Πανεπιστήμιο Πειραιώς

Τμήμα Ψηφιακών Συστημάτων

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ

Κατεύθυνση: Μεγάλα Δεδομένα Και Αναλυτική

ΔΙΑΓΝΩΣΗ ΔΙΑΒΗΤΗ ΜΕ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΣΑΛΙΑΡΗΣ ΧΑΡΙΛΑΟΣ

ΑΜ: ΜΕ1822

Μεταπτυχιακή Διπλωματική Εργασία

Επιβλέπων: Ανδριάνα Πρέντζα, Καθηγήτρια Τμήματος Ψηφιακών Συστημάτων

Πειραιάς 2020

Περίληψη

Ο τομέας της υγείας και ειδικότερα ο τομέας της υγειονομικής περίθαλψης έχει από καιρό υιοθετήσει και επωφεληθεί σημαντικά από την τεχνολογική πρόοδο. Τα τελευταία χρόνια, η μηχανική μάθηση (ένα υποσύνολο τεχνητής νοημοσύνης) διαδραματίζει βασικό ρόλο σε πολλές περιοχές που σχετίζονται με την υγεία, συμπεριλαμβανομένης της ανάπτυξης νέων ιατρικών διαδικασιών, του χειρισμού των δεδομένων και των αρχείων των ασθενών, της πρόβλεψης και της θεραπείας των χρόνιων ασθενειών, όπως είναι ο σακχαρώδης διαβήτης.

Η συγκεκριμένη διπλωματική εργασία αφορά τη συγκριτική μελέτη αλγορίθμων μηχανικής μάθησης, καθώς και την αξιολόγηση της επίδοσης των αλγορίθμων για τη διάγνωση του σακχαρώδη διαβήτη μέσω του συνόλου δεδομένων Pima Indians Diabetes Database.

Για κάθε ένα από τα μοντέλα μηχανικής μάθησης παράγονται μετρήσεις οι οποίες εν συνεχεία συγκρίνονται για να διαπιστωθεί ποιο από τα μοντέλα είναι το πιο αποτελεσματικό. Επίσης, πραγματοποιήθηκε αναδρομή και σύγκριση με άλλες έρευνες οι οποίες έχουν γίνει με χρήση του ίδιου συνόλου δεδομένων, καθώς και των ίδιων αλγορίθμων.

Τέλος, με τη βοήθεια του προγράμματος που αναπτύχθηκε, ο κάθε χρήστης δύναται να εισάγει αποδεκτές τιμές στα απαιτούμενα πεδία, ώστε να διαπιστωθεί αν ο εν λόγω χρήστης πάσχει από σακχαρώδη διαβήτη ή όχι.

Summary

The health sector, and in particular the healthcare sector, has long adopted and benefited greatly from technological progress. In recent years, mechanical learning (a subset of artificial intelligence) has played a key role in many health-related areas, including the development of new medical procedures, patient data and records handling, the prevention and treatment of chronic diseases, such as diabetes mellitus.

This diplomatic thesis concerns the comparative study of machine learning algorithms and evaluating the performance of algorithms for the diagnosis of diabetes mellitus through the Pima Indians Diabetes Database.

Measurements are then produced for each of the machine learning algorithms which are then compared to determine which of the models is most effective. Also, a comparison was made with other surveys using the same data set and the same algorithms.

Finally, with the assistance of the program developed, each user can enter acceptable values in the required fields to determine if the user is suffering from diabetes mellitus or not.

Ευχαριστίες

Για την υλοποίηση της διπλωματικής εργασίας θα ήθελα να ευχαριστήσω θερμά την κα. Ανδριάνα Πρέντζα, η οποία είναι καθηγήτρια του τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς, τόσο για την επίβλεψή της εργασίας, όσο και για την πολύτιμη βοήθειά της καθ΄ όλη την διάρκεια της υλοποίησης της εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τον στενά οικογενειακό και φιλικό μου κύκλο για την στήριξη που μου πρόσφεραν σε όλη την εκπαιδευτική μου πορεία μέχρι σήμερα και ελπίζω να συνεχίσουν να το κάνουν και στις επόμενες επιλογές μου.

Πίνακας Περιεχομένων

ΛΙΣΤΑ ΕΙΚΟΝΩΝ

Λίστα Πινάκων

ΚΕΦΑΛΑΙΟ 1 - ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγή

Η μηχανική μάθηση είναι μία μέθοδος ανάλυσης δεδομένων που αυτοματοποιεί την ανάπτυξη αναλυτικού μοντέλου. Είναι ένας κλάδος της τεχνητής νοημοσύνης που βασίζεται στην ιδέα ότι τα συστήματα μπορούν να μάθουν από τα δεδομένα, να προσδιορίσουν τα πρότυπα και να λάβουν αποφάσεις με ελάχιστη ανθρώπινη παρέμβαση.[1]

Ο λόγος που η μηχανική μάθηση τυγχάνει περαιτέρω χρήσης τα τελευταία χρόνια είναι απλός: τεράστιες ποσότητες δεδομένων καταγράφονται καθημερινά και πλέον είναι διαθέσιμες και σε ψηφιακή μορφή. Η επεξεργασία αυτού του όγκου δεδομένων έχει πλέον γίνει προσιτή, καθώς η υπολογιστική δύναμη διατίθεται σε όλο και πιο χαμηλές τιμές. Τα παραπάνω συνδυάζονται με την εξάπλωση δομών (frameworks) ανοιχτού κώδικα, βιβλιοθηκών και εργαλείων που απλοποιούν περαιτέρω την όλη διαδικασία επεξεργασίας των δεδομένων.[2]

Χαρακτηριστική είναι η δήλωση του επιστήμονα υπολογιστών Sebastian Thrum στο περιοδικό “The New Yorker” σε ένα πρόσφατο άρθρο με τίτλο "A.I. Versus M.D.”: «Όπως οι μηχανές κάνουν τους ανθρώπινους μύες χίλιες φορές ισχυρότερους, έτσι και οι μηχανές θα κάνουν τον ανθρώπινο εγκέφαλο χίλιες φορές πιο ισχυρό».[3]

Κάθε μέρα παράγονται μεγάλες ποσότητες δεδομένων σε όλους του κλάδους. Ένας σημαντικός κλάδος με πολλά δεδομένα είναι και ο τομέας της υγείας, και με τη χρήση της μηχανικής μάθησης μπορεί να επιτευχθούν σημαντικά αποτελέσματα όσων αφορά την πρόληψη και την αντιμετώπιση των ασθενειών που προσβάλουν τους ανθρώπους.

Η μηχανική μάθηση συμβάλει στην απλοποίηση των διαχειριστικών υπηρεσιών στα νοσοκομεία, στη χαρτογράφηση, την πρόβλεψη και τη θεραπεία νοσημάτων και στην εξατομίκευση των ιατρικών θεραπειών.

Ειδικότερα στην υγειονομική περίθαλψη, η μηχανική εκμάθηση έχει οδηγήσει σε συναρπαστικές νέες εξελίξεις που θα μπορούσαν να επαναπροσδιορίσουν τη διάγνωση και τη θεραπεία σοβαρών νοσημάτων όπως είναι ο καρκίνος τα επόμενα χρόνια. Επίσης, η μηχανική εκμάθηση μπορεί να αυξήσει την πρόσβαση στις θεραπείες σε αναπτυσσόμενες χώρες που δεν έχουν αρκετούς εξειδικευμένους γιατρούς που μπορούν να θεραπεύσουν ορισμένες ασθένειες, μπορεί να βελτιώσει την ακρίβεια της πρόβλεψης και μπορεί να βοηθήσει στην εξατομίκευση της θεραπείας. Στην πλειοψηφία των περιπτώσεων μπορεί να αυξήσει θεαματικά την αποτελεσματικότητα της ροής εργασίας στα νοσοκομεία. Οι πιθανότητες μοιάζουν ατελείωτες.[2]

1.2 Ορισμός του προβλήματος

Ο Σακχαρώδης Διαβήτης είναι μία από τις μεγαλύτερες παγκόσμιες απειλές υγείας του 21ου αιώνα. Κάθε χρόνο όλο και περισσότεροι άνθρωποι αναγκάζονται να ζήσουν στη νέα πραγματικότητα της νόσου αυτής, η οποία μπορεί να καταλήξει σε επιπλοκές που αλλάζουν τον τρόπο ζωής τους. Υπολογίζεται πως πάνω από 463 εκατομμύρια ενήλικων νοσούν σήμερα από σακχαρώδη διαβήτη παγκοσμίως καταδεικνύοντας αυξημένο κίνδυνο επιπολασμού του σακχαρώδους διαβήτη σε όλο τον κόσμο.[4]

Τελευταία γίνεται προσπάθεια με τη βοήθεια της μηχανικής μάθησης να γίνει πρόβλεψη ασθενειών χρησιμοποιώντας σύνολα δεδομένων. Στη συγκεκριμένη διπλωματική εργασία γίνεται χρήση του συνόλου δεδομένων Pima Indians Diabetes Database, το οποίο προσφέρεται για την πρόβλεψη του σακχαρώδη διαβήτη με χρήση αλγορίθμων, κατόπιν κατάλληλης επεξεργασίας. Οι αλγόριθμοι που θα χρησιμοποιηθούν, θα αναλυθούν και θα συγκριθούν μεταξύ τους και θα επιλεχθεί ο καλύτερος σε αποτελεσματικότητα.

Με βάση το καλύτερο μοντέλο πρόβλεψης σακχαρώδη διαβήτη, δημιουργήθηκε μία εφαρμογή με την οποία δίνεται η δυνατότητα στον χρήστη να εισάγει τα αντίστοιχα χαρακτηριστικά του με σκοπό την πρόβλεψη αν είναι θετικός στην ασθένεια του σακχαρώδη διαβήτη ή αρνητικός.

1.3 Δομή της μεταπτυχιακής διπλωματικής εργασίας

Η διπλωματική εργασία αποτελείται από εφτά (7) κεφάλαια και τις επιμέρους ενότητες του κάθε ένα.

Αρχικά, στην εισαγωγή καθορίζεται το πρόβλημα, παρουσιάζεται η δομή της μεταπτυχιακής διπλωματικής εργασίας και η συνεισφορά της. Εν συνεχεία γίνεται μία συνοπτική αναφορά όσων αφορά το σακχαρώδη διαβήτη, καθώς και τη χρήση των αλγορίθμων μηχανικής μάθησης τόσο στη θεραπεία όσο και στην πρόβλεψη του σακχαρώδη διαβήτη.

Στο δεύτερο κεφάλαιο γίνεται μία λεπτομερής αναφορά στον σακχαρώδη διαβήτη, τη μηχανική μάθηση, καθώς επίσης και σε έρευνες-δημοσιεύσεις που έχουν χρησιμοποιήσει το σύνολο δεδομένων που πραγματεύεται η συγκεκριμένη μεταπτυχιακή εργασία. Πιο αναλυτικά για τον σακχαρώδη διαβήτη αναφέρονται οι τύποι του, τα χαρακτηριστικά συμπτώματα που εμφανίζει και οι τρόποι αντιμετώπισής του. Για τη μηχανική μάθηση γίνεται αναφορά στον ορισμό, τις κατηγορίες στις οποίες διαχωρίζεται, καθώς και τις εφαρμογές της στην σημερινή εποχή. Τέλος, παρουσιάζονται τρεις έρευνες όπου αναλύονται η επεξεργασία των δεδομένων και τα αποτελέσματά τους.

Στο τρίτο κεφάλαιο παρουσιάζονται συνοπτικά οι αλγόριθμοι μηχανικής μάθησης που επιλέχθηκαν, καθώς και η γλώσσα και η πλατφόρμα που χρησιμοποιήθηκαν για την υλοποίηση της διπλωματικής εργασίας. Στη συνέχεια, γίνεται αναλυτική παρουσίαση του συνόλου δεδομένων με τη βοήθεια στατιστικών στοιχείων και γραφικών παραστάσεων. Επίσης, γίνεται η κατάλληλη προ-επεξεργασία των δεδομένων με βάση τόσο των συμπερασμάτων από την παραπάνω διαδικασία, όσο και με την ιατρική τεχνογνωσία.

Στο τέταρτο κεφάλαιο, αναλύονται και εφαρμόζονται οι πέντε αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται στην υλοποίηση της εργασίας. Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται είναι οι εξής: λογιστική παλινδρόμηση (Logistic Regression), δέντρα αποφάσεων (Decision Tree

Classifier), τυχαία δάση αποφάσεων (Random forest), μηχανές διανυσμάτων υποστήριξης (Support Vector Machines, SVM) και ο αλγόριθμος K-Πλησιέστερων Γειτόνων (K-Nearest Neighbors, KNN).

Στο πέμπτο κεφάλαιο, γίνεται παρουσίαση των αποτελεσμάτων και εύρεση του καλύτερου αλγορίθμου για την πρόβλεψη του σακχαρώδη διαβήτη. Γίνεται επίσης και σύγκριση με παρόμοιες έρευνες που χρησιμοποιούν το ίδιο σύνολο δεδομένων καθώς και τους ίδιους αλγόριθμους.

Στο έκτο κεφάλαιο, παρουσιάζεται λεπτομερής περιγραφή της εφαρμογής που έχει υλοποιηθεί, κατά την οποία μπορεί ο χρήστης να εισάγει τα χαρακτηριστικά του και να γίνει πρόβλεψη για το αν η διάγνωση στον σακχαρώδη διαβήτη είναι θετική ή αρνητική.

Στο έβδομο και τελευταίο κεφάλαιο, παρουσιάζονται τα συμπεράσματα των αποτελεσμάτων σε σύγκριση με τις έρευνες που αναφέρθηκαν στο κεφάλαιο 2. Επίσης γίνεται αξιολόγηση της εφαρμογής που υλοποιήθηκε και τέλος αναφέρονται βελτιωτικές αλλαγές όσων αφορά το σύνολο δεδομένων, την εφαρμογή των αλγορίθμων και την εφαρμογή.

1.4 Συνεισφορά της μεταπτυχιακής διπλωματικής εργασίας

Η χρήση της μηχανικής μάθησης στον τομέα της υγείας διαπιστώνουμε ότι είναι κάτι παραπάνω από αναγκαία. Οι επιστήμονες θα είναι σε θέση να προβούν σε καλύτερες μεθόδους θεραπείας των ασθενών, καθώς και στον πρόωρο εντοπισμό τους, συμβάλλοντας στην καλύτερη αντιμετώπισή τους.

Η συγκεκριμένη μεταπτυχιακή διπλωματική εργασία συνεισφέρει στον εντοπισμό του καλύτερου αλγορίθμου μηχανικής μάθησης για την πρόβλεψη του σακχαρώδη διαβήτη.

Τέλος, με την υλοποίηση της εφαρμογής μπορεί εύκολα και γρήγορα κάποιος χρήστης να διαπιστώσει αν πιθανότατα πάσχει από σακχαρώδη διαβήτη. Φυσικά, η εφαρμογή αυτή είναι μία πρώτη εκτίμηση πιθανότητας να πάσχει

κάποιος από σακχαρώδη διαβήτη, ωστόσο για τη διάγνωση της νόσου απαιτείται επίσκεψη σε ιατρό, καθώς και οι απαιτούμενες εξετάσεις στις οποίες θα υποβληθεί ο ασθενής.

ΚΕΦΑΛΑΙΟ 2 - ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ

Σε αυτό το κεφάλαιο αρχικά γίνεται λεπτομερής αναφορά στον σακχαρώδη διαβήτη και ειδικότερα στους τύπους διαβήτη που υπάρχουν, τα χαρακτηριστικά συμπτώματα που εκδηλώνονται, καθώς επίσης και στους τρόπους αντιμετώπισης της ασθένειας. Εν συνεχεία, γίνεται αναφορά στη μηχανική μάθηση και πιο συγκεκριμένα στις διάφορες κατηγορίες και στη χρήση της. Τέλος, παρουσιάζονται και αναλύονται δημοσιεύσεις και έρευνες στις οποίες έχει γίνει χρήση του ίδιου συνόλου δεδομένων όπως και στην εν λόγω μεταπτυχιακή εργασία.

2.1 Σακχαρώδης διαβήτης

2.1.1 Ορισμός

Ο Σακχαρώδης Διαβήτης είναι ένα σύνδρομο με ετερογενές και πολυπαραγοντικό υπόστρωμα. Αποτελεί μια χρόνια νόσο του παγκρέατος που χαρακτηρίζεται από διαταραχή του μεταβολισμού των υδατανθράκων με αποτέλεσμα την αύξηση της γλυκόζης στο αίμα. Η διαταραχή αυτή αφορά και άλλα βασικά συστατικά των τροφών, όπως τα λίπη και οι πρωτεΐνες και οφείλεται κυρίως στην ελλιπή παραγωγή της ινσουλίνης. Η ανεπάρκεια αυτή αφορά είτε μειωμένη παραγωγή της ινσουλίνης ή μη επαρκή κάλυψη των αυξημένων αναγκών του μεταβολισμού. Η κύρια έκφραση της διαταραχής του μεταβολισμού στον σακχαρώδη διαβήτη είναι η αύξηση των επιπέδων γλυκόζης στο αίμα.[5]

2.1.2 Ταξινόμηση

Η ταξινόμηση γίνεται ανάλογα με τη διαταραχή της ινσουλίνης στις ακόλουθες κατηγορίες, ως εξής [5]:

1. Σακχαρώδης διαβήτης τύπου 1 ή νεανικός ή ινσουλινο-εξαρτώμενος: εμφανίζεται συνήθως σε νεαρά άτομα ηλικίας κάτω των

40 ετών και ουσιαστικά οφείλεται στην καταστροφή των β-κυττάρων του παγκρέατος που παράγουν την ινσουλίνη. Τα άτομα που πάσχουν από τη νόσο αυτή θα πρέπει να χρησιμοποιούν ενέσιμη ινσουλίνη για ελέγχουν τα επίπεδα της γλυκόζης στο αίμα τους.

2. Σακχαρώδης διαβήτης τύπου 2: είναι ο συχνότερος τύπος και χαρακτηρίζεται από τη συνύπαρξη διαταραχής της έκκρισης και της δράσης της ινσουλίνης. Ουσιαστικά η ινσουλίνη που παράγει το πάγκρεας είτε δεν επαρκεί ή δεν λειτουργεί σωστά ή και τα δύο. Παλιότερα ονομάζονταν σακχαρώδης διαβήτης των ενηλίκων μιας και εμφανίζεται σε άτομα μεγαλύτερης ηλικίας και μη ινσουλινο-εξαρτώμενος σακχαρώδης διαβήτης, ωστόσο σε ορισμένες περιπτώσεις μπορεί να χρειαστεί χορήγηση ενέσιμης ινσουλίνης.

3. Διαβήτης της κύησης: ορίζεται ως διαταραχή του μεταβολισμού των υδατανθράκων ποικίλης βαρύτητας με έναρξη ή πρώτη αναγνώριση στην παρούσα εγκυμοσύνη ή ακόμα και κατά την διάρκεια της εγκυμοσύνης.

4. Άλλοι ειδικοί τύποι διαβήτη:

    α) Σακχαρώδης διαβήτης που προκαλείται από γενετικές διαταραχές που αφορούν την δράση της ινσουλίνης.

    β) Σακχαρώδης διαβήτης που προκαλείται από ενδοκρινοπάθειες, λοιμώξεις, νόσους του εξωκρινούς παγκρέατος ή άλλες νόσους.

    γ) Σακχαρώδης διαβήτης που προκαλείται από φάρμακα ή χημικές ουσίες.

    δ) Σακχαρώδης διαβήτης που προκαλείται από γενετικά σύνδρομα.

    ε) Σακχαρώδης διαβήτης που προκαλείται από γενετικές διαταραχές της λειτουργίας των β-κυττάρων (έκκριση ινσουλίνης).

2.1.3 Σακχαρώδης διαβήτης τύπου 2

Το συγκεκριμένο σύνολο δεδομένων που χρησιμοποιείται για την υλοποίηση της μεταπτυχιακής διπλωματικής εργασίας αφορά αποκλειστικά τον σακχαρώδη διαβήτη τύπου 2.

Ο σακχαρώδης διαβήτης τύπου 2 οφείλεται σε μία μείωση των υποδοχέων της ινσουλίνης στην περιφέρεια (λιπώδης και μυϊκός ιστός) με αρχική αύξηση της έκκρισης ινσουλίνης για την αντιμετώπιση της κατάστασης που επιφέρει εκφύλιση των β-κυττάρων του παγκρέατος και αυξημένη απόπτωση των β-κυττάρων με αποτέλεσμα τη μείωση των επιπέδων ινσουλίνης στο αίμα. Ο σακχαρώδης διαβήτης τύπου 2 αφορά πάνω από το 90% των ατόμων με σακχαρώδη διαβήτη.

Υπολογίζεται πως 463 εκατομμύρια άνθρωποι πάσχουν από σακχαρώδη διαβήτη παγκοσμίως, με το 90-95% να αφορά σε σακχαρώδη διαβήτη τύπου 2. Το ολοένα και αυξανόμενο ποσοστό εμφάνισης της νόσου οφείλεται σε πολλούς παράγοντες της καθημερινότητας όπως η καθιστική ζωή, οι διατροφικές συνήθειες που αλλάζουν, αλλά και στον τρόπο με τον οποίο ζουν οι άνθρωποι σήμερα.[5]

Ο σακχαρώδης διαβήτης τύπου 2 αναπτύσσεται συνήθως μετά τα 40, αλλά έχει αναφερθεί αύξηση των περιστατικών του σακχαρώδη διαβήτη τύπου 2 στα παιδιά, ακόμη και από την ηλικία των 4 ετών, καθώς και στους εφήβους. Παρότι οι ορμόνες στην εφηβεία συνδέονται με μικρές αυξήσεις της ινσουλινοαντοχής, οι κυριότεροι παράγοντες που συμβάλλουν στην αυξημένη επίπτωση διαβήτη στους νέους είναι η παχυσαρκία και ο καθιστικός τρόπος ζωής.

2.1.4 Διάγνωση του σακχαρώδους διαβήτη

Τα κριτήρια με τα οποία εξετάζεται αν κάποιο άτομο έχει σακχαρώδη διαβήτη είναι τα εξής [4-5]:

i) Γλυκόζη πλάσματος νηστείας ≥ 126 mg/dL. Για τη συγκριμένη εξέταση, απαιτείται η μη λήψη τροφής από το άτομο τουλάχιστον για 8 ώρες. Η εξέταση γίνεται το πρωί μετά από την έγερση του ατόμου.

ii) Γλυκόζη πλάσματος 2 ωρών ≥ 200 mg/dL. Κατά τη συγκεκριμένη διαδικασία, το άτομο λαμβάνει γλυκόζη 75 γραμμαρίων διαλυμένο σε νερό. Η γλυκόζη πλάσματος πρέπει να είναι ≥ 200 mg/dL.

iii) Τυχαία μέτρηση γλυκόζης πλάσματος ≥ 200 mg/dL σε ασθενή με τυπικά συμπτώματα υπεργλυκαιμίας. Ως συμπτώματα υπεργλυκαιμίας θεωρούνται η πολυουρία, η πολυδιψία, η πολυφαγία και η ανεξήγητη απώλεια βάρους.

Ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ) και η Αμερικανική Διαβητολογική Εταιρεία (ADA) συνιστούν να τίθεται η διάγνωση του σακχαρώδη διαβήτη με βάση τα αποτελέσματα σε τουλάχιστον δύο δοκιμασίες ελέγχου γλυκόζης πλάσματος / αίματος.

Υπάρχουν και συμπληρωματικές εξετάσεις που χρησιμοποιούνται για την παρακολούθηση του σακχαρώδη διαβήτη. Οι εξετάσεις αυτές λειτουργούν συμπληρωματικά ως προς τον έλεγχο της γλυκόζης αίματος, παρέχοντας εναλλακτικές πληροφορίες σχετικά με το γλυκαιμικό έλεγχο. Τέτοιες εξετάσεις είναι η ανάλυση ούρων, η μέτρηση κετονών και το λιπιδαιμικό προφίλ.

Η εξέταση της ανάλυσης ούρων αφορά την ανάλυση της χημικής σύστασης των ούρων που περιλαμβάνει και τον έλεγχο για την παρουσία γλυκόζης.

Λιπιδαιμικό προφίλ: Ο σακχαρώδης διαβήτης τύπου 2 σχετίζεται με τον κίνδυνο εμφάνισης στεφανιαίας νόσου και είναι 2 έως 4 φορές μεγαλύτερος από αυτόν των ατόμων που δεν έχουν σακχαρώδη διαβήτη. Αυτός ο αυξημένος κίνδυνος οφείλεται πιθανώς στα μη φυσιολογικά επίπεδα λιπιδίων (δυσλιπιδαιμία). Τα ελεύθερα λιπαρά οξέα, τα τριγλυκερίδια και συνήθως η χοληστερόλη ανευρίσκονται συχνά σε αυξημένες συγκεντρώσεις στο αίμα.

2.1.5 Τρόποι αντιμετώπισης του σακχαρώδους διαβήτη

Δεν υπάρχει τυποποιημένη προσέγγιση στη θεραπεία του σακχαρώδους διαβήτη, αλλά κατά κανόνα, οι ασθενείς με σακχαρώδη διαβήτη τύπου 2 καθοδηγούνται αρχικά να ακολουθήσουν πρόγραμμα δίαιτας και σωματικής άσκησης. Η δίαιτα και η άσκηση είναι ζωτικής σημασίας για τον καλό έλεγχο του σακχάρου, ανεξάρτητα από τον τύπο φαρμακοθεραπείας που θα χρησιμοποιηθεί τελικά. Επειδή μόνο με δίαιτα και άσκηση συνήθως δεν επιτυγχάνεται ο γλυκαιμικός στόχος του ασθενούς, ξεκινά η χορήγηση από το στόμα κάποιου αντιδιαβητικού δισκίου. Με τον καιρό, μπορεί να απαιτηθεί συνδυασμός φαρμάκων από το στόμα αλλά και έναρξη ενέσιμης αγωγής.

Για την αντιμετώπιση του σακχαρώδους διαβήτη συνίσταται [4-5]:

i) Δίαιτα: Τα άτομα με σακχαρώδη διαβήτη πρέπει να διαμορφώσουν ένα πρόγραμμα γευμάτων σε συνεννόηση με το γιατρό τους. Το πρόγραμμα αυτό προσαρμόζεται στις ανάγκες και τον τρόπο ζωής του κάθε ατόμου. Το πρόγραμμα γευμάτων περιλαμβάνει τρεις συνιστώσες: την ποσότητα (υπολογισμός της πρόσληψης θερμίδων και της πρόσληψης υδατανθράκων), τους τύπους των τροφών και το χρόνο κατανάλωσης των γευμάτων.

ii) Σωματική άσκηση: Η μέτρια, συστηματική άσκηση βοηθά σημαντικά στον έλεγχο των επιπέδων της γλυκόζης και στη διατήρηση της καρδιαγγειακής υγείας. Η άσκηση βελτιώνει την ευαισθησία των ιστών στην ινσουλίνη και μειώνει τα επίπεδα ινσουλίνης σε κατάσταση νηστείας και μεταγευματικά. Η προσαρμογή του τρόπου ζωής μέσω της άσκησης μπορεί να είναι εξαιρετικά σημαντική για τη βελτίωση του γλυκαιμικού ελέγχου στα άτομα με σακχαρώδη διαβήτη τύπου 2. Η άσκηση ελαττώνει επίσης τον κίνδυνο καρδιαγγειακής νόσου βελτιώνοντας το λιπιδαιμικό προφίλ και βελτιώνοντας τα επίπεδα αρτηριακής πίεσης στα διαβητικά άτομα. Όταν συνδυαστεί με τροποποιήσεις στη δίαιτα, η άσκηση μπορεί να μειώσει αποτελεσματικά το σωματικό λίπος. Αυτός είναι συνήθως ένας σημαντικός παράγοντας

για την επίτευξη του γλυκαιμικού ελέγχου στα άτομα με σακχαρώδη διαβήτη τύπου 2.

iii) Φαρμακευτική αγωγή: Η φαρμακευτική αγωγή για το σακχαρώδη διαβήτη μπορεί να ξεκινήσει με διάφορους τρόπους και θα πρέπει να εξατομικεύεται με βάση τις ανάγκες του ασθενούς. Τα αντιδιαβητικά δισκία χρησιμοποιούνται για τη θεραπεία μόνο του σακχαρώδη διαβήτη τύπου 2. Συνταγογραφούνται συνήθως όταν με το πρόγραμμα διαιτητικής αγωγής και άσκησης δεν επιτυγχάνεται ο έλεγχος των επιπέδων γλυκόζης. Ωστόσο, καθώς ο σακχαρώδης διαβήτης τύπου 2 διαγιγνώσκεται συνήθως χρόνια μετά την έναρξή του και η υπεργλυκαιμία είναι σημαντική, ορισμένοι ασθενείς ξεκινούν αμέσως με αντιδιαβητική αγωγή από το στόμα, ή και με ινσουλίνη. Τα τελευταία χρόνια, αυξήθηκε σημαντικά ο αριθμός και η ποικιλία των αντιδιαβητικών φαρμάκων από του στόματος για τη θεραπεία του σακχαρώδη διαβήτη τύπου 2. Σήμερα υπάρχουν αρκετές κατηγορίες τέτοιων φαρμάκων που συνιστώνται στη θεραπεία του σακχαρώδη διαβήτη τύπου 2. Κάθε κατηγορία έχει διαφορετικό μηχανισμό δράσης, ενώ μπορούν να χρησιμοποιηθούν ταυτόχρονα περισσότεροι τύποι. Ο θεράπων ιατρός είναι εκείνος που θα αποφασίσει ποια κατηγορία φαρμάκων ή συνδυασμός είναι ο καταλληλότερος για τον κάθε ασθενή με βάση τον τρόπο ζωής του, τις ανάγκες του και τις τιμές που καταγράφονται στις μετρήσεις του.

2.2 Μηχανική μάθηση

2.2.1 Ορισμός

Η Μηχανική Μάθηση (Machine Learning) είναι ένας τομέας της επιστήμης των υπολογιστών άρρηκτα συνδεδεμένος με την τεχνητή νοημοσύνη (Artificial Intelligence) που έχει ως στόχο την κατασκευή και χρήση αλγορίθμων για τη μετατροπή εμπειρικών δεδομένων σε μοντέλα που μπορούν εν συνεχεία να χρησιμοποιηθούν στο μέλλον. Το κύριο χαρακτηριστικό της μηχανικής μάθησης

είναι η ικανότητα του συστήματος, μέσω της χρήσης των αλγορίθμων, να αποκτά και να ενσωματώνει γνώση από τα σύνολα δεδομένων και να βελτιώνεται και να εκπαιδεύεται με αποτέλεσμα να μπορεί να προβεί σε σχετικές προβλέψεις.[6,7]

Αν και το πεδίο της μηχανικής μάθησης είναι πολύ ευρύ και ολοένα εξελισσόμενο και ως εκ τούτου είναι πολύ δύσκολο να το ορίσει κανείς, αρχικά ο Arthur Samuel το 1959, πραγματοποίησε μία αναφερόμενη δήλωση για τη μηχανική μάθηση δηλώνοντας: «Η μηχανική μάθηση είναι το πεδίο σπουδών που δίνει στους υπολογιστές τη δυνατότητα να μάθουν χωρίς να προγραμματίζονται ρητά».[7]

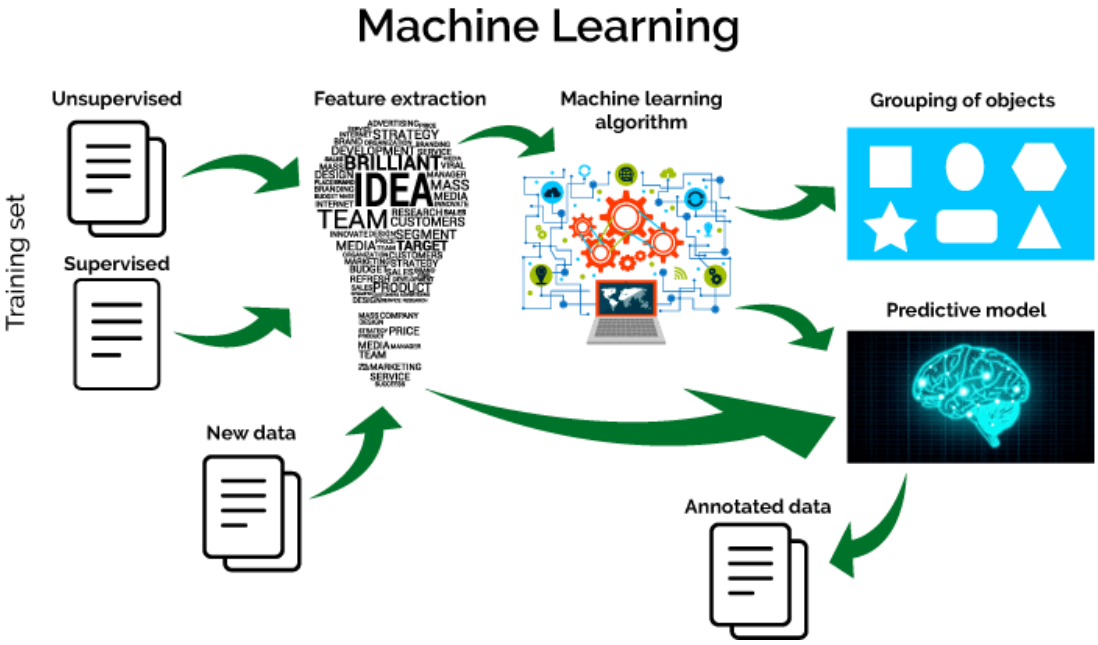
Ένας άλλος σχετικός γενικός ορισμός μηχανικής μάθησης δίνεται από τον Mitchell το 1997: «Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία Ε ως προς κάποια κλάση εργασιών Τ και μέτρο απόδοσης Ρ, αν η απόδοσή του σε εργασίες από το Τ, όπως μετριέται από το Ρ, βελτιώνεται μέσω της εμπειρίας Ε».[7]

Οι επιστήμονες του χώρου της πληροφορικής και της Τεχνητής Νοημοσύνης (ΤΝ) ορίζουν τη λεγόμενη μηχανική μάθηση ως *το φαινόμενο κατά το οποίο ένα σύστημα βελτιώνει την απόδοσή του κατά την εκτέλεση μίας συγκεκριμένης εργασίας, χωρίς να υπάρχει ανάγκη να προγραμματιστεί εκ νέου.* Βάσει του ορισμού αυτού, η μηχανική μάθηση αποσκοπεί στη δημιουργία μηχανών ικανών να μαθαίνουν, να βελτιώνουν την απόδοσή τους σε κάποιους τομείς μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας.

Η μηχανική μάθηση είναι στενά συνδεδεμένη με την υπολογιστική στατιστική και παράλληλα έχει ισχυρούς δεσμούς με τη μαθηματική βελτιστοποίηση, η οποία της παρέχει μεθόδους, θεωρία και τομείς εφαρμογής.

Η μηχανική μάθηση μπορεί να χρησιμοποιηθεί στις υπολογιστικές εφαρμογές, στις οποίες ο σχεδιασμός και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Ωστόσο, διαφέρει από την εξόρυξη δεδομένων, η οποία επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση των δεδομένων (μη επιβλεπόμενη μάθηση).

Η μηχανική μάθηση συμβάλλει σημαντικά και στο πεδίο της ανάλυσης δεδομένων, καθώς είναι μία μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη (Εικόνα 1). Τα αναλυτικά μοντέλα επιτρέπουν στους ερευνητές, τους επιστήμονες δεδομένων, τους μηχανικούς και τους αναλυτές να παράγουν αξιόπιστες αποφάσεις και αποτελέσματα και να αναδείξουν αλληλοσυσχετίσεις μέσω της μάθησης από ιστορικές σχέσεις και τάσεις στα δεδομένα.



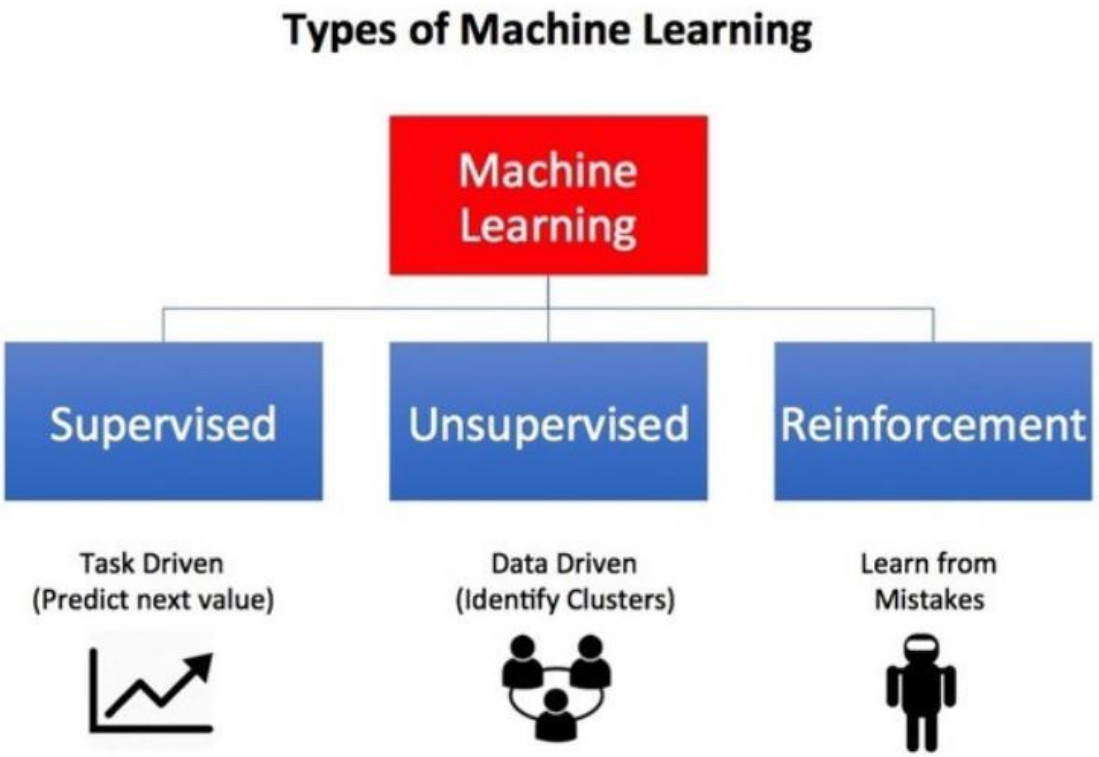
Εικόνα 1: Πλαίσιο μηχανικής μάθησης [9]

2.2.2 Κατηγοριοποίηση μηχανικής μάθησης

Τα τελευταία χρόνια έχουν αναπτυχθεί πολλές τεχνικές μηχανικής μάθησης, οι οποίες χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και εμπίπτουν σε ένα από τα τρία παρακάτω είδη (Εικόνα 2):

I. Επιβλεπόμενη μάθηση (Supervised Learning)
II. Μη Επιβλεπόμενη μάθηση (Unsupervised Learning)
III. Ενισχυτική μάθηση (Reinforcement Learning)

Η κύρια διαφορά μεταξύ των δύο τύπων της επιβλεπόμενης και μη επιβλεπόμενης μηχανικής μάθησης είναι ότι στην εποπτευόμενη μάθηση το σύστημα που σχεδιάζεται και προγραμματίζεται εκπαιδεύεται χρησιμοποιώντας σύνολα δεδομένων των οποίων γνωρίζουμε εκ των προτέρων τις τιμές εξόδων τους. Στόχος της εποπτευόμενης μάθησης είναι να δημιουργήσει μία συνάρτηση με την οποία να προσεγγίσει καλύτερα τη σχέση μεταξύ εισόδου και εξόδου που παρατηρείται στο σύνολο δεδομένων που της δίνεται. Αντιθέτως η μη επιβλεπόμενη μάθηση δεν λαμβάνει πληροφορίες σχετικά με τις τιμές εξόδων και στόχος της είναι να ανακαλύπτει συσχετίσεις ή ομάδες χρησιμοποιώντας τη φυσική δομή που υπάρχει στο σύνολο δεδομένων που της δίνεται.[10]



Εικόνα 2: Κατηγορίες μηχανικής μάθησης [11]

2.2.2.1 Επιβλεπόμενη μάθηση

Η επιβλεπόμενη μάθηση είναι το πιο δημοφιλές παράδειγμα της μηχανικής μάθησης και είναι το πιο εύκολο να κατανοηθεί και το πιο απλό στην εφαρμογή.

Ουσιαστικά ο αλγόριθμος μηχανικής μάθησης εκπαιδεύεται ώστε να κατασκευάσει μία συνάρτηση η οποία αντιστοιχίζει τις εισόδους ενός συνόλου δεδομένων με τις επιθυμητές εξόδους. Εν συνεχεία μπορεί να προβεί σε πρόβλεψη τιμών λαμβάνοντας νέες εισόδους για τις οποίες δεν είναι γνωστές οι τιμές εξόδων τους. Για να επιτευχθεί αυτό οι νέες είσοδοι θα ταξινομηθούν βασισμένες σε προηγούμενα δεδομένα εκπαίδευσης.[12]

Η επιβλεπόμενη μάθηση περιγράφεται συχνά ως προσανατολισμένη στην εργασία (task-driven), τροφοδοτώντας όλο και περισσότερα παραδείγματα στον αλγόριθμο έως ότου μπορεί να εκτελέσει με ακρίβεια αυτήν την εργασία.[11]

Η επιβλεπόμενη μάθηση χρησιμοποιείται συνήθως στο πλαίσιο της ταξινόμησης και της παλινδρόμησης. Η ταξινόμηση (classification) αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων, όπως για παράδειγμα αν κάποιος είναι θετικός στον σακχαρώδη διαβήτη τύπου 2 ή είναι αρνητικός στην εν λόγω πάθηση. Η παλινδρόμηση (regression) αφορά στην χαρτογράφηση της εισόδου σε μία αριθμητική τιμή.[12]

Οι κυριότερες τεχνικές Μηχανικής Μάθησης με επίβλεψη είναι:

- Μάθηση Εννοιών (Concept Learning)
- Δέντρα Απόφασης (Decision Trees)
- Μάθηση Κανόνων (Rule Learning)
- Μάθηση κατά Περίπτωση (Instance Based Learning)
- Μάθηση κατά Bayes
- Γραμμική Παρεμβολή (Linear Regression)
- Νευρωνικά Δίκτυα (Neural Networks)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

2.2.2.2 Μη επιβλεπόμενη μάθηση

Σε αυτήν την κατηγορία μηχανικής μάθησης, το σύστημα πρέπει μόνο του να ανακαλύψει "συσχετίσεις" και "ομάδες" και να δημιουργήσει εκ νέου πρότυπα από ένα σύνολο δεδομένων που του παρέχεται, μόνο που το σύνολο

δεδομένων αυτή τη φορά δεν θα περιέχει πληροφορίες εξόδου. Οι κανόνες συσχέτισης (association rules) και οι ομάδες (clusters), οι οποίες προκύπτουν από τη διαδικασία της ομαδοποίησης βασίζονται αποκλειστικά και μόνο στις ιδιότητες του συνόλου και αποτελούν χαρακτηριστικά παραδείγματα της μη επιβλεπόμενης μάθησης.[11]

Αυτό που κάνει τη μη επιβλεπόμενη μάθηση τόσο ενδιαφέρουσα είναι ότι η συντριπτική πλειονότητα των δεδομένων δεν φέρει ετικέτα. Έχοντας ευφυείς αλγόριθμους που μπορούν να πάρουν πληθώρα δεδομένων (terabyte of data) χωρίς ετικέτα και να τα κατανοήσουν, είναι μια τεράστια πηγή δυνητικού κέρδους για πολλές βιομηχανίες. Αυτό από μόνο του θα μπορούσε να βοηθήσει στην αύξηση της παραγωγικότητας σε διάφορους τομείς.

Επειδή η μη επιβλεπόμενη μάθηση βασίζεται στα δεδομένα και τις ιδιότητές τους, προκύπτει ότι η μη επιβλεπόμενη μάθηση έχει ως βάση τα δεδομένα (data driven). Τα αποτελέσματα από μια μη επιβλεπόμενη μαθησιακή εργασία ελέγχονται από τα δεδομένα και τον τρόπο με τον οποίο διαμορφώνονται.[11]

2.2.2.3 Ενισχυτική μάθηση

Εκτός από τις παραπάνω τεχνικές μηχανικής μάθησης, υπάρχει και η ενισχυτική μάθηση (Reinforcement Learning) κατά την οποία ο αλγόριθμος μαθαίνει μία στρατηγική ενεργειών μέσα από την άμεση αλληλεπίδραση με το περιβάλλον. Ο αλγόριθμος ενίσχυσης μαθαίνει δοκιμάζοντας διάφορους τρόπους για την επίλυση ενός προβλήματος και ενώ στην αρχή κάνει πολλά λάθη, εν συνεχεία ο αλγόριθμος μάθησης μαθαίνει να κάνει λιγότερα λάθη κάθε φορά που παρουσιάζεται το ίδιο πρόβλημα. Για το λόγο αυτό η ενισχυτική μάθηση είναι γνωστή και ως μάθηση από λάθη. Χρησιμοποιείται κυρίως σε προβλήματα σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.[11]

2.2.2.4 Άλλες υποκατηγορίες

Ανάμεσα στις παραπάνω κατηγορίες εισήχθη και η έννοια της *ημι-επιβλεπόμενης μάθησης* (Semi-Supervised Learning). Σε αυτόν τον τύπο

εκμάθησης, ο αλγόριθμος εκπαιδεύεται βάσει ενός συνδυασμού δεδομένων με ετικέτα και χωρίς σήμανση. Συνήθως αυτός ο συνδυασμός θα περιέχει μια πολύ μικρή ποσότητα δεδομένων με επισήμανση και μια πολύ μεγάλη ποσότητα δεδομένων χωρίς ετικέτα. Η βασική διαδικασία είναι ότι πρώτα, ο προγραμματιστής θα συγκεντρώσει παρόμοια δεδομένα χρησιμοποιώντας έναν αλγόριθμο μάθησης χωρίς επίβλεψη και, στη συνέχεια, θα χρησιμοποιήσει τα υπάρχοντα δεδομένα με ετικέτα για να επισημάνει τα υπόλοιπα δεδομένα χωρίς σήμανση. Οι τυπικές περιπτώσεις χρήσης ενός τέτοιου τύπου αλγορίθμου έχουν μια κοινή ιδιότητα μεταξύ τους, στο ότι η απόκτηση δεδομένων χωρίς ετικέτα είναι σχετικά φθηνή, ενώ η επισήμανση των εν λόγω δεδομένων είναι πολύ ακριβή.[13]

Η *μεταγωγή* είναι μια ειδική περίπτωση της αρχής αυτής, όπου το σύνολο των καταστάσεων του προβλήματος είναι γνωστό κατά το χρόνο εκμάθησης, όμως ένα μέρος των στόχων λείπουν.

Επιπλέον, υπάρχει η *διαδικασία εκμάθησης* (Meta Learning) που μαθαίνει στη μηχανή τις δικές της επαγωγικές μεθόδους, βασιζόμενη στην προηγούμενη εμπειρία.

Τέλος, υπάρχει η αναπτυξιακή μάθηση ή αναπτυξιακή ρομποτική (Developmental robotics), η οποία έχει αναπτυχθεί για την εκμάθηση των ρομπότ. Το ρομπότ χρησιμοποιώντας ένα σύνολο εσωστρεφών αναπτυξιακών αρχών, το οποίο ρυθμίζει την αλληλοεπίδραση τόσο με τους ανθρώπους εκπαιδευτές, όσο και με το κοινωνικό περιβάλλον του μπορεί να αποκτήσει μία ποικιλία διανοητικών δεξιοτήτων. [14]

2.2.2.5 Διάκριση μηχανικής μάθησης με βάση το επιθυμητό αποτέλεσμα

Η μηχανική μάθηση διακρίνεται επίσης και σε περιπτώσεις ανάλογα με το επιθυμητό αποτέλεσμα του συστήματος στα παρακάτω είδη:

- Κατηγοριοποίηση (Classification): Η κατηγοριοποίηση (classification) είναι ένας τύπος εποπτευόμενης μάθησης και αποτελεί μία από τις πιο

διαδεδομένες ενέργειες της μηχανικής μάθησης. Καθορίζει την κλάση στην οποία ανήκουν τα στοιχεία δεδομένων και χρησιμοποιείται καλύτερα όταν η έξοδος έχει πεπερασμένες και διακριτές τιμές. Πιο αναλυτικά τα δεδομένα εισόδου χωρίζονται σε δύο ή περισσότερες κλάσεις, και η μηχανή πρέπει να κατασκευάσει ένα μοντέλο, το οποίο θα αντιστοιχίζει τα δεδομένα σε μία ή περισσότερες (multi-label) κλάσεις, συνεπώς αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών).[15]

- Παλινδρόμηση (Regression): Η ανάλυση παλινδρόμησης (regression) χρησιμοποιείται σε στατιστικά στοιχεία για την εύρεση τάσεων στα δεδομένα. Πιο συγκεκριμένα είναι μια αξιόπιστη μέθοδος προσδιορισμού των μεταβλητών που επηρεάζουν ένα θέμα ενδιαφέροντος. Η διαδικασία εκτέλεσης μιας παλινδρόμησης επιτρέπει να προσδιοριστεί με σιγουριά ποιοι παράγοντες έχουν μεγαλύτερη σημασία, ποιοι παράγοντες μπορούν να αγνοηθούν και πώς αυτοί οι παράγοντες επηρεάζουν ο ένας τον άλλον.[16]
- Συσταδοποίηση (Clustering): Η συσταδοποίηση (clustering) είναι μία τεχνική της μηχανικής μάθησης που περιλαμβάνει την ομαδοποίηση σημείων δεδομένων. Είναι ένας τύπος μεθόδου μάθησης χωρίς επίβλεψη. Χρησιμοποιείται ως διαδικασία για την εύρεση ουσιαστικής δομής, επεξηγηματικών υποκείμενων διαδικασιών, γενετικών χαρακτηριστικών και συσταδοποιήσεων που μπορεί να υπάρχουν σε ένα σύνολο παραδειγμάτων. Δεδομένου ενός συνόλου σημείων δεδομένων, μπορούμε να χρησιμοποιήσουμε έναν αλγόριθμο συσταδοποίησης για να ταξινομήσουμε κάθε σημείο δεδομένων σε μια συγκεκριμένη ομάδα. Θεωρητικά, τα σημεία δεδομένων που βρίσκονται στην ίδια ομάδα πρέπει να έχουν παρόμοιες ιδιότητες ή χαρακτηριστικά, ενώ τα σημεία δεδομένων σε διαφορετικές ομάδες θα πρέπει να έχουν πολύ ανόμοιες ιδιότητες ή χαρακτηριστικά. Στην Επιστήμη Δεδομένων, μπορεί να χρησιμοποιηθεί η ανάλυση συμπλέγματος για να εξορυχθούν πολύτιμες πληροφορίες από τα δεδομένα, βλέποντας σε ποιες ομάδες εμπίπτουν τα σημεία δεδομένων όταν εφαρμοστεί ένας αλγόριθμος ομαδοποίησης.[17-18]

- Εκτίμηση πυκνότητας (Density Estimation): είναι η κατασκευή ενός εκτιμητή της συνάρτησης πυκνότητας με βάση την κατανομή των δεδομένων.[19]
- Μείωση διαστασιμότητας (Dimensionality Reduction): Είναι η διαδικασία με την οποία επιτυγχάνεται η μείωση του αριθμού των υπό εξέταση χαρακτηριστικών μεταβλητών του συνόλου δεδομένων. Καθώς ο αριθμός των χαρακτηριστικών αυξάνεται, το μοντέλο γίνεται πιο περίπλοκο. Όσο υψηλότερος είναι ο αριθμός των χαρακτηριστικών, τόσο πιο δύσκολα γίνεται η απεικόνιση του συνόλου εκπαίδευσης και στη συνέχεια και η επεξεργασία του. Ένα μοντέλο μηχανικής εκμάθησης που εκπαιδεύεται σε μεγάλο αριθμό χαρακτηριστικών, εξαρτάται όλο και περισσότερο από τα δεδομένα στα οποία εκπαιδεύτηκε και με τη σειρά του υπερ-εκπαιδευτεί, με αποτέλεσμα την κακή απόδοση σε πραγματικά δεδομένα, ξεπερνώντας τον σκοπό του. Όσο λιγότερα χαρακτηριστικά διαθέτει το σύνολο δεδομένων εκπαίδευσης, τόσο λιγότερες υποθέσεις κάνει το μοντέλο και τόσο απλό γίνεται το μοντέλο. Με τη μείωση διαστάσεων επιτυγχάνονται τα εξής [20-21]:

  - Λιγότερα παραπλανητικά δεδομένα που σημαίνει ότι η ακρίβεια του μοντέλου βελτιώνεται.
  - Λιγότερες διαστάσεις, που σημαίνει μικρότερη υπολογιστική δύναμη.
  - Λιγότερα δεδομένα που σημαίνει ότι οι αλγόριθμοι εκπαιδεύονται πιο γρήγορα.
  - Λιγότερα δεδομένα που σημαίνει λιγότερος απαιτούμενος χώρος αποθήκευσης.
  - Λιγότερες διαστάσεις που επιτρέπουν τη χρήση αλγορίθμων ακατάλληλων για μεγάλο αριθμό διαστάσεων.
  - Αφαιρούνται οι περιττές λειτουργίες και ο θόρυβος.

2.2.3 Εφαρμογές της μηχανικής μάθησης

Κατά την τελευταία δεκαετία, η μηχανική μάθηση χρησιμοποιήθηκε σε μια πληθώρα από εφαρμογές όπως αυτοκίνητα αυτο-οδήγησης, πρακτική

αναγνώριση ομιλίας, αποτελεσματική αναζήτηση στο διαδίκτυο και μια πολύ βελτιωμένη κατανόηση του ανθρώπινου γονιδιώματος. Η μηχανική μάθηση είναι τόσο διαδεδομένη σήμερα που πιθανότατα χρησιμοποιείται δεκάδες φορές την ημέρα χωρίς αυτό να γίνεται αντιληπτό. Πολλοί ερευνητές πιστεύουν επίσης ότι είναι ο καλύτερος τρόπος για να σημειωθεί πρόοδος προς την ανθρώπινη τεχνητή νοημοσύνη.

Τα τελευταία χρόνια όλο και περισσότερες εταιρίες τεχνολογίας όπως η Apple, η Google, το Facebook και η Microsoft έχουν εντάξει την μηχανική μάθηση για την υποστήριξη των υπηρεσιών τους και τα αποτελέσματα είναι θεαματικά. Πιο αναλυτικά, η μηχανική μάθηση χρησιμοποιείται σε συστήματα προτάσεων όπως το Netflix, το YouTube και το Spotify, μηχανές αναζήτησης όπως το Google και το Baidu, ροές κοινωνικών μέσων όπως το Facebook και το Twitter, βοηθοί φωνής όπως η Siri και η Alexa και η λίστα απλά ολοένα και επεκτείνεται.[22]

2.3 Προσέγγιση μέσα από άλλες έρευνες

Στην υποενότητα αυτή παρουσιάζονται τρεις έρευνες-δημοσιεύσεις, οι οποίες έχουν εκπονηθεί χρησιμοποιώντας το ίδιο σύνολο δεδομένων Pima Indians Diabetes Database ως προς τη μεθοδολογία και τα αποτελέσματά τους.

2.3.1 Πρώτη έρευνα: Performance Evaluation of Machine Learning Models for Diabetes Prediction

Για την υλοποίηση της έρευνας με τίτλο “Performance Evaluation of Machine Learning Models for Diabetes Prediction” χρησιμοποιήθηκε το εν λόγω σύνολο δεδομένων, λαμβάνοντας υπόψιν ολόκληρο το σύνολο, ενώ δεν έγινε κάποια προεπεξεργασία των δεδομένων για τη χρήση των αλγορίθμων μηχανικής μάθησης. Οι αλγόριθμοι μηχανικής μάθησης που εφαρμόστηκαν στη συγκεκριμένη δημοσίευση είναι οι εξής: KNN Classification (KNN), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression και Random Forest Classification (RF). Οι παραπάνω αλγόριθμοι αξιολογήθηκαν με βάση διάφορες μετρήσεις απόδοσης όπως το ποσοστό

ακρίβειας (Accuracy), την ανάκληση (Recall), τη βαθμολογία f1 (f1-score), το ποσοστό εσφαλμένης ταξινόμησης (Misclassification Rate) και τη βαθμολογία ROC-AUC (ROC-AUC-Score). Τα αποτελέσματα των αλγορίθμων παρουσιάζονται στον παρακάτω πίνακα:

Classifier	Accuracy %	Recall	f1-Score	Misclassification Rate	ROC-AUC-Score
KNN	**73.43**	69	69	31.1	69.8
Decision Tree	**70.31**	72	72	28.57	69.2
Naive Bayes	75.52	74	74	25.97	70.2
Support Vector Machine	**65.63**	64	49	36.36	60.5
Logistic Regression	**77.6**	76	75	23.8	73.6
Random Forest	**74.30**	69	69	29.0	70.1

Πίνακας 1: Πίνακας αποτελεσμάτων πρώτης έρευνας-δημοσίευσης

Τα αποτελέσματα με έντονο χρώμα (bold) αναφέρονται στους αλγορίθμους που χρησιμοποιήθηκαν και στη συγκεκριμένη διπλωματική έρευνα.

Από τον παραπάνω πίνακα αποτελεσμάτων διαπιστώθηκε ότι ο Logistic Regression είναι ο αλγόριθμος που επιτυγχάνει υψηλότερη αποτελεσματικότητα στην έρευνα με ποσοστό ακρίβειας 77.6%.[23]

2.3.2 Δεύτερη έρευνα: Prediction of Onset Diabetes using Machine Learning Techniques

Στη συγκεκριμένη δημοσίευση με τίτλο “Prediction of Onset Diabetes using Machine Learning Technique“ προκειμένου να μην μειωθεί αισθητά το σύνολο

δεδομένων, η προεπεξεργασία των δεδομένων που χρησιμοποιήθηκε ήταν η αφαίρεση των εγγραφών εκείνων που περιείχαν τιμές εκτός των φυσιολογικών ορίων. Πιο συγκεκριμένα, αφαιρέθηκαν οι εγγραφές όπου περιείχαν πολλές μηδενικές τιμές στα χαρακτηριστικά τους. Από τις 768 εγγραφές διατηρήθηκαν για την έρευνα οι 755, συνεπώς έγινε αφαίρεση 13 εγγραφών. Για την κατηγοριοποίηση της εν λόγω έρευνας χρησιμοποιήθηκε το εργαλείο Weka (Waikato Environment for Knowledge Analysis).

Οι αλγόριθμοι που χρησιμοποιήθηκαν στην εν λόγω δημοσίευση είναι οι εξής: Naïve Bayes, Logistic Regression, Multilayer perception, Support Vector Machine, IBK, AdaBoostM1, Bagging, OneR, J48 και Random Forrest. Οι παραπάνω αλγόριθμοι εφαρμόστηκαν με τη μέθοδο του cross validation ορίζοντας το k fold ίσο με 10. Στον παρακάτω πίνακα παρουσιάζονται αναλυτικά τα αποτελέσματα των αλγορίθμων που χρησιμοποιήθηκαν.

Category	Classifier Name	Sensitivity	Specificity	PPV	NPV	AUC	Total Accuracy
Bayes	NaiveBayes	0.80	0.67	0.84	0.61	0.815	75.76
Function	LR	0.80	0.74	0.89	0.58	0.833	**78.01**
	MLP	0.81	0.68	0.85	0.63	0.817	76.82
	SVM	0.78	0.74	0.90	0.53	0.716	**77.08**
Lazy	IBK	0.77	0.59	0.79	0.55	0.668	70.99
Meta	AdaBoostM1	0.78	0.66	0.85	0.56	0.805	74.70
	Bagging	0.79	0.66	0.84	0.59	0.822	75.25
Rules	OneR	0.73	0.64	0.88	0.40	0.642	71.25
Trees	J48	0.81	0.63	0.79	0.66	0.763	74.30
	Random Forest	0.78	0.67	0.86	0.54	0.808	**74.83**

Πίνακας 2: Πίνακας αποτελεσμάτων δεύτερης έρευνας-δημοσίευσης

Τα αποτελέσματα με έντονο χρώμα (bold) αναφέρονται στους αλγορίθμους που χρησιμοποιήθηκαν και στην συγκεκριμένη διπλωματική έρευνα.

Από τον παραπάνω πίνακα αποτελεσμάτων διαπιστώθηκε ότι ο Logistic Regression είναι ο αλγόριθμος που επιτυγχάνει υψηλότερη αποτελεσματικότητα στην έρευνα με ποσοστό ακρίβειας 78.01%. [24]

2.3.3 Τρίτη έρευνα: Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm

Σε αυτήν την έρευνα με τίτλο " Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm" χρησιμοποιήθηκε το σύνολο δεδομένων Pima Indian για την εκπαίδευση αλγορίθμων μηχανικής μάθησης για την πρόβλεψη του σακχαρώδη διαβήτη. Οι αλγόριθμοι που χρησιμοποιήθηκαν στην εν λόγω έρευνα είναι οι εξής: Support Vector Machine (SVM), K- Nearest Neighbors(KNN), Naive Bayes Algorithm (GNB) και Artificial Neural Network (ANN). Πιο συγκεκριμένα όσον αφορά το τεχνητό νευρωνικό δίκτυο (ANN) γίνεται χρήση του MLP (multilayer perceptron), δηλαδή ενός επιβλεπόμενου αλγόριθμου μηχανικής μάθησης.

Αρχικά, οι αλγόριθμοι εφαρμόστηκαν χωρίς καμία προ επεξεργασία των δεδομένων για διαφορετικά μεγέθη συνόλου δεδομένων για την εκπαίδευσή τους και τα αποτελέσματα των αλγορίθμων παρουσιάζονται στον παρακάτω πίνακα:

Training Dataset Size / Category	**368**	**468**	**568**	**668**	**Average Accurracy**
SVM	63	63	63	63	**63**
KNN	64	68	68	66	**66.5**
GNB	76	77	76	76	76.25
ANN	63	63	63	63	63

Πίνακας 3: Πίνακας αποτελεσμάτων τρίτης έρευνας-δημοσίευσης χωρίς προεπεξεργασία δεδομένων

Στη συνέχεια, για τη βελτίωση της αποτελεσματικότητας των αλγορίθμων μηχανικής μάθησης έγινε κανονικοποίηση (Normalization) των δεδομένων με χρήση του μοντέλου κανονικοποίησης Min Max Scaler (MMS). Το MMS περιορίζει τα δεδομένα στο εύρος [0, 1] ή [-1, 1]. Με τη χρήση της μεθόδου προεπεξεργασίας Min Max Scaler (MMS) επιτεύχθηκαν υψηλότερα αποτελέσματα, όπως παρουσιάζονται στον παρακάτω πίνακα.

Training Dataset Size / Category	368	468	568	668	Average Accuracy
SVM+ MMS	78	78	78.2	78	**78.05**
KNN+ MMS	75.3	76	75.1	75.6	**75.5**
GNB+ MMS	79.1	79.3	79.1	79.5	79.3
ANN+ MMS	81.8	82.3	82.9	82.4	82.35

Πίνακας 4: Πίνακας αποτελεσμάτων τρίτης έρευνας-δημοσίευσης με προεπεξεργασία δεδομένων

Τα αποτελέσματα με έντονο χρώμα (bold) αναφέρονται στους αλγορίθμους που χρησιμοποιήθηκαν και στη συγκεκριμένη διπλωματική έρευνα.

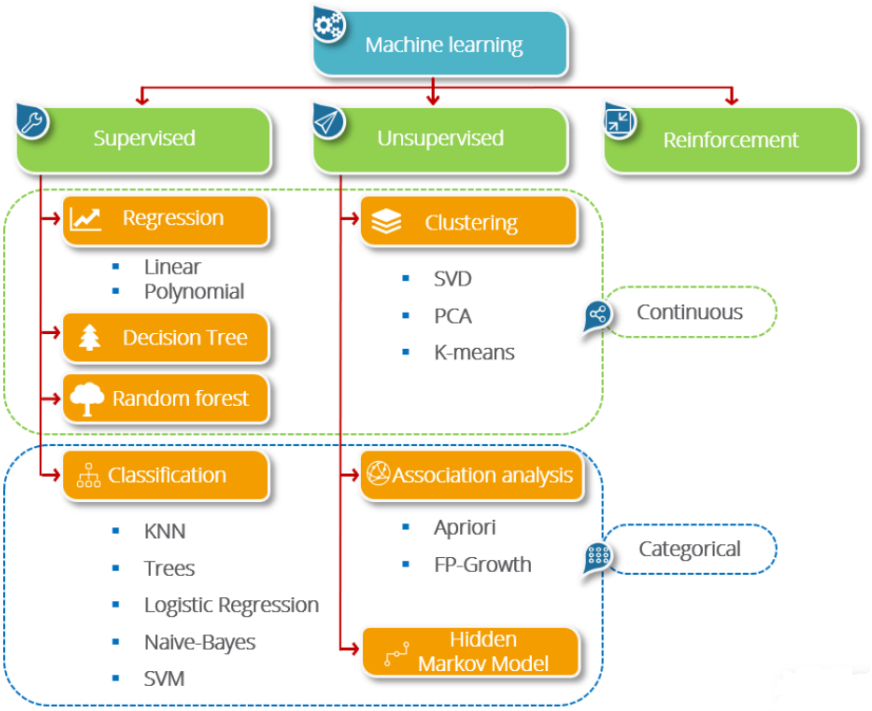
Από τον παραπάνω πίνακα αποτελεσμάτων, διαπιστώθηκε ότι ο Artificial Neural Network (ANN) είναι ο αλγόριθμος που επιτυγχάνει υψηλότερη αποτελεσματικότητα στην έρευνα με ποσοστό μέσης ακρίβειας 82.35%. [25]

ΚΕΦΑΛΑΙΟ 3 – ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΑΝΑΛΥΣΗ

Στο κεφάλαιο αυτό θα περιγραφεί η επιλογή των αλγορίθμων μηχανικής μάθησης, και θα αναλυθούν τόσο τα μέσα που χρειάστηκαν (γλώσσα προγραμματισμού και πλατφόρμα) όσο και η ανάλυση των δεδομένων. Επίσης, θα περιγραφεί λεπτομερώς η προ-επεξεργασία των δεδομένων με την οποία θα καταστήσει τα δεδομένα κατάλληλα για χρήση από αλγόριθμους μηχανικής μάθησης.

3.1 Επιλογή αλγορίθμων

Λαμβάνοντας υπόψιν τις κατηγορίες μηχανικής μάθησης που αναλύθηκαν στο προηγούμενο κεφάλαιο, καθώς και το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα μεταπτυχιακή διπλωματική εργασία για την πρόβλεψη του σακχαρώδη διαβήτη τύπου 2, είναι σαφές ότι η επιβλεπομένη μηχανική μάθηση είναι η καταλληλότερη επιλογή. Παρακάτω συνοψίζονται οι πιο διαδεδομένοι αλγόριθμοι ανά βασική κατηγορία μηχανικής μάθησης (Εικόνα 3).



Εικόνα 3: Κατηγορίες και αλγόριθμοι μηχανικής μάθησης [26]

Οι πέντε αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης που επιλέχθηκαν είναι οι εξής:

- Λογιστική παλινδρόμηση (Logistic Regression): Η λογιστική παλινδρόμηση χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι δυαδική. Είναι ένα μη γραμμικό μοντέλο, τα σφάλματα, του οποίου δεν υπακούν στην κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή. Όπως όλες οι αναλύσεις παλινδρόμησης, η λογιστική παλινδρόμηση είναι μια προγνωστική ανάλυση. Η λογιστική παλινδρόμηση χρησιμοποιείται για την περιγραφή δεδομένων και για την εξήγηση της σχέσης μεταξύ μιας εξαρτώμενης δυαδικής μεταβλητής και μιας ή περισσότερων ονομαστικών, κανονικών, ή ανεξάρτητων επιπέδων αναλογίας. Στη συγκεκριμένη περίπτωση, όπου υφίστανται δύο κατηγορίες, είτε θα έχει σακχαρώδη διαβήτη, είτε δεν θα έχει, προσφέρεται κατάλληλα η χρήση της λογιστικής παλινδρόμησης.[27]

- Δέντρα αποφάσεων (Decision Tree Classifier): είναι από τα πιο ισχυρά και δημοφιλή εργαλεία ταξινόμησης και πρόβλεψης. Ένα δέντρο απόφασης είναι ένα διάγραμμα ροής που έχει τη δομή ενός δέντρου, στον οποίο κάθε εσωτερικός κόμβος δηλώνει μία δοκιμή σε ένα χαρακτηριστικό, κάθε κλάδος αντιπροσωπεύει ένα αποτέλεσμα της δοκιμής και κάθε κόμβος φύλλων κρατά μια ετικέτα κλάσης.[28]

- Τυχαία δάση αποφάσεων (Random Forest): Ο αλγόριθμος του δέντρου αποφάσεων είναι αρκετά εύκολος να κατανοηθεί και να ερμηνευθεί. Ωστόσο, συχνά ένα μόνο δέντρο δεν επαρκεί για την παραγωγή αποτελεσματικών συμπερασμάτων. Για αυτό τον λόγο γίνεται και η χρήση των τυχαίων δέντρων αποφάσεων (Random Forest), ο οποίος είναι ένας αλγόριθμος μηχανικής εκμάθησης με βάση τα δέντρα, που αξιοποιεί τη δύναμη των δέντρων πολλαπλών αποφάσεων για τη λήψη αποφάσεων.[29]

- Μηχανές διανυσμάτων υποστήριξης (Support Vector Machines, SVM): προσπαθούν να βρουν το καλύτερο υπερ-επίπεδο για τον διαχωρισμό

των διαφορετικών κατηγοριών μεγιστοποιώντας την απόσταση μεταξύ των σημείων δείγματος και του υπερ-επίπεδο. Με τη χρήση του συγκεκριμένου αλγορίθμου η κατανομή των δεδομένων γίνεται σε ένα μεγαλύτερο επίπεδο και τα δεδομένα μπορούν να διαχωριστούν με πιο αποτελεσματικό τρόπο.[30]

- Ο αλγόριθμος K-πλησιέστερων γειτόνων (K-Nearest Neighbors, KNN): είναι ένας τύπος επιβλεπόμενου αλγορίθμου μηχανικής μάθησης που χρησιμοποιείται για την ταξινόμηση και την παλινδρόμηση. Τόσο στην ταξινόμηση όσο και στην παλινδρόμηση η είσοδος απαρτίζεται από τα πιο κοντινά σε απόσταση παραδείγματα εκπαίδευσης των χαρακτηριστικών. Ανάλογα με την χρήση του αλγόριθμου καθορίζεται και η έξοδός του. Στη συγκεκριμένη περίπτωση χρησιμοποιείται για την κατηγοριοποίηση των δεδομένων.[31]

3.2 Γλώσσα που χρησιμοποιήθηκε

Η γλώσσα προγραμματισμού που επιλέχθηκε για την υλοποίηση της εργασίας είναι η Python. Αναλυτικότερα είναι μια υψηλού επιπέδου γλώσσα προγραμματισμού, που κύριος στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της. Περιέχει ένα μεγάλο πλήθος από βιβλιοθήκες που διευκολύνουν τους προγραμματιστές να εκφράσουν έννοιες σε λιγότερες γραμμές κώδικα και σε πιο γρήγορο χρόνο σε σύγκριση με άλλες γλώσσες προγραμματισμού, όπως η C++.[32]

Ο κώδικας της Python μπορεί να λειτουργεί σε διαφορετικές πλατφόρμες και εκτελείται σε σύστημα διερμηνέων, που σημαίνει ότι ο κώδικας μπορεί να εκτελεστεί μόλις γραφτεί χωρίς να απαιτείται η εγκατάσταση κάποιου διερμηνευτή της Python.

3.3 Πλατφόρμα που χρησιμοποιήθηκε

Η πλατφόρμα που χρησιμοποιήθηκε για την υλοποίηση της εργασίας είναι το Jupyter. Το Jupyter Notebook είναι μια εφαρμογή web ανοιχτού κώδικα που μας επιτρέπει να δημιουργούμε και να μοιραζόμαστε έγγραφα που περιέχουν ζωντανό κώδικα, εξισώσεις, οπτικοποιήσεις και κείμενο αφήγησης. Οι χρήσεις περιλαμβάνουν: τον καθαρισμό και τον μετασχηματισμό δεδομένων, την αριθμητική προσομοίωση, τη στατιστική μοντελοποίηση, την οπτικοποίηση δεδομένων, την εκμάθηση μηχανών και πολλά άλλα.[33] Με άλλα λόγια, είναι μία δημοφιλής εφαρμογή που μας επιτρέπει να επεξεργαζόμαστε, να εκτελούμε και να μοιραζόμαστε τον κώδικα Python σε μια προβολή ιστού. Μας επιτρέπει να τροποποιήσουμε και να εκτελέσουμε εκ νέου τμήματα του κώδικά μας με πολύ ευέλικτο τρόπο.[34]

3.4 Περιγραφή δεδομένων

Για την υλοποίηση της διπλωματικής εργασίας χρησιμοποιήθηκε το σύνολο δεδομένων Pima Indians Diabetes Database. Το συγκεκριμένο σύνολο δεδομένων δημιουργήθηκε από το Εθνικό Ινστιτούτο Διαβήτη και Πεπτικού και Νεφροπάθειες στις ΗΠΑ (National Institute of Diabetes and Digestive and Kidney Diseases) και είναι διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

Το συγκεκριμένο σύνολο δεδομένων προσφέρεται για ανάπτυξη μοντέλων πρόβλεψης για την εύρεση του σακχαρώδη διαβήτη. Όλοι οι ασθενείς αυτής της βάσης δεδομένων είναι Pima-ινδικές γυναίκες τουλάχιστον 21 ετών και ζουν κοντά στο Φοίνιξ (Phoenix) της Αριζόνα. Επίσης, το συγκεκριμένο σύνολο δεδομένων αφορά τον τύπο 2 του σακχαρώδη διαβήτη.

Πιο αναλυτικά, περιέχει 768 εγγραφές ατόμων και 8 μεταβλητές, οι οποίες κρίθηκαν ως σημαντικές παράγοντες για την εμφάνιση διαβήτη στον ασθενή. Οι μεταβλητές που περιέχονται είναι οι εξής:

Χαρακτηριστικά	Επεξήγηση
Pregnancies	Συνολικός αριθμός που έχει μείνει έγκυος.
Glucose	Συγκέντρωση γλυκόζης πλάσματος 2 ώρες σε δοκιμασία ανοχής γλυκόζης από το στόμα.
BloodPressure	Διαστολική αρτηριακή πίεση (mmHg).
SkinThickness	Το πάχος της πτυχής του δέρματος στον σημείο του τρικέφαλου μυ (mm).
Insulin	Η ινσουλίνη ορού 2 ωρών (mu U / ml).
BMI	Ο δείκτης μάζας σώματος (kg / m^2).
DiabetesPedigreeFunction	Λειτουργία γενεαλογικού διαβήτη.
Age	Ηλικία του ασθενή.

Πίνακας 5: Χαρακτηριστικές μεταβλητές

Επίσης, περιέχεται μία επιπλέον στήλη με την οποία χαρακτηρίζεται ο ασθενής ως θετικός (1) ή αρνητικός (2) στην ασθένεια του σακχαρώδη διαβήτη.

Παρακάτω παρουσιάζεται στιγμιότυπο του συνόλου δεδομένων που θα χρησιμοποιηθούν για την υλοποίηση της διπλωματικής εργασίας (Εικόνα 4).

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0

Εικόνα 4: Απεικόνιση δεδομένων

Το συγκεκριμένο σύνολο δεδομένων έχει χρησιμοποιηθεί ευρέως σε πειράματα μηχανικής μάθησης και έχει χρησιμοποιηθεί σε αρκετές μελέτες, όπως σε αυτές που περιγράφονται στις παρακάτω δημοσιεύσεις:

- «Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks» από τους Kamer Kayaer και Tulay Yildirim.

- «Region based Support Vector Machine Algorithm for Medical Diagnosis on Pima Indian Diabetes DataSet» από τους Savvas Karatsiolis και Christos N. Schizas.

- «Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients» από τους Asha Gowda Karegowda, M.A. Jayaram και A.S. Manjunath.

3.5 Ανάλυση δεδομένων

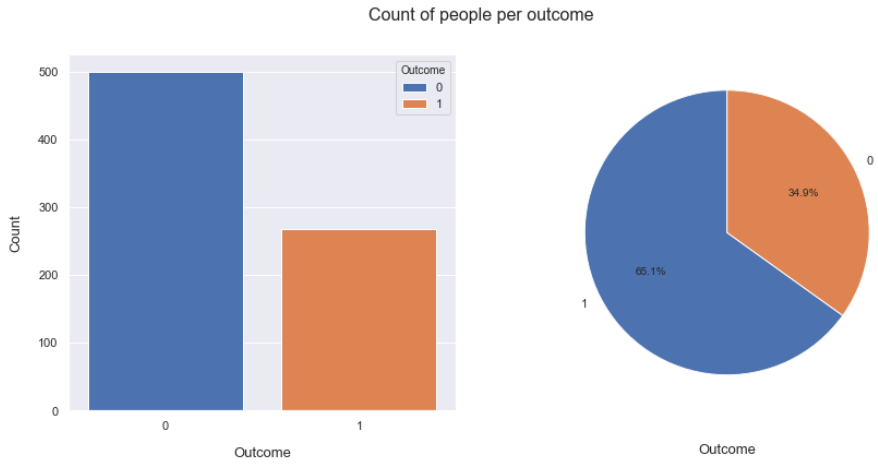
3.5.1 Γενικά στατιστικά στοιχεία

Το σύνολο δεδομένων, όπως έχει αναφερθεί, περιέχει 768 εγγραφές με 9 στήλες χαρακτηριστικών, εκ των οποίων η τελευταία στήλη χαρακτηρίζει αν το άτομο με τα συγκεκριμένα χαρακτηριστικά έχει ή δεν έχει διαγνωστεί με τη νόσου του σακχαρώδη διαβήτη.

Για την καλύτερη μελέτη και κατανόηση των δεδομένων δημιουργήθηκαν κάποια στατιστικά στοιχεία με τη χρήση μεθόδων που προέρχονται από τη χρήση της γλώσσας Python.

Στο συγκεκριμένο σύνολο δεδομένων, το σύνολο ατόμων που έχουν διαγνωστεί με σακχαρώδη διαβήτη είναι 268 άτομα, ενώ 500 άτομα δεν έχουν διαγνωστεί με την ασθένεια του σακχαρώδη διαβήτη.

Αρχικά, απεικονίζονται μερικές γραφικές παραστάσεις με τη μορφή ιστογράμματος και πίτας του συνολικού αριθμού ατόμων που έχουν διαγνωστεί με διαβήτη, καθώς επίσης και του συνολικού αριθμού ατόμων που δεν έχουν διαγνωστεί με διαβήτη (Εικόνα 5).



Εικόνα 5: Απεικόνιση του πλήθους ατόμων που έχουν και δεν έχουν διαβήτη

Με τη βοήθεια των παραπάνω γραφημάτων, είναι σαφές ότι τα άτομα που δεν έχουν διαγνωστεί με διαβήτη είναι αρκετά περισσότερα από αυτά που έχουν διαγνωστεί με την εν λόγω ασθένεια.

Πιο αναλυτικά, το 65,1% δεν έχουν διαγνωστεί με σακχαρώδη διαβήτη, ενώ μόλις το 34,9% έχουν διαγνωστεί με σακχαρώδη διαβήτη.

Στη συνέχεια, έγινε έλεγχος για το αν τα δεδομένα περιέχουν κενές (null) τιμές, κάτι το οποίο δεν επιβεβαιώθηκε. Άρα, όλα τα πεδία του συνόλου δεδομένων είναι συμπληρωμένα.

Η μέθοδος describe χρησιμοποιήθηκε για την παραγωγή στατιστικών λεπτομερειών όπως φαίνεται παρακάτω (Εικόνα 6).

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Εικόνα 6: Στατιστικά στοιχεία

Τα παραπάνω στατιστικά στοιχεία φανερώνουν ότι η ελάχιστη τιμή (min) σε συγκεκριμένα χαρακτηριστικά εμπεριέχει μηδενικές τιμές, κάτι το οποίο δεν είναι δυνατόν.

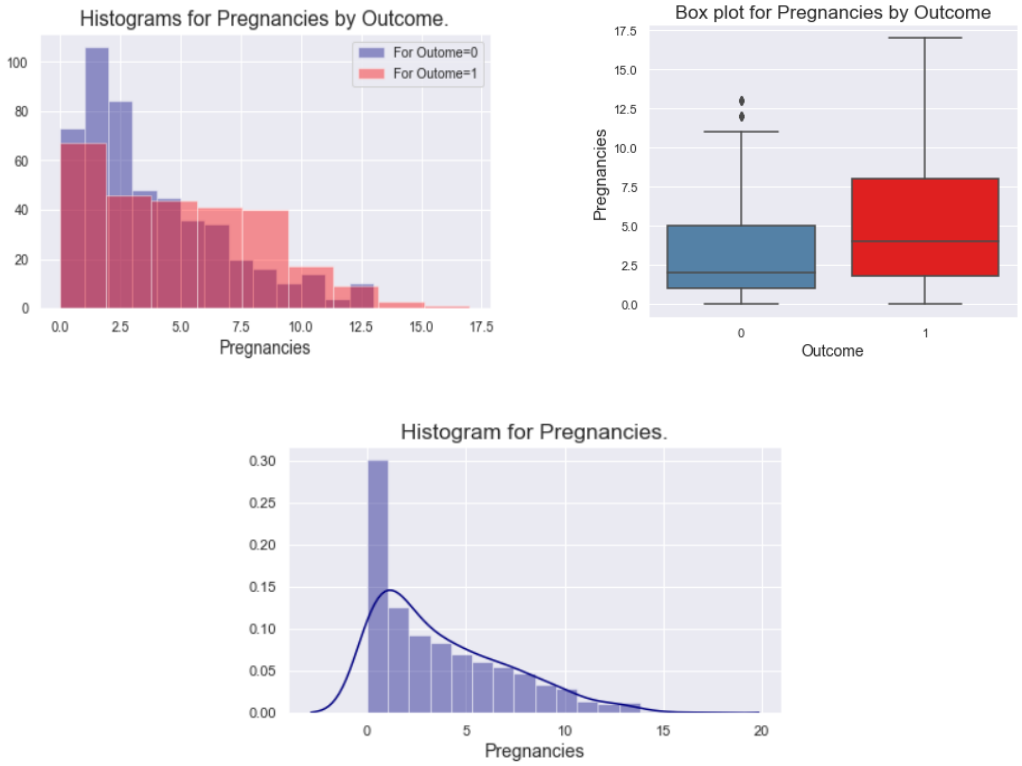
Πιο αναλυτικά, η συγκέντρωση γλυκόζης πλάσματος (Glucose), η διαστολική αρτηριακή πίεση (BloodPressure), το πάχος της πτυχής του δέρματος (SkinThickness), η ινσουλίνη (Insulin) καθώς και ο δείκτης μάζας σώματος (BMI) δεν μπορούν να περιέχουν μηδενικές τιμές. Η αντιμετώπιση των μηδενικών τιμών κρίνεται αναγκαία και θα αναλυθεί περεταίρω στη συνέχεια.

Για την πλήρη κατανόηση του συνόλου των δεδομένων κρίνεται απαραίτητη η μελέτη καθενός από τα χαρακτηριστικά ξεχωριστά με τη χρήση γραφικών παραστάσεων.

3.5.2 Γραφικές παραστάσεις για κάθε χαρακτηριστικό

Εγκυμοσύνες (Pregnancies)

Το χαρακτηριστικό «εγκυμοσύνες (Pregnancies)» αναφέρεται στον συνολικό αριθμό εγκυμοσύνων. Όπως διαπιστώθηκε ήδη από πριν, η ελάχιστη τιμή που περιέχεται στο σύνολο δεδομένων είναι το μηδέν (0) και η μέγιστη είναι το δεκαεφτά (17). Παρακάτω εμφανίζονται γραφικές παραστάσεις σε μορφή ιστογράμματος και θηκογράμματος για την καλύτερη κατανόηση του συγκεκριμένου χαρακτηριστικού (Εικόνα 7).

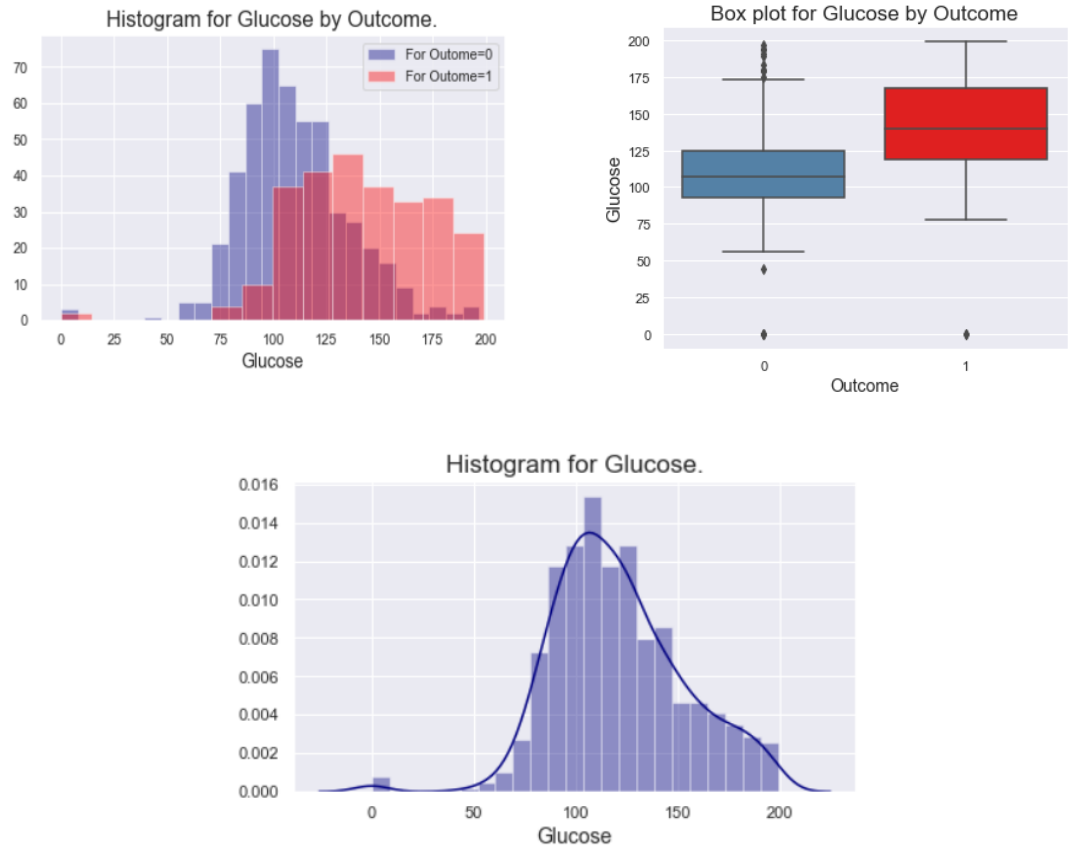


Εικόνα 7: Γραφικές παραστάσεις για το χαρακτηριστικό Pregnancies

Από τις παραπάνω γραφικές παραστάσεις, διαπιστώνεται ότι ένα μεγάλο ποσοστό γυναικών δεν έχει κάποια εγκυμοσύνη, κάτι το οποίο είναι αποδεκτό, αφού το σύνολο δεδομένων περιέχει άτομα από 21 ετών και πάνω. Τα άτομα που είναι κοντά στις ηλικίες των 21 είναι πολύ συνηθισμένο να μην έχουν παιδιά και συνεπώς και εγκυμοσύνες. Επίσης, από το πρώτο ιστόγραμμα συμπεραίνεται ότι οι γυναίκες με περισσότερες εγκυμοσύνες έχουν περισσότερες πιθανότητες να έχουν σακχαρώδη διαβήτη. Από το θηκόγραμμα (boxplot) διαπιστώνεται ότι υπάρχουν κάποιες λίγες εγγραφές εκτός τιμών (Outliers).

Γλυκόζη (Glucose)

Για το χαρακτηριστικό «γλυκόζη (Glucose)», η ελάχιστη τιμή που περιέχεται στο σύνολο δεδομένων είναι το μηδέν (0) και η μέγιστη είναι το δεκαεφτά (17). Παρακάτω εμφανίζονται γραφικές παραστάσεις σε μορφή ιστογράμματος και θηκογράμματος για την καλύτερη κατανόηση του συγκεκριμένου χαρακτηριστικού (Εικόνα 8).

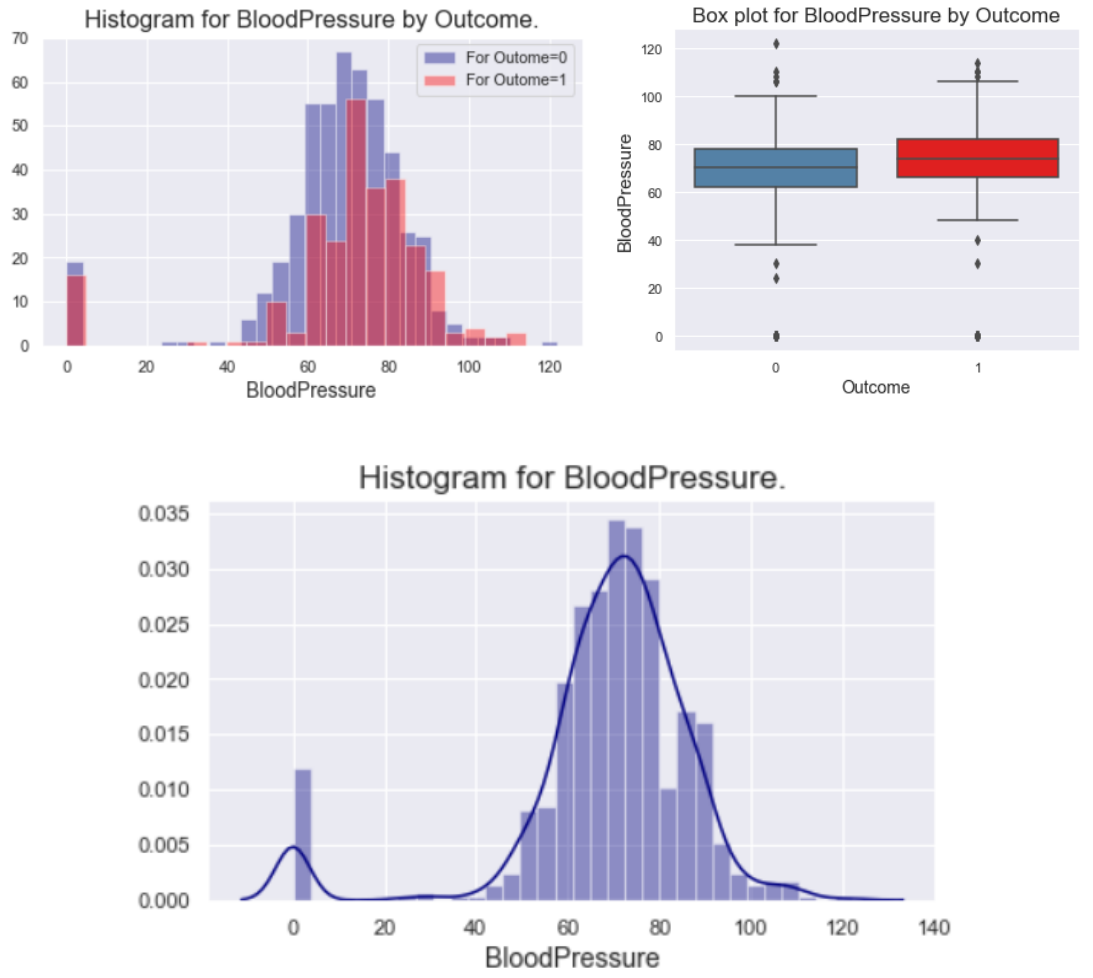


Εικόνα 8: Γραφικές παραστάσεις για το χαρακτηριστικό Glucose

Από τις παραπάνω γραφικές παραστάσεις διαπιστώνεται ότι υπάρχει μία αριστερή ασυμμετρία και ότι τα υψηλότερα επίπεδά γλυκόζης συνδέονται με τα άτομα που έχουν σακχαρώδη διαβήτη. Από το θηκόγραμμα (boxplot) διαπιστώνεται ότι υπάρχουν κάποιες λίγες εγγραφές εκτός τιμών (Outliers).

Διαστολική αρτηριακή πίεση (Blood Pressure)

Για το χαρακτηριστικό «διαστολική αρτηριακή πίεση (Blood Pressure)», η ελάχιστη τιμή που περιέχεται στο σύνολο δεδομένων είναι το μηδέν (0) και η μέγιστη ισούται με 112. Η μονάδα μέτρησης της πτυχής του δέρματος είναι σε χιλιοστά στήλης υδραργύρου (mmHg). Παρακάτω εμφανίζονται γραφικές παραστάσεις σε μορφή ιστογράμματος και θηκογράμματος για την καλύτερη κατανόηση του συγκεκριμένου χαρακτηριστικού (Εικόνα 9).

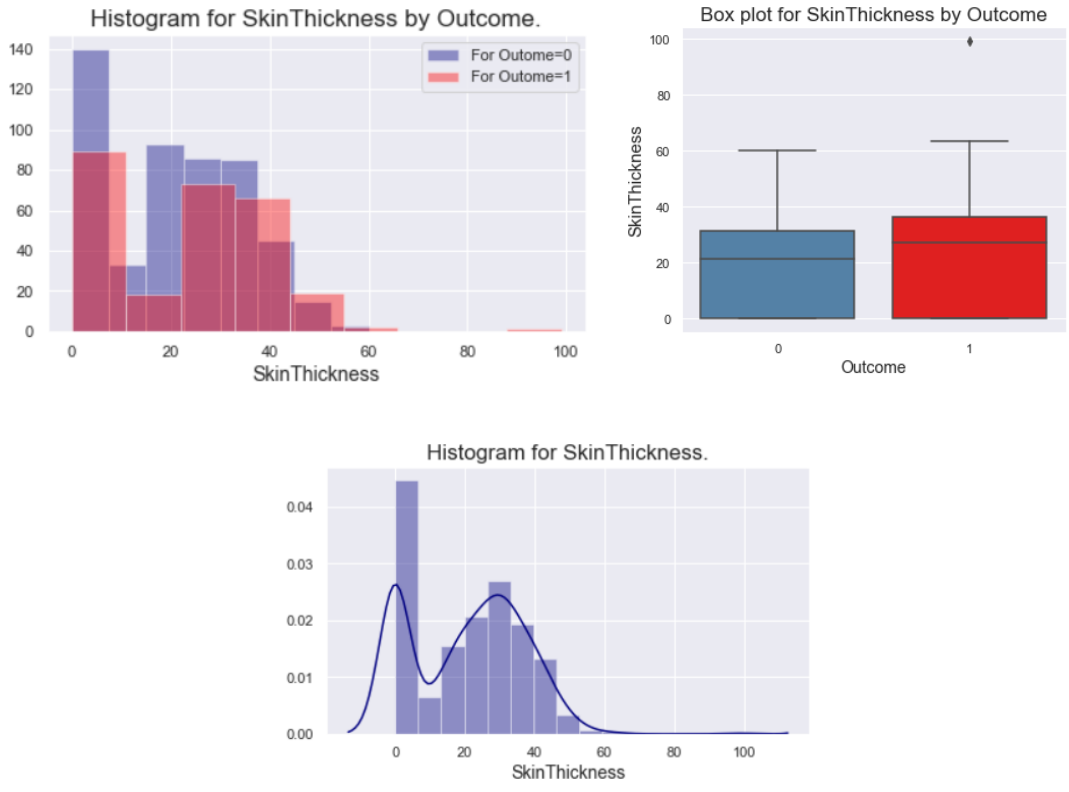


Εικόνα 9: Γραφικές παραστάσεις για το χαρακτηριστικό Blood Pressure

Από τις παραπάνω γραφικές παραστάσεις διαπιστώνεται ότι η κατανομή φαίνεται κανονική, αν και υπάρχει μία αριστερή ασυμμετρία εξαιτίας των μηδενικών τιμών που υπάρχουν στο σύνολο δεδομένων. Οι μηδενικές τιμές θα πρέπει να αντιμετωπιστούν. Οι γυναίκες που έχουν διαγνωστεί με σακχαρώδη διαβήτη έχουν υψηλότερη διαστολική αρτηριακή πίεση, σε σχέση με τις γυναίκες που είναι υγιείς. Επίσης, από το θηκόγραμμα (boxplot) διαπιστώνεται ότι υπάρχουν κάποιες λίγες εγγραφές εκτός τιμών (Outliers).

Πάχος πτυχής δέρματος (SkinThickness)

Για το χαρακτηριστικό «πάχος της πτυχής του δέρματος (SkinThickness)», η ελάχιστη τιμή που περιέχεται στο σύνολο δεδομένων είναι το μηδέν (0) και η μέγιστη ισούται με 99. Η μονάδα μέτρησης της πτυχής του δέρματος είναι σε χιλιοστά (mm). Παρακάτω εμφανίζονται γραφικές παραστάσεις σε μορφή ιστογράμματος και θηκογράμματος για την καλύτερη κατανόηση του συγκεκριμένου χαρακτηριστικού (Εικόνα 10).



Εικόνα 10: Γραφικές παραστάσεις για το χαρακτηριστικό SkinThickness

Από τις παραπάνω γραφικές παραστάσεις διαπιστώνεται ότι υπάρχει μία δεξιά ασυμμετρία. Επίσης, από το θηκόγραμμα (boxplot) είναι εμφανές ότι υπάρχουν ελάχιστες εγγραφές εκτός τιμών (Outliers). Από ιατρικής απόψεως είναι γνωστό ότι οι παχύσαρκοι άνθρωποι έχουν μία προδιάθεση για εμφάνιση σακχαρώδη διαβήτη και συνεπώς μεγαλύτερο πάχος δέρματος. Επίσης, από το θηκόγραμμα (boxplot) διαπιστώνεται ότι υπάρχουν ελάχιστες εγγραφές εκτός τιμών (Outliers).

Ινσουλίνη (Insulin)

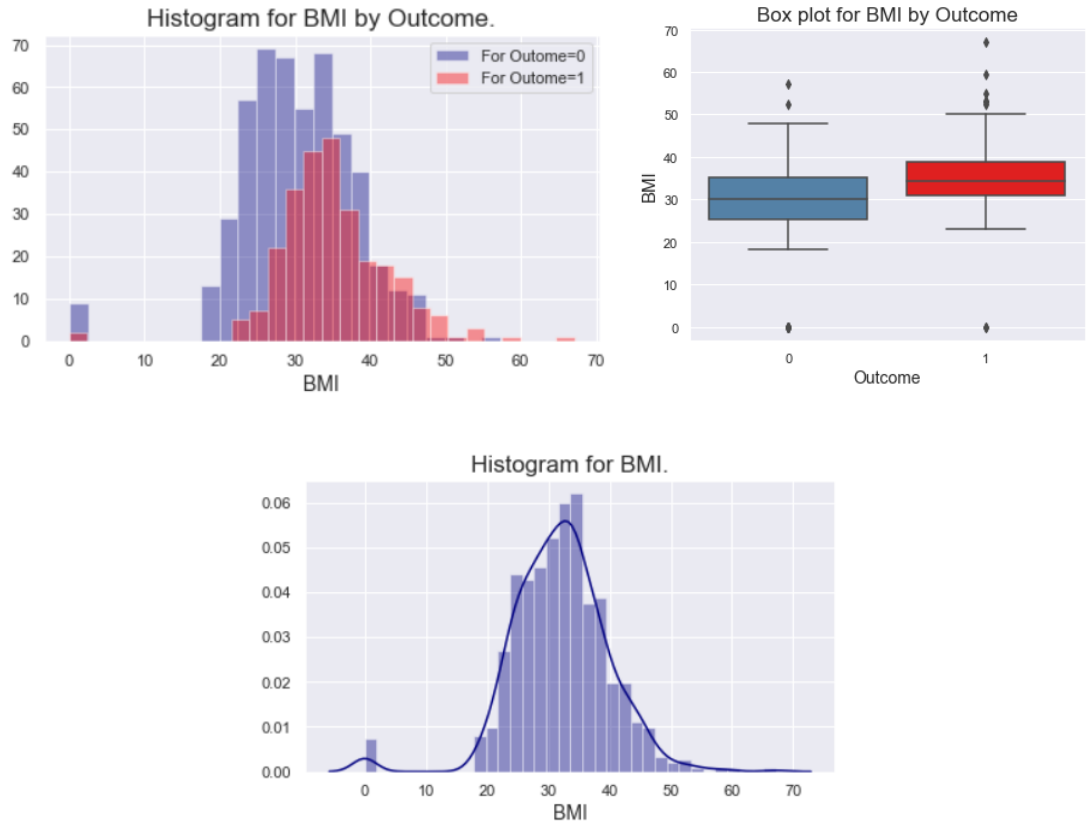
Για το χαρακτηριστικό «ινσουλίνη (Insulin)», η ελάχιστη τιμή που περιέχεται στο σύνολο δεδομένων είναι το μηδέν (0) και η μέγιστη ισούται με 846. Η μονάδα μέτρησης της ινσουλίνη (Insulin) είναι σε mu U/ml. Παρακάτω εμφανίζονται γραφικές παραστάσεις σε μορφή ιστογράμματος και θηκογράμματος για την καλύτερη κατανόηση του συγκεκριμένου χαρακτηριστικού (Εικόνα 11).

Εικόνα 11: Γραφικές παραστάσεις για το χαρακτηριστικό Insulin

Από τις παραπάνω γραφικές παραστάσεις διαπιστώνεται ότι οι άνθρωποι που έχουν διαγνωστεί με σακχαρώδη διαβήτη έχουν μεγαλύτερη ποσότητα ινσουλίνης, σε σχέση με τους ανθρώπους που είναι υγιείς. Από ιατρικής απόψεως είναι γνωστό ότι η ινσουλίνη κυμαίνεται από 16 έως και 166. Οπότε διαπιστώνεται ότι το συγκεκριμένο χαρακτηριστικό περιέχει αρκετές τιμές εκτός ορίων (Outliers) και απαιτείται η αντιμετώπισή τους.

Δείκτης μάζας σώματος (BMI)

Για το χαρακτηριστικό «δείκτης μάζας σώματος (BMI)», η ελάχιστη τιμή που περιέχεται στο σύνολο δεδομένων είναι το μηδέν (0) και η μέγιστη ισούται με 67. Η μονάδα μέτρησης του δείκτης μάζας σώματος (BMI) είναι kg/m^2. Παρακάτω εμφανίζονται γραφικές παραστάσεις σε μορφή ιστογράμματος και θηκογράμματος για την καλύτερη κατανόηση του συγκεκριμένου χαρακτηριστικού (Εικόνα 12).

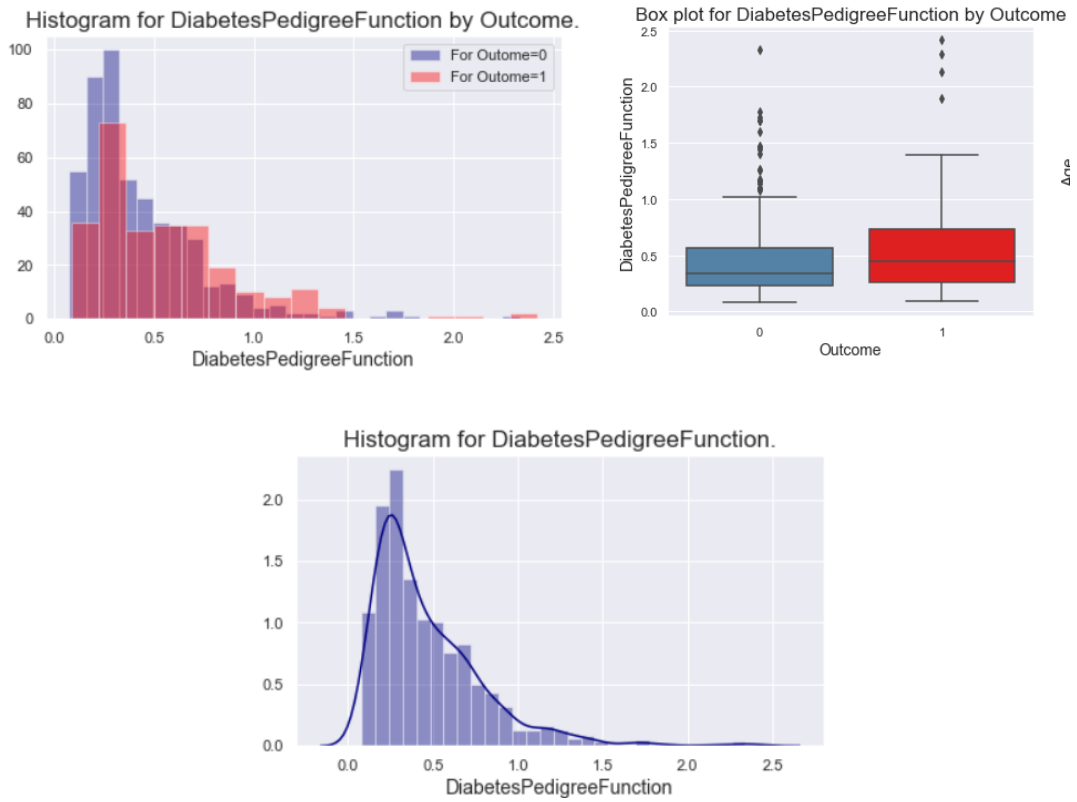


Εικόνα 12: Γραφικές παραστάσεις για το χαρακτηριστικό BMI

Από τις παραπάνω γραφικές παραστάσεις διαπιστώνεται ότι οι περισσότεροι άνθρωποι χαρακτηρίζονται ως παχύσαρκοι, αφού το κανονικό εύρος του δείκτη μάζας σώματος κυμαίνεται από 18 και 25. Επίσης, παρατηρείται ότι οι διαβητικοί άνθρωποι έχουν υψηλότερες τιμές δείκτη μάζας σώματος. Τέλος, από το θηκόγραμμα (boxplot) διαπιστώνεται ότι υπάρχουν ελάχιστες εγγραφές εκτός τιμών (Outliers).

Λειτουργία γενεαλογικού διαβήτη (DiabetesPedigreeFunction)

Για το χαρακτηριστικό «λειτουργία γενεαλογικού διαβήτη (DiabetesPedigreeFunction)», περιέχονται στο σύνολο δεδομένων δεκαδικές τιμές και πιο συγκριμένα οι τιμές κυμαίνονται από 0.078 έως και 2.42. Παρακάτω εμφανίζονται γραφικές παραστάσεις σε μορφή ιστογράμματος και θηκογράμματος για την καλύτερη κατανόηση του συγκεκριμένου χαρακτηριστικού (Εικόνα 13).



Εικόνα 13: Γραφικές παραστάσεις για το χαρακτηριστικό DiabetesPedigreeFunction

Από τις παραπάνω γραφικές παραστάσεις διαπιστώνεται ότι υπάρχει μία έντονη δεξιά ασυμμετρία. Αν και δεν είναι δυνατόν να προβεί κάποιος σε ασφαλή αποτελέσματα, το μόνο που είναι σαφές είναι ότι όσο αυξάνεται η τιμή της λειτουργίας γενεαλογικού διαβήτη, τόσο πιο πιθανό είναι να έχει διαγνωστεί το άτομο με σακχαρώδη διαβήτη.

Ηλικία (Age)

Για το χαρακτηριστικό «ηλικία (Age)», περιέχονται στο σύνολο δεδομένων τιμές από την ελάχιστη τιμή των 21 ετών έως και τη μέγιστη τιμή των 81 ετών. Παρακάτω εμφανίζονται γραφικές παραστάσεις σε μορφή ιστογράμματος και θηκογράμματος για την καλύτερη κατανόηση του συγκεκριμένου χαρακτηριστικού (Εικόνα 14).

Εικόνα 14: Γραφικές παραστάσεις για το χαρακτηριστικό Age

Από τις παραπάνω γραφικές παραστάσεις διαπιστώνεται ότι οι περισσότερες εγγραφές αφορούν νέους ηλικιακά ανθρώπους (κάτω των 30 ετών) και ότι όσο

ο άνθρωπος μεγαλώνει τόσο αυξάνεται η συχνότητα εμφάνισης σακχαρώδους διαβήτη.

3.5.3 Εύρεση μηδενικών τιμών

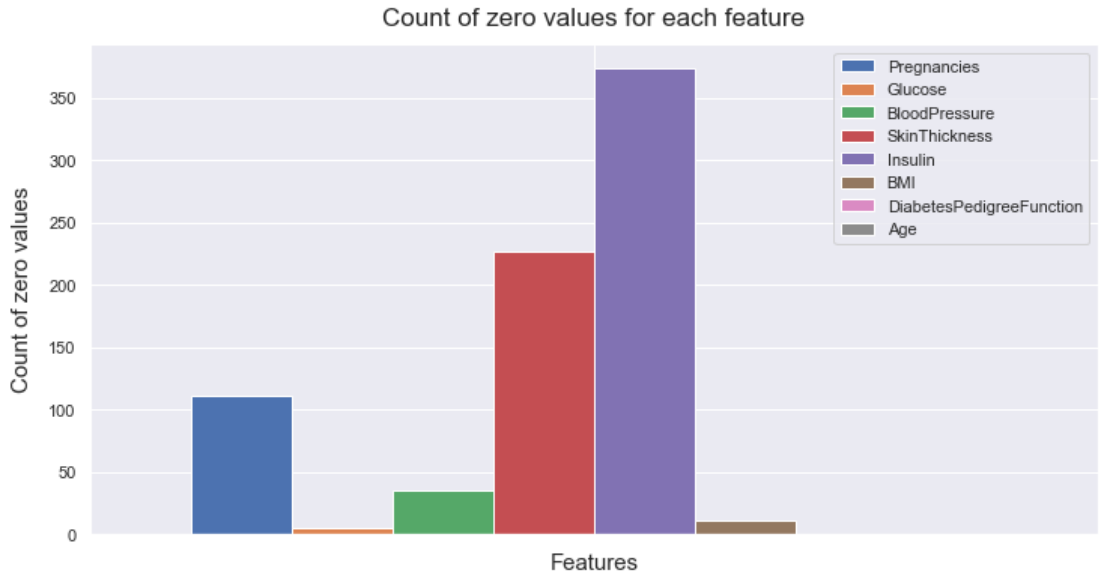
Όπως αποδείχτηκε τόσο από τη μέθοδο describe, όσο και από τις γραφικές παραστάσεις, το σύνολο δεδομένων περιέχει μηδενικές τιμές, οι οποίες για συγκεκριμένα χαρακτηριστικά είναι μη αποδεκτές.

Για να αντιμετωπιστούν οι μηδενικές τιμές θα πρέπει αρχικά να διευκρινιστεί πόσες μηδενικές τιμές περιέχει κάθε ένα από τα χαρακτηριστικά. Χρησιμοποιώντας γραφικές παραστάσεις καθώς και προγραμματιστικές εντολές γίνεται εύρεση του πλήθους μηδενικών τιμών για κάθε χαρακτηριστικό. Το πλήθος των μηδενικών τιμών για κάθε ένα χαρακτηριστικό παρουσιάζεται τον παρακάτω πίνακα.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
111	5	35	227	374	11	0	0

Πίνακας 6: Σύνολο μηδενικών τιμών για κάθε χαρακτηριστικό

Τα παραπάνω αποτελέσματα απεικονίζονται υπό τη μορφή γραφικής παράστασης όπως φαίνεται παρακάτω (Εικόνα 15).



Εικόνα 15: Γραφική απεικόνιση των μηδενικών τιμών κάθε χαρακτηριστικού

3.6 Επεξεργασία δεδομένων

Έχοντας μία πλήρη εικόνα του συνόλου δεδομένων Pima Indians Diabetes Database, εν συνεχεία γίνεται επεξεργασία του συνόλου δεδομένων ώστε να είναι κατάλληλο για να χρησιμοποιηθεί από τους αλγόριθμους μηχανικής μάθησης για την πρόβλεψη του σακχαρώδους διαβήτη.

3.6.1 Αντιμετώπιση μηδενικών τιμών

Στα χαρακτηριστικά του συνόλου δεδομένων, όπως έχει ήδη αναφερθεί υπάρχουν αρκετές μηδενικές τιμές. Η ερώτηση που προκύπτει από αυτό είναι η εξής: Είναι δυνατόν τα χαρακτηριστικά του συνόλου δεδομένων να περιέχουν μηδενικές τιμές;

Για να απαντηθεί το παραπάνω ερώτημα πρέπει να ληφθεί υπόψιν το ιατρικό υπόβαθρο, δηλαδή ποια χαρακτηριστικά μπορούν να δεχθούν μηδενικές τιμές και για ποια δεν είναι δυνατή η μηδενική τιμή. Συνεπώς, μηδενικές τιμές δεν μπορούν να λάβουν τα εξής χαρακτηριστικά του συνόλου δεδομένων: συγκέντρωση γλυκόζης πλάσματος (Glucose), διαστολική αρτηριακή πίεση

(BloodPressure), πάχος της πτυχής του δέρματος (SkinThickness), ινσουλίνη (Insulin) και δείκτης μάζας σώματος (BMI). Αντιθέτως, τα χαρακτηριστικά εγκυμοσύνες (Pregnancies) και η λειτουργία γενεαλογικού διαβήτη (DiabetesPedigreeFunction) μπορούν να λάβουν μηδενικές τιμές και δεν δημιουργούν κάποιο πρόβλημα στην έρευνα της μεταπτυχιακής διπλωματικής εργασίας. Σχετικά με την ηλικία του ασθενή (Age), η μηδενική τιμή θεωρητικά είναι αποδεκτή, ωστόσο το σύνολο δεδομένων περιέχει ως ελάχιστη τιμή την ηλικία των 21 ετών.

Για την αντιμετώπιση των μηδενικών τιμών που εμπεριέχονται στα συγκεκριμένα χαρακτηριστικά του συνόλου δεδομένων που αναφέρθηκαν, υπάρχουν τέσσερις (4) περιπτώσεις που μπορούν να εφαρμοστούν. Η πρώτη περίπτωση είναι να παραμείνουν οι μηδενικές τιμές ως έχουν, η δεύτερη περίπτωση να αφαιρεθούν οι εγγραφές όπου περιέχουν μηδενικές τιμές, η τρίτη περίπτωση να αντικατασταθούν με το μέσο όρο κάθε χαρακτηριστικού και η τελευταία περίπτωση και αυτή που χρησιμοποιήθηκε στη συγκεκριμένη μεταπτυχιακή διπλωματική εργασία είναι η αντικατάσταση με το μέσο όρο του κάθε χαρακτηριστικού αλλά ανά κλάση αποτελέσματος (θετικό ή αρνητικό αποτέλεσμα στον σακχαρώδη διαβήτη). Δηλαδή, αν μία εγγραφή περιέχει μηδενικές τιμές και το αποτέλεσμα της εγγραφής είναι ότι είναι θετική στον σακχαρώδη διαβήτη, τότε η μηδενική τιμή θα αντικατασταθεί με το μέσο όρο του συγκεκριμένου χαρακτηριστικού για τις εγγραφές που έχουν διαγνωστεί με σακχαρώδη διαβήτη.

Στη συνέχεια παρουσιάζονται αναλυτικά οι περιπτώσεις για τα συγκεκριμένα χαρακτηριστικά που δεν μπορούν να λάβουν μηδενικές τιμές.

3.6.1.1 Διατήρηση μηδενικών τιμών

Σε περίπτωση που ένα σύνολο δεδομένων περιέχει μηδενικές τιμές οι οποίες δεν καλύπτουν μεγάλο εύρος των δεδομένων και δεν επηρεάζουν το αποτέλεσμα, τότε μπορούν να διατηρηθούν ως έχειν. Στη συγκεκριμένη περίπτωση όμως, οι μηδενικές τιμές καλύπτουν μεγάλο μέρος του συνόλου

δεδομένων και συνεπώς δεν μπορούν να διατηρηθούν και η αντικατάστασή τους είναι απαραίτητη.

3.6.1.2 Αφαίρεση μηδενικών τιμών

Μία από τις περιπτώσεις για να αντιμετωπιστούν οι μηδενικές τιμές στο σύνολο δεδομένων είναι να αφαιρεθούν οι εγγραφές που περιέχουν μηδενικές τιμές. Το πρόβλημα που προκύπτει από την αφαίρεση των εγγραφών είναι ότι οι εγγραφές που περιέχουν μηδενικές τιμές καλύπτουν ένα πολύ μεγάλο μέρος από το σύνολο δεδομένων με αποτέλεσμα το πλήθος του συνόλου δεδομένων να μειώνεται αισθητά, όπως φαίνεται και από τον παρακάτω πίνακα.

	Με μηδενικές τιμές	Με αφαίρεση μηδενικών τιμών
Σύνολο εγγραφών	768	392

Πίνακας 7: Σύνολο εγγραφών με μηδενικές τιμές και χωρίς

3.6.1.3 Αντικατάσταση με το μέσο όρο κάθε χαρακτηριστικού

Ένας άλλος τρόπος για την αντιμετώπιση των μηδενικών τιμών είναι η αντικατάστασή τους με τον μέσο όρο του κάθε χαρακτηριστικού που βρίσκονται. Δηλαδή, αν για παράδειγμα το χαρακτηριστικό διαστολική αρτηριακή πίεση (BloodPressure) περιέχει μηδενικές τιμές, τότε αυτές θα αντικατασταθούν με το μέσο όρο τιμών του χαρακτηριστικού αυτού, που στην συγκεκριμένη περίπτωση ισούται με 69. Στον πίνακα παρουσιάζονται όλες οι τιμές μέσου όρου κάθε χαρακτηριστικού.

Χαρακτηριστικά	Μέσο όρος χαρακτηριστικού
Συγκέντρωση γλυκόζης πλάσματος (Glucose)	121
Διαστολική αρτηριακή πίεση (BloodPressure)	69
Πάχος της πτυχής του δέρματος (SkinThickness)	21
Ινσουλίνη (Insulin)	80
Δείκτης μάζας σώματος (BMI)	32

Πίνακας 8: Μέσος όρος κάθε χαρακτηριστικού

Θεωρητικά αυτή η μέθοδος θα μπορούσε να χρησιμοποιηθεί αν το σύνολο των δεδομένων περιείχε ισάριθμες εγγραφές για κάθε μία από τις δύο περιπτώσεις, δηλαδή περίπου ίσο αριθμό θετικών και αρνητικών αποτελεσμάτων στον σακχαρώδη διαβήτη.

3.6.1.4 Αντικατάσταση με το μέσο όρο κάθε χαρακτηριστικού ανά κλάση

Ένας άλλος τρόπος για την αντιμετώπιση των μηδενικών τιμών, ο οποίος είναι παρεμφερής με τον προηγούμενο, αλλά στη δεδομένη περίσταση πιο αποτελεσματικός, είναι η αντικατάστασή τους με τον μέσο όρο του κάθε χαρακτηριστικού που βρίσκονται. Δηλαδή, αν για παράδειγμα το χαρακτηριστικό διαστολική αρτηριακή πίεση (BloodPressure) περιέχει μηδενικές τιμές, τότε αυτές θα αντικατασταθούν με το μέσο όρο τιμών της κάθε κλάσης που ανήκουν για το συγκεκριμένο χαρακτηριστικό, που στη συγκεκριμένη περίπτωση αν ανήκει στην κλάση 0, δηλαδή αρνητικό αποτέλεσμα στον σακχαρώδη διαβήτη, ισούται με 68, ενώ αν ανήκει στην

κλάση 1, δηλαδή θετικό αποτέλεσμα στον σακχαρώδη διαβήτη, ισούται με 71. Στον πίνακα παρουσιάζονται όλες οι τιμές μέσου όρου κάθε χαρακτηριστικού ανά κλάση.

Χαρακτηριστικά	**Μέσο όρος χαρακτηριστικού για Output 0**	**Μέσο όρος χαρακτηριστικού για Output 1**
Συγκέντρωση γλυκόζης πλάσματος (Glucose)	110	141
Διαστολική αρτηριακή πίεση (BloodPressure)	68	71
Πάχος της πτυχής του δέρματος (SkinThickness)	20	22
Ινσουλίνη (Insulin)	69	100
Δείκτης μάζας σώματος (BMI)	30	35

Πίνακας 9: Μέσος όρος χαρακτηριστικού ανά κλάση

3.6.2 Ιατρικό υπόβαθρο για κάθε χαρακτηριστικό

Το ιατρικό υπόβαθρο κρίνεται απαραίτητη αφού η διπλωματική εργασία αφορά μία πολύ σοβαρή ασθένεια και συγκριμένα αυτή του σακχαρώδους διαβήτη. Σύμφωνα με επιστημονικά βιβλία και άρθρα, διαπιστώθηκαν οι δυνατές τιμές που μπορεί να πάρει κάθε ένα χαρακτηριστικό, με σκοπό την αφαίρεση των μη αποδεκτών τιμών (Outliers) του συνόλου δεδομένων. Αναλυτικά για κάθε χαρακτηριστικό συμπεράναμε τα εξής:

Συγκέντρωση γλυκόζης πλάσματος (Glucose):	Για το συγκεκριμένο χαρακτηριστικό οι φυσιολογικές τιμές ενός ανθρώπου κυμαίνονται από 101 μέχρι και 140 mg/dl. Οι τιμές από 141 μέχρι και 199 mg/dl υποδεικνύουν ότι ο άνθρωπος έχει μία προδιάθεση εμφάνισης της ασθένειας του σακχαρώδη διαβήτη. Πάνω από 199 mg/dl ο άνθρωπος είναι διαβητικός και πάνω από 500 mg/dl ο άνθρωπος πιθανόν να βρίσκεται σε κώμα. Για τιμές κάτω από 101 mg/dl το άτομο παρουσιάζει αντιδραστική υπογλυκαιμία. Ως κατώτατη τιμή θεωρούμε την τιμή 60 mg/dl. Από τα παραπάνω θεωρούμε ως δυνατές τιμές, τις τιμές που κυμαίνονται από 60 μέχρι και 500 mg/dl. [35]
Διαστολική αρτηριακή πίεση (BloodPressure):	Για το συγκεκριμένο χαρακτηριστικό οι φυσιολογικές τιμές ενός ανθρώπου κυμαίνονται από 50 μέχρι και 80 mmHg. Οι τιμές από 81 μέχρι και 89 mmHg υποδεικνύουν ότι ο άνθρωπος έχει μία προδιάθεση για υπέρταση. Οι τιμές από 90 μέχρι και 99 mmHg υποδεικνύουν ο άνθρωπος έχει διαστολική υπέρταση πρώτου (1ο) βαθμού. Οι τιμές από 100 mmHg μέχρι και 109 mmHg ο άνθρωπος έχει διαστολική υπέρταση δευτέρου (2ο) βαθμού. Πάνω από 110 mmHg ο άνθρωπος βρίσκεται σε κατάσταση υπερτασικής κρίσης. Από τα παραπάνω θεωρούμε ως δυνατές τιμές, τις τιμές που κυμαίνονται από 30 μέχρι και 110 mmHg. [35]

Πάχος της πτυχής του δέρματος (SkinThickness):	Για το συγκεκριμένο χαρακτηριστικό οι φυσιολογικές τιμές των γυναικών κυμαίνονται από 10 μέχρι και 40 mm. Από τα παραπάνω θεωρούμε ως δυνατές τιμές, τις τιμές που κυμαίνονται από 10 μέχρι και 69 mm. [36]
Ινσουλίνη (Insulin):	Για το συγκεκριμένο χαρακτηριστικό οι φυσιολογικές τιμές ενός ανθρώπου κυμαίνονται από 6 mu U / ml μέχρι και 29 mu U / ml. Για τιμές μικρότερες από το 6 mu U / ml υποδεικνύουν ότι ο άνθρωπος έχει διαγνωστεί με σακχαρώδη διαβήτη. Από τα παραπάνω θεωρούμε ως δυνατές τιμές, τις τιμές που κυμαίνονται από 6 μέχρι και 500 mu U / ml. [35]
Δείκτης μάζας σώματος (BMI):	Το συγκεκριμένο χαρακτηριστικό μετριέται με βάση τον τύπο «βάρος σε kg / (ύψος σε m)^2» και οι φυσιολογικές τιμές ενός ανθρώπου κυμαίνονται από 19 μέχρι και 25 kg/m^2. Οι τιμές από 26 μέχρι και 30 kg/m^2 υποδεικνύουν ότι ο άνθρωπος είναι υπέρβαρος. Οι τιμές από 31 μέχρι και 40 kg/m^2 ο άνθρωπος είναι παχύσαρκος. Πάνω από 41kg/m^2 ο άνθρωπος είναι υπερβολικά παχύσαρκος, ενώ κάτω από 19 kg/m^2 ο άνθρωπος είναι υπερβολικά αδύνατος. Από τα παραπάνω θεωρούμε ως δυνατές τιμές, τις τιμές που κυμαίνονται λίγο κάτω από 19 μέχρι και 60 kg/m^2. [35]

Πίνακας 10: Δυνατές τιμές ανά χαρακτηριστικό

3.6.3 Αντιμετώπιση outliers με βάση την ιατρική

Στα χαρακτηριστικά του συνόλου δεδομένων, εκτός από τις μηδενικές τιμές που περιγράφηκαν στην παραπάνω ενότητα, παρατηρήθηκε ότι υπάρχουν και ακραίες τιμές. Οι τιμές που αναφέρονται ως ακραίες τιμές είναι γνωστές με τον όρο «Outliers» και καθορίζονται από τις δυνατές τιμές που μπορεί να πάρει κάθε χαρακτηριστικό βασισμένο στην επιστήμη της ιατρικής. Συχνά οι ακραίες τιμές προκαλούν προβλήματα στην αποτελεσματικότητα των αλγορίθμων της μηχανικής μάθησης. Για τον λόγο αυτόν κρίνεται απαραίτητη η αντιμετώπισή τους. Αναλυτικά, ακραίες τιμές περιέχονται στα εξής χαρακτηριστικά: συγκέντρωση γλυκόζης πλάσματος (Glucose), διαστολική αρτηριακή πίεση (BloodPressure), πάχος της πτυχής του δέρματος (SkinThickness), ινσουλίνη (Insulin) και δείκτης μάζας σώματος (BMI).

Για την αντιμετώπιση των ακραίων τιμών (Outliers) που εμπεριέχονται στα χαρακτηριστικά του συνόλου δεδομένων που αναφέρθηκαν, υπάρχουν δύο (2) περιπτώσεις που μπορούμε να εφαρμόσουμε. Η πρώτη περίπτωση είναι να διατηρηθούν οι ακραίες τιμές ως έχουν, με άλλα λόγια να συμπεριληφθούν στην έρευνα. Η δεύτερη και πιο αποτελεσματική περίπτωση είναι να αφαιρεθούν όλες οι ακραίες τιμές των χαρακτηριστικών σύμφωνα με τη βοήθεια της ιατρικής επιστήμης μιας και οδηγεί σε πιο ασφαλή συμπεράσματα όσον αφορά την πρόβλεψη του σακχαρώδη διαβήτη.

Στη συνέχεια παρουσιάζονται αναλυτικά οι περιπτώσεις αντιμετώπισης των ακραίων τιμών των χαρακτηριστικών του συνόλου δεδομένου.

3.6.3.1 Διατήρηση όλων των ακραίων τιμών

Σε περίπτωση που ένα σύνολο δεδομένων περιέχει πολλές ακραίες τιμές στα χαρακτηριστικά των δεδομένων του και δεν επηρεάζουν την έρευνα, τότε μπορούν να διατηρηθούν ως έχειν. Στη συγκεκριμένη περίπτωση, οι ακραίες τιμές των χαρακτηριστικών είναι αρκετές με αποτέλεσμα να κρίνεται απαραίτητη η αντιμετώπισή τους.

3.6.3.2 Αφαίρεση όλων των ακραίων τιμών

Ένας από τους πιο συνηθισμένους τρόπους αντιμετώπισης των ακραίων τιμών (Outliers) είναι να αφαιρέσουμε όλες τις ακραίες τιμές των χαρακτηριστικών του συνόλου δεδομένων σύμφωνα με το ιατρικό υπόβαθρο που αναφέρθηκε παραπάνω. Ως ακραίες τιμές θωρούνται οι τιμές που είναι μη αποδεκτές για το εκάστοτε χαρακτηριστικό.

Αναλυτικά, σύμφωνα με το ιατρικό υπόβαθρο βρέθηκαν οι εξής ακραίες (μη αποδεκτές) τιμές στα αντίστοιχα χαρακτηριστικά, όπως παρουσιάζονται στον παρακάτω πίνακα.

Χαρακτηριστικά	**Αποδεκτές τιμές**	**Σύνολο μη αποδεκτών εγγραφών**
Συγκέντρωση γλυκόζης πλάσματος (Glucose)	60 – 500 mg/dl	4
Διαστολική αρτηριακή πίεση (BloodPressure)	30 – 110 mmHg	3
Πάχος της πτυχής του δέρματος (SkinThickness)	10 - 69 mm	5
Ινσουλίνη (Insulin)	6 – 500 mu U / ml	9
Δείκτης μάζας σώματος (BMI)	19 – 60 kg/m^2	5

Πίνακας 11: Πλήθος μη αποδεκτών τιμών ανά χαρακτηριστικό

Με βάση τον παραπάνω πίνακα, αφαιρέθηκαν οι 26 τιμές που ήταν εκτός του ορίου των αποδεκτών τιμών. Με την εφαρμογή της αφαίρεσης όλων των ακραίων τιμών του συνόλου για τα συγκεκριμένα χαρακτηριστικά, το σύνολο δεδομένων καταλήγει να περιέχει 742 εγγραφές.

Απομένει λοιπόν να υπολογιστεί η συσχέτιση που έχουν τα χαρακτηριστικά ως προς τον επηρεασμό του αποτελέσματος.

3.7 Εύρεση καλύτερων χαρακτηριστικών

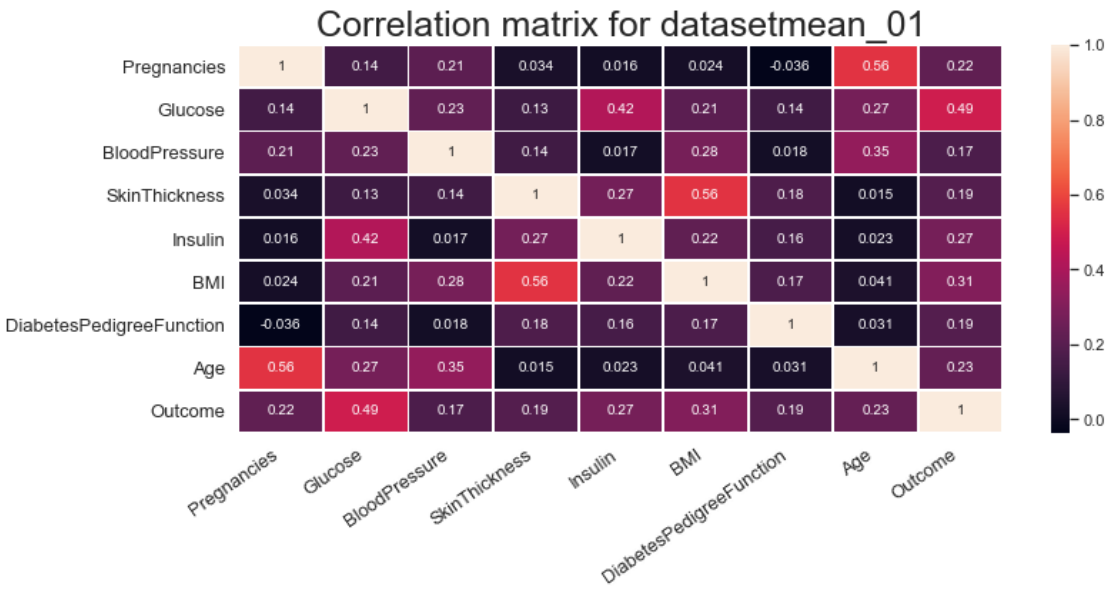
Μετά την ολοκλήρωση της επεξεργασίας των δεδομένων, πρέπει να εξεταστεί η αλληλεπίδραση των χαρακτηριστικών κυρίως ως προς το αποτέλεσμα (Outcome). Με άλλα λόγια, πρέπει να διαπιστωθεί πόσο επηρεάζει κάθε χαρακτηριστικό το αποτέλεσμα (Outcome), αν δηλαδή κάποιος έχει ή δεν έχει σακχαρώδη διαβήτη.

Με αυτήν τη διαδικασία στόχος είναι να επιλεχθούν τα καλύτερα χαρακτηριστικά. Για την επίτευξη του στόχου αυτού χρησιμοποιήθηκε η μέθοδος της μήτρας συσχέτισης (correlation matrix), στην οποία γίνεται πλήρης αναφορά παρακάτω, καθώς και πληροφορίες από το ιατρικό υπόβαθρο.

3.7.1 Μήτρα συσχέτισης (Correlation matrix)

Ένας πίνακας συσχετισμού είναι ένας πίνακας που εμφανίζει συντελεστές συσχέτισης μεταξύ των διαφόρων χαρακτηριστικών. Κάθε χαρακτηριστικό στον πίνακα συσχετίζεται με κάθε ένα άλλο χαρακτηριστικό που επίσης βρίσκεται στον πίνακα. Η διαγώνια του πίνακα είναι πάντοτε ένα σύνολο από αυτές, επειδή η συσχέτιση μεταξύ μιας μεταβλητής και της ίδιας είναι πάντα 1. Οι τιμές κυμαίνονται από -1 έως 1 και όσο πιο κοντά η τιμή είναι στο 1, τόσο μεγαλύτερη είναι η σχέση (συσχέτισης) μεταξύ δύο χαρακτηριστικών.[37]

Εφαρμόζοντας την παραπάνω μέθοδο διαπιστώνεται ποια από τα χαρακτηριστικά επηρεάζουν περισσότερο το εν λόγω αποτέλεσμα. Τα αποτελέσματα παρουσιάζονται στην παρακάτω εικόνα με χρήση γραφικής παράστασης (Εικόνα 16).



Εικόνα 16: Απεικόνιση των συσχετίσεων μεταξύ των χαρακτηριστικών και του αποτελέσματος

Με βάση τα παραπάνω αποτελέσματα, είναι σαφές ότι τα χαρακτηριστικά που επηρεάζουν περισσότερο το αποτέλεσμα (Outcome), σε σειρά από το σημαντικότερο (1), στο λιγότερο σημαντικό (8), είναι τα εξής:

1. Glucose
2. BMI
3. Insulin
4. Age
5. Pregnancies
6. DiabetesPedigreeFunction
7. SkinThickness
8. BloodPressure

Πίνακας 12: Απεικονίζονται τα χαρακτηριστικά με σειρά σημαντικότητας

3.7.2 Ιατρικό υπόβαθρο καλύτερων χαρακτηριστικών

Σύμφωνα με το ιατρικό υπόβαθρο ο βαθμός επηρεασμού κάθε χαρακτηριστικού προς το αποτέλεσμα (Outcome), δηλαδή του αν κάποιος έχει ή όχι σακχαρώδη διαβήτη, ακολουθεί την παρακάτω σειρά προσεγγιστικά, από τον ισχυρότερο επηρεασμό προς τον λιγότερο.

1. Glucose
2. Insulin
3. Pregnancies
4. DiabetesPedigreeFunction
5. BMI
6. Age
7. BloodPressure
8. SkinThickness

Πίνακας 13: Ιατρικό υπόβαθρο σειρά σημαντικότητας των χαρακτηριστικών

3.7.3 Χαρακτηριστικά που επιλέγονται

Η σύγκριση του πίνακα της μήτρας συσχέτισης (Πίνακας 8) και το ιατρικό υπόβαθρο καλύτερων χαρακτηριστικών (Πίνακας 9) υποδεικνύει ότι υπάρχουν δύο χαρακτηριστικά που επηρεάζουν λιγότερο το αποτέλεσμα (διαστολική αρτηριακή πίεση, BloodPressure και πάχος της πτυχής του δέρματος, SkinThickness), τα οποία και θα αφαιρεθούν από το τελικό σύνολο δεδομένων.

3.8 Τελικό σύνολο δεδομένων

Με την εφαρμογή των παραπάνω διεργασιών δημιουργήθηκε το τελικό σύνολο δεδομένων που θα χρησιμοποιηθεί από τους αλγόριθμους μηχανικής μάθησης για την πρόβλεψη του σακχαρώδη διαβήτη. Πιο συγκεκριμένα, το τελικό σύνολο δεδομένων περιέχει 742 εγγραφές, από τις οποίες οι 258 αφορούν άτομα που έχουν διαγνωστεί με την ασθένεια του σακχαρώδη διαβήτη, ενώ οι υπόλοιπες 484 εγγραφές αφορούν υγιή άτομα. Επίσης, το τελικό σύνολο αποτελείται από εφτά (7) στήλες. Η τελευταία στήλη υποδηλώνει το αποτέλεσμα (Outcome), αν

δηλαδή το άτομο έχει σακχαρώδη διαβήτη ή όχι, ενώ οι υπόλοιπες έξη (6) στήλες αποτελούν τα χαρακτηριστικά του ατόμου. Τα χαρακτηριστικά που διατηρήθηκαν στην έρευνα είναι τα εξής: ο συνολικός αριθμός που έχει μείνει έγκυος (Pregnancies) κάθε άτομο, η συγκέντρωση γλυκόζης πλάσματος (Glucose), η ινσουλίνη (Insulin), ο δείκτης μάζας σώματος (BMI), η λειτουργία γενεαλογικού διαβήτη (DiabetesPedigreeFunction) και η ηλικία του ατόμου (Age).

ΚΕΦΑΛΑΙΟ 4 – ΑΝΑΛΥΣΗ ΚΑΙ ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ

Πριν από την εφαρμογή των αλγορίθμων μηχανικής μάθησης απαιτείται ένα τελευταίο βήμα που σχετίζεται με το διαχωρισμό του τελικού συνόλου δεδομένων σε δύο (2) υποσύνολα δεδομένων, ένα υποσύνολο που θα αποτελείται από τα χαρακτηριστικά (features) και ένα που θα αποτελείται από το αποτέλεσμα (labels).

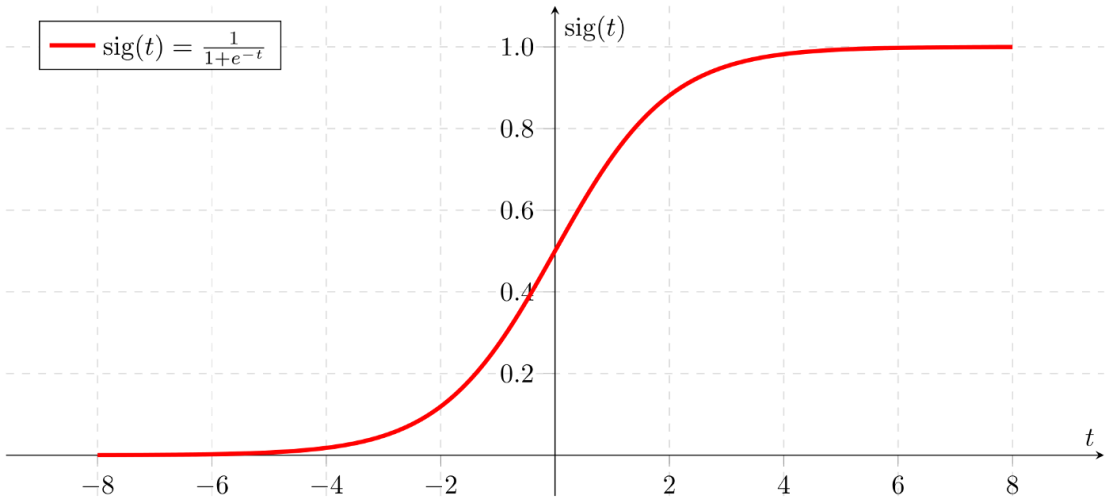
Πιο αναλυτικά, το πρώτο υποσύνολο δεδομένων θα περιέχει τις πρώτες έξι (6) στήλες του τελικού συνόλου δεδομένων που δημιουργήθηκε πιο πάνω και είναι οι εξής: ο συνολικός αριθμός που έχει μείνει έγκυος (Pregnancies), η συγκέντρωση γλυκόζης πλάσματος (Glucose), η ινσουλίνη (Insulin), ο δείκτης μάζας σώματος (BMI), η λειτουργία γενεαλογικού διαβήτη (DiabetesPedigreeFunction) και η ηλικία του ασθενή (Age), ενώ το δεύτερο σύνολο δεδομένων θα περιέχει την τελευταία στήλη με το αποτέλεσμα κάθε εγγραφής (Outcome).

Αφού έχουν δημιουργηθεί τα δύο υποσύνολα γίνεται η εφαρμογή των αλγορίθμων μηχανικής μάθησης που έχουν επιλεχθεί όπως αναφέρθηκαν στο κεφάλαιο 3. Στο κεφάλαιο αυτό θα αναλυθούν και θα εφαρμοστούν λεπτομερώς οι αλγόριθμοι αυτοί.

4.1 Λογιστική παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση (Logistic Regression), γνωστή και ως Logit Regression ή Logit Model, είναι ένα στατιστικό μοντέλο που στη βασική της μορφή χρησιμοποιεί μια λογική λειτουργία (logistic function), για να μοντελοποιήσει μια δυαδική εξαρτώμενη μεταβλητή. Η λογιστική παλινδρόμηση περιγράφει και εκτιμά τη σχέση μεταξύ μιας εξαρτώμενης δυαδικής μεταβλητής και ανεξάρτητων μεταβλητών.

Η λογική λειτουργεία, που ονομάζεται επίσης και η σιγμοειδής λειτουργία (sigmoid function), αναπτύχθηκε από τους στατιστικολόγους για να περιγράψει τις ιδιότητες της πληθυσμιακής ανάπτυξης στην οικολογία, να αυξηθεί γρήγορα και να μεγιστοποιηθεί η φέρουσα ικανότητα του περιβάλλοντος. Είναι μια καμπύλη σχήματος «S» που μπορεί να πάρει οποιοδήποτε πραγματικό αριθμό και να το χαρτογραφήσει σε τιμή μεταξύ 0 και 1, αλλά ποτέ δεν είναι ακριβώς στα όρια αυτά (Εικόνα 17).



Εικόνα 17: Γράφημα σιγμοειδής λειτουργείας

Πιο αναλυτικά, ο αλγόριθμος Logistic Regression είναι ένας από τους πιο απλούς και συνηθισμένους αλγόριθμους Machine Learning για ταξινόμηση δύο κατηγοριών. Οι τιμές εισόδου (x) συνδυάζονται γραμμικά χρησιμοποιώντας είτε βάρη, είτε τιμές συντελεστών (που αναφέρονται ως το ελληνικό κεφαλαίο γράμμα Βήτα) για την πρόβλεψη μιας τιμής εξόδου (y). Μία διαφορά κλειδιού από τη γραμμική παλινδρόμηση είναι ότι η τιμή εξόδου που διαμορφώνεται είναι δυαδικές τιμές (0 ή 1) και όχι αριθμητική τιμή. Έτσι δίνεται ένα χαρακτηριστικό x προσπαθεί να μάθει αν συμβαίνει κάποιο συμβάν y ή όχι. Επομένως y μπορεί να είναι 0 ή 1. Στην περίπτωση όπου συμβαίνει το συμβάν, y δίνεται η τιμή 1. Εάν το συμβάν δεν συμβεί, τότε το y δίνεται στην τιμή 0.[38]

Στη περίπτωση της μεταπτυχιακής διπλωματικής εργασίας εξετάζεται αν ένας άνθρωπος με συγκεκριμένα χαρακτηριστικά μπορεί να έχει διαβήτη (1), ή είναι

υγιής (0). Τα χαρακτηριστικά με τα οποία θα κριθεί το αποτέλεσμα του μοντέλου είναι τα εξής: Pregnancies (εγκυμοσύνες), Glucose (συγκέντρωση γλυκόζης πλάσματος), BloodPressure (διαστολική αρτηριακή πίεση), SkinThickness (πάχος της πτυχής του δέρματος), Insulin (ινσουλίνη), BMI (δείκτης μάζας σώματος), DiabetesPedigreeFunction (λειτουργία γενεαλογικού διαβήτη) και Age (ηλικία). Τα χαρακτηριστικά θα αναλυθούν στο επόμενο κεφάλαιο λεπτομερώς.

4.2 Δέντρο απόφασης (Decision Tree Classifier)

Ο αλγόριθμος δένδρων απόφασης (Decision Tree Classifier) ανήκει στην οικογένεια αλγορίθμων μάθησης υπό επίβλεψη. Σε αντίθεση με άλλους αλγόριθμους εποπτευόμενης μάθησης, ο αλγόριθμος δέντρων αποφάσεων μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων παλινδρόμησης και ταξινόμησης.

Τα δέντρα απόφασης είναι μία από τις μεθόδους πρόβλεψης που χρησιμοποιούνται στη στατιστική, την εξόρυξη δεδομένων και την εκμάθηση μηχανών. Η μάθηση δέντρων αποφάσεων χρησιμοποιείται συνήθως στην εξόρυξη δεδομένων. Ο στόχος της χρήσης ενός δέντρου αποφάσεων είναι να δημιουργηθεί ένα μοντέλο εκπαίδευσης που μπορεί να χρησιμοποιηθεί για την πρόβλεψη της τάξης ή της αξίας της μεταβλητής στόχου με την εκμάθηση απλών κανόνων αποφάσεων που προκύπτουν από προηγούμενα δεδομένα (δεδομένα εκπαίδευσης). Πιο συγκεκριμένα χρησιμοποιείται ως μοντέλο πρόβλεψης με σκοπό να μεταβεί από παρατηρήσεις ενός στοιχείου (που αντιπροσωπεύεται στους κλάδους) σε συμπεράσματα, σχετικά με την τιμή στόχου του αντικειμένου (που αναπαρίσταται στα φύλλα). Τα μοντέλα δένδρων όπου η μεταβλητή στόχος μπορεί να πάρει ένα διακριτό σύνολο τιμών ονομάζονται δέντρα ταξινόμησης, σε αυτές τις δομές δέντρων, τα φύλλα αντιπροσωπεύουν ετικέτες τάξεων και κλάδων, αντιπροσωπεύοντας συζεύξεις χαρακτηριστικών που οδηγούν σε αυτές τις ετικέτες τάξης.

Αναλυτικότερα, στα δέντρα απόφασης για την πρόβλεψη μιας ετικέτας κλάσης για ένα αρχείο ξεκινάμε από τη ρίζα του δέντρου. Συγκρίνουμε τις τιμές της

ιδιότητας ρίζας με το χαρακτηριστικό της εγγραφής. Με βάση τη σύγκριση, ακολουθούμε τον κλάδο που αντιστοιχεί στην τιμή και μεταβούμε στον επόμενο κόμβο. Έτσι, τα δέντρα αποφάσεων σχηματίζουν ένα δέντρο με ιεραρχικό τρόπο, με κάθε κόμβο να έχει μια απόφαση όριο να προχωρήσει προς τα κάτω. Το δέντρο σταματά να διακλαδώνεται στο επίπεδο όπου βρίσκεται ότι δεν υπάρχουν πλέον διαχωρίσεις. Οι εσωτερικοί κόμβοι αντιπροσωπεύουν μεταβλητές εισόδου με ακμές σε κάθε ένα από τα παιδιά. Τα παιδιά χωρίζουν τις τιμές από τη μεταβλητή εισόδου. Το κάνουν αυτό χωρίζοντας τα δεδομένα σε κάθε επίπεδο με κόμβους που διανέμονται στα παιδιά. Αυτή η συμπεριφορά είναι γνωστή ως επαναληπτικός διαχωρισμός. Τα δέντρα αποφάσεων είναι εύκολο να ερμηνευτούν και να λειτουργούν αποτελεσματικά, και ως εκ τούτου μπορούν να λειτουργήσουν καλά με μεγάλα σύνολα δεδομένων. Τα δέντρα αποφάσεων μπορούν επίσης να χειριστούν και τα δύο αριθμητικά και κατηγορηματικά δεδομένα, δηλαδή, παλινδρόμηση σε περίπτωση αριθμητικής και ταξινομημένης σε περίπτωση κατηγορηματικών δεδομένων. Ωστόσο, η ακρίβεια των δέντρων αποφάσεων δεν είναι τόσο καλή όσο αυτή που παράγονται από άλλους αλγόριθμους ταξινόμησης μηχανικής μάθησης. Επιπλέον, τα δέντρα απόφασης γενικεύουν ιδιαίτερα στο σύνολο δεδομένων κατάρτισης και έτσι είναι ιδιαίτερα ευαίσθητα στην υπερφόρτωση. Το δέντρο αποφάσεων στοχεύει στη διαίρεση των δεδομένων έτσι ώστε κάθε ένα των διαμοιρασμένων περιπτώσεων έχει παρόμοιες / ομοιογενείς τιμές. Οι αλγόριθμοι ID3 είναι που χρησιμοποιείται για τον υπολογισμό της ομοιογένειας ενός δείγματος και αν είναι εντελώς ομοιογενής μεταφράζεται σε μια εντροπία 0 και σε τιμή 1 ή αντίστροφα. Ένα δέντρο απόφασης είναι μια μορφή μιας παραμετρικής εποπτευόμενης μεθόδου εκμάθησης, με παραμετρικό τρόπο εννοούμε ότι μπορεί να εφαρμοστεί σε οποιαδήποτε δεδομένα ανεξάρτητα από την υποκείμενη διανομή τους.[39]

4.3 Τυχαία δάση (Random Forest Classifier)

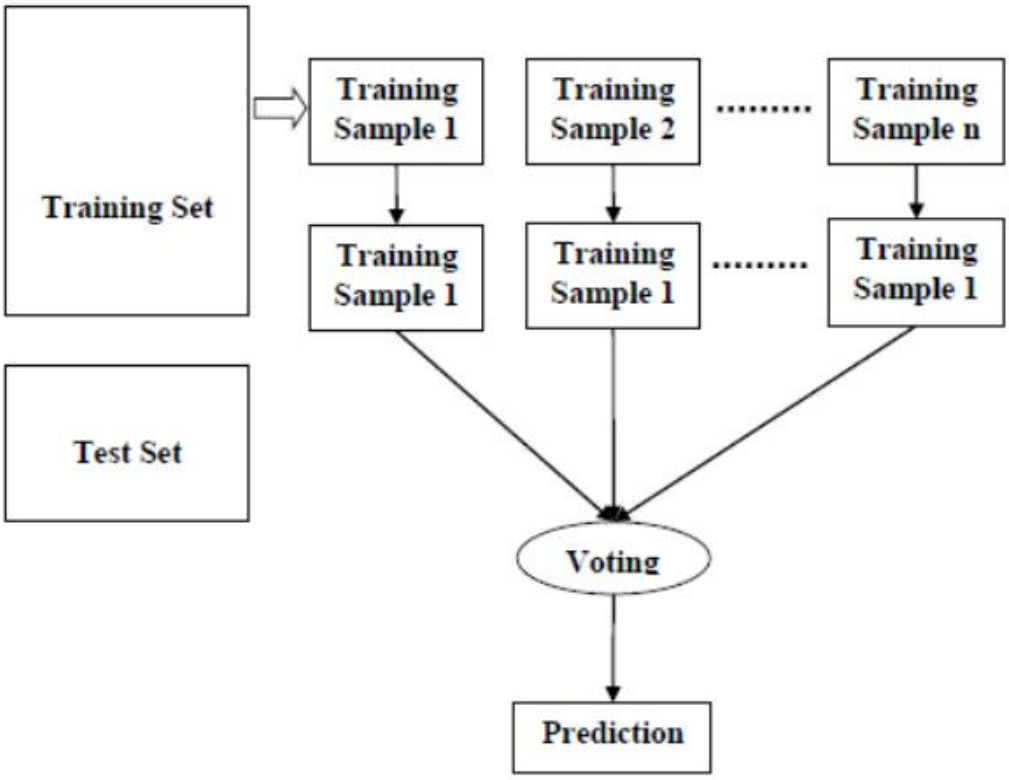
Τα τυχαία δάση ή τα τυχαία δάση αποφάσεων (Random Forest Classifier) είναι ένας αλγόριθμος εποπτευόμενης μάθησης που χρησιμοποιείται τόσο για ταξινόμηση, όσο και για παλινδρόμηση. Τις περισσότερες φορές χρησιμοποιείται για τον πρώτο λόγο που αναφέρθηκε, δηλαδή την ταξινόμηση. Όπως μπορεί να διαπιστωθεί και από την ονομασία του συγκριμένου αλγορίθμου αποτελείται από πολλά επιμέρους δέντρα, τα οποία δημιουργούν ένα δάσος και συνεπώς όσα περισσότερα δέντρα, τόσο πιο ισχυρό δάσος.

Με τον ίδιο ακριβώς τρόπο που λειτουργεί το δέντρο απόφασης (Decision Tree), έτσι και ο αλγόριθμος του τυχαίου δάσους δημιουργεί δέντρα αποφάσεων από τα δείγματα δεδομένων και στη συνέχεια παίρνει την πρόβλεψη από καθένα από αυτά και τελικά επιλέγει την καλύτερη λύση μέσω ψηφοφορίας.

Ο κυριότερος λόγος που προτιμάτε ο αλγόριθμος του τυχαίου δάσους από του δέντρου απόφασης είναι ότι μειώνεται αισθητά η υπερεκπαίδευση (overfitting) του μοντέλου.

Ο αλγόριθμος τυχαίου δάσους (Random Forest) λειτουργεί με τον εξής τρόπο. Αρχικά, επιλέγει τυχαία δείγματα από το σύνολο δεδομένων. Στη συνέχεια, ο αλγόριθμος θα κατασκευάσει ένα δέντρο απόφασης για κάθε ένα δείγμα σχηματίζοντας με αυτόν τον τρόπο το δάσος (Forest). Από κάθε ένα δέντρο απόφασης παίρνει το αποτέλεσμα πρόβλεψής του με τη διαδικασία της ψηφοφορία για κάθε ένα από τα δέντρα απόφασης. Τέλος, επιλέγεται ως αποτέλεσμα πρόβλεψης του δάσους εκείνο με την υψηλότερη ψηφοφορία.[40]

Η διαδικασία λειτουργίας του τυχαίου δάσους (Random Forest) αναπαρίσταται και στην ακόλουθη εικόνα (Εικόνα 18).



Εικόνα 18: Τρόπος λειτουργίας του τυχαίου δάσους (Random Forest)

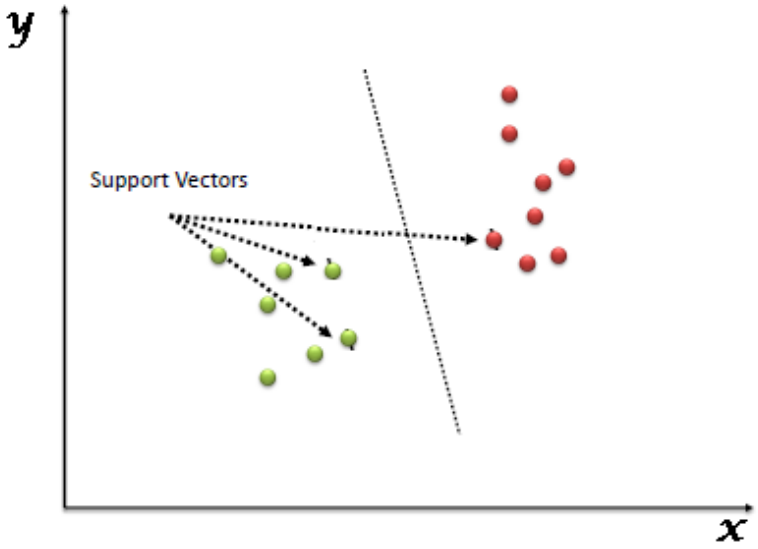
4.4 Μηχανές διανυσμάτων υποστήριξης (SVM)

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines, SVM) είναι ένας εποπτευόμενος αλγόριθμος μηχανικής μάθησης με συναφείς αλγόριθμους εκμάθησης που αναλύουν δεδομένα που χρησιμοποιούνται τόσο για ανάλυση ταξινόμησης όσο και για προκλήσεις παλινδρόμησης. Ωστόσο, χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης.

Λαμβάνοντας υπόψη ένα σύνολο εκπαιδευτικών παραδειγμάτων, κάθε ένα από τα οποία χαρακτηρίζεται ότι ανήκει είτε στη μία κατηγορία, είτε στην άλλη κατηγορία από τις συνολικά δύο κατηγορίες, ένας αλγόριθμος κατάρτισης SVM δημιουργεί ένα μοντέλο που εκχωρεί νέα παραδείγματα είτε στη μία κατηγορία,

είτε στην άλλη κατηγορία, καθιστώντας τον έναν μη πιθανοτικό δυαδικό γραμμικό ταξινομητή.

Πιο αναλυτικά, ο συγκεκριμένος αλγόριθμος, σχεδιάζει κάθε στοιχείο δεδομένων ως σημείο σε n-διάστατο χώρο (όπου n είναι ο αριθμός των χαρακτηριστικών) με την τιμή κάθε χαρακτηριστικού να είναι η τιμή μιας συγκεκριμένης συντεταγμένης. Στη συνέχεια, πραγματοποιεί ταξινόμηση βρίσκοντας το υπερ-επίπεδο (hyper-plane) που διαφοροποιεί τις δύο κατηγορίες πολύ καλά. Ένα μοντέλο SVM είναι μια αναπαράσταση των παραδειγμάτων ως σημεία στο διάστημα, χαρτογραφημένα έτσι ώστε τα παραδείγματα των ξεχωριστών κατηγοριών να χωρίζονται από ένα σαφές κενό που είναι όσο το δυνατόν ευρύτερο. Με άλλα λόγια οι μηχανές διανυσμάτων υποστήριξης (SVM) είναι ένα όριο που διαχωρίζει καλύτερα τις δύο κλάσεις (υπερ-επίπεδο / γραμμή), όπως παρουσιάζεται παρακάτω (Εικόνα 19).[41]



Εικόνα 19: Στιγμιότυπο της εφαρμογής του αλγορίθμου SVM

4.5 Κ-πλησιέστερος γείτονας (K-Nearest Neighbors KNN)

Ο Κ-πλησιέστερος γείτονας (K-Nearest Neighbor), γνωστός ως KNN, είναι ένας αλγόριθμος με τον οποίο γίνεται η προσέγγιση ταξινόμησης δεδομένων εκτιμώντας πόσο πιθανό είναι ένα σημείο δεδομένων να είναι μέλος της μίας ομάδας ή της άλλης, ανάλογα με την ομάδα στην οποία βρίσκονται τα πλησιέστερα σημεία δεδομένων.

Πιο αναλυτικά, ο αλγόριθμος KNN προσπαθεί να προσδιορίσει σε ποια ομάδα βρίσκεται ένα σημείο δεδομένων κοιτάζοντας τα σημεία δεδομένων γύρω από αυτό. Κοιτάζοντας ένα σημείο σε ένα πλέγμα, προσπαθεί να προσδιορίσει εάν ένα σημείο βρίσκεται στην ομάδα Α ή Β, κοιτάζοντας τις καταστάσεις των σημείων που βρίσκονται κοντά του. Το εύρος καθορίζεται αυθαίρετα, αλλά το θέμα είναι να ληφθεί ένα δείγμα των δεδομένων. Εάν η πλειοψηφία των σημείων ανήκει στην ομάδα Α, τότε είναι πιθανό το εν λόγω σημείο δεδομένων να είναι Α, αντί για Β και αντίστροφα.

Τέλος, ο αλγόριθμος KNN είναι ένα παράδειγμα αλγορίθμου "τεμπέλη μαθητή" ("lazy learner"), επειδή δεν δημιουργεί εκ των προτέρων ένα μοντέλο του συνόλου δεδομένων. Οι μόνοι υπολογισμοί που πραγματοποιούνται είναι όταν τους ζητηθεί να πραγματοποιήσουν δημοσκοπήσεις στους γείτονες του σημείου δεδομένων. Αυτό καθιστά το KNN πολύ εύκολο στην εφαρμογή για την εξόρυξη δεδομένων.[42]

4.6 Χρήση αλγορίθμων μηχανικής μάθησης

Με την ολοκλήρωση της ανάλυσης του συνόλου δεδομένων Pima Indians Diabetes Database, καθώς και με την κατάλληλη επεξεργασία τους, και με το διαχωρισμό του συνόλου δεδομένου γίνεται η χρήση των αλγορίθμων της μηχανικής μάθησης. Παρακάτω παρουσιάζεται αναλυτικότερα η εφαρμογή των αλγορίθμων μηχανικής μάθησης. Οι αλγόριθμοι μηχανικής μάθησης που θα

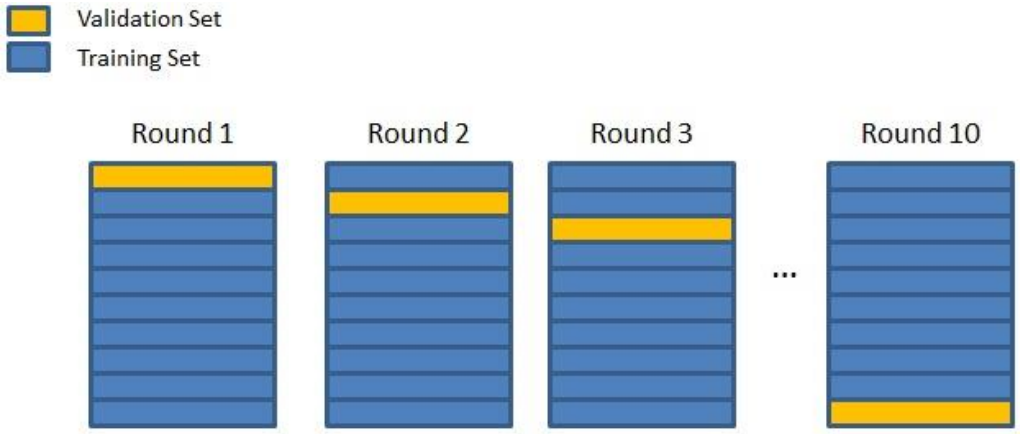
χρησιμοποιηθούν για την πρόβλεψη του σακχαρώδη διαβήτη, όπως ήδη έχει αναφερθεί σε προηγούμενο κεφάλαιο είναι οι εξής: λογιστική παλινδρόμηση (Logistic Regression), δέντρων αποφάσεων (DecisionTreeClassifier), τυχαία δάση αποφάσεων (Random forest), μηχανές διανυσμάτων υποστήριξης (SVM) και ο αλγόριθμος Κ-πλησιέστερων γειτόνων (K Nearest Neighbors, KNN). Για την εκπαίδευση των αλγορίθμων χρησιμοποιήθηκε η μέθοδος του Cross validation όπως περιγράφεται στη συνέχεια.

4.6.1 Cross validation

Για κάθε έναν από τους αλγόριθμους χρησιμοποιήθηκε η μέθοδος του cross validation κατά την οποία τα δεδομένα χωρίζονται σε επιμέρους μικρότερα σύνολα. Με τη μέθοδο αυτή επιτυγχάνεται η αποφυγή της υπερ-εκπαίδευσης του μοντέλου.

Κατά τη μέθοδο αυτή τα δεδομένα χωρίζονται με τέτοιο τρόπο, ώστε να παρέχονται άφθονα δεδομένα όχι μόνο για την εκπαίδευση του μοντέλου, αλλά και για την επικύρωσή του. Το σύνολο των δεδομένων χωρίζεται σε k υποσύνολα. Η τιμή του k κυμαίνεται από πέντε (5) έως δέκα (10) συνήθως, ανάλογα με το εκάστοτε σύνολο δεδομένων.

Η συγκεκριμένη διαδικασία επαναλαμβάνεται k φορές, έτσι ώστε κάθε φορά, ένα από τα υποσύνολα k να χρησιμοποιείται ως το σετ δοκιμής / επικύρωσης και τα άλλα υπο-σύνολα k-1 να συντίθενται για να σχηματίσουν ένα σύνολο εκπαίδευσης. Όπως μπορεί να φανεί και από την εικόνα παρακάτω (Εικόνα 20), κάθε σημείο δεδομένων χρησιμοποιείται σε μια επικύρωση που ορίζεται ακριβώς μία φορά, και χρησιμοποιείται σε ένα σετ κατάρτισης k-1 φορές, με αποτέλεσμα να μειώνεται σημαντικά η μεροληψία του μοντέλου, καθώς χρησιμοποιούνται τα περισσότερα δεδομένα για την τοποθέτηση. Παράλληλα όμως μειώνεται σημαντικά η διακύμανση, καθώς τα περισσότερα από τα δεδομένα χρησιμοποιούνται επίσης στο σετ επικύρωσης. Η εκτίμηση σφάλματος υπολογίζεται κατά μέσο όρο σε όλες τις δοκιμές k για να επιτευχθεί η συνολική αποτελεσματικότητα του μοντέλου.[43]



Εικόνα 20: Cross validation

Για την υλοποίηση της διπλωματικής εργασίας, το k ορίστηκε με τιμή ίση με δέκα (10). Με άλλα λόγια το τελικό σύνολο δεδομένων θα χωριστεί σε δέκα (10) υποσύνολα και ο κάθε αλγόριθμος θα επαναληφθεί δέκα (10) φορές.

4.6.2 Στατιστικές μετρήσεις

Για την καλύτερη κατανόηση της αποτελεσματικότητας των αλγορίθμων μηχανικής μάθησης γίνεται χρήση του υπολογισμού στατιστικών μετρήσεων. Για κάθε έναν αλγόριθμο μηχανικής μάθησης τόσο χωρίς τη χρήση παραμέτρων, όσο και με τη χρήση παραμέτρων υπολογίζονται τα εξής: η μέση τυπική απόκλιση, η μέση εκτίμηση σφάλματος και το μέσο ποσοστό ακρίβειας της αποτελεσματικότητάς του. Οι παραπάνω μετρήσεις συμπεριλαμβάνουν τα αποτελέσματα και των k (10) υπο-φακέλων με τη χρήση της μεθόδου cross validation, από την οποία προκύπτουν και οι μέσοι όροι αποτελεσμάτων. Επιπλέον για κάθε έναν υπό-φάκελο ξεχωριστά έχει υπολογιστεί το ποσοστό ακρίβειας εκμάθησης του αλγορίθμου, καθώς και το ποσοστό ακρίβειας της αποτελεσματικότητάς του.

Τα αποτελέσματα που υπολογίζονται απεικονίζονται ενδεικτικά στο παρακάτω στιγμιότυπο που ακολουθεί (Εικόνα 21).

The Results per each 10 folds of DecisionTreeClassifier Customized are:

	Algorithm	Parameter	Repeats	Training Accuracy	Test Accuracy
1	DecisionTreeClassifier	Customized	1	89.66	84.00
2	DecisionTreeClassifier	Customized	2	91.75	88.00
3	DecisionTreeClassifier	Customized	3	90.87	85.14
4	DecisionTreeClassifier	Customized	4	91.02	83.78
5	DecisionTreeClassifier	Customized	5	90.42	89.19
6	DecisionTreeClassifier	Customized	6	89.97	86.49
7	DecisionTreeClassifier	Customized	7	89.67	90.54
8	DecisionTreeClassifier	Customized	8	90.87	89.19
9	DecisionTreeClassifier	Customized	9	90.12	90.54
10	DecisionTreeClassifier	Customized	10	90.72	87.84

The Average of Standard Deviation for DecisionTreeClassifier Customized is: 3.0%

The Average of Validation Error for DecisionTreeClassifier Customized is: 13.0%

The Average of Test Accuracy for DecisionTreeClassifier Customized is: 87.0%

Εικόνα 21: Ενδεικτικά αποτελέσματα του αλγόριθμο DecisionTreeClassifier με χρήση παραμέτρων

Αρχικά, υπολογίζονται και παρουσιάζονται τα ποσοστά τόσο εκπαίδευσης και όσο και επαλήθευσης για όλα τα υποσύνολα του k cross validation. Επίσης, περιέχονται πληροφορίες σχετικά με τον αλγόριθμο, με τις παραμέτρους και με το υποσύνολο που χρησιμοποιείτε κάθε φορά.

Για τον υπολογισμό της τυπικής απόκλισης χρησιμοποιήθηκε μία συνάρτηση, που βρίσκεται μέσα στη βιβλιοθήκη statistics, την stdev(). Η συνάρτηση stdev() υπολογίζει μόνο την τυπική απόκλιση από ένα δείγμα δεδομένων παρά από ολόκληρο τον πληθυσμό. Η τυπική απόκλιση είναι ένα μέτρο διάδοσης στα στατιστικά στοιχεία που χρησιμοποιείται για την ποσοτικοποίηση του μέτρου της εξάπλωσης, της μεταβολής ενός συνόλου τιμών δεδομένων. Είναι πολύ παρόμοια με τη διακύμανση, δίνει το μέτρο απόκλισης, ενώ η διακύμανση παρέχει την τετραγωνική τιμή. Ένα χαμηλό μέτρο τυπικής απόκλισης υποδεικνύει ότι τα δεδομένα είναι λιγότερο διαδεδομένα, ενώ μια υψηλή τιμή της τυπικής απόκλισης δείχνει ότι τα δεδομένα σε ένα σύνολο διαδίδονται εκτός από τις μέσες τιμές τους. Μια χρήσιμη ιδιότητα της τυπικής απόκλισης είναι ότι σε αντίθεση με τη διακύμανση, εκφράζεται στις ίδιες μονάδες με τα δεδομένα.[44]

Για τον υπολογισμό της εκτίμησης σφάλματος υπολογίζεται αρχικά το ποσοστό αποτελεσματικότητας για όλες τις δοκιμές k και στην συνέχεια υπολογίζεται ο

μέσος όρος όλων των ποσοστών ακριβείας της αποτελεσματικότητας του αλγορίθμου που έχουν υπολογιστεί με τη συνάρτηση Average() που έχει δημιουργηθεί και τέλος ο μέσος όρος αφαιρείται από την μονάδα (1) και στρογγυλοποιείται.

Τέλος, για τον υπολογισμό του μέσου ποσοστού ακρίβειας της αποτελεσματικότητας του αλγορίθμου υπολογίζεται ο μέσος όρος ακρίβειας της αποτελεσματικότητας του αλγόριθμου για όλα τα υποσύνολα του k cross validation με χρήση της συνάρτησης Average().

4.6.3 Ανάλυση παραμέτρων των αλγορίθμων

Παρακάτω θα αναλυθούν οι παράμετροι που χρησιμοποιήθηκαν σε κάθε έναν από τους αλγόριθμους μηχανικής μάθησης για την βελτίωση της αποτελεσματικότητας των αλγορίθμων. Οι αλγόριθμοι στην πρώτη εφαρμογή τους χρησιμοποιούνται χωρίς να μεταβάλλεται κάποια από τις παραμέτρους τους είναι δηλαδή ήδη καθορισμένοι (default). Για την υλοποίηση της εργασίας χρησιμοποιήθηκε η βιβλιοθήκη scikit-learn και συγκεκριμένα η έκδοση: 0.22.2.

4.6.3.1 Παράμετροι λογιστικής παλινδρόμησης (Logistic Regression)

Για την προσπάθεια βελτίωσης της αποτελεσματικότητας του αλγορίθμου της λογιστικής παλινδρόμησης (Logistic Regression) χρειάστηκε να μεταβληθεί η πατάμετρος max_iter. Ο αλγόριθμος με την χρήση default παραμέτρων πετυχαίνει μέσο ποσοστό ακρίβειας ίσο με 77%, με random_state ίσο με 16.

Η παράμετρος class_weight είναι ορισμένη εξαρχής στην τιμή «None». Χρειάστηκε να αλλαχθεί η τιμή σε «balanced». Η "ισορροπημένη" (balanced) λειτουργία χρησιμοποιεί τις τιμές του y για την αυτόματη προσαρμογή των βαρών αντιστρόφως ανάλογων με τις συχνότητες κλάσης στα δεδομένα εισόδου ως n_samples / (n_classes * np.bincount (y)). Η παράμετρος

class_weight μπορεί να πάρει εκτός τις τιμές που αναφέρθηκαν «None» και «balanced» και την τιμή «dict».

Η παράμετρος max_iter δηλώνει το μέγιστο αριθμό επαναλήψεων για τη σύγκλιση των επιλυτών.[45]

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των δοκιμών που χρησιμοποιήθηκαν για τις παραπάνω παραμέτρους με βάση την μέση ακρίβεια αποτελεσματικότητας του αλγορίθμου.

Παράμετροι		**Αποτέλεσμα**
class_weight	**max_iter**	**Test Accuracy (%)**
None	92	77
balanced	107	76
dict	92	77

Πίνακας 14: Αποτελέσματα παραμέτρων λογιστικής παλινδρόμησης

Από τα παραπάνω αποτελέσματα συμπεραίνει κανείς ότι η καλύτερη αποτελεσματικότητα του αλγορίθμου επιτυγχάνεται όταν οι παράμετροι έχουν τις προκαθορισμένες τιμές (default), για τις οποίες το class_weight ισούται με «None» και το max_iter είναι ίσο με 92, και το μέσο ποσοστό ακρίβειας είναι ίσο με 77%.

4.6.3.2 Παράμετροι δέντρων αποφάσεων (DecisionTreeClassifier)

Για την βελτίωση της αποτελεσματικότητας του αλγορίθμου των δέντρων αποφάσεων (DecisionTreeClassifier) χρειάστηκε να μεταβληθεί η παράμετρος max_depth, θέτοντάς την από «None» σε «4». Ο αλγόριθμος με την χρήση

default παραμέτρων πετυχαίνει μέσο ποσοστό ακρίβειας ίσο με 84%, με random_state ίσο με 16.

Η παράμετρος max_depth έχει προκαθορισμένη την τιμή «None» και μπορεί να δεχθεί ακέραιες αριθμητικές τιμές. Η παράμετρος max_depth υποδεικνύει το μέγιστο βάθος που μπορεί να επεκταθεί το δέντρο.

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των δοκιμών που χρησιμοποιήθηκαν για τις παραπάνω παραμέτρους με βάση την ακρίβεια του αλγορίθμου.

Παράμετροι	**Αποτέλεσμα**
max_depth	**Test Accuracy (%)**
1	80
2	83
3	86
4	87

Πίνακας 15: Αποτέλεσμα παραμέτρου δέντρων αποφάσεων

Αλλάζοντας την τιμή της παραμέτρου max_depth από «None» σε «4» αυξήθηκε η αποτελεσματικότητα του αλγορίθμου με μέσο ποσοστό ακρίβειας 87%.

4.6.3.3 Παράμετροι τυχαία δάση αποφάσεων (RandomForestClassifier)

Για τη βελτίωση της αποτελεσματικότητας του αλγορίθμου τυχαία δάση αποφάσεων (RandomForestClassifier) χρειάστηκε να αυξηθεί η παράμετρος

n_estimators από 100 σε 110. Ο αλγόριθμος με την χρήση default παραμέτρων πετυχαίνει μέσο ποσοστό ακρίβειας ίσο με 88%, με random_state ίσο με 16.

Η συγκεκριμένη παράμετρος n_estimators είναι ο αριθμός των δέντρων που θα χρησιμοποιηθούν στο δάσος. Δεδομένου ότι το Random Forest είναι μια μέθοδος συνόλου που περιλαμβάνει τη δημιουργία δέντρων πολλαπλών αποφάσεων, χρησιμοποιείται η εν λόγω παράμετρος για τον έλεγχο του αριθμού των δέντρων που θα χρησιμοποιηθούν στη διαδικασία.

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των δοκιμών που χρησιμοποιήθηκαν για τις παραπάνω παραμέτρους με βάση την ακρίβεια του αλγορίθμου.

Παράμετροι	**Αποτέλεσμα**
n_estimators	**Average Test Accuracy (%)**
100	88
110	89
120	89

Πίνακας 16: Αποτελέσματα παραμέτρων για τα τυχαία δάση αποφάσεων

Αλλάζοντας την τιμή της παραμέτρου n_estimators από 100 σε 110 αυξήθηκε η αποτελεσματικότητα του αλγορίθμου με μέσο ποσοστό ακρίβειας 89%.

4.6.3.4 Παράμετροι μηχανές διανυσμάτων υποστήριξης (SVM)

Για τη βελτίωση της αποτελεσματικότητας του αλγορίθμου μηχανές διανυσμάτων υποστήριξης (SVM) χρειάστηκε να μεταβληθεί η τιμή της παραμέτρου C από 1.0 σε 4.0. Ο αλγόριθμος με την χρήση default παραμέτρων πετυχαίνει μέσο ποσοστό ακρίβειας ίσο με 81%, με random_state ίσο με 16.

Η παράμετρος C υποδεικνύει στη βελτιστοποίηση SVM κατά πόσο μπορεί να αποφευχθεί η εσφαλμένη ταξινόμηση κάθε παραδείγματος εκπαίδευσης. Για μεγάλες τιμές C, η βελτιστοποίηση θα επιλέξει ένα μικρότερο περιθώριο υπερ-επιπέδου εάν αυτό το υπερ-επίπεδο βοηθά στην καλύτερη ταξινόμηση όλων των σημείων εκπαίδευσης. Αντιστρόφως, μια πολύ μικρή τιμή του C θα κάνει τον βελτιστοποιητή να αναζητήσει ένα μεγαλύτερο περιθώριο υπερ-επίπεδο διαχωρισμού, ακόμα και αν αυτό το υπερ-επίπεδο ταξινομεί εσφαλμένα περισσότερα σημεία. Για πολύ μικρές τιμές του C, αναμένονται εσφαλμένα ταξινομημένα παραδείγματα, συχνά ακόμη και αν τα δεδομένα εκπαίδευσης είναι γραμμικά διαχωρίσιμα.[46]

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των δοκιμών που χρησιμοποιήθηκαν για τις παραπάνω παραμέτρους με βάση την ακρίβεια του αλγορίθμου.

Παράμετροι	**Αποτέλεσμα**
C	**Average Test Accuracy (%)**
1	81
2	83
3	83
4	84

Πίνακας 17: Αποτελέσματα παραμέτρων για το Υποστηρικτικό μηχάνημα υποστήριξης

Αλλάζοντας την τιμή της παραμέτρου C από 1.0 σε 4.0 αυξήθηκε η αποτελεσματικότητα του αλγορίθμου με μέσο ποσοστό ακρίβειας 84%.

4.6.3.5 Παράμετροι K-πλησιέστερων γειτόνων (K Nearest Neighbors)

Για τη βελτίωση της αποτελεσματικότητας του αλγορίθμου κ-πλησιέστερων γειτόνων (K Nearest Neighbors) χρειάστηκε να μειωθεί η τιμή της παραμέτρου p από 2 σε 1. Ο αλγόριθμος με την χρήση default παραμέτρων πετυχαίνει μέσο

ποσοστό ακρίβειας ίσο με 85%, χωρίς την χρήση της παραμέτρου random_state καθώς δεν υποστηρίζεται από τον συγκεκριμένο αλγόριθμο.

Το p είναι η παράμετρος ισχύος για τη μέτρηση Minkowski. Όταν το p είναι ίσο με 1, αυτό ισοδυναμεί με τη χρήση απόστασης Μανχάταν (L1), ενώ όταν το p είναι ίσο με 2 τότε ισοδυναμεί με τη χρήση της ευκλείδειας απόστασης (L2).[47]

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των δοκιμών που χρησιμοποιήθηκαν για τις παραπάνω παραμέτρους με βάση την ακρίβεια του αλγορίθμου.

Παράμετροι	**Αποτέλεσμα**
p	**Test Accuracy**
1	86
2	85
3	85

Πίνακας 18: Αποτελέσματα παραμέτρων του K-πλησιέστερος γειτόνων

Αλλάζοντας την τιμή της παραμέτρου p από 2.0 σε 1.0 αυξήθηκε η αποτελεσματικότητα του αλγορίθμου με μέσο ποσοστό ακρίβειας 86%.

ΚΕΦΑΛΑΙΟ 5 – ΑΠΟΤΕΛΕΣΜΑΤΑ

Μετά την ολοκλήρωση της διαδικασίας για τη βελτίωση των ποσοστών ακρίβειας και της αποτελεσματικότητας των αλγορίθμων μηχανικής μάθησης σειρά έχει η παρουσίαση των αποτελεσμάτων τους. Στη συνέχεια, γίνεται εύρεση του καλύτερου αλγόριθμου μηχανικής μάθησης με βάση το μέσο ποσοστό ακρίβειας που πετυχαίνει.

5.1 Αποτελέσματα ανά αλγόριθμο

Όλοι οι αλγόριθμοι μηχανικής μάθησης, όπως αναφέρθηκε και πιο πάνω, εκπαιδεύτηκαν σε δύο (2) στάδια, ένα με τη χρήση των προκαθορισμένων τιμών των παραμέτρων και ένα με την προσαρμογή των τιμών των παραμέτρων. Παρακάτω ακολουθούν τα αποτελέσματα για κάθε ένα αλγόριθμο.

5.1.1 Λογιστική παλινδρόμηση (Logistic Regression)

Παρακάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου μηχανικής μάθησης λογιστικής παλινδρόμησης (Logistic Regression).

- Με προκαθορισμένες τιμές παραμέτρων:
  LogisticRegression(random_state=16)

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	6%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	23%
Μέσος όρος ακρίβειας του αλγορίθμου	77%

Πίνακας 19: Αποτελέσματα λογιστικής παλινδρόμησης με προκαθορισμένες τιμές παραμέτρων

- Με προσαρμογή παραμέτρων:
  LogisticRegression(random_state=16,max_iter=92)

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	6%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	23%
Μέσος όρος ακρίβειας του αλγορίθμου	77%

Πίνακας 20: Αποτελέσματα λογιστικής παλινδρόμησης με προσαρμογή παραμέτρων

5.1.2 Δέντρα αποφάσεων (DecisionTreeClassifier)

Παρακάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου μηχανικής μάθησης των δέντρων αποφάσεων (DecisionTreeClassifier).

- Με προκαθορισμένες τιμές παραμέτρων:
  DecisionTreeClassifier(random_state=16)

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	2%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	16%
Μέσος όρος ακρίβειας του αλγορίθμου	84%

Πίνακας 21: Αποτελέσματα δέντρων αποφάσεων με προκαθορισμένες τιμές παραμέτρων

- Με προσαρμογή παραμέτρων:
  DecisionTreeClassifier(random_state=16, max_depth=4)

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	3%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	13%
Μέσος όρος ακρίβειας του αλγορίθμου	87%

Πίνακας 22: Αποτελέσματα δέντρων αποφάσεων με προσαρμογή παραμέτρων

5.1.3 Τυχαία δάση αποφάσεων (RandomForestClassifier)

Παρακάτω παρουσιάζονται τα αποτελέσματα από τα τυχαία δάση αποφάσεων (RandomForestClassifier).

- Με προκαθορισμένες τιμές παραμέτρων:
  RandomForestClassifier(random_state=16)

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	4%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	12%
Μέσος όρος ακρίβειας του αλγορίθμου	88%

Πίνακας 23: Αποτελέσματα από τα τυχαία δάση αποφάσεων με προκαθορισμένες τιμές παραμέτρων

- Με προσαρμογή παραμέτρων:
  RandomForestClassifier(random_state=16, n_estimators=110)

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	4%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	11%
Μέσος όρος ακρίβειας του αλγορίθμου	89%

Πίνακας 24: Αποτελέσματα από τα τυχαία δάση αποφάσεων με προσαρμογή παραμέτρων

5.1.4 Μηχανές διανυσμάτων υποστήριξης (SVM)

Παρακάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου μηχανές διανυσμάτων υποστήριξης (SVM).

- Με προκαθορισμένες τιμές παραμέτρων: SVC(random_state=16)

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	6%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	19%
Μέσος όρος ακρίβειας του αλγορίθμου	81%

Πίνακας 25: Αποτελέσματα από τις μηχανές διανυσμάτων υποστήριξης με προκαθορισμένες τιμές παραμέτρων

- Με προσαρμογή παραμέτρων: SVC(random_state=16, C=4)

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	5%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	16%
Μέσος όρος ακρίβειας του αλγορίθμου	84%

Πίνακας 26: Αποτελέσματα από τις μηχανές διανυσμάτων υποστήριξης με προσαρμογή παραμέτρων

5.1.5 Κ-πλησιέστερων γειτόνων (K Nearest Neighbors)

Παρακάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου μηχανικής μάθησης Κ-πλησιέστερων γειτόνων (K Nearest Neighbors ή αλλιώς KNN).

- Με προκαθορισμένες τιμές παραμέτρων: KNeighborsClassifier()

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	5%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	15%
Μέσος όρος ακρίβειας του αλγορίθμου	85%

Πίνακας 27: Αποτελέσματα Κ-πλησιέστερων γειτόνων με προκαθορισμένες τιμές παραμέτρων

- Με προσαρμογή παραμέτρων: KNeighborsClassifier(p=1)

Παρακάτω παρατίθενται τα αποτελέσματα που επιτυγχάνονται:

Μέση τυπική απόκλιση του αλγορίθμου	5%
Μέσο σφάλμα επικύρωσης του αλγορίθμου	14%
Μέσος όρος ακρίβειας του αλγορίθμου	86%

Πίνακας 28: Αποτελέσματα Κ-πλησιέστερων γειτόνων με προσαρμογή παραμέτρων

5.2 Εύρεση καλύτερου αλγορίθμου

Παρακάτω παρουσιάζονται συγκεντρωτικά όλες οι στατιστικές μετρήσεις όλων των αλγορίθμων τόσο με τη χρήση παραμέτρων, όσο και χωρίς την χρήση αυτών.

- Μέση τυπική απόκλιση

Ο παρακάτω πίνακας περιέχει όλους τους μέσους όρους τυπικής απόκλισης των αλγορίθμων συγκεντρωτικά.

	Algorithm	Parameter	Average of Standard Deviation
1	DecisionTreeClassifier	Default	2
2	DecisionTreeClassifier	Customized	3
3	RandomForestClassifier	Customized	4
4	RandomForestClassifier	Default	4
5	SVM	Customized	5
6	KNN	Customized	5
7	KNN	Default	5
8	LogisticRegression	Customized	6
9	LogisticRegression	Default	6
10	SVM	Default	6

Πίνακας 29: Συγκεντρωτικός πίνακας μέσου όρου τυπικής απόκλισης αλγορίθμων

- Μέσο σφάλμα επικύρωσης

Ο παρακάτω πίνακας περιέχει όλους τους μέσους όρους σφάλματος επικύρωσης των αλγορίθμων συγκεντρωτικά.

	Algorithm	Parameter	Average of Validation Error
1	RandomForestClassifier	Customized	11
2	RandomForestClassifier	Default	12
3	DecisionTreeClassifier	Customized	13
4	KNN	Customized	14
5	KNN	Default	15
6	DecisionTreeClassifier	Default	16
7	SVM	Customized	16
8	SVM	Default	19
9	LogisticRegression	Customized	23
10	LogisticRegression	Default	23

Πίνακας 30: Συγκεντρωτικός πίνακας μέσου όρου σφάλματος επικύρωσης αλγορίθμων

- Μέσος όρος ακρίβειας

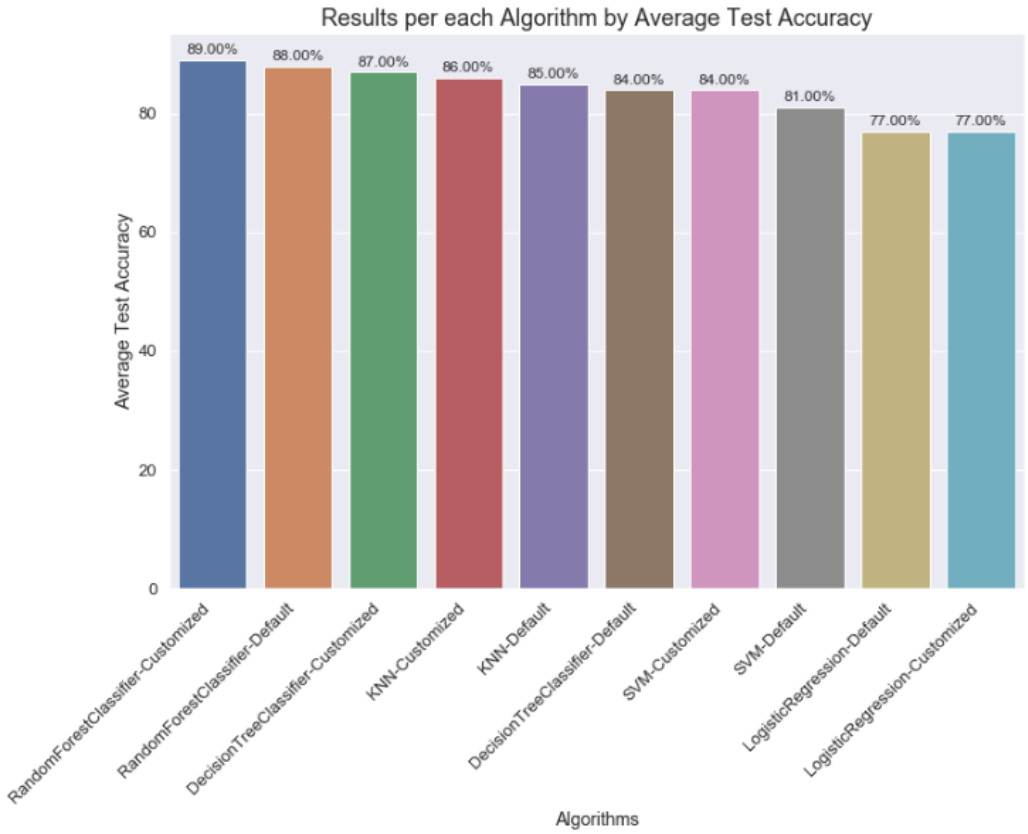
Ο παρακάτω πίνακας περιέχει όλους τους μέσους όρους ακρίβειας των αλγορίθμων συγκεντρωτικά τόσο με τη χρήση παραμέτρων, όσο και χωρίς τη χρήση αυτών. Ο πίνακας έχει ταξινομηθεί κατά φθίνουσα σειρά (στην πρώτη θέση βρίσκεται ο καλύτερος αλγόριθμος) με βάση το μέσο ποσοστό ακρίβειας

της αποτελεσματικότητας των αλγορίθμων.

	Algorithm	Parameter	Average Test Accuracy
1	RandomForestClassifier	Customized	89
2	RandomForestClassifier	Default	88
3	DecisionTreeClassifier	Customized	87
4	KNN	Customized	86
5	KNN	Default	85
6	DecisionTreeClassifier	Default	84
7	SVM	Customized	84
8	SVM	Default	81
9	LogisticRegression	Default	77
10	LogisticRegression	Customized	77

Πίνακας 31: Συγκεντρωτικός πίνακας μέσου όρου ακρίβειας αλγορίθμων

Τα παραπάνω αποτελέσματα αποτυπώνονται και στην παρακάτω γραφική αναπαράσταση για την καλύτερη απεικόνιση των αποτελεσμάτων (Εικόνα 22).



Εικόνα 22: Γραφική αναπαράσταση των αλγορίθμων με τα καλύτερα ποσοστά ακρίβειας

Συνοψίζοντας όλα τα παραπάνω στατιστικά αποτελέσματα με την χρήση των αλγορίθμων μηχανικής μάθησης καταλήγει κανείς στο συμπέρασμα ότι ο καλύτερος αλγόριθμος είναι τα τυχαία δάση αποφάσεων (RandomForestClassifier) με χρήση παραμέτρων, με μέσο ποσοστό ακριβείας 89%.

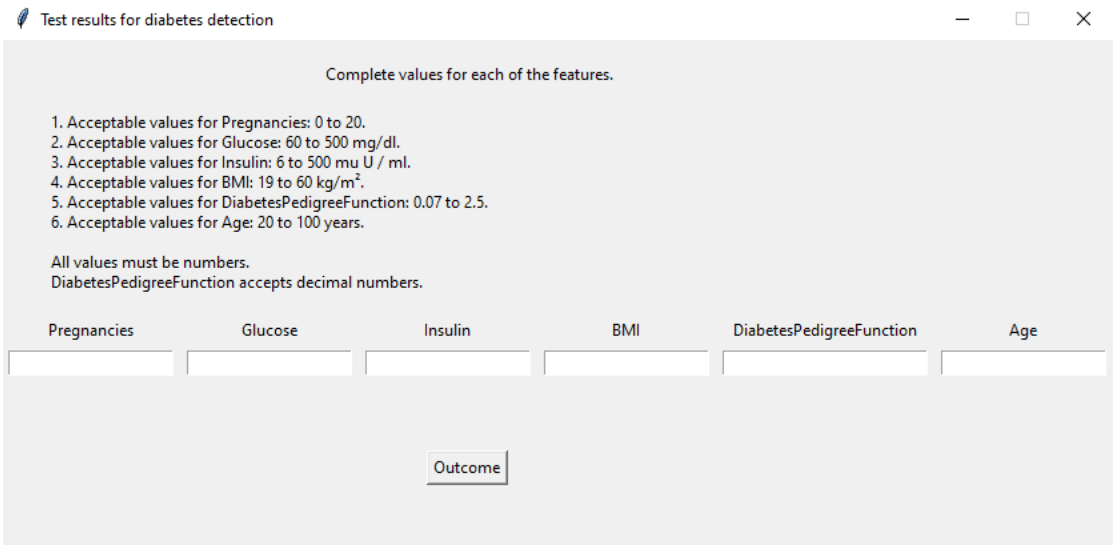
Στη συνέχεια της διπλωματικής εργασίας ο αλγόριθμος που έχει επιλεχθεί ως ο καλύτερος (τα τυχαία δάση αποφάσεων) εκπαιδεύεται εκ νέου σε όλο το σύνολο δεδομένων διατηρώντας τις επιλεγμένες παραμέτρους και χρησιμοποιείται στην εφαρμογή που έχει αναπτυχθεί για την πρόβλεψη του σακχαρώδη διαβήτη τύπου 2.

ΚΕΦΑΛΑΙΟ 6 – ΑΝΑΠΤΥΞΗ ΕΦΑΡΜΟΓΗΣ

Στο τελευταίο σκέλος της μεταπτυχιακής διπλωματικής εργασίας αναπτύχθηκε μία εφαρμογή κατά την οποία ο χρήστης έχει τη δυνατότητα να ορίσει τις τιμές των χαρακτηριστικών στα κατάλληλα πεδία και με βάσει αυτές θα προβλέπεται αν έχει διαγνωστεί με σακχαρώδη διαβήτη ή είναι υγιής.

6.1 Λεπτομερής περιγραφή εφαρμογής

Πιο αναλυτικά, κατά την εκτέλεση της εφαρμογής εμφανίζεται ένα παράθυρο, στο οποίο ο χρήστης καλείται να συμπληρώσει τα κατάλληλα πεδία για την πρόβλεψη του σακχαρώδη διαβήτη, όπως απεικονίζεται στην παρακάτω εικόνα (Εικόνα 23).



Εικόνα 23: Η εφαρμογή που έχει αναπτυχθεί

Ο χρήστης καλείται να συμπληρώσει τα πεδία για κάθε ένα από τα χαρακτηριστικά, σύμφωνα με τους κανόνες που βρίσκονται στην αρχή της εφαρμογής. Οι κανόνες έχουν προκύψει με βάση τις επιτρεπτές μέγιστες και ελάχιστες τιμές όπως αυτές ορίστηκαν με την μέθοδο της προεπεξεργασίας των δεδομένων για κάθε ένα χαρακτηριστικό.

Οι κανόνες περιγράφονται αναλυτικά παρακάτω:

i) Οι εγκυμοσύνες (Pregnancies) μπορούν να πάρουν τιμές από 0 έως και 20.

ii) Η συγκέντρωση γλυκόζης πλάσματος (Glucose) μπορεί να πάρει τιμές από 60 έως και 500 mg/dl.

iii) Η ινσουλίνη (Insulin) μπορεί να πάρει τις τιμές από 6 έως και 500 mu U / ml.

iv) Ο δείκτης μάζας σώματος (BMI) μπορεί να πάρει τις τιμές από 19 έως και 60.

v) Λειτουργία γενεαλογικού διαβήτη (DiabetesPedigreeFunction) μπορεί να πάρει τις τιμές από 0,07 έως και 2,5. Οι δυνατές τιμές που μπορεί να πάρει το συγκεκριμένο χαρακτηριστικό δεν είναι προσδιορισμένες, για αυτό τον λόγο κυμαινόμαστε στα πλαίσια τιμών που περιέχονται στο σύνολο δεδομένων.

vi) Ηλικία του ασθενή (Age) μπορεί να πάρει τις τιμές από 20 έως και 100 έτη. Κάτω από 20 ετών ο σακχαρώδης διαβήτης εμφανίζεται σε άτομα που χαρακτηρίζονται ως παχύσαρκα.

Επιπλέον, όλα τα πεδία πρέπει να συμπληρωθούν με θετικές αριθμητικές τιμές. Τα πεδία που αφορούν τον δείκτη μάζας σώματος (BMI) και τη λειτουργία γενεαλογικού διαβήτη (DiabetesPedigreeFunction) μπορούν να πάρουν και δεκαδικά νούμερα. Σε όλα τα υπόλοιπα πεδία πρέπει να συμπληρωθούν με ακέραιους αριθμούς. Οποιαδήποτε άλλη εκχώρηση δεν γίνεται αποδεκτή.

Για να εκχωρηθούν οι τιμές, καθώς και για να γίνει ο κατάλληλος έλεγχος των τιμών σύμφωνα με τα παραπάνω κριτήρια απαιτείται από το χρήστη να πατήσει το κουμπί «Outcome» (Αποτέλεσμα).

Σε περίπτωση που ο χρήστης συμπληρώσει λανθασμένα κάποιο από τα πεδία ή αν ακόμα αφήσει κάποιο κενό, εμφανίζεται σχετικό μήνυμα σφάλματος. Οι περιπτώσεις λανθασμένης εκχώρησης αναφέρονται λεπτομερώς παρακάτω.

Αν κάποιο από τα πεδία μείνει κενό, δηλαδή δεν του εκχωρηθεί κάποιος χαρακτήρας, τότε το μήνυμα που εμφανίζεται κάτω από το αντίστοιχο πεδίο είναι το εξής: «Enter value.».

Αν συμπληρωθεί κάποιο πεδίο με μη αριθμητικούς χαρακτήρες, όπως για παράδειγμα αν ο χρήστης προσπαθήσει να εισάγει μία λέξη ή περιέχονται σημεία στίξης (εκτός των σημείων στίξης «,» και «.» όπου δηλώνονται οι δεκαδικοί αριθμοί), τότε εμφανίζεται το εξής μήνυμα: «DiabetesPedigreeFunction must be a number.», όπου DiabetesPedigreeFunction είναι το πεδίο που συμπληρώθηκε λανθασμένα.

Αν συμπληρωθεί κάποιο από τα πεδία με αρνητική τιμή, τότε το μήνυμα που εμφανίζεται κάτω από το αντίστοιχο πεδίο είναι το εξής: «DiabetesPedigreeFunction must be a positive number.», όπου DiabetesPedigreeFunction είναι το πεδίο που συμπληρώθηκε με αρνητική τιμή.

Αν συμπληρωθεί κάποιο από τα πεδία Pregnancies, Glucose, Insulin και Age με δεκαδικές τιμές, τότε το μήνυμα που εμφανίζεται κάτω από το αντίστοιχο πεδίο είναι το εξής: «Age must be an integer.», όπου Age είναι το πεδίο που συμπληρώθηκε με δεκαδική τιμή. Η δήλωση των δεκαδικών τιμών γίνεται με τη χρήση του σημείου στίξης κόμμα (,) και του σημείου στίξης τελεία (.).

Αν συμπληρωθεί κάποιο από τα πεδία με τιμές μικρότερες ή μεγαλύτερες των επιτρεπτών, τότε το μήνυμα που εμφανίζεται κάτω από το αντίστοιχο πεδίο υπενθυμίζει το εύρος των επιτρεπτών τιμών για το εκάστοτε πεδίο και είναι το εξής: «DiabetesPedigreeFunction must be between 0.07 to 2.5.», όπου DiabetesPedigreeFunction είναι το πεδίο που συμπληρώθηκε με τιμή που δεν εμπίπτει στο επιτρεπτό εύρος τιμών.

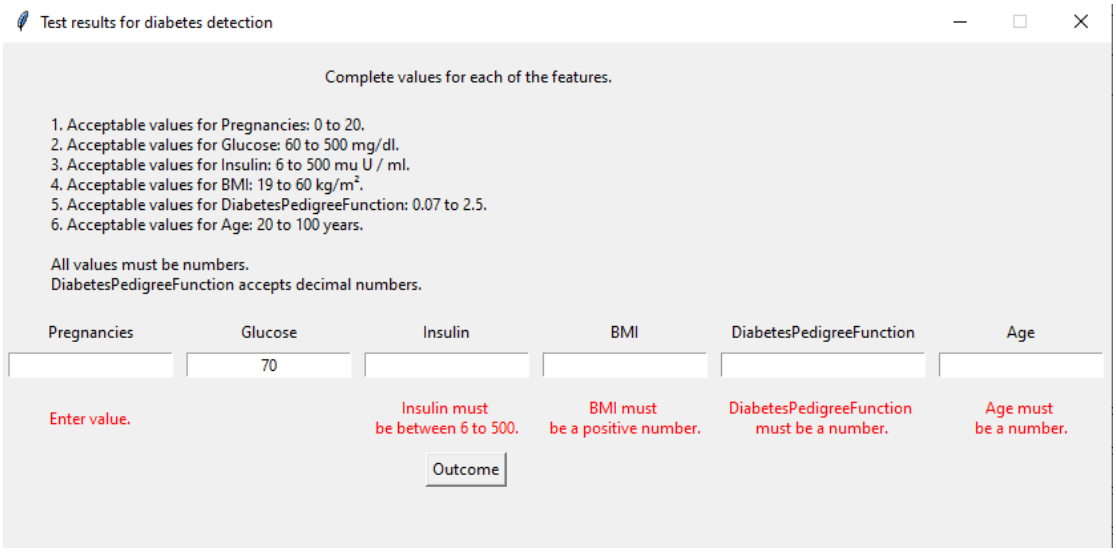
Στην περίπτωση που κάποιο από τα πεδία συμπληρώνεται σύμφωνα με τους κανόνες του εκάστοτε πεδίου, τότε δεν εμφανίζεται κάποιο μήνυμα λάθους και η εκχώρηση της τιμής γίνεται αποδεκτή.

Μέχρι να συμπληρωθούν όλα τα πεδία με αποδεκτές τιμές, ο χρήστης έχει τη

δυνατότητα να μεταβάλει όχι μόνο τα πεδία με τις μη αποδεκτές τιμές, αλλά ακόμα και τα πεδία που έχουν ήδη συμπληρωθεί με αποδεκτές τιμές.

Στη συνέχεια παρουσιάζεται ένα στιγμιότυπο (Εικόνα 24) κατά το οποίο έχουν εκχωρηθεί όλοι οι δυνατοί συνδυασμοί εισόδων με τα αντίστοιχα μηνύματα ελέγχου. Αναλυτικά για κάθε πεδίο δόθηκαν ενδεικτικά οι παρακάτω τιμές:

- Στο πεδίο Pregnancies (εγκυμοσύνες) εκχωρήθηκε η τιμή: (κενό)
- Στο πεδίο Glucose (συγκέντρωση γλυκόζης πλάσματος) εκχωρήθηκε η τιμή: 70
- Στο πεδίο Insulin (ινσουλίνη) εκχωρήθηκε η τιμή: 3
- Στο πεδίο BMI (δείκτης μάζας σώματος) εκχωρήθηκε η τιμή: -30
- Στο πεδίο DiabetesPedigreeFunction (λειτουργία γενεαλογικού διαβήτη) εκχωρήθηκε η τιμή: 0.#2
- Στο πεδίο Age (ηλικία) εκχωρήθηκε η τιμή: 26a

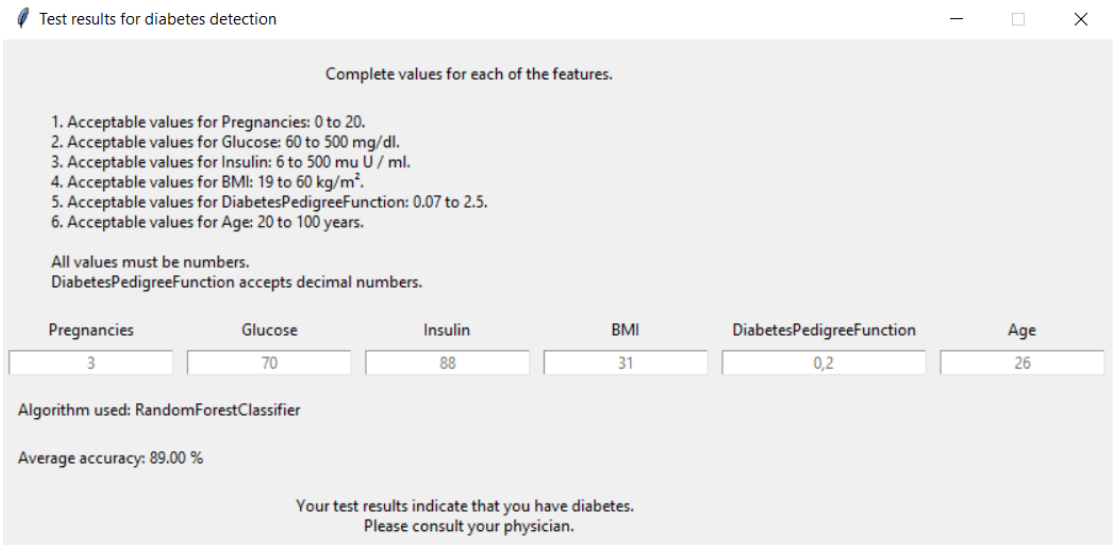


Εικόνα 24: Εμφάνιση όλων των μηνυμάτων που εμφανίζονται κατά την εκχώρηση τιμών

Με την συμπλήρωση των αποδεκτών εκχωρήσεων για κάθε ένα από τα πεδία, η εφαρμογή με βάσει τις εκχωρήσεις των χαρακτηριστικών και με τη χρήση του αποθηκευμένου μοντέλου πραγματοποιεί την πρόβλεψη για το αν ο συγκεκριμένος χρήστης έχει σακχαρώδη διαβήτη ή δεν έχει. Κατά την εκτέλεση της εφαρμογής παρουσιάζεται το όνομα του μοντέλου που χρησιμοποιήθηκε,

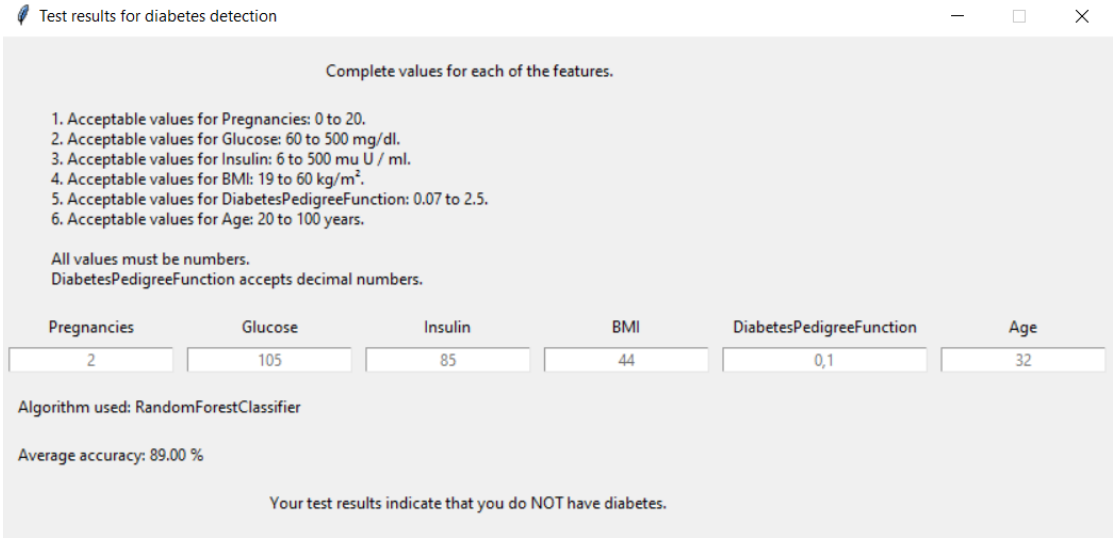
καθώς επίσης και το ποσοστό επιτυχίας όσων αφορά την ακρίβεια της αποτελεσματικότητας του μοντέλου.

Στο παρακάτω στιγμιότυπο (Εικόνα 25) φαίνεται το αποτέλεσμα της πρόβλεψης, όταν αυτή είναι θετική στη διάγνωση του σακχαρώδους διαβήτη.



Εικόνα 25: Εμφάνιση αποτελέσματος όταν είναι θετική η πρόβλεψη του σακχαρώδη διαβήτη

Στο παρακάτω στιγμιότυπο (Εικόνα 26) φαίνεται το αποτέλεσμα της πρόβλεψης, όταν αυτή είναι αρνητική στη διάγνωση του σακχαρώδους διαβήτη.



Εικόνα 26: Εμφάνιση αποτελέσματος όταν είναι αρνητική η πρόβλεψη του σακχαρώδη διαβήτη

ΚΕΦΑΛΑΙΟ 7 – ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Στόχος αυτής της μεταπτυχιακής διπλωματικής εργασίας είναι να συγκρίνει αλγορίθμους μηχανικής μάθησης και να αξιολογήσει την επίδοση των αλγορίθμων για τη διάγνωση του σακχαρώδη διαβήτη τύπου 2 μέσω του συνόλου δεδομένων Pima Indians Diabetes Database.

Στο τελευταίο αυτό κεφάλαιο, πραγματοποιήθηκε αναδρομή και σύγκριση με άλλες αντίστοιχες έρευνες οι οποίες έχουν γίνει με χρήση του ίδιου συνόλου δεδομένων, καθώς και των ίδιων αλγορίθμων. Επίσης, πραγματοποιήθηκε ανάλυση της εφαρμογής που αναπτύχθηκε για τους σκοπούς αυτής της εργασίας. Τέλος, παρατίθενται βελτιωτικές προτάσεις τόσο στην επιλογή και παραμετροποίηση των αλγορίθμων μηχανικής μάθησης όσο και της εφαρμογής.

7.1 Αξιολόγηση και σύγκριση αλγορίθμων

7.1.1 Σύγκριση αποτελεσμάτων πρώτης έρευνας με την παρούσα διπλωματική εργασία

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των αλγορίθμων μηχανικής μάθησης για την πρόβλεψη του σακχαρώδη διαβήτη τόσο της παρούσας διπλωματικής εργασίας, όσο και της πρώτης έρευνας που έχει αναλυθεί στο κεφάλαιο 2. Στην πρώτη αυτή έρευνα δεν πραγματοποιήθηκε διαδικασία προ-επεξεργασίας των δεδομένων. Καλύτερα αποτελέσματα επιτυγχάνει ο αλγόριθμος λογιστικής παλινδρόμησης (Logistic Regression) με ποσοστό ακρίβειας 77,6%. Όσον αφορά την συγκεκριμένη μεταπτυχιακή διπλωματική εργασία παρόλο που στην λογιστική παλινδρόμηση το ποσοστό ακρίβειας είναι περίπου το ίδιο, στους άλλους αλγόριθμους μηχανικής μάθησης παρατηρείται σημαντική αύξηση της αποτελεσματικότητας όπως φαίνεται παρακάτω.

Αλγόριθμοι	Πρώτη έρευνα Accuracy %	Διπλωματική εργασία (default) Accuracy %	Διπλωματική εργασία (customized) Accuracy %
Logistic Regression	77.6	77	77
Decision Tree	70.31	84	87
Random Forest	74.30	88	89
Support Vector Machine	65.63	81	84
KNN	73.43	85	86

Πίνακας 32: Αποτελέσματα σύγκρισης 1ης έρευνας με διπλωματική εργασία

7.1.2 Σύγκριση αποτελεσμάτων δεύτερης έρευνας με την παρούσα διπλωματική εργασία

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των αλγορίθμων μηχανικής μάθησης για την πρόβλεψη του σακχαρώδη διαβήτη τόσο της παρούσας διπλωματικής εργασίας, όσο και της δεύτερης έρευνας που έχει αναλυθεί στο κεφάλαιο 2. Για την σύγκριση επιλέχθηκαν οι αλγόριθμοι που χρησιμοποιήθηκαν από κοινού. Στην συγκεκριμένη έρευνα έγινε αφαίρεση 13 εγγραφών οι οποίες περιείχαν μηδενικές τιμές σε χαρακτηριστικά που δεν το επέτρεπαν. Ο αλγόριθμος της λογιστικής παλινδρόμησης πέτυχε υψηλότερο ποσοστό ακρίβειας ίσο με 78,01%. Οι μεγαλύτερες διαφορές παρουσιάζονται στους αλγόριθμους τυχαία δάση αποφάσεων (Random Forest) και μηχανές διανυσμάτων υποστήριξης (Support Vector Machine), όπου η διαφορά κυμαίνεται περίπου στις 15 ποσοστιαίες μονάδες και στις 7, αντίστοιχα.

Αλγόριθμοι	Δεύτερη έρευνα Accuracy %	Διπλωματική εργασία (default) Accuracy %	Διπλωματική εργασία (customized) Accuracy %
Logistic Regression	78.01	77	77
Random Forest	74.83	88	89
Support Vector Machine	77.08	81	84

Πίνακας 33: Αποτελέσματα σύγκρισης 2ης έρευνας με διπλωματική εργασία

7.1.3 Σύγκριση αποτελεσμάτων τρίτης έρευνας με την παρούσα διπλωματική εργασία

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των αλγορίθμων μηχανικής μάθησης για την πρόβλεψη του σακχαρώδη διαβήτη τόσο της παρούσας διπλωματικής εργασίας, όσο και της τρίτης έρευνας που έχει αναλυθεί στο κεφάλαιο 2. Τα αποτελέσματα που αφορούν την έρευνα αυτή είναι κατόπιν προ-επεξεργασίας, πραγματοποιώντας κανονικοποίηση (Normalization) των δεδομένων με χρήση του μοντέλου κανονικοποίησης Min Max Scaler (MMS). Η διαφορά ανάμεσα στην παρούσα μεταπτυχιακή διπλωματική εργασία και στην συγκεκριμένη έρευνα για τους αλγόριθμους μηχανικής μάθησης μηχανές διανυσμάτων υποστήριξης (Support Vector Machine) και Κ-πλησιέστερων γειτόνων (KNN) κυμαίνεται στις 6 ποσοστιαίες μονάδες και στις 11, αντίστοιχα.

Αλγόριθμοι	Τρίτη έρευνα Accuracy %	Διπλωματική εργασία (default) Accuracy %	Διπλωματική εργασία (customized) Accuracy %
Support Vector Machine	78.05	81	84
KNN	75.5	85	86

Πίνακας 34: Αποτελέσματα σύγκρισης 3ης έρευνας με διπλωματική εργασία

7.1.4 Συμπεράσματα

Τα αποτελέσματα των αλγόριθμων μηχανικής μάθησης αναφέρθηκαν αναλυτικά στο κεφάλαιο 5. Από τα αποτελέσματα καταλήγουμε ότι ο καλύτερος αλγόριθμος για την πρόβλεψη της ασθένειας του σακχαρώδη διαβήτη τύπου 2 είναι τα τυχαία δάση αποφάσεων (RandomForestClassifier) με ποσοστό ακρίβειας 89%.

Από τη σύγκριση των αποτελεσμάτων μεταξύ των τριών επιλεγμένων μελετών και της παρούσας μεταπτυχιακής διπλωματικής εργασίας, συμπεραίνει κανείς ότι η προ-επεξεργασία του συνόλου δεδομένων επηρεάζει σε μεγάλο βαθμό το ποσοστό ακρίβειας του κάθε αλγορίθμου μηχανικής μάθησης. Τα αποτελέσματα της παρούσας διπλωματικής εργασίας επιτυγχάνουν υψηλότερα ποσοστά ακρίβειας σε σύγκριση με τα αποτελέσματα των τριών ερευνών, κυρίως λόγω προ-επεξεργασίας που έχει πραγματοποιηθεί στο σύνολο των δεδομένων. Τη μόνη εξαίρεση αποτελεί ο αλγόριθμος της λογιστικής παλινδρόμησης που παρέμεινε στα ίδια περίπου επίπεδα ακρίβειας, όπως φαίνεται στον παρακάτω πίνακα.

Έρευνα / Αλγόριθμος	Πρώτη έρευνα Accuray %	Δεύτερη έρευνα Accuray %	Τρίτη έρευνα Accuray %	Διπλωματική εργασία (default) Accuray %	Διπλωματική εργασία (customized) Accuray %
Logistic Regression	77.6	**78.01**	-	77	77
Decision Tree	70.31	-	-	84	**87**
Random Forest	74.3	74.83	-	88	**89**
Support Vector Machine	65.63	77.08	78.05	81	**84**
KNN	73.43	-	75.5	85	**86**

Πίνακας 35: Συγκεντρωτικά αποτελέσματα των τριών ερευνών με διπλωματική εργασία

7.2 Αξιολόγηση εφαρμογής

Η εφαρμογή που αναπτύχθηκε για τις ανάγκες της υλοποίησης αυτής της μεταπτυχιακής διπλωματικής εργασίας παρέχει την δυνατότητα στον χρήστη να κάνει μία πρόβλεψη για το αν είναι θετικός στην εμφάνιση του σακχαρώδη διαβήτη τύπου 2 ή όχι. Ωστόσο, για να εξακριβώσει κάποιος αν έχει σακχαρώδη διαβήτη ή όχι και να γίνει η διάγνωση θα πρέπει να συμβουλευτεί τον ιατρό του και να πραγματοποιήσει τις απαιτούμενες εξετάσεις.

7.3 Μελλοντικές βελτιώσεις

Τόσο οι αλγόριθμοι μηχανικής μάθησης, όσο και η εφαρμογή που αναπτύχθηκε μπορούν και να βελτιωθούν και να εξελιχθούν. Παρακάτω παρατίθενται μερικοί τρόποι βελτίωσης.

Όσον αφορά τους αλγόριθμους μηχανικής μάθησης μπορούν να εφαρμοστούν με διαφορετικές παραμέτρους και να επιτευχθούν καλύτερα και πιο αποτελεσματικά συμπεράσματα. Επίσης, σημαντικό θα ήταν να αυξηθούν οι εγγραφές του συνόλου δεδομένων, με σκοπό την καλύτερη εκπαίδευση των αλγορίθμων.

Τέλος, όσον αφορά τη βελτίωση της εφαρμογής που έχει αναπτυχθεί μπορούν να γίνουν βελτιωτικές αλλαγές, τόσο στην λειτουργία της εφαρμογής όσο και στην αλληλεπίδραση με τον χρήστη. Θα ήταν καλό ο χρήστης να είναι σε θέση να επιλέγει τον αλγόριθμο μηχανικής μάθησης που επιθυμεί και να μην είναι ήδη προεπιλεγμένος. Επίσης, θα ήταν πιο ελκυστικό προς τον χρήστη αν γινόταν προσθήκη γραφικών στην εφαρμογή.

Η συνεχής εξέλιξη της τεχνολογίας και η ευρύτερη χρήση της στον ιατρικό τομέα σίγουρα μπορεί να επιφέρει θεαματικά αποτελέσματα στο εγγύς μέλλον όσων αφορά την έγκαιρη διάγνωση και ως εκ τούτου αντιμετώπιση της νόσου.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Sas, “Machine Learning. What it is and why it matters”, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.sas.com/en_us/insights/analytics/machine-learning.html [Τελευταία πρόσβαση 19 Μαΐου 2020]

[2] Dr. Manjiri Bakre, “5 Ways Machine learning is Redefining Healthcare”, October 31 2019, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.entrepreneur.com/article/341626 [Τελευταία πρόσβαση 19 Μαΐου 2020]

[3] Siddhartha Mukherjee, A.I. Versus M.D., March 27 2017, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.newyorker.com/magazine/2017/04/03/ai-versus-md [Τελευταία πρόσβαση 19 Μαΐου 2020]

[4] Nam H. Cho, Rhys Williams, “IDF DIABETES ATLAS”, Ninth edition 2019, ISBN: 978-2-930229-87-4

[5] Ελληνική Διαβητολογική Εταιρεία (ΕΔΕ), «Κατευθυντήριες Οδηγίες για τη Διαχείριση του Διαβητικού Ασθενούς», 2019

[6] Thomas W. Edgar, David O. Manz, “Research Methods for Cyber Security”, 2017

[7] Wilson RA, Keil FC. “The MIT encyclopedia of the cognitive sciences”, MIT Press, 1999

[8] Nick McCrea, “An Introduction to Machine Learning Theory and Its Applications: A Visual Tutorial with Examples”, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.toptal.com/machine-learning/machine-learning-theory-an-introductory-primer [Τελευταία πρόσβαση 19 Μαΐου 2020]

[9] Pantech,”Top 10 Machine Learning Projects”, June 26 2018

[10] Devin Soni, towardsdatascience, “Supervised vs. Unsupervised Learning”, March 22 2018, Διαθέσιμο στην ηλεκτρονική διεύθυνση:

https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d [Τελευταία πρόσβαση 19 Μαΐου 2020]

[11] Hunter Heidenreich , “What are the types of machine learning?”, December 5 2018, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f [Τελευταία πρόσβαση 19 Μαΐου 2020]

[12] Aidan Wilson, towardsdatascience, “ A Brief Introduction to Supervised Learning”, September 29 2019, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590 [Τελευταία πρόσβαση 19 Μαΐου 2020]

[13] AlindGupta, GeeksforGeeks, “ML | Semi-Supervised Learning”, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.geeksforgeeks.org/ml-semi-supervised-learning/ [Τελευταία πρόσβαση 19 Μαΐου 2020]

[14] Angelo Cangelosi, Matthew Schlesinger, “Developmental Robotics”, January 2015, ISBN: 9780262028011

[15] Simplilearn, “Classification - Machine Learning”. Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.simplilearn.com/classification-machine-learning-tutorial [Τελευταία πρόσβαση 19 Μαΐου 2020]

[16] Ben Foley, Surveygizmo, “What is Regression Analysis and Why Should I Use It?”, February 14 2018, Διαθέσιμο στην ηλεκτρονική διεύθυνση: [Τελευταία πρόσβαση 19 Μαΐου 2020]

[17] Surya Priy, GeeksforGeeks, “Clustering in Machine Learning”, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.geeksforgeeks.org/clustering-in-machine-learning/ [Τελευταία πρόσβαση 19 Μαΐου 2020]

[18] George Seif, towardsdatascience, “The 5 Clustering Algorithms Data Scientists Need to Know”, February 5 2018, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 [Τελευταία πρόσβαση 19 Μαΐου 2020]

[19] B. W. Silverman," Density estimation for statistics and data analysis",1986

[20] Anannya Uberoi, GeeksforGeeks, "Introduction to Dimensionality Reduction", Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.geeksforgeeks.org/dimensionality-reduction/ [Τελευταία πρόσβαση 19 Μαΐου 2020]

[21] Judy T Raj, towardsdatascience, "A beginner's guide to dimensionality reduction in Machine Learning", March 11 2019, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e [Τελευταία πρόσβαση 19 Μαΐου 2020]

[22] Karen Haoarchive, technologyreview, "What is machine learning?", November 17 2018, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/ [Τελευταία πρόσβαση 19 Μαΐου 2020]

[23] Aishwarya Jakka, Vakula Rani J, "Performance Evaluation of Machine Learning Models for Diabetes Prediction", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019

[24] Md. Aminul Islam, Nusrat Jahan, "Prediction of Onset Diabetes using Machine Learning Techniques", International Journal of Computer Applications (0975 – 8887), Volume 180 – No.5, December 2017

[25] Samrat Kumar Dey, Ashraf Hossain, Md. Mahbubur Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm", 21st International Conference of Computer and Information Technology (ICCIT), 21-23 December 2018

[26] Atul, edureka, "What is Machine Learning? Machine Learning For Beginners", Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.edureka.co/blog/what-is-machine-learning/ [Τελευταία πρόσβαση 19 Μαΐου 2020]

[27] Statisticssolutions, “What is Logistic Regression?”, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.statisticssolutions.com/what-is-logistic-regression/ [Τελευταία πρόσβαση 20 Ιουνίου 2020]

[28] Saloni Gupta, GeeksforGeeks, Decision Tree, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.geeksforgeeks.org/decision-tree-introduction-example/?ref=rp [Τελευταία πρόσβαση 20 Ιουνίου 2020]

[29] Abhishek Sharma, analyticsvidhya, “Decision Tree vs. Random Forest – Which Algorithm Should you Use?”, May 12 2020, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ [Τελευταία πρόσβαση 20 Ιουνίου 2020]

[30] Mohtadi Ben Fraj, medium, “In Depth: Parameter tuning for SVC”, January 5 2018, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769 [Τελευταία πρόσβαση 20 Ιουνίου 2020]

[31] Onel Harrison, towardsdatascience, “Machine Learning Basics with the K-Nearest Neighbors Algorithm”, September 10 2018, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 [Τελευταία πρόσβαση 19 Μαΐου 2020]

[32] K. R. Srinath, “Python – The Fastest Growing Programming Language”, December 2017

[33] Jupyter, “The Jupyter Notebook”, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://jupyter.org [Τελευταία πρόσβαση 19 Μαΐου 2020]

[34] Charles Bochet, “how to install PySpark and Jupyter Notebook in 3 Minutes”, January 20, 2020, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.sicara.ai/blog/2017-05-02-get-started-pyspark-jupyter-notebook-3-minutes [Τελευταία πρόσβαση 19 Μαΐου 2020]

[35] Χαράλαμπος Δ. Τούντας, “Σακχαρώδης Διαβήτης Θεωρία Πράξη”, 1995, ISBN: 960-90332-0-2

[36] Κοντογιάννη, Μερόπη Γιαννακούλια, Μαρία Καράτζη, Καλλιόπη-Ζαφειρένια Φάππα, Ευαγγελία, “Ανθρωπομετρία”, 2015, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://repository.kallipos.gr/bitstream/11419/1945/1/02_chapter_6.pdf [Τελευταία πρόσβαση 19 Μαΐου 2020]

[37] statisticshowto, correlation matrix, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.statisticshowto.datasciencecentral.com/correlation-matrix/ [Τελευταία πρόσβαση 19 Μαΐου 2020]

[38] Jason Brownlee, “Logistic Regression for Machine Learning”, August 12 2019, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://machinelearningmastery.com/logistic-regression-for-machine-learning/ [Τελευταία πρόσβαση 19 Μαΐου 2020]

[39] Danish Haroon, “Python Machine Learning Case Studies”, Apress, Oktober 27 2017

[40] tutorialspoint, “Classification Algorithms - Random Forest”, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm [Τελευταία πρόσβαση 19 Μαΐου 2020]

[41] Sunil Ray, “Understanding Support Vector Machine(SVM) algorithm from examples (along with code)”,September 13 2017, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ [Τελευταία πρόσβαση 19 Μαΐου 2020]

[42] techopedia, “K-Nearest Neighbor (K-NN)”, March 14 2017, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.techopedia.com/definition/32066/k-nearest-neighbor-k-nn [Τελευταία πρόσβαση 19 Μαΐου 2020]

[43] Prashant Gupta, “Cross-Validation in Machine Learning”, Jun 5 2017, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f [Τελευταία πρόσβαση 19 Μαΐου 2020]

[44] geeksforgeeks, Standard Deviation, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://www.geeksforgeeks.org/python-statistics-stdev/ [Τελευταία πρόσβαση 19 Μαΐου 2020]

[45] scikit-learn, DecisionTreeClassifier, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html [Τελευταία πρόσβαση 19 Μαΐου 2020]

[46] stackexchange, C parameter of SVM, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://stats.stackexchange.com/questions/31066/what-is-the-influence-of-c-in-svms-with-linear-kernel [Τελευταία πρόσβαση 19 Μαΐου 2020]

[47] scikit-learn, KNeighborsClassifier, Διαθέσιμο στην ηλεκτρονική διεύθυνση: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html [Τελευταία πρόσβαση 19 Μαΐου 2020]