



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΗΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ &
ΠΛΗΡΟΦΟΡΙΚΗΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Ανάπτυξη Μοντέλων Μηχανικής Μάθησης στο Διαβήτη»

Μαρίνα Ακαρέπη

Επιβλέπων: Θεμιστοκλής Έξαρχος, Αναπληρωτής Καθηγητής

ΚΕΡΚΥΡΑ, 2023

Abstract

This thesis focuses on the construction of two machine learning models for predicting the hospital readmission rate and length of stay of diabetic patients. The dataset used is the "Diabetes 130-Hospitals Dataset" and includes clinical information on more than 100,000 diabetic patients from 130 hospitals. After preprocessing the data, the optimal features were selected using the WSAE and CSAE methods for the classification and regression problem respectively. Eight algorithms were used for the classification problem with Voting (NB, LR, DT), Bagging (LR) and Stacking (LR) being the 3 best algorithms where the first algorithm performed well for the class "<30", the second algorithm for the class "No" and the third algorithm for the class ">30" while all of them had similar average performance. For the regression problem, 3 different algorithms were used, where the K-Nearest Neighbors Regressor (K-NNR) algorithm showed the best performance with the lowest mean absolute error (MAE) and root mean square error (RMSE) values. Furthermore, a web application was implemented for the use of the above models by diabetic patients and health professionals. In conclusion, it turns out that the use of machine learning turned out to be quite beneficial as it not only helps in the personalized and early care of diabetic patients by preventing readmissions or death rates due to delayed care but also helps in the better allocation of financial resources of hospitals.

Περίληψη

Αυτή η πτυχιακή εργασία επικεντρώνεται στην κατασκευή δύο μοντέλων μηχανικής μάθησης για την πρόβλεψη της επανεισαγωγής διαβητικών ασθενών στο νοσοκομείο και τη διάρκεια της παραμονής τους. Το σύνολο δεδομένων που χρησιμοποιήθηκε είναι το "Diabetes 130-Hospitals Dataset" και περιλαμβάνει κλινικές πληροφορίες για περισσότερους από 100.000 διαβητικούς ασθενείς από 130 νοσοκομεία. Κατόπιν τη προεπεξεργασίας των δεδομένων, έγινε επιλογή των βέλτιστων χαρακτηριστικών, μετά από σύγκριση, με τις μεθόδους WSAE και CSAE για το πρόβλημα της ταξινόμησης και της παλινδρόμησης αντίστοιχα. Χρησιμοποιήθηκαν 8 αλγόριθμοι για το πρόβλημα ταξινόμησης με τους Voting (NB, LR, DT), Bagging (LR) και Stacking (LR) να είναι οι 3 καλύτεροι όπου ο πρώτος είχε καλή απόδοση για την κλάση "<30", ο δεύτερος για την κλάση "No" και ο τρίτος αλγόριθμος για την κλάση ">30" ενώ όλοι είχαν παρόμοια μέση απόδοση. Για το πρόβλημα παλινδρόμησης χρησιμοποιήθηκαν 3 διαφορετικοί αλγόριθμοι, όπου ο αλγόριθμος K-Nearest Neighbors Regressor (K-NNR) παρουσίασε την καλύτερη απόδοση με τις χαμηλότερες τιμές μέσου απόλυτου σφάλματος (MAE) και μέσου τετραγωνικού σφάλματος (RMSE). Ακόμη, έγινε υλοποίηση διαδικτυακής εφαρμογής για την χρήση των παραπάνω μοντέλων από τους διαβητικούς ασθενείς αλλά και τους επαγγελματίες υγείας. Γενικότερα, η χρήση μηχανικής μάθησης αποδείχθηκε πως είναι αρκετά ωφέλιμη καθώς όχι μόνο βοηθά στην εξατομικευμένη και πρόωρη φροντίδα των διαβητικών ασθενών προλαμβάνοντας τις επανεισαγωγές ή και τα ποσοστά θανάτων λόγω αργοπορημένης φροντίδας αλλά βοηθά και στην ομαλότερη κατανομή των οικονομικών πόρων των νοσοκομείων.

Περιεχόμενα

ΚΕΦΆΛΑΙΟ 1° - Εισαγωγή	6
1.1.Εισαγωγή	6
1.2. Ορισμός του Σακχαρώδη Διαβήτη	7
1.3. Ταξινόμηση	7
1.4. Συμπτώματα και Διάγνωση	8
1.5. Επιδημιολογία	9
1.6. Αντιμετώπιση του Σακχαρώδη Διαβήτη	11
1.7. Η Αναγκαιότητα εφαρμογής μηχανικής μάθησης στο πρόβλημα.	12
ΚΕΦΆΛΑΙΟ 2° - Βιβλιογραφική Ανασκόπηση	14
2.1. Karlsburger Diabetes-Management System (KADIS)	14
2.2. Πρόβλεψη διαβήτη με ταξινομητές μηχανικής μάθησης και ρύθμιση των υπερπαραμέτρων τους.....	15
2.3. Σύστημα GlucoTab	16
2.4. Εφαρμογή Diabits	17
2.5. Ανασκόπηση στους παράγοντες κινδύνου επανεισαγωγής νοσηλευόμενων διαβητικών	19
2.6. Παράγοντες πρόβλεψης των επανεισαγωγών και του LoS για ασθενείς με σακχαρώδη διαβήτη.....	20
2.7. Μάθηση για Κατηγοριοποίηση Κινδύνων στον Διαβήτη Τύπου 2	21
2.8. Πρόβλεψη του διαβήτη τύπου 2 με αλγόριθμο μηχανικής μάθησης.....	23
2.9. Πρόβλεψη των επανεισαγωγών διαβητικών ασθενών ανεξάρτητα από την HbA1c... ..	24
2.10. Ένα οντολογικά βασισμένο και ερμηνεύσιμα ασαφές DSS για τη διάγνωση του διαβήτη.....	25
2.11. Μοντέλο ML για την πρόβλεψη του LoS και της θνησιμότητας των νοσηλευόμενων ασθενών με διαβήτη και υπέρταση	27
2.12. Αξιολόγηση υπηρεσιών φροντίδας διαβητικών, της διάρκειας παραμονής και του 30ήμερου ποσοστού επανεισαγωγής	28
2.13. Πρόβλεψη της διάρκειας παραμονής στο νοσοκομείο χρησιμοποιώντας μια γαλλική ιατρική βάση δεδομένων	29
ΚΕΦΆΛΑΙΟ 3° - Περιγραφή Προβλήματος και Δεδομένων.....	31
3.1. Εισαγωγή	31
3.3. Περιγραφή προβλήματος και δεδομένων.....	31
ΚΕΦΆΛΑΙΟ 4° - Μεθοδολογία	38
4.1. Προεπεξεργασία δεδομένων	38
4.1.1 Ταξινόμηση.....	39
4.1.2 Παλινδρόμηση	41
4.2. Επιλογή χαρακτηριστικών	42

4.2.1. Ταξινόμηση.....	42
4.2.2. Παλινδρόμηση.....	46
4.3. Αξιολόγηση επιδόσεων	47
4.3.1. Ταξινόμηση.....	47
4.3.2. Παλινδρόμηση	48
4.4. Δημιουργία μοντέλου – Επιλογή Αλγορίθμων και συντονισμός παραμέτρων.....	50
4.4.1. Naïve Bayes.....	50
4.4.2. Decision Tree.....	51
4.4.3. Logistic Regression	52
4.4.4. Random Forest.....	53
4.4.5. Bagging	54
4.4.6. Stacking	55
4.4.7. AdaBoostM1.....	55
4.4.8. Linear Regression.....	56
4.4.9. K-Nearest Neighbor Regressor.....	56
4.4.10. Random Forest Regressor.....	57
ΚΕΦΆΛΑΙΟ 5° - Πειράματα και Αποτελέσματα	59
5.1. Πρόβλημα Ταξινόμησης	59
5.2. Πρόβλημα Παλινδρόμησης	65
ΚΕΦΆΛΑΙΟ 6° – Υλοποίηση Διαδικτυακής Εφαρμογής	68
6.1. Τεχνολογική Προσέγγιση.....	68
6.2. Παρουσίαση εφαρμογής.....	68
ΚΕΦΆΛΑΙΟ 7° - Συμπεράσματα	79
ΑΝΑΦΟΡΕΣ	81

ΚΕΦΑΛΑΙΟ 1^ο - Εισαγωγή

1.1. Εισαγωγή

Η μηχανική μάθηση είναι ένας κλάδος της τεχνητής νοημοσύνης που αφορά τον σχεδιασμό, την ανάπτυξη και την εφαρμογή αλγορίθμων που επιτρέπουν σε υπολογιστές να μαθαίνουν από τα δεδομένα και να προβλέπουν τα μελλοντικά αποτελέσματα χωρίς να είναι προγραμματισμένοι ειδικά γι' αυτά. Αυτή η επιστημονική προσέγγιση έχει επανασχεδιάσει τον τρόπο με τον οποίο αντιμετωπίζουμε προβλήματα πρόβλεψης και ανάλυσης δεδομένων, εισάγοντας τη δυνατότητα του υπολογιστή να εξάγει αυτόματα συμπεράσματα από πληθώρα πληροφοριών. Στον τομέα της υγείας, η μηχανική μάθηση έχει αναδειχθεί ως ένα ισχυρό εργαλείο για την κατανόηση και την αντιμετώπιση πολλών ασθενειών. Συγκεκριμένα, στον τομέα του Σακχαρώδους Διαβήτη, παρέχει τη δυνατότητα ανάπτυξης προγνωστικών μοντέλων που μπορούν να προβλέψουν την εξέλιξη της νόσου και να προσφέρουν προσαρμοσμένες θεραπείες και όχι μόνο.

Ο Σακχαρώδης διαβήτης, ως μία από τις πιο διαδεδομένες χρόνιες ασθένειες που επηρεάζει εκατομμύρια ανθρώπους παγκοσμίως, αντιμετωπίζεται με σοβαρότητα από τον επιστημονικό και ιατρικό χώρο καθώς αποτελεί ένα από τα πιο κοινά χρόνια νοσήματα παγκοσμίως και απαιτεί συνεχή παρακολούθηση και φροντίδα. Η μηχανική μάθηση μπορεί να εκμεταλλευτεί τα δεδομένα από ασθενείς, ιατρούς και νοσοκομεία για τη δημιουργία προγνωστικών μοντέλων που εντοπίζουν την πιθανότητα επιπλοκών, όπως υπογλυκαιμίες ή υπεργλυκαιμίες, καθώς και τη διάρκεια της νοσηλείας ή των επανεισαγωγών. Οι επανεισαγωγές διαβητικών ασθενών στο νοσοκομείο αποτελούν μία πραγματική πρόκληση για τον τομέα της υγείας, καθώς επιφέρουν τόσο αυξημένο κόστος όσο και ανθρώπινο πόνο. Η πρόβλεψη αυτών των επιστροφών και η βελτιστοποίηση της διάρκειας παραμονής των ασθενών στο νοσοκομείο αποτελούν προτεραιότητα για τη βελτίωση της ποιότητας της φροντίδας τους.

Σε αυτό το κεφάλαιο, θα εξετάσουμε τι ακριβώς είναι ο Σακχαρώδης Διαβήτης, σε ποιες κατηγορίες χωρίζεται τονίζοντας με παρουσίαση επιδημιολογικών στοιχείων την σημαντικότητά του ως ένα από τα πιο σοβαρά χρονικά νοσήματα της εποχής μας. Επιπλέον θα αναλύσουμε τη σημασία της μηχανικής μάθησης για την πρόβλεψη των πιθανών επανεισαγωγών και της διάρκειας παραμονής στο νοσοκομείο των διαβητικών

ασθενών ως παράγοντες για την αποτελεσματική και πρόωρη αντιμετώπιση του διαβήτη.

1.2. Ορισμός του Σακχαρώδη Διαβήτη

Ο σακχαρώδης διαβήτης είναι μια χρόνια μεταβολική διαταραχή γνωστή από αρχαιολόγους χρόνων. Οφείλεται αφενός μεν σε σχετική ή απόλυτη έλλειψη της ινσουλίνης, ορμόνης που παράγεται από τα β-κύτταρα των νησιδίων του Langerhans του παγκρέατος και αφετέρου στην αδυναμία της υπάρχουσας ινσουλίνης να ασκήσει τις δράσεις της στους ιστούς στόχους λόγω αντίστασης στην ινσουλίνη που παρουσιάζεται στους ιστούς αυτούς. Αποτέλεσμα αυτών είναι να εμφανίζεται διαταραχή στον μεταβολισμό της γλυκόζης, των λιπαρών οξέων και των αμινοξέων. Με συνέπεια την αύξηση της γλυκόζης στο αίμα και την δημιουργία των επιπλοκών του διαβήτη που είναι χαρακτηριστικές για αυτόν και αφορούν την εμφάνιση μικροαγγειοπάθειας με εκδηλώσεις από τους οφθαλμούς, τους νεφρούς και τα περιφερικά νεύρα κατά κύριο λόγο, αλλά και επιπλοκές από την μακροαγγειοπάθεια με εκδηλώσεις από την καρδιά και τα αγγεία που παρουσιάζονται συχνότερα στα άτομα με σακχαρώδη διαβήτη από ότι στον γενικό πληθυσμό [1].

1.3. Ταξινόμηση

Υπάρχουν 4 τύποι σακχαρώδους διαβήτη, ο Σακχαρώδης διαβήτης τύπου 1 (ΣΔτ1), ο Σακχαρώδης διαβήτης τύπου 2 (ΣΔτ2), ο Σακχαρώδης διαβήτης της κύησης και άλλοι τύποι σακχαρώδους διαβήτη. Συγκεκριμένα

Διαβήτης τύπου 1 (ΣΔ Ι): Επίσης γνωστός ως νεανικός διαβήτης ή ινσουλινοεξαρτώμενος διαβήτης. Σε αυτόν τον τύπο, το ανοσοποιητικό σύστημα καταστρέφει τα παγκρεατικά κύτταρα που παράγουν ινσουλίνη. Αυτό οδηγεί σε απόλυτη έλλειψη ινσουλίνης, και οι ασθενείς χρειάζεται να χρησιμοποιούν ενέσιμη ινσουλίνη από την αρχή για ελέγχουν τα επίπεδα της γλυκόζης στο αίμα τους. Συνήθως επηρεάζει νέους ανθρώπους κάτω των 40 ετών.

Διαβήτης τύπου 2 (ΣΔ ΙΙ): Παλιότερα ονομάζονταν σακχαρώδης διαβήτης των ενηλίκων μιας και εμφανίζεται σε άτομα μεγαλύτερης ηλικίας ή μη ινσουλινοεξαρτώμενος σακχαρώδης διαβήτης. Στον τύπο αυτό, ο οργανισμός

αντιμετωπίζει αυξημένη αντίσταση στην ινσουλίνη, και η παραγωγή ινσουλίνης δεν είναι επαρκής για να καλύψει τις μεταβολικές ανάγκες. Στα αρχικά στάδια, τα επίπεδα ινσουλίνης μπορεί να είναι φυσιολογικά ή αυξημένα, και οι ασθενείς αντιμετωπίζονται με αντιδιαβητικά φάρμακα. Σε προχωρημένα στάδια, η θεραπεία με ινσουλίνη μπορεί να απαιτηθεί. Ο διαβήτης τύπου 2 έχει κληρονομικό χαρακτήρα και αποτελεί τον πιο συχνό τύπο διαβήτη.

Διαβήτης της εγκυμοσύνης: Αυτός ο τύπος διαβήτη είναι η διαταραχή του μεταβολισμού των υδατανθράκων ποικίλης βαρύτητας με έναρξη ή πρώτη αναγνώριση στην παρούσα εγκυμοσύνη ή ακόμα και κατά την διάρκεια της εγκυμοσύνης.

Άλλοι ειδικοί τύποι διαβήτη: Υπάρχουν διάφοροι άλλοι τύποι διαβήτη που μπορεί να οφείλονται σε φάρμακα, ορμονικές διαταραχές, νοσήματα του παγκρέατος, γενετικά σύνδρομα, κ.λπ. Πιο ειδικά:

- Σακχαρώδης διαβήτης που προκαλείται από γενετικές διαταραχές που αφορούν την δράση της ινσουλίνης.
- Σακχαρώδης διαβήτης που προκαλείται από ενδοκρινοπάθειες, λοιμώξεις, νόσους του εξωκρινούς παγκρέατος ή άλλες νόσους.
- Σακχαρώδης διαβήτης που προκαλείται από φάρμακα ή χημικές ουσίες.
- Σακχαρώδης διαβήτης που προκαλείται από γενετικά σύνδρομα.
- Σακχαρώδης διαβήτης που προκαλείται από γενετικές διαταραχές της λειτουργίας των β-κυττάρων.

Η διάγνωση και η διαχείριση του διαβήτη απαιτούν συνήθως τη συνεργασία με έναν ιατρό για να επιλεγεί η κατάλληλη θεραπεία και να παρακολουθείται η κατάσταση του ασθενούς.

1.4. Συμπτώματα και Διάγνωση

Στα αρχικά στάδιά της η νόσος μπορεί να είναι εντελώς ασυμπτωματική και συνεπώς ο ασθενής μπορεί να μην γνωρίζει ότι πάσχει. Παρόλα αυτά οι επιπλοκές του διαβήτη εξελίσσονται από αυτά τα στάδια και πολλές φορές αποτελούν αυτές τις πρώτες εκδηλώσεις της νόσου. Τα συμπτώματα του σακχαρώδη διαβήτη που αφορούν τις αυξημένες τιμές σακχάρου στο αίμα είναι:

- Πολυουρία

- Πολυδιψία
- Ξηροστομία
- Πολυφαγία
- Απώλεια βάρους
- Θάμβος όρασης
- Κόπωση

Έχει αποδειχθεί ότι οι έγκαιρη διάγνωση του διαβήτη και πρώιμες παρεμβάσεις μειώνουν τον κίνδυνο ανάπτυξης επιπλοκών και βελτιώνουν την ποιότητα ζωής των ασθενών. Η διάγνωση του διαβήτη γίνεται με τα παρακάτω κριτήρια.

Η διάγνωση του ΣΔ τίθεται όταν:

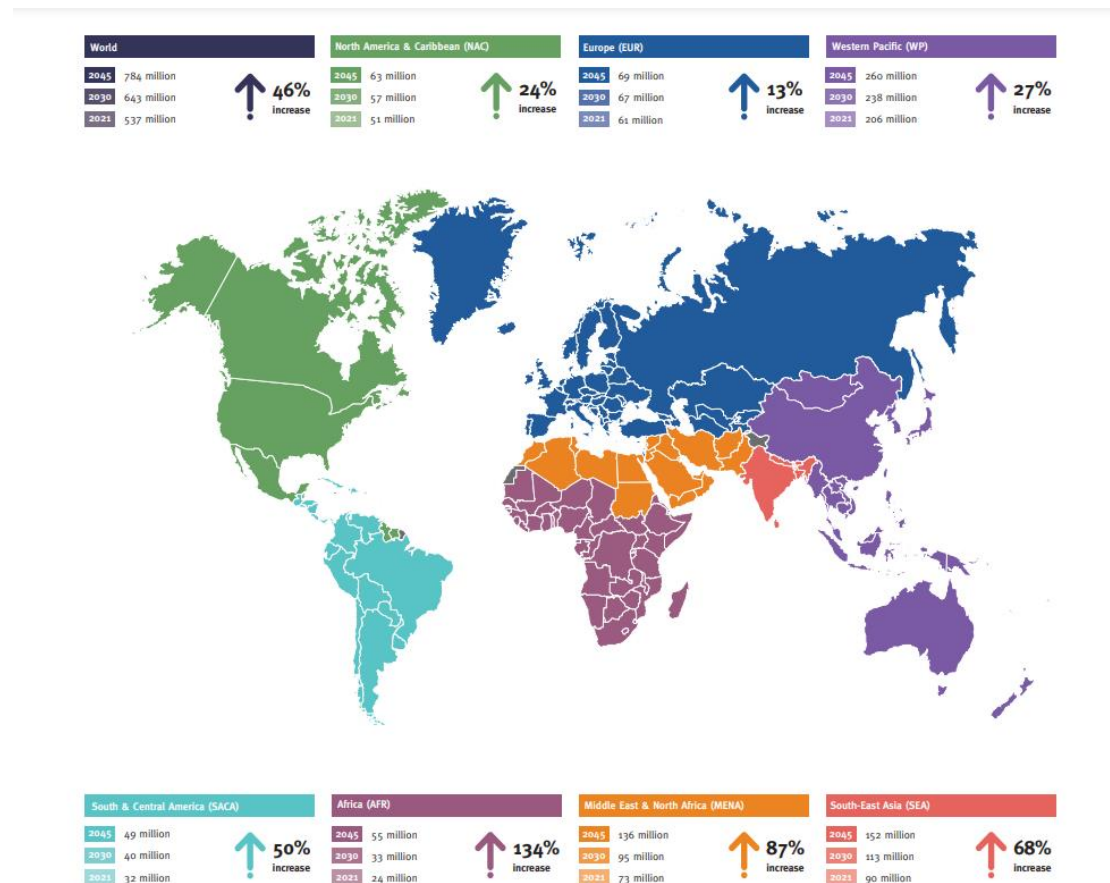
- Ο ασθενής έχει συμπτώματα (πολυουρία, πολυδιψία, απώλεια βάρους) και τιμή σακχάρου στο αίμα ίση ή μεγαλύτερη από 200 mg/dl
- Σάκχαρο αίματος μετά 8 ώρες νηστεία ίσο ή μεγαλύτερο από 126 mg/dl ή
- Σάκχαρο αίματος 2 ώρες μετά λήψη 75γρ γλυκόζης (καμπύλη σακχάρου) ίσο ή μεγαλύτερο από 200 mg/dl

Ο Παγκόσμιος Οργανισμός Υγείας και η Αμερικανική Διαβητολογική Εταιρεία (ADA) συνιστούν να τίθεται η διάγνωση του σακχαρώδη διαβήτη με βάση τα αποτελέσματα σε τουλάχιστον δύο δοκιμασίες ελέγχου γλυκόζης πλάσματος / αίματος. Στις οδηγίες της ελληνικής διαβητολογικής εταιρείας δεν περιλαμβάνεται αυτή η δοκιμασία στα διαγνωστικά κριτήρια του διαβήτη και έτσι μόνο ως ενδεικτική μπορεί να χρησιμοποιείται. Φυσικά, υπάρχουν και συμπληρωματικές εξετάσεις που χρησιμοποιούνται για την παρακολούθηση του σακχαρώδη διαβήτη, όπως είναι η ανάλυση ούρων (εξετάζεται εάν υπάρχει γλυκόζη), η μέτρηση κετονών και το λιπιδαιμικό προφίλ.

1.5. Επιδημιολογία

Περίπου 60 εκατομμύρια άτομα στην Ευρώπη, δηλαδή περίπου το 10,3% των ανδρών και το 9,6% των γυναικών άνω των 25 ετών είναι διαβητικοί. Το 21% των ατόμων άνω των 60 ετών έχει συμπτωματικό διαβήτη, ενώ το 45% των νεοδιαγνωσθεισών περιπτώσεων είναι παιδιά και έφηβοι. Η παγκόσμια πρόβλεψη επιπολασμού για τα έτη 2017-2045 είναι ότι το 2017 υπήρχαν 425 εκατομμύρια διαβητικοί, το 2045

προβλέπεται ότι θα αυξηθούν στα 629 εκατομμύρια, ενώ τα πράγματα γίνονται χειρότερα αν προσθέσει κανείς σε αυτούς και τα άτομα υψηλού κινδύνου για διαβήτη, όπου συνολικά το 2017 υπολογίστηκαν σε 777 εκατομμύρια, ενώ το 2045 αναμένεται να φτάσουν τα 1.257 εκατομμύρια (Εικόνα 1.1).



Εικόνα 1.1. Συγκριτικός επιπολασμός ΣΔ σε παγκόσμια κλίμακα για το 2021-2045 (20-79 ετών). [Πηγή: International Diabetes Federation. IDF Diabetes Atlas, 10th edn. Brussels, Belgium: International Diabetes Federation, 2021. (<http://www.diabetesatlas.org>)

Ο διαβήτης προκάλεσε 6,7 εκατομμύρια θανάτους το 2021 και συγκεκριμένα 1 κάθε 5 δευτερόλεπτα. Υπολογίζεται ότι το 44% των ενηλίκων που ζουν με διαβήτη (240 εκατομμύρια άνθρωποι) δεν έχουν διαγνωστεί. Σχεδόν το 90% αυτών των ανθρώπων ζουν σε χώρες χαμηλού και μεσαίου εισοδήματος. Η συχνότητα του σακχαρώδους διαβήτη στην Ελλάδα (διαγνωσμένου και αδιάγνωστου) σύμφωνα με επιδημιολογικές μελέτες είναι της τάξης του 7-8%. Από το σύνολο των ατόμων με διαβήτη, το περίπου 95% έχει ΣΔτ2 και το υπόλοιπο περίπου 5% έχει ΣΔτ1. Διαχρονικά παρατηρείται μια συνεχής αύξηση της συχνότητας του σακχαρώδους διαβήτη στην Ελλάδα (2,4% το 1974, 3,1% το 1990 και 7-8% σήμερα).

1.6. Αντιμετώπιση του Σακχαρώδη Διαβήτη

Μετά τη διάγνωση του διαβήτη η δίαιτα και η άσκηση εξακολουθούν και αποτελούν τον ακρογωνιαίο λίθο για την αντιμετώπιση που μαζί με την υπογλυκαιμική αγωγή, τις προληπτικές παρεμβάσεις (πχ. αντιγριπικός εμβολιασμός) και τον τακτικό έλεγχο για εμφάνιση επιπλοκών (καρδιογράφημα, βυθοσκόπηση, κ.α.) εξασφαλίζουν μια καλή ποιότητα ζωής για τα άτομα με διαβήτη). Για την αντιμετώπιση του σακχαρώδους διαβήτη συνίσταται:

Α) Δίαιτα: Τα άτομα με σακχαρώδη διαβήτη πρέπει να διαμορφώσουν ένα πρόγραμμα γευμάτων σε συνεννόηση με το γιατρό τους προσαρμοσμένο στις ανάγκες και τον τρόπο ζωής του κάθε ατόμου. Ο σχεδιασμός του διατροφικού προγράμματος για τον διαβήτη περιλαμβάνει τρεις σημαντικούς παράγοντες:

Ποσότητα και Θερμίδες: Ο υπολογισμός της πρόσληψης θερμίδων και των υδατανθράκων είναι απαραίτητος για τη διαχείριση του διαβήτη. Αυτός ο υπολογισμός βοηθά στη διατήρηση των επιπέδων γλυκόζης στο αίμα σε ασφαλές επίπεδο.

Τύποι Τροφών: Η επιλογή των τροφών παίζει σημαντικό ρόλο στον έλεγχο του διαβήτη. Οι υγιεινές επιλογές όπως φρούτα, λαχανικά, πλήρεις σιτηρέσιοι, λευκά κρέατα και ψάρια αποτελούν τμήμα της ισορροπημένης διατροφής.

Χρονική Κατανάλωση: Η διάταξη των γευμάτων και η χρονική τους κατανάλωση επηρεάζουν τον έλεγχο της γλυκόζης. Η διατήρηση του χρόνου κατανάλωσης γευμάτων μπορεί να βοηθήσει στη διατήρηση σταθερών επιπέδων γλυκόζης στο αίμα.

Β) Σωματική άσκηση: Η τακτική, μέτρια άσκηση αποτελεί ένα σημαντικό μέσο για τον έλεγχο των επιπέδων γλυκόζης στο αίμα και την διατήρηση της καρδιαγγειακής υγείας. Η άσκηση βελτιώνει την ευαισθησία των ιστών στην ινσουλίνη και μειώνει τα επίπεδα ινσουλίνης και της γλυκόζης στο αίμα τόσο κατά τη νηστεία όσο και μετά τα γεύματα. Η υιοθέτηση ενός υγιούς τρόπου ζωής μέσω της άσκησης είναι ιδιαίτερα σημαντική για τους ανθρώπους που πάσχουν από σακχαρώδη διαβήτη τύπου 2. Η άσκηση συμβάλλει επίσης στη μείωση του κινδύνου καρδιαγγειακών προβλημάτων, βελτιώνοντας το λιπιδαιμικό προφίλ και ρυθμίζοντας τα επίπεδα αρτηριακής πίεσης στα άτομα με διαβήτη. Σε συνδυασμό με τις αλλαγές στη διατροφή, η άσκηση αποτελεί αποτελεσματικό μέσο για τη μείωση του σωματικού λίπους. Αυτό αποτελεί συνήθως κρίσιμο παράγοντα για την επίτευξη καλού γλυκαιμικού ελέγχου στους ασθενείς με σακχαρώδη διαβήτη τύπου 2.

Γ) Φαρμακευτική αγωγή: Η θεραπεία του σακχαρώδη διαβήτη προσαρμόζεται ατομικά σύμφωνα με τις ανάγκες και την κατάσταση του κάθε ασθενούς. Για τον σακχαρώδη διαβήτη τύπου 2, η φαρμακευτική αγωγή μπορεί να ξεκινήσει διάφορους τρόπους. Συνήθως, οι γιατροί συνιστούν τη χρήση αντιδιαβητικών δισκίων όταν η διατροφή και η άσκηση δεν αρκούν για τον έλεγχο των επιπέδων γλυκόζης στο αίμα. Στην πράξη, λόγω του χρόνου που περνά από τη διάγνωση του διαβήτη τύπου 2 και της σημασίας της έγκαιρης θεραπείας λόγω της υπεργλυκαιμίας, ορισμένοι ασθενείς ξεκινούν αμέσως με αντιδιαβητική αγωγή από το στόμα ή ακόμη και με ινσουλίνη. Τα τελευταία χρόνια, ο αριθμός και η ποικιλία των αντιδιαβητικών φαρμάκων από το στόμα έχει αυξηθεί σημαντικά, και υπάρχουν διάφορες κατηγορίες φαρμάκων με διαφορετικούς τρόπους δράσης. Σε πολλές περιπτώσεις, οι γιατροί επιλέγουν τον κατάλληλο συνδυασμό φαρμάκων για κάθε ασθενή, λαμβάνοντας υπόψη τον τρόπο ζωής, τις ανάγκες και τις μετρήσεις των επιπέδων γλυκόζης στο αίμα του καθενός

1.7. Η Αναγκαιότητα εφαρμογής μηχανικής μάθησης στο πρόβλημα.

Ο Σακχαρώδης διαβήτης είναι μια σοβαρή νόσος που επηρεάζει τον τρόπο με τον οποίο ο οργανισμός χειρίζεται τη γλυκόζη και αποτελεί μία από τις πιο διαδεδομένες χρόνιες ασθένειες σε παγκόσμιο επίπεδο. Αυξημένα επίπεδα ζάχαρου στο αίμα μπορούν να προκαλέσουν σοβαρές επιπλοκές, όπως καρδιαγγειακά προβλήματα, νευρολογικές διαταραχές, προβλήματα όρασης, προβλήματα στους νεφρούς κ.λ.π. Η νόσος απαιτεί συνεχή παρακολούθηση, διατροφικές προσαρμογές και συχνά φαρμακευτική θεραπεία, καθιστώντας τον διαβήτη ένα σημαντικό δημόσιο υγείας ζήτημα. Η σοβαρότητά της αλλά και συχνότητά της, καθιστά επιτακτική την ανάγκη να εξετάσουμε τους τρόπους με τους οποίους η μηχανική μάθηση μπορεί να συμβάλει στην αντιμετώπισή της. Οι διαβητικοί ασθενείς επισκέπτονται συχνά τα νοσοκομεία για επανεξετάσεις, θεραπευτική ρύθμιση, ή ακόμη και επείγουσες καταστάσεις καθώς είναι ευάλωτοι σε επιπλοκές. Συνεπώς, η πρόβλεψη της επιστροφής διαβητικών ασθενών στο νοσοκομείο μέσα σε συγκεκριμένο χρονικό διάστημα, καθώς και η εκτίμηση του χρόνου που θα παραμείνουν εκεί αποτελεί ένα πολύ μεγάλο βήμα στην πρόωρη αντιμετώπιση του αλλά και στη διαχείρισή του γενικότερα. Η μηχανική μάθηση προσφέρει μια ισχυρή εργαλειοθήκη για την αντιμετώπιση αυτής της πρόκλησης. Χρησιμοποιώντας δεδομένα διαβητικών ασθενών, όπως το ιατρικό τους ιστορικό, τα επίπεδα γλυκόζης στο αίμα, τους γενετικούς παράγοντες ή άλλες

παραμέτρους και εφαρμόζοντας διάφορους αλγορίθμους μηχανικής μάθησης μπορούμε να δημιουργήσουμε προγνωστικά μοντέλα. Αυτά τα μοντέλα θα είναι σε θέση να προβλέπουν με υψηλή ακρίβεια εάν ένα άτομο είναι πιθανό να ξαναμπει στο νοσοκομείο εντός ενός συγκεκριμένου χρονικού πλαισίου (σε περισσότερο από 30 μέρες ή σε λιγότερο από 30 ημέρες) ή δεν θα ξαναμπει αλλά και πόσες μέρες θα παραμείνει εκεί. Τα οφέλη αυτής αυτών των προβλέψεων είναι πολλαπλά. Καταρχάς, επιτρέπει στους ιατρούς να λαμβάνουν προληπτικά μέτρα για τη βελτίωση της φροντίδας του διαβητικού ασθενούς, προσφέροντας την κατάλληλη παρακολούθηση και θεραπεία ενώ προσδιορίζονται πρόωρα και οι διαβητικοί ασθενείς που κινδυνεύουν περισσότερο από την ασθένεια κάνοντας επανεισαγωγή. Επιπλέον, βοηθά στην εξοικονόμηση πόρων του νοσοκομείου, καθώς μπορεί να μειώσει τον αριθμό των αναγκαίων επισκέψεων και νοσηλείων. Τέλος, προσφέρει στους ασθενείς τη δυνατότητα για πιο προσωποποιημένη και αποτελεσματική θεραπεία, βοηθώντας στη βελτίωση της ποιότητας της ζωής τους. Συνεπώς, η εφαρμογή της μηχανικής μάθησης για την πρόβλεψη των επιστροφών διαβητικών ασθενών στο νοσοκομείο αλλά και της διάρκειας παραμονής τους, δεν παρέχει μόνο τη δυνατότητα βελτιωμένης διαχείρισης της νόσου και της υγείας των ατόμων, προσφέροντας προσωποποιημένη, αποτελεσματική, και στοχευμένη περίθαλψη αλλά έχει και οικονομικά οφέλη για τα νοσοκομεία.

ΚΕΦΑΛΑΙΟ 2^ο - Βιβλιογραφική Ανασκόπηση

Παρακάτω παρουσιάζονται μελέτες και διάφορα συστήματα υποστήριξης αποφάσεων για την πρόβλεψη του διαβήτη όπως το σύστημα *KADIS* ή το *GlucoTab*, της επανεισδοχής διαβητικών ασθενών στο νοσοκομείο αλλά και της διάρκειας παραμονής τους.

2.1. Karlsburger Diabetes-Management System (KADIS)

Στη μελέτη, εξετάστηκε κατά πόσον ένα DSS για τον διαβήτη μπορεί να βελτιώσει τον μακροχρόνιο ή/και βραχυχρόνιο γλυκαιμικό έλεγχο καθώς ο αυξανόμενος επιπολασμός του σακχαρώδη διαβήτη τύπου 2 (T2D) και η έλλειψη εξειδικευμένων ιατρών έχει προκαλέσει ένα κενό υγειονομικής περίθαλψης που είναι δυνατό να καλυφθεί με ένα σύστημα υποστήριξης αποφάσεων. Πρόκειται για την εφαρμογή ενός εξατομικευμένου συστήματος υποστήριξης αποφάσεων που ονομάζεται Karlsburger Diabetes-Management System (KADIS) [2]. Συγκεκριμένα, γίνεται μια αναδρομική μελέτη κοορτής του προγράμματος Diabetiva, το οποίο προσέφερε το παραπάνω DSS μία φορά το χρόνο. Ο γλυκαιμικός έλεγχος αναλύθηκε κατά την έναρξη και μετά από 12 μήνες σε 452 άτομα με T2D. Ο χρόνος στο εύρος (TIR- γλυκόζη 3,9-10 mmol/L) και το Q-Score, μια σύνθετη μετρική που αναπτύχθηκε για την ανάλυση συνεχών προφίλ γλυκόζης, ήταν βραχυπρόθεσμα και η HbA1c μακροπρόθεσμα μέτρα γλυκαιμικού ελέγχου. Μετρήθηκε επίσης η μεταβλητότητα της γλυκόζης (GV). Στην αρχή, το ένα τρίτο των ασθενών είχε καλό βραχυπρόθεσμο και μακροπρόθεσμο γλυκαιμικό έλεγχο. Το Q-Score εντόπισε ανεπαρκή βραχυπρόθεσμο γλυκαιμικό έλεγχο στο 17,9% των ασθενών με HbA1c <6,5%, κυρίως λόγω υπογλυκαιμίας. Η GV και η υπεργλυκαιμία ήταν υπαίτιες σε ασθενείς με HbA1c >7,5% και >8%, αντίστοιχα. Η εφαρμογή του DSS κατά την έναρξη βελτίωσε τον βραχυπρόθεσμο και μακροπρόθεσμο γλυκαιμικό έλεγχο, όπως φάνηκε από τη μείωση του Q-Score, της GV και της HbA1c μετά από 12 μήνες. Η πολλαπλή παλινδρόμηση έδειξε ότι η συνολική επίδραση στην GV προέκυψε από τις μεμονωμένες επιδράσεις όλων των παραμέτρων που επηρεάζουν (Εικόνα 2.1).

Parameter	No DSS			DSS		
	Baseline	12 months	P	Baseline	12 months	P
Quality of glycemic control						
HbA1c	6.93 ± 0.79	7.31 ± 0.84	<.00001	7.02 ± 1.01	6.69 ± 0.75	<.001
Q-Score	8.86 ± 3.33	9.07 ± 3.32		7.74 ± 3.49	7.07 ± 2.98	<.001
TIR	78.0 ± 18.1	74.5 ± 23.0		82.0 ± 20.2	85.5 ± 15.3	<.001
Q-Score parameters						
MSG (mmol/L)	7.92 ± 1.53	8.23 ± 1.76	.049	7.67 ± 1.80	7.37 ± 1.21	<.001
Range (mmol/L)	8.20 ± 2.99	8.25 ± 2.61		7.24 ± 2.79	6.83 ± 2.60	.006
MODD (mmol/L)	1.95 ± 0.80	2.03 ± 0.84		1.72 ± 0.83	1.65 ± 0.75	
%Hyper	12.9 (5.6; 26.3)	16.4 (5.9; 35.8)	.013	7.1 (1.5; 19.3)	5.5 (0.7; 14.7)	.001
%Hypo	0.2 (0.0; 2.8)	0.1 (0.0; 2.8)		0.3 (0.0; 3.3)	0.1 (0.0; 2.4)	
Glucose variability						
SD (mmol/L)	1.91 ± 0.78	1.94 ± 0.70		1.66 ± 0.75	1.56 ± 0.66	.008
CV	24.5 ± 9.5	23.8 ± 7.7		21.8 ± 8.8	21.1 ± 8.1	
IQR (mmol/L)	2.62 ± 1.17	2.68 ± 1.15		2.25 ± 1.21	2.11 ± 1.01	.036
MAGE (mmol/L)	5.19 ± 2.27	5.23 ± 2.06		4.49 ± 2.10	4.17 ± 1.88	.003
CONGA (mmol/l)	2.84 ± 1.24	2.85 ± 1.15		2.46 ± 1.22	2.28 ± 1.03	.003

Εικόνα 2.1. Σύγκριση του γλυκαιμικού ελέγχου και της μεταβλητότητας της γλυκόζης σε ασθενείς που έλαβαν θεραπεία με (n = 322) ή χωρίς (n = 130) DSS. Τα δεδομένα δίνονται ως μέσος όρος(SD) ή διάμεσος

Αποδείχθηκε πως, το DSS μπορεί να βελτιώσει τον βραχυπρόθεσμο και μακροπρόθεσμο γλυκαιμικό έλεγχο σε άτομα με T2D χωρίς να αυξάνει την υπογλυκαιμία. Ακόμη, το Q-Score επιτρέπει τον εντοπισμό ατόμων με ανεπαρκή γλυκαιμικό έλεγχο.

2.2. Πρόβλεψη διαβήτη με ταξινομητές μηχανικής μάθησης και ρύθμιση των υπερπαραμέτρων τους

Η παρούσα μελέτη [3] παρουσιάζει τη βελτίωση της απόδοσης των ταξινομητών για την πρόβλεψη του διαβήτη με την επιλογή κατάλληλων μεθόδων προεπεξεργασίας και τη ρύθμιση των υπερπαραμέτρων των ταξινομητών. Χρησιμοποιούνται ταξινομητές μηχανικής μάθησης σε Python για το σύνολο δεδομένων διαβήτη PIMA. Το σύνολο δεδομένων προεπεξεργάζεται με τέσσερις διαφορετικές μεθόδους όπως βλέπουμε στην Εικόνα 2.2, για τον χειρισμό των ελλিপών τιμών και των ακραίων τιμών.

Dataset	Shape of Datasets	No. of (+ve /-ve) Samples	Preprocessing Activities
D1- Actual Pima Diabetes Dataset	768 X 9	268 / 500	Preprocessing is not performed
D2- Dataset filling missing values with column's mean value	768 X 9	268 / 500	Fill all unknown values by corresponding column's mean value
D3- Dataset removing missing values rows	393 X 9	130 / 263	Remove all rows which have unknown value
D4- Dataset removing outliers and rows having missing values	337 X 9	98 / 239	Remove rows having missing values and outliers

Εικόνα 2.2. Διαφορετικές εκδόχες του PIMA Dataset μετά από προεπεξεργασία με διαφορετικές μεθόδους.

Οι ταξινομητές που χρησιμοποιούνται είναι οι KNN, Decision Tree, Random Forest και Support Vector Machine ενώ συντονίζονται και οι παράμετροί τους. Η απόδοση

των ταξινομητών αξιολογείται με τη χρήση διαφόρων μετρικών απόδοσης όπως F1score, Accuracy, Precision και Recall (Εικόνα 2.3).

Metrics	Max. (%)	Classifiers Arrange in Non-decreasing Order of Their Score				Min. (%)
		Rank 1	Rank 2	Rank 3	Rank 4	
F1 Score	75.68	RF	DT	SVM	KNN	61.53
Accuracy	88.61	RF	DT	SVM	KNN	81.01
Precision	100.00	RF	SVM	DT	KNN	75.00
Recall	73.91	DT	SVM	RF	KNN	52.17
Specificity	100.00	RF	KNN	SVM	DT	91.07

Εικόνα 2.3. Κατάταξη των ταξινομητών για το σύνολο δεδομένων D3

Performance Metrics	Dataset D1 : Actual PIMA Diabetes Dataset		Dataset D2 : Missing value are filled by Mean		Dataset D3: Removing Rows having Missing Values		Dataset D4 : Removing Rows having Missing Values & Outliers	
	Max. Score	*BPC	Max. Score	BPC	Max. Score	BPC	Max.	BPC
F1 Score	72.73	RF	73.39	RF	75.68	RF	62.5	SVM
Accuracy	80.52	RF	81.17	RF	88.61	RF	82.35	SVM
Precision	74.47	DT	76.09	DT	100.00	RF	83.33	SVM
Recall	72.73	RF	72.72	RF	73.91	DT	59.26	DT
Specificity	90.74	SVM	90.74	SVM	100.00	RF	95.83	SVM

Εικόνα 2.4. Προβολή των επιδόσεων όλων των εκδοχών των PIMA Datasets που δείχνουν τον ταξινομητή με τις καλύτερες επιδόσεις.

Τα αποτελέσματα δείχνουν ότι (Εικόνα 2.4), ο καλύτερος ταξινομητής για την πρόβλεψη του διαβήτη είναι ο Random Forest. Πετυχαίνει τα υψηλότερα F1score και Accuracy με τιμές 75,68 % και 88,61 % αντίστοιχα για το μοντέλο D3 του συνόλου δεδομένων που λαμβάνεται με την αφαίρεση των δειγμάτων που έχουν ελλιπή ή άγνωστες τιμές από το σύνολο δεδομένων PIMA.. Επίσης, φαίνεται ότι η απόδοση των ταξινομητών μπορεί να βελτιωθεί με την επιλογή κατάλληλων μεθόδων προεπεξεργασίας και τη ρύθμιση των υπερπαραμέτρων των ταξινομητών.

2.3. Σύστημα GlucoTab

Στην εργασία γίνεται αναφορά των αποτελεσμάτων μιας κλινικής μελέτης στην οποία δοκιμάστηκε το φορητό client-server σύστημα (GlucoTab®) για την ηλεκτρονική υποστήριξη αποφάσεων και ροών εργασίας ενώ εξετάζεται ο αντίκτυπός του στην πρόληψη της υπογλυκαιμίας [4]. Το σύστημα δοκιμάστηκε σε 30 ασθενείς με σακχαρώδη διαβήτη τύπου 2 (T2DM). Το σύστημα στηρίζει τους HCP κατά την εκτέλεση μιας basal-bolus ινσουλινοθεραπείας. Καταγράφηκαν η θεραπεία του διαβήτη, οι ανεπιθύμητες ενέργειες, τα σφάλματα του λογισμικού και η

ανατροφοδότηση του χρήστη. Διερευνήθηκαν η ασφάλεια, η αποτελεσματικότητα και η αποδοχή του συστήματος από τους χρήστες.

Glycemic management		Treatment process support	
Mean premeal BG value (patient-day-weighted mg/dl)		Documented treatment activities	
		Tablet/eCRF, n (%)	
Overall	154.9 ± 31.6	BG measurements	866/883 (98.1%)
Morning	150.8 ± 37.5	Basal injections	229/235 (97.4%)
Noon	176.3 ± 50.7	Bolus injections	817/874 (93.5%)
Evening	141.0 ± 29.9	New total daily dose	198/207 (95.7%)
Bedtime	147.2 ± 32.6	System uptime	Hours supported/total hours (%)
Premeal BG values in range (%)	%		5154/5295 (97.3%)
0 mg/dl to <40 mg/dl	0.0	Treatment activities (mean number of premeal activities/day)	Mean ± SD (max)
40 mg/dl to <70 mg/dl	1.3	BG measurements	3.4 ± 0.3 (4)
70 mg/dl to <100 mg/dl	11.8	Basal injections	0.9 ± 0.1 (1)
100 mg/dl to <140 mg/dl	36.4	Bolus injections	3.3 ± 0.3 (4)
>140 mg/dl to <180 mg/dl	24.8	Adherence to the insulin dose suggestions (%)	Section 1.01 %
180 mg/dl to <300 mg/dl	23.1	Bolus	96.2
≥300 mg/dl	2.6	Basal	95.7
Hypoglycemic events (n)	n	Adjustment of the total daily dose	98.7
Mild (≥40 mg/dl and <70 mg/dl)	12		
Severe (<40 mg/dl)	0		
Adverse events (n)	n		
Adverse events not device related	20		
Adverse device-related events	0		
Serious adverse events	4		
Serious adverse device-related events	0		
Anticipated serious adverse device related events	0		
Unexpected serious adverse device related events	0		
Insulin therapy (patient-day-weighted) [U]	Mean ± SD		
Daily ordered basal	20.8 ± 12.2		
Daily ordered bolus	26.3 ± 15.2		

Εικόνα 2.5. Γλυκαιμική διαχείριση, ανεπιθύμητες ενέργειες, θεραπεία Ινσουλίνης και Εικόνα 2.6. Υποστήριξη της θεραπευτικής διαδικασίας και τήρηση.

Μόνο το 1,3% των μετρήσεων γλυκόζης αίματος (BG) ήταν <70 mg/dl και μόνο το 2,6% ήταν >300 mg/dl (Εικόνα 2.5). Η διαθεσιμότητα του συστήματος (97,3%) και το ποσοστό των θεραπευτικών δραστηριοτήτων που καταγράφηκαν με το σύστημα (>93,5%) ήταν υψηλά. Μόνο λίγες προτάσεις του συστήματος απορρίφθηκαν από τους χρήστες (>95,7% τήρησης) (Εικόνα 2.6). Με την αξιολόγηση 3 ανώνυμων ερωτηματολογίων διαπιστώθηκε ότι πλειονότητα των χρηστών πίστευε ότι, οι εσφαλμένες θεραπείες θα μπορούσαν να αποφευχθούν με τη χρήση αυτού του συστήματος. Γενικότερα, τα δεδομένα της μελέτης έδειξαν σημαντική μείωση της υπογλυκαιμίας με την εφαρμογή ενός τέτοιου συστήματος για τη διαχείριση του διαβήτη, σε σύγκριση με μια διαδικασία που γίνεται στο χαρτί ενώ έγινε αποδεκτό από τους ιατρούς καθώς οι χρήστες τηρούσαν τις προτάσεις δόσης ινσουλίνης που έκανε το σύστημα

2.4. Εφαρμογή Diabits

Στην εργασία παρουσιάζονται οι διάφορες δυσκολίες της προγνωστικής παρακολούθησης της γλυκαιμίας και εξετάζονται η ακρίβεια και τα αποτελέσματα του

ελέγχου της γλυκόζης αίματος του Diabits, μιας εφαρμογής για έξυπνα κινητά τηλέφωνα που βοηθά διαβητικούς στην παρακολούθηση και στη διαχείριση των επιπέδων γλυκόζης σε πραγματικό χρόνο [5].

Χρησιμοποιώντας δεδομένα από τις συσκευές συνεχούς παρακολούθησης γλυκόζης (CGM) και την καταχώρηση του χρήστη, η εφαρμογή χρησιμοποιεί τεχνικές μηχανικής μάθησης για τη δημιουργία εξατομικευμένων μοντέλων ασθενών και προβλέπει τις αυξομειώσεις της γλυκόζης στο αίμα έως και 60 λεπτά εκ των προτέρων. Αυτές οι προβλέψεις βοηθούν τους ασθενείς να κρατήσουν τα επίπεδα γλυκόζης εντός ορίων. Εξετάζονται με βάση τα δεδομένα πραγματικών χρηστών, η ακρίβεια πρόβλεψης του Diabits και η συσχέτιση μεταξύ της καθημερινής χρήσης της εφαρμογής και των μετρήσεων ελέγχου της γλυκόζης αίματος. Επιπλέον, υπολογίστηκε και έγινε σύγκριση με την ακρίβεια των προβλέψεων το 2018 Ohio T1DM (Σακχαρώδης διαβήτης τύπου 1). Με βάση 6,8 εκατομμύρια δεδομένα, οι προβλέψεις 30 λεπτών Diabits που αξιολογήθηκαν με τη μετρική Parkes Error Grid μετρικής βρέθηκαν να είναι 86,89% κλινικά ακριβείς (ζώνη A) και 99,56% κλινικά αποδεκτές (ζώνες A και B), ενώ οι προβλέψεις 60 λεπτών ήταν 70,56% κλινικά ακριβείς και 97,49% κλινικά αποδεκτές.

Daily sessions ^a	0	1-5	6-10	>10
Average blood glucose (mg/dL)	154.0	150.7; $P_0 < .001$; $P_{\text{fewer}} < .001$	145.6; $P_0 < .001$; $P_{\text{fewer}} < .001$	141.6; $P_0 < .001$; $P_{\text{fewer}} < .001$
Standard deviation (mg/dL)	47.6	45.3; $P_0 < .001$; $P_{\text{fewer}} < .001$	42.1; $P_0 < .001$; $P_{\text{fewer}} < .001$	41.5; $P_0 < .001$; $P_{\text{fewer}} = .07$
Time in euglycemic range, as % of all data	67.52	69.39%; $P_0 < .001$; $P_{\text{fewer}} < .001$	73.05%; $P_0 < .001$; $P_{\text{fewer}} < .001$	74.28%; $P_0 < .001$; $P_{\text{fewer}} = .004$
GMI ^c (%)	6.99	6.91%; $P_0 < .001$; $P_{\text{fewer}} < .001$	6.79%; $P_0 < .001$; $P_{\text{fewer}} < .001$	6.70%; $P_0 < .001$; $P_{\text{fewer}} < .001$
HbG1 ^d (<4.5: low risk; 4.5-9.0: moderate risk; >9.0: high risk) [67]	4.63	4.20; $P_0 < .001$; $P_{\text{fewer}} < .001$	3.62; $P_0 < .001$; $P_{\text{fewer}} < .001$	3.13; $P_0 < .001$; $P_{\text{fewer}} < .001$
LBGI ^e (<1.1: minimal risk; 1.1-2.5: low risk; 2.5-5.0: moderate risk; >5.0: high risk) [69]	0.42	0.45; $P_0 < .001$; $P_{\text{fewer}} < .001$	0.46; $P_0 = .007$; $P_{\text{fewer}} = .32$	0.59; $P_0 < .001$; $P_{\text{fewer}} < .001$

Εικόνα 2.7. Μετρήσεις του ελέγχου της γλυκόζης στο αίμα σε συνδυασμό με τη συχνότητα της ημερήσιας χρήσης του Diabits.

Αναλύοντας στατιστικά στοιχεία καθημερινής χρήσης και δεδομένα CGM για τους 280 πιο μακροχρόνιους χρήστες του Diabits (Εικόνα 2.7), διαπιστώθηκε ότι, υπό συνθήκες ελεύθερης διαβίωσης, οι μετρήσεις ελέγχου της γλυκόζης στο αίμα βελτιώθηκαν με την αύξηση της συχνότητας χρήσης της εφαρμογής. Π.χ , ο μέσος όρος της γλυκόζης αίματος για τις ημέρες που οι χρήστες δεν αλληλεπιδρούσαν με την εφαρμογή ήταν 154,0 (SD 47,2) mg/dL, με το 67,52% του χρόνου να διανύεται στο υγιές εύρος 70 έως 180 mg/dL. Για τις ημέρες με 10 ή περισσότερες φορές χρήσης της εφαρμογής, η μέση γλυκόζη αίματος μειώθηκε σε 141,6 (SD 42,0) mg/dL ($P < .001$), ενώ ο χρόνος σε

ευγλυκαιμικό εύρος αυξήθηκε σε 74,28% ($P<.001$), παρατηρώντας βελτίωση στον έλεγχο της γλυκόζης τις ημέρες με συχνότερη χρήση Diabits Στο σύνολο δεδομένων T1DM του Οχάιο με 6 ασθενείς με διαβήτη τύπου 1, οι προβλέψεις 30 λεπτών του βασικού μοντέλου Diabits είχαν RMSE 18,68 (SD 2,19) mg/dL, το οποίο αποτελεί βελτίωση σε σχέση με τα δημοσιευμένα αποτελέσματα της τελευταίας τεχνολογίας για αυτό το Dataset.

2.5. Ανασκόπηση στους παράγοντες κινδύνου επανεισαγωγής νοσηλευόμενων διαβητικών

Η εργασία κάνει ανασκόπηση για να εντοπίσει με συστηματικό τρόπο τους γνωστούς παράγοντες κινδύνου για επανεισαγωγή στο νοσοκομείο μεταξύ ατόμων με διαβήτη, για να γίνει καλύτερα αυτή την δαπανηρή επιλογή [6]. Γενικά, υπάρχει περιορισμένη κατανόηση των παραγόντων κινδύνου που συμβάλλουν στα αυξημένα ποσοστά επανεισδοχής μεταξύ των ατόμων που παίρνουν εξιτήριο από το νοσοκομείο με διαβήτη. Η ανασκόπηση καταχωρήθηκε προοπτικά στη βάση δεδομένων PROSPERO. Οι παράγοντες κινδύνου εντοπίστηκαν από 2 συγγραφείς μέσω συστηματικής ανασκόπησης της βιβλιογραφίας στις βάσεις δεδομένων PubMed, EMBASE & SCOPUS, η οποία πραγματοποιήθηκε πριν από την εξαγωγή των δεδομένων και εκτιμά την ποιότητα και την ημι-ποσοτική σύνθεση σύμφωνα με τις οδηγίες PRISMA (Εικόνα 2.8).

Table 1: Characteristics of studies identifying statistically significant risk factors for readmission in generalised diabetes populations (ordered by definition of readmission)

Ref	First author	Year	Readmission definition (d = day, m = month, yr = year)	Sample size	Country	Study Design	No. of variables		Primary outcome [* = Multiple primary or composite outcomes]
							Assessed	Significant	
[23]	X.Liu	2015	7d to >90d	37,620	China	Retrospective cohort analysis	13	13	Readmission
[24]	S.Mokhtar	2012	28d	1125	Saudi Arabia	Retrospective cohort & case-control	11	3	Readmission
[25]	J.Albrecht	2012	30d	26,878	USA	Retrospective cohort analysis	7	7	Readmission
[8]	D.Rubin	2016	30d	44,203	USA	Development & validation of risk tool	46	37	Readmission
[26]	H.Sonmez	2017	30d	102,694	USA	Retrospective cohort analysis	2	2	Readmission and association with admission diagnosis
[27]	J.Robbins	2006	30d	291,752	USA	Retrospective cohort analysis	31	31	Readmission
[28]	F.Zaccardi	2017	30d	101,475	UK	Retrospective case-control study	1	1	Admission, Readmission, length of stay, mortality*
[29]	J.Chen	2012	30d	30,139	USA	Retrospective cohort analysis	39	33	Readmission
[18]	J.Swami	2018	30d	70	USA	prospective observational study	5	2	Readmission
[30]	K.Lipska	2014	30d	33,952,331	USA	Retrospective observational	5	2	Hyper/hypoglycaemia, hospitalization, mortality & readmission*
[31]	K.Bennett	2012	30d	94,121	USA	Retrospective cohort analysis	24	13	Readmission
[32]	D.Rubin	2018	30d	105,791	USA	Retrospective cohort analysis	43	40	Readmission
[33]	A.Karunakaran	2018	30d	17284	USA	Retrospective cohort analysis	51	46	Readmission
[34]	S.Healy	2013	30d & 180d	2265	USA	Retrospective cohort analysis	26	11	Readmission
[35]	X.Liu	2017	30d, 60d & 90d	73,144	China	Cross sectional analysis	18	16	Readmission
[36]	L.Chwastiak	2014	1m	82,060	USA	Retrospective cohort analysis	46	29	Readmission
[37]	H.Kim	2010	3m	124,967	USA	Retrospective cohort analysis	32	18	Scheduled and unscheduled readmissions*
[38]	J.Ena	2018	90d	1977	Spain	Retrospective cohort analysis	21	10	Readmission
[39]	Y.Nishino	2015	Study period	445,504	UK	Cross-sectional analysis	36	24	Admission & readmission*
[22]	S.Cramer	2010	Not declared	2633	USA	Retrospective cohort analysis	10	10	Readmission

Εικόνα 2.8. Στοιχεία των μελετών που εντοπίζουν στατιστικά σημαντικούς παράγοντες κινδύνου για επανεισαγωγή σε γενικευμένους πληθυσμούς με διαβήτη (ταξινόμηση με βάση τον ορισμό της επανεισαγωγής)

Επιλέχθηκαν για συμπερίληψη 83 μελέτες, κυρίως από τις Ηνωμένες Πολιτείες, οι οποίες χρησιμοποίησαν αναδρομική ανάλυση τοπικών ή περιφερειακών Datasets. Εντοπίστηκαν 76 διακριτοί στατιστικά σημαντικοί παράγοντες κινδύνου σε 48 μελέτες. Οι παράγοντες κινδύνου που εντοπίστηκαν συνήθως ήταν: η συν-νοσηρότητα, η ηλικία, η φυλή και ο ασφαλιστικός τύπος. Λίγες μελέτες πραγματοποίησαν εκτιμήσεις ισχύος ενώ υπολογίστηκαν μεγέθη επίδρασης για την πλειονότητα των στατιστικά σημαντικών κινδύνων.

2.6. Παράγοντες πρόβλεψης των επανεισαγωγών και του LoS για ασθενείς με σακχαρώδη διαβήτη

Στη μελέτη, γίνεται πρόβλεψη των επανεισαγωγών και της διάρκειας παραμονής των ασθενών που πάσχουν από διαβήτη, προκειμένου να βελτιωθεί η ποιότητα της υγειονομικής περίθαλψης και να κατανεμηθούν αποτελεσματικότερα οι πόροι [7].

Η προεπεξεργασία των δεδομένων περιλαμβάνει την αφαίρεση χαρακτηριστικών με υψηλό ποσοστό ελλিপών τιμών όπως το "weight", ο "payer code" και η "medical speciality". Επίσης, αντιμετωπίστηκαν τα ασύμμετρα δεδομένα. Χαρακτηριστικά με υψηλή ασυμμετρία μετασχηματίστηκαν με λογαριθμικό μετασχηματισμό, συγκεκριμένα $\log(x+1)$ για να ληφθούν υπόψη οι μηδενικές τιμές. Τα ονομαστικά χαρακτηριστικά έγιναν αριθμητικά και οι αριθμητικές στήλες τυποποιήθηκαν με τη χρήση του z-score ώστε να συμμορφώνονται με μια γκαουσιανή κατανομή.

Για την επιλογή των χαρακτηριστικών χρησιμοποιήθηκαν η Recursive Feature Elimination (RFE), η feature importance based ranking και διάφοροι αλγόριθμοι μηχανικής μάθησης (Random Forest, Support Vector Machine (SVM), XGBoost, KNN και Logistic Regression). Στο στάδιο εκπαίδευσης και αξιολόγησης του μοντέλου χρησιμοποιήθηκε 10 k-fold cross validation και ως μετρικές αξιολόγησης χρησιμοποιήθηκαν τα Accuracy, Area Under the ROC Curve (AUC), Precision, Recall, and F1 score. Για να αντιμετωπιστεί το ζήτημα των ανισοβαρών κλάσεων, εφαρμόστηκε η τεχνική υπερδειγματοληψίας Synthetic Minority Over-sampling Technique (SMOTE). Επιλέον βήμα ήταν η κωδικοποίηση-διακριτοποίηση χαρακτηριστικών.

	Accuracy	Precision	Recall	F1-score
RF	0.9482	0.9522	0.9481	0.9480
LR	0.5862	0.5868	0.5861	0.5854
XGboost	0.9486	0.9533	0.9485	0.9485
SVM	0.9185	0.9209	0.9185	0.9184
KNN	0.8615	0.8730	0.8616	0.8604

Εικόνα 2.9. Μετρήσεις επιδόσεων για την πρόβλεψη επανεισδοχής.

	Accuracy	Precision	Recall	F1-score
RF	0.7984	0.8123	0.7984	0.7949
LR	0.5476	0.5371	0.5476	0.5403
XGboost	0.8102	0.8104	0.8095	0.8081
SVM	0.8748	0.8765	0.8748	0.8739
KNN	0.8546	0.8679	0.8546	0.8482

Εικόνα 2.10. Μετρικές επιδόσεων για την πρόβλεψη της διάρκειας παραμονής.

Τα αποτελέσματα της μελέτης έδειξαν ότι ο Random Forest πέτυχε accuracy 94,8% στην πρόβλεψη των επανεισαγωγών (Εικόνα 2.9) Ο αλγόριθμος SVM είχε το υψηλότερο στατιστικό AUC 0,97, υποδεικνύοντας την αποτελεσματικότητά του στην πρόβλεψη επανεισαγωγών. Για την πρόβλεψη της διάρκειας παραμονής, ο SVM είχε ακρίβεια πρόβλεψης 78,5% (Εικόνα 2.10).

2.7. Μάθηση για Κατηγοριοποίηση Κινδύνων στον Διαβήτη Τύπου 2

Η εργασία διερευνά τους προβλεπτικούς παράγοντες του γλυκαιμικού ελέγχου και στον προσδιορισμό της καταλληλότητας των μεθόδων μηχανικής μάθησης για την ταξινόμηση ασθενών με σακχαρώδη διαβήτη τύπου 2 με κακό και φυσιολογικό γλυκαιμικό έλεγχο [8].

Οι ερευνητές συνέλεξαν δεδομένα όπως η ηλικία, το φύλο και ο BMI, καθώς και μετρήσεις καταθλιπτικών συμπτωμάτων (χρησιμοποιώντας το Patient Health Questionnaire-9) και συμπτωμάτων που σχετίζονται με το άγχος (χρησιμοποιώντας το Generalized Anxiety Disorder-7). Τα σήματα ηλεκτροκαρδιογραφίας (ΗΚΓ) συλλέχθηκαν επίσης με τη χρήση φορητής συσκευής παρακολούθησης ECG και λήφθηκαν δείγματα αίματος για τη μέτρηση των λιπιδίων (HDL, LDL και TG) και των επιπέδων HbA1c. Έγινε έλεγχος της κυματομορφής του ECG και την αφαίρεση τυχόν

αρρυθμιών και κινήσεων. Στη συνέχεια, τα δεδομένα του διαστήματος μεταξύ των παλμών (IBI) μεταφορτώθηκαν και εισήχθησαν στο CardioPro Infiniti HRV Analysis Module για περαιτέρω ανάλυση. Από τα δεδομένα IBI προέκυψαν τα πεδία χρόνου και συχνότητας της μεταβλητότητας του καρδιακού ρυθμού (HRV). Το πεδίο του χρόνου περιλάμβανε παραμέτρους όπως η τυπική απόκλιση των κανονικών προς κανονικά διαστήματα (SDNN), η μέση τετραγωνική ρίζα των διαδοχικών διαφορών (RMSSD), ο αριθμός των ζευγών διαδοχικών NN που διαφέρουν περισσότερο από 50 ms (NN50) και το ποσοστό των NN50 (pNN50). Το πεδίο συχνοτήτων περιλάμβανε την πολύ χαμηλή συχνότητα (VLF), τη χαμηλή συχνότητα (LF), την υψηλή συχνότητα (HF), τη συνολική ισχύ (TP) και τον λόγο LF/HF. Οι λοξές κατανομές HRV μετασχηματίστηκαν με τη χρήση φυσικών λογαρίθμων.

Χρησιμοποιήθηκε ανάλυση δέντρων παλινδρόμησης για να εντοπίσουν τους πολλαπλούς παράγοντες κινδύνου που θα μπορούσαν να προβλέψουν τις τιμές της HbA1c και μέθοδοι μηχανικής μάθησης όπως ο k-nearest neighbors, οι αλγόριθμοι νευρωνικών δικτύων, ο SVM και ο Random forest για την ταξινόμηση των ασθενών σε ομάδες κακού και φυσιολογικού γλυκαιμικού ελέγχου. Η απόδοση αξιολογήθηκε με τη χρήση των accuracy, sensitivity, specificity και AUC-ROC (Εικόνα 2.11).

Model %	Testing Data			
	Sensitivity	Specificity	Area under Curves	Accuracy Rate
Boosting	51%	63%	63%	57%
Support vector machine	52%	65%	58%	58%
Classification tree	59%	80%	69%	69%
Neural network	69%	80%	60%	74%
K-nearest neighbors	63%	93%	78%	78%
Random forest	77%	91%	95%	84%

Εικόνα 2.11. Σύγκριση των μεθόδων ταξινόμησης μηχανικής μάθησης.

Τα αποτελέσματα της μελέτης έδειξαν ότι η ανάλυση δέντρων παλινδρόμησης εντόπισε πολλαπλούς παράγοντες κινδύνου για τον κακό γλυκαιμικό έλεγχο σε ασθενείς με T2DM. Αυτοί περιλαμβάνουν δημογραφικά στοιχεία (ηλικία, φύλο και ΔΜΣ), φυσιολογικές διαστάσεις (προφίλ λιπιδίων και δείκτες HRV) και συναισθηματικές διαστάσεις (κατάθλιψη και άγχος). Οι SVM και Random forest, συγκρίθηκαν και διαπιστώθηκε ότι το random forest υπερείχε του SVM σε όλες τις μετρικές με τιμές 84% accuracy, 77% sensitivity, 91% specificity και AUC-ROC 91%.

2.8. Πρόβλεψη του διαβήτη τύπου 2 με αλγόριθμο μηχανικής μάθησης

Αυτή η εργασία, αναπτύσσει ένα μοντέλο μηχανικής μάθησης για την πρόβλεψη της εμφάνισης του διαβήτη τύπου 2, για το επόμενο έτος χρησιμοποιώντας μεταβλητές στο τρέχον έτος που γίνεται η μελέτη [9]. Το Dataset προέρχεται από ηλεκτρονικά αρχεία υγείας από το 2013 έως το 2018. Για την κατασκευή του μοντέλου πρόβλεψης, επιλέχθηκαν αρχικά τα βασικά χαρακτηριστικά με τη χρήση δοκιμών ANOVA, δοκιμών chi-squared και αναδρομικών μεθόδων εξάλειψης χαρακτηριστικών. Το χαρακτηριστικά που προκύψαν ήταν η γλυκόζη πλάσματος νηστείας (FPG), η HbA1c, τα τριγλυκερίδια, ο BMI, το gamma-GTP, η ηλικία, το ουρικό οξύ, το φύλο, το κάπνισμα, η κατανάλωση αλκοόλ, η σωματική δραστηριότητα και το οικογενειακό ιστορικό. Οι αλγόριθμοι που χρησιμοποιήσαν ήταν logistic regression, random forest, SVM, XGBoost και αλγόριθμους μηχανικής μάθησης ensemble (Εικόνα 2.12) με βάση αυτές τις μεταβλητές για να προβλέψουμε το αποτέλεσμα ως κανονικό, προ-διαβήτη ή διαβήτη (Εικόνα 2.13).

	Accuracy	Precision	Recall	F1-score	MCC	KC
LR	0.71	0.71	0.71	0.71	0.56	0.56
RF	0.73	0.74	0.73	0.74	0.60	0.60
XGBoost	0.72	0.74	0.72	0.73	0.58	0.58
SVM	0.73	0.74	0.74	0.74	0.60	0.60
CIM	0.73	0.73	0.73	0.73	0.59	0.59
Stacking classifier	0.72	0.75	0.72	0.73	0.58	0.58
Soft voting	0.73	0.74	0.73	0.73	0.59	0.59

LR = logistic regression, RF = random forest, SVM = support vector machine, CIM = confusion matrix-based classifier integration approach, MCC = Matthews Correlation Coefficient, and KC = Cohen's kappa score.

Εικόνα 2.12. Σύγκριση επιδόσεων των μοντέλων πρόβλεψης στο σύνολο δεδομένων δοκιμής

	Normal	Prediabetes	Diabetes
Normal	148	58	4
Prediabetes	51	126	29
Diabetes	1	16	167

Εικόνα 2.13. Πίνακας σύγχυσης του μοντέλου RF.

Συγκεκριμένα, το accuracy των μοντέλων κυμάνθηκε από 71% έως 73%. Οι τιμές του precision για τις κατηγορίες φυσιολογικό, προ-διαβήτη και διαβήτη ήταν 70%, 61% και 90% αντίστοιχα. Οι υψηλότερες ψευδώς θετικές περιπτώσεις προέρχονταν από την κατηγορία προδιαβήτη. Το μοντέλο γενικότερα αποδείχθηκε αρκετά καλό στην πρόβλεψη της εμφάνισης του T2D στον κορεατικό πληθυσμό. Τα αποτελέσματα του cross validation έδειξαν ότι τα ensemble μοντέλα είχαν καλύτερη απόδοση από τα μεμονωμένα μοντέλα.

2.9. Πρόβλεψη των επανεισαγωγών διαβητικών ασθενών ανεξάρτητα από την HbA1c

Το πρόβλημα που πραγματεύεται η εργασία είναι η πρόβλεψη του κινδύνου επανεισαγωγής για διαβητικούς ασθενείς εντός 30 ημερών μετά την έξοδο από το νοσοκομείο. Γίνεται ανάπτυξη ενός πλαισίου μηχανικής μάθησης για τον εντοπισμό ασθενών υψηλού κινδύνου για επανεισδοχή, χρησιμοποιώντας δημογραφικές και διαγνωστικές πληροφορίες, πληροφορίες σχετικά με κλινικές διαδικασίες και τη φαρμακευτική αγωγή για όλες τις ηλικιακές ομάδες [10].

Το σύνολο δεδομένων χωρίζεται τυχαία σε ένα σύνολο εκπαίδευσης (75%) και ένα σύνολο δοκιμής (25%). Οι παράμετροι επιλέγονται με βάση τις επιδόσεις ταξινόμησης που αξιολογούνται με 10 k-fold cross validation στο σύνολο εκπαίδευσης. Επίσης έγινε χρήση του εργαλείου βαθμολόγησης του δείκτη LACE για την εκτίμηση κινδύνου επανεισδοχής στο νοσοκομείο. Γίνεται δημιουργία μοντέλων για κάθε ηλικιακή ομάδα ανεξάρτητα.

Για την ηλικιακή ομάδα [0-30), χρησιμοποιούνται τα μοντέλα Extreme Gradient Boosted Trees, Gradient Boosted Greedy Trees Classifier και ExtraTrees Classifier. Οι προβλεπόμενες πιθανότητες από αυτά τα μοντέλα υπολογίζονται κατά μέσο όρο με τη χρήση μιας μεθόδου ensemble (Εικόνα 2.14). Για την ηλικιακή ομάδα [30-70], χρησιμοποιούνται τα μοντέλα: Random Forest Classifier, Extreme Gradient Boosted Trees Classifier with Early Stopping, Nystroem Kernel SVM Classifier, Balanced Random Forest Classifier, Gradient Boosted Trees Classifier with Early Stopping, Gradient Boosted Greedy Trees Classifier with Early Stopping, Balanced ExtraTrees Classifier και Extreme Gradient Boosted Classifier με Early Stopping (Εικόνα 2.15). Οι προβλεπόμενες πιθανότητες από αυτά τα μοντέλα υπολογίζονται επίσης κατά μέσο όρο με τη χρήση μιας μεθόδου ensemble. Για την ηλικιακή ομάδα [70-100), το πλαίσιο μηχανικής μάθησης παρήγαγε ένα άλλο σύνολο 3 μοντέλων και είναι τα : Balanced ExtraTrees Classifier (Gin), Extreme Gradient Boosted Trees Classifier with Early Stopping, Extreme Gradient Boosted Trees Classifier with Early Stopping (Εικόνα 2.16).

Για την αξιολόγηση της απόδοσης χρησιμοποιούνται οι μετρικές F1 Score, Sensitivity, False Positive Rate, Specificity, Precision, Negative Predictive Value, Accuracy, and Matthews Correlation Coefficient (Εικόνα 2.17)

Measure	Value
F1 Score	0.4213
Sensitivity	0.4978
False Positive Rate	0.1081
Specificity	0.8919
Precision	0.3651
Negative Predictive Value	0.9343
Accuracy	0.8481
Matthews Correlation Coefficient	0.3416

Measure	Value
F1 Score	0.3001
Sensitivity	0.4363
False Positive Rate	0.1738
Specificity	0.8262
Precision	0.2288
Negative Predictive Value	0.9254
Accuracy	0.785
Matthews Correlation Coefficient	0.2012

Measure	Value
F1 Score	0.2694
Sensitivity	0.4902
False Positive Rate	0.2889
Specificity	0.7111
Precision	0.1857
Negative Predictive Value	0.9121
Accuracy	0.6849
Matthews Correlation Coefficient	0.1403

Εικόνες 2.14 (αριστερά), 2.15 (μέση) και 2.16 (δεξιά). Σύνοψη των μέτρων διαμόρφωσης για τις ομάδες [0-30), [30-70) και [70-100) αντίστοιχα με 10-fold cross validation

Age Group	Model Description	AUC
[0-30)	Ensemble Average	0.79
[30-70)	Ensemble Average	0.70
[70-100)	Meta Model (Elastic Net)	0.65
All Age Groups	LACE	0.56

Εικόνα 2.17. Σύγκριση της απόδοσης των μοντέλων

Συμπεραίνεται πως, η συγκεκριμένη προσέγγιση απέδωσε 26% βελτίωση χρησιμοποιώντας πάνω από 100.000 περιστατικά ασθενών από 130 νοσοκομεία των ΗΠΑ για μια περίοδο 10 ετών σε σύγκριση με το LACE, το οποίο προέκυψε από 4.800 ασθενείς για μια περίοδο 4 ετών.

2.10. Ένα οντολογικά βασισμένο και ερμηνεύσιμα ασαφές DSS για τη διάγνωση του διαβήτη

Η εργασία αποσκοπεί στην ανάπτυξη ενός ακριβούς, δυναμικού, σημασιολογικά ευφούς και ερμηνεύσιμου συστήματος υποστήριξης κλινικών αποφάσεων για τη διάγνωση του διαβήτη και προτείνει ένα οντολογικό ασαφές DSS που χρησιμοποιεί πολλαπλές πτυχές της γνώσης - ασαφή συμπερασματολογία, οντολογική συλλογιστική και μια ασαφή αναλυτική ιεραρχική διαδικασία (FAHP) για να παρέχει έναν πιο διαισθητικό και ακριβή σχεδιασμό [11].

Στην προεπεξεργασία έγινε αντιμετώπιση των ελλειπών τιμών, εντοπισμός των ακραίων τιμών, κωδικοποίηση κατηγορικών ή ονομαστικών χαρακτηριστικών και σημασιολογική κωδικοποίηση των ιατρικών εννοιών σύμφωνα με την τυποποιημένη ορολογία SCT. Το Dataset ελήφθη από νοσοκομεία και περιλάμβανε 39 χαρακτηριστικά. Ύστερα έγινε καθορισμός γλωσσικών μεταβλητών και ασαφή κατάτμηση. Καθορίζονται οι κρίσιμες γλωσσικές μεταβλητές, τα εύρη τους και η

λεπτομερής διάρθρωση, καθώς και τα ταατα και οι παράμετροι των συναρτήσεων συμμετοχής. Ο σχεδιασμός ερμηνεύσιμων ασαφών κατατμήσεων βασίζεται σε διάφορους περιορισμούς, όπως η διακριτότητα, ένας δικαιολογημένος αριθμός ασαφών συνόλων, η ομοιομορφία της κάλυψης, η κανονικοποίηση, η επικάλυψη και η ορθογωνιότητα. Επίσης, δε μπορεί να κατασκευαστεί ενιαίο FISS (Fuzzy Inference System Structure) λόγω του μεγάλου αριθμού χαρακτηριστικών εισόδου, οπότε σχεδιάζεται ένα ιεραρχικό σύστημα. Οι ισχυρές ασαφείς κατατμήσεις που σχετίζονται με όλα τα γλωσσικά χαρακτηριστικά προσδιορίζονται με τη χρήση μεθόδων κατάτμησης όπως οι k-means και HFP. Η ποιότητα κάθε ασαφούς διαμέρισης μετράται με τη χρήση τριών δεικτών ποιότητας (PC, PE και CI). Η καλύτερη κατάτμηση πρέπει να μεγιστοποιεί τα PC και CI ενώ ελαχιστοποιεί το PE. Στην Εικόνα 2.18 φαίνονται τα αποτελέσματα.

Dataset	Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Glucose level dataset	Naïve Bayes	96.67	96.7	96.7	96.7
	SVM	96.67	96.7	96.7	96.7
	Logistic regression	95	95.5	95	95
	KNN (k=3)	95	95.1	95	95
	Decision tree (C4.5)	100	100	100	100
	ANN	91.67	92.1	91.7	91.7
	Random forest	98.33	98.4	98.3	98.3
	FRBS 1: RP5-SP-FDT-S (3.2 rules)	98.33	98.75	98.33	98.29
Kidney function dataset	Naïve Bayes	38.3	31	38.3	32.9
	SVM	45	26.2	45	33.1
	Logistic regression	41.67	38.5	41.7	38.8
	KNN (k=3)	43.3	43.1	43.3	39.1
	Decision tree (C4.5)	51.67	28	51.7	36.3
	ANN	43.33	41.7	43.3	41.8
	Random forest	35	35.2	35	34.6
	FRBS 2: RP3-SP-WM (13.8 rules)	48.35	47.01	48.35	46.64
Liver function dataset	Naïve Bayes	53.3	58.9	53.3	48.9
	SVM	50	46.4	50	43.6
	Logistic regression	46.67	46.7	46.7	46.7
	KNN (k=3)	48.3	49.1	48.3	48
	Decision tree (C4.5)	48.3	47	48.3	46.6
	ANN	51.6	52.5	51.7	51.4
	Random forest	46.67	47.1	46.7	46.7
	FRBS 3: RP3-SP-WM-S (6.1 rules)	56.66	60.49	56.66	59.08
Lipid profile	Naïve Bayes	51.67	53.6	51.7	50.1
	SVM	51.67	48.2	51.7	43.1
	Logistic regression	53.3	53	53.3	52.9
	KNN (k=3)	46.67	46.1	46.7	46.2
	Decision tree (C4.5)	58.3	61.4	58.3	52.4
	ANN	50	48.7	50	47.9
	Random forest	55	54.5	55	54
	FRBS 4: RP3-SP-FDT-S (5.9 rules)	65	58.94	65	59.26
Symptoms	Naïve Bayes	56.67	56.7	56.7	56.7
	SVM	51.67	51.4	51.7	51.4
	Logistic regression	61.67	61.5	61.7	61.4
	KNN (k=3)	45	44.9	45	44.9
	Decision tree (C4.5)	60	60	60	60
	ANN	60	60.5	60	60
	Random forest	60	59.9	60	59.9
	FRBS 5: RP5-SP-FDT (52.4)	65.01	72.32	65.01	65.14
Complications	Naïve Bayes	46.67	47.3	46.7	46.5
	SVM	41.67	38.5	41.7	38.8
	Logistic regression	45	45.6	45	44.7
	KNN (k=3)	41.67	39.4	41.7	35.4
	Decision tree (C4.5)	38.3	35.7	38.3	36.2
	ANN	48.3	49.1	48.3	48
	Random forest	48.3	49.3	48.3	47.7
	FRBS 6: FDT-S (11.2 rules)	46.67	45.67	46.67	52.03

Εικόνα 2.18. Αποτελέσματα των υποσυστημάτων FRBS σε σύγκριση με αλγόριθμους μηχανικής μάθησης.

Οι μέθοδοι Kmeans και HFP ήταν αποτελεσματικές, ωστόσο η Kmeans είναι η καλύτερη μέθοδος κατάτμησης στις περισσότερες περιπτώσεις ενώ υπήρχε περίπτωση που η HFP ξεπέρασε την Kmeans. Γενικότερα, το σύστημα πέτυχε accuracy 90% και confidence 100% χρησιμοποιώντας όλα τα υποσυστήματα. Όταν απουσίαζαν οι εξετάσεις για τα επίπεδα γλυκόζης, το σύστημα επέδειξε accuracy 71,6% και

confidence 77,47%. Η απουσία τόσο της γλυκόζης όσο και των συμπτωμάτων οδήγησε σε περαιτέρω μείωση του accuracy κατά 65%. Η απουσία γλυκόζης, συμπτωμάτων και επιπλοκών επέδειξε τα χειρότερα αποτελέσματα.

2.11. Μοντέλο ML για την πρόβλεψη του LoS και της θνησιμότητας των νοσηλευόμενων ασθενών με διαβήτη και υπέρταση

Η παρούσα μελέτη αναπτύσσει ένα μοντέλο μηχανικής μάθησης για την πρόβλεψη της διάρκειας παραμονής (LoS) και της θνησιμότητας μεταξύ των νοσηλευόμενων ασθενών με T2DM και HTN καθώς η Ταϊβάν συγκαταλέγεται μεταξύ των χωρών με τα υψηλότερα ποσοστά διαβήτη τύπου 2 (T2DM) και υπέρτασης (HTN) και αυτό δημιουργεί επιβάρυνση στα νοσοκομεία και επηρεάζει τη συνολική διαχείριση και διοίκηση των νοσοκομείων [12].

Χρησιμοποιώντας την Εθνική Ερευνητική Βάση Δεδομένων Ασφάλισης Υγείας της Ταϊβάν (NHIRD), αυτή η μελέτη περιελάμβανε 58.618 ασθενείς, όπου 25.868 είχαν T2DM, 32.750 είχαν HTN και 6419 είχαν τόσο T2DM όσο και HTN. Αφού έγινε ανάλυση των δεδομένων με διαφορετικά μοντέλα μηχανικής μάθησης για την πρόβλεψη της LoS και της θνησιμότητας. Η αξιολόγηση έγινε με τη σχεδίαση περιγραφικών στατιστικών γραφημάτων, τη σημασία των χαρακτηριστικών, την καμπύλη ακρίβειας-ανάκλησης, τα διαγράμματα ακρίβειας και την AUC. Τα δεδομένα εκπαίδευσης και δοκιμής ορίστηκαν σε αναλογία 8:2 πριν από την εφαρμογή των αλγορίθμων.

Model	MSE	RMSE	MAE	R2
SVM	0.393	0.510	0.121	0.486
LR	0.570	0.755	0.065	0.172
GBM	0.584	0.755	0.004	0.397
XGBoost	0.312	0.386	0.123	0.633
RF	0.261	0.401	0.027	0.591

Εικόνα 2.19. Απόδοση πρόβλεψης LoS των διαφόρων μοντέλων.

Classifier	Accuracy Score	Balanced Accuracy Score	Test Score	Precision	Recall	AUC	AUPR
LoR	0.9779	0.9719	0.9728	0.9432	0.9786	0.97	0.93
RC	0.9736	0.9592	0.9692	0.9312	0.9463	0.94	0.89
SVM	0.7899	0.7562	0.7332	0.7599	0.6524	0.88	0.89

Εικόνα 2.20. Επιδόσεις πρόβλεψης θνησιμότητας διαφόρων μοντέλων

Βλέπουμε ότι, ο XGBoost παρουσίασε την καλύτερη απόδοση στην πρόβλεψη της LoS με τιμές μετρικών $R^2=0,633$, $RMSE=0,386$ και $MAE=0,123$, ενώ ο Random Forest είχε ελαφρώς χαμηλότερη απόδοση με τιμές $R^2=0,591$, $RMSE=0,401$ και $MAE=0,027$ (Εικόνα 2.19) Η λογιστική παλινδρόμηση (LoR) είχε την καλύτερη απόδοση στην πρόβλεψη της θνησιμότητας με τιμές $CV\ Score=0,9779$, $Test\ Score=0,9728$, $Precision=0,9432$, $Recall=0,9786$, $AUC=0,97$ και $AUPR=0,93$, ενώ ακολούθησε με μικρή διαφορά ο ταξινομητής Ridge με τιμές $CV\ Score=0,9736$, $Test\ Score=0,9692$, $Precision=0,9312$, $Recall=0,9463$, $AUC=0,94$ και $AUPR=0,89$ (Εικόνα 2.20).

2.12. Αξιολόγηση υπηρεσιών φροντίδας διαβητικών, της διάρκειας παραμονής και του 30ήμερου ποσοστού επανεισαγωγής.

Η εργασία πραγματεύεται το κατά πόσον η συνδιαχείριση ασθενών με διαβήτη από την ομάδα ενδονοσοκομειακών υπηρεσιών διαχείρισης του διαβήτη (IDMS) μειώνει τη διάρκεια παραμονής στο νοσοκομείο (LoS) και το ποσοστό επανεισδοχής 30 ημερών (30DR) [13].

Η ανάλυση έγινε μεταξύ ασθενών με διαβήτη που εισήχθησαν σε κοινοτικό νοσοκομείο. Η ιατρική ομάδα IDMS αποτελούνταν από ενδοκρινολόγο, νοσηλεύτη και εκπαιδευτή διαβήτη. Η ομάδα σύγκρισης αποτελούνταν από νοσηλευόμενους ασθενείς με διαβήτη υπό τη συνήθη φροντίδα των θεραπόντων ιατρών. Η σχέση μεταξύ των ομάδων μελέτης και των μεταβλητών έκβασης αξιολογήθηκε με τη χρήση μοντέλων γενικευμένης εκτιμητικής εξίσωσης (Generalized Estimating Equation models).

	Model 1 *		Model 2 **		Model 3 †	
	n = 4654		n = 4576		n = 4576	
	Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value
Study Group						
IDMS team	0.4665 (0.081)	<0.0001	0.4769 (0.081)	<0.0001	0.3456 (0.069)	<0.0001
Study Month	-0.0003 (0.005)	0.953	0.0006 (0.005)	0.887	0.0037 (0.004)	0.308
IDMS team over time	-0.0173 (0.008)	0.025	-0.0170 (0.008)	0.028	-0.0130 (0.007)	0.046

Length of stay of study patients. * Model 1 = Study Group, Study Month, and Interaction term between Study Group and Study Month only; ** Model 2 = Model 1 + Gender, Age, Race, and Ethnicity; † Model 3 = Model 2 + Number of Comorbidities, Diabetes Type, and Disease Severity; SE, standard error; IDMS, inpatient diabetes management service.

Εικόνα 2.21. Ανάλυση της διάρκειας παραμονής

	Model 1 *		Model 2 **		Model 3 †	
	n = 4654		n = 4576		n = 4576	
	Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value
Study Group						
IDMS team	0.6999 (0.238)	0.003	0.7380 (0.239)	0.002	0.7190 (0.242)	0.003
Study Month	-0.0002 (0.012)	0.986	0.0006 (0.012)	0.961	0.0068 (0.012)	0.571
IDMS team over time	-0.0430 (0.024)	0.071	-0.0443 (0.024)	0.063	-0.0472 (0.024)	0.048

30-Day Readmissions of study patients. * Model 1 = Study Group, Study Month, and Interaction term between Study Group and Study Month only; ** Model 2 = Model 1 + Gender, Age, Race, and Ethnicity; † Model 3 = Model 2 + Number of Comorbidities, Diabetes Type, and Disease Severity; SE, standard error; IDMS, inpatient diabetes management service.

Εικόνα 2.22 . Ανάλυση των 30 ημερών επανεισαγωγών

Βρέθηκε πώς, 4.654 ασθενείς με διαβήτη ($70,8 \pm 0,2$ ετών) εισήχθησαν μεταξύ Ιανουάριο 2016 και Μάιο 2017. Η ομάδα IDMS συνδιαχειρίστηκε το 18,3% των ασθενών όπου η νόσος ήταν σε πιο βαριά κατάσταση ($p < 0,0001$). Η μέση διάρκεια νοσηλείας (LOS) στους παραπάνω ασθενείς μειώθηκε κατά 27%. Το μέσο LoS μειώθηκε με την πάροδο του χρόνου στην ομάδα IDMS ($p = 0,046$), ενώ δεν παρατηρήθηκε σημαντική μείωση στην ομάδα σύγκρισης (Εικόνα 2.21). Ο μέσος όρος 30DR στους ασθενείς που συνδιαχειρίστηκε η ομάδα IDMS μειώθηκε κατά 10,71%. Η διάμεσος (median) 30DR μειώθηκε μεταξύ των ασθενών που συν-χειρίστηκαν από το IDMS παίρνοντας τιμή $p = 0,048$ (Εικόνα 2.22). Προκύπτει πως η διάρκεια παραμονής και η 30DR μειώθηκαν σημαντικά στους ασθενείς που συν-διαχειρίστηκαν από μια εξειδικευμένη ομάδα διαβήτη. Αυτές οι αλλαγές θα οδηγούσαν σίγουρα σε σημαντική εξοικονόμηση κόστους.

2.13. Πρόβλεψη της διάρκειας παραμονής στο νοσοκομείο χρησιμοποιώντας μια γαλλική ιατρική βάση δεδομένων

Η μελέτη επικεντρώνεται στη διερεύνηση των μοντέλων μηχανικής μάθησης που προβλέπουν καλύτερα τη παρατεταμένη διάρκεια παραμονής στο νοσοκομείο PLoS καθώς είναι μια σημαντική ένδειξη επιδείνωσης της αποτελεσματικότητας της ποιότητας φροντίδας στους ασθενείς [14].

Το Dataset συλλέχθηκε από τη γαλλική βάση δεδομένων Ιατρικής-Διοίκησης (PMSI) μελετώντας τα εξιτήρια που πάρθηκαν το έτος 2015 από ένα μεγάλο πανεπιστημιακό νοσοκομείο στην Γαλλία (APHM). Το LoS έγινε δυαδική μεταβλητή (μακροχρόνια vs. βραχυχρόνια) σύμφωνα με το 90ο εκατοστημόριο (14 ημέρες). Εφαρμόστηκαν οι αλγόριθμοι Logistic regression (LR), classification and regression trees (CART), Random forest (RF), Gradient boosting (GB) and Neural networks (NN). Η προγνωστική απόδοση των μοντέλων αξιολογήθηκε με τη χρήση της καμπύλης AUC.

	100 sampling experiments Mean AUC
Logistic Regression (LR)	0,7947
Classification Trees (CT)	0,7858
Random Forest (RF)	0,8086
Gradient Boosting (GB)	0,8101
Neural Networks (NNET)	0,8085

Εικόνα 2.23 . Απόδοση των ρυθμισμένων ταξινομητών για 100 επαναλήψεις πειραμάτων δειγματοληψίας

Η ανάλυσή μας περιλάμβανε 73.182 νοσηλείες, εκ των οποίων 7.341 (10,0%) οδήγησαν σε PLoS. Ο ταξινομητής GB ήταν το πιο αποδοτικό μοντέλο με την υψηλότερη AUC (0,810), υπερτερώντας έναντι όλων των άλλων μοντέλων (όλες οι τιμές $p < 0,0001$). Η απόδοση των μοντέλων RF, GB και NN με την AUC να εκτείνεται από 0,808 έως 0,810, ήταν ανώτερη από εκείνη του μοντέλου LR με τιμή AUC = 0,795 με όλες τις τιμές $p < 0,0001$.

Αντίθετα, η επίδοση του LR ήταν ανώτερη από του CART με AUC = 0,786), $p < 0,0001$. Η μεταβλητή που προβλέπει περισσότερο την PLoS ήταν ο προορισμός του ασθενούς μετά τη νοσηλεία σε άλλα ιδρύματα. Το τυπικό κλινικό προφίλ αυτών των ασθενών (17,5 % του δείγματος) ήταν του ηλικιωμένου ασθενή, που εισήχθη επειγόντως, για τραύμα, νευρολογική ή καρδιαγγειακή ή προβλήματα ψυχικής υγείας, άνοιας και ημιπληγίας. Η χρήση του αλγορίθμου GB να ήταν χρήσιμη για ασθενείς που κινδυνεύουν με PLoS (Εικόνα 2.23).

ΚΕΦΑΛΑΙΟ 3^ο - Περιγραφή Προβλήματος και Δεδομένων

3.1. Εισαγωγή

Το κεφάλαιο αυτό επικεντρώνεται στη λεπτομερή περιγραφή και ανάλυση του συνόλου δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση και δοκιμή των αλγορίθμων μηχανικής μάθησης που θα εφαρμόσουμε παρακάτω για την κατασκευή των μοντέλων μας αλλά και στην περιγραφή των δύο προβλημάτων που καλούμαστε να λύσουμε. Δόθηκε αρκετή προσοχή στην επιλογή του συνόλου δεδομένων καθώς ήταν απαραίτητο να μην ήταν πολύ μικρό σε μέγεθος αλλά αρκετό για να εκπαιδευτούν οι αλγόριθμοι σε όσα περισσότερα παραδείγματα γίνεται. Έτσι το μοντέλο θα ήταν πιο αξιόπιστο. Επίσης, ούτε ο αριθμός των χαρακτηριστικών θα έπρεπε να είναι πολύ μικρός για τον ίδιο λόγο. Φυσικά δεν υπάρχει κατάλληλο σύνολο δεδομένων σε πραγματικά δεδομένα. Κάποιες φορές, ένα μικρό σύνολο δεδομένων μπορεί να είναι πιο αξιολογίσιμο από ένα μεγαλύτερο και αυτό έγκειται στην ποιότητα των χαρακτηριστικών.

3.3. Περιγραφή προβλήματος και δεδομένων

Ο στόχος της πτυχιακής εργασίας είναι η κατασκευή δύο μοντέλων μηχανικής μάθησης. Το πρώτο θα είναι ένα μοντέλο ταξινόμησης όπου θα προβλέπει αν ένας διαβητικός ασθενής εισέρθει ξανά στο νοσοκομείο σε λιγότερο από 30 ημέρες ή σε περισσότερο από 30 ημέρες ή δεν θα ξαναμπει. Το δεύτερο θα είναι ένα μοντέλο παλινδρόμησης όπου θα προβλέπει τη διάρκεια παραμονής του διαβητικού ασθενή στο νοσοκομείο.

Το σύνολο δεδομένων "Diabetes 130-Hospitals Dataset" αποτελεί κλινική περίθαλψη διάρκειας δέκα ετών (1999-2008) σε 130 νοσοκομεία και ολοκληρωμένα δίκτυα παροχής υπηρεσιών στις ΗΠΑ. Ο Strack και οι λοιποί ομο-συντάκτες, χρησιμοποίησαν τα δεδομένα για να διερευνήσουν τον αντίκτυπο της μέτρησης της HbA1c στα ποσοστά επανεισαγωγής στο νοσοκομείο. Τα δεδομένα συλλέχθηκαν συγκεκριμένα από τη βάση δεδομένων Health Facts, η οποία είναι μια εθνική αποθήκη δεδομένων στις Ηνωμένες Πολιτείες. Μόλις οι παραπάνω ολοκλήρωσαν την έρευνά τους, το σύνολο δεδομένων υποβλήθηκε στο αποθετήριο μηχανικής μάθησης του UCI, έτσι ώστε να καταστεί διαθέσιμο για μεταγενέστερη χρήση. Περιλαμβάνει 101,766 παραδείγματα και πάνω από 50 χαρακτηριστικά που αντιπροσωπεύουν τα αποτελέσματα των ασθενών και των

νοσοκομείων. Οι πληροφορίες εξήχθησαν από τη βάση δεδομένων για τις περιπτώσεις που πληρούσαν τα ακόλουθα κριτήρια.

- Πρόκειται για μια ενδονοσοκομειακή συνάντηση (εισαγωγή σε νοσοκομείο).
- Πρόκειται για διαβητική συνάντηση, δηλαδή μια συνάντηση κατά την οποία οποιοδήποτε είδος διαβήτη καταχωρήθηκε στο σύστημα ως διάγνωση.
- Η διάρκεια παραμονής ήταν τουλάχιστον 1 ημέρα και το πολύ 14 ημέρες.
- Κατά τη διάρκεια της συνάντησης πραγματοποιήθηκαν εργαστηριακές εξετάσεις.
- Κατά τη διάρκεια της συνάντησης χορηγήθηκαν φάρμακα.

Τα δεδομένα περιέχουν χαρακτηριστικά όπως ο αριθμός του ασθενούς, η φυλή, το φύλο, η ηλικία, ο τύπος εισαγωγής, ο χρόνος παραμονής στο νοσοκομείο, η ιατρική ειδικότητα του γιατρού που τον δέχθηκε, ο αριθμός των εργαστηριακών εξετάσεων που πραγματοποιήθηκαν, το αποτέλεσμα της εξέτασης HbA1c, η διάγνωση, ο αριθμός των φαρμάκων, τα διαβητικά φάρμακα, ο αριθμός των επισκέψεων σε εξωτερικά ιατρεία, σε νοσοκομεία και σε επείγοντα περιστατικά κατά το έτος πριν από τη νοσηλεία κ.λπ. Στον παρακάτω πίνακα απεικονίζεται ο κατάλογος των χαρακτηριστικών αλλά και η λεπτομερή περιγραφή τους (Πίνακας 3.1).

Όνομα χαρακτηριστικού	Τύπος	Περιγραφή και τιμές	% λείπουν
Encounter ID	Αριθμητικό	Μοναδικό αναγνωριστικό ενός περιστατικού	0%
Patient number	Αριθμητικό	Μοναδικό αναγνωριστικό ασθενούς	0%
Race	Ονομαστικό	Τιμές: Καυκάσιος, Ασιάτης, Αφροαμερικανός, Ισπανός και άλλοι	2%
Gender	Ονομαστικό	Τιμές: άνδρας, γυναίκα και άγνωστος/άκυρος	0%
Age	Ονομαστικό	Ομαδοποίηση σε διαστήματα 10 ετών: [0, 10), [10, 20), . . . , [90, 100)	0%
Weight	Αριθμητικό	Βάρος σε κιλά	97%
Admission type	Ονομαστικό	Ακέραιο αναγνωριστικό που αντιστοιχεί σε 9 διαφορετικές τιμές, για παράδειγμα, έκτακτο, επείγον, προαιρετικό, νεογέννητο και μη διαθέσιμο	0%
Discharge disposition	Ονομαστικό	Ακέραιο αναγνωριστικό που αντιστοιχεί σε 29 διαφορετικές τιμές, για παράδειγμα, πήρε εξιτήριο, εξέπνευσε και δεν είναι διαθέσιμο	0%
Admission source	Ονομαστικό	Ακέραιο αναγνωριστικό που αντιστοιχεί σε 21 διαφορετικές τιμές, για παράδειγμα, παραπομπή από ιατρό, επείγοντα περιστατικά και μεταφορά από νοσοκομείο	0%
Time in hospital	Αριθμητικό	Ακέραιος αριθμός ημερών μεταξύ εισαγωγής και εξιτηρίου	0%
Payer code	Ονομαστικό	Ακέραιο αναγνωριστικό που αντιστοιχεί σε 23 διαφορετικές τιμές, για παράδειγμα, Blue Cross\Blue Shield, Medicare και αυτοπληρωμή	52%

Medical specialty	Ονομαστικό	Ακέραιο αναγνωριστικό της ειδικότητας του ιατρού που εισήχθη, που αντιστοιχεί σε 84 διαφορετικές τιμές, για παράδειγμα, καρδιολογία, εσωτερική ιατρική, οικογενειακή/γενική πρακτική και χειρουργός	53%
Number of lab procedures	Αριθμητικό	Αριθμός εργαστηριακών εξετάσεων που πραγματοποιήθηκαν κατά τη διάρκεια της συνάντησης	0%
Number of procedures	Αριθμητικό	Αριθμός διαδικασιών (εκτός από εργαστηριακές εξετάσεις) που πραγματοποιήθηκαν κατά τη διάρκεια της συνάντησης	0%
Number of medications	Αριθμητικό	Αριθμός διαφορετικών γενικών ονομάτων που χορηγήθηκαν κατά τη διάρκεια της συνάντησης	0%
Number of outpatient visits	Αριθμητικό	Αριθμός επισκέψεων του ασθενούς στα εξωτερικά ιατρεία κατά το έτος που προηγείται της συνάντησης	0%
Number of emergency visits	Αριθμητικό	Αριθμός επείγουσών επισκέψεων του ασθενούς κατά το έτος που προηγείται της συνάντησης	0%
Number of inpatient visits	Αριθμητικό	Αριθμός ενδονοσοκομειακών επισκέψεων του ασθενούς κατά το έτος που προηγείται της συνάντησης	0%
Diagnosis 1	Ονομαστικό	Η πρωταρχική διάγνωση (κωδικοποιημένη ως τα τρία πρώτα ψηφία του ICD9) με 848 διαφορετικές τιμές	0%
Diagnosis 2	Ονομαστικό	Δευτερεύουσα διάγνωση (κωδικοποιημένη ως τα τρία πρώτα ψηφία του ICD9) με 923 διαφορετικές τιμές	0%
Diagnosis 3	Ονομαστικό	Πρόσθετη δευτερεύουσα διάγνωση (κωδικοποιημένη ως τα τρία πρώτα ψηφία του ICD9) με 954 διαφορετικές τιμές	1%
Number of diagnoses	Αριθμητικό	Αριθμός διαγνώσεων που καταχωρήθηκαν στο σύστημα	0%
Glucose serum test result	Ονομαστικό	Υποδεικνύει το εύρος του αποτελέσματος ή αν δεν έγινε η εξέταση. Τιμές: ">200", ">300", "φυσιολογικό" και "κανένα" εάν δεν μετρήθηκε	0%
A1c test result	Ονομαστικό	Δείχνει το εύρος του αποτελέσματος ή αν δεν έγινε η εξέταση. Τιμές: ">8" εάν το αποτέλεσμα ήταν μεγαλύτερο από 8%, ">7" εάν το αποτέλεσμα ήταν μεγαλύτερο από 7% αλλά μικρότερο από 8%, "φυσιολογικό" εάν το αποτέλεσμα ήταν μικρότερο από 7% και "κανένα" εάν δεν μετρήθηκε	0%
Change of medications	Ονομαστικό	Δείχνει εάν υπήρξε αλλαγή στα διαβητικά φάρμακα (είτε δοσολογία είτε γενική ονομασία). Τιμές: "αλλαγή" και "καμία αλλαγή"	0%
Diabetes medications	Ονομαστικό	Δείχνει εάν υπήρξε συνταγογράφηση διαβητικών φαρμάκων. Τιμές: "ναι" και "όχι"	0%
24 features for medications	Ονομαστικό	Για τις γενικές ονομασίες: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, το χαρακτηριστικό υποδεικνύει εάν το φάρμακο συνταγογραφήθηκε ή υπήρξε αλλαγή στη δοσολογία. Τιμές: "up" εάν η δοσολογία αυξήθηκε κατά τη διάρκεια της συνάντησης, "down" εάν η δοσολογία μειώθηκε, "steady" εάν η δοσολογία δεν άλλαξε και "no" εάν το φάρμακο δεν συνταγογραφήθηκε	0%

Readmitted	Ονομαστικό	Ημέρες έως την επανεισαγωγή σε νοσοκομείο. Τιμές: "<30" εάν ο ασθενής επανεισήχθη σε λιγότερο από 30 ημέρες, ">30" εάν ο ασθενής επανεισήχθη σε περισσότερες από 30 ημέρες, και "Όχι" εάν δεν υπάρχει καταγραφή επανεισδοχής	0%
------------	------------	--	----

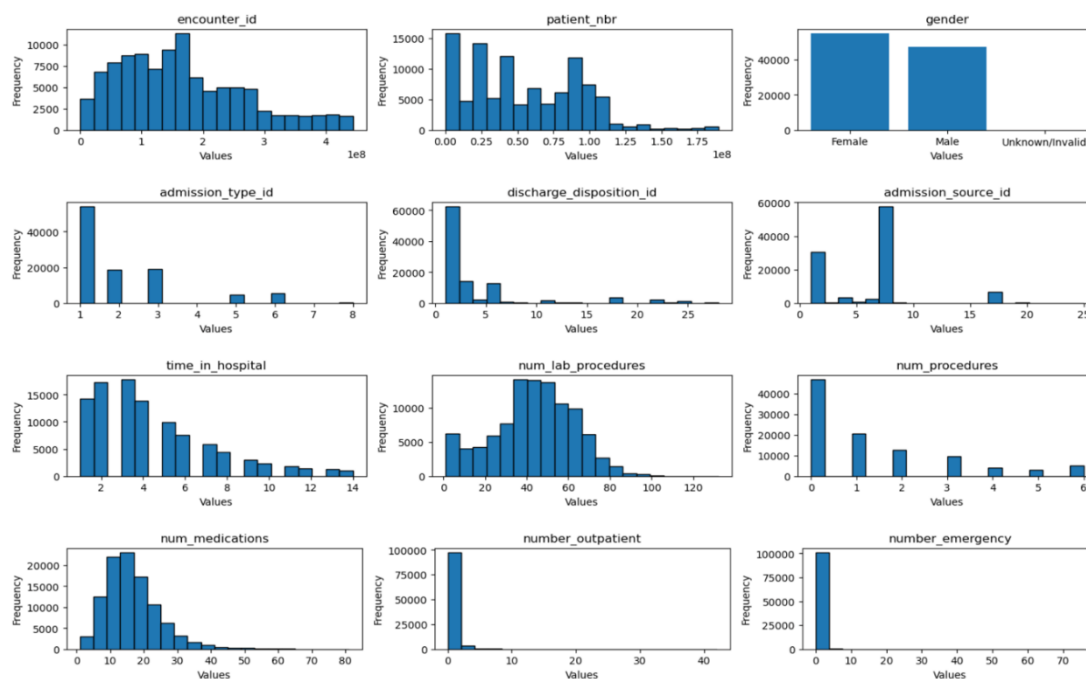
Πίνακας 3.1. Κατάλογος και περιγραφή χαρακτηριστικών

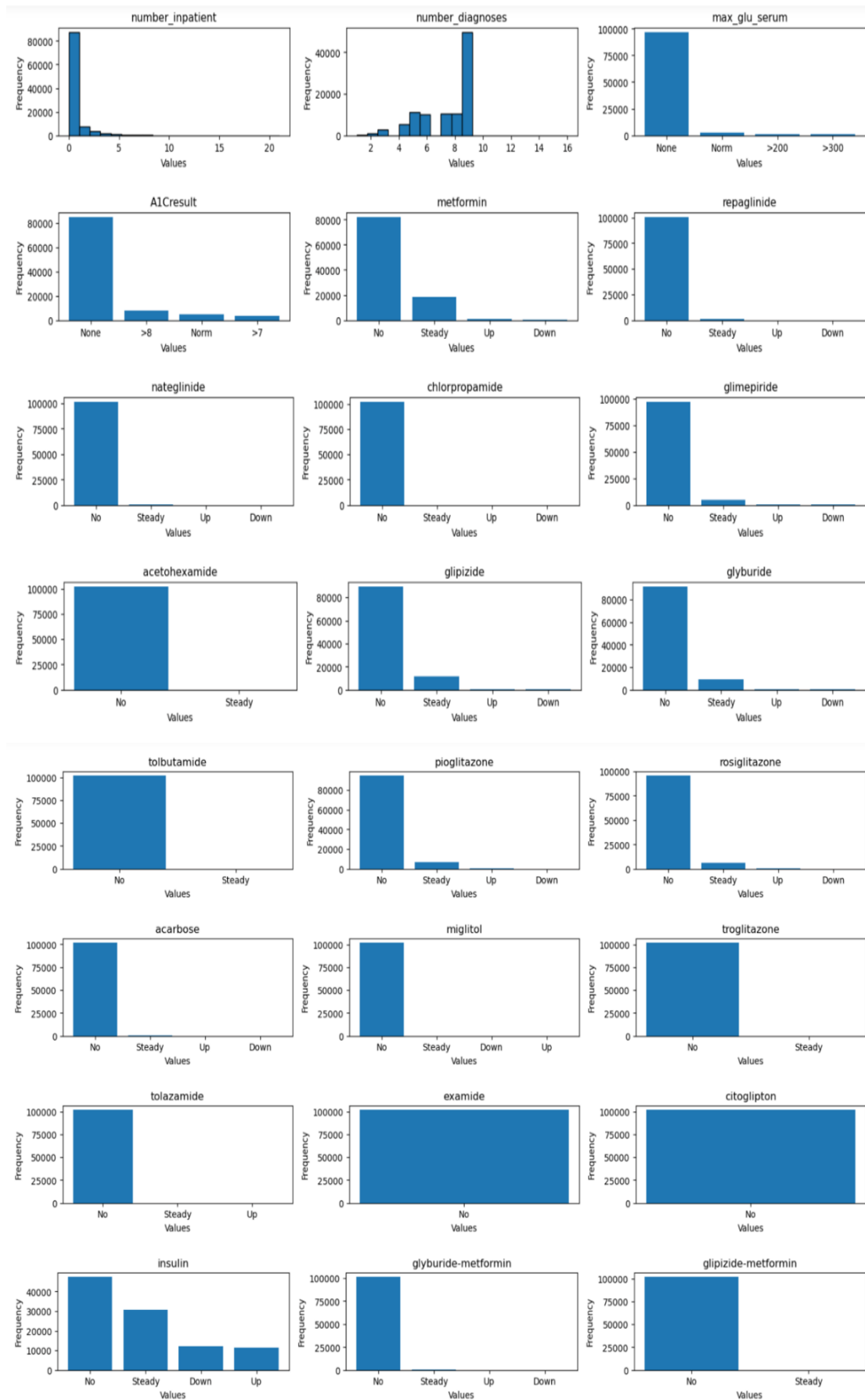
Τα χαρακτηριστικά-στόχοι του συνόλου δεδομένων είναι:

A) Το χαρακτηριστικό “readmitted” για την πρόβλεψη του εάν ο ασθενής έχει υποστεί επανεισαγωγή στο νοσοκομείο και αν αυτή έχει συμβεί εντός διαστήματος μικρότερου των 30 ημερών ή μεγαλύτερου των 30 ημερών.

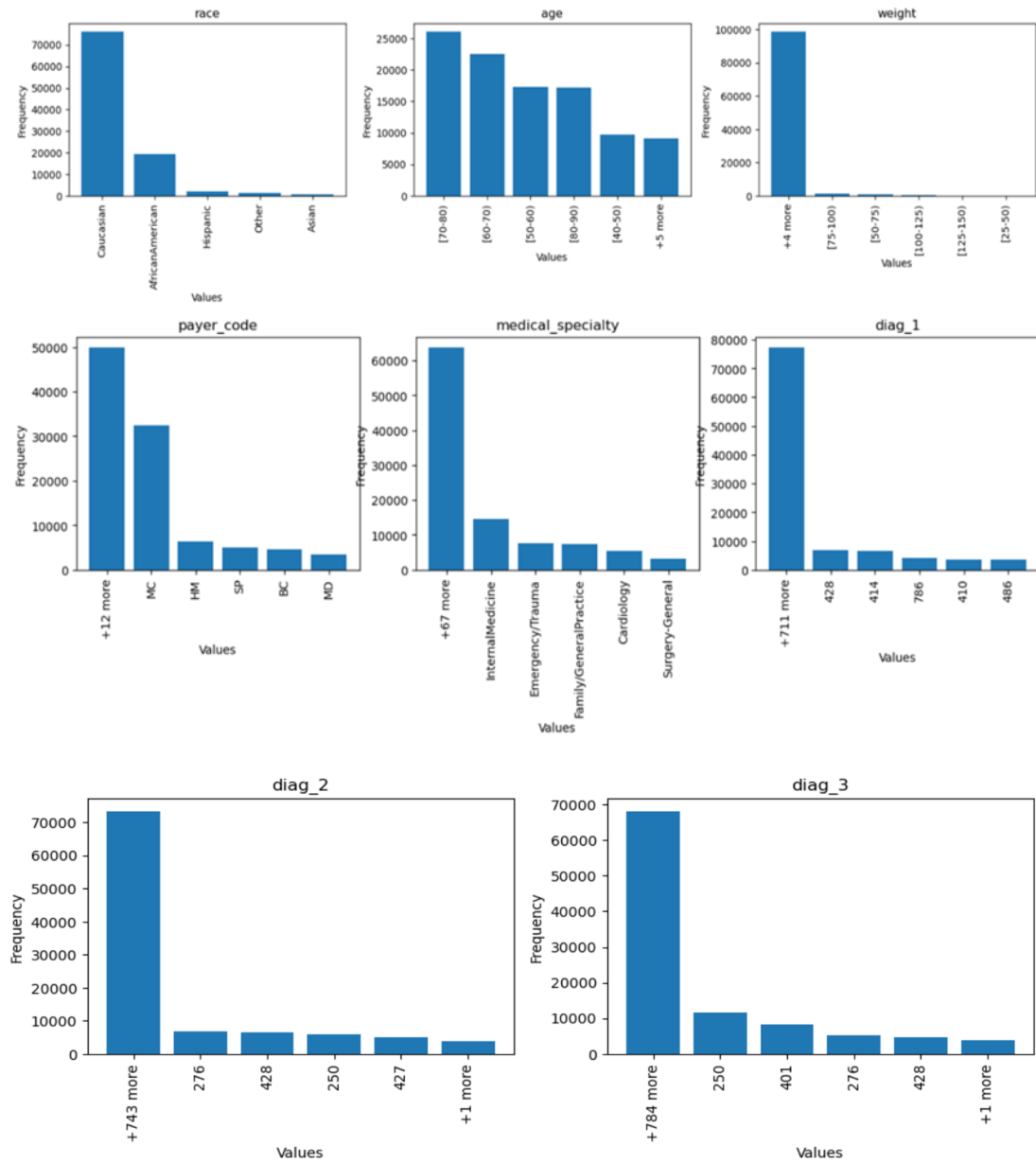
B) Το χαρακτηριστικό "time_in_hospital" για τη μέτρηση του χρόνου παραμονής του ασθενούς στο νοσοκομείο.

Στο ακόλουθο Σχήμα 3.1, παρουσιάζεται η κατανομή των χαρακτηριστικών του συνόλου δεδομένων.





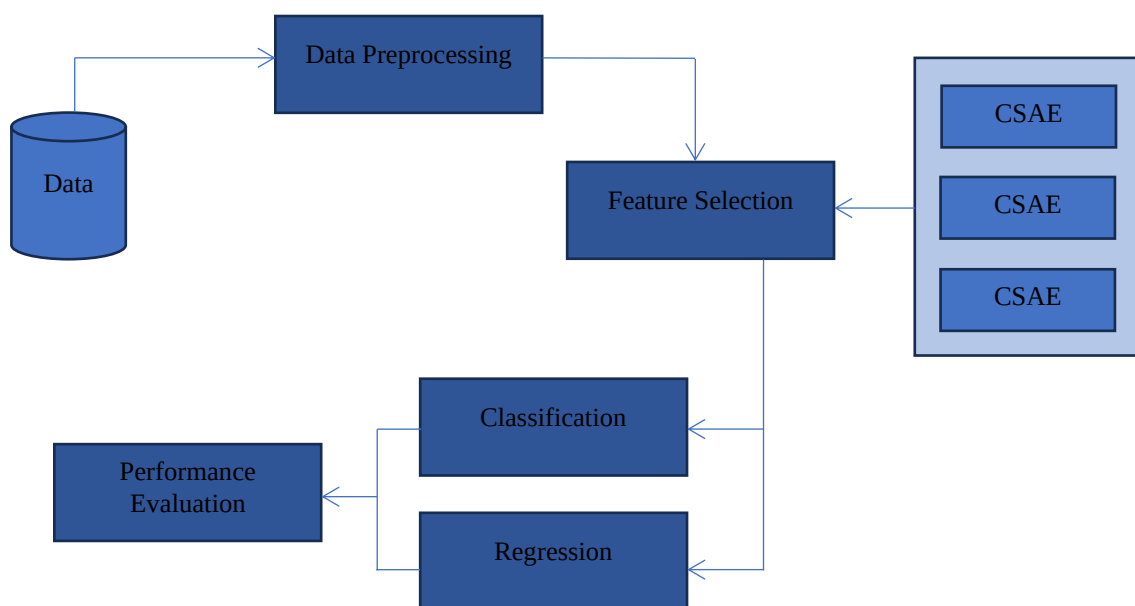




Σχήμα 3.1. Κατανομή χαρακτηριστικών

ΚΕΦΑΛΑΙΟ 4^ο - Μεθοδολογία

Το κεφάλαιο της μεθοδολογίας περιλαμβάνει τα πάντα σχετικά με την προεπεξεργασία των δεδομένων, τη διαχείριση των χαρακτηριστικών, τη δημιουργία του μοντέλου και την αξιολόγηση. Η έρευνα ακολουθεί την ακόλουθη μεθοδολογία που παρουσιάζεται στο σχήμα 4.1 :



Σχήμα 4.1. Μεθοδολογία του μέρους της ταξινόμησης

4.1. Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων είναι απαραίτητη για την κατασκευή οποιουδήποτε μοντέλου μηχανικής μάθησης για την επεξεργασία των δεδομένων σύμφωνα με τον τύπο του προβλήματος, ενώ ο κατάλληλος καθαρισμός των δεδομένων αυξάνει τις πιθανότητες επίτευξης καλύτερου αποτελέσματος. Στην εργασία αυτή, αναλύουμε ένα πρόβλημα ταξινόμησης (classification) και ένα πρόβλημα παλινδρόμησης (regression). Συνεπώς, η διαδικασία επεξεργασίας των δεδομένων θα πρέπει να χωριστεί σε δύο μέρη για το κάθε πρόβλημα ξεχωριστά. Η προεπεξεργασία γίνεται με τις ακόλουθες προσεγγίσεις:

4.1.1 Ταξινόμηση

- **Αφαίρεση περιττών χαρακτηριστικών:** Τα χαρακτηριστικά "encounter_id" και "patient_nbr" έχουν αφαιρεθεί επειδή δεν παρέχουν κάποια χρήσιμη πληροφορία για την κατασκευή του μοντέλου. Τα χαρακτηριστικά "examide" και "citoglipton" έχουν επίσης αφαιρεθεί επειδή έχουν μία μόνο τιμή.
- **Μετατροπή χαρακτηριστικών από αλφαριθμητικά σε ονομαστικά:** Ορισμένα από τα χαρακτηριστικά: diag_1, diag_2, diag_3 είναι χαρακτηριστικά συμβολοσειράς. Για την περαιτέρω επεξεργασία τους πρέπει να μετατραπούν σε ονομαστικά. Η μετατροπή γίνεται με τη χρήση του φίλτρου "StringToNominal" στο Weka. Συγκεκριμένα, η παράμετρος "attribute range" πρέπει να οριστεί σε 14-16 που είναι ο αριθμός των ακόλουθων γνωρισμάτων diag_1, diag_2, diag_3. Έτσι, τα χαρακτηριστικά είναι πλέον δυνατό να χρησιμοποιηθούν.
- **Χειρισμός ελλιπών τιμών/διπλοτύπων:** Ορισμένα από τα χαρακτηριστικά έχουν ελλιπή παραδείγματα. Τα χαρακτηριστικά είναι τα εξής: race (2%), weight (97%), payer_code(40%), medical_specialty (49%), diag_1 (0%), diag_2 (0%), diag_3(1%). Από τα παραπάνω έχουν αφαιρεθεί τα χαρακτηριστικά weight, payer_code και medical_specialty επειδή έχουν υψηλότερο ποσοστό ελλιπών τιμών. Τα άλλα χαρακτηριστικά αντιμετωπίζονται με τη χρήση του φίλτρου "ReplaceMissingValues" του λογισμικού Weka. Το φίλτρο αντικαθιστά τις ελλειπόμενες τιμές με τη μέση τιμή για τα αριθμητικά χαρακτηριστικά και την επικρατούσα τιμή για τα ονομαστικά χαρακτηριστικά. Όλες οι περιπτώσεις του συνόλου δεδομένων είναι μοναδικές, επομένως, δεν χρειάζεται να σκεφτούμε τις διπλές περιπτώσεις.
- **Μετατροπή σε ονομαστικό τύπο:** Ορισμένα από τα χαρακτηριστικά είναι διακριτά, αλλά ο τύπος δεδομένων είναι αριθμητικός. Καθώς το λογισμικό Weka μπορεί να χειριστεί ονομαστικά χαρακτηριστικά στον αλγόριθμο ταξινόμησης, αυτά τα χαρακτηριστικά έχουν μετατραπεί σε ονομαστικά. Τα διακριτά χαρακτηριστικά που εξετάστηκαν είναι τα εξής: 'admission_type_id', 'discharge_disposition_id', 'admission_source_id', 'time_in_hospital',

'num_procedures'. Για τη μετατροπή χρησιμοποιήθηκε το φίλτρο NumericToNominal. Για την εφαρμογή του φίλτρου, πρέπει να ρυθμιστεί η παράμετρος "attribute indices" για να καθοριστεί ποια χαρακτηριστικά μετατρέπονται.

- **Χειρισμός ανισορροπίας μεταξύ των κλάσεων:** Η κατανομή της κλάσης-στόχος (target class), η οποία είναι η μεταβλητή που προσπαθούμε να προβλέψουμε έχει προβλήματα ανισορροπίας, ο πίνακας 4.1 δείχνει την κατανομή των δεδομένων:

Τιμές	Μετρήσεις
No	54.864
>30	35.545
<30	11.357

Πίνακας 4.1. Μετρήσεις τιμών του Target class

Το ζήτημα της ανισορροπίας των κλάσεων πρέπει να ελέγχεται για την ύπαρξη ενός γενικευμένου αποτελέσματος για το μοντέλο, διαφορετικά το αποτέλεσμα θα είναι μεροληπτικό για μια συγκεκριμένη τιμή κλάσης. Το label της κλάσης μειοψηφίας "<30" έχει 11.357 παραδείγματα, τα οποία είναι αρκετά για τη δημιουργία οποιουδήποτε μοντέλου μηχανικής μάθησης. Ως εκ τούτου, είναι ασφαλές να επιλέξουμε την προσέγγιση της υποδειγματοληψίας. Η προσέγγιση υπερδειγματοληψίας παράγει εικονικά παραδείγματα και μειώνει την αξιοπιστία του μοντέλου και μπορεί να δώσει κακή πρόβλεψη σε πραγματικά δεδομένα. Η τεχνική υποδειγματοληψίας δεν έχει τέτοιου είδους προβλήματα. Το φίλτρο "SpreadSubsample" χρησιμοποιείται για την υποδειγματοληψία του συνόλου δεδομένων στο λογισμικό Weka. Ορίζουμε την παράμετρο "spread" της κατανομής με την τιμή -1 για την υποδειγματοληψία των πλειοψηφικών κλάσεων. Ως αποτέλεσμα, όλες οι κλάσεις έχουν 11.357 περιπτώσεις.

- **Τυποποίηση:** Το φίλτρο "Standardize" εφαρμόζεται στο σύνολο δεδομένων για να πραγματοποιήσει προ-επεξεργασία δεδομένων, ιδίως κλιμάκωση ή τυποποίηση χαρακτηριστικών. Αυτή η σημαντική τεχνική χρησιμοποιείται ευρέως στη μηχανική μάθηση για να διασφαλιστεί ότι όλα τα γνωρίσματα του συνόλου δεδομένων είναι ομοιόμορφα κλιμακωμένα. Η διαδικασία περιλαμβάνει την

προσαρμογή των τιμών κάθε γνωρίσματος έτσι ώστε να μοιράζονται μια κοινή κλίμακα, με μέσο όρο 0 και τυπική απόκλιση 1. Με τον τρόπο αυτό, αντιμετωπίζεται το ζήτημα των διαφορετικών κλιμάκων ή μονάδων μέτρησης μεταξύ των γνωρισμάτων, αποτρέποντας οποιοδήποτε χαρακτηριστικό να κυριαρχήσει στη διαδικασία μάθησης. Η διαδικασία τυποποίησης χρησιμοποιεί τις προεπιλεγμένες ρυθμίσεις παραμέτρων για την επίτευξη αυτού του μετασχηματισμού, προετοιμάζοντας αποτελεσματικά το σύνολο δεδομένων για περαιτέρω ανάλυση και μοντελοποίηση.

4.1.2 Παλινδρόμηση

Η αφαίρεση περιττών χαρακτηριστικών, η μετατροπή χαρακτηριστικών από αλφαριθμητικά σε ονομαστικά και ο χειρισμός ελλιπών τιμών/διπλοτύπων έγινε με τον ίδιο ακριβώς τρόπο επάνω στα ίδια χαρακτηριστικά, όπως και στην προ-επεξεργασία δεδομένων για το πρόβλημα της ταξινόμησης.

- **Επαναδειγματοληψία:** Καθώς το σύνολο δεδομένων είναι πολύ μεγάλο και έχει περίπου 101.766 παραδείγματα, δεν είναι εύκολα διαχειρίσιμο από το Weka. Ως εκ τούτου, λήφθηκε δείγμα 25% του συνόλου δεδομένων με τη χρήση του φίλτρου "Resample" για την ανάλυση του προβλήματος της παλινδρόμησης.
- **Μετατροπή από τακτικά σε δυαδικά χαρακτηριστικά:** Καθώς ο στόχος είναι να γίνει ανάλυση για παλινδρόμηση, τα ονομαστικά χαρακτηριστικά είναι μη χρήσιμα μέχρι να μετατραπούν σε δυαδικά ή αριθμητικά χαρακτηριστικά. Η δυαδική μετατροπή είναι καταλληλότερη όταν τα χαρακτηριστικά έχουν μόνο 2 κλάσεις. Καθώς όμως υπάρχουν πολλαπλές τιμές σε κάθε χαρακτηριστικό, η αριθμητική μετατροπή είναι η καλύτερη επιλογή. Το φίλτρο 'OrdinalToNumeric' χρησιμοποιείται για τη μετατροπή τους σε αριθμητικά.
- **Κανονικοποίηση:** Η μέθοδος αυτή μας επιτρέπει να αλλάξουμε τις τιμές των αριθμητικών στηλών στο σύνολο δεδομένων σε μια κοινή κλίμακα, συγκεκριμένα μεταξύ 0 και 1. Η κανονικοποίηση χρησιμοποιείται συχνά όταν κάνουμε ανάλυση παλινδρόμησης, καθώς βοηθά τη βελτιστοποίηση που βασίζεται στη κλίση κατά τη φάση της εκπαίδευσης διατηρώντας έτσι την σταθερότητα του μοντέλου. Δεν απαιτεί κάθε σύνολο δεδομένων κανονικοποίηση παρά μόνο όταν τα χαρακτηριστικά έχουν διαφορετικά εύρη τιμών. Για να την εφαρμόσουμε χρησιμοποιούμε το φίλτρο 'Normalize' στο λογισμικό Weka.

4.2. Επιλογή χαρακτηριστικών

Η επιλογή χαρακτηριστικών είναι μια διαδικασία που χρησιμοποιείται ευρέως στο χώρο της μηχανικής μάθησης και αποσκοπεί στην μείωση του αριθμού των μεταβλητών εισόδου κατά την ανάπτυξη ενός μοντέλου πρόβλεψης, κρατώντας μόνο εκείνες που εκτιμάται ότι είναι πιο χρήσιμες για την πρόβλεψη της μεταβλητής-στόχος. Οι τεχνικές επιλογής χαρακτηριστικών μειώνουν τα χαρακτηριστικά, συμβάλλοντας στην μείωση του υπολογιστικού κόστους και σε ορισμένες περιπτώσεις στην βελτίωση της απόδοσης του μοντέλου. Υπάρχουν διάφορες τεχνικές επιλογής χαρακτηριστικών που μπορούν να χρησιμοποιηθούν. Στην παρούσα εργασία, χρησιμοποιήθηκαν η μέθοδος αναλογίας-κέρδους (Gain Ratio Attribute Evaluation-GRAE), η μέθοδος επιλογής χαρακτηριστικών βασισμένη στην αλληλοσυσχέτιση (CFS Subset Attribute Evaluation-CSAE) και η μέθοδος περιτυλίγματος υποσυνόλου χαρακτηριστικών (Wrapper Subset Attribute Evaluation-WSAE).

4.2.1. Ταξινόμηση

- **Gain Ratio Attribute Evaluation.** Η μέθοδος της αξιολόγησης χαρακτηριστικών με βάση τον λόγο κέρδους είναι ένα πολύτιμο μέτρο αξιολόγησης χαρακτηριστικών που χρησιμοποιείται συνήθως σε αλγόριθμους που βασίζονται σε δέντρα αποφάσεων για την επιλογή και την ταξινόμηση χαρακτηριστικών [15]. Σκοπός του είναι να αντιμετωπίσει την ενδόμυχη προκατάληψη προς τα χαρακτηριστικά με μεγάλο αριθμό τιμών ή επιπέδων, που συχνά συναντάται στις παραδοσιακές μετρήσεις του κέρδους πληροφορίας (Information Gain). Κατά τη διαδικασία της δημιουργίας δέντρων αποφάσεων, ένα χαρακτηριστικό με πολλά επίπεδα μπορεί να φαίνεται ότι έχει υψηλότερο κέρδος πληροφορίας αποκλειστικά και μόνο λόγω της ικανότητάς του να χωρίζει τα δεδομένα σε μεγαλύτερο αριθμό υποσυνόλων. Ωστόσο, αυτό μπορεί να οδηγήσει σε υπερπροσαρμογή και σε λιγότερο σημαντικές διαιρέσεις στο δέντρο που προκύπτει. Ο λόγος κέρδους (Gain Ratio) μετριάζει αποτελεσματικά αυτό το ζήτημα λαμβάνοντας υπόψη τόσο τον αριθμό των επιπέδων ενός χαρακτηριστικού όσο και την κατανομή των ετικετών κλάσης (labels) σε κάθε επίπεδο. Με την ενσωμάτωση της Πληροφορίας Διαχωρισμού (Split Information), η οποία μετρά τις ενδεχόμενες πληροφορίες που δημιουργούνται από τη διάσπαση των δεδομένων με βάση τα επίπεδα του

χαρακτηριστικού, το Gain Ratio "τιμωρεί" τα χαρακτηριστικά με υψηλή πληθικότητα. Ως αποτέλεσμα, παρέχει μια πιο δίκαιη αξιολόγηση των χαρακτηριστικών, επιτρέποντας μια πιο ισορροπημένη και τεκμηριωμένη επιλογή των χαρακτηριστικών κατά τη διαδικασία κατασκευής του δέντρου αποφάσεων. Αυτή η βελτιωμένη προσέγγιση αξιολόγησης των χαρακτηριστικών συμβάλλει τελικά στη δημιουργία εγκυρότερων και πιο αξιόπιστων δέντρων αποφάσεων, ιδίως σε σενάρια όπου τα χαρακτηριστικά παρουσιάζουν ποικίλους αριθμούς επιπέδων.

Στην παρούσα εργασία, η GRAE εφαρμόστηκε με τη μέθοδο αναζήτησης Ranker. Ως βασικός ταξινομητής χρησιμοποιήθηκε το δέντρο απόφασης για την εύρεση του βέλτιστου αριθμού χαρακτηριστικών. Αντί να επιλεγεί ένας συγκεκριμένος αριθμός χαρακτηριστικών, η μελέτη επικεντρώνεται στην εύρεση του ιδανικότερου συνόλου χαρακτηριστικών που μπορεί να παράγει υψηλές επιδόσεις. Ως εκ τούτου, για την επιλογή της παραμέτρου του αριθμού των χαρακτηριστικών πραγματοποιήθηκε τυχαία δοκιμή και σφάλμα. Στον ακόλουθο πίνακα 4.2 παρουσιάζονται τα αποτελέσματα των πειραμάτων:

Αριθμός χαρακτηριστικών	F-measure
40	0.430
35	0.432
30	0.419
25	0.424
20	0.419
15	0.426
10	0.440
5	0.331

Πίνακας 4.2. Αριθμός επιλεγμένων χαρακτηριστικών με GRAE

Παρατηρούμε ότι, για 10 χαρακτηριστικά το μοντέλο παράγει το υψηλότερο αποτέλεσμα, αν και αυτό δε φαίνεται ιδιαίτερα ικανοποιητικό. Το τελικό σύνολο δεδομένων αποθηκεύτηκε ως σύνολο δεδομένων GRAE και θα χρησιμοποιηθεί παρακάτω για τη δημιουργία του μοντέλου.

- **CFS Subset Attribute Evaluation (CSAE):** Η μέθοδος της αξιολόγησης των χαρακτηριστικών με βάση την αλληλοσυσχέτιση (Correlation-based Feature Selection) είναι μια ευρέως διαδεδομένη μέθοδος αξιολόγησης χαρακτηριστικών στο πλαίσιο της επιλογής χαρακτηριστικών στη μηχανική μάθηση [16]. Στοχεύει στον εντοπισμό και τη διατήρηση των πιο σχετικών και κατατοπιστικών χαρακτηριστικών σε ένα σύνολο δεδομένων, μειώνοντας παράλληλα τον πλεονασμό μεταξύ τους. Η CFS αξιολογεί τα χαρακτηριστικά με βάση δύο βασικά κριτήρια: την ατομική συσχέτιση κάθε χαρακτηριστικού με την κλάση-στόχο και τον πλεονασμό μεταξύ διαφορετικών χαρακτηριστικών.

Για να ξεκινήσει η αξιολόγηση, η CFS υπολογίζει τη συσχέτιση κάθε χαρακτηριστικού με την κλάση-στόχο, αναζητώντας χαρακτηριστικά που παρουσιάζουν ισχυρή σχέση με τις ετικέτες της κλάσης. Δίνει μεγαλύτερη σημασία στα χαρακτηριστικά που μπορούν να κάνουν καλή διάκριση μεταξύ διαφορετικών κλάσεων, καθώς αυτά είναι πιο πιθανό να συμβάλουν σημαντικά στη διαδικασία μάθησης.

Επιπλέον, η CFS εξετάζει επίσης τη συσχέτιση μεταξύ των ίδιων των χαρακτηριστικών. Στοχεύει στον εντοπισμό υποσυνόλων χαρακτηριστικών που συσχετίζονται σε μεγάλο βαθμό με την κλάση-στόχο, αλλά έχουν χαμηλή συσχέτιση μεταξύ τους. Ελαχιστοποιώντας τον πλεονασμό μεταξύ των επιλεγμένων χαρακτηριστικών, η CFS διασφαλίζει ότι τα διατηρούμενα χαρακτηριστικά προσφέρουν μοναδικές και ανεξάρτητες πληροφορίες στον αλγόριθμο μάθησης.

Η CFS Subset είναι μια αποτελεσματική τεχνική για την αξιολόγηση χαρακτηριστικών, καθώς επιτυγχάνει ισορροπία μεταξύ σχετικότητας και πλεονασμού, επιτρέποντας την εξαγωγή ενός συμπαγούς συνόλου χαρακτηριστικών που βελτιστοποιούν τη διαδικασία μάθησης. Αυτό όχι μόνο ενισχύει την ακρίβεια και την ερμηνευσιμότητα του μοντέλου, αλλά μειώνει επίσης την υπολογιστική πολυπλοκότητα του αλγορίθμου μηχανικής μάθησης, καθιστώντας την, ένα πολύτιμο εργαλείο για τις διαδικασίες επιλογής χαρακτηριστικών.

Στην παρούσα εργασία, η CSAE εφαρμόζεται με τη μέθοδο αναζήτησης Πρώτα στο Καλύτερο (Best-First), η οποία επιλέγει τα 4 καλύτερα χαρακτηριστικά: "discharge_disposition_id", "number_emergency", "number_inpatient",

"diag_1". Το τελικό σύνολο δεδομένων αποθηκεύτηκε ως σύνολο δεδομένων CSAE και θα χρησιμοποιηθεί παρακάτω για τη δημιουργία του μοντέλου.

- **Wrapper Subset Attribute Evaluation:** Η μέθοδος αξιολόγησης περιτυλίσματος υποσυνόλου χαρακτηριστικών είναι μια πολύ ισχυρή μέθοδος που χρησιμοποιείται για την επιλογή χαρακτηριστικών στη μηχανική μάθηση [17]. Σε αντίθεση με τις μεθόδους φίλτρου ή διήθησης (filter methods) που αξιολογούν τα χαρακτηριστικά ανεξάρτητα από τον αλγόριθμο μάθησης, οι μέθοδοι περιτυλίσματος ή ενσωμάτωσης (wrapper methods) ενσωματώνουν τον πραγματικό αλγόριθμο μάθησης στη διαδικασία αξιολόγησης. Αυτό σημαίνει ότι η wrapper subset αξιολογεί υποσύνολα χαρακτηριστικών εκπαιδεύοντας και δοκιμάζοντας το μοντέλο επαναληπτικά σε διαφορετικούς συνδυασμούς χαρακτηριστικών.

Η συγκεκριμένη μέθοδος ακολουθεί μια προσέγγιση βασισμένη στην αναζήτηση, εξερευνώντας διάφορα υποσύνολα χαρακτηριστικών και αξιολογώντας την απόδοσή τους χρησιμοποιώντας τον επιλεγμένο αλγόριθμο μάθησης. Χρησιμοποιεί στρατηγικές αναζήτησης, όπως η εμπρόσθια επιλογή (forward selection), η οπισθοδρομική εξάλειψη (backward elimination) ή ένας συνδυασμός και των δύο (bidirectional selection), για την αποτελεσματική διάσχιση του χώρου των χαρακτηριστικών και τον εντοπισμό του πιο σχετικού υποσυνόλου.

Κατά τη διάρκεια κάθε επανάληψης, ο αλγόριθμος μάθησης εκπαιδεύεται και δοκιμάζεται στο σύνολο δεδομένων χρησιμοποιώντας ένα συγκεκριμένο υποσύνολο χαρακτηριστικών. Η μετρική αξιολόγησης, που συνήθως είναι η ακρίβεια (accuracy) ή άλλα μέτρα απόδοσης, χρησιμοποιείται για τη μέτρηση της απόδοσης του μοντέλου. Η διαδικασία αναζήτησης συνεχίζεται έως ότου βρεθεί το βέλτιστο υποσύνολο χαρακτηριστικών, το οποίο οδηγεί στην καλύτερη απόδοση του μοντέλου.

Το υποσύνολο περιτυλίσματος έχει το πλεονέκτημα ότι αποτυπώνει τις αλληλεπιδράσεις και τις εξαρτήσεις μεταξύ των χαρακτηριστικών που αφορούν ειδικά τον επιλεγμένο αλγόριθμο μάθησης. Ωστόσο, ένα από τα μειονεκτήματα της μεθόδου περιτυλίσματος είναι το μεγάλο υπολογιστικό κόστος, ιδίως όταν έχουμε μεγάλα σύνολα δεδομένων και πολύπλοκους αλγορίθμους μάθησης.

Παρά το μειονέκτημα αυτό, η εφαρμογή τη επιτρέπει τον προσδιορισμό συνδυαζόμενων χαρακτηριστικών που βελτιστοποιούν πραγματικά τις προγνωστικές ικανότητες του μοντέλου.

Η μέθοδος περιτυλίγματος χρησιμοποιείται για κάθε έναν από τους ταξινομητές που έχουν χρησιμοποιηθεί στην παρούσα εργασία. Το τελικό σύνολο δεδομένων έχει αποθηκευτεί ως σύνολο δεδομένων WSAE και θα χρησιμοποιηθεί παρακάτω για τη δημιουργία του μοντέλου. Στον πίνακα 4.3 παρουσιάζονται τα επιλεγμένα χαρακτηριστικά.

Χαρακτηριστικά
age
discharge_disposition_id
number_outpatient
number_emergency
number_inpatient
diag_1
number_diagnoses
nateglinide
acarbose
tolazamide
glipizide-metformin
change

Πίνακας 4.3. Αριθμός επιλεγμένων χαρακτηριστικών με WSAE

4.2.2. Παλινδρόμηση

Για την ανάλυση παλινδρόμησης χρησιμοποιείται μόνο η μέθοδος **CFS Subset Evaluation**. Η τεχνική επιλογής υποσυνόλου χαρακτηριστικών με βάση την αλληλοσυσχέτιση έχει ήδη αναλυθεί στην παραπάνω ενότητα. Η CFS Subset επιλέγει

μόνο 11 χαρακτηριστικά και αυτά είναι τα εξής: "gender", "discharge disposition id", "num_lab_procedures", "num_medication", "diag_1", "diag_2", "diag_3", "number_diagnosis", "a1cresult", "repaglinide", "readmitted".

4.3. Αξιολόγηση επιδόσεων

Προκειμένου να αξιολογηθεί η αποτελεσματικότητα του μοντέλου στο σύνολο δεδομένων και να συγκριθούν τα αποτελέσματά τους, χρησιμοποιούνται διάφορες μετρικές αξιολόγησης.

4.3.1. Ταξινόμηση

Για την ταξινόμηση λαμβάνονται υπόψη οι μετρικές αξιολόγησης της ορθότητας (Precision), της Ανάκλησης (Recall) και του Αρμονικού Μέσου (F-measure ή F1-score), ενώ η Ακρίβεια (Accuracy) δεν λαμβάνεται υπόψη λόγω των προβλημάτων ανισορροπίας που υπάρχει στο σύνολο δεδομένων. Ο λόγος για τον οποίο δεν υπολογίζεται η ακρίβεια είναι η πιθανότητα παραπλάνησης όταν το σύνολο δεδομένων περιέχει ανισόρροπες κλάσεις.

Ορθότητα: Η ορθότητα είναι μια μετρική που αναφέρεται στο ποσοστό των πραγματικά θετικών προβλέψεων (True Positives-TP) σε σχέση με το συνολικό αριθμό των προβλέψεων που το μοντέλο θεωρεί ως θετικές. Αποκτά ιδιαίτερη σημασία σε σενάρια όπου τα ψευδώς θετικά αποτελέσματα (False Positives-FP), δηλαδή τα δείγματα που έχουν εσφαλμένα κατηγοριοποιηθεί ως θετικά ενώ είναι αρνητικά, έχουν σημαντικές συνέπειες ή είναι ανεπιθύμητα. Χρησιμοποιείται ως πολύτιμος δείκτης για την αξιολόγηση της ικανότητας του μοντέλου να αποφεύγει τη λανθασμένη πρόβλεψη των θετικών δειγμάτων. Ο υπολογισμός της ορθότητας περιλαμβάνει έναν συγκεκριμένο τύπο, που επιτρέπει την ακριβέστερη αξιολόγηση της απόδοσης του μοντέλου.

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

Ανάκληση: Το Recall, γνωστό και ως ευαισθησία (sensitivity), αξιολογεί την ικανότητα του μοντέλου να ανιχνεύει το ποσοστό των πραγματικά θετικών προβλέψεων σε σχέση με το συνολικό αριθμό των πραγματικά θετικών παραδειγμάτων

στα δεδομένα. Αυτή η μετρική αξιολόγησης καθίσταται ιδιαίτερα πολύτιμη σε σενάρια όπου τα ψευδώς αρνητικά αποτελέσματα, δηλαδή τα δείγματα που ταξινομούνται εσφαλμένα ως αρνητικά (False Negatives-FN) έχουν σημαντικές συνέπειες ή είναι ανεπιθύμητα. Μετρώντας την ανάκληση, μπορούμε να αξιολογήσουμε την αποτελεσματικότητα του μοντέλου στην ικανότητά του να αποφύγει τις προβλέψεις πραγματικών θετικών περιπτώσεων ως αρνητικές, εξασφαλίζοντας έτσι μια πιο ολοκληρωμένη αξιολόγηση της απόδοσής του.

$$Precision = \frac{TP}{(TP + FN)}$$

Αρμονικός Μέσος: Το F-measure, που συχνά αναφέρεται ως F1-score, είναι μια μετρική αξιολόγησης που χρησιμοποιείται στη μηχανική μάθηση για την επίτευξη ισορροπίας μεταξύ ακρίβειας και ανάκλησης. Το F-measure συνδυάζει αυτές τις δύο μετρικές σε μια ενιαία τιμή, παρέχοντας μια ολοκληρωμένη αξιολόγηση της συνολικής απόδοσης του μοντέλου. Αποδεικνύεται ιδιαίτερα χρήσιμος σε περιπτώσεις όπου η επίτευξη υψηλής ακρίβειας και ανάκλησης είναι εξίσου κρίσιμες. Υπολογίζοντας τον αρμονικό μέσο όρο (Harmonic Mean) των παραπάνω δύο μετρικών, το F-measure προσφέρει ένα αξιόπιστο μέτρο της αποτελεσματικότητας του μοντέλου όσον αφορά την επίτευξη μιας ισορροπημένης μέτρησης μεταξύ ακρίβειας και ανάκλησης, ειδικά σε περιπτώσεις ανισορροπίας μεταξύ των κλάσεων.

$$F - Measure = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

Το F-Measure έχει χρησιμοποιηθεί στο μέρος του ρύθμισης των παραμέτρων, επειδή το σύνολο δεδομένων ήταν λίγο ανισόρροπο και η ακρίβεια δεν αποτελεί καλή επιλογή για όταν έχουμε ανισόρροπα σύνολα δεδομένων.

4.3.2. Παλινδρόμηση

Για την ανάλυση παλινδρόμησης χρησιμοποιήθηκαν 3 διαφορετικές μετρικές επιδόσεων και είναι το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error-MAE), η Τετραγωνική Ρίζα του Τετραγωνικού Σφάλματος (Root Mean Squared Error-RMSE) και το Σχετικό Απόλυτο Σφάλμα (Relative Absolute Error-RAE).

Μέσο Απόλυτο Σφάλμα: Το μέσο απόλυτο σφάλμα (MAE) είναι μια μετρική που χρησιμοποιείται για τη μέτρηση της μέσης απόλυτης διαφοράς μεταξύ των προβλεπόμενων και των πραγματικών τιμών σε ένα μοντέλο παλινδρόμησης.

$$MAE = \frac{1}{n} * \sum |y_i - \bar{y}|$$

όπου n είναι ο αριθμός των σημείων δεδομένων, y_i είναι η πραγματική τιμή και $\bar{y}$ είναι η προβλεπόμενη τιμή για κάθε σημείο δεδομένων i .

Μέσο Τετραγωνικό Σφάλμα: Το μέσο τετραγωνικό σφάλμα (RMSE) είναι μια ευρέως διαδεδομένη μετρική στην παλινδρόμηση για την αξιολόγηση της ακρίβειας των προβλέψεων, μετρώντας την τετραγωνική ρίζα των μέσων τετραγωνικών διαφορών μεταξύ των προβλεπόμενων και των πραγματικών τιμών.

$$RMSE = \sqrt{\left(\frac{1}{n}\right) * \sum (y_i - \bar{y})^2}$$

όπου n είναι ο αριθμός των σημείων δεδομένων, y_i είναι η πραγματική τιμή και $\bar{y}$ είναι η προβλεπόμενη τιμή για κάθε σημείο δεδομένων i .

Σχετικό Απόλυτο Σφάλμα: Το σχετικό απόλυτο σφάλμα (RAE) είναι μια μετρική απόδοσης που χρησιμοποιείται στην παλινδρόμηση για την αξιολόγηση της ακρίβειας των προβλέψεων σε σχέση με το μέσο όρο των πραγματικών τιμών. Μετράει τη μέση απόλυτη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών, κανονικοποιημένη με το μέσο όρο των πραγματικών τιμών.

$$RAE = \left(\frac{1}{n}\right) * \sum |y_i - \hat{y}| / \left(\frac{1}{n}\right) * \sum |y_i - \bar{y}|$$

όπου n είναι ο αριθμός των σημείων δεδομένων, y_i είναι η πραγματική τιμή, $\bar{y}$ είναι η προβλεπόμενη τιμή για κάθε σημείο δεδομένων i και $\bar{y}$ είναι ο μέσος όρος των πραγματικών τιμών.

4.4. Δημιουργία μοντέλου – Επιλογή Αλγορίθμων και συντονισμός παραμέτρων

Η διαδικασία προεπεξεργασίας και επιλογής χαρακτηριστικών έχει ολοκληρωθεί και τώρα μπορούν να εφαρμοστούν τα μοντέλα μηχανικής μάθησης. Εφαρμόστηκαν 8 διαφορετικά μοντέλα για να γίνει σύγκριση μεταξύ τους και να βρεθεί το καλύτερο μοντέλο. Οι 8 αλγόριθμοι είναι οι Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Bagging, AdaBoostM1, Stacking και Voting. Για όλους τους αλγόριθμους η ρύθμιση των παραμέτρων πραγματοποιείται στο προεπεξεργασμένο σύνολο δεδομένων χρησιμοποιώντας το 66% του μέρους εκπαίδευσης(training split).

Διασταυρωμένη επικύρωση (Cross Validation): Για την αξιολόγηση της απόδοσης του αλγορίθμου, χρησιμοποιήθηκε 10-fold cross validation. Η μέθοδος της διασταυρούμενης επικύρωσης μειώνει τα προβλήματα υπερπροσαρμογής και βελτιώνει την απόδοση γενίκευσης του μοντέλου. Ως εκ τούτου, η μέθοδος αυτή είναι πιο αξιόπιστη από τη μέθοδο ποσοστιαίου διαχωρισμού (percentage split). Η προσέγγιση με διασταυρούμενη επικύρωση επιφέρει χαμηλότερη διακύμανση από εκείνη του ενός μόνο συνόλου παρακράτησης (Hold-out set), επομένως, παράγεται ένα αρκετά γενικευμένο αποτέλεσμα. Τα αποτελέσματα του cross validation είναι πιο αξιόπιστα από ένα hold-out σύνολο αποτελεσμάτων. Μειώνει επίσης τις πιθανότητες υπερπροσαρμογής.

4.4.1. Naïve Bayes

Ο Naïve Bayes είναι ένας δημοφιλής και αρκετά χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης σε προβλήματα ταξινόμησης. Βασίζεται στο θεώρημα Bayes, το οποίο αξιοποιεί τις πιθανότητες για να κάνει προβλέψεις [18]. Η "αφελής" υπόθεση στον Naïve Bayes είναι ότι τα χαρακτηριστικά είναι υπό όρους ανεξάρτητα, δηλαδή ότι κάθε χαρακτηριστικό συμβάλλει ανεξάρτητα στην τελική απόφαση ταξινόμησης. Αυτό απλοποιεί τον υπολογισμό και καθιστά τον αλγόριθμο αποδοτικό, ιδίως με σύνολα δεδομένων πολλών διαστάσεων.

Ένα από τα σημαντικότερα πλεονεκτήματα του Naïve Bayes είναι η απλότητα και η ευκολία εφαρμογής του. Μπορεί να λειτουργήσει καλά ακόμη και με περιορισμένα δεδομένα εκπαίδευσης, καθιστώντας τον κατάλληλο για σενάρια όπου δεν είναι διαθέσιμες μεγάλες ποσότητες επισημασμένων δεδομένων. Ωστόσο, η "αφελής" παραδοχή της ανεξαρτησίας των χαρακτηριστικών μπορεί να αποτελέσει περιορισμό σε ορισμένες εφαρμογές του πραγματικού κόσμου. Σε περιπτώσεις όπου τα χαρακτηριστικά συσχετίζονται, η υπόθεση αυτή μπορεί να μην ισχύει, οδηγώντας σε μειωμένη ακρίβεια. Επιπλέον, ο Naïve Bayes είναι γνωστό ότι είναι ένας "προκατειλημμένος" ταξινομητής, ευνοώντας ορισμένες κλάσεις έναντι άλλων, ειδικά όταν υπάρχει ανισορροπία στην κατανομή των κλάσεων.

Η ρύθμιση των παραμέτρων που διενεργείται για την εύρεση του καλύτερου συνόλου παραμέτρων παρουσιάζεται στον πίνακα 4.4:

Παράμετροι	F-Measure
default	0.418
Use kernel estimator = True	0.455
Use supervised discretization = True	0.461

Πίνακας 4.4. Ρύθμιση παραμέτρων του Naïve Bayes

4.4.2. Decision Tree

Το Δέντρο Αποφάσεων (DT) είναι ένα δενδροειδές μοντέλο, όπου κάθε εσωτερικός κόμβος αντιπροσωπεύει μια απόφαση βάσει ενός χαρακτηριστικού, κάθε κλαδί αντιπροσωπεύει το αποτέλεσμα της απόφασης και κάθε κόμβος φύλλου αντιστοιχεί σε μια ετικέτα κλάσης ή μια προβλεπόμενη τιμή [19].

Η δύναμη των δέντρων απόφασης έγκειται στην ερμηνευσιμότητά τους, καθώς η διαδικασία λήψης αποφάσεων απεικονίζεται οπτικά με απλό τρόπο. Μπορούν να χειριστούν τόσο αριθμητικά όσο και κατηγορικά δεδομένα και είναι σχετικά ανθεκτικά στις ακραίες τιμές. Ωστόσο, τα δέντρα αποφάσεων μπορεί να υποφέρουν από υπερπροσαρμογή, κατά την οποία απομνημονεύουν τα δεδομένα εκπαίδευσης και αποδίδουν ελάχιστα σε νέα, άγνωστα δεδομένα. Τεχνικές όπως το κλάδεμα, ο καθορισμός ενός ελάχιστου αριθμού δειγμάτων ανά φύλλο ή η χρήση μεθόδων συνόλου, όπως το Random Forest ή το Gradient Boosting, μπορούν να

χρησιμοποιηθούν για την αντιμετώπιση της υπερπροσαρμογής και τη βελτίωση της απόδοσης του αλγορίθμου.

Η διαδικασία ρύθμισης γίνεται με την παρατήρηση των παραμέτρων minNumObj και unpruned που παρουσιάζεται στον πίνακα 4.5.

Παράμετροι	F-Measure
default	0.433
Min num obj = 2	0.431
Min num obj = 5	0.433
Min num obj = 8	0.432
Min num obj = 7	0.440
Min num obj = 7, unpruned = true	0.417

Πίνακας 4.5. Ρύθμιση παραμέτρων του δέντρου αποφάσεων

4.4.3. Logistic Regression

Η λογιστική παλινδρόμηση (LR) είναι ένας ευρέως διαδεδομένος αλγόριθμος μηχανικής μάθησης για εργασίες δυαδικής ταξινόμησης (binary classification) [20]. Η απλότητα και η ερμηνευσιμότητά του τον καθιστούν κατάλληλο για βασικά προβλήματα ταξινόμησης. Ο αλγόριθμος χειρίζεται αποτελεσματικά μεγάλα σύνολα δεδομένων και παρέχει πιθανολογικές εξόδους, επιτρέποντας την καλύτερη κατανόηση των προβλέψεων. Επιπλέον, η λογιστική παλινδρόμηση μπορεί να αναδείξει τη σημασία των χαρακτηριστικών, ενισχύοντας την επιλογή και την ανάλυση χαρακτηριστικών. Ωστόσο, περιορίζεται στη δυαδική ταξινόμηση και προϋποθέτει γραμμικά όρια απόφασης, μειώνοντας ενδεχομένως την αποτελεσματικότητά του για πολύπλοκες συσχετίσεις. Ο αλγόριθμος είναι επίσης ευαίσθητος στις ακραίες τιμές και η υπόθεση της ανεξαρτησίας των χαρακτηριστικών μπορεί να μην ισχύει πάντα. Παρά τους περιορισμούς αυτούς, η λογιστική παλινδρόμηση παραμένει πολύτιμη για εργασίες δυαδικής ταξινόμησης όπου η ερμηνευσιμότητα και η αποτελεσματικότητα είναι απαραίτητες. Στον πίνακα 4.6 παρουσιάζεται η ρύθμιση των παραμέτρων.

Παράμετροι	F-Measure
default	0.437
Use AIC	0.436
Num boosting iteration = 10	0.434
Num boosting iteration = 30	0.437
Num boosting iteration = 50	0.437
Num boosting iteration = 80	0.435

Πίνακας 4.6 Ρύθμιση παραμέτρων της λογιστικής παλινδρόμησης

4.4.4. Random Forest

Το αλγόριθμος Τυχαίου Δάσους είναι ένας αλγόριθμος μάθησης συνόλου που δημιουργεί πολλαπλά δέντρα απόφασης κατά τη διάρκεια της εκπαίδευσης, όπου κάθε δέντρο εκπαιδεύεται σε ένα τυχαίο υποσύνολο χαρακτηριστικών και δειγμάτων [21]. Η τελική πρόβλεψη λαμβάνεται μέσω ενός μηχανισμού ψηφοφορίας για ταξινόμηση ή μέσου όρου για παλινδρόμηση. Το βασικό πλεονέκτημα του Random Forest είναι η ικανότητά του να μειώνει την υπερπροσαρμογή και να ενισχύει τη γενίκευση συνδυάζοντας τις προβλέψεις διαφορετικών δέντρων. Αυτό τον καθιστά ιδιαίτερα αποδοτικό στο χειρισμό πολύπλοκων συνόλων δεδομένων και στην επίτευξη ακριβών αποτελεσμάτων με ελάχιστο συντονισμό παραμέτρων.

Για τον συντονισμό του αλγορίθμου random forest έχει ληφθεί υπόψη η παράμετρος max depth (μέγιστο βάθος των ατομικών δέντρων) και numFeatures (αριθμός των τυχαία επιλεγμένων χαρακτηριστικών για τη δημιουργία κάθε δέντρου) για την ανάλυση που παρουσιάζεται στον πίνακα 4.7:

Παράμετροι	F-Measure
default	0.436
Max depth = 2	0.431
Max depth = 5	0.436

Max depth = 10	0.440
Max depth = 15	0.436
Max depth = 10, num features = 2	0.442
Max depth = 10, num features = 3	0.441
Max depth = 10, num features = 5	0.439

Πίνακας 4.7. Ρύθμιση παραμέτρων του Random Forest

4.4.5. Bagging

Το Bootstrap Aggregating (Bagging) ή αλλιώς "Σακούλιασμα" είναι μια τεχνική συνένωσης που επιτρέπει τη βελτίωση της αξιοπιστίας του μοντέλου [22]. Λειτουργεί με τη δημιουργία πολλαπλών αντιγράφων του ίδιου βασικού αλγορίθμου. Καθένα από αυτά τα αντίγραφα εκπαιδεύεται στη συνέχεια ανεξάρτητα σε διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης, χρησιμοποιώντας μια μέθοδο που ονομάζεται bootstrapping (δειγματοληψία με επανατοποθέτηση). Το κύριο πλεονέκτημα του bagging έγκειται στην ικανότητά του να μειώνει τη διακύμανση, η οποία συμβάλλει στην αποφυγή της υπερπροσαρμογής, οδηγώντας τελικά σε καλύτερες επιδόσεις γενίκευσης. Το bagging είναι ιδιαίτερα πολύτιμο σε περιπτώσεις όπου οι επιμέρους αλγόριθμοι, οι οποίοι χρησιμοποιούνται για τη δημιουργία των αντιγράφων, τείνουν να κάνουν λάθη ή να παρουσιάζουν υψηλή μεταβλητότητα στις προβλέψεις τους.

Για τον συντονισμό των παραμέτρων έχει ληφθεί υπόψη ο αριθμός των επαναλήψεων (iterations), η αναπαράσταση αντιγράφων με χρήση βαρών (represent copies using weights) και το ποσοστό των περιπτώσεων που θα συμπεριφθούν σε κάθε δείγμα bootstrap (bagSizePercent), όπως φαίνεται στον πίνακα 4.8.

Παράμετροι	F-Measure
default	0.447
Represent copies using weights = True	0.449
iteration = 50	0.444
iteration = 20	0.437

bagSizePercent = 80	0.445
bagSizePercent = 50	0.434
bagSizePercent = 40	0.449

Πίνακας 4.8. Ρύθμιση παραμέτρων του ταξινομητή bagging

4.4.6. Stacking

Το Stacking (Στοιίβαξη), μια γνωστή τεχνική μάθησης συνόλου, περιλαμβάνει το συνδυασμό πολλαπλών ταξινομητών βάσης (base classifiers) για τη βελτίωση της απόδοσης πρόβλεψης [23]. Ακολουθεί μια διαδικασία δύο βημάτων: πρώτα, οι ταξινομητές βάσης εκπαιδεύονται στα αρχικά δεδομένα- στη συνέχεια, χρησιμοποιείται ένας μετα-ταξινομητής για να μάθει από τις προβλέψεις των ταξινομητών βάσης. Ο μετα-ταξινομητής χρησιμοποιεί τις εξόδους των βασικών ταξινομητών ως χαρακτηριστικά για να κάνει την τελική πρόβλεψη. Η τεχνική Stacking αξιοποιεί την ποικιλομορφία των ταξινομητών βάσης για να ενισχύσει την ακρίβεια και την αξιοπιστία, καθιστώντας την ιδιαίτερα χρήσιμη σε απαιτητικά προβλήματα ταξινόμησης. Επιτρέπει στους χρήστες να δημιουργήσουν πιο ισχυρά μοντέλα συνδυάζοντας αποτελεσματικά τα πλεονεκτήματα των διαφορετικών ταξινομητών.

Για τον ταξινομητή στοιίβαξης έχει χρησιμοποιηθεί το προεπιλεγμένο σύνολο παραμέτρων. Αλλά η παράμετρος του ταξινομητή στοιίβαξης ελήφθη με βάση τον έλεγχο της απόδοσης των 4 κύριων ταξινομητών. Όπως είδαμε πιο πάνω, η λογιστική παλινδρόμηση είχε καλή απόδοση, επομένως, θα χρησιμοποιηθεί στον αλγόριθμο στοιίβαξης.

4.4.7. AdaBoostM1

Ο AdaBoostM1 (AdaBoost.Multi-class Version 1), είναι μια εκδοχή του αλγορίθμου AdaBoost που έχει σχεδιαστεί για προβλήματα πολυκατηγορικής (multi-class) ταξινόμησης [24]. Λειτουργεί εκπαιδεύοντας διαδοχικά μια σειρά αδύναμων ταξινομητών, αποδίδοντας μεγαλύτερα βάρη σε περιπτώσεις που ταξινομούνται εσφαλμένα σε κάθε επανάληψη. Το τελικό μοντέλο είναι ένας συνδυασμός αυτών των αδύναμων ταξινομητών, με βάρη που καθορίζονται από τις επιμέρους επιδόσεις τους. Το σημαντικότερο ίσως προτέρημα του AdaBoostM1 βρίσκεται στην ικανότητά του να

επικεντρώνεται σε δύσκολες περιπτώσεις που είναι δύσκολο να ταξινομηθούν. Η μέθοδος αυτή ενισχύει αποτελεσματικά τη συνολική ακρίβεια του ensemble μοντέλου. Μέσω επαναληπτικών προσαρμογών των βαρών των λανθασμένα ταξινομημένων περιπτώσεων, ο AdaBoostM1 μπορεί να δώσει προτεραιότητα στη μάθηση από πολύπλοκα σημεία δεδομένων, οδηγώντας σε ένα πιο ισχυρό και ακριβές μοντέλο. Ο AdaBoostM1 εφαρμόζεται και αξιολογείται με τις προεπιλεγμένες παραμέτρους χρησιμοποιώντας τον αλγόριθμο Logistic Regression.

4.4.8. Linear Regression

Η γραμμική παλινδρόμηση είναι μια στατιστική τεχνική που χρησιμοποιείται για τη μοντελοποίηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Υποθέτει μια γραμμική σχέση, στοχεύοντας στην εύρεση της γραμμής που ταιριάζει καλύτερα και ελαχιστοποιεί τη διαφορά μεταξύ παρατηρούμενων και προβλεπόμενων τιμών. Η απλότητα και η ερμηνευσιμότητα της γραμμικής παλινδρόμησης είναι τα κύρια πλεονεκτήματά της, καθιστώντας την προσιτή τόσο σε αρχάριους όσο και σε έμπειρους χρήστες. Οι συντελεστές στην εξίσωση της παλινδρόμησης έχουν σαφείς ερμηνείες, επιτρέποντάς μας να κατανοήσουμε την επίδραση κάθε ανεξάρτητης μεταβλητής στην εξαρτημένη μεταβλητή. Είναι ευέλικτη και μπορεί να εφαρμοστεί σε διάφορους τομείς για την πρόβλεψη, την πρόγνωση και τον εντοπισμό σχέσεων μεταξύ μεταβλητών. Ωστόσο, η γραμμική παλινδρόμηση έχει και περιορισμούς. Βασίζεται στην υπόθεση της γραμμικότητας, η οποία μπορεί να μην ισχύει για όλα τα σύνολα δεδομένων, οδηγώντας σε ανακριβείς προβλέψεις. Επιπλέον, είναι ευαίσθητη στις ακραίες τιμές και ενδέχεται να μην καταγράφει πολύπλοκες αλληλεπιδράσεις μεταξύ μεταβλητών. Παρά τα μειονεκτήματα αυτά, η γραμμική παλινδρόμηση παραμένει μια θεμελιώδης και ευρέως χρησιμοποιούμενη τεχνική στην ανάλυση δεδομένων και χρησιμεύει ως σημαντικό σημείο εκκίνησης για πιο προηγμένα μοντέλα παλινδρόμησης.

Η γραμμική παλινδρόμηση χρησιμοποιήθηκε με τα προεπιλεγμένα σύνολα παραμέτρων.

4.4.9. K-Nearest Neighbor Regressor

Ο K-NN (k-Πλησιέστερος γείτονας) Regressor είναι ένας μη παραμετρικός αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για εργασίες παλινδρόμησης [25]. Προβλέπει

την τιμή-στόχο ενός νέου σημείου δεδομένων με τον μέσο όρο των τιμών-στόχων των k πλησιέστερων γειτόνων του στο σύνολο εκπαίδευσης. Η απλότητα και η ευελιξία του KNN Regressor τον καθιστούν προσιτό σε αρχάριους και χρήσιμο για διάφορους τύπους συνόλων δεδομένων, αποτυπώνοντας μη γραμμικές σχέσεις μεταξύ μεταβλητών χωρίς να υποθέτει κάποια συγκεκριμένη λειτουργική μορφή. Είναι ανθεκτικός στις ακραίες τιμές, καθώς οι προβλέψεις βασίζονται στους μέσους όρους των γειτονικών σημείων δεδομένων, μειώνοντας την επιρροή των ακραίων τιμών. Ωστόσο, η υπολογιστική του ένταση μπορεί να αποτελέσει μειονέκτημα για μεγάλα σύνολα δεδομένων και η επιλογή της βέλτιστης τιμής για το " k " είναι κρίσιμη για την αποφυγή της υπερπροσαρμογής ή της υπερ-εξομάλυνσης. Επιπλέον, ο KNN Regressor αντιμετωπίζει όλα τα χαρακτηριστικά ισότιμα, καθιστώντας τον ευαίσθητο σε μη συναφή χαρακτηριστικά και υστερώντας σε μια σαφή αναπαράσταση του μοντέλου, γεγονός που μπορεί να επηρεάσει την ερμηνευσιμότητα. Η προσεκτική εξέταση των διακυμάνσεων του είναι απαραίτητη όταν επιλέγεται να χρησιμοποιείται ο συγκεκριμένος αλγόριθμος.

Με τη χρήση της παραμέτρου k (αριθμός γειτόνων), συντονίστηκε η τιμή του k , με τη βέλτιστη να είναι $k = 3$. Η διαδικασία συντονισμού των παραμέτρων παρουσιάζεται συνοπτικά στον πίνακα 4.9:

Παράμετροι	RMSE
default	3.437
$K=3$	2.804
$K=5$	2.842
$K=7$	2.890
$K=9$	2.921

Πίνακας 4.9. Συντονισμός παραμέτρων του K-NN Regressor

4.4.10. Random Forest Regressor

Ο Random Forest Regressor είναι ένας αλγόριθμος μηχανικής μάθησης συνόλου που χρησιμοποιείται για προβλήματα παλινδρόμησης. Συνδυάζει πολλαπλά δέντρα απόφασης για να κάνει πιο ακριβείς προβλέψεις δημιουργώντας ένα πλήθος δέντρων κατά τη διαδικασία εκπαίδευσης και υπολογίζοντας κατά μέσο όρο τις εξόδους τους.

Αυτή η προσέγγιση χειρίζεται αποτελεσματικά τις μη γραμμικές σχέσεις και τις αλληλεπιδράσεις μεταξύ των μεταβλητών, καθιστώντας την κατάλληλη για σύνθετα σύνολα δεδομένων με πολλά χαρακτηριστικά. Ένα από τα κύρια πλεονεκτήματα του Random Forest Regressor είναι η υψηλή ακρίβεια πρόβλεψης, καθώς η συνάθροιση των δέντρων μειώνει την υπερπροσαρμογή και τη διακύμανση του μοντέλου. Παρουσιάζει επίσης ανθεκτικότητα σε ακραίες τιμές και θορυβώδη δεδομένα. Επιπλέον, ο αλγόριθμος μπορεί να εντοπίσει αυτόματα τα πιο σχετικά χαρακτηριστικά, απλοποιώντας τη διαδικασία επιλογής χαρακτηριστικών. Ωστόσο, η υπολογιστική πολυπλοκότητα μπορεί να αποτελέσει μειονέκτημα, ιδίως για μεγάλα σύνολα δεδομένων. Η φύση του συνόλου του μοντέλου μπορεί επίσης να μειώσει την ερμηνευσιμότητα, καθώς η κατανόηση των σχέσεων μεταξύ των μεταβλητών γίνεται πρόκληση. Παρ' όλα αυτά, ο Regressor Random Forest παραμένει ένα ισχυρό και ευέλικτο εργαλείο, ικανό να χειρίζεται αποτελεσματικά ένα ευρύ φάσμα προβλημάτων παλινδρόμησης.

Η παράμετρος μέγιστο βάθος (max depth) και αριθμός χαρακτηριστικών (numFeatures) συντονίστηκε, αλλά όπως φαίνεται στον πίνακα 4.10, η προεπιλεγμένες ρυθμίσεις είχαν την καλύτερη απόδοση.

Παράμετροι	RMSE
default	2.2806
Max depth = 3	2.5783
Max depth = 5	2.5043
Max depth = 7	2.4635
Max depth = 17	2.3047
Num features = 3	2.2858
Num features = 5	2.2914

Πίνακας 4.10. Ρύθμιση παραμέτρων του Random Forest Regressor

ΚΕΦΑΛΑΙΟ 5^ο - Πειράματα και Αποτελέσματα

5.1. Πρόβλημα Ταξινόμησης

Στην εν λόγω εργασία, χρησιμοποιήθηκαν 8 διαφορετικοί ταξινομητές για το πρόβλημα της ταξινόμησης και συγκρίθηκαν μεταξύ τους για να βρεθεί ο καλύτερος. Οι Bagging, AdaBoostM1, Stacking και Voting εξετάζονται επίσης εδώ για την προσπάθεια βελτίωσης της απόδοσης του ταξινομητή και είναι μεταξύ των 8. Επίσης, για να έχουμε διάφορες ομάδες δεδομένων, χρησιμοποιήθηκαν 3 διαφορετικές τεχνικές επιλογής χαρακτηριστικών για να ελεγχθεί η δυναμική της των μεθόδων επιλογής χαρακτηριστικών και να προσδιοριστεί ποια είναι η καλύτερη ομάδα που μπορεί να χρησιμοποιηθεί για αυτά τα συγκεκριμένα δεδομένα. Στον ακόλουθο πίνακα 5.1 παρουσιάζονται τα αποτελέσματα των αλγορίθμων (ανάλυση σε επίπεδο κλάσης) του συνόλου δεδομένων GRAE.

Αλγόριθμος	Κλάση	Precision	Recall	F-measure
Naïve Bayes	NO	0.452	0.650	0.533
	>30	0.414	0.341	0.374
	<30	0.503	0.370	0.426
	Avg.	0.456	0.454	0.445
Decision Tree	NO	0.452	0.630	0.527
	>30	0.412	0.331	0.367
	<30	0.490	0.393	0.436
	Avg.	0.451	0.452	0.443
Logistic Regression	NO	0.441	0.697	0.540
	>30	0.418	0.305	0.352
	<30	0.514	0.355	0.420
	Avg.	0.458	0.452	0.438
Random Forest	NO	0.450	0.643	0.530
	>30	0.413	0.327	0.365
	<30	0.495	0.386	0.434
	Avg.	0.453	0.452	0.443

Πίνακας 5.1. Απόδοση διαφορετικών αλγορίθμων στο σύνολο δεδομένων GRAE

Οι αριθμοί με έντονα γράμματα στον παραπάνω πίνακα δείχνουν την καλύτερη απόδοση για κάθε κλάση. Στο σύνολο δεδομένων GRAE, ο Naïve Bayes πέτυχε το

υψηλότερο μέτρο $F\text{-measure}=0,445$ κατά μέσο όρο σε όλες τις κλάσεις. Παρουσίασε υποσχόμενο precision και recall για την κλάση ">30", υποδεικνύοντας την ικανότητά του να εντοπίζει σωστά τους ασθενείς με υψηλότερο κίνδυνο επανεισαγωγής μετά από 30 ημέρες. Ωστόσο, η απόδοσή του για την κατηγορία "<30" ήταν συγκριτικά χαμηλότερη. Το Decision Tree ακολούθησε σχεδόν την πορεία του Naïve Bayes όσον αφορά το $F\text{-measure}=0,443$ κατά μέσο όρο. Παρουσίασε υψηλότερα αποτελέσματα για την κλάση "<30", ενώ οι τιμές των precision και recall του για την κλάση ">30" ήταν σχετικά χαμηλότερες. Ο αλγόριθμος Logistic Regression επέδειξε ισορροπημένες τιμές precision και recall, επιτυγχάνοντας μέσο $F\text{-measure}=0,438$. Παρουσίασε τις καλύτερες επιδόσεις για την κλάση "OXI", αλλά οι τιμές των precision και recall για τις άλλες δύο κλάσεις δεν ήταν τόσο υψηλές. Ο Random Forest πέτυχε μέσο $F\text{-measure}=0,443$, παρόμοιο με το Decision Tree. Η απόδοσή του ήταν συνεπής σε όλες τις κλάσεις, με ανταγωνιστικά αποτελέσματα τόσο για την κλάση με label "<30" όσο και για την κλάση με label ">30", αλλά σίγουρα δεν ήταν τα καλύτερα για καμία από τις κλάσεις.

Αλγόριθμος	Κλάση	Precision	Recall	F-measure
Naïve Bayes	NO	0.463	0.596	0.521
	>30	0.415	0.360	0.386
	<30	0.484	0.409	0.444
	Avg.	0.454	0.455	0.450
Decision Tree	NO	0.461	0.543	0.498
	>30	0.401	0.322	0.357
	<30	0.461	0.469	0.465
	Avg.	0.441	0.445	0.440
Logistic Regression	NO	0.453	0.647	0.533
	>30	0.421	0.333	0.372
	<30	0.503	0.392	0.442
	Avg.	0.459	0.458	0.449
Random Forest	NO	0.444	0.564	0.497
	>30	0.387	0.354	0.369
	<30	0.454	0.369	0.407
	Avg.	0.428	0.429	0.425

Πίνακας 5.2. Απόδοση διαφορετικών αλγορίθμων στο σύνολο δεδομένων CSAE

Στο σύνολο δεδομένων CSAE (Πίνακας 5.2), ο Naïve Bayes παρουσίασε το υψηλότερο μέσο F-measure με τιμή 0,450. Παρουσίασε σχετικά καλό precision και recall για την κατηγορία ">30", γεγονός που υποδηλώνει την ικανότητά του να εντοπίζει ασθενείς με υψηλότερο κίνδυνο επανεισαγωγής μετά από 30 ημέρες, όπως και στο σύνολο δεδομένων GRAE. Το Decision Tree ακολουθεί ξανά την πορεία του Naïve Bayes, με μέσο μέτρο F-measure = 0.440. Ενώ οι τιμές των precision και recall του για την κλάση με label ">30" ήταν σχετικά χαμηλές, παρουσίασε ανταγωνιστικά αποτελέσματα για τις άλλες δύο κλάσεις. Ο Logistic Regression πέτυχε μέσο μέτρο F-measure=0,449. Παρουσίασε υψηλό precision και recall για την κλάση "OXI", αλλά οι επιδόσεις του για τις άλλες δύο κλάσεις ήταν μέτριες. Ο Random Forest πέτυχε μέσο F-measure=0,425. Όπως συνέβει και στο σύνολο δεδομένων CRAE, οι τιμές των recall και precision ήταν καλά ισοροπημένες, υποδεικνύοντας συνεπή απόδοση σε όλες τις κλάσεις, αλλά δεν ήταν οι βέλτιστες σε καμία από τις κλάσεις.

Αλγόριθμος	Κλάση	Precision	Recall	F-measure
Naïve Bayes	NO	0.475	0.596	0.529
	>30	0.427	0.376	0.400
	<30	0.479	0.415	0.445
	Avg.	0.460	0.462	0.458
Decision Tree	NO	0.467	0.583	0.519
	>30	0.412	0.363	0.386
	<30	0.481	0.420	0.448
	Avg.	0.453	0.455	0.451
Logistic Regression	NO	0.467	0.614	0.530
	>30	0.427	0.377	0.400
	<30	0.501	0.401	0.446
	Avg.	0.465	0.464	0.459
Random Forest	NO	0.461	0.480	0.470
	>30	0.381	0.378	0.380
	<30	0.431	0.415	0.423
	Avg.	0.424	0.425	0.424

Πίνακας 5.3. Απόδοση διαφορετικών αλγορίθμων στο σύνολο δεδομένων WSAE

Σχετικά με το σύνολο δεδομένων WSAE (Πίνακας 5.3), ο αλγόριθμος logistic regression επέδειξε το υψηλότερο μέσο F-measure με τιμή 0,459, υποδεικνύοντας τη συνολική αποτελεσματικότητά του στην ταξινόμηση των ασθενών σε διαφορετικές κατηγορίες επανεισαγωγής. Ξεπέρασε τους άλλους αλγορίθμους όσον αφορά το precision, το recall αλλά και το μέτρο F-measure για τις κατηγορίες "OXI" και ">30", καθιστώντας τον μια πολλά υποσχόμενη επιλογή για αυτές τις κατηγορίες. Ωστόσο, παρουσίασε ελαφρώς χαμηλότερες επιδόσεις στην κατηγορία "<30", με F-measure 0,446, γεγονός που υποδηλώνει ότι μπορεί να μην είναι τόσο αποτελεσματικός στον εντοπισμό ασθενών με επανεισαγωγές εντός 30 ημερών. Από την άλλη πλευρά, ο αλγόριθμος Decision Tree απέδωσε το υψηλότερο μέτρο F-measure=0,448, για την κατηγορία "<30", αναδεικνύοντας τα πλεονεκτήματά του στην πρόβλεψη των αποτελεσμάτων επανεισαγωγής εντός 30 ημερών. Επίσης, επέδειξε ανταγωνιστικές επιδόσεις για τις κατηγορίες "OXI" και ">30", με τιμές μέσου F-measure 0,519 και 0,386, αντίστοιχα. Ωστόσο, το συνολικό μέσο F-measure=0,451 υποδεικνύει ότι ενδέχεται να μην έχει τόσο ολοκληρωμένη απόδοση όσο ο logistic regression.

Ο αλγόριθμος Naïve Bayes παρουσίασε σταθερή απόδοση σε όλες τις κλάσεις, με μέσο F-measure=0,458. Παρουσίασε ισορροπημένες τιμές precision και recall για όλες τις κατηγορίες, γεγονός που το καθιστά αξιόπιστη επιλογή για την ταξινόμηση των ασθενών σε διαφορετικές κατηγορίες επανεισδοχής. Ο Random Forest, με μέσο F-measure=0.424, είχε ελαφρώς χαμηλότερες επιδόσεις από τους άλλους αλγορίθμους όσον αφορά τη συνολική απόδοση στην ταξινόμηση. Ενώ παρουσίασε ανταγωνιστικά αποτελέσματα για τις κατηγορίες "OXI" και "<30", η απόδοσή του για την κατηγορία ">30" ήταν σχετικά χαμηλή με τιμή F-measure=0,380.

Μεταξύ αυτών των 3 συνόλων δεδομένων, με κανένα από τα αυτά οι αλγόριθμοι δεν είχαν πραγματικά καλά αποτελέσματα και οι επιδόσεις τους είναι αρκετά παρόμοιες. Όμως παρατηρούμε ότι, όλοι οι αλγόριθμοι απέδωσαν καλύτερα για το σύνολο δεδομένων WSAE. Ο αλγόριθμος Logistic regression μεταξύ των 4 αλγορίθμων είχε την καλύτερη απόδοση με τιμή F-measure=0,459. Ο αλγόριθμος Naïve Bayes αποδίδει επίσης πολύ κοντά στον Logistic regression. Ο αλγόριθμος Decision Tree είχε καλή απόδοση ειδικά για την κατηγορία "<30". Έτσι, η λήψη μιας απόφασης από αυτό το σημείο είναι πραγματικά δύσκολη. Για αυτόν το λόγο, θα δοκιμάσουμε κάποιους αλγορίθμους bagging και boosting με Logistic regression και έναν voting ταξινομητή επίσης με Logistic regression, Decision tree και Naïve bayes.

Ο πίνακας 5.4 δείχνει τα αποτελέσματα των ταξινομητών bagging, boosting και stacking με τον αλγόριθμο Logistic regression. Φυσικά, καταλήγουμε μετά από τις παραπάνω αναλύσεις στην επιλογή και περαιτέρω χρήση του συνόλου δεδομένων WSAE, καθώς είχε καλή απόδοση σε σχέση με τις υπόλοιπες τεχνικές.

Αλγόριθμος	Κλάση	Precision	Recall	F-measure
Logistic Regression	NO	0.467	0.614	0.530
	>30	0.427	0.377	0.400
	<30	0.501	0.401	0.446
	Avg.	0.465	0.464	0.459
Bagging (Logistic Regression)	NO	0.469	0.617	0.533
	>30	0.431	0.365	0.396
	<30	0.497	0.415	0.453
	Avg.	0.466	0.466	0.460
Stacking (Logistic Regression)	NO	0.470	0.604	0.529
	>30	0.422	0.393	0.407
	<30	0.505	0.395	0.443
	Avg.	0.466	0.464	0.460
AdaBoostM1 (Logistic Regression)	NO	0.468	0.570	0.514
	>30	0.417	0.384	0.400
	<30	0.481	0.413	0.445
	Avg.	0.455	0.456	0.453

Πίνακας 5.4. Bagging, Stacking και Boosting (AdaboostM1) με Logistic regression

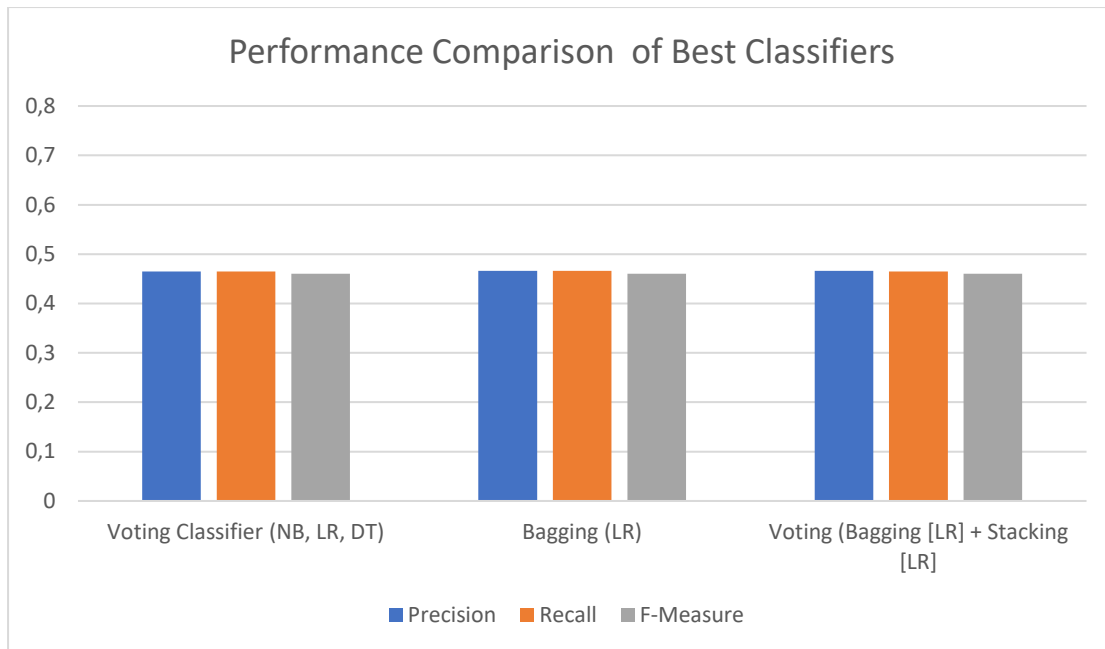
Το Bagging και το Stacking θα μπορούσαν να βελτιώσουν ελαφρώς την απόδοση του Logistic regression αλγορίθμου αλλά χωρίς κάποια θεαματικά αποτελέσματα. Από την άλλη πλευρά, το AdaboostM1 μειώνει την απόδοση. Εδώ, το Bagging βελτίωσε την απόδοση της κλάσης "<30", σημειώνοντας F-measure=0,453, της κλάσης "OXI" με F-measure =0,533, καθώς και τη βεβαρημένη μέση απόδοση με τιμή F-measure=0,460. Από την άλλη πλευρά, το Stacking βελτιώνει την απόδοση της κλάσης ">30" έχοντας τιμή F-measure=0,407. Ας προσπαθήσουμε τώρα να εφαρμόσουμε και τον ταξινομητή ψηφοφορίας (Voting) με τη χρήση των αλγορίθμων Logistic regression, Naïve bayes και Decision tree, εφαρμόζοντας επίσης τις τεχνικές των bagging και stacking που μόλις παρουσιάστηκε παραπάνω. Στον πίνακα 5.5 παρουσιάζονται τα αποτελέσματα του

ταξινομητή voting και του ταξινομητή bagging με την καλύτερη μέχρι στιγμής απόδοση για να γίνει σύγκριση:

Αλγόριθμος	Κλάση	Precision	Recall	F-measure
Voting Classifier (NB, LR, DT)	NO	0.470	0.612	0.532
	>30	0.430	0.364	0.395
	<30	0.494	0.420	0.454
	Avg.	0.465	0.465	0.460
Bagging (LR)	NO	0.469	0.617	0.533
	>30	0.431	0.365	0.396
	<30	0.497	0.415	0.453
	Avg.	0.466	0.466	0.460
Voting (Bagging [LR] + Stacking [LR])	NO	0.469	0.612	0.531
	>30	0.428	0.379	0.402
	<30	0.500	0.409	0.447
	Avg.	0.466	0.465	0.460

Πίνακας 5.5. Επίδοση ταξινομητή Voting και σύγκριση με τον ταξινομητή Bagging

Από τα παραπάνω αποτελέσματα φαίνεται ότι και οι 3 αλγόριθμοι είχαν περίπου την ίδια μέση απόδοση και ο καθένας από αυτούς είχε καλύτερη απόδοση για την πρόβλεψη μιας συγκεκριμένης κατηγορίας. Έτσι, εδώ υπάρχει ένας συμβιβασμός μεταξύ των αλγορίθμων και των κλάσεων, ο οποίος δείχνει ότι ο καλύτερος δυνατός αλγόριθμος μπορεί να είναι διαφορετικός για διαφορετικές περιπτώσεις χρήσης. Εδώ υπάρχει ένα ενδιαφέρον γεγονός, ο ταξινομητής Voting με τους Stacking και Bagging παρουσιάζει καλύτερη απόδοση για την κλάση ">30", αλλά παρουσίαζε καλύτερες τιμές μόνο με τον ταξινομητή Stacking. Το Σχήμα 5.1 δείχνει τη βεβαρημένη μέση απόδοση των καλύτερων μοντέλων:



Σχήμα 5.1. Σύγκριση επιδόσεων των καλύτερων ταξινομητών

Κατόπιν όλων αυτών των πειραμάτων, οι 3 καλύτεροι αλγόριθμοι που βρέθηκαν είναι οι Voting (NB, LR, DT), Bagging (LR) και Stacking (LR) όπου ο voting είχε καλή απόδοση για την κλάση "<30", ο bagging για την κλάση "No" και ο stacking αλγόριθμος για την κλάση ">30" και όλοι είχαν παρόμοια μέση απόδοση. Η μέθοδος επιλογής χαρακτηριστικών Wrapper subset attribute evaluation (WSAE) βρέθηκε να είναι η καλύτερη μέθοδος αξιολόγησης χαρακτηριστικών, σε σχέση με τις μεθόδους GRAE και CSAE. Κατά μέσο όρο, όλοι οι αλγόριθμοι σημείωσαν τιμή F-measure=0,460, το οποίο σίγουρα δεν είναι ένα καλό σκορ για χρήση στην παραγωγή και αποδεικνύει ότι τα δεδομένα έχουν πολύ λιγότερες τάσεις ή μοτίβα για την επίτευξη υψηλότερου σκορ. Παρόλο που σε διαφορετικά σύνολα δεδομένων, τα αποτελέσματα θα μπορούσαν να βελτιωθούν αν μπορούσαν να βρεθούν αρκετά μοτίβα. Άλλωστε, μπορούν να εφαρμοστούν προηγμένες τεχνικές βαθιάς μάθησης για τη βελτίωση των επιδόσεων.

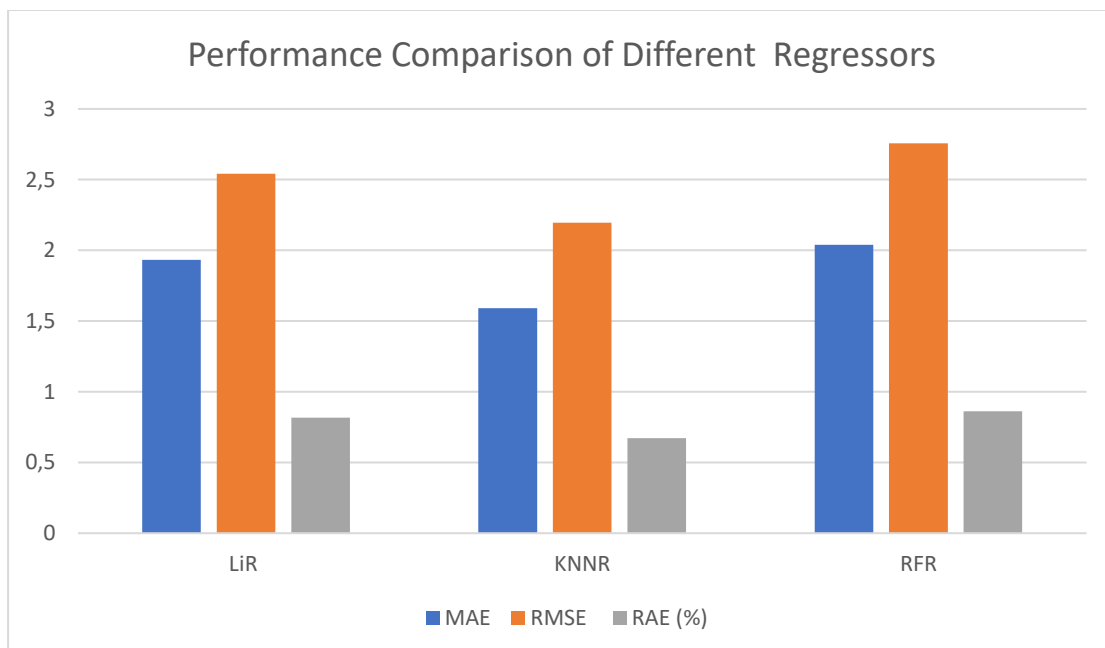
5.2. Πρόβλημα Παλινδρόμησης

Για την ανάλυση του προβλήματος της παλινδρόμησης εφαρμόστηκαν 3 διαφορετικοί αλγόριθμοι οι οποίοι είναι ο Linear Regression (LiR), ο Random Forest Regressor (RFR) και ο K-Nearest Neighbor Regressor (K-NNR). Στον ακόλουθο πίνακα 5.6 παρουσιάζονται τα αποτελέσματα των αλγορίθμων:

Αλγόριθμος	MAE	RMSE	RAE (%)
LiR	1.9336	2.5421	81.79%
KNNR	1.5903	2.1944	67.27%
RFR	2.0391	2.7576	86.26%

Πίνακας 5.6. Σύγκριση επιδόσεων του κάθε Regressor

Εξετάζοντας τα αποτελέσματα, ο K-NNR υπερτερεί τόσο του LiR αλγορίθμου όσο και του RFR όσον αφορά τις MAE και RMSE μετρικές, με χαμηλότερες τιμές που υποδηλώνουν καλύτερη ακρίβεια πρόβλεψης. Αυτό υποδηλώνει ότι ο KNNR παρέχει εγκυρότερες προβλέψεις και παρουσιάζει εξαιρετική απόδοση στην ελαχιστοποίηση των απόλυτων και τετραγωνικών διαφορών μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Επιπλέον, ο K-NNR παρουσιάζει επίσης σημαντικά χαμηλότερο ποσοστό RAE , υποδεικνύοντας μειωμένο σχετικό σφάλμα στις προβλέψεις σε σύγκριση με τους αλγορίθμους LiR και RFR. Το σχήμα 5.2 παρουσιάζει τη σύγκριση με γραφικό τρόπο:



Σχήμα 5.2. Σύγκριση επιδόσεων των Regressors

Παρατηρούμε πως, η απόδοση του LiR τοποθετείται μεταξύ των K-NNR και RFR. Ενώ παράγει ανταγωνιστικά αποτελέσματα όσον αφορά τις MAE και RMSE μετρικές σε σύγκριση με τον RFR, υστερεί σε σχέση με το K-NNR, υποδεικνύοντας ότι ο αλγόριθμος LiR ενδέχεται να μην καταγράφει τις πολύπλοκες συσχετίσεις τόσο αποτελεσματικά όσο ο K-NNR.

Από την άλλη πλευρά, ο αλγόριθμος RFR παρουσιάζει τα υψηλότερα ποσοστά στις MAE, RMSE και RAE μετρικές, γεγονός που υποδηλώνει ότι είναι ο λιγότερο αξιόπιστος από τους τρεις αλγορίθμους στην πρόβλεψη της μεταβλητής-στόχου. Είναι απαραίτητο να ερμηνεύσουμε αυτά τα ευρήματα με προσοχή, καθώς τα υψηλά ποσοστά ακρίβειας του RFR θα μπορούσαν να αποδοθούν σε υπερπροσαρμογή, ιδίως εάν του έχει επιτραπεί να αναπτύξει βαθιά και πολύπλοκα δέντρα.

Κατά την αξιολόγηση των αλγορίθμων, είναι σημαντικό να λαμβάνονται υπόψη τα χαρακτηριστικά του συνόλου δεδομένων και η ερμηνευσιμότητα των μοντέλων. Ενώ ο K-NNR παρουσιάζει ανώτερη ακρίβεια πρόβλεψης, ενδέχεται να μην είναι τόσο ερμηνεύσιμος όσο ο LiR, ο οποίος προσφέρει σαφή κατανόηση των σχέσεων μεταξύ των μεταβλητών μέσω των συντελεστών. Ο RFR, παρά τη χαμηλότερη ακρίβεια πρόβλεψης, μπορεί να εξακολουθεί να είναι πολύτιμος για τη καταγραφή πολύπλοκων αλληλεπιδράσεων και μη γραμμικών μοτίβων στα δεδομένα.

ΚΕΦΑΛΑΙΟ 6^ο – Υλοποίηση Διαδικτυακής Εφαρμογής

6.1. Τεχνολογική Προσέγγιση

Η εφαρμογή επικεντρώνεται στην αξιοποίηση τεχνικών επιβλεπόμενης μάθησης (supervised learning) για την εκπαίδευση προγνωστικών μοντέλων στο επιλεγμένο σύνολο δεδομένων ασθενών με διαβήτη. Η προεπεξεργασία των δεδομένων, η ανάλυση των χαρακτηριστικών, η δημιουργία μοντέλων, η αξιολόγηση και η ανάπτυξη εκτελέστηκαν με τη χρήση βιβλιοθηκών της προγραμματιστικής γλώσσας Python, συμπεριλαμβανομένων των Pandas, Scikit-learn, Matplotlib και Django. Επίσης, χρησιμοποιήθηκε η περιγραφική γλώσσα HTML (Hyper Text Markup Language) για την κατασκευή της εφαρμογής. Είναι μια υλοποίηση end-to-end, δηλαδή η δημιουργία ενός πλήρους λειτουργικού συστήματος από την αρχή μέχρι το τέλος χωρίς αυτό να χρειάζεται να προμηθευτεί τίποτα από κάποιον τρίτο. Ακόμη, η διαδικτυακή αυτή εφαρμογή, επιτρέπει τη δημιουργία προβλέψεων σε πραγματικό χρόνο με τη χρήση των αναπτυγμένων μοντέλων μέσω μιας διεπαφής χρήστη.

Το Django είναι ένα δημοφιλές Python web framework που χρησιμοποιείται για την ανάπτυξη ιστοσελίδων και εφαρμογών. Παρέχει ένα σύνολο εργαλείων και βιβλιοθηκών για τη διευκόλυνση της δημιουργίας ασφαλών, επεκτάσιμων και γρήγορων ιστοσελίδων με λειτουργική διαχείριση της βάσης δεδομένων και υποστήριξη για αλληλεπίδραση με τον χρήστη. Συγκεκριμένα, το Django παρέχει μια διεπαφή όπου οι χρήστες μπορούν να κάνουν εισαγωγή δεδομένων και να δουν τις προβλέψεις των μοντέλων. Τα μοντέλα φορτώθηκαν στο Django χρησιμοποιώντας το Pickle module.

6.2. Παρουσίαση εφαρμογής

Η εφαρμογή έχει δύο κύριες οθόνες, την αρχική σελίδα για την παλινδρόμηση και τη αρχική σελίδα για την ταξινόμηση (Εικόνα). Στο πάνω μέρος της σελίδας εμφανίζονται το όνομα της εφαρμογής (Readmission and LoS for Diabetics) και οι επιλογές του χρήστη που είναι οι «Readmission Prediction» και «Length of Stay Prediction». Πατώντας επάνω στο κουμπί «Readmission Prediction», εμφανίζονται στο χρήστη κάποια πεδία τα οποία οποία και θα πρέπει να συμπληρώσει για να γίνει η πρόβλεψη. Τα περισσότερα υπάρχουν ήδη σαν επιλογές που παρουσιάζονται στον χρήστη εκτός

από τα χαρακτηριστικά Diagnosis 1, 2 και 3. Παρακάτω, παρουσιάζονται με τη χρήση στιγμιοτύπων οι λειτουργίες της εφαρμογής.

Readmission and LoS for Diabetics

Length of Stay Prediction

Readmission Prediction

Predict Readmission to Hospital

Enter user data to predict the admission of the patient to the hospital.

Age:

[0-10] ▾

Discharge Disposition ID:

1 ▾

Number of Outpatient Visits:

1 ▾

Number of Emergency Visits:

1 ▾

Number of Inpatient Visits:

1 ▾

Diagnosis 1:

Number of Diagnoses:

1 ▾

Nateglinide:

No ▾

Acarbose:

No ▾

Tolazamide:

No ▾

Glipizide-Metformin:

No ▾

Change:

No ▾

Predict

Readmission and LoS for Diabetics
Length of Stay Prediction
Readmission Prediction

Predict Time in Hospital

Enter user data to predict the time in the hospital.

Gender:

Female

Discharge Disposition ID:

1

Number of Lab Procedures:

1

Number of Medications:

1

Diagnosis 1:

Diagnosis 2:

Diagnosis 3:

Number of Diagnoses:

1

A1C Result:

None

Repaglinide:

No

Readmitted:

NO

Predict

Εικόνα 6.1 και 6.2. Αρχική οθόνη πρόβλεψης επανεισαγωγής (αριστερά) και αρχική οθόνη πρόβλεψης της διάρκειας παραμονής (δεξιά).

Στην οθόνη για την πρόβλεψη της επανεισαγωγής σε νοσοκομείο, εισάγονται τα παρακάτω χαρακτηριστικά:

1) Αναλυτικά, στο πεδίο «**Age**» ο χρήστης επιλέγει το όριο που ανήκει ηλικιακά ο διαβητικός ασθενής.

Age:

[0-10]

[0-10]

[10-20]

[20-30]

[30-40]

[40-50]

[50-60]

[60-70]

[70-80]

[80-90]

[90-100]

Εικόνα 6.3. Χαρακτηριστικό «Age»

2) Στο επόμενο πεδίο «**Discharge Disposition ID**» εισάγεται η κατάσταση ή ο προορισμός του ασθενούς μετά την έξοδο του από το νοσοκομείο (1-Πήγε σπίτι, 2-

Μεταφορά σε άλλο νοσοκομείο, 3- Πήγε σε νοσοκομείο βραχυπρόθεσμης διάρκειας, κ.λ.π). Οι τιμές κυμαίνονται από 1 έως 28.

Discharge Disposition ID:

1
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
27
28

Εικόνα 6.4. Χαρακτηριστικό «Discharge Disposition ID»

3) Κατόπιν, στο πεδίο «**Number of Outpatient Visits**» εισάγεται ο αριθμός επισκέψεων του ασθενούς στα εξωτερικά ιατρεία τον προηγούμενο χρόνο και κυμαίνεται από 0 έως 42.

Number of Outpatient Visits:

1
22
23
24
25
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

Εικόνα 6.5. Χαρακτηριστικό «Number of Outpatient Visits»

4) Στο πεδίο «**Number of Emergency Visits**» εισάγεται ο ακέραιος αριθμός επείγουσών επισκέψεων του ασθενούς τον προηγούμενο χρόνο και κυμαίνεται από 0 έως 76.

Number of Emergency Visits:

1
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76

Εικόνα 6.6. Χαρακτηριστικό «Number of Emergency Visits»

5) Μετά, στο πεδίο «**Number of Inpatient Visits**» εισάγεται ο ακέραιος αριθμός ενδονοσοκομειακών επισκέψεων του ασθενούς κατά το προηγούμενο έτος και κυμαίνεται από 0 έως 21.

Number of Inpatient Visits:

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

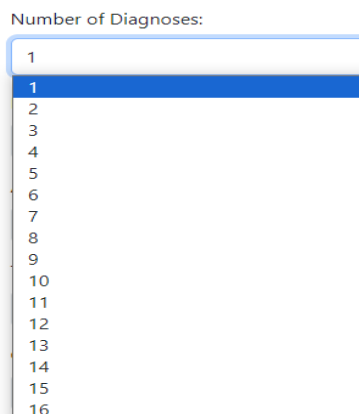
Εικόνα 6.7. Χαρακτηριστικό «Number of Inpatient Visits»

6) Στο πεδίο «**Diagnosis 1**» εισάγεται η πρωταρχική διάγνωση, η οποία είναι κωδικοποιημένη ως τα τρία πρώτα ψηφία του ICD9, με 848 διαφορετικές τιμές. Καθώς οι τιμές είναι πάρα πολλές, το πεδίο είναι κενό και ο χρήστης συμπληρώνει τον κατάλληλο κωδικό.

Diagnosis 1:

Εικόνα 6.8. Χαρακτηριστικό «Diagnosis 1»

7) Ύστερα, στο πεδίο «**Number of Diagnoses**» εισάγεται ο ακέραιος αριθμός διαγνώσεων που καταχωρήθηκαν στο σύστημα και κυμαίνεται από 0 έως 16.

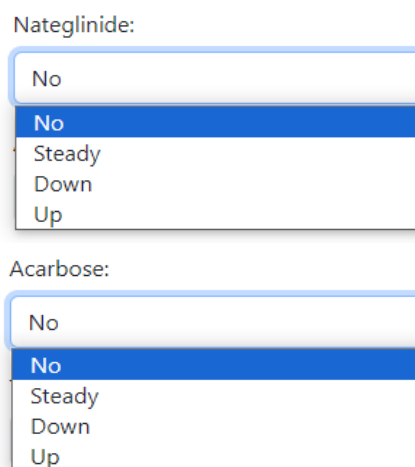


Number of Diagnoses:

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16

Εικόνα 6.9. Χαρακτηριστικό «Number of Diagnoses»

8) Στα πεδία «**Nateglinide**» και «**Acarbose**» ο χρήστης βρίσκεται ανάμεσα σε 4 επιλογές. Οι τιμές είναι: "up" εάν η δοσολογία αυξήθηκε κατά τη διάρκεια της συνάντησης, "down" εάν η δοσολογία μειώθηκε, "steady" εάν η δοσολογία δεν άλλαξε και "no" εάν το φάρμακο δεν συνταγογραφήθηκε.



Nateglinide:

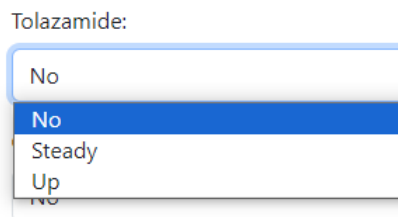
- No
- Steady
- Down
- Up

Acarbose:

- No
- Steady
- Down
- Up

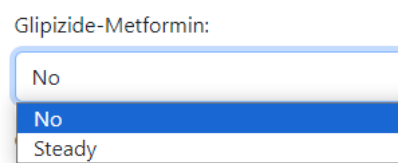
Εικόνα 6.10. Χαρακτηριστικά «Nateglinide» και «Acarbose»

9) Στο πεδίο «**Tolazamide**» γίνεται ακριβώς ότι και στο πεδία «Nateglinide» και «Acarbose» με τη διαφορά ότι ο χρήστης βρίσκεται ανάμεσα σε μόνο 3 επιλογές που είναι οι «up», «steady» και «no».



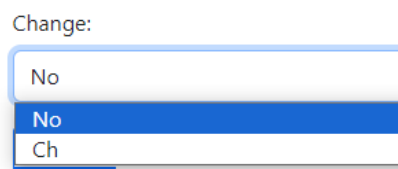
Εικόνα 6.12. Χαρακτηριστικό «Tolazamide»

10) Στο πεδίο «**Glipizide-Metformin**» γίνεται ότι και στα παραπάνω 2 πεδία, όμως ο χρήστης βρίσκεται ανάμεσα σε 2 επιλογες, την «steady» και την «no».



Εικόνα 6.13. Χαρακτηριστικό «Glipizide-Metformin»

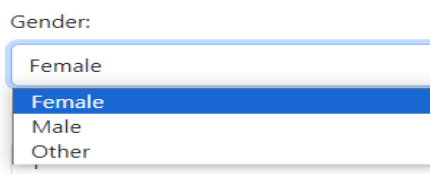
11) Και τέλος, στο πεδίο «**Change**» το οποίο δείχνει εάν υπήρξε αλλαγή στα διαβητικά φάρμακα, ο χρήστης βρίσκεται ανάμεσα σε 2 επιλογες, την «no» και την «Ch», με την πρώτη να δηλώνει ότι δεν υπήρξε αλλαγή και την δεύτερη να υπήρξε αλλαγή.



Εικόνα 6.14. Χαρακτηριστικό «Change»

Στην οθόνη για την πρόβλεψη της διάρκειας παραμονής (LoS), εισάγονται τα παρακάτω χαρακτηριστικά:

1) Αναλυτικά, στο πεδίο «**Gender**», ο χρήστης εισάγει το φύλο του διαβητικού ασθενή και επιλέγει ανάμεσα σε «Female», «Male» και «Other».



Εικόνα 6.15. Χαρακτηριστικό «Gender»

- 2) Το πεδίο «**Discharge Disposition ID**» αναλύθηκε πιο πάνω.
- 3) Στο πεδίο «**Number of Lab Procedures**», εισάγεται ο ακέραιος αριθμός εργαστηριακών εξετάσεων που έγιναν στο νοσοκομείο και κυμαίνεται από 1 έως 132.

Number of Lab Procedures:

1
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132

Εικόνα 6.16. Χαρακτηριστικό «Number of Lab Procedures»

- 4) Στο πεδίο «**Number of Medications**» εισάγεται ο ακέραιος αριθμός διαφορετικών φαρμάκων που χορηγήθηκαν κατά τη διάρκεια της συνάντησης και κυμαίνεται από 1 έως 81.

Number of Medications:

1
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81

Εικόνα 6.17. Χαρακτηριστικό «Number of Medications»

- 5) Μετά, στα πεδία «**Diagnosis 1**», «**Diagnosis 2**» και «**Diagnosis 3**» εισάγεται η πρωταρχική, η δευτερεύουσα και η πρόσθετη δευτερεύουσα διάγνωση, οι οποίες είναι κωδικοποιημένες ως τα τρία πρώτα ψηφία του ICD9 με τιμές 848, 923 και 954 αντίστοιχα. Καθώς οι τιμές είναι πάρα πολλές, τα πεδία είναι κενά και ο χρήστης συμπληρώνει τον κατάλληλο κωδικό.

Diagnosis 1:

Diagnosis 2:

Diagnosis 3:

Εικόνα 6.18. Χαρακτηριστικά «Diagnosis 1», «Diagnosis 2» και «Diagnosis 3»

6) Το πεδίο «**Number of Diagnosis**» αναλύθηκε πιο πάνω .

7) Στο πεδίο «**A1C Result**» εισάγεται το εύρος του αποτελέσματος ή αν δεν έγινε η εξέταση. Ο χρήστης μπορεί να επιλέξει ανάμεσα στις τιμές: ">8" εάν το αποτέλεσμα ήταν μεγαλύτερο από 8%, ">7" εάν το αποτέλεσμα ήταν μεγαλύτερο από 7% αλλά μικρότερο από 8%, "Norm" εάν το αποτέλεσμα ήταν μικρότερο από 7% και "None" εάν δεν μετρήθηκε.

A1C Result:

None

None

>7

>8

Norm

Εικόνα 6.19. Χαρακτηριστικό «A1C Result»

8) Στο πεδίο «**Repaglinide**» ο χρήστης βρίσκεται ανάμεσα σε 4 επιλογές. Οι τιμές είναι: "up" εάν η δοσολογία αυξήθηκε κατά τη διάρκεια της συνάντησης, "down" εάν η δοσολογία μειώθηκε, "steady" εάν η δοσολογία δεν άλλαξε και "no" εάν το φάρμακο δεν συνταγογραφήθηκε.

Repaglinide:

No

No

Up

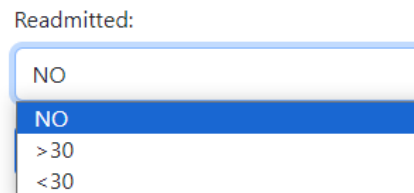
Steady

Down

Εικόνα 6.20. Χαρακτηριστικό «Repaglinide»

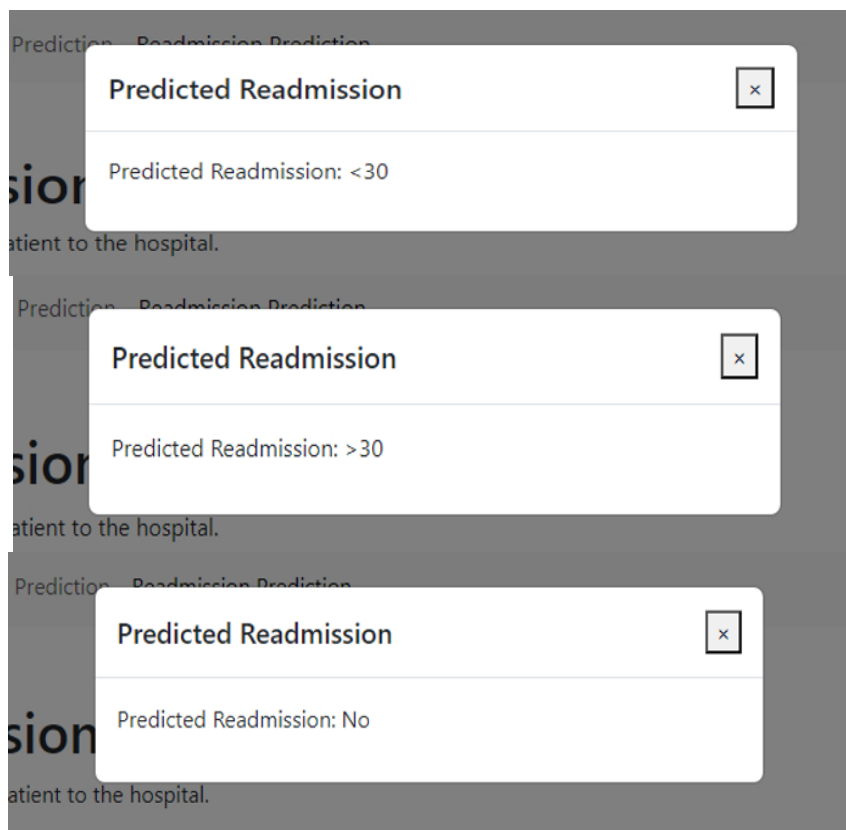
9) Και τέλος στο πεδίο «**Readmitted**», εισάγονται οι ημέρες έως την επανεισαγωγή του διαβητικού ασθενή σε νοσοκομείο. Ο χρήστης μπορεί να επιλέξει ανάμεσα στις

τιμές: "<30" εάν ο ασθενής επανεισήχθη σε λιγότερο από 30 ημέρες, ">30" εάν ο ασθενής επανεισήχθη σε περισσότερες από 30 ημέρες, και "No" εάν δεν υπάρχει καταγραφή επανεισδοχής.

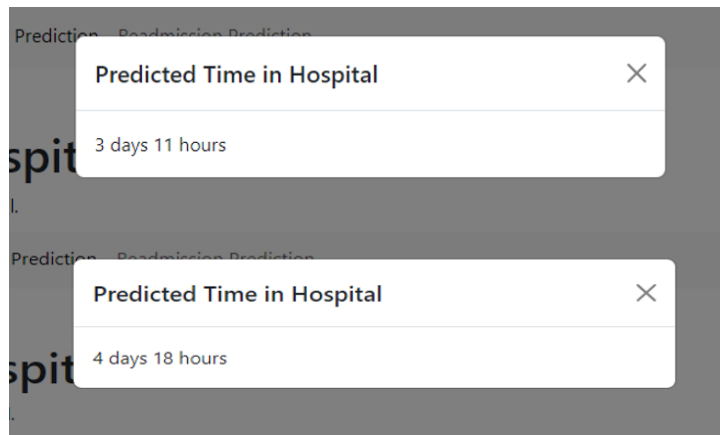


Εικόνα 6.21. Χαρακτηριστικό «Readmitted»

Αφότου συμπληρωθούν όλα τα πεδία, ο χρήστης πατάει το κουμπί «Predict» το οποίο βρίσκεται κάτω αριστερά στην οθόνη και εμφανίζεται στο πάνω μέρος της οθόνης ένα μικρό παράθυρο με την πρόβλεψη που έχει γίνει από το σύστημα. Στην Εικόνα 1 παρουσιάζονται τα 3 αποτελέσματα για την πρόβλεψη επανεισαγωγής στο νοσοκομείο και στην Εικόνα 2, το αποτέλεσμα της πρόβλεψης της διάρκειας παραμονής.



Εικόνα 6.22. Μήνυμα αποτελέσματος για την πρόβλεψη της επανεισαγωγής (<30, >30, No)



Εικόνα 6.23. Μήνυμα αποτελέσματος για την πρόβλεψη της διάρκειας νοσηλείας (σε μέρες και ώρες)

ΚΕΦΑΛΑΙΟ 7^ο - Συμπεράσματα

Σε αυτή τη μελέτη, διάφορες προσεγγίσεις μηχανικής μάθησης χρησιμοποιήθηκαν για την διαχείριση ασθενών με διαβήτη. Πιο συγκεκριμένα, έγινε πρόβλεψη της επανεισαγωγής διαβητικών ασθενών στο νοσοκομείο για 3 χρονικά πλαίσια (επανεισαγωγή σε περισσότερο από 30 μέρες, σε λιγότερο από 30 μέρες ή καθόλου) αλλά και πρόβλεψη της διάρκειας παραμονής τους. Για το πρόβλημα της ταξινόμησης του διαστήματος επανεισδοχής, χρησιμοποιήθηκαν 8 διαφορετικοί ταξινομητές ώστε να συγκριθούν και να βρεθεί ο καλύτερος. Τελικά, επιλέχθηκαν οι 3 καλύτεροι ταξινομητές οποίοι εμφάνιζαν τα ίδια αποτελέσματα για τη σταθμισμένη μέση βαθμολογία με τιμή F-measure = 0,460, αλλά διαφορετικά για την καλύτερη πρόβλεψη μεταξύ των διαφόρων κατηγοριών. Γενικότερα, η αξιοποίηση της δύναμης των τεχνικών συνόλου (ensemble) σε αυτή την περίπτωση θα βοηθούσε πάρα πολύ για την απόκτηση ενός γενικευμένου μοντέλου.

Στο κομμάτι του προβλήματος παλινδρόμησης, χρησιμοποιήθηκαν 3 διαφορετικοί αλγόριθμοι, όπου ο αλγόριθμος K-Nearest Neighbors Regressor (K-NNR) παρουσίασε την καλύτερη απόδοση με τις χαμηλότερες τιμές μέσου απόλυτου σφάλματος (MAE) και μέσου τετραγωνικού σφάλματος (RMSE). Η απόδοση του K-NNR στην πρόβλεψη του αριθμού ημερών παραμονής ενός διαβητικού ασθενούς στο νοσοκομείο, παρέχει πολύτιμες πληροφορίες για τους επαγγελματίες του τομέα της υγειονομικής περίθαλψης (ιατροί, νοσηλευτές κ.λ.π), επιτρέποντάς τους να προγραμματίσουν την κατανομή των πόρων και να βελτιστοποιούν την παροχή φροντίδας στους διαβητικούς ασθενείς που συντρέχουν μεγαλύτερο κίνδυνο επιπλοκών.

Επιπλέον, για να μπορέσουν οι ιατροί αλλά και οι διαβητικοί ασθενείς να έχουν πρόσβαση στις παραπάνω προβλέψεις καθώς ο τελικός στόχος της εργασίας είναι να γίνει χρήση των προβλεπτικών μοντέλων από τους επαγγελματίες υγείας και τους ασθενείς, αναπτύχθηκε μια εφαρμογή. Συγκεκριμένα, με την εισαγωγή των χαρακτηριστικών που έχουν επιλεγεί ως βέλτιστα από τις τεχνικές WSAE για το πρόβλημα ταξινόμησης και CSAE για το πρόβλημα παλινδρόμησης, ο ασθενής λαμβάνει σαν μήνυμα εάν θα γίνει επανεισαγωγή του στο νοσοκομείο. Αν το αποτέλεσμα είναι θετικό, τότε εισάγει επιπλέον χαρακτηριστικά και εμφανίζεται στην οθόνη ποιο θα είναι το διάστημα παραμονής του σε μέρες και ώρες. Η εφαρμογή αυτή θα μπορούσε να είναι ένα σημαντικό εργαλείο στην υποστήριξη λήψης αποφάσεων για

τη διαχείριση των διαβητικών ασθενών, παρέχοντας πολλαπλά οφέλη τόσο για τους ίδιους τους ασθενείς, τους επαγγελματίες υγείας αλλά και τους πόρους των νοσοκομείων.

Συμπερασματικά, η εφαρμογή της μηχανικής μάθησης στη διαχείριση του σακχαρώδη διαβήτη προσφέρει αρκετά πλεονεκτήματα. Με την πρόβλεψη των επιστροφών διαβητικών ασθενών στο νοσοκομείο και της διάρκειας παραμονής τους, όχι μόνο οι ιατροί είναι σε θέση να λάβουν προληπτικά μέτρα για τη φροντίδα αυτών με προσωποποιημένη και άμεση παρακολούθηση μειώνοντας τον αριθμό των επανεισαγωγών και των θανάτων αλλά και το νοσοκομείο εξοικονομεί οικονομικούς πόρους για την ομαλότερη διαχείριση της φροντίδας των διαβητικών ασθενών. Ακόμη, ο ίδιος ο ασθενής έχει τη δυνατότητα να είναι σε ετοιμότητα αλλά και να «προετοιμαστεί» ψυχολογικά για μια επανεισαγωγή του στο νοσοκομείο, ρυθμίζοντας τους οικονομικούς πόρους του.

ΑΝΑΦΟΡΕΣ

- [1] Χαρατσή - Γιωτάκη, Ε., Γιωτάκη, Ζ., & Καρανάτσης, Ν. (2023). *Σακχαρώδης Διαβήτης* [Μονογραφία]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. <https://dx.doi.org/10.57713/kallipos-136>
- [2] Augstein, P., Heinke, P., Vogt, L., Kohnert, K. D., & Salzsieder, E. (2022). Patient-Tailored Decision Support System Improves Short- and Long-Term Glycemic Control in Type 2 Diabetes. *Journal of diabetes science and technology*, 16(5), 1159–1166. <https://doi.org/10.1177/19322968211008871>
- [3] Gupta, S. C., & Goel, N. (2023). Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques. *Procedia Computer Science*, 218, 1257–1269. <https://doi.org/10.1016/j.procs.2023.01.104>
- [4] Spat, S., Donsa, K., Beck, P., Höll, B., Mader, J. K., Schaupp, L., Augustin, T., Chiarugi, F., Lichtenegger, K. M., Plank, J., & Pieber, T. R. (2017). A Mobile Computerized Decision Support System to Prevent Hypoglycemia in Hospitalized Patients With Type 2 Diabetes Mellitus. *Journal of diabetes science and technology*, 11(1), 20–28. <https://doi.org/10.1177/1932296816676501>
- [5] Kriventsov, S., Lindsey, A., & Hayeri, A. (2020). The Diabits App for Smartphone-Assisted Predictive Monitoring of Glycemia in Patients with Diabetes: Retrospective Observational Study. *JMIR diabetes*, 5(3), e18660. <https://doi.org/10.2196/18660>
- [6] Robbins, T. D., Lim Choi Keung, S. N., Sankar, S., Randeva, H., & Arvanitis, T. N. (2019). Risk factors for readmission of inpatients with diabetes: A systematic review. *Journal of Diabetes and Its Complications*, 33(5), 398–405. <https://doi.org/10.1016/j.jdiacomp.2019.01.004>
- [7] L. Alturki, K. Aloraini, A. Aldughayshim and S. Albahli, "Predictors of Readmissions and Length of Stay for Diabetes Related Patients," *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, Abu Dhabi, United Arab Emirates, 2019, pp. 1-8, doi: 10.1109/AICCSA47632.2019.9035280.
- [8] Cheng, Y. L., Wu, Y. R., Lin, K. D., Lin, C. R., & Lin, I. M. (2023). Using Machine Learning for the Risk Factors Classification of Glycemic Control in Type 2 Diabetes

- Mellitus. *Healthcare (Basel, Switzerland)*, 11(8), 1141. <https://doi.org/10.3390/healthcare11081141>
- [9] Deberneh, H. M., & Kim, I. (2021). Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *International Journal of Environmental Research and Public Health*, 18(6), 3317. <https://doi.org/10.3390/ijerph18063317>
- [10] Damian M (2017). Predicting Diabetic Readmission Rates: Moving Beyond Hba1c. *Current Trends in Biomedical Engineering & Biosciences*; 7(3):555707. 007 DOI: [10.19080/CTBEB.2017.07.555715](https://doi.org/10.19080/CTBEB.2017.07.555715).
- [11] S. El-Sappagh, J. M. Alonso, F. Ali, A. Ali, J. -H. Jang and K. -S. Kwak. (2018). "An Ontology-Based Interpretable Fuzzy Decision Support System for Diabetes Diagnosis," in *IEEE Access*, vol. 6, pp. 37371-37394, DOI: [10.1109/ACCESS.2018.2852004](https://doi.org/10.1109/ACCESS.2018.2852004).
- [12] Barsasella, D., Bah, K., Mishra, P., Uddin, M., Dhar, E., Suryani, D. L., Setiadi, D., Masturoh, I., Sugiarti, I., Jonnagaddala, J., & Syed-Abdul, S. (2022). A Machine Learning Model to Predict Length of Stay and Mortality among Diabetes and Hypertension Inpatients. *Medicina*, 58(11), 1568. <https://doi.org/10.3390/medicina58111568>
- [13] Mandel, S. R., Langan, S., Mathioudakis, N. N., Sidhaye, A. R., Bashura, H., Bie, J. Y., Mackay, P., Tucker, C., Demidowich, A. P., Simonds, W. F., Jha, S., Ebinuwa, I., Kantsiper, M., Howell, E. E., Wachter, P., Golden, S. H., & Zilbermint, M. (2019). Retrospective study of inpatient diabetes management service, length of stay and 30-day readmission rate of patients with diabetes at a community hospital. *Journal of community hospital internal medicine perspectives*, 9(2), 64–73. <https://doi.org/10.1080/20009666.2019.1593782>
- [14] Jaotombo, F., Pauly, V., Fond, G., Orleans, V., Auquier, P., Ghattas, B., & Boyer, L. (2022). Machine-learning prediction for hospital length of stay using a French medico-administrative database. *Journal of market access & health policy*, 11(1), 2149318. <https://doi.org/10.1080/20016689.2022.2149318>

- [15] Karegowda, Asha & Manjunath, A. & M.A, Jayaram. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, Volume 2, No. 2, pp. 271-277, <https://csjournals.com/IJITKM/PDF%203-1/19.pdf>
- [16] Hall A. Mark (1999). *Correlation-based Feature Selection for Machine Learning*, University of Waikato, Hamilton. <https://www.cs.waikato.ac.nz/~mhall/thesis.pdf>
- [17] Ron Kohavi, George H. John. (1997). Wrappers for feature subset selection, *Artificial Intelligence*, Volume 97, Issues 1–2, Pages 273-324, ISSN 0004-3702, [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X).
- [18] John, G.H., & Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo*, Pages 338-345, <https://doi.org/10.48550/arXiv.1302.4964>
- [19] Ross Quinlan (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA
- [20] Sumner, M., Frank, E., & Hall, M.A. (2005). Speeding Up Logistic Model Tree Induction. *European Conference on Principles of Data Mining and Knowledge Discovery*. DOI:[10.1007/11564126_72](https://doi.org/10.1007/11564126_72)
- [21] Leo Breiman (2001). Random Forests. *Machine Learning*. 45(1), 5-32, <https://doi.org/10.1023/A:1010933404324>
- [22] Leo Breiman (1996). Bagging predictors. *Machine Learning*. 24(2), 123-140, <https://doi.org/10.1007/BF00058655>
- [23] D. H. Wolpert, “Stacked Generalization,” *Neural Networks*, Vol. 5, No. 2, 1992, pp. 241-259. [http://dx.doi.org/10.1016/S0893-6080\(05\)80023-1](http://dx.doi.org/10.1016/S0893-6080(05)80023-1)
- [24] Yoav Freund, Robert E. Schapire (1996). Experiments with a new boosting algorithm. In: *Thirteenth International Conference on Machine Learning*, San Francisco, 148-156, <https://api.semanticscholar.org/CorpusID:1836349>
- [25] Aha, W. & Kibler, Dennis & Albert, Marc. (1991). Instance-Based Learning Algorithms. *Machine Learning*. 6. 37-66, DOI:[10.1023/A:1022689900470](https://doi.org/10.1023/A:1022689900470).

