009345794

Specificity: 90.9090909090909

Positive Predictive Value: 92.70833333333334

Negative Predictive Value: 79.54545454545455
```

Εικόνα 4.20 Αφελής Μπεϋζιανός – Μετρικές train set πίνακα σύγχυσης

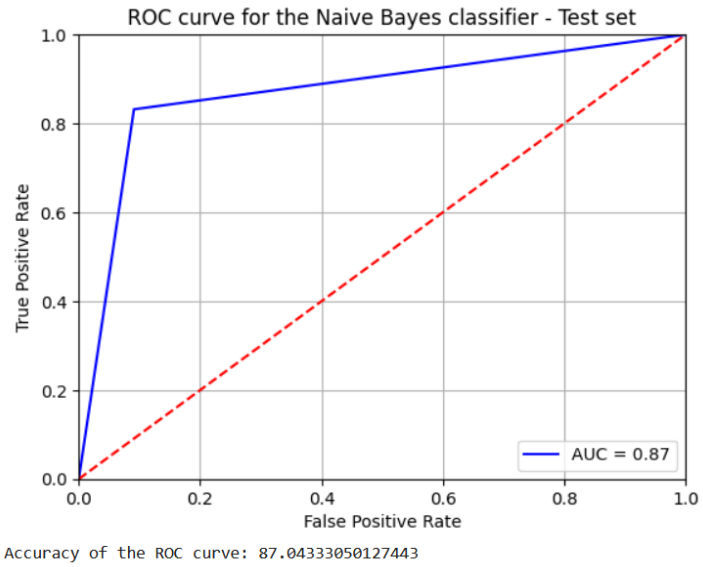
Από τις παραπάνω μετρήσεις καταλαβαίνουμε ότι το μοντέλο του αφελή Μπεϋζιανού έχει μεγαλύτερο ποσοστό specificity από sensitivity. Αυτό σημαίνει ότι το μοντέλο είναι σε θέση να αναγνωρίσει το 90,9% των ασθενών που πάσχουν από καρδιακές παθήσεις και το 83,1% των ασθενών που δεν πάσχουν. Η υψηλή Θετική Προβλεπτική Τιμή (Positive Predictive Value) μπορεί να δώσει ικανοποιητικά αποτελέσματα 92,7% ότι ο ασθενής έχει καρδιακή νόσο σε περίπτωση θετικού αποτελέσματος, ενώ η Αρνητική Προβλεπτική Τιμή (Negative Predictive Value) μπορεί σίγουρα να πει το 79,5% της πιθανότητας ότι ο ασθενής δεν έχει τη νόσο.

Έπειτα, έγινε αναπαράσταση της καμπύλης ROC για να απεικονιστεί η απόδοση του μοντέλου του αφελή Μπεϋζιανού στο σύνολο δεδομένων εκπαίδευσης και για το σύνολο δεδομένων ελέγχου αντίστοιχα.



Εικόνα 4.21 Αφελής Μπεϋζιανός – Καμπύλη ROC για το train set

Από το παραπάνω γράφημα παίρνουμε τις ακόλουθες πληροφορίες, που αφορούν το σύνολο δεδομένων εκπαίδευσης του αφελή Μπεϋζιανού. Ο άξονας των x (False Positive Rate – FPR) δείχνει το ποσοστό των αρνητικών δειγμάτων που ταξινομήθηκαν λανθασμένα ως θετικά. Ο άξονας των y (True Positive Rate – TPR) δείχνει το ποσοστό των θετικών δειγμάτων που ταξινομήθηκαν σωστά. η κόκκινη γραμμή αντιπροσωπεύει την τυχαία κατάταξη – δηλαδή έναν ταξινομητή χωρίς διακριτική ικανότητα. Η μπλε γραμμή αντιπροσωπεύει την απόδοση του ταξινομητή στις διάφορες τιμές κατωφλίων. Εδώ η τιμή του AUC = 0.86 σημαίνει ότι υπάρχει 86% πιθανότητα ο ταξινομητής να κατατάξει ένα θετικό δείγμα πάνω από ένα αρνητικό. Η απόδοση αυτή θεωρείται πολύ καλή (μια τιμή από 0.8 και πάνω θεωρείται καλή). Ο ταξινομητής έχει πολύ καλή απόδοση στο σύνολο δεδομένων εκπαίδευσης. Υπάρχει καλή ισορροπία μεταξύ των TPR και των FPR σε μικρές τιμές κατωφλίου.



Εικόνα 4.22 Αφελής Μπεϋζιανός – Καμπύλη ROC για το test set

Στην παραπάνω εικόνα βλέπουμε την καμπύλη ROC για το σύνολο δεδομένων ελέγχου του αφελή Μπεϋζιανού. Η καμπύλη ROC (μπλε γραμμή) δείχνει τη σχέση ανάμεσα στο True Positive Rate (TPR) και του False Positive Rate (FPR) για διαφορετικές τιμές κατωφλίου. Η κόκκινη γραμμή αντιπροσωπεύει την τυχαία κατάταξη. Η απόδοση του ταξινομητή AUC = 0.87. Η καμπύλη ROC σημειώνει έντονη καμπυλότητα προς το επάνω αριστερό μέρος των αξόνων, το οποίο σημαίνει πολύ υψηλό TPR ~ 0.83 με πολύ χαμηλό FPR ~ (0.05 - 0.1). ο ταξινομητής μπορεί να κάνει αποτελεσματική διάκριση των θετικών από τα αρνητικά δείγματα. Η τιμή του AUC = 0.87 μεταφράζεται ως 87% πιθανότητα ένα τυχαίο δείγμα να ταξινομηθεί σωστά πάνω από ένα αρνητικό. Πρόκειται για μια εξαιρετικά καλή απόδοση. Η απόδοση των δεδομένων ελέγχου είναι λίγο καλύτερη και δείχνει ότι ο ταξινομητής δεν κάνει overfitting στο σύνολο δεδομένων εκπαίδευσης. Η χρήση του αφελή Μπεϋζιανού στο συγκεκριμένο πρόβλημα δείχνει να είναι πολύ αποτελεσματική.

5. Συμπεράσματα

Ολοκληρώνοντας τις απαραίτητες διαδικασίες με κάθε ταξινομητή, μπορούμε να κανουμε μια σύνοψη των αποτελεσμάτων που μας έδωσε κάθε ένας ξεχωριστά και να βγάλουμε συμπερασμάτα από την σύγκριση των μοντέλων των ταξινομητών. Ο πίνακας σύγχυσης και η αναφορά των αποτελεσμάτων για τα δεδομένα ελέγχου του μοντέλου της λογιστικής παλινδόμησης φαίνονται στην παρακάτω εικόνα:

```
Predicted   0    1  All
Actual
0          41   36   77
1          39   68  107
All        80  104  184

Classification report of LR  Confusion matrix - Test set:
              precision    recall  f1-score   support

           0       0.51      0.53      0.52        77
           1       0.65      0.64      0.64       107

    accuracy                           0.59       184
   macro avg       0.58      0.58      0.58       184
weighted avg       0.59      0.59      0.59       184
```

Εικόνα 5.1 Λογιστική Παλινδρόμηση – τελική αναφορά

Από την παραπάνω εικόνα προκύπτει ο πίνακας σύγχυσης (confusion matrix) με τις προβλεπόμενες και πραγματικές τιμές, όπου 41 ασθενείς προβλέφθηκαν ότι είναι πιθανό να πάθουν καρδιακή προσβολή και η πρόβλεψη είναι ΣΩΣΤΗ (TP). 68 ασθενείς προβλέφθηκαν ότι δεν θα έχουν καρδιακές παθήσεις και η πρόβλεψη είναι ΣΩΣΤΗ (TN). 36 ασθενείς προβλέφθηκαν ότι θα έχουν καρδιακές παθήσεις και είναι πιθανό να πάθουν έμφραγμα, αλλά η πρόβλεψη είναι ΛΑΘΟΣ (FP). 39 ασθενείς προβλέφθηκαν ότι δεν θα υποφέρουν από καρδιακές παθήσεις, όμως η πρόβλεψη είναι ΛΑΘΟΣ (FN).

Το μοντέλο της Λογιστικής Παλινδρόμησης παρουσιάζει μέτρια απόδοση με συνολική ορθότητα (accuracy) 59%, υποδεικνύοντας ότι κάνει σωστές προβλέψεις σε λίγο παραπάνω από τις μισές περιπτώσεις. Η κλάση 1 αποδίδει καλύτερα, με F1-score 0.64, ακρίβεια (precision) 0.65 και ανάκληση (recall) 0.64, σε αντίθεση με την κλάση 0 που έχει F1-score 0.52, ακρίβεια 0.51 και ανάκληση 0.53. Αυτές οι τιμές δείχνουν ότι το μοντέλο έχει την τάση να προβλέπει καλύτερα τις περιπτώσεις της κλάσης 1, ενώ υστερεί στην αναγνώριση της κλάσης 0. Οι μέσοι όροι macro και weighted επιβεβαιώνουν τη γενική μέτρια ισορροπία απόδοσης του μοντέλου. Ο μέσος όρος macro avg υπολογίζει το μέσο όρο των F1-score για όλες τις κλάσεις ταξινόμησης [38] και ο μέσος όρος weighted avg είναι ο σταθμισμένος μέσος όρος του F1-score, ο οποίος υπολογίζεται λαμβάνοντας τον μέσο όρο όλων των F1-scores ανά κατηγορία και λαμβάνοντας υπόψη τον δείκτη support κάθε κατηγορίας [38]. Συνολικά, το μοντέλο

δεν είναι ιδιαίτερα αποδοτικό και ενδέχεται να χρειάζεται βελτίωση μέσω τροποποίησης χαρακτηριστικών, ρύθμισης παραμέτρων ή επιλογής διαφορετικού αλγορίθμου.

Όμοια για τις Μηχανές Υποστήριξης Διανυσμάτων έχουμε τα αποτελέσματα του πίνακα σύγχυσης για τα δεδομένα ελέγχου όπως παρουσιάζονται στην ακόλουθη εικόνα:

```
Predicted   0   1  All
Actual
0          56  21   77
1          32  75  107
All        88  96  184

Classification report of SVM Confusion matrix - Test set:
              precision    recall  f1-score   support

           0       0.64      0.73      0.68        77
           1       0.78      0.70      0.74       107

    accuracy                           0.71       184
   macro avg       0.71      0.71      0.71       184
weighted avg       0.72      0.71      0.71       184
```

Εικόνα 5.2 Μηχανές Υποστήριξης Διανυσμάτων – τελική αναφορά

Όπως προηγουμένως, με τον ίδιο τρόπο λαμβάνουμε συμπεράσματα για τον ταξινομητή των Μηχανών Υποστήριξης Διανυσμάτων. Το μοντέλο εμφανίζει σημαντικά καλύτερη απόδοση από το μοντέλο της Λογιστικής Παλινδρόμησης, με συνολική ακρίβεια 71% έναντι 59%. Συγκεκριμένα, καταφέρνει να εντοπίσει πιο αποτελεσματικά και ισορροπημένα και τις δύο κατηγορίες, με F1-score 0.68 για την κλάση 0 και 0.74 για την κλάση 1, ενώ η ακρίβεια (precision) και η ανάκληση (recall) είναι επίσης υψηλότερες σε σχέση με τη Λογιστική Παλινδρόμηση, καθώς προβλέπει σωστά το 73% των περιπτώσεων της κλάσης 0 και το 70% της κλάσης 1, δείχνοντας πιο σταθερή και αξιόπιστη συμπεριφορά. Συνολικά, η αξιολόγηση δείχνει ότι το μοντέλο είναι καλύτερα προσαρμοσμένο στα δεδομένα και αποτελεί πιο κατάλληλη επιλογή ταξινόμησης για το συγκεκριμένο πρόβλημα.

Τέλος, για τον αφελή Μπεϋζιανό η πλήρης αναφορά του πίνακα σύγχυσης για τα δεδομένα ελέγχου παρουσιάζεται στην ακόλουθη εικόνα:

```
Predicted   0   1  All
Actual
0          70   7   77
1          18  89  107
All        88  96  184

Classification report of NB Confusion matrix - Test set:
              precision    recall  f1-score   support

           0       0.80      0.91      0.85        77
           1       0.93      0.83      0.88       107

    accuracy                           0.86       184
   macro avg       0.86      0.87      0.86       184
weighted avg       0.87      0.86      0.86       184
```

Εικόνα 5.3 Αφελής Μπεϋζιανός - τελική αναφορά

Το μοντέλο εμφανίζει πολύ καλή απόδοση με συνολική ακρίβεια 86%, σημαντικά υψηλότερη σε σύγκριση με τα δύο προηγούμενα μοντέλα. Καταγράφει υψηλές τιμές ακρίβειας (precision), ανάκλησης (recall) και F1-score και για τις δύο κατηγορίες. Με F1-score 0.85 για την κλάση 0 και 0.88 για την κλάση 1, υποδεικνύει ότι είναι ικανό να ταξινομεί σωστά και τις δύο κλάσεις με σχετική ισορροπία. Ειδικά στην κλάση 1, το μοντέλο πετυχαίνει ακρίβεια (precision) 0.93, δείχνοντας μεγάλη ακρίβεια στις θετικές προβλέψεις. Οι υψηλές τιμές μέσου όρου macro και weighted 0.86 και 0.87 αντίστοιχα, ενισχύουν την εικόνα της συνολικής αποτελεσματικότητας και σταθερότητας του μοντέλου [38]. Συμπερασματικά, το μοντέλου του αφελή Μπεϋζιανού είναι το καλύτερο από τα μοντέλα που εξετάστηκαν, προσφέροντας αξιόπιστη και ισορροπημένη απόδοση στο συγκεκριμένο πρόβλημα ταξινόμησης.

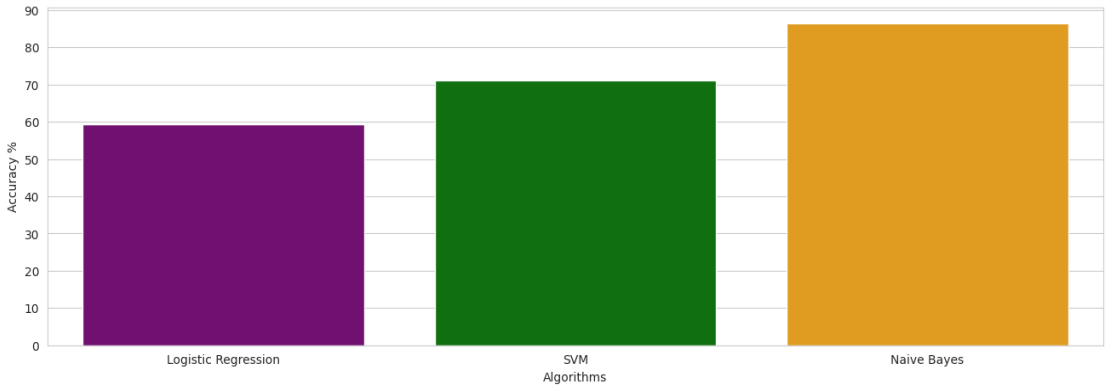
5.1 Σύγκριση αποτελεσμάτων αλγορίθμων ταξινόμησης

Για όλους τους πίνακες σύγχυσης, όπως απεικονίζεται και στις παραπάνω εικόνες για όλα τα μοντέλα, οι πραγματικές τιμές του συνόλου δοκιμών είναι: 77 ασθενείς που δεν έχουν καρδιακή νόσο (0) και 107 ασθενείς που έχουν καρδιακή νόσο (1). Συγκεκριμένα για την Λογιστική Παλινδρόμηση προκύπτει ότι ο αριθμός των ασθενών που προβλέφθηκε ότι δεν έχουν καρδιακή νόσο είναι 41/77 και ο αριθμός των ασθενών που προβλέφθηκε ότι έχουν καρδιακή νόσο είναι 68/107. Όμοια για τις Μηχανές Υποστήριξης Διανυσμάτων ο αριθμός των ασθενών που προβλέφθηκε ότι δεν είχαν καρδιακή νόσο είναι 56/77 και ο αριθμός των ασθενών που προβλέφθηκε ότι είχαν καρδιακή νόσο είναι 75/107. Για τον αφελή Μπεϋζιανό, ο αριθμός των ασθενών

που προβλέφθηκε ότι δεν είχαν καρδιακή νόσο είναι 70/77 και ο αριθμός των ασθενών που προβλέφθηκε ότι έχουν καρδιακή νόσο είναι 89/107.

Η ακρίβεια (precision) είναι η μέτρηση των ασθενών που αναγνωρίζονται σωστά ως πάσχοντες από καρδιακή νόσο σε σχέση με όλους τους ασθενείς που την έχουν πραγματικά. Ο αφελής Μπεϋζιανός επιτυγχάνει την υψηλότερη ακρίβεια (93%), ο ταξινομητής Μηχανών Υποστήριξης Διανυσμάτων ακολουθεί στην δεύτερη θέση με ακρίβεια (78%) και το μοντέλο της λογιστικής παλινδρόμησης δίνει την μικρότερη ακρίβεια (65%). Η ανάκληση (recall) είναι η μέτρηση για τον ορθό εντοπισμό ενός ασθενούς που έχει πράγματι καρδιακή νόσο και για τον αφελή Μπεϋζιανό η τιμή της είναι η υψηλότερη (83%), για τις Μηχανές Υποστήριξης Διανυσμάτων η τιμή της ανάκλησης είναι αρκετά υψηλή (70%), ενώ για τη λογιστική παλινδρόμηση είναι λίγο χαμηλότερη (64%). Για το σύνολο δεδομένων που αφορούν την καρδιά, η επίτευξη υψηλής ανάκλησης (recall) είναι πιο σημαντική από την επίτευξη υψηλής ακρίβειας (precision), καθώς είναι ζωτικής σημασίας να εντοπιστούν όσο το δυνατόν περισσότεροι ασθενείς που πάσχουν από καρδιακές παθήσεις. Όμως, η υψηλή ακρίβεια είναι επίσης απαραίτητη όταν πρόκειται για προβλήματα υγείας, καθώς ένας ασθενής μπορεί να πάσχει από καρδιακή νόσο και ταυτόχρονα από κάποια άλλη πάθηση. Έτσι, ένα υψηλό F1-score θα υποδηλώνει τόσο καλή ακρίβεια (precision) όσο και ανάκληση (recall). Ο αφελής Μπεϋζιανός επιτυγχάνει το υψηλότερο F1-score με ποσοστό 88%, στη συνέχεια ο ταξινομητής Μηχανών Υποστήριξης Διανυσμάτων έχει ποσοστό 74% και τελευταίο το μοντέλο Λογιστικής Παλινδρόμησης με ποσοστό 64%.

Επίσης, είναι σημαντικό να αναπαραστήσουμε και την ακρίβεια (accuracy) των τριών μοντέλων για τα δεδομένα των δεδομένων ελέγχου (test set).



Εικόνα 5.4 Σύγκριση της ακρίβειας των ταξινομητών

Παρά το γεγονός ότι η χρησιμοποιήθηκε η ίδια αναλογία στα σύνολα εκπαίδευσης (train set) και δοκιμής (test set) με αναλογία 80% - 20% αντίστοιχα και για τους τρεις ταξινομητές, είναι προφανές ότι οι τρεις αλγόριθμοι συμπεριφέρονται διαφορετικά και παράγουν διαφορετικά αποτελέσματα ακρίβειας. Ο αφελής

Μπεϋζιανός δίνει την καλύτερη απόδοση και πέτυχε καλύτερα αποτελέσματα σχεδόν για όλες τις μετρικές. Έχει τη μεγαλύτερη ακρίβεια σε σχέση με τους άλλους δύο ταξινομητές για το σύνολο των δοκιμών (test set) σε ποσοστό 87.04333050127443%. Οι Μηχανές Υποστήριξης Διανυσμάτων είχαν ακρίβεια σε ποσοστό 71.41036533559898%, ενώ το μοντέλο της Λογιστικής Παλινδρόμησης είχε την χειρότερη απόδοση με ακρίβεια 58.39907755795607%. Παρόλο που ο αφελής Μπεϋζιανός έχει καλύτερες επιδόσεις από τους άλλους στο συγκεκριμένο σύνολο δεδομένων, υπάρχουν και άλλοι παράμετροι που ενδεχομένως επηρεάζουν και μεταβάλλουν τις επιδόσεις των μοντέλων. Τέτοιες παράμετροι και στρατηγικές κανονικοποίησης, όπως για παράδειγμα η παράμετρος C softening στο μοντέλο Μηχανών Υποστήριξης Διανυσμάτων, μπορούν να ρυθμιστούν με διαφορετικούς τρόπους σε κάθε έναν από αυτούς τους αλγορίθμους, προκειμένου να επιτευχθεί καλύτερη ακρίβεια. Το μοντέλο Μηχανών Υποστήριξης Διανυσμάτων επεκτάθηκε με τη χρήση γραμμικών πυρήνων και πυρήνων συναρτήσεων ακτινικής βάσης και βρέθηκαν οι βέλτιστες παράμετροι, μετατρέποντας το σύνολο δεδομένων σε έναν χώρο χαρακτηριστικών, ο οποίος θα μπορούσε να βελτιωθεί ίσως με περισσότερα δεδομένα εκπαίδευσης.

6. Προτάσεις βελτίωσης

Ακολουθούν μερικές προτάσεις βελτίωσης για περαιτέρω εξερεύνηση των δεδομένων. Αρχικά, ας λάβουμε υπόψη ότι το σύνολο δεδομένων με το οποίο πειραματιστήκαμε δεν είναι πολύ μεγάλο. Τα συμπεράσματα και η επιλογή των αλγορίθμων ταξινόμησης έγιναν με βάση αυτό το σύνολο δεδομένων. Πριν λοιπόν προχωρήσουμε στην υλοποίηση του μοντέλου μας, πρέπει να το επικυρώσουμε με σύνολα δεδομένων που έχουν περισσότερα δεδομένα.

Μία άλλη σκέψη είναι στο υπάρχον σύνολο δεδομένων να αλλάξουμε το ποσοστό εκπαίδευσης και ελέγχου σε 70% - 30% αντίστοιχα. Θα ακολουθήσουμε την ίδια διαδικασία, επιλέγοντας τους ίδιους αλγορίθμους, μετρικές απόδοσης και αξιολόγησης και στο τέλος θα λάβουμε τα αποτελέσματα. Με αυτόν τον τρόπο, θα έχουμε τη δυνατότητα και να κάνουμε σύγκριση των αποτελεσμάτων των αλγορίθμων για το ποσοστό εκπαίδευσης και ελέγχου 70%-30% αλλά και να πραγματοποιήσουμε σύγκριση των τελικών αποτελεσμάτων που θα προκύψουν για τα σύνολα 80%-20% και 70%-30% αντίστοιχα.

Επίσης, η εφαρμογή σε σύνολο πραγματικών ιατρικών δεδομένων είναι ίσως και η πιο αξιόπιστη επιλογή που μπορούμε να κάνουμε και να προχωρήσουμε στην συνέχεια σε αφαίρεση δεδομένων που παρουσιάζουν μικρή συσχέτιση μεταξύ τους ώστε να μην επηρεαστούν τα υπόλοιπα δεδομένά μας.

Τέλος, μπορούμε να εφαρμόσουμε και άλλους αλγορίθμους ταξινόμησης στα δεδομένα μας και να δούμε πώς ανταποκρίνονται στο πρόβλημα που θέλουμε να επιλύσουμε. Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε τους Κ-πλησιέστερους γείτονες (KNN) και τα Δέντρα Απόφασης (Decision Trees).

Βιβλιογραφία

1. Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

2. http://repfiles.kallipos.gr/html_books/93/00eintroduction.html#_idTextAnchor000 / Κεντρική Ομάδα Υποστήριξης. (2022). *Εισαγωγή* [Κεφάλαιο 1]. Στο *Οδηγός για συγγραφείς ΧΞLATEX* (σελ. 1–10). Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://repfiles.kallipos.gr/html_books/93/00eintroduction.html#_idTextAnchor000

3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

4. Garson, G. D. (2006). *Statnotes: Topics in multivariate analysis*. Retrieved December 4, 2006, from https://www.statnotes.com/

5. McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

6. Κόκκινος, Γ. (2011). *Παράλληλοι αλγόριθμοι εξόρυξης γνώσης από βάσεις δεδομένων με τεχνητά νευρωνικά δίκτυα και μηχανές διανυσμάτων υποστήριξης* [Μεταπτυχιακή διπλωματική εργασία, Πανεπιστήμιο Μακεδονίας]. http://dspace.lib.uom.gr/bitstream/2159/14399/1/KokkinosIoannisMsc2011.pdf

7. Βαζιργιάννης, Μ., & Χαλκίδη, Μ. (χ.χ.). *Εξόρυξη γνώσης από βάσεις δεδομένων και τον Παγκόσμιο Ιστό*. Εκδόσεις Gutenberg.

8. Ντάλλα, Μ. (2009). *Εφαρμογή αλγορίθμων επαγωγικού λογικού προγραμματισμού στη σχεσιακή εξόρυξη δεδομένων* [Μεταπτυχιακή διπλωματική εργασία, Πανεπιστήμιο Πατρών]. http://nemertes.lis.upatras.gr/jspui/bitstream/10889/2656/1/Nimertis_Dalla.pdf

9. Κεχαγιά-Παρδάλη, Ε. (2006). *Αλγόριθμοι εξόρυξης χωρικών δεδομένων: Εφαρμογή σε αλγόριθμους συσταδοποίησης* [Διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο]. http://www.dbnet.ece.ntua.gr/~kpatro/theses/geodb/Kehagia_thesis.pdf

10. Ξενή, Μ. (χ.χ.). *Λογιστική παλινδρόμηση & διαχωριστική ανάλυση* [Διπλωματική εργασία, Πανεπιστήμιο Πατρών]. http://nemertes.lis.upatras.gr/jspui/bitstream/10889/5174/1/Diplwmatiki.pdf

11. Understanding logistic regression. (n.d.). *GeeksforGeeks*. Retrieved June 2, 2025, from https://www.geeksforgeeks.org/understanding-logistic-regression/?ref=header_outind

12. Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

13. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

15. Lee, W.-M. (2019). *Python machine learning*. John Wiley & Sons.

16. Tan, A. C., & Gilbert, D. (2003). Machine learning and its application to bioinformatics: An overview. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btg005

17. Javatpoint. (2021). *Machine learning tutorial*. Retrieved June 2, 2025, from https://www.javatpoint.com/machine-learning

18. W.-M. Lee, Python machine learning. John Wiley \& Sons, 2019.

19. Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov.

20. Murel, J., & Kavlakoglu, E. (2024, January 19). *What is a confusion matrix?* IBM. https://www.ibm.com/think/topics/confusion-matrix

21. Google Developers. (2024, October 16). *Classification: ROC and AUC*. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

22. NumPy Developers. (n.d.). *NumPy: The fundamental package for scientific computing with Python*. Retrieved June 2, 2025, from https://numpy.org/

23. The pandas development team. (n.d.). *pandas: Python data analysis library*. Retrieved June 2, 2025, from https://pandas.pydata.org/

24. Matplotlib Development Team. (n.d.). *Matplotlib: Visualization with Python*. Retrieved June 2, 2025, from https://matplotlib.org/

25. The MathWorks, Inc. (n.d.). *MATLAB: The language of technical computing*. Retrieved June 2, 2025, from https://www.mathworks.com/products/matlab.html

26. Langley, P., & Carbonell, J. G. (1987). *Machine learning: Techniques and foundations*. UC Irvine: Donald Bren School of Information and Computer Sciences. Retrieved from https://escholarship.org/uc/item/8160p39f

27. Marion Neumann. 2019. AI profiles: an interview with Thomas Dietterich. May 2019, Τόμ. 5, σσ. 7-9. / Neumann, M. (2019). AI profiles: An interview with Thomas Dietterich. *[Όνομα περιοδικού ή συλλογής]*, 5, 7–9.

28. Sah, S. (2020). Machine learning: A review of learning types. *Preprints*. https://doi.org

29. Kubat, M. (έτος). *Εισαγωγή στη μηχανική μάθηση* (Κεφάλαιο 2, Ενότητες 2.1–2.4)

30. Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristic curves. *Canadian Journal of Psychiatry,* 52(2), 121–128. https://doi.org/10.1177/070674370705200206

31. Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography, 17*(2), 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

32. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

33. NETZSCH Analyzing & Testing. (2021, Δεκέμβριος 3). Από τα έξυπνα δεδομένα στην τεχνητή νοημοσύνη. https://analyzing-testing.netzsch.com/el/blog/2021/apo-ta-exupna-dedomena-sten-tekhnete-noemosune

34. VectorMine. (2022, Οκτώβριος 30). *Τύποι μηχανικής μάθησης με διάγραμμα περιγράμματος ταξινόμησης αλγορίθμων. Σηματοδοτημένο*

εκπαιδευτικό σύστημα με εποπτευόμενες, χωρίς επίβλεψη και ενίσχυση μεθόδους τεχνητής νοημοσύνης διανυσματική απεικόνιση. Depositphotos. https://depositphotos.com/gr/vector/types-machine-learning-algorithms-classification-outline-diagram-labeled-educational-scheme-618897540.html

35. Καμπύλη λειτουργικού χαρακτηριστικού δέκτη (ROC curve). (2025). Στην *Μεθοδολογία Έρευνας και Στατιστική Ανάλυση με Χρήση Ανοικτού Λογισμικού* (σελ. 1–10). Δημοκρίτειο Πανεπιστήμιο Θράκης. https://utopia.duth.gr/epdiaman/files/kedivim/section7/ROC%20Curve.pdf

36. Simplilearn. (2025, Μάρτιος 26). *Τι είναι ο πίνακας σύγχυσης στη μηχανική μάθηση;* https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning

37. Κομπούτος, Ν. (2025). *Συσχέτιση* [Έγγραφο μαθημάτων, Δημοκρίτειο Πανεπιστήμιο Θράκης]. https://eclass.uth.gr/modules/document/file.php/PE_P_170/Correlation.pdf

38. Leung, K. (2023, Ιανουάριος 4). *Micro, macro & weighted averages of F1 score, clearly explained*. KDnuggets. https://www.kdnuggets.com/2023/01/micro-macro-weighted-averages-f1-score-clearly-explained.html

39. Βικιπαίδεια. (2023, Ιανουάριος 29). *Σιγμοειδής συνάρτηση*. https://el.wikipedia.org/wiki/Σιγμοειδής_συνάρτηση

40. Author Unknown. (n.d.). *Support Vector Machine with Scikit-Learn: A Friendly Introduction*. Towards Data Science. https://towardsdatascience.com/support-vector-machine-with-scikit-learn-a-friendly-introduction-a2969f2ff00d/