Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων & Μηχανικων Υπολογιστων
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ
ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Ανάπτυξη Μοντέλων Τεχνητής Νοημοσύνης για την Πρόβλεψη της Μετεγχειρητικής Πορείας Ασθενών με Κάταγμα Ισχύου

Διπλωματικη Εργασια

της

Αγγελικής Πνευματικού

Επιβλέπων: Γεώργιος Ματσόπουλος
Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2024

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ
ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Ανάπτυξη Μοντέλων Τεχνητής Νοημοσύνης για την Πρόβλεψη της Μετεγχειρητικής Πορείας Ασθενών με Κάταγμα Ισχύου

Διπλωματική Εργασία

της

Αγγελικής Πνευματικού

Επιβλέπων: Γεώργιος Ματσόπουλος
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 8η Ιουλίου 2024.

Αθήνα, Ιούλιος 2024

Γ. Ματσόπουλος	Π. Τσανάκας	Αθ. Παναγόπουλος
Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ

Αγγελική Πνευματικού
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών ΕΜΠ

Περίληψη

Τα κατάγματα ισχίου αποτελούν ένα πολύ συχνό φαινόμενο, κυρίως στον ηλικιωμένο πληθυσμό. Πολλοί ασθενείς τραυματίζονται και υποβάλλονται σε χειρουργείο προκειμένου να αποκατασταθούν. Η περίθαλψη και ο τρόπος αντιμετώπισης κάθε ασθενούς είναι αρκετά κοστοβόρα και ο κίνδυνος θνησιμότητας είναι υψηλός. Για τον λόγο αυτό, υπάρχει έντονη ανάγκη να εντοπιστούν οι παράγοντες που αυξάνουν τον κίνδυνο θνησιμότητας, ώστε ο κλινικός να μπορεί να εκτιμήσει την κατάλληλη πορεία περίθαλψης.

Συλλέχθηκαν 400 περιστατικά από ασθενείς που είχαν εισαχθεί στην Πανεπιστημιακή Κλινική του νοσοκομείου ΚΑΤ που περιείχαν δημογραφικές πληροφορίες, ασθένειες και άλλες πληροφορίες που περιγράφουν την κατάσταση των ασθενών. Τα δεδομένα αυτά πέρασαν από διαδικασίες επιλογής χαρακτηριστικών με διάφορες μεθόδους, ώστε να βρεθούν τα υποσύνολα του αρχικού συνόλου δεδομένων και στη συνέχεια τροφοδοτήθηκαν σε διάφορα μοντέλα μηχανικής μάθησης.

Οι μέθοδοι επιλογής χαρακτηριστικών που χρησιμοποιήθηκαν ήταν: συσχέτιση, δέντρα αποφάσεων, forward feature selection, fisher scores, κέρδος πληροφορίας, LASSO regularization L1, random forest, recursive feature elimination, sequential feature selection και variance inflation factor. Τα μοντέλα που χρησιμοποιήθηκαν ήταν: artificial neural network, ένα νευρωνικό δίκτυο με έναν απλό αλγόριθμο back-propagation, δέντρα απόφασης, gradient boosting, knn, γραμμική παλινδρόμηση, naive bayes, PCA, penalized logistic regression, random forest, stochastic gradient descent, support vector machine και ο αλγόριθμος XGBoost.

Το μοντέλο που πέτυχε τη μεγαλύτερη ακρίβεια (92.5%) ήταν το back propagation σε συνδυασμό με τη μέθοδο επιλογής χαρακτηριστικών FFS, η οποία είχε επιλέξει ως παράγοντες υψηλού ρίσκου: φύλο, garden, AO fracture classification, evans fracture classification, είδος αναισθησίας, είδος χειρουργείου, atrial fibrillation, chronic kidney disease, διαβήτης, frailty score, post-operative physical ability, periprosthetic infection, pneumonia, deep vein thrombosis, επανεισαγωγή, είδος κατάγματος και infection risk score.

Λέξεις-κλειδιά: κατάγματα ισχίου, επιλογή χαρακτηριστικών, μηχανική μάθηση, θνησιμότητα

Abstract

Hip fractures are a very common phenomenon, mainly in the elderly population. Many patients are injured and undergo surgery in order to be restored. The care and way of managing each patient are quite costly and the risk of mortality is high. For this reason, there is an intense need to identify the factors that increase the risk of mortality so that the clinician can estimate the appropriate course of care.

400 cases were collected from patients who had been admitted to the University Clinic of KAT Hospital, containing demographic information, diseases, and other information describing the condition of the patients. This data went through feature selection processes using various methods to find subsets of the original data set and then fed into various machine learning models.

The feature selection methods used were: correlation, decision trees, forward feature selection, fisher scores, information gain, LASSO regularization L1, random forest, recursive feature elimination, sequential feature selection, and variance inflation factor. The models used were: artificial neural network, a neural network with a simple back-propagation algorithm, decision trees, gradient boosting, knn, linear regression, naive bayes, PCA, penalized logistic regression, random forest, stochastic gradient descent, support vector machine, and the XGBoost algorithm.

The model that achieved the highest accuracy (92.5%) was back propagation combined with the feature selection method FFS, which had selected as high-risk factors: gender, garden, AO fracture classification, evans fracture classification, type of anesthesia, type of surgery, atrial fibrillation, chronic kidney disease, diabetes, frailty score, post-operative physical ability, periprosthetic infection, pneumonia, deep vein thrombosis, readmission, type of fracture, and infection risk score.

Keywords: hip fractures, feature selection, machine learning, mortality

Ευχαριστίες

Καταρχάς, θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη υπεύθυνο καθηγητή μου, τον Καθηγητή Γεώργιο Ματσόπουλο

Στη συννέχεια θα ήθελα να εκφράσω την ευγωμοσύνη μου στον σύμβουλο μου, Ιωάννη Βεζάκη, για τη συνεχή υποστήριξή του. Η καθοδήγησή του με βοήθησε καθ' όλη τη διάρκεια της έρευνας και της συγγραφής αυτής της διπλωματικής.

Θα ήθελα, επίσης, να ευχαριστήσω τα μέλη της επιτροπής της διπλωματικής μου: τον Καθηγητή Παναγιώτη Τσανάκα και τον Καθηγητή Αθανάσιο Δ. Παναγόπουλο..

Είμαι ευγνώμων στην Πανεπιστημιακή Κλινική του νοσοκομείου ΚΑΤ για την παροχή των δεδομένων που ήταν απαραίτητα για την έρευνά μου.

Οι ειλικρινείς μου ευχαριστίες επίσης πηγαίνουν στους υποψήφιους διδάκτωρες, Γεώργιος Τσάλιμας και Κυριάκος Καντονίδης, στην Ιατρική Σχολή του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών (ΕΚΠΑ) που είχαν αναλάβει τη διαδικασία συλλογής των δεδομένων από τους ασθενείς.

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

Τα κατάγματα ισχύος αποτελούν συχνή πρόκληση στην ορθοπεδική πρακτική και επιβάλλουν σημαντική ετήσια επιβάρυνση στα συστήματα υγειονομικής περίθαλψης. Οι επιπλοκές που προκύπτουν από αυτούς τους τραυματισμούς επηρεάζουν κυρίως τους ηλικιωμένους με οστεοπόρωση, καθώς έχει παρατηρηθεί αξιοσημείωτη συσχέτιση μεταξύ τέτοιων καταγμάτων με αυξημένη αναπηρία, νοσηρότητα και θνησιμότητα, ειδικά κατά τον πρώτο χρόνο μετά τον τραυματισμό. Με την αύξηση του προσδόκιμου ζωής, ο προβλεπόμενος αριθμός περιστατικών αυτών αναμένεται επίσης να αυξηθεί, καθώς εκτιμάται ότι θα φτάσει τα 3 εκατομμύρια έως το 2025 και περαιτέρω προβλέπεται ότι θα κλιμακωθεί σε 4,5 έως 6,3 εκατομμύρια έως το 2050. Αυτό θα οδηγήσει σε ακόμα μεγαλύτερη οικονομική επιβάρυνση του συστήματος υγείας καθώς τα κόστη των επεμβάσεων και μετέπειτα της ανάρρωσης αναμένονται να διπλασιαστούν. Επιπλέον, παράλληλα με τον αυξανόμενο αριθμό περιπτώσεων, υπάρχει σημαντικός επιπολασμός ανεπιθύμητων κλινικών εκβάσεων και θανάτων, που θα μπορούσαν να αποδοθούν στην ανεπαρκή θεραπεία και φροντίδα των ατόμων αυτών. Με αφορμή αυτές τις συνθήκες, έχουν διεξαχθεί πολλαπλές έρευνες που μελετούν τους παράγοντες που μπορουν να βοηθήσουν τους ειδικούς να εκτιμήσουν τα αποτελέσματα των επεμβάσεων για τα κατάγματα ισχύου με σκοπό την πιο εύστοχη και έγκαιρη αντιμετώπιση των περιστατικών αυτών αλλά και των επιπλοκών που μπορεί να προκύψουν.

Κατά τη διεξαγωγή των ερευνών αυτών αναπτύσσονται μοντέλα μηχανικής μάθησης τα οποία τροφοδοτούνται με πολλαπλά ιατρικά δεδομένα των ασθενών, όπως ηλικία, φύλο, είδος κατάγματος, χρόνος νοσηλείας και πολλά ακόμα, με στόχο να εντοπίσουν αλληλοσχετίσεις μεταξύ των δεδομένων αυτών και να προβάλλουν εκτιμήσεις για την μετεχγειριτική περίοδο του κάθε ασθενούς. Προηγούμενες έρευνες έχουν εντοπίσει ότι παράγοντες όπως η ηλικία, το φύλο και η παρουσία συννοσηροτήτων, συμπεριλαμβανομένου του διαβήτη, της πνευμονίας και άλλων λοιμώξεων, παρουσιάζουν ισχυρότερη συσχέτιση με τη δυσμενή έκβαση σε αυτές τις περιπτώσεις. Τέτοια μοντέλα μπορούν να βοηθήσουν σημαντικά τη διαδικασία επιλογής του είδους παροχής ιατρικής φροντίδας των ασθενών καθώς θα μπορούν να προβλεφθούν ή να εκτιμηθούν τυχόν επιπλοκές, με αποτέλεσμα την πιο στοχευμένη αξιοποίηση

των πόρων του νοσοκομείου και την καταλληλότερη λήψη αποφάσεων, αφού θα βασίσονται πάνω στα δεδομένα.

Αυτή η μελέτη στοχεύει να εμβαθύνει στην υπάρχουσα έρευνα για να εντοπίσει χαρακτηριστικά που ενισχύουν την κατανόησή μας για τις περιόδους κατά την ανάρρωσή των ασθενών. Στο πλαίσιο αυτής της μελέτης, χρησιμοποιώντας κατάλληλες μεθοδολογίες από την υπάρχουσα βιβλιογραφία, θα εντοπίσουμε και θα αξιολογήσουμε τις συσχετίσεις μεταξύ των καθορισμένων παραγόντων του κάθε ασθενούς τα οποία, κατόπιν, θα τα αξιοποιήσουμε για περεταίρω εκτιμήσεις. Στη συνέχεια, θα αναπτύξουμε, θα διερευνήσουμε και θα αξιολογήσουμε μοντέλα μηχανικής μάθησης για να αναλύσουμε αυτά τα περιστατικά, εστιάζοντας στην αξιολόγηση της ακρίβειας και της αξιοπιστίας τους στην πρόβλεψη του χρονοδιαγράμματος ανάρρωσης ενός ασθενούς, με κύρια εστίαση στην πρόβλεψη της θνησιμότητας. Η εμπειρική βάση για τα μοντέλα μας προέρχεται από δεδομένα που παρέχονται από την Πανεπιστημιακή Κλινική του Νοσοκομείου ΚΑΤ, που αφορούν ασθενείς που υποβάλλονται σε θεραπεία για κατάγματα ισχύου. Ο στόχος μας εκτείνεται πέρα από την απλή κατηγοριοποίηση χαρακτηριστικών στην αξιολόγηση και εξέταση των μοντέλων για να διασφαλιστεί μια πιο αξιόπιστη αξιολόγηση των αποτελεσμάτων.

Κεφάλαιο 2

Ιατρικά Δεδομένα - Κάταγμα Ισχύου

2.1 Κάταγμα Ισχύου

Το κάταγμα ισχύου πρόκειται για ένα πρόβλημα της δημόσιας υγείας που οδηγεί σε νοσηλεία, μακροχρόνια αποκατάσταση, μειωμένη ποιότητα ζωής, υψηλή θνησιμότητα στους ασθενείς καθώς και συνοδεύεται από μεγάλα έξοδα υγειονομικής περίθαλψης [1]. Η βελτίωση του βιωτικού επιπέδου οδηγεί στην αύξηση του αριθμού των περιστατικών κατάγματος ισχύου το οποίο συνοδεύεται σε αύξηση των επιπλοκών, με ό,τι αυτό συνεπάγεται για τα συστήματα υγείας. Πολλά από αυτά τα περιστατικά αντιμετωπίζονται με χειρουργικές επεμβάσεις και συνοδεύονται και από ένα μεγάλο ρίσκο θνησιμότητας. Για την ακρίβεια, έχει εκτιμηθεί ότι οι θνησιμότητα εντός των 30 ημερών μετά τον τραυματισμό ανέρχεται στο 5.3% [2]. Οι ασθενείς που διατρέχουν μεγαλύτερο κίνδυνο καταγμάτων από τον γενικό πληθυσμό είναι οι ηλικιωμένοι λόγω του αυξημένου κινδύνου πτώσης σε συνδυασμό με την οστεοπόρωση. Επιπλέον, η ακατάλληλη ή μη-έγκαιρη ιατρική παρέμβαση συμβάλλει σημαντικά στα αρνητικά αποτελέσματα που σχετίζονται με αυτές τις περιπτώσεις. Είναι, συνεπώς, κρίσημο να βρεθούν τρόποι με τους οποίους να γίνεται ακριβή εκτίμηση της αναρρωτικής περιόδου του ασθενούς πριν αποφασιστεί κάποια θεραπεία με στόχο την καταλληλότερη και εξατομικευμένη αντιμετώπιση του κάθε ασθενούς.

Βάσει των σύχγρονων ερευνών, έχει εκτιμηθεί ότι κατά προσέγγιση τα κατάγματα ισχύου επηρεάζουν το 18% του γυναικείου πληθυσμού και το 6% του ανδρικού [3]. Συνήθως, αυτά τα περιστατικά απαιτούν χειρουργική παρέμβαση με στόχο την γρήγορη επαναφορά της κινητικότητας προκειμένου να μπορεί ο ασθενής να επιτελεί καθημερινές δραστηριότητες. Έχει, ωστόσο, παρατηρηθεί ότι όσον αφορά το υποκεφαλικό κάταγμα ισχύου, η θνησιμότητα κατά το πρώτο έτος μετά το ατύχημα πλησιάζει το 30%. Οι περισσότεροι θάνατοι οφείλονται σε συννοσηρότητες και μετεγχειρετικές επιπλοκές, όπως λοιμώξεις της χειρουργικής περιοχής (SSI)[4]. Όταν προκύπτουν λοιμώξεις, σε πολλές περιπτώσεις, είναι απαραίτητη η επανεισαγωγή και επανεχγείριση για την ανταλλαγή ή αφαίρεση κάποιου χειρουργικού υλικού ή

εξαρτήματος που είχε τοποθετηθεί για την αποκατάσταση του τραύματος, πράγμα που αυξάνει την νοσηρότητα για τον ασθενή [3]. Αυτό κατά συνέπεια, πέραν ότι συμπεριλαμβάνει επαναληπτικό χειρουργείο, μπορεί να απαιτήσει και αύξηση της κατανάλωσης αντιβιοτικών, παρατεταμένη διαμονή στο νοσοκομείο, τα οποία επιβαρύνουν σημαντικά το νοσοκομείο από πλευράς οικονομικής άποψης εφόσον αυξάνει σημαντικά τα έξοδα που πραγματοποιούνται για την περίθαλψη των ασθενών αυτών. Από άλλες μελέτες, όμως, έχει αποδειχθεί ότι ο κίνδυνος θνησιμότητας αυξάνεται σε περιπτώσεις που οι ασθενείς δεν έχουν υποβληθεί επέμβαση. Πιο συγκεκριμένα, οι ασθενείς με κάταγμα ισχύου που υποβλήθηκαν σε θεραπεία μη-εγχειρητικά είχαν υψηλότερο κίνδυνο θνησιμότητας σε έναν χρόνο κατά 29.8% και σε δύο χρόνια 45.6%, δηλαδή ήταν τέσσερις φορές υψηλότερος σε ένα έτος και τρείς φορές υψηλότερος σε δύο χρόνια από την ομάδα που χειρουργήθηκε [5].

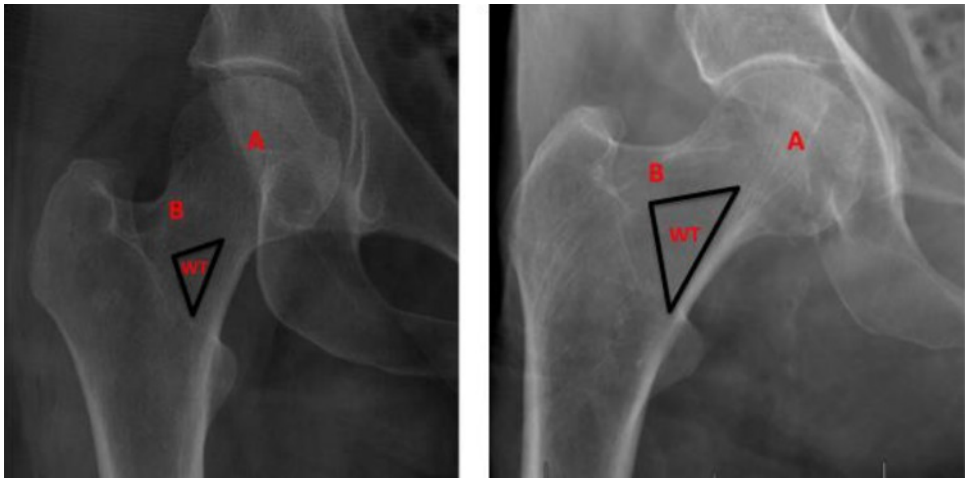
Βάσει μίας οικονομικής ανάλυσης, είχε επισημανθεί ότι, παρόλο που τα κατάγματα ισχύου αποτελούν χονδρικά το 14% από όλα τα κατάγματα ευθραυστότητας, αυτά τα περιστατικά αντιπροσωπεύουν αξιοσημείωτη δαπάνη με εκτιμώμενο κόστος 15 δισεκατομμύρια δολάρια ετησίως στις Ηνωμένες Πολιτείες της Αμερικής. Η θεραπεία αυτών των καταγμάτων είχε ταξινομηθεί ως 13η πιο ακριβή διάγνωση από το Medicare για το 2011, ενώ ταυτόχρονα, αποδείχθηκε ότι παρόλο που ένα κάταγμα ισχύου εκτιμήθηκε ότι κόστιζε περίπου 10.000 δολάρια για την αρχική νοσηλεία, το εκτιμώμενο κόστος υγειονομικής περίθαλψης και κοινωνικού κόστους ενός έτους προσδιορίζεται κοντά στα 43.000 δολάρια και πιθανότατα οφείλεται στην αυξανόμενη ανάγκη για πρόσθετη φροντίδα και επίβλεψη μετά την επέμβαση [6]. Αυτό μπορεί να υποστηριχθεί και περαιτέρω από μελέτες που δείχνουν ότι, για ένα μεγάλο ποσοστό ασθενών με κατάγματα ισχύου για τους οποίος απαιτείται μακροχρόνια τοποθέτηση σε μονάδα φροντίδας, απαιτείται κατά μέσον όρο σχετικό κόστος κυμαίνεται από 19.000 εώς 66.000 δολάρια. Αυτό υπογραμμίζει ότι η διαχείρηση των καταγμάτων αυτών θα παραμείνει μία σημαντική πτυχή της γηριατρικής υγειονομικής περίθαλψης [7].

2.2 Ανατομία του Ισχύου

Για την καλύτερη κατανόηση των καταγμάτων ισχύου, αξίζει να αναλύσουμε την ανατομία του ίδιου του ισχύου. Η άρθρωση του ισχύου είναι τύπου σφαίρας και λαβωτής τύπου διάρρθωσης (ball-in-socket) και αποτελείται από τη μηριαία κεφαλή και τον αυχένα. Η σταθερότητα της άρθρωσης στηρίζεται κυρίως στην οστική αρχιτεκτονική όπου απαρτίζεται από την Κοτύλη (the socket of the joint) και τη μηριαία κεφαλή (the ball of the joint) [8]. Το εγγυής μηριαίο αποτελείται από τη μηριαία κεφαλή, τον αυχένα του μηριαίου και τον μείζονα και ελάσσονα τροχαντήρα. Η μηριαία κεφαλή συνδέεται προς τα κάτω με τον άξονα μέσω του μηριαίου αυχένα, ο οποίος βρίσκεται μεταξύ του μείζονος και του ελάσσονα τροχαντήρα. Η γωνία που σχηματίζεται από τον αυχένα του μηριαίου και την έσω όψη του μηριαίου άξονα είναι

περίπου 127 μοίρες με έυρος από 120 εώς 140. Η μηριαία έκδοση σχηματίζεται από τη γωνία του άξονα μεταξύ του μηριαίου αυχένα και του διακονδύλιου μηριαίου άξονα. Μία σηματνική δομή, γνωστή και ως calcar femorale, είναι μια πυκνή σπογγώδης δοκός που εκτείνεται από τις οπίσθιες πλευρές του αυχένα του μηριαίου εώς τον οπίσθιο εγγύς μηριαίο άξονα. Αυτή η δομή παίζει κρίσιμο ρόλο στην παροχή δομικής υποστήριξης και στην κατανομή της πίεσης από την κεφαλή του μηριαίου στο εγγύς μηριαίο οστό. Έτσι, η παρουσία ή απουσία της παίζει σημαντικό ρόλο στην κατάλληλη επιλογή εμφυτεύματος για τη θεραπεία του κατάγματος ισχύου [7].

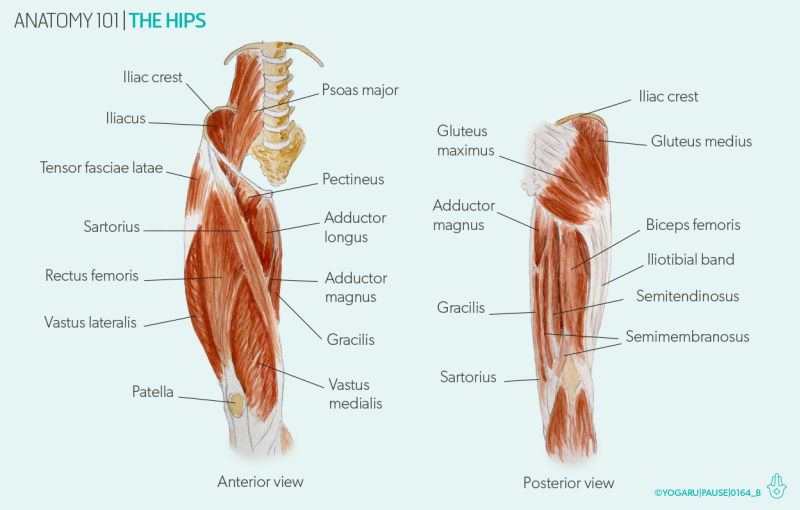
Μέσα στον αυχένα του μηριαίου οστού βρίσκονται οι συμπιεστικές και εφελκυστικές δοκίδες, οι οποίες σχηματίζουν το τρίγωνο Ward που συνδέεται προς τα επάνω από τις εφελκυστικές δοκίδες και κατωτέρω από τις συμπιεστικές δοκίδες. Πρόκειται για μία περιοχή χαμηλής οστικής πυκνότητας. Πρόσφατες μελέτες έχουν δείξει ότι ο εκφυλισμός των δοκίδων σχετίζεται στενά με την εμφάνιση καταγμάτων του αυχένα του μηριαίου και της διεύρυνσης του τριγώνου του Ward [7].



Σχήμα 2.1: (Αριστερά) Ακτινογραφία δεξιού ισχίου μιας γυναίκας 30 ετών. (Δεξιά) Ακτινογραφία δεξιού ισχίου ενός άνδρα 98 ετών. Το μέγεθος του τριγώνου Ward (WT) είναι σημαντικά μεγαλύτερο στη δεξιά εικόνα σε σύγκριση με την αριστερή εικόνα και υπάρχει μεγαλύτερος εκφυλισμός των βασικών συμπιεστικών δοκίδων (A) και των βασικών εφελκυστικών δοκίδων (B) στη δεξιά εικόνα σε σύγκριση με την αριστερή εικόνα [3]

Η γνώση της κατανομής και δράσης των μυϊκών ομάδων στο εγγύς είναι το κλειδί για την κατανόηση του τρόπου με τον οποίο οι δυνάμεις δρουν και παρεκτοπίζουν τα κατάγματα του εγγύς μηριαίου ώστε να επιτευχθεί ανάταξη και επιλογή οστεοσύνθεσης. Οι μύες του λαγονοψοΐτη, ο ορθός μηριαίος και ο σκαληνός είναι οι μύες που εμπλέκονται στην κίνηση του ισχίου. Οι μύες που εμπλέκονται στην επέκταση του ισχίου είναι ο μείζων γλουτιαίος και οι ισχίοι (ο ημιυμενώδης, ο ημιτενοντώδης και το μακρύ κεφάλι του δικέφαλου μηριαίου). Στη συνέχεια, η απομάκρυνση του ισχίου οφείλεται κυρίως στις ενέργειες του μέσου και ελάσσων γλουτιαίος, ενώ ο τείνων της πλατείας περιτονία βοηθά, επίσης, στην απομάκρυνση

του κάμπτοντος ισχίου [8]. Από την άλλη, η προσέγγιση του ισχίου οφείλεται κυρίως στην ενέργεια του ελάσσων προσαγωγού μυ, του μακρύ προσαγωγού μυ, του μέγα προσαγωγού, του κτενίτης, και του γρακιλίου. Η εξωτερική περιστροφή του ισχίου προκύπτει από την ενέργεια των έσω, έξω θυροειδής, διδύμου άνω και διδύμου κάτω, του τετρακέφαλου μηριαίου και του απιοειδή. Τέλος, η εσωτερική περιστροφή του ισχίου προέρχεται από τις δευτερεύουσες ενέργειες των προσόψεων ινών του μέσου γλουτέου και του ελάχιστου γλουτέου, καθώς και από τον τείνων της πλατείας περιτονία, τον ημιυμενώδη, τον ημιτενοντώδη, τον κτενίτη, και του πίσω μέρους του μεγάλου προσαγωγού [8], [9].



Σχήμα 2.2: Οι βασικοί μύες που είναι υπεύθυνοι για τις κινήσεις του ισχύου [10]

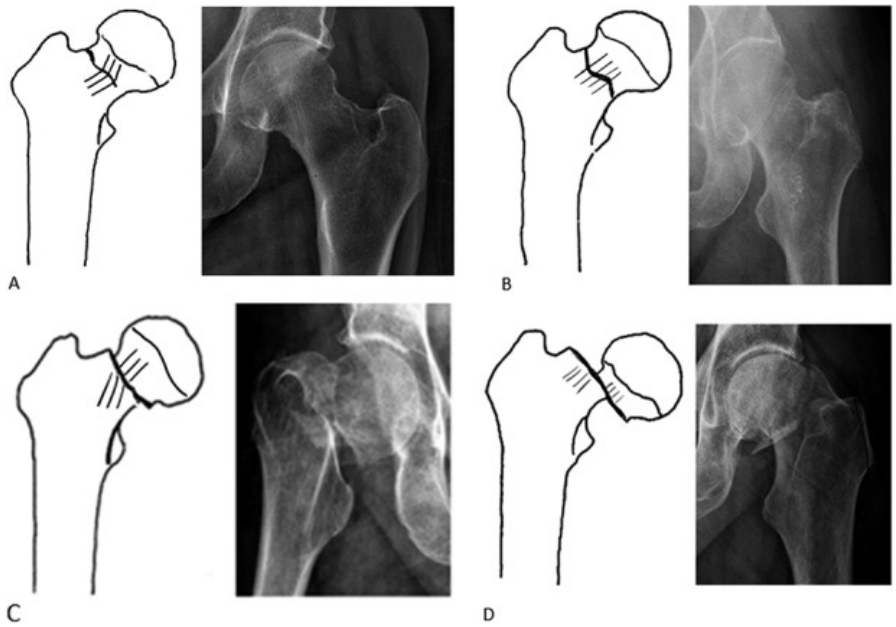
Η κατανόηση της παροχής αίματος στο εγγύς μηριαίο και ιδιαίτερα στη μηριαία κεφαλή είναι, επίσης, μεγάλης σημασίας για την εκτίμηση του κινδύνου οστεοκνέκρωσης. Η κύρια παροχή αίματος προέρχεται από την έσω μηριαία περισπωμένη αρτηρία που αρδεύει την φορτιζόμενη περιοχή της μηριαίας κεφαλής. Επιπλέον, η αρτηρία του στρογγυλού συνδέσμου προερχόμενη από τη θυροειδή αρτηρία ή την έσω μηριαία περισπωμένη είναι υπεύθυνη για την αιμάτωση [8].

2.3 Υποκεφαλικό Κάταγμα Ισχύου

Τα κατάγματα ισχίου ταξινομούνται ανάλογα με τη θέση τους σε σχέση με το θύλακο του ισχίου, οδηγώντας σε δύο κύριους τύπους: ενδοθυλακικά κατάγματα, όπως αυτά του αυχένα του μηριαίου, και τα εξωθυλακικά κατάγματα, συμπεριλαμβανομένων των διατροχαντήριων και υποτροχαντηριων καταγμάτων. Ακολούθως, θα εξετάσουμε συγκεκριμένα τα μοτίβα του αυχένα του μηριαίου και των διατροχαντηρικών καταγμάτων.

Τα υποκεφαλικά κατάγματα ταξινομούνται σύμφωνα με τα χαρακτηριστικά τους. Οι πιο συχνά αναφερόμενες ταξινομήσεις είναι του Garden ή του Pauwels. Η

ταξινόμηση Garden είναι ένα ευρέως αναγνωρισμένο σύστημα για την ταξινόμηση των υποκεφαλικών καταγμάτων ισχίου, και αξιολογεί την παρεκτόπιση του κατάγματος όπως αυτή προκύπτει από μία πρόσθια-οπίσθια και μία πλάγια μια ακτινογραφία ισχίου. Παραδοσιακά ο ταξινομητής αυτός κατηγοριοποιεί τα κατάγματα σε τέσσερις τύπους: Τα κατάγματα τύπου 1 είναι ατελή και ωθούνται σε βλασσότητα, τα κατάγματα τύπου 2 είναι πλήρη αλλά δεν έχουν παρεκτοπιστεί, τα κατάγματα τύπου 3 είναι εν μέρει παρεκτοπισμένα και τα κατάγματα τύπου 4 είναι πλήρως. Μια απλοποιημένη έκδοση αυτής της ταξινόμησης χωρίζει τα κατάγματα του αυχένα του μηριαίου οστού σε μη παρεκτοπισμένα ή όχι. Η ταξινόμηση αυτή παρουσιάζει αξιοπιστία και επαναληψιμότητα δικαιολογώντας έτσι την ευρεία χρήση της. Κυρίως, αυτό το σύστημα βοηθά στην επιλογή χειρουργικών θεραπειών, καθώς τα παρεκτοπισμένα κατάγματα συνήθως απαιτούν αντικατάσταση της άρθρωσης [7].

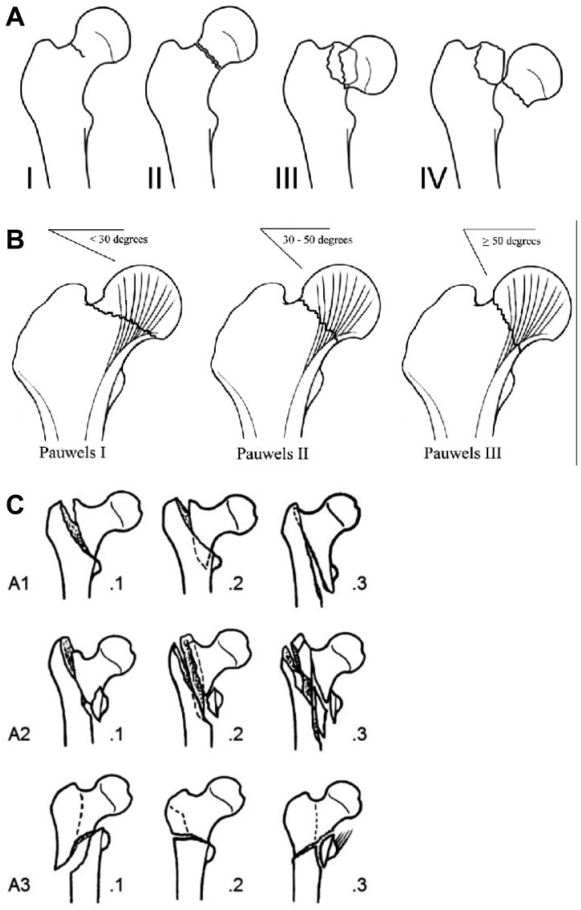


Σχήμα 2.3: Ακτινογραφίες που απεικονίζουν κατάγματα του μηριαίου αυχένα των τύπων Garden (Α) I, (Β) II, (Γ) III και (Δ) I΄΄ [11]

Η ταξινόμηση Pauwel κατηγοριοποιεί τα κατάγματα με βάση τη γωνία της καταγματικής γραμμής σε σχέση με το οριζόντιο επίπεδο. Υποδηλώνει ότι τα κατάγματα με μεγαλύτερη γωνία είναι πιο ασταθή επειδή υπόκεινται σε μεγαλύτερες διατμητικές δυνάμεις και ως εκ τούτου έχουν αυξημένες πιθανότητες οστεονέκρωσης. Αυτή η αστάθεια αυξάνει τον κίνδυνο οστικού θανάτου (οστεονέκρωση) μετά την επέμβαση. Σύμφωνα με αυτή την ταξινόμηση, τα κατάγματα τύπου I έχουν γωνία μικρότερη από 30 μοίρες, υποδηλώνοντας μικρότερη πιθανότητα παρεκτόπισης. Τα κατάγματα τύπου II κυμαίνονται από 30 έως 50 μοίρες, παρουσιάζοντας σχετικά μεγαλύτερη πιθανότητα αστάθειας, ενώ, τέλος, τα κατάγματα τύπου III, με γωνία

μεγαλύτερη από 50 μοίρες, θεωρούνται εξαιρετικά ασταθή [12].

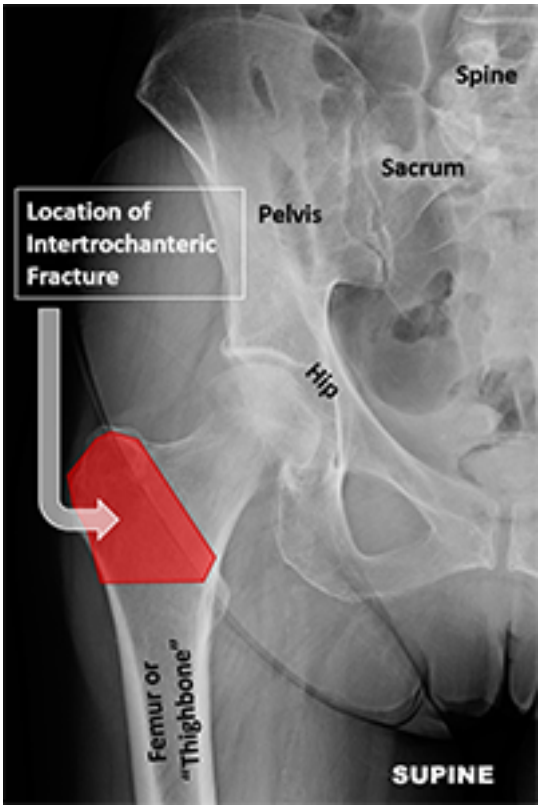
Μελέτες σχετικά με αυτές τις ταξινομήσεις έχουν δείξει, ωστόσο, ότι υπάρχει ανακολουθία αναφορικά με την αξιοπιστία των των ταξινομήσεων μεταξύ των ιατρών. Για παράδειγμα, μια μελέτη που περιελάμβανε χειρουργούς τραύματος και ορθοπαιδικούς που αξιολογούσαν τις ακτινογραφίες απέδειξε ότι οι απόψεις των ειδικών δεν συμφωνούσαν απόλυτα με την ταξινόμηση Pauwels. Μια άλλη μελέτη συνέκρινε την αξιοπιστία των συστημάτων Garden, Pauwels και AO, κατέληξε ότι η ταξινόμηση Garden είχε την υψηλότερη αξιοπιστία. Παρά την ευρεία εφαρμογή τους και τις πολυάριθμες μελέτες για τα κατάγματα του ισχίου, υπάρχει περιορισμένη έρευνα σχετικά με την αξιοπιστία αυτών των συστημάτων ταξινόμησης, ειδικά μεταξύ έμπειρων γιατρών. Στη μελέτη αυτή, θα επικεντρωθούμε περισσότερο στις ταξινομήσεις Garden.



Σχήμα 2.4: Ταξινόμηση του υποκεφαλικού κατάγματος ισχύου βάσει των ταξινομητών: Garden (A), Pauwels (B), και AO/OTA (C) [13]

2.4 Διατροχαντήρια Κατάγματα

Τα Διατροχαντήρια κατάγματα είναι εξωαρθρικά. Αυτή η περιοχή χαρακτηρίζεται από την παρουσία σπογγώδους οστού. Στον μείζωνα τροχαντήρα καταρύονται αρκετοί μύες, συμπεριλαμβανόμενου και του μέσου γλυοτιαίου και του μικρού γλουτιαίου και είναι, επίσης, το σημείο έκλυσης για έναν άλλο μυ, τον έξω πλάγιο μυ. Ο ελάσσον τροχαντήρας είναι το σημείο που καταρύεται ο λαγονοψοΐτης. Η ευρεία περιοχή γύρω από το σημείο του κατάγματος έχει καλή παροχή αίματος, η οποία βοηθά στην επούλωση και μειώνει τον κίνδυνο οστεονέκρωσης σε σύγκριση με τα κατάγματα στον αυχένα του μηριαίου οστού [9]. Τα διατροχαντήρια και υποτροχαντήρια κατάγματα είναι συγκεκριμένοι τύποι και εμπίπτουν στην ευρύτερη κατηγορία των περιτροχαντηρίων καταγμάτων.



Σχήμα 2.5: Απεικόνιση της ανατομίας όπου φαίνεται η περιοχή του Μεσοτροχαντηρίου [14]

Ο συγκεκριμένος τύπος κατάγματος είναι πιο κοινός στον ηλικιωμένο πληθυσμό λόγω της οστεοπόρωσης σε συνδυασαμό με τον μηχανισμό χαμλής ενέργειας. Βάσει της σύχγρονης βιβλιογραφίας, έχει παρατηρηθεί ότι ο λόγος των περιστατικών με γυναίκες προς των περιστατικών με άντρες είναι ανάμεσα στο 2:1 με 8:1. Συνήθως, αυτά τα κατάγματα προκύπτουν σε ασθενείς που υποφέρουν από ισχαιμικό κάταγμα ισχύου, ενώ για τον νεότερο πληθυσμό προκύπτουν από μηχανισμούς

υψηλής ενέργειας [9].

Όπως τα υποκεφαλικά κατάγματα ισχύου, τα κατάγματα μεσοτροχαντηρίου μηριαίου οστού συσχετίζονται, επίσης, με υψηλή νοσηρότητα και θνησιμότητα. Έχει υπολογιστεί ότι προκύπτουν χονδρικά 180.000 κατάγματα τέτοιου είδους ετησίως, και μέχρι το 2040 εκτιμάται να αυξηθούν στα 500.000 περιστατικά [9].

Για τα διατροχαντήρια κατάγματα του ισχίου, χρησιμοποιείται η ταξινόμηση Evans η οποία εξετάζει τη θέση, την κατεύθυνση και τη σταθερότητα του κατάγματος. Ωστόσο, το να προσιοριστεί εάν το κάταγμα είναι σταθερό ή ασταθές είναι ζωτικής σημασίας γιατί επηρεάζει το χειρουργικό εμφύτευμα που θα χρησιμοποιηθεί. Η σταθερότητα εξαρτάται από την κατάσταση του μηριαίου ασβεστίου ή του πίσω μέρους του άνω άκρου του μηριαίου οστού. Κατάγματα με αντίστροφη κλίση ή εγκάρσια κατάγματα κατά μήκος του τροχαντήρα συνήθως θεωρούνται ασταθή επειδή τείνουν να μετατοπίζονται προς τα μέσα [7].

Το AO/OTA ταξινομεί αυτά τα κατάγματα ως 31-A, διαιρώντας τα περαιτέρω με τη σταθερότητα και τις συγκεκριμένες λεπτομέρειες κατάγματος: 31-A1 για σταθερά κατάγματα, 31-A2 για ασταθή και 31-A3 για κατάγματα ανάστροφης κλίσης ή εκείνα που επηρεάζουν την εξωτερική πλευρά του οστού. Ενώ το σύστημα AO/OTA είναι λεπτομερές και αξιόπιστο, χρησιμοποιείται κυρίως για έρευνα [7].

2.5 Ο Ρόλος της Οστεοπόρωσης στα Κατάγματα Ισχύου

Καθώς οι άνθρωποι μεγαλώνουν, ο μηριαίος λαιμός τους υφίσταται αλλαγές που αυξάνουν τον κίνδυνο καταγμάτων. Το οστό γίνεται πιο πορώδες, αυξάνοντας από 4% πορώδες στα νεαρά άτομα σε σχεδόν 50% στους ηλικιωμένους. Επιπλέον, μικρές ρωγμές συσσωρεύονται στο οστό με την πάροδο του χρόνου, με αυτή τη διαδικασία να εμφανίζεται πιο γρήγορα στις γυναίκες από ότι στους άνδρες. Ένας άλλος παράγοντας που διακυβεύει την ακεραιότητα των οστών είναι η μη ενζυματική σκλήρυνση του κολλαγόνου των οστών. Αυτές οι αλλαγές καθιστούν το οστό λιγότερο ελαστικό, πιο αδύναμο και πιο επιρρεπές σε κατάγματα ακόμη και από κρούσεις χαμηλής ενέργειας. Το οστό στον αυχένα του μηριαίου οστού γίνεται, επιπλέον, πιο λεπτό με την ηλικία, ειδικά στην επάνω πλευρά, το οποίο είναι λιγότερο καταπονημένο και επομένως πιο ευαίσθητο σε κατάγματα [7].

2.6 Θεραπεία

Ο στόχος της χειρουργικής αντιμετώπισης των καταγμάτων ισχίου στους ηλικιωμένους ασθενείς είναι να επιτρέψει την άμεση κινητοποίηση και φόρτιση βάρους καθώς με αυτόν τον τρόπο μειώνει σημαντικά τον κίνδυνο να προκύψουν μετεγχειρητικές επιπλοκές και βελτιώνει το ποσοστό θνησιμότητας. Κατά την πλειο-

ψηφία των περιπτώσεων, ο ασθενής οδηγείται σε επέμβαση με εξαίρεση συγκεκριμενες περιπτώσεις κατά τις οποίες παρουσιάζει σοβαρές συννοσηρότητες που αυξάνουν σημαντικά τον κίνδυνο θνησιμότητας για τον ασθενή. Οι περισσότερες έρευνες έχουν ωστόσο αποδείξει ότι τα ποσοστά ανάρρωσης των ασθενών που έχουν υποβληθεί χειρουργική επέμβαση είναι μεγαλύτερα από εκείνα των ασθενών που είχαν μη-επεμβατική αντιμετώπιση.

Χειρουργική Θεραπεία για τα Υποκεφαλικά Κατάγματα Ισχύου

Η θεραπεία που θα ακολουθήσει προκειμένου να αντιμετωπιστεί κάποιο υποκεφαλικό κάταγμα εξαρτάται από το αν το κάταγμα έχει μετακιηθεί από τη θέση του ή όχι. Κατά κανόνα, τα κατάγματα που έχουν μετατοπισθεί αντιμετωπίζονται με χειρουργική επέμβαση, συγκεκριμένα με την αντικατάσταση της άρθρωσης στους ηλικιωμένους. Ωστόσο, για περιπτώσεις όπου το κάταγμα έχει μετακινηθεί ελάχιστα εώς και καθόλου (όπως στις περιπτώσεις Garden I και Garden II) μπορεί να αντιμετωπιστεί με κατάλληλες βίδες ή μία συσκευή που ονομάζεται συρόμενη βίδα ισχύου. Οι βίδες είναι συνήθως διατεταγμένες σε σχήμα τριγώνου για να παρέχουν ισχυρή και σταθερή στήριξη που επιτυχγάνεται με το να αγγίζουν το εξωτερικό στρώμα του οστού. Προτείνεται, ακολούθως, η χρήση μεταλλικών ροδέλων με τις βίδες σε άτομα με αδύναμα οστά με στόχο την διαμέριση των δυνάμεων και την βελτίωση της αποτελεσματικότητας της θεραπείας, μειώνοντας τηνν πιθανότητα χαλάρωσης των βιδών. Παρόλο που σε συγκεκριμένες περιπτώσεις περίπλοκων καταγμάτων η προσθήκη μίας επιπρόσθετης βίδας μπορεί να βοηθήσει, η χρήση αυτής της μεθόδου ακόμα εξετάζεται από πολλούς ειδικούς [7]. Η συρόμενη βίδα, η οποία ευνοείται για τη σταθερότητα της, λειτουργεί επτρέποντας σε μία βίδα μέσα σε ένα μεταλλικό χιτώνιο να κινηθεί, βοηθώντας στη συμπίεση και την επούλωση των οστών [15]. Μελέτες προτείνουν ότι αυτή η μέθοδος μπορεί να προσφέρει καλύτερη σταθεροποίηση, ειδικά για κατάγματα με υψηλό κίνδυνο μετατόπισης κοντά στη βάση του αυχένα του μηριαίου.

Πολλαπλές έρευνες έχουν εξετάσει τη χρήση της συρόμενης βίδας που συνοδεύονται από αντίστοιχες επιτυχίες σε σχέση με άλλες μεθόδους, ωστόσο, παρουσίασαν ιδιαίτερο προβάδισμα για συγκεκριμένες ομάδες ασθενών, όπως σε καπνιστές ή σε ασθενείς με κατάγματα σε συγκεκριμένες περιοχές. Παρόλα αυτά, δεν έχει εδραιωθεί κάποια συγκεκριμένη στρατηγική και προσέγγιση που να αφορά την αντιμετώπιση του κατάγματος πέραν ότι βασίζεται κυρίως από την κρίση του ειδικού που θα διαχειριστεί το περιστατικό.

Για κατάγματα με μεγάλες μετατοπίσεις, η αρθροπλαστική (αντικατάσταση της άρθρωσης) προτιμάται στους ηλικιωμένους για να ελαχιστοποιηθεί ο κίνδυνος οστικού θανάτου στη θέση του κεφαλιού του μηριαίου. Μερικές από τις επιλογές που μπορεί να ακολουθήσει ο ειδικός είναι η ολική αρθροπλαστική (THA) [16], η οποία αντικαθιστά τα πιο πολλά συστατικά της άρθρωσης και μπορεί να προσφέρει καλύτερη λειτουργία σε εκείνους τους ασθενείς που είναι πιο δραστήριοι και συνήθως

νεότεροι [17], ή η ημιαρθροπλαστική (ΗΑ), η οποία είναι απλούστερη και λιγότερο κοστοβόρα, αλλά ενδέχεται να χρειαστεί στο μέλλον επιπλέον επέμβαση. Στη συνέχεια, όσον αφορά την αρθροπλαστική, υπάρχουν δύο είδη υλικών που μπορούν να χρησιμοποιηθούν: τα τσιμεντοειδή υλικά και τα μη-τσιμεντοειδή. Η επιλογή μεταξύ τσιμεντοειδούς και μη τσιμεντοειδούς ημιαρθρωπλαστικής εξαρτάται από τον κίνδυνο επιπλοκών που σχετίζονται με την επέμβαση και των μακροπρόθεσμων αποτελεσμάτων, έχοντας υπόψην ότι στις περισσότερες περιπτώσεις, η τσιμεντοειδής προσέγγιση οδηγεί σε λιγότερα προβλήματα που σχετίζονται με τα εμφυτεύματα [18].

Χειρουργική Αντιμετώπιση Μεσοτροχαντηριακών Καταγμάτων Μηριαίου Οστών

Η επιλογή εμφυτεύματος προκειμένου να αντιμετωπιστεί το μεσοτροχαντηριακό κάταγμα του μηριαίου οστού εξαρτάται από το πόσο σταθερή είναι η κατάσταση του εξωτερικού στρώματος του ίδιου του οστού [7]. Για τα πιο σταθερά κατάγματα, επιλογές όπως το συρόμενη βίδα είναι αρκετά αποτελεσματική, ενώ σε μη σταθερά κατάγματα, συνήθως απαιτείται μία ενδομυελική συσκευή (σαν μία εσωτερική ράβδο) με στόχο τη σταθεροποίηση του κατάγματος [19]. Η ενδομυελική συσκευή προσφέρει καλύτερη σταθερότητα λόγω του ότι ταιριάζει καλύτερα με τις φυσικές δυνάμεις του μηρού, με αποτέλεσμα να προσφέρει ισχυρότερη αντίσταση στην πίεση και να μειώνει την πιθανότητα μετακίνησης του κατάγματος [20], [21].

Σχήμα 2.6: Σταθερό Κάταγμα [22]

Σε περιπτώσεις που το κάταγμα επηρεάζει και την εξωτερική πλευρά του μηριαίου οστού, μία ενδομυελική συσκευή μπορεί να αποτελέσει κρίσιμη λόγω του ρόλου της ως εσωτερική στήριξη που αποτρέπει το κάταγμα από το να χειροτερεύσει. Πολλαπλές μελέτες έχουν οδηγηθεί στο συμπέρασμα ότι για συγκεκριμένους τύπους

Σχήμα 2.7: Ασταθές Κάταγμα [22]

καταγμάτων, όπως εκείνα που εξαπλώνονται και σε άλλες περιοχές ή συμπεριλαμβάνουν και το πλευρικό τοίχωμα των οστών, μπορεί η χρήση των συρόμενων βιδών να οδηγήσουν στη χειροτέρευση του κατάγματος καθώς ενδέχεται να προκαλέσει επιπλοκές, όπως, για παράδειγμα, την μετατόπιση προς τα έσω του μηριαίου οστού [23]. Μία κοινή επιπλοκή της συρόμενης βίδας είναι η μετατόπιση του μηριαίου οστού σε μία πιο γωνιακή θέση, με αποτέλεσμα η βίδα να καρφωθεί μέσα στο οστό.

Ουσιαστικά, σε αυτές τις περιπτώσεις, η πιο συνήθης προσέγγιση είναι η επέμβαση, συγκερκιμένα με την τοποθέτηση συσκευών και εμφυτευμάτων, όσον αφορά την αντιμετώπιση τους. Εκεί που διαφέρει η προσέγγιση, ωστόσο, είναι στην επιλογή των υλικών που θα αξιοποιηθούν και θα τοποθετηθούν στον ασθενή κατά τη διάρκεια του χειρουγείου. Η κατάλληλη επιλογή εξαρτάται από την σοβαρότητα του κατάγματος αλλά και άλλων χρακτηριστικών του ίδιου του κατάγματος. Συνολικά, και για τις δύο περιπτώσεις καταγμάτων, η επιλογή θεραπείας εξαρτάται από τον τύπο του κατάγματος, την ηλικία του ασθενούς, το επίπεδο δραστηριότητας και την κατάσταση της υγείας του ασθενούς [6].

2.7 Μηχανική Μάθηση στη Βιοϊατρική

Είναι γεγονός ότι η Τεχνητή Νοημοσύνη, και κατά επέκταση η Μηχανική Μάθηση, δυεισδύει ολοένα και περισσότερο στην καθημερινότητα μας και εφαρμόζεται σε πολλαπλούς τομείς και σύντομα θα γίνει αναπόσπαστο κομμάτι οποιασδήποτε επιστήμης. Πρόκειται για έναν κλάδο της επιστήμης των υπολογιστών, ο οποίος επιδιώκει μιμηθεί τη διαδικασία μάθησης που ακολουθούν τα έμβια όντα [24]. Μπορεί να υποστηριχθεί ότι ο κλάδος αυτός εξελίχθηκε από τη διαδικασία αναγνώρισης προτύπων και μοτίβων και την υπολογιστική θεωρία μάθησης της τεχνητής νοημοσύνης [25]. Έτσι, η μηχανική μάθηση κατασκευάζει αλγόριθμους που μπορούν να αντλήσουν χρήσιμη πληροφορία από έναν μεγάλο όγκο δεδομένων και να παράγει

εκτιμήσεις για τα δεδομένα αυτά. Οι αλγόριθμοι αυτοί εκπαιδεύονται μέσω των στατιστικών συσχετίσεων ή άλλων μοτίβων που βρίσκουν από τα δεδομένα που τους τροφοδοτούνται [26]. Η μηχανική μάθηση χρησιμοποιείται για την επίλυση απλών ως και πολύπλοκων προβλημάτων αλλά και την υλοποίηση διαφόρων μεθόδων και διεργασιών, όπως στην αναγνώριση εικόνας, ανίχνευση αντικειμένων, αναγνώριση προσώπου, επεξεργασία ιατρικής εικόνας, και πολλά ακόμα. Πολλοί υποστηρίζουν ότι η μηχανική μάθηση δεν είναι ένας καινούργιος κλάδος της επιστήμης των υπολογιστών, αλλά προυπήρχε και διαρκώς εξελισσόταν καθώς αναπτύσσονταν νέοι αλγόριθμοι, νέες τεχνικές και τομείς όπως το Big Data που έχει αλλάξει ριζικά τον κόσμο της Τεχνητής Νοημοσύνης.

Συγκεκριμένα στον τομέα της Βιοϊατρικής υφίστανται πολλαπλές μέθοδοι μηχανικής μάθησης, όπου κατά κύριο λόγο αξιοποιούνται μοντέλα βαθιάς μάθησης [25]. Οι κλινικοί έρχονται αντιμέτωποι με διάφορες πηγές πληροφορίας για την υγεία των ασθενών, για παράδειγμα από ηλεκτροκαρδιογραφήματα, εξετάσεις αίματος, επίπεδα οξυγόνου, ακτινογραφίες, και πολλά άλλα. Παρόλα αυτά, αποτελεί πρόκληση να αξιοποιήσουν κάθε δεδομένο που τους παρέχεται για κάθε ασθενή προκειμένου να πραγματοποιήσει κάποια διάγνωση ή να αποφασίσει ποια θερπαεία θα ακολουθήσει ο ασθενής. Έτσι, συνήθως, καταλήγει ο κλινικός να αξιοποιεί ένα ελάχιστο κομμάτι από όλον τον όγκο πληροφορίας που έχει συλλεγεί για κάθε ασθενή [27]. Σε αυτό το σημείο μπορεί να συμβάλλει η μηχανική μάθηση που μπορεί να προσφέρει μία αυτοματοποιημένη ανάλυση λαμβάνοντας υπόψη ένα μείγμα από ετερογενείς πηγές μετρήσεων, βελτιώνοντας έτσι την ακρίβεια της διάγνωσης αλλά και βοηθούν τους ειδικούς να λαμβάνουν καταλληλότερες αποφάσεις. Τελευταία, οι ερευνητές εστιάζουν σε τεχνικές επεξεργασίας ιατρικών εικόνων για την εύρεση και ταξινόμηση διαφόρων καρκίνων. Άλλες έρευνες επικεντρώνονται σε προβλέψεις εμφάνισης ασθενειών, όπως του Πάρκινσον, όπου εκεί τροφοδοτούνται σε πολλά μοντέλα μηχανικής μάθησης δεδομένα με διάφορα χαρακτηριστικά για κάθε ασθενή και συγκρίνονται οι επιδόσεις τους. Μερικά από τα πιο χρησιμοποιημένα μοντέλα στον τομέα της βιοϊατρικής είναι: k nearest-neighbors (KNN), Naïve Bayes (NB), regression treets (RT), Νευρωνικά Δίκτυα, όπως τα CNN, AlexNet και πολλά ακόμα, και Support Vector Machines (SVM).

2.8 Τρέχουσα κατανόηση των προβλεπτικών παραγόντων για τη μετεγχειρητική αποκατάσταση σε ασθενείς με κάταγμα ισχίου

Έχουν διεξαχθεί αρκετές state-of-the-art έρευνες που αποσκοπούν στον εντοπισμό των παραγόντων που μπορούν να συμβάλλουν στην πρόβλεψη της μετεγχειρητικής περιόδου ενός ασθενούς με κάταγμα ισχύου κατόπιν επέμβασης αλλά και να εκτιμήσουν ορισμένες περιπλοκές που μπορούν να προκύψουν, όπως, για

παράδειγμα ο θάνατος του ασθενούς. Οι περισσότερες έρευνες αυτές αξιοποιούν μοντέλα τεχνητής νοημοσύνης τα οποία και τροφοδοτούν με ιατρικά δεδομένα που παρέχουν γενικές και ειδικές πληροφορίες από ασθενείς που έχουν υποβληθεί θεραπεία για την αντιμετώπιση των καταγμάτων αυτών. Βάσει των ερευνών αυτών, οι πιο δημοφιλείς προβλέψεις που κατάφεραν να εκτιμήσουν είναι η διάρκεια διαμονής του ασθενούς στο νοσοκομείο κατόπιν του χειρουργείου και ο θάνατος μέχρι και έναν χρόνο μετά το χειρουργείο [28].

Προκειμένου να μπορέσουν να κάνουν αυτές τις εκτιμήσεις, οι ερευνητές ανέπτυξαν και χρησιμοποίησαν πολλαπλά μοντέλα μηχανικής μάθησης, με το πιο δημοφιλές από αυτά να είναι το στατιστικό μοντέλο Logistic Regression, με στόχο την ταξινόμηση των δεδομένων εισόδου σε πιθανά μετεγχειρητικά σενάρια, όπως αυτό του θανάτου. Άλλα μοντέλα τα οποία αξιοποιήθηκαν και τα συνέκριναν μεταξύ τους ήταν τα Δέντρα Απόφασης, τα Νευρωνικά Δίκτυα, τα Random Forests, ο αλγόριθμος των Κ Πλησιέστερων Γειτόνων, τα Μηχανήματα των Διανυσμάτων Υποστήριξης, το μοντέλο Gradient Boosting και ο ταξινομητής Naive Bayes.

Μέσω των μοντέλων αυτών αποδείχθηκε ότι σε πολλές έρευνες παράγοντες όπως το βάρος, το BMI, η ηλικία, το κάπνισμα, πιθανές λοιμώξεις από την επέμβαση (SSI), το φύλο, η τεχνική που χρησιμοποιήθηκε κατά το χειρουργείο, η νοητική έκπτωση, η διάρκεια από την ώρα του τραυματισμού μέχρι τη νοσηλεία, και άλλα επικείμενα νοσήματα που είχε ο κάθε ασθενής παίζουν καθοριστικό ρόλο στην ανάρρωσή του. Ένας ακόμη σημαντικός παράγοντας αποτελεί και η κινητικότητα του ασθενούς πριν τον τραυματισμό και, προφανώς, ύστερα από την επέμβαση.

Πέραν της εκτίμησης της μετεγχειρητικής περιόδου του κάθε ασθενούς, ο εντοπισμός των παραγόντων που μπορούν να μας προϊδεάσουν για την κάθε έκβαση είναι εξίσου σημαντικός. Η αξιοποίηση των εκτιμήσεων αυτών αλλά και η κατανόηση της βαρύτητας του κάθε παράγοντα μπορούν να βοηθήσουν τους ιατρούς και ειδικούς να λαμβάνουν ακόμα πιο ενημερωμένες αποφάσεις βάσει δεδομένων όσον αφορά την αντιμετώπιση του κάθε περιστατικού.

Κεφάλαιο 3

Υλικό και Μέθοδοι

3.1 Επιβλεπόμενη Μάθηση

Οι πιο γνωστές κατηγορίες που ταξινομείται η μηχανική μάθηση είναι η επιβλεπόμενη και η μη-επιβλεπόμενη μάθηση. Η ειδοποιός διαφορά μεταξύ των δύο βρίσκεται στον τρόπο που εκπαιδεύεται ένας αλγόριθμος μηχανικής μάθησης. Οι αλγόριθμοι της επιβλεπόμενης μάθησης παράγουν μία συνάρτηση που συνδέει τις εισόδους με τις επιθυμητές εξόδους. Δηλαδή, ένας αλγόριθμος χρησιμοποιεί ήδη προσδιορισμένες ετικέτες προκειμένου να ερμηνεύσει τη σχέση που υπάρχει μεταξύ των δεδομένων εισόδου με την έξοδο. Τα πιο συνήθη προβλήματα που επιλύουν οι αλγόριθμοι επιβλεπόμενης μάθησης είναι τα προβλήματα ταξινόμησης. Στα προβλήματα ταξινόμησης, οι κλάσεις στις οποίες θα ταξινομηθούν τα δεδομένα είναι προκαθορισμένες. Σε αυτό το κομμάτι είναι χρήσιμο να διακρίνουμε και τη διαφορά μεταξύ των δύο κύριων μοντέλων επιβλεπόμενης μάθησης: τα μοντέλα ταξινόμησης και τα μοντέλα παλινδρόμησης. Η δεύτερη κατηγορία μοντέλων αντιστοιχίζουν τις εισόδους σε πεδία πραγματικών αριθμών, ενώ οι ταξινομητές, όπως αναφέρθηκε και παραπάνω, αντιστιχίζουν τις εισόδους σε προκαθορισμένες κλάσεις.

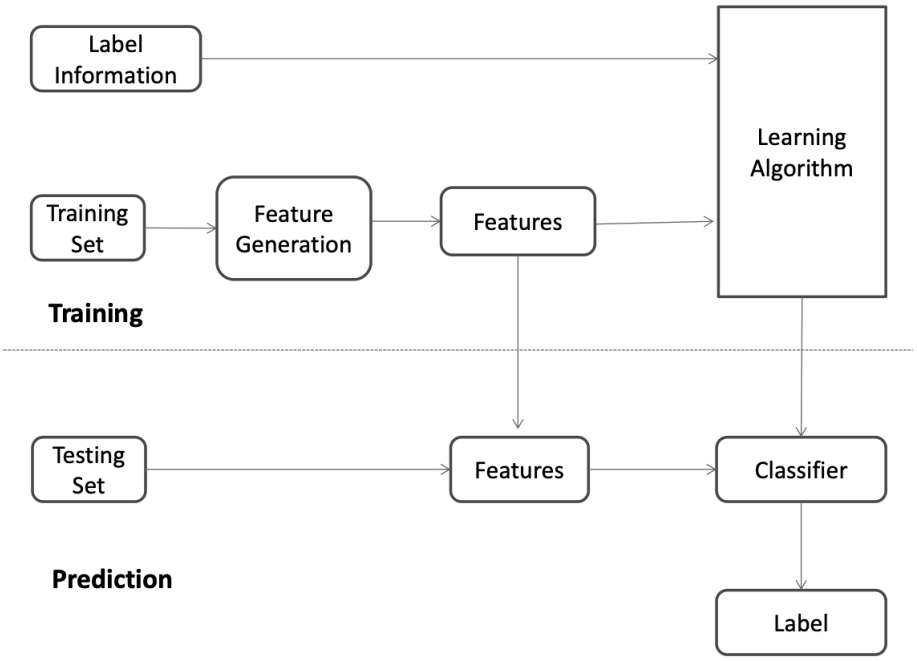
Στην επιβλεπόμενη μάθηση, ο αλγόριθμος ακολουθεί μία διαδικασία κατά την οποία προσπαθεί να προβλέψει την ετικέτα ενός αντικειμένου βάσει κάποιου συνόλου χαρακτηριστικών. Πιο συγκεκριμένα, ο αλγόριθμος προσπαθεί να κατασκευάσει μαθηματικά μοντέλα προκειμένου να ¨ταιριάξει¨ τα δεδομένα εισόδου στις αντίστοιχες κλάσεις αυτές. Στη συνέχεια, ο αλγόριθμος δέχεται ένα σύνολο χαρακτηριστικών ως εισόδο μαζί με τις ¨σωστές¨ εξόδους και ¨εκπαιδεύεται¨ συγκρίνοντας την ¨παραγόμενη¨ έξοδο του με την ¨σωστή¨ (προκαθορισμένη) έξοδο με στόχο να εντοπίσει τα λάθη. Βάσει των λαθών αυτών, ο αλγόριθμος προσαρμόζεται. Τα μοντέλα, τελικά, που έχει αναπαράξει ο αλγόριθμος αξιολογούνται βάσει κάποιων μετρικών που χρησιμοποιούνται συνήθως στη σταστική [26].

3.2 Επιλογή Χαρακτηριστικών

Στον τομέα της μηχανικής μάθησης, μια αρκετά σημαντική και ξεχωριστή διαδικασία από την διαδικασία εκπαίδευσης και πρόβλεψης μίας κατάστασης και πραγματοποιείται κατά την προεπεξεργασία των δεδομένων είναι η επιλογή των χαρακτηρστικών [29], [30]. Κατά τη διαδικασία αυτή επιλέγονται τα χαρακτηριστικά ή αλλιώς στήλες που θα τροφοδοτηθούν στα μοντέλα μηχανικής μάθησης. Ο λόγος για τον οποίον είναι μερικές φορές απαραίτητη η επιλογή χαρακτηριστικών είναι επειδή πολλές φορές, η αναπαράσταση των δεδομένων γίνεται με πολλαπλά χαρακτηριστικά, τα οποία μπορεί να μην σχετίζονται όλα άμεσα με την έννοια-στόχο [29]. Είναι βασικό να εμπεδώσουμε ότι η πολλή πληροφορία δεν μεταφράζεται σε ποιοτική πληροφορία καθώς μπορούν συγκεκριμένα χαρακτηριστικά να μην συμβάλλουν στην αναπαραγωγή κάποιας πρόβλεψης ή και να θεωρούνται περιττά, με αποτέλεσμα να μην μας κατευθύνουν προς τα σωστά συμπεράσματα. Για τη διαδικασία αυτήν υπάρχουν πολλές μέθοδοι οι οποίες βάσει κάποιων συναρτήσεων ή μέσω του υπολογισμού των συσχετίσεων των χαρακτηρσιτικών με την επιθυμητή έξοδο, επιλέγουν τα χαρακτηρισιτικά τα οποία θεωρούν πιο ¨χρήσιμα¨ και θα οδηγήσουν στην ακριβέστερη πρόβλεψη του κάθε μοντέλου. Ειδικά σε σύνολα δεδομένων πολλών διαστάσεων η διαδικασία επιλογής χαρακτηριστικών είναι ένας αποτελεσματικός τρόπος περιορισμού της περιττής πληροφορίας και μπορεί να μειώσει τον χρόνο υπολογισμού, να βελτιώσει την ακρίβεια της εκμάθησης και να συμβάλλει στην καλύτερη κατανόηση των δεδομένων [31]. Επομένως, μία πολύ βασική διαδικασία προετοιμασίας των δεδομένων πρωτού γίνει κάποια πρόβλεψη από τα μοντέλα είναι η επιλογή της σχετικής πληροφορίας.

Είναι σημαντικό να διευκρινιστεί η σημασία του ρόλου της διαδικασίας επιλογής χαρακτηριστικών σε προβλήματα ταξινόμησης και για αυτό είναι κρίσιμο να κατανοήσουμε τον ρόλο των ίδιων των χαρακτηριστικών στη διαδικασία εκπαίδευσης ενός μοντέλου που προτίθεται για ταξινόμηση. Καταρχάς, όπως ήδη είναι γνωστό, ένα πρόβλημα ταξινόμησης στοχεύει να χαρτογραφήσει μία νέα παρατήρηση σε κάποια από τις υπάρχουσες κατηγορίες βάσει την εκπαίδευση που έχει πραγματοποιήσει το μοντέλο πάνω στα δεδομένα εκπαίδευσης των οποίων η κατηγορίες είναι γνωστές εκ των προτέρων. Στη φάση εκπαίδευσης, τα δεδομένα αναλύονται σε ένα σύνολο χαρακτηριστικών με βάση τα μοντέλα δημιουργίας χαρακτηριστικών, όπως το μοντέλο διανυσματικού χώρου για δεδομένα κειμένου. Αυτά τα χαρακτηριστικά μπορεί είτε να είναι κατηγορικά (για παράδειγμα για τύπο αίματος), κανονικά (για παράδειγμα ¨μεγάλο¨, ¨μεσαίο¨ ή ¨μικρό¨), ακέραια (για παράδειγμα ο αριθμός των εμφανίσεων ενός μέρος λέξη σε ένα email) ή δεκαδικά (για παράδειγμα μια μέτρηση της πίεσης αίματος). Αφού αναπαραστήσει τα δεδομένα μέσω αυτών των εξαγόμενων χαρακτηριστικών, ο αλγόριθμος μάθησης θα χρησιμοποιήσει τις πληροφορίες των ετικετών καθώς και τα ίδια τα δεδομένα για να μάθει μια συνάρτηση χαρτογράφησης f (ή έναν ταξινομητή) από τα χαρακτηριστικά στις ετικέτες ως εξής: $f(features) \rightarrow labels$. Στη φάση της πρόβλεψης, τα δεδομένα αναπαρίστανται α-

πό ένα σύνολο εξαγόμενων χαρακτηριστικών που παράχθηκε κατά τη διαδικασία της εκπαίδευσης και στη συνέχεια η χαρτογράφηση θα εφαρμοστεί πάνω στα νέα δεδομένα προκειμένου να τους αντιστοιχίσει κάποια ετικέτα.



Σχήμα 3.1: Αναπαράσταση διαδικασίας ταξινόμησης

Η διαδικασία επιλογής χαρακτηριστικών υφίσταται σε διάφορους τομείς της μηχανικής μάθησης, όπως στην αναγνώριση εικόνων, στην ανάκτηση εικόνας, στην εξόρυξη κειμένου, στην ανάλυση δεδομένων βιοπληροφορικής, και σε πολλούς ακόμα [31]. Υπάρχουν τρείς κατηγορίες που εφαρμόζονται για την επιλογή χαρακτηριστικών: η επιβλεπόμενη, η μη-επιβλεπόμενη και η ημι-επιβλεπόμενη, όπως και οι κατηγορίες της μηχανικής μάθησης. Η πρώτη κατηγορία, η επιβλεπόμενη επιλογή χαρακτηριστικών, χρησιμοποεί σαν βασικό κριτήριο τη συσχέτιση και τη σχετικότητα μεταξύ των χαρακτηριστικών και της ετικέτας εξόδου και συχνά χρησιμοποιείται για προβλήματα ταξινόμησης. Η σημασία (importance) των χαρακτηριστικών μπορεί να αξιολογηθεί από ποικίλες μετρικές και ο στόχος της είναι να βρεθεί ο βέλτιστος συνδυασμός χαρακτηριστικών, ο οποίος αποτελεί υποσύνολο του αρχικού συνόλου, και να μεγιστοποιήσει την ακρίβεια της ταξινόμησης. Από την άλλη, οι μη επιβλεπόμενες μέθοδοι επιλογής χαρακτηριστικών στοχεύουν να ανακαλύψουν την φύση των ταξινομημένων δεδομένων και να βελτιώσουν την ακρίβεια συσταδοποίησης μέσω της εύρεσης ενός υποσυνόλου χαρακτηριστικών βάσει είτε κάποιου αλγόριθμου συσταδοποίησης ή μέσω κάποιου κριτηρίου αξιολόγησης. Τέλος, η ημι-επιβλεπόμενη επιλογή χαρακτηριστικών χρησιμοποιείται κυρίως για την ημι-επιβλεπόμενη μάθηση κατά την οποία υπάρχουν δείγματα δεδομένων με D_i και χωρίς D_u ετικέτες, και το δείγμα D_u χρησιμοποιείται προκειμένου να βελτιώσει την διαδικασία εκμάθησης του μοντέλου που έχει ήδη προ-εκπαιδευτεί από το σύνολο D_i [31].

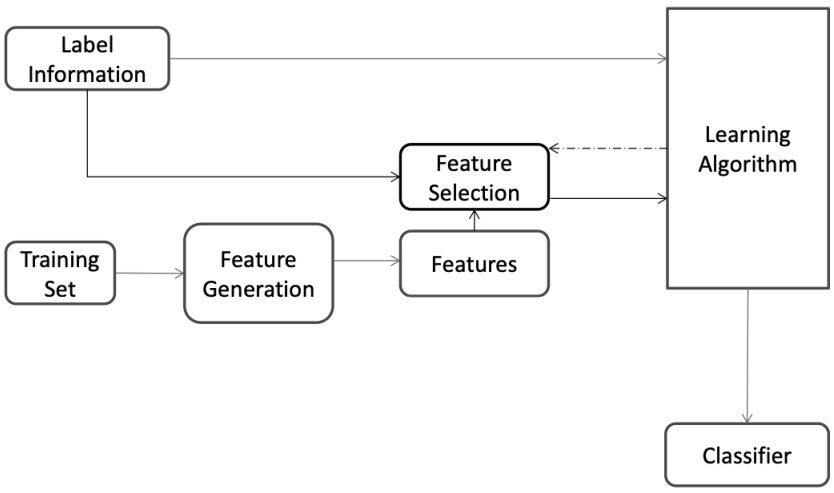
Η κατηγορία της επιβλεπόμενης επιλογής χαρακτηριστικών μπορεί να κα-

τανεμηθεί σε υποκατηγορίες μοντέλων οι οποίες είναι οι εξής: τα μοντέλα φίλτρα, τα μοντέλα περιτύλιξης και τα ενσωματωμένα μοντέλα. Τα μοντέλα φίλτρα ξεχωρίζουν τη διαδικασία επιλογής χαρακτηριστικών από την ίδια την διαδικασία ταξινόμησης έτσι ώστε η προκατάληψη του αλγόριθμου εκάθησης να μην επηρεάζει τον αλγόριθμο επιλογής των χαρακτηριστικών. Βασίζεται κυρίως σε μετρικές των γενικών χαρακτηριστικών όπως η απόσταση, η συνέπεια, η εξάρτηση, η πληροφόρηση και η συσχέτιση. Τέτοια μοντέλα είναι το Fisher Score και το Information Gain, τα οποία θα δούμε και στη συνέχεια. Τα μοντέλα περιτύλιξης χρησιμοποιούν την προγνωστική ακρίβεια ενός προκαθορισμένου αλγόριθμου μάθησης για να καθορίσει την ποιότητα των επιλεγμένων χαρακτηριστικών. Αυτά τα μοντέλα είναι υπολογιστικά ακριβά όταν αναλύουν ένα μεγάλο σύνολο χαρακτηριστικών. Εδώ, έρχεται το ενσωματωμένο μοντέλο το οποίο αποτελεί η γεφύρωση των δύο προηγούμενων μοντέλων καθώς περιλαμβάνει στατιστικές μετρικές και κριτήρια όπως το μοντέλο φίλτρου, επιλέγει τα υποψήφια χαρακτηριστικά βάσει του αλγόριθμου επιλογής χαρακτηριστικών και, στη συνέχεια, επιλέγει ένα υποσύνολο από αυτά με τα οποία επιτυγχάνει τη μέγιστη ακρίβεια ταξινόμησης. Σε αυτήν την περίπτωση μοντέλου, η διαδικασία επιλογής χαρακτηριστικών πραγματοποιείται κατά τη διάρκεια της εκμάθησης [32].

Ουσιαστικά, η επιλογή χαρακτηρισικών αποτελεί πρόβλημα βελτιστοποίησης του οποίου η βέλτιστη λύση μπορεί να βρεθεί μόνο μέσω μίας εξαντλητικής αναζήτησης βάσει κάποιου κριτηρίου αξιολόγησης ή αναζήτησης. Ωστόσο, πέραν από την επιλογή των σχετικών χαρακτηρισιτικών, αρκετά σημαντικό ρόλο παίζει και ο συνδυασμός των επιλεγμένων χαρακτηριστικών, καθώς για διάφορους συνδυασμούς, μπορούν να προκύψουν διαφορετικές ακρίβειες από κάθε μοντέλο [33]. Εξετάζοντας τα ενδεχόμενα, οι αλγόριθμοι επαγωγής μπορούν να χρησιμοποιηθούν προκειμένου να πραγματοποιήσουν κάποια ταξινόμηση με ή και χωρίς επιλογή των χαρακτηριστικών. Από το ένα άκρο, υπάρχουν διάφοροι αλγόριθμοι επαγωγής που χρησιμοποιούνται για απλή ταξινόμηση, ξεκινώντας από τον απλούστερο με τους K πλησιέστερους γείτονες. Ο αλγόριθμος αυτός μπορεί να επιλέγει να ταξινομεί το δείγμα δοκιμής παίρνοντας το κοντινότερο αποθηκευμένο παράδειγμα εκπαίδευσης, αξιοποιώντας όλα τα διαθέσιμα χαρακτηριστικά. Παρόλο που αυτή η μέθοδος έχει υψηλή ασυμπτωτική ακρίβεια, η παρουσία ενός μη σχετικού χαρακτηριστικού μπορεί να μειώσει σημαντικά τον χρόνο εκμάθησης. Από την άλλη, χρησιμοποιούνται επαγωγικές μέθοδοι οι οποίες έχουν αυτοσκοπό να επιλέξουν τα σχετικά χαρακτηριστικά και να απορρίψουν τα μη-σχετικά. Έχουν αναπτυχθεί αρκετές μέθοδοι επιλογής χαρακτηριστικών οι οποίες αποδεικνύεται ότι βελτιώνουν την επίδοση των επαγωγικών αλγόριθμων όπως εκείνον των K-πλησιέστερων γειτόνων. Αυτό συμβαίνει καθώς με την ελάττωση των χαρακτηριστικών που χρησιμοποιεί κάθε μοντέλο για την εκπαίδευση, ελαττώνεται και ο αριθμός των υποθέσεων που λαμβάνονται υπόψη ενώ, ταυτόχρονα, μειώνεται και το μέγεθος του δείγματος με αποτέλεσμα να πραγματοποιεί καλύτερη γενίκευση. Στο ενδιάμεσο υπάρχουν και οι μέθοδοι που αντιστιχίζουν σε κάθε χαρακτηριστικό ένα διαφορετικό βάρος, με αποτέλεσμα

αντί να επιλέγει ένα υποσύνολο των χαρακτηριστικών, στοχεύουν στην επίτευξη μίας καλής κλιμάκωσης, δηλαδή να δίνεται διαφορετικό βάρος σε κάθε χαρακτηριστικά ανάλογα με τη σημασία του καθενός.

Τυπικά, οι αλγόριθμοι επιλογής χαρακτηρισιτκών έχουν την εξής δομή: αρχικά, διαθέτουν κάποιον αλγόριθμο αναζήτησης, ο οποίος αναζητά τον χώρο των υποσυνόλων των χαρακτηριστικών. Στη συνέχεια, διαθέτουν τη συνάρτηση αξιολόγησης, στην οποία εκχωρείται ένα υποσύνολο χαρακτηριστικών και η έξοδο της αποτελεί κάποια αριθμητική αξιολόγηση. Είναι σημαντικό να διευκρινιστεί ότι ο στόχος του αλγόριθμου είναι η μεγιστοποίηση της τιμής της συνάρτησης αυτής. Τέλος, υπάρχει και η συνάρτηση επίδοσης, η οποία συνήθως είναι μία διαδικασία ταξινόμησης και βάσει την ακρίβεια που επιτυχγάνει, αξιολογεί και τη μέθοδο επιλογής χαρακτηριστικών [33]. Γενικότερα, τα βασικά βήματα που ακολουθεί ένας αλγόριθμος επιλογής χαρακτηριστικών είναι η επιλογή υποσυνόλου χαρακτηριστικών, στη συνέχεια ακολουθεί η αξιολόγηση του υποσυνόλου αυτού. Ύστερα, το επόμενο βήμα είναι ότι εφαρμόζει κάποιο κριτίριο διακοπής, και τέλος είναι η επικύρωση του αποτελέσματος [32]. Η διαδικασία επιλογής των χαρακτηριστικών επηρεάζει την ταξινόμηση αφού ο αλγόριθμος ταξινόμησης πλέον εκπαιδεύεται βάσει των επιλεγμένων χαρακτηριστικών αντί για ολόκληρο το σύνολο των χαρακτηριστικών που περιείχε το σύνολο δεδομένων.



Σχήμα 3.2: Αναπαράσταση διαδικασίας επιλογής χαρακτηριστικών [32]

Οι αλγόριθμοι αναζήτησης διακρίνονται σε τρεις κατηγορίες: τους εκθετικούς, τους τυχαιοποιημένους (randomized) και τους διαδοχικούς. Οι εκθετικοί έχουν εκθετική πολυπλοκότητα και είναι αρκετά επιβαρυντικοί προγραμματιστικά, με πολυπλοκότητα $\mathcal{O}(2^d)$ όπου το d είναι ο αριθμός των χαρακτηριστικών. Οι τυχαιοποιημένοι αλγόριθμοι συμπεριλαμβάνουν μεθόδους αναζήτησης γενετικής και προσομοιωμένης ανόπτησης και επιτυγχάνουν υψηλές ακρίβειες αλλά απαιτούν προκαταλήψεις (biases) για να αποδώσουν μικρά υποσύνολα. Τέλος, οι διαδοχικοί αλγόριθμοι αναζήτησης έχουν πολυωνυμική πολυπλοκότητα $\mathcal{O}(d^2)$, προσθέτουν ή αφαιρούν

χαρακτηριστικά και χρησιμοποιούν μια στρατηγική αναζήτησης με αναρρίχηση σε λόφους (hill climbing search strategy). Οι πιο συνήθεις αλγόριθμοι διαδοχικής αναζήτησης είναι η προς τα εμπρός διαδοχική επιλογή (Forward Sequential Selection, FSS) και η προς τα πίσω διαδοχική επιλογή (Backward Sequential Selection, BSS). Ο τρόπος με τον οποίο τα μοντέλα αυτά αξιολογούν κάποιο χαρακτηριστικό είναι με τη βοήθεια κάποιας ευρετικής συνάρτησης.

Μία άλλη μέθοδος διαδοχικής αναζήτησης που είναι αρκετά διαδεδομένη για μικρό όγκο δεδομένων εκπαίδευσης αλλά για ένα σύνολο δεδομένων που περιέχει πολλά χαρακτηριστικά και συμβάλλει σημαντικά στην μείωση του κινδύνου της υπερπροσαρμογής του μοντέλου μηχανικής μάθησης είναι η αναδρομική απαλοιφή χαρακτηριστικών (Recursive Feature Elimination (RFE). Η μέθοδος RFE αξιοποιεί τη δυνατότητα γενίκευσης που ενσωματώνεται στις μηχανές διανυσμάτων υποστήριξης και είναι, επομένως, κατάλληλη για προβλήματα μικρών δειγμάτων. Παρά τις καλές επιδόσεις της, η RFE τείνει να απορρίπτει ¨αδύναμα¨ χαρακτηριστικά, τα οποία μπορεί να παρέχουν σημαντική βελτίωση της απόδοσης όταν συνδυάζονται με άλλα χαρακτηριστικά [34]. Το παραδοσιακό RFE αφαιρεί διαδοχικά το χειρότερο χαρακτηριστικό που προκαλεί πτώση της ¨ακρίβειας ταξινόμησης¨ μετά τη δημιουργία ενός μοντέλου ταξινόμησης [35].

Στη συνέχεια, μία άλλη χρήσιμη μορφή επιλογής χαρακτηριστικών είναι η μείωση διαστάσεων. Η μείωση διαστάσεων μπορεί να κατηγοριοποιηθεί σε εκχείλιση χαρακτηριστικών και σε επιλογή χαρακτηριστικών. Η εκχείλιση χαρακτηριστικών προσεγγίζει την προβολή των χαρακτηριστικών σε ένα νέο χώρο χαρακτηριστικών με χαμηλότερη διάσταση κατασκευάζοντας έτσι νέα χαρακτηριστικά που συνήθως είναι συνδυασμοί των αρχικών χαρακτηριστικών. Παραδείγματα τεχνικών εξαγωγής χαρακτηριστικών περιλαμβάνουν την ανάλυση κύριων συνιστωσών (PCA), τη γραμμική διακριτική ανάλυση (LDA) και την ανάλυση κανονικής συσχέτισης (CCA), ενώ επιπλέον μέθοδοι επιλογλής χαρακτηριστικών αποτελούν και το κέρδος πληροφορίας (IG), το Fisher Score και το Lasso. Η εκχείλιση αντιστοιχεί τον αρχικό χώρο χαρακτηριστικών σε ένα νέο χώρο χαρακτηριστικών με χαμηλότερες διαστάσεις [32]. Η αντίστοιχιση αυτή, ωστόσο, γίνεται αρκετά κοστοβόρα και για αυτό προτιμούνται οι μέθοδοι επιλογής χαρακτηριστικών, αφού επιλέγουν ένα υποσύνολο των αρχικών χαρακτηριστικών χωρίς να προϋποθέτουν κάποιον μετασχηματισμό στην ήδη υπάρχουσα πληροφορία και ταυτόχρονα διατηρεί τη φυσική σημασία των αρχικών χαρακτηριστικών αυτών.

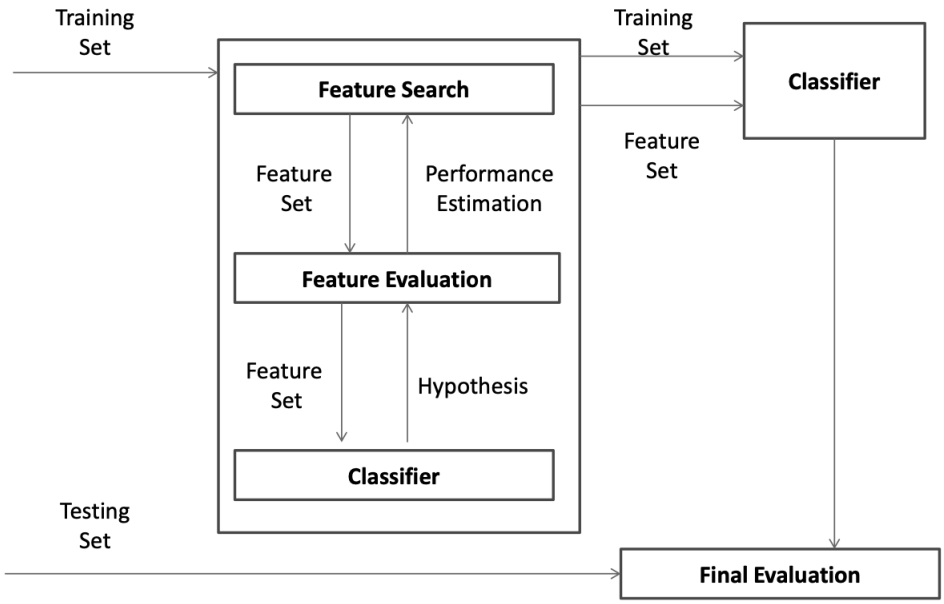
Από την κατηγορία των μοντέλων φίλτρα, το Fisher Score αξιολογεί κάθε χαρακτηριστικό ανεξάρτητα από τα υπόλοιπα αναθέτοντας του παρόμοιες τιμές σε περιπτώσεις της ίδιας κλάσης και διαφορετικές τιμές σε περιπτώσεις από διαφορετικές κλάσεις. Η τιμή αυτή καθορίζεται από τον τύπο $S_i = \frac{\sum_{k=1}^{K} n_k(\mu_{ki}-\mu_i)^2}{\sum_{k=1}^{K} n_k \sigma_{ki}^2}$, όπου το μ_{ij} και σ_{ij} είναι η μέση τιμή και η απόκλιση του i-στού στοιχείου από το j-στού στοιχείου και το n_j είναι το πλήθος των στοιχείων της κλάσης j και το μ_i είναι η μέση τιμή του χαρακτηριστικού i. Το γεγονός ότι η συγκεκριμένη μέθοδος αξιολογεί κάθε

χαρακτηριστικό ξεχωριστά, καθιστά αυτή τη μέθοδο ακατάλληλη για την αντιμετώπιση περιττών χαρακτηριστικών. Μία εναλλαγή της μεθόδου αυτής προκειμένου να μπορεί να επιλέγει χαρακτηριστικά είναι μέσω του καθορισμού ενός κατωφλίου της χαμηλότερης αποδεκτής τιμής fisher ώστε να επιλέγεται υποσύνολο από τα χαρακτηριστικά τα οποία οι τιμές fisher είναι μεγαλύτερες ή ίσες με την τιμή κατωφλίου αυτή.

Το κέρδος πληροφορίας είναι μία πολύ δημοφιλής μέθοδος επιλογής χαρακτηριστικών και όπως και στα δέντρα αποφάσεων, χρησιμοποιείται για να ανακαλύψει την εξάρτηση μεταξύ των χαρακτηριστικών και των ετικετών. Προφανώς, ένα χαρακτηριστικό όσο υψηλότερη τιμή κέρδους πληροφορίας έχει, τόσο πιο σχετικό είναι το χαρακτηριστικό. Παρόμοια όπως με το Fisher Score, έτσι και το IG, αποτυγχάνει να διαχειρίζεται τα περιττά χαρακτηριστικά αφού απλά αναθέτει σε κάθε χαρακτηριστικό κάποια τιμή. Ωστόσο, πάλι, εφαρμόζοντας κάποιο κατώφλι ως κάτω όριο, μπορεί να τροποποιηθεί ο αλγόριθμος με τέτοιον τρόπο ώστε να διαμορφώνει ένα υποσύνολο χαρακτηριστικών, των οποίων οι τιμές ξεπερνούν ή ισούνται με το κατώφλι αυτό.

Τα μοντέλα περιτύλιξης, όπως αναφέρθηκαν και προηγουμένως λειτουργούν ανεξάρτητα από την υπόλοιπη διαδικασία ταξινόμησης, ωστόσο, ένα μειονέκτημα τους είναι ότι αγνοούν την επίδοση του επιλεγμένου υποσυνόλου χαρακτηριστικών κατά την διάρκεια της ίδιας της ταξινόμησης. Ο βέλτιστός συνδυασμός χαρακτηριστικών που προκύπτει από αυτά βασίζεται συνήθως από συγκεκριμένες προκαταλήψεις και ευρετικές συναρτήσεις του αλγόριθμου τους. Βάσει αυτής της υπόθεσης, αυτά τα μοντέλα αξιοπούν συγκεκριμένους ταξινομητές προκειμένου να αξιολογήσουν την ποιότητα των επιλεγμένων χαρακτηριστικών και δίνουν έναν απλό αλλά δυνατό τρόπο αντιμετώπισης του προβλήματος κατάλληλης επιλογής χαρακτηριστικών ανεξάρτητα της επιλεγμένης μεθόδου εκμάθησης. Δεδομένου οτι διαθέτει και έναν προκαθορισμένο ταξινομητή, ένα τυπικό μοντέλο περιτύλιξης ακολουθεί τα εξής βήματα: αρχικά αναζητά το υποσύνολο των χαρακτηριστικών, επαναλαμβάνεται στη συνέχεια αυτή η διαδικασία όσο αξιολογεί το επιλεγμένο υποσύνολο μέχρι να επιτευχθεί η επιθυμητή ποιότητα. Ο προεπιλεγμένος ταξινομητής λειτουργεί σαν ένα μαύρο κουτί ο οποίος κάνει μία εκτίμηση της επίδοσης του αλγόριθμου και την επιστρέφει πίσω στη διαδικασία αναζήτησης χαρακτηριστικών μέχρις ότου να επιλεγεί το τελικό σύνολο χαρακτηριστικών που έχει την υψηλότερη εκτιμώμενη επίδοση, το οποίο στη συνέχεια τροφοδοτείται για την κανονική εκπαίδευση στον αλγόριθμο εκμάθησης [31], [32].

Τα ενσωματωμένα μοντέλα αξιοποιούν και αυτά προκαθορισμένους ταξινομητές προκειμένου να αξιολογήσουν την ποιότητα των επιλεγμένων χαρακτηριστικών και οι προκαταλήψεις του ταξινομητή αποφεύγονται κατά την επιλογή των χαρακτηριστικών. Υπάρχουν τρία είδη τέτοιων μοντελών, η πρώτη κατηγορία είναι οι μέθοδοι κλαδέματος κατά τις οποίες χρησιμοποιούν όλα τα χαρακτηριστικά για την εκπαίδευση του μοντέλου και μετά προσπαθούν να απορρίψουν μερικά από τα λιγότερο



Σχήμα 3.3: Αναπαράσταση διαδικασίας επιλογής χαρακτηριστικών με τα μοντέλα περιτύλιξης [32].

σχετικά χαρακτηριστικά ενώ ταυτόχρονα επιθυμούν να διατηρήσουν την επίδοση του αρχικού μοντέλου. Σε αυτήν την κατηγορία ανήκει και η μέθοδος RFE. Η δεύτερη κατηγορία είναι εκείνες οι μέθοδοι που έχουν ενσωματωμένο έναν μηχανισμό για την επιλογή των χαρακτηριστικών όπως το ID3. Τέλος, το τρίτο είδος είναι τα μοντέλα κανονικοποίησης που διαθέτουν αντικειμενικές συναρτήσεις που ελαχιστοποιούν τα σφάλματα προσαρμογής και αναγκάζουν τους συντελεστές να είναι μικροί ή να είναι ακριβώς μηδενικοί. Τα χαρακτηριστικά που έχουν συντελεστές κοντά στην τιμή 0 στη συνέχεια εξαλείφονται. Ένα τέτοιο μοντέλο είναι το μοντέλο LASSO Regularization το οποίο πέραν ότι χρησιμοποιείται σε προβλήματα ταξινόμησης, μπορεί να αξιοποιηθεί και στη διαδικασία επιλογής χαρακτηριστικών. Η λογική από πίσω είναι ίδια με εκείνη που χρησιμοποιείται κατά την ταξινόμηση και εκμεταλλεύεται τον τρόπο που λειτουργεί με τα πέναλτυ. Με άλλα λόγια, η επιλογή χαρακτηριστικών επιτυχγάνεται εκτιμώντας τους γραμμικούς ταξινομητές w με κατάλληλα ρυθμισμένες ποινές. Ο ταξινομητής w περιέχει ένα διάνυσμα από συντελεστές όπου κάθε συντελεστής αντιστοιχεί σε κάποιο χαρακτηριστικό και μπορούν μερικοί από αυτούς να διαθέτουν μηδενική τιμή. Οι συντελεστές με μηδενική τιμή, λοιπόν, αντιστοιχούν στα χαρακτηριστικά που μπορούν να παραλειφθούν [31], [36].

Αναλυτικότερα, οι μέθοδοι που χρησιμοποιήθηκαν για τη διαδικασία επιλογής χαρακτηριστικών εξετάζονται παρακάτω:

Η Επιλογή Χαρακτηριστικών προς τα Εμπρός (Forward Feature Selection) είναι μια σταδιακή προσέγγιση στην επιλογή χαρακτηριστικών που ξεκινά με ένα κενό σύνολο χαρακτηριστικών και προσθέτει διαδοχικά ένα χαρακτηριστικό κάθε φορά. Σε κάθε βήμα, το χαρακτηριστικό που βελτιώνει περισσότερο την απόδοση

του μοντέλου, σύμφωνα με ένα καθορισμένο κριτήριο, προστίθεται στο μοντέλο. Η διαδικασία αυτή συνεχίζεται έως ότου η προσθήκη περισσότερων χαρακτηριστικών δεν βελτιώνει σημαντικά την απόδοση του μοντέλου ή έως ότου επιτευχθεί ο επιθυμητός αριθμός χαρακτηριστικών.

Σε αυτή την υλοποίηση, η διαδικασία αρχίζει με τη φόρτωση και την προετοιμασία του συνόλου δεδομένων. Η μεταβλητή-στόχος, Death, διαχωρίζεται από τα χαρακτηριστικά και καθορίζονται τυχόν πρόσθετες στήλες που πρέπει να εξαιρεθούν. Ένας `RandomForestClassifier` αρχικοποιείται ως μοντέλο για την αξιολόγηση της σημασίας κάθε χαρακτηριστικού. Ο `SequentialFeatureSelector` από το `sklearn.feature_selection` χρησιμοποιείται για την εκτέλεση της FFS. Αυτός ο επιλογέας αρχικοποιείται με τον `RandomForestClassifier` και έχει ρυθμιστεί ώστε να επιλέγει χαρακτηριστικά με κατεύθυνση προς τα εμπρός, βελτιστοποιώντας την ακρίβεια. Ο επιλογέας προσαρμόζει το μοντέλο στο σύνολο δεδομένων, προσθέτοντας διαδοχικά τα χαρακτηριστικά που συμβάλλουν περισσότερο στη βελτίωση της ακρίβειας του μοντέλου.

Η Βαθμολογία Fisher (Fisher Score) είναι μια μέθοδος επιλογής χαρακτηριστικών που αξιολογεί τη διακριτική ικανότητα μεμονωμένων χαρακτηριστικών. Υπολογίζει τον λόγο της διακύμανσης μεταξύ των κλάσεων προς τη διακύμανση εντός των κλάσεων για κάθε χαρακτηριστικό, προσδιορίζοντας τα χαρακτηριστικά που μπορούν να διαχωρίσουν καλύτερα τις διαφορετικές κλάσεις.

Η εφαρμογή ξεκινά με τη φόρτωση και την προετοιμασία του συνόλου δεδομένων. Η μεταβλητή-στόχος, Death, διαχωρίζεται από τα χαρακτηριστικά. Στη συνέχεια χρησιμοποιείται η συνάρτηση fisher_score για τον υπολογισμό του Fisher Score για κάθε χαρακτηριστικό. Η συνάρτηση αυτή υπολογίζει τον μέσο όρο κάθε κλάσης και τον συνολικό μέσο όρο για ένα δεδομένο χαρακτηριστικό. Στη συνέχεια υπολογίζει τη διακύμανση μεταξύ των κλάσεων, η οποία μετρά τη μεταβλητότητα των μέσων των κλάσεων από το συνολικό μέσο, και τη διακύμανση εντός των κλάσεων, η οποία μετρά τη μεταβλητότητα των τιμών των χαρακτηριστικών εντός κάθε κλάσης.

Η βαθμολογία Fisher για κάθε χαρακτηριστικό υπολογίζεται ως ο λόγος της διακύμανσης μεταξύ των κλάσεων προς τη διακύμανση εντός των κλάσεων. Εάν η διακύμανση εντός της κλάσης είναι μηδέν, η βαθμολογία Fisher τίθεται στο μηδέν για την αποφυγή σφαλμάτων διαίρεσης με το μηδέν. Οι βαθμολογίες αποθηκεύονται σε ένα λεξικό με τα ονόματα των χαρακτηριστικών ως κλειδιά.

Το Πληροφοριακό Κέρδος (Information Gain - IG) είναι μια μέθοδος επιλογής χαρακτηριστικών που μετρά την ποσότητα πληροφορίας που παρέχει ένα χαρακτηριστικό σχετικά με την ετικέτα της τάξης. Αξιολογεί πόσο καλά κάθε χαρακτηριστικό διαχωρίζει τις διαφορετικές τάξεις, με μεγαλύτερες τιμές να υποδεικνύουν μεγαλύτερη σχετικότητα.

Η υλοποίηση του Κέρδους Πληροφορίας για την επιλογή χαρακτηριστικών ξεκινά με τη φόρτωση και προετοιμασία του συνόλου δεδομένων, διαχωρίζοντας τη μεταβλητή στόχο, Death, από τα χαρακτηριστικά. Η αμοιβαία πληροφορία μεταξύ

κάθε χαρακτηριστικού και της μεταβλητής στόχου υπολογίζεται χρησιμοποιώντας τη συνάρτηση `mutual_info_classif` από το `sklearn.feature_selection`. Αυτή η συνάρτηση υπολογίζει την εξάρτηση μεταξύ του χαρακτηριστικού και της μεταβλητής στόχου, παράγοντας μια βαθμολογία σημασίας για κάθε χαρακτηριστικό.

Οι υπολογισμένες σημασίες χαρακτηριστικών αποθηκεύονται σε μια σειρά pandas, με τα ονόματα των χαρακτηριστικών ως δείκτες. Αυτές οι σημασίες ταξινομούνται κατά φθίνουσα σειρά για να εντοπιστούν τα πιο ενημερωτικά χαρακτηριστικά. Δημιουργείται ένα ραβδόγραμμα των σημασιών των χαρακτηριστικών για να απεικονιστεί οπτικά η σχετικότητα κάθε χαρακτηριστικού. Για την επιλογή των κορυφαίων χαρακτηριστικών, χρησιμοποιείται η μέθοδος SelectKBest, καθορίζοντας το `mutual_info_classif` ως τη συνάρτηση βαθμολόγησης και το k ως τον αριθμό των κορυφαίων χαρακτηριστικών που θα επιλεγούν.

Η Κανονικοποίηση LASSO (Least Absolute Shrinkage and Selection Operator) είναι μια μέθοδος επιλογής χαρακτηριστικών που εφαρμόζει κανονικοποίηση $L1$ για να περιορίσει τους συντελεστές των λιγότερο σημαντικών χαρακτηριστικών στο μηδέν. Με αυτόν τον τρόπο, επιλέγονται μόνο τα πιο σημαντικά χαρακτηριστικά που συμβάλλουν στη μοντελοποίηση.

Η υλοποίηση της κανονικοποίησης LASSO ξεκινά με τη φόρτωση και προετοιμασία του συνόλου δεδομένων, διαχωρίζοντας τη μεταβλητή στόχο Death από τα χαρακτηριστικά. Η κανονικοποίηση LASSO χρησιμοποιεί το μοντέλο Lasso από το `sklearn.linear_model`, όπου το παράμετρο alpha καθορίζει τον βαθμό της κανονικοποίησης.

Το μοντέλο LASSO εφαρμόζεται στα δεδομένα, προσαρμόζοντας τους συντελεστές των χαρακτηριστικών. Μόνο τα χαρακτηριστικά με μη μηδενικούς συντελεστές επιλέγονται, υποδεικνύοντας τη σημαντικότητά τους.

Η Αναδρομική Εξάλειψη Χαρακτηριστικών (Recursive Feature Elimination - RFE) είναι μια μέθοδος επιλογής χαρακτηριστικών που χρησιμοποιεί έναν εκτιμητή για να επιλέξει επαναληπτικά τις πιο σημαντικές δυνατότητες μέχρι να φτάσει στον επιθυμητό αριθμό χαρακτηριστικών.

Η υλοποίηση της RFE ξεκινά με τη φόρτωση και προετοιμασία του συνόλου δεδομένων, διαχωρίζοντας τη μεταβλητή στόχο, Death, από τα χαρακτηριστικά. Χρησιμοποιείται ένας εκτιμητής SVR (Support Vector Regressor) με γραμμικό πυρήνα ως το βασικό μοντέλο για την αξιολόγηση της σημασίας των χαρακτηριστικών. Η RFE αρχικοποιείται με τον εκτιμητή SVR και τον επιθυμητό αριθμό χαρακτηριστικών προς επιλογή. Η RFE προσαρμόζεται στα δεδομένα και επαναληπτικά εξαλείφει τα λιγότερο σημαντικά χαρακτηριστικά, διατηρώντας μόνο τα πιο σημαντικά.

Η Αναδρομική Εξάλειψη Χαρακτηριστικών (Backward Elimination) είναι μια μέθοδος επιλογής χαρακτηριστικών που ξεκινά με όλα τα χαρακτηριστικά και σταδιακά αφαιρεί τα λιγότερο σημαντικά χαρακτηριστικά μέχρι να βρεθεί το βέλτιστο υποσύνολο.

Η υλοποίηση της Αναδρομικής Εξάλειψης ξεκινά πάλι με τη φόρτωση και

προετοιμασία του συνόλου δεδομένων, διαχωρίζοντας τη μεταβλητή στόχο, Death, από τα χαρακτηριστικά. Χρησιμοποιείται ένας Random Forest Classifier ως το βασικό μοντέλο για την αξιολόγηση της σημασίας των χαρακτηριστικών. Η διαδικασία ξεκινά με την αρχικοποίηση ενός SequentialFeatureSelector (SFS) από το `mlxtend.feature_selection`, με τον Random Forest Classifier ως τον εκτιμητή. Η SFS ρυθμίζεται σε αναδρομική (backward) επιλογή χαρακτηριστικών και χρησιμοποιεί διασταυρούμενη επικύρωση (cross-validation) για να αξιολογήσει την απόδοση του μοντέλου. Η SFS εφαρμόζεται στα δεδομένα, αφαιρώντας επαναληπτικά τα λιγότερο σημαντικά χαρακτηριστικά και διατηρώντας μόνο τα πιο σημαντικά.

Ο Συντελεστής Διόγκωσης Διακύμανσης (Variance Inflation Factor - VIF) είναι μια μέθοδος που χρησιμοποιείται για τον εντοπισμό και την εξάλειψη πολυδιγραμμικότητας στα χαρακτηριστικά ενός συνόλου δεδομένων. Ο VIF μετρά πόσο η διακύμανση ενός εκτιμώμενου συντελεστή παλινδρόμησης αυξάνεται λόγω της γραμμικής εξάρτησης με άλλα χαρακτηριστικά.

Η υλοποίηση της επιλογής χαρακτηριστικών με VIF ξεκινά με τη φόρτωση και προετοιμασία του συνόλου δεδομένων, διαχωρίζοντας τη μεταβλητή στόχο Death από τα χαρακτηριστικά. Επιλέγονται μόνο οι αριθμητικές στήλες του συνόλου δεδομένων για την ανάλυση VIF, ώστε να αποφεύγονται προβλήματα με μη αριθμητικά δεδομένα. Οι τιμές VIF υπολογίζονται για κάθε χαρακτηριστικό του συνόλου δεδομένων χρησιμοποιώντας τη συνάρτηση `variance_inflation_factor` από το `statsmodels.stats.outliers_influence`.

Για όλες τις παραπάνω μεθόδους, δημιουργείται ένα νέο DataFrame που περιέχει μόνο τα επιλεγμένα χαρακτηριστικά. Τα ονόματα των επιλεγμένων χαρακτηριστικών εκτυπώνονται για επαλήθευση και αποθηκεύονται χρησιμοποιώντας τη μονάδα pickle για μελλοντική αναφορά. Στη συνέχεια, τα επιλεγμένα χαρακτηριστικά συνδυάζονται με τη μεταβλητή στόχο για να σχηματιστεί ένα νέο σύνολο δεδομένων, το οποίο αποθηκεύεται για περαιτέρω ανάλυση ή μοντελοποίηση. Αυτή η διαδικασία διασφαλίζει ότι περιλαμβάνονται τα πιο σημαντικά χαρακτηριστικά, βελτιώνοντας την απόδοση και την ερμηνευσιμότητα του μοντέλου.

Μία ακόμα μέθοδος επιλογής χαρακτηριστικών πραγματοποιήθηκε κατά την προεπεξεργασία των δεδομένων, και ήταν η επιλογή των χαρακτηριστικών βάσει της συσχέτισης τους με την στήλη ¨Θάνατος¨. Με άλλα λόγια, η μέθοδος επιλογής χαρακτηριστικών που χρησιμοποιήθηκε εδώ επικεντρώνεται στην αναγνώριση χαρακτηριστικών που εμφανίζουν τη μεγαλύτερη συσχέτιση με τη στήλη "Death" εντός του συνόλου δεδομένων. Η ανάλυση συσχέτισης είναι μια βασική στατιστική τεχνική που χρησιμοποιείται για να ποσοτικοποιήσει τη σχέση μεταξύ μεταβλητών. Σε αυτήν την περίπτωση, υπολογίζεται ο συντελεστής συσχέτισης Pearson ανάμεσα σε κάθε χαρακτηριστικό και τη μεταβλητή στόχο "Death".

Η πρώτη ενέργεια είναι η υπολογισμός του πίνακα συσχέτισης για όλα τα χαρακτηριστικά στο σύνολο δεδομένων. Στη συνέχεια, εξάγονται και ταξινομούνται με φθίνουσα σειρά οι συντελεστές συσχέτισης ως προς τη στήλη "Death". Αυτή η ταξι-

νόμηση διατάσσει τα χαρακτηριστικά βάσει της έντασης της γραμμικής τους σχέσης με τη μεταβλητή στόχο, όπου υψηλότερες απόλυτες τιμές υποδεικνύουν ισχυρότερες συσχετίσεις.

Για τη στενότερη επιλογή των επιρροητικότερων χαρακτηριστικών, επιλέγονται τα δέκα κορυφαία χαρακτηριστικά με τις υψηλότερες τιμές συσχέτισης, συμπεριλαμβανομένης της ίδιας της μεταβλητής "Death". Αυτά τα χαρακτηριστικά θεωρούνται πιο πιθανό να έχουν προγνωστική ισχύ σχετικά με τη μεταβλητή αποτελέσματος "Death".

Τέλος, δημιουργείται ένα νέο σύνολο για κάθε μέθοδο δεδομένων αποκλειστικά με αυτά τα κορυφαία χαρακτηριστικά από το αρχικό σύνολο δεδομένων. Αυτό το υποσύνολο αναλύεται περαιτέρω ή χρησιμοποιείται απευθείας σε εργασίες μοντελοποίησης, με στόχο τη βελτίωση της απόδοσης του μοντέλου μέσω της εστίασης στους πιο σημαντικούς προβλέποντες.

3.3 Γραμμική Παλινδρόμηση (Linear Regression)

Ένα από τα πιο διαδεδομένα μοντέλα επιβλεπόμενης μάθησης είναι η γραμμική παλινδρόμηση. Πιο συγκεκριμένα πρόκειται για ένα από τα πιο δημοφιλή μαθηματικά μοντέλα ποσοτικής ανάλυσης που χρησιμοποιείται ευρέως σε πολλούς τομείς για την επίλυση διαφόρων προβλημάτων. Το πρόβλημα της παλινδρόμησης, ουσιαστικά, προσπαθεί να εντοπίσει κάποια συνάρτηση f που παράγει την τιμή στόχο t_i για κάθε πρότυπο εισόδου

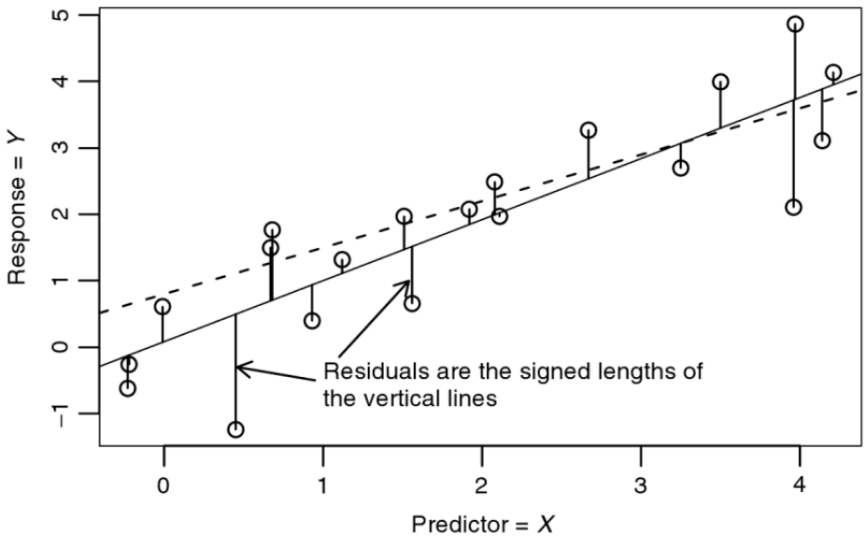
$$x_i$$

. Στα προβλήματα παλινδρόμησης οι επιθυμητές τιμές είναι πραγματικοί αριθμοί, ενώ είναι ανώφελο να αναφερόμαστε στην ακρίβεια εφόσον είναι αδύνατο να ισχύει το $y = t$ ακριβώς αλλά ένα μοντέλο παλινδρόμησης επιτυγχάνει να προσεγγίζει τη σχέση αυτήν. Για τον λόγο αυτόν, αξιοποιούνται κριτήρια μέτρησης της απόστασης μεταξύ της εξόδου του μοντέλου και της επιθυμητής τιμής αυτής [37]. Μερικά κριτήρια μέτρησης είναι το μέσο τετραγωνικό σφάλμα, η μέθοδος των ελαχίστων τετραγώνων, το μέσο απόλυτο σφάλμα, την ομοιότητα συνημητόνου και την ομοιότητα Pearson. Το κριτήριο μέσο τετραγωνικό σφάλμα μετράει τη μέση ευκλείδεια απόσταση μεταξύ του διανύσματος εξόδου y_p του μοντέλου και του διανύσματος στόχου t_p, με τη βέλτιστη τιμή να είναι το μηδέν και περιγράφεται με την ακόλουθη σχέση:

$$J_{MSE} = \sum_{p=1}^{N} \parallel t_p - y_p \parallel^2 = \sum_{p=1}^{N} \sum_{i=1}^{m} (t_{p,i} - y_{p,i})^2$$

Η μέθοδος των ελαχίστων τετραγώνων (LSM) χρησιμοποιείται για να βρεθεί η καλύτερη προσέγγιση μιας καμπύλης σε ένα σύνολο σημείων δεδομένων με τη μείωση του συνολικού τετραγωνικού σφάλματος των σημείων της καμπύλης. Η μέθοδος αυτή στη γραμμική παλινδρόμηση χρησιμοποιείται για να βρει τις προβλέψεις b0 και b1 έτσι ώστε η συσσωρευτική τετραγωνική απόσταση από την πραγματική απόκριση yi

να πλησιάζει το ελάχιστο δυνατό με τους συντελεστές παλινδρόμησης b0 και b1: $(b_0, b_1) = \underset{b_0,b_1}{\operatorname{argmin}} \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$



Σχήμα 3.4: Παράδειγμα της μεθόδου των ελαχίστων τετραγώνων [38]

Το κίνητρο πίσω από την προσέγγιση των ελαχίστων τετραγώνων είναι να βρεθούν οι εκτιμήσεις των παραμέτρων χρησιμοποιώντας την καλύτερη εφαρμογή της γραμμής στα σημεία δεδομένων (x_i, y_i) [38].

Με παρόμοια λογική, το μέσο απόλυτο σφάλμα υπολογίζει την απόλυτη απόσταση του διανύσματος εξόδου y_p από την επιθυμητό διάνυσμα εξόδου t_p, όπως φαίνεται και στην ακόλουθη σχέση:

$$\mathrm{J}_{MSE} = \sum_{p=1}^{N} \mid \mathrm{t}_p - \mathrm{y}_p \mid$$

Συνεχίζοντας με την ομοιότητα συνημιτόνου, το κριτήριο αυτό υφίσταται σε ζευγάρια διανυσμάτων t, y και υπολογίζει την ομοιότητα μεταξύ τους βάσει του κανονικοποιημένου εσωτερικού τους γινομένου:

$$\mathrm{J}_{cos} = \frac{\mathbf{t}^T \cdot \mathbf{y}}{\|\mathbf{t}\|\|\mathbf{y}\|} = \frac{\sum_{i=1}^{m} t_i y_i}{\sqrt{\sum_{i=1}^{m} t_i^2}\sqrt{\sum_{i=1}^{m} y_i^2}}$$

Το κριτήριο αυτό συνήθως χρησιμοποιείται σε προβλήματα εξόρυξης δεδομένων, σε συστήματα συστάσεων και σε προβλήματα συσταδοποίησης κειμένων. Τέλος, όσον αφορά την ομοιότητα Pearson, όπως και στην ομοιότητα συνημιτόνου, έτσι και εδώ αξιοποιείται το ζευγάρι διανυσμάτων t, y, και συγκεκριμένα, πρόκειται για κριτήριο συσχέτισης των διανυσμάτων αυτών. Ορίζεται όπως παρουσιάζεται ακολούθως:

$$\mathrm{J}_P = \frac{\sum (t_i - \bar{t}_i)(y_i - \bar{y}_i)}{\sqrt{\sum (t_i - \bar{t}_i)^2 \sum (y_i - \bar{y}_i)^2}}.$$

όπου οι $\bar{t}$, $\bar{y}$ είναι οι μέσες τιμές των t και y.

Για την υλοποίηση του μοντέλου πρόβλεψης χρησιμοποιώντας τον αλγόριθμο λογιστικής παλινδρόμησης (Linear Regression), ακολουθήθηκε μια σειρά σημαντικών βημάτων για να διασφαλιστεί η ακρίβεια και η αξιοπιστία των προβλέψεων. Αρχικά, το σύνολο δεδομένων φορτώθηκε και προετοιμάστηκε χρησιμοποιώντας τη βιβλιοθήκη pandas. Τα δεδομένα περιλάμβαναν διάφορα χαρακτηριστικά και τη μεταβλητή στόχο Death, η οποία αποτέλεσε το κύριο αντικείμενο πρόβλεψης.

Ακολούθησαν διάφορες μέθοδοι επιλογής χαρακτηριστικών, όπως η συσχέτιση, τα δέντρα αποφάσεων, η επιλογή χαρακτηριστικών προς τα εμπρός (FFS), οι βαθμολογίες Fisher, το πληροφοριακό κέρδος (IG), η κανονικοποίηση LASSO, τα τυχαία δάση, η αναδρομική εξάλειψη χαρακτηριστικών (RFE), η αλληλουχική επιλογή χαρακτηριστικών (SFS) και ο συντελεστής διόγκωσης διακύμανσης (VIF). Για κάθε μέθοδο δημιουργήθηκε και αποθηκεύτηκε ένα σύνολο δεδομένων με τα επιλεγμένα χαρακτηριστικά.

Για την εκπαίδευση του μοντέλου λογιστικής παλινδρόμησης, χρησιμοποιήθηκε η διασταυρούμενη επικύρωση `Stratified K-Folds` με 2 διαιρέσεις. Μετά την επιλογή των καλύτερων υπερπαραμέτρων, το μοντέλο λογιστικής παλινδρόμησης αξιολογήθηκε μέσω της διαδικασίας διασταυρούμενης επικύρωσης, όπου υπολογίστηκαν διάφορες μετρικές απόδοσης όπως η ακρίβεια του μοντέλου, η αναφορά ταξινόμησης και ο πίνακας σύγχυσης.

Η ακρίβεια του μοντέλου και η λεπτομερής αναφορά ταξινόμησης που περιλάμβανε την ακρίβεια, την ανάκληση και το F1-score για κάθε κατηγορία αποθηκεύτηκαν για κάθε μέθοδο επιλογής χαρακτηριστικών. Ο πίνακας σύγχυσης δημιουργήθηκε και αποθηκεύτηκε ως αρχείο CSV για περαιτέρω ανάλυση.

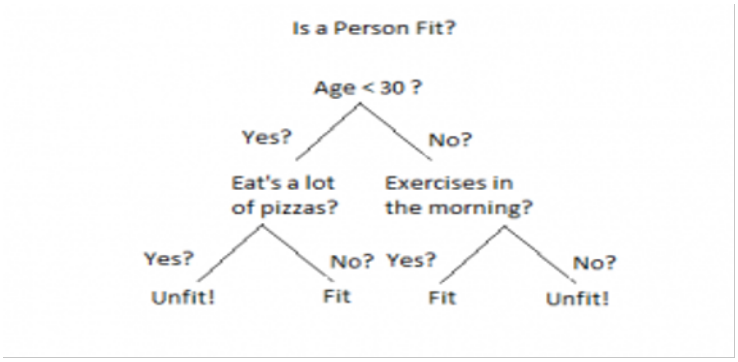
Επιπλέον, σχεδιάστηκε και αποθηκεύτηκε η καμπύλη ROC για να δείξει τη σχέση μεταξύ του ποσοστού αληθινά θετικών και του ποσοστού ψευδώς θετικών σε διάφορες ρυθμίσεις κατωφλίου, καθώς και η περιοχή κάτω από την καμπύλη (AUC). Για την εκτίμηση της ακρίβειας των πιθανοτήτων, υπολογίστηκε και αποθηκεύτηκε η απώλεια λογαρίθμου (log loss) για κάθε διαίρεση κατά τη διασταυρούμενη επικύρωση.

Τα αποτελέσματα από τις αξιολογήσεις, συμπεριλαμβανομένων των βαθμολογιών ακρίβειας, των αναφορών ταξινόμησης, των πινάκων σύγχυσης, των καμπυλών ROC και των διαγραμμάτων απώλειας λογαρίθμου, αποθηκεύτηκαν για κάθε μέθοδο επιλογής χαρακτηριστικών. Αυτά τα αποτελέσματα αποθηκεύτηκαν σε μορφές JSON και CSV για εύκολη πρόσβαση και ανάλυση.

Η διαδικασία αυτή διασφάλισε ότι το μοντέλο λογιστικής παλινδρόμησης ήταν βελτιστοποιημένο και αξιόπιστο, παρέχοντας ακριβείς προβλέψεις για το σύνολο δεδομένων. Οι λεπτομερείς μετρήσεις αξιολόγησης και οι οπτικοποιήσεις παρείχαν πολύτιμες πληροφορίες για τις επιδόσεις του μοντέλου.

3.4 Δέντρα Αποφάσεων (Decision Trees)

Τα δέντρα αποφάσεων είναι μια μορφή επιβλεπόμενης μάθησης κατά την οποία το σύνολο δεδομένων χωρίζεται σε υποσύνολα βάσει κάποιας παραμέτρου ενώ ταυτόχρονα αναπτύσσεται ένα σχετικό δέντρο αποφάσεων. Ένα δέντρο αποφάσεων μπορεί να χαρακτηριστεί από δύο οντότητες: τους κόμβους και τα φύλλα. Τα φύλλα αποτελούν αποφάσεις ή οι έξοδοι ενός δέντρου απόφασης, ενώ οι κόμβοι αποτελούν τα σημεία στα οποία χωρίζονται τα δεδομένα [39].



Σχήμα 3.5: Δυαδικό δέντρο που περιγράφει απλοϊκά τη λογική πίσω από τα δέντρα αποφάσεων [40]

Στην εικόνα περιγράφεται η διαδικασία πρόβλεψης αν ένας άνθρωπος θα είναι σε καλή φυσική κατάσταση αξιοποιώντας δεδομένα όπως την ηλικία, τη διατροφή και την σωματική άσκηση. Τα δεδομένα αυτά αποτελούν κόμβοι απόφασης του δέντρου, ενώ τα αποτελέσματα “fit” και “unfit” που είναι και οι έξοδοι του δέντρου μας είναι τα φύλλα του. Στο συγκεκριμένο παράδειγμα, πρόκειται για ένα δέντρο απόφασης που αποτελεί δυαδικής ταξινόμησης τύπου ναι/όχι αποφάσεων.

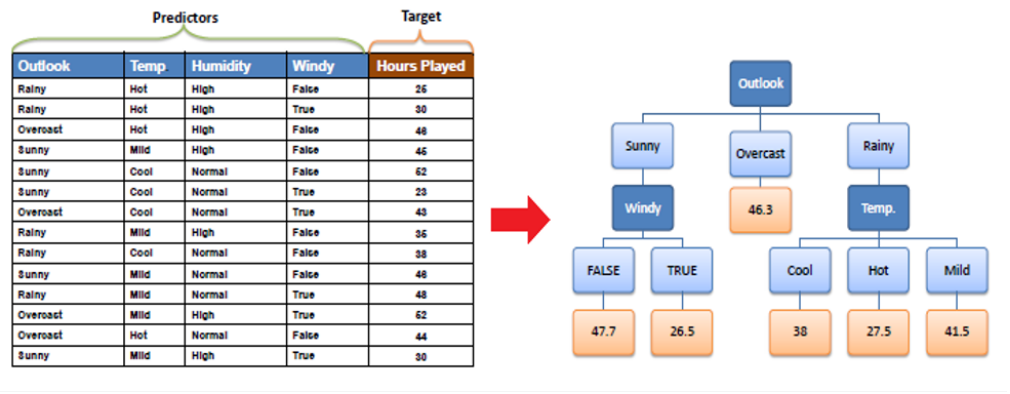
Γενικότερα, υπάρχουν δύο είδη δέντρων αποφάσεων: τα δέντρα παλινδρόμησης και τα δέντρα ταξινόμησης. Τα δέντρα παλινδρόμησης εξάγουν συνεχή αποτελέσματα, δηλαδή προκύπτει ένα αριθμητικό αποτέλεσμα που περιγράφει κάποια πρόβλεψη.

Αντίθετα, τα δέντρα ταξινόμησης, όπως του παραδείγματος της εικόνας, ταξινομούν τα δεδομένα με τρόπο δυαδικό και οι μεταβλητές είναι κατηγορηματικές (categorical). Μία βασική έννοια των δέντρων αποφάσεων αποτελεί η εντροπία (entropy), η οποία αντιπροσωπεύει το μέτρο του ποσού αβεβαιότητας ή τυχαιότητας των δεδομένων. Ο τύπος της είναι ο εξής:

$$H(S) = \sum_{x \in X} P(x) \log_2 \frac{1}{P(x)}$$

.

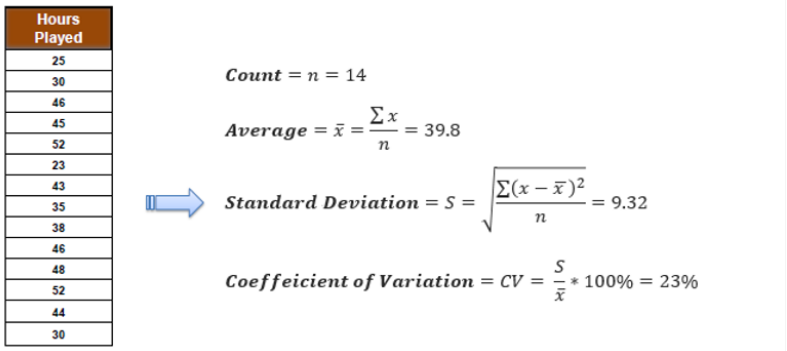
Όσο μικρότερη η τιμή της εντροπίας, τόσο μικρότερη είναι η αβεβαιότητα του συνόλου δεδομένων. Μία ακόμη στοιχειώδης έννοια αποτελεί και το κέρδος πληροφορίας (information gain), συμβολίζεται ως IG(S,A) και αντιπροσωπεύει τη μεταβολή της εντροπίας κάποιου συνόλου S κατόπιν μιας απόφασης βάσει κάποιου χα-



Σχήμα 3.6: Δέντρο παλινδρόμησης όπου η έξοδος είναι ένα αριθμητικό δεδομένο, στη συγκεκριμένη περίπτωση αποτελεί ένας προβλεπόμενος αριθμός ωρών που θα παίξει κάποιος τένις βάσει των καιρικών συνθηκών [41]

ρακτηριστικού A. Ουσιαστικά, μετρά τη σχετική μεταβολή της εντροπίας σε σχέση με τις ανεξάρτητες μεταβλητές. Ο απλός τύπος του είναι: $IG(S,A) = H(S) - H(S,A)$ ή αλλιώς, $IG(S,A) = H(S) - \sum_{i=0}^{n} P(x) \cdot H(x)$, όπου IG(S,A) είναι το κέρδος πληροφορίας εφαρμόζοντας το χαρακτηριστικό A στο σύνολο S, το $H(S)$ αντιπροσωπεύει την εντροπία του συνόλου S πριν την εφαρμογή του χαρακτηριστικού A και το δεύτερο κομμάτι της εξίσωσης υπολογίζει την εντροπία ύστερα από την εφαρμογή του χαρακτηριστικού. Τέλος, μία ακόμη έννοια είναι η τυπική απόκλιση (standard deviation).

Ένα δέντρο αποφάσεων είναι χτισμένο από πάνω προς τα κάτω από έναν ριζικό κόμβο και κατά μήκος του περιλαμβάνει τη διαίρεση των δεδομένων σε υποσύνολα που περιέχουν στιγμιότυπα με παρόμοιες τιμές (ομογενή). Η τυπική απόκλιση χρησιμοποιείται προκειμένου να υπολογιστεί η ομοιογένεια ενός αριθμητικού δείγματος. Αν ένα δείγμα είναι πλήρως ομογενές, τότε η τυπική απόκλιση ισούται με μηδέν [5].



Σχήμα 3.7: Παράδειγμα υπολογισμού του Standard Deviation (S) για ένα χαρακτηριστικό που χρησιμοποιείται για το χτίσιμο του δέντρου, όπου το CV είναι ο συντελεστής της απόκλισης που καθορίζει το τέλος των διακλαδώσεων, το ν είναι το πλήθος των δεδομένων και το avg αντιστοιχεί στην τιμή των κόμβων-φύλλων [42]

Ο συνηθέστερος αλγόριθμος που χρησιμοποιείται για ένα δέντρο απόφασης

είναι ο Iterative Dichotomiser 3 (ID3) που εκτελεί μία από πάνω προς τα κάτω (top-down) άπληστη αναζήτηση μέσω του χώρου των πιθανών κλαδιών χωρίς παλινδρόμηση. Στην περίπτωση ενός δέντρου παλινδρόμησης, αντί να χρησιμοποιείται ο παράγοντας του κέρδους πληροφορίας (information gain), αξιοποιείται η τυπική απόκλιση (SD). Ο αλγόριθμος ακολουθεί την εξής επαναληπτική διαδικασία: Πρώτα, δημιουργείται ένας κόμβος ρίζα για το δέντρο. Στη συνέχεια, αν όλα τα παραδείγματα είναι θετικά, τότε επιστρέφεται ένα φύλλο με ετικέτα "yes", ενώ αν είναι αρνητικά, επιστρέφεται ένα φύλλο με ετικέτα "no". Αν δεν υπάρχουν περισσότερα χαρακτηριστικά προς εξέταση, τότε επιστρέφεται η ετικέτα που παρουσιάζει τη μεγαλύτερη συχνότητα.

Στην παρούσα εργασία χρησιμοποιήθηκαν τα δέντρα αποφάσεων και για τη διαδικασία επιλογής χαρακτηριστικών αλλά και για την ίδια την ταξινόμηση και πρόβλεψη του αποτελέσματος που θέλαμε να εκτιμήσουμε. Η διαδικασία επιλογής χαρακτηριστικών με τη χρήση δέντρων απόφασης περιλαμβάνει αρχικά το διαχωρισμό της μεταβλητής-στόχου από τα χαρακτηριστικά στο σύνολο δεδομένων. Στη συνέχεια εκπαιδεύεται ένας ταξινομητής δέντρων απόφασης στα δεδομένα. Το μοντέλο υπολογίζει τη σημασία κάθε χαρακτηριστικού με βάση το πόσο αποτελεσματικά διαχωρίζει τα δεδομένα για την πρόβλεψη της μεταβλητής-στόχου. Στη συνέχεια, αυτές οι εισαγωγές χαρακτηριστικών ταξινομούνται σε φθίνουσα σειρά. Τα κορυφαία χαρακτηριστικά, σύμφωνα με τη σημασία τους, επιλέγονται με βάση τον καθορισμένο αριθμό χαρακτηριστικών που πρέπει να διατηρηθούν. Αυτά τα επιλεγμένα χαρακτηριστικά χρησιμοποιούνται για τη δημιουργία ενός νέου συνόλου δεδομένων, το οποίο περιλαμβάνει μόνο τα πιο σημαντικά χαρακτηριστικά μαζί με τη μεταβλητή-στόχο. Αυτό το σύνολο δεδομένων αποθηκεύεται για περαιτέρω ανάλυση και τα ονόματα των επιλεγμένων χαρακτηριστικών αποθηκεύονται επίσης για αναφορά. Η διαδικασία αυτή βοηθά στη μείωση της διαστατικότητας των δεδομένων, ενώ παράλληλα διατηρεί τα πιο σημαντικά χαρακτηριστικά για την προγνωστική μοντελοποίηση.

Η διαδικασία πρόβλεψης με τη χρήση ενός ταξινομητή δέντρων απόφασης περιλαμβάνει διάφορα βήματα για να εξασφαλιστεί ένα ακριβές και αξιόπιστο μοντέλο. Πρώτον, το σύνολο δεδομένων φορτώνεται και καθαρίζεται, με τη μεταβλητή-στόχο να διαχωρίζεται από τα χαρακτηριστικά. Στη συνέχεια, το σύνολο δεδομένων υποβάλλεται σε επιλογή χαρακτηριστικών με βάση διάφορες μεθόδους (π.χ. συσχέτιση, δέντρα απόφασης, FFS, Fisher Scores κ.λπ.), κάθε μία από τις οποίες παράγει ένα υποσύνολο χαρακτηριστικών που είναι πιο συναφή για το έργο πρόβλεψης.

Για κάθε επιλεγμένο σύνολο χαρακτηριστικών, ένα μοντέλο Δέντρου Απόφασης υφίσταται ρύθμιση υπερπαραμέτρων μέσω GridSearchCV με StratifiedKFold cross-validation για την εύρεση του καλύτερου συνδυασμού παραμέτρων. Οι υπερπαράμετροι περιλαμβάνουν το κριτήριο για τη διάσπαση, τον τύπο του διαχωριστή, το μέγιστο βάθος του δέντρου, τα ελάχιστα δείγματα που απαιτούνται για τη διάσπαση ενός εσωτερικού κόμβου, τα ελάχιστα δείγματα που απαιτούνται για να είναι κόμβος φύλλου και τα μέγιστα χαρακτηριστικά που λαμβάνονται υπόψη για τη διάσπαση.

Μετά τον προσδιορισμό των καλύτερων παραμέτρων του μοντέλου, το μοντέλο αξιολογείται χρησιμοποιώντας προβλέψεις διασταυρούμενης επικύρωσης με 3 διαιρέσεις. Υπολογίζονται διάφορες μετρικές, συμπεριλαμβανομένης της ακρίβειας, μιας λεπτομερούς έκθεσης ταξινόμησης και ενός πίνακα σύγχυσης, ο οποίος αποθηκεύεται για περαιτέρω ανάλυση. Η απόδοση του μοντέλου απεικονίζεται με τη χρήση καμπυλών ROC και διαγραμμάτων απώλειας λογαρίθμου. Η καμπύλη ROC αξιολογεί την ικανότητα του μοντέλου να διακρίνει μεταξύ των κλάσεων, ενώ το διάγραμμα απώλειας λογαρίθμου δείχνει την ασφάλεια πρόβλεψης του μοντέλου σε διάφορες αναδιπλώσεις.

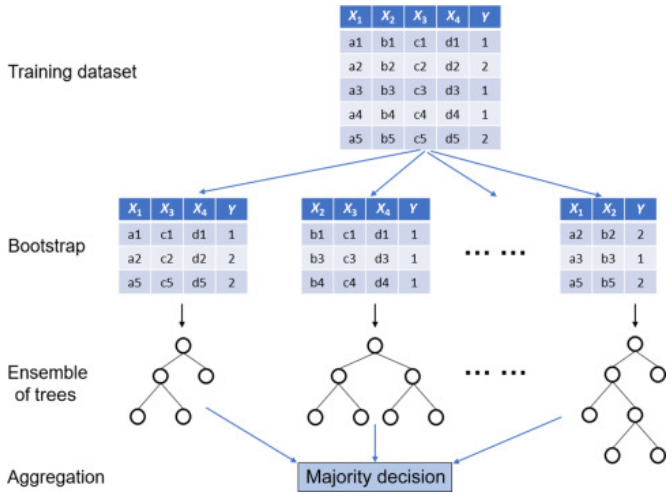
Τέλος, όλα τα αποτελέσματα, συμπεριλαμβανομένων των καλύτερων παραμέτρων, των αναφορών, των πινάκων σύγχυσης και των δεδομένων της καμπύλης ROC, αποθηκεύονται σε μορφές JSON και CSV.Οι βαθμολογίες ακρίβειας για κάθε μέθοδο συγκεντρώνονται σε ένα DataFrame και αποθηκεύονται, παρέχοντας μια ολοκληρωμένη επισκόπηση της απόδοσης του μοντέλου σε διάφορες μεθόδους επιλογής χαρακτηριστικών. Αυτή η συστηματική προσέγγιση εξασφαλίζει ισχυρή αξιολόγηση και επιλογή μοντέλου για αξιόπιστες προβλέψεις.

Για τον ταξινομητή δέντρου αποφάσεων, εστιάσαμε σε μετρικές ταξινόμησης όπως η ακρίβεια, η έκθεση ταξινόμησης, ο πίνακας σύγχυσης, οι καμπύλες ROC και τα διαγράμματα λογαριθμικής απώλειας. Αυτές οι μετρικές παρέχουν πληροφορίες σχετικά με την ικανότητα του μοντέλου να ταξινομεί σωστά τις περιπτώσεις και την απόδοσή του σε διάφορα κατώτατα όρια.

3.5 Random Forests (RF)

Ο ταξινομμητής Random Forests είναι μία μέθοδος που εκπαιδεύει παράλληλα πολλαπλά δέντρα αποφάσεων εφαρμόζοντας την τεχνική bootstrapping η οποία στη συνέχεια συνοδεύεται από την τεχνική aggregation. Η πρώτη τεχνική εξασφαλίζει ότι μερικά δέντρα αποφάσεων εκπαιδεύονται παράλληλα σε διάφορα υποσύνολα δεδομένων του γενικότερου συνόλου δεδομένου αξιοποιώντας διαφορετικά υποσύνολα των διαθέσιμων χαρακτηριστικών. Η τεχνική bootstrapping εξασφαλίζει ότι κάθε δέντρο αποφάσεων που ανήκει στο random forest είναι ξεχωριστό και στέλνει σε κάθε δέντρο ένα υποσύνολο του συνόλου δεδομένων, βέβαια, δεν είναι απαραίτητο τα δεδομένα να βρίσκονται αποκλειστικά σε ένα υποσύνολο [43], [44]. Επομένως, η διαφορά που έχουν τα RF δέντρα σε σχέση με τα απλά δέντρα αποφάσεων είναι ότι αναλαμβάνουν έναν μικρό αριθμό δεδομένων σε κάθε τερματικό κόμβο [43]. Στη συνέχεια, για το τελικό αποτέλεσμα, ο ταξινομητής RF συγκεντρώνει το σύνολο των εξόδων από κάθε δέντρο. Αυτό έχει ως αποτέλεσμα, ο συγκεκριμένος ταξινομητής να μπορεί να κάνει καλή γενίκευση [44], [45]. Γενικότερα, πρόκειται για έναν ακριβή αλγόριθμο που έχει την ικανότητα να χειρίζεται χιλιάδες μεταβλητές χωρίς απώλειες και τη χειροτέρευση της ακρίβειας του.

Το Random Forest αξιοποιήθηκε και για τη διαδικασία επιλογής χαρακτη-



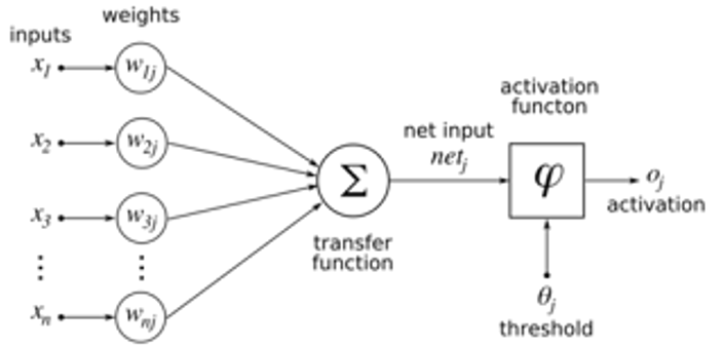
Σχήμα 3.8: Μία αναπαράσταση του ταξινομητή RF και της διδικασίας που ακολουθεί κατά την επεξεργασία των δεδομένων προκειμένου να παράξει κάποιο αποτέλεσμα [46]

ριστικών αλλά και για την υλοποίηση εκτιμήσεων. Η διαδικασία επιλογής χαρακτηριστικών ξεκίνησε με τη φόρτωση και τον καθαρισμό του συνόλου δεδομένων, διασφαλίζοντας ότι η μεταβλητή-στόχος διαχωρίστηκε από τα χαρακτηριστικά. Στη συνέχεια χρησιμοποιήθηκε ένας ταξινομητής Random Forest για την εκπαίδευση του μοντέλου στα δεδομένα. Το μοντέλο υπολόγισε τη σημασία κάθε χαρακτηριστικού, υποδεικνύοντας πόσο αποτελεσματικά συνέβαλε κάθε χαρακτηριστικό στην πρόβλεψη. Αυτές οι σημαντικότητες (importances) των χαρακτηριστικών ταξινομήθηκαν και επιλέχθηκαν τα κορυφαία χαρακτηριστικά με βάση τις βαθμολογίες αυτές. Στη συνέχεια, το σύνολο δεδομένων βελτιώθηκε ώστε να περιλαμβάνει μόνο αυτά τα επιλεγμένα χαρακτηριστικά, τα οποία αποθηκεύτηκαν για περαιτέρω ανάλυση.

Μετά την επιλογή των χαρακτηριστικών, χρησιμοποιήθηκε ο ταξινομητής Random Forest για τη μοντελοποίηση πρόβλεψης. Το σύνολο δεδομένων φορτώθηκε και η ρύθμιση των υπερπαραμέτρων πραγματοποιήθηκε με τη χρήση του GridSearchCV με διασταυρωμένη επικύρωση StratifiedKFold. Το βήμα αυτό περιελάμβανε τη δοκιμή διαφόρων υπερπαραμέτρων για τον εντοπισμό της καλύτερης διαμόρφωσης του μοντέλου. Μόλις βρέθηκαν οι βέλτιστες παράμετροι, το μοντέλο αξιολογήθηκε πάλι με τη χρήση προβλέψεων διασταυρούμενης επικύρωσης. Υπολογίστηκαν μετρικές όπως η ακρίβεια, οι εκθέσεις ταξινόμησης και οι πίνακες σύγχυσης. Επιπλέον, δημιουργήθηκαν καμπύλες ROC και διαγράμματα λογαριθμικής απώλειας για την οπτικοποίηση της απόδοσης του μοντέλου. Όλες οι μετρικές αξιολόγησης, συμπεριλαμβανομένων των καλύτερων παραμέτρων και των αναφορών επιδόσεων, αποθηκεύτηκαν σε μορφές JSON και CSV για μελλοντική αναφορά και ανάλυση.

3.6 Multilayered Perceptrons (MLP)

Ο όρος νευρωνικά δίκτυα αναφέρεται σε υπολογιστικά συστήματα με διασυνδεδεμένους κόμβους οι οποίοι προσεγγίζουν τους νευρώνες του ανθρώπινου εγκεφάλου. Με τη χρήση αλγορίθμων, τα νευρωνικά δίκτυα αποκτούν την ικανότητα να αναγνωρίζουν μοτίβα και συσχετίσεις ανάμεσα σε ακατέργαστα δεδομένα, μπορούν να τα ομαδοποιούν και να τα ταξινομούν, και με τον χρόνο, αναπτύσσονται διαρκώς, και βελτιώνονται. Με άλλα λόγια, τα νευρωνικά δίκτυα, προσομοιάζοντας την λειτουργία των εγκεφαλικών νευρώνων, οι οποίοι αναπτύσσουν συνάψεις με στόχο την μεταφορά των νευρικών ώσεων, προσπαθούν να μιμηθούν τη συμπεριφορά του ανθρώπινου εγκεφάλου. Η λειτουργία τους χαρακτηρίζεται ως τέτοια, καθώς η ενεργοποίηση του κάθε κόμβου, εξαπλώνεται κατά μήκος όλου του δικτύου, δημιουργώντας με αυτόν τον τρόπο μία απάντηση στην έξοδό του. Οι συνδέσεις που δημιουργούνται επιτρέπουν την διαπέραση των σημάτων από κόμβο σε κόμβο, καθώς σε κάθε τμήμα του δικτύου περνάει από διάφορα επίπεδα επεξεργασίας μέχρις ότου φτάσει στην τελική έξοδο. Το μοντέλο ενός τεχνητού νευρώνα (perceptron) παρουσιάζεται στο σχήμα παρακάτω:

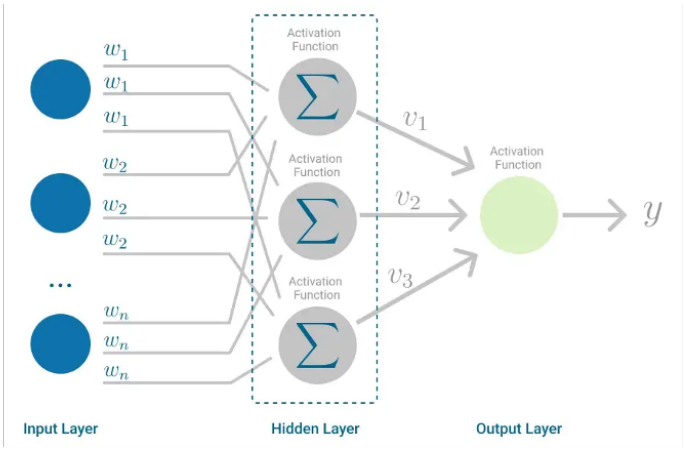


Σχήμα 3.9: Το μοντέλο ενός τεχνητού νευρώνα (perceptron) με τις επαυξημένες εισόδους, τα βάρη, τη συνάρτηση ενεργοποίησης και την έξοδο που περιγράφονται παρακάτω [47].

Η πιο δημοφιλής δομή νευρωνικού δικτύου είναι το Multilayer Perceptron που περιλαμβάνει κατά σειρά το στρώμα εισόδου, τα βάρη κάθε στρώματος, τα ενδιάμεσα ¨κρυμμένα¨ στρώματα τα οποία περιλαμβάνουν μία συνάρτηση ενεργοποίησης και τέλος, το στρώμα εξόδου (ή αλλιώς στόχου). Η δομή αυτή προκύπτει από τη διαδοχική συνένωση πολλών Perceptrons. Τα επιμέρους στρώματα ενώνονται μεταξύ τους μέσω των κόμβων και με αυτόν τον τρόπο σχηματίζονται τα ¨δίκτυα¨. Με τη προσθήκη περισσότερων στρωμάτων (layers) στην εσωτερική δομή των νευρωνικών δικτύων, δημιουργούνται τα βαθιά νευρωνικά δίκτυα.

Ο αλγόριθμος που ακολουθεί το MLP είναι ο ακόλουθος: Καταρχάς, οι είσοδοι $(x_1, x_2, \ldots, x_n)$ προωθούνται μέσω του MLP λαμβάνοντας το γινόμενο της εισόδου επί τα βάρη $(w_{1,}w_2, \ldots, w_n)$ που υπάρχουν μεταξύ του στρώματος εισόδου και του κρυφού στρώματος. Στα βάρη, πριν αρχίσει να τρέχει ο αλγόριθμος, αναθέτονται τυχαίες τιμές, οι οποίες μετά από κάθε επανάληψη προσαρμόζονται έτσι

ώστε να εξασφαλιστεί το χαμηλότερο δυνατό σφάλμα στα δεδομένα εκπαίδευσης [24], [27]. Αυτό πραγματοποιείται με έναν αλγόριθμο back propagation που επιδιώκει να ελαχιστοποιήσει το mean squared error [24]. Τα χαρακτηριστικά περνούν στο επόμενο στάδιο της συνάρτησης εισόδου u ως: $u(x) = \sum_{i=1}^{n} w_i x_i$ [24]. Στη συνέχεια, αξιοποιούνται οι συναρτήσεις ενεργοποίησης σε κάθε ένα από τα κρυφά επίπεδα. Υστερα, η υπολογισμένη έξοδος από το κρυφό στρώμα θα προωθηθεί μέσω της συνάρτησης ενεργοποίησης προς το επόμενο επίπεδο, πάλι υπολογίζοντας το γινόμενο της εξόδου επί τα αντίστοιχα βάρη του επόμενου επιπέδου. Ο υπολογισμός της εξόδου αυτής μπορεί να εκφραστεί από τη σχέση: $h_j = a_1 \left[\sum_{i=1}^{n} w_{ij} x_i + \theta_j\right]$, όπου τα w_{ij} αντιπροσωπεύουν τα βάρη, το x_i τις εισόδους και τα θ_j τα biases που ανήκουν στο j-th κρυμμένο στρώμα. Η έξοδος από κάθε νευρώνα είναι η ακόλουθη: $y = f(u(x)) = 1$, αν $u(x) > \theta$ ή 0, αλλιώς, όπου θ είναι κάποιο κατώφλι [18]. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου η πληροφορία να φτάσει στο τελικό επίπεδο εξόδου. Ουσιαστικά, ο νευρώνας θα καθορίζει αν η σχέση: $w_1x_1 + w_2x_2 + \cdots + w_nx_n - \theta > 0$ είναι αληθής ή όχι. Η εξίσωση $w_1x_1 + w_2x_2 + \cdots + w_nx_n - \theta = 0$ ορίζει ένα υπερεπίπεδο και ο νευρώνας επιστρέφει ¨1¨ εάν η είσοδος βρίσκεται πάνω από αυτό και ¨0¨ αν είναι κάτω. Για τον λόγο αυτό ένας τεχνητός νευρώνας ανήκει στην κατηγορία των γραμμικών ταξινομητών [24].



Σχήμα 3.10: Ένα μοντέλο MLP όπου [στα αριστερά με μπλε χρώμα] φαίνονται οι είσοδοι, οι οποίες πολλαπλασιαζόμενες επί τα βάρη οδηγούνται στις αντίστοιχες συναρτήσεις ενεργοποίησης των κρυφών στρωμάτων [με γκρι χρώμα] που δρομολογούν τις υπολογισμένες εξόδους ως προς τη συνάρτηση ενεργοποίησης του στρώματος εξόδου [με πράσινο χρώμα] και τέλος στην έξοδο y.

Βασική λειτουργία των νευρωνικών δικτύων είναι να συλλέγουν την πληροφορία, να αξιολογούν την επάρκειά της και αναλόγως να συνεχίσουν να τη διαπερνούν μέσα από τους κόμβους, επιφέροντας περαιτέρω επεξεργασία. Στα κρυμμένα στρώματα πραγματοποιείται η επεξεργασία μέσω ενός συστήματος σταθμισμένων συνδέσεων. Οι κόμβοι στο εσωτερικό συνδυάζουν τα δεδομένα εισόδου με ένα σύνολο συντελεστών με κατάλληλους συντελεστές βαρύτητας. Εφόσον το προκύπτουν σταθμισμένο άθροισμα πληροί τον απαραίτητο όγκο πληροφορίας, το νευρωνικό

δίκτυο αποφαίνεται για την προώθησή του εντός του δικτύου, μέσω των συναρτήσεων ενεργοποίησης. Οι εν λόγω συναρτήσεις καθορίζουν -μεταξύ άλλων- την κανονικοποίηση που εφαρμόζεται στην είσοδο του επόμενου κόμβου. Διακρίνονται διάφορες περιπτώσεις συναρτήσεων ενεργοποίησης, όπως είναι επί παραδείγματι η σιγμοειδής συνάρτηση, η softmax (συνήθης εφαρμογή σε ταξινόμηση κατηγοριών, που δίνει μία πιθανότητα ένταξης στην εκάστοτε κλάση), η υπερβολική εφαπτομένη και η ReLu (Rectified Linear Unit)[48], Ακολούθως, τα ενδιάμεσα στρώματα του νευρωνικού δικτύου διασφαλίζουν την μετάβαση της πληροφορίας στην έξοδο.

Η υλοποίηση του μοντέλου πρόβλεψης με τη χρήση ενός πολυεπίπεδου αντιληπτικού (MLP classifier) περιλάμβανε διάφορα βήματα για να εξασφαλιστεί η ακρίβεια και η αξιοπιστία των προβλέψεων. Εφαρμόστηκαν διάφορες μέθοδοι επιλογής χαρακτηριστικών για να εντοπιστούν τα πιο σχετικά χαρακτηριστικά για το μοντέλο πρόβλεψης. Αυτές οι μέθοδοι περιλάμβαναν τη συσχέτιση, τα δέντρα αποφάσεων, την επιλογή χαρακτηριστικών προς τα εμπρός (FFS), τις βαθμολογίες Fisher, το πληροφοριακό κέρδος (IG), την κανονικοποίηση LASSO, τα τυχαία δάση, την αναδρομική εξάλειψη χαρακτηριστικών (RFE), την αλληλουχική επιλογή χαρακτηριστικών (SFS), και τον συντελεστή διόγκωσης διακύμανσης (VIF). Για κάθε μέθοδο, δημιουργήθηκε και αποθηκεύτηκε ένα σύνολο δεδομένων που περιείχε τα επιλεγμένα χαρακτηριστικά για περαιτέρω ανάλυση.

Για τη βελτιστοποίηση του MLP classifier, πραγματοποιήθηκε αναζήτηση πλέγματος χρησιμοποιώντας GridSearchCV από το sklearn.model_selection. Το πλέγμα παραμέτρων περιλάμβανε διάφορες ρυθμίσεις για τα
hidden_layer_sizes, τις συναρτήσεις activation, τους αλγόριθμους solver, τις αρχικές τιμές μάθησης learning_rate_init, και τις μέγιστες επαναλήψεις max_iter. Η αναζήτηση πλέγματος χρησιμοποίησε διασταυρούμενη επικύρωση Stratified K-Folds με 5 διαίρεση για να εξασφαλίσει μια αξιόπιστη αξιολόγηση του μοντέλου σε διαφορετικά υποσύνολα των δεδομένων.

Το καλύτερο μοντέλο που εντοπίστηκε από την αναζήτηση πλέγματος αξιολογήθηκε χρησιμοποιώντας διασταυρούμενη επικύρωση με 5 διαιρέσεις. Μετρήθηκαν η ακρίβεια του μοντέλου και δημιουργήθηκε μια λεπτομερής αναφορά ταξινόμησης που περιλάμβανε την ακρίβεια, την ανάκληση και το F1-score για κάθε κλάση. Δημιουργήθηκε ένας πίνακας σύγχυσης για να οπτικοποιηθούν οι επιδόσεις του μοντέλου όσον αφορά τις αληθινά θετικές, αληθινά αρνητικές, ψευδώς θετικές και ψευδώς αρνητικές προβλέψεις. Ο πίνακας αυτός αποθηκεύτηκε ως αρχείο CSV για περαιτέρω επιθεώρηση.

Επιπλέον, σχεδιάστηκε και αποθηκεύτηκε η καμπύλη ROC για να δείξει τη σχέση μεταξύ του ποσοστού αληθινά θετικών και του ποσοστού ψευδώς θετικών σε διάφορες ρυθμίσεις κατωφλίου. Ο υπολογισμός της περιοχής κάτω από την καμπύλη (AUC) περιγράφηκε και απεικονίστηκε. Για την εκτίμηση της ακρίβειας των πιθανοτήτων, υπολογίστηκε και αποθηκεύτηκε η απώλεια λογαρίθμου (log loss) για κάθε αναδίπλωση κατά τη διασταυρούμενη επικύρωση.

Τέλος, τα αποτελέσματα από τις αξιολογήσεις, συμπεριλαμβανομένων των βαθμολογιών ακρίβειας, των αναφορών ταξινόμησης, των πινάκων σύγχυσης, των καμπυλών ROC και των διαγραμμάτων απώλειας λογαρίθμου, αποθηκεύτηκαν για κάθε μέθοδο επιλογής χαρακτηριστικών. Αυτά τα αποτελέσματα αποθηκεύτηκαν σε μορφές JSON και CSV για εύκολη πρόσβαση και ανάλυση.

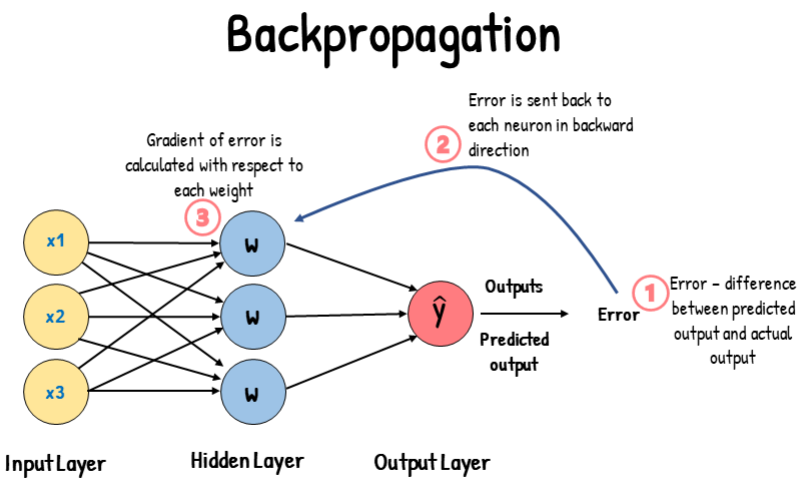
Αυτή η διαδικασία εξασφάλισε μια ολοκληρωμένη αξιολόγηση των επιδόσεων του MLP classifier σε διαφορετικές μεθόδους επιλογής χαρακτηριστικών, παρέχοντας ένα αξιόπιστο πλαίσιο για ακριβείς προβλέψεις. Η χρήση διασταυρούμενης επικύρωσης και εκτεταμένης βελτιστοποίησης παραμέτρων βοήθησε στον εντοπισμό της βέλτιστης διαμόρφωσης του μοντέλου, ενώ οι λεπτομερείς μετρήσεις αξιολόγησης και οι οπτικοποιήσεις παρείχαν βαθιές πληροφορίες για τις επιδόσεις του μοντέλου.

3.7 Back Propagation Network (BP)

Από όλους τους αλγόριθμους επιβλεπόμενης μάθησης, ο αλγόριθμος BP πιθανώς να είναι ο πιο διαδεδομένος από αυτούς, πράγμα που οφείλεται στην απλότητα του. Ουσιαστικά, πρόκειται για έναν αλγόριθμο που έχει παρόμοια λογική με αυτήν του κανόνα αλυσίδας και του gradient descent [49]. Ο κύριος στόχος του αλγόριθμου αυτού είναι η διόρθωση των λαθών ξεκινώντας από το τέλος, δηλαδή της εξόδους, προς την είσοδο, δηλαδή επιδιώκει να ελαχιστοποιήσει τη διαφορά μεταξύ των επιθυμητών στόχων - εξόδων και της εξόδου που επιτυγχάνει. Ο αλγόριθμος BP αποτελείται από δίκτυα των οποίων η συνάρτηση κόστους τους έχει μία τάση να διαμοιράζεται μεταξύ των κόμβων τους με στόχο την διόρθωση του αλγόριθμου. Αυτό σημαίνει ότι τα επίπδεα του BP, δεδομένου ότι θα έχει τουλάχιστον ένα κρυφό επίπεδο, διαμοιράζονται μεταξύ τους κατά την επεξεργασία κάθε χαρακτηριστικού του διανύσματος εισόδου. Η διόρθωση του αλγόριθμου επιτυγχάνεται με τη βοήθεια διάφορων μετρικών και ελέγχων, και συγκερκιμένα σε πολλές περιπτώσεις, με τον αλγόριθμο gradient descent ή το delta rule, οι οποίοι ρυθμίζουν το σύστημα προσαρμόζοντας τις τιμές των βαρών που υπάρχει σε κάθε νευρώνα-κόμβο ώστε να φέρουν τις τιμές εξόδου που παράγει ο αλγόριθμος πιο κοντά στις επιθυμητές εξόδους. Με αυτόν τον τρόπο η εκμάθηση του αλγόριθμου γίνεται με έναν πιο αποτελεσματικό τρόπο, παρόλου που η ¨λογική¨ πίσω από την αναπροσαρμογή των βαρών των επιπέδων δεν είναι προφανής [50], [51].

Ο αλγόριθμος gradient descent χρησιμοποιεί μία σταδιακή διαδικασία που παρέχει την πληροφορία που χρειάζεται ο BP προκειμένου να διαμορφώσει τις τιμές των βαρών των κόμβων με τέτοιον τρόπο ώστε να προκύψει η επιθυμητή έξοδος. Για τη διαδικασία αυτή, υπάρχει η συνάρτηση κόστους που είναι υπεύθυνη να υπολογίσει το σφάλμα που προκύπτει κατά την αναπαραγωγή της εξόδου, δηλαδή, το σφάλμα υπολογίζεται στο τελικό επίπεδο του δικτύου. Το σφάλμα είναι η προαναφερόμενη διαφορά, ή απόσταση, μεταξύ των τιμών της παραγόμενης εξόδου απο τον αλγόριθμο και της επιθυμητής προκαθορισμένης εξόδου. Επομένως, ο αλγόριθμος

αυτός επιδιώκει να ελαχιστοποιήσει όσο το δυνατόν περισσότερο γίνεται το σφάλμα αυτό και να αυξήσει την ακρίβεια. Κατά την οπισθοδιάδοση, το σφάλμα διαδίδεται προς τα πίσω, δηλαδή από τον κόμβο εξόδου προς τον κόμβο εισόδου, διαπερνώντας και από τους ενδιάμεσους, κρυφούς κόμβους (εφόσον υπάρχουν) έτσι ώστε κάθε κόμβος να προσαρμόζει τις τιμές των βαρών του ανάλογα, σε περίπτωση που έπαιξε ρόλο στην παραγωγή αυτού του σφάλματος. Ο κάθε κόμβος επηρεάζει την έξοδο εφόσον του το επιτρέπει η συνάρτηση ενεργοποίησης που υπάρχει σε κάθε κόμβο, η οποία σε περίπτωση που δεχθεί κάποια τιμή που ξεπερνάει το κατώφλι της, τότε ενεργοποιεί τον νευρώνα να αναπαράγει κάποια έξοδο και να το στέλνει στο επόμενο επίπεδο μέχρι να φτάσει τον κόμβο εξόδου όπου εκεί υπολογίζεται η τελική έξοδος (30). Αυτή η διαδικασία επαναλαμβάνεται μέχρι να βελτιωθεί ικανοποιητικά η διαφορά μεταξύ της εξόδου του δικτύου με τον στόχο [51].



Σχήμα 3.11: Αναπαράσταση της διαδικασίας BP [52]

Γενικότερα, αυτός ο αλγόριθμος χρησιμοποιείται με διάφορους τρόπους στον κόσμο της μηχανικής μάθησης. Συνήθως βρίσκεται ενσωματωμένους σε πολυπλοκότερα μοντέλα της μηχανικής μάθησης όπως σε νευρωνικά δίκτυα, αλλά μπορούν επίσης να χρησιμοποιηθούν και σε προβλήματα ταξινόμησης και παλινδρόμησης [53].

Στην έρευνα αυτή, κατασκευάστηκε ένας αλγόριθμος ενός τυπικού νευρωνικού δικτύου με back propagation. Αντί να αξιοποιηθούν έτοιμες συναρτήσεις από τις βιβλιοθήκες της Python, κατασκευάστηκαν κατάλληλες συναρτήσεις ενεργοποίησης βάσει της θεωρίας, συγκεκριμένα, η συνάρτηση που χρησιμοποιήθηκε ως συνάρτηση ενεργοποίησης ήταν η σιγμοϊδή και δημιουργήθηκαν κλάσεις που κατασκευάζουν το νευρωνικό δίκτυο για τις οποίες δοκιμάστηκαν διαφορετικές αρχιτεκτονικές (αριθμός επιπέδων και κόμβων ανά επιπέδων) με σκοπό να βρεθεί το βέλτιστο μοντέλο που να πραγματοποιεί τις ακριβέστερες προβλέψεις. Η διαδικασία αυτή περιλάμβανε πολλά βήματα για να εξασφαλιστεί η σωστή εκτέλεση του μοντέλου.

Το πρώτο βήμα αφορούσε τη φόρτωση του καθαρού συνόλου δεδομένων από το καθορισμένο μονοπάτι χρησιμοποιώντας τη βιβλιοθήκη pandas. Το σύνολο δεδομένων περιλάμβανε διάφορα χαρακτηριστικά και τη μεταβλητή στόχο Death που ήταν το επίκεντρο της πρόβλεψης.

Ακολούθησαν διάφορες μέθοδοι επιλογής χαρακτηριστικών για να εντοπιστούν τα πιο σχετικά χαρακτηριστικά για το μοντέλο πρόβλεψης. Αυτές οι μέθοδοι περιλάμβαναν τη συσχέτιση, τα δέντρα αποφάσεων, την επιλογή χαρακτηριστικών προς τα εμπρός (FFS), τις βαθμολογίες Fisher, το πληροφοριακό κέρδος (IG), την κανονικοποίηση LASSO, τα τυχαία δάση, την αναδρομική εξάλειψη χαρακτηριστικών (RFE), την αλληλουχική επιλογή χαρακτηριστικών (SFS), και τον συντελεστή διόγκωσης διακύμανσης (VIF). Για κάθε μέθοδο, δημιουργήθηκε και αποθηκεύτηκε ένα σύνολο δεδομένων που περιείχε τα επιλεγμένα χαρακτηριστικά για περαιτέρω ανάλυση.

Στη συνέχεια, χρησιμοποιήθηκαν διάφορες αρχιτεκτονικές του νευρωνικού δικτύου και εύρη επαναλήψεων (epochs) για την εκπαίδευση του μοντέλου. Οι αρχιτεκτονικές περιλάμβαναν διατάξεις όπως [5, 5], [10, 5], [15, 10], [20, 10], [20, 15], [30, 15], [30, 20], [40, 15], [40, 20], [50, 25], και [10, 10], ενώ οι επαναλήψεις κυμαίνονταν από 100 έως 1200.

Για κάθε συνδυασμό μεθόδου επιλογής χαρακτηριστικών και αρχιτεκτονικής νευρωνικού δικτύου, πραγματοποιήθηκε διασταυρούμενη επικύρωση χρησιμοποιώντας Stratified K-Folds με 5 διαίρεση για να εξασφαλιστεί η αξιοπιστία των αποτελεσμάτων. Η εκπαίδευση των νευρωνικών δικτύων πραγματοποιήθηκε με τη χρήση της μεθόδου οπισθοδιάδοσης, όπου το δίκτυο προσαρμόστηκε στα δεδομένα εκπαιδευσης μέσω πολλαπλών επαναλήψεων.

Μετά την εκπαίδευση, το μοντέλο αξιολογήθηκε χρησιμοποιώντας διάφορες μετρικές. Υπολογίστηκε η ακρίβεια του μοντέλου και δημιουργήθηκε μια αναλυτική αναφορά ταξινόμησης που περιλάμβανε την ακρίβεια, την ανάκληση και το F1-score για κάθε κατηγορία. Δημιουργήθηκε ένας πίνακας σύγχυσης για να απεικονιστούν οι επιδόσεις του μοντέλου σε πραγματικά θετικά, πραγματικά αρνητικά, ψευδώς θετικά και ψευδώς αρνητικά αποτελέσματα. Ο πίνακας αυτός αποθηκεύτηκε ως αρχείο CSV για περαιτέρω ανάλυση.

Επιπλέον, σχεδιάστηκε και αποθηκεύτηκε η καμπύλη ROC για να δείξει τη σχέση μεταξύ του ποσοστού αληθινά θετικών και του ποσοστού ψευδώς θετικών σε διάφορες ρυθμίσεις κατωφλίου. Υπολογίστηκε επίσης η περιοχή κάτω από την καμπύλη (AUC). Για την εκτίμηση της ακρίβειας των πιθανοτήτων, υπολογίστηκε και αποθηκεύτηκε η απώλεια λογαρίθμου (log loss) για κάθε διαίρεση κατά τη διασταυρούμενη επικύρωση.

Τα αποτελέσματα από τις αξιολογήσεις, συμπεριλαμβανομένων των βαθμολογιών ακρίβειας, των αναφορών ταξινόμησης, των πινάκων σύγχυσης, των καμπυλών ROC και των διαγραμμάτων απώλειας λογαρίθμου, αποθηκεύτηκαν για κάθε μέθοδο επιλογής χαρακτηριστικών. Αυτά τα αποτελέσματα αποθηκεύτηκαν σε μορ-

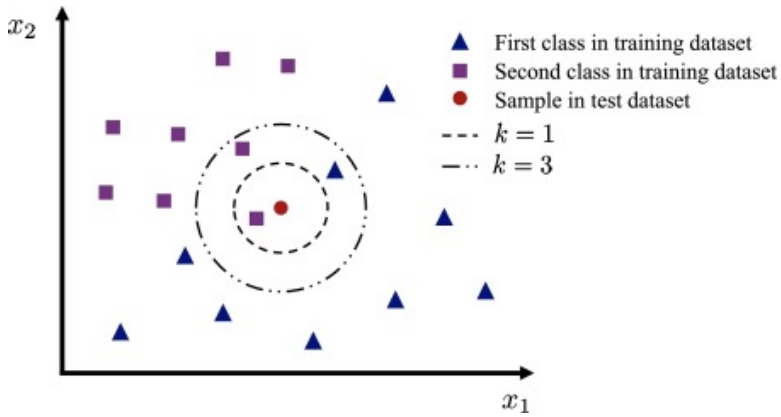
φές JSON και CSV για εύκολη πρόσβαση και ανάλυση.

3.8 K-Nearest Neighbors (KNN)

Ένας ακόμα πολύ δημοφιλής αλγόριθμος επιβλεπόμενης μάθησης είναι ο αλγόριθμος Κ πλησιέστερων γειτόνων. Πρόκειται για μια κλασική μη-παραμετρική μέθοδο που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση [54]. Η βασική λογική πίσω από τον αλγόριθμο αυτόν είναι ο υπολογισμός και ο εντοπισμός της απόστασης των δεδομένων ελέγχου προκειμένου να τα αντιστοιχίσει σε κάποια κατηγορία. Η μέθοδος θεωρείται στοιχειώδης, δηλαδή δεν χρησιμοποιεί μάθηση καθώς δεν υπάρχει καμία παράμετρος που να απαιτεί αυτορρύθμιση [55] και διαθέτει μόνο μία υπερπαράμετρο k που καθορίζεται από τον χρήστη εκ των προτέρων.

Η βασική αρχή του αλγόριθμου αυτού είναι ότι κατά τη διάρκεια της ταξινόμησης, τα δείγματα συγκρίνονται τοπικά με τα k γειτονικά δείγματα εκπαίδευσης σε έναν μεταβλητό χώρο, και η κατηγορία τους αποφασίζεται βάσει της ταξινόμησης των k πλησιέστερων γειτόνων τους. Ένας από τους πιο συνήθης τρόπους υπολογισμού των αποστάσεων μεταξύ των γειτόνων είναι μέσω των ευκλείδιων αποστάσεων [30], [55] μεταξύ του υποψήφιου προς ταξινόμηση δείγματος και των k γειτόνων του. Η εκτίμηση, και κατέπέκταση ταξινόμηση σε κάποια κατηγορία, καθορίζεται από την κατηγορία στην οποία βρίσκεται η πλειοψηφία των κοντινότερων γειτόνων του δείγματος.

Πιο αναλυτικά, έστω ότι υπάρχει ένα σύνολο δεδομένων εκπαίδευσης $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ και ένα δείγμα ελέγχου το x_o, ο στόχος του αλγόριθμου είναι να εκτιμήσει την κατηγορία του x_0. Κατά την διαδικασία εκπαίδευσης, το σύνολο δεδομένων D φορτώνεται και αποθηκεύεται και στη συνέχεια η διαδικασία ελέγχου αναζητά τους k πλησιέστερους γείτονες από το σύνολο δεδομένων βάσει της ευκλείδιας απόστασης $d(\mathbf{x}_0, \mathbf{x}_i) = ||\mathbf{x}_0 - \mathbf{x}_i||_2$. Η εκτίμηση εξαρτάται από την κατηγορία στην οποία θα ανήκει το μεγαλύτερο πλήθος των δεδομένων εκπαίδευσης.



Σχήμα 3.12: Η απεικόνιση του K-NN για τιμές της υπερπαραμέτρου k = 1 και k = 3 [56]

Αυτό που πρέπει να σημειωθεί είναι ότι παίζει μεγάλο ρόλο στην απόφαση

ταξινόμησης η κατάλληλη ανάθεση της υπερπαραμέτρου k, δηλαδή η επιλογή του πλήθους των γειτόνων. Αυτό μπορεί να αποτελέσει πρόβλημα, καθώς σε πολλές περιπτώσεις, η τιμή του πλήθους καθορίζεται αυθάιρετα από τον χρήστη [54], [55], και ο μόνος τρόπος πορκειμένου να βρεθεί η βέλτιστη τιμή είναι μέσω της διαδικασίας δοκιμής και λάθους (trial and error).

Για την εφαρμογή του μοντέλου πρόβλεψης χρησιμοποιώντας τον αλγόριθμο K-Nearest Neighbors (KNN), ακολούθησαν διάφορα στάδια για να εξασφαλιστεί η αξιοπιστία και η ακρίβεια των αποτελεσμάτων. Αρχικά, φορτώθηκε το σύνολο δεδομένων από το καθορισμένο μονοπάτι χρησιμοποιώντας τη βιβλιοθήκη pandas. Το σύνολο δεδομένων περιλάμβανε διάφορα χαρακτηριστικά και τη μεταβλητή στόχο Death, που αποτέλεσε το επίκεντρο της πρόβλεψης.

Ακολούθησαν διάφορες μέθοδοι επιλογής χαρακτηριστικών, όπως η συσχέτιση, τα δέντρα αποφάσεων, η επιλογή χαρακτηριστικών προς τα εμπρός (FFS), οι βαθμολογίες Fisher, το πληροφοριακό κέρδος (IG), η κανονικοποίηση LASSO, τα τυχαία δάση, η αναδρομική εξάλειψη χαρακτηριστικών (RFE), η αλληλουχική επιλογή χαρακτηριστικών (SFS), και ο συντελεστής διόγκωσης διακύμανσης (VIF). Για κάθε μέθοδο, δημιουργήθηκε και αποθηκεύτηκε ένα σύνολο δεδομένων με τα επιλεγμένα χαρακτηριστικά.

Για την εκπαίδευση του μοντέλου KNN, χρησιμοποιήθηκε αναζήτηση πλέγματος (GridSearchCV) με διασταυρούμενη επικύρωση Stratified K-Folds με 5 διαιρέσεις. Η αναζήτηση πλέγματος επέτρεψε τη δοκιμή διάφορων συνδυασμών υπερπαραμέτρων, όπως ο αριθμός των γειτόνων (n_neighbors), τα βάρη (weights), ο αλγόριθμος (algorithm), το μέγεθος φύλλου (leaf_size), και η παράμετρος p.

Μετά την επιλογή των καλύτερων υπερπαραμέτρων, το μοντέλο KNN αξιολογήθηκε μέσω διασταυρούμενης επικύρωσης. Υπολογίστηκαν διάφορες μετρικές, όπως η ακρίβεια του μοντέλου, η οποία αποθηκεύτηκε μαζί με μια λεπτομερή αναφορά ταξινόμησης που περιλάμβανε την ακρίβεια, την ανάκληση και το F1-score για κάθε κατηγορία. Δημιουργήθηκε επίσης ένας πίνακας σύγχυσης για να απεικονιστούν οι επιδόσεις του μοντέλου σε πραγματικά θετικά, πραγματικά αρνητικά, ψευδώς θετικά και ψευδώς αρνητικά αποτελέσματα. Ο πίνακας αυτός αποθηκεύτηκε ως αρχείο CSV για περαιτέρω ανάλυση.

Επιπλέον, σχεδιάστηκε και αποθηκεύτηκε η καμπύλη ROC για να δείξει τη σχέση μεταξύ του ποσοστού αληθινά θετικών και του ποσοστού ψευδώς θετικών σε διάφορες ρυθμίσεις κατωφλίου, καθώς και η περιοχή κάτω από την καμπύλη (AUC). Για την εκτίμηση της ακρίβειας των πιθανοτήτων, υπολογίστηκε και αποθηκεύτηκε η απώλεια λογαρίθμου (log loss) για κάθε διαίρεση κατά τη διασταυρούμενη επικύρωση.

Τα αποτελέσματα από τις αξιολογήσεις, συμπεριλαμβανομένων των βαθμολογιών ακρίβειας, των αναφορών ταξινόμησης, των πινάκων σύγχυσης, των καμπυλών ROC και των διαγραμμάτων απώλειας λογαρίθμου, αποθηκεύτηκαν για κάθε μέθοδο επιλογής χαρακτηριστικών. Αυτά τα αποτελέσματα αποθηκεύτηκαν σε μορ-

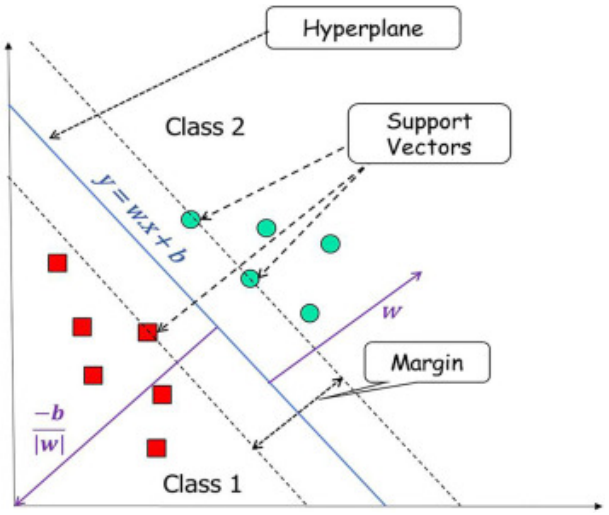
φές JSON και CSV για εύκολη πρόσβαση και ανάλυση.

Η διαδικασία αυτή διασφάλισε ότι το μοντέλο KNN ήταν βελτιστοποιημένο και αξιόπιστο, παρέχοντας ακριβείς προβλέψεις για το σύνολο δεδομένων. Οι λεπτομερείς μετρήσεις αξιολόγησης και οι οπτικοποιήσεις παρείχαν πολύτιμες πληροφορίες για τις επιδόσεις του μοντέλου.

3.9 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine)

Το μοντέλο SVM χρησιμοποιείται και για ταξινόμηση και για παλινδρόμηση και στόχος του είναι να εντοπίσει το καταλληκότερο υπερεπίπεδο για ταξινόμηση με τη βοήθεια των διανυσμάτων υποστήριξης [57]. Το SVM μπορεί να είναι είτε γραμμικό είτε μη γραμμικό και μπορεί να ταξινομηθεί σε hard margin και σε soft margin. Στην πρώτη κατηγορία, τα δεδομένα είναι πλήρως διαχωρίσιμα από ένα υπερεπίπεδο. Το ίδιο δεν ισχύει σε ένα soft margin SVM.

Ξεκινώντας από την απλούστερη περίπτωση του προβλήματος ταξινόμησης δύο γραμμικά διαχωρίσιμων κλάσεων. Στις δύο διαστάσεις η γραμμική συνάρτηση διαχωρισμού είναι διαχωριστική ευθεία, στις τρείς είναι διαχωριστικό επίπεδο και σε περισσότερες από τρείς είναι υπερεπίπεδο. Επομένως, είναι σαφές ότι δεν υπάρχει μοναδική λύση σε αυτό το πρόβλημα ταξινόμησης καθώς ενδεχομένως υπάρχουν πολλές - πιθανώς και άπειρες - γραμμικές συναρτήσεις που μπορούν να διαχωρίσουν τις δύο κλάσεις αυτές. Σε αυτό το σημείο, έρχεται το κριτήριο αξιολόγησης των λύσεων αυτών, το περιθώριο ταξινόμησης (margin) γ μεταξύ των κλάσεων, το οποίο ορίζεται ως η ελάχιστη απόσταση οποιουδήποτε προτύπου από τη διαχωριστική επιφάνεια [55]. Τα πρότυπα των κλάσεων που ικανοποιούν τις ισότητες που αναπαριστούν τις σχέσεις μεταξύ των διανυσμάτων w και της πόλωσης w_0 ονομάζονται διανύσματα υποστήριξης.



Σχήμα 3.13: Η απεικόνιση των στοιχείων ενός SVM [58]

Όταν το πρόβλημα είναι soft margin, το μοντέλο χρησιμοποιεί τις μεταβλητές χαλάρωσης ξ που αντιπροσωπεύουν το πέναλτι των σημείων που βρίσκονται στη λάθος πλευρά του περιθωρίου ταξινόμησης. Οι τιμές των μεταβλητών αυτών αυξάνονται όταν αυξάνεται η απόσταση από το περιθώριο ταξινόμησης [57]. Όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, χρησιμοποιείται το μη-γραμμικό SVM και ο αρχικός χώρος εισόδου αντιστιχίζεται σε έναν χώρο χαρακτηριστικών υψηλότερων διαστάσεων με τη χρήση των συναρτήσεων πυρήνων (kernel function) στο σύνολο εκπαίδευσης.

Για την υλοποίηση του μοντέλου πρόβλεψης με χρήση του Support Vector Machine (SVM), ακολουθήθηκε η ακόλουθη μεθοδολογία, η οποία ενσωματώνει βήματα για την επιλογή χαρακτηριστικών και τη βελτιστοποίηση των υπερπαραμέτρων.

Αρχικά, φορτώθηκαν τα δεδομένα από το αρχείο που περιέχει τις απαραίτητες πληροφορίες για την ανάλυση. Τα δεδομένα καθαρίστηκαν και προετοιμάστηκαν ώστε να περιλαμβάνουν τις μεταβλητές που απαιτούνται για την πρόβλεψη της μεταβλητής-στόχου Death.

Για κάθε μέθοδο επιλογής χαρακτηριστικών, δημιουργήθηκαν διαφορετικά σύνολα δεδομένων. Οι μέθοδοι περιλαμβάνουν τις correlation, Decision Trees, FFS, Fisher Scores, IG, LASSO Regularization, Random Forest, RFE, SFS, και VIF. Τα σύνολα δεδομένων για κάθε μέθοδο φορτώθηκαν από αρχεία pickle.

Η διαδικασία βελτιστοποίησης των υπερπαραμέτρων για το SVM περιλάμβανε τη χρήση του GridSearchCV. Ο GridSearchCV εκτελέστηκε με διασταυρούμενη επικύρωση (cross-validation) 5-πτώσεων, χρησιμοποιώντας ένα σύνολο από προκαθορισμένες υπερπαραμέτρους.

Το GridSearchCV εντοπίζει τον συνδυασμό υπερπαραμέτρων που προσφέρει την καλύτερη απόδοση με βάση την ακρίβεια. Τα καλύτερα μοντέλα για κάθε μέθοδο επιλογής χαρακτηριστικών στη συνέχεια αξιολογήθηκαν με χρήση διασταυρούμενης επικύρωσης 5-πτώσεων.

Η αξιολόγηση των μοντέλων περιλάμβανε τον υπολογισμό της ακρίβειας, της αναφοράς ταξινόμησης (classification report), του πίνακα σύγχυσης (confusion matrix), των καμπυλών ROC και της απώλειας καταγραφής (log loss). Τα αποτελέσματα αποθηκεύτηκαν και οπτικοποιήθηκαν για να διευκολύνουν την ανάλυση και την ερμηνεία τους.

Τέλος, αποθηκεύτηκαν οι καλύτερες υπερπαράμετροι για κάθε μέθοδο, οι πίνακες σύγχυσης, τα δεδομένα των καμπυλών ROC, και οι απώλειες καταγραφής για μελλοντική αναφορά και σύγκριση. Αυτή η μεθοδολογία εξασφαλίζει ότι το μοντέλο SVM είναι βελτιστοποιημένο και αξιολογείται με ακρίβεια, λαμβάνοντας υπόψη διαφορετικές προσεγγίσεις επιλογής χαρακτηριστικών και υπερπαραμέτρων.

3.10 Αφελής Bayes (Naive Bayes)

Το μοντέλο Naive Bayes χρησιμοποιείται για δυαδική ταξινόμηση και πρόκειται για ένα απλό αλλά σημαντικό πιθανοτικό μοντέλο [59]. Το μοντέλο αυτό βασίζεται στον κανόνα του Bayes και χρησιμοποιείται προκειμένου να εξηγήσει τη μέθοδο εκτίμησης της μέγιστης πιθανοφάνειας όταν τα δεδομένα είναι ¨πλήρως παρατηρούμενα¨ είτε είναι ¨μερικώς παρατηρούμενα¨, και ο τύπος είναι ο ακόλουθος:

$$P(C = c_k|X = x) = \frac{P(X = x|C = c_k)P(C = c_k)}{P(X = x)}$$

και θεωρεί ότι όλα τα πιθανά γεγονότα πέφτουν σε μία ακριβώς κλάση [60]. Στη σχέση αυτή, η μεταβλητή C είναι μία τυχαία μεταβλητή της οποίας η τιμές είναι οι κλάσεις, ενώ η X αναπαριστά ένα διάνυσμα για κάθε στοιχείο. Η πιθανότητα $P(C = c_k|X = x$ είναι η υπο συνθήκη πιθανότητα ότι το στοιχείο ανήκει στην κλάση c_k δεδομένου ότι έχει το χαρακτηριστικό διάνυσμα x. Ο κανόνας αυτός υποδεικνύει πώς μπορεί να υπολογιστεί την υποσυνθήκη πιθανότητα ενός στοιχείου να βρίσκεται σε κάποια δεδομένη κλάση μέσω των υπόλοιπων υπο συνθήκη πιθανοτήτων των χαρακτηριστικών διανυσμάτων κάθε κλάσης. Δεδομένου αυτού, μπορούμε να απλοποιήσουμε την παραπάμνω σχέση στην ακόλουθη:

$$P(c_k|x) = P(c_k) \cdot \frac{P(x|c_k)}{P(x)}$$

Θεωρούμε στην αρχή ένα σύνολο εκπαίδευσης που αποτελείται από δεδομένα $(x(i), y(i))$ όπου κάθε $x(i)$ είναι ένα διάνυσμα και κάθε $y(i)$ είναι στο $1, 2, .., k$. Στη συνέχεια, υποθέτουμε ότι κάθε διάνυσμα x ανήκει στο σύνολο $\{-1, +1\}^d$, όπου d είναι ο αριθός των χαρατκηριστικών στο μοντέλο. Το μοντέλο Naive Bayes προκύπτει από τις ακόλουθες υποθέσεις. Υποθέτουμε, καταρχάς τις τυχαίες μεταβλητές Y και $X_1, ..., X_d$ που αντιστοιχούν στην ετικέτα y και στα διανύσματα $x_1, ..., x_d$. Ο στόχος είναι η μοντελοποίηση της κοινής πιθανότητας $P(Y = y, X_1) = x_1, X_2 = x_2, ..., X_d = x_d$ για οποιαδήποτε ετικέτα y που συνδυάζεται με τις τιμές των χψαρακτηριστικών $x_1, .., x_d$. Μία βασική ιδέα στο μοντέλο αυτό είναι η υπόθεση ότι η ακόλουθη σχέση προκύπτει από τις υποθέσεις ανεξαρτησίας: $P(Y = y, X_1 = x_1, ..., X_d = x_d) = P(Y = y) \times \prod P(X_j = x_j|Y = y)$, ότι δηλαδή κάθε τιμή X_j είναι ανεξάρτητη από όλες τις υπόλοιπες τιμές των χαρακτηριστικών όταν του ανατίθεται η ετικέτα του Y. Αξιοποιώντας την υπόθεση αυτή, το μοντέλο έχει δύο τύπους παραμέτρων: $q(y)$ για $y \in \{1, \ldots, k\}$, με $P(Y = y) = q(y)$, και $q_j(x|y)$ για $j \in \{1, \ldots, d\}$, $x \in \{-1, +1\}$, $y \in \{1, \ldots, k\}$, με $P(X_j = x|Y = y) = q_j(x|y)$. Με αυτόν τον τρόπο, το μοντέλο μπορεί να περιγραφεί από τις παραπάνω παραμέτρους και την πιθανότητα $p(y, x_1, ..., x_d) = q(y) \times \prod q_j(x_j|y)$. Επομένως, το μοντέλο αποτελείται από ακέραιους που επισημάνουν το πλήθος των ετικετών και των χαρακτηριστικών μαζί με τις παραμέτρους που περιγράφουν την πιθανότητα κατανομής των ετικετών και των πιθανοτήτων των χαρακτηριστικών που έχουν αντιστοιχηθεί σε

κάθε χαρακτηριστικό μίας δεδομένης ετικέτας. Οι παράμετροι αυτοί εκτιμώνται από τα δεδομένα εκπαίδευσης και το μοντέλο μετά εφαρμόζεται για να ταξινομήσει τα καινούργια δεδομένα ελέγχου.

Για την υλοποίηση του μοντέλου πρόβλεψης χρησιμοποιώντας τον αλγόριθμο Naive Bayes, ακολουθήθηκε μια σειρά από σημαντικά βήματα που στόχευαν στην εξασφάλιση της ακρίβειας και της αξιοπιστίας των προβλέψεων. Αρχικά, το σύνολο δεδομένων φορτώθηκε και προετοιμάστηκε χρησιμοποιώντας τη βιβλιοθήκη pandas. Τα δεδομένα περιλάμβαναν διάφορα χαρακτηριστικά και τη μεταβλητή στόχο Death, η οποία αποτέλεσε το κύριο αντικείμενο πρόβλεψης.

Χρησιμοποιήθηκαν διάφορες μέθοδοι επιλογής χαρακτηριστικών, όπως η συσχέτιση, τα δέντρα αποφάσεων, η επιλογή χαρακτηριστικών προς τα εμπρός (FFS), οι βαθμολογίες Fisher, το πληροφοριακό κέρδος (IG), η κανονικοποίηση LASSO, τα τυχαία δάση (Random Forest), η αναδρομική εξάλειψη χαρακτηριστικών (RFE), η αλληλουχική επιλογή χαρακτηριστικών (SFS) και ο συντελεστής διόγκωσης διακύμανσης (VIF). Για κάθε μέθοδο, δημιουργήθηκε και αποθηκεύτηκε ένα σύνολο δεδομένων με τα επιλεγμένα χαρακτηριστικά.

Κατα την εκπαίδευση του μοντέλου, με τη βοήθεια και της διασταυρούμενης επικύρωσης, επιλέχθηκαν οι βέλτιστες υπερπαραμέτροι για το μοντέλο αυτό για κάθε μέθοδο ξεχωριστά. Η μοναδική υπερπαράμετρος που εξετάστηκε του μοντέλου Naive Bayes είναι η μεταβλητή ομαλοποίησης (var_smoothing) η οποία βοηθάει στις περιπτώσεις που υπάρχει μηδεική προβλεψιμότητα, της οποίας οι υποψήφιες τιμές ήταν: 1e-9, 1e-8, 1e-7, 1e-6, 1e-5.

Για την εκπαίδευση του μοντέλου Naive Bayes, χρησιμοποιήθηκε η διασταυρούμενη επικύρωση Stratified K-Folds. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε με βάση την ακρίβεια, την αναφορά ταξινόμησης και τον πίνακα σύγχυσης. Η ακρίβεια του μοντέλου και η λεπτομερής αναφορά ταξινόμησης που περιλάμβανε την ακρίβεια, την ανάκληση και το F1-score για κάθε κατηγορία αποθηκεύτηκαν για κάθε μέθοδο επιλογής χαρακτηριστικών.

Ο πίνακας σύγχυσης δημιουργήθηκε και αποθηκεύτηκε ως αρχείο CSV για περαιτέρω ανάλυση. Επίσης, σχεδιάστηκε και αποθηκεύτηκε η καμπύλη ROC για να δείξει τη σχέση μεταξύ του ποσοστού αληθινά θετικών και του ποσοστού ψευδώς θετικών σε διάφορες ρυθμίσεις κατωφλίου, καθώς και η περιοχή κάτω από την καμπύλη (AUC).

Για την εκτίμηση της ακρίβειας των πιθανοτήτων, υπολογίστηκε και αποθηκεύτηκε η απώλεια λογαρίθμου (log loss) για κάθε διαίρεση κατά τη διασταυρούμενη επικύρωση. Τα αποτελέσματα από τις αξιολογήσεις, συμπεριλαμβανομένων των βαθμολογιών ακρίβειας, των αναφορών ταξινόμησης, των πινάκων σύγχυσης, των καμπυλών ROC και των διαγραμμάτων απώλειας λογαρίθμου, αποθηκεύτηκαν για κάθε μέθοδο επιλογής χαρακτηριστικών.

Αυτή η διαδικασία διασφάλισε ότι το μοντέλο Naive Bayes ήταν βελτιστοποιημένο και αξιόπιστο, παρέχοντας ακριβείς προβλέψεις για το σύνολο δεδομένων.

3.11 Gradient Boosting

Το Gradient Boosting είναι ένα πολύ χρήσιμο και δυνατό μοντέλο μηχανικής μάθησης που έχει αποκτήσει μεγάλη δημοτικότητα καθώς μπορεί να εφαρμοστεί για την επίλυση διαφόρων προβλημάτων. Αυτή η μέθοδος επιτυγχάνει να ταιριάζει νέα μοντέλα που να προσφέρουν μία πιο ακριβή εκτίμηση μίας παραμέτρου απάντησης. Αυτή η διαδικασία βελτιστοποιεί τέτοιες προβλέψεις του συνόλου για να ταιριάζουν καλύτερα με τα δεδομένα εκπαίδευσης.

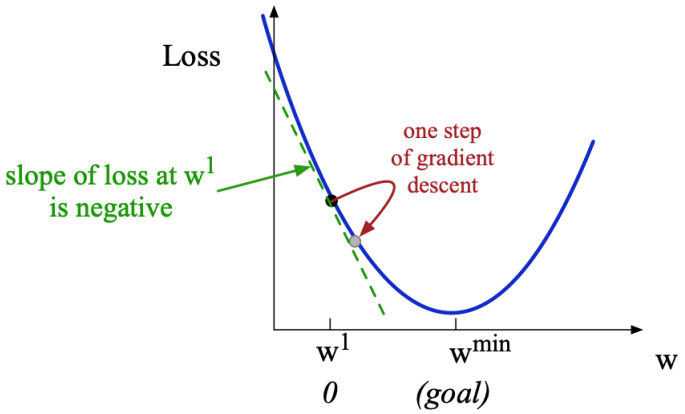
Η ιδέα πίσω από αυτό το μοντέλο συμπεριλαμβάνει μία επαναληπτική προσέγγιση στη μάθηση ενσωμάτωσης (ensemble learning), όπου οι ¨αδύναμοι μαθητές¨ (weak learners) συνδυάζονται για να δημιουργήσουν έναν ïσχυρό μαθητή¨, όπου οι μαθητές που αναφέρονται είναι τα επιμέρους μοντέλα που αξιοποιούνται προκειμένου να διαμορφώσουν ένα ïσχυρό¨ μοντέλο. Αυτά τα μοντέλα που συνδυάζουν άλλα μοντέλα για να αναπαράξουν ένα ισχυρότερο μοντέλο βασίζονται στην απλή συνάθροιση των μοντέλων αυτών κατά την συνένωση τους. Αυτή η τεχνική λεγόμενη boosting ουσιαστικά συνδυάζει τα μοντέλα σειριακά και σε κάθε επανάληψη, ένας νέος αδύναμος, βασικός μαθητής εκπαιδεύεται βάσει του σφάλματος που έχει υπολογιστεί από ολόκληρο το συνόλο που έχει εκπαιδευτεί μέχρι εκείνο το σημείο. Γενικότερα, οι αλγόριθμοι ενίσχυσης (boosting) αποτελούν συνδυαστικές μεθόδους μάθησης που επιτυγχάνουν καλύτερες επιδόσεις με τη δημιουργία μιας σειράς εκτιμητών ή ταξινομητών που εκπαιδεύονται με διαφορετικές κατανομές των δεδομένων εκπαίδευσης [55]. Η τεχνική αυτή συνοδεύεται από τη μέθοδο gradient-descent με αποτέλεσμα να δημιουργούνται έτσι τα gradient boosting machines. Αυτά τα gradient boosting machines ακολουθούν μία διαδικασία εκμάθησης κατα την οποία προσαρμόζει συνεχόμενα νέα μοντέλα με στόχο να παρέχει μία πιο ακριβή πρόβλεψη. Η ιδέα πίσω από αυτόν τον αλγόριθμο είναι να δημιουργεί νέους βασικούς-μαθητές ώστε να είναι συσχετισμένοι κατά το μέγιστο με την αρνητική κλίση της συνάρτησης απώλειας που σχετίζεται με ολόκληρο το σύνολο. Οι συναρτήσεις απώλειας επιλέγονται βάσει την κρίση του χρήστη και συνήθως γίνεται εμπειρικά [61].

Ο στόχος της τεχνικής gradient descent στις περισσότερες περιπτώσεις είναι η εύρεση των βέλτιστων βαρών των μοντέλων ελαχιστοποιώντας τη συνάρτηση κόστους [62] που έχει καθοριστεί για κάθε μοντέλο. Η συναρτήσεις κόστους, ή αλλιώς συναρτήσεις απώλειας, είναι οι συναρτήσεις που υπολογίζουν το σφάλμα, δηλαδή τη διαφορά ανάμεσα στην παραγόμενη από το μοντέλο έξοδο και την επιθυμητή έξοδο που είναι προκαθορισμένη. Επομένως, ο στόχος είναι να βρεθεί ένα σύνολο βαρών που ελαχιστοποιεί τη συνάρτηση απώλειας κατα μέσο όρο για όλα τα παραδείγματα :

$$\theta^* = \arg\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} L_{CE}(f(x^{(i)}; \theta), y^{(i)})$$

Ο τρόπος με τον οποίο η μέθοδος gradient descent υπολογίζει τα βέλιστα βάρη αυτά είναι μέσω της εύρεσης του ελάχιστου μίας συνάρτησης υπολογίζοντας την

κατεύθυνση της κλίσης της συνάρτησης που αυξάνεται πιο απότομα και επιλέγοντας να κινηθεί προς την αντίθετη κατεύθυνση [62].



Σχήμα 3.14: Το πρώτο βήμα στην επαναληπτική εύρεση του ελαχίστου αυτής της συνάρτησης απωλειών είναι η μετακίνηση του w προς την αντίθετη κατεύθυνση από την κλίση της συνάρτησης (αν η κλίση είναι αρνητική, πρέπει να μετακινήσουμε το w προς τη θετική κατεύθυνση)

Η ποσότητα με την οποία θα κινείται προς οποιαδήποτε κατεύθυνση καθορίζεται από την κλίση $\frac{d}{dw}L(\hat{y}, y)$ που πολλαπλασιάζεται με τον ρυθμό εκμάθησης (learning rate) η. Όσο μεγαλύτερος είναι ο ρυθμός αυτός, τόσο πιο πολύ θα μετακινείται το w ανά βήμα. Αν θεωρήσουμε ότι η συνάρτηση απώλειας συμβολίζεται ως $L(\hat{y}, y)$, τότε, η αλλαγή που πραγματοποιεί σε κάθε επανάληψη ο αλγόριθμος είναι η κλίση επί τον ρυθμό μάθησης, όπως φαίνεται παρακάτω:

$$w^{t+1} = w^t - \eta \frac{d}{dw} L(\hat{y}, y)$$

Μία παραλλαγή και επέκταση του μοντέλου αυτού αποτελεί το *Stochastic Gradient Boosting* του οποίου η διαφορά είναι ότι υπολογίζεται η κλίση (gradient) αξιοποιώντας ένα τυχαίο σύνολο από ολόκληρο το σύνολο παρατηρήσεων. Σε πολλές περιπτώσεις, μελέτες ισχυρίζονται ότι η προσθήκη της τυχαιότητας στη διαδικασία εκπαίδευσης του μοντέλου του gradient boosting μπορεί να βελτιώσει την ακρίβεια και την ταχύτητα του μοντέλου αυτού [63]. Με αυτόν τον τρόπο, ένα υποσύνολο από τα δεδομένα εκπαίδευσης επιλέγεται τυχαία, και αυτό το δείγμα χρησιμοποιείται αντί για ολόκληρο το σύνολο των δεδομένων προκειμένου να προσαρμοστεί στον base learner και να ανανεώσει το μοντέλο στην συγκεκριμένη επανάληψη [63]. Πιο αναλυτικά, στην περίπτωση του stochastic gradient descent ο αλγόριθμος ελαχιστοποιεί την συνάρτηση απώλειας βάσει ενός υποσυνόλου των δεδομένων αντί να τροποποιεί τα βάρη κατόπιν επεξεργασίας ολόκληρου του συνόλου δεδομένων [62].

Για την εφαρμογή του μοντέλου πρόβλεψης, χρησιμοποιήθηκε ο αλγόριθμος ενίσχυσης βαθμίδων (Gradient Boosting), ο οποίος απαιτούσε την υλοποίηση μιας σειράς σημαντικών βημάτων.

Το πρώτο βήμα αφορούσε τη φόρτωση του καθαρού συνόλου δεδομένων

από το καθορισμένο μονοπάτι χρησιμοποιώντας τη βιβλιοθήκη pandas. Το σύνολο δεδομένων περιλάμβανε διάφορα χαρακτηριστικά και τη μεταβλητή στόχο Death, που ήταν το επίκεντρο της πρόβλεψης.

Ακολούθησαν διάφορες μέθοδοι επιλογής χαρακτηριστικών για να εντοπιστούν τα πιο σχετικά χαρακτηριστικά για το μοντέλο πρόβλεψης. Αυτές οι μέθοδοι περιλάμβαναν τη συσχέτιση, τα δέντρα αποφάσεων, την επιλογή χαρακτηριστικών προς τα εμπρός (FFS), τις βαθμολογίες Fisher, το πληροφοριακό κέρδος (IG), την κανονικοποίηση LASSO, τα τυχαία δάση, την αναδρομική εξάλειψη χαρακτηριστικών (RFE), την αλληλουχική επιλογή χαρακτηριστικών (SFS), και τον συντελεστή διόγκωσης διακύμανσης (VIF). Για κάθε μέθοδο, δημιουργήθηκε και αποθηκεύτηκε ένα σύνολο δεδομένων που περιείχε τα επιλεγμένα χαρακτηριστικά για περαιτέρω ανάλυση.

Στη συνέχεια, καθορίστηκε ένα σύνολο παραμέτρων για δοκιμή, όπως ο αριθμός των εκτιμητών (n_estimators), ο ρυθμός μάθησης (learning_rate), το μέγιστο βάθος (max_depth), το υποδείγμα (subsample), ο ελάχιστος αριθμός δειγμάτων για διαίρεση (min_samples_split), και ο ελάχιστος αριθμός δειγμάτων σε φύλλο (min_samples_leaf).

Αν επιλεγόταν η εύρεση των βέλτιστων υπερπαραμέτρων, πραγματοποιήθηκε αναζήτηση πλέγματος (GridSearchCV) χρησιμοποιώντας Stratified K-Folds διασταυρούμενη επικύρωση με 5 διαίρεση για να εξασφαλιστεί η αξιοπιστία των αποτελεσμάτων. Ο καλύτερος μοντέλο που προέκυψε από την αναζήτηση πλέγματος αξιολογήθηκε χρησιμοποιώντας διάφορες μετρικές.

Η αξιολόγηση του μοντέλου περιλάμβανε την ακρίβεια του μοντέλου, η οποία υπολογίστηκε και αποθηκεύτηκε μαζί με μια λεπτομερή αναφορά ταξινόμησης που περιλάμβανε την ακρίβεια, την ανάκληση και το F1-score για κάθε κατηγορία. Δημιουργήθηκε επίσης ένας πίνακας σύγχυσης για να απεικονιστούν οι επιδόσεις του μοντέλου σε πραγματικά θετικά, πραγματικά αρνητικά, ψευδώς θετικά και ψευδώς αρνητικά αποτελέσματα. Ο πίνακας αυτός αποθηκεύτηκε ως αρχείο CSV για περαιτέρω ανάλυση.

Επιπλέον, σχεδιάστηκε και αποθηκεύτηκε η καμπύλη ROC για να δείξει τη σχέση μεταξύ του ποσοστού αληθινά θετικών και του ποσοστού ψευδώς θετικών σε διάφορες ρυθμίσεις κατωφλίου. Υπολογίστηκε επίσης η περιοχή κάτω από την καμπύλη (AUC). Για την εκτίμηση της ακρίβειας των πιθανοτήτων, υπολογίστηκε και αποθηκεύτηκε η απώλεια λογαρίθμου (log loss) για κάθε διαίρεση κατά τη διασταυρούμενη επικύρωση.

Τα αποτελέσματα από τις αξιολογήσεις, συμπεριλαμβανομένων των βαθμολογιών ακρίβειας, των αναφορών ταξινόμησης, των πινάκων σύγχυσης, των καμπυλών ROC και των διαγραμμάτων απώλειας λογαρίθμου, αποθηκεύτηκαν για κάθε μέθοδο επιλογής χαρακτηριστικών. Αυτά τα αποτελέσματα αποθηκεύτηκαν σε μορφές JSON και CSV για εύκολη πρόσβαση και ανάλυση.

Αυτή η διαδικασία εξασφάλισε μια ολοκληρωμένη αξιολόγηση των επιδόσε-

ων του μοντέλου Gradient Boosting σε διαφορετικές μεθόδους επιλογής χαρακτηριστικών, παρέχοντας ένα αξιόπιστο πλαίσιο για ακριβείς προβλέψεις. Η χρήση διασταυρούμενης επικύρωσης και εκτεταμένης βελτιστοποίησης παραμέτρων βοήθησε στον εντοπισμό της βέλτιστης διαμόρφωσης του μοντέλου, ενώ οι λεπτομερείς μετρήσεις αξιολόγησης και οι οπτικοποιήσεις παρείχαν βαθιές πληροφορίες για τις επιδόσεις του μοντέλου.

Στην έρευνα αυτή, αξιοποιήθηκε και μία παραλλαγή του Gradient Boosting, το Stochastic Gradient Boosting. Το Stochastic Gradient Boosting είναι μια τεχνική μηχανικής μάθησης που κατασκευάζει προσθετικά μοντέλα παλινδρόμησης προσαρμόζοντας διαδοχικά μια απλή παραμετρική συνάρτηση, γνωστή ως βασικός μαθητής, στα τρέχοντα ¨ψευδο¨-υπολείμματα χρησιμοποιώντας τη μέθοδο των ελαχίστων τετραγώνων σε κάθε επανάληψη. Τα ψευδο-υπολείμματα αντιπροσωπεύουν την κλίση της συναρτησιακής απώλειας που ελαχιστοποιείται, και αξιολογούνται σε κάθε σημείο εκπαίδευσης στο τρέχον βήμα. Η τυχαιοποίηση ενσωματώνεται στη διαδικασία μέσω της επιλογής ενός τυχαίου υποσυνόλου των δεδομένων εκπαίδευσης (χωρίς αντικατάσταση) σε κάθε επανάληψη, αντί της χρήσης του πλήρους συνόλου δεδομένων [63]. Αυτή η προσέγγιση όχι μόνο βελτιώνει την ακρίβεια της προσέγγισης και την ταχύτητα εκτέλεσης του Gradient Boosting, αλλά αυξάνει επίσης την ανθεκτικότητα του μοντέλου απέναντι στην υπερεκτίμηση της ικανότητας του βασικού μαθητή, καθιστώντας το πιο ανθεκτικό σε υπερβολική προσαρμογή.

Η υλοποίηση του μοντέλου πρόβλεψης Stochastic Gradient Boosting πραγματοποιήθηκε με τη χρήση του αλγορίθμου HistGradientBoosting και περιλάμβανε μια σειρά από βήματα για τη διασφάλιση της ακρίβειας και της αξιοπιστίας των προβλέψεων. Αρχικά, το σύνολο δεδομένων φορτώθηκε και προετοιμάστηκε, με τη μεταβλητή-στόχο να είναι η Death, ενώ τα υπόλοιπα χαρακτηριστικά χρησιμοποιήθηκαν ως ανεξάρτητες μεταβλητές.

Η διαδικασία επιλογής χαρακτηριστικών περιλάμβανε τη χρήση διαφόρων μεθόδων, όπως correlation, Decision Trees, FFS, Fisher Scores, IG, LASSO Regularization, Random Forest, RFE, SFS και VIF. Για κάθε μέθοδο επιλογής χαρακτηριστικών, το αντίστοιχο σύνολο δεδομένων φορτώθηκε και προετοιμάστηκε.

Στη συνέχεια, πραγματοποιήθηκε αναζήτηση βέλτιστων υπερπαραμέτρων (GridSearchCV) για το μοντέλο HistGradientBoosting, χρησιμοποιώντας τον StratifiedKFold για διασταυρούμενη επικύρωση. Οι υπερπαράμετροι που εξετάστηκαν περιλάμβαναν την παράμετρο max_iter με τιμές από 100 έως 300, την παράμετρο learning_rate με τιμές από 0.01 έως 0.5, το max_depth με τιμές από 3 έως 10 και την παράμετρο min_samples_leaf με τιμές από 1 έως 4.

Για κάθε σύνολο δεδομένων που προέκυψε από τη διαδικασία επιλογής χαρακτηριστικών, η αναζήτηση υπερπαραμέτρων προσδιόρισε τις βέλτιστες παραμέτρους για το HistGradientBoosting. Αυτές οι παράμετροι στη συνέχεια χρησιμοποιήθηκαν για την εκπαίδευση και αξιολόγηση του μοντέλου. Η αξιολόγηση των μοντέλων πραγματοποιήθηκε χρησιμοποιώντας διασταυρούμενη επικύρωση (cross-

validation) και την μέθοδο cross_val_predict για την πρόβλεψη των τιμών της μεταβλητής στόχου.

Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση περιλάμβαναν την ακρίβεια (accuracy), την αναφορά ταξινόμησης (classification report), τον πίνακα σύγχυσης (confusion matrix), τις καμπύλες ROC και τις τιμές log loss. Η καμπύλη ROC παρείχε μια γραφική απεικόνιση της απόδοσης του μοντέλου, ενώ οι τιμές log loss παρείχαν μια ποσοτική μέτρηση της ακρίβειας των προβλέψεων του μοντέλου.

Η διαδικασία αξιολόγησης περιλάμβανε τη δημιουργία και αποθήκευση πινάκων σύγχυσης και καμπυλών ROC, καθώς και τον υπολογισμό και την αποθήκευση των τιμών log loss για κάθε αναδίπλωση της διασταυρούμενης επικύρωσης. Τα αποτελέσματα της αξιολόγησης, συμπεριλαμβανομένων των καλύτερων υπερπαραμέτρων, της ακρίβειας και των αναφορών ταξινόμησης, καταγράφηκαν και αποθηκεύτηκαν για κάθε μέθοδο επιλογής χαρακτηριστικών.

Συνολικά, η χρήση του HistGradientBoosting σε συνδυασμό με τη βελτιστοποίηση των υπερπαραμέτρων και την προσεκτική επιλογή χαρακτηριστικών παρείχε ένα ισχυρό πλαίσιο για την ανάπτυξη ακριβών και αξιόπιστων μοντέλων πρόβλεψης, επιτρέποντας τη βέλτιστη αξιοποίηση των δεδομένων και την επίτευξη υψηλής ακρίβειας στις προβλέψεις.

3.12 XGBoost και Δέντρα Συγκροτημάτων

Μία επέκταση του μοντέλου Gradient Boosting είναι το XGBoost. Οι μέθοδοι μηχανικής μάθησης και οι προσεγγίσεις που βασίζονται σε δεδομένα γίνονται ολοένα και πιο σημαντικές σε πολλούς τομείς. Υπάρχουν δύο σημαντικοί παράγοντες που οδηγούν αυτές τις επιτυχημένες εφαρμογές: η χρήση αποτελεσματικών (στατιστικών) μοντέλων που συλλαμβάνουν τις περίπλοκες εξαρτήσεις δεδομένων και τα συστήματα εκμάθησης που είναι σε θέση να μάθουν το μοντέλο από μεγάλα σύνολα δεδομένων. Μεταξύ των μεθόδων μηχανικής μάθησης που χρησιμοποιούνται στην πράξη, η ενίσχυση βαθμιδωτών δέντρων (gradient tree boosting) ξεχωρίζει σε πολλές εφαρμογές. Η XGBoost είναι ένα κλιμακούμενο σύστημα μηχανικής μάθησης για ενίσχυση δέντρων, που έχει αποδειχθεί εξαιρετικά αποτελεσματικό σε ένα ευρύ φάσμα προβλημάτων [63].

Η XGBoost (Extreme Gradient Boosting) είναι ένας εξαιρετικά αποδοτικός αλγόριθμος ενίσχυσης βαθμιδωτών δέντρων, ο οποίος χρησιμοποιείται ευρέως για την επίτευξη κορυφαίων αποτελεσμάτων σε προβλήματα ταξινόμησης και παλινδρόμησης. Χρησιμοποιεί μια σύγκροτη μέθοδο που συνδυάζει πολλά απλά μοντέλα (trees) για να κατασκευάσει ένα ισχυρότερο συνολικό μοντέλο.

Η ενίσχυση βαθμιδωτών δέντρων κατασκευάζει προσθετικά μοντέλα παλινδρόμησης προσαρμόζοντας διαδοχικά μια απλή παραμετρική συνάρτηση (base learner) στα τρέχοντα ¨ψευδο"-υπολείμματα χρησιμοποιώντας την ελάχιστη τετραγωνική απόκλιση σε κάθε επανάληψη. Τα ψευδο-υπολείμματα είναι οι βαθμίδες

της συνάρτησης απώλειας που ελαχιστοποιούνται, ως προς τις τιμές του μοντέλου σε κάθε σημείο των δεδομένων εκπαίδευσης που αξιολογούνται στο τρέχον βήμα. Η ενσωμάτωση τυχαιοποίησης στη διαδικασία βελτιώνει την ακρίβεια και την ταχύτητα εκτέλεσης του αλγορίθμου. Συγκεκριμένα, σε κάθε επανάληψη, ένα υποσύνολο των δεδομένων εκπαίδευσης επιλέγεται τυχαία και χρησιμοποιείται για την προσαρμογή της συνάρτησης βάσης και την ενημέρωση του μοντέλου. Αυτή η προσέγγιση αυξάνει επίσης την ανθεκτικότητα κατά της υπερεφαρμογής [34].

Το συγκρότημα δέντρων (Tree Ensemble Model) χρησιμοποιεί K προσθετικές συναρτήσεις για να προβλέψει την έξοδο. Το μοντέλο εκπαιδεύεται με την ελαχιστοποίηση ενός τακτοποιημένου στόχου, ο οποίος περιλαμβάνει έναν όρο απώλειας που μετρά τη διαφορά μεταξύ των προβλέψεων και των πραγματικών τιμών και έναν όρο που τιμωρεί την πολυπλοκότητα του μοντέλου για την αποφυγή υπερβολικής προσαρμογής. Συγκεκριμένα, χρησιμοποιούνται συναρτήσεις δέντρων παλινδρόμησης (regression trees) όπου κάθε δέντρο περιέχει ένα συνεχές σκορ σε κάθε φύλλο, το οποίο προστίθεται για την τελική πρόβλεψη [34].

Η εξίσωση που χρησιμοποιείται για την πρόβλεψη είναι η ακόλουθη:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F} \tag{3.1}$$

Όπου η $f(x)$ είναι η συνάρτηση παλινδρόμησης για κάθε στιγμιότυπο, T είναι ο αριθμός των φύλλων και κάθε f_k αντιστοιχεί σε μία ανεξάρτητη δομή δέντρου που περιέχει ένα σκορ για κάθε κόμβο. Για την τελική πρόβλεψη, υπολογίζεται το άθροισμα των πόντων κάθε κόμβου ως ακολούθως:

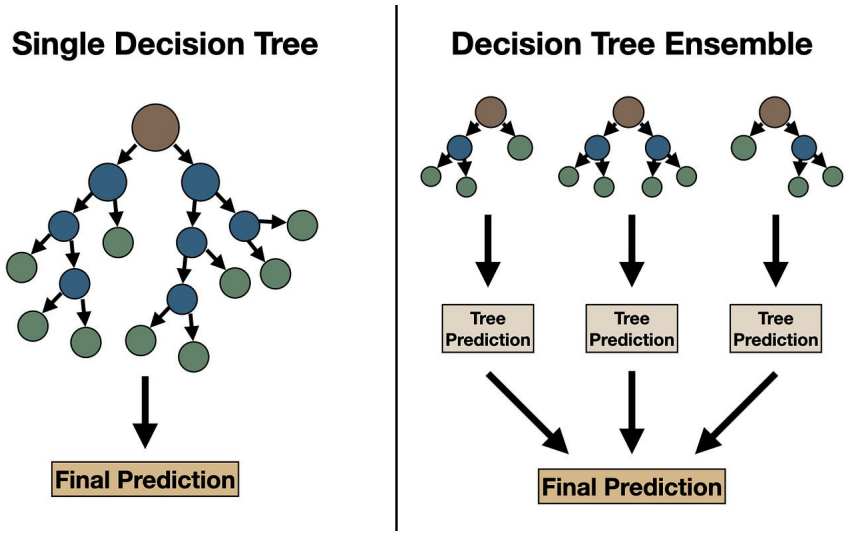
$$L(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k) \tag{3.2}$$

Όπου ο όρος Ω είναι μια συνάρτηση τακτοποίησης που τιμωρεί την πολυπλοκότητα του μοντέλου, και ο όρος l είναι μια διαφορίσιμη κυρτή συνάρτηση απώλειας που μετρά τη διαφορά μεταξύ των προβλέψεων και των πραγματικών τιμών. Ο τακτοποιημένος στόχος επιδιώκει να επιλέξει ένα μοντέλο που χρησιμοποιεί απλές και προβλεπτικές συναρτήσεις, αποφεύγοντας την υπερβολική προσαρμογή. Ο στόχος τακτοποιείται με έναν πρόσθετο όρο που βοηθά στην εξομάλυνση των τελικών βαρών.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{3.3}$$

Αυτός ο πρόσθετος όρος τακτοποίησης επιλέγει ένα μοντέλο που χρησιμοποιεί απλές και προβλεπτικές συναρτήσεις, αποφεύγοντας την υπερβολική προσαρμογή.

Η XGBoost είναι εξαιρετικά κλιμακούμενη και μπορεί να επεξεργαστεί εκατομμύρια παραδείγματα γρήγορα, χρησιμοποιώντας καινοτόμες βελτιστοποιήσεις όπως παράλληλη και κατανεμημένη επεξεργασία, καινοτόμο αλγόριθμο εκμάθησης



Σχήμα 3.15: Δομή tree ensemble: το τελικό αποτέλεσμα πρόβλεψης προκύπτει από το άθροισμα των προβλέψεων από κάθε δέντρο [64]

δέντρων για την αντιμετώπιση αραιών δεδομένων, και δομή μπλοκ που εκμεταλλεύεται την προσωρινή μνήμη για την εκμάθηση δέντρων εκτός πυρήνα [34].

Η XGBoost επιτρέπει σε επιστήμονες δεδομένων και ερευνητές να κατασκευάζουν ισχυρές παραλλαγές αλγορίθμων ενίσχυσης δέντρων και έχει χρησιμοποιηθεί επιτυχώς σε πολλούς διαγωνισμούς και πραγματικές εφαρμογές, αποδεικνύοντας την αποτελεσματικότητα και την αξιοπιστία της.

Η υλοποίηση του XGBoost για προγνωστικό μοντελοποίησης περιλαμβάνει μια δομημένη μεθοδολογία με στόχο τη βελτιστοποίηση της απόδοσης του μοντέλου και τη διασφάλιση της αξιοπιστίας των προβλέψεων. Η διαδικασία ξεκινά με τη φόρτωση και την αρχική ανάλυση του συνόλου δεδομένων, το οποίο προήλθε από ένα αρχείο CSV με καθαρισμένα δεδομένα. Η κατανόηση της δομής και των χαρακτηριστικών του συνόλου δεδομένων είναι κρίσιμη καθώς αποτελεί τη βάση για τα επόμενα βήματα μοντελοποίησης.

Για τη βελτιστοποίηση της προγνωστικής ικανότητας του αλγορίθμου XGBoost, εφαρμόστηκαν διάφορες μέθοδοι επιλογής χαρακτηριστικών. Αυτές οι μέθοδοι, όπως ανάλυση συσχέτισης, δέντρα απόφασης, επιλογή χαρακτηριστικών με βάση προηγούμενα δεδομένα, βαθμολογίες Fisher, κέρδος πληροφορίας, κανονικοποίηση LASSO, επιλογή με βάση δέντρα τυχαίου δάσους, αναδρομική επιλογή χαρακτηριστικών (RFE), συνεχόμενη επιλογή χαρακτηριστικών (SFS)| και παράγοντας φλερτότητας διαστρωμάτωσης (VIF), εφαρμόστηκαν συστηματικά στο σύνολο δεδομένων. Κάθε μέθοδος περιελάμβανε τη φόρτωση προεπεξεργασμένων συνόλων δεδομένων που είχαν προσαρμοστεί ειδικά για τη συγκεκριμένη τεχνική επιλογής χαρακτηριστικών, τα οποία ανήκαν σε σειριοποιημένα αρχεία.

Για κάθε επιλεγμένο σύνολο δεδομένων, ο μοντέλο XGBoost προετοιμάστηκε και ρυθμίστηκε με χρήση αναζήτησης πλέγματος διασταυρούμενης επικύρωσης. Αυτή η προσέγγιση ελέγχει συστηματικά ένα προκαθορισμένο πλέγμα υπερπαρα-

μέτρων, συμπεριλαμβανομένου του αριθμού των εκτιμητών, του ρυθμού μάθησης, του μέγιστου βάθους δέντρου, του λόγου υποδειγματοληψίας και του ελάχιστου βάρους παιδιού. Η τεχνική διασταυρούμενης επικύρωσης stratified k-fold με πέντε διαιρέσεις χρησιμοποιήθηκε για να εξασφαλιστεί η αξιόπιστη αξιολόγηση σε διαφορετικά υποσύνολα των δεδομένων. Αυτή η διαδικασία βοηθά στην εντοπισμό του βέλτιστου συνδυασμού υπερπαραμέτρων που μεγιστοποιεί την απόδοση του μοντέλου, όπως αξιολογείται από μετρικές όπως η ακρίβεια, η ακρίβεια, η ανάκληση και η περιοχή κάτω από την καμπύλη ROC (AUC).

Αφού εντοπίστηκαν οι βέλτιστες υπερπαράμετροι για κάθε μέθοδο επιλογής χαρακτηριστικών, τα τελικά μοντέλα XGBoost εκπαιδεύτηκαν χρησιμοποιώντας ολόκληρο το σύνολο δεδομένων. Τα μοντέλα αξιολογήθηκαν εκτενώς χρησιμοποιώντας διάφορες μετρικές απόδοσης, συμπεριλαμβανομένων ακρίβειας, πινάκων σύγχυσης, καμπύλες ROC και τιμές log loss. Αυτές οι αξιολογήσεις παρέχουν εισηγήσεις σχετικά με τη δυνατότητα του μοντέλου να ταξινομήσει σωστά τις περιπτώσεις και τη διακριτική του ικανότητα σε διαφορετικά κατώφλια.

Επιπλέον, τα αποτελέσματα από κάθε επανάληψη του μοντέλου αποθηκεύτηκαν συστηματικά σε δομημένες μορφές.

Συνοψίζοντας, η συστηματική εφαρμογή του XGBoost με αυστηρή ρύθμιση υπερπαραμέτρων και μεθοδολογίες επιλογής χαρακτηριστικών διασφαλίζει την ανάπτυξη αξιόπιστων προγνωστικών μοντέλων. Αυτή η προσέγγιση όχι μόνο ενισχύει την ακρίβεια του μοντέλου αλλά παρέχει και εισηγήσεις σχετικά με τη σημασία των χαρακτηριστικών και την ερμηνευσιμότητα του μοντέλου, ουσιώδη για τη λήψη ενημερωμένων αποφάσεων σε διάφορες εργασίες προγνωστικής μοντελοποίησης.

3.13 Principal Component Regression

Ένα μοντέλο παρόμοιας φύσεως της γραμμικής παλινδρόμησης είναι το Principal Component Regression και διαχειρίζεται περιπτώσεις που ο αριθμός των χαρακτηριστικών (ή συνιστωσών) είναι υψηλός ανάλογα με τον αριθμό των παρατηρήσεων. Στο PCR, τα στοιχεία θεωρείται να έχουν χαμηλής διάστασης αναπαράστασης, ακόμα και εάν υπάρχουν σε χώρο υψηλών διαστάσεων. Η μέθοδος αυτή πρώτα εφαρμόζει τη μέθοδο ανάλυση κυρίων συνιστωσών (Principal Component Analysis, (PCA)) στα παρατηρούμενα χαρακτηριστικά με στόχο να μειώσει τις διαστάσεις τους [63].

Η μέθοδος PCA βασίζεται στην ιδέα ότι σε πολλά συστήματα με, έστω, n τυχαίες μεταβλητές, οι βαθμοί ελευθερίας m είναι λιγότεροι. Με αυτόν τον τρόπο, ενώ οι διαστάσεις του διανύσματος παρατήρησης είναι n, το σύστημα μπορεί να εκφραστεί από m ασυσχέτιστες αλλά κρυφές τυχαίες μεταβλητές [55]. Οι m μεταβλητές αυτές ονομάζονται παράγοντες ή χαρακτηριστικά του διανύσματος παρατήρησης και ο αριθμός αυτός είναι η ουσιαστική διάσταση του τυχαίου διανύσματος παρατήρησης, ενώ ο αριθμός n ονομάζεται επιφανειακή διάσταση του διανύσματος. Αν,

βέβαια, οι δύο διαστάσεις αυτές είναι ίδιες, τότε οι n μεταβλητές είναι ασυσχέτιστες, και όσο πιο ασυσχέτιστες είναι, τόσο λιγότερα χαρακτηριστικά απαιτούνται προκειμένου να περγράψουν τις τυχαίες παρατηρήσεις. Επομένως, το PCA προβάλλει κάθε συνιστώσα για να βρεί μία αναπαράσταση του συστήματος σε μικρότερη διάσταση. Στη συνέχεια, το μοντέλο PCR εφαρμόζει γραμμική παλινδρόμηση αξιοποιώντας το μειωμένο σε διάσταση σύνολο συνιστωσών [63].

Η εφαρμογή του μοντέλου πρόβλεψης με χρήση της Principal Component Analysis (PCA) και διάφορων αλγορίθμων μηχανικής μάθησης περιλάμβανε μια σειρά από βήματα για την εξασφάλιση της βέλτιστης απόδοσης και αξιοπιστίας.

Αρχικά, το σύνολο δεδομένων φορτώθηκε και προετοιμάστηκε με τη βιβλιοθήκη pandas. Η μεταβλητή-στόχος ήταν η Death, ενώ τα υπόλοιπα χαρακτηριστικά χρησιμοποιήθηκαν ως μεταβλητές εισόδου. Η τυποποίηση των δεδομένων ήταν απαραίτητη για να διασφαλιστεί ότι όλα τα χαρακτηριστικά είχαν την ίδια κλίμακα, χρησιμοποιώντας τον StandardScaler.

Στη συνέχεια, εφαρμόστηκε η μέθοδος PCA για τη μείωση της διαστατικότητας των δεδομένων, διατηρώντας το 95% της συνολικής διακύμανσης. Αυτή η μείωση της διαστατικότητας όχι μόνο βελτίωσε την απόδοση των μοντέλων αλλά και επιτάχυνε τη διαδικασία εκπαίδευσης και πρόβλεψης. Για τη διαδικασία όμως εκπαίδευσης και πρόβλεψης, δεν χρησιμοποιήθηκε μόνο η κλασική μορφή του Principal Component Regression που, ουσιαστικά, περιλαμβάνει τη μέθοδο μείωσης διαστάσεων PCA που αναφέρθηκε προηγουμένως, σε συνδυασμό με τη Γραμμική παλινδρόμηση, αλλά, δοκιμάστηκε σε συνδυασμό με τα παραπάνω μοντέλα, όπως, τα δέντρα αποφάσεων, random forests, gradient boosting, k-nn, logistic regression, stochastic gradient descent, xgboost, support vector machine και ο αφελής bayes.

Τα δεδομένα χωρίστηκαν σε σύνολα εκπαίδευσης και δοκιμής με αναλογία 80/20 χρησιμοποιώντας τη μέθοδο `train_test_split`. Αυτό εξασφάλισε ότι το μοντέλο είχε αρκετά δεδομένα για να μάθει και να γενικεύσει τις προβλέψεις του.

Για κάθε μοντέλο, πραγματοποιήθηκε αναζήτηση βέλτιστων υπερπαραμέτρων (GridSearchCV). Αυτή η διαδικασία περιλάμβανε την αναζήτηση στον χώρο των υπερπαραμέτρων για κάθε αλγόριθμο, προκειμένου να βρεθούν οι συνδυασμοί που μεγιστοποιούν την απόδοση του μοντέλου. Για παράδειγμα, για τον αλγόριθμο Gradient Boosting, οι υπερπαράμετροι που εξετάστηκαν περιλάμβαναν τον αριθμό των δέντρων (`n_estimators`), τον ρυθμό μάθησης (`learning_rate`), το μέγιστο βάθος των δέντρων (`max_depth`), το ποσοστό των δειγμάτων που χρησιμοποιήθηκαν για την εκπαίδευση κάθε δέντρου (subsample), καθώς και το ελάχιστο αριθμό δειγμάτων που απαιτούνται για να διαχωριστεί ένας κόμβος (`min_samples_split`) και για να είναι φύλλο (`min_samples_leaf`).

Για κάθε αλγόριθμο, όπως Decision Trees, K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, Stochastic Gradient Descent (SGD), XGBoost, Support Vector Machine (SVM), Linear Regression και Naive Bayes, οι υπερπαράμετροι προσαρμόστηκαν για να επιτευχθεί η καλύτερη δυνατή απόδοση.

Για παράδειγμα, στην περίπτωση του SVM, οι υπερπαράμετροι περιλάμβαναν την παράμετρο κανονικοποίησης (C), τον τύπο πυρήνα (kernel) και τη γάμμα (gamma).

Η αξιολόγηση των μοντέλων πραγματοποιήθηκε χρησιμοποιώντας τη μέθοδο της διασταυρούμενης επικύρωσης (cross-validation) με Stratified K-Folds και 5 διαιρέσεις. Αυτή η μέθοδος εξασφάλισε ότι τα δεδομένα εκπαιδεύτηκαν και αξιολογήθηκαν σε διάφορες διαιρέσεις, αποφεύγοντας τυχόν μεροληψίες και εξασφαλίζοντας πιο σταθερά και αξιόπιστα αποτελέσματα.

Η ακρίβεια των μοντέλων, οι αναφορές ταξινόμησης, οι πίνακες σύγχυσης και οι καμπύλες ROC υπολογίστηκαν και αναλύθηκαν για να αξιολογηθεί η απόδοση κάθε μοντέλου. Τα αποτελέσματα έδειξαν ότι η χρήση της PCA σε συνδυασμό με τη βελτιστοποίηση των υπερπαραμέτρων και τη διασταυρούμενη επικύρωση παρείχε ένα σταθερό πλαίσιο για την ανάπτυξη ακριβών και αξιόπιστων μοντέλων πρόβλεψης.

3.14 Penalized Logistic Regression

Σε πολλά προβλήματα ταξινόμησης, τα δεδομένα είναι υψηλών διαστάσεων που δημιουργούν δυσκολίες κατά τη στατιστική ανάλυση, με αποτέλεσμα πολλές κλασικές στατιστικές μέθοδοι, όπως το logistic regression, η οποία είναι μία μέθοδος κανονικοποίησης των μοντέλων γραμμικής παλινδρόμησης που μειώνει τα σφάλματα που προκύπτουν από την υπερπροσαρμογή του μοντέλου, να μην είναι τόσο αποτελεσματικές στη διαχείριση τους. Σε αυτό το σημείο, έχουν διεξαχθεί πολλαπλές έρευνες με στόχο την μείωση των διαστάσεων, και ένας από τους τρόπους που επιτυχγάνεται αυτό είναι με τη μείωση των δεδομμένων. Προκειμένου να επιτευχθεί αυτό, αξιοποιούνται πολλές ποινοικοποιημένες μέθοδοι (penalized methods) [62].

Η καλή κατανόηση του μοντέλου αυτού είναι αξιοσημείωτο να διασαφηνιστεί ο τρόπος λειτουργίας του απλού μοντέλου logistic regression, το οποίο είναι ένα στατιστικό μοντέλο του οποίου η συνάρτηση έχει μία μη γραμμική σχέση με τον γραμμικό συνδυασμό των μεταβλητών. Επειδή χρησιμοποιείται σε προβλήματα δυαδικής ταξινόμησης, η μεταβλητή απάντησης μπορεί να έχει είτε την τιμή 0 είτε την τιμή 1 [65]. Αρχικά, το logistic regression για κάθε δεδομένο εισόδου παράγει μία ετικέτα $\hat{y}$, μία εκτίμηση της πραγματικής ετικέτας y [65]. Ο αλγόριθμος υπολογίζει την απόσταση μεταξύ της παραγόμενης εξόδου με την επιθυμητή έξοδο, η οποία ονομάζεται συνάρτηση απώλειας ή συνάρτηση κόστους. Στη συνέχεια, απαιτείται ένας αλγόριθμος βελτιστοποίησης ο οποίος σε κάθε επανάληψη ανανεώνει τα βάρη προκειμένου να ελαχιστοποιήσει την συνάρτηση κόστους, και η πιο δημοφιλής συνάρτηση είναι η cross-entropy loss. Όπως έχουμε ήδη αναφέρει παραπάνω, ο πιο κλασικός αλγόριθμος που πραγματοποιεί αυτή τη διαδικασία είναι ο gradient descent. Στην περίπτωση του logistic regression η συνάρτηση απώλειας είναι convex. Το βασικό χαρακτηρσιτικό αυτού του είδους συνάρτησης είναι ότι έχει το πολύ ένα ελάχιστο, επομένως, δεν υπάρχουν τοπικά ελάχιστα που να μπορεί να εγκλωβιστεί ο αλγόριθμος. Με αυτόν τον τρόπο, η μέθοδος gradient descent μπορεί να

ξεκινήσει από οποιοδήποτε σημείο και είναι εγγυημένο ότι θα βρεί το ελάχιστο της συνάρτησης [62].

Στη μέθοδο logistic regression η παράμετρος w είναι πολύ πιο πολύπλοκη σε σχέση με τις υπόλοιπες μεθόδους καθώς επεξεργάζεται πολλές διαστάσεις και οφείλει να αντιστοιχίζει κάθε βάρος w_i για κάθε είσοδο x_i. Για κάθε διάσταση/μεταβλητή w_i, το gradient θα διαθέτει μία πληροφορία που μας ενημερώνει την κλίση για κάθε μεταβλητή και σε κάθε διάσταση εκφράζεται η κλίση ως μερική παράγωγος $\frac{\theta}{\theta(w_i)}$ της συνάρτησης απώλειας. Συνολικά, το gradient μίας συνάρτησης πολλαπλών μεταβλητών είναι ένα διάνυσμα όπου κάθε συνιστώσα εκφράζει μία μερική παράγωγο της συνάρτησης για κάθε μεταβλητή:

$$\nabla L(f(x;\theta),y) = \begin{bmatrix} \frac{\partial}{\partial\theta_1}L(f(x;\theta),y) \\ \frac{\partial}{\partial\theta_2}L(f(x;\theta),y) \\ \dots \\ \frac{\partial}{\partial\theta_n}L(f(x;\theta),y) \\ \frac{\partial}{\partial\theta_b}L(f(x;\theta),y) \end{bmatrix}$$

και έτσι η συνάρτηση που υπολογίζει το θ, δηλαδή η αντικειμενική συνάρτηση γίνεται:

$$\theta^{t+1} = \theta^t - \eta\nabla L(f(x;\theta),y)$$

Σε πολλές περιπτώσεις υπάρχει ο κίνδυνος το μοντέλο να προσαρμόσει τα βάρη σε υπερβολικό βαθμό βάσει των δεδομένων εκπαίδευσης έτσι ώστε να τα ταξινομεί στις κατάλληλες κλάσεις, με αποτέλεσμα να υπάρχει υπερπροσαρμογή. Αυτό συμβαίνει καθώς το μοντέλο θα αναθέτει υψηλές τιμές στα βάρη για συγκεκριμένες εισόδους με σκοπό να τις ταξινομεί αποκλειστικά σε μία και μοναδική κλάση. Προκειμένου να αποφευχθεί αυτό το φαινόμενο, ένα καλό μοντέλο πρέπει να έχει την ικανότητα να γενικεύει και για αυτό εισάγονται τεχνικές κανονικοποίησης και έτσι, ένας νέος όρος προστίθεται στην αντικειμενική συνάρτηση, ο $R(\theta)$. Αυτός ο όρος κανονικοποίησης χρησιμοποιείται προκειμένου να "τιμωρεί" τα μεγάλα βάρη, έτσι ώστε όταν τα βάρη ταιριάζουν σε μεγάλο βαθμό με τα δεδομένα εκπαίδευσης, θα τιμωρούνται έτσι ώστε να "ταιριάζουν" λιγότερο στα δεδομένα αυτά, οπότε κατά συνέπεια, θα χρησιμοποιεί μικρότερα σε τιμή βάρη [62]. Έτσι, η σχέση για ένα σύνολο m δειγμάτων η παραπάνω σχέση γίνεται:

$$\theta = argmax_\theta \left(\sum_{i=1}^{m} \log P(y^{(i)}|x^{(i)}) - \alpha R(\theta) \right)$$

Υπάρχουν δύο συνήθεις τρόποι υπολογισμού του όρου κανονικοποίησης $R(\theta)$. Ο πρώτος είναι ο L2 regularization που είναι μία τετραγωνική συνάρτηση των τιμών των βαρών, και το όνομά του το οφείλει στο γεγονός ότι αξιοποιεί την τετραγωνική νόρμα των τιμών βάρους, η οποία είναι ίδια με την Ευκλείδια απόσταση

του διανύσματος θ από την αρχή, και αν θεωρούμε ότι το θ έχει n βάρη, τότε, έχουμε:

$$R(\theta) = \|\theta\|_2^2 = \sum_{j=1}^{n} \theta_j^2$$

Ο δεύτερος τρόπος είναι με L1 regularization, ή αλλιώς LASSO, που είναι μία γραμμική συνάρτηση των τιμών των βαρών, και το όνομά του προκύπτει από την νόρμα $\|W\|_1$ η οποία υπολογίζει το άθροισμα των απόλυτων τιμών των βαρών, ή αλλιώς γνωστή και ως η απόσταση Μανχάτταν:

$$R(\theta) = \|\theta\|_1 = \sum_{i=1}^{n} |\theta_i|$$

Ο τρόπος που λειτουγεί ο αλγόριθμος με την κανονικοποίηση είναι ο εξής: συνήθως, σε δεδομένα υψηλών διαστάσεων οι μεταβλητές συσχετίζονται. Έτσι, με την κανονικοποίηση, όπως για παράδειγμα την LASSO, από μία ομάδα με υψηλά συσχετισμένες μεταβλητές, θα επιλεχθεί τυχαία μόνο μία από αυτές, και οι υπόλοιπες θα αφαιρεθούν. Έτσι, λοιπόν, το μοντέλο PLR προσθέτει έναν μη-αρνητικό όρο τιμωρίας προκειμένου να περιορίσει τις διαστάσεις και να χειρίζεται περιπτώσεις όπου οι μεταβλητές έχουν υψηλή συσχέτιση μεταξύ τους και να αποφύγει την υπερπροσαρμογή του μοντέλου.

Η υλοποίηση του μοντέλου πρόβλεψης με τη χρήση της κανονικοποιημένης λογιστικής παλινδρόμησης (Penalized Logistic Regression) περιλάμβανε μια σειρά από βήματα για την εξασφάλιση της βέλτιστης απόδοσης και ακρίβειας. Αρχικά, το σύνολο δεδομένων φορτώθηκε και προετοιμάστηκε χρησιμοποιώντας τη βιβλιοθήκη pandas. Η μεταβλητή-στόχος ήταν η Death, ενώ τα υπόλοιπα χαρακτηριστικά χρησιμοποιήθηκαν ως μεταβλητές εισόδου.

Η διαδικασία επιλογής χαρακτηριστικών περιλάμβανε τη χρήση διαφόρων μεθόδων, όπως correlation, Decision Trees, FFS, Fisher Scores, IG, LASSO Regularization, Random Forest, RFE, SFS και VIF. Για κάθε μέθοδο επιλογής χαρακτηριστικών, το αντίστοιχο σύνολο δεδομένων φορτώθηκε και προετοιμάστηκε.

Στη συνέχεια, πραγματοποιήθηκε αναζήτηση βέλτιστων υπερπαραμέτρων (GridSearchCV) για την κανονικοποιημένη λογιστική παλινδρόμηση, χρησιμοποιώντας τον StratifiedKFold για διασταυρούμενη επικύρωση με 5 διαίρεσεις. Οι υπερπαράμετροι που εξετάστηκαν περιλάμβαναν την παράμετρο κανονικοποίησης (penalty), είτε L1 είτε L2, και την αντίστροφη δύναμη κανονικοποίησης (C), με τιμές που κυμαίνονταν από 0.001 έως 10.

Για κάθε σύνολο δεδομένων που προέκυψε από τη διαδικασία επιλογής χαρακτηριστικών, η αναζήτηση υπερπαραμέτρων προσδιόρισε τις βέλτιστες παραμέτρους για την κανονικοποιημένη λογιστική παλινδρόμηση. Αυτές οι παράμετροι στη συνέχεια χρησιμοποιήθηκαν για την εκπαίδευση και αξιολόγηση του μοντέλου.

Η αξιολόγηση των μοντέλων πραγματοποιήθηκε χρησιμοποιώντας διασταυρούμενη επικύρωση (cross-validation) και τη μέθοδο cross_val_predict για την

πρόβλεψη των τιμών της μεταβλητής στόχου. Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση περιλάμβαναν την ακρίβεια (accuracy), την αναφορά ταξινόμησης (classification report), τον πίνακα σύγχυσης (confusion matrix), τις καμπύλες ROC και τις τιμές log loss.

Η διαδικασία αξιολόγησης περιλάμβανε τη δημιουργία και αποθήκευση πινάκων σύγχυσης και καμπυλών ROC, καθώς και τον υπολογισμό και την αποθήκευση των τιμών log loss για κάθε αναδίπλωση της διασταυρούμενης επικύρωσης. Η καμπύλη ROC παρείχε μια γραφική απεικόνιση της απόδοσης του μοντέλου, ενώ οι τιμές log loss παρείχαν μια ποσοτική μέτρηση της ακρίβειας των προβλέψεων του μοντέλου.

Συνολικά, η χρήση της κανονικοποιημένης λογιστικής παλινδρόμησης σε συνδυασμό με τη βελτιστοποίηση των υπερπαραμέτρων και την προσεκτική επιλογή χαρακτηριστικών παρείχε ένα ισχυρό πλαίσιο για την ανάπτυξη ακριβών και αξιόπιστων μοντέλων πρόβλεψης.

3.15 Δεδομένα

Ο σκοπός της έρευνας ήταν να εκτιμηθεί, με τη βοήθεια των παραπάνω ταξινομητών, η θνησιμότητα ενός ασθενούς με ένα από τα δύο συγκερκιμένα είδη κατάγματος ισχύου κατόπιν της εγχείρισης. Τα δεδομένα τα οποία αξιοποιήσαμε ήταν δοσμένα από την Πανεπιστημιακή Κλινική του νοσοκομείου ΚΑΤ. Ο λόγος που επιλέχθηκαν τα συγκεκριμένα μοντέλα ήταν επειδή αντίστοιχες έρευνες αξιοποιούσαν παρόμοια αλλά και, επίσης, έπρεπε να λάβουμε υπόψη μας ότι το σύνολο δεδομένων που λάβαμε από το νοσοκομείο δεν ήταν αρκετά μεγάλου όγκου ώστε να μπορούμε με επιτυχία να το αξιοποιήσουμε για την εκπαίδευση ενός πιο πολύπλοκου μοντέλου μηχανικής μάθησης, όπως για παράδειγμα ένα πολυεπίπεδο μοντέλο νευρωνικού δικτύου. Πέραν αυτών όμως, τα δεδομένα μας ήταν σε μορφή csv αρχείων, πράγμα που έκανε την φόρτωση τους στα μοντέλα αυτά εύκολη.

Από το σύνολο δεδομένων συλλέχθηκαν 400 έγκυρα περιστατικά ασθενών που είχαν ένα από τα δύο κατάγματα ισχύου: περιτροχαντήριο κάταγμα ισχύου με κωδικό 72.1 ή κάταγμα του αυχένα μηριαίου με κωδικό 72.0. Τα χαρακτηριστικά που συμπεριλαμβάνονταν στο σύνολο δεδομένων ήταν η ηλικία, το φύλο, ημέρες νοσηλείας (length of stay), ημερομηνία εισαγωγής, ημερομηνία εξόδου, ακτινολογική ταξινόμηση καταγμάτων αυχένα μηριαίου (Garden) και ακτινολογική ταξινόμηση καταγμάτων κατά ΑΟ με τρείς κατηγορίες 31A-A1, 31A-A2, 31A-A3 με απλούστερη ονομασία 1,2,3 αντιστοίχως αλλά και ακτινολογική ταξινόμηση περιτροχαντηρίων καταγωμάτων κατά Evans με τις τιμές 1 για σταθερά κατάγματα, 2 για ασταθή και 3 για κατάγματα ανάστροφης λοξότητας. Στη βάση δεδομένων υπήρχαν επίσης, οι συνολικές μέρες από την ημέρα εισαγωγής μέχρι την ημέρα του χειρουργείου όπου η τιμή 0 σημαίνει ότι ο ασθενής χειρουργήθηκε την ημέρα της εισαγωγής του. Ακόμη, συμπεριλαμβάνονται το είδος της αναισθησίας με τις ακόλουθες τιμές 1 ως

επισκληρίδιος (ραχιαία) αναισθησία και 2 ως γενική αναισθησία, το είδος του χειρουργείου όπου η τιμή 0 αντιστοιχεί στο ότι δεν χειρουργήθηκε, η τιμή 1 αντιστοιχεί στο ενδομυελικό ήλος μηριαίου, ένα είδος χειρουργείου που συνήθως υφίσταται σε περιπτώσεις περιτροχαντηρίων καταγμάτων, 2 είναι η ημιολική διπολική αρθροπλαστική, 3 η ημιολική αρθροπλαστική, 4 ολική αρθροπλαστική και τέλος 5 για την κοχλίωση. Τα 2,3,4 και 5 είναι συνήθως για την αντιμετώπιση κατάγματος του αυχένα του μηριαίου. Επιπροσθέτως, συμπεριλαμβάνονται στήλες που περιέχουν πληροφορίες όπως τον αριθμό των λευκών αιοσφαιρίων προ εξιτηρίου, τον αριθμό αιοσφαιρίνης προ εξιτηρίου που οι φυσιολογικές τιμές είναι 13.5-14.5 για τους άνδρες και 12.5-13.5 για τις γυναίκες αλλά και, επίσης, στήλες που ανάλογα με το εάν η τιμή τους είναι 1 ή 0, διευκρινίζουν αν ο ασθενής είχε κάποιο επικείμενο νόσημα από το ατομικό ιστροικό του, όπως για παράδειγμα πνευμονία, άννοια, καρδιακή ανεπάρκεια, διαβήτη, κολπική πυροδότηση, δυσλιπιδαιμία, υπέρταση και χρόνια νεφρική νόσο.

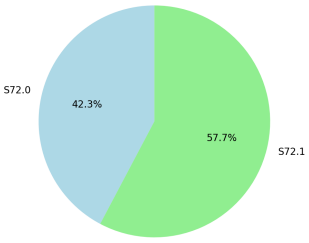
Άλλα στοιχεία που περιλαμβάνονταν είναι η βαθμολογία αδυναμίας (frailty score (1-9)) με την τιμή 1 να αντιστοιχεί σε έναν ασθενή που είναι πολύ ενεργός και σε καλή φόρμα εώς την τιμή 9 που αντιστοιχεί σε έναν ασθενή που είναι θανάσιμα άρρωστος. Είναι σημαντικό να επισημανθεί ότι η πληροφορία αυτή είναι μία εκτίμηση προ κατάγματος. Επίσης, συμπεριλαμβάνεται και η κατάσταση προ κατάγματος με αξιολόγηση 1 εώς και 4 όπου η τιμμή για την πλήρη κινητοποίηση είναι 1, για μερική κινητοποίηση με δυνατότητα αυτοεξυπηρέτησης είναι 2, για μερική κινητοποίηση χωρίς τη δυνατότητα αυτοεξυπηρέτησης 3, και, τέλος, για κλινοστατισμό 4. Μία επιπρόσθετη στήλη αντιστοιχεί στην πληροφορία για την τελική φυσική κατάσταση μετά το χειρουργείο με αξιολόγηση από 1 εώς 4, όπου η τιμή 1 αντιστοιχεί στην πλήρη αποκατάσταση ενώ η 4 στον κλινοστατισμό και οι υπόλοιπες στήλες περιλαμβάνουν πληροφορίες που αφορούν πιθανές επιπλοκές μετά την επέμβαση όπως την περιπροσθετική λοίμωξη, τη λοίμωξη ουρουποιητικού, πνευμονία, εν τω βάθη γλεβοθρόμβωση, χειρουργείο αναθεώρησης, εξάρθρημα ισχύου. Τέλος, υπάρχουν και οι στήλες που δηλώνουν το διάστημα που ο ασθενής απεβίωσε από το χειρουργείο, η τιμή 0 σε όλες τις στήλες που αντιστοιχούν σε μήνες υποδηλώνουν ότι ο ασθενής βρίσκεται ακόμα εν ζωή. Για την ευκολότερη διαχείριση σε μερικές περιπτώσεις, διαμορφώθηκε κατά την έρευνα μία τελευταία στήλη που υποδήλωνε εάν ο ασθενής είχε αποβιώσει ή όχι, μη λαμβάνοντας υπόψη το χρονικό διάστημα από την ημέρα της επέμβασης.

3.16 Μεθοδολογία: Επεξεργασία και Προετοιμασία Δεδομένων

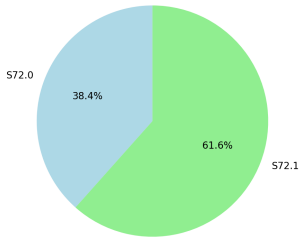
Όπως αναφέρθηκε παραπάνω, το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα μελέτη προήλθε από μια εκτεταμένη βάση δεδομένων για τα κατάγματα ισχίου από την πανεπιστημιακή ορθοπεδική κλινική του νοσοκομείου ΚΑΤ, η

οποία περιλάμβανε ένα ευρύ φάσμα μεταβλητών, όπως δημογραφικά στοιχεία των ασθενών, ιατρικές διαγνώσεις και χειρουργικές λεπτομέρειες. Η φάση προεπεξεργασίας των δεδομένων ήταν μεγάλης σημασίας για την προετοιμασία του συνόλου δεδομένων για ανάλυση και περιελάμβανε διάφορα βασικά βήματα.

Για να διασφαλιστεί η ακεραιότητα του συνόλου δεδομένων, απαλείφθηκαν οι γραμμές με ελλείπουσες τιμές σε κρίσιμες στήλες όπως «Ηλικία», «Φύλο» και «Χρόνος μέχρι τη χειρουργική επέμβαση». Το βήμα αυτό ήταν απαραίτητο για τη διατήρηση της ακρίβειας των επακόλουθων αναλύσεων. Επιπλέον, επειδή τα είδη καταγμάτων ήταν δύο με τις ακόλουθες ονομασίες (S72.0) Fracture of the femoral neck Hip bone fracture MCA και (S72.1) Pertrochanteric fracture Intratrochanteric fracture Trochanteric fracture, το αριθμητικό μέρος των πληροφοριών διάγνωσης εξήχθη με τη χρήση κανονικών εκφράσεων, δημιουργώντας μια νέα στήλη «Αριθμός διάγνωσης» για πιο προσιτό χειρισμό των δεδομένων.



(α') Τα δύο είδη καταγμάτων του συνόλου δεδομένων



(β') Τα δύο είδη καταγμάτων του συνόλου δεδομένων για τα περιστατικά θανάτου

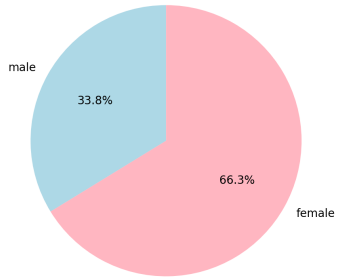
Σχήμα 3.16: Είδη καταγμάτων

Από όλα τα περιστατικά του συνόλου δεδομένων, το 42% των ασθενών διαγνώστηκε με (S72.0) Fracture of the femoral neck Hip bone fracture MCA και το υπόλοιπο 58% με (S72.1) Pertrochanteric fracture Intratrochanteric fracture Trochanteric fracture.

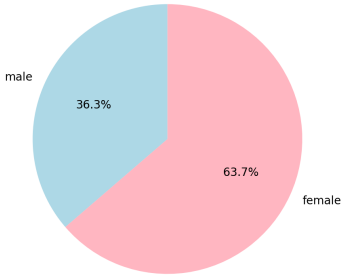
Από τα 400 περιστατικά, το 34% είναι άντρες και το υπόλοιπο 66% γυναίκες. Στη συνέχεια, στο διάγραμμα 2.17(β') φαίνεται από τα περιστατικά θανάτου, το 36% είναι άντρες και το υπόλοιπο 64% γυναίκες. Από αυτά τα νούμερα, θα μπορούσε να δημιουργηθεί η υπόθεση ότι το φύλο δεν παίζει σημαντικό ρόλο όσον αφορά την εκτίμηση των θανάτων.

Οι ημερομηνίες εισαγωγής και εξόδου μετατράπηκαν σε μορφή datetime για να διευκολυνθεί η χρονική ανάλυση. Δημιουργήθηκαν νέες στήλες για την καταγραφή του μήνα και του έτους τόσο της εισαγωγής όσο και της απόλυσης, επιτρέποντας την εξέταση εποχιακών τάσεων. Μη σχετικές στήλες, όπως ο Serial Number, ο Registration Number και αρκετές άλλες, αφαιρέθηκαν για να εξορθολογιστεί το σύνολο δεδομένων και να επικεντρωθεί σε πιο σχετικές μεταβλητές.

Για τη βελτίωση του συνόλου δεδομένων, εφαρμόστηκαν διάφορες τεχνικές μηχανικής χαρακτηριστικών. Οι αριθμητικές τιμές των μηνών για τις εισαγωγές



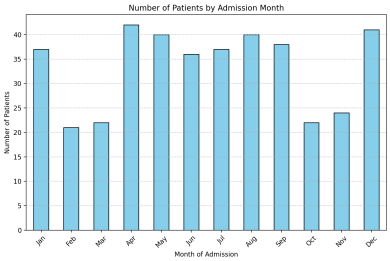
(α΄) Κατανομή φύλων



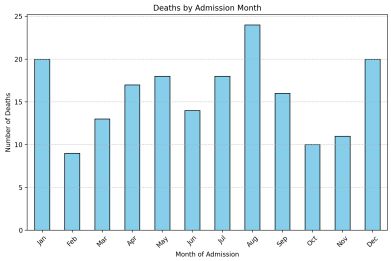
(β΄) Κατανομή φύλων για τα περιστατικά θανάτου

Σχήμα 3.17: Φύλα συνόλου δεδομένου

μετασχηματίστηκαν με τη χρήση συναρτήσεων ημιτόνου και συνημιτόνου για να αποτυπωθεί η περιοδική φύση των δεδομένων, απαραίτητη για την ανάλυση εποχικών προτύπων.



(α΄) Περιστατικά ανά τους μήνες



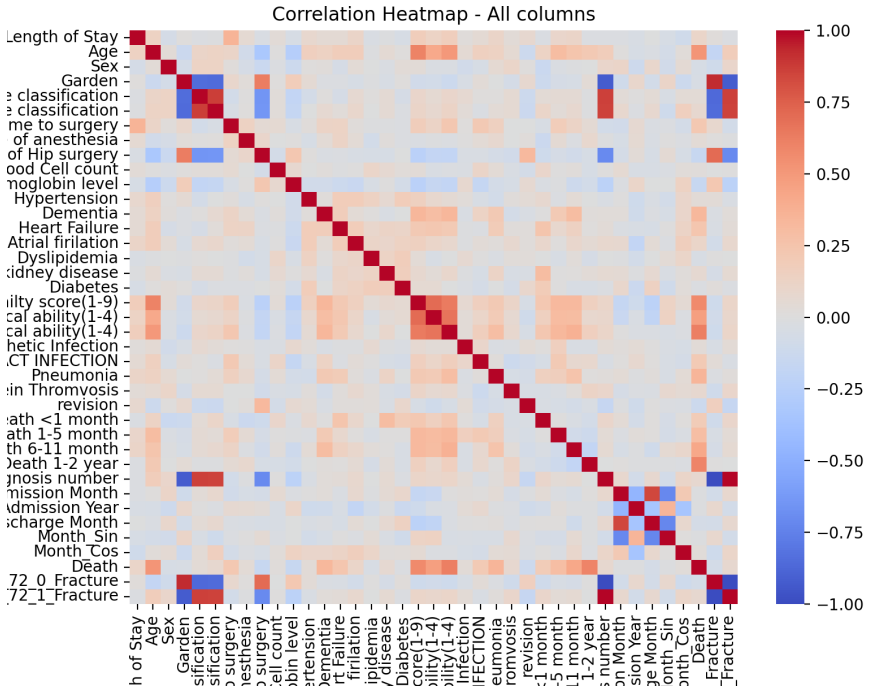
(β΄) Περιστατικά θανάτων ανά τους μήνες

Σχήμα 3.18: Περιστατικά και περιστατικά θανάτων ανά τους μήνες

Μπορεί από τα παρακάτω διαγράμματα να αποφαθνεί η αναγκαιότητα του μετασχηματισμού των μηνών σε συναρτήσεις ημιτόνου και συνημιτόνου καθώς παρουσιάζεται μία περιοδικότητα μεταξύ τους. Αναλυτικότερα, η κυκλική φύση των μηνών σημαίνει ότι ο Δεκέμβριος (12ος μήνας) είναι ¨δίπλα¨ στον Ιανουάριο του επόμενου έτους (1ο μήνα), πράγμα που δεν αντικατοπτρίζεται στην περίπτωση που αντιμετωπίζουμε τους μήνες ως γραμμικό χαρακτηριστικό. Σε μία γραμμική ανα-

παράσταση των δεδομένων, η αριθμητική διαφορά μεταξύ των μηνών Ιανουάριο και Δεκέμβριο είναι μεγάλη, ενώ στην πραγματικότητα είναι προσωρινά παρακείμενοι, πράγμα που επαληθεύεται και από το δεύτερο γράφημα που αναπαριστά τους θανάτους ανά μήνα, και κατά το οποίο στους μήνες Ιανουάριο και Δεκέμβριο, φαίνεται να παρατηρείται παρόμοιο μοτίβο θανάτων. Έτσι, με τον μετασχηματισμό αυτό, η κυκλική φύση των δεδομένων μας παραμένει αμετάβλητη και διασφαλίζει ότι η μετάβαση από τον Δεκέμβριο στον Ιανουάριο γίνεται ομαλά και με συνεχή τρόπο.

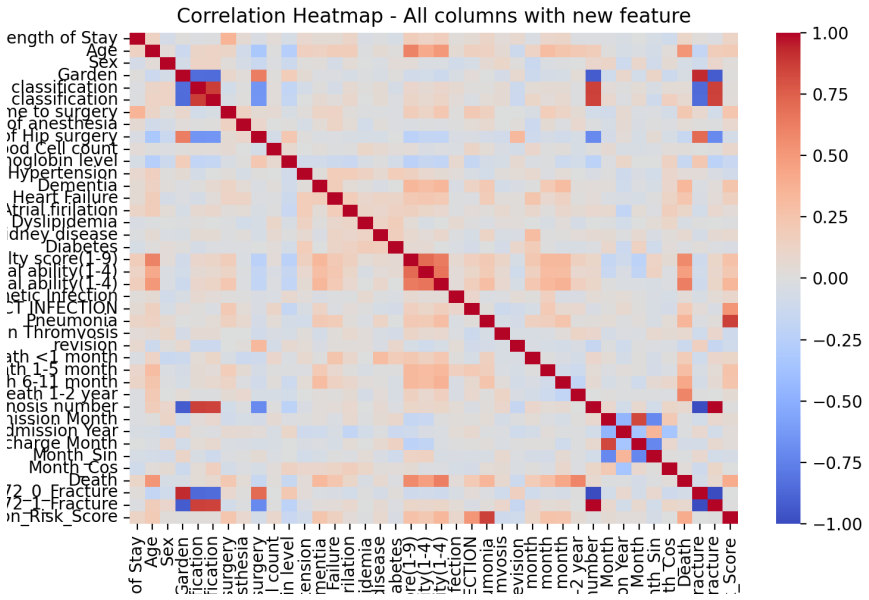
Αρκετές κατηγορικές στήλες μετατράπηκαν σε ακέραιους τύπους για να εξασφαλιστεί η συμβατότητα με τα αναλυτικά μοντέλα. Αυτές περιλάμβαναν ιατρικές ταξινομήσεις, τύπους χειρουργικών επεμβάσεων και διάφορους δείκτες ιατρικής κατάστασης. Δημιουργήθηκε μια νέα δυαδική στήλη για να δηλώνει αν ένας ασθενής πέθανε μετά την επέμβαση, ενοποιώντας διάφορες στήλες που σχετίζονται με το θάνατο σε έναν ενιαίο δείκτη.



Σχήμα 3.19: Θερμικός χάρτης για ολόκληρο το σύνολο δεδομένων

Υπολογίστηκαν, στη συνέχεια, οι συντελεστές συσχέτισης μεταξύ της βαθμολογίας κινδύνου λοίμωξης και των διαφόρων περιόδων θνησιμότητας για να εκτιμηθεί η προγνωστική ισχύς αυτού του σύνθετου χαρακτηριστικού. Δημιουργήθηκαν Heatmaps για την οπτικοποίηση του πίνακα συσχέτισης μετά την προσθήκη του νέου χαρακτηριστικού, παρέχοντας πληροφορίες σχετικά με τον αντίκτυπό του στο σύνολο δεδομένων.

Ο χάρτης αυτός μας εμφανίζει όλες τις συσχετίσεις μεταξύ των χαρακτηριστικών, και μας ενδιαφέρει κυρίως να εντοπίσουμε τα χαρακτηριστικά που συσχετίζονται περισσότερο με τη στήλη ¨θάνατος¨, δηλαδή εκείνα που έχουν τα πιο



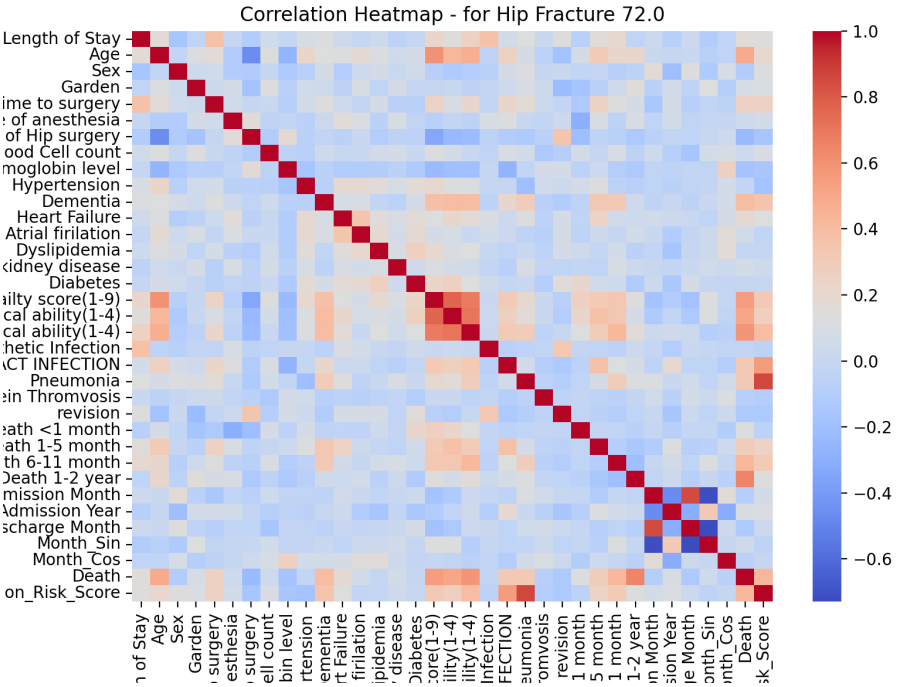
Σχήμα 3.20: Θερμικός χάρτης με το επιπλέον χαρακτηριστικό

έντονα χρώματα πάνω στον χάρτη, είτε είναι κόκκινο, που σημαίνει ότι υπάρχει έντονη θετική συσχέτιση, είτε είναι μπλε, πράγμα που σημαίνει ότι υπάρχει αρνητική συσχέτιση. Αυτά τα χαρακτηριστικά είναι τα frailty scores, η ηλικία, το είδος του χειρουργείου, η πνευμονία, η θρόμβωση, η άνοια, η καρδιακή ανεπάρκεια, και παρατηρούμε μία αντίστροφη συσχέτιση του θανάτου με τα hemoglobin levels.

Δημιουργήθηκε μια βαθμολογία κινδύνου λοίμωξης συνδυάζοντας διάφορες μεταβλητές που σχετίζονται με τη λοίμωξη (π.χ. «Πνευμονία», «Λοίμωξη της Ουρητικής Οδού», «Περιπροθετική Λοίμωξη») χρησιμοποιώντας βελτιστοποιημένα βάρη για τη μεγιστοποίηση της συσχέτισης με τη θνησιμότητα. Στο σημείο αυτό, αναπαράχθηκε άλλος ένας θερμικός χάρτης προκειμένου να εξεταστεί η συσχέτιση του νέου χαρακτηριστικού αυτού σε σχέση με τον θάνατο.

Είναι εμφανές ότι το καινούργιο χαρακτηριστικό, το οποίο αποτελεί συνδυασμό μερικών άλλων, συσχετίζεται αρκετά με τη στήλη θάνατος καθώς στον χάρτη απεικονίζεται η συσχέτιση του με τον θάνατο με έντονο πορτοκαλί χρώμα. Βάσει των συντελεστών συσχετίσεων, μία από τις απλές μεθόδους επιλογής χαρακτηριστικών βασίστηκε στις τιμές των συντελεστών αυτοσυσχέτισης μεταξύ των χαρακτηριστικών αυτών και της στήλης του θανάτου.

Δημιουργήθηκαν, επίσης, δυαδικοί δείκτες για συγκεκριμένους τύπους καταγμάτων, επιτρέποντας μια πιο λεπτομερή ανάλυση των διαφόρων τύπων καταγμάτων στο σύνολο δεδομένων, προκειμένου να εξεταστεί εάν κάποιο από τα δύο είδη καταγμάτων ήταν πιο ¨βαρύ¨ και συσχετιζόταν περισσότερο με τον θάνατο σε σχέση με το άλλο, για τη διαδικασία επιλογής χαρακτηριστικών. Κατ' επέκταση, παράχθηκαν διάφορες οπτικοποιήσεις για την απόκτηση βαθύτερης κατανόησης του συνόλου δεδομένων, συμπεριλαμβανόμενοι θερμικοί χάρτες συσχετίσεων για την οπτικοποίηση των σχέσεων μεταξύ διαφορετικών μεταβλητών, βοηθώντας στον εντοπισμό ι-



Σχήμα 3.21: Θερμικός χάρτης για το (S72.0) Fracture of the femoral neck Hip bone fracture MCA

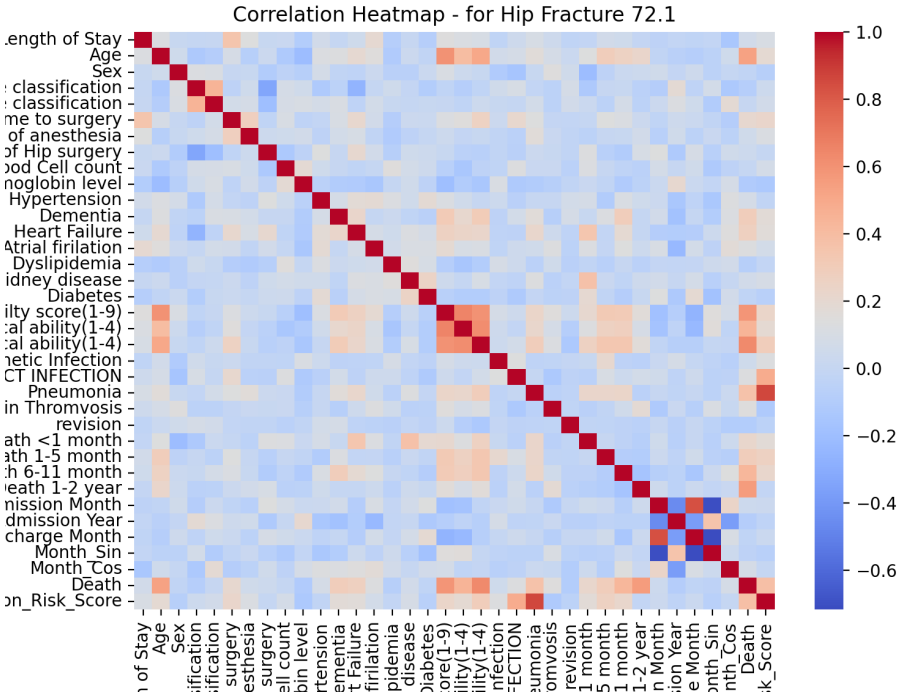
σχυρών συσχετίσεων και πιθανών προβλεπτικών παραγόντων για τη θνησιμότητα και άλλες εκβάσεις. Έτσι, πραγματοποιήθηκαν διαχωρισμένες αναλύσεις με βάση συγκεκριμένους τύπους καταγμάτων (S72.0 και S72.1), με τη δημιουργία ξεχωριστών θερμικών χαρτών συσχέτισης για κάθε τύπο κατάγματος, ώστε να κατανοηθούν οι μοναδικές σχέσεις εντός αυτών των υποομάδων.

Από ότι απεικονίζεται και στους δύο παραπάνω χάρτες μπορούμε εύκολα να συμπεράνουμε ότι οι συσχετίσεις μεταξύ των χαρακτηριστικών και των διαφορετικών ειδών καταγμάτων δεν διαφέρουν πολύ.

Σχεδιάστηκαν ιστογράμματα για να εξεταστούν οι κατανομές βασικών μεταβλητών, όπως η ηλικία, οι φυσικές ικανότητες πριν και μετά το κάταγμα, ο αριθμός των λευκών αιμοσφαιρίων και τα επίπεδα αιμοσφαιρίνης. Ξεχωριστά ιστογράμματα για τις περιπτώσεις θανάτου παρείχαν συγκριτική ανάλυση για τον εντοπισμό τάσεων και διαφορών.

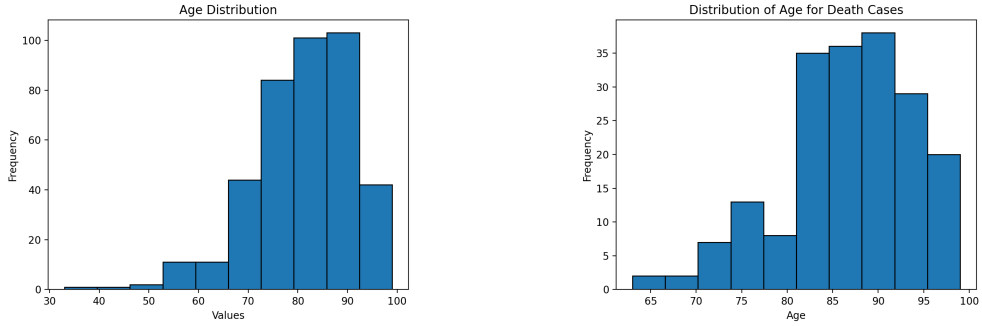
Χρησιμοποιήθηκαν ραβδογράμματα για την απεικόνιση του αριθμού των εισαγωγών ασθενών και των θανάτων ανά μήνα, αποκαλύπτοντας τυχόν εποχιακές τάσεις στα ποσοστά εισαγωγών και θνησιμότητας. Η αναλογία ανδρών και γυναικών ασθενών εξετάστηκε μέσω ιστογραμμάτων, τόσο για το σύνολο των δεδομένων όσο και ειδικά για τις περιπτώσεις θανάτου.

Αυτή η προσέγγιση για την προεπεξεργασία δεδομένων και τη μηχανική των χαρακτηριστικών εξασφάλισε ότι το σύνολο δεδομένων προετοιμάστηκε σχολαστικά για τη μετέπειτα ανάλυση και επιλογή χαρακτηριστικών και, στη συνέχεια, εκπα-



Σχήμα 3.22: Θερμικός χάρτης για το (S72.1) Pertrochanteric fracture Intratrochanteric fracture Trochanteric fracture

ίδευση των μοντέλων με σκοπό την εκτίμηση της θνησιμότητας κάποιου ασθενούς.

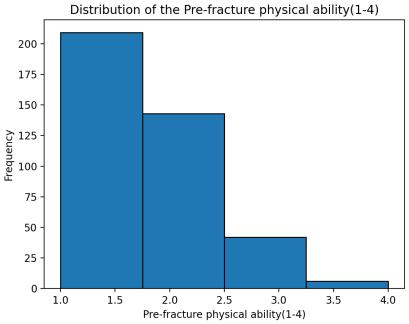


(α΄) Ηλικίες (β΄) Θάνατοι ανά ηλικία

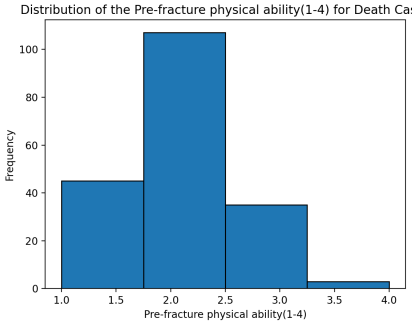
Σχήμα 3.23: Ηλικίες συνόλου δεδομένων και θάνατοι ανα ηλικία

Για περεταίρω δημογραφικά στοιχεία όσον αφορά τους ασθενείς του συνόλου δεδομένων από τα διαγράμματα φαίνεται ότι οι ασθενείς είναι μεγάλης ηλικίας με ελάχιστα περιστατικά κάτω των 70 ετών και ακόμα λιγότερα κάτω των 50. Οι περισσότεροι ασθενείς είναι κοντά στο διάστημα 70-100 ετών. Αντίστοιχα, από τους ασθενείς που απεβίωσαν μετά την επέμβαση, η πλειοψηφία των περιστατικών είναι για ασθενείς άνω των 80 ετών.

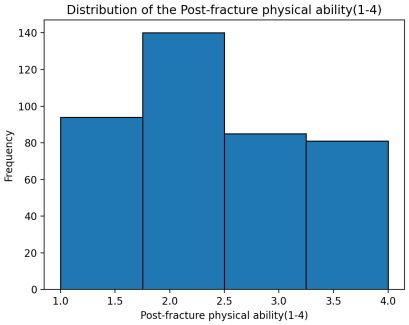
Ένα ακόμα χαρακτηριστικό το οποίο φάνηκε μετέπειτα αρκετά καθοριστικό βάσει μερικών μεθόδων επιλογής χαρακτηριστικών είναι η κινητικότητα και ικανότητα κίνησης του ασθενούς πριν το κάταγμα, δηλαδή το pre-fracture physical



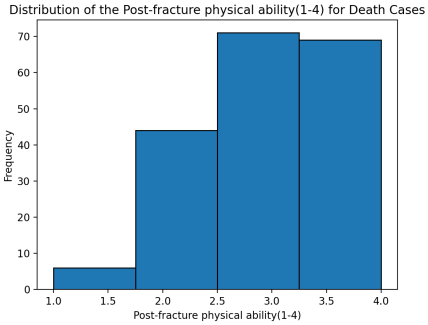
(α') Κατανομή των περιστατικών με την κινητοποίηση των ασθενών πριν το κάταγμα όλου του συνόλου δεδομένων



(β') Κατανομή των περιστατικών θανάτων με την κινητοποίηση των ασθενών πριν το κάταγμα



(γ') Κατανομή των περιστατικών με την κινητοποίηση των ασθενών μετά το κάταγμα όλου του συνόλου δεδομένων

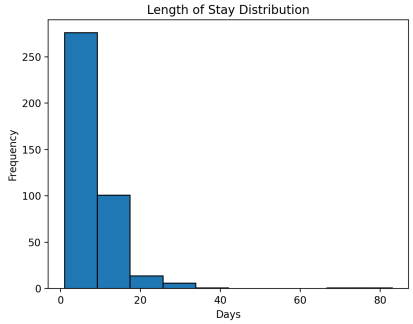


(δ') Κατανομή των περιστατικών θανάτων με την κινητοποίηση των ασθενών μετά το κάταγμα

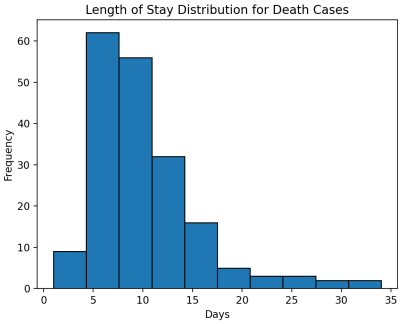
Σχήμα 3.24: Κινητικότητα ασθενών πριν και μετά το κάταγμα

ability. Παρακάτω, στα σχήματα 3.24 (α') και (β') απεικονίζονται οι κατανομές των ασθενών όσον αφορά την κινητικότητα τους με το 1 να είναι πλήρης κινητοποίηση και το 4 να αντιστοιχεί σε κλινοστατισμός, δηλαδή μηδενική κινητοποίηση. Επιπλέον, αναλύθηκε και η στήλη που αφορά την φυσική κατάσταση του ασθενούς μετά την επέμβαση, post-fracture physical ability, με την ίδια βαθμονόμηση όπως και στην κατάσταση του ασθενούς πριν το κάταγμα. Το οποίο φαίνεται να παίζει σημαντικό ρόλο, όσον αφορά τη συσχέτιση με τη θνησιμότητα του ασθενούς. Τέλος, αναλύθηκε και το frailty score των ασθενών, αφού απεικονίστηκε η κατανομή του τόσο για όλο το σύνολο δεδομένων όσο και για το φιλτραρισμένο σύνολο δεδομένων από όλους τους ασθενείς που απεβίωσαν. Από τους χάρτες είναι εμφανές ότι υπάρχει μεγάλη συσχέτιση με το frailty score και τον θάνατο, αφού, αναδεικνύει και την ευαισθησία του ασθενούς.

Όσον αφορά την κινητική ικανότητα των ασθενών πριν και μετά το κάταγμα, αυτό που μπορεί να ερμηνευθεί από τα διαγράμματα 3.24, είναι ότι ενω η κατάσταση και η ικανότητα του ασθενούς να μπορεί να κινηθεί πριν το κάταγμα δεν παίζει καθοριστικό ρόλο όσον αφορά την πιθανότητα του να πεθάνει ή όχι, καθώς στο διάγραμμα φαίνεται ότι οι περισσότεροι ασθενείς που έχασαν τη ζωή τους είχαν βαθμό κινητικότητας 2, αυτό που έχει σημασία είναι η κινητικότητα μετά το κάταγμα. Αυτό



(α') Κατανομή ημερών παραμονής στο νοσοκομείο



(β') Κατανομή ημερών παραμονής στο νοσοκομείο για τα περιστατικά θανάτου

Σχήμα 3.25: Ημέρες παραμονής στο νοσοκομείο

μπορεί να φανεί και στο διάγραμα 2.22(β') στο οποίο τα περισσότερα περιστατικά θανάτων ήταν ασθενείς με κινητικότητα μετά το κάταγμα με βαθμό μεγαλύτερο από 2.

Επίσης, αναπαράχθηκαν διαγράμματα που απεικονίζουν τις ημέρες παραμονής των ασθενών τόσο ολόκληρου του συνόλου δεδομένων όσο και των περιστατικών με θανάτους. Αυτό που παρατηρείται είναι ότι τα περισσότερα περιστατικά θανάτων συνέβησαν για τους ασθενείς που είχαν παραμείνει στο νοσοκομείο από 5 μέχρι 15 μέρες, πράγμα που είναι λογικό, εφόσον η πλειοψηφία όλων των ασθενών είχαν παραμείνει σε αυτό το διάστημα στο νοσοκομείο.

Κεφάλαιο 4

Αποτελέσματα

4.1 Επιλογής Χαρακτηριστικών

Κατά την εκτέλεση των μεθόδων επιλογής χαρακτηριστικών, για κάθε μέθοδο προέκυψαν διαφορετικοί συνδυασμοί χαρακτηριστικών από το αρχικό σύνολο δεδομένων μας. Αρχικά από τη μέθοδο επιλογής χαρακτηριστικών βάσει του κριτηρίου της συσχέτισης, δηλαδή να επιλεχθούν τα πρώτα δέκα χαρακτηριστικά που έχουν μεγαλύτερη συσχέτιση με τη στήλη ¨Θάνατος¨ του συνόλου δεδομένων, τα χαρακτηριστικά που αποτέλεσαν υποσύνολο για την τροφοδότηση στο μοντέλο μηχανικής μάθησης είναι τα ακόλουθα: Post-opperative physical ability(1-4), Frailty score(1-9), Age, Pre-fracture physical ability(1-4), Infection Risk Score, Pneumonia, Dementia, Heart Failure, Time to surgery, Urinary Tract Infection.

Στη συνέχεια, η επόμενη μέθοδος δέντρο αποφάσεων, επέλεξε τα εξής χαρακτηριστικά: Post-opperative physical ability(1-4), Age, Hemoglobin level, Length of Stay, Frailty score(1-9), white Blood Cell count, Infection Risk Score, Month Sin, Pre-fracture physical ability(1-4), Discharge Month.

Το μοντέλο Forward Feature Selection επέλεξε ως χαρακτηριστικά τα: Sex, Garden, AO fracture classification, Evans Fracture classification, Type of anesthesia, Type of Hip surgery, Atrial fibrilation, Chronic kidney disease, Diabetes, Frailty score(1-9), Post-opperative physical ability(1-4), Periprosthetic Infection, Pneumonia, Deep Vein Thrombosis, Revision, Diagnosis number, Infection Risk Score.

Η μέθοδος Fisher Score κατέληξε στα χαρακτηριστικά: Frailty score(1-9), Post-opperative physical ability(1-4), Age, Pre-fracture physical ability(1-4), Infection Risk Score, Dementia, Pneumonia, Heart Failure, Hemoglobin level, Type of Hip surgery.

Η μέθοδος Random Forest επέλεξε τα: Age, Sex, AO fracture classification, Type of anesthesia, Hypertension, Dementia, Heart Failure, Atrial fibrilation, Dyslipidemia, Chronic kidney disease, Pre-fracture physical ability(1-4), Post-opperative physical ability(1-4), Admission Month, Type 72 0 Fracture, Infection Risk Score.

Η μέθοδος κέρδους πληροφορίας IG επέλεξε τα: Age, Frailty score(1-9), Pre-fracture physical ability(1-4), Post-opperative physical ability(1-4), Urinary Tract Infection.

Η μέθοδος LASSO Regularization επέλεξε τα: Age, white Blood Cell count.

Η μέθοδος Recursive Feature Elimination επέλεξε τα: Dementia, Heart Failure, Atrial fibrilation, Frailty score(1-9), Pre-fracture physical ability(1-4), Post-opperative physical ability(1-4), Urinary Tract Infection, Pneumonia, Month Cos, Type 72 1 Fracture.

Η μέθοδος Sequential Feature Elimination επέλεξε τα: Length of Stay, Age, Sex, Garden, AO fracture classification, Time to surgery, Type of anesthesia, Type of Hip surgery, white Blood Cell count, Hemoglobin level, Hypertension, Dementia, Atrial fibrilation, Dyslipidemia, Chronic kidney disease, Frailty score(1-9), Post-opperative physical ability(1-4), Periprosthetic Infection, Urinary Tract Infection, Pneumonia, Deep Vein Thrombosis, Revision, Diagnosis number, Admission Month, Admission Year, Discharge Month, Month Sin, Month Cos.

Τέλος, η μέθοδος Variance Inflation Factor επέλεξε τα: Length of Stay, Age, Sex, Time to surgery, Type of anesthesia, Type of Hip surgery, white Blood Cell count, Hemoglobin level, Hypertension, Dementia, Heart Failure, Atrial fibrilation, Dyslipidemia, Chronic kidney disease, Diabetes, Pre-fracture physical ability(1-4), Post-opperative physical ability(1-4), Deep Vein Thrombosis, Revision, Admission Year, Month Cos.

Στη συνέχεια, θα αναλυθούν τα αποτελέσματα προβλέψεων κάθε μοντέλου.

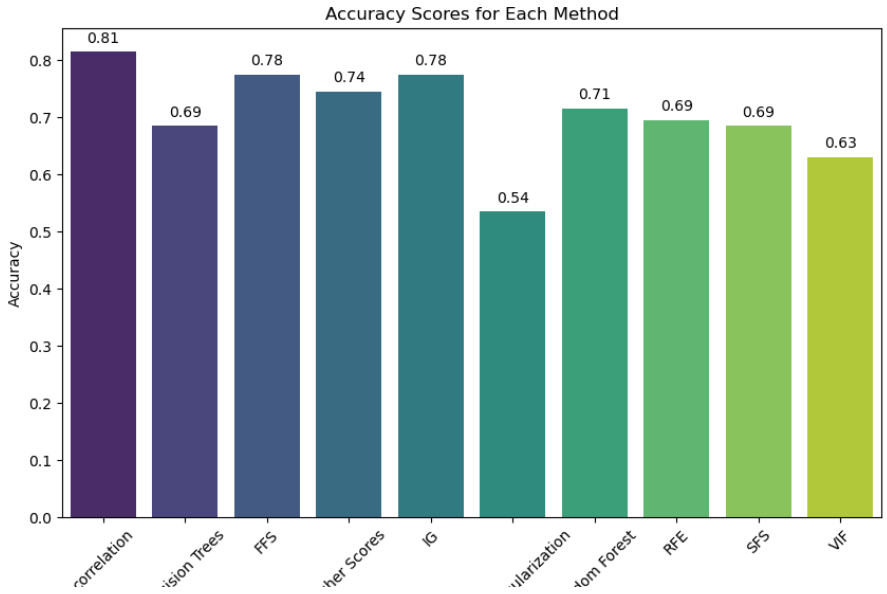
4.2 Δέντρα Αποφάσεων

Για τα δέντρα αποφάσεων, αρχικά, κατά την προσομοίωση τους επιλέχθηκαν οι κατάλληλες αρχιτεκτονικές, με άλλα λόγια συνδυασμοί υπερπαραμέτρων με τις οποίες πέτυχαν τις υψηλότερες ακρίβειες το μοντέλο για κάθε μέθοδο επιλογής χαρακτηριστικών. Πιο συγκερκιμένα, για κάθε μέθοδο επιλέχθηκαν οι ακόλουθοι συνδυασμοί:

Method	Criterion	Max Depth	Min Samples Leaf	Min Samples Split
Correlation	entropy	10	4	10
Decision Trees	entropy	10	4	10
FFS	entropy	10	4	10
Fisher Scores	entropy	10	4	10
IG	entropy	10	4	10
LASSO	entropy	10	4	10
Random Forest	entropy	10	4	10
RFE	entropy	10	4	10
SFS	entropy	10	4	10
VIF	entropy	10	4	10

Πίνακας 4.1: Υπερπαράμετροι για κάθε μέθοδο

Ακολουθούν αναλυτικότερα αποτελέσματα ως προς την επίδοση του μοντέλου.



Σχήμα 4.1: Οι ακρίβειες του μοντέλου Δέντρων Αποφάσεων για κάθε μία μέθοδο

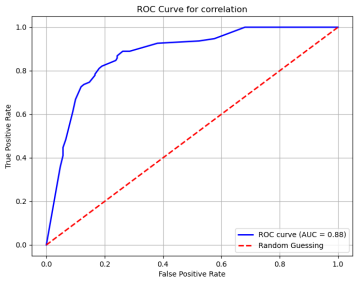
Τη μέγιση ακρίβεια την πέτυχε το μοντέλο με τα χαρακτηριστικά που είχαν επιλεχθεί με το κριτήριο της συσχέτισης, η οποία ήταν 81%, ενώ την ελάχιστη την πέτυχε με τη μέθοδο LASSO Regularization, η οποία ήταν 54%.

Στη συνέχεια, απεικονίστηκαν τα διαγράμματα ROC Curves προκειμένου να ελεχθεί περεταίρω το μοντέλο ως προς την επίδοση του. Όπως φαίνεται και από τα διαγράμματα παραπάνω, είναι εμφανές ότι για όλες τις μεθόδους επιλογής χαρακτηριστικών, το μοντέλο έκανε εκτιμήσεις πολύ καλύτερα από το random guessing που αντιστοιχεί στην κόκκινη διακεκομένη γραμμή. Στην αξιολόγηση των AUC βαθμολογιών για τις διάφορες μεθόδους επιλογής χαρακτηριστικών, παρατηρούμε τα ακόλουθα αποτελέσματα:

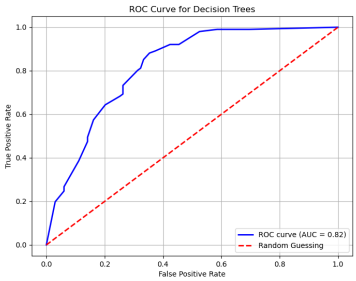
Method	AUC
Correlation	0.88
Decision Trees	0.82
Forward Feature Selection	0.82
Fisher Scores	0.80
Information Gain	0.83
Lasso Regularization	0.61
Random Forest	0.77
Recursive Feature Elimination	0.79
Sequential Feature Selection	0.73
Variance Inflation Factor	0.74

Πίνακας 4.2: AUC βαθμολογίες για κάθε μέθοδο

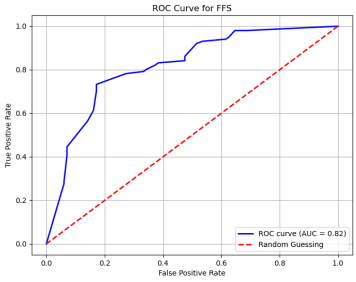
Για την περεταίρω ανάλυση των αποτελεσμάτων των Δέντρων Αποφάσεων,



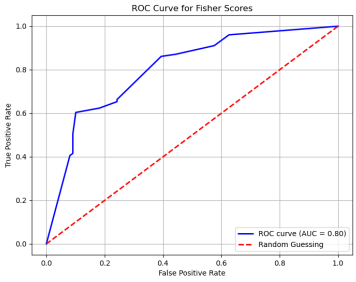
(α′) Correlation



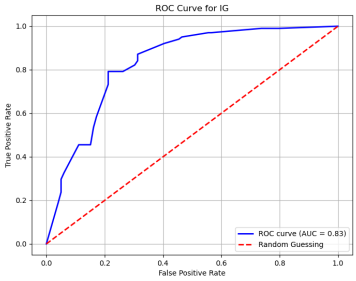
(β′) Decision Trees



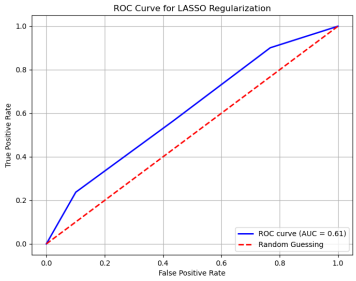
(γ′) Forward Feature Selection



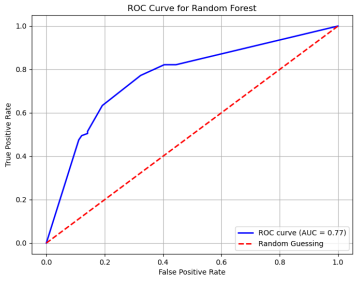
(δ′) Fisher Scores



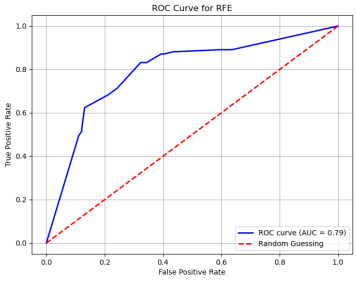
(ε′) Information Gain



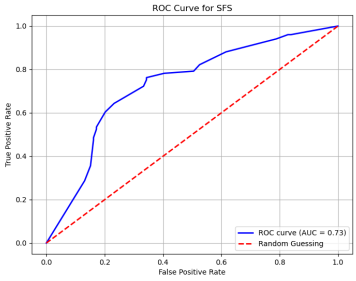
(ϛ′) Lasso Regularization



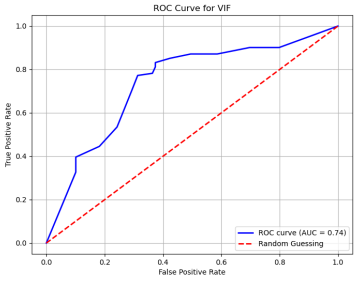
(ζ) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection



(ι′) Variance Inflation Factor

Σχήμα 4.2: ROC Curves για όλες τις μεθόδους για το μοντέλο Δέντρα Αποφάσεων

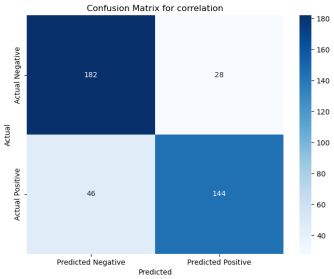
αξιοποιήσαμε τη μήτρα σύγχυσης, η οποία παρέχει μια αναλυτική παρουσίαση των αποτελεσμάτων της ταξινόμησης, καταγράφοντας τις προβλέψεις του μοντέλου έναντι των πραγματικών τιμών. Επειδή, ωστόσο, για την εκπαίδευση των μοντέλων χρησιμοποιήσαμε και τη μέθοδο cross validation, μπορεί να παρατηρηθεί πώς για κάθε μέθοδο επιλογής χαρακτηριστικών, το άθροισμα όλων στοιχείων των στηλών και των γραμμών της μήτρας δεν είναι το ίδιο. Αυτό οφείλεται στο γεγονός ότι το cross validation επιλέγει τυχαία τμήματα ως προς το μέγεθος και τα στοιχεία για κάθε μέθοδο, επομένως έτσι ο αριθμός των δειγμάτων που εξετάζεται και εκτιμάται ανά μέθοδο διαφέρει.

Method	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
Correlation	81.5%	81.8%	81.2%	81.3%
Decision Trees	68.5%	68.7%	68.4%	68.4%
FFS	77.5%	77.8%	77.4%	77.4%
Fisher Scores	74.5%	74.5%	74.5%	74.5%
IG	77.5%	77.5%	77.5%	77.5%
LASSO Regularization	53.5%	54.4%	53.2%	50.0%
Random Forest	71.5%	72.5%	71.6%	71.2%
RFE	69.5%	69.7%	69.5%	69.4%
SFS	68.5%	68.5%	68.5%	68.5%
VIF	63.0%	63.4%	63.1%	62.8%

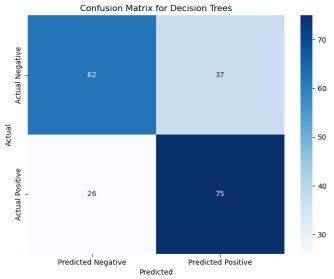
Πίνακας 4.3: Μετρικές Απόδοσης για κάθε μέθοδο

Από ότι μπορεί να παρατηρηθεί από τα διαγράμματα που απεικονίζουν τις συναρτήσεις log-loss για κάθε μέθοδο επιλογής χαρακτηριστικών ως προς τον αριθμό των folds, παρατηρούμε ότι παρόλο που το κύριο μοντέλο είναι το ίδιο (Δέντρα Αποφάσεων), η συμπεριφορά του μοντέλου για κάθε συνδυασμό χαρακτηριστικών διαφέρει. Μερικοί συνδυασμοί παρουσιάζουν παρόμοια συμπεριφορά και καλή επίδοση για folds ίσα με 1 ή 3 και πολύ κακή επίδοση (υψηλότερο log-loss) για folds ίσα με 2. Ενώ άλλα μοντέλα έχουν την ακριβώς αντίθετη συμπεριφορά, δηλαδή πολύ καλή επίδοση για 2 folds και κακές επιδόσεις για folds ίσα με 1 και 3 και σε άλλα το log-loss αυξάνεται γραμμικά, ανάλογα με τα folds και σε άλλα να μειώνεται ανάλογα κατα μήκος των folds.

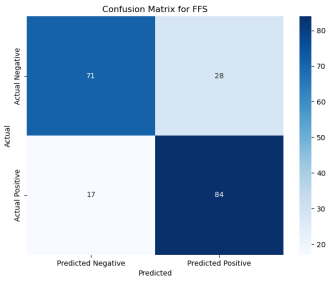
Συνολικά, το μοντέλο των Δέντρων Αποφάσεων παρουσιάζει παρόμοια αποτελέσματα για τη πλειοψηφία των μεθόδων επιλογής χαρακτηριστικών.



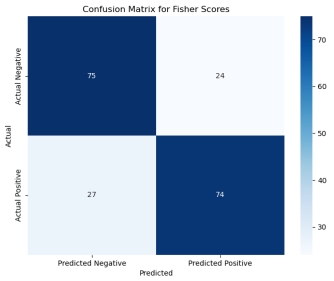
(α′) Correlation



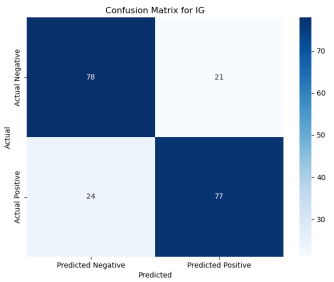
(β′) Decision Trees



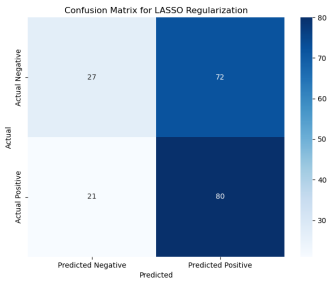
(γ′) Forward Feature Selection



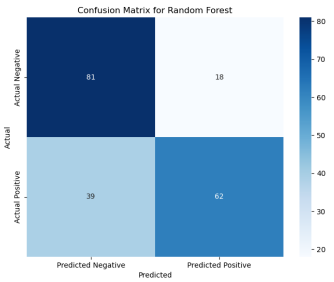
(δ′) Fisher Scores



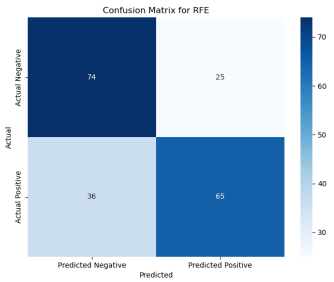
(ε′) Information Gain



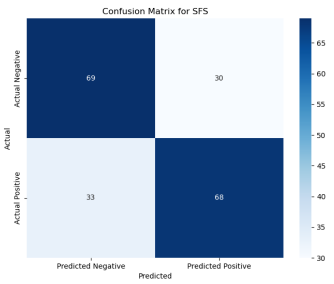
(ϛ′) Lasso Regularization



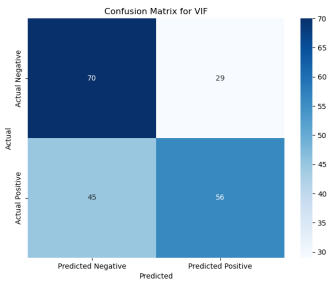
(ζ′) Random Forest



(η′) Recursive Feature Elimination

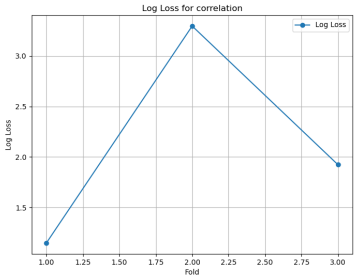


(θ′) Sequential Feature Selection

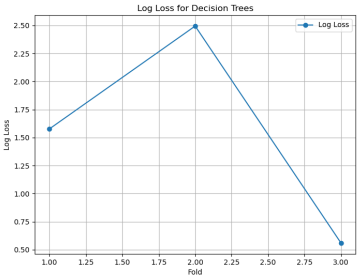


(ι′) Variance Inflation Factor

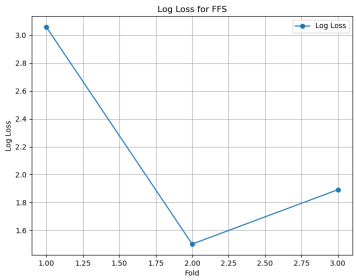
Σχήμα 4.3: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Δέντρα Αποφάσεων



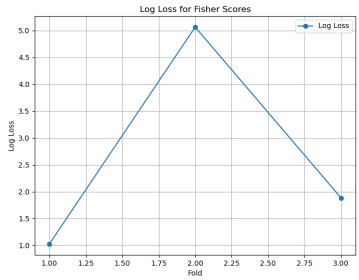
(α΄) Correlation



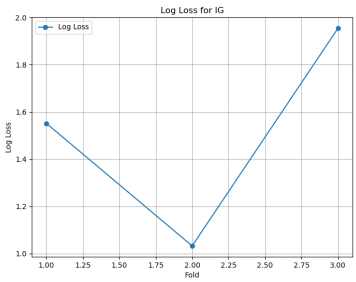
(β΄) Decision Trees



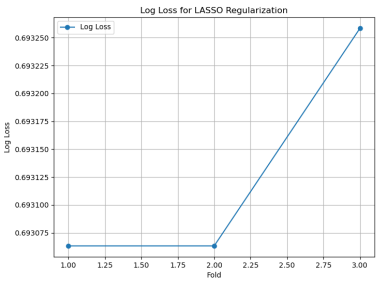
(γ΄) Forward Feature Selection



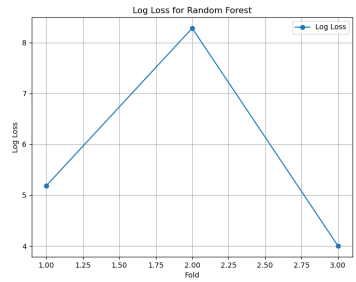
(δ΄) Fisher Scores



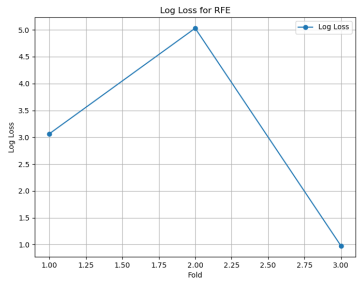
(ε΄) Information Gain



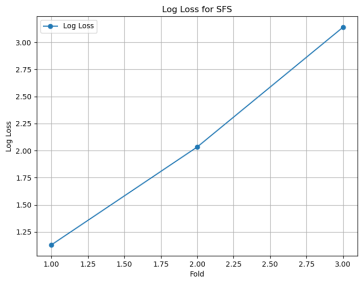
(ϛ΄) Lasso Regularization



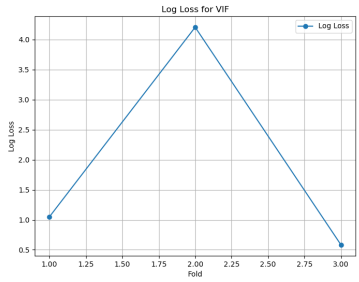
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination



(θ΄) Sequential Feature Selection



(ι΄) Variance Inflation Factor

Σχήμα 4.4: Log-Loss ως προς τον αριθμό των folds συναρτήσεις για όλες τις μεθόδους για το μοντέλο Δέντρα Αποφάσεων

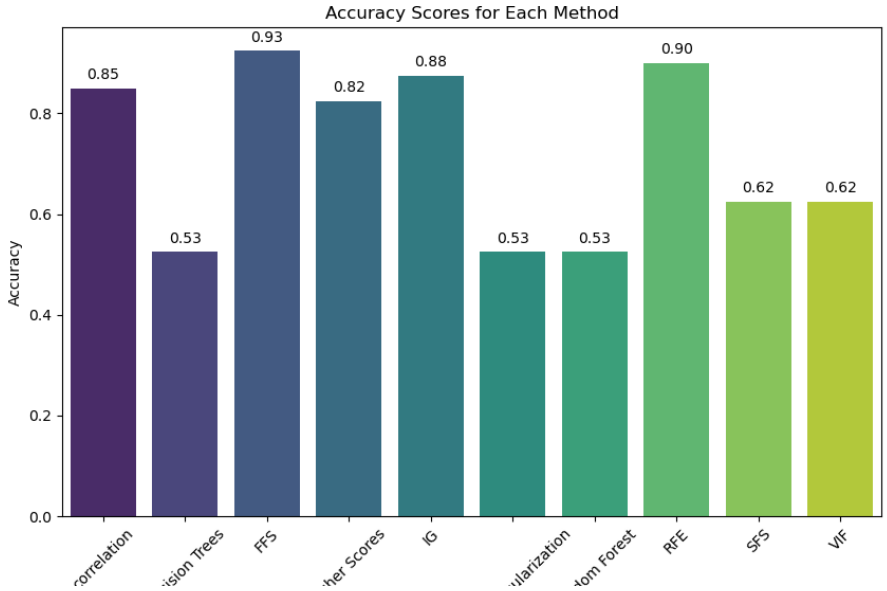
4.3 Back Propagation

Για το μοντέλο αυτό, εξετάστηκαν πολλαπλές αρχιτεκτονικές (συνδυασμοί κρυφών επιπέδων και αριθμό κόμβων) και για κάθε μέθοδο επιλέχθηκε η βέλτιστη αρχιτεκτονική με την οποία επιτύγχανε τη μέγιστη ακρίβεια. Όλα τα μοντέλα έχουν ένα επίπεδο εισόδου, δύο κρυφά επίπεδα και ένα επίπεδο εξόδου και ανάλογα την αρχιτεκτονική, μετά έχουν διαφορετικό πλήθος νευρώνων. Ξεκινώντας με τη μέθοδο επιλογής χαρακτηριστικών βάσει του κριτηρίου της συσχέτισης, ο αλγόριθμος επέλεξε ως βέλτιστη αρχιτεκτονικές την εξής με δύο κρυφά επίπεδα:

- Χρησιμοποιεί το πρώτο κρυφό επίπεδο με 20 νευρώνες, το δεύτερο επίπεδο με 10 νευρώνες, αριθμός εποχών ίσος με 1100.
- Χρησιμοποιεί πρώτο επίπεδο με 5 νευρώνες, δεύτερο επίπεδο με 5 νευρώνες, αριθμός εποχών ίσος με 100.
- Με το Forward Feature Selection χρησιμοποιεί το πρώτο κρυφό επίπεδο με 50 νευρώνες, το δεύτερο κρυφό επίπεδο με 25 νευρώνες, αριθμός εποχών ίσος με 1000.
- Με τα Fisher Scores χρησιμοποιεί το πρώτο κρυφό επίπεδο με 40 νευρώνες, το δεύτερο κρυφό επίπεδο με 20 νευρώνες, αριθμός εποχών ίσος με 1100.
- Με το κέρδος πληροφορίας χρησιμοποιεί το πρώτο κρυφό επίπεδο με 30 νευρώνες, το δεύτερο κρυφό επίπεδο με 20 νευρώνες, αριθμός εποχών ίσος με 800.
- Με τη μέθοδο Lasso Regularization χρησιμοποιεί το πρώτο κρυφό επίπεδο με 30 νευρώνες, το δεύτερο κρυφό επίπεδο με 20 νευρώνες, αριθμός εποχών ίσος με 1200.
- Με τη μέθοδο Random Forest χρησιμοποιεί το πρώτο κρυφό επίπεδο με 5 νευρώνες, το δεύτερο κρυφό επίπεδο με 5 νευρώνες, αριθμός εποχών ίσος με 100.
- Με τη Recursive Feature Elimination χρησιμοποιεί το πρώτο κρυφό επίπεδο με 30 νευρώνες, το δεύτερο κρυφό επίπεδο με 20 νευρώνες, αριθμός εποχών ίσος με 1200.
- Με τη Sequential Feature Selection χρησιμοποιεί το πρώτο κρυφό επίπεδο με 40 νευρώνες, το δεύτερο κρυφό επίπεδο με 5 νευρώνες, αριθμός εποχών ίσος με 900.

Συνολικά, οι ακρίβειες που πέτυχε το μοντέλο αυτό με κάθε μέθοδο επιλογής χαρακτηριστικών απεικονίζεται παρακάτω και μπορεί να αποφανθεί ότι σε με μερικές μεθόδους η επίδοση του μοντέλου είναι πολύ υψηλή, συγκεκριμμένα για την μέθοδο επιλογής βάσει της συσχέτισης με 85%, το Forward Feature Selection με 93%, με

τα Fisher Scores 82%, το κέρδος πληροφορίας 88% και το Recursive Feature Elimination με 90%. Αλλά μπορεί να είναι και πολύ χαμηλή, όπως στις υπόλοιπες μεθόδους με χαμηλότερη ακρίβεια ίση με 53%.



Σχήμα 4.5: Οι ακρίβειες του μοντέλου Back Propagation για κάθε μία μέθοδο

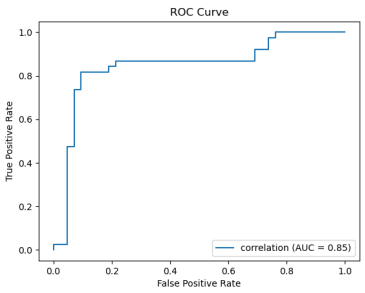
Ακολουθούν και τα ROC διαγράμματα και όπως και με την ακρίβεια, το μοντέλο σε συνδυασμό με πολλές μεθόδους φαινεται να έχει πολύ καλή επίδοση, ενώ με μερικές άλλες μεθόδους φαίνεται να έχει ίδια επίδοση με αυτή που θα είχε αν έκανε εκτιμήσεις τυχαία (random guessing). Πιο συγκεκριμένα:

Method	AUC
Correlation	0.85
Decision Trees	0.5
Forward Feature Selection	0.89
Fisher Scores	0.87
Information Gain	0.91
Lasso Regularization	0.48
Random Forest	0.48
Recursive Feature Elimination	0.90
Sequential Feature Selection	0.62
Variance Inflation Factor	0.61

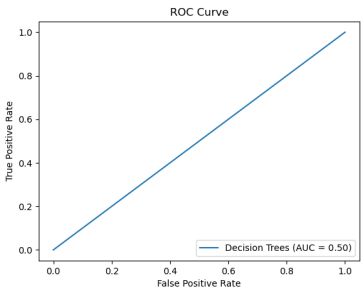
Πίνακας 4.4: AUC σορες φορ εαση μετηοδ

Παρακάτω εμφανίζονται και οι μήτρες σύγχυσης για καθεμία από τις μεθόδους.

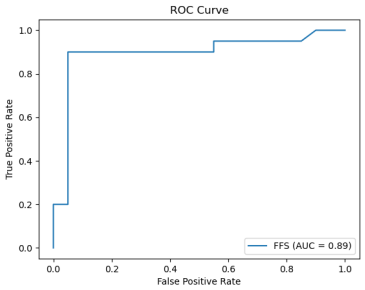
Σε αυτή την ενότητα, αξιολογούμε την απόδοση του αλγορίθμου ανάδρασης (back propagation) χρησιμοποιώντας διαφορετικές μεθόδους επιλογής χαρακτηριστικών. Η αξιολόγηση περιλαμβάνει την ακρίβεια, τη διαμόρφωση, αναλυτική αναφορά ταξινόμησης και τη μήτρα σύγχυσης για κάθε μέθοδο.



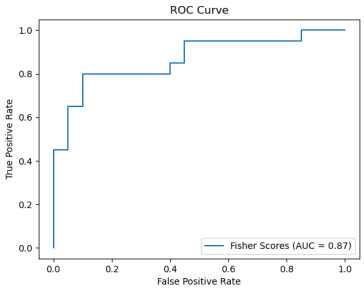
(α΄) Correlation



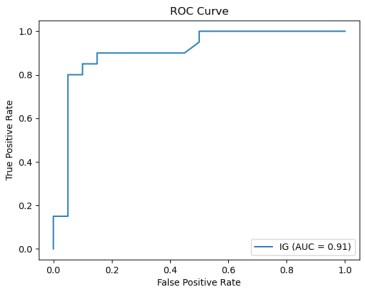
(β΄) Decision Trees



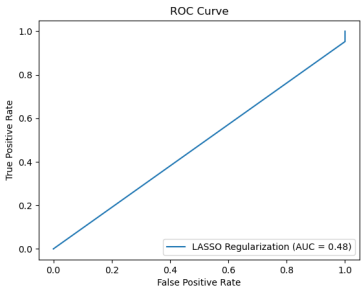
(γ΄) Forward Feature Selection



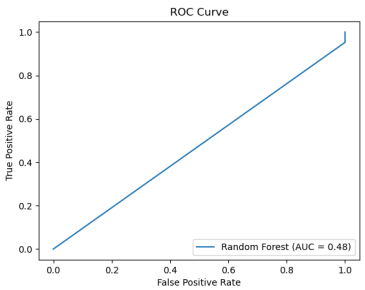
(δ΄) Fisher Scores



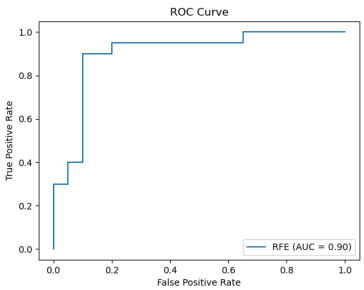
(ε΄) Information Gain



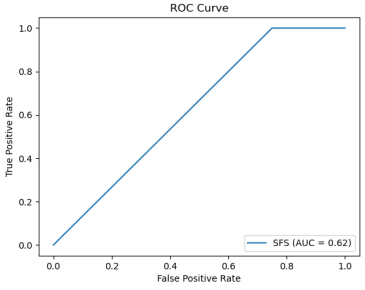
(ϛ΄) Lasso Regularization



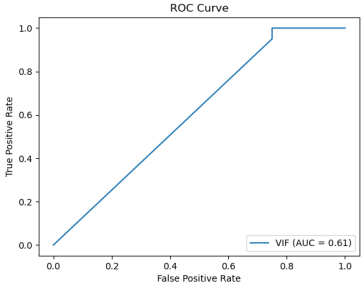
(ζ) Random Forest



(η΄) Recursive Feature Elimination

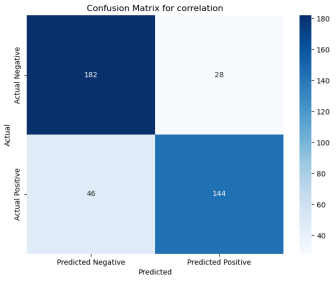


(θ΄) Sequential Feature Selection

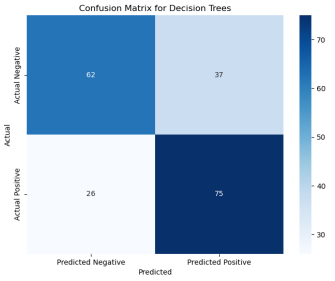


(ι΄) Variance Inflation Factor

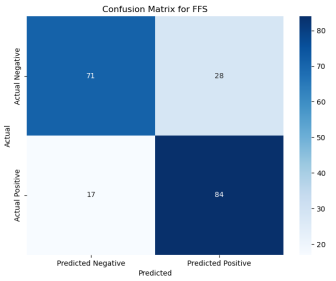
Σχήμα 4.6: ROC Curves για όλες τις μεθόδους για το μοντέλο Back Propagation



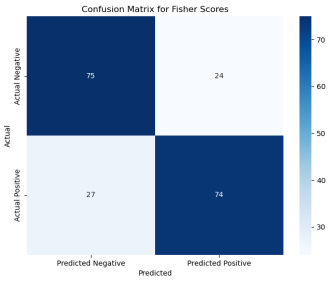
(α΄) Correlation



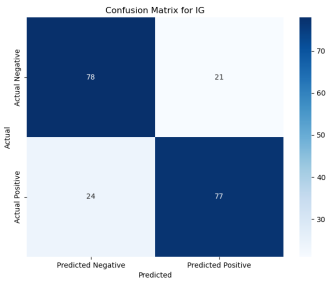
(β΄) Decision Trees



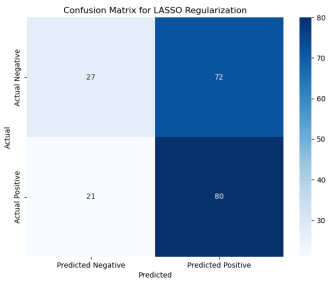
(γ΄) Forward Feature Selection



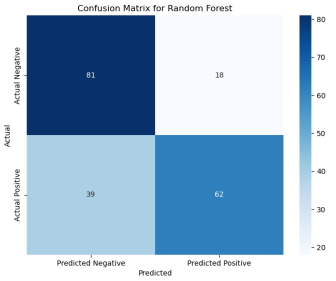
(δ΄) Fisher Scores



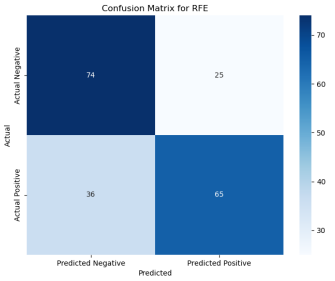
(ε΄) Information Gain



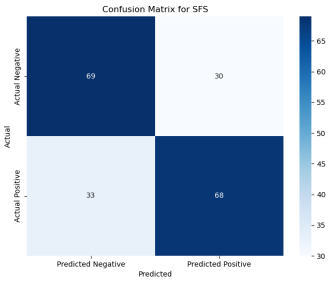
(ϛ΄) Lasso Regularization



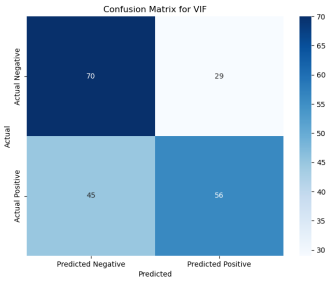
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination



(θ΄) Sequential Feature Selection



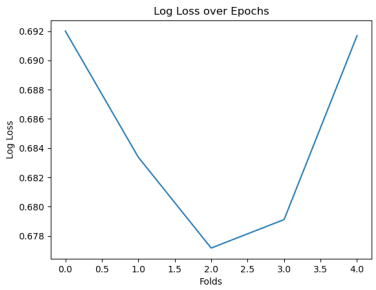
(ι΄) Variance Inflation Factor

Σχήμα 4.7: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Δέντρα Αποφάσεων

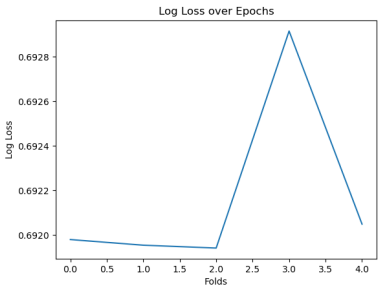
Method	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
Correlation	85.0%	85.4%	84.7%	84.8%
Decision Trees	52.5%	26.3%	50.0%	34.4%
FFS	92.5%	92.6%	92.5%	92.5%
Fisher Scores	82.5%	83.2%	82.5%	82.4%
IG	87.5%	87.6%	87.5%	87.5%
LASSO Regularization	52.5%	26.3%	50.0%	34.4%
Random Forest	52.5%	26.3%	50.0%	34.4%
RFE	90.0%	90.0%	90.0%	90.0%
SFS	62.5%	78.6%	62.5%	56.4%
VIF	62.5%	78.6%	62.5%	56.4%

Πίνακας 4.5: Μετρικές Απόδοσης του μοντέλου για κάθε μέθοδο

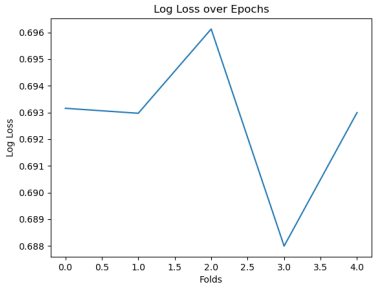
Παρακάτω απεικονίζονται και οι γραφικές παραστάσεις των συναρτήσεων log-loss ως προς τα folds. Αυτό που μπορεί να παρατηρηθεί είναι ότι το μοντέλο back propagation συπμεριφέρεται διαφορετικά σε κάθε fold για κάθε διαφορετική μέθοδο επιλογής χαρακτηριστικών. Ακόμα και τα μοντέλα με τις μέγιστες επιδόσεις έχουν διαφορετική συμπεριφορά μεταξύ και σε κάποια folds έχουν πολύ υψηλό λογ-λοσς και σε κάποια έχουν πολύ χαμηλά.



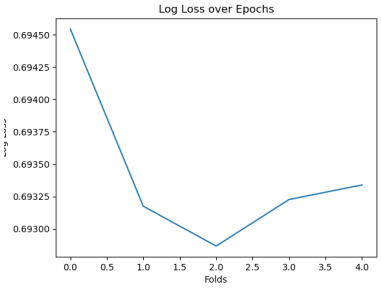
(α΄) Correlation



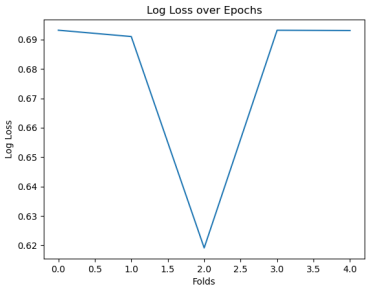
(β΄) Decision Trees



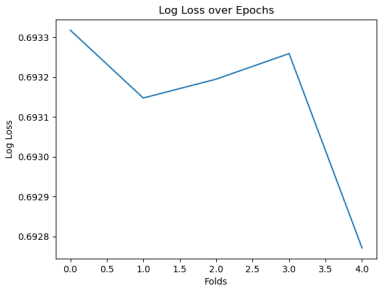
(γ΄) Forward Feature Selection



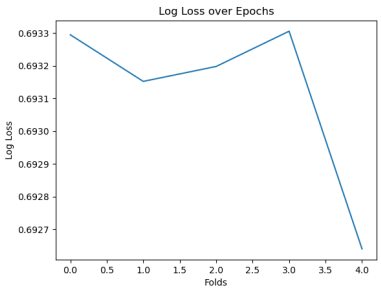
(δ΄) Fisher Scores



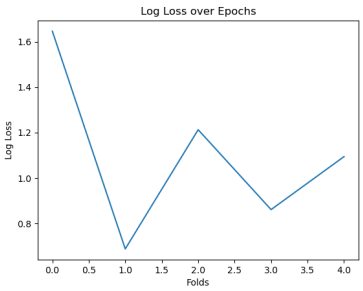
(ε΄) Information Gain



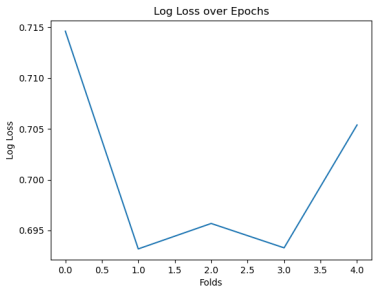
(ϛ΄) Lasso Regularization



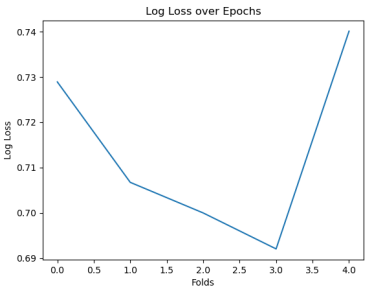
(ζ) Random Forest



(η΄) Recursive Feature Elimination



(θ΄) Sequential Feature Selection



(ι΄) Variance Inflation Factor

Σχήμα 4.8: Log-Loss ως προς τον αριθμό των folds συναρτήσεις για όλες τις μεθόδους για το μοντέλο Back Propagation

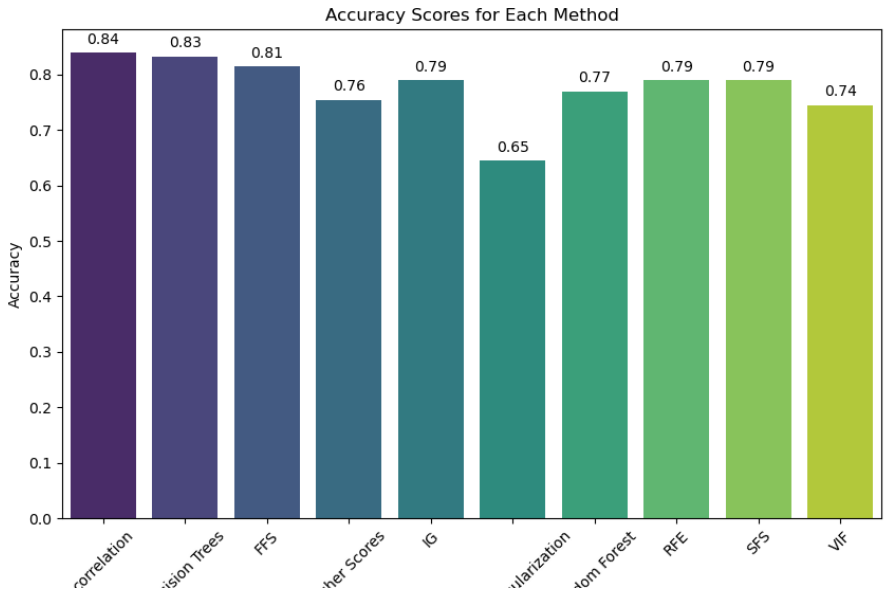
4.4 Gradient Boosting

Στην διαδικασία εκπαίδευσης του μοντέλου αυτού συνδυάζοντας διαφορετικές μεθόδους επιλογής χαρακτηριστικών, επιλέχθηκαν για κάθε μέθοδο ο βέλτιστος αριθμός υπερπαραμέτρων. Έτσι, για κάθε μέθοδο προέκυψαν οι παρακάτω συνδυασμοί:

Method	Learning Rate	Max Depth	Min Samples Leaf	Min Samples Split	N Estimators	Subsample
Decision Trees	0.01	5	2	5	200	1.0
Forward Feature Selection	0.01	5	1	5	100	0.8
Fisher Scores	0.1	3	2	5	100	0.9
Information Gain	0.01	3	1	2	200	1.0
LASSO L1 Regularization	0.01	3	4	5	300	0.9
Random Forest	0.01	5	4	5	300	0.8
Recursive Feature Elimination	0.01	5	1	2	200	0.9
Sequential Feature Selection	0.01	5	2	2	100	0.9
Variance Inflation Factor	0.01	3	4	2	100	1.0

Πίνακας 4.6: Υπερπαράμετροι για κάθε μέθοδο

Όσον αφορά την απόδοση του μοντέλου, οι μέγιστες ακρίβειες που πέτυχε με κάθε μέθοδο απεικονίζονται στο ακόλουθο διάγραμμα:

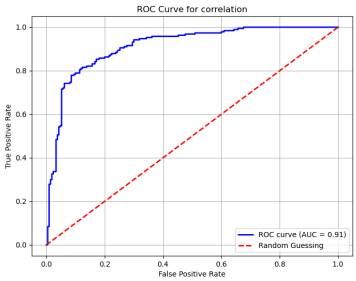


Σχήμα 4.9: Οι ακρίβειες του μοντέλου Gradient Boosting για κάθε μία μέθοδο

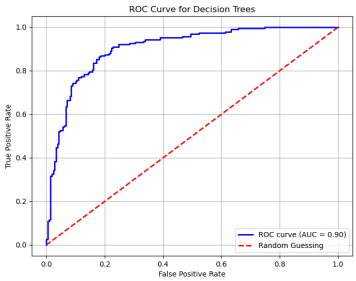
Οι ακρίβειες που επιτυγχάνει το μοντέλο για την πλειοψηφία των μεθόδων επιλογής χαρακτηριστικών είναι αρκετά ικανοποιητικές και παρόμοιες μεταξύ τους με ελάχιστες εξαιρέσεις. Για την περεταίρω ανάλυση όμως του μοντέλου, αναπαράχθηκαν και τα ROC Curves.

Από ό,τι φαίνεται, το μοντέλο αποδίδει πολύ καλύτερα από ότι θα απέδιδε το random guessing για κάθε μέθοδο επιλογής χαρακτηριστικών. Πιο συγκεκριμένα, για την αξιολόγηση των AUC βαθμολογιών για τις διάφορες μεθόδους επιλογής χαρακτηριστικών, παρατηρούμε τα ακόλουθα αποτελέσματα:

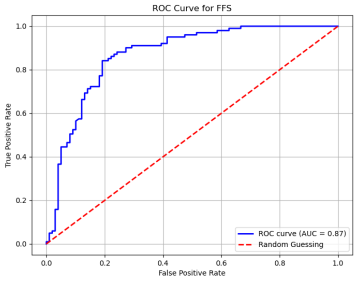
Στη συνέχεια, ακολούθησαν και οι απεικονίσεις των μητρών σύγχυσης προκειμένου να αναλυθεί λίγο καλύτερα η απόδοση του μοντέλου για κάθε μέθοδο επιλογής χαρακτηριστικών.



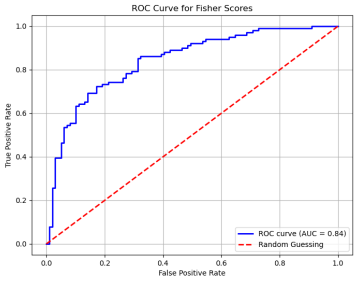
(α′) Correlation



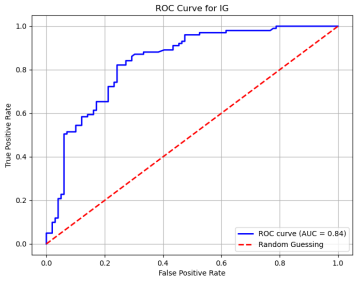
(β′) Decision Trees



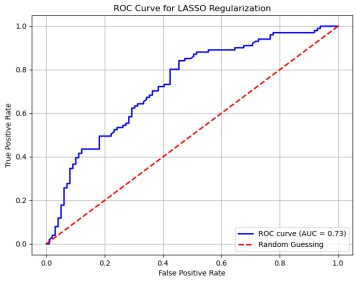
(γ′) Forward Feature Selection



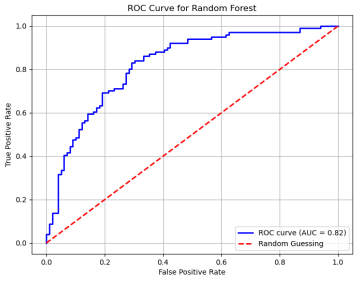
(δ′) Fisher Scores



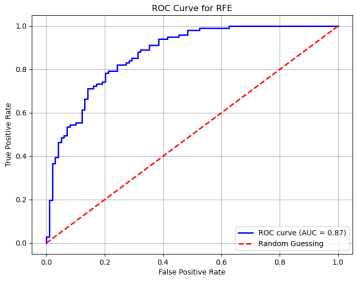
(ε′) Information Gain



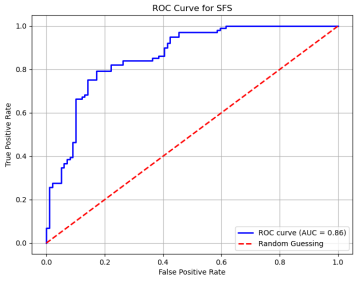
(ϛ′) Lasso Regularization



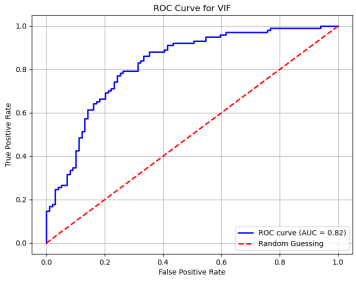
(ζ′) Random Forest



(η′) Recursive Feature Elimination

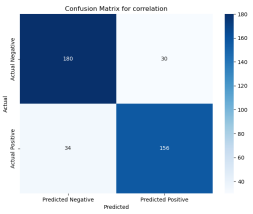


(θ′) Sequential Feature Selection

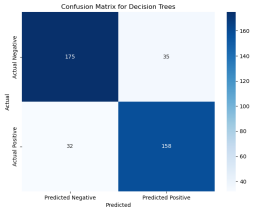


(ι′) Variance Inflation Factor

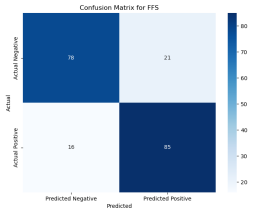
Σχήμα 4.10: ROC Curves για όλες τις μεθόδους για το μοντέλο Gradient Boosting



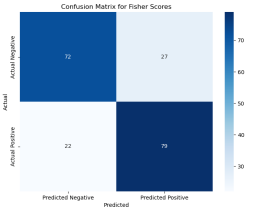
(α′) Correlation



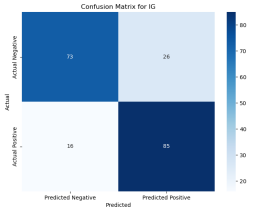
(β′) Decision Trees



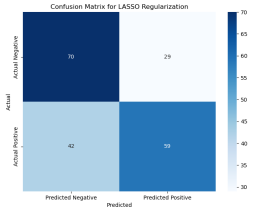
(γ′) Forward Feature Selection



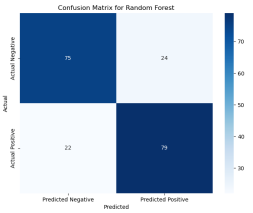
(δ′) Fisher Scores



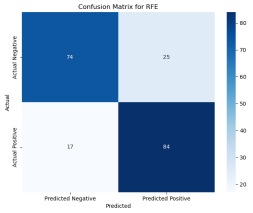
(ε′) Information Gain



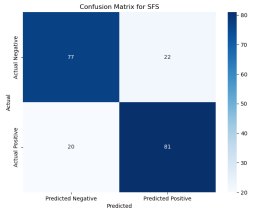
(ϛ′) Lasso Regularization



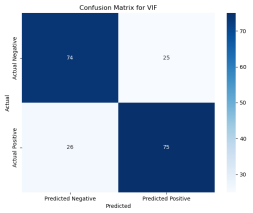
(ζ′) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection



(ι′) Variance Inflation Factor

Σχήμα 4.11: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Gradient Boosting

Method	AUC Score
Correlation	0.91
Decision Trees	0.90
Forward Feature Selection	0.87
Fisher Scores	0.84
Information Gain	0.84
Lasso Regularization	0.73
Random Forest	0.82
Recursive Feature Elimination	0.87
Sequential Feature Selection	0.86
Variance Inflation Factor	0.82

Πίνακας 4.7: AUC Scores για κάθε μέθοδο

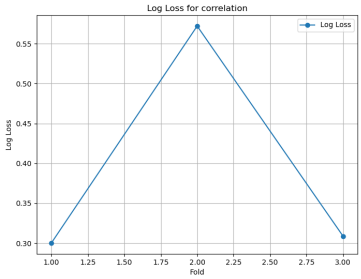
Σε συνέχεια, με τα διαγράμματα των μητρών σύγχυσης μπορεί να αναλυθεί περεταίρω η συμπεριφορά του μοντέλου σε συνδυασμό με άλλες μετρικές οι οποίες απορρέουν από τα διαγράμματα αυτά.

Method	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
Correlation	84%	84.0%	83.9%	83.9%
Decision Trees	81%	81.5%	81.0%	81.0%
FFS	83%	84.0%	83.0%	84.0%
Fisher Scores	76%	75.5%	75.0%	75.0%
Information Gain	79%	79.0%	78.0%	79.0%
Lasso Regularization	65%	65.0%	65.0%	65.0%
Random Forest	77%	77.0%	77.0%	77.0%
RFE	79%	79.0%	79.0%	79.0%
SFS	79%	79.0%	79.0%	79.0%
VIF	75%	75.0%	75.0%	74.0%

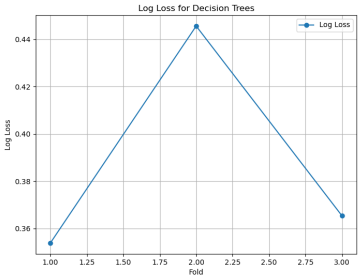
Πίνακας 4.8: Μετρικές Απόδοσης για κάθε μέθοδο

Οι πληροφορίες αυτές συνοδεύονται, ακόμη, και με τις γραφικές παραστάσεις των log-loss σφαλμάτων συναρτήσει τον αριθμό των αναδιπλώσεων (folds) προκειμένου να αξιολογηθεί και η συμμπεριφορά των μοντέλων κατά τη διάρκεια του cross validation.

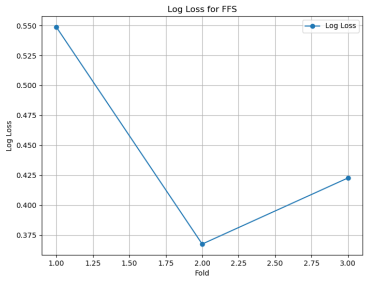
Όπως μπορεί να παρατηρηθεί, μερικά μοντέλα μεταξύ τους παρουσιάζουν παρόμοιες συμπεριφορές ως προς τα σφάλματα μεταξύ τους. Συγκεκριμένα, το μοντέλο με το correlation, το μοντέλο με τα δέντρα αποφάσεων και εκείνο με το Lasso Regularization παρουσιάζουν μία τριγωνική συμπεριφορά, δηλαδή για 1 και 3 διαιρέσεις ελαχιστοποιείται το log-loss ενώ για 2 διαιρέσεις, κλιμακώνεται το log-loss. Αντίθετα, τα μοντέλα με τις μεθόδους forward feature selection και sequential feature selection φαίνεται να έχουν ακριβώς αντίθετη συμπεριφορά. Τέλος, τα υπόλοιπα μοντέλα παρουσιάζουν κοινή καθοδική συμπεριφορά, δηλαδή αντιστρόφως ανάλογη τον αριθμό των folds.



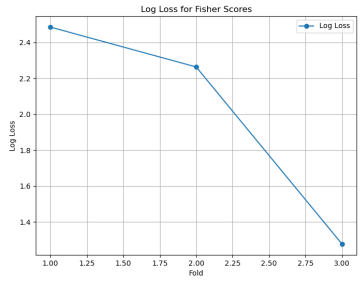
(α′) Correlation



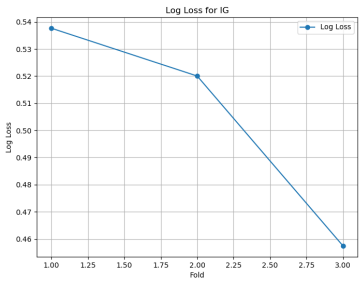
(β′) Decision Trees



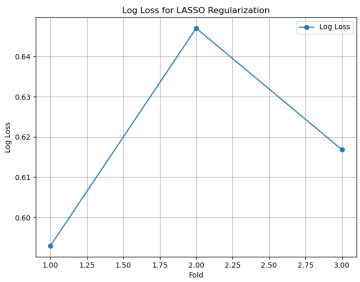
(γ′) Forward Feature Selection



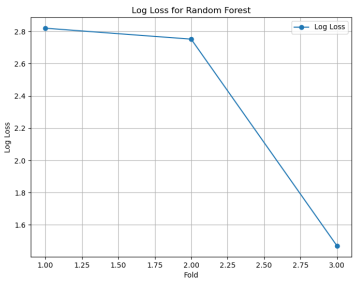
(δ′) Fisher Scores



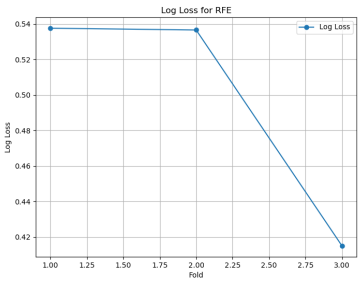
(ε′) Information Gain



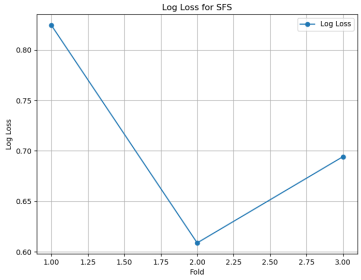
(ϛ′) Lasso Regularization



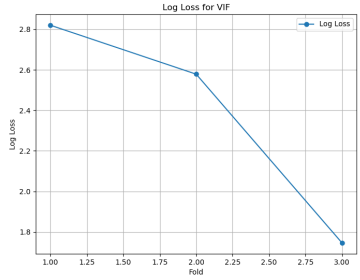
(ζ′) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection



(ι′) Variance Inflation Factor

Σχήμα 4.12: Log-Loss ως προς τον αριθμό των folds συναρτήσεις για όλες τις μεθόδους για το μοντέλο Gradient Boosting

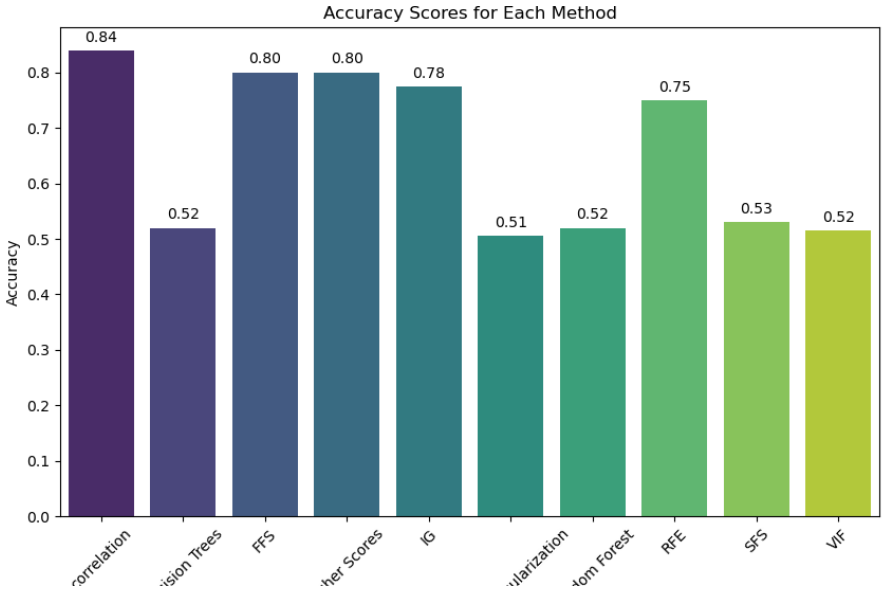
4.5 K-Nearest Neighbors

Για το μοντέλο αυτό πάλι πραγματοποιήθηκε κατά την εκπαίδευση του έλεγχος προκειμένου να βρεθεί ο βέλτιστος συνδυασμός υπερπαραμέτρων για κάθε μέθοδο. Από τη διαδικασία αυτή, προέκυψαν οι ακόλουθες αρχιτεκτονικές:

Method	Algorithm	Leaf Size	N Neighbors	P	Weights
Correlation	ball_tree	40	5	1	distance
Decision Trees	ball_tree	20	3	1	uniform
Forward Feature Selection	ball_tree	20	8	1	uniform
Fisher Scores	auto	20	8	1	uniform
Information Gain	auto	20	9	2	distance
LASSO L1 Regularization	auto	20	3	1	uniform
Random Forest	auto	20	3	1	uniform
Recursive Feature Elimination	brute	20	3	2	uniform
Sequential Feature Selection	auto	20	3	1	distance
Variance Inflation Factor	auto	20	3	1	uniform

Πίνακας 4.9: Υπερπαράμετροι για κάθε μέθοδο

Οι ακρίβειες των μοντέλων αυτών παρατίθενται στη συνέχεια:



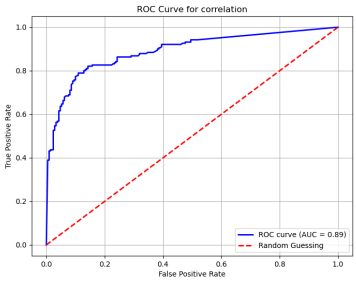
Σχήμα 4.13: Οι ακρίβειες του μοντέλου KNN για κάθε μία μέθοδο

Από το διάγραμμα υποδηλώνεται ότι στη πλειοψηφία των περιπτώσεων, ο αλγόριθμος Κ κοντινότεροι γείτονες είχε αρκετά ικανοποιητικές ακρίβειες αλλά και σε μερικές περιπτώσεις αρκετά χαμηλές.

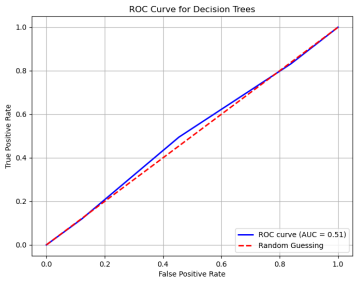
Προκειμένου να αναλυθεί περεταίρω η συμπεριφορά του μοντέλου με κάθε μέθοδο, συμπεριλαμβάνονται και οι καμπύλες ROC.

Από τα διαγράμματα αυτά είναι εύκολο να αξιολογηθεί η ικανότητα του μοντέλου ταξινόμησης με τη βοήθεια της υπολογισμένης AUC βαθμολογίας.

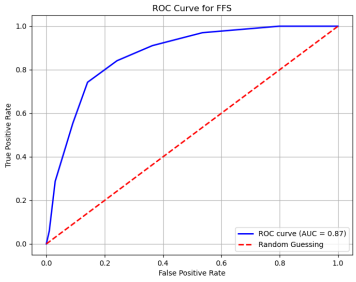
Στη συνέχεια, ακολουθούν και οι μήτρες σύγχυσης και οι υπόλοιπες μετρι-



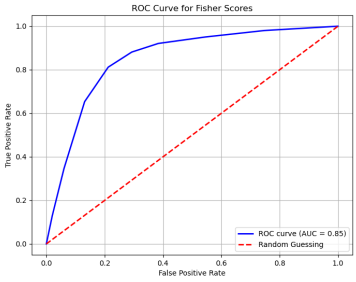
(α′) Correlation



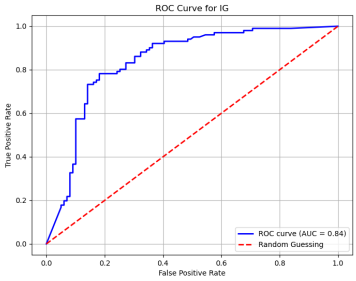
(β′) Decision Trees



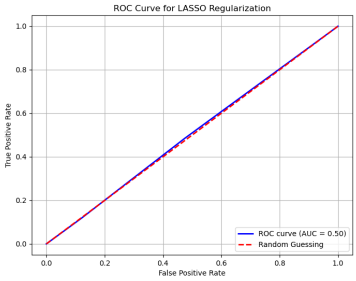
(γ′) Forward Feature Selection



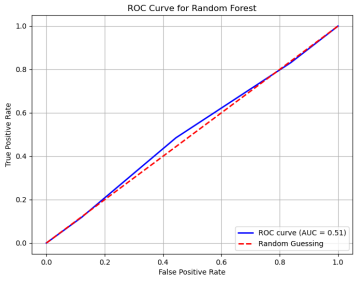
(δ′) Fisher Scores



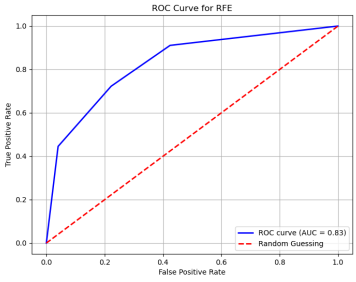
(ε′) Information Gain



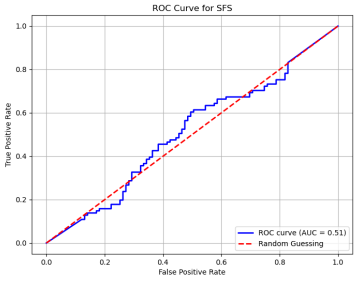
(ϛ′) Lasso Regularization



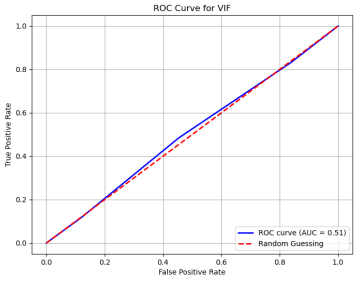
(ζ) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection



(ι′) Variance Inflation Factor

Σχήμα 4.14: ROC Curves για όλες τις μεθόδους για το μοντέλο K-Nearest Neighbors

Method	AUC Score
Correlation	0.89
Decision Trees	0.51
Forward Feature Selection	0.87
Fisher Scores	0.85
Information Gain	0.84
Lasso Regularization	0.50
Random Forest	0.51
Recursive Feature Elimination	0.83
Sequential Feature Selection	0.51
Variance Inflation Factor	0.51

Πίνακας 4.10: AUC Scores για κάθε μέθοδο

κές για την καλύτερη και αναλυτικότερη αξιολόγηση του μοντέλου αυτού σε συνδυασμό με τις διαφορετικές μεθόδους επιλογής χαρακτηριστικών.

Βάσει των παραπάνω, λοιπόν, και τις στατιστικές μετρικές, υπάρχουν οι ακόλουθες πληροφορίες για το κάθε μοντέλο:

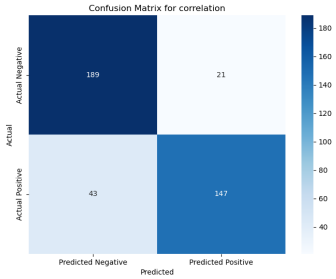
Method	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
Correlation	84%	84.5%	83.7%	83.8%
Decision Trees	52%	52.0%	52.0%	51.9%
Forward Feature Selection	80%	80.4%	80.1%	80.0%
Fisher Scores	80%	80.0%	80.0%	80.0%
Information Gain	77.5%	77.8%	77.4%	77.4%
Lasso Regularization	51%	50.5%	50.5%	50.5%
Random Forest	52%	52.0%	52.0%	51.9%
Recursive Feature Elimination	75%	75.1%	75.0%	75.0%
Sequential Feature Selection	53%	53.0%	53.0%	53.0%
Variance Inflation Factor	51.5%	51.5%	51.5%	51.5%

Πίνακας 4.11: Μετρικές Απόδοσης του μοντέλου για κάθε μέθοδο

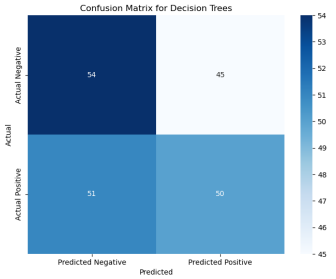
Ακολουθούν και οι γραφικές παραστάσεις των log-loss συναρτήσεων για να αξιολογηθεί και η συμπεριφορά τνω μοντέλων για διαφορετικό αριθμό folds που αντιστοιχούν στον άξονα x'x.

Αυτό που είναι αξιοσημείωτο να αναφερθεί είναι ότι το μοντέλο αυτό για κάθε μέθοδο επιλογής χαρακτηριστικών παρουσιάζει παρόμοιες συναρτήσεις log-loss καθώς φαίνονται να είναι γραμμικές. Κατα τη πλειψηφία των μεθόδων, είναι γραμμικές αντιστρόφως ανάλογες ως προς τον αριθμό των folds και για τις δύο μεθόδους είναι πάλι γραμμική η συνάρτηση αλλά ανάλογη του αριθμού των folds.

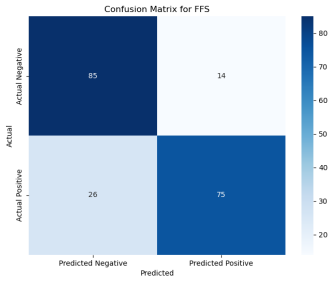
Συνολικά, το μοντέλο για μερικές μεθόδους επιτυγχάνει υψηλή απόδοση ενώ για μερικές άλλες αρκετά χαμηλή καθώς φαίνεται να έχει παρόμοια συμπεριφορά με το random guessing σε εκείνες που έχουν πολύ χαμηλή ακρίβεια (δηλαδή κοντά στο 50%).



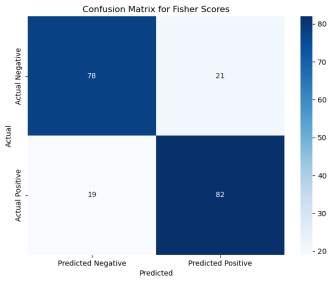
(α′) Correlation



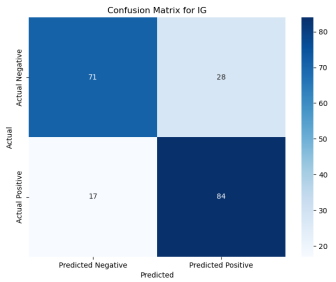
(β′) Decision Trees



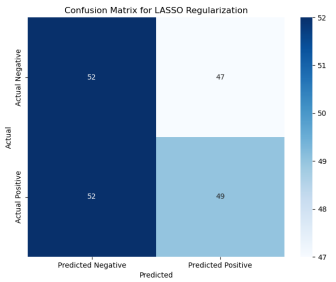
(γ′) Forward Feature Selection



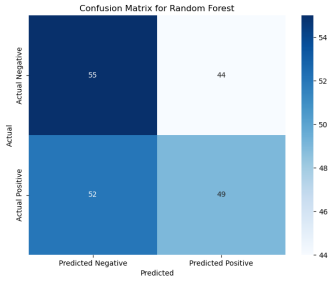
(δ′) Fisher Scores



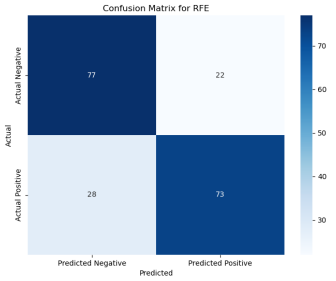
(ε′) Information Gain



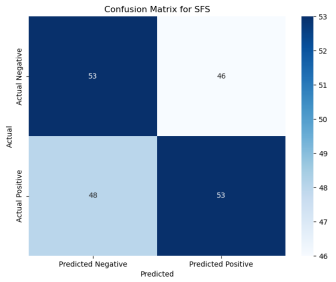
(ϛ′) Lasso Regularization



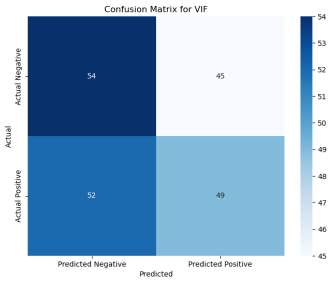
(ζ′) Random Forest



(η′) Recursive Feature Elimination

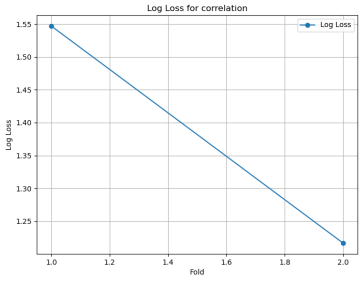


(θ′) Sequential Feature Selection

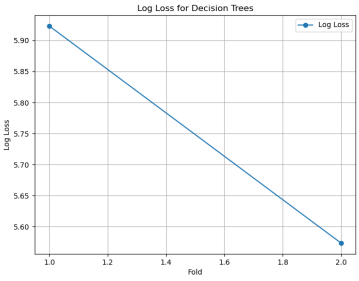


(ι′) Variance Inflation Factor

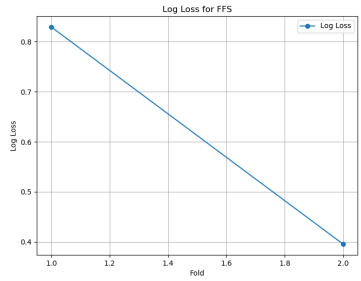
Σχήμα 4.15: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο K-Nearest Neighbors



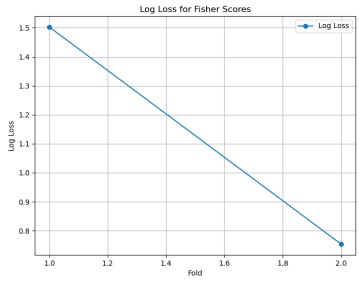
(α΄) Correlation



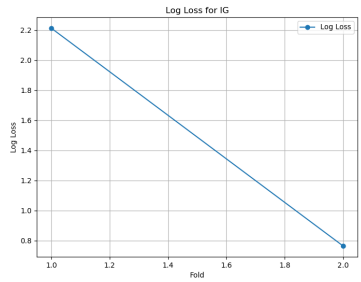
(β΄) Decision Trees



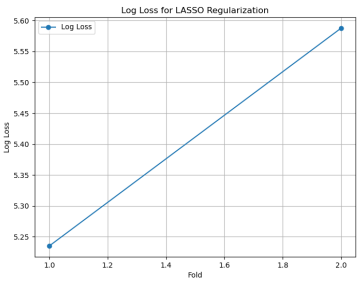
(γ΄) Forward Feature Selection



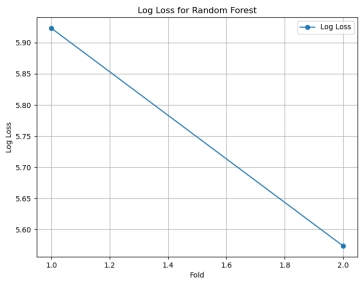
(δ΄) Fisher Scores



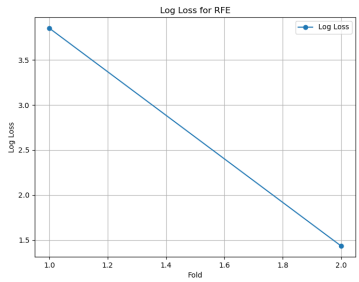
(ε΄) Information Gain



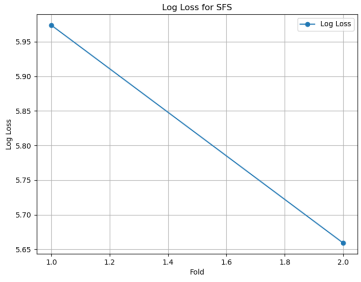
(ϛ΄) Lasso Regularization



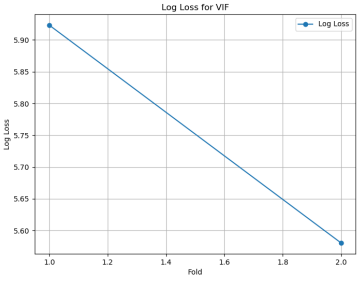
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination



(θ΄) Sequential Feature Selection



(ι΄) Variance Inflation Factor

Σχήμα 4.16: Log-Loss ως προς τον αριθμό των folds συναρτήσεις για όλες τις μεθόδους για το μοντέλο K-Nearest Neighbors

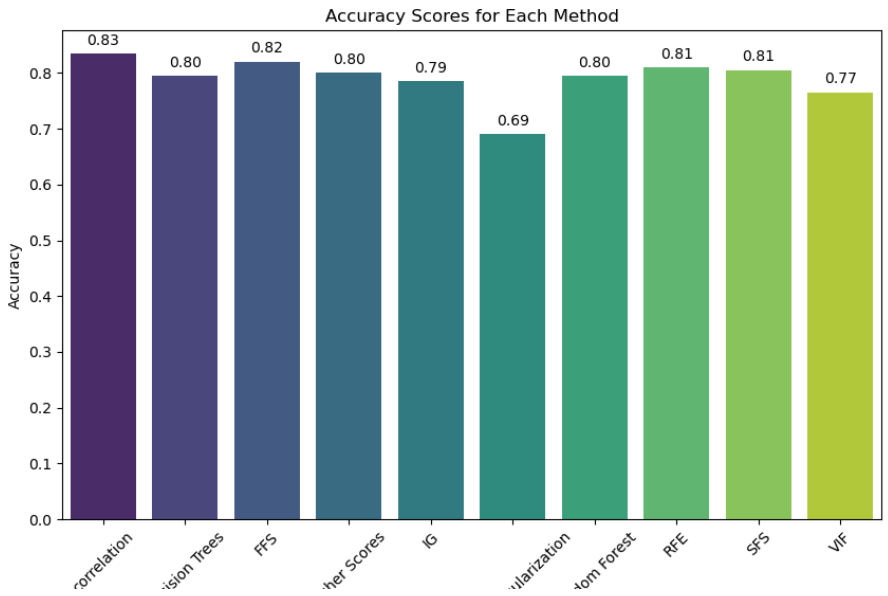
4.6 Penalized Logistic Regression

Οι παρακάτω παράμετροι αντιστοιχούν στις βέλτιστες ρυθμίσεις που προέκυψαν από την διαδικασία Grid Search για κάθε μέθοδο επιλογής χαρακτηριστικών.

Method	N Estimators	Learning Rate	Max Depth	Subsample	Min Samples Split	Min Samples Leaf
Correlation	200	0.1	5	0.9	5	2
Decision Trees	300	0.01	10	0.8	2	1
FFS	100	0.5	3	1.0	10	4
Fisher Scores	200	0.01	5	0.8	5	2
IG	150	0.1	4	0.85	4	2
LASSO Regularization	250	0.05	6	0.9	6	3
Random Forest	300	0.01	10	0.8	2	1
RFE	200	0.1	5	0.9	5	2
SFS	150	0.2	4	0.85	4	2
VIF	250	0.05	6	0.9	6	3

Πίνακας 4.12: Υπερπαράμετροι για κάθε μέθοδο

Οι ακρίβειες του μοντέλου αυτού απεικονίζονται ακολούθως:



Σχήμα 4.17: Οι ακρίβειες του μοντέλου Penalized Logistic Regression για κάθε μία μέθοδο

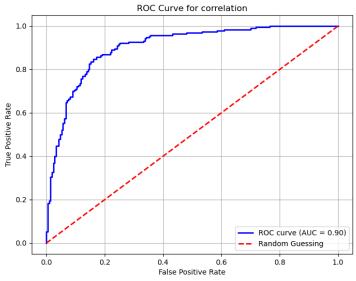
Παρατίθενται, επίσης, και οι καμπύλες ROC για κάθε μέθοδο:

Από τις καμπύλες αυτές και τη βοήθεια της βαθμολογίας AUC προέκυψαν τα εξής αποτελέσματα:

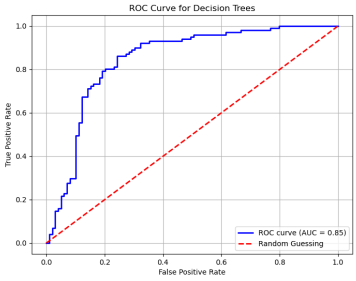
Οι μήτρες σύγχυσης έχουν ως εξής: Από τις μήτρες και τις μετρικές που υπολόγισε το μοντέλο όσον αφορά την επίδοση του με κάθε μέθοδο, συνεχίζει η αξιολόγηση του ως εξής:

Από τα παραπάνω αποφαίνεται ότι το μοντέλο αυτό μμε μικρή εξάιρεση με τη μέθοδο Lasso Regularization έχει πολύ ικανοποιητική ικανότητα ταξινόμησης και διάκρισης των θετικών και αντίστοιχα αρνητικών περιστατικών.

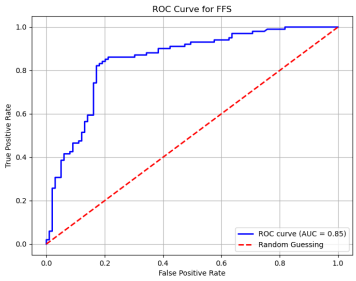
Στη συνέχεια, υπάρχουν και οι αναπαραστάσεις των συναρτήσεων log-loss για την περεταίρω αξιολόγηση και ερμήνευση της συμπεριφοράς του μοντέλου με



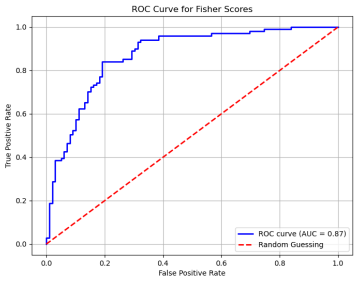
(α΄) Correlation



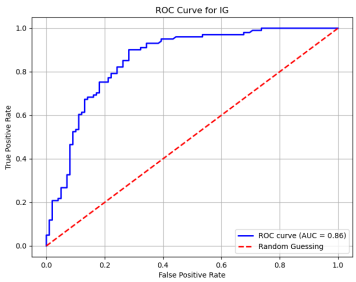
(β΄) Decision Trees



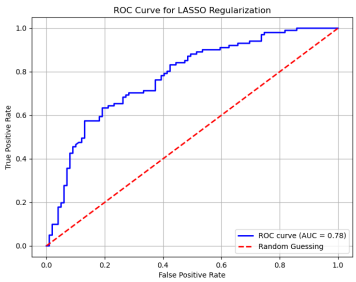
(γ΄) Forward Feature Selection



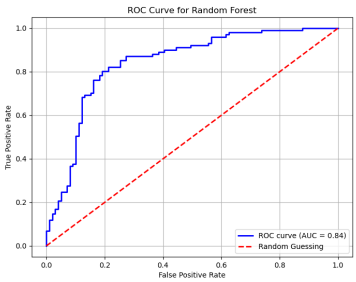
(δ΄) Fisher Scores



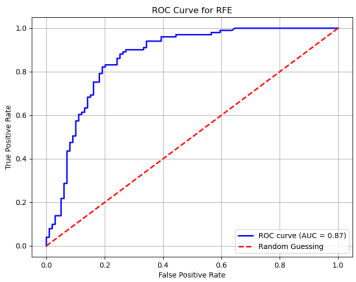
(ε΄) Information Gain



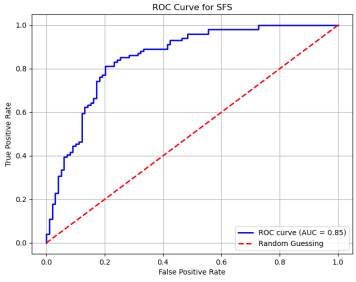
(ϛ΄) Lasso Regularization



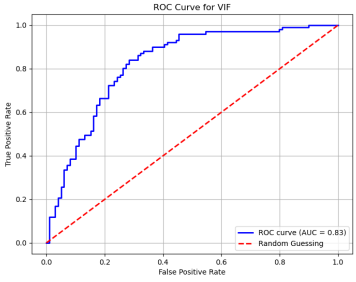
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination

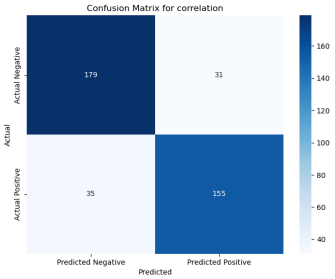


(θ΄) Sequential Feature Selection

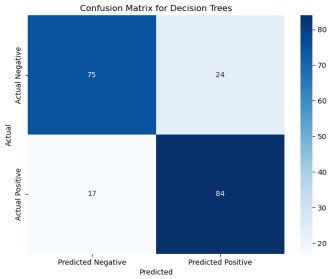


(ι΄) Variance Inflation Factor

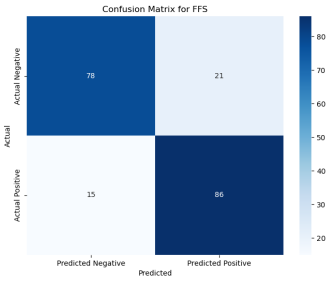
Σχήμα 4.18: ROC Curves για όλες τις μεθόδους για το μοντέλο Penalized Logistic Regression



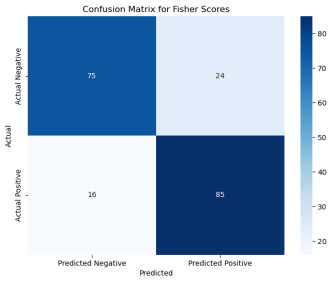
(α′) Correlation



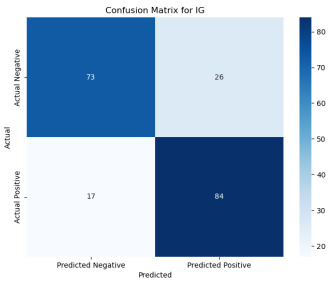
(β′) Decision Trees



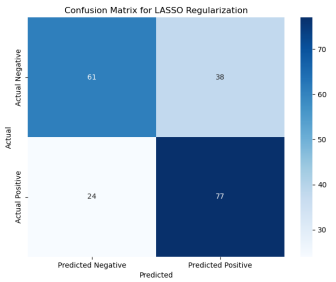
(γ′) Forward Feature Selection



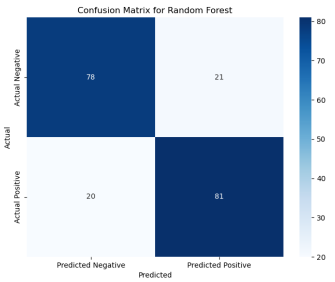
(δ′) Fisher Scores



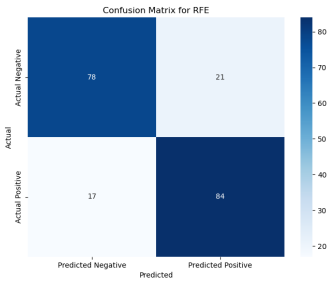
(ε′) Information Gain



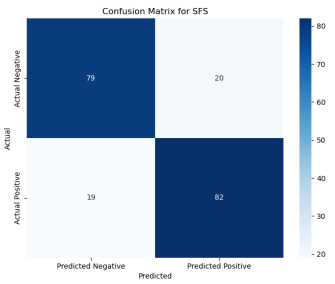
(ϛ′) Lasso Regularization



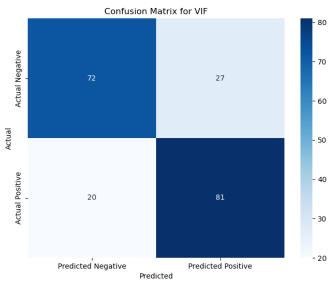
(ζ′) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection



(ι′) Variance Inflation Factor

Σχήμα 4.19: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Penalized Logistic Regression

Method	AUC Score
Correlation	0.90
Decision Trees	0.85
Forward Feature Selection	0.85
Fisher Scores	0.87
Information Gain	0.86
Lasso Regularization	0.78
Random Forest	0.84
Recursive Feature Elimination	0.87
Sequential Feature Selection	0.85
Variance Inflation Factor	0.83

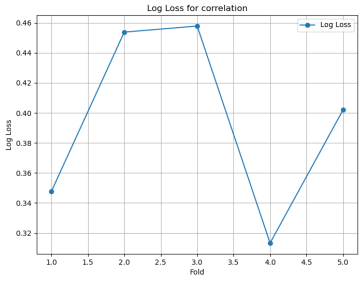
Πίνακας 4.13: AUC Scores για κάθε μέθοδο

Method	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
Correlation	83.5%	83.5%	83.4%	83.4%
Decision Trees	79.5%	83.2%	82.8%	82.9%
Forward Feature Selection	82%	82.1%	81.9%	81.9%
Fisher Scores	80%	80.2%	80.0%	80.0%
Information Gain	78.5%	78.7%	78.5%	78.4%
Lasso Regularization	69%	69.4%	68.9%	68.8%
Random Forest	79.5%	79.5%	79.5%	79.5%
Recursive Feature Elimination	81%	81.1%	81.0%	81.0%
Sequential Feature Selection	80.5%	80.5%	80.5%	80.5%
Variance Inflation Factor	76.5%	76.6%	76.5%	76.4%

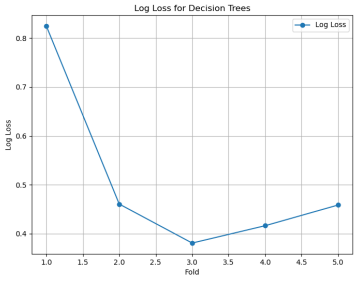
Πίνακας 4.14: Μετρικές Απόδοσης για κάθε μέθοδο

κάθε μέθοδο επιλογής χαρακτηριστικών, πάλι συναρτήσει του αριθμού των folds

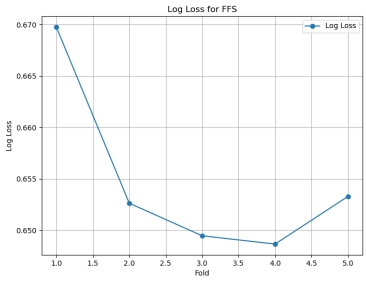
Σε αυτό το σημείο μπορεί να σχολιαστεί ότι στις περισσότερες περιπτώσεις παρουσιάζει παρόμοιες συναρτήσεις log-loss με ελάχιστες εξαιρέσεις όπως στην περίπτωση του correlation, Lasso Regularization, Sequential Feature Selection και Variance Inflation Factor.



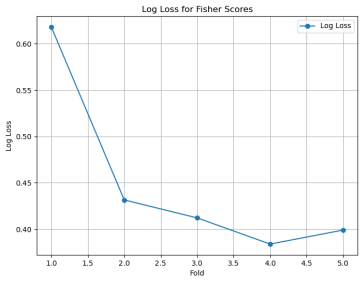
(α′) Correlation



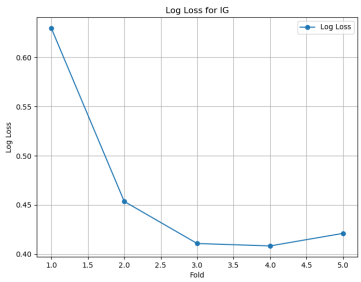
(β′) Decision Trees



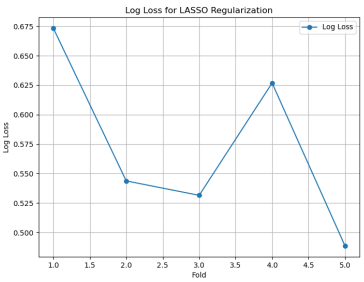
(γ′) Forward Feature Selection



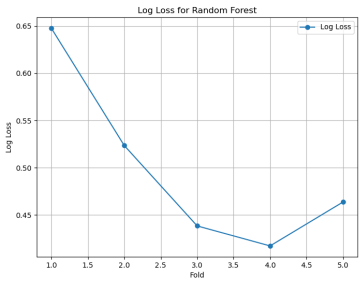
(δ′) Fisher Scores



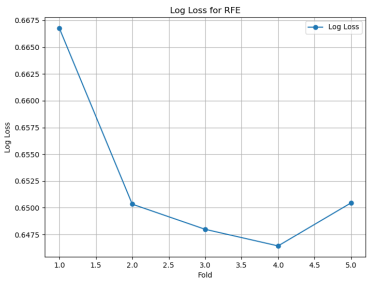
(ε′) Information Gain



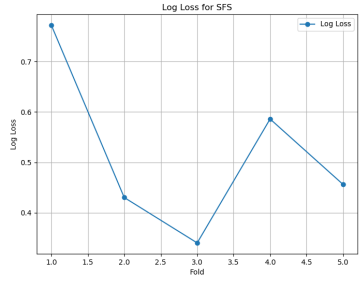
(ϛ′) Lasso Regularization



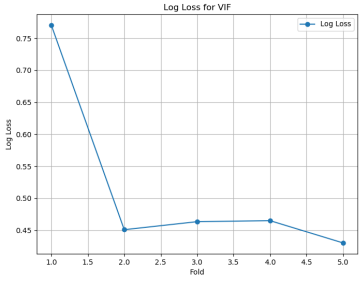
(ζ′) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection

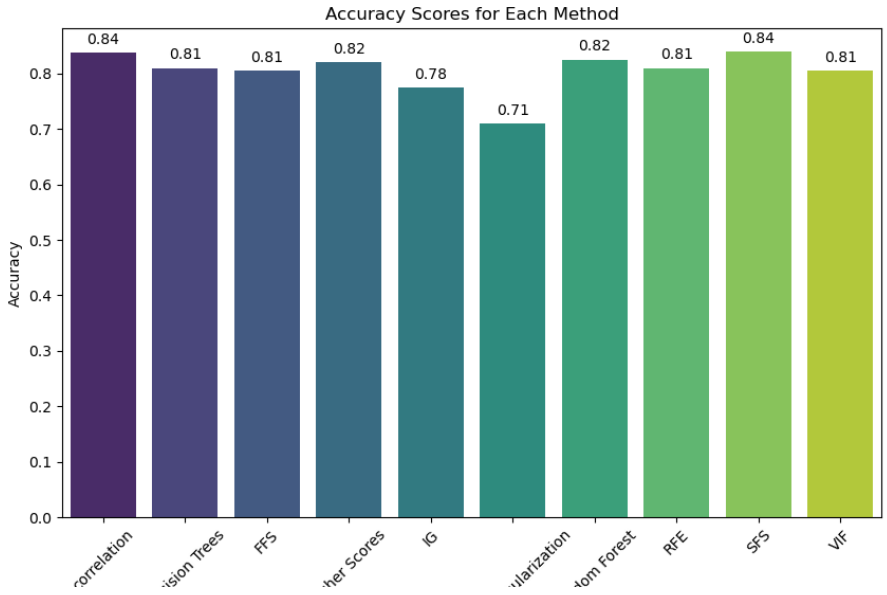


(ι′) Variance Inflation Factor

Σχήμα 4.20: Log-Loss ως προς τον αριθμό των folds συναρτήσεις για όλες τις μεθόδους για το μοντέλο Penalized Logistic Regression

4.7 Γραμμική Παλινδρόμηση

Στην γραμμική παλινδρόμηση δεν προηγήθηκε η εύρεση του βέλτιστου συνδυασμού υπερπαραμέτρων. Αυτό συμβαίνει γιατί η γραμμική παλινδρόμηση αυτόματα υπολογίζει το καλύτερο ταίριασμα ελαχιστοποιώντας το άθροισμα των τετραγωνικών σφαλμάτων. Οπότε, απευθείας για την αξιολόγηση του μοντέλου, απεικονίζονται παρακάτω οι ακρίβειες που πέτυχε για κάθε μέθοδο:



Σχήμα 4.21: Οι ακρίβειες του μοντέλου Linear Regression για κάθε μία μέθοδο

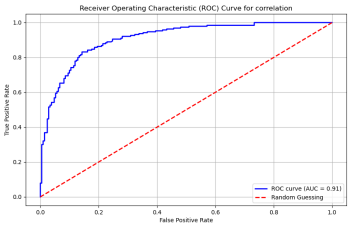
Από μία ματιά είναι εμφανές ότι όλες οι μέθοδοι επιτυγχάνουν υψηλές τιμές ακρίβειας όπου οι περισσότερες κυμαίνονται στο 78% με 84%.

Στη συνέχεια, ακολουθούν και οι γραφικές παραστάσεις των καμπύλων ROC μαζί με την υπολογισμένη βαθμολογία AUC για παραπάνω ανάλυση της κάθε εκδοχής του μοντέλου.

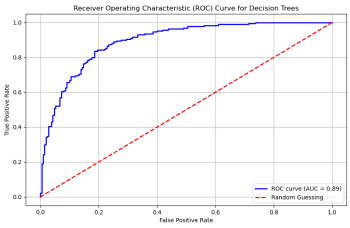
Από τα διαγράμματα αυτά δίνονται οι ακόλουθες πληροφορίες όσον αφορά τη βαθμολογία AUC:

Method	AUC Score
Correlation	0.91
Decision Trees	0.89
Forward Feature Selection	0.84
Fisher Scores	0.87
Information Gain	0.85
Lasso Regularization	0.74
Random Forest	0.83
Recursive Feature Elimination	0.85
Sequential Feature Selection	0.82
Variance Inflation Factor	0.78

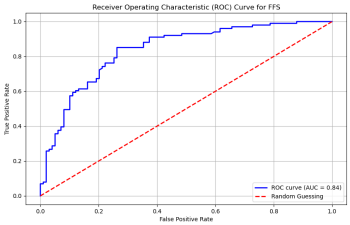
Πίνακας 4.15: ΑΥΣ Σχορες φορ Εαχη Μετηοδ



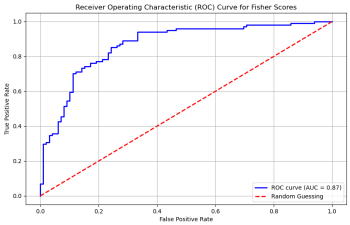
(α΄) Correlation



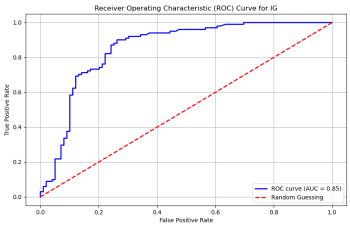
(β΄) Decision Trees



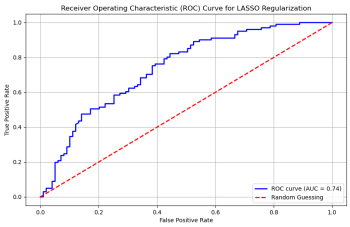
(γ΄) Forward Feature Selection



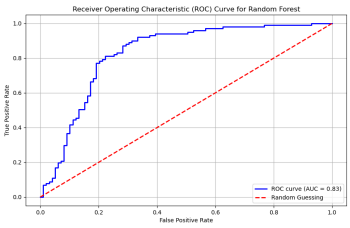
(δ΄) Fisher Scores



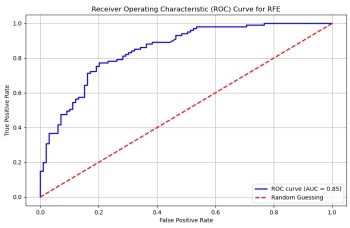
(ε΄) Information Gain



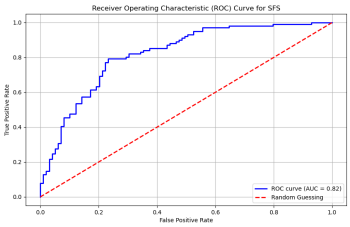
(ϛ΄) Lasso Regularization



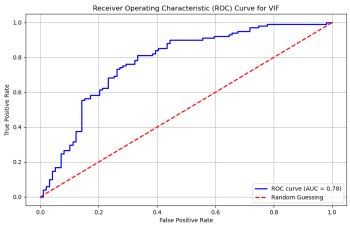
(ζ) Random Forest



(η΄) Recursive Feature Elimination



(θ΄) Sequential Feature Selection



(ι΄) Variance Inflation Factor

Σχήμα 4.22: ROC Curves για όλες τις μεθόδους για το μοντέλο Linear Regression

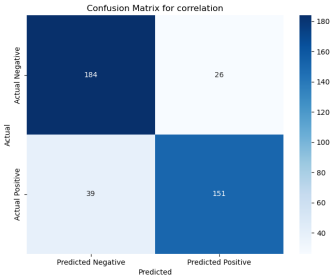
Τα διαγράμματα αυτά συνοδεύονται από τις μήτρες σύγχυσης για κάθε μέθοδο: Από τα οποία απορρέουν τα ακόλουθα:

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Correlation	83.75	83.91	83.55	83.64
Decision Trees	83.00	83.04	82.86	82.92
Forward Feature Selection	80.50	80.58	80.47	80.48
Fisher Scores	82.00	82.01	82.00	82.00
Information Gain	77.50	77.57	77.47	77.47
Lasso Regularization	71.00	71.41	70.93	70.81
Random Forest	82.50	82.50	82.50	82.50
Recursive Feature Elimination	81.00	81.01	80.99	80.99
Sequential Feature Selection	84.00	84.00	84.00	84.00
Variance Inflation Factor	80.50	80.58	80.47	80.48

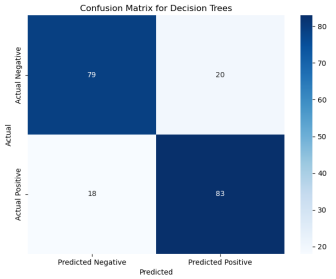
Πίνακας 4.16: Μετρικές Απόδοσης για κάθε μέθοδο

Τέλος, παρουσιάζονται και οι γραφικές παραστάσεις των συναρτήσεων log-loss της κάθε μεθόδου που υπολογίζεται συναρτήσει τον αριθμό των επαναλήψεων (iterations).

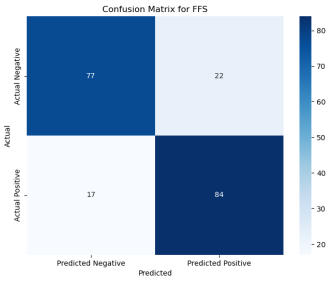
Αυτό που μπορεί να παρατηρηθεί από τα διαγράμματα log-loss είναι ότι για όλες τις μεθόδους, το μοντέλο μειώνει σημαντικά το σφάλμα όσο αυξάνονται οι επαναλήψεις. Συγκεκριμένα, στις 2000 επαναλήψεις σημειώνεται η πτώση του log-loss για όλες τις μεθόδους, όπου και επιτυγχάνει το ελάχιστο σφάλμα.



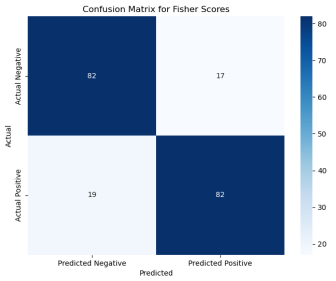
(α΄) Correlation



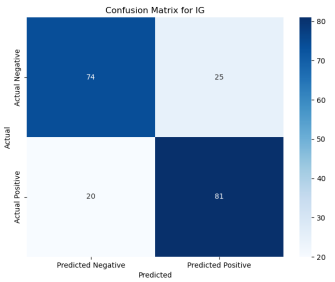
(β΄) Decision Trees



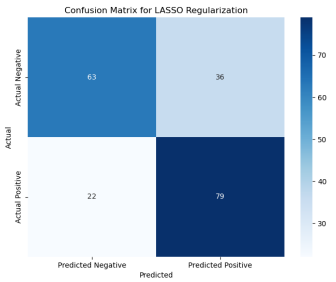
(γ΄) Forward Feature Selection



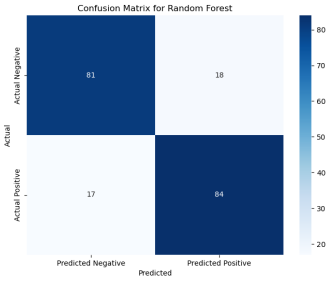
(δ΄) Fisher Scores



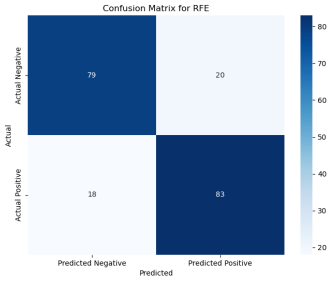
(ε΄) Information Gain



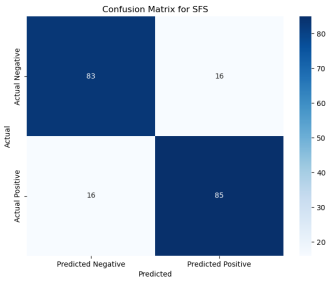
(ϛ΄) Lasso Regularization



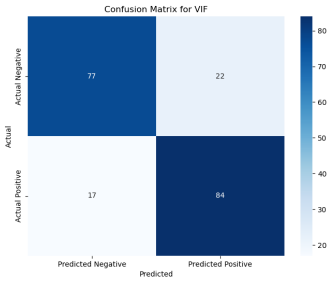
(ζ) Random Forest



(η΄) Recursive Feature Elimination

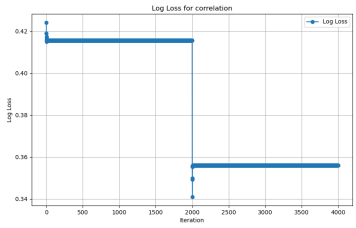


(θ΄) Sequential Feature Selection

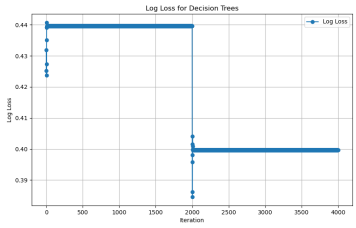


(ι΄) Variance Inflation Factor

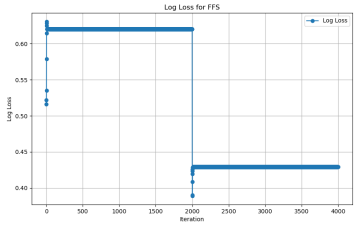
Σχήμα 4.23: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Linear Regression



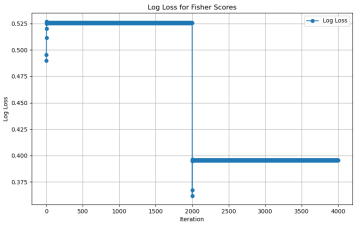
(α΄) Correlation



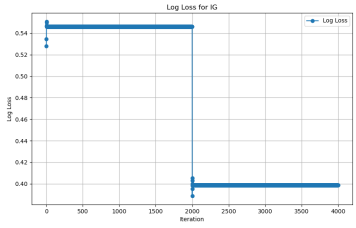
(β΄) Decision Trees



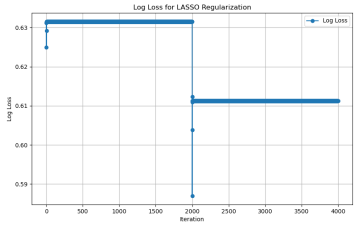
(γ΄) Forward Feature Selection



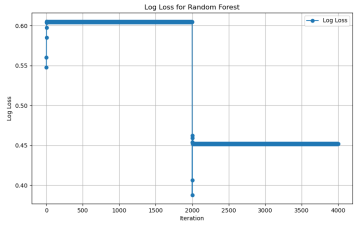
(δ΄) Fisher Scores



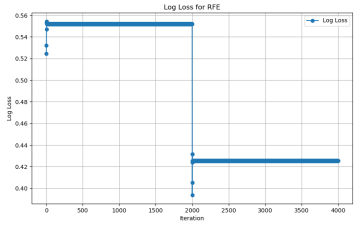
(ε΄) Information Gain



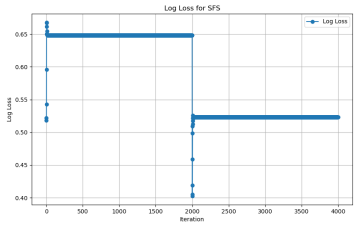
(ϛ΄) Lasso Regularization



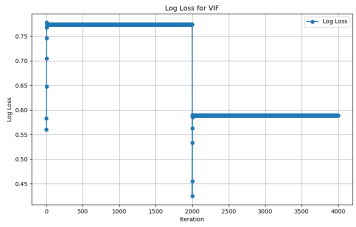
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination



(θ΄) Sequential Feature Selection



(ι΄) Variance Inflation Factor

Σχήμα 4.24: Log-Loss ως προς τον αριθμό των επαναλήψεων συναρτήσεις για όλες τις μεθόδους για το μοντέλο Linear Regression

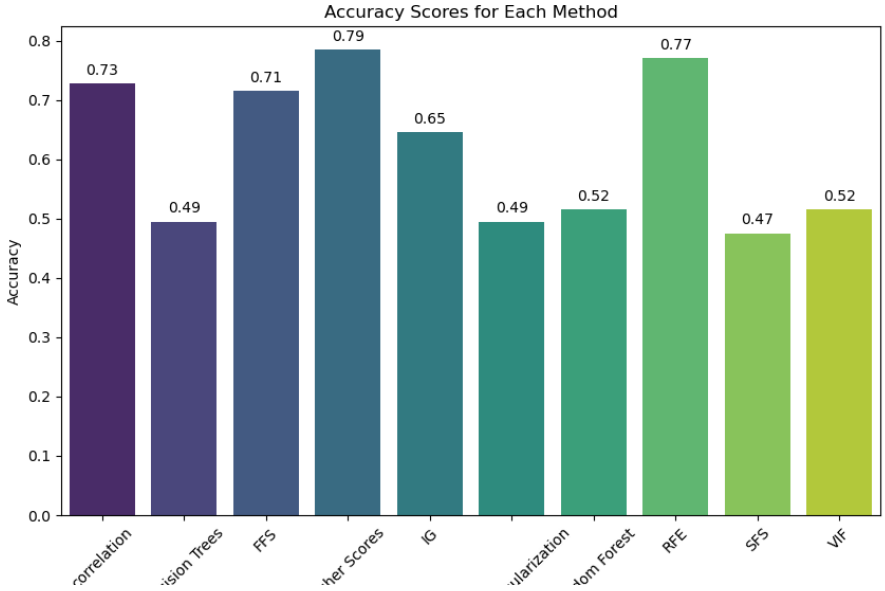
4.8 Multilayered Perceptron

Όπως σε όλα τα μοντέλα, έτσι και σε αυτό, κατά την εκπαίδευση του προηγήθηκε η επιλογή του βέλτιστου συνδυασμού υπερπαραμέτρων για κάθε μέθοδο βάσει της υψηλότερης ακρίβειας. Επομένως, για κάθε μέθοδο προέκυψε η ακόλουθη αρχιτεκτονική:

Method	Activation	Hidden Layer Size	Initial Learning Rate	Max Iterations	Solver
Correlation	relu	100	0.1	1000	adam
Decision Trees	relu	50	0.001	1000	sgd
FFS	tanh	50	0.001	1000	sgd
Fisher Scores	relu	50	0.01	1000	adam
IG	relu	50	0.1	1000	adam
LASSO Regularization	relu	50	0.01	1000	sgd
Random Forest	relu	50	0.001	1000	adam
RFE	relu	100	0.1	1000	sgd
SFS	tanh	100	0.001	1000	adam
VIF	tanh	50	0.001	1000	adam

Πίνακας 4.17: Υπερπαράμετροι για κάθε μέθοδο

Όσον αφορά τις ακρίβειες που επιτυγχάνει το μοντέλο για κάθε μέθοδο, απεικονίζονται στο εξής διάγραμμα: Από μια ματιά μόνο, είναι εμφανές ότι το μοντέλο



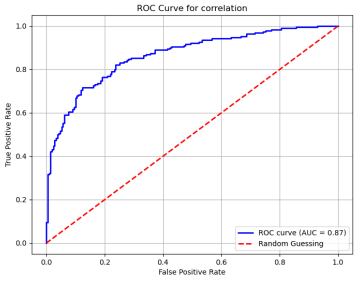
Σχήμα 4.25: Οι ακρίβειες του μοντέλου Multilayered Perceptron για κάθε μία μέθοδο

δεν έχει πετύχει πολύ ψηλές αποδόσεις κρίνοντας από τις ακρίβειες που απεικονίζονται και αυτό μπορεί να επιβεβαιωθεί αναλύοντας περεταίρω και τις καμπύλες ROC μαζί με την βαθολογία AUC.

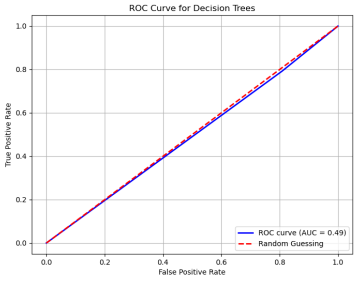
Από τις καμπύλες, προκύπτουν οι ακόλουθες βαθμολογίες AUC:

Ενώ σε μερικές μεθόδους είναι αρκετά υψηλή η βαθμολογία αυτή, στις περισσότερες είναι κοντά στο 50%.

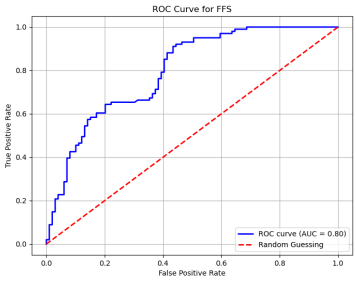
Ακολουθούν και τα διαγράμματα που απεικονίζουν. τις μήτρες σύγχυσης



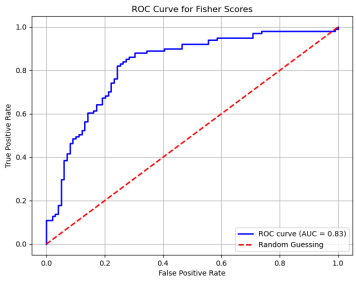
(α΄) Correlation



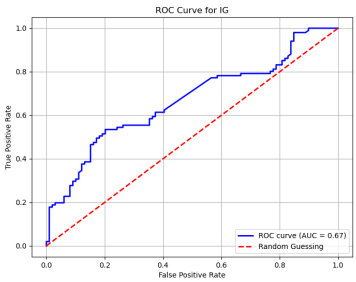
(β΄) Decision Trees



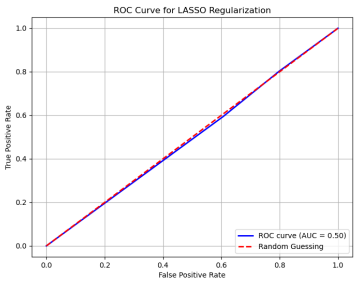
(γ΄) Forward Feature Selection



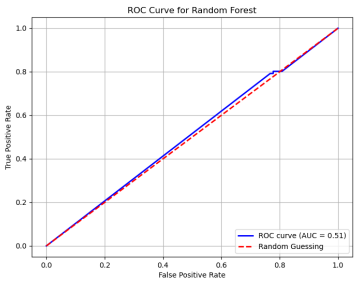
(δ΄) Fisher Scores



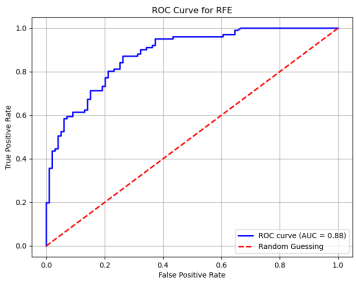
(ε΄) Information Gain



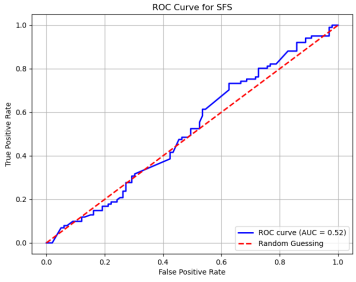
(ϛ΄) Lasso Regularization



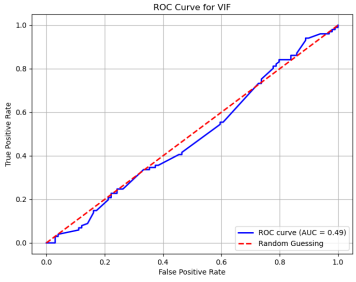
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination



(θ΄) Sequential Feature Selection



(ι΄) Variance Inflation Factor

Σχήμα 4.26: ROC Curves για όλες τις μεθόδους για το μοντέλο Multilayered Perceptron

Method	AUC Score
Correlation	0.87
Decision Trees	0.49
Forward Feature Selection	0.80
Fisher Scores	0.83
Information Gain	0.67
Lasso Regularization	0.50
Random Forest	0.51
Recursive Feature Elimination	0.88
Sequential Feature Selection	0.52
Variance Inflation Factor	0.49

Πίνακας 4.18: ΑΥΣ Σχορες φορ Εαχη Μετηοδ

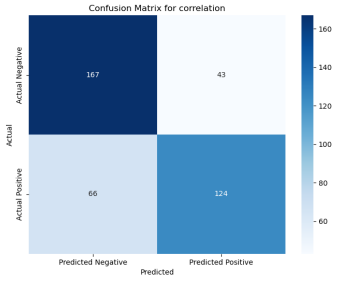
αλλά και παραπάνω πληροφορίες όσον αφορά τις υπόλοιπες στατιστικές μετρικές για την αξιολόγηση του μοντέλου. Προκύπτουν οι ακόλουθες πληροφορίες:

Method	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
Correlation	72.75%	72.96%	72.39%	72.43%
Decision Trees	49.5%	49.58%	49.60%	49.04%
Forward Feature Selection	71.5%	72.64%	71.38%	71.05%
Fisher Scores	78.5%	78.57%	78.47%	78.47%
Information Gain	64.5%	66.59%	64.67%	63.51%
LASSO Regularization	49.5%	49.38%	49.40%	48.94%
Random Forest	51.5%	51.46%	51.45%	51.35%
Recursive Feature Elimination	77%	77.10%	76.97%	76.96%
Sequential Feature Selection	47.5%	46.86%	47.30%	45.42%
Variance Inflation Factor	51.5%	51.87%	51.20%	46.53%

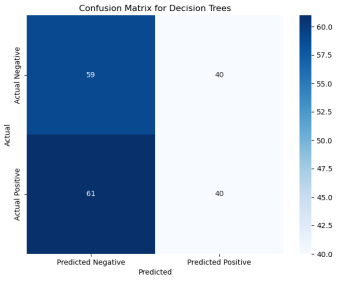
Πίνακας 4.19: Μετρικές Απόδοσης για κάθε μέθοδο

Ακόμη, απεικονίζονται και οι γραφικές παραστάσεις των log-loss συναρτήσεων προκειμένου να μπορεί να ερμηνευτεί η συμπεριφορά του μοντέλου με καθεμία από τις μεθόδους συναρτήσει των folds, τα οποία είναι 5 για αυτό το μοντέλο.

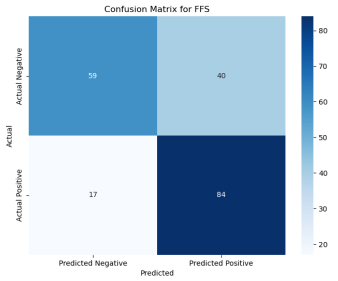
Γενικά, οι συμπεριφορές του μοντέλου με κάθε μέθοδο είναι διαφορετικές. Ωστόσο, αξίζει να σημειωθεί ότι τα μοντέλα που πέτυχαν τις υψηλότερες ακρίβειες συγκριτικά με τα υπόλοιπα, στα 5 folds έχουν χαμηλότερο σφάλμα, και στις περισσότερες περιπτώσεις έχει φθίνουσα συμπεριφορά. Η εξαίρεση είναι με το recursive feature elimination, το οποίο έχει από τις υψηλότερες επιδόσεις, αλλά φαίνεται να είναι αύξουσα η συμπεριφορά του για folds ίσα ε 5.



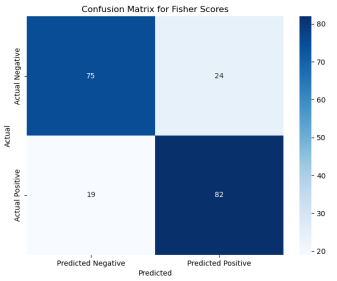
(α΄) Correlation



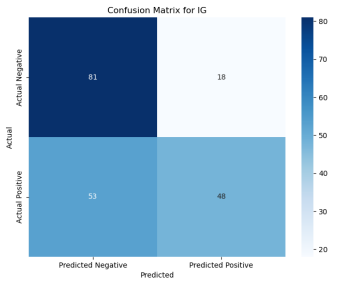
(β΄) Decision Trees



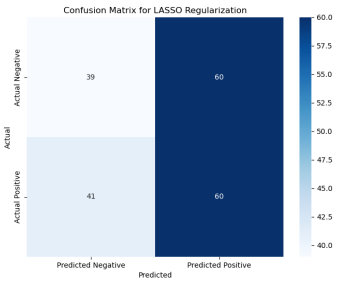
(γ΄) Forward Feature Selection



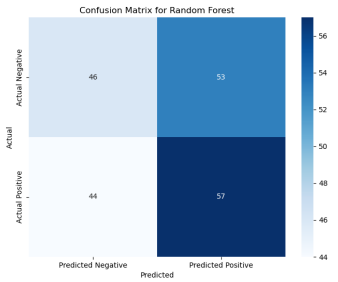
(δ΄) Fisher Scores



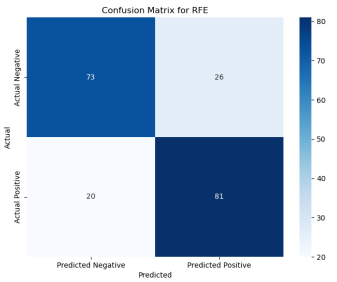
(ε΄) Information Gain



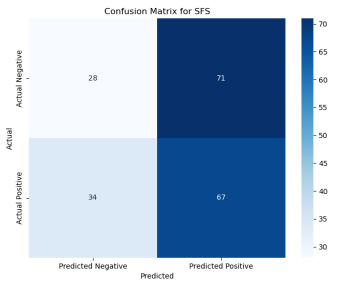
(ϛ΄) Lasso Regularization



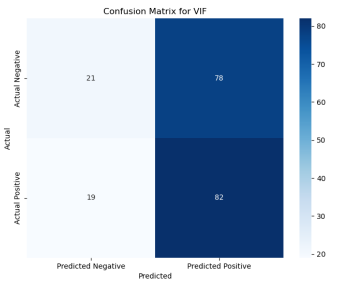
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination

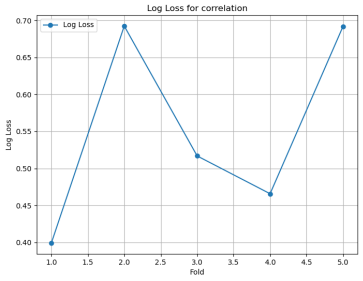


(θ΄) Sequential Feature Selection

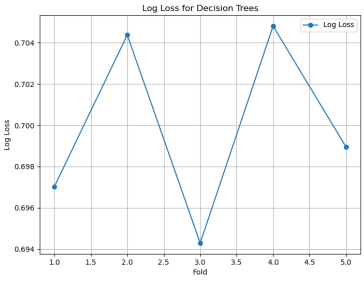


(ι΄) Variance Inflation Factor

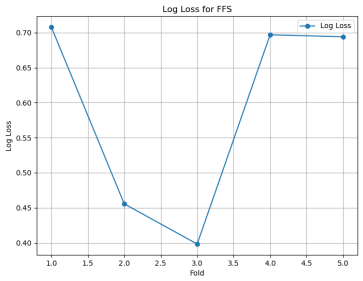
Σχήμα 4.27: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Multilayered Perceptron



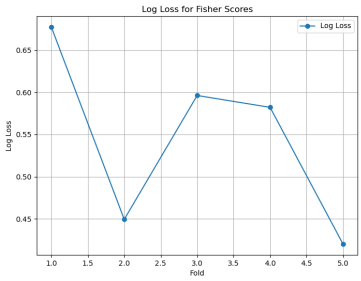
(α′) Correlation



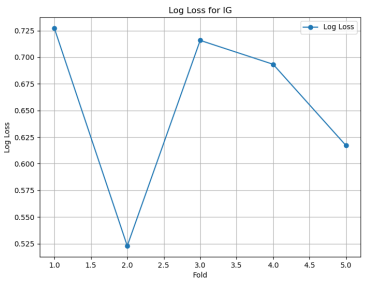
(β′) Decision Trees



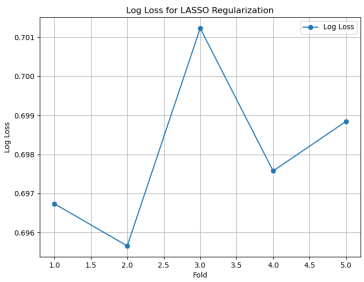
(γ′) Forward Feature Selection



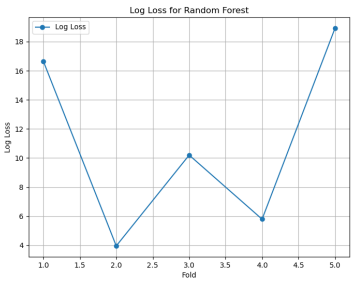
(δ′) Fisher Scores



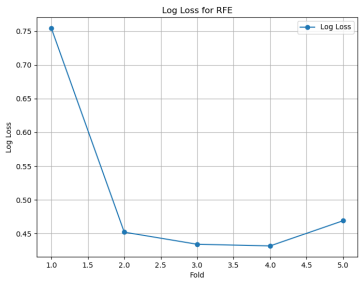
(ε′) Information Gain



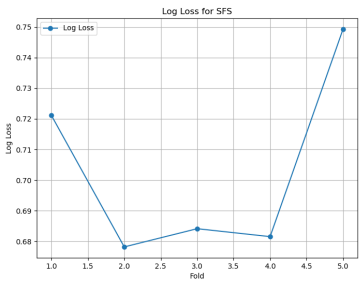
(ϛ′) Lasso Regularization



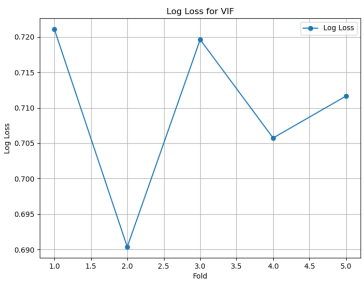
(ζ′) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection



(ι′) Variance Inflation Factor

Σχήμα 4.28: Log-Loss ως προς τον αριθμό των επαναλήψεων συναρτήσεις για όλες τις μεθόδους για το μοντέλο Multilayered Perceptron

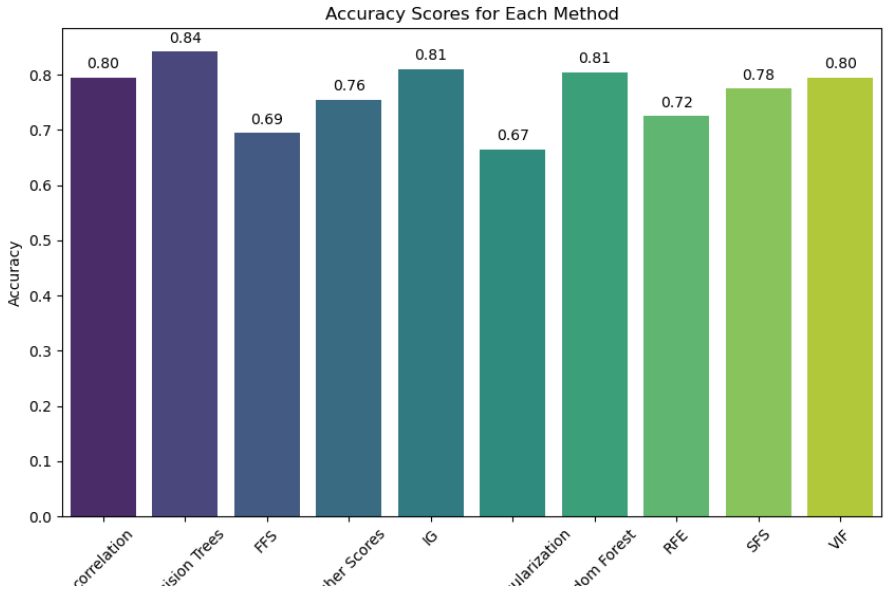
4.9 Naive Bayes

Από τη διαδικασία εκπαίδευσης, όπου επιλέχθηκε η κατάλληλη υπερπαράμετρος για αυτό το μοντέλο, η var_smoothing, προέκυψαν οι παρακάτω αρχιτεκτονικές:

Method	Var Smoothing
Correlation	1×10^{-9}
Decision Trees	1×10^{-9}
FFS	1×10^{-5}
Fisher Scores	1×10^{-7}
IG	1×10^{-9}
LASSO Regularization	1×10^{-7}
Random Forest	1×10^{-9}
RFE	1×10^{-6}
SFS	1×10^{-9}
VIF	1×10^{-8}

Πίνακας 4.20: Υπερπαράμετροι για κάθε μέθοδο

Το διάγραμμα που απεικονίζει τις συνολικές και μέγιστες ακρίβειες που πετυχαίνει το μοντέλο για όλες τις μεθόδους επιλογής χαρακτηριστικών φαίνεται παρακάτω:

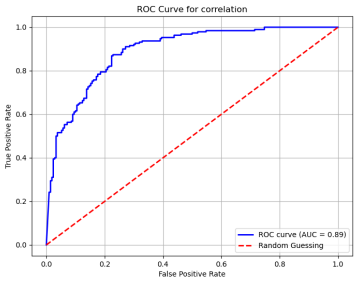


Σχήμα 4.29: Οι ακρίβειες του μοντέλου Naive Bayes για κάθε μία μέθοδο

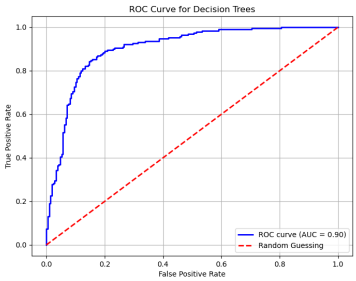
Ακολουθούν και οι καμπύλες ROC μαζί με τις βαθμολογίες AUC: Από τα οποία απορρέουν οι εξής βαθμολογίες:

και οι απεικονίσεις των μητρών σύγχυσης: Από τα οποία προκύπτουν τα εξής συμπεράσματα μαζί με τις υπόλοιπες στατιστικές μετρικές που αξιολογούν το μοντέλο.

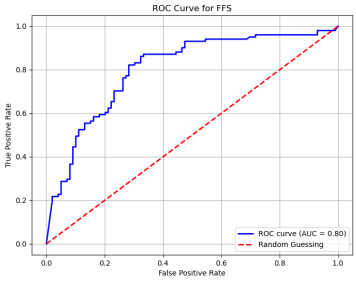
Τέλος, διαθέτονται και οι συναρτήσεις log-loss για την καλύτερη ανάλυση και ερμήνευση της συπεριφοράς του μοντέλου για κάθε διαίρεση (fold).



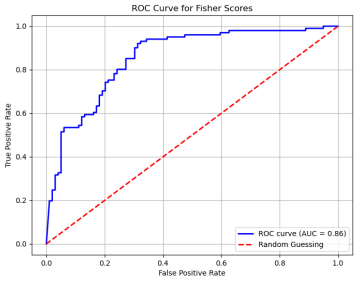
(α′) Correlation



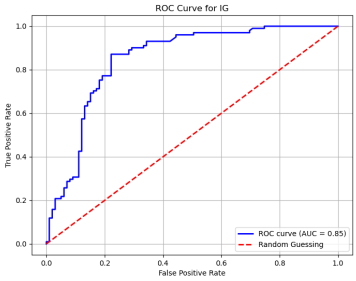
(β′) Decision Trees



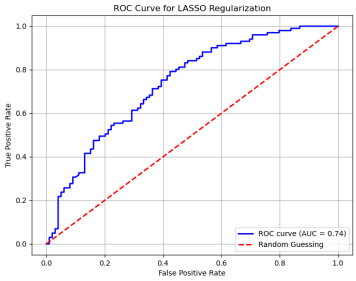
(γ′) Forward Feature Selection



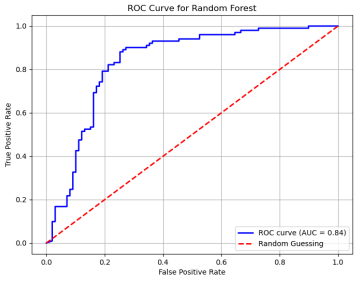
(δ′) Fisher Scores



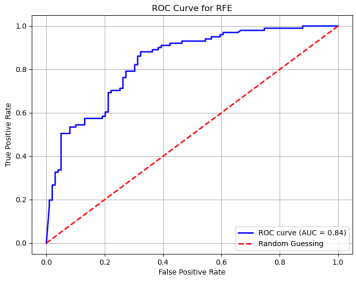
(ε′) Information Gain



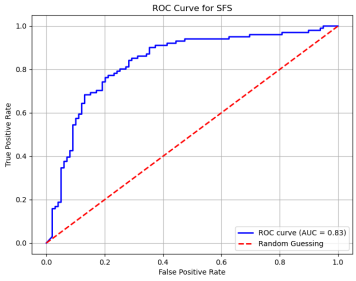
(ϛ′) Lasso Regularization



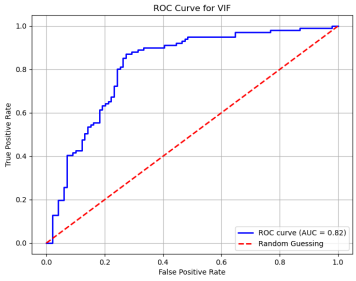
(ζ) Random Forest



(η′) Recursive Feature Elimination

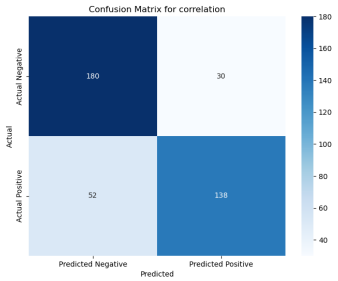


(θ′) Sequential Feature Selection

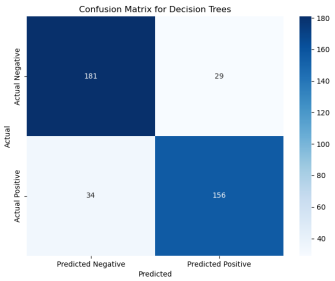


(ι′) Variance Inflation Factor

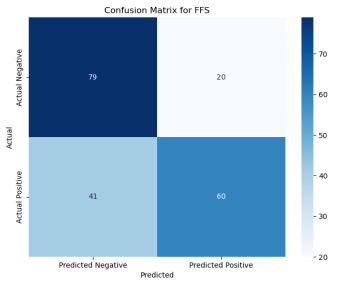
Σχήμα 4.30: ROC Curves για όλες τις μεθόδους για το μοντέλο Naive Bayes



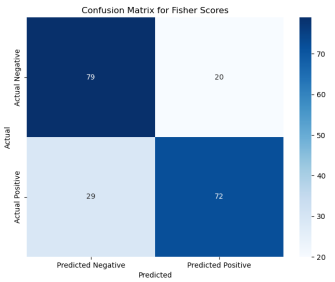
(α΄) Correlation



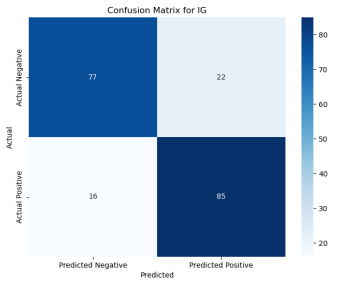
(β΄) Decision Trees



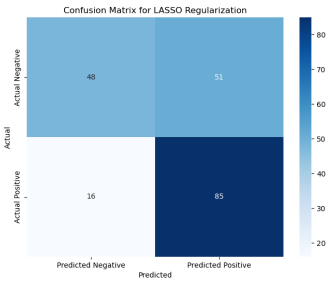
(γ΄) Forward Feature Selection



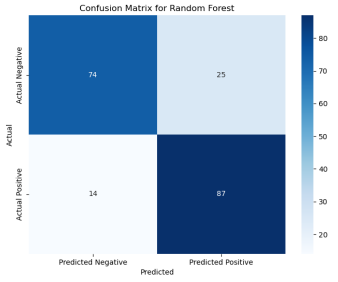
(δ΄) Fisher Scores



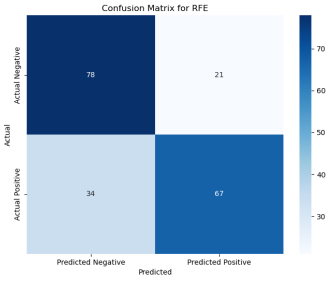
(ε΄) Information Gain



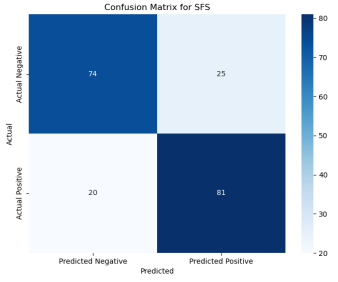
(ϛ΄) Lasso Regularization



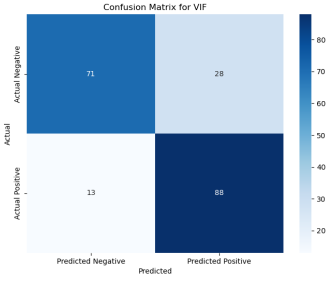
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination

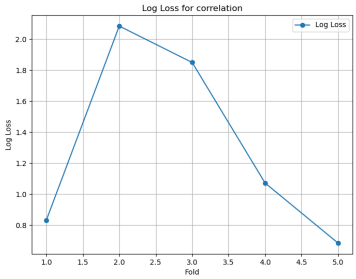


(θ΄) Sequential Feature Selection

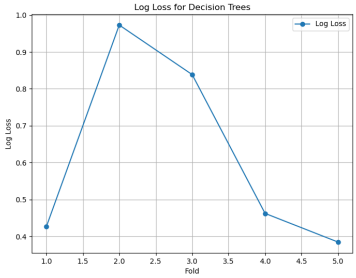


(ι΄) Variance Inflation Factor

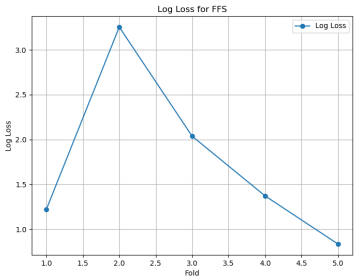
Σχήμα 4.31: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Naive Bayes



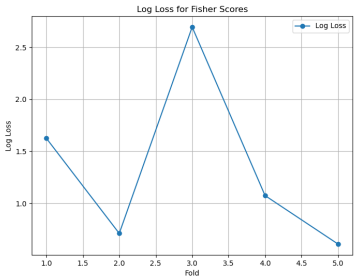
(α′) Correlation



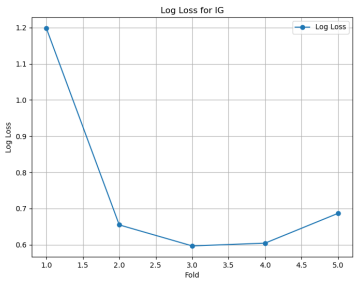
(β′) Decision Trees



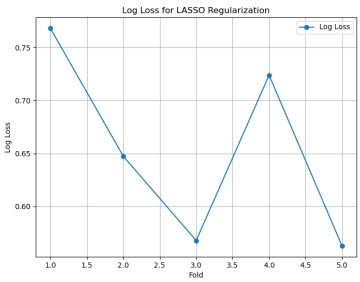
(γ′) Forward Feature Selection



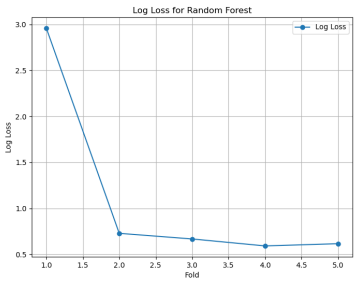
(δ′) Fisher Scores



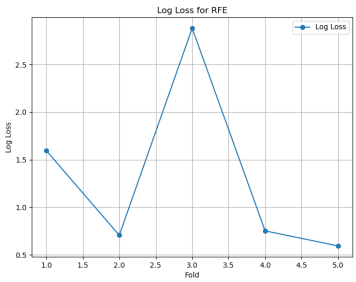
(ε′) Information Gain



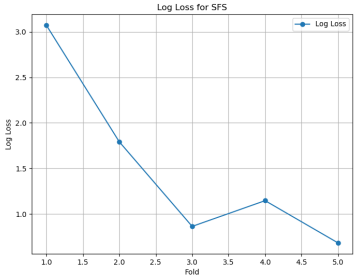
(ϛ′) Lasso Regularization



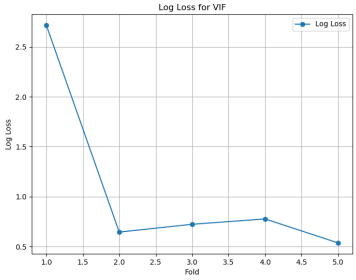
(ζ′) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection



(ι′) Variance Inflation Factor

Σχήμα 4.32: Log-Loss ως προς τον αριθμό των επαναλήψεων συναρτήσεις για όλες τις μεθόδους για το μοντέλο Naive Bayes

Method	AUC Score
Correlation	0.89
Decision Trees	0.9
Forward Feature Selection	0.8
Fisher Scores	0.86
Information Gain	0.85
Lasso Regularization	0.74
Random Forest	0.84
Recursive Feature Elimination	0.84
Sequential Feature Selection	0.83
Variance Inflation Factor	0.82

Πίνακας 4.21: AUC Scores για κάθε μέθοδο

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Correlation	79.5	79.86	79.17	79.27
Decision Trees	84.25	84.26	84.15	84.19
FFS	69.5	70.42	69.60	69.22
Fisher Scores	75.5	75.70	75.54	75.47
IG	81.0	81.12	80.98	80.97
LASSO Regularization	66.5	68.75	66.32	65.31
Random Forest	80.5	80.88	80.44	80.42
RFE	72.5	72.89	72.56	72.42
SFS	77.5	77.57	77.47	77.47
VIF	79.5	80.19	79.42	79.35

Πίνακας 4.22: Μετρικές Απόδοσης για κάθε μέθοδο

Από τα διαγράμματα αυτά, παρατηρείται από όλες τις μεθόδους, φθίνουσα συμπεριοφορά ως προς τα τελευταία folds. Παρόλο που μπορεί να σημειώεται αύξηση του log-loss σε μερικές διαιρέσεις, όλα τα μοντέλα, επιτυγχάουν ελαχιστοποίηση του log-loss, με εξαίρεση τα random forest, στην 5η διαίρεση.

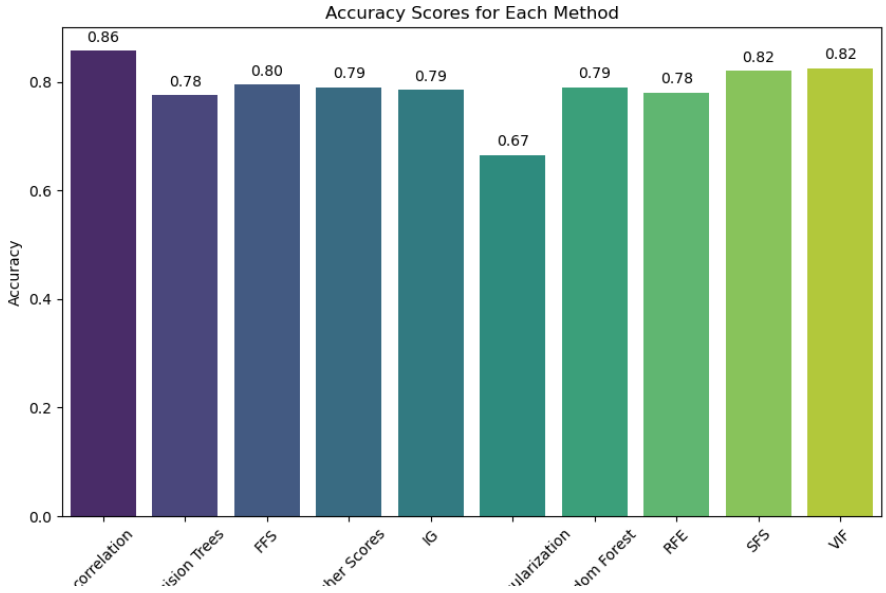
4.10 Random Forest

Για το μοντέλο αυτό κατά διαδικασία επιλογής των βέλτιστων υπερπαραμέτρων, ο αλγόριθμος επέλεξε τις ακόλουθες αρχιτεκτονικές:

Method	Criterion	Max Depth	Max Features	Min Samples Leaf	Min Samples Split	N Estimators
Correlation	gini	null	sqrt	1	2	200
Decision Trees	gini	20	sqrt	2	5	50
FFS	gini	20	sqrt	4	2	50
Fisher Scores	gini	20	log2	1	10	100
IG	entropy	5	sqrt	1	2	50
LASSO Regularization	entropy	5	sqrt	4	5	50
Random Forest	gini	5	log2	2	10	50
RFE	entropy	5	sqrt	4	5	50
SFS	entropy	5	log2	1	10	50
VIF	gini	5	log2	4	5	200

Πίνακας 4.23: Υπερπαράμετροι για κάθε μέθοδο

Οι ακρίβειες που επιτυγχάνονται από το μοντέλο για κάθε μέθοδο απεικονίζονται στο ακόλουθο διάγραμμα:



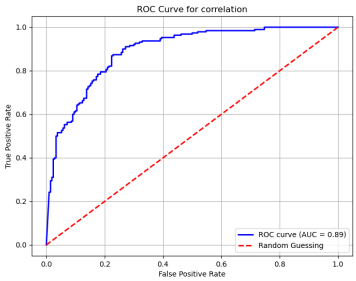
Σχήμα 4.33: Οι ακρίβειες του μοντέλου Random Forest για κάθε μία μέθοδο

Επιπλέον, απεικονίζοται και οι καμπύλες ROC: Οι βαθμολογίες AUC για κάθε μέθοδο

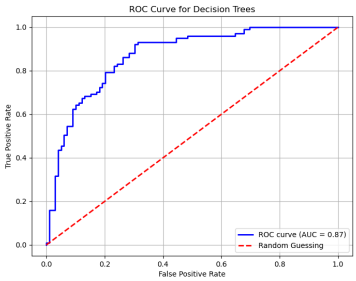
Οι απεικονίσεις των μητρών σύγχυσης για κάθε μέθοδο: Υπόλοιπες μετρικές αξιολόγησης και αποτελέσματα μητρών σύγχυσης:

Επιπλέον, οι συναρτήσεις log-loss απεικονίζονται στα ακόλουθα διαγράμματα:

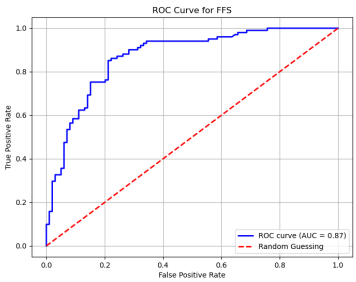
Όλες οι μέθοδοι παρουσιάζουν παρόμοιες συμπεριφορές μεταξύ τους, για παράδειγμα οι μέθοδοι random forest, recursive feature elimination, information gain, fisher scores έχουν παρόμοιες καμπύλες, αντίστοιχα, οι decision trees, lasso regularization, sequential feature elimination, επίσης, έχουν παρόμοιες καμπύλες.



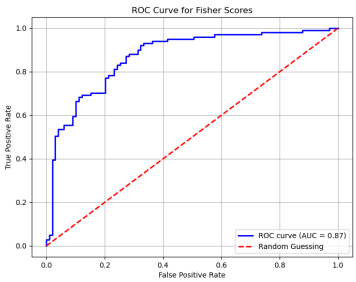
(α΄) Correlation



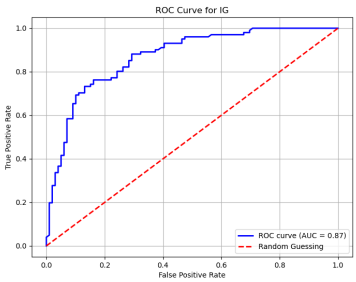
(β΄) Decision Trees



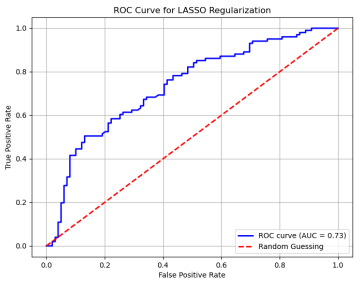
(γ΄) Forward Feature Selection



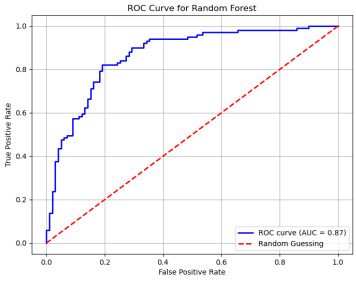
(δ΄) Fisher Scores



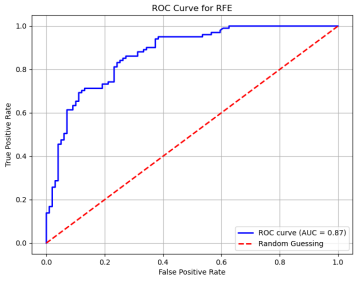
(ε΄) Information Gain



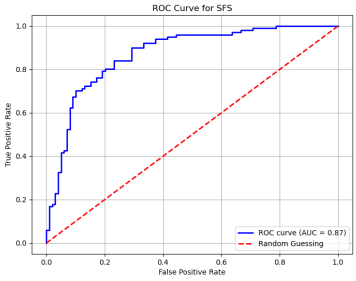
(ϛ΄) Lasso Regularization



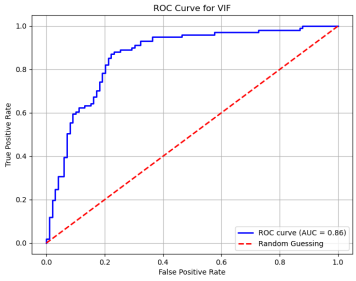
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination

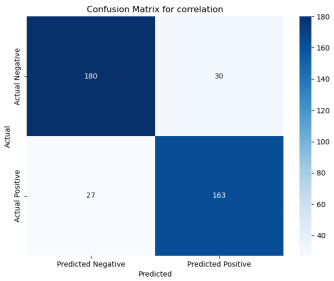


(θ΄) Sequential Feature Selection

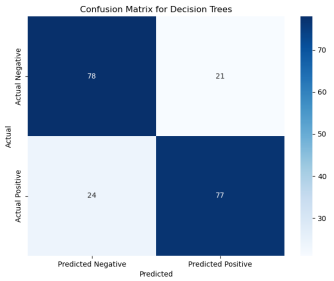


(ι΄) Variance Inflation Factor

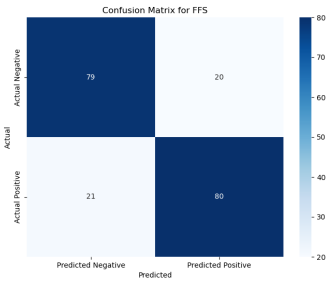
Σχήμα 4.34: ROC Curves για όλες τις μεθόδους για το μοντέλο Random Forest



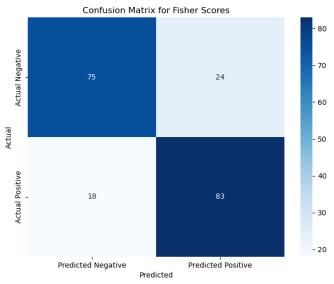
(α′) Correlation



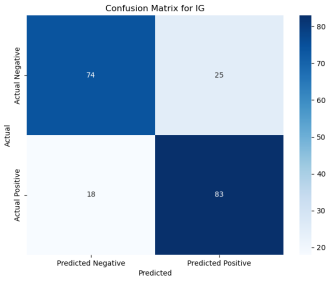
(β′) Decision Trees



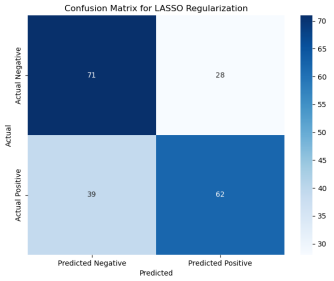
(γ′) Forward Feature Selection



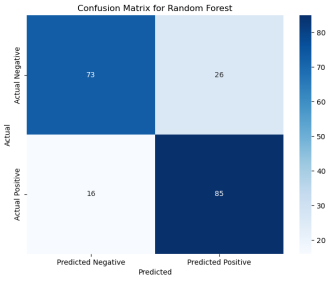
(δ′) Fisher Scores



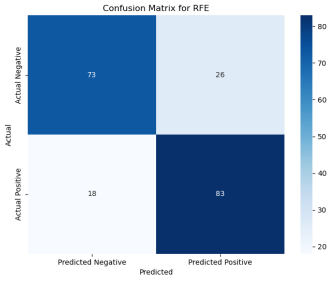
(ε′) Information Gain



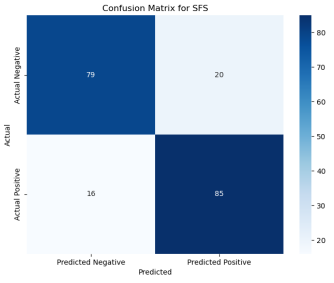
(ϛ′) Lasso Regularization



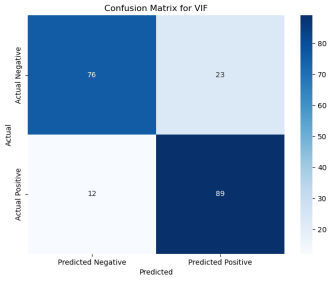
(ζ) Random Forest



(η′) Recursive Feature Elimination

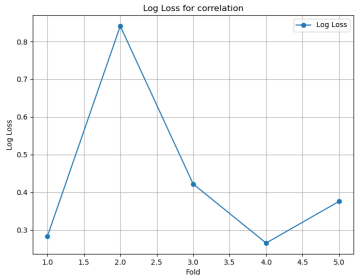


(θ′) Sequential Feature Selection

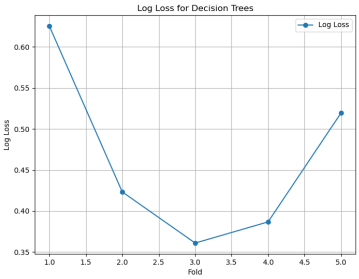


(ι′) Variance Inflation Factor

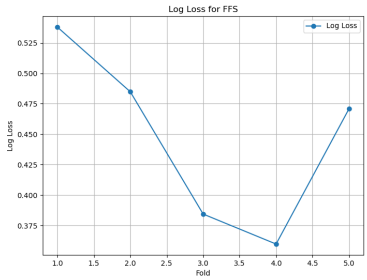
Σχήμα 4.35: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Random Forest



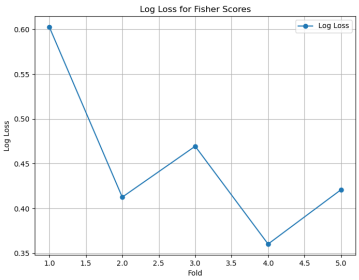
(α΄) Correlation



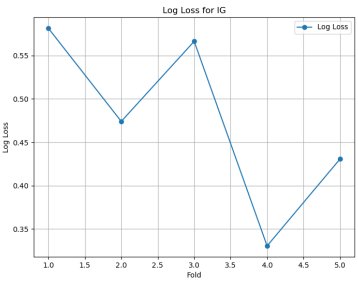
(β΄) Decision Trees



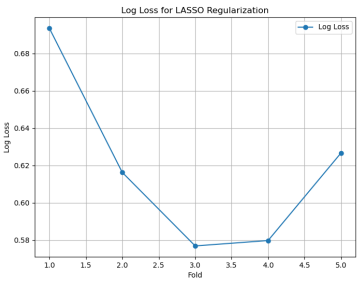
(γ΄) Forward Feature Selection



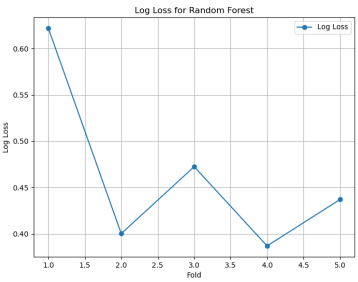
(δ΄) Fisher Scores



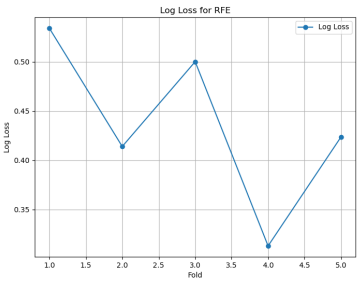
(ε΄) Information Gain



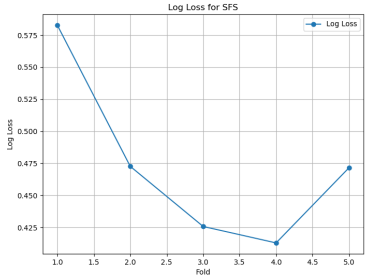
(ϛ΄) Lasso Regularization



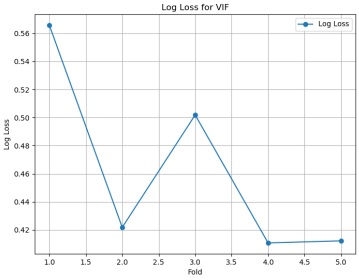
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination



(θ΄) Sequential Feature Selection



(ι΄) Variance Inflation Factor

Σχήμα 4.36: Log-Loss ως προς τον αριθμό των επαναλήψεων συναρτήσεις για όλες τις μεθόδους για το μοντέλο Random Forest

Method	AUC Score
Correlation	0.91
Decision Trees	0.87
Forward Feature Selection	0.87
Fisher Scores	0.87
Information Gain	0.87
Lasso Regularization	0.73
Random Forest	0.87
Recursive Feature Elimination	0.87
Sequential Feature Selection	0.87
Variance Inflation Factor	0.86

Πίνακας 4.24: AUC Scores για κάθε μέθοδο

Method	Overall Accuracy	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
Correlation	85.75%	85.71%	85.75%	85.72%
Decision Trees	77.5%	77.52%	77.51%	77.49%
FFS	79.5%	79.5%	79.50%	79.40%
Fisher Scores	79.0%	79.11%	78.97%	78.96%
IG	78.5%	78.65%	78.46%	78.45%
LASSO Regularization	66.5%	66.72%	66.55%	66.43%
Random Forest	79.0%	79.30%	78.90%	78.92%
RFE	78.0%	78.18%	77.95%	77.94%
SFS	82.0%	82.05%	81.97%	81.98%
VIF	82.5%	82.90%	82.44%	82.40%

Πίνακας 4.25: Μετρικές Απόδοσης για κάθε μέθοδο

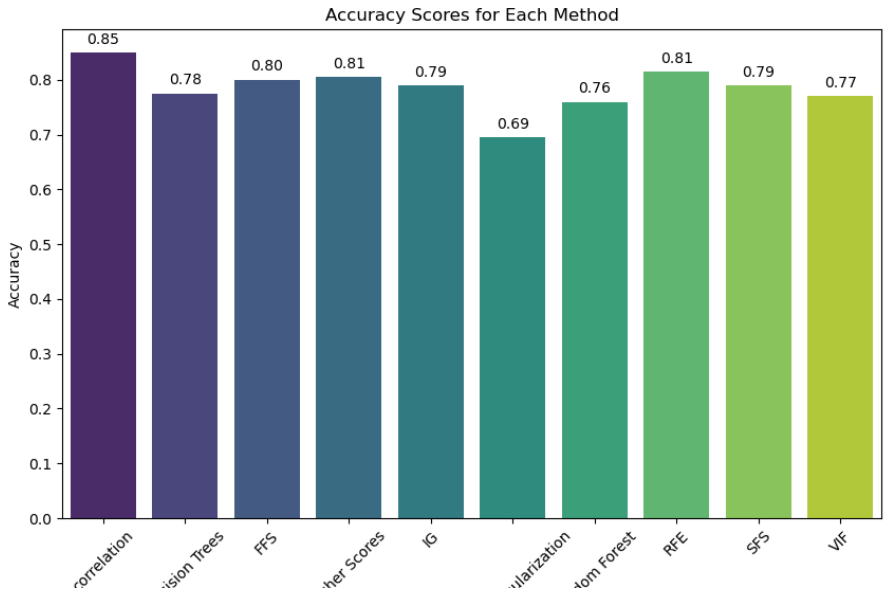
4.11 Stochastic Gradient Descent

Από τη διαδικασία επιλογής των βέλτιστων συνδυασμών υπερπαραμέτρων για κάθε μέθοδο, προέκυψαν οι ακόλουθες αρχιτεκτονικές:

Method	Learning Rate	Max Depth	Max Iter	Min Samples Leaf
Correlation	0.01	3	300	1
Decision Trees	0.01	3	200	4
FFS	0.01	3	200	1
Fisher Scores	0.1	5	200	4
IG	0.01	5	100	4
LASSO Regularization	0.01	3	200	4
Random Forest	0.01	5	200	1
RFE	0.1	3	100	2
SFS	0.01	3	300	1
VIF	0.1	3	200	1

Πίνακας 4.26: Υπερπαράμετροι για κάθε μέθοδο

Η ακρίβειες που επιτυγχάνονται από το μοντέλο απεικονίζονται στο παρακάτω διάγραμμα: Επιπλέον, απεικονίζοται και οι καμπύλες ROC: Οι βαθμολογίες



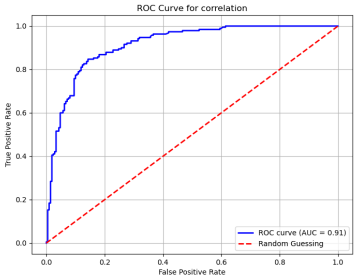
Σχήμα 4.37: Οι ακρίβειες του μοντέλου Stochastic Gradient Descent για κάθε μία μέθοδο

AUC για κάθε μέθοδο:

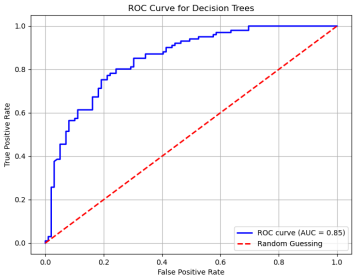
Οι απεικονίσεις των μητρών σύγχυσης για κάθε μέθοδο:

Τα αποτελέσματα που απορρέουν από αυτά και τις υπόλοιπες μετρικές αξιολόγησης των μοντέλων είναι τα ακόλουθα:

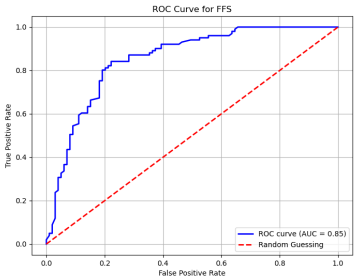
Και οι συναρτήσεις log-loss: Γενικότερα, παρατηρείται ότι δεν υπάρχει μία συγκεκριμένη συμπεριφορά για όλες τις μεθόδους, κάθε μέθοδο αποδίδει καλύτερα σε διαφορετικές διαιρέσεις (folds).



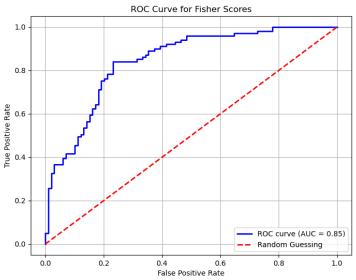
(α΄) Correlation



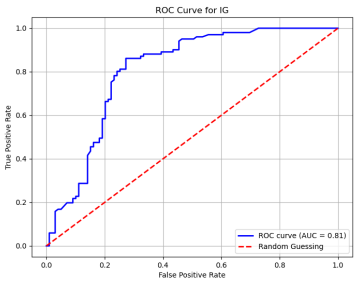
(β΄) Decision Trees



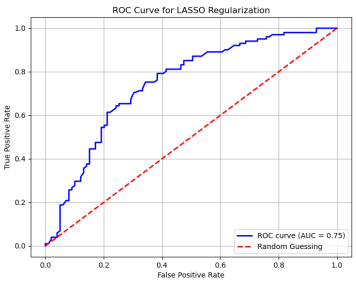
(γ΄) Forward Feature Selection



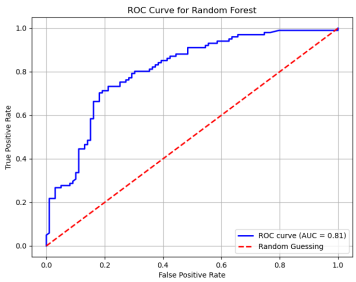
(δ΄) Fisher Scores



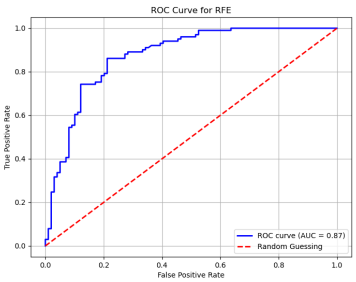
(ε΄) Information Gain



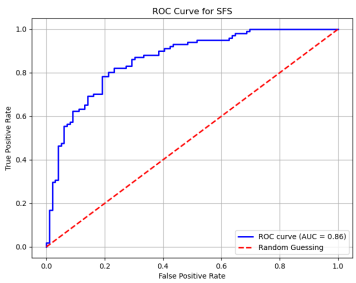
(ϛ΄) Lasso Regularization



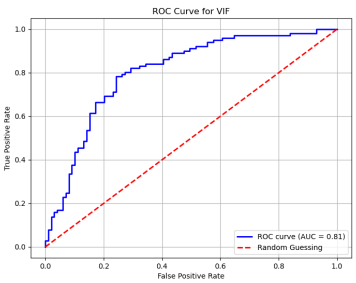
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination

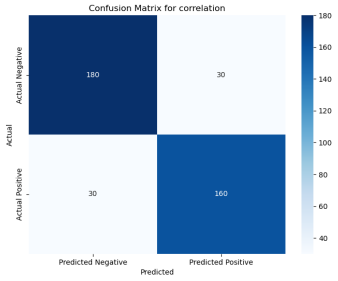


(θ΄) Sequential Feature Selection

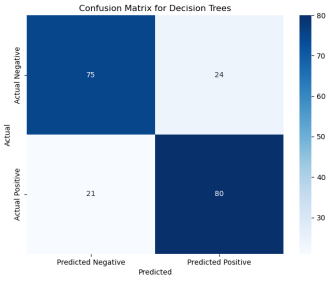


(ι΄) Variance Inflation Factor

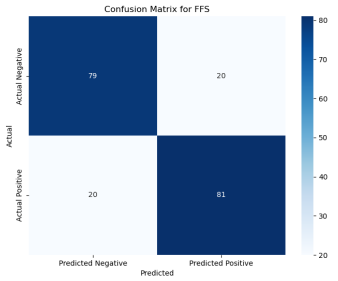
Σχήμα 4.38: ROC Curves για όλες τις μεθόδους για το μοντέλο Stochastic Gradient Descent



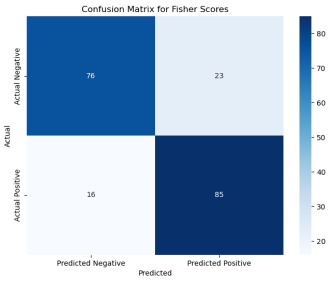
(α΄) Correlation



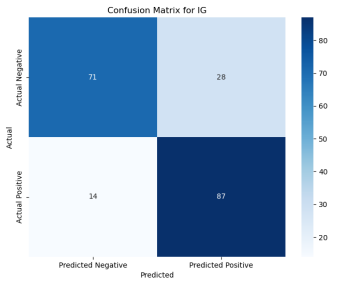
(β΄) Decision Trees



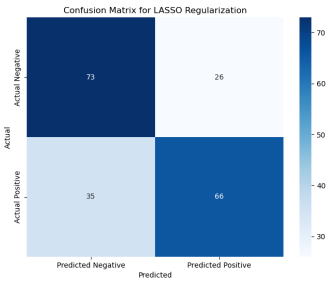
(γ΄) Forward Feature Selection



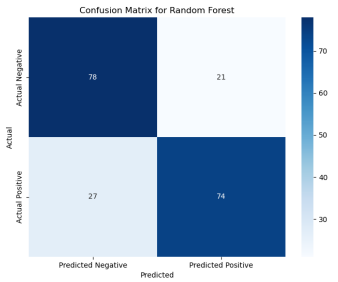
(δ΄) Fisher Scores



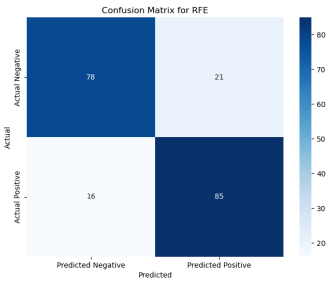
(ε΄) Information Gain



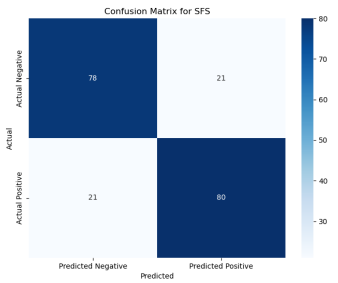
(ϛ΄) Lasso Regularization



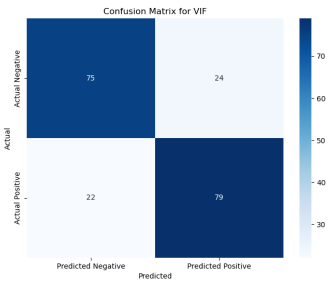
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination

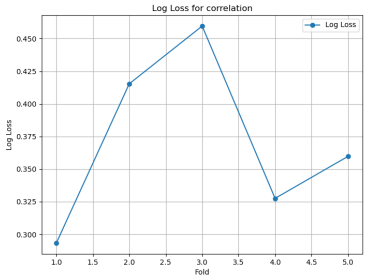


(θ΄) Sequential Feature Selection

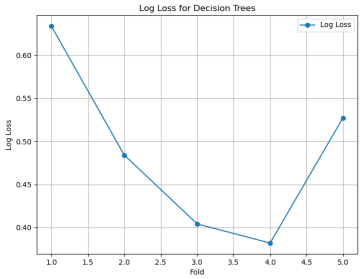


(ι΄) Variance Inflation Factor

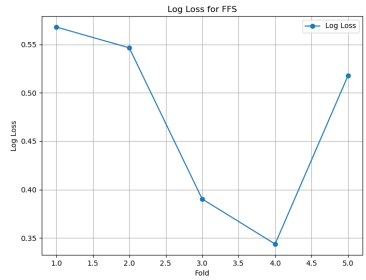
Σχήμα 4.39: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Stochastic Gradient Descent



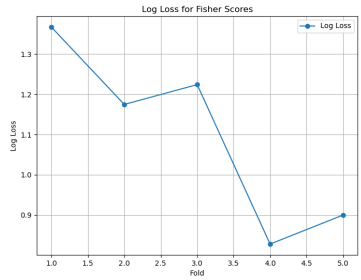
(α′) Correlation



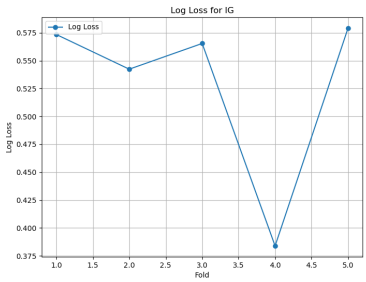
(β′) Decision Trees



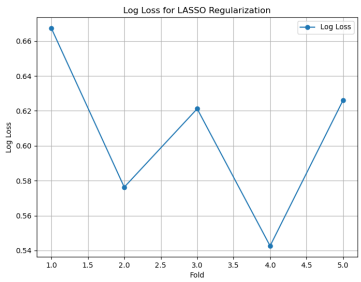
(γ′) Forward Feature Selection



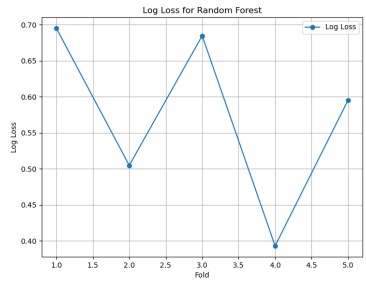
(δ′) Fisher Scores



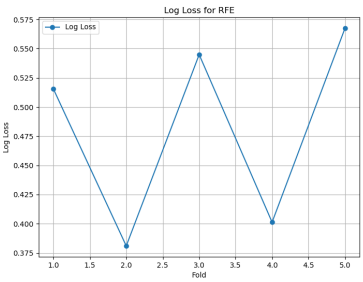
(ε′) Information Gain



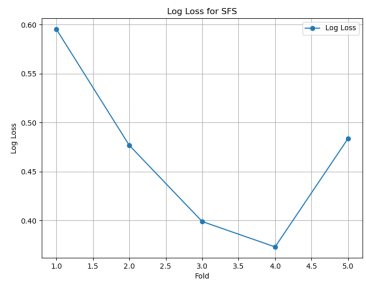
(ϛ′) Lasso Regularization



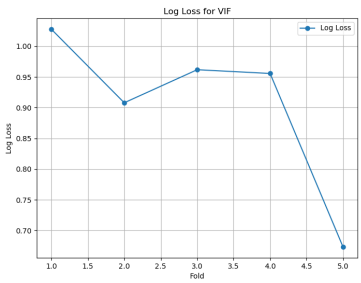
(ζ) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection



(ι′) Variance Inflation Factor

Σχήμα 4.40: Log-Loss ως προς τον αριθμό των επαναλήψεων συναρτήσεις για όλες τις μεθόδους για το μοντέλο Stochastic Gradient Descent

Method	AUC Score
Correlation	0.91
Decision Trees	0.85
Forward Feature Selection	0.85
Fisher Scores	0.85
Information Gain	0.81
Lasso Regularization	0.75
Random Forest	0.81
Recursive Feature Elimination	0.87
Sequential Feature Selection	0.86
Variance Inflation Factor	0.81

Πίνακας 4.27: AUC Scores για κάθε μέθοδο

Method	Overall Accuracy	Precision	Recall	F1-Score
Correlation	85%	84.9%	84.2%	84.2%
Decision Trees	77.5%	77.52%	77.48%	77.48%
FFS	80%	80.1%	80.1%	80.19%
Fisher Scores	80.5%	80.65%	80.46%	80.46%
IG	79%	75.59%	78.9%	78.8%
LASSO Regularization	69.5%	69.6%	69.5%	69.4%
Random Forest	76.0%	76.02%	76.07%	75.99%
RFE	81.5%	81.58%	81.47%	81.47%
SFS	79%	78.99%	78.99%	78.99%
VIF	77%	76.9 %	76.9%	76.9%

Πίνακας 4.28: Μετρικές Απόδοσης για κάθε μέθοδο

4.12 Support Vector Machine

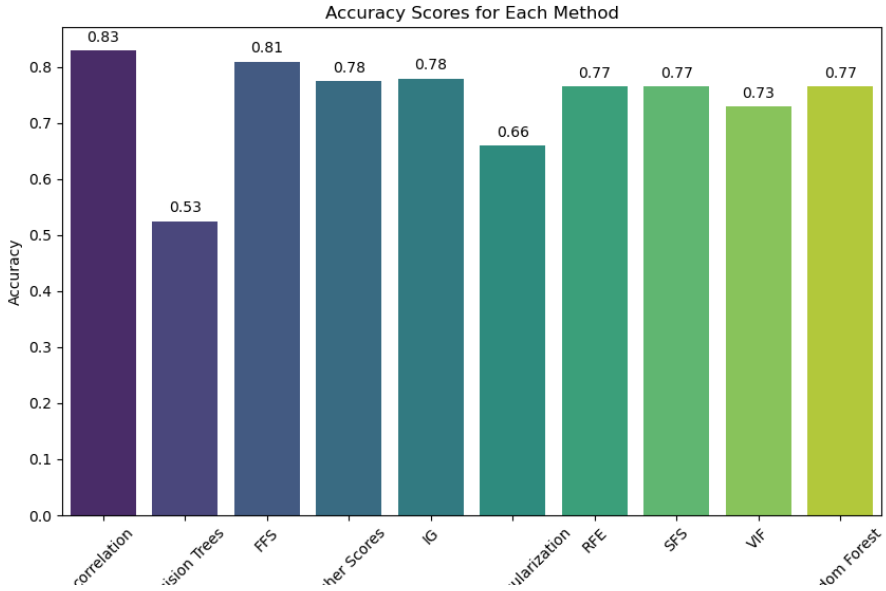
Κατόπιν της διαδικασίας επιλογής των βέλτιστων υπερπαραμέτρων για κάθε μέθοδο του μοντέλου, προέκυψαν οι ακόλουθες αρχιτεκτονικές:

Μέθοδος	C	Degree	Gamma	Kernel
Decision Trees	0.1	2	scale	rbf
Correlation	0.1	2	scale	linear
FFS	0.1	2	scale	linear
Fisher Scores	0.8	2	scale	linear
IG	0.1	2	scale	linear
LASSO Regularization	0.1	2	scale	linear
Random Forest	0.1	2	scale	linear
RFE	0.1	2	scale	rbf
SFS	0.1	2	scale	linear
VIF	1	2	scale	linear

Πίνακας 4.29: Υπερπαράμετροι για κάθε μέθοδο

Οι μέγιστες συνολικές ακρίβειες που επέτυχε το μοντέλο για κάθε μέθοδο παρουσιάζονται στο ακόλουθο διάγραμμα: Επιπλέον, απεικονίζοται και οι καμπύλες ROC: Οι βαθμολογίες AUC για κάθε μέθοδο:

Οι απεικονίσεις των μητρών σύγχυσης για κάθε μέθοδο: Από τα οποία

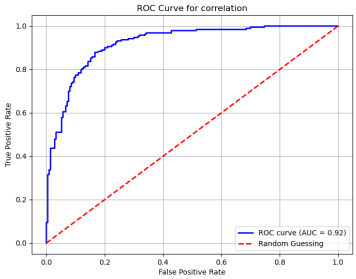


Σχήμα 4.41: Οι ακρίβειες του μοντέλου Support Vector Machine για κάθε μία μέθοδο

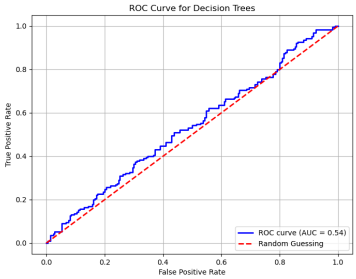
Μέθοδος	AUC Βαθμολογία
Correlation	0.92
Decision Trees	0.54
Forward Feature Selection	0.88
Fisher Scores	0.89
Information Gain	0.87
Lasso Regularization	0.80
Random Forest	0.87
Recursive Feature Elimination	0.87
Sequential Feature Selection	0.88
Variance Inflation Factor	0.87

Πίνακας 4.30: AUC Βαθμολογία για κάθε μέθοδο

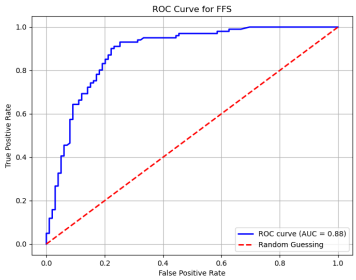
προκύπτουν οι εξής πληροφορίες όσον αφορά τις αποδόσεις του μοντέλου με κάθε μέθοδο:



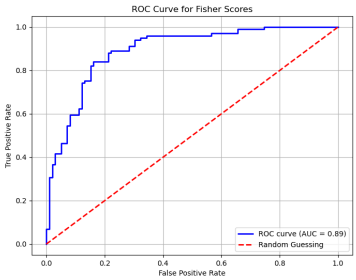
(α΄) Correlation



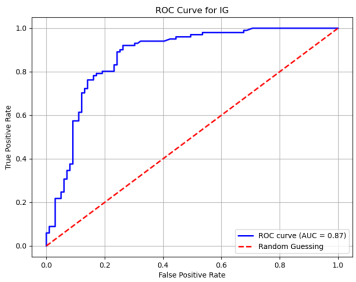
(β΄) Decision Trees



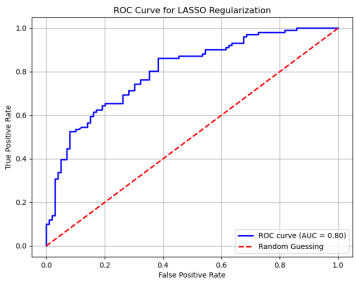
(γ΄) Forward Feature Selection



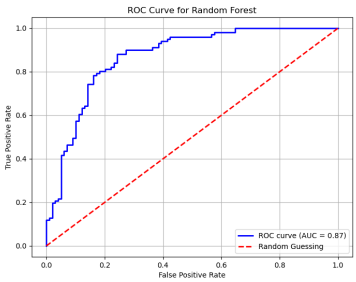
(δ΄) Fisher Scores



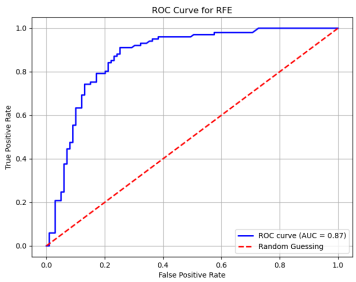
(ε΄) Information Gain



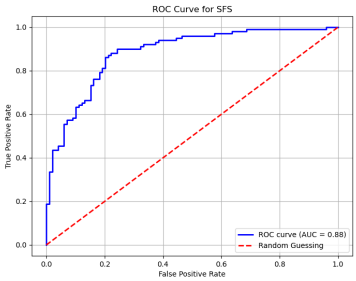
(ϛ΄) Lasso Regularization



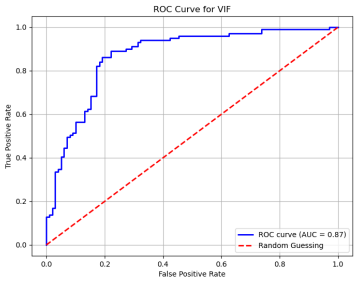
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination

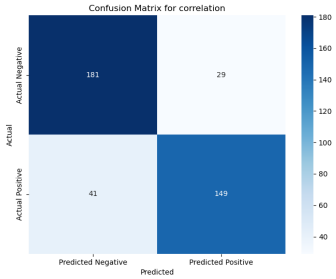


(θ΄) Sequential Feature Selection

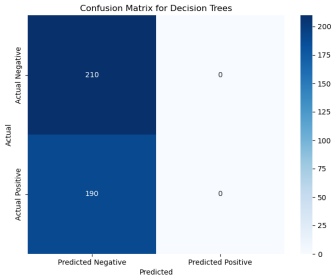


(ι΄) Variance Inflation Factor

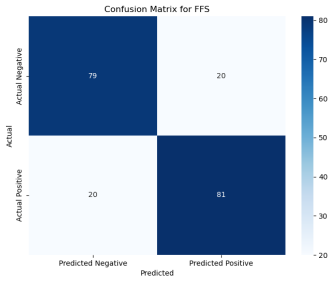
Σχήμα 4.42: ROC Curves για όλες τις μεθόδους για το μοντέλο Support Vector Machine



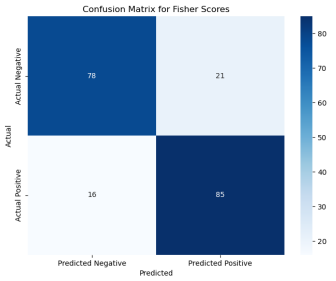
(α΄) Correlation



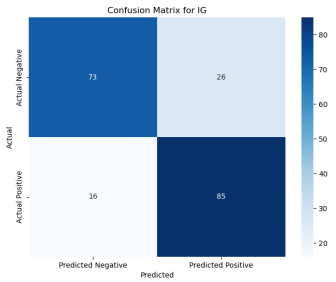
(β΄) Decision Trees



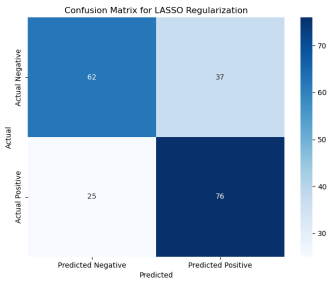
(γ΄) Forward Feature Selection



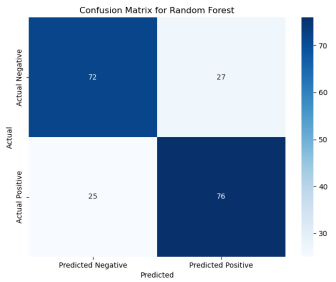
(δ΄) Fisher Scores



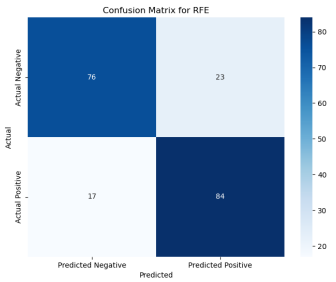
(ε΄) Information Gain



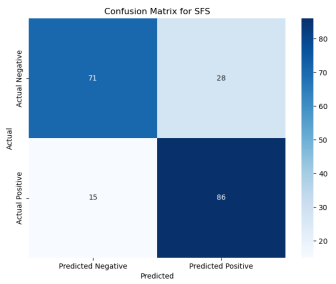
(ϛ΄) Lasso Regularization



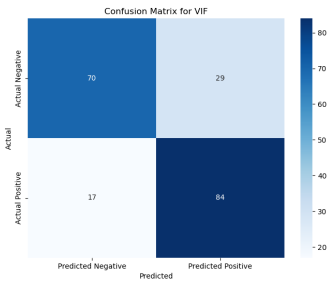
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination

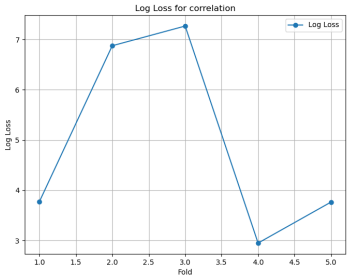


(θ΄) Sequential Feature Selection

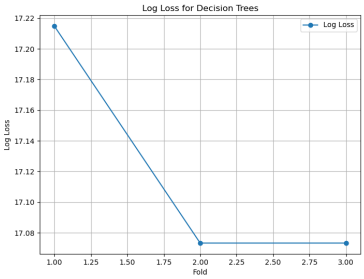


(ι΄) Variance Inflation Factor

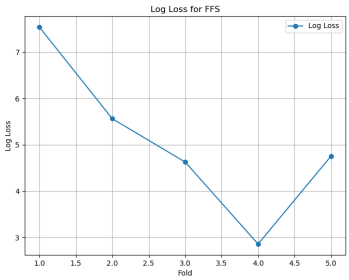
Σχήμα 4.43: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο Support Vector Machine



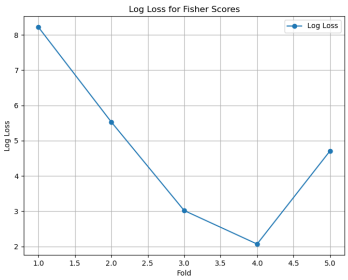
(α΄) Correlation



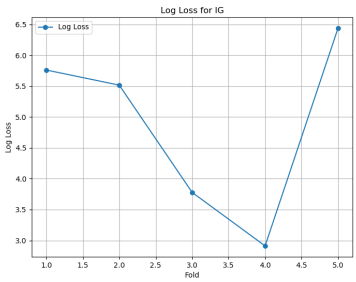
(β΄) Decision Trees



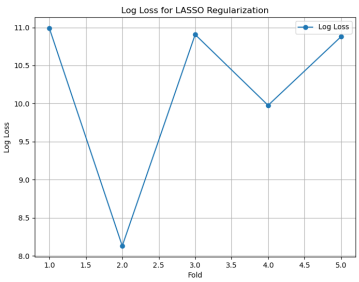
(γ΄) Forward Feature Selection



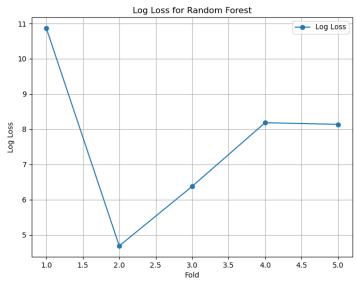
(δ΄) Fisher Scores



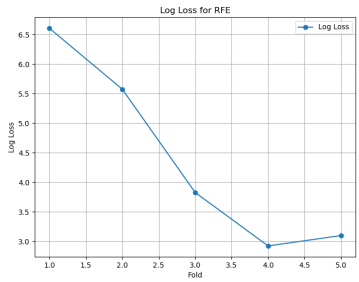
(ε΄) Information Gain



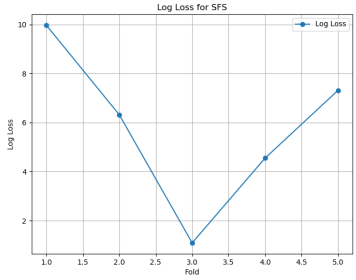
(ϛ΄) Lasso Regularization



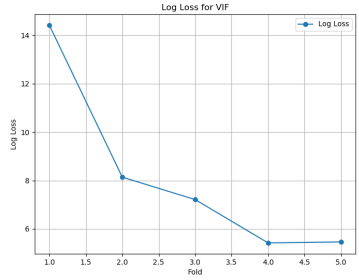
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination



(θ΄) Sequential Feature Selection



(ι΄) Variance Inflation Factor

Σχήμα 4.44: Log-Loss ως προς τον αριθμό των επαναλήψεων συναρτήσεις για όλες τις μεθόδους για το μοντέλο Support Vector Machine

Μέθοδος	Accuracy	Precision	Recall	F1-Score
Correlation	83.0%	83.1%	82.8%	82.9%
Decision Trees	52.5%	26.25%	50%	34%
FFS	81.0%	81.1%	81.0%	81.0%
Fisher Scores	77.5%	77.6%	77.5%	77.5%
IG	78.0%	78.1%	77.9%	77.9%
LASSO Regularization	66.0%	66.2%	65.9%	65.8%
Random Forest	76.5%	76.5%	76.5%	76.5%
RFE	79.0%	79.1%	79.0%	79.0%
SFS	76.5%	76.6%	76.5%	76.5%
VIF	73.0%	74.4%	72.9%	72.5%

Πίνακας 4.31: Μετρικές Απόδοσης για κάθε μέθοδο

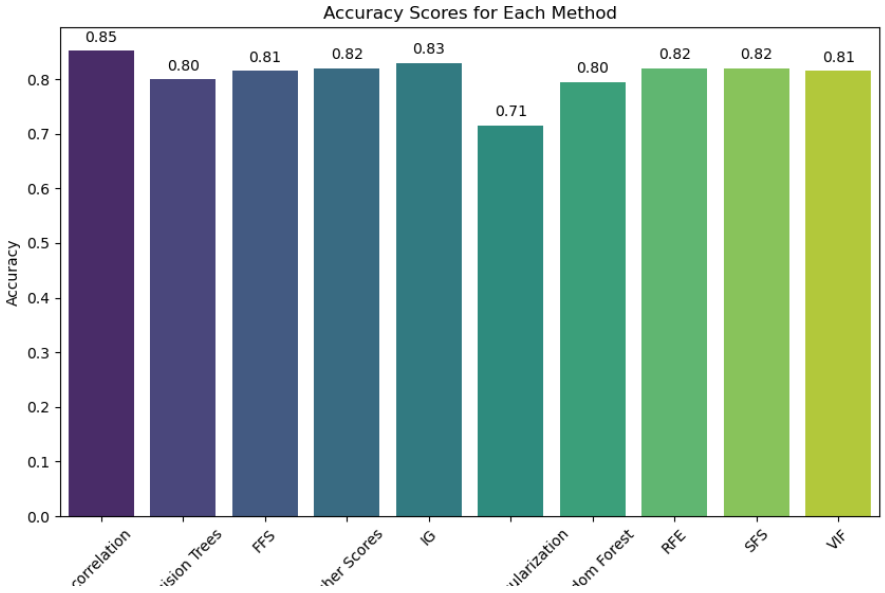
4.13 XGBoost

Για τις βέλτιστες υπερπαραμέτρους βρέθηκαν οι εξής αρχιτεκτονικές για κάθε μέθοδο:

Μέθοδος	Learning Rate	Max Depth	Min Child Weight	N Estimators	Subsample
Correlation	0.1	10	5	200	1.0
Decision Trees	0.01	3	5	100	1.0
FFS	0.01	3	3	300	1.0
Fisher Scores	0.5	10	1	100	0.8
IG	0.5	3	5	300	0.9
LASSO Regularization	0.01	3	1	300	1.0
Random Forest	0.5	3	5	300	0.8
RFE	0.1	3	1	100	1.0
SFS	0.01	3	5	200	1.0
VIF	0.01	10	1	200	0.8

Πίνακας 4.32: Υπερπαράμετροι για κάθε μέθοδο

Οι μέγιστες συνολικές ακρίβειες που επέτυχε το μοντέλο για κάθε μέθοδο παρουσιάζονται στο ακόλουθο διάγραμμα: Επιπλέον, απεικονίζοται και οι κα-

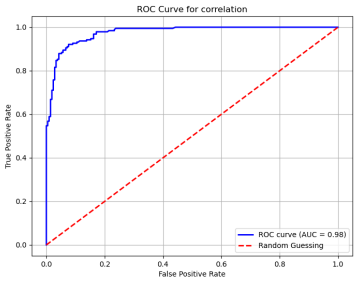


Σχήμα 4.45: Οι ακρίβειες του μοντέλου XGBoost για κάθε μία μέθοδο

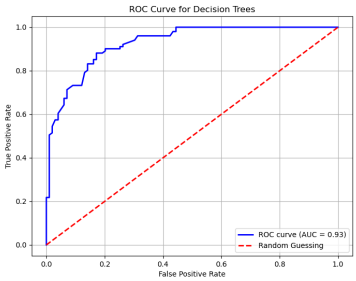
μπύλες ROC: Οι βαθμολογίες AUC για κάθε μέθοδο:

Οι απεικονίσεις των μητρών σύγχυσης για κάθε μέθοδο: Από τις στατιστικές μετρικές και τις μήτρες σύγχυσης απορρέουν τα εξής αποτελέσματα:

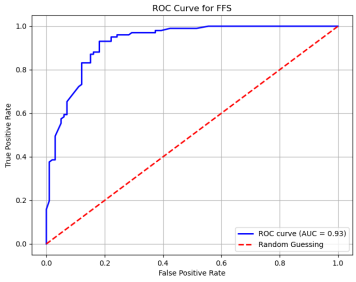
Τέλος, παρατίθενται και τα διαγράμματα των συναρτήσεων λογ-λοσς για κάθε μέθοδο: Από τα διαγράμματα αυτά φαίνεται ότι τα περισσότερα μοντέλα έχουν παρόμοιες συμπεριφορές για κάθε μέθοδο, όπως για παράδειγμα τα decision trees με FFS, τα fisher scores με τα IG και lasso regularization και τα VIF, RFE και Random Forest.



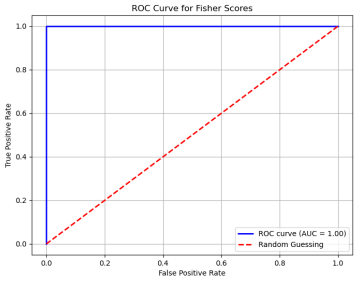
(α′) Correlation



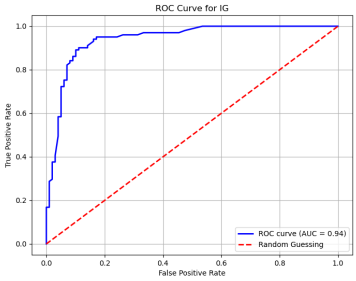
(β′) Decision Trees



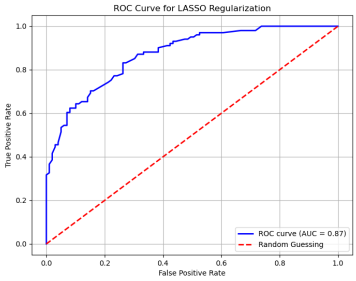
(γ′) Forward Feature Selection



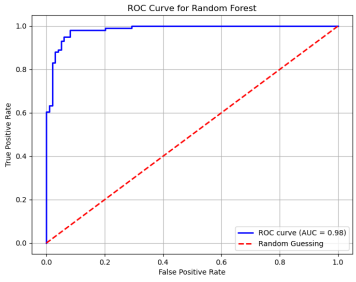
(δ′) Fisher Scores



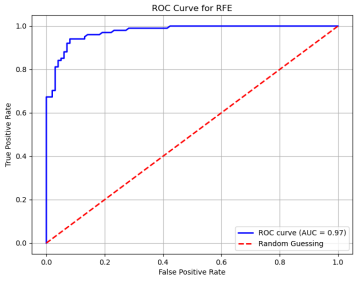
(ε′) Information Gain



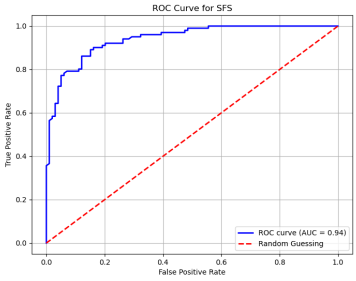
(ϛ′) Lasso Regularization



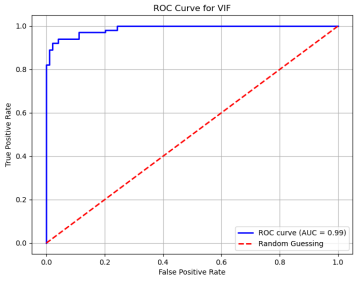
(ζ) Random Forest



(η′) Recursive Feature Elimination

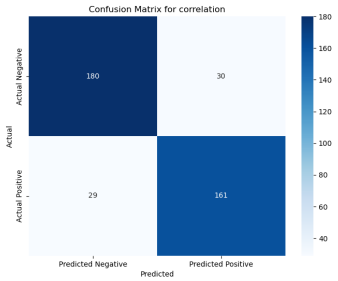


(θ′) Sequential Feature Selection

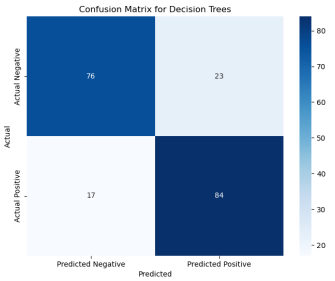


(ι′) Variance Inflation Factor

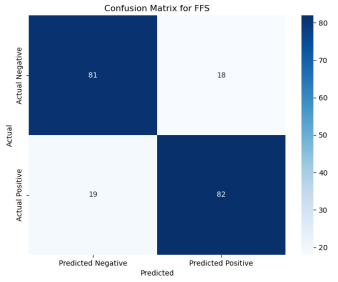
Σχήμα 4.46: ROC Curves για όλες τις μεθόδους για το μοντέλο XGBoost



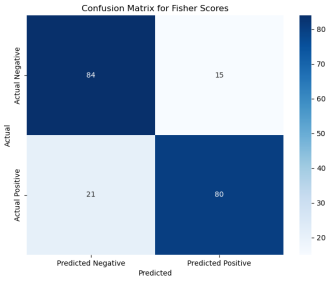
(α΄) Correlation



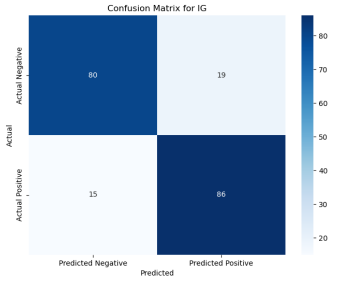
(β΄) Decision Trees



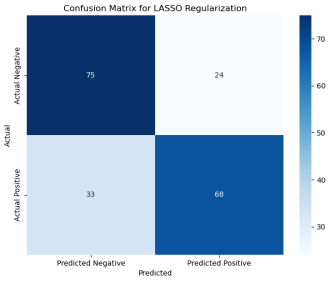
(γ΄) Forward Feature Selection



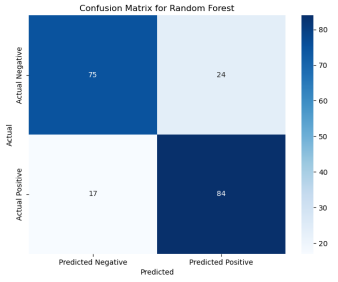
(δ΄) Fisher Scores



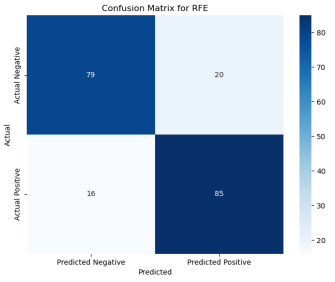
(ε΄) Information Gain



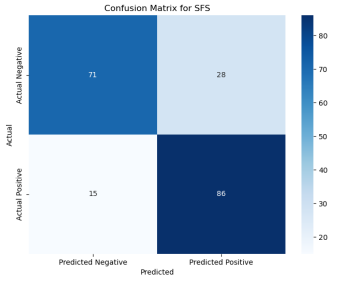
(ϛ΄) Lasso Regularization



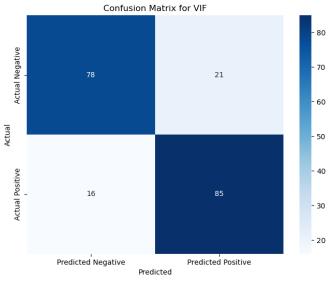
(ζ΄) Random Forest



(η΄) Recursive Feature Elimination

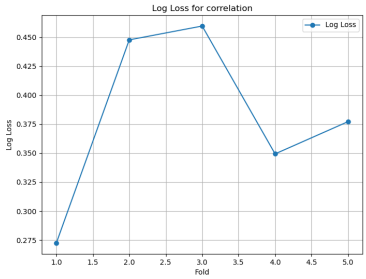


(θ΄) Sequential Feature Selection

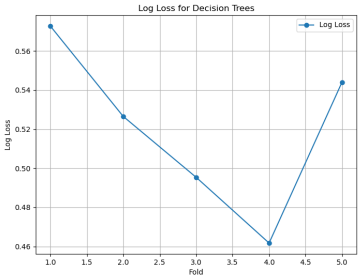


(ι΄) Variance Inflation Factor

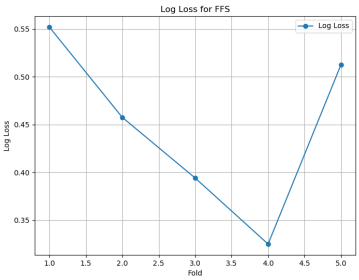
Σχήμα 4.47: Μήτρες σύγχυσης για όλες τις μεθόδους για το μοντέλο XGBoost



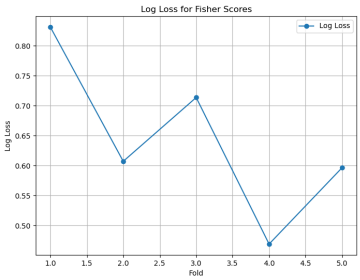
(α′) Correlation



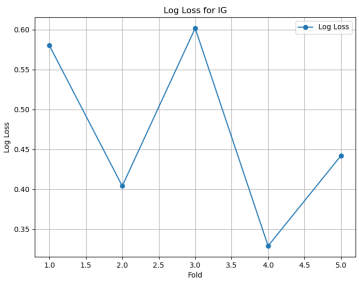
(β′) Decision Trees



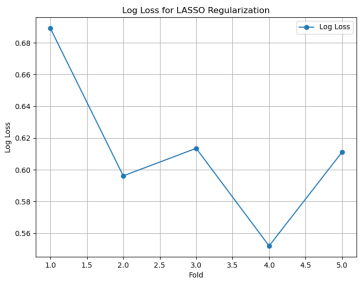
(γ′) Forward Feature Selection



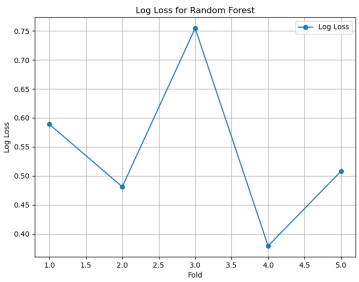
(δ′) Fisher Scores



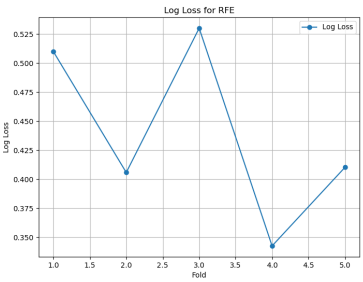
(ε′) Information Gain



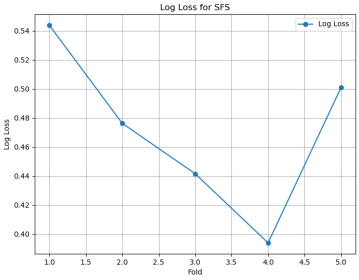
(ϛ′) Lasso Regularization



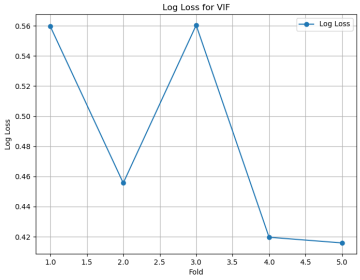
(ζ′) Random Forest



(η′) Recursive Feature Elimination



(θ′) Sequential Feature Selection



(ι′) Variance Inflation Factor

Σχήμα 4.48: Log-Loss ως προς τον αριθμό των επαναλήψεων συναρτήσεις για όλες τις μεθόδους για το μοντέλο XGBoost

Μέθοδος	AUC
Correlation	0.98
Decision Trees	0.93
Forward Feature Selection	0.93
Fisher Scores	1.00
Information Gain	0.94
Lasso Regularization	0.87
Random Forest	0.98
Recursive Feature Elimination	0.97
Sequential Feature Selection	0.94
Variance Inflation Factor	0.99

Πίνακας 4.33: AUC βαθμολογία για κάθε μέθοδο

Μέθοδος	Ακρίβεια (%)	Precision (%)	Recall (%)	F1-Score (%)
Correlation	85.25	85.21	85.75	85.72
Decision Trees	80.00	80.86	80.00	77.49
FFS	81.50	81.50	81.50	79.50
Fisher Scores	82.00	82.11	82.00	78.96
IG	83.00	83.06	83.00	78.45
LASSO Regularization	71.50	71.68	71.50	66.43
Random Forest	79.50	79.65	79.50	78.92
RFE	82.00	82.06	82.00	77.94
SFS	82.00	82.00	82.00	79.45
VIF	81.50	81.59	81.50	79.45

Πίνακας 4.34: Μετρικές Απόδοσης για κάθε μέθοδο

4.14 Principal Component Analysis σε συνδυασμό με τα μοντέλα

Για κάθε μοντέλο εξετάστηκαν και επιλέχθηκαν οι βέλτιστοι συνδυασμοί υπερπαραμέτρων βάσει των αποδόσεων στις ακρίβειες συνδυάζοντας τη μέθοδο PCA για τη μείωση των διαστάσεων. Για κάθε μοντέλο προέκυψαν οι εξής αρχιτεκτονικές:

Hyperparameter	Value
criterion	entropy
max depth	10
min samples leaf	4
min samples split	10

Πίνακας 4.35: Υπερπαράμετροι για Decision Trees

Hyperparameter	Value
learning rate	0.01
max depth	3
min samples leaf	4
min samples split	2
n estimators	300
subsample	0.8

Πίνακας 4.36: Υπερπαράμετροι για Gradient Boosting

Hyperparameter	Value
algorithm	auto
leaf size	20
n neighbors	7
p	2
weights	uniform

Πίνακας 4.37: Υπερπαράμετροι για KNN

Hyperparameter	Value
C	0.01
max iter	500
penalty	l2
solver	lbfgs

Πίνακας 4.38: Υπερπαράμετροι για Logistic Regression

Οι μέγιστες συνολικές ακρίβειες που επέτυχε το μοντέλο για κάθε μέθοδο παρουσιάζονται στο ακόλουθο διάγραμμα: Επιπλέον, απεικονίζοται και οι καμπύλες ROC: Οι βαθμολογίες AUC για κάθε μοντέλο:

Οι απεικονίσεις των μητρών σύγχυσης για κάθε μέθοδο: Από τις υπόλοιπες στατιστικές μετρικές υπάρχουν και οι ακόλουθες πληροφορίες για την αξιολόγηση των μοντέλων:

Hyperparameter	Value
criterion	entropy
max depth	5
max features	log2
min samples leaf	4
min samples split	10
n estimators	50

Πίνακας 4.39: Υπερπαράμετροι για Random Forest

Hyperparameter	Value
alpha	0.001
eta0	0.1
learning rate	constant
loss	hinge
max iter	200
penalty	l1
tol	1ε-05

Πίνακας 4.40: Υπερπαράμετροι για Stochastic Gradient Descent

Hyperparameter	Value
learning rate	0.01
max depth	3
min child weight	3
n estimators	100
subsample	0.8

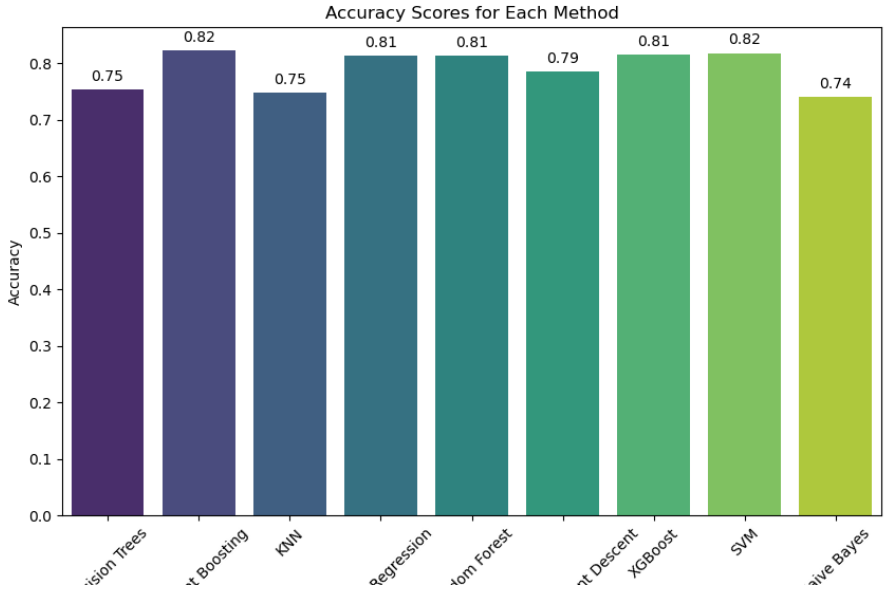
Πίνακας 4.41: Υπερπαράμετροι για XGBoost

Hyperparameter	Value
C	0.8
degree	2
gamma	scale
kernel	rbf

Πίνακας 4.42: Υπερπαράμετροι για SVM

Μέθοδος	AUC
Decision Trees	0.76
Gradient Boosting	0.89
K-Nearest Neighbors	0.84
Logistic Regression	0.9
Naive Bayes	0.84
Penalized Logistic Regression	0.93
Random Forest	0.86
Stochastic Gradient Descent	0.85
Support Vector Machine	0.89
XGBoost	0.88

Πίνακας 4.43: AUC βαθμολογία για κάθε μέθοδο

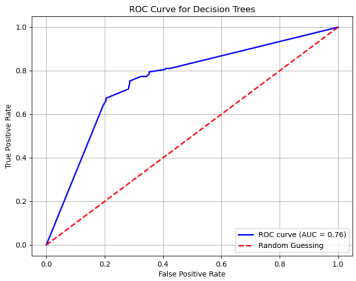


Σχήμα 4.49: Οι ακρίβειες του για κάθε μοντέλο με το Principal Component Analysis

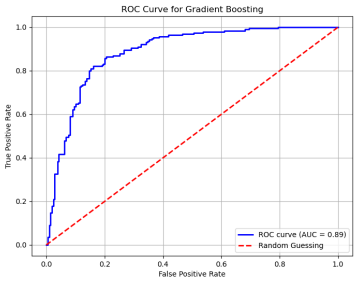
Μέθοδος	Ακρίβεια	Precision	Recall	F1-Score
Decision Trees	74%	74.1%	73.7%	73.7%
Gradient Boosting	82.5%	82.5%	82.5%	82.5%
K-Nearest Neighbors (KNN)	74.75%	75.4%	74.2%	74.2%
Logistic Regression	85%	87.3%	85%	85.3%
Random Forest	80.75%	80.7%	80.7%	80.7%
Stochastic Gradient Descent (SGD)	80.25%	80.3%	80.1%	80.2%
XGBoost	81.75%	81.8%	81.6%	81.7%
Support Vector Machine (SVM)	81.5%	81.5%	81.4%	81.5%

Πίνακας 4.44: Μετρικές Απόδοσης για κάθε μέθοδο

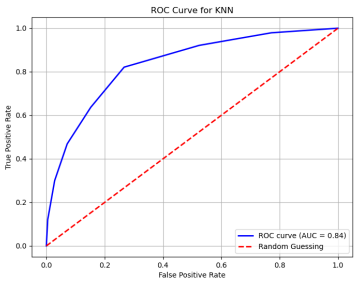
Τέλος, παρατίθενται και τα διαγράμματα των συναρτήσεων log-loss για κάθε μέθοδο:



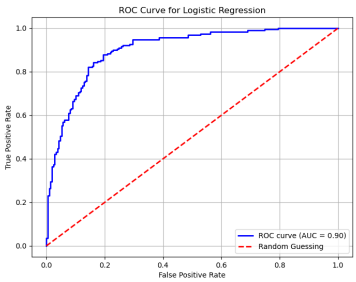
(α') Decision Trees



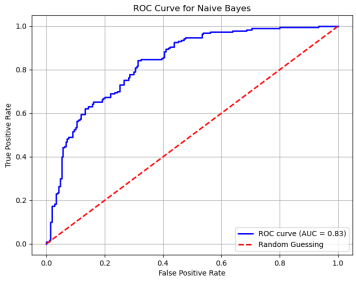
(β') Gradient Boosting



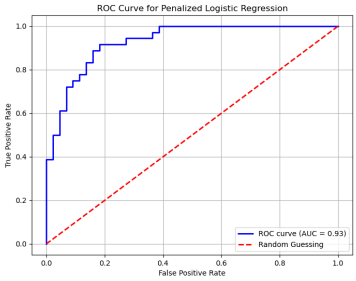
(γ') K-Nearest Neighbors



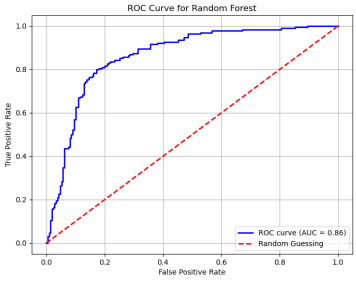
(δ') Logistic Regression



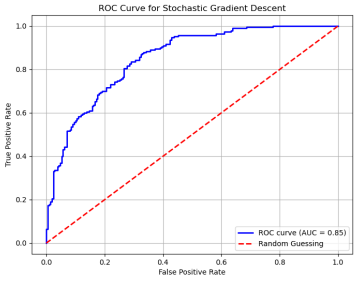
(ε') Naive Bayes



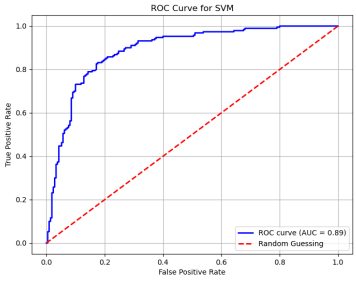
(ς') Penalized Logistic Regression



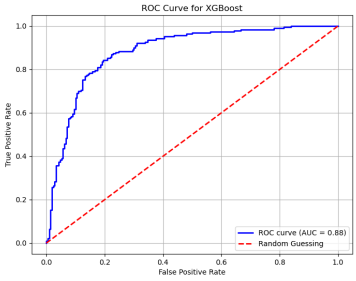
(ζ) Random Forest



(η') Stochastic Gradient Descent

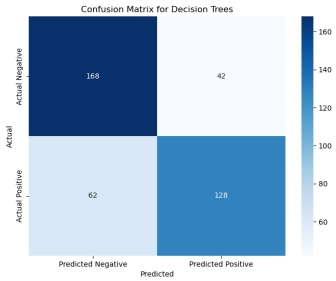


(θ') Support Vector Machine

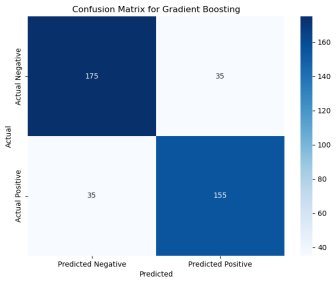


(ι') XGBoost

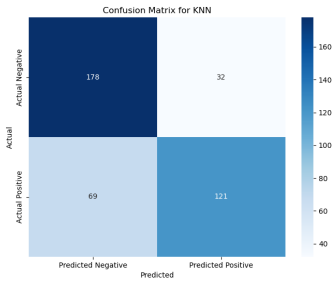
Σχήμα 4.50: ROC Curves για τα μοντέλα με το Principal Component Analysis



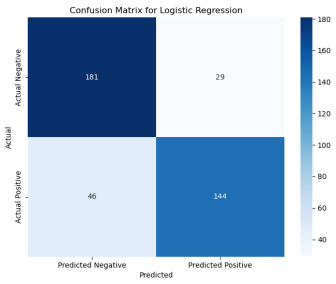
(α΄) Decision Trees



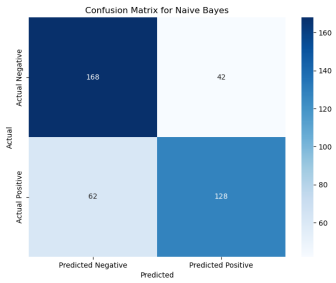
(β΄) Gradient Boosting



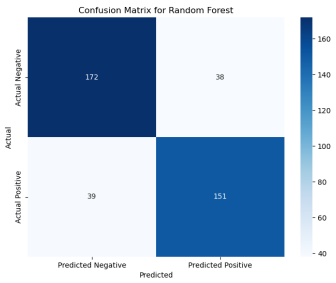
(γ΄) K-Nearest Neighbors



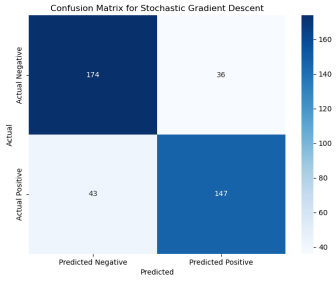
(δ΄) Logistic Regression



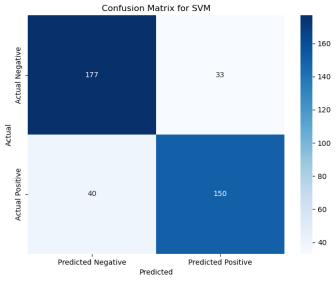
(ε΄) Naive Bayes



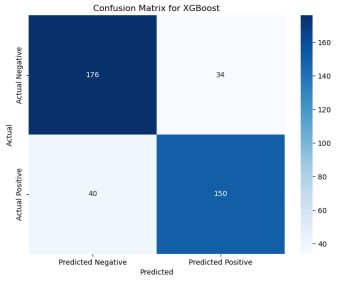
(ς΄) Random Forest



(ζ) Stochastic Gradient Descent

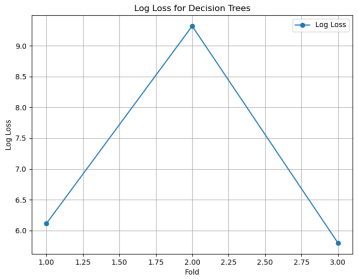


(η΄) Support Vector Machine

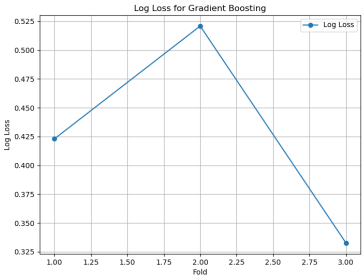


(θ΄) XGBoost

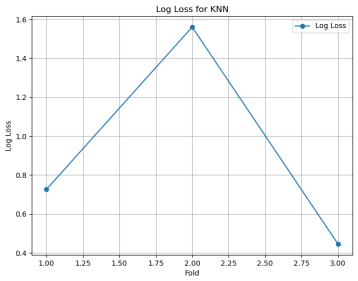
Σχήμα 4.51: Μήτρες σύγχυσης για όλα τα μοντέλα με τη εφαρμμογή PCA



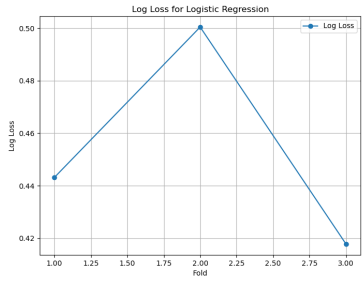
(α') Decision Trees



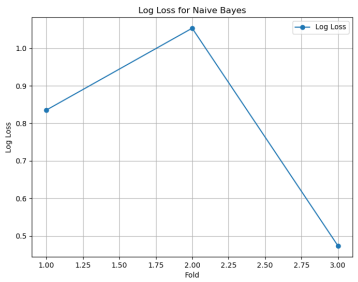
(β') Gradient Boosting



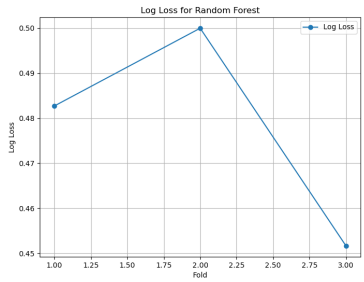
(γ') K-Nearest Neighbors



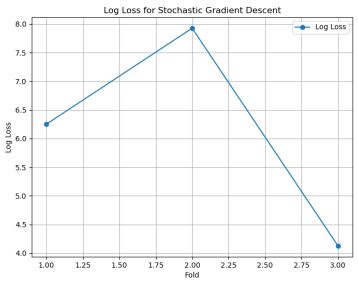
(δ') Logistic Regression



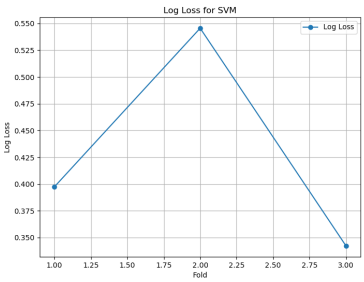
(ε') Naive Bayes



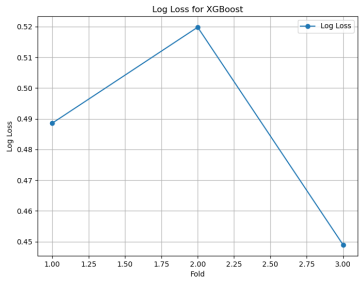
(ϛ') Random Forest



(ζ) Stochastic Gradient Descent



(η') Support Vector Machine



(θ') XGBoost

Σχήμα 4.52: Log-Loss ως προς τον αριθμό των επαναλήψεων συναρτήσεις για όλα τα μοντέλα με την εφαρμογή PCA

Κεφάλαιο 5

Συζήτηση Αποτελεσμάτων

5.1 Ανάλυση Αποτελεσμάτων

Από την ανάλυση των αποτελεσμάτων των Δέντρων Αποφάσεων, παρατηρήθηκαν σημαντικές διακυμάνσεις στην απόδοση του μοντέλου ανάλογα με τις μεθόδους επιλογής χαρακτηριστικών. Κάθε μοντέλο με διαφορετική μέθοδο επιλογής χαρακτηριστικών παρείχε διαφορετικές επιδόσεις με βάση τις μετρικές ακρίβειας, ανάκλησης, F1-score και τη μήτρα σύγχυσης.

Το μοντέλο Δέντρων Απόφασης πέτυχε την υψηλότερη ακρίβεια με τη μέθοδο Correlation, φτάνοντας το 81.5%, ενώ η χαμηλότερη ακρίβεια παρατηρήθηκε με τη μέθοδο LASSO Regularization, μόλις 53.5%. Οι άλλες μέθοδοι κυμαίνονταν μεταξύ αυτών των δύο ακραίων τιμών, χωρίς να δείχνουν πολύ υψηλές αποδόσεις. Από τις μετρικές και τη μήτρα σύγχυσης, δίνεται μια καλύτερη εικόνα της ικανότητας του κάθε μοντέλου να διακρίνει μεταξύ των κλάσεων. Η μέθοδος LASSO Regularization εμφάνισε σημαντικές αδυναμίες στην ταξινόμηση, όπως και η πλειοψηφία των μοντέλων, καθώς δεν παρουσίασαν υψηλές διακριτικές ικανότητες, ακόμα και η υψηλότερη σε απόδοση μέθοδος, που πέτυχε ακρίβεια 81.5%.

Όσον αφορά το μοντέλο Back Propagation, αντίστοιχα, παρατηρήθηκαν μικτές αποδόσεις από τις διαφορετικές μεθόδους επιλογής χαρακτηριστικών. Καταρχάς, το μοντέλο αυτό με τη μέθοδο Forward Feature Selection πέτυχε με συντριπτική διαφορά την υψηλότερη ακρίβεια (92.5%), στη συνέχεια ακολούθησαν οι μέθοδοι Recursive Feature Elimination, Information Gain, Correlation και Fisher Scores με αρκετά υψηλές αποδόσεις με ακρίβειες 90%, 87.5%, 85% και 82.5% αντίστοιχα. Τα υπόλοιπα μοντέλα σημείωσαν πολύ χαμηλές ακρίβειες με τη χαμηλότερη να φτάνει στο 52.5%.

Οι μετρικές και η μήτρα σύγχυσης ενισχύουν την αποτελεσματικότητα και την αξιοπιστία των ¨δυνατότερων¨ μοντέλων, καθώς παρουσιάζουν πολύ υψηλά σκορ σε AUC, precision, recall και f1-score αναδεικνύοντας ότι τα μοντέλα αυτά έχουν πολύ καλή διακριτική ικανότητα. Για τα μοντέλα που δεν πέτυχαν υψηλές ακρίβειες, όπως τα Decision Trees, LASSO, Random Forest, Sequential Feature Selection και Variance Inflation Factor, φαίνεται να έχουν μικτές αποδόσεις παρουσιάζοντας

υψηλή ανάκληση και χαμηλή ακρίβεια θετικής κλάσης. Για παράδειγμα, τα SFS και VIF παρουσίασαν ανάκληση 100% και ακρίβεια θετικής κλάσης ίση με 57.1%. Τα Lasso και Random Forest επίσης, από τις καμπύλες ROC, φαίνονται να αποδίδουν χειρότερα και από το random guessing.

Για το μοντέλο Gradient Boosting, τα περισσότερα μοντέλα κυμαίνονται μεταξύ ακρίβειες 84% με 74% με εξαίρεση το μοντέλο με τη μέθοδο Lasso Regularization που φτάνει μέγιστη ακρίβεια ίση με περίπου 65%. Την υψηλότερη ακρίβεια την επιτυγχάνει η μέθοδος Correlation και ακολουθούν οι μέθοδοι Decision Trees και FFS. Από τα διαγράμματα φαίνεται πως όλα τα μοντέλα αποδίδουν καλύτερα από το random guessing και έχουν καλές ισορροπίες μεταξύ της ανάκλησης και της ακρίβειας θετικής κλάσης.

Ακολουθεί το μοντέλο K-Nearest Neighbors, το οποίο επιτυγχάνει, ανάλογα με τη μέθοδο, σχετικά υψηλές ακρίβειες (του βελησνεκούς του προηγούμενου μοντέλου, Gradient Boosting) και πολύ χαμηλές, όπως το 51%. Την υψηλότερη ακρίβεια την πετυχαίνει για άλλη μία φορά με τη μέθοδο επιλογής χαρακτηριστικών βάσει της συσχέτισης με 84% και τη χαμηλότερη πάλι με τη LASSO. Άλλα μοντέλα που πετυχαίνουν σχετικά υψηλές αποδόσεις είναι εκείνα με την FFS, IG, Fisher Scores με 80% ακρίβειες, και πολύ χαμηλές ακρίβειες φτάνουν τα μοντέλα με μεθόδους τα Decision Trees, Random Forest, SFS, VIF. Ακόμη και εάν δεν έχουν όλα τα μοντέλα πολύ υψηλή απόδοση και εμφανίζουν μέτρια διακριτική ικανότητα μεταξύ των κλάσεων, δεν υπάρχουν μικτές συμπεριφορές όσον αφορά την ανάκληση και την ακρίβεια θετικής κλάσης, αλλά φαίνεται να διατηρείται μια ισορροπία μεταξύ αυτών των δύο. Γενικότερα, από τα μοντέλα αυτά, εκείνα που πετυχαίνουν ακρίβειες άνω του 80% φαίνεται να είναι αρκετά αξιόπιστα και αποδίδουν αρκετά καλύτερα στην διαδικασία ταξινόμησης.

Προχωρώντας στο μοντέλο Penalized Logistic Regression, τα μοντέλα με τις διαφορετικές μεθόδους φαίνεται να σημειώνουν παρόμοιες ακρίβειες με τις περισσότερες να κυμαίνονται μεταξύ του 83% με 77% με εξαίρεση τη Lasso Regularization, η οποία φτάνει μέγιστη ακρίβεια 69%. Την υψηλότερη την πετυχαίνει η μέθοδος Correlation με 83%. Οι υπόλοιπες μετρικές επιβεβαιώνουν τις διακριτικές ικανότητες των μοντέλων, τα οποία όλα φαίνεται να παρουσιάζουν καλή ισορροπία και ικανοποιητική ικανότητα να διακρίνουν τις κλάσεις μεταξύ τους.

Παρόμοια αποτελέσματα καταγράφει και το μοντέλο της Γραμμικής Παλινδρόμησης με μέγιστη ακρίβεια το 84% που πετυχαίνει με τις μεθόδους Correlation και Sequential Feature Selection. Τη χαμηλότερη ακρίβεια την έχει με τη μέθοδο Lasso Regularization που είναι 71%. Τα διαγράμματα ROC και οι υπόλοιπες μετρικές πάλι επιβεβαιώνουν ότι τα μοντέλα έχουν καλή ισορροπία και φαίνεται να έχουν αρκετά καλή διακριτική ικανότητα, ξεχωρίζοντας ικανοποιητικά τα θετικά από τα αρνητικά περιστατικά.

Το μοντέλο Multilayered Perceptron παρουσιάζει μεγάλες διακυμάνσεις μεταξύ των ακριβειών των μοντέλων, με τη μέγιστη ακρίβεια να φτάνει στο 79% με τη

μέθοδο Fisher Scores και ελάχιστη ακρίβεια 47% με τη μέθοδο Sequential Feature Selection. Αρκετά χαμηλές ακρίβειες έχουν, επίσης, και τα μοντέλα με τις μεθόδους LASSO, IG, Random Forest, VIF, Decision Trees με ακρίβειες κοντά ή και κάτω από το 50%. Για εκείνα τα μοντέλα, φαίνεται και από την καμπύλη ROC ότι αν το μοντέλο επέλεγε να ταξινομήσει κάποιο περιστατικό τυχαία, θα απέδιδε καλύτερα από αυτά, αλλά παρόλα αυτά, φαίνεται οι μετρικές ανάκλησης και ακρίβειας θετικής κλάσης να μην έχουν μεγάλες διαφορές.

Στη συνέχεια, το μοντέλο Naive Bayes παρουσιάζει μέτριες ακρίβειες με την υψηλότερη να είναι στο 84% με τη μέθοδο Decision Trees και τη χαμηλότερη να είναι 67% με τη μέθοδο Lasso Regularization. Οι υπόλοιπες κυμαίνονται μεταξύ των δύο και με τις μετρικές φαίνεται ότι έχουν ικανοποιητικές ικανότητες όσον αφορά την ταξινόμηση των περιστατικών σε κλάσεις, και παρουσιάζουν και κάποια σταθερότητα όσον αφορά τις συμπεριφορές, δηλαδή να διατηρούν μια σχετικά καλή ισορροπία μεταξύ των μετρικών χωρίς μεγάλες αποκλίσεις μεταξύ της ανάκλησης και της ακρίβειας θετικής κλάσης.

Το μοντέλο Random Forest σημειώνει αρκετά υψηλή ακρίβεια με τη μέθοδο Correlation (86%) και οι υπόλοιπες μέθοδοι έχουν ακρίβειες από 82% (με τη μέθοδο VIF και SFS) έως και 67% με τη LASSO. Με τις μετρικές και τις καμπύλες ROC επιβεβαιώνεται η διακριτική ικανότητα των μοντέλων που πετυχαίνουν υψηλά σκορ στην ακρίβεια.

Ακόμη, το μοντέλο Stochastic Gradient Descent φτάνει ακρίβεια 85% με τη μέθοδο Correlation με αμέσως επόμενη υψηλότερη ακρίβεια τη 81% με τα μοντέλα με μεθόδους Fisher Scores και RFE. Πάλι, τα μοντέλα φαίνεται να έχουν ικανοποιητική ικανότητα διάκρισης μεταξύ των κλάσεων.

Το μοντέλο Support Vector Machine επιτυγχάνει μέγιστη ακρίβεια με τη μέθοδο Correlation με 83%. Οι υπόλοιπες μέθοδοι επιτυγχάνουν ακρίβειες λίγο χαμηλότερες από αυτό που είναι, ωστόσο, μέτριες.

Ακόμη, το μοντέλο XGBoost φαίνεται να έχει καλά αποτελέσματα, καθώς επιτυγχάνει μέγιστη ακρίβεια 85% με τη μέθοδο Correlation και τα υπόλοιπα μοντέλα, με εξαίρεση το Lasso, το οποίο επιτυγχάνει μέγιστη ακρίβεια ίση με 71%, κυμαίνονται σε ακρίβειες του 83% με 80% που είναι αρκετά καλές. Η ικανότητα του κάθε μοντέλου να ξεχωρίζει τις κλάσεις του συνόλου δεδομένων είναι ικανοποιητική, κυρίως της μεθόδου με τη μέγιστη ακρίβεια.

Τέλος, με τη μέθοδο PCA τα αποτελέσματα που προκύπτουν ανά μοντέλο είναι αρκετά ενδιαφέροντα καθώς επιτυγχάνεται ως μέγιστη ακρίβεια το 82% των μοντέλων Gradient Boosting και Support Vector Machine αλλά και τα υπόλοιπα μοντέλα δεν πέφτουν από το 74%, το οποίο είναι μέτριο σαν απόδοση. Παρόλα αυτά, έχουν όλα τα μοντέλα μια μέτρια προς καλή ικανότητα διάκρισης μεταξύ των κλάσεων και φαίνονται να έχουν μια ισορροπία μεταξύ των μετρικών της ανάκλησης και ακρίβειας θετικής κλάσης.

5.2 Σύγκριση με State of the Art Αποτελέσματα

Παρόμοιες έρευνες έχουν διεξαχθεί με σκοπό να εκτιμηθεί η πιθανότητα θανάτου ή την εκτίμηση των μετεγχειρητικών επιπλοκών κάποιου ασθενούς που έχει υποβληθεί σε χειρουργείο προκειμένου να αντιμετωπίσει ένα κάταγμα ισχίου. Οι έρευνες αυτές διέφεραν σε συγκεκριμένους τομείς από την παρούσα έρευνα. Καταρχάς, υπήρξαν διαφορές στον όγκο των συνόλων δεδομένων. Μερικές έρευνες είχαν συλλέξει από 1000 μέχρι και 550.000 περιστατικά [2]. Ωστόσο, σε αυτά τα σύνολα δεδομένων ήταν ασθενείς ή περιστατικά που δεν είχαν διαγνωστεί με τα συγκεκριμένα κατάγματα που αναφέρονται στην εργασία αυτή, αλλά γενικότερα ήταν ηλικιωμένοι που είχαν υποστεί κάποιο χειρουργείο για οποιοδήποτε κάταγμα ισχίου χαμηλής ενέργειας (low-energy hip fracture) [66]. Υπήρξαν και έρευνες που είχαν συλλέξει κοντά στους 700 ασθενείς [67] που όμως είχαν διαγνωστεί με τα συγκεκριμένα κατάγματα ισχίου. Στις έρευνες αυτές, όπως και στην παρούσα, ένας από τους κύριους στόχους ήταν να εντοπίσουν χαρακτηριστικά τα οποία θεωρούνται μεγαλύτεροι παράγοντες κινδύνου (risk factors) για τους ασθενείς αυτούς. Με άλλα λόγια, ποιοι παράγοντες που διαθέτουν οι ασθενείς αυξάνουν τον κίνδυνο θνησιμότητάς τους. Πολλά από τα χαρακτηριστικά/παράγοντες που εντοπίστηκαν ως παράγοντες κινδύνου ήταν τα ακόλουθα: χρόνος μέχρι το χειρουργείο από τη στιγμή του κατάγματος, η διάρκεια διαμονής στο νοσοκομείο, κατάσταση διαμονής (σε σπίτι ή σε κέντρο), καρδιαγγειακά νοσήματα [68], διαβήτης, λοίμωξη χειρουργικού τραύματος (SSI - Surgical Site Infection), η περίοδος που πραγματοποιήθηκε το χειρουργείο, ο δείκτης μάζας σώματος (BMI - Body Mass Index), η αναιμία, η λήψη θεραπείας με κορτικοστεροειδή (corticosteroids), αυξημένη χρήση αντιβιοτικών και τα επίπεδα αιμοσφαιρίνης (hemoglobin levels), παχυσαρκία (συγκεκριμένα ο λόγος της περιμέτρου της μέσης με την περίμετρο της περιφέρειας) [67]. Ακόμη, άλλοι παράγοντες που φάνηκαν να είναι σημαντικοί ήταν η κινητικότητα πριν το κάταγμα (pre-fracture mobility), η γνωστική εξασθένηση (cognitive impairment), το μετεγχειρητικό παραλήρημα (post-operative delirium), ο χρόνος αποκατάστασης, η χειρουργική τεχνική, η θεραπεία για οστεοπόρωση (osteoporosis), η επανεισαγωγή στο νοσοκομείο (readmission)[69]. Επιπλέον, σημαντικοί παράγοντες ήταν οι πνευμονικές νόσοι (lung disease), η άνοια (dementia), το σκορ CCI - Charlson Comorbidity Index, η σαρκοπενία (sarcopenia), άλλες παθήσεις όπως η οστεοαρθρίτιδα (osteoarthritis), η δύναμη λαβής (grip strength), η ευαισθησία, το κάπνισμα (smoking) και το είδος αναισθησίας (anesthesia type).

Διεξάχθηκαν και έρευνες που εξέταζαν τους παράγοντες που μπορούσαν να βοηθήσουν στην εκτίμηση της διάρκειας διαμονής του ασθενούς στο νοσοκομείο. Μερικοί παράγοντες που βρέθηκαν για αυτό περιλαμβάνουν την κοινωνική και οικονομική κατάσταση του ασθενούς, το είδος του κατάγματος, το φύλο, η ηλικία [70], η εθνικότητα [71], δείκτες όπως το ασβέστιο (calcium) πριν την επέμβαση, το ποσοστό των λεμφοκυττάρων (lymphocyte percentage), η ενδοεγχειρητική αιμορραγία (intraoperative bleeding), η χορήγηση γλυκόζης και χλωριούχου νατρίου (sodium

chloride) μετά την επέμβαση, και το σκορ Charlson Comorbidity Index [72]. Ένας ακόμη παράγοντας που εντοπίστηκε από μία έρευνα ήταν η διάρκεια του χειρουργείου **γαρσια2012πατιεντ**. Σε μερικές έρευνες ο αλγόριθμος απαιτούσε να εισάγει ο χρήστης μόνο τα χαρακτηριστικά ηλικία, φύλο, εθνικότητα και τα comorbidity scores [73], [74] προκειμένου α εκτιμήσει τη διάρκεια παραμονής στο νοσοκομείο. Οι παράγοντες αυτοί είχαν βρεθεί, επίσης, και ως καθοριστικοί παράγοντες κινδύνου για τον θάνατο.

Πολλοί από τους παραπάνω παράγοντες είχαν βρεθεί και στην παρούσα έρευνα, καθώς όπως είχε αναφερθεί και παραπάνω, κάθε μέθοδος επιλογής χαρακτηριστικών εντόπιζε διαφορετικά χαρακτηριστικά βάσει του αντίστοιχου κριτηρίου της μεθόδου αυτής.

Σε μερικές έρευνες που έκαναν απόπειρα να εκτιμήσουν με τη χρήση μοντέλων μηχανικής μάθησης (machine learning) την πιθανότητα θανάτου του ασθενούς, οι έρευνες αυτές πειραματίστηκαν με μοντέλα όπως: Gradient Boosting Classifier (GB), Random Forests Classifier (RF), Artificial Neural Network Classifier (ANN), Logistic Regression Classifier (LR), Naive Bayes Classifier (NB), Support Vector Machine Classifier (SVM) και K-Nearest Neighbors Classifier (KNN) [75]. Από αυτά, μερικά αποτελέσματα που έβγαλαν είχαν τις εξής ακρίβειες: GB model = 93%, RF = 95%, ANN = 94%, LR = 91%, NB = 89%, SVM = 90% και KNN = 90%, με υψηλές τιμές στη βαθμολογία της καμπύλης χαρακτηριστικής λειτουργίας του δέκτη (AUC - Area Under the Curve) (από 81% έως και 99%) και με πολύ υψηλή αρνητική προβλεπτική τιμή (Negative Predictive Value). Μία ακόμη έρευνα, εκτίμησε με τη βοήθεια των μοντέλων Artificial Neural Networks, Logistic Regression Naive Bayes την πιθανότητα θνησιμότητας ενός ασθενούς εντός των 30 ημερών κατά την οποία πέτυχε ακρίβειες 92%, 87% και 83% αντίστοιχα [2]. Άλλη έρευνα που χρησιμοποίησε παρόμοια μοντέλα, μερικά για την εκτίμηση της διάρκειας παραμονής στο νοσοκομείο και άλλα για να εκτιμήσουν τον θάνατο του ασθενούς, πέτυχαν ακρίβειες που κυμαίνονταν από το 68% μέχρι και το 95% [28]. Η έρευνα που ήθελε να εκτιμήσει τη διάρκεια παραμονής στο νοσοκομείο LOS, με τη χρήση του αλγόριθμου Naive Bayes πέτυχε ακρίβεια στο 87%.

Είναι, επομένως, εμφανές ότι υπάρχουν πολλοί παράγοντες που μπορούν να επηρεάσουν τις επιπλοκές αλλά και τη θνησιμότητα ενός ασθενούς όσον αφορά την έκβαση ενός χειρουργείου για την αποκατάσταση από ένα κάταγμα ισχίου.

5.3 Συμπεράσματα

Καλύτεροι Εκτελεστές: Τα μοντέλα Back Propagation με Forward Feature Selection και Recursive Feature Elimination είναι οι καλύτεροι και πιο αξιόπιστοι εκτελεστές, παρουσιάζοντας ακρίβειες 92.5% και 90% αντίστοιχα. Επίσης, το Gradient Boosting και Logistic Regression με Correlation και SFS έδειξαν εξαιρετικές αποδόσεις οι οποίες είναι συγκρίσιμες με εκείνες των καλύτερων όσον αφορά την

απόδοση μοντέλων από τις προαναφερόμενες έρευνες.

Χειρότερες Αποδόσεις: Τα Decision Trees και LASSO Regularization είχαν τις χειρότερες αποδόσεις, με το Decision Trees να πετυχαίνει ακρίβεια μόλις 52.5% και LASSO να παρουσιάζει παρόμοια απόδοση σε πολλά μοντέλα. Η τόση μεγάλη αποτυχία της μεθόδου LASSO L1 Regularization μπορεί να οφείλεται στο γεγονός ότι η μέθοδος αυτή επιλέγει μόνο δύο χαρακτηριστικά από όλα τα διαθέσιμα που υπάρχουν στο σύνολο δεδομένων, εφόσον έχει οριστεί κάποιο κατώφλι που έχει τεθεί ως ¨αποδεκτές¨ τιμές που μπορεί να αποδώσει η μέθοδος αυτή για να θεωρήσει το χαρακτηριστικό ¨χρήσιμο¨ ή ¨ποιοτικό¨ για περεταίρω επεξεργασία.

Ισορροπία Μεταξύ Μετρικών: Τα περισσότερα μοντέλα που πετυχαίνουν υψηλή ακρίβεια διατηρούν μια καλή ισορροπία μεταξύ των μετρικών ανάκλησης και ακρίβειας θετικής κλάσης, υποδηλώνοντας την ικανότητά τους να διακρίνουν μεταξύ των κλάσεων με ακρίβεια.

Επιλογή Χαρακτηριστικών: Οι μέθοδοι επιλογής χαρακτηριστικών έχουν σημαντικό αντίκτυπο στην απόδοση των μοντέλων. Η σωστή επιλογή μπορεί να βελτιώσει σημαντικά την ακρίβεια και άλλες μετρικές απόδοσης του μοντέλου. Πέραν αυτού, παίζει πολύ ρόλο ο συνδυασμός μεθόδου επιλογής χαρακτηριστικών με το ίδιο το μοντέλο. Με άλλα λόγια, μερικές μέθοδοι ενώ αποδίδουν αρκετά χαμηλά με τα περισσότερα μοντέλα, όπως για παράδειγμα τα δέντρα αποφάσεων με τη πλειοψηφία των μοντέλων δεν πετύχαινε υψηλές ακρίβειες, με μερικά μοντέλα ¨ταιριάζουν¨ και επιτυγχάνουν μεγάλες, ή έστω μεγαλύτερες από άλλες μεθόδους ακρίβειες, όπως στην περίπτωση με το μοντέλο XGBoost και Naive Bayes.

Αποδόσεις με PCA: Η μέθοδος PCA παρείχε καλές αποδόσεις σε πολλά μοντέλα, με μέγιστη ακρίβεια 82%. Παρόλο που δεν είναι η κορυφαία επιλογή, παρουσιάζει σταθερότητα και καλή ισορροπία μεταξύ των μετρικών.

Πολλά από τα αποτελέσματα από πολλά μοντέλα της παρούσας έρευνας ήταν πολύ ικανοποιητικά, ιδίως των καλύτερων σε απόδοση μοντέλων (*βασκ προπαγατιον με ΦΦΣ*), λαμβάνοντας υπόψη ότι εκπαιδεύτηκαν με ένα σύνολο δεδομένων των 400 δειγμάτων και τη μέθοδο *χροσσ-αλιδατιον*. Από τις αντίστοιχες έρευνες, τα μοντέλα αυτά μπορούν να συγκριθούν με τις ακρίβειες και τα αποτελέσματα των καλύτερων μοντέλων των ερευνών αυτών, καθώς τα ξεπερνούν για ένα ελάχιστο ποσοστό ακρίβειας. Πιο συγκεκριμένα, το μοντέλο *ΡΦ* της έρευνας [28] πέτυχε 2.5% μεγαλύτερη ακρίβεια, το μοντέλο *ΑΝΝ* της [28] πέτυχε 1.5% παραπάνω και το μοντέλο *ΓΒ* της έρευνας [28] πέτυχε 0.5% πιο ψηλή ακρίβεια από το μοντέλο *ΒΠ με ΦΦΣ* της παρούσας έρευνας.

5.4 Περιορισμοί

Στην παρούσα εργασία υπήρξαν μερικοί περιορισμοί. Ο πιο βασικός περιορισμός ήταν η έλλειψη δεδομένων. Συγκριτικά με τις περισσότερες έρευνες, η συγκεκριμένη είχε 400 δείγματα χωρίς missing values. Ο λόγος που ήταν τόσα ήταν

λόγω του ότι αφορούσαν αποκλειστικά περιστατικά που είχαν έρθει στο νοσοκομείο ΚΑΤ τα τελευταία χρόνια από το 2019. Επίσης, η ίδια η διαδικασία συλλογής των δεδομένων αποτέλεσε έναν περιορισμό, καθώς όσοι συμμετείχαν στη συλλογή τους, προκειμένου να συλλέξουν συγκεκριμένες πληροφορίες, έπρεπε να επικοινωνήσουν τηλεφωνικώς με τον ασθενή ή με κάποιο μέλος της οικογένειάς του, με αποτέλεσμα να υπάρχει κίνδυνος ύπαρξης ανακριβειών. Από τους 700 συνολικά ασθενείς που είχαν περάσει από το νοσοκομείο ΚΑΤ, μόνο τα 400 αποτέλεσαν έγκυρα δείγματα, με την έννοια να μην απουσιάζουν πληροφορίες.

5.5 Μελλοντικές Επεκτάσεις

Αξίζει να σημειωθεί ότι, εφόσον με τα δεδομένα αυτά έχουν υπάρξει μοντέλα που επιστρέφουν αποτελέσματα μεγάλης ακρίβειας, η πρόβλεψη της θνησιμότητας μπορεί να γίνει ακόμα πιο ακριβής με τη χρήση εικόνων και τη βοήθεια πιο πολύπλοκων μοντέλων μηχανικής μάθησης, όπως τα συνελικτικά νευρωνικά δίκτυα (CNN - Convolutional Neural Networks) και πολλά ακόμα. Επίσης, ενδεχομένως να μπορεί να επεκταθεί περαιτέρω η έρευνα ώστε, αντί να ταξινομεί τα περιστατικά σε δυαδική έξοδο (1: θάνατος, 0: όχι), να εκτιμά όχι μόνο τη θνησιμότητα ή ένα συγκεκριμένο χαρακτηριστικό, αλλά και άλλες επιπλοκές που μπορεί να παρουσιάσει ο ασθενής μετά την επέμβαση, εφόσον το σύνολο δεδομένων είναι αρκετά μεγάλο.

Βιβλιογραφία

[1] Sadeghi, Omid, et al., «Abdominal Obesity and Risk of Hip Fracture: A Systematic Review and Meta-Analysis of Prospective Studies,» *Advances in Nutrition*, τόμ. 8, αρθμ. 5, σσ. 728–738, 2017.

[2] DeBaun, Michael R. and others, «Artificial Neural Networks Predict 30-Day Mortality After Hip Fracture: Insights From Machine Learning,» *Journal of the American Academy of Orthopaedic Surgeons*, τόμ. 29, αρθμ. 22, σσ. 977–983, Νοέ. 2021.

[3] Lu, Young and Uppal, Harmeeth S., «Hip Fractures: Relevant Anatomy, Classification, and Biomechanics of Fracture and Fixation,» *Geriatric Orthopaedic Surgery & Rehabilitation*, JulyJuly 2019.

[4] Pollard, T. C., et al., «Deep Wound Infection after Proximal Femoral Fracture: Consequences and Costs,» *The Journal of Hospital Infection*, AprApr 2006.

[5] Tay, Eileen, «Hip Fractures in the Elderly: Operative versus Nonoperative Management,» *Singapore Medical Journal*, AprApr 2016.

[6] Williamson, S., et al., «Costs of Fragility Hip Fractures Globally: A Systematic Review and Meta-Regression Analysis,» *Osteoporosis International*, τόμ. 28, σσ. 1617–1628, 2017.

[7] Young, Alexander L., et al., «Mutual Information: Measuring Nonlinear Dependence in Longitudinal Epidemiological Data,» *PLOS ONE*, JulyJuly 2011.

[8] Mark D. Miller and Stephen R. Thompson and Jennifer Hart, *Review of Orthopaedics*, 6η έκδοση. Saunders, Μάι. 2012.

[9] Attum, Basem, *Intertrochanteric Femur Fracture*. U.S. National Library of Medicine, 2023.

[10] Ruth Delahunty, *Anatomy 101 - The hips*, Accessed: 2023-07-04, MayMay 2022.

[11] Kazley, Jillian, Banerjee, Samik, Abousayed, Mostafa και Rosenbaum, Andrew, «Classifications in Brief: Garden Classification of Femoral Neck Fractures,» *Clinical Orthopaedics and Related Research*, τόμ. 476, 441–445, FebFeb 2022.

[12] Kazley, Jillian, «Femoral Neck Fractures,» *StatPearls [Internet]*, MayMay 2023.

[13] Lu, Young and Uppal, Harmeeth, «Hip Fractures: Relevant Anatomy, Classification, and Biomechanics of Fracture and Fixation,» *Geriatric Orthopaedic Surgery Rehabilitation*, τόμ. 10, σ. 215 145 931 985 913, Ιούλ. 2019.

[14] Christopher Doro, MD, *Anatomy 101 - The hips*, Accessed: 2023-07-04, 2022.

[15] Booth, K. C., et al., «Femoral Neck Fracture Fixation: A Biomechanical Study of Two Cannulated Screw Placement Techniques,» *Orthopedics*, NovNov 1998.

[16] Guyen, O., «Hemiarthroplasty or Total Hip Arthroplasty in Recent Femoral Neck Fractures?» *Orthopaedics & Traumatology, Surgery & Research: OTSR*, FebFeb 2019.

[17] Parker, M. J., et al., «Hemiarthroplasty versus Internal Fixation for Displaced Intracapsular Hip Fractures in the Elderly. A Randomised Trial of 455 Patients,» *The Journal of Bone and Joint Surgery. British Volume*, NovNov 2002.

[18] Veldman, H. D., et al., «Cemented versus Cementless Hemiarthroplasty for a Displaced Fracture of the Femoral Neck: A Systematic Review and Meta-Analysis of Current Generation Hip Stems,» *The Bone & Joint Journal*, AprApr 2017.

[19] Gupta, R. K., et al., «Unstable Trochanteric Fractures: The Role of Lateral Wall Reconstruction,» *National Center for Biotechnology Information*, FebFeb 2010.

[20] Tawari, A. A., et al., «What Makes an Intertrochanteric Fracture Unstable in 2015? Does the Lateral Wall Play a Role in the Decision Matrix?» *Journal of Orthopaedic Trauma*, AprApr 2015.

[21] Haidukewych, G. J., et al., «Reverse Obliquity Fractures of the Intertrochanteric Region of the Femur,» *The Journal of Bone and Joint Surgery. American Volume*, MayMay 2001.

[22] Mark Karadsheh, MD and Daniel Tarazona, MD, *Intertrochanteric Fractures*, Updated: Feb 29, 2024. Accessed: 2023-07-04, FebFeb 2024.

[23] O'Neill, F., et al., «Dynamic Hip Screw versus DHS Blade: A Biomechanical Comparison of the Fixation Achieved by Each Implant in Bone,» *The Journal of Bone and Joint Surgery. British Volume*, MayMay 2011.

[24] Samuel, A. L., «Some Studies in Machine Learning Using the Game of Checkers,» *IBM Journal of Research and Development*, τόμ. 3, αρθμ. 3, σσ. 210–229, 1959.

[25] Park, Cheolsoo, et al., «Machine Learning in Biomedical Engineering - Biomedical Engineering Letters,» *SpringerLink*, FebFeb 2018.

[26] Nasteski, Vladimir, «An Overview of the Supervised Machine Learning Methods,» DecDec 2017.

[27] Bashar Rajoub, «Chapter 3 - Supervised and unsupervised learning,» στο *Biomedical Signal Processing and Artificial Intelligence in Healthcare*, σειρά Developments in Biomedical Engineering and Bioelectronics, Walid Zgallai, επιμελητής, Academic Press, 2020, σσ. 51–89.

[28] Lex, Johnathan R., et al., «Artificial Intelligence for Hip Fracture Detection and Outcome Prediction: A Systematic Review and Meta-analysis,» *JAMA Network Open*, τόμ. 6, αρθμ. 3, ε233391, 2023.

[29] Kira, Kenji and Rendell, Larry A., «A Practical Approach to Feature Selection,» *ScienceDirect*, JuneJune 2014.

[30] Bouchefry, Khadija El and de Souza, Rafael S., «A Practical Approach to Feature Selection,» *ScienceDirect*, AprApr 2020.

[31] Jie, Cai, et al., «Feature Selection in Machine Learning: A New Perspective,» *ScienceDirect*, MarMar 2018.

[32] Tang, Jiliang, Alelyani, Salem και Liu, Huan, «Feature Selection for Classification: A Review,» *Michigan State University*, 2014.

[33] Blum, Avrim L. and Langley, Pat, «Selection of Relevant Features and Examples in Machine Learning,» *ScienceDirect*, 1998.

[34] Chen, Tianqi and Guestrin, Carlos, «XGBoost: A Scalable Tree Boosting System,» JuneJune 2016.

[35] Jeon, Hyelynn and Oh, Sejong, «Hybrid-Recursive Feature Elimination for Efficient Feature Selection,» *MDPI*, MayMay 2020.

[36] Zhao, Xi, et al., «Feature Selection with Attributes Clustering by Maximal Information Coefficient,» *ScienceDirect*, MayMay 2013.

[37] Ying-Ao Wang and Qin Huang and Zhigang Yao and Ye Zhang, «On a class of linear regression methods,» *Journal of Complexity*, τόμ. 82, σ. 101 826, 2024.

[38] Maulud, Dastan Hussen and Abdulazeez, Adnan Mohsin, «A Review on Linear Regression Comprehensive in Machine Learning,» *Journal of Applied Science and Technology Trends*, 2020.

[39] Castillo, D., «Decision trees for classification: A machine learning algorithm,» *Seldon Blog*, 2021.

[40] Mayur Kulkarni, *Decision Trees For Classification: A Machine Learning Algorithm*, Accessed: 2023-07-04, SeptemberSeptember 2017.

[41] Saed Sayad, *Decision Tree*, Accessed: 2023-07-04, 2023.

[42] Unknown, *Decision Tree*, Accessed: 2023-07-04, MarchMarch 2021.

[43] Breiman, Leo, «Consistency for a Simple Model of Random Forests,» SeptSept 2004.

[44] Pijush Dutta and Shobhandeb Paul and Asok Kumar, «Chapter 25 - Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19,» στο *Electronic Devices, Circuits, and Systems for Biomedical Applications*, Suman Lata Tripathi and Valentina E. Balas and S.K. Mohapatra and Kolla Bhanu Prakash and Janmenjoy Nayak, επιμελητής, Academic Press, 2021, σσ. 521–540.

[45] Uduak A. Umoh and Imo J. Eyoh and Vadivel S. Murugesan and Emmanuel E. Nyoho, «Chapter 14 - Fuzzy-machine learning models for the prediction of fire outbreaks: A comparative analysis,» στο *Artificial Intelligence and Machine Learning for EDGE Computing*, Rajiv Pandey and Sunil Kumar Khatri and Neeraj kumar Singh and Parul Verma, επιμελητής, Academic Press, 2022, σσ. 207–233.

[46] Ajitesh Kumar, *Random Forest Classifier - Sklearn Python Example*, Accessed: 2023-07-04, Last updated: 13th Dec, 2023, DecemberDecember 2023.

[47] Arjun Chandran, «How artificial neural networks work, from the math up,» *Berkeley Scientific Journal*, DecemberDecember 2019, UC Berkeley's Premier Undergraduate Science Journal.

[48] Sathya, R. and Abraham, Annamma, «Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification,» *Ijarai*, τόμ. 2, αρθμ. 2, 2013.

[49] Lecun, Yann, «A Theoretical Framework for Back-Propagation,» AugAug 2001.

[50] Buscema, Massimo, «Back Propagation Neural Networks,» *Taylor & Francis*, JulyJuly 2009.

[51] Hong, Won-Kee, *Artificial Intelligence-Based Design of Reinforced Concrete Structures: Artificial Neural Networks for Engineering Applications*. Woodhead Publishing, 2023.

[52] Mbali Kalirane, *Gradient Descent vs. Backpropagation: What's the Difference?* Accessed: 2023-07-04, AprilApril 2023.

[53] Hong, Won-Kee, «Factors Influencing Network Trainings,» JuneJune 2023.

[54] Shi, Yuanming, et al., «Primer on Artificial Intelligence,» SeptSept 2021.

[55] Δημήτρης, *Μηχανική Μάθηση*, Α. Εκδόσεις Κλειδάριθμος, 2019.

[56] Yuanming Shi, Yong Zhou κ.ά., «Primer on artificial intelligence,» στο *Mobile Edge Artificial Intelligence*, Science Direct, 2022, κεφ. k-Nearest neighbors method.

[57] Rani, Alka, et al., «Machine Learning for Soil Moisture Assessment,» JanJan 2022.

[58] Anshul Saini, *Support Vector Machines (SVM) - A Complete Guide for Beginners*, Accessed: 2023-07-04, OctoberOctober 2021.

[59] Collins, Michael, «The Naive Bayes Model, Maximum-Likelihood Estimation,» 2002.

[60] Lewis, David D., «Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval,» *SpringerLink*, JanJan 1998.

[61] Natekin, Alexey and Knoll, Alois, «Gradient Boosting Machines, a Tutorial,» *Frontiers*, OctOct 2013.

[62] Algamal, Zakariya Yahya and Lee, Muhammad Hisyam, «High Dimensional Logistic Regression Model using Adjusted Elastic Net Penalty,» *Pakistan Journal of Statistics and Operation Research*, τόμ. 11, σσ. 667–676, 2015.

[63] Friedman, Jerome H., «Stochastic Gradient Boosting,» *ScienceDirect*, JanJan 2002.

[64] Shaw Talebi, *10 Decision Trees are Better Than 1: Breaking down bagging, boosting, Random Forest, and AdaBoost*, Accessed: 2023-07-04, FebruaryFebruary 2023.

[65] Daniel Jurafsky and James H. Martin, *Speech and Language Processing*. Stanford University, 2023, κεφ. 5.

[66] Simunovic, Nicole, et al., «Effect of Early Surgery after Hip Fracture on Mortality and Complications: Systematic Review and Meta-Analysis,» *CMAJ*, τόμ. 182, αριθμ. 15, σσ. 1609–1616, 2010.

[67] Ji, Chenni, et al., «Incidence and Risk of Surgical Site Infection after Adult Femoral Neck Fractures Treated by Surgery: A Retrospective Case-Control Study,» *Medicine*, MarMar 2019.

[68] Chang, Wenli and others, «Preventable risk factors of mortality after hip fracture surgery: Systematic review and meta-analysis,» *International Journal of Surgery (London, England)*, τόμ. 52, σσ. 320–328, 2018.

[69] Smith, Toby, et al., «Pre-Operative Indicators for Mortality Following Hip Fracture Surgery: A Systematic Review and Meta-Analysis,» *Age and Ageing*, τόμ. 43, αριθμ. 4, σσ. 464–471, 2014.

[70] Vaseenon, Tanawat, et al., «Long-Term Mortality after Osteoporotic Hip Fracture in Chiang Mai, Thailand,» *Journal of Clinical Densitometry*, τόμ. 13, αριθμ. 1, σσ. 63–67, 2010.

[71] Xu, Donghui, et al., «Concrete and Steel Bridge Structural Health Monitoring-Insight into Choices for Machine Learning Applications,» *Construction and Building Materials*, AugAug 2023.

[72] Zhong, Hao, et al., «The Application of Machine Learning Algorithms in Predicting the Length of Stay Following Femoral Neck Fracture,» *International Journal of Medical Informatics*, τόμ. 155, σ. 104 572, 2021.

[73] Ramkumar, Prem N. and others, «Development and Validation of a Machine Learning Algorithm After Primary Total Hip Arthroplasty: Applications to Length of Stay and Payment Models,» *Τηε Θουρναλ οφ Αρτηροπλαστψ*, τόμ. 34, αρθμ. 4, σσ. 632–637, 2019.

[74] Hu, Fangke and others, «Preoperative predictors for mortality following hip fracture surgery: a systematic review and meta-analysis,» *Injury*, τόμ. 43, αρθμ. 6, σσ. 676–685, 2012.

[75] Kitcharanant, Nitchanant, et al., «Development and internal validation of a machine-learning-developed model for predicting 1-year mortality after fragility hip fracture,» *BMC Geriatrics*, τόμ. 22, αρθμ. 1, σ. 451, 2022.