ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ»

ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

Μεταπτυχιακή Διπλωματική Εργασία

Πρόβλεψη Σακχαρώδους Διαβήτη με Μεθόδους Μηχανικής Μάθησης

Κολοκυθάς Κωνσταντίνος

ΑΜ: ΜΕ1930

Επιβλέπων καθηγητής: Φιλιππάκης Μιχαήλ

Πειραιάς, Νοέμβριος, 2021

Ευχαριστίες

Με την παρούσα διπλωματική εργασία ολοκληρώνονται οι σπουδές μου στο μεταπτυχιακό πρόγραμμα σπουδών «Πληροφοριακά Συστήματα & Υπηρεσίες» του Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς. Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου κ. Μιχαήλ Φιλιππάκη για την εμπιστοσύνη που μου έδειξε στην εκπόνηση της παρούσας διπλωματικής εργασίας, καθώς και για όλες τις υποδείξεις και συμβουλές του.

Στις σπουδές μου ήταν καθοριστική η συμβολή των καθηγητών μου στα γνωστικά αντικείμενα που παρακολούθησα, στους οποίους καιοφείλω να εκφράσω τις ειλικρινείς μου ευχαριστίες για τη συμβολή τους στην ολοκλήρωση των σπουδών μου. Επίσης θα ήθελα να ευχαριστήσω τη Δρ. Μαρία Ελένη Πούλου για την πολύτιμή βοήθεια της στην επίβλεψη της διπλωματικής.

Ιδιαίτερα θερμές ευχαριστίες θέλω να δώσω στην οικογένεια μου για την συνεχείς συμπαράσταση τους, για τις πολύτιμές συμβουλές τους και για όλα όσα μου έχουν προσφέρει όλα αυτά τα χρόνια της ζωής μου αλλά και των σπουδών μου.

Περίληψη

Ο σακχαρώδης διαβήτης είναι μία από τις σοβαρότερες ασθένειες σε παγκόσμιο επίπεδο και, εάν δεν διαγνωστεί έγκαιρα, μπορεί να δημιουργήσει σοβαρά προβλήματα υγείας και να αυξήσει τον κίνδυνο της θνησιμότητας. Μάλιστα, οι ασθενείς είναι ιδιαίτερα ευάλωτοι στην ασθένεια του COVID-19. Η χρήση προγνωστικών μεθόδων με τη βοήθεια της εξόρυξης δεδομένων και ιδιαίτερα της Μηχανικής Μάθησης μπορεί να συμβάλει στην έγκαιρη διάγνωση, ώστε να αποφευχθούν οι σοβαρές επιπλοκές στην υγεία των ασθενών.

Στην παρούσα διπλωματική εργασία διερευνάται το πρόβλημα της πρόβλεψης του σακχαρώδους διαβήτη βάσει συμπτωμάτων με την βοήθεια αλγόριθμων Μηχανικής Μάθησης για δυαδική ταξινόμηση, με σκοπό τη σύγκρισή τους. Η υλοποίηση έγινε με τη βοήθεια της βιβλιοθήκης scikit-learn της γλώσσας προγραμματισμού Python με ένα συγκεκριμένο σύνολο δεδομένων που διατίθεται δημόσια από το αποθετήριο μηχανικής μάθησης του Πανεπιστημίου της Καλιφόρνια. Το σύνολο δεδομένων αφορά παρατηρήσεις που συλλέχθηκαν από 520 ασθενείς βάσει ερωτηματολογίου, το οποίο περιλαμβάνει τα συχνότερα συμπτώματα της ασθένειας.

Οι αλγόριθμοι μηχανικής μάθησης που εφαρμόστηκαν στο σύνολο δεδομένων είναι η λογιστική παλινδρόμηση (Logistic Regression), οι μηχανές υποστήριξης διανυσμάτων (Support Vector Machines), ο απλοϊκός Bayes (Naive Bayes), το πολυεπίπεδο perceptron (Multi-Layer Perceptron) και τα τυχαία δάση (Random Forests). Η σύγκριση των αποτελεσμάτων έγινε με διάφορες μετρικές απόδοσης που αφορούν προβλήματα ταξινόμησης. Τα συγκριτικά αποτελέσματα υποδεικνύουν ως καλύτερο αλγόριθμο για το συγκεκριμένο σύνολο τα τυχαία δάση. Το σχετικό μοντέλο παρουσιάζει τις καλύτερες μετρικές απόδοσης, με το μοντέλο των μηχανών υποστήριξης διανυσμάτων να ακολουθεί.

Λέξεις – Κλειδιά

Σακχαρώδης Διαβήτης, Εξόρυξη Δεδομένων, Μηχανική Μάθηση, Δυαδική Ταξινόμηση

Abstract

Diabetes mellitus is one of the most serious diseases worldwide and, if not diagnosed early, can cause serious health problems and increase the risk of mortality. In fact, patients are particularly vulnerable to COVID-19 disease. The use of prognostic methods with the help of data mining and especially Machine Learning can help in early diagnosis, to avoid serious complications in the health of patients.

In this work, the problem of predicting diabetes based on symptoms is investigated with the help of Machine Learning algorithms for binary classification, to compare them. The implementation was done with scikit-learn library which is based on Python programming language with a specific dataset, publicly available from UCI machine learning repository. The dataset refers to observations collected from 520 patients based on a questionnaire, which includes the most common symptoms of the disease.

The Machine Learning algorithms applied to the dataset are Logistic Regression, Support Vector Machines, Naive Bayes, Multi-Layer Perceptron, and Random Forests. The results were compared with various performance metrics related to classification problems. The comparative results indicate as the best algorithm for this dataset the random forests. The relevant model presents the best performance metrics, with the model of vector support machines following.

Keywords

Diabetes Mellitus, Data Mining, Machine Learning, Binary Classification

Περιεχόμενα

Πίνακας Σχημάτων

Πίνακας Πινάκων

Πίνακας Συντομογραφιών & Ακρωνυμίων

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CV	Cross Validation
DL	Deep Learning
EDA	Exploratory Data Analysis
LG	Logistic Regression
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MLP	Multi Layer Perceptron
NB	Naïve Bayes
RF	Random Forest
ROC	Receiving Operating Characteristic
SVM	Support Vector Machine
ΒΜ	Βαθιά Μάθηση
ΜΜ	Μηχανική Μάθηση
ΠΟΥ	Παγκόσμιος Οργανισμός Υγείας
ΣΔ	Σακχαρώδης Διαβήτης
ΤΝ	Τεχνητή Νοημοσύνη
ΤΝΔ	Τεχνητό Νευρωνικό Δίκτυο

1 Εισαγωγή

1.1 Εισαγωγή

Στην εποχή των Μεγάλων Δεδομένων, ιδιαίτερη αξία έχει αποκτήσει η εξαγωγή χρήσιμης πληροφορίας από τον μεγάλο όγκο δεδομένων, η Εξόρυξη Δεδομένων. Η Μηχανική Μάθηση (ΜΜ) είναι πλέον η κυρίαρχη μέθοδος για την εξόρυξη δεδομένων και αναπτύσσεται ραγδαία, παράλληλα με την εξέλιξη των δυνατοτήτων των υπολογιστικών συστημάτων σε υλικό και λογισμικό [1].

Η ΜΜ μπορεί να διευκολύνει πολύ περίπλοκες και χρονοβόρες εργασίες και βρίσκει εφαρμογή στην επίλυση περίπλοκων προβλημάτων σε ένα ευρύ φάσμα εφαρμογών του πραγματικού κόσμου που αφορούν τους τομείς της ιατρικής, της ψυχαγωγίας, της εκπαίδευσης κλπ. [1]. Στον ιατρικό τομέα, η ΜΜ μπορεί να διευκολύνει πολύ περίπλοκες και χρονοβόρες εργασίες και ιδιαίτερη θέση, μεταξύ των άλλων, κατέχει η διάγνωση ασθενειών σε πρώιμα στάδια, όπως για παράδειγμα της ασθένειας του σακχαρώδους διαβήτη [2].

Ο σακχαρώδης διαβήτης είναι μια χρόνια ασθένεια η οποία πλήττει ένα μεγάλο μέρος του παγκόσμιου πληθυσμού. Εάν δεν διαγνωστεί έγκαιρα προκειμένου ο ασθενής να ακολουθήσει την ενδεδειγμένη θεραπευτική αγωγή, μπορεί να προκληθούν σοβαρές βλάβες στον οργανισμό που υποβαθμίζουν την ποιότητα ζωής του ασθενούς, τον καθιστούν ευάλωτο σε άλλες ασθένειες και μπορεί να οδηγήσουν ακόμη και στον θάνατο [3].

Η αυξητική τάση περιπτώσεων σακχαρώδους διαβήτη, έχει ως αποτέλεσμα την παραγωγή Μεγάλων Δεδομένων τα οποία έχουν σημαντική χρήση στην προώθηση της κλινικής και ιατρικής έρευνας, και πολύ περισσότερο στην εφαρμογή της εξόρυξης των δεδομένων και της ΜΜ στον προαναφερόμενο τομέα. Ως εκ τούτου, η δημιουργία αξιόπιστων συστημάτων πρόγνωσης με τη βοήθεια του υπολογιστή και της ΜΜ είναι σημαντική βοήθεια για τον ιατρικό κόσμο, ώστε η πρόγνωση να είναι ταχύτερη και ευκολότερη, προκειμένου να λαμβάνονται όσο το δυνατόν πιο σωστές αποφάσεις για την υγεία του ασθενούς, χωρίς να απαιτείται θεωρητικό και τεχνητό υπόβαθρο για την ΜΜ [2].

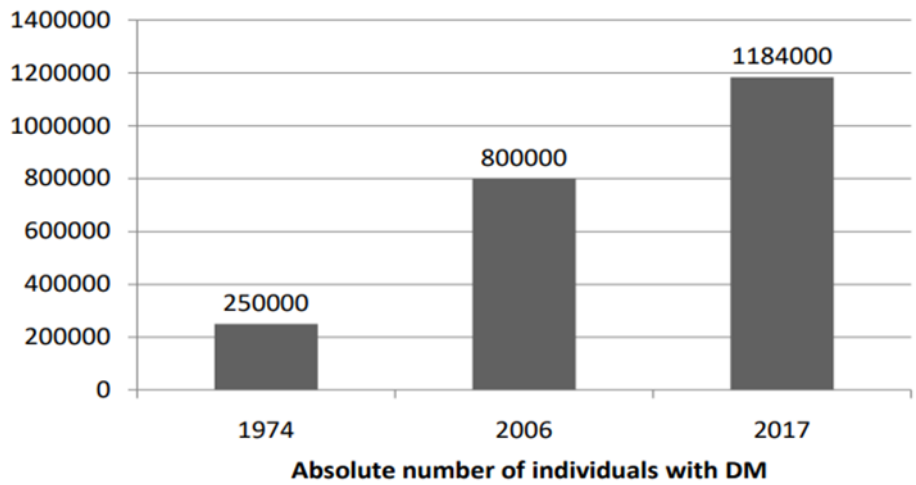
1.2 Ορισμός του προβλήματος

Ο σακχαρώδης διαβήτης είναι μία χρόνια ασθένεια η οποία οφείλεται στην αδυναμία του ανθρώπινου οργανισμού στη διαχείριση μιας ορμόνης, της ινσουλίνης, η

οποία είναι ορμόνη που παράγεται από το πάγκρεας και ρυθμίζει το σάκχαρο του αίματος. Η αδυναμία μπορεί να οφείλεται στο ότι το πάγκρεας δεν παράγει αρκετή ινσουλίνη ή δεν μπορεί να διαχειριστεί αποτελεσματικά την ινσουλίνη που παράγει. Σε κάθε περίπτωση το αποτέλεσμα είναι η υπεργλυκαιμία, δηλαδή το αυξημένο σάκχαρο στο αίμα, η οποία με την πάροδο του χρόνου προκαλεί σοβαρές βλάβες στον οργανισμό. Ο διαβήτης είναι μία από της κύριες αιτίες της τύφλωσης, της νεφρικής ανεπάρκειας, των καρδιακών προσβολών, των εγκεφαλικών επεισοδίων και του ακρωτηριασμού των κάτω άκρων [3].

Σύμφωνα με επίσημα στοιχεία του Παγκόσμιου Οργανισμού Υγείας (ΠΟΥ), ο αριθμός των ατόμων με διαβήτη αυξήθηκε από 108 εκατομμύρια το 1980 σε 422 εκατομμύρια το 2014 και αναμένεται να φτάσει στα 642 εκατομμύρια το 2040, πράγμα που σημαίνει ότι 1 στους 10 ενήλικες θα έχει την ασθένεια [3], [4]. Το 2019 περίπου 1,5 εκατομμύρια θάνατοι οφειλόταν άμεσα στον διαβήτη, παρουσιάζοντας μια αύξηση της τάξης του 70% παγκόσμια μεταξύ του 2000 και του 2019, με αύξηση των θανάτων κατά 80% στους άνδρες. Αξίζει να σημειωθεί ότι, στην Ανατολική Μεσόγειο οι θάνατοι από διαβήτη έχουν υπερδιπλασιαστεί και αντιπροσωπεύουν την μεγαλύτερη ποσοστιαία αύξηση σε παγκόσμιο επίπεδο [5].

Σύμφωνα με τα στοιχεία της Διεθνούς Ομοσπονδίας για τον Διαβήτη για την Ελλάδα που αναφέρει η Μήτρου [4], το 2015 ο επιπολασμός του σακχαρώδη διαβήτη υπολογίζονταν σε 7,5%, οι θάνατοι που σχετίζονται με το διαβήτη σε 4.963 και το μέσο ετήσιο κόστος ανά ασθενή σε 2.562 ευρώ. Οι Lupa et al. [6] στην επισκόπησή τους από επιδημιολογικές μελέτες που έχουν γίνει στην Ελλάδα, εκτιμούν ότι, το 2017 οι ασθενείς με διαβήτη ανέρχονται σε 1.184 εκατομμύρια, παρουσιάζοντας αύξηση της τάξης του 50% σε σχέση με το 2006, όπως φαίνεται στο Σχήμα 1.



Σχήμα 1. Εκτιμώμενος Αριθμός Ασθενών στην Ελλάδα το 2017 [6]

Από τον Δεκέμβριο του 2019 που εμφανίστηκε η σοβαρή ασθένεια του COVID-19 και εξελίχθηκε σε πανδημία, μέχρι τον Απρίλιο του 2021 έχασαν τη ζωή τους περίπου 3 εκατομμύρια άνθρωποι [7]. Υπάρχουν στοιχεία που καταδεικνύουν ότι, οι ασθενείς με διαβήτη διατρέχουν μεγαλύτερο κίνδυνο επιπλοκών και θνησιμότητας λόγω του COVID-19 [8].

Από όλα τα παραπάνω, γίνεται αντιληπτό πόσο σοβαρό είναι το πρόβλημα της ασθένειας του διαβήτη και πόσο σημαντική είναι η πρόβλεψη της ασθένειας σε πρώιμα στάδια.

1.3 Δομή Μεταπτυχιακής Διπλωματικής Εργασίας

Το Κεφάλαιο 1 της εργασίας αποτελεί την εισαγωγή στο θέμα της εργασίας. Παρουσιάζεται συνοπτικά το πρόβλημα της ασθένειας του σακχαρώδους διαβήτη, η δομή της εργασίας, καθώς και η συνεισφορά της.

Στο Κεφάλαιο 2, αρχικά δίνεται συνοπτικά το θεωρητικό υπόβαθρο για την ασθένεια του σακχαρώδους διαβήτη. Στη συνέχεια αναφέρονται συνοπτικά βασικά στοιχεία για τη ΜΜ, παρουσιάζονται οι κατηγορίες της ΜΜ, οι κυριότερες εφαρμογές της και οι σημαντικότεροι αλγόριθμοι δυαδικής ταξινόμησης.

Στο Κεφάλαιο 3 παρουσιάζονται οι σχετικές ακαδημαϊκές εργασίες για την πρόβλεψη του διαβήτη με μεθόδους ΜΜ, που αφορούν το ίδιο σύνολο δεδομένων με αυτό που χρησιμοποιήθηκε στην παρούσα εργασία.

Στο Κεφάλαιο 4 παρουσιάζεται το περιβάλλον υλοποίησης σε Python, και η μεθοδολογία για την ανάπτυξη του έργου.

Στο Κεφάλαιο 5 παρουσιάζονται τα συγκριτικά αποτελέσματα της υλοποίησης των αλγόριθμων και συζητούνται τα αποτελέσματα.

Στο Κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα και οι μελλοντικές κατευθύνσεις.

1.4 Συνεισφορά Διπλωματικής Εργασίας

Η διάγνωση του σακχαρώδους διαβήτη σε πρώιμα στάδια είναι καθοριστικής σημασίας για την αντιμετώπιση της ασθένειας. Με δεδομένο ότι η ασθένεια πλήττει ένα μεγάλο μέρος του πληθυσμού και παρουσιάζει αυξητικές τάσεις, εάν δεν γίνει έγκαιρη διάγνωση οι

πιθανότητες για βλάβες στην υγεία αυξάνονται, με αποτέλεσμα πρώτιστα να επηρεάζεται η ποιότητα της ζωής των ασθενών και κατά δεύτερο λόγο να επιβαρύνεται οικονομικά τόσο ο ίδιος ο ασθενής αλλά και το σύστημα υγείας. Είναι χαρακτηριστικό το γεγονός ότι, ο ΠΟΥ προτρέπει τις κυβερνήσεις, μεταξύ των άλλων, να διασφαλίσουν καλύτερες μεθόδους διάγνωσης.

Για την παρούσα διπλωματική εργασία επιλέχθηκε ως σύνολο δεδομένων το σύνολο «Early stage diabetes risk prediction dataset» [9], το οποίο διατίθεται από το αποθετήριο ΜΜ του πανεπιστημίου UCI [10], το οποίο διατέθηκε στο αποθετήριο το 2020 και το οποίο περιέχει δεδομένα που συλλέχθηκαν βάσει ερωτηματολογίου από 520 ασθενείς του Sylhet Diabetes Hospital του Bangladesh και αφορούν τα συμπτώματα της ασθένειας.

Τα προγνωστικά μοντέλα που βασίζονται σε μεθόδους ΜΜ προσφέρουν έναν γρήγορο και φθηνό τρόπο για να διαγνωσθεί σε πρώτη φάση η ασθένεια, σύμφωνα με τα βασικά συμπτώματα. Η σύγκριση των διαθέσιμων αλγόριθμων ΜΜ για την εύρεση του καταλληλότερου αλγόριθμου που θα εφαρμοστεί για τη δημιουργία του προγνωστικού μοντέλου συνεισφέρει στην επίλυση του προβλήματος.

2 Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζεται συνοπτικά η ασθένεια του σακχαρώδους διαβήτη. Στη συνέχεια παρουσιάζονται βασικά στοιχεία για τη Μηχανική Μάθηση, τις κατηγορίες της και τις κυριότερες εφαρμογές της.

2.1 Η Ασθένεια του Σακχαρώδους Διαβήτη

Ο σακχαρώδης διαβήτης είναι μία χρόνια ασθένεια η οποία οφείλεται στην αδυναμία του ανθρώπινου οργανισμού στη διαχείριση μιας ορμόνης, της ινσουλίνης, η οποία είναι ορμόνη που παράγεται από το πάγκρεας και ρυθμίζει το σάκχαρο του αίματος. Η αδυναμία μπορεί να οφείλεται στο ότι το πάγκρεας δεν παράγει αρκετή ινσουλίνη ή δεν μπορεί να διαχειριστεί αποτελεσματικά την ινσουλίνη που παράγει. Σε κάθε περίπτωση το αποτέλεσμα είναι η υπεργλυκαιμία, δηλαδή το αυξημένο σάκχαρο στο αίμα, η οποία με την πάροδο του χρόνου προκαλεί σοβαρές βλάβες στον οργανισμό. Ο διαβήτης είναι μία από της κύριες αιτίες της τύφλωσης, της νεφρικής ανεπάρκειας, των καρδιακών προσβολών, των εγκεφαλικών επεισοδίων και του ακρωτηριασμού των κάτω άκρων [3].

2.1.1 Βασικοί τύποι της ασθένειας

Υπάρχουν αρκετοί τύποι σακχαρώδους διαβήτη, αλλά οι κυριότεροι είναι οι Τύπου 1, Τύπου 2.

Ο διαβήτης τύπου 1 (ΣΔ1), που παλαιότερα ήταν γνωστός ως εξαρτώμενος από ινσουλίνη, νεανικός ή παιδικής ηλικίας, χαρακτηρίζεται από ανεπαρκή παραγωγή ινσουλίνης από το πάγκρεας και απαιτεί καθημερινή χορήγηση ινσουλίνης. Δεν είναι γνωστή ούτε η αιτία του, ούτε τα μέσα για την πρόληψή του. Ο ΣΔ1 πλήττει ένα ποσοστό 7%-12% των ασθενών [2].

Ο διαβήτης τύπου 2 (ΣΔ2), που παλαιότερα ήταν γνωστός ως μη ινσουλινοεξαρτώμενος ή διαβήτης ενηλίκων, οφείλεται στην αδυναμία του παγκρέατος να διαχειριστεί αναποτελεσματικά την ποσότητα ινσουλίνης που παράγει. Η πλειονότητα των ατόμων με διαβήτη έχουν ΣΔ2 σε ποσοστό 87%-91% επί του συνόλου των πασχόντων [4]. Ο ΣΔ2 οφείλεται σε ένα μεγάλο βαθμό στην αλλαγή του τρόπου ζωής και της διατροφής που παρατηρείται παγκόσμια τις τελευταίες δεκαετίες και είναι κατά κύριο λόγο αποτέλεσμα του υπερβολικού σωματικού βάρους και της σωματικής αδράνειας [3]. Μέχρι πρόσφατα, αυτός ο τύπος διαβήτη είχε παρατηρηθεί μόνο σε ενήλικες, αλλά εμφανίζεται όλο και πιο συχνά στα παιδιά [3].

Οι άλλοι τύποι του διαβήτη, όπως για παράδειγμα ο διαβήτης εγκυμοσύνης που εμφανίζεται σε γυναίκες σε κυοφορία, αφορούν ένα ποσοστό 1-3% των νοσούντων [4].

2.1.2 Επιπτώσεις στην Υγεία

Σύμφωνα με τον ΠΟΥ [3], ο διαβήτης μπορεί να προκαλέσει βλάβες στην καρδιά, τα αιμοφόρα αγγεία, τα μάτια, τα νεφρά και τα νεύρα. Έτσι, οι ενήλικες με διαβήτη έχουν δύο έως τρεις φορές αυξημένο κίνδυνο καρδιακών προσβολών και εγκεφαλικών επεισοδίων. Σε συνδυασμό με μειωμένη ροή του αίματος, η βλάβη των νεύρων στα πόδια αυξάνει την πιθανότητα έλκους στα πόδια, της λοίμωξης και πολλές φορές οδηγεί στην ανάγκη ακρωτηριασμού των άκρων. Η διαβητική αμφιβληστροειδοπάθεια είναι μια σημαντική αιτία τύφλωσης και εμφανίζεται ως αποτέλεσμα μακροχρόνιας συσσωρευμένης βλάβης στα μικρά αιμοφόρα αγγεία στον αμφιβληστροειδή. Ο διαβήτης είναι η αιτία του 2,6% της παγκόσμιας τύφλωσης. Επίσης, ο διαβήτης είναι από τις κύριες αιτίες νεφρικής ανεπάρκειας.

Σύμφωνα με τους Κοντοάγγελος et al. [11], ο διαβήτης δεν αποτελεί μόνο σοβαρό ιατρικό πρόβλημα, αλλά συνοδεύεται συχνά από νευροψυχολογικά προβλήματα, όπως είναι το άγχος, η κατάθλιψη, η έλλειψη προσοχής και συγκέντρωσης, διαταραχές ύπνου, κοινωνική αποξένωση, σεξουαλικές διαταραχές κλπ. Η ασθένεια του COVID-19 ανέδειξε ακόμη ένα μεγάλο πρόβλημα για τους ασθενείς με διαβήτη: το ότι είναι πιο ευάλωτοι από τον υπόλοιπο πληθυσμό εάν νοσήσουν από τον COVID-19 [5]. Σύμφωνα με τους Kumar et al [8], ο διαβήτης σε ασθενείς με COVID-19 σχετίζεται με διπλάσια αύξηση της θνησιμότητας, καθώς και σοβαρότερα συμπτώματα του COVID-19, σε σύγκριση με τους μη διαβητικούς.

2.1.3 Συμπτώματα και Διάγνωση της Ασθένειας

Σύμφωνα με τον ΠΟΥ, στα συμπτώματα του διαβήτη είναι παρόμοια για τον ΣΔ1 και ΣΔ2 και περιλαμβάνουν υπερβολική απέκκριση ούρων (πολυουρία), δίψα (πολυδιψία), συνεχή πείνα που οδηγεί σε πολυφαγία, απώλεια βάρους, αλλαγές στην όραση και κόπωση. Τα συμπτώματα μπορεί να εμφανιστούν ξαφνικά. Στον ΣΔ2 τα συμπτώματα είναι λιγότερο έντονα, με αποτέλεσμα- εάν δεν εκτιμηθούν σωστά, η ασθένεια να διαγνωστεί αρκετά χρόνια μετά την έναρξή της, αφού έχουν ήδη προκύψει επιπλοκές [3]. Επίσης, έχουν αναφερθεί λιγότερο συχνά συμπτώματα, όπως, για παράδειγμα η μυϊκή ακαμψία, η καθυστερημένη επούλωση τραυμάτων κλπ., τα οποία παρατίθενται από τους Islam et al. [12].

Η έγκαιρη διάγνωση της ασθένειας μειώνει τον κίνδυνο των επιπλοκών και των σοβαρότερων προβλημάτων στην υγεία, συνεπώς συνιστάται σε αυτούς που εμφανίζουν κάποια από τα κύρια συμπτώματα να προχωρούν σε αιματολογικές εξετάσεις μέτρησης σακχάρου στο αίμα. Η διάγνωση του διαβήτη είναι συνήθως θετική όταν [13]:

- Οι ασθενείς εμφανίζουν συμπτώματα πολυουρίας, πολυδιψίας, απώλειας βάρους και τιμή σακχάρου στο αίμα ≥ 200 mg/dl
- Σάκχαρο αίματος μετά 8 ώρες νηστεία ≥ 126 mg/dl ή
- Σάκχαρο αίματος 2 ώρες μετά λήψη 75γρ γλυκόζης (καμπύλη σακχάρου) ≥ 200 mg/dl

Με δεδομένα τη σοβαρότητα της ασθένειας, το ότι ο ΣΔ2 αφορά το 90% των πασχόντων και προέρχεται κύρια από την αλλαγή του τρόπου ζωής, καθώς και το γεγονός ότι τα συμπτώματα είναι λιγότερο έντονα από αυτά του ΣΔ1, οι οργανισμοί υγείας και ειδικοί από τον ιατρικό κόσμο συνιστούν στις κυβερνήσεις καμπάνιες για την ενημέρωση του πληθυσμού όσον αφορά την αλλαγή του τρόπου ζωής και την αυτό-αξιολόγηση των συμπτωμάτων. Η χρήση του Διαδικτύου από το μεγαλύτερο μέρος του παγκόσμιου πληθυσμού αναμφισβήτητα βοηθά σε αυτές τις καμπάνιες. Η τεράστια πρόοδος στις Τεχνολογίες Πληροφοριών και Επικοινωνιών έχει ως αποτέλεσμα την παροχή των κατάλληλων προγνωστικών ψηφιακών εργαλείων- πλατφορμών στους χρήστες του Διαδικτύου, ώστε να μπορούν να πραγματοποιούν μία πρώτη εκτίμηση του προσωπικού κινδύνου με βάση τα βασικά, αλλά και δευτερεύοντα συμπτώματα, προκειμένου στη συνέχεια να προχωρήσουν σε ιατρικές εξετάσεις. Πίσω από αυτά τα εργαλεία κρύβεται η Εξόρυξη Δεδομένων από τον τεράστιο όγκο δεδομένων που συλλέγεται από τα εκατομμύρια των ασθενών και αποθηκεύεται σε βάσεις δεδομένων. Η κύρια μέθοδος πλέον για την Εξόρυξη Δεδομένων, είναι αναμφισβήτητα η Μηχανική Μάθηση.

Μία πρώτη εκτίμηση του προσωπικού κινδύνου μέσω ψηφιακών πλατφορμών, έστω και εάν δεν είναι εντελώς σωστή, αποτελεί το κίνητρο για την διεξαγωγή εξετάσεων για τη διάγνωση της ασθένειας, ώστε, σε περίπτωση που υπάρχει, να αντιμετωπιστεί κατάλληλα για να αποφευχθούν οι σοβαρότερες επιπλοκές για την υγεία.

2.2 Μηχανική Μάθηση

Στην εποχή των Μεγάλων Δεδομένων, ιδιαίτερη αξία έχει η εξαγωγή χρήσιμης πληροφορίας από όλον αυτό τον όγκο των δεδομένων. Η Μηχανική Μάθηση (ΜΜ) παρέχει

την τεχνική βάση για την εξόρυξη των δεδομένων και χρησιμοποιείται για την εξαγωγή πληροφορίας από ακατέργαστα δεδομένα σε βάσεις δεδομένων, πληροφορίας που μπορεί να εκφραστεί σε κατανοητή μορφή και να χρησιμοποιηθεί για διάφορους σκοπούς. Η Μηχανική Μάθηση είναι υποσύνολο της Τεχνητής Νοημοσύνης (ΤΝ), μιας ευρύτερης έννοιας της επιστήμης των υπολογιστών που αφορά τη δημιουργία έξυπνων μηχανών που μπορούν να προσομοιώσουν τις ανθρώπινες ικανότητες.

Η Τεχνητή Νοημοσύνη (Artificial Intelligence) ως επιστήμη των υπολογιστών γεννήθηκε στην δεκαετία του 1950 όταν πρωτοπόροι στο νεοεμφανιζόμενο πεδίο της επιστήμης των υπολογιστών αναρωτήθηκαν εάν οι υπολογιστές μπορούν να σκεφθούν. Το 1990 οι Rich & Knight δίνουν έναν συνοπτικό ορισμό της ΤΝ : *«Τεχνητή Νοημοσύνη είναι η μελέτη του πώς να κάνουμε τους υπολογιστές ικανούς να κάνουν πράγματα στα οποία προς το παρόν οι άνθρωποι τα καταφέρνουν καλύτερα»* [14].

Για ένα μεγάλο χρονικό διάστημα υπήρχε η πεποίθηση ότι η τεχνητή νοημοσύνη μπορεί να επιτευχθεί με τη συγγραφή προγραμμάτων, όπου περιλαμβάνεται ένα σύνολο ρητών κανόνων για τη διαχείριση της γνώσης. Αυτή η προσέγγιση είναι γνωστή ως συμβολική ΤΝ ή ΤΝ βασισμένη στη γνώση και ήταν το κυρίαρχο παράδειγμα από τη δεκαετία του 1950 έως τα τέλη της δεκαετίας του 1980. Η συμβολική ΤΝ είναι κατάλληλη για την επίλυση σαφώς καθορισμένων λογικών προβλημάτων, όπως για παράδειγμα ενός παιχνιδιού όπως το σκάκι, όπου οι κανόνες είναι συγκεκριμένοι και μπορούν εύκολα να διατυπωθούν [1].

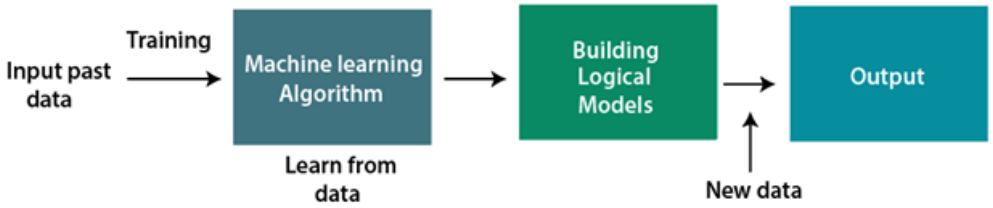
Η πραγματική πρόκληση για την ΤΝ αποδείχτηκε ότι ήταν η επίλυση προβλημάτων που είναι διαισθητικά εύκολες για τον άνθρωπο, όπως για παράδειγμα η αναγνώριση της ομιλίας και της εικόνας. Το ερώτημα εάν ένας υπολογιστής μπορεί να προγραμματιστεί έτσι ώστε, να εκτελεί μια εργασία μαθαίνοντας τους κανόνες μόνο μέσα από την εξέταση των δεδομένων οδήγησε σε μια άλλη προσέγγιση: τη Μηχανική Μάθηση (Machine Learning). Το 1959 ο Arthur Samuel διατύπωσε έναν γενικό ορισμό: *«Η Μηχανική Μάθηση είναι το πεδίο μελέτης το οποίο δίνει στους υπολογιστές τη δυνατότητα να μαθαίνουν χωρίς να έχουν ρητά προγραμματιστεί»*. Το 1997 ο Tom Mitchell έδωσε έναν πιο περιεκτικό ορισμό, ο οποίος ουσιαστικά προσδιορίζει τι είναι μάθηση: *«Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία **E** (Experience) σε σχέση με κάποια εργασία **T** (Task) και κάποιο μέτρο απόδοσης **P** (Performance), εάν βελτιωθεί η απόδοσή του στο **T**, όπως μετράται από το **P**, με εμπειρία **E**»* [1].

Για παράδειγμα, για μια εργασία που αφορά την πρόβλεψη μιας ασθένειας ο στόχος είναι η απάντηση εάν ένας άνθρωπος έχει μια ασθένεια ή όχι. Για μια τέτοια εργασία **T**, οι

ειδικοί του τομέα συλλέγουν περιπτώσεις, δηλαδή παραδείγματα, όπου ένας ασθενής είχε ή δεν είχε την εν λόγω ασθένεια με την καταγραφή μιας σειράς χαρακτηριστικών (features) που σύμφωνα με τη γνώση που διαθέτουν θα βοηθούσαν στην πρόβλεψη. Τέτοια χαρακτηριστικά μπορεί να είναι η ηλικία και το φύλο του ασθενούς, τα αποτελέσματα από μια σειρά διαγνωστικών εξετάσεων όπως η αρτηριακή πίεση, το σάκχαρο του αίματος και ούτω καθεξής. Για κάθε παράδειγμα θέτουν μια ετικέτα (label) όπου δηλώνεται το εάν υπάρχει η ασθένεια ή όχι. Για τα δεδομένα αυτά, μπορεί να γίνει η κατάλληλη αναπαράσταση (representation), έτσι ώστε να είναι σε μορφή κατάλληλη για τον υπολογιστή. Το σύνολο των παραδειγμάτων σε αυτή τη μορφή αποτελούν ένα σύνολο δεδομένων (dataset) ή αλλιώς σύνολο εκπαίδευσης (training set) μέσω του οποίου θα αποκτηθεί η εμπειρία **E**. Ο αλγόριθμος ΜΜ προσδιορίζει πώς θα βγει το συμπέρασμα εάν ο ασθενής έχει την ασθένεια ή όχι με τη γενίκευση από τα δεδομένα. Ο αλγόριθμος ΜΜ με συνεχείς δοκιμές και επαναλήψεις καλείται να βρει μια όσο το δυνατόν απλούστερη μαθηματική συνάρτηση με βάση τα χαρακτηριστικά που του δίνονται από τα παραδείγματα, για να εξάγει το ορθό αποτέλεσμα για τον στόχο. Το απαιτούμενο επίπεδο ορθότητας (accuracy) του αλγόριθμου εξαρτάται από την εκάστοτε εργασία και είναι το μέτρο απόδοσης **P** του αλγόριθμου. Συνήθως, το ενδιαφέρον επικεντρώνεται στο πόσο καλά αποδίδει ένας αλγόριθμος ΜΜ σε δεδομένα που δεν έχει ξαναδεί, γιατί έτσι προσδιορίζεται πόσο αποτελεσματικά θα χρησιμοποιηθεί σε εφαρμογές του πραγματικού κόσμου. Έτσι, η απόδοση του αλγόριθμου αποτιμάται με την χρήση ενός συνόλου δοκιμής (test set), το οποίο διαχωρίζεται από τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση του συστήματος ΜΜ. Η δυνατότητα της καλής απόδοσης ενός αλγόριθμου σε νέα δεδομένα, τα οποία το σύστημα δεν έχει ξαναδεί ονομάζεται γενίκευση (generalization) [1].

Το σύστημα ΜΜ μαθαίνει από ιστορικά δεδομένα, δημιουργεί μοντέλα πρόβλεψης με σκοπό όταν θα λάβει νέα δεδομένα τα οποία δεν έχει ξαναδεί, να προβλέψει την έξοδο για αυτά τα δεδομένα. Η ακρίβεια της προβλεπόμενης εξόδου εξαρτάται από την ποσότητα των δεδομένων, καθώς η τεράστια ποσότητα δεδομένων βοηθά στη δημιουργία ενός καλύτερου μοντέλου που προβλέπει την έξοδο με μεγαλύτερη ακρίβεια. Έτσι, σε ένα περίπλοκο πρόβλημα, όπως αυτό της διάγνωσης μιας ασθένειας, αντί να γραφεί ένα πρόγραμμα για την επίλυσή του, χρειάζεται μόνο η τροφοδοσία με δεδομένα σε γενικούς αλγόριθμους και με τη βοήθεια αυτών των αλγορίθμων, η μηχανή δημιουργεί τη λογική σύμφωνα με τα δεδομένα και προβλέπει την έξοδο [1].

Συνοπτικά, η διαδικασία μάθησης στην ΜΜ φαίνεται στο Σχήμα 2.



Σχήμα 2. Η Διαδικασία Μάθησης στην Μηχανική Μάθηση [15]

Η Βαθιά Μάθηση (ΒΜ) είναι μια κατηγορία αλγορίθμων ΜΜ εμπνευσμένη από τη δομή του ανθρώπινου εγκεφάλου που δίνει τη δυνατότητα στους υπολογιστές να βελτιώνονται με την εμπειρία και τα δεδομένα. Ο πυρήνας της ΒΜ είναι τα Τεχνητά Νευρωνικά Δίκτυα - ΤΝΔ (Artificial Neural Networks - ANN). Ο όρος νευρωνικό δίκτυο ξεκινά από τη νευροβιολογία, αλλά στην ουσία μόνο οι βασικές έννοιες της ΒΜ αναπτύχθηκαν αντλώντας την έμπνευση από τη λειτουργία του ανθρώπινου εγκεφάλου. Η βασική αρχή για την ΒΜ είναι η μάθηση μέσω πολλαπλών επιπέδων σύνθεσης από μη γραμμικούς μετασχηματισμούς δεδομένων εισόδου. Οι βασικές ιδέες θεμελιώδεις ιδέες για την ΒΜ, όπως για παράδειγμα ο αλγόριθμος οπισθοδιάδοσης που διατυπώθηκε το 1986 και η υπολογιστική όραση με τη βοήθεια των Συνελικτικών Νευρωνικών Δικτύων που υλοποιήθηκε το 1989. Η ΒΜ άρχισε να αναπτύσσεται μετά το 2006 και απογειώνεται μετά το 2012 λόγω της προόδου στο υλικό (hardware) των υπολογιστών, της αύξησης των συνόλων των δεδομένων και των αλγοριθμικών προόδων [1].

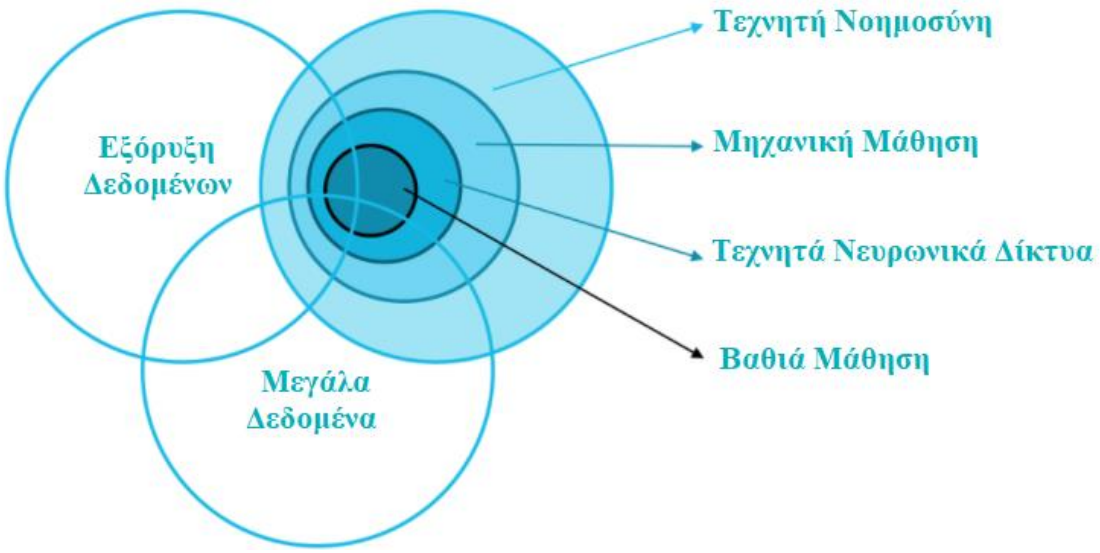
Η ΜΜ βασίζεται σε πειραματικά ευρήματα και οι αλγόριθμοι εξελίσσονται μόνο όταν είναι διαθέσιμα κατάλληλα δεδομένα, υλικό αλλά και λογισμικό. Έτσι, ένας σημαντικός παράγοντας στην ανάπτυξη της ΜΜ είναι η δημιουργία και η συνεχής ανάπτυξη βιβλιοθηκών λογισμικού γλωσσών προγραμματισμού, όπως η scikit-learn[1] και η TensorFlow[2] της γλώσσας προγραμματισμού python[3] και διάφορων άλλων βιβλιοθηκών με βάση άλλες γλώσσες προγραμματισμού (R, Java κλπ.) [1].

Στο Σχήμα 2 δίνεται η συσχέτιση των τομέων της ΤΝ με τα Μεγάλα Δεδομένα και την Εξόρυξη Δεδομένων.

[1] https://scikit-learn.org/stable/

[2] https://www.tensorflow.org/

[3] https://www.python.org/



Σχήμα 3. Μεγάλα Δεδομένα, Εξόρυξη Δεδομένων και Τομείς Τεχνητής Νοημοσύνης

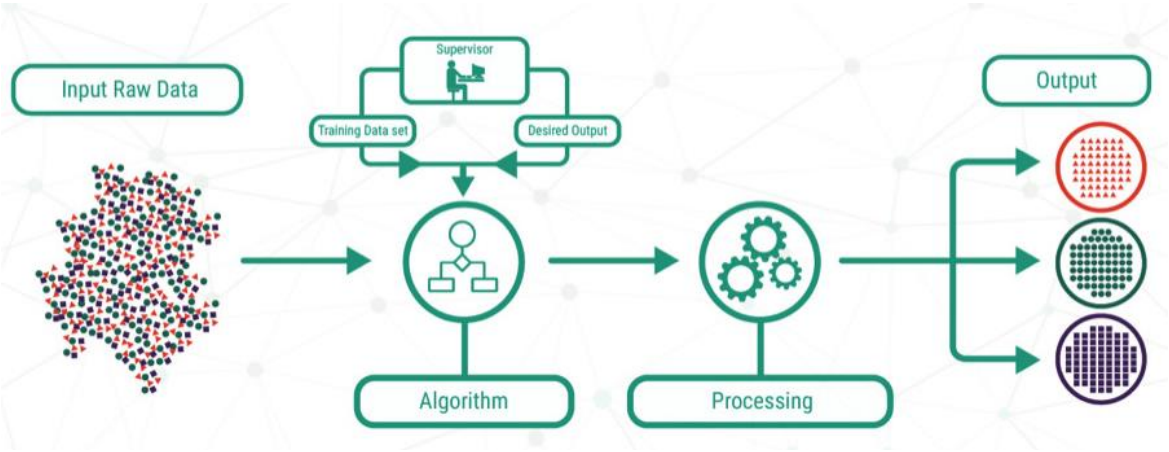
2.2.1 Κατηγορίες Μηχανικής Μάθησης

Η βασική κατηγοριοποίηση των αλγόριθμων ΜΜ γίνεται ανάλογα με το είδος της εμπειρίας που έχουν κατά την εκπαίδευση. Οι αλγόριθμοι αποκτούν την εμπειρία, δηλαδή εκπαιδεύονται, με βάση το σύνολο εκπαίδευσης (training set). Έτσι, κατηγοριοποιούνται ανάλογα με το εάν η εκπαίδευση γίνεται με ανθρώπινη επίβλεψη ή όχι. Σύμφωνα με αυτό το κριτήριο, υπάρχουν τρεις κύριες κατηγορίες αλγορίθμων μάθησης [1]:

- Επιβλεπόμενη ή εποπτευόμενη μάθηση (supervised learning)
- Μη επιβλεπόμενη ή μη εποπτευόμενη μάθηση (unsupervised learning)
- Ενισχυτική μάθηση (reinforcement learning)

Στην επιβλεπόμενη μάθηση παρέχεται σε ένα σύστημα ΜΜ ένα σύνολο ακατέργαστων δεδομένων με βάση το οποίο θα αποκτήσει εμπειρία ο αλγόριθμος. Το σύνολο των δεδομένων περιέχει τα παραδείγματα με τα χαρακτηριστικά τους και κάθε παράδειγμα - στοιχείο του συνόλου συνδέεται με ένα στόχο (target) ή αλλιώς ετικέτα (label). Τέτοιου είδους δεδομένα, ονομάζονται δεδομένα με ετικέτα (labeled). Πχ, στο πρόβλημα της πρόβλεψης μιας ασθένειας, κάθε παράδειγμα έχει τις τιμές των χαρακτηριστικών σε μορφή διανύσματος $\boldsymbol{x}$ και μια συνδεμένη ετικέτα-στόχο y με διακριτή τιμή, η οποία μπορεί να έχει την τιμή 0 για την περίπτωση που δεν υπήρχε ασθένεια και την τιμή 1 εάν υπήρχε. Η επιβλεπόμενη μάθηση βασίζεται στην επίβλεψη και πηγάζει από την θεώρηση ότι ο στόχος

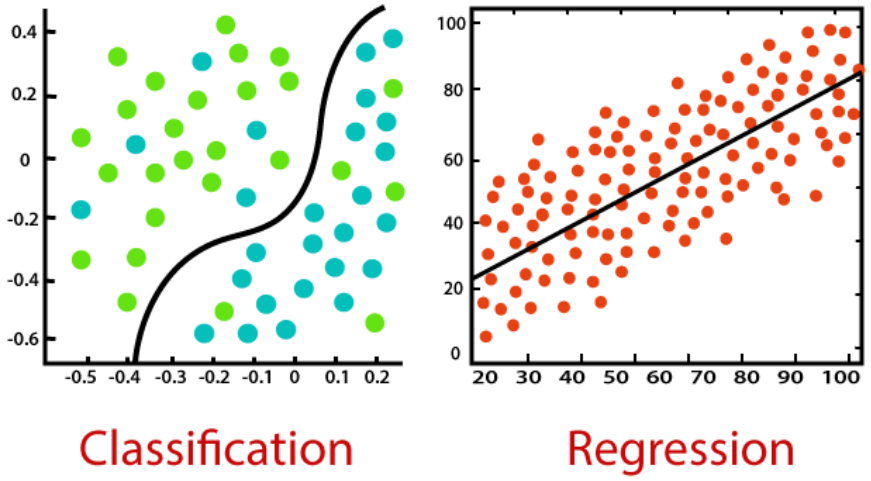
y δίνεται από έναν δάσκαλο που δείχνει στον μαθητή τι ακριβώς περιμένει από αυτόν να κάνει. Στο Σχήμα 4 φαίνεται η λειτουργία ενός συστήματος επιβλεπόμενης ΜΜ.



Σχήμα 4. Επιβλεπόμενη Μάθηση [16]

Οι κύριες εργασίες της επιβλεπόμενης μάθησης, όπως φαίνεται στο Σχήμα , είναι δύο:

- η **ταξινόμηση ή κατηγοριοποίηση (classification)**, όπου από το σύστημα αναμένεται η πρόβλεψη για την ταξινόμηση των δεδομένων σε διάφορες εκ των προτέρων γνωστές κατηγορίες. Σε περίπτωση που οι κατηγορίες ταξινόμησης είναι δύο, τότε έχουμε τη **δυαδική ταξινόμηση (binary classification)**.
- η **παλινδρόμηση (regression),** όπου από το σύστημα αναμένεται η πρόβλεψη μιας τιμής.

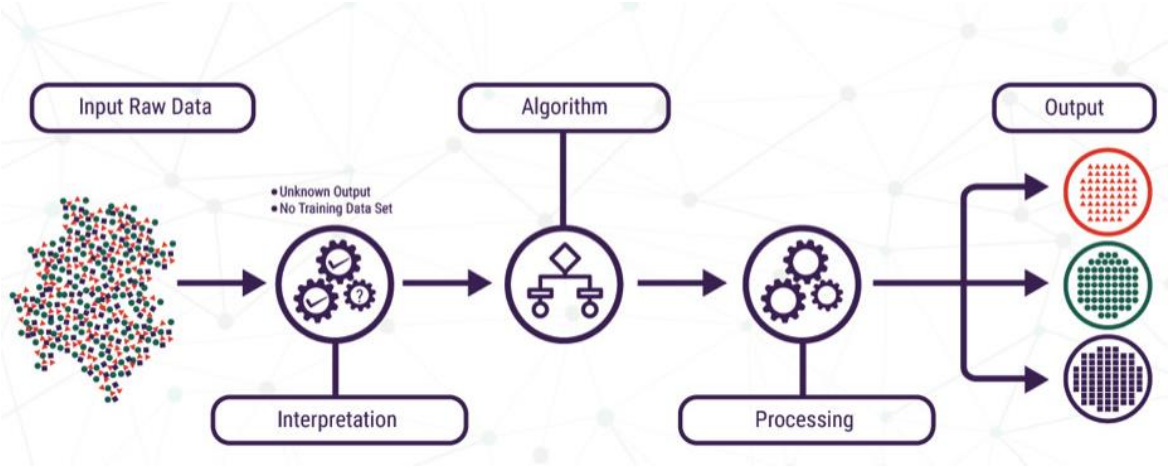


Σχήμα 5. Ταξινόμηση και Παλινδρόμηση [17]

Οι πιο σημαντικοί αλγόριθμοι δυαδικής ταξινόμησης είναι [18]:

- Η λογιστική παλινδρόμηση (logistic regression)
- Οι μηχανές υποστήριξης διανυσμάτων (Support Vector Machines – SVM)
- Τα δέντρα απόφασης (Decision Trees) και τα τυχαία δάση (Random Forests)
- Ο απλοϊκός Bayes (Naïve Bayes)
- Τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks -ANN)
- Οι k-πλησιέστεροι γείτονες (k-Nearest Neighbors -KNN)

Στην **μη-επιβλεπόμενη μάθηση** ο αλγόριθμος εκπαιδεύεται στο σύνολο των δεδομένων που περιέχει τα χαρακτηριστικά, μαθαίνοντας τις χρήσιμες ιδιότητες της δομής του συνόλου δεδομένων. Τα παραδείγματα δεν περιέχουν ετικέτες – στόχους (unlabeled). Η μη επιβλεπόμενη μάθηση αφορά την παρατήρηση διάφορων παραδειγμάτων και προσπαθεί έμμεσα ή άμεσα να μάθει την κατανομή πιθανότητας *p(x)* ή τις ενδιαφέρουσες ιδιότητες αυτής της κατανομής. Στο Σχήμα 6 φαίνεται η λειτουργία ενός συστήματος μη-επιβλεπόμενης ΜΜ.



Σχήμα 6. Μη-επιβλεπόμενη Μάθηση [16]

Οι πιο συνηθισμένες εργασίες μη επιβλεπόμενης μάθησης είναι [18]:

- Η **συσταδοποίηση (clustering)** όπου από το σύστημα αναμένεται ο διαχωρισμός των δεδομένων σε ομοειδείς συστάδες. Σημαντικοί αλγόριθμοι αυτής της κατηγορίας είναι ο k-Means και η ιεραρχική ανάλυση συστάδων (Hierarchical Cluster Analysis – HCA).
- Η **μείωση διαστάσεων (dimensionality reduction)** όπου από το σύστημα αναμένεται η εξαγωγή χρήσιμων ιδιοτήτων των δεδομένων με απλοποίηση των δεδομένων χωρίς να χαθεί όμως χρήσιμη πληροφορία. Σημαντικοί αλγόριθμοι

αυτής της κατηγορίας είναι η ανάλυση κυρίων συνιστωσών (Principal Component Analysis - PCA), ο kernel PCA και η τοπική-γραμμική ενσωμάτωση (Locally-Linear Embedding - LLE).

- Η **εκμάθηση κανόνα συσχέτισης (association rule learning)**, όπου ο στόχος είναι η ανακάλυψη συσχετίσεων μεταξύ των χαρακτηριστικών ενός πολύ μεγάλου όγκου δεδομένων. Σημαντικοί αλγόριθμοι αυτής της κατηγορίας είναι ο αλγόριθμος Apriori και ο αλγόριθμος Eclat.

Η **ενισχυτική μάθηση** είναι μια τελείως διαφορετική κατηγορία αλγορίθμων. Υπάρχει ένα σύστημα εκμάθησης, ο πράκτορας (agent) που μπορεί να παρατηρεί το περιβάλλον, να επιλέγει και να εκτελεί ενέργειες. Εάν επιλέγει την ορθή ενέργεια επιβραβεύεται και στην αντίθετη περίπτωση τιμωρείται. Σε αυτή την περίπτωση ο πράκτορας θα πρέπει να μαθαίνει μόνος του και ο στόχος είναι η συνεχής επιβράβευση. Στο Σχήμα 7 φαίνεται η λειτουργία ενός συστήματος επιβλεπόμενης ΜΜ. Τα τελευταία χρόνια, λόγω της ανάπτυξης των ΤΝΔ και της ΒΜ, η ενισχυτική μάθηση γνωρίζει ιδιαίτερη ανάπτυξη.

Σχήμα 7. Ενισχυτική Μάθηση [16]

Επίσης, στις κατηγορίες της ΜΜ εντάσσεται και η **ημι-επιβλεπόμενη μάθηση,** η οποία είναι ένας συνδυασμός επιβλεπόμενης και μη επιβλεπόμενης μάθησης, όπου το σύστημα καλείται να εκπαιδευτεί σε σύνολο δεδομένων που περιέχει δομημένα παραδείγματα, αλλά και αδόμητα.

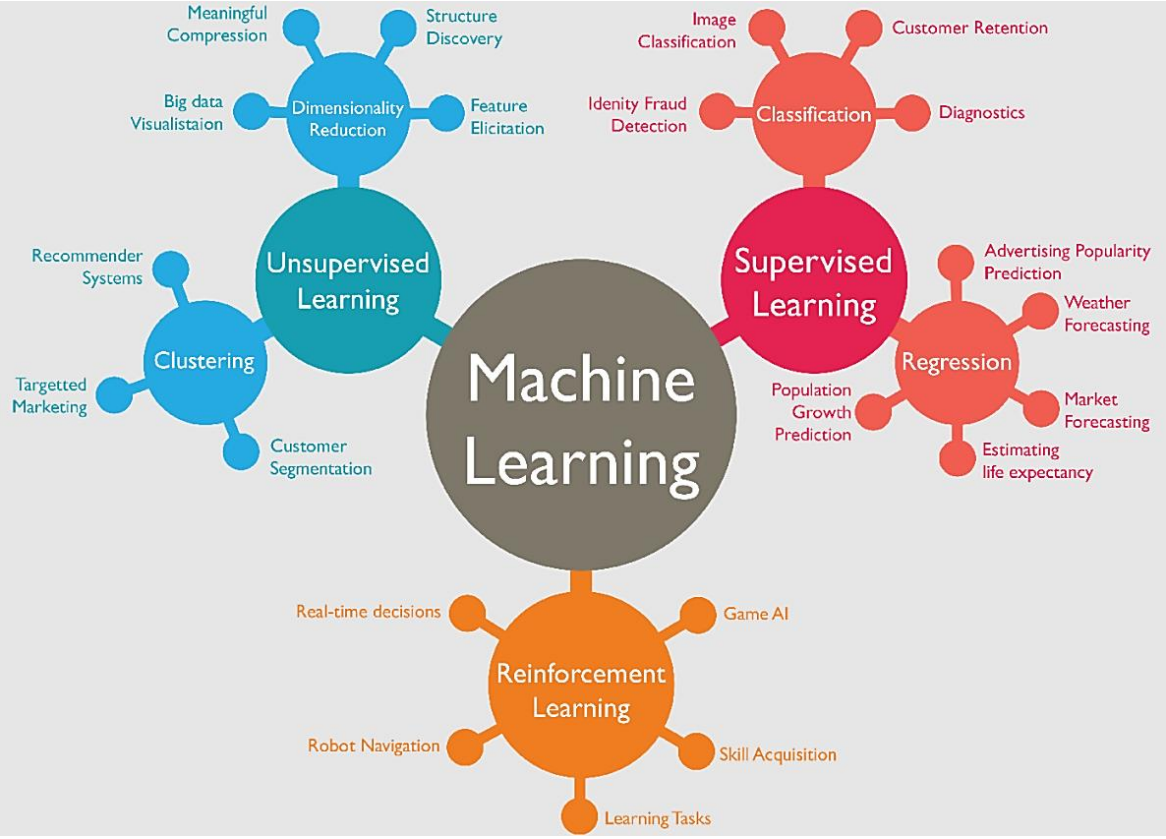
2.2.2 Εφαρμογές Μηχανικής Μάθησης

Στις μέρες μας οι εφαρμογές της ΜΜ καλύπτουν όλους τους τομείς της επιστήμης και της τεχνολογίας του πραγματικού κόσμου όπως είναι, για παράδειγμα, η ιατρική, το εμπόριο, η

ψυχαγωγία, η εκπαίδευση κλπ. Χαρακτηριστικά παραδείγματα εφαρμογών της ΜΜ αποτελούν [1]:

- Η κατηγοριοποίηση εικόνων
- Η αναγνώριση ομιλίας
- Η διάγνωση ασθενειών
- Η πρόγνωση του καιρού
- Τα συστήματα συστάσεων
- Η τμηματοποίηση των πελατών

Στο Σχήμα 8, φαίνονται ενδεικτικές εφαρμογές των αλγορίθμων ΜΜ, ανάλογα με την κατηγορία ΜΜ και το είδος των εργασιών.



Σχήμα 8. Εφαρμογές Μηχανικής Μάθησης [19]

2.2.3 Σημαντικοί Αλγόριθμοι Ταξινόμησης

Η εξέλιξη της ΜΜ συμβαδίζει με την εξέλιξη των ηλεκτρονικών υπολογιστών και ουσιαστικά χρονολογείται από το 1943, όταν οι McCulloch and Pitts ανέπτυξαν τη θεωρία του τεχνητού νευρώνα, εμπνευσμένοι από τη λειτουργία του βιολογικού νευρώνα του

ανθρώπινου εγκεφάλου και αρχίζει να αναπτύσσεται την δεκαετία του 1950, όπου πρωτοπόροι στον τομέα αρχίζουν και αναπτύσσουν τους πρώτους αλγόριθμους ΜΜ, όπως ο Rosenblatt το 1958 που παρουσίασε τον αλγόριθμο του perceptron, που αφορούσε την εκπαίδευση ενός τεχνητού νευρώνα [20] και ο Cox την ίδια χρονιά που παρουσίασε τον αλγόριθμο της λογιστικής παλινδρόμησης [21]. Η μεγάλη ανάπτυξη της ΜΜ ξεκινά από τη δεκαετία του 1980, παράλληλα με την ανάπτυξη σε υλικό και λογισμικό των υπολογιστών. Η πρόταση του αλγόριθμου της οπισθοδιάδοσης από τους Rumelhart et al. το 1986 [22] σηματοδοτεί την αρχή της ανάπτυξης των αλγόριθμων των ΤΝΔ, με κυρίαρχο το πολυεπίπεδο perceptron (Multi-Layer Perceptron -MLP) που αποτελεί την πεμπτουσία της ΒΜ και η οποία κυριαρχεί πλέον έναντι των άλλων μεθόδων στην εξόρυξη δεδομένων από τεράστιες σε όγκο δεδομένων βάσεις [1]. Την ίδια χρονιά ο Quinlan αναπτύσσει τους αλγόριθμους των δέντρων απόφασης (decision trees) [23]. Στη δεκαετία του 1990 οι Cortes και Vapnik επεκτείνουν τη θεωρία του Vapnik για τα διανύσματα υποστήριξης και εισάγουν τον αλγόριθμο των SVM [24]. Στην ίδια δεκαετία προτείνεται και ο αλγόριθμος του Naive Bayes [25], καθώς και ο αλγόριθμος για τα τυχαία δάση (Random Forests -RF) από τον Ho [26].

Η παράθεση των πρωτοποριακών μεθόδων ΜΜ δεν σταματά εδώ, ούτε μπορεί να είναι εξαντλητική στα πλαίσια της παρούσας εργασίας. Το πρόβλημα της διάγνωσης του σακχαρώδους διαβήτη, είναι ένα πρόβλημα επιβλεπόμενης μάθησης για δυαδική ταξινόμηση δεδομένων. Έτσι, παρουσιάζονται συνοπτικά οι σημαντικότεροι αλγόριθμοι που βρίσκουν ιδιαίτερη εφαρμογή σε προβλήματα δυαδικής ταξινόμησης.

Γενικά, στη δυαδική ταξινόμηση έχουμε ένα σύνολο δεδομένων D με K παραδείγματα με ετικέτα $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^{K}$ όπου κάθε στοιχείο του $\boldsymbol{x}_k$ είναι ένα διάνυσμα n χαρακτηριστικών του οποίου κάθε διάσταση $i= 1,2, \dots n$ περιέχει μια τιμή που περιγράφει το παράδειγμα και το y_k είναι η ετικέτα που παίρνει μόνο δύο τιμές από το σύνολο {0,1}, που αντιστοιχούν στις τάξεις-κλάσεις για τις οποίες είναι επιθυμητό από τον αλγόριθμο να μάθει πως θα τις διαχωρίζει αυτόματα. Δηλαδή, είναι επιθυμητή η δημιουργία ενός μοντέλου-ταξινομητή που θα μάθει να προβλέπει το y από το διάνυσμα $\boldsymbol{x}$. Ένας τρόπος για να λυθεί το πρόβλημα είναι η προσέγγιση με μια συνάρτηση $y = f(\boldsymbol{x})$ με μια άγνωστη συνάρτηση f και ο στόχος είναι η εκμάθηση της συνάρτησης για το δοσμένο σύνολο και στι συνέχεια να γίνουν προβλέψεις χρησιμοποιώντας την $\hat{y} = \hat{f}(\boldsymbol{x})$ ως εκτίμηση. Η πρόβλεψη σε κάθε περίπτωση βασίζεται στη θεωρία των πιθανοτήτων και η βελτιστοποίηση της λύσης στην μεγιστοποίηση της μέγιστης αληθοφάνειας [1].

Μια κατηγορία μοντέλων είναι τα διακριτά μοντέλα, όπου από το μοντέλο εκτιμάται η πιθανότητα της περίπτωσης ενός παραδείγματος να ανήκει στην κλάση 0 ή 1 δοθέντος του *x*, δηλαδή η πιθανότητα $p(y|x)$. Εάν η εκτίμηση της πιθανότητας είναι μεγαλύτερη του 50%, τότε το μοντέλο προβλέπει ότι το στιγμιότυπο ανήκει στην κλάση 1 (ή αλλιώς στη θετική κλάση), αλλιώς προβλέπει ότι ανήκει στην κλάση 0 (αρνητική κλάση) [27].

Η κατανομή της υπό συνθήκη πιθανότητας σε πιθανές ετικέτες για το διάνυσμα εισόδου $\boldsymbol{x}$ και το σύνολο εκπαίδευσης *D* για τη δυαδική ταξινόμηση δίνεται από τη σχέση $p(y = 1 \mid x, D)$ καθώς $p(\, y = 1 \mid x, D\,) + p(\, y = 0 \mid x, D\,) = 0$.

Δοθείσας μιας πιθανοτικής εξόδου, υπολογίζεται η αρχική πρόβλεψη σε σχέση με την πραγματική τιμή χρησιμοποιώντας τη σχέση

$$\hat{y} = \hat{f}\,(\boldsymbol{x}) = \arg\max_{c}\ p(y = c \mid x, D)$$

που είναι η μέγιστη a posteriori[4] (Maximum A Posteriori - MAP) εκτίμηση.

Τα διακριτά μοντέλα καλούνται να μάθουν ένα όριο απόφασης (decision boundary) που θα διαχωρίζει τις τάξεις των δεδομένων.

Οι δύο τάξεις των δεδομένων διαχωρίζονται ιδανικά με γραμμικό όριο απόφασης, ένα υπερεπίπεδο, που η εξίσωσή του δίνεται από τη σχέση:

$$w_0x_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n = 0$$

η οποία σε μορφή πινάκων γράφεται ως $\boldsymbol{w^T x} = 0$

Ο όρος w_0 αντιπροσωπεύει την μετατόπιση ως προς την πηγή (0,0) του καρτεσιανού συστήματος συντεταγμένων και αναφέρεται συνήθως ως πόλωση (bias) και κατά σύμβαση τίθεται το $x_0 = 1$. Το στοιχείο w_i είναι ο συντελεστής που πολλαπλασιάζεται με το χαρακτηριστικό x_i. Οι συντελεστές αυτοί ονομάζονται επίσης βάρη γιατί ορίζουν πως το κάθε χαρακτηριστικό επηρεάζει την πρόβλεψη για το y και είναι οι παράμετροι που πρέπει να μάθει το σύστημα.

Στην πράξη κατά κανόνα το όριο απόφασης είναι μη γραμμικό. Το γραμμικό μοντέλο μπορεί να μετασχηματιστεί σε μη γραμμικό, με τη βοήθεια των μαθηματικών και συγκεκριμένα των μη γραμμικών συναρτήσεων που θα εφαρμόζονται σε κάθε είσοδο $\boldsymbol{x}$. Τέτοιες συναρτήσεις είναι, για παράδειγμα, η σιγμοειδής συνάρτηση και η συνάρτηση υπερβολικής εφαπτομένης. Η μη γραμμική συνάρτηση ονομάζεται μη γραμμικός

[4] Με τον όρο a prosteriori αναφέρεται η γνώση που προέρχεται από την εμπειρία, ενώ με τον όρο a priori χαρακτηρίζεται η έμφυτη γνώση- η εκ των προτέρων γνώση

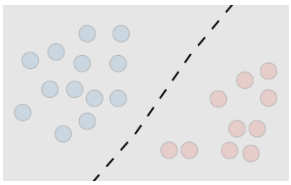
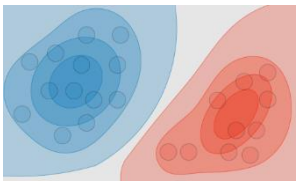
μετασχηματισμός χαρακτηριστικού, επειδή μετασχηματίζει τα χαρακτηριστικά της εισόδου $\boldsymbol{x}$. Ετσι, το όριο απόφασης, είναι $w_0 + w_1 f_1(\boldsymbol{x}) + w_2 f_2(\boldsymbol{x}) + \cdots + w_B f_B(\boldsymbol{x}) = 0$.

Μία άλλη κατηγορία μοντέλων είναι τα παραγωγικά (generative) μοντέλα. Ένα παραγωγικό μοντέλο πρώτα προσπαθεί να μάθει πως παράγονται τα δεδομένα με εκτίμηση της πιθανότητας $p(x|y)$, η ο οποία χρησιμοποιείται για την εκτίμηση της πιθανότητας $p(y|x)$ με τη χρήση του κανόνα του Bayes [27].

Στον Πίνακα 1 δίνεται μια σύγκριση των διακριτών και των παραγωγικών μοντέλων.

Άλλες κατηγορίες μοντέλων είναι τα δέντρα απόφασης και τα συλλογικά (ensemble) μοντέλα. Στα συλλογικά μοντέλα ανήκουν τα τυχαία δάση (random forests). Πρόκειται για μια τεχνική που βασίζεται σε δέντρα και χρησιμοποιεί μεγάλο αριθμό δέντρων αποφάσεων που δημιουργούνται από τυχαία επιλεγμένα σύνολα χαρακτηριστικών.

Πίνακας 1. Σύγκριση Διακριτικού και Παραγωγικού Μοντέλου [27]

	Διακριτικό Μοντέλο	**Παραγωγικό Μοντέλο**
Στόχος	Απευθείας εκτίμηση $P(y\|x)$	Εκτίμηση $P(x\|y)$ για την εξαγωγή του συμπεράσματος $P(y\|x)$
Στόχος μάθησης	Όριο απόφασης	Κατανομές πιθανοτήτων των δεδομένων
Απεικόνιση		
Αλγόριθμοι	LR, SVMs, MLPs	NB

Ανεξάρτητα από την κατηγορία του μοντέλου που θα επιλεγεί, εφαρμόζονται κάποιες βασικές έννοιες για την εκτίμηση του λάθους στην πρόβλεψη.

Η συνάρτηση λάθους (error function) L: $(\hat{y}, y) \in R \times Y \longmapsto L(\hat{y}, y) \in R$ παίρνει ως είσοδο την τιμή πρόβλεψης $\hat{y}$ και την αντίστοιχη πραγματική τιμή y και εξάγει το πόσο

διαφορετικές είναι. Ανάλογα με τον αλγόριθμο, ορίζεται και η αντίστοιχη συνάρτηση κόστους.

Η συνάρτηση κόστους (cost function) J χρησιμοποιείται για να αξιολογηθεί η απόδοση ενός μοντέλου και ορίζεται σε σχέση με την συνάρτηση λάθους ως

$$J(w) = \sum_{i=1}^{k} L((\hat{y}(i)), y(i))$$

Ο σκοπός του συστήματος είναι να ελαχιστοποιηθεί η συνάρτηση κόστους μετά από διαδοχικές δοκιμές και να αποτιμηθεί το σύστημα με βάση το μέτρο απόδοσης που έχει προσδιοριστεί και που συνήθως είναι η ορθότητα (accuracy), δηλαδή το ποσοστό των ορθών προβλέψεων επί του συνόλου των παραδειγμάτων.

2.2.3.1 Λογιστική Παλινδρόμηση (Logistic Regression-LG)

Η λογιστική παλινδρόμηση είναι ένα γραμμικό διακριτικό μοντέλο που, ανεξάρτητα από την ονομασία του χρησιμοποιείται για ταξινόμηση και όχι για παλινδρόμηση. Είναι γνωστή και ως logit regression, ταξινόμηση μέγιστης εντροπίας (maximum-entropy) ή ως log-linear ταξινομητής. Στο μοντέλο αυτό οι πιθανότητες που περιγράφουν τις πιθανές εξόδους μιας απλής δοκιμής μοντελοποιούνται χρησιμοποιώντας τη λογιστική συνάρτηση.

Η λογιστική συνάρτηση είναι η σιγμοειδής συνάρτηση, η οποία με παράμετρο μια λογιστική πραγματική τιμή z εξάγει ως αποτέλεσμα μία πραγματική τιμή που ανήκει στο διάστημα [0,1] και δίνεται από τον τύπο:

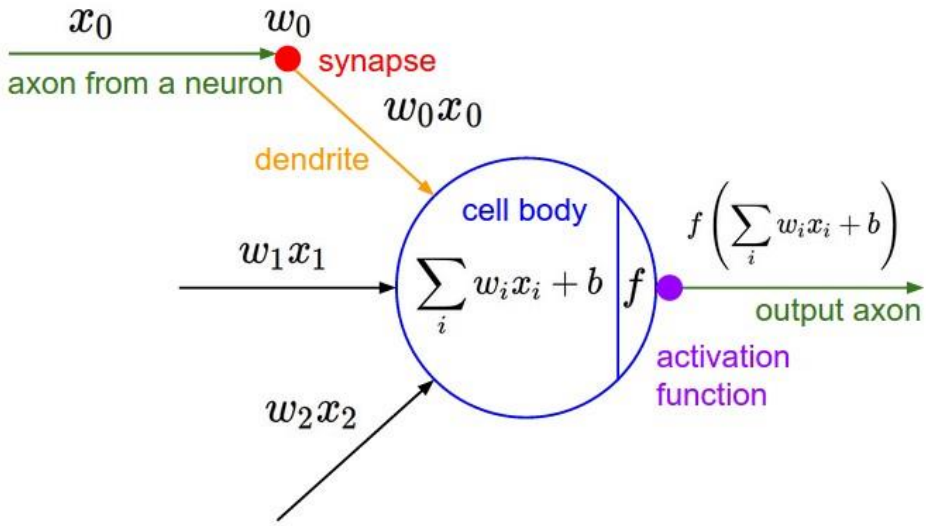
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Στο μοντέλο της λογιστικής παλινδρόμησης, υπολογίζεται το σταθμισμένο άθροισμα
$z = w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$ των χαρακτηριστικών της εισόδου, και στο z εφαρμόζεται η σιγμοειδής συνάρτηση, η οποία επιστρέφει τιμές στο εύρος [0,1], οπότε η πιθανότητα $p = \sigma(z)$, όπου z το σταθμισμένο άθροισμα κάθε εισόδου [1]. Συνεπώς, για ένα στιγμιότυπο του συνόλου εκπαίδευσης x η πρόβλεψη της εξόδου $\hat{y}$ είναι:

$$\hat{y} = \begin{cases} 0 \text{ εάν } \sigma(z) < 0.5 \\ 1 \text{ εάν } \sigma(z) \geq 0.5 \end{cases}$$

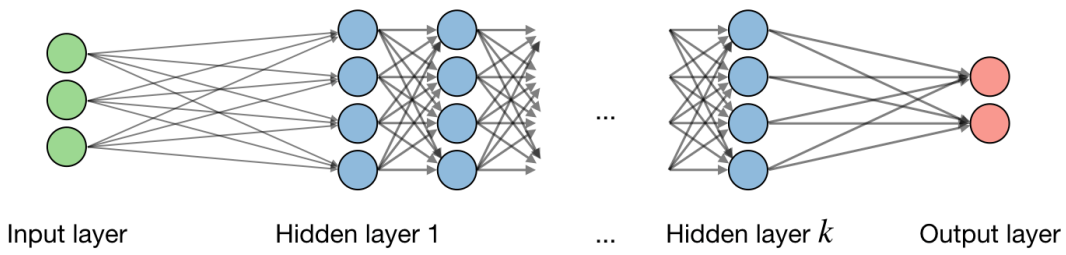
2.2.3.2 Πολυεπίπεδο Perceptron (Multi-Layer Perceptron – MLP)

Τα Τεχνητά Νευρωνικά Δίκτυα – ΤΝΔ (Artificial neural networks -ANNs), είναι συστήματα εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα τα οποία συνιστούν τον ανθρώπινο εγκέφαλο. Το μοντέλο του τεχνητού νευρώνα δίνεται στο Σχήμα 9.



Σχήμα 9. Το Μοντέλο του Τεχνητού Νευρώνα [28]

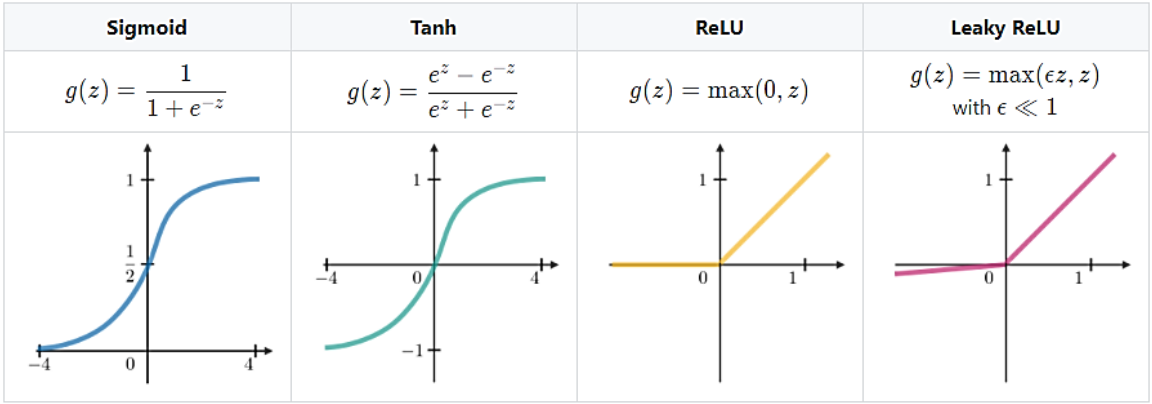
Ένα ΤΝΔ βασίζεται στη συλλογή συνδεμένων τεχνητών νευρώνων που ονομάζονται μονάδες ή κόμβοι. Κάθε σύνδεση, όπως οι συνάψεις στον βιολογικό εγκέφαλο, μπορεί να μεταφέρει το σήμα σε άλλους νευρώνες. Ο τεχνητός νευρώνας που λαμβάνει το σήμα το μεταφέρει σε άλλους νευρώνες. Το σήμα σε μια σύνδεση είναι ένας πραγματικός αριθμός και η έξοδος κάθε νευρώνα υπολογίζεται από κάποια μη-γραμμική συνάρτηση από το άθροισμα των εισόδων του. Οι συνδέσεις συνήθως ονομάζονται ακμές. Οι κόμβοι και οι ακμές τυπικά έχουν κάποια βάρη που προσαρμόζονται όσο προχωρά η μάθηση. Οι νευρώνες διατάσσονται σε επίπεδα, όπως φαίνεται στο Σχήμα 10. Εάν όλοι οι κόμβοι κάθε επιπέδου συνδέονται με όλους τους κόμβους του επόμενου επίπεδου, τότε τα ΤΝΔ ονομάζονται πλήρως συνδεμένα (fully connected). Τα διαφορετικά επίπεδα μπορούν να εκτελούν διαφορετικές μετατροπές στις εισόδους τους. Τα σήματα εισέρχονται και ταξιδεύουν από το πρώτο επίπεδο, το οποίο ονομάζεται επίπεδο εισόδου (input layer) μέχρι το τελευταίο επίπεδο, το οποίο ονομάζεται επίπεδο εξόδου (output layer) και περνούν από τα ενδιάμεσα επίπεδα, τα οποία ονομάζονται κρυφά επίπεδα (hidden layers). Εάν η κατεύθυνση των σημάτων είναι πάντα προς τα εμπρός τότε τα ΤΝΔ ονομάζονται εμπροσθο-τροφοδοτούμενα (feedforward) [29].



Σχήμα 10. Εμπροσθο-τροφοδοτούμενο Τεχνητό Νευρωνικό Δίκτυο [30]

Το πολυεπίπεδο perceptron (Multilayer Perceptron -MLP) είναι μία κατηγορία εμπροσθο-τροφοδοτούμενων ΤΝΔ που η βασική τους μονάδα είναι το perceptron, ένας αλγόριθμος δυαδικής ταξινόμησης που λειτουργεί όπως η λογιστική παλινδρόμηση (Σχήμα 10) και αποτελείται από τουλάχιστον ένα κρυφό επίπεδο.

Το MLP μαθαίνει τη συνάρτηση $f(\cdot): R^n \rightarrow R^o$, εκπαιδεύοντας ένα σύνολο δεδομένων, όπου n είναι ο αριθμός διαστάσεων στο επίπεδο εισόδου και o είναι ο αριθμός των διαστάσεων για την έξοδο, που στην περίπτωση της δυαδικής ταξινόμησης είναι o=2. Με δεδομένο ένα σύνολο χαρακτηριστικών και έναν στόχο, προσεγγίζει τη συνάρτηση. Το επίπεδο εισόδου αποτελείται από $\{x_i | x_1, x_2, \ldots, x_n\}$ νευρώνες, που αναπαριστά τα χαρακτηριστικά της εισόδου. Κάθε νευρώνας στο κρυφό επίπεδο μετατρέπει τις τιμές από το προηγούμενο επίπεδο με ένα γραμμικό άθροισμα $w_1x_1 + w_2x_2 + \ldots + w_nx_n$ που ακολουθείται από μια μη-γραμμική συνάρτηση ενεργοποίησης $g(\cdot): R \rightarrow R$, η οποία μπορεί να είναι η σιγμοειδής, η συνάρτηση γραμμικής εφαπτομένης, η συνάρτηση ράμπας (Rectified Linear Unit -ReLU) και η Leaky ReLU που φαίνονται στο Σχήμα 11. Το επίπεδο εξόδου λαμβάνει τις τιμές από το τελευταίο κρυφό επίπεδο και τις μετατρέπει σε τιμές εξόδου. Το μεγάλο πλεονέκτημα των MLPs είναι ότι μπορούν να μάθουν οποιαδήποτε μη γραμμική συνάρτηση [29].



Σχήμα 11. Συναρτήσεις Ενεργοποίησης MLP [30]

Οι βασικές έννοιες για το MLP σύμφωνα με την [30] περιγράφονται αμέσως παρακάτω. Η συνάρτηση κόστους που εφαρμόζεται στην δυαδική ταξινόμηση είναι η συνάρτηση της εγκάρσιας εντροπίας (cross entropy) που δίνεται από τον τύπο:

$$L(z, y) = -[y\log(z) + (1 - y)\log(1 - z)], \text{όπου } z = \hat{y}.$$

Ο ρυθμός μάθησης (learning rate), που συμβολίζεται συνήθως με α, είναι ο ρυθμός με τον οποίο ενημερώνονται τα βάρη. Η ενημέρωση των βαρών γίνεται με τη μέθοδο της οπισθο-διάδοσης (backpropagation) κατά την οποία λαμβάνονται υπόψη η πραγματική έξοδος σε σχέση με την επιθυμητή έξοδο, Η παράγωγος σε σχέση με το βάρος w υπολογίζεται με την χρήση του αλυσιδωτού κανόνα για την παραγώγιση και έχει την ακόλουθη μορφή:

$$\frac{\partial L(z, y)}{\partial w} = \frac{\partial L(z, y)}{\partial \alpha} \cdot \frac{\partial \alpha}{\partial z} \cdot \frac{\partial z}{\partial \mathrm{w}}$$

και τα βάρη ενημερώνονται ως εξής:

$$w \leftarrow w - a\frac{\partial L(z, y)}{\partial w}$$

Κάθε επανάληψη για την ενημέρωση των βαρών όλου του συνόλου των δεδομένων ονομάζεται εποχή (epoch).

Η απλοποιημένη διαδικασία για την ενημέρωση των βαρών έχει ως εξής:

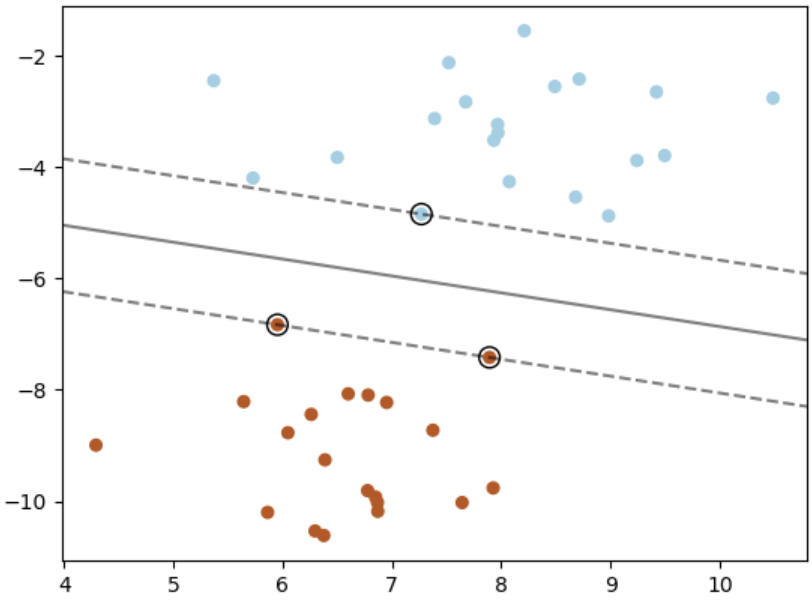
- Βήμα 1: Λαμβάνεται μια δέσμη δεδομένων από το σύνολο των δεδομένων
- Βήμα 2: Εφαρμόζεται η προς τα εμπρός διάδοση των βαρών για να υπολογιστεί η σχετική απώλεια.
- Βήμα 3: Υπολογίζεται με τη μέθοδο της οπισθοδιάδοσης η παράγωγος της απώλειας ως προς τα βάρη
- Βήμα 4: Ενημέρωση των βαρών

Η διαδικασία επαναλαμβάνεται μέχρι να ελαχιστοποιηθεί η συνάρτηση κόστους, σύμφωνα με τα κριτήρια που έχουν τεθεί.

2.2.3.3 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVMs)

Η μηχανή υποστήριξης διανυσμάτων (Support Vector Machine – SVM) κατασκευάζει ένα υπερ-επίπεδο ή ένα σύνολο υπερ-επιπέδων σε έναν υψηλό ή άπειρο διαστάσεων χώρο, ο οποίος μπορεί να χρησιμοποιηθεί για ταξινόμηση, παλινδρόμηση ή άλλες εργασίες. Διαισθητικά, ένας καλός διαχωρισμός επιτυγχάνεται από το υπερ-επίπεδο που έχει τη μεγαλύτερη απόσταση από τα πλησιέστερα σημεία δεδομένων εκπαίδευσης οποιασδήποτε

κατηγορίας (το λεγόμενο λειτουργικό περιθώριο), καθώς γενικά όσο μεγαλύτερο είναι το περιθώριο τόσο μικρότερο είναι το σφάλμα γενίκευσης του ταξινομητή. Στο Σχήμα 12 φαίνεται η συνάρτηση απόφασης για ένα γραμμικά διαχωρίσιμο πρόβλημα, με τρία δείγματα στα όρια περιθωρίου, που ονομάζονται "διανύσματα υποστήριξης". Γενικά, όταν το πρόβλημα δεν διαχωρίζεται γραμμικά, τα διανύσματα υποστήριξης είναι τα δείγματα εντός των ορίων του περιθωρίου [29].



Σχήμα 12. Τα Διανύσματα Υποστήριξης [29]

Με δοσμένα τα διανύσματα $x_i \in \mathbb{R}^p$, i=1,…, n ενός συνόλου εκπαίδευσης σε ένα πρόβλημα δυαδικής ταξινόμησης και ένα διάνυσμα $y \in \{1, -1\}^n$, ο στόχος είναι να βρεθούν τα $w \in \mathbb{R}^p$ και $b \in \mathbb{R}$ έτσι ώστε η πρόβλεψη που δίνεται από το $\text{sign}(w^T\phi(x) + b)$, όπου in το πρόσημο, να είναι σωστή για τα περισσότερα δείγματα. Ο ταξινομητής SVM για δυαδική ταξινόμηση λύνει το πρόβλημα

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

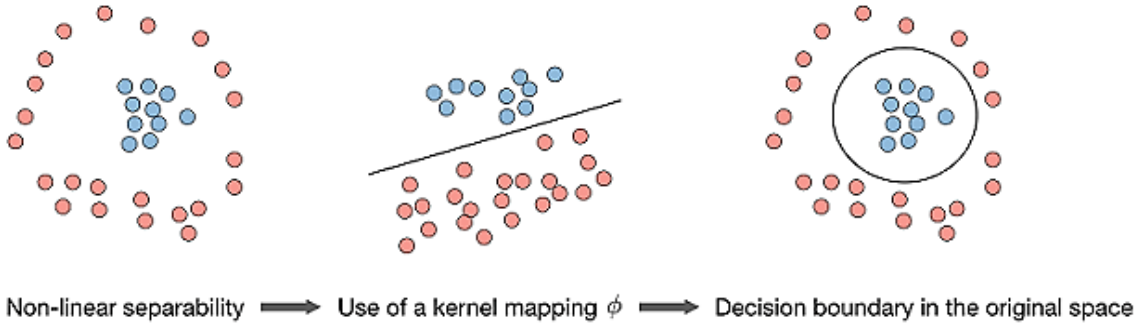
που υπόκειται στο $y_i(w^T\phi(x_i) + b) \geq 1 - \zeta_i,$

$$\zeta_i \geq 0, i = 1, \ldots, n$$

Το $w^T w$ είναι η νόρμα Manhattan, η ζ είναι η συνάρτηση κόστους και C είναι η ποινή. Διαισθητικά, προσπαθούμε να μεγιστοποιήσουμε το περιθώριο, ελαχιστοποιώντας το $||w||^2 = w^T w$, ενώ υφίσταται ποινή όταν ένα δείγμα δεν έχει ταξινομηθεί σωστά ή εντός του ορίου του περιθωρίου. Στην ιδανική περίπτωση, η τιμή $y_i(w^T\phi(x_i) + b)$ είναι

επιθυμητό να είναι ≥ 1 για όλα τα παραδείγματα, πράγμα που σημαίνει μια τέλεια πρόβλεψη. Αλλά τα προβλήματα συνήθως δεν είναι πάντοτε απόλυτα διαχωρίσιμα με υπερεπίπεδα, οπότε επιτρέπουμε σε μερικά δείγματα να βρίσκονται σε απόσταση ζ_i από το σωστό όριο περιθωρίου. Ο όρος ποινής C ελέγχει την ισχύ αυτής της ποινής, και ως αποτέλεσμα, ενεργεί ως αντίστροφη παράμετρος ομαλοποίησης (regularization) [29].

Εάν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, τότε στα διανύσματα των χαρακτηριστικών εφαρμόζεται η συνάρτηση του πυρήνα (kernel). Η συνάρτηση του πυρήνα είναι μία συνάρτηση που χαρτογραφεί τα διανύσματα των χαρακτηριστικών σε μια μεγαλύτερη διάσταση, προκειμένου να γίνουν γραμμικά διαχωρίσιμα, όπως φαίνεται στο Σχήμα 13 και εκφράζεται από τη σχέση $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ Ουσιαστικά, τα χαρακτηριστικά αντικαθίστανται από τον πυρήνα. Η συνάρτηση του πυρήνα μπορεί να είναι γραμμική, πολυωνυμική ή ακτινικής βάσης (radial basis function – RBF) [29].



Σχήμα 13. Εφαρμογή του Πυρήνα σε μη Γραμμικά Διαχωρίσιμα Χαρακτηριστικά [27]

2.2.3.4 Απλοϊκός Bayes (Naïve Bayes - NB)

Ο Naive Bayes (NB) είναι αλγόριθμος επιβλεπόμενης ΜΜ και βασίζεται στην εφαρμογή του θεωρήματος του Bayes με την «αφελή» υπόθεση της υπό όρους ανεξαρτησίας μεταξύ του διανύσματος των χαρακτηριστικών $\boldsymbol{x}$ και της αντίστοιχης μεταβλητής κλάσης y.

Το θεώρημα του Bayes δηλώνει την ακόλουθη σχέση, δεδομένης της μεταβλητής κλάσης ταξινόμησης y και το διάνυσμα με τις τιμές των n χαρακτηριστικών x_1 έως x_n:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n \mid y)}{P(x_1, \dots, x_n)}$$

Χρησιμοποιώντας την απλοϊκή υπόθεση της υπό όρους ανεξαρτησίας η οποία είναι:

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y),$$

για όλα τα i η παραπάνω σχέση απλοποιείται σε:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)}$$

Καθώς η $P(x_1, \dots, x_n)$ είναι σταθερή δοθείσης της εισόδου $\boldsymbol{x}$, εφαρμόζεται ο ακόλουθος κανόνας ταξινόμησης για την πρόβλεψη $\hat{y}$:

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

Στη συνέχεια, χρησιμοποιείται η μέγιστη εκ των υστέρων (Maximum A Posteriori-MAP) εκτίμηση για την εκτίμηση της $P(y)$ που είναι η σχετική συχνότητα της κλάσης y στο σύνολο εκπαίδευσης και την εκτίμηση της $P(x_i \mid y)$. Οι διαφορετικοί NB ταξινομητές διαφέρουν κυρίως με βάση την υπόθεση σε σχέση με την κατανομή της $P(x_i \mid y)$, δηλαδή εάν είναι Gaussian ή πολυωνυμική [29].

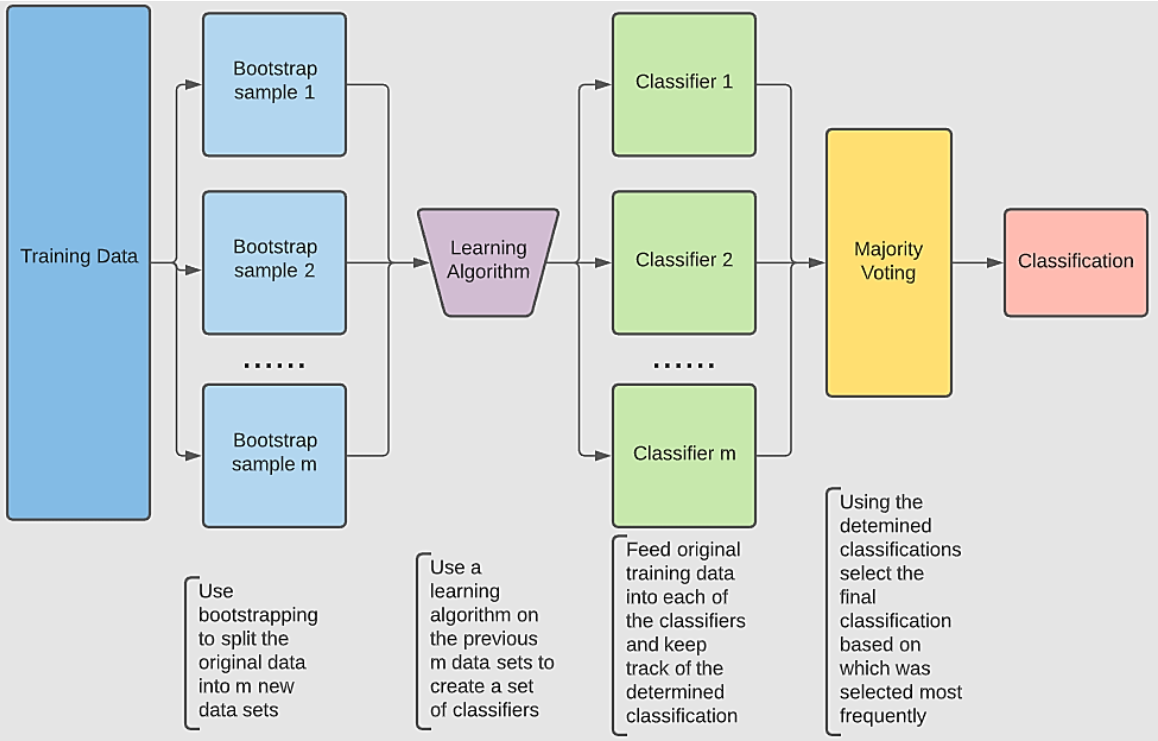
Παρά τις προφανώς υπερβολικά απλοποιημένες παραδοχές τους, οι NB ταξινομητές λειτουργούν αρκετά καλά σε πολλές πραγματικές καταστάσεις, όπως στην ταξινόμηση εγγράφων και το φιλτράρισμα μηνυμάτων spam του ηλεκτρονικού ταχυδρομείου. Απαιτούν μια μικρή ποσότητα δεδομένων εκπαίδευσης για την εκτίμηση των απαραίτητων παραμέτρων.

2.2.3.5 Τυχαία Δάση (Random Forests – RF)

Τα τυχαία δάση είναι μία μέθοδος συλλογικής (ensemble) μάθησης που βασίζεται στον αλγόριθμο των δέντρων αποφάσεων. Η τεχνική για τη δημιουργία αυτού του συλλογικού μοντέλου είναι το bagging, όρος που χρησιμοποιείται ως συντόμευση για την αθροιστική εκκίνηση (bootstrap aggregating). Το bagging εφαρμόζεται για την βελτίωση της σταθερότητας και της ακρίβειας αλγόριθμων μηχανικής μάθησης. Η τεχνική του bagging φαίνεται στο Σχήμα 14 και έχει ως εξής [31]:

Δοθέντος ενός συνόλου δεδομένων εκπαίδευσης *D* με *K* παραδείγματα, το bagging δημιουργεί *m* νέα σύνολα εκπαίδευσης D_i με δειγματοληψία από το *D* με ομοιόμορφη κατανομή και με αντικατάσταση. Κατά τη δειγματοληψία κάποια παραδείγματα μπορεί να

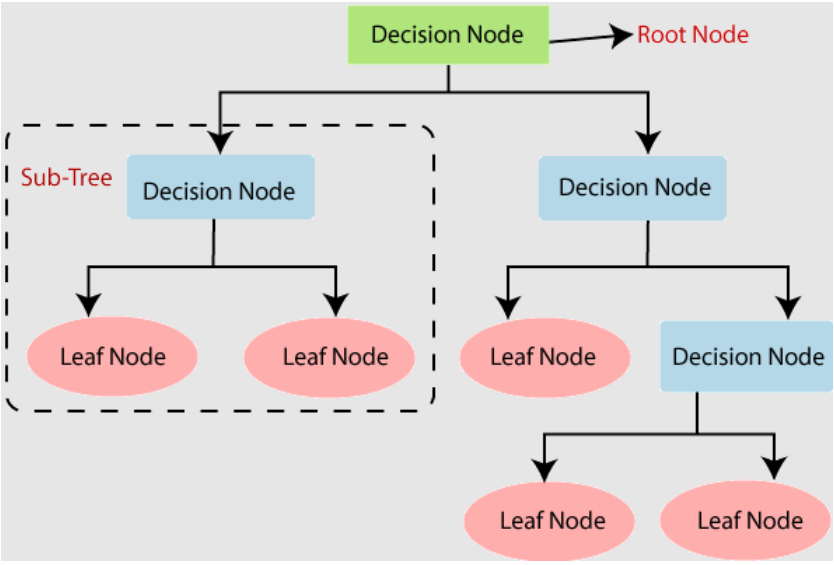
επαναλαμβάνονται σε κάθε D_i . Κάθε ένα από αυτά τα δείγματα είναι ένα δείγμα εκκίνησης (bootstrap sample). Για κάθε ένα από αυτά τα δείγματα εφαρμόζεται ένας από τους αλγόριθμους επιβλεπόμενης ΜΜ για την εκπαίδευσή τους. Στην περίπτωση των τυχαίων δασών, ο αλγόριθμος που εφαρμόζεται είναι τα δέντρα απόφασης (που δημιουργούν το δάσος). Μετά την εκπαίδευση, λαμβάνεται η απόφαση από κάθε δέντρο και με τη μέθοδο της ψηφοφορίας κατά πλειοψηφία, εφόσον πρόκειται για εργασία ταξινόμησης, προβλέπεται η τελική απόδοση και γίνεται η ταξινόμηση.



Σχήμα 14. Η Τεχνική Bagging [31]

Η βάση για την κατανόηση του αλγόριθμου του τυχαίου δάσους, είναι η κατανόηση του αλγόριθμου των δέντρων απόφασης, που είναι μία μη παραμετρική μέθοδος επιβλεπόμενης μάθησης. Τα δέντρα απόφασης ονομάζονται έτσι επειδή, παρόμοια με ένα δέντρο, ο αλγόριθμος ξεκινά από έναν κόμβο-ρίζα που αναπτύσσεται σε κλαδιά και έτσι δημιουργείται μια δομή που μοιάζει με δέντρο, όπως φαίνεται στο Σχήμα 15.

Ο στόχος του αλγόριθμου του δέντρου απόφασης είναι η δημιουργία ενός μοντέλου που προβλέπει την τιμή της μεταβλητής στόχου μαθαίνοντας απλούς κανόνες απόφασης που συμπεραίνονται από τα χαρακτηριστικά των δεδομένων. Μαθηματικά, το πρόβλημα διατυπώνεται όπως ακολουθεί [29].



Σχήμα 15. Το Δέντρο Απόφασης [15]

Με δοσμένα τα διανύσματα εκπαίδευσης $x_i \in R^n$, $i=1,\ldots, l$ και το διάνυσμα ετικέτας $y \in R^l$, το δέντρο απόφασης διαχωρίζει αναδρομικά τον χώρο των χαρακτηριστικών έτσι ώστε δείγματα με την ίδια ετικέτα να ομαδοποιούνται μαζί. Έστω τα δεδομένα στον κόμβο m αναπαρίστανται με Q_m με N_m δείγματα. Για κάθε υποψήφιο διαχωρισμό $\theta = (j, t_m)$ που αποτελείται από ένα χαρακτηριστικό j και ένα κατώφλι t_m, τα δεδομένα διαχωρίζονται σε $Q_m^{left}(\theta)$ και $Q_m^{right}(\theta)$ υποσύνολα:

$$Q_m^{left}(\theta) = \{(x, y) | x_j <= t_m\}$$
$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta)$$

Η ποιότητα ενός υποψήφιου διαχωρισμού του κόμβου m υπολογίζεται χρησιμοποιώντας μια συνάρτηση μίξης (impurity function) ή συνάρτηση κόστους $H()$

$$G(Q_m, \theta) = \frac{N_m^{left}}{N_m} H(Q_m^{left}(\theta)) + \frac{N_m^{right}}{N_m} H(Q_m^{right}(\theta))$$

και επιλέγονται οι παράμετροι που ελαχιστοποιούν την μίξη:

$$\theta^* = \text{argmin}_\theta \, G(Q_m, \theta)$$

Η διαδικασία εκτελείται αναδρομικά για τα υποσύνολα $Q_m^{left}(\theta^*)$ και $Q_m^{right}(\theta^*)$ μέχρι το μέγιστο επιτρεπτό βάθος πλησιάσει το $N_m < \min_{samples}$ ή $N_m = 1$.

Στην περίπτωση της ταξινόμησης, εάν η ετικέτα λαμβάνει τιμές *0,1,…,K-1*, για τον κόμβο m, τότε

$$p_{mk} = 1/N_m \sum_{y \in Q_m} I(y = k)$$

είναι η αναλογία παρατηρήσεων της κλάσης k στον κόμβο m. Εάν ο m είναι τερματικός κόμβος, τότε η p_{mk} είναι η πιθανότητα πρόβλεψης.

Οι πιο συνηθισμένες συναρτήσεις κόστους είναι η gini και η εντροπία. Η gini δίνεται από τη σχέση:

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

Η εντροπία δίνεται από τη σχέση:

$$H(Q_m) = -\sum_k p_{mk}\log(p_{mk})$$

3 Σχετικές Εργασίες

Στο κεφάλαιο αυτό γίνεται μια ανασκόπηση της σχετικής βιβλιογραφίας, όπου παρουσιάζονται εργασίες που έχουν συνεισφέρει στην πρόγνωση του διαβήτη με εξόρυξη δεδομένων βάσει μεθόδων ΜΜ γενικά, και ειδικότερα εργασίες που βασίζονται στο σύνολο δεδομένων που επιλέχθηκε για την παρούσα εργασία.

3.1 Η Μηχανική Μάθηση στην Πρόβλεψη του Διαβήτη

Η πρόωρη διάγνωση του διαβήτη είναι ένας σημαντικός παράγοντας για τη θεραπεία της ασθένειας και μειώνει τις πιθανότητες για εμφάνιση άλλων επιπλοκών. Έτσι, έχουν αναπτυχθεί διάφορες υπολογιστικές μέθοδοι για την πρόβλεψη της ασθένεια που βασίζονται στην εξόρυξη δεδομένων με μεθόδους ΜΜ. Το θέμα έχει απασχολήσει ιδιαίτερα την ακαδημαϊκή κοινότητα. Ενδεικτικά αναφέρονται οι εργασίες των Chaki et al. με τίτλο «Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review» [32] και των Jaiswal et al. « A review on current advances in machine learning based diabetes prediction» [33].

Υπάρχουν διάφορα σύνολα δεδομένων με διαφορετικά χαρακτηριστικά και μεγέθη, που προέρχονται από διάφορες πηγές και έχουν χρησιμοποιηθεί για τη δημιουργία μοντέλων ΜΜ, όπως αναφέρεται στις [32] - [33]. Το πιο δημοφιλές σύνολο δεδομένων είναι το Pima Indian Diabetes [34] που προέρχεται από παρατηρήσεις σε 768 γυναίκες ασθενείς ηλικίας τουλάχιστον 21 ετών και περιέχει αριθμητικά χαρακτηριστικά που αφορούν τις εγκυμοσύνες, η συγκέντρωση πλάσματος στο αίμα, τη διαστολική αρτηριακή πίεση, το πάχος του δέρματος, το σάκχαρο, τον δείκτη μάζας σώματος, την γενεαλογική λειτουργία του διαβήτη, την ηλικία των ασθενών, καθώς και το αποτέλεσμα των παρατηρήσεων για το εάν έχουν την ασθένεια ή όχι.

Το σύνολο δεδομένων «Early stage diabetes risk prediction dataset» [9] που επιλέχθηκε για την παρούσα εργασία, έχει διατεθεί πρόσφατα στο αποθετήριο ΜΜ του UCI [10] προς πειραματισμούς από τους Islam et al.[12], και συνεπώς υπάρχουν πολύ λίγες όσον αφορά τον αριθμό σχετικές εργασίες, οι οποίες παρουσιάζονται στην επόμενη παράγραφο. Το σύνολο δεδομένων περιέχει παρατηρήσεις από 520 ασθενείς και των δύο φύλλων, διαφόρων ηλικιών. Τα χαρακτηριστικά του συνόλου αφορούν κατά κύριο λόγο 14 συμπτώματα της ασθένειας και τα δεδομένα συλλέχθηκαν βάσει ερωτηματολογίου - που,

σύμφωνα με τους δωρητές, εγκρίθηκε από ιατρούς- με αρνητικές ή θετικές απαντήσεις όσον αφορά την εμφάνιση των συμπτωμάτων. Το σύνολο συμπληρώνεται με παρατηρήσεις για το φύλο και την ηλικία των ασθενών, καθώς και εάν έχουν την ασθένεια ή όχι.

3.2 Εργασίες Σχετικές με το Σύνολο Δεδομένων των Islam et al.

Οι Islam et al. στην εργασία τους με τίτλο «Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques» [12], παρουσιάζουν για πρώτη φορά την εφαρμογή τεχνικών εξόρυξης δεδομένων με μεθόδους ΜΜ στο σύνολο δεδομένων με παρατηρήσεις που συνέλεξαν οι ίδιοι. Οι συγγραφείς αναλύουν το σύνολο δεδομένων με τους αλγόριθμους Naive Bayes (NB), Logistic Regression (LR) και Random Forest (RF) και στη συνέχεια εφαρμόζουν τις τεχνικές αποτίμησης της διασταυρούμενη επικύρωσης 10 τμημάτων (10-fold Cross Validation) και του ποσοστού διαχωρισμού (Percentage Split) 80-20. Τα καλύτερα αποτελέσματα έχει ο αλγόριθμος RF με ορθότητα 97,4% με την πρώτη τεχνική αποτίμησης και 99% με την δεύτερη. Στην εργασία δεν αναφέρεται το περιβάλλον υλοποίησης των δοκιμών και οι παράμετροι του κάθε μοντέλου.

Οι Das et al. στην εργασία τους με τίτλο «Prognostic Biomarkers Identification for Diabetes Prediction by Utilizing Machine Learning Classifiers» [35] εστιάζουν στον εντοπισμό των σημαντικότερων χαρακτηριστικών του συνόλου δεδομένων και επιλέγουν να εφαρμόσουν τους πειραματισμούς στα 10 από τα 15 χαρακτηριστικά του συνόλου, τα οποία εντοπίζουν με τις μεθόδους επιλογής χαρακτηριστικών Chi-Square, Minimum Redundancy Maximum Relevance (mRMR), και την Recursive Feature Elimination που βασίζεται στον Random Forest (RFE-RF) με τεχνική αποτίμησης τον διαχωρισμό 80-20. Στη συνέχεια, χρησιμοποιούν τους αλγόριθμους Decision Tree (DT), K-Nearest Neighbors (KNN), LR, NB, RF, Neural Network (NN), και SVM (Support Vector Machine). Τα καλύτερα αποτελέσματα είχε ο αλγόριθμος SVM με RBF πυρήνα με ορθότητα 95,192% με τη μέθοδο Chi-Square, 94,238% με τη μέθοδο mRMR και 98,077% με την μέθοδο RFE-RF. Επίσης, οι συγγραφείς πειραματίστηκαν με τους αλγόριθμους για τα 6 κοινά χαρακτηριστικά που εξήγαγαν και από τις τρεις μεθόδους. Τα καλύτερα αποτελέσματα είχε ο αλγόριθμος SVM με RBF πυρήνα με ορθότητα 89,423%. Στην εργασία δεν αναφέρεται το περιβάλλον υλοποίησης των δοκιμών, ο τρόπος προεπεξεργασίας των αριθμητικών και δυαδικών δεδομένων και οι παράμετροι του κάθε μοντέλου.

Οι Gamara et al. στην εργασία τους με τίτλο «Early Stage Diabetes Likelihood Prediction using Artificial Neural Networks» [36] πραγματοποιούν τους πειραματισμούς τους στο περιβάλλον ΜΜ του MATLAB [5] χρησιμοποιώντας τον αλγόριθμο NN. Χρησιμοποιούν την τεχνική αποτίμησης ποσοστού διαχωρισμού 70-15-15, όπου το πρώτο υποσύνολο χρησιμοποιείται για την εκπαίδευση, το δεύτερο για την επικύρωση και το τρίτο για τις δοκιμές του μοντέλου. Οι συγγραφείς παραμετροποιούν το TNN με ένα κρυφό επίπεδο 40 νευρώνων με συνάρτηση ενεργοποίησης την σιγμοειδή συνάρτηση και γραμμική συνάρτηση εξόδου. Τα αποτελέσματα των δοκιμών τους δίνουν ορθότητα 99,2% για το σύνολο εκπαίδευσης, 98,37% για το σύνολο επικύρωσης και 100% για το σύνολο δοκιμής.

Οι Nurjahan et al. στην εργασία τους με τίτλο «Mining Significant Features of Diabetes through Employing Various Classification Methods» [37] πραγματοποιούν τους πειραματισμούς τους επιλέγοντας τα σημαντικότερα χαρακτηριστικά στο περιβάλλον ΜΜ του Weka[6] και στη συνέχεια αναλύουν τα δεδομένα με τους αλγόριθμους Decision tree, K-Nearest Neighbour, NB, SVM, LR, extreme Gradient Boosting, MLP και RF με την βοήθεια της βιβλιοθήκης scikit-learn εφαρμόζοντας την τεχνική αποτίμησης της διασταυρούμενης επικύρωσης 10 τμημάτων. Επίσης, περιλαμβάνουν στις δοκιμές τους και το σύνολο δεδομένων Pima Indian Diabetes. Για το σύνολο δεδομένων Early stage diabetes risk prediction dataset, τα καλύτερα αποτελέσματα των δοκιμών τους έχει ο αλγόριθμος RF με ορθότητα 97,5%. Στην εργασία δεν αναφέρεται ποια είναι τα σημαντικότερα χαρακτηριστικά που επιλέχθηκαν και οι παράμετροι του κάθε μοντέλου.

Τέλος, οι Chaves and Marques στην εργασία τους με τίτλο «Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study» [38] οι συγγραφείς αναλύουν το σύνολο δεδομένων με τους αλγόριθμους NB, NN, AdaBoost, KNN και SVM εφαρμόζοντας την τεχνική αποτίμησης της διασταυρούμενης επικύρωσης 10 τμημάτων σε όλο το σύνολο δεδομένων και με το μοντέλο του NN να παρουσιάζει τα καλύτερα αποτελέσματα με ορθότητα 98,3%. Οι συγγραφείς πραγματοποίησαν τα πειράματά τους στο περιβάλλον ΜΜ Orange[7] που βασίζεται στη γλώσσα Python. Οι συγγραφείς παραθέτουν τον τρόπο προεπεξεργασίας των δεδομένων, καθώς και τις παραμέτρους του κάθε μοντέλου.

[5] https://www.mathworks.com/products/matlab.html

[6] http://old-www.cms.waikato.ac.nz/~ml/weka/

[7] https://orangedatamining.com/

Να σημειώσουμε ότι, στην πλειοψηφία των προαναφερόμενων εργασιών οι συγγραφείς παραθέτουν και άλλες μετρικές απόδοσης των ταξινομητών τους, εκτός από την ορθότητα, όπως για παράδειγμα η ακρίβεια (precision), η ανάκληση (recall) κ.α.. Απλά, συνηθίζεται να αναφέρεται ως κύρια μετρική απόδοσης για τη σύγκριση των μοντέλων πρώτα η ορθότητα και στη συνέχεια να εξετάζονται οι τιμές των υπόλοιπων μετρικών, προκειμένου να υπάρχει μια καλύτερη εικόνα για τη συνολική αποτίμηση ενός ταξινομητή. Τα μέτρα απόδοσης ενός ταξινομητή θα παρουσιαστούν αναλυτικά στο επόμενο κεφάλαιο.

4 Μεθοδολογία

Το κεφάλαιο αυτό αφορά τη μεθοδολογία υλοποίησης του έργου. Αρχικά παρουσιάζεται το περιβάλλον υλοποίησης σε Python. Στη συνέχεια, παρουσιάζονται τα βασικά βήματα της υλοποίησης.

4.1 Περιβάλλον Υλοποίησης

Για τη δημιουργία του έργου επιλέχθηκε η γλώσσα προγραμματισμού Python, η οποία στις μέρες μας είναι η δημοφιλέστερη για προβλήματα ΒΜ και ανάλυσης δεδομένων.

Η Python είναι μια διερμηνευόμενη (interpreted), υψηλού επιπέδου και γενικού σκοπού γλώσσα προγραμματισμού, δημιουργήθηκε από τον Guido van Rossum και παρουσιάστηκε το 1991. Η φιλοσοφία σχεδιασμού της Python δίνει έμφαση στην αναγνωσιμότητα του κώδικα. Οι γλωσσικές δομές και η αντικειμενοστραφής προσέγγιση στοχεύουν να βοηθήσουν τους προγραμματιστές να γράψουν σαφή, λογικό κώδικα για μικρά και μεγάλα έργα. Η Python είναι ισχυρή και γρήγορη, τρέχει παντού, είναι φιλική και εύχρηστη και είναι δωρεάν και ανοιχτού κώδικα. Η έκδοση 2 της Python επίσημα σταμάτησε να υποστηρίζεται το 2020 και αυτή τη στιγμή υπάρχει υποστήριξη για τις εκδόσεις 3.6.x και τις μεταγενέστερες.

Η υλοποίηση έγινε στο περιβάλλον της ελεύθερης διανομής του Anaconda[8] (anaconda individual edition 2021.05-64bits), το οποίο είναι ένα περιβάλλον βασισμένο στην Python 3.8 και είναι μία από τις δημοφιλέστερες πλατφόρμες για την επιστήμη των δεδομένων και περιέχει πολλές προεγκατεστημένες βιβλιοθήκες ανοιχτού κώδικα διαφορετικών αλγορίθμων.

Τα αρχεία δημιουργήθηκαν στο Jupyter[9] ως notebooks. Το Jupyter είναι ένα έργο ανοιχτού κώδικα που δημιουργήθηκε για την υποστήριξη αλληλεπίδρασης της επιστήμης των δεδομένων και του επιστημονικού υπολογισμού σε διάφορες γλώσσες προγραμματισμού. Το Jupyter προσφέρει ένα web-based περιβάλλον για εργασία με σημειωματάρια (notebooks) που περιέχουν κώδικα, δεδομένων και κείμενο. Τα Jupyter notebooks είναι τα πρότυπα workspace για την επιστήμη δεδομένων με Python.

[8] https://www.anaconda.com/

[9] https://jupyter.org/

4.1.1 Βιβλιοθήκες Python

Οι σημαντικότερες βιβλιοθήκες της γλώσσας προγραμματισμού Python που χρησιμοποιήθηκαν στο έργο είναι ελεύθερες και ανοιχτού κώδικα (free and open-source), υποστηρίζονται από μεγάλες κοινότητες προγραμματιστών και διαθέτουν πολύ καλή τεκμηρίωση (documentation). Οι βιβλιοθήκες που χρησιμοποιήθηκαν και περιλαμβάνονται στη διανομή του Anaconda περιγράφονται συνοπτικά στη συνέχεια.

A. SciPy: Η βασική βιβλιοθήκη για επιστημονικούς και τεχνικούς υπολογισμούς [39].

Η SciPy περιέχει λειτουργικές μονάδες για βελτιστοποίηση, γραμμική άλγεβρα, ολοκλήρωση, παρεμβολή, ειδικές λειτουργίες και άλλες εργασίες κοινές στην επιστήμη και τη μηχανική. Βασίζεται στο αντικείμενο των πινάκων NumPy, αποτελεί μέρος της στοίβας NumPy και το οικοσύστημά της περιλαμβάνει ένα σύνολο επιστημονικών βιβλιοθηκών. Τα βασικά πακέτα- βιβλιοθήκες του οικοσυστήματος της SciPy είναι τα παρακάτω:

NumPy[10]: Βασική βιβλιοθήκη για N-διάστατους πίνακες.

Η NumPy προσθέτει υποστήριξη για μεγάλους, πολυδιάστατους πίνακες και μητρώα, μαζί με μια μεγάλη συλλογή μαθηματικών συναρτήσεων υψηλού επιπέδου για λειτουργίες γραμμικής άλγεβρας. Δημιουργήθηκε το 2005, με βάση τις πρώτες εργασίες των βιβλιοθηκών Numerical και Numarray.

pandas[11]: Βιβλιοθήκη για δομές δεδομένων και ανάλυση.

Η βιβλιοθήκη pandas είναι για χειρισμό και ανάλυση δεδομένων. Συγκεκριμένα, προσφέρει δομές δεδομένων και λειτουργίες για χειρισμό αριθμητικών πινάκων και χρονοσειρών. Το όνομα προέρχεται από τον όρο "panel data", έναν όρο οικονομετρίας για σύνολα δεδομένων που περιλαμβάνει παρατηρήσεις σε πολλαπλές χρονικές περιόδους για τα ίδια αντικείμενα.

Matplotlib[12]: Βιβλιοθήκη για ολοκληρωμένη 2-D σχεδίαση.

Η matplotlib είναι η βιβλιοθήκη σχεδίασης και η αριθμητική επέκταση μαθηματικών της NumPy. Παρέχει μια αντικειμενοστραφή διεπαφή προγραμματισμού (Application Programming Interface – API) για την ενσωμάτωση σχεδίων σε εφαρμογές με χρήση εργαλείων γενικής χρήσης γραφικών διεπαφών χρήστη (Graphical User Interface – GUI). Η matplotlib δεν υποστηρίζει την Python 2 μετά το 2020. Η Pyplot είναι μια ενότητα της matplotlib που παρέχει διεπαφή τύπου MATLAB[13]. Έχει σχεδιαστεί έτσι ώστε, να μπορεί

[10] https://numpy.org/

[11] https://pandas.pydata.org/

[12] https://matplotlib.org/

[13] https://www.mathworks.com/products/matlab.html

να χρησιμοποιηθεί όπως το MATLAB, με τη δυνατότητα χρήσης της Python και το πλεονέκτημα του ότι είναι δωρεάν και ανοιχτού κώδικα. Η βιβλιοθήκη **Seaborn** είναι για την οπτικοποίηση δεδομένων και βασίζεται στην matplotlib. Παρέχει διεπαφή υψηλού επιπέδου για τη σχεδίαση ελκυστικών και ενημερωτικών στατιστικών γραφικών.

IPython: Η βελτιωμένη διαδραστική κονσόλα

Η IPython παρέχει μια πλούσια αρχιτεκτονική για διαδραστικούς υπολογισμούς με ισχυρό διαδραστικό κέλυφος (shell) και είναι πυρήνας (kernel) για το Jupyter. Το Jupyter παρέχει τη λειτουργικότητα IPython στο πρόγραμμα περιήγησης στο Web, δίνοντας τη δυνατότητα τεκμηρίωσης των υπολογισμών. Επίσης, υποστηρίζει τη διαδραστική οπτικοποίηση δεδομένων και τη χρήση εργαλείων GUI και εύκολα στη χρήση εργαλεία υψηλής απόδοσης για παράλληλους υπολογισμούς**.**

B. scikit-learn: Βιβλιοθήκη για Μηχανική Μάθηση [29]

Η βιβλιοθήκη scikit learn[14] διαθέτει διάφορους αλγόριθμους ΜΜ για ταξινόμηση, παλινδρόμηση, συσταδοποίηση, MLP, προεπεξεργασία δεδομένων (εξαγωγή χαρακτηριστικών και εξομάλυνση) και άλλους. Έχει σχεδιαστεί για να λειτουργεί με τις αριθμητικές και επιστημονικές βιβλιοθήκες NumPy και SciPy.

4.2 Μεθοδολογία Υλοποίησης

Η πρόωρη πρόβλεψη του διαβήτη με τη χρήση τεχνικών εξόρυξης δεδομένων εφαρμόζοντας αλγόριθμους ΜΜ από ένα σύνολο δεδομένων όπου υπάρχουν οι παρατηρήσεις στοιχείων ασθενών που έχουν ή δεν έχουν τη νόσο, είναι ένα πρόβλημα επιβλεπόμενης ΜΜ που ανήκει στην κατηγορία της δυαδικής ταξινόμησης. Με βάση τα καταγραμμένα χαρακτηριστικά, είτε πρόκειται για κλινικές εξετάσεις, είτε για καταγραφή συμπτωμάτων, καλούμαστε να διερευνήσουμε ποιος είναι ο καταλληλότερος αλγόριθμος ΜΜ, ώστε να δημιουργήσουμε το καλύτερο μοντέλο πρόβλεψης για νέα δεδομένα με τα ίδια χαρακτηριστικά. Εφόσον έχει προσδιοριστεί το πρόβλημα, ακολουθούνται συγκεκριμένα βήματα που φαίνονται στο Σχήμα 16 και που αφορούν:

- Την επιλογή του συνόλου δεδομένων
- Την εξερεύνηση του συνόλου δεδομένων
- Την επιλογή των μέτρων απόδοσης των μοντέλων που θα δημιουργηθούν

[14] https://scikit-learn.org/stable/

- Την επιλογή της τεχνικής αποτίμησης του συνόλου δεδομένων κατά την εκπαίδευση
- Την επιλογή των αλγόριθμων που θα εφαρμοστούν για τη δημιουργία των μοντέλων

Το σύνολο δεδομένων που έχει επιλεγεί για την επίλυση του προβλήματος είναι ένα σύνολο δεδομένων με 520 παρατηρήσεις- στιγμιότυπα, τα οποία έχουν συλλεχθεί χρησιμοποιώντας ερωτηματολόγια από ασθενείς ενός νοσοκομείου για διαβητικούς της πόλης Sylhet στο Bangladesh από τους Islam et al.[12] και που στη συνέχεια οι συγγραφείς το διέθεσαν στο αποθετήριο μηχανικής μάθησης του πανεπιστημίου της Καλιφόρνια (UCI) των ΗΠΑ [10].

Σχήμα 16. Η Μεθοδολογία Υλοποίησης

Στη συνέχεια, παρουσιάζονται αναλυτικά τα περιεχόμενα του συνόλου δεδομένων και την επεξηγηματική ανάλυση δεδομένων.

4.3 Εξερεύνηση Συνόλου Δεδομένων

Η εξερεύνηση του συνόλου δεδομένων αφορά αφενός το περιεχόμενό του και, αφετέρου, την διερευνητική ανάλυση των δεδομένων. Η διερευνητική ανάλυση δεδομένων (Exploratory Data Analysis – EDA) είναι μια προσέγγιση για την ανάλυση ενός συνόλου δεδομένων για τη σύνοψη των κύριων χαρακτηριστικών των δεδομένων με τη χρήση γραφικών παραστάσεων και μεθόδους οπτικοποίησης δεδομένων. Πιο απλά, είναι ένας τρόπος για την εικόνα των δεδομένων και τις μεταξύ τους συσχετίσεις [18].

4.3.1 Περιγραφή του Συνόλου Δεδομένων

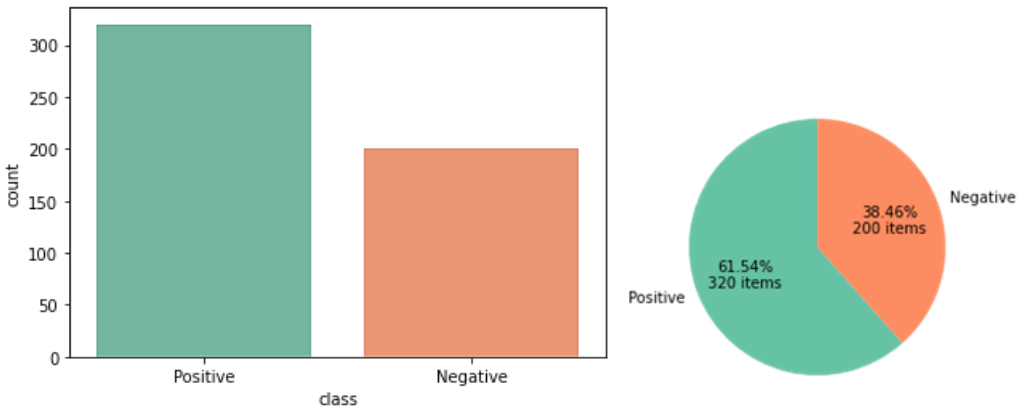
Το σύνολο δεδομένων περιέχει παρατηρήσεις από 520 ασθενείς με 16 χαρακτηριστικά για κάθε στιγμιότυπο και μία μεταβλητή που περιγράφει εάν ο ασθενής έχει την ασθένεια (positive) ή δεν την έχει (negative) που είναι και η κλάση ταξινόμησης.

Πίνακας 2. Περιγραφή Στοιχείων Στιγμιότυπων

	Χαρακτηριστικά	**Τιμές**	**Αριθμητικές Τιμές**
1	Ηλικία (Age)	16-90	16-90
2	Φύλο (Gender)	Male, Female	1, 0
3	Πολυουρία (Polyuria)	Yes, No	1, 0
4	Πολυδιψία (Polydipsia)	Yes, No.	1, 0
5	Ξαφνική απώλεια βάρους (sudden weight loss)	Yes, No	1, 0
6	Αδυναμία (weakness)	Yes, No	1, 0
7	Πολυφαγία (Polyphagia)	Yes, No	1, 0
8	Γενετική μυκητίαση (Genital thrush)	Yes, No	1,0
9	Θολή όραση (visual blurring)	Yes, No	1, 0
10	Φαγούρα (Itching)	Yes, No	1, 0
11	Ανησυχία (Irritability)	Yes, No	1, 0
12	Καθυστερημένη επούλωση (delayed healing)	Yes, No	1, 0
13	Μερική πάρεση (partial paresis)	Yes, No	1, 0
14	Μυική αδυναμία (muscle stiffness)	Yes, No	1, 0
15	Αλωπεκία (Alopecia)	Yes, No	1, 0
16	Παχυσαρκία (Obesity)	Yes, No	1, 0
17	Κλάση (class)	Positive, Negative	1, 0

Η δομή των δεδομένων του συνόλου, καθώς και οι τιμές που παίρνουν τα στιγμιότυπα περιγράφεται αναλυτικά στον Πίνακα 2. Μέσα σε παρένθεση αναγράφονται οι ετικέτες των χαρακτηριστικών, όπως αναφέρονται στο αρχικό σύνολο. Επίσης, επειδή στην ΜΜ εργαζόμαστε με αριθμητικές τιμές, στην τελευταία στήλη του Πίνακα 2 φαίνεται η αρχική αντιστοιχία των αριθμητικών τιμών χωρίς τεχνικές κλιμάκωσης.

Να σημειώσουμε ότι, οι συγγραφείς στο αρχικό τους άρθρο αναφέρουν ότι υπήρχαν ελλείπεις τιμές στα στιγμιότυπα και απομάκρυναν 20 εγγραφές, με συνέπεια να μείνουν 500 εγγραφές. Στο σύνολο που διατίθεται στο αποθετήριο, δεν υπάρχουν ελλιπείς τιμές, συνεπώς τα στοιχεία αφορούν 520 εγγραφές. Από αυτές τις εγγραφές, προκύπτει ότι 320 στιγμιότυπα και σε ποσοστό 61,54% ανήκουν στην κλάση Positive, ενώ 200 στιγμιότυπα και ποσοστό 38,46% ανήκουν στην κλάση Negative, όπως φαίνεται στο Σχήμα 17.



Σχήμα 17. Κατανομή Κλάσης Ταξινόμησης

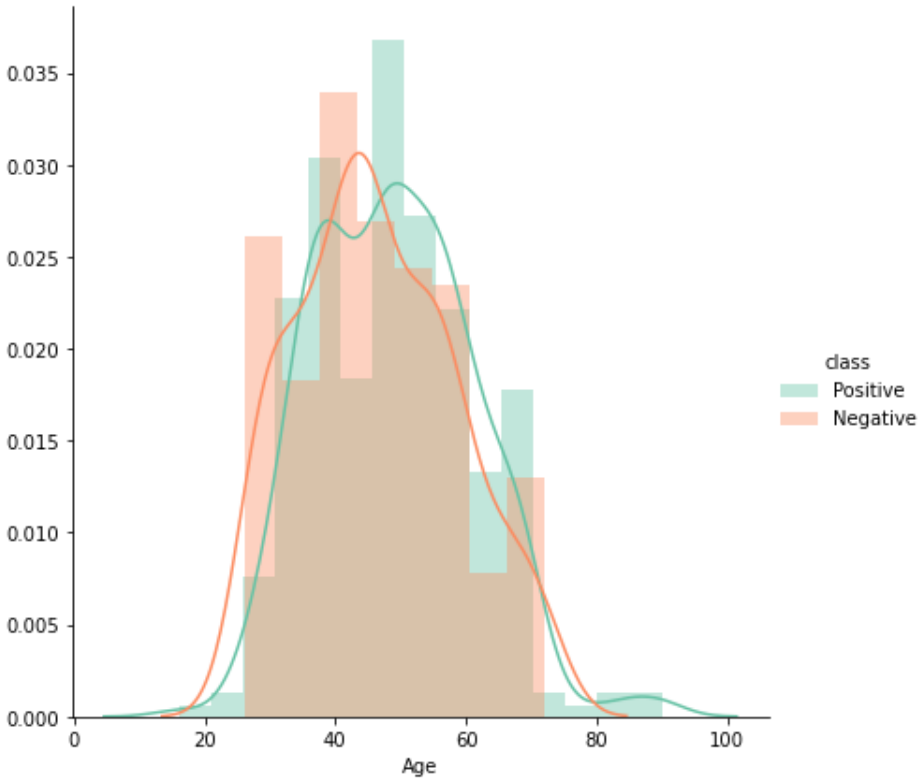
4.3.2 Διερευνητική Ανάλυση Δεδομένων

Μετά την εξερεύνηση του περιεχομένου του συνόλου των δεδομένων, ακολουθεί η διερεύνηση των κατανομών διάφορων χαρακτηριστικών και η μεταξύ τους συσχέτιση. Στα πλαίσια της παρούσας εργασίας, η ανάλυση αυτή είναι ενδεικτική και όχι εξαντλητική.

4.3.2.1 Κατανομή Ηλικίας

Η ηλικιακή κατανομή φαίνεται στο Σχήμα 18, όπου παρατηρούμε ότι η πλειοψηφία των δειγμάτων αφορά τις ηλικίες από 36-65 ετών. Από αυτή την κατανομή θα μπορούσαμε να συμπεράνουμε ότι, εφόσον η έρευνα πραγματοποιήθηκε σε νοσοκομείο για διαβήτη, η πλειοψηφία των ασθενών που ίσως έχουν την ασθένεια είναι ενήλικες. Σύμφωνα με τα στατιστικά στοιχεία που αφορούν τις ηλικίες που πλήττονται από την ασθένεια, ειδικά όσον

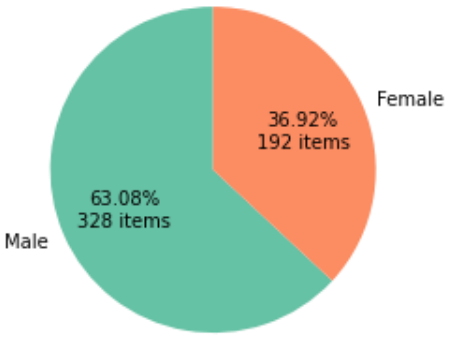
αφορά τον Τύπο 2 του ΣΔ, η πλειοψηφία του γενικού συνόλου των ασθενών ανήκουν σε αυτό το εύρος των ηλικιών.



Σχήμα 18. Κατανομή του Χαρακτηριστικού Ηλικία

4.3.2.2 Κατανομή Φύλου

Η κατανομή των παρατηρήσεων όσον αφορά το χαρακτηριστικό φύλο φαίνεται στο Σχήμα 19. Από εδώ, δεν μπορούμε να εξάγουμε με ασφάλεια κάποιο συμπέρασμα, παραμόνο ότι η πλειοψηφία των παρατηρήσεων αφορά άντρες σε ποσοστό 63,08%, δηλαδή τα 2/3 των παρατηρήσεων του συνόλου. Ίσως να ήταν πιο χρήσιμο, από τη στιγμή που υπάρχουν διαφορετικά γενετικά χαρακτηριστικά σε κάθε φύλο, το σύνολο των παρατηρήσεων να ήταν μόνο για ένα από τα δύο φύλα.



Σχήμα 19. Κατανομή του Χαρακτηριστικού Φύλο

4.3.2.3 Συσχέτιση Χαρακτηριστικών με την Κλάση Ταξινόμησης

Η συσχέτιση του κάθε δυαδικού χαρακτηριστικού του συνόλου δεδομένων σε σχέση με την κλάση ταξινόμησης φαίνεται στο Σχήμα 20.

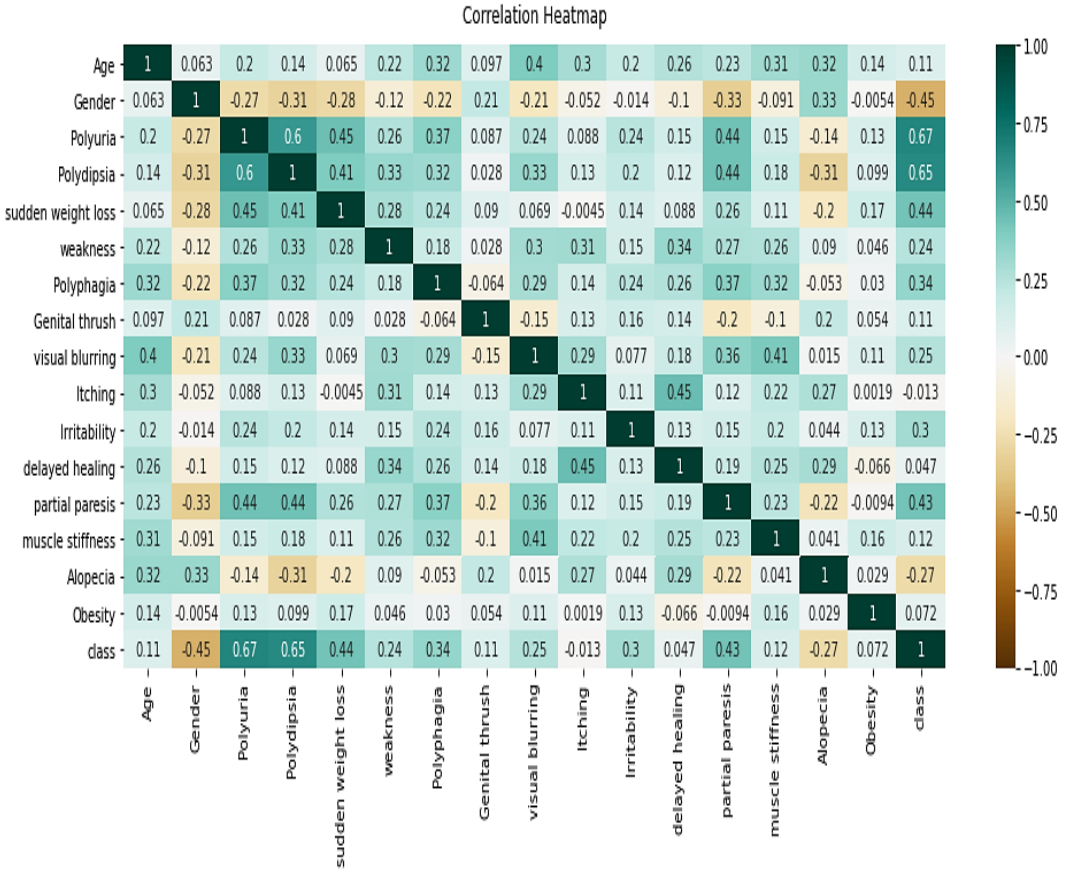
Σχήμα 20. Συσχέτιση Χαρακτηριστικών με την Κλάση Ταξινόμησης

Παρατηρούμε ότι, στο σύνολο δεδομένων υπάρχει μια ανισορροπία όσον αφορά τα θετικά δείγματα για τους άνδρες και τις γυναίκες. Η πλειοψηφία των στιγμιότυπων είναι άνδρες και οι περισσότεροι είναι αρνητικοί, ενώ η μειοψηφία των στιγμιότυπων είναι γυναίκες και οι περισσότερες είναι θετικές.

Επίσης, παρατηρούμε ότι, τα κυριότερα συμπτώματα των παρατηρήσεων της θετικής κλάσης είναι η πολυουρία, η πολυδιψία και η ξαφνική απώλεια βάρους. Ανατρέχοντας στην παράγραφο 2.1.3 Συμπτώματα και Διάγνωση της Ασθένειας, διαπιστώνουμε ότι αυτά τα συμπτώματα ταυτίζονται με αυτά που αναφέρονται ως κύρια συμπτώματα για τη διάγνωση του διαβήτη [13].

4.3.2.4 Συσχετίσεις Χαρακτηριστικών

Μια χρήσιμη πληροφορία για τη συσχέτιση των χαρακτηριστικών του συνόλου δεδομένων ανά δύο σε ένα διάγραμμα, δίνεται από το διάγραμμα «θερμότητας» (heatmap) του μητρώου συσχέτισης, εφόσον όλες οι τιμές του συνόλου είναι αριθμητικές. Έτσι, αφού μετατρέψουμε τις τιμές σε αριθμητικές, μπορούμε να εξάγουμε το μητρώο συσχέτισης για όλο το σύνολο δεδομένων σε μορφή heatmap φαίνεται στο Σχήμα 21.



Σχήμα 21. Πίνακας Θερμότητας Συσχέτισης Χαρακτηριστικών

Οι συντελεστές συσχέτισης στο μητρώο κυμαίνονται από το -1 έως το 1. Όταν οι συντελεστές είναι κοντά στο 1, σημαίνει ότι υπάρχει μια ισχυρή θετική συσχέτιση. Όταν οι συντελεστές είναι κοντά στο -1, σημαίνει ότι υπάρχει μια αρνητική συσχέτιση. Όταν οι συντελεστής είναι κοντά στο 0, σημαίνει ότι δεν υπάρχει γραμμική συσχέτιση [18].

Τα συμπεράσματα της συσχέτισης της κλάσης με τα συμπτώματα είχαν εξαχθεί από τα διαγράμματα της προηγούμενης παραγράφου και εδώ επιβεβαιώνονται. Μια επιπλέον πληροφορία που μπορούμε να εξάγουμε αφορά τη συσχέτιση της ηλικίας με την κλάση. Παρατηρούμε ότι, η κλάση με την ηλικία έχουν θετική συσχέτιση, δηλαδή όταν μεγαλώνει η ηλικία, η κλάση τείνει να είναι θετική. Άλλες πληροφορίες που εξάγουμε είναι η συσχέτιση των συμπτωμάτων ανά δύο. Για παράδειγμα, βλέπουμε τη θετική συσχέτιση της πολυδιψίας με την πολυουρία και την ξαφνική απώλεια βάρους. Επίσης, παρατηρούμε την αρνητική συσχέτιση του φύλου με την κλάση, κάτι που ήταν αναμενόμενο λόγω της άνισης κατανομής του δείγματος σε άνδρες και γυναίκες, καθώς και της αναλογίας των θετικών και των αρνητικών χαρακτηριστικών.

4.4 Επιλογή Μέτρων Απόδοσης

Η αποτίμηση της ποιότητας ενός μοντέλου για δυαδική ταξινόμηση, δηλαδή ενός **δυαδικού ταξινομητή** γίνεται με βάση κάποια μέτρα απόδοσης. Γενικά, σε μια εργασία ταξινόμησης γενικά, ακολουθούνται οι εξής συμβάσεις για την ονοματολογία και τους συμβολισμούς [40]:

- Ο κλάσεις είναι δύο: η *θετική (Positive – P)* και η *αρνητική (Negative – N).*
- Ο αριθμός των προβλέψεων που ανήκουν στη *θετική* κλάση, ονομάζονται *αληθώς θετικά (True Positives- TP),* δηλαδή είναι ο αριθμός των σωστών προβλέψεων για ασθένεια.
- Ο αριθμός των προβλέψεων που ανήκουν στην αρνητική κλάση, ονομάζονται *αληθώς αρνητικά (True Negatives- TN),* δηλαδή είναι ο αριθμός των σωστών προβλέψεων για μη ασθένεια.
- Ο αριθμός προβλέψεων που ταξινομήθηκαν στη θετική κλάση, ενώ στην πραγματικότητα ανήκουν στην αρνητική, ονομάζονται *ψευδώς αρνητικά (False Positive- FP).*

- Ο αριθμός προβλέψεων που ταξινομήθηκαν στην αρνητική κλάση ενώ στην πραγματικότητα ανήκουν στη θετική, ονομάζονται *ψευδώς αρνητικά (False Negatives - FN).*

4.4.1 Μετρικές απόδοσης

Οι μετρικές απόδοσης για τη δυαδική ταξινόμηση είναι πολλές και η βιβλιοθήκη scikit-learn διαθέτει τις περισσότερες από αυτές που θα παρουσιαστούν στη συνέχεια.

Η *ορθότητα (accuracy)* είναι το πιο συνηθισμένο μέτρο απόδοσης ενός μοντέλου. Η ορθότητα είναι η αναλογία των παραδειγμάτων για τα οποία έχει γίνει η σωστή πρόβλεψη επί του συνόλου των προβλέψεων και με βάση τις παραπάνω συμβάσεις και ισοδυναμεί με τον δείκτη λάθους (error rate) του μοντέλου [1]. Η ορθότητα δίνεται από τον τύπο:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Η ορθότητα ως μέτρο απόδοσης είναι η αναλογία των παραδειγμάτων για τα οποία έγινε σωστή πρόβλεψη. Όμως, σε προβλήματα του πραγματικού κόσμου οι κλάσεις των δεδομένων στα σύνολα δεν κατανεμημένες με ισορροπία. Για παράδειγμα, σε ένα σύνολο δεδομένων που αφορούν μια πολύ σπάνια περίπτωση ασθένειας μπορεί το 90% των παραδειγμάτων να ανήκουν στην αρνητική κλάση και μόνο το 10% στη θετική. Σε τέτοια περίπτωση, μπορεί να έχουμε ένα μοντέλο με υψηλή ορθότητα, αλλά το μοντέλο δεν έχει προγνωστική δύναμη. Συνεπώς, η ορθότητα ως μετρική απόδοσης δεν αρκεί. Για τον σκοπό αυτό, στην ταξινόμηση χρησιμοποιούνται και άλλες μετρικές απόδοσης, όπως περιγράφονται αναλυτικά από τους συγγραφείς στις [18], [40], [41] και παρουσιάζονται στη συνέχεια.

Η *ακρίβεια* (*precision*), είναι η αναλογία του κάθε παραδείγματος που προβλέπεται ως θετικό και είναι πραγματικά θετικό. Μπορεί να θεωρηθεί ως μέτρο θορύβου στις προβλέψεις, δηλαδή την πιθανότητα της ορθότητας για μια θετική πρόβλεψη. Η ακρίβεια υπολογίζεται από τον τύπο:

$$Precision = \frac{TP}{FP + TP}$$

Η *ανάκληση* (*recall*), γνωστή και ως ευαισθησία (sensitivity) ή δείκτης αληθώς θετικών (True Positive Rate – TPR), είναι η αναλογία κάθε θετικού παραδείγματος που

είναι πραγματικά θετικό. Η ανάκληση μετρά την ικανότητα του μοντέλου να αναγνωρίζει ένα παράδειγμα της θετικής κλάσης και υπολογίζεται από τον τύπο:

$$Recall = TPR = \frac{TP}{TP + FN}$$

Συνοπτικά, η ακρίβεια είναι το κλάσμα των προβλέψεων που αναφέρθηκαν από το μοντέλο ως σωστές και η ανάκληση είναι το κλάσμα των πραγματικών παραδειγμάτων που έχουν προβλεφθεί.

Σε πολλές περιπτώσεις, εάν θέλουμε να συνοψίσουμε την απόδοση του μοντέλου, επιδιώκοντας ένα είδος ισορροπίας μεταξύ της ακρίβειας και την ανάκλησης, χρησιμοποιούμε μια άλλη μετρική που συνδυάζει τις δύο μετρικές, τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης που ονομάζεται *F1 αποτέλεσμα* (*F1* score) και υπολογίζεται από τον τύπο:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

Η *προσδιοριστικότητα (specificity)*, γνωστή και ως δείκτης αληθώς αρνητικών (True Negative Rate – TNR) μετρά τις ορθά αρνητικές προβλέψεις στο σύνολο των ορθών αρνητικών παραδειγμάτων και υπολογίζεται από τον τύπο:

$$Specificity = TNR = \frac{TN}{TN + FP}$$

Η μετρική του δείκτη ψευδώς αρνητικών (False Positive Rate- *FPR*) αντιστοιχεί στην αναλογία των αρνητικών παραδειγμάτων που θεωρήθηκαν ως θετικά, σε σχέση με όλα τα αρνητικά παραδείγματα. Όσο μεγαλύτερος είναι ο *FPR*, πόσα περισσότερα αρνητικά παραδείγματα έχουν ταξινομηθεί λάθος. Ο *FPR* υπολογίζεται από τον τύπο:

$$FPR = \frac{FP}{FP + TN} = 1 - specificity = 1 - TNR$$

Όταν χρησιμοποιούνται ως μετρικές εκτίμησης απόδοσης του δυαδικού ταξινομητή η ακρίβεια και η ανάκληση, συνηθίζεται να απεικονίζονται γραφικά με την PR καμπύλη (PR curve) με την ακρίβεια να αντιστοιχεί στον άξονα των y και την ανάκληση στον άξονα των x.

4.4.2 Πίνακας ταξινόμησης

Για μια εργασία ταξινόμησης συνηθίζεται να δημιουργείται *ο πίνακας σύγχυσης (confusion matrix)*, που συνήθως αναφέρεται ως πίνακας ταξινόμησης**.** Ο πίνακας ταξινόμησης είναι ένας δισδιάστατος πίνακας ο οποίος παρέχει τη δυνατότητα οπτικοποίησης της απόδοσης ενός μοντέλου [40], [42].

		Πραγματική Κλάση		
	Συνολικός πληθυσμός	Κλάση θετική	Κλάση αρνητική	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$
Κλάση πρόγνωσης	Κλάση πρόγνωσης θετική	**True positive (TP)**	**False positive (FP)**	Precision $\frac{TP}{FP+TP}$
	Κλάση πρόγνωσης αρνητική	**False negative (FP)**	**True negative (TN)**	
		True positive rate (TPR), Recall, Sensitivity $\frac{TP}{TP+FN}$	False positive rate (FPR) $\frac{FP}{FP+TN}$	F1 score = $\frac{2*Precision*Recall}{Precision+Recall}$
		False negative rate (FNR) $1-TPR$	Specificity True negative rate (TNR) $\frac{TN}{TN+FP}$	

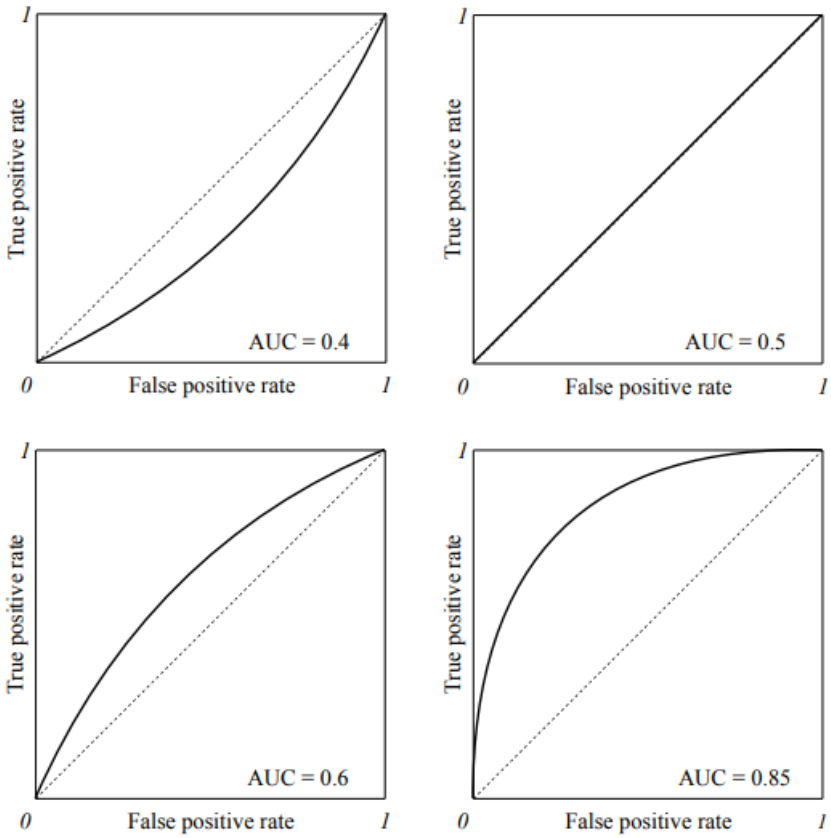
Σχήμα 22. Ο Πίνακας Ταξινόμησης και οι Μετρικές Απόδοσης [42]

Στην περίπτωση της δυαδικής ταξινόμησης είναι ένα μητρώο 2x2, όπως δίνεται στο Σχήμα 22 και κάθε γραμμή παριστάνει τα παραδείγματα της κλάσης που έχει προβλέψει ο ταξινομητής (κλάση πρόγνωσης), ενώ κάθε στήλη παριστάνει τα παραδείγματα της κλάσης που ανήκουν πραγματικά τα παραδείγματα (πραγματική κλάση). Ο όρος σύγχυση προκύπτει από το γεγονός από το ότι μπορούμε να δούμε εύκολα εάν το μοντέλο συγχέει τις δύο κλάσεις, δηλαδή εάν ταξινομεί λάθος σε μία κλάση από αυτή που αντιστοιχεί πραγματικά σε ένα παράδειγμα. Ο πίνακας ταξινόμησης συνήθως οπτικοποιείται με τη μορφή heatmap και παρέχει τη δυνατότητα να δούμε όχι μόνο εάν το μοντέλο είναι λάθος, αλλά και τι πήγε λάθος. Από τον πίνακα ταξινόμησης μπορούν να εξαχθούν με κατάλληλους

υπολογισμούς όλες οι μετρικές απόδοσης που αναφέρθηκαν προηγουμένως, όπως φαίνεται στο Σχήμα 23.

4.4.3 Ποιοτικοί Δείκτες

Ένας εύκολος σχετικά γραφικός τρόπος και ο πιο συνηθισμένος για την ποιοτική εκτίμηση της απόδοσης ενός ταξινομητή είναι η λήψη χαρακτηριστικής καμπύλης λειτουργίας (Receiving Operating Characteristic curve – ROC curve). Η ROC καμπύλη είναι η σχεδίαση της μετρικής *TPR*, δηλαδή της ανάκλησης, σε σχέση με την μετρική *FPR* και αναπαριστά τα *TP* και τα *FP* δηλαδή την πιθανότητα στην οποία ένα παράδειγμα προβλέπεται να ανήκει σε μια κλάση. Η ROC καμπύλη χρησιμοποιείται ως μια γενική μετρική του μοντέλου. Όσο καλύτερο είναι ένα μοντέλο, τόσο υψηλότερα είναι η καμπύλη και συνεπώς τόσο μεγαλύτερη η επιφάνεια κάτω από την καμπύλη. Στις βιβλιοθήκες MM υπάρχει συνάρτηση για τον υπολογισμό της επιφάνειας κάτω από την ROC καμπύλη (Area Under the Curve – AUC) και ονομάζεται ROC AUC ή απλά AUC. Ο τέλειος ταξινομητής έχει ROC AUC ίση με 1, ενώ ένας κακός ταξινομητής θα έχει 0.5, ίσως και λιγότερο [18]. Συνεπώς, όσο πιο κοντά η ROC AUC στο 1, τόσο καλύτερος είναι ο ταξινομητής. Παραδείγματα των ROC και ROC AUC φαίνονται στο Σχήμα 22.



Σχήμα 23. ROC Καμπύλες και AUC [41]

Ένας άλλος συντελεστής ποιοτικός συντελεστής που χρησιμοποιείται στην ταξινόμηση, είναι ο συντελεστής συσχέτισης του Matthews (Matthews Correlation Coefficient - MCC). Ο MCC χρησιμοποιείται στη ΜΜ ως μέτρο της ποιότητας της ταξινόμησης. Λαμβάνει υπόψη τα αληθώς και ψευδώς θετικά και αρνητικά και θεωρείται γενικά ως ισορροπημένο μέτρο που μπορεί να χρησιμοποιηθεί ακόμη και αν οι τάξεις έχουν πολύ διαφορετικά μεγέθη [43]. Ο MCC είναι ουσιαστικά μια τιμή συντελεστή συσχέτισης μεταξύ -1 και +1. Ένας συντελεστής +1 αντιπροσωπεύει μια τέλεια πρόβλεψη, 0 μια μέση τυχαία πρόβλεψη και -1 μια αντίστροφη πρόβλεψη . Ο MCC υπολογίζεται από τον τύπο:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4.5 Τεχνικές Αποτίμησης

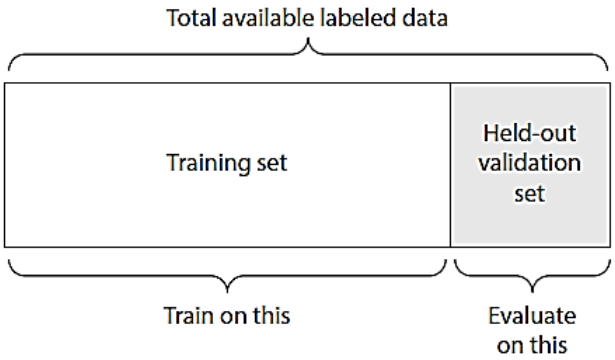
Η διαδικασία δημιουργίας ενός προγνωστικού μοντέλου για μια εργασία και ένα συγκεκριμένο σύνολο δεδομένων είναι η εκπαίδευση. Ο αλγόριθμος ΜΜ με συνεχείς δοκιμές και επαναλήψεις καλείται να εξάγει το ορθό αποτέλεσμα για τον στόχο.

Μετά από ορισμένες επαναλήψεις, το μοντέλο αποτιμάται με βάση τα μέτρα απόδοσης. Εφόσον η απόδοση δεν είναι ικανοποιητική, αναζητούνται οι βέλτιστες παράμετροι και το μοντέλο εκπαιδεύεται ξανά. Εφόσον η απόδοση είναι ικανοποιητική, με βάση τις αναζητούνται βέλτιστες παραμέτρους, το μοντέλο αποτιμάται τελικά σε ένα σύνολο δεδομένων που δεν έχει ξαναδεί. Η δυνατότητα του μοντέλου να αποδίδει καλά σε νέα δεδομένα ονομάζεται γενίκευση (generalization) [1].

Η εκμάθηση των παραμέτρων μιας συνάρτησης πρόβλεψης και η δοκιμή της στα ίδια δεδομένα είναι μεθοδολογικό λάθος: ένα μοντέλο που θα επαναλάμβανε τις προβλέψεις στα ίδια δεδομένα θα είχε τέλεια απόδοση, αλλά δεν θα μπορούσε να κάνει σωστές προβλέψεις σε νέα δεδομένα. Αυτή η κατάσταση ονομάζεται υπερταίριασμα ή υπερπροσαρμογή (overfitting), όπου το μοντέλο διαχωρίζει μεν πολύ σωστά τα δεδομένα, αλλά με μεγάλη λεπτομέρεια και θα αποτύχει στην περίπτωση εισαγωγής νέων δεδομένων όπου θα δημιουργηθεί μεγάλο λάθος γενίκευσης. Για την αποφυγή της υπερπροσαρμογής, μια συνηθισμένη πρακτική κατά την εκτέλεση ενός εποπτευόμενου πειράματος ΜΜ το μοντέλο να εκπαιδεύεται σε ένα μέρος του συνόλου δεδομένων και η απόδοσή του να μετράται, δηλαδή να επικυρώνεται κατά κάποιο τρόπο, στο υπόλοιπο του συνόλου δεδομένων. Αφού ρυθμιστούν οι παράμετροι του μοντέλου με βάση τα δεδομένα της

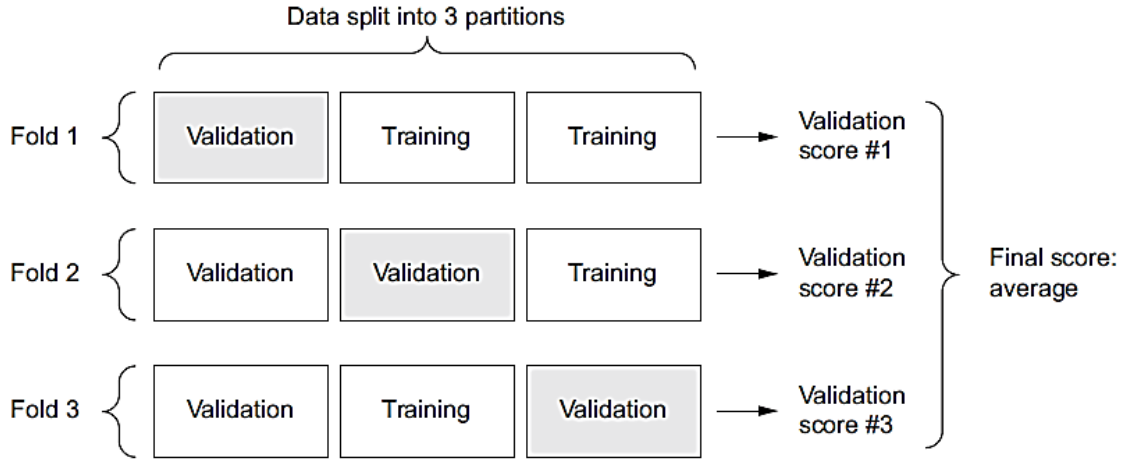
εκπαίδευσης (training data), το μοντέλο αποτιμάται για τα δεδομένα της δοκιμής (test data) [1], [18].

Μια συνηθισμένη τεχνική επικύρωσης (validation) όταν διατίθεται ένα συγκεκριμένο σύνολο δεδομένων για πειραματισμό, είναι το σύνολο των δεδομένων να διαχωρίζεται (split) σε δύο υποσύνολα σε μια αναλογία πχ 80%-20%, όπου το μοντέλο θα εκπαιδευτεί στο 80% των παραδειγμάτων και στο 20% των παραδειγμάτων θα μετρηθεί η απόδοση του μοντέλου [44]. Αυτή η τεχνική ονομάζεται hold-out validation και φαίνεται στο Σχήμα 24.



Σχήμα 24. Η Τεχνική Επικύρωσης Hold-out Validation [44]

Μια άλλη τεχνική επικύρωσης, είναι η επικύρωση k τμημάτων (k-fold validation) που φαίνεται στο Σχήμα 25 για k= 3. Το σύνολο δεδομένων χωρίζεται σε k ίσα τμήματα (folds). Για κάθε τμήμα i το μοντέλο εκπαιδεύεται στα υπόλοιπα k-1 τμήματα. Η τελική απόδοση είναι ο μέσος όρος των k επιδόσεων. Σε αντίθεση με την hold-out validation, δεν απαιτείται διαχωρισμός του συνόλου [44].



Σχήμα 25. Η Τεχνική Επικύρωσης 3k-fold Validation [44]

Άλλες τεχνικές επικύρωσης, αναφέρονται λεπτομερώς στις [18] και [41].

Για την παρούσα εργασία ως τεχνική επικύρωσης επιλέχθηκε η τεχνική διαχωρισμού του αρχικού συνόλου των δεδομένων σε ποσοστά 80%-20%. Σημαντικός παράγοντας κατά τον διαχωρισμό των δεδομένων είναι, να διασφαλίζεται κάθε φορά να λαμβάνεται κάθε φορά το ίδιο υποσύνολο στιγμιότυπων των δεδομένων και για την εκπαίδευσης και για την δοκιμή- ειδικότερα όταν πρόκειται να συγκριθούν διάφορα μοντέλα. Η δυνατότητα αυτή παρέχεται από την βιβλιοθήκη scikit-learn, μέσω της παραμέτρου random_state, η οποία συνήθως λαμβάνει τιμές από 0 έως 42. Επιλέχθηκε η τιμή 42. Μια άλλη σημαντική παράμετρος για τον διαχωρισμό του συνόλου που διαθέτει η scikit-learn, είναι η παράμετρος stratify (διαστρωμάτωση), η οποία δίνει τη δυνατότητα του διαχωρισμού του συνόλου με βάση κάποιο χαρακτηριστικό ή τον στόχο/κλάση. Αυτό πρακτικά σημαίνει ότι, εάν η αναλογία πχ των θετικών δειγμάτων σε μία κλάση είναι 80% και των αρνητικών 20%, τότε το σύνολο δοκιμής θα έχει 80% θετικά δείγματα και 20% αρνητικά. Λόγω του ότι το σύνολο περιέχει 320 στιγμιότυπα και σε ποσοστό 61,54% ανήκουν στην κλάση Positive, ενώ 200 στιγμιότυπα και ποσοστό 38,46% ανήκουν στην κλάση Negative, αποφασίστηκε να δοκιμαστεί και αυτή η παράμετρος με stratify ως προς την κλάση.

4.6 Επιλογή Αλγόριθμων ΜΜ

Για τη δημιουργία του καταλληλότερου προγνωστικού μοντέλου με βάση το σύνολο των δεδομένων που περιέχει κατά βάση τα συμπτώματα της ασθένειας, επιλέχθηκαν οι πιο αντιπροσωπευτικοί από τους δημοφιλέστερους αλγόριθμους δυαδικής ταξινόμησης που διαθέτει η βιβλιοθήκη scikit-learn [29].

Πιο συγκεκριμένα, επιλέχθηκαν για τις προβλέψεις πέντε συνολικά αλγόριθμοι:

1. Η λογιστική παλινδρόμηση (LR)
2. Ο απλοϊκός Bayes (NB)
3. Η μηχανή υποστήριξης διανυσμάτων (SVM)
4. Το πολυεπίπεδο Peceptron και (MLP)
5. Τα τυχαία δάση (RF)

Κάθε αλγόριθμος που διατίθεται από την scikit-learn διατίθεται ως κλάση της python, και υπάρχουν εξ'ορισμού (default) τιμές για τις σχετικές υπερπαραμέτρους του- αναφέρονται απλά ως παράμετροι-, όπου μπορεί κανείς να τις δει συνοπτικά στο documentation του ιστότοπου της βιβλιοθήκης. Στο documentation δίνεται η περιγραφή της

κάθε παραμέτρου και οι δυνατές τιμές που μπορεί να πάρει, εφόσον ο χρήστης επιθυμεί να αλλάξει την παράμετρο. Για κάθε αλγόριθμο και κάθε σχετική του παράμετρο, με τις όλες τις δυνατές επιλογές υπάρχει διαθέσιμος ο κώδικας αναλυτικά στο αποθετήριο του github [45].

Για παράδειγμα, η κλάση της μηχανής υποστήριξης διανυσμάτων για ταξινόμηση (C-Support Vector Classification) έχει τις εξής default παραμέτρους:

class sklearn.svm.**SVC**(*, *C=1.0*, *kernel='rbf'*, *degree=3*, *gamma='scale'*, *coef0=0.0*, *shrinking=True*, *probability=False*, *tol=0.001*, *cache_size=200*, *class_weight=None*, *verbose=False*, *max_iter=1*, *decision_function_shape='ovr'*, *break_ties=False*, *random_state=None*)

και για την παράμετρο kernel, η οποία αναφέρεται στην επιλογή του πυρήνα, δίνονται οι δυνατές τιμές που μπορεί να πάρει, καθώς και η default τιμή:

kernel*{'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'}, default='rbf'*

Στην κλάση της LR, οι παράμετροι και οι default τιμές στην scikit-learn ορίζονται ως εξής:

class sklearn.linear_model.**LogisticRegression**(*penalty='l2'*, *, *dual=False*, *tol=0.0001*, *C=1.0*, *fit_intercept=True*, *intercept_scaling=1*, *class_weight=None*, *random_state=None*, *solver='lbfgs'*, *max_iter=100*, *multi_class='auto'*, *verbose=0*, *warm_start=False*, *n_jobs=None*, *l1_ratio=None*)

Η κλάση GaussianNB εφαρμόζει τον αλγόριθμο Gaussian NB για δυαδική ταξινόμηση. Η πιθανότητα των χαρακτηριστικών θεωρείται ότι είναι Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Οι παράμετροι σ_y και μ_y εκτιμώνται με τη μέγιστη πιθανότητα και οι παράμετροι ορίζονται ως εξής:

class sklearn.naive_bayes.**GaussianNB**(*, *priors=None*, *var_smoothing=1e-09*)

Στην κλάση του MLP, οι παράμετροι και οι default τιμές στην scikit-learn ορίζονται ως εξής:

class sklearn.neural_network.**MLPClassifier**(*hidden_layer_sizes=100*, *activation='relu'*, *, *solver='adam'*, *alpha=0.0001*, *batch_size='auto'*, *learning_rate='constant'*, *learning_rate_init=*

0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e08, n_iter_no_change=10, max_fun=15000)

Στην κλάση του ταξινομητή RF, οι παράμετροι και οι default τιμές στην scikit-learn ορίζονται ως εξής:

class sklearn.ensemble.**RandomForestClassifier**(*n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0,max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None*)

Στην παρούσα εργασία, για την εκπαίδευση των μοντέλων για κάθε έναν από τους αλγόριθμους που επιλέχθηκαν ως αρχικές παράμετροι αφέθηκαν οι εξ'ορισμού παράμετροι της scikit-learn, με το σκεπτικό ότι, αυτές οι παράμετροι μπορούν στη συνέχεια να αλλάξουν εφόσον για κάποιο μοντέλο που δεν αποδίδει καλά, μπορούν να αλλάξουν προκειμένου να βελτιστοποιηθεί.

Αξίζει να σημειωθεί ότι, για πολλούς αλγόριθμους ΜΜ προκειμένου να έχουν καλή απόδοση, είναι πολύ σημαντική η κλιμάκωση των χαρακτηριστικών (feature scaling). Κάποιοι αλγόριθμοι, όπως ο SVM και ο MLP, δεν αποδίδουν καλά εάν οι αριθμητικές τιμές εκτείνονται σε μεγάλο εύρος. Για παράδειγμα, στο σύνολο δεδομένων μας το χαρακτηριστικό ηλικία παίρνει τιμές από 16 έως 90, ενώ όλα τα άλλα χαρακτηριστικά που παίρνουν τις τιμές Yes και No μετατρέπονται αντίστοιχα στις αριθμητικές τιμές 0 και 1.

Ο πιο συνηθισμένος τρόπος για την κλιμάκωση των αριθμητικών τιμών στο ίδιο εύρος είναι η κανονικοποίηση (normalization), όπου οι αριθμητικές τιμές μετατρέπονται στο εύρος [0, 1]. Η scikit-learn παρέχει τη συνάρτηση MinMaxScaler για τον σκοπό αυτό. Ένας άλλος τρόπος είναι η προτυποποίση (standardization) κατά την οποία οι τιμές των χαρακτηριστικών κλιμακώνονται έτσι ώστε να έχουν τις ιδιότητες της πρότυπης κανονικής κατανομής με μέση τιμή μ = 0 και τυπική απόκλιση σ = 1. Η scikit-learn παρέχει τη συνάρτηση StandardScaler για τον σκοπό αυτό.

Συνεπώς, η κλιμάκωση των δεδομένων κρίνεται απαραίτητη, θα εφαρμοστούν και οι δοκιμές γίνονται και με τις δύο τεχνικές κλιμάκωσης.

5 Αποτελέσματα Υλοποίησης και Συζήτηση

Στο κεφάλαιο αυτό παρουσιάζονται τα συγκριτικά αποτελέσματα της υλοποίησης των αλγόριθμων ΜΜ που εφαρμόστηκαν στο σύνολο δεδομένων «Early stage diabetes risk prediction dataset» και η συζήτηση για τα αποτελέσματα.

5.1 Αποτελέσματα Υλοποίησης

Τα πειράματα πραγματοποιήθηκαν με τη δημιουργία μοντέλων με εφαρμογή των αλγόριθμων LR, NB, SVM, MLP και RF για δύο περιπτώσεις κλιμάκωσης δεδομένων για την αριθμητική τιμή του χαρακτηριστικού της ηλικίας, μία με κανονικοποίηση της τιμής και μία με προτυποποίηση της τιμής. Σε κάθε περίπτωση για το διαχωρισμό του συνόλου σε ποσοστό 80-20 εφαρμόστηκε ο διαχωρισμός με τυχαίο σπόρο με τιμή 42, χωρίς διαστρωμάτωση και με διαστρωμάτωση.

5.1.1 Αποτελέσματα

Στον Πίνακα 3 παρατίθενται τα αποτελέσματα όλων των τεσσάρων δοκιμών που πραγματοποιήθηκαν, σύμφωνα με την κλιμάκωση των αριθμητικών χαρακτηριστικών και την εφαρμογή ή όχι της διαστρωμάτωσης κατά τον διαχωρισμό του συνόλου δεδομένων για τη δημιουργία των συνόλων εκπαίδευσης και επικύρωσης. Το καλύτερο μοντέλο για κάθε δοκιμή σημειώνεται με έντονα γράμματα.

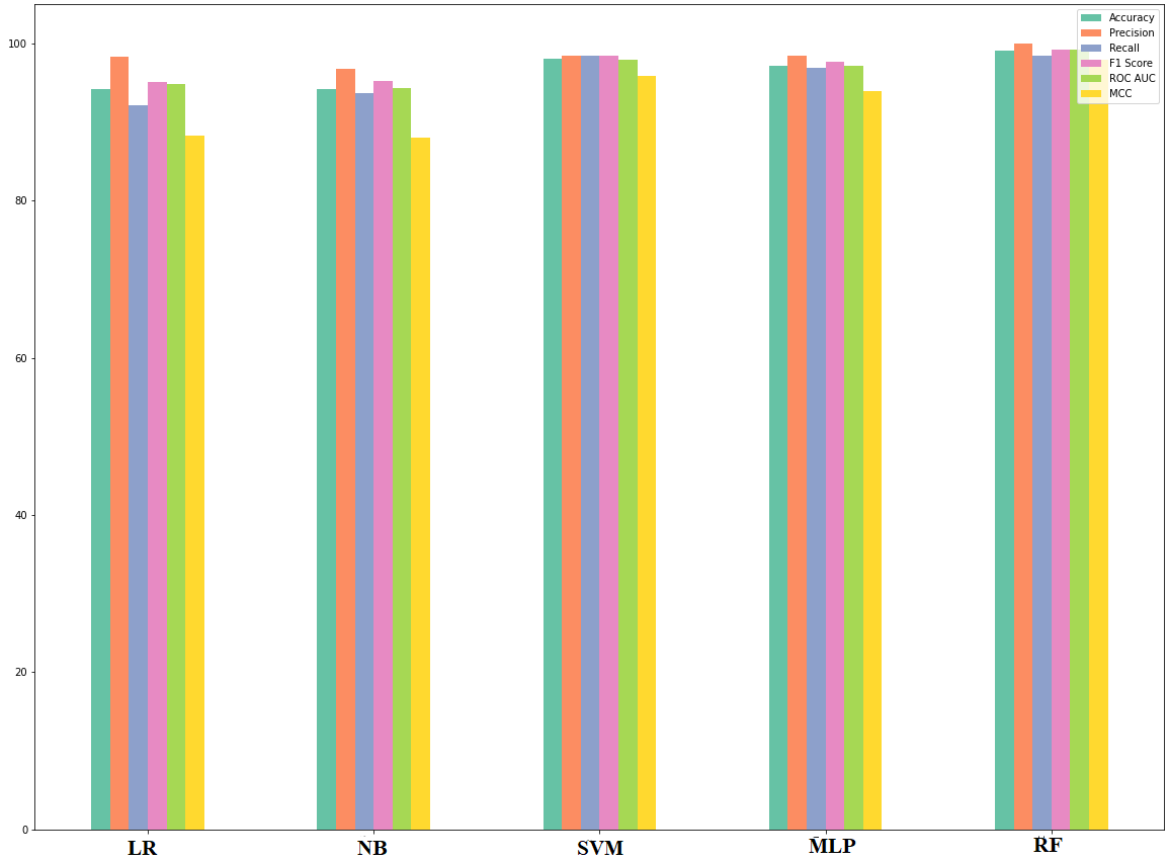
Από ότι παρατηρούμε στον Πίνακα 3, τις καλύτερες επιδόσεις συνολικά εμφανίζουν τα μοντέλα στα οποία εφαρμόστηκε κλιμάκωση με προτυποποίηση και διαστρωμάτωση (ομάδα Δ.). Συνεπώς, θα παρουσιάσουμε περισσότερα στοιχεία για τα μοντέλα αυτής της ομάδας.

Στο Σχήμα 26 δίνεται η γραφική σύγκριση όλων των μοντέλων και στο Σχήμα 27 η γραφική σύγκριση ανά μετρική επίδοσης.

Στο Σχήμα 28 δίνονται οι πίνακες ταξινόμησης και οι μετρικές για κάθε μοντέλο ξεχωριστά.

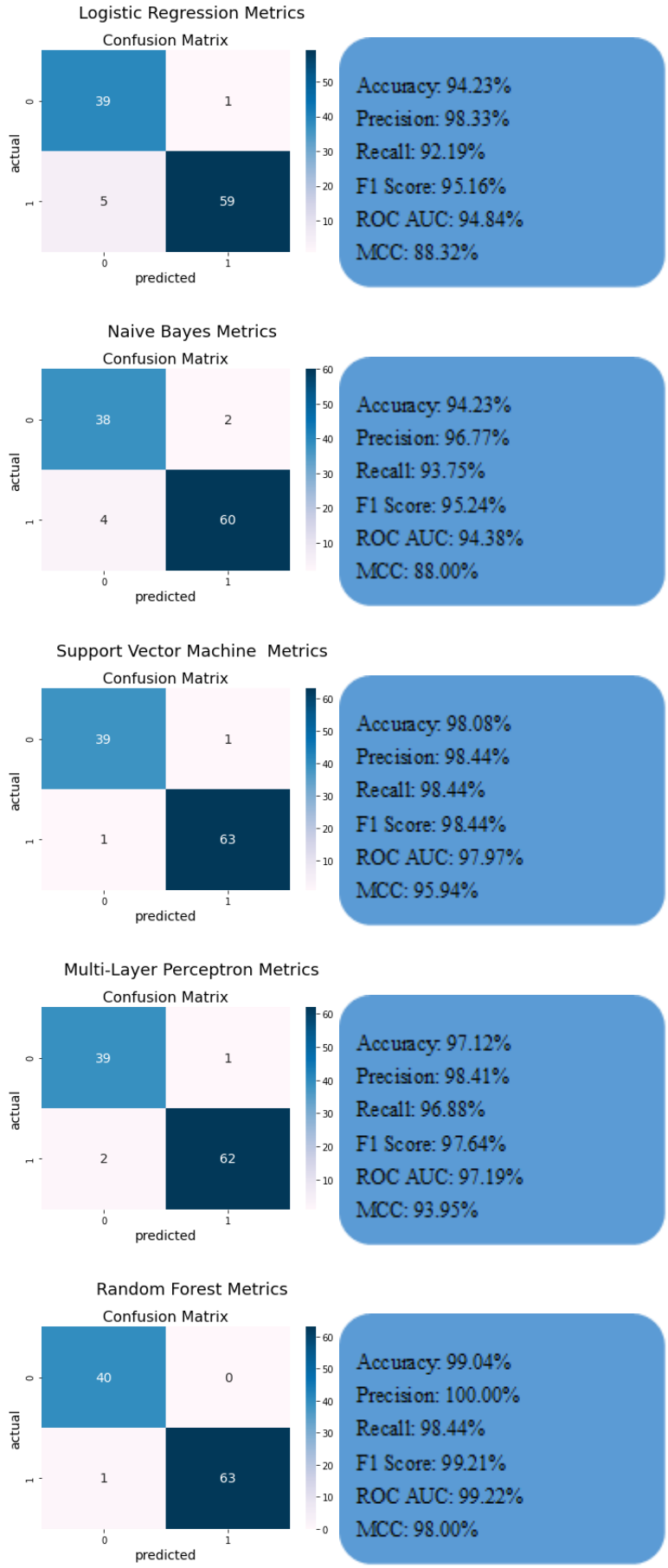
Πίνακας 3. Συγκριτικά Αποτελέσματα Όλων των Δοκιμών

Μοντέλο	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC(%)	MCC (%)
Α. Αποτελέσματα με κανονικοποίηση						
LR	93,27	94,44	95,77	95,10	91,83	84,36
NB	91,35	93,06	94,37	93,71	89,61	79,88
SVM	97,12	98,57	97,18	97,87	97,08	93,42
MLP	94,23	95,77	95,77	95,77	93,34	86,68
RF	**99,04**	**100,00**	**98,59**	**99,29**	**99,30**	**97,82**
Β. Αποτελέσματα με κανονικοποίηση και διαστρωμάτωση						
LR	94,23	98,33	92,19	95,16	94,84	88,32
NB	94,23	96,77	93,75	95,24	94,38	88,00
SVM	98,08	98,44	98,44	98,44	97,97	95,94
MLP	96,15	98,39	95,31	96,83	96,41	92,03
RF	**99,04**	**98,46**	**100,00**	**99,22**	**98,75**	**97,98**
Γ. Αποτελέσματα με προτυποποίηση						
LR	92,31	93,15	95,77	94,44	90,31	82,04
NB	91,35	93,06	94,37	93,71	89,61	79,88
SVM	99,04	98,61	100,00	99,30	98,48	97,79
MLP	97,11	98,57	97,18	97,87	97,08	93,42
RF	**99,04**	**100,00**	**98,59**	**99,29**	**99,30**	**97,82**
Δ. Αποτελέσματα με προτυποποίηση και διαστρωμάτωση						
LR	94,23	98,33	92,19	95,16	94,84	88,32
NB	94,23	96,77	93,75	95,24	94,38	88,00
SVM	98,08	98,44	98,44	98,44	97,97	95,94
MLP	97,11	98,41	96,88	97,64	97,19	93,95
RF	**99,04**	**100,00**	**98,44**	**99,21**	**99,22**	**98,00**



Σχήμα 26. Σύγκριση Μοντέλων

Σχήμα 27. Σύγκριση Μετρικών Μοντέλων



Σχήμα 28. Πίνακες Ταξινόμησης και Μετρικές Μοντέλων

5.2 Συζήτηση

Η αναλυτική παρουσίαση όλων των δοκιμών κρίθηκε απαραίτητη γιατί τα κριτήρια επιλογής της εφαρμογής ενός αλγόριθμου για τη δημιουργία ενός προγνωστικού μοντέλου διαφέρουν ανάλογα με τα μέτρα σύγκρισης που θα επιλέξει κάποιος.

Για παράδειγμα, εάν η επιλογή του μοντέλου γίνει με μέτρο απόδοσης την ακρίβεια, η επιλογή του καταλληλότερου μοντέλου θα είναι άλλη από αυτή που θα γίνει εάν ως μέτρο σύγκρισης χρησιμοποιηθεί η ανάκληση. Ωστόσο, σε γενικές γραμμές το μέτρο σύγκρισης που προτείνεται είναι το F1 αποτέλεσμα. Το ανάλογο ισχύει και για την περίπτωση που επιλεγεί ως μέτρο σύγκρισης κάποιος από τους δύο ποιοτικούς δείκτες, δηλαδή την AUC και τον MCC.

Άλλη παράμετρος επιλογής θα μπορούσε να είναι η απλότητα του μοντέλου και οι διαθέσιμοι υπολογιστικοί πόροι. Βέβαια, το συγκεκριμένο σύνολο δεδομένων που χρησιμοποιήθηκε για τις δοκιμές θεωρείται πάρα πολύ μικρό ως προς τις παρατηρήσεις και η εκτέλεση των υπολογισμών δεν είναι χρονοβόρα ούτε απαιτητικά από άποψη πόρων. Όμως, σε περιπτώσεις του πραγματικού κόσμου τα σύνολα δεδομένων κατά πολύ μεγαλύτερα, αυτοί οι παράγοντες έχουν ιδιαίτερο λόγο. Για παράδειγμα, ο αλγόριθμος NB έχει υπολογιστική πολυπλοκότητα $O(nP)$, ο SVM μεταξύ $O(n^2P)$ και $O(n^3P)$ και ο MLP $O(2^n)$, όπου n ο αριθμός των στιγμιότυπων του συνόλου δεδομένων και P ο αριθμός των χαρακτηριστικών [46].

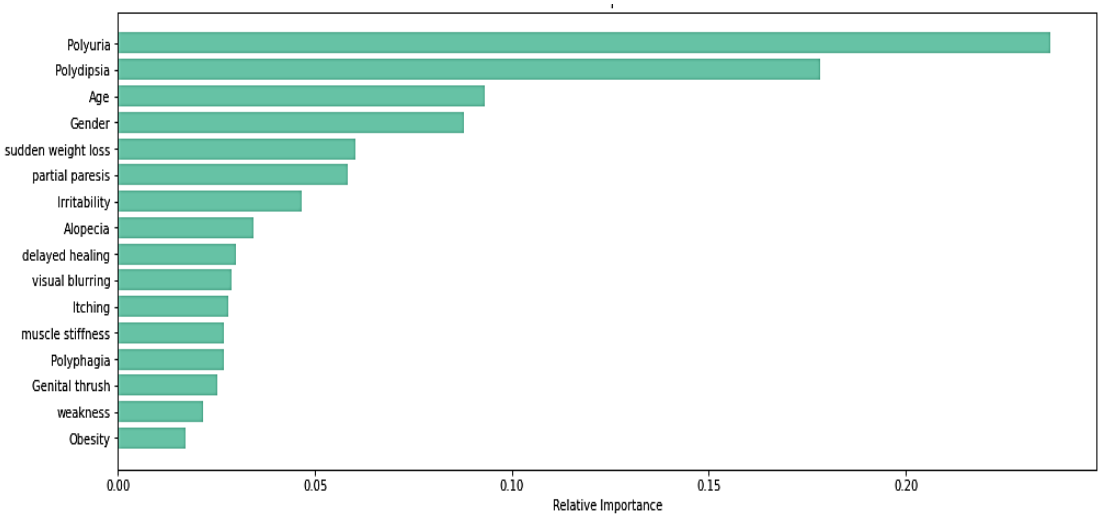
Όσον αφορά τα αποτελέσματα των δοκιμών στο συγκεκριμένο σύνολο δεδομένων, το οποίο είναι πολύ μικρό και προσφέρεται μόνο για πειραματισμούς και με βάση τις παραδοχές που αναφέρθηκαν στο Κεφάλαιο 4 που αφορά τη μεθοδολογία και τη σύγκριση των αποτελεσμάτων που παρατίθεται στον Πίνακα 3 παρατηρούμε τα εξής:

- Τα μοντέλα που προέκυψαν με εφαρμογή διαστρωμάτωσης έχουν καλύτερες επιδόσεις από αυτά χωρίς διαστρωμάτωση και μάλιστα με τις ίδιες περίπου επιδόσεις ανεξάρτητα από την μέθοδο κλιμάκωσης.
- Τα μοντέλα που δημιουργήθηκαν με την εφαρμογή των αλγορίθμων NB και LR, εμφανίζουν σε κάθε περίπτωση τις χειρότερες μετρικές, συνεπώς θεωρούμε ότι δεν απαιτείται περαιτέρω διερεύνηση για την βελτιστοποίησή τους.
- Καλά αποτελέσματα εμφανίζει ο αλγόριθμος MLP, ωστόσο θεωρούμε ότι δεν απαιτείται περαιτέρω διερεύνηση για τις βέλτιστες παραμέτρους του, γιατί η

scikit-learn δεν διαθέτει τις δυνατότητες επιλογής παραμέτρων που διαθέτει, για παράδειγμα η βιβλιοθήκη keras[15], η οποία είναι αποκλειστικά για ΤΝΔ.

- Οι αλγόριθμοι SVM και RF εμφανίζουν τα καλύτερα αποτελέσματα σε όλες τις περιπτώσεις σε σχέση με τους υπόλοιπους τρεις αλγόριθμους.
- Ο αλγόριθμος RF με τις εξ'ορισμού παραμέτρους της scikit-learn σε όλες τις δοκιμές εμφανίζει τα ίδια άριστα αποτελέσματα για όλες τις μετρικές.

Με βάση τα παραπάνω, ο καλύτερος αλγόριθμος για τη δημιουργία προγνωστικού μοντέλου για το συγκεκριμένο σύνολο δεδομένων εμφανίζει ο αλγόριθμος RF. Σε γενικές γραμμές είναι γνωστό ότι ο RF λειτουργεί πολύ καλά για σύνολα δεδομένων με πολλά χαρακτηριστικά και μάλιστα για σύνολα όπου τα δεδομένα και οι κλάσεις δεν είναι ομοιόμορφα κατανεμημένες [46]. Δεν είναι τυχαίο το γεγονός ότι, ο αλγόριθμος RF χρησιμοποιείται συχνά ως μέθοδος εύρεσης των σπουδαιότερων χαρακτηριστικών ενός συνόλου δεδομένων. Στο Σχήμα παρατίθεται η σπουδαιότητα των χαρακτηριστικών για την πρόβλεψη της κλάσης σύμφωνα με το μοντέλο που δημιουργήθηκε από τον RF. Ανατρέχοντας στην EDA στο Κεφάλαιο 4, παρατηρούμε ότι, ο αλγόριθμος αξιολογεί ως τα τρία σπουδαιότερα χαρακτηριστικά την πολυουρία, την πολυδιψία και την ξαφνική απώλεια βάρους, λαμβάνοντας παράλληλα υπόψη και τον παράγοντα της ηλικίας. Για το χαρακτηριστικό του φύλου, το οποίο ο αλγόριθμος το αξιολογεί ως σπουδαίο, σχετίζεται άμεσα με την αρχική μας παρατήρηση ότι, θα έπρεπε να υπάρχει γενικότερα διαχωρισμός σε διαφορετικά σύνολα δεδομένων με βάση το φύλο.



Σχήμα 29. Η Σπουδαιότητα των Χαρακτηριστικών Σύμφωνα με τον RF

[15] https://keras.io/

Με δεδομένο ότι, για κάθε αλγόριθμο και για κάθε συγκεκριμένο σύνολο δεδομένων εφόσον το μοντέλο που δημιουργείται εμφανίζει σχετικά καλές αποδόσεις υπάρχει περίπτωση το μοντέλο να βελτιστοποιηθεί με αναζήτηση καλύτερων παραμέτρων [18] και εφόσον ο SVM εμφανίζει τα αμέσως καλύτερα αποτελέσματα στις περισσότερες δοκιμές, αποφασίστηκε να χρησιμοποιηθεί η δυνατότητα που δίνει η scikit-learn για την αναζήτηση τυχόν καλύτερων παραμέτρων κατά την εκπαίδευση του μοντέλου παρέχοντας την κλάση GridSearchCV και παρουσιάζουμε τη δοκιμή στην επόμενη παράγραφο. Άλλωστε, κατά κοινή ομολογία, η ΜΜ είναι μία σειρά συνεχών πειραματισμών και προσπαθειών για την εύρεση του καλύτερου προγνωστικού μοντέλου.

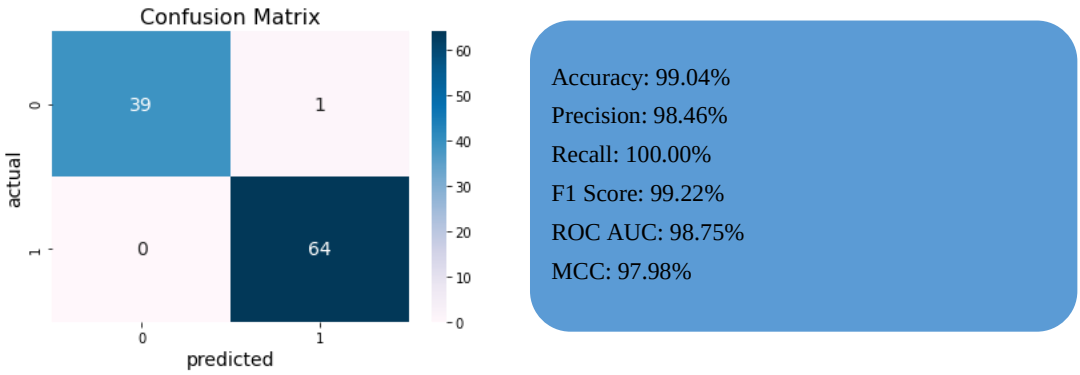
5.3 Βελτιστοποίηση Μοντέλου SVM

Οι παράμετροι με την μεγαλύτερη σπουδαιότητα κατά την εφαρμογή ενός αλγόριθμου SVM είναι ο πυρήνας, η παράμετρος C που αντιπροσωπεύει τον όρο ποινής και ενεργεί ως παράμετρος ομαλοποίησης και η παράμετρος gamma του πυρήνα που ορίζει πόσο μεγάλη επιρροή έχει ένα παράδειγμα της εκπαίδευσης[16]. Όσο μεγαλύτερο είναι το gamma τόσο περισσότερο επηρεάζονται τα πιο κοντινά παραδείγματα. Όσο μεγαλύτερο είναι το C, τόσο περισσότερο βοηθά στη σωστή ταξινόμηση όλων των παραδειγμάτων.

Η GridSearchCV εκτιμά τις βέλτιστες παραμέτρους από ένα σετ παραμέτρων συνδυάζοντας αυτές μεταξύ τους κατά την εκπαίδευση με διασταυρούμενη επικύρωση. Η εφαρμογή της GridSearchCV για την εύρεση του βέλτιστου πυρήνα, του C και του gamma, για κλιμάκωση με κανονικοποίση είχε ως αποτέλεσμα πυρήνα *Radial Basis Function* (RBF), C= 10 και gamma= 1.

Εφαρμόζοντας αυτές τις παραμέτρους στην εκπαίδευση με hold-out validation και κανονικοποίηση και διαστρωμάτωση στα δεδομένα, τα αποτελέσματα που προέκυψαν φαίνονται στο Σχήμα 30 και είναι ακριβώς ίδια με αυτά του μοντέλου που προέκυψε από την εφαρμογή του RF (Σχήμα 28).

[16] https://scikit-learn.org/stable/modules/svm.html



Σχήμα 30. Πίνακας Ταξινόμησης και Μετρικές Βελτιστοποιημένου SVM Μοντέλου

Η εφαρμογή της GridSearchCV για την εύρεση του βέλτιστου πυρήνα, του C και του gamma, για κλιμάκωση με προτυποποίηση πάλι είχε ως αποτέλεσμα πυρήνα *Radial Basis Function* (RBF), C= 10 και gamma= 1. Ωστόσο, εφαρμόζοντας αυτές τις παραμέτρους με εκπαίδευση hold-out validation, τα αποτελέσματα που προέκυψαν δεν ήταν καλύτερα από αυτά του μοντέλου με τις default παραμέτρους. Έτσι, ως συμπέρασμα προέκυψε ότι, δεν βελτιώνεται περαιτέρω το μοντέλο με την αλλαγή των παραμέτρων για εκπαίδευση με hold-out validation.

6 Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Στο κεφάλαιο αυτό παρουσιάζονται τα συμπεράσματα από την παρούσα εργασία και οι μελλοντικές κατευθύνσεις.

6.1.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία διερευνήθηκε το θέμα της εξόρυξης δεδομένων με τη βοήθεια μεθόδων ΜΜ για την πρόωρη πρόβλεψη της ασθένειας του σακχαρώδους διαβήτη με βάση τα συμπτώματά της, το οποίο αποτελεί ένα πρόβλημα δυαδικής ταξινόμησης.

Αρχικά, παρουσιάστηκε το απαιτούμενο θεωρητικό υπόβαθρο για την ασθένεια και την ΜΜ. Το θέμα της πρόωρης πρόβλεψης της ασθένειας έχει απασχολήσει την ακαδημαϊκή κοινότητα και για τον σκοπό αυτό έγινε αναφορά στις σχετικές εργασίες.

Στη συνέχεια, παρουσιάστηκε η μεθοδολογία για την υλοποίηση μοντέλων ΜΜ με τη βοήθεια της βιβλιοθήκης scikit-learn που βασίζεται στη γλώσσα προγραμματισμού Python, με σκοπό την εύρεση του καταλληλότερου μοντέλου για το σύνολο δεδομένων «Early stage diabetes risk prediction dataset» που διατίθεται στο αποθετήριο της ΜΜ του UCI. Πραγματοποιήθηκε εξερεύνηση των δεδομένων και επιλέχθηκαν τα μέτρα απόδοσης, οι τεχνικές αποτίμησης και οι αλγόριθμοι ΜΜ. Πιο συγκεκριμένα, επιλέχθηκαν για τις προβλέψεις πέντε συνολικά αλγόριθμοι: η λογιστική παλινδρόμηση, ο απλοϊκός Bayes, η μηχανή υποστήριξης διανυσμάτων, το πολυεπίπεδο Peceptron και το τυχαίο δάσος.

Τα μοντέλα δημιουργήθηκαν με την εφαρμογή των σχετικών επιλογών, με σκοπό να προσδιοριστεί ο αλγόριθμος που δίνει το καλύτερο προγνωστικό μοντέλο για το συγκεκριμένο σύνολο δεδομένων. Τα καλύτερα αποτελέσματα για όλα τα μοντέλα έδωσε η επιλογή της κλιμάκωσης του χαρακτηριστικού της ηλικίας με προτυποποίηση και με την εφαρμογή διαστρωμάτωσης ως προς το χαρακτηριστικό της κλάσης.

Το μοντέλο με τις καλύτερες μετρικές απόδοσης και τους καλύτερους ποιοτικούς δείκτες σε όλες τις δοκιμές για το συγκεκριμένο σύνολο δεδομένων, είναι το μοντέλο που δημιουργήθηκε με τον αλγόριθμο συλλογικής μάθησης RF με κλιμάκωση αριθμητικών δεδομένων με προτυποποίση και διαστρωμάτωση. Το μοντέλο έχει ορθότητα 99,04%, ακρίβεια 100%, ανάκληση 98,4% και F1 αποτέλεσμα 99,21%. Όσον αφορά τους ποιοτικούς δείκτες έχουν AUC 99,22% και συντελεστή συσχέτισης Matthews 98%.

Με την βελτιστοποίηση των παραμέτρων, τα μοντέλα του SVM και του RF μπορούν να συγκριθούν για την περίπτωση της κλιμάκωσης με κανονικοποίηση και διαστρωμάτωση, όπου εμφανίζουν ακριβώς τις ίδιες μετρικές: ορθότητα 99,04%, ακρίβεια 98,46%, ανάκληση 100% και F1 αποτέλεσμα 99,22%. Όσον αφορά τους ποιοτικούς δείκτες έχουν AUC 98,75% και συντελεστή συσχέτισης Matthews 97,98%.

Τα αποτελέσματα της παρούσας εργασίας, δεν μπορούν να συγκριθούν απευθείας με τα αποτελέσματα των υπόλοιπων σχετικών εργασιών, γιατί στην πλειοψηφία αυτών δεν αναφέρονται πλήρη στοιχεία για το περιβάλλον υλοποίησης και τις παραμέτρους κάθε μοντέλου. Ωστόσο, θα μπορούσαμε να αναφέρουμε ότι, σε καμία από τις σχετικές εργασίες όπου ακολουθείται η τεχνική του διαχωρισμού 80-20 δεν αναφέρονται καλύτερα αποτελέσματα για τον SVM και τον RF σε σχέση με την παρούσα εργασία.

Συνεπώς, θεωρούμε ότι, η παρούσα διπλωματική εργασία πέτυχε τους στόχους της.

6.1.2 Μελλοντικές κατευθύνσεις

Η διερεύνηση τεχνικών εξόρυξης δεδομένων με μεθόδους ΜΜ είναι ένα επιστημονικό πεδίο, το οποίο παρουσιάζει ιδιαίτερο ενδιαφέρον και εξελίσσεται συνεχώς. Το πρόβλημα της πρόωρης πρόγνωσης του διαβήτη θα μπορούσε να διερευνηθεί στο μέλλον περισσότερο:

- Με εφαρμογή της διασταυρούμενης επικύρωσης και σύγκριση σε σχέση με την τεχνική διαχωρισμού που εφαρμόστηκε στην παρούσα εργασία.
- Με πειραματισμούς στα MLP, με δεδομένο ότι η ΒΜ κυριαρχεί πλέον στην ΜΜ και ότι υπάρχουν ειδικές βιβλιοθήκες για ΒΜ, οι οποίες παρέχουν περισσότερες δυνατότητες πειραματισμών για την εύρεση του καταλληλότερου μοντέλου, όπως για παράδειγμα η βιβλιοθήκη της keras.
- Με boosting αλγόριθμους συλλογικής ΜΜ.

Βιβλιογραφία

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[2] A. Smiti, "When machine learning meets medical world: Current status and future challenges," *Comput. Sci. Rev.*, vol. 37, p. 100280, 2020, doi: https://doi.org/10.1016/j.cosrev.2020.100280.

[3] World Health Organization, "Diabetes," 2021. https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed Apr. 25, 2021).

[4] Π. Μήτρου, "Νεότερα δεδομένα στα μεταβολικά νοσήματα: Σακχαρώδης Διαβήτης," *Επιστημονικά Χρονικά*, vol. 22, no. S1, pp. 83–91, 2017.

[5] World Health Organization, "WHO reveals leading causes of death and disability worldwide: 2000-2019," 2020. https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019 (accessed Apr. 25, 2021).

[6] C. V Loupa, S. Kalantzi, and A. Maris, "Trends in epidemiology of diabetes mellitus in Greece. Review of the major epidemiological studies," 2017.

[7] World Health Organization, "Coronavirus disease (COVID-19) pandemic," 2021. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/ (accessed Apr. 25, 2020).

[8] A. Kumar, A. Arora, P. Sharma, and S. Anil, "Is diabetes mellitus associated with mortality and severity of COVID- 19? A meta-analysis," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. January, pp. 535–545, 2020, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1871402120301090?via%3Dihub.

[9] "Early stage diabetes risk prediction dataset," 2020. https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. (accessed Apr. 01, 2021).

[10] D. Dua and C. Graff, "UC Irvine Machine Learning Repository." 2021, [Online]. Available: http://archive.ics.uci.edu/ml.

[11] Κ. Κοντοάγγελος *et al.*, "Σακχαρώδης διαβήτης και ψυχοπαθολογία," *Αρχεία Ελληνικής Ιατρικής*, vol. 30, no. 6, pp. 688–699, 2013.

[12] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, *Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques*, vol. 992. Springer

Singapore, 2020.

[13] SOS Ιατροί, "ΣΑΚΧΑΡΩΔΗΣ ΔΙΑΒΗΤΗΣ," 2020. https://www.sosiatroi.gr/iatrikes-symvoules/pathologika/sakxarodis-diabitis/ (accessed Apr. 25, 2021).

[14] Α. Γεωργούλη, *Τεχνητή Νοημοσύνη. [ηλεκτρ. βιβλ.]*. Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[15] Javatpoint, "Machine Learning Tutorial." 2021, [Online]. Available: https://www.javatpoint.com/machine-learning.

[16] R. van Loon, "Machine Learning Explained: Understanding Supervised, Unsupervised, and Reinforcement Learning," 2018. https://www.linkedin.com/pulse/machine-learning-explained-understanding-supervised-ronald-van-loon (accessed Apr. 25, 2020).

[17] Javatpoint, "Regression vs. Classification in Machine Learning." 2021, [Online]. Available: https://www.javatpoint.com/machine-learning.

[18] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[19] 7wData, "10 Companies Using Machine Learning in Cool Ways," 2021. https://www.7wdata.be/big-data/10-companies-using-machine-learning-in-cool-ways-2/ (accessed Apr. 25, 2021).

[20] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychol. Rev.*, vol. 65, no. 6, p. 386, 1958.

[21] D. R. Cox, "The regression analysis of binary sequences," *J. R. Stat. Soc. Ser. B*, vol. 20, no. 2, pp. 215–232, 1958.

[22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, doi: 10.1038/323533a0.

[23] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/BF00116251.

[24] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[25] H. Zhang and J. Su, "Naive bayesian classifiers for ranking," in *European conference on machine learning*, 2004, pp. 501–512.

[26] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, vol. 1, pp. 278–282 vol.1, doi:

10.1109/ICDAR.1995.598994.

[27] A. Amidi and S. Amidi, “Supervised Learning cheatsheet,” 2019. https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning# (accessed Apr. 25, 2021).

[28] M. Mishra, “Artificial Neural Network(The basic idea behind machine’s brain),” 2018. https://analyticsmitra.wordpress.com/2018/02/05/artificial-neural-network-the-basic-idea-behind-machines-brain/ (accessed Apr. 30, 2021).

[29] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[30] A. Amidi and S. Amidi, “Deep Learning cheatsheet,” 2019. https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-deep-learning (accessed Apr. 25, 2021).

[31] Wikipedia the free encyclopedia, “Bootstrap aggregating,” 2021. https://en.wikipedia.org/wiki/Bootstrap_aggregating (accessed Apr. 25, 2021).

[32] J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan, “Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review,” *J. King Saud Univ. - Comput. Inf. Sci.*, 2020, doi: https://doi.org/10.1016/j.jksuci.2020.06.013.

[33] V. Jaiswal, A. Negi, and T. Pal, “A review on current advances in machine learning based diabetes prediction,” *Prim. Care Diabetes*, vol. 15, no. 3, pp. 435–443, 2021, doi: https://doi.org/10.1016/j.pcd.2021.02.005.

[34] Kaggle, “Pima Indians Diabetes. Database Predict the onset of diabetes based on diagnostic measures.” https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed Apr. 30, 2021).

[35] U. Das, A. Yakin Srizon, M. Ansarul Islam, D. Sikder Tonmoy, and M. Al Mehedi Hasan, “Prognostic Biomarkers Identification for Diabetes Prediction by Utilizing Machine Learning Classifiers,” in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, 2020, pp. 1–6, doi: 10.1109/STI50764.2020.9350498.

[36] R. P. C. Gamara, A. A. Bandala, P. J. M. Loresco, and R. R. P. Vicerra, “Early Stage Diabetes Likelihood Prediction using Artificial Neural Networks,” in *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 2020, pp. 1–5, doi: 10.1109/HNICEM51456.2020.9400075.

[37] Nurjahan, M. A. T. Rony, M. S. Satu, and M. Whaiduzzaman, “Mining Significant Features of Diabetes through Employing Various Classification Methods,” in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 240–244, doi: 10.1109/ICICT4SD50815.2021.9397006.

[38] L. Chaves and G. Marques, “Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study,” *Appl. Sci.*, vol. 11, no. 5, 2021, doi: 10.3390/app11052218.

[39] P. Virtanen *et al.*, “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nat. Methods*, vol. 17, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.

[40] W.-M. Lee, *Python machine learning*. John Wiley \& Sons, 2019.

[41] A. Burkov, *The Hundred-Page Machine Learning Book*. Andriy Burkov (January 12, 2019), 2019.

[42] Wikipedia the free encyclopedia, “Confusion matrix.” 2021, [Online]. Available: https://en.wikipedia.org/wiki/Confusion_matrix.

[43] Wikipedia the free encyclopedia, “Matthews correlation coefficient.” 2021, [Online]. Available: https://en.wikipedia.org/wiki/Matthews_correlation_coefficient.

[44] F. Chollet and others, *Deep learning with Python*, vol. 361. Manning New York, 2018.

[45] C. Lorentzen, “scikit-learn: machine learning in Python.” 2021, [Online]. Available: https://github.com/scikit-learn/scikit-learn.

[46] F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, “Machine Learning in IoT Security: Current Solutions and Future Challenges,” *IEEE Commun. Surv. Tutorials*, vol. 22, no. 3, pp. 1686–1721, 2020, doi: 10.1109/COMST.2020.2986444.