ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Πρόβλεψη εμφάνισης στεφανιαίας νόσου με τη χρήση αλγορίθμων μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ανδρέα Β. Γαλάτη

Επιβλέπων: Γιώργος Ματσόπουλος
Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Πρόβλεψη εμφάνισης στεφανιαίας νόσου με τη χρήση αλγορίθμων μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ανδρέα Β. Γαλάτη

Επιβλέπων: Γιώργος Ματσόπουλος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 7η Νοεμβρίου 2023

……...	……...	……...
Γιώργος Ματσόπουλος Καθηγητής Ε.Μ.Π.	Αθανάσιος Παναγόπουλος Καθηγητής Ε.Μ.Π	Παναγιώτης Τσανάκας Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2023

.....

Ανδρέας Γαλάτης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

Περίληψη

Στον τομέα της υγείας, η μηχανική μάθηση έχει σημαντικό ρόλο στην ανάλυση μεγάλου όγκου δεδομένων, βοηθώντας τους επαγγελματίες υγείας στη διάγνωση ή πρόληψη ασθενειών. Οι καρδιαγγειακές παθήσεις είναι χρόνιες ασθένειες με παγκόσμια εξάπλωση και σοβαρές επιπλοκές για τους πάσχοντες και αποτελούν την κύρια αιτία θανάτου παγκοσμίως. Η συχνότερη μεταξύ αυτών είναι η στεφανιαία νόσος. Με στόχο την πρόβλεψη και κατ' επέκταση την πρόληψη της εμφάνισης στεφανιαίας νόσου και των επιπλοκών της, στην παρούσα εργασία αναπτύχθηκε ερμηνεύσιμο μοντέλο εκτίμησης του κινδύνου εμφάνισης καρδιακής νόσου με την χρήση αλγορίθμων μηχανικής μάθησης. Χρησιμοποιήθηκαν δεδομένα 319.795 ατόμων που συλλέχθηκαν από το Κέντρο Ελέγχου και Πρόληψης Ασθενειών των Η.Π.Α. (CDC) και περιλαμβάνουν 18 χαρακτηριστικά που σχετίζονται με τη γενική υγεία και τις συνήθειες των συμμετεχόντων. Για την αποτελεσματική διαχείριση της μη ισορροπημένης φύσης των δεδομένων διερευνήθηκαν ποικίλες τεχνικές εξισορρόπησής τους. Η ανάπτυξη του μοντέλου υλοποιήθηκε σε γλώσσα προγραμματισμού Python με χρήση των αλγορίθμων K-Nearest Neighbors, Naive Bayes, Logistic Regression, Decision Tree, Random Forest και Multi-layer Perceptron (MLP). Το κάθε μοντέλο αξιολογήθηκε ως προς τη διακριτική του ικανότητα, ενώ οι μέθοδοι ερμηνευσιμότητας Permutation Feature Importance και LIME προσέφεραν πολύτιμες πληροφορίες για την βαρύτητα κάθε χαρακτηριστικού στην πρόβλεψη. Μεταξύ των ανωτέρω μοντέλων τα καλύτερα αποτελέσματα έδωσαν τα μοντέλα με Logistic Regression, Random Forest και MLP με ευαισθησία 81% και ειδικότητα 72%, 71% και 72% αντίστοιχα.

Λέξεις Κλειδιά: στεφανιαία νόσος, παράγοντες κινδύνου στεφανιαίας νόσου, αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης, προεπεξεργασία δεδομένων, υποδειγματοληψία, υπερδειγματοληψία, καμπύλες εκπαίδευσης, ερμηνευσιμότητα μοντέλων μηχανικής μάθησης

Abstract

In the healthcare sector, machine learning has an important role in analyzing big data, helping healthcare professionals to diagnose or prevent diseases. Cardiovascular diseases are chronic diseases with a global prevalence and serious complications for patients and are the leading cause of death worldwide. The most common among them is coronary heart disease. In order to predict and thus prevent the occurrence of coronary artery disease and its complications, an interpretive model for estimating the risk of heart disease using machine learning algorithms was developed in this paper. Data of 319.795 individuals collected from the U.S. Centers for Disease Control and Prevention (CDC) were used regarding 18 characteristics related to the participants' general health and habits. A variety of balancing techniques were explored to effectively manage the unbalanced nature of the data. The model was developed in Python programming language and several supervised learning algorithms were used. Each model was evaluated for its discriminative ability, while the Permutation Feature Importance and LIME interpretability methods provided valuable insights into the weight of each feature in the prediction. Among the models tested, the best results were obtained by the models with Logistic Regression, Random Forest and MLP with sensitivity 81% and specificity 72%, 71% and 72% respectively.

Keywords: coronary heart disease, coronary heart disease risk factors, supervised machine learning algorithms, data preprocessing, undersampling, oversampling, learning curves, interpretability of machine learning models

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ. Γιώργο Ματσόπουλο για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου αυτό το εξαιρετικά ενδιαφέρον θέμα. Ευχαριστώ επίσης το μέλος ΕΔΙΠ κα Ουρανία Πετροπούλου για τις πολύτιμες συμβουλές της.

Επιπλέον θα ήθελα να ευχαριστήσω την υποψήφια Διδάκτορα κα Ολυμπία Γιαννακοπούλου για τη συνεχή καθοδήγηση και υποστήριξη που μου προσέφερε, καθώς και για τον πολύτιμο χρόνο που διέθεσε για την υλοποίηση αυτής της εργασίας.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου για την συνεχή συμπαράσταση που μου προσέφερε καθ' όλη τη διάρκεια των σπουδών μου.

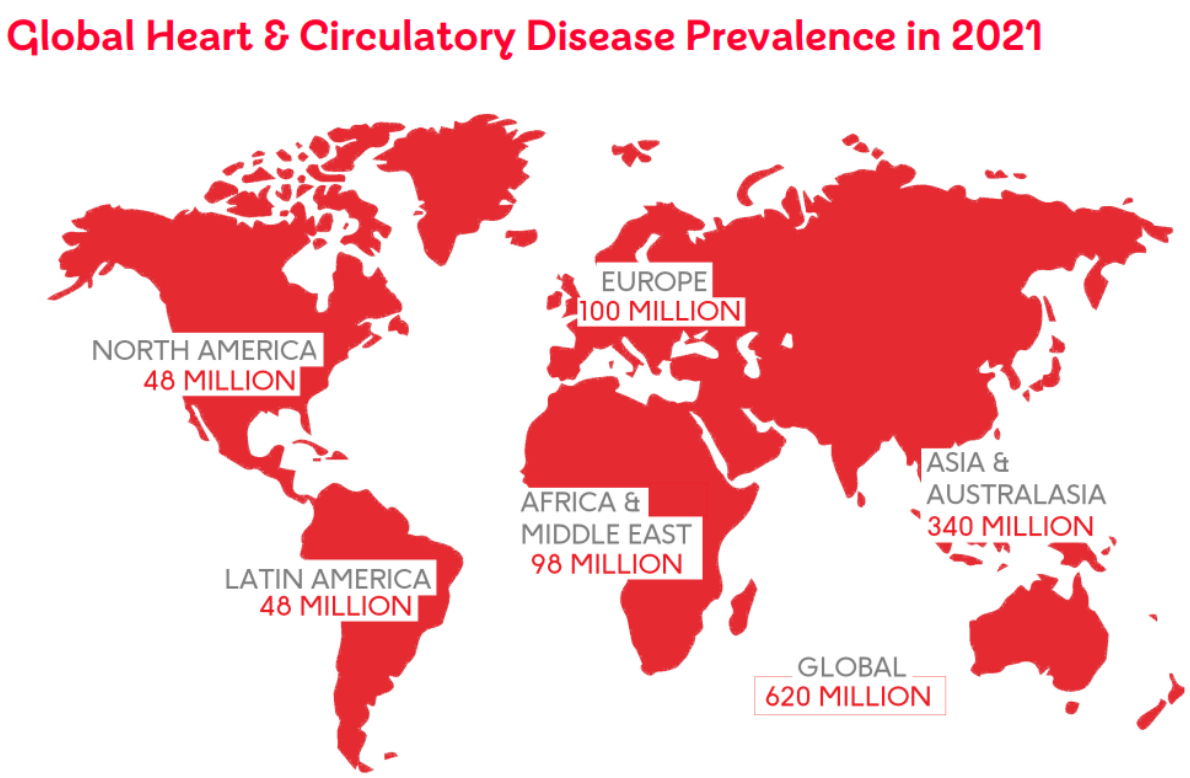
Περιεχόμενα

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

Κεφάλαιο 1: Στατιστικά στοιχεία καρδιαγγειακών νοσημάτων

1.1 Στατιστικά στοιχεία παγκοσμίως



***Εικόνα 1**: Παγκόσμιος επιπολασμός ΚΝ. 620 εκατομμύρια άνθρωποι πάσχουν από ΚΝ παγκοσμίως* [1]

Σύμφωνα με στατιστικά στοιχεία του 2021, περίπου 620 εκατομμύρια άνθρωποι, δηλ 1 στους 13, πάσχουν από καρδιαγγειακά νοσήματα (ΚΝ) σε όλο τον κόσμο [2] (Εικόνα 1). Το 2019 ήταν αντίστοιχα 550 εκατομμύρια, από τους οποίους 290 εκατομμύρια ήταν γυναίκες (53%) και 260 εκατομμύρια άνδρες (47%) [3].

Ο αριθμός αυτός αυξάνεται διαρκώς λόγω της αλλαγής του τρόπου ζωής, της γήρανσης, του αυξανόμενου πληθυσμού και των βελτιωμένων ποσοστών επιβίωσης από καρδιακές προσβολές και εγκεφαλικά επεισόδια και θα συνεχίσουν να αυξάνονται εάν συνεχιστούν αυτές οι τάσεις. Το 1990 285 εκατομμύρια άνθρωποι έπασχαν από ΚΝ παγκοσμίως. Ο αριθμός αυτός αυξήθηκε σε 350 εκατομμύρια το 2000 και περισσότερο από 430 εκατομμύρια το 2010. Αυτό σημαίνει ότι από το 1990 έως σήμερα ο αριθμός των ανθρώπων με ΚΝ παγκοσμίως έχει αυξηθεί κατά 93% [4].

Από τις καρδιαγγειακές παθήσεις το 2019, οι πιο συχνές είναι η στεφανιαία (ισχαιμική) καρδιακή νόσος (παγκόσμια επίπτωση εκτιμάται σε 200 εκατομμύρια), η περιφερική αρτηριακή (αγγειακή) νόσος (110 εκατομμύρια), εγκεφαλικό επεισόδιο (100 εκατομμύρια) και κολπική μαρμαρυγή (60 εκατομμύρια). Κάθε χρόνο περίπου 60 εκατομμύρια άνθρωποι σε όλο τον κόσμο αναπτύσσουν ΚΝ [3].

Επιπλέον τα ΚΝ σύμφωνα με τα δεδομένα του Παγκόσμιου Οργανισμού Υγείας (ΠΟΥ) [5] είναι η κύρια αιτία θανάτου παγκοσμίως. Εκτιμάται ότι 17,9 εκατομμύρια άνθρωποι πέθαναν

από ΚΝ το 2019, που αντιπροσωπεύουν το 32% όλων των θανάτων παγκοσμίως. Από αυτούς τους θανάτους, το 85% οφειλόταν σε καρδιακή προσβολή και εγκεφαλικό επεισόδιο. Πάνω από τα τρία τέταρτα των θανάτων από ΚΝ παρατηρούνται σε χώρες χαμηλού και μεσαίου εισοδήματος. Από τα 17 εκατομμύρια πρόωρους θανάτους (κάτω των 70 ετών) λόγω μη μεταδοτικών ασθενειών το 2019, το 38% προκλήθηκε από ΚΝ.

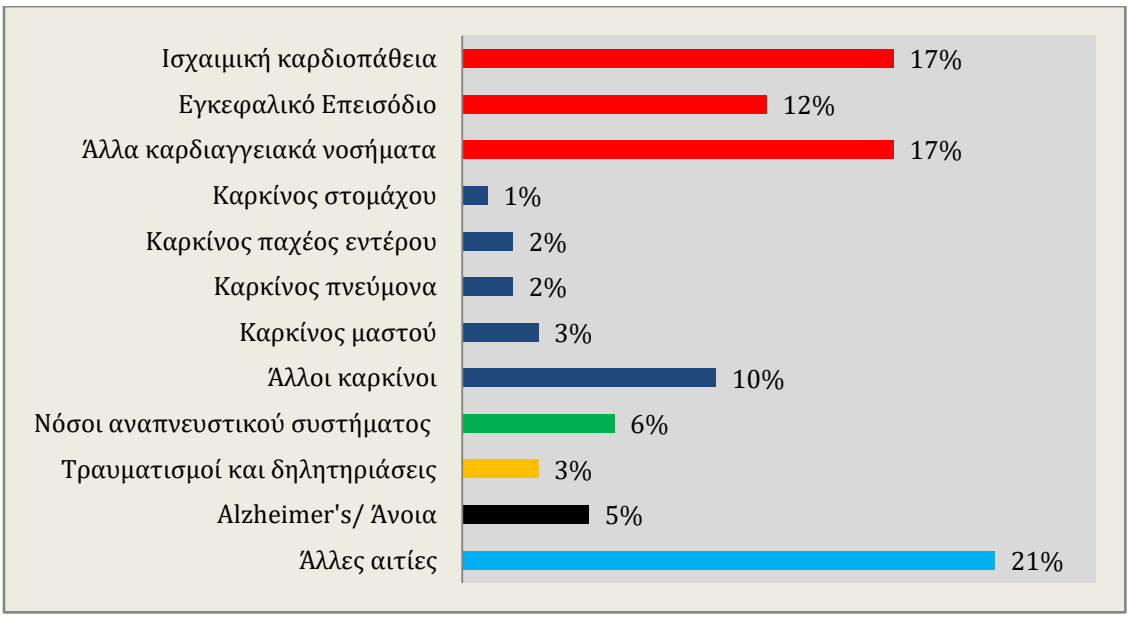
1.2 Στατιστικά στοιχεία στην Ευρώπη

Στην Ευρώπη για το 2019, σύμφωνα με τον ΠΟΥ [6], [7], σε σύνολο 9,7 εκατομμύρια θανάτους, οι παθήσεις της καρδιάς και του κυκλοφορικού συστήματος ήταν επίσης η κύρια αιτία θανάτου και προκάλεσαν περίπου 4,1 εκατομμύρια θανάτους, ή το 42% όλων των θανάτων. Στους άνδρες, τα ΚΝ ευθύνονται για 1,9 εκατομμύρια θανάτους (39% όλων των θανάτων των ανδρών), ενώ στις γυναίκες ευθύνονται για 2,2 εκατομμύρια θανάτους (45% όλων των θανάτων των γυναικών). Συγκριτικά, ο καρκίνος - η επόμενη συχνότερη αιτία θανάτου - ευθύνεται για μόλις 1,1 εκατομμύριο θανάτους (22%) στους άνδρες και μόλις 900.000 θανάτους (18%) στις γυναίκες. Οι συχνότερες μορφές ΚΝ είναι η ισχαιμική καρδιοπάθεια ή στεφανιαία νόσος και το εγκεφαλικό επεισόδιο. (Εικόνα 2,3).

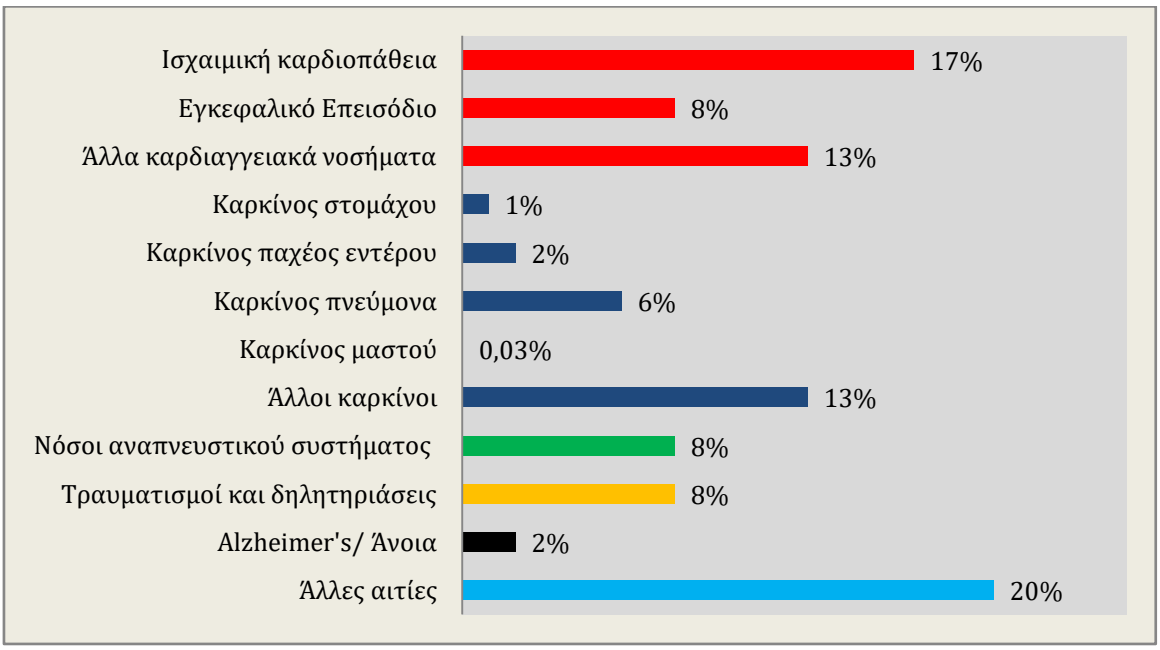
Η στεφανιαία νόσος είναι η πρώτη αιτία θνησιμότητας στην Ευρώπη, υπεύθυνη για 833.000 θανάτους μεταξύ των ανδρών (17% του συνόλου των θανάτων) και 816.000 θανάτους μεταξύ των γυναικών (17%). Το εγκεφαλικό επεισόδιο είναι η δεύτερη συχνότερη αιτία θανάτου με 392.000 θανάτους (8%) στους άνδρες και 576.000 θανάτους (12%) στις γυναίκες. (Εικόνα 2,3)

Η σύγκριση της θνησιμότητας λόγω ΚΝ στις επιμέρους ευρωπαϊκές χώρες αποκαλύπτει σημαντικές διαφορές, με μεγαλύτερη επιβάρυνση συνήθως στις χώρες της Κεντρικής και Ανατολικής Ευρώπης σε σύγκριση με τις χώρες της Βόρειας, Νότιας και Δυτικής Ευρώπης. Αξίζει επίσης να σημειωθεί ότι σε ορισμένες χώρες ο αριθμός των θανάτων από καρκίνο είναι πλέον μεγαλύτερος από τους θανάτους από καρδιαγγειακά νοσήματα [7].

Οι πρόωροι θάνατοι παρουσιάζουν επίσης ενδιαφέρον δεδομένου ότι πολλοί από αυτούς θεωρούνται ότι μπορούν να προληφθούν μέσω της μειωμένης έκθεσης σε συμπεριφορικούς παράγοντες κινδύνου και της έγκαιρης και αποτελεσματικής θεραπείας.



***Εικόνα 2**: Αιτίες θανάτου γυναικών στην Ευρώπη σε σύνολο 4,8 εκατομ. θανάτων (2019)* [7]



Εικόνα 3: *Αιτίες θανάτου ανδρών στην Ευρώπη σε σύνολο 4,9 εκατομ. θανάτων (2019)* [7]

1.3 Στατιστικά στοιχεία στις Ηνωμένες Πολιτείες Αμερικής

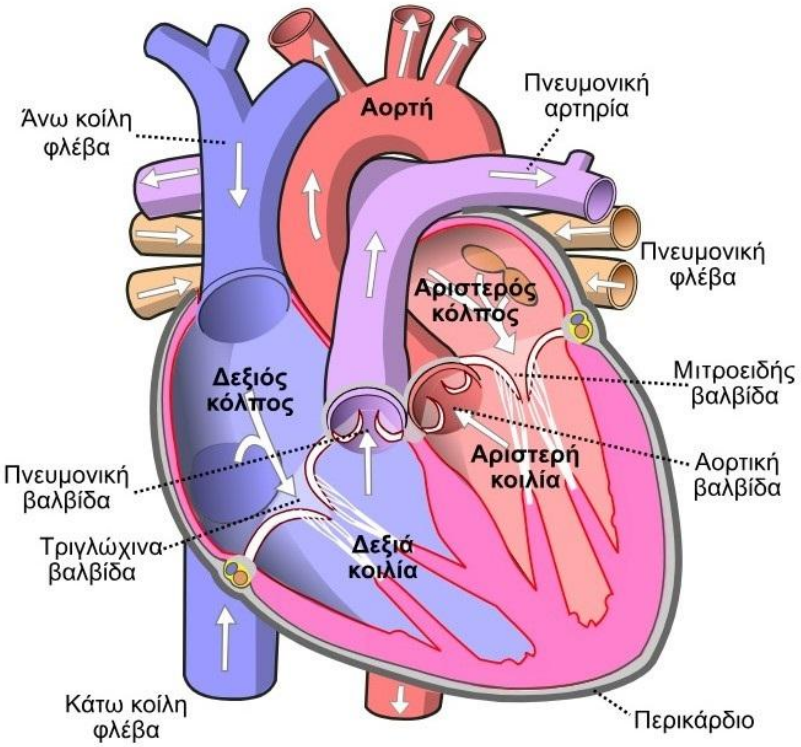
Στις Ηνωμένες Πολιτείες, χώρα προέλευσης των δεδομένων της παρούσης εργασίας, σύμφωνα με τα δεδομένα του αμερικάνικου CDC (Centers for Disease Control and Prevention), τα ΚΝ είναι επίσης η κύρια αιτία θανάτου. Ένα άτομο πεθαίνει από αυτά κάθε 34 δευτερόλεπτα. Περίπου 695.000 άνθρωποι πέθαναν από ΚΝ το 2021 - δηλαδή 1 στους 5 θανάτους [8], [9]. Στους άνδρες ευθύνονται για περίπου 385.000 θανάτους το 2021 (25% όλων των θανάτων των ανδρών), ενώ στις γυναίκες για 311.000 θανάτους (20% όλων των θανάτων των γυναικών) [8].

Τα ΚΝ κόστισαν στις Ηνωμένες Πολιτείες περίπου 239,9 δισεκατομμύρια δολάρια το χρόνο για το 2018 και το 2019 [10]. Αυτό περιλαμβάνει το κόστος των υπηρεσιών υγειονομικής περίθαλψης, των φαρμάκων και της χαμένης παραγωγικότητας λόγω θανάτων.

Η στεφανιαία νόσος, που είναι ο συνηθέστερος τύπος ΚΝ, προκάλεσε 375.476 θανάτους το 2021 [9]. Περίπου 1 στους 20 ενήλικες ηλικίας 20 ετών και άνω πάσχουν από στεφανιαία νόσο (περίπου 5%) [11] και περίπου 2 στους 10 θανάτους συμβαίνουν σε ενήλικες ηλικίας κάτω των 65 ετών [8].

Κάθε χρόνο, περίπου 805.000 άνθρωποι στις Ηνωμένες Πολιτείες παθαίνουν έμφραγμα, δηλαδή ένας κάθε 40 δευτερόλεπτα [9].

Κεφάλαιο 2: Το καρδιαγγειακό σύστημα



Εικόνα 4: *Ανατομία της καρδιάς και πορεία του αίματος μέσω των καρδιακών κοιλοτήτων* [12]

Η καρδιά και το σύνολο των αιμοφόρων αγγείων αποτελούν το καρδιαγγειακό σύστημα, μέσω του οποίου το αίμα κυκλοφορεί αδιάκοπα στον ανθρώπινο οργανισμό. Με την κυκλοφορία του αίματος μεταφέρονται σε όλα τα κύτταρα του σώματος οξυγόνο και άλλα χρήσιμα συστατικά, ενώ από τα κύτταρα μεταφέρεται διοξείδιο του άνθρακα και τοξικά προϊόντα του μεταβολισμού προς τα κατάλληλα όργανα για να αποβληθούν.

Η καρδιά είναι το κεντρικό όργανο της κυκλοφορίας. Είναι ένα κοίλο μυώδες όργανο με σχήμα ανάστροφου κώνου. Ο καρδιακός μυς (μυοκάρδιο) είναι ο μοναδικός γραμμωτός μυς, η λειτουργία του οποίου κατ' εξαίρεση δεν ελέγχεται από τη θέλησή μας, όπως συμβαίνει με τους υπόλοιπους γραμμωτούς μυς. Βρίσκεται μέσα στην θωρακική κοιλότητα ανάμεσα στους δύο πνεύμονες, και εξωτερικά περιβάλλεται από έναν υμένα που ονομάζεται περικάρδιο. Εσωτερικά η καρδιά χωρίζεται σε τέσσερις κοιλότητες: τους δύο κόλπους (αριστερός και δεξιός) στο πάνω μέρος και δύο κοιλίες (αριστερή και δεξιά) στο κάτω μέρος. Επικοινωνία υπάρχει μεταξύ αριστερού κόλπου και αριστερής κοιλίας και μεταξύ δεξιού κόλπου και δεξιάς κοιλίας μέσω της τριγλώχινος και διγλώχινος βαλβίδας αντίστοιχα. Το αίμα διοχετεύεται από τις κοιλίες στην αορτή και στην πνευμονική αρτηρία μέσω της αορτικής και της πνευμονικής βαλβίδας αντίστοιχα (Εικόνα 4) [13].

Οι αρτηρίες που τροφοδοτούν με αίμα το μυοκάρδιο ξεκινούν από την αορτή, αναπτύσσονται στην επιφάνεια της καρδιάς και ονομάζονται στεφανιαίες, διότι την αγκαλιάζουν σαν στεφάνι. Είναι οι πρώτοι κλάδοι της αορτής και είναι δύο: η δεξιά και η αριστερή στεφανιαία αρτηρία .

Η λειτουργία της καρδιάς χαρακτηρίζεται από την περιοδικότητα του καρδιακού παλμού που οφείλεται στα ερεθίσματα που παράγονται από τον φλεβόκομβο, τον φυσικό

βηματοδότη της καρδιάς ο οποίος έχει έναν ενδογενή ρυθμό παραγωγής παλμών. Η συχνότητα των καρδιακών παλμών σε ηρεμία κυμαίνεται σε έναν ενήλικα 70-80 παλμούς ανά λεπτό. Η καρδιά σε κατάσταση ηρεμίας προωθεί περίπου 5.25 λίτρα αίμα ανά λεπτό (καρδιακή παροχή) με εύρος 4-8 λίτρα ανά λεπτό. Σε κατάσταση άσκησης η ποσότητα αυτή μπορεί να φτάσει τα 19.5 λίτρα, δηλαδή 4-5 φορές μεγαλύτερη [13].

Τα αιμοφόρα αγγεία μεταφέρουν το αίμα από την καρδιά προς τα όργανα του σώματος και από τα όργανα πίσω προς την καρδιά. Αποτελούν ένα κλειστό σύστημα “σωλήνων” μέσω των οποίων γίνεται η κυκλοφορία του αίματος. Διακρίνονται σε αρτηρίες, που μεταφέρουν το αίμα από την καρδιά προς τα διάφορα όργανα, φλέβες που μεταφέρουν το αίμα από τα διάφορα όργανα προς την καρδιά και τα τριχοειδή αγγεία που είναι ένα δίκτυο μικροσκοπικών αγγείων μεταξύ αρτηριών και φλεβών, μέσα από τα οποία γίνεται η ανταλλαγή των αερίων και των διαφόρων ουσιών στα όργανα του σώματος.

Το αίμα μεταφέρεται προς τους πνεύμονες για να αποβάλλει το διοξείδιο του άνθρακα και να προμηθευτεί οξυγόνο μέσω της μικρής (ή πνευμονικής) κυκλοφορίας και μεταφέρει το οξυγόνο και τις χρήσιμες ουσίες προς τα όργανα του σώματος μέσω της μεγάλης κυκλοφορίας. Η μικρή κυκλοφορία αρχίζει από τη δεξιά κοιλία της καρδιάς με την πνευμονική αρτηρία, που οδηγεί το αίμα στους πνεύμονες και ειδικά στις κυψελίδες, όπου αποβάλλεται το διοξείδιο του άνθρακα και εμπλουτίζεται με οξυγόνο. Εν συνεχεία, το αίμα μέσω των πνευμονικών φλεβών επιστρέφει στον αριστερό κόλπο της καρδιάς. Αφού περάσει από εκεί στην αριστερή κοιλία, ξεκινά η μεγάλη κυκλοφορία, όπου μέσω της αορτής το αίμα στέλνεται σε όλο το σώμα, για να αποδώσει το οξυγόνο και επιστρέφει μέσω των φλεβών στον δεξιό κόλπο της καρδιάς. Ο δεξιός κόλπος συστέλλεται και στέλνει το αίμα στη δεξιά κοιλία για να επαναληφθεί ο ίδιος κύκλος. [13]

Κεφάλαιο 3: Καρδιαγγειακά νοσήματα

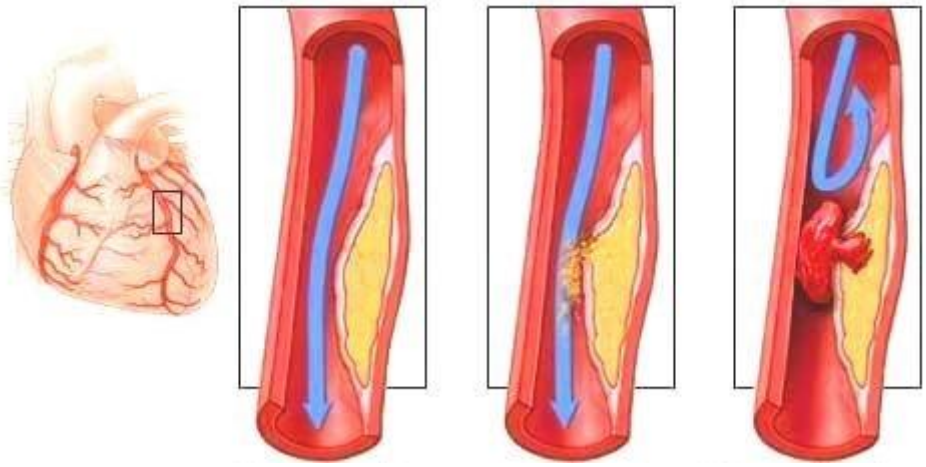
Τα καρδιαγγειακά νοσήματα είναι μια ομάδα παθήσεων της καρδιάς και των αιμοφόρων αγγείων. Σε αυτές περιλαμβάνονται:

3.1 Στεφανιαία νόσος / Καρδιακή Νόσος

3.1.1 Αιτιολογία

Η στεφανιαία νόσος προκαλείται από το σχηματισμό αθηρωματικών πλακών στο τοίχωμα των στεφανιαίων αρτηριών (Εικόνα 5). Οι αθηρωματικές πλάκες σχηματίζονται από εναποθέσεις κυρίως χοληστερίνης και ασβεστίου στον εσωτερικό χιτώνα των αρτηριών.
Έχει παρατηρηθεί ότι ο σχηματισμός αθηρωματικών πλακών ευνοείται από την υπερλιπιδαιμία, την αρτηριακή υπέρταση, το κάπνισμα, τον σακχαρώδη διαβήτη, και από γενετικούς παράγοντες (κληρονομική προδιάθεση). Επομένως η στεφανιαία νόσος έχει πολυπαραγοντική αιτιολογία. Οι περισσότεροι από τους προδιαθεσικούς παράγοντες χαρακτηρίζουν τις σύγχρονες βιομηχανικές κοινωνίες, στις οποίες η υπερκατανάλωση λίπους και άλατος ευνοούν την δυσλιπιδαιμία και την υπέρταση αντίστοιχα ενώ είναι πολύ διαδεδομένο και το κάπνισμα. Η υπερκατανάλωση κορεσμένου λίπους προκύπτει κυρίως από την περίσσεια κόκκινου κρέατος, γαλακτοκομικών προϊόντων και συνθετικών λιπών (trans) που χαρακτηρίζει τις βιομηχανικές κοινωνίες. Παρ’ όλα αυτά υπάρχουν ευρήματα

από μούμιες σε διάφορες περιοχές που δείχνουν ότι η στεφανιαία νόσος υπάρχει από παλιά έως και περισσότερα από 4000 χρόνια πριν [15], όμως η εξάπλωση της στις σύγχρονες βιομηχανικές κοινωνίες δεν έχει προηγούμενο και την έχει καταστήσει πρώτη αιτία θανάτου.



Εικόνα 5: *Ανάπτυξη αθηρωματικής πλάκας στις αρτηρίες ενός ατόμου με καρδιακή νόσο.Στένωση της αρτηρίας (αριστερά), σπάσιμο αθηρωματικής πλάκας (κέντρο) και δημιουργία θρόμβου (δεξιά)* [14] *(Προσαρμογή)*

Η αθηρωματική πλάκα με τη συνεχή εναπόθεση χοληστερίνης σταδιακά μεγεθύνεται και στενεύει τον αυλό της πάσχουσας αρτηρίας. Όταν το εύρος της αρτηρίας μειωθεί αρκετά, περί το 70%, εμφανίζονται συμπτώματα στην προσπάθεια γιατί η παροχή αίματος δεν μπορεί να αυξηθεί ανάλογα με τις ανάγκες του μυοκαρδίου. Η σημαντικότερη συνέπεια είναι ότι, ακόμη (συνηθέστερα) και πριν μεγαλώσει, η αθηρωματική πλάκα υπόκειται σε μικροτραυματισμούς που είναι το έναυσμα για τοπική δημιουργία θρόμβου. Η κατάληξη συνήθως είναι πλήρης ή ατελής απόφραξη της στεφανιαίας αρτηρίας και αυτή είναι η χειρότερη εξέλιξη, γιατί αυτού του είδους η απόφραξη προκαλεί έμφραγμα ή οξύ στεφανιαίο σύνδρομο, καταστάσεις απειλητικές για τη ζωή και την ακεραιότητα της καρδιάς [16].

3.1.2 Συμπτώματα

Το βασικό σύμπτωμα που θα φέρει τον ασθενή στον γιατρό είναι η στηθάγχη, δηλαδή πόνος ή αίσθημα καύσου ή βάρους στο στήθος που αντανακλάται συνήθως στο αριστερό χέρι στην εσωτερική (ωλένιο) πλευρά μέχρι το μικρό δάχτυλο ή/και στη ρίζα του λαιμού και στην άνω γνάθο. Μερικές φορές η κύρια ενόχληση μπορεί να εντοπίζεται στο επιγάστριο (πάνω από το στομάχι). Η στηθάγχη έχει δύο βασικές μορφές την στηθάγχη προσπαθείας και την στηθάγχη ηρεμίας. Μπορεί όμως η πρώτη εκδήλωση της στεφανιαίας νόσου να είναι το οξύ έμφραγμα του μυοκαρδίου ή και ο θάνατος χωρίς πρότερη συμπτωματολογία. Η στηθάγχη προσπαθείας εμφανίζεται κατά την προσπάθεια και την άσκηση όταν η καρδιά δεν μπορεί να ανταποκριθεί επαρκώς στις αυξημένες απαιτήσεις παροχής αίματος. Έχει μικρή διάρκεια, μερικών λεπτών, και υποχωρεί με τη διακοπή της προσπάθειας. Η στηθάγχη ηρεμίας εμφανίζεται κατά την ηρεμία και προκαλείται από απότομη σημαντική στένωση ή παροδική απόφραξη στην αρτηρία. Διαρκεί συνήθως μερικά λεπτά και υποχωρεί για να επανέλθει αργότερα. Είναι πολύ σοβαρό σύμπτωμα διότι η πιθανότητα να εξελιχθεί σε οξύ έμφραγμα του μυοκαρδίου είναι μεγάλη. Στο οξύ έμφραγμα του μυοκαρδίου ο στηθαγχικός

πόνος αποκτά μόνιμο χαρακτήρα, είναι περισσότερο έντονος, η διάρκεια του ξεπερνά τα 20-30 λεπτά και συνήθως διαρκεί 2-3 ώρες ή και περισσότερο. Μπορεί να συνυπάρχουν συνοδά συμπτώματα όπως ιδρώτας και έμετος και ανάλογα με τη σοβαρότητα του εμφράγματος δύσπνοια ή μεγάλη πτώση της αρτηριακής πίεσης [16].

3.1.3 Διάγνωση

Η αρχική διάγνωση της στεφανιαίας νόσου γίνεται από τα χαρακτηριστικά συμπτώματα και πρέπει να ακολουθήσει περαιτέρω διερεύνηση με ειδικές διαγνωστικές εξετάσεις [16]:

- Ηλεκτροκαρδιογράφημα (ΗΚΓ). Μετρά την ηλεκτρική δραστηριότητα, τον ρυθμό και την κανονικότητα του καρδιακού παλμού με ειδική συσκευή που ονομάζεται ηλεκτροκαρδιογράφος.
- Ηχοκαρδιογράφημα/Triplex. Τα ηχοκαρδιογραφικά μηχανήματα χρησιμοποιούν τα υπερηχητικά κύματα (ειδικά ηχητικά κύματα) με διάφορες τεχνικές ώστε να μελετούν τη δομή και τη λειτουργικότητα τόσο του συνόλου όσο και των επιμέρους τμημάτων της καρδιάς.
- Τεστ κοπώσεως. Ο ασθενής υποβάλλεται σε έντονη άσκηση πάνω σε κυλιόμενο διάδρομο, ενώ ταυτόχρονα βρίσκεται σε ηλεκτροκαρδιογραφική παρακολούθηση. Αυτό βοηθά να προσδιοριστεί πόσο καλά λειτουργεί η καρδιά όταν βρίσκεται σε συνθήκες αυξημένων αναγκών, δηλαδή διερευνά τυχόν ύπαρξη δυσαναλογίας μεταξύ απαιτούμενου και προσφερόμενου οξυγόνου στο μυοκάρδιο.
- Ακτινογραφία θώρακος. Απεικονίζει σε δύο διαστάσεις τη θωρακική κοιλότητα, τους πνεύμονες, την καρδιά και τα μεγάλα αγγεία. Από την ακτινογραφία θώρακος αντλούμε πληροφορίες σχετικά με το μέγεθος και το σχήμα της καρδιάς.
- Καρδιακός καθετηριασμός. Ελέγχει το εσωτερικό των αρτηριών για απόφραξη, εισάγοντας έναν λεπτό, εύκαμπτο σωλήνα μέσα από μια αρτηρία στη βουβωνική χώρα, το χέρι ή το λαιμό για να φτάσει στην καρδιά. Οι επαγγελματίες υγείας μπορούν να μετρήσουν την αρτηριακή πίεση εντός της καρδιάς, την καρδιακή παροχή μέσω των κοιλοτήτων της καρδιάς, τον κορεσμό του οξυγόνου, καθώς και να διενεργήσουν στεφανιογραφία.
- Στεφανιογραφία. Πρόκειται για μία ελάχιστα επεμβατική τεχνική απεικόνισης των στεφανιαίων αρτηριών με ακτίνες Χ, μετά από έγχυση ακτινοσκιερού υλικού κατά τη διάρκεια καρδιακού καθετηριασμού.
- Απεικόνιση της καρδιά με μαγνητική τομογραφία (Magnetic resonance imaging – MRI). Στηρίζεται στο φαινόμενο του μαγνητικού συντονισμού. Με τη βοήθεια ηλεκτρονικών υπολογιστών προσδιορίζεται το μαγνητικό σώμα στο χώρο με αποτέλεσμα τη δυνατότητα λήψης τομών σε τρία επίπεδα.
- Αξονική τομογραφία για ανίχνευση ασβεστίου στεφανιαίας αρτηρίας. Η ειδική αυτή αξονική τομογραφία θώρακα, έχει διάρκεια μόλις λίγων δευτερολέπτων και με ελάχιστη ακτινοβολία μετρά τη συγκέντρωση του ασβεστίου στις στεφανιαίες αρτηρίες συμβάλλοντας στον εντοπισμό πιθανής στεφανιαίας νόσου προτού εμφανιστούν συμπτώματα [17].
- Τομογραφία εκπομπής ποζιτρονίων (Positron Emission Tomography - PET). Η σάρωση PET της καρδιάς είναι μια μη επεμβατική εξέταση ακτινοδιάγνωσης της πυρηνικής ιατρικής. Χρησιμοποιεί ραδιενεργούς ιχνηθέτες για να παράγει τρισδιάστατες εικόνες του καρδιακού μυός με σκοπό τη διάγνωση της στεφανιαίας

νόσου και της βλάβης λόγω καρδιακής προσβολής. Οι σαρώσεις PET μπορούν να δείξουν υγιή και κατεστραμμένο καρδιακό μυ και ανάλογα με τα αποτελέσματα βοηθούν στην επιλογή της σωστής θεραπείας [18].

3.1.4 Θεραπεία

Η θεραπευτική αντιμετώπιση γίνεται με δύο τρόπους [16]:

- Φαρμακευτική αγωγή σε περιπτώσεις μέτριας βαρύτητας με τριπλό στόχο, την ανακούφιση από τη στηθάγχη, την αποφυγή οξέων συμβάντων και την αναχαίτιση της στεφανιαίας νόσου.
- Χειρουργική με αγγειοπλαστική των στεφανιαίων αρτηριών (μπαλονάκι) ή αορτοστεφανιαία παράκαμψη (bypass) σε πιο σοβαρές περιπτώσεις. Οι επεμβάσεις αυτές αντιμετωπίζουν την απόφραξη, αλλά δεν θεραπεύουν την αιτία της δημιουργίας των στενώσεων. Για το λόγο αυτό, για να μην δημιουργηθούν νέες στενώσεις ο ασθενής θα πρέπει να ενημερώνεται για τους παράγοντες κινδύνου και να προσαρμόζει ανάλογα τον τρόπο ζωής του.

Πρόσφατα η συνεργασία μεταξύ διαφορετικών επιστημονικών κλάδων όπως η βιοτεχνολογία και η μηχανική των ιστών (tissue engineering) έχει οδηγήσει στην ανάπτυξη νέων θεραπευτικών στρατηγικών όπως τα βλαστοκύτταρα (stem cells), η νανοτεχνολογία, η ρομποτική χειρουργική και άλλες εξελίξεις (τρισδιάστατη εκτύπωση καρδιάς και νεότερα φάρμακα). Αυτές οι μέθοδοι θεραπείας υπόσχονται καλύτερα αποτελέσματα στη διαχείριση της στεφανιαίας νόσου [19].

3.2 Λοιπές παθήσεις

3.2.1 Παθήσεις αιμοφόρων αγγείων του εγκεφάλου

Ο εγκέφαλός για να λειτουργεί σωστά, χρειάζεται οξυγόνο. Οι αρτηρίες παρέχουν αίμα πλούσιο σε οξυγόνο σε όλα τα μέρη του εγκεφάλου. Εάν συμβεί κάτι που εμποδίζει τη ροή του αίματος, τα εγκεφαλικά κύτταρα αρχίζουν να πεθαίνουν μέσα σε λίγα λεπτά. Αυτό προκαλεί εγκεφαλικό επεισόδιο (stroke), το οποίο μπορεί να οδηγήσει σε μόνιμη εγκεφαλική βλάβη, μακροχρόνια αναπηρία ή ακόμα και στο θάνατο. Υπάρχουν δύο τρόποι μείωσης της αιμάτωσης του εγκεφάλου είτε απόφραξη κάποιας αρτηρίας από θρόμβο αίματος (ισχαιμικό εγκεφαλικό επεισόδιο), είτε ρήξη κάποιου αγγείου του εγκεφάλου (αιμορραγικό εγκεφαλικό επεισόδιο). Συχνότερα είναι τα ισχαιμικά εγκεφαλικά επεισόδια, που αποτελούν το 85% όλων των περιστατικών. Το εγκεφαλικό είναι μια σοβαρή ιατρική κατάσταση που απαιτεί επείγουσα φροντίδα και θεραπεία ανάλογα με την αιτία που το προκάλεσε [20], [21].

3.2.2 Παθήσεις των περιφερικών αρτηριών

Οι περιφερικές αρτηρίες αιματώνουν το υπόλοιπο σώμα. Η ανάπτυξη της αθηρωματικής πλάκας, λόγω συσσώρευσης λιπιδίων και ασβεστίου στο αρτηριακό τοίχωμα προκαλεί στένωση ή απόφραξη των περιφερικών αρτηριών, με αποτέλεσμα μείωση της ροής του αίματος και κατά συνέπεια του οξυγόνου και των θρεπτικών ουσιών που χρειάζονται τα όργανα. Είναι πιο συχνό στις αρτηρίες που τροφοδοτούν τα κάτω άκρα με αποτέλεσμα κυρίως εκδήλωση πόνου και δυσκολία στη βάδιση, που βελτιώνεται μετά την ανάπαυση [22].

3.2.3 Ρευματική νόσος της καρδιάς

Είναι η βλάβη του καρδιακού μυός και των καρδιακών βαλβίδων από ρευματικό πυρετό, που προκαλείται από στρεπτοκοκκική λοίμωξη [5].

3.2.4 Συγγενείς καρδιοπάθειες

Οφείλονται σε βλάβες της καρδιάς εκ γενετής εξαιτίας δυσπλασιών, που επηρεάζουν την φυσιολογική ανάπτυξη και τη λειτουργία της [5].

3.2.5 Θρόμβωση των εν τω βάθει φλεβών και πνευμονική εμβολή

Η εν τω βάθει φλεβική θρόμβωση είναι μια κατάσταση κατά την οποία δημιουργείται θρόμβος αίματος στις εν τω βάθει φλέβες, συνήθως των κάτω άκρων. Αυτή οφείλεται σε φλεβική στάση, βλάβη του αγγειακού τοιχώματος ή αύξηση της πηκτικότητας του αίματος. Όταν ένα τμήμα του θρόμβου αποκολληθεί μπορεί να μεταφερθεί στην καρδιά και από εκεί στον πνεύμονα και να προκαλέσει πνευμονική εμβολή, δηλαδή απόφραξη αρτηριών στον πνεύμονα, κατάσταση που μπορεί να είναι απειλητική για τη ζωή [5].

Κεφάλαιο 4: Παράγοντες κινδύνου στεφανιαίας νόσου και πρόληψη

Οι κυριότεροι παράγοντες κινδύνου για στεφανιαία νόσο είναι η αυξημένη αρτηριακή πίεση, η αυξημένη γλυκόζη αίματος, η αυξημένη χοληστερόλη αίματος , το αυξημένο βάρος, η παχυσαρκία, η κληρονομικότητα, η ηλικία, το φύλο, η φυλή, η ψυχική υγεία και η χρόνια νεφρική νόσος. Ορισμένοι συμπεριφορικοί παράγοντες, όπως η ανθυγιεινή διατροφή, η έλλειψη σωματικής άσκησης, η υπερκατανάλωση αλκοόλ, το κάπνισμα και η διάρκεια του ύπνου μπορούν να προκαλέσουν κάποιους από τους ανωτέρω παράγοντες κινδύνου [23], [24]. Πολλά από τα προαναφερθέντα μπορούν να επηρεαστούν με τη λήψη κατάλληλων προληπτικών μέτρων.

4.1 Παράγοντες κινδύνου στεφανιαίας νόσου

4.1.1 Υψηλή αρτηριακή πίεση (υπέρταση)

Η υψηλή αρτηριακή πίεση είναι ένας σημαντικός παράγοντας κινδύνου για τις καρδιακές παθήσεις. Είναι μία χρόνια πάθηση κατά την οποία η πίεση του αίματος στις αρτηρίες είναι υψηλότερη του φυσιολογικού. Ο έλεγχος της αρτηριακής πίεσης περιλαμβάνει δύο μετρήσεις, τη συστολική και τη διαστολική, που εξαρτώνται από το εάν ο καρδιακός μυς συστέλλεται (συστολή) ή χαλαρώνει μεταξύ των παλμών (διαστολή). Η φυσιολογική αρτηριακή πίεση σε κατάσταση ηρεμίας κυμαίνεται από 100 έως 140 mmHg η συστολική (ανώτατη μέτρηση) και από 60 έως 90 mmHg η διαστολική (κατώτατη μέτρηση). Θεωρείται υψηλή η αρτηριακή πίεση εάν είναι μονίμως σε επίπεδα 140/90 mmHg ή παραπάνω και εάν δεν ελέγχεται, μπορεί να επηρεάσει την καρδιά και άλλα κύρια όργανα του σώματός,

συμπεριλαμβανομένων των νεφρών και του εγκεφάλου. Ονομάζεται συχνά «σιωπηλός δολοφόνος» επειδή συνήθως δεν έχει συμπτώματα, γι' αυτό ο μόνος τρόπος εντοπισμού της γίνεται με τη μέτρησή της. Υπάρχει ισχυρή σχέση μεταξύ της υπέρτασης και της στεφανιαίας νόσου. Η υψηλότερη πίεση στα τοιχώματα των αγγείων μπορεί να τα καταστρέψει, καθιστώντας ευκολότερη τη δημιουργία αθηρωματικής πλάκας, προκαλώντας επικίνδυνη μείωση της αιμάτωσης του μυοκαρδίου. Επίσης αναγκάζει την καρδιά να αντλεί πιο δυνατά για να κυκλοφορήσει επαρκώς το αίμα και αυτή η υπερκόπωση μπορεί να οδηγήσει σε καρδιακή ανεπάρκεια [25].

Οι αλλαγές στον τρόπο ζωής, όπως η μείωση πρόσληψης αλατιού με την τροφή, ή η φαρμακευτική αγωγή μπορεί να οδηγήσει στη μείωση της υπέρτασης, επομένως και του κινδύνου καρδιακών παθήσεων και καρδιακής προσβολής.

4.1.2 Υψηλά επίπεδα χοληστερόλης στο αίμα

Η χοληστερόλη είναι ένα οργανικό μόριο που ανήκει σε μια κατηγορία ουσιών που ονομάζονται λιπίδια. Βιοσυντίθεται από όλα τα ζωικά κύτταρα και αποτελεί βασικό δομικό συστατικό των ζωικών κυτταρικών μεμβρανών. Όταν απομονώνεται χημικά, είναι ένα κιτρινωπό κρυσταλλικό στερεό σώμα. Η χοληστερόλη αποτελεί πρόδρομη ουσία των στεροειδών ορμονών, της βιταμίνης D και των χολικών οξέων, που βοηθούν στη διαδικασία της πέψης και βοηθούν στην απορρόφηση του λίπους της διατροφής. Επειδή είναι μια υδρόφοβη ένωση δεν μπορεί να μεταφερθεί από μόνη της στην κυκλοφορία του αίματος και ως εκ τούτου ενώνεται με πρωτεΐνες, γνωστές ως λιποπρωτεΐνες. Μεταξύ των λιποπρωτεϊνών, οι λιποπρωτεΐνες υψηλής πυκνότητας (HDL) ή «καλή χοληστερόλη», οι λιποπρωτεΐνες χαμηλής πυκνότητας (LDL) ή «κακή χοληστερόλη» και η λιποπρωτεΐνη πολύ χαμηλής πυκνότητας (VLDL) περιέχουν τις υψηλότερες ποσότητες κυκλοφορούσας χοληστερόλης στο σώμα. Πλήθος μελετών έχουν δείξει ότι υψηλά επίπεδα ολικής χοληστερόλης (πάνω από 200 mg/dL) και LDL χοληστερόλης (πάνω από 130mg/dL) συσχετίζονται με αυξημένο κίνδυνο εμφάνισης καρδιαγγειακών νοσημάτων (CVD) [26]. Αντίθετα η χοληστερόλη HDL (πάνω από 40 mg/dL) θεωρείται ότι παρέχει κάποια προστασία έναντι των καρδιακών παθήσεων [27]. Οι κύριες διατροφικές πηγές χοληστερόλης περιλαμβάνουν το κόκκινο κρέας , τους κρόκους και τα ολόκληρα αυγά, το συκώτι, τα νεφρά, τα εντόσθια, το ιχθυέλαιο και το βούτυρο. Εάν λαμβάνουμε περισσότερη χοληστερόλη από αυτή που μπορεί να χρησιμοποιήσει το σώμα, η επιπλέον χοληστερόλη μπορεί να συσσωρευτεί στα τοιχώματα των αρτηριών, συμπεριλαμβανομένων αυτών της καρδιάς. Αυτό οδηγεί σε στένωση των αρτηριών και μπορεί να μειώσει τη ροή του αίματος στην καρδιά, τον εγκέφαλο, τα νεφρά και άλλα μέρη του σώματος. Η υψηλή χοληστερόλη στο αίμα συνήθως δεν έχει σημεία ή συμπτώματα γι' αυτό πρέπει να ελέγχεται προληπτικά το επίπεδό της στο αίμα. Διατήρηση της χοληστερόλης σε φυσιολογικά επίπεδα επιτυγχάνεται επιλέγοντας υγιεινό τρόπο ζωής, όπως υγιεινή διατροφή με λιγότερα κορεσμένα λιπαρά, σωματική άσκηση, μείωση σωματικού βάρους και διακοπή του καπνίσματος. Σε περίπτωση υψηλών επιπέδων LDL χοληστερόλης και τριγλυκεριδίων, συνιστάται φαρμακευτική αγωγή μετά από συμβουλή ιατρού.

4.1.3 Σακχαρώδης διαβήτης

Η γλυκόζη είναι η ουσία που αποτελεί την πηγή ενέργειας για τον ανθρώπινο οργανισμό. Η ινσουλίνη είναι μία ορμόνη, η οποία παράγεται από το πάγκρεας και παίζει ρόλο στη ρύθμιση της γλυκόζης του αίματος, βοηθώντας την είσοδο της στα κύτταρα. Ο σακχαρώδης

διαβήτης προκαλείται όταν ο οργανισμός δεν μπορεί να παράγει επαρκή ποσότητα ινσουλίνης ή δεν μπορεί να τη χρησιμοποιήσει όπως πρέπει. Οι κυριότερες κλινικές εκδηλώσεις περιλαμβάνουν πολυδιψία, πολυουρία, αδυναμία και ανεξήγητη απώλεια βάρους. Υπάρχουν δύο τύποι σακχαρώδους διαβήτη. Ο σακχαρώδης διαβήτης τύπου Ι ή νεανικός διαβήτης, στον οποίο το πάγκρεας δεν παράγει καθόλου ινσουλίνη και ο σακχαρώδης διαβήτης τύπου ΙΙ, στον οποίο το πάγκρεας παράγει μεν ινσουλίνη, η οποία όμως είτε δεν επαρκεί, είτε δεν μπορεί να χρησιμοποιηθεί αποτελεσματικά από τον οργανισμό. Ο σακχαρώδης διαβήτης τύπου ΙΙ, μπορεί να προληφθεί με φυσική δραστηριότητα, υγιεινή διατροφή και καταπολέμηση της παχυσαρκίας. Ο διαβήτης προκαλεί τη συσσώρευση σακχάρου στο αίμα. Για τους υγιείς ενήλικες, οι τιμές του σακχάρου νηστείας θα πρέπει να κυμαίνονται από 72-100mg/dL και οι τιμές του μεταγευματικού σακχάρου (περίπου δύο ώρες μετά το γεύμα) δεν θα πρέπει να ξεπερνούν τα 140mg/dL. Πληθυσμιακές μελέτες έχουν δείξει ότι τα άτομα που πάσχουν από σακχαρώδη διαβήτη διατρέχουν τριπλάσιο κίνδυνο να προσβληθούν από ισχαιμική καρδιοπάθεια, σε σχέση με τους υγιείς. Τα άτομα αυτά έχουν υψηλότερο κίνδυνο πρόωρης, ταχέως εξελισσόμενης στεφανιαίας νόσου. Αυτό σημαίνει ότι σε σύγκριση με εκείνους που δεν έχουν διαβήτη, τα τοιχώματα των αρτηριών τους έχουν περισσότερες εναποθέσεις λίπους και αρχίζουν να σκληραίνουν νωρίτερα και χωρίς προειδοποιητικά συμπτώματα, καθιστώντας τη θεραπεία δύσκολη και προκαλώντας ταχύτερη εξέλιξη της νόσου. Επιπλέον τα άτομα με διαβήτη έχουν αυξημένο κίνδυνο επαναλαμβανόμενων καρδιακών προσβολών και ουλών στον καρδιακό μυ, γεγονός που αυξάνει τον κίνδυνο αιφνίδιου καρδιακού θανάτου [28], [4]. Ειδική περίπτωση είναι ο διαβήτης κατά τη διάρκεια της κύησης (Gestational Diabetes) ο οποίος προκαλεί προβλήματα στο έμβρυο, αλλά συνήθως υποχωρεί μετά τον τοκετό. Παρ' ολ' αυτά, ακόμα κι αν εξαφανιστεί μετά τη γέννηση του μωρού, οι μισές από τις γυναίκες που είχαν διαβήτη κύησης αναπτύσσουν διαβήτη τύπου ΙΙ αργότερα, γι' αυτό πρέπει να ελέγχονται [29].

4.1.4 Παχυσαρκία

Η παχυσαρκία είναι το υπερβολικό σωματικό λίπος και συνδέεται με υψηλότερα επίπεδα «κακής» χοληστερόλης και τριγλυκεριδίων αλλά και με χαμηλότερα επίπεδα «καλής» χοληστερόλης. Μπορεί να οδηγήσει σε υψηλή αρτηριακή πίεση, διαβήτη καθώς και καρδιακές παθήσεις [30], [31]. Συνήθως εκτιμάται με τον Δείκτη Μάζας Σώματος (ΔΜΣ) / Body Mass Index (BMI). Ο ΔΜΣ είναι μία γενική ιατρική ένδειξη για τον υπολογισμό του βαθμού παχυσαρκίας ενός ατόμου. Λόγω του εύκολου υπολογισμού του είναι ένα ευρέως διαδεδομένο διαγνωστικό εργαλείο των πιθανών προβλημάτων υγείας ενός ατόμου σε σχέση με το βάρος του. Δημιουργήθηκε το 1832 από τον Adolphe Quetelet [32] και γι'αυτό ονομάζεται και Quetelet index. Στο Διεθνές Σύστημα μονάδων (SI) υπολογίζεται πολύ εύκολα από τον τύπο:

$$\text{ΔΜΣ} = \frac{\text{Βάρος (kg)}}{\text{Ύψος}^2\ (\text{m}^2)} \quad (1)$$

Παγκοσμίως έχει γίνει αποδεκτή η εξής κατηγοριοποίηση (Πίνακας 1):

Πίνακας 1: Βασικές κατηγορίες BMI

Κατηγορία	BMI (kg/m^2)
Ελλιποβαρής	<18.5
Φυσιολογικό βάρος	18.5 – 24.9
Υπέρβαρος	25 – 29.9
Παχυσαρκία	≥ 30

4.1.5 Αλλεργικό Άσθμα

Το άσθμα είναι μια κοινή χρόνια φλεγμονώδης πάθηση του αναπνευστικού συστήματος που χαρακτηρίζεται από μεταβλητή στένωση των αεραγωγών οδών και βρογχόσπασμο. Στα συνήθη συμπτώματα περιλαμβάνονται: συριγμός, βήχας, αίσθημα σύσφιξης στο θώρακα και δύσπνοια. Με την πάροδο του χρόνου, μπορεί επίσης να έχει αντίκτυπο στην καρδιά επειδή προκαλεί φλεγμονή στους αεραγωγούς, η οποία δεν περιορίζεται μόνο στους πνεύμονες αλλά οι δείκτες φλεγμονής είναι αυξημένοι συστηματικά. Έτσι το σώμα βρίσκεται σε μια χρόνια φλεγμονώδη κατάσταση και η φλεγμονή είναι γνωστό ότι επιδεινώνει τη καρδιαγγειακή νόσο. Συνεπώς το αλλεργικό άσθμα αποτελεί σημαντικό παράγοντα κινδύνου για καρδιαγγειακή νόσο. Εάν δε συνυπάρχει με αυξημένη χοληστερόλη μεγαλώνει το κίνδυνο βλάβης του τοιχώματος των αρτηριών [33].

4.1.6 Οικογενειακό ιστορικό

Όταν τα μέλη μιας οικογένειας μεταφέρουν χαρακτηριστικά από τη μια γενιά στην άλλη μέσω γονιδίων, αυτή η διαδικασία ονομάζεται κληρονομικότητα. Οι γενετικοί παράγοντες πιθανότατα παίζουν κάποιο ρόλο στην υψηλή αρτηριακή πίεση [34], κάποιες καρδιακές παθήσεις και άλλες σχετικές παθήσεις όπως υπερχοληστερολαιμία και σακχαρώδης διαβήτης [35], [36]. Ωστόσο, είναι επίσης πιθανό τα άτομα με οικογενειακό ιστορικό καρδιακών παθήσεων να έχουν κοινό περιβάλλον και άλλες κοινές συνήθειες που μπορεί να αυξήσουν τον κίνδυνο. Ο κίνδυνος για καρδιακή νόσο μπορεί να αυξηθεί ακόμη περισσότερο όταν η κληρονομικότητα συνδυάζεται με λανθασμένες επιλογές τρόπου ζωής, όπως το κάπνισμα και η ανθυγιεινή διατροφή.

4.1.7 Ηλικία και φύλο

Οι καρδιακές παθήσεις είναι ο νούμερο ένα δολοφόνος τόσο των ανδρών όσο και των γυναικών [37]. Η καρδιακή νόσος μπορεί να συμβεί σε οποιαδήποτε ηλικία, αλλά ο κίνδυνος αυξάνεται όσο αυξάνεται η ηλικία. [38], [39].

4.1.8 Φυλή και εθνικότητα

Οι θάνατοι από καρδιακές παθήσεις διαφέρουν ανάλογα με τη φυλή. Στις Ηνωμένες Πολιτείες είναι η κύρια αιτία θανάτου για τα άτομα των περισσότερων φυλετικών ομάδων, συμπεριλαμβανομένων των Αφροαμερικανών, των Ινδιάνων της Αμερικής, των ιθαγενών της Αλάσκας, των ισπανόφωνων και των λευκών ανδρών. Παρακάτω παρουσιάζονται τα ποσοστά όλων των θανάτων που προκλήθηκαν από καρδιακές παθήσεις το 2021,

καταγεγραμμένα ανά φυλετική ομάδα των Η.Π.Α, όπου υπάρχουν αντίστοιχα στοιχεία [8] (Πίνακας 2).

Πίνακας 2: *Ποσοστά θανάτων που προκλήθηκαν από καρδιακές παθήσεις το 2021, ανά φυλετική ομάδα στις ΗΠΑ.*

Φυλετικές ομάδες	**Ποσοστό των θανάτων**
Αμερικανοί ινδιάνοι ή ιθαγενείς της Αλάσκας	15.5
Ασιάτες	18.6
Μαύροι (μη ισπανόφωνοι)	22.6
Ιθαγενείς της Χαβάης ή άλλοι νησιώτες του Ειρηνικού	18.3
Λευκοί (Μη ισπανόφωνοι)	18.0
Ισπανόφωνοι	11.9
Συνολικά	17.4

4.1.9 Ψυχική Υγεία

Οι διαταραχές ψυχικής υγείας που σχετίζονται συχνότερα με καρδιακές παθήσεις ή σχετικούς παράγοντες κινδύνου περιλαμβάνουν:

- **Διαταραχές της διάθεσης**, όπως η μείζονα κατάθλιψη ή η διπολική διαταραχή
- **Αγχώδεις Διαταραχές**, όπως γενικευμένο άγχος, κοινωνικό άγχος, διαταραχές πανικού και φοβίες.
- **Διαταραχή Μετατραυματικού Στρες,** που μπορεί να εμφανιστεί μετά από μια τραυματική εμπειρία ζωής, όπως πόλεμος, φυσική καταστροφή ή οποιοδήποτε άλλο σοβαρό περιστατικό.
- **Χρόνιο Στρες,** δηλαδή το στρες που είναι σταθερό και επιμένει για μεγάλο χρονικό διάστημα.

Ένας μεγάλος αριθμός ερευνών δείχνει ότι η ψυχική υγεία σχετίζεται με παράγοντες κινδύνου για καρδιακή νόσο. Αυτές οι επιπτώσεις μπορεί να προκύψουν τόσο άμεσα, μέσω βιολογικών οδών, όσο και έμμεσα, μέσω επικίνδυνων συμπεριφορών για την υγεία [40].

Τα άτομα που βιώνουν κατάθλιψη, άγχος, ακόμη και μετατραυματικό στρες για μεγάλο χρονικό διάστημα μπορεί να παρουσιάσουν ορισμένες διαταραχές στον οργανισμό, όπως αυξημένη καρδιακή αντίδραση (π.χ. αυξημένος καρδιακός ρυθμός και αρτηριακή πίεση), μειωμένη ροή αίματος στην καρδιά και αυξημένα επίπεδα κορτιζόλης. Με την πάροδο του χρόνου, αυτές μπορεί να οδηγήσουν σε εναπόθεση ασβεστίου στις αρτηρίες, μεταβολικές ασθένειες και καρδιακές παθήσεις [41], [42].

Σημειώνεται επίσης η επίδραση των φαρμάκων που χρησιμοποιούνται για τη θεραπεία διαταραχών ψυχικής υγείας στον κίνδυνο καρδιομεταβολικής νόσου. Η χρήση ορισμένων αντιψυχωσικών φαρμάκων έχει συσχετιστεί με παχυσαρκία, αντίσταση στην ινσουλίνη, διαβήτη, καρδιακές προσβολές, κολπική μαρμαρυγή, εγκεφαλικό επεισόδιο και θάνατο [43].

Διαταραχές ψυχικής υγείας όπως το άγχος και η κατάθλιψη μπορεί επίσης να αυξήσουν την πιθανότητα υιοθέτησης συμπεριφορών όπως το κάπνισμα, ο ανενεργός τρόπος ζωής ή η αποτυχία λήψης κατάλληλων φαρμάκων. Αυτό συμβαίνει επειδή τα άτομα που αντιμετωπίζουν μια διαταραχή ψυχικής υγείας μπορεί να έχουν μειωμένη ικανότητα υγιούς αντιμετώπισης στρεσογόνων καταστάσεων, καθιστώντας δύσκολο το να κάνουν επιλογές υγιεινού τρόπου ζωής για να μειώσουν τον κίνδυνο καρδιακής νόσου [40].

4.1.10 Προβλήματα νεφρικής λειτουργίας – Χρόνια νεφρική νόσος

Ως Χρόνια Νεφρική Νόσος (ΧΝΝ) ή Χρόνια Νεφρική Ανεπάρκεια χαρακτηρίζεται η μη αναστρέψιμη δυσλειτουργία-ανεπάρκεια των νεφρών, που προκαλείται από βλάβη των νεφρών ποικίλης αιτιολογίας. Η ΧΝΝ είναι ένας σημαντικός παράγοντας κινδύνου για εμφάνιση στεφανιαίας νόσου, η οποία αποτελεί την κύρια αιτία νοσηρότητας και θανάτου των ασθενών που πάσχουν από αυτήν. Οι πάσχοντες παρουσιάζουν αύξηση της ουρίας του αίματος (ουραιμία), η οποία προκαλεί υπέρταση, διαβήτη, φλεγμονή, οξειδωτικό στρες και μη φυσιολογικό μεταβολισμό ασβεστίου-φωσφόρου, τα οποία αποτελούν παράγοντες κινδύνου εμφάνισης ΣΝ. Η ΧΝΝ και η νεφρική νόσος τελικού σταδίου τροποποιούν επίσης την κλινική εικόνα και τα κύρια συμπτώματα της ΣΝ. Η διαχείριση της ΣΝ σε αυτούς τους ασθενείς είναι περίπλοκη, λόγω της πιθανότητας εμφάνισης και συννοσηρών καταστάσεων και πιθανών παρενεργειών κατά τη διάρκεια της θεραπείας [44].

4.2 Παράγοντες τρόπου ζωής που αυξάνουν τον κίνδυνο καρδιακών παθήσεων.

4.2.1 Ανθυγιεινή διατροφή

Η κατανάλωση τροφών με υψηλή περιεκτικότητα σε κορεσμένα λιπαρά και χοληστερόλη έχει συνδεθεί με καρδιακές παθήσεις και σχετικές παθήσεις, όπως η αθηροσκλήρωση. Επίσης η ποσότητα του αλατιού (χλωριούχου νατρίου) που καταναλώνεται είναι καθοριστικός παράγοντας για τα επίπεδα της αρτηριακής πίεσης και της υπέρτασης και του συνολικού καρδιαγγειακού κινδύνου. Πρόσληψη αλατιού μικρότερη από 5 γραμμάρια (περίπου 2 γραμμάρια νατρίου) ανά άτομο την ημέρα συνιστάται από τον ΠΟΥ για την πρόληψη των καρδιαγγειακών παθήσεων [45]. Ωστόσο, δεδομένα από διάφορες χώρες δείχνουν ότι οι περισσότεροι πληθυσμοί καταναλώνουν πολύ περισσότερο αλάτι από το συνιστώμενο. Η υπερβολική πρόσληψη αλατιού εκτιμάται ότι προκαλεί περίπου 5 εκατομμύρια θανάτους παγκοσμίως κάθε χρόνο [46].

4.2.2 Έλλειψη σωματικής δραστηριότητας

Η έλλειψη σωματικής δραστηριότητας μπορεί να οδηγήσει σε καρδιακές παθήσεις. Μπορεί επίσης να αυξήσει τις πιθανότητες εμφάνισης άλλων ιατρικών παθήσεων που αποτελούν παράγοντες κινδύνου, όπως η παχυσαρκία, η υψηλή αρτηριακή πίεση, η υψηλή χοληστερόλη και ο διαβήτης. Η τακτική σωματική δραστηριότητα μπορεί να μειώσει τον κίνδυνο για καρδιακές παθήσεις [47].

4.2.3 Υπερβολική κατανάλωση αλκοόλ

Η αρτηριακή πίεση αυξάνεται με την τακτική κατανάλωση αλκοόλ με δοσοεξαρτώμενο τρόπο. Οι γυναίκες δεν πρέπει να πίνουν περισσότερο από 1 ποτό την ημέρα (7-14 g καθαρής αιθανόλης). Οι άνδρες δεν πρέπει να πίνουν περισσότερα από 2 ποτά την ημέρα (14–28 g αιθανόλης). Σημαντική μείωση στις μετρήσεις της αρτηριακής πίεσης παρατηρείται μετά από μόλις 1 μήνα αποχής από το αλκοόλ. Η χρόνια πρόσληψη αλκοόλ σε μεγάλη ποσότητα, πάνω από 14 g αλκοόλ την ημέρα, σχετίζεται και με καρδιακές αρρυθμίες, με συχνότερη την κολπική μαρμαρυγή [48].

4.2.4 Κάπνισμα

Το κάπνισμα μπορεί να βλάψει την καρδιά και τα αιμοφόρα αγγεία, γεγονός που αυξάνει τον κίνδυνο για καρδιακές παθήσεις όπως η αθηροσκλήρωση και η καρδιακή προσβολή μέσω πολλών μηχανισμών [49], [50]: Αυξάνει τα επίπεδα χοληστερόλης, με αποτέλεσμα στένωση των αγγείων, και ειδικά των στεφανιαίων αγγείων, οδηγώντας στην εκδήλωση στεφανιαίας νόσου 2 έως 4 φορές περισσότερο από τους μη καπνιστές [51], [52]. Η νικοτίνη αυξάνει την αρτηριακή πίεση. Το μονοξείδιο του άνθρακα από τον καπνό του τσιγάρου μειώνει την ποσότητα οξυγόνου που μπορεί να μεταφέρει το αίμα. Η έκθεση στο παθητικό κάπνισμα έχει συνδεθεί με σημαντική αύξηση στον κίνδυνο εμφάνισης στεφανιαίας νόσου [53]. Το κάπνισμα είναι μία από τις σημαντικότερες αιτίες πρόωρου θανάτου παγκοσμίως, προκαλώντας περισσότερους από 8 εκατομμύρια θανάτους κάθε χρόνο σε όλο τον κόσμο σύμφωνα με τον ΠΟΥ, ενώ η χρήση του είναι εφικτό να περιοριστεί με τη λήψη κατάλληλων μέτρων [54].

4.2.5 ΄Υπνος

Τόσο η διάρκεια όσο και η ποιότητα του ύπνου έχουν συσχετιστεί με τον κίνδυνο εμφάνισης στεφανιαίας νόσου. Συγκεκριμένα, διάρκεια νυχτερινού ύπνου λιγότερο από 7 ώρες ή περισσότερο από 8 ώρες έδειξε αυξημένο κίνδυνο ΣΝ κατά 13% σε σύγκριση με τον ύπνο αναφοράς (7-8 ώρες). Επιπλέον, τα άτομα με κακή ποιότητα ύπνου διέτρεχαν μεγαλύτερο κίνδυνο για ΣΝ από εκείνα με ύπνο καλής ποιότητας [55].

4.3 Πρόληψη

Οι παράγοντες που αυξάνουν τον κίνδυνο για στεφανιαία νόσο και εγκεφαλικό επεισόδιο διακρίνονται σε αυτούς που μπορούν να τροποποιηθούν, όπως λιπιδικές διαταραχές, υπέρταση, κάπνισμα και σε αυτούς που δεν μπορούν, όπως ηλικία, φύλο και οικογενειακό ιστορικό πρώιμης στεφανιαίας νόσου. Εντυπωσιακές μειώσεις στα ποσοστά θνησιμότητας λόγω καρδιακών παθήσεων και εγκεφαλικών επεισοδίων έχουν επιτευχθεί σε όλες τις ηλικιακές ομάδες στη Βόρεια Αμερική από το 1980 έως το 2015 [56], σε μεγάλο βαθμό μέσω της βελτίωσης των τροποποιήσιμων παραγόντων κινδύνου: μείωση στο κάπνισμα, βελτίωση στα επίπεδα λιπιδίων και συστηματική ανίχνευση και θεραπεία της υπέρτασης. Ο ρόλος του προσυμπτωματικού ελέγχου για καρδιαγγειακό κίνδυνο και η χρήση αποτελεσματικών θεραπειών για τη μείωσή του είναι ζωτικής σημασίας [57]. Παρακάτω αναφέρονται τα κυριότερα προληπτικά μέτρα που μπορούν να ληφθούν προς αυτήν την κατεύθυνση, τα οποία προτείνει η Ειδική Ομάδα Προληπτικών Υπηρεσιών των ΗΠΑ (U.S. Preventive Services Task Force – USPSTF) [58] :

- **Πίεση αίματος:** Συνιστάται ο έλεγχος για υπέρταση σε ενήλικες 18 ετών και άνω με μέτρηση της αρτηριακής πίεσης από ειδικό αλλά και λήψη μετρήσεων εκτός του κλινικού περιβάλλοντος για διαγνωστική επιβεβαίωση πριν από την έναρξη της θεραπείας.
- **Έλεγχος λιπιδίων ορού και χρήση στατινών για την πρόληψη:** Συνιστάται στους ενήλικες χωρίς ιστορικό καρδιαγγειακής νόσου να χρησιμοποιούν χαμηλή έως μέτρια δόση στατίνης για την πρόληψη συμβάντων καρδιαγγειακής νόσου και θνησιμότητας όταν πληρούνται όλα τα ακόλουθα κριτήρια:
    1. είναι ηλικίας 40–75 ετών,
    2. έχουν έναν ή περισσότερους παράγοντες κινδύνου για καρδιαγγειακή νόσο

δηλαδή δυσλιπιδαιμία, σακχαρώδη διαβήτη, υπέρταση ή κάπνισμα και
3. έχουν υπολογισμένο 10ετές κίνδυνο για καρδιαγγειακό επεισόδιο 10% ή μεγαλύτερο [59].

Ο προσδιορισμός της δυσλιπιδαιμίας και ο υπολογισμός του 10ετούς κινδύνου εκδήλωσης καρδιαγγειακής νόσου απαιτεί καθολικό έλεγχο λιπιδίων σε ενήλικες ηλικίας 40-75 ετών. Τα τρέχοντα στοιχεία είναι ανεπαρκή για την αξιολόγηση της ισορροπίας των οφελών και των βλαβών από την έναρξη χρήσης στατινών για την πρωτογενή πρόληψη συμβάντων καρδιαγγειακής νόσου και θνησιμότητας σε ενήλικες ηλικίας 76 ετών και άνω χωρίς ιστορικό καρδιακής προσβολής ή εγκεφαλικού.

- **Χρήση ασπιρίνης:** Συνιστάται η έναρξη χρήσης χαμηλής δόσης ασπιρίνης για την πρωτογενή πρόληψη της καρδιαγγειακής νόσου (CVD):
    1. Σε ενήλικες ηλικίας 50–59 ετών που έχουν 10% ή μεγαλύτερο κίνδυνο καρδιαγγειακής νόσου για 10 χρόνια, δεν διατρέχουν αυξημένο κίνδυνο αιμορραγίας, έχουν προσδόκιμο ζωής τουλάχιστον 10 χρόνια και είναι πρόθυμοι να λαμβάνουν χαμηλή δόση ασπιρίνης καθημερινά για τουλάχιστον 10 χρόνια.
    2. Σε ενήλικες ηλικίας 60–69 ετών που έχουν 10% ή μεγαλύτερο κίνδυνο καρδιαγγειακής νόσου για 10 χρόνια η αντιμετώπιση θα πρέπει να είναι ατομική: Άτομα που δεν διατρέχουν αυξημένο κίνδυνο αιμορραγίας, έχουν προσδόκιμο ζωής τουλάχιστον 10 ετών και είναι πρόθυμα να λαμβάνουν χαμηλή δόση ασπιρίνης καθημερινά για τουλάχιστον 10 χρόνια έχουν περισσότερες πιθανότητες να ωφεληθούν. Επομένως, τα άτομα που δίνουν μεγαλύτερη αξία στα πιθανά οφέλη από τις πιθανές βλάβες μπορούν να επιλέξουν να ξεκινήσουν χαμηλή δόση ασπιρίνης.
    3. Σε ενήλικες κάτω των 50 ετών ή άνω των 70 ετών τα τρέχοντα στοιχεία είναι ανεπαρκή για την αξιολόγηση της ισορροπίας των οφελών και των βλαβών από την έναρξη της χρήσης ασπιρίνης για την πρωτογενή πρόληψη της καρδιαγγειακής νόσου.
- **Συμβουλευτική για υγιεινή διατροφή και σωματική δραστηριότητα για πρόληψη καρδιαγγειακής νόσου:** Συνιστάται σε ενήλικες με παράγοντες κινδύνου καρδιαγγειακής νόσου συμβουλευτικές παρεμβάσεις για την προώθηση υγιεινής διατροφής και σωματικής δραστηριότητας με εξατομικευμένο τρόπο.
- **Έλεγχος για σακχαρώδη διαβήτη:** Συνιστάται έλεγχος του πληθυσμού για λανθάνοντα διαβήτη και διαβήτη τύπου 2 σε ενήλικες ηλικίας 35-70 ετών που είναι υπέρβαροι ή παχύσαρκοι. Οι κλινικοί γιατροί θα πρέπει να προτείνουν σε ασθενείς με λανθάνοντα διαβήτη αποτελεσματικές προληπτικές παρεμβάσεις.
- **Έλεγχος κάπνισματος και συμβουλευτική για τη διακοπή του:** Συνιστάται στους κλινικούς γιατρούς να ρωτούν όλους τους ενήλικες σχετικά με τη χρήση καπνού, να τους συμβουλεύουν να διακόψουν το κάπνισμα και να παρέχουν στους καπνιστές συμπεριφορικές παρεμβάσεις και συνταγογράφηση φαρμακοθεραπείας εγκεκριμένης από τον οργανισμό φαρμάκων των ΗΠΑ (FDA) σε ενήλικες, εκτός των εγκύων.

Μια ανάλογη πρόταση για πρόληψη των καρδιαγγειακών παθήσεων είχε γίνει και στην Ελλάδα το 2008 από επιστημονική επιτροπή ειδικών υπό την αιγίδα του Υπουργείου Υγείας και Κοινωνικής Αλληλεγγύης (Εθνικό Σχέδιο Δράσης για τα Καρδιαγγειακά Νοσήματα 2008 – 2012) [60].

Τα τελευταία χρόνια σημαντική υπήρξε η συνεισφορά της τεχνικής νοημοσύνης στον τομέα της υγείας, τόσο στην πρόληψη όσο και στη διάγνωση και τη θεραπεία των ασθενειών. Η δεκαετία του 2010 έφερε αύξηση στον αριθμό των μελετών και των εργασιών που αναφέρονται στο ρόλο της τεχνητής νοημοσύνης και της μηχανικής μάθησης στην ιατρική και την υγειονομική περίθαλψη. Ο αριθμός των εργασιών βιοεπιστημών που αφορούν σε αυτούς τους τομείς αυξήθηκε από 596 το 2010 σε 12.422 το 2019, ενώ βρισκόμαστε ακόμη στην αρχή αυτής της εποχής [61].

Στον τομέα της καρδιολογίας υπάρχουν σημαντικές εξελίξεις με τον FDA στις ΗΠΑ να έχει ήδη εγκρίνει 57 ιατρικές συσκευές και αλγορίθμους που βασίζονται στην τεχνητή νοημοσύνη και στη μηχανική μάθηση (Artificial Intelligence and Machine Learning (AI/ML) - Enabled Medical Devices) [62] που μπορούν να επικουρήσουν τους επαγγελματίες υγείας, βελτιώνοντας τις παρεχόμενες υπηρεσίες υγείας στους ασθενείς και ευρύτερα στο κοινωνικό σύνολο. Ειδικότερα η χρήση της μηχανικής μάθησης στην καρδιολογία, με τη δημιουργία μοντέλων βασιζόμενα σε αλγόριθμους που έχουν την ικανότητα να εκπαιδεύονται σε έναν μεγάλο όγκο δεδομένων και να εξάγουν σημαντική γνώση από αυτά, μπορούν να αξιοποιηθούν από επαγγελματίες υγείας στη παροχή εξατομικευμένων συμβουλών στο κοινό, βάσει προγνωστικών παραγόντων στην πρόληψη ασθενειών όπως η στεφανιαία νόσος [63]. Προς αυτήν την κατεύθυνση έχει εστιάσει και η παρούσα εργασία. Προσπαθεί να προβλέψει εάν κάποιος εκδηλώσει ΚΝ με κριτήρια τον τρόπο ζωής του και συνυπάρχουσες παθολογικές καταστάσεις, ώστε να κατευθυνθεί στη λήψη μέτρων που θα προστατέψουν την υγεία του από μελλοντική εκδήλωση ΚΝ.

Κεφάλαιο 5: Μηχανική Μάθηση

5.1 Εισαγωγή

Η μηχανική μάθηση είναι ένα κλάδος της τεχνητής νοημοσύνης (Artificial Intelligence), η οποία με την ευρεία έννοια ορίζεται ως η ικανότητα μιας μηχανής να μιμείται την ευφυή ανθρώπινη συμπεριφορά για την εκτέλεση σύνθετων εργασιών με τρόπο παρόμοιο με τον τρόπο με τον οποίο οι άνθρωποι επιλύουν προβλήματα. Ο όρος μηχανική μάθηση χρησιμοποιήθηκε για πρώτη φορά τη δεκαετία του 1950 από τον πρωτοπόρο της τεχνητής νοημοσύνης Arthur Samuel ως «το πεδίο σπουδών που δίνει στους υπολογιστές τη δυνατότητα να μαθαίνουν χωρίς να είναι ρητά προγραμματισμένοι» [64]. Επικεντρώνεται στη διδασκαλία των υπολογιστών να μαθαίνουν από τα δεδομένα και να βελτιώνονται με την εμπειρία χωρίς να έχουν προγραμματιστεί με συγκεκριμένους κανόνες. Στη μηχανική μάθηση, η επίλυση των προβλημάτων επιτυγχάνεται με τη δημιουργία μοντέλων, τα οποία βασιζόμενα σε αλγόριθμους, εκπαιδεύονται για να βρίσκουν μοτίβα και συσχετίσεις σε μεγάλα σύνολα δεδομένων για να λαμβάνουν τις καλύτερες αποφάσεις και προβλέψεις με βάση αυτή την ανάλυση. Τα μοντέλα της μηχανικής μάθησης βελτιώνονται με τη χρήση και γίνονται πιο ακριβή όσο έχουν πρόσβαση σε περισσότερα δεδομένα. Το εκπαιδευμένο μοντέλο εξαρτάται επίσης από την ποιότητα των δεδομένων που χρησιμοποιούνται για την εκπαίδευσή του. Εάν τα δεδομένα είναι προκατειλημμένα, η έξοδος του μοντέλου θα είναι επίσης προκατειλημμένη. Ένα σύστημα μηχανικής μάθησης μπορεί να χρησιμοποιήσει τα δεδομένα για να εξηγήσει τι συνέβη (descriptive), να προβλέψει τι θα συμβεί (predictive) ή για να κάνει προτάσεις σχετικά με τι πρέπει να συμβεί (prescriptive) [65]. Η μηχανική μάθηση χρησιμοποιείται σε μεγάλη ποικιλία εφαρμογών, όπως στην ιατρική, το φιλτράρισμα e-mail,

την αναγνώριση ομιλίας, τη γεωργία και την όραση υπολογιστών, όπου είναι δύσκολο ή ανέφικτο να αναπτυχθούν συμβατικοί αλγόριθμοι για την εκτέλεση των απαραίτητων εργασιών.

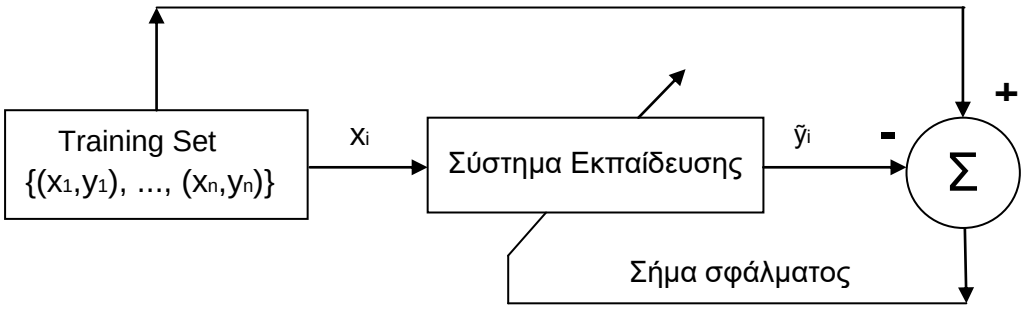
Υπάρχουν τρεις υποκατηγορίες μηχανικής μάθησης:

- **Επιβλεπόμενη μάθηση (supervised learning).** Τα εποπτευόμενα μοντέλα μηχανικής εκμάθησης εκπαιδεύονται με σύνολα δεδομένων με ετικέτα/κλάση που είναι ήδη γνωστή, τα οποία επιτρέπουν στα μοντέλα να μαθαίνουν και να αναπτύσσονται με μεγαλύτερη ακρίβεια στην πάροδο του χρόνου, με σκοπό να λύνουν προβλήματα ταξινόμησης (classification) ή παλινδρόμησης (regression). Η επιβλεπόμενη μηχανική εκμάθηση είναι ο πιο κοινός τύπος που χρησιμοποιείται σήμερα.
- **Μη επιβλεπόμενη ή αυτό-οργανούμενη μάθηση (unsupervised learning).** Σε αντίθεση με την επιβλεπόμενη μάθηση, εδώ ένας αλγόριθμος αναζητά μοτίβα σε δεδομένα χωρίς ετικέτα. Ένα σύστημα χωρίς επίβλεψη μπορεί να προσδιορίσει κοινά χαρακτηριστικά και να ομαδοποιήσει τα δεδομένα εισόδου σε ξεχωριστές κατηγορίες. Μέχρι στιγμής, αυτή η προσέγγιση έχει χρησιμοποιηθεί αποτελεσματικά σε εφαρμογές φυσικής γλώσσας και στη ρομποτική.
- **Ενισχυτική μάθηση**. Η ενισχυτική μηχανική μάθηση εκπαιδεύει τις μηχανές μέσω δοκιμής και λάθους ώστε να κάνουν την καλύτερη ενέργεια με τη δημιουργία ενός συστήματος ανταμοιβής. Η ενισχυτική μάθηση μπορεί να εκπαιδεύσει μοντέλα να παίζουν παιχνίδια ή να εκπαιδεύει αυτόνομα οχήματα να οδηγούν λέγοντας στο μηχάνημα πότε πήρε τις σωστές αποφάσεις, κάτι που το βοηθά να μάθει με την πάροδο του χρόνου ποιες ενέργειες πρέπει να κάνει.

Στην συγκεκριμένη διπλωματική εργασία χρησιμοποιείται η επιβλεπόμενη μάθηση και ειδικότερα σε πρόβλημα δυαδικής ταξινόμησης (binary classification).

5.2 Επιβλεπόμενη μάθηση

Οι αλγόριθμοι μηχανικής μάθησης αντιμετωπίζουν κάθε παράδειγμα ενός συνόλου δεδομένων (dataset) ως μια συλλογή χαρακτηριστικών. Αυτά τα χαρακτηριστικά μπορεί να είναι δυαδικά (binary), κατηγορικά (categorical), ή συνεχούς χαρακτήρα (continuous). Η επιβλεπόμενη μάθηση περιλαμβάνει την εκπαίδευση του μοντέλου σε δεδομένα με ετικέτα και τη δοκιμή του σε δεδομένα χωρίς ετικέτα. Τα μοντέλα κατασκευάζουν τις συναρτήσεις που απεικονίζουν δεδομένες εισόδους σε γνωστές, επιθυμητές εξόδους (σύνολο εκπαίδευσης), με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Αρχικά γίνεται η συλλογή δεδομένων, τα οποία στη συνέχεια διαιρούνται σε δεδομένα εκπαίδευσης (training set) και δεδομένα δοκιμών (test set). Μετά τα δεδομένα υφίστανται προεπεξεργασία. Τα προκύπτοντα χαρακτηριστικά τροφοδοτούνται σε έναν αλγόριθμο και στη συνέχεια το μοντέλο εκπαιδεύεται για να μάθει τα χαρακτηριστικά που σχετίζονται με κάθε ετικέτα. Τέλος, στο μοντέλο εισάγονται τα δεδομένα δοκιμής και οι έξοδοι συγκρίνονται με τις πραγματικές εξόδους και αξιολογούνται με τις κατάλληλες μετρικές.



Εικόνα 6: *Σχηματικό διάγραμμα που απεικονίζει τη διαδικασία της επιβλεπόμενης μάθησης* [66] *(Προσαρμογή)*

Η Εικόνα 6 δείχνει ένα σχηματικό διάγραμμα που απεικονίζει τη διαδικασία της επιβλεπόμενης μάθησης. Σε αυτό το διάγραμμα, το διάνυσμα (x_i,y_i) είναι ένα δείγμα εκπαίδευσης, όπου το 'x' αντιπροσωπεύει την είσοδο του συστήματος, το 'y' αντιπροσωπεύει την έξοδο του συστήματος (δηλαδή, την επίβλεψη ή την ετικέτα της εισόδου x) και το 'i' είναι ο δείκτης του δείγματος εκπαίδευσης. Κατά τη διάρκεια της διαδικασίας μια είσοδος εκπαίδευσης xi τροφοδοτείται στο σύστημα εκπαίδευσης και αυτό δημιουργεί μια έξοδο $\tilde{y}_i$. Η έξοδος του συστήματος εκμάθησης $\tilde{y}_i$ συγκρίνεται στη συνέχεια με τη γνωστή ετικέτα εισόδου y_i από έναν “διαιτητή” που υπολογίζει τη διαφορά μεταξύ τους. Η διαφορά, που ονομάζεται σήμα σφάλματος σε αυτό το διάγραμμα, αποστέλλεται στη συνέχεια στο σύστημα εκπαίδευσης για προσαρμογή των παραμέτρων του εκπαιδευόμενου μοντέλου. Η διαδικασία εκμάθησης αποτελεί δηλαδή ένα σύστημα ανάδρασης κλειστού βρόχου. Ο στόχος αυτής της μαθησιακής διαδικασίας είναι να αποκτήσει ένα σύνολο βέλτιστων παραμέτρων του συστήματος εκμάθησης που μπορούν να ελαχιστοποιήσουν τις διαφορές μεταξύ $\tilde{y}_i$ και y_i για όλα τα i, δηλαδή ελαχιστοποιώντας το συνολικό σφάλμα σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης [66].

Υπάρχουν δύο γενικοί τύποι επιβλεπόμενης μάθησης: η ταξινόμηση και η παλινδρόμηση.

5.2.1 Ταξινόμηση

Στη επιβλεπόμενη μάθηση όταν η έξοδος είναι σε κατηγορική μορφή, το πρόβλημα αναφέρεται ως πρόβλημα ταξινόμησης. Γενικά, σε προβλήματα ταξινόμησης, τα δείγματα ενός συνόλου δεδομένων κατηγοριοποιούνται σε καθορισμένες κλάσεις βάσει των τιμών των χαρακτηριστικών τους. Ένας αλγόριθμος ταξινόμησης, που εναλλακτικά αναφέρεται ως ταξινομητής (classifier), εκπαιδεύεται στο training set και αντιστοιχίζει κάθε νέο δείγμα σε μια συγκεκριμένη κλάση. Υπάρχουν δύο διαφορετικοί τύποι ταξινόμησης βάσει αριθμού κλάσεων κατηγοριοποίησης: η δυαδική (binary), όταν η ταξινόμηση γίνεται σε δύο κατηγορίες και η πολλαπλών κλάσεων (multi-class), όταν αυτή γίνεται σε περισσότερες από δύο κατηγορίες.

5.2.2 Παλινδρόμηση

Η παλινδρόμηση είναι ένα από τα πιο βασικά εργαλεία στον τομέα της μηχανικής μάθησης, που χρησιμοποιείται για προβλήματα πρόβλεψης. Εμπίπτει στην επιβλεπόμενη μάθηση όπου ο αλγόριθμος προβλέπει στην έξοδο συνεχείς αριθμητικές τιμές, για κάθε νέο

δείγμα στην είσοδο. Βοηθά στη δημιουργία μιας σχέσης μεταξύ τυχαίων μεταβλητών, εκτιμώντας πώς η μία επηρεάζει την άλλη. Συγκεκριμένα, είναι η διαδικασία εκτίμησης της σχέσης μεταξύ ανεξάρτητων μεταβλητών, τους παλινδρομητές και μιας εξαρτώμενης μεταβλητής, της απόκρισης. Η εξάρτηση της απόκρισης από τους παλινδρομητές περιλαμβάνει έναν προσθετικό όρο σφάλματος, το προσδοκώμενο σφάλμα, μέσω του οποίου συνυπολογίζονται οι αβεβαιότητες στον τρόπο με τον οποίο διατυπώνεται αυτή η εξάρτηση. Υπάρχουν δύο κατηγορίες μοντέλων παλινδρόμησης: τα γραμμικά και τα μη γραμμικά. Αυτή η προσαρμογή της λειτουργίας εξυπηρετεί δύο σκοπούς: Εκτίμηση των δεδομένων που λείπουν εντός του γνωστού εύρους δεδομένων (Interpolation) και εκτίμηση μελλοντικών δεδομένων εκτός του γνωστού εύρους των δεδομένων (Extrapolation). Μερικά παραδείγματα χρήσης αυτών των μοντέλων περιλαμβάνουν την πρόβλεψη της τιμής ενός σπιτιού με δεδομένα χαρακτηριστικά σπιτιού, την πρόβλεψη πωλήσεων, την πρόβλεψη της θερμοκρασίας κ.ά. [67].

5.3 Αλγόριθμοι επιβλεπόμενης μάθησης

Υπάρχουν πολλοί αλγόριθμοι που εφαρμόζονται στη επιβλεπόμενη μηχανική μάθηση. Η επιλογή του αλγορίθμου εξαρτάται από το συγκεκριμένο πρόβλημα, τη φύση των δεδομένων, τα επιθυμητά αποτελέσματα και τους διαθέσιμους υπολογιστικούς πόρους. Οι συνηθέστεροι αλγόριθμοι που χρησιμοποιούνται είναι:

5.3.1 Κ-Πλησιέστεροι Γείτονες (K-Nearest Neighbors)

Ο αλγόριθμος K-Nearest Neighbors (KNN) είναι ένας απλός μη παραμετρικός αλγόριθμος επιβλεπόμενης μάθησης, που αναπτύχθηκε για πρώτη φορά από την Evelyn Fix και τον Joseph Hodges το 1951 [68] και αργότερα επεκτάθηκε από τον Thomas Cover [69]. Ανήκει στην κατηγορία των αλγορίθμων μάθησης που βασίζονται σε στιγμιότυπα (instance- based learning ή memory-based learning), η οποία είναι μια οικογένεια αλγορίθμων μάθησης που, αντί να εκτελούν ρητή γενίκευση, συγκρίνουν νέα στιγμιότυπα με αυτά που εμφανίζονται στην εκπαίδευση, που έχουν αποθηκευτεί στη μνήμη. Αποτελεί ένα από τους lazy-learning αλγόριθμους (“τεμπέληδες” στην εκμάθηση) [70], οι οποίοι απαιτούν λιγότερο υπολογιστικό χρόνο κατά τη φάση της εκπαίδευσης, αλλά περισσότερο κατά τη διαδικασία ταξινόμησης. Μπορεί να χρησιμοποιηθεί τόσο για προβλήματα παλινδρόμησης όσο και για προβλήματα ταξινόμησης. Στην ταξινόμηση ένα αντικείμενο ταξινομείται με βάση την πλειοψηφική τάξη των k πλησιέστερων γειτόνων του στο σύνολο εκπαίδευσης. Το k είναι ένας θετικός ακέραιος, συνήθως μικρός και περιττός. Αν k = 1, τότε το αντικείμενο απλώς εκχωρείται στην κλάση που ανήκει ο πλησιέστερος γείτονας. Στην παλινδρόμηση, η έξοδος είναι ο μέσος όρος των τιμών των k πλησιέστερων γειτόνων. Εάν k = 1, τότε η έξοδος εκχωρείται απλώς στην τιμή που έχει ο πλησιέστερος γείτονας. Πολύ σημαντική απόφαση κατά τη χρήση του αλγόριθμου KNN είναι η επιλογή της κατάλληλης μετρικής απόστασης για την εύρεση των πλησιέστερων γειτόνων. Αυτή γίνεται συνήθως με τη χρήση της απόστασης Minkowski. Εάν x,y δύο διανύσματα και n το πλήθος των χαρακτηριστικών, τότε:

- Η απόσταση Minkowski υπολογίζεται από τον τύπο:

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \quad (2)$$

- Η απόσταση Manhattan προκύπτει από τον τύπο Minkowski για p=1:

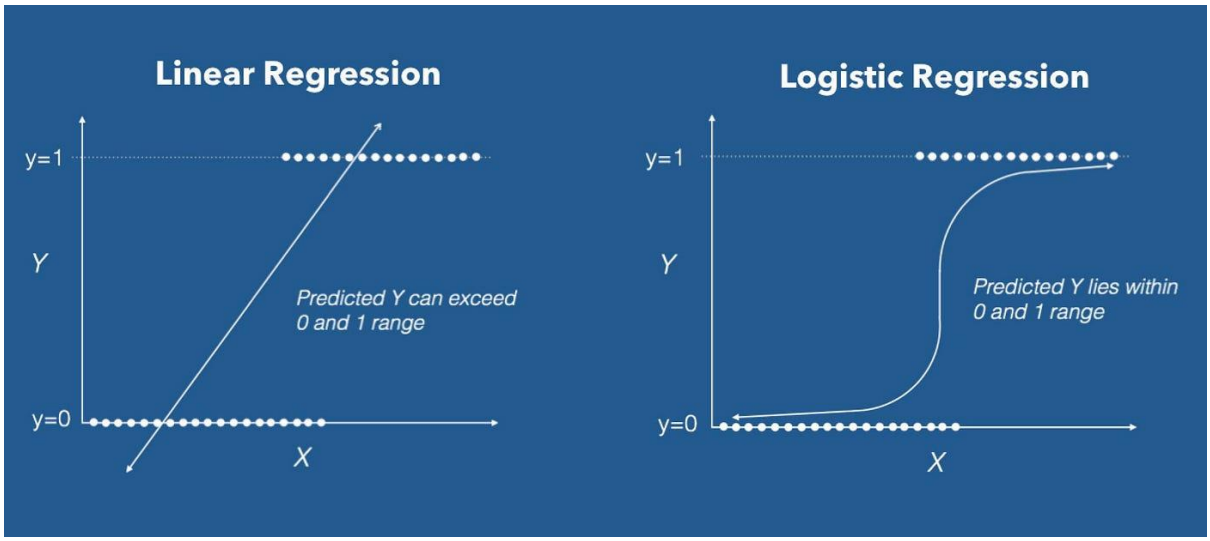
$$d_M(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (3)$$

- Η ευκλείδεια απόσταση προκύπτει από τον τύπο Minkowski για p=2:

$$d_e(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4)$$

Ο KNN είναι αλγόριθμος εύκολος στην εφαρμογή, αποτελεσματικός τόσο για μικρά όσο και για μεγάλα σύνολα δεδομένων ακόμα και εάν περιέχουν θόρυβο. Όμως έχει υψηλό υπολογιστικό κόστος, επειδή χρειάζεται να προσδιοριστεί το βέλτιστο k [71] και να υπολογιστούν οι αποστάσεις μεταξύ κάθε νέου στιγμιότυπου και όλων των δειγμάτων εκπαίδευσης [72].

5.3.2 Λογιστική Παλινδρόμηση (Logistic Regression)



Εικόνα 7: *Γραμμική και λογιστική παλινδρόμηση* [73]

Το μοντέλο της λογιστικής παλινδρόμησης ανήκει στην οικογένεια των στατιστικών μοντέλων παλινδρόμησης, τα οποία είναι γνωστά ως γενικευμένα γραμμικά μοντέλα (Generalized Linear Models – GLM) [74]. Η έννοια της λογιστικής παλινδρόμησης χρονολογείται από τον 19ο αιώνα, αλλά η συγκεκριμένη διατύπωση και ανάπτυξη του αλγορίθμου λογιστικής παλινδρόμησης όπως τον γνωρίζουμε σήμερα μπορεί να αποδοθεί σε πολλά άτομα. Συνδέεται με το όνομα του στατιστικολόγου David Cox [75], ενός βρετανού στατιστικολόγου που συνέβαλε σημαντικά στην ανάπτυξη της λογιστικής παλινδρόμησης στις δεκαετίες του 1950 και του 1960. Ο Cox επέκτεινε την έννοια της ανάλυσης παλινδρόμησης από τη γραμμική παλινδρόμηση στα προβλήματα δυαδικής ταξινόμησης, εισάγοντας τη λογιστική συνάρτηση για να μοντελοποιήσει τη σχέση μεταξύ των μεταβλητών πρόβλεψης και της πιθανότητας να ανήκουν σε μια συγκεκριμένη κλάση. Έκτοτε ο αλγόριθμος λογιστικής παλινδρόμησης έχει βασιστεί στην εργασία πολλών άλλων ερευνητών και στατιστικολόγων που ακολούθησαν.

Η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης βασίζεται στη φύση της επιλεγμένης μεταβλητής εξόδου, η οποία στην μεν πρώτη είναι

διακριτή (discrete) στη δε δεύτερη λαμβάνει συνεχείς αριθμητικές τιμές (continuous numerical) (Εικόνα 7). Διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης ανάλογα με την ιδιαίτερη φύση της εξαρτημένης κατηγορικής μεταβλητής, η οποία μπορεί να είναι:

- Δυαδική (binary) μεταβλητή.
- Τακτική (ordinal) μεταβλητή. Η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες μεταξύ των οποίων ισχύει η έννοια της ανισότητας.
- Ονομαστική (nominal) ή πολυωνυμική (polynomial) ή κατηγορική χωρίς διαβάθμιση (non-ordered categorical). Περιέχει τρεις ή περισσότερες κατηγορίες χωρίς κάποια φυσική διαβάθμιση.

Η δυαδική λογιστική παλινδρόμηση [76], [77] χρησιμοποιεί τη θεωρία πιθανοτήτων για να προβλέψει την πιθανότητα επιλογής της κλάσης εξόδου Y_{i}, που ακολουθεί την κατανομή Bernoulli (λαμβάνει δύο τιμές 0 ή 1), με χρήση των χαρακτηριστικών (προγνωστικών παραγόντων) πλήθους k, αριθμητικών ή/και κατηγορικών, ενός στιγμιοτύπου εισόδου $\widetilde{\mathrm{X}}_{\mathrm{i}} = (\mathrm{x}_1, \mathrm{x}_{2,} \mathrm{x}_3, \dots, \mathrm{x}_{\mathrm{k}})$, εισάγοντάς το στη λογιστική συνάρτηση (logistic function). Η λογιστική συνάρτηση χρησιμοποιείται συνήθως για τη μοντελοποίηση κάθε θετικού στιγμιότυπου $\widetilde{\mathrm{X}}_{\mathrm{i}}$ με την αναμενόμενη δυαδική έξοδο να δίνεται από την εξής σχέση:

$$P\left(Y_{\mathrm{i}} = 1 | \widetilde{\mathrm{X}}_{\mathrm{i}}\right) = p_i = f(z) = \frac{1}{1+e^{-z}} \quad (5)$$

ή ισοδύναμα:

$$logit(p_i) = ln\frac{p_i}{1-p_i} = b_0 + b_1 x_1 + b_2 x_2 + \ \dots + b_k x_k = z \quad (6)$$

Όπου:

- $logit(p_i)$ ο λογαριθμικός μετασχηματισμός του p_i, ο οποίος εδώ είναι ο λογάριθμος των πιθανοτήτων θετικής απόκρισης,
- $z = b_0 + b_1 x_1 + b_2 x_2 + \ \dots + b_k x_k$ είναι ο γραμμικός συνδυασμός των τιμών των χαρακτηριστικών του στιγμιοτύπου,
- b_o το ύψος της κλίσης της γραμμής παλινδρόμησης και
- b_i οι συντελεστές παλινδρόμησης (coefficients) καθένας εκ των οποίων εκφράζει το μέγεθος συνεισφοράς του αντίστοιχου χαρακτηριστικού x_j.

Οι συντελεστές της παλινδρόμησης b_i υπολογίζονται κατά τη διάρκεια της εκπαίδευσης με τη βοήθεια της εκτίμησης της μέγιστης πιθανοφάνειας (Maximum Likelihood Estimate – MLE), δηλαδή της μεγιστοποίησης της συνάρτησης πιθανοφάνειας L :

$$L = \prod_{i=1}^{n} f(\widetilde{\mathrm{X}}_{\mathrm{i}} | \, \theta) \quad (7)$$

ή προτιμότερα της λογαριθμικής έκδοσης αυτής:

$$\ln L = \sum_{i=1}^{n} \ln f\left(\widetilde{\mathrm{X}}_{\mathrm{i}} | \theta\right) \quad (8)$$

όπου n το πλήθος των στιγμιοτύπων στο training set και θ είναι μια παράμετρος της μεταβλητής η οποία μπορεί να μεταβάλλεται ελεύθερα. Η προβλεπόμενη τιμή για κάθε παρατήρηση (μέση λογαριθμική πιθανοφάνεια) θα ισούται με:

$$\hat{l} = \frac{1}{n} lnL(b_i) \quad (9)$$

Η συνάρτηση της πιθανοφάνειας έκβασης ενός γεγονότος (likelihood) δείχνει πόσο κατάλληλα ένα παρατηρούμενο δείγμα περιγράφεται από κάποιες τιμές παραμέτρων π.χ. μέσος όρος, τυπική απόκλιση. Άρα, η μεγιστοποίηση της συνάρτησης της πιθανότητας έκβασης καθορίζει τις παραμέτρους εκείνες που είναι οι πλέον ικανές να παράγουν τα παρατηρούμενα στοιχεία. Για τη εύρεση της MLE χρησιμοποιούνται αριθμητικές μέθοδοι βελτιστοποίησης που αρχίζουν με μία υπόθεση και κάνουν επαναλήψεις για να βελτιώσουν αυτήν την υπόθεση. Η πιο κοινή είναι η μέθοδος Newton-Raphson [77].

Η λογιστική παλινδρόμηση είναι πολύ χρήσιμη για την κατανόηση της βαρύτητας της επιρροής κάθε προγνωστικού παράγοντα στην επιλογή της κλάσης στην έξοδο, επειδή κάθε ένας από αυτούς έχει και έναν συντελεστή παλινδρόμησης. Επειδή δεν υπολογίζει αποστάσεις μεταξύ των χαρακτηριστικών, μπορεί να μαθαίνει γρηγορότερα από μεθόδους όπως ο KNN. Όμως η ισχύς της περιορίζεται στην ταξινόμηση μόνο γραμμικά διαχωρίσιμων δεδομένων [74], [78].

5.3.3 Αφελής Μπεϋζιανός (Naive Bayes)

Η προέλευση του Naive Bayes Classifier (NB) μπορεί να εντοπιστεί στον 18ο αιώνα και στο έργο του αιδεσιμότατου Thomas Bayes, οι σημειώσεις του οποίου δημοσιεύθηκαν μετά το θάνατό του από τον Richard Price [79], στις οποίες περιγράφεται μια ειδική περίπτωση του θεωρήματος που φέρει πλέον το όνομά του και χρησιμοποιείται για των υπολογισμό των υπό συνθήκη πιθανοτήτων. Το θεώρημα αυτό περιγράφει την πιθανότητα να συμβεί ένα γεγονός βάσει προϋπάρχουσας γνώσης που μπορεί να σχετίζεται με αυτό.

Σε αυτό βασίζεται και ο ταξινομητής Naive Bayes [74]. Για τους σκοπούς της ταξινόμησης, υπολογίζεται η πιθανότητα επιλογής μιας ετικέτας κλάσης y για ένα στιγμιότυπο εισόδου x με δεδομένες τις τιμές των χαρακτηριστικών του. Αυτή μπορεί να αναπαρασταθεί ως P(y|x), η οποία είναι γνωστή ως η εκ των υστέρων ή αναθεωρημένη πιθανότητα (posterior probability) της κλάσης επιλογής y. Χρησιμοποιώντας το θεώρημα Bayes, μπορούμε να αναπαραστήσουμε την πιθανότητα αυτή ως:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad (10)$$

όπου το P(x|y) είναι η πιθανότητα κλάσης υπό όρους (class-conditional probability) του x, με δεδομένη τη επιλογή κλάσης y. Η P(x|y) μετρά τη πιθανότητα να παρατηρηθεί το x από την κατανομή των στιγμιοτύπων που ανήκουν στην κατηγορία y. Εάν το x ανήκει πράγματι στην κατηγορία y, τότε θα πρέπει να περιμένουμε ότι το P(x|y) θα είναι υψηλό. Από αυτή την άποψη, η χρήση πιθανοτήτων κλάσης υπό όρους επιχειρεί να αποτυπώσει τη διαδικασία από την οποία δημιουργήθηκαν τα στιγμιότυπα δεδομένων.

Ο δεύτερος όρος στον αριθμητή της σχέσης (10) P(y) είναι η εκ των προτέρων πιθανότητα (prior probability) να έχω έξοδο y ανεξάρτητα από την είσοδο x. Δηλαδή περιλαμβάνει τις

προηγούμενες πεποιθήσεις μας σχετικά με την κατανομή των ετικετών κλάσεων, ανεξάρτητα από τις παρατηρούμενες τιμές χαρακτηριστικών. (Αυτή είναι η άποψη του Bayes.) Η εκ των προτέρων πιθανότητα μπορεί να ληφθεί με τη χρήση προηγούμενων ειδικών γνώσεων.

Ο παρονομαστής στην σχέση (10) περιλαμβάνει την πιθανότητα απόδειξης, P(x) που εκφράζει την πιθανότητα να έχω είσοδο x ανεξάρτητα από την έξοδο y. Σημειώνεται ότι αυτός ο όρος δεν εξαρτάται από την επιλογή κλάσης και επομένως μπορεί να αντιμετωπιστεί ως σταθερά κανονικοποίησης στον υπολογισμό των μεταγενέστερων πιθανοτήτων. Περαιτέρω, η τιμή του P(x) μπορεί να υπολογιστεί ως

$$P(x) = \sum_i P(x|y_i)P(y_i) \quad (11)$$

Ο Naive Bayes υποθέτει ότι η πιθανότητα κλάσης υπό όρους όλων των χαρακτηριστικών του στιγμιοτύπου X μπορεί να συνυπολογιστεί ως γινόμενο των πιθανοτήτων κλάσης υπό όρους για κάθε χαρακτηριστικό x_i, όπως περιγράφεται στην ακόλουθη εξίσωση [74]:

$$P(X|y) = \prod_{i=1}^{d} P(x_i|y) \quad (12)$$

όπου κάθε στιγμιότυπο δεδομένων X αποτελείται από d χαρακτηριστικά, $\{x_1, x_{2,} \dots x_d\}$. Η βασική υπόθεση πίσω από την σχέση (12) είναι ότι οι τιμές των χαρακτηριστικών x_i είναι υπό όρους ανεξάρτητες η μία από την άλλη, δεδομένης της κλάσης y (γι' αυτό άλλωστε λέγεται και αφελής) . Αυτό σημαίνει ότι τα χαρακτηριστικά επηρεάζονται μόνο από την κλάση και αν τη γνωρίζουμε, μπορούμε να θεωρήσουμε ότι τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο.

Χρησιμοποιώντας την υπόθεση Bayes, χρειάζεται μόνο να υπολογιστεί η υπό όρους πιθανότητα για κάθε x_i δεδομένου του y ξεχωριστά, αντί να υπολογιστεί για κάθε συνδυασμό τιμών χαρακτηριστικών. Για παράδειγμα, εάν n_i^0και n_j^0 είναι το πλήθος των στιγμιοτύπων που ανήκουν στην κλάση 0 με τιμές χαρακτηριστικών $x_1 = c_i$ και $x_2 = c_j$, αντίστοιχα, τότε η πιθανότητα κλάσης υπό όρους μπορεί να εκτιμηθεί ως:

$$P(x_1 = c_i\, , x_2 = c_j\, |y = 0) = \frac{n_i^0}{n^0} * \frac{n_j^0}{n^0} \quad (13)$$

Στην σχέση (13), χρειάζεται μόνο να μετρηθεί ο αριθμός των περιπτώσεων εκπαίδευσης για κάθε μία από τις k διαφορετικές τιμές ενός χαρακτηριστικού x_i, ανεξάρτητα από τις τιμές άλλων χαρακτηριστικών. Ως εκ τούτου, ο αριθμός των παραμέτρων που απαιτούνται για την εκμάθηση πιθανοτήτων κλάσης υπό όρους μειώνεται από d^k σε $d * k$. Αυτό απλοποιεί πολύ τον υπολογισμό της πιθανότητας κλάσης υπό όρους και καθιστά πιο επιδεκτικό τον εκτιμητή στην εκμάθηση παραμέτρων και στην πραγματοποίηση προβλέψεων, ακόμη και σε χώρους υψηλών διαστάσεων.

Ο ταξινομητής Naive Bayes υπολογίζει την αναθεωρημένη πιθανότητα επιλογής κλάσης y για μια δοκιμαστική περίπτωση στιγμιοτύπου x χρησιμοποιώντας την ακόλουθη εξίσωση:

$$P(y|x) = \frac{P(y)\prod_{i=1}^{d} P(x_i|y)}{P(x)} \quad (14)$$

Ο ταξινομητής αυτός χρησιμοποιείται στη μηχανική μάθηση για εργασίες ταξινόμησης, ιδιαίτερα στην κατηγοριοποίηση κειμένου και το φιλτράρισμα ανεπιθύμητων μηνυμάτων. Είναι απλός στην εφαρμογή του, απαιτεί λίγα δεδομένα εκπαίδευσης και μπορεί να χειριστεί μεγάλα σύνολα δεδομένων αποτελεσματικά. Δεν επηρεάζεται από απομονωμένα στιγμιότυπα που αποτελούν θόρυβο. Δεν λαμβάνει υπ' όψιν του χαρακτηριστικά που δεν σχετίζονται με το πρόβλημα, καθώς για αυτά η πιθανότητα $P(x_i|y)$ γίνεται σχεδόν ομοιόμορφα κατανεμημένη για κάθε κλάση y. Είναι εξαιρετικά γρήγορος σε σύγκριση με πιο εξελιγμένες μεθόδους, καθώς μπορεί εύκολα και σε σύντομο χρόνο να βγάλει ικανοποιητικά αποτελέσματα σε προβλήματα με πολλές διαστάσεις. Πρέπει όμως πάντα τα χαρακτηριστικά να είναι ανεξάρτητα μεταξύ τους με δεδομένη την επιλογή κλάσης, κάτι που πολύ συχνά δεν ισχύει στην πραγματικότητα. Σε αυτές τις περιπτώσεις ο Naive Bayes δεν έχει καλές επιδόσεις [74], [80].

5.3.4 Μηχανές διανυσμάτων υποστήριξης - Support Vector Machine (SVM)

Ο αλγόριθμος Support Vector Machines (SVM) είναι αλγόριθμος επιβλεπόμενης μάθησης που χρησιμοποιείται τόσο για εργασίες ταξινόμησης όσο και παλινδρόμησης. Αναπτύχθηκε από τον Vladimir Vapnik και τους συνεργάτες του [81] και βασίζεται στη θεωρία Vapnik–Chervonenkis (VC theory) , η οποία επιχειρεί να εξηγήσει τη μαθησιακή διαδικασία από στατιστική άποψη [82]. Ο SVM λειτουργεί βρίσκοντας ένα βέλτιστο υπερεπίπεδο ως επιφάνεια απόφασης σε ένα χώρο χαρακτηριστικών πολλών διαστάσεων που διαχωρίζει στο μέγιστο δυνατό βαθμό τα σημεία δεδομένων διαφορετικών κλάσεων. Μπορεί να χρησιμοποιηθεί για διαχωρισμό δεδομένων που είτε διαχωρίζονται γραμμικά είτε όχι [67].

Όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα ο SVM στοχεύει να βρει το βέλτιστο υπερεπίπεδο που διαχωρίζει τα σημεία δεδομένων διαφορετικών κλάσεων με το μέγιστο περιθώριο. Το περιθώριο ορίζεται ως η απόσταση μεταξύ του υπερεπίπεδου και των πλησιέστερων σημείων δεδομένων από κάθε κατηγορία. Ο SVM προσδιορίζει ένα υποσύνολο σημείων δεδομένων που ονομάζονται διανύσματα υποστήριξης (support vectors), τα οποία είναι τα σημεία που βρίσκονται πιο κοντά στο όριο απόφασης ή στο υπερεπίπεδο. Αυτά τα διανύσματα υποστήριξης παίζουν κρίσιμο ρόλο στον προσδιορισμό του βέλτιστου υπερεπίπεδου, επηρεάζοντας τη θέση και την κατεύθυνσή του. Το πρόβλημα βελτιστοποίησης του υπερεπίπεδου μπορεί να διατυπωθεί ως πρόβλημα τετραγωνικού προγραμματισμού και επιλύεται χρησιμοποιώντας τεχνικές όπως οι πολλαπλασιαστές Lagrange με στόχο να ελαχιστοποιηθεί το σφάλμα ταξινόμησης και να μεγιστοποιηθεί το περιθώριο ταυτόχρονα [83].

Όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα αλλά αυτό οφείλεται σε θόρυβο ή σε λίγες λάθος ταξινομήσεις ο SVM μπορεί να το αντιμετωπίσει με τη χρήση των μεταβλητών χαλαρότητας εισάγοντας τη παράμετρο C που επιλέγεται από το χρήστη και είναι το βάρος του κόστους των λανθασμένων ταξινομήσεων. Η υπερπαράμετρος C ελέγχει την αντιστάθμιση μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης του σφάλματος ταξινόμησης στα δεδομένα εκπαίδευσης. Μικρότερη τιμή C επιλέγεται όταν το δείγμα εκπαίδευσης θεωρείται θορυβώδες και οδηγεί σε μεγαλύτερο περιθώριο αλλά επιτρέπει περισσότερες εσφαλμένες ταξινομήσεις. Αντίστροφα, μεγαλύτερη τιμή C επιλέγεται όταν υπάρχει μεγάλη εμπιστοσύνη στο δείγμα εκπαίδευσης και οδηγεί σε μικρότερο περιθώριο αλλά λιγότερες εσφαλμένες ταξινομήσεις [67].

Για την ταξινόμηση σε δύο κλάσεις δεδομένων όταν αυτά είναι εγγενώς μη γραμμικά διαχωρίσιμα, είναι απαραίτητο να αντιστοιχιστούν σε έναν χώρο υψηλότερων διάστασεων για να επιτραπεί ο γραμμικός διαχωρισμός των κλάσεων. Αυτή η διαδικασία γίνεται στον SVM με τη εφαρμογή του Τεχνάσματος Πυρήνα (kernel trick) [67], που βασίζεται στο θεώρημα Cover [84], σύμφωνα με το οποίο, κάθε πολυδιάστατος χώρος με μη γραμμικά διαχωρίσιμα πρότυπα, μπορεί να μετασχηματιστεί σε ένα νέο χώρο στον οποίο τα πρότυπα είναι γραμμικά διαχωρίσιμα με υψηλή πιθανότητα, αρκεί ο μετασχηματισμός να είναι μη γραμμικός και ο νέος αυτός χώρος να έχει την απαραίτητη διάσταση. Αυτό επιτυγχάνεται με τη χρήση συναρτήσεων πυρήνα όπως η Γκαουσιανή RBF, η Πολυωνυμική, η Σιγμοειδής και η αντίστροφη πολυτετραγωνική.

Μόλις επιτευχθεί το βέλτιστο υπερεπίπεδο, ο SVM μπορεί να κάνει προβλέψεις σε νέα, άγνωστα σημεία δεδομένων. Ο αλγόριθμος υπολογίζει την απόσταση μεταξύ του νέου σημείου δεδομένων και του υπερεπίπεδου. Ανάλογα σε ποια πλευρά του υπερεπίπεδου πέφτει το σημείο, του εκχωρείται μια ετικέτα κλάσης.

Ο SVM λειτουργεί σχετικά καλά όταν υπάρχει ένα σαφές περιθώριο διαχωρισμού μεταξύ των κλάσεων ή όταν εμπλέκονται μη γραμμικά όρια απόφασης. Είναι πιο αποτελεσματικός σε χώρους υψηλών διαστάσεων και σε περιπτώσεις όπου ο αριθμός των διαστάσεων είναι μεγαλύτερος από τον αριθμό των δειγμάτων.

Από τη άλλη πλευρά όμως, ο αλγόριθμος SVM δεν είναι κατάλληλος για μεγάλα σύνολα δεδομένων, διότι απαιτεί την επίλυση ενός προβλήματος τετραγωνικής βελτιστοποίησης, γεγονός που τον καθιστά υπολογιστικά ακριβό και απαιτητικό σε μνήμη. Καθώς ο ταξινομητής διανύσματος υποστήριξης λειτουργεί βάζοντας σημεία δεδομένων, πάνω και κάτω από το υπερεπίπεδο ταξινόμησης, δεν υπάρχει πιθανολογική εξήγηση για την ταξινόμηση, άρα και διαφάνεια [67], [83].

5.3.5 Δένδρο απόφασης (Decision Tree)

Ένα δέντρο απόφασης είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιείται για εργασίες ταξινόμησης και παλινδρόμησης. Χτίζει από πάνω προς τα κάτω μια ιεραρχική δομή αποφάσεων που μοιάζει με δέντρο, αντιπροσωπεύοντας ένα μοντέλο αποφάσεων και τα πιθανά αποτελέσματά τους, που μοιάζει με διάγραμμα ροής. Ο κορυφαίος κόμβος απόφασης σε ένα δέντρο αντιστοιχεί στον καλύτερο προγνωστικό παράγοντα και ονομάζεται ρίζα του δένδρου (root node). Διαφορετικοί αλγόριθμοι χρησιμοποιούν διαφορετικές μετρικές για την εύρεση του "καλύτερου". Αυτές γενικά μετρούν την ομοιογένεια της μεταβλητής στόχου εντός των υποσυνόλων. Κάθε εσωτερικός κόμβος στο δέντρο αντιπροσωπεύει μια απόφαση που βασίζεται σε ένα χαρακτηριστικό και κάθε κόμβος φύλλο αντιπροσωπεύει μια ετικέτα κλάσης ή μια προβλεπόμενη τιμή. Τα δέντρα αποφάσεων μπορούν να χειριστούν τόσο κατηγορικά όσο και αριθμητικά δεδομένα [85]. Για να κάνει προβλέψεις σε νέα άγνωστα δεδομένα, ο αλγόριθμος διασχίζει το δέντρο αποφάσεων με βάση τις τιμές των χαρακτηριστικών εισόδου, ακολουθώντας τη διαδρομή από τον ριζικό κόμβο σε έναν κόμβο φύλλο. Η τελική πρόβλεψη είναι η ετικέτα ή η τιμή κλάσης που σχετίζεται με τον καταληκτικό κόμβο φύλλο.

Ανάλογα με την υποκείμενη μέτρηση, η απόδοση διαφόρων ευρετικών αλγορίθμων για τη μάθηση δέντρων αποφάσεων μπορεί να ποικίλλει σημαντικά. Υπάρχουν δύο βασικοί αλγόριθμοι για τη δημιουργία δέντρων απόφασης:

- Ο C4.5 από τον J. R. Quinlan [86] (επέκταση του ID3 του ιδίου), ο οποίος χρησιμοποιείται για ταξινόμηση, εφαρμόζοντας μια άπληστη αναζήτηση από πάνω

προς τα κάτω μέσα στο χώρο των πιθανών κλάδων. Ο ID3 [87] χρησιμοποιεί την ιδέα του κέρδους πληροφορίας (Information Gain), που βασίζεται στην εντροπία (Entropy) για την κατασκευή ενός δέντρου αποφάσεων. Τον όρο εντροπία εισήγαγε στην Θεωρία Πληροφορίας πρώτα ο Claude Shannon το 1948 και γι' αυτό συχνά αναφέρεται και ως εντροπία Shannon [88]. Εντροπία μιας τυχαίας μεταβλητής στη θεωρία πληροφορίας είναι η μέση αβεβαιότητα ή έκπληξη που ενυπάρχει στις πιθανές τιμές που αυτή έχει. Εάν *T* ένα χαρακτηριστικό ενός συνόλου δεδομένων, τότε η εντροπία του υπολογίζεται ως :

$$Entropy(T) = E(T) = \sum_{i=1}^{n} p(t_i) log_2 p(t_i) \quad (15),$$

όπου n το πλήθος των κλάσεων ταξινόμησης και $p(t_i)$ η συχνότητα εμφάνισης της κλάσης i στο σύνολο των στιγμιοτύπων.

Το κέρδος πληροφορίας βασίζεται στη μείωση της εντροπίας μετά τον διαχωρισμό ενός συνόλου δεδομένων βάσει ενός χαρακτηριστικού. Η κατασκευή ενός δέντρου απόφασης, δηλαδή η δημιουργία ενός νέου κλάδου, έχει να κάνει με την εύρεση του χαρακτηριστικού που επιστρέφει το υψηλότερο κέρδος πληροφορίας. Ένας κόμβος με εντροπία 0 είναι κόμβος φύλλο. Ένας κόμβος με εντροπία μεγαλύτερη από 0 χρειάζεται περαιτέρω διαχωρισμό. Γενικά ισχύει:

$$\text{Information Gain}(\mathrm{T}, \alpha) = \text{Εντροπία γονικού κόμβου T} - \text{Συνολική Εντροπία παιδιών του T} \quad (16)$$

Πιο συγκεκριμένα:

$$\text{Information Gain}(\mathrm{T}, \alpha) = \mathrm{IG}(\mathrm{T}, \alpha) = E(T) - \sum_{\mathrm{i}=1}^{|\alpha|} \frac{|\alpha_\mathrm{i}|}{|\mathrm{T}|} \mathrm{E}(\alpha_i) \quad (17)$$

όπου T ο γονικός κόμβος-χαρακτηριστικό, α το σύνολο των διαφορετικών τιμών του T, $|\alpha|$ το πλήθος του συνόλου α, $|\mathrm{T}|$ το πλήθος του συνόλου των στιγμιοτύπων και $|\alpha_\mathrm{i}|$ το πλήθος των στιγμιοτύπων για τα οποία $\mathrm{T} = \alpha_\mathrm{i}$.

Το κύριο μειονέκτημα του ID3 είναι ότι τείνει να επιλέγει ως ρίζα/επόμενο κόμβο το χαρακτηριστικό που έχει τις περισσότερες μοναδικές τιμές. Αυτό αντιμετωπίστηκε στον αλγόριθμο C4.5, η βασική ιδέα του οποίου είναι η χρησιμοποίηση του λόγου κέρδους πληροφορίας (Information Gain Ratio) αντί του κέρδους πληροφορίας. Συγκεκριμένα απλώς προσθέτει μια ποινή στο κέρδος πληροφορίας διαιρώντας με την εντροπία του γονικού κόμβου:

$$Information\ Gain\ Ratio = \frac{\text{Information Gain}(\mathrm{T}, \alpha)}{E(T)} \quad (18)$$

- Ο CART (Classification And Regression Tree) από τον Leo Breiman και τους συνεργάτες του το 1984 [89], ο οποίος χρησιμοποιείται τόσο για ταξινόμηση όσο και για παλινδρόμηση. Χρησιμοποιεί μόνο δυαδικά δένδρα, που έχουν δύο κλάδους σε κάθε κόμβο. Σε προβλήματα παλινδρόμησης χρησιμοποιεί ως κριτήριο χωρισμού ενός κόμβου τη μείωση της διασποράς (variance decrease), ενώ σε προβλήματα ταξινόμησης τη μη καθαρότητα Gini (Gini impurity), που οφείλει το όνομά της στον

ιταλό μαθηματικό Corrado Gini [90], [91]. Η μη καθαρότητα Gini μετρά την πιθανότητα εσφαλμένης ταξινόμησης ενός τυχαία επιλεγμένου στοιχείου στο σύνολο [92]. Όσο μεγαλύτερη είναι η τιμή Gini σε έναν κόμβο, τόσο μεγαλύτερη μη καθαρότητα, δηλαδή αβεβαιότητα, σημαίνει ότι έχει. Φτάνει στο ελάχιστο του (μηδέν) όταν όλες οι περιπτώσεις στον κόμβο εμπίπτουν σε μία κατηγορία. Δίνεται από τον τύπο :

$$GINI(t) = 1 - \sum_{i=0}^{j} [p^2(C_i|t)] \quad (19)$$

όπου το j αντιπροσωπεύει τον αριθμό των κλάσεων της ταξινόμησης και το $p(C_i|t)$ την αναλογία της κλάσης C_i στην τιμή t του κόμβου-χαρακτηριστικού Τ.

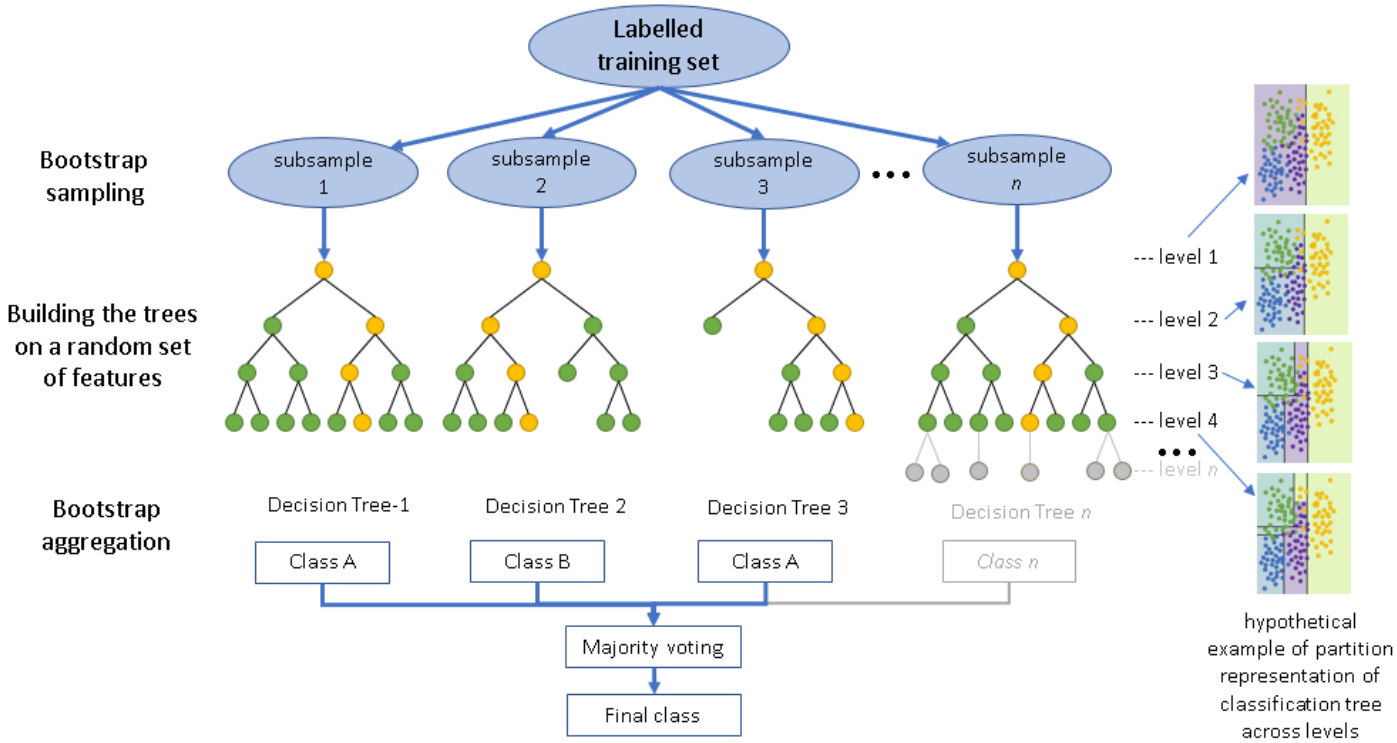
Ο αλγόριθμος CART επιλέγει την ρίζα του δένδρου ή και οποιοδήποτε εσωτερικό κόμβο, χρησιμοποιώντας τον ακόλουθο τύπο:

$$GINI(T) = \frac{|t_{left}|}{|T|} Gini(t_{left}) + \frac{|t_{right}|}{|T|} Gini(t_{right}) \quad (20)$$

όπου $|t_{left}|$,| $|t_{right}|$ είναι το πλήθος των στιγμιοτύπων των κόμβων στα αριστερά και δεξιά αντίστοιχα και |T| είναι το μέγεθος δείγματος του γονικού κόμβου Τ.

Ο αλγόριθμος του δέντρου αποφάσεων είναι σχετικά εύκολο να ερμηνευθεί. Ωστόσο, μπορεί να υποφέρει από υπερεκπαίδευση (overfitting – υπερβολική προσαρμογή στο training set βλ. κεφάλαιο 5.6) εάν το δέντρο γίνει πολύ περίπλοκο ή τα δεδομένα εκπαίδευσης είναι θορυβώδη. Τεχνικές όπως το κλάδεμα (pruning - αφαίρεση κόμβων που συμβάλλουν ελάχιστα στη συνολική ακρίβεια ή προγνωστική δύναμη του δέντρου), ο καθορισμός ορίων βάθους ή η χρήση συνδυαστικών μεθόδων (ensemble methods), όπως ο Random Forest, μπορούν να βοηθήσουν στον μετριασμό της υπερεκπαίδευσης και στη βελτίωση της γενίκευσης [86].

5.3.6 Τυχαίο Δάσος (Random Forest)

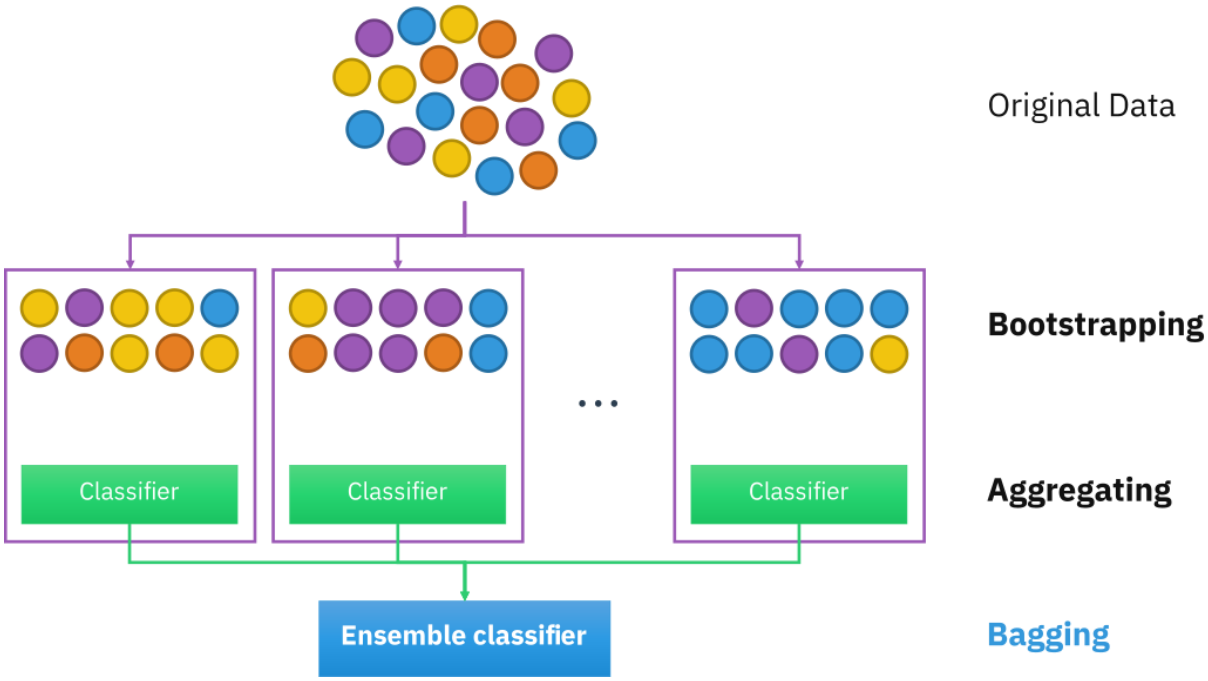


Εικόνα 8: *Απλοποιημένη σχηματοποίηση του Random Forest* [93]

Ο αλγόριθμος Random Forest (Εικόνα 8) χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης. Ανήκει στην κατηγορία των συνδυαστικών μεθόδων μηχανικής μάθησης, οι οποίες συνδυάζουν τις προβλέψεις πολλών μεμονωμένων μοντέλων για να κάνουν μια τελική πρόβλεψη. Αποτελείται από πολλά δέντρα απόφασης. Ο πρώτος αλγόριθμος για δάση τυχαίας απόφασης δημιουργήθηκε το 1995 από την Tin Kam Ho [94]. Μια επέκταση αυτού του αλγορίθμου, αναπτύχθηκε από τον Leo Breiman το 2001 [95].

Ο Random Forest προσπαθεί να βελτιώσει τη γενίκευση κατασκευάζοντας ένα σύνολο δέντρων απόφασης που δεν παρουσιάζουν οιαδήποτε συσχέτιση μεταξύ τους. Τα τυχαία δάση βασίζονται στην ιδέα του bagging για να χρησιμοποιήσουν ένα διαφορετικό δείγμα bootstrap των δεδομένων εκπαίδευσης για την εκμάθηση δέντρων αποφάσεων. Το Bagging, το οποίο είναι επίσης γνωστό ως bootstrap aggregating (Εικόνα 9), είναι μια τεχνική που αρχικά λαμβάνει επανειλημμένα δείγματα (με αντικατάσταση) από ένα σύνολο δεδομένων σύμφωνα με μια ομοιόμορφη κατανομή πιθανοτήτων. Επειδή η δειγματοληψία γίνεται με αντικατάσταση, ορισμένα δείγματα μπορεί να εμφανιστούν πολλές φορές στο ίδιο σετ εκπαίδευσης, ενώ άλλα μπορεί να παραλείπονται τελείως. Ύστερα κάθε δείγμα bootstrap, που έχει το ίδιο μέγεθος με τα αρχικά δεδομένα, χρησιμοποιείται για εκπαίδευση ενός ταξινομητή. Το σύνολο των επιμέρους ταξινομητών δημιουργεί έναν ensemble ταξινομητή και κάθε πρόβλεψη για νέα άγνωστα δεδομένα γίνεται με εύρεση του μέσου όρου των προβλέψεων των επιμέρους ταξινομητών για regression, ή με ψηφοφορία για classification. Ωστόσο, ένα βασικό χαρακτηριστικό διάκρισης του τυχαίου δάσους από την μέθοδο bagging είναι ότι σε κάθε δέντρο, το καλύτερο κριτήριο διαχωρισμού επιλέγεται μεταξύ ενός μικρού συνόλου τυχαία επιλεγμένων χαρακτηριστικών, εξαλείφοντας οποιαδήποτε συσχέτιση που θα μπορούσε να υπάρχει μεταξύ τους. Με αυτόν τον τρόπο, κατασκευάζεται ένα σύνολο

δέντρων απόφασης όχι μόνο με χειρισμό των δεδομένων εκπαίδευσης (χρησιμοποιώντας δείγματα bootstrap παρόμοια με το bagging), αλλά και των χαρακτηριστικών (χρησιμοποιώντας διαφορετικά υποσύνολα χαρακτηριστικών σε κάθε δένδρο). Όπως και στο bagging, μόλις κατασκευαστεί ένα σύνολο δέντρων απόφασης, η πρόβλεψη για νέα δεδομένα σε ταξινόμηση γίνεται με ψηφοφορία, ενώ σε παλινδρόμηση με το μέσο όρο των επιμέρους προβλέψεων των δέντρων απόφασης [74].



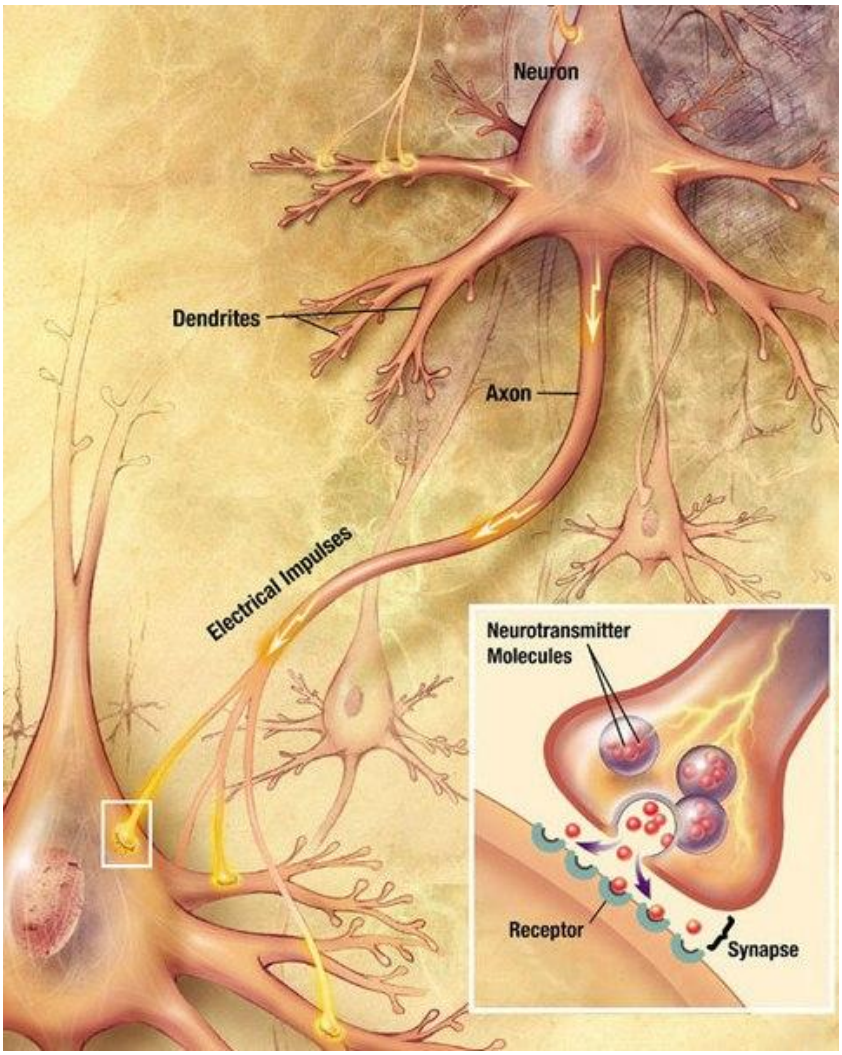
*Εικόνα 9: Μέθοδος Bagging (**B**ootstrap **agg**regat**ing**)* [96]

Κατά την υλοποίηση του Random Forest, κάθε δέντρο απόφασης κατασκευάζεται εφαρμόζοντας έναν αλγόριθμο στο εκάστοτε training set, όπως πχ C4.5. Στην αυθεντική μέθοδο που προτάθηκε από τον Breiman, κατά την οποία κάθε δέντρο είναι ένα κλασσικό δέντρο CART, ως κριτήριο χωρισμού χρησιμοποιείται η μη καθαρότητα GINI, σύμφωνα με την οποία επιλέγεται κάθε φορά το σημαντικότερο χαρακτηριστικό [97].

Σημειώνεται ότι τα δέντρα απόφασης που εμπλέκονται σε ένα τυχαίο δάσος είναι δέντρα που δεν έχουν κλαδευτεί, καθώς τους επιτρέπεται να μεγαλώσουν στο μεγαλύτερο δυνατό μέγεθός τους μέχρι κάθε φύλλο να είναι καθαρό. Ως εκ τούτου, οι βασικοί ταξινομητές του τυχαίου δάσους αντιπροσωπεύουν ασταθείς ταξινομητές που έχουν χαμηλή μεροληψία (bias) αλλά και υψηλή διασπορά (variance) (βλ κεφάλαιο 5.6), που οδηγεί σε υπερεκπαίδευση, λόγω του μεγάλου μεγέθους τους.

Συγκεντρώνοντας τις προβλέψεις ενός συνόλου ισχυρών και ασυσχέτιστων δέντρων απόφασης, τα τυχαία δάση είναι σε θέση να μειώσουν τη διασπορά των δέντρων χωρίς να επηρεάσουν αρνητικά τη χαμηλή μεροληψία τους. Αυτό καθιστά τα τυχαία δάση αρκετά ανθεκτικά στην υπερεκπαίδευση. Επιπλέον, λόγω της ικανότητάς τους να λαμβάνουν υπόψη μόνο ένα μικρό υποσύνολο χαρακτηριστικών σε κάθε εσωτερικό κόμβο, τα τυχαία δάση είναι υπολογιστικά γρήγορα και ισχυρά ακόμη και σε προβλήματα πολλών διαστάσεων [74].

5.3.7 Τεχνητά Νευρωνικά Δίκτυα (ANN - Artificial Neural Networks)



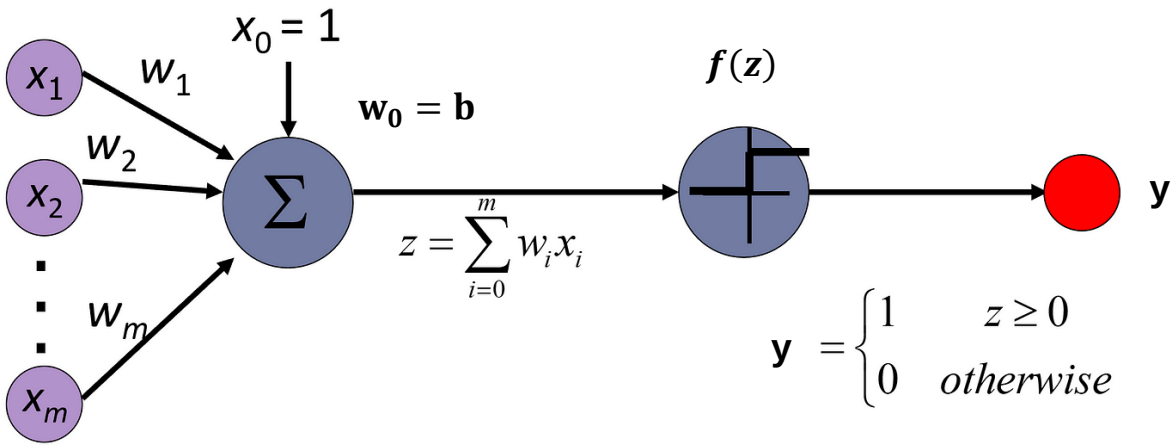
Εικόνα 10: *Αναπαράσταση νευρώνα και μιας συναπτικής σύνδεσης με έναν γειτονικό του* [98]

Τα Τεχνητά νευρωνικά δίκτυα είναι ισχυρά μοντέλα που χρησιμοποιούνται για ταξινόμηση και παλινδρόμηση, τα οποία μπορούν να μάθουν πολύπλοκα και μη γραμμικά όρια απόφασης αμιγώς από τα δεδομένα. Έχουν ευρύ φάσμα εφαρμογών σε τομείς όπως όραση, ομιλία και επεξεργασία γλώσσας. Ιστορικά, η μελέτη των ANN εμπνεύστηκε από προσπάθειες για να προσομοιωθούν βιολογικά νευρωνικά συστήματα. Ο ανθρώπινος εγκέφαλος αποτελείται κυρίως από νευρικά κύτταρα, τους νευρώνες, τα οποία συνδέονται μεταξύ τους μέσω ινωδών σχηματισμών που λέγονται νευράξονες (Εικόνα 10). Οποτεδήποτε ένας νευρώνας ερεθίζεται, απαντώντας σε ένα ερέθισμα, μεταδίδει την ενεργοποίηση σε άλλους νευρώνες μέσω του νευράξονα. Η συλλογή των σημάτων ή νευρικών ώσεων γίνεται μέσω αποφυάδων που ονομάζονται δενδρίτες, οι οποίοι προέρχονται από το κυτταρόπλασμα του νευρώνα. Η ισχύς του σημείου επαφής μεταξύ δενδρίτη και νευράξονα, γνωστή και ως σύναψη, καθορίζει τη συνδεσιμότητα μεταξύ των νευρώνων. Έχει ανακαλυφθεί ότι ο ανθρώπινος εγκέφαλος μαθαίνει αλλάζοντας την ισχύ αυτής της συναπτικής σύνδεσης μεταξύ των νευρώνων κατόπιν επανάληψης του ερεθισμού από την ίδια νευρική ώση. Ο ανθρώπινος εγκέφαλος αποτελείται περίπου από 100 δισεκατομμύρια νευρώνες που συνδέονται μεταξύ τους με πολύπλοκους τρόπους,

καθιστώντας δυνατή την μάθηση νέων και τακτικών δραστηριοτήτων. Σε αυτήν την ιδέα βασίζεται η κατασκευή των τεχνητών νευρωνικών δικτύων [74].

Ανάλογα με τη δομή του ανθρώπινου εγκεφάλου, ένα τεχνητό νευρωνικό δίκτυο αποτελείται από έναν αριθμό μονάδων επεξεργασίας, τους κόμβους, που συνδέονται μεταξύ τους μέσω κατευθυνόμενων συνδέσμων. Οι κόμβοι αντιστοιχούν στους νευρώνες που εκτελούν τις βασικές μονάδες υπολογισμού, ενώ οι κατευθυνόμενοι σύνδεσμοι αντιστοιχούν στις συνδέσεις μεταξύ νευρώνων, δηλαδή στους νευράξονες και δενδρίτες. Επιπλέον, το βάρος μιας κατευθυνόμενης σύνδεσης μεταξύ δύο νευρώνων αντιπροσωπεύει τη δύναμη της συναπτικής σύνδεσης μεταξύ των νευρώνων. Όπως και στα βιολογικά νευρωνικά συστήματα, ο πρωταρχικός στόχος των ANN είναι να προσαρμόσουν τα βάρη των συνδέσμων μέχρι να μάθουν τις σχέσεις εισόδου-εξόδου των υποκείμενων δεδομένων [74].

5.3.7.1 Το Perceptron του Rosenblatt



Εικόνα 11: *Ισοδύναμο γράφημα ροής σήματος του Perceptron* [99] *(Προσαρμογή)*

Ο Rosenblatt το 1958 [100], βασιζόμενος στο μοντέλο ενός νευρώνα του McCulloch-Pitts [101], πρότεινε το Perceptron ως πρώτο μοντέλο μάθησης με τη συνδρομή ενός «εκπαιδευτή» (δηλαδή επιβλεπόμενη μάθηση). Το Perceptron είναι η απλούστερη δυνατή μορφή ενός νευρωνικού δικτύου, αποτελείται από ένα μόνο νευρώνα και χρησιμοποιείται μόνο για την ταξινόμηση προτύπων, τα οποία είναι γραμμικά διαχωρίσιμα. Περιορίζεται στην ταξινόμηση προτύπων που ανήκουν σε δύο μόνο κλάσεις [67].

Ένα perceptron περιλαμβάνει δύο τύπους κόμβων: τους κόμβους εισόδου, που χρησιμοποιούνται για την είσοδο των χαρακτηριστικών και έναν κόμβο εξόδου, που χρησιμοποιείται για την δυαδική έξοδο του μοντέλου. Στην Εικόνα 11 απεικονίζεται η βασική αρχιτεκτονική ενός perceptron που παίρνει n χαρακτηριστικά εισόδου, x_1, x_2,…, x_n, και παράγει μια δυαδική έξοδο y. Ο κόμβος εισόδου που αντιστοιχεί σε ένα χαρακτηριστικό x_i συνδέεται μέσω ενός συνδέσμου με ρυθμιζόμενο βάρος w_i στον κόμβο εξόδου. Ο σύνδεσμος αυτός με το βάρος του χρησιμοποιείται για να προσομοιάσει την ισχύ μιας συναπτικής σύνδεσης ανάμεσα σε νευρώνες. Ο κόμβος εξόδου υπολογίζει ένα σταθμισμένο άθροισμα των εισόδων z, αφού προσθέσει έναν παράγοντα πόλωσης $w_0 = b$, και στη συνέχεια παράγει την έξοδο y ως εξής:

$$y = f(z) = \begin{cases} 1, & \text{εάν } z > 0 \\ -1, & \text{αλλιώς} \end{cases} \quad (21)$$

$$\text{όπου } z = \sum_{i=0}^{m} w_i x_i = \sum_{i=1}^{m} w_i x_i + b \quad (22)$$

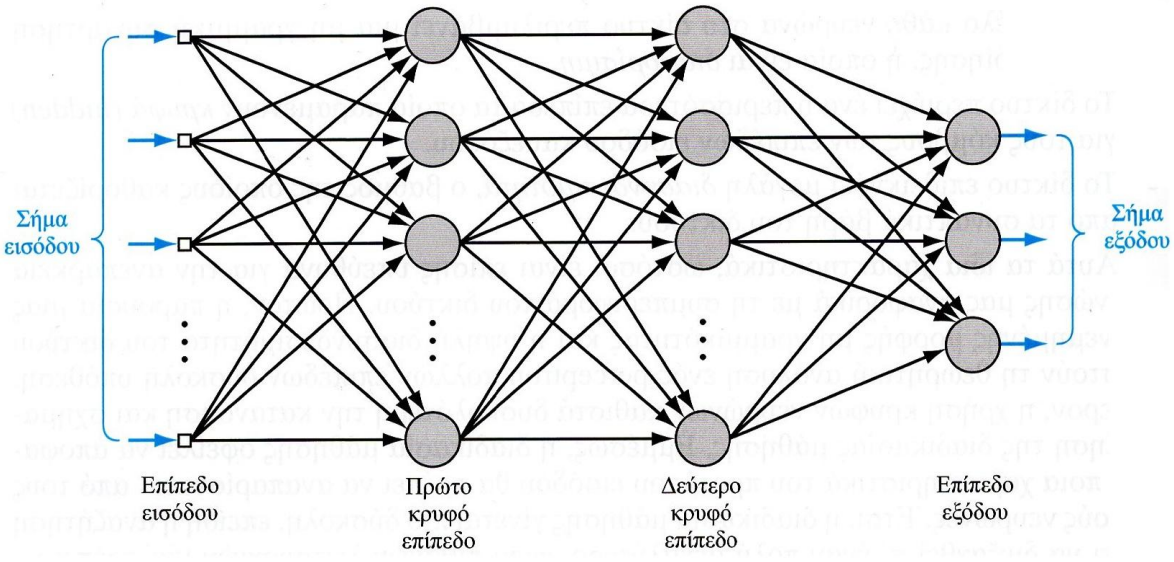
και f η συνάρτηση προσήμου sign(x) που λειτουργεί ως συνάρτηση ενεργοποίησης (activation function).

Κατά την εκπαίδευση του Perceptron, είναι σημαντικός ο καθορισμός των παραμέτρων w_i (βαρών) ώστε η έξοδος y να πλησιάσει την πραγματική έξοδο d για κάθε δείγμα εκπαίδευσης. Αυτό επιτυγχάνεται χρησιμοποιώντας τον τύπο επαναληπτικής ενημέρωσης των βαρών για κάθε δείγμα εισόδου:

$$w_i\,(k+1) = w_i\,(k) + \lambda\big(d - y(k)\big)x_i \quad (23)$$

όπου $w_i\,(k)$ είναι το βάρος που σχετίζεται με το σύνδεσμο i μετά την κ επανάληψη, λ είναι μια παράμετρος που ονομάζεται παράμετρος μάθησης (learning rate) και x_i είναι η τιμή του χαρακτηριστικού i του δείγματος εισόδου. Η παράμετρος μάθησης παίρνει τιμές από 0 έως 1 και ελέγχει το βαθμό προσαρμογής σε κάθε επανάληψη. Μεγάλη τιμή λ οδηγεί σε γρήγορη προσαρμογή. Ο αλγόριθμος σταματά όταν ο μέσος αριθμός των διαφορών γίνει μικρότερος από ένα όριο γ (threshold) [74].

5.3.7.2 Πολυεπίπεδο Perceptron (MLP - Multilayer Perceptron)



Εικόνα 12: *Αρχιτεκτονικός γράφος ενός perceptron πολλών επιπέδων με δύο κρυφά επίπεδα* [67]

Όπως αναφέρθηκε, οι δυνατότητες του perceptron περιορίζονται στην ταξινόμηση γραμμικά διαχωρίσιμων προτύπων. Για να ξεπεραστούν οι πρακτικοί περιορισμοί του perceptron, δημιουργήθηκε μια άλλη δομή νευρωνικού δικτύου το perceptron πολλαπλών

επιπέδων, που είναι ικανό και για εκμάθηση μη γραμμικών ορίων απόφασης. Το πρώτο MLP δημοσιεύτηκε από τους Ivakhnenko και Lapa το 1965 [102], αναπτύσσοντας μια μέθοδο επαγωγικής στατιστικής μάθησης γνωστή ως GMDH (Group Method of Data Handling), γι' αυτό ο Ivakhnenko συχνά αναφέρεται και ως «πατέρας της βαθιάς μάθησης».

Η γενική αρχιτεκτονική του MLP (Εικόνα 12) αποτελείται από επίπεδα που οργανώνονται σε μορφή αλυσίδας, ώστε κάθε επίπεδο να λειτουργεί ως έξοδος του προηγούμενου επιπέδου. Το δίκτυο επιδεικνύει μεγάλη διασυνδεσιμότητα, ο βαθμός της οποίας καθορίζεται από τα συναπτικά βάρη του δικτύου. Το πρώτο στρώμα ονομάζεται επίπεδο εισόδου (input layer) και αναπαριστά την εισαγωγή των χαρακτηριστικών στο μοντέλο. Οι είσοδοι αυτοί τροφοδοτούν ενδιάμεσα στρώματα που ονομάζονται κρυφά επίπεδα (hidden layers), που αποτελούνται από επεξεργαστικές μονάδες, τους κρυφούς κόμβους. Κάθε κρυφός κόμβος λειτουργεί με σήματα που λαμβάνονται από τους κόμβους εισόδου ή άλλους κρυφούς κόμβους προηγούμενου επιπέδου και παράγει μια τιμή ενεργοποίησης που μεταδίδεται στο επόμενο επίπεδο. Γενικά όσο μεγαλύτερος είναι ο αριθμός των κρυφών επιπέδων, τόσο πιο βαθιά είναι η ιεραρχία των χαρακτηριστικών που μαθαίνει το δίκτυο. Τέτοια μοντέλα ANN με μακριές αλυσίδες κρυφών επιπέδων είναι γνωστά ως βαθιά νευρωνικά δίκτυα (deep neural networks). Το τελικό επίπεδο, που λέγεται επίπεδο εξόδου (output layer) επεξεργάζεται τις τιμές ενεργοποίησης από το προηγούμενο επίπεδο για να παράξει προβλέψεις των μεταβλητών εξόδου. Για δυαδική ταξινόμηση, το επίπεδο εξόδου περιέχει έναν μόνο κόμβο που αντιπροσωπεύει την ετικέτα δυαδικής κλάσης. Σε αυτή την αρχιτεκτονική, δεδομένου ότι τα σήματα διαδίδονται μόνο προς τα εμπρός από το στρώμα εισόδου στο επίπεδο εξόδου, ονομάζονται επίσης δίκτυα τροφοδότησης προς τα εμπρός (Feedforward Neural Networks - FNN) [74].

Κάθε νευρώνας στο δίκτυο περιλαμβάνει μια μη γραμμική συνάρτηση ενεργοποίησης, η οποία είναι διαφορίσιμη. Οι πιο συνηθέστερα χρησιμοποιούμενες είναι οι σιγμοειδείς (sigmoid) συναρτήσεις σ(x) και συγκεκριμένα η λογιστική (logistic) και η υπερβολική εφαπτομένη (hyperbolic tangent):

$$\text{Λογιστική συνάρτηση: } \sigma(\mathrm{x}) = \frac{1}{1+\mathrm{e}^{-\mathrm{x}}} \quad (24)$$

$$\text{Συνάρτηση υπερβολικής εφαπτομένης: } \sigma(\mathrm{x}) = \tanh(\mathrm{x}) \quad (25)$$

Έστω $w_{ij}(l)$ το βάρος του συνδέσμου από τον κόμβο j του επιπέδου (l-1) στον κόμβο i του επιπέδου l και $b_i(l)$ ο όρος bias σε αυτόν τον κόμβο. Η τιμή ενεργοποίησης $a_i(l)$ είναι:

$$a_i(l) = f\big(z_i(l)\big) = f\left(\sum_j w_{ij}(l) a_j(l-1) + b_i(l)\right) \quad (26)$$

όπου f(.) η συνάρτηση ενεργοποίησης που μετατρέπει το z στο α. Επίσης, βάσει ορισμού, $a_j(0) = x_j$ στο επίπεδο εισόδου και $a(L) = y$ στον κόμβο εξόδου.

Τα βάρη και ο όρος bias (w,b) του μοντέλου ANN μαθαίνονται κατά τη διάρκεια της εκπαίδευσης έτσι ώστε οι προβλέψεις σε στιγμιότυπα εκπαίδευσης να ταιριάζουν με τις πραγματικές ετικέτες. Αυτό επιτυγχάνεται με τη χρήση μιας συνάρτησης σφάλματος (loss function):

$$E(w,b) = \sum_{k=1}^{n} Loss(d_k, y_k) \quad (27)$$

όπου d_k είναι η αληθινή ετικέτα του k - στιγμιότυπου εκπαίδευσης και y_k είναι ίσο με $a(L)$, που παράγεται χρησιμοποιώντας το x_k . Μια τυπική επιλογή της συνάρτησης σφάλματος είναι η συνάρτηση τετραγωνικού σφάλματος:

$$Loss(d_k, y_k) = (d_k - y_k)^2 \quad (28)$$

Η $E(w, b)$ είναι συνάρτηση των παραμέτρων του μοντέλου (w, b) επειδή η τιμή ενεργοποίησης εξόδου $\sigma_\kappa(z)$ εξαρτάται από τα βάρη και τους όρους πόλωσης. Σκοπός είναι η επιλογή των (w, b) τα οποία ελαχιστοποιούν το σφάλμα εκπαίδευσης $E(w, b)$. Δυστυχώς, λόγω της χρήσης κρυφών κόμβων με μη γραμμικές συναρτήσεις ενεργοποίησης, το E(w, b) δεν είναι μια κυρτή συνάρτηση των w και b, πράγμα που σημαίνει ότι το E(w, b) μπορεί να έχει τοπικά ελάχιστα που δεν είναι συνολικά βέλτιστα. Ωστόσο, μπορούμε να εφαρμόσουμε τεχνικές βελτιστοποίησης, όπως η μέθοδος gradient descent [103], [104] για να καταλήξουμε σε μια τοπικά βέλτιστη λύση. Συγκεκριμένα, οι παράμετροι $w_{ij}(l) \kappa \alpha \iota\, b_i(l)$ ενημερώνονται επαναληπτικά σύμφωνα με τις παρακάτω εξισώσεις:

$$w_{ij}(l) \leftarrow w_{ij}(l) - \lambda \frac{\partial E}{\partial w_{ij}(l)} \quad (29)$$

$$b_i(l) \leftarrow b_i(l) - \lambda \frac{\partial E}{\partial b_i(l)} \quad (30)$$

όπου λ η υπερπαράμετρος μάθησης (learning rate). Διαισθητικά, σύμφωνα με αυτή, τα βάρη κινούνται προς την κατεύθυνση μείωσης του σφάλματος εκπαίδευσης. Εάν προσεγγιστεί κάποιο ελάχιστο κατά τη διάρκεια αυτής της μεθόδου, η κλίση του σφάλματος εκπαίδευσης θα είναι κοντά στο 0 εξαλείφοντας τον δεύτερο όρο, έχοντας ως αποτέλεσμα την σύγκλιση των βαρών [74].

Για να γίνει η ενημέρωση των βαρών στις σχέσεις (29) και (30), πρέπει να γίνει ο υπολογισμός της μερικής παραγώγου του E σε σχέση με το $\partial w_{ij}(l)$. Αυτός ο υπολογισμός είναι μη τετριμμένος ειδικά σε κρυφά επίπεδα (l < L), αφού το $w_{ij}(l)$ δεν επηρεάζει άμεσα το y = a(L) (και επομένως το σφάλμα εκπαίδευσης), αλλά έχει πολύπλοκες αλυσίδες επιρροών μέσω των τιμών ενεργοποίησης στα επόμενα επίπεδα. Για την αντιμετώπιση αυτού του προβλήματος, αναπτύχθηκε μια τεχνική γνωστή ως οπισθοδιάδοση (backpropagation) [105], η οποία διαδίδει τις παραγώγους προς τα πίσω, από το επίπεδο εξόδου προς τα κρυφά επίπεδα. Με αυτόν τον τρόπο, τα βάρη προσαρμόζονται ανάστροφα για τα στρώματα L-1 έως 1, ανάλογα με το πόσο συνεισέφεραν στο συνολικό σφάλμα του δικτύου (credit assignment). Ο πίνακας των βαρών επιστρέφεται όταν κατά την εκτέλεση του αλγόριθμου της backpropagation το σφάλμα εκπαίδευσης είναι μικρότερο από μια μικρή προκαθορισμένη τιμή ε. Εάν έχει οριστεί μέγιστος αριθμός εποχών (επαναλήψεων) και αυτό δεν έχει επιτευχθεί, επιστρέφεται σφάλμα καθώς δεν βρέθηκε λύση.

Το MLP μπορεί να χρησιμοποιηθεί σε προβλήματα τόσο ταξινόμησης πολλών κλάσεων όσο και παλινδρόμησης. Όπως αναφέρθηκε μπορούν να αξιοποιηθούν για την εκμάθηση πολύπλοκων ορίων απόφασης σε μια μεγάλη γκάμα εφαρμογών. Όμως η μεγάλη πολυπλοκότητα αυτών των κλασικών ANN μοντέλων τα καθιστά ευάλωτα σε υπερεκπαίδευση, πρόβλημα που αντιμετωπίζεται με τεχνικές βαθειάς μάθησης, όπως η μέθοδος dropout, κατά την οποία κατά τη διάρκεια της εκπαίδευσης αφαιρούνται τυχαία κόμβοι εισόδου και κρυφοί κόμβοι. Ένα πλεονέκτημα των MLP είναι ότι μπορούν να διαχειριστούν μη σχετικά ή περιττά χαρακτηριστικά που δεν βοηθούν στη μείωση του

σφάλματος εκπαίδευσης, χρησιμοποιώντας μηδενικά βάρη. Βέβαια, εάν υπάρχει μεγάλος αριθμός τέτοιων χαρακτηριστικών το ANN μοντέλο μπορεί να υποφέρει από υπερεκπαίδευση, υστερώντας σημαντικά στην γενίκευση. Σημειώνεται ότι η εκπαίδευση ενός MLP μπορεί να απαιτεί αρκετό χρόνο ειδικά όταν το πλήθος των κρυφών κόμβων είναι μεγάλος [74].

5.4 Μετρικές αξιολόγησης απόδοσης μοντέλων μηχανικής μάθησης

Η απόδοση ενός μοντέλου ταξινόμησης μπορεί να αξιολογηθεί συγκρίνοντας τις προβλεπόμενες ετικέτες με τις πραγματικές ετικέτες των περιπτώσεων. Αυτές οι πληροφορίες μπορούν να συνοψιστούν σε έναν πίνακα που ονομάζεται πίνακας σύγχυσης (confusion matrix). Ο Πίνακας 3 απεικονίζει τον πίνακα σύγχυσης για ένα πρόβλημα δυαδικής ταξινόμησης [74].

Πίνακας 3: Πίνακας σύγχυσης για ένα πρόβλημα δυαδικής ταξινόμησης.

		Προβλεπόμενη κλάση	
		Class 0	**Class 1**
Πραγματική κλάση	**Class 0**	**TN**	**FP**
	Class 1	**FN**	**TP**

Οι τέσσερις σχετικές εγγραφές που περιλαμβάνονται σε έναν πίνακα σύγχυσης για έναν δυαδικό ταξινομητή είναι:

- **TN (True Negative)** - αληθώς αρνητικά, που δηλώνει τον αριθμό των σωστά ταξινομημένων αρνητικών δειγμάτων.
- **TP (True Positive)** – αληθώς θετικά, που δηλώνει τον αριθμό των σωστά ταξινομημένων θετικών δειγμάτων.
- **FN (False Negative)** - ψευδώς αρνητικά, που δηλώνει τον αριθμό των θετικών δειγμάτων που ταξινομήθηκαν εσφαλμένα ως αρνητικά.
- **FP (False Positive)** - ψευδώς θετικά, που δηλώνει τον αριθμό των αρνητικών δειγμάτων που ταξινομήθηκαν εσφαλμένα ως θετικά.

Αν και ένας πίνακας σύγχυσης παρέχει τις πληροφορίες που απαιτούνται για τον προσδιορισμό του πόσο καλά αποδίδει ένα μοντέλο ταξινόμησης, συνοψίζοντας αυτές τις πληροφορίες σε έναν μόνο αριθμό καθιστά ευκολότερη τη σύγκριση της σχετικής απόδοσης διαφορετικών μοντέλων. Αυτό μπορεί να γίνει χρησιμοποιώντας τις ακόλουθες μετρικές αξιολόγησης, οι οποίες παίρνουν τιμές μεταξύ [0,1] [106]. (Οι σχέσεις εδώ αφορούν προβλήματα δυαδικής ταξινόμησης).

- **Ορθότητα (Accuracy)**. Η ορθότητα είναι το ποσοστό των σωστά ταξινομημένων δειγμάτων στο συνολικό αριθμό των δειγμάτων του test set. Αυτή η μετρική είναι από τις πιο συχνά χρησιμοποιούμενες σε εφαρμογές της ML στην ιατρική, αλλά είναι επίσης γνωστό ότι είναι παραπλανητική στην περίπτωση dataset με μη ισορροπημένες κλάσεις, αφού απλώς καταχωρώντας όλα τα δείγματα στην επικρατούσα κλάση είναι ένας εύκολος τρόπος για την επίτευξη υψηλής ορθότητας:

$$\text{Accuracy} = \frac{\#\ \text{σωστά ταξινομημένων δειγμάτων}}{\#\ \text{όλων των δειγμάτων}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (31)$$

- **Ευαισθησία (sensitivity).** Η ευαισθησία, επίσης γνωστή ως ανάκληση (recall) ή ποσοστό αληθώς θετικών (True Positive Rate - TPR), δηλώνει το ποσοστό θετικών δειγμάτων που έχουν ταξινομηθεί σωστά στο σύνολο των θετικών δειγμάτων στο test set. Αυτή η μετρική θεωρείται ως μια από τις πιο σημαντικές για ιατρικές μελέτες, καθώς είναι επιθυμητό να χάνονται όσο το δυνατόν λιγότερα θετικά περιστατικά, που μεταφράζεται σε υψηλή ευαισθησία:

$$\text{Sensitivity} = \frac{\#\text{αληθώς θετικών δειγμάτων}}{\#\ \text{όλων των θετικών δειγμάτων}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (32)$$

- **Ειδικότητα (Specificity).** Η ειδικότητα, ή το ποσοστό των αληθώς αρνητικών (True Negative Rate), είναι η αντίστοιχη μετρική της ευαισθησίας για την αρνητική όμως κλάση και δηλώνει το ποσοστό των αρνητικών δειγμάτων που έχουν ταξινομηθεί σωστά:

$$\text{Specificity} = \frac{\#\ \text{αληθώς αρνητικών δειγμάτων}}{\#\ \text{όλων των αρνητικών δειγμάτων}} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (33)$$

- **Recall macro.** Η μετρική recall macro προκύπτει από τη μέση τιμή της ευαισθησίας και της ειδικότητας:

$$Recall_{macro} = \frac{Sensitivity + Specificity}{2} \quad (34)$$

  Σε περίπτωση μη ισορροπημένου dataset, εάν μία κλάση είναι σπάνια, αλλά πάρα πολύ σημαντική, η μετρική recall macro είναι καλύτερη επειδή αντιμετωπίζει ισότιμα κάθε κλάση.

- **Ακρίβεια (Precision).** Η ακρίβεια για κάθε κλάση δηλώνει το ποσοστό των δειγμάτων που ταξινομήθηκαν σωστά στο σύνολο των δειγμάτων που ταξινομήθηκαν σε κάθε κατηγορία. Για την ακρίβεια της θετικής κλάσης, που ονομάζεται Positive Predictive Value ισχύει:

$$\text{Positive Predicted Value} = \frac{\#\ \text{αληθώς θετικών δειγμάτων}}{\#\ \text{πλήθος θετικά ταξινομημένων δειγμάτων}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (35)$$

  Αντίστοιχα για την αρνητική κλάση, που ονομάζεται Negative Predicted Value ισχύει:

$$\text{Negative Predicted Value} = \frac{\#\ \text{αληθώς αρνητικών δειγμάτων}}{\#\ \text{αρνητικά ταξινομημένων δειγμάτων}} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (36)$$

- **F1 Score**: Το F1 score είναι ο αρμονικός μέσος των precision και recall, το οποίο σημαίνει ότι «τιμωρεί» τις ακραίες τιμές αυτών.

$$\text{F1} = 2\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (37)$$

- **Καμπύλη ROC και AUC(ROC) score**: Η καμπύλη πιθανότητας ROC (Receiver Operating Characteristic) είναι η γραφική παράσταση της sensitivity (TPR) στον άξονα y σε σχέση με το 1-specificity (False Positive Rate ή FPR) στον άξονα x. Η περιοχή κάτω από την καμπύλη λέγεται AUC(ROC) score (Area Under Curve(ROC)) και χρησιμοποιείται ως μετρική απόδοσης σε προβλήματα ταξινόμησης. Η καμπύλη ROC εκτιμά την ικανότητα του κάθε μοντέλου να ξεχωρίζει τις κλάσεις μεταξύ τους. Όσο υψηλότερη είναι η AUC, τόσο καλύτερο είναι το μοντέλο στην πρόβλεψη. Αυτή η μετρική έχει χρησιμοποιηθεί εκτενώς στην ιατρική για την αξιολόγηση της διαγνωστικής ικανότητας των μοντέλων στην διάκριση των ασθενών από τα υγιή άτομα [107], [108]. Με βάση ένα πρόχειρο σύστημα ταξινόμησης, η AUC μπορεί να ερμηνευθεί ως εξής: 0.90 -1 = εξαιρετικό, 0.80 – 0.90 = καλό, 0.70 – 0.80 = μέτριο, 0.60 – 0.70= φτωχό, 0.50 – 0.60 = αποτυχημένο [109].

- **Καμπύλη PRC και AUC(PRC) score**: Η καμπύλη PRC (Precision- Recall Curve) είναι η γραφική παράσταση της precision (άξονας y) σε σχέση με το recall (άξονας x). Η περιοχή κάτω από την καμπύλη λέγεται AUC(PRC) score και χρησιμοποιείται ως μετρική απόδοσης σε προβλήματα ταξινόμησης. Η χρήση αυτής της μετρικής ενδείκνυται σε dataset που είναι έντονα μη ισορροπημένα στις κλάσεις τους, με το πλήθος των δειγμάτων στην αρνητική κλάση να ξεπερνά κατά πολύ το πλήθος αυτών της θετικής κλάσης. Αυτό συμβαίνει διότι τα διαγράμματα PRC αξιολογούν το ποσοστό των αληθώς θετικών μεταξύ των θετικών προβλέψεων, το οποίο δεν κάνουν τα διαγράμματα ROC. Τέτοιες περιπτώσεις συναντώνται πολύ συχνά σε προβλήματα δυαδικής ταξινόμησης στη βιοπληροφορική [110].

5.5 Βελτιστοποίηση υπερπαραμέτρων μοντέλων

Εικόνα 13: *10-fold Cross Validation* [111]

Για να επιτευχθεί η καλύτερη δυνατή απόδοση ενός μοντέλου μηχανικής μάθησης στη γενίκευση, χρειάζεται να πραγματοποιηθεί η διαδικασία της βελτιστοποίησης των υπερπαραμέτρων του, δηλαδή παραμέτρων που πρέπει να οριστούν ρητά πριν την εκπαίδευσή του. Μια από τις συνηθέστερα χρησιμοποιούμενες μεθόδους για το σκοπό αυτό είναι η k-fold Cross Validation (Διασταυρούμενη Επικύρωση με k «πτυχές»). Στη Cross Validation (Εικόνα 13) αρχικά χωρίζουμε τυχαία το training set σε k αριθμό "πτυχών" (folds). Στη συνέχεια, για κάθε k-fold, θεωρούμε ότι τα k-1 folds είναι το training set και ότι το fold που έμεινε έξω είναι το test set. Η μετρική σφάλματος υπολογίζεται σε αυτό το fold. Η διαδικασία επαναλαμβάνεται για τα k folds για κάθε τιμή των υπερπαραμέτρων και υπολογίζεται η μέση τιμή της επιλεγμένης μετρικής απόδοσης E_i. Με αυτό τον τρόπο, αφενός επιτυγχάνεται αμεροληψία στην αξιολόγηση αφήνοντας τελείως έξω το test set και αφετέρου χρησιμοποιούνται αποτελεσματικά τα δεδομένα εκπαίδευσης, δηλαδή χρησιμοποιούνται όλα και παίρνοντας τη μέση τιμή εξαλείφονται πιθανές ανωμαλίες στα δεδομένα.

Μια καλή επιλογή του k είναι το 10, όπως έχει αποδειχθεί εμπειρικά [112]. Παρόλα αυτά εάν δουλεύουμε σε σχετικά μικρά datasets μπορεί να είναι χρήσιμη η αύξηση του αριθμού των πτυχών. Με αύξηση του k, περισσότερα δεδομένα εκπαίδευσης θα χρησιμοποιηθούν σε κάθε επανάληψη, που έχει ως αποτέλεσμα χαμηλότερη μεροληψία (bias) κατά την εκτίμηση της απόδοσης στη γενίκευση. Ωστόσο, μεγάλες τιμές του k θα αυξήσουν επίσης τον χρόνο εκτέλεσης του αλγορίθμου διασταυρούμενης επικύρωσης και θα προκαλέσει εκτίμηση με υψηλότερη διασπορά (variance), καθώς οι πτυχές εκπαίδευσης θα μοιάζουν περισσότερο μεταξύ τους. Από την άλλη πλευρά, σε μεγάλα σύνολα δεδομένων, μπορεί να επιλεγεί μικρότερη τιμή για το k, για παράδειγμα k = 5, και να προκύψει μια ακριβής εκτίμηση της μέσης απόδοσης του μοντέλου, ενώ παράλληλα μειώνεται το υπολογιστικό κόστος της επανεκπαίδευσης και αξιολόγησης του μοντέλου στις διάφορες πτυχές [113].

Σημειώνεται ότι και οι μέθοδοι προεπεξεργασίας που αναπτύσσονται στο επόμενο κεφάλαιο έχουν υπερπαραμέτρους που χρήζουν βελτιστοποίησης, επίσης με cross validation.

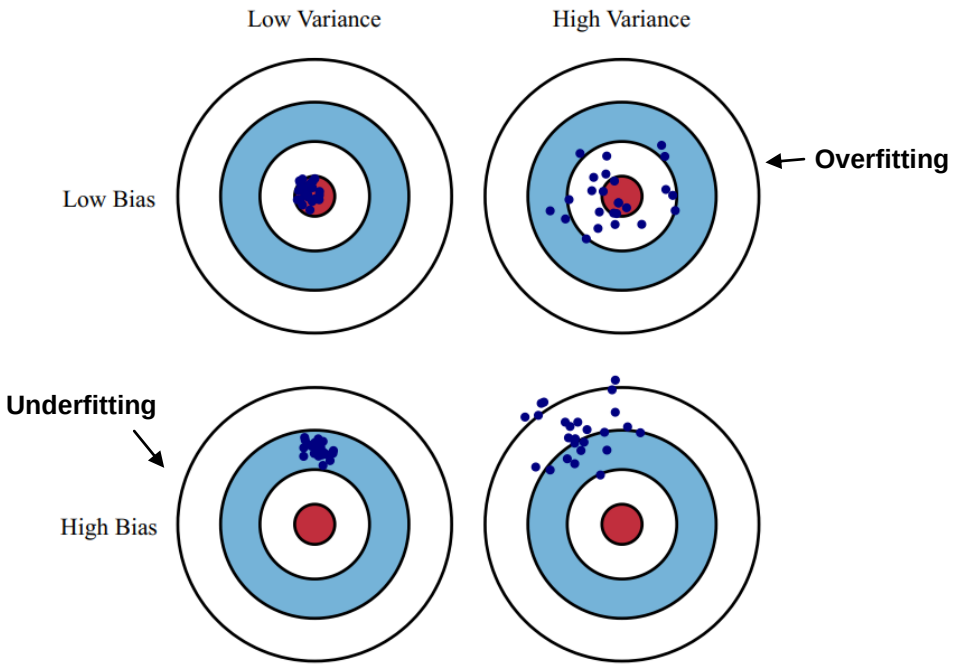
5.6 Υπερεκπαίδευση (Overfitting) και υποεκπαίδευση (Underfitting) μοντέλων

Τα σφάλματα πρόβλεψης που εμφανίζονται στα μοντέλα επιβλεπόμενης μάθησης, μπορούν να χωριστούν σε δύο βασικές κατηγορίες βάσει της αιτίας που τα προκαλεί: το σφάλμα λόγω μεροληψίας (bias error) και το σφάλμα λόγω διασποράς (variance error)[114] :

- **Bias error:** Το σφάλμα λόγω μεροληψίας ορίζεται ως η διαφορά μεταξύ της αναμενόμενης (ή μέσης) πρόβλεψης του μοντέλου και της σωστής τιμής που προσπαθεί αυτό να προβλέψει και οφείλεται στις απλουστευτικές υποθέσεις που κάνει, προκειμένου να γίνει η εκπαίδευσή του πιο εύκολη. Γενικά οι γραμμικοί αλγόριθμοι, όπως η γραμμική παλινδρόμηση, εμφανίζουν υψηλό bias error. Αυτό τους καθιστά γρήγορους στην εκπαίδευση και εύκολα κατανοήσιμους, αλλά από την άλλη πλευρά εμφανίζουν χαμηλή επίδοση σε πολύπλοκα προβλήματα.
- **Variance error**: Το σφάλμα λόγω διασποράς λαμβάνεται ως η μεταβλητότητα μιας πρόβλεψης ενός μοντέλου για ένα στιγμιότυπο με χρήση διαφορετικών δεδομένων εκπαίδευσης. Οι μη γραμμικοί αλγόριθμοι που εμφανίζουν γενικά υψηλό variance error είναι ο Decision tree, ο KNN και ο SVM.

Τα μοντέλα που παρουσιάζουν υψηλό bias και χαμηλό variance εμφανίζουν υποεκπαίδευση (underfitting), δηλαδή αποτυγχάνουν στη μοντελοποίηση των δεδομένων εκπαίδευσης με αποτέλεσμα να έχουν κακή επίδοση τόσο στα δεδομένα εκπαίδευσης όσο και σε νέα άγνωστα δεδομένα κατά τη γενίκευση. Για αυτόν το λόγο η ύπαρξη υποεκπαίδευσης είναι εύκολο να διαπιστωθεί.

Τα μοντέλα που παρουσιάζουν υψηλό variance και χαμηλό bias εμφανίζουν υπερεκπαίδευση (overfitting). Υπερεκπαίδευση υπάρχει όταν το μοντέλο μαθαίνει τις λεπτομέρειες και το θόρυβο που υπάρχει στα δεδομένα εκπαίδευσης σε τέτοια έκταση ώστε να επηρεάζεται η επίδοση του σε νέα δεδομένα, δηλαδή στη γενίκευση. Οι bias και variance μπορούν να απεικονιστούν με χρήση bull's eye diagram. Στο διάγραμμα αυτό το κέντρο του στόχου αντιστοιχεί σε ένα μοντέλο που προβλέπει τέλεια τις σωστές τιμές. Όσο απομακρυνόμαστε από το κέντρο οι προβλέψεις γίνονται ολοένα και χειρότερες. Κάθε «χτύπημα» (μπλε κουκίδα) στον «στόχο» αντιπροσωπεύει μια διαφορετική υλοποίηση του μοντέλου, δεδομένης της φυσικής μεταβλητότητας (natural variability) στα δεδομένα εκπαίδευσης που κάθε φορά συγκεντρώνονται. Κάποιες φορές τα δεδομένα εκπαίδευσης θα έχουν καλή κατανομή, οπότε η υλοποίηση αυτή του μοντέλου θα έχει καλή επίδοση και θα απεικονίζεται κοντά στο κέντρο του διαγράμματος, ενώ σε άλλες τα δεδομένα μπορεί να είναι γεμάτα ακραίες ή μη κανονικές τιμές και να οδηγούν σε χειρότερες προβλέψεις. Αυτές οι διαφορετικές υλοποιήσεις έχουν ως αποτέλεσμα διεσπαρμένα χτυπήματα στο στόχο. Στην Εικόνα 14 παρουσιάζονται τέσσερα bull's eye diagram σε τέσσερις υλοποιήσεις μοντέλων που παρουσιάζουν διαφορετικούς συνδυασμούς ψηλών ή χαμηλών τιμών bias και variance. Στο πρώτο διάγραμμα απεικονίζεται ένα μοντέλο που κάνει επιτυχημένες προβλέψεις, στο δεύτερο ένα μοντέλο που παρουσιάζει υπερεκπαίδευση, ενώ στο τρίτο ένα που παρουσιάζει υποεκπαίδευση.



Εικόνα 14: *Γραφική απεικόνιση bias και variance* [114]

Ο στόχος κάθε αλγόριθμου επιβλεπόμενης μηχανικής μάθησης είναι η επίτευξη χαμηλού bias και χαμηλού variance error, ώστε να έχει καλή επίδοση στην πρόβλεψη νέων

περιπτώσεων. Η παραμετροποίηση των αλγορίθμων είναι μια προσπάθεια εξισορρόπησης των δύο αυτών σφαλμάτων. Για παράδειγμα, ο αλγόριθμος KNN έχει χαμηλό bias και υψηλό variance. Αύξηση στην τιμή του k, δηλαδή στον αριθμό των πλησιέστερων γειτόνων που συνεισφέρουν στην πρόβλεψη, θα οδηγήσει σε μείωση της variance και σε αύξηση του bias, φτάνοντας σε ένα σημείο ισορροπίας. Υπερβολική αύξηση όμως της τιμής του k θα οδηγήσει σε μεγάλο bias, που επίσης δεν είναι επιθυμητό. Υπάρχει δηλαδή ένα ισοζύγιο: αύξηση του bias θα οδηγήσει σε μείωση του variance και η αύξηση της variance θα έχει ως αποτέλεσμα τη μείωση του bias.

Ένα εργαλείο για τον έλεγχο ύπαρξης υπερεκπαίδευσης και υποεκπαίδευσης είναι η σχεδίαση των καμπυλών εκπαίδευσης (Learning Curves) [115]. Οι καμπύλες εκπαίδευσης δείχνουν την επίδραση της προσθήκης περισσότερων δειγμάτων κατά τη διάρκεια της εκπαιδευτικής διαδικασίας. Το αποτέλεσμα απεικονίζεται ελέγχοντας τη στατιστική απόδοση του μοντέλου ως προς μια μετρική απόδοσης τόσο στην εκπαίδευση όσο και στην επικύρωση, σχεδιάζοντας δύο καμπύλες, μία για τη καθεμία. Όσο αυτές οι καμπύλες συγκλίνουν με την αύξηση των δειγμάτων εκπαίδευσης τόσο μειώνεται η ύπαρξη υπερεκπαίδευσης. Εάν οι καμπύλες με την αύξηση των δειγμάτων εκπαίδευσης αποκλίνουν, με αυτήν της εκπαίδευσης να δείχνει αύξηση της επίδοσης, ενώ της επικύρωσης μείωση, τότε έχουμε αύξηση της υπερεκπαίδευσης.

Κεφάλαιο 6: Ποιότητα δεδομένων (Data Quality)- Προεπεξεργασία (Preprocessing)

Η συλλογή των δεδομένων και η προεπεξεργασία τους είναι το πρώτο βήμα στην ανάπτυξη ενός μοντέλου ταξινόμησης. Επειδή τα προβλήματα ποιότητας δεδομένων συνήθως δεν μπορούν να αποφευχθούν, η εξόρυξη δεδομένων εστιάζει στον εντοπισμό και τη διόρθωσή τους (καθαρισμός δεδομένων - data cleaning), αλλά και στη χρήση αλγορίθμων που μπορούν να ανεχθούν κακή ποιότητα δεδομένων. Τα συνηθέστερα ζητήματα ποιότητας που προκύπτουν κατά τη μέτρηση και τη συλλογή των δεδομένων και η αντιμετώπισή τους, καθώς και οι μέθοδοι προεπεξεργασίας αναφέρονται παρακάτω.

6.1 Θόρυβος

Ο θόρυβος είναι η τυχαία συνιστώσα ενός σφάλματος μέτρησης. Συνήθως περιλαμβάνει την παραμόρφωση μιας τιμής ή την προσθήκη ψευδών περιπτώσεων. Ο όρος θόρυβος χρησιμοποιείται συχνά σε σχέση με δεδομένα που έχουν χωρική ή χρονική συνιστώσα. Σε τέτοιες περιπτώσεις, τεχνικές από την επεξεργασία σήματος ή εικόνας μπορούν συχνά να χρησιμοποιηθούν για τη μείωση του θορύβου και, ως εκ τούτου, να βοηθήσουν στην ανακάλυψη μοτίβων (σημάτων) που μπορεί να «είναι χαμένα στο θόρυβο». Ωστόσο, η εξάλειψη του θορύβου είναι συχνά δύσκολη και πολλή δουλειά στην εξόρυξη δεδομένων επικεντρώνεται στην επινόηση ισχυρών αλγορίθμων που παράγουν αποδεκτά αποτελέσματα ακόμη και όταν υπάρχει θόρυβος [74].

6.2 Ορθότητα, Μεροληψία και Ακρίβεια των δεδομένων

Στη στατιστική και γενικά στις πειραματικές επιστήμες, η ποιότητα της διαδικασίας λήψης μιας μέτρησης και τα δεδομένα που προκύπτουν αξιολογείται με την ορθότητα (precision) και τη μεροληψία (bias). Η ορθότητα χρησιμοποιείται για την περιγραφή της εγγύτητας που έχουν μεταξύ τους επαναλαμβανόμενες μετρήσεις της ίδιας υποκείμενης ποσότητας και συνήθως μετριέται με την τυπική απόκλιση που αυτές έχουν.

Η μεροληψία είναι η συστηματική διασπορά που παρατηρείται στις επαναλαμβανόμενες μετρήσεις της ίδιας ποσότητας και υπολογίζεται βρίσκοντας τη διαφορά ανάμεσα στη μέση τιμή των μετρήσεων αυτών και μιας γνωστής (αντικειμενικής) μέτρησης της ποσότητας αυτής. Η μεροληψία μπορεί να προσδιοριστεί μόνο για αντικείμενα των οποίων η μετρούμενη ποσότητα είναι γνωστή με μέσα εκτός της τρέχουσας διαδικασίας.

Συχνά χρησιμοποιείται ο γενικότερος όρος, ακρίβεια (accuracy), για να περιγραφεί ο βαθμός του σφάλματος μέτρησης στα δεδομένα. Η ακρίβεια είναι η εγγύτητα των μετρήσεων με την πραγματική τιμή της μετρούμενης ποσότητας. Μια σημαντική πτυχή της ακρίβειας είναι η χρήση σημαντικών ψηφίων (significant digits). Ο στόχος είναι να χρησιμοποιηθούν τόσα ψηφία για να αντιπροσωπεύεται το αποτέλεσμα μιας μέτρησης, όσα δικαιολογούνται από την ορθότητα των δεδομένων.

Ζητήματα όπως τα σημαντικά ψηφία, η ορθότητα, η μεροληψία και η ακρίβεια μερικές φορές παραβλέπονται, αλλά είναι σημαντικά στην εξόρυξη δεδομένων καθώς και για τη στατιστική. Πολλές φορές, τα σύνολα δεδομένων δεν συνοδεύονται από πληροφορίες σχετικά με την ορθότητα που τα διέπει, και επιπλέον, τα προγράμματα που χρησιμοποιούνται για την ανάλυση επιστρέφουν αποτελέσματα χωρίς καμία τέτοια πληροφορία. Ωστόσο, χωρίς κάποια κατανόηση της ακρίβειας των δεδομένων και των αποτελεσμάτων, ένας αναλυτής διατρέχει τον κίνδυνο να διαπράξει σοβαρά σφάλματα στην ανάλυσή τους [74].

6.3 Επεξεργασία κατηγορικών μεταβλητών

Κατηγορικές (categorical) είναι οι μεταβλητές που έχουν ως τιμές ετικέτες και όχι αριθμούς. Οι αλγόριθμοι ταξινόμησης για να εκπαιδευθούν απαιτούν στα δεδομένα τα χαρακτηριστικά να μην έχουν κατηγορική μορφή. Γι' αυτό είναι αναγκαία η κωδικοποίησή τους σε αριθμητική μορφή. Οι κατηγορικές μεταβλητές μπορούν να χωριστούν σε διατεταγμένες (ordinal) και μη διατεταγμένες που μπορούν να παίρνουν δύο ή περισσότερες μοναδικές τιμές (nominal). Η διαχείρισή τους μπορεί να γίνει ως εξής :

- Στις διατεταγμένες κατηγορικές μεταβλητές σε κάθε μοναδική τιμή ανατίθεται ένα αριθμός αντίστοιχος της βαρύτητάς της ώστε να μη χαθεί η ταξική φύση της μεταβλητής και να μη γίνει απώλεια πληροφορίας στον εκτιμητή . Τέτοια μπορεί να είναι η ηλικία που παίρνει ως τιμές ίσα χρονικά διαστήματα. Αυτή η προσέγγιση προϋποθέτει ότι η απόσταση μεταξύ των τιμών είναι περίπου ίση.
- Στις μη διατεταγμένες που παίρνουν πολλές τιμές συνήθως εφαρμόζονται οι εξής μέθοδοι:
    - **Η μέθοδος One-Hot Encoding**. Σύμφωνα με αυτή, για κάθε διαφορετική τιμή δημιουργείται ένα πρόσθετο χαρακτηριστικό και σε κάθε στιγμιότυπο μπαίνει 1 στο χαρακτηριστικό που έχει αντιστοιχιστεί στην τιμή που είχε και 0 στα υπόλοιπα. Τέτοια παραδείγματα είναι η φυλή και η χώρα προέλευσης.

- **Η μέθοδος Dummy Encoding.** Όταν χρησιμοποιείται η μέθοδος One-Hot Encoding, πρέπει να λαμβάνεται υπ' όψιν ότι προκαλείται ένα πρόβλημα, το οποίο είναι γνωστό ως dummy variable trap (παγίδα της «ψεύτικης» μεταβλητής). Αυτό οφείλεται στην εισαγωγή πολυσυγγραμμικότητας (multicollinearity), δηλαδή στην ύπαρξη χαρακτηριστικών που έχουν υψηλή συσχέτιση μεταξύ τους και το ένα προβλέπει την τιμή άλλων. Αυτό επηρεάζει αρνητικά ορισμένες μεθόδους, όπως για παράδειγμα αυτές που απαιτούν αντιστροφή πίνακα. Εάν τα χαρακτηριστικά έχουν υψηλή συσχέτιση, οι πίνακες είναι υπολογιστικά δύσκολο να αντιστραφούν, γεγονός που μπορεί να οδηγήσει σε αριθμητικά ασταθείς εκτιμήσεις. Για να μειωθεί η συσχέτιση μεταξύ των μεταβλητών, η μέθοδος Dummy Encoding αφαιρεί μια στήλη από τον κωδικοποιημένο πίνακα που δημιουργήθηκε με τη μέθοδο one-hot encoding. Επισημαίνεται ότι με αυτόν τον τρόπο δεν υπάρχει απώλεια σημαντικής πληροφορίας ενώ αντίθετα θα μειωθεί ο χρόνος που χρειάζεται για τη διαδικασία εκπαίδευσης του αλγορίθμου και θα απαιτηθεί λιγότερη μνήμη [113].

6.4 Χωρισμός dataset σε training και test set

Το dataset χωρίζεται σε training και test set ώστε να γίνει η εκπαίδευση των εκτιμητών στα δεδομένα του training set και ύστερα η αξιολόγησή τους στα δεδομένα του test set. Όταν γίνεται αυτός ο χωρισμός, πρέπει να λαμβάνεται υπ' όψιν ότι αποκρύπτονται πολύτιμες πληροφορίες από τις οποίες θα μπορούσε το μοντέλο να επωφεληθεί κατά τη διαδικασία της εκπαίδευσης. Επομένως, δεν είναι επιθυμητή η διάθεση μεγάλου όγκου δεδομένων στο test set. Ωστόσο, όσο μικρότερο είναι το test set, τόσο πιο ανακριβής είναι η εκτίμηση του σφάλματος γενίκευσης (generalization error). Η διαίρεση ενός συνόλου δεδομένων σε training και test set έχει να κάνει με την εξισορρόπηση αυτού του ισοζυγίου. Στην πράξη, οι διαχωρισμοί που χρησιμοποιούνται πιο συχνά είναι 60:40, 70:30 ή 80:20, ανάλογα με το μέγεθος του αρχικού συνόλου δεδομένων. Ωστόσο, για μεγάλα σύνολα δεδομένων, οι διαχωρισμοί 90:10 ή 99:1 είναι επίσης κατάλληλοι. Για παράδειγμα, εάν το σύνολο δεδομένων περιέχει περισσότερα από 100.000 παραδείγματα εκπαίδευσης, μπορεί να είναι καλό να κρατηθούν μόνο 10.000 στιγμιότυπα για test set, προκειμένου να ληφθεί μια καλή εκτίμηση της απόδοσης στη γενίκευση. Πρέπει πάντα να αποφεύγεται η διάχυση πληροφορίας από το training στο test set ώστε να είναι ορθότερη η εκτίμηση της ικανότητας γενίκευσης του μοντέλου [113].

6.5 Απουσιάζουσες τιμές (Missing Values)

Δεν είναι ασυνήθιστο για ένα στιγμιότυπο να λείπουν μία ή περισσότερες τιμές από τα χαρακτηριστικά του. Σε ορισμένες περιπτώσεις, οι πληροφορίες δεν συλλέχθηκαν, όπως η άρνηση ενός ερωτηθέντος να δώσει μια απάντηση. Σε άλλες περιπτώσεις, ορισμένα χαρακτηριστικά δεν ισχύουν για όλα τα αντικείμενα. Για παράδειγμα, συχνά, οι φόρμες έχουν μέρη υπό όρους που συμπληρώνονται μόνο όταν ένα άτομο απαντά σε μια προηγούμενη ερώτηση με συγκεκριμένο τρόπο, αλλά για λόγους απλότητας, όλα τα πεδία αποθηκεύονται. Είναι σημαντικό οι απουσιάζουσες τιμές, εάν εντοπιστούν, να λαμβάνονται υπόψη κατά την ανάλυση δεδομένων. Υπάρχουν διάφορες στρατηγικές για την αντιμετώπιση αυτού του ζητήματος, καθεμία από τις οποίες χρησιμοποιείται ανάλογα με την περίπτωση. Αυτές οι

στρατηγικές παρατίθενται στη συνέχεια, μαζί με μια σύνοψη των πλεονεκτημάτων και των μειονεκτημάτων τους [74]:

- Αφαίρεση στιγμιοτύπων με απουσιάζουσες τιμές. Απλή και αποτελεσματική στρατηγική, ωστόσο, εάν ένα στιγμιότυπο με απουσιάζουσες τιμές περιέχει σημαντικές πληροφορίες και εάν πολλά έχουν τιμές που λείπουν, τότε μια αξιόπιστη ανάλυση μπορεί να είναι δύσκολη ή αδύνατη. Εάν όμως ένα σύνολο δεδομένων έχει μόνο λίγες εγγραφές στις οποίες λείπουν τιμές, τότε ίσως είναι σκόπιμο να παραληφθούν.
- Συμπλήρωση των τιμών που απουσιάζουν, ειδικά όταν είναι διαθέσιμο μικρό πλήθος εγγραφών. Μερικές φορές τα δεδομένα που λείπουν μπορούν να εκτιμηθούν αξιόπιστα με χρήση των τιμών των στιγμιοτύπων που βρίσκονται πλησιέστερα σε αυτό με την απουσιάζουσα τιμή. Εάν το χαρακτηριστικό είναι συνεχές, τότε χρησιμοποιείται η μέση τιμή του χαρακτηριστικού των πλησιέστερων γειτόνων. Εάν είναι κατηγορικό, τότε μπορεί να ληφθεί η πιο συχνά εμφανιζόμενη τιμή. Όλες οι μέθοδοι αντικατάστασης των χαμένων τιμών, μπορεί να προκαλέσουν απόκλιση στα δεδομένα, αφού η τιμή αντικατάστασης κατά πάσα πιθανότητα δεν είναι η σωστή, αλλά στις περισσότερες των περιπτώσεων δεν υπάρχει εναλλακτική λύση.
- Πολλά μοντέλα ταξινόμησης μπορούν να τροποποιηθούν ώστε να λειτουργούν με τιμές που λείπουν.

6.6 Ανακόλουθες τιμές (Inconsistent Values)

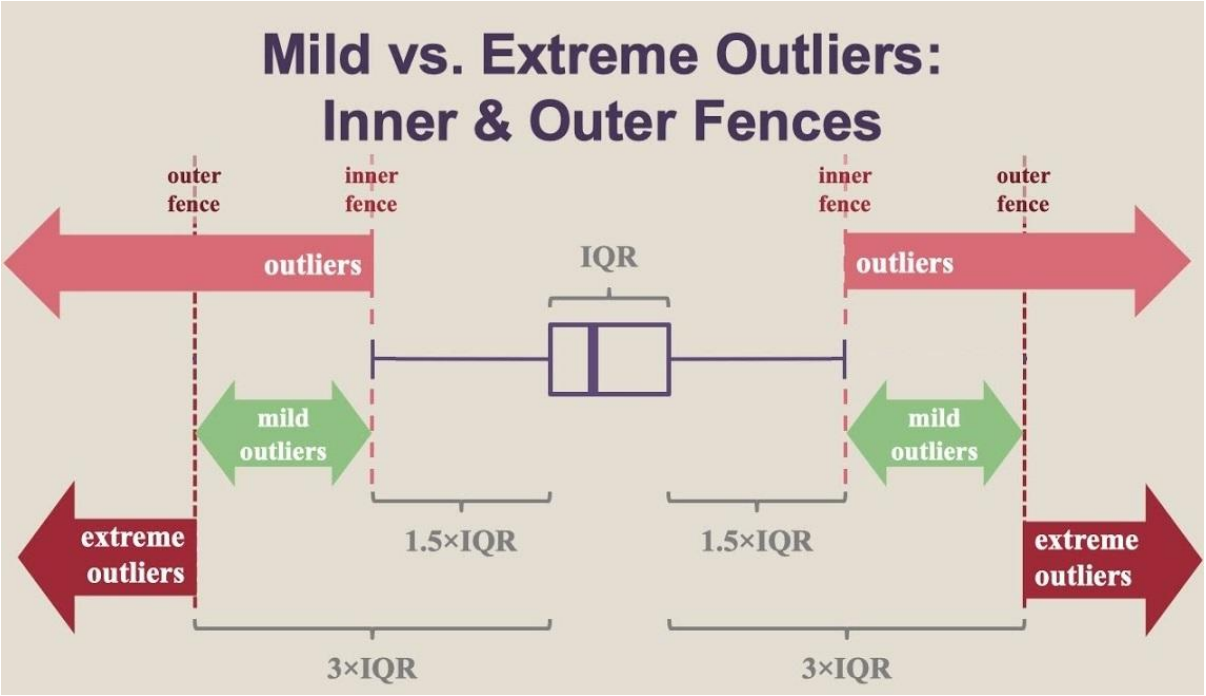
Μερικές φορές τα δεδομένα ενδέχεται να περιέχουν μη αποδεκτές τιμές. Ανεξάρτητα από την αιτία που προέκυψε αυτό, είναι σημαντικό να εντοπιστούν και, εάν είναι δυνατόν, να διορθωθούν. Ορισμένοι τύποι μη αποδεκτών τιμών είναι εύκολο να εντοπιστούν, όπως για παράδειγμα, το ύψος ενός ατόμου δεν πρέπει να είναι αρνητικό. Σε άλλες περιπτώσεις, μπορεί να χρησιμοποιηθεί μια εξωτερική πηγή πληροφοριών τόσο για τον εντοπισμό όσο και για τη διόρθωση τέτοιων τιμών [74].

6.7 Διπλότυπα δεδομένα (Duplicate Data)

Ένα σύνολο δεδομένων μπορεί να περιλαμβάνει στιγμιότυπα δεδομένων που είναι διπλότυπα. Αυτά μπορεί είτε να αντιπροσωπεύουν διαφορετικά στιγμιότυπα που τυχαίνει να έχουν τις ίδιες τιμές χαρακτηριστικών είτε να προέκυψαν από λάθη κατά τη συλλογή των δεδομένων. Εάν διαπιστωθεί ότι αυτά είναι πολλαπλά αντίγραφα του ίδιου στιγμιοτύπου, πρέπει να αφαιρεθούν και να διατηρηθεί μόνο ένα. Εάν αυτά αντιπροσωπεύουν διαφορετικές εγγραφές, άρα είναι και τα συχνότερα εμφανιζόμενα, τότε πρέπει να διατηρούνται ώστε το μοντέλο να δίνει περισσότερο βάρος στη σωστή τους πρόβλεψη κατά την εκπαίδευση του. Ενδέχεται όμως να προκαλέσουν προβλήματα στην εκπαίδευση ορισμένων αλγόριθμων, εάν η πιθανότητα πανομοιότυπων στιγμιότυπων δεν λαμβάνεται ειδικά υπόψη στη σχεδίασή τους. Για παράδειγμα στον KNN μπορεί η ομάδα των K-πλησιέστερων γειτόνων ενός δείγματος να αποτελείται μόνο από duplicates ειδικά εάν αυτά αποτελούν μεγάλο μέρος του dataset. Μια προσέγγιση αντιμετώπισης του προβλήματος αυτού είναι πάλι η διατήρηση μόνο ενός δείγματος για κάθε ομάδα διπλών αντικειμένων [74].

6.8 Ακραίες τιμές (Outliers)

Ακραίες είναι οι τιμές ενός χαρακτηριστικού που είναι ασυνήθιστες σε σχέση με τις τυπικές τιμές για αυτό το χαρακτηριστικό. Σε αντίθεση με το θόρυβο, οι ακραίες τιμές μπορεί να είναι πραγματικές τιμές χαρακτηριστικών. Κατ' αρχήν οι ακραίες τιμές θα πρέπει να ορισθούν και να εντοπισθούν. Μια από τις δημοφιλέστερες μεθόδους εύρεσης των ακραίων τιμών είναι η Τεχνική του Ενδοτεταρτημοριακού Εύρους (Interquartile Range - IQR) (Εικόνα 15) [116].



Εικόνα 15: Boxplot με ενδοτεταρτημοριακό εύρος (Προσαρμογή) [117]

Τα δεδομένα χωρίζονται σε 4 ίσα μέρη (τεταρτημόρια) με το ενδοτεταρτημοριακό εύρος (IQR) να περιλαμβάνει το ενδιάμεσο 50% των παρατηρήσεων. Αν ορίσουμε ως Q_1 την τιμή που αντιστοιχεί στη θέση 25% κάτω από τη διάμεσο (μεσαία τιμή) των δεδομένων και Q_3 την τιμή που αντιστοιχεί στη θέση 75% πάνω από τη διάμεσο των δεδομένων τότε για τον εντοπισμό των ακραίων τιμών ισχύει:

$$\text{Αν} \quad \text{IQR} = Q_3 - Q_1 \qquad (38)$$

Τότε:

$$\text{Άνω ήπια ακραία τιμή (Upper mild outlier)} = Q_3 + 1.5 * \text{IQR} \qquad (39)$$
$$\text{Κάτω ήπια ακραία τιμή (Lower mild outlier)} = Q_3 - 1.5 * \text{IQR} \qquad (40)$$

$$\text{Άνω εξαιρετικά ακραία τιμή (Upper extreme outlier)} = Q_3 + 3 * \text{IQR} \qquad (41)$$
$$\text{Κάτω εξαιρετικά ακραία τιμή (Lower extreme outlier)} = Q_3 - 3 * \text{IQR} \qquad (42)$$

Ο χειρισμός των ακραίων τιμών περιλαμβάνει [118]:

- Απόρριψη των δειγμάτων αν το πλήθος των ακραίων τιμών είναι μικρός .
- Αντικατάσταση, η οποία μπορεί να γίνει :

- ή με τον μέσο όρο της στήλης, που είναι και η πιο κοινή μέθοδος
- ή με την μικρότερη τιμή του χαρακτηριστικού αν είναι κάτω ακραία τιμή και με την μεγαλύτερη αν είναι άνω ακραία τιμή.

6.9 Κανονικοποίηση (Scaling)

Όταν σε ένα σύνολο δεδομένων πρόκειται να χρησιμοποιηθούν διαφορετικές μεταβλητές μαζί, που εμφανίζουν διαφορετικά επίπεδα διακυμάνσεων στις αριθμητικές τιμές τους, τότε είναι απαραίτητο να γίνει ένας κατάλληλος μετασχηματισμός, η κανονικοποίηση, ώστε να αποφευχθεί η υπερεκτίμηση των μεταβλητών με μεγάλες τιμές αλλά και η υποεκτίμηση αυτών με μικρές τιμές στα αποτελέσματα της ανάλυσης. Με τη κανονικοποίηση, προσαρμόζονται οι τιμές των μεταβλητών σε μία κοινή κλίμακα (π.χ. μεταξύ 0 και 1) με αποτέλεσμα να επιτυγχάνεται η ομοιόμορφη εκτίμησή τους [74]. Υπάρχουν δύο συνήθεις τρόποι που επιτυγχάνεται αυτό:

- **Κλίμακα ελαχίστου μεγίστου (Min-Max Scaling)**, στην οποία όλες οι τιμές των χαρακτηριστικών έρχονται στο εύρος [0,1], με την εφαρμογή της σχέσης (43).

$$\mathrm{x_{new}} = \frac{\mathrm{x} - \mathrm{X_{min}}}{\mathrm{X_{max}} - \mathrm{X_{min}}} \quad (43)$$

Όπου X_{max} και X_{min} η μέγιστη και η ελάχιστη τιμή αντίστοιχα κάθε χαρακτηριστικού Χ.

- **Standard scaling**, όπου υπολογίζεται το standard score για κάθε τιμή χαρακτηριστικού, ως εξής:

$$\mathrm{x_{new}} = \frac{\mathrm{x} - \mu}{\sigma} \quad (44)$$

Όπου μ είναι ο μέσος όρος των τιμών του κάθε χαρακτηριστικού και σ είναι η τυπική του απόκλιση. Με αυτό τον τρόπο κάθε χαρακτηριστικό μετατρέπεται σε μια νέα μεταβλητή που έχει μέσο όρο 0 και τυπική απόκλιση 1.

Σημειώνεται ότι η κανονικοποίηση πρέπει να γίνεται στο training set και το test set πρέπει να κανονικοποιείται χρησιμοποιώντας τις παραμέτρους που προέκυψαν από το training set, για να μην υπάρχει διαρροή πληροφορίας από το test set στην εκπαίδευση του ταξινομητή.

6.10 Μείωση διαστατικότητας (Dimensionality Reduction)

Μια πολύ σημαντική παράμετρος για την απόδοση των ταξινομητών είναι η διασταστικότητα των δεδομένων, ιδιαίτερα σε σχέση με τον διαθέσιμο αριθμό δειγμάτων. Γενικά και ανεξάρτητα από το μοντέλο του ταξινομητή, η απόδοση αυξάνεται όσο αυξάνεται το πλήθος και η ποιότητα των δεδομένων και όσο μειώνεται η διαστατικότητα. Αντίστροφα, τα προβλήματα δυσκολεύουν όσο η διαστατικότητα αυξάνεται και τα δείγματα δεν επαρκούν για να καλύψουν όλες τις κατηγορίες του προβλήματος. Το πρόβλημα αυτό αναφέρεται ως η κατάρα της διαστατικότητας (the curse of dimensionality): όσο αυξάνει η διαστατικότητα, τόσο τα διαθέσιμα δεδομένα γίνονται αραιά (sparse). Σε γενικές γραμμές λοιπόν, οι πολύ μεγάλες διαστάσεις του χώρου εισόδου (χαρακτηριστικών) κάνουν δυσκολότερο για τον ταξινομητή να υπολογίσει το σύνορο απόφασης μεταξύ των κλάσεων και αυξάνουν τις

απαιτήσεις χώρου και χρόνου εκπαίδευσης ή/και ταξινόμησης.

Για να μειωθεί η διαστατικότητα των δεδομένων χρησιμοποιούμε τεχνικές μείωσης διαστατικότητας (dimensionality reduction). Η μείωση διαστατικότητας γίνεται με εφαρμογή είτε τεχνικών επιλογής χαρακτηριστικών (feature selection) όπου ουσιαστικά αφαιρούνται κάποια χαρακτηριστικά με βάση ένα κριτήριο (πχ variance) χωρίς μετασχηματισμό των τιμών τους, είτε τεχνικών εξαγωγής χαρακτηριστικών (feature extraction), όπου μετασχηματίζουμε τις τιμές των χαρακτηριστικών σε νέες (εξάγουμε δηλαδή νέα χαρακτηριστικά) αλλά σε ένα χώρο μικρότερων διαστάσεων. Η βασικότερη τεχνική feature extraction είναι η ανάλυση σε κύριες συνιστώσες (Principal Components Analysis - PCA). [74]. Σημειώνεται ότι η μέθοδος αυτή μπορεί να χρησιμοποιηθεί και για οπτικοποίηση των δεδομένων είτε σε δύο είτε σε τρεις διαστάσεις.

6.11 Έλεγχος συσχέτισης χαρακτηριστικών (Attributes' Correlation)

Η συσχέτιση χρησιμοποιείται για τη μέτρηση της γραμμικής σχέσης μεταξύ των τιμών δύο χαρακτηριστικών στα δείγματα ενός dataset (πχ ύψος και βάρος) ή μεταξύ δύο αντικειμένων του ίδιου χαρακτηριστικού (πχ ζεύγος χρονοσειρών θερμοκρασίας). Συνεπώς, η συσχέτιση αντιπροσωπεύει τη μέτρηση της ομοιότητας μεταξύ των χαρακτηριστικών, οι τιμές των οποίων μπορούν να έχουν διαφορετικούς τύπους και κλίμακες. Υπάρχουν πολλοί τρόποι υπολογισμού της συσχέτισης . Για τις αριθμητικές τιμές συχνά χρησιμοποιείται η συσχέτιση του Pearson (σχέση 45) [74]. Εάν X και Y δύο διανύσματα αριθμητικών τιμών, ισχύει:

$$corr(X,Y) = \frac{covariance(X,Y)}{standard\_deviation(X) * standard\_deviation(Y)} \qquad (45)$$

Η συσχέτιση παίρνει τιμές από -1 έως 1. Θετική συσχέτιση δείχνει ότι οι μεταβλητές αυξάνονται ή μειώνονται μαζί, ενώ αρνητική υποδηλώνει ότι εάν η μία αυξάνεται, η άλλη μειώνεται. Μηδενική συσχέτιση σημαίνει ότι δεν υπάρχει καμία σχέση ή σύνδεση μεταξύ των μεταβλητών. Όταν δύο χαρακτηριστικά εμφανίζουν υψηλή συσχέτιση (ισχυρά γραμμική σχέση), συνήθως μεγαλύτερη από 0.7 ή μικρότερη από -0,7, τότε λέγεται ότι εμφανίζουν συγγραμικότητα (collinearity), ενώ όταν αυτό συμβαίνει σε τρία ή περισσότερα χαρακτηριστικά, τότε λέγεται πολυσυγγραμικότητα (multicollinearity). Για τον υπολογισμό της συσχέτισης μεταξύ δύο χαρακτηριστικών ενός dataset μπορεί να χρησιμοποιηθεί ο πίνακας συσχέτισης (correlation matrix), ενώ μεταξύ τριών ή περισσότερων ο παράγοντας πληθωριστικής διασποράς (Variance Inflation Factor - VIF). Σημειώνεται ότι υψηλή συσχέτιση μεταξύ μίας μεταβλητής που χρησιμοποιείται για την πραγματοποίηση πρόβλεψης και της μεταβλητής ταξινόμησης είναι κάτι επιθυμητό. Δεν είναι όμως επιθυμητό όταν αυτό συμβαίνει μεταξύ δύο χαρακτηριστικών που χρησιμοποιούνται ως προβλεπτικοί παράγοντες. Η ύπαρξη προβλεπτικών παραγόντων που εμφανίζουν υψηλή συσχέτιση σε ένα dataset αποτελεί πρόβλημα στα γραμμικά κυρίως μοντέλα, όπως η γραμμική ή η λογιστική παλινδρόμηση, καθώς δυσκολεύει τον υπολογισμό των συνιστωσών (coefficients) του μοντέλου.

Για να αντιμετωπιστεί αυτό το πρόβλημα μπορεί να αφαιρεθεί ένα από τα δύο χαρακτηριστικά, συνήθως αυτό με τη μικρότερη συσχέτιση με τη μεταβλητή ταξινόμησης. Άλλη προσέγγιση είναι η εφαρμογή της μεθόδου PCA στο dataset, ώστε να μετατραπεί το σύνολο των ισχυρά συσχετιζόμενων προβλεπτών σε ένα σύνολο μη συσχετιζόμενων γραμμικά μεταβλητών. Αυτά μπορούν να οδηγήσουν σε ένα απλούστερο και γρηγορότερα εκπαιδευόμενο μοντέλο ταξινόμησης [119].

6.12 Ανισορροπία στις κλάσεις (Class Imbalance problem)

Ένα dataset είναι ισορροπημένο όταν το πλήθος των δειγμάτων που ανήκουν σε κάθε κλάση είναι περίπου ίσο. Σε πολλά σύνολα δεδομένων όμως υπάρχει δυσανάλογος αριθμός περιπτώσεων που ανήκουν στις διάφορες κλάσεις, μια ιδιότητα γνωστή ως ανισορροπία κλάσεων. Τέτοια παραδείγματα μοντελοποίησης σπάνιων συμβάντων είναι η ανίχνευση περιπτώσεων απάτης, η διάγνωση σοβαρής ασθένειας κ.α.

Ένας ταξινομητής που εκπαιδεύεται σε ένα μη ισορροπημένο σύνολο δεδομένων εμφανίζει μεροληψία (bias) στη βελτίωση της απόδοσής του σε σχέση με την πλειοψηφική κλάση, το οποίο δεν είναι επιθυμητό. Ως αποτέλεσμα, όλοι οι ταξινομητές επηρεάζονται εάν η μειοψηφική τάξη δεν εκπροσωπείται καλά στο training set, αν και ορισμένοι από αυτούς είναι πιο αποτελεσματικοί στο χειρισμό της ανισορροπίας από κάποιους άλλους, όπως ο k-NN. Εξαιτίας αυτού, η ακρίβεια, που είναι η παραδοσιακή μετρική αξιολόγησης της απόδοσης των ταξινομητών, δεν είναι αξιόπιστη σε αυτήν την περίπτωση. Αυτό συμβαίνει διότι, αν και εμφανίζει υψηλή τιμή, δεν αντιπροσωπεύεται η επιτυχία του μοντέλου στον εντοπισμό της μειοψηφούσας κλάσης σε νέα δεδομένα [74].

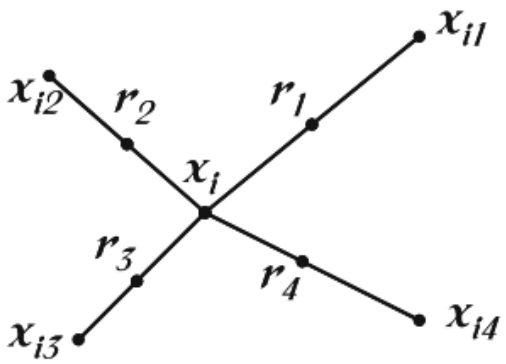
Για την αντιμετώπιση του προβλήματος αυτού πρέπει να γίνει κατάλληλη επεξεργασία στο training set, ώστε να υπάρχει επαρκής εκπροσώπηση της μειοψηφούσας κλάσης. Οι βασικές προσεγγίσεις είναι [120]:

1) Επαναδειγματοληψία δεδομένων : Τα στιγμιότυπα εκπαίδευσης τροποποιούνται με τέτοιο τρόπο, ώστε να παράγουν μια πιο ισορροπημένη κατανομή των κλάσεων στο dataset, το οποίο επιτρέπει στους ταξινομητές να αποδίδουν με παρόμοιο τρόπο, όπως με τα ισορροπημένα dataset. Οι τεχνικές επαναδειγματοληψίας (resampling) μπορούν να κατηγοριοποιηθούν σε τρεις ομάδες:

- **Υποδειγματοληψία (Undersampling)**, όπου δημιουργείται ένα υποσύνολο του αρχικού dataset με αφαίρεση δειγμάτων της πλειοψηφούσας κλάσης. Οι δύο συνηθέστεροι μέθοδοι υποδειγματοληψίας είναι:
  - **Τυχαία υποδειγματοληψία (Random Undersampling).** Στην τυχαία υποδειγματοληψία, οι περιπτώσεις της πλειοψηφικής κατηγορίας αφαιρούνται τυχαία μέχρι να επιτευχθεί μια πιο ισορροπημένη κατανομή κλάσεων. Δίνει συχνά καλά αποτελέσματα με πολύ μικρό χρόνο απόκρισης . Ωστόσο, ένας περιορισμός είναι ότι ορισμένα από τα χρήσιμα δείγματα από τη πλειοψηφία (π.χ. αυτά που είναι πιο κοντά στο πραγματικό όριο απόφασης) ενδέχεται να μην επιλέγονται για εκπαίδευση, καταλήγοντας επομένως σε ένα κατώτερο μοντέλο ταξινόμησης. Αυτό αντιμετωπίζεται εν μέρει με τεχνικές που χρησιμοποιούν ευρετικές διαδικασίες για να αποφασίσουν στοχευμένα τα δείγματα που πρέπει να αφαιρεθούν ή να διατηρηθούν.
  - **Near Miss:** Επιλέγει από τη πλειοψηφούσα κλάση ποια δείγματα θα διατηρηθούν, βάσει της απόστασης (π.χ. της ευκλείδειας) των δειγμάτων πλειοψηφίας με αυτά της μειοψηφίας [121].
- **Υπερδειγματοληψία (Oversampling)**, όπου δημιουργείται ένα υπερσύνολο του αρχικού dataset, αναπαράγοντας κάποια δείγματα της μειοψηφούσας κλάσης ή δημιουργώντας νέα τεχνητά δείγματα βάσει των υπαρχόντων. Οι δύο συνηθέστεροι μέθοδοι υπερδειγματοληψίας είναι:
  - **Τυχαία υπερδειγματοληψία (Random Oversampling)**. Στοχεύει στην εξισορρόπηση της κατανομής κλάσεων μέσω της τυχαίας αναπαραγωγής στιγμιοτύπων της κλάσης μειοψηφίας. Το βασικό μειονέκτημα αυτής της

μεθόδου είναι ότι μπορεί να αυξήσει την πιθανότητα υπερεκπαίδευσης, καθώς δημιουργεί πανομοιότυπα αντίγραφα στιγμιοτύπων της μειοψηφίας.

- **SMOTE (Synthetic Minority Over-sampling Technique)** [122]. Βασίζεται στη δημιουργία νέων τεχνητών στιγμιοτύπων που ανήκουν στη μειοψηφούσα κλάση, παρεμβάλλοντάς τα ανάμεσα σε στιγμιότυπα της μειοψηφούσας κλάσης (Εικόνα 16). Αρχικά επιλέγεται ένα δείγμα που ανήκει στην μειοψηφούσα κλάση (x_i) ως βάση. Βάσει μιας μετρικής απόστασης, επιλέγονται τέσσερα δείγματα (x_{i1} έως x_{i4}) της ίδιας κλάσης από το training set που βρίσκονται πλησιέστερα στο x_i. Ύστερα σχεδιάζονται τα ευθύγραμμα τμήματα που τα ενώνουν και τέλος δημιουργούνται με τυχαία παρεμβολή τέσσερα συνθετικά δείγματα (r_1 έως r_4) σε κάποιο σημείο κατά μήκος αυτών των τμημάτων. Αυτή η τεχνική τείνει να διευρύνει τα όρια απόφασης που σχετίζονται με τις σπάνιες περιπτώσεις, σε αντίθεση με την υπερεκπαίδευση που σχετίζεται με την τυχαία υπερδειγματοληψία.



Εικόνα 16: *Απλή απεικόνιση δημιουργίας νέων συνθετικών στιγμιοτύπων με χρήση της τεχνικής SMOTE* [120]

- **ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning)** [123]**.** Η βασική ιδέα του ADASYN είναι να η χρησιμοποίηση μιας σταθμισμένης κατανομής για διαφορετικά δείγματα της κλάσης της μειοψηφίας, ανάλογα με το επίπεδο δυσκολίας στη εκμάθησή τους, παράγοντας περισσότερα συνθετικά δεδομένα για δείγματα της μειοψηφούσας κλάσης που είναι πιο δύσκολη η εκμάθηση τους σε σύγκριση με εκείνα που είναι πιο εύκολη. Ως αποτέλεσμα, η προσέγγιση ADASYN βελτιώνει την εκπαίδευση σε σχέση με τις κατανομές δεδομένων με δύο τρόπους: (1) μειώνοντας την μεροληψία που εισάγεται από την ανισορροπία τάξης και (2) μετατοπίζοντας προσαρμοστικά το όριο απόφασης ταξινόμησης προς τα δύσκολα παραδείγματα.

- **Υβριδικές μέθοδοι**, που συνδυάζουν τεχνικές υποδειγματοληψίας και υπερδειγματοληψίας.

2) Τροποποίηση των αλγορίθμων (Algorithmic Modification) : Προσανατολίζεται στην τροποποίηση των βασικών αλγορίθμων, ώστε να είναι πιο προσαρμοσμένοι σε ζητήματα ανισορροπίας κλάσεων.

3) Εκπαίδευση με ευαισθησία στο κόστος (Cost-sensitive learning): Ενσωματώνει προσεγγίσεις σε επίπεδο δεδομένων ή σε αλγοριθμικό επίπεδο ή και στα δύο επίπεδα από κοινού, αποδίδοντας υψηλότερο κόστος για την εσφαλμένη ταξινόμηση των δειγμάτων της

μειοψηφούσας κλάσης σε σχέση με της πλειοψηφικής, στοχεύοντας στην ελαχιστοποίηση των σφαλμάτων υψηλότερου κόστους.

Κεφάλαιο 7: Ερμηνευσιμότητα (Interpretability) μοντέλων μηχανικής μάθησης

7.1 Ορισμός και ταξινόμηση

Ερμηνευσιμότητα κατά τον Miller [124] είναι ο βαθμός κατά τον οποίο ένας άνθρωπος μπορεί να κατανοήσει την αιτία μιας απόφασης. Όσο μεγαλύτερη είναι η ερμηνευσιμότητα ενός μοντέλου μηχανικής μάθησης, τόσο πιο εύκολο είναι για κάποιον να κατανοήσει γιατί έχουν ληφθεί ορισμένες αποφάσεις ή προβλέψεις. Ένα μοντέλο είναι καλύτερα ερμηνεύσιμο από ένα άλλο μοντέλο, εάν οι αποφάσεις του είναι ευκολότερα κατανοητές από τις αποφάσεις του άλλου.

Η ερμηνευσιμότητα ενός μοντέλου είναι απαραίτητη, διότι σε ορισμένα προβλήματα δεν αρκεί η πρόβλεψη αλλά απαιτείται και η εξήγηση του πώς κατέληξε σε αυτή τη πρόβλεψη. Μια σωστή πρόβλεψη μπορεί να λύνει μόνο εν μέρει το αρχικό πρόβλημα. Όταν ένα πρόβλημα εντάσσεται σε περιβάλλον χαμηλού κινδύνου, ένα λάθος δεν θα έχει σοβαρές συνέπειες, (π.χ. σύστημα σύστασης ταινίας) ούτε απαιτεί ιδιαίτερες επεξηγήσεις. Όταν όμως το πρόβλημα αφορά σε θέματα υψηλού ρίσκου, όπως πχ σε ζητήματα υγείας, η λανθασμένη πρόβλεψη κοστίζει περισσότερο και απαιτείται καλύτερη γνώση των παραγόντων που συνέβαλαν στη λήψη της [125].

Ένα μοντέλο μαύρου κουτιού (black box model) είναι ένα σύστημα που δεν αποκαλύπτει τους εσωτερικούς μηχανισμούς του. Στη μηχανική μάθηση, ως μαύρα κουτιά περιγράφονται μοντέλα που δεν μπορούν να γίνουν κατανοητά κοιτάζοντας τις παραμέτρους τους (π.χ. ένα νευρωνικό δίκτυο). Αυτό επιδιώκουν να αντιμετωπίσουν οι μέθοδοι για την ερμηνευσιμότητα της μηχανικής μάθησης, οι οποίες ταξινομούνται σύμφωνα με διάφορα κριτήρια. Μια συνήθης ταξινόμηση είναι βάσει του μοντέλου που χρησιμοποιήθηκε [125]:

- **Μέθοδοι που αφορούν συγκεκριμένο μοντέλο (model-specific).** Εφαρμόζονται σε μεμονωμένο μοντέλο ή ομάδα μοντέλων και εξαρτώνται σε μεγάλο βαθμό από τη λειτουργία και τις δυνατότητες του. Πχ. η χρήση των συνιστωσών παλινδρόμησης (coefficients) σε μοντέλο λογιστικής παλινδρόμησης είναι model-specific.
- **Μέθοδοι ανεξαρτήτως μοντέλου (model-agnostic).** Μπορούν να χρησιμοποιηθούν σε οποιοδήποτε μοντέλο μηχανικής μάθησης και εφαρμόζονται μετά την εκπαίδευση του μοντέλου (post hoc). Αυτές οι μέθοδοι συνήθως λειτουργούν με την ανάλυση των ζευγών χαρακτηριστικών εισόδου και εξόδου. Εξ ορισμού, δεν μπορούν να έχουν πρόσβαση σε εσωτερικά στοιχεία των μοντέλων όπως βάρη ή δομικές πληροφορίες. Οι model-agnostic μέθοδοι μπορούν να ταξινομηθούν σε δύο κατηγορίες:
    - **Καθολικές μέθοδοι ανεξαρτήτως μοντέλου (global model-agnostic methods).** Χαρακτηρίζουν συνολικά τη συμπεριφορά ενός μοντέλου μηχανικής μάθησης και είναι ιδιαίτερα χρήσιμες όταν είναι επιθυμητή η κατανόηση των γενικών μηχανισμών που υπάρχουν στα δεδομένα ή η διόρθωση ενός μοντέλου. Ένα παράδειγμα τέτοιας μεθόδου είναι η μετάθεση σημαντικότητας χαρακτηριστικών (Permutation feature importance) [126].
    - **Τοπικές μέθοδοι ανεξαρτήτως μοντέλου (local model-agnostic methods).** Επεξηγούν μεμονωμένες προβλέψεις ενός μοντέλου. Ένα παράδειγμα τέτοιας

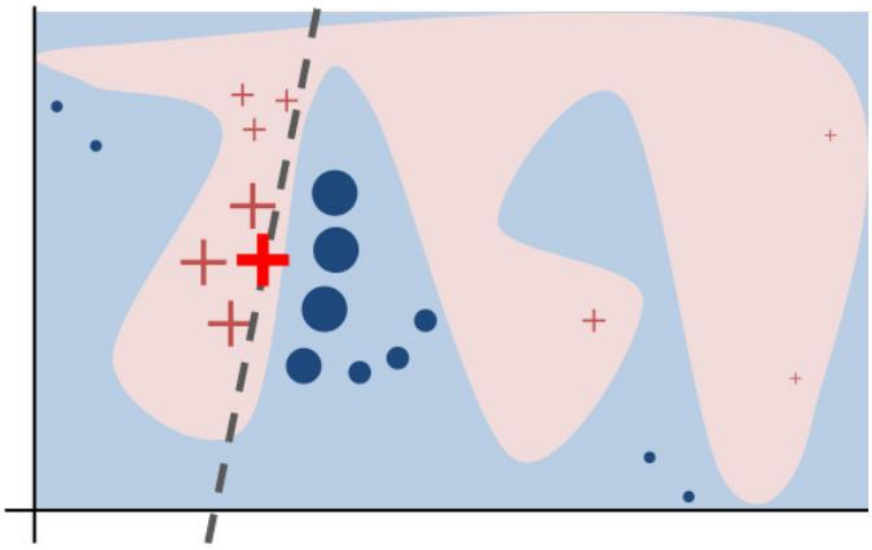
μεθόδου είναι η μέθοδος με τοπικά ερμηνεύσιμες εξηγήσεις ανεξαρτήτως μοντέλου (Local Interpretable Model-agnostic Explanations - LIME) [127].

7.2 Σημαντικότητα χαρακτηριστικών με μετάθεση (Permutation feature Importance)

Σύμφωνα με αυτή τη μέθοδο, η σημαντικότητα ενός χαρακτηριστικού υπολογίζεται μετρώντας την αύξηση του σφάλματος πρόβλεψης του μοντέλου μετά τη τυχαία αναδιάταξη/ανακάτεμα των τιμών του κατά προτίμηση στο test set [125]. Πρώτα υπολογίζεται το αρχικό σφάλμα του μοντέλου και μετά για κάθε χαρακτηριστικό υπολογίζεται ξανά το σφάλμα αφού έχει προηγηθεί ανακάτεμα των τιμών του. Τέλος, η σημαντικότητα κάθε χαρακτηριστικού υπολογίζεται ως το πηλίκο ή η διαφορά των δύο σφαλμάτων. Ένα χαρακτηριστικό είναι "σημαντικό" εάν το ανακάτεμα των τιμών του αυξάνει το σφάλμα μοντέλου, επειδή σε αυτήν την περίπτωση το μοντέλο βασίστηκε στο χαρακτηριστικό αυτό για την πρόβλεψη. Ένα χαρακτηριστικό είναι «ασήμαντο» εάν το ανακάτεμα των τιμών του αφήνει αμετάβλητο το σφάλμα του μοντέλου, επειδή σε αυτήν την περίπτωση το μοντέλο αγνόησε το χαρακτηριστικό αυτό για την πραγματοποίηση της πρόβλεψης. Μια εναλλακτική υλοποίηση είναι αντί της αύξησης στο σφάλμα να χρησιμοποιηθεί η μείωση στην τιμή μιας μετρικής (πχ accuracy, sensitivity) μετά το ανακάτεμα των τιμών σε κάθε χαρακτηριστικό [128].

Η μέτρηση της permutation feature importance εισήχθη από τον Breiman το 2001 [95] για τον αλγόριθμο τυχαίου δάσους. Με βάση αυτή την ιδέα, οι Fisher, Rudin, and Dominici το 2018 [126] πρότειναν μια ανεξαρτήτως μοντέλου έκδοση της σημαντικότητας των χαρακτηριστικών και την ονόμασαν αξιοπιστία μοντέλου (model reliance).

7.3 Τοπικά ερμηνεύσιμες εξηγήσεις ανεξαρτήτως μοντέλου (LIME)



Εικόνα 17: *Απλή εφαρμογή του LIME σε ένα μη γραμμικό μοντέλο δυαδικής ταξινόμησης* [127]

Ένας σημαντικός εκπρόσωπος της κατηγορίας local model-agnostic είναι η μέθοδος Local Interpretable Model-agnostic Explanations (LIME) [127], η οποία χρησιμοποιείται για την εξήγηση μεμονωμένων προβλέψεων Η κύρια ιδέα του LIME είναι να εξηγήσει μια πρόβλεψη ενός σύνθετου μοντέλου f_M, π.χ. ενός βαθέως νευρωνικού δικτύου, εκπαιδεύοντας ένα τοπικό υποκατάστατο μοντέλο (local surrogate model) f_S, του οποίου οι προβλέψεις είναι εύκολο να εξηγηθούν.

Τεχνικά, το LIME δημιουργεί δείγματα στη γειτονιά N_{x_i} της εισόδου ενδιαφέροντος x_i, τα αξιολογεί χρησιμοποιώντας το μοντέλο-στόχο και στη συνέχεια το προσεγγίζει σε αυτήν την τοπική γειτονιά με μια απλή γραμμική συνάρτηση, δηλαδή ένα υποκατάστατο μοντέλο που είναι εύκολο να ερμηνευτεί. Έτσι, το LIME δεν εξηγεί άμεσα την πρόβλεψη του μοντέλου-στόχου $f_M(x_i)$, αλλά μάλλον τις προβλέψεις ενός υποκατάστατου μοντέλου $f_S(x_i)$, το οποίο προσεγγίζει τοπικά το μοντέλο-στόχο (δηλαδή, $f_M(x) \approx f_S(x)$ για $x \in N_{x_i}$) [129].

Στην Εικόνα 17 φαίνεται ένα απλό παράδειγμα εφαρμογής της μεθόδου. Η σύνθετη συνάρτηση απόφασης του black box model f_M, (άγνωστη στο LIME) αντιπροσωπεύεται από το μπλε/ροζ φόντο, το οποίο δεν μπορεί να προσεγγιστεί καλά από ένα γραμμικό μοντέλο. Ο έντονος κόκκινος σταυρός είναι το παράδειγμα που εξηγείται. Το LIME δειγματοληπτεί νέα στιγμιότυπα, λαμβάνει προβλέψεις χρησιμοποιώντας το f_M, και τις ζυγίζει με βάση την εγγύτητα στο στιγμιότυπο ενδιαφέροντος (που αναπαρίσταται εδώ κατά μέγεθος). Η διακεκομμένη γραμμή είναι η ερμηνεία του f_M, που είναι αξιόπιστη σε τοπικό επίπεδο (αλλά όχι συνολικά για όλο το μοντέλο).

Κεφάλαιο 8: Ανάπτυξη μοντέλου μηχανικής μάθησης για την πρόβλεψη εμφάνισης Στεφανιαίας Νόσου

Η υλοποίηση του μοντέλου της παρούσης εργασίας έγινε με τη χρήση της γλώσσας προγραμματισμού Python σε μορφή Jupiter notebook στην πλατφόρμα Colaboratory της Google. Χρησιμοποιήθηκαν επίσης βιβλιοθήκες όπως Scikit-learn, Pandas, Numpy και Seaborn.

8.1 Επισκόπηση και οπτικοποίηση δεδομένων

Τα δεδομένα της εργασίας προέρχονται από το Σύστημα Παρακολούθησης Συμπεριφορικών Παραγόντων Κινδύνου (Behavioral Risk Factor Surveillance System: BRFSS) του αμερικάνικου CDC. Το BRFSS είναι το κορυφαίο σύστημα τηλεφωνικών ερευνών σχετικά με την υγεία των κατοίκων των ΗΠΑ που συλλέγει δεδομένα, τα οποία αφορούν τις συμπεριφορές κινδύνου που σχετίζονται με την υγεία, τις χρόνιες παθήσεις και τη εφαρμογή προληπτικών μέτρων. Ολοκληρώνει περίπου 400.000 συνεντεύξεις ενηλίκων κάθε χρόνο, καθιστώντας το το μεγαλύτερο σύστημα ερευνών υγείας που λειτουργεί αδιάλειπτα σε όλο τον κόσμο.

Συγκεκριμένα, το dataset αποτελεί υποσύνολο του αρχικού dataset που προέκυψε το 2020 μέσω του αντίστοιχου ερωτηματολογίου [130] και περιλαμβάνει 319.795 εγγραφές (άτομα) και 18 χαρακτηριστικά από τα συνολικά 279 (feature selection), που σχετίζονται έμμεσα ή άμεσα με την καρδιαγγειακή υγεία. Ο πίνακας 4 δείχνει την πλήρη επεξήγηση κάθε

χαρακτηριστικού και τις δυνατές τιμές του. Τα δεδομένα είναι αποθηκευμένα σε μορφή αρχείου με τιμές χωρισμένες με κόμμα (comma separated values CSV format).

Πίνακας 4: *Περιγραφή χαρακτηριστικών και οι διαφορετικές τιμές που λαμβάνουν*

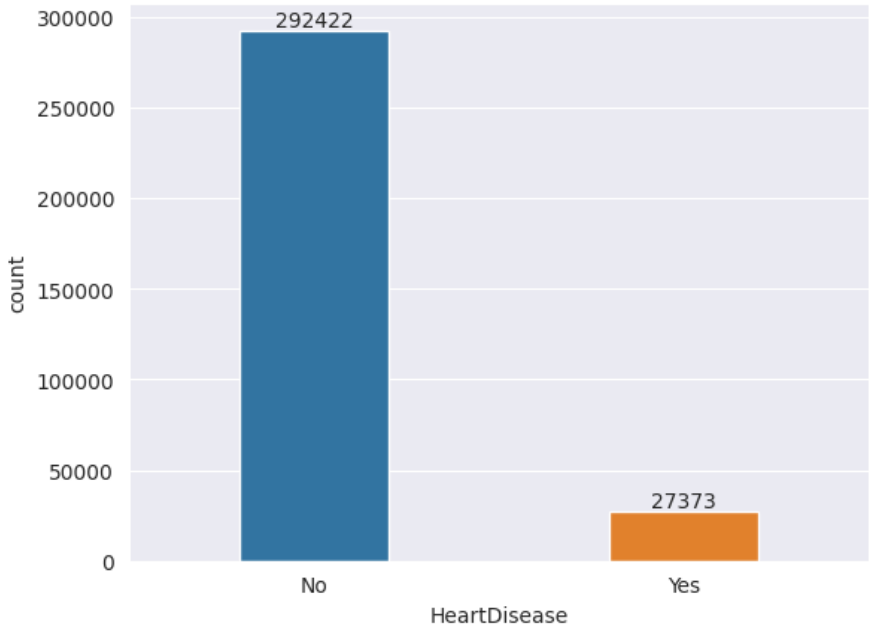
A/A	Περιγραφή Χαρακτηριστικών	Όνομα Χαρακτηριστικού	Τύπος Χαρακτηριστικού	Διαφορετικές Τιμές
0	΄Υπαρξη στηθάγχης ή στεφανιαίας νόσου.	HeartDisease	Κατηγορικό	Ναι, Όχι
1	Δείκτης Μάζας Σώματος (βάσει σχέσης 1)	BMI	Αριθμητικό	Δεκαδικές τιμές από 12.02 έως 94.85
2	Κάπνισμα: Ναι εάν κάποιος έχει καπνίσει τουλάχιστον 100 τσιγάρα (5 πακέτα) στη ζωή του. Διαφορετικά όχι	Smoking	Κατηγορικό	Ναι, Όχι
3	Κατανάλωση αλκοόλ: Ναι εάν κάποιος έχει καταναλώσει τουλάχιστον ένα ποτό (7-14 gr. αιθυλικής αλκοόλης) τις τελευταίες 30 ημέρες.	AlcoholDrinking	Κατηγορικό	Ναι, Όχι
4	Εάν το άτομο έχει περάσει εγκεφαλικό επεισόδιο	Stroke	Κατηγορικό	Ναι, Όχι
5	Αριθμός ημερών τον τελευταίο μήνα, στις οποίες δεν ήταν καλή η σωματική υγεία (δηλαδή υπήρχε σωματική ασθένεια ή τραυματισμός)	PhysicalHealth	Αριθμητικό	Ακέραιες τιμές από 0 έως 30
6	Αριθμός ημερών τον τελευταίου μήνα, στις οποίες δεν ήταν καλή η ψυχική υγεία, (δηλαδή υπήρχε άγχος, κατάθλιψη ή συναισθηματικά προβλήματα)	MentalHealth	Αριθμητικό	Ακέραιες τιμές από 0 έως 30
7	΄Υπαρξη δυσκολίας στο περπάτημα ή στο ανέβασμα σκάλας	DiffWalking	Κατηγορικό	Ναι, Όχι
8	Βιολογικό φύλλο	Sex	Κατηγορικό	Άνδρας, Γυναίκα
9	Προσδιορισμός ηλικίας	AgeCategory	Κατηγορικό Διατεταγμένο	13 ηλικιακές ομάδες: 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80 or older
10	Προσδιορισμός φυλής	Race	Κατηγορικό	Λευκός, Μαύρος, Ασιάτης, Αμερικανός Ινδιάνος/ Ιθαγενής από την Αλάσκα, Ισπανόφωνος, Άλλο
11	΄Υπαρξη διαβήτη, οριακού διαβήτη ή διαβήτη κατά τη διάρκεια εγκυμοσύνης.	Diabetic	Κατηγορικό Διατεταγμένο	Ναι, Όχι, Όχι (οριακός διαβήτης), Ναι (κατά τη διάρκεια εγκυμοσύνης)

12	Συμμετοχή κατά τις τελευταίες 30 μέρες σε άλλες φυσικές δραστηριότητες, εκτός της τακτικής εργασίας, όπως τρέξιμο, καλλισθενική γυμναστική, γκολφ, κηπουρική ή περπάτημα για άσκηση.	PhysicalActivity	Κατηγορικό	Ναι, Όχι
13	Πως οι ερωτηθέντες αυτοπροσδιορίζουν γενικά την υγεία τους	GenHealth	Κατηγορικό Διατεταγμένο	Κακή, Μέτρια, Καλή, Πολύ Καλή, Εξαιρετική
14	Ώρες ύπνου ενός 24ώρου κατά μέσο όρο.	SleepTime	Αριθμητικό	Ακέραιες τιμές από 1 έως 24
15	Ύπαρξη αλλεργικού άσθματος.	Asthma	Κατηγορικό	Ναι, Όχι
16	Πρόβλημα νεφρικής λειτουργίας. Δεν περιλαμβάνονται η νεφρολιθίαση, η ουρολοίμωξη και η ακράτεια.	KidneyDisease	Κατηγορικό	Ναι, Όχι
17	Ύπαρξη στο ιστορικό καρκίνου του δέρματος	SkinCancer	Κατηγορικό	Ναι, Όχι

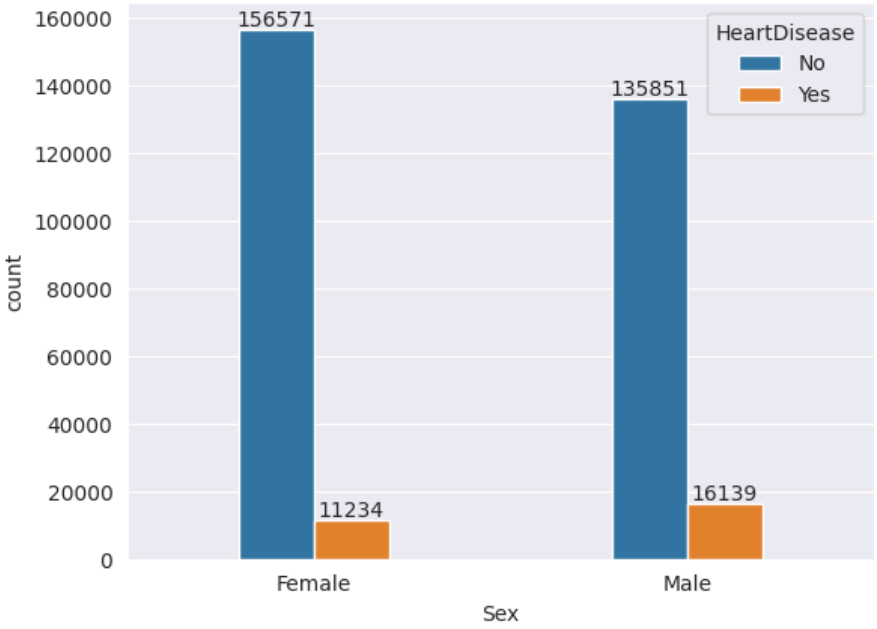
Το χαρακτηριστικό 0 (HeartDisease), δηλαδή η ύπαρξη στεφανιαίας νόσου αποτελεί και τη μεταβλητή ταξινόμησης (target variable). Από το σύνολο των 319.795 στιγμιοτύπων, τα 292.422 ήταν αρνητικά για στεφανιαία νόσο, ενώ 27.373 ήταν θετικά, δηλαδή τα θετικά αποτελούσαν το 9,4% του συνόλου. Είναι εμφανές δηλαδή ότι το dataset είναι μη ισορροπημένο. Η συχνότητα καρδιακής νόσου τόσο στο σύνολο όσο και ανά φύλο φαίνεται στον Πίνακα 5 και απεικονίζεται στις Εικόνες 18 και 19. Παρατηρείται ότι το 7% των γυναικών και το 11% των ανδρών ήταν θετικοί.

Πίνακας 5: *Συχνότητα εμφάνισης στεφανιαίας νόσου συνολικά και ανά φύλο*

Στεφανιαία Νόσος	**Φύλο**		**Σύνολο**
	Άνδρες	**Γυναίκες**	
Ναι	16.139	11.234	27.373
Όχι	135.851	156.571	292.422
Σύνολο	151.990	167.805	319.795



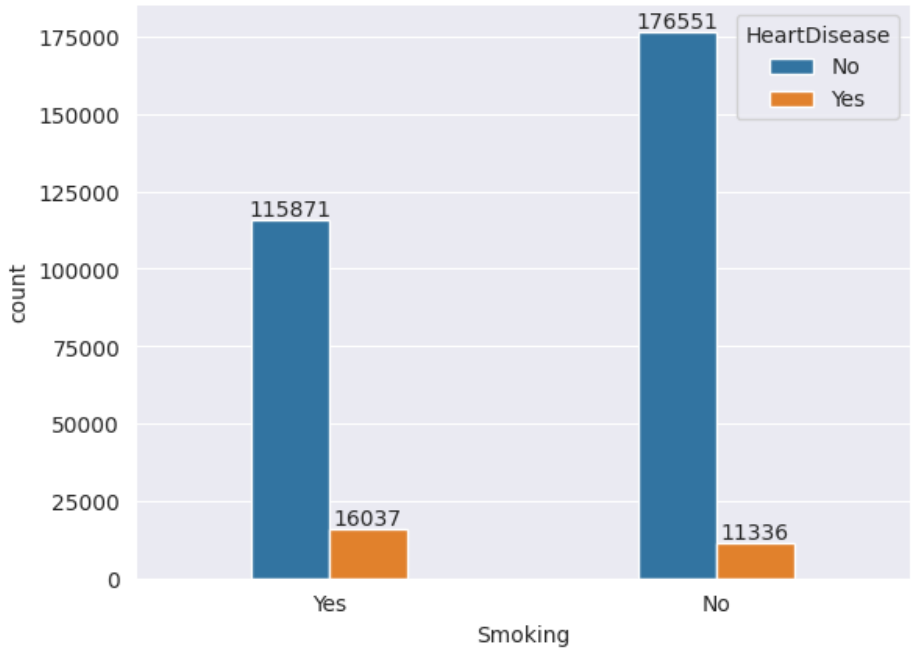
Εικόνα 18: Πλήθος δειγμάτων ανά κλάση



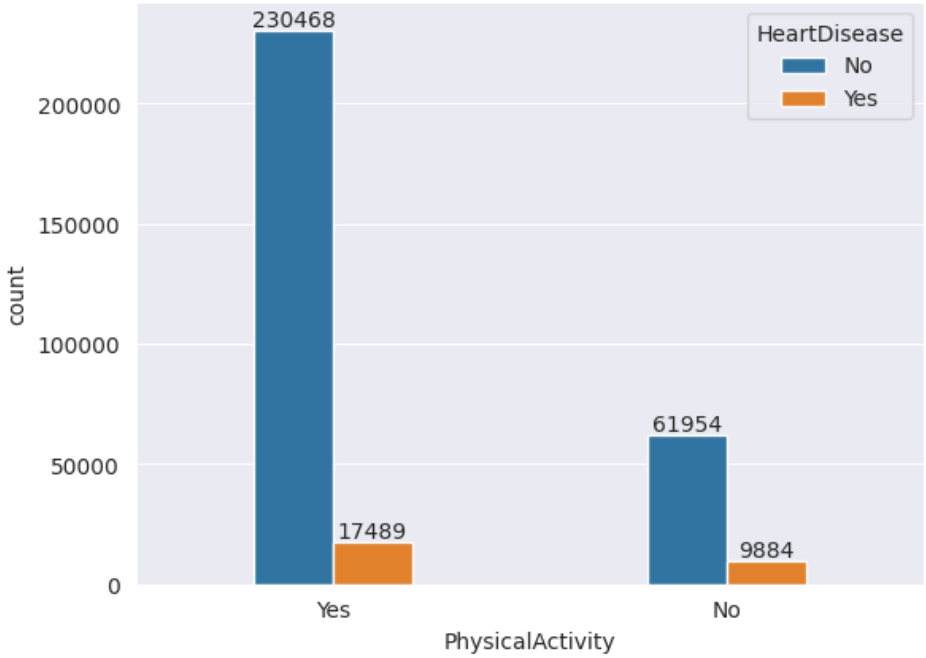
Εικόνα 19: Πλήθος δειγμάτων ανά κλάση και φύλο

Στις Εικόνες 20, 21, 22, 23, 24, 25 και 26 απεικονίζεται η συσχέτιση της στεφανιαίας νόσου με το κάπνισμα, τη φυσική δραστηριότητα, το διαβήτη, την ηλικία, την κατανάλωση αλκοόλ το ιστορικό εγκεφαλικού επεισοδίου και τη φυλή αντίστοιχα. Παρατηρείται ότι το 12% των καπνιζόντων ήταν πάσχοντες, ενώ το αντίστοιχο ποσοστό των μη καπνιζόντων ήταν 6% δηλαδή το μισό. Σχετικά με τη φυσική δραστηριότητα, από αυτούς που δεν ασκούνταν έπασχε το 14%, ενώ από αυτούς που ασκούνταν το 7% δηλαδή το μισό. Όσον αφορά το διαβήτη οι διαβητικοί έπασχαν σε ποσοστό 22%, ενώ οι μη διαβητικοί σε 6%. Υπάρχει επίσης αύξηση των πασχόντων αυξανομένης της ηλικίας. Αντίθετα παρατηρείται ότι όσοι δεν έπιναν καθόλου αλκοόλ έπασχαν σε ποσοστό 10%, ενώ όσοι έπιναν τουλάχιστον ένα

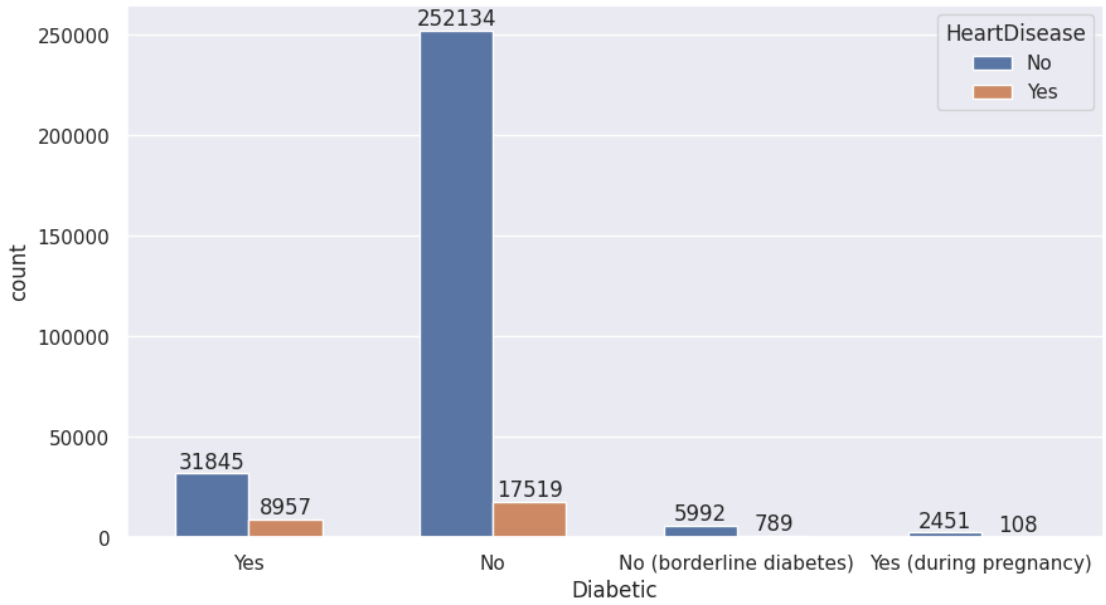
ποτό το μήνα σε ποσοστό 6%. Όσοι είχαν ιστορικό εγκεφαλικού επεισοδίου είχαν και στεφανιαία νόσο σε ποσοστό 36%, ενώ όσοι δεν είχαν ιστορικό μόνο 7.5%. Τέλος, ως προς τη φυλή, οι Ινδιάνοι και οι Ιθαγενείς της Αλάσκας είχαν ΣΝ σε ποσοστό 10.4%, οι λευκοί 9.2%, οι μαύροι 7.5%, οι ισπανόφωνοι 5.3% και οι ασιάτες 3.3%.



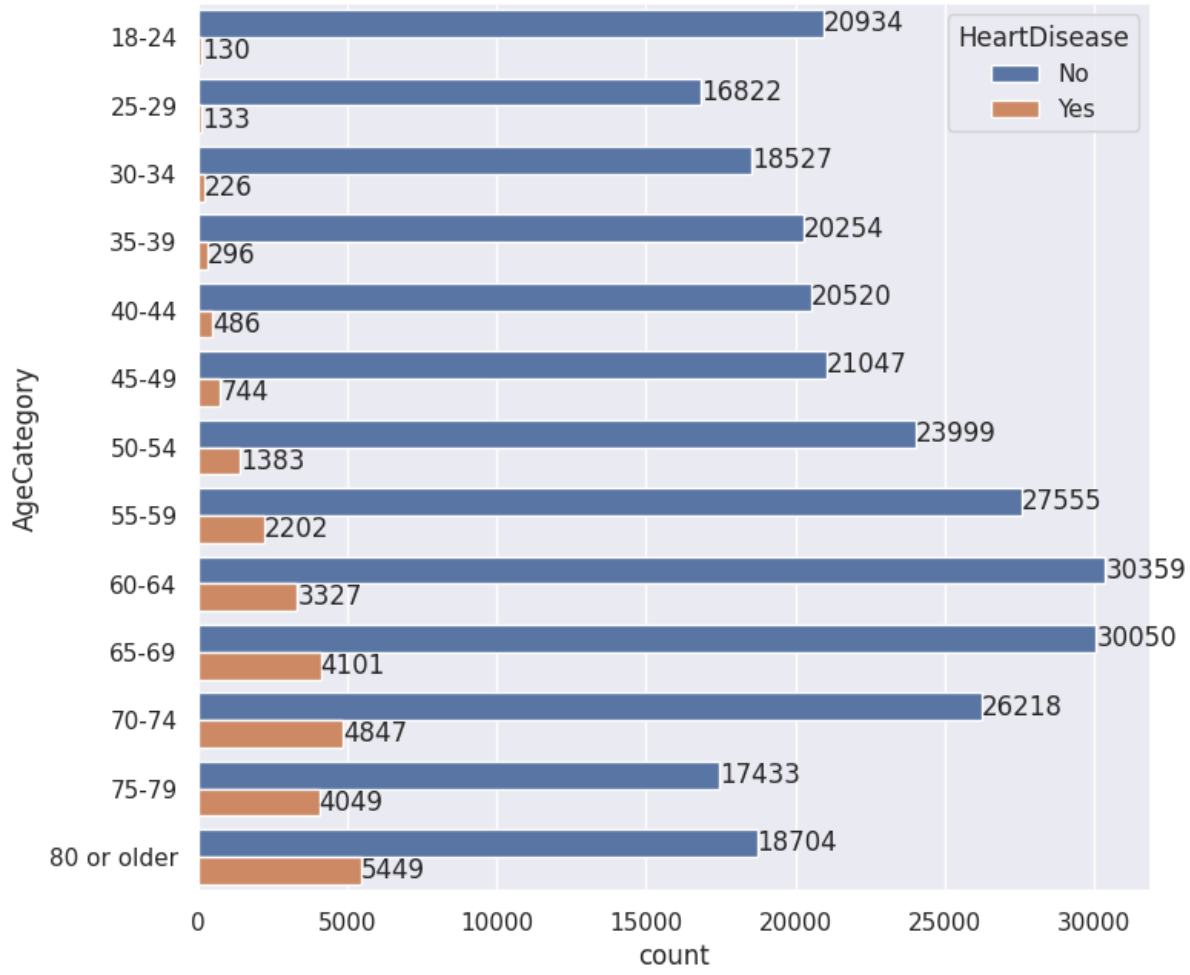
Εικόνα 20: Συσχέτιση στεφανιαίας νόσου και καπνίσματος



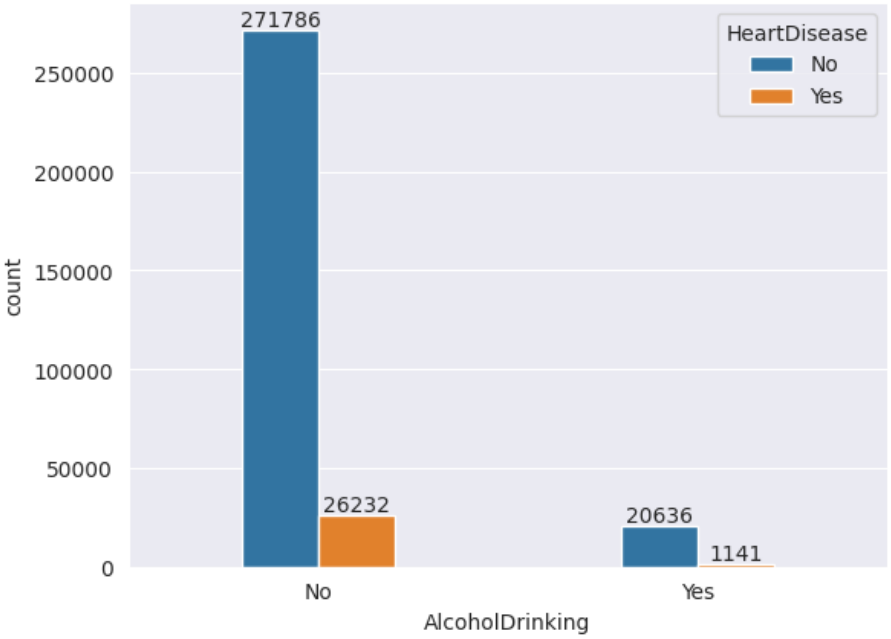
Εικόνα 21: Συσχέτιση στεφανιαίας νόσου και φυσικής δραστηριότητας



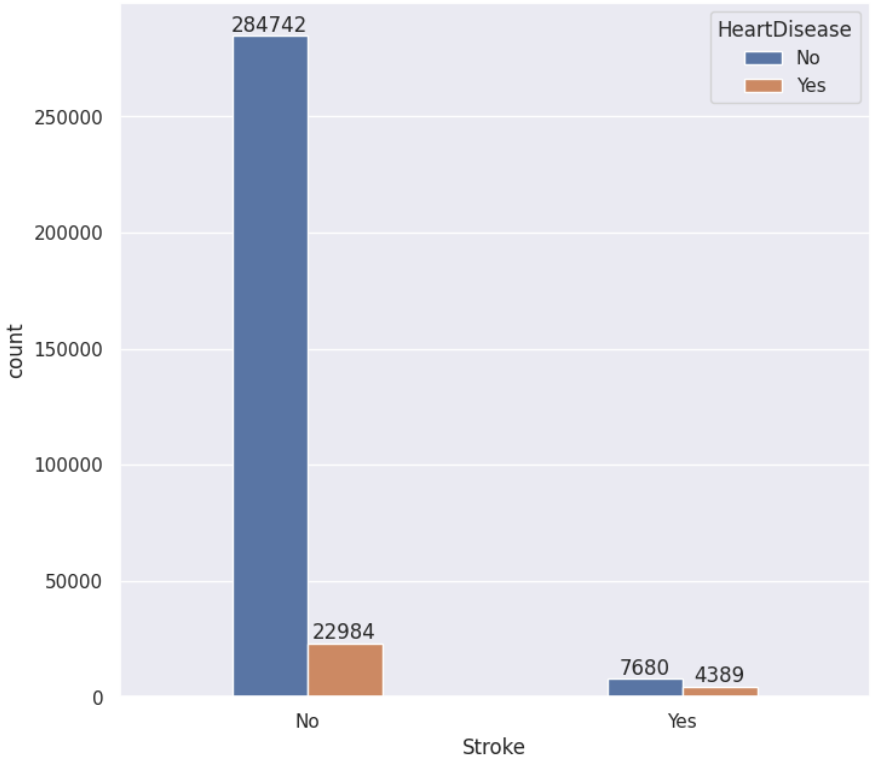
Εικόνα 22: Συσχέτιση στεφανιαίας νόσου και διαβήτη



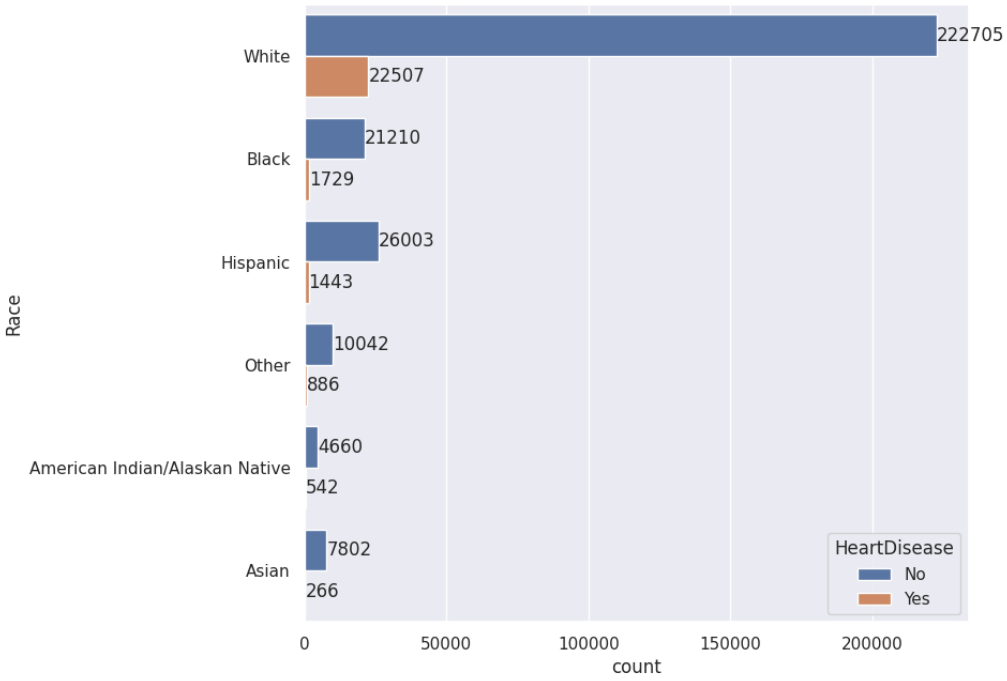
Εικόνα 23: Συσχέτιση στεφανιαίας νόσου και ηλικιακών ομάδων



Εικόνα 24: Συσχέτιση στεφανιαίας νόσου και κατανάλωσης αλκοόλ



Εικόνα 25: Συσχέτιση στεφανιαίας νόσου και εγκεφαλικού επεισοδίου



Εικόνα 26: *Συσχέτιση στεφανιαίας νόσου και φυλής*

Κατά τον έλεγχο του dataset δεν βρέθηκαν απουσιάζουσες ή ανακόλουθες τιμές.

Ο έλεγχος για διπλότυπες εγγραφές στο dataset έδειξε 29.930 διπλότυπες εγγραφές. Αυτές αφορούσαν 11.852 εγγραφές που επαναλαμβάνονταν μία ή περισσότερες φορές και δεν εμφανίζονταν σε συνεχόμενες σειρές. Επειδή τα δεδομένα είχαν συλλεχθεί από αξιόπιστη πηγή (CDC) με αυστηρό πρωτόκολλο δειγματοληψίας μέσω τηλεφωνικών αριθμών και ανά περιοχή, θεωρήθηκαν πραγματικές εγγραφές που αντιστοιχούν σε διαφορετικά άτομα και γι' αυτό δεν αφαιρέθηκαν.

8.2 Προεπεξεργασία δεδομένων (Preprocessing)

Η διαδικασία που ακολουθήθηκε για την προεπεξεργασία των δεδομένων περιλαμβάνει τα παρακάτω στάδια.

8.2.1 Διαχείριση κατηγορικών χαρακτηριστικών

Σχετικά με τα κατηγορικά χαρακτηριστικά εφαρμόστηκαν οι εξής μέθοδοι:

- **Ordinal Encoding** για τα χαρακτηριστικά AgeCategory, GenHealth και Diabetic.
- **Dummy Encoding** για τα χαρακτηριστικά HeartDisease (μεταβλητή ταξινόμησης), Smoking, AlcoholDrinking, Stroke, DiffWalking, PhysicalActivity, Asthma, KidneyDisease, SkinCancer και Sex.
- **One-Hot-Encoding** για τη μεταβλητή Race και ύστερα αφαίρεση της στήλης 'Race_Other' για αποφυγή της multicollinearity. Αναλυτικότερα παρουσιάζονται στον Πίνακα 6 ανά χαρακτηριστικό.

Πίνακας 6: Κωδικοποίηση κατηγορικών μεταβλητών

Χαρακτηριστικό	Κωδικοποίηση
AgeCategory	18-24→1, 25-29→2, 30-34→3 , 35-39→4 , 40-44→5, 45-49→6, 50-54→7 , 55-59→8 , 60-64→9, 65-69→10, 70-74→11, 75-79→12, 80 or older→13
GenHealth	Excellent→ 1, Very good→ 2, Good→ 3, Fair→4, Poor→5
Diabetic	No→0, Yes (during pregnancy)→1, No (borderline diabetes)→2, Yes→3
HeartDisease	Yes→ 1, No→ 0
Smoking	
AlcoholDrinking	
Stroke	
DiffWalking	
PhysicalActivity	
Asthma	
KidneyDisease	
SkinCancer	
Sex	Female→0 Male→1
Race	Πέντε στήλες Race_American Indian/Alaskan Native, Race_Asian, Race_Black, Race_Hispanic, Race_White με 1 στην αντίστοιχη στήλη

8.2.2 Χωρισμός σε training και test set

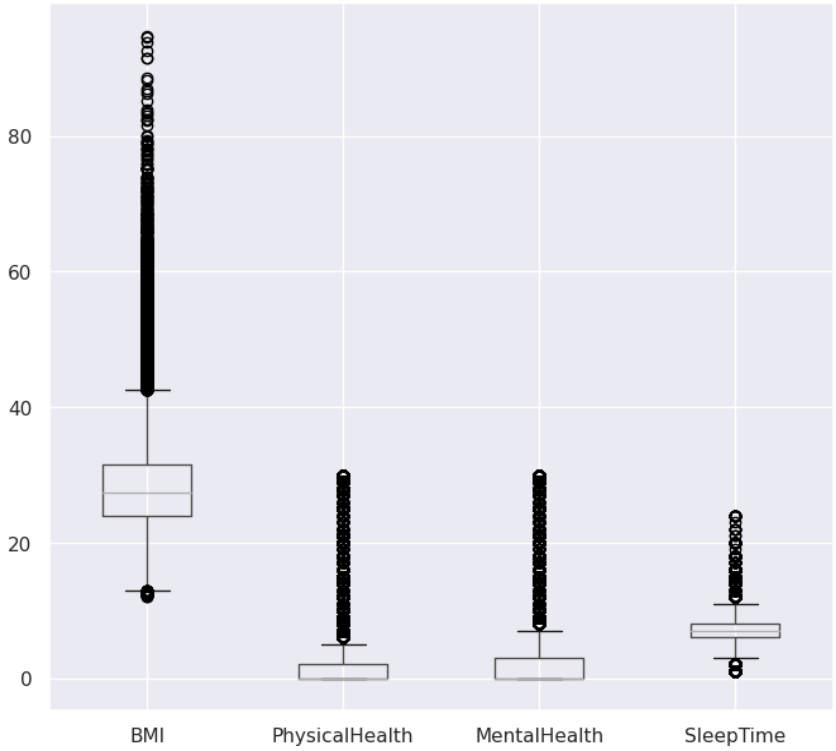
Ακολουθεί χωρισμός του dataset σε training και test set πριν οποιαδήποτε άλλη επεξεργασία, ώστε να μην υπάρξει διάχυση πληροφορίας από το test set στο training set και να αποφευχθεί η υπερεκπαίδευση του εκάστοτε εκτιμητή. Ο διαχωρισμός έγινε με αναλογία training set 70% και test set 30% και διατηρήθηκε η αναλογία των 2 κλάσεων και στα δύο set (Πίνακας 7).

Πίνακας 7: Πλήθος δειγμάτων ανά κλάση στα training και test set

	HeartDisease		Σύνολο
	Yes	**No**	
Training Set	19.082	204.774	223.856
Test Set	8.291	87.648	95.939

8.2.3 Διαχείριση ακραίων τιμών

Για τη διαχείριση των ακραίων τιμών εφαρμόστηκε η μέθοδος του Ενδοτεταρτημοριακού Εύρους για τα αριθμητικά χαρακτηριστικά BMI, PhysicalHealth, MentalHealth, SleepTime. Στην Εικόνα 27 παρουσιάζεται το boxplot για το training set.



Εικόνα 27: Boxplot των αριθμητικών χαρακτηριστικών του training set

Στον Πίνακα 8 φαίνονται οι ακραίες τιμές για αυτές τις μεταβλητές.

Πίνακας 8: Outliers για τα αριθμητικά χαρακτηριστικά στο training set

Χαρακτηριστικά	Mild Outliers		Extreme Outliers	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
BMI	12,90	42,58	1,77	53,71
PhysicalHealth	0	5	0	8
MentalHealth	0	7,5	0	12
SleepTime	3	11	0	14

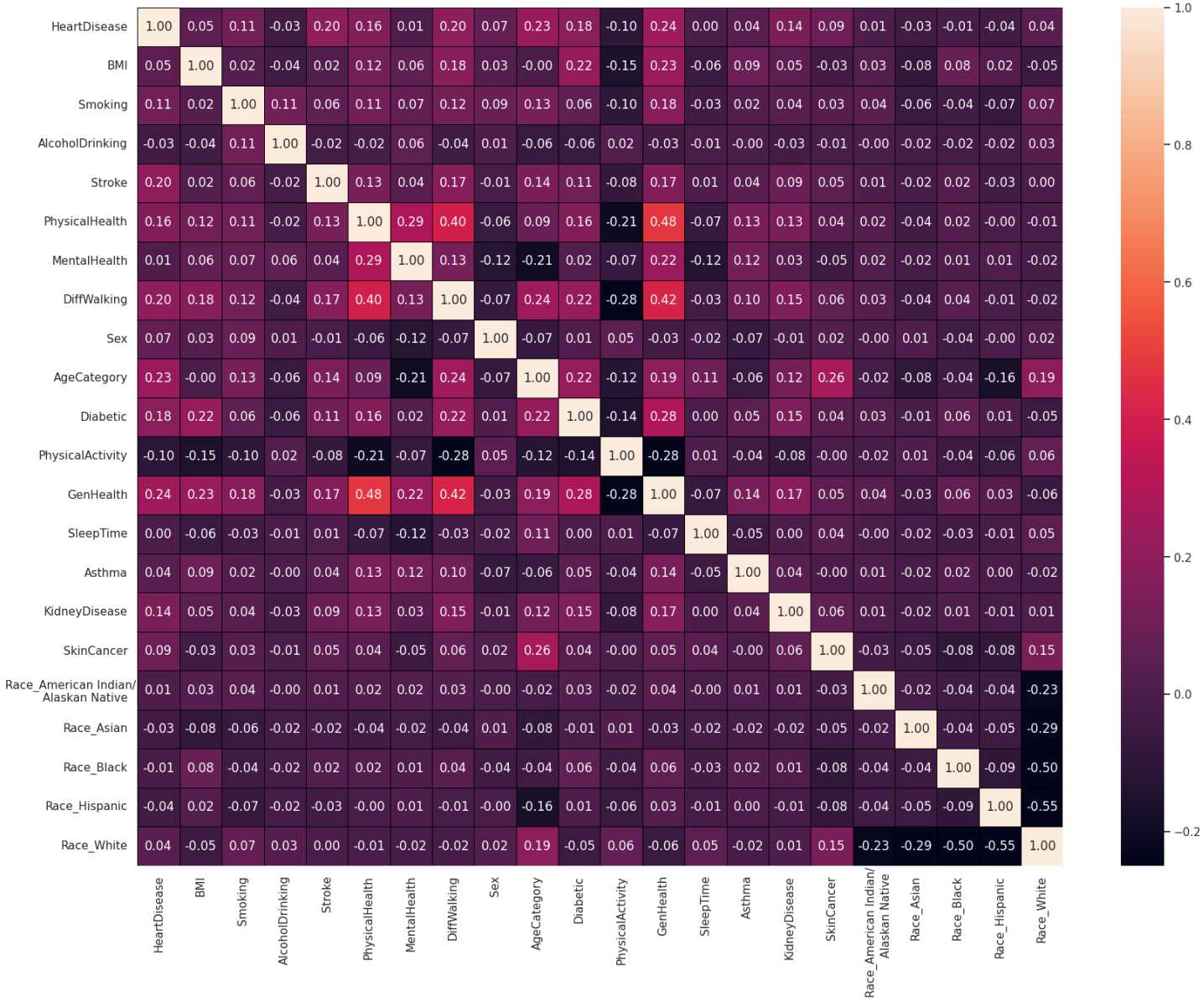
Στα δείγματα του training set που αυτά τα χαρακτηριστικά είχαν μεγαλύτερες τιμές από τα άνω extreme outliers αντικαταστάθηκαν με την άνω ακραία τιμή, ενώ αυτά που είχαν μικρότερες από τα κάτω extreme outliers αντικαταστάθηκαν με την κάτω ακραία τιμή. Με τη χρήση των outliers του training set τροποποιήθηκε και το test set.

8.2.4 Κανονικοποίηση

Για την κανονικοποίηση των τιμών των χαρακτηριστικών χρησιμοποιήθηκαν οι τεχνικές Min Max Scaling και Standard scaling για να ελεγχθεί ποια από τις δύο θα οδηγούσε σε καλύτερα αποτελέσματα. Τελικά επελέγη η τεχνική Min Max Scaling. Με τη χρήση των αντίστοιχων μετασχηματιστών που κανονικοποίησαν το training set τροποποιήθηκε και το test set.

8.2.5 Υπολογισμός συσχέτισης χαρακτηριστικών

Στην Εικόνα 28 απεικονίζεται ο heat map με τις τιμές της συσχέτισης Pearson των χαρακτηριστικών μαζί με την μεταβλητή ταξινόμησης Heart Disease στο training set. Παρατηρείται ότι δεν υπάρχει τιμή μεγαλύτερη από 0,7 ή μικρότερη από -0,7.



Εικόνα 28: *Heat map της συσχέτισης των χαρακτηριστικών του training set, συμπεριλαμβανομένης της μεταβλητής ταξινόμησης*

8.2.6 Διαχείριση των μη ισορροπημένων κλάσεων του training set

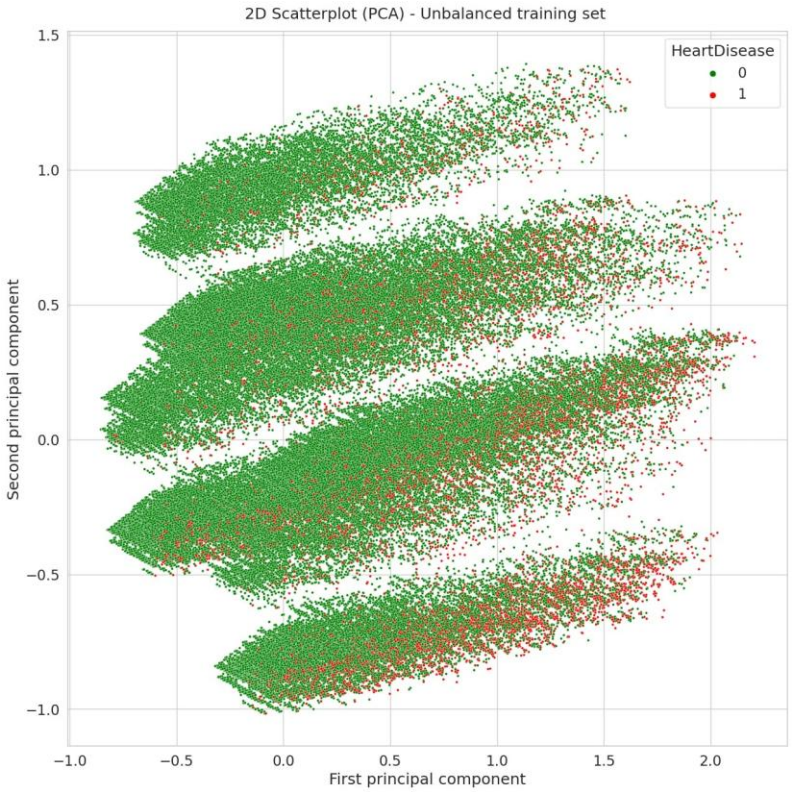
Όπως ήδη αναφέρθηκε, το dataset (άρα και το training set) είναι μη ισορροπημένο με την αρνητική κλάση να εκπροσωπείται με δεκαπλάσιο περίπου αριθμό στιγμιοτύπων. Για την εξισορρόπηση του πλήθους των στιγμιοτύπων του training set δοκιμάστηκαν οι μέθοδοι Random Oversampling, SMOTE, ADASYN, Random Undersampling και Near Miss (Πίνακας 9). Επίσης δοκιμάστηκε η cost-sensitive μέθοδος της τροποποίησης της υπερπαραμέτρου class _weight στους αλγόριθμους που το υποστηρίζουν.

Πίνακας 9: Πλήθος δειγμάτων κλάσεων στο training set μετά την εξισορρόπηση

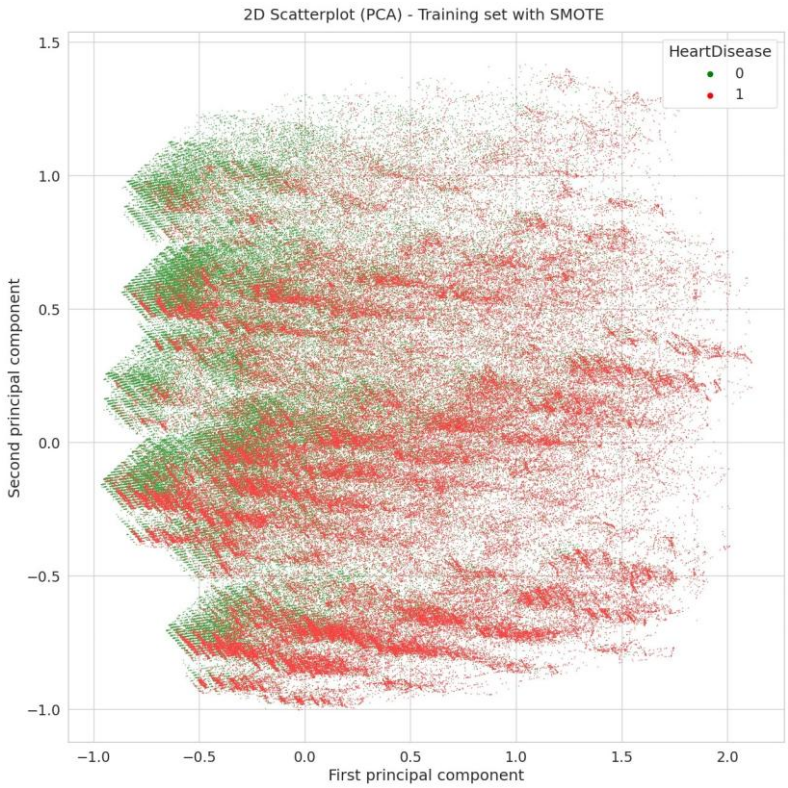
Μέθοδος εξισορρόπησης	Πλήθος δειγμάτων κλάσης 1 (Training set)	Πλήθος δειγμάτων κλάσης 0 (Training set)
Random Oversampling	204.774	204.774
SMOTE		
ADASYN	208.088	204.774
Random Undersampling	19.082	19.082
Near Miss		

8.2.7 Οπτικοποίηση training set με χρήση της μεθόδου PCA

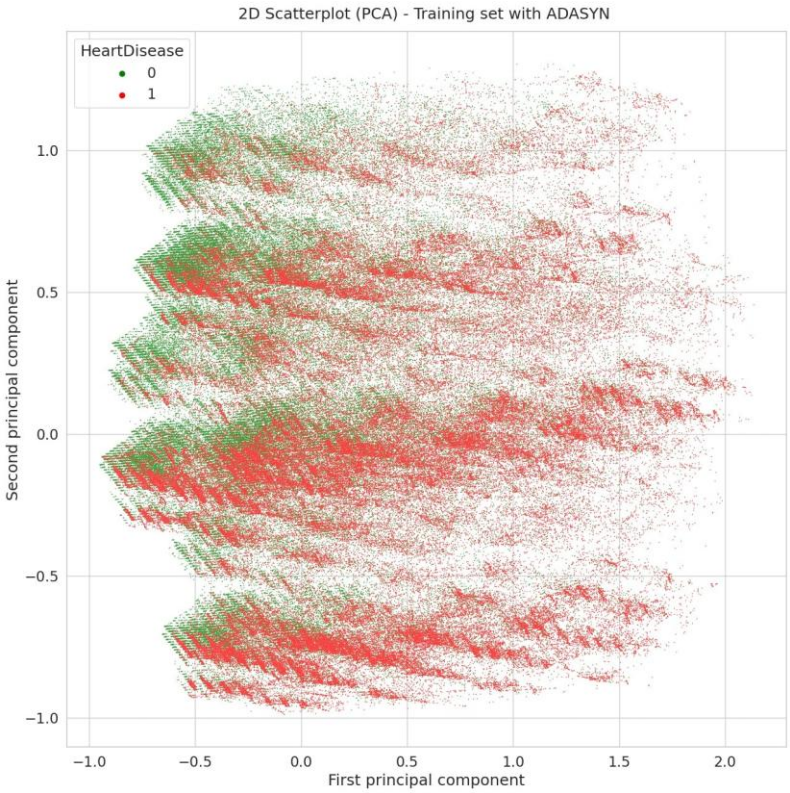
Στις Εικόνες 29, 30, 31 και 32 εφαρμόστηκε η μέθοδος PCA με σκοπό την οπτικοποίηση των στιγμιοτύπων του μη ισορροπημένου training set και των training set που προέκυψαν με τις μεθόδους εξισορρόπησης SMOTE, ADASYN και Random Undersampling.



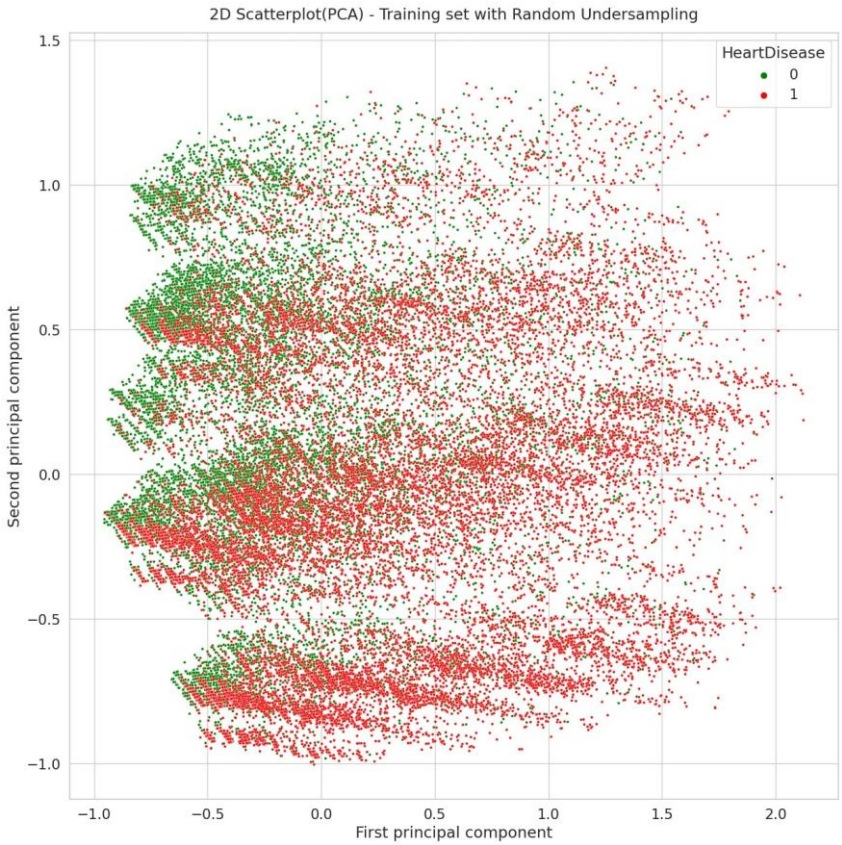
Εικόνα 29: 2D Scatterplot των στιγμιοτύπων του μη ισορροπημένου training set με τη μέθοδο PCA (2 - principal components)



Εικόνα 30: *2D Scatterplot των στιγμιοτύπων του ισορροπημένου training set με SMOTE (PCA - 2 principal components)*



Εικόνα 31: *2D Scatterplot των στιγμιοτύπων του ισορροπημένου training set με ADASYN (PCA - 2 principal components)*



Εικόνα 32: *2D Scatterplot των στιγμιοτύπων του ισορροπημένου training set με Random Undersampling (PCA -2 principal components)*

Κεφάλαιο 9: Αποτελέσματα και συζήτηση

Το σημαντικότερο πρόβλημα στην ανάπτυξη μοντέλου πρόβλεψης εμφάνισης στεφανιαίας νόσου με το παρόν σύνολο δεδομένων ήταν η ανισορροπία στις δύο κλάσεις, με την αρνητική κλάση να έχει δεκαπλάσια αντιπροσώπευση σε σχέση με τη θετική. Αυτό οφείλεται στο χαμηλό επιπολασμό της στεφανιαίας νόσου στον γενικό πληθυσμό από όπου προέρχονται τα δεδομένα. Εξαιτίας αυτού του γεγονότος τα μοντέλα που εκπαιδεύονταν σε μη ισορροπημένα δεδομένα εμφάνιζαν ιδιαίτερα χαμηλές τιμές στην ευαισθησία, δηλαδή στον εντοπισμό θετικών περιπτώσεων σε νέα άγνωστα δεδομένα. Αυτό αντιμετωπίστηκε με χρήση διαφόρων τεχνικών που ήδη αναφέρθηκαν.

Στο πλαίσιο της επιλογής του καταλληλότερου αλγορίθμου μηχανικής μάθησης για τα συγκεκριμένα δεδομένα δοκιμάστηκαν οι εξής ταξινομητές (Classifiers): KNN, Naive Bayes, Logistic Regression, Decision Tree, Random Forest και MLP. Κάθε αλγόριθμος εκπαιδεύεται κάθε φορά με τα αντίστοιχα training set που προέκυψαν από τις διαφορετικές τεχνικές αντιμετώπισης του προβλήματος της ανισορροπίας των κλάσεων κατά την προεπεξεργασία και ύστερα αξιολογούνται οι προβλέψεις του στο test set.

Ως μετρική βελτιστοποίησης επελέγη η μετρική sensitivity που εστιάζει στο ποσοστό των True Positive από τα συνολικά θετικά που υπήρχαν στο test set, γιατί το μοντέλο πρέπει να είναι αποτελεσματικό στο να προβλέπει κυρίως εάν κάποιος είναι πιθανό να εμφανίσει στεφανιαία νόσο εξαιτίας του τρόπου ζωής του και άλλων συνυπαρχουσών νόσων που δε σχετίζονται άμεσα με την καρδιά. Παράλληλα, ελήφθη υπ' όψιν η διατήρηση σε υψηλά επίπεδα και της τιμής της specificity, ώστε να μην είναι μεγάλος ο αριθμός των ψευδώς

θετικών προβλέψεων. Η αύξηση του sensitivity γίνεται με τίμημα τη μείωση της precision, αφού είναι μετρικές αντιστρόφως ανάλογες μεταξύ τους.

Για κάθε αλγόριθμο παρουσιάζεται συγκεντρωτικός πίνακας αξιολόγησης των προβλέψεων ταξινόμησης των δειγμάτων του test set με τις κατάλληλες μετρικές, αφού έχει προηγηθεί εκπαίδευση με τις default τιμές των υπερπαραμέτρων του εκτιμητή. Στο training set, από το οποίο προκύπτουν τα καλύτερα αποτελέσματα, γίνεται βελτιστοποίηση των υπερπαραμέτρων του κάθε αλγορίθμου με χρήση της διασταυρούμενης επικύρωσης. Στο καλύτερο μοντέλο για κάθε περίπτωση παρουσιάζεται ο πίνακας σύγχυσης, οι καμπύλες ROC και οι καμπύλες εκπαίδευσης (Learning Curves). Ως μετρική για τον υπολογισμό των Learning Curves στα training set επιλέχθηκε η μετρική recall macro, που αξιολογεί την εκπαίδευση και στις δύο κλάσεις.

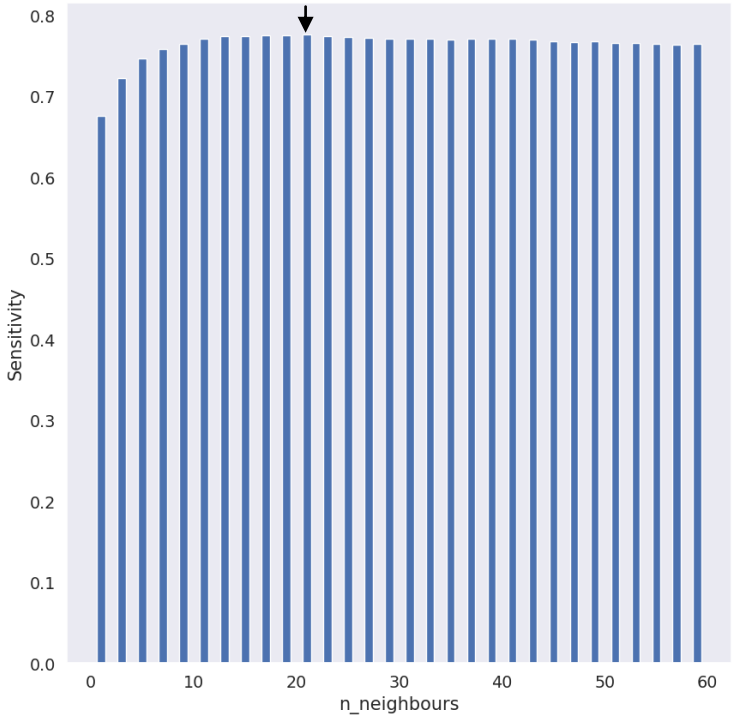
9.1 Μοντέλο με K-Nearest Neighbors

Πίνακας 10: *Αξιολόγηση προβλέψεων του KNN στο test set, μετά από εκπαίδευση με χρήση διαφόρων μεθόδων αντιμετώπισης της ανισορροπίας στις κλάσεις και βελτιστοποίηση*

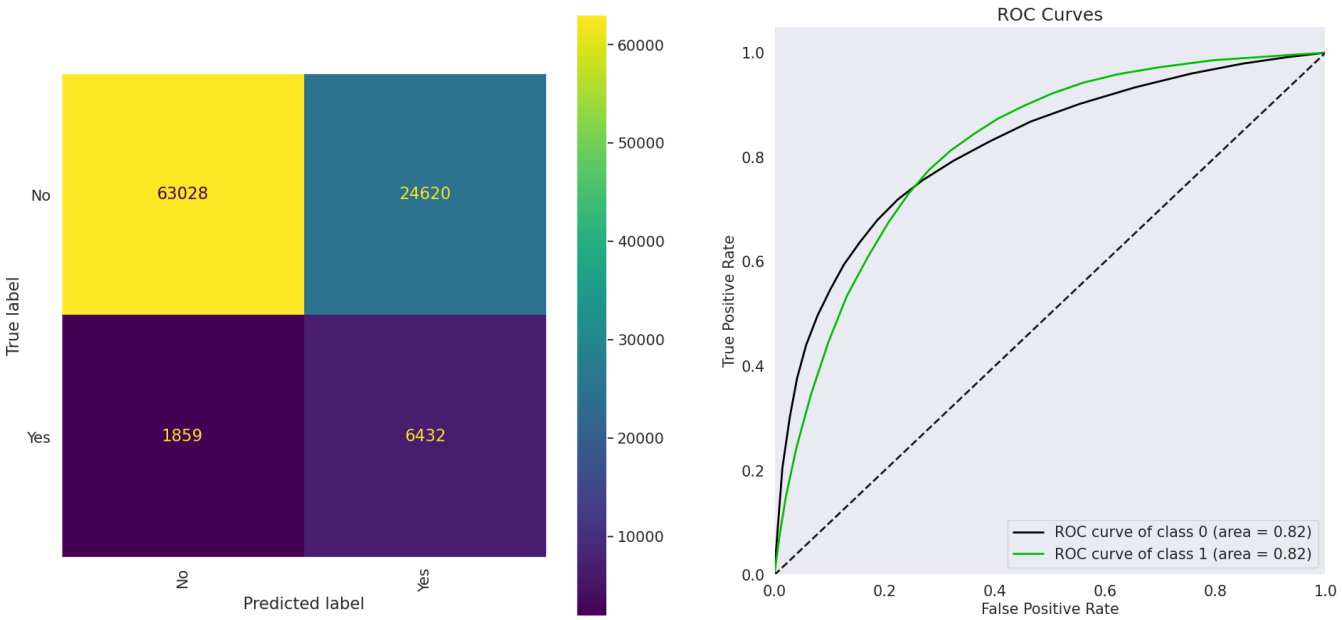
Μέθοδος αντιμετώπισης ανισορροπίας	Sensitivity	Specificity	Accuracy	AUC(ROC) score
Χωρίς (Unbalanced)	0.14	0.98	0.91	0.71
Random OverSampling	0.49	0.83	0.80	0.70
SMOTE	0.54	0.80	0.78	0.72
ADASYN	0.55	0.79	0.77	0.72
Random Undersampling	0.75	0.71	0.71	0.78
Near Miss	0.84	0.24	0.29	0.55
Βελτιστοποίηση: Random Undersampling με n_neighbors = 21	0.78	0.72	0.72	0.82

Στον Πίνακα 10 παρατηρείται ότι το μοντέλο που είχε εκπαιδευτεί στο μη ισορροπημένο training set έδωσε την χαμηλότερη τιμή sensitivity (0.14), ενώ την ψηλότερη (0.84) αυτό που προέκυψε από τη μέθοδο υποδειγματοληψίας Near Miss. Εξαιτίας όμως των πολύ χαμηλών τιμών specificity (0.24) και accuracy (0.29) με την Near Miss, επιλέχθηκε η μέθοδος τυχαίας υποδειγματοληψίας για να γίνει η βελτιστοποίηση της υπερπαραμέτρου n_neighbors (κοντινότεροι γείτονες), η οποία είχε τιμές sensitivity 0.75, specificity και accuracy 0.71 αλλά και AUC(ROC) score 0.78.

Η βελτιστοποίηση με 5-fold cross validation στο train set έδωσε βέλτιστη τιμή n_neighbors= 21 (default τιμή 5) με μέση τιμή ευαισθησίας 0.78 (Εικόνα 33). Με εφαρμογή n_neighbors= 21 στο test set προέκυψε sensitivity 0.78, specificity 0.72, accuracy 0.72 και AUC(ROC) score 0.82. Στην Εικόνα 34 παρουσιάζονται ο πίνακας σύγχυσης και οι καμπύλες ROC του βελτιστοποιημένου μοντέλου.

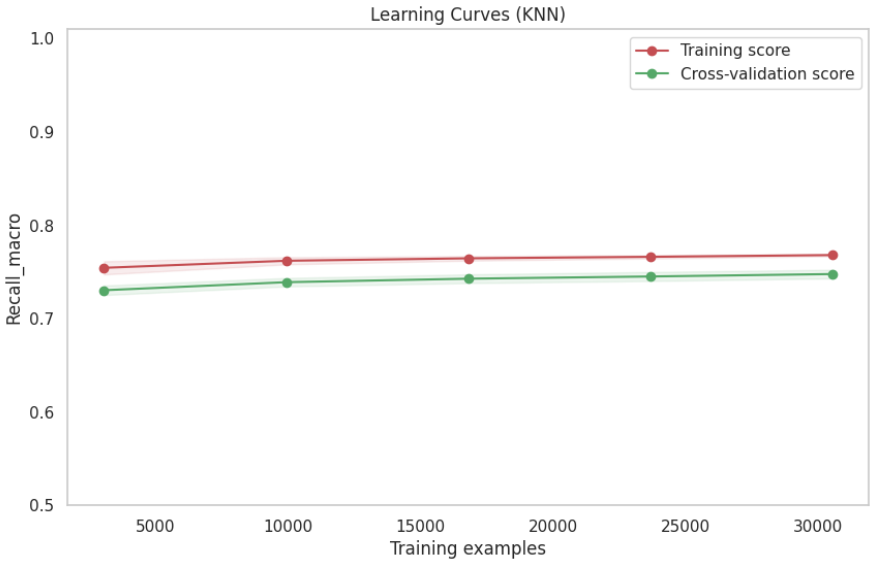


Εικόνα 33: *Οι τιμές της sensitivity στις διαφορετικές τιμές της υπερπαραμέτρου n_neighbors του KNN (βέλτιστη τιμή n_neighbors=21)*



Εικόνα 34: *Confusion matrix και ROC curves της αξιολόγησης του βελτιστοποιημένου μοντέλου με τον KNN στο test set*

Στην Εικόνα 35 έχουν σχεδιαστεί οι καμπύλες εκπαίδευσης του βέλτιστου μοντέλου με τον αλγόριθμο KNN. Παρατηρείται ότι δεν υπάρχει υπερεκπαίδευση, καθώς οι δύο καμπύλες συγκλίνουν με την αύξηση των δειγμάτων εκπαίδευσης και μειώνεται η τυπική απόκλιση (οι σκιές στις δυο καμπύλες).



Εικόνα 35: *Learning Curves του βελτιστοποιημένου μοντέλου με τον KNN*

9.2 Μοντέλο με Gaussian Naive Bayes

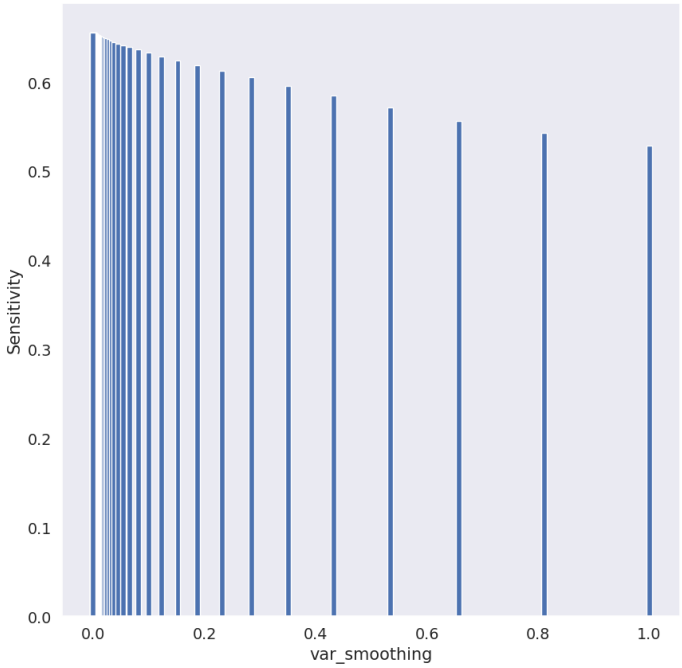
Πίνακας 11: *Αξιολόγηση προβλέψεων του Gaussian Naive Bayes στο test set, μετά από εκπαίδευση με χρήση διαφόρων μεθόδων αντιμετώπισης της ανισορροπίας στις κλάσεις και βελτιστοποίηση*

Μέθοδος αντιμετώπισης ανισορροπίας	Sensitivity	Specificity	Accuracy	AUC(ROC) score
Χωρίς (Unbalanced)	0.52	0.87	0.84	0.80
Random OverSampling	0.67	0.78	0.77	0.80
SMOTE	0.68	0.78	0.77	0.80
ADASYN	0.70	0.76	0.75	0.81
Random Undersampling	0.66	0.78	0.77	0.80
Near Miss	0.76	0.36	0.39	0.56
Βελτιστοποίηση: ADASYN και $\mathbf{Var\_smoothing} = \mathbf{10^{-5}}$	0.70	0.76	0.76	0.81

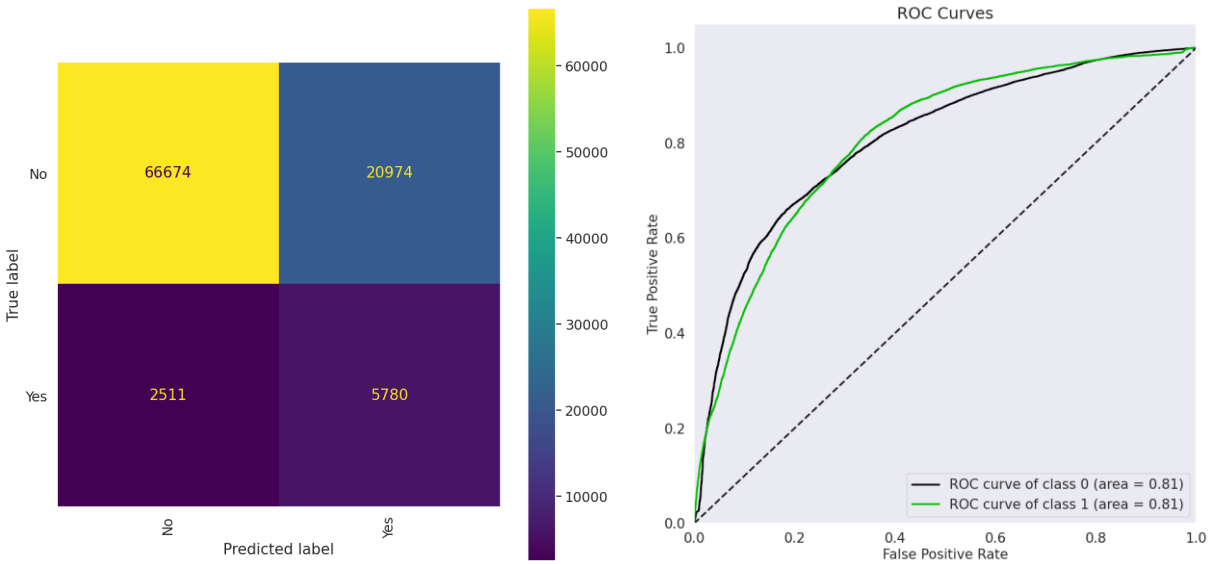
Στον Πίνακα 11 παρατηρείται ότι την χειρότερη επίδοση ως προς την sensitivity είχε το unbalanced training set (0.52) ενώ την καλύτερη (0.76) αυτό που προέκυψε από τη μέθοδο Near Miss. Επειδή όμως με αυτήν οι τιμές των specificity (0.36) και accuracy (0.39) είναι χαμηλές, η βελτιστοποίηση έγινε στο υπερδειγματοληπτημένο training set με τη μέθοδο ADASYN, με τιμές sensitivity 0.70, specificity 0.76, accuracy 0.75 και AUC(ROC) score 0.81.

Βελτιστοποιήθηκε με 5-fold cross validation η υπερπαράμετρος var_smoothing του αλγορίθμου (Εικόνα 36). Προέκυψε sensitivity 0.65 με βέλτιστη τιμή $\mathrm{Var\_smoothing} = 10^{-5}$, η οποία όμως κατά τη εφαρμογή στο test set οδήγησε σε βελτίωση μόνο της accuracy (0.76). Η υπερπαράμετρος var_smoothing ισούται με το τμήμα της μεγαλύτερης διασποράς όλων των χαρακτηριστικών που προστίθεται στις υπόλοιπες για σταθερότητα στον υπολογισμό, ώστε να λαμβάνονται υπ' όψιν δείγματα που είναι μακριά από τη μέση τιμή της κατανομής. Η

default τιμή της είναι 10^{-9}. Στην Εικόνα 37 παρουσιάζονται ο πίνακας σύγχυσης και οι καμπύλες ROC του βελτιστοποιημένου μοντέλου.



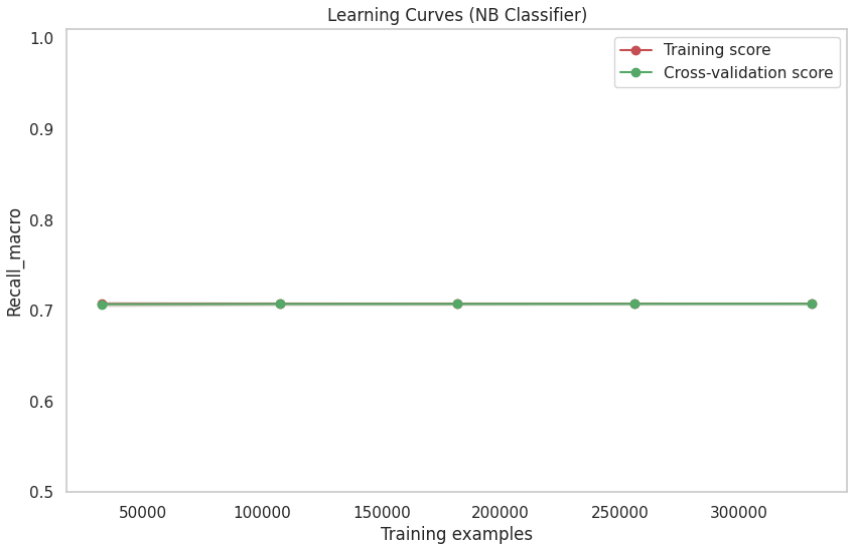
Εικόνα 36: *Οι τιμές της sensitivity στις διαφορετικές τιμές της υπερπαραμέτρου var_smoothing του Gaussian Naive Bayes (βέλτιστη τιμή var_smoothing* $\mathbf{10^{-5}}$*)*



Εικόνα 37: *Confusion matrix και ROC curves της αξιολόγησης του βελτιστοποιημένου μοντέλου με τον Gaussian Naive Bayes στο test set*

Στην Εικόνα 38, στην οποία έχουν σχεδιαστεί οι καμπύλες εκπαίδευσης του βέλτιστου μοντέλου με τον αλγόριθμο Gaussian Naive Bayes, παρατηρείται ότι υπάρχει πλήρης σύγκλιση και επιπέδωση των δύο καμπυλών από πολύ νωρίς χωρίς να υπάρχει

υπερεκπαίδευση. Όσα δείγματα και αν προστεθούν δεν θα βελτιώσουν την επίδοση του μοντέλου.



Εικόνα 38: *Learning Curves του βελτιστοποιημένου μοντέλου με τον Gaussian NB*

9.3 Μοντέλο με Logistic Regression

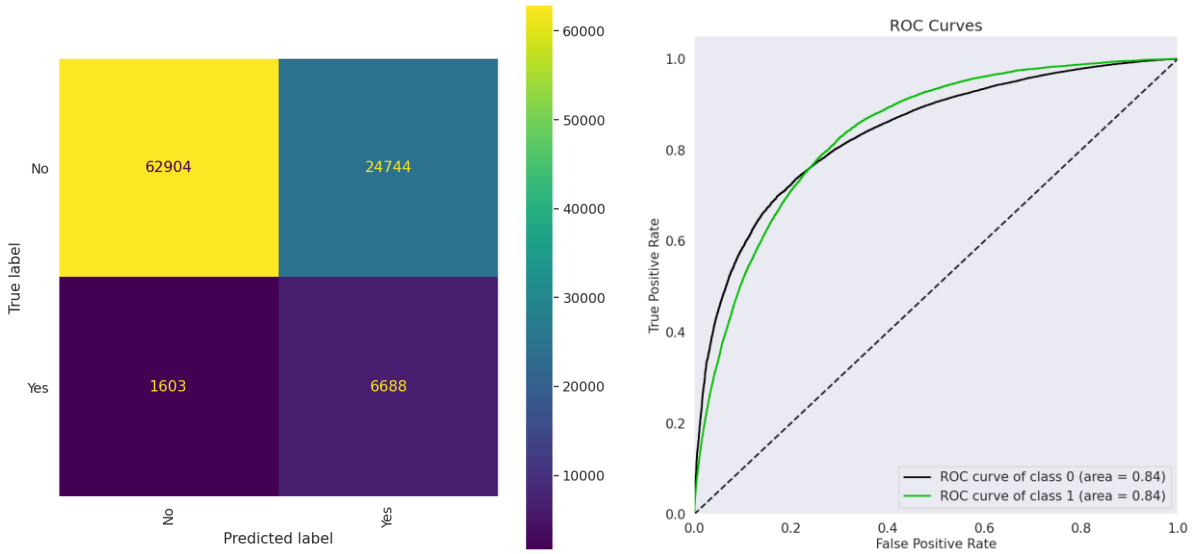
Πίνακας 12: *Αξιολόγηση προβλέψεων του Logistic Regression στο test set, μετά από εκπαίδευση με χρήση διαφόρων μεθόδων αντιμετώπισης της ανισορροπίας στις κλάσεις και βελτιστοποίηση*

Μέθοδος αντιμετώπισης ανισορροπίας	Sensitivity	Specificity	Accuracy	AUC(ROC) score
Χωρίς (Unbalanced)	0.11	0.99	0.91	0.84
Random OverSampling	0.78	0.75	0.75	0.84
SMOTE	0.78	0.75	0.75	0.84
ADASYN	0.81	0.72	0.73	0.84
Random Undersampling	0.78	0.75	0.75	0.84
Near Miss	0.82	0.33	0.37	0.65
Με class_weight =balanced	0.78	0.75	0.75	0.84
Βελτιστοποίηση: ADASYN	0.81	0.72	0.73	0.84

Όπως φαίνεται στον Πίνακα 12, το μοντέλο που εκπαιδεύτηκε στο μη ισορροπημένο test set είχε την χειρότερη επίδοση ως προς τη sensitivity (0.11). Την καλύτερη τιμή είχε αυτό που προέκυψε μετά από υποδειγματοληψία Near Miss (0.82), το οποίο όμως είχε χαμηλές τιμές specificity (0.33) και accuracy (0.37).

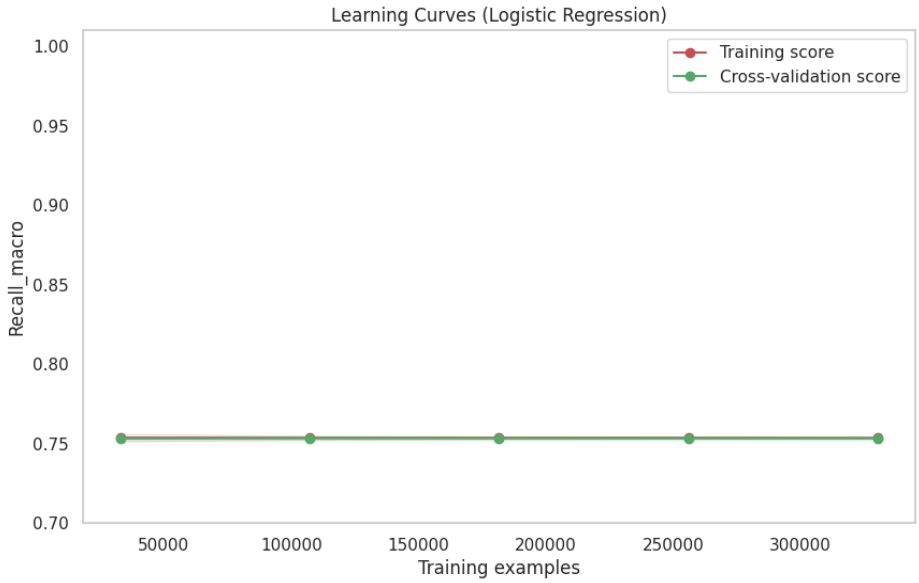
Για βελτιστοποίηση των υπερπαραμέτρων του εκτιμητή επιλέχθηκε αυτό που εκπαιδεύτηκε σε ισορροπημένο train set με υπερδειγματοληψία ADASYN με τιμές sensitivity 0.81, specificity 0.72, accuracy 0.73 και AUC(ROC) score 0.84. Βελτιστοποιήθηκαν οι

υπερπαράμετροι solver, C και penalty της Λογιστικής Παλινδρόμησης με grid search 5–fold cross validation και προέκυψε βέλτιστη τιμή sensitivity 0.79 με τιμές υπερπαραμέτρων: $C = 0.184207,\ \text{solver} = \text{sag},\ \text{penalty} = \text{l2}$. Κατά τη αξιολόγηση του μοντέλου με το test set δεν προέκυψε περαιτέρω βελτίωση των τιμών των μετρικών. Στην Εικόνα 39 παρουσιάζονται ο πίνακας σύγχυσης και οι καμπύλες ROC του βελτιστοποιημένου μοντέλου.



Εικόνα 39: *Confusion matrix και ROC curves της αξιολόγησης του βελτιστοποιημένου μοντέλου με τον Logistic Regression στο test set*

Στην Εικόνα 40, που έχουν σχεδιαστεί οι καμπύλες εκπαίδευσης του βέλτιστου μοντέλου με τον αλγόριθμο της Λογιστικής Παλινδρόμησης, παρατηρείται ότι υπάρχει από νωρίς πλήρης σύγκλιση και επιπέδωση των δύο καμπυλών χωρίς να υπάρχει υπερεκπαίδευση. Πρόσθεση περαιτέρω δειγμάτων δεν θα βελτιώσει την επίδοση του μοντέλου.



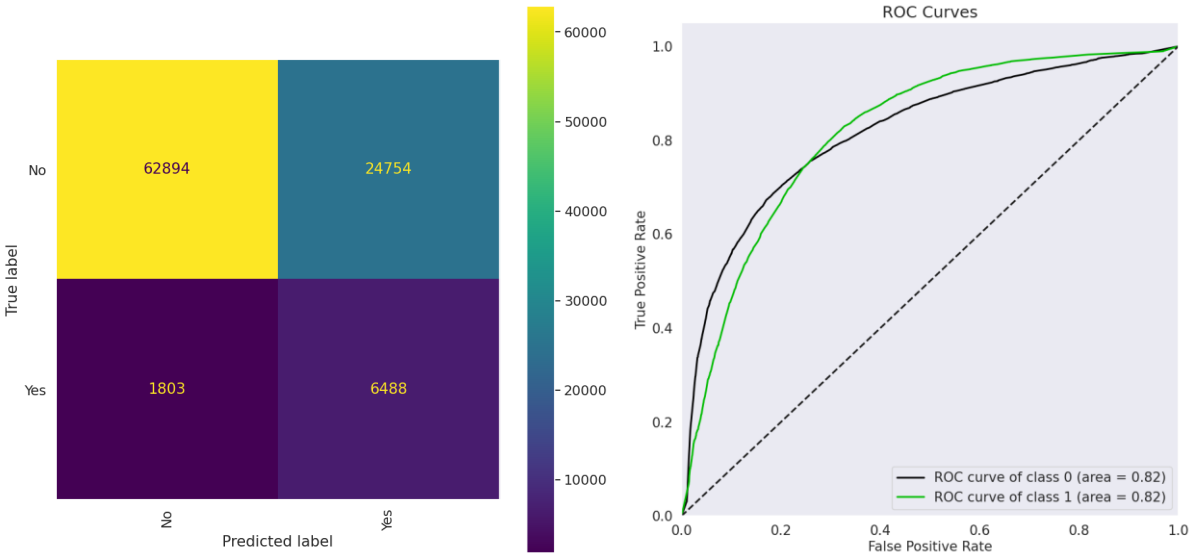
Εικόνα 40: *Learning Curves του βελτιστοποιημένου μοντέλου με Logistic Regression*

9.4 Μοντέλο με Decision Tree

Πίνακας 13: Αξιολόγηση προβλέψεων του Decision Tree στο test set, μετά από εκπαίδευση με χρήση διαφόρων μεθόδων αντιμετώπισης της ανισορροπίας στις κλάσεις και βελτιστοποίηση

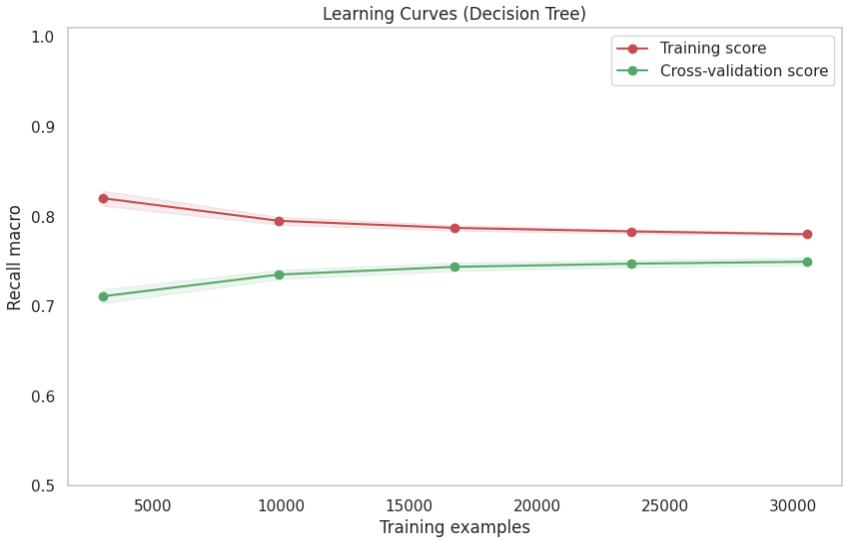
Μέθοδος αντιμετώπισης ανισορροπίας	Sensitivity	Specificity	Accuracy	AUC(ROC) score
Χωρίς (Unbalanced)	0.25	0.92	0.86	0.58
Random OverSampling	0.23	0.93	0.87	0.58
SMOTE	0.28	0.90	0.85	0.59
ADASYN	0.28	0.90	0.85	0.59
Random Undersampling	0.67	0.68	0.68	0.67
Near Miss	0.87	0.17	0.23	0.52
Με class_weight =balanced	0.23	0.93	0.87	0.58
Βελτιστοποίηση: Undersampling	0.78	0.72	0.72	0.82

Από τον Πίνακα 13 φαίνεται ότι η καλύτερη τιμή sensitivity (0.67) σε συνδυασμό με τις specificity (0.68) και accuracy (0.68) ήταν μετά από εκπαίδευση στο τυχαία υποδειγματοληπτημένο training set. Αυτό βελτιστοποιήθηκε με grid search 5-fold cross validation ως προς τις υπερπαραμέτρους criterion, max_depth, min_samples_leaf και min_samples_split του εκτιμητή. Προέκυψε βέλτιστη τιμή sensitivity 0.72 με τις εξής τιμές των υπερπαραμέτρων: criterion='entropy', max_depth=10, min_samples_leaf=4, min_samples_split=10. Με εφαρμογή αυτών των τιμών των υπερπαραμέτρων στην αξιολόγηση με το test set προέκυψε sensitivity 0.78, specificity 0.72, accuracy 0.72 και AUC(ROC) score 0.82. Στην Εικόνα 41 παρουσιάζονται ο πίνακας σύγχυσης και οι καμπύλες ROC του βελτιστοποιημένου μοντέλου.



Εικόνα 41: Confusion matrix και ROC curves της αξιολόγησης της βελτιστοποιημένου εκδοχής του μοντέλου με τον Decision Tree στο test set

Στην Εικόνα 42, στην οποία έχουν σχεδιαστεί οι καμπύλες εκπαίδευσης του βέλτιστου μοντέλου με τον αλγόριθμο Decision Tree, παρατηρείται ότι με την αύξηση των δειγμάτων εκπαίδευσης αυξάνεται η τιμή της μετρικής στο validation score και μειώνεται στο training score, υπάρχει δηλαδή σύγκλιση στις δύο καμπύλες και δεν υπάρχει υπερεκπαίδευση. Επίσης οι δύο καμπύλες σταθεροποιούνται σε κάποιο επίπεδο.



Εικόνα 42: Learning Curves του βελτιστοποιημένου μοντέλου με Decision Tree

9.5 Μοντέλο με Random Forest

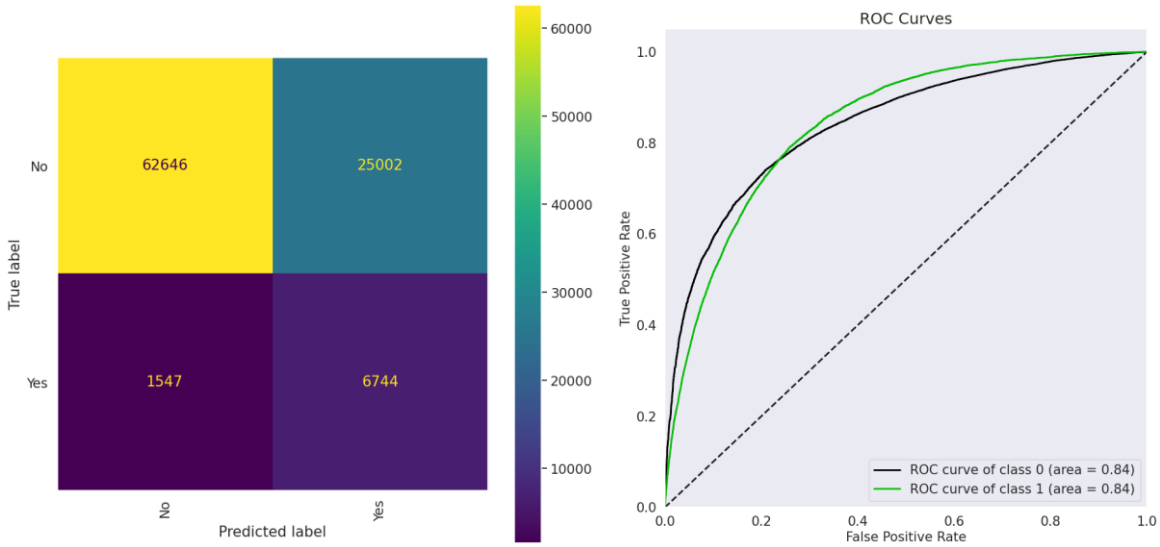
Πίνακας 14: Αξιολόγηση προβλέψεων του Random Forest στο test set, μετά από εκπαίδευση με χρήση διαφόρων μεθόδων αντιμετώπισης της ανισορροπίας στις κλάσεις και βελτιστοποίηση

Μέθοδος αντιμετώπισης ανισορροπίας	Sensitivity	Specificity	Accuracy	AUC(ROC) score
Χωρίς (Unbalanced)	0.12	0.98	0.91	0.79
Random OverSampling	0.22	0.95	0.89	0.79
SMOTE	0.24	0.95	0.88	0.80
ADASYN	0.23	0.94	0.88	0.79
Random Undersampling	0.77	0.71	0.72	0.81
Near Miss	0.88	0.16	0.22	0.58
Με class_weight =balanced	0.11	0.98	0.90	0.79
Βελτιστοποίηση: Undersampling	0.81	0.71	0.72	0.84

Στον Πίνακα 14 παρατηρείται ότι τόσο η εκπαίδευση με το μη ισορροπημένο train set όσο και η εκπαίδευση με τις μεθόδους υπερδειγματοληψίας και class weight, δεν έδωσαν ικανοποιητικά αποτελέσματα ως προς το sensitivity. Μεταξύ των μεθόδων υποδειγματοληψίας του training set Near Miss και Random Undersampling επιλέχθηκε η

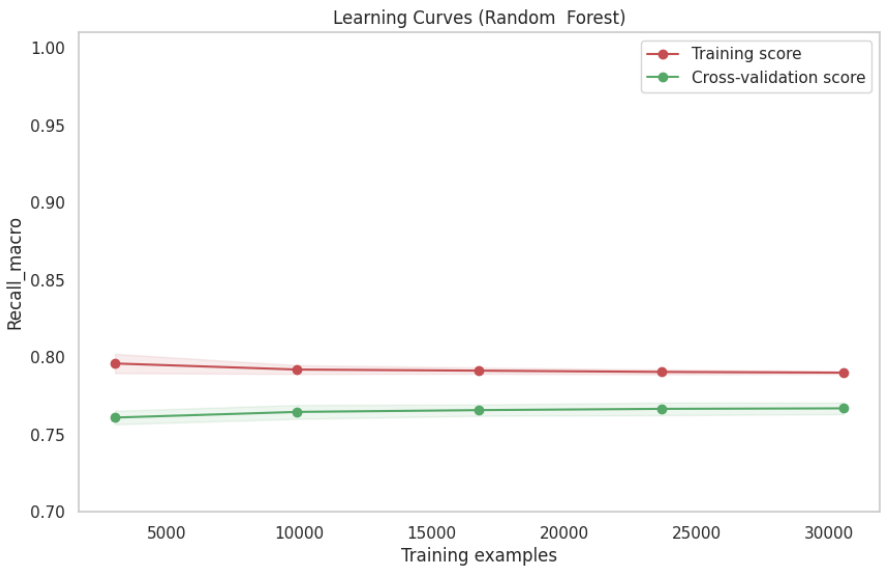
δεύτερη για τη βελτιστοποίηση των υπερπαραμέτρων του εκτιμητή λόγω υψηλότερης τιμής στη specificity, με τιμές sensitivity 0.77, specificity 0.71, accuracy 0.72 και AUC(ROC) score 0.81.

Συγκεκριμένα, χρησιμοποιήθηκε η μέθοδος grid search 5-fold cross validation και έδωσε τις εξής βελτιστοποιημένες τιμές των υπερπαραμέτρων που επιλέχθηκαν: bootstrap = True, max_depth = 30, max_features = log2, min_samples_leaf = 10, min_samples_split = 4, n_estimators = 610 και βέλτιστη τιμή sensitivity 0.82. Κατά την αξιολόγηση του μοντέλου με το test set προέκυψαν τιμές sensitivity 0.81 , specificity 0.71, accuracy 0.72 και AUC(ROC) score 0.84. Στην Εικόνα 43 παρουσιάζονται ο πίνακας σύγχυσης και οι καμπύλες ROC του βελτιστοποιημένου μοντέλου.



Εικόνα 43: *Confusion matrix και ROC curves της αξιολόγησης του βελτιστοποιημένου μοντέλου με τον Random Forest στο test set*

Στην Εικόνα 44, στην οποία έχουν σχεδιαστεί οι καμπύλες εκπαίδευσης του βέλτιστου μοντέλου με τον αλγόριθμο Random Forest, παρατηρείται μερική σύγκλιση μεταξύ των καμπυλών εκπαίδευσης, επιπέδωση και των δύο και απουσία υπερεκπαίδευσης.



Εικόνα 44: *Learning Curves του βελτιστοποιημένου μοντέλου με Random Forest*

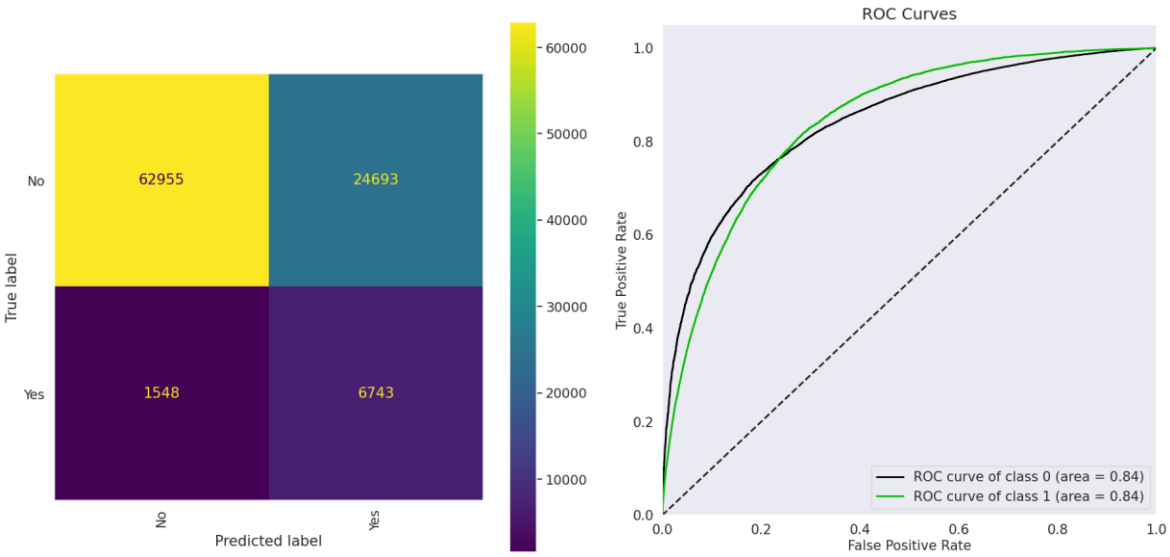
9.6 Μοντέλο με MLP

Πίνακας 15: Αξιολόγηση προβλέψεων του MLP στο test set, μετά από εκπαίδευση με χρήση διαφόρων μεθόδων αντιμετώπισης της ανισορροπίας στις κλάσεις και βελτιστοποίηση

Μέθοδος αντιμετώπισης ανισορροπίας	Sensitivity	Specificity	Accuracy	AUC(ROC) score
Χωρίς (Unbalanced)	0.08	0.99	0.91	0.84
Random OverSampling	0.75	0.76	0.76	0.83
SMOTE	0.70	0.77	0.77	0.81
ADASYN	0.73	0.75	0.75	0.81
Random Undersampling	0.79	0.73	0.73	0.84
Near Miss	0.86	0.20	0.26	0.59
Βελτιστοποίηση Undersampling	0.81	0.72	0.73	0.84

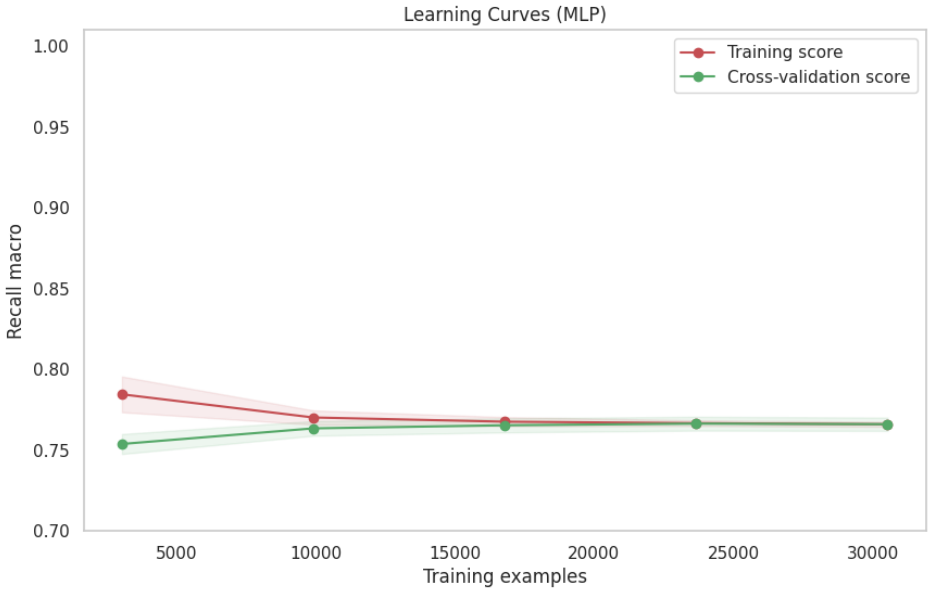
Στον Πίνακα 15 παρατηρείται ότι ο καλύτερος συνδυασμός τιμών sensitivity και specificity προέκυψε με τη χρήση της μεθόδου της τυχαίας υποδειγματοληψίας του training set (0.79 και 0.73 αντίστοιχα) και γι' αυτό επιλέχτηκε για βελτιστοποίηση των υπερπαραμέτρων του MLP.

Με εφαρμογή της μεθόδου grid search 5-fold cross validation προέκυψαν οι εξής βέλτιστες τιμές υπερπαραμέτρων: activation = tanh, alpha = 0.05, hidden_layer_sizes = (50, 50, 50), learning_rate = adaptive, solver = adam και βέλτιστη τιμή sentivity 0.81. Κατά την αξιολόγηση του μοντέλου με το test set προέκυψαν τιμές sensitivity 0.81 , specificity 0.72, accuracy 0.73 και AUC(ROC) score 0.84. Στην Εικόνα 45 παρουσιάζονται ο πίνακας σύγχυσης και οι καμπύλες ROC του βελτιστοποιημένου μοντέλου.



Εικόνα 45: Confusion matrix και ROC curves της αξιολόγησης του βελτιστοποιημένου μοντέλου με τον MLP στο test set

Στην Εικόνα 46, που έχουν σχεδιαστεί οι καμπύλες εκπαίδευσης του βέλτιστου μοντέλου με τον αλγόριθμο MLP, παρατηρείται απόλυτη σύγκλιση και επιπέδωση των καμπυλών εκπαίδευσης με την αύξηση των δειγμάτων εκπαίδευσης και απουσία υπερεκπαίδευσης.



Εικόνα 46: *Learning Curves του βελτιστοποιημένου μοντέλου με MLP*

9.7 Σύγκριση αποτελεσμάτων μοντέλων

Εξετάζοντας τους πίνακες των αποτελεσμάτων όλων των αλγορίθμων που εφαρμόστηκαν, παρατηρείται ότι όταν χρησιμοποιήθηκε για εκπαίδευση το μη ισορροπημένο training set, η τιμή της ευαισθησίας ήταν σε όλους τους αλγόριθμους πολύ χαμηλή, όπως αναμενόταν. Υπήρχε όμως αξιοσημείωτη βελτίωση της τιμής της μετά την αντιμετώπιση της ανισορροπίας των δύο κλάσεων.

Με την εφαρμογή τεχνικών υπερδειγματοληψίας υπήρχε σημαντική αύξηση στην τιμή της ευαισθησίας, εξαιρουμένων όμως των μοντέλων με Decision Tree και Random Forest. Αυτό οφείλεται στο γεγονός ότι από προεπιλογή δεν τίθεται όριο στο μέγιστο βάθος που μπορεί να φθάσει το κάθε δέντρο απόφασης, με αποτέλεσμα το μοντέλο που προκύπτει να υπερεκπαιδεύεται και ως συνέπεια να έχει υψηλή επίδοση στο training set, αλλά χαμηλή επίδοση κατά τη γενίκευση στην πρόβλεψη θετικών δειγμάτων. Με τον καθορισμό τιμής στην παράμετρο του μέγιστου βάθους βελτιώνεται η τιμή της ευαισθησίας χωρίς όμως να ξεπερνά την αντίστοιχη της υποδειγματοληψίας. Αυτό εξηγεί επίσης γιατί οι ίδιοι αλγόριθμοι είχαν χαμηλές τιμές ευαισθησίας και όταν εφαρμόστηκε η μέθοδος class weight. Η Λογιστική Παλινδρόμηση αντίθετα, είχε καλύτερα αποτελέσματα με τη μέθοδο class weight.

Η μέθοδος υποδειγματοληψίας Near Miss σε όλους τους αλγόριθμους αύξησε σημαντικά την ευαισθησία με παράλληλη όμως μεγάλη μείωση στην ειδικότητα και γι' αυτό δεν επιλέχθηκε προς βελτιστοποίηση. Σε όλους τους αλγόριθμους η μέθοδος υποδειγματοληψίας Random Undersampling έδωσε την υψηλότερη τιμή ευαισθησίας, εκτός από την Λογιστική Παλινδρόμηση και τον Αφελή Μπεϋζιανό, όπου λίγο καλύτερη τιμή έδωσε η μέθοδος υπερδειγματοληψίας Adasyn. Κανένα μοντέλο μετά την βελτιστοποίηση δεν έδωσε τιμή sensitivity μεγαλύτερη από 0.81.

9.8 Επιλογή βέλτιστου μοντέλου

Στον Πίνακα 16 παρουσιάζονται συγκεντρωτικά ανά μοντέλο οι τιμές των μετρικών που προέκυψαν μετά τη βελτιστοποίηση και στον Πίνακα 17 οι αντίστοιχες τιμές των πινάκων σύγχυσης.

Πίνακας 16: Συγκριτικός πίνακας με την αξιολόγηση των βελτιστοποιημένων μοντέλων ανά αλγόριθμο

Βέλτιστα μοντέλα ανά αλγόριθμο	Μέθοδος Εξισορρόπησης	Sensitivity	Specificity	Accuracy	AUC(ROC) score
KNN	Random Undersampling	0.78	0.72	0.72	0.82
Gaussian Naïve Bayes	Adasyn	0.70	0.76	0.76	0.81
Logistic Regression	Adasyn	0.81	0.72	0.73	0.84
Decision Tree	Random Undersampling	0.78	0.72	0.72	0.82
Random Forest	Random Undersampling	0.81	0.71	0.72	0.84
MLP	Random Undersampling	0.81	0.72	0.73	0.84

Πίνακας 17: Συγκριτικός πίνακας των τιμών του confusion matrix των βελτιστοποιημένων μοντέλων ανά αλγόριθμο

Βέλτιστα μοντέλα ανά αλγόριθμο	Μέθοδος Εξισορρόπησης	TP	TN	FP	FN
KNN	Random Undersampling	6432	63028	24620	1859
Gaussian Naïve Bayes	Adasyn	5780	66674	20974	2511
Logistic Regression	Adasyn	6688	62904	24744	1603
Decision Tree	Random Undersampling	6488	62894	24754	1803
Random Forest	Random Undersampling	6744	62646	25002	1547
MLP	Random Undersampling	6743	62955	24693	1548

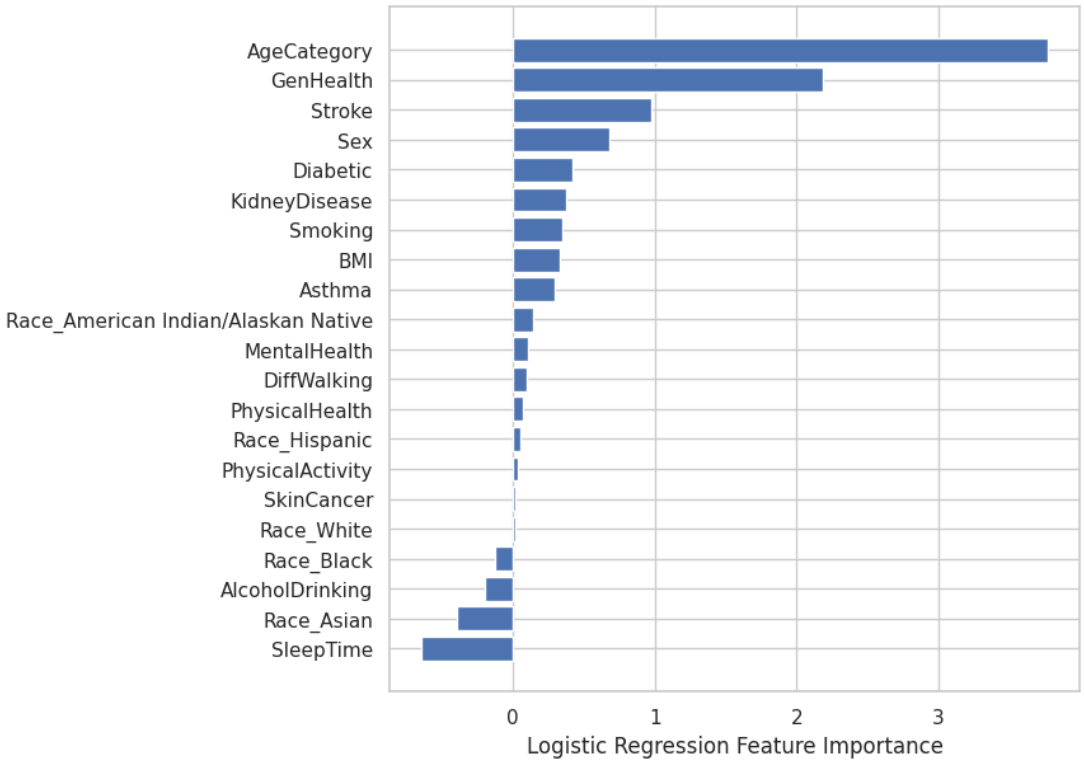
Στους πίνακες αυτούς παρατηρείται ότι λαμβάνοντας υπ' όψιν κυρίως τις τιμές της sensitivity (0.81), αλλά και δευτερευόντως των specificity και AUC(ROC) score, την καλύτερη επίδοση με μικρές αποκλίσεις είχαν τα μοντέλα με χρήση των αλγορίθμων Logistic Regression, Random Forest και MLP με ευαισθησία 81% και ειδικότητα 72%, 71% και 72% αντίστοιχα. Για αυτόν το λόγο προτείνονται και τα τρία για να χρησιμοποιηθούν στη πρόβλεψη εμφάνισης ΣΝ.

9.9 Ερμηνευσιμότητα βέλτιστων μοντέλων

Για τη διερεύνηση της σημαντικότητας του κάθε χαρακτηριστικού στον καθορισμό της πρόβλεψης κάθε στιγμιοτύπου, εφαρμόστηκαν στα τρία βέλτιστα μοντέλα:

- Οι model - specific μέθοδοι:
  - Στο Random Forest μέσω της μέσης μείωσης στην καθαρότητα (Mean Decrease in Impurity – MDI).
  - Στη Λογιστική Παλινδρόμηση με τη χρήση των συντελεστών παλινδρόμησης (coefficients).
- Οι model – agnostic μέθοδοι:
  - Η global model-agnostic μέθοδος Permutation Feature Importance και στα τρία μοντέλα. Η σημαντικότητα σε αυτά καθορίστηκε βάσει της μείωσης της μετρικής recall macro, προκειμένου να ερμηνευθεί η επίδραση που έχει κάθε χαρακτηριστικό στη ταξινόμηση δειγμάτων που ανήκουν τόσο στη θετική όσο και στην αρνητική κλάση. Η μέθοδος αυτή εφαρμόστηκε στο test set που θεωρείται πιο αξιόπιστη επιλογή σε σχέση με το training set.
  - Η local model-agnostic μέθοδος LIME, η οποία εφαρμόστηκε για να ερμηνεύσει την ταξινόμηση τριών στιγμιοτύπων του test set, τα ίδια και στα τρία μοντέλα.

9.9.1 Logistic Regression

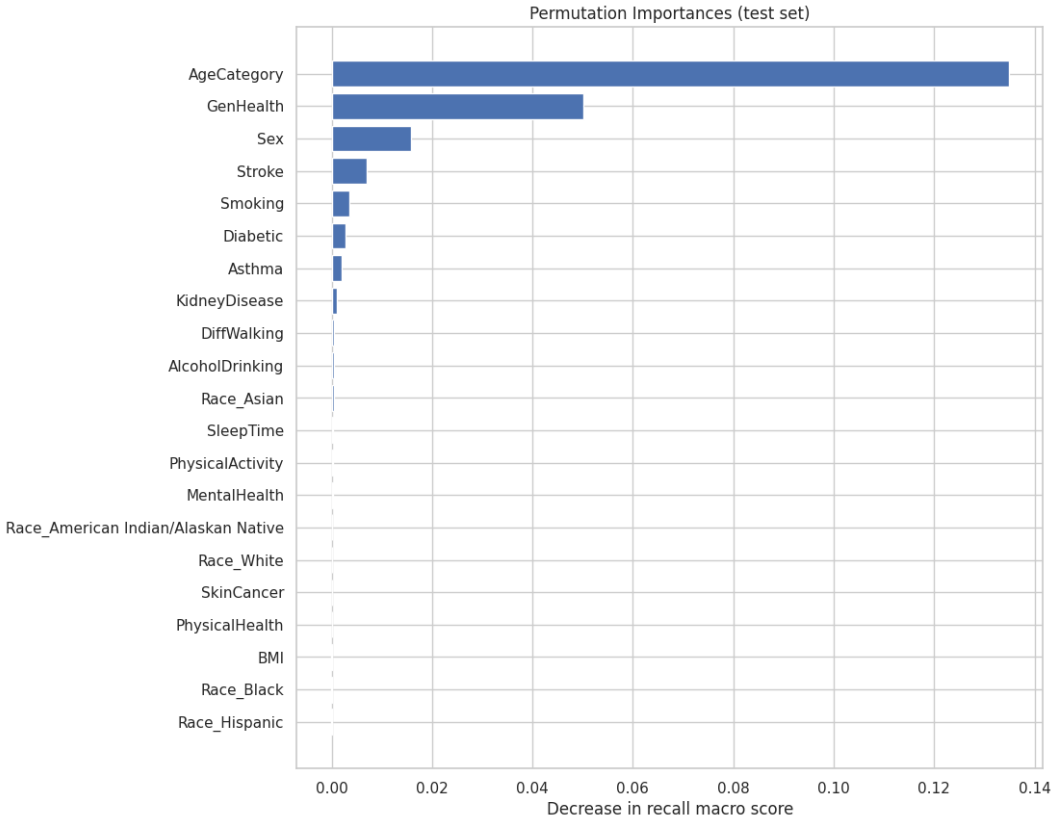


Εικόνα 47: *Αξιολόγηση της σημαντικότητας των χαρακτηριστικών του dataset στο μοντέλο με Logistic Regression χρησιμοποιώντας τους συντελεστές παλινδρόμισης (coefficients)*

Στην Εικόνα 47 παρουσιάζεται η επιρροή του κάθε χαρακτηριστικού στην ταξινόμηση ενός στιγμιοτύπου σύμφωνα με τη χρήση των συντελεστών παλινδρόμησης του μοντέλου με τη Λογιστική Παλινδρόμηση. Η πλειοψηφία των χαρακτηριστικών με την αύξηση της τιμής τους οδηγεί στην αύξηση της πιθανότητας εμφάνισης ΣΝ με βαρύτητα που καθορίζεται από το

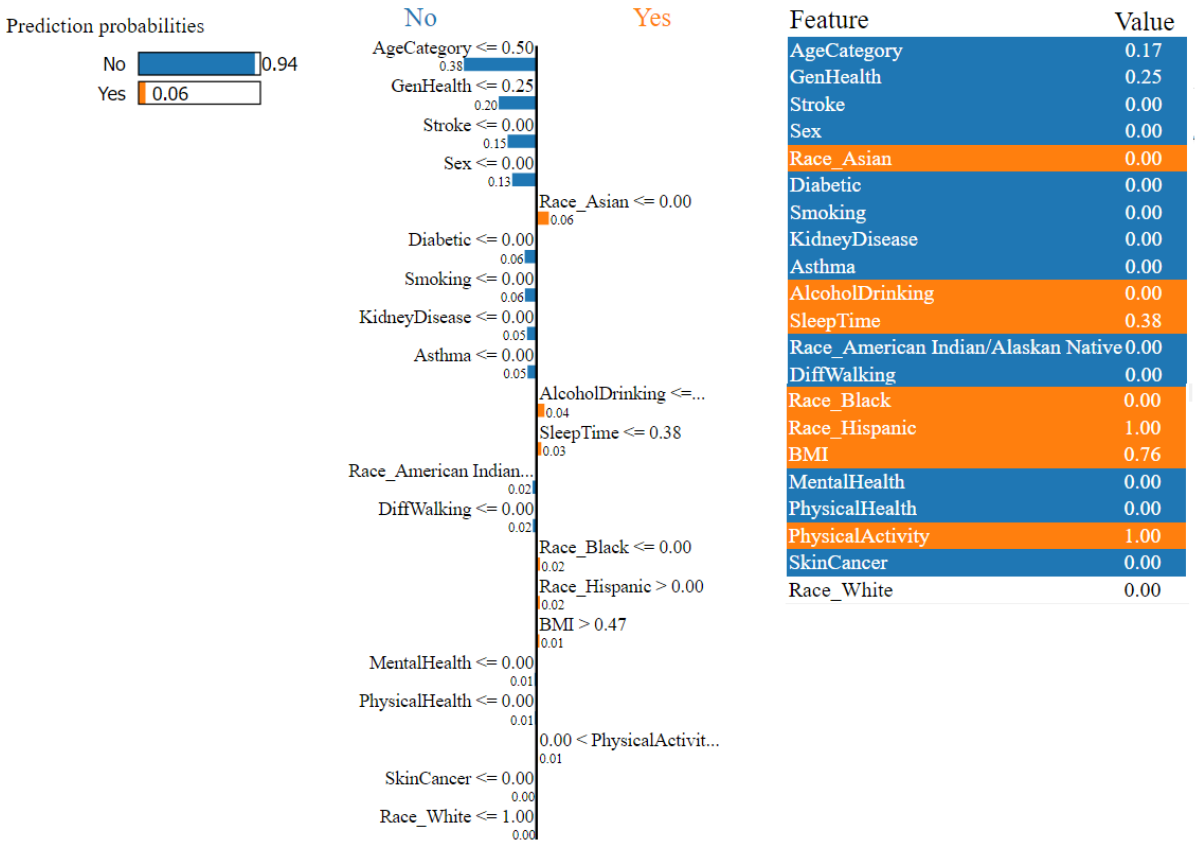
ύψος του αντίστοιχου συντελεστή. Συγκεκριμένα, η ηλικιακή κατηγορία εμφανίζει την μεγαλύτερη επίδραση, ακολουθεί η γενική υγεία, το ιστορικό εγκεφαλικού επεισοδίου, το φύλο (εάν είναι άνδρας), ο διαβήτης, το πρόβλημα νεφρικής λειτουργίας, το κάπνισμα, ο BMI και ακολουθούν τα υπόλοιπα. Ο καρκίνος του δέρματος και η λευκή φυλή έχουν σχεδόν μηδενική επίδραση. Τέλος, μείωση της πιθανότητας εμφάνισης ΣΝ παρατηρείται με αύξηση των ωρών ύπνου, με λήψη αλκοόλ και εάν κάποιος ανήκει στη μαύρη ή ασιατική φυλή. Η αρνητική επίδραση της λήψης αλκοόλ μπορεί να δικαιολογηθεί από το γεγονός ότι το κριτήριο για να έχει τιμή 1 ήταν η λήψη τουλάχιστον ενός μόνο ποτού το μήνα.

Στην Εικόνα 48 απεικονίζεται η επιρροή των χαρακτηριστικών στο ίδιο μοντέλο χρησιμοποιώντας τη μέθοδο Permutation Feature Importance. Παρατηρείται περίπου η ίδια σειρά βαρύτητας, εξαιρουμένου του BMI, που το αξιολογεί με πολύ μικρή επίδραση.

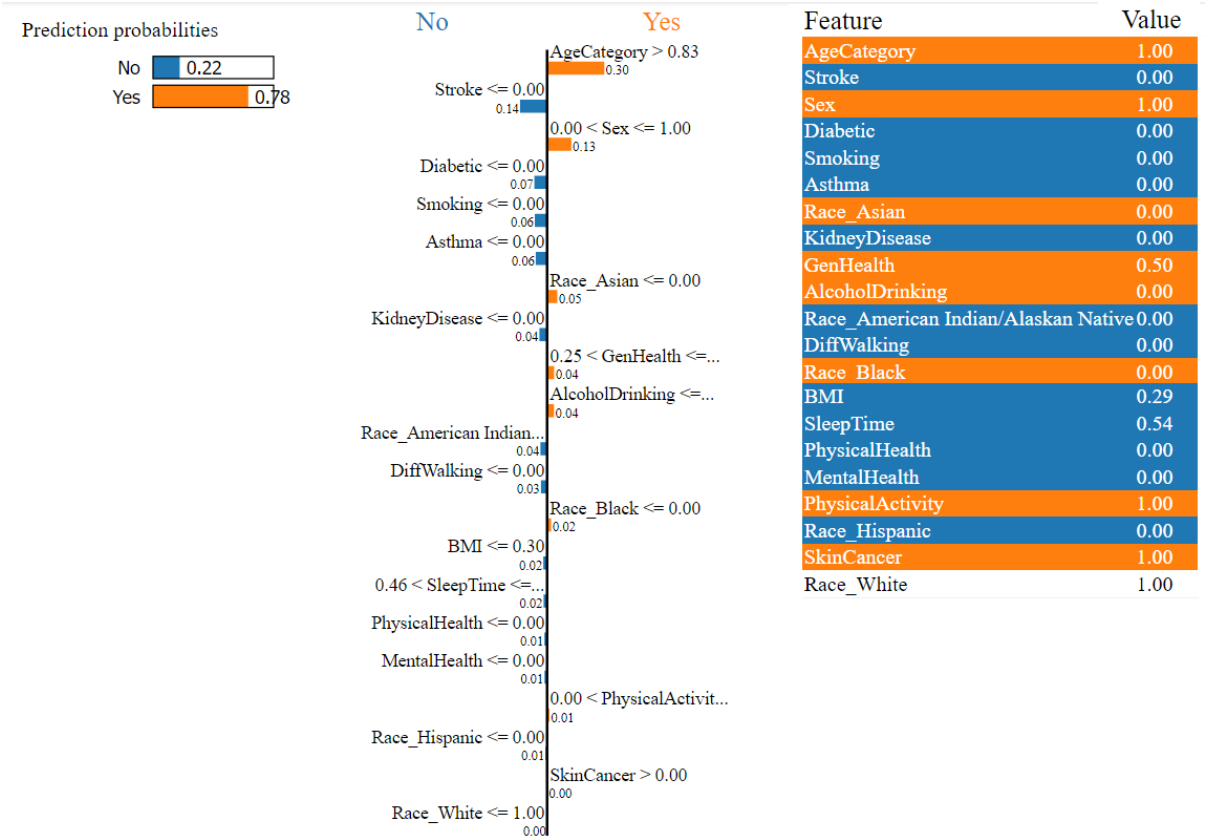


Εικόνα 48: *Αξιολόγηση της σημαντικότητας των χαρακτηριστικών του dataset στο μοντέλο με Logistic Regression χρησιμοποιώντας τη global model-agnostic μέθοδο Permutation Feature importance στο test set βάσει της μείωσης της μετρικής recall macro*

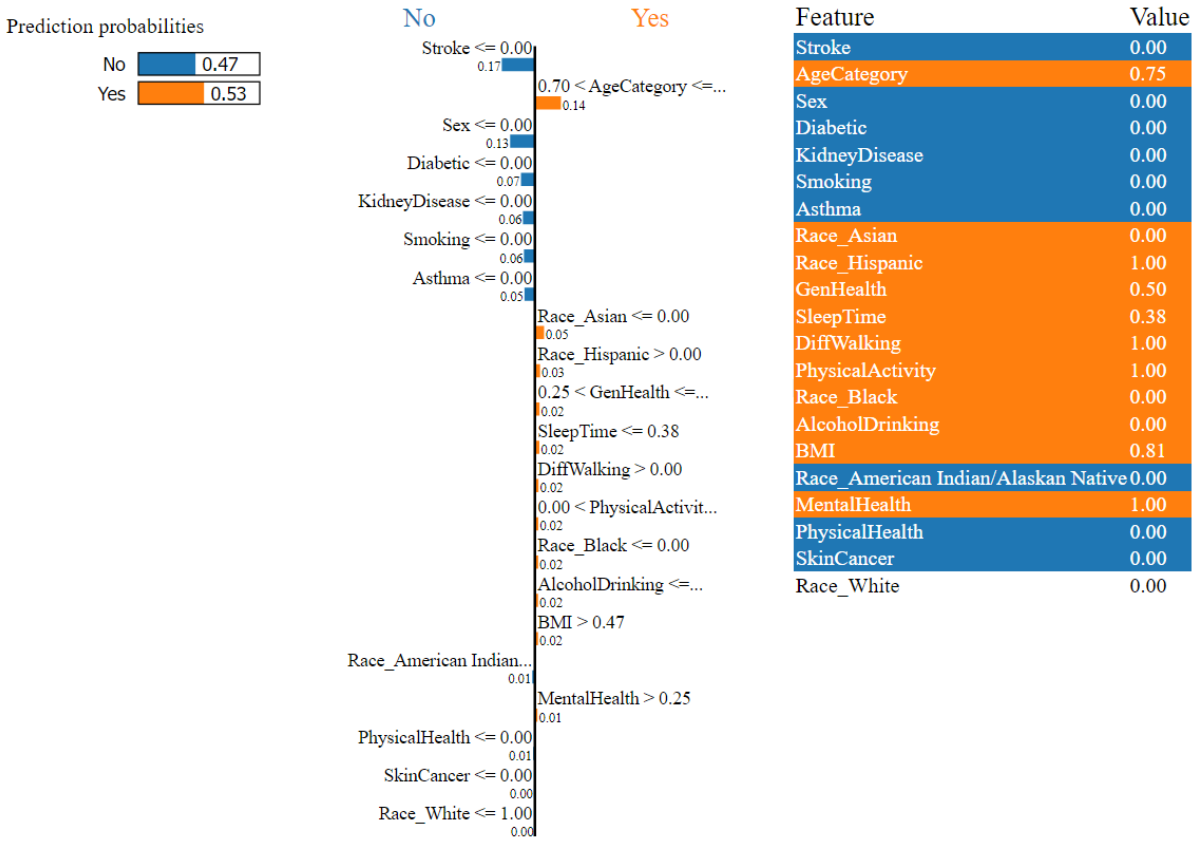
Στις Εικόνες 49, 50 και 51 παρουσιάζεται με τη μέθοδο LIME η ερμηνεία της ταξινόμησης ενός αρνητικού και ενός θετικού στιγμιοτύπου, που ταξινομήθηκαν ορθά και ενός αρνητικού που ταξινομήθηκε εσφαλμένα ως θετικό. Στα δύο πρώτα, η ορθή ταξινόμηση βασίστηκε κυρίως στις τιμές των χαρακτηριστικών ηλικία, γενική υγεία, ύπαρξη ιστορικού εγκεφαλικού, φύλο, διαβήτης, κάπνισμα, χρόνια νεφρική νόσος και άσθμα αξιολογώντας τα είτε με θετική είτε με αρνητική συμβολή στην απόφαση. Το αρνητικό ταξινομήθηκε ως τέτοιο με πιθανότητα 94% και το θετικό με πιθανότητα 78%. Το τρίτο αξιολογήθηκε λανθασμένα ως θετικό με οριακή πιθανότητα 0.53% λόγω της μεγάλης ηλικίας και της μέτριας γενικής υγείας.



Εικόνα 49: *Ερμηνεία με τη μέθοδο LIME επιτυχούς ταξινόμησης αρνητικού στιγμιοτύπου στο μοντέλο με λογιστική παλινδρόμηση*



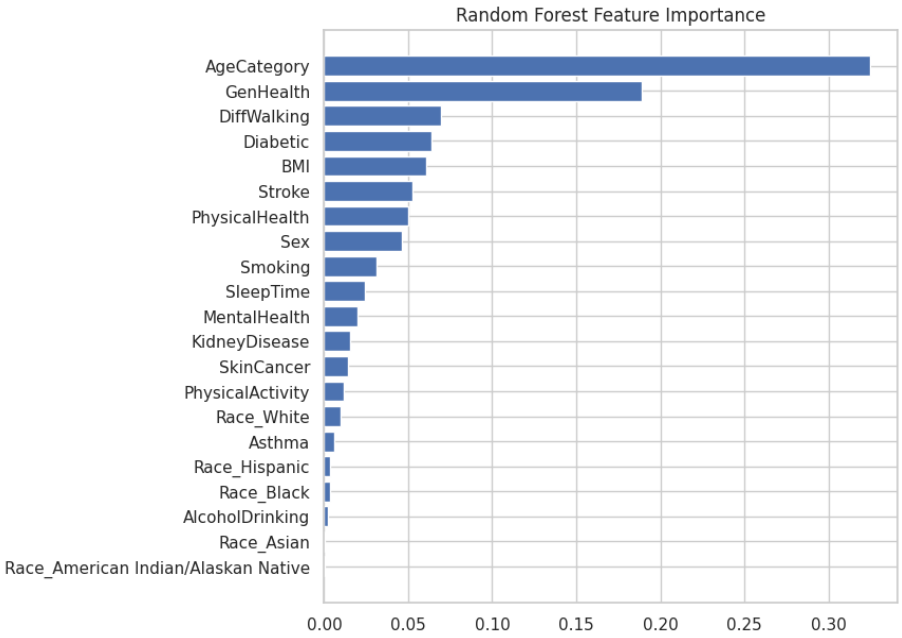
Εικόνα 50: *Ερμηνεία με τη μέθοδο LIME επιτυχούς ταξινόμησης θετικού στιγμιοτύπου στο μοντέλο με λογιστική παλινδρόμηση*



Εικόνα 51: *Ερμηνεία με τη μέθοδο LIME εσφαλμένης ταξινόμησης αρνητικού στιγμιοτύπου στο μοντέλο με λογιστική παλινδρόμηση*

9.9.2 Random Forest

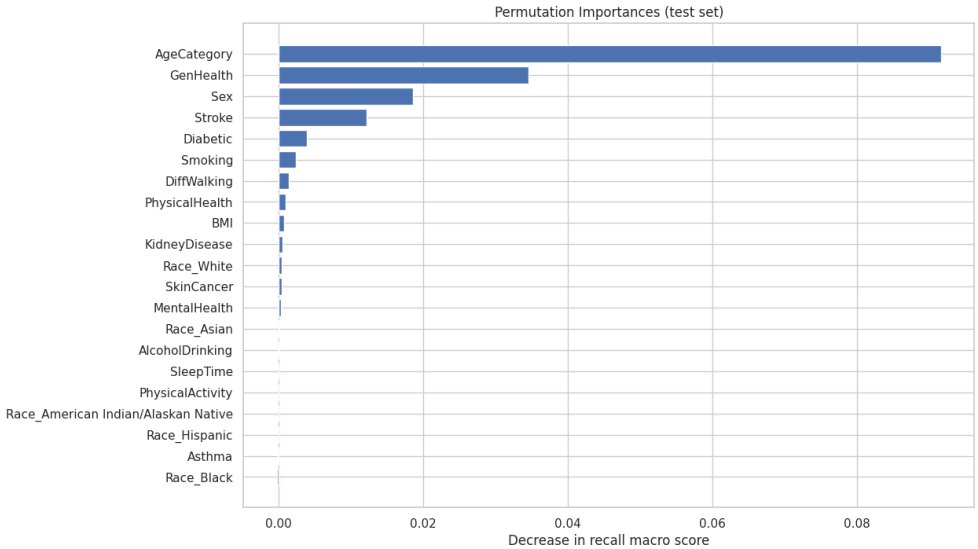
Στην Εικόνα 52 παρουσιάζεται η σημαντικότητα κάθε χαρακτηριστικού στην ταξινόμηση ενός στιγμιοτύπου με χρήση της model specific μεθόδου του Random Forest με αξιολόγηση σημαντικότητας χαρακτηριστικού βάσει της μέσης μείωσης στην μη καθαρότητα (Tree's Feature Importance from Mean Decrease in Impurity - MDI). Η επίδραση κάθε χαρακτηριστικού υπολογίζεται ως η μέση τιμή της συσσώρευσης της αύξησης της καθαρότητας (ή μείωσης της μη καθαρότητας) μέσα σε κάθε δένδρο. Εάν ένα χαρακτηριστικό είναι χρήσιμο στην ταξινόμηση ενός δείγματος, τείνει στο να χωρίζει κόμβους με δείγματα που ανήκουν σε διαφορετικές κλάσεις σε κόμβους με δείγματα που ανήκουν μόνο σε μία κλάση. Η μέθοδος αυτή μπορεί να είναι παραπλανητική για χαρακτηριστικά που παίρνουν πολλές μοναδικές τιμές (high cardinality), όπως για παράδειγμα οι συνεχείς αριθμητικές μεταβλητές, καθώς αυξάνει την σημαντικότητα που αντιστοιχίζει σε αυτές [131]. Ένα άλλο μειονέκτημα της MDI είναι ότι βασίζεται στα στατιστικά που παράγονται από το training set, με αποτέλεσμα να αξιολογούνται ως σημαντικά χαρακτηριστικά που δεν συμβάλλουν στη σωστή ταξινόμηση, όσο το μοντέλο έχει τη δυνατότητα να τα χρησιμοποιεί για να υπερεκπαιδευτεί. Εναλλακτικά μπορεί να εφαρμοστεί η μέθοδος Permutation Feature Importance, η οποία θεωρείται πιο αξιόπιστη, αφού μετριάζει αυτούς τους περιορισμούς [132].



Εικόνα 52: *Αξιολόγηση της σημαντικότητας των χαρακτηριστικών του dataset στο μοντέλο με Random Forest χρησιμοποιώντας την model-specific μεθοδο Random Forest Feature Importance (Mean Decrease in Impurity - MDI)*

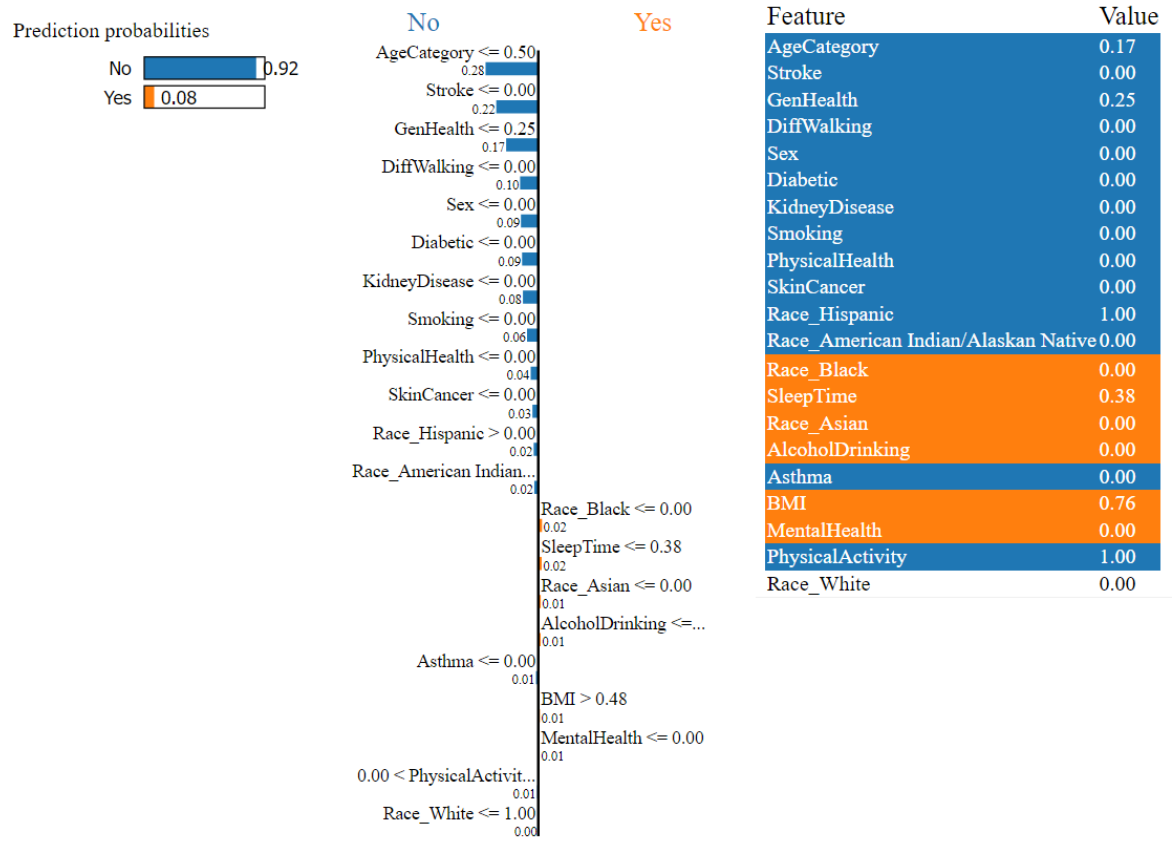
Όπως φαίνεται στην Εικόνα 52, η ηλικιακή κατηγορία εμφανίζει την μεγαλύτερη σημαντικότητα, ακολουθεί η γενική υγεία, η δυσκολία στο περπάτημα, ο διαβήτης, ο BMI, το ιστορικό εγκεφαλικού επεισοδίου, η σωματική υγεία, το φύλο, το κάπνισμα και ακολουθούν τα υπόλοιπα.

Στην Εικόνα 53 παρουσιάζεται η βαρύτητα των χαρακτηριστικών με χρήση της μεθόδου Permutation Feature Importance. Η ιεράρχηση των χαρακτηριστικών βάσει της σημαντικότητας είναι ίδια με την προηγούμενη για τα δύο πρώτα, αλλά παρατηρούνται διαφορές ως προς τα υπόλοιπα. Παρατηρείται μείωση στην επίδραση του BMI, που ίσως οφείλεται στην υπερεκτίμηση που κάνει η MDI λόγω των πολλών μοναδικών τιμών που αυτό έχει. Επίσης το φύλο αξιολογείται στην τρίτη θέση από την όγδοη που αξιολογούνταν με την MDI.

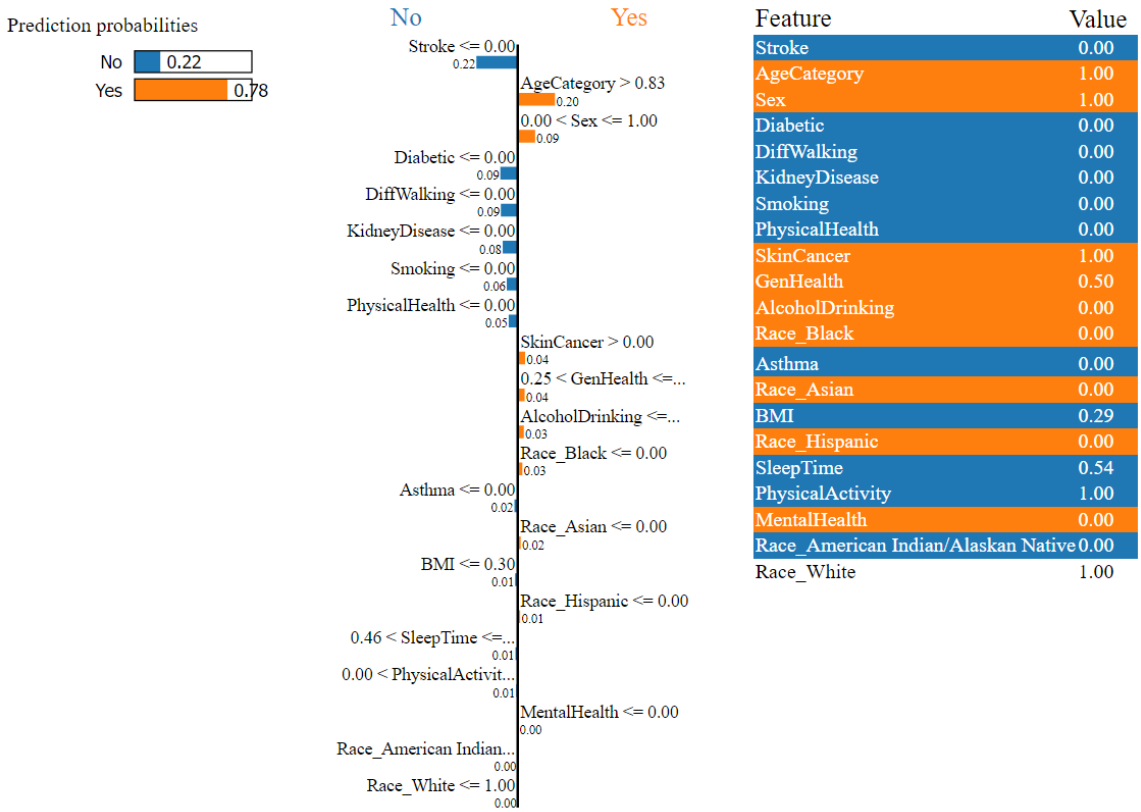


Εικόνα 53: *Αξιολόγηση της σημαντικότητας των χαρακτηριστικών του dataset στο μοντέλο με Random Forest χρησιμοποιώντας τη global model-agnostic μέθοδο Permutation Feature importance στο test set βάσει της μείωσης της μετρικής recall macro*

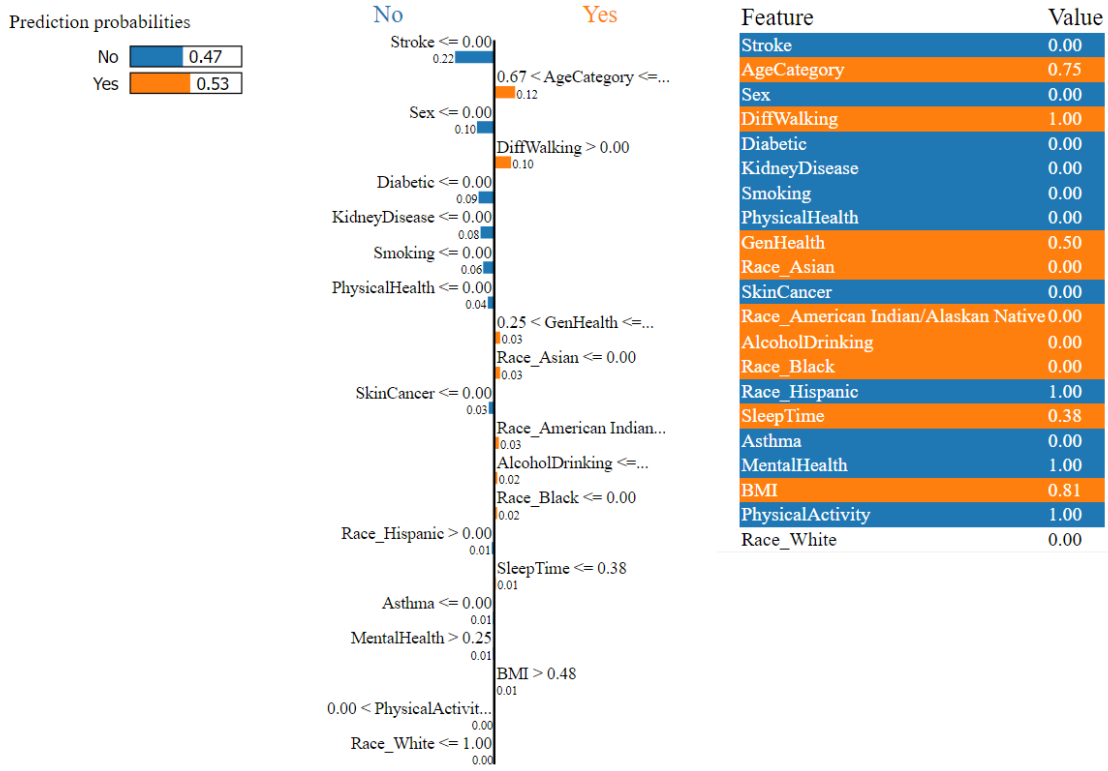
Στις Εικόνες 54, 55, 56 παρουσιάζεται με τη μέθοδο LIME η ερμηνεία της ταξινόμησης από τον RF ενός αρνητικού και ενός θετικού στιγμιοτύπου, που ταξινομήθηκαν ορθά και ενός αρνητικού που ταξινομήθηκε εσφαλμένα ως θετικό (τα ίδια που χρησιμοποιήθηκαν στο προηγούμενο μοντέλο). Στα δύο πρώτα, η ορθή ταξινόμηση βασίστηκε κυρίως στις τιμές των χαρακτηριστικών ηλικία, ύπαρξη ιστορικού εγκεφαλικού, γενική υγεία, δυσκολία στο περπάτημα, φύλο, διαβήτης, χρόνια νεφρική νόσος και κάπνισμα με θετική ή αρνητική συμβολή, ανάλογα με τις τιμές τους. Το αρνητικό ταξινομήθηκε ως τέτοιο με πιθανότητα 92% και το θετικό με πιθανότητα 78%. Το τρίτο αξιολογήθηκε λανθασμένα ως θετικό με οριακή πιθανότητα 0.53% λόγω της μεγάλης ηλικίας, της μέτριας γενικής υγείας και της δυσκολίας στο περπάτημα.



Εικόνα 54: *Ερμηνεία με τη μέθοδο LIME επιτυχούς ταξινόμησης αρνητικού στιγμιοτύπου στο μοντέλο με Random Forest*



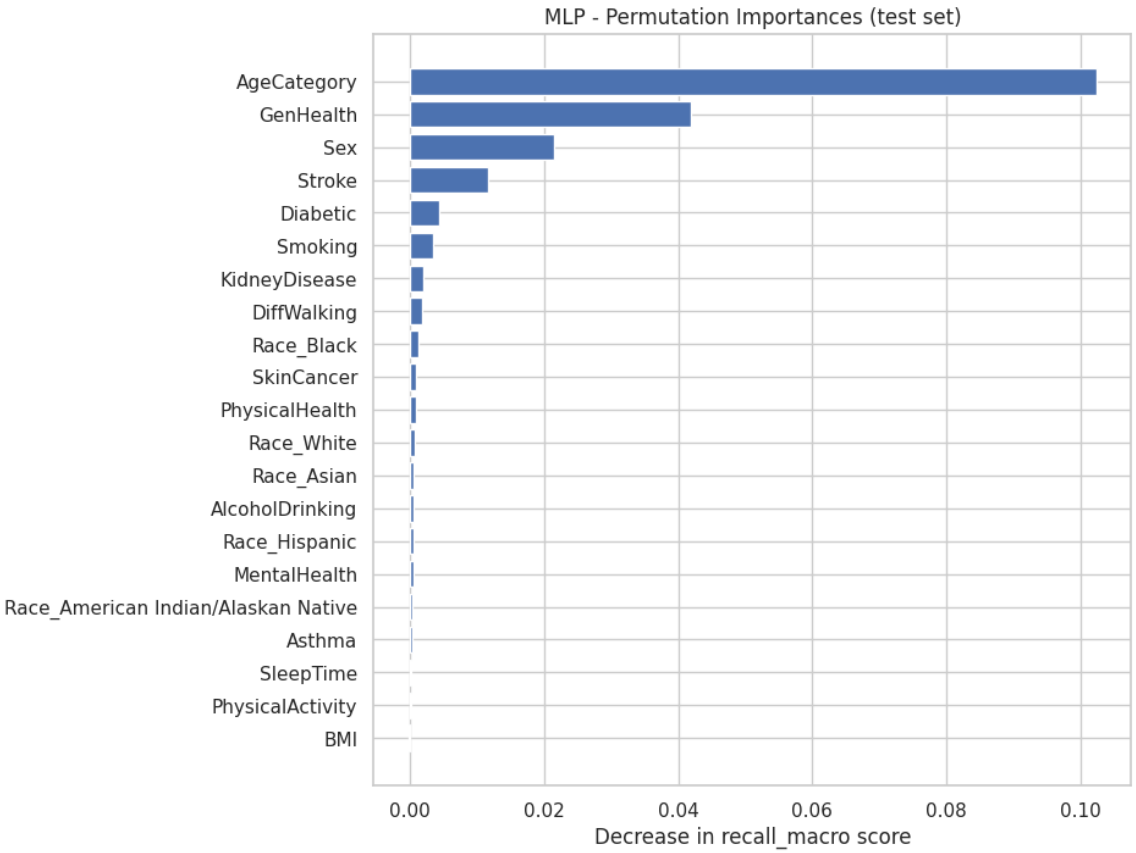
Εικόνα 55: *Ερμηνεία με τη μέθοδο LIME επιτυχούς ταξινόμησης θετικού στιγμιοτύπου στο μοντέλο με Random Forest*



Εικόνα 56: *Ερμηνεία με τη μέθοδο LIME εσφαλμένης ταξινόμησης αρνητικού στιγμιοτύπου στο μοντέλο με Random Forest*

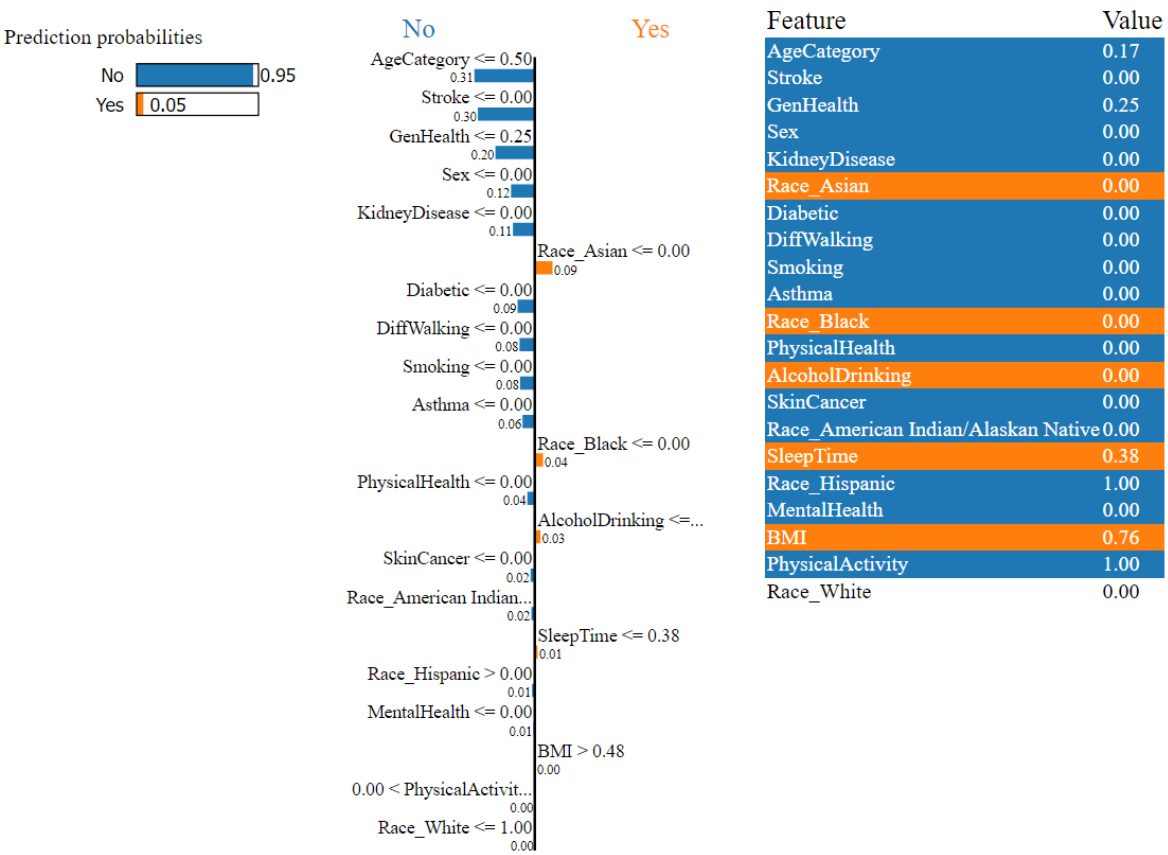
9.9.3 MLP

Στο μοντέλο με MLP εφαρμόστηκε η μέθοδος αξιολόγησης της σημαντικότητας των χαρακτηριστικών Permutation Feature Importance, που απεικονίζεται στην Εικόνα 57. Σύμφωνα με αυτήν το μοντέλο βασίζεται για πρόβλεψη ύπαρξης ΣΝ με φθίνουσα βαρύτητα στα χαρακτηριστικά ηλικιακή κατηγορία, γενική υγεία, φύλο, ιστορικό εγκεφαλικού, διαβήτη, κάπνισμα, χρόνια νεφρική νόσο, δυσκολία στο περπάτημα, μαύρη φυλή και ακολουθούν τα υπόλοιπα.

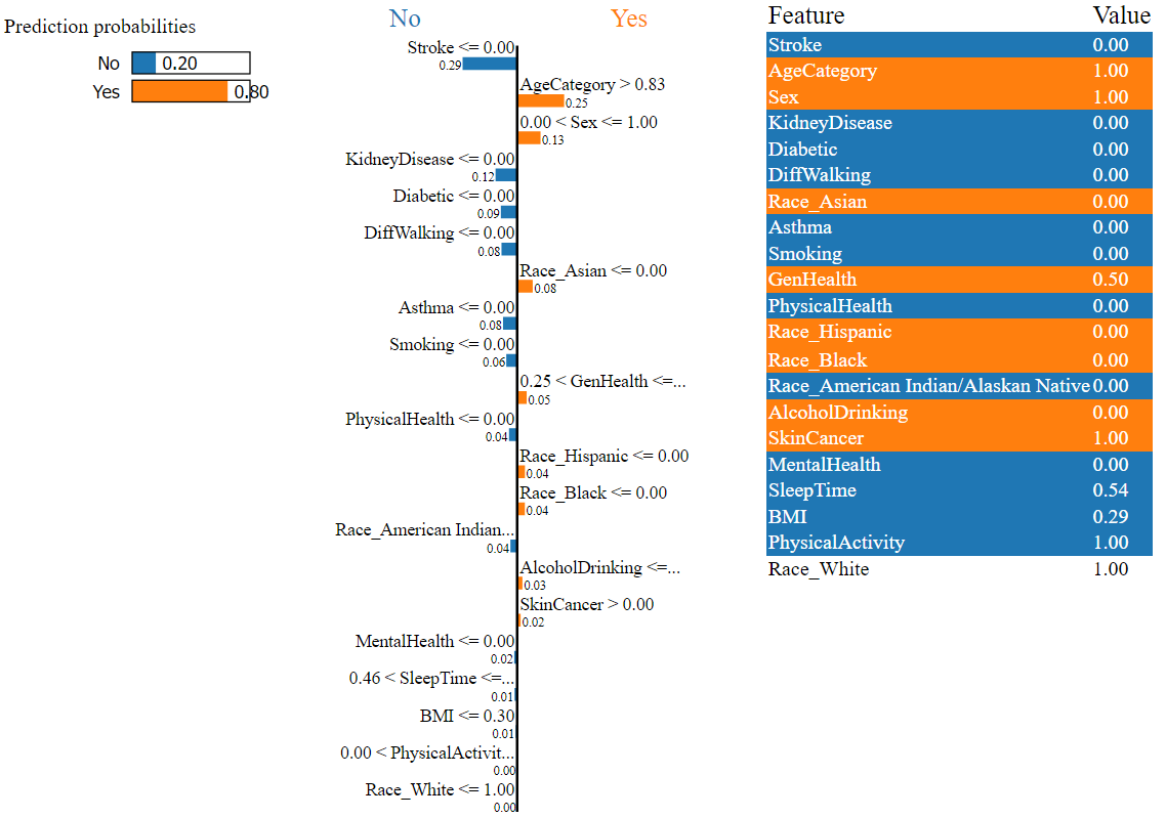


Εικόνα 57: *Αξιολόγηση της σημαντικότητας των χαρακτηριστικών του dataset στο μοντέλο με MLP χρησιμοποιώντας τη global model-agnostic μέθοδο Permutation Feature importance στο test set βάσει της μείωσης της μετρικής recall macro*

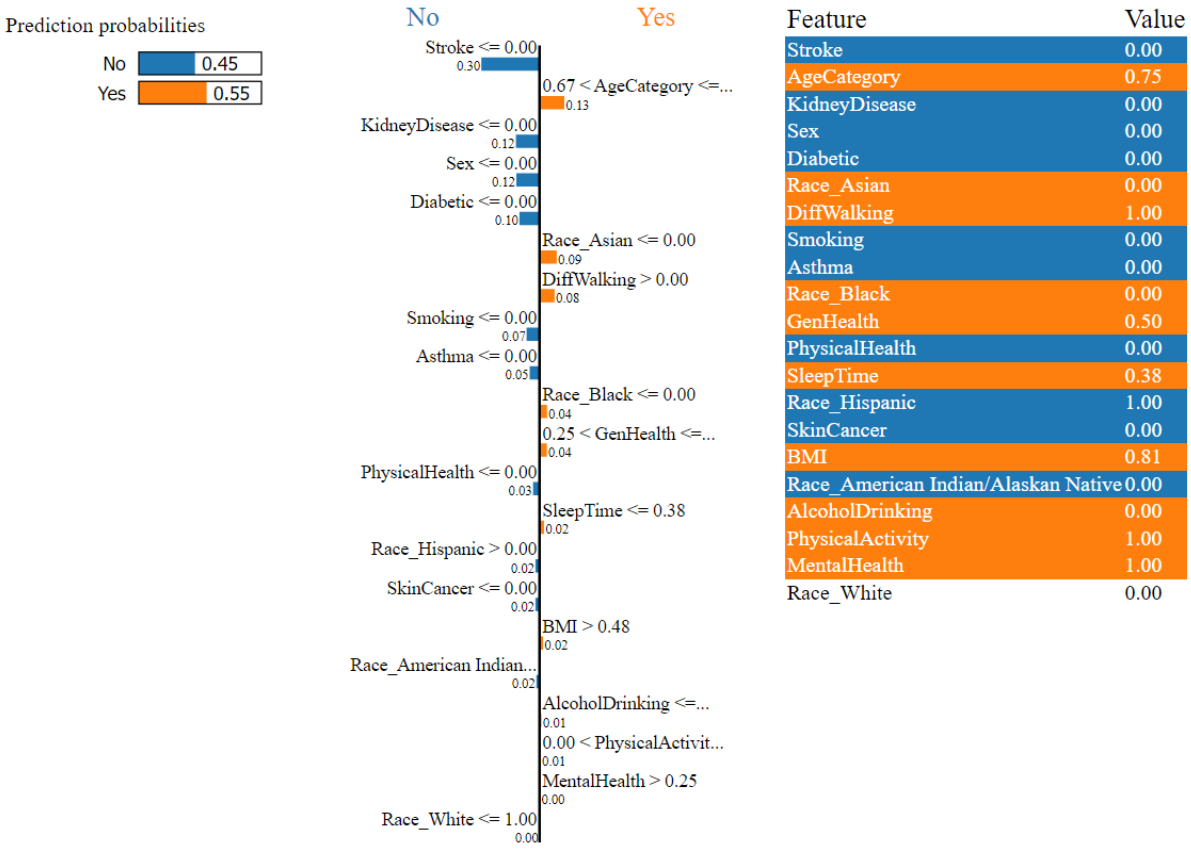
Στις Εικόνες 58, 59 και 60 καταγράφεται με τη μέθοδο LIME η ερμηνεία της ταξινόμησης από τον MLP ενός αρνητικού και ενός θετικού στιγμιοτύπου, που ταξινομήθηκαν σωστά και ενός αρνητικού που ταξινομήθηκε λάθος ως θετικό (τα ίδια που χρησιμοποιήθηκαν στα προηγούμενα δύο μοντέλα). Στα δύο πρώτα, η ορθή ταξινόμηση βασίστηκε κυρίως στις τιμές των χαρακτηριστικών ηλικία, ύπαρξη ιστορικού εγκεφαλικού, φύλο, γενική υγεία, διαβήτης, χρόνια νεφρική νόσος και δυσκολία στο περπάτημα με θετική ή αρνητική συμβολή, ανάλογα με τις τιμές τους. Το αρνητικό ταξινομήθηκε ως τέτοιο με πιθανότητα 95% και το θετικό με πιθανότητα 80%. Το τρίτο αξιολογήθηκε λανθασμένα ως θετικό με οριακή πιθανότητα 0.55% λόγω της μεγάλης ηλικίας, της μέτριας γενικής υγείας, της δυσκολίας στο περπάτημα και της φυλής.



Εικόνα 58: *Ερμηνεία με τη μέθοδο LIME επιτυχούς ταξινόμησης αρνητικού στιγμιοτύπου στο μοντέλο με MLP*



Εικόνα 59: *Ερμηνεία με τη μέθοδο LIME επιτυχούς ταξινόμησης θετικού στιγμιοτύπου στο μοντέλο με MLP*



Εικόνα 60: *Ερμηνεία με τη μέθοδο LIME εσφαλμένης ταξινόμησης αρνητικού στιγμιοτύπου στο μοντέλο με MLP*

9.9.4 Σύγκριση μεθόδων ερμηνευσιμότητας μοντέλων

Σε όλες τις μεθόδους ερμηνευσιμότητας που χρησιμοποιήθηκαν στα τρία τελικά μοντέλα για την αξιολόγηση της συμβολής των χαρακτηριστικών στην ταξινόμηση κάθε στιγμιοτύπου, η ηλικία και η γενική υγεία είχαν την μεγαλύτερη βαρύτητα.

Η μέθοδος model-agnostic Permutation Feature Importance αξιολόγησε τα χαρακτηριστικά με παρόμοια βαρύτητα και στα τρία διαφορετικά μοντέλα. Συγκεκριμένα μετά τα δύο πρώτα που ήδη αναφέρθηκαν, σημαντική επίδραση απέδωσε στο φύλο, στο ιστορικό εγκεφαλικού επεισοδίου, στο διαβήτη, στο κάπνισμα, στη δυσκολία στο περπάτημα και στη χρόνια νεφρική νόσο. Στα υπόλοιπα χαρακτηριστικά δεν αποδόθηκε αξιοσημείωτη βαρύτητα.

Το χαρακτηριστικό BMI αξιολογήθηκε με μεγαλύτερη σημαντικότητα στο μοντέλο με Random Forest και με τις δύο μεθόδους (MDI και Permutation Feature Importance), καθώς επίσης και στη Λογιστική παλινδρόμηση με τη μέθοδο που χρησιμοποιεί τους συντελεστές παλινδρόμησης. Αντίθετα δεν θεωρήθηκε σημαντικό από τη μέθοδο Permutation Feature Importance στον MLP και στη Λογιστική Παλινδρόμηση.

Με τη μέθοδο LIME αξιολογήθηκαν ως σημαντικότερα χαρακτηριστικά για ταξινόμηση ενός δείγματος τα ίδια που εκτιμήθηκαν και από τις υπόλοιπες global μεθόδους. Σημειώνεται ότι από αυτήν δεν μπορούν να εξαχθούν καθολικά συμπεράσματα για την επίδραση των χαρακτηριστικών, διότι λειτουργεί σε τοπικό επίπεδο.

Κεφάλαιο 10: Σύγκριση με βιβλιογραφία, συμπεράσματα και μελλοντικές προκλήσεις

10.1 Σύγκριση με βιβλιογραφία

Τα τελευταία χρόνια, πολλές τεχνολογίες εξόρυξης δεδομένων έχουν χρησιμοποιηθεί για το σχεδιασμό ενός μοντέλου για την πρόβλεψη καρδιακής νόσου ή ανίχνευσή της σε πρώιμο στάδιο. Στη βιβλιογραφία υπάρχουν πολλές επιστημονικές εργασίες με αντίστοιχα θέματα.

Οι Gonsalves και συνεργάτες [133], χρησιμοποιώντας χαρακτηριστικά κυρίως σχετικά με τρόπο ζωής και συμπεριφοράς, ανέπτυξαν μοντέλα με χρήση των αλγορίθμων SVM, Decision Tree και NB. Τα καλύτερα αποτελέσματα προέκυψαν με χρήση του NB με sensitivity 0.63 και specificity 0.76. Η χαμηλότερη τιμή sensitivity από αυτήν της παρούσας εργασίας θα μπορούσε να οφείλεται στο γεγονός ότι χρησιμοποιήθηκε για εκπαίδευση ένα μη ισορροπημένο dataset, με 302 αρνητικά και 160 θετικά δείγματα, χωρίς να αντιμετωπιστεί το πρόβλημα της ανισορροπίας των κλάσεων.

Οι Obasi και συνεργάτες [134] σε dataset 1990 παρατηρήσεων με 993 θετικές και 997 αρνητικές, ανέπτυξαν μοντέλα χρησιμοποιώντας τους αλγορίθμους Logistic Regression, NB και Random Forest. Καλύτερα αποτελέσματα είχε ο Random Forest με sensitivity 0.93 και specificity 0.92. Οι υψηλότερες τιμές των μετρικών σε σχέση με αυτές της παρούσας εργασίας ίσως οφείλονται αφενός στο πλήρως ισορροπημένο dataset, αφετέρου στην επιλογή των χαρακτηριστικών που συμπεριελάμβαναν όχι μόνο σχετικά με τον τρόπο ζωής, όπως κάπνισμα, αλκοόλ και παχυσαρκία, αλλά και χαρακτηριστικά που αφορούσαν άμεσα την καρδιακή λειτουργία όπως καρδιογράφημα, αρτηριακή πίεση και στηθάγχη. Παρόμοια χαρακτηριστικά χρησιμοποίησαν οι Chandrasekhar και Peddakrishna σε μία εκτεταμένη μελέτη σε δύο dataset με επίσης πολύ υψηλές τιμές στα αποτελέσματα [135].

10.2 Συμπεράσματα και μελλοντικές προκλήσεις

Το αντικείμενο της παρούσης εργασίας ήταν η δημιουργία ενός μοντέλου μηχανικής μάθησης με στόχο την πρόβλεψη εμφάνισης στεφανιαίας νόσου. Ακριβώς επειδή σκοπός είναι η πρόληψη και όχι η διάγνωση της πάθησης, χρησιμοποιήθηκαν ως παράγοντες πρόβλεψης χαρακτηριστικά που σχετίζονται κυρίως με την γενική κατάσταση, συνυπάρχουσες νόσους, τον τρόπο ζωής και τις συνήθειες ενός ατόμου και όχι με στοιχεία που έχουν άμεση σχέση με την παθολογία της καρδιάς όπως στηθάγχη, υπέρταση, καρδιογράφημα, χοληστερόλη κτλ.

Όταν η συλλογή ιατρικών δεδομένων που αφορούν ασθένειες προέρχεται από τον γενικό πληθυσμό, στον οποίο ο επιπολασμός είναι πολύ χαμηλός, τα δεδομένα συνήθως εμφανίζουν μεγάλη ανισορροπία μεταξύ θετικής και αρνητικής κλάσης, με την θετική να έχει πολύ μικρότερη εκπροσώπηση. Αυτό καθιστά δύσκολη την εκπαίδευση ενός μοντέλου με χρήση αλγορίθμων μηχανικής μάθησης στην αναγνώριση νέων θετικών περιπτώσεων στη γενίκευση. Το ίδιο συνέβαινε και στα δεδομένα της παρούσης εργασίας και αντιμετωπίστηκε με διάφορες μεθόδους είτε επαναδειγματοληψίας είτε αλγοριθμικές, που βελτίωσαν σημαντικά την επίδοση των μοντέλων.

Ως μετρική αξιολόγησης χρησιμοποιήθηκε η ευαισθησία και δευτερευόντως η ειδικότητα, ώστε το μοντέλο που θα προκύψει να εστιάζει στον εντοπισμό όσο το δυνατόν περισσότερων αληθώς θετικών περιπτώσεων, χωρίς όμως να αυξάνονται πολύ οι ψευδώς

θετικές. Μικρή αύξηση των ψευδώς θετικών, τα οποία σχετίζονται κυρίως με οριακές περιπτώσεις, δεν αποτελεί σημαντικό πρόβλημα, διότι στόχος είναι η πρόληψη, δηλαδή μετά από κατάλληλη συμβουλευτική παρέμβαση, να βοηθηθεί κάποιος στην αλλαγή του τρόπου ζωής του, για να αποτραπεί η ανάπτυξη της νόσου (βλ. Κεφ. 4.3).

Τελικά αναπτύχθηκαν τρία μοντέλα που είχαν εξίσου καλές επιδόσεις, με την χρήση των αλγορίθμων Logistic Regression, Random Forest και MLP με εντοπισμό του 81% των θετικών περιστατικών από αυτά που έπασχαν από τη νόσο. Τα μοντέλα έγιναν ερμηνεύσιμα με χρήση τεχνικών όπως Permutation Feature Importance και LIME. Η προβλεπτική ικανότητα των μοντέλων κρίνεται ως αξιόπιστη, εάν ληφθούν υπ' όψιν οι καμπύλες εκπαίδευσης, καθώς και το γεγονός ότι και τα τρία ιεραρχούν παρόμοια την βαρύτητα των χαρακτηριστικών τόσο συνολικά όσο και στην ερμηνεία μεμονωμένων περιπτώσεων.

Για περαιτέρω βελτίωση της επίδοσης των μοντέλων θα μπορούσαν να χρησιμοποιηθούν σύνολα δεδομένων προερχόμενα και από άλλες χώρες εκτός ΗΠΑ από τις οποίες προέρχονταν τα δεδομένα αυτής της εργασίας. Σχετικά με τα χαρακτηριστικά, θα ήταν πιο βοηθητική, όπου αυτό είναι δυνατόν, η ποσοτικοποίηση κάποιων εξ' αυτών όπως του καπνίσματος και της λήψης αλκοόλ για να ενισχυθεί η διακριτική ικανότητα των μοντέλων. Επιπρόσθετα, θα μπορούσαν να χρησιμοποιηθούν για την πρόβλεψη και άλλα χαρακτηριστικά όπως χοληστερόλη, οικογενειακό ιστορικό, στοιχεία προσωπικότητας, stress, ιδιαίτερες συνθήκες τόπου κατοικίας όπως ηχορύπανση και μόλυνση ατμόσφαιρας κά. Σκόπιμη θεωρείται και η δοκιμή υβριδικού ensemble μοντέλου που θα συνδυάζει τα τρία βέλτιστα μοντέλα που προέκυψαν και η απόφαση να λαμβάνεται με ψηφοφορία με soft voting. Μπορούν επίσης να χρησιμοποιηθούν και άλλες τεχνικές ερμηνευσιμότητας, όπως η SHAP (SHapley Additive exPlanations) από την οποία, με βάση τις παραγόμενες τοπικές εξηγήσεις, μπορούν να εξαχθούν πιο αξιόπιστα συμπεράσματα σχετικά με τη συνολική ερμηνεία του κάθε μοντέλου. Τέλος, για να χρησιμοποιηθούν αυτά τα μοντέλα στην πράξη για πρόβλεψη στεφανιαίας νόσου σε πραγματικό χρόνο από ειδικούς υγείας, είναι αναγκαία η ενσωμάτωσή τους σε μια πρακτική εφαρμογή, όπως web ή mobile application.

Συμπερασματικά, η χρήση αλγορίθμων μηχανικής μάθησης μπορεί να αποτελέσει χρήσιμο εργαλείο στα χέρια των ειδικών τόσο για την πρόβλεψη όσο και για την πρόληψη της ΣΝ. Η λήψη προληπτικών μέτρων θα έχει μεγάλο όφελος για τη ζωή κάθε ατόμου, την ποιότητα ζωής του, το οικογενειακό και κοινωνικό του περιβάλλον, αλλά βεβαίως και για τα ασφαλιστικά συστήματα των κρατών. Η μελέτη αυτή συνηγορεί μαζί με πολλές άλλες ότι αυτό είναι εφικτό με την επιλογή κατάλληλων χαρακτηριστικών και αλγορίθμων μηχανικής μάθησης.

Βιβλιογραφία

[1] British Heart Foundation, “BHF Statistics Factsheet - Global,” 2023. [Online]. Available: https://www.bhf.org.uk/-/media/files/for-professionals/research/heart-statistics/bhf-cvd-statistics-global-factsheet.pdf.

[2] M. Lindstrom *et al.*, “Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990-2021,” *J. Am. Coll. Cardiol.*, vol. 80, no. 25, pp. 2372–2425, 2022, doi: https://doi.org/10.1016/j.jacc.2022.11.001.

[3] Institute for Health Metrics and Evaluation (IHME), “Global Burden of Disease (2020) data - estimates for 2019,” 2020. [Online]. Available: https://vizhub.healthdata.org/gbd-results/.

[4] S. Yusuf, S. Reddy, S. Ounpuu, and S. Anand, “Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization.,” *Circulation*, vol. 104, no. 22, pp. 2746–2753, Nov. 2001, doi: 10.1161/hc4601.099487.

[5] World Health Organization, “Cardiovascular diseases (CVDs) - Key facts,” 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[6] World Health Organization, “WHO Mortality Database,” *2021*. [Online]. Available: https://www.who.int/data/data-collection-tools/who-mortality-database.

[7] A. Timmis *et al.*, “European Society of Cardiology: cardiovascular disease statistics 2021: Executive Summary,” *Eur. Hear. J. - Qual. Care Clin. Outcomes*, vol. 8, no. 4, pp. 377–382, 2022, doi: 10.1093/ehjqcco/qcac014.

[8] National Center for Health Statistics, “Multiple Cause of Death 2018–2021 on CDC WONDER Database.,” 2023. [Online]. Available: https://wonder.cdc.gov/mcd.html.

[9] C. W. Tsao *et al.*, “Heart Disease and Stroke Statistics-2023 Update: A Report From the American Heart Association.,” *Circulation*, vol. 147, no. 8, pp. e93–e621, Feb. 2023, doi: 10.1161/CIR.0000000000001123.

[10] Agency for Healthcare Research and Quality, “Medical Expenditure Panel Survey (MEPS): household component summary tables: medical conditions, United States,” *2021*. [Online]. Available: https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp.

[11] National Center for Health Statistics, “Percentage of coronary heart disease for adults aged 18 and over, United States, 2019—2021,” 2023.

[12] Wikipedia, “Καρδιά.” [Online]. Available: https://el.wikipedia.org/wiki/Καρδιά.

[13] J. G. Betts, *Anatomy & Physiology*, 1st ed. Rice University, 2013.

[14] S. Swetha, “Pathophysiology of myocardial infarction,” 2017. [Online]. Available: https://www.semanticscholar.org/paper/Clinical-Study-of-Cardiac-Markers-in-Post-Patients-Swetha/a3ccc7932ec4e21a1250eb4f103830b25875eab9/figure/0.

[15] R. C. Thompson *et al.*, “Atherosclerosis across 4000 years of human history: the Horus study of four ancient populations.,” *Lancet (London, England)*, vol. 381, no. 9873, pp. 1211–1222, Apr. 2013, doi: 10.1016/S0140-6736(13)60598-X.

[16] Δημήτριος Κρεμαστινός, *Καρδιολογία*, 1st ed. Πασχαλίδης, 2005.

[17] D. S. Berman, Y. Arnson, and A. Rozanski, “Coronary Artery Calcium Scanning: The Agatston Score and Beyond∗ ,” *JACC Cardiovasc. Imaging*, vol. 9, no. 12, pp. 1417–1419, 2016, doi: https://doi.org/10.1016/j.jcmg.2016.05.020.

[18] J. de Almeida, S. Martinho, L. Gonçalves, and M. Ferreira, “Positron Emission Tomography in Coronary Heart Disease,” *Applied Sciences*, vol. 12, no. 9. 2022, doi: 10.3390/app12094704.

[19] E. Kandaswamy and L. Zuo, “Recent advances in treatment of coronary artery disease:

Role of science and technology," *Int. J. Mol. Sci.*, vol. 19, no. 2, 2018, doi: 10.3390/ijms19020424.
[20] National Institutes of Health (NIH), "Stroke," 2023. [Online]. Available: https://www.nhlbi.nih.gov/health/stroke.
[21] Mayo Clinic, "Stroke - Symptoms and causes," 2023.
[22] M. A. Creager and J. Loscalzo, "Arterial Diseases of the Extremities," in *Harrison's Principles of Internal Medicine, 20e*, J. L. Jameson, A. S. Fauci, D. L. Kasper, S. L. Hauser, D. L. Longo, and J. Loscalzo, Eds. New York, NY: McGraw-Hill Education, 2018.
[23] H. Tunstall-Pedoe, M. Woodward, R. Tavendale, R. A'Brook, and M. K. McCluskey, "Comparison of the prediction by 27 different factors of coronary heart disease and death in men and women of the Scottish Heart Health Study: cohort study.," *BMJ*, vol. 315, no. 7110, pp. 722–729, Sep. 1997, doi: 10.1136/bmj.315.7110.722.
[24] G. Andrikopoulos *et al.*, "Epidemiological characteristics, management and early outcome of acute myocardial infarction in Greece: the HELlenic Infarction Observation Study.," *Hellenic J. Cardiol.*, vol. 48, no. 6, pp. 325–334, 2007.
[25] E. Escobar, "Hypertension and coronary heart disease.," *J. Hum. Hypertens.*, vol. 16 Suppl 1, pp. S61-3, Mar. 2002, doi: 10.1038/sj.jhh.1001345.
[26] J. D. Neaton *et al.*, "Serum Cholesterol Level and Mortality Findings for Men Screened in the Multiple Risk Factor Intervention Trial," *Arch. Intern. Med.*, vol. 152, no. 7, pp. 1490–1500, Jul. 1992, doi: 10.1001/archinte.1992.00400190110021.
[27] J.-P. Després, I. Lemieux, G.-R. Dagenais, B. Cantin, and B. Lamarche, "HDL-cholesterol as a marker of coronary heart disease risk: the Québec cardiovascular study," *Atherosclerosis*, vol. 153, no. 2, pp. 263–272, 2000, doi: https://doi.org/10.1016/S0021-9150(00)00603-1.
[28] M. Chiha, M. Njeim, and E. G. Chedrawy, "Diabetes and Coronary Heart Disease: A Risk Factor for the Global Epidemic," *Int. J. Hypertens.*, vol. 2012, p. 697240, 2012, doi: 10.1155/2012/697240.
[29] Centers for Disease Control and Prevention, "Gestational Diabetes and Pregnancy," 2022. [Online]. Available: https://www.cdc.gov/pregnancy/diabetes-gestational.html.
[30] N. Katta, T. Loethen, C. J. Lavie, and M. A. Alpert, "Obesity and Coronary Heart Disease: Epidemiology, Pathology, and Coronary Artery Imaging," *Curr. Probl. Cardiol.*, vol. 46, no. 3, p. 100655, 2021, doi: https://doi.org/10.1016/j.cpcardiol.2020.100655.
[31] R. H. Eckel, "Obesity and heart disease: a statement for healthcare professionals from the Nutrition Committee, American Heart Association.," *Circulation*, vol. 96, no. 9, pp. 3248–3250, Nov. 1997, doi: 10.1161/01.cir.96.9.3248.
[32] G. Eknoyan, "Adolphe Quetelet (1796-1874)--the average man and indices of obesity.," *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.*, vol. 23, no. 1, pp. 47–51, Jan. 2008, doi: 10.1093/ndt/gfm517.
[33] J. Guo, Y. Zhang, T. Liu, B. D. Levy, P. Libby, and G.-P. Shi, "Allergic asthma is a risk factor for human cardiovascular diseases," *Nat. Cardiovasc. Res.*, vol. 1, no. 5, pp. 417–430, 2022, doi: 10.1038/s44161-022-00067-z.
[34] J. A. Diamond and R. A. Phillips, "Hypertensive Heart Disease," *Hypertens. Res.*, vol. 28, no. 3, pp. 191–202, 2005, doi: 10.1291/hypres.28.191.
[35] J. A. Böök, "Heredity and heart disease," *Am. J. Public Heal. Nations Heal.*, vol. 50, no. 3_Pt_2, pp. 1–4, 1960.
[36] R. Aherrahrou *et al.*, "Genetic Regulation of SMC Gene Expression and Splicing Predict Causal CAD Genes.," *Circ. Res.*, vol. 132, no. 3, pp. 323–338, Feb. 2023, doi: 10.1161/CIRCRESAHA.122.321586.

[37] A. H. E. M. Maas and Y. E. A. Appelman, “Gender differences in coronary heart disease.,” *Netherlands Hear. J. Mon. J. Netherlands Soc. Cardiol. Netherlands Hear. Found.*, vol. 18, no. 12, pp. 598–602, Dec. 2010, doi: 10.1007/s12471-010-0841-y.
[38] P. Jousilahti, E. Vartiainen, J. Tuomilehto, and P. Puska, “Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14 786 middle-aged men and women in Finland.,” *Circulation*, vol. 99, no. 9, pp. 1165–1172, Mar. 1999, doi: 10.1161/01.cir.99.9.1165.
[39] S. Sidney, C. Lee, J. Liu, S. S. Khan, D. M. Lloyd-Jones, and J. S. Rana, “Age-Adjusted Mortality Rates and Age and Risk–Associated Contributions to Change in Heart Disease and Stroke Mortality, 2011-2019 and 2019-2020,” *JAMA Netw. Open*, vol. 5, no. 3, pp. e223872–e223872, Mar. 2022, doi: 10.1001/jamanetworkopen.2022.3872.
[40] M. A. Abed, M. I. Kloub, and D. K. Moser, “Anxiety and adverse health outcomes among cardiac patients: a biobehavioral model.,” *J. Cardiovasc. Nurs.*, vol. 29, no. 4, pp. 354–363, Jul. 2014, doi: 10.1097/JCN.0b013e318292b235.
[41] J. D. Bremner *et al.*, “Brain Correlates of Mental Stress-Induced Myocardial Ischemia.,” *Psychosom. Med.*, vol. 80, no. 6, pp. 515–525, 2018, doi: 10.1097/PSY.0000000000000597.
[42] D. Edmondson, I. M. Kronish, J. A. Shaffer, L. Falzon, and M. M. Burg, “Posttraumatic stress disorder and risk for coronary heart disease: A meta-analytic review,” *Am. Heart J.*, vol. 166, no. 5, pp. 806–814, 2013, doi: https://doi.org/10.1016/j.ahj.2013.07.031.
[43] WHO, “Management of physical health conditions in adults with severe mental disorders - WHO Guidelines,” 2018. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/275718/9789241550383-eng.pdf.
[44] S. M. J. *et al.*, “Chronic Kidney Disease and Coronary Artery Disease,” *J. Am. Coll. Cardiol.*, vol. 74, no. 14, pp. 1823–1838, Oct. 2019, doi: 10.1016/j.jacc.2019.08.1017.
[45] WHO, “Salt intake.” [Online]. Available: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3082.
[46] R. W. Hunter, N. Dhaun, and M. A. Bailey, “The impact of excessive salt intake on human health,” *Nat. Rev. Nephrol.*, vol. 18, no. 5, pp. 321–335, 2022, doi: 10.1038/s41581-021-00533-0.
[47] A. Muscella, E. Stefàno, and S. Marsigliante, “The effects of exercise training on lipid metabolism and coronary heart disease,” *Am. J. Physiol. Circ. Physiol.*, vol. 319, no. 1, pp. H76–H88, May 2020, doi: 10.1152/ajpheart.00708.2019.
[48] E. Day and J. H. F. Rudd, “Alcohol use disorders and the heart,” *Addiction*, vol. 114, no. 9, pp. 1670–1678, Sep. 2019, doi: https://doi.org/10.1111/add.14703.
[49] C. J. Smith and T. H. Fischer, “Particulate and vapor phase constituents of cigarette mainstream smoke and risk of myocardial infarction.,” *Atherosclerosis*, vol. 158, no. 2, pp. 257–267, Oct. 2001, doi: 10.1016/s0021-9150(01)00570-6.
[50] A. D. Blann, U. Kirkpatrick, C. Devine, S. Naser, and C. N. McCollum, “The influence of acute smoking on leucocytes, platelets and the endothelium.,” *Atherosclerosis*, vol. 141, no. 1, pp. 133–139, Nov. 1998.
[51] National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health., “The Health Consequences of Smoking—50 Years of Progress A Report of the Surgeon General,” 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK179276/.
[52] R. Heidel, “Reducing the Health Consequences of Smoking: 25 Years of Progress. A Report of the Surgeon General.,” *Popul. Dev. Rev.*, vol. 15, no. 1, p. 165, 1989, doi: 10.2307/1973420.

[53] A. Dunbar, W. Gotsis, and W. Frishman, “Second-hand tobacco smoke and cardiovascular disease risk: an epidemiological review.,” *Cardiol. Rev.*, vol. 21, no. 2, pp. 94–100, 2013, doi: 10.1097/CRD.0b013e31827362e4.
[54] WHO, “WHO report on the global tobacco epidemic 2021,” *2021*. [Online]. Available: https://www.who.int/publications/i/item/9789240032095.
[55] R. Zeng, Y. ting Jiang, T. wu Chen, D. dan Guo, and R. Li, “Longitudinal associations of sleep duration and sleep quality with coronary heart disease risk among adult population: classical meta-analysis and Bayesian network meta-analysis,” *Sleep Biol. Rhythms*, vol. 19, no. 3, pp. 265–276, 2021, doi: 10.1007/s41105-021-00312-1.
[56] G. A. Mensah *et al.*, “Decline in Cardiovascular Mortality: Possible Causes and Implications.,” *Circ. Res.*, vol. 120, no. 2, pp. 366–380, Jan. 2017, doi: 10.1161/CIRCRESAHA.116.309115.
[57] M. A. Papadakis, S. J. McPhee, and M. W. Rabow, *CURRENT Medical Diagnosis & Treatment*. McGraw Hill, 2023.
[58] USPSTF, “Recommendations for Cardiovascular Disorders,” 2023. [Online]. Available: https://www.uspreventiveservicestaskforce.org/uspstf/topic_search_results?topic_status=P.
[59] A. C. of Cardiology, “ASCVD Risk Estimator +,” 2023. [Online]. Available: https://tools.acc.org/ascvd-risk-estimator-plus/#!/calculate/estimate/.
[60] Γ. Κυριόπουλος, “Εθνικό Σχέδιο Δράσης για τα Καρδιαγγειακά Νοσήματα 2008-2012,” *2008*. [Online]. Available: https://www.moh.gov.gr/articles/health/domes-kai-draseis-gia-thn-ygeia/ethnika-sxedia-drashs/95-ethnika-sxedia-drashs?fdl=228.
[61] S. Benjamens, P. Dhunnoo, and B. Meskó, “The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database,” *npj Digit. Med.*, vol. 3, no. 1, pp. 1–8, 2020, doi: 10.1038/s41746-020-00324-0.
[62] FDA, “Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices,” *2022*. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.
[63] J. Azmi, M. Arif, M. T. Nafis, M. A. Alam, S. Tanweer, and G. Wang, “A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data.,” *Med. Eng. Phys.*, vol. 105, p. 103825, Jul. 2022, doi: 10.1016/j.medengphy.2022.103825.
[64] A. L. Samuel, “Some studies in machine learning using the game of checkers,” *IBM J. Res. Dev.*, vol. 44, no. 1.2, pp. 206–226, doi: 10.1147/rd.441.0206.
[65] T. W. MALONE and P. J. MCGOVERN, “ARTIFICIAL INTELLIGENCE AND THE FUTURE OF WORK,” 2020. [Online]. Available: https://workofthefuture.mit.edu/wp-content/uploads/2020/12/2020-Research-Brief-Malone-Rus-Laubacher2.pdf.
[66] Q. Liu and Y. Wu, “Supervised Learning,” Jan. 2012, doi: 10.1007/978-1-4419-1428-6_451.
[67] S. Haykin, *Neural networks and learning machines*, 3rd ed. Pearson Prentice Hall, 2009.
[68] E. Fix and J. L. Hodges, “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties,” *International Statistical Review / Revue Internationale de Statistique*, 1989. [Online]. Available: https://apps.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf.
[69] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[70] T. Mitchell, “Machine Learning,” McGraw-Hill Education, 1997.
[71] D. Wettschereck, D. W. Aha, and T. Mohri, “A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms,” *Artif. Intell. Rev.*, vol. 11, no. 1, pp. 273–314, 1997, doi: 10.1023/A:1006593614256.
[72] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN model-based approach in classification,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2888, pp. 986–996, 2003, doi: 10.1007/978-3-540-39964-3_62.
[73] “Linear Regression VS Logistic Regression Graph.” [Online]. Available: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148.
[74] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, “Introduction to Data Mining,” 2nd ed., Pearson, 2019.
[75] D. R. Cox, “The Regression Analysis of Binary Sequences,” *J. R. Stat. Soc. Ser. B*, vol. 20, no. 2, pp. 215–242, Jun. 1958.
[76] Δ. Πετρίδης, *Ανάλυση Πολυμεταβλητων Τεχνικών, Εφαρμογές Περιπτώσεων*. Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα/ Κάλλιπος, 2015.
[77] M. Maalouf, “Logistic regression in data analysis: An overview,” *Int. J. Data Anal. Tech. Strateg.*, vol. 3, no. 3, pp. 281–299, 2011, doi: 10.1504/IJDATS.2011.041335.
[78] R. J. Rossi, *Applied Biostatistics for the Health Sciences*, 2nd ed. Wiley, 2022.
[79] T. Bayes and null Price, “LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S,” *Philos. Trans. R. Soc. London*, vol. 53, pp. 370–418, Jan. 1763, doi: 10.1098/rstl.1763.0053.
[80] I. Rish, “An Empirical Study of the Naïve Bayes Classifier,” *IJCAI 2001 Work Empir Methods Artif Intell*, vol. 3, Jan. 2001.
[81] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.
[82] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 2000.
[83] D. Anguita, A. Ghio, N. Greco, L. Oneto, and S. Ridella, “Model selection for support vector machines: Advantages and disadvantages of the Machine Learning Theory,” in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1–8, doi: 10.1109/IJCNN.2010.5596450.
[84] T. M. Cover, “Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition,” *IEEE Trans. Electron. Comput.*, vol. EC-14, no. 3, pp. 326–334, 1965, doi: 10.1109/PGEC.1965.264137.
[85] L. Rokach and O. Maimon, “Top-down induction of decision trees classifiers - a survey,” *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 35, no. 4, pp. 476–487, 2005, doi: 10.1109/TSMCC.2004.843247.
[86] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
[87] J. R. Quinlan, “Induction of Decision Trees,” *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1023/A:1022643204877.
[88] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.
[89] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. Routledge, 1984.
[90] C. Gini, *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche*. Tipogr. di P. Cuppini, 1912.
[91] L. Ceriani and P. Verme, “The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini,” *J. Econ. Inequal.*, vol. 10, no. 3, pp. 421–443, 2012, doi: 10.1007/s10888-011-9188-x.

[92] T. Daniya, M. Geetha, and K. S. Kumar, “Classification and regression trees with gini index,” *Adv. Math. Sci. J.*, vol. 9, no. 10, pp. 8237–8247, 2020, doi: 10.37418/amsj.9.10.53.
[93] “Simplified Schematization of Random Trees.” [Online]. Available: https://catalyst.earth/catalyst-system-files/help/concepts/focus_c/oa_classif_intro_rt.html.
[94] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, vol. 1, pp. 278–282 vol.1, doi: 10.1109/ICDAR.1995.598994.
[95] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[96] Wikipedia, “Ensemble Bagging - Bootstrap aggregating.” [Online]. Available: https://en.wikipedia.org/wiki/Bootstrap_aggregating#/media/File:Ensemble_Bagging.svg.
[97] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, “Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics,” *WIREs Data Min. Knowl. Discov.*, vol. 2, no. 6, pp. 493–507, Nov. 2012, doi: https://doi.org/10.1002/widm.1072.
[98] Wikimedia Commons, “Chemical synapse schema cropped.” [Online]. Available: https://commons.wikimedia.org/wiki/File:Chemical_synapse_schema_cropped.jpg.
[99] D. R. Yehoshua, “Perceptrons: The First Neural Network Model.” [Online]. Available: https://towardsdatascience.com/perceptrons-the-first-neural-network-model-8b3ee4513757.
[100] F. ROSENBLATT, “The perceptron: a probabilistic model for information storage and organization in the brain.,” *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, Nov. 1958, doi: 10.1037/h0042519.
[101] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943, doi: 10.1007/BF02478259.
[102] A. G. Ivakhnenko and V. G. Lapa, *Cybernetics and forecasting techniques*, 1st ed. American Elsevier Pub. Co, 1967.
[103] H. Robbins and S. Monro, “A Stochastic Approximation Method,” *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–407, Sep. 1951, doi: 10.1214/aoms/1177729586.
[104] S. Amari, “A Theory of Adaptive Pattern Classifiers,” *IEEE Trans. Electron. Comput.*, vol. EC-16, no. 3, pp. 299–307, 1967, doi: 10.1109/PGEC.1967.264666.
[105] S. Linnainmaa, “Taylor Expansion of the Accumulated Rounding Error,” *BIT*, vol. 16, no. 2, pp. 146–160, 1976, doi: 10.1007/BF01931367.
[106] S. A. Hicks *et al.*, “On evaluation metrics for medical applications of artificial intelligence,” *Sci. Rep.*, vol. 12, no. 1, p. 5979, 2022, doi: 10.1038/s41598-022-09954-8.
[107] K. Hajian-Tilaki, “Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.,” *Casp. J. Intern. Med.*, vol. 4, no. 2, pp. 627–635, 2013.
[108] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/radiology.143.1.7063747.
[109] S. Safari, A. Baratloo, M. Elfil, and A. Negida, “Evidence Based Emergency Medicine; Part 5 Receiver Operating Curve and Area under the Curve.,” *Emerg. (Tehran, Iran)*, vol. 4, no. 2, pp. 111–113, 2016.
[110] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the

ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, pp. 1–21, 2015, doi: 10.1371/journal.pone.0118432.
[111] K. Rosaen, "Diagram of k-fold cross-validation with k = 10."
[112] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, 1995, pp. 1137–1143.
[113] S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed. Packt Publishing, 2019.
[114] S. Fortmann-Roe, "Understanding the Bias-Variance Tradeoff," 2012. [Online]. Available: http://scott.fortmann-roe.com/docs/BiasVariance.html.
[115] Scikit-Learn, "Validation curves_ plotting scores to evaluate models — scikit-learn 1.3.1 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/learning_curve.html#learning-curve.
[116] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset BT - Information and Decision Sciences," 2018, pp. 511–518.
[117] DrGreggHarbaugh, "The 1.5×IQR Rule To Locate Outliers & Modified Box-&-Whisker Plots." [Online]. Available: https://www.youtube.com/watch?v=0YZKL160EDM&t=597s&ab_channel=DrGreggHarbaugh.
[118] C. Albon, *Machine Learning with Python Cookbook*. O'Reilly Media, Inc., 2018.
[119] Ankit songara, "Correlation vs Collinearity vs Multicollinearity," 2022. [Online]. Available: https://medium.com/@songaraankit/correlation-vs-collinearity-vs-multicollinearity-b8e4391617af.
[120] A. Fernández, S. García, M. Galar, and R. C. Prati, *Learning from Imbalanced Data Sets*. 2018.
[121] I. Mani and J. Zhang, "Knn approach to unbalanced data distributions: a case study involving information extraction," in *In Proceedings of workshop on learning from imbalanced dataset, volume 126.*, 2003.
[122] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. February 2017, pp. 321–357, 2002, doi: 10.1613/jair.953.
[123] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.
[124] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.
[125] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Independently published (2022); eBook (Creative Commons Licensed), 2022.
[126] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously.," *J. Mach. Learn. Res.*, vol. 20, 2019.
[127] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[128] Scikit-Learn, "4.2. Permutation feature importance." [Online]. Available: https://scikit-learn.org/stable/modules/permutation_importance.html#permutation-importance.

[129] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, “Explainable AI Methods - A Brief Overview BT - xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers,” A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds. Cham: Springer International Publishing, 2022, pp. 13–38.
[130] CDC, “2020 BRFSS Questionnaire,” 2021. [Online]. Available: https://www.cdc.gov/brfss/questionnaires/pdf-ques/2020-BRFSS-Questionnaire-508.pdf.
[131] Scikit-Learn, “Feature importances with a forest of trees — scikit-learn 1.3.2 documentation.” [Online]. Available: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html.
[132] Scikit-Learn, “Permutation Importance vs Random Forest Feature Importance (MDI) — scikit-learn 1.3.2 documentation.” [Online]. Available: https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html.
[133] A. H. Gonsalves, F. Thabtah, R. M. A. Mohammad, and G. Singh, “Prediction of coronary heart disease using machine learning: An experimental analysis,” *ACM Int. Conf. Proceeding Ser.*, pp. 51–56, 2019, doi: 10.1145/3342999.3343015.
[134] T. Obasi and M. Omair Shafiq, “Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases,” *Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019*, pp. 2393–2402, 2019, doi: 10.1109/BigData47090.2019.9005488.
[135] N. Chandrasekhar and S. Peddakrishna, “Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization,” *Processes*, vol. 11, no. 4, 2023, doi: 10.3390/pr11041210.