

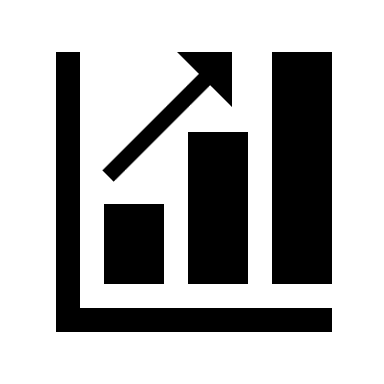
ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΣΤΑΤΙΣΤΙΚΑ ΠΡΟΓΡΑΜΜΑΤΑ 2

ΖΗΚΟΥ ΚΩΝΣΤΑΝΤΙΝΑ-ΓΕΩΡΓΙΑ (1547-70, 8ο εξάμηνο)

ΠΑΣΧΑΛΙΔΟΥ ΚΥΡΙΑΚΗ (1547-291 4ο εξάμηνο)



ΓΡΕΒΕΝΑ 2023

**ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ**

**Περίληψη ………………………………………………………..σελ. 3**

**Εισαγωγή …………………………………………………………σελ. 4**

**Περιγραφική Στατιστική Ανάλυση…………………….σελ. 6**

**Έλεγχος Στατιστικών Υποθέσεων………………………σελ. 16**

**Προβλεπτικά μοντέλα………………………………………σελ. 17**

**Συζήτηση………………………………………………………….σελ. 27**

**Περιορισμοί……………………………………………………..σελ. 28**

**Βιβλιογραφικές Αναφορές……………………………….σελ. 29**

**Παράρτημα………………………………………………………σελ. 30**

**ΠΕΡΙΛΗΨΗ**

**Η παρούσα εργασία έχει ως στόχος την υλοποίηση αλγορίθμων μηχανικής μάθησης με το μοντέλο logistic regression. Αυτό γίνεται με χρήση του προγραμματιστικού εργαλείου της R στον υπολογιστή. Για την υλοποίηση των αλγορίθμων απώτερος στόχος είναι η στατιστική ανάλυση και επεξεργασία των δεδομένων του συνόλου δεδομένων Heart failure clinic records.**

**ΕΙΣΑΓΩΓΗ**

«Εδώ κάνετε μια αναλυτική περιγραφή της βάσης δεδομένων και γράφετε ποιος είναι ο σκοπός της παρούσας μελέτης»

Η βάση δεδομένων που χρησιμοποιήθηκε είναι η βάση από το UCI Machine Learning Repository και αφορά το σύνολο δεδομένων Heart Failure Clinic Records. Το συγκεκριμένο σύνολο δεδομένων αφορά 299 παρατηρήσεις δηλαδή δεδομένα των ασθενών με 13 διαφορετικά χαρακτηριστικά-μεταβλητές. ΟΙ μεταβλητές αυτές είναι οι εξής:

- age: age of the patient (years)

- anaemia: decrease of red blood cells or hemoglobin (boolean)

- high blood pressure: if the patient has hypertension (boolean)

- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)

- diabetes: if the patient has diabetes (boolean)

- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)

- platelets: platelets in the blood (kiloplatelets/mL)

- sex: woman or man (binary)

- serum creatinine: level of serum creatinine in the blood (mg/dL)

- serum sodium: level of serum sodium in the blood (mEq/L)

- smoking: if the patient smokes or not (boolean)

- time: follow-up period (days)

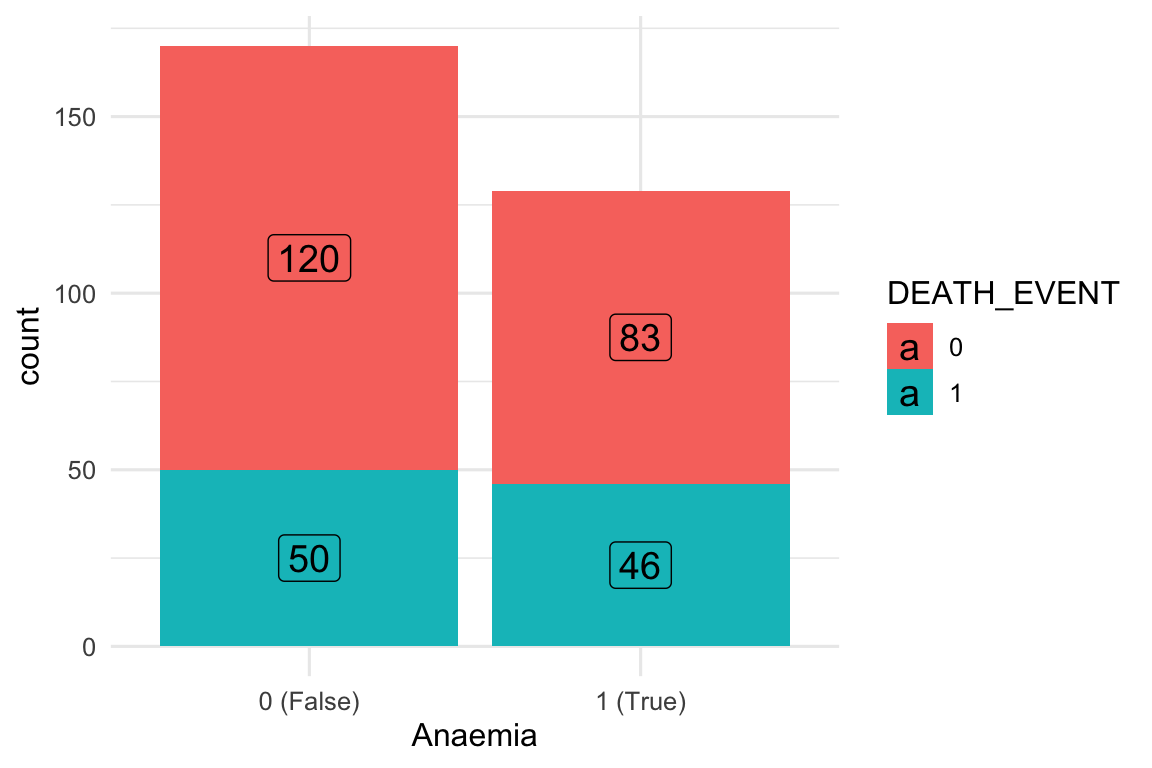
- [target] death event: if the patient deceased during the follow-up period (boolean)

Στην μεταβλητή death event την ορίζουμε ως target διότι είναι η μεταβλητή-χαρακτηριστικό που θέλουμε να προβλέψουμε.

**Περιγραφική Στατιστική Ανάλυση**

«Περιγράφετε στατιστικά όλες τις μεταβλητές της βάσης δεδομένων χρησιμοποιώντας τα κατάλληλα περιγραφικά μέτρα και διαγράμματα»

Αρχικά, περιγράφεται η μεταβλητή anaemia. Στην μεταβλητή αναιμία το πλήθος της μεταβλητής αυτής είναι 50 να είναι False και 120,ενώ True είναι 46 και 83.

0 Fals

Στην συνέχεια παρουσιάζεται η μεταβλητή diabetes. Για την μεταβλητή αυτή ως 0 False εμφανίζονται 56 και 188 τιμές στο πλήθος ενώ ως 0 True εμφανίζονται 40 και 85.

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Ακολουθεί η μεταβλητή high blood pressure. Για την μεταβλητή αυτή το πλήθος των 0 False είναι 57 και 137 αντίστοιχα, ενώ για την τιμή 1 True είναι 39 και 66 αντίστοιχα.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Στη συνέχεια ακολουθεί η μεταβλητή sex. Για την μεταβλητή sex το πληθος της τιμής 0 False είναι 34 και 71 αντίστοιχα, ενώ για την μεταβλητή 1 True το πλήθος των τιμών είναι 62 και 132 αντίστοιχα.

A graph with red and blue squares

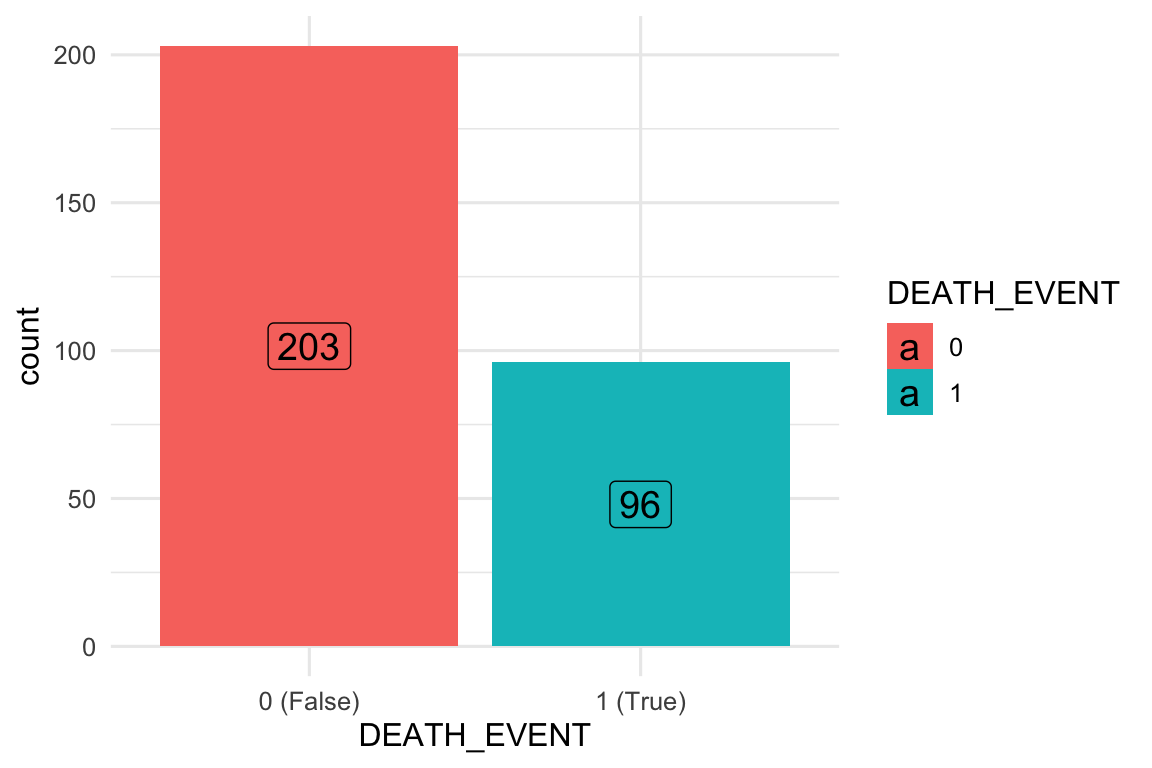
Description automatically generated with low confidence

Στη συνέχεια ακολουθεί η μεταβλητή smoking του συνόλου δεδομένων. Για την μεταβλητή αυτή το πλήθος είναι0 False για τις τιμές 66 και 137, ενώ είναι 1 True για τις μεταβλητές 30 και 66 αντίστοιχα.

A picture containing text, screenshot, diagram, font

Description automatically generated

Στη συνέχεια ακολουθεί η μεταβλητή Death event. Για την μεταβλητή death event το πλήθος είναι 203 0 False και 1 True για τιμές 96 του συνόλου δεδομένων.



Στη συνέχεια παρουσιάζονται τα θηκογράμματα για κάθε μια μεταβλητή όπως έγιναν στην R.

Ακολουθεί το θηκόγραμμα της μεταβλητής age σε σχέση με την μεταβλητή Death event. Παρατηρούμε ότι η διάμεσος είναι στο 60 για την τιμή 0 της μεταβλητής death event ενώ, είναι περίπου 65 για την μεταβλητή death event 1 True.

A picture containing diagram, screenshot, line, rectangle

Description automatically generated

Στη συνέχεια ακολουθεί το θηκόγραμμα της μεταβλητής creatinine\_phosphokinase. Σε αυτή την μεταβλητή παρατηρούμε ότι υπάρχουν πολλές ακραίες τιμές οι οποίες είναι εκτός του θηκογράμματος. Οι τιμές της διαμέσου είναι 200 και 250 αντίστοιχα.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Στη συνέχεια ακολουθεί το θηκόγραμμα της μεταβλητής ejection\_fraction. Σε αυτά τα δύο θηκοκγράμματα που έγιναν ξεχωριστά για τις δύο τιμές της μεταβλητής death\_event 0 και 1, παρατηρούμε ότι υπάρχουν πολλές ακραίες τιμές όπως απεικονίζονται στο θηκόγραμμα. Οι τιμές της διαμέσου είναι 38 και 30 αντίστοιχα.

A picture containing text, diagram, screenshot, line

Description automatically generated

Στη συνέχεια παρουσιάζεται το θηκόγραμμα της μεταβλητής platelets. Συγκριτικά με τα προηγούμενα θηκογράμματα δεν υπάρχουν τόσες πολλές ακραίες τιμές, δηλαδή τιμές εκτός του θηκογράμματος, όμως παραμένουν να υπάρχουν ακραίες τιμές.

A picture containing text, diagram, screenshot, line

Description automatically generated

Στη συνέχεια παρουσιάζεται το θηκόγραμμα της μεταβλητής serum\_creatinine. . Σε αυτά τα δύο θηκογράμματα ανάλογα με την τιμή 0 ή 1 τότε παρουσιάζονται τα αντίστοιχα θηκογράμματα όπου υπάρχουν ακραίες τιμές. Οι τιμές της διαμέσου είναι περίπου 1,5 και 1,8.

A picture containing text, diagram, screenshot, line

Description automatically generated

Στη συνέχεια παρουσιάζονται οι ακραίες τιμές της μεταβλητής serum\_sodium. Σε αυτά τα θηκογράμματα παρουσιάζεται πέρα από τ διάμεσο η οποία είναι περίπου 138 και 137 αντίστοιχα, οι ακραίες τιμές.

A picture containing text, diagram, screenshot, line

Description automatically generated

**Έλεγχος Στατιστικών Υποθέσεων**

«Μελετάτε όλες τις ανά δύο σχέσεις μεταξύ των μεταβλητών»

Η ανά δύο σχέσεις μεταξύ των μεταβλητών μελετήθηκαν στην προηγούμενη ενότητα απεικονίζοντας τα θηκογράμματα. Το μοντέλο που που χρησιμοποιούμε είναι μοντέλο μηχανικής μάθησης και δεν απαιτεί παραδοσιακή περιγραφική στατιστική.

**Προβλεπτικά μοντέλα**

«Χρησιμοποιώντας το μοντέλο της γραμμικής ή λογιστικής παλινδρόμησης βρίσκετε από ποιους παράγοντες επηρεάζεται η μεταβλητή ενδιαφέροντος»

Η μηχανική μάθηση είναι πλέον ένας διαδεδομένος μηχανισμός ο οποίος έχει ως απώτερο σκοπό να προβλέπει. Η μηχανική μάθηση είναι η επιστήμη της πληροφορικής η οποία αλγοριθμικά βελτιώνει αυτόματα με βάση την εμπειρία και τη χρήση των δεδομένων. Ένας ακόμη ορισμός είναι ότι η μηχανική μάθηση είναι η επιστήμη με την οποία ο υπολογιστής δρά χωρίς αποκλειστικά να είναι προγραμματισμένος. Τι σημαίνει όμως αυτό; Αυτό σημαίνει ότι ο υπολογιστής δρα δηλαδή κάνει προβλέπεις ενός φαινομένου βρίσκοντας μοτίβα στα δεδομένα όπου οι υπολογιστές είναι πλέον ικανοί να ανακαλύψουν. Ένα πολύ γνωστό παράδειγμα που το βρίσκουμε στη βιβλιογραφία είναι αυτό που προτείνονται ταινίες και συστάσεις. Χωρίς τη μηχανική μάθηση θα ήταν απαραίτητο να προγραμματίζονται χειροκίνητα οι συστάσεις και να βρίσκονται μονοπάτια και κανόνες για να δημιουργηθούν οι προτάσεις. Σε αυτό το παράδειγμα αν έχουμε ένα παράδειγμα χρήστη είναι εφικτό να γίνει, αν ο χρήστης έχει δει την ταινία α και β τότε είναι πιθανό να δει την ταινία γ. Τι γίνεται στην περίπτωση που έχουν εκατομμύρια χρήστες και δισεκατομμύρια ταινίες; Τότε είναι που εισάγεται η μηχανική μάθηση για να αυτοματοποιήσει και να μάθει το σύστημα να κάνει συστάσεις σε κάθε χρήστη.

Πόσοι τύποι μηχανικής μάθησης υπάρχουν; Υπάρχουν δύο βασικοί τύποι μηχανικής μάθησης. Η μηχανική μάθηση υπό επίβλεψη και η χωρίς. Στα αγγλικά λέγεται supervised and unsupervised machine learning. Στην μηχανική μάθηση με επίβλεψη έχουμε δεδομένα όπου υπάρχουν φωτογραφίες με ετικέτες και αυτό σημαίνει ότι για παράδειγμα ότι θέλω να προβλέψω και να κατηγοριοποιήσω τις φωτογραφίες που μου δόθηκαν με ετικέτες από κάποιον ειδικό. Ο αλγόριθμος της μηχανικής μάθησης θα μου πει την ετικέτα που θα προβλεφθεί από αυτόν τον αλγόριθμο. Ο δεύτερος τύπος μηχανικής μάθησης που υπάρχει είναι η μηχανική μάθηση χωρίς επίβλεψη. Αυτό σημαίνει ότι χρησιμοποιείται όπως λέμε για clustering για να βρει μοτίβα στα δεδομένα σε περίπτωση που έχω κατεργασμένα δεδομένα έτσι ώστε να τα τοποθετήσει σε διαφορετικές ομάδες και να βρει τα μοτίβα. Για παράδειγμα έχω ένα σύνολο δεδομένων με δύο παρατηρήσεις με ζώα όπου μοιάζουν μεταξύ τους και έχω άλλες δύο όπου τα ζώα δεν μοιάζουν μεταξύ τους. Η μηχανή καταλαβαίνει λοιπόν ότι ανήκουν σε διαφορετικές ομάδες και βρίσκει τα κρυμμένα μοτίβα στα δεδομένα.

Οι αλγόριθμοι μηχανικής μάθησης δεν εφευρέθηκαν τυχαία, δηλαδή στηρίζονται στα μαθηματικά και έχουν δυνατές και ισχυρές βάσεις. Είναι ουσιαστικά η βάση από τα μαθηματικά και τη στατιστική. Μία ακόμη κατηγορία της μηχανικής μάθησης είναι deep learning. Το deep learning είναι υποκατηγορία της μηχανικής μάθησης όπου χρησιμοποιεί νευρωνικά δίκτυα όπου μοντελοποιούνται ανθρώπινοι εγκέφαλοι και στόχο έχουν να κάνουν προβλέψεις. Για παράδειγμα αν κάνεις ένα ηχογράφημα τότε για να προβλέψεις τον lv φραγμό στον ήχο τότε μπορείς να χρησιμοποιήσεις νευρωνικά δίκτυα. Επίσης, τα νευρωνικά δίκτυα χρησιμοποιούνται πολύ στην αναγνώριση εικόνας και στα δεδομένα για τον ήχο και την επεξεργασία φυσικής γλώσσας. Γιατί χρειαζόμαστε μηχανική μάθηση; Η μηχανική μάθηση είναι παντού πλέον την καθημερινή μας ζωή. Δηλαδή, ένα απλό παράδειγμα είναι η κατηγοριοποίηση των μηνυμάτων του ηλεκτρονικού ταχυδρομείου σε span και σε όχι spam. Η μηχανική μάθηση εμφανίζεται και στην ιατρική. Χρησιμοποιείται για να κάνει προβλέψεις στην ιατρική για να προβλέπει για παράδειγμα διαγνώσεις μέσα από τις εικόνες και τον ήχο που λαμβάνουν οι αλγόριθμοι μηχανικής μάθησης από τα δεδομένα. Αυτοί οι αλγόριθμοι είναι πλέον ικανοί να προβλέψουν και να αναλύσουν μεγάλα multi-modal δεδομένα για τα οποία είναι απαραίτητο να τρέχουν σε πραγματικό χρόνο. Αυτά είναι τα πλεονεκτήματα της μηχανικής μάθησης και πάμε να εξετάσουμε την μηχανική μάθηση σε heart failure. Γιατί μας χρειάζεται η μηχανική μάθηση σε ένα τέτοιο σύνολο δεδομένων; Πρώτα απ’όλα χρειαζόμαστε την μηχανική μάθηση για να προβλέψουμε υψηλού κινδύνου επιβλαβές της καρδιάς οι οποίες είναι απαραίτητες για να θεραπεύσουν τυχόν heart failures. Μερικά παραδείγματα είναι το καρδιογενές σοκ, μπαλονάκια και μεταμόσχευση καρδιάς. Χρειαζόμαστε τη μηχανική μάθηση γιατί κάθε μία τέτοια βλάβη της καρδιάς δεν απαιτεί την ίδια θεραπεία. Αν εφαρμοστεί για παράδειγμα η ίδια θεραπεία για κάθε ασθενή τότε δεν χρειάζεται να μιλάμε για πρόβλεψη της καρδιακής θεραπείας και να παίρνει το σύστημα αποφάσεις, τότε μιλάμε απλά για την ίδια θεραπεία. Είναι αδύνατο να έχω απλούς αλγορίθμους οι οποίοι λειτουργούν με κανόνες για τέτοιου είδους αποφάσεις. Μερικές ασθένειες της καρδιάς τις σκεφτόμαστε ως ετερογενείς μηχανισμούς και η μηχανική μάθηση πραγματικά μας βοηθά να ανακαλύψουμε αυτούς τους μηχανισμούς και πιθανά να βρούμε νέες θεραπείες για άλλα ζητήματα τα οποία είναι μοναδικά όσον αφορά της επιπλοκές της καρδιάς. Οι αλγόριθμοι μηχανικής μάθησης πλέον έχουν τη δύναμη να εκπαιδευτούν και να κάνουν προβλέψεις. Τέτοιου είδους ζητήματα πλέον είναι ακαδημαικής σημασίας και έχουν κλονίσει τον ακαδημαικό κλάδο. Έτσι μπορούμε να εκπαιδεύσουμε και να κάνουμε προβλέψεις για τυχόν επιπλοκές. Επιπρόσθετα το πρόβλημα μπορεί να γίνει πιο σύνθετο προσθέτοντας κι άλλες μεταβλητές και χαρακτηριστικά. Όταν ακούμε λοιπόν για μηχανική μάθηση εννοούμε δεδομένα σε heart failure για τα οποία αλγόριθμοι μηχανικής μάθησης θα μας κάνουν προβλέψεις και δεν εννούμε πλέον βασικά στατιστικά γραμμικά μοντέλα. Όπως έγινε αντιληπτό από τα παραπάνω υπάρχουν τρία διαφορετικά ζητήματα για να εξετάσουμε. Είναι τα δεδομένα, η πρόβλεψη και η υλοποίηση των αλγορίθμων μηχανικής μάθησης. Πρώτα από όλα έχουμε πρόβλημα με την πρόβλεψη από τους αλγορίθμους μηχανικής μάθησης. Δηλαδή είναι σημαντικό να εξακριβώσουμε ποιους αλγορίθμους μηχανικής μάθησης για χρησιμοποιήσουμε στο σύνολο δεδομένων για να μας κάνει προβλέψεις. Κάθε χαρακτηριστικό ή μεταβλητή πραγματικά χρησιμοποιείται από τη θεωρία παιγνίων και έχει πολύ δυνατό background θεωρητικό από τα μαθηματικά. Μετέπειτα είναι το πρόβλημα των δεδομένων. Σύμφωνα με αυτό, είναι η συγκέντρωση των δεδομένων. Στην ιατρική, αντίθετα με το παράδειγμα των ταινιών και των συστάσεων, είναι πολύ δύσκολο να βρεθούν εκατομμύρια και δισεκατομμύρια παρατηρήσεις για να προβλεφθεί το μοντέλο να γίνουν οι κατάλληλες συστάσεις. Στην ιατρική συνήθως τα σύνολα δεδομένων ακόμη και σήμερα είναι της τάξεως των εκατοντάδων και των χιλιάδων παρατηρήσεων. Τέτοιου είδους αλγόριθμοι όμως απαιτούν μεγάλο αριθμό ασθενών για να γίνουν σωστά οι προβλέψεις. Το τρίτο πρόβλημα όπως προαναφέρθηκε είναι η επίλυση του προβλήματος της υλοποίησης του αλγορίθμου. Σε αυτό το στάδιο υπάρχουν τέσσερα domain όπως μου έρχονται στο μυαλό όπου είναι και τα πιο προκλητικά και ανταγωνιστικά σε τέτοιου είδους ιατρικής φύσεως μηχανικής μάθησης πρόβλημα. Είναι η αποτελεσματικότητα του αλγορίθμου μηχανικής μάθησης. Είναι η βελτίωση του αλγορίθμου μηχανικής μάθησης σε τέτοιου είδους με πραγματικά δεδομένα τα οποία χρειάζονται και εξέταση με το στάδιο testing αφού γίνει το training. Μετέπειτα είναι να λυθεί το πρόβλημα transparency και interpretability όπου φτιάχνουμε το πρόβλημα μηχανικής μάθησης ανοιχτού κώδικα έτσι ώστε ο καθένας να μπορεί να έχει πρόσβαση για να εξετάσει τον κώδικα.

Το μοντέλο που χρησιμοποιήθηκε για το σκοπό αυτής της εργασίας είναι το μοντέλο της logistic\_regression. Έτσι, τα αποτελέσματα που μας έδωσε η R είναι τα εξής όπου το μοντέλο που μας δόθηκε τοποθετήθηκαν όλες οι μεταβλητές.

##

## Call:

## glm(formula = DEATH\_EVENT ~ ., family = binomial, data = train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.3355 -0.4834 -0.1286 0.3220 2.3089

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 2.414e+01 8.579e+00 2.814 0.004901 \*\*

## age 5.887e-02 2.124e-02 2.771 0.005581 \*\*

## anaemia1 4.765e-01 4.608e-01 1.034 0.301096

## creatinine\_phosphokinase 2.542e-04 2.070e-04 1.228 0.219598

## diabetes1 3.236e-01 4.295e-01 0.753 0.451157

## ejection\_fraction -8.585e-02 2.247e-02 -3.820 0.000133 \*\*\*

## high\_blood\_pressure1 1.396e-01 4.569e-01 0.306 0.759896

## platelets -3.864e-06 2.749e-06 -1.406 0.159866

## serum\_creatinine 8.774e-01 3.093e-01 2.837 0.004558 \*\*

## serum\_sodium -1.712e-01 6.042e-02 -2.834 0.004594 \*\*

## sex1 -7.796e-01 5.556e-01 -1.403 0.160627

## smoking 7.599e-01 5.626e-01 1.351 0.176828

## time -2.514e-02 4.261e-03 -5.899 3.65e-09 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 281.32 on 223 degrees of freedom

## Residual deviance: 143.06 on 211 degrees of freedom

## AIC: 169.06

##

## Number of Fisher Scoring iterations: 6

Ενώ ο confusion matrix μας δίνει το εξής:

## Actual

## Predict 0 1

## 0 47 12

## 1 4 12

Δηλαδή παίρνουμε σφάλμα απόδοσης στην τάξη του 21.3%. Αυτό είναι ένα μεγάλο ποσοστό misclassification rate όπως λέμε στην γλώσσα της μηχανικής μάθησης.

Στη συνέχεια για να κάνουμε το μοντέλο μας πιο εύρωστο και σωστό αφαιρέθηκαν κάποιες μεταβλητές έτσι ώστε οι μεταβλητές να είναι πιο εύστοχες στο ζητούμενο μας. Τα αποτελέσματα που πήραμε από την R είναι τα εξής:

## Call:

## glm(formula = DEATH\_EVENT ~ age + ejection\_fraction + serum\_creatinine +

## serum\_sodium + time, family = binomial, data = train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.2260 -0.5155 -0.1359 0.4383 2.4919

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 21.08017 7.73251 2.726 0.00641 \*\*

## age 0.05328 0.01994 2.672 0.00755 \*\*

## ejection\_fraction -0.08383 0.02110 -3.973 7.09e-05 \*\*\*

## serum\_creatinine 0.78924 0.25852 3.053 0.00227 \*\*

## serum\_sodium -0.15099 0.05518 -2.737 0.00621 \*\*

## time -0.02491 0.00401 -6.214 5.18e-10 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 281.32 on 223 degrees of freedom

## Residual deviance: 148.94 on 218 degrees of freedom

## AIC: 160.94

##

## Number of Fisher Scoring iterations: 6

Στη συνέχεια παρουσιάζεται ο confusion matrix ο οποίος μας δίνει τα εξής συμπεράσματα.

## Actual

## Predict 0 1

## 0 47 12

## 1 4 12

Όπου πάλι παίρνουμε σφάλμα απόδοσης misclassification rate της τάξεως του 21.3%. Δηλαδή το μοντέλο μας και που αφαιρέσαμε τις μεταβλητές δεν βελτιώθηκε.

Το ερώτημα που μπορούμε να θίξουμε είναι μπορούμε να κάνουμε το μοντέλο μας πιο εύρωστο και δυνατό; Η απάντηση είναι πώς μπορούμε να το κάνουμε αυτό κανονικοποιώντας τα δεδομένα μας. Έτσι λοιπόν κανονικοποιώντας τα δεδομένα μας παίρνουμε αύξηση στην επίδοση του μοντέλου από 23 % λανθασμένες απαντήσεις στο 16% μείωση. Δηλαδή έχουμε ακρίβεια στην τάξη του 84%. Τα αποτελέσματα που μου δόθηκαν από την R είναι τα εξής:

##

## Call:

## glm(formula = DEATH\_EVENT ~ age + ejection\_fraction + serum\_creatinine +

## serum\_sodium + time, family = binomial, data = train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.0899 -0.5970 -0.2395 0.5317 2.8157

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.2357 0.2221 -5.564 2.64e-08 \*\*\*

## age 0.4411 0.2027 2.176 0.02956 \*

## ejection\_fraction -0.9697 0.2188 -4.432 9.32e-06 \*\*\*

## serum\_creatinine 0.7159 0.2117 3.382 0.00072 \*\*\*

## serum\_sodium -0.1218 0.2047 -0.595 0.55195

## time -1.5884 0.2589 -6.134 8.55e-10 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 284.23 on 223 degrees of freedom

## Residual deviance: 169.45 on 218 degrees of freedom

## AIC: 181.45

##

## Number of Fisher Scoring iterations: 5

Ενώ ο πίνακας confusion είναι ο εξής:

## Actual

## Predict 0 1

## 0 47 6

## 1 6 16

Παρατηρούμε ότι το μοντέλο μας αφού κανονικοποιήθηκαν τα δεδομένα έγινε πιο εύρωστο, σταθερό και δυνατό έτσι ώστε οι καινούριες παρατηρήσεις που θα εισαχθούν να προβλεφθούν σωστά.

Έτσι το μοντέλο μας το τελικό αποτελείται από τις μεταβλητές age, ejection\_fraction, serum\_creatinine, serum\_sodium, time και μας δίνει misclassification rate 16%. Δηλαδή, η ακρίβεια που παίρνουμε είναι 84%.

**ΣΥΖΗΤΗΣΗ**

«Αυτό είναι το σημαντικότερο μέρος της εργασίας. Εδώ παραθέτετε τα συμπεράσματα σας σχετικά με τα εξαγόμενα αποτελέσματα. Μπορείτε εάν θέλετε να τα συγκρίνεται και με αντίστοιχα άρθρα της Βιβλιογραφίας»

Τα συμπεράσματα μας είναι τα εξής. Το μοντέλο που μελετήθηκε είναι το μοντέλο μηχανικής μάθησης logistic regression και η υλοποίηση αυτού πραγματοποιήθηκε με χρήση της R. Παρατηρούμε ότι αρχικά το μοντέλο μηχανικής μας μάθησης με όλες τις μεταβλητές και τις 12 για να προβλέψουμε την death event μας δίνει misclassification rate 21.3%. Δηλαδή, ακρίβεια 79.7%. Μετέπειτα ακολουθώντας τον μηχανισμό της αφαίρεσης των μη απαιτούμενων μεταβλητών το misclassification rate παρέμεινε στο 21.3%. Αυτό σημαίνει ότι δεν πραγματοποιήθηκε κάποια βελτίωση. Μετέπειτα, για να μειώσουμε το misclassification rate και να αυξήσουμε την ακρίβεια του μοντέλου μας κανονικοποιούμε τα δεδομένα μας με τέτοιο τρόπο ώστε πετύχαμε misclassification rate 16%, δηλαδή ακρίβεια 84%. Οπότε το τελικό μοντέλο που επιλέγουμε για το συγκεκριμένο σύνολο δεδομένων μας είναι αυτό με τα κανονικοποιημένα δεδομένα.

**ΠΕΡΙΟΡΙΣΜΟΙ**

(προαιρετικό)

**ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ**

**ΠΑΡΑΡΤΗΜΑ**

Στο Παράρτημα έχουμε την επιλογή να παραθέσουμε

1. Κώδικα από το περιβάλλον της R σε αυτή την μορφή:

**ΕΝΤΟΛΗ 1.** data <- read.csv(file=”path…./mydataset.csv”)

**ΕΝΤΟΛΗ 1 library**(dplyr)

**ΕΝΤΟΛΗ 2** heart = **read.csv**("~/Desktop/UCALGARY/606/project/heart\_failure\_clinical\_records\_dataset.csv")

**ΕΝΤΟΛΗ 3 glimpse**(heart)

**ΕΝΤΟΛΗ 4 library**(ggplot2)

**ΕΝΤΟΛΗ 5** df=heart

**ΕΝΤΟΛΗ 6 ggplot**(df, **aes**(x = anaemia, fill = DEATH\_EVENT)) **+**

**geom\_bar**(stat = "count", position = "stack", show.legend = FALSE) **+**

**scale\_x\_discrete**(labels = **c**("0 (False)", "1 (True)"))**+**

**labs**(x = "Anaemia") **+**

**theme\_minimal**(base\_size = 12) **+**

**geom\_label**(stat = "count", **aes**(label = ..count..), position = **position\_stack**(just = 0.5),

size = 5)

**ΕΝΤΟΛΗ 7 ggplot**(df, **aes**(x = diabetes, fill = DEATH\_EVENT)) **+**

**geom\_bar**(stat = "count", position = "stack", show.legend = FALSE) **+**

**scale\_x\_discrete**(labels = **c**("0 (False)", "1 (True)")) **+**

**labs**(x = "Diabetes") **+**

**theme\_minimal**(base\_size = 12) **+**

**geom\_label**(stat = "count", **aes**(label = ..count..), position = **position\_stack**(vjust = 0.5),

size = 5)

**ΕΝΤΟΛΗ 8 ggplot**(df, **aes**(x = high\_blood\_pressure, fill = DEATH\_EVENT)) **+**

**geom\_bar**(stat = "count", position = "stack", show.legend = FALSE) **+**

**scale\_x\_discrete**(labels = **c**("0 (False)", "1 (True)")) **+**

**labs**(x = "High blood pressure") **+**

**theme\_minimal**(base\_size = 12) **+**

**geom\_label**(stat = "count", **aes**(label = ..count..), position = **position\_stack**(vjust = 0.5),

size = 5)

**ΕΝΤΟΛΗ 9 ggplot**(df, **aes**(x = sex, fill = DEATH\_EVENT)) **+**

**geom\_bar**(stat = "count", position = "stack", show.legend = FALSE) **+**

**scale\_x\_discrete**(labels = **c**("0 (Female)", "1 (Male)")) **+**

**labs**(x = "Sex") **+**

**theme\_minimal**(base\_size = 12) **+**

**geom\_label**(stat = "count", **aes**(label = ..count..), position = **position\_stack**(vjust = 0.5),

size = 5)

**ΕΝΤΟΛΗ 10 ggplot**(df, **aes**(x = smoking, fill = DEATH\_EVENT)) **+**

**geom\_bar**(stat = "count", position = "stack", show.legend = FALSE) **+**

**scale\_x\_discrete**(labels = **c**("0 (False)", "1 (True)")) **+**

**labs**(x = "Smoking") **+**

**theme\_minimal**(base\_size = 12) **+**

**geom\_label**(stat = "count", **aes**(label = ..count..), position = **position\_stack**(vjust = 0.5),

size = 5)

**ΕΝΤΟΛΗ 11 ggplot**(df, **aes**(x = DEATH\_EVENT, fill = DEATH\_EVENT)) **+**

**geom\_bar**(stat = "count", position = "stack", show.legend = TRUE) **+**

**scale\_x\_discrete**(labels = **c**("0 (False)", "1 (True)")) **+**

**labs**(x = "DEATH\_EVENT") **+**

**theme\_minimal**(base\_size = 12) **+**

**geom\_label**(stat = "count", **aes**(label = ..count..), position = **position\_stack**(vjust = 0.5),

size = 5)

**ΕΝΤΟΛΗ 12 ggplot**(df, **aes**(x=age, fill = DEATH\_EVENT)) **+**

**geom\_boxplot**()**+** **coord\_flip**() **+theme\_minimal**(base\_size = 12) **+**

**stat\_boxplot**( position = "dodge2")

**ΕΝΤΟΛΗ 13 ggplot**(df, **aes**(x=creatinine\_phosphokinase, fill = DEATH\_EVENT)) **+**

**geom\_boxplot**()**+** **coord\_flip**()**+** **theme\_minimal**(base\_size = 12) **+**

**stat\_boxplot**(position = "dodge2")

**ΕΝΤΟΛΗ 14 ggplot**(df, **aes**(x=creatinine\_phosphokinase, fill = DEATH\_EVENT)) **+**

**geom\_boxplot**()**+** **coord\_flip**()**+** **theme\_minimal**(base\_size = 12) **+**

**stat\_boxplot**(position = "dodge2")

**ΕΝΤΟΛΗ 15 ggplot**(df, **aes**(x=platelets, fill = DEATH\_EVENT)) **+**

**geom\_boxplot**()**+** **coord\_flip**() **+** **theme\_minimal**(base\_size = 12) **+**

**stat\_boxplot**( position = "dodge2")

**ΕΝΤΟΛΗ 16 ggplot**(df, **aes**(x=serum\_creatinine, fill = DEATH\_EVENT)) **+**

**geom\_boxplot**()**+** **coord\_flip**() **+** **theme\_minimal**(base\_size = 12) **+**

**stat\_boxplot**( position = "dodge2")

**ΕΝΤΟΛΗ 17 ggplot**(df, **aes**(x=serum\_sodium, fill = DEATH\_EVENT)) **+**

**geom\_boxplot**()**+** **coord\_flip**() **+theme\_minimal**(base\_size = 12) **+**

**stat\_boxplot**( position = "dodge2")

**ΕΝΤΟΛΗ 18** data = heart

**ΕΝΤΟΛΗ 19 library**(sampling)

**ΕΝΤΟΛΗ 20 set.seed**(5)

**ΕΝΤΟΛΗ 21** idx=sampling**:::strata**(heart, stratanames = **c**('DEATH\_EVENT'), size = **c**(3**/**4**\***96,3**/**4**\***203), method = 'srswor')

**ΕΝΤΟΛΗ 22** train=heart[idx**$**ID\_unit,]

**ΕΝΤΟΛΗ 23** test=heart[**-**idx**$**ID\_unit,]

**ΕΝΤΟΛΗ 24** lr<-**glm**(DEATH\_EVENT**~**., family=binomial, data=train)

**ΕΝΤΟΛΗ 25 summary**(lr)

**ΕΝΤΟΛΗ 26** preds<-**predict**(lr, test, type="response")

**ΕΝΤΟΛΗ 27** Predict<-**rep**(0,**dim**(test)[1])

**ΕΝΤΟΛΗ 28** Predict[preds**>=**0.5]=1

**ΕΝΤΟΛΗ 29** Actual<-test**$**DEATH\_EVENT

**ΕΝΤΟΛΗ 30 table**(Predict, Actual)

**ΕΝΤΟΛΗ 31** lr<-**glm**(DEATH\_EVENT**~**age**+**ejection\_fraction**+**serum\_creatinine**+**serum\_sodium**+**time, family=binomial, data=train)

**ΕΝΤΟΛΗ 32 summary**(lr)

**ΕΝΤΟΛΗ 33** preds<-**predict**(lr, test, type="response")

**ΕΝΤΟΛΗ 34** Predict<-**rep**(0,**dim**(test)[1])

**ΕΝΤΟΛΗ 35** Predict[preds**>=**0.5]=1

**ΕΝΤΟΛΗ 36** Actual<-test**$**DEATH\_EVENT

**ΕΝΤΟΛΗ 37 table**(Predict, Actual)

**ΕΝΤΟΛΗ 38** features = **subset**(data,select=**c**(age, ejection\_fraction, serum\_creatinine, serum\_sodium, time))

**ΕΝΤΟΛΗ 39** scaled<-**scale**(features)

**ΕΝΤΟΛΗ 40** DEATH\_EVENT = data**$**DEATH\_EVENT

**ΕΝΤΟΛΗ 41** new\_df = **data.frame**(scaled, DEATH\_EVENT)

**ΕΝΤΟΛΗ 42** idx=**sample**(1**:nrow**(new\_df),3**/**4**\*nrow**(new\_df))

**ΕΝΤΟΛΗ 43** train=new\_df[idx,]

**ΕΝΤΟΛΗ 44** test=new\_df[**-**idx,]

**ΕΝΤΟΛΗ 45** new\_lr<-**glm**(DEATH\_EVENT**~**age**+**ejection\_fraction**+**serum\_creatinine**+**serum\_sodium**+**time, family=binomial, data=train)

**ΕΝΤΟΛΗ 46 summary**(new\_lr)

**ΕΝΤΟΛΗ 47** preds<-**predict**(new\_lr, test, type="response")

**ΕΝΤΟΛΗ 48** Predict<-**rep**(0,**dim**(test)[1])

**ΕΝΤΟΛΗ 49** Predict[preds**>=**0.5]=1

**ΕΝΤΟΛΗ 50** Actual<-test**$**DEATH\_EVENT

**ΕΝΤΟΛΗ 51 table**(Predict, Actual)

**ΕΝΤΟΛΗ 2.** …….

2. Πίνακες

**Πίνακας 1.** … Τίτλος …

**Πίνακας 2.** … Τίτλος …

3. Διαγράμματα

**Διάγραμμα 1.** …. Τίτλος …

**Διάγραμμα 2.** …. Τίτλος …

4. Ό,τι άλλο πιστεύετε ότι θα στηρίξει το Κύριο Μέρος της εργασίας σας

Εδώ να σημειώσετε το κάθε μέλος της ομάδας τι συνεισφορά είχε στην εργασία. Περιγράψτε ο καθένας ξεχωριστά τον ρόλο που είχε στην ομάδα.

Και τα δύο μέλη της ομάδας έχουν συνεισφέρει επαρκώς και μεθοδικά καθώς η εργασία πραγματοποιήθηκε κάτω από το ομαδικό πνεύμα με πολύ μεράκι και χαρά.

**Οδηγίες**

1. Το μέγεθος της εργασίας να μην είναι μικρότερο των **20** σελίδων αλλά ούτε και μεγαλύτερο των **50** σελίδων (μαζί με το Παράρτημα) γραμματοσειράς **Calibri (14)** και **1.5** διάστιχο με πλήρης στοίχιση.

2. Στο κύριο μέρος της εργασίας **ΔΕΝ** παραθέτουμε κώδικα της R ή copy – paste πίνακες και αποτελέσματα από το περιβάλλον της R, πέρα μόνον από τα Διαγράμματα και τους Πίνακες που θεωρείτε σημαντικότερα για να στηρίξουν αυτά που γράφετε. Τα Διαγράμματα και οι Πίνακες πρέπει να έχουν την κατάλληλη μορφοποίηση και όχι απλό copy – paste από το περιβάλλον της R.

**Δείτε αντίστοιχες εργασίες εδώ:** https://pubmed.ncbi.nlm.nih.gov/

**Μερικά παραδείγματα:**

https://pubmed.ncbi.nlm.nih.gov/29204219/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7353331/

<https://www.cambridge.org/core/journals/british-journal-of-nutrition/article/comparison-of-statistical-and-machinelearning-techniques-in-evaluating-the-association-between-dietary-patterns-and-10year-cardiometabolic-risk-20022012-the-attica-study/3C5B2ABA276B6441643162DF52051972>

3. Στο Παράρτημα βάζουμε Διαγράμματα και Πίνακες τα οποία θεωρούνται δευτερεύοντα αλλά κρίνετε ότι πρέπει να υπάρχουν στην εργασία σας ώστε να στηρίξουν τη δουλειά σας. Δεν είστε υποχρεωμένοι να τα μορφοποιήσετε αλλά αρκεί ένα απλό copy – paste από το περιβάλλον της R. Παρόλα αυτά θα εκτιμηθεί ιδιαίτερα η προσπάθεια βελτίωσης τους.

Στην εργασία θα χρησιμοποιηθεί η βάση δεδομένων: **Heart failure clinical records** η οποία είναι διαθέσιμη εδώ: *Heart failure clinical records. (2020). UCI Machine Learning Repository*.

Η συγκεκριμένη βάση δεδομένων περιλαμβάνει 13 κλινικά χαρακτηριστικά από 299 ασθενείς.

**Οι μεταβλητές είναι:**

1. Ηλικία (age)

2. Αναιμία (anaemia)

3. Υπέρταση (high blood pressure)

4. Φωσφοκινάση κρεατίνης (creatinine phosphokinase)

5. Διαβήτης (diabetes)

6. Κλάσμα εξώθησης (ejection fraction)

7. Αιμοπετάλια (platelets)

8. Φύλο (sex)

9. Κρεατινίνη ορού (serum creatinine)

10. Επίπεδα νατρίου στο αίμα (serum sodium)

11. Καπνιστικές συνήθειες (smoking)

12. Χρόνος παρακολούθησης (time)

13. Αποτέλεσμα (Death event)