Πανεπιστήμιο Πειραιώς
Τμήμα Ψηφιακών Συστημάτων

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ
Κατεύθυνση: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

Διάγνωση διαβήτη με χρήση αλγορίθμων μηχανικής μάθησης

Γρηγόριος Καλαντώνης
ΑΜ: ΜΕ1711

Μεταπτυχιακή Διπλωματική Εργασία

Επιβλέπων: Ανδριάνα Πρέντζα, Καθηγήτρια

Πειραιάς 2019

Περίληψη

Η ιλιγγιώδης εξέλιξη της επιστήμης των δεδομένων, τις τελευταίες δεκαετίες, καθώς και η εξειδίκευσή της στο χώρο των μεγάλων δεδομένων, αποτέλεσε ένα σημαντικό εργαλείο για το σύνολο των επιστημών, θετικών, φυσικών, ανθρωπιστικών και κοινωνικών. Οι επιστήμες αυτές από τη μία «χρησιμοποίησαν» την επιστήμη των μεγάλων δεδομένων, προκειμένου να εξελιχθούν και να βελτιώσουν τα εμπειρικά τους αποτελέσματα, και από την άλλη αποτέλεσαν πηγή δεδομένων που εφοδίασε την επιστήμη των μεγάλων δεδομένων με πληθώρα δεδομένων για να πειραματιστεί και να βελτιώσει τις δικές της μεθόδους. Η αλληλεπίδραση αυτή ήταν και είναι σημαντική και για τα δύο μέρη. Σήμερα λοιπόν είμαστε σε θέση να μετράμε μία αναρίθμητη πια σειρά ψηφιακών μηχανών και αισθητήρων από διαφορετικά ερευνητικά πεδία που παράγουν δεδομένα.

Η βιοτεχνολογία είναι ένας από τους κλάδους που τροφοδότησε και εξακολουθεί να τροφοδοτεί την επιστήμη των μεγάλων δεδομένων με σημαντικό υλικό. Τεχνολογίες όπως η ψηφιακή μικροσκοπία υψηλή ανάλυση, η φασματομετρία μάζα, η απεικόνιση μαγνητικού συντονισμού (MRI) παράγουν καθημερινά αμέτρητα δεδομένα. Πρόκειται ωστόσο για πρωτογενή δεδομένα που δεν παρέχουν καμία ανάλυση, ερμηνεία ή εξαγωγή γνώσης. Αυτό το κενό προσπάθησε να καλύψει ο νέος τομέας της Βιολογικής Εξόρυξης Δεδομένων ή αλλιώς η «Ανακάλυψη της Γνώσης» στα βιολογικά δεδομένα, που ουσιαστικά δεν συλλέγει απλά δεδομένα, αλλά επιπλέον τα επεξεργάζεται και εξάγει συμπεράσματα από αυτά.

«Πρωταρχικός στόχος της “Βιολογικής Εξόρυξης Δεδομένων” είναι να εμβαθύνει στα γρήγορα αναπτυσσόμενα βιολογικά δεδομένα και να θέσει τη βάση που ενισχύει τις απαντήσεις σε θεμελιώδη ζητήματα των επιστημών της βιολογίας και της ιατρικής». (Kavakiotis I., et al., 2017).

Στόχος της παρούσας μεταπτυχιακής διπλωματικής εργασίας είναι η συγκριτική ανάλυση μιας σειράς αλγορίθμων μηχανικής μάθησης (εποπτευόμενης μάθησης) της προβλεπτικής τους ικανότητας στην εφαρμογή τους στην έρευνα της πάθησης του διαβήτη. Συγκεκριμένα, το «πρόβλημα» εντοπίστηκε στο πώς μπορεί να γίνει πρόβλεψη για τη διάγνωση του διαβήτη κατά τη διάρκεια της εγκυμοσύνης, με τη χρήση των συγκεκριμένων αλγορίθμων, και κατά πόσον αυτοί οι αλγόριθμοι διαφοροποιούνται στα αποτελέσματά τους.

Η παρουσίαση των μεθόδων και των αποτελεσμάτων άλλων μελετών για το ίδιο θέμα κρίθηκε αναγκαία. Για τον σκοπό της μελέτης μας ελήφθησαν δεδομένα από το διαδίκτυο και στη συνέχεια αναλύθηκαν τόσο με τη χρήση του εργαλείου weka, όσο και με τη χρήση της γλώσσας προγραμματισμού R.

Summary

Continuous and important developments in biotechnology contribute to the easy and inexpensive production of data, thus leading the science of applied biology to the area of big data.

Today, there are plenty of digital machines and sensors from various research fields that produce data, including high resolution digital microscopy, mass spectrometry, magnetic resonance imaging (MRI), etc. Although these technologies produce a wealth of data, they do not provide any analysis, interpreting or extracting knowledge. For this purpose, the field of Biological Data Mining or otherwise the "discovery of knowledge" in biological data is more than ever necessary and important. The primary objective is to deepen the rapidly growing biological data and to set the basis for the answers to fundamental questions of the biology and medical sciences (Kavakiotis I., et al., 2017).

The purpose of this study is to review the applications of machine learning, techniques and data mining tools in the field of diabetes research during pregnancy in relation to its prediction and diagnosis. A wide range of machine learning algorithms and specifically supervised learning were used.

Data was obtained from the internet and analyzed using the Weka tool. After a comparative review of the results of the methods, we try to draw some conclusions in order to find out which method yields better results for the given data set.

Finally, once we have concluded with the learning methods, using R programming language, we perform them by taking their results.

Ευχαριστίες

Αρχικά, θέλω να ευχαριστήσω την κα. Πρέντζα Ανδριάνα, καθηγήτρια του τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς, για την επίβλεψη, τη πολύτιμη βοήθεια της και την υπομονή που έδειξε καθ' όλη τη διάρκεια της διπλωματικής μου εργασίας και μέχρι την ολοκλήρωση της, με πολύτιμες παρατηρήσεις και συμβουλές.
Ευχαριστώ πολύ, επίσης τον κ. Φιλιππάκη Μιχαήλ καθηγητή και μέλος της επιτροπής εξέτασης της μεταπτυχιακής μου εργασίας, όπως και τους υπόλοιπους καθηγητές του τμήματος για τη βοήθεια που μου προσέφεραν και για τις χρήσιμες παρατηρήσεις τους, καθ' όλη τη διάρκεια του μεταπτυχιακού.
Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους γονείς μου που με στήριξαν και με βοήθησαν σε όλη τη διάρκεια των μεταπτυχιακών σπουδών μου καθώς και στην σύντροφό μου που παρέμεινε στο πλευρό μου διακριτικά και υπομονετικά στηρίζοντας με αγάπη την προσπάθεια μου αυτή.

Πίνακας Περιεχομένων

Πίνακας Εικόνων/Σχημάτων

Πίνακας Πινάκων

Πίνακας Συντομογραφιών

ARFF	Attribute-Relation File Format
Caret	Classification and Regression Training
GDM	Gestational Diabetes Mellitus
GGT	Glucose Tolerance Test
IGT	Impaired Glucose Tolerance
KDD	Knowledge Discovery in Databases
MAP	Maximum A Posteriori
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
UCI	UCI Machine Learning repository
WEKA	Waikato Environment for Knowledge Analysis
ΔΜΣ	Δείκτης Μάζας Σώματος
ΜΔΕ	Μεταπτυχιακή Διπλωματική Εργασία
ΠΟΥ	Παγκόσμιος Οργανισμός Υγείας

Κεφάλαιο 1: Εισαγωγή

1.1 Εισαγωγή

Η ισχύς, η αποτελεσματικότητα και η χρησιμότητα των αλγορίθμων μηχανικής μάθησης προέρχονται πρωταρχικά από την ικανότητά τους να εξάγουν πρότυπα και να δημιουργούν, από τα πρωτογενή δεδομένα, μοντέλα ικανά να προβλέπουν και να οδηγούν σε ασφαλή συμπεράσματα. Το γεγονός αυτό είναι αναμφίβολα ιδιαίτερα σημαντικό. Οι αλγόριθμοι μηχανικής μάθησης μπόρεσαν να οργανώσουν και να αξιοποιήσουν ένα τεράστιο σύνολο δεδομένων, με μέγεθος που μπορεί να φτάσει σε terabyte ή petabytes δεδομένων. Η αφθονία των δεδομένων σαφώς ενίσχυσε σημαντικά αυτή τη διαδικασία σε όλους τους επιστημονικούς τομείς αλλά κυρίως στον τομέα της βιολογίας.

Σε αυτό το υβριδικό πεδίο, μία από τις σημαντικότερες ερευνητικές εφαρμογές είναι η πρόγνωση και η διάγνωση ασθενειών που απειλούν την ανθρώπινη ζωή ή επηρεάζουν την ποιότητα ζωής των ανθρώπων. Μία τέτοια ασθένεια είναι και ο σακχαρώδης διαβήτης (Kavakiotis I., et al., 2017).

Η εφαρμογή μεθόδων μηχανικής μάθησης και εξόρυξης δεδομένων στην έρευνα του σακχαρώδους διαβήτη χρησιμοποιεί πλήθος δεδομένων, προκειμένου να εξάγει σημαντική γνώση για την ασθένεια αυτή. Πέρα από τις επιπτώσεις σε ατομικό επίπεδο, η ασθένεια του σακχαρώδους διαβήτη έχει και σημαντικές κοινωνικές προεκτάσεις. Οι σοβαρές κοινωνικές επιπτώσεις της συγκεκριμένης ασθένειας καθιστούν τον σακχαρώδη διαβήτη μία από τις κύριες προτεραιότητες στην έρευνα των ιατρικών επιστημών, η οποία αναπόφευκτα δημιουργεί τεράστιο όγκο δεδομένων.

Για το λόγο αυτόν έγινε βιβλιογραφική ανασκόπηση και παρουσίαση σημαντικών προσπαθειών που έχουν γίνει μέχρι σήμερα στο συγκεκριμένο θέμα.

Στη συνέχεια, χρησιμοποιώντας διαθέσιμα δεδομένα που αφορούν τον σακχαρώδη διαβήτη, έγινε προσπάθεια να αναζητηθεί η καλύτερη μέθοδος ταξινόμησης των δεδομένων αυτών. Αναλυτικότερα, στην αναζήτηση της καλύτερης μεθόδου ακολουθήθηκε διαδικασία αλλαγών σε βασικές παραμέτρους των ταξινομητών, προκειμένου να εντοπιστεί η καλύτερη μέθοδος που θα καταφέρει να ταξινομήσει αποτελεσματικότερα τα προς μελέτη δεδομένα.

1.2 Ορισμός προβλήματος

Πολλά προβλήματα του πραγματικού κόσμου μπορούν να λυθούν με τη μηχανική μάθηση στους τομείς της αναγνώρισης προτύπων, της επεξεργασίας σημάτων και της ιατρικής διάγνωσης. Η παρούσα μελέτη ασχολείται με διάφορες τεχνικές ταξινόμησης για τη βελτίωση της ακρίβειας σε διαφορετικούς αλγορίθμους.

Ο διαβήτης είναι σοβαρό θέμα δημόσιας υγείας, που, κατά πολλούς, τείνει να πάρει παγκοσμίως επιδημικές διαστάσεις. Γενικότερα, φαίνεται ότι ο επιπολασμός χρόνιων, μη μεταδοτικών ασθενειών αυξάνεται με ανησυχητικό ρυθμό. Περίπου 18 εκατομμύρια άνθρωποι πεθαίνουν κάθε χρόνο από καρδιαγγειακές παθήσεις, για τις οποίες ο διαβήτης και η υπέρταση αποτελούν σημαντικούς παράγοντες προδιάθεσης (Qassim, 2007).

Σύμφωνα με πρόσφατα στατιστικά στοιχεία της Διεθνούς Ομοσπονδίας Διαβήτη (IDF), ο αριθμός των ατόμων με Σακχαρώδη Διαβήτη παγκοσμίως ανέρχεται σε 415 εκατομμύρια

(επιπολασμός 8,8%), ενώ αναμένεται να φτάσει τα 642 εκατομμύρια το 2040. Αντιστοίχως, στην Ευρώπη ο αριθμός των ατόμων με Σακχαρώδη Διαβήτη είναι 59,8 εκατομμύρια (επιπολασμός 9,1%), ενώ αναμένεται να φτάσει τα 71,1 εκατομμύρια το 2040. Με απλά λόγια, ένας στους έντεκα ενήλικες πάσχει σήμερα από διαβήτη, ενώ το 2040 αναμένεται να πάσχει από διαβήτη ένας στους δέκα (ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ Δ/ΝΣΗ ΠΡΩΤΟΒΑΘΜΙΑΣ ΦΡΟΝΤΙΔΑΣ ΥΓΕΙΑΣ, 2016). Κάθε χρόνο 7 εκατομμύρια άνθρωποι αναπτύσσουν διαβήτη, και οι πιο δραματικές αυξήσεις του διαβήτη τύπου 2 εμφανίστηκαν σε πληθυσμούς με ταχείες αλλαγές στον τρόπο ζωής τους.

Ένα άτομο με διαβήτη τύπου 2 είναι δύο έως τέσσερεις φορές πιο πιθανό να πάθει καρδιαγγειακή νόσο, και το 80% των ατόμων με διαβήτη θα πεθάνει από αυτό. Η πρόωρη θνησιμότητα που προκαλείται από το διαβήτη έχει ως αποτέλεσμα 12 έως 14 χρόνια απώλειας ζωής. Ένα άτομο με διαβήτη απαιτείται να δαπανά για ιατρικά έξοδα δύο έως πέντε φορές περισσότερα από αυτά ενός ατόμου χωρίς διαβήτη και ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ) εκτιμά ότι μέχρι 15% των ετήσιων προϋπολογισμών υγείας δαπανώνται για ασθένειες που σχετίζονται με το διαβήτη. Το ετήσιο κόστος άμεσης υγειονομικής περίθαλψης του διαβήτη παγκοσμίως για άτομα ηλικίας 20-79 ετών εκτιμάται ότι ανέρχεται σε 286 δισεκατομμύρια δολάρια (Qassim, 2007).

1.3 Δομή Μεταπτυχιακής Διπλωματικής Εργασίας (ΜΔΕ)

Στο πρώτο κεφάλαιο γίνεται εισαγωγή στην έννοια και τη χρησιμότητα της μηχανικής μάθησης, κυρίως όμως για τη συνεισφορά της στον τομέα της ιατρικής. Γίνεται αναφορά στον ορισμό του προβλήματος του διαβήτη, τις επιπτώσεις του, και τις κοινωνικοοικονομικές του προεκτάσεις. Ενώ παράλληλα, παρουσιάζεται η δομή και η συνεισφορά της ΜΔΕ

Στο δεύτερο κεφάλαιο παρουσιάζεται συνοπτικά η ασθένεια του διαβήτη και τα χαρακτηριστικά των ασθενών.

Στο τρίτο κεφάλαιο εξετάζονται οι διαφορετικές μέθοδοι μάθησης καθώς και οι αλγόριθμοι εποπτευόμενης μάθησης, οι οποίες θα παρουσιαστούν μαζί με τρεις μελέτες που έχουν γίνει για το θέμα που μελετά η συγκεκριμένη ΜΔΕ.

Στο τέταρτο κεφάλαιο γίνεται παρουσίαση του συνόλου δεδομένων που εξετάζονται και ανασκόπηση των χαρακτηριστικών τους, όπως επίσης αναλύεται και η μεθοδολογία που ακολουθήθηκε για να αντιμετωπιστεί το πρόβλημα της πρόβλεψης του διαβήτη. Ένα κεφάλαιο, όσο εκτεταμένο και αν είναι, δεν μπορεί να αποτελέσει πλήρη επισκόπηση των τεχνικών προεπεξεργασίας δεδομένων και αξιολόγηση αποτελεσμάτων εποπτευόμενων αλγορίθμων μηχανικής μάθησης. Παρόλα αυτά, ελπίζουμε ότι τα στοιχεία που αναφέρονται καλύπτουν τα σημαντικότερα θεωρητικά ζητήματα και καθοδηγούν τον ερευνητή σε ενδιαφέροντα συμπεράσματα.

Στο πέμπτο κεφάλαιο, με τη βοήθεια του εργαλείου WEKA, εντοπίζεται η καλύτερη μέθοδος ταξινόμησης των δεδομένων, κατηγοριοποιώντας τα σε προερχόμενα από ασθενείς ή μη ασθενείς.

Στο έκτο κεφάλαιο, επαναλαμβάνεται η διαδικασία του προηγούμενου κεφαλαίου, χρησιμοποιώντας τη γλώσσα προγραμματισμού R.

Τέλος, στο έβδομο και τελευταίο κεφάλαιο, πραγματοποιείται σύγκριση της παρούσας ΜΔΕ με αντίστοιχες μελέτες και παρατίθενται προτάσεις για περαιτέρω διερευνηση του πρόβλημματος της πρόβλεψης της ασθένειας του σακχαρώδη διαβήτη.

1.4 Συνεισφορά Μεταπτυχιακής Διπλωματικής Εργασίας (ΜΔΕ)

Ο διαβήτης είναι η τέταρτη κύρια αιτία θανάτου στις περισσότερες ανεπτυγμένες χώρες. Τα συμπτώματα από το διαβήτη, όπως η στεφανιαία αρτηρία και οι περιφερικές αγγειακές παθήσεις, το εγκεφαλικό επεισόδιο, η διαβητική νευροπάθεια, οι ακρωτηριασμοί, η νεφρική ανεπάρκεια και η τύφλωση, έχουν ως αποτέλεσμα την αύξηση της αναπηρίας για κάθε κοινωνία. Ο διαβήτης είναι ένα από τα προβλήματα υγείας, που αποτελούν πρόκληση για τους αναλυτές δεδομένων στον 21ο αιώνα (Qassim, 2007).

Τόσο η σοβαρότητα του θέματος όσο και η αναγκαιότητα μελετών στο πεδίο αυτό είναι αυταπόδεικτες.

Η **επιστημονική συνεισφορά της ΜΔΕ** μπορεί να συνοψιστεί στα ακόλουθα:

- Εντοπισμός του καλύτερου δυνατού μοντέλου μάθησης. Το μοντέλο αυτό στοχεύουμε να μπορεί να συνεισφέρει θετικά στην πρόβλεψη της διαβητικής νεφροπάθειας πριν από την πραγματική διάγνωση, με υψηλές επιδόσεις πρόβλεψης.

- Πιθανή πηγή για γιατρούς και ερευνητές πληροφοριών σχετικών με την ανάλυση του παράγοντα κινδύνου. Ως εκ τούτου, η αυτόματη, κατά κάποιο τρόπο, και έγκαιρη προειδοποίηση για τους παράγοντες κινδύνου κάθε ασθενούς είναι ικανή να ενισχύσει τη διαδικασία της πρόληψης και επίσης να διευκολύνει το σχεδιασμό αποτελεσματικών και σωστών στρατηγικών θεραπείας.

Ως αποτέλεσμα της έγκαιρης διάγνωσης του διαβήτη, όπως άλλωστε και στο σύνολο των ασθενειών, είναι η αποτελεσματικότερη αντιμετώπιση της νόσου τόσο για την υγεία του ίδιου του ασθενούς όσο και για τα συστήματα υγείας. Είναι σαφές ότι η έγκαιρη αντιμετώπιση οδηγεί σε χαμηλότερες δαπάνες για ασθενείς και κρατικούς μηχανισμούς.

Κεφάλαιο 2: Θεωρητικό Υπόβαθρο Ασθένειας Σακχαρώδους Διαβήτη

2.1 Η ασθένεια του Σακχαρώδους διαβήτη

Ο Σακχαρώδης διαβήτης είναι ένα από τα σημαντικότερα προβλήματα υγείας σε όλον τον κόσμο. Μπορεί να ταξινομηθεί σε δύο κατηγορίες: διαβήτης τύπου 1 και διαβήτης τύπου 2.

Ο διαβήτης τύπου 1 είναι αυτοάνοση ασθένεια. Στην περίπτωση αυτή, το σώμα καταστρέφει τα κύτταρα που είναι απαραίτητα για την παραγωγή ινσουλίνης που θα διευκολύνει την απορρόφηση της γλυκόζης, απαραίτητης στην παραγωγή ενέργειας για τον ανθρώπινο οργανισμό. Αυτός ο τύπος μπορεί να προκληθεί ασχέτως εάν υπάρχει παχυσαρκία. Η παχυσαρκία είναι η αύξηση του δείκτη μάζας σώματος (ΔΜΣ) περισσότερο από το κανονικό. Ο διαβήτης τύπου 1 μπορεί να εμφανιστεί σε παιδική ή εφηβική ηλικία. Ο διαβήτης τύπου 2 συνήθως επηρεάζει τους ενήλικες που είναι παχύσαρκοι. Σε αυτόν τον τύπο, το σώμα αποτυγχάνει να παράγει ινσουλίνη. Ο διαβήτης τύπου 2 εμφανίζεται συχνότερα στις μεσαίες ή ηλικιωμένες ομάδες πληθυσμού. Επιπλέον, υπάρχουν και άλλες αιτίες για το διαβήτη, όπως βακτηριακή ή ιική μόλυνση, τοξικά ή χημικά περιεχόμενα στα τρόφιμα, αυτοάνοση αντίδραση, παχυσαρκία, κακή διατροφή, αλλαγή τρόπου ζωής, διατροφική συνήθεια, ρύπανση του περιβάλλοντος κ.λπ. (Dr. Asir Antony Gnana Singh D., κ.ά, 2017).

Τα άτομα με διαβήτη έχουν αυξημένο κίνδυνο εμφάνισης σειράς σοβαρών προβλημάτων υγείας. Τα σταθερά υψηλά επίπεδα γλυκόζης στο αίμα μπορούν να οδηγήσουν σε σοβαρές ασθένειες που επηρεάζουν την καρδιά και τα αιμοφόρα αγγεία, τα μάτια, τα νεφρά, τα νεύρα και τα δόντια. Επίσης, τα άτομα με διαβήτη έχουν υψηλότερο κίνδυνο εμφάνισης λοιμώξεων. Σε όλες σχεδόν τις αναπτυγμένες χώρες, ο διαβήτης είναι η κύρια αιτία καρδιαγγειακών παθήσεων, τύφλωσης, νεφρικής ανεπάρκειας και ακρωτηριασμού των κάτω άκρων (Pradeep Kandhasamy J., Balamurali S., 2015). Συνεπώς, είναι πολύ σημαντικό να αναπτυχθούν προγνωστικά μοντέλα, χρησιμοποιώντας τους παράγοντες κινδύνου για την εμφάνιση του διαβήτη.

2.1.1 Διαβήτης κατά την περίοδο της εγκυμοσύνης

Τα διαγνωστικά κριτήρια για το διαβήτη σε μη επίτοκες γυναίκες βασίζονται στη σχέση των τιμών του πλάσματος γλυκόζης και του κινδύνου ειδικών διαβητικών μικροαγγειακών επιπλοκών (Κωτσιαντής Σ., 2006).

Όλες οι μορφές διαβήτη αυξάνουν τον κίνδυνο μακροχρόνιων επιπλοκών. Αυτά συνήθως αναπτύσσονται μετά από πολλά χρόνια (δέκα έως είκοσι), αλλά μπορεί να είναι το πρώτο σύμπτωμα σε εκείνους που διαφορετικά διαγνώστηκαν νωρίτερα. Τα κριτήρια για τη διάγνωση του διαβήτη κατά τη διάρκεια της εγκυμοσύνης έχουν δοθεί (Pradeep Kandhasamy J., Balamurali S., 2015) από τον ΠΟΥ το 2006.

Τα κριτήρια αυτά έχουν ως εξής:

- γλυκόζη πλάσματος νηστείας ≥ 7,0 mmol/l (126 mg/dl)
- 2ωρη μεταγευματική γλυκόζη πλάσματος ≥ 11,1 mmol/l(200 mg/dl) μετά τη χορήγηση από στόματος 75 g
- φορτίο γλυκόζης

- τυχαία μέτρηση γλυκόζης πλάσματος ≥ 11,1 mmol/l (200 mg/dl) μετά την παρουσία συμπτωμάτων διαβήτη.

Τα διαγνωστικά κριτήρια για το διαβήτη σε μη εγκύους βασίζονται στη σχέση μεταξύ των τιμών γλυκόζης στο πλάσμα και του κινδύνου ειδικών για διαβήτη μικροαγγειακών επιπλοκών (Pradeep Kandhasamy J., Balamurali S., 2015).

2.1.2 Χαρακτηριστικά ατόμων

Τα χαρακτηριστικά των ατόμων με διαβήτη, έτσι όπως τα ανέδειξε ο παθολόγος - διαβητολόγος, Διευθυντής της Α΄ Παθολογικής Κλινικής του Γενικού Νοσοκομείου Δράμας κ. Σ. Μπακατσέλος στο 21ο Συνέδριο Διαβητολογικής Εταιρείας Βορείου Ελλάδος, με θέμα «Σακχαρώδης Διαβήτης Κύησης», είναι τα ακόλουθα:

- **<u>Χαμηλού κινδύνου</u>,** *για τα οποία δεν απαιτείται οποιαδήποτε δοκιμασία ανοχής γλυκόζης, αν πληρούνται τα εξής:*

Ηλικία <25 ετών

Φυσιολογικό βάρος πριν την εγκυμοσύνη

Αρνητικό ιστορικό διαβήτη σε άτομα πρώτου βαθμού συγγένειας

Αρνητικό ιστορικό αποτυχημένης εγκυμοσύνης

Αρνητικό ιστορικό διαταραχής στη γλυκόζη

Φυσιολογικό βάρος γεννήσεως

Μέλος Εθνικής Ομάδας με χαμηλό επιπολασμό GDM (Διαβήτης Κύησης)

- **<u>Μέσου κινδύνου:</u>** *δοκιμασία ανοχής γλυκόζης στην 24η-28η εβδομάδα με:*

  Δοκιμασία 2 σταδίων:

  τεστ ανοχής με 50 gr γλυκόζης (GTT) η οποία ακολουθείται από OGTT στις επίτοκες που πληρούν τα κριτήρια υποψίας για διάγνωση της GTT (>140 ή >130mg/dl)

  Δοκιμασία 1 σταδίου:

  OGTT σε όλες τις επίτοκες μέσου κινδύνου

- **<u>Υψηλού κινδύνου:</u>** *δοκιμασία ανοχής γλυκόζης ενός ή δύο σταδίων γίνεται το συντομότερο δυνατό αν ένα ή περισσότερα από τα παρακάτω ισχύουν:*

Σοβαρή παχυσαρκία

Ισχυρό οικογενειακό ιστορικό διαβήτη τύπου 2

Προηγούμενο ιστορικό: GDM, IGT (εξασθενημένη ανοχή γλυκόζης), γλυκοζουρία (Μπακατσέλος Σ., 2007).

ΚΕΦΑΛΑΙΟ 3: Μηχανική Μάθηση

3.1 Μέθοδοι μάθησης

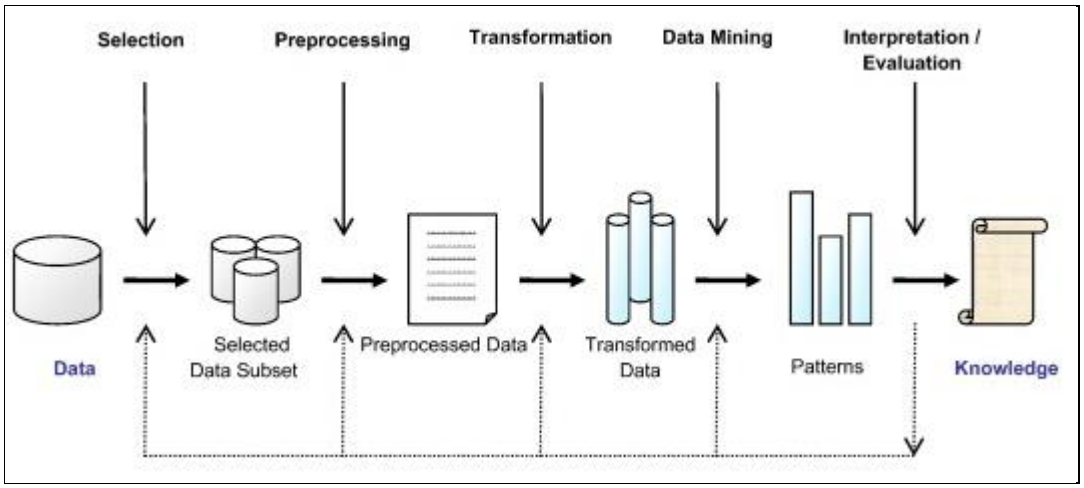
Η εξόρυξη δεδομένων είναι μια από τις διεργασίες της «ανακάλυψης γνώσης σε βάσεις δεδομένων». Ο γενικός στόχος της διαδικασίας εξόρυξης δεδομένων είναι η εξαγωγή πληροφοριών από σύνολο δεδομένων και η μετατροπή τους σε κατανοητή δομή για περαιτέρω χρήση. Αυτή η διαδικασία γίνεται όλο και πιο διαδεδομένη σε όλους τους τομείς της έρευνας στις ιατρικές επιστήμες. Τα προβλήματα εξόρυξης δεδομένων συχνά επιλύονται χρησιμοποιώντας διαφορετικές προσεγγίσεις από την επιστήμη των υπολογιστών, όπως οι πολυδιάστατες βάσεις δεδομένων, η μηχανική μάθηση, η οπτικοποίηση δεδομένων και οι στατιστικές, συμπεριλαμβανομένων των δοκιμασιών υποθέσεων, των ομαδοποιήσεων, της ταξινόμησης και των τεχνικών παλινδρόμησης. Τα τελευταία χρόνια, η εξόρυξη δεδομένων έχει χρησιμοποιηθεί ευρέως στους τομείς της επιστήμης και της μηχανικής, όπως η βιοπληροφορική, η γενετική, η ιατρική και η εκπαίδευση. Η εξόρυξη δεδομένων γενικά τροφοδοτεί κάθε φορά μια διαδικασία μοντελοποίησης, διά μέσου της οποίας μπορούμε να προδιαγράψουμε τη μελλοντική κατάσταση. Η διαδικασία της μοντελοποίησης είναι η διαδικασία με την οποία δημιουργείται ένα μοντέλο με σκοπό την πρόβλεψη αποτελέσματος. Η διαδικασία εξόρυξης δεδομένων για τη διάγνωση του διαβήτη μπορεί να κατηγοριοποιηθεί. Καθώς οι υποκείμενες αρχές και οι τεχνικές που χρησιμοποιούνται για την εξόρυξη δεδομένων για το διαβήτη και τους πάσχοντες από αυτόν μπορεί να διαφέρουν από χώρα σε χώρα, να υπόκεινται σε διαφορετικές νομοθεσίες ή να τροποποιούνται ανά διαφορετικά προγράμματα, η κατηγοριοποίηση της εξόρυξης δεδομένων φαίνεται αναπόφευκτη.

Η μηχανική μάθηση, εξάλλου, είναι ισχυρό εργαλείο τεχνητής νοημοσύνης που μας δίνει τη δυνατότητα να επεξεργαζόμαστε petabytes δεδομένων, με σκοπό να αποκτούν νόημα σε έναν περίπλοκο κόσμο. Είναι πλέον σε συνεχή χρήση, με όλο και περισσότερες εφαρμογές που δεν μπορούμε ενδεχομένως να φανταστούμε. Οι περισσότεροι άνθρωποι πιθανόν ήδη γνωρίζουν ότι ο πάροχος ηλεκτρονικού ταχυδρομείου χρησιμοποιεί αλγόριθμο εκμάθησης μηχανών για τον εντοπισμό ανεπιθύμητων μηνυμάτων. Τα τελευταία χρόνια, εταιρείες όπως η Google και η Tesla κατασκευάζουν συστήματα αυτόματης οδήγησης, τα οποία σύντομα θα μειώσουν ή θα αντικαταστήσουν τους οδηγούς. Επιπλέον, η Amazon και η Braintree χρησιμοποιούν τη μηχανική σε συνδυασμό με άλλα εργαλεία για να σταματήσουν υποκλοπές πιστωτικών καρτών.

Η εξόρυξη είναι μία από τις πιο σημαντικές εφαρμογές της μηχανικής μάθησης. Πολύ συχνά, κατά τη διάρκεια αναλύσεων ή ενδεχομένως κατά την προσπάθεια δημιουργίας σχέσεων μεταξύ πολλαπλών χαρακτηριστικών ή μεταβλητών, οι άνθρωποι είναι επιρρεπείς σε λάθη. Ως αποτέλεσμα αυτού είναι δύσκολη η επίλυση πολλών προβλημάτων. Σε αυτό το σημείο έρχεται η μηχανική μάθηση, η οποία μπορεί να χρησιμοποιηθεί για την επιτυχή επίλυση αυτών των προβλημάτων, βελτιώνοντας την αποτελεσματικότητα των δεδομένων. Το ίδιο σύνολο χαρακτηριστικών αντιπροσωπεύει κάθε εμφάνιση οποιουδήποτε συνόλου δεδομένων που χρησιμοποιείται από τους αλγόριθμους μηχανικής μάθησης. Τα χαρακτηριστικά μπορούν να είναι συνεχή, διακριτά ή και δυαδικής μορφής (Pradeep Kandhasamy J., Balamurali S., 2015).

Η ανακάλυψη της γνώσης σε βάσεις δεδομένων είναι πεδίο που περιλαμβάνει θεωρίες, μεθόδους και τεχνικές, προσπαθώντας να κάνει κατανοητά τα δεδομένα και να οδηγήσει στην εξαγωγή χρήσιμων γνώσεων από αυτές. Όπως απεικονίζεται και παρακάτω (εικόνα 1), είναι διαδικασία πολλαπλών σταδίων (επιλογής, προεπεξεργασίας, μετασχηματισμού,

εξόρυξης δεδομένων, ερμηνείας-αξιολόγησης). Τέλος, το πιο σημαντικό βήμα σε ολόκληρη την αλυσίδα της εικόνας είναι η εξόρυξη των δεδομένων, που αποτελεί παράδειγμα της εφαρμογής αλγορίθμων μηχανικής μάθησης κατά την ανάλυση των δεδομένων (Kavakiotis I., et al., 2017).



Εικόνα 1: Αναπαράσταση των σταδίων για την ανακάλυψη της γνώσης σε βάσεις δεδομένων (Kavakiotis I., et al., 2017).

3.1.1 Εποπτευόμενη Μάθηση

Η πλειονότητα της πρακτικής μηχανικής μάθησης χρησιμοποιεί εποπτευόμενη μάθηση. Ονομάζεται «εποπτευόμενη μάθηση» επειδή η διαδικασία ενός αλγορίθμου, ο οποίος μαθαίνει από το σύνολο δεδομένων κατάρτισης, μπορεί να θεωρηθεί ως δάσκαλος που εποπτεύει τη διαδικασία μάθησης. Γνωρίζοντας τις σωστές απαντήσεις, ο αλγόριθμος κάνει επανειλημμένα προβλέψεις για τα δεδομένα εκπαίδευσης και διορθώνεται από το δάσκαλο. Η μάθηση σταματά όταν ο αλγόριθμος επιτυγχάνει αποδεκτό επίπεδο ως προς την απόδοση.

Στην εποπτευόμενη μάθηση έχουμε μεταβλητές εισόδου (X) και μία μεταβλητή εξόδου (Y), χρησιμοποιώντας έναν αλγόριθμο με σκοπό να ανιχνεύσουμε τη σχέση μεταξύ εισόδου και εξόδου.

$$Y = f(X)$$

Ο στόχος είναι να προσεγγίσουμε τη σχέση των δύο μεταβλητών τόσο καλά ώστε όταν έχουμε νέα δεδομένα εισόδου (X) να μπορούμε να προβλέψουμε τις μεταβλητές εξόδου (Y) για τα δεδομένα αυτά.

Τα εποπτευόμενα μαθησιακά προβλήματα μπορούν να ομαδοποιηθούν περαιτέρω σε προβλήματα παλινδρόμησης και ταξινόμησης.

Παραδείγματα:

Ταξινόμηση: Πρόβλημα κατάταξης έχουμε όταν η μεταβλητή εξόδου είναι κατηγορική τιμή, όπως άσπρο-μαύρο ή ασθενείς-μη ασθενείς.
Παλινδρόμηση: Πρόβλημα παλινδρόμησης έχουμε όταν η μεταβλητή εξόδου είναι πραγματική τιμή, όπως δολάρια ή βάρος.

Ορισμένοι τύποι προβλημάτων που στηρίζονται στην ταξινόμηση και στην παλινδρόμηση περιλαμβάνουν πρόβλεψη κατηγοριοποίησης και χρονοσειρών.

Μερικά δημοφιλή παραδείγματα εποπτευόμενων αλγορίθμων μηχανικής μάθησης είναι:

- Γραμμική παλινδρόμηση για προβλήματα παλινδρόμησης.
- Τυχαία Δάση για προβλήματα ταξινόμησης και παλινδρόμησης.
- Μηχανές διανυσμάτων στήριξης για προβλήματα ταξινόμησης.

3.1.2 Μη Εποπτευόμενη Μάθηση

Στη μη εποπτευόμενη μάθηση έχουμε μόνο δεδομένα εισόδου (X) και καμία αντίστοιχη μεταβλητή εξόδου.

Ο στόχος της μάθησης χωρίς επίβλεψη είναι να μοντελοποιήσει την υποκείμενη δομή ή την κατανομή στα δεδομένα, προκειμένου ο αλγόριθμος να μάθει όσο το δυνατόν περισσότερα για αυτά.

Η μάθηση αυτή ονομάζεται «χωρίς επίβλεψη» διότι, σε αντίθεση με την εποπτευόμενη μάθηση, που αναφέρθηκε, δεν υπάρχουν σωστές απαντήσεις και δεν υπάρχει δάσκαλος. Οι αλγόριθμοι αφήνονται μόνοι τους να ανακαλύψουν και να παρουσιάσουν τη δομή των δεδομένων.

Τα προβλήματα μη εποπτευόμενης μάθησης μπορούν να ομαδοποιηθούν περαιτέρω σε προβλήματα ομαδοποίησης (clustering) και συσχέτισης (association).

> Ομαδοποίηση: Πρόβλημα ομαδοποίησης είναι αυτό στο οποίο θέλουμε να ανακαλύψουμε τις εγγενείς ομαδοποιήσεις των δεδομένων, όπως ομαδοποίηση πελατών με αγοραστική συμπεριφορά.
> Συσχέτιση: Πρόβλημα συσχέτισης είναι αυτό στο οποίο θέλουμε να ανακαλύψουμε κανόνες που περιγράφουν μεγάλα τμήματα των δεδομένων μας, όπως οι άνθρωποι που αγοράζουν το X επίσης τείνουν να αγοράζουν το Y.

Μερικά δημοφιλή παραδείγματα αλγορίθμων μάθησης χωρίς επίβλεψη είναι:

- K-means, για προβλήματα ομαδοποίησης.
- Αλγόριθμος A-priori, για προβλήματα συσχέτισης.

3.1.3 Ημιεποπτευόμενη Μάθηση

Σε προβλήματα με μεγάλη ποσότητα δεδομένων εισόδου (X), όπου μόνο μερικά από τα δεδομένα είναι επισημασμένα (Y), ονομάζονται ημιεποπτευόμενα μαθησιακά προβλήματα.

Αυτά τα προβλήματα βρίσκονται μεταξύ της μάθησης υπό επίβλεψη και της μάθησης χωρίς επίβλεψη.

Καλό παράδειγμα είναι αρχείο φωτογραφιών, όπου μόνο μερικές από τις εικόνες είναι επισημασμένες (π.χ. σκύλος, γάτα, πρόσωπο), ενώ η πλειοψηφία δεν έχει επισημανθεί.

Πολλά προβλήματα μηχανικής μάθησης στον πραγματικό κόσμο ανήκουν στην παραπάνω κατηγορία. Αυτό μπορεί να οφείλεται στο γεγονός ότι η επισήμανση δεδομένων είναι δαπανηρή ή χρονοβόρα, καθώς και στο ότι μπορεί να απαιτείται πρόσβαση σε εμπειρογνώμονες του τομέα, η οποία δεν είναι πάντοτε εφικτή. Τα μη επισημασμένα

δεδομένα είναι φθηνά και προσβάσιμα, ενώ μπορούν να γίνουν αντικείμενο συλλογής και αποθήκευσης (Brownlee J., 2016).

3.2 Αλγόριθμοι Εποπτευόμενης Μάθησης

Είναι σημαντικό στην προσπάθεια να επιλυθεί κάποιο πρόβλημα να υπάρχει αντιπροσωπευτικό δείγμα αλγορίθμων για εκμάθηση.

Καλός κανόνας είναι «μερικοί αλγόριθμοι από κάθε τύπο». Για παράδειγμα, στην περίπτωση της δυαδικής ταξινόμησης που καλούμαστε να επιλύσουμε, θα παρουσιαστούν αλγόριθμοι από τους εξής τρείς τύπους:

- Δένδρα Απόφασης: J48
- Σύνολα Δένδρων: Τυχαία Δάση
- Μη Γραμμικές Μέθοδοι: Τροφοδοτούμενα προς τα εμπρός Δίκτυα Πολλαπλών Στρωμάτων, Λογιστική Παλινδρόμηση, Μηχανές Διανυσμάτων Υποστήριξης (SVM), k Πλησιέστεροι Γείτονες (Knn) και Μπεϊζιανός Αλγόριθμος (Naive Bayes).

Τέλος, θα μπορούσε να παρουσιαστεί και η μέθοδος της γραμμικής παλινδρόμισης, η οποία όμως δεν εξετάζετε στην παρούσα ΜΔΕ καθώς δεν εμφάνιζε ικανοποιητικά αποτελέσματα στο εξετασθέν σύνολο δεδομένων, με σκοπό να καλύφθούν και οι Γραμμικές Μέθοδοι.

3.2.1 Αλγόριθμος J48

Ο αλγόριθμος J48 (ή αλλιώς C4.5) ανήκει στη μεγάλη κατηγορία των αλγορίθμων ταξινόμησης οι οποίοι δημιουργούν δεντρικά μοντέλα ταξινομητών (δένδρα απόφασης). Αποτελεί απόγονο του αλγορίθμου ID3. Ένα δένδρο απόφασης αποτελείται από κόμβους που αντιστοιχούν σε κάποιο χαρακτηριστικό του συνόλου εκπαίδευσης ο καθένας και διακρίνονται στους εξής:

- Ρίζα: κόμβος ο οποίος βρίσκεται στην κορυφή του δένδρου και χωρίζει το σύνολο εκπαίδευσης σε δύο ή περισσότερα υποσύνολα.
- Εσωτερικοί κόμβοι: ενδιάμεσοι κόμβοι οι οποίοι με τη σειρά τους χωρίζουν το κάθε υποσύνολο του υποδένδρου σε μικρότερα υποσύνολα.
- Φύλλα: αποτελεί τον τερματικό κόμβο και αντιπροσωπεύει μια κλάση από το διακριτό σύνολο κλάσεων του συνόλου εκπαίδευσης.

Όλοι οι κόμβοι, εκτός από τα φύλλα, έχουν εξερχόμενες ακμές, οι οποίες αντιστοιχούν σε μια συνθήκη βάσει της οποία γίνεται η διάσπαση των δεδομένων. Η συνθήκη αυτή ονομάζεται «συνθήκη διάσπασης». Το βασικό ποσοτικό μέτρο που χρησιμοποιείται για επιλογή των διαχωριστών είναι το κέρδος πληροφορίας, το οποίο βασίζεται στην εντροπία πληροφορίας. Έχουν προταθεί αρκετές παραλλαγές του C4.5, όπως C4.5-no-pruning, C4.5-rules, οι οποίες προσπαθούν να βελτιώσουν την επίδοση του αλγορίθμου εκτελώντας διάφορες επιπρόσθετες λειτουργίες (π.χ. κλάδεμα δένδρου) (Qassim, 2007) (Han J., KanberM. Pei J., 2012).

3.2.2 Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση αποτελεί στην ουσία μοντέλο ταξινόμησης των τιμών μιας μεταβλητής εξόδου Y με βάση τη θεωρία των πιθανοτήτων. Στο μοντέλο αυτό, όπου η μεταβλητή Y συνήθως έχει δυαδικό χαρακτήρα (λαμβάνει δύο τιμές), στοχεύεται η πρόβλεψη του αποτελέσματός της από πλήθος προβλεπτικών μεταβλητών που μπορεί να είναι ονομαστικές, τακτικές ή ποσοτικές.

Πολλές φορές, η λογιστική παλινδρόμηση συγχέεται λανθασμένα με τη γραμμική παλινδρόμηση. Η σημαντικότερη διαφοροποίηση μεταξύ τους βασίζεται στη φύση της επιλεγμένης μεταβλητής εξόδου. Στην πρώτη η μεταβλητή εξόδου μπορεί να είναι κατηγορική, ενώ στη δεύτερη είναι αποκλειστικά ποσοτική. Κατά την κλασική γραμμική παλινδρόμηση, η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο των ελάχιστων τετραγώνων, κατά τη λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο του λόγου πιθανοφάνειας, δηλαδή επιλέγονται οι πιο πιθανοφανείς τιμές των παραμέτρων, προκειμένου να οδηγήσουν σε κάποια αποτελέσματα.

Η λογιστική παλινδρόμηση επινοήθηκε ως εναλλακτική επιλογή της γραμμικής διακριτικής ανάλυσης για την ταξινόμηση των στοιχείων της εξαρτημένης, με ευρεία απήχηση σε πολλά διαφορετικά επιστημονικά πεδία και κυρίως στην ιατρική και στις κοινωνικές επιστήμες (Agresti A, 1996). Χαρακτηριστικά, χρησιμοποιείται στην πρόβλεψη:

• εμφάνισης ή μη μιας νόσου (π.χ. διαβήτης) από σύνολο διαφορετικών χαρακτηριστικών του πάσχοντος ατόμου (ηλικία, φύλο, αιματολογικά, ηλεκτροκαρδιογράφημα, κ.λπ.),

• επιλογής ενός πολιτικού κόμματος με βάση την καταγραφή των δημογραφικών στοιχείων των πολιτών (ηλικία, φύλο, φυλή, τόπος διαμονής, εισόδημα προηγούμενης ψηφοφορίας,κ.λπ.),

• πιθανότητας αποτυχίας στη διεργασία παραγωγής προϊόντος σε εργοστάσιο τροφίμων,

• πρόθεσης αγοράς αγαθού από καταναλωτή (έρευνα αγοράς),

• πιθανότητας αθέτησης δανειολήπτη ως προς την αποπληρωμή δανείου (Tabaei B., Herman W., 2002).

3.2.3 Τυχαία Δάση

Ο αλγόριθμος αυτός αναπτύχθηκε από τους Leo Breiman και Adele Cutler το 2001. Είναι από τους δημοφιλέστερους αλγορίθμους στην κατηγορία του, κυρίως για την ταχύτητα αλλά και για την ακρίβειά του. Σύμφωνα με τους δημιουργούς του (NongyaoNai-arun, RungruttikarnMoungmai, 2015) (Sittidech P., Nai-arun N., 2014), προσφέρει:

- Καλύτερη ακρίβεια μεταξύ των υπαρχόντων αλγορίθμων.
- Η ταχύτητά του είναι πολύ καλή, ακόμα και σε πολύ μεγάλα σύνολα δεδομένων εκπαίδευσης.
- Μπορεί να είναι αποδοτικός σε πάρα πολύ μεγάλο αριθμό χαρακτηριστικών (ακόμα και σε χιλιάδες).
- Δίνει εκτίμηση για το ποια χαρακτηριστικά είναι τα πιο σημαντικά στην κατηγοριοποίηση.
- Δεν χρειάζεται τη χρήση διαφορετικού συνόλου δεδομένων για τον έλεγχο ακριβείας, δεν είναι δηλαδή απαραίτητη η διασταυρούμενη επικύρωση (cross-validation), καθώς η εκτίμηση του λάθους γενίκευσης γίνεται από τον ίδιο τον αλγόριθμο κατά την εκτέλεσή του.
- Μπορεί να είναι αποδοτικός σε ελλιπή δεδομένα.
- Δεν παρουσιάζει φαινόμενα υπερεκπαίδευσης.

3.2.4 Τροφοδοτούμενα προς τα εμπρός δίκτυα πολλαπλών στρωμάτων (Multilayer Perceptron)

Αυτή η κατηγορία δικτύων περιλαμβάνει περισσότερα από ένα κρυφά στρώματα υπολογιστικών νευρώνων, τα οποία ονομάζονται και «κρυφοί νευρώνες». Κρυφά στρώματα είναι εκείνα τα οποία βρίσκονται μεταξύ του επιπέδου των κόμβων εισόδου καθώς και του στρώματος εξόδου. Με την προσθήκη επιπλέον στρωμάτων νευρώνων, ο αλγόριθμος μπορεί να οδηγηθεί στην πραγματοποίηση πιο σύνθετων υπολογισμών και άρα στην εξαγωγή υψηλότερης τάξης αποτελεσμάτων. Η ικανότητα αυτή των δικτύων είναι ιδιαίτερα χρήσιμη όταν οι κόμβοι στο επίπεδο εισόδου είναι πολλοί. Η λειτουργία ενός δικτύου πολλαπλών στρωμάτων έχει ως εξής:

Αρχικά το επίπεδο με τους κόμβους εισόδου τροφοδοτεί το πρώτο κρυφό στρώμα με το διάνυσμα εισόδου, ενεργοποιώντας το. Έπειτα πραγματοποιούνται οι υπολογισμοί στους νευρώνες του πρώτου κρυφού στρώματος και παράγονται σήματα εξόδου. Αυτά τα σήματα εξόδου χρησιμοποιούνται ως είσοδοι στο δεύτερο κρυφό στρώμα νευρώνων που, με τη σειρά του, θα τροφοδοτήσει το επόμενο στρώμα. Η διαδικασία συνεχίζεται με τον ίδιο τρόπο μέχρι το σήμα να φτάσει στο στρώμα εξόδου, από το οποίο εξάγεται η απόκριση του δικτύου στο σήμα εισόδου που του δόθηκε μέσω των κόμβων στο επίπεδο εισόδου. Αυτά τα δίκτυα καλούνται Πολυστρωματικοί Αισθητήρες (Multilayer Perceptron), όπως αναφέρει και ο Bishop (Bishop C., 1995).

3.2.5 Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Είναι ένας εποπτευόμενος αλγόριθμος εκμάθησης μηχανής, ο οποίος μπορεί να χρησιμοποιηθεί για προκλήσεις ταξινόμησης και παλινδρόμησης. Ωστόσο, χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης. Σε αυτόν τον αλγόριθμο, σχεδιάζουμε κάθε στοιχείο δεδομένων ως σημείο σε n-διάστατο χώρο (όπου n είναι ο αριθμός των χαρακτηριστικών), με την αξία κάθε χαρακτηριστικού να είναι η τιμή συγκεκριμένης συντεταγμένης. Στη συνέχεια, πραγματοποιούμε την ταξινόμηση βρίσκοντας το υπερ-επίπεδο που διαφοροποιεί πολύ καλά τις δύο κατηγορίες.

Το SVM έχει μια τεχνική που λέγεται kernel trick. Αυτή η τεχνική παίρνει ως είσοδο έναν μικρής διάστασης χώρο και τον αναγάγει σε μεγαλύτερης διάστασης. Για παράδειγμα, μετατρέπει ένα μη διαχωρίσιμο πρόβλημα σε διαχωρίσιμο πρόβλημα, ανάγοντάς το σε μεγαλύτερης διάστασης επίπεδο. Με απλά λόγια, κάνει μερικούς εξαιρετικά περίπλοκους μετασχηματισμούς δεδομένων και, στη συνέχεια, ανακαλύπτει τη διαδικασία διαχωρισμού των δεδομένων με βάση τις ετικέτες ή τις εξόδους που έχουν οριστεί (Sunil R., 2017).

Η εκπαίδευση των SVM γίνεται με αργό ρυθμό, ειδικά σε μεγάλα προβλήματα. Οι αλγόριθμοι εκπαίδευσής τους είναι πολύπλοκοι και μερικές φορές είναι δύσκολο να εφαρμοστούν.

Ο αλγόριθμος SVM στο weka ονομάζεται SMO. Αυτή η εφαρμογή αντικαθιστά όλες τις ελλιπείς τιμές και μετατρέπει τα ονομαστικά χαρακτηριστικά σε δυαδικά. Κανονικοποιεί επίσης όλα τα χαρακτηριστικά από προεπιλογή.

Σε έναν SVM ψάχνουμε για δύο πράγματα, ένα διαχωρισμό με το μεγαλύτερο ελάχιστο περιθώριο και ταυτόχρονα μία γραμμή που θα διαχωρίζει σωστά όσο το δυνατόν περισσότερες περιπτώσεις. Το πρόβλημα είναι ότι δεν μπορούμε πάντοτε να πετύχουμε και τα δύο. Η παράμετρος c καθορίζει το δεύτερο διαχωρισμό.

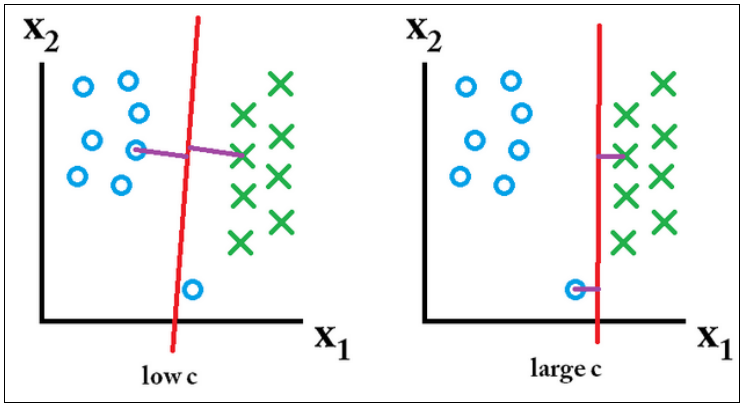
Η σταθερά C είναι θετικός αριθμός και εκφράζει την εξισορρόπηση σημαντικότητας μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης των λάθος ταξινομήσεων. Έτσι, αφού το σφάλμα είναι ανάλογο με το C και τις μεταβλητές χαλαρότητας, όσο υψηλότερη

είναι η τιμή του C, τόσο περισσότερο θα τιμωρηθούν τα λάθος ταξινομημένα σημεία και τα σημεία στο εσωτερικό του περιθωρίου. Προκύπτει δηλαδή ότι:

• Αν C $\rightarrow$0 τότε σημαίνει ότι αγνοούνται οι μεταβλητές χαλαρότητας

• Αν C $\rightarrow +\infty$ τότε προσδίδεται μεγάλη ποινή στις λάθος ταξινομήσεις. Η τιμή της C είναι επιλογή του χρήστη και γίνεται μετά την αξιολόγηση διάφορων δοκιμών μέσω της διαδικασίας επικύρωσης. (Παπαποστόλου Σ., 2017)

Στη παρακάτω εικόνα αριστερά έχουμε ένα χαμηλό c που μας δίνει ένα αρκετά μεγάλο ελάχιστο περιθώριο (μοβ). Ωστόσο, αυτό απαιτεί να αγνοήσουμε την απόκλιση που δεν καταφέραμε να την ταξινομήσουμε σωστά. Στην ίδια εικόνα δεξιά έχουμε υψηλό c (Kent Munthe Caspersen, 2015).



Εικόνα 2: Μεταβολή τιμής παραμέτρου διαχωρισμού περιπτώσεων στον SVM (Kent Munthe Caspersen, 2015).

3.2.6 Μπεϊζιανός Αλγόριθμος

Είναι ένας ταξινομητής που χρησιμοποιεί το θεώρημα του Bayes. Το θεώρημα αυτό πήρε το όνομά του από τον Rev. Thomas Bayes. Λειτουργεί με υπό όρους πιθανότητα. Σύμφωνα με αυτό, υποδεικνυόμενη πιθανότητα είναι η πιθανότητα κάτι να συμβεί, δεδομένου ότι κάτι άλλο έχει ήδη συμβεί. Χρησιμοποιώντας την πιθανότητα υπό όρους, μπορούμε να υπολογίσουμε την πιθανότητα γεγονότος χρησιμοποιώντας τις προηγούμενες γνώσεις του.

Παρακάτω είναι ο τύπος για τον υπολογισμό της υπό όρους πιθανότητας:

$$P(H|\mathrm{E}) = \frac{P(E|\mathrm{H}) * P(H)}{P(E)}$$

όπου

- P (H) είναι η πιθανότητα της υπόθεσης H να είναι αληθής. Αυτό είναι γνωστό ως η προηγούμενη πιθανότητα (priori probability).
- P (E) είναι η πιθανότητα των στοιχείων (ανεξάρτητα από την υπόθεση).
- P (E|H) είναι η πιθανότητα των αποδεικτικών στοιχείων, δεδομένου ότι η υπόθεση είναι αληθής.

Προβλέπει πιθανότητες συμμετοχής για κάθε κλάση, όπως η πιθανότητα ότι μια δεδομένη εγγραφή ή ένα σημείο δεδομένων ανήκει σε συγκεκριμένη κλάση. Η κατηγορία με την

υψηλότερη πιθανότητα θεωρείται ως η πιο πιθανή κατηγορία. Αυτό είναι επίσης γνωστό ως Μέγιστο εκ των Υστέρων (MAP).

Το MAP για μια υπόθεση είναι:

$MAP(H) = \max(P(H \mid E)) = \max((P(E \mid H) * P(H))/P(E) = \max(P(E \mid H)*P(H))$

Όπου :

- $P(H \mid E)$,είναι η πιθανότητα της υπόθεσης, δεδομένου ότι υπάρχουν τα στοιχεία.
- P(E) είναι πιθανότητα απόδειξης και χρησιμοποιείται για την κανονικοποίηση του αποτελέσματος. Παραμένει το ίδιο, έτσι η αφαίρεση δεν θα το επηρεάσει.

Ο ταξινομητής υποθέτει ότι όλες οι λειτουργίες δεν σχετίζονται μεταξύ τους. Η παρουσία ή η απουσία χαρακτηριστικού δεν επηρεάζει την παρουσία ή την απουσία οποιουδήποτε άλλου (Saxena R., 2017).

Η τεχνική Κατηγοριοποίησης του Μπεϊζιανού Αλγορίθμου είναι ιδιαίτερα κατάλληλη όταν η συχνότητα των εισροών είναι υψηλή. Παρά του ότι είναι απλή η λειτουργία του, ο αλγόριθμος αυτός συχνά μπορεί να υπερβαίνει τις πιο εξελιγμένες μεθόδους ταξινόμησης (TIBCO Statistica™, 2018).

Πλεονεκτήματα και μειονεκτήματα:

Πλεονεκτήματα

- Είναι γρήγορος και εξαιρετικά κλιμακωτός αλγόριθμος.
- Μπορεί να χρησιμοποιηθεί για ταξινόμηση τόσο σε δυαδικές όσο και σε περισσότερες κλάσεις.
- Παρέχει διαφορετικούς τύπους αλγορίθμων, όπως GaussianNB, MultinomialNB, BernoulliNB.
- Είναι απλός αλγόριθμος που εξαρτάται από το να κάνεις δέσμη μετρήσεων.
- Παρέχει μεγάλη επιλογή για προβλήματα ταξινόμησης κειμένου. Είναι δημοφιλής στην ταξινόμηση μηνυμάτων spam.
- Μπορεί εύκολα να εκπαιδευτεί και σε μικρό σύνολο δεδομένων.

Μειονεκτήματα

Θεωρεί ότι όλα τα χαρακτηριστικά είναι ασυσχέτιστα, οπότε δεν μπορεί να μάθει τη σχέση μεταξύ τους. Μπορεί να μάθει μεμονωμένα χαρακτηριστικά γνωρίσματα χωρίς όμως να μπορεί να καθορίσει τη μεταξύ τους σχέση (Saxena R., 2017).

3.2.7 k Πλησιέστεροι Γείτονες (kNN)

Ο αλγόριθμος k Πλησιέστεροι Γείτονες (kNN), επίσης γνωστός ως Collaborative Filtering ή Instance-Based Learning, είναι χρήσιμη τεχνική εξόρυξης δεδομένων που μας επιτρέπει να χρησιμοποιήσουμε τις προηγούμενες εμφανίσεις δεδομένων με γνωστές τιμές εξόδου για να προβλέψουμε άγνωστη τιμή εξόδου νέου στιγμιότυπου δεδομένων.

Ο Αλγόριθμος kNN είναι μία πολύ γνωστή και ευρεία χρησιμοποιούμενη τεχνική κατηγοριοποίησης που στηρίζεται στη χρήση μέτρων βασισμένων στην απόσταση. Η κεντρική ιδέα είναι πως η τιμή της συνάρτησης-στόχου για ένα νέο στιγμιότυπο βασίζεται αποκλειστικά και μόνο στις αντίστοιχες τιμές των k πιο «κοντινών» στιγμιότυπων εκπαίδευσης, τα οποία αποτελούν τους «γείτονές» του.

Η περιγραφή του φαίνεται να είναι παρόμοια τόσο με την Παλινδρόμηση όσο και με την Ταξινόμηση. Ωστόσο, η Παλινδρόμηση μπορεί να χρησιμοποιηθεί μόνο για αριθμητικές εξόδους. Αυτό διαφοροποιεί αμέσως την Παλινδρόμηση από τον αλγόριθμο kNN. Επιπλέον, η Ταξινόμηση στα Δέντρα Απόφασης χρησιμοποιεί ως παράδειγμα στιγμιότυπο δεδομένων για να δημιουργήσει Δέντρο, το οποίο διατρέχουμε για να βρούμε την απάντησή μας. Αυτό μπορεί να είναι δύσκολο σε ορισμένες περιπτώσεις. Για παράδειγμα, μια εταιρεία, όπως η Amazon, και ένα πολύ σύνηθες χαρακτηριστικό «Πελάτες που αγόρασαν το Χ, επίσης αγόρασαν και το Υ». Εάν η Amazon δημιουργήσει ένα Δέντρο Ταξινόμησης, πόσους κλάδους και κόμβους θα μπορούσε να έχει; Υπάρχουν ίσως μερικές εκατοντάδες χιλιάδες προϊόντα. Πόσο μεγάλο θα ήταν αυτό το Δέντρο; Πόσο ακριβές θα ήταν ένα μεγάλο Δέντρο; Ακόμα κι αν κατέληγε κάπου, πιθανόν θα εμφάνιζε ότι υπήρχαν μόνο τρία προϊόντα. Η σελίδα της Amazon θέλει να έχει δώδεκα προϊόντα σε αυτό για να τα συστήσει. Οπότε είναι ένα κακό μοντέλο εξόρυξης δεδομένων για αυτή την περίπτωση.

Ο αλγόριθμος kNN διορθώνει όλα τα παραπάνω προβλήματα με πολύ αποτελεσματικό τρόπο, ειδικά σε περιπτώσεις όπως το παράδειγμα που χρησιμοποιήθηκε για την Amazon. Δεν περιορίζεται σε οποιονδήποτε αριθμό συγκρίσεων. Είναι εξίσου κλιμακωτός για βάση δεδομένων είκοσι πελατών, όπως και για βάση δεδομένων είκοσι εκατομμυρίων πελατών. Μπορούμε ακόμα να ορίσουμε τον αριθμό των αποτελεσμάτων που θέλουμε να βρούμε (Abernethy M., 2010).

Ο Ταξινομητής είναι ευρέως διαθέσιμος, δεδομένου ότι είναι μη παραμετρικός, δηλαδή δεν κάνει καμία υποκειμενική υπόθεση για την κατανομή των δεδομένων, σε αντίθεση με άλλους αλγορίθμους, οι οποίοι υποθέτουν κανονική κατανομή των δεδομένων (GeeksforGeeks, n.d.).

Ο kNN στο εργαλείο WEKA ονομάζεται IBk (το IB σημαίνει Instance Based, δηλαδή Βάσει Υποθέσεων, και το k επιτρέπει να καθοριστεί ο αριθμός των Γειτόνων που θα εξετάσουμε).

3.3 Θεωρίες και Προσεγγίσεις του προβλήματος μέσα από άλλες μελέτες

Η προσέγγιση της σχετικής με το θέμα της ΜΔΕ βιβλιογραφίας δείχνει ότι έχουν γίνει αξιόλογες προσπάθειες. Η αξιολόγηση της αποτελεσματικότητας αλγορίθμων για τον εντοπισμό και για την πρόβλεψη ασθενειών φαίνεται να είναι πρόβλημα που απασχόλησε και εξακολουθεί να απασχολεί πολλούς ερευνητές.

Τρεις μελέτες που ασχολήθηκαν με το ίδιο ακριβώς θέμα της ΜΔΕ, δηλαδή με τον εντοπισμό κατάλληλου μοντέλου που θα προβλέπει με μεγάλη ακρίβεια την ασθένεια του σακχαρώδους διαβήτη, παρουσιάζονται συνοπτικά παρακάτω.

Και στις τρεις μελέτες που θα παρουσιαστούν τα δείγματα δεδομένων έχουν ληφθεί από το χώρο αποθήκευσης συνόλων δεδομένων μηχανικής μάθησης «UCI» (UCI Machine Learning repository). Συγκεκριμένα χρησιμοποιούν το σύνολο δεδομένων «PIMA Indians», το οποίο και θα παρουσιαστεί λεπτομερώς στο κεφάλαιο 4, καθώς είναι το ίδιο σύνολο δεδομένων που εξετάζει η παρούσα εργασία.

Αυτός ήταν και ο βασικότερος λόγος επιλογής των μελετών αυτών, προκειμένου να μπορεί να γίνει αντιπαραβολή των αποτελεσμάτων τους με αυτά της παρούσας εργασίας.

3.3.1 Μεθοδολογία και αποτελέσματα των άλλων μελετών

Πρώτη μελέτη

Κύριος στόχος της μελέτης των Nabi Meraj, Wahid Abdul και Kumar Pradeep με τίτλο *Performance Analysis of Classification Algorithms in Predicting Diabetes,*[1] ήταν η σύγκριση των παρακάτω ταξινομητών ως προς τη δυνατότητά τους να κατηγοριοποιούν ασθενείς με σακχαρώδη διαβήτη:

- Δένδρα Απόφασης (J48)
- Μπεϊζιανός Αλγόριθμος
- Λογιστική Παλινδρόμηση
- Τυχαία Δάση

Οι επιδόσεις των αλγορίθμων μετρήθηκαν και η σύγκριση των τεσσάρων αλγορίθμων έγινε ως προς την ακρίβειά τους.

Η προεπεξεργασία των δεδομένων επικεντρώθηκε στο πώς να αντιμετωπιστούν οι ελλιπείς τιμές δεδομένων. Η μέθοδος που επιλέχθηκε για να καλυφθούν τα ελλιπή δεδομένα ήταν η χρήση του μέσου όρου.

Τα προεπεξεργασμένα δεδομένα χρησιμοποιήθηκαν για την εκπαίδευση και για τον έλεγχο του κάθε ταξινομητή. Σε όλες τις περιπτώσεις, το 70% των δεδομένων χρησιμοποιήθηκε από τον ταξινομητή για την εκπαίδευσή του και το υπόλοιπο 30% για την αξιολόγησή του.

Στο παρακάτω πίνακα απεικονίζονται τα αποτελέσματα της ακρίβειας των εξεταζόμενων ταξινομητών:

Πίνακας 1: Πίνακας Αποτελεσμάτων Πρώτης Μελέτης

Ταξινομητές	Ακρίβεια
Μπεϊζιανός Αλγόριθμος	76,95%
Λογιστική Παλινδρόμηση	**80,43%**
J48	76,52%
Τυχαία Δάση	76,52%

Παρατηρείται ότι τα καλύτερα αποτελέσματα επιτυγχάνονται με τη χρήση της Λογιστικής Παλινδρόμησης.

Δεύτερη Μελέτη

Στην εργασία των Amit kumar Dewangan και Pragati Agrawal *Classification of Diabetes Mellitus Using Machine Learning Techniques* (Amit kumar Dewangan, Pragati Agrawal, 2015) προτείνεται χρήση υβριδικού μοντέλου που συνδυάζει δύο ταξινομητές, το Τροφοδοτούμενο προς τα εμπρός Δίκτυο Πολλαπλών Στρωμάτων και τον Μπεϊζιανό Αλγόριθμο, με σκοπό τη μέτρηση της ακρίβειας, της ευαισθησίας και της ειδικότητας των μετρήσεων για τη διάγνωση του σακχαρώδους διαβήτη. Επιπλέον, αφαιρεί πλήρως δύο

[1] *Nabi, Meraj; Wahid, Abdul; Kumar, Pradeep 'Performance* Analysis of Classification Algorithms in Predicting Diabetes', *International Journal of Advanced Research in Computer Science,* 8, 3, 2017, σσ. 456-461.

απο τα οκτώ χαρακτηριστικά του συνόλου δεδομένων, και συγκεκριμένα τη διαστολική αρτηριακή πίεση (σε mm) και τη γενεαλογική λειτουργία του διαβήτη, κρίνοντας ότι επηρεάζουν αρνητικά την εκπαίδευση των ταξινομητών.

Τα αποτελέσματα αυτής της συνδυαστικής χρήσης των δύο ταξινομητών είναι τα ακόλουθα:

Πίνακας 2 : Πίνακας Αποτελεσμάτων Δεύτερης Μελέτης

Ακρίβεια	81,89%
Ευαισθησία	**64,10%**
Ειδικότητα	90,90%

Μολονότι επιτυγχάνεται ικανοποιητικό ποσοστό ακρίβειας πρόβλεψης, παρατηρούμε πολύ χαμηλό ποσοστό ευαισθησίας, που σημαίνει ότι αυτό το μοντέλο ταξινόμησης δεν είναι ικανό να ταξινομεί σωστά τους ασθενείς.

Τρίτη Μελέτη

Η μελέτη των Amatul Zehra, Tuty Asmawaty, M.A MAznan *A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset* (Amatul Zehra,Tuty Asmawaty,M.A M. Aznan, 2013) συγκρίνει την ακρίβεια ταξινόμησης τόσο σε μη επεξεργασμένα όσο και σε προεπεξεργασμένα δεδομένα.

Σχετικά με τα δεδομένα που εμφάνιζαν μηδενική τιμή, αποφασίστηκε να καταργηθούν όλες αυτές οι εγγραφές. Επομένως, όλες οι ελλιπείς τιμές έχουν απαλειφθεί. Στη συνέχεια, ακολουθήθηκε η διαδικασία της διακριτοποίησης (discretization). Η διακριτοποίηση των δεδομένων ορίζεται ως διαδικασία μετατροπής των συνεχών τιμών σε πεπερασμένο σύνολο διαστημάτων. Τα αποτελέσματα των αλγορίθμων σε επεξεργασμένα και σε μη επεξεργασμένα δεδομένα φαίνονται στον πίνακα που ακολουθεί:

Πίνακας 3: Πίνακας Αποτελεσμάτων Τρίτης Μελέτης

Ταξινομητές	Ακρίβεια σε δεδομένα χωρίς προ-επεξεργασία	Ακρίβεια σε προ-επεξεργασμένα δεδομένα
Μπεϊζιανός Αλγόριθμος	76,30%	80,30%
Multilayer Perceptron	75,39%	81%
Decision Table	71,22%	85,20%
J48	73,82%	80%
Simple Cart	75,13%	79,60%

Τα αποτελέσματα έδειξαν σαφώς ότι τα προεπεξεργασμένα δεδομένα παρέχουν καλύτερη ακρίβεια ταξινόμησης.

Η μελέτη εστίασε στη διερεύνηση ενός μοντέλου πρόβλεψης και διάγνωσης του διαβήτη μακροπρόθεσμα. Τα περισσότερα από τα μοντέλα που έχουν αναπτυχθεί για την πρόβλεψη και τη διάγνωση του διαβήτη φαίνεται να λειτουργούν βραχυπρόθεσμα. Ωστόσο, σύμφωνα με τη μελέτη, σπανίως υπάρχουν αναπτυγμένα συστήματα για την πρόβλεψη της εμφάνισης του διαβήτη μακροπρόθεσμα.

ΚΕΦΑΛΑΙΟ 4: Μεθοδολογία

4.1 Διαδικασία επίλυσης του προβλήματος Μηχανικής Μάθησης

Στην παρούσα ενότητα θα περιγραφεί η πορεία που ακολουθήθηκε για τη σχεδίαση των μοντέλων ταξινόμησης, με σκοπό την πρόβλεψη του σακχαρώδους διαβήτη. Ήταν απαραίτητο να ολοκληρωθούν συγκεκριμένα στάδια, προτού καταλήξουμε σε ικανοποιητικό αποτέλεσμα καθώς «η διαδικασία λύσεως ενός προβλήματος Μηχανικής Μάθησης είναι μία αλληλεπιδραστική και επαναληπτική διαδικασία, η οποία περιλαμβάνει πλήθος βημάτων κατά τη διάρκεια των οποίων θα πρέπει να ληφθούν αποφάσεις από τον εκάστοτε χρήστη» (Κωτσιαντής Σ., 2006).

Αρχικά, επιλέχθηκε ένα αντιπροσωπευτικό σύνολο δεδομένων. Τα δεδομένα αυτά χρησιμοποιήθηκαν για την εκπαίδευση και για την αξιολόγηση των Ταξινομητών. Επιπλέον, προσδιορίστηκαν οι συνιστώσες του προβλήματος που διαχειριστήκαμε. Η προεπεξεργασία αυτών των δεδομένων ήταν πολύ σημαντική, καθώς βοήθησε στη μείωση σφαλμάτων και στην εξάλειψη ελλιπών εγγραφών.

Αφού έγινε η επιλογή των κατάλληλων Ταξινομητών για τα δεδομένα, πραγματοποιήθηκαν πειράματα στις παραμέτρους και ως προς τον τρόπο εκπαίδευσης των Ταξινομητών και τελικώς έγινε η αξιολόγηση των αποτελεσμάτων που εξήχθησαν από αυτούς.

4.1.1 Εργαλεία που χρησιμοποιήθηκαν

Στην υλοποίηση της εκπαίδευσης και αξιολόγησης των Ταξινομητών χρησιμοποιήθηκε το εργαλείο WEKA και η γλώσσα προγραμματισμού R. Το WEKA χρησιμοποιήθηκε στον έλεγχο και στην επιλογή των κατάλληλων Ταξινομητών, αφενός λόγω της ευελιξίας που παρέχει σε σχέση με τις γλώσσες προγραμματισμού, καθώς δεν απαιτείται να γραφτεί κώδικας για την υλοποίηση ενός αλγορίθμου και αφετέρου επειδή, σε σχέση με άλλα εργαλεία, όπως το Rapidminer, παρέχεται δωρεάν για μόνιμη εγκατάσταση.

Ωστόσο, το μειονέκτημα αυτής της ευκολίας στη χρήση είναι ότι το WEKA αποδίδει λιγότερο στη στατιστική ανάλυση και στην εξερεύνηση των δεδομένων. Επιπλέον, αν θελήσουμε να διαχειριστούμε μεγάλο όγκο δεδομένων, η εκπαίδευση θα είναι εξαιρετικά αργή έως ανέφικτη. Συνεπώς για τέτοιου είδους περιπτώσεις θα πρέπει να χρησιμοποιηθεί κάποια γλώσσα προγραμματισμού. Μία από τις πλέον κατάλληλες για παρουσίαση δεδομένων, και παράλληλα εύκολη γλώσσα λόγω των πολλών βιβλιοθηκών που παρέχονται για μηχανική μάθηση, είναι η γλώσσα προγραμματισμού R.

Με βάση την παραπάνω επιχειρηματολογία, επιλέχθηκε η αξιολόγηση των Ταξινομητών να γίνει με τη χρήση του WEKA και την υλοποίησή τους με τη γλώσσα προγραμματισμού R.

4.1.2 Επιλογή και επισκόπηση Δεδομένων

Στο πρώτο αυτό βήμα θα πρέπει να επιλεγεί το κατάλληλο σύνολο των δεδομένων. Γενικά, το σύστημα υγειονομικής περίθαλψης περιλαμβάνει μεγάλη ποσότητα πληροφοριών για ασθενείς, που μπορεί να χρησιμοποιηθεί στην εξόρυξη δεδομένων και στην εξαγωγή κρυφών μοτίβων. Ανεξάρτητα από το γεγονός ότι το πρόβλημα στο σακχαρώδη διαβήτη είναι ότι δεν μεταβολίζεται η γλυκόζη, διάφοροι άλλοι παράγοντες, όπως το ύψος, το βάρος

και η ινσουλίνη, συμβάλλουν στην εμφάνισή του. Θα πρέπει λοιπόν να αναζητηθεί κάποιο σύνολο δεδομένων που να περιέχει και αυτές τις μεταβλητές.

Έπειτα από αναζήτηση στο διαδίκτυο, επιλέχθηκε ως πιο αντιπροσωπευτικό δείγμα σύνολο δεδομένων διαθέσιμο στην ιστοσελίδα: http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff. Αυτό περιέχει δείγμα γυναικών (PIMA Indians) ηλικίας τουλάχιστον 21 ετών. Κάθε στοιχείο αποτελεί ένα πολυδιάστατο διάνυσμα, όπου κάθε διάσταση εκφράζει και μία μεταβλητή.

Παρακάτω παρουσιάζονται οι κατηγορίες των μεταβλητών :

Ποσοτικές μεταβλητές

i. Συνεχείς (π.χ. το βάρος ενός ανθρώπου)
ii. Διακριτές (π.χ. το πλήθος των φοιτητών)

Ποιοτικές μεταβλητές

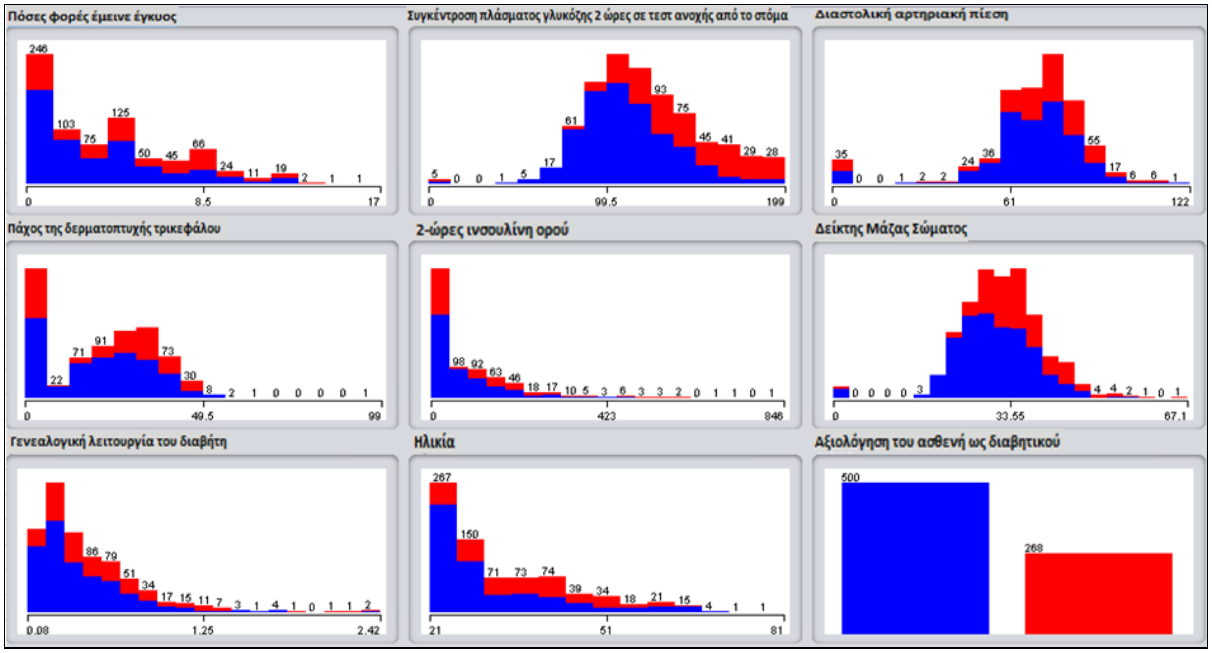
i. Κατηγορικές (π.χ. το φύλο [1 = άνδρας, 2 = γυναίκα])
ii. Διάταξης (π.χ. η επίδοση ενός φοιτητή: Καλώς/Λίαν καλώς/ Άριστα)

Το σύνολο δεδομένων μας αποτελείται από 768 στιγμιότυπα, όπου κάθε στιγμιότυπο περιλαμβάνει εννέα χαρακτηριστικά (όλα αριθμητικά) τα οποία είναι τα εξής :

Πίνακας 4: Πίνακας χαρακτηριστικών του συνόλου δεδομένων PIMA Indians

Α/Α	Μεταβλητή	Επεξήγηση	Τύπος Μεταβλητής
1	Number of times pregnant	Πόσες φορές έμεινε έγκυος	Διακριτή
2	Plasma glucose concentration a two hours in an oral glucose tolerance test	Συγκέντρωση πλάσματος γλυκόζης 2 ώρες σε τεστ ανοχής γλυκόζης από το στόμα	Διακριτή
3	Diastolic blood pressure (mm Hg)	Διαστολική αρτηριακή πίεση (σε mm)	Διακριτή
4	Triceps skin fold thickness (mm)	Πάχος της δερματοπτυχής τρικέφαλου (σε mm)	Διακριτή
5	2-Hour serum insulin (mu U/ml)	2 ώρες ινσουλίνη ορού (σε mu U/ml)	Διακριτή
6	Body mass index (weight in kg/(height in m)^2)	Δείκτης Μάζας Σώματος (βάρος σε κιλά/(ύψος σε μέτρα)^2])	Συνεχής
7	Diabetes pedigree function	Γενεαλογική λειτουργία του διαβήτη	Συνεχής
8	Age (years)	Ηλικία (σε έτη)	Διακριτή
9	Class variable (0 or 1) 268 of 768 are 1, the others are 0	Αξιολόγηση του ασθενή ως διαβητικού (τιμή 1) ή μη (τιμή 0)	Κατηγορική

Στο παρακάτω γράφημα που προέκυψε από το εργαλείο Weka απεικονίζονται ιστογράμματα όλων των χαρακτηριστικών:



Εικόνα 3 : Ιστογράμματα χαρακτηριστικών του συνόλου δεδομένων PIMA Indians

Τα παραπάνω ιστογράμματα δείχνουν την κατανομή των χαρακτηριστικών και τις ετικέτες των δεδομένων μας, παρέχοντας τις ακόλουθες πρώτες πληροφορίες:

- Η κλάση 0 (μπλε χρώμα στον πίνακα των χαρακτηριστικών) με 500 περιπτώσεις αντιπροσωπεύει ασθενείς που βγήκαν αρνητικοί στην εξέταση και η κλάση 1 (κόκκινο χρώμα στον πίνακα των χαρακτηριστικών) με 268 περιπτώσεις αντιπροσωπεύει τους ασθενείς που βγήκαν θετικοί στην εξέταση. Παρατηρούμε ανισορροπία στα δεδομένα, που ενδεχομένως θα δυσκολέψει την εκπαίδευση των Ταξινομητών, οδηγώντας σε παραπλανητικό αποτέλεσμα υπέρ των μη ασθενών.
- Η συγκέντρωση πλάσματος γλυκόζης, η διαστολική αρτηριακή πίεση και ο Δείκτης Μάζας Σώματος κατανέμονται κανονικά.
- Όλα τα υπόλοιπα χαρακτηριστικά φαίνεται να μην κατανέμονται κανονικά, παρουσιάζοντας κυρτότητα προς τα αριστερά.

4.1.3 Επεξεργασία Δεδομένων

Τα δεδομένα φαίνεται ότι δεν περιλαμβάνουν ελλιπείς τιμές, καθώς αυτές είναι συμπληρωμένες. Ωστόσο, θα πρέπει να εξετάσουμε λεπτομερώς τα δεδομένα μας για να αποφευχθούν τυχόν σφάλματα κατά την εκπαίδευση των αλγορίθμων μηχανικής μάθησης, καθώς δεν είναι σπάνια η περίπτωση να υπάρχουν σφάλματα στις τιμές των χαρακτηριστικών. Το φαινόμενο αυτό ονομάζεται θόρυβος και συναντάται κατά τη συλλογή δεδομένων από τις πειραματικές μετρήσεις και γενικότερα, όπου επεμβαίνει ο ανθρώπινος παράγοντας στη δημιουργία δεδομένων εκπαίδευσης.

Είναι αναμενόμενο η εκτεταμένη παρουσία θορύβου να αποπροσανατολίζει τον αλγόριθμο μάθησης. Αυτό ωστόσο δεν συμβαίνει. Στην περίπτωση που παρουσιάζεται θόρυβος ίδιας μορφής στα δεδομένα αξιολόγησης, χρειάζεται να αξιολογηθεί καθώς ανάγεται σε «χαρακτηριστικό» του προβλήματος μάθησης (Κωτσιαντής Σ., 2006).

Πραγματοποιήθηκε περαιτέρω έλεγχος στα δεδομένα, με σκοπό να εξεταστεί ύπαρξη θορύβου. Τα δεδομένα που λείπουν ενδέχεται να έχουν διαφορετικές πηγές, όπως θάνατο

ασθενών, δυσλειτουργίες εξοπλισμού, άρνηση απαντήσεων σε συγκεκριμένες ερωτήσεις κ.λπ. Έπειτα από εξέταση όλων των τιμών των χαρακτηριστικών, παρατηρείται ότι το σύνολο των δεδομένων περιέχει ορισμένα ασυνεπή δεδομένα, καθώς σε όλα τα χαρακτηριστικά, πλην της γενεαλογικής λειτουργίας του διαβήτη και της ηλικίας, υπάρχουν τιμές ίσες με το μηδέν. Αυτό, εκτός του χαρακτηριστικού για το πόσες φορές έχει μείνει έγκυος κάποια γυναίκα, το οποίο ενδέχεται να είναι ίσο με το μηδέν, δεν μπορεί να ευσταθεί. Συγκεκριμένα, στο ποσοστό 49% των περιπτώσεων, δηλαδή στις 376 από τις συνολικά 768, λείπουν τιμές ενός ή περισσότερων από τα χαρακτηριστικά.

Το ποσοστό ελλειπουσών τιμών ανά χαρακτηριστικό παρατίθεται στον παρακάτω πίνακα:

Πίνακας 5 : Ποσοστό ελλειπουσών τιμών χαρακτηριστικών του συνόλου δεδομένων PIMA Indians

Μεταβλητή	Ελλιπείς Τιμές	Ποσοστό Ελλειπουσών Τιμών
Πόσες φορές έμεινε έγκυος	0	0,00%
Συγκέντρωση Πλάσματος	5	0,65%
Διαστολική αρτηριακή πίεση	35	4,55%
Πάχος της δερματοπτυχής τρικέφαλου	227	29,52%
2 ώρες ινσουλίνη ορού	374	**48,63%**
Δείκτης Μάζας Σώματος	11	1,43%
Γενεαλογική λειτουργία του διαβήτη	0	0,00%
Ηλικία	0	0,00%

Καταλήγουμε στο συμπέρασμα ότι θα ήταν ωφέλιμο να χρησιμοποιηθεί κάποια τεχνική προεπεξεργασίας στα δεδομένα μας, με σκοπό να βελτιώσουμε τα εξαγόμενα αποτελέσματα. Χρησιμοποιώντας το φίλτρο Replace with missing values (Αντικατάσταση με ελλείπουσες τιμές), όπως προτείνουν οι Pradeep Kandhasamy J., Balamurali S. (Pradeep Kandhasamy J., Balamurali S., 2015), θα μειώσουμε τον θόρυβο. Στην τεχνική αυτή οι μηδενικές τιμές στο σύνολο των δεδομένων αντιμετωπίζονται ως ελλείπουσες τιμές, εκτός από το πρώτο χαρακτηριστικό (πόσες φορές έμεινε έγκυος), και όλες οι άλλες έχουν αντικατασταθεί από μέσες τιμές, χρησιμοποιώντας τον αλγόριθμο k Πλησιέστεροι Γείτονες.

Ωστόσο, με τη χρήση του φίλτρου στα δεδομένα, δεν παρατηρείται βελτίωση στα αποτελέσματα ταξινόμησης. Αποφασίστηκε λοιπόν να γίνει προσπάθεια συμπλήρωσης των ελλειπουσών τιμών με χρήση άλλη τεχνικής. Το φίλτρο Replace Missing Values, το οποίο αντικαθιστά τις χαμένες τιμές μιας ονομαστικής στήλης με την πιο συνηθισμένη τιμή ή μιας αριθμητικής με τη μέση τιμή χαρακτηριστικού, φάνηκε να είναι καλή λύση.

Παρατηρώντας καλύτερα τα δεδομένα μας, γίνεται εύκολα αντιληπτό ότι οι κλάσεις τους δεν είναι ισορροπημένες. Έγινε λοιπόν προσπάθεια εξισορρόπησής τους ώστε να εκπαιδευτούν καλύτερα οι Ταξινομητές και να βελτιώσουμε τα αποτελέσματα ταξινόμησής τους. Ακολουθήθηκε προεπεξεργασία στα δεδομένα, έτσι ώστε να αυξήσουμε τα δεδομένα με την ετικέτα «μη ασθενής», στοχεύοντας στην ισορροπία μεταξύ των δύο κλάσεων.

Έπειτα από αναζήτηση για την αντιμετώπιση της ιδιομορφίας του συνόλου δεδομένων, όπου οι κλάσεις δεν είναι ισορροπημένες (65,1% τα αρνητικά – 34,9% τα θετικά), έγινε χρήση του φίλτρου SMOTE.

Αυτό το φίλτρο για τα δεδομένα μειοψηφίας δημιουργεί νέα συνθετική περίπτωση δεδομένων Α, λαμβάνοντας τη διαφορά μεταξύ του διανύσματος χαρακτηριστικών της Α και του πλησιέστερου γείτονα Β που ανήκουν στην ίδια κλάση, πολλαπλασιάζοντας τη διαφορά τους με τυχαίο αριθμό μεταξύ 0 και 1 και κατόπιν προσθέτοντάς τη στην Α. Αυτό δημιουργεί τυχαίο τμήμα γραμμής μεταξύ κάθε ζεύγους υπαρχόντων χαρακτηριστικών από τις περιπτώσεις Α και Β, με αποτέλεσμα τη δημιουργία νέας στιγμής μέσα στο σύνολο δεδομένων. Αυτή η διαδικασία επαναλαμβάνεται για τους άλλους γείτονες k-1 της ιεραρχίας μειονότητας. Το φίλτρο SMOTE δημιουργεί γενικότερες περιοχές από την τάξη μειοψηφίας και οι ταξινομητές των Δέντρων Αποφάσεων μπορούν να χρησιμοποιήσουν το σύνολο δεδομένων για καλύτερες γενικεύσεις. Προκειμένου να αποφευχθεί το πρόβλημα υπερεκπαίδευσης, το SMOTE δημιουργεί νέες περιπτώσεις. Οι τιμές του νέου στιγμιότυπου εξακολουθούν να έχουν σχέση με το σύνολο δεδομένων. Για κάθε περίπτωση τάξης μειοψηφίας, το SMOTE παρεμβάλλει τιμές με την τεχνική kNN και δημιουργεί τιμές χαρακτηριστικών για νέες περιπτώσεις δεδομένων (Pears R., Connor A., 2014).

Τέλος, συνδυάζοντας τα φίλτρα Replace Missing Values και SMOTE, προσπαθήσαμε να πετύχουμε όσο το δυνατόν καλύτερη ακρίβεια στις προβλέψεις μας.

4.1.4 Διαδικασία Επιλογής Ταξινομητών προς εκπαίδευση

Η επιλογή του καταλληλότερου αλγορίθμου μάθησης για το εκάστοτε πρόβλημα περιλαμβάνει αποφάσεις για το ποιο μοντέλο και ποιες παράμετροι είναι τα πλέον κατάλληλα να χρησιμοποιηθούν. Ένα πρόβλημα μπορεί να λυθεί συγκριτικά καλά χρησιμοποιώντας ποικίλες μεθόδους, π.χ. Τεχνητά Νευρωνικά Δίκτυα, Δέντρα Απόφασης, Μπεϋζιανούς Ταξινομητές, κ.λπ. Επιπλέον, συνήθως παρουσιάζεται μια μέθοδος η οποία προβλέπει καλύτερα ορισμένα μέρη των περιπτώσεων σε σχέση με τις άλλες. Για παράδειγμα, στην περίπτωσή μας, ένας Ταξινομητής θα μπορούσε να προβλέπει καλύτερα τους ασθενείς, ενώ ένας άλλος τους μη ασθενείς. Έτσι, ανάλογα με το τι είναι σημαντικότερο για τον αναλυτή να προβλέψει, επιλέγεται και ο κατάλληλος Ταξινομητής. Κατά συνέπεια, η επιλογή της καταλληλότερης μεθόδου για την τελική λύση είναι ένα περίπλοκο πρόβλημα (Κωτσιαντής Σ., 2006).

Είναι σημαντικό, κατά την προσπάθεια επίλυσης ενός προβλήματος, να υπάρχει αντιπροσωπευτικό δείγμα αλγορίθμων για εκμάθηση. Όπως αναφέρθηκε ήδη και στο προηγούμενο κεφάλαιο, ένας καλός κανόνας είναι «μερικοί από κάθε τύπο», οπότε, στην περίπτωσή μας, που καλούμαστε να αντιμετωπίσουμε πρόβλημα δυαδικής ταξινόμησης, θα επιλέξουμε τουλάχιστον έναν Ταξινομητή από κάθε κύριο είδος.

Έτσι, καταλήξαμε στους παρακάτω επτά Ταξινομητές:

- Μη Γραμμικές Μέθοδοι Εκπαίδευσης: Multilayer Perceptron, SVM, kNN και Naive Bayes, Λογιστική Παλινδρόμηση
- Δέντρα Απόφασης: J48
- Σύνολα Δέντρων: Τυχαία Δάση

4.1.5 Τροποποίηση παραμέτρων και εκπαίδευση Ταξινομητών

Είναι πολύ σημαντικό κατά την εκτέλεση των αλγορίθμων να πειραματιστούμε, μεταβάλλοντας τις τιμές των παραμέτρων των Ταξινομητών, με σκοπό να βελτιώσουμε την εκπαίδευσή τους και τα αποτελέσματά τους. Μεταβάλαμε λοιπόν την παράμετρο c στον Ταξινομητή SVM, τα τυχαία επιλεγμένα χαρακτηριστικά στον Ταξινομητή Τυχαία Δάση, τον αριθμό των εξεταζόμενων Γειτόνων k στον Ταξινομητή kNN, τον ελάχιστο αριθμό περιπτώσεων ανά φύλλο στα Δέντρα Απόφασης, τον αριθμό επαναλήψεων στη Λογιστική Παλινδρόμηση και τον αριθμό κρυφών κόμβων στα Τροφοδοτούμενα προς τα εμπρός Δίκτυα Πολλαπλών Στρωμάτων. Για την εκπαίδευση των Ταξινομητών ήταν απαραίτητο να χωριστεί το σύνολο των δεδομένων μας σε ένα σύνολο εκπαίδευσης και σε ένα αξιολόγησης. Όπως αναφέρουν και τα ονόματά τους το πρώτο θα χρησιμεύσει στην εκπαίδευση των Ταξινομητών και το δεύτερο στην αξιολόγησή τους. Οι τεχνικές εκπαίδευσης που χρησιμοποιήθηκαν είναι οι ακόλουθες:

- Ποσοστό διάσπασης: μέσω αυτού, κατακερματίζουμε τα δεδομένα σε δεδομένα εκπαίδευσης και σε δεδομένα αξιολόγησης ανάλογα με το ποσοστό που έχουμε ορίσει. Για παράδειγμα, όταν ορίσουμε 70% ποσοστό διάσπασης, θα έχουμε το 70% των δεδομένων που έχουμε εισαγάγει ως δεδομένα εκπαίδευσης και το υπόλοιπο 30% ως δεδομένα αξιολόγησης.

- Διασταυρούμενη επικύρωση: Σύμφωνα με αυτήν, τα δεδομένα χωρίζονται σε τμήματα, ανάλογα με τον αριθμό τμημάτων που έχουν οριστεί. Αν, για παράδειγμα, ορίσουμε αριθμό τμημάτων ίσο με 10 τότε παίρνουμε τα 9 για εκπαίδευση και το 1 για έλεγχο, επαναλαμβάνοντας την ίδια διαδικασία για όλες τις διαφορετικές περιπτώσεις εκπαίδευσης και δοκιμής, μέχρις ότου όλα τα τμήματα χρησιμοποιηθούν έστω και μία φορά για αξιολόγηση.

Κατά τη μεταβολή των παραμέτρων, παρατηρήθηκαν σημαντικές αλλαγές στα εξαγόμενα αποτελέσματα και επιλέχθηκαν αυτές με τα ακριβέστερα και τα πιο αντικειμενικά αποτελέσματα ανά Ταξινομητή για να αξιολογηθούν.

4.1.6 Αξιολόγηση παραγόμενης γνώσης

Τέλος, μετά από την επιλογή των Ταξινομητών, το έργο της εκπαίδευσης έχει ολοκληρωθεί και μπορούμε να προχωρήσουμε στην αξιολόγησή των αποτελεσμάτων τους. Στον τομέα της Ιατρικής, το κόστος μιας διάγνωσης μπορεί να έχει βαρύτατες συνέπειες. Η εφαρμογή αλγορίθμων μηχανικής μάθησης για να βρεθούν οι καλύτεροι ταξινομητές μπορεί να διαδραματίσει σημαντικό ρόλο συμβάλλοντας στην ελαχιστοποίηση των λαθών αυτών. Κατά την πειραματική εξέλιξη του συστήματος, ωστόσο, θα πρέπει να ακολουθηθεί το στάδιο της αξιολόγησης. Αυτό έχει δύο κυρίως σκοπούς:

1. Την εκτίμηση της χρησιμότητας του συστήματος για τους υποψήφιους χρήστες του, με τη βοήθεια κατάλληλα ορισμένων μέτρων αποτελεσματικότητας και τεχνικών εκτίμησης των μέτρων αυτών.
2. Τη ρύθμιση διαφόρων παραμέτρων του συστήματος έτσι ώστε να βελτιστοποιηθεί η απόδοσή του ως προς τα επιλεγμένα μέτρα (Κωτσιαντής Σ., 2006).

Η ακρίβεια ταξινόμησης που εξετάστηκε στα πειράματά μας αποτελεί την αφετηρία μας. Είναι ο αριθμός των σωστών προβλέψεων που διαχωρίζονται από τον συνολικό αριθμό των προβλέψεων πολλαπλασιασμένων επί 100 για να μετατραπούν σε ένα ποσοστό. Ωστόσο, ένας καθαρός και αδιαμφισβήτητος τρόπος για να παρουσιάσουμε τα αποτελέσματα πρόβλεψης ενός ταξινομητή είναι να χρησιμοποιήσουμε έναν πίνακα σύγχυσης (confusion matrix).

Πίνακας 6: Πίνακας Σύγχυσης

Πίνακας Σύγχυσης		Δεδομένα Πρόβλεψης	
		Θετικά	Αρνητικά
Πραγματικά Δεδομένα	Θετικά	Σωστά Ταξινομημένα ως Θετικά (TP)	Λανθασμένα Ταξινομημένα ως Θετικά (FN)
	Αρνητικά	Λανθασμένα Ταξινομημένα ως Αρνητικά (FP)	Σωστά Ταξινομημένα ως Αρνητικά (TN)

Οι καταχωρίσεις στον πίνακα έχουν την ακόλουθη σημασία στο πλαίσιο της εργασίας:

- TP: είναι ο αριθμός των σωστών προβλέψεων όταν τα δείγματα είναι θετικά.
- FN είναι ο αριθμός των εσφαλμένων προβλέψεων όταν τα δείγματα είναι αρνητικά.
- FP: είναι ο αριθμός των εσφαλμένων προβλέψεων όταν τα δείγματα είναι θετικά.
- TN: είναι ο αριθμός των σωστών προβλέψεων όταν τα δείγματα είναι αρνητικά.

Για ένα δυαδικό πρόβλημα ταξινόμησης ο πίνακας έχει δύο γραμμές και δύο στήλες. Στην κορυφή είναι οι παρατηρούμενες ετικέτες κλάσης και στην πλευρά τους οι προβλεπόμενες ετικέτες κατηγορίας. Κάθε κελί περιέχει τον αριθμό των προβλέψεων του ταξινομητή που εμπίπτουν σε αυτό το κελί.

Είναι ένας χρήσιμος πίνακας που παρουσιάζει τόσο την κατανομή τάξης στα δεδομένα όσο και την ταξινομημένη προβλεπόμενη κατανομή τάξης με ανάλυση των τύπων σφαλμάτων.

Από τον πίνακα σύγχυσης, με τη χρήση του παρακάτω τύπου λαμβάνουμε την ακρίβεια ενός Ταξινομητή.

Ακρίβεια Ταξινόμησης:

$$Ακρίβεια = (TP + TN) \,/\, (TP + FN + FP + TN)$$

Παρόλα αυτά, μερικές φορές μπορεί να είναι επιθυμητό να επιλέξουμε ένα μοντέλο με χαμηλότερη ακρίβεια επειδή έχει μεγαλύτερη προγνωστική δύναμη στο πρόβλημα. Για παράδειγμα, σε ένα πρόβλημα όπου υπάρχει μεγάλη ανισορροπία τάξης, όπως συμβαίνει στην περίπτωσή μας, ένα μοντέλο μπορεί να κατατάσσει όλα τα στιγμιότυπα στην τάξη πλειοψηφίας και συνεπώς να επιτυγχάνει υψηλή ακρίβεια ταξινόμησης. Το πρόβλημα είναι ότι αυτό το μοντέλο δεν είναι επαρκές στην επίλυση τέτοιων προβλημάτων. Αυτό ονομάζεται Accuracy Paradox. Για προβλήματα όπως αυτά, απαιτούνται πρόσθετα μέτρα για την αξιολόγηση ενός ταξινομητή. Τέτοια μέτρα αξιολόγησης είναι η ευαισθησία, η ειδικότητα και το μέτρο F.

Η ευαισθησία υπολογίζει το ποσοστό των ψευδώς αρνητικών και η ειδικότητα το ποσοστό των ψευδώς θετικών αποτελεσμάτων όταν ελέγχεται μεγάλος αριθμός θετικών και αρνητικών δειγμάτων.

Ευαισθησία (recall/sensitivity):

$$Recall = TP \,/\, (TP + FN)$$

Ειδικότητα (specificity):

$$Precision = TP \ / \ (TP + FP)$$

Τέλος, αν προσπαθούσαμε να επιλέξουμε ένα μοντέλο βασισμένο στην ισορροπία μεταξύ ακρίβειας και ανάκλησης, τότε το μέτρο της F είναι το κατάλληλο (Brownlee J., 2014).

Μέτρο F (F-measure):

$$F = \ 2 \ / \ (1 \ / \ Precision \ + \ 1 \ / \ Recall)$$

ΚΕΦΑΛΑΙΟ 5: Χρήση εργαλείου Weka για την επεξεργασία δεδομένων και την υλοποίηση αλγορίθμων

5.1 Εργαλείο Weka

Το WEKA (Waikato Environment for Knowledge Analysis) είναι σουίτα λογισμικού για μηχανική μάθηση και εξόρυξη δεδομένων. Αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Ν. Ζηλανδίας και πήρε το όνομά του από το Weka, ένα μικρό και υπό εξαφάνιση πουλί της Ν. Ζηλανδίας. Είναι γραμμένο σε Java και ανήκει στην κατηγορία του λεγόμενου ελεύθερου λογισμικού, επιτρέποντας στους χρήστες να χρησιμοποιούν αλλά και να τροποποιούν ελεύθερα το λογισμικό του.

Είναι ένα από τα πιο διαδεδομένα λογισμικά εξόρυξης δεδομένων και έχει χρησιμοποιηθεί σε μεγάλο αριθμό επιστημονικών εργασιών. Η δημοφιλία του οφείλεται στα ειδικά χαρακτηριστικά του και στις δυνατότητες που προσφέρει. Συγκεκριμένα, παρέχει ποικιλία μεθόδων για κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, κανόνες συσχέτισης, προεπεξεργασία δεδομένων, καθώς και εργαλεία οπτικοποίησης (Κύρκος Ε., 2015). Η πιο πρόσφατη έκδοση, η οποία έχει χρησιμοποιηθεί και στην παρούσα εργασία, είναι η Weka 3.8 και υπάρχει και η Weka 3.9 development έκδοση.

5.1.1 Εισαγωγή δεδομένων στο weka

Αρχικά έγινε εξαγωγή των δεδομένων από τη σελίδα (http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff), όπου και επιλέχθηκε αρχείο με κατάληξη .arff που είναι η μορφή αρχείων την οποία μπορεί να εκτελέσει το weka. Τα αρχεία ARFF είναι απλά αρχεία κειμένου που περιέχουν ενότητα κεφαλίδας με λίστα τύπων (αριθμητική, ονομαστική συμβολοσειρά, ημερομηνία) και ενότητα δεδομένων με τις τιμές (FileInfo - The File Extensions Database, 2012).

5.2 Αποτελέσματα

Με τη βοήθεια της μηχανικής μάθησης για την επεξεργασία των δεδομένων του Pima Indian, δημιουργήθηκε μοντέλο πρόβλεψης. Αυτό το μοντέλο πρέπει να προβλέψει με μεγάλη πιθανότητα ποιοι άνθρωποι είναι πιθανό να αναπτύξουν διαβήτη, χρησιμοποιώντας Δέντρο Απόφασης (J48), Λογιστική Παλινδρόμηση, Τυχαία Δάση, Multilayer perceptron, SVM, Μπεϊζιανός Αλγόριθμος και kNN, ώστε να θεωρηθεί ένα καλό μοντέλο.

Για καθέναν από τους Ταξινομητές έγιναν τρεις διαφορετικές διαδικασίες αξιολόγησης. Οι Ταξινομητές αξιολογήθηκαν σε τρία διαφορετικά επίπεδα της μεταβλητής «φίλτρο». Το πρώτο επίπεδο ήταν χωρίς τη χρήση κάποιου φίλτρου, το δεύτερο ήταν το φίλτρο «αντικατάσταση με ελλιπείς τιμές», ενώ το τρίτο ήταν ο συνδυασμός των φίλτρων «αντικατάσταση των ελλιπών τιμών» με το φίλτρο «SMOTE».

Σε καθένα από τα διαφορετικά επίπεδα του φίλτρου ο Ταξινομητής μετρήθηκε ως προς την ακρίβεια, την ειδικότητα, την ευαισθησία και το μέτρο F.

Προκειμένου να έχουμε πληρέστερη εικόνα για τις ικανότητες του κάθε Ταξινομητή και για να υπολογίσουμε τους πιο πάνω δείκτες, μετρήσαμε για κάθε Ταξινομητή τις συχνότητες (απόλυτες και σχετικές) των σωστά ταξινομημένων ασθενών, είτε αυτοί εμφανίζουν την ασθένεια είτε όχι, αλλά και τις συχνότητες των λανθασμένα ταξινομημένων ασθενών, είτε έχουν την ασθένεια είτε όχι.

Ένα από τα ζητήματα που μας απασχόλησε κατά τη διαδικασία των πειραμάτων ήταν αν θα εκπαιδεύαμε και θα αξιολογούσαμε τα δεδομένα μας μέσω διασταυρούμενης επικύρωσης ή μέσω δεδομένων εκπαίδευσης και αξιολόγησης.

Η διαίρεση των παρατηρήσεων χρησιμοποιώντας διασταυρούμενη επικύρωση K-fold παίρνει k-φορές περισσότερο από ένα σύνολο δεδομένων αξιολόγησης, το οποίο είναι το μειονέκτημά του. Επιπλέον, εάν διαχωρίσαμε τα δεδομένα μας χρησιμοποιώντας σύνολο δεδομένων αξιολόγησης, ίσως μέσω ορισμένων χρήσιμων δεδομένων στο δοκιμαστικό μέρος, η διασταυρούμενη επικύρωση αποδίδει καλύτερα, επειδή κάθε παρατήρηση χρησιμοποιείται τόσο για εκπαίδευση όσο και για δοκιμή (Daniel Lee, works at Booz Allen Hamilton, 2017). Παρατηρήθηκε, όπως φαίνεται και στον παρακάτω πίνακα, ότι οι ταξινομητές J48, Τυχαία Δάση και ο Μπεϊζιανός Αλγόριθμος παρουσίαζαν καλύτερα αποτελέσματα με διάσπαση των δεδομένων σε δεδομένα εκπαίδευσης και αξιολόγησης 78%, 76% και 66% αντίστοιχα.

Πίνακας 7: Διαφορά ποσοστών Ακρίβειας, Ειδικότητας και Ευαισθησίας και μέτρου F ανάλογα με τον τρόπο εκπαίδευσης

Ταξινομητές	Τρόπος εκπαίδευσης/Διαφορά	Ακρίβεια	Ειδικότητα (Precision)	Ευαισθησία (recall)	Μέτρο F
J48	Ποσοστό εκπαίδευσης 78%	81,22%	79%	87,13%	82,63%
	Διασταυρούμενη επικύρωση = 10	74,78%	69%	77,02%	72,97%
	Διαφορά	**6,44%**	**9,25%**	**10,11%**	**9,66%**
Τυχαία Δάση	Ποσοστό εκπαίδευσης 76%	80,93%	79,82%	83,49%	81,61%
	Διασταυρούμενη επικύρωση = 10	79,24%	75,24%	79,04%	77,09%
	Διαφορά	**1,69%**	**4,58%**	**4,45%**	**4,52%**
Μπεϊζιανός Αλγόριθμος	Ποσοστό εκπαίδευσης 66%	73,44%	76,34%	66,67%	71,17%
	Διασταυρούμενη επικύρωση = 10	72,32%	71,89%	61,36%	66,21%
	Διαφορά	**1,12%**	**4,44%**	**5,30%**	**4,96%**

Ωστόσο, προτιμήθηκε να γίνει η εκπαίδευση και ο έλεγχος μέσω της διασταυρούμενης επικύρωσης γιατί το αρχείο δεδομένων μας ήταν σχετικά μικρό, με πολλές ελλιπείς τιμές, που δυσκολεύουν τη γενίκευση των αποτελεσμάτων. Με αυτόν τον τρόπο, τα αποτελέσματά μας είναι πιο αντικειμενικά, ως προς ένα άλλο άγνωστο σύνολο δεδομένων. Παρέχεται καλύτερη εμπιστοσύνη στην ακρίβεια της πρόβλεψής μας. Ως αποτέλεσμα, υλοποιήθηκαν πειράματα για τους Ταξινομητές, με τις παρακάτω παραμέτρους και με κοινή μέθοδο εκπαίδευσης τη Διασταυρούμενη Επικύρωση = 10.

Πίνακας 8 : Παράμετροι Εκπαίδευσης συνόλου δεδομένων

Ταξινομητές	Μέθοδος εκπαίδευσης	Μεταβολή παραμέτρων
J48	Διασταυρούμενη Επικύρωση= 10	Προεπιλεγμένες
Λογιστική Παλινδρόμηση	Διασταυρούμενη Επικύρωση= 10	Προεπιλεγμένες
Τυχαία Δάση	Διασταυρούμενη Επικύρωση= 10	Προεπιλεγμένες
Multilayer perceptron	Διασταυρούμενη Επικύρωση= 10	Προεπιλεγμένες
SVM	Διασταυρούμενη Επικύρωση= 10	c= 1
Μπεϊζιανός Αλγόριθμος	Διασταυρούμενη Επικύρωση= 10	Προεπιλεγμένες
kNN	Διασταυρούμενη Επικύρωση= 10	Αριθμός πλησιέστερων γειτόνων= 7

Στον παρακάτω πίνακα απεικονίζεται ανά Ταξινομητή ο απόλυτος αριθμός των σωστών και των λανθασμένα ταξινομημένων περιπτώσεων (ως θετικές και αρνητικές αντίστοιχα) και το ποσοστό τους ανά κατηγορία:

Πίνακας Αποτελεσμάτων 1									
Ταξινομητές	Φίλτρα	Σωστά Ταξινομημένα ως Αρνητικά (TN)	Λανθασμένα Ταξινομημένα ως Αρνητικά (FP)	Λανθασμένα Ταξινομημένα ως Θετικά (FN)	Σωστά Ταξινομημένα ως Θετικά (TP)	Ακρίβεια	Ειδικότητα (Precision)	Ευαισθησία (recall)	Μέτρο F
J48	Χωρίς Φίλτρο	407	93	108	160	73,83%	63%	60%	61%
		79%	37%	21%	63%				
	Αντικατάσταση με Ελλιπείς Τιμές	380	71	132	106	70,54%	60%	45%	51%
		74%	40%	26%	60%				
	Αντικατάσταση Ελλιπών Τιμών & SMOTE	365	135	91	305	74,78%	69%	77%	**73%**
		80%	31%	20%	69%				
Λογιστική Παλινδρόμηση	Χωρίς Φίλτρο	440	60	115	153	77,21%	72%	57%	64%
		79%	28%	21%	72%				
	Αντικατάσταση με Ελλιπείς Τιμές	392	59	109	129	75,62%	69%	54%	61%
		78%	31%	22%	69%				
	Αντικατάσταση Ελλιπών Τιμών & SMOTE	408	92	137	259	74,44%	74%	65%	**69%**
		75%	26%	25%	74%				
Τυχαία Δάση	Χωρίς Φίλτρο	418	82	104	164	75,78%	67%	61%	64%
		80%	33%	20%	67%				
	Αντικατάσταση με Ελλιπείς Τιμές	391	60	114	124	74,75%	67%	52%	59%
		77%	33%	23%	67%				
	Αντικατάσταση Ελλιπών Τιμών & SMOTE	397	103	83	313	79,24%	75%	79%	**77%**
		83%	25%	17%	75%				

Multilayer perceptron	Χωρίς Φίλτρο	416	84	105	163	75,39%	66%	61%	63%
		80%	34%	20%	66%				
	Αντικατάσταση με Ελλιπείς Τιμές	374	77	114	124	72,28%	62%	52%	56%
		77%	38%	23%	62%				
	Αντικατάσταση Ελλιπών Τιμών & SMOTE	364	136	94	302	74,33%	69%	76%	**72%**
		79%	31%	21%	69%				
SVM	Χωρίς Φίλτρο	449	51	123	145	77,34%	74%	54%	63%
		78%	26%	22%	74%				
	Αντικατάσταση με Ελλιπείς Τιμές	406	45	125	113	75,33%	72%	47%	57%
		76%	28%	24%	72%				
	Αντικατάσταση Ελλιπών Τιμών & SMOTE	410	90	155	241	72,66%	73%	61%	**66%**
		73%	27%	27%	73%				
Μπεϊζιανός Αλγόριθμος	Χωρίς Φίλτρο	422	78	104	164	76,30%	68%	61%	64%
		80%	32%	20%	68%				
	Αντικατάσταση με Ελλιπείς Τιμές	378	73	102	136	74,60%	65%	57%	61%
		79%	35%	21%	65%				
	Αντικατάσταση Ελλιπών Τιμών & SMOTE	405	95	153	243	72,32%	72%	61%	**66%**
		73%	28%	27%	72%				
kNN	Χωρίς Φίλτρο	397	103	126	142	70,18%	58%	53%	55%
		76%	42%	24%	58%				
	Αντικατάσταση με Ελλιπείς Τιμές	322	129	116	122	64,44%	49%	51%	50%
		74%	51%	26%	49%				
	Αντικατάσταση Ελλιπών Τιμών & SMOTE	369	131	83	313	76,12%	70%	79%	**75%**
		82%	30%	18%	70%				

Αρχικά παρατηρήθηκε ότι η χρήση του φίλτρου «Αντικατάσταση με Ελλιπείς Τιμές» στα δεδομένα μας, δεν απέδωσε καρπούς. Καθώς η ακρίβεια δεν ήταν η αναμενόμενη, το φίλτρο απορρίφθηκε. Αντίθετα, παρατηρήθηκε ότι για όλους τους αλγορίθμους η χρήση του φίλτρου «SMOTE» λειτούργησε με βελτιωμένα αποτελέσματα. Μερικοί αλγόριθμοι μηχανικής μάθησης φαίνεται να είναι πιο επιρρεπείς από άλλους στην ανισορροπία δεδομένων, ωστόσο στην πραγματικότητα παρατηρήθηκε ότι όλοι ήταν ευάλωτοι σε αυτήν, άλλοι περισσότερο και άλλοι λιγότερο.
Παρατηρήθηκε λοιπόν ότι πιο ευάλωτοι αλγόριθμοι ήταν εκείνοι που κλίνουν προς τη λεγόμενη «τεμπέλικη μάθηση» (lazy learning). Όσο πιο «τεμπέλης» είναι ένας αλγόριθμος, τόσο πιο ευάλωτος είναι σε δεδομένα που παρουσιάζουν έλλειψη ισορροπίας. Ως αποτέλεσμα αυτής της κατάστασης, οι αλγόριθμοι εκμάθησης που κάνουν επιπλέον εργασία για να εξαγάγουν σημαντικές αναπαραστάσεις, προκειμένου να κωδικοποιηθούν τα δεδομένα με έναν ειδικό τρόπο, εμφανίστηκαν λιγότερο ευάλωτοι.

Για παράδειγμα, η βελτίωση της απόδοσης του «τεμπέλη» Ταξινομητή kNN ήταν αισθητή μετά από την εξισορρόπηση των κλάσεων, σε αντίθεση με τον SVM του οποίου η απόδοση ήταν μειωμένη. Αυτό οφείλεται στο γεγονός ότι το SVM επηρεάζεται μόνο από εκείνα τα σημεία δεδομένων που βρίσκονται πλησιέστερα στο περιθώριο, και επομένως οτιδήποτε πέρα από αυτό δεν τον επηρεάζει. Το SVM κάνει πολύ σκληρή δουλειά για να βρει αυτούς τους φορείς υποστήριξης, σε αντίθεση με τον «τεμπέλη» kNN Ταξινομητή, που είναι σαν κάποιον που προσπαθεί να καταλάβει πράγματα σε σχέση με κάποιον που προσπαθεί να τα απομνημονεύσει. Επομένως, υπάρχει μεγαλύτερη ευπάθεια εάν τα δεδομένα αποθηκεύονται απλά και κωδικοποιούνται με ειδικό παραμετρικό ή μη παραμετρικό τρόπο (Abhinav Maurya, PhD Student at CMU, 2016).

Για την καλύτερη αξιολόγηση των αποτελεσμάτων μας χρησιμοποιήθηκαν και άλλες μετρικές, όπως της ευαισθησία και της ειδικότητας εξάγοντας τα εξής: Παρατηρήσαμε ότι τα Τυχαία Δάση και ο kNN έχουν τα καλύτερα ποσοστά ευαισθησίας, ενώ τα Τυχαία Δάση τα καλύτερα και από πλευράς ειδικότητας. Έτσι, καταλήγουμε στο συμπέρασμα ότι Τυχαία Δάση και ο kNN μπορούν και ταξινομούν το ίδιο καλά τους ασθενείς, ενώ τα Τυχαία Δάση παρουσιάζουν μεγαλύτερη ακρίβεια στα αποτελέσματά του σε σχέση με τους υπόλοιπους Ταξινομητές.

Στην περίπτωσή μας είναι πιο σημαντικό να διαγνώσουμε κάποιον ασθενή ως ασθενή παρά κάποιον μη ασθενή ως ασθενή. Ωστόσο τα ποσοστά των δύο αλγορίθμων είναι αρκετά κοντά, με αποτέλεσμα η διαφορά τους να μην είναι καθοριστικής σημασίας.

Επειδή συχνά η ειδικότητα και η ευαισθησία δεν βοηθούν για να εκτιμήσουμε την καταλληλότητα ενός Ταξινομητή, χρησιμοποιήσαμε και ένα άλλο μέτρο εκτίμησης, το μέτρο F. Το μέτρο F είναι αρμονικό μέσο μεταξύ ευαισθησίας και ειδικότητας, όπου, όταν έχει υψηλή τιμή, και τα δύο μέτρα είναι ικανοποιητικά. Στα αποτελέσματα παρατηρείται ότι ο αλγόριθμος Τυχαία Δάση έχει το μεγαλύτερο δείκτη μέτρου F, ενώ ο SVM και ο Μπεϊζιανός Αλγόριθμος τους μικρότερους δείκτες.

ΚΕΦΑΛΑΙΟ 6: Υλοποίηση των Μεθόδων Μάθησης με τη γλώσσα προγραμματισμού R

6.1 Γλώσσα Προγραμματισμού R

Η R είναι γλώσσα προγραμματισμού και περιβάλλον λογισμικού για στατιστική ανάλυση, γραφική αναπαράσταση και αναφορά. Δημιουργήθηκε από τους Ross Ihaka και Robert Gentleman στο πανεπιστήμιο του Auckland της Νέας Ζηλανδίας και αναπτύσσεται από την ομάδα του R Development Core Team. Ονομάστηκε R, με βάση το πρώτο γράμμα του ονόματος των δύο συγγραφέων R (Robert Gentleman και Ross Ihaka). Είναι ελεύθερα διαθέσιμη υπό τη Γενική Άδεια Δημόσιας Χρήσης του GNU και προσφέρει προκατασκευασμένες δυαδικές εκδόσεις για διάφορα λειτουργικά συστήματα, όπως Linux, Windows και Mac. (Tutorialspoint).

Παρακάτω αναφέρονται τα σημαντικότερα χαρακτηριστικά της γλώσσας προγραμματισμού R:

- Είναι καλά αναπτυγμένη, απλή και αποτελεσματική γλώσσα προγραμματισμού που περιλαμβάνει όρους, βρόγχους, επαναλαμβανόμενες λειτουργίες που ορίζονται από το χρήστη και εγκαταστάσεις εισόδου και εξόδου.
- Διαθέτει αποτελεσματικό σύστημα διαχείρισης και αποθήκευσης δεδομένων.
- Παρέχει ομάδα χειριστών για υπολογισμούς σε πίνακες, λίστες, διανύσματα και μήτρες.
- Παρέχει μεγάλη, συνεκτική και ολοκληρωμένη συλλογή εργαλείων για την ανάλυση δεδομένων.
- Παρέχει γραφικές διευκολύνσεις για ανάλυση δεδομένων και εμφάνιση είτε απευθείας στον υπολογιστή είτε εκτύπωση.

Ως συμπέρασμα, η R είναι η πιο διαδεδομένη για μηχανική μάθηση γλώσσα προγραμματισμού στον κόσμο. Είναι η πρώτη επιλογή των επιστημόνων δεδομένων και υποστηρίζεται από ενεργητική και ταλαντούχο κοινότητα συνεργατών. Η R διδάσκεται στα πανεπιστήμια και αναπτύσσεται σε επιχειρησιακές εφαρμογές κρίσιμης σημασίας (Tutorialspoint).

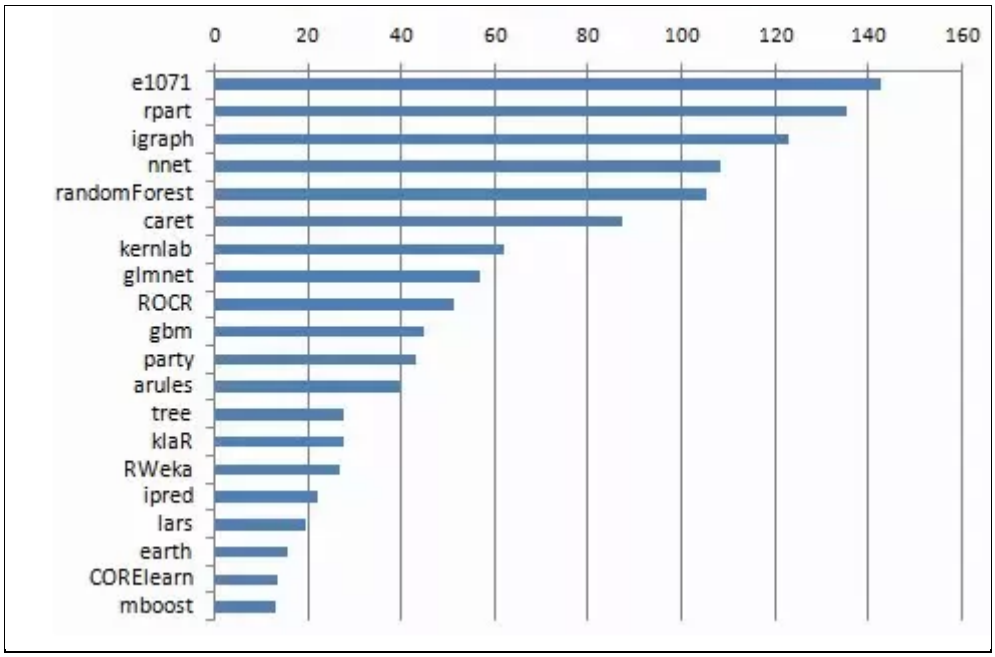
6.2 Εκτέλεση Αλγορίθμων Μηχανικής Μάθησης στη γλώσσα προγραμματισμού R

Στην παρούσα ενότητα θα παρουσιάσουμε πώς εκτελέστηκα στην γλώσσα προγραμματισμού R οι επτά εξεταζόμενοι αλγόριθμοι του προηγούμενου κεφαλαίου.

1. j48
2. Λογιστική Παλινδρόμηση
3. Τυχαία Δάση
4. Multilayer perceptron
5. SVM
6. Μπεϊζιανός Αλγόριθμος
7. Knn

Υπάρχουν πολλά πακέτα στη γλώσσα προγραμματισμού R τα οποία παρέχουν λύσεις για τεχνικές μάθησης. Βάσει του άρθρου της Geethika Bhavya Peddibhotla το 2015, τα πρώτα

είκοσι πακέτα R για Μηχανική Μάθηση που βασίζονται σε λήψεις απεικονίζονται στο παρακάτω γράφημα (Geethika Bhavya Peddibhotla, 2015):



Εικόνα 4: Τα πρώτα 20 πακέτα στην R για Μηχανική Μάθηση
(Geethika Bhavya Peddibhotla, 2015)

Στο πλαίσιο της ΜΔΕ, επιλέχθηκαν οι βιβλιοθήκες CARET και RWeka. Το πακέτο CARET (Classification and Regression Training) είναι ένα σύνολο λειτουργιών που προσπαθούν να εξορθολογήσουν τη διαδικασία δημιουργίας προγνωστικών μοντέλων. Το πακέτο περιέχει εργαλεία για:

- Διάσπαση δεδομένων
- Προεπεξεργασία
- Επιλογή χαρακτηριστικών
- Μοντελοποίηση μοντέλου χρησιμοποιώντας αναδειγματοληψία
- Εκτίμηση μεταβλητής σημασίας

καθώς και για άλλες λειτουργίες (Kuhn, 2018).

Το πακέτο RWeka κάνει διασύνδεση της γλώσσας προγραμματισμού R με το Weka και συγκεκριμένα την έκδοση 3.9.3 που περιέχει εργαλεία για την προεπεξεργασία δεδομένων, την ταξινόμηση, την παλινδρόμηση και την ομαδοποίηση. Το πακέτο RWeka περιέχει τον κώδικα διεπαφής μεταξύ των δύο προγραμμάτων (Kurt Hornik,Christian Buchta,Torsten Hothorn,Alexandros Karatzoglou,David Meyer,Achim Zeileis, 2018).

Επιπλέον, από το εργαλείο του Weka έχει εξαχθεί το αρχείο με τα επεξεργασμένα δεδομένα (φίλτρο SMOTE) το οποίο αντιγράφθηκε σε αρχείο excel, και με τη χρήση του εργαλείου text to columns του excel τα δεδομένα χωρίστηκαν σε στήλες, με σκοπό να αποθηκευτεί σε μορφή .xls.

Στον παρακάτω πίνακα απεικονίζονται τα αποτελέσματα των ταξινομητών στη γλώσσα προγραμματισμού R σε σχέση με αυτά του εργαλείου WEKA:

Πίνακας Αποτελεσμάτων 2									
Ταξινομητές	Φίλτρα	Σωστά Ταξινομημένα ως Αρνητικά (TN)	Λανθασμένα Ταξινομημένα ως Αρνητικά (FP)	Λανθασμένα Ταξινομημένα ως Θετικά (FN)	Σωστά Ταξινομημένα ως Θετικά (TP)	Ακρίβεια	Ειδικότητα (Precision)	Ευαισθησία (recall)	Μέτρο F
J48	R	402	98	94	302	78,57%	76%	76%	76%
		81%	25%	19%	76%				
	WEKA	365	135	91	305	74,78%	69%	77%	73%
		80%	31%	20%	69%				
Λογιστική Παλινδρόμηση	R	487	13	258	138	69,75%	91%	35%	50%
		65%	9%	35%	91%				
	WEKA	408	92	137	259	74,44%	74%	65%	69%
		75%	26%	25%	74%				
Τυχαία Δάση	R	431	105	69	291	80,58%	73%	81%	77%
		86%	27%	14%	73%				
	WEKA	397	103	83	313	79,24%	75%	79%	77%
		83%	25%	17%	75%				
Multilayer perceptron	R	490	260	10	136	69,87%	34%	93%	50%
		98%	66%	2%	34%				
	WEKA	367	133	71	325	77,23%	71%	82%	76%
		84%	29%	16%	71%				

SVM	R	486	14	267	129	68,64%	90%	33%	48%
		65%	10%	35%	90%				
	WEKA	410	90	155	241	72,66%	73%	61%	66%
		73%	27%	27%	73%				
Μπεϊζιανός Αλγόριθμος	R	496	269	4	127	69,53%	32%	97%	48%
		99%	68%	1%	32%				
	WEKA	405	95	153	243	72,32%	72%	61%	66%
		73%	28%	27%	72%				
kNN	R	402	123	98	273	75,33%	69%	74%	71%
		80%	31%	20%	69%				
	WEKA	369	131	83	313	76,12%	70%	79%	75%
		82%	30%	18%	70%				

Παρατηρούμε ότι ο ταξινομητής Τυχαία Δάση εμφανίζει τα καλύτερα αποτελέσματα και στα δύο προγράμματα.

Στον παρακάτω πίνακα απεικονίζονται οι διαφορές σε απόλυτες τιμές της ακρίβειας μεταξύ R και Weka για κάθε αλγόριθμο:

Πίνακας 9: Διαφορά Ακρίβειας σε απόλυτες τιμές μεταξύ R και WEKA

Ταξινομητές	Weka	R	Διαφορά σε απόλυτες τιμές
j48	74,78%	78,57%	3,79%
Λογιστική Παλινδρόμηση	74,44%	69,75%	4,69%
Τυχαία Δάση	79,24%	80,58%	1,34%
Multilayer perceptron	74,33%	69,88%	4,45%
SVM	72,66%	68,64%	4,02%
Μπεϊζιανός Αλγόριθμος	72,32%	69,53%	2,79%
kNN	76,12%	75,33%	0,79%
Μέσος Όρος Διαφοράς σε Απόλυτη Τιμή			3,12%

Παρατηρούμε ότι έχουμε ποσοστό απόκλισης περίπου 3% μεταξύ των αποτελεσμάτων του εργαλείου WEKA και της γλώσσας προγραμματισμού R, που θα μπορούσε να χαρακτηριστεί ως αποδεκτό ποσοστό απόκλισης, αν σκεφτούμε ότι συχνά υπάρχουν μικρές διαφοροποιήσεις στα αποτελέσματα ταξινόμησης κάθε φορά που εκτελούμε αλγόριθμο είτε στη γλώσσα προγραμματισμού R είτε στο εργαλείο WEKA.

Τέλος, έχει παρατηρηθεί σε αντίστοιχη έρευνα των Ahmad Al-Khoder και Hazar Harmouch ότι το εργαλείο του WEKA μπορεί και επιτυγχάνει μεγαλύτερα αποτελέσματα ακριβείας στο ίδιο σύνολο δεδομένων όχι μόνο από τη γλώσσα προγραμματισμού R αλλά και από άλλα εργαλεία, όπως το RapidMiner και το KNIME (Ahmad Al-Khoder, Hazar Harmouch).

Αυτή η μελέτη μπορεί να χρησιμοποιηθεί για να επιλεγεί ο καλύτερος Ταξινομητής για την πρόβλεψη του σακχαρώδους διαβήτη. Στο μέλλον μπορούμε να χρησιμοποιήσουμε αυτό το είδος μελέτης για οποιαδήποτε άλλη ασθένεια με τα κατάλληλα σύνολα δεδομένων.

ΚΕΦΑΛΑΙΟ 7: Συμπεράσματα και Μελλοντικές Κατευθύνσεις

7.1 Συγκριτική Παρουσίαση με άλλες μελέτες

Η προσπάθεια της παρούσας ΜΔΕ έγινε με σκοπό να συγκρίνει επτά αλγορίθμους για την ικανότητά τους να προβλέπουν με σημαντική ακρίβεια την ύπαρξη ή όχι του σακχαρώδους διαβήτη.

Οι αλγόριθμοι που συγκρίθηκαν ήταν:

- Αλγόριθμος J48
- Λογιστική Παλινδρόμηση
- Τυχαία Δάση
- Multilayer Perceptron
- SVM
- Μπεϊζιανός Αλγόριθμος
- kNN

Η σύγκριση των αλγορίθμων έγινε ως προς την ακρίβεια, την ευαισθησία, την εξειδίκευση και την προβλεπτική ικανότητά τους.

Για να γίνει η σύγκριση αυτή χρησιμοποιήθηκε το ίδιο σύνολο δεδομένων με τις μελέτες που αναφέρθηκαν στο κεφάλαιο 3. Τα δεδομένα αυτά χρησιμοποιήθηκαν χωρίς καμία προεπεξεργασία, στη συνέχεια έγινε προεπεξεργασία των δεδομένων με σειρά τεχνικών, προκειμένου να ελέγξουμε αν τα αποτελέσματα των αλγορίθμων θα ήταν περισσότερο βελτιωμένα.

Η πρώτη μελέτη δούλεψε πάνω σε τέσσερεις αλγορίθμους, μελετώντας την ακρίβεια, την ευαισθησία και την εξειδίκευσή τους τόσο σε προεπεξεργασμένα, όσο και σε μη επεξεργασμένα δεδομένα. Η επεξεργασία των δεδομένων έγινε αντικαθιστώντας τις ελλιπείς τιμές με τη διαδικασία του μέσου όρου. Τα αποτελέσματα των προεπεξεργασμένων δεδομένων ήταν καλύτερα.

Στην παρούσα ΜΔΕ παρατηρούμε ότι, εκπαιδεύοντας και αξιολογώντας τα μοντέλα μας με Διασταυρούμενη Επικύρωση ίση με 10, έχουμε καλύτερα αποτελέσματα μόνο στον Ταξινομητή Τυχαία Δάση (ακρίβεια= 79,24%), και αυτό μόνο σε σχέση με τους ταξινομητές Δέντρων Απόφασης και τον Μπεϊζιανό Αλγόριθμο. Ωστόσο, αν χρησιμοποιήσουμε τεχνικές εκπαίδευσης με χρήση δεδομένων εκπαίδευσης και αξιολόγησης, στους J48, Τυχαία Δάση και Μπεϊζιανό Αλγόριθμο με ποσοστό εκπαίδευσης 78%, 76% και 66% αντίστοιχα, που δεν προτιμήθηκε ως μοντέλο στην παρούσα ΜΔΕ, λόγω του ότι η επιδίωξη ήταν τα πιο αντικειμενικά αποτελέσματα, τότε, όπως φαίνεται και από τον παρακάτω πίνακα, μπορούμε να πετύχουμε μέγιστη ακρίβεια (81,22%) με τα Δέντρα Απόφασης, μεγαλύτερη από την πρώτη μελέτη. Επιπλέον, καθώς στα αποτελέσματα της έρευνας των Pradeep Kandhasamy J. και Balamurali S. δεν υπάρχουν πίνακες σύγχυσης για να δούμε πώς προκύπτουν τα ποσοστά ακρίβειας των ταξινομητών ώστε να έχουμε πλήρη εικόνα και εάν, για παράδειγμα, υπάρχουν φαινόμενα υπερεκπαίδευσης.

Πίνακας 10: Αποτελέσματα Σύγκρισης Πρώτης Μελέτης με παρούσα Εργασία

Ταξινομητής	Ακρίβεια με μέθοδο Pradeep Kandhasamy J. , Balamurali S.	Ακρίβεια στην παρούσα ΜΔΕ με χρήση δεδομένων εκπαίδευσης και αξιολόγησης	Ακρίβεια στην παρούσα ΜΔΕ με διασταυρούμενη επικύρωση= 10
Μπεϊζιανός Αλγόριθμος	**76.95%**	73,44%	72,32%
Λογιστική Παλινδρόμηση	**80.43%**	74,44%	74,44%
J48	76.52%	**81,22%**	74,78%
Τυχαία Δάση	76.52%	**80,93%**	**79,24%**

Στη δεύτερη μελέτη χρησιμοποιήθηκαν μόνο δύο αλγόριθμοι, προκειμένου να φτιαχτεί υβριδικό μοντέλο ικανό για πρόβλεψη που εκπαιδεύει τα δεδομένα και με τον Multilayer perceptron και με τον Μπεϊζιανό Αλγόριθμο, επιτυγχάνοντας την υψηλότερη ακρίβεια (81,89), μεγαλύτερη από το δικό μας μέγιστο (J48). Ωστόσο, αυτό το αποτέλεσμα έχει πολύ υψηλά επίπεδα ταξινόμησης των μη ασθενών και όχι των ασθενών, παρατηρείται δηλαδή το φαινόμενο της υπερεκπαίδευσης, καθώς ταξινομεί σωστά σχεδόν όλα τα αρνητικά που είναι η πλειονότητα, χωρίς όμως να προβλέπει τους ασθενείς, που είναι ο στόχος μας.

Υπολογίζοντας λοιπόν το μέτρο F, καθώς στη δεύτερη μελέτη δίνονται οι τιμές της ειδικότητας και της ευαισθησίας, βλέπουμε ότι το μέτρο F είναι ίσο με 75%, δηλαδή μικρότερο από αυτό το οποίο έχει επιτευχθεί στην παρούσα ΜΔΕ.

Πίνακας 11: Αποτελέσματα Σύγκρισης Δεύτερης Μελέτης με παρούσα Εργασία

Ταξινομητές	**Μέτρο F με χρήση δεδομένων εκπαίδευσης και αξιολόγησης**	**Μέτρο F με διασταυρούμενη επικύρωση = 10**
J48	**83%**	73%
Τυχαία Δάση	**82%**	**77%**
Multilayer perceptron	**76%**	**76%**
kNN	75%	75%

Από όσα παρουσιάστηκαν στην παρούσα ΜΔΕ γίνεται εύκολα αντιληπτό ότι τα δεδομένα εκπαίδευσης αποτελούν την πιο σημαντική παράμετρο για την επιτυχή πρόβλεψη του εκάστοτε μοντέλου. Ως αποτέλεσμα, η προεπεξεργασία των δεδομένων εξασφαλίζει την καλή ή μη καλή επίδοση του κάθε μοντέλου πρόβλεψης.

7.2 Προτάσεις για περαιτέρω βελτίωση των παραγόμενων αποτελεσμάτων

Στην παρούσα ΜΔΕ σκοπός ήταν η σύγκριση διαφορετικών ταξινομητών ως προς την πρόβλεψη του σακχαρώδους διαβήτη. Ωστόσο θα παρουσίαζε ιδιαίτερο ενδιαφέρον να γίνει χρήση ενός υβριδικού μοντέλου, έτσι όπως προτάθηκε στη δεύτερη μελέτη που παρουσιάστηκε, όπου θα συμμετείχε ο ταξινομητής Τυχαία Δάση. Ο ταξινομητής αυτός παρουσίασε τα καλύτερα αποτελέσματα στην έρευνά μας, αφού βέβαια έχει προηγηθεί

προεπεξεργασία στα δεδομένα, τόσο ως προς τη συμπλήρωση ελλιπών τιμών όσο και ως προς την εξισορρόπηση των κλάσεων, όπως παρουσιάστηκε στην παρούσα ΜΔΕ.

Τέλος, θα μπορούσε να γίνει περαιτέρω μελέτη σε περισσότερους ταξινομητές και τεχνικές μάθησης ώστε να βελτιωθούν ακόμα περισσότερο η ακρίβεια και η δυνατότητα διάγνωσης και εντοπισμού της ασθένειας.

Παράρτημα R

Εγκατάσταση πακέτων

```
> install.packages("caret")

> library(caret)

> install.packages("RWeka")

> library(RWeka)
```

Εντολή εισαγωγής δεδομένων χωρίς επεξεργασία στη R

```
> replacemissingvalues_smote <- read_excel("C:/Users/Grigorios Kalan
tonis/Desktop/replacemissingvalues_smote.xlsx",
+     col_types = c("numeric", "numeric", "numeric",
+         "numeric", "numeric", "numeric",
+         "numeric", "numeric", "text"))
```

Ορίζουμε ως τρόπο εκπαίδευσης τη Διασταυρούμενη Επικύρωση ίση με 10 και ως μέτρο αξιολόγησης την Ακρίβεια:

```
> control <- trainControl(method="cv", number=10)
> metric <- "Accuracy"
```

Τέλος, μερικοί αλγόριθμοι εκτελούν πολύ καλύτερα ορισμένες βασικές προεπεξεργασίες δε -δομένων. Για παράδειγμα, πολλοί αλγόριθμοι με βάση τα στιγμιότυπα λειτουργούν πολύ καλύτερα εάν όλες οι μεταβλητές εισόδου έχουν την ίδια κλίμακα.

Η συνάρτηση train () επιτρέπει να καθορίσουμε την προεπεξεργασία των δεδομένων που πρέπει να εκτελεστεί πριν από την εκπαίδευση. Αυτοί οι μετασχηματισμοί παρέχονται στη μεταβλητή preProcess:

```
> preProcess=c("center", "scale")
```

Το πακέτο RWeka απαιτεί η στήλη της κλάσης να αντιπροσωπεύεται από χαρακτήρα όχι αριθμητικής τιμής. Για να το αλλάξουμε αυτό χρησιμοποιούμε την παρακάτω μέθοδο:

```
> numOnly$class=factor(numOnly$class,levels = c(1,0),labels = c("pos
itive","negative"))
```

Παρατήρηση δεδομένων

Με τις παρακάτω εντολές θα ερευνήσουμε τα δεδομένα του αρχείου. Θα δούμε τον αριθμό περιπτώσεων που ανήκουν σε κάθε κατηγορία ως απόλυτη μέτρηση και ως ποσοστό.

Μετά από τη χρήση του φίλτρου SMOTE, οι κλάσεις έχουν ισορροπήσει.

```
> Percentage<- prop.table(table (replacemissingvalues_smote$class))*100
> cbind(freq=table(replacemissingvalues_smote$class),percentage=Percentage)
               freq percentage
tested_negative  500   55.80357
tested_positive  396   44.19643
```

Τέλος, με την εντολή summary μπορούμε να δούμε την περίληψη κάθε ιδιότητας.

Αυτό περιλαμβάνει τις μέσες, τις ελάχιστες και τις μέγιστες τιμές, καθώς και τα εκατοστημόρια (π.χ. 25η, 50η ή μέση και 75η τιμές σε αυτά τα σημεία):

```
> summary(replacemissingvalues_smote)
      preg                plas                pres                skin
 Min.   :       0   Min.   :       44   Min.   :       24   Min.   :       7
 1st Qu.:       1   1st Qu.:      102   1st Qu.:       66   1st Qu.:      27
 Median :       4   Median :      124   Median :       75   Median :      39
 Mean   :  570070   Mean   : 17614394   Mean   : 11617810   Mean   : 4371223
 3rd Qu.:       8   3rd Qu.:      163   3rd Qu.:       88   3rd Qu.: 2915342
 Max.   :13294468   Max.   :190562606   Max.   :104584477   Max.   :98768841
      insu                mass                pedi                age
 Min.   :       14   Min.   :       18   Min.   :0.000e+00   Min.   :       21
 1st Qu.:      140   1st Qu.:       28   1st Qu.:0.000e+00   1st Qu.:       25
 Median :155548223   Median :       34   Median :0.000e+00   Median :       32
 Mean   : 90648017   Mean   :  4948033   Mean   :1.182e+07   Mean   :  4565887
 3rd Qu.:155548223   3rd Qu.:       41   3rd Qu.:1.000e+00   3rd Qu.:       47
 Max.   :602819594   Max.   :49187921   Max.   :2.271e+09   Max.   :57013234
    class
 Length:896
 Class :character
 Mode  :character
```

Παρατηρούμε ότι οι εγγραφές του αρχείου είναι περισσότερες (896 από 768 του μη επεξεργασμένου αρχείου), καθώς, όπως αναφέρθηκε, το φίλτρο SMOTE επαναλαμβάνει σύνολο δεδομένων με την τεχνική συνθετικής μειοψηφίας για να ισορροπήσει τις δύο κλάσεις.

Εκτέλεση Ταξινομητών

Ταξινομητής J48

Με την παρακάτω εντολή κάνουμε fit το μοντέλο μας χρησιμοποιώντας τη μέθοδο j48:

```
> diabetes_j48 <- J48(class ~ ., data = numOnly)
```

Στη συνέχεια εμφανίζουμε τα αποτελέσματά του στην παρακάτω εικόνα:

```
> summary(diabetes_j48)

=== Summary ===

Correctly Classified Instances         780               87.0536 %
Incorrectly Classified Instances       116               12.9464 %
Kappa statistic                          0.7347
Mean absolute error                      0.1943
Root mean squared error                  0.3117
Relative absolute error                 39.3867 %
Root relative squared error             62.7597 %
Total Number of Instances              896

=== Confusion Matrix ===

   a   b   <-- classified as
 318  78 |   a = positive
  38 462 |   b = negative
```

```
> eval_j48 <- evaluate_Weka_classifier(diabetes_j48, numFolds = 10,
complexity = FALSE,seed = 1, class = TRUE)
```

```
> eval_j48
=== 10 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances         704               78.5714 %
Incorrectly Classified Instances       192               21.4286 %
Kappa statistic                          0.566
Mean absolute error                      0.2675
Root mean squared error                  0.4026
Relative absolute error                 54.2222 %
Root relative squared error             81.0749 %
Total Number of Instances              896

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,763    0,196    0,755      0,763    0,759      0,566    0,844     0,773     positive
                 0,804    0,237    0,810      0,804    0,807      0,566    0,844     0,860     negative
Weighted Avg.    0,786    0,219    0,786      0,786    0,786      0,566    0,844     0,822

=== Confusion Matrix ===

   a   b   <-- classified as
 302  94 |   a = positive
  98 402 |   b = negative
```

Για το J48 πετυχαίνουμε **Ακρίβεια ίση με 78,57%.**

Ταξινομητής Random Forest

Όπως και στον προηγούμενο αλγόριθμο, εκπαιδεύουμε το μοντέλο μας με τη μέθοδο rf:

```
> fit.rf <- train(class~., data=replacemissingvalues_smote, method="
rf", metric=metric, trControl=control)
```

Και παίρνουμε τα ακόλουθα αποτελέσματα:

```
> fit.rf
Random Forest

896 samples
  8 predictor
  2 classes: 'tested_negative', 'tested_positive'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 806, 807, 806, 806, 806, 807, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  2     0.8058427  0.6020064
  5     0.7946816  0.5813399
  8     0.7957803  0.5837853

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

Παρατηρούμε ότι η μέθοδος kNN του πακέτου CARET εμφανίζει την Ακρίβεια με το καλύτερο αποτέλεσμα και μας προτείνει αριθμό γειτόνων ίσο με 2, με το οποίο επιτυγχάνεται **Ακρίβεια ίση με 80%.**

Ο Πίνακας Σύγχυσης εμφανίζεται με χρήση της ομώνυμης συνάρτησης:

```
> rfCM <- confusionMatrix(fit.rf, norm = "none")
```

```
> rfCM
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

                 Reference
Prediction        tested_negative tested_positive
  tested_negative             431             105
  tested_positive              69             291

 Accuracy (average) : 0.8058
```

Ταξινομητής Multilayer Perceptron

Με την κατάλληλη μέθοδο, κάνουμε fit το μοντέλο μας για τον Multilayer Perceptron:

```
> fit.mlp <- train(class~., data=numOnly,method='mlp', metric=metric
, preProc=c("center", "scale"), trControl=control, fit=FALSE,maxit=5
0,hidden1=1)
```

```
> fit.mlp
Multi-Layer Perceptron

896 samples
  8 predictor
  2 classes: 'positive', 'negative'

Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 806, 807, 806, 806, 807, 806, ...
Resampling results across tuning parameters:

  size  Accuracy      Kappa
  1     0.6975655431  0.3439495467
  3     0.6920099875  0.3320843801
  5     0.6986766542  0.3461885621

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was size = 5.
```

Εμφανίζοντας και τον Πίνακα Σύγχυσης:

```
> mlpCM <- confusionMatrix(fit.mlp, norm = "none")
```

```
> mlpCM
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

          Reference
Prediction positive negative
  positive      136       10
  negative      260      490

 Accuracy (average) : 0.6987
```

Το βέλτιστο αποτέλεσμα είναι **Ακρίβεια ίση με 69,87%.**

Ταξινομητής Naïve Bayes

Γίνεται χρήση της μεθόδου "nb":

```
> fit.nb <- train(class~., data=replacemissingvalues_smote, method="
nb", metric=metric, trControl=control)
```

```
> fit.nb
Naive Bayes

896 samples
  8 predictor
  2 classes: 'tested_negative', 'tested_positive'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 807, 807, 807, 806, 807, 806, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.6953184  0.33525114
   TRUE      0.5903620  0.08028351

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter
 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = FALSE and adjust = 1.
```

Τέλος, εμφανίζουμε τον Πίνακα Σύγχυσης με τις σωστές και με τις λανθασμένες ταξινομήσεις ανά κλάση:

```
> nbCM <- confusionMatrix(fit.nb, norm = "none")
```

```
> nbCM
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

                 Reference
Prediction        tested_negative tested_positive
  tested_negative             496             269
  tested_positive               4             127

 Accuracy (average) : 0.6953
```

Για c= 1 η **Ακρίβεια είναι ίση με 69,53%,** το οποίο είναι και το καλύτερο αποτέλεσμα.

Ταξινομητής kNN

Με την παρακάτω συνάρτηση εκπαιδεύουμε το μοντέλο μας:

```
> fit.knn <- train(class~., data=replacemissingvalues_smote, method=
"knn", metric=metric, preProc=c("center", "scale"), trControl=contro
l)
```

Τα αποτελέσματα είναι ορατά στις παρακάτω εικόνες:

```
> fit.knn
k-Nearest Neighbors

896 samples
  8 predictor
  2 classes: 'tested_negative', 'tested_positive'

Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 807, 807, 806, 806, 807, 806, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.7477154  0.4861955
  7  0.7532834  0.4962190
  9  0.7377403  0.4645834

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 7.
```

```
> knnCM <- confusionMatrix(fit.knn, norm = "none")
```

```
> knnCM
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

                Reference
Prediction       tested_negative tested_positive
  tested_negative             402             123
  tested_positive              98             273

 Accuracy (average) : 0.7533
```

Παίρνουμε **Ακρίβεια ίση με 75,33%.**

Ταξινομητής Logistic Regression με χρήση πακέτου r-weka

Με την παρακάτω συνάρτηση εκπαιδεύουμε το μοντέλο μας:

```
> logistic<-Logistic(class ~ ., data = numOnly)
```

```
> summary(logistic)

=== Summary ===

Correctly Classified Instances         630               70.3125 %
Incorrectly Classified Instances       266               29.6875 %
Kappa statistic                          0.3573
Mean absolute error                      0.3857
Root mean squared error                  0.4389
Relative absolute error                 78.1983 %
Root relative squared error             88.3744 %
Total Number of Instances              896

=== Confusion Matrix ===

   a   b   <-- classified as
 142 254 |   a = positive
  12 488 |   b = negative
```

```
> eval_logistic<- evaluate_Weka_classifier(logistic, numFolds = 10,
complexity = FALSE,seed = 1, class = TRUE)
```

Τα αποτελέσματα φαίνονται στην παρακάτω εικόνα:

```
> eval_logistic
=== 10 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances         625               69.7545 %
Incorrectly Classified Instances       271               30.2455 %
Kappa statistic                          0.3447
Mean absolute error                      0.388
Root mean squared error                  0.4416
Relative absolute error                 78.6577 %
Root relative squared error             88.9163 %
Total Number of Instances              896

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,348    0,026    0,914      0,348    0,505      0,428    0,686     0,720     positive
                 0,974    0,652    0,654      0,974    0,782      0,428    0,686     0,674     negative
Weighted Avg.    0,698    0,375    0,769      0,698    0,660      0,428    0,686     0,694

=== Confusion Matrix ===

   a   b   <-- classified as
 138 258 |   a = positive
  13 487 |   b = negative
```

Αποτέλεσμα ακρίβειας ίσο με 69,75%.

Ταξινομητής SVM με χρήση πακέτου R-Weka

Με την παρακάτω συνάρτηση εκπαιδεύουμε το μοντέλο μας:

```
> svm<-SMO(class ~ ., data = numOnly)
```

```
> summary(svm)

=== Summary ===

Correctly Classified Instances         621               69.308  %
Incorrectly Classified Instances       275               30.692  %
Kappa statistic                          0.3327
Mean absolute error                      0.3069
Root mean squared error                  0.554
Relative absolute error                 62.2203 %
Root relative squared error            111.5547 %
Total Number of Instances              896

=== Confusion Matrix ===

   a   b   <-- classified as
 130 266 |   a = positive
   9 491 |   b = negative
```

```
> eval_svm<- evaluate_Weka_classifier(svm, numFolds = 10, complexity
= FALSE,seed = 1, class = TRUE)
```

Τα αποτελέσματα είναι ορατά στη παρακάτω εικόνα:

```
> eval_svm
=== 10 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances         615               68.6384 %
Incorrectly Classified Instances       281               31.3616 %
Kappa statistic                          0.319
Mean absolute error                      0.3136
Root mean squared error                  0.56
Relative absolute error                 63.5774 %
Root relative squared error            112.7646 %
Total Number of Instances              896

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,326    0,028    0,902      0,326    0,479      0,404    0,649     0,592     positive
                 0,972    0,674    0,645      0,972    0,776      0,404    0,649     0,643     negative
Weighted Avg.    0,686    0,389    0,759      0,686    0,644      0,404    0,649     0,620

=== Confusion Matrix ===

   a   b   <-- classified as
 129 267 |   a = positive
  14 486 |   b = negative
```

Ποσοστό ακρίβειας ίσο με 68,63%.

Αναφορές

Geethika Bhavya Peddibhotla. (2015). *Top 20 R Machine Learning and Data Science packages*. Ανάκτηση 12 20, 2018, από https://www.kdnuggets.com/2015/06/top-20-r-machine-learning-packages.html

Abernethy M. (2010). *Nearest Neighbor and server-side library*. Ανάκτηση 08 01, 2018, από https://www.ibm.com/developerworks/library/os-weka3/index.html

Abhinav Maurya, PhD Student at CMU. (2016). *Are some ML algorithms more vulnerable to unbalanced training sets than others? Why?* Ανάκτηση 10 1, 2018, από https://www.quora.com/Are-some-ML-algorithms-more-vulnerable-to-unbalanced-training-sets-than-others-Why

Agresti A. (1996). *An Introduction to Categorical Data Analysis* (2nd ed. εκδ.). New York: Wiley.

Ahmad Al-Khoder, Hazar Harmouch. (n.d.). *Evaluating four of the most popular Open Source and Free Data*. Ανάκτηση 01 15, 2019, από https://pdfs.semanticscholar.org/94d7/c3e183b5a5513c087cf9fff3e9d41d75a78a.pdf

Amatul Zehra,Tuty Asmawaty,M.A M. Aznan. (2013). *A comparative study on the pre-processing and mining*. Ανάκτηση 12 20, 2018, από https://pdfs.semanticscholar.org/2450/721b16959b477c3504759729bc782f4a8d0e.pdf

Amit kumar Dewangan, Pragati Agrawal. (2015). *Classification of Diabetes Mellitus Using Machine Techniques*. Ανάκτηση 11 5, 2018, από https://www.ijeas.org/download_data/IJEAS0205060.pdf

Bishop C. (1995). Neural Networks for Pattern Recognition. Oxford: Clarendon Press.

Brownlee J. (2014). *Classification Accuracy is Not Enough: More Performance Measures You Can Use*. Ανάκτηση 10 04, 2018, από https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/

Brownlee J. (2016). *Supervised and Unsupervised Machine Learning Algorithms*. Ανάκτηση 07 29, 2018, από 5. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Daniel Lee, works at Booz Allen Hamilton. (2017). *Is it better using training-test split or k-fold CV, when we are working with large datasets?* Ανάκτηση 12 17, 2018, από https://www.quora.com/Is-it-better-using-training-test-split-or-k-fold-CV-when-we-are-working-with-large-datasets

Dr. Asir Antony Gnana Singh D., κ.ά. (2017). *Diabetes Prediction Using Medical Data*. Ανάκτηση 08 10, 2018, από https://www.researchgate.net/publication/316432650_Diabetes_Prediction_Using_Medical_Data?enrichId=rgreq-6712e4e2a0fb2fb98ca2553234f8a518-XXX&enrichSource=Y292ZXJQYWdlOzMxNjQzMjY1MDtBUzo0ODY2MTQ4NDY3MDk3NjBAMTQ5MzAyOTQyODM5Mg%3D%3D&el=1_x_2&_esc=publicati

FileInfo - The File Extensions Database. (2012). *.ARFF File Extension*. Ανάκτηση 09 01, 2018, από https://fileinfo.com/extension/arff

GeeksforGeeks, n.d. (n.d.). *K-Nearest Neighbours*. Ανάκτηση 08 02, 2018, από https://www.geeksforgeeks.org/k-nearest-neighbours/

Han J., KanberM. Pei J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufman.

Kavakiotis I., et al. (2017). *Diabetes Prediction Using Medical Data*. Ανάκτηση 08 23, 2018, από Diabetes Prediction Using Medical Data: https://www.researchgate.net/publication/316432650_Diabetes_Prediction_Using_Medical_Data?enrichId=rgreq-6712e4e2a0fb2fb98ca2553234f8a518-XXX&enrichSource=Y292ZXJQYWdlOzMxNjQzMjY1MDtBUzo0ODY2MTQ4NDY3MDk3NjBAMTQ5MzAyOTQyODM5Mg%3D%3D&el=1_x_2&_esc=publicati

Kent Munthe Caspersen. (2015). *What is the influence of C in SVMs with linear kernel?* Ανάκτηση 07 14, 17, από https://stats.stackexchange.com/questions/31066/what-is-the-influence-of-c-in-svms-with-linear-kernel

Kuhn, M. (2018). *The caret Package*. Ανάκτηση 09 24, 2018, από http://topepo.github.io/caret/index.html

Kurt Hornik,Christian Buchta,Torsten Hothorn,Alexandros Karatzoglou,David Meyer,Achim Zeileis. (2018). *Package 'RWeka'*. Ανάκτηση 12 20, 2018, από https://cran.r-project.org/web/packages/RWeka/RWeka.pdf

NongyaoNai-arun, RungruttikarnMoungmai. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. *Original Research Article Procedia Computer Science*, σσ. 132-142.

Pears R., Connor A. (2014). *Synthetic Minority Over-sampling TEchnique*. Ανάκτηση 10 03, 2018, από https://arxiv.org/ftp/arxiv/papers/1407/1407.2330.pdf

Pradeep Kandhasamy J., Balamurali S. (2015). *Performance Analysis of Classifier Models to Predict Diabetes Mellitus*. Ανάκτηση 07 25, 2018, από https://www.sciencedirect.com/science/article/pii/S1877050915004500

Qassim. (2007). *Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century?* Ανάκτηση 08 20, 2018, από https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3068646/

Quinlan JR. (1986). Machine Learning 1. Στο *Induction of decision tree* (σσ. 81-106). Khuwer Academic Publisher.

Saxena R. (2017). *HOW THE NAIVE BAYES CLASSIFIER WORKS IN MACHINE LEARNING*. Ανάκτηση 08 12, 2018, από http://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/

Sittidech P., Nai-arun N. (2014). Random Forest Analysis on Diabetes Complication Data. *Proceeding of the IASTED International Conference*, σσ. 315-320.

Sunil R. (2017). *Understanding Support Vector Machine algorithm from examples (along with code)*. Ανάκτηση 08 23, 2018, από https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Tabaei B., Herman W. (2002). *A Multivariate logistic regression equation to screen for diabetes*. Ανάκτηση 08 02, 2018, από http://care.diabetesjournals.org/content/diacare/25/11/1999.full.pdf

TIBCO Statistica™. (2018). *Naive Bayes Classifier*. Ανάκτηση 07 19, 2018, από http://www.statsoft.com/Textbook/Naive-Bayes-Classifier#index

Tutorialspoint. (n.d.). *R Tutorial*. Ανάκτηση 08 14, 2018, από https://www.tutorialspoint.com/r/index.htm

Twain, J. (2014). *Cohen's kappa in plain English*. Ανάκτηση 10 03, 2018, από https://stats.stackexchange.com/questions/82162/cohens-kappa-in-plain-english

UCI Machine Learning repository. (n.d.). Ανάκτηση 04 19, 2019, από https://archive.ics.uci.edu/ml/index.php

Κύρκος Ε. (2015). *Αποθετήριο Κάλλιπος : Οδηγός WEKA*. Ανάκτηση 06 15, 2018, από https://repository.kallipos.gr/bitstream/11419/1239/2/Kef._13.pdf

Κωτσιαντής Σ. (2006). *Ομάδες Ταξινομητών για την αύξηση της ακρίβειας των μεθόδων Μηχανικής Μάθησης και Εξόρυξης Γνώσης*. Ανάκτηση 07 10, 2018, από http://nemertes.lis.upatras.gr/jspui/bitstream/10889/244/1/326.pdf

Μπακατσέλος Σ. (2007). *Σακχαρώδης Διαβήτης κύησης ,21ο συνέδριο διαβητολογικής εταιρείας βορείου Ελλάδος*. Ανάκτηση 08 15, 2018, από http://medicalrecords.gr/debe_21/speakers/diafaneies/pempti/18.30-19.30/mpakatselos.ppt

Παπαποστόλου Σ. (2017). *ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΜΕ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ*. Ανάκτηση 06 24, 2019, από https://ikee.lib.auth.gr/record/288589/files/GRI-2017-18958.pdf

ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ Δ/ΝΣΗ ΠΡΩΤΟΒΑΘΜΙΑΣ ΦΡΟΝΤΙΔΑΣ ΥΓΕΙΑΣ. (2016). *Παγκόσμια Ημέρα Διαβήτη (World Diabetes Day) - 14η Νοεμβρίου 2016*. Ανάκτηση 01 05, 2019, από http://www.3ype.gr/uploads/xartis/PagkosmiesHmeres/Diavitis.pdf