ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΤΕΧΝΟΛΟΓΙΕΣ ΠΟΛΥΜΕΣΩΝ και ΓΡΑΦΙΚΑ
ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

1η ΕΡΓΑΣΙΑ – CLASSIFICATION PROBLEMS

Εργασία

του

Εμμανουηλίδη Αθανάσιου
ΑΜ: ics21190

Θεσσαλονίκη, 12 Δεκεμβρίου 2023

ΠΡΟΒΛΕΨΗ ΕΓΚΕΦΑΛΙΚΟΥ

Αθανάσιος Εμμανουηλίδης

Απαλλακτική Εργασία

υποβαλλόμενη για την εκπλήρωση των απαιτήσεων του

Μαθήματος: ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ
του Τμήματος ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Επιβλέπων Καθηγητής
Ευτύχιος Πρωτοπαπαδάκης

Περίληψη

Στην παρούσα μελέτη γίνεται η ανάλυση δεδομένων με στόχο την πρόβλεψη εγκεφαλικών σε ασθενείς. Στα δεδομένα υπάρχουν δείκτες απόδοσης και χαρακτηριστικών των ασθενών, καθώς και η τελική τους κατάσταση (αν υπέστη εγκεφαλικό ένας άνθρωπος ή όχι). Γίνεται εμβάθυνση στην αναγνώριση και αξιολόγηση της αποδοτικότητας διάφορων αλγορίθμων ταξινόμησης, με στόχο την κατανόηση τεχνικών που προσφέρουν την καλύτερη επίδοση. Μέσω την ανάλυσης των αποδόσεων τον αλγορίθμων γίνεται η επαλήθευση της εγκυρότητας τους με στόχο την αποτύπωση πραγματικών προβλέψεων.

Η ικανότητα πρόβλεψης ενός εγκεφαλικού συμβάλει ουσιαστικά σε αποφάσεις για την ζωή και τις επιλογές του ατόμου. Τα αποτελέσματα βασίζονται σε δείκτες και ως αποτέλεσμα ένας χρήστης αλλάζοντας κάποιες συνήθειες μπορεί να μειώσει την πιθανότητα να πάθει εγκεφαλικό. Στην αναφορά γίνεται σύγκριση των αλγορίθμων ώστε να γίνει εκτίμηση των αποδόσεων τους.

Επιπλέον γίνεται εφαρμογή και αξιολόγηση συγκεκριμένων μοντέλων με στόχο την σωστή ερμηνεία τους στο συγκεκριμένο πρόβλημα, υπογραμμίζοντας τη σημασία της επιλογής του καταλληλότερου. Με αυτόν τον τρόπο, η μελέτη αυτή συμβάλλει στην περαιτέρω κατανόηση της ουσίας της επιστήμης δεδομένων στον τομέα της ιατρικής. Είναι απαραίτητο να σημειωθεί ότι στον πραγματικό κόσμο υπάρχουν πολύ παράγοντες που συμβάλουν στην αύξηση της πιθανότητας ενός εγκεφαλικού, για αυτό τα αποτελέσματα που θα εμφανιστούν θα βασιστούν μόνο στα συγκεκριμένα δεδομένα που παρέχει το data set.

Λέξεις Κλειδιά: εγκεφαλικά, πρόβλεψη, ανάλυση δεδομένων (Exploratory Data Analysis – EDA), δείκτες απόδοσης, μέθοδοι/αλγόριθμοι ταξινόμησης, αξιολόγηση μοντέλων, πρόληψη εγκεφαλικού, προ επεξεργασία δεδομένων.

Abstract

In the present study, data analysis is performed with the aim of predicting strokes in patients. In the data are indicators of patient performance and characteristics, as well as their final status (whether a person had a stroke or not). It deepens the identification and evaluation of the efficiency of various classification algorithms, with the aim of understanding the techniques that offer the best performance. Through the analysis of the performance of the algorithms, the validity of the algorithms is verified with the aim of capturing real predictions.

The ability to predict a stroke significantly contributes to decisions about the person's life and choices. The results are based on indicators and as a result a user by changing some habits can reduce the chance of having a stroke. The report compares the algorithms in order to estimate their performance.

In addition, specific models are applied and evaluated with the aim of their correct interpretation in the specific problem, underlining the importance of choosing the most appropriate one. In this way, this study contributes to the further understanding of the importance of data science in the field of medicine. It is necessary to note that in the real world there are many factors that contribute to increasing the probability of a stroke, therefore the results that will be displayed will only be based on the specific data provided by the data set.

Keywords: stroke, prediction, Exploratory Data Analysis – EDA, performance indicators, classification methods, model evaluation, stroke prevention, data preprocessing.

Περιεχόμενα

Κατάλογος Εικόνων (αν υπάρχουν)

Κατάλογος Πινάκων (αν υπάρχουν)

1 Εισαγωγή

1.1 Θεωρητικό υπόβαθρο

Λόγω των εγκεφαλικών επεισοδίων τα ποσοστά θανάτων αυξάνονται κάθε χρόνο με μεγάλο ρυθμό. Λύσεις σε αυτό το πρόβλημα μπορεί να δώσει η μηχανική μάθηση η οποία προσφέροντας τεχνικές, μπορεί να βελτιώσει την ακρίβεια της πρόγνωσης και των μεθόδων για την αναγνώριση εγκεφαλικών επεισοδίων (Schwartz et al., *Stroke mortality prediction using machine learning: Systematic review* 2023). Αυτό θα διαδραματίσει καθοριστικό ρόλο, ώστε να ληφθούν κατάλληλα μέτρα για τη θεραπεία και την αποκατάσταση των ασθενών που μπορεί να διαγνωστούν από το μοντέλο μηχανικής μάθησης ότι έχουν μεγάλη πιθανότητα να πάθουν εγκεφαλικό βραχυπρόθεσμα. Αρχικά, οι αλγόριθμοι της μηχανικής μάθησης για την πρόβλεψη ενός εγκεφαλικού επεισοδίου βασίζονται στην επεξεργασία ιατρικών δεδομένων και γενικώς στην ανάλυση δεδομένων από ηλεκτρονικά αρχεία υγείας (EHR). Η λειτουργία ενός μοντέλου βασίζεται στην παραγωγή μοτίβων και συσχετίσεων τα οποία μπορεί να μη φανερωθούν με παραδοσιακές μεθόδους που χρησιμοποιούνται μέχρι τώρα σε κλινικές. Σε αρκετά μοντέλα έχουν εφαρμοστεί τέτοιες μέθοδοι ώστε να επιτευχθεί η πρόληψη του εγκεφαλικού επεισοδίου, αυτά τα μοντέλα χρησιμοποιούν ένα ευρύ φάσμα τεχνικών συμπεριλαμβανομένων των μεθόδων ταξινόμησης logistic regression, random forests, και neural networks στοχεύοντας στην ακρίβεια σε αυτές τις προβλέψεις. Μερικές από τις βασικές μεταβλητές που προσδιορίστηκαν περιλαμβάνουν την ηλικία και τον δείκτη μάζας σώματος (BMI) (οι επιλογές αυτές έγιναν με βάση το Εθνικό Ινστιτούτο Υγείας (NIHSS)). Αυτοί οι παράγοντες ευθυγραμμίζονται με την κατανόηση μιας παραδοσιακής κλινικής, δίνοντας έμφαση στη μηχανική μάθηση η οποία βοηθάει στο να αυξήσει τα υπάρχοντα προγνωστικά μοντέλα .

Σε ένα πρόβλημα υπάρχουν πάντα πολλές προκλήσεις, οι οποίες δυσκολεύουν την τελική αποτύπωση του. Η υψηλή μεταβλητότητα στις

μετρήσεις απόδοσης του μοντέλου και η μεροληψία στο σχεδιασμό και τη μελέτη είναι κάποιες από αυτές. Σε μια σωστή εφαρμογή μοντέλων μηχανικής μάθησης σε κλινικά περιβάλλοντα πρέπει να μην υπάρχει χάσμα μεταξύ της εφαρμογής των μοντέλων στον πραγματικό κόσμο και στον δεδομένων που έχουν φτιαχτεί για την ερευνητική πρόοδο.

Ως συμπέρασμα είναι φανερό ότι τα μοντέλα της μηχανικής μάθησης είναι υποσχόμενα για την ανάπτυξη των κλινικών καθώς υπάρχει μεγάλη βελτίωση των προβλέψεων θνησιμότητας από εγκεφαλικά επεισόδια. Παρ όλα αυτά είναι απαραίτητο να σημειωθεί ότι στόχος είναι να υπάρχει μεροληψία και τα μοντέλα αυτά να λειτουργούν σε μεγάλης κλίμακας μελέτες ώστε να επικυρώνεται η αποτελεσματικότητά τους. Η εργασία αυτή επιχειρεί να αξιολογήσει διαφορετικά μοντέλα μηχανικής μάθησης για την πρόβλεψη εγκεφαλικού στους ανθρώπους. Για να γίνει το θεωρητικό υπόβαθρο αρχικά θα αναφερθούν οι μέθοδοι που εφαρμόστηκαν, καθώς και μία περιγραφή, εξηγώντας τις παραμέτρους που επηρεάζουν την απόδοση των τιμών.

Αρχικά η μέθοδοι ταξινόμησης που εφαρμόστηκαν στον κώδικα επίλυσης του προβλήματος είναι οι:

- AdaBoost
- K-Nearest Neighbors (KNN)
- Naive Bayes
- Linear Discriminant Analysis (LDA)
- MPL Classifier
- Support Vector Machine (SVM)
- Decision Tree Classifier
- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier
- Gaussian Naive Bayes
- Neural Network

Πέραν των μεθόδων ταξινόμησης, στον κώδικα περιέχονται και οι παρακάτω μέθοδοι οι οποίοι χρησιμοποιήθηκαν για να γίνει η υλοποίηση της λύσης του προβλήματος. Οι μέθοδοι συνοδεύονται από μια περιγραφή της χρησιμότητας και λειτουργικότητας τους:

- Ανάγνωση Δεδομένων από Excel: Χρήση της βιβλιοθήκης pandas για την ανάγνωση δεδομένων από ένα αρχείο Excel.
- Οπτικοποίηση Δεδομένων: Εφαρμογή βιβλιοθηκών όπως matplotlib και plotly για τη δημιουργία διαγραμμάτων και γραφημάτων που απεικονίζουν τις διαφορετικές πτυχές των δεδομένων, όπως τον αριθμό των ατόμων που υπέστησαν εγκεφαλικό ή όχι.
- Εκπαίδευση Μοντέλων και Αξιολόγηση Επιδόσεων: Χρήση διαφόρων μετρικών όπως ακρίβεια, ανάκληση, F1 score και ROC-AUC για την αξιολόγηση των μοντέλων ταξινόμησης

1.2 Σκοπός – Στόχοι

Ο θάνατος με αιτία το εγκεφαλικό είναι πλέον στις μέρες μας πολύ συχνός. Ένα μεγάλο μέρος ανθρώπων, είτε μπορεί να βρούνε το θάνατο, είτε να έχουν κάποιου είδους αναπηρία. Η ανίχνευση του εγκεφαλικού είναι πλέον κρίσιμη και απαραίτητη. Στόχος αυτής της εργασίας είναι να βρει ένα μοντέλο το οποίο να προσφέρει κατάλληλα αποτελέσματα για την πρόβλεψη της πιθανότητας εγκεφαλικού επεισοδίου. Μια μέθοδος η οποία θα βοηθήσει είναι μέσω της μηχανικής μάθησης, η οποία θα εκπαιδεύει ένα μοντέλο και με βάση κάποια πρότυπα θα γίνεται μια πρόβλεψη για τα άτομα που μπορεί να βρίσκονται πολύ κοντά στο να πάθουν εγκεφαλικό στο μέλλον. Οι παραδοσιακοί τρόποι δεν είχαν τόσο καλά αποτελέσματα, για αυτό το λόγο η μηχανική μάθηση θα προσπαθήσει να δώσει μια λύση.

1.3 Διάρθρωση της μελέτης

Η εργασία αυτή δομείται σε 7 κεφάλαια τα οποία θα αναλυθούν παρακάτω. Το κάθε κεφάλαιο είναι ξεχωριστό και χρησιμεύει ώστε να μπορεί να γίνει μια πλήρης ανάλυση και σωστή διατύπωση του τρόπου αντιμετώπισης καθώς και η διεξαγωγή των αποτελεσμάτων της εργασίας.

Κεφάλαιο 2: Στο κεφάλαιο αυτό θα γίνει εισαγωγή στην επιλογή του dataset δηλαδή γενικές πληροφορίες της προέλευσης των χαρακτηριστικών και το είδους των δεδομένων που χρησιμοποιήθηκαν.

Κεφάλαιο 3: Πολύ σημαντικό είναι να γίνει μια προ επεξεργασία των δεδομένων. Τα βήματα που ακολουθήθηκαν ώστε να καταφέρει να γίνει αυτή η προ επεξεργασία αναφέρονται σε αυτό το κεφάλαιο. Διαδικασίες όπως ο καθαρισμός των δεδομένων, η επεξεργασία των τιμών που λείπουν, η κωδικοποίηση κάποιον τιμών για να μπορέσει να λειτουργεί σωστά ο κώδικας είναι κάποιες από τις μεθόδους που θα αναλυθούν.

Κεφάλαιο 4: Επίσης πολύ σημαντικό σημείο στην ανάλυση των δεδομένων είναι η οπτικοποίηση τους. Στο κεφάλαιο αυτό ο αναγνώστης θα έχει τη δυνατότητα να δει οπτικοποιημένα κάποια αποτελέσματα τα οποία βοήθησαν πριν την τελική φάση της μοντελοποίησης, καθώς προσφέρουν και γενικά στοιχεία που υπήρξαν εμπόδιο κατά την υλοποίηση αυτού του Project.

Κεφάλαιο 5: Σε αυτό το κεφάλαιο θα αναφερθούν οι classifiers που χρησιμοποιήθηκαν. Θα αναφερθεί επίσης ο τρόπος ώστε να γίνει σωστά η προσέγγιση για την πρόβλεψη του εγκεφαλικού καθώς και ποια κριτήρια και ποια μοντέλα αποτέλεσαν σημαντικό παράγοντα στην διεξαγωγή των αποτελεσμάτων.

Κεφάλαιο 6: Σε αυτό το κεφάλαιο θα γίνει η εμφάνιση των αποτελεσμάτων καθώς και θα φανεί η απόδοση τον μοντέλων στο dataset που επιλέχθηκε. Θα αναφερθούν ποιες από τις τιμές που χρησιμοποιήθηκαν βοήθησαν στην αναγνώριση της ποιότητας του κάθε μοντέλου.

Κεφάλαιο 7: Τέλος, υπάρχουν τα συμπεράσματα στα οποία θα γίνει μια σύντομη αναφορά με βάση τα προηγούμενα κεφάλαια καθώς και η επιλογή του σωστού μοντέλου το οποίο τελικά θα φανεί κατάλληλο στην πρόβλεψη του εγκεφαλικού.

2 Περιγραφή ανάλυση Dataset

2.1 Πηγή

Το dataset επιλέχθηκε μέσω του Kaggle μιας σελίδας που παρέχει ένα εύρος συνόλου δεδομένων. Δημιουργήθηκε από τον fedesoriano και χρησιμοποιείται κυρίως για έργα που αφορούν τη μηχανική μάθηση και στόχος είναι οι πρόβλεψη των εγκεφαλικών επεισοδίων. Κατά την έρευνα αυτού του dataset δεν βρέθηκε αν αυτά τα δεδομένα είναι φτιαχτά από τον fedesoriano ή παρέχονται μέσα από κάποια βάση δεδομένων κάποιου οργανισμού.

2.2 Χαρακτηριστικά

Τα δεδομένα αποτελούνται από 12 στήλες οι οποίες παρέχουν κλινικά χαρακτηριστικά. Όλες οι στήλες εκτός του id του ασθενούς θα μπορούσαν να είναι ενδεικτικά για ένα μελλοντικό εγκεφαλικό. Κατά κύριο λόγο είναι αριθμητικά δεδομένα παρόλα αυτά υπάρχουν και κάποια αλφαριθμητικά δεδομένα τα οποία στην πορεία υπέστησαν επεξεργασία. Πιο αναλυτικά τα χαρακτηριστικά αυτά περιλαμβάνουν:

1. id: μοναδικό αναγνωριστικό
2. gender: "Άνδρας", "Γυναίκα" ή "Άλλος"
3. age: ηλικία του ασθενούς
4. hypertension: 0 εάν ο ασθενής δεν έχει υπέρταση, 1 εάν ο ασθενής έχει υπέρταση
5. heart_disease: 0 εάν ο ασθενής δεν έχει καμία καρδιακή νόσο, 1 εάν ο ασθενής έχει καρδιακή νόσο
6. ever_married: "Όχι" ή "Ναι"
7. work_type: "παιδιά", "Ιδιωτικός Υπάλληλος", "Ποτέ δεν δούλεψε", "Ιδιώτης" ή "Αυτοαπασχολούμενος"
8. Residence_type: "Αγροτικό" ή "Αστικό"
9. avg_glucose_level: μέσο επίπεδο γλυκόζης στο αίμα
10. bmi: δείκτης μάζας σώματος
11. smoking_status: "πρώην καπνιστής", "ποτέ δεν κάπνιζε", "καπνίζει" ή "Άγνωστο"*

12. stroke: 1 εάν ο ασθενής έπαθε εγκεφαλικό ή 0 εάν όχι

**Σημείωση: "Άγνωστο" στο smoking_status σημαίνει ότι οι πληροφορίες δεν είναι διαθέσιμες για αυτόν τον ασθενή*

Τα χαρακτηριστικά αυτά έχουνε σχέση με πιθανή αιτία εγκεφαλικού για αυτόν τον λόγο και επιλέχτηκαν.

2.3 Μεταβλητή στόχος

Η μεταβλητή η οποία επιλέχθηκε μεταξύ των άλλων είναι η stroke. Ο λόγος της επιλογής αυτής της μεταβλητής είναι προφανής, καθώς δηλώνει αν ένας ασθενής έχει υποστεί εγκεφαλικό επεισόδιο ή όχι. Είναι απαραίτητο να θέσουμε μια μεταβλητή ως το κεντρικό στόχο του προβλήματός μας, είναι κρίσιμη η κατανόηση της επειδή επηρεάζει την επιλογή των μοντέλων τα οποία πρέπει να εκπαιδευτούν. Επίσης θα πρέπει να γίνουν περισσότερες αναλύσεις με βάση αυτόν τον στόχο καθώς και ειδική μελέτη μέσω γραφημάτων για παρατηρήσεις όπως το γεγονός αν το σύνολο δεδομένων είναι ισορροπημένο.

2.4 Αρχικές Παρατηρήσεις

Για καλύτερη κατανόηση οι αρχικές παρατηρήσεις θα χωριστούνε σε κάποιες μορφές. Ποιο συγκεκριμένα έχουμε:

Ανισορροπία κλάσης: Κατά την ανάλυση του συνόλου των δεδομένων, παρατηρήθηκε έντονη διαφορά μεταξύ στα άτομα που έχουν πάθει εγκεφαλικό και στα μη. Πιο συγκεκριμένα 249 άτομα έχουν υποστεί εγκεφαλικό ενώ 4861 όχι. Αυτή η διαφορά αποτελεί ένα ποσοστό 4.9% των ατόμων που έπαθαν εγκεφαλικό και 95.1% αυτών που δεν έπαθαν. Αυτή η ανισορροπία απαιτεί κάποιες ενέργειες, καθώς θα είναι αδύνατη η σωστή αξιολόγηση του κάθε μοντέλου.

Έλλειψη τιμών: Κατά την επεξεργασία το δεδομένο βρέθηκε ότι υπήρχαν τιμές Nan. Η έλλειψη των τιμών μπορεί να επηρεάσει σημαντικά τα μοντέλα καθώς μερικά δεν μπορούν να τρέξουν έχοντας να επεξεργαστούνε δεδομένα που δεν υπάρχουν. Στο παρόν project η αντιμετώπιση αυτού του προβλήματος γίνεται με πολλούς και διαφορετικούς τρόπους οι οποίοι θα αναλυθούν παρακάτω.

Κατανομή χαρακτηριστικών: Η ηλικία και το μέσο επίπεδο γλυκόζης είναι στήλες οι οποίες χρειάστηκαν περισσότερη έρευνα για τον εντοπισμό μοτίβων η προβλημάτων τα οποία θα μπορούσαν να επηρεάσουν τα μοντέλα.

3 Προ επεξεργασία Δεδομένων

3.1 Καθαρισμός(cleaning)

Ο συνολικός αριθμός των τιμών που έλειπαν ήταν 201 οι οποίες βρισκόντουσαν όλες στην στήλη bmi. Ένα από τα πιο κρίσιμα τμήματα είναι να γίνει σωστά ο καθαρισμός των δεδομένων. Αυτή η διαδικασία γίνεται με την χρήση 4 διαφορετικών στρατηγικών. Ποιο συγκεκριμένα έγινε χρήση:

Dropping Missing Values: Μέσο της εντολής df.dropna() γίνεται η αφαίρεση των ελλιπών τιμών. Δηλαδή μέσω αυτής της εντολής αφαιρείται ολόκληρη η γραμμή διασφαλίζοντας ότι τα μοντέλα θα εκπαιδευτούν χωρίς να λάβουν υπόψη τα συγκεκριμένα δεδομένα που αφαιρέθηκαν. Βέβαια είναι σημαντικό να αναγνωριστεί ότι η απώλεια κάποιων πληροφοριών μπορεί να έχει σημαντικό ρόλο στο τελικό αποτέλεσμα.

Linear Regression Imputation: Η μέθοδος αυτή είναι η εκπαίδευση ενός μοντέλου στο τμήμα του συνόλου bmi χρησιμοποιώντας τα υπόλοιπα χαρακτηριστικά που βρίσκονται στη βάση γίνεται πρόβλεψη της τιμής που λείπει. Μέσω αυτής της μεθόδου διατηρείται η ακεραιότητα των δεδομένων και δεν υπάρχει απώλεια πληροφοριών.

Backfilling: Αυτή η στρατηγική είναι πολύ απλή γιατί ουσιαστικά η κενή τιμή συμπληρώνεται με βάση την επόμενη έγκυρη παρατήρηση. Αυτός ο τρόπος είναι γρήγορος και αρκετά αποτελεσματικός παρέχοντας μια λογική εκτίμηση, αποφεύγοντας την διαγραφή των κοινών τιμών.

Iterative Imputer: αποτελεί μια προηγμένη τεχνική διαχείρισης των τιμών που λείπουνε. Πιο συγκεκριμένα γίνεται μοντελοποίηση κάθε χαρακτηριστικού με τις τιμές που λείπουν ως συνάρτηση άλλων χαρακτηριστικών με κυκλικό τρόπο. Αυτή μέθοδος είναι αποτελεσματική και υποκειμενική καθώς αξιοποιεί όλο το σύνολο των δεδομένων

και προσπαθεί να κάνει μια εκτίμηση των τιμών που λείπουνε με μεγαλύτερη ακρίβεια από άλλες απλούστερες μεθόδους

3.2 Κωδικοποίηση Μεταβλητών (Encoding Categorical Variables)

Αρχικά έγινε ανάλυση των μεταβλητών. Παρατηρήθηκε ότι τα δεδομένα από 5 κατηγορίες είναι αλφαριθμητικά. Οι στήλες αυτές ήταν: το φύλο, αν ο ασθενής είναι παντρεμένος, ο τόπος διαμονής και η κατάσταση καπνίσματος. Αυτό οδηγεί σε μετατροπή αυτών τον κελιών σε αριθμητικές τιμές. Για να γίνει αυτή η μετατροπή έγινε χρήση του LabelEncoder από το Scikit-learn. Οι μετατροπή αυτή είναι απαραίτητη, λόγο του ότι οι αλγόριθμοι μηχανικής μάθησης δεν μπορούν να λειτουργήσουν με αλφαριθμητικές τιμές και απαιτούν αριθμητική είσοδο για την καλύτερη λειτουργία τους.

3.3 Ανισορροπία κλάσης

Λόγο της μεγάλης ανισορροπίας της μεταβλητής stroke τα αποτελέσματα των μεθόδων ήταν δεδομένο ότι δεν θα ήταν σωστά. Η αντιμετώπιση του προβλήματος ανισορροπίας κλάσης έγινε με την χρήση της SMOTE. Με αυτό τον τρόπο διασφαλίζεται ότι το μοντέλο είναι ισορροπημένο και η προγνωστική απόδοση πολύ καλύτερη.

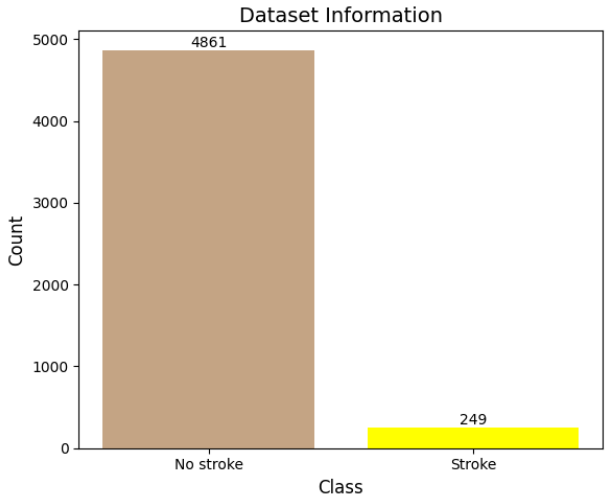
4 Ανάλυση Δεδομένων

Το επόμενο βήμα εφόσον τελείωσε η προ επεξεργασία είναι η υλοποίηση κανόνων και μεθόδων για την εύρεση λύσης στο πρόβλημα. Για την καλύτερη κατανόηση του προβλήματος έγιναν οπτικοποιήσεις οι οποίες βοηθάνε σε γρηγορότερα συμπεράσματα, ενώ προσφέρουν μια καλύτερη εικόνα για το σύνολο του προβλήματος και των αποτελεσμάτων.

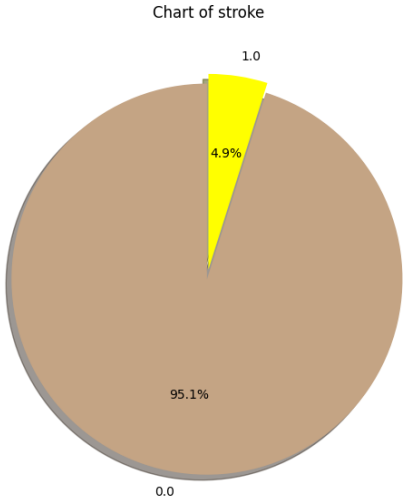
4.1 Οπτικοποιήσεις (visualizations)

Οι οπτικοποιήσεις θα αφορούν κλάδους και προβλήματα που αναφέρθηκαν σε προηγούμενα κεφάλαια. Όσον αφορά το πρόβλημα ανισορροπίας κλάσης παρατηρείτε από το διάγραμμα πίτας (εικόνα 1) ότι απεικονίζεται ένα μεγάλο ποσοστό ατόμων που δεν έχουν πάθει εγκεφαλικό σε σύγκριση με το ελάχιστο πλήθος ατόμων που έχουν υποστεί. Αν και τα ποσοστά φανερώνουν την ανάγκη για την εξισορρόπηση του datast είναι

απαραίτητη η υπογράμμιση του αριθμού των ατόμων που έπαθαν και όχι εγκεφαλικό. Το γράφημα ράβδων (εικόνα 2) δείχνει ότι 4861 ασθενείς ανήκουν στην κατηγορία “όχι εγκεφαλικό” ενώ 249 στην κατηγορία εγκεφαλικό. Αυτό μας οδηγεί πλέον στην σίγουρη ανάγκη για την χρήση της SMOTE που αναφέρθηκε νωρίτερα για την ισορροπία της κλάσης.



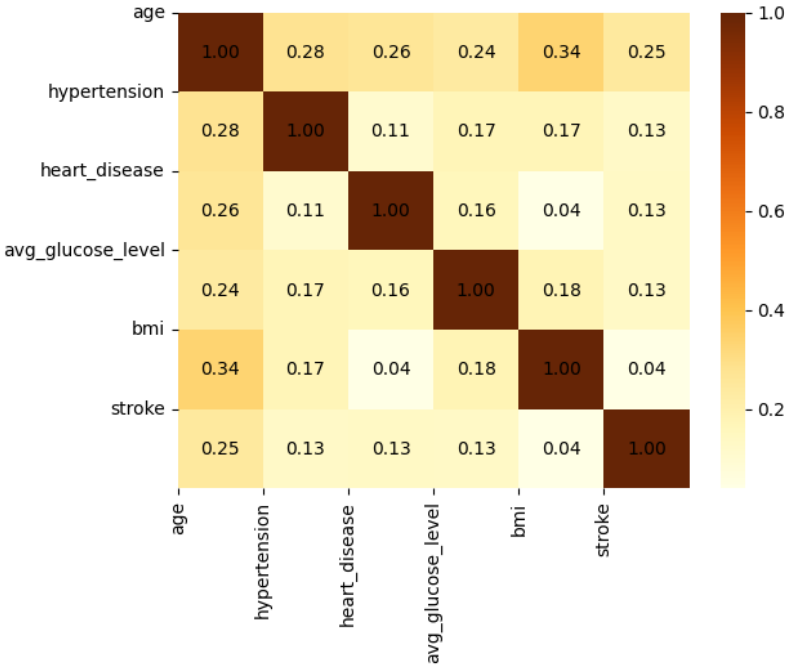
Εικόνα 2: Γράφημα ράβδων, απεικόνιση αριθμού των περιπτώσεων για τάξεις εγκεφαλικού επεισοδίου έναντι χωρίς εγκεφαλικό



Εικόνα 1: Γράφημα πίτας, απεικόνιση ανισορροπία κατηγορίας στο σύνολο δεδομένων Stroke

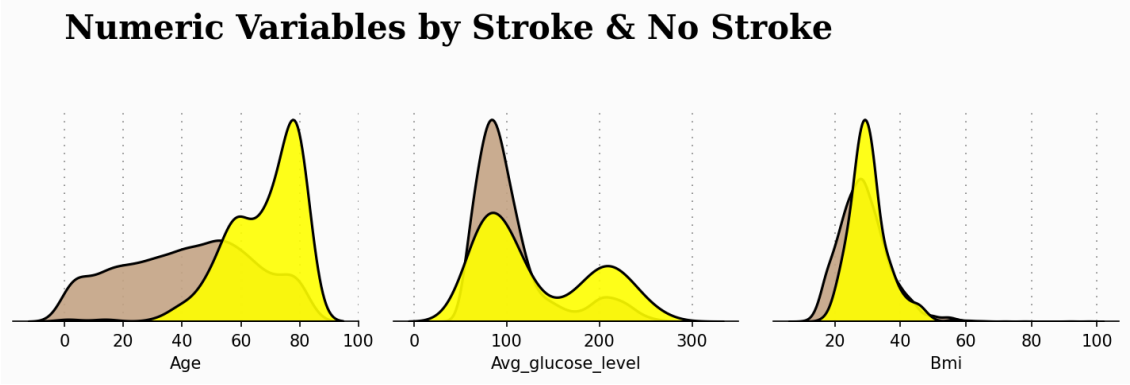
Ένα βήμα που βοηθάει στην ανάλυση του προβλήματος είναι η αναπαράσταση του πίνακα συσχέτισης. Η ανάλυση συσχέτισης αποτελεί μία μέθοδο που χρησιμοποιείτε για τον εντοπισμό προτύπων και πιθανόν συσχετίσεων στα δεδομένα. Στο συγκεκριμένο πρόβλημα καθώς ο στόχος είναι το εγκεφαλικό επεισόδιο, διερευνήθηκε η σχέση μεταξύ του στόχου και των δεικτών υγείας. Με βάση το Heatmap (εικόνα 3) οι τιμές κυμαίνονται από το 0 μέχρι το 1, όπου στο 0 δεν υπάρχει καμία γραμμική σχέση και στο 1 δείχνει μια τέλεια θετική γραμμική σχέση. Εξερευνώντας το, τονίζονται οι σχέσεις μεταξύ των διαφόρων δεικτών υγείας και εμφάνισης εγκεφαλικού επεισοδίου. Κάποιες παρατηρήσεις που μπορούμε να βγάλουμε από τον χάρτη, είναι ότι η ηλικία, οι καρδιακές παθήσεις, η υπέρταση και το μέσο επίπεδο γλυκόζης έχουν μία συσχέτιση με το εγκεφαλικό επεισόδιο. Αυτό οδηγεί στην παρατήρηση ότι οι παράγοντες αυτοί τείνουν προς εξέταση και πιο

συγκεκριμένα ο παράγοντας ηλικία η οποία φανερώνει την μεγαλύτερη συσχέτιση με το εγκεφαλικό.



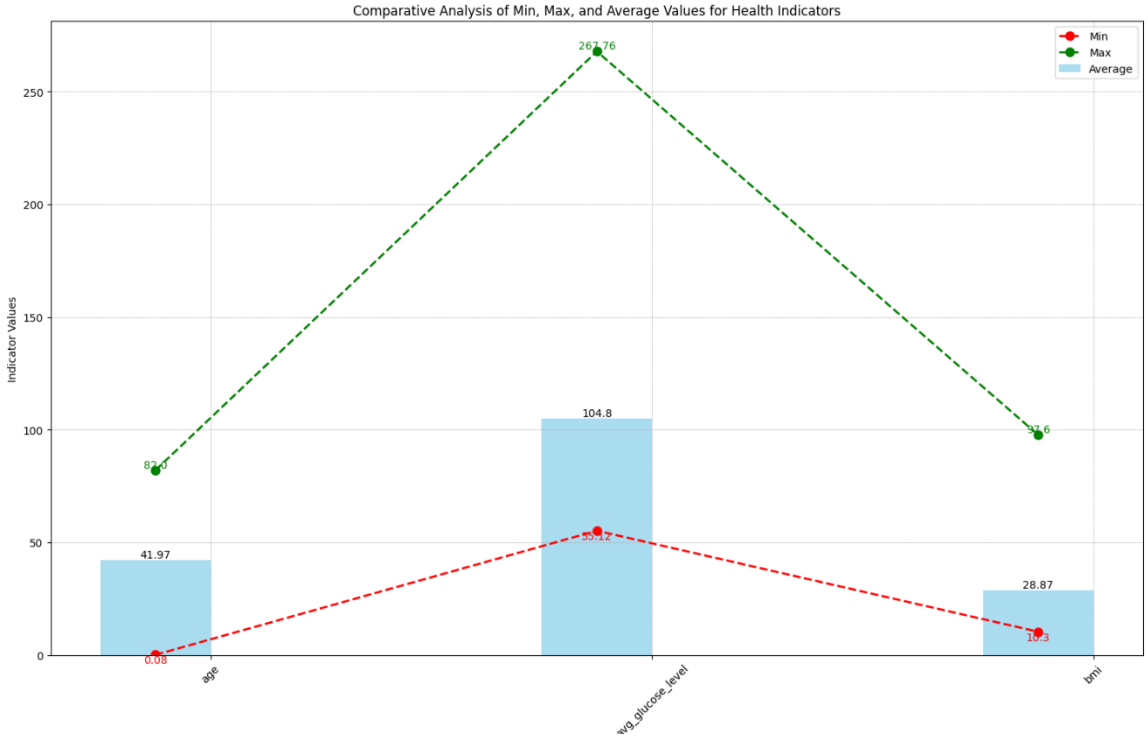
Εικόνα 3: Χάρτης θερμότητας των συντελεστών συσχέτισης μεταξύ δεικτών υγείας και εγκεφαλικού επεισοδίου

Εφόσον έγιναν αυτές οι παρατήρησης έγινε η παραγωγή διαγραμμάτων KDE (Kernel Density Estimation). Πλέον υπάρχει μια πιο προφανής απάντηση για τον ρόλο που διαδραματίζει η ηλικία ενός ατόμου ως αίτιο θανάτου. Τα άλλα δύο διαγράμματα (Avg_glucose_level, Bmi)(εικόνα 4)δεν δείχνουν κάτι προφανές σε αντίθεση με την ηλικία, η οποία εκφράζει ότι άτομα στην ηλικία 80 χρονών έχουν πολλές πιθανότητες να πάθουν εγκεφαλικό σε αντίθεση με άτομα κάτω των 40 χρονών.

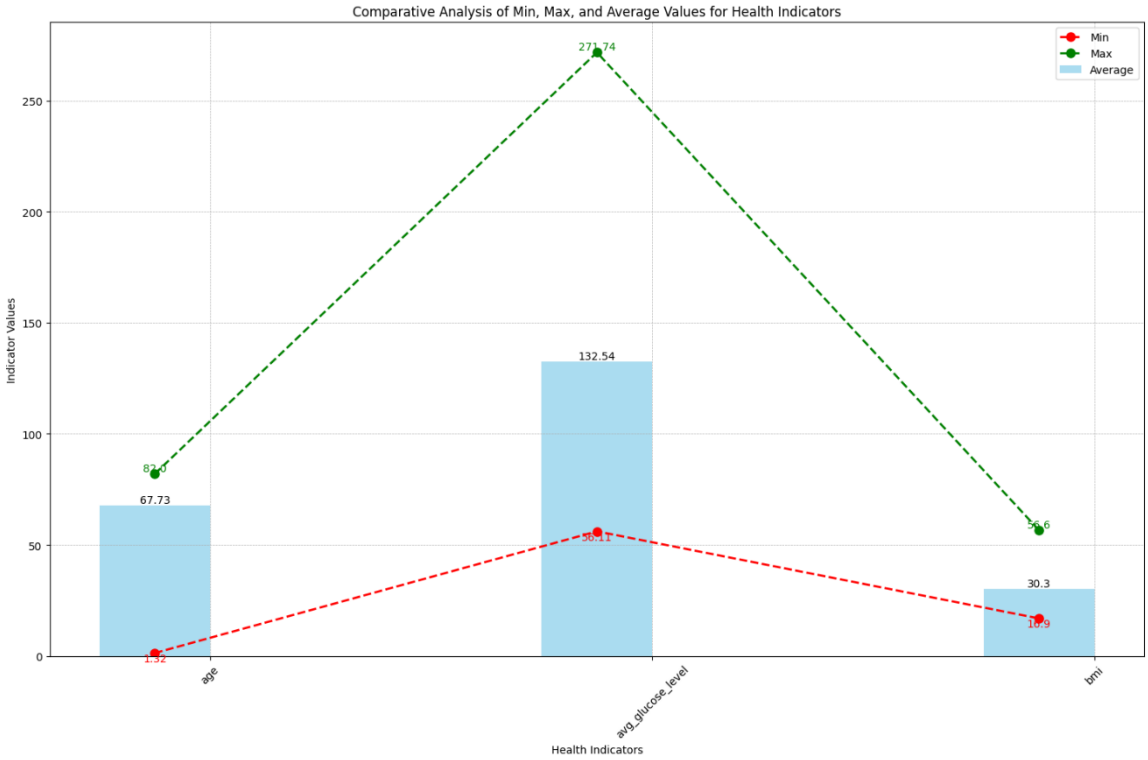


Εικόνα 4: Εκτιμήσεις πυκνότητας πυρήνα της ηλικίας, του επιπέδου γλυκόζης και του ΔΜΣ κατά κατάσταση εγκεφαλικού επεισοδίου

Στην συνέχεια γίνεται μια οπτικοποίηση συγκριτικών γραφημάτων. Στόχος αυτών τον γραφημάτων είναι η σύγκριση ελάχιστων, μέγιστων και μέσων τιμών. Οι τιμές αυτές διεξήχθησαν για τις τιμές ηλικία, μέσου επίπεδου γλυκόζης και Bmi για άτομα με και χωρίς εγκεφαλικό (εικόνα 5 – εικόνα 6). Τα γραμμικά γραφήματα, σε συνδυασμό με τα γραφήματα ράβδων παρέχουν την έντονη διαφορά των τιμών, που μπορεί να βοηθήσει στο να φτιαχτεί το εύρος των μεταβλητών σε σχέση με την εμφάνιση εγκεφαλικού επεισοδίου.

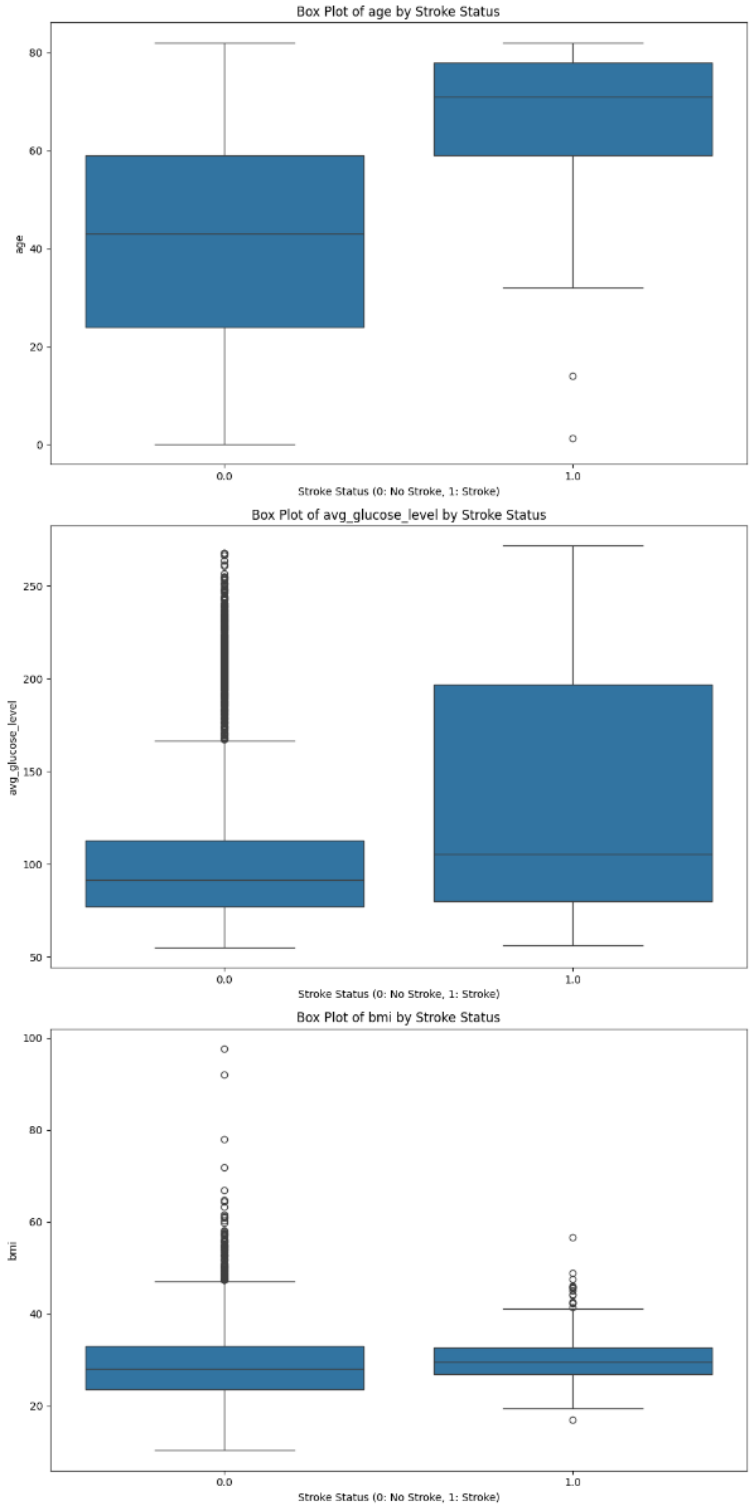


Εικόνα 5: Συγκριτικό διάγραμμα ράβδων ελάχιστων, μέγιστων και μέσες τιμές για βασικούς δείκτες υγείας μεταξύ των ασθενών που δεν έχουν εγκεφαλικό επεισόδιο



Εικόνα 6: Συγκριτικό ραβδόγραμμα ελάχιστων, μέγιστων και μέσων τιμών για βασικούς δείκτες υγείας μεταξύ ασθενών με εγκεφαλικό επεισόδιο

Καθώς η ηλικία, το μέσο επίπεδο γλυκόζης και το Bmi αποτελούν μεταβλητές αρκετά σημαντικές για το πρόβλημα, καθίσταται αναγκαία η εύρεση της κεντρικής τάσης των μεταβλητών μαζί με πιθανές ακραίες τιμές. Κάτι τέτοιο μπορεί να γίνει με την χρήση Box plots. Στην εικόνα 4.7 εντοπίζονται οι διαφορές στην κατανομή των βασικών δεικτών μεταξύ ασθενών που έχουν υποστεί εγκεφαλικό και αυτών που δεν έχουν.



Εικόνα 7: Box plots που απεικονίζουν την κατανομή της ηλικίας, του επιπέδου γλυκόζης και του Bmi κατά κατάσταση εγκεφαλικού επεισοδίου

5 Μοντελοποίηση και αξιολόγηση Μοντέλων

Στο Κεφάλαιο αυτό θα γίνει μια αναφορά στα μοντέλα τα οποία επιλέχτηκαν να δοκιμαστούν σε αυτό το πρόβλημα. Θα γίνουν αναφορές στις μετρήσεις των μοντέλων σε διαφορετικές περιπτώσεις που εκπαιδεύτηκαν.

5.1 Επιλογή Μοντέλων

Σε αυτή την μελέτη έγινε χρήση συνολικά 13 μοντέλων μηχανικής μάθησης. Λόγο της κρίσιμης φύσης της εργασίας καθίσταται απαραίτητη η εύρεση του καλύτερου μοντέλου. Αυτά τα οποία επιλέχθηκαν για να δοκιμαστούν και να εκπαιδευτούν είναι:

- Ο Logistic Regression και Gaussian Naive Bayes για την απλότητα και την ερμηνευτικότητά
- K-Nearest Neighbors (KNN) για τη μη παραμετρική τους φύση
- Support Vector Machines (SVM) για την αποτελεσματικότητά τους σε χώρους υψηλών διαστάσεων.
- Τα Decision Trees για την ευκολία κατανόησης και ερμηνείας τους
- Random Forest, το Gradient Boosting και το AdaBoost επιλέχθηκαν για την ισχυρή απόδοσή τους στην ταξινόμηση. Συνδυάζουν πολλούς αδύναμους learners (“μαθητές”) παράγοντας έναν ισχυρό learner.
- Ο ταξινομητής Multi-layer Perceptron (MLP) συμπεριλήφθηκε για τη διερεύνηση τεχνικών deep learning (βαθιάς μάθησης).

5.2 Εκπαίδευση (training)

Για την εκπαίδευση τον μοντέλων χρειάστηκε προ επεξεργασία των δεδομένων. Τα στάδια της προεργασίας αναλύθηκαν στο κεφάλαιο 3. Εν συντομία έγινε χειρισμός των τιμών που έλειπαν, κωδικοποίηση μεταβλητών ορισμένων κατηγοριών και κανονικοποίηση των χαρακτηριστικών. Μετά μέσω της SMOTE αντιμετωπίστηκε η ανισορροπία κλάσης.

Όσον αφορά το κομμάτι της εκπαίδευσης, αρχικά το σύνολο δεδομένων χωρίστηκε σε train και test set δηλαδή σύνολα εκπαίδευσης και δοκιμών. Αυτό βοηθάει στην αξιολόγηση της γενίκευσης του μοντέλου. Κάθε μοντέλο εκπαιδεύτηκε στο training set

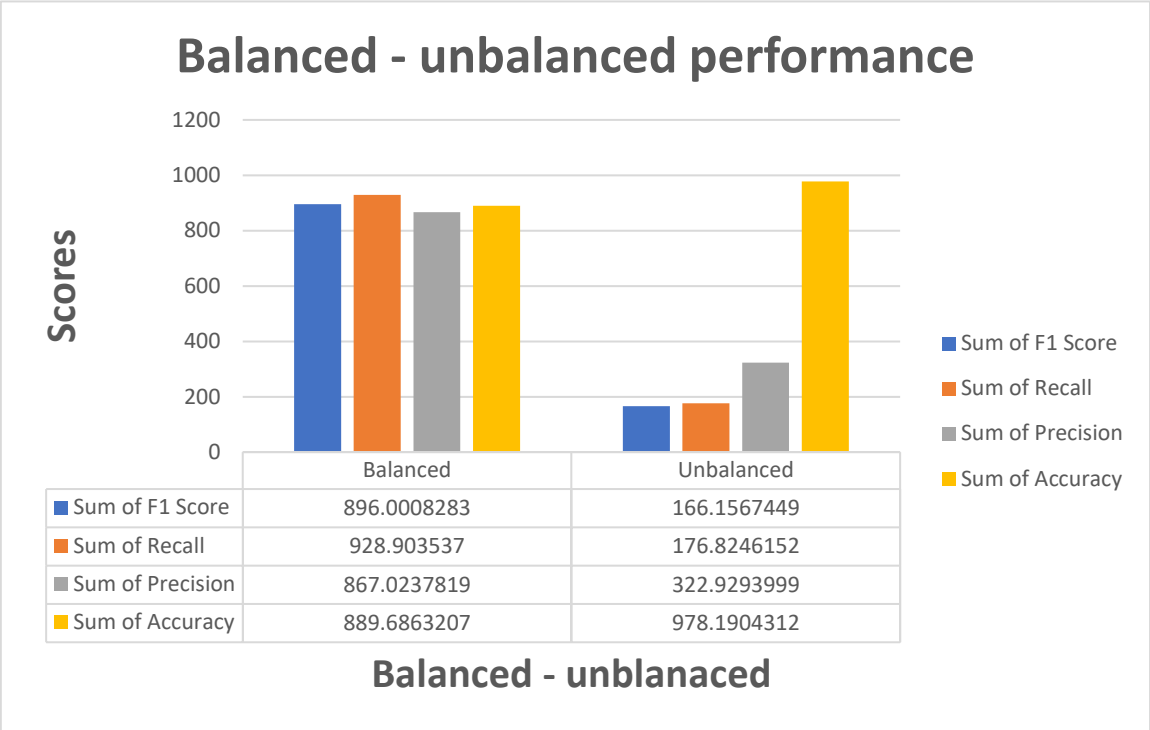
ενώ έγινε χρήση υπερπαραμέτρων για να βελτιστοποιηθεί. Επιπλέον, χάρι στην χρήση τεχνικών cross-validation υπάρχει αποφυγή overfitting (υπεπροσαρμογή) διασφαλίζοντας ότι τα μοντέλα θα μπορούν να γενικευτούν καλά και σε μη ορατά δεδομένα.

Επίσης, είναι σημαντικό να σημειωθεί ότι για την εξαγωγή αποτελεσμάτων οι μέθοδοι δοκιμάστηκαν και σε διαφορετικές καταστάσεις. Η καταστάσεις αυτές βοήθησαν ώστε να γίνει η σύγκριση με και χωρίς αυτές τις ενέργειες και πόσο σημαντικά άλλαξαν το τελικό αποτέλεσμα της εκπαίδευσης των μοντέλων. Παρακάτω θα αναφερθούν ονομαστικά μόνο ενώ σε επόμενο κεφάλαιο θα γίνει εμφάνιση των μετρήσεων και ανάλυση των αποτελεσμάτων:

- **2 επαναλήψεις**: Για δοκιμή και σε balanced-unbalanced (χρήση SMOTE)
- **4 επαναλήψεις**: Για handling missing values 3 + 1 remove (Bfill, Iterative Imputer, Linear Regression + Drop Values)
- **10 επαναλήψεις:** Για κάθε πείραμα (για την διαχείριση της περίπτωσης όπου κάποιο αποτέλεσμα θα εμφανίσει ασυνήθιστα αποτελέσματα)

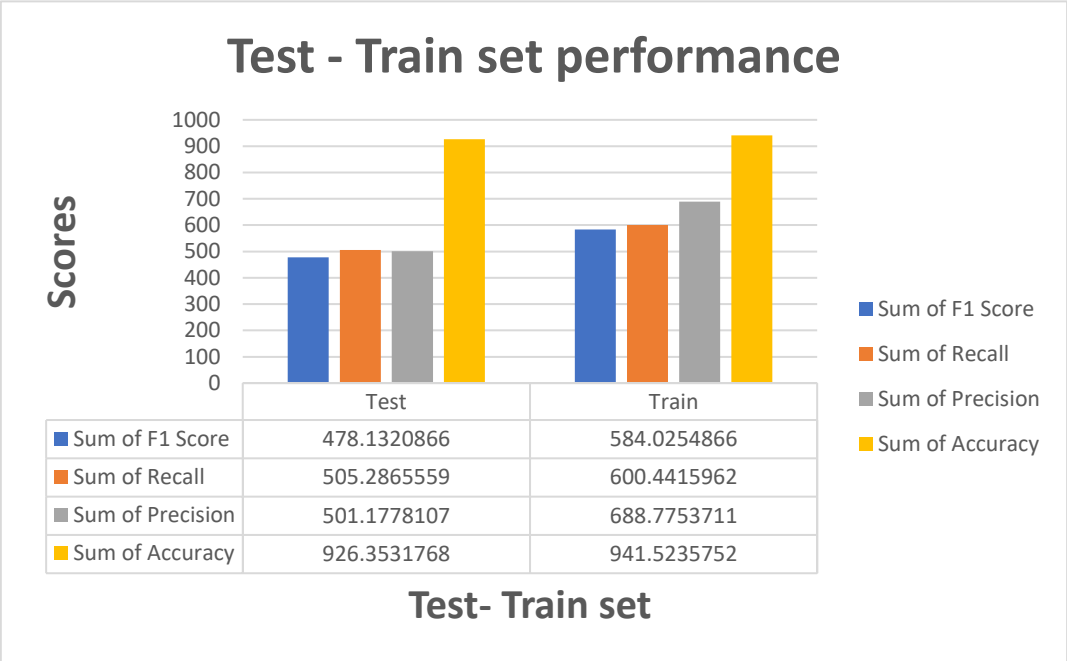
5.3 Εμφάνιση Μετρήσεων

Τα αποτελέσματα μετά την εκπαίδευση των μοντέλων αποθηκεύτηκαν σε ένα κεντρικό Excel με συνολικά 2080 εγγραφές (από την εκπαίδευση υπάρχουν 80 επαναλήψεις, 13 είναι οι μέθοδοι και 2 από το train και test set. Οπότε το τελικό αποτέλεσμα είναι 2080 εγγραφές). Λόγο του όγκου των γραφημάτων το κάθε γράφημα θα αποτελείτε μόνο από έναν τίτλο ενώ ο σχολιασμός θα εκφραστεί αναλυτικότερα στο κεφάλαιο 6.



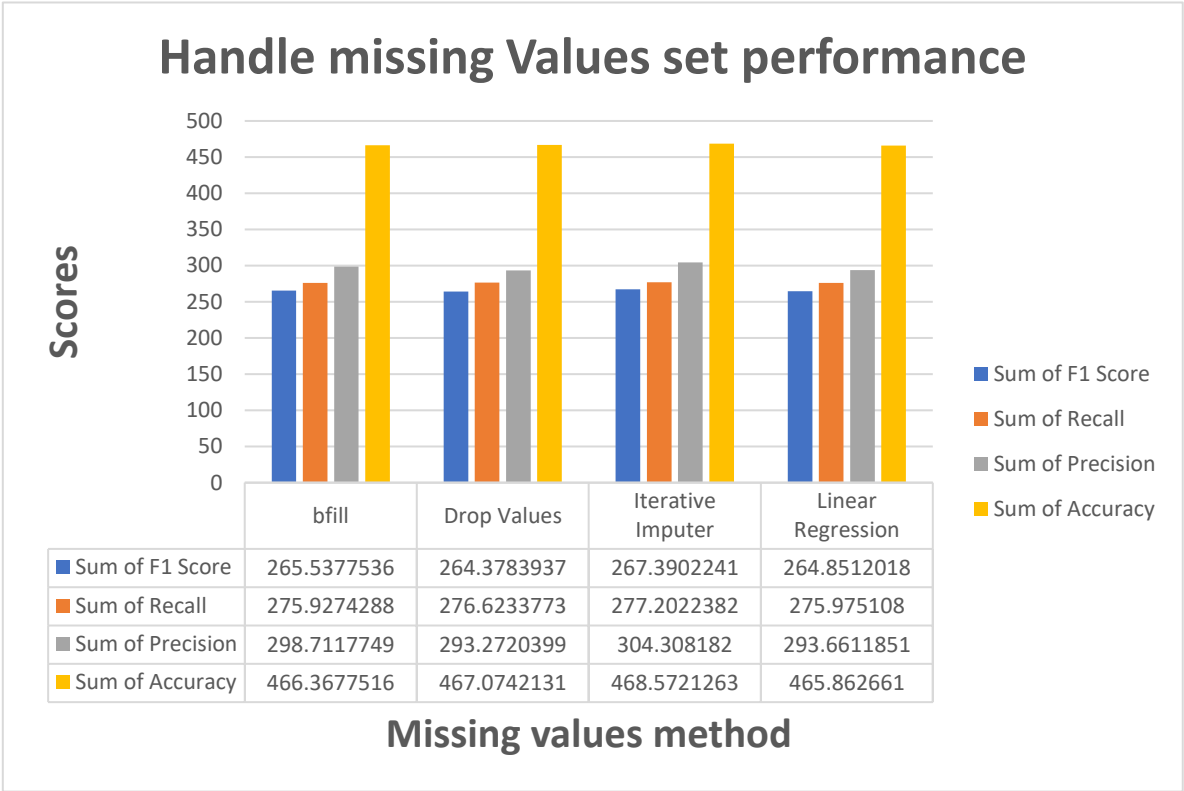
	Balanced	Unbalanced
Sum of F1 Score	896.0008283	166.1567449
Sum of Recall	928.903537	176.8246152
Sum of Precision	867.0237819	322.9293999
Sum of Accuracy	889.6863207	978.1904312

Εικόνα 8: Διαφορά μεταξύ Balanced – Unbalanced σε Recall, Accuracy, Precision, F1 Score, ROC AUC Score.



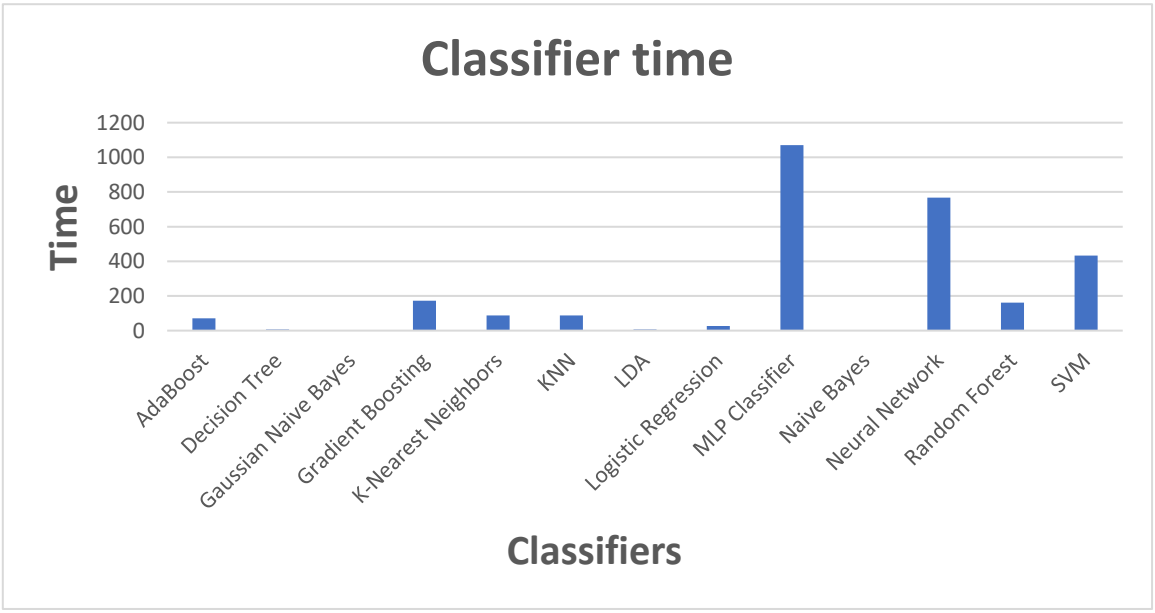
	Test	Train
Sum of F1 Score	478.1320866	584.0254866
Sum of Recall	505.2865559	600.4415962
Sum of Precision	501.1778107	688.7753711
Sum of Accuracy	926.3531768	941.5235752

Εικόνα 9: Διαφορά μεταξύ test – train set σε Recall, Accuracy, Precision, F1 Score, ROC AUC Score

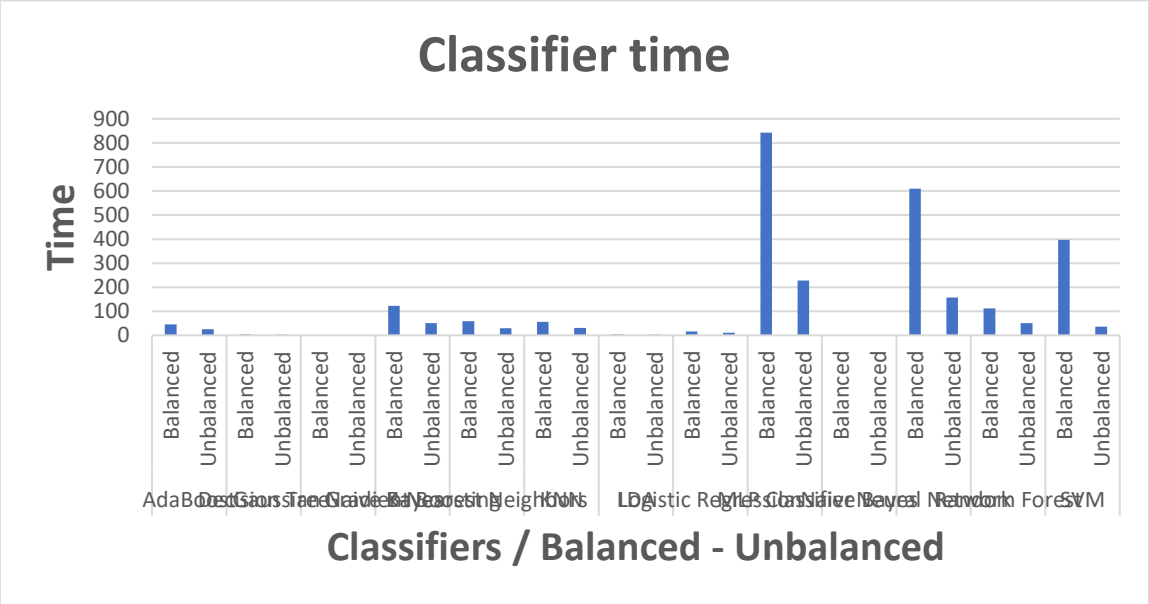


	bfill	Drop Values	Iterative Imputer	Linear Regression
Sum of F1 Score	265.5377536	264.3783937	267.3902241	264.8512018
Sum of Recall	275.9274288	276.6233773	277.2022382	275.975108
Sum of Precision	298.7117749	293.2720399	304.308182	293.6611851
Sum of Accuracy	466.3677516	467.0742131	468.5721263	465.862661

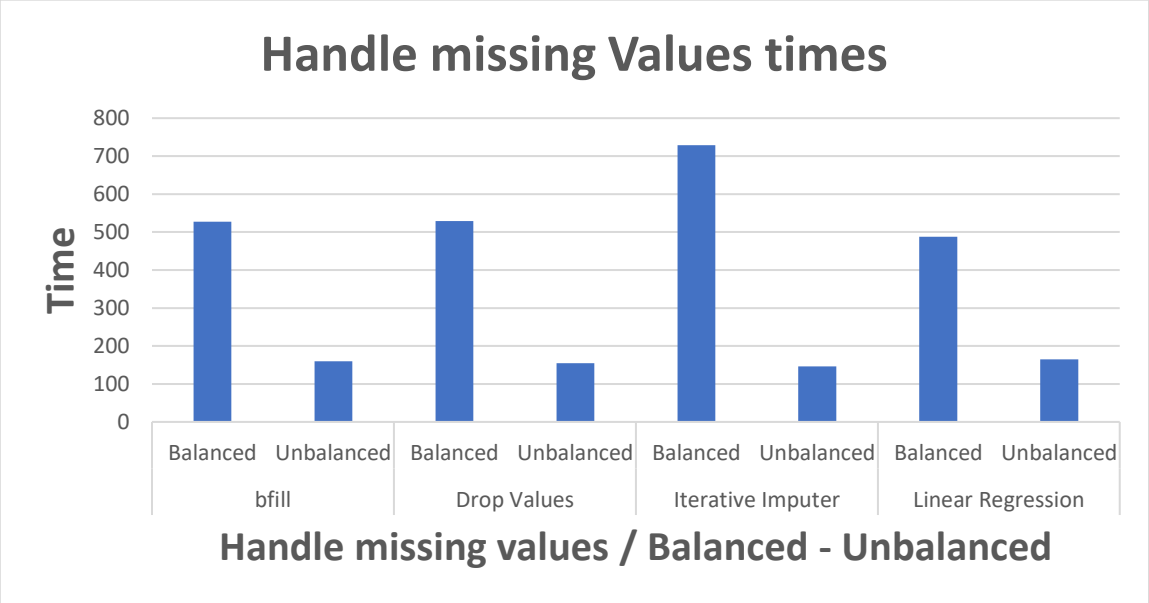
Εικόνα 10: Διαφορά μεταξύ μεθόδων διαχείρισης των τιμών που λείπουν σε Recall, Accuracy, Precision, F1 Score, ROC AUC Score



Εικόνα 11: Διαφορά χρόνων μεταξύ των μοντέλων ταξινόμησης

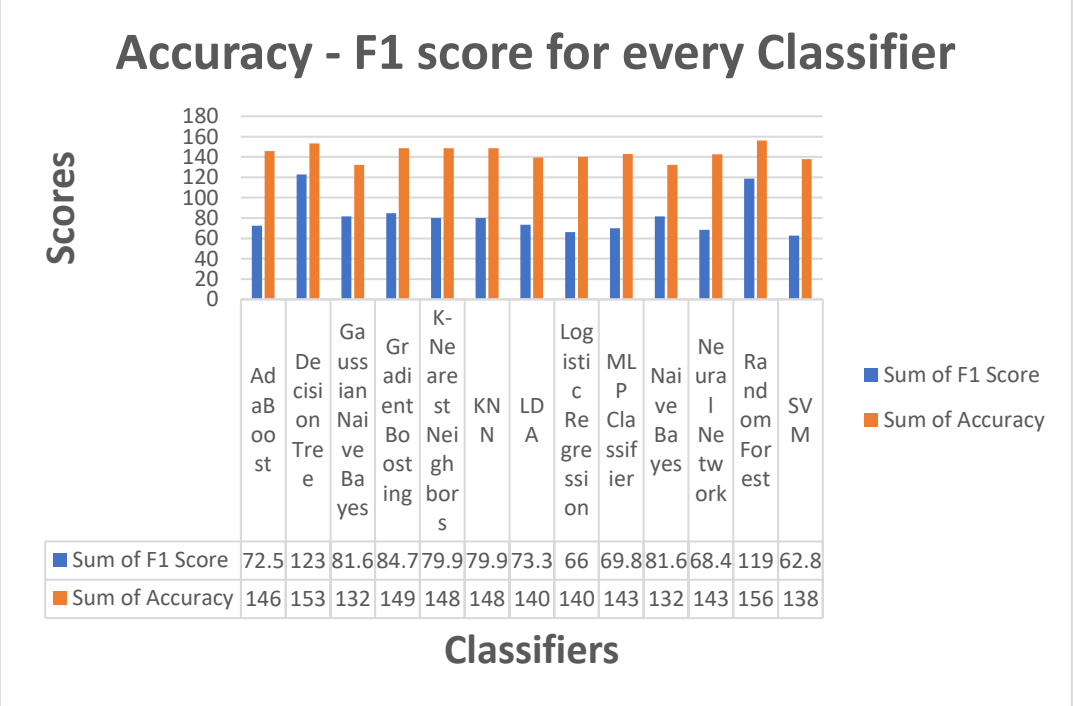


Εικόνα 12: Διαφορά χρόνων μεταξύ των μοντέλων ταξινόμησης πριν και μετά την χρήση SMOTE για ισορροπία της κλάσης

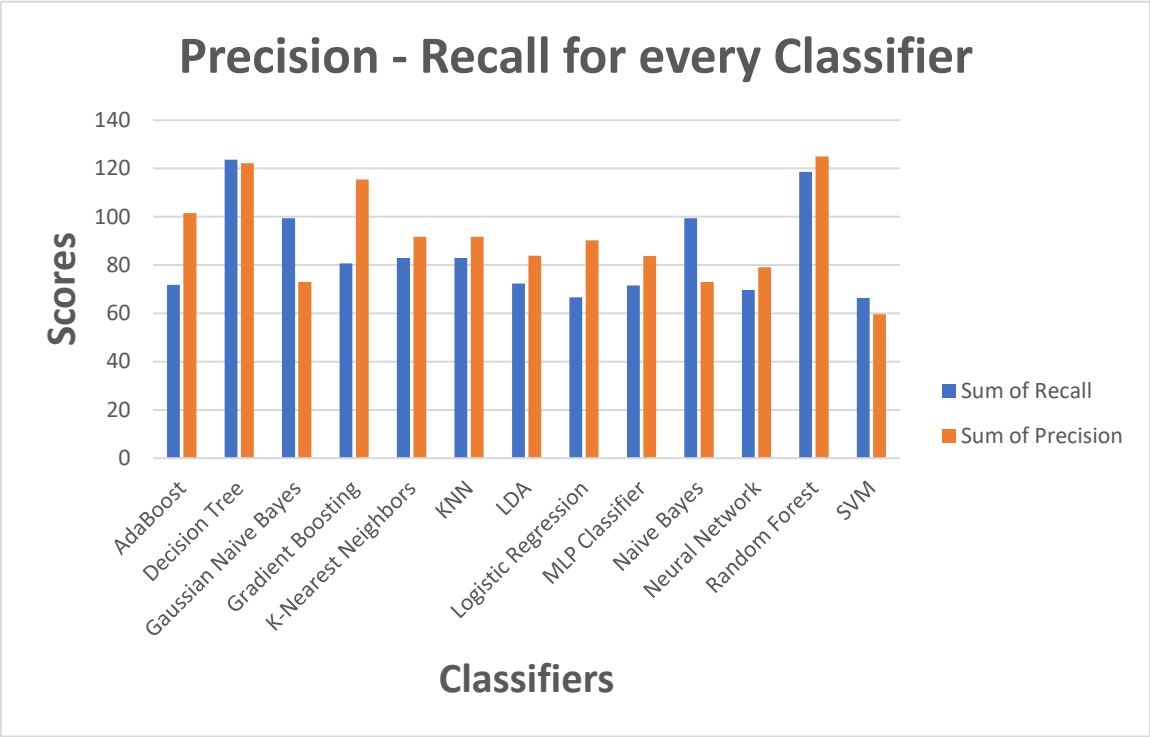


Εικόνα 13: Διαφορά χρόνων σε balanced-unbalance data για κάθε διαφορετική μέθοδο διαχείρησης χαμένων τιμών

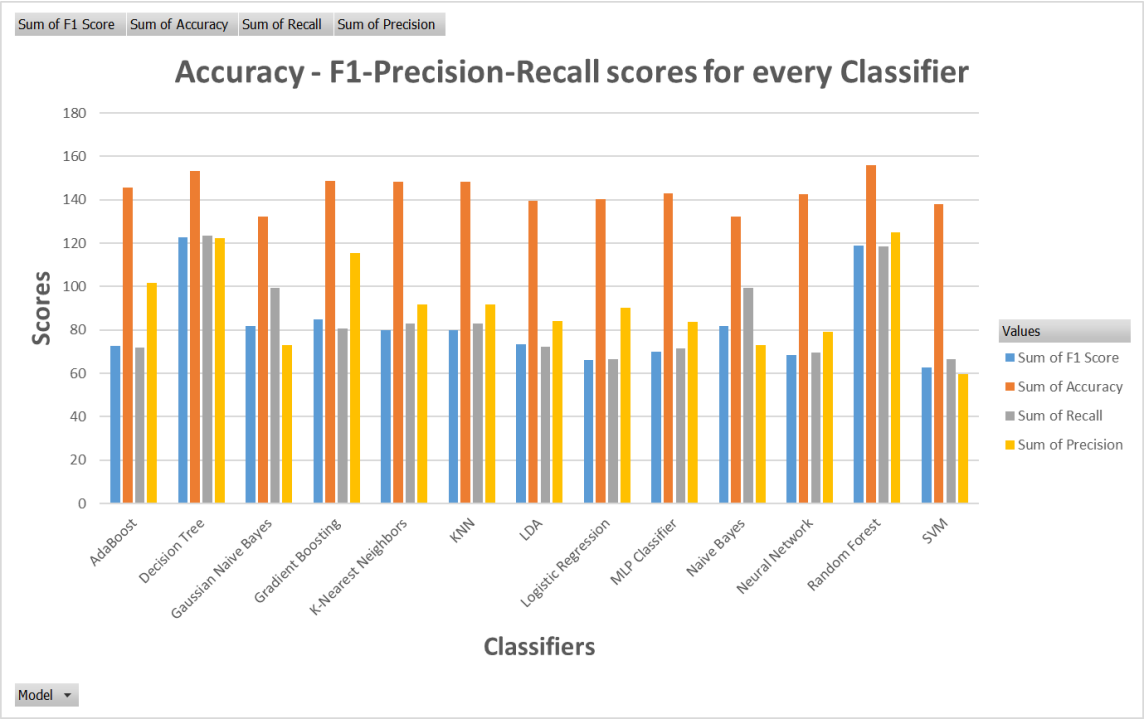
Παρακάτω θα γίνει η εμφάνιση των αποτελεσμάτων των μεθόδων ταξινόμησης για τα score Accuracy, F1, Precision, Recall:



Εικόνα 14: Accuracy – F1 score για κάθε μέθοδο ταξινόμησης



Εικόνα 15: Precision – Recall για κάθε μέθοδο ταξινόμησης



Εικόνα 16: Accuracy - F1-Precision-Recall για κάθε μέθοδο ταξινόμησης

6 Αποτελέσματα

Στο κεφάλαιο αυτό θα γίνει ανασκόπηση τον αποτελεσμάτων, καθώς και ανάλυση και αξιολόγηση των μοντέλων. Θα αναλυθεί η σημασία των μετρήσεων και η σημαντικότητα των τιμών και των γραφημάτων που αναφέρθηκαν στο κεφάλαιο 5.

6.1 Αξιολόγηση Μοντέλων

Για την καλύτερη ανάλυση και ακρίβεια των μοντέλων έγινε επισκόπηση των μετρήσεων F1 Score, Recall, Precision, Accuracy. Αυτές οι μετρήσεις προσφέρουν μια άποψη για την αποτελεσματικότητα των μοντέλων ταξινόμησης στα περιστατικά εγκεφαλικών επεισοδίων.

Επιπλέον αρκετά σημαντικό ρόλο διαδραμάτισαν και άλλοι παράγοντες όπως ο παράγοντας που αφορά την ισορροπία των δεδομένων (εικόνα 8). Όλες οι τιμές που αναφέρονται παραπάνω εκτός του accuracy, σημειώνουν ότι τα μοντέλα δεν μπορούν να παράγουν ικανοποιητικές προβλέψεις ενός εγκεφαλικού επεισοδίου σε μη ισορροπημένο σύνολο δεδομένων σε αντίθεση με ένα ισορροπημένο.

Αναμενόμενα είναι και τα αποτελέσματα του train και test set (εικόνα 9) των οποίων οι τιμές βρίσκονται αρκετά κοντά, με μια μικρή διαφορά στο Precision που σημειώνει την μεγαλύτερη διαφορά τιμής. Το αποτέλεσμα αυτό δείχνει καλά γενικευμένα μοντέλα.

Όσον αφορά την διαχείριση των τιμών που λείπουν, φαίνεται να μην επηρεάζει σημαντικά την εξαγωγή των αποτελεσμάτων. Οι τιμές είναι αρκετά κοντά, με την μόνη μέθοδο που ξεχωρίζει με μικρή διαφορά την **Iterative Imputer** (εικόνα 10) οι οποία σημειώνει τις καλύτερες τιμές σε κάθε μέτρηση.

Ο χρόνος σε τέτοια προβλήματα είναι επίσης ένας παράγοντας που επηρεάζει την απόφαση επιλογής κατάλληλου ταξινομητή, διότι ένα καλό μοντέλο θα πρέπει να μπορεί να παράγει γρήγορα και αξιόπιστα αποτελέσματα ώστε να μπορεί να λειτουργεί και με μεγαλύτερο όγκο δεδομένων στο μέλλον χωρίς να είναι χρονοβόρο. Με βάση τους χρόνους εκτέλεσης (εικόνα 11) ακολουθούν κάποια συμπεράσματα για τους ποιο γρήγορους και αργούς ταξινομητές.

Ξεκινώντας με τους γρηγορότερους ταξινομητές ο Gaussian Naive Bayes που είναι μια παραλλαγή του Naive Bayes και ο Naive Bayes είναι εξαιρετικά γρήγοροι ενώ χρειάζονται περίπου 2,21 και 2,26 δευτερόλεπτα, αντίστοιχα. Αυτό οφείλετε στο ότι είναι ανεξάρτητοι μεταξύ των χαρακτηριστικών και χρειάζονται λιγότερους υπολογισμούς. Στην κατηγορία των αμέσως πιο γρήγορων ταξινομητών κατατάσσονται οι:

- LDA (Linear Discriminant Analysis): Με χρόνο εκτέλεσης περίπου 6,65 δευτερόλεπτα. Ο LDA λόγο της χρήσης γραμμικών υπολογισμών για τον διαχωρισμό των κλάσεων αποτελεί συνήθως έναν γρήγορος ταξινομητής.
- Decision Tree: Ο τέταρτος στην λίστα με τους ποιο γρήγορους ταξινομητές με χρόνο περίπου 6,93 δευτερόλεπτα αποτελεί ο Decision Tree ο οποίος εκπαιδεύεται και κάνει προβλέψεις γρήγορα λόγο της απλής δομής του.
- Logistic Regression: Συγκριτικά με τους 4 προηγούμενους ταξινομητές ο Logistic Regression έχει έναν μέτριο χρόνο, περίπου 27,36 δευτερόλεπτα. Είναι σχετικά αποτελεσματικός αν και ο χρόνος μπορεί να αυξηθεί αν αυξηθεί και ο αριθμός των χαρακτηριστικών.

Συνεχίζοντας θα γίνει μια σύντομη ανάλυση και των ταξινομητών με τους μεγαλύτερους χρόνους ενώ ενδιάμεσες τιμές δεν θα σχολιαστούν. Ο πιο αργός με διαφορά είναι ο MLP Classifier με χρόνο 1069,61 δευτερόλεπτα. Αυτό οφείλετε στο γεγονός ότι τα νευρωνικά δίκτυα απαιτούν αρκετό χρόνο λόγο της πολυπλοκότητας τους και της ανάγκης τους για backpropagation για την εκπαίδευση σύνδεσης πολλαπλών επιπέδων. Στην κατηγορία εμμέσως πιο αργών ταξινομητών κατατάσσονται οι:

- Neural Network: Λόγο του χρόνου 767,72 δευτερολέπτων συγκαταλέγεται στους πιο αργούς ταξινομητές. Αυτό οφείλετε στην επαναληπτική διαδικασία της εκπαίδευσης του, εφόσον χρειάζεται περάσματα προς τα εμπρός και προς τα πίσω για κάθε επίπεδο
- SVM (Support Vector Machine): Το τρίτο και τελευταίο σε αυτήν την κατηγορία αποτελεί το SVM με χρόνο περίπου 433,58 δευτερόλεπτα. Ο λόγος είναι ότι κατά την διάρκεια της εκπαίδευσης λαμβάνει παραπάνω χρόνο για την βελτιστοποίηση που απαιτείται ώστε να μεγιστοποιήσει το περιθώριο μεταξύ των κλάσεων.

Σημείωση: Όλοι οι χρόνοι που αναφέρθηκαν παραπάνω βρίσκονται στον πίνακα 1 ο οποίος παρήγαγε το ραβδόγραμμα 11.

Πίνακας 1: Χρόνοι εκτέλεσης για κάθε classifier

Classifiers	Sum of Execution Time (seconds)
AdaBoost	71.85016394
Decision Tree	6.929674625
Gaussian Naive Bayes	2.212697983
Gradient Boosting	173.3759112
K-Nearest Neighbors	87.72803402
KNN	86.92702246
LDA	6.654024601
Logistic Regression	27.3576026
MLP Classifier	1069.606361
Naive Bayes	2.262599945
Neural Network	767.7170663
Random Forest	162.3097982
SVM	433.5806961

Με παρόμοιο τρόπο εξηγούνται και οι χρόνοι των ταξινομητών σε ισοζυγισμένα και μη σύνολα δεδομένων (πίνακας 12). Τα αποτελέσματα για το unbalanced έχουν χαμηλότερους χρόνους από το balanced ακολουθώντας την ίδια λίστα βαθμίδων που εξηγήθηκε αναλυτικά παραπάνω.

Οι ταξινομητές στο σύνολο τους εκτελούνται πολύ πιο γρήγορα σε μη ισοζυγισμένα σύνολα δεδομένων (εικόνες 12 – 13). Αν και αυτό είναι καλό δεν παίζει καθοριστικό ρόλο, καθώς τα αποτελέσματα για σωστή πρόβλεψη είναι τόσο χαμηλά που καθιστούν τα μοντέλα ανίκανα να βρούνε αν κάποιος ασθενής θα πάθει εγκεφαλικό ή όχι.

Στο σημείο αυτό θα γίνει ανάλυση των Accuracy, F1 score, Precision και Recall συγκριτικά για κάθε ταξινομητή μόνο στο test set καθώς αυτό αντιπροσωπεύει πόσο καλά ένα μοντέλο λειτουργεί σε ξένα δεδομένα, ένα μοντέλο που αποδίδει καλά στο σετ δοκιμών είναι αξιόπιστο ώστε να κάνει σωστές προβλέψεις σε πραγματικό περιβάλλον. Μέσο αυτών των τιμών θα γίνει η τελική αξιολόγηση και επιλογή του καλύτερου μοντέλου.

Στο accuracy και στο F1 score υπάρχουν δύο classifiers που διαφέρουν σε σχέση με τους άλλους. Ο Random Forest και ο Decision Tree φαίνεται να αποδίδουν καλύτερα υποδηλώνοντας ότι υπάρχει απόδοση και αξιοπιστία. Ωστόσο είναι σημαντικό να σημειωθεί ότι πρώτο ανέρχεται με καλύτερο F1 Score το Decision Tree με τιμή περίπου 42.83 (πίνακας 2), ενώ στο accuracy με μικρή διαφορά στην τιμή το Random Forest υπερτερεί με τιμή περίπου 76.08 (πίνακας 2).

Συνεχίζοντας στο Precision και στο Recall παρατηρείτε ότι τις καλύτερες τιμές έχουν οι ταξινομητές Naive Bayes, Gaussian Naive Bayes και Random Forest. Στο Recall με τιμή 49.63 περίπου ο Naive Bayes και ο Gaussian Naive Bayes φανερώνουν την ίδια μεγαλύτερη τιμή μεταξύ των υπολοίπων, ενώ στο Precision με τιμή 45 περίπου βρίσκεται στην πρώτη θέση ο Random Forest. (πίνακας 3).

Πίνακας 2: Τιμές F1 score - Accuracy για κάθε Classifier

Classifiers	Sum of F1 Score	Sum of Accuracy	Max of Accuracy
AdaBoost	35.28278128	72.74335466	76.08506957
Test	35.28278128	72.74335466	
Decision Tree	**42.83944184**	73.34324899	Max of F1 score
Test	**42.83944184**	73.34324899	42.83944184
Gaussian Naive Bayes	40.98085523	66.03464832	
Test	40.98085523	66.03464832	
Gradient Boosting	36.71514027	73.82746717	
Test	36.71514027	73.82746717	
K-Nearest Neighbors	36.66546211	73.53328462	
Test	36.66546211	73.53328462	
KNN	36.66546211	73.53328462	
Test	36.66546211	73.53328462	
LDA	36.32443374	69.73152621	
Test	36.32443374	69.73152621	
Logistic Regression	33.05033644	70.07425137	
Test	33.05033644	70.07425137	
MLP Classifier	34.59005674	71.22815809	
Test	34.59005674	71.22815809	
Naive Bayes	40.98085523	66.03464832	
Test	40.98085523	66.03464832	
Neural Network	33.90227162	71.1866677	
Test	33.90227162	71.1866677	
Random Forest	38.69727754	**76.08506957**	

Test	38.69727754	**76.08506957**	
SVM	31.43771242	68.99756717	
Test	31.43771242	68.99756717	
Grand Total	478.1320866	926.3531768	

Πίνακας 3: Τιμές για Precision και Recall για κάθε Classifier

Row Labels	Sum of Recall	Sum of Precision	Max of Precision
AdaBoost	35.27823689	39.51721867	45.00758002
Test	35.27823689	39.51721867	
Decision Tree	43.62458693	42.22551628	**Max of Recall**
Test	43.62458693	42.22551628	49.6370307
Gaussian Naive Bayes	**49.6370307**	36.71120815	
Test	**49.6370307**	36.71120815	
Gradient Boosting	36.86316277	39.88298392	
Test	36.86316277	39.88298392	
K-Nearest Neighbors	39.62080671	35.1742985	
Test	39.62080671	35.1742985	
KNN	39.62080671	35.1742985	
Test	39.62080671	35.1742985	
LDA	35.85866726	41.57997642	
Test	35.85866726	41.57997642	
Logistic Regression	33.27339894	43.04078706	
Test	33.27339894	43.04078706	
MLP Classifier	35.50554968	39.53591977	
Test	35.50554968	39.53591977	
Naive Bayes	**49.6370307**	36.71120815	
Test	**49.6370307**	36.71120815	
Neural Network	34.59732644	36.71606542	
Test	34.59732644	36.71606542	
Random Forest	38.62167481	**45.00758002**	
Test	38.62167481	**45.00758002**	
SVM	33.14827739	29.90074983	
Test	33.14827739	29.90074983	
Grand Total	505.2865559	501.1778107	

6.2 Σημασία αποτελεσμάτων/Αξιολόγηση Μετρήσεων

Πριν την τελική αξιολόγηση θα γίνει μια σύντομη ανασκόπηση της σημασίας κάθε μετρικής στο συγκεκριμένο πρόβλημα:

Accuracy: Είναι το ποσοστό των συνολικών προβλέψεων (που έχουν πάθει ή όχι εγκεφαλικό). Αυτό σημαίνει ότι σε ένα σύνολο ατόμων παραδείγματος χάρη 100 αν το μοντέλο έχει accuracy 76% μπορεί να προσδιορίσει σωστά 76 ασθενείς αν έχουν πάθει εγκεφαλικό ή όχι. Αυτή η μετρική δεν αρκεί όμως για μια τέτοιά πρόβλεψη, λόγο της πιθανότητας ο αριθμός των περιπτώσεων χωρίς εγκεφαλικό να υπερβαίνει κατά πολύ τις περιπτώσεις εγκεφαλικού.

F1 Score: Είναι ο μέσος της ακρίβειας (Precision) και της ανάκλησης (Recall). Παρέχει ισορροπία μεταξύ των δύο αυτών μετρικών όταν η κατανομή κλάσης είναι άνιση. Μια υψηλή βαθμολογία F1 υποδηλώνει ότι το μοντέλο μπορεί να καταγράψει τις περισσότερες από τις πραγματικές περιπτώσεις εγκεφαλικού επεισοδίου. Επίσης η υψηλή τιμή δηλώνει ότι οι τιμές για Recall και Precision είναι δεκτές, δηλαδή οι προβλέψεις είναι σωστές και δεν έχουμε ψευδή αποτελέσματα.

Recall: Είναι η ικανότητα του μοντέλου να εντοπίζει τα πραγματικά θετικά (True positive), που στο πρόβλημα αυτό είναι οι ασθενείς που θα υποστούν εγκεφαλικό. Μια υψηλή βαθμολογία recall σημαίνει ότι θα προβλέψει σωστά τους ασθενείς που διατρέχουν τον κίνδυνο εγκεφαλικού. Μια υψηλή τιμή recall μπορεί να σώσει ζωές ανθρώπων που μπορεί να πάθουν εγκεφαλικό και χρειάζονται άμεση ιατρική φροντίδα.

Precision: Είναι η ακρίβεια των θετικών προβλέψεων. Μια υψηλή τιμή Precision δείχνει ότι η πρόβλεψη ενός ασθενή για εγκεφαλικό είναι πιθανόν σωστή. Η τιμή του Precision εκφράζεται ελεύθερα ως το σύνολο των ασθενών που έγινε η πρόβλεψη ότι θα πάθουν εγκεφαλικό σε σχέση με αυτούς που έπαθαν όντως.

Συνοψίζοντας από τα παραπάνω, το καλύτερο μοντέλο για την πρόβλεψη εγκεφαλικού πρέπει να έχει ιδανικά για όλες αυτές τις μετρήσεις υψηλές τιμές. Αν και οι τιμές των μοντέλων ήταν αρκετά κοντά το Random Forest ξεχώρισε για την σταθερή του απόδοση. Η σταθερότητα και η συνέπεια είναι απαραίτητες σε ιατρικές εφαρμογές όπου η αξιοπιστία είναι ζωτικής σημασίας. Είχε καλές τιμές στο F1 Score και Accuracy, ωστόσο αυτό που το ξεχώρισε από τα υπόλοιπα είναι η μεγαλύτερη τιμή precision και recall που είχε, υποδηλώνοντας ότι λειτουργεί καλά σε διαφορετικά υποσύνολα δεδομένων.

7 Σύνοψη και συμπεράσματα

7.1 Επίλογος

Αυτό το project ξεκίνησε με στόχο την πρόβλεψη εγκεφαλικών επεισοδίων χρησιμοποιώντας ταξινομητές μηχανικής μάθησης. Εφόσον καθορίστηκε το πρόβλημα έγινε προ επεξεργασία και διερευνητική ανάλυση των δεδομένων (Exploratory Data Analysis – EDA). Εφόσον έγινε καλύτερη κατανόηση της φύσης των δεδομένων, καθοριστικό ρόλο στις αποφάσεις διαδραμάτισε και η οπτικοποίηση κάποιων ιατρικών τιμών των ασθενών. Στην συνέχεια έγινε η επιλογή των μοντέλων αξιολόγησης τα οποία αξιολογήθηκαν αυστηρά κάτω από ορισμένες συνθήκες. Η πορεία αυτής της εργασίας αποκάλυψε τις διαφορές μεταξύ των μετρήσεων αξιολόγησης και της σημασία επιλογής ενός ισορροπημένου μοντέλου.

7.2 Συμπεράσματα

Συνολικά υπήρχαν 13 ταξινομητές που αξιολογήθηκαν. Μετά την επεξεργασία και της κατάλληλες τροποποιήσεις έγινε ανάλυση των αποτελεσμάτων φανερώνοντας για αυτή την εργασία ως καλύτερο τον ταξινομητή Random Forest. Παρείχε ισορροπία μεταξύ της ακρίβειας (precision) και ανάκλησης (recall), μεταξύ των υπολοίπων δοκιμασμένων μοντέλων. Η παρακάτω εντολή δοκιμάστηκε σε κάθε μέθοδο διαχείρισης τιμών που έλειπαν ώστε να βρεθεί η μέθοδος η οποία θα παρείχε τις καλύτερες τιμές δηλαδή προβλέψεις. (καθώς τα μοντέλα λειτουργούσαν με διαφορά καλύτερα μετά την χρήση της SMOTE οι τιμές είναι μόνο σε ισοζυγισμένο μοντέλο)

```
=AVERAGEIFS([Precision Column], [Model Column], "Random Forest", [Set Column], "Test")
```

Τα αποτελέσματα φαίνονται στους πίνακες 4-7 και με κόκκινο είναι η μεγαλύτερη τιμή μεταξύ των διαφορετικών τεχνικών. Η τεχνική Iterative Imputer ξεχωρίζει έχοντας τις μεγαλύτερες τιμές στο Precision, στο Accuracy και στο F1 Score και για αυτόν τον λόγο προτιμάται. Συμπερασματικά χρησιμοποιώντας τον ταξινομητή Random Forest, την

τεχνική Iterative Imputer και δημιουργώντας μια ισορροπία στα δεδομένα με την χρήση της SMOTE το μοντέλο:

- Προβλέπει ένα εγκεφαλικό, αυτό είναι σωστό περίπου 99,31% των περιπτώσεων
- Είναι σε θέση να προσδιορίσει σωστά τη περίπτωση εγκεφαλικού στο 94,57% των περιπτώσεων (μικρότερη πιθανότητα να χαθούν θετικές περιπτώσεις)
- Δείχνει ότι είναι αξιόπιστο στην διάκριση μεταξύ περιπτώσεων εγκεφαλικού και μη εγκεφαλικού με ακρίβεια 96,93%.
- Υποδηλώνει ότι είναι ισορροπημένο όσον αφορά την ακρίβεια και την ανάκληση με τιμή 0,96. Όσο ποιο κοντά είναι η τιμή στο 1 τόσο ποιο ισορροπημένο είναι το μοντέλο.

Τα αποτελέσματα αυτά υποδηλώνουν ένα αξιόπιστο και αποτελεσματικό μοντέλο.

Πίνακας 4: Τιμές μετρικών για τεχνική Bfill

Precision		**Recall**
0.931949		0.961094
Accuracy		**F1 Score**
0.945039		0.946286

Πίνακας 5: Τιμές μετρικών για τεχνική Drop Missing Values

Precision		**Recall**
0.935135		0.965987
Accuracy		**F1 Score**
0.949574		0.950287

Πίνακας 6: Τιμές μετρικών για τεχνική Iterative Imputer

Precision		**Recall**
0.993142		0.945672
Accuracy		**F1 Score**
0.969357		0.968812

Πίνακας 7: Τιμές μετρικών για τεχνική Linear Regression

Precision		**Recall**
0.931245		0.957639
Accuracy		**F1 Score**
0.943033		0.944243

8 Βιβλιογραφία

Σχετική έρευνα για το θεωρητικό υπόβαθρο:

Schwartz, L., Anteby, R., Klang, E., & Soffer, S. (2023). Stroke mortality prediction using machine learning: Systematic review. *Journal of the Neurological Sciences, 444*, 120529. https://doi.org/10.1016/j.jns.2022.120529

Stroke Prediction Data Set: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

---

Για visualizations και κώδικα:

https://www.kaggle.com/code/gaetanlopez/how-to-make-clean-visualizations

https://www.kaggle.com/code/faraahanwaaar/stroke-data-analysis-and-prediction

https://www.kaggle.com/code/ahmedterry/stroke-prediction-eda-classification-models

https://www.kaggle.com/code/mdromzanalom/data-mining-project

https://www.kaggle.com/code/mohamedelnahry/stroke-predicition-accuracy-95

https://www.kaggle.com/code/joshuaswords/predicting-a-stroke-shap-lime-explainer-eli5

Για ανάλυση των μετρικών:

https://arize.com/blog-course/f1-score/

Classifiers (κώδικας και πληροφορίες για τον τρόπο λειτουργίας):

AdaBoost:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

KNN:

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Naive Bayes:

https://www.geeksforgeeks.org/naive-bayes-classifiers/

LDA:

https://scikit-learn.org/stable/modules/lda_qda.html

MPL Classifier:

https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

SVM:

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

Decision Tree:

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Random Forest:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Logistic Regression:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Gradient Boosting:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

Gaussian Naive Bayes:

https://builtin.com/artificial-intelligence/gaussian-naive-bayes

Neural Network:

https://scikit-learn.org/stable/modules/neural_networks_supervised.html

---

Μέθοδοι αντιμετώπισης των τιμών που έλειπαν:

Bfill: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.bfill.html

Iterative imputer:

https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

Drop Missing Values: https://scikit-learn.org/stable/modules/impute.html

Linear Regression:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Αντιμετώπιση ανισότητας:

SMOTE: https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/