Εθνικό Μετσόβιο Πολυτεχνείο

Ανάλυση Βιοδεδομένων
Εξαμηνιαία Εργασία: Ανάλυση Παραγόντων Νόσησης από Διαβήτη Τύπου ΙΙ

Βλασόπουλος Μιχάλης: 03400204

Ιωάννα Μάλλη: 03400224

Αλέξης Μηλιώνης : 03400226

Αντώνης Προμπονάς : 03400232

ΔΠΜΣ
Επιστήμη Δεδομένων και Μηχανική Μάθηση

1 Abstract

Στην παρούσα εργασία ερευνήθηκαν οι παράγοντες που οδηγούν στην εμφάνιση διαβήτη τύπου 2, με ιδιαίτερη έμφαση σε εκείνους που σχετίζονται με τις επιλογές του τρόπου ζωής. Για τον σκοπό αυτό, αξιοποιήθηκαν σύνολα δεδομένων από τηλεφωνικές έρευνες, τα οποία συνενώθηκαν για τη δημιουργία νέου dataset. Πραγματοποιήθηκε διερευνητική ανάλυση των δεδομένων για την εξαγωγή χρήσιμων μοτίβων. Στη συνέχεια, δημιουργήθηκαν μοντέλα Random Forest για την ταξινόμηση των πιθανών ασθενών σε κατηγορίες Υγιών και Διαβητικών, με έμφαση στην επίτευξη υψηλού recall καθώς η ακριβής αναγνώριση των διαβητικών ασθενών είναι κρίσιμη στον ιατρικό τομέα. Για την αντιμετώπιση του προβλήματος της ανισορροπίας των κλάσεων, εξετάστηκαν μέθοδοι data oversampling και data undersampling.

2 Εισαγωγή

Ο διαβήτης είναι μια χρόνια πάθηση που εμφανίζεται όταν το σώμα δεν μπορεί να παράγει αρκετή ινσουλίνη ή δεν μπορεί να χρησιμοποιήσει αποτελεσματικά την ινσουλίνη που παράγει. Αποτελεί ένα διαδεδομένο πρόβλημα υγείας παγκοσμίως που μπορεί να οδηγήσει σε σοβαρές επιπλοκές εάν δεν ελεγχθεί σωστά, όπως καρδιακές παθήσεις, νεφρική ανεπάρκεια, νευροπάθεια, προβλήματα όρασης και άλλες σοβαρές καταστάσεις υγείας. Η διαχείριση του διαβήτη περιλαμβάνει τακτική παρακολούθηση των επιπέδων σακχάρου στο αίμα, υγιεινή διατροφή, άσκηση, και φαρμακευτική αγωγή εάν είναι απαραίτητο. Ο διαβήτης έχει δυο διαφορετικές μορφές εμφάνισης. Ο διαβήτης τύπου 1 είναι μια αυτοάνοση πάθηση όπου το ανοσοποιητικό σύστημα επιτίθεται και καταστρέφει τα β-κύτταρα του παγκρέατος που παράγουν ινσουλίνη. Άτομα με διαβήτη τύπου 1 χρειάζονται καθημερινή χορήγηση ινσουλίνης για να επιβιώσουν. Είναι πιο κοινός σε παιδιά και νεαρούς ενήλικες, αλλά μπορεί να εμφανιστεί σε οποιαδήποτε ηλικία.

Ο διαβήτης τύπου 2, αποτελεί την πιο συνηθισμένη μορφή διαβήτη και όπως εμφανίζεται όταν το σώμα δεν μπορεί να χρησιμοποιήσει την ινσουλίνη αποτελεσματικά (αντίσταση στην ινσουλίνη) ή όταν το πάγκρεας δεν παράγει αρκετή ινσουλίνη. Συχνά συνδέεται με τον τρόπο ζωής και μπορεί να προληφθεί ή να καθυστερήσει με υγιεινή διατροφή, τακτική άσκηση και διατήρηση υγιούς σωματικού βάρους.

Επιπρόσθετα , εμφανίζεται και η κατηγορία του προ-διαβήτη. Οι προδιαβητικού εμφανίζουν υψηλά επίπεδα σακχάρου, ψηλότερα από το φυσιολογικό αλλα όχι αρκετά υψηλά ώστε να θεωρούνται διαβητικοί. Οι κύριοι παράγοντες κινδύνου για έναν προδιαβητικό είναι η παχυσαρκία,ο ανενεργός τρόπος ζωής σε συνδυασμό με κακές διατροφικές συνήθειες κάθως και το οικογενειακό ιστορικό, σε περίπτωση που υπάρχουν μέλη της οικογένειας που έχουν διαβήτη τύπου 2.

Στην παρούσα εργασία, λοιπόν ερευνάται ο διαβήτης και επιχειρείται η κατανόηση των παραγόντων που τον προκαλούν. Συγκεκριμένα, μέσα από ένα dataset μεγάλου πλήθους δεδομένων προερχόμενων από τηλεφωνική έρευνα η οποία πραγματοποιήθηκε στις ΗΠΑ το 2014, 2015 και 2022, αντλούνται πληροφορίες που συνδέεουν τον διαβήτη με καθημερινές συνήθειες όπως το κάπνισμα και η διατροφη, με την ψυχική υγεία και την οικονομική κατάσταση του ατόμου. Αποτελείται από 253 χιλιάδες δείγματα, περιλαμβάνει 30 χαρακτηριστικά που σχετίζονται με την υγεία και τον τρόπο ζωής και κατηγοριοποιεί τους συμμετέχοντες σε 3 κατηγορίες: Διαβητικούς, Προδιαβητικούς και Υγιείς.

Βασικός μας σκοπός είναι αρχικά, η καλύτερη κατανόηση της νόσου και ιδιαίτερα η αποσαφήνιση των επικίνδυνων παράγοντων του lifestyle που προκαλούν την εμφάνιση και την εξέλιξη της. Με αυτόν τον τρόπο, καθίσταται δυνατή η πρόληψη από τη δημιουργία προδιάθεσης ή και η εξέλιξη της νόσου.

3 Ανάλυση συνόλου δεδομένων

Το σύνολο δεδομένων που επεξεργαζόμαστε είναι το "`diabetes_dataset_ total.csv`". Προέκυψε από την συνένωση 3 διαφορετικών αρχείων *.csv*, που περιείχαν στοιχεία από 3 διαφορετικές χρονιές(2014-2015-2022) και αποτελείται συνολικά από 30 μεταβλητές που έχουν πλήθος εγγραφών 247598, η κάθε μία. Όλες οι παρακάτω μεταβλητές , έχουν τροποποιηθεί με τέτοιο τρόπο, ώστε να είναι ακέραιες.

Οι πρώτες μεταβλητες που εξετάζουμε προσδιορίζουν το γενετικό φύλο του ανθρώπου (Άντρας(0)-Γυναίκα(1)-(sex)) καθώς και την ηλικιακή ομάδα στην οποία ανήκει (age group). Οι ηλικιακές ομάδες χωρίζονται σε 6 κατηγορίες και αφορούν μόνο περιπτώσεις ανθρώπων σε διάφορες φάσεις της ενήλικης ζωής τους. Έπειτα, έχουμε στοιχεία που μας βοηθούν να εξέτασουμε την κοινωνική κατάσταση των ανθρώπων. Συγκεριμένα, το dataset περιλαμβάνει τις μεταβλητές `education`, `has insurance`, `income_group`, οι οποίες περιλαμβάνουν στοιχεία σχετικά με το επίπεδο σπουδών των ανθρώπων , αν έχουν ασφάλεια ή όχι, κάθως και σε ποια οικονομική κλίμακα ανήκουν με βάση το εισόδημα τους. Κλείνοντας, με τις μεταβλητές που εξετάζουν το κοινωνικό υπόβαθρο έχουμε τις μεταβλητές `race` και `marital_status`. Η `race` δείχνει την καταγωγή ή την φυλή των ανθρώπων , ενώ η `marital_status` περιγράφει τη κάτασταση του άτομου όσο αναφορά τη συντροφική ή συζηγική του ζωή.

Index	**Age Group**
0	18-24
1	25-34
2	35-44
3	45-54
4	55-64
5	65 and above

Index	**Education**
0	did not graduate High School
1	Graduated High School
2	Attended College or Technical School
3	Graduated from College or Technical School

Table 1: Education Levels

Index	**Has Insurance**
0	Do not have some form of health insurance
1	Have some form of health insurance

Table 2: Insurance Status

Index	**Income Group**
0	Less than 15,000
1	15,000 to $<$ 25,000
2	25,000 to $<$ 35,000
3	35,000 to $<$ 50,000
4	50,000 to $<$ 100,000
5	100,000 to $<$ 200,000
6	200,000 or more

Table 3: Income Groups

Οι παραπάνω μεταβλητές είναι γενικού χαρακτήρα και αν και προσφέρουν πλούσια πληροφορία σχετικά με την κοινωνική κατάταση του ατόμου δεν σχετίζονται άμεσα με την πρόκληση διαβήτη.

Index	Race
0	White
1	Black only, non-Hispanic
2	American Indian or Alaskan Native only, Non-Hispanic
3	Asian only, non-Hispanic
4	Native Hawaiian or other Pacific Islander only, Non-Hispanic
5	Multiracial, non-Hispanic
6	Hispanic

Table 4: Racial Groups

Index	Marital Status
0	Married
1	Divorced
2	Widowed
3	Separated
4	Never married
5	Member of unmarried couple

Table 5: Marital Status

Η general health είναι μία ακέραια μεταβλητή που πέρνει τιμές από 0 έως 4. Ένα άτομο που έχει τιμή 0 είναι εξαιρετικά υγιής, ενώ ένα άτομο με τιμή 4, δεν είναι υγιής.

Index	General Health
0	Excellent health
1	Very good health
2	Good health
3	Fair health
4	Poor health

Table 6: General Health Status

Έπειτα, έχουμε την `exercise lately` που είναι μία δίτιμη μεταβλήτη και δείχνει αν το άτομο έχει κάνει γυμναστική τουλάχιστον 1 φορά ή όχι τις τελευταίες 30 ημέρες. Ακολουθούν, οι μεταβλητές smoking, heavy drinker.

Η `smoking`: Παίρνει τιμές από 0 έως 4, ως εξής:

Index	Smoking Status
0	Current smoker, everyday
1	Current smoker, occasionally
2	Former smoker
3	Never smoked

Table 7: Smoking Status

Η μεταβλήτη `heavy_drinker` προσδιορίζει αν τα άτομα, πίνουν πολύ. Heavy drinker θεωρείται ο άντρας που πίνει πάνω από 14 ποτά την εβδομάδα ή η γυναίκα που πίνει πάνω από 7 ποτά την εβδομάδα.

Οι μεταβλητές `height`, `weight`, `sleep_time` , δείχνουν το ύψος, το βάρος και το πόσες ώρες την ημέρα κοιμάται το άτομο.
Μία ακόμη σημντική παράμετρος , είναι η μεταβλητή `bmi`. Ο bmi είναι ένας δείκτης που χρησιμοποιείται για την εκτίμηση του σωματικού λίπους με βάση το ύψος και το βάρος ενός ατόμου.Χωρίζεται σε 4 κατηγορίες ως εξής:
`bmi_groups`: Four-categories of Body Mass Index (BMI)

Οι μεταβλητες `has_doctor`, `affords_doctor`, σχετίζονται με την ύπαρξη σε

Index	BMI Category
0	Underweight (BMI $<$ 18.5)
1	Normal weight (18.5 $\leq$ BMI $<$ 25)
2	Overweight (25 $\leq$ BMI $<$ 30)
3	Obese (BMI $\geq$ 30)

Table 8: BMI Categories

προσωπικό γιατρό έχει το κάθε άτομο και την πρόσβαση σε περίθαλψη από οικονομικής άποψης.

Index	Have Personal Health Care Provider?
0	Yes, only one
1	More than one
2	No

Table 9: Personal Health Care Provider Status

Index	Could Not Afford To See Doctor
0	No
1	Yes

Table 10: Affordability to See Doctor

Η μεταβλητή `last_check_up` δείχνει το πόσος καιρός έχει περάσει από την τελευταία φορά που το άτομο έκανε εξέταση. Οι τιμές διανέμονται από 0 έως 3, ως εξής:

Index	Length of Time Since Last Routine Checkup
0	Within past year
1	Within past 2 years
2	Within past 5 years
3	5 or more years ago
4	Never

Table 11: Length of Time Since Last Routine Checkup

Η μεταβλητή `heart_problem` δείχνει αν το άτομο έχει υποστεί ποτέ καρδιακό επεισόδιο (1-NAI—0-OXI), ενώ η μεταβλητή `heart_history` δείχνει το ιστορικό του ατόμου με το άσθμα.Οι τιμές κυμαίνονται από 0 έως 2 , ως εξής: 0, ποτέ. 1, το άτομο έχει τώρα άσθμα. 2, το άτομο το έχει ξεπεράσει.

Η μεταβλητή stroke δείχνει αν το άτομο έχει περάσει ποτέ εγκεφαλικό (1-NAI——0-OXI), ενώ η μεταβλητή `depression` δείχνει αν το άτομο έχει περάσει ποτέ κατάθλιψη (1-NAI——0-OXI).

Υπάρχουν κάποιες εξετάσεις, οι οποίες οφείλουν να πραγματοποιούνται τακτικά από τους ανθρώπους, διότι μπορούν να ανακάλυψουν κάποιο σημαντικό πρόβλημα υγείας και να βοηθήσουν στην πρόληψη και στην αντιμετώπιση περαιτέρω προβλημάτων. Οι εξετάσεις αυτές είναι οι εξετάσεις αίματος και στο σύνολο δεδομένων μας εμφανίζονται με τη μορφή της μεταβλητής blood sugar. Η μεταβλητή αυτή δείχνει πότε ήταν η τελευταία φορά που το άτομο έκανε εξετάσεις αίματος για να ελέγξει το επίπεδο της γλυκόζης στον οργανισμό του. Παίρνει τιμές από 0 έως 6, ως εξής:

Index	Last Eye Exam Where Pupils Were Dilated
0	Never
1	Within the past month
2	Within the past year
3	Within the past 2 years
4	2 or more years ago

Table 12: Last Eye Exam with Pupil Dilation

Η μεταβλητή prediabetes δείχνει αν το άτομο έχει διαγνωσθεί πότε από γιατρό ως διαβητικός (0-OXI——1-NAI), ενώ η μεταβλητή diabetes type δείχνει τη κατάσταση του ατόμου όσο αναφόρα τον διαβήτη. Η μεταβλήτη αυτή παίρνει τιμές απο 0 έως 2, ως εξής:
$0 - >$ μη διαβητικός
$1 - >$ διαβήτης τύπου 1
$2 - >$ διαβήτης τύπου 2

Η μεταβλητή currently insulin προσδιορίζει αν στο άτομο χορηγείται τώρα ινσουλίνη (1-NAI——0-OXI), ενώ η μεταβλητή eye exam δείχνει πότε ήταν η τελευταία φορά που το άτομο επισκεφθηκε τον οφθαλμιάτρο.Η eye exam παίρνει τιμές από 0 έως 4, ως εξής:

Index	Last Eye Exam Where Pupils Were Dilated
0	Never
1	Within the past month
2	Within the past year
3	Within the past 2 years
4	2 or more years ago

Table 13: Last Eye Exam with Pupil Dilation

Η μεταβλητή eye photo προσδιορίζει πότε ήταν η τελευταία φορά που τράβηξαν φωτογραφία από το πίσω μέρος του ματιού. Η eye photo παίρνει τιμές από 0 έως 4, ως εξής:

Index	When was the last time they took a photo of the back of your eye?
0	Never
1	Within the past month
2	Within the past year
3	Within the past 2 years
4	2 or more years ago

Table 14: Eye Photo History

Η μεταβλητή diabetes education προσδιορίζει πότε ήταν η τελευταία φορά που το άτομο παρακολούθησε κάποια διάλεξη ή μάθημα σχετικά με το πως να αντιμετωπίζει το διαβήτη. Η diabetes education παίρνει τιμές από 0 έως 6, ως εξής:

Ολοκληρώνοντας με τη παρουσίαση του συνόλου δεδομένων έχουμε 2 τελευταίες μεταβλητές.Τη μεταβλητή sore feet και τη μεταβλητή diabetes. Η sore feet, προσδιορίζει αν το άτομο είχε πόνο ή ενοχλήσεις στα πόδια για περισσότερο από 4 εβδομάδες (1-NAI——0-OXI).Η diabetes αποκλύπτει αν το άτομο έχει διαβήτη,

Index	When was the last time you took a course or class in how to manage your diabetes?
0	Never
1	Within the past year
2	Within the last 2 years
3	Within the last 3 years
4	Within the last 5 years
5	Within the last 10 years
6	10 years ago or more

Table 15: Diabetes Education

και αν ναι, από μορφή της ασθένειας πάσχει. Η μεταβλητή αυτή παίρνει 3 τιμές, από 0 έως 2, ως εξής:
$0 \rightarrow$ Όχι, δεν έχει ή έχει άλλα είναι γυναίκα κατά τη διάρκει της εγκυμοσύνης της.
$1 \rightarrow$ Προ-διαβητικός ή στο border line του διαβήτη
$2 \rightarrow$ Ναι, έχει διαβήτη

4 Exploratory Data Analysis

Απαραίτητο βήμα τόσο στην ανάλυση των δεδομένων του dataset, όσο και στην μετέπειτα δημιουργία μοντέλων μηχανικής μάθησης για την πρόβλεψη διαβήτη είναι η κατανόηση του συνόλου δεδομένων και συγκεκριμένα των επεξηγηματικών μεταβλητών και της σχέσης αυτών με την εξαρτημένη μεταβλητή. Εκπονούμε λοιπόν, μια διερευνητική ανάλυση των δεδομένων (exploratory data analysis - EDA).

4.1 Τύπος δεδομένων

Τα περισσοτερα features τα οποία περιέχονται στο dataset είναι κατηγορικές μεταβλητές. Το γεγονός αυτό, εισάγει κάποιες δυσκολίες στη δημιουργία μοντέλων μηχανικής μάθησης. Μία από τις πιο σημαντικές από αυτές είναι ότι αποκλείει τις περισσότερες από τις πιο δημοφιλής μεθόδους clustering, οι οποίες θα μπορούσαν να χρησιμοποιηθούν για την εύρεση χρήσιμων patterns σε υποσύνολα των δεδομένων.

Επιπλέον, η επιλογή από τους δημιουργούς του dataset (που πραγματοποίησαν την τηλεφωνική δειγματοληψία) να καταγράψουν τα αποτελέσματα σε διακριτές κατηγορίες, π.χ. ηλικιακα groups αντί για ηλικία, σημαίνει πως το granularity των δεδομένων μας δεν είναι το επιθυμητό και σημαντικο ποσοστό του variance των κατανομών χάθηκε. Αυτό το έξτρα variance δυνητικά θα μπορούσε να συμβάλει καταλυτικά στην δημιουργία ενός μοντέλου με καλύτερη προβλεπτική ικανότητα.

4.2 Class Imbalance

Στο παρόν dataset, υπάρχει έντονη ανισσοροπία των κλάσεων, με την κλάση των υγειών να έχει σημαντικά περισσότερα δείγματα από τις υπόλοιπες. Το φαινόμενο αυτό, βέβαια είναι και το αναμενόμενο καθώς οι διαβητικόι και προδιαβητικοί ειναι μικρό ποσοστό το πληθυσμού. Σε επόμενα κομμάτια της παρούσας εργασίας, θα εξερευνήσουμε τεχνικές για διόρθωση αυτής της ανισσοροπίας, όπως είναι ο εμπλουτισμός των μη υγιών κλάσεων με δεδομένα από άλλες χρονιές που πραγματοποιήθηκε η ίδια έρευνα.

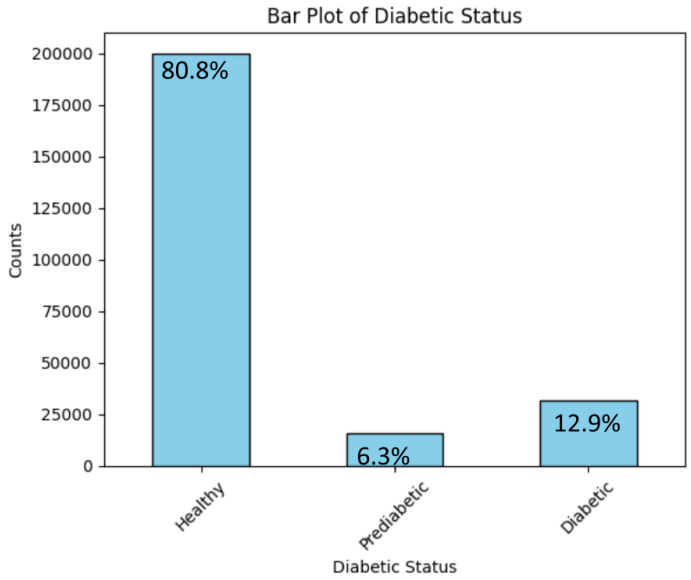


Figure 1: Συχνότητα κλάσεων στο δείγμα

4.3 Ανάλυση Μιας Μεταβλητής

Εξετάζεται η κατανομή των επεξηγηματικων κατανομών στις κλάσεις τους. Τα αποτελέσματα παρουσιάζονται στην Εικόνα **??**.

Από τα παραπάνω μπορούμε να εξάγουμε διάφορες ενδιαφέρουσες παρατηρήσεις:

Κάποιες από τις επεξηγηματικές μεταβλητές παρουσιάζουν κατανομή η οποία προσομοιάζει την κανονική. Τέτοιες είναι το BMI, ύψος και βάρος.

Ενδιαφέρον παρουσιάζει η κατανομή των Age Groups, η οποία είναι μετατοπισμένη προς τις κατηγορίες 4 και 5, που αντιπροσωπεύουν ηλικίες άνω των 55 ετών. Αυτό ίσως να οφείλεται στον τρόπο δειγματοληψίας μέσω τηλεφώνου: νεότερες γενίες πιθανότατα να συμμετήχαν πιο εύκολα σε κάποια έρευνα πχ μέσω διαδικτύου. Επιπλέον, στην κατηγορία Race υπάρχει υπερεκπροσώπηση της κατηγορίας των Λευκών, με άλλες εθνοτικές ομάδες όπως οι Αφροαμερικάνοι και οι Λατίνοι να μην εμφανίζονται σε αντίστοιχο ποσοστό. Τέτοιου είδους ανισσοροπίες στα δεδομένα, εισάγουν bias στα μοντέλα ταξινόμησης, κάτι το οποίο είναι ανεπιθύμητο.

Τέλος, η πλειοψηφία των ερωτηθέντων δηλώνει πως διαθέτει προσωπικό/ οικογενειακό γιατρό και έχει πρόσβαση σε ιατρική περίθαλψη. Αυτό διαφένεται στα διαγράμματα των μεταβλητών '`has_doctor`' και '`afford_doctor`'.

4.4 Σχέσεις ανεξάρτητων μεταβλητών με την έξοδο - Bivariate Analysis

Συνεχίζουμε την ανάλυση εξετάζοντας την κατανομή επιλεγμένων επεξηγηματικών μεταβλητών, ανάλογα με την τρέχουσα κατάσταση διαβήτη του εκάστοτε ατόμου. Εστιάζουμε κυρίως σε επεξηγηματικές μεταβλητές που σχετίζονται με το lifestyle και κοινωνικους-οικονομικούς παράγοντες. Τα αποτελέσματα δίνονται στα γράφηματα **??** και **??**.

Στα παραπάνω διαγράμματα, διαφένεται ένα βασικό πρόβλημα του συγκεκριμένου dataset, το οποίο δυσχεραίνει σημαντικά τη διαδικασία ταξινόμησης. Συγκεκριμένα, τα διάφορα features, εμφανίζουν μεγάλο βαθμό αλληλοεπικάλυψης μεταξύ τους για διαφορετικές τιμές της εξόδου (που στην προκειμένη περίπτωση είναι η νόσηση η όχι από διαβήτη η προδιαβήτη). Επομένως, οι discriminative αλγόριθμοι δε μπορούν να διακρίνουν το όριο απόφασης, ώστε να παράγουν ασφαλείς προβλέψεις.

Χαρακτηριστικό παράδειγμα αποτελεί το bmi, το οποίο αν και έχει διάφορα επιθυμητα χαρακτηριστικά όπως μεγάλο feature importance όταν χρησιμοποείται σε μοντέλα ταξινόμησης και υψηλή τιμή του συντελεστή συσχέτισης, εξακολουθεί

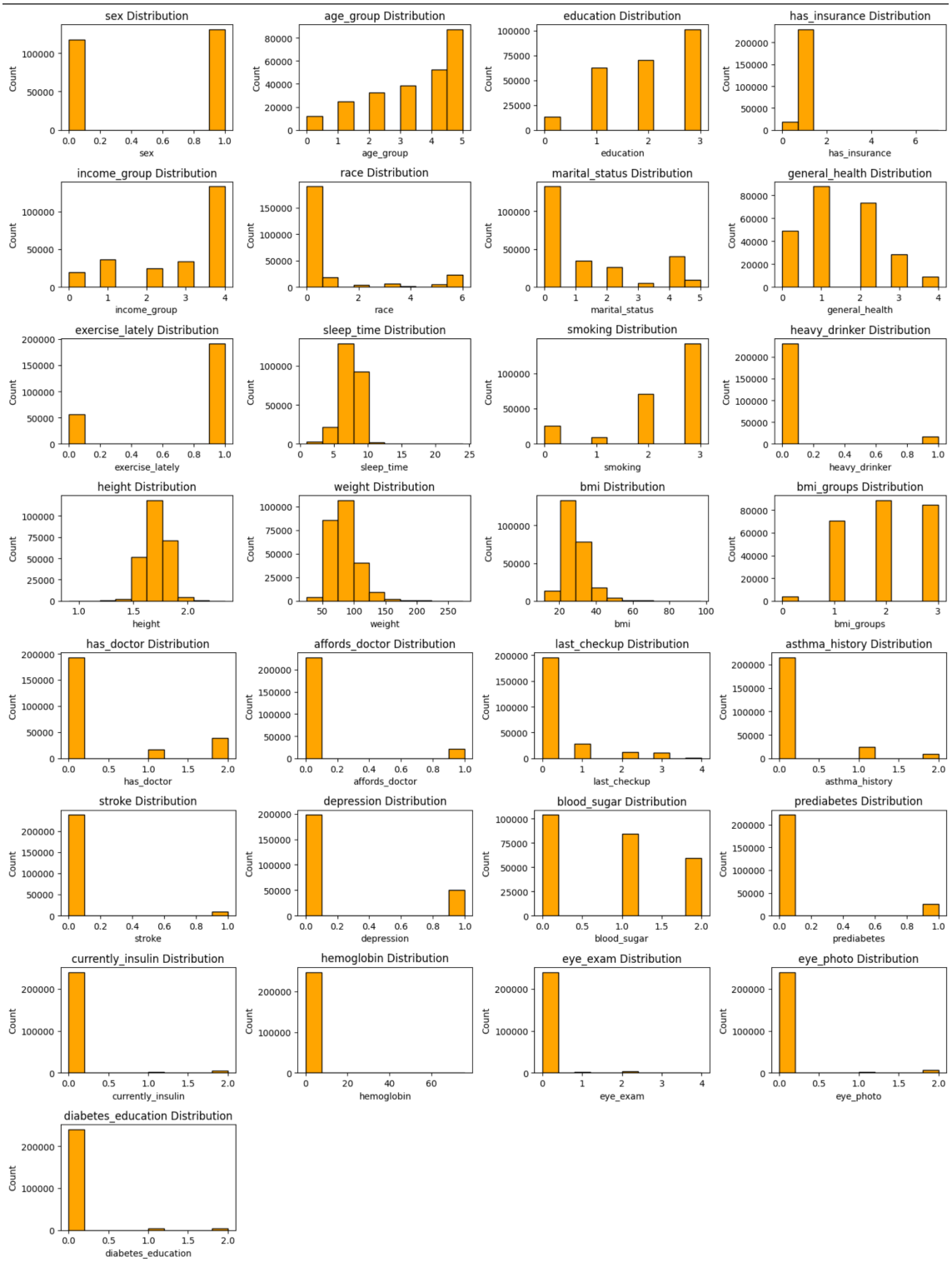


Figure 2: Κατανομές εξαρτημένων μεταβλητών

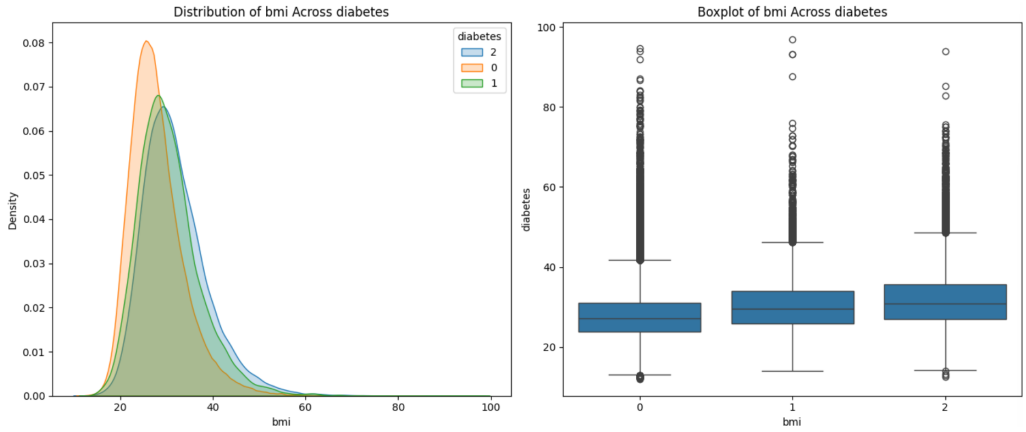


Figure 3: Κατανομή BMI για υγιής, διαβητικούς και προδιαβητικούς

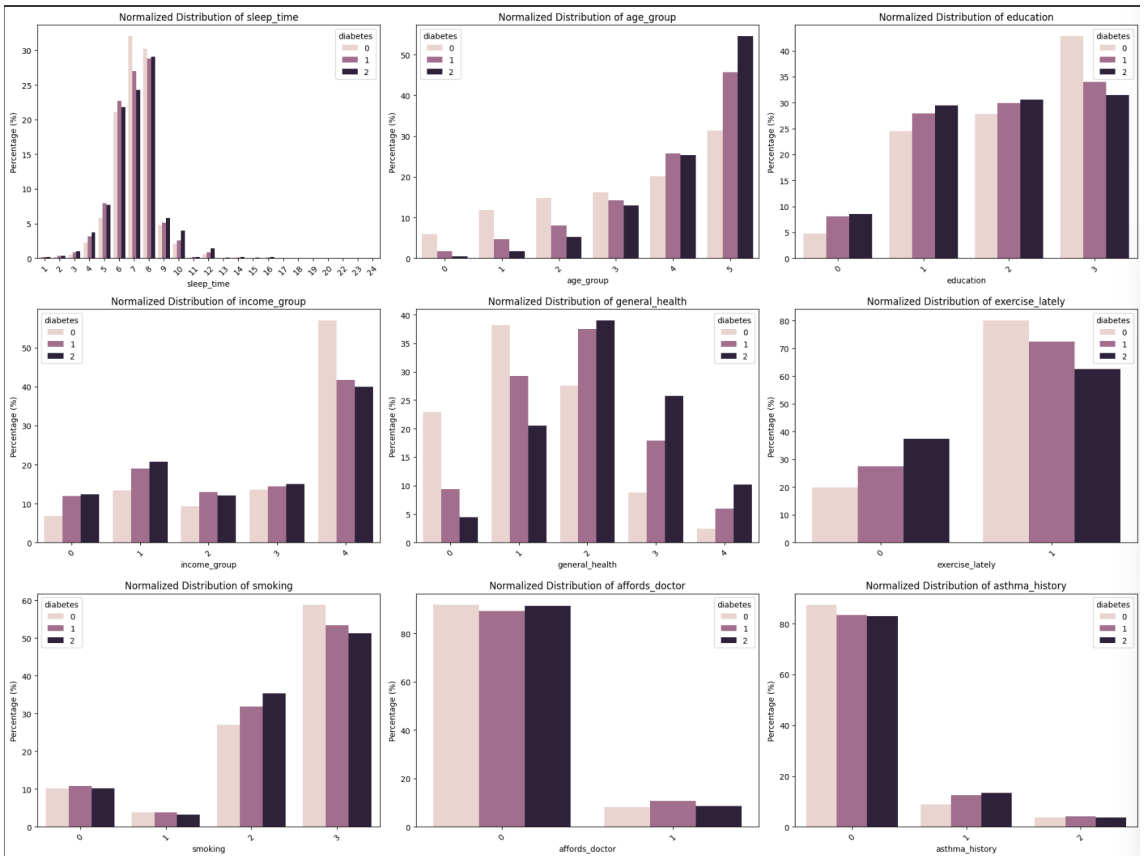


Figure 4: Κατανομή επιλεγμένων κατηγορικών μεταβλητών για υγιής, διαβητικούς και προδιαβητικούς

να παρουσιάζει μεγάλη επικάλυψη στις τρεις κατανομές για τις 3 κλάσεις της μεταβλητής εξόδου.

Θα επιχειρήσουμε σε αυτό το σημείο να δώσουμε μια πιθανή εξήγηση για αυτό το φαινόμενο. Από την ιατρική βιβλιογραφία, είναι κοινα αποδεκτό ότι η κακή διατροφή και παχυσαρκία αυξάνουν την πιθανότητα εμφάνισης διαβήτη. Ωστόσο, τα δεδομένα μας αποτελούν ένα στιγμιότυπο στη ζωή ενός ατόμου. Δεν διαθετουμε δεδομένα που παρακολουθούν έναν ασθενή σε βάθος χρόνου. Επομένως, δεν αποτυπώνεται με κάποιον τρόπο, πως ένα άτομο που αυτή τη στιγμή έχει επιβαρυντικο lifestyle αλλά είναι υγιές, μπορεί να εμφανίσει στο μέλλον διαβήτη. Αντίστοιχα, ένα άτομο που έχει διαβήτη αλλά έχει προσαρμώσει τον τρόπο ζωής του σε υγιεινές συνήθειες, μπορεί να μετακινεί τα δεδομένα με τέτοιο τρόπο ώστε το "προφανές" συμπέρασμα ότι κακό lifestyle προκαλέι διαβήτη να μην μεταφράζεται στο μοντέλο μας.

Προκύπτει επομένως το ερώτημα: Ποιο ποσοστό των ανθρώπων οι οποίοι έχουν διαγνωστεί με διαβήτη ή προδιαβήτη συνεχίζουν να έχουν στην ζωή τους κάποιο επιβαρυντικό συνήθειο, όπως η υψηλή κατανάλωση αλκοολ, η έλλειψη άσκησης, το υψηλό bmi και το κάπνισμα; Η ανάλυση μας δείχνει πως το 90.2% των διαβητικών και προδιαβητικών συμμετέχει σε μία από αυτές τις δραστηριότητες, οι οποίες δυσχεραίνουν την αντιμετώπιση της ασθένειας τους. Αντίθετα, το ποσοστό αυτό είναι 76.6% στους υγιής, το οποίο δείχνει, πως εκ πρώτης όψεως οι διαβητικοί και προδιαβητικοί έχουν πιο ανθυγιεινά συνήθεια. Ωστόσο, ένα ποσοστό της τάξης του 10% των διαβητικών και προδιαβητικών, ακολουθούν ιδιαίτερα ισορροπήμένο και υγιές lifestyle.

Συνεπιπρόσθετα, έχουν συμπεριληφθεί και διαγράμματα που εξετάζουν τη συσχέτιση του bmi, του φύλου και της κατάστασης διαβήτη. Γενικά, από την βιβλιογραφία ([**1**], [**2**]) γνωρίζουμε πως οι άντρες είναι πιο επιρρεπής στην εμφάνιση της νόσου σε σχέση με τις γυναίκες και μάλιστα εμφανίζουν τη νόσο για χαμηλότερο bmi. Αυτό διαφένεται στο διάγραμμα **??**, όπου το median τον αντρών στην κατηγορία προδιαβητικών και διαβητικών είναι ελαφρώς χαμηλότερο από των γυναικών. Αντίθετα, στην κατηγορία των υγειών, οι άντρες έχουν ελαφρως υψηλότερο bmi.

Τέλος, στο γράφημα **??**, παρατηρούμε την κατανομή προδιαβητικων, διαβητικών και υγειών για διαφορετικούς συνδυασμούς ηλικίας και bmi. Για λόγους ευκρίνειας και δεδομένου ότι η μεταβλητή age group είναι κατηγορική, έχει προστεθεί jitter. Οι παρατηρήσεις που μπορούν να εξαχθούν εδώ, είναι πως η ηλικία φαίνεται να είναι σημαντικός παράγοντας στην εκδήλωση της νόσου. Προφανώς, υγιή άτομα εμφανίζονται σε όλο το φάσμα ηλικιών και bmi. Για πολύ υψηλές τιμές του bmi και γηραιότερες ηλικίες παρατηρούμε μεγάλο ποσοστό των συμμετεχόντων στην έρευνα να δηλώνουν πως έχουν διαγνωστεί με διαβήτη η προδιαβήτη. Επομένως, υγιή άτομα τα οποία ανήκουν σε κατηγορίες με υψηλή συχνότητα εμφάνισης διαβήτη, βρίσκονται σε υψηλό κίνδυνο να εμφανίσουν και αυτοι σημάδια της ασθένειας και θα πρέπει να παρακολουθούν την υγεία τους.

4.5 Συσχέτιση και Πολυσυγγραμικότητα

Μελετώνται ο συντέλεστης συσχέτισης Pearson και το VIF των επεξηγηματικών μεταβλητών, τόσο για την μεταξύ τους συσχέτιση και πολυσυγγραμικότητα αλλα και με την συσχέτιση με την μεταβλητή εξόδου.

Οι μεταβλητές `height`, `weight`, και `bmi`, έχουν υψηλή εξάρτηση μεταξύ τους και υψηλό δείκτη πολυσυγγραμικότητας (VIF¿5). Το bmi είναι δείκτης που προκύπτει από το βάρος και το ύψος. Συνεπώς, στα μοντέλα ταξινόμησης που θα ακολουθήσουν θα χρησιμοποιηθεί είτε το `bmi_groups` είτε το `bmi` και όχι οι μεταβλητές που αφορούν το βάρος και το ύψος.

Επεξηγηματικές μεταβλητές όπως το `eye_photo`, `currently_insulin`, `eye_exam` εμφανίζουν σημαντικό correlation με την μεταβλητή εξόδου, διότι σχετίζονται για εξετάσεις και θεραπείες που άτομα με διάγνωση διαβήτη η κίνδυνο διαβήτη θα κάνουν πιο συχνά από το μέσο άτομο.

Ακόμα, στο διάγραμμα **??**, που περιέχει συχνότητες κανονικοποιημένες ως προς το status διαβήτη, μπορούμε να παρατηρήσουμε τις σχετικές συχνότητες των κατηγοριών των επεξηγηματικών μεταβλητών για τις τρεις τιμές της εξόδου. Αυτό οδηγεί στην εξαγωγή χρήσιμων παρατηρήσεων για τα δεδομένα μας:

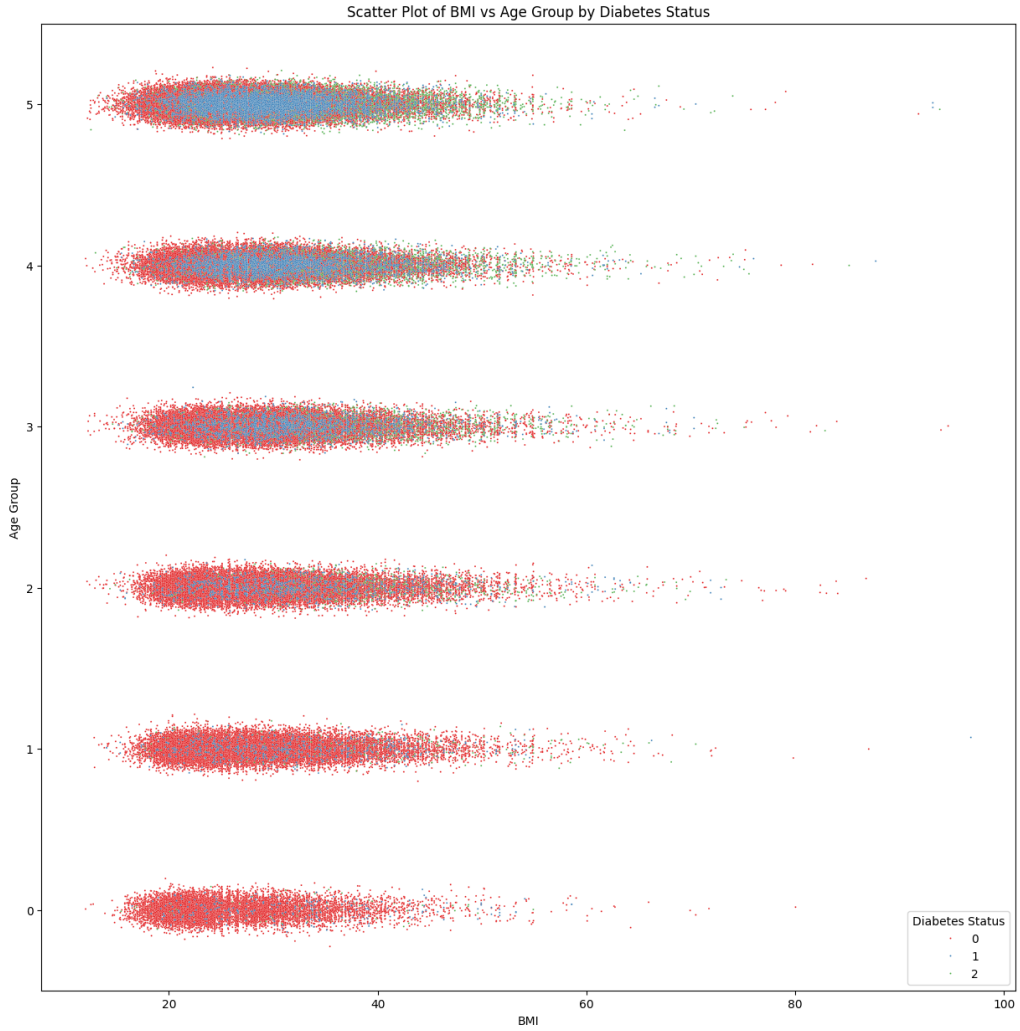


Figure 5: Scatter plot των BMI vs κατηγορία ηλικίας, όπου το χρώμα δηλώνει το στάτους διαβήτη. Στο γράφημα έχει χρησιμοποιηθεί gaussian jitter για υψηλότερη ευκρίνεια

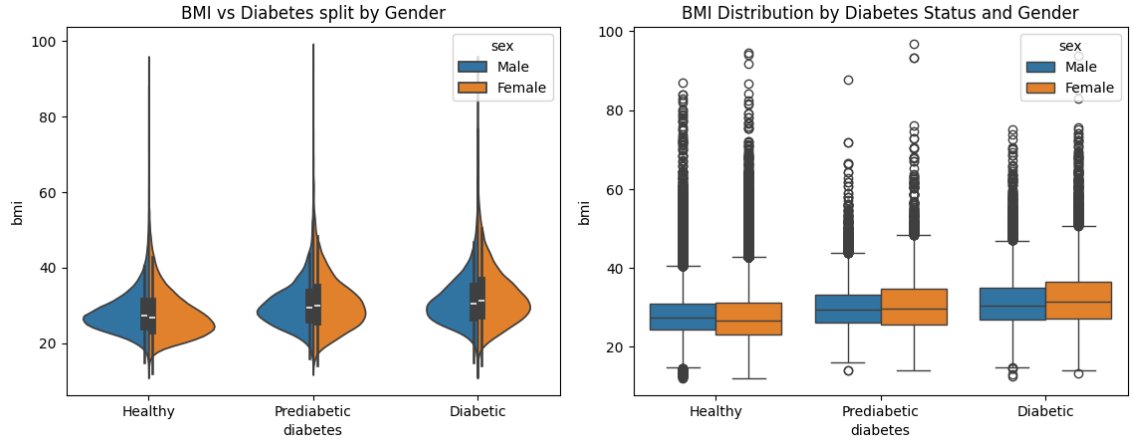


Figure 6: Violin plot των κατανομών BMI για τις διαφορετικές κατηγορίες διαβήτη, για τα 2 φύλα

Figure 7: Boxplot των κατανομών του BMI σε διαφορετικά status διαβήτη, για τα 2 φύλα

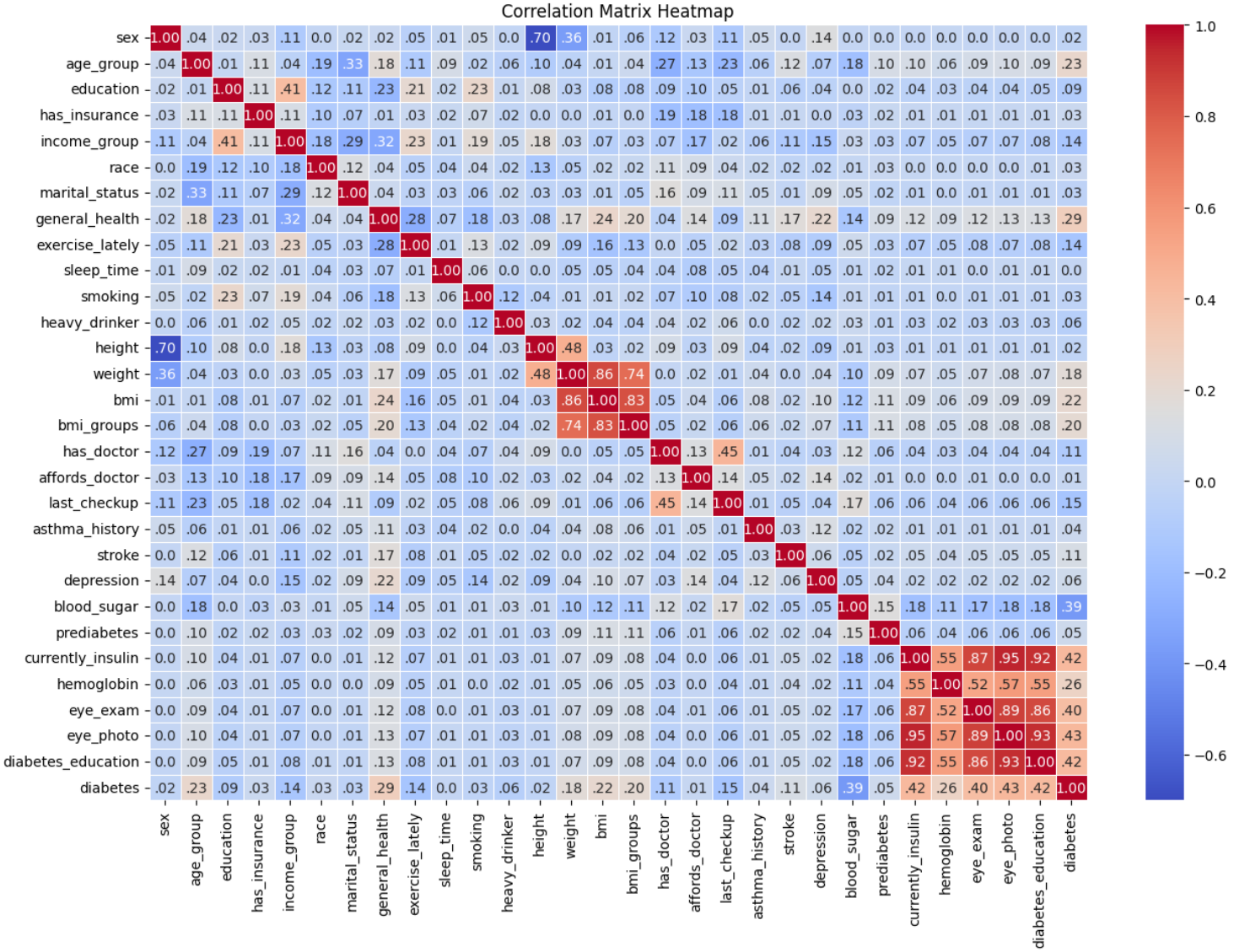


Figure 8: Πίνακας συσχέτισης του dataset

- 75% των ατόμων που δηλώνουν διαβητικοί, δεν παίρνουν αγωγή ινσουλίνης. Δεν έχουμε στοιχεία για να συμπεράνουμε αν αυτό συμβαίνει γιατί αντιμετωπίζουν το διαβήτη τους με άλλες μεθόδους, αν δεν έχουν πρόσβαση σε θεραπεία ή αν υπάρχει σημαντικό ποσοστό παραπλανήτικών η λανθασμένων απαντήσεων.

- Ένας παράγοντας ενδιαφέροντος για τη μέλετη μας, είναι η επιρροή του εισοδήματος στην εμφανισή και νόσηση από διαβήτη. Παρατηρούμε, λοιπόν και σε συνδυασμό με προηγούμενα γραφήματα, πως η κατηγορία των διαβητικών είναι πιο ομοιόμορφα κατανεμημένη στα διάφορα κοινωνικά στρώματα. Η κατηγορία των υγειών, έχει μια κατανομή που είναι μετατοπισμένη ελαφρώς προς μεγαλύτερα εισοδήματα.

- Τα άτομα με προδιαβήτη και διαβήτη δηλώνουν, κατα μέσο όρο, χειρότερη κατάσταση υγείας στην μεταβλητη `general_health`, σε σύγκριση με τους υγιής. Η υπόθεση, που μπορεί να δημιουργηθεί εδώ, είναι πως η εμφάνιση του διαβήτη μπορεί να συνοδευτεί από άλλες παθογένειες ή και ότι οι παράγοντες που οδηγούν στην ανάπτυξη διαβήτη όπως το αυξημένο βάρος και το κάπνισμα συντελούν στην εμφάνιση και άλλων ασθενιών όπως οι καρδιολογικές νόσοι.

- Οι διαβητικοί, εκ πρώτης όψεως φαίνεται να ασκούνται λιγότερο από τους

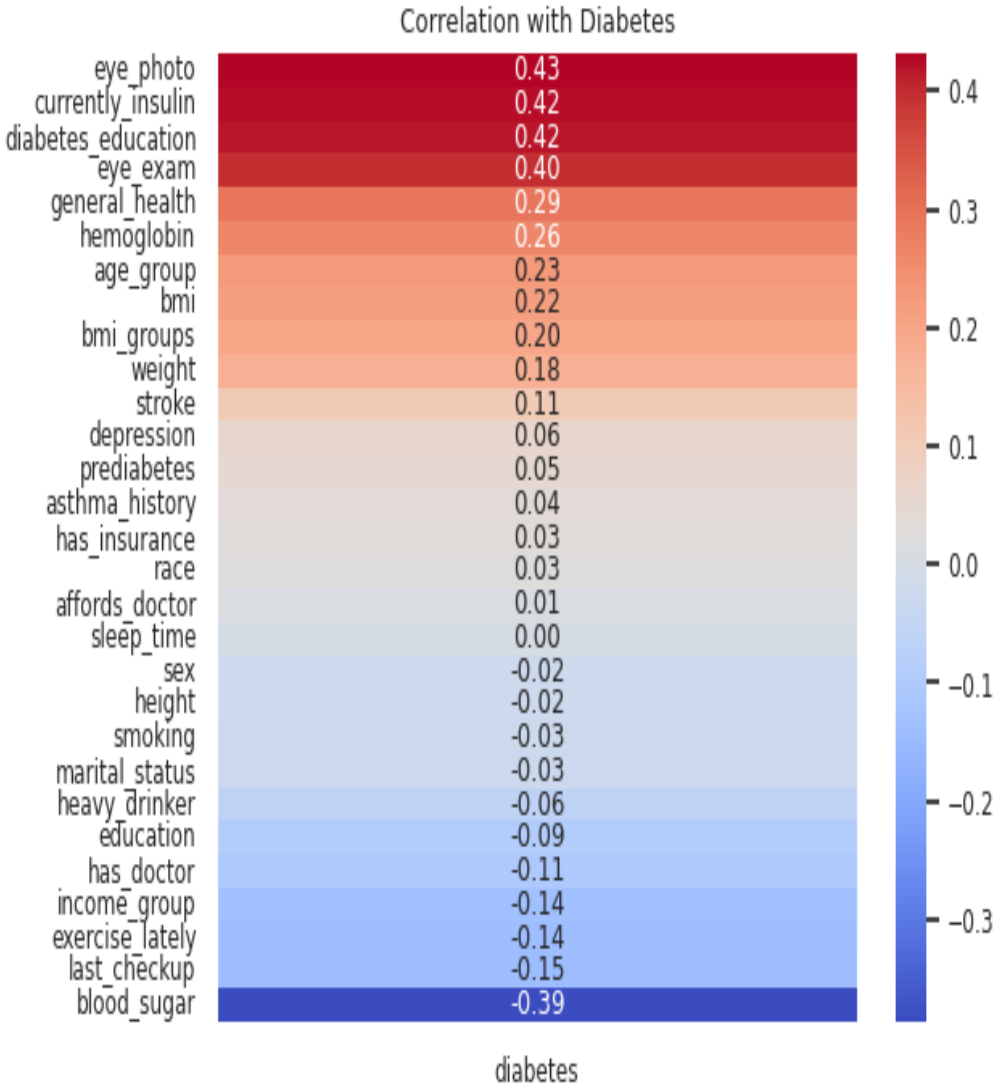


Figure 9: Correlation επεξηγηματικών μεταβλητών με την εξαρτημένη

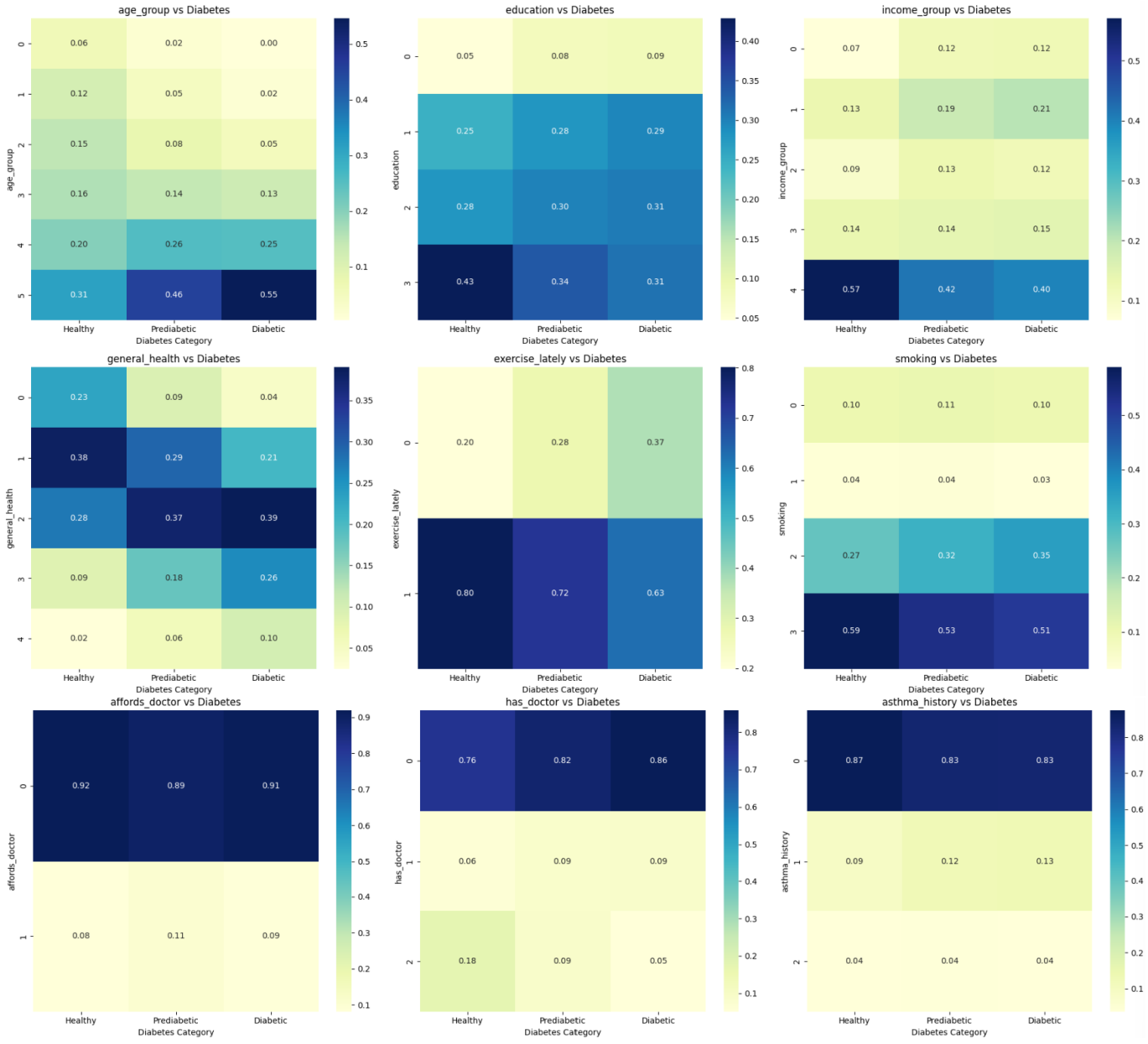


Figure 10: Contigency tables για ορισμένες επεξηγηματικές μεταβλητές με την μεταβλητή εξόδου

Table 16: Variance Inflation Factor (VIF) για επεξηγηματικές μεταβλητές

Μεταβλητές	**VIF**
const	5061.153045
weight	68.348726
bmi	52.721388
height	19.546290
eye_photo	14.569118
currently_insulin	12.409035
diabetes_education	8.575374

υγιής. Επιπλέον, εμφανίζουν υψηλότερα ποσοστά στην κατηγορία της καθημερινής κατανάλωσης αλκοολ στην μεταβλητή `heavy_drinking` από τους υγιής και υψηλότερα ποσοστά στην κατηγορία των υπέρβαρων στην μεταβλητή `bmi_groups`.

5 Preprocessing

Το preprocessing των δεδομενων αποτελείται από 3 διαδοχικά στάδια, το Feature, το Label και το Model, τα οποία θα αναλύσουμε παρακάτω.

5.1 Feature Preprocessing

Στη προεπεξεργασία των χαρακτηριστικών, δημιουργούμε νέα features με σκοπό να ανακαλύψουμε κρυμμένα υπάρχοντα μοτίβα στα δεδομένα. Δίνουμε ιδιαίτερη έμφαση σε πιθανούς συνδυασμούς του bmi με άλλα χαρακτηριστικά, καθώς έχει το μεγαλύτερο feature importance, όπως θα φανεί από τα αποτελέσματα. Να σημειωθεί ότι τα σύνθετα χαρακτηριστικά που δημιουργήσαμε, είναι αποτέλεσμα διαδικασίας trial & error. Συγκεκριμένα, δημιουργήθηκαν τα εξής σύνθετα χαρακτηριστικά:

- bmi groups: δημιουργεί περισσότερες κατηγορίες bmi, και συγκεκριμένα διαχωρίζει τα παχύσαρκα άτομα σε παχύσαρκα και νοσηρά παχύσαρκα.
- general health: μειώνει τις υπάρχοντες κατηγορίες του χαρακτηριστικού general health
- bmi age: δημιουργεί συνδυασμούς των bmi και age, δίνοντας έμφαση σε χρήστες με συνδυασμό υψηλών bmi και age
- bmi exercise: ενοποιεί τα χαρακτηριστικά bmi (i) και exercise lately (j), σε ένα ενιαίο χαρακτηριστικό της μορφής (i_j). Έπειτα διαγράφεται το exercise lately
- smoking alcohol: ενοποιεί τα χαρακτηριστικά smoking (i) και alcohol (j), σε ένα ενιαίο χαρακτηριστικό της μορφής (i_j). Έπειτα διαγράφονται τα επιμέρους χαρακτηριστικά
- education income: ενοποιεί τα χαρακτηριστικά education (i) και income (j), σε ένα ενιαίο χαρακτηριστικό της μορφής (i_j). Έπειτα διαγράφονται τα επιμέρους χαρακτηριστικά
- insurance healthcare: ενοποιεί τα χαρακτηριστικά insurance (i), healthcare (j) και has doctor (k), σε ένα ενιαίο χαρακτηριστικό της μορφής (i_j_k). Έπειτα διαγράφονται τα επιμέρους χαρακτηριστικά
- health problems: ενοποιεί τα χαρακτηριστικά stroke (i), depression (j) και asthma history (k), σε ένα ενιαίο χαρακτηριστικό της μορφής (i_j_k). Έπειτα διαγράφονται τα επιμέρους χαρακτηριστικά

Επιπλέον, στην προεπεξεργασία των χαρακτηριστικών, διαγράφονται χαρακτηριστικά του dataset, τα οποία αφορούν αποκλειστικά άτομα με διαβήτη και δεν μας είναι χρήσιμα (prediabetes, eye exam, currently insulin, blood sugar, hemoglobin, eye photo, diabetes education) καθώς και τα περιττά χαρακτηριστικά height και weight που εμπεριέχονται στο bmi και έχουν υψηλή συσχέτιση με αυτό.

5.2 Label Preprocessing

Στη στάδιο της προεπεξεργασίας των labels, μετατρέπουμε τα υπάρχοντα labels τριών διακριτών τιμών (μη-διαβητικοί/προδιαβητικοί/διαβητικοί) σε labels δύο διακριτών τιμών, ενοποιώντας τους προδιαβητικούς με τους διαβητικούς. Αιτία αυτού είναι η ύπαρξη ελάχιστων labels για την προδιαβητική κατηγορία, και η αδυναμία των ταξινομητών να διακρίνουν μεταξύ προδιαβητικών και διαβητικών.

5.3 Model Preprocessing

Στο τελευταίο στάδιο της προεπεξεργασίας, τα δεδομένα προετοιμάζονται για εκπαίδευση. Αρχικά, διαχωρίζονται σε train/test sets. Στη συνέχεια, αναλόγως την εκτέλεση, γίνεται εξισορροπηση του πλήθους δεδομένων ανά κατηγορία, με την τεχνική του Rare Class Sampling [**3**]. Δοκιμάστηκαν για balancing οι τεχνικές Smote, undersampling και Rare Class Sampling και από αυτές, επιλέξαμε να χρησιμοποιήσουμε στα τελικά πειράματα το Rare Class Sampling, καθώς προσφέρει τα καλύτερα αποτελέσματα ταξινόμησης. Οι τεχνικές αυτές εξηγούνται παρακάτω. Τέλος, γίνεται shuffle και Standard Scaling, ώστε τα δεδομένα να ανήκουν σε συγκεκριμένο εύρος, που βοηθά στην εκπαίδευση του μοντέλου.

6 Data Balancing

Το challenge με μη ισορροπημένα σύνολα δεδομένων είναι ότι οι περισσότερες τεχνικές μηχανικής μάθησης θα αγνοήσουν τη κλάση μειοψηφίας και στη συνέχεια θα έχουν κακή απόδοση σε αυτή, ενω υπάρχει πιθανότητα να αποτελεί και μια απο τις σημαντικότερες κλάσεις. Επιβάλλεται επομένως, η χρήση τεχνικών data augmentation και over/undersampling.

6.1 Συγχώνευση datasets

Όπως έχει ήδη αναφερθεί οι κλάσεις των προδιαβητικών και διαβητικών εμφανίζονταν σε πολύ μικρότερη συχνότητα από την τάξη των υγειών. Για την εξομάλυνση αυτού του ζητήματος, δημιουργήθηκε ένα νέο dataset, βασισμένο σε αυτό του 2022, στο οποίο προστέθηκαν επιλεκτικά δείγματα προδιαβητικών (της πλέον underrepresented κλάσης) από άλλες προηγούμενες χρονιές του 2014 και 2015. Για τη διατήρηση της συνοχής του νέου dataset και δεδομένου πως τα features διαφοροποιούντουσαν από χρονιά σε χρονιά, στο τελικό dataset έγινε χρήση της τομής των features από τα dataset και των τριών χρονιών. Μέσα από αυτή τη μέθοδο παρατηρήσαμε σημαντική βελτίωση στα αποτελέσματα των μοντέλων ταξινόμησης μας.

6.2 Smote

Μια προσέγγιση για την αντιμετώπιση των unbalanced datasets είναι η υπερδειγματοληψία του minority class. Η απλούστερη προσέγγιση περιλαμβάνει την αντιγραφή παραδειγμάτων στην κλάση μειοψηφίας, αν και αυτά τα παραδείγματα δεν προσθέτουν νέες πληροφορίες στο μοντέλο. Αντίθετα, νέα παραδείγματα μπορούν να συντεθούν από τα ήδη υπάρχοντα. Αυτός είναι ένα έιδος data augmentation τεχνικής για τη κλάση μειοψηφίας και αναφέρεται ως Τεχνική Υπερδειγματοληψίας Συνθετικής Μειονότητας ή SMOTE για συντομία.

Λεπτομερέστερα:

- Πώς το SMOTE συνθέτει νέα samples για το minority class.
- Πώς προσαρμόζονται και αξιολογούνται σωστά τα μοντέλα μηχανικής μάθησης σε σύνολα δεδομένων εκπαίδευσης που έχουν μετασχηματιστεί με SMOTE.
- Πώς μπορούν να χρησιμοποιηθούν επεκτάσεις του SMOTE που δημιουργούν συνθετικά παραδείγματα κατά μήκος του decision boundary της κλάσης.

Το SMOTE λειτουργεί επιλέγοντας samples που βρίσκονται κοντά στο χώρο χαρακτηριστικών, σχεδιάζοντας μια γραμμή μεταξύ των παραδειγμάτων στο χώρο χαρακτηριστικών και σχεδιάζοντας ένα νέο δείγμα σε ένα σημείο κατά μήκος αυτής της γραμμής. Συγκεκριμένα, αρχικά επιλέγεται ένα τυχαίο παράδειγμα από την κλάση μειοψηφίας. Στη συνέχεια, βρίσκονται οι k πλησιέστερους γείτονες για αυτό το παράδειγμα (συνήθως k=5). Επιλέγεται τυχαία ένας από τους k γείτονες και δημιουργείται ένα συνθετικό παράδειγμα σε ένα τυχαία επιλεγμένο σημείο μεταξύ των δύο αυτών παραδειγμάτων στον χώρο χαρακτηριστικών. Αυτή η διαδικασία μπορεί να χρησιμοποιηθεί για να δημιουργηθούν όσα συνθετικά παραδείγματα απαιτούνται για το minority class. Συνήθως, προτείνεται πρώτα η χρήση τυχαίας υποδειγματοληψίας για να τριμαριστεί ο αριθμός των samples στην κλάση πλειοψηφίας και μετά να χρησιμοποιηθεί το SMOTE για υπερδειγματοληψία της κλάσης μειοψηφίας, ώστε να προκύψει τελικά ένα πιο ισορροπημένο αποτέλεσμα.

Η σωστή εφαρμογή της υπερδειγματοληψίας κατά το k-fold είναι η εφαρμογή της μεθόδου μόνο στο σύνολο δεδομένων εκπαίδευσης και, στη συνέχεια, η αξιολόγηση του μοντέλου στο stratified αλλά μη μετασχηματισμένο σύνολο δοκιμών. Ακόμα θα μπορούσε να εξερευνηθεί η δοκιμή διαφορετικών αναλογιών της κλασης μειοψηφίας και της κλάσης πλειοψηφίας για να φανεί εάν είναι δυνατή μια περαιτέρω αύξηση στην απόδοση. Μια άλλη περιοχή που μπορεί να εξερευνηθεί θα ήταν να δοκιμαστούν διαφορετικές τιμές των k-πλησιέστερων γειτόνων που επιλέγονται στη διαδικασία SMOTE όταν δημιουργείται κάθε νέο συνθετικό παράδειγμα. Η προεπιλογή είναι k=5, αν και μεγαλύτερες ή μικρότερες τιμές θα επηρεάσουν τους τύπους των samples που δημιουργούνται και με τη σειρά τους μπορεί να επηρεάσουν την απόδοση του μοντέλου. Προτείνε η χρήση grid search για την εκάστοτε περίπτωση μοντέλων και συνόλου δεδομένων.

Μια δημοφιλής επέκταση στο SMOTE περιλαμβάνει την επιλογή εκείνων των περιπτώσεων της κατηγορίας μειοψηφίας που είναι εσφαλμένα ταξινομημένες, για παράδειγμα με χρήση μοντέλου k-πλησιέστερου γείτονα. Στη συνέχεια, μπορούμε να γίνει υπερδειγματοληψία μόνο αυτών των δύσκολων περιπτώσεων, παρέχοντας περισσότερη ανάλυση μόνο όπου μπορεί να απαιτείται για το συγκεκριμένο dataset. Τα samples που είναι εσφαλμένα ταξινομημένα είναι πιθανότατα διφορούμενα και μπορεί να βρίσκονται σε μια περιοχή του άκρου ή πάνω στο άκρο του ορίου απόφασης, όπου είναι δυνατό να υπάρχει overlap των δύο κλάσεων προς τα συγκεκριμένα σημεία.

6.3 Undersampling

Εδώ πραγματοποιείται απλώς με τυχαίο τρόπο υποδειγματοληψία του majority class, όσο θεωρείται πως απαιτείται από την φύση των δεδομένων. Γίνεται αντιληπτό πως, λόγω της τυχαιότητας της συγκεκριμένης μεθόδου, τα αποτελέσματα πιθανώς να διαφοροποιούνται αισθητά μεταξύ τους όταν αλλάζει το seed που παράγει τη ψευδοτυχαία επιλογή. Πολύ μεγάλη σημασία έχει αφενώς το πόσα δείγματα θα επιλεχθούν, ή αντίθετα το πόσα δείγματα θα "λείψουν" και άρα πόσο θα μεταβληθεί η υπάρχουσα κατανομή της κλάσης αυτής και αφετέρου, ακόμα και για επαρκή αριθμό εναπομείναντων δειγμάτων, αν αυτά συνεχίζουν να αντιπροσωπεύουν ορθά την αρχική κατανομή του majority class.

6.4 Rare Class Sampling

Στο dataset που επιλέξαμε, ακόμη και με την συγχώνευση προδιαβητικών με διαβητικούς, υπάρχει μεγάλη ανισότητα μεταξύ των 2 ετικετών. Για να αντιμετωπίσουμε το πρόβλημα αυτό, χρησιμοποιούμε την τεχνική Rare Class Sampling (RCS). Σε κάθε κλάση ανατίθεται μια συχνότητα

$f_c = \frac{\text{Total number of class } i \text{ labels}}{\text{Total number of labels}}$. Με βάση τις παραπάνω συχνότητες, οι πιθανότητες δειγματοληψίας κάθε κλάσης c, ορίζονται ως: $P(c) = \frac{e^{(1-f_c)/T}}{\sum_{c'=1}^{C} e^{(1-f_{c'})/T}}$. Επομένως, οι κατηγορίες με μικρότερη συχνότητα θα έχουν υψηλότερη πιθανότητα δειγματοληψίας. Η θερμοκρασία T ελέγχει την ομαλότητα της κατανομής. Υψηλότερο T οδηγεί σε μια πιο ομοιόμορφη κατανομή, ενώ χαμηλότερο T οδηγεί σε μεγαλύτερη εστίαση στις σπάνιες κατηγορίες με μικρό f_c.

7 Ταξινομητές

7.1 Random Forest

Ένας ταξινομητής Random Forest είναι μια μη-παραμετρική μέθοδος ταξινόμησης. Λειτουργεί κατασκευάζοντας πολλαπλά δέντρα αποφάσεων κατά τη διάρκεια της εκπαίδευσης και εξάγοντας τη συχνότερη κλάση των ατομικών δέντρων (ταξινόμηση). Κάθε δέντρο κατασκευάζεται χρησιμοποιώντας ολόκληρο το σύνολο των δεδομένων εκπαίδευσης αντί για ένα τυχαίο υποσύνολο, αλλά εξακολουθεί να χρησιμοποιεί ένα τυχαίο υποσύνολο χαρακτηριστικών, γεγονός που βοηθά στη μείωση της υπερεφαρμογής και στη βελτίωση της γενίκευσης. Όταν εισάγεται ένα νέο σημείο δεδομένων, περνάει μέσα από κάθε δέντρο στο δάσος, και κάθε δέντρο παρέχει μια ψήφο ταξινόμησης. Η κλάση με τις περισσότερες ψήφους από όλα τα δέντρα επιλέγεται ως η τελική πρόβλεψη. Αυτή η μέθοδος βελτιώνει την ακρίβεια και τη σταθερότητα του μοντέλου σε σύγκριση με ένα μόνο δέντρο αποφάσεων.

7.2 Balanced Random Forest

Το Balanced Random Forest είναι μια παραλλαγή του κλασικού ταξινομητή Random Forest που έχει σχεδιαστεί για να αντιμετωπίζει την ανισορροπία κατηγοριών στα σύνολα δεδομένων. Σε έναν τυπικό Random Forest, η ανισορροπία κατηγοριών μπορεί να οδηγήσει σε biased μοντέλα που ευνοούν την πλειοψηφική κατηγορία. Το Balanced Random Forest μετριάζει αυτό το πρόβλημα διασφαλίζοντας ότι κάθε δέντρο στο δάσος κατασκευάζεται με ένα ισορροπημένο υποσύνολο των δεδομένων εκπαίδευσης. Συγκεκριμένα, για κάθε δέντρο, γίνεται τυχαία υποδειγματοληψία της πλειοψηφικής κατηγορίας και/ή υπερδειγματοληψία της μειοψηφικής κατηγορίας για να δημιουργηθεί ένα ισορροπημένο σύνολο δεδομένων. Αυτή η διαδικασία διασφαλίζει ότι κάθε κατηγορία εκπροσωπείται εξίσου στην εκπαίδευση κάθε δέντρου, οδηγώντας σε πιο ισορροπημένες και δίκαιες προβλέψεις. Η κύρια διαφορά μεταξύ BRF και RF έγκειται σε αυτή τη στρατηγική δειγματοληψίας, η οποία βοηθά στη βελτίωση της απόδοσης του μοντέλου σε ανισομερή σύνολα δεδομένων, μειώνοντας τη μεροληψία προς την πλειοψηφική κατηγορία και βελτιώνοντας την ακρίβεια της ταξινόμησης της μειοψηφικής κατηγορίας.

8 Πειράματα

Σε αυτήν τη μελέτη, πραγματοποιήσαμε τρία ξεχωριστά πειράματα για να αξιολογήσουμε την απόδοση διαφορετικών τεχνικών data balancing και model configurations.

Η κύρια μέτρική αξιολόγησης για όλα τα πειράματα ήταν το recall, η οποία μετρά την ικανότητα του μοντέλου να αναγνωρίζει σωστά τις θετικές περιπτώσεις από το σύνολο των πραγματικών θετικών περιπτώσεων. Ιδιαίτερα μας ενδιαφέρει η μεγιστοποίηση του recall της κλάσης των διαβητικών (minority class), που

ισοδυναμεί με την ικανότητα του μοντέλου να ταξινομεί ως διαβητικούς, τους πραγματικά διαβητικούς, διότι θέλουμε να μηδενίσουμε την πιθανότητα να είναι κάποιος διαβητικός, και το μοντέλο να τον ταξινομήσει ως υγιή.

Οι ταξινομητές που χρησιμοποιήθηκαν ήταν οι Random Forest και Balanced Random Forest, που υλοποιήθηκαν με χρήση των βιβλιοθηκών sklearn και imblearn αντίστοιχα. Επιλέξαμε να χρησιμοποιήσουμε αποκλειστικά Tree Classifiers, καθώς το dataset μας αποτελείται από σχεδόν εξ' ολοκλήρου κατηγορικά δεδομένα. Επίσης, τα δένδρα παρέχουν τη δυνατότητα οπτικοποίησης της σημαντικότητας των features, γεγονός που βοηθάει στην εξαγωγή συμπερασμάτων. Μέσω των Bar Plots για τα feature importances, μπορούμε να παρατηρήσουμε για κάθε πείραμα ποια χαρακτηριστικά ήταν τα επιδραστικότερα για τον καθορισμό του μοντέλου.

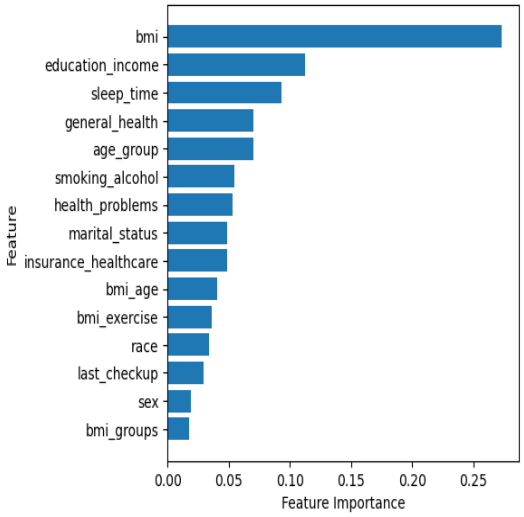
Σε κάθε πείραμα παραθέτουμε το classification report, καθώς και το barplot των Feature Importances αλλά και το confusion matrix.

8.1 Πείραμα 1

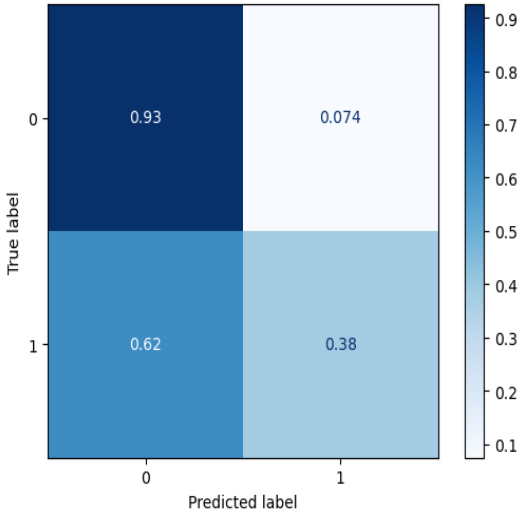
Στο πρώτο πείραμα, διατηρούμε το dataset ανέπαφο, και χρησιμοποιούμε Random Forest ως classifier. Τα αποτελέσματα μας είναι:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0 (Non-diabetics)	0.86	0.93	0.89	20003
1 (Diabetics)	0.55	0.38	0.45	4757
Accuracy			0.82	24760
Macro avg	0.70	0.65	0.67	24760
Weighted avg	0.80	0.82	0.81	24760

Table 17: Classification Report για πείραμα 1



() Feature Importance Bar Plot



() Confusion Matrix

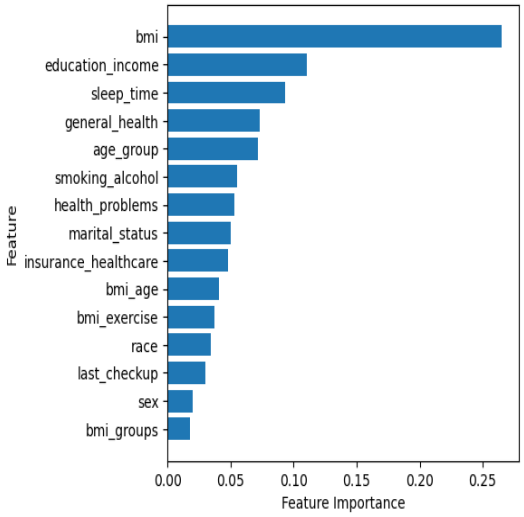
Figure 11: Αποτελέσματα πειράματος 1
lofsubfigure“numberline()“centering Feature Importance Bar Plotlofsubfigure“numberline()“centering Confusion Matrix

8.2 Πείραμα 2

Στο δεύτερο πείραμα, χρησιμοποιούμε το Rare Class Sampling για δειγματοληψία από το αρχικό dataset, και δημιουργούμε μοντέλο Random Forest. Τα αποτελέσματά μας είναι τα εξής:

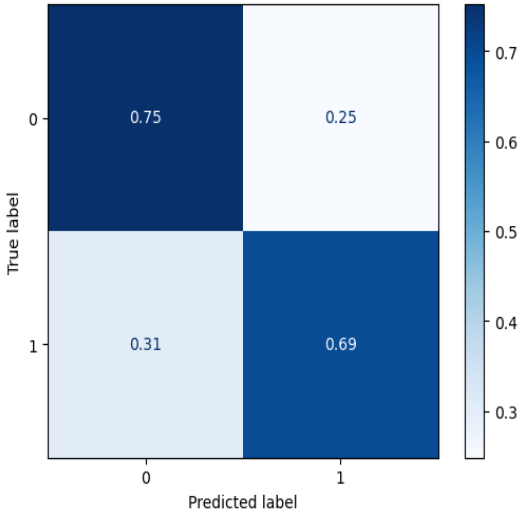
Class	Precision	Recall	F1-Score	Support
0 (Non-diabetics)	0.91	0.75	0.82	20003
1 (Diabetics)	0.40	0.69	0.51	4757
Accuracy			0.74	24760
Macro avg	0.66	0.72	0.67	24760
Weighted avg	0.81	0.74	0.76	24760

Table 18: Classification Report για πείραμα 2



Δ

() Feature Importance Bar Plot



Δ

() Confusion Matrix

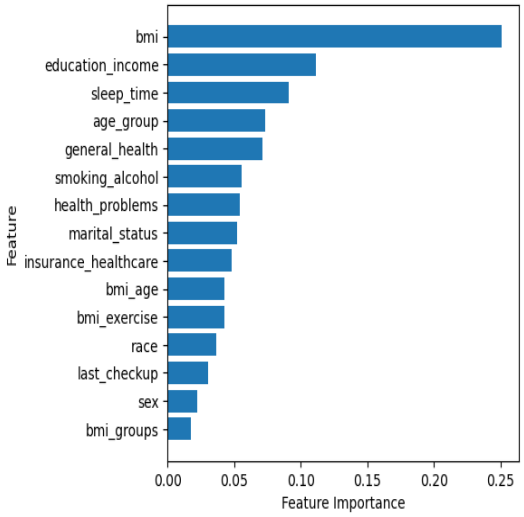
Figure 12: Αποτελέσματα πειράματος 2
lofsubfigure“numberline()“centering Feature Importance Bar Plotlofsubfigure“numberline()“centering Confusion Matrix

8.3 Πείραμα 3

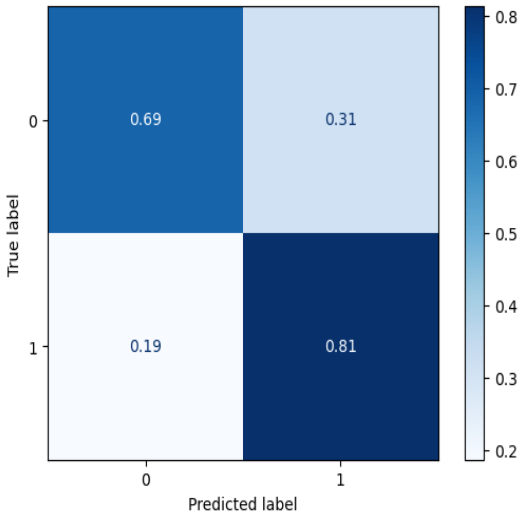
Στο τελευταίο πείραμα, χρησιμοποιούμε εκ νέου το αρχικό dataset, και χρησιμοποιούμε το Balanced Random Forest. Τα αποτελέσματά μας είναι τα εξής:

Class	Precision	Recall	F1-Score	Support
0 (Non-diabetics)	0.94	0.69	0.79	20003
1 (Diabetics)	0.38	0.81	0.52	4757
Accuracy			0.71	24760
Macro avg	0.66	0.75	0.66	24760
Weighted avg	0.83	0.71	0.74	24760

Table 19: Classification Report για πείραμα 3



() Feature Importance Bar Plot



() Confusion Matrix

Figure 13: Αποτελέσματα πειράματος 3
lofsubfigure“numberline()“centering Feature Importance Bar Plotlofsubfigure“numberline()“centering Confusion Matrix

9 Συμπεράσματα

Μέσα από τα μοντέλα ταξινόμησης μας συμπεραίνουμε τα εξής:

- Παρατηρούμε ότι το πείραμα 2, με data balancing, βελτιώνει το recall του πρώτου πειράματος για τη κατηγορία διαβητικών από 38% σε 69%, γεγονός που αναμέναμε. Ωστόσο, το βέλτιστο recall συναντάται στο πείραμα 3, με 81%, με balanced random forest και χωρίς να κάνουμε data balancing. Αυτό οφείλεται ενδεχομένως στον τρόπο που χειρίζεται την ανισότητα των κλάσεων εσωτερικά η βιβλιοθήκη imblearn, που είναι αποδοτικότερος από το manual resampling που υλοποιούμε στο πείραμα 2
- Για την εργασία μας, η σειρά σημαντικότητας των χαρακτηριστικών μένει σταθερή, ανεξαρτήτως πειράματος, με σημαντικότερα χαρακτηριστικά να είναι τα bmi (όπως αναμενόταν), το education income και το sleep time.
- Με την αύξηση του recall, που είναι η μετρική ενδιαφέροντος για την εργασία μας, παρατηρούμε αντίστοιχη μείωση του accuracy, γεγονός που οφείλεται

στην προσπάθεια του μοντέλου να "μάθει" καλύτερα το minority class, και έχει άμεσο αντίκτυπο στο accuracy του majority class, και άρα στο ολικό accuracy του μοντέλου.

Εν κατακλείδι, το φαινόμενο του διαβήτη είναι πολυσύνθετο, με αποτέλεσμα η ταξινόμηση ατόμων σε διαβητικούς η υγιής, να είναι ιδιαίτερα δύσκολή χωρίς την πρόσβαση σε ευαίσθητα ιατρικά δεδομένα, όπως αυτά που προκύπτουν από ιατρικές εξετάσεις.

Επιπλέον έρευνα, θα οφελούταν της πρόσβασης σε ιατρικά εξειδικευμένα δεδομένα, σε δεδομένα χρονοσειρών που παρακολουθούν την εξέλιξη της νόσησης σε ασθενείς σε βάθος χρόνου είτε της εστίασης σε συγκεκριμένες υποομάδες εντός του συνόλου δεδομένων μας που εμφανίζουν πιο ομοιόμορφη συμπεριφορά.

References

[1] Anna Nordström, Jenny Hadrévi, Tommy Olsson, Paul W. Franks, Peter Nordström, Higher Prevalence of Type 2 Diabetes in Men Than in Women Is Associated With Differences in Visceral Fat Mass, *The Journal of Clinical Endocrinology & Metabolism*, Volume 101, Issue 10, 1 October 2016, Pages 3740–3746, blue`https://doi.org/10.1210/jc.2016-1915`.

[2] Ciarambino, T., Crispino, P., Leto, G., Mastrolorenzo, E., Para, O., Giordano, M., Influence of Gender in Diabetes Mellitus and Its Complications, *International Journal of Molecular Sciences*, 2022 Aug 9;23(16):8850, blue`https://doi.org/10.3390/ijms23168850`.

[3] Lukas Hoyer, Dengxin Dai, Luc Van Gool, DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation, blue`https://arxiv.org/abs/2111.14887`.