ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Τμημα Πληροφορικης

IONIAN UNIVERSITY

A Stroke Prediction Study

Στέργιος Μουτζίκος
ΑΜ: inf2021149

Μάθημα: ΑΠΟΘΗΚΕΣ ΚΑΙ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ
Διδάσκοντες: κα. Κερμανίδου, κ. Έξαρχος

Μάιος 2025

Περιεχόμενα

.1 Περίληψη

Η παρούσα μελέτη διερευνά την αποτελεσματικότητα τεσσάρων αλγορίθμων μηχανικής μάθησης (Logistic Regression, Random Forest, Support Vector Machine και XGBoost) στην πρόβλεψη εγκεφαλικών επεισοδίων χρησιμοποιώντας ανισόρροπα (imbalanced) ιατρικά δεδομένα. Δόθηκε ιδιαίτερη έμφαση στη χρήση τεχνικών υποδειγματοληψίας για την αντιμετώπιση της ανισορροπίας των κλάσεων και στην αξιολόγηση των μοντέλων με βάση ποικίλες μετρικές απόδοσης, όπως ακρίβεια, ανάκληση, ROC AUC και RMSE. Τα αποτελέσματα δείχνουν ότι ο Logistic Regression προσφέρει την καλύτερη ισορροπία ανάμεσα σε αναγνωσιμότητα και προγνωστική ακρίβεια, ενώ ο SVM ξεχωρίζει σε διαχωριστική ικανότητα και ελαχιστοποίηση σφαλμάτων. Προτείνονται κατευθύνσεις για μελλοντικές βελτιώσεις, όπως η αξιοποίηση περισσότερων χαρακτηριστικών, η εφαρμογή εξελιγμένων τεχνικών εξισορρόπησης και η ενσωμάτωση ερμηνεύσιμων μοντέλων. Η εργασία στοχεύει να συμβάλει στην έγκαιρη και αξιόπιστη διάγνωση εγκεφαλικών επεισοδίων με τη βοήθεια της τεχνητής νοημοσύνης.

.2 Εισαγωγή

Το εγκεφαλικό επεισόδιο αποτελεί μία από τις κύριες αιτίες θνησιμότητας και μακροχρόνιας αναπηρίας παγκοσμίως. Πρόκειται για μία ιατρική κατάσταση κατά την οποία η ροή του αίματος προς μια περιοχή του εγκεφάλου διακόπτεται, με αποτέλεσμα τα εγκεφαλικά κύτταρα να στερούνται οξυγόνου και να καταστρέφονται μέσα σε λίγα λεπτά.
Η έγκαιρη πρόβλεψη και διάγνωση του εγκεφαλικού είναι ζωτικής σημασίας, καθώς κάθε λεπτό καθυστέρησης μπορεί να οδηγήσει σε σοβαρές και μη αναστρέψιμες βλάβες. Τα τελευταία χρόνια, η μηχανική μάθηση έχει αναδειχθεί ως ένα ισχυρό εργαλείο για την προγνωστική ανάλυση στην ιατρική, προσφέροντας τη δυνατότητα για έγκαιρη αναγνώριση των ατόμων με υψηλό κίνδυνο εμφάνισης εγκεφαλικού μέσω της ανάλυσης δημογραφικών, κλινικών και ιστορικών δεδομένων.

.2.1 Εισαγωγή στον Ερευνητικό Χώρο

- **Τι είναι η πρόβλεψη εγκεφαλικού επεισοδίου**
  Η πρόβλεψη εγκεφαλικού επεισοδίου (stroke prediction) αφορά την εκτίμηση της πιθανότητας εμφάνισης ενός εγκεφαλικού επεισοδίου σε έναν ασθενή, με βάση κλινικά, δημογραφικά και συμπεριφορικά χαρακτηριστικά. Πρόκειται για ένα πεδίο εφαρμογής της

μηχανικής μάθησης, στο οποίο αλγόριθμοι εκπαιδεύονται σε ιστορικά ιατρικά δεδομένα με σκοπό να ανιχνεύσουν πρότυπα και συσχετίσεις που σχετίζονται με τον κίνδυνο εγκεφαλικού. Η δυνατότητα πρόβλεψης ενός τέτοιου επεισοδίου μπορεί να συνδράμει στην έγκαιρη παρέμβαση, τη στοχευμένη πρόληψη και τη βελτίωση της έκβασης για τους ασθενείς.

- **Γιατί είναι σημαντικό πρόβλημα στη δημόσια υγεία**
  Το εγκεφαλικό επεισόδιο αποτελεί μια από τις κύριες αιτίες θνησιμότητας και μακροχρόνιας αναπηρίας παγκοσμίως. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, εκατομμύρια άνθρωποι υφίστανται εγκεφαλικά κάθε χρόνο, με σημαντικές κοινωνικές και οικονομικές επιπτώσεις. Η έγκαιρη ανίχνευση ατόμων υψηλού κινδύνου είναι ζωτικής σημασίας για την πρόληψη, καθώς πολλές περιπτώσεις εγκεφαλικών θα μπορούσαν να αποφευχθούν μέσω παρεμβάσεων στον τρόπο ζωής ή με φαρμακευτική αγωγή. Επομένως, η ανάπτυξη ακριβών και αποδοτικών συστημάτων πρόβλεψης έχει ιδιαίτερη σημασία για τα σύγχρονα συστήματα υγείας.
- **Δεδομένα που συνήθως χρησιμοποιούνται για την πρόβλεψη**
  Για την πρόβλεψη εγκεφαλικού, χρησιμοποιούνται συνήθως κλινικά δεδομένα και δημογραφικά χαρακτηριστικά των ασθενών. Τυπικά γνωρίσματα περιλαμβάνουν:
  1. **Ηλικία και φύλο**
  2. **Αρτηριακή πίεση**
  3. **Επίπεδα γλυκόζης**
  4. **Καρδιακές παθήσεις**
  5. **Ιστορικό καπνίσματος**
  6. **BMI (Δείκτης Μάζας Σώματος)**
  7. **Εργασιακή κατάσταση και τόπος διαμονής**

.2.2 Βασικές προσεγγίσεις επίλυσης

Η πρόβλεψη εγκεφαλικού ανήκει στον ευρύτερο χώρο της εφαρμοσμένης μηχανικής μάθησης (applied machine learning) σε ιατρικά δεδομένα. Η προσέγγιση βασίζεται στη μοντελοποίηση του προβλήματος ως δυαδική ταξινόμηση (binary classification),

όπου ο στόχος είναι να προβλεφθεί αν ένα άτομο είναι πιθανό να υποστεί εγκεφαλικό (κλάση 1) ή όχι (κλάση 0).

Για την επίλυση του προβλήματος, χρησιμοποιούνται αλγόριθμοι μηχανικής μάθησης που εκπαιδεύονται πάνω σε ιατρικά και "ιστορικά" δεδομένα.

- **Σύντομη εισαγωγή στους αλγορίθμους που χρησιμοποιήθηκαν:**

  1. **Logistic Regression**
     Η λογιστική παλινδρόμηση είναι ένα απλό αλλά ιδιαίτερα ερμηνεύσιμο γραμμικό μοντέλο για δυαδική ταξινόμηση.

  2. **Random Forest**
     Ο Random Forest είναι ένα σύνολο αποφασιστικών δέντρων (ensemble), το οποίο βασίζεται στη λήψη αποφάσεων μέσω πλειοψηφικής ψήφου από πολλαπλά δέντρα.

  3. **Support Vector Machine (SVM)**
     Ο SVM επιδιώκει να εντοπίσει το "threshold" που διαχωρίζει τις δύο κατηγορίες με το μεγαλύτερο δυνατό περιθώριο.

  4. **XGBoost**
     Ο XGBoost είναι μια εξελιγμένη τεχνική gradient boosting που ενσωματώνει regularization, early stopping και άλλες βελτιστοποιήσεις.

.2.3 Συνεισφορά της εργασίας

Η παρούσα εργασία εστιάζει στην αξιολόγηση και σύγκριση διαφορετικών αλγορίθμων μηχανικής μάθησης για την πρόβλεψη εγκεφαλικού επεισοδίου, αξιοποιώντας ένα ισορροπημένο υποσύνολο δεδομένων (249 περιπτώσεις με εγκεφαλικό και 249 χωρίς). Σε αντίθεση με πολλές προγενέστερες μελέτες που βασίζονται σε μη ισορροπημένα σύνολα, η εργασία αυτή εφαρμόζει τεχνικές υποδειγματοληψίας (undersampling) για την αντιμετώπιση του προβλήματος της μη ισορροπημένης ταξινόμησης, κάτι που συχνά υποβαθμίζει τη συνολική ακρίβεια σε πραγματικές εφαρμογές.

Συγκεκριμένα, η εργασία:

1. Παρουσιάζει συστηματική αξιολόγηση τεσσάρων δημοφιλών αλγορίθμων (Logistic Regression, Random Forest, SVM, XGBoost) με τη χρήση 10-πλής διασταυρού-

μενης επικύρωσης (10-fold cross-validation).

2. Συμπεριλαμβάνει ευρύ φάσμα μετρικών αξιολόγησης πέρα από την ακρίβεια (accuracy), όπως Precision, Recall, F1-score, ROC AUC, Cohen's Kappa, RMSE, RAE και SSE, για μια ολιστική κατανόηση της απόδοσης των μοντέλων.

3. Εξετάζει την επίδραση των υπερπαραμέτρων σε κάθε μοντέλο, προτείνοντας βελτιστοποιημένες ρυθμίσεις που ισορροπούν την ακρίβεια με την ευαισθησία στην ανίχνευση των εγκεφαλικών περιστατικών.

4. Προσφέρει πρακτικές συστάσεις για την επιλογή κατάλληλου μοντέλου σε περιβάλλοντα δημόσιας υγείας, όπου η ελαχιστοποίηση των ψευδώς αρνητικών (false negatives) είναι ιδιαίτερα κρίσιμη.

Η συνεισφορά της εργασίας ενισχύει τη γνώση γύρω από την εφαρμογή της μηχανικής μάθησης σε ιατρικά προβλήματα και προσφέρει χρήσιμες ενδείξεις για πραγματική χρήση τέτοιων μοντέλων σε κλινικά ή προγνωστικά συστήματα υποστήριξης αποφάσεων.

.3 Ερευνητικός χώρος με βιβλιογραφική επισκόπηση στην πρόβλεψη εγκεφαλικού με μηχανική μάθηση

Η πρόβλεψη εγκεφαλικού επεισοδίου αποτελεί κρίσιμο πεδίο έρευνας στη βιοϊατρική πληροφορική, με στόχο την έγκαιρη ανίχνευση ατόμων υψηλού κινδύνου και την πρόληψη σοβαρών επιπλοκών. Η χρήση αλγορίθμων μηχανικής μάθησης (ML) έχει αναδειχθεί ως αποτελεσματική προσέγγιση για την ανάλυση πολύπλοκων και ανισόρροπων ιατρικών δεδομένων.

(a) **An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data**

Christos Kokkotis, Georgios Giarmatzis, Erasmia Giannakou, κ.ά. (2022)

Η μελέτη αυτή προτείνει έναν ερμηνεύσιμο αγωγό μηχανικής μάθησης που αντιμετωπίζει την πρόκληση των ανισόρροπων δεδομένων. Χρησιμοποιήθηκαν

τεχνικές διαχείρισης κλάσεων και επιλέχθηκαν μοντέλα με δυνατότητα ερμηνείας, ώστε να επιτυγχάνεται τόσο ακρίβεια όσο και κατανόηση των αποτελεσμάτων. Τα τελικά μοντέλα παρουσίασαν ακρίβεια 73.5% και ποσοστό false negative 18.6%.

(b) **Predictive Modelling and Identification of Key Risk Factors for Stroke Using Machine Learning Approaches** *Συγγραφείς: Ahmad Hassan, Saima Gulzar Ahmad, Ehsan Ullah Munir, Imtiaz Ali Khan, Naeem Ramzan (2024)*

Η έρευνα αξιολόγησε πολλαπλούς αλγορίθμους ML σε εκτεταμένα datasets, εντοπίζοντας κρίσιμους παράγοντες κινδύνου, όπως ηλικία, BMI, επίπεδα γλυκόζης, υπέρταση και καρδιοπάθειες. Τα μοντέλα συγκρίθηκαν σε balanced και unbalanced σύνολα, αποδεικνύοντας την ανάγκη για κατάλληλες τεχνικές προεπεξεργασίας και επιλογής χαρακτηριστικών.

(c) **A Comparative Analysis of Machine Learning Classifiers for Stroke Prediction** *Nitish Biswas, Khandaker Mohammad Mohi Uddin, Sarreha Tasmin Rikta, Samrat Kumar Dey (2022)*

Η μελέτη αξιολόγησε την απόδοση διάφορων ταξινομητών (classifiers) όπως AdaBoost και Gradient Boost σε ανισόρροπα δεδομένα. Μέσω της τεχνικής Random Over Sampling (ROS), επιτεύχθηκε εξισορρόπηση των τάξεων, οδηγώντας σε καλύτερη αναγνώριση ατόμων υψηλού κινδύνου.

(d) **Predicting Stroke Risk Using Ensemble Learning Models on EHR Data** *Tashkova et al. (2025)*

Διερευνάται η εφαρμογή προηγμένων αλγορίθμων μηχανικής μάθησης για την πρόβλεψη του κινδύνου εγκεφαλικού επεισοδίου, βασιζόμενη σε δεδομένα ηλεκτρονικών ιατρικών φακέλων (EHR). Η έρευνα επικεντρώθηκε στην αξι-

ολόγηση της αποτελεσματικότητας διαφόρων μοντέλων ταξινόμησης, συμπεριλαμβανομένων της Logistic Regression, Random Forest και XGBoost, με στόχο τόσο τη βελτίωση της προγνωστικής ακρίβειας όσο και την κατανόηση των κρισιμότερων παραγόντων κινδύνου. Το σύνολο δεδομένων περιελάμβανε ένα ευρύ φάσμα δημογραφικών και κλινικών μεταβλητών, όπως ηλικία, φύλο, παρουσία υπέρτασης, διαβήτη, δείκτη μάζας σώματος (BMI), επίπεδα γλυκόζης αίματος και ιστορικό καρδιακών παθήσεων. Για την αντιμετώπιση της ανισορροπίας μεταξύ των κλάσεων (εγκεφαλικό επεισόδιο έναντι μη-εγκεφαλικού), οι ερευνητές εφάρμοσαν προηγμένες τεχνικές όπως η συνθετική δειγματοληψία SMOTE και η στάθμιση βαρών των κατηγοριών (class weighting). Παράλληλα, χρησιμοποιήθηκαν μέθοδοι πολλαπλής συμπλήρωσης για τη διαχείριση ελλιπών δεδομένων.

Τα κύρια ευρήματα της μελέτης είναι:

- Το μοντέλο XGBoost επέδειξε την υψηλότερη προγνωστική απόδοση, με F1-score περίπου 0.83 και AUC άνω του 0.87, καθιστώντας το την βέλτιστη επιλογή για αυτό το πρόβλημα ταξινόμησης.
- Ο Random Forest παρουσίασε επίσης ικανοποιητική απόδοση, με ισορροπημένα μετρικά precision και recall.
- Ο Logistic Regression, αν και ελαφρώς κατώτερος σε απόδοση, προσέφερε σημαντικά πλεονεκτήματα σε ό,τι αφορά την ερμηνευσιμότητα των αποτελεσμάτων - ένα κρίσιμο στοιχείο για κλινικές εφαρμογές.

.4 Αλγόριθμοι Μηχανικής Μάθησης

Στην παρούσα εργασία εφαρμόστηκαν τέσσερις ευρέως χρησιμοποιούμενοι αλγόριθμοι μηχανικής μάθησης για την πρόβλεψη εγκεφαλικού επεισοδίου. Κάθε μοντέλο επιλέχθηκε βάσει των χαρακτηριστικών του σε προβλήματα δυαδικής ταξινόμησης και της αποδοτικότητάς του σε ιατρικά δεδομένα.

- **Logistic Regression**
  Η λογιστική παλινδρόμηση είναι ένα γραμμικό μοντέλο που χρησιμοποιείται για προβλήματα δυαδικής ταξινόμησης. Υπολογίζει την πιθανότητα ένα δείγμα να ανήκει σε μία από δύο κατηγορίες (π.χ., "εγκεφαλικό" ή

"όχι εγκεφαλικό") μέσω της συνάρτησης sigmoid, η οποία χαρτογραφεί τις τιμές σε εύρος [0,1].

Παράμετροι:

i. max_iter=1000: Μέγιστος αριθμός επαναλήψεων για τη σύγκλιση του αλγορίθμου. Αυξάνεται σε σχέση με την προεπιλογή ώστε να διασφαλιστεί η σύγκλιση του μοντέλου.

ii. C=0.5: Αντιστρόφως ανάλογο της δύναμης κανονικοποίησης (regularization). Μικρότερη τιμή ενισχύει το regularization, μειώνοντας τον κίνδυνο του overfitting.

Πλεονεκτήματα:

i. Ερμηνευσιμότητα: Παρέχει συντελεστές που εξηγούν τη βαρύτητα κάθε χαρακτηριστικού.

ii. Απλή υλοποίηση, κατάλληλη για μικρότερα datasets.

– **Random Forest**

Ο Random Forest είναι ένας αλγόριθμος σύνολου (ensemble), που αποτελείται από πολλαπλά δέντρα αποφάσεων. Το τελικό αποτέλεσμα προκύπτει από ψηφοφορία (majority voting) των επιμέρους δέντρων. Η τυχαιοποίηση κατά την κατασκευή κάθε δέντρου οδηγεί σε μεγαλύτερη γενίκευση και μικρότερο κίνδυνο υπερπροσαρμογής.

Παράμετροι:

i. n_estimators=250: Αριθμός δέντρων. Περισσότερα δέντρα οδηγούν συνήθως σε πιο σταθερά αποτελέσματα.

ii. max_depth=None: Το βάθος κάθε δέντρου είναι απεριόριστο, επιτρέποντας στο μοντέλο να μάθει και πολύπλοκες σχέσεις. Αν και αυξάνει τη δυνατότητα πρόβλεψης, απαιτεί προσοχή για overfitting.

Πλεονεκτήματα:

i. Ανθεκτικότητα στον θόρυβο και στα μη σημαντικά χαρακτηριστικά.

ii. Μπορεί να χειριστεί μη γραμμικές σχέσεις.

– **Support Vector Machine (SVM)**

O SVM στοχεύει στην εύρεση του υπερεπιπέδου που διαχωρίζει καλύτερα τις δύο κατηγορίες, με το μέγιστο περιθώριο μεταξύ των κοντινότερων σημείων (support vectors). Ιδανικό για προβλήματα υψηλής διαστατικότητας.

Παράμετροι:

i. kernel='rbf': Χρησιμοποιεί μη γραμμικό Radial Basis Function, επιτρέποντας στο μοντέλο να διαχωρίζει πολύπλοκες κατανομές.

ii. C=0.5: Παράγοντας regularization. Μικρότερες τιμές επιτρέπουν μεγαλύτερα περιθώρια με κόστος ορισμένων λαθών.

iii. probability=True: Ενεργοποιεί την πρόβλεψη πιθανοτήτων, απαραίτητη για την εξαγωγή ROC AUC

Πλεονεκτήματα:

i. Υψηλή ακρίβεια σε datasets με πολύπλοκες δομές.

ii. Κατάλληλο για μη γραμμικά προβλήματα.

iii. Σταθερή απόδοση ακόμα και με περιορισμένα δεδομένα.

– **XGBoost (Extreme Gradient Boosting)**

Ο XGBoost είναι ένας εξελιγμένος αλγόριθμος gradient boosting, ο οποίος δημιουργεί διαδοχικά μοντέλα για να διορθώσει τα λάθη των προηγούμενων. Είναι ιδιαίτερα αποδοτικός σε datasets με ανισόρροπες κλάσεις και υψηλή πολυπλοκότητα.

Παράμετροι:

i. learning_rate=0.1: Ελέγχει το μέγεθος του βήματος κατά την ενημέρωση του μοντέλου. Χαμηλότερη τιμή προσφέρει πιο αργή αλλά σταθερή εκμάθηση.

ii. scale_pos_weight: Παράμετρος για την εξισορρόπηση των βαρών των θετικών και αρνητικών παραδειγμάτων. Χρήσιμο σε περιπτώσεις όπως

το stroke dataset όπου υπάρχει αρχική ανισορροπία.

iii. use_label_encoder=False, eval_metric='logloss': Απενεργοποιεί το παλιό label encoder του XGBoost και καθορίζει τη μετρική απώλειας για την εκπαίδευση.

Πλεονεκτήματα:

i. Αντιμετωπίζει την ανισορροπία αποτελεσματικά.

ii. Πολύ καλή ακρίβεια και ευελιξία.

iii. Ενσωματώνει τακτικές πρόληψης υπερπροσαρμογής..

.5 Μεθοδολογική διαδικασία

Το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα μελέτη προέρχεται από ηλεκτρονικά ιατρικά αρχεία και περιλαμβάνει δημογραφικές, κλινικές και συμπεριφορικές μεταβλητές σχετικές με την πιθανότητα εμφάνισης εγκεφαλικού επεισοδίου. Κάθε γραμμή του συνόλου αντιπροσωπεύει έναν ασθενή και περιλαμβάνει τα παρακάτω χαρακτηριστικά:

- `id`: Μοναδικό αναγνωριστικό ασθενούς (δεν χρησιμοποιήθηκε στο μοντέλο).
- `gender`: Φύλο (Male, Female).
- `age`: Ηλικία ασθενούς (εύρος: 0.08 – 82 έτη).
- `hypertension`: Ιστορικό υπέρτασης (0: Όχι, 1: Ναι).
- `heart_disease`: Παρουσία καρδιακής νόσου (0: Όχι, 1: Ναι).
- `ever_married`: Έχει παντρευτεί ποτέ (Yes, No).
- `work_type`: Επαγγελματική κατηγορία (Private, Self-employed, Other).
- `Residence_type`: Τύπος κατοικίας (Urban, Rural).
- `avg_glucose_level`: Μέσο επίπεδο γλυκόζης στο αίμα (εύρος: 55.1 – 272).

- `bmi`: Δείκτης Μάζας Σώματος (BMI), με κάποιες ελλιπείς τιμές (missing values).
- `smoking_status`: Κατάσταση καπνίσματος (smokes, formerly smoked, never smoked, unknown).
- `stroke`: Μεταβλητή-στόχος, δηλώνει αν ο ασθενής είχε εγκεφαλικό επεισόδιο (0: Όχι, 1: Ναι).

.5.1 Περιγραφή Προεπεξεργασίας

- **Διαχείριση Ελλιπών Τιμών (Missing Values)**
  Η μοναδική στήλη με ελλιπείς τιμές ήταν η bmi (201 missing values), όπου αντικαταστάθηκαν με τον μέσο όρο της στήλης:

  Μέση τιμή BMI = 28.89
- **Διαγραφή της Id Στήλης**
  Η στήλη `id` αφαιρέθηκε καθώς δεν περιείχε πληροφορία σχετική με την πρόβλεψη.
- **Κωδικοποίηση Κατηγορικών Μεταβλητών**
  Οι δυαδικές κατηγορικές μεταβλητές κωδικοποιήθηκαν ως εξής:

  * gender: Male → 1, Female → 0
  * ever_married: Yes → 1, No → 0
  * Residence_type: Urban → 1, Rural → 0

  Οι "πολυκατηγορικές" μεταβλητές μετατράπηκαν με one-hot encoding:

  * `work_type`: Private, Self-employed, Never_worked, Govt_job, Other
  * `smoking_status`: smokes, formerly smoked, never smoked, unknown

  Μετά την κωδικοποίηση, έγινε και ανακατάταξη/μετονομασία ορισμένων στηλών για ευκολότερη ανάγνωση και επεξεργασία.

– **Αντιμετώπιση Ακραίων Τιμών (Outliers)**
Για τον εντοπισμό και τον χειρισμό των ακραίων τιμών (outliers), εφαρμόστηκε το Z-score method σε αριθμητικές στήλες , με όριο $Z > 3$.

Αποτελέσματα:

* avg_glucose_level: 49 ακραίες τιμές
* bmi: 59 ακραίες τιμές
* age και gender: Δεν εντοπίστηκαν ακραίες τιμές

Διόρθωση (Capping):
Για κάθε ακραία τιμή, εφαρμόστηκε capping: οι τιμές εκτός του ανεκτού εύρους αντικαταστάθηκαν με τα όρια του z_score =3, δηλαδή στην ανώτερη "επιτρεπτή" τιμή.

Παράδειγμα:

* Στη στήλη avg_glucose_level, τιμή 252.72 → capped σε 241.99
* Στη στήλη bmi, τιμή 56.6 → capped σε 51.99

Με αυτό τον τρόπο, διατηρήθηκαν όλες οι εγγραφές, ενώ περιορίστηκε η επίδραση των ακραίων τιμών.

.5.2 Γραφήματα

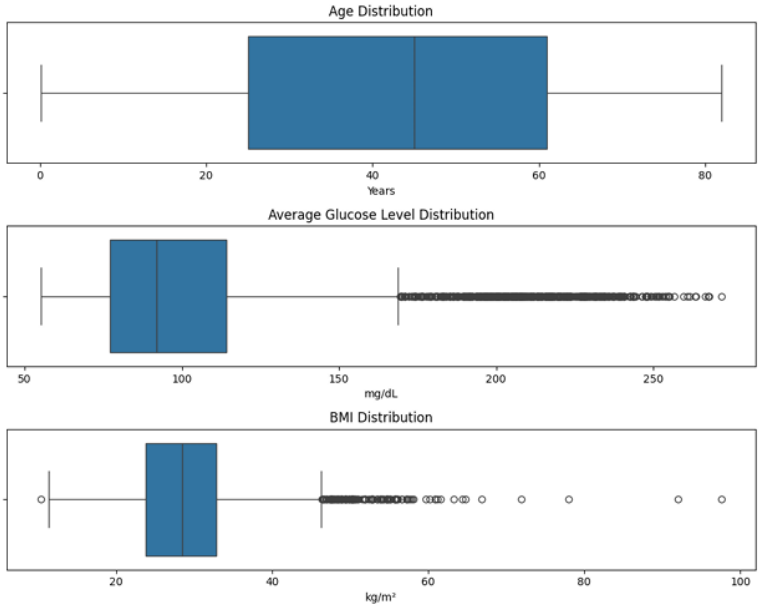

Figure 1: ***Κατανομές τιμών των μεταβλητών age, Average glucose level, BMI πριν να "χειριστούν" οι ακραίες τιμές τους***

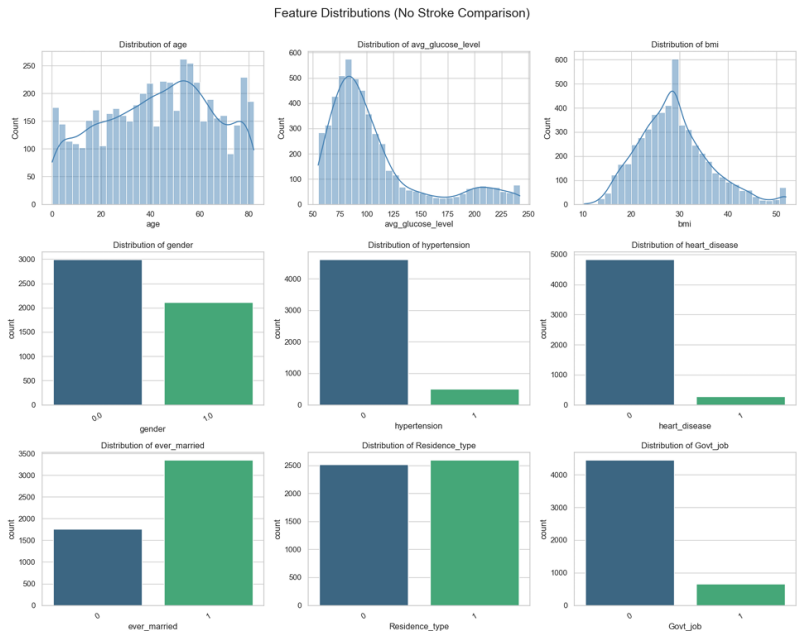

Figure 2: ***Κατανομές τιμών όλων των μεταβλητών, μετά το outlier handling (1/2)***

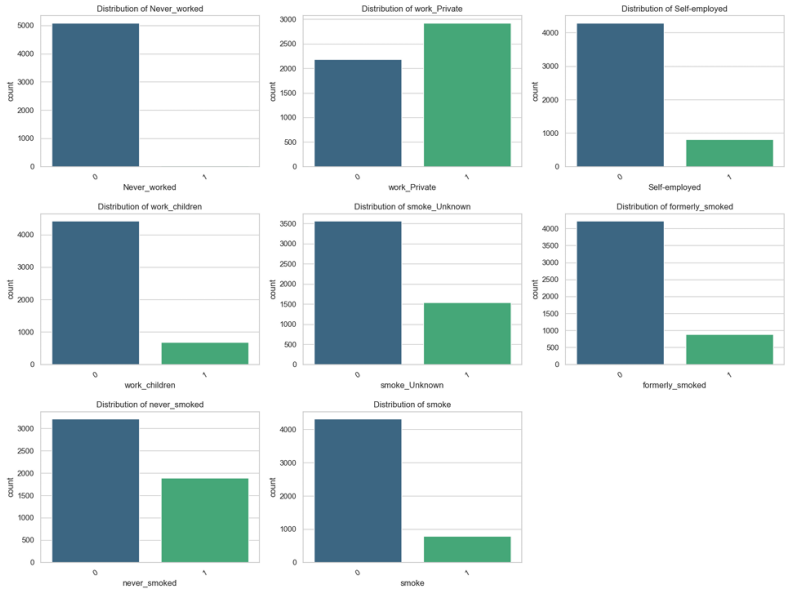

Figure 3: ***Κατανομές τιμών όλων των μεταβλητών, μετά το outlier handling (2/2)***

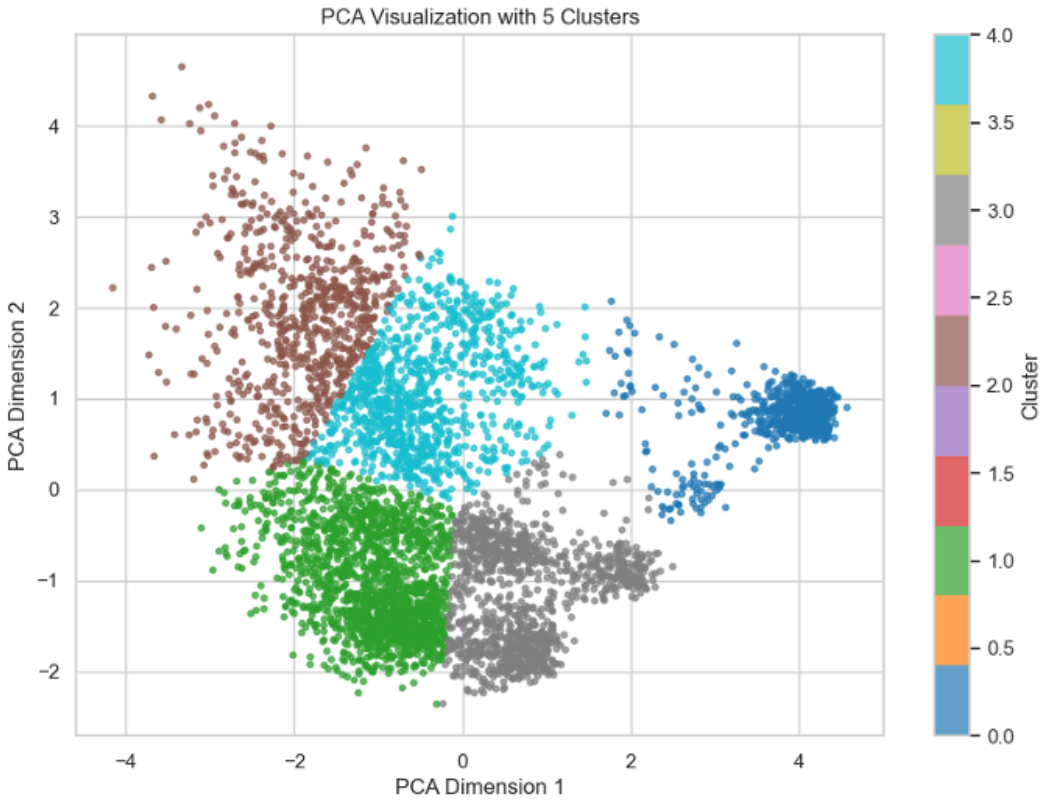

Figure 4: ***Πολυδιάστατο visualization με την χρήση του PCA***

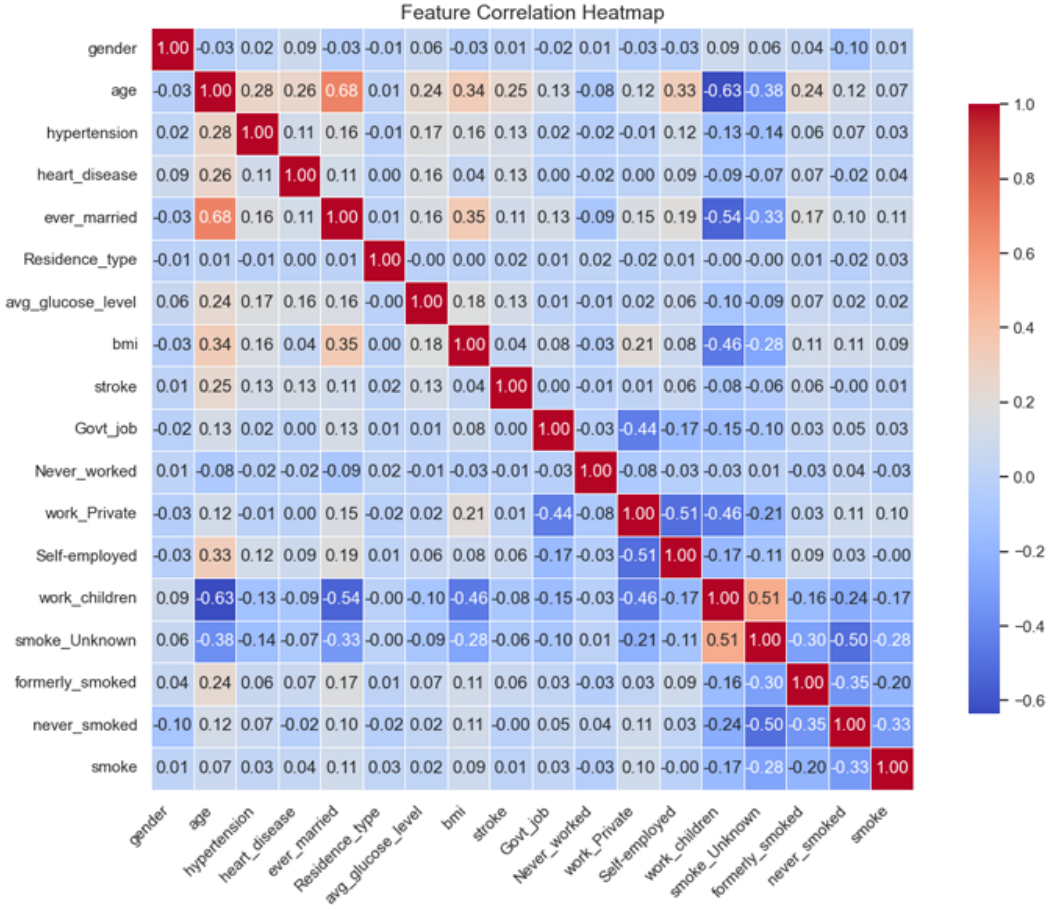

Figure 5: ***Feature correlation heatmap***

.5.3 Πειράματα μάθησης και σχολιασμός αποτελεσμάτων

Ο κύριος στόχος της φάσης των πειραμάτων ήταν η εκπαίδευση και αξιολόγηση εποπτευόμενων αλγορίθμων μηχανικής μάθησης για την πρόβλεψη εμφάνισης εγκεφαλικού επεισοδίου, χρησιμοποιώντας ένα ισορροπημένο (balanced) dataset (μέσω undersampling).

– **Αντιμετώπιση Imbalanced Data**

Το αρχικό dataset παρουσίαζε σοβαρή ανισορροπία στις κλάσεις, με σημαντικά λιγότερα περιστατικά εγκεφαλικού (5%) σε σχέση με τα μη εγκεφαλικά (95%). Για την αντιμετώπιση του προβλήματος χρησιμοποιήθηκε η τεχνική Random Undersampling, με αποτέλεσμα να εξισσοροπηθούν τα παραδείγματα της της πλειοψηφικής κλάσης (μη εγκεφαλικά) με αυτά της μειοψηφικής κλάσης (εγκεφαλικά), διατηρώντας 249 δείγματα για κάθε κατηγορία, με την χρήση του RandomUnderSampler από τη βιβλιοθήκη imblearn.

– **Διαδικασία Εκπαίδευσης**

Η αξιολόγηση πραγματοποιήθηκε με 10-Fold Stratified Cross-Validation, διατηρώντας την αναλογία των κλάσεων σε κάθε fold.

i. Διαχωρισμός σε training/test set.

ii. Εφαρμογή undersampling στο training set.

iii. Εκπαίδευση κάθε μοντέλου στα undersampled δεδομένα.

iv. Πρόβλεψη πιθανοτήτων στο test set.

v. Εύρεση βέλτιστου threshold που μεγιστοποιεί το F1-score.

vi. Υπολογισμός μετρικών απόδοσης.

– **Αξιολόγηση Μοντέλων**

Οι μετρικές που χρησιμοποιήθηκαν είναι:

i. Accuracy: Συνολική ακρίβεια προβλέψεων.

ii. Precision: Ποσοστό σωστών θετικών προβλέψεων.

iii. Recall: Ποσοστό ανίχνευσης θετικών περιστατικών.

iv. Recall: Ποσοστό ανίχνευσης θετικών περιστατικών.

v. F1-score: Συνδυασμός precision και recall.

vi. ROC AUC: Καμπύλη αποδοτικότητας ταξινόμησης.

vii. Kappa score: (0 = το μοντέλο προβλέπει τυχαία, 1 =το μοντέλο προβλέπει άριστα).

viii. RMSE: Μέσος όρος σφαλμάτων).

ix. RAE: Σχετικό απόλυτο σφάλμα (σύγκριση με μοντέλο baseline).

x. SSE: μετρά τη συνολική απόκλιση του μοντέλου.

Τα αποτελέσματα των 10 επαναλήψεων συγκεντρώθηκαν σε πίνακα, και υπολογίστηκαν οι μέσοι όροι ανά μοντέλο.

Πραγματοποιήθηκε απεικόνιση της μέσης ROC καμπύλης για κάθε μοντέλο, υπολογίζοντας τον μέσο όρο των TPR και FPR ανά fold. Η μέση AUC (Area Under Curve) για κάθε μοντέλο αποτέλεσε βασικό δείκτη απόδοσης, δείχνοντας την ικανότητα του μοντέλου να διακρίνει μεταξύ θετικών και αρνητικών παραδειγμάτων. Παρακάτω ακολουθούν οι σχετικοί πίνακες:

```
============================================================
10-Fold CV Evaluation with Undersampling (Without random_state -> results
slightly vary each time):
                     Accuracy  Precision  Recall     F1  ROC AUC  Kappa  \
Model
Logistic Regression     0.811      0.759   0.920  0.830    0.847  0.622
Random Forest           0.777      0.724   0.908  0.804    0.826  0.554
SVM                     0.807      0.770   0.884  0.821    0.851  0.614
XGBoost                 0.773      0.728   0.892  0.798    0.821  0.547

                      RMSE    RAE    SSE  Threshold
Model
Logistic Regression  0.399  0.311  7.932      0.367
Random Forest        0.412  0.332  8.479      0.376
SVM                  0.397  0.317  7.891      0.417
XGBoost              0.433  0.294  9.380      0.269
```

Figure 6: *Αποτελέσματα μετρικών*

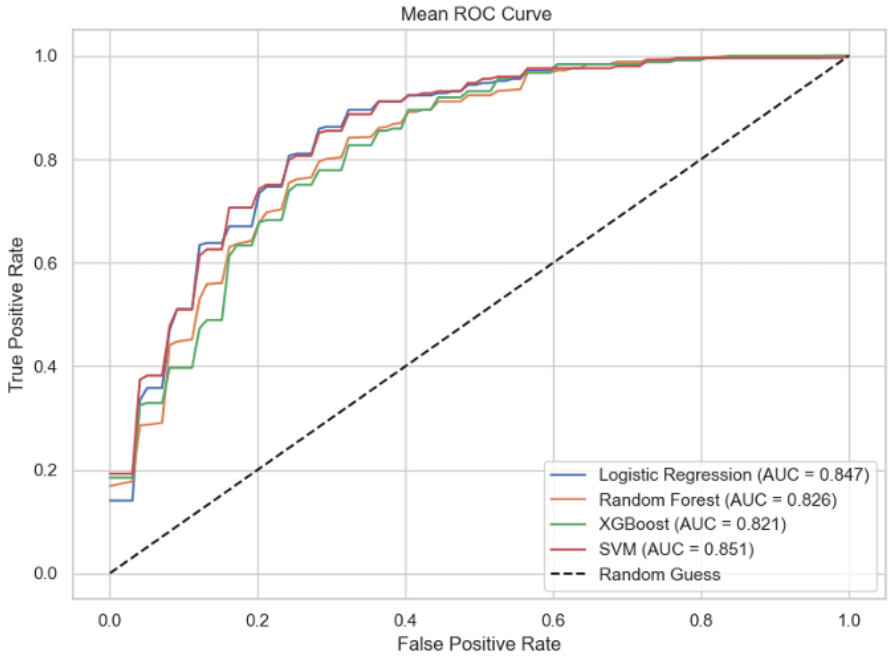

Figure 7: *Μέση καμπύλη ROC*

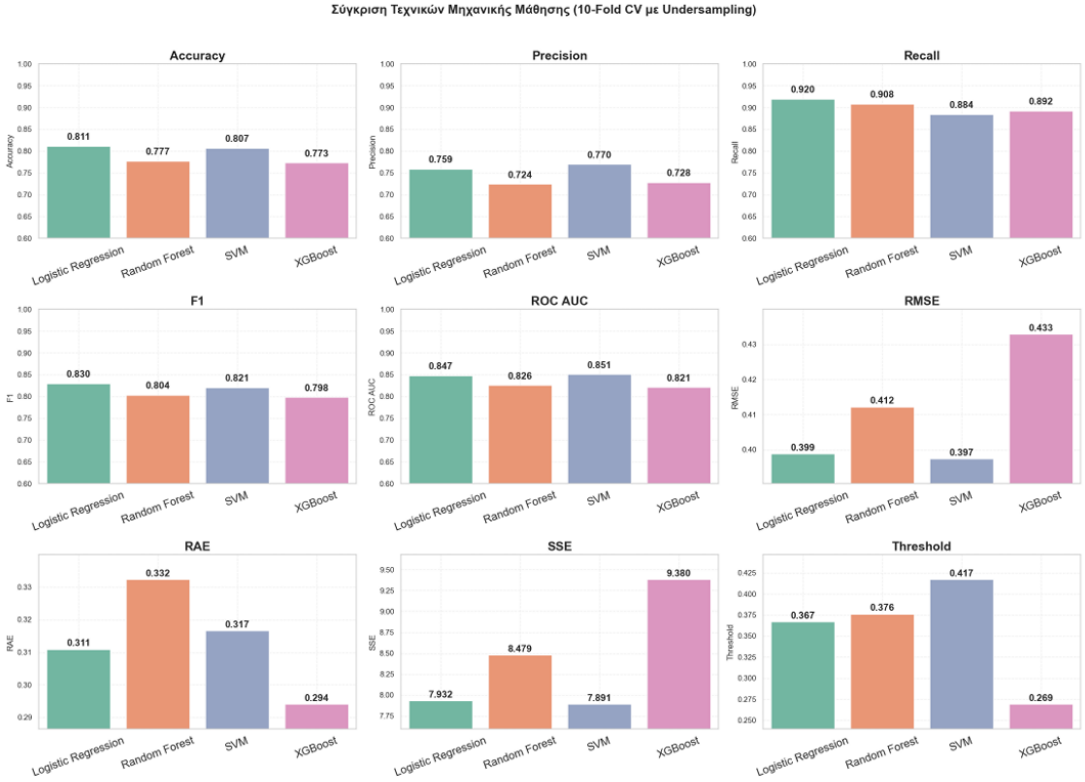

Figure 8: *Σύγκριση τεχνικών μηχανικής μάθησης*

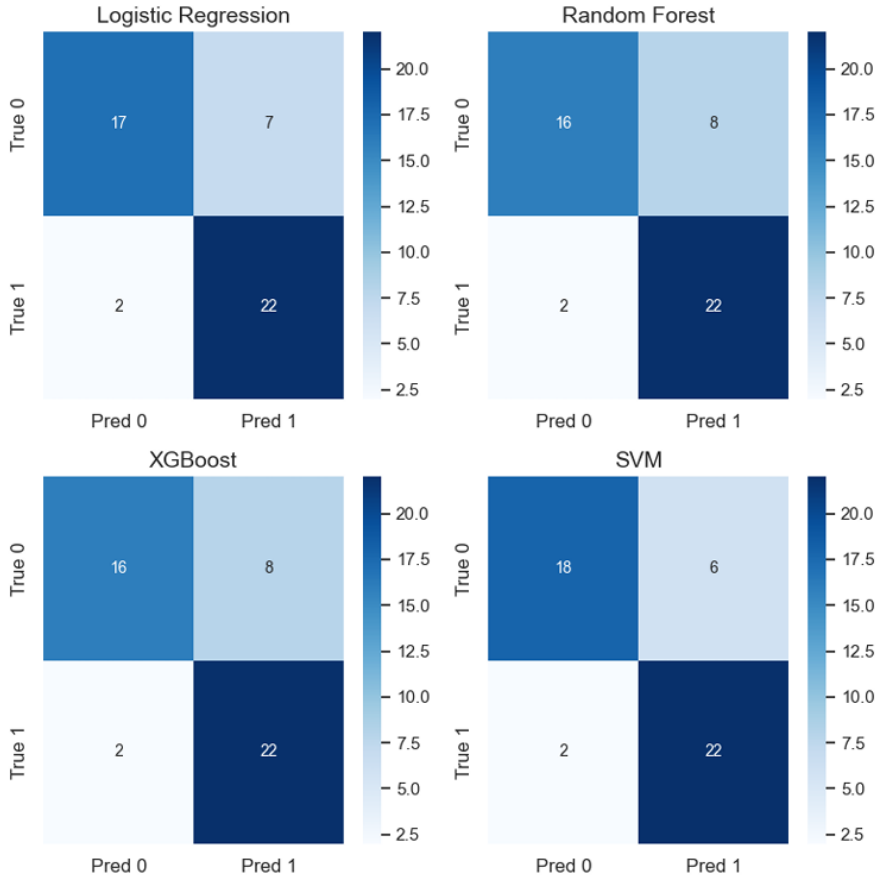

Figure 9: *Confusion Matrices*

– **Σχολιασμός Αποτελεσμάτων**

Ο Logistic Regression παρουσίασε τη συνολικά καλύτερη επίδοση, επιτυγχάνοντας την υψηλότερη ακρίβεια (0.811), recall (0.920), και F1 score (0.830), καθιστώντας τον εξαιρετική επιλογή για εφαρμογές υγειονομικού ενδιαφέροντος όπου απαιτείται υψηλή ερμηνευσιμότητα και ελαχιστοποίηση των false negative.

Ο SVM ακολούθησε στενά σε ακρίβεια (0.807) και είχε τις καλύτερες επιδόσεις σε ROC AUC (0.851), Precision (0.770), καθώς και τα χαμηλότερα σφάλματα (RMSE = 0.397, SSE = 7.891), καταδεικνύοντας την ικανότητά του να διαχωρίζει πιο αποτελεσματικά τις περιπτώσεις εγκεφαλικού από τις μη.

Οι Random Forest και XGBoost σημείωσαν επίσης ικανοποιητικά αποτελέσματα, με υψηλές τιμές recall και F1, αν και ελαφρώς χαμηλότερες από τους δύο πρώτους. Ωστόσο, ο Random Forest παραμένει αξιόπιστη επιλογή για σταθερά αποτελέσματα, ενώ ο XGBoost προσφέρει ισχυρή απόδοση σε πιο σύνθετα μοντέλα με δυνατότητες ρύθμισης παραμέτρων.

Δεν χρησημοποιήθηκε random state στον κώδικα, συνεπώς με κάθε "run" του κώδικα τα αποτελέσματα θα είναι ελαφρώς διαφοροποιημένα.

.6 Συμπεράσματα και προτάσεις για μελλοντικές βελτιώσεις

Η παρούσα μελέτη αξιολόγησε τέσσερις δημοφιλείς αλγορίθμους μηχανικής μάθησης για την πρόβλεψη περιστατικών εγκεφαλικού, χρησιμοποιώντας ένα ισορροπημένο dataset μέσω τεχνικής undersampling. Τα αποτελέσματα ανέδειξαν τον Logistic regression ως τον πιο αξιόπιστο και ερμηνεύσιμο ταξινομητή, με κορυφαίες επιδόσεις σε recall, F1 score και ακρίβεια. Ο SVM εμφάνισε την υψηλότερη ικανότητα διαχωρισμού μέσω

ROC AUC και χαμηλότερα σφάλματα πρόβλεψης, καθιστώντας το κατάλληλο για εφαρμογές με έμφαση στην ακρίβεια πιθανοτήτων.

Ωστόσο, οι Random Forest και XGBoost, αν και υστέρησαν ελαφρώς, παρουσίασαν αξιοσημείωτη σταθερότητα και δυνατότητα για εντοπισμό σύνθετων μη γραμμικών σχέσεων.

Προτάσεις για Μελλοντικές Βελτιώσεις

* Επέκταση χαρακτηριστικών (feature engineering):Η δημιουργία νέων, πιο ερμηνεύσιμων χαρακτηριστικών μπορεί να βελτιώσει την απόδοση των μοντέλων, ιδίως εκείνων με απλή δομή όπως ο logistic Regression.
* Δοκιμή με περισσότερα μοντέλα ή βελτιωμένες εκδοχές (π.χ. LightGBM, CatBoost): Προσφέρουν υψηλή απόδοση και καλύτερη διαχείριση κατηγορηματικών χαρακτηριστικών.
* Ρύθμιση υπερπαραμέτρων (hyperparameter tuning): Ειδικά για πιο σύνθετα μοντέλα (XGBoost, Random Forest), η χρήση Grid Search ή Bayesian Optimization μπορεί να οδηγήσει σε καλύτερα αποτελέσματα.
* Χρήση εξωτερικών συνόλων δεδομένων για επικύρωση (external validation): Η αξιολόγηση σε διαφορετικά datasets μπορεί να αποδείξει τη γενικευσιμότητα των μοντέλων.
* Ερμηνευσιμότητα μοντέλων (model explainability): Τεχνικές όπως SHAP ή LIME μπορούν να βοηθήσουν στη κατανόηση των αποφάσεων των μοντέλων, κάτι ιδιαίτερα σημαντικό σε ιατρικές εφαρμογές.

Bibliography

[1] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[2] Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(1), 559–563.

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

[5] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[6] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.

[7] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

[8] Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC.

[9] Kokkotis, C., Giarmatzis, G., Giannakou, E., et al. (2022). An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data.

[10] Hassan, A., Ahmad, S. G., Munir, E. U., Khan, I. A., & Ramzan, N. (2024). Predictive Modelling and Identification of Key Risk Factors for Stroke Using Machine Learning Approaches.

[11] Biswas, N., Uddin, K. M. M., Rikta, S. T., & Dey, S. K. (2022). A Comparative Analysis of Machine Learning Classifiers for Stroke Prediction.