

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Πτυχιακή Εργασία

Ανάπτυξη Μοντέλων Πρόβλεψης στην Καρδιακή Ανεπάρκεια

Αδάμ Στυλιανή
Π2020160

10 Ιουνίου 2024

Τριμελής Επιτροπή

Έξαρχος Θεμιστοκλής
Αναπληρωτής Καθηγητής
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Βραχάτης Αριστείδης
Επίκουρος Καθηγητής
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Κερμανίδου Κάτια - Λήδα
Αναπληρώτρια Καθηγήτρια
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Περιεχόμενα

Περιεχόμενα	2
Περίληψη	3
1. Εισαγωγή	4
1.1 Κυκλοφορικό Σύστημα	4
1.2 Καρδιά	4
1.3 Καρδιακή Ανεπάρκεια	6
1.4 Γιατί μελετάμε την Καρδιακή Ανεπάρκεια	8
2. Βιβλιογραφική Ανασκόπηση	9
2.1 Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction[6]	9
2.2 Artificial intelligence in the diagnosis and detection of heart failure: the past, present, and future[7]	11
2.3 Prevalence, outcomes and costs of a contemporary, multinational population with heart failure[8]	12
2.4 Development of heart failure risk prediction models based on a multi-marker approach using random forest algorithms[9]	13
2.5 Development and validation of prediction models for incident atrial fibrillation in heart failure[10]	14
2.6 Heart failure mortality prediction using PRISM score and development of a classification and regression tree model to refer patients for palliative care consultation[11]	15
2.7 Heart Failure Development Prediction using Stochastic Gradient Descent Optimization[12]	16
2.8 Prediction of patients with heart failure after myocardial infarction[13]	17
2.9 Heart Failure Prediction by Feature Ranking Analysis in Machine Learning[14]	18
2.10 Derivation and validation of a machine learning-based risk prediction model in patients with acute heart failure[15]	19
3. Περιγραφή Δεδομένων	20
4. Μεθοδολογία	26
4.1 Decision Trees	26
4.2 Random Forest	27
4.3 Gradient Boosting	28
4.4 XGBoost (Extreme Gradient Boosting)	29
4.5 LightGBM	30
4.6 AdaBoost	31
4.7 CatBoost	31
4.8 Extra Tree	31
4.9 RusBoost	32
4.10 Πειράματα	32
5. Αποτελέσματα	41
6. Εφαρμογή	50
7. Συμπεράσματα	55
Παράρτημα	56
Βιβλιογραφία	63

Περίληψη

Η παρούσα πτυχιακή εργασία εστιάζει στην ανάπτυξη και αξιολόγηση μοντέλων πρόβλεψης για την καρδιακή ανεπάρκεια χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Μέσα από τη μελέτη των χαρακτηριστικών των ασθενών και τη χρήση αλγορίθμων όπως Random Forest, Gradient Boosting και άλλους, επιτεύχθηκαν υψηλής ακρίβειας προβλέψεις. Τα αποτελέσματα της έρευνας δείχνουν ότι τα μοντέλα μηχανικής μάθησης μπορούν να προσφέρουν αξιόπιστα εργαλεία για την έγκαιρη διάγνωση και πρόληψη της καρδιακής ανεπάρκειας. Η τελική εφαρμογή, που αναπτύχθηκε σε Django, αποδείχθηκε αποτελεσματική σε πραγματικά δεδομένα, προσφέροντας έτσι μια σημαντική βοήθεια στους επαγγελματίες υγείας. Παρά τους περιορισμούς της μελέτης, τα ευρήματα υπογραμμίζουν τη δυναμική της μηχανικής μάθησης στην ιατρική πρόβλεψη και προτείνουν κατευθύνσεις για μελλοντική έρευνα.

1. Εισαγωγή

Η καρδιακή ανεπάρκεια αποτελεί μια από τις σημαντικότερες αιτίες θνησιμότητας και νοσηρότητας παγκοσμίως. Η έγκαιρη διάγνωση και παρέμβαση είναι κρίσιμη για τη βελτίωση της ποιότητας ζωής και την επιβίωση των ασθενών. Με την ανάπτυξη της μηχανικής μάθησης, προκύπτουν νέες δυνατότητες για την ανάλυση μεγάλων συνόλων δεδομένων και την πρόβλεψη ιατρικών καταστάσεων. Η παρούσα εργασία εστιάζει στη χρήση διαφόρων αλγορίθμων μηχανικής μάθησης για την ανάπτυξη μοντέλων πρόβλεψης της καρδιακής ανεπάρκειας. Η έρευνα περιλαμβάνει την ανάλυση των δεδομένων, την επιλογή των κατάλληλων αλγορίθμων, την εκπαίδευση και αξιολόγηση των μοντέλων, καθώς και την ανάπτυξη μιας εφαρμογής που θα χρησιμοποιεί το τελικό μοντέλο για πρόβλεψη σε πραγματικό χρόνο. Στόχοι λοιπόν της έρευνας είναι:

1. Να αναλυθούν τα χαρακτηριστικά των ασθενών με καρδιακή ανεπάρκεια
2. Να αναπτυχθούν και να αξιολογηθούν μοντέλα πρόβλεψης χρησιμοποιώντας διάφορους αλγορίθμους μηχανικής μάθησης
3. Να δημιουργηθεί μια εφαρμογή που θα χρησιμοποιεί το τελικό μοντέλο για την πρόβλεψη της καρδιακής ανεπάρκειας σε νέους ασθενείς

Με αυτές τις προκλήσεις και στόχους, η παρούσα εργασία στοχεύει να συμβάλει στη βελτίωση των μεθόδων πρόβλεψης και στη διάθεση εργαλείων υποστήριξης για τους επαγγελματίες υγείας.

1.1 Κυκλοφορικό Σύστημα

Ο ανθρώπινος οργανισμός απαρτίζεται από μια πληθώρα συστημάτων, τα οποία λειτουργούν αρμονικά προκειμένου να υπάρξει ζωή. Ένα από αυτά τα συστήματα είναι και το κυκλοφορικό με το οποίο μεταφέρονται οι θρεπτικές ουσίες στα κύτταρα των ιστών αλλά και απομακρύνονται. Αποτελείται από την καρδιά, τα αιμοφόρα αγγεία και το αίμα. Άρρηκτα συνδεδεμένο με το κυκλοφορικό σύστημα είναι και το λεμφικό, στο οποίο κυκλοφορεί η λέμφος.

Συνοπτικά, το λεμφικό σύστημα αποτελείται από τα λεμφαγγεία, τη λέμφο, τους λεμφαδένες, τον σπλήνα και τον θυμό αδένα. Οι τελευταίοι είναι δομές που αποτελούνται από μια αρκετά συγκεκριμένη και εξειδικευμένη μορφή συνδετικού ιστού, ο οποίος ονομάζεται λεμφικός ιστός όπου περιέχονται τα λεμφοκύτταρα. Η σπουδαιότητα του έγκειται στην διεργασία που εκτελεί παραλαμβάνοντας το πλεόνασμα του μεσοκυττάριου υγρού και επαναφέροντας το στο καρδιαγγειακό σύστημα. Ταυτόχρονα μεταφέρει λιπαρές ουσίες απ' το λεπτό έντερο στο αίμα και εν τέλει συμβάλλει στην άμυνα του οργανισμού καταστρέφοντας παθογόνους μικροοργανισμούς και καρκινικά κύτταρα.

1.2 Καρδιά

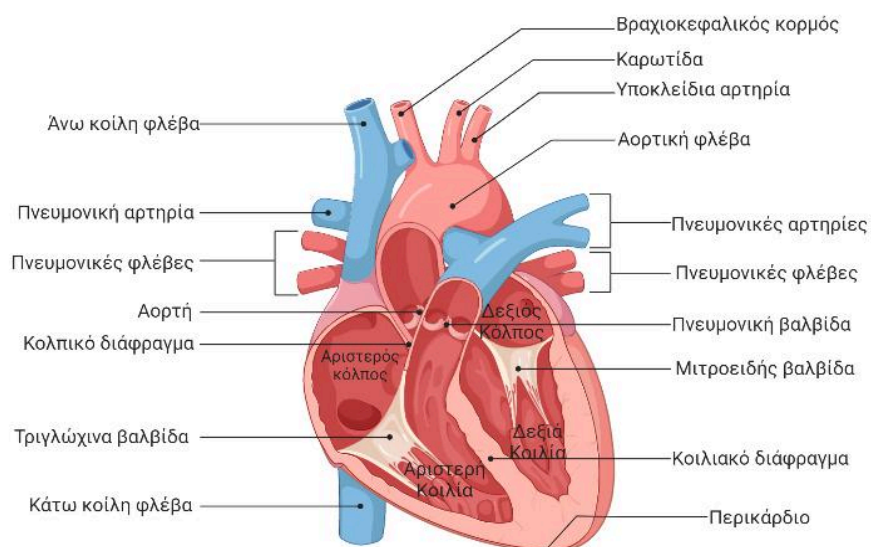
Το σημαντικότερο όργανο του κυκλοφορικού συστήματος είναι η καρδιά. Η λειτουργία της είναι κατανοητή και καίριας σημασίας. Εντοπίζεται ανάμεσα στους δύο πνεύμονες πίσω από το στήθος, έχει κωνικό σχήμα και είναι αποτελούμενη από μυϊκό ιστό το οποίο με την σειρά του, αποτελεί το μυοκάρδιο και έχει μέγεθος γροθιάς.

Η καρδιά είναι μια μυώδης αντλία και απαρτίζεται από τον καρδιακό μυ. Η κατάλληλη σύνδεση των μυϊκών ινών του μυοκαρδίου επιτρέπει την σύσπαση τους. Γενικότερα η καρδιά

είναι τετράχωρη. Αποτελείται από δύο κόλπους και δύο κοιλίες. Οι κόλποι βρίσκονται στο ανώτερο τμήμα της και οι κοιλίες στο κατώτερο. Οι κόλποι έχουν λεπτά τοιχώματα και ονομάζονται αριστερός και δεξιός κόλπος αντίστοιχα, ενώ οι κοιλίες έχουν παχύτερα τοιχώματα και τα ονόματά τους προκύπτουν από την θέση τους, δηλαδή, αριστερή και δεξιά κοιλία. Οι δύο κοιλίες διαχωρίζονται από το μεσοκοιλιακό διάφραγμα και οι κόλποι με το μεσοκολπικό διάφραγμα. Συνεπώς ούτε οι δύο κοιλίες ούτε οι δύο κόλποι επικοινωνούν μεταξύ τους. Η παραπάνω αναφορά στα τοιχώματα των κοιλιών, σχετίζεται με την λειτουργία που εκτελούν, δηλαδή η αριστερή κοιλία έχει παχύτερα τοιχώματα από την δεξιά, αφού στέλνει το αίμα σε μεγαλύτερη απόσταση (ακόμη και σε όλο το σώμα). Επιπλέον τόσο μεταξύ των κόλπων όσο και των κοιλιών υπάρχουν βαλβίδες, οι οποίες καθορίζουν την μονόδρομη ροή του αίματος σε κάθε σύσπαση της καρδιάς. Επομένως η καρδιά είναι τόσο αναρροφητική, αφού συγκεντρώνει το αίμα από τα τριχοειδή του σώματος μέσω των φλεβών, όσο και συμπιεστική δεδομένου ότι στέλνει το αίμα στα τριχοειδή όλου του σώματος μέσω των αρτηριών που ξεκινούν από τις κοιλίες της.

Η κίνηση του αίματος είναι από τους κόλπους προς τις κοιλίες, οι οποίες με την σειρά τους συστέλλονται και προωθούν το αίμα στις δύο αρτηρίες. Ο δεξιός κόλπος συγκεντρώνει το αίμα από την περιφέρεια του σώματος, περιέχοντας το διοξειδίο του άνθρακα, σε αντίθεση με τον αριστερό κόλπο στον οποίο το αίμα που φτάνει έχει ήδη περάσει από τους πνεύμονες και περιέχει οξυγόνο. Κατά την συστολή των κόλπων το αίμα διοχετεύεται στις κοιλίες, με τις βαλβίδες ανάμεσα σε κόλπο και κοιλία να κλείνουν για να αποφευχθεί η παλινδρόμηση, και εν τέλει να φτάνει στις αρτηρίες. Η αορτή εντοπίζεται στην αριστερή κοιλία και κινείται προς την περιφέρεια του σώματος ενώ η δεξιά κοιλία συνδέεται με την πνευμονική αρτηρία και καταλήγει στους πνεύμονες. Φυσικά τόσο στην αορτή όσο και στην πνευμονική αρτηρία υπάρχουν βαλβίδες που ρυθμίζουν την ροή του αίματος.

Για την ζωή ενός ανθρώπου η καρδιά χτυπάει συνεχώς, δημιουργώντας τους παλμούς. Οι παλμοί οφείλονται σε συστολές και χαλαρώσεις του μυοκαρδίου και ο φυσιολογικός τους αριθμός είναι περίπου 60-80 / λεπτό. Βέβαια η λειτουργία της μπορεί να επηρεαστεί από γεγονότα που συμβαίνουν στον οργανισμό όπως για παράδειγμα η σωματική άσκηση είναι ικανή να αυξήσει τους παλμούς. Σε κάθε περίπτωση η καρδιά θα χτυπήσει περίπου 100.000 φορές την μέρα και θα αντλήσει περίπου 13.000 λίτρα αίμα.



Εικόνα 1. Ανατομία Καρδιάς

1.3 Καρδιακή Ανεπάρκεια

Υπάρχουν πολλά και διαφορετικά προβλήματα καρδιάς. Κάποια είναι κοινά και συχνά εμφανιζόμενα και άλλα πιο σπάνια. Το καθένα έχει διαφορετικά χαρακτηριστικά, αίτια και τρόπους αντιμετώπισης.

Η καρδιακή ανεπάρκεια είναι μια συχνή και σοβαρή πάθηση η οποία μάλιστα χειροτερεύει με την πάροδο του χρόνου και δεν έχει ριζική θεραπεία. Πρόκειται για μια κατάσταση κατά την οποία η καρδιά δεν αντλεί το απαραίτητο αίμα για τις ανάγκες του σώματος. Συγκεκριμένα η καρδιά δεν μπορεί να ανταπεξέλθει αποτελεσματικά στην απαίτηση του οργανισμού για αίμα και οξυγόνο και έτσι δεν αντλεί αρκετό αίμα ή δεν το κάνει με αρκετή πίεση, οδηγώντας σε ανεπάρκεια της κυκλοφορίας του αίματος στο σώμα.

Πιο αναλυτικά, η καρδιά είναι αδύνατη και δεν μπορεί να στείλει αρκετή ποσότητα αίματος σε όλο το σώμα, με αποτέλεσμα το σώμα να μην παίρνει την ποσότητα του οξυγόνου και τις θρεπτικές ουσίες που χρειάζεται για να δουλέψει φυσιολογικά. Το αίμα και το οξυγόνο δεν επαρκούν για να θρέψουν τους μυς και τα όργανα του σώματος με αποτέλεσμα οι μύες να κουράζονται και ο ασθενής να εμφανίζει εύκολη κόπωση. Ταυτόχρονα δεν επιστρέφει αρκετό αίμα στην καρδιά, η οποία με την σειρά της, περιμένει ένα δυνατό τράβηγμα που δεν έρχεται. Τότε μαζεύεται ολόένα και περισσότερο αίμα στις φλέβες, περιμένοντας να γυρίσει στην καρδιά και έτσι αυξάνεται η πίεση μέσα σε αυτές, οπότε αρχίζουν να βγαίνουν υγρά στους γύρω ιστούς, κυρίως στα πόδια και την κοιλιά. Υγρό μαζεύεται και στους πνεύμονες και προκαλεί συμφόρηση και δύσπνοια. Στην κατάσταση αυτή το σώμα προσπαθεί να ανταπεξέλθει κάνοντας κάποιες προσαρμογές αλλά μόνο για λίγο χρονικό διάστημα τα καταφέρνει.

Παρατηρείται πως η καρδιακή ανεπάρκεια αναπτύσσεται επειδή ο/η ασθενής πάσχει από μια άλλη καρδιακή πάθηση όπως για παράδειγμα στεφανιαία νόσο, καρδιακή προσβολή, υπέρταση, βαλβιδική νόσος, αρρυθμία. Άλλες αιτίες περιλαμβάνουν προβλήματα με τις βαλβίδες της καρδιάς, καρδιομυοπάθειες (παθήσεις του μυός της καρδιάς), αρτηριοσκλήρυνση, ή αυξημένο φορτίο στην καρδιά λόγω κατανάλωσης αλκοόλ ή χρήσης ορισμένων φαρμάκων.

Συμπτώματα της καρδιακής ανεπάρκειας αποτελούν η κόπωση, το πρήξιμο στα πόδια και στην κοιλιά, η δύσπνοια, οι ζαλάδες, η απώλεια όρεξης, η αύξηση του σωματικού βάρους λόγω πρόσκρουσης υγρών.

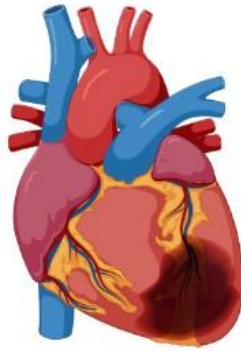
Εμφανίζεται και αναπτύσσεται σε οποιαδήποτε ηλικία αλλά εμφανώς είναι πιο συνήθης σε μεγαλύτερες ηλικίες. Περίπου το 1% των ατόμων κάτω των 65 ετών υποφέρει από καρδιακή ανεπάρκεια, αλλά το 7% των ατόμων ηλικίας 75-84 ετών υποφέρει από καρδιακή ανεπάρκεια και το ποσοστό αυξάνεται στο 15% σε άτομα άνω των 85 ετών. Είναι η συνηθέστερη αιτία νοσηλείας σε ασθενείς ηλικίας άνω των 65 ετών.

Όπως προαναφέρθηκε η καρδιακή ανεπάρκεια είναι μια συχνή και σοβαρή πάθηση, η οποία μάλιστα χειροτερεύει με την πάροδο του χρόνου. Ο ρυθμός με τον οποίο χειροτερεύει δεν είναι προβλέψιμος και διαφέρει από ασθενή σε ασθενή. Σε ορισμένους τα συμπτώματα παραμένουν σταθερά για μήνες ή χρόνια ενώ σε άλλους χειροτερεύουν προοδευτικά χρόνο με το χρόνο. Υπάρχουν και περιπτώσεις που ενώ διατηρούνται οι ασθενείς ξαφνικά χειροτερεύουν είτε γιατί εμφανίστηκε μια αρρυθμία, είτε μια λοίμωξη, είτε μια καρδιακή προσβολή (ένα έμφραγμα). Συνηθισμένη αιτία επιδείνωσης αποτελεί και η μη συμμόρφωση του ασθενούς στη θεραπευτική αγωγή ή στους διαιτητικούς περιορισμούς που έχει ορίσει ο γιατρός. Η καρδιακή ανεπάρκεια μπορεί να διαχειριστεί με φάρμακα, διατροφικές αλλαγές, ασκήσεις και σε ορισμένες περιπτώσεις, με χειρουργική επέμβαση.

Συχνός παράγοντας που συνδέεται με την καρδιακή ανεπάρκεια, όπως προαναφέρθηκε είναι η στεφανιαία νόσος. Πιο συγκεκριμένα η στεφανιαία νόσος προκύπτει όταν τα κεντρικά αιμοφόρα αγγεία, που τροφοδοτούν την καρδιά με αίμα και οξυγόνο υποστούν βλάβη. Η ροή του αίματος μειώνεται, αφού στα τοιχώματα των αρτηριών σχηματίζονται αθηρωματικές πλάκες. Ωστόσο μια ολική απόφραξη στην αρτηρία θα εμποδίσει την καρδιά να λάβει το απαραίτητο οξυγόνο με αποτέλεσμα να προκληθεί έμφραγμα του μυοκαρδίου.

Απ' τα πρώτα συμπτώματα της νόσου είναι η πίεση, ο πόνος, η δυσφορία και το σφίξιμο στο στήθος (συνήθως στην μέση και αριστερά του). Η δύσπνοια και η έντονη κόπωση εξίσου μπορούν να παρουσιαστούν ως προειδοποίηση, όπως και η ταχυκαρδία, η ναυτία και η ζαλάδα.

Με τον όρο “έμφραγμα του μυοκαρδίου” ορίζεται η νέκρωση μιας περιοχής του καρδιακού μυ που προκύπτει από την μειωμένη ροή του αίματος λόγω των αθηρωματικών πλακών. Η αντιμετώπιση της στεφανιαίας νόσου εξαρτάται από την σοβαρότητα του επεισοδίου αλλά και την έκταση της ζημιάς που έχει προκληθεί. Πάντως η καρδιά ως όργανο μπορεί να συνεχίσει να λειτουργεί ακόμη και αν ένα κομμάτι της έχει υποστεί ζημιές.



Εικόνα 2. Έμφραγμα του μυοκαρδίου



Εικόνα 3. Στένωση Αρτηρίας

1.4 Γιατί μελετάμε την Καρδιακή Ανεπάρκεια

Οι παθήσεις που σχετίζονται με την καρδιά είναι πολλές. Η καρδιακή ανεπάρκεια αποτελεί μια συχνά εμφανιζόμενη ασθένεια, πολύπλευρη και κατέχει υψηλά ποσοστά θνησιμότητας. Πρόσφατες στατιστικές δείχνουν ότι η καρδιακή ανεπάρκεια εξακολουθεί να αποτελεί σοβαρό πρόβλημα υγείας σε παγκόσμιο επίπεδο όντας υπεύθυνη ακόμα και για το 1/3 των θανάτων. Οι ασθενείς από καρδιαγγειακά νοσήματα αυξάνονται ραγδαία τα τελευταία χρόνια, με αποτέλεσμα εκατομμύρια άνθρωποι να πεθαίνουν. Οι προσπάθειες για την κατανόηση και αντιμετώπιση συνεχίζονται ενεργά μέσω επιστημονικών ερευνών και θεραπευτικών προσεγγίσεων.

Ενδεικτικά σε επίπεδο στατιστικών αποδείχθηκε πως η καρδιακή ανεπάρκεια επηρεάζει περισσότερους από 64 εκατομμύρια ανθρώπους παγκοσμίως το έτος 2023 ^{[1],[2]} ενώ μέχρι και το 2021^[3] παγκόσμια έπασχαν από αυτή 26 εκατομμύρια άνθρωποι. Πιο ειδικά στην Ευρώπη ετησίως έχουν αναφερθεί περίπου 4 εκατομμύρια κρούσματα καρδιακής ανεπάρκειας, ενώ στην Ελλάδα νοσούν 8.500 άνθρωποι^[4].

Από την άλλη πλευρά, ο Παγκόσμιος Οργανισμός Υγείας αναφέρει πως οι καρδιαγγειακές παθήσεις κοστίζουν την ζωή σε 17.9 εκατομμύρια ανθρώπους ανά έτος εκ των οποίων οι 1.5 εκατομμύρια θάνατοι προέρχονται από την Ευρώπη και στην Ελλάδα μέχρι το 2021 το 8% του πληθυσμού έπασχε από καρδιακή ανεπάρκεια. Στον όρο καρδιαγγειακές παθήσεις εντάσσονται η στεφανιαία νόσος, η εγκεφαλοαγγειακή νόσος, η ρευματική καρδιοπάθεια και πάρα πολλές ακόμη^[5].

Η μελέτη της καρδιακής ανεπάρκειας, λοιπόν, είναι κρίσιμη για πολλαπλούς λόγους και ονομαστικά θα μπορούσαμε να αναφέρουμε την υψηλή προδιάθεση και επίπτωση. Πιο αναλυτικά, η καρδιακή ανεπάρκεια είναι μία από τις πιο συχνές καρδιαγγειακές παθήσεις, που επηρεάζει εκατομμύρια ανθρώπους παγκοσμίως, όπως αναφέραμε. Ακόμα και στις ανεπτυγμένες κοινωνίες συχνά αντιμετωπίζουν αυξανόμενο πρόβλημα με την καρδιακή ανεπάρκεια, ενώ αυτή η πάθηση αντιπροσωπεύει σημαντικό ποσοστό των θανάτων σε παγκόσμιο επίπεδο. Παράλληλα οι καρδιαγγειακές παθήσεις επιφέρουν σημαντικό οικονομικό κόστος στα συστήματα υγείας και την κοινωνία γενικότερα. Η αντιμετώπιση της καρδιακής ανεπάρκειας απαιτεί σημαντικούς πόρους για τη θεραπεία, την έρευνα και την πρόληψη. Βέβαια δεν μπορεί να παραλειφθεί το γεγονός ότι η καρδιακή ανεπάρκεια συνδέεται συχνά με άλλες καρδιαγγειακές παθήσεις, όπως η στεφανιαία νόσος και η εγκεφαλοαγγειακή νόσος, καθιστώντας την ακόμα πιο σοβαρή και πολυπλοκότερη στη διαχείριση της και φυσικά μπορεί να έχει σημαντικές επιπτώσεις στην ποιότητα ζωής των ασθενών, περιορίζοντας την κινητικότητά τους, επηρεάζοντας την ανεξαρτησία τους και δυσχεραίνοντας τις καθημερινές τους δραστηριότητες. Έτσι η εξέλιξη της ιατρικής και η επιστημονική έρευνα επιτρέπουν την ανάπτυξη νέων θεραπευτικών προσεγγίσεων και τεχνολογιών για τη διαχείριση και την πρόληψη της καρδιακής ανεπάρκειας. Ταυτόχρονα η συνεχής έρευνα και μελέτη της καρδιακής ανεπάρκειας είναι κρίσιμη για τη βελτίωση της πρόληψης, της διάγνωσης και της θεραπείας της πάθησης, με στόχο τη μείωση του φορτίου της νόσου τόσο σε επίπεδο ατόμων όσο και κοινωνικών συστημάτων.

2. Βιβλιογραφική Ανασκόπηση

2.1 Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction^[6]

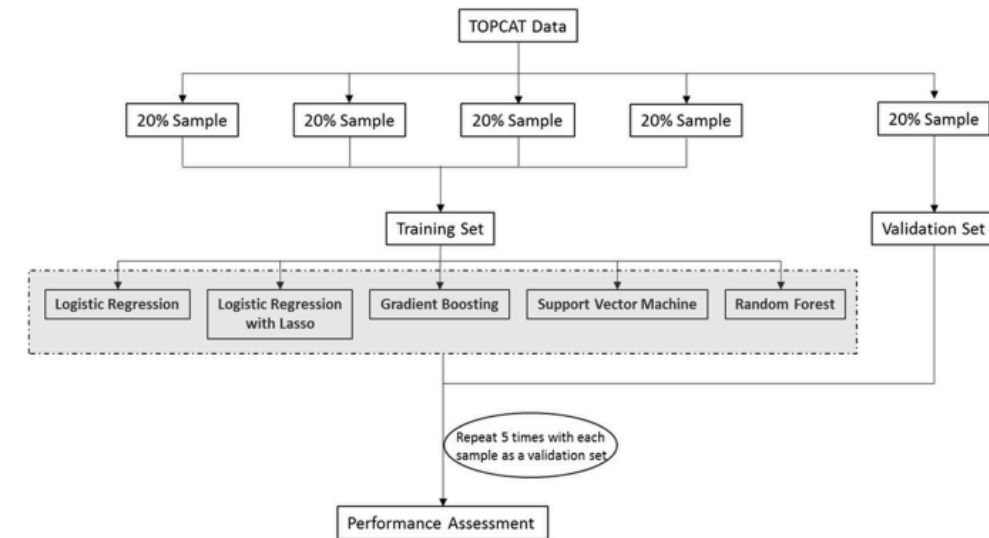
Μεγάλος αριθμός ανθρώπων παγκοσμίως παρουσιάζει καρδιακή ανεπάρκεια με τους μισούς εξ' αυτών να έχουν διατηρημένο το κλάσμα εξώθησης (HFpEF - Heart failure with preserved ejection fraction). Το κλάσμα εξώθησης είναι το ποσοστό του αίματος στον κύριο θάλαμο άντλησης που εξωθείται με κάθε παλμό. Υπάρχει ο όρος HFpEF και ο όρος HFrEF. Η διαφορά του δεύτερου όρου με τον πρώτο είναι η εξής HFpEF = καρδιακή ανεπάρκεια με διατηρημένο κλάσμα εξώθησης και HFrEF = καρδιακή ανεπάρκεια με μειωμένο κλάσμα εξώθησης.

Αυτή η μελέτη με τίτλο "Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction" στοχεύει στην ανάπτυξη μοντέλων για την πρόβλεψη της θνησιμότητας και της νοσηλείας σε ασθενείς με καρδιακή ανεπάρκεια με διατηρημένο κλάσμα εξώθησης (HFpEF) στη δοκιμή TOPCAT. Πρόκειται για μια δομική που διαδραματίστηκε στις ΗΠΑ και συμμετείχαν ασθενείς ηλικίας 50 ετών και άνω από διαφορετικές χώρες με καρδιακή ανεπάρκεια και κλάσμα εξώθησης αριστερής κοιλίας 45% ή υψηλότερο. Σε αυτή συμμετείχαν 1.767 ασθενείς με τους 387 ασθενείς (22%) να πεθαίνουν κατά τη διάρκεια της δοκιμαστικής παρακολούθησης ενώ οι 268 πέθαναν σε παρακολούθηση τριών ετών. Σε αυτή την παρακολούθηση, το 24,6% των ασθενών πέθαναν και το 31,5% νοσηλεύτηκε για έξαρση της καρδιακής ανεπάρκειας. Το ποσοστό επιβίωσης θνησιμότητας ήταν 0,92 σε ένα έτος και 0,82 σε παρακολούθηση δύο ετών, ενώ το ποσοστό επιβίωσης νοσηλείας με καρδιακή ανεπάρκεια ήταν 0,78 σε ένα έτος και 0,69 σε παρακολούθηση δύο ετών.

Συνολικά χρησιμοποιήθηκαν πέντε μέθοδοι: logistic regression with a forward selection of variables (λογιστική παλινδρόμηση), logistic regression with a lasso regularization for variable selection (λογιστική παλινδρόμηση με μπροστική επιλογή μεταβλητών), Random Forest (RF), Gradient Boosting και Support Vector Machine (μηχανή υποστήριξης διανυσμάτων). Τα μοντέλα εκπαιδεύτηκαν για τρία χρόνια και επικυρώθηκαν χρησιμοποιώντας 5-fold cross-validation (πενταπλάσια διασταυρούμενη επικύρωση). Η διάκριση και η βαθμονόμηση του μοντέλου εκτιμήθηκαν χρησιμοποιώντας χαρακτηριστικές καμπύλες λειτουργίας δέκτη και βαθμολογίες Brier. Οι κορυφαίες μεταβλητές πρόβλεψης αξιολογήθηκαν χρησιμοποιώντας τα μοντέλα με την καλύτερη απόδοση.

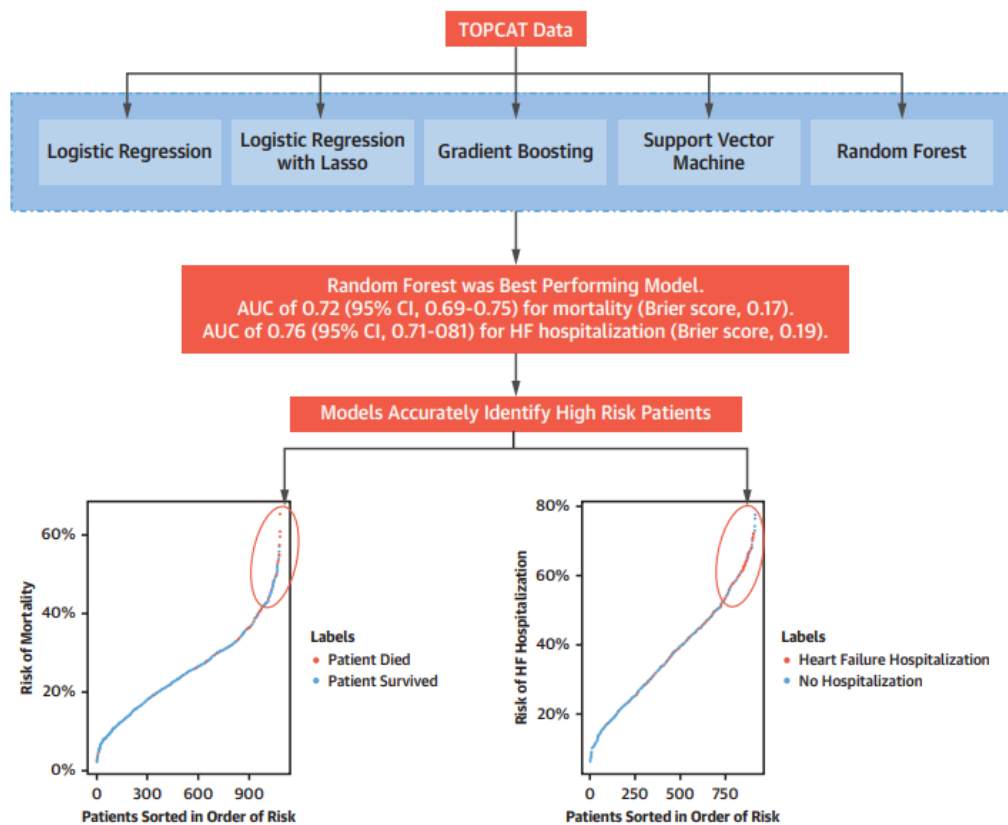
Το μοντέλο Random Forest (RF) είχε την καλύτερη απόδοση, με μέση στατιστική C 0,72 για την πρόβλεψη της θνησιμότητας και 0,76 για νοσηλεία HF. Τα επίπεδα αζώτου της ουρίας στο αίμα, ο δείκτης μάζας σώματος και οι βαθμολογίες υποκλίμακας KCCQ συσχετίστηκαν ισχυρά με τη θνησιμότητα, ενώ το επίπεδο αιμοσφαιρίνης, το άζωτο ουρίας στο αίμα, ο χρόνος από την προηγούμενη νοσηλεία της καρδιακής ανεπάρκειας και οι βαθμολογίες KCCQ ήταν οι πιο σημαντικοί προγνωστικοί παράγοντες της νοσηλείας της καρδιακής ανεπάρκειας.

Αυτά τα μοντέλα υπογραμμίζουν τη σημασία των δεδομένων κατάστασης υγείας για τον προσδιορισμό της πρόγνωσης σε ασθενείς με HFpEF και τονίζουν τη σημασία των δεδομένων κατάστασης υγείας για τον προσδιορισμό της πρόγνωσης.



TOPCAT = Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist trial.

Εικόνα 4. Ροή ανάλυσης για την ανάπτυξη και αξιολόγηση μοντέλων



Angraal, S. et al. *J Am Coll Cardiol HF*. 2020;8(1):12-21.

These models accurately identify high risk patients who died or were hospitalized for heart failure exacerbation. AUC = area under the curve; HF = heart failure; TOPCAT = Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist trial.

Εικόνα 5. Μοντέλα πρόβλεψης για την αξιολόγηση των κινδύνων θνησιμότητας και καρδιακής ανεπάρκειας νοσηλείας σε ασθενείς με καρδιακή ανεπάρκεια με διατηρημένο κλάσμα εξώθησης

2.2 Artificial intelligence in the diagnosis and detection of heart failure: the past, present, and future^[7]

Η Τεχνητή Νοημοσύνη (AI) έχει εξελιχθεί τα τελευταία χρόνια και έχει κατορθώσει να μεταμορφώσει σημαντικά τον τομέα της καρδιαγγειακής ιατρικής, επιτρέποντας την έγκαιρη ανίχνευση της καρδιακής ανεπάρκειας, η οποία έχει υψηλό επιπολασμό και ποσοστό θνησιμότητας. Το μοντέλο Νευρωνικού Δικτύου (NN), με ποσοστό ακρίβειας 85%, έπαιξε καθοριστικό ρόλο. Το AI μπορεί επίσης να αναλύσει ακατέργαστα δεδομένα εικόνες από τεχνικές καρδιακής απεικόνισης και καταγραφές ηλεκτροκαρδιογραφήματος, ενσωματώνοντας αλγόριθμους.

Τα δέντρα αποφάσεων και οι μέθοδοι λογιστικής παλινδρόμησης (logistic regression) έχουν χρησιμοποιηθεί για τη δημιουργία μοντέλων λήψης αποφάσεων για τη διάγνωση της συμφορητικής καρδιακής ανεπάρκειας. Έτσι ο ρόλος του AI στην έγκαιρη ανίχνευση μελλοντικών επεισοδίων θνησιμότητας και αποσταθεροποίησης έχει επίσης βελτιώσει σημαντικά τα αποτελέσματα των καρδιαγγειακών παθήσεων. Γενικότερα οι εξελίξεις της τεχνητής νοημοσύνης έχουν αλλάξει σημαντικά την πρόληψη, τη διάγνωση και τη διαχείριση της καρδιακής ανεπάρκειας, αποδεικνύοντας τις δυνατότητές της στη μετατροπή διαφόρων πτυχών της υγειονομικής περίθαλψης.

Η Μηχανική Μάθηση (ML) είναι ο πυρήνας της τεχνητής νοημοσύνης, χρησιμοποιώντας μοντέλα που βασίζονται σε δεδομένα εκπαίδευσης για τη λήψη αποφάσεων και αλγόριθμους προγραμματισμού για την επίλυση προβλημάτων.

Τα κοινά μοντέλα ταξινόμησης περιλαμβάνουν Ταξινόμηση Binary, Multi-class, Multi-label και Imbalanced. Η δυαδική ταξινόμηση χρησιμοποιεί αλγόριθμους όπως Logistic Regression, k-πλησιέστερους γείτονες (kNN), δέντρο αποφάσεων (decision tree), διανυσματικό μηχάνημα υποστήριξης (support vector machine) και naive bayes για την ταξινόμηση εργασιών με δύο ετικέτες.

Το Deep Learning (DL) είναι μια κατηγορία αλγορίθμων ML (Μηχανική Μάθηση) που χρησιμοποιεί χαρακτηριστικά υψηλότερου επιπέδου όπως τα νευρωνικά δίκτυα που προέρχονται από ένα μοντέλο ανθρώπινου εγκεφάλου για την ανάγνωση, τη δημιουργία και την εκμάθηση πολύπλοκων ιεραρχικών αναπαραστάσεων.

Η τεχνητή νοημοσύνη (AI), που βασίζεται σε ML, έχει αποδείξει ότι προβλέπει τα ποσοστά 5ετούς επιβίωσης σε ασθενείς με καρδιαγγειακά νοσήματα με μεγαλύτερη ακρίβεια (80%) από τους κλινικούς γιατρούς (60%). Η καρδιακή ανεπάρκεια είναι μια σημαντική καρδιαγγειακή διαταραχή με ποσοστό θνησιμότητας 10,4% στις 30 ημέρες, 22% σε 1 έτος και 42,3% στα 5 έτη.

2.3 Prevalence, outcomes and costs of a contemporary, multinational population with heart failure^[8]

Η έρευνα αφορά τον επιπολασμό, τα αποτελέσματα και το κόστος υγειονομικής περίθαλψης της καρδιακής ανεπάρκειας σε μεγάλο μέρος του πληθυσμού σε πολλές χώρες. Συγκεκριμένα πρόκειται για την Καρδιονεφρική και Μεταβολική Νόσο (CaReMe) που αποσκοπεί στον προσδιορισμό του παγκόσμιου επιπολασμού της καρδιακής ανεπάρκειας (ΣΥ) σε 11 χώρες. Στην έρευνα τονίζεται ιδιαίτερα η ανάγκη για βελτιωμένες στρατηγικές διαχείρισης και πρόληψης.

Τα δεδομένα από 629.624 άτομα με διαγνωσμένη καρδιακή ανεπάρκεια συγκεντρώθηκαν από ψηφιακά συστήματα υγειονομικής περίθαλψης. Τα δεδομένα αποκάλυψαν υψηλό κίνδυνο θνησιμότητας, συννοσηρότητες (συνύπαρξη δυο νόσων ή δύο παθολογικών καταστάσεων στον ίδιο ασθενή) και παραλλαγές στα πρότυπα θεραπείας σε διάφορες χώρες.

Σύμφωνα με την έρευνα, τα ποσοστά επιπολασμού της καρδιακής ανεπάρκειας σύμφωνα με τους αυστηρούς και γενικούς ορισμούς στη μελέτη καρδιακής ανεπάρκειας της Καρδιονεφρικής και Μεταβολικής νόσου (CaReMe) ήταν 1,05% και 2,01% αντίστοιχα.

Για να αντιμετωπιστούν οι συγκρούσεις συμφερόντων μεταξύ των ερευνητών η μελέτη για την καρδιακή ανεπάρκεια και τις μυοκαρδιοπάθειες ανέφερε ότι οι ασθενείς και/ή το κοινό δεν συμμετείχαν στο σχεδιασμό, τη διεξαγωγή, την αναφορά ή τα σχέδια διάδοσης της έρευνας. Επιπλέον, ανέφερε ότι η συγκατάθεση του ασθενούς για δημοσίευση δεν ίσχυε. Η μελέτη παρέχει επίσης λεπτομέρειες σχετικά με την έγκριση δεοντολογίας από τις χώρες όπου ισχύει, υποδεικνύοντας τη διαδικασία έγκρισης για την έρευνα.

Μερικοί περιορισμοί που αναγνωρίστηκαν στη μελέτη περιλαμβάνουν την άγνωστη γενίκευση των αποτελεσμάτων σε πληθυσμούς με διαφορετικές συνθήκες όσον αφορά τη φυλή, τους πόρους ή τη φροντίδα, καθώς και την υποεκτίμηση του συνολικού κόστους υγειονομικής περίθαλψης που οφείλεται να επικεντρωθεί κυρίως στη νοσοκομειακή περίθαλψη.

Η συντριπτική πλειοψηφία των ασθενών πάσχει από ισχαιμική καρδιοπάθεια, διαβήτη και χρόνια νεφρική νόσο. Η μελέτη ανακάλυψε ότι η καρδιονεφρική νόσος και η θνησιμότητα από κάθε αιτία έχουν τα μεγαλύτερα ποσοστά συμβάντων, ενώ τα καρδιονεφρικά νοσήματα έχουν τις υψηλότερες δαπάνες νοσοκομειακής περίθαλψης.

Τέλος, η μελέτη τονίζει τη σημασία της διαχείρισης της καρδιακής ανεπάρκειας και του δυναμικού της δημόσιας υγείας για την κατανόηση του αντίκτυπου της.

2.4 Development of heart failure risk prediction models based on a multi-marker approach using random forest algorithms^[9]

Σκοπός της μελέτης ήταν να αναπτύξει ένα μοντέλο πρόβλεψης κινδύνου καρδιακής ανεπάρκειας (ΚΑ) χρησιμοποιώντας μια προσέγγιση πολλαπλών δεικτών με τυχαίους αλγόριθμους. Περιλάμβανε 193 συμμετέχοντες και εντόπισε έναν συνδυασμό τεσσάρων καρδιακών βιοδεικτών (CK-MB, BNP, Gal-3 και sST2) που βελτίωσαν την ακρίβεια πρόβλεψης για καρδιακή ανεπάρκεια. Ο αλγόριθμος του random forest τόνισε τη σημασία αυτών των βιοδεικτών και την προγνωστική τους απόδοση, υποδηλώνοντας ότι ο συνδυασμός θα μπορούσε να βοηθήσει στην έγκαιρη αναγνώριση του κινδύνου καρδιακής ανεπάρκειας.

Από τον Ιούνιο έως τον Δεκέμβριο του 2017, συμμετείχαν συνολικά 193 άτομα, όπως προαναφέρθηκε, εκ των οποίων οι 80 νοσούσαν με καρδιακή ανεπάρκεια και οι 113 ήταν υγιείς μάρτυρες. Μεταξύ των καρδιακών βιοδεικτών και των ηχοκαρδιογραφικών παραμέτρων, πραγματοποιήθηκαν αναλύσεις συσχέτισης και παλινδρόμησης.

Σχετικά με τους συμμετέχοντες στη μελέτη με καρδιακή ανεπάρκεια είχαν υψηλότερη συχνότητα οι άνδρες σε σύγκριση με τα άτομα ελέγχου. Επιπλέον, οι ασθενείς με καρδιακή ανεπάρκεια εμφάνισαν σημαντικά υψηλότερα επίπεδα CK-MB, BNP, Gal-3 και sST2 σε σύγκριση με τα άτομα ελέγχου.

Ο πιο σημαντικός προγνωστικός παράγοντας που προσδιορίστηκε από τον αλγόριθμο random forest στην πρόβλεψη της καρδιακής ανεπάρκειας ήταν το BNP (νατριουρητικό πεπτιδίο τύπου Β). Παρουσίασε αξιοσημείωτα υψηλότερη μέση ακρίβεια μείωσης και Gini σε σύγκριση με άλλους βιοδείκτες.

Οι ασθενείς με καρδιακή ανεπάρκεια είχαν σημαντικά υψηλότερα επίπεδα CK-MB, BNP, Gal-3 και sST2. Το BNP έδειξε μια ισχυρή ανεξάρτητη προγνωστική ικανότητα για ΚΑ, αλλά τα CK-MB, sST2 και Gal-3 είχαν περιορισμένο διαγνωστικό δυναμικό για ΚΑ. Με μεγαλύτερη μέση ακρίβεια μείωσης και Gini, BNP ήταν η πιο σχετική μεταβλητή. Το μοντέλο πολλαπλών δεικτών αποκάλυψε μια συνολική άνοδο.

Τέλος η μελέτη για την πρόβλεψη της καρδιακής ανεπάρκειας χρησιμοποιώντας μια προσέγγιση πολλαπλών δεικτών με αλγόριθμους όπως ο random forest υποστηρίχθηκε από την Επιτροπή Δεοντολογίας του Νοσοκομείου Anzhen του Πεκίνου.

2.5 Development and validation of prediction models for incident atrial fibrillation in heart failure^[10]

Η μελέτη αφορά την ανάπτυξη και την επικύρωση μοντέλων πρόβλεψης για περιστατική κολπική μαρμαρυγή (AF) μεταξύ ασθενών με καρδιακή ανεπάρκεια (KA) χρησιμοποιώντας δεδομένα. Τα μοντέλα έδειξαν μέτρια προγνωστική απόδοση, με παρόμοια αποτελέσματα για μοντέλα παλινδρόμησης random forest και ειδικών αιτιών Cox. Οι μελέτες εντόπισαν προγνωστικές μεταβλητές όπως η ηλικία, το φύλο, οι συννοσηρότητες και οι παράγοντες του τρόπου ζωής για την πρόβλεψη του κινδύνου εμφάνισης κολπικής μαρμαρυγής εντός ενός έτους σε ασθενείς με καρδιακή ανεπάρκεια. Παράλληλα, στην μελέτη τονίζεται η σημασία της έγκαιρης παρέμβασης και της παρακολούθησης ατόμων υψηλού κινδύνου για τη βελτίωση των αποτελεσμάτων σε ασθενείς με καρδιακή ανεπάρκεια, με συστάσεις για προσυμπτωματικό έλεγχο σε ασθενείς με καρδιακή ανεπάρκεια και βελτιστοποίηση της φροντίδας της καρδιακής ανεπάρκειας για την πρόληψη της έναρξης της κολπικής μαρμαρυγής.

Οι προγνωστικές μεταβλητές που εντοπίστηκαν στη μελέτη που στόχευε στην ανάπτυξη και επικύρωση μοντέλων κλινικής πρόβλεψης για τον κίνδυνο ενός έτους κολπικής μαρμαρυγής σε ασθενείς με καρδιακή ανεπάρκεια ήταν το φύλο, η ηλικία κατά τη διάγνωση καρδιακής ανεπάρκειας, η κατηγορία New York Heart Association (NYHA), η υπέρταση, ο σακχαρώδης διαβήτης, η χρόνια νεφρική νόσος, η αποφρακτική άπνοια ύπνου, η χρόνια αποφρακτική πνευμονοπάθεια και το έμφραγμα του μυοκαρδίου.

Τα μοντέλα που αναπτύχθηκαν χρησιμοποιήσαν δεδομένα στην πρόβλεψη περιστατικού κολπικής μαρμαρυγής σε ασθενείς με καρδιακή ανεπάρκεια. Συγκεκριμένα αναφέρεται στην μελέτη, ότι είχαν προγνωστική απόδοση με περιοχή εξαρτώμενη από το χρόνο κάτω από την καμπύλη 65,7% και βαθμολογία Brier 7,0% στην πρόβλεψη περιστατικού κολπικής μαρμαρυγής σε ασθενείς με καρδιακή ανεπάρκεια.

Η μελέτη διεξήχθη χρησιμοποιώντας το Δανικό Μητρώο Καρδιακής Ανεπάρκειας και διαπίστωσε ότι ο κίνδυνος κολπικής μαρμαρυγής (KM) για ένα έτος σε ασθενείς με καρδιακή ανεπάρκεια είναι 9,3% για τους άνδρες και 6,4% για τις γυναίκες δεδομένων όλων των παραγόντων κινδύνου και του NYHA III/IV.

Επίσης η μελέτη διαπίστωσε ότι ο προβλεπόμενος κίνδυνος ήταν 5,1% για την απουσία παραγόντων κινδύνου και κατηγορίας I NYHA. Η μελέτη υποδηλώνει ότι η εφαρμογή του μοντέλου σε κλινικές ρυθμίσεις ρουτίνας θα μπορούσε να βελτιώσει την ακρίβεια της πρόβλεψης του κινδύνου AF μεταξύ ασθενών με καρδιακή ανεπάρκεια.

Ο εντοπισμός ατόμων υψηλού κινδύνου για έγκαιρη παρέμβαση και παρακολούθηση σε ασθενείς με καρδιακή ανεπάρκεια είναι σημαντικός, καθώς επιτρέπει την έναρξη προληπτικών προσπαθειών για τη μείωση των τροποποιήσιμων παραγόντων κινδύνου για κολπική μαρμαρυγή (AF) και καθοδηγεί την έγκαιρη παρακολούθηση για την ανίχνευση μη διαγνωσμένης κολπικής μαρμαρυγής. Αυτή η προληπτική προσέγγιση μπορεί να οδηγήσει στην έγκαιρη ανίχνευση και διαχείριση της κολπικής μαρμαρυγής, βελτιώνοντας δυνητικά τα αποτελέσματα για ασθενείς με καρδιακή ανεπάρκεια. Επιπλέον, η έγκαιρη αναγνώριση ατόμων υψηλού κινδύνου μπορεί να προκαλέσει την έναρξη έγκαιρης φαρμακευτικής αγωγής ελέγχου του ρυθμού και κατάλυσης με καθετήρα, παρεμβάσεις που έχουν δείξει οφέλη.

2.6 Heart failure mortality prediction using PRISM score and development of a classification and regression tree model to refer patients for palliative care consultation^[11]

Η μελέτη στόχευε να αξιολογήσει τη θνησιμότητα ενός έτους σε ασθενείς με καρδιακή ανεπάρκεια χρησιμοποιώντας PRISM, μια βαθμολογία διαστρωμάτωσης μη ειδικού κινδύνου για τη νόσο και τροποποιημένο μοντέλο καρδιακής ανεπάρκειας Seattle (SHFM) για να καθοδηγήσει την επιλογή ασθενών για συμβουλευτική παρηγορητικής φροντίδας. Η μελέτη συνέκρινε τη διακριτική ικανότητα της κατηγορικής βαθμολογίας PRISM και της τροποποιημένης βαθμολογίας SHFM στην πρόβλεψη της θνησιμότητας ενός έτους σε ασθενείς με καρδιακή ανεπάρκεια. Η διακριτική ικανότητα της κατηγορικής βαθμολογίας PRISM δεν διέφερε σημαντικά από την τροποποιημένη βαθμολογία SHFM, αλλά βελτιώθηκε σημαντικά όταν η βαθμολογία PRISM συνδυάστηκε με την τροποποιημένη βαθμολογία SHFM.

Ένα μοντέλο Classification and Regression Tree (CART) αναπτύχθηκε για την ταξινόμηση των ασθενών με βάση τον κίνδυνο θνησιμότητας χρησιμοποιώντας αυτές τις βαθμολογίες. Η μελέτη προτείνει ότι αυτά τα εργαλεία μπορούν να εντοπίσουν αποτελεσματικά ασθενείς με υψηλό κίνδυνο θνησιμότητας και να βοηθήσουν στην καθοδήγηση των συζητήσεων σχετικά με τους στόχους περίθαλψης. Επιπλέον, η ενσωμάτωση της παρηγορητικής φροντίδας στη διαχείριση ασθενών με καρδιοπάθεια συνιστάται για τη βελτίωση της ποιότητας ζωής, με διάφορα εργαλεία πρόβλεψης θνησιμότητας διαθέσιμα για να βοηθήσουν στην πρόγνωση και τη διαστρωμάτωση κινδύνου για την παρηγορητική φροντίδα.

Διεξήχθη μια αναδρομική μελέτη 689 ασθενών με καρδιακή ανεπάρκεια από το 2012 έως το 2014, με τη θνησιμότητα ενός έτους να υπολογίζεται χρησιμοποιώντας Pmort30/PRISM και τροποποιημένες βαθμολογίες SHFM.

Σύμφωνα με τη μελέτη, παράγοντες που συμβάλλουν στην υποχρησιμοποίηση της παρηγορητικής φροντίδας σε ασθενείς με καρδιοπάθεια περιλαμβάνουν αβεβαιότητα στην πορεία της νόσου, κακή επικοινωνία, έλλειψη γνώσης και περίπλοκες αποφάσεις θεραπείας όπως εμφύτευση και απενεργοποίηση ICD, LVAD και μεταμόσχευση καρδιάς.

Οι προβλεπόμενες βαθμολογίες επικυρώθηκαν χρησιμοποιώντας την περιοχή κάτω από την καμπύλη ROC. Ένα μοντέλο δέντρου CART αναπτύχθηκε για την ταξινόμηση των ασθενών με βάση τον κίνδυνο θνησιμότητας τους. Τα αποτελέσματα έδειξαν ότι η διακριτική ικανότητα της κατηγορικής βαθμολογίας PRISM δεν διέφερε σημαντικά από το τροποποιημένο SHFM, αλλά βελτιώθηκε σημαντικά με το συνδυασμό του PRISM και του τροποποιημένου SHFM. Η προγνωστική ικανότητα του μοντέλου δέντρου CART ήταν 72,2% μετά από διασταυρούμενη επικύρωση.

Η μελέτη πρότεινε τη χρήση εργαλείων διαστρωμάτωσης κινδύνου όπως το PRISM και το SHFM για τον εντοπισμό ασθενών με υψηλότερο κίνδυνο θνησιμότητας και καθοδήγηση των στόχων των συζητήσεων περίθαλψης, οδηγώντας ενδεχομένως σε βελτιωμένα αποτελέσματα σε ασθενείς με καρδιακή ανεπάρκεια. Ωστόσο κατέληξε στο συμπέρασμα ότι το PRISM και το τροποποιημένο SHFM είχαν καλή απόδοση για την πρόβλεψη θνησιμότητας ενός έτους και όταν συνδυάζονται σε ένα μοντέλο CART, μπορούν να βοηθήσουν στον εντοπισμό του πληθυσμού με τον υψηλότερο κίνδυνο θνησιμότητας και να καθοδηγήσουν τους στόχους συζήτησης περίθαλψης.

2.7 Heart Failure Development Prediction using Stochastic Gradient Descent Optimization^[12]

Η έρευνα που χρηματοδοτείται από το Εθνικό Ίδρυμα Ερευνών της Ουκρανίας συμβάλλει στην ανάπτυξη ευφυών τεχνολογιών για την αξιολόγηση επιδημικών καταστάσεων εστιάζοντας στη δημιουργία ενός μοντέλου πρόβλεψης για ασθενείς με ύποπτη καρδιακή ανεπάρκεια με βάση μεθόδους μηχανικής μάθησης. Αυτό το μοντέλο στοχεύει να προβλέψει την ανάπτυξη της καρδιακής ανεπάρκειας στα πρώιμα στάδια, επιτρέποντας τον προσδιορισμό των κινδύνων επιπλοκών και θανάτων από τη νόσο σε πρώιμο στάδιο.

Η μηχανική μάθηση και η βελτιστοποίηση της στοχαστικής κλίσης κατάβασης διαδραματίζουν κρίσιμο ρόλο στην πρόβλεψη της ανάπτυξης καρδιακής ανεπάρκειας στους ασθενείς. Η μελέτη χρησιμοποίησε ένα μοντέλο γραμμικής παλινδρόμησης βελτιστοποιημένο με τη μέθοδο στοχαστικής κλίσης καθόδου για να προβλέψει την ανάπτυξη καρδιακής ανεπάρκειας σε ασθενείς με υποψία καρδιακής ανεπάρκειας. Αυτή η προσέγγιση επέτρεψε την προσαρμογή των παραμέτρων για την ελαχιστοποίηση μιας συγκεκριμένης συνάρτησης κόστους επαναληπτικά, με αποτέλεσμα ένα μοντέλο με επαρκή ακρίβεια για χρήση στην ιατρική πρακτική.

Το μοντέλο χρησιμοποιεί ένα μοντέλο γραμμικής παλινδρόμησης βελτιστοποιημένο με μια μέθοδο στοχαστικής κλίσης. Η μελέτη χρησιμοποιεί ένα ανοικτό σύνολο δεδομένων 1504 εγγραφών με 13 χαρακτηριστικά.

Το μοντέλο έχει επαρκείς δείκτες απόδοσης για έγκαιρη πρόβλεψη καρδιακής ανεπάρκειας στην κλινική πράξη, ωφελώντας πάνω από 64 εκατομμύρια ανθρώπους παγκοσμίως.

2.8 Prediction of patients with heart failure after myocardial infarction^[13]

Η καρδιακή ανεπάρκεια είναι μια διαδεδομένη αιτία θανάτου παγκοσμίως, με ένα υψηλό ποσοστό ασθενών να διαγιγνώσκεται σε σύντομο χρονικό διάστημα μετά το έμφραγμα του μυοκαρδίου. Αυτή η μελέτη είχε ως στόχο να μεταφράσει υπάρχοντα δεδομένα από τον πραγματικό κόσμο σε πραγματικές αποδείξεις διερευνώντας συσχετίσεις μεταξύ ιστορικών συννοσηροτήτων και ασθενειών καρδιακής ανεπάρκειας.

Η μελέτη χρησιμοποίησε μοντέλα μηχανικής μάθησης για να προβλέψει την ανάπτυξη καρδιακής ανεπάρκειας σε ασθενείς ηλικίας κάτω των 71 ετών που είχαν υποστεί έμφραγμα του μυοκαρδίου. Συγκρίθηκαν διάφορα μοντέλα, με τον random forest να έχει την καλύτερη απόδοση και η ηλικιακή διαστρωμάτωση βρέθηκε να ενισχύει την ακρίβεια πρόβλεψης. Η έρευνα υπογραμμίζει τη σημασία της ενσωμάτωσης της ηλικίας στο προγνωστικό μοντέλο για ασθενείς μετά από έμφραγμα του μυοκαρδίου για τη βελτίωση των αποτελεσμάτων.

Οι τεχνολογίες μηχανικής μάθησης χρησιμοποιήθηκαν για να προβλέψουν την ανάπτυξη καρδιακής ανεπάρκειας σε ασθενείς με έμφραγμα του μυοκαρδίου και να τους υπενθυμίσουν να ενισχύσουν την προσωπική αυτοφροντίδα για να αποφύγουν τη μετάβαση στην καρδιακή ανεπάρκεια ή να αναβάλουν την εμφάνιση προηγούμενων συμβάντων.

Η μελέτη χρησιμοποίησε τρεις διαφορετικές τεχνολογίες μηχανικής μάθησης, συγκεκριμένα την λογιστική παλινδρόμηση, τον random forest και τον XGBoost, για τη δημιουργία μοντέλων πρόβλεψης για ασθενείς μετά από έμφραγμα του μυοκαρδίου. Αυτά τα μοντέλα αξιολογήθηκαν χρησιμοποιώντας μια πενταπλάσια μέθοδο διασταυρούμενης επικύρωσης (5 fold cross validation) για την αξιολόγηση της ακρίβειας και της σταθερότητας της πρόβλεψης.

Η έρευνα χρησιμοποίησε την Εθνική Ερευνητική Βάση Δεδομένων Ασφάλισης Υγείας της Ταϊβάν (NHIRD) για να αναλύσει συννοσηρότητες και παράγοντες που σχετίζονται με τη μετάβαση στην καρδιακή ανεπάρκεια σε ασθενείς μετά από έμφραγμα του μυοκαρδίου.

Τα αποτελέσματα έδειξαν ότι η κατασκευή μοντέλων πρόβλεψης χωρίς ηλικιακή διαστρωμάτωση παρείχε κατώτερη απόδοση σε σύγκριση με τις στρωματοποιημένες ομάδες που χρησιμοποιούν πανομοιότυπα χαρακτηριστικά.

2.9 Heart Failure Prediction by Feature Ranking Analysis in Machine Learning^[14]

Οι καρδιακές παθήσεις είναι μια κύρια αιτία θνησιμότητας παγκοσμίως και η πρόβλεψη τους είναι μια κρίσιμη πρόκληση στην ανάλυση κλινικών δεδομένων. Οι εξελίξεις στη μηχανική μάθηση (ML), στην τεχνητή νοημοσύνη (AI) και στην επιστήμη δεδομένων έχουν αποδειχθεί αποτελεσματικές στη λήψη αποφάσεων και στις προβλέψεις από μεγάλους όγκους δεδομένων που παράγονται από τη βιομηχανία υγειονομικής περίθαλψης. Η εφαρμογή τεχνικών μηχανικής μάθησης για την πρόβλεψη καρδιακών παθήσεων είναι καθοριστική.

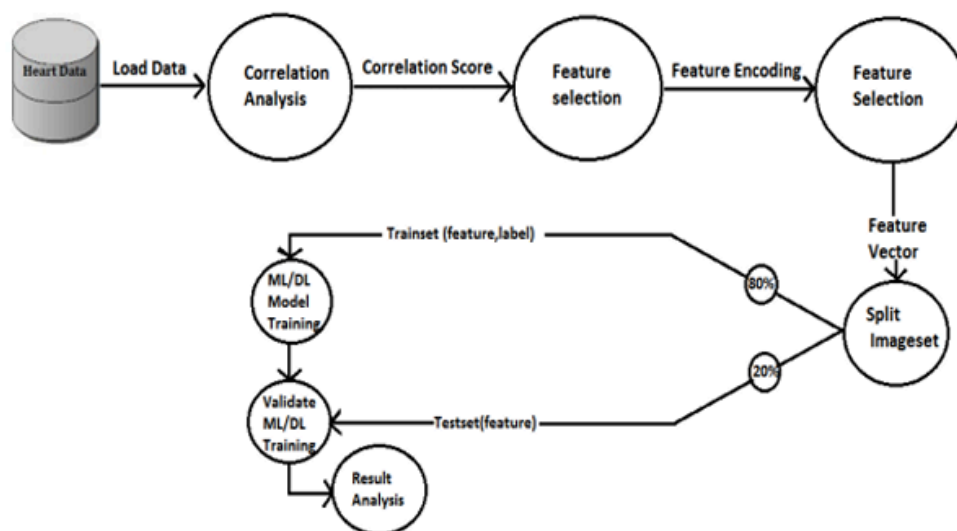
Οι προσεγγίσεις της μηχανικής μάθησης (ML) έχουν διευρύνει τη μελέτη σε ιατρικούς τομείς, αναγνωρίζοντας μοτίβα στο ανθρώπινο σώμα χρησιμοποιώντας αλγόριθμους και τεχνικές συσχέτισης. Μια τέτοια πραγματικότητα είναι η στεφανιαία νόσος, η οποία μπορεί να προβλεφθεί χρησιμοποιώντας τεχνικές μηχανικής μάθησης (ML).

Διάφοροι αλγόριθμοι και τεχνικές συσχέτισης χρησιμοποιούνται για τον εντοπισμό σημαντικών χαρακτηριστικών που επηρεάζουν την καρδιακή ανεπάρκεια. Η έρευνα υπογραμμίζει τη σημασία της έγκαιρης ανίχνευσης των καρδιακών παθήσεων για τη μείωση των ποσοστών θνησιμότητας και τονίζει τον αντίκτυπο της μηχανικής μάθησης στη βοήθεια των επαγγελματιών υγείας.

Αρχικά, η μηχανική μάθηση (ML) χρησιμοποιήθηκε για τον προσδιορισμό του βαθμού καρδιακής ανεπάρκειας και τον εντοπισμό σημαντικών χαρακτηριστικών που επηρεάζουν την καρδιακή νόσο χρησιμοποιώντας τεχνικές συσχέτισης. Αυτή η εργασία προτείνει μια μέθοδο εύρεσης σημαντικών χαρακτηριστικών χρησιμοποιώντας τεχνικές μηχανικής μάθησης (ML), σχεδιάζοντας και αναπτύσσοντας πρόβλεψη καρδιακών παθήσεων χρησιμοποιώντας κατάταξη χαρακτηριστικών.

Το SVM (Support Vector Machine) έδειξε την υψηλότερη ακρίβεια 95% για την πρόβλεψη καρδιακών παθήσεων.

Μερικοί από τους βασικούς παράγοντες που επηρεάζουν την καρδιακή νόσο που αναφέρονται σε ένα από τα έγγραφα περιλαμβάνουν το φύλο, την αιμοσφαιρίνη, την ασθένεια του θυρεοειδούς, το H_B, το H_C, την κίρρωση, τα τριγλυκερίδια και το CRF.



Εικόνα 6. Σχεδιάγραμμα μεθοδολογίας

2.10 Derivation and validation of a machine learning-based risk prediction model in patients with acute heart failure^[15]

Αυτή η μελέτη στοχεύει να αναπτύξει ένα φιλικό προς τον χρήστη μοντέλο πρόβλεψης κινδύνου χρησιμοποιώντας μεθόδους μηχανικής μάθησης, που ονομάζεται 4 V-RS, για την πρόβλεψη της ενδονοσοκομειακής θνησιμότητας σε ασθενείς με οξεία καρδιακή ανεπάρκεια (AHF). Ένα μοντέλο κινδύνου που βασίζεται στη μηχανική μάθηση αναπτύχθηκε χρησιμοποιώντας παλινδρόμηση χειριστή ελάχιστης απόλυτης συρρίκνωσης και επιλογής (LASSO) για τον εντοπισμό παραγόντων πρόβλεψης ενδονοσοκομειακής θνησιμότητας.

Το μοντέλο ενσωματώνει τη συστολική αρτηριακή πίεση, το άζωτο της ουρίας του αίματος, το χλωρίδιο του ορού και την C-αντιδρώσα πρωτεΐνη ως μεταβλητές και έδειξε καλή διάκριση και βαθμονόμηση στην πρόβλεψη των αποτελεσμάτων. Το μοντέλο 4 V-RS ξεπέρασε τα υπάρχοντα μοντέλα κινδύνου, όπως το μοντέλο κινδύνου ADHERE, και βρέθηκε ότι είναι φιλικό προς τον χρήστη και εύκολα εφαρμόσιμο στην κλινική πράξη.

Οι μεταβλητές που περιλαμβάνονται στο μοντέλο 4 V-RS για την πρόβλεψη της ενδονοσοκομειακής θνησιμότητας σε ασθενείς με οξεία καρδιακή ανεπάρκεια είναι η συστολική αρτηριακή πίεση, το BUN (άζωτο ουρίας αίματος), το χλωριούχο ορό και η CRP (C-αντιδρώσα πρωτεΐνη).

Το μοντέλο επικυρώθηκε εξωτερικά στην κοόρτη επικύρωσης και συγκρίθηκε με δύο προϋπάρχοντα μοντέλα κινδύνου: τη βαθμολογία κινδύνου Get With The Guidelines που ενσωματώνει εγκεφαλικό νατριουρητικό πεπτιδίο και υποχλωραιμία (GWTG-BNP-CI-RS) και τον κίνδυνο οξείας μη αντιρροπούμενης καρδιακής ανεπάρκειας στο εθνικό μητρώο (EMMENΩ).

Το μοντέλο έδειξε υψηλότερη περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη (AUC) από το μοντέλο κινδύνου ADHERE και σημαντική βελτίωση στην καθαρή αναταξινόμηση. Επομένως, απέδωσε καλύτερα από το συγκριτικά απλό μοντέλο κινδύνου ADHERE στην πρόβλεψη των αποτελεσμάτων για ασθενείς με οξεία καρδιακή ανεπάρκεια. Σημειώνεται πως είχε παρόμοια απόδοση με το GWTG-BNP-CI-RS όσον αφορά τη διάκριση και την καθαρή αναταξινόμηση. Το μοντέλο 4V-RS, το οποίο περιείχε μόνο τέσσερα άμεσα διαθέσιμα σημεία δεδομένων τη στιγμή της εισαγωγής, είχε παρόμοια απόδοση με το πιο περίπλοκο προϋπάρχον μοντέλο κινδύνου σε ασθενείς με AHF.

3. Περιγραφή Δεδομένων

Στην παρούσα εργασία όλα υλοποιούνται βασισμένα σε ένα dataset. Πρόκειται για το dataset “Heart Failure Prediction” (“heart_failure_clinical_records_dataset”) που βρίσκεται στο Kaggle^[16]. Χρησιμοποιείται για προβλέψεις σχετικά με την καρδιακή ανεπάρκεια και περιέχει κλινικά και δημογραφικά χαρακτηριστικά ασθενών που πάσχουν από αυτήν. Παράλληλα μπορεί να χρησιμοποιηθεί για να προβλέψει αν ένας ασθενής είναι πιθανό να πάσχει από καρδιακή ανεπάρκεια στο μέλλον.

Για κάθε ασθενή δηλώνονται τα εξής χαρακτηριστικά:

- ❖ Ηλικία (age)
- ❖ Αναιμία (anaemia)
- ❖ Κρεατινίνη Φωσφοκινάσης/Επίπεδο του ενζύμου CPK στο αίμα (mcg/L) (creatinine phosphokinase)
- ❖ Διαβήτης (diabetes)
- ❖ Ποσοστό αίματος που φεύγει από την καρδιά σε κάθε συστολή (ejection_fraction)
- ❖ Υψηλή πίεση / Υπέρταση (high_blood_pressure)
- ❖ Αιμοπετάλια (platelets)
- ❖ Κρεατινίνη (serum_creatinine)
- ❖ Νάτριο (serum_sodium)
- ❖ Φύλο (sex)
- ❖ Κάπνισμα (smoking)
- ❖ Χρόνος νοσηλείας (time)

Όλα τα παραπάνω χαρακτηριστικά αναλύονται με στόχο να βρεθεί/εντοπιστεί κάποιο πιθανό μοντέλο αλλά και συσχετίσεις μεταξύ των χαρακτηριστικών των ασθενών για τις πιθανότητες να αναπτύξουν καρδιακή ανεπάρκεια. Με βάση αυτά τα δεδομένα θα δημιουργηθούν μοντέλα πρόβλεψης για να προβλέπουν τον κίνδυνο της καρδιακής ανεπάρκειας για μελλοντικούς ασθενείς.

Για την ανάλυση και τη χρήση του συνόλου δεδομένων απαιτούνται κατάλληλες τεχνικές μηχανικής μάθησης και στατιστικής, και πρέπει να γίνουν με προσοχή και ακρίβεια για να εξαλειφθούν αξιόπιστα συμπεράσματα.

Τα δείγματα είναι οι επιμέρους εγγραφές ή εγγραφές δεδομένων που αναφέρονται στο σύνολο δεδομένων. Σε αυτήν την περίπτωση, τα δείγματα αναφέρονται σε διαφορετικούς ασθενείς που υποβάλλονται σε κλινική μελέτη ή έχουν παρακολουθηθεί για κάποια χρονική περίοδο και φτάνουν τα τριακόσια (300) σε αριθμό. Επομένως το σύνολο των δεδομένων αποτελείται από τριακόσιους (300) ασθενείς, δώδεκα (12) χαρακτηριστικά και την στήλη έξοδο (tag) με τιμές μηδέν (0) και ένα (1) αν έζησε ή πέθανε ο ασθενής.

Κάποια από τα χαρακτηριστικά αναλύονται ακολούθως:

1. Υψηλή πίεση / Υπέρταση (high_blood_pressure)

Η υπέρταση είναι μια κατάσταση κατά την οποία οι αρτηριακές πιέσεις στα αγγεία του αίματος είναι μόνιμα υψηλές. Αυτό μπορεί να οφείλεται σε διάφορους παράγοντες, όπως τον τρόπο ζωής, τη διατροφή, τη γενετική προδιάθεση και την υγεία των νεφρών και του καρδιαγγειακού συστήματος. Η υπέρταση μπορεί να είναι επικίνδυνη, καθώς αυξάνει τον

κίνδυνο εμφάνισης καρδιαγγειακών προβλημάτων, εγκεφαλικών επεισοδίων και νεφρικών προβλημάτων.

2. Διαβήτης (diabetes)

Ο διαβήτης είναι μια χρόνια κατάσταση που σχετίζεται με την ανεπάρκεια της ινσουλίνης ή την ανεπαρκή χρήση της από το σώμα. Η ινσουλίνη είναι ορμόνη που παράγεται από το πάγκρεας και είναι υπεύθυνη για τη ρύθμιση της σακχάρου στο αίμα. Αν ο διαβήτης δεν διαχειρίζεται σωστά, μπορεί να προκαλέσει πολλά σοβαρά προβλήματα υγείας, όπως καρδιαγγειακές παθήσεις, νευρολογικά προβλήματα και προβλήματα με τα μάτια, τους νεφρούς και τα άκρα. Χαρακτηρίζεται από τα υψηλά επίπεδα γλυκόζης στο αίμα, είτε λόγω έλλειψης ινσουλίνης (Τύπος 1) είτε λόγω αντοχής στην ινσουλίνη (Τύπος 2).

3. Κρεατινίνη φωσφοκινάσης / Επίπεδο του ενζύμου CPK στο αίμα (creatinine phosphokinase)

Η κρεατινίνη φωσφοκινάσης είναι μια εξέταση του αίματος που χρησιμοποιείται για να μετρήσει το επίπεδο της κρεατινίνης στο αίμα. Η κρεατινίνη είναι ένα προϊόν του μεταβολισμού της κρεατίνης, που βρίσκεται στους μύες. Η μέτρηση της κρεατινίνης φωσφοκινάσης είναι σημαντική για την αξιολόγηση της νεφρικής λειτουργίας, καθώς το υψηλό επίπεδο μπορεί να υποδηλώνει προβλήματα με τα νεφρά. Το ένζυμο CPK βρίσκεται κυρίως στους μύες και τον εγκέφαλο, και τα αυξημένα επίπεδα του στο αίμα μπορεί να υποδηλώνουν μυϊκή βλάβη ή άλλα προβλήματα υγείας.

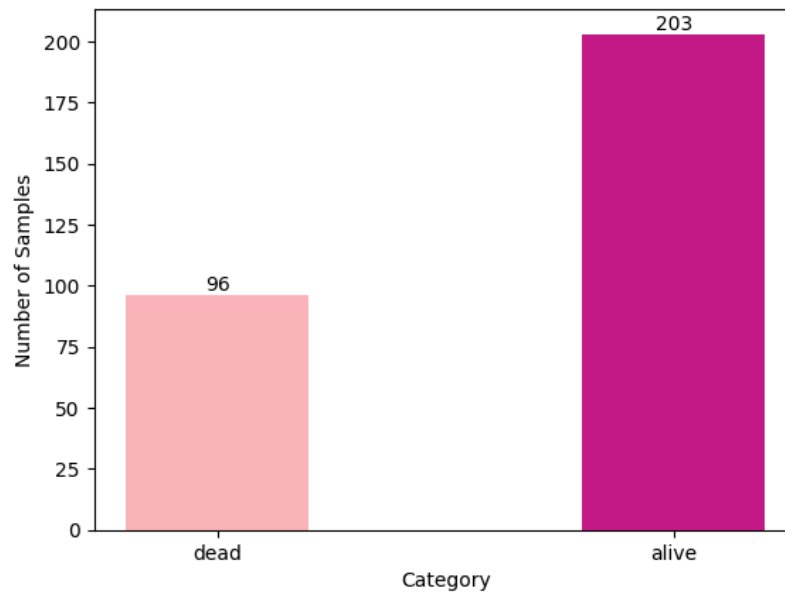
4. Αναιμία (anaemia)

Πρόκειται για μια κατάσταση κατά την οποία το αίμα δεν έχει επαρκή αριθμό υγιών ερυθρών αιμοσφαιρίων ή αιμοσφαιρίνης. Αυτό μπορεί να οδηγήσει σε κόπωση, αδυναμία και άλλα συμπτώματα λόγω της μειωμένης ικανότητας του αίματος να μεταφέρει οξυγόνο στους ιστούς.

5. Ποσοστό αίματος που φεύγει από την καρδιά σε κάθε συστολή (ejection fraction)

Ένας δείκτης της καρδιακής λειτουργίας που μετρά το ποσοστό του αίματος που εξωθείται από την αριστερή κοιλία της καρδιάς σε κάθε συστολή. Τα χαμηλά ποσοστά μπορεί να υποδηλώνουν καρδιακή ανεπάρκεια.

Το σύνολο των δεδομένων (dataset) όπως προαναφέρθηκε είναι επισημειωμένο με μηδέν (0) και ένα (1) και απαρτίζεται από δώδεκα (12) χαρακτηριστικά. Στην εικόνα 7 απεικονίζεται η κατανομή των δύο κατηγοριών και στην συνέχεια στην εικόνα 8 ένα διάγραμμα στήλης (bar plot) που απεικονίζει τη σημασία των χαρακτηριστικών (feature importance) σύμφωνα με την μέθοδο Borda Consensus.



Εικόνα 7. Αριθμός δειγμάτων ανά κατηγορία

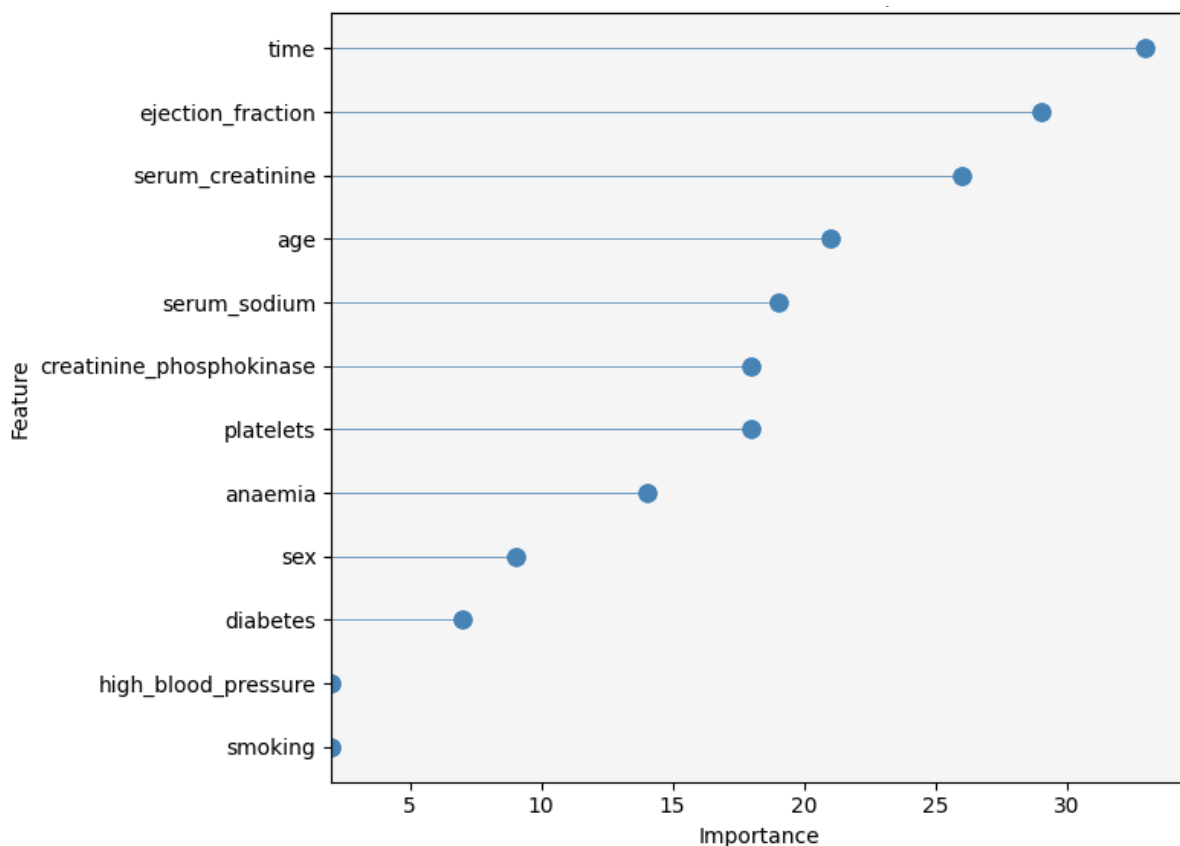
Το διάγραμμα στήλης (bar plot) απεικονίζει τη σημασία των χαρακτηριστικών (feature importance) σύμφωνα με την μέθοδο Borda Consensus. Η μέθοδος αυτή συνδυάζει διαφορετικές μεθόδους αξιολόγησης της σημασίας των χαρακτηριστικών για να καταλήξει σε μια συναινετική κατάταξη.

Στον κατακόρυφο άξονα (y-axis) αναγράφονται τα χαρακτηριστικά που αξιολογούνται, όπως time, ejection_fraction, serum_creatinine, age, serum_sodium, creatinine_phosphokinase, platelets, anaemia, sex, diabetes, high_blood_pressure και smoking.

Στον οριζόντιο άξονα (x-axis) αναγράφεται η τιμή της σημασίας του κάθε χαρακτηριστικού. Οι τιμές αυτές δείχνουν πόσο σημαντικό είναι κάθε χαρακτηριστικό για την πρόβλεψη του αποτελέσματος (DEATH_EVENT) βάσει του συναινετικού μοντέλου.

Και τέλος τα χαρακτηριστικά ταξινομούνται από το πιο σημαντικό στο λιγότερο σημαντικό. Το χαρακτηριστικό "time" είναι το πιο σημαντικό, ενώ το "smoking" είναι το λιγότερο σημαντικό σύμφωνα με την μέθοδο Borda Consensus. Μπορούμε να καταλήξουμε στο συμπέρασμα πως ο χρόνος (time) είναι το πιο σημαντικό χαρακτηριστικό, κάτι που μπορεί να υποδηλώνει ότι ο χρόνος παρακολούθησης ή νοσηλείας παίζει καθοριστικό ρόλο στην πρόβλεψη της επιβίωσης. Στην συνέχεια το ποσοστό εξώθησης (ejection_fraction) και η κρεατινίνη (serum_creatinine) επίσης εμφανίζονται ως πολύ σημαντικά χαρακτηριστικά σε αντίθεση με το κάπνισμα (smoking) και την υψηλή πίεση (high_blood_pressure) που εμφανίζονται με χαμηλότερη σημασία στην πρόβλεψη του αποτελέσματος θανάτου σε αυτό το σύνολο δεδομένων.

Συνοπτικά, λοιπόν, η εικόνα 8 παρέχει μια οπτική αναπαράσταση της συνεισφοράς κάθε χαρακτηριστικού στο μοντέλο πρόβλεψης.



Εικόνα 8. Σημαντικότητα Χαρακτηριστικών

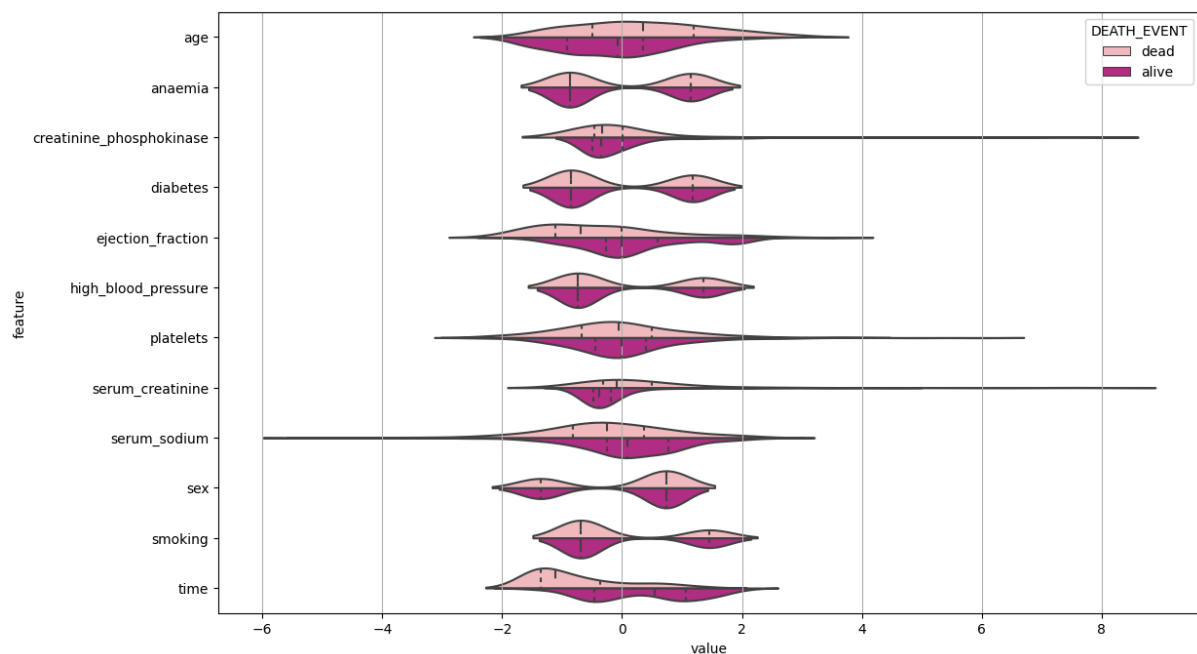
Ένα violin plot (γράφημα βιολιού) είναι ένα γραφικό εργαλείο που χρησιμοποιείται για την απεικόνιση της κατανομής δεδομένων, καθώς παρέχει μια πλούσια και λεπτομερή εικόνα των δεδομένων. Συγκεκριμένα, ένα violin plot δείχνει την πυκνότητα κατανομής, τις κεντρικές τάσεις, την συμμετρία ή την ασυμμετρία των κατανομών και τέλος συγκρίνει τις κατανομές. Είναι ιδιαίτερα χρήσιμο για την κατανόηση της πλήρους κατανομής των δεδομένων. Παρέχει οπτική αναπαράσταση της πυκνότητας και της συμμετρίας της κατανομής, επιτρέποντας έτσι την εύκολη σύγκριση μεταξύ διαφορετικών συνόλων δεδομένων.

Στην εικόνα 9 το violin plot συγκρίνει διάφορα χαρακτηριστικά (features) μεταξύ δύο ομάδων ατόμων αυτών που είναι ζωντανοί (alive) και αυτών που έχουν πεθάνει (dead). Πιο αναλυτικά, στον κατακόρυφο άξονα (y-axis) αναγράφονται τα χαρακτηριστικά που εξετάζονται (π.χ., age, anaemia, creatinine_phosphokinase, diabetes, κλπ.) και στον οριζόντιο άξονα (x-axis) αναγράφονται οι τιμές των χαρακτηριστικών σε κανονικοποιημένη κλίμακα.

Για κάθε χαρακτηριστικό διακρίνεται η κατανομή των τιμών του για τις δύο ομάδες. Οι φαρδιές περιοχές δείχνουν όπου οι τιμές είναι πιο πυκνές (υπάρχουν περισσότερα δεδομένα). Φυσικά, το χρώμα διαχωρίζει τις δύο ομάδες - το ροζ χρώμα αντιπροσωπεύει τα άτομα που έχουν πεθάνει (dead), ενώ το μοβ χρώμα αντιπροσωπεύει τα άτομα που είναι ζωντανά (alive). Το violin plot παρουσιάζει συμμετρία γύρω από την κεντρική κάθετη γραμμή του κάθε χαρακτηριστικού. Αυτό σημαίνει ότι για κάθε χαρακτηριστικό, η κατανομή της πυκνότητας των δεδομένων για τις δύο ομάδες ("dead" και "alive") είναι αντικατοπτρική ως προς την κεντρική γραμμή του κάθε "βιολιού".

Αν οι κατανομές είναι συμμετρικές, οι δύο πλευρές του "βιολιού" θα έχουν παρόμοιο σχήμα και εύρος. Για παράδειγμα, το χαρακτηριστικό "serum_sodium" φαίνεται σχετικά συμμετρικό, καθώς οι δύο πλευρές του "βιολιού" είναι σχεδόν καθρέφτες η μία της άλλης. Απ' την άλλη πλευρά αν οι κατανομές είναι ασύμμετρες, οι δύο πλευρές του "βιολιού" θα έχουν διαφορετικό σχήμα ή εύρος. Για παράδειγμα, το χαρακτηριστικό "creatinine_phosphokinase" δείχνει κάποιες ασυμμετρίες, καθώς οι δύο πλευρές του "βιολιού" δεν είναι ίδιες. Γενικότερα η συμμετρία ή ασυμμετρία της κατανομής μπορεί να παρέχει πληροφορίες για την φύση και την κατανομή των δεδομένων για κάθε χαρακτηριστικό, βοηθώντας στην ανάλυση και την κατανόηση των διαφορών μεταξύ των δύο ομάδων.

Επομένως μπορούμε να δούμε πως διαφέρουν οι κατανομές των χαρακτηριστικών μεταξύ των δύο ομάδων. Για παράδειγμα, για το χαρακτηριστικό "age", παρατηρούμε ότι οι ζωντανοί (alive) έχουν ευρύτερη κατανομή στην μέση τιμή σε σχέση με τους νεκρούς (dead). Παρόμοιες παρατηρήσεις μπορούν να γίνουν για κάθε χαρακτηριστικό, βοηθώντας στην κατανόηση των διαφορών στις κατανομές μεταξύ των δύο ομάδων.



Εικόνα 9. Violin Plot

Στην εικόνα 10 παρατηρούμε έναν θερμικό χάρτη (heatmap) που απεικονίζει τον πίνακα συσχετίσεων (correlation matrix) μεταξύ των χαρακτηριστικών του συνόλου δεδομένων. Ο θερμικός χάρτης χρησιμοποιείται για να δείξει την δύναμη και την κατεύθυνση της συσχέτισης μεταξύ των ζευγών χαρακτηριστικών.

Τα χαρακτηριστικά του συνόλου δεδομένων αναγράφονται τόσο στον οριζόντιο (x-axis) όσο και στον κατακόρυφο άξονα (y-axis). Αυτά περιλαμβάνουν την ηλικία (age), την αναιμία (anaemia), την κρεατινίνη φωσφοκινάσης (creatinine_phosphokinase), τον διαβήτη (diabetes), το ποσοστό εξώθησης (ejection_fraction), την υπέρταση (high_blood_pressure), τα αιμοπετάλια (platelets), την κρεατινίνη (serum_creatinine), το νάτριο (serum_sodium), το φύλο (sex), το κάπνισμα (smoking), τον χρόνο (time), και το αποτέλεσμα θανάτου (DEATH_EVENT).

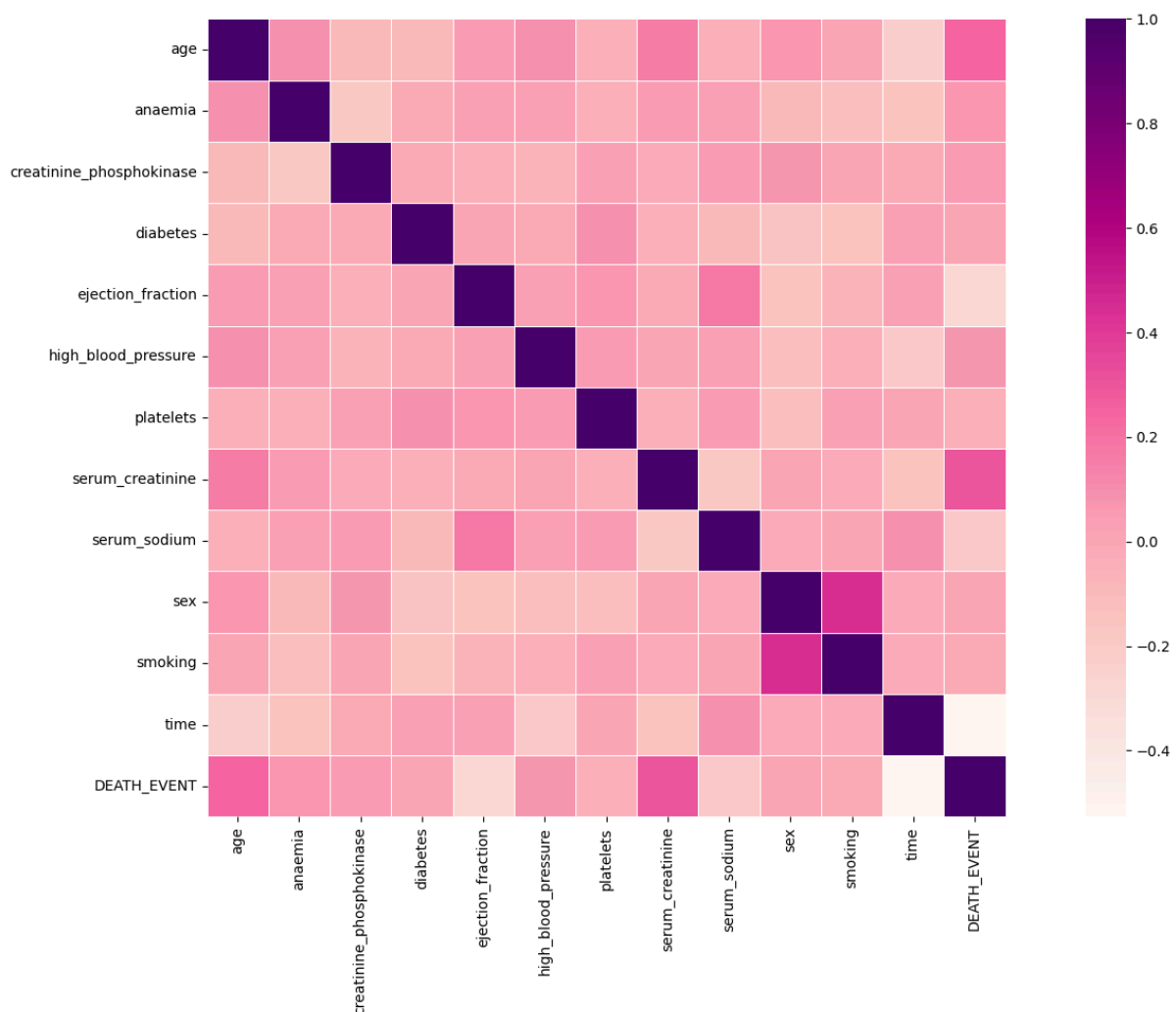
Η χρωματική κλίμακα στα δεξιά δείχνει την ένταση της συσχέτισης, από -1 έως 1. Τα σκούρα μοβ χρώματα υποδεικνύουν ισχυρές αρνητικές συσχετίσεις (κοντά στο -1), ενώ τα

σκούρα ροζ ή κόκκινα χρώματα δείχνουν ισχυρές θετικές συσχετίσεις (κοντά στο 1). Οι ανοιχτές αποχρώσεις δείχνουν χαμηλές ή μηδενικές συσχετίσεις (κοντά στο 0).

Κάθε κελί στον χάρτη δείχνει τη συσχέτιση μεταξύ του χαρακτηριστικού στη σειρά και του χαρακτηριστικού στη στήλη. Για παράδειγμα, το κελί στην τομή της στήλης "age" και της στήλης "DEATH_EVENT" δείχνει τη συσχέτιση μεταξύ ηλικίας και του γεγονότος θανάτου.

Αναλυτικότερα:

- ❖ Τα κελιά κατά μήκος της διαγωνίου είναι σκούρα μοβ και έχουν τιμή 1, καθώς κάθε χαρακτηριστικό είναι απόλυτα συσχετισμένο με τον εαυτό του.
- ❖ Οι θετικές συσχετίσεις παρουσιάζονται με πιο έντονα ροζ χρώματα, ενώ οι αρνητικές συσχετίσεις με πιο σκούρα μοβ χρώματα.
- ❖ Κάποιες συσχετίσεις είναι πιο ισχυρές από άλλες, π.χ., η σχέση μεταξύ του "ejection_fraction" και του "DEATH_EVENT" είναι αρνητική και σχετικά ισχυρή.



Εικόνα 10. Heatmap

4. Μεθοδολογία

4.1 Decision Trees

Τα δέντρα απόφασης είναι δομές που χωρίζουν επαναληπτικά το σύνολο δεδομένων σε υποσύνολα βάσει χαρακτηριστικών για να δημιουργήσουν ένα μοντέλο πρόβλεψης. Διαιρούν ουσιαστικά τα δεδομένα σε διακλαδώσεις βάσει κανόνων. Ο αλγόριθμος decision trees είναι βασικός αλγόριθμος και μπορεί να χρησιμοποιηθεί για ταξινόμηση και παλινδρόμηση. Υπάρχει και η πιθανότητα του overfitting (υπερ προσαρμογής), ειδικά αν το δέντρο είναι πολύ βαθύ ή αν δημιουργούνται πολύπλοκοι κανόνες, και αυτό μπορεί να μειώσει τη γενικευτική ικανότητα του μοντέλου. Προκειμένου να αντιμετωπιστεί αυτό το πρόβλημα, συχνά χρησιμοποιούνται τεχνικές όπως ο περιορισμός του βάθους του δέντρου ή η αποκοπή των φύλλων.

Σε κάθε βήμα, επιλέγεται το χαρακτηριστικό που διαχωρίζει καλύτερα τα δεδομένα σύμφωνα με μετρικές όπως Gini impurity ή Entropy και το επιλεγμένο χαρακτηριστικό χρησιμοποιείται για τη διαίρεση του συνόλου δεδομένων σε υποσύνολα. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να επιτευχθεί μια προκαθορισμένη ελάχιστη καθαρότητα (purity) ή μέχρι να μην υπάρχουν άλλες διακρίσιμες διακλαδώσεις.

Αναλυτικότερα η λειτουργία του αλγορίθμου είναι η εξής:

α) Διαίρεση Δεδομένων

Ξεκινώντας από τη ρίζα του δέντρου, τα δεδομένα διαιρούνται σε διακλαδώσεις βάσει κάποιων κριτηρίων. Το κάθε κριτήριο αντιπροσωπεύει μια ερώτηση που οδηγεί στον διαχωρισμό των δεδομένων σε υποσύνολα.

β) Επιλογή Κριτηρίων

Το κριτήριο που χρησιμοποιείται για να επιλεγεί η βέλτιστη διαίρεση είναι συνήθως μία μετρική που μετράει το πόσο καλά ο διαχωρισμός βοηθά στην αύξηση της ομοιότητας των δεδομένων μέσα στις νέες κλάδους. Για παράδειγμα, στην περίπτωση ταξινόμησης, αυτό μπορεί να είναι η "εντροπία" ή το "Gini impurity".

γ) Αναδρομική Διαδικασία

Η παραπάνω διαδικασία επαναλαμβάνεται για κάθε νέο υποσύνολο, δηλαδή δημιουργούνται νέα διακλαδώσεις με βάση τα κριτήρια που επιλέγονται.

δ) Τερματισμός

Η διαδικασία συνεχίζεται μέχρι να επιτευχθεί κάποιο κριτήριο τερματισμού. Αυτό μπορεί να είναι, για παράδειγμα, όταν ο αριθμός των δειγμάτων σε έναν κλάδο είναι πολύ μικρός ή όταν δεν υπάρχουν περισσότερες ερωτήσεις που θα μπορούσαν να διαχωρίσουν τα δεδομένα με βάση τα διαθέσιμα χαρακτηριστικά.

ε) Απόφαση

Κάθε φύλλο του δέντρου αντιπροσωπεύει μια απόφαση ή μια κλάση για τα δεδομένα που φτάνουν εκεί. Κατά την ταξινόμηση, το φύλλο αποδίδεται σε μια συγκεκριμένη κατηγορία, ενώ κατά την παλινδρόμηση μπορεί να προβλέπει μια τιμή.

4.2 Random Forest

Είναι ένα σύνολο από πολλά δέντρα αποφάσεων που εκπαιδεύονται σε διαφορετικά υποσύνολα των δεδομένων. Στη συνέχεια, συνδυάζουν τις προβλέψεις από τα δέντρα για να βελτιώσουν την απόδοση και τη γενικότητα.

Ο συνδυασμός πολλών δέντρων αποφάσεων βοηθά στη μείωση της τυχαιότητας και της υπερεκπαίδευσης που μπορεί να προκύψει από ένα μόνο δέντρο αποφάσεων. Επιπλέον, ο Random Forest μπορεί να αντιμετωπίσει και επιλύσει προβλήματα πολυκατηγορικής ταξινόμησης.

Η τυχαιότητα στην επιλογή των δειγμάτων και των χαρακτηριστικών για την εκπαίδευση και η συνδυαστική φύση των δέντρων αποφάσεων καθιστούν τον Random Forest έναν ισχυρό αλγόριθμο που παρέχει καλή απόδοση και γενίκευση σε ποικίλα προβλήματα μηχανικής μάθησης.

Αναλυτικότερα η λειτουργία του αλγορίθμου είναι η εξής:

α) Επιλογή Δειγμάτων

Επιλέγεται ένα τυχαίο υποσύνολο των δεδομένων εκπαίδευσης. Αυτό το υποσύνολο λέγεται "δείγμα εκπαίδευσης". Η τυχαιότητα είναι σημαντική για να δημιουργηθεί ποικιλία στα δέντρα και να αποφευχθεί η υπερεκπαίδευση.

β) Εκπαίδευση Δέντρων

Για κάθε δείγμα εκπαίδευσης, δημιουργείται ένα δέντρο αποφάσεων. Τα δέντρα αποφάσεων είναι δομές που λαμβάνουν αποφάσεις με βάση τις χαρακτηριστικές παραμέτρους των δεδομένων.

γ) Πρόβλεψη από Δέντρα

Αφού τα δέντρα εκπαιδευτούν, μπορούν να χρησιμοποιηθούν για πρόβλεψη για νέα δεδομένα. Κάθε δέντρο παράγει μια πρόβλεψη για το κάθε νέο δείγμα.

δ) Συνδυασμός Προβλέψεων

Οι προβλέψεις από όλα τα δέντρα συνδυάζονται. Στην περίπτωση της ταξινόμησης, συνήθως χρησιμοποιείται η πλειοψηφία των προβλέψεων ως την τελική πρόβλεψη.

4.3 Gradient Boosting

Ο Gradient Boosting είναι ένας αλγόριθμος boosting που συνδυάζει αδύναμους ταξινομητές, προσθέτοντας σταδιακά μοντέλα που διορθώνουν τα σφάλματα των προηγούμενων. Είναι μια τεχνική μηχανικής μάθησης για προβλήματα ταξινόμησης και παλινδρόμησης. Ουσιαστικά ξεκινάει με ένα μοντέλο (συνήθως απλό) και στη συνέχεια προσπαθεί να βελτιώσει την ακρίβεια των προβλέψεων προσθέτοντας νέα μοντέλα, τα οποία εστιάζονται στα σημεία όπου το προηγούμενο μοντέλο απέτυχε.

Εκπαιδεύει σειριακά δέντρα αποφάσεων, με κάθε δέντρο να προσπαθεί να διορθώσει τα λάθη του προηγούμενου. Είναι ιδιαίτερα αποτελεσματικός και συνήθως παρουσιάζει υψηλή απόδοση.

Ο στόχος του είναι να κατασκευάσει ένα συνδυαστικό μοντέλο, το οποίο αποτελείται από τη συνδυαστική εκτίμηση πολλών αδύναμων μοντέλων. Τα αδύναμα αυτά μοντέλα είναι συνήθως απλά δέντρα αποφάσεων.

Η διαδικασία ξεκινά με την κατασκευή ενός πρώτου αδύναμου μοντέλου, συνήθως ενός απλού δέντρου αποφάσεων, που προσπαθεί να επιλύσει το πρόβλημα. Στη συνέχεια, τα υπολείπονται σφάλματα αυτού του μοντέλου υπολογίζονται και χρησιμοποιούνται για να εκπαιδεύσουν το δεύτερο δέντρο, το οποίο προσπαθεί να διορθώσει τα λάθη του πρώτου.

Η διαδικασία συνεχίζεται, με κάθε νέο δέντρο προστίθεται στο συνδυαστικό μοντέλο προσπαθώντας να διορθώσει τα λάθη του προηγούμενου. Κατά την εκπαίδευση των δέντρων, τα βάρη των δειγμάτων προσαρμόζονται έτσι ώστε τα δέντρα που ακολουθούν να επικεντρώνονται περισσότερο στα δείγματα που προκαλούν τα μεγαλύτερα λάθη.

Το τελικό συνδυαστικό μοντέλο προκύπτει από το άθροισμα των προβλέψεων των αδύναμων μοντέλων, με κάθε πρόβλεψη να διορθώνεται σε κάθε επανάληψη με βάση το σφάλμα που παράχθηκε από το προηγούμενο δέντρο.

Οπότε συνοπτικά:

1. Ξεκινά με ένα απλό μοντέλο, συνήθως έναν σταθερό προβλεπτή
2. Υπολογίζονται τα σφάλματα των προβλέψεων (residuals)
3. Ένα νέο δέντρο απόφασης εκπαιδεύεται για να προβλέψει τα σφάλματα
4. Το νέο μοντέλο προστίθεται στο συνολικό με μικρό συντελεστή μάθησης (Learning rate)
5. Η διαδικασία επαναλαμβάνεται για έναν προκαθορισμένο αριθμό βημάτων

4.4 XGBoost (Extreme Gradient Boosting)

Ο XGBoost είναι μια βελτιστοποιημένη υλοποίηση του Gradient Boosting που είναι σχεδιασμένη να είναι εξαιρετικά αποδοτική και ισχυρή. Ανήκει στην κατηγορία των ensemble methods, και συγκεκριμένα των gradient boosting methods. Παρέχει βελτιστοποιημένη υλοποίηση, ενσωματώνει τη ρύθμιση υπερπαραμέτρων και είναι αποτελεσματικός σε πολλά προβλήματα.

Διαθέτει τα εξής κύρια χαρακτηριστικά:

α) Υποστήριξη της Παραγωγής Κόστους: Ο XGBoost μπορεί να διαχειριστεί περιπτώσεις όπου οι ανεπιθύμητες προβλέψεις έχουν διαφορετική επίπτωση κόστους. Με άλλα λόγια, μπορεί να αντιμετωπίσει ανισορροπίες στα δεδομένα.

β) Δέντρα Αποφάσεων Μικρού Βάθους: Ο αλγόριθμος δημιουργεί πολλά μικρά δέντρα αποφάσεων, γνωστά και ως "weak learners". Αυτά τα δέντρα έχουν μικρό βάθος, το οποίο προστατεύει από την υπερεκπαίδευση (overfitting) και τα καθιστά πιο απλά.

γ) Regularization (Ρύθμιση): Ο XGBoost χρησιμοποιεί μια στρατηγική ρύθμισης για τον περιορισμό της πολυπλοκότητας των μοντέλων. Αυτό βοηθάει στην αποφυγή της υπερεκπαίδευσης και στη βελτίωση της γενικοποίησης.

δ) Επεξεργασία Λείων Συναρτήσεων Κόστους (Smooth Loss Functions): Ο XGBoost χρησιμοποιεί λείες συναρτήσεις κόστους για την ελαχιστοποίηση των λαθών. Αυτό βοηθάει τον αλγόριθμο να συγκλίνει πιο γρήγορα και σε καλύτερα αποτελέσματα.

ε) Συνδυασμός Δέντρων: Το XGBoost συνδυάζει τα δέντρα που δημιουργεί με σκοπό να βελτιώσει την ακρίβεια και τη γενίκευση του μοντέλου.

στ) Βελτιστοποίηση Υπερπαραμέτρων: Ο XGBoost παρέχει ενσωματωμένα εργαλεία για την επιλογή και ρύθμιση των υπερπαραμέτρων του, κάνοντας τον εύκολο στη χρήση και αποτελεσματικό στην εύρεση των καλύτερων παραμέτρων για το συγκεκριμένο πρόβλημα.

Συνοπτικά η μεθοδολογία του:

1. Ξεκινά με ένα απλό μοντέλο
2. Υπολογίζονται τα σφάλματα των προβλέψεων (residuals)
3. Ένα νέο δέντρο απόφασης εκπαιδεύεται για να προβλέψει τα σφάλματα, χρησιμοποιώντας παραλληλισμό και τεχνικές regularization
4. Το νέο μοντέλο προστίθεται στο συνολικό μοντέλο με έναν μικρό συντελεστή μάθησης
5. Η διαδικασία επαναλαμβάνεται για έναν προκαθορισμένο αριθμό βημάτων

4.5 LightGBM

Ο LightGBM είναι βασισμένος στον Gradient Boosting που είναι σχεδιασμένος για να είναι γρήγορος και εξαιρετικά αποδοτικός, ειδικά για μεγάλα σύνολα δεδομένων. Χρησιμοποιεί έναν αλγόριθμο που βασίζεται σε ισορροπημένη διαίρεση φύλλων για να πετύχει καλύτερη απόδοση.

Χαρακτηριστικά του αποτελούν:

α) Υπολογιστική Αποδοτικότητα: Ο LightGBM χρησιμοποιεί την τεχνική του "Histogram-Based Learning", όπου τα δεδομένα χωρίζονται σε ιστογράμματα για γρηγορότερη επεξεργασία. Αυτό επιτρέπει την εκτέλεση των βασικών βημάτων του Gradient Boosting (όπως η εύρεση του βέλτιστου split) σε πολύ μικρότερο χρόνο.

β) Αναθεώρηση Φύλλων (Leaf-wise Splitting): Αντί να υπολογίζει σταθερό αριθμό διαμερισμάτων σε κάθε βήμα, ο LightGBM επιλέγει δυναμικά το κόμβο που θα διαιρεθεί βασισμένο στο ποιο κόμβο θα οδηγήσει στη μεγαλύτερη μείωση της συνάρτησης κόστους. Αυτό οδηγεί σε καλύτερη εκμετάλλευση των δεδομένων και γρηγορότερη σύγκλιση.

γ) Διαχείριση Κατηγορικών Μεταβλητών: Ο LightGBM μπορεί να χειριστεί κατηγορικές μεταβλητές χωρίς την ανάγκη για προεπεξεργασία όπως η μετατροπή τους σε αριθμητικές. Αυτό γίνεται μέσω της τεχνικής του "Συγχώνευσης Κατηγορικών Μεταβλητών" (Categorical Feature Split).

δ) Ενσωματωμένη ρύθμιση παραμέτρων: Ο LightGBM παρέχει αυτόματη ρύθμιση παραμέτρων μέσω της τεχνικής του "Optimal Split Finding". Αυτό σημαίνει ότι μπορεί να εκτελέσει έναν αυτοματοποιημένο εντοπισμό βέλτιστων παραμέτρων.

Συνοπτικά η μεθοδολογία του:

1. Ξεκινά με ένα απλό μοντέλο
2. Υπολογίζονται τα σφάλματα των προβλέψεων (residuals)
3. Χρησιμοποιούνται δέντρα απόφασης, με την εκπαίδευση να γίνεται με τεχνικές όπως Gradient-based One-Side Sampling (GOSS) και Exclusive Feature Bundling (EFB)
4. Το νέο μοντέλο προστίθεται στο συνολικό μοντέλο με έναν μικρό συντελεστή μάθησης
5. Η διαδικασία επαναλαμβάνεται για έναν προκαθορισμένο αριθμό βημάτων

4.6 AdaBoost

Ο AdaBoost είναι ένας αλγόριθμος boosting που δημιουργεί μια σειρά από αδύναμους ταξινομητές, συνήθως δέντρα αποφάσεων (weak classifiers), και τους συνδυάζει σε ένα ισχυρό μοντέλο. Κατά την εκπαίδευση, δίνεται μεγαλύτερη βαρύτητα στα λανθασμένα δείγματα, ώστε ο επόμενος "αδύναμος" μαθητής να επικεντρωθεί σε αυτά τα δείγματα. Τελικά, οι προβλέψεις των αδύναμων μαθητών συνδυάζονται για να δημιουργήσουν μια συνολική πρόβλεψη.

Συνοπτικά η μεθοδολογία του:

1. Δίνεται ίσο βάρος σε κάθε παρατήρηση στο σύνολο εκπαίδευσης
2. Ένας αδύναμος ταξινομητής εκπαιδεύεται στα δεδομένα και η ακρίβεια του αξιολογείται
3. Οι παρατηρήσεις που ταξινομήθηκαν λάθος λαμβάνουν αυξημένο βάρος, ενώ οι σωστές μειωμένο
4. Οι προβλέψεις των αδύναμων ταξινομητών συνδυάζονται, με τους ακριβέστερους ταξινομητές να λαμβάνουν μεγαλύτερη βαρύτητα

4.7 CatBoost

Ο CatBoost είναι ένας αλγόριθμος boosting που έχει σχεδιαστεί για να διαχειρίζεται κατηγορηματικά δεδομένα χωρίς την ανάγκη μετατροπής τους σε αριθμητικά χαρακτηριστικά. Βασίζεται σε δέντρα αποφάσεων και χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης και παλινδρόμησης. Μια από τις κύριες τεχνικές που χρησιμοποιεί είναι η "κατηγορική κωδικοποίηση" που επιτρέπει την αντιμετώπιση κατηγορικών χαρακτηριστικών χωρίς προεπεξεργασία.

4.8 Extra Tree

Ο αλγόριθμος ExtraTree είναι παρόμοιος με τα δέντρα απόφασης αλλά εισάγει περισσότερη τυχαιότητα για να βελτιώσει την ποικιλομορφία. Είναι μια παραλλαγή του αλγορίθμου Random Forest και η βασική διαφορά είναι ότι για τον καθορισμό του διαχωρισμού σε κάθε κόμβο του δέντρου, χρησιμοποιούνται τυχαίες τιμές για τα χαρακτηριστικά, αντί να επιλεγούν οι βέλτιστες τιμές.

Συνοπτικά η μεθοδολογία του:

1. Τυχαία Επιλογή Διακλαδώσεων: Σε κάθε κόμβο, επιλέγονται τυχαία κάποια χαρακτηριστικά και τυχαία σημεία διαχωρισμού (split points)
2. Δημιουργία Διακλαδώσεων: Οι τυχαίες διακλαδώσεις χρησιμοποιούνται για τον διαχωρισμό των δεδομένων
3. Επαναληψιμότητα: Η διαδικασία επαναλαμβάνεται μέχρι να επιτευχθούν οι συνθήκες τερματισμού

4.9 RusBoost

Ο αλγόριθμος RusBoost (Random Under-Sampling Boosting) συνδυάζει την τεχνική του undersampling με το AdaBoost για να αντιμετωπίσει το πρόβλημα των μη ισορροπημένων συνόλων δεδομένων.

Συνοπτικά η μεθοδολογία του:

1. Αρχικοποίηση: Δίνεται ίσο βάρος σε κάθε παρατήρηση στο σύνολο εκπαίδευσης
2. Δημιουργία Υποσυνόλων: Δείγματα από την κατηγορία με την πλειοψηφία αφαιρούνται τυχαία για να ισορροπήσουν το σύνολο δεδομένων
3. Εκπαίδευση Αδύναμων Ταξινομητών: Ένας αδύναμος ταξινομητής εκπαιδεύεται στα εξισορροπημένα δεδομένα και η ακρίβεια του αξιολογείται
4. Ενημέρωση Βαρών
5. Συνδυασμός Ταξινομητών: Οι προβλέψεις των αδύναμων ταξινομητών συνδυάζονται, με τους ακριβέστερους ταξινομητές να λαμβάνουν μεγαλύτερη βαρύτητα

4.10 Πειράματα

Στην παρούσα πτυχιακή εργασία επιλέχθηκαν οι εννέα (9) αλγόριθμοι που προηγήθηκαν. Η επιλογή έγινε από ένα πείραμα με το Pycaret ^[17]. Σε αυτό το πείραμα δοκιμάστηκαν πολλοί διαφορετικοί ταξινομητές και ελέγχθηκαν μέσω πληθώρας διαφορετικών μετρικών (π.χ accyracy, f1-score, precision, recall).

Για να στηθεί το πείραμα χρειάστηκε το εβδομήντα τοις εκατό (70%) του συνόλου δεδομένων (dataset) να χρησιμοποιηθεί για εκπαίδευση (training) και το τριάντα τοις εκατό (30%) ως “hold-out” για τον έλεγχο (testing). Ουσιαστικά το εβδομήντα τοις εκατό (70%) χρησιμοποιήθηκε για την αξιολόγηση των αλγορίθμων μέσω διασταυρωτικής αξιολόγησης (stratified k-fold cross validation). Ακολούθως απεικονίζεται η μορφή του πειράματος στην εικόνα 11.

	Description	Value
0	Session id	123
1	Target	DEATH_EVENT
2	Target type	Binary
3	Original data shape	(299, 13)
4	Transformed data shape	(299, 13)
5	Transformed train set shape	(209, 13)
6	Transformed test set shape	(90, 13)
7	Numeric features	12
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Fold Generator	StratifiedKFold
13	Fold Number	10
14	CPU Jobs	-1
15	Use GPU	False
16	Log Experiment	False
17	Experiment Name	clf-default-name
18	USI	262e

Εικόνα 11. Πείραμα pycaret

Έτσι λοιπόν η σύγκριση των μοντέλων η οποία και βρίσκεται στην εικόνα 12, αιτιολογεί την επιλογή των εννέα (9) αλγορίθμων που υλοποιήθηκαν. Σημειώνεται πως με την κίτρινη υπογράμμιση στην εικόνα φαίνονται οι καλύτερες τιμές ανά στήλη.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.8181	0.9258	0.6595	0.7976	0.6804	0.5635	0.5816	0.0660
catboost	CatBoost Classifier	0.8138	0.9096	0.6024	0.8167	0.6565	0.5415	0.5709	0.7450
et	Extra Trees Classifier	0.8088	0.8891	0.5548	0.8342	0.6383	0.5214	0.5524	0.0590
lightgbm	Light Gradient Boosting Machine	0.8040	0.8649	0.6119	0.7621	0.6562	0.5247	0.5465	0.0320
gbc	Gradient Boosting Classifier	0.7990	0.8701	0.6143	0.7835	0.6444	0.5138	0.5482	0.0330
xgboost	Extreme Gradient Boosting	0.7940	0.8791	0.6000	0.7538	0.6485	0.5076	0.5281	0.0220
lda	Linear Discriminant Analysis	0.7855	0.8719	0.6310	0.7076	0.6519	0.5013	0.5137	0.0100
lr	Logistic Regression	0.7805	0.8387	0.5452	0.6598	0.5766	0.4520	0.4699	0.8270
nb	Naive Bayes	0.7805	0.8355	0.4976	0.7083	0.5642	0.4393	0.4596	0.0090
ada	Ada Boost Classifier	0.7802	0.8264	0.5833	0.7251	0.6302	0.4778	0.4938	0.0320
ridge	Ridge Classifier	0.7712	0.0000	0.5857	0.6976	0.6190	0.4608	0.4771	0.0100
dt	Decision Tree Classifier	0.7557	0.7236	0.6381	0.6218	0.6204	0.4417	0.4499	0.0110
qda	Quadratic Discriminant Analysis	0.7464	0.7814	0.4619	0.6304	0.5106	0.3609	0.3775	0.0130
dummy	Dummy Classifier	0.6795	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0090
knn	K Neighbors Classifier	0.6217	0.5446	0.1452	0.3167	0.1811	-0.0119	-0.0101	0.0210
svm	SVM - Linear Kernel	0.6033	0.0000	0.2000	0.0619	0.0944	0.0000	0.0000	0.0140

Εικόνα 12. Σύγκριση Αλγορίθμων

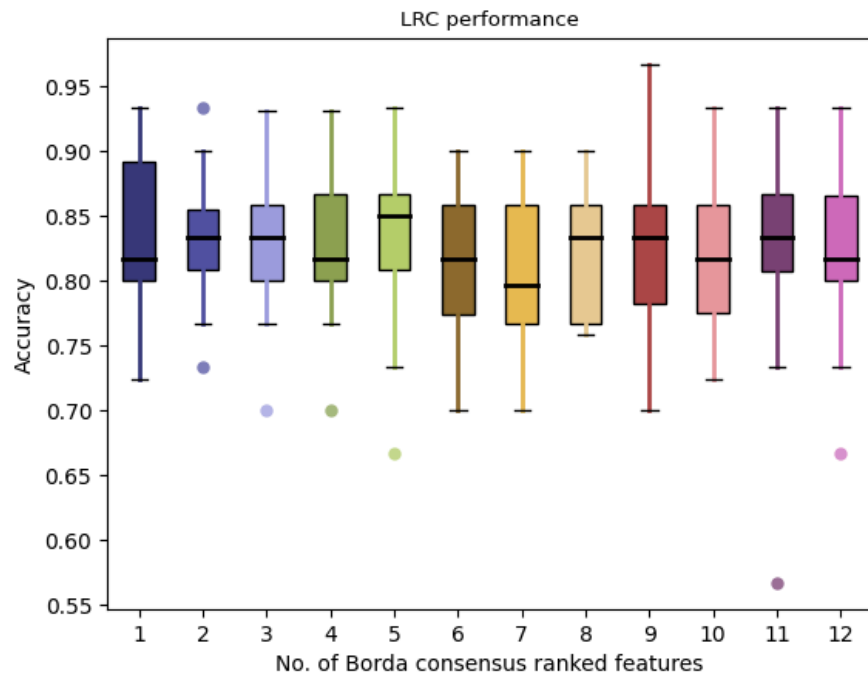
Όπως φάνηκε στην εικόνα 7, το σετ δεδομένων (dataset) που χρησιμοποιήθηκε στην παρούσα πτυχιακή, παρουσιάζει ένα σοβαρό πρόβλημα ανισορροπίας. Τα δείγματα για τους ανθρώπους που πέθαναν είναι ελάχιστα συγκριτικά με όσους έζησαν, αφού είναι λιγότεροι από το ένα τρίτο. Προκειμένου να ξεπεραστεί αυτό το πρόβλημα, εφαρμόστηκε “undersampling” μιας και πρόκειται για ζωντανούς ανθρώπους και έτσι “μίκρυνε” το σετ δεδομένων (dataset). Παράλληλα εφαρμόστηκε RFE (Recursive Feature Elimination), έτσι ώστε να πραγματοποιηθεί ένα feature selection και να απομακρυνθούν τα λιγότερο σημαντικά χαρακτηριστικά (εικόνα 8).

Για να επιτευχθεί η εύρεση της χρυσής τομής στο ερώτημα πόσα χαρακτηριστικά (features) θα κρατηθούν βοήθησαν τέσσερις (4) αλγόριθμοι και τρεις (3) διαφορετικές τιμές accuracy, F1 και ROC_AUC. Οι αλγόριθμοι ήταν:

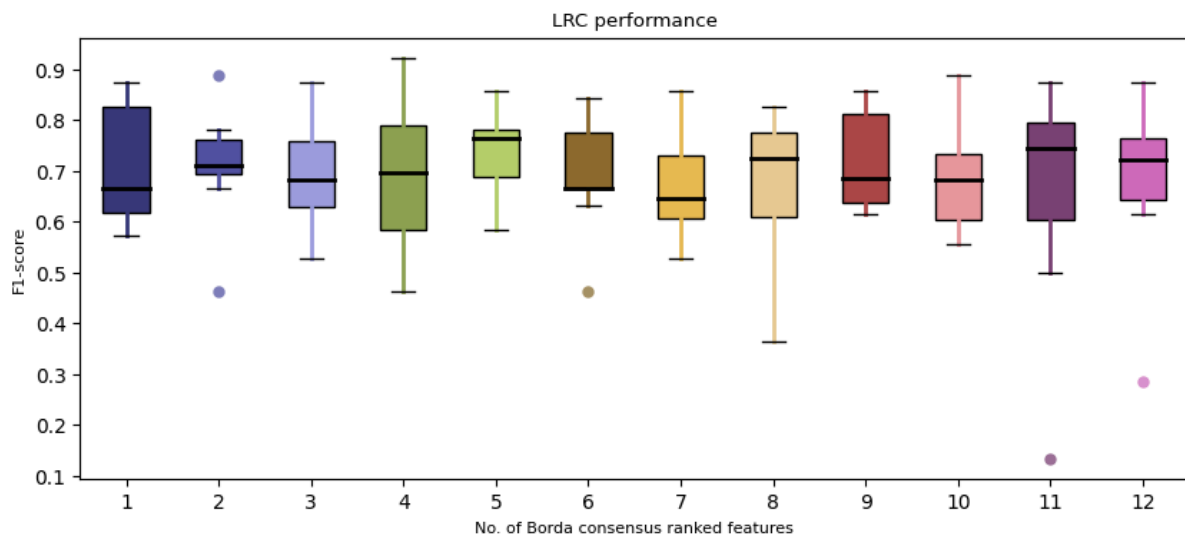
1. LRC (Logistic Regression Classifier)
2. LDA (Linear Discriminant Analysis)
3. DTC (Decision Tree Classifier)
4. RFC (Random Forest Classifier)

Στα γραφήματα που προέκυψαν παρατηρείται η διαφοροποίηση στις τιμές κάθε φορά που προστίθεται ένα χαρακτηριστικό. Στον οριζόντιο άξονα (x-axis) βρίσκεται ο αριθμός των χαρακτηριστικών και στον κατακόρυφο (y-axis) οι τιμές για τις μετρικές accuracy, F1 και ROC_AUC. Η μαύρη οριζόντια γραμμή σε κάθε ορθογώνιο πλαίσιο αποτελεί την διάμεσο.

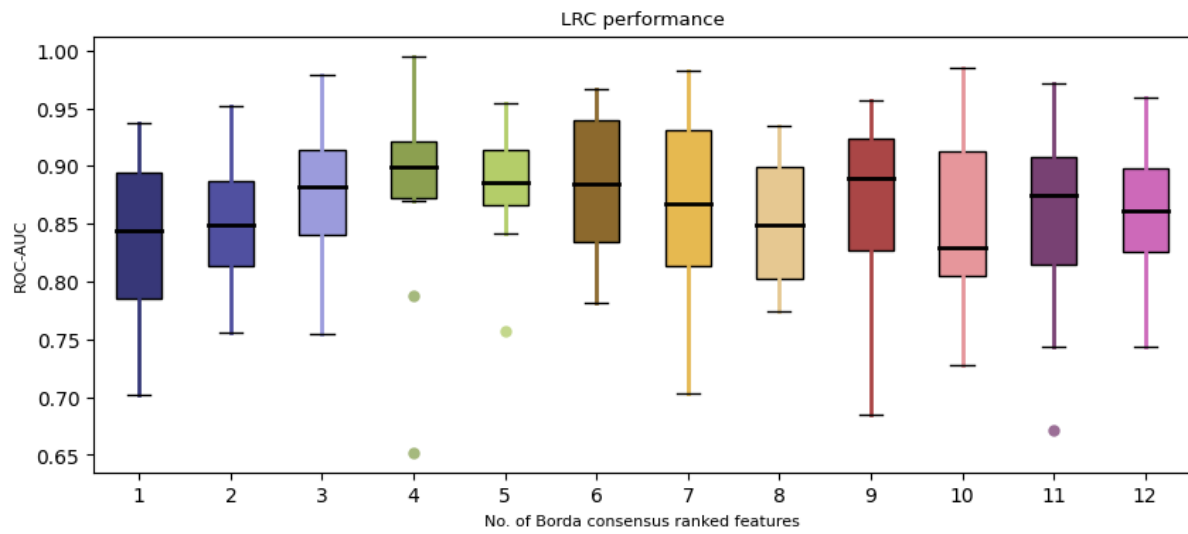
1. LRC (Logistic Regression Classifier)



Εικόνα 13. Accuracy - LRC

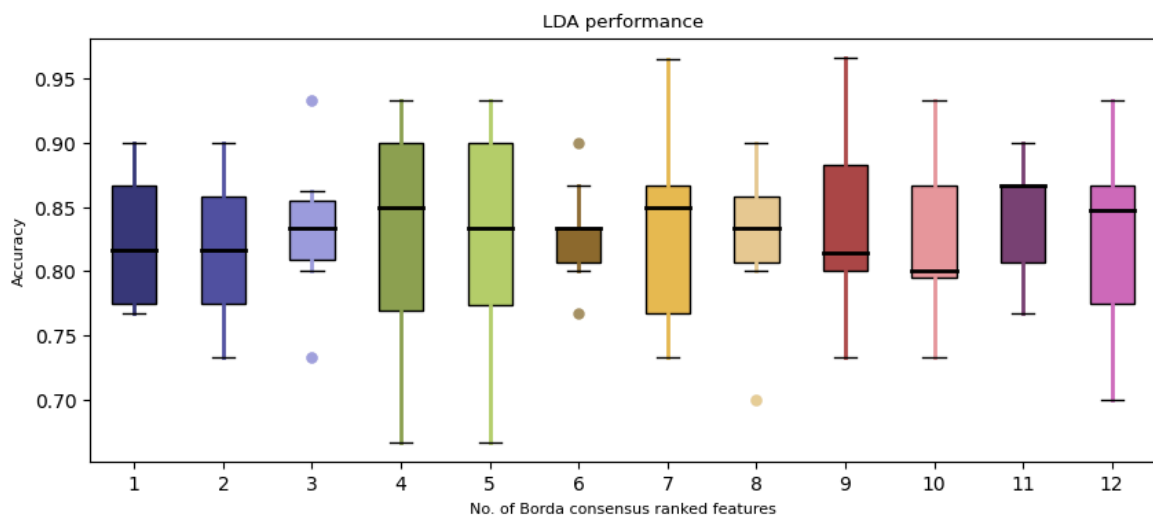


Εικόνα 14. F1 Score - LRC

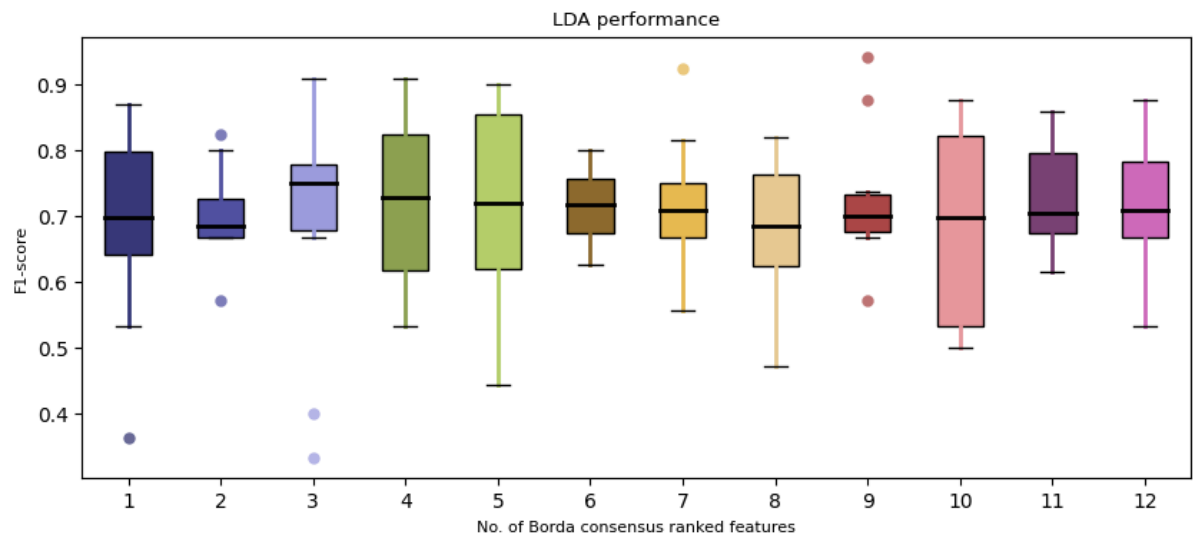


Εικόνα 15. ROC_AUC - LRC

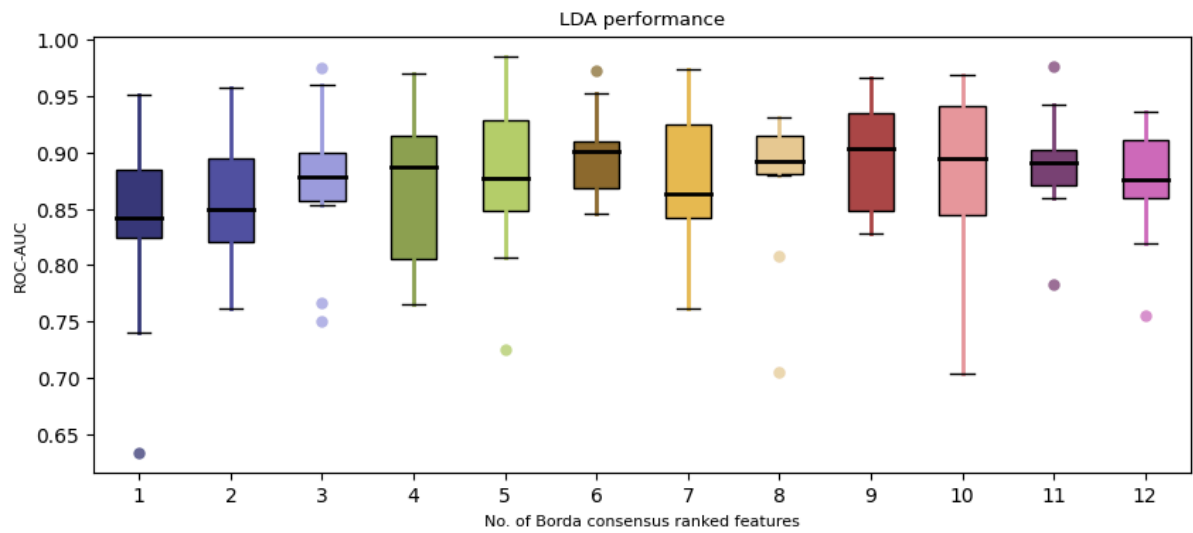
2. LDA (Linear Discriminant Analysis)



Εικόνα 16. Accuracy - LDA

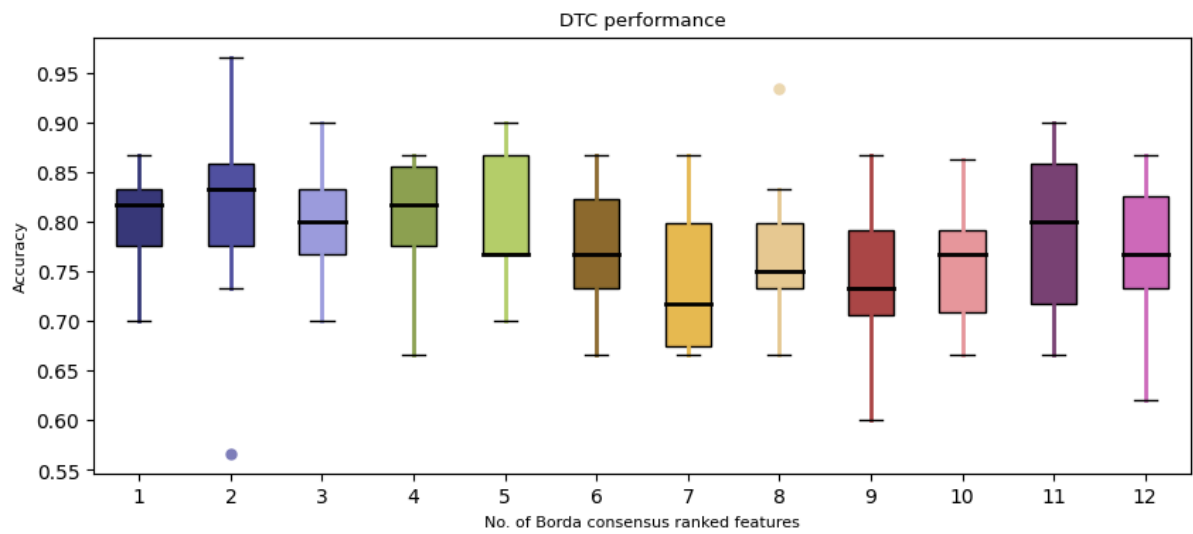


Εικόνα 17. F1 Score - LDA

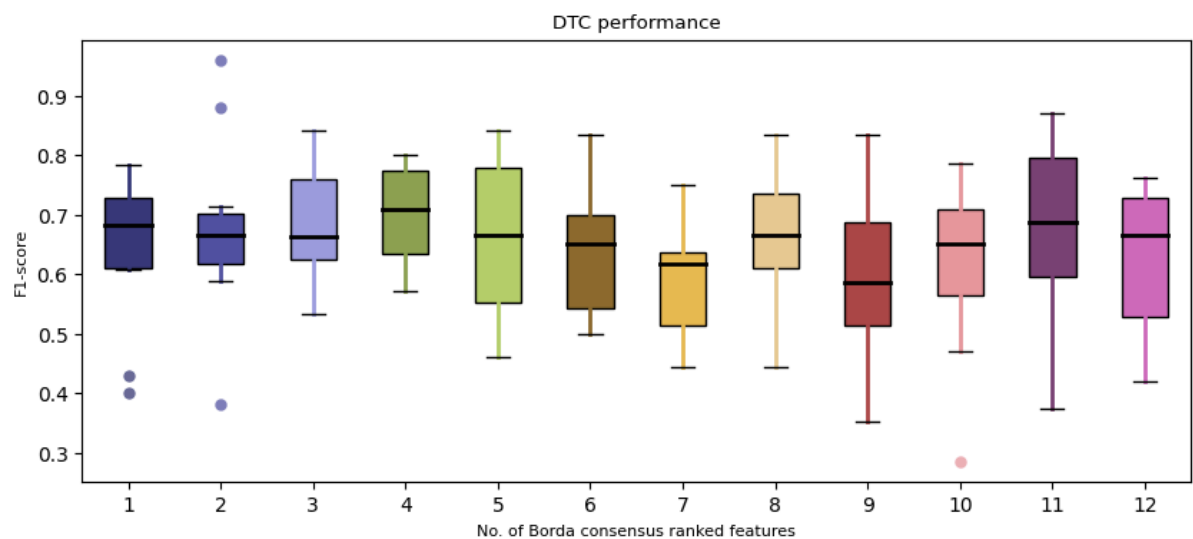


Εικόνα 18. ROC_AUC - LDA

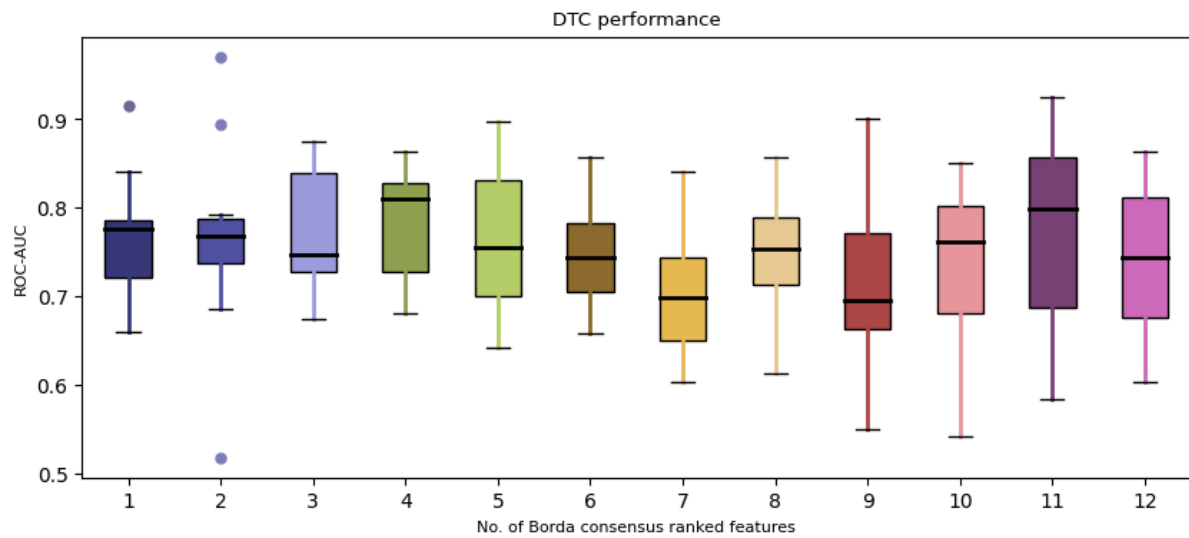
3. DTC (Decision Tree Classifier)



Εικόνα 19. Accuracy - DTC

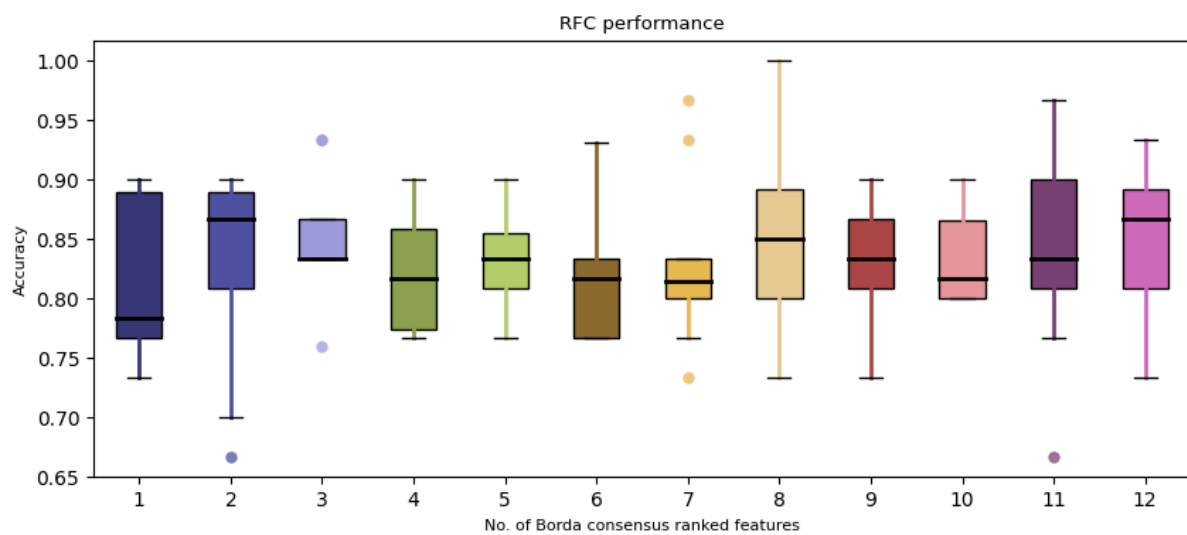


Εικόνα 20. F1 Score - DTC

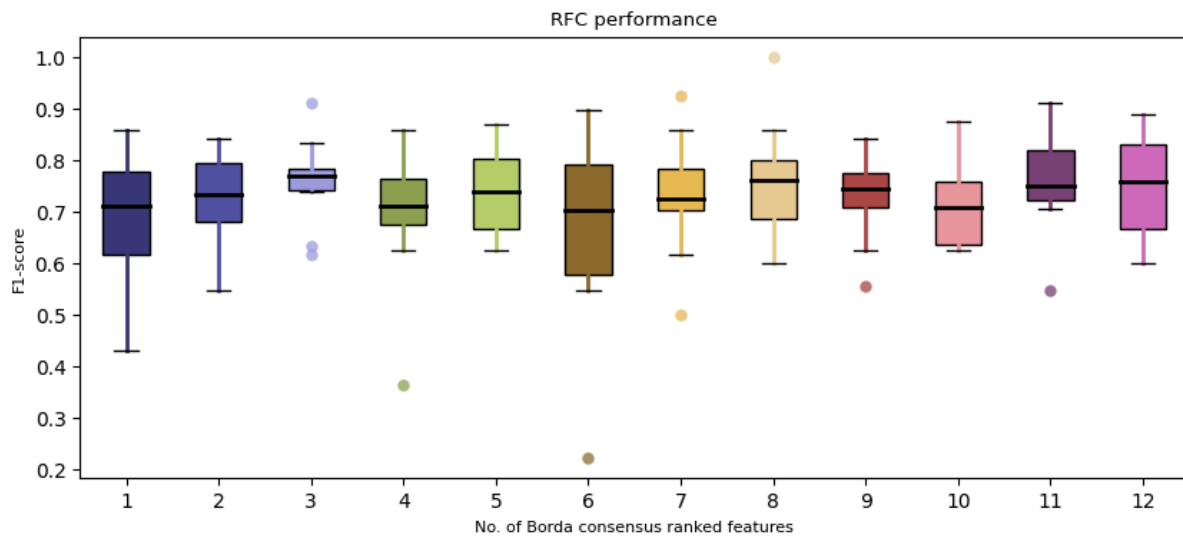


Εικόνα 21. ROC_AUC - DTC

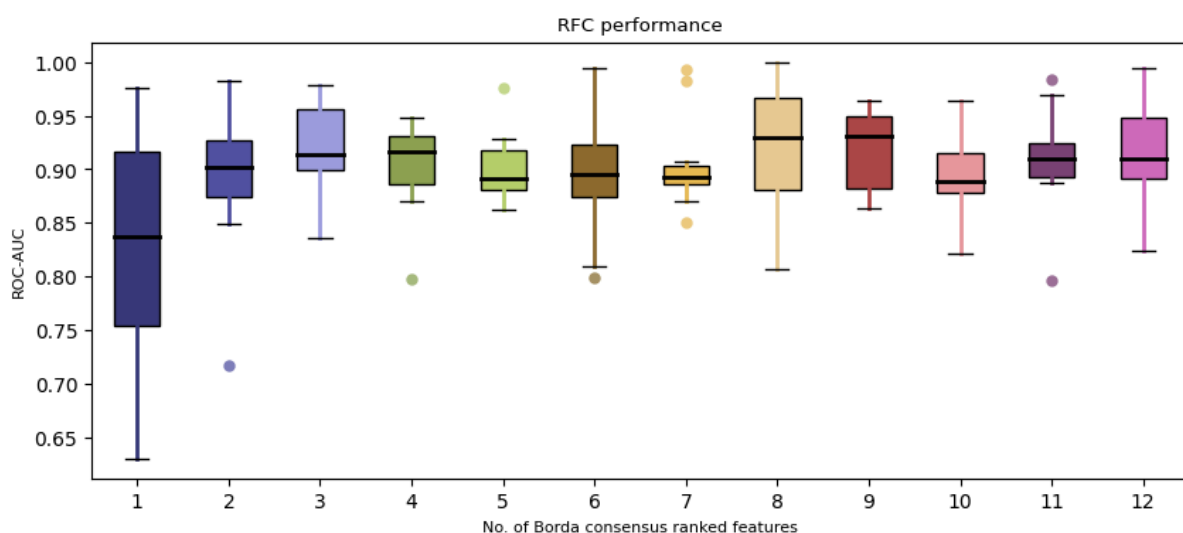
4. RFC (Random Forest Classifier)



Εικόνα 22. Accuracy - RFC



Εικόνα 23. F1 Score - RFC



Εικόνα 24. ROC_AUC - RFC

Οι εννέα (9) αλγόριθμοι υλοποιήθηκαν στο περιβάλλον ανάπτυξης Visual Studio και σε τέσσερα (4) διαφορετικά πειράματα, τα οποία απεικονίζονται στην εικόνα 25. Τα πειράματα χωρίστηκαν σύμφωνα με το σετ δεδομένων (dataset). Ουσιαστικά :

- 1ο πείραμα: Ολόκληρο το σετ δεδομένων (dataset) και αυτούσια features
- 2ο πείραμα: Ολόκληρο το σετ δεδομένων (dataset) και feature selection
- 3ο πείραμα: Balanced σετ δεδομένων (dataset) και αυτούσια features
- 4ο πείραμα: Balanced σετ δεδομένων (dataset) και feature selection

Σημειώνεται πως στα παραπάνω πειράματα όταν πραγματοποιείται feature selection από τα δώδεκα (12) χαρακτηριστικά κρατήθηκαν τα δέκα (10).



Εικόνα 25. Πειράματα

Κατά την υλοποίηση αυτών των αλγορίθμων, ο στόχος ήταν η εξέταση της απόδοσης τους σε διάφορα πειράματα και η αξιολόγηση των αποτελεσμάτων.

Τα πειράματα περιελάμβαναν την εκτέλεση των αλγορίθμων σε διαφορετικά σενάρια και τη σύγκριση των αποτελεσμάτων. Μέσω των πειραμάτων, παρατηρήθηκε η συμπεριφορά των αλγορίθμων σε διαφορετικές συνθήκες και οι επιδόσεις τους αξιολογήθηκαν με βάση τους προκαθορισμένους στόχους. Η σύγκριση των αλγορίθμων ανέδειξε τα πλεονεκτήματα και τα μειονεκτήματά τους σε σχέση με την απόδοση και την αποτελεσματικότητα.

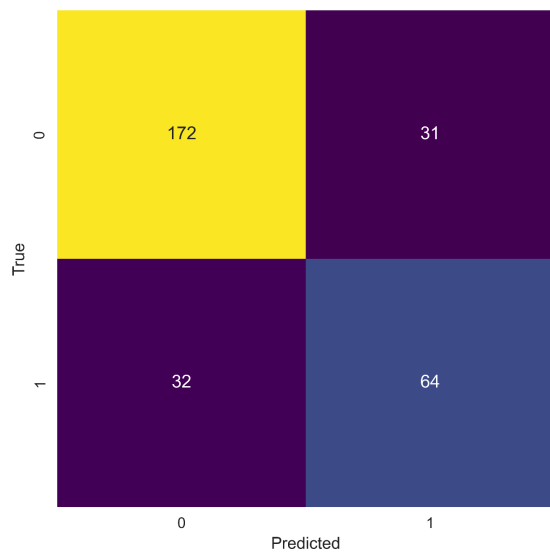
Συνολικά, τα αποτελέσματα των πειραμάτων παρέχουν σημαντική ενημέρωση για την επιλογή του κατάλληλου αλγορίθμου για συγκεκριμένες εφαρμογές και προβλήματα.

5. Αποτελέσματα

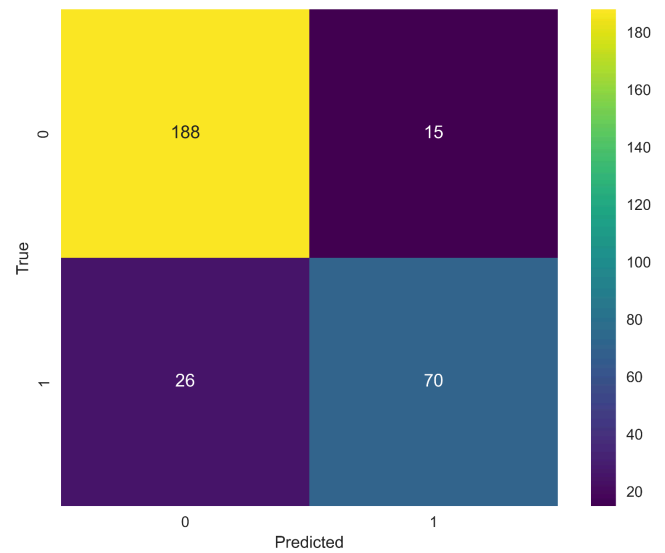
Στο πρώτο πείραμα όπως προαναφέρθηκε, το σετ δεδομένων (dataset) ήταν ολόκληρο και αυτούσια τα χαρακτηριστικά (features) δηλαδή και τα δώδεκα (12). Στην εικόνα 26 απεικονίζονται τα αποτελέσματα συγκεντρωτικά και έπειτα ακολουθούν οι πίνακες σύγχυσης (confusion matrix).

Algorithm		Random Forest	AdaBoost	CatBoost	Decision Tree	Extra Tree	Gradient Boosting	RusBoost	XGB	LGBM
Metric	Accuracy	0,86	0,82	0,83	0,78	0,79	0,82	0,78	0,84	0,86
	F1 Score	0,77	0,71	0,71	0,66	0,62	0,71	0,66	0,74	0,78
	ROC_AUC	0,91	0,84	0,92	0,77	0,89	0,9	0,85	0,91	0,91
Standard Deviation		0,08	0,07	0,07	0,07	0,08	0,04	0,06	0,07	0,05
		0,13	0,1	0,12	0,1	0,12	0,08	0,09	0,11	0,09
		0,06	0,08	0,05	0,08	0,06	0,05	0,07	0,05	0,05

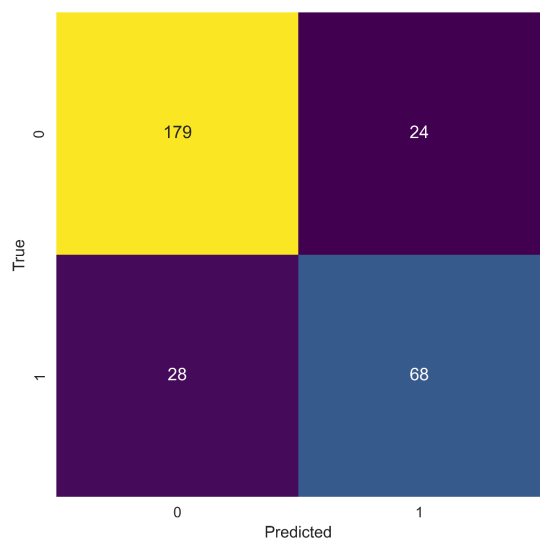
Εικόνα 26. Αποτελέσματα Πειράματος 1



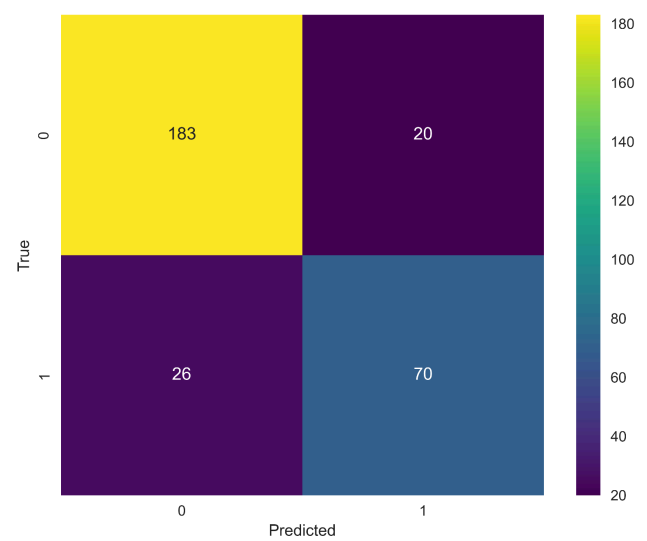
Εικόνα 27. Confusion Matrix Decision Trees 1



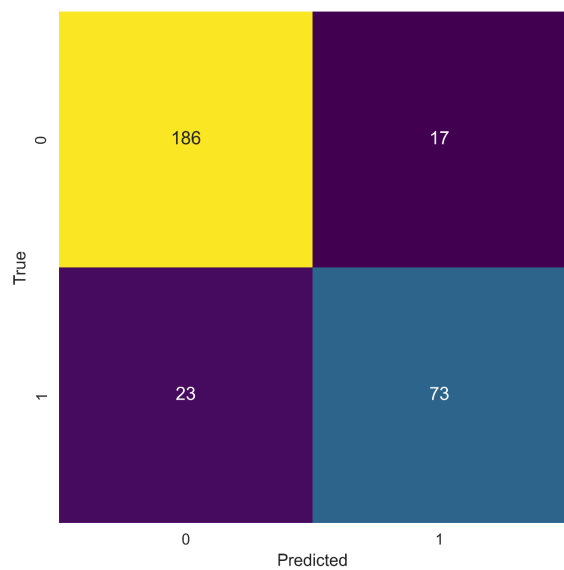
Εικόνα 28. Confusion Matrix Random Forest 1



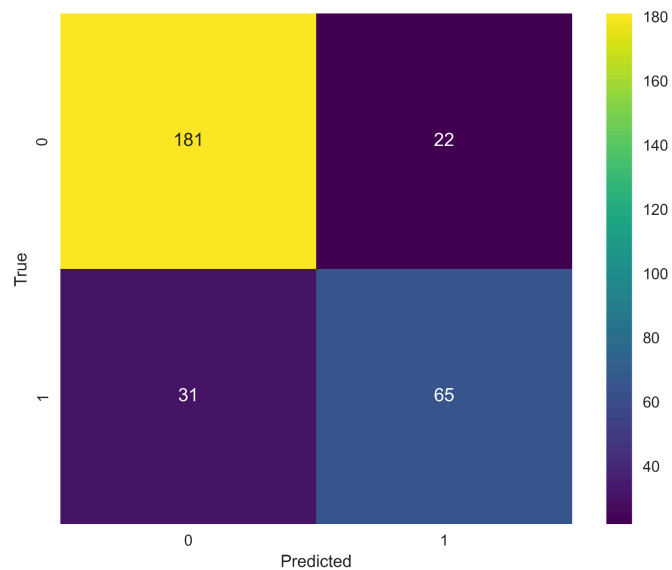
Εικόνα 29. Confusion Matrix Gradient Boosting 1



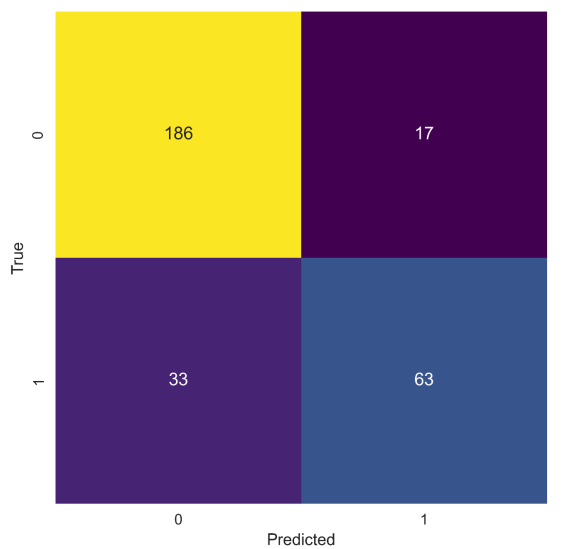
Εικόνα 30. Confusion Matrix XGBoost 1



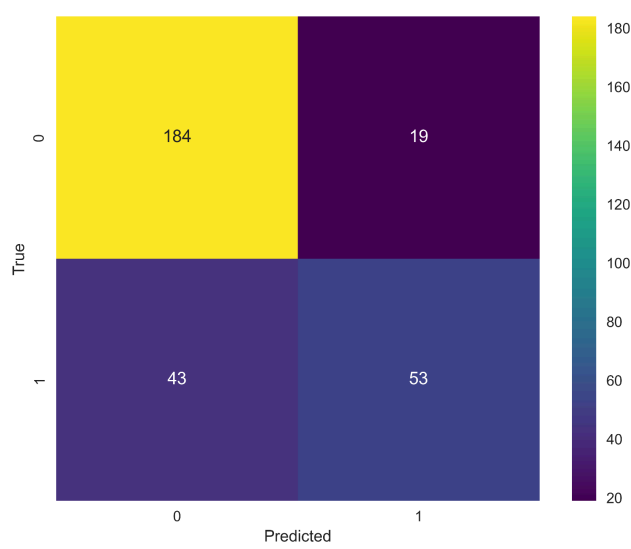
Εικόνα 31. Confusion Matrix LightGBM 1



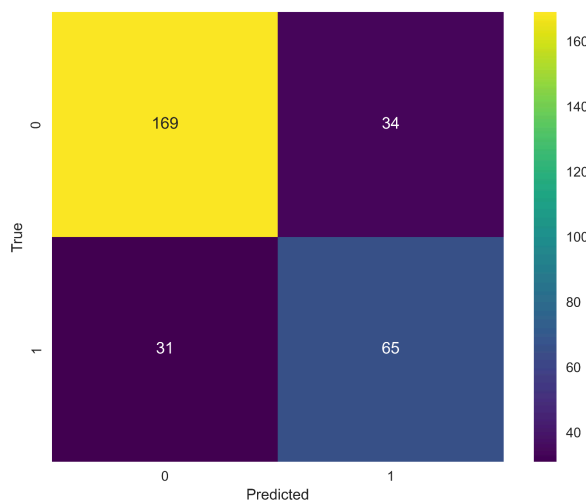
Εικόνα 32. Confusion Matrix AdaBoost 1



Εικόνα 33. Confusion Matrix CatBoost 1



Εικόνα 34. Confusion Matrix Extra Tree 1

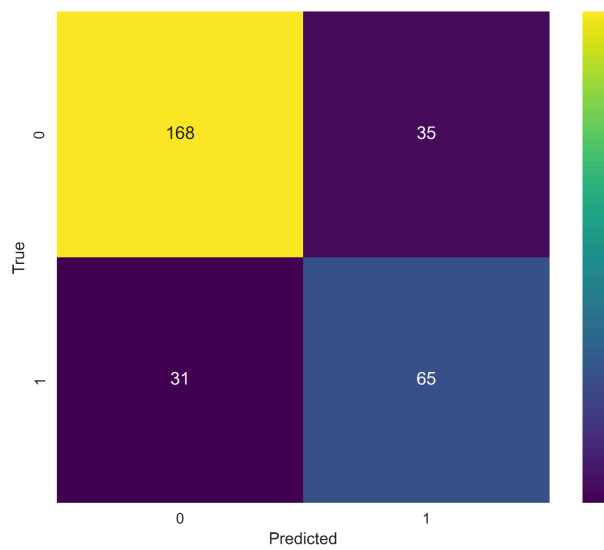


Εικόνα 35. Confusion Matrix RusBoost 1

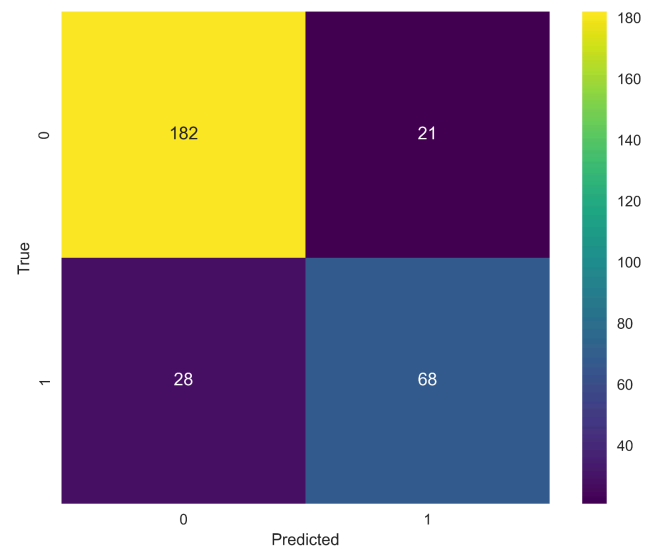
Στο δεύτερο πείραμα όπως προαναφέρθηκε, το σετ δεδομένων (dataset) ήταν ολόκληρο και στα χαρακτηριστικά (features) πραγματοποιήθηκε feature selection. Στην εικόνα 36 απεικονίζονται τα αποτελέσματα συγκεντρωτικά και έπειτα ακολουθούν οι πίνακες σύγχυσης (confusion matrix) για το συγκεκριμένο πείραμα.

Algorithm		Random Forest	AdaBoost	CatBoost	Decision Tree	Extra Tree	Gradient Boosting	RusBoost	XGB	LGBM
Metric	Accuracy	0,83	0,84	0,83	0,77	0,81	0,83	0,79	0,83	0,85
	F1 Score	0,73	0,73	0,71	0,66	0,66	0,73	0,68	0,72	0,76
	ROC_AUC	0,91	0,91	0,92	0,76	0,9	0,9	0,83	0,91	0,91
Standard Deviation		0,06	0,07	0,07	0,11	0,07	0,05	0,04	0,06	0,06
		0,11	0,12	0,12	0,14	0,14	0,1	0,06	0,12	0,11
		0,06	0,06	0,05	0,12	0,06	0,05	0,06	0,05	0,05

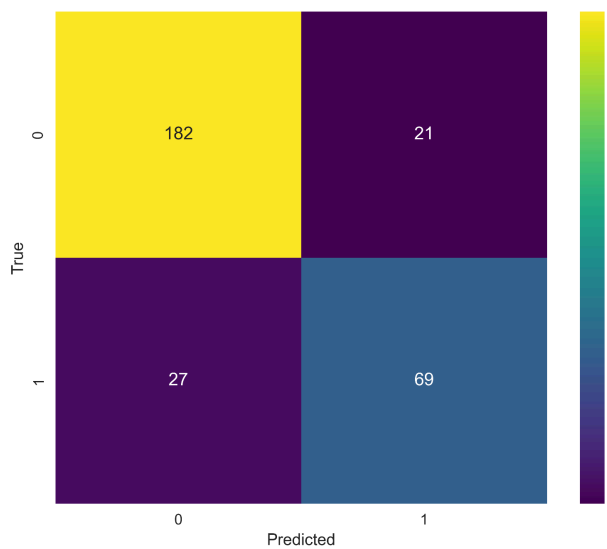
Εικόνα 36. Αποτελέσματα Πειράματος 2



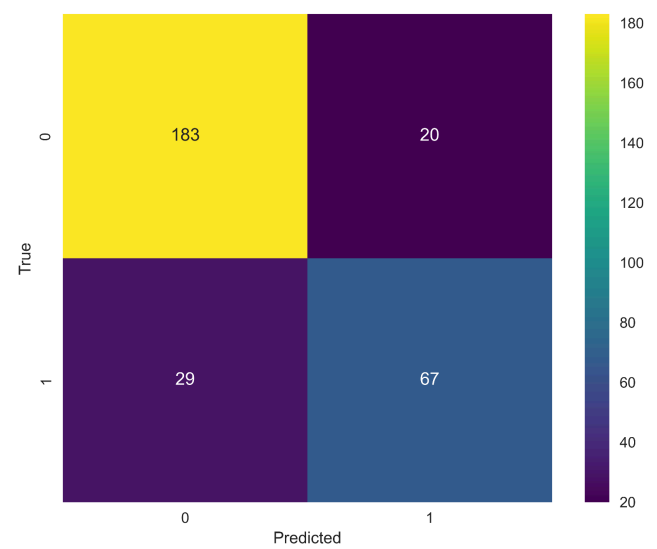
Εικόνα 37. Confusion Matrix Decision Tree 2



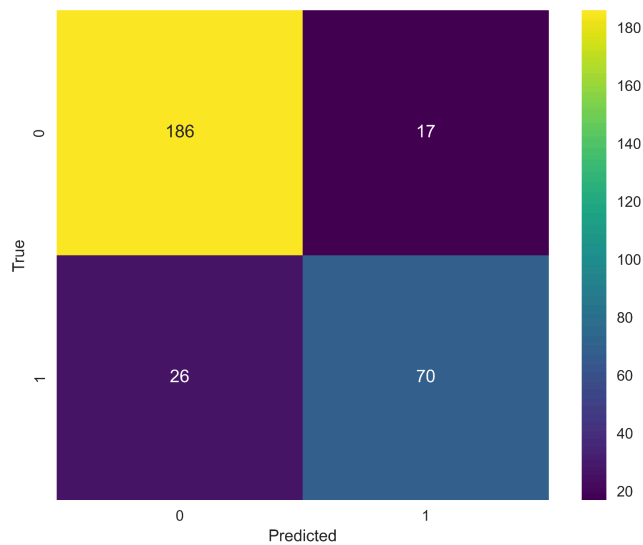
Εικόνα 38. Confusion Matrix Random Forest 2



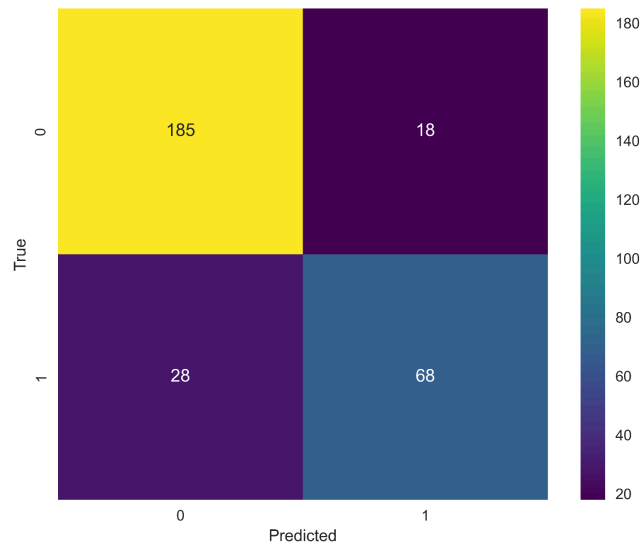
Εικόνα 39. Confusion Matrix Gradient Boosting 2



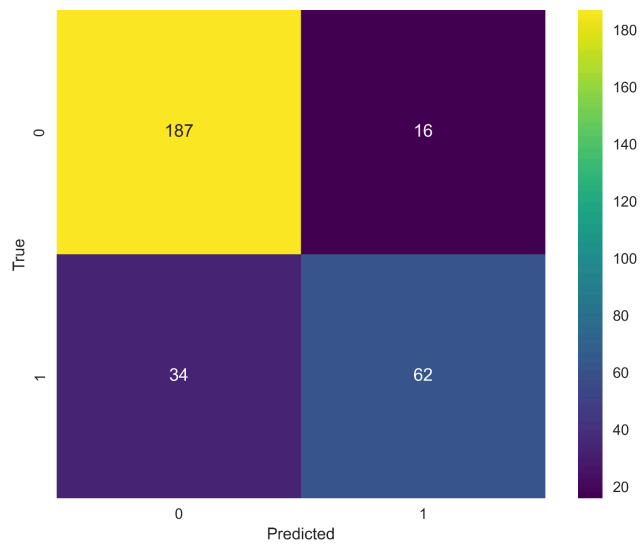
Εικόνα 40. Confusion Matrix XGBoost 2



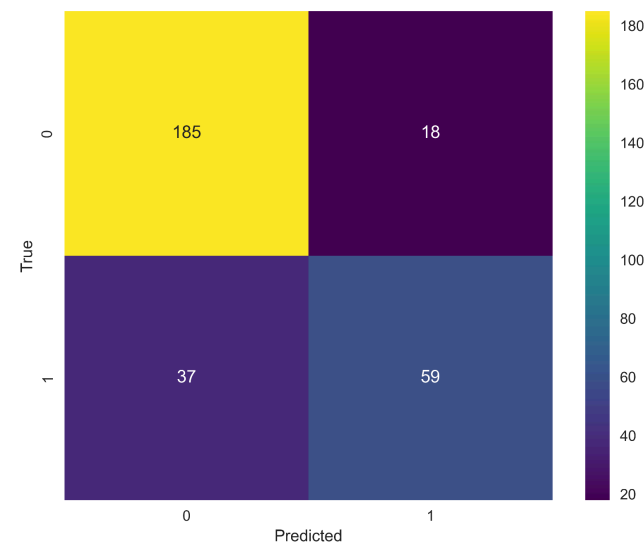
Εικόνα 41. Confusion Matrix LightGBM 2



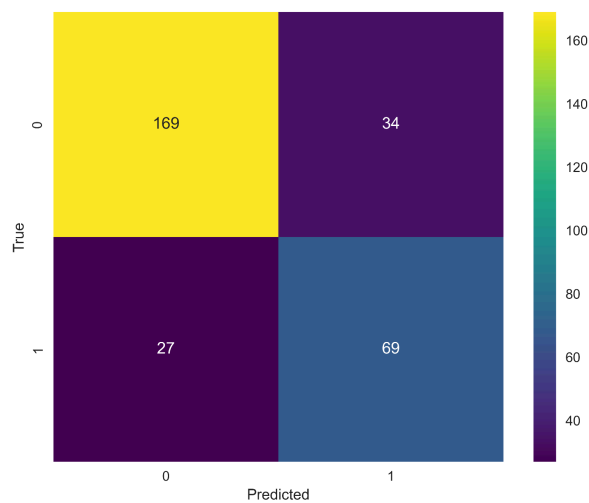
Εικόνα 42. Confusion Matrix AdaBoost 2



Εικόνα 43. Confusion Matrix CatBoost 2



Εικόνα 44. Confusion Matrix Extra Tree 2

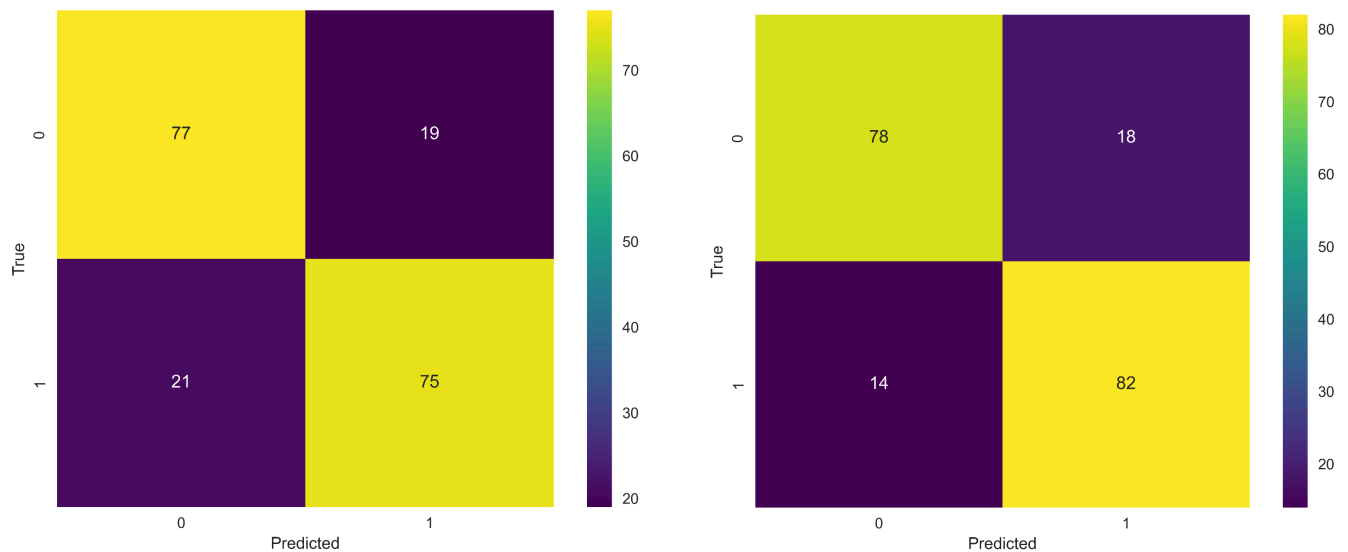


Εικόνα 45. Confusion Matrix RusBoost 2

Στο τρίτο πείραμα όπως προαναφέρθηκε, το σετ δεδομένων (dataset) ήταν ισορροπημένο (balanced) με εκατόν ενενήντα δύο (192) δείγματα και αυτούσια παρέμειναν τα χαρακτηριστικά (features) δηλαδή και δώδεκα (12) διατηρήθηκαν. Στην εικόνα 46 απεικονίζονται τα αποτελέσματα συγκεντρωτικά και έπειτα ακολουθούν οι πίνακες σύγχυσης (confusion matrix) για το συγκεκριμένο πείραμα.

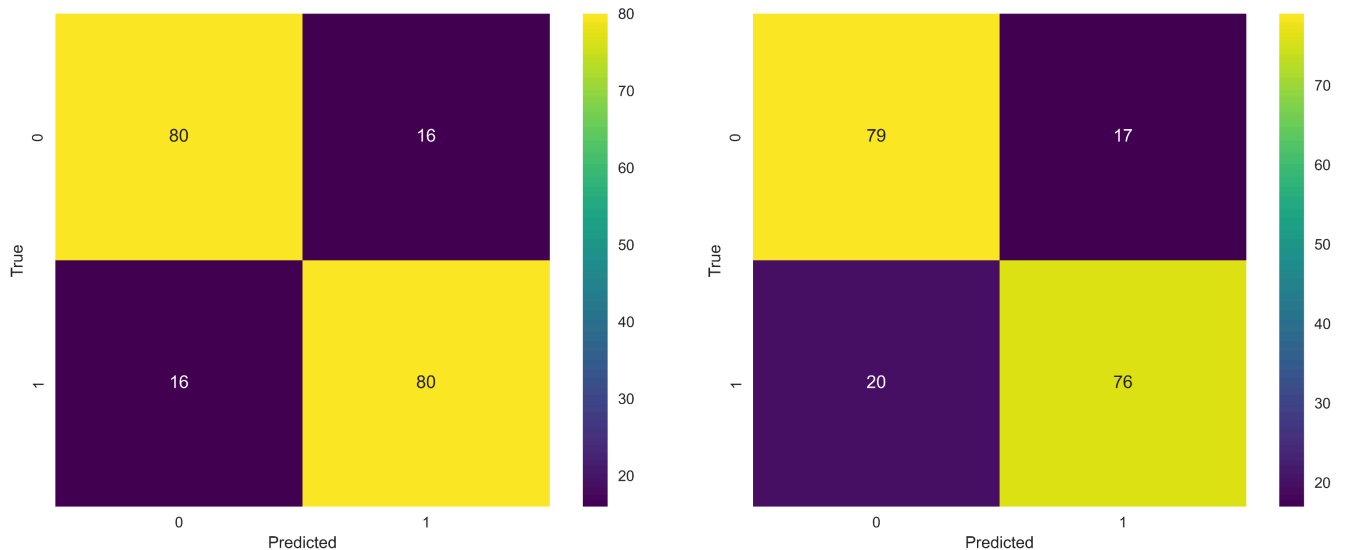
Algorithm		Random Forest	AdaBoost	CatBoost	Decision Tree	Extra Tree	Gradient Boosting	RusBoost	XGB	LGBM
Metric	Accuracy	0,83	0,78	0,83	0,79	0,8	0,83	0,79	0,8	0,79
	F1 Score	0,82	0,75	0,81	0,77	0,78	0,81	0,77	0,79	0,78
	ROC_AUC	0,92	0,83	0,92	0,78	0,9	0,91	0,82	0,9	0,89
Standard Deviation		0,08	0,07	0,05	0,07	0,09	0,06	0,06	0,09	0,08
		0,1	0,12	0,08	0,1	0,11	0,09	0,09	0,11	0,1
		0,06	0,05	0,05	0,07	0,06	0,06	0,05	0,06	0,08

Εικόνα 46. Αποτελέσματα Πειράματος 3



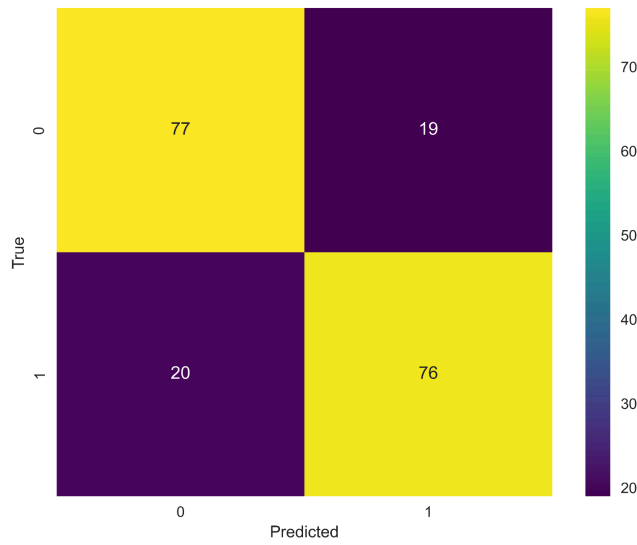
Εικόνα 47. Confusion Matrix Decision Trees 3

Εικόνα 48. Confusion Matrix Random Forest 3

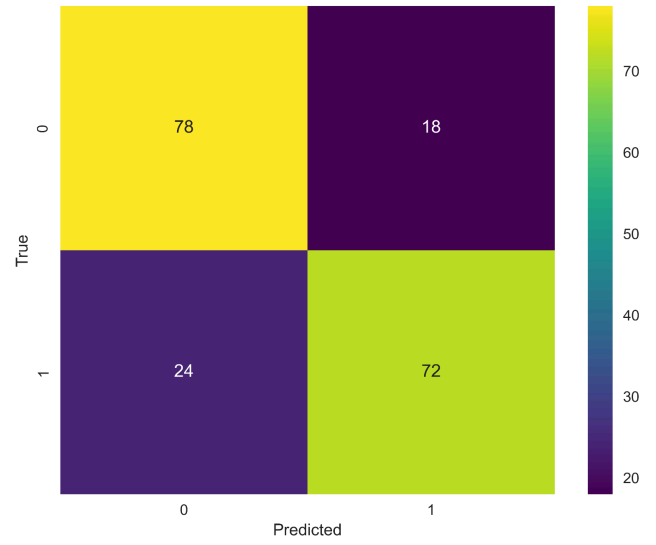


Εικόνα 49. Confusion Matrix Gradient Boosting 3

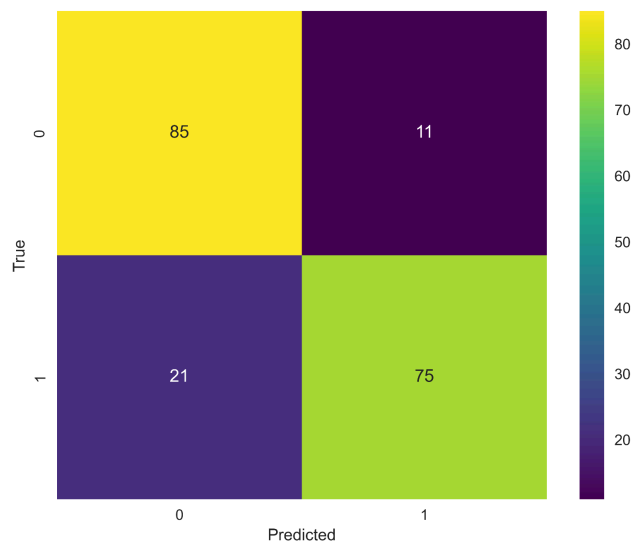
Εικόνα 50. Confusion Matrix XGBoost 3



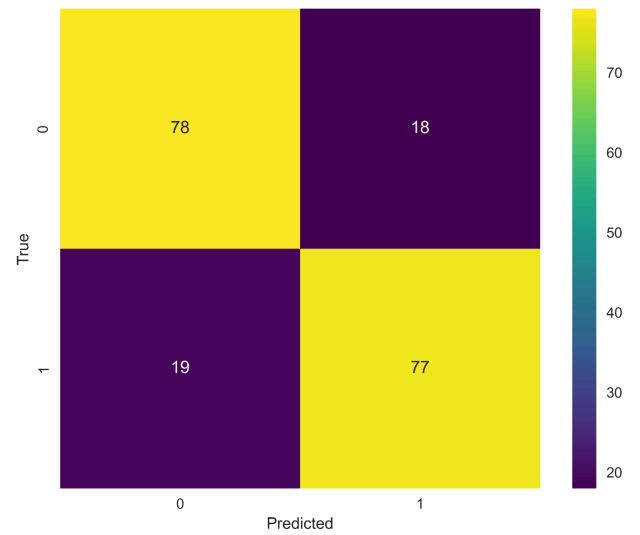
Εικόνα 51. Confusion Matrix LightGBM 3



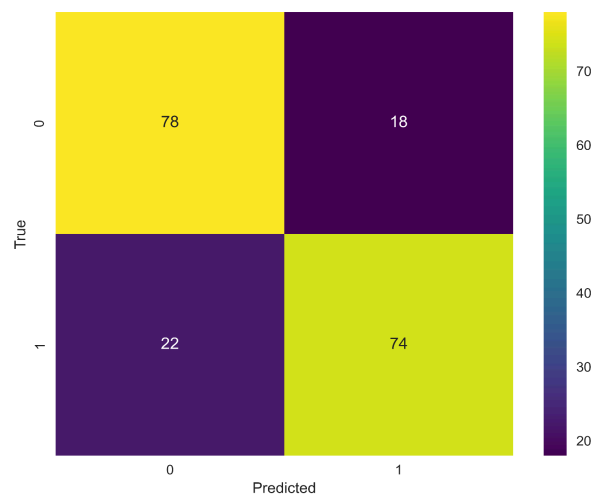
Εικόνα 52. Confusion Matrix AdaBoost 3



Εικόνα 53. Confusion Matrix CatBoost 3



Εικόνα 54. Confusion Matrix Extra Tree 3

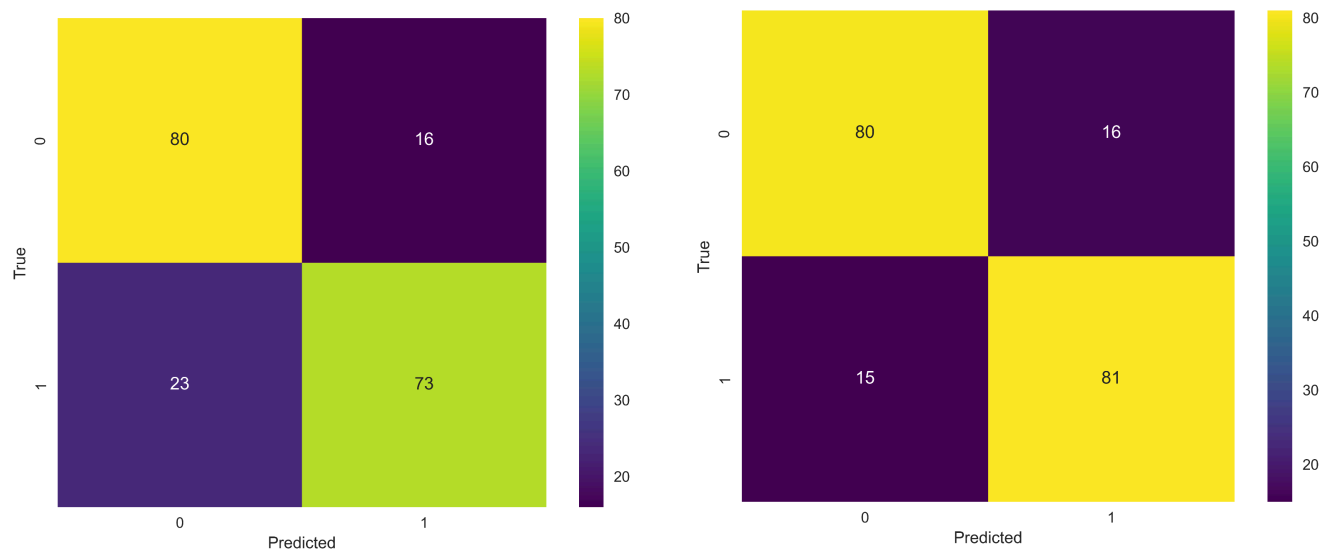


Εικόνα 55. Confusion Matrix RusBoost 3

Στο τέταρτο πείραμα όπως προαναφέρθηκε, το σετ δεδομένων (dataset) ήταν ισορροπημένο (balanced) με εκατόν ενενήντα δύο (192) δείγματα και στα χαρακτηριστικά (features) πραγματοποιήθηκε feature selection. Στην εικόνα 56 απεικονίζονται τα αποτελέσματα συγκεντρωτικά και έπειτα ακολουθούν οι πίνακες σύγχυσης (confusion matrix) για το συγκεκριμένο πείραμα.

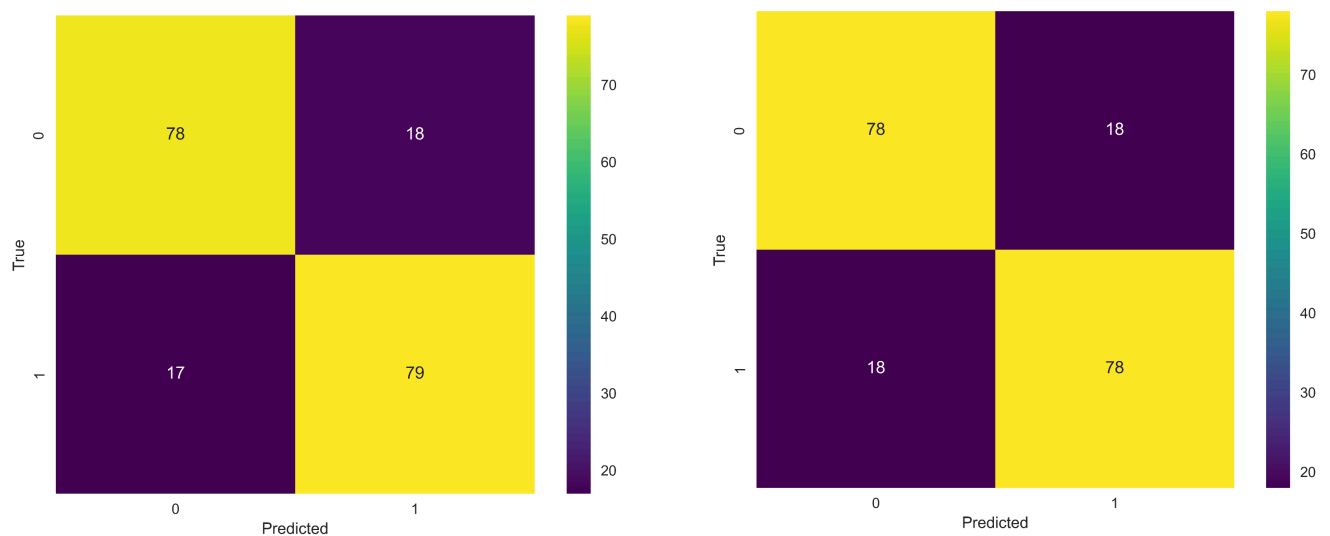
Algorithm		Random Forest	AdaBoost	CatBoost	Decision Tree	Extra Tree	Gradient Boosting	RusBoost	XGB	LGBM
Metric	Accuracy	0,83	0,77	0,83	0,79	0,8	0,81	0,78	0,81	0,8
	F1 Score	0,82	0,74	0,82	0,78	0,78	0,8	0,76	0,8	0,78
	ROC_AUC	0,92	0,84	0,92	0,79	0,91	0,9	0,83	0,89	0,91
Standard Deviation		0,09	0,07	0,04	0,06	0,08	0,07	0,03	0,09	0,08
		0,11	0,12	0,07	0,07	0,12	0,1	0,07	0,1	0,11
		0,06	0,06	0,05	0,06	0,06	0,07	0,05	0,07	0,07

Εικόνα 56. Αποτελέσματα Πειράματος 4



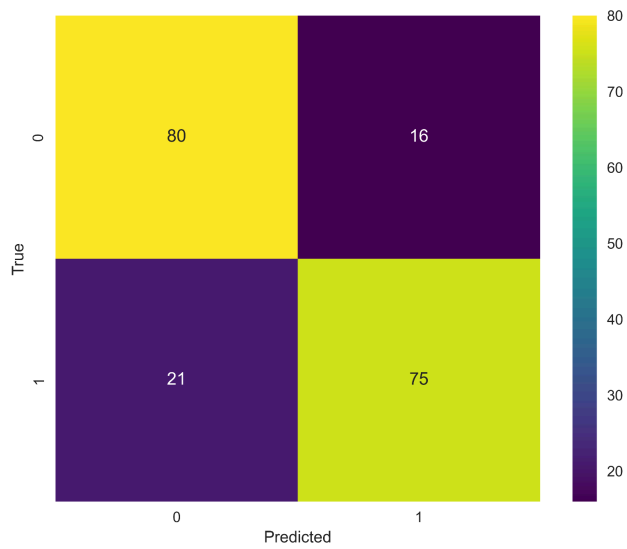
Εικόνα 57. Confusion Matrix Decision Trees 4

Εικόνα 58. Confusion Matrix Random Forest 4

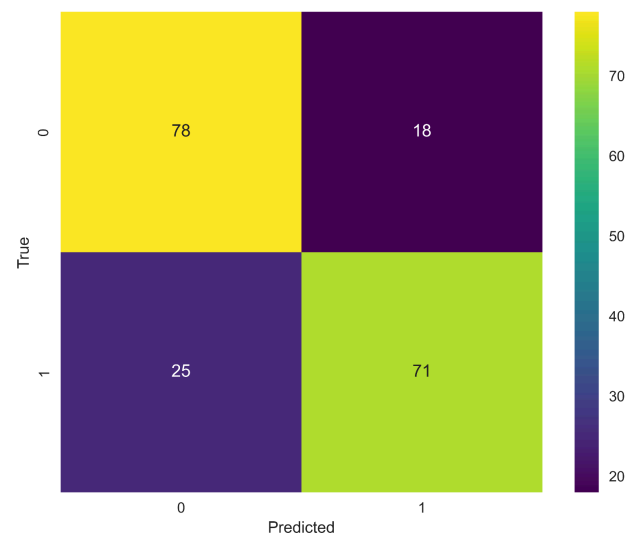


Εικόνα 59. Confusion Matrix Gradient Boosting 4

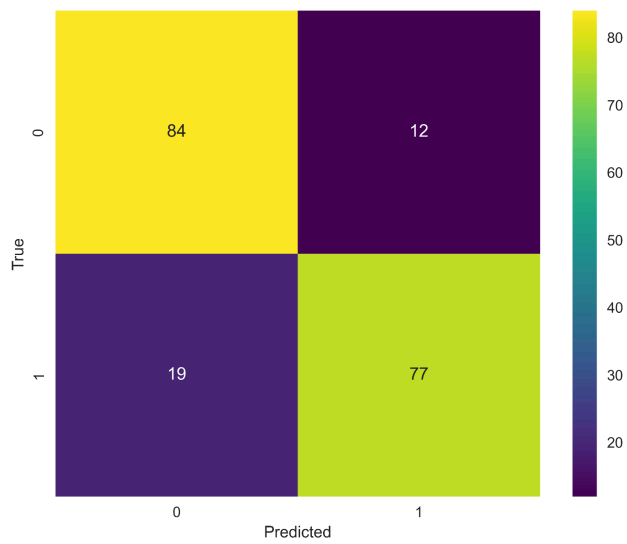
Εικόνα 60. Confusion Matrix XGBoost 4



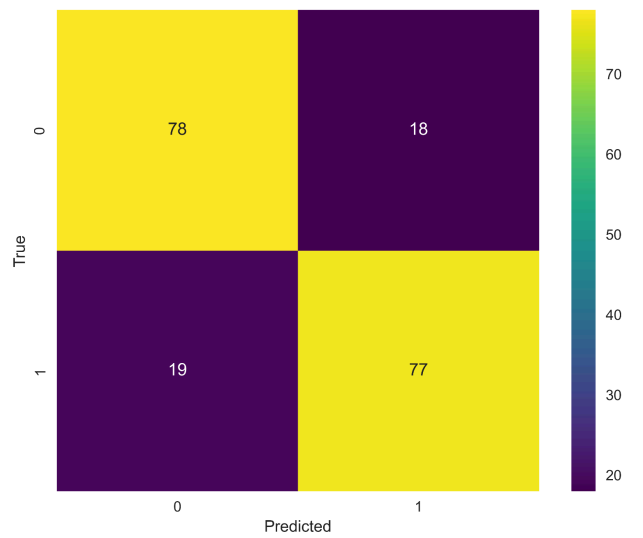
Εικόνα 61. Confusion Matrix LightGBM 4



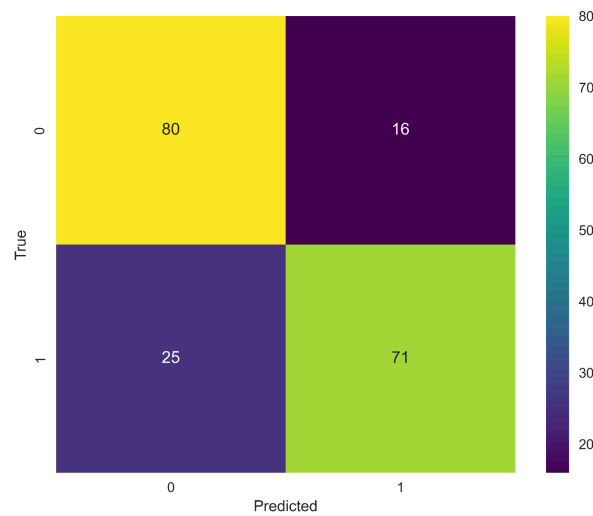
Εικόνα 62. Confusion Matrix AdaBoost 4



Εικόνα 63. Confusion Matrix CatBoost 4



Εικόνα 64. Confusion Matrix Extra Tree 4



Εικόνα 65. Confusion Matrix RusBoost 4

6. Εφαρμογή

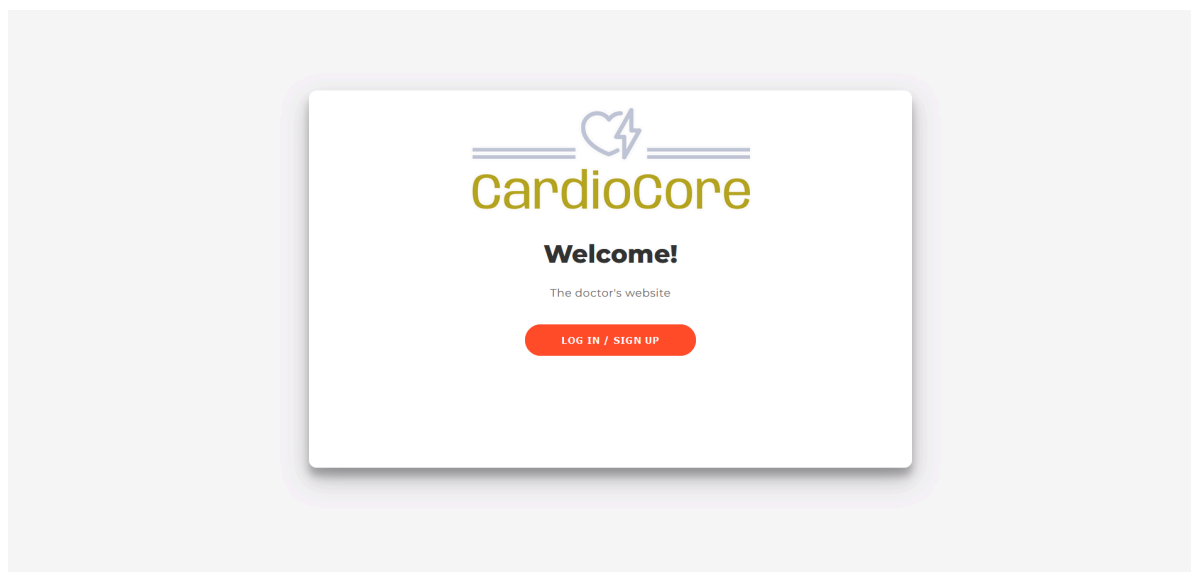
Για την ολοκλήρωση της παρούσας πτυχιακής, δημιουργήθηκε μια εφαρμογή για αποκλειστική χρήση γιατρού. Η εφαρμογή έχει ως στόχο να υποστηρίξει τους γιατρούς στην πρόβλεψη της καρδιακής ανεπάρκειας, βασιζόμενη σε αλγόριθμους μηχανικής μάθησης. Δεν προορίζεται να αντικαταστήσει την ιατρική διάγνωση, αλλά να λειτουργήσει ως εργαλείο υποστήριξης λήψης αποφάσεων.

Η εφαρμογή επιτρέπει στους γιατρούς να εισάγουν δεδομένα νέων ασθενών και να λαμβάνουν πρόβλεψη για την πιθανότητα καρδιακής ανεπάρκειας. Ουσιαστικά πραγματοποιείται μια σύγκριση των ασθενών που αποτελούν το σετ δεδομένων (dataset) με τον νέο ασθενή και γίνεται μια πρόβλεψη σχετικά με την κατάσταση στην οποία βρίσκεται ο ασθενής.

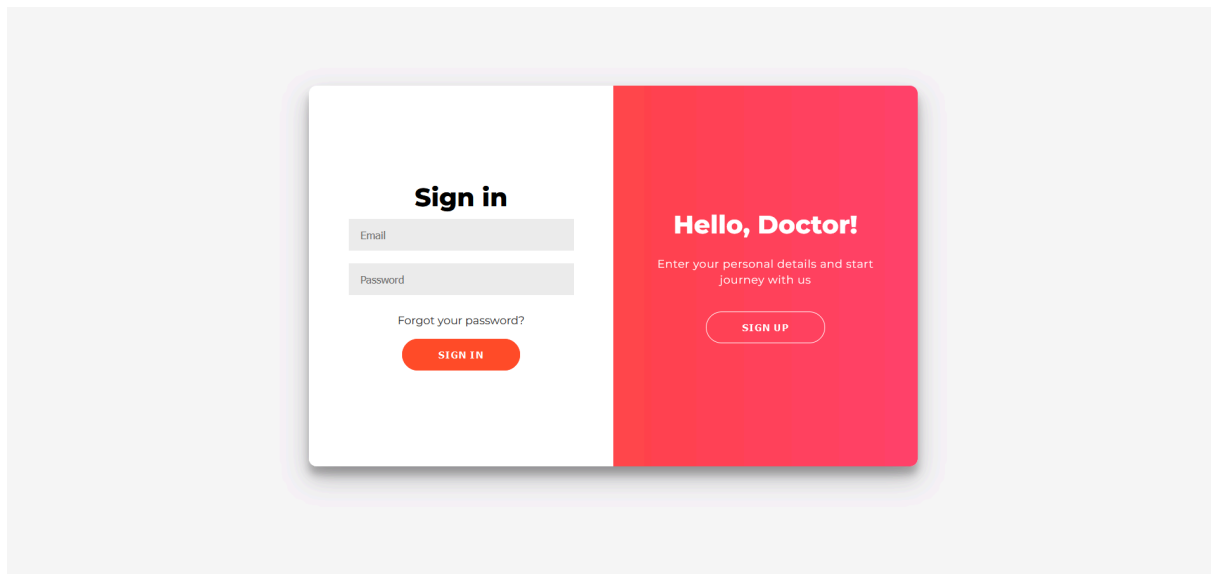
Αναπτύχθηκε, λοιπόν, με στόχο να είναι εύχρηστη και φιλική προς τον χρήστη, ειδικά για τους γιατρούς που θα την χρησιμοποιούν. Για τον σχεδιασμό της λήφθηκαν υπόψη οι πιθανές απαιτήσεις των τελικών χρηστών (γιατροί) για να καθοριστούν οι λειτουργίες που θα πρέπει να περιλαμβάνει η εφαρμογή και εν τέλει δημιουργήθηκε ένα φιλικό περιβάλλον χρήστη, λαμβάνοντας υπόψη και τις ανάγκες και τις δυνατότητες των γιατρών.

Χρησιμοποιήθηκαν τεχνολογίες όπως η Python για την υλοποίηση των αλγορίθμων μηχανικής μάθησης, django^[18] για την ανάπτυξη του backend της εφαρμογής, και HTML/ CSS/ JavaScript για το frontend. Παράλληλα, ενσωματώθηκε ο αλγόριθμος μηχανικής μάθησης που αναπτύχθηκε και αξιολογήθηκε στο πλαίσιο της έρευνας, ο Random Forest. Στα αρχεία του django υπάρχει ο “server-side” κώδικας που διαχειρίζεται τις εισερχόμενες αιτήσεις, επεξεργάζεται τα δεδομένα των ασθενών και επιστρέφει τις προβλέψεις στους χρήστες. Φυσικά, για να είναι εύχρηστη η εφαρμογή, υπάρχουν φόρμες για να συμπληρώσει ο γιατρός κάνοντας έτσι εύκολη την εισαγωγή των δεδομένων των ασθενών και την προβολή των αποτελεσμάτων.

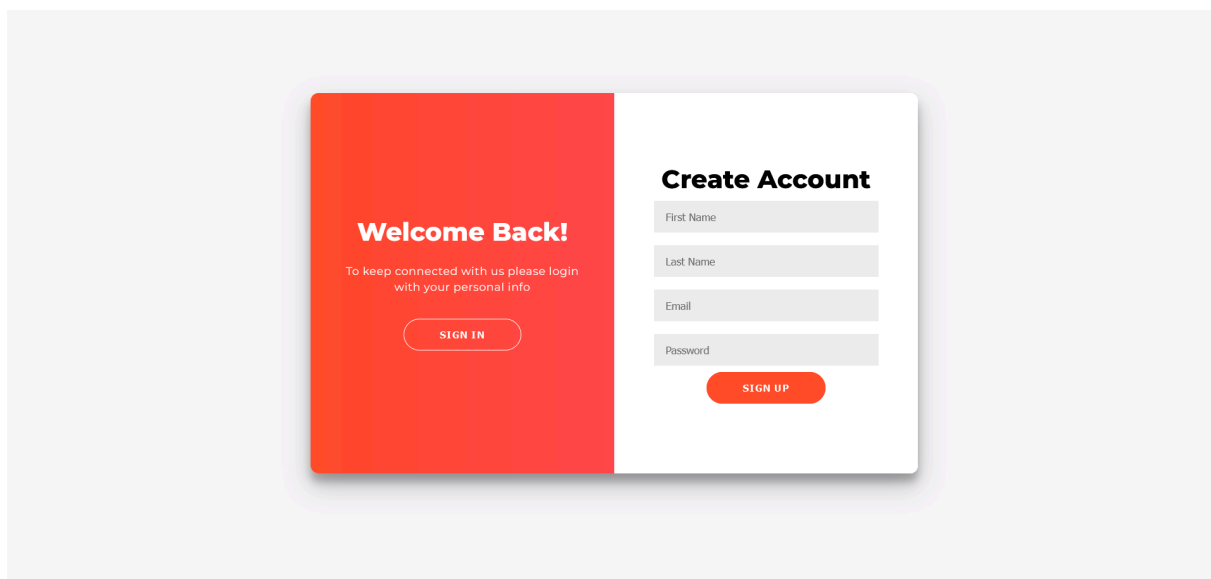
Συνοπτικά, λοιπόν, ο γιατρός εισάγει τα δεδομένα και η εφαρμογή χρησιμοποιεί τον εκπαιδευμένο αλγόριθμο για να προβλέψει την πιθανότητα καρδιακής ανεπάρκειας. Τα αποτελέσματα παρουσιάζονται με έναν κατανοητό τρόπο, επιτρέποντας στον γιατρό να λάβει υποστηριζόμενες αποφάσεις. Ακολούθως στις εικόνες απεικονίζεται η εφαρμογή.



Εικόνα 67. Σελίδα Εισόδου



Εικόνα 68. Σελίδα για Σύνδεση σε Υπάρχων Λογαριασμό



Εικόνα 69. Σελίδα για Δημιουργία Λογαριασμού

Welcome, Doctor!

This is your dashboard where you can manage your diagnoses for your patients, view patient information, and add a new patient to your database.

Add New Patient & Diagnosis

A new patient and a diagnosis

[New Patients & Diagnosis](#)

Patients

View and manage patient information.

[Go to Patients](#)

Εικόνα 70. Αρχική Σελίδα

Profile

Doctor Information

Name: Stella Adam

Email: stellaadam58@gmail.com

Update Profile

First Name:

Last Name:

Email:

[Update Profile](#)

Εικόνα 71. Σελίδα Profile

Add New Patient & Diagnosis

Add a new patient to your database and perform a diagnosis.

Age:

Anaemia:

Creatinine phosphokinase:

Diabetes:

Ejection fraction:

High blood pressure:

Platelets:

Serum creatinine:

Serum sodium:

Sex:

Smoking:

Time:

Add Patient to Database & Diagnose

Εικόνα 72. Σελίδα Προσθήκης Ασθενή & Διάγνωσης

Patients

#	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	s
1	75	0	582	0	20	1	265000	1.9	1
2	55	0	7861	0	38	0	263358.03	1.1	1
3	65	0	146	0	20	0	162000	1.3	1
4	50	1	111	0	20	0	210000	1.9	1
5	65	1	160	1	20	0	327000	2.7	1
6	90	1	47	0	40	1	204000	2.1	1
7	75	1	246	0	15	0	127000	1.2	1
8	60	1	315	1	60	0	454000	1.1	1
9	65	0	157	0	65	0	263358.03	1.5	1
10	80	1	123	0	35	1	388000	9.4	1
11	75	1	81	0	38	1	368000	4	1
12	62	0	231	0	25	1	253000	0.9	1
13	45	1	981	0	30	0	136000	1.1	1
14	50	1	168	0	38	1	276000	1.1	1

Εικόνα 73. Σελίδα Πίνακα Ασθενών και Αποτέλεσμα Διάγνωσης (υπάρχοντες ασθενείς)

297	45	0	2060	1	60	0	742000	0.8	1
298	45	0	2413	0	38	0	140000	1.4	1
299	50	0	196	0	45	0	395000	1.6	1

New Patients

View and manage the patients information that you added.

#	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum
1	80	0	1	0	38	0	25000.0	2.0	138
2	50	0	3	0	28	0	30000.0	1.2	138
3	81	0	3	0	28	0	35000.0	1.3	139
4	81	0	3	0	28	0	35000.0	1.3	139
5	96	0	3	0	38	0	25000.0	0.0	0
6	53	0	63	1	60	0	368000.0	0.8	135
7	86	0	3	0	135	0	36000.0	1.3	369
8	22	0	2	0	38	0	365000.0	1.2	138
9	58	0	1	0	38	0	350000.0	2.1	138
10	22	0	3	0	38	0	350000.0	2.0	139
11	90	0	7861	0	40	0	127000.0	1.9	116

Εικόνα 74. Σελίδα Πίνακα Ασθενών και Αποτέλεσμα Διάγνωσης (ασθενείς που προστέθηκαν)

7. Συμπεράσματα

Η παρούσα πτυχιακή εργασία είχε ως στόχο την ανάπτυξη και αξιολόγηση μοντέλων πρόβλεψης για την καρδιακή ανεπάρκεια χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Με βάση την ανάλυση των δεδομένων και τα αποτελέσματα των πειραμάτων, καταλήγουμε σε συμπεράσματα.

Αρχικά, η ανάλυση των χαρακτηριστικών των δεδομένων έδειξε σημαντικές διαφορές μεταξύ των ασθενών που επιβίωσαν και αυτών που δεν επιβίωσαν. Για παράδειγμα, η ηλικία φαίνεται να έχει ευρύτερη κατανομή στους επιζώντες ασθενείς. Οι θερμικοί χάρτες συσχετίσεων (heatmap) βοήθησαν στην αναγνώριση των ισχυρότερων και ασθενέστερων συσχετίσεων μεταξύ των χαρακτηριστικών, επιτρέποντας τη βελτιστοποίηση των μοντέλων πρόβλεψης.

Παράλληλα χρησιμοποιήθηκαν διάφορα μοντέλα μηχανικής μάθησης, όπως Decision Trees, Random Forest, Gradient Boosting, XGBoost, LightGBM, AdaBoost και άλλα. Κάθε μοντέλο εκπαιδεύτηκε και αξιολογήθηκε. Ο αλγόριθμος Random Forest, ο οποίος χρησιμοποιήθηκε και στην τελική εφαρμογή, απέδωσε πολύ καλά, προσφέροντας υψηλή ακρίβεια και ευρωστία στις προβλέψεις.

Η τελική εφαρμογή αναπτύχθηκε σε Django και ενσωμάτωσε τον αλγόριθμο Random Forest. Δοκιμάστηκε σε πραγματικά δεδομένα και απέδειξε την αποτελεσματικότητά της στην πρόβλεψη της καρδιακής ανεπάρκειας, ως εργαλείο υποστήριξης αποφάσεων για τους επαγγελματίες υγείας, δηλαδή τους γιατρούς.

Οι περιορισμοί της παρούσας έρευνας περιλαμβάνουν την ποιότητα και το μέγεθος των δεδομένων, καθώς και την πιθανή μεροληψία των μοντέλων λόγω ανισομερούς κατανομής των χαρακτηριστικών. Επίσης, η εστίαση σε μια συγκεκριμένη γεωγραφική περιοχή μπορεί να περιορίζει τη γενικότητα των αποτελεσμάτων.

Μελλοντικά, προτείνεται η χρήση μεγαλύτερων και πιο ποικίλων συνόλων δεδομένων για τη βελτίωση της γενικευσιμότητας των μοντέλων. Πιθανότατα η ενσωμάτωση περισσότερων χαρακτηριστικών και η δοκιμή νέων αλγορίθμων μηχανικής μάθησης θα μπορούσε να βελτιώσει περαιτέρω την ακρίβεια των προβλέψεων. Η ανάπτυξη ενός διαδραστικού περιβάλλοντος για τους χρήστες, που θα επιτρέπει την εύκολη τροποποίηση των παραμέτρων και την άμεση ανατροφοδότηση, θα μπορούσε να ενισχύσει την αποδοτικότητα της εφαρμογής.

Συνολικά, η εργασία αυτή απέδειξε ότι οι τεχνικές μηχανικής μάθησης μπορούν να προσφέρουν αξιόπιστα εργαλεία πρόβλεψης για την καρδιακή ανεπάρκεια, συμβάλλοντας σημαντικά στην πρόληψη και την έγκαιρη παρέμβαση. Η συνέχιση της έρευνας σε αυτόν τον τομέα είναι απαραίτητη για την περαιτέρω βελτίωση των τεχνικών και την ενσωμάτωσή τους στην καθημερινή κλινική πρακτική.

Παράρτημα

Με την χρήση του ακόλουθου κώδικα, απεφάνθει η σημαντικότητα των χαρακτηριστικών (feature importance) και το γράφημα που προκύπτει είναι η εικόνα 8 .

```
import pandas as pd
from catboost import CatBoostClassifier
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns

# Function to assign a rank to each element in a list with the
first element being the most important.
def rank_list(lst:list)->dict:
    # Returns a dictionary comprehension where each list element
    is a key and its rank is a value.
    # The rank is calculated as the length of the list minus one
    minus the element's index.
    return {elm:len(lst) - 1 - i for i, elm in enumerate(lst)}

# Function to perform Borda count aggregation for a list of
ranked lists.
def borda_aggregation(loflists: list[list]) -> dict:
    # Convert each individual list into a dictionary of ranks.
    list_ranks = [rank_list(l) for l in loflists]
    # Create a set of all unique elements across all the lists.
    feature_set = {i for i in [el for nl in loflists for el in
nl]}

    # Return a dictionary where each element's score is the sum
    of its ranks across all the lists.
    return {e:sum([lr.get(e, 0) for lr in list_ranks]) for e in
feature_set}

# Function to create a sorted DataFrame from a dictionary of
results.
def create_sorted_df(result:dict):
    # Create a DataFrame from the dictionary.
    df = pd.DataFrame(list(result.items()), columns=['Feature',
'Borda Rank'])

    # Sort the DataFrame based on the 'Borda Rank' column in
    descending order.
```

```

        df.sort_values(by='Borda Rank', ascending=False,
inplace=True)

        # Reset the DataFrame's index and drop the old index.
        df.reset_index(drop=True, inplace=True)

        # Return the sorted DataFrame.
        return df

    # Function to create a DataFrame from a list of lists using Borda
count aggregation.
    def borda_df(loflists: list[list]):
        # Aggregate the lists into a Borda count dictionary.
        borda_dict = borda_aggregation(loflists)

        # Create and return a sorted DataFrame from the Borda count
dictionary.
        return create_sorted_df(borda_dict)

    # Function to create a DataFrame showing the importance of
features from a model.
    def feature_importance_df_creation(model, model_label,
feature_names):
        # Create a DataFrame from the model's feature importances.
        importance_df = pd.DataFrame(model.feature_importances_)
        # Rename the column to the provided model label.
        importance_df = importance_df.rename(columns={0:
model_label})

        # Add the feature names as a column in the DataFrame.
        importance_df["feature_name"] = feature_names

        # Sort the DataFrame based on the importance scores in
descending order.
        importance_df = importance_df.sort_values(by=model_label,
ascending=False)

        # Return the sorted DataFrame.
        return importance_df

    # DATA_PATH =
"./datasets/greek_pts_numerical_thyroid_plus_outcomes_v1.csv"
    df= pd.read_csv('./heart_failure_clinical_records_dataset.csv',
sep=";")

    tag = pd.DataFrame(df["DEATH_EVENT"], columns=["DEATH_EVENT"])
    df = df.drop(['DEATH_EVENT'], axis=1)

```

```

xgbclass = XGBClassifier(eval_metric='logloss', verbosity=0,
importance_type='gain', use_label_encoder=False,
objective='binary:logistic')

cbc = CatBoostClassifier(iterations=100, verbose=0)
lgbm = LGBMClassifier(importance_type='split',
objective="binary")

xgbclass.fit(df, tag["DEATH_EVENT"])

cbc.fit(df_scaled, tag["DEATH_EVENT"])

lgbm.fit(df_scaled, tag["DEATH_EVENT"])

xgb_importance = feature_importance_df_creation(xgbclass,
"XgBoost_importance", list(df_scaled.columns))
cbc_importance = feature_importance_df_creation(cbc,
"CatBoost_importance", list(df_scaled.columns))
lgbm_importance = feature_importance_df_creation(lgbm,
"LightGBM_importance", list(df_scaled.columns))

xgboost_imp_lst = list(xgb_importance["feature_name"])
catboost_imp_lst = list(cbc_importance["feature_name"])
lgbm_imp_lst = list(lgbm_importance["feature_name"])

borda_results = borda_df([xgboost_imp_lst, catboost_imp_lst,
lgbm_imp_lst])
borda_results

plt.figure(figsize=(8, 6))
ax = sns.scatterplot(data=borda_results, x="Borda Rank",
y="Feature", s=100, color="steelblue", label=" ")
x_start = borda_results["Borda Rank"].min()
ax.set_xlim(left=x_start)
ax.set_facecolor("whitesmoke")
for _, row in borda_results.iterrows():
    ax.plot([x_start, row["Borda Rank"]], [row["Feature"],
row["Feature"]], color="steelblue", linewidth=0.5)
ax.set_title("Borda Consensus Feature Importance")
ax.set_xlabel("Importance")
ax.set_ylabel("Feature")
ax.legend_.remove()
plt.tight_layout()

```

```

plt.savefig("./figures_anyadr/borda_feature_impotance_imbalanced
.pdf", bbox_inches="tight")
plt.show()

borda_results[["Feature"]].to_csv("./feature_lst.csv",
index=False)

```

Ο ακόλουθος κώδικας εμφανίζει τους πίνακες σύγχυσης (confusion matrix), που βρίσκονται συγκεντρωτικά στο Κεφάλαιο 5, καθώς επίσης και τα αποτελέσματα για τις μετρικές σε κάθε αλγόριθμο, που βρίσκονται στο Κεφάλαιο 5 στις εικόνες 26, 36, 46 και 56.

Ο κώδικας εκτελείται για κάθε αλγόριθμο του 4ου Κεφαλαίου και για τα τέσσερα (4) πειράματα αλλάζοντας μονάχα το αρχείο csv που “διαβάζει”. Στις ακόλουθες γραμμές υλοποιείται μονάχα ο αλγόριθμος Random Forest και σημειώνεται πως η ίδια υλοποίηση ισχύει και για τους υπόλοιπους οκτώ (8) αλγορίθμους.

```

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, f1_score,
roc_auc_score, confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import KFold
import numpy as np
import pandas as pd
from sklearn.feature_selection import RFE
from sklearn.model_selection import cross_validate
from pycaret import classification as clf
import warnings
warnings.filterwarnings('ignore')

df = pd.read_csv("./heart_failure_clinical_records_dataset.csv",
sep=";")
df

tag = pd.DataFrame(df["DEATH_EVENT"], columns=["DEATH_EVENT"])
df = df.drop(['DEATH_EVENT'], axis=1)

def perform_kfold_cv_with_metrics(df, tag, model, n_splits=10,
plot_cm=True):
    """
    Perform K-Fold cross-validation, calculate metrics, including
    standard deviations,
    and optionally plot and save the confusion matrix.

```

```

Parameters:
- df: DataFrame containing the features.
- tag: Series containing the target labels.
    - model: The machine learning model to be trained and
evaluated.
- n_splits: Number of splits for K-Fold cross-validation.
- plot_cm: Boolean, whether to plot the confusion matrix.

Returns:
- Confusion matrix, average accuracy, average F1 score,
average ROC-AUC score, and their standard deviations.
"""
kf = KFold(n_splits=n_splits, shuffle=True, random_state=42)

accuracies = []
f1_scores = []
roc_aucs = []
aggregate_predictions = []
aggregate_actuals = []

for train_index, test_index in kf.split(df):
    X_train, X_test = df.iloc[train_index],
df.iloc[test_index]
    y_train, y_test = tag.iloc[train_index],
tag.iloc[test_index]

    model.fit(X_train, y_train)
    predictions = model.predict(X_test)

    aggregate_predictions.extend(predictions)
    aggregate_actuals.extend(y_test.values.tolist())

    roc_auc = None
    if hasattr(model, "predict_proba"):
        proba_predictions = model.predict_proba(X_test)[:, 1]
        roc_auc = roc_auc_score(y_test, proba_predictions)
    elif hasattr(model, "decision_function"):
        proba_predictions = model.decision_function(X_test)
        roc_auc = roc_auc_score(y_test, proba_predictions)

    acc = accuracy_score(y_test, predictions)
    f1 = f1_score(y_test, predictions, average='binary')
    if roc_auc is not None:

```

```

        roc_aucs.append(roc_auc)

    accuracies.append(acc)
    f1_scores.append(f1)

    conf_matrix = confusion_matrix(aggregate_actuals,
aggregate_predictions)

    if plot_cm:
        plt.figure(figsize=(8, 6))
        sns.heatmap(conf_matrix, annot=True, fmt="d",
cmap="viridis", square=True, cbar=True,
                    xticklabels=np.unique(aggregate_actuals),
yticklabels=np.unique(aggregate_actuals))
        plt.xlabel('Predicted')
        plt.ylabel('True')
        plt.title('Confusion Matrix')
        plt.savefig('confusion_matrix.png', dpi=600) # Save the
confusion matrix as a PNG file with 600 DPI
        plt.show()

    avg_accuracy = np.mean(accuracies)
    avg_f1_score = np.mean(f1_scores)
    avg_roc_auc = np.mean(roc_aucs) if roc_aucs else None

    # Calculate standard deviations
    std_accuracy = np.std(accuracies)
    std_f1_score = np.std(f1_scores)
    std_roc_auc = np.std(roc_aucs) if roc_aucs else None

    return conf_matrix, avg_accuracy, avg_f1_score, avg_roc_auc,
std_accuracy, std_f1_score, std_roc_auc

    rfc = RandomForestClassifier(n_estimators=100)
    conf_matrix, avg_accuracy, avg_f1_score, avg_roc_auc,
std_accuracy, std_f1_score, std_roc_auc =
perform_kfold_cv_with_metrics(df, tag, rfc)

    print(f"Average Accuracy: {avg_accuracy} with a standard
deviation: {std_accuracy}")
    print(f"Average F1 Score: {avg_f1_score} with a standard
deviation: {std_f1_score}")

```

```

    print(f"Average ROC-AUC: {avg_roc_auc} with a standard deviation:
{std_roc_auc}", "Not calculated" if avg_roc_auc is None else
avg_roc_auc)
    selector = RFE(rfc, n_features_to_select=10, step=1)

    selector = selector.fit(df, tag)

    selected_features = df.columns[selector.support_]
    df_selected = df[selected_features]
    df_selected

    conf_matrix,      avg_accuracy,      avg_f1_score,      avg_roc_auc,
std_accuracy,        std_f1_score,        std_roc_auc        =
perform_kfold_cv_with_metrics(df_selected , tag, rfc)

    print(f"Average Accuracy: {avg_accuracy} with a standard
deviation: {std_accuracy}")
    print(f"Average F1 Score: {avg_f1_score} with a standard
deviation: {std_f1_score}")
    print(f"Average ROC-AUC: {avg_roc_auc} with a standard deviation:
{std_roc_auc}", "Not calculated" if avg_roc_auc is None else
avg_roc_auc)

```

Βιβλιογραφία

[1] British Heart Foundation. (2023). BHF Statistics Compendium 2023

[2] World Heart Federation. (2023). World Heart Report 2023

[3] Connie W. Tsao, MD, MPH, FAHA, Chair, Aaron W. Aday, MD, MSc, FAHA, Zaid I. Almarzooq, MB, BCH, Alvaro Alonso, MD, PhD, FAHA, Andrea Z. Beaton, MD, MS, FAHA, Marcio S. Bittencourt, MD, PhD, MPH, FAHA, Amelia K. Boehme, PhD, MSPH, Alfred E. Buxton, MD, April P. Carson, PhD, MSPH, FAHA, Yvonne Commodore-Mensah, PhD, MHS, RN, FAHA, Mitchell S.V. Elkind, MD, MS, FAHA, Kelly R. Evenson, PhD, MS, FAHA, Chete Eze-Nliam, MD, MPH, Jane F. Ferguson, PhD, FAHA, Giuliano Generoso, MD, PhD, Jennifer E. Ho, MD, FAHA, Rizwan Kalani, MD, Sadiya S. Khan, MD, MSc, FAHA, Brett M. Kissela, MD, MS, FAHA, Kristen L. Knutson, PhD, Deborah A. Levine, MD, MPH, Tené T. Lewis, PhD, FAHA, Junxiu Liu, PhD, Matthew Shane Loop, PhD, FAHA, Jun Ma, , MD, PhD, FAHA, Michael E. Mussolino, PhD, FAHA, Sankar D. Navaneethan, MD, MS, MPH, Amanda Marma Perak, MD, MS, FAHA, Remy Poudel, MS, MPH, CPH, Mary Rezk-Hanna, PhD, FAHA, Gregory A. Roth, MD, MPH, FAHA, Emily B. Schroeder, MD, PhD, FAHA, Svati H. Shah, MD, MHS, FAHA, Evan L. Thacker, PhD, Lisa B. VanWagner, MD, MSc, FAHA, Salim S. Virani, MD, PhD, FAHA, Jenifer H. Voecks, PhD, Nae-Yuh Wang, PhD, MS, FAHA, Kristine Yaffe, MD, Seth S. Martin, MD, MHS, FAHA, Vice Chair, on behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee 2022

<https://www.ahajournals.org/doi/epub/10.1161/CIR.0000000000001052>

[4] European Heart Journal, "Epidemiology of Heart Failure in Europe: A Comprehensive Review", Volume 42, Issue 5, May 2021.

[5] Bowen RES, Graetz TJ, Emmert DA, Avidan MS. Statistics of heart failure and mechanical circulatory support in 2020. *Ann Transl Med.* 2020 Jul;8(13):827. doi: 10.21037/atm-20-1127. PMID: 32793672; PMCID: PMC7396255.

[6] Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, Masoudi FA, Spertus JA, Krumholz HM. Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction. *JACC Heart Fail.* 2020 Jan;8(1):12-21. doi: 10.1016/j.jchf.2019.06.013. Epub 2019 Oct 9. PMID: 31606361.

[7] Yasmin F, Shah SMI, Naeem A, Shujaiddin SM, Jabeen A, Kazmi S, Siddiqui SA, Kumar P, Salman S, Hassan SA, Dasari C, Choudhry AS, Mustafa A, Chawla S, Lak HM. Artificial intelligence in the diagnosis and detection of heart failure: the past, present, and future. *Rev Cardiovasc Med.* 2021 Dec 22;22(4):1095-1113. doi: 10.31083/j.rcm2204121. PMID: 34957756.

[8] Norhammar A, Bodegard J, Vanderheyden M, Tangri N, Karasik A, Maggioni AP, Svein KA, Taveira-Gomes T, Botana M, Hunziker L, Thuresson M, Banerjee A, Sundström J, Bollmann A. Prevalence, outcomes and costs of a contemporary, multinational population

with heart failure. *Heart*. 2023 Mar 10;109(7):548-556. doi: 10.1136/heartjnl-2022-321702. PMID: 36781285; PMCID: PMC10086499.

[9] Yuan H, Fan XS, Jin Y, He JX, Gui Y, Song LY, Song Y, Sun Q, Chen W. Development of heart failure risk prediction models based on a multi-marker approach using random forest algorithms. *Chin Med J (Engl)*. 2019 Apr 5;132(7):819-826. doi: 10.1097/CM9.000000000000149. PMID: 30829708; PMCID: PMC6595865.

[10] Vinter N, Gerds TA, Cordsen P, Valentin JB, Lip GYH, Benjamin EJJ, Johnsen SP, Frost L. Development and validation of prediction models for incident atrial fibrillation in heart failure. *Open Heart*. 2023 Jan;10(1):e002169. doi: 10.1136/openhrt-2022-002169. PMID: 36639191; PMCID: PMC9843222.

[11] Avula S, LaFata M, Nabhan M, Allana A, Toprani B, Scheidel C, Suneja A. Heart failure mortality prediction using PRISM score and development of a classification and regression tree model to refer patients for palliative care consultation. *Int J Cardiol Heart Vasc*. 2019 Dec 13;26:100440. doi: 10.1016/j.ijcha.2019.100440. Erratum in: *Int J Cardiol Heart Vasc*. 2020 Dec 19;32:100699. PMID: 31886404; PMCID: PMC6921143.

[12] H. Padalko, K. Bazilevych and M. Butkevych, "Heart Failure Development Prediction using Stochastic Gradient Descent Optimization," 2022 IEEE 9th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T), Kharkiv, Ukraine, 2022, pp. 297-300, doi: 10.1109/PICST57299.2022.10238625.

[13] P. -Y. Liang, L. -J. Wang, Y. -S. Wu, T. -W. Pai, C. -H. Wang and M. -H. Liu, "Prediction of patients with heart failure after myocardial infarction," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 2009-2014, doi: 10.1109/BIBM49941.2020.9313253.

[14] T. P. Pushpavathi, S. Kumari and N. K. Kubra, "Heart Failure Prediction by Feature Ranking Analysis in Machine Learning," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 915-923, doi: 10.1109/ICICT50816.2021.9358733.

[15] T. P. Pushpavathi, S. Kumari and N. K. Kubra, "Heart Failure Prediction by Feature Ranking Analysis in Machine Learning," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 915-923, doi: 10.1109/ICICT50816.2021.9358733.

[16] <https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data>

[17] <https://pycaret.org/>

[18] <https://www.djangoproject.com/>

[19] Οι εικόνες του Κεφαλαίου 1 δημιουργήθηκαν στο biorender