



ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Διάγνωση Διαβήτη με Χρήση Τεχνικών Μηχανικής Μάθησης

Δημήτριος Χατζηθανασίου Καρακάσης
A.M. 71141209

Εισηγητής: Νικόλαος Βασιλάς, ΔΕΠ ΠΑΔΑ

ΑΙΓΑΛΕΩ, 2024

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Διάγνωση Διαβήτη με Χρήση Τεχνικών Μηχανικής Μάθησης

Δημήτριος Χατζηαθανασίου Καρακάσης
A.M. 711141209

Επιβλέπων Καθηγητής: Νικόλαος Βασιλάς

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

.....

.....

.....

Ημερομηνία εξέτασης: Ιούλιος 2024

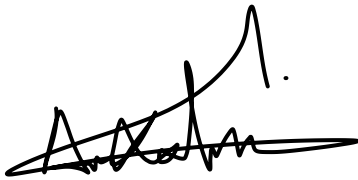
ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Δημήτριος Χατζηαθανασίου - Καρακάσης του Κωνσταντίνου, με αριθμό μητρώου 711141209 φοιτητής του Πανεπιστημίου Δυτικής Αττικής, της Σχολής Μηχανικών, του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο Δηλών



.....
Δημήτριος Χατζηαθανασίου - Καρακάσης

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου Νικόλαο Βασιλά, για την ομαλή συνεργασία και πολύτιμη καθοδήγηση του καθ' όλη τη διάρκεια αυτής της προσπάθειας.

Επιπλέον, ένα μεγάλο ευχαριστώ στην οικογένειά μου, τους συναδέλφους και φίλους μου για τη συνεχή τους συμπαράσταση και ενθάρρυνση κατά τη διάρκεια των σπουδών μου.

ΠΕΡΙΛΗΨΗ

Στη διπλωματική αυτή θα εξεταστούν διάφορες τεχνικές μηχανικής μάθησης όπως πολυστρωματικά νευρωνικά δίκτυα (MLPs), μηχανές διανυσμάτων υποστήριξης (SVM) και δίκτυα ακτινικής βάσης (RBF networks) για τη διάγνωση της ασθένειας του διαβήτη. Για την εκπαίδευση των μοντέλων και την αποτίμηση της γενικευτικής τους ικανότητας θα χρησιμοποιηθεί το Diabetes Prediction Dataset που περιλαμβάνει 100.000 δείγματα. Τέλος, θα δοθεί έμφαση στην επιλογή των υπερπαραμέτρων των διαφόρων μοντέλων (π.χ. ρυθμός εκμάθησης, dropout rate, κλπ.) και θα δοκιμαστούν μέθοδοι συνόλων bagging, boosting για την βελτίωση των αποτελεσμάτων.

ΕΠΙΣΤΗΜΟΝΙΚΗ ΠΕΡΙΟΧΗ: Νευρωνικά Δίκτυα Βαθιάς Μάθησης, Αρχιτεκτονική Νευρωνικών Δικτύων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Python, Μηχανική Μάθηση, Νευρωνικά Δίκτυα, Βαθιά Μάθηση, Πολυστρωματικά Perceptron, Bagging, Πρόβλεψη Διαβήτη

ABSTRACT

This thesis examines various machine learning techniques such as Multilayer Perceptrons (MLPs), Support Vector Machines (SVMs), and Radial Basis Function (RBF) networks for diagnosing diabetes. The Diabetes Prediction Dataset containing 100,000 samples will be used for training the models and evaluating their generalization ability. Furthermore, emphasis will be placed on selecting the hyperparameters of the different models (e.g., learning rate, dropout rate, etc.), and bagging and boosting methods will be experimented with to improve the results.

RESEARCH AREA: Deep Learning Neural Networks, Neural Network Architecture

KEYWORDS: Python, Machine Learning, Neural Networks, Deep Learning, Multilayer Perceptron, Bagging, Diabetes Prediction

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1	15
1. ΕΙΣΑΓΩΓΗ.....	15
1.1 Ο Ρόλος Της Μηχανικής Μάθησης Στον Ιατρικό Κλάδο.....	16
1.2 Προκλήσεις και Βέλτιστες Πρακτικές.....	17
ΚΕΦΑΛΑΙΟ 2	19
2. ΠΑΡΟΥΣΙΑΣΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ.....	19
2.1 Χαρακτηριστικά Δεδομένων.....	19
2.2 Προεπεξεργασία Δεδομένων.....	24
2.2.1 Κωδικοποίηση Παραμέτρων.....	24
2.2.2 Synthetic Minority Over-sampling Technique (SMOTE).....	25
2.2.3 Ανίχνευση Ανωμαλιών και Ενδοτεταρτημοριακό Εύρος (IQR).....	27
2.3 Διερευνητική Ανάλυση Δεδομένων.....	29
2.3.1 Κατανόηση Των Δεικτών Στη Διάγνωση Του Διαβήτη.....	29
2.3.2 Έλεγχος Γραμμικού Διαχωρισμού.....	31
ΚΕΦΑΛΑΙΟ 3	33
3. ΜΟΝΤΕΛΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ.....	33
3.1 Πολυστρωματικά Νευρωνικά Δίκτυα (MLPs).....	36
3.1.1 Εκπαίδευση Μοντέλου MLP.....	37
3.1.2 Εκμάθηση Συνόλου (Ensemble Learning).....	40
3.1.2.1 Μέθοδος Boosting.....	40
3.1.2.2 Διαδικασία Bootstrap.....	41
3.1.2.3 Μέθοδος Bagging.....	41
3.2 Δίκτυα Ακτινικής Βάσης (RBF networks).....	44
3.2.1 Αρχιτεκτονική των Δικτύων RBF.....	44
3.2.2 Εκπαίδευση Μοντέλου RBF.....	45
3.3 Μηχανές Διανυσμάτων Υποστήριξης (SVMs).....	48
3.3.1 Ανάλυση Λειτουργίας SVM Μοντέλων.....	48
3.3.2 Εκπαίδευση Μοντέλων SVM.....	49
3.4 Επιπλέον Παραδείγματα Εκπαίδευσης Μοντέλων.....	51
3.4.1 Δέντρα Απόφασης (Decision Trees).....	51
3.4.2 Ταξινομητής Τυχαίου Δάσους (Random Forest).....	51
ΚΕΦΑΛΑΙΟ 4	53
4. ΣΥΜΠΕΡΑΣΜΑΤΑ.....	53
4.1 Ανασκόπηση και Βασικά Ευρήματα.....	53
4.1.1 Προεπεξεργασία Δεδομένων.....	53
4.1.2 Εκπαίδευση Μοντέλων.....	53
4.2 Συμπεράσματα και Τελικές Παρατηρήσεις.....	54
4.3 Μελλοντικές Εργασίες.....	54
ΒΙΒΛΙΟΓΡΑΦΙΑ	56

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 2.1: Γραφήματα συχνότητας Φύλου/Ηλικίας.....	21
Εικόνα 2.2: Γραφήματα συχνότητας Υπέρτασης/Καρδιακής Νόσου.....	21
Εικόνα 2.3: Γραφήματα συχνότητας Καπνίσματος/BMI.....	22
Εικόνα 2.4: Γραφήματα συχνότητας Επιπέδων HbA1c/Γλυκόζης.....	22
Εικόνα 2.5: Γράφημα συχνότητας Διαβήτη.....	23
Εικόνα 2.6: Αποτέλεσμα Κωδικοποίησης Παραμέτρων.....	25
Εικόνα 2.7: Τροποποίηση Δεδομένων Μετά Την Χρήση SMOTE.....	26
Εικόνα 2.8: Γενικό παράδειγμα διαχωρισμού δεδομένων με IQR.....	27
Εικόνα 2.9: Συγκεντρωτική απεικόνιση δεδομένων (κόκκινο χρώμα θετικά δείγματα, μπλε χρώμα αρνητικά).....	30
Εικόνα 2.10: Μήτρα σύγχυσης μοντέλου λογιστικής παλινδρόμησης.....	32
Εικόνα 3.1: Γράφημα αρχιτεκτονικής ενός MLP με δύο κρυφά επίπεδα [10].....	37
Εικόνα 3.2: Αποτελέσματα ευστοχίας MLP μοντέλου με 10 εποχές.....	38
Εικόνα 3.3: Αποτελέσματα ευστοχίας MLP μοντέλου με 50 εποχές.....	39
Εικόνα 3.4: Βασική αρχιτεκτονική Boosting [13].....	41
Εικόνα 3.5: Αποτελέσματα ευστοχίας τεχνικής Bagging με 6 MLP.....	42
Εικόνα 3.6: Αποτελέσματα ευστοχίας RBF μοντέλου με απλή συνάρτηση ενεργοποίησης.....	46
Εικόνα 3.7: Αποτελέσματα σφάλματος RBF μοντέλου με απλή συνάρτηση ενεργοποίησης.....	46
Εικόνα 3.8 : Αρχιτεκτονική SVM βασισμένο σε δίκτυο RBF [10].....	48
Εικόνα 3.9 : Παράδειγμα από επιφάνειες απόφασης SVM μοντέλων με linear(αριστερά) και RBF(δεξιά) πυρήνα [16].....	49
Εικόνα 3.10 : Αποτελέσματα ευστοχίας με πυρήνα linear και RBF	50
Εικόνα 3.11 : Αποτελέσματα ευστοχίας Decision Tree (αριστερά) και Random Forest (δεξιά)....	52

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 3.1: Συγκεντρωτικά αποτελέσματα ευστοχίας όλων των τεχνικών MLPs43

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

SMOTE Synthetic Minority Over-sampling Technique

IQR Interquartile Range

MLPs Multi-layer Perceptrons

SVM Support Vector Machine

RBF Radial Basis Function Network

ReLU Rectified Linear Unit

ΚΕΦΑΛΑΙΟ 1

1. ΕΙΣΑΓΩΓΗ

Ο διαβήτης αποτελεί ένα σοβαρό και διαδεδομένο πρόβλημα υγείας που αντιμετωπίζουν πολλοί άνθρωποι παγκοσμίως. Η ασθένεια αυτή επηρεάζει την ικανότητα του οργανισμού να ρυθμίζει τα επίπεδα γλυκόζης στο αίμα, και μπορεί να οδηγήσει σε σοβαρές επιπτώσεις στην υγεία όπως καρδιαγγειακές παθήσεις. Σύμφωνα με έρευνα που διεξήχθη σε ελληνικό δείγμα ατόμων, η συνολική επιδημιολογική προσβολή από διαβήτη ανέρχεται στο 11.9% [1]. Αυτό το υψηλό ποσοστό επισημαίνει την ανάγκη για αποτελεσματικές μεθόδους διάγνωσης και διαχείρισης του διαβήτη. Στο πλαίσιο αυτό, έχουν προταθεί και αναπτύχθει διάφορα συστήματα και μέθοδοι που χρησιμοποιούν μηχανική μάθηση ή υπολογιστικά συστήματα για τη διάγνωση του διαβήτη.

Στη διπλωματική αυτή εξετάζεται η χρήση διαφόρων τεχνικών μηχανικής μάθησης για τη διάγνωση του διαβήτη, επικεντρώνοντας σε σύγχρονες προσεγγίσεις και την αξιολόγηση της αποτελεσματικότητάς τους. Επιπλέον, δίνεται έμφαση στην επιλογή και βελτιστοποίηση των υπερπαραμέτρων των μοντέλων, καθώς και η αξιολόγηση των μεθόδων Bagging, Boosting για την βελτίωση της ακρίβειας και της αξιοπιστίας των αποτελεσμάτων.

Πιο συγκεκριμένα, για την εκπαίδευση χρησιμοποιήθηκε συλλογή ηλεκτρονικών ιατρικών εγγράφων που περιλαμβάνει 100.000 δείγματα από ιατρικές εξετάσεις και δημογραφικά χαρακτηριστικά ατόμων. Τα μοντέλα που εκπαιδεύτηκαν είναι πολυστρωματικά νευρωνικά δίκτυα (MLPs), μηχανές διανυσμάτων υποστήριξης (SVM) και δίκτυα ακτινικής βάσης (RBF Networks). Οι τεχνολογίες και τα εργαλεία που χρησιμοποιήθηκαν είναι η γλώσσα Python, βιβλιοθήκες όπως sklearn για τα μοντέλα, pandas για στατιστική ανάλυση των δεδομένων.

1.1 Ο Ρόλος Της Μηχανικής Μάθησης Στον Ιατρικό Κλάδο

Το μεγαλύτερο εμπόδιο στην ανάπτυξη στον τομέα της μηχανικής μάθησης και των νευρωνικών δικτύων ήταν ο μικρός όγκος διαθέσιμης πληροφορίας. Τα τελευταία χρόνια όμως έχει σημειωθεί μεγάλη προσπάθεια από ιδιωτικούς και εθνικούς οργανισμούς για την συλλογή και παραγωγή πληροφοριών σε όλους τους τομείς. Ο κλάδος της ιατρικής είναι ιδιαίτερα αντιμετώπισε αρκετά προβλήματα σε αυτή τη διαδικασία αφού έπρεπε να γίνει ψηφιοποίηση των βάσεων δεδομένων, που μέχρι τότε είχε είτε έγγραφη μορφή ή βρισκόταν σε μορφή δύσκολη να δημοσιευτεί, για παράδειγμα δισκέτες. Επιπλέον υπήρχε το ζήτημα της προστασίας των προσωπικών δεδομένων των ασθενών. Πλέον όμως έχουν δημιουργηθεί διεθνείς βάσεις δεδομένων και πολλά μεγάλα νοσοκομεία μοιράζονται ανωνυμοποιημένα δεδομένα ασθενών, ένα τέτοιο σύνολο δεδομένων χρησιμοποιείται και σε αυτή τη διπλωματική εργασία.

Πλέον οι μηχανές και τα πληροφοριακά συστήματα είναι από τα βασικά εργαλεία του ιατρικού προσωπικού, τόσο για την διεξαγωγή εξετάσεων όσο και για την αρχειοθέτηση δεδομένων ασθενών. Επόμενο λοιπόν ήταν να εμφανιστεί ενδιαφέρον για ανάπτυξη συστημάτων τεχνητής νοημοσύνης στον τομέα της ιατρικής, μάλιστα σε μια δημοσίευση του 2021 φαίνεται εκθετική αύξηση στον αριθμό σχετικών δημοσιεύσεων [2].

Η διάγνωση του διαβήτη συγκεκριμένα, από την οπτική της μηχανικής μάθησης αποτελεί ένα πρόβλημα ταξινόμησης δειγμάτων σε δύο κλάσεις, θετική και αρνητική. Η προσέγγιση αυτής της διπλωματικής εργασίας είναι η διάγνωση διαβήτη βάση εργαστηριακών και δημογραφικών δεδομένων που έχουν συλλεχθεί από τυχαίο πληθυσμό με χρήση διαφορετικών μοντέλων μηχανικής μάθησης και η παραμετροποίηση τους. Παρόμοια παραδείγματα υπάρχουν αρκετά στον επιστημονικό χώρο όπως το άρθρο που παρουσιάστηκε στο διεθνές συνέδριο για τις σύγχρονες τάσεις στην προηγμένη πληροφορική (ICRTAC) το 2019 [3], όπου πέτυχαν ποσοστά επιτυχίας μέχρι και 98.8%.

1.2 Προκλήσεις και Βέλτιστες Πρακτικές

Οι μέθοδοι μηχανικής μάθησης, ενώ είναι ισχυρές, έρχονται με το δικό τους σύνολο προκλήσεων που πρέπει να αντιμετωπιστούν για να εξασφαλιστεί η αποτελεσματικότητα και η αξιοπιστία των μοντέλων.

Αρχικά, η ποιότητα και η ποσότητα των δεδομένων διαδραματίζουν κρίσιμο ρόλο στην απόδοση των μοντέλων μηχανικής μάθησης. Ανεπαρκή ή μεροληπτικά δεδομένα μπορεί να οδηγήσουν σε ανακριβείς προβλέψεις και μεροληπτικά αποτελέσματα. Είναι σημαντικό να διασφαλιστεί ότι τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση είναι αντιπροσωπευτικά, ποικίλα και απαλλαγμένα από σφάλματα. Η επιλογή του σωστού συνόλου χαρακτηριστικών που αντιπροσωπεύουν αποτελεσματικά τα υποκείμενα μοτίβα στα δεδομένα είναι κρίσιμης σημασίας για την απόδοση του μοντέλου. Τεχνικές οι οποίες περιλαμβάνουν τη δημιουργία νέων χαρακτηριστικών ή τον μετασχηματισμό υπάρχοντων, μπορούν να επηρεάσουν σημαντικά την ικανότητα του μοντέλου να μαθαίνει και να γενικεύει.

Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης ρυθμίζονται από υπερπαραμέτρους που πρέπει να συντονιστούν για τη βελτιστοποίηση της απόδοσης του μοντέλου. Η εύρεση του σωστού συνδυασμού υπερπαραμέτρων μπορεί να είναι δύσκολη και συχνά απαιτεί εκτεταμένους πειραματισμούς και διαδικασίες επικύρωσης.

Χρησιμοποιώντας τις παραπάνω διαδικασίες μπορούν να αποφευχθούν σενάρια υπερ-προσαρμογής και υπο-προσαρμογής (overfitting και underfitting) των μοντέλων. Η υπερ-προσαρμογή συμβαίνει όταν ένα μοντέλο “μαθαίνει” να καταγράφει το θόρυβο από τα δεδομένα εκπαίδευσης αντί για το υποκείμενο μοτίβο, με αποτέλεσμα την κακή γενίκευση σε περιπτώσεις που δεν εμφανίζονται στο σύνολο δεδομένων εκπαίδευσης. Από την άλλη πλευρά, η υπο-προσαρμογή συμβαίνει όταν ένα μοντέλο είναι πολύ απλό για να συλλάβει την υποκείμενη δομή των δεδομένων, οδηγώντας σε χαμηλή απόδοση τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα ελέγχου.

Επιπλέον, πολλοί αλγόριθμοι μηχανικής μάθησης, ειδικά μοντέλα βαθιάς μάθησης, απαιτούν σημαντικούς υπολογιστικούς πόρους κατά την διαδικασία εκπαίδευσης και εξαγωγής συμπερασμάτων. Η διασφάλιση της ύπαρξης και παραχώρησης επαρκών υπολογιστικών πόρων, είναι απαραίτητη για την αποτελεσματική διεξαγωγή πειραμάτων και την κλιμάκωση των μοντέλων.

Τέλος, καθώς τα μοντέλα μηχανικής μάθησης χρησιμοποιούνται όλο και περισσότερο στις διαδικασίες λήψης αποφάσεων, η κατανόηση του τρόπου με τον οποίο αυτά τα μοντέλα φτάνουν στις προβλέψεις τους είναι ζωτικής σημασίας.

Η ερμηνεία των μοντέλων είναι απαραίτητη για την οικοδόμηση εμπιστοσύνης και διαφάνειας στα συστήματα τεχνητής νοημοσύνης, ειδικά σε ευαίσθητους τομείς όπως η υγειονομική περίθαλψη και τα οικονομικά.

Το επόμενο κεφάλαιο ασχολείται με το σύνολο δεδομένων που θα χρησιμοποιηθεί, στην επεξεργασία που υπέστει καθώς και στην κατανόηση των χαρακτηριστικών του και της αλληλεπίδρασης μεταξύ τους.

ΚΕΦΑΛΑΙΟ 2

2. ΠΑΡΟΥΣΙΑΣΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Το σύνολο δεδομένων που χρησιμοποιήθηκε περιλαμβάνει ποσοτικά και ποιοτικά δεδομένα ιατρικών και δημογραφικών χαρακτηριστικών ασθενών, μαζί με την ένδειξη αν ήταν θετικοί στον διαβήτη ή όχι. Τα δεδομένα έχουν συλλεχθεί από περιστατικά ασθενών που είχαν προγραμματισμένες εξετάσεις αλλά και από προσελεύσεις τους στα επείγοντα. Τα δεδομένα χωρίστηκαν σε 80% δεδομένα εκπαίδευσης, με τα οποία θα εκπαιδευτούν τα μοντέλα και 20% δεδομένα ελέγχου.

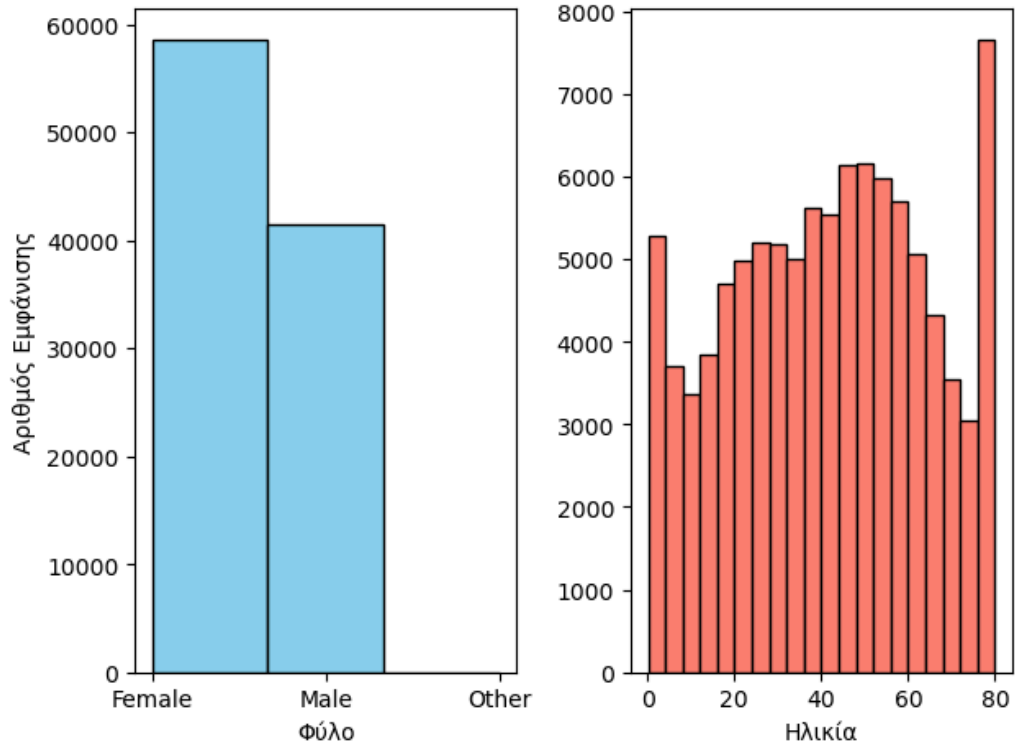
2.1 Χαρακτηριστικά Δεδομένων

Τα χαρακτηριστικά που εμφανίζονται στα δεδομένα είναι εννέα συμπεριλαμβάνοντας την ένδειξη εάν ο ασθενής είχε όντως διαβήτη, η οποία είναι απαραίτητη για την εκπαίδευση των μοντέλων στη συνέχεια.

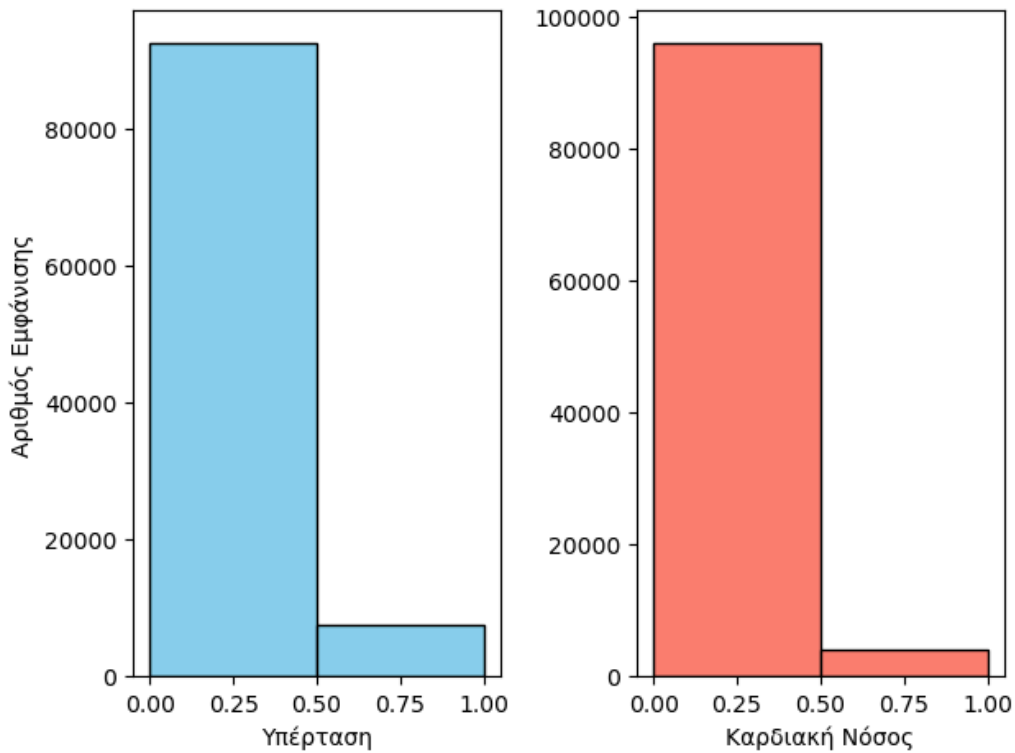
1. **Φύλο:** Αναφέρεται στο βιολογικό φύλο του ατόμου. Υπάρχουν τρεις κατηγορίες: αρσενικό, θηλυκό και άλλο.
2. **Ηλικία:** Η ηλικία είναι ένας σημαντικός παράγοντας καθώς ο διαβήτης διαγιγνώσκεται συνήθως σε ενήλικες μεγαλύτερης ηλικίας.
3. **Υπέρταση:** Έχει τιμές 0 ή 1, όπου 0 υποδεικνύει ότι οι ασθενείς δεν έχουν υπέρταση και 1 σημαίνει ότι έχουν υπέρταση.
4. **Καρδιακή νόσος:** Έχει τιμές 0 ή 1, όπου 0 υποδεικνύει ότι οι ασθενείς δεν έχουν καρδιακή νόσο και 1 σημαίνει ότι έχουν καρδιακή νόσο.
5. **Ιστορικό καπνίσματος:** Το ιστορικό καπνίσματος θεωρείται επίσης παράγοντας κινδύνου για το διαβήτη και μπορεί να επιδεινώσει τις επιπλοκές που σχετίζονται με το διαβήτη. Στο σύνολο δεδομένων υπάρχουν 6 κατηγορίες: Never, Not current, Former, Current, Ever, No Info
6. **Δείκτης μάζας σώματος (BMI):** Ο Δείκτης Μάζας Σώματος (BMI) είναι μια μέτρηση του λίπους του σώματος βασισμένη στο βάρος και το ύψος.
7. **Επίπεδο Γλυκοζυλιωμένης Αιμοσφαιρίνης (HbA1c):** Το επίπεδο HbA1c (Αιμοσφαιρίνη A1c) είναι μια έμμεση μέτρηση του μέσου επιπέδου σακχάρου στο αίμα ενός ατόμου τους τελευταίους 2-3 μήνες.

8. **Επίπεδο γλυκόζης στο αίμα:** Το επίπεδο γλυκόζης στο αίμα αναφέρεται στην ποσότητα γλυκόζης στην κυκλοφορία του αίματος σε μια συγκεκριμένη χρονική στιγμή.
9. **Διαβήτης:** Ο διαβήτης είναι η κατηγορία που προβλέπεται, με τιμές 1 που υποδεικνύει την παρουσία διαβήτη και 0 που υποδεικνύουν την απουσία.

Διάγνωση Διαβήτη με Χρήση Τεχνικών Μηχανικής Μάθησης

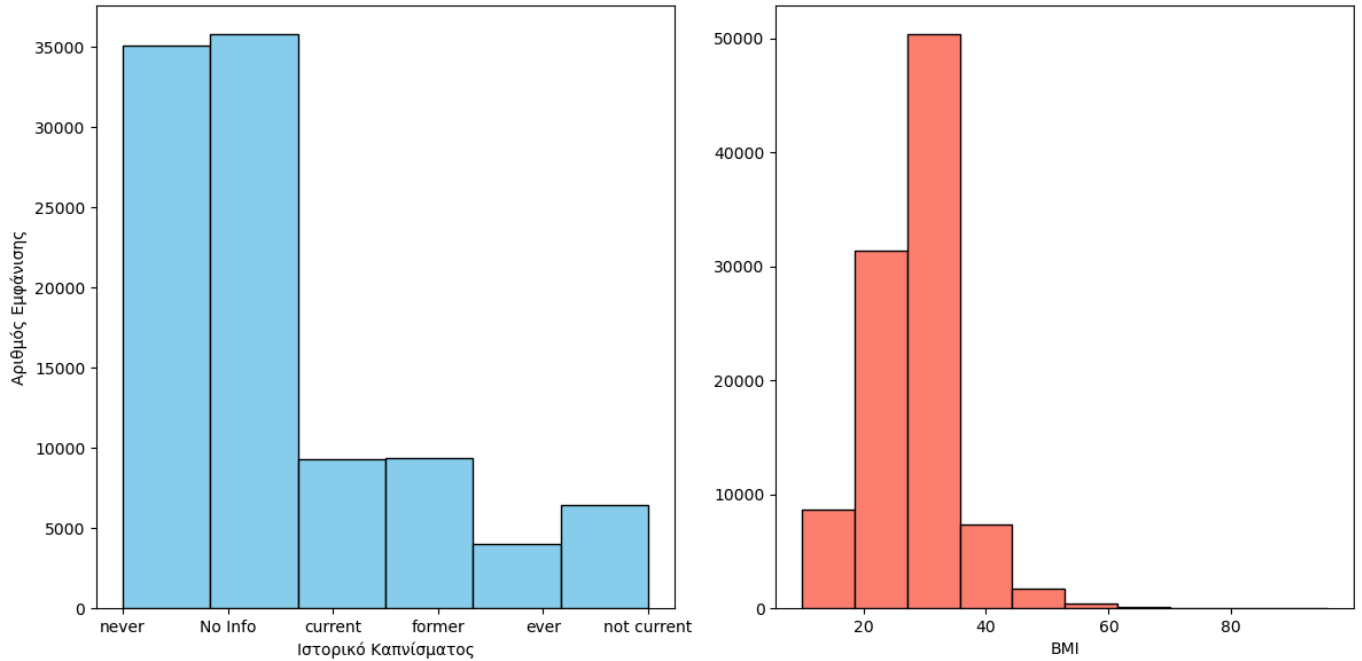


Εικόνα 2.1: Γραφήματα συχνότητας Φύλου/Ηλικίας

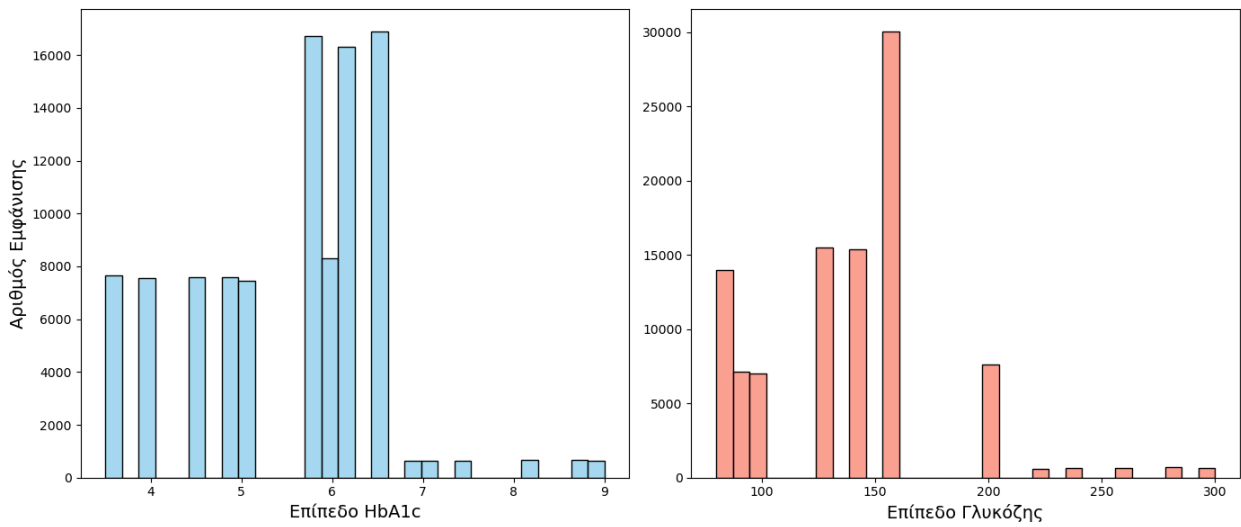


Εικόνα 2.2: Γραφήματα συχνότητας Υπέρτασης/Καρδιακής Νόσου

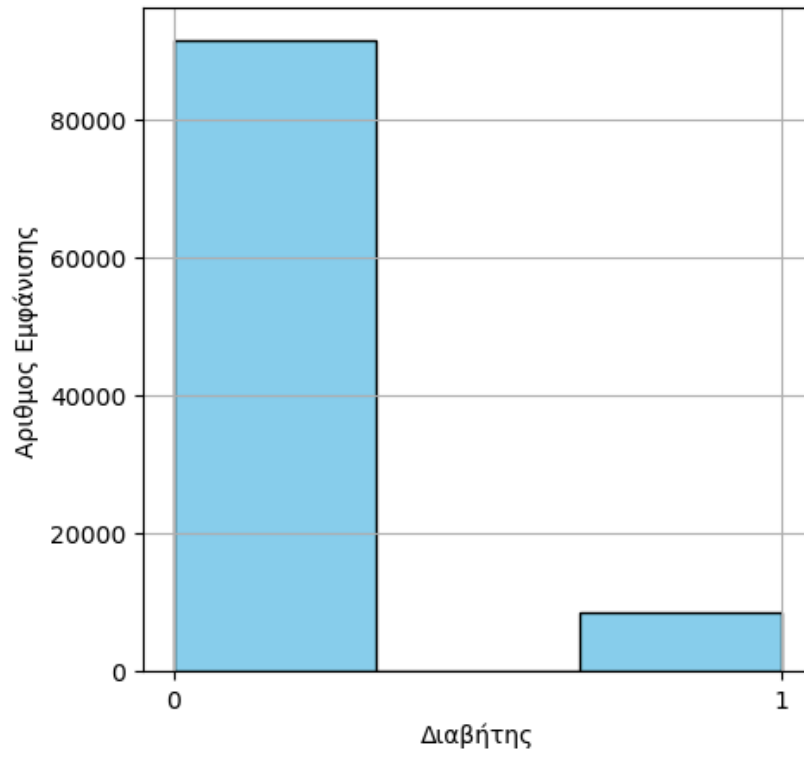
Διάγνωση Διαβήτη με Χρήση Τεχνικών Μηχανικής Μάθησης



Εικόνα 2.3: Γραφήματα συχνότητας Καπνίσματος/BMI



Εικόνα 2.4: Γραφήματα συχνότητας Επιπέδων HbA1c/Γλυκόζης



Εικόνα 2.5: Γράφημα συχνότητας Διαβήτη

2.2 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων αποτελεί ένα κρίσιμο βήμα στην ανάπτυξη ακριβών και αποτελεσματικών μοντέλων μηχανικής μάθησης. Η διαδικασία αυτή περιλαμβάνει μια σειρά από μεθόδους και τεχνικές που εφαρμόζονται στα δεδομένα προκειμένου να βελτιωθεί η ποιότητά τους και να επιτευχθεί καλύτερη απόδοση του μοντέλου που εκπαιδεύεται [4].

Ιδιαίτερα σε κατηγορίες μοντέλων που καλούνται να λύσουν προβλήματα ταξινόμησης, τόσο οι πηγές που αντλούνται τα δεδομένα, όσο και η ποιότητα τους μπορεί να αποτρέψει προβλήματα υπερπροσαρμογής ή υποπροσαρμογής των μοντέλων. Έτσι, για να μεγιστοποιηθεί η απόδοση, η ευστοχία και η ικανότητα του μοντέλου να γενικεύει την γνώση είναι απαραίτητο μεγάλο και ποικίλο πλήθος δεδομένων, τα οποία περιέχουν όσους περισσότερους συνδυασμούς παραμέτρων είναι δυνατό [5].

2.2.1 Κωδικοποίηση Παραμέτρων

Η κωδικοποίηση των παραμέτρων αποτελεί ένα σημαντικό βήμα στην προεπεξεργασία των δεδομένων, ειδικά όταν το σύνολο δεδομένων περιλαμβάνει τόσο ποιοτικές όσο και ποσοτικές παραμέτρους. Η διαδικασία αυτή επιτρέπει τη μετατροπή των κατηγορικών δεδομένων σε αριθμητικές τιμές, κάτι που απαιτείται από πολλούς αλγόριθμους μηχανικής μάθησης για την αποτελεσματική τους λειτουργία. Για παράδειγμα, ένα χαρακτηριστικό όπως το φύλο μπορεί να κωδικοποιηθεί ως [Female, Male, Other] \Rightarrow [0, 1, 2], όπου ο αλγόριθμος αντιστοιχεί την αρχική κατηγορία με έναν αριθμό για την επεξεργασία του. Αυτή η μετατροπή επιτρέπει στο μοντέλο να αντιληφθεί και να αντιμετωπίσει τα κατηγορικά χαρακτηριστικά του συνόλου δεδομένων με τον ίδιο τρόπο με τις ποσοτικές παραμέτρους. Αυτό βοηθάει στη διατήρηση της συνοχής και της ομοιογένειας στην επεξεργασία των δεδομένων, προσφέροντας έτσι μία ολοκληρωμένη και αξιόπιστη ανάλυση.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
...
99995	Female	80.0	0	0	No Info	27.32	6.2	90	0
99996	Female	2.0	0	0	No Info	17.37	6.5	100	0
99997	Male	66.0	0	0	former	27.83	5.7	155	0
99998	Female	24.0	0	0	never	35.42	4.0	100	0
99999	Female	57.0	0	0	current	22.43	6.6	90	0

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	0	80.0	0	1	4	25.19	6.6	140	0
1	0	54.0	0	0	0	27.32	6.6	80	0
2	1	28.0	0	0	4	27.32	5.7	158	0
3	0	36.0	0	0	1	23.45	5.0	155	0
4	1	76.0	1	1	1	20.14	4.8	155	0
...
99995	0	80.0	0	0	0	27.32	6.2	90	0
99996	0	2.0	0	0	0	17.37	6.5	100	0
99997	1	66.0	0	0	3	27.83	5.7	155	0
99998	0	24.0	0	0	4	35.42	4.0	100	0
99999	0	57.0	0	0	1	22.43	6.6	90	0

Εικόνα 2.6: Αποτέλεσμα Κωδικοποίησης Παραμέτρων

2.2.2 Synthetic Minority Over-sampling Technique (SMOTE)

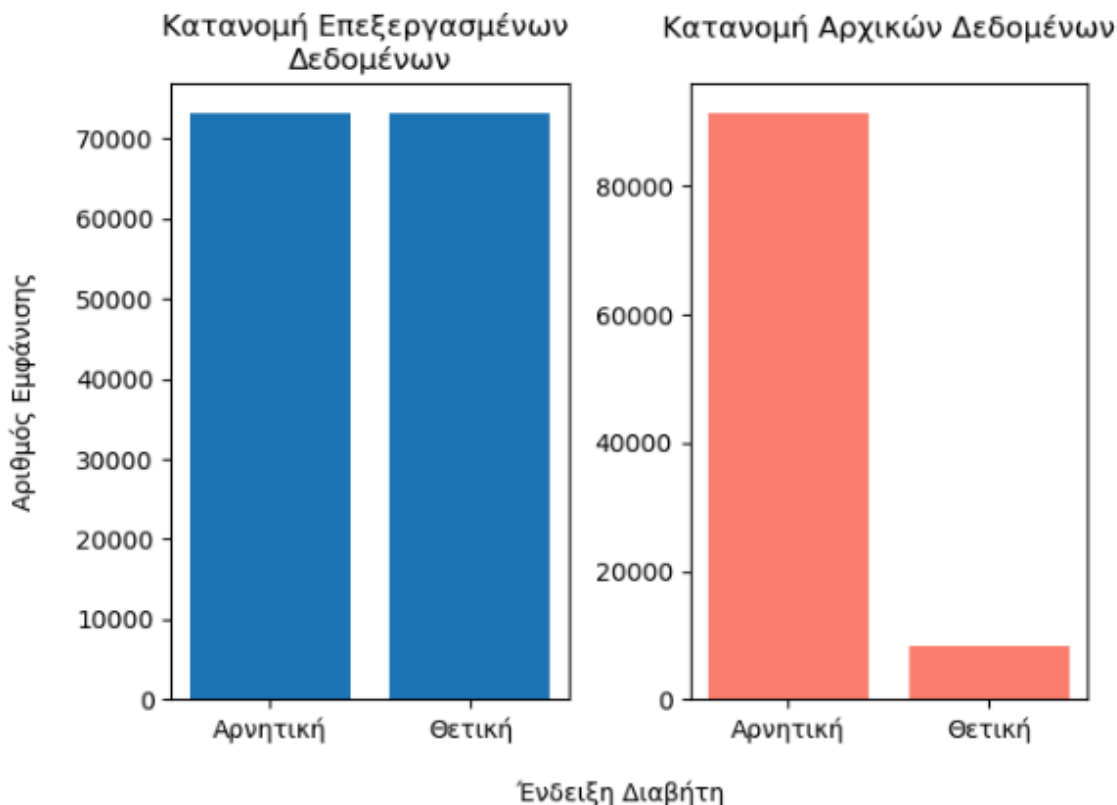
Συχνό φαινόμενο είναι σε ένα σύνολο δεδομένων κάποια κλάση να μην αντιπροσωπεύεται όσο άλλες. Αυτό μπορεί να έχει σαν αποτέλεσμα το μοντέλο να μην μπορεί να διαχωρίσει αυτή τη κλάση σωστά, αφού κατά την διαδικασία εκπαίδευσης δεν υπήρχαν αρκετά παραδείγματα.

Μια τεχνική, ιδιαίτερα διαδεδομένη σε σενάρια όπου μια κλάση ξεπερνά σημαντικά τις άλλες, είναι το SMOTE. Το οποίο με την εισαγωγή συνθετικών δειγμάτων της κλάσης που αποτελεί μειονότητα, στοχεύει να αντιμετωπίσει αυτό το πρόβλημα και να βελτιώσει τη συνολική απόδοση των μοντέλων μηχανικής μάθησης. Έτσι λοιπόν προστίθενται νέα πλασματικά δεδομένα, άρα και το μέγεθος του συνόλου αυξάνεται και έτσι ισορροπείται η αντιπροσώπηση των κλάσεων, αποφεύγοντας έτσι πιθανότητες overfitting του μοντέλου [6].

Το SMOTE έχει εφαρμογές σε διάφορους τομείς, όπως η ιατρική διάγνωση, η ανίχνευση απάτης, και η οικονομική ανάλυση. Για παράδειγμα, στη διάγνωση ασθενειών, η χρήση του SMOTE μπορεί να ενισχύσει την ανίχνευση σπάνιων

περιστατικών, ενώ στην ανίχνευση απάτης, μπορεί να βοηθήσει στην αναγνώριση σπάνιων περιπτώσεων απάτης που διαφορετικά θα περνούσαν απαρατήρητες λόγω της έλλειψης επαρκών δεδομένων.

Σημαντικό να σημειωθεί ότι η τεχνική αυτή εφαρμόζεται μόνο στο υποσύνολο των δεδομένων στο οποίο θα εκπαιδευτεί το μοντέλο και όχι σε αυτό που χρησιμοποιείται για τον έλεγχο του. Αυτό είναι σημαντικό για να ελέγχεται κάθε μοντέλο σε πραγματικά δεδομένα.



Εικόνα 2.7: Τροποποίηση Δεδομένων Μετά Την Χρήση SMOTE

Στο παραπάνω σχήμα βλέπουμε ότι η κλάση των θετικών σε διαβήτη ασθενών έχει φτάσει σε αριθμό την κλάση των αρνητικών, με αποτέλεσμα να αυξηθεί όμως σημαντικά ο αριθμός των δειγμάτων.

Το SMOTE λοιπόν, αποτελεί μια αποτελεσματική τεχνική για την αντιμετώπιση των προβλημάτων ανισορροπίας των κλάσεων στα σύνολα δεδομένων. Μέσω της δημιουργίας συνθετικών δειγμάτων, μπορεί να βελτιωθεί η απόδοση των μοντέλων μηχανικής μάθησης, ειδικά σε σενάρια όπου η μειονότητα των δεδομένων έχει κρίσιμη σημασία.

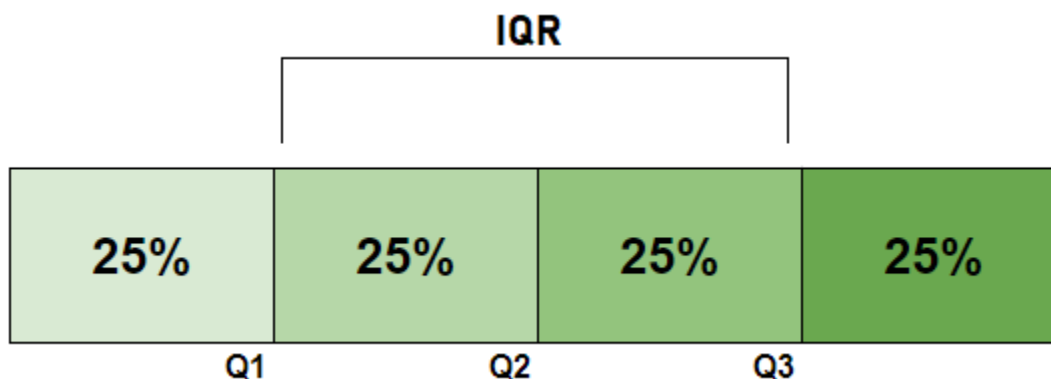
2.2.3 Ανίχνευση Ανωμαλιών και Ενδοτεταρτημοριακό Εύρος (IQR)

Η χρήση του IQR είναι μια σημαντική τεχνική προεπεξεργασίας δεδομένων που βοηθά στη βελτίωση της ποιότητας των δεδομένων που χρησιμοποιούνται για την εκπαίδευση μοντέλων μηχανικής μάθησης, αφαιρώντας ακραίες τιμές (outliers) από το σύνολο δεδομένων.

Οι ακρέες αυτές τιμές μπορούν να προκαλέσουν παραμορφώσεις στα στατιστικά χαρακτηριστικά των δεδομένων, να επηρεάσουν αρνητικά την ακρίβεια και την απόδοση των μοντέλων μηχανικής μάθησης, οδηγώντας σε μη αξιόπιστες προβλέψεις [7].

Αναλυτικότερα, τα δεδομένα χωρίζονται σε τέσσερα τέταρτα, βάση των τιμών τους. Έπειτα με τη χρήση ενός ορίου (threshold). Ένα μέρος των τιμών χαρακτηρίζονται ως ακραίες τιμές και διαγράφονται από το σύνολο δεδομένων με την παρακάτω λογική:

$$Q1 - threshold * IQR > Outliers > Q3 + threshold * IQR$$



Εικόνα 2.8: Γενικό παράδειγμα διαχωρισμού δεδομένων με IQR

Η επιλογή του κατάλληλου threshold για τον καθορισμό των ακραίων τιμών είναι κρίσιμη και εξαρτάται από την εκάστοτε περίπτωση χρήσης. Σε κάποιες εφαρμογές, όπως η ανάλυση ιατρικών δεδομένων, μπορεί να απαιτείται ένα αυστηρότερο threshold λόγω της ανάγκης για ακριβή ανάλυση και πρόγνωση. Αντίθετα, σε άλλες εφαρμογές, όπως η ανάλυση καταναλωτικών προτύπων, μπορεί να επιτρέπεται μεγαλύτερη ανοχή σε ακραίες τιμές, προκειμένου να αναγνωριστούν ενδιαφέροντα πρότυπα συμπεριφοράς.

Η χρήση του IQR δεν περιορίζεται μόνο στην ανίχνευση ανωμαλιών αλλά έχει και σημαντικές επιπτώσεις στη στατιστική ανάλυση και την επεξεργασία

δεδομένων. Μέσω της αφαίρεσης των ακραίων τιμών, βελτιώνεται η ακρίβεια των στατιστικών μέτρων, όπως η μέση τιμή (mean) και η τυπική απόκλιση (standard deviation). Επιπλέον, η μείωση της παραμόρφωσης που προκαλούν οι ακραίες τιμές επιτρέπει την πιο αξιόπιστη ανάλυση τάσεων και την καλύτερη κατανόηση των δεδομένων.

Η τεχνική του IQR αποτελεί ένα ισχυρό εργαλείο στην ανίχνευση outliers. Μέσω της αφαίρεσης ακραίων τιμών, βελτιώνεται η ποιότητα των δεδομένων και κατά συνέπεια η απόδοση των μοντέλων μηχανικής μάθησης. Επιπλέον, η εφαρμογή του IQR σε συνδυασμό με άλλες τεχνικές και η προσαρμογή του threshold, είναι ιδιαίτερα ευέλικτη και αποτελεσματική σε ποικίλα πεδία εφαρμογής.

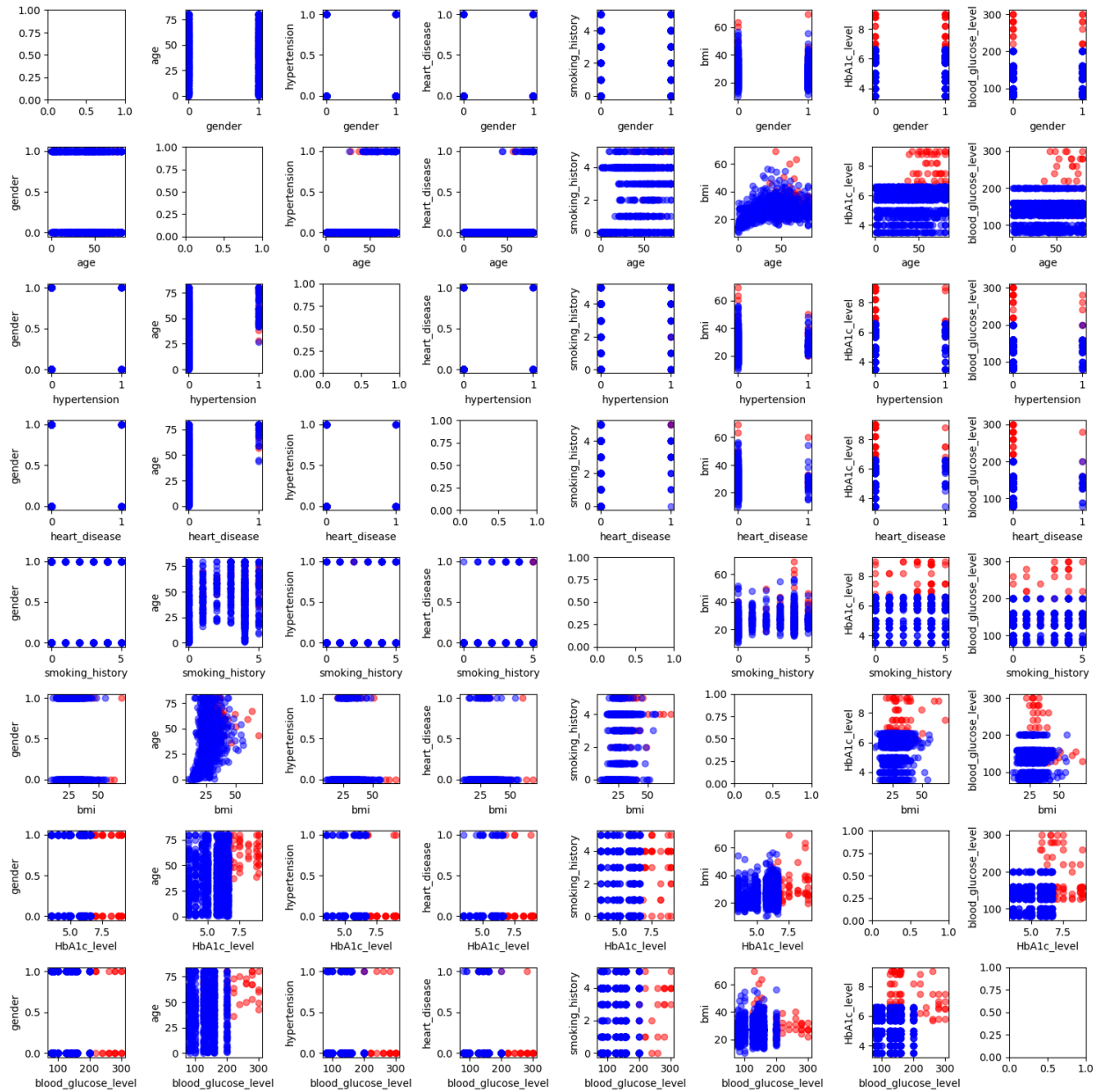
2.3 Διερευνητική Ανάλυση Δεδομένων

Σε κάθε περίπτωση, πολύ σημαντικό βήμα πριν από την εκπαίδευση μοντέλων είναι η δημιουργία ενός πλαισίου, βάση του οποίου θα κριθούν τα αποτελέσματα που θα προκύψουν. Βασική προϋπόθεση για αυτό είναι η κατανόηση τόσο του προβλήματος προς επίλυση όσο και του συνόλου δεδομένων, η προέλευση του, τα χαρακτηριστικά του και η σχέσεις μεταξύ αυτών.

2.3.1 Κατανόηση Των Δεικτών Στη Διάγνωση Του Διαβήτη

Αρχικά πρέπει να σημειωθεί ότι η θετική ή αρνητική διάγνωση γίνεται από ιατρικό προσωπικό βάση μιας συγκεκριμένης μεθοδολογίας βασιζόμενη σε εξετάσεις. Όπως αναφέρει και το CDC (Centers for Disease Control and Prevention) στην επίσημη ιστοσελίδα του, ικανά κριτήρια για τη διάγνωση του διαβήτη είναι το ποσοστό γλυκοζυλιωμένης αιμοσφαιρίνης (HbA1c) μεγαλύτερο των 6.5% και το επίπεδο γλυκόζης μεγαλύτερο από 126 mg/dL στο αίμα [8]. Αυτά είναι δύο χαρακτηριστικά που υπάρχουν στα δεδομένα και θα έπρεπε όλα τα δείγματα που πληρούν τις παραπάνω προϋποθέσεις να ανήκουν στην κλάση των θετικών περιπτώσεων διαβήτη. Παρακάτω παρουσιάζεται μια ανάλυση των δεδομένων, συγκρίνοντας κάθε χαρακτηριστικό του δείγματος με όλα τα άλλα.

Διάγνωση Διαβήτη με Χρήση Τεχνικών Μηχανικής Μάθησης



Εικόνα 2.9: Συγκεντρωτική απεικόνιση δεδομένων (κόκκινο χρώμα θετικά δείγματα, μπλε χρώμα αρνητικά)

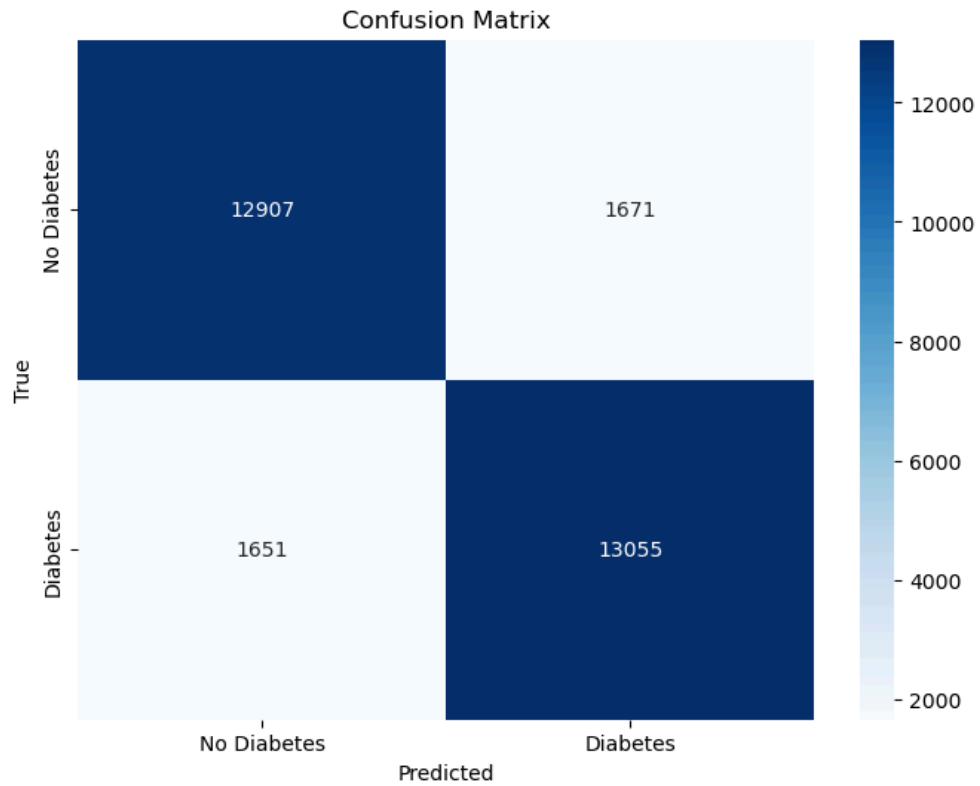
Από την παραπάνω εικόνα υπάρχει σαφής συσχέτιση μεταξύ των τιμών HbA1c (έβδομη στήλη), επίπεδο γλυκόζης (όγδοη στήλη) και της διάγνωσης. Μάλιστα για τα δύο χαρακτηριστικά που αναφέρθηκαν τα δεδομένα είναι γραμμικώς διαχωρίσιμα ακριβώς στα όρια που χρησιμοποιούνται στην πράξη για τη διάγνωση.

2.3.2 Έλεγχος Γραμμικού Διαχωρισμού

Γραμμικά διαχωρίσιμα πρότυπα ονομάζονται τα πρότυπα τα οποία βρίσκονται σε αντίθετες πλευρές ενός υπερεπιπέδου. Ένας από τους τρόπους για να ελεγχθεί εάν είναι πράγματι γραμμικά διαχωρίσιμα, είναι η εκπαίδευση ενός γραμμικού μοντέλου μηχανικής μάθησης και ο έλεγχος της απόδοσης του. Εάν το ποσοστό ευστοχίας του είναι κοντά σε αυτό της τυχαίας επιλογής τότε δεν είναι γραμμικά διαχωρίσιμο και αντίθετα εάν είναι 100% τότε είναι πλήρως γραμμικά διαχωρίσιμο. Στην πράξη βέβαια συνθήκες όπως ο θόρυβος ή λάθη στη διαδικασία παραγωγής/συλλογής δεδομένων μπορεί να χαμηλώσουν αυτή τη γραμμικότητα.

Για τον έλεγχο των δεδομένων που θα χρησιμοποιηθούν στην συγκεκριμένη διπλωματική εργασία, εκπαιδεύτηκαν δύο διαφορετικά γραμμικά μοντέλα. Ένα SVM με γραμμικό πυρήνα το οποίο κατάφερε να κατασκευάσει υπερεπίπεδο που διαχώρισε τα δεδομένα με ποσοστό επιτυχίας άνω των 95%. Το άλλο είναι ένα μοντέλο λογιστικής παλινδρόμησης (Logistic regression) που πέτυχε ποσοστό επιτυχίας 89%. Συνεπώς τα πρότυπά είναι γραμμικά διαχωρίσιμα σε ένα μεγάλο βαθμό. Με άλλα λόγια γραμμικά μοντέλα μπορούν να πετύχουν υψηλές αποδόσεις εκπαιδευόμενα με αυτό το σύνολο δεδομένων. Πιθανή αιτία για την γραμμικότητα αυτή είναι η βαρύτητα των χαρακτηριστικών που αναφέρθηκαν παραπάνω, αφού από μόνα τους είναι, θεωρητικά και πρακτικά, ικανές συνθήκες για την ταξινόμηση των κλάσεων.

Διάγνωση Διαβήτη με Χρήση Τεχνικών Μηχανικής Μάθησης



Εικόνα 2.10: Μήτρα σύγχυσης μοντέλου λογιστικής παλινδρόμησης

ΚΕΦΑΛΑΙΟ 3

3. ΜΟΝΤΕΛΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Η ταξινόμηση είναι μια διαδικασία, στη μηχανική μάθηση, με στόχο να κατηγοριοποιηθούν τα δεδομένα εισόδου σε μία από προκαθορισμένες κλάσεις ή κατηγορίες. Σε αυτή τη διπλωματική εργασία, οι κατηγορίες του συνόλου δεδομένων είναι δύο: είτε ο ασθενής έχει διαβήτη είτε όχι.

Για την επίλυση αυτού του προβλήματος, χρησιμοποιήθηκαν διάφορα μοντέλα μηχανικής μάθησης, το καθένα με τα δικά του χαρακτηριστικά και τεχνικές. Ακολουθεί μια επισκόπηση των χαρακτηριστικών αυτών των μοντέλων:

- **Δεδομένα με ετικέτα:** Όλα τα μοντέλα μηχανικής εκμάθησης που θα χρησιμοποιηθούν για ταξινόμηση απαιτούν δεδομένα με ετικέτα για εκπαίδευση. Τα επισημασμένα δεδομένα αποτελούνται από δείγματα εισόδου μαζί με τις αντίστοιχες ετικέτες τους. Στην συγκεκριμένη περίπτωση, τα δείγματα εισόδου είναι τα ιατρικά και δημογραφικά χαρακτηριστικά των ασθενών και οι ετικέτες είναι δυαδικοί δείκτες για το εάν έχουν ή όχι διαβήτη.
- **Επιβλεπόμενη μάθηση:** Οι εργασίες ταξινόμησης συνήθως εμπίπτουν στην ομπρέλα της επιβλεπόμενης μάθησης, όπου τα μοντέλα μαθαίνουν να χαρτογραφούν τα χαρακτηριστικά εισόδου σε ετικέτες στόχευσης με βάση παραδείγματα που παρέχονται κατά τη διάρκεια της εκπαίδευσης. Τα μοντέλα εκπαιδεύονται σε ένα επισημασμένο σύνολο δεδομένων, όπου παρέχονται οι σωστές ετικέτες κλάσεων για κάθε δείγμα εισόδου [9].
- **Ρυθμός μάθησης:** Ο ρυθμός μάθησης καθορίζει το μέγεθος των βημάτων που λαμβάνονται κατά τη διάρκεια της διαδικασίας βελτιστοποίησης πριν την ενημέρωση των παραμέτρων του μοντέλου. Ένας κατάλληλος ρυθμός μάθησης είναι ζωτικής σημασίας για την αποτελεσματική σύγκλιση και την απόδοση του μοντέλου.
- **Μαζική και On-Line Μάθηση:** Ορισμένα μοντέλα μηχανικής μάθησης, όπως οι μηχανές διανυσμάτων υποστήριξης, μπορούν να εκπαιδευτούν χρησιμοποιώντας μαζική εκμάθηση, όπου ο κύκλος

προσαρμογής των βαρών του μοντέλου επαναλαμβάνεται μετά του συνόλου των παραδειγμάτων των δεδομένων (batch size). Σε αντίθεση με την On-Line μάθηση όπου ο κύκλος επαναλαμβάνεται μετά από κάθε παράδειγμα [10].

- **Συναρτήσεις ενεργοποίησης:** Εκτός από τα αναφερόμενα χαρακτηριστικά, η επιλογή των συναρτήσεων ενεργοποίησης διαδραματίζει κρίσιμο ρόλο στη συμπεριφορά και την απόδοση των νευρωνικών δικτύων. Οι συναρτήσεις ενεργοποίησης εισάγουν μη γραμμικότητα στο δίκτυο, επιτρέποντάς του να μαθαίνει πολύπλοκα μοτίβα και σχέσεις στα δεδομένα. Κάθε συνάρτηση ενεργοποίησης έχει τις δικές της ιδιότητες και είναι κατάλληλη για διαφορετικούς τύπους εργασιών και αρχιτεκτονικών. Η επιλογή της κατάλληλης συνάρτησης ενεργοποίησης είναι απαραίτητη για τη διασφάλιση αποτελεσματικής μάθησης και σύγκλισης των μοντέλων νευρωνικών δικτύων.
- **Σφάλμα/Ποσοστό λάθους (loss/error):** το σφάλμα ή το ποσοστό λάθους αναφέρεται στο μέτρο του πόσο καλά ή ανεπαρκώς ταιριάζουν οι προβλέψεις ενός μοντέλου με τα πραγματικά δεδομένα. Προσδιορίζει ποσοτικά τη διαφορά μεταξύ των προβλεπόμενων τιμών που παράγονται από το μοντέλο και των πραγματικών τιμών από το σύνολο δεδομένων. Ο στόχος της εκπαιδευτικής διαδικασίας είναι να ελαχιστοποιηθεί αυτή η απώλεια, βελτιώνοντας έτσι την απόδοση του μοντέλου. Οι συνήθεις συναρτήσεις απώλειας περιλαμβάνουν το μέσο τετραγωνικό σφάλμα (MSE) για εργασίες παλινδρόμησης και τη δυαδική διασταυρούμενη εντροπία για εργασίες δυαδικής ταξινόμησης.
- **Ακρίβεια (accuracy):** η ακρίβεια, από την άλλη πλευρά, είναι μια μέτρηση που χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης. Αντιπροσωπεύει την αναλογία των σωστά προβλεπόμενων παρουσιών από το σύνολο δεδομένων. Η ακρίβεια υπολογίζεται ως ο αριθμός των αληθινών θετικών και των αληθινών αρνητικών προβλέψεων διαιρεμένος με τον συνολικό αριθμό των προβλέψεων. Αν και η ακρίβεια είναι ένα χρήσιμο μέτρο, είναι σημαντικό να λαμβάνεται υπόψη σε συνδυασμό με άλλες μετρήσεις,

ειδικά σε περιπτώσεις μη ισορροπημένων συνόλων δεδομένων όπου η ακρίβεια από μόνη της μπορεί να είναι παραπλανητική.

Παρακάτω παρουσιάζεται η εκπαίδευση των μοντέλων μηχανικής μάθησης. Επιπλέον δίνεται σημαντική προσοχή στην παραμετροποίηση τους και πώς αυτή μαζί με διάφορες άλλες τεχνικές επηρεάζουν την απόδοσή τους.

3.1 Πολυστρωματικά Νευρωνικά Δίκτυα (MLPs)

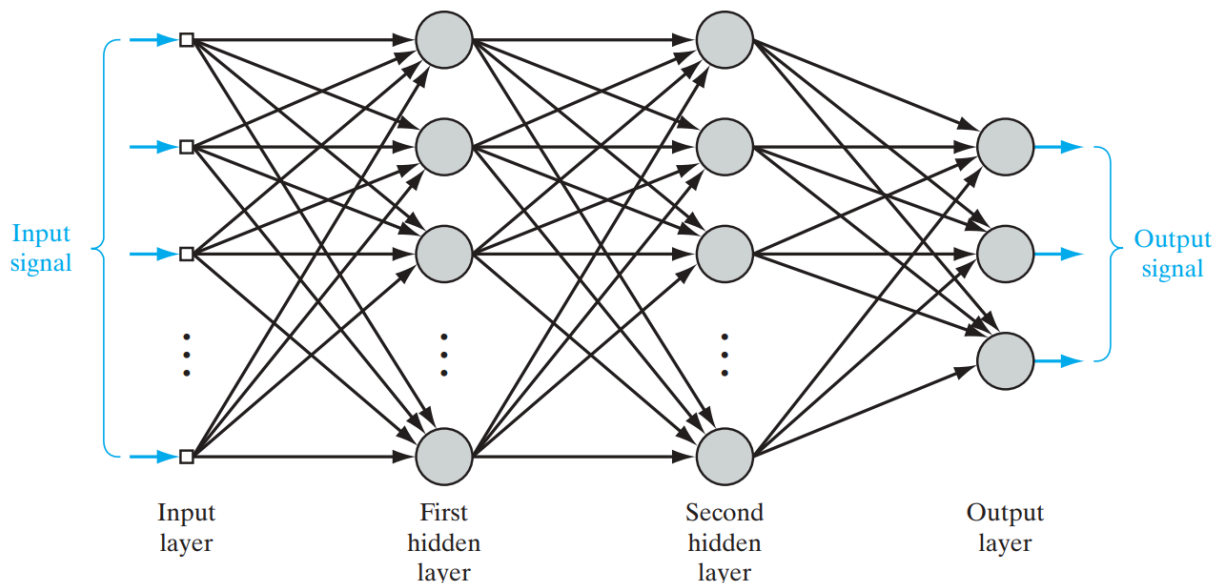
Τα πολυστρωματικά νευρωνικά δίκτυα είναι μια υποκατηγορία των τεχνητών νευρωνικών δικτύων που χρησιμοποιούνται συχνά σε προβλήματα κατηγοριοποίησης. Η αρχιτεκτονική τους αποτελείται από ένα επίπεδο εισόδου, από το οποίο εισέρχεται το σήμα εισόδου (ερέθισμα). Έναν αριθμό κρυφών επιπέδων, όπου σε κάθε ένα από αυτά υπάρχει ένας αριθμός κρυφών νευρώνων [10]. Τέλος υπάρχει το επίπεδο εξόδου από το οποίο αντλείται η πρόβλεψη/αντίδραση του μοντέλου για κάθε ερέθισμα. Πιο αναλυτικά:

Το επίπεδο εισόδου είναι το σημείο από το οποίο εισέρχεται το σήμα εισόδου στο δίκτυο. Τα δεδομένα εισόδου τροφοδοτούνται στους νευρώνες του επιπέδου αυτού και κάθε νευρώνας στο επίπεδο εισόδου αντιστοιχεί σε ένα χαρακτηριστικό του δεδομένου εισόδου.

Τα κρυφά επίπεδα αποτελούνται από έναν αριθμό κρυφών νευρώνων οι οποίοι λειτουργούν ως ανιχνευτές χαρακτηριστικών (feature detectors). Κάθε κρυφός νευρώνας λαμβάνει σήματα από τους νευρώνες του προηγούμενου επιπέδου και εκτελεί μαθηματικούς υπολογισμούς για να ανιχνεύσει μοτίβα ή χαρακτηριστικά. Κάθε κρυφός νευρώνας εφαρμόζει μια συνάρτηση ενεργοποίησης στα σήματα που λαμβάνει, όπως η ReLU (Rectified Linear Unit). Πολλαπλά κρυφά επίπεδα μπορούν να συνδυαστούν για να δημιουργήσουν βαθιά νευρωνικά δίκτυα, τα οποία είναι ικανά να μάθουν και να αναπαραστήσουν σύνθετες σχέσεις στα δεδομένα.

Το επίπεδο εξόδου είναι το τελικό επίπεδο του δικτύου, από το οποίο αντλείται η πρόβλεψη ή η αντίδραση του μοντέλου για κάθε ερέθισμα. Ο αριθμός των νευρώνων στο επίπεδο εξόδου εξαρτάται από τον αριθμό των κλάσεων στο πρόβλημα κατηγοριοποίησης. Για παράδειγμα, σε ένα πρόβλημα δυαδικής ταξινόμησης, όπως στο πρόβλημα με το οποίο ασχολείται η συγκεκριμένη διπλωματική εργασία, υπάρχει ένας νευρώνας εξόδου με συνάρτηση ενεργοποίησης Sigmoid που παράγει μια τιμή μεταξύ 0 και 1. Σε προβλήματα πολυκατηγοριακής ταξινόμησης, χρησιμοποιείται συχνά η συνάρτηση ενεργοποίησης Softmax, η οποία παρέχει μια πιθανότητα για κάθε κλάση, και ο νευρώνας με τη μεγαλύτερη πιθανότητα καθορίζει την τελική πρόβλεψη.

Παρακάτω φαίνεται γράφημα της αρχιτεκτονικής ενός πολυστρωματικού νευρωνικού δικτύου με δύο κρυφά στρώματα.



Εικόνα 3.1: Γράφημα αρχιτεκτονικής ενός MLP με δύο κρυφά επίπεδα [10].

3.1.1 Εκπαίδευση Μοντέλου MLP

Η διαδικασία εκπαίδευσης των πολυστρωματικών νευρωνικών δικτύων αποτελεί ένα σημαντικό κομμάτι της μηχανικής μάθησης και περιλαμβάνει συγκεκριμένα βήματα για την απόκτηση της ικανότητας του δικτύου να παράγει ακριβείς προβλέψεις.

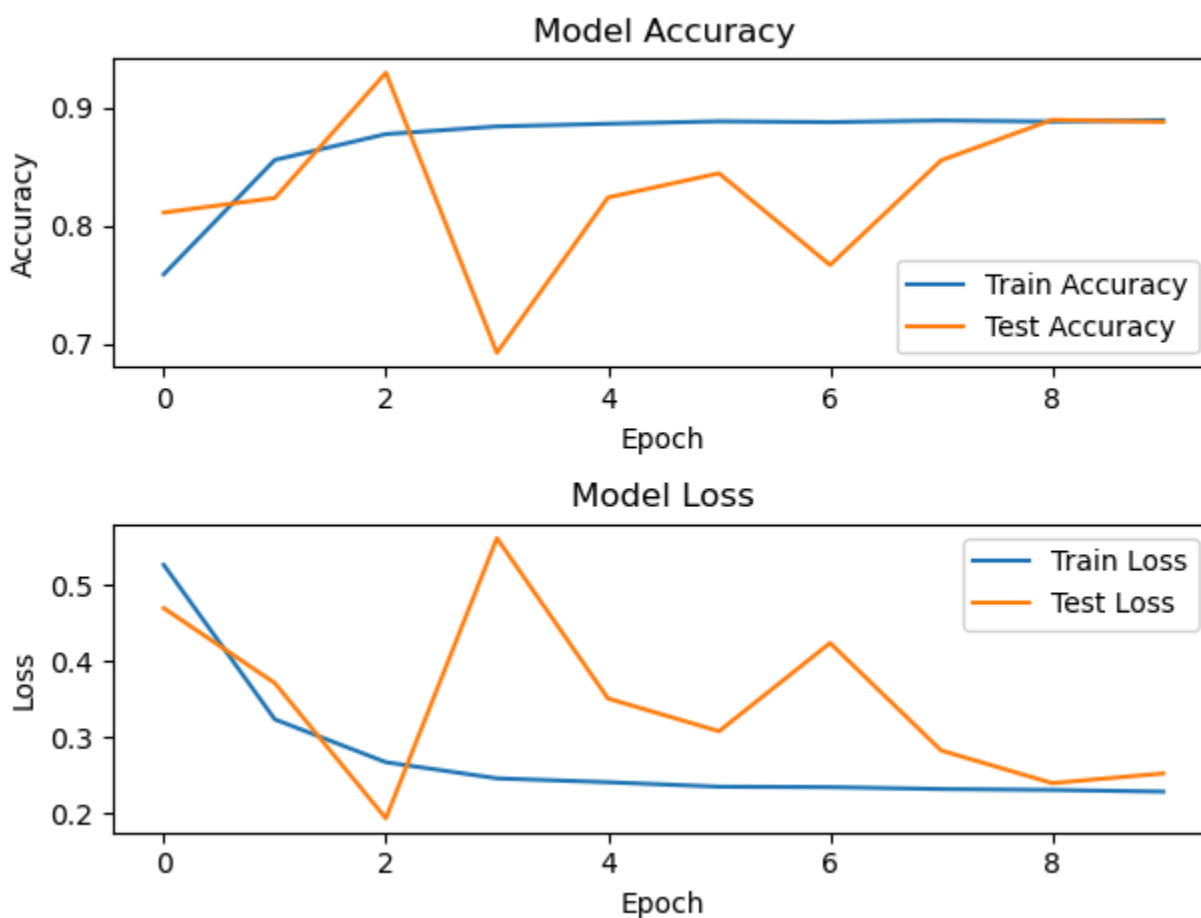
Κατά τη διαδικασία αυτή, τα δεδομένα εισόδου περνούν μέσα από το δίκτυο σταδιακά, επίπεδο-επίπεδο, με τη διαδικασία της εμπρόσθιας διάδοσης (Forward Propagation). Κατά τη διαδικασία αυτή, οι νευρώνες κάθε επιπέδου υπολογίζουν τα βάρη των εισόδων τους και τις συναρτήσεις ενεργοποίησης για να παράγουν την έξοδο, η οποία γίνεται η είσοδος για το επόμενο επίπεδο.

Στη συνέχεια, υπολογίζεται το σφάλμα με την σύγκριση της προβλεπόμενης εξόδου με την πραγματική τιμή στόχο, χρησιμοποιώντας μια συνάρτηση απώλειας, όπως η Cross-Entropy Loss για κατηγοριοποίηση.

Έπειτα, ακολουθεί η διαδικασία της οπίσθιας διάδοσης (Backpropagation), όπου το σφάλμα διαδίδεται πίσω μέσω του δικτύου για να υπολογιστούν οι βαθμίδες των βαρών και να ενημερωθούν αναλόγως. Τα βάρη των συνδέσεων μεταξύ των νευρώνων ενημερώνονται με βάση τις βαθμίδες που υπολογίστηκαν κατά την οπίσθια διάδοση. Η διαδικασία αυτή επαναλαμβάνεται για πολλούς κύκλους (epochs) μέχρι το δίκτυο να συγκλίνει σε μια λύση που ελαχιστοποιεί το σφάλμα. Κάθε επανάληψη αυτής της διαδικασίας βελτιώνει την ακρίβεια των προβλέψεων και συνεπώς την απόδοση του δικτύου.

Παρακάτω παρουσιάζεται η εκπαίδευση ενός MLP μοντέλου και η παραμετροποίηση του. Επιπλέον υπάρχουν παραδείγματα της διαδικασίας επιλογής υπαραμέτρων και θα δοκιμαστούν μέθοδοι συνόλων για την βελτίωση των αποτελεσμάτων.

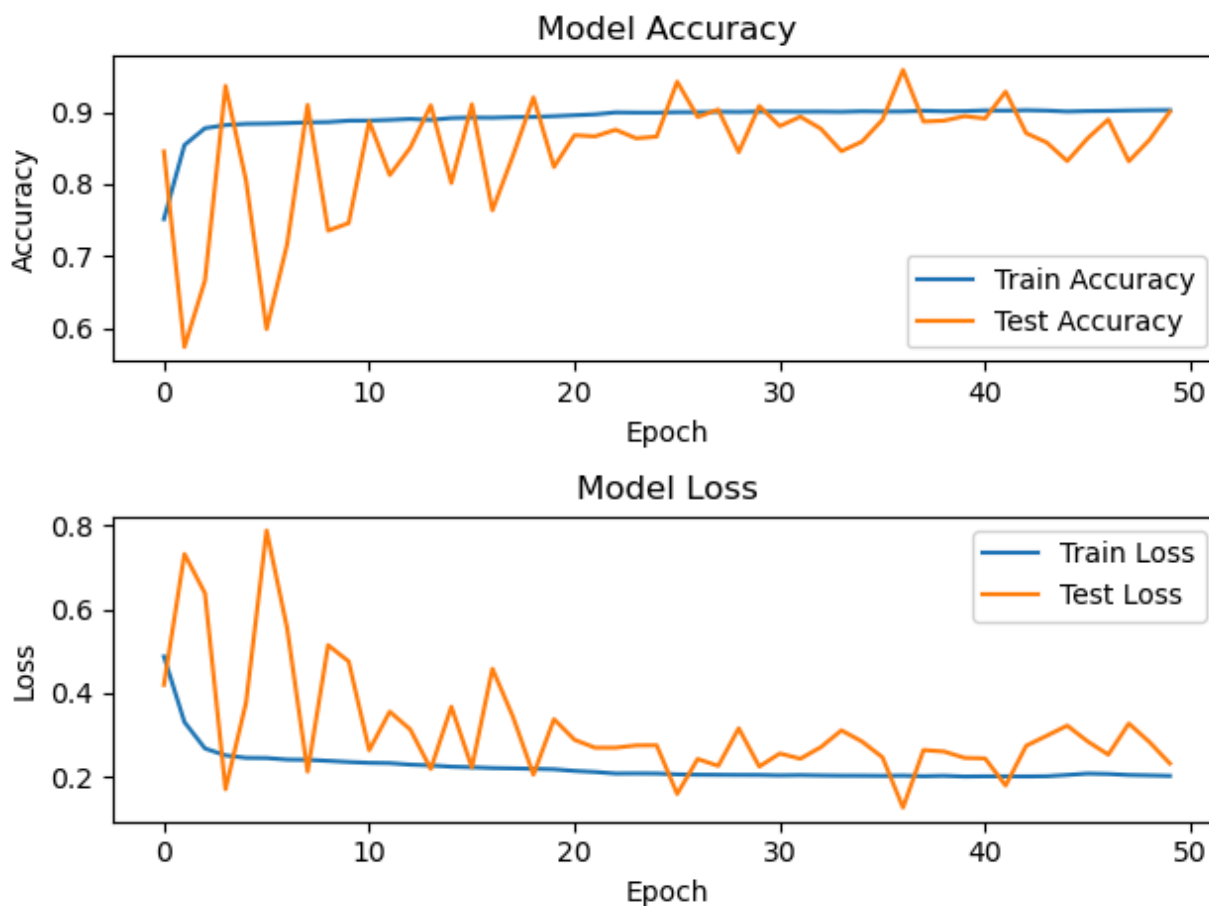
Αρχικά, η αρχιτεκτονική του μοντέλου αποτελείται από ένα επίπεδο εισόδου, δύο κρυφά επίπεδα: ένα 32 νευρωνίων με συνάρτηση ενεργοποίησης ReLu και ένα 16 νευρωνίων επίσης με συνάρτηση ενεργοποίησης ReLu. Τέλος έχει ένα επίπεδο εξόδου με ένα νευρώνιο αποτελέσματος με συνάρτηση ενεργοποίησης sigmoid. Ο ρυθμός εκμάθησης είναι 0.001 και batch size = 32.



Εικόνα 3.2: Αποτελέσματα ευστοχίας MLP μοντέλου με 10 εποχές.

Η τελική απόδοση του μοντέλου είναι 86.45% η οποία ενδεχομένως να είναι αποδεκτή για κάποιες εφαρμογές. Παρόλα αυτά από το παραπάνω γράφημα φαίνεται ότι το ποσοστό ευστοχίας είναι σε ανοδική πορεία και παράλληλα, το ποσοστό σφάλματος σε καθοδική πορεία στο σετ επικύρωσης.

Αυτή η βελτίωση διαρκεί για λίγες εποχές πριν τη λήξη της διαδικασίας εκπαίδευσης, πράγμα που ίσως να σημαίνει ότι το μοντέλο μπορεί να βελτιώσει τα αποτελέσματα του περαιτέρω. Για αυτό τον λόγο παρακάτω διεξάγεται η εκπαίδευση νέου μοντέλου με τη διαφορά ότι οι ο αριθμός εποχών αυξάνεται στις 50.



Εικόνα 3.3: Αποτελέσματα ευστοχίας MLP μοντέλου με 50 εποχές.

Η τελική απόδοση του μοντέλου αυξήθηκε κατά 4.13% και έφτασε τα 90.58%. Επιπλέον από τα παραπάνω γραφήματα παρατηρούμε ότι όντως το μοντέλο είχε ακόμα περιθώριο βελτίωσης αφού η διαφορά τόσο του ποσοστού της ευστοχίας όσο και του σφάλματος μεταξύ εποχών μειώνεται περισσότερο με το πέρας κάθε εποχής μετά της δέκατης.

3.1.2 Εκμάθηση Συνόλου (Ensemble Learning)

Στα σύγχρονα παραδείγματα μηχανικής μάθησης, η τεχνική εκμάθησης συνόλου έχει αναδειχθεί ως βασική στρατηγική για τη βελτίωση της προγνωστικής απόδοσης. Οι μεθοδολογίες συνόλου περιλαμβάνουν τη συγχώνευση πολλαπλών μοντέλων για τη δημιουργία ενός σύνθετου προγνωστικού δείκτη που ξεπερνά κάθε μεμονωμένο μοντέλο [11]. Η συγχώνευση συνήθως γίνεται υπολογίζοντας τον μέσο όρο των αποφάσεων, βεβαρημένων ή μη. Ιδιαίτερα σημαντικές βελτιώσεις εμφανίζονται σε περιπτώσεις όπου το κάθε μεμονωμένο μοντέλο έχει χαμηλό ποσοστό επιτυχίας.

3.1.2.1 Μέθοδος Boosting

Το Boosting είναι ένα εξέχον υποσύνολο των μεθοδολογιών Ensemble Learning, που χαρακτηρίζεται από την ικανότητά του να εκπαιδεύει διαδοχικά μια σειρά σχετικά “αδύναμων” μοντέλων για τη δημιουργία ενός ισχυρότερου μοντέλου πρόβλεψης υψηλής απόδοσης. Είναι μια τεχνική επιβλεπόμενης μάθησης με βασικό σκοπό την ελαχιστοποίηση του bias.

Πιο συγκεκριμένα κάθε ατομικό μοντέλο κατά τη διαδικασία εκπαίδευσης του, για κάθε δείγμα εισόδου υπολογίζει την πρόβλεψη του και σε περίπτωση που είναι λάθος τροποποιεί τα δεδομένα με τα σχετικά τους βάρη. Έτσι κάθε μοντέλο δίνει έμφαση στα προηγούμενα λάθη οπότε όταν δοθεί η συγκεντρωτική απόφαση θα καλύπτει το ένα τα σφάλματα του άλλου.

Ένα πολύ σημαντικό πλεονέκτημα της μεθοδολογίας αυτής είναι ότι έχει καλά αποτελέσματα με σχετικά χαμηλό αριθμό δεδομένων εκπαίδευσης. Παρόλα αυτά είναι ευαίσθητα στον θόρυβο και επιρρεπή σε over-fitting. Ένα από τα πιο γνωστά παραδείγματα Boosting είναι το AdaBoost, από τους Yoav Freund και Robert E. Schapire [12], το οποίο θεωρητικά μπορεί να μειώσει σημαντικά το σφάλμα από κάθε μέθοδο μάθησης η οποία παράγει ταξινομητή με επίδοση λίγο καλύτερη της τυχαίας επιλογής.



Εικόνα 3.4: Βασική αρχιτεκτονική Boosting [13].

3.1.2.2 Διαδικασία Bootstrap

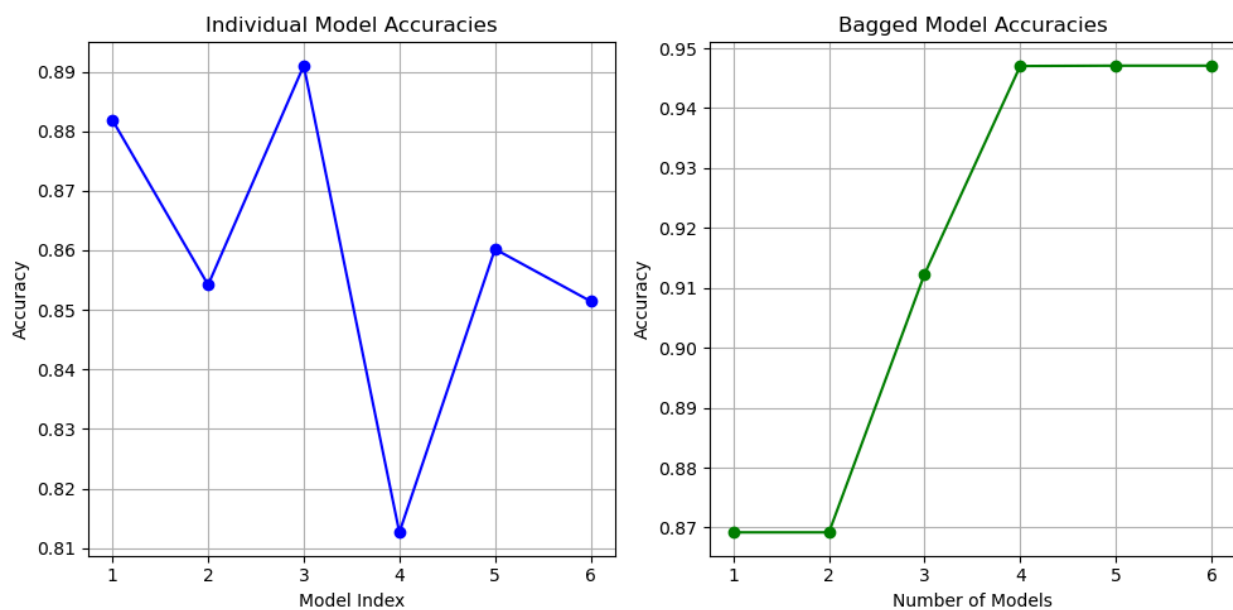
Το bootstrapping είναι μια ισχυρή τεχνική αναδειγματοληψίας που χρησιμοποιείται ευρέως στη μηχανική μάθηση για την αξιολόγηση της σταθερότητας και της αξιοπιστίας των στατιστικών εκτιμήσεων, ιδιαίτερα σε σενάρια όπου τα δεδομένα είναι περιορισμένα. Η μέθοδος περιλαμβάνει τυχαία δειγματοληψία σημείων δεδομένων με αντικατάσταση από το αρχικό σύνολο δεδομένων για τη δημιουργία πολλαπλών νέων συνόλων δεδομένων του ίδιου μεγέθους με το αρχικό. Κάθε νέο σύνολο δεδομένων, μπορεί να περιέχει διπλότυπα στιγμιότυπα και να παραλείπει άλλα, μιμούμενο τη μεταβλητότητα και την τυχαιότητα που υπάρχει στα αρχικά δεδομένα. Το bootstrapping είναι ιδιαίτερα χρήσιμο σε καταστάσεις όπου οι παραδοσιακές στατιστικές παραδοχές μπορεί να μην ισχύουν ή όταν η ακριβής εκτίμηση της μεταβλητότητας είναι ζωτικής σημασίας για τη λήψη αποφάσεων σε εργασίες μηχανικής μάθησης.

3.1.2.3 Μέθοδος Bagging

Bagging ονομάζεται η μέθοδος κατά την οποία δημιουργούνται πολλαπλές εκδόσεις ενός μοντέλου οι οποίες χρησιμοποιούνται για τη λήψη ενός συγκεντρωτικού προγνωστικού. Πιο συγκεκριμένα, η τελική πρόβλεψη για κάθε ερέθισμα είναι το αποτέλεσμα της ψηφοφορίας που γίνεται μεταξύ των μοντέλων, όπου η ψήφος τους βασίζεται στην ατομική τους πρόβλεψη.[14] Τα μοντέλα

εκπαιδεύονται σε διαφορετικά, τυχαία επιλεγμένα κομματιατα, των δεδομένων εκπαίδευσης και με αυτό τον τρόπο επιτυγχάνεται έμμεσα μια καλύτερη ειδίκευση των μοντέλων αφού η πλειοψηφία μπορεί να καλύψει τυχόν ελλείψεις ή “τυφλά σημεία” των ατομικών μοντέλων.

Παρακάτω εφαρμόζεται η μέθοδος bagging στο πρώτο μοντέλο της ενότητας 3.1.1. και θα κατασκευαστούν και θα εκπαιδευτούν 6 ίδια μοντέλα. Συνεπώς, η αρχιτεκτονική κάθε μοντέλου θα είναι η εξής: ένα επίπεδο εισόδου, δύο κρυφά επίπεδα : ένα 32 νευρωνίων με συνάρτηση ενεργοποίησης ReLu και ένα 16 νευρωνίων με συνάρτηση ενεργοποίησης επίσης ReLu. Τέλος θα έχουν από ένα επίπεδο εξόδου με ένα νευρώνιο αποτελέσματος με συνάρτηση ενεργοποίησης sigmoid. Ο ρυθμός εκμάθησης είναι 0.001 και batch size = 32 και θα τρέχουν για 10 εποχές.



Εικόνα 3.5: Αποτελέσματα ευστοχίας τεχνικής Bagging με 6 MLP

Στην παραπάνω εικόνα βλέπουμε σημαντική βελτίωση με την τεχνική Bagging. Η απόδοση είναι 94.70% πολύ μεγαλύτερη από οποιοδήποτε μοντέλο ατομικά.

Πίνακας 3.1: Συγκεντρωτικά αποτελέσματα ευστοχίας όλων των τεχνικών MLPs

Έκδοση Μοντελου	MLP με 10 εποχές	MLP με 50 εποχές	Bagging 6 MLPs
Απόδοση	86.45%	90.58%	94.70%

3.2 Δίκτυα Ακτινικής Βάσης (RBF networks)

Τα Δίκτυα Με Ακτινικές Βάσεις (Radial Basis Function Networks, RBFNs) είναι μια κατηγορία τεχνητών νευρωνικών δικτύων, κατάλληλα για αναγνώριση προτύπων και ταξινόμησης λόγω της ικανότητάς τους να μοντελοποιούν σύνθετες σχέσεις. Το παρόν κεφάλαιο παρέχει μια λεπτομερή επισκόπηση των RBF μοντέλων, αναλύοντας την αρχιτεκτονική τους, τον μηχανισμό λειτουργίας τους, τα πλεονεκτήματα και τα μειονεκτήματά τους.

3.2.1 Αρχιτεκτονική των Δικτύων RBF

Ένα Δίκτυο RBF αποτελείται από τρία επίπεδα: το επίπεδο εισόδου, ένα κρυφό επίπεδο και το επίπεδο εξόδου.

- **Επίπεδο Εισόδου:** Αυτό το επίπεδο αποτελείται από κόμβους που μεταφέρουν τα χαρακτηριστικά εισόδου απευθείας στο κρυφό επίπεδο. Κάθε κόμβος σε αυτό το επίπεδο αντιπροσωπεύει μια μεταβλητή εισόδου.
- **Κρυφό Επίπεδο:** Το κρυφό επίπεδο αποτελείται από νευρώνες με συναρτήσεις, ακτινικής βάσης, ενεργοποίησης. Η πιο κοινή είναι η Γκαουσιανή συνάρτηση. Η έξοδος κάθε κρυφού νευρώνα καθορίζεται από την απόσταση μεταξύ του διανύσματος εισόδου και του κέντρου του νευρώνα, η οποία τροποποιείται από μια παράμετρο πλάτους (spread).
- **Επίπεδο Εξόδου:** Το επίπεδο εξόδου είναι ένα γραμμικό επίπεδο όπου κάθε νευρώνας υπολογίζει ένα σταθμισμένο άθροισμα των εξόδων από το κρυφό επίπεδο. Στο παράδειγμα παρακάτω χρησιμοποιήθηκε η σιγμοειδής συνάρτηση για την έκφραση της πρόβλεψης με 0 ή 1.

3.2.2 Εκπαίδευση Μοντέλου RBF

Σε αυτή την υποενότητα θα αναλυθεί η διαδικασία εκπαίδευσης ενός RBF δικτύου, η επιλογή υπερπαραμέτρων και κατανόηση των βασικών χαρακτηριστικών.

Η εκπαίδευση ενός Δικτύου RBF περιλαμβάνει δύο βασικά βήματα:

1. **Καθορισμός των Υπερπαραμέτρων του Κρυφού Επιπέδου:** Αυτό το βήμα περιλαμβάνει την επιλογή των κέντρων και των παραμέτρων πλάτους. Τα κέντρα μπορούν να επιλεγούν χρησιμοποιώντας μεθόδους όπως η κ-μέσοι (k-means) clustering, τυχαία επιλογή από τα δεδομένα εκπαίδευσης, ή άλλους αλγορίθμους συσταδοποίησης. Οι παράμετροι πλάτους μπορούν να οριστούν ομοιόμορφα ή να βελτιστοποιηθούν μέσω διασταυρούμενης επικύρωσης.
2. **Εκπαίδευση των Συντελεστών Εξόδου:** Αφού καθοριστούν τα κέντρα και τα πλάτη, υπολογίζονται οι συντελεστές. Αυτό το βήμα συνήθως περιλαμβάνει την επίλυση ενός προβλήματος γραμμικής παλινδρόμησης, το οποίο μπορεί να γίνει αποδοτικά χρησιμοποιώντας τεχνικές όπως η εκτίμηση ελαχίστων τετραγώνων.

Στην παρακάτω ενότητα φαίνεται η βελτίωση που μπορεί να σημειωθεί με την σωστή επιλογή υπερπαραμέτρων και συγκεκριμένα της συνάρτησης ενεργοποίησης. Και στις δύο περιπτώσεις το μοντέλο έχει παραμείνει ίδιο με μόνη διαφορά την επιλογή της συνάρτησης ενεργοποίησης. Συγκεκριμένα, το μοντέλο έχει την συγκεκριμένη δομή:

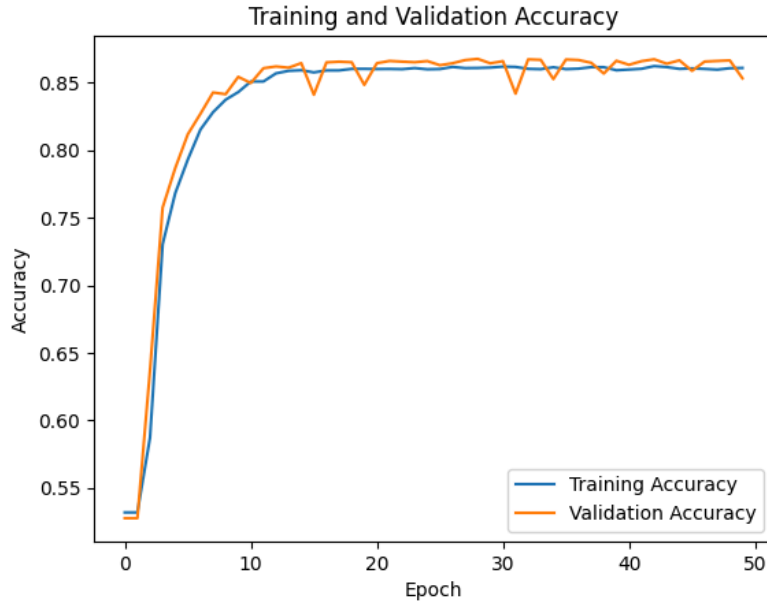
Ένα επίπεδο εισόδου το οποίο δέχεται το διάνυσμα των δεδομένων εκπαίδευσης. Το κρυφό επίπεδο το οποίο απαρτίζεται από 32 κέντρα και καταλήγει στο επίπεδο εξόδου το οποίο εκτελεί μια σιγμοειδή συνάρτηση για ομαλοποίηση της πρόβλεψης σε 0 και 1.

Επιπλέον, το μοντέλο εκπαιδεύεται για 50 εποχές, με batch size 120.

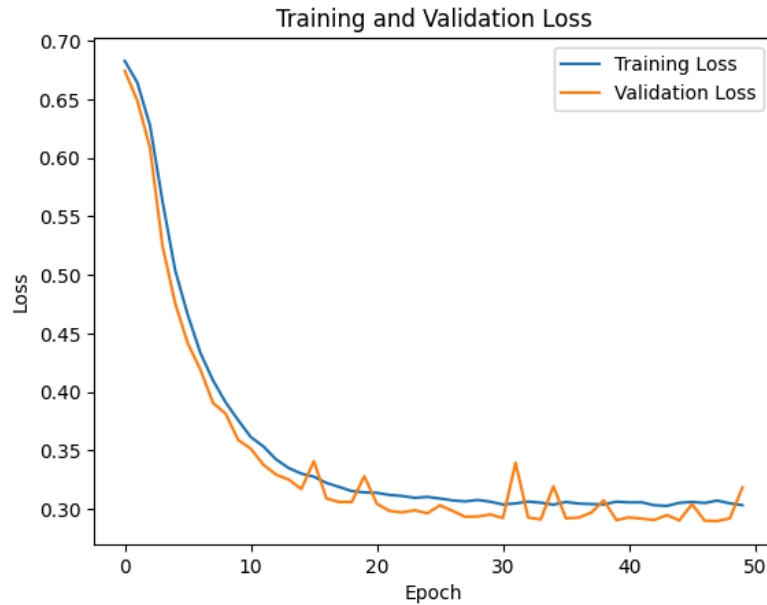
Η αρχική συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε είναι η εξής:

$$\varphi(x) = \exp(-x^2) \quad (3.1),$$

και η ευστοχία που επιτεύχθηκε έφτασε τα 83.25%.



Εικόνα 3.6: Αποτελέσματα ευστοχίας RBF μοντέλου με απλή συνάρτηση ενεργοποίησης.



Εικόνα 3.7: Αποτελέσματα σφάλματος RBF μοντέλου με απλή συνάρτηση ενεργοποίησης.

Στη συνέχεια δοκιμάστηκε μια πιά σύνθετη συνάρτηση ενεργοποίησης η οποία είναι βασισμένη στον παρακάτω μαθηματικό τύπο:

$$\varphi_j(x) = \exp\left(-\frac{1}{2\sigma_j^2}\|x - x_j\|^2\right), j = 1, 2, \dots, N \quad (3.2)$$

όπου $f_j(x)$ είναι η έξοδος του j -οστού κρυφού νευρώνα, x είναι το διάνυσμα εισόδου, x_j είναι το κέντρο του j -οστού νευρώνα, και σ_j είναι η παράμετρος πλάτους.

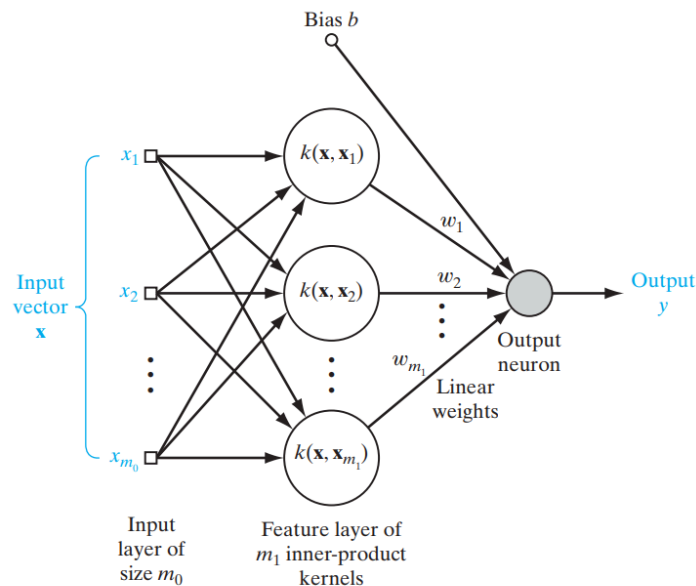
Αυτή η αλλαγή ήταν αρκετή για να σημειωθεί σημαντική βελτίωση της τάξης του 8.21%. Δηλαδή το μοντέλο με την συνάρτηση ενεργοποίησης (3.2) πέτυχε ποσοστό ευστοχίας 91.46%.

3.3 Μηχανές Διανυσμάτων Υποστήριξης (SVMs)

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVMs) είναι ισχυρά μοντέλα εκμάθησης που χρησιμοποιούνται συχνά για εργασίες ταξινόμησης. Παρακάτω, θα αναλυθούν τα βασικά χαρακτηριστικά των μοντέλων SVM, η εκπαίδευση δύο μοντέλων και η σύγκριση της απόδοσης τους.

3.3.1 Ανάλυση Λειτουργίας SVM Μοντέλων

Τα μοντέλα SVM κατασκευάζουν ένα υπερπεδίο το οποίο δρα ως επιφάνεια απόφασης η οποία διαχωρίζει τα δείγματα σε κλάσεις. Η μεθοδολογία αυτή καλύπτει προβλήματα γραμμικά και μη γραμμικά διαχωρίσιμα. Ο διαχωρισμός αυτός επιτυγχάνεται με τη χρήση μεθόδων πυρήνα [15]. Έχουν αναπτυχθεί πολλές μέθοδοι, με διαφορετικές ιδιότητες η κάθε μια και η επιλογή της ιδανικής, για την εφαρμογή, μπορεί να βελτιώσει την απόδοση του μοντέλου σημαντικά. Παρακάτω θα δοκιμαστούν δύο συναρτήσεις, linear και RBF.



Εικόνα 3.8 : Αρχιτεκτονική SVM βασισμένο σε δίκτυο RBF [10].

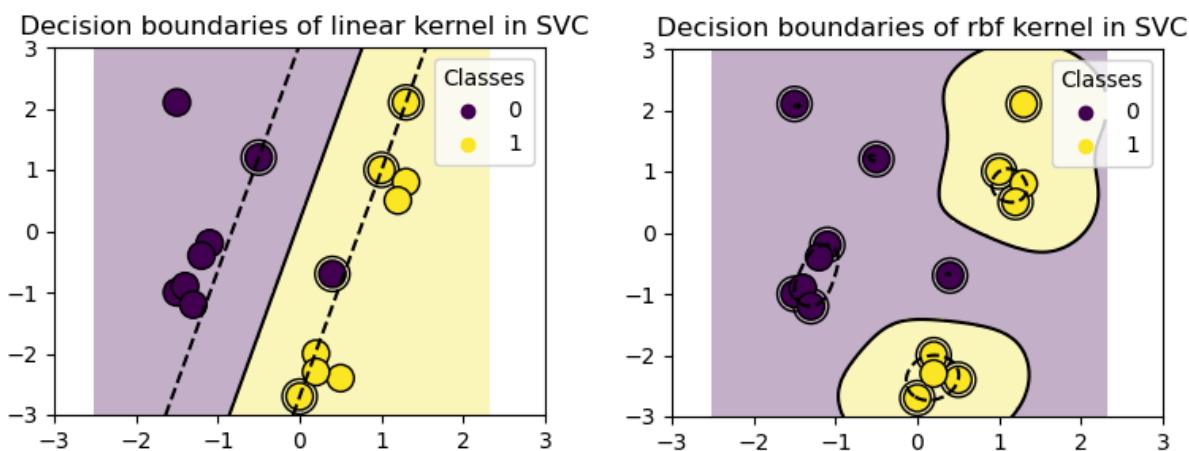
3.3.2 Εκπαίδευση Μοντέλων SVM

Αρχικά να σημειωθεί ότι τα SVM μοντέλα έχουν πολύ καλά αποτελέσματα με μικρό αριθμό δεδομένων από άλλα, αυτό βέβαια επιτυγχάνεται με το κόστος πιο πολύπλοκων υπολογισμών. Για αυτό το λόγο, σαν δεδομένα εκπαίδευσης δόθηκαν δέκα χιλιάδες δείγματα (10%).

Ξεκινώντας με το πρώτο μοντέλο του οποίου ο πυρήνας είναι τύπου linear, ένας απλός σχετικά πυρήνας ο οποίος λειτουργεί βάση του παρακάτω τύπου:

$$K(x_1, x_2) = x_1 \cdot x_2 \text{ όπου } x_2 \cdot x_2 \text{ το εσωτερικό γινόμενο των δύο σημείων.}$$

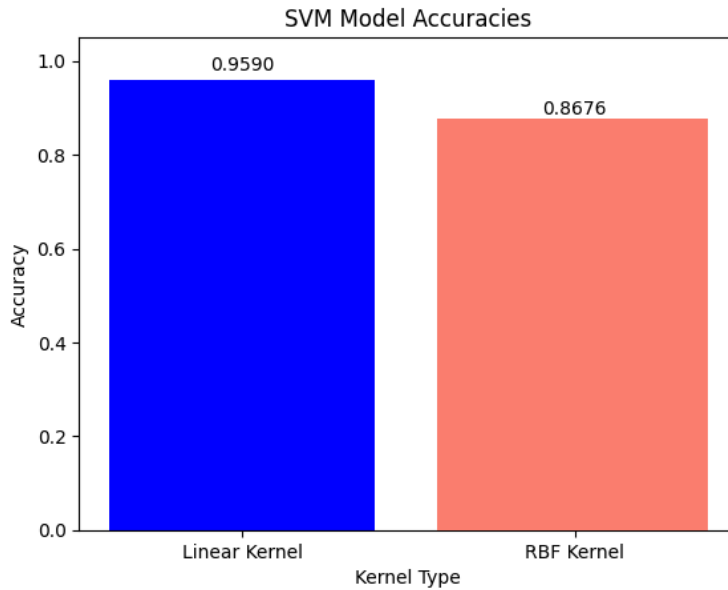
Το δεύτερο μοντέλο έχει έναν RBF πυρήνα με $\gamma = 1/\text{αριθμός_δειγμάτων}$. Ο μαθηματικός τύπος που υπολογίζει είναι: $K(x_1, x_2) = \exp(-\gamma \cdot \|x_1 - x_2\|^2)$, όπου γ (gamma) ελέγχει την επιρροή του κάθε δείγματος στην επιφάνεια/όριο απόφασης και $\|x_1 - x_2\|^2$ η ευκλείδεια απόσταση των δύο σημείων. Από τον παραπάνω τύπο φαίνεται πώς όσο μεγαλύτερη η απόσταση των δύο σημείων τόσο πιο κοντά φτάνει η μέθοδος πυρήνα στο μηδέν και άρα τα σημεία είναι ανόμοια.



Εικόνα 3.9 : Παράδειγμα από επιφάνειες απόφασης SVM μοντέλων με linear(αριστερά) και RBF(δεξιά) πυρήνα [16].

Τα αποτελέσματα της ευστοχίας των μοντέλων φαίνονται στην παρακάτω εικόνα. Το μοντέλο με τον linear πυρήνα είχε ποσοστό ευστοχίας 95.90% ενώ το

μοντέλο με τον πυρήνα RBF 86.76%. Αυτό συμβαίνει γιατί τα περισσότερα ζευγάρια κλάσεων του προβλήματος είναι γραμμικά διαχωρίσιμα μεταξύ τους με αποτέλεσμα ο linear πυρήνας να είναι πολύ αποδοτικός. Για παράδειγμα το ζευγάρι age-heart_disease. Εδώ αξίζει επίσης να σημειωθεί ότι οι κλάσεις HbA1c_level, Blood_glucose_level και BMI χωρίζουν το πρόβλημα γραμμικά αφού αυτοί είναι και δείκτες οι οποίοι χρησιμοποιούνται για την διάγνωση του διαβήτη[17].



Εικόνα 3.10 : Αποτελέσματα ευστοχίας με πυρήνα linear και RBF .

3.4 Επιπλέον Παραδείγματα Εκπαίδευσης Μοντέλων

Σε αυτή την ενότητα θα αναλυθούν αναφορικά δύο ακόμα προσεγγίσεις για την επίλυση του προβλήματος της πρόβλεψης του διαβήτη χρησιμοποιώντας δύο επιπλέον μοντέλα μηχανικής μάθησης που σημείωσαν υψηλό ποσοστό ευστοχίας κατά την ερευνητική διαδικασία.

3.4.1 Δέντρα Απόφασης (Decision Trees)

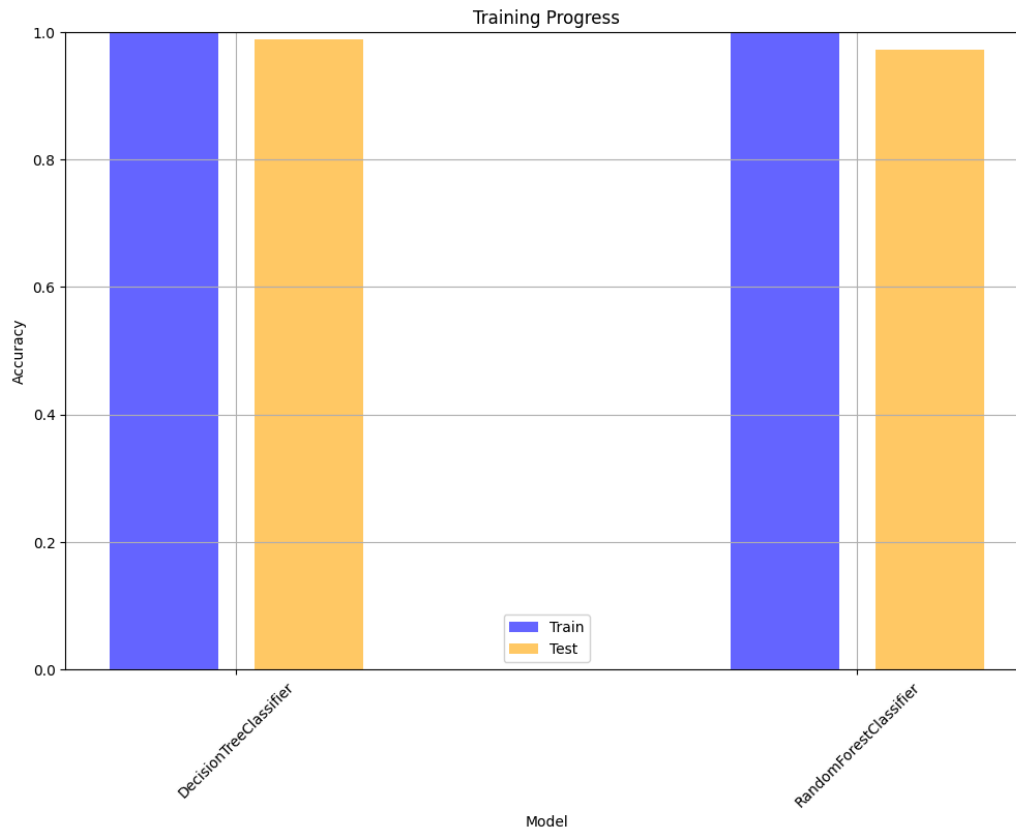
Τα δέντρα απόφασης αποτελούν μια από τις απλούστερες και πιο επιτυχημένες μορφές επιβλεπόμενων αλγορίθμων μάθησης. Δεχονται σαν είσοδο το διάλυμα δεδομένων εκπαίδευσης και επιστρέφουν μια δυαδική (0,1) πρόβλεψη. Εσωτερικά κατασκευάζεται ένα σύμπλεγμα από κόμβους (δέντρο) οι οποίοι αντιπροσωπεύουν αποφάσεις βασισμένες σε χαρακτηριστικά των δεδομένων. Οι τελικοί κόμβοι (φύλλα) αποτελούν τις κλάσεις του προβλήματος.

Ο ταξινομητής που υλοποιήθηκε έφερε ποσοστό ευστοχίας 98.94%.

3.4.2 Ταξινομητής Τυχαίου Δάσους (Random Forest)

Οι ταξινομητές Random Forest υλοποιούν μια μέθοδο εκμάθησης συνόλου που βασίζεται στα δέντρα απόφασης. Στο στάδιο εκπαίδευσης, κατασκευάζουν πολλαπλά δέντρα απόφασης και εξάγουν την συγκεντρωτική τους πρόβλεψη. Αναλυτικότερα, τα δεδομένα εισόδου προκύπτουν από τυχαίες δειγματοληψίες από το σύνολο δεδομένων, στις οποίες επιτρέπονται και οι επιλογές δείγματος περισσότερες από μία φορές (Bootstrapping). Δηλαδή κάποια δείγματα δεν θα εμφανίζονται καθόλου ενώ κάποια ενδεχομένως να εμφανίζονται πολλές φορές. Έπειτα κάθε δειγματοληψία δίνεται σαν είσοδος σε διαφορετικό ταξινομητή δέντρου απόφασης. Ο κάθε ταξινομητής εκπαιδεύεται χωριστά και στο τέλος κάθε δέντρο-ταξινομητής συμμετέχει σε ψηφοφορία για επιλογή της κλάσης. Το αποτέλεσμα αυτής της ψηφοφορίας είναι και η τελική πρόβλεψη του συγκεντρωτικού ταξινομητή.

Για το παράδειγμα αυτό κατασκευάστηκε ταξινομητής μεγέθους 100 εσωτερικών δέντρων και επιτεύχθηκε ποσοστό ακρίβειας 97.29%



Εικόνα 3.11 : Αποτελέσματα ευστοχίας Decision Tree (αριστερά) και Random Forest (δεξιά).

Η μικρή μείωση του ποσοστού ευστοχίας του ταξινομητή Random Forest ίσως να είναι σημάδι over-fitting. Επιπλέον στην περίπτωση που εξετάζουμε υπάρχει μεγάλο πλήθος δεδομένων και το ένα decision tree πετυχαίνει μόνο του πολύ μεγάλη ακρίβεια οπότε είναι μια περίπτωση που ίσως να μην είναι ιδανική η μεθοδολογία Random Forest αφού είναι και πιο υπολογιστικά “ακριβή”.

ΚΕΦΑΛΑΙΟ 4

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτή την διπλωματική εργασία, εξετάστηκαν και αξιολογήθηκαν διάφορες τεχνικές μηχανικής μάθησης για τη διάγνωση του διαβήτη. Τα μοντέλα που χρησιμοποιήθηκαν περιλάμβαναν Πολυστρωματικά Νευρωνικά Δίκτυα (MLPs), Δίκτυα Ακτινικής Βάσης (RBF networks) και Μηχανές Διανυσμάτων Υποστήριξης (SVMs). Επιπλέον, ερευνήθηκαν μέθοδοι εκμάθησης συνόλου όπως το bagging και το boosting για τη βελτίωση της απόδοσης των μοντέλων. Η ανάλυση πραγματοποιήθηκε χρησιμοποιώντας το σύνολο δεδομένων Diabetes Prediction Dataset που περιέχει 100,000 δείγματα.

4.1 Ανασκόπηση και Βασικά Ευρήματα

4.1.1 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων περιλάμβανε κωδικοποίηση των παραμέτρων, εφαρμογή της τεχνικής SMOTE για την αντιμετώπιση της ανισορροπίας στα δεδομένα, και ανίχνευση ανωμαλιών με τη χρήση του ενδοτεταρτημοριακού εύρους (IQR). Αυτές οι διαδικασίες ήταν κρίσιμες για τη βελτίωση της ποιότητας των δεδομένων και την εξασφάλιση της αξιοπιστίας των αποτελεσμάτων.

4.1.2 Εκπαίδευση Μοντέλων

Πολυστρωματικά Νευρωνικά Δίκτυα (MLPs):

Με 10 εποχές εκπαίδευσης, το MLP πέτυχε απόδοση 86.45%.

Με 50 εποχές εκπαίδευσης, η απόδοση αυξήθηκε στο 90.58%.

Η χρήση της μεθόδου bagging με 6 MLPs οδήγησε σε απόδοση 94.70%.

Δίκτυα Ακτινικής Βάσης (RBF networks):

Το απλό RBF μοντέλο είχε απόδοση 83.25%.

Χρησιμοποιώντας πιο προχωρημένη συνάρτηση ενεργοποίησης, η απόδοση αυξήθηκε στο 91.46%.

Μηχανές Διανυσμάτων Υποστήριξης (SVMs):

Το SVM με γραμμικό πυρήνα πέτυχε απόδοση 95.90%.

Το SVM με RBF πυρήνα είχε απόδοση 86.76%.

Άλλα Μοντέλα:

Τα δέντρα απόφασης (Decision Trees) είχαν απόδοση 98.94%.

Ο ταξινομητής τυχαίου δάσους (Random Forest) πέτυχε την υψηλότερη απόδοση, 97.29%.

4.2 Συμπεράσματα και Τελικές Παρατηρήσεις

Η έρευνα αυτή ανέδειξε την αποτελεσματικότητα των διαφορετικών τεχνικών μηχανικής μάθησης στη διάγνωση του διαβήτη. Τα αποτελέσματα δείχνουν ότι τα πιο εξελιγμένα μοντέλα, όπως τα δέντρα απόφασης και οι ταξινομητές τυχαίου δάσους, μπορούν να επιτύχουν πολύ υψηλές αποδόσεις. Τα SVMs επίσης παρουσιάζουν εξαιρετική απόδοση, ειδικά με τη χρήση του γραμμικού πυρήνα, ενισχύοντας την αρχική υπόθεση του γραμμικού διαχωρισμού των κλάσεων του συνόλου δεδομένων.

Επιπλέον, παρατηρήθηκαν τα πλεονεκτήματα που προσφέρουν τεχνικές Bagging στη κατηγορία των πολυστρωματικών νευρωνικών δικτύων, πετυχαίνοντας το υψηλότερο ποσοστό ακρίβειας από τις υπόλοιπες παραλλαγές που δοκιμάστηκαν.

4.3 Μελλοντικές Εργασίες

Για μελλοντικές έρευνες, προτείνεται η διερεύνηση επιπλέον τεχνικών και παραμετροποιήσεων για τα μοντέλα που εξετάστηκαν. Επίσης, η χρήση δεδομένων από διαφορετικές πηγές και η δοκιμή αυτών των μοντέλων σε πραγματικά κλινικά περιβάλλοντα θα μπορούσε να παρέχει περαιτέρω πολύτιμα συμπεράσματα και να βελτιώσει τη γενίκευση των αποτελεσμάτων.

Με την εξέλιξη των τεχνικών μηχανικής μάθησης και την αυξανόμενη διαθεσιμότητα δεδομένων, οι δυνατότητες για βελτίωση της ιατρικής διάγνωσης συνεχίζουν να αυξάνονται. Μια αξιόπιστη και έγκυρη λύση θα μπορούσε να διευκολύνει τη διαδικασία διάγνωσης, μέχρι και σε περιοχές που η παρουσία ιατρικού προσωπικού ίσως να είναι δύσκολη. Ακόμα με τη χρήση επιπλέον δεδομένων και πιο σύνθετων τεχνικών θα μπορούσε να επιτευχθεί η έγκαιρη

πρόβλεψη και πρόγνωση ασθενειών και παθήσεων πριν εμφανιστούν. Δηλαδή η παραγωγή ενός μοντέλου που θα ήταν ικανό να υπολογίζει πιθανότητες εμφάνισης πολλών παθήσεων στο μέλλον για κάθε ασθενή.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] K. Makrilakis *et al.*, “Prevalence of diabetes and pre-diabetes in Greece. Results of the First National Survey of Morbidity and Risk Factors (EMENO) study,” *Diabetes Res. Clin. Pract.*, vol. 172, p. 108646, Feb. 2021.
- [2] M. M. Ahsan, S. A. Luna, and Z. Siddique, “Machine-Learning-Based Disease Diagnosis: A Comprehensive Review,” *Healthc. Pap.*, vol. 10, no. 3, p. 541, Mar. 2022.
- [3] “Diabetes Prediction using Machine Learning Algorithms,” *Procedia Comput. Sci.*, vol. 165, pp. 292–299, Jan. 2019.
- [4] S. Kotsiantis *et al.*, “Data Preprocessing for Supervised Learning,” *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE*, vol. 1, p. 7.
- [5] “Jabbar, H., and Rafiqul Zaman Khan. ‘Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study).’ *Computer Science, Communication and Instrumentation Devices* 70.10.3850 (2015): 978-981”.
- [6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *jair*, vol. 16, pp. 321–357, Jun. 2002.
- [7] H. P. Vinutha, B. Poornima, and B. M. Sagar, “Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset,” *Information and Decision Sciences*, pp. 511–518, 2018.
- [8] CDC, “Testing for Diabetes,” Diabetes. Accessed: May 31, 2024. [Online]. Available: <https://www.cdc.gov/diabetes/diabetes-testing/index.html>
- [9] S. Haykin, *Νευρωνικά Δίκτυα και Μηχανική Μάθηση*. Παπασωτηρίου, 2010, pp. 35–36.
- [10] S. Haykin, “Νευρωνικά Δίκτυα και Μηχανική Μάθηση,” in *To Perceptron Πολλαπλών Επιπέδων*, Παπασωτηρίου, 2010.
- [11] T. G. Dietterich, “Ensemble Methods in Machine Learning,” *Multiple Classifier Systems*, pp. 1–15, 2000.
- [12] “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” *J. Comput. System Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.
- [13] “Website.” [Online]. Available: R. F. Improve, “Boosting in Machine Learning,” GeeksforGeeks. Accessed: May 30, 2024. [Online]. Available: <https://www.geeksforgeeks.org/boosting-in-machine-learning-boosting-and-adaboost/>
- [14] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
- [15] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
- [16] “Plot classification boundaries with different SVM Kernels,” scikit-learn. Accessed: May 16, 2024. [Online]. Available: https://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html
- [17] American Diabetes Association, “Understanding Diabetes Diagnosis,” American Diabetes Association. Accessed: May 22, 2024. [Online]. Available: <https://diabetes.org/about-diabetes/diagnosis>