

Διεθνές Πανεπιστήμιο της Ελλάδος

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Αξιολόγηση Τεχνικών Εκπαίδευσης Μοντέλων ML
για Πρόβλεψη Εγκεφαλικού μέσω Συνθετικών
Δεδομένων»



Του φοιτητή
Στιβακτάκη Ιωάννη
Αρ. Μητρώου: 2019158

Επιβλέπουσα
Ελβίρα–Μαρία
Αρβανίτου
Επίκουρη Καθηγήτρια

Ημερομηνία 05/09/2026

Τίτλος Δ.Ε. : Αξιολόγηση Τεχνικών Εκπαίδευσης Μοντέλων ML για πρόβλεψη Εγκεφαλικού Μέσω
Συνθετικών Δεδομένων

Κωδικός Δ.Ε.: 26180

Ονοματεπώνυμο φοιτητή: Στιβακτάκης Ιωάννης

Ονοματεπώνυμο εισηγητή: Ελβίρα-Μαρία Αρβανίτου

Ημερομηνία ανάληψης Δ.Ε. 12/03/2026

Ημερομηνία περάτωσης Δ.Ε. 05/09/2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Στιβακτάκη Ιωάννη που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας, δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση (downloading), ανάρτηση (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραίτητα και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος

Πρόλογος

Η επιλογή της συγκεκριμένης διπλωματικής εργασίας βασίστηκε στο ενδιαφέρον για τη μηχανική μάθηση και στη δυνατότητα εφαρμογής της σε πραγματικά προβλήματα του χώρου της υγείας. Η μηχανική μάθηση αποτελεί έναν τομέα με ιδιαίτερη σημασία, καθώς επιτρέπει την ανάλυση δεδομένων και την εξαγωγή χρήσιμων συμπερασμάτων. Όταν συνδυάζεται με ιατρικά προβλήματα, η αξία της γίνεται ακόμη μεγαλύτερη, καθώς μπορεί να συμβάλει στην καλύτερη κατανόηση, πρόβλεψη και υποστήριξη αποφάσεων που σχετίζονται με την υγεία.

Μέσα από την εκπόνηση της παρούσας εργασίας δόθηκε η δυνατότητα εμβάθυνσης σε βασικές έννοιες της μηχανικής μάθησης και καλύτερης κατανόησης των δυσκολιών που εμφανίζονται όταν τα δεδομένα προέρχονται από πραγματικά προβλήματα. Παράλληλα, αναδείχθηκε η σημασία της σωστής αξιολόγησης των μοντέλων, της ποιότητας των δεδομένων και της υπεύθυνης χρήσης τέτοιων τεχνικών σε εφαρμογές υγείας.

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει την πρόβλεψη εγκεφαλικού επεισοδίου με χρήση μοντέλων μηχανικής μάθησης, δίνοντας ιδιαίτερη έμφαση στο πρόβλημα της ανισορροπίας των δεδομένων. Για την πειραματική διαδικασία χρησιμοποιήθηκε το Stroke Prediction Dataset, το οποίο περιλαμβάνει 5.110 εγγραφές ασθενών και παρουσιάζει έντονη ανισορροπία, καθώς οι περιπτώσεις εγκεφαλικού αποτελούν μικρό ποσοστό του συνολικού συνόλου δεδομένων.

Στην εργασία αξιολογήθηκαν τα μοντέλα Logistic Regression, Random Forest και XGBoost σε διαφορετικά σενάρια εκπαίδευσης. Αρχικά εξετάστηκαν baseline μοντέλα και στη συνέχεια εφαρμόστηκαν τεχνικές αντιμετώπισης της ανισορροπίας, όπως class weights, SMOTE, Borderline-SMOTE, SMOTE-Tomek και παραγωγή συνθετικών δεδομένων μέσω CTGAN. Η αξιολόγηση πραγματοποιήθηκε κυρίως με μετρικές όπως Recall, F1-score και PR-AUC, καθώς η απλή ακρίβεια μπορεί να είναι παραπλανητική σε ανισόρροπα ιατρικά δεδομένα.

Τα αποτελέσματα έδειξαν ότι τα baseline μοντέλα δυσκολεύτηκαν να εντοπίσουν τη θετική κλάση, παρά το υψηλό Accuracy. Η Logistic Regression σε συνδυασμό με Borderline-SMOTE παρουσίασε την πιο ισορροπημένη συμπεριφορά στο πραγματικό test set, ενώ το CTGAN δεν κατάφερε να ξεπεράσει τις SMOTE-based τεχνικές, πιθανώς λόγω του περιορισμένου αριθμού πραγματικών θετικών δειγμάτων. Επιπλέον, εξετάστηκε προσομοίωση Federated Learning με FedAvg σε IID και Non-IID σενάρια, όπου το Local SMOTE παρουσίασε απόδοση κοντά στην αντίστοιχη centralized προσέγγιση.

Συνολικά, η εργασία αναδεικνύει τη σημασία των τεχνικών εξισορρόπησης και των κατάλληλων μετρικών αξιολόγησης στην πρόβλεψη εγκεφαλικού επεισοδίου, καθώς και τη δυνατότητα αξιοποίησης του Federated Learning σε περιβάλλοντα υγείας όπου η ιδιωτικότητα των δεδομένων είναι κρίσιμη.

Abstract

This diploma thesis examines stroke prediction using machine learning models, with particular emphasis on the problem of class imbalance. The experimental process was based on the Stroke Prediction Dataset, which contains 5.110 patient records and presents a highly imbalanced class distribution, since stroke cases represent only a small percentage of the total dataset.

In this study, three machine learning models were evaluated, Logistic Regression, Random Forest and XGBoost. Initially, baseline models were trained without any imbalance-handling technique. Afterwards, several approaches were applied, including class weights, SMOTE, Borderline-SMOTE, SMOTE-Tomek and synthetic data generation through CTGAN. The evaluation focused mainly on Recall, F1-score and PR-AUC, since Accuracy alone can be misleading in imbalanced medical datasets.

The results showed that the baseline models struggled to identify the positive class, despite achieving high Accuracy. Logistic Regression combined with Borderline-SMOTE demonstrated the most balanced performance on the real test set. In contrast, CTGAN did not outperform the SMOTE-based techniques, possibly due to the limited number of real positive samples available for training the generative model. In addition, a Federated Learning simulation was implemented using the FedAvg algorithm under IID and Non-IID scenarios. In this setting, Local SMOTE achieved performance close to the corresponding centralized SMOTE approach.

Overall, this thesis highlights the importance of imbalance-handling techniques and appropriate evaluation metrics in stroke prediction. It also shows that Federated Learning can be a useful approach in healthcare environments where data privacy is a critical requirement.

Περιεχόμενα

Πρόλογος.....	iii
Περίληψη.....	iv
Abstract	v
Περιεχόμενα	vi
Κατάλογος Σχημάτων	viii
Κατάλογος Πινάκων.....	ix
Συντομογραφίες.....	x
Κεφάλαιο 1ο: Εισαγωγή	11
1.1 Εισαγωγή στο Πρόβλημα	11
1.2 Κίνητρο της Εργασίας.....	12
1.3 Στόχοι της Διπλωματικής.....	14
1.4 Δομή της Εργασίας.....	15
Κεφάλαιο 2ο: Θεωρητικό Υπόβαθρο.....	18
2.1 Machine Learning στην Ιατρική.....	18
2.2 Imbalanced Datasets.....	21
2.3 Τεχνικές Αντιμετώπισης Imbalance	23
2.3.1 Class Weighting.....	23
2.3.2 SMOTE.....	23
2.3.3 Borderline-SMOTE	24
2.3.4 SMOTE-Tomek.....	24
2.4 Generative Synthetic Data	25
2.4.1 Synthetic Data Generation.....	25
2.4.2 Generative Adversarial Networks (GANs).....	25
2.4.3 CTGAN	26
Κεφάλαιο 3ο: Περιγραφή Δεδομένων και Προεπεξεργασία	27
3.1 Εισαγωγή.....	27
3.2 Περιγραφή Dataset	27
3.3 Εξερευνητική Ανάλυση (EDA).....	30
3.4 Προεπεξεργασία	33
3.5 Μετρικές Αξιολόγησης	35
Κεφάλαιο 4ο: Μεθοδολογία & Πειράματα.....	39
4.1 Περιγραφή Μοντέλων	39
4.2 Πειράματα με Τεχνικές Εξισορρόπησης	41

4.2.1	Baseline αποτελέσματα	43
4.2.2	Χρήση Class Weights	43
4.2.3	Εφαρμογή SMOTE.....	44
4.2.4	Εφαρμογή Borderline-SMOTE	45
4.2.5	Εφαρμογή SMOTE-Tomek	45
4.2.6	Παραγωγή συνθετικών δεδομένων μέσω CTGAN	46
4.2.7	Συνολική πειραματική διαδικασία.....	47
4.3	Ανάλυση Συμπεριφοράς Μοντέλων.....	48
4.3.1	Συμπεριφορά της Logistic Regression.....	48
4.3.2	Συμπεριφορά του Random Forest.....	53
4.3.3	Συμπεριφορά του XGBoost.....	55
4.3.4	Συνολική σύγκριση και ερμηνεία	58
4.4	Πειραματική Ανάλυση με Εξισορρόπηση του Test Set	60
Κεφάλαιο 5ο:	Federated Learning	66
5.1	Σχεδιασμός Προσομοίωσης.....	66
5.2	Σενάρια Κατανομής.....	67
5.3	Υλοποίηση FedAvg.....	69
5.4	Federated Learning με Τοπική Εξισορρόπηση.....	71
Κεφάλαιο 6ο:	Συγκριτική Αξιολόγηση και Συζήτηση.....	76
6.1	Συνολικά Αποτελέσματα.....	76
6.2	SMOTE vs CTGAN	80
6.3	Ερμηνεία Αποτελεσμάτων	82
Κεφάλαιο 7ο:	Συμπεράσματα και Μελλοντική Έρευνα	86
7.1	Σύνοψη Ευρημάτων	86
7.2	Περιορισμοί της Έρευνας.....	87
7.3	Προτάσεις για Μελλοντική Επέκταση	90
BIBΛΙΟΓΡΑΦΙΑ.....		93

Κατάλογος Σχημάτων

Σχήμα 1.1: Είδη εγκεφαλικού [1].....	11
Σχήμα 3.1 Δεδομένα stroke – no stroke.	29
Σχήμα 3.2: Σύγκριση μεταβλητών	32
Σχήμα 3.3: Mean – Median BMI.....	34
Σχήμα 4.1: Απόδοση SMOTE Ratios.....	50
Σχήμα 4.2: FP-FN των SMOTE Ratios.....	51
Σχήμα 4.3: Σύγκριση αποτελεσμάτων της Logistic Regression.....	52
Σχήμα 4.4: Confusion Matrix Random Forest Boredeline SMOTE	54
Σχήμα 4.5: Random Forest σύγκριση τεχνικών	55
Σχήμα 4.6: XGBoost σύγκριση τεχνικών.....	57
Σχήμα 4.7: Real test set – SMOTE test set Logistic Regression.....	61
Σχήμα 4.8: Real test set – SMOTE test set Random Forest	62
Σχήμα 4.9: Real test set – SMOTE test set XGBoost.....	63
Σχήμα 6.1: Αποτελέσματα Federated Learning στα IID σενάρια.	78
Σχήμα 6.2: Αποτελέσματα Federated Learning Non IID	79
Σχήμα 6.3: Σύγκριση SMOTE Centarilezed – Federated Learning	79

Κατάλογος Πινάκων

Πίνακας 2.1: Σύνοψη σχετικών εργασιών και σύγκριση με την παρούσα ΔΕ.....	21
Πίνακας 3.1: Μεταβλητές του dataset.....	28
Πίνακας 3.2: Γενική μορφή του πίνακα σύγχυσης.....	37
Πίνακας 4.1: Σύνοψη πειραματικών σεναρίων και τεχνικών εξισορρόπησης.....	42
Πίνακας 4.2: Αριθμητικά δεδομένα Logistic Regression Baseline.....	49
Πίνακας 4.3: Confusion Matrix – Centralized LR Baseline.....	49
Πίνακας 4.4: Αριθμητικά δεδομένα Logistic Regression Class_Weight.....	49
Πίνακας 4.5: Αριθμητικά δεδομένα Logisitic Regression SMOTE.....	51
Πίνακας 4.6: Αριθμητικά δεδομένα Logistic Regression CTGAN.....	52
Πίνακας 4.7: Σύγκριση τεχνικών εξισορρόπησης για τη Logistic Regression.....	53
Πίνακας 4.8: Σύγκριση τεχνικών εξισορρόπησης για το Random Forest.....	55
Πίνακας 4.9: Σύγκριση τεχνικών εξισορρόπησης για το XGBoost.....	57
Πίνακας 4.10: Σύγκριση βέλτιστων επιδόσεων ανά μοντέλο (Real Test Set).....	59
Πίνακας 4.11: Σύγκριση Real Test Set – SMOTE Test Set.....	65
Πίνακας 5.1: Κατανομή δειγμάτων ανά client στο Non-IID σενάριο.....	69
Πίνακας 5.2: Συγκεντρωτική παρουσίαση των σεναρίων Federated Learning.....	74

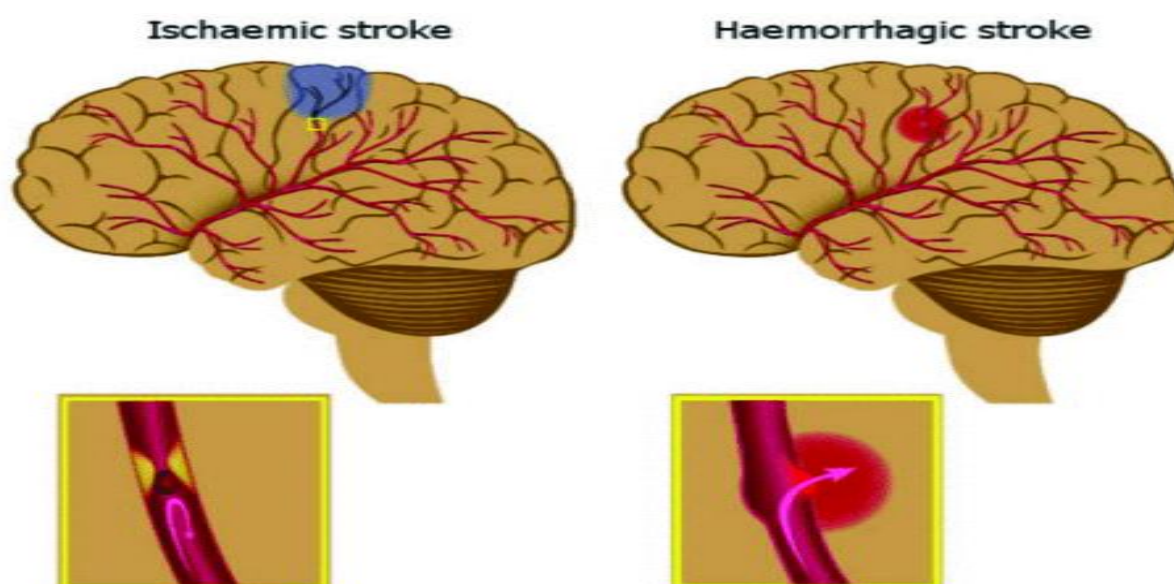
Συντομογραφίες

Δ.Ε.	Διπλωματική Εργασία
ΔΙΠΙΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
SVM	Support Vector Machines
SDG	Synthetic Data Generation
PR-AUC	Καμπύλη precision-recall
SMOTE	Synthetic Minority Over-sampling Technique
GANs	Generative Adversarial Networks
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
LR	Logistic Regression
RF	Random Forest
FedAvg	Federated Averaging
BMI	Body Mass Index
SVM	Support Vector Machine
KNN	K-Nearest Neighbors

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εισαγωγή στο Πρόβλημα

Το εγκεφαλικό επεισόδιο αποτελεί μία από τις πιο σημαντικές προκλήσεις για τη δημόσια υγεία σε παγκόσμιο επίπεδο, προκαλώντας σημαντικές επιπτώσεις τόσο στα ποσοστά επιβίωσης των ασθενών όσο και στην αποκατάσταση και τη μετέπειτα ποιότητα ζωής τους. Δεν πρόκειται απλώς για ένα ξαφνικό ιατρικό περιστατικό, αλλά για μία κατάσταση που μπορεί να επηρεάσει καθοριστικά τη λειτουργικότητα, την αυτονομία και την καθημερινότητα του ατόμου. Σύμφωνα με διεθνείς οργανισμούς υγείας, καταγράφονται περίπου δώδεκα εκατομμύρια νέα περιστατικά ετησίως, ενώ ένα σημαντικό ποσοστό των επιζώντων αντιμετωπίζει μόνιμες νευρολογικές βλάβες και λειτουργικούς περιορισμούς [1]. Αξίζει να σημειωθεί, ότι περισσότερα από εκατό εκατομμύρια άνθρωποι σε ολόκληρο τον κόσμο έχουν υποστεί αναπηρίες λόγω εγκεφαλικού επεισοδίου [2]. Πιο συγκεκριμένα, το εγκεφαλικό επεισόδιο το 2021 ήταν η τρίτη κύρια αιτία θανάτου και αναπηρίας παγκοσμίως, με περίπου 93,8 εκατομμύρια υπάρχουσες περιπτώσεις και 11,9 εκατομμύρια νέα περιστατικά μέσα στο ίδιο έτος σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας [3]. Επιπλέον, ο κίνδυνος εμφάνισης εγκεφαλικού κατά τη διάρκεια της ζωής του ανθρώπου έχει αυξηθεί σημαντικά τις τελευταίες δεκαετίες με εκτιμήσεις να αναφέρουν ότι τα ποσοστά των ανθρώπων, που θα υποστούν ένα τέτοιου είδους επεισόδιο, να αγγίζουν το 25% [4]. Είναι σημαντικό να κατανοήσουμε ότι, ο έγκαιρος εντοπισμός και η πρόληψή του εγκεφαλικού, αποτελούν πρωταρχικό στόχο για τα σύγχρονα συστήματα υγείας. Η σημασία αυτή δεν οφείλεται μόνο στο γεγονός ότι το εγκεφαλικό συγκαταλέγεται μεταξύ των κυριότερων αιτιών θνησιμότητας και αναπηρίας παγκοσμίως, αλλά και στο ότι η σοβαρότητα του προβλήματος δεν περιορίζεται αποκλειστικά στον αριθμό των περιστατικών. Επιπροσθέτως, είναι ιδιαίτερα κρίσιμο καθώς συχνά εμφανίζεται ξαφνικά και απαιτεί άμεση ιατρική αντιμετώπιση. Παράλληλα, είναι αρκετές οι περιπτώσεις όπου το εγκεφαλικό οδηγεί σε μόνιμες βλάβες, ανάγκη για μακροχρόνια αποκατάσταση, εξάρτηση από τρίτα άτομα και αυξημένο κόστος για τα συστήματα υγείας.



Σχήμα 1.1: Είδη εγκεφαλικού [1]

Το εγκεφαλικό επεισόδιο σχετίζεται με τη διακοπή της αιματικής ροής προς τον εγκέφαλο, οδηγώντας σε ανεπαρκή παροχή οξυγόνου και ζωτικών θρεπτικών συστατικών στα εγκεφαλικά κύτταρα του

ασθενή. Ως αποτέλεσμα, η κατάσταση αυτή προκαλεί μη αναστρέψιμες βλάβες, συχνά μέσα σε πολύ σύντομο χρονικό διάστημα.

Όπως φαίνεται στο σχήμα 1.1, το εγκεφαλικό επεισόδιο μπορεί γενικά να διακριθεί σε δύο βασικές κατηγορίες οι οποίες είναι το ισχαιμικό και το αιμορραγικό. Το ισχαιμικό εγκεφαλικό εμφανίζεται όταν η αιματική ροή προς ένα τμήμα του εγκεφάλου διακόπτεται ή μειώνεται, συνήθως λόγω απόφραξης αγγείου από θρόμβο. Ως αποτέλεσμα, τα εγκεφαλικά κύτταρα δεν λαμβάνουν την απαραίτητη ποσότητα οξυγόνου και θρεπτικών συστατικών, γεγονός που μπορεί να οδηγήσει σε σοβαρή βλάβη. Αντίθετα, το αιμορραγικό εγκεφαλικό προκαλείται όταν ένα αιμοφόρο αγγείο στον εγκέφαλο υποστεί ρήξη και υπάρξει αιμορραγία. Και στις δύο περιπτώσεις, το εγκεφαλικό αποτελεί επείγουσα κατάσταση, καθώς η έγκαιρη αναγνώριση και αντιμετώπιση είναι καθοριστική για τον περιορισμό της βλάβης και την καλύτερη αποκατάσταση του ασθενούς [2], [4]. Συνεπώς, η άμεση ιατρική παρέμβαση είναι καθοριστικής σημασίας, καθώς ακόμη και μικρές καθυστερήσεις στη διάγνωση αλλά και στη φροντίδα των ασθενών, μπορούν να επηρεάσουν αρνητικά την έκβαση της υγείας τους. Αξίζει να σημειωθεί πως όσα επεισόδια αντιμετωπίστηκαν σε μικρότερο χρονικό διάστημα των τριών ωρών, οι ασθενείς είτε ανάρρωσαν πλήρως είτε χρειάστηκαν επιπλέον περίθαλψη, χωρίς όμως να τους προκληθεί χρόνια αναπηρία [6]. Επιπλέον, η εμφάνιση εγκεφαλικού επεισοδίου συνδέεται με ένα σύνολο παραγόντων κινδύνου, οι οποίοι διακρίνονται σε τροποποιήσιμους και μη τροποποιήσιμους. Στην πρώτη κατηγορία, περιλαμβάνονται η υπέρταση, ο σακχαρώδης διαβήτης, το κάπνισμα, η παχυσαρκία και ο καθιστικός τρόπος ζωής, ενώ στη δεύτερη ανήκουν η ηλικία, το φύλο και η γενετική προδιάθεση. Η κατανόηση και η ανάλυση των παραγόντων αυτών αποτελούν κρίσιμο βήμα για την ανάπτυξη αξιόπιστων μοντέλων μηχανικής μάθησης με στόχο την πρόβλεψη ασθενειών. Η ταχεία εξέλιξη της τεχνητής νοημοσύνης και ιδιαίτερα της μηχανικής μάθησης έχουν δημιουργήσει νέες δυνατότητες στην ανάλυση ιατρικών δεδομένων προσφέροντας έτσι, πληθώρα πλεονεκτήματα στην ανίχνευση και την αντιμετώπιση ασθενειών όπως το εγκεφαλικό επεισόδιο. Πιο συγκεκριμένα, οι αλγόριθμοι μηχανικής μάθησης έχουν τη δυνατότητα να επεξεργάζονται μεγάλους όγκους μη γραμμικών δεδομένων και να εντοπίζουν σύνθετα μοτίβα που δεν είναι εύκολα ανιχνεύσιμα μέσω παραδοσιακών στατιστικών μεθόδων. Στον τομέα της πρόβλεψης εγκεφαλικού επεισοδίου, οι τεχνικές αυτές επιτρέπουν την εκτίμηση του κινδύνου εμφάνισης της νόσου σε ατομικό επίπεδο, ενισχύοντας τις κλινικές διαδικασίες λήψης αποφάσεων.

Η σοβαρότητα του εγκεφαλικού επεισοδίου δεν περιορίζεται μόνο στο ίδιο το περιστατικό, αλλά επεκτείνεται και στις μακροχρόνιες συνέπειες που μπορεί να προκαλέσει στη ζωή του ασθενούς. Πολλοί ασθενείς μετά από ένα εγκεφαλικό επεισόδιο αντιμετωπίζουν δυσκολίες στην κίνηση, στην ομιλία, στη μνήμη ή στην αυτοεξυπηρέτηση, γεγονός που επηρεάζει σημαντικά την καθημερινότητά τους. Παράλληλα, η ανάγκη για αποκατάσταση, ιατρική παρακολούθηση και υποστήριξη από τρίτα άτομα δημιουργεί σημαντική επιβάρυνση τόσο για τις οικογένειες όσο και για τα συστήματα υγείας. Για τον λόγο αυτό, η πρόληψη και η έγκαιρη αναγνώριση ατόμων που παρουσιάζουν αυξημένο κίνδυνο εμφάνισης εγκεφαλικού αποτελούν ιδιαίτερα σημαντικούς στόχους. Η αξιοποίηση διαθέσιμων δημογραφικών, κλινικών και συμπεριφορικών χαρακτηριστικών μπορεί να συμβάλει στην καλύτερη κατανόηση των παραγόντων που σχετίζονται με τη νόσο. Σε αυτό το πλαίσιο, οι υπολογιστικές μέθοδοι και ειδικότερα η μηχανική μάθηση μπορούν να χρησιμοποιηθούν ως υποστηρικτικά εργαλεία για την ανάλυση τέτοιων δεδομένων και την ανάπτυξη προγνωστικών μοντέλων.

1.2 Κίνητρο της Εργασίας

Το κίνητρο της παρούσας διπλωματικής εργασίας προκύπτει από την ανάγκη αξιοποίησης σύγχρονων τεχνικών μηχανικής μάθησης σε ένα πρόβλημα με ιδιαίτερη ιατρική και κοινωνική σημασία, όπως είναι η πρόβλεψη του εγκεφαλικού επεισοδίου. Όπως αναφέρθηκε και στην προηγούμενη ενότητα, το

εγκεφαλικό αποτελεί μία από τις βασικότερες αιτίες θανάτου και αναπηρίας παγκοσμίως, ενώ η έγκαιρη ανίχνευση ατόμων που παρουσιάζουν αυξημένο κίνδυνο μπορεί να συμβάλει ουσιαστικά στην πρόληψη και στην καλύτερη διαχείριση της νόσου. Στο πλαίσιο αυτό, η χρήση υπολογιστικών μεθόδων για την ανάλυση ιατρικών δεδομένων παρουσιάζει ιδιαίτερο ενδιαφέρον, καθώς μπορεί να προσφέρει επιπλέον εργαλεία υποστήριξης στην ιατρική αξιολόγηση και στη λήψη αποφάσεων.

Η επιλογή του συγκεκριμένου θέματος δεν βασίζεται μόνο στη σημασία του εγκεφαλικού ως ιατρικού προβλήματος, αλλά και στις τεχνικές προκλήσεις που παρουσιάζει η πρόβλεψή του μέσω δεδομένων. Παρά τη σημαντική πρόοδο που έχει σημειωθεί τα τελευταία χρόνια στην εφαρμογή τεχνικών μηχανικής μάθησης για την πρόγνωση ασθενειών, η αξιόπιστη πρόβλεψη της εμφάνισης εγκεφαλικού επεισοδίου εξακολουθεί να αποτελεί ένα ιδιαίτερα απαιτητικό ερευνητικό πρόβλημα. Ένας από τους βασικότερους παράγοντες που περιορίζουν την αποτελεσματικότητα των προγνωστικών μοντέλων είναι ότι, σε αντίθεση με πιο απλά προβλήματα ταξινόμησης, όπου οι διαθέσιμες κατηγορίες μπορεί να είναι σχετικά ισορροπημένες, η φύση των διαθέσιμων ιατρικών δεδομένων, τα οποία χαρακτηρίζονται συχνά από έντονη ανισορροπία μεταξύ των κλάσεων καθώς και η έλλειψη στη ποσότητα των δεδομένων λόγω των ευαίσθητων προσωπικών δεδομένων των ασθενών. Στον τομέα του εγκεφαλικού επεισοδίου, τα θετικά περιστατικά εμφανίζονται σε σημαντικά μικρότερο ποσοστό σε σχέση με τα αρνητικά, γεγονός που δυσκολεύει την σωστή εκπαίδευση μοντέλων με ικανοποιητική ικανότητα γενίκευσης, με συνέπεια τα αποτελέσματα των μετρήσεων τους να μην είναι ικανοποιητικά.

Στην πλειονότητα της υπάρχουσας βιβλιογραφίας, η αξιολόγηση των μοντέλων βασίζεται κατά κύριο λόγο σε μετρικές όπως η συνολική ακρίβεια (accuracy) [1], [3], [4]. Ωστόσο, η ανισορροπία που υπάρχει στα ιατρικά δεδομένα δημιουργεί ένα σημαντικό πρόβλημα στην εκπαίδευση των μοντέλων μηχανικής μάθησης με αποτέλεσμα το μοντέλο να εμφανίζει υψηλό ποσοστό συνολικής ακρίβειας προβλέποντας σχεδόν πάντα ότι ένας ασθενής δεν θα εμφανίσει εγκεφαλικό επεισόδιο. Πρέπει να γίνει κατανοητό ότι, παρόλο που μια τέτοια συμπεριφορά μπορεί να φαίνεται ικανοποιητική με βάση το Accuracy, στην πραγματικότητα είναι προβληματική καθώς οι μετρικές αυτές δεν αποτυπώνουν επαρκώς την πραγματική απόδοση σε περιβάλλοντα με ανισορροπία δεδομένων καθώς σπάνια ανιχνεύουν την ύπαρξη προδιάθεσης της νόσου. Το σημείο αυτό, αποτελεί ένα από τα πιο κρίσιμα σημεία της εργασίας, καθώς δείχνει ότι η υψηλή αριθμητική απόδοση ενός μοντέλου δεν σημαίνει απαραίτητα ότι το μοντέλο είναι πραγματικά χρήσιμο. Ειδικά σε προβλήματα υγείας, η αποτυχία εντοπισμού ενός πιθανού περιστατικού μπορεί να είναι πολύ πιο σημαντική από ένα απλό στατιστικό λάθος. Τα παραπάνω αποτελούν γεγονός που αναδεικνύει την ανάγκη για πιο κατάλληλες προσεγγίσεις αξιολόγησης και βελτιστοποίησης των μοντέλων, με έμφαση σε μετρικές όπως η ανάκληση (recall), η F1-score και η καμπύλη precision-recall.

Γι' αυτό, η παρούσα εργασία δεν εστιάζει απλώς στην εκπαίδευση ενός μοντέλου πρόβλεψης, αλλά στη συστηματική αξιολόγηση τεχνικών που μπορούν να βελτιώσουν τη συμπεριφορά των μοντέλων απέναντι στη μειοψηφική κλάση. Το ενδιαφέρον επικεντρώνεται σε μεθόδους που επιχειρούν να αντιμετωπίσουν την ανισορροπία των δεδομένων, είτε τροποποιώντας τον τρόπο με τον οποίο το μοντέλο μαθαίνει, είτε μεταβάλλοντας την κατανομή των δεδομένων εκπαίδευσης. Στην πρώτη περίπτωση ανήκει η χρήση βαρών κλάσεων, όπου δίνεται μεγαλύτερη σημασία στα δείγματα της μειοψηφικής κλάσης κατά την εκπαίδευση. Στη δεύτερη περίπτωση, ανήκουν τεχνικές υπερδειγματοληψίας, όπως το SMOTE και οι παραλλαγές του, οι οποίες δημιουργούν επιπλέον συνθετικά δείγματα για την κλάση που εμφανίζεται σπανιότερα. Οι προσεγγίσεις αυτές επιδιώκουν την εξισορρόπηση της κατανομής των δεδομένων μέσω της ενίσχυσης της μειοψηφικής κλάσης κατά τη φάση της εκπαίδευσης.

Επιπλέον, ιδιαίτερο κίνητρο για την εργασία αποτέλεσε η διερεύνηση πιο σύγχρονων μεθόδων δημιουργίας συνθετικών δεδομένων, όπως το CTGAN. Η παραγωγή συνθετικών δεδομένων έχει αποκτήσει σημαντικό ενδιαφέρον τα τελευταία χρόνια, ειδικά στον χώρο της υγείας, όπου τα πραγματικά δεδομένα είναι συχνά περιορισμένα, ευαίσθητα ή δύσκολα προσβάσιμα. Σε αντίθεση με απλές τεχνικές υπερδειγματοληψίας, τα γενετικά μοντέλα επιδιώκουν να μάθουν την κατανομή των πραγματικών δεδομένων και να δημιουργήσουν νέα δείγματα που μοιάζουν με αυτά. Στο πλαίσιο της παρούσας εργασίας, το CTGAN χρησιμοποιείται με στόχο την παραγωγή επιπλέον δειγμάτων για τη μειοψηφική κλάση του εγκεφαλικού, ώστε να εξεταστεί κατά πόσο τα συνθετικά δεδομένα μπορούν να βοηθήσουν τα μοντέλα να αναγνωρίζουν καλύτερα τα πραγματικά περιστατικά. Πρέπει να σημειωθεί όμως ότι, η χρήση τέτοιου είδους μεθόδων συνοδεύεται συνήθως από σημαντικές προκλήσεις όπως η ποιότητα των παραγόμενων δεδομένων και η επίδρασή τους στην απόδοση των μοντέλων. Η ποιότητα τους δεν είναι πάντα προβλέψιμη και εξαρτάται από παράγοντες όπως η αναλογία των συνθετικών δειγμάτων και η πολυπλοκότητα του συνόλου δεδομένων.

Ένα ακόμη σημαντικό κίνητρο σχετίζεται με το ζήτημα της ιδιωτικότητας των ιατρικών δεδομένων. Τα δεδομένα υγείας θεωρούνται ιδιαίτερα ευαίσθητα, καθώς περιλαμβάνουν πληροφορίες που αφορούν την κατάσταση, το ιστορικό και τα χαρακτηριστικά των ασθενών. Συνεπώς, η συλλογή, η διαμοίραση και η κεντρική αποθήκευσή τους συνοδεύονται από αυστηρούς περιορισμούς και σημαντικές ηθικές και τεχνικές απαιτήσεις. Η ανάγκη αξιοποίησης τέτοιων δεδομένων για την εκπαίδευση μοντέλων έρχεται συχνά σε σύγκρουση με την ανάγκη προστασίας της ιδιωτικότητας. Αυτό καθιστά απαραίτητη τη διερεύνηση προσεγγίσεων που μπορούν να αξιοποιήσουν τη γνώση που υπάρχει στα δεδομένα, χωρίς να απαιτούν απαραίτητα τη συγκέντρωσή τους σε ένα ενιαίο κεντρικό σημείο. Σε αυτό το πλαίσιο, το Federated Learning παρουσιάζει ιδιαίτερο ενδιαφέρον, καθώς επιτρέπει την εκπαίδευση μοντέλων σε κατανεμημένα δεδομένα, όπως για παράδειγμα δεδομένα που βρίσκονται σε διαφορετικά νοσοκομεία ή φορείς υγείας, χωρίς να απαιτείται η άμεση μεταφορά των δεδομένων αυτών σε έναν κεντρικό server. Με αυτόν τον τρόπο, η εκπαίδευση μπορεί να πραγματοποιείται τοπικά, ενώ στον κεντρικό μηχανισμό αποστέλλονται μόνο οι ενημερώσεις του μοντέλου. Η λογική αυτή είναι ιδιαίτερα σημαντική για ιατρικές εφαρμογές, όπου η προστασία των προσωπικών δεδομένων αποτελεί βασική προϋπόθεση. Αυτό σημαίνει ότι, το κίνητρο της εργασίας επεκτείνεται πέρα από την απλή βελτίωση των μετρικών απόδοσης και αγγίζει και το ζήτημα της ασφαλέστερης και πιο ρεαλιστικής αξιοποίησης ιατρικών δεδομένων.

Με βάση όσα αναφέρθηκαν παραπάνω, γίνεται σαφές ότι το βασικό κίνητρο της παρούσας διπλωματικής εργασίας είναι να εξεταστεί κατά πόσο διαφορετικές τεχνικές εκπαίδευσης και εξισορρόπησης μπορούν να βελτιώσουν την πρόβλεψη του εγκεφαλικού επεισοδίου σε ένα περιβάλλον όπου τα δεδομένα είναι περιορισμένα, ανισόρροπα και ευαίσθητα. Η εργασία δεν αντιμετωπίζει τη μηχανική μάθηση ως μία απλή διαδικασία παραγωγής προβλέψεων, αλλά ως ένα εργαλείο που πρέπει να αξιολογείται προσεκτικά, ειδικά όταν εφαρμόζεται σε προβλήματα υγείας.

1.3 Στόχοι της Διπλωματικής

Ο βασικός στόχος της παρούσας διπλωματικής εργασίας είναι η *αξιολόγηση διαφορετικών τεχνικών εκπαίδευσης μοντέλων μηχανικής μάθησης για την πρόβλεψη του εγκεφαλικού επεισοδίου*, με ιδιαίτερη έμφαση στη χρήση τεχνικών εξισορρόπησης και συνθετικών δεδομένων. Το πρόβλημα της πρόβλεψης εγκεφαλικού επεισοδίου προσεγγίζεται ως ένα πρόβλημα δυαδικής ταξινόμησης, όπου το μοντέλο καλείται να προβλέψει αν ένας ασθενής ανήκει στην κλάση εμφάνισης εγκεφαλικού επεισοδίου ή όχι, με βάση δημογραφικά, κλινικά και συμπεριφορικά χαρακτηριστικά. Ωστόσο, όπως αναφέρθηκε και στις προηγούμενες ενότητες, η συγκεκριμένη διαδικασία δεν είναι απλή, καθώς τα ιατρικά δεδομένα

που αφορούν σπάνια περιστατικά παρουσιάζουν συνήθως έντονη ανισορροπία μεταξύ των κλάσεων. Ένας ακόμη στόχος της διπλωματικής είναι η *σύγκριση της συμπεριφοράς διαφορετικών αλγορίθμων μηχανικής μάθησης κάτω από τις ίδιες τεχνικές εξισορρόπησης*. Συγκεκριμένα, εξετάζονται μοντέλα όπως η Logistic Regression, το Random Forest και το XGBoost. Η επιλογή αυτών των μοντέλων επιτρέπει τη σύγκριση διαφορετικών προσεγγίσεων ταξινόμησης, από ένα πιο απλό και ερμηνεύσιμο γραμμικό μοντέλο, όπως η Logistic Regression, έως πιο σύνθετα μοντέλα. Με τον τρόπο αυτό, η εργασία δεν περιορίζεται στην αξιολόγηση μίας μόνο μεθόδου, αλλά επιχειρεί να εξετάσει συνολικά πώς διαφορετικοί αλγόριθμοι ανταποκρίνονται όταν εκπαιδεύονται με πραγματικά, εξισορροπημένα ή συνθετικά δεδομένα. Επίσης, σημαντικό στοιχείο της εργασίας αποτελεί και η επιλογή των κατάλληλων μετρικών αξιολόγησης. Δεδομένου ότι το Accuracy μπορεί να είναι παραπλανητικό σε ανισόρροπα σύνολα δεδομένων, η αξιολόγηση των πειραμάτων δεν βασίζεται αποκλειστικά στη συνολική ακρίβεια. Αντίθετα, δίνεται ιδιαίτερη έμφαση σε μετρικές όπως η ανάκληση, το F1-score και το PR-AUC. Τέλος, η εργασία έχει ως στόχο τη διερεύνηση της εκπαίδευσης μοντέλων σε διαφορετικά υπολογιστικά σενάρια. Πέρα από την κεντριοποιημένη εκπαίδευση, όπου όλα τα δεδομένα θεωρείται ότι συγκεντρώνονται σε ένα ενιαίο σημείο, εξετάζεται και η προσέγγιση του Federated Learning. Στο πλαίσιο της εργασίας, το Federated Learning χρησιμοποιείται ως προσομοίωση ενός περιβάλλοντος όπου διαφορετικοί φορείς, όπως νοσοκομεία ή κλινικές μονάδες, διαθέτουν διαφορετικά τμήματα των δεδομένων και συμμετέχουν στη διαδικασία εκπαίδευσης μέσω ανταλλαγής ενημερώσεων του μοντέλου.

Συμπερασματικά, ο στόχος της διπλωματικής δεν είναι να παρουσιαστεί απλώς ένα μοντέλο πρόβλεψης εγκεφαλικού επεισοδίου, αλλά να πραγματοποιηθεί μία συγκριτική και συστηματική αξιολόγηση τεχνικών που σχετίζονται με την εκπαίδευση μοντέλων σε δύσκολα και ρεαλιστικά ιατρικά δεδομένα. Η εργασία εξετάζει πώς συμπεριφέρονται διαφορετικά μοντέλα όταν εκπαιδεύονται με αρχικά ανισόρροπα δεδομένα, με τεχνικές εξισορρόπησης, με συνθετικά δεδομένα και σε καταναμημένα σενάρια εκπαίδευσης.

1.4 Δομή της Εργασίας

Η παρούσα διπλωματική εργασία οργανώνεται σε επτά κεφάλαια, με στόχο να παρουσιαστεί με συστηματικό τρόπο τόσο το θεωρητικό υπόβαθρο όσο και η πειραματική διαδικασία που ακολουθήθηκε για την αξιολόγηση τεχνικών μηχανικής μάθησης στην πρόβλεψη του εγκεφαλικού επεισοδίου. Η δομή της εργασίας έχει σχεδιαστεί έτσι ώστε ο αναγνώστης να μπορεί αρχικά να κατανοήσει το ιατρικό και τεχνικό πρόβλημα, στη συνέχεια να εξοικειωθεί με τις βασικές έννοιες και μεθόδους που χρησιμοποιούνται, και τέλος να παρακολουθήσει την πειραματική αξιολόγηση και τη συζήτηση των αποτελεσμάτων.

Στο Κεφάλαιο 1, παρουσιάζεται η εισαγωγή της εργασίας. Αρχικά γίνεται αναφορά στο εγκεφαλικό επεισόδιο ως ένα κρίσιμο ζήτημα δημόσιας υγείας, με σημαντικές επιπτώσεις στη θνησιμότητα, την αναπηρία και την ποιότητα ζωής των ασθενών. Στη συνέχεια, αναλύεται το κίνητρο της διπλωματικής, το οποίο σχετίζεται με την ανάγκη έγκαιρης πρόβλεψης του εγκεφαλικού, αλλά και με τις προκλήσεις που εμφανίζονται κατά την αξιοποίηση ιατρικών δεδομένων, όπως η ανισορροπία των κλάσεων και η προστασία της ιδιωτικότητας. Επιπλέον, καθορίζεται ο βασικός στόχος της εργασίας, ο οποίος είναι η συγκριτική αξιολόγηση τεχνικών εξισορρόπησης, παραγωγής συνθετικών δεδομένων και διαφορετικών μεθόδων εκπαίδευσης μοντέλων μηχανικής μάθησης.

Στο Κεφάλαιο 2, παρουσιάζεται η βιβλιογραφική ανασκόπηση και το θεωρητικό υπόβαθρο της εργασίας. Το κεφάλαιο αυτό περιλαμβάνει αρχικά μία σύντομη αναφορά στη χρήση της μηχανικής

μάθησης στον χώρο της υγείας και στις δυνατότητες που προσφέρει για την ανάλυση ιατρικών δεδομένων. Στη συνέχεια, αναλύεται το πρόβλημα της ανισορροπίας κλάσεων, το οποίο αποτελεί ένα από τα βασικά ζητήματα της παρούσας εργασίας, καθώς επηρεάζει άμεσα την αξιοπιστία των μοντέλων πρόβλεψης. Ιδιαίτερη έμφαση δίνεται στο γιατί η συνολική ακρίβεια μπορεί να είναι παραπλανητική σε τέτοια προβλήματα και γιατί απαιτείται η χρήση καταλληλότερων μετρικών αξιολόγησης. Επιπλέον, παρουσιάζονται οι τεχνικές υπερδειγματοληψίας, όπως το SMOTE, το Borderline-SMOTE και το SMOTE-Tomek, καθώς και οι γενετικές προσεγγίσεις παραγωγής συνθετικών δεδομένων, με έμφαση στο CTGAN.

Στο Κεφάλαιο 3, περιγράφονται τα δεδομένα που χρησιμοποιήθηκαν και τα βήματα προεπεξεργασίας που εφαρμόστηκαν πριν από την εκπαίδευση των μοντέλων. Αρχικά παρουσιάζεται το σύνολο δεδομένων και τα βασικά χαρακτηριστικά του, όπως η ηλικία, ο δείκτης μάζας σώματος, το μέσο επίπεδο γλυκόζης, η υπέρταση, η καρδιακή νόσος και άλλες μεταβλητές που σχετίζονται με τον κίνδυνο εμφάνισης εγκεφαλικού επεισοδίου. Στη συνέχεια, πραγματοποιείται εξερευνητική ανάλυση των δεδομένων, με στόχο την κατανόηση των κατανομών, των συσχετίσεων και κυρίως της έντονης ανισορροπίας μεταξύ των δύο κλάσεων. Το κεφάλαιο συνεχίζεται με την παρουσίαση της προεπεξεργασίας, η οποία περιλαμβάνει τη διαχείριση ελλειπών τιμών, την κωδικοποίηση κατηγορικών χαρακτηριστικών και την κλιμάκωση των μεταβλητών. Τέλος, παρουσιάζονται οι μετρικές αξιολόγησης που χρησιμοποιούνται στην εργασία, με έμφαση στο Recall, το F1-score και το PR-AUC, καθώς οι μετρικές αυτές θεωρούνται πιο κατάλληλες για την αξιολόγηση μοντέλων σε ανισόρροπα ιατρικά δεδομένα.

Στο Κεφάλαιο 4, παρουσιάζεται η μεθοδολογία και η πειραματική διαδικασία της κεντριοποιημένης εκπαίδευσης. Αρχικά περιγράφονται τα μοντέλα μηχανικής μάθησης που χρησιμοποιήθηκαν, δηλαδή η Logistic Regression, το Random Forest και το XGBoost. Στη συνέχεια παρουσιάζονται τα πειράματα που πραγματοποιήθηκαν με διαφορετικές τεχνικές αντιμετώπισης της ανισορροπίας, ξεκινώντας από τα baseline αποτελέσματα χωρίς κάποια παρέμβαση στα δεδομένα. Έπειτα εξετάζεται η χρήση βαρών κλάσεων, οι τεχνικές υπερδειγματοληψίας όπως το SMOTE και οι παραλλαγές του, καθώς και η παραγωγή συνθετικών δεδομένων μέσω CTGAN. Στο ίδιο κεφάλαιο αναλύεται η συμπεριφορά των μοντέλων στο πραγματικό test set, το οποίο διατηρείται ανισόρροπο ώστε να προσομοιάζει περισσότερο ένα ρεαλιστικό περιβάλλον αξιολόγησης.

Στο Κεφάλαιο 5, εξετάζεται η προσέγγιση του Federated Learning. Το κεφάλαιο αυτό επικεντρώνεται στον σχεδιασμό μιας προσομοίωσης κατανεμημένης εκπαίδευσης, όπου τα δεδομένα θεωρείται ότι κατανέμονται σε διαφορετικούς clients, οι οποίοι μπορούν να αντιστοιχούν ενδεικτικά σε διαφορετικά νοσοκομεία ή φορείς υγείας. Επιπλέον, περιγράφεται η υλοποίηση του αλγορίθμου FedAvg, μέσω του οποίου οι ενημερώσεις των τοπικών μοντέλων συνδυάζονται σε ένα συνολικό μοντέλο. Τέλος, εξετάζεται η εφαρμογή τεχνικών τοπικής εξισορρόπησης στο πλαίσιο του Federated Learning, με στόχο να διερευνηθεί κατά πόσο η αντιμετώπιση της ανισορροπίας σε επίπεδο client μπορεί να βελτιώσει την τελική απόδοση.

Στο Κεφάλαιο 6, πραγματοποιείται η συγκριτική αξιολόγηση και η συζήτηση των αποτελεσμάτων. Σε αυτό το κεφάλαιο παρουσιάζονται συνολικά τα αποτελέσματα των πειραμάτων, μέσα από πίνακες και γραφήματα που επιτρέπουν τη σύγκριση διαφορετικών μοντέλων, τεχνικών εξισορρόπησης και μεθόδων εκπαίδευσης. Ιδιαίτερη έμφαση δίνεται στη σύγκριση μεταξύ κεντριοποιημένης και κατανεμημένης εκπαίδευσης, καθώς και στη σύγκριση μεταξύ SMOTE και CTGAN, ώστε να αξιολογηθεί ποια μέθοδος παρήγαγε πιο χρήσιμα δεδομένα για την εκπαίδευση των μοντέλων. Επιπλέον, γίνεται ερμηνεία των αποτελεσμάτων με βάση τις επιλεγμένες μετρικές αξιολόγησης,

δίνοντας ιδιαίτερη βαρύτητα στην ικανότητα των μοντέλων να εντοπίζουν τη μειοψηφική κλάση του εγκεφαλικού επεισοδίου.

Τέλος, στο Κεφάλαιο 7, παρουσιάζονται τα συμπεράσματα και οι προτάσεις για μελλοντική έρευνα. Αρχικά γίνεται σύνοψη των βασικών ευρημάτων της εργασίας και αναδεικνύονται τα σημαντικότερα συμπεράσματα που προέκυψαν από την πειραματική διαδικασία. Στη συνέχεια, αναφέρονται οι περιορισμοί της έρευνας, όπως το μέγεθος και η φύση του συνόλου δεδομένων, η ποιότητα των συνθετικών δεδομένων και οι περιορισμοί της προσομοίωσης Federated Learning. Τέλος, προτείνονται πιθανές κατευθύνσεις για μελλοντική επέκταση, όπως η χρήση μεγαλύτερων και πιο αντιπροσωπευτικών ιατρικών συνόλων δεδομένων, η αξιοποίηση πιο σύνθετων γενετικών μοντέλων, η εφαρμογή επιπλέον τεχνικών προστασίας ιδιωτικότητας και η αξιολόγηση της μεθοδολογίας σε πραγματικά καταναεμημένα περιβάλλοντα υγείας.

Κεφάλαιο 2ο: Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζεται το θεωρητικό υπόβαθρο που απαιτείται για την κατανόηση της παρούσας εργασίας. Αρχικά, στην Ενότητα 2.1 γίνεται εισαγωγή στη μηχανική μάθηση και στις βασικές εφαρμογές της στον τομέα της υγείας, με έμφαση στην πρόβλεψη εγκεφαλικού επεισοδίου. Έπειτα, στην Ενότητα 2.2 αναλύονται οι βασικές τεχνικές μηχανικής μάθησης που χρησιμοποιούνται στη βιβλιογραφία, καθώς και τα κύρια προβλήματα που εμφανίζονται κατά την εφαρμογή τους σε ευαίσθητα ιατρικά δεδομένα. Επιπροσθέτως, δίνεται ιδιαίτερη έμφαση στο πρόβλημα της ανισορροπίας των δεδομένων, το οποίο επηρεάζει σε πολύ μεγάλο βαθμό την απόδοση και την αξιολόγηση των μοντέλων. Τέλος, στην Ενότητα 2.3 παρουσιάζονται οι τεχνικές αντιμετώπισης της ανισορροπίας, όπως η στάθμιση κλάσεων, το SMOTE, το Borderline-SMOTE και το SMOTE-Tomek. Τέλος, στην Ενότητα 2.4 παρουσιάζονται οι μέθοδοι παραγωγής συνθετικών δεδομένων, με έμφαση στα GANs και στο CTGAN. Το κεφάλαιο αυτό αποτελεί τη θεωρητική βάση για την ανάλυση και την πειραματική διαδικασία που ακολουθεί στα επόμενα κεφάλαια.

2.1 Machine Learning στην Ιατρική

Η μηχανική μάθηση τα τελευταία χρόνια έχει αποκτήσει ιδιαίτερη σημασία στον τομέα της υγείας, καθώς επιτρέπει την ανάλυση μεγάλου όγκου ιατρικών δεδομένων και την εξαγωγή χρήσιμων συμπερασμάτων που υποστηρίζουν τόσο τη διάγνωση όσο και την πρόγνωση ασθενειών. Ενδεικτικά, σχετικές μελέτες των Satapathy et al. [1], Thakur και Gupta [23] και Suryawanshi et al. [25] έχουν εξετάσει την εφαρμογή μοντέλων μηχανικής μάθησης στην πρόβλεψη εγκεφαλικού επεισοδίου, παρουσιάζοντας ικανοποιητικά αποτελέσματα απόδοσης [4]. Αντίθετα, από τις παραδοσιακές στατιστικές μεθόδους, οι αλγόριθμοι μηχανικής μάθησης μπορούν να εντοπίσουν πολύπλοκες και μη γραμμικές σχέσεις μεταξύ μεταβλητών, ακόμη και όταν αυτές δεν είναι άμεσα εμφανείς. Το χαρακτηριστικό αυτό είναι ιδιαίτερα σημαντικό σε ιατρικά δεδομένα, τα οποία είναι συχνά πολυπαραγοντικά και επηρεάζονται από αλληλεπιδράσεις μεταξύ δημογραφικών, κλινικών και βιολογικών παραμέτρων. Η εφαρμογή της μηχανικής μάθησης στην υγειονομική περίθαλψη καλύπτει ένα ευρύ φάσμα προβλημάτων. Ενδεικτικά, χρησιμοποιείται στην ανάλυση ιατρικών εικόνων, στην ταξινόμηση ασθενειών, στην εκτίμηση κινδύνου, στην πρόβλεψη της πορείας της υγείας ενός ασθενούς, καθώς και στη λήψη αποφάσεων σχετικά με τη θεραπεία του. Στην πραγματικότητα, τέτοια συστήματα λειτουργούν ως υποστηρικτικά εργαλεία για τους επαγγελματίες υγείας, παρέχοντας επιπλέον πληροφορία και αναδεικνύοντας μη γραμμικές σχέσεις που δεν είναι δυνατό να εντοπιστούν με τις υπάρχουσες συμβατικές μεθόδους ανάλυσης. Σχετικά με την πρόβλεψη εγκεφαλικού επεισοδίου, η χρήση της μηχανικής μάθησης παρουσιάζει μεγάλο ενδιαφέρον, καθώς η έγκαιρη αναγνώριση ατόμων υψηλού κινδύνου μπορεί να συμβάλει ουσιαστικά στην πρόληψη σοβαρών επιπλοκών. Τα περισσότερα μοντέλα βασίζονται σε δομημένα δεδομένα ασθενών, όπως ηλικία, φύλο, δείκτης μάζας σώματος, επίπεδα γλυκόζης, υπέρταση, καρδιακές παθήσεις και συνήθειες όπως το κάπνισμα. Μέσα από την εκπαίδευση σε ιστορικά δεδομένα, τα μοντέλα μαθαίνουν να συσχετίζουν συγκεκριμένους συνδυασμούς χαρακτηριστικών με αυξημένο ή μειωμένο κίνδυνο εμφάνισης της νόσου. Έχουν προταθεί πολλές διαφορετικές τεχνικές για την αντιμετώπιση του προβλήματος, όπως η Logistic Regression, οι Support Vector Machines (SVM), τα Decision Trees, τα Random Forests και τα Neural Networks, καθώς και πιο σύνθετες προσεγγίσεις όπως deep learning μέθοδοι [1]. Ωστόσο, πολλές από τις μελέτες αυτές αναφέρουν υψηλές τιμές ακρίβειας, συχνά μεγαλύτερες από 90% αλλά αν και τα αποτελέσματα αυτά φαίνονται αρχικά ενθαρρυντικά, απαιτούν πιο προσεκτική ερμηνεία καθώς ανιχνεύουν σπάνια της προδιάθεση εγκεφαλικού.

Ένα βασικό χαρακτηριστικό των ιατρικών δεδομένων, και ειδικά των δεδομένων πρόβλεψης εγκεφαλικού επεισοδίου, είναι η ανισορροπία μεταξύ των κλάσεων. Οι περιπτώσεις στις οποίες εμφανίζεται εγκεφαλικό αποτελούν συνήθως μικρό ποσοστό του συνόλου, γεγονός που μπορεί να επηρεάσει σημαντικά την απόδοση των μοντέλων. Σε τέτοιες περιπτώσεις, η χρήση της ακρίβειας ως μοναδικού δείκτη αξιολόγησης μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα. Ένα μοντέλο μπορεί να εμφανίζει υψηλή ακρίβεια απλώς προβλέποντας την πλειοψηφική κλάση, χωρίς να εντοπίζει αποτελεσματικά τις κρίσιμες περιπτώσεις της μειονοτικής κλάσης. Λόγω αυτού, η αξιολόγηση των μοντέλων σε τέτοιου είδους προβλήματα απαιτεί τη χρήση πιο κατάλληλων μετρικών, όπως η ανάκληση (recall), η F1-score, η καμπύλη precision-recall (PR-AUC), καθώς και η ανάλυση του πίνακα σύγχυσης. Επιπλέον, σε ιατρικό πλαίσιο, ένα ψευδώς αρνητικό αποτέλεσμα μπορεί να έχει πολύ μεγαλύτερες συνέπειες από ένα ψευδώς θετικό, καθώς στην πρώτη περίπτωση το σύστημα αποτυγχάνει να εντοπίσει έναν ασθενή που βρίσκεται σε πραγματικό κίνδυνο. Συμπερασματικά, η μηχανική μάθηση παίζει ένα ιδιαίτερα σημαντικό ρόλο στον χώρο της υγείας και πιο συγκεκριμένα στην πρόβλεψη εγκεφαλικού επεισοδίου. Πρέπει να τονιστεί ότι, παρά τον μεγάλο αριθμό προτεινόμενων μοντέλων και τις υψηλές επιδόσεις που συχνά αναφέρονται στη βιβλιογραφία, η πραγματική τους αξία εξαρτάται από την ικανότητά τους να αποδίδουν αξιόπιστα σε ρεαλιστικά και μη ισορροπημένα δεδομένα. Για αυτόν τον λόγο, η εργασία δίνει ιδιαίτερη έμφαση όχι μόνο στα ίδια τα μοντέλα, αλλά κυρίως στο πρόβλημα της ανισορροπίας των δεδομένων και πώς αυτό μπορεί να αντιμετωπιστεί, το οποίο επηρεάζει άμεσα και σημαντικά την εξαγωγή των αποτελεσμάτων.

Στη βιβλιογραφία έχουν παρουσιαστεί αρκετές προσεγγίσεις μηχανικής μάθησης για την πρόβλεψη εγκεφαλικού επεισοδίου ή για την εκτίμηση της έκβασης των ασθενών μετά από εγκεφαλικό. Οι περισσότερες μελέτες χρησιμοποιούν δημογραφικά, κλινικά και συμπεριφορικά χαρακτηριστικά, όπως ηλικία, υπέρταση, καρδιακή νόσο, μέσο επίπεδο γλυκόζης, δείκτη μάζας σώματος και κάπνισμα, με σκοπό την εκπαίδευση μοντέλων ταξινόμησης. Παρότι οι εργασίες αυτές δείχνουν ότι η μηχανική μάθηση μπορεί να συμβάλει στην έγκαιρη αναγνώριση ασθενών υψηλού κινδύνου, αρκετές από αυτές δίνουν μεγαλύτερη έμφαση στη συνολική ακρίβεια του μοντέλου και λιγότερο στην ικανότητα εντοπισμού της θετικής κλάσης, η οποία είναι ιδιαίτερα σημαντική σε προβλήματα όπως η πρόβλεψη εγκεφαλικού.

Οι Satapathy et al. [1] χρησιμοποίησαν ένα σύνολο δεδομένων περίπου 5.000 εγγραφών και εφάρμοσαν τέσσερις αλγορίθμους μηχανικής μάθησης: Logistic Regression, Random Forest, Decision Tree και Support Vector Machine. Τα χαρακτηριστικά που χρησιμοποιήθηκαν περιλάμβαναν μεταξύ άλλων την ηλικία, την υπέρταση, την καρδιακή νόσο, το BMI, το κάπνισμα και το μέσο επίπεδο γλυκόζης. Σύμφωνα με τα αποτελέσματά τους, οι Logistic Regression, Random Forest και SVM πέτυχαν ακρίβεια 94,50%, ενώ το Decision Tree παρουσίασε χαμηλότερη ακρίβεια, 91,03%. Παρόλο που η εργασία αναφέρει υψηλές τιμές ακρίβειας, η αξιολόγηση βασίζεται κυρίως σε κλασικές μετρικές, χωρίς να αναλύεται σε βάθος η επίδραση της ανισορροπίας των κλάσεων. Αυτό είναι σημαντικό, γιατί σε ιατρικά προβλήματα με λίγες θετικές περιπτώσεις η υψηλή ακρίβεια δεν σημαίνει απαραίτητα ότι το μοντέλο εντοπίζει αποτελεσματικά τους ασθενείς που πραγματικά κινδυνεύουν.

Αντίστοιχα, οι Thakur και Gupta [23] μελέτησαν την πρόβλεψη εγκεφαλικού χρησιμοποιώντας Logistic Regression, Random Forest και Decision Tree. Το σύνολο δεδομένων περιλάμβανε δημογραφικά στοιχεία, ιατρικό ιστορικό και παράγοντες τρόπου ζωής. Στα αποτελέσματά τους, η Logistic Regression πέτυχε ακρίβεια 78%, ενώ το Random Forest και το Decision Tree πέτυχαν ακρίβεια 91%. Οι συγγραφείς αναφέρουν επίσης Precision, Recall και F1-score, γεγονός που δίνει πιο ολοκληρωμένη εικόνα σε σχέση με την απλή ακρίβεια. Ωστόσο, και σε αυτή την περίπτωση δεν δίνεται ιδιαίτερη

έμφαση σε τεχνικές εξισορρόπησης των δεδομένων, όπως το SMOTE, ούτε εξετάζεται πώς αλλάζει η απόδοση του μοντέλου όταν η θετική κλάση είναι αρκετά μικρότερη από την αρνητική.

Μία διαφορετική προσέγγιση παρουσιάζεται από τους Suryawanshi et al. [25], οι οποίοι εστίασαν στη χρήση Logistic Regression για δυαδική πρόβλεψη εγκεφαλικού. Η εργασία τους βασίζεται στην ιδέα ότι ορισμένα χαρακτηριστικά, όπως η ηλικία και το μέσο επίπεδο γλυκόζης, μπορεί να έχουν μη γραμμική σχέση με την πιθανότητα εμφάνισης εγκεφαλικού. Για τον λόγο αυτό, εφάρμοσαν λογαριθμικούς μετασχηματισμούς σε συγκεκριμένες μεταβλητές, με στόχο τη βελτίωση της απόδοσης του μοντέλου. Το τροποποιημένο μοντέλο Logistic Regression πέτυχε AUC 0,92, ενώ το βασικό μοντέλο πέτυχε AUC 0,91. Η συγκεκριμένη εργασία είναι χρήσιμη για την παρούσα μελέτη, επειδή δείχνει ότι ακόμη και απλούστερα και ερμηνεύσιμα μοντέλα, όπως η Logistic Regression, μπορούν να αποδώσουν ικανοποιητικά όταν συνδυάζονται με κατάλληλη προεπεξεργασία των χαρακτηριστικών.

Οι Deepthi et al. [2] ασχολήθηκαν επίσης με την πρόβλεψη εγκεφαλικού επεισοδίου χρησιμοποιώντας Random Forest και K-Nearest Neighbors. Στην εργασία τους αναφέρονται ως σημαντικοί παράγοντες η ηλικία, οι καρδιακές παθήσεις, το μέσο επίπεδο γλυκόζης και η υπέρταση. Η μελέτη υποστηρίζει ότι οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την ανάλυση ιατρικών δεδομένων και την πρόβλεψη κινδύνου εγκεφαλικού. Ωστόσο, η παρουσίαση των αποτελεσμάτων επικεντρώνεται κυρίως στη γενική πρόβλεψη και στην ακρίβεια, χωρίς εκτενή συζήτηση για το πρόβλημα της ανισορροπίας των κλάσεων ή για το κατά πόσο το μοντέλο εντοπίζει ικανοποιητικά τα πραγματικά περιστατικά εγκεφαλικού.

Σε πιο κλινικό πλαίσιο, οι Chiu et al. [9] δεν ασχολήθηκαν απλώς με την πρόβλεψη εμφάνισης εγκεφαλικού, αλλά με την πρόβλεψη της λειτουργικής έκβασης ασθενών με οξύ ισχαιμικό εγκεφαλικό μετά από θεραπεία επαναιμάτωσης. Στη μελέτη τους χρησιμοποιήθηκαν αλγόριθμοι Logistic Regression, Support Vector Machine, Random Forest και Extreme Gradient Boosting, ενώ τα αποτελέσματα συγκρίθηκαν με το παραδοσιακό DRAGON score. Τα μοντέλα μηχανικής μάθησης παρουσίασαν καλύτερη συνολική απόδοση σε σχέση με το κλινικό score ως προς μετρικές όπως Accuracy, Precision, Recall και Specificity. Η εργασία αυτή δείχνει ότι η μηχανική μάθηση μπορεί να αξιοποιηθεί όχι μόνο για την αρχική πρόβλεψη κινδύνου, αλλά και για την εκτίμηση της μελλοντικής πορείας του ασθενούς.

Συνολικά, οι παραπάνω μελέτες δείχνουν ότι η μηχανική μάθηση έχει εφαρμοστεί εκτενώς στην πρόβλεψη εγκεφαλικού επεισοδίου, κυρίως με μοντέλα όπως Logistic Regression, Random Forest, Decision Tree, SVM, KNN και XGBoost. Παρόλα αυτά, ένα κοινό σημείο που παρατηρείται είναι ότι αρκετές εργασίες εστιάζουν κυρίως στην ακρίβεια, ενώ η ανισορροπία των δεδομένων δεν εξετάζεται πάντα με συστηματικό τρόπο. Σε προβλήματα όπως το εγκεφαλικό, όπου οι θετικές περιπτώσεις είναι συνήθως λιγότερες από τις αρνητικές, η απλή ακρίβεια μπορεί να δημιουργήσει παραπλανητική εικόνα για την πραγματική απόδοση του μοντέλου. Για τον λόγο αυτό, η παρούσα εργασία διαφοροποιείται εξετάζοντας όχι μόνο κλασικά μοντέλα ταξινόμησης, αλλά και τεχνικές αντιμετώπισης της ανισορροπίας, όπως class weights, SMOTE-based μέθοδοι και CTGAN, με έμφαση σε μετρικές όπως Recall, F1-score και PR-AUC. Επιπλέον, επεκτείνει την ανάλυση σε περιβάλλον Federated Learning, ώστε να εξεταστεί η συμπεριφορά της πρόβλεψης εγκεφαλικού και σε κατανεμημένο πλαίσιο εκπαίδευσης.

Όπως φαίνεται στον Πίνακα 2.1, οι περισσότερες σχετικές εργασίες επικεντρώνονται κυρίως στη σύγκριση κλασικών αλγορίθμων μηχανικής μάθησης, όπως Logistic Regression, Random Forest, Decision Tree, SVM (Support Vector Machine) και KNN (K-Nearest Neighbors). Παρότι αρκετές από αυτές παρουσιάζουν υψηλές τιμές Accuracy, η αντιμετώπιση της ανισορροπίας των δεδομένων δεν

αναλύεται πάντα συστηματικά. Η παρούσα ΔΕ διαφοροποιείται από τις παραπάνω προσεγγίσεις, καθώς δεν περιορίζεται μόνο στη σύγκριση μοντέλων, αλλά εξετάζει διαφορετικές τεχνικές εξισορρόπησης, όπως class weights, SMOTE, Borderline-SMOTE, SMOTE-Tomek και CTGAN. Επιπλέον, η αξιολόγηση επεκτείνεται και σε περιβάλλον Federated Learning, στοιχείο που δεν εμφανίζεται στις συγκεκριμένες μελέτες.

Πίνακας 2.1: Σύνοψη σχετικών εργασιών και σύγκριση με την παρούσα ΔΕ.

Έρευνα	Μοντέλα / τεχνικές	Βασικό αποτέλεσμα	Τεχνικές που δεν εξετάζονται
Satapathy et al. [1]	Logistic Regression, Random Forest, Decision Tree, SVM	Οι LR, RF και SVM πέτυχαν Accuracy 94,50%, ενώ το Decision Tree 91%	class weights, Borderline-SMOTE, SMOTE-Tomek, CTGAN και Federated Learning
Thakur & Gupta [23]	Logistic Regression, Random Forest, Decision Tree	Η Logistic Regression πέτυχε Accuracy 78%, ενώ Random Forest και Decision Tree πέτυχαν 91%	class weights, SMOTE-based, CTGAN και Federated Learning
Suryawanshi et al. [25]	Logistic Regression	Το τροποποιημένο μοντέλο πέτυχε AUC 0,92, έναντι 0,91 του βασικού μοντέλου	Random Forest/XGBoost (πειραματική σύγκριση), class weights, SMOTE-based, CTGAN και Federated Learning
Deepthi et al. [2]	KNN, Random Forest	Αναδείχθηκαν παράγοντες όπως ηλικία, καρδιακή νόσος κλπ.	Logistic Regression/XGBoost (βασικά συγκριτικά μοντέλα), class weights, SMOTE-based, CTGAN και Federated Learning
Chiu et al. [9]	Logistic Regression, SVM, Random Forest, XGBoost, DRAGON score	Τα μοντέλα ML παρουσίασαν καλύτερη απόδοση από το DRAGON score	CTGAN, SMOTE-Tomek, Borderline-SMOTE και Federated Learning

2.2 Imbalanced Datasets

Σε πολλά προβλήματα μηχανικής μάθησης, και ιδιαίτερα στον τομέα της υγείας, τα διαθέσιμα δεδομένα δεν κατανέμονται ομοιόμορφα μεταξύ των κατηγοριών. Το φαινόμενο αυτό είναι γνωστό ως

ανισορροπία δεδομένων (class imbalance) και εμφανίζεται όταν η μία κλάση υπερισχύει σημαντικά σε αριθμό δειγμάτων σε σχέση με την άλλη. Στην περίπτωση της πρόβλεψης εγκεφαλικού επεισοδίου, τα άτομα που εμφανίζουν τη νόσο αποτελούν συνήθως ένα μικρό ποσοστό του συνολικού πληθυσμού, γεγονός που οδηγεί σε έντονα μη ισορροπημένα σύνολα δεδομένων. Η ύπαρξη ανισορροπίας αποτελεί μία από τις βασικότερες προκλήσεις στην ανάπτυξη αξιόπιστων μοντέλων πρόβλεψης. Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης έχουν σχεδιαστεί με την υπόθεση ότι τα δεδομένα είναι σχετικά ισορροπημένα. Ως αποτέλεσμα, κατά τη διαδικασία εκπαίδευσης, τα μοντέλα τείνουν να δίνουν μεγαλύτερη έμφαση στην πλειοψηφική κλάση, καθώς αυτή κυριαρχεί αριθμητικά. Αυτό οδηγεί σε μοντέλα που εμφανίζουν υψηλή συνολική απόδοση, αλλά χαμηλή ικανότητα εντοπισμού της μειονοτικής κλάσης, η οποία στις περισσότερες ιατρικές εφαρμογές είναι και η πιο σημαντική. Ένα από τα κύρια προβλήματα που προκύπτουν σε τέτοιες περιπτώσεις είναι η μεροληψία του μοντέλου. Το μοντέλο “μαθαίνει” ότι είναι πιο ασφαλές να προβλέπει την πλειοψηφική κλάση, καθώς έτσι μειώνει το συνολικό σφάλμα κατά τη διάρκεια της εκπαίδευσης. Για παράδειγμα, σε ένα σύνολο δεδομένων όπου το 95% των δειγμάτων ανήκει στην αρνητική κλάση (no stroke), ένα μοντέλο μπορεί να επιτύχει ακρίβεια της τάξης του 95% απλώς προβλέποντας ότι κανένα άτομο δεν θα εμφανίσει εγκεφαλικό επεισόδιο. Παρότι το αποτέλεσμα αυτό φαίνεται εντυπωσιακό, στην πράξη το μοντέλο είναι ουσιαστικά ανούσιο, καθώς αποτυγχάνει να εντοπίσει τις περιπτώσεις που έχουν πραγματική σημασία. Το φαινόμενο αυτό αναδεικνύει τον κίνδυνο χρήσης της ακρίβειας ως μοναδικού δείκτη αξιολόγησης σε μη ισορροπημένα δεδομένα. Η ακρίβεια υπολογίζει το ποσοστό των σωστών προβλέψεων στο σύνολο των δειγμάτων, χωρίς να λαμβάνει υπόψη τη σχετική σημασία των διαφορετικών τύπων σφαλμάτων. Στα ιατρικά προβλήματα, ωστόσο, τα σφάλματα δεν έχουν την ίδια βαρύτητα. Ένα ψευδώς αρνητικό αποτέλεσμα (false negative), δηλαδή η αποτυχία εντοπισμού ενός ασθενούς που βρίσκεται σε κίνδυνο, μπορεί να έχει σοβαρές συνέπειες, ενώ ένα ψευδώς θετικό αποτέλεσμα (false positive) είναι συνήθως λιγότερο κρίσιμο. Αυτό δείχνει ότι, σε προβλήματα με ανισορροπία δεδομένων απαιτείται η χρήση πιο κατάλληλων μετρικών αξιολόγησης, οι οποίες εστιάζουν στην απόδοση του μοντέλου ως προς τη μειονοτική κλάση. Η ανάκληση (recall) αποτελεί μία από τις σημαντικότερες μετρικές, καθώς μετρά το ποσοστό των πραγματικών θετικών περιπτώσεων που εντοπίζονται σωστά από το μοντέλο. Υψηλή τιμή ανάκλησης σημαίνει ότι το μοντέλο είναι ικανό να εντοπίζει αποτελεσματικά τα περιστατικά της νόσου, κάτι που είναι ιδιαίτερα σημαντικό σε ιατρικά πλαίσια. Παράλληλα, η F1-score συνδυάζει την ακρίβεια (precision) και την ανάκληση, παρέχοντας μια πιο ισορροπημένη εικόνα της απόδοσης του μοντέλου. Η χρήση της είναι χρήσιμη όταν υπάρχει ανάγκη για συμβιβασμό μεταξύ των δύο μετρικών. Ωστόσο, σε περιπτώσεις έντονης ανισορροπίας, ακόμη πιο κατάλληλη θεωρείται η καμπύλη precision-recall και το αντίστοιχο εμβαδόν κάτω από αυτή (PR-AUC). Σε αντίθεση με την ROC καμπύλη, η οποία μπορεί να εμφανίζεται αισιόδοξη σε μη ισορροπημένα δεδομένα, η PR καμπύλη εστιάζει αποκλειστικά στη συμπεριφορά του μοντέλου ως προς τη θετική κλάση και παρέχει πιο ρεαλιστική εικόνα της απόδοσης. Εκτός από τις μετρικές αξιολόγησης, σημαντικό ρόλο παίζει και η ανάπτυξη κατάλληλων τεχνικών για την αντιμετώπιση της ανισορροπίας. Οι τεχνικές αυτές στοχεύουν στην εξισορρόπηση της κατανομής των δεδομένων, είτε μέσω υποδειγματοληψίας της πλειοψηφικής κλάσης είτε μέσω υπερδειγματοληψίας της μειονοτικής. Μεταξύ των πιο διαδεδομένων μεθόδων περιλαμβάνεται το SMOTE (Synthetic Minority Over-sampling Technique), το οποίο δημιουργεί συνθετικά δείγματα της μειονοτικής κλάσης, καθώς και οι παραλλαγές του, όπως το Borderline-SMOTE και το SMOTE-Tomek. Πιο πρόσφατα, έχουν προταθεί και παραγωγικές προσεγγίσεις, όπως τα Generative Adversarial Networks (GANs), που επιτρέπουν τη δημιουργία πιο ρεαλιστικών συνθετικών δεδομένων. Ένα χαρακτηριστικό παράδειγμα αποτελεί το CTGAN, το οποίο έχει σχεδιαστεί ειδικά για δομημένα δεδομένα και μπορεί να χρησιμοποιηθεί για την ενίσχυση της μειονοτικής κλάσης. Ωστόσο, η

αποτελεσματικότητα αυτών των μεθόδων εξαρτάται από παράγοντες όπως η ποιότητα των παραγόμενων δεδομένων και η επιλογή των κατάλληλων παραμέτρων [7].

Συνοψίζοντας, η ανισορροπία δεδομένων αποτελεί ένα από τα σημαντικότερα εμπόδια στην ανάπτυξη αξιόπιστων μοντέλων μηχανικής μάθησης στον τομέα της υγείας. Η ύπαρξή της επηρεάζει τόσο τη διαδικασία εκπαίδευσης όσο και την αξιολόγηση των μοντέλων, καθιστώντας απαραίτητη τη χρήση κατάλληλων μετρικών και τεχνικών αντιμετώπισης. Η κατανόηση και η σωστή διαχείριση του προβλήματος αυτού είναι κρίσιμη για την ανάπτυξη μοντέλων που δεν είναι μόνο ακριβή, αλλά και πρακτικά χρήσιμα σε πραγματικές συνθήκες.

2.3 Τεχνικές Αντιμετώπισης Imbalance

Η ανισορροπία δεδομένων αποτελεί ένα από τα βασικότερα εμπόδια στην ανάπτυξη αξιόπιστων μοντέλων μηχανικής μάθησης, ιδιαίτερα σε εφαρμογές υγείας όπου η μειονοτική κλάση είναι συχνά και η πιο κρίσιμη. Για την αντιμετώπιση αυτού του προβλήματος έχουν αναπτυχθεί διάφορες τεχνικές, οι οποίες στοχεύουν είτε στην τροποποίηση της διαδικασίας εκπαίδευσης είτε στην αναδιάρθρωση της κατανομής των δεδομένων. Οι προσεγγίσεις αυτές μπορούν γενικά να διακριθούν σε τρεις κατηγορίες. Η πρώτη αποτελείται από μεθόδους που προσαρμόζουν το ίδιο το μοντέλο, όπως η στάθμιση κλάσεων, η δεύτερη περιλαμβάνει τεχνικές δειγματοληψίας, όπως η υπερδειγματοληψία και η υποδειγματοληψία και τέλος, πιο σύνθετες μεθόδους που δημιουργούν συνθετικά δεδομένα, όπως το SMOTE και οι παραλλαγές του. Στη συνέχεια παρουσιάζονται οι βασικές τεχνικές που χρησιμοποιούνται στην παρούσα εργασία.

2.3.1 Class Weighting

Μία από τις πιο άμεσες προσεγγίσεις για την αντιμετώπιση της ανισορροπίας είναι η προσαρμογή των βαρών των κλάσεων κατά τη διαδικασία εκπαίδευσης του μοντέλου. Η βασική ιδέα είναι ότι τα σφάλματα που σχετίζονται με τη μειονοτική κλάση πρέπει να έχουν μεγαλύτερο κόστος σε σχέση με αυτά της πλειοψηφικής. Στην πράξη, αυτό επιτυγχάνεται αναθέτοντας μεγαλύτερο βάρος στα δείγματα της μειονοτικής κλάσης κατά τον υπολογισμό της συνάρτησης κόστους. Έτσι, το μοντέλο “τιμωρείται” περισσότερο όταν αποτυγχάνει να εντοπίσει ένα θετικό περιστατικό, γεγονός που το ωθεί να δώσει μεγαλύτερη προσοχή σε αυτές τις περιπτώσεις.

Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη σε αλγορίθμους όπως η Logistic Regression, όπου η συνάρτηση κόστους μπορεί εύκολα να προσαρμοστεί ώστε να λαμβάνει υπόψη διαφορετικά βάρη για κάθε κλάση. Επιπλέον, δεν απαιτεί καμία τροποποίηση στο ίδιο το σύνολο δεδομένων, γεγονός που την καθιστά απλή και υπολογιστικά αποδοτική. Ωστόσο, η αποτελεσματικότητα της μεθόδου εξαρτάται σε μεγάλο βαθμό από τη σωστή επιλογή των βαρών. Εάν τα βάρη είναι υπερβολικά υψηλά, το μοντέλο μπορεί να οδηγηθεί σε υπερπροσαρμογή (overfitting) στη μειονοτική κλάση, ενώ αν είναι πολύ χαμηλά, το πρόβλημα της ανισορροπίας παραμένει. Συχνά, τα βάρη καθορίζονται αυτόματα με βάση τη συχνότητα των κλάσεων, ωστόσο σε πολλές περιπτώσεις απαιτείται εμπειρική προσαρμογή.

Συνολικά, η στάθμιση κλάσεων αποτελεί μια απλή αλλά αποτελεσματική τεχνική η οποία χρησιμοποιείται συχνά ως baseline προσέγγιση πριν την εφαρμογή πιο σύνθετων μεθόδων.

2.3.2 SMOTE

Το SMOTE (Synthetic Minority Over-sampling Technique) αποτελεί μία από τις πιο ευρέως χρησιμοποιούμενες τεχνικές για την αντιμετώπιση της ανισορροπίας δεδομένων. Σε αντίθεση με την απλή υπερδειγματοληψία, η οποία δημιουργεί αντίγραφα υπαρχόντων δειγμάτων, το SMOTE

δημιουργεί νέα συνθετικά δείγματα βασισμένα στη γειτονιά των υπαρχόντων δεδομένων. Η βασική ιδέα του SMOTE είναι ότι για κάθε δείγμα της μειονοτικής κλάσης, επιλέγονται κάποιοι από τους κοντινότερους γείτονές του (k -nearest neighbors) και δημιουργούνται νέα δείγματα κατά μήκος της ευθείας που συνδέει τα σημεία αυτά. Με τον τρόπο αυτό, παράγονται νέα δεδομένα που βρίσκονται εντός του χώρου της μειονοτικής κλάσης, αντί να αναπαράγονται υπάρχοντα δείγματα. Με την προσέγγιση αυτή δημιουργούνται περισσότερες και συνεχείς περιοχές στον χώρο των χαρακτηριστικών, διευκολύνοντας το μοντέλο να μάθει καλύτερα τα όρια μεταξύ των κλάσεων. Επιπλέον, μειώνει τον κίνδυνο υπερπροσαρμογής που σχετίζεται με την απλή αντιγραφή δεδομένων. Παρά τα πλεονεκτήματα αυτά, το SMOTE παρουσιάζει και ορισμένους περιορισμούς. Ένα βασικό πρόβλημα είναι ότι δεν λαμβάνει υπόψη τη δομή των ορίων μεταξύ των κλάσεων. Ως αποτέλεσμα, μπορεί να δημιουργήσει συνθετικά δείγματα σε περιοχές όπου οι δύο κλάσεις επικαλύπτονται, εισάγοντας θόρυβο στο σύνολο δεδομένων. Αυτό μπορεί να επηρεάσει αρνητικά την απόδοση του μοντέλου, ιδιαίτερα όταν τα δεδομένα είναι ήδη πολύπλοκα. Επιπλέον, η επιλογή των παραμέτρων του SMOTE, όπως ο αριθμός των γειτόνων και ο βαθμός υπερδειγματοληψίας, επηρεάζει σημαντικά τα αποτελέσματα [8]. Η ακατάλληλη επιλογή μπορεί να οδηγήσει είτε σε ανεπαρκή εξισορρόπηση είτε σε υπερβολική αλλοίωση της αρχικής κατανομής.

Παρά τους περιορισμούς του, το SMOTE αποτελεί μια ισχυρή και ευρέως αποδεκτή μέθοδο, η οποία χρησιμοποιείται συχνά ως σημείο αναφοράς σε πειραματικές μελέτες.

2.3.3 Borderline-SMOTE

Το Borderline-SMOTE αποτελεί επέκταση του SMOTE και στοχεύει στη βελτίωση της ποιότητας των συνθετικών δεδομένων, εστιάζοντας στις πιο “δύσκολες” περιοχές του χώρου χαρακτηριστικών. Συγκεκριμένα, η μέθοδος εντοπίζει τα δείγματα της μειονοτικής κλάσης που βρίσκονται κοντά στα όρια μεταξύ των κλάσεων, δηλαδή εκεί όπου υπάρχει μεγαλύτερη πιθανότητα λανθασμένης ταξινόμησης. Αντί να δημιουργεί συνθετικά δείγματα για όλα τα σημεία της μειονοτικής κλάσης, το Borderline-SMOTE επικεντρώνεται μόνο σε αυτά τα κρίσιμα σημεία. Με τον τρόπο αυτό, ενισχύει τη “διαχωριστική ικανότητα” του μοντέλου, καθώς προσθέτει περισσότερη πληροφορία στις περιοχές όπου τα όρια των κλάσεων είναι ασαφή. Η προσέγγιση αυτή έχει αποδειχθεί ιδιαίτερα αποτελεσματική σε περιπτώσεις όπου η απλή εφαρμογή του SMOTE οδηγεί σε δημιουργία θορύβου. Ωστόσο, η μέθοδος εξαρτάται από τον σωστό εντοπισμό των borderline δειγμάτων, κάτι που μπορεί να επηρεαστεί από τη δομή των δεδομένων [9].

2.3.4 SMOTE-Tomek

Η μέθοδος SMOTE-Tomek συνδυάζει την υπερδειγματοληψία με τεχνικές καθαρισμού δεδομένων. Αρχικά εφαρμόζεται το SMOTE για τη δημιουργία συνθετικών δειγμάτων και στη συνέχεια χρησιμοποιούνται τα Tomek links για την αφαίρεση αμφίβολων ή θορυβωδών ζευγών δειγμάτων που βρίσκονται κοντά στα όρια των κλάσεων [10]. Τα Tomek links είναι ζεύγη δειγμάτων διαφορετικών κλάσεων που είναι πολύ κοντά μεταξύ τους και θεωρούνται ενδείξεις περιοχών επικάλυψης. Η αφαίρεσή τους συμβάλλει στη δημιουργία πιο καθαρών ορίων μεταξύ των κλάσεων, βελτιώνοντας την απόδοση του μοντέλου. Ο συνδυασμός αυτός επιτρέπει όχι μόνο την ενίσχυση της μειονοτικής κλάσης, αλλά και τη μείωση του θορύβου, καθιστώντας τη μέθοδο ιδιαίτερα αποτελεσματική σε πρακτικά σενάρια.

2.4 Generative Synthetic Data

Η ανάγκη για αποτελεσματική αντιμετώπιση της ανισορροπίας δεδομένων έχει οδηγήσει στην ανάπτυξη πιο σύνθετων προσεγγίσεων, οι οποίες δεν περιορίζονται απλώς στην ανακατανομή των υπάρχοντων δεδομένων, αλλά επιχειρούν τη δημιουργία νέων, συνθετικών δειγμάτων. Οι τεχνικές αυτές εντάσσονται στο ευρύτερο πεδίο της παραγωγικής μοντελοποίησης και στοχεύουν στην αναπαράσταση της υποκείμενης κατανομής των δεδομένων, ώστε να παραχθούν νέα δείγματα που να είναι όσο το δυνατόν πιο ρεαλιστικά. Σε αντίθεση με τις κλασικές μεθόδους υπερδειγματοληψίας, όπως το SMOTE, οι οποίες βασίζονται σε απλές γραμμικές παρεμβολές μεταξύ υπάρχοντων σημείων, οι παραγωγικές μέθοδοι προσπαθούν να “μάθουν” τη συνολική δομή των δεδομένων. Με τον τρόπο αυτό, μπορούν θεωρητικά να δημιουργήσουν πιο πλούσια και πιο ρεαλιστικά δείγματα, τα οποία ενσωματώνουν καλύτερα τις πολύπλοκες σχέσεις μεταξύ των χαρακτηριστικών. Η σημασία αυτών των τεχνικών είναι ιδιαίτερα έντονη σε ιατρικά δεδομένα, όπου οι μεταβλητές είναι συχνά ετερογενείς και παρουσιάζουν μη γραμμικές εξαρτήσεις. Η δυνατότητα δημιουργίας συνθετικών δεδομένων που διατηρούν αυτές τις ιδιότητες μπορεί να συμβάλει ουσιαστικά στη βελτίωση της απόδοσης των μοντέλων πρόβλεψης.

2.4.1 Synthetic Data Generation

Η δημιουργία συνθετικών δεδομένων (Synthetic Data Generation – SDG) αναφέρεται στη διαδικασία παραγωγής τεχνητών δεδομένων που μιμούνται τα στατιστικά χαρακτηριστικά ενός πραγματικού συνόλου δεδομένων. Τα δεδομένα αυτά δεν αντιστοιχούν σε πραγματικές παρατηρήσεις, αλλά έχουν σχεδιαστεί έτσι ώστε να διατηρούν τις βασικές ιδιότητες της αρχικής κατανομής. Η βασική ιδέα πίσω από το SDG είναι ότι, εφόσον μπορεί να μοντελοποιηθεί επαρκώς η κατανομή των δεδομένων, είναι δυνατή η παραγωγή νέων δειγμάτων που ακολουθούν την ίδια κατανομή. Αυτό επιτρέπει την ενίσχυση της μειονοτικής κλάσης χωρίς την απλή αντιγραφή υπάρχοντων δεδομένων, μειώνοντας τον κίνδυνο υπερπροσαρμογής. Στον τομέα της υγείας, η χρήση συνθετικών δεδομένων παρουσιάζει επιπλέον πλεονεκτήματα. Πέρα από την αντιμετώπιση της ανισορροπίας, μπορεί να συμβάλει στην προστασία ευαίσθητων προσωπικών δεδομένων, καθώς τα συνθετικά δείγματα δεν αντιστοιχούν σε πραγματικούς ασθενείς. Παράλληλα, διευκολύνει την ανταλλαγή δεδομένων για ερευνητικούς σκοπούς. Ωστόσο, η ποιότητα των συνθετικών δεδομένων αποτελεί κρίσιμο παράγοντα. Εάν τα παραγόμενα δεδομένα δεν αντικατοπτρίζουν επαρκώς την πραγματική κατανομή, ενδέχεται να οδηγήσουν σε λανθασμένη εκπαίδευση των μοντέλων. Συνεπώς, απαιτείται προσεκτική αξιολόγηση της καταλληλότητάς τους πριν από τη χρήση τους σε πειραματικά πλαίσια.

2.4.2 Generative Adversarial Networks (GANs)

Τα Generative Adversarial Networks (GANs) αποτελούν μία από τις πιο σημαντικές εξελίξεις στον τομέα της παραγωγικής μοντελοποίησης. Εισήχθησαν με στόχο τη δημιουργία δεδομένων υψηλής πιστότητας και βασίζονται σε μια ανταγωνιστική διαδικασία εκπαίδευσης μεταξύ δύο νευρωνικών δικτύων: του generator και του discriminator. Ο generator έχει ως στόχο να δημιουργήσει συνθετικά δείγματα που μοιάζουν όσο το δυνατόν περισσότερο με τα πραγματικά δεδομένα, ενώ ο discriminator προσπαθεί να διακρίνει αν ένα δείγμα είναι πραγματικό ή συνθετικό. Κατά τη διάρκεια της εκπαίδευσης, τα δύο δίκτυα “ανταγωνίζονται” μεταξύ τους, με αποτέλεσμα ο generator να βελτιώνεται σταδιακά και να παράγει ολοένα και πιο ρεαλιστικά δεδομένα. Η προσέγγιση αυτή έχει αποδειχθεί ιδιαίτερα επιτυχημένη σε δεδομένα εικόνας, όπου τα GANs μπορούν να δημιουργήσουν εντυπωσιακά ρεαλιστικά αποτελέσματα. Ωστόσο, η εφαρμογή τους σε δομημένα (tabular) δεδομένα παρουσιάζει σημαντικές δυσκολίες. Τα tabular δεδομένα περιλαμβάνουν συχνά τόσο συνεχείς όσο και κατηγορικές

μεταβλητές, ενώ οι σχέσεις μεταξύ των χαρακτηριστικών μπορεί να είναι σύνθετες και μη γραμμικές. Επιπλέον, η εκπαίδευση των GANs είναι γνωστό ότι είναι ασταθής και απαιτεί προσεκτική ρύθμιση υπερπαραμέτρων [11]. Προβλήματα όπως το mode collapse, όπου ο generator παράγει περιορισμένη ποικιλία δειγμάτων, μπορούν να επηρεάσουν αρνητικά την ποιότητα των παραγόμενων δεδομένων. Παρά τις προκλήσεις αυτές, τα GANs αποτελούν τη βάση για πιο εξειδικευμένες αρχιτεκτονικές που έχουν σχεδιαστεί ειδικά για tabular δεδομένα, όπως το CTGAN.

2.4.3 CTGAN

Το CTGAN (Conditional Tabular GAN) αποτελεί μια εξειδικευμένη αρχιτεκτονική GAN που έχει σχεδιαστεί ειδικά για τη δημιουργία συνθετικών δεδομένων σε πίνακες (tabular data). Σε αντίθεση με τα κλασικά GANs, το CTGAN λαμβάνει υπόψη τις ιδιαιτερότητες των δομημένων δεδομένων, όπως η ύπαρξη κατηγορικών μεταβλητών και η ανομοιογένεια στις κατανομές των χαρακτηριστικών. Ένα από τα βασικά χαρακτηριστικά του CTGAN είναι η χρήση condition vectors, οι οποίοι επιτρέπουν στο μοντέλο να εστιάζει σε συγκεκριμένες κατηγορίες κατά τη διαδικασία παραγωγής δεδομένων. Με τον τρόπο αυτό, μπορεί να ελεγχθεί η κατανομή των παραγόμενων δειγμάτων, κάτι που είναι ιδιαίτερα χρήσιμο για την ενίσχυση της μειονοτικής κλάσης. Επιπλέον, το CTGAN εφαρμόζει τεχνικές κανονικοποίησης για τη διαχείριση συνεχών μεταβλητών με πολύπλοκες κατανομές, καθώς και ειδικούς μηχανισμούς για την καλύτερη αναπαράσταση κατηγορικών δεδομένων. Αυτές οι βελτιώσεις επιτρέπουν στο μοντέλο να παράγει πιο ρεαλιστικά και στατιστικά συνεπή δεδομένα σε σύγκριση με τα παραδοσιακά GANs. Παρά τα πλεονεκτήματά του, το CTGAN δεν είναι απαλλαγμένο από περιορισμούς. Η ποιότητα των παραγόμενων δεδομένων εξαρτάται από το μέγεθος και την ποιότητα του αρχικού συνόλου δεδομένων, ενώ η διαδικασία εκπαίδευσης παραμένει απαιτητική υπολογιστικά. Επιπλέον, η χρήση συνθετικών δεδομένων μπορεί να οδηγήσει σε υπεραισιοδοξία στα αποτελέσματα, εάν δεν αξιολογηθεί προσεκτικά η γενικευσιμότητα του μοντέλου [12].

Στο πλαίσιο της παρούσας εργασίας, το CTGAN αξιοποιείται ως μια εναλλακτική προσέγγιση για την αντιμετώπιση της ανισορροπίας δεδομένων, επιτρέποντας τη δημιουργία συνθετικών δειγμάτων της μειονοτικής κλάσης. Η σύγκρισή του με παραδοσιακές μεθόδους, όπως το SMOTE, αναμένεται να προσφέρει χρήσιμα συμπεράσματα σχετικά με την αποτελεσματικότητα των παραγωγικών τεχνικών σε προβλήματα πρόβλεψης εγκεφαλικού επεισοδίου.

Κεφάλαιο 3ο: Περιγραφή Δεδομένων και Προεπεξεργασία

3.1 Εισαγωγή

Η παρούσα ενότητα περιγράφει τη μεθοδολογική προσέγγιση που ακολουθήθηκε για την υλοποίηση και αξιολόγηση των μοντέλων μηχανικής μάθησης στο πρόβλημα της πρόβλεψης εγκεφαλικού επεισοδίου. Σε αντίθεση με το προηγούμενο κεφάλαιο, το οποίο εστιάζει στο θεωρητικό υπόβαθρο, το Κεφάλαιο 3 επικεντρώνεται στην πρακτική εφαρμογή των μεθόδων και στον τρόπο με τον οποίο αυτές ενσωματώνονται σε ένα συνεκτικό πειραματικό πλαίσιο. Ιδιαίτερη έμφαση δίνεται στη διαμόρφωση μιας αυστηρά καθορισμένης διαδικασίας, η οποία διασφαλίζει τη συγκρισιμότητα των αποτελεσμάτων και την αξιοπιστία των συμπερασμάτων. Η ανάλυση ξεκινά με την περιγραφή του χρησιμοποιούμενου dataset και των βασικών χαρακτηριστικών του, ακολουθούμενη από τα στάδια προεπεξεργασίας των δεδομένων. Στη συνέχεια, παρουσιάζεται ο πειραματικός σχεδιασμός, με στόχο την αντικειμενική σύγκριση διαφορετικών τεχνικών αντιμετώπισης της ανισορροπίας. Επιπλέον, περιγράφεται το πλαίσιο αξιολόγησης που χρησιμοποιείται για την αποτίμηση της απόδοσης των μοντέλων, καθώς και οι λεπτομέρειες υλοποίησης που διασφαλίζουν την επαναληψιμότητα των πειραμάτων. Μέσω αυτής της δομημένης προσέγγισης, το κεφάλαιο θέτει τις βάσεις για την ανάλυση των αποτελεσμάτων που ακολουθεί, επιτρέποντας την εξαγωγή τεκμηριωμένων και αξιόπιστων συμπερασμάτων.

3.2 Περιγραφή Dataset

Η παρούσα μελέτη βασίζεται σε ένα δομημένο σύνολο δεδομένων που αφορά την πρόβλεψη εγκεφαλικού επεισοδίου (Stroke Prediction Dataset)¹, το οποίο είναι διαθέσιμο μέσω της πλατφόρμας Kaggle² και χρησιμοποιείται ευρέως σε ερευνητικά πλαίσια για την ανάπτυξη και αξιολόγηση μοντέλων μηχανικής μάθησης στον τομέα της υγείας. Το συγκεκριμένο dataset αποτελείται από 5.110 εγγραφές, καθεμία από τις οποίες αντιστοιχεί σε έναν ασθενή και περιέχει δημογραφικά και κλινικά χαρακτηριστικά που σχετίζονται με τον κίνδυνο εμφάνισης εγκεφαλικού επεισοδίου. Επίσης, περιλαμβάνει αρχικά 12 στήλες, εκ των οποίων η μία αντιστοιχεί στη μεταβλητή-στόχο. Η μεταβλητή στόχος είναι δυαδική και υποδηλώνει την παρουσία (1) ή απουσία (0) εγκεφαλικού επεισοδίου. Ο βασικός σκοπός της χρήσης του συγκεκριμένου συνόλου δεδομένων είναι η εκπαίδευση και αξιολόγηση μοντέλων μηχανικής μάθησης, τα οποία καλούνται να ταξινομήσουν έναν ασθενή είτε στην κλάση εμφάνισης εγκεφαλικού επεισοδίου είτε στην κλάση μη εμφάνισης εγκεφαλικού.

Το dataset περιλαμβάνει χαρακτηριστικά που μπορούν να χωριστούν σε τρεις βασικές κατηγορίες: δημογραφικά, κλινικά και χαρακτηριστικά τρόπου ζωής. Στα δημογραφικά χαρακτηριστικά περιλαμβάνονται μεταβλητές όπως το φύλο (*gender*), η ηλικία (*age*), η οικογενειακή κατάσταση (*ever married*), ο τύπος εργασίας (*work type*) και ο τύπος κατοικίας (*residence type*). Στα κλινικά χαρακτηριστικά περιλαμβάνονται μεταβλητές όπως η ύπαρξη υπέρτασης (*hypertension*), η ύπαρξη καρδιακής νόσου (*heart disease*), το μέσο επίπεδο γλυκόζης (*avg glucose level*) στο αίμα και ο δείκτης μάζας σώματος (*BMI*). Τέλος, στα χαρακτηριστικά τρόπου ζωής περιλαμβάνεται κυρίως η καπνιστική συνήθεια (*smoking status*) του ασθενούς. Ο συνδυασμός αυτών των μεταβλητών επιτρέπει τη μελέτη διαφορετικών παραγόντων που μπορεί να σχετίζονται άμεσα ή έμμεσα με την εμφάνιση εγκεφαλικού επεισοδίου. Πιο συγκεκριμένα, η στήλη *gender* περιγράφει το φύλο του ασθενούς. Η μεταβλητή αυτή είναι κατηγορική και περιλαμβάνει τιμές που αντιστοιχούν σε άνδρες, γυναίκες και σε πολύ περιορισμένες περιπτώσεις, άλλη κατηγορία. Η στήλη *age* περιγράφει την ηλικία του ατόμου και

¹ Stroke Prediction Dataset, Kaggle: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

² Η πλατφόρμα Kaggle: <https://www.kaggle.com>

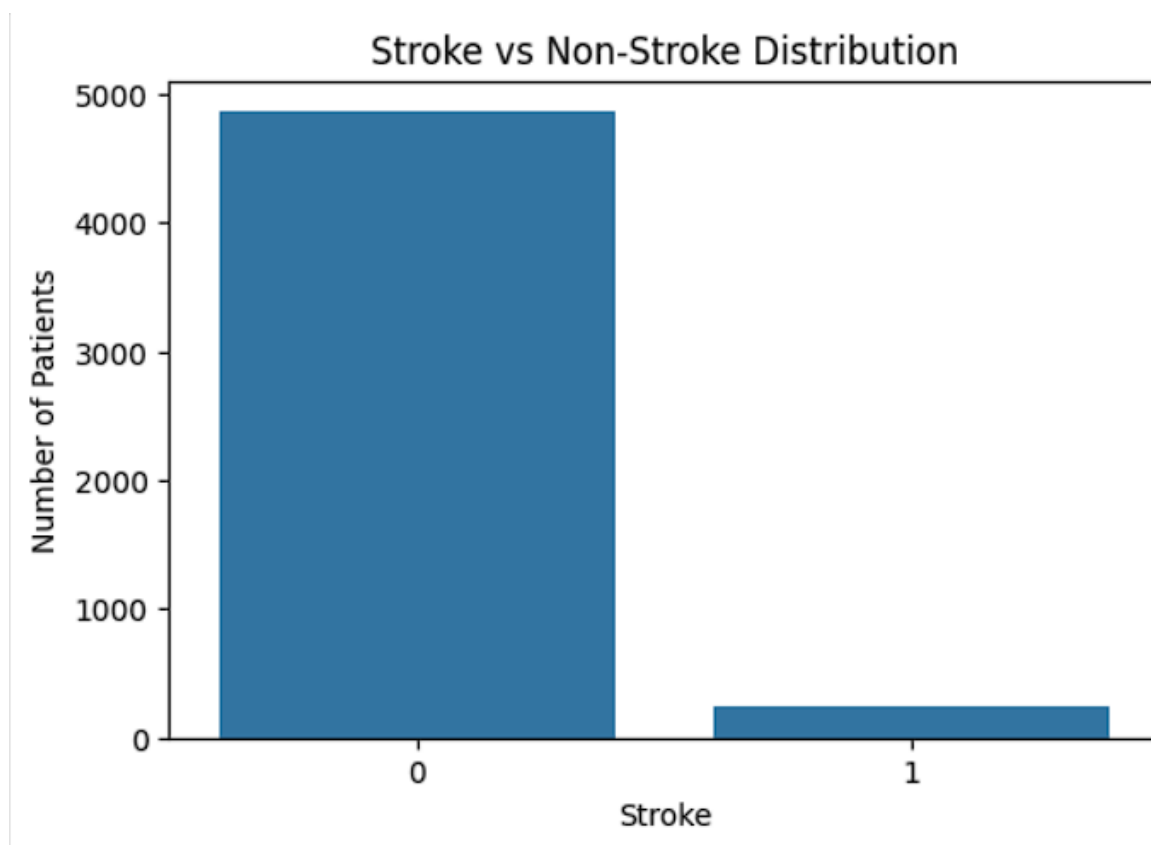
αποτελεί ένα από τα σημαντικότερα χαρακτηριστικά του dataset, καθώς η ηλικία είναι γνωστός παράγοντας κινδύνου για την εμφάνιση εγκεφαλικού επεισοδίου. Στην παρούσα εργασία, η ηλικία αντιμετωπίζεται ως αριθμητική μεταβλητή και χρησιμοποιείται ως βασικό χαρακτηριστικό εισόδου στα μοντέλα. Η μεταβλητή *hypertension* δηλώνει αν ο ασθενής έχει υπέρταση. Πρόκειται για δυαδική μεταβλητή, όπου η τιμή 0 δηλώνει απουσία υπέρτασης και η τιμή 1 δηλώνει ύπαρξη υπέρτασης. Αντίστοιχα, η μεταβλητή *heart_disease* δηλώνει αν ο ασθενής έχει ιστορικό καρδιακής νόσου. Και αυτή η μεταβλητή είναι δυαδική, με την τιμή 0 να αντιστοιχεί σε απουσία καρδιακής νόσου και την τιμή 1 σε ύπαρξη καρδιακής νόσου. Η στήλη *ever_married* δείχνει αν ο ασθενής έχει υπάρξει παντρεμένος, ενώ η μεταβλητή *work_type* περιγράφει τον τύπο εργασίας του ατόμου. Η συγκεκριμένη μεταβλητή περιλαμβάνει κατηγορίες όπως ιδιωτική εργασία, αυτοαπασχόληση, εργασία στον δημόσιο τομέα, οικιακά και περιπτώσεις όπου το άτομο δεν έχει εργαστεί. Παρόλο που οι μεταβλητές αυτές δεν αποτελούν άμεσα ιατρικούς δείκτες, μπορούν να συνδέονται έμμεσα με κοινωνικούς, οικονομικούς ή ηλικιακούς παράγοντες, οι οποίοι ενδέχεται να επηρεάζουν την κατάσταση υγείας ενός ατόμου. Η μεταβλητή *Residence_type* περιγράφει τον τύπο περιοχής κατοικίας του ασθενούς, δηλαδή αν πρόκειται για αστική ή αγροτική περιοχή. Η πληροφορία αυτή μπορεί να έχει ενδιαφέρον, καθώς ο τόπος κατοικίας μπορεί να σχετίζεται με διαφορετικά επίπεδα πρόσβασης σε υπηρεσίες υγείας, διαφορετικές συνήθειες ζωής ή διαφορετικούς περιβαλλοντικούς παράγοντες. Η στήλη *avg_glucose_level* αναφέρεται στο μέσο επίπεδο γλυκόζης στο αίμα και αποτελεί αριθμητικό χαρακτηριστικό. Η συγκεκριμένη μεταβλητή είναι σημαντική, καθώς τα αυξημένα επίπεδα γλυκόζης μπορεί να σχετίζονται με μεταβολικές διαταραχές, όπως ο σακχαρώδης διαβήτης, ο οποίος αποτελεί γνωστό παράγοντα κινδύνου για καρδιαγγειακά συμβάντα. Η στήλη *bmi* αντιστοιχεί στον δείκτη μάζας σώματος του ασθενούς. Ο δείκτης μάζας σώματος χρησιμοποιείται συχνά ως ένδειξη για την αξιολόγηση του σωματικού βάρους σε σχέση με το ύψος και μπορεί να συνδέεται με παράγοντες όπως η παχυσαρκία και ο γενικότερος μεταβολικός κίνδυνος. Στο συγκεκριμένο dataset, η μεταβλητή *bmi* παρουσιάζει ελλιπείς τιμές, γεγονός που απαιτεί κατάλληλη διαχείριση στο στάδιο της προεπεξεργασίας. Η επιλογή του τρόπου συμπλήρωσης των τιμών αυτών είναι σημαντική, καθώς η λανθασμένη διαχείριση ελλιπών δεδομένων μπορεί να επηρεάσει την ποιότητα της εκπαίδευσης των μοντέλων. Η μεταβλητή *smoking_status* περιγράφει την καπνιστική συνήθεια του ασθενούς. Περιλαμβάνει κατηγορίες που δηλώνουν αν το άτομο καπνίζει, αν κάπνιζε στο παρελθόν, αν δεν έχει καπνίσει ποτέ ή αν η πληροφορία είναι άγνωστη. Το κάπνισμα αποτελεί έναν σημαντικό τροποποιήσιμο παράγοντα κινδύνου για πολλές καρδιαγγειακές παθήσεις και επομένως η παρουσία αυτής της μεταβλητής είναι χρήσιμη για την πρόβλεψη του εγκεφαλικού επεισοδίου. Τέλος, η στήλη *id* αποτελεί μοναδικό αναγνωριστικό κάθε εγγραφής. Η στήλη αυτή δεν περιέχει πληροφορία που σχετίζεται με την υγεία ή τα χαρακτηριστικά του ασθενούς και για αυτόν τον λόγο αφαιρείται πριν από την εκπαίδευση των μοντέλων.

Πίνακας 3.1: Μεταβλητές του dataset

Όνομα Μεταβλητής	Περιγραφή Μεταβλητής	Κατηγορία Μεταβλητής
ID	Αναγνωριστικό εγγραφής	-
Gender	Φύλο του ασθενούς	Δημογραφικά χαρακτηριστικά
Age	Ηλικία του ατόμου	Δημογραφικά χαρακτηριστικά
Hypertension	Ύπαρξη υπέρτασης	Κλινικά χαρακτηριστικά
Heart Disease	Ύπαρξη καρδιακής νόσου	Κλινικά χαρακτηριστικά
Ever Married	Παντρεμένος ή όχι	Δημογραφικά χαρακτηριστικά

Όνομα Μεταβλητής	Περιγραφή Μεταβλητής	Κατηγορία Μεταβλητής
Work Type	Τύπος εργασίας	Δημογραφικά χαρακτηριστικά
Residence Type	Τύπος περιοχής κατοικίας	Δημογραφικά χαρακτηριστικά
<i>Avg Glucose Level</i>	Μέσο επίπεδο γλυκόζης	Κλινικά χαρακτηριστικά
BMI	Δείκτης μάζας σώματος	Κλινικά χαρακτηριστικά
Smoking Status	Καπνιστική συνήθεια	Χαρακτηριστικά τρόπου ζωής
Stroke	Ύπαρξη εγκεφαλικού	Μεταβλητή στόχος

Ένα από τα πιο σημαντικά χαρακτηριστικά του συγκεκριμένου dataset είναι η έντονη ανισορροπία μεταξύ των δύο κλάσεων της μεταβλητής stroke. Όπως φαίνεται και στο σχήμα 3.1, από τις συνολικά 5.110 εγγραφές, οι 4.861 αντιστοιχούν σε ασθενείς που δεν έχουν εμφανίσει εγκεφαλικό επεισόδιο, ενώ μόλις 249 αντιστοιχούν σε ασθενείς που έχουν εμφανίσει εγκεφαλικό. Αυτό σημαίνει ότι η αρνητική κλάση αντιπροσωπεύει περίπου το 95% του συνόλου δεδομένων, ενώ η θετική κλάση αντιστοιχεί μόνο στο 5%. Η κατανομή αυτή αποτελεί βασική πρόκληση για την εκπαίδευση των μοντέλων, καθώς ένα μοντέλο μπορεί να επιτύχει υψηλή συνολική ακρίβεια προβλέποντας σχεδόν πάντα την πλειοψηφική κλάση, χωρίς όμως να εντοπίζει αποτελεσματικά τα πραγματικά περιστατικά εγκεφαλικού.



Σχήμα 3.1 Δεδομένα stroke – no stroke.

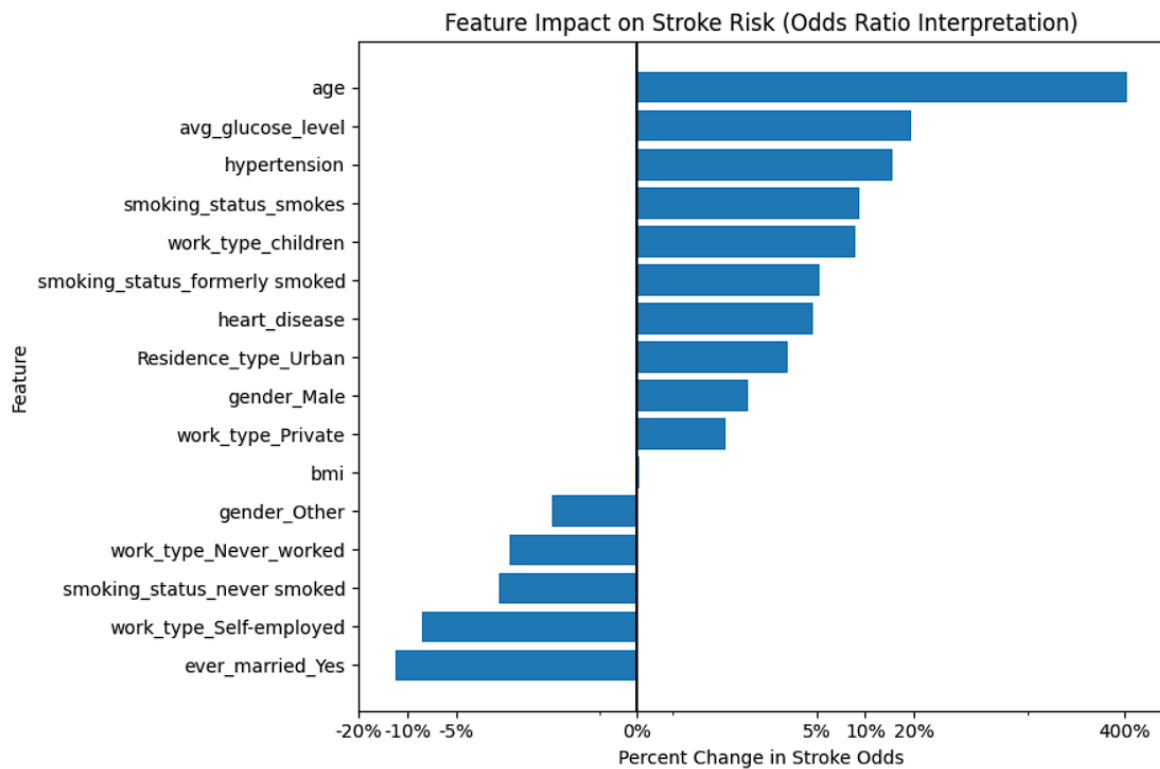
Η ύπαρξη αυτής της ανισορροπίας καθιστά το dataset ιδιαίτερα κατάλληλο για τη διερεύνηση τεχνικών εξισορρόπησης και παραγωγής συνθετικών δεδομένων. Στο πλαίσιο της παρούσας εργασίας, η θετική κλάση, δηλαδή οι περιπτώσεις εγκεφαλικού επεισοδίου, αποτελεί την κλάση με τη μεγαλύτερη σημασία από πλευράς πρόβλεψης, αλλά ταυτόχρονα και την κλάση με τα λιγότερα διαθέσιμα δείγματα. Αυτό δημιουργεί την ανάγκη εφαρμογής τεχνικών όπως τα βάρη κλάσεων, το SMOTE, το Borderline-SMOTE, το SMOTE-Tomek και το CTGAN, οι οποίες εξετάζονται στα επόμενα κεφάλαια. Η συνολική εικόνα δείχνει ότι, το dataset που χρησιμοποιείται στην παρούσα εργασία προσφέρει ένα ρεαλιστικό πλαίσιο μελέτης για την πρόβλεψη του εγκεφαλικού επεισοδίου. Περιλαμβάνει μεταβλητές που σχετίζονται με γνωστούς παράγοντες κινδύνου, διαθέτει σαφώς ορισμένη μεταβλητή-στόχο και παρουσιάζει έντονη ανισορροπία μεταξύ των κλάσεων.

3.3 Εξερευνητική Ανάλυση (EDA)

Η εξερευνητική ανάλυση δεδομένων αποτελεί ένα βασικό στάδιο πριν από την εκπαίδευση των μοντέλων μηχανικής μάθησης, καθώς επιτρέπει την καλύτερη κατανόηση της δομής, των χαρακτηριστικών και των πιθανών προβλημάτων που υπάρχουν στο σύνολο δεδομένων. Στην παρούσα εργασία, η ανάλυση αυτή πραγματοποιείται με στόχο να εξεταστεί η κατανομή των μεταβλητών, η σχέση τους με τη μεταβλητή-στόχο stroke, καθώς και η έντονη ανισορροπία που υπάρχει μεταξύ των δύο κλάσεων. Η διαδικασία αυτή είναι ιδιαίτερα σημαντική, επειδή τα ιατρικά δεδομένα συχνά παρουσιάζουν ιδιαιτερότητες που μπορούν να επηρεάσουν άμεσα την απόδοση των μοντέλων, όπως ελλιπείς τιμές, ακραίες τιμές, μη ισορροπημένες κλάσεις και μεταβλητές με διαφορετικές κλίμακες. Αρχικά, εξετάστηκε η γενική μορφή του dataset, το οποίο αποτελείται από 5.110 εγγραφές και 12 αρχικές στήλες. Κάθε γραμμή αντιστοιχεί σε έναν ασθενή, ενώ οι στήλες περιλαμβάνουν χαρακτηριστικά όπως το φύλο, η ηλικία, η ύπαρξη υπέρτασης, η ύπαρξη καρδιακής νόσου, η οικογενειακή κατάσταση, ο τύπος εργασίας, ο τύπος κατοικίας, το μέσο επίπεδο γλυκόζης, ο δείκτης μάζας σώματος, η καπνιστική συνήθεια και η μεταβλητή-στόχος. Η πρώτη αυτή εικόνα του dataset είναι απαραίτητη, καθώς βοηθά στον εντοπισμό των τύπων των μεταβλητών και στον διαχωρισμό τους σε αριθμητικές και κατηγορικές. Οι αριθμητικές μεταβλητές, όπως η age, η avg_glucose_level και η bmi, μπορούν να αναλυθούν μέσω στατιστικών μέτρων και κατανομών, ενώ οι κατηγορικές μεταβλητές, όπως η gender, η work_type, η Residence_type και η smoking_status, απαιτούν διαφορετικό χειρισμό, κυρίως στο στάδιο της κωδικοποίησης. Ένα από τα πρώτα σημεία που αναδείχθηκαν κατά την εξερευνητική ανάλυση είναι η ύπαρξη ελλিপών τιμών στη μεταβλητή bmi. Ο δείκτης μάζας σώματος αποτελεί ένα χαρακτηριστικό που μπορεί να σχετίζεται με τη γενικότερη κατάσταση υγείας ενός ασθενούς και με παράγοντες κινδύνου, όπως η παχυσαρκία και οι μεταβολικές διαταραχές. Η παρατήρηση αυτή δείχνει ότι, η ύπαρξη ελλিপών τιμών στη συγκεκριμένη μεταβλητή δεν μπορεί να αγνοηθεί. Αντίθετα, απαιτείται κατάλληλη διαχείριση στο στάδιο της προεπεξεργασίας, ώστε να μη χαθούν χρήσιμες εγγραφές από το dataset. Η παρουσία των ελλিপών τιμών στο bmi αποτελεί επομένως ένα σημαντικό εύρημα της αρχικής ανάλυσης και λαμβάνεται υπόψη στην επόμενη φάση, όπου εφαρμόζεται συμπλήρωση των ελλিপών τιμών. Ιδιαίτερη σημασία έχει η κατανομή της μεταβλητής-στόχου stroke. Η κατανομή αυτή επιβεβαιώνει ότι το dataset είναι έντονα ανισόρροπο. Η συγκεκριμένη παρατήρηση είναι καθοριστική για ολόκληρη την εργασία, καθώς επηρεάζει τόσο την επιλογή των τεχνικών εκπαίδευσης όσο και την επιλογή των μετρικών αξιολόγησης. Σε ένα τέτοιο περιβάλλον, ένα μοντέλο μπορεί να εμφανίζει υψηλή συνολική ακρίβεια απλώς προβλέποντας σχεδόν πάντα την πλειοψηφική κλάση, χωρίς όμως να εντοπίζει αποτελεσματικά τους ασθενείς που έχουν πραγματικά εμφανίσει εγκεφαλικό επεισόδιο. Η ανισορροπία αυτή είναι ιδιαίτερα σημαντική επειδή η θετική κλάση, αν και αριθμητικά μικρή, είναι η πιο κρίσιμη από πλευράς πρόβλεψης. Στην παρούσα εργασία, το ενδιαφέρον δεν βρίσκεται κυρίως στο να αναγνωριστούν σωστά οι ασθενείς που δεν έχουν εμφανίσει

εγκεφαλικό, καθώς αυτοί αποτελούν τη συντριπτική πλειονότητα του dataset. Αντίθετα, μεγαλύτερη σημασία έχει η ικανότητα των μοντέλων να εντοπίζουν τους ασθενείς που ανήκουν στη μειοψηφική κλάση και αυτό είναι ένα από τα πιο ουσιαστικά σημεία της ανάλυσης, επειδή δείχνει ότι η απλή παρατήρηση ενός υψηλού Accuracy δεν αρκεί για να θεωρηθεί ένα μοντέλο αποτελεσματικό. Σε ένα ιατρικό πρόβλημα, η αποτυχία εντοπισμού ενός πιθανού περιστατικού εγκεφαλικού μπορεί να είναι πολύ πιο σημαντική από την επίτευξη υψηλής συνολικής ακρίβειας. Στη συνέχεια, εξετάστηκαν οι βασικές αριθμητικές μεταβλητές του dataset. Η μεταβλητή age παρουσιάζει ιδιαίτερο ενδιαφέρον, καθώς η ηλικία αποτελεί γνωστό παράγοντα κινδύνου για την εμφάνιση εγκεφαλικού επεισοδίου. Από την ανάλυση προκύπτει ότι τα περιστατικά εγκεφαλικού εμφανίζονται συχνότερα σε μεγαλύτερες ηλικίες, γεγονός που είναι αναμενόμενο και συμβαδίζει με τη γενική ιατρική γνώση γύρω από τη νόσο. Η ηλικία, αναμένεται να αποτελέσει ένα από τα πιο σημαντικά χαρακτηριστικά για τα μοντέλα πρόβλεψης. Παρόλα αυτά, η ύπαρξη συσχέτισης μεταξύ ηλικίας και εγκεφαλικού δεν σημαίνει ότι η ηλικία από μόνη της αρκεί για την πρόβλεψη. Αντίθετα, η πρόβλεψη εξαρτάται από τον συνδυασμό πολλών παραγόντων, γεγονός που καθιστά χρήσιμη την εφαρμογή τεχνικών μηχανικής μάθησης. Η μεταβλητή avg_glucose_level, η οποία αναφέρεται στο μέσο επίπεδο γλυκόζης στο αίμα, είναι επίσης σημαντική, καθώς υψηλές τιμές γλυκόζης μπορεί να σχετίζονται με μεταβολικές διαταραχές και αυξημένο καρδιαγγειακό κίνδυνο. Κατά την εξερευνητική ανάλυση, η μεταβλητή αυτή εξετάζεται ως αριθμητικό χαρακτηριστικό που μπορεί να διαφοροποιείται μεταξύ των δύο κλάσεων. Αν και δεν μπορεί να θεωρηθεί από μόνη της απόλυτος δείκτης εμφάνισης εγκεφαλικού, η παρουσία αυξημένων τιμών σε ορισμένες περιπτώσεις μπορεί να συνεισφέρει στη συνολική εικόνα κινδύνου ενός ασθενούς. Πέρα από τις αριθμητικές μεταβλητές, εξετάστηκαν και οι κατηγορικές μεταβλητές του dataset όπως η μεταβλητή gender, η ever_married, η work_type, η Residence_type και η smoking_status. Οι μεταβλητές αυτές δεν έχουν όλες την ίδια βαρύτητα, αλλά μπορούν να προσφέρουν χρήσιμη πληροφορία στα μοντέλα, ειδικά όταν συνδυάζονται με τα υπόλοιπα χαρακτηριστικά. Για παράδειγμα, το κάπνισμα αποτελεί γνωστό τροποποιήσιμο παράγοντα κινδύνου για καρδιαγγειακά νοσήματα, ενώ ο τύπος εργασίας ή κατοικίας μπορεί να συνδέεται έμμεσα με διαφορετικά επίπεδα τρόπου ζωής, πρόσβασης σε υπηρεσίες υγείας ή καθημερινών συνηθειών. Σημαντικό μέρος της εξερευνητικής ανάλυσης αποτελεί και η διερεύνηση πιθανών συσχετίσεων μεταξύ των χαρακτηριστικών.

Μέσω της ανάλυσης συσχετίσεων μπορεί να εντοπιστεί αν ορισμένες μεταβλητές έχουν ισχυρότερη σχέση με τη μεταβλητή-στόχο ή αν υπάρχουν σχέσεις μεταξύ ανεξάρτητων μεταβλητών που πρέπει να ληφθούν υπόψη. Σε προβλήματα υγείας, οι σχέσεις αυτές δεν είναι πάντα γραμμικές ή άμεσα εμφανείς. Πιο συγκεκριμένα, από το Σχήμα 3.2 φαίνεται ότι η ηλικία αποτελεί το χαρακτηριστικό με τη μεγαλύτερη θετική επίδραση στον εκτιμώμενο κίνδυνο εμφάνισης εγκεφαλικού επεισοδίου. Θετική επίδραση παρουσιάζουν επίσης μεταβλητές όπως το μέσο επίπεδο γλυκόζης, η υπέρταση, το κάπνισμα και η ύπαρξη καρδιακής νόσου, στοιχεία που συμβαδίζουν με γνωστούς παράγοντες κινδύνου. Αντίθετα, ορισμένες μεταβλητές εμφανίζουν αρνητική επίδραση, ωστόσο τα αποτελέσματα αυτά πρέπει να ερμηνεύονται με προσοχή, καθώς προκύπτουν από το συγκεκριμένο dataset και δεν αποτελούν από μόνα τους αιτιώδη σχέση.



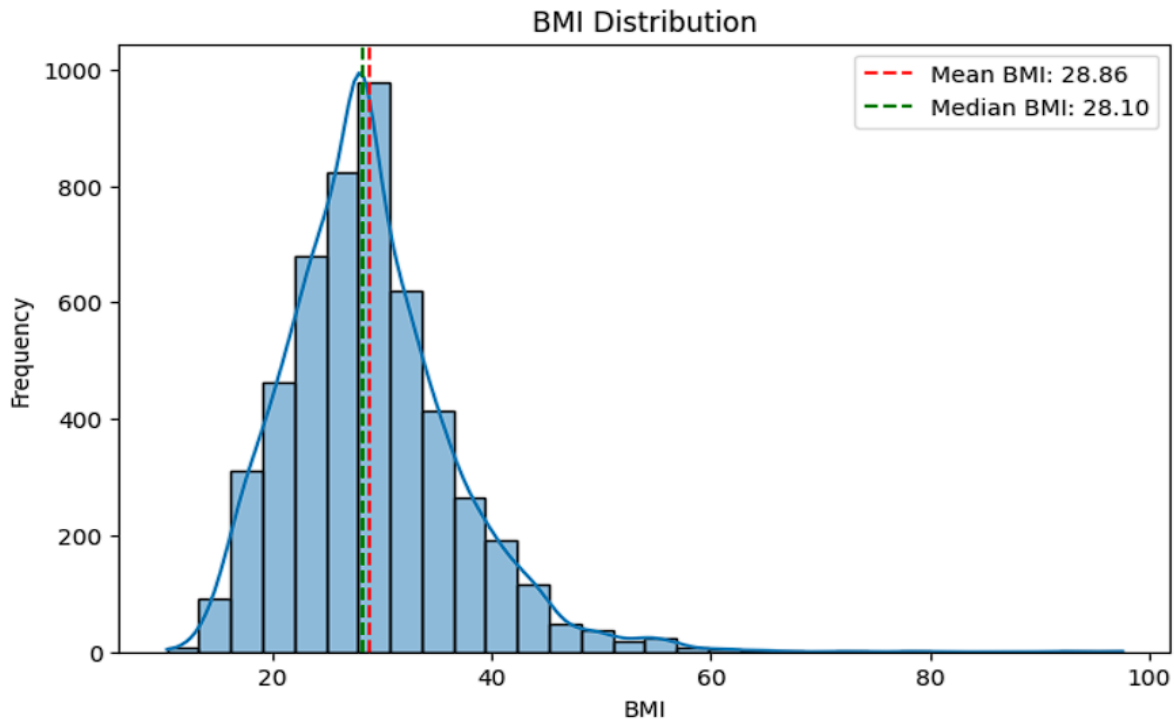
Σχήμα 3.2: Σύγκριση μεταβλητών

Για παράδειγμα, η ηλικία, η υπέρταση, η καρδιακή νόσος, τα επίπεδα γλυκόζης και ο δείκτης μάζας σώματος μπορεί να αλληλεπιδρούν μεταξύ τους με σύνθετο τρόπο. Αυτός είναι ένας ακόμη λόγος για τον οποίο η χρήση μοντέλων μηχανικής μάθησης παρουσιάζει ενδιαφέρον, καθώς τα μοντέλα μπορούν να αξιοποιήσουν συνδυασμούς χαρακτηριστικών για να εντοπίσουν μοτίβα στα δεδομένα. Παράλληλα, η εξερευνητική ανάλυση ανέδειξε την ανάγκη για κατάλληλη προεπεξεργασία πριν από την εκπαίδευση των μοντέλων. Η ύπαρξη κατηγορικών μεταβλητών απαιτεί μετατροπή τους σε αριθμητική μορφή, ώστε να μπορούν να χρησιμοποιηθούν από τους αλγόριθμους μηχανικής μάθησης. Επιπλέον, οι αριθμητικές μεταβλητές βρίσκονται σε διαφορετικές κλίμακες, γεγονός που καθιστά απαραίτητη την εφαρμογή κλιμάκωσης, ειδικά για μοντέλα όπως η Logistic Regression. Η αφαίρεση της στήλης id είναι επίσης αναγκαία, καθώς πρόκειται για μοναδικό αναγνωριστικό και δεν προσφέρει ουσιαστική πληροφορία για την πρόβλεψη. Έτσι, η εξερευνητική ανάλυση δεν λειτουργεί μόνο περιγραφικά, αλλά καθοδηγεί τα επόμενα βήματα της προεπεξεργασίας. Ως γενικό συμπέρασμα, η εξερευνητική ανάλυση επιβεβαιώνει ότι το συγκεκριμένο dataset παρουσιάζει τα βασικά χαρακτηριστικά ενός πραγματικού και απαιτητικού ιατρικού προβλήματος πρόβλεψης. Περιλαμβάνει μεταβλητές που σχετίζονται με γνωστούς παράγοντες κινδύνου, διαθέτει ελλιπείς τιμές που πρέπει να διαχειριστούν κατάλληλα και, κυρίως, παρουσιάζει έντονη ανισορροπία μεταξύ της αρνητικής και της θετικής κλάσης. Η ανισορροπία αυτή αποτελεί τον κεντρικό λόγο για τον οποίο στην παρούσα εργασία εξετάζονται τεχνικές όπως τα βάρη κλάσεων, το SMOTE, οι παραλλαγές του και το CTGAN. Κατά συνέπεια, η ανάλυση του dataset δεν αποτελεί απλώς ένα προκαταρκτικό βήμα, αλλά τεκμηριώνει την ανάγκη για τις μεθοδολογικές επιλογές που ακολουθούν στα επόμενα κεφάλαια.

3.4 Προεπεξεργασία

Η προεπεξεργασία των δεδομένων αποτελεί ένα ιδιαίτερα σημαντικό στάδιο στη διαδικασία ανάπτυξης μοντέλων μηχανικής μάθησης, καθώς επηρεάζει άμεσα την ποιότητα της εκπαίδευσης και την αξιοπιστία των αποτελεσμάτων. Πριν από την εφαρμογή οποιουδήποτε αλγορίθμου είναι απαραίτητο, τα δεδομένα να μετατραπούν σε κατάλληλη μορφή, ώστε να μπορούν να αξιοποιηθούν αποτελεσματικά από τα μοντέλα. Στην παρούσα εργασία, η προεπεξεργασία περιλαμβάνει τη διαχείριση ελλিপών τιμών, την αφαίρεση μη χρήσιμων χαρακτηριστικών, την κωδικοποίηση κατηγορικών μεταβλητών, τον διαχωρισμό των δεδομένων σε σύνολο εκπαίδευσης και ελέγχου, καθώς και την κλιμάκωση των αριθμητικών χαρακτηριστικών. Αρχικά, από το σύνολο δεδομένων αφαιρέθηκε η στήλη *id*. Η συγκεκριμένη στήλη αποτελεί μοναδικό αναγνωριστικό για κάθε εγγραφή και δεν περιέχει κάποια ουσιαστική πληροφορία που να σχετίζεται με την πιθανότητα εμφάνισης εγκεφαλικού επεισοδίου. Η διατήρησή της θα μπορούσε να δημιουργήσει θόρυβο στη διαδικασία εκπαίδευσης, καθώς το μοντέλο ενδέχεται να προσπαθήσει να εντοπίσει μοτίβα σε μια μεταβλητή που στην πραγματικότητα δεν έχει προγνωστική αξία. Έτσι, κρίθηκε απαραίτητη η αφαίρεσή της πριν από την εκπαίδευση των μοντέλων. Η επιλογή αυτή είναι σημαντική, καθώς σε προβλήματα μηχανικής μάθησης δεν αρκεί απλώς να χρησιμοποιούνται όλες οι διαθέσιμες στήλες, αλλά πρέπει να εξετάζεται αν κάθε χαρακτηριστικό προσφέρει πραγματική πληροφορία στο πρόβλημα που μελετάται. Στη συνέχεια, εξετάστηκε η ύπαρξη ελλিপών τιμών στο *dataset*. Από την αρχική διερεύνηση προέκυψε ότι η μεταβλητή *bmi*, η οποία αντιστοιχεί στον δείκτη μάζας σώματος, περιείχε ελλειπείς τιμές. Η μεταβλητή αυτή είναι σημαντική, καθώς ο δείκτης μάζας σώματος μπορεί να σχετίζεται με παράγοντες κινδύνου όπως η παχυσαρκία, οι μεταβολικές διαταραχές και γενικότερα η καρδιαγγειακή υγεία. Έτσι, η απλή αφαίρεση των εγγραφών με ελλειπές *bmi* δεν θεωρήθηκε κατάλληλη επιλογή, καθώς θα οδηγούσε σε απώλεια δεδομένων από ένα ήδη περιορισμένο και έντονα ανισόρροπο σύνολο. Ιδιαίτερα σε ένα πρόβλημα όπου η θετική κλάση, δηλαδή οι περιπτώσεις εγκεφαλικού επεισοδίου, είναι αριθμητικά πολύ μικρή, η διατήρηση όσο το δυνατόν περισσότερων εγγραφών είναι σημαντική.

Για τη διαχείριση των ελλিপών τιμών στη μεταβλητή *bmi*, εφαρμόστηκε συμπλήρωση με τη διάμεση τιμή. Η επιλογή της διαμέσου θεωρήθηκε πιο κατάλληλη, όπως φαίνεται και στο Σχήμα 3.3 παρακάτω, σε σχέση με τον μέσο όρο, επειδή η διάμεσος επηρεάζεται λιγότερο από ακραίες τιμές. Σε ιατρικά δεδομένα, μεταβλητές όπως ο δείκτης μάζας σώματος μπορεί να παρουσιάζουν τιμές που αποκλίνουν σημαντικά από το συνηθισμένο εύρος. Σε τέτοιες περιπτώσεις, ο μέσος όρος μπορεί να μετατοπιστεί από τις ακραίες τιμές και να μην αντιπροσωπεύει ικανοποιητικά την κεντρική τάση των δεδομένων. Αντίθετα, η διάμεσος παρέχει μια πιο σταθερή και ανθεκτική εκτίμηση, γεγονός που την καθιστά κατάλληλη επιλογή για τη συμπλήρωση των ελλিপών τιμών. Κατ' αυτόν τον τρόπο, διατηρήθηκε το σύνολο των εγγραφών χωρίς να αλλοιωθεί υπερβολικά η κατανομή της συγκεκριμένης μεταβλητής.



Σχήμα 3.3: Mean – Median BMI

Μετά τη διαχείριση των ελλিপών τιμών, πραγματοποιήθηκε ο διαχωρισμός των χαρακτηριστικών εισόδου από τη μεταβλητή-στόχο. Η μεταβλητή stroke ορίστηκε ως η εξαρτημένη μεταβλητή, δηλαδή η τιμή που καλούνται να προβλέψουν τα μοντέλα. Όλες οι υπόλοιπες μεταβλητές, μετά την αφαίρεση του id, χρησιμοποιήθηκαν ως χαρακτηριστικά εισόδου. Ο διαχωρισμός αυτός είναι βασικό βήμα σε κάθε πρόβλημα επιβλεπόμενης μάθησης, καθώς καθορίζει ποια δεδομένα θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου και ποια τιμή θα αποτελεί το αποτέλεσμα της πρόβλεψης.

Έπειτα, εφαρμόστηκε κωδικοποίηση των κατηγορικών μεταβλητών. Το dataset περιλαμβάνει αρκετές μη αριθμητικές μεταβλητές, όπως το gender, το ever_married, το work_type, το Residence_type και το smoking_status. Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης δεν μπορούν να επεξεργαστούν άμεσα δεδομένα σε μορφή κειμένου, επομένως οι κατηγορικές αυτές μεταβλητές έπρεπε να μετατραπούν σε αριθμητική μορφή. Για τον σκοπό αυτό χρησιμοποιήθηκε η τεχνική One-Hot Encoding, με την οποία κάθε κατηγορική μεταβλητή μετατρέπεται σε ένα σύνολο δυαδικών στηλών. Κάθε νέα στήλη δηλώνει την παρουσία ή απουσία μιας συγκεκριμένης κατηγορίας. Στην παρούσα εργασία, το One-Hot Encoding εφαρμόστηκε με χρήση της επιλογής drop_first=True. Η επιλογή αυτή αφαιρεί την πρώτη κατηγορία κάθε κατηγορικής μεταβλητής, ώστε να αποφευχθεί η δημιουργία πλήρως εξαρτημένων στηλών. Με αυτόν τον τρόπο περιορίζεται το πρόβλημα της πολυσυγγραμμικότητας, το οποίο μπορεί να επηρεάσει κυρίως γραμμικά μοντέλα, όπως η Logistic Regression. Μετά την εφαρμογή της κωδικοποίησης, το σύνολο δεδομένων μετατράπηκε πλήρως σε αριθμητική μορφή και μπορούσε πλέον να χρησιμοποιηθεί από τα μοντέλα μηχανικής μάθησης. Με βάση την τελική μορφή του dataset, μετά την αφαίρεση του id και την εφαρμογή του One-Hot Encoding, προέκυψαν συνολικά 16 χαρακτηριστικά εισόδου και η μεταβλητή-στόχος stroke.

Στη συνέχεια, τα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Ο διαχωρισμός αυτός πραγματοποιήθηκε με αναλογία 80% για εκπαίδευση και 20% για έλεγχο. Το σύνολο εκπαίδευσης χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων και την εφαρμογή των τεχνικών εξισορρόπησης, ενώ το σύνολο ελέγχου διατηρήθηκε αποκλειστικά για την τελική αξιολόγηση της

απόδοσης. Η ύπαρξη ξεχωριστού test set είναι απαραίτητη, καθώς επιτρέπει την εκτίμηση της ικανότητας γενίκευσης των μοντέλων σε δεδομένα που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση. Ιδιαίτερη σημασία δόθηκε στο γεγονός ότι ο διαχωρισμός πραγματοποιήθηκε με stratification ως προς τη μεταβλητή stroke. Η επιλογή αυτή εξασφαλίζει ότι η αναλογία των δύο κλάσεων διατηρείται περίπου ίδια τόσο στο training set όσο και στο test set. Δεδομένου ότι το dataset είναι έντονα ανισόρροπο, ένας απλός τυχαίος διαχωρισμός θα μπορούσε να οδηγήσει σε ακόμη μικρότερο αριθμό θετικών περιστατικών στο test set ή στο training set, γεγονός που θα επηρέαζε αρνητικά την εκπαίδευση και την αξιολόγηση. Με τη χρήση του stratified split, η κατανομή των κλάσεων παραμένει πιο αντιπροσωπευτική του αρχικού dataset. Συγκεκριμένα, το σύνολο εκπαίδευσης περιλαμβάνει 4.088 εγγραφές, ενώ το σύνολο ελέγχου περιλαμβάνει 1.022 εγγραφές. Στο test set διατηρείται η πραγματική ανισορροπία, με 972 περιπτώσεις χωρίς εγκεφαλικό και 50 περιπτώσεις με εγκεφαλικό. Η διατήρηση του test set σε ανισόρροπη μορφή αποτελεί μια συνειδητή και σημαντική επιλογή της εργασίας. Παρόλο που εφαρμόζονται τεχνικές εξισορρόπησης στο training set, το test set δεν μεταβάλλεται κατά την κύρια αξιολόγηση. Μέσω αυτής της διαδικασίας, η απόδοση των μοντέλων εξετάζεται σε ένα πιο ρεαλιστικό περιβάλλον, όπου τα θετικά περιστατικά είναι πράγματι πολύ λιγότερα από τα αρνητικά. Αν το test set εξισορροπούσαν από την αρχή, η αξιολόγηση θα ήταν λιγότερο αντιπροσωπευτική της πραγματικής κατανομής του προβλήματος. Με αυτή τη λογική, όλες οι βασικές συγκρίσεις των μοντέλων πραγματοποιούνται πάνω στο πραγματικό, ανισόρροπο test set. Η πειραματική εξισορρόπηση του test set εξετάζεται ξεχωριστά σε επόμενη ενότητα και δεν αποτελεί την κύρια αξιολόγηση της εργασίας.

Τέλος, εφαρμόστηκε κλιμάκωση των αριθμητικών χαρακτηριστικών με τη χρήση του StandardScaler. Η κλιμάκωση είναι απαραίτητη κυρίως για μοντέλα όπως η Logistic Regression, τα οποία επηρεάζονται από τη διαφορετική κλίμακα των μεταβλητών. Για παράδειγμα, η ηλικία, το μέσο επίπεδο γλυκόζης και ο δείκτης μάζας σώματος εκφράζονται σε διαφορετικά αριθμητικά εύρη. Αν οι μεταβλητές αυτές χρησιμοποιηθούν χωρίς κλιμάκωση, ένα χαρακτηριστικό με μεγαλύτερες αριθμητικές τιμές μπορεί να επηρεάσει δυσανάλογα τη διαδικασία εκπαίδευσης. Με την εφαρμογή του StandardScaler, κάθε χαρακτηριστικό μετασχηματίζεται έτσι ώστε να έχει μέση τιμή κοντά στο μηδέν και τυπική απόκλιση ίση με τη μονάδα. Σημαντικό είναι ότι η κλιμάκωση εφαρμόστηκε με σωστό τρόπο, ώστε να αποφευχθεί διαρροή πληροφορίας από το test set προς το training set. Συγκεκριμένα, ο scaler προσαρμόστηκε μόνο στα δεδομένα εκπαίδευσης και στη συνέχεια εφαρμόστηκε τόσο στο training set όσο και στο test set. Έτσι, οι παράμετροι της κλιμάκωσης υπολογίζονται αποκλειστικά από τα δεδομένα εκπαίδευσης, ενώ το test set παραμένει πραγματικά ανεξάρτητο κατά τη διαδικασία αξιολόγησης. Η αποφυγή data leakage είναι ιδιαίτερα σημαντική, καθώς οποιαδήποτε πληροφορία από το test set χρησιμοποιηθεί κατά την προετοιμασία του training set μπορεί να οδηγήσει σε υπερεκτίμηση της απόδοσης των μοντέλων.

Συμπερασματικά, η διαδικασία προεπεξεργασίας είχε ως στόχο τη δημιουργία ενός καθαρού, αριθμητικού και κατάλληλα διαμορφωμένου συνόλου δεδομένων για την εκπαίδευση των μοντέλων μηχανικής μάθησης. Η αφαίρεση της στήλης id, η συμπλήρωση των ελλειπόν τιμών του bmi με τη διάμεσο, η κωδικοποίηση των κατηγορικών μεταβλητών, ο διαχωρισμός των δεδομένων και η κλιμάκωση των χαρακτηριστικών αποτέλεσαν βασικά βήματα για την ορθή προετοιμασία του dataset. Μετά την ολοκλήρωση αυτών των βημάτων, το σύνολο δεδομένων είναι κατάλληλο για να χρησιμοποιηθεί στα πειράματα που παρουσιάζονται στα επόμενα κεφάλαια.

3.5 Μετρικές Αξιολόγησης

Η επιλογή των κατάλληλων μετρικών αξιολόγησης αποτελεί ένα από τα πιο σημαντικά στάδια σε κάθε εργασία μηχανικής μάθησης, καθώς οι μετρικές καθορίζουν τον τρόπο με τον οποίο ερμηνεύεται η απόδοση των μοντέλων. Στην παρούσα διπλωματική εργασία, το συγκεκριμένο ζήτημα έχει ιδιαίτερη

σημασία, επειδή το πρόβλημα της πρόβλεψης εγκεφαλικού επεισοδίου βασίζεται σε ένα έντονα ανισόρροπο σύνολο δεδομένων. Όπως αναφέρθηκε στις προηγούμενες ενότητες, η πλειονότητα των εγγραφών αντιστοιχεί σε ασθενείς που δεν έχουν εμφανίσει εγκεφαλικό επεισόδιο, ενώ οι θετικές περιπτώσεις αποτελούν ένα πολύ μικρό ποσοστό του συνολικού dataset. Συγκεκριμένα, η αρνητική κλάση αντιστοιχεί περίπου στο 95% των δεδομένων, ενώ η θετική κλάση, δηλαδή οι περιπτώσεις εγκεφαλικού, αντιστοιχεί περίπου στο 5%. Σε ένα τέτοιο περιβάλλον, η χρήση της συνολικής ακρίβειας, δηλαδή του Accuracy, δεν είναι αρκετή για την αξιόπιστη αξιολόγηση των μοντέλων. Το Accuracy δείχνει το ποσοστό των συνολικά σωστών προβλέψεων ενός μοντέλου, χωρίς όμως να εξετάζει αν οι σωστές προβλέψεις αφορούν την πλειοψηφική ή τη μειοψηφική κλάση. Αυτό μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα. Για παράδειγμα, σε ένα dataset όπου περίπου το 95% των περιπτώσεων δεν αφορά εγκεφαλικό επεισόδιο, ένα μοντέλο που προβλέπει σχεδόν πάντα την αρνητική κλάση μπορεί να εμφανίζει πολύ υψηλό Accuracy, χωρίς όμως να αναγνωρίζει αποτελεσματικά τους ασθενείς που πραγματικά έχουν εμφανίσει εγκεφαλικό. Συνεπώς, παρόλο που η συνολική ακρίβεια μπορεί αριθμητικά να φαίνεται υψηλή, η πρακτική χρησιμότητα του μοντέλου σε ένα ιατρικό πρόβλημα μπορεί να είναι περιορισμένη. Στα αρχικά baseline αποτελέσματα, χωρίς την εφαρμογή κάποιας τεχνικής αντιμετώπισης της ανισορροπίας, τα μοντέλα που χρησιμοποιήθηκαν, δηλαδή το Logistic Regression, το Random Forest και το XGBoost, μπορούσαν να εμφανίζουν υψηλό Accuracy, επειδή αναγνώριζαν σωστά τη μεγάλη πλειονότητα των αρνητικών περιπτώσεων. Ωστόσο, η ικανότητά τους να εντοπίζουν τις θετικές περιπτώσεις ήταν πολύ περιορισμένη. Αυτό δείχνει ότι το Accuracy από μόνο του δεν μπορεί να θεωρηθεί επαρκής δείκτης αξιολόγησης για το συγκεκριμένο πρόβλημα. Στο πλαίσιο της παρούσας ανάλυσης, το σημείο αυτό αποτελεί ένα από τα πιο ουσιαστικά σημεία της εργασίας, καθώς αναδεικνύει ότι ένα μοντέλο δεν πρέπει να αξιολογείται μόνο με βάση το πόσες συνολικά προβλέψεις κάνει σωστά, αλλά κυρίως με βάση το αν εντοπίζει τις περιπτώσεις που έχουν πραγματική σημασία. Με βάση τα παραπάνω, στην παρούσα εργασία χρησιμοποιούνται επιπλέον μετρικές αξιολόγησης, οι οποίες είναι πιο κατάλληλες για προβλήματα με ανισόρροπες κλάσεις. Οι βασικές μετρικές που αξιοποιούνται είναι το Precision, το Recall, το F1-score, το ROC-AUC και το PR-AUC. Παράλληλα, για την καλύτερη ερμηνεία των αποτελεσμάτων χρησιμοποιείται και ο πίνακας σύγχυσης ή confusion matrix, καθώς επιτρέπει την αναλυτική παρατήρηση των σωστών και λανθασμένων προβλέψεων ανά κλάση.

Ο πίνακας σύγχυσης αποτελεί βασικό εργαλείο για την κατανόηση της συμπεριφοράς ενός μοντέλου ταξινόμησης. Στην περίπτωση της δυαδικής ταξινόμησης, οι προβλέψεις χωρίζονται σε τέσσερις κατηγορίες: True Positives, False Positives, True Negatives και False Negatives. Τα True Positives (TP) είναι οι περιπτώσεις όπου το μοντέλο προβλέπει σωστά ότι ένας ασθενής έχει εμφανίσει εγκεφαλικό επεισόδιο. Τα True Negatives (TN) είναι οι περιπτώσεις όπου το μοντέλο προβλέπει σωστά ότι ένας ασθενής δεν έχει εμφανίσει εγκεφαλικό. Τα False Positives (FP) είναι οι περιπτώσεις όπου το μοντέλο προβλέπει λανθασμένα ότι ένας ασθενής έχει εμφανίσει εγκεφαλικό, ενώ στην πραγματικότητα δεν έχει. Τέλος, τα False Negatives (FN) είναι οι περιπτώσεις όπου το μοντέλο προβλέπει λανθασμένα ότι ένας ασθενής δεν έχει εμφανίσει εγκεφαλικό, ενώ στην πραγματικότητα ανήκει στη θετική κλάση. Στο συγκεκριμένο πρόβλημα, τα False Negatives έχουν ιδιαίτερη σημασία, επειδή αντιστοιχούν σε πραγματικά περιστατικά εγκεφαλικού τα οποία το μοντέλο δεν κατάφερε να εντοπίσει. Σε ένα ιατρικό πλαίσιο, αυτό είναι σημαντικότερο από ένα απλό λάθος ταξινόμησης, καθώς η μη αναγνώριση ενός περιστατικού υψηλού κινδύνου μπορεί να οδηγήσει σε καθυστέρηση στην περαιτέρω διερεύνηση ή στην προληπτική αντιμετώπιση. Από την άλλη πλευρά, τα False Positives επίσης δεν είναι αμελητέα, καθώς μπορεί να οδηγήσουν σε περιττή ανησυχία ή επιπλέον εξετάσεις. Παρόλα αυτά, στην παρούσα εργασία δίνεται μεγαλύτερη έμφαση στη μείωση των False Negatives και στην αύξηση της ικανότητας εντοπισμού της θετικής κλάσης, επειδή το ζητούμενο είναι η αναγνώριση των περιστατικών

εγκεφαλικού. Η γενική μορφή του πίνακα σύγχυσης για το πρόβλημα της δυαδικής ταξινόμησης παρουσιάζεται στον Πίνακα 3.2.

Πίνακας 3.2: Γενική μορφή του πίνακα σύγχυσης

Actual classes	Predicted classes	
	No Stroke	Stroke
No Stroke	True Negative (TN)	False Positive (FP)
Stroke	False Negative (FN)	True Positive (TP)

Η μετρική Recall, ή αλλιώς ανάκληση, αποτελεί μία από τις σημαντικότερες μετρικές για την παρούσα εργασία. Το Recall δείχνει το ποσοστό των πραγματικά θετικών περιπτώσεων που κατάφερε να εντοπίσει σωστά το μοντέλο. Με άλλα λόγια, απαντά στο ερώτημα, πόσους ασθενείς κατάφερε να αναγνωρίσει το μοντέλο, από όλους τους ασθενείς που πράγματι έχουν εμφανίσει εγκεφαλικό επεισόδιο. Η μαθηματική του μορφή είναι [13]:

$$Recall = \frac{TP}{TP + FN} \quad (3.1)$$

Η συγκεκριμένη μετρική είναι ιδιαίτερα κρίσιμη στο πρόβλημα της πρόβλεψης εγκεφαλικού επεισοδίου, επειδή δίνει άμεση εικόνα για την ικανότητα του μοντέλου να εντοπίζει τη μειοψηφική κλάση. Ένα χαμηλό Recall σημαίνει ότι το μοντέλο χάνει πολλά πραγματικά θετικά περιστατικά, ακόμη κι αν εμφανίζει υψηλό Accuracy. Για τον λόγο αυτό, η ανάκληση χρησιμοποιείται ως βασικός δείκτης αξιολόγησης σε όλα τα πειράματα της εργασίας.

Η μετρική Precision, ή αλλιώς ακρίβεια θετικών προβλέψεων, δείχνει το ποσοστό των προβλέψεων που ταξινομήθηκαν ως θετικές και ήταν πράγματι θετικές. Πιο συγκεκριμένα η μετρική αυτή μας δείχνει, πόσα ήταν όντως θετικά περιστατικά από όλους τους ασθενείς που το μοντέλο προέβλεψε ως περιστατικά εγκεφαλικού. Η μαθηματική συνάρτηση της είναι [13]:

$$Precision = \frac{TP}{TP + FP} \quad (3.2)$$

Το Precision είναι σημαντικό επειδή δείχνει πόσο αξιόπιστες είναι οι θετικές προβλέψεις του μοντέλου. Σε περίπτωση που ένα μοντέλο αυξάνει πολύ το Recall, αλλά ταυτόχρονα παράγει πάρα πολλά False Positives, τότε το Precision μπορεί να μειωθεί σημαντικά. Αυτό σημαίνει ότι το μοντέλο εντοπίζει περισσότερα πραγματικά περιστατικά, αλλά ταυτόχρονα χαρακτηρίζει λανθασμένα πολλούς ασθενείς ως θετικούς. Άρα, το Precision βοηθά στην κατανόηση της ισορροπίας μεταξύ ευαισθησίας και αξιοπιστίας των θετικών προβλέψεων.

Για να συνδυαστούν οι πληροφορίες του Precision και του Recall, χρησιμοποιείται το F1-score. Το F1-score αποτελεί τον αρμονικό μέσο των δύο αυτών μετρικών και παρέχει μία συνολική εικόνα της ισορροπίας μεταξύ τους. Η μαθηματική του μορφή είναι:

$$F1 - score = 2 \frac{Precision \times Recall}{Precision + Recall} \quad (3.3)$$

Η συγκεκριμένη μετρική είναι χρήσιμη σε περιπτώσεις όπου δεν αρκεί να εξετάζεται μόνο το Recall ή μόνο το Precision. Για παράδειγμα, ένα μοντέλο μπορεί να πετυχαίνει υψηλό Recall, επειδή προβλέπει πολλές περιπτώσεις ως θετικές, αλλά ταυτόχρονα να έχει χαμηλό Precision λόγω πολλών λανθασμένων

θετικών προβλέψεων. Αντίστοιχα, ένα μοντέλο μπορεί να έχει υψηλό Precision αλλά χαμηλό Recall, αν προβλέπει ως θετικές μόνο λίγες περιπτώσεις με μεγάλη βεβαιότητα. Το F1-score βοηθά να αξιολογηθεί πιο συνολικά αυτή η ισορροπία. Στην παρούσα εργασία, το F1-score χρησιμοποιείται ως συμπληρωματική μετρική, ώστε να εξεταστεί αν η αύξηση του Recall συνοδεύεται από αποδεκτή συνολική συμπεριφορά του μοντέλου.

Εκτός από τις παραπάνω μετρικές, χρησιμοποιείται και το ROC-AUC, το οποίο εκφράζει την επιφάνεια κάτω από την καμπύλη ROC. Η καμπύλη ROC δείχνει τη σχέση μεταξύ του True Positive Rate και του False Positive Rate για διαφορετικά κατώφλια ταξινόμησης [14]. Το ROC-AUC μπορεί να δώσει μια γενική εικόνα της ικανότητας του μοντέλου να διαχωρίζει τις δύο κλάσεις. Ωστόσο, σε έντονα ανισόρροπα δεδομένα, το ROC-AUC μπορεί επίσης να δώσει πιο αισιόδοξη εικόνα από την πραγματική απόδοση στη θετική κλάση, επειδή επηρεάζεται από τον μεγάλο αριθμό των True Negatives. Η παρατήρηση αυτή μας οδηγεί στο συμπέρασμα ότι, το ROC-AUC δεν πρέπει να αγνοείται, αλλά δεν αποτελεί τη βασική μετρική πάνω στην οποία στηρίζεται η ερμηνεία των αποτελεσμάτων.

Αντίθετα, ιδιαίτερη σημασία δίνεται στο PR-AUC, δηλαδή στην επιφάνεια κάτω από την καμπύλη Precision-Recall. Η καμπύλη Precision-Recall εξετάζει τη σχέση μεταξύ Precision και Recall για διαφορετικά κατώφλια απόφασης και θεωρείται ιδιαίτερα χρήσιμη σε προβλήματα όπου η θετική κλάση είναι σπάνια [14]. Στην παρούσα εργασία, το PR-AUC είναι σημαντικό επειδή εστιάζει περισσότερο στη συμπεριφορά του μοντέλου απέναντι στη θετική κλάση, δηλαδή στα περιστατικά εγκεφαλικού. Σε αντίθεση με το Accuracy ή ακόμη και το ROC-AUC, το PR-AUC μπορεί να αποτυπώσει καλύτερα αν ένα μοντέλο είναι πράγματι ικανό να εντοπίζει τη μειοψηφική κλάση χωρίς να παράγει υπερβολικά πολλές λανθασμένες θετικές προβλέψεις.

Η επιλογή των Recall, F1-score και PR-AUC ως βασικών μετρικών αξιολόγησης συνδέεται άμεσα με τη φύση του προβλήματος. Στην πρόβλεψη εγκεφαλικού επεισοδίου, το ζητούμενο δεν είναι απλώς να μεγιστοποιηθεί η συνολική ακρίβεια, αλλά να εξεταστεί αν τα μοντέλα μπορούν να εντοπίσουν αποτελεσματικά τα πραγματικά θετικά περιστατικά. Ένα μοντέλο που προβλέπει σωστά τη μεγάλη πλειονότητα των αρνητικών περιπτώσεων αλλά αποτυγχάνει να εντοπίσει τα περιστατικά εγκεφαλικού δεν μπορεί να θεωρηθεί επαρκές για το συγκεκριμένο πρόβλημα. Επομένως, η αξιολόγηση πρέπει να εστιάζει στη μειοψηφική κλάση και όχι μόνο στη συνολική συμπεριφορά του μοντέλου.

Στην παρούσα εργασία, το Accuracy εξακολουθεί να παρουσιάζεται στους πίνακες αποτελεσμάτων, ώστε να υπάρχει μια συνολική εικόνα της απόδοσης κάθε μοντέλου. Ωστόσο, η ερμηνεία των αποτελεσμάτων δεν βασίζεται κυρίως σε αυτό. Αντίθετα, δίνεται μεγαλύτερη βαρύτητα στο Recall, επειδή δείχνει την ικανότητα εντοπισμού των πραγματικών περιστατικών εγκεφαλικού, στο F1-score, επειδή αποτυπώνει την ισορροπία μεταξύ σωστών θετικών προβλέψεων και κάλυψης της θετικής κλάσης, και στο PR-AUC, επειδή είναι πιο κατάλληλο για περιβάλλοντα με έντονη ανισορροπία.

Λαμβάνοντας υπόψη τα παραπάνω, η επιλογή των μετρικών αξιολόγησης στην παρούσα εργασία έγινε με βάση τη φύση του προβλήματος και όχι με βάση μια γενική λογική μέτρησης της απόδοσης. Η πρόβλεψη εγκεφαλικού επεισοδίου αποτελεί ένα πρόβλημα όπου η θετική κλάση είναι σπάνια αλλά ιδιαίτερα σημαντική. Για αυτόν τον λόγο, οι μετρικές που επιλέχθηκαν επιτρέπουν μια πιο ουσιαστική και ρεαλιστική αξιολόγηση των μοντέλων. Η προσέγγιση αυτή βοηθά ώστε, η ανάλυση των αποτελεσμάτων στα επόμενα κεφάλαια δεν περιορίζεται στο αν ένα μοντέλο εμφανίζει υψηλή συνολική ακρίβεια, αλλά εξετάζει αν μπορεί να συμβάλει πραγματικά στον εντοπισμό των περιπτώσεων εγκεφαλικού επεισοδίου, που αποτελεί και τον βασικό στόχο της παρούσας διπλωματικής εργασίας.

Κεφάλαιο 4ο: Μεθοδολογία & Πειράματα

4.1 Περιγραφή Μοντέλων

Στο πλαίσιο της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκαν τρία διαφορετικά μοντέλα μηχανικής μάθησης για την πρόβλεψη του εγκεφαλικού επεισοδίου. Η *Logistic Regression*, το *Random Forest* και το *XGBoost*. Η επιλογή των συγκεκριμένων μοντέλων έγινε με στόχο να αξιολογηθούν διαφορετικές προσεγγίσεις ταξινόμησης, από ένα πιο απλό και ερμηνεύσιμο γραμμικό μοντέλο μέχρι πιο σύνθετες μεθόδους που βασίζονται σε δέντρα αποφάσεων και τεχνικές ensemble learning [24]. Με αυτή τη λογική, η πειραματική διαδικασία δεν περιορίζεται σε έναν μόνο αλγόριθμο, αλλά εξετάζει πώς διαφορετικοί τύποι μοντέλων ανταποκρίνονται στο ίδιο πρόβλημα πρόβλεψης, ειδικά όταν τα δεδομένα παρουσιάζουν έντονη ανισορροπία μεταξύ των κλάσεων.

Το πρόβλημα που εξετάζεται είναι πρόβλημα δυαδικής ταξινόμησης, καθώς η μεταβλητή στόχος (stroke) λαμβάνει δύο πιθανές τιμές, την τιμή 0 όταν ο ασθενής δεν έχει εμφανίσει εγκεφαλικό επεισόδιο, και την τιμή 1, όταν ο ασθενής έχει εμφανίσει εγκεφαλικό επεισόδιο. Επομένως, κάθε μοντέλο εκπαιδεύεται με στόχο να διαχωρίσει τις δύο αυτές κατηγορίες με βάση τα διαθέσιμα χαρακτηριστικά του dataset, όπως η ηλικία, η υπέρταση, η καρδιακή νόσος, το μέσο επίπεδο γλυκόζης, ο δείκτης μάζας σώματος, η καπνιστική συνήθεια και άλλα δημογραφικά ή συμπεριφορικά δεδομένα. Η βασική πρόκληση, όμως, είναι ότι η θετική κλάση, δηλαδή τα περιστατικά εγκεφαλικού, αντιπροσωπεύει πολύ μικρότερο ποσοστό του συνόλου δεδομένων σε σχέση με την αρνητική κλάση. *Με βάση τα παραπάνω, τα μοντέλα δεν αξιολογούνται μόνο ως προς τη συνολική τους ακρίβεια, αλλά κυρίως ως προς την ικανότητά τους να εντοπίζουν τη μειοψηφική κλάση.*

Το πρώτο μοντέλο που χρησιμοποιήθηκε είναι η Logistic Regression. Παρά το όνομά της, η Logistic Regression αποτελεί αλγόριθμο ταξινόμησης και όχι απλής παλινδρόμησης. Χρησιμοποιείται συχνά σε προβλήματα δυαδικής ταξινόμησης, επειδή υπολογίζει την πιθανότητα μια παρατήρηση να ανήκει σε μία από τις δύο κλάσεις. Στην περίπτωση της παρούσας εργασίας, το μοντέλο εκτιμά την πιθανότητα ένας ασθενής να ανήκει στην κλάση εμφάνισης εγκεφαλικού επεισοδίου. Η LR βασίζεται σε έναν γραμμικό συνδυασμό των χαρακτηριστικών εισόδου και στη συνέχεια εφαρμόζει τη λογιστική συνάρτηση, ώστε η έξοδος να εκφράζεται ως πιθανότητα μεταξύ 0 και 1. Ένα βασικό πλεονέκτημα της Logistic Regression είναι η απλότητα και η ερμηνευσιμότητά της. Σε προβλήματα υγείας, η δυνατότητα ερμηνείας ενός μοντέλου είναι ιδιαίτερα σημαντική, καθώς δεν αρκεί ένα μοντέλο να παράγει προβλέψεις, αλλά είναι χρήσιμο να μπορεί να εξηγηθεί σε κάποιο βαθμό και ο τρόπος με τον οποίο λαμβάνει τις αποφάσεις του. Η LR επιτρέπει την κατανόηση της επίδρασης των χαρακτηριστικών στην τελική πρόβλεψη, γεγονός που την καθιστά κατάλληλη ως βασικό σημείο αναφοράς. Παράλληλα, επειδή είναι σχετικά απλό μοντέλο, μπορεί να λειτουργήσει ως baseline για τη σύγκριση με πιο σύνθετους αλγόριθμους [15]. Ωστόσο, η Logistic Regression έχει και περιορισμούς καθώς βασίζεται σε γραμμική σχέση μεταξύ των χαρακτηριστικών και της πιθανότητας ταξινόμησης. Αυτό έχει ως αποτέλεσμα, ενδεχόμενη δυσκολία στο να αποτυπώσει πιο σύνθετες μη γραμμικές σχέσεις που πιθανώς υπάρχουν στα δεδομένα. Στην περίπτωση της πρόβλεψης εγκεφαλικού επεισοδίου, οι παράγοντες κινδύνου μπορεί να αλληλεπιδρούν μεταξύ τους με πολύπλοκο τρόπο. Για παράδειγμα, η ηλικία, η υπέρταση, το επίπεδο γλυκόζης και ο δείκτης μάζας σώματος μπορεί να επηρεάζουν συνδυαστικά την πιθανότητα εμφάνισης εγκεφαλικού. Παρ' όλα αυτά, η LR παραμένει σημαντική για την εργασία, επειδή είναι ιδιαίτερα κατάλληλη για την υλοποίηση Federated Learning, καθώς αποτελεί ένα απλό μοντέλο με μικρό αριθμό παραμέτρων και χαμηλές υπολογιστικές απαιτήσεις. Αυτό διευκολύνει την εκπαίδευσή της σε διαφορετικούς clients και κάνει πιο άμεση τη διαδικασία συνδυασμού των τοπικών

βαρών μέσω του αλγορίθμου FedAvg. Για τον λόγο αυτό, χρησιμοποιείται στην παρούσα εργασία ως βασικό μοντέλο για την προσομοίωση της κατανεμημένης εκπαίδευσης.

Το δεύτερο μοντέλο που χρησιμοποιήθηκε είναι το Random Forest όπου αποτελεί μέθοδο ensemble learning και βασίζεται στη δημιουργία πολλών δέντρων αποφάσεων. Κάθε δέντρο εκπαιδεύεται σε διαφορετικό υποσύνολο των δεδομένων και των χαρακτηριστικών, και η τελική πρόβλεψη προκύπτει από τον συνδυασμό των προβλέψεων όλων των δέντρων. Στα προβλήματα ταξινόμησης, η τελική απόφαση συνήθως προκύπτει μέσω πλειοψηφικής ψηφοφορίας μεταξύ των δέντρων [16]. Η βασική ιδέα πίσω από το Random Forest είναι ότι ο συνδυασμός πολλών επιμέρους μοντέλων μπορεί να οδηγήσει σε πιο σταθερή και αξιόπιστη πρόβλεψη σε σχέση με ένα μεμονωμένο δέντρο αποφάσεων.

Ένα σημαντικό πλεονέκτημα του Random Forest είναι ότι μπορεί να αποτυπώσει μη γραμμικές σχέσεις και αλληλεπιδράσεις μεταξύ των χαρακτηριστικών. Σε αντίθεση με τη Logistic Regression, δεν περιορίζεται σε γραμμικούς συνδυασμούς των μεταβλητών, αλλά μπορεί να διαχωρίσει τα δεδομένα με πιο σύνθετο τρόπο. Αυτό είναι ιδιαίτερα χρήσιμο σε ιατρικά δεδομένα, όπου οι σχέσεις μεταξύ των παραγόντων κινδύνου και της εμφάνισης μιας νόσου δεν είναι πάντα απλές ή γραμμικές. Επιπλέον, το Random Forest είναι σχετικά ανθεκτικό στον θόρυβο και μπορεί να διαχειριστεί αποτελεσματικά διαφορετικούς τύπους χαρακτηριστικών μετά την κατάλληλη προεπεξεργασία. Παρόλα αυτά, το Random Forest μπορεί να αντιμετωπίσει δυσκολίες όταν η θετική κλάση είναι πολύ μικρή σε σχέση με την αρνητική. Σε ένα έντονα ανισόρροπο dataset, όπως αυτό που χρησιμοποιείται στην παρούσα εργασία, υπάρχει κίνδυνος τα περισσότερα δέντρα να μάθουν κυρίως τη συμπεριφορά της πλειοψηφικής κλάσης. Αυτό μπορεί να οδηγήσει σε υψηλή συνολική ακρίβεια, αλλά χαμηλή ικανότητα εντοπισμού των περιστατικών εγκεφαλικού. Για αυτόν τον λόγο, το Random Forest αξιολογείται όχι μόνο στην αρχική του μορφή, αλλά και σε συνδυασμό με τεχνικές αντιμετώπισης της ανισορροπίας, όπως η χρήση class weights, το SMOTE, οι παραλλαγές του και τα συνθετικά δεδομένα που παράγονται μέσω CTGAN.

Το τρίτο μοντέλο που χρησιμοποιήθηκε είναι το XGBoost το οποίο αποτελεί έναν από τους πιο δημοφιλείς και ισχυρούς αλγορίθμους gradient boosting, ιδιαίτερα σε προβλήματα ταξινόμησης με δομημένα δεδομένα. Σε αντίθεση με το Random Forest, όπου τα δέντρα δημιουργούνται σε μεγάλο βαθμό ανεξάρτητα μεταξύ τους, στο XGBoost τα δέντρα δημιουργούνται διαδοχικά. Κάθε νέο δέντρο προσπαθεί να διορθώσει τα λάθη των προηγούμενων, με αποτέλεσμα το μοντέλο να βελτιώνεται σταδιακά κατά τη διάρκεια της εκπαίδευσης. Η διαδικασία αυτή επιτρέπει στο XGBoost να πετυχαίνει υψηλή προβλεπτική απόδοση σε πολλά πρακτικά προβλήματα.

Το XGBoost παρουσιάζει ιδιαίτερο ενδιαφέρον για την παρούσα εργασία, επειδή μπορεί να μοντελοποιήσει σύνθετες σχέσεις μεταξύ των χαρακτηριστικών και να χειριστεί αποτελεσματικά μη γραμμικές αλληλεπιδράσεις. Επιπλέον, προσφέρει δυνατότητες ρύθμισης που είναι χρήσιμες σε προβλήματα ανισορροπίας, όπως η παράμετρος 'scale_pos_weight', μέσω της οποίας μπορεί να δοθεί μεγαλύτερη βαρύτητα στη μειοψηφική κλάση. Αυτό είναι σημαντικό στο συγκεκριμένο πρόβλημα, καθώς ο βασικός στόχος δεν είναι μόνο να επιτευχθεί υψηλή συνολική απόδοση, αλλά να βελτιωθεί η ικανότητα του μοντέλου να εντοπίζει τα πραγματικά περιστατικά εγκεφαλικού. Ένα ακόμη πλεονέκτημα του XGBoost είναι ότι συνήθως παρουσιάζει υψηλή απόδοση σε πίνακες δεδομένων με αριθμητικά και κωδικοποιημένα κατηγορικά χαρακτηριστικά, όπως το dataset που χρησιμοποιείται στην εργασία. Μετά την εφαρμογή της προεπεξεργασίας και του One-Hot Encoding, τα δεδομένα βρίσκονται σε μορφή κατάλληλη για την εκπαίδευση του συγκεκριμένου αλγορίθμου. Παράλληλα, το XGBoost περιλαμβάνει μηχανισμούς κανονικοποίησης και ελέγχου της πολυπλοκότητας του μοντέλου, οι οποίοι μπορούν να περιορίσουν τον κίνδυνο υπερπροσαρμογής στα δεδομένα εκπαίδευσης [17].

Ωστόσο, όπως και τα υπόλοιπα μοντέλα, το XGBoost δεν είναι απαλλαγμένο από περιορισμούς. Η υψηλή του απόδοση εξαρτάται σε σημαντικό βαθμό από την κατάλληλη ρύθμιση των παραμέτρων του και από την ποιότητα των δεδομένων εκπαίδευσης. Σε έντονα ανισόρροπα προβλήματα, υπάρχει πάντα ο κίνδυνος το μοντέλο να δώσει μεγαλύτερη έμφαση στην πλειοψηφική κλάση, αν δεν εφαρμοστούν κατάλληλες τεχνικές αντιμετώπισης της ανισορροπίας. Συνεπώς, στην παρούσα εργασία το XGBoost αξιολογείται τόσο σε baseline μορφή όσο και σε συνδυασμό με διαφορετικές τεχνικές εξισορρόπησης, ώστε να εξεταστεί πώς μεταβάλλεται η συμπεριφορά του.

Η χρήση των τριών αυτών μοντέλων επιτρέπει μια πιο ολοκληρωμένη αξιολόγηση της μεθοδολογίας. Η Logistic Regression προσφέρει ένα απλό και ερμηνεύσιμο μοντέλο αναφοράς, το Random Forest εισάγει μια πιο σύνθετη προσέγγιση βασισμένη σε πολλά δέντρα αποφάσεων, ενώ το XGBoost αντιπροσωπεύει μια πιο προηγμένη τεχνική boosting με υψηλή προβλεπτική ικανότητα. Με τον τρόπο αυτό, είναι δυνατό να εξεταστεί αν οι τεχνικές εξισορρόπησης και τα συνθετικά δεδομένα επηρεάζουν με τον ίδιο τρόπο όλα τα μοντέλα ή αν η αποτελεσματικότητά τους εξαρτάται από τον αλγόριθμο που χρησιμοποιείται. Στην πειραματική διαδικασία, κάθε μοντέλο αξιολογείται αρχικά χωρίς κάποια τεχνική παρέμβαση στην ανισορροπία των δεδομένων, ώστε να δημιουργηθεί ένα baseline αποτέλεσμα. Στη συνέχεια, τα ίδια μοντέλα αξιολογούνται με χρήση class weights, με τεχνικές υπερδειγματοληψίας όπως το SMOTE, το Borderline-SMOTE και το SMOTE-Tomek, καθώς και με συνθετικά δεδομένα που παράγονται μέσω CTGAN. Η προσέγγιση αυτή συμβάλλει στη συστηματική και δίκαιη σύγκριση των αποτελεσμάτων, καθώς κάθε τεχνική εφαρμόζεται στα ίδια μοντέλα και η αξιολόγηση πραγματοποιείται στο ίδιο πραγματικό test set.

Γενικά, η επιλογή της Logistic Regression, του Random Forest και του XGBoost εξυπηρετεί τον βασικό στόχο της εργασίας, ο οποίος είναι η αξιολόγηση διαφορετικών τεχνικών εκπαίδευσης και εξισορρόπησης για την πρόβλεψη του εγκεφαλικού επεισοδίου. Τα τρία μοντέλα αντιπροσωπεύουν διαφορετικά επίπεδα πολυπλοκότητας και διαφορετικούς τρόπους μάθησης από τα δεδομένα. Γι' αυτό, η σύγκρισή τους επιτρέπει την καλύτερη κατανόηση του τρόπου με τον οποίο η ανισορροπία των κλάσεων και οι τεχνικές αντιμετώπισής της επηρεάζουν την τελική απόδοση σε ένα ιατρικό πρόβλημα δυαδικής ταξινόμησης.

4.2 Πειράματα με Τεχνικές Εξισορρόπησης

Στο πειραματικό μέρος της παρούσας εργασίας, ο βασικός στόχος ήταν να αξιολογηθεί η επίδραση διαφορετικών τεχνικών αντιμετώπισης της ανισορροπίας στην απόδοση μοντέλων μηχανικής μάθησης για την πρόβλεψη του εγκεφαλικού επεισοδίου. Όπως αναδείχθηκε στο προηγούμενο κεφάλαιο, το σύνολο δεδομένων παρουσιάζει έντονη ανισορροπία μεταξύ των δύο κλάσεων της μεταβλητής stroke, καθώς οι περιπτώσεις μη εμφάνισης εγκεφαλικού είναι πολύ περισσότερες από τις περιπτώσεις εμφάνισης εγκεφαλικού. Το γεγονός αυτό καθιστά δύσκολη την εκπαίδευση μοντέλων που να μπορούν να αναγνωρίζουν αποτελεσματικά τη θετική κλάση, δηλαδή τα περιστατικά εγκεφαλικού επεισοδίου. Λόγω αυτού, η ΔΕ δεν περιορίστηκε στην εκπαίδευση των μοντέλων πάνω στο αρχικό ανισόρροπο σύνολο δεδομένων. Αντίθετα, προτιμήθηκε μια σταδιακή πειραματική διαδικασία, με σκοπό να εξεταστεί πώς επηρεάζεται κάθε μοντέλο όταν εφαρμόζονται διαφορετικές τεχνικές εξισορρόπησης. Η διαδικασία αυτή περιλάμβανε αρχικά την εκπαίδευση baseline μοντέλων, χωρίς καμία παρέμβαση στην κατανομή των κλάσεων. Στη συνέχεια εφαρμόστηκαν τεχνικές που επηρεάζουν τον τρόπο εκπαίδευσης του μοντέλου, όπως η χρήση βαρών κλάσεων, καθώς και τεχνικές που μεταβάλλουν το training set, όπως το SMOTE, το Borderline-SMOTE, το SMOTE-Tomek και η παραγωγή συνθετικών δεδομένων μέσω CTGAN.

Η βασική αρχή που ακολουθήθηκε σε όλα τα πειράματα ήταν ότι οι τεχνικές εξισορρόπησης εφαρμόστηκαν αποκλειστικά στο σύνολο εκπαίδευσης και όχι στο κύριο σύνολο ελέγχου. Το test set διατηρήθηκε στην αρχική, πραγματική και ανισόρροπη μορφή του, ώστε η αξιολόγηση των μοντέλων να πραγματοποιηθεί σε μια κατανομή που προσεγγίζει περισσότερο ένα ρεαλιστικό ιατρικό περιβάλλον. Η εφαρμογή της μεθόδου αυτής θεωρήθηκε ιδιαίτερα σημαντική, καθώς η εξισορρόπηση του test set πριν από την κύρια αξιολόγηση θα μπορούσε να οδηγήσει σε υπερβολικά αισιόδοξα και λιγότερο αντιπροσωπευτικά αποτελέσματα.

Συνεπώς, το test set χρησιμοποιήθηκε ως σταθερό σημείο αξιολόγησης για όλα τα πειράματα. Μετά τον διαχωρισμό των δεδομένων σε training και test set, το test set περιλάμβανε 1.022 εγγραφές, από τις οποίες οι 972 αντιστοιχούσαν στην αρνητική κλάση και οι 50 στη θετική κλάση. Η διατήρηση αυτής της κατανομής ήταν απαραίτητη, διότι ο στόχος της εργασίας δεν ήταν να αξιολογηθεί η απόδοση των μοντέλων σε τεχνητά ισορροπημένες συνθήκες, αλλά να εξεταστεί κατά πόσο μπορούν να εντοπίσουν πραγματικά περιστατικά εγκεφαλικού σε ένα περιβάλλον όπου αυτά εμφανίζονται σπάνια.

Για την καλύτερη οργάνωση της πειραματικής διαδικασίας, στον Πίνακα 4.1 παρουσιάζεται συνοπτικά το σύνολο των πειραματικών σεναρίων που εξετάστηκαν στην παρούσα ΔΕ. Ο πίνακας περιλαμβάνει την τεχνική εξισορρόπησης που εφαρμόστηκε, τα μοντέλα στα οποία χρησιμοποιήθηκε, τις βασικές παραμέτρους κάθε προσέγγισης και τον κύριο στόχο του κάθε πειράματος.

Πίνακας 4.1: Σύνοψη πειραματικών σεναρίων και τεχνικών εξισορρόπησης.

ID Πειράματος	Τεχνική Εξισορρόπησης	Μοντέλα	Παράμετροι	Στόχος
1	Baseline	LR, RF, XGBoost	Χωρίς εξισορρόπηση	Καταγραφή της αρχικής συμπεριφοράς των μοντέλων
2	Class Weights	LR, RF, XGBoost	class_weight='balanced', scale_pos_weight	Ενίσχυση της μειοψηφικής κλάσης μέσω μεγαλύτερου βάρους στα θετικά δείγματα
3	SMOTE	LR, RF, XGBoost	Sampling ratios: 0.5, 0.7, 1.0	Δημιουργία συνθετικών δειγμάτων για τη θετική κλάση στο training set
4	Borderline-SMOTE	LR, RF, XGBoost	Sampling ratio: 0.7	Ενίσχυση των πιο δύσκολων θετικών δειγμάτων κοντά στο όριο απόφασης
5	SMOTE-Tomek	LR, RF, XGBoost	Sampling ratios: 0.7	Υπερδειγματοληψία και καθαρισμός θορύβου των δειγμάτων
6	CTGAN	LR, RF, XGBoost	Ratios: 0.7, 0.5, 0.3, Epochs: 30	Παραγωγή συνθετικών δεδομένων για τη μειοψηφική κλάση μέσω generative model

4.2.1 Baseline αποτελέσματα

Το πρώτο στάδιο της πειραματικής διαδικασίας ήταν η εκπαίδευση των μοντέλων χωρίς καμία τεχνική αντιμετώπισης της ανισορροπίας. Το στάδιο αυτό ήταν απαραίτητο, διότι παρείχε ένα αρχικό σημείο αναφοράς, με βάση το οποίο μπορούσαν να συγκριθούν όλες οι επόμενες τεχνικές. Με άλλα λόγια, πριν εφαρμοστεί οποιαδήποτε παρέμβαση στα δεδομένα ή στη διαδικασία μάθησης, έπρεπε πρώτα να καταγραφεί η φυσική συμπεριφορά των μοντέλων πάνω στο αρχικό training set.

Στο baseline στάδιο εκπαιδεύτηκαν τα τρία μοντέλα που παρουσιάστηκαν στην προηγούμενη ενότητα, τα οποία είναι, η Logistic Regression, το Random Forest και το XGBoost. Κάθε μοντέλο εκπαιδεύτηκε χρησιμοποιώντας το αρχικό training set, όπως αυτό προέκυψε μετά την προεπεξεργασία, χωρίς χρήση class weights, χωρίς oversampling και χωρίς παραγωγή συνθετικών δεδομένων. Η αξιολόγηση πραγματοποιήθηκε στο πραγματικό test set, το οποίο παρέμεινε ανισόρροπο.

Η λογική πίσω από τα baseline πειράματα ήταν να διαπιστωθεί αν τα μοντέλα μπορούν, χωρίς κάποια επιπλέον βοήθεια, να μάθουν επαρκώς τα χαρακτηριστικά της μειοψηφικής κλάσης. Σε προβλήματα όπως η πρόβλεψη εγκεφαλικού επεισοδίου, αυτό είναι ιδιαίτερα σημαντικό, επειδή ένα μοντέλο μπορεί να επιτύχει υψηλή συνολική ακρίβεια προβλέποντας κυρίως την πλειοψηφική κλάση. Επομένως, τα baseline πειράματα δεν είχαν ως στόχο να δώσουν απαραίτητα την καλύτερη απόδοση, αλλά να δείξουν ποια είναι η αρχική δυσκολία του προβλήματος.

Στην πράξη, τα baseline αποτελέσματα χρησιμοποιήθηκαν ως βάση σύγκρισης για όλα τα επόμενα πειράματα. Ως αποτέλεσμα, κάθε βελτίωση ή μεταβολή στην απόδοση των μοντέλων μπορούσε να αποδοθεί πιο καθαρά στην αντίστοιχη τεχνική που εφαρμόστηκε. Για παράδειγμα, αν ένα μοντέλο εμφάνιζε καλύτερη ικανότητα εντοπισμού της θετικής κλάσης μετά την εφαρμογή SMOTE ή class weights, αυτή η διαφορά μπορούσε να συγκριθεί με τη baseline συμπεριφορά του ίδιου μοντέλου.

4.2.2 Χρήση Class Weights

Μετά την εκπαίδευση των baseline μοντέλων, εφαρμόστηκε η τεχνική της στάθμισης των κλάσεων, δηλαδή η χρήση class weights. Η τεχνική αυτή αποτελεί έναν τρόπο αντιμετώπισης της ανισορροπίας χωρίς να μεταβάλλεται το ίδιο το σύνολο δεδομένων. Αντί να δημιουργηθούν νέα δείγματα ή να αφαιρεθούν υπάρχοντα, η ανισορροπία αντιμετωπίζεται μέσα από τη διαδικασία εκπαίδευσης του μοντέλου.

Η βασική ιδέα της τεχνικής είναι ότι τα σφάλματα στη μειοψηφική κλάση πρέπει να έχουν μεγαλύτερο κόστος από τα σφάλματα στην πλειοψηφική κλάση. Στο συγκεκριμένο πρόβλημα, αυτό σημαίνει ότι το μοντέλο πρέπει να δώσει μεγαλύτερη βαρύτητα στα δείγματα όπου stroke = 1, επειδή αυτά είναι λιγότερα αλλά πιο σημαντικά για τον στόχο της πρόβλεψης. Με αυτόν τον τρόπο, κατά την εκπαίδευση, το μοντέλο δεν αντιμετωπίζει όλες τις εγγραφές ως ισοδύναμες, αλλά λαμβάνει υπόψη τη δυσανάλογη κατανομή των δύο κλάσεων.

Στην Logistic Regression εφαρμόστηκε η επιλογή class_weight='balanced'. Η επιλογή αυτή υπολογίζει αυτόματα τα βάρη των κλάσεων με βάση τη συχνότητά τους στο training set. Επειδή η θετική κλάση εμφανίζεται πολύ σπανιότερα από την αρνητική, λαμβάνει μεγαλύτερο βάρος κατά την εκπαίδευση. Έτσι, το μοντέλο ενθαρρύνεται να δώσει μεγαλύτερη προσοχή στα περιστατικά εγκεφαλικού, αντί να ευνοεί σχεδόν αποκλειστικά την πλειοψηφική κλάση.

Αντίστοιχη προσέγγιση εφαρμόστηκε και στο Random Forest, όπου χρησιμοποιήθηκε στάθμιση κλάσεων ώστε τα δέντρα αποφάσεων του μοντέλου να λάβουν υπόψη την ανισορροπία. Παρόλο που το Random Forest είναι πιο σύνθετο μοντέλο από τη Logistic Regression, η ύπαρξη πολύ λίγων θετικών

δειγμάτων μπορεί να επηρεάσει τον τρόπο με τον οποίο δημιουργούνται οι διαχωρισμοί στα δέντρα. Η παρατήρηση αυτή δείχνει ότι, η χρήση `class weights` εξετάστηκε ως ένας τρόπος να δοθεί μεγαλύτερη σημασία στη θετική κλάση κατά την εκπαίδευση.

Στην περίπτωση του XGBoost, η αντίστοιχη αντιμετώπιση της ανισορροπίας πραγματοποιήθηκε μέσω της παραμέτρου `scale_pos_weight`. Η συγκεκριμένη παράμετρος χρησιμοποιείται σε προβλήματα δυαδικής ταξινόμησης με ανισορροπες κλάσεις και αυξάνει τη βαρύτητα των θετικών δειγμάτων κατά τη διαδικασία εκπαίδευσης. Η χρήση της ήταν ιδιαίτερα σημαντική, καθώς το XGBoost διαθέτει ενσωματωμένους μηχανισμούς που μπορούν να αξιοποιήσουν τέτοιου είδους στάθμιση στη διαδικασία βελτιστοποίησης.

Τα πειράματα με `class weights` σχεδιάστηκαν έτσι ώστε να εξεταστεί αν η αλλαγή της διαδικασίας μάθησης, χωρίς παραγωγή νέων δεδομένων, είναι αρκετή για να βελτιώσει την ικανότητα των μοντέλων να αναγνωρίζουν τη θετική κλάση. Μέσω της διαδικασίας, η τεχνική των βαρών κλάσεων αποτέλεσε ένα ενδιάμεσο βήμα ανάμεσα στα απλά `baseline` μοντέλα και στις πιο παρεμβατικές τεχνικές εξισορρόπησης, όπως η υπερδειγματοληψία και η παραγωγή συνθετικών δεδομένων.

4.2.3 Εφαρμογή SMOTE

Στη συνέχεια εφαρμόστηκε η τεχνική SMOTE, η οποία αποτελεί μία από τις πιο διαδεδομένες τεχνικές υπερδειγματοληψίας για προβλήματα ανισορροπής ταξινόμησης [8]. Σε αντίθεση με τα `class weights`, το SMOTE δεν αλλάζει απλώς τον τρόπο με τον οποίο το μοντέλο μαθαίνει, αλλά μεταβάλλει το ίδιο το σύνολο εκπαίδευσης. Συγκεκριμένα, δημιουργεί νέα συνθετικά δείγματα για τη μειοψηφική κλάση, με στόχο να αυξηθεί η παρουσία της μέσα στο `training set`.

Η τεχνική SMOTE δεν βασίζεται στην απλή αντιγραφή υπαρχόντων θετικών δειγμάτων. Αντίθετα, δημιουργεί νέα δείγματα μέσω παρεμβολής μεταξύ γειτονικών παρατηρήσεων της μειοψηφικής κλάσης. Κατ' αυτόν τον τρόπο, επιχειρείται η παραγωγή νέων σημείων που βρίσκονται στον χώρο των χαρακτηριστικών μεταξύ πραγματικών θετικών περιπτώσεων. Η λογική αυτή είναι σημαντική, επειδή η απλή επανάληψη των ίδιων θετικών εγγραφών θα μπορούσε να οδηγήσει σε υπερπροσαρμογή, ενώ η δημιουργία νέων συνθετικών παρατηρήσεων μπορεί να βοηθήσει το μοντέλο να μάθει πιο γενικευμένα μοτίβα της θετικής κλάσης.

Στην έρευνα που πραγματοποιήθηκε, το SMOTE εφαρμόστηκε αποκλειστικά στο `training set`, μετά τον διαχωρισμό των δεδομένων σε `training` και `test set`. Η συγκεκριμένη πρακτική, έγινε για να αποφευχθεί η διαρροή πληροφορίας από το `test set` προς τη διαδικασία εκπαίδευσης. Αν η υπερδειγματοληψία εφαρμοζόταν πριν από τον διαχωρισμό των δεδομένων, τότε υπήρχε πιθανότητα συνθετικά δείγματα που σχετίζονται με εγγραφές του `test set` να καταλήξουν στο `training set` ή αντίστροφα. Αυτό θα αλλοίωνε την ανεξαρτησία του `test set` και θα μπορούσε να οδηγήσει σε μη ρεαλιστικά αποτελέσματα.

Επιπλέον, εξετάστηκαν διαφορετικές αναλογίες υπερδειγματοληψίας, ώστε να αξιολογηθεί αν η μερική ή η πλήρης εξισορρόπηση οδηγεί σε καλύτερη συμπεριφορά των μοντέλων. Οι βασικές αναλογίες που χρησιμοποιήθηκαν ήταν 0,5, 0,7 και 1,0. Η αναλογία αυτή εκφράζει τη σχέση του αριθμού των δειγμάτων της μειοψηφικής κλάσης προς τον αριθμό των δειγμάτων της πλειοψηφικής κλάσης μετά την εφαρμογή του SMOTE. Για παράδειγμα, `ratio 0,7` σημαίνει ότι η θετική κλάση αυξάνεται ώστε να αντιστοιχεί περίπου στο 70% της αρνητικής κλάσης στο `training set`.

Η χρήση πολλών `ratios` ήταν σημαντική, επειδή η πλήρης εξισορρόπηση δεν είναι απαραίτητα πάντα η καλύτερη επιλογή. Σε ορισμένα προβλήματα, η υπερβολική παραγωγή συνθετικών δειγμάτων μπορεί να αλλοιώσει τη φυσική κατανομή των δεδομένων ή να οδηγήσει το μοντέλο σε υπερβολική ευαισθησία

απέναντι στη θετική κλάση. Για τον λόγο αυτό, δεν εξετάστηκε μόνο το σενάριο πλήρους εξισορρόπησης, αλλά και σενάρια μερικής εξισορρόπησης, ώστε να παρατηρηθεί ποια αναλογία οδηγεί σε πιο ισορροπημένη συμπεριφορά.

Μετά την εφαρμογή του SMOTE, τα μοντέλα Logistic Regression, Random Forest και XGBoost εκπαιδεύτηκαν εκ νέου πάνω στα τροποποιημένα training sets. Η αξιολόγηση πραγματοποιήθηκε και πάλι στο ίδιο πραγματικό και ανισόρροπο test set. Με τον τρόπο αυτό, τα αποτελέσματα των SMOTE πειραμάτων ήταν άμεσα συγκρίσιμα τόσο με τα baseline μοντέλα όσο και με τα μοντέλα που εκπαιδεύτηκαν με class weights.

4.2.4 Εφαρμογή Borderline-SMOTE

Εκτός από το κλασικό SMOTE, εφαρμόστηκε και η τεχνική Borderline-SMOTE. Η συγκεκριμένη μέθοδος αποτελεί παραλλαγή του SMOTE και σχεδιάστηκε ώστε να δίνει μεγαλύτερη έμφαση στα δείγματα της μειοψηφικής κλάσης που βρίσκονται κοντά στο όριο απόφασης μεταξύ των δύο κλάσεων. Η βασική ιδέα είναι ότι δεν έχουν όλα τα θετικά δείγματα την ίδια χρησιμότητα για την εκπαίδευση του μοντέλου. Ορισμένα δείγματα είναι πιο εύκολα διαχωρίσιμα από την αρνητική κλάση, ενώ άλλα βρίσκονται σε περιοχές όπου οι δύο κλάσεις αλληλεπικαλύπτονται.

Τα δείγματα που βρίσκονται κοντά στο όριο μεταξύ των δύο κλάσεων θεωρούνται πιο δύσκολα για το μοντέλο, επειδή είναι πιθανότερο να ταξινομηθούν λανθασμένα. Συνεπώς, το Borderline-SMOTE επικεντρώνεται περισσότερο σε αυτές τις περιοχές και δημιουργεί συνθετικά δείγματα γύρω από πιο «δύσκολες» περιπτώσεις της μειοψηφικής κλάσης. Κατ' αυτόν τον τρόπο, επιχειρείται η ενίσχυση των περιοχών όπου το μοντέλο χρειάζεται περισσότερη πληροφορία για να μάθει καλύτερα τον διαχωρισμό μεταξύ των δύο κλάσεων.

Στην εργασία αυτή, το Borderline-SMOTE εφαρμόστηκε επίσης αποκλειστικά στο training set, ακολουθώντας την ίδια λογική που εφαρμόστηκε και στο απλό SMOTE. Το test set δεν τροποποιήθηκε, ώστε η αξιολόγηση να παραμείνει ρεαλιστική και συγκρίσιμη με τα υπόλοιπα πειράματα. Η εφαρμογή της τεχνικής αυτής είχε ως στόχο να εξεταστεί αν μια πιο στοχευμένη μορφή υπερδειγματοληψίας μπορεί να είναι πιο αποτελεσματική από το απλό SMOTE για την ενίσχυση της θετικής κλάσης.

Μετά την παραγωγή των συνθετικών δειγμάτων μέσω Borderline-SMOTE, τα ίδια μοντέλα εκπαιδεύτηκαν εκ νέου πάνω στο νέο training set. Η αξιολόγηση πραγματοποιήθηκε με τις ίδιες μετρικές που χρησιμοποιήθηκαν σε όλα τα πειράματα, δηλαδή Accuracy, Precision, Recall, F1-score, ROC-AUC και PR-AUC. Έτσι, η συμπεριφορά των μοντέλων με Borderline-SMOTE μπορούσε να συγκριθεί άμεσα με τη baseline εκπαίδευση, τη χρήση class weights και το απλό SMOTE.

4.2.5 Εφαρμογή SMOTE-Tomek

Στη συνέχεια εφαρμόστηκε η τεχνική SMOTE-Tomek, η οποία συνδυάζει δύο διαφορετικές λογικές. Αφενός, την υπερδειγματοληψία της μειοψηφικής κλάσης και αφετέρου τον καθαρισμό του συνόλου δεδομένων. Αρχικά, η μέθοδος δημιουργεί νέα συνθετικά δείγματα για τη θετική κλάση μέσω SMOTE. Έπειτα, εφαρμόζει Tomek Links, με στόχο την αφαίρεση ορισμένων δειγμάτων που βρίσκονται σε περιοχές επικάλυψης μεταξύ των δύο κλάσεων.

Η λογική της τεχνικής αυτής είναι ότι η απλή αύξηση των δειγμάτων της μειοψηφικής κλάσης μπορεί να μην είναι αρκετή, αν στο dataset υπάρχουν σημεία που προκαλούν σύγχυση στο μοντέλο. Τα Tomek Links βοηθούν στον εντοπισμό ζευγών δειγμάτων διαφορετικών κλάσεων που βρίσκονται πολύ κοντά μεταξύ τους στον χώρο των χαρακτηριστικών. Τέτοιες περιπτώσεις μπορεί να θεωρηθούν αμφίβολες ή

δύσκολες για τον διαχωρισμό των κλάσεων. Με την αφαίρεση ορισμένων από αυτά τα σημεία, το training set μπορεί να γίνει πιο καθαρό ως προς τα όρια μεταξύ των δύο κλάσεων.

Στην έρευνα αυτή, το SMOTE-Tomek εφαρμόστηκε αποκλειστικά στο σύνολο εκπαίδευσης, όπως και όλες οι υπόλοιπες τεχνικές εξισορρόπησης. Αυτό έχει ως αποτέλεσμα, να διατηρηθεί η ανεξαρτησία του test set και να αποφευχθεί η διαρροή της πληροφορίας. Η τεχνική αυτή εξετάστηκε επειδή συνδυάζει την ενίσχυση της μειοψηφικής κλάσης με έναν μηχανισμό καθαρισμού, κάτι που θεωρητικά μπορεί να βοηθήσει τα μοντέλα να μάθουν πιο καθαρά όρια απόφασης.

Αφού ολοκληρώθηκε η εφαρμογή του SMOTE-Tomek, τα μοντέλα εκπαιδεύτηκαν ξανά και αξιολογήθηκαν στο ίδιο test set. Η σύγκριση με τις υπόλοιπες τεχνικές επέτρεψε να εξεταστεί αν ο συνδυασμός oversampling και καθαρισμού οδηγεί σε καλύτερη απόδοση σε σχέση με το απλό SMOTE ή το Borderline-SMOTE. Η τεχνική αυτή ήταν ιδιαίτερα χρήσιμη ως επιπλέον πειραματικό σενάριο, επειδή δεν περιορίζεται μόνο στην αύξηση των θετικών δειγμάτων, αλλά επιχειρεί να βελτιώσει και τη συνολική ποιότητα του training set.

4.2.6 Παραγωγή συνθετικών δεδομένων μέσω CTGAN

Το τελευταίο στάδιο της πειραματικής διαδικασίας αφορούσε τη χρήση συνθετικών δεδομένων που δημιουργήθηκαν μέσω CTGAN. Αυτή η μέθοδος επιλέχθηκε επειδή τα generative models έχουν τη δυνατότητα να μάθουν τη στατιστική κατανομή των πραγματικών δεδομένων και να δημιουργήσουν νέα δείγματα που μοιάζουν με αυτά. Σε αντίθεση με τις κλασικές τεχνικές υπερδειγματοληψίας, όπως το SMOTE, οι οποίες βασίζονται κυρίως σε παρεμβολή μεταξύ υπαρχόντων σημείων, το CTGAN προσπαθεί να μοντελοποιήσει πιο σύνθετες σχέσεις μεταξύ των χαρακτηριστικών.

Εν προκειμένω, το CTGAN δεν χρησιμοποιήθηκε για την παραγωγή ενός πλήρως νέου dataset, αλλά για την ενίσχυση της μειοψηφικής κλάσης. Πιο συγκεκριμένα, εκπαιδεύτηκε πάνω στα διαθέσιμα πραγματικά δεδομένα και χρησιμοποιήθηκε για την παραγωγή επιπλέον συνθετικών δειγμάτων που αντιστοιχούν στην κλάση stroke = 1. Το συγκεκριμένο βήμα, έγινε επειδή το βασικό πρόβλημα του dataset είναι η μικρή διαθεσιμότητα θετικών περιστατικών εγκεφαλικού. Επομένως, στόχος δεν ήταν να αυξηθεί συνολικά το dataset, αλλά να ενισχυθεί η κλάση που εμφανίζεται σπανιότερα και έχει μεγαλύτερη σημασία για την πρόβλεψη.

Η εκπαίδευση του CTGAN πραγματοποιήθηκε με 30 epochs. Στα πειράματα με CTGAN εξετάστηκαν διαφορετικές αναλογίες ενίσχυσης της μειοψηφικής κλάσης, συγκεκριμένα ratios 0,7, 0,5 και 0,3. Η λογική ήταν παρόμοια με αυτή των SMOTE πειραμάτων να εξεταστεί δηλαδή, αν διαφορετικός αριθμός συνθετικών θετικών δειγμάτων επηρεάζει τη συμπεριφορά των μοντέλων. Για κάθε ratio υπολογίστηκε ο επιθυμητός αριθμός δειγμάτων της θετικής κλάσης σε σχέση με την αρνητική κλάση του training set και στη συνέχεια δημιουργήθηκε ο αντίστοιχος αριθμός synthetic samples.

Στο training set υπήρχαν 3.889 δείγματα της αρνητικής κλάσης και 199 δείγματα της θετικής κλάσης. Για το CTGAN ratio 0,7 ο στόχος ήταν η θετική κλάση να φτάσει περίπου τα 2.722 δείγματα, δηλαδή το 70% της αρνητικής κλάσης. Άρα, δημιουργήθηκαν 2.523 συνθετικά θετικά δείγματα, τα οποία προστέθηκαν στα 199 πραγματικά θετικά δείγματα του training set. Το τελικό augmented training set για το συγκεκριμένο σενάριο περιείχε 6.611 εγγραφές.

Αντίστοιχα, για το CTGAN ratio 0,5 ο στόχος ήταν η θετική κλάση να φτάσει τα 1.944 δείγματα. Επειδή υπήρχαν ήδη 199 πραγματικά θετικά δείγματα, δημιουργήθηκαν 1.745 συνθετικά δείγματα. Το τελικό training set σε αυτή την περίπτωση περιείχε 5.833 εγγραφές. Για το CTGAN ratio 0,3 ο στόχος ήταν η

θετική κλάση να φτάσει τα 1.166 δείγματα, οπότε δημιουργήθηκαν 967 συνθετικά θετικά δείγματα και το τελικό training set περιείχε 5.055 εγγραφές.

Μετά την παραγωγή των συνθετικών δεδομένων, τα νέα δείγματα ενώθηκαν με το αρχικό training set και τα μοντέλα εκπαιδεύτηκαν εκ νέου. Η αξιολόγηση έγινε και πάλι στο ίδιο πραγματικό και ανισόρροπο test set. Η διαδικασία αυτή ήταν απαραίτητη, επειδή η παραγωγή συνθετικών δεδομένων δεν εγγυάται από μόνη της καλύτερη απόδοση. Τα synthetic samples πρέπει να είναι αρκετά ρεαλιστικά και χρήσιμα, ώστε να βοηθήσουν το μοντέλο να μάθει καλύτερα τη μειοψηφική κλάση και να γενικεύσει σε πραγματικά δεδομένα.

Συνεπώς, τα πειράματα με CTGAN είχαν διπλό σκοπό. Πρώτον, να εξεταστεί αν μια πιο σύγχρονη μέθοδος παραγωγής συνθετικών δεδομένων μπορεί να αποτελέσει εναλλακτική στις κλασικές τεχνικές υπερδειγματοληψίας. Δεύτερον, να αξιολογηθεί αν τα συνθετικά δεδομένα που παράγονται από το CTGAN είναι πραγματικά χρήσιμα για τα μοντέλα Logistic Regression, Random Forest και XGBoost στο συγκεκριμένο πρόβλημα πρόβλεψης εγκεφαλικού επεισοδίου.

4.2.7 Συνολική πειραματική διαδικασία

Συνοψίζοντας, η πειραματική διαδικασία της ενότητας αυτής οργανώθηκε με τρόπο ώστε κάθε τεχνική να αξιολογείται με όσο το δυνατόν πιο δίκαιο και συγκρίσιμο τρόπο. Όλα τα μοντέλα εκπαιδεύτηκαν αρχικά στο ίδιο προεπεξεργασμένο training set και αξιολογήθηκαν στο ίδιο test set. Στη συνέχεια, κάθε τεχνική εξισορρόπησης εφαρμόστηκε μόνο στο training set και η αξιολόγηση παρέμεινε σταθερή στο πραγματικό test set. Ως αποτέλεσμα, η σύγκριση μεταξύ των τεχνικών δεν επηρεάζεται από αλλαγές στο σύνολο ελέγχου.

Για λόγους επαναληψιμότητας της πειραματικής διαδικασίας, η υλοποίηση πραγματοποιήθηκε σε περιβάλλον Python, μέσω notebook, ώστε όλα τα στάδια της ανάλυσης να εκτελούνται με συγκεκριμένη και ελεγχόμενη σειρά. Για την επεξεργασία και διαχείριση των δεδομένων χρησιμοποιήθηκαν οι βιβλιοθήκες pandas και NumPy, ενώ για την υλοποίηση των μοντέλων μηχανικής μάθησης και των μετρικών αξιολόγησης χρησιμοποιήθηκε η βιβλιοθήκη scikit-learn. Οι τεχνικές υπερδειγματοληψίας, όπως το SMOTE, το Borderline-SMOTE και το SMOTE-Tomek, υλοποιήθηκαν μέσω της βιβλιοθήκης imbalanced-learn, ενώ για το μοντέλο XGBoost χρησιμοποιήθηκε η αντίστοιχη βιβλιοθήκη xgboost. Για την παραγωγή συνθετικών δεδομένων με CTGAN αξιοποιήθηκε κατάλληλη βιβλιοθήκη παραγωγής συνθετικών tabular δεδομένων, ενώ για την υλοποίηση του Federated Learning στο επόμενο κεφάλαιο χρησιμοποιήθηκε η βιβλιοθήκη PyTorch. Η επαναληψιμότητα των πειραμάτων ενισχύθηκε με τη χρήση σταθερών ρυθμίσεων σε όλα τα στάδια της διαδικασίας. Ο διαχωρισμός των δεδομένων πραγματοποιήθηκε με αναλογία 80/20 σε training και test set, με stratification ως προς τη μεταβλητή στόχο, ώστε να διατηρηθεί η αναλογία των κλάσεων και στα δύο σύνολα. Επιπλέον, όπου υποστηριζόταν από τις βιβλιοθήκες, χρησιμοποιήθηκε σταθερή τιμή random_state, ώστε οι τυχαίες διαδικασίες, όπως ο διαχωρισμός των δεδομένων και η παραγωγή συνθετικών δειγμάτων μέσω SMOTE, να μπορούν να επαναληφθούν με τον ίδιο τρόπο.

Η σειρά των πειραμάτων είχε επίσης συγκεκριμένη λογική. Αρχικά χρησιμοποιήθηκαν baseline μοντέλα, ώστε να αποτυπωθεί η αρχική συμπεριφορά χωρίς παρέμβαση. Έπειτα εφαρμόστηκαν class weights, δηλαδή μια τεχνική που δεν αλλάζει το dataset αλλά αλλάζει τη διαδικασία μάθησης. Στη συνέχεια εξετάστηκαν τεχνικές υπερδειγματοληψίας, οι οποίες μεταβάλλουν το training set δημιουργώντας νέα δείγματα για τη μειοψηφική κλάση. Τέλος, εφαρμόστηκε CTGAN, ώστε να εξεταστεί μια πιο σύνθετη προσέγγιση παραγωγής συνθετικών δεδομένων.

Για κάθε πείραμα υπολογίστηκαν οι ίδιες μετρικές αξιολόγησης και η ανάλυση των αποτελεσμάτων βασίστηκε κυρίως στο Recall, το F1-score και το PR-AUC. Ιδιαίτερη σημασία δόθηκε στο Recall, καθώς δείχνει την ικανότητα του μοντέλου να εντοπίζει τα πραγματικά περιστατικά εγκεφαλικού.

Με αυτή τη διαδικασία, η εργασία δεν εξετάζει απλώς αν ένα μοντέλο μπορεί να προβλέψει το εγκεφαλικό επεισόδιο, αλλά αξιολογεί συστηματικά πώς διαφορετικές τεχνικές εκπαίδευσης και εξισορρόπησης επηρεάζουν τη συμπεριφορά του. Η αναλυτική ερμηνεία των αποτελεσμάτων και η σύγκριση της συμπεριφοράς των μοντέλων παρουσιάζονται στην επόμενη ενότητα.

Για την ενίσχυση της αναπαραγωγιμότητας της εργασίας, τα παραγόμενα datasets της πειραματικής διαδικασίας διατέθηκαν σε δημόσιο GitHub repository³. Το repository περιλαμβάνει τα datasets που χρησιμοποιήθηκαν μετά την προεπεξεργασία, καθώς και τα σύνολα εκπαίδευσης που δημιουργήθηκαν από τις τεχνικές SMOTE, Borderline-SMOTE, SMOTE-Tomek και CTGAN. Η διάθεση των αρχείων αυτών επιτρέπει την ανεξάρτητη επιθεώρηση των δεδομένων και διευκολύνει την επανάληψη ή επέκταση των πειραμάτων.

4.3 Ανάλυση Συμπεριφοράς Μοντέλων

Μετά την ολοκλήρωση των πειραμάτων με τις διαφορετικές τεχνικές εξισορρόπησης, πραγματοποιήθηκε συγκριτική ανάλυση της συμπεριφοράς των μοντέλων στο πραγματικό test set. Η συγκεκριμένη ανάλυση αποτελεί βασικό μέρος της εργασίας, καθώς δεν αρκεί απλώς να καταγραφούν οι αριθμητικές τιμές των μετρικών, αλλά πρέπει να ερμηνευθεί το πώς και το γιατί κάθε μοντέλο ανταποκρίθηκε διαφορετικά στις τεχνικές που εφαρμόστηκαν. Στο παρόν κείμενο, το ενδιαφέρον επικεντρώνεται κυρίως στην ικανότητα των μοντέλων να εντοπίζουν τη θετική κλάση, δηλαδή τα περιστατικά εγκεφαλικού επεισοδίου, τα οποία αποτελούν τη μειοψηφική αλλά και την πιο κρίσιμη κατηγορία του προβλήματος.

Η αξιολόγηση όλων των μοντέλων πραγματοποιήθηκε στο ίδιο test set, το οποίο δεν είχε υποστεί κάποια τεχνική εξισορρόπησης στο κύριο πειραματικό μέρος. Το test set περιλάμβανε 1.022 εγγραφές, από τις οποίες οι 972 αντιστοιχούσαν σε ασθενείς χωρίς εγκεφαλικό επεισόδιο και οι 50 σε ασθενείς με εγκεφαλικό επεισόδιο. Η επιλογή αυτή ήταν σημαντική, διότι επέτρεψε την αξιολόγηση των μοντέλων σε μια κατανομή που παραμένει κοντά στην πραγματική φύση του προβλήματος. Επομένως, τα αποτελέσματα που παρουσιάζονται στην παρούσα ενότητα δεν εκφράζουν την απόδοση των μοντέλων σε ένα τεχνητά ισορροπημένο περιβάλλον, αλλά την ικανότητά τους να γενικεύουν σε πραγματικά ανισόροπα δεδομένα.

Κατά την ανάλυση των αποτελεσμάτων, ιδιαίτερη έμφαση δόθηκε στις μετρικές Recall, F1-score και PR-AUC. Το Accuracy καταγράφηκε σε όλα τα πειράματα, αλλά δεν χρησιμοποιήθηκε ως βασικό κριτήριο αξιολόγησης, επειδή σε ένα τόσο ανισόροπο dataset μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα. Αυτό φάνηκε ήδη από τα baseline αποτελέσματα, όπου ορισμένα μοντέλα εμφάνιζαν υψηλή συνολική ακρίβεια, χωρίς όμως να εντοπίζουν ουσιαστικά τα περιστατικά εγκεφαλικού. Για αυτόν τον λόγο, η ανάλυση επικεντρώνεται κυρίως στο κατά πόσο κάθε μοντέλο κατάφερε να αυξήσει την ικανότητα εντοπισμού της θετικής κλάσης μετά την εφαρμογή των τεχνικών εξισορρόπησης.

4.3.1 Συμπεριφορά της Logistic Regression

Η Logistic Regression παρουσίασε μία από τις πιο ενδιαφέρουσες συμπεριφορές στην πειραματική διαδικασία, καθώς επηρεάστηκε έντονα από τις τεχνικές αντιμετώπισης της ανισορροπίας. Όπως

³ GitHub repository: https://github.com/Giannis-Stivaktakis/stroke_prediction

φαίνεται και στον πίνακα 4.1, στο baseline σενάριο, δηλαδή χωρίς καμία παρέμβαση στην κατανομή των κλάσεων ή στη διαδικασία εκπαίδευσης, το μοντέλο εμφάνισε πολύ υψηλό Accuracy, περίπου 0,9521. Ωστόσο, η τιμή αυτή ήταν παραπλανητική, καθώς το Recall για την κλάση του εγκεφαλικού ήταν μόλις 0,02. Αυτό σημαίνει ότι από τις 50 πραγματικές περιπτώσεις εγκεφαλικού στο test set, το μοντέλο κατάφερε να εντοπίσει σωστά μόνο μία.

Πίνακας 4.2: Αριθμητικά δεδομένα Logistic Regression Baseline

Μετρική	Accuracy	Precision	Recall	F1 - Score	ROC-AUC	PR-AUC
Τιμή	0.952	1.0	0.02	0.039	0.842	0.269

Με βάση τον πίνακα 4.2 ο οποίος παρουσιάζει το confusion matrix του baseline μοντέλου, επιβεβαιώνεται αυτή τη συμπεριφορά, καθώς το μοντέλο ταξινομήσε σωστά σχεδόν όλες τις αρνητικές περιπτώσεις, αλλά απέτυχε σχεδόν πλήρως στη θετική κλάση. Συγκεκριμένα, υπήρχαν 972 true negatives, 0 false positives, 49 false negatives και μόνο 1 true positive. Το αποτέλεσμα αυτό ήταν ιδιαίτερα σημαντικό, γιατί έδειξε στην πράξη ότι ένα μοντέλο μπορεί να φαίνεται πολύ καλό αν εξεταστεί μόνο το Accuracy, αλλά να είναι ανεπαρκής για τον πραγματικό στόχο της εργασίας, δηλαδή τον εντοπισμό των περιστατικών εγκεφαλικού.

Πίνακας 4.3: Confusion Matrix – Centralized LR Baseline

Actual classes	Predicted classes	
	No Stroke	Stroke
No Stroke	972	0
Stroke	49	1

Η χρήση βαρών κλάσεων άλλαξε σημαντικά τη συμπεριφορά της Logistic Regression. Με την επιλογή `class_weight='balanced'`, το μοντέλο έδωσε μεγαλύτερη σημασία στη μειοψηφική κλάση κατά την εκπαίδευση. Όπως φαίνεται και στον πίνακα 4.3, το αποτέλεσμα ήταν η θεαματική αύξηση του Recall από 0,02 σε περίπου 0,80. Αυτό σημαίνει ότι το μοντέλο κατάφερε πλέον να εντοπίσει 40 από τις 50 πραγματικές περιπτώσεις εγκεφαλικού. Η βελτίωση αυτή είναι πολύ σημαντική, επειδή δείχνει ότι η Logistic Regression, παρόλο που είναι ένα σχετικά απλό γραμμικό μοντέλο, μπορεί να γίνει πολύ πιο ευαίσθητη απέναντι στη θετική κλάση όταν η ανισορροπία αντιμετωπιστεί κατάλληλα.

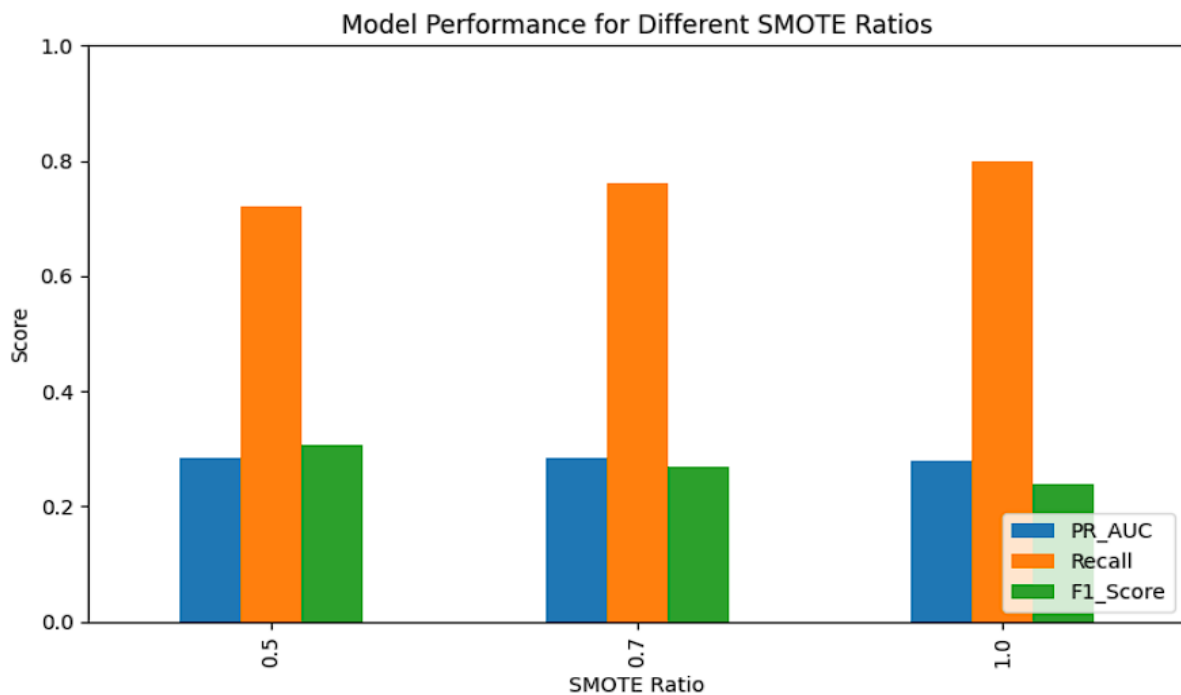
Πίνακας 4.4: Αριθμητικά δεδομένα Logistic Regression Class_Weight

Μετρική	Accuracy	Precision	Recall	F1 - Score	ROC-AUC	PR-AUC
Τιμή	0.745	0.137	0.8	0.235	0.843	0.268

Ωστόσο, η αύξηση του Recall συνοδεύτηκε από σημαντική αύξηση των false positives. Το μοντέλο άρχισε να προβλέπει περισσότερες περιπτώσεις ως θετικές, με αποτέλεσμα να εντοπίζει περισσότερα πραγματικά περιστατικά εγκεφαλικού, αλλά ταυτόχρονα να χαρακτηρίζει λανθασμένα αρκετούς ασθενείς ως θετικούς. Αυτό οδήγησε σε χαμηλότερο Precision και σε F1-score περίπου 0,2353. Το αποτέλεσμα αυτό δείχνει ξεκάθαρα την ανταλλαγή μεταξύ Recall και Precision. Σε ένα ιατρικό

πρόβλημα, η αύξηση του Recall είναι σημαντική, όμως δεν μπορεί να εξετάζεται απομονωμένα, καθώς η μεγάλη αύξηση των λανθασμένων θετικών προβλέψεων μπορεί να περιορίσει την πρακτική χρησιμότητα του μοντέλου.

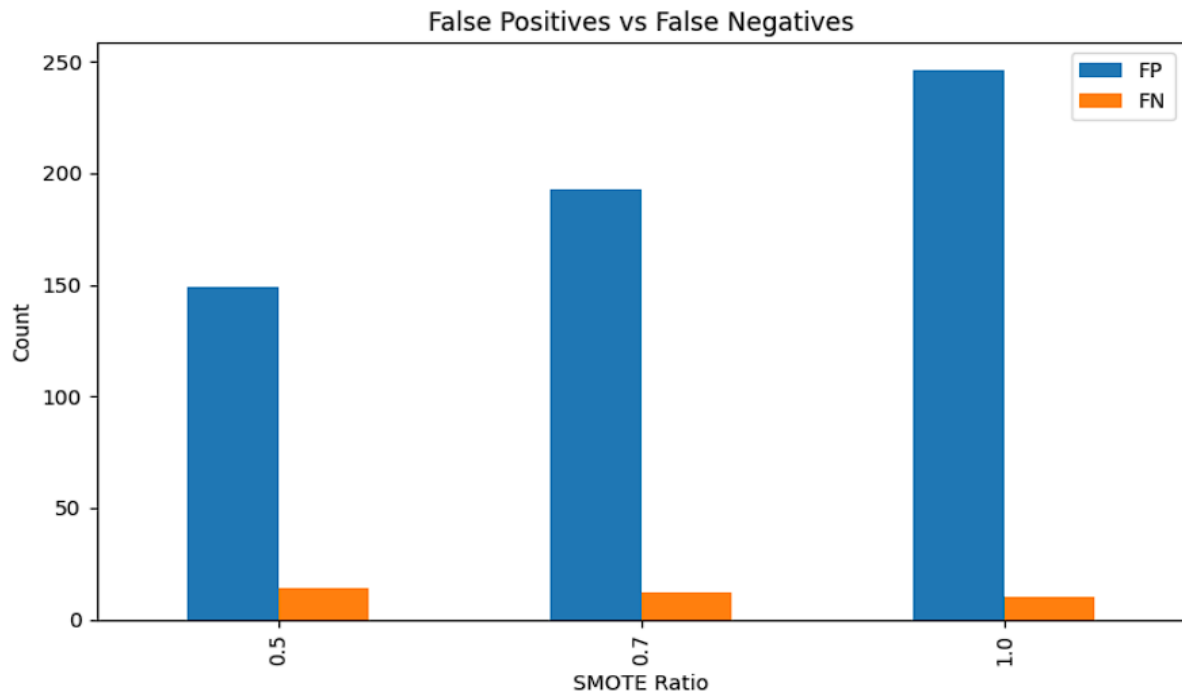
Στα πειράματα με SMOTE, η Logistic Regression παρουσίασε επίσης σημαντική βελτίωση σε σχέση με το baseline. Ειδικά στο σενάριο SMOTE με ratio 0,7 το Recall έφτασε περίπου στο 0,76 ενώ το F1-score ήταν περίπου 0,2705 και το PR-AUC περίπου 0,2896. Η συμπεριφορά αυτή δείχνει ότι η δημιουργία επιπλέον συνθετικών δειγμάτων για τη μειοψηφική κλάση βοήθησε το μοντέλο να μάθει καλύτερα τα χαρακτηριστικά που σχετίζονται με την εμφάνιση εγκεφαλικού. Σε σχέση με τα class weights, το Recall ήταν ελαφρώς χαμηλότερο, όμως η συνολική ισορροπία του μοντέλου ήταν καλύτερη, όπως φαίνεται από τη βελτίωση του F1-score και του PR-AUC.



Σχήμα 4.1: Απόδοση SMOTE Ratios

Από το Σχήμα 4.2 παρατηρείται ότι η αύξηση του SMOTE ratio οδηγεί σε βελτίωση του Recall, καθώς το μοντέλο εντοπίζει περισσότερα περιστατικά της θετικής κλάσης. Ωστόσο, η πλήρης εξισορρόπηση με ratio 1.0 δεν οδηγεί απαραίτητα στην καλύτερη συνολική συμπεριφορά, καθώς το F1-score μειώνεται σε σχέση με τα μικρότερα ratios. Για τον λόγο αυτό, το SMOTE ratio 0.7 θεωρείται πιο ισορροπημένη επιλογή, καθώς διατηρεί υψηλό Recall χωρίς τόσο μεγάλη απώλεια στη συνολική ισορροπία του μοντέλου.

Παράλληλα, το Σχήμα 4.3 παρουσιάζει τη μεταβολή των False Positives και False Negatives για τα διαφορετικά SMOTE ratios. Παρατηρείται ότι όσο αυξάνεται το ratio, τα False Negatives μειώνονται ελαφρώς, γεγονός που σημαίνει ότι το μοντέλο χάνει λιγότερα πραγματικά περιστατικά εγκεφαλικού. Ωστόσο, η μείωση αυτή συνοδεύεται από σημαντική αύξηση των False Positives, ειδικά στο ratio 1.0. Αυτό δείχνει ότι η πλήρης εξισορρόπηση κάνει το μοντέλο πιο ευαίσθητο στη θετική κλάση, αλλά ταυτόχρονα αυξάνει τις λανθασμένες θετικές προβλέψεις.



Σχήμα 4.2: FP-FN των SMOTE Ratios

Ακόμη πιο ισορροπημένη συμπεριφορά παρατηρήθηκε με το Borderline-SMOTE. Η συγκεκριμένη τεχνική οδήγησε σε Recall περίπου 0,78 και F1-score περίπου 0,2910. Το αποτέλεσμα αυτό ήταν ιδιαίτερα ενδιαφέρον, καθώς το Borderline-SMOTE δεν αύξησε απλώς την ευαισθησία του μοντέλου στη θετική κλάση, αλλά πέτυχε και καλύτερη ισορροπία μεταξύ Recall και Precision. Η ερμηνεία που προκύπτει είναι ότι η ενίσχυση των πιο δύσκολων δειγμάτων κοντά στο όριο απόφασης βοήθησε τη Logistic Regression περισσότερο από την πιο γενική παραγωγή συνθετικών δειγμάτων του απλού SMOTE.

Από τον Πίνακα 4.4 φαίνεται ότι το Borderline-SMOTE παρουσίασε την πιο ισορροπημένη συμπεριφορά σε σχέση με τις υπόλοιπες SMOTE-based τεχνικές. Συγκεκριμένα, πέτυχε το υψηλότερο Recall με τιμή 0,78, ενώ ταυτόχρονα είχε και το καλύτερο F1-score με τιμή 0,29. Επιπλέον, σε σχέση με το απλό SMOTE, μείωσε τα False Positives από 193 σε 179 και τα False Negatives από 12 σε 11, γεγονός που δείχνει ότι βοήθησε το μοντέλο να εντοπίσει περισσότερα πραγματικά περιστατικά εγκεφαλικού με ελαφρώς καλύτερη συνολική ισορροπία.

Πίνακας 4.5: Αριθμητικά δεδομένα Logistic Regression SMOTE

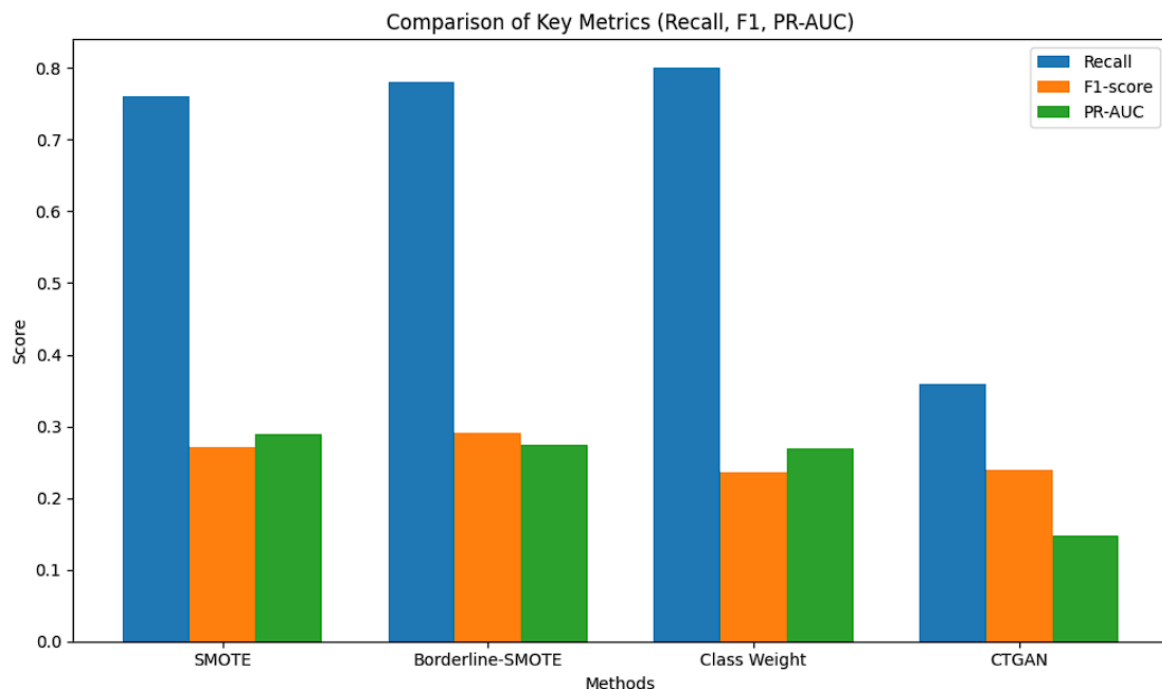
Μέθοδος	Precision	Recall	F1	ROC_AUC	PR_AUC	FP	FN	TP	TN
SMOTE (0.7)	0.16	0.76	0.27	0.84	0.28	193	12	38	779
Borderline SMOTE	0.17	0.78	0.29	0.84	0.27	179	11	39	793
SMOTE Tomek	0.16	0.76	0.27	0.84	0.28	193	12	38	779

Αντίθετα, τα πειράματα με CTGAN δεν οδήγησαν σε καλύτερη συμπεριφορά για τη Logistic Regression σε σχέση με τις SMOTE-based τεχνικές. Όπως φαίνεται από τον πίνακα 4.5, στο σενάριο CTGAN με ratio 0,7 το Recall ήταν 0,36 με F1-score περίπου 0,2 και PR-AUC περίπου 0,15. Η τιμή αυτή είναι καλύτερη από το baseline, αλλά αισθητά χαμηλότερη από τα αποτελέσματα με class weights, SMOTE και Borderline-SMOTE. Αυτό δείχνει ότι τα συνθετικά δεδομένα που παρήχθησαν μέσω CTGAN δεν ήταν τόσο χρήσιμα για τη Logistic Regression όσο τα δείγματα που δημιουργήθηκαν με τις κλασικές τεχνικές υπερδειγματοληψίας.

Πίνακας 4.6: Αριθμητικά δεδομένα Logistic Regression CTGAN

Μετρική	Accuracy	Precision	Recall	F1 - Score	ROC-AUC	PR-AUC
Τιμή	0.832	0.13	0.36	0.2	0.72	0.15

Συμπερασματικά, η Logistic Regression ήταν το μοντέλο που επωφελήθηκε περισσότερο από τις τεχνικές αντιμετώπισης της ανισορροπίας. Από ένα baseline μοντέλο που σχεδόν αγνοούσε τη θετική κλάση, μετατράπηκε σε μοντέλο με υψηλό Recall όταν εφαρμόστηκαν class weights και SMOTE-based τεχνικές. Το αποτέλεσμα αυτό δείχνει ότι, παρά την απλότητά της, η LR μπορεί να αποτελέσει ισχυρό εργαλείο σε ανισόρροπα προβλήματα, αρκεί να συνδυαστεί με κατάλληλη τεχνική εξισορρόπησης. Από το Σχήμα 4.4 φαίνεται ότι οι τεχνικές SMOTE, Borderline-SMOTE και class weights βελτίωσαν σημαντικά το Recall της Logistic Regression σε σχέση με το baseline μοντέλο. Η χρήση class weights πέτυχε το υψηλότερο Recall, όμως το Borderline-SMOTE παρουσίασε καλύτερη συνολική ισορροπία, καθώς είχε υψηλό Recall και το μεγαλύτερο F1-score. Αντίθετα, το CTGAN παρουσίασε χαμηλότερη απόδοση στις βασικές μετρικές, γεγονός που δείχνει ότι τα συνθετικά δεδομένα που παρήχθησαν δεν βοήθησαν το μοντέλο στον ίδιο βαθμό με τις SMOTE-based τεχνικές.



Σχήμα 4.3: Σύγκριση αποτελεσμάτων της Logistic Regression

Πίνακας 4.7: Σύγκριση τεχνικών εξισορρόπησης για τη Logistic Regression.

Τεχνική	Accuracy	Precision	Recall	F1 - Score	ROC-AUC	PR-AUC
Baseline	0.95	1.0	0.02	0.03	0.84	0.26
Class weight	0.74	0.13	0.8	0.23	0.84	0.26
SMOTE	0.79	0.16	0.76	0.27	0.84	0.28
Borderline SMOTE	0.81	0.17	0.78	0.29	0.84	0.27
SMOTE Tomek	0.79	0.16	0.76	0.27	0.84	0.28
CTGAN	0.83	0.13	0.36	0.2	0.72	0.15

Από τον Πίνακα 4.7 φαίνεται ότι η εφαρμογή τεχνικών εξισορρόπησης βελτίωσε σημαντικά τη συμπεριφορά της Logistic Regression σε σχέση με το baseline μοντέλο. Παρόλο που η χρήση class weights πέτυχε το υψηλότερο Recall, το Borderline-SMOTE παρουσίασε την πιο ισορροπημένη συνολική απόδοση, καθώς συνδύασε υψηλό Recall με το καλύτερο F1-score. Για τον λόγο αυτό, θεωρείται η καταλληλότερη τεχνική για τη Logistic Regression στο συγκεκριμένο πειραματικό πλαίσιο.

4.3.2 Συμπεριφορά του Random Forest

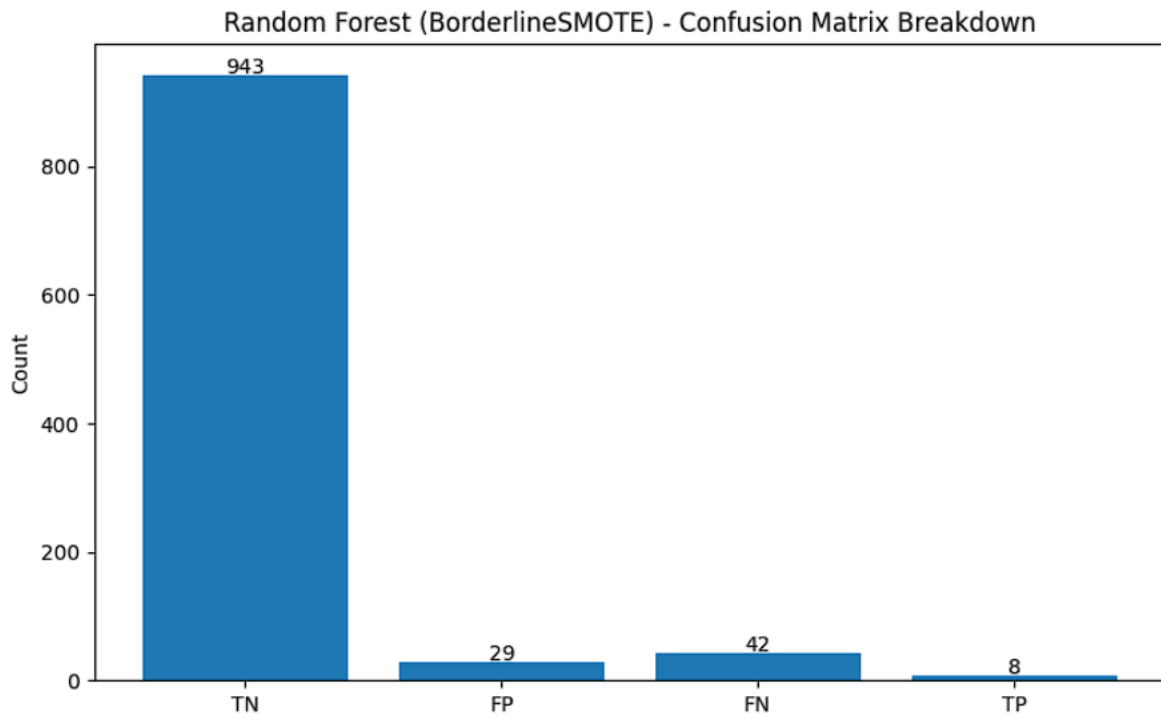
Το Random Forest παρουσίασε αρκετά διαφορετική συμπεριφορά σε σχέση με τη Logistic Regression. Παρόλο που θεωρητικά αποτελεί πιο σύνθετο μοντέλο και έχει τη δυνατότητα να αποτυπώνει μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών, στην πράξη δυσκολεύτηκε σημαντικά να εντοπίσει τη θετική κλάση του εγκεφαλικού. Στο baseline σενάριο, το μοντέλο εμφάνισε υψηλό Accuracy, όμως το Recall ήταν 0,00. Αυτό σημαίνει ότι δεν αναγνώρισε καμία από τις 50 πραγματικές περιπτώσεις εγκεφαλικού στο test set.

Το αποτέλεσμα αυτό δείχνει ότι η πολυπλοκότητα ενός μοντέλου δεν εγγυάται απαραίτητα καλύτερη απόδοση σε ανισόρροπα δεδομένα. Το Random Forest φαίνεται ότι έμαθε κυρίως τη συμπεριφορά της πλειοψηφικής κλάσης, δηλαδή των ασθενών χωρίς εγκεφαλικό επεισόδιο. Επειδή η θετική κλάση είχε πολύ μικρή παρουσία στο training set, το μοντέλο δεν κατάφερε να δημιουργήσει αποτελεσματικούς κανόνες ταξινόμησης για τα περιστατικά εγκεφαλικού. Αυτό αποτελεί ένα σημαντικό εύρημα, καθώς δείχνει ότι ένα πιο σύνθετο μοντέλο δεν είναι απαραίτητα πιο κατάλληλο όταν το βασικό πρόβλημα είναι η κατανομή των δεδομένων.

Η χρήση class weights στο Random Forest δεν βελτίωσε ουσιαστικά τη συμπεριφορά του μοντέλου. Το Recall παρέμεινε 0,00 γεγονός που δείχνει ότι η απλή αύξηση της βαρύτητας της θετικής κλάσης δεν ήταν αρκετή για να αλλάξει τον τρόπο με τον οποίο το Random Forest μάθαινε από τα δεδομένα. Αυτό διαφοροποιείται σημαντικά από τη Logistic Regression, όπου η ίδια γενική λογική στάθμισης είχε οδηγήσει σε πολύ μεγάλη αύξηση του Recall. Επομένως, από τα αποτελέσματα προκύπτει ότι η αποτελεσματικότητα μιας τεχνικής εξαρτάται άμεσα από το είδος του μοντέλου.

Με την εφαρμογή του SMOTE, το Random Forest παρουσίασε κάποια βελτίωση, αλλά η βελτίωση αυτή παρέμεινε περιορισμένη. Στο σενάριο SMOTE με ratio 0,7 το Recall έφτασε περίπου στο 0,12. Αυτό σημαίνει ότι το μοντέλο κατάφερε να εντοπίσει 6 από τις 50 πραγματικές θετικές περιπτώσεις. Αν και η τιμή αυτή είναι καλύτερη από το baseline και το class weight σενάριο, παραμένει χαμηλή για ένα

πρόβλημα όπου ο βασικός στόχος είναι η αναγνώριση όσο το δυνατόν περισσότερων περιστατικών εγκεφαλικού. Το αποτέλεσμα δείχνει ότι η απλή αύξηση των θετικών δειγμάτων μέσω SMOTE δεν ήταν αρκετή για να επιτρέψει στο Random Forest να γενικεύσει αποτελεσματικά στο test set.

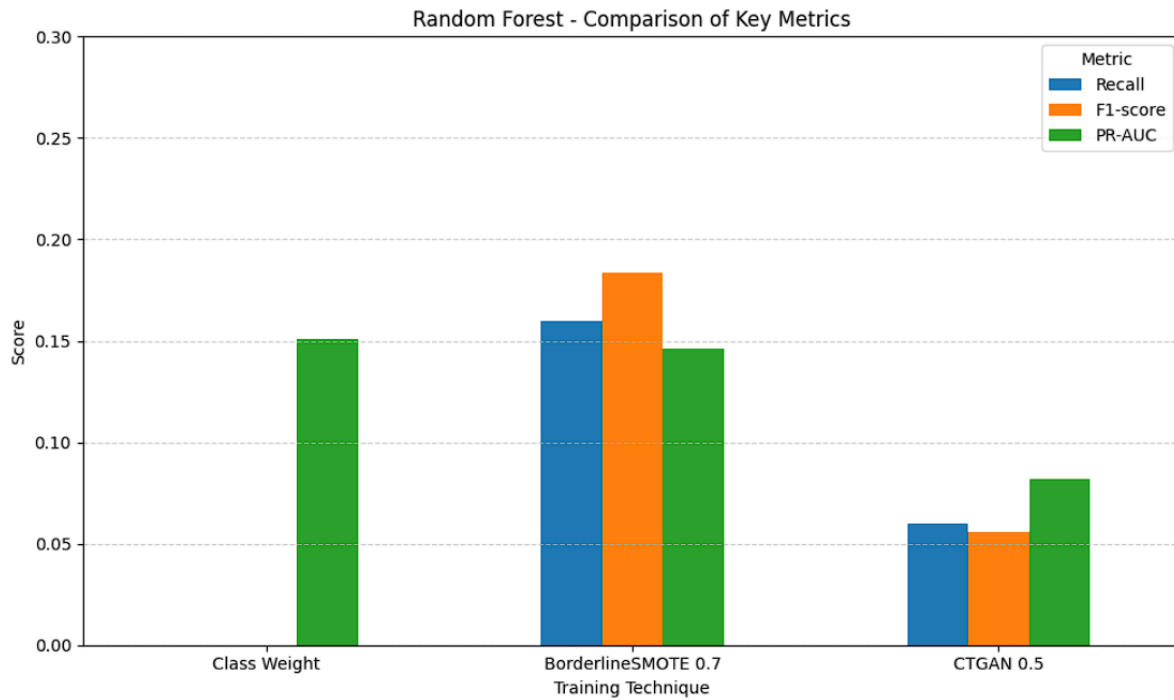


Σχήμα 4.4: Confusion Matrix Random Forest Boredeline SMOTE

Όπως γίνεται αντιληπτό και από το Σχήμα 4.5, η καλύτερη συμπεριφορά για το Random Forest παρατηρήθηκε με το Borderline-SMOTE. Με τη συγκεκριμένη τεχνική, το Recall έφτασε περίπου στο 0,16 το F1-score περίπου στο 0,1839 και το PR-AUC περίπου στο 0,1459. Παρόλο που οι τιμές αυτές δεν είναι ιδιαίτερα υψηλές, αποτέλεσαν τη βέλτιστη συμπεριφορά του Random Forest μεταξύ των τεχνικών που εξετάστηκαν. Αυτό δείχνει ότι η στοχευμένη ενίσχυση των δειγμάτων κοντά στο όριο απόφασης βοήθησε περισσότερο από το απλό SMOTE. Παρ' όλα αυτά, το μοντέλο εξακολούθησε να δυσκολεύεται σημαντικά με τη θετική κλάση.

Τα πειράματα με CTGAN δεν βελτίωσαν ουσιαστικά τη συμπεριφορά του Random Forest. Στο σενάριο CTGAN με ratio 0,5 το Recall ήταν περίπου 0,06 ενώ το F1-score περίπου 0,0506 και το PR-AUC περίπου 0,081. Οι τιμές αυτές δείχνουν ότι τα συνθετικά δεδομένα που δημιουργήθηκαν μέσω CTGAN δεν παρείχαν αρκετά χρήσιμη πληροφορία στο Random Forest ώστε να βελτιώσει την ικανότητα εντοπισμού των περιστατικών εγκεφαλικού. Αντίθετα, το μοντέλο συνέχισε να εμφανίζει περιορισμένη ευαισθησία στη μειοψηφική κλάση.

Συνολικά, το Random Forest ήταν το μοντέλο που παρουσίασε τη μεγαλύτερη δυσκολία στο συγκεκριμένο πρόβλημα. Παρόλο που δοκιμάστηκαν διαφορετικές τεχνικές αντιμετώπισης της ανισορροπίας, η απόδοσή του ως προς τη θετική κλάση παρέμεινε χαμηλή. Με βάση το Σχήμα 4.6 γίνεται αντιληπτό ότι, η καλύτερη συμπεριφορά του παρατηρήθηκε με το Borderline-SMOTE, όμως ακόμη και τότε το Recall παρέμεινε αρκετά περιορισμένο. Από την ανάλυση αυτή προκύπτει ότι το Random Forest, στο συγκεκριμένο dataset, δεν αξιοποίησε αποτελεσματικά ούτε τη στάθμιση κλάσεων ούτε τα συνθετικά δεδομένα στον ίδιο βαθμό με άλλα μοντέλα.



Σχήμα 4.5: Random Forest σύγκριση τεχνικών

Πίνακας 4.8: Σύγκριση τεχνικών εξισορρόπησης για το Random Forest.

Τεχνική	Accuracy	Precision	Recall	F1 - Score	ROC-AUC	PR-AUC
Baseline	0.95	0	0	0	0.81	0.16
Class weight	0.95	0	0	0	0.80	0.15
SMOTE	0.92	0.15	0.12	0.13	0.77	0.13
Borderline SMOTE	0.93	0.21	0.16	0.18	0.79	0.14
SMOTE Tomek	0.92	0.15	0.14	0.15	0.77	0.13
CTGAN	0.90	0.05	0.06	0.05	0.69	0.081

Συμπερασματικά, όπως γίνεται αντιληπτό από τον Πίνακα 4.8, το Random Forest δυσκολεύτηκε σημαντικά να εντοπίσει τη θετική κλάση. Τόσο στο baseline σενάριο όσο και με τη χρήση class weights, το Recall παρέμεινε μηδενικό, γεγονός που δείχνει ότι το μοντέλο δεν αναγνώρισε πραγματικά περιστατικά εγκεφαλικού. Αντίθετα, οι τεχνικές υπερδειγματοληψίας βελτίωσαν σε κάποιο βαθμό τη συμπεριφορά του μοντέλου, με το Borderline-SMOTE να παρουσιάζει την καλύτερη συνολική απόδοση, καθώς πέτυχε το υψηλότερο Precision, Recall και F1-score. Παρόλα αυτά, η απόδοση του Random Forest παρέμεινε χαμηλότερη σε σχέση με τη Logistic Regression.

4.3.3 Συμπεριφορά του XGBoost

Το XGBoost παρουσίασε ενδιάμεση συμπεριφορά σε σχέση με τη Logistic Regression και το Random Forest. Στο baseline σενάριο, χωρίς καμία τεχνική αντιμετώπισης της ανισορροπίας, το μοντέλο

εμφάνισε Recall περίπου 0,04. Αυτό σημαίνει ότι εντόπισε μόνο ένα πολύ μικρό μέρος των πραγματικών περιστατικών εγκεφαλικού. Παρόλο που το XGBoost θεωρείται ισχυρός αλγόριθμος για προβλήματα ταξινόμησης με δομημένα δεδομένα, τα baseline αποτελέσματα έδειξαν ότι και αυτό επηρεάζεται έντονα όταν η θετική κλάση είναι πολύ μικρή.

Η πιο σημαντική βελτίωση για το XGBoost παρατηρήθηκε με τη χρήση της παραμέτρου `scale_pos_weight`. Η συγκεκριμένη τεχνική αύξησε τη βαρύτητα της θετικής κλάσης κατά την εκπαίδευση και βοήθησε το μοντέλο να δώσει μεγαλύτερη σημασία στα περιστατικά εγκεφαλικού. Με αυτή τη ρύθμιση, το Recall έφτασε περίπου στο 0,50 δηλαδή το μοντέλο κατάφερε να εντοπίσει 25 από τις 50 πραγματικές θετικές περιπτώσεις. Το confusion matrix για το συγκεκριμένο πείραμα έδειξε 860 true negatives, 112 false positives, 25 false negatives και 25 true positives.

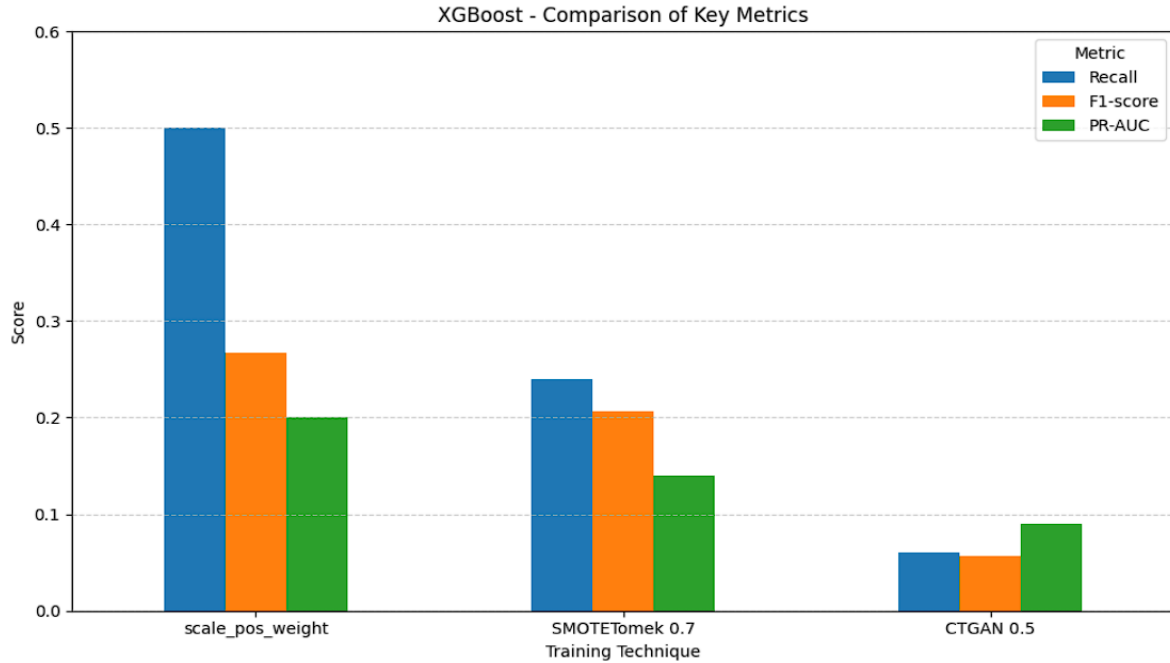
Το αποτέλεσμα αυτό ήταν σημαντικό, διότι έδειξε ότι το XGBoost ανταποκρίθηκε καλύτερα στην εσωτερική στάθμιση της θετικής κλάσης παρά στις τεχνικές παραγωγής επιπλέον συνθετικών δειγμάτων. Με το `scale_pos_weight`, το μοντέλο δεν χρειάστηκε να εκπαιδευτεί σε νέο τροποποιημένο dataset, αλλά άλλαξε ο τρόπος με τον οποίο υπολόγιζε τη σημασία των σφαλμάτων κατά την εκπαίδευση. Η προσέγγιση αυτή φάνηκε πιο αποτελεσματική για το συγκεκριμένο μοντέλο, καθώς βελτίωσε σημαντικά το Recall σε σχέση με το baseline.

Στα πειράματα με SMOTE, το XGBoost παρουσίασε βελτίωση σε σχέση με το baseline, αλλά όχι στον ίδιο βαθμό με το `scale_pos_weight`. Με SMOTE ratio 0,7 το Recall ήταν περίπου 0,20. Το αποτέλεσμα αυτό σημαίνει ότι το μοντέλο εντόπισε περισσότερα περιστατικά εγκεφαλικού σε σχέση με το baseline, αλλά λιγότερα σε σχέση με τη στάθμιση της θετικής κλάσης. Το Borderline-SMOTE είχε παρόμοια συμπεριφορά, με Recall περίπου 0,18 ενώ το SMOTE-Tomek έφτασε περίπου στο 0,24. Από τις SMOTE-based τεχνικές, το SMOTE-Tomek φάνηκε να είναι η καλύτερη για το XGBoost, χωρίς όμως να ξεπερνά το `scale_pos_weight`.

Η συμπεριφορά αυτή δείχνει ότι το XGBoost δεν επωφελήθηκε τόσο πολύ από την προσθήκη νέων συνθετικών δειγμάτων όσο από την προσαρμογή της διαδικασίας μάθησης. Μία πιθανή ερμηνεία είναι ότι το XGBoost, ως αλγόριθμος boosting, εκπαιδεύεται διαδοχικά προσπαθώντας να διορθώσει τα λάθη των προηγούμενων δέντρων. Κατά συνέπεια, όταν αυξάνεται η βαρύτητα των θετικών δειγμάτων, το μοντέλο μπορεί να κατευθύνει καλύτερα τη διαδικασία εκπαίδευσης προς τη μειοψηφική κλάση. Αντίθετα, τα συνθετικά δείγματα από SMOTE ή CTGAN δεν φαίνεται να πρόσφεραν στο συγκεκριμένο μοντέλο την ίδια βελτίωση.

Τα αποτελέσματα με CTGAN στο XGBoost ήταν περιορισμένα. Στο σενάριο CTGAN με ratio 0,5 το Recall ήταν περίπου 0,06 το F1-score περίπου 0,05 και το PR-AUC περίπου 0,09. Η απόδοση αυτή ήταν χαμηλή και δεν έδειξε ουσιαστική βελτίωση σε σχέση με τις υπόλοιπες τεχνικές. Αυτό σημαίνει ότι τα συνθετικά δεδομένα που δημιουργήθηκαν από το CTGAN δεν βοήθησαν το XGBoost να αναγνωρίσει αποτελεσματικότερα τη θετική κλάση.

Το Σχήμα 4.7 παρουσιάζει τη σύγκριση βασικών τεχνικών που εφαρμόστηκαν στο XGBoost. Παρατηρείται ότι η χρήση της παραμέτρου `scale_pos_weight` οδήγησε στην καλύτερη συμπεριφορά, καθώς πέτυχε το υψηλότερο Recall και F1-score σε σχέση με τις υπόλοιπες τεχνικές. Αντίθετα, το SMOTE-Tomek και ιδιαίτερα το CTGAN είχαν χαμηλότερη απόδοση, γεγονός που δείχνει ότι για το XGBoost η στάθμιση της θετικής κλάσης ήταν πιο αποτελεσματική από την παραγωγή ή τροποποίηση των δεδομένων εκπαίδευσης.



Σχήμα 4.6: XGBoost σύγκριση τεχνικών

Με βάση τη συνολική εικόνα, το XGBoost παρουσίασε καλύτερη συμπεριφορά από το Random Forest, αλλά δεν έφτασε τα υψηλά επίπεδα Recall που πέτυχε η Logistic Regression με class weights ή SMOTE-based τεχνικές. Το σημαντικότερο εύρημα για το XGBoost είναι ότι ανταποκρίθηκε καλύτερα στη στάθμιση της θετικής κλάσης παρά στην παραγωγή συνθετικών δεδομένων. Αυτό δείχνει ότι η επιλογή της κατάλληλης τεχνικής εξισορρόπησης πρέπει να γίνεται με βάση το μοντέλο και όχι με την υπόθεση ότι μια μέθοδος θα λειτουργήσει το ίδιο καλά σε όλους τους αλγορίθμους.

Πίνακας 4.9: Σύγκριση τεχνικών εξισορρόπησης για το XGBoost.

Τεχνική	Accuracy	Precision	Recall	F1 - Score	ROC-AUC	PR-AUC
Baseline	0.94	0.28	0.04	0.07	0.82	0.20
Class weight	0.86	0.18	0.5	0.26	0.82	0.20
SMOTE	0.91	0.15	0.2	0.17	0.78	0.13
Borderline SMOTE	0.91	0.17	0.18	0.17	0.79	0.15
SMOTE Tomek	0.91	0.18	0.24	0.20	0.79	0.14
CTGAN	0.90	0.05	0.06	0.05	0.72	0.09

Από τον Πίνακα 4.9 προκύπτει ότι, το XGBoost παρουσίασε την καλύτερη συμπεριφορά όταν χρησιμοποιήθηκε η τεχνική class weight με την παράμετρο scale_pos_weight. Συγκεκριμένα, η τεχνική αυτή οδήγησε στο υψηλότερο Recall, με τιμή 0.5, και στο καλύτερο F1-score, με τιμή 0.26, σε σχέση με τις υπόλοιπες τεχνικές. Το baseline μοντέλο εμφάνισε σχετικά υψηλό Accuracy και Precision, όμως το Recall ήταν πολύ χαμηλό, γεγονός που δείχνει ότι εντόπισε ελάχιστα πραγματικά περιστατικά

εγκεφαλικού. Αντίθετα, οι τεχνικές SMOTE βελτίωσαν σε κάποιο βαθμό την ικανότητα εντοπισμού της θετικής κλάσης, χωρίς όμως να ξεπεράσουν τη χρήση `scale_pos_weight`. Το CTGAN παρουσίασε τη χαμηλότερη απόδοση, κάτι που δείχνει ότι τα συνθετικά δεδομένα δεν βοήθησαν ουσιαστικά το XGBoost στο συγκεκριμένο πειραματικό πλαίσιο.

4.3.4 Συνολική σύγκριση και ερμηνεία

Η συνολική σύγκριση των αποτελεσμάτων δείχνει ότι η αντιμετώπιση της ανισορροπίας είχε καθοριστική επίδραση στη συμπεριφορά των μοντέλων, αλλά η επίδραση αυτή δεν ήταν ίδια για όλους τους αλγορίθμους. Η Logistic Regression ήταν το μοντέλο που παρουσίασε τη μεγαλύτερη βελτίωση όταν εφαρμόστηκαν τεχνικές όπως τα `class weights`, το SMOTE και το Borderline-SMOTE. Παρόλο που στο baseline σενάριο σχεδόν αγνοούσε τη θετική κλάση, μετά την εφαρμογή των τεχνικών εξισορρόπησης κατάφερε να εντοπίσει πολύ μεγαλύτερο ποσοστό των πραγματικών περιστατικών εγκεφαλικού.

Το Random Forest, αντίθετα, δυσκολεύτηκε περισσότερο. Η χρήση `class weights` δεν οδήγησε σε βελτίωση, ενώ οι τεχνικές υπερδειγματοληψίας προσέφεραν μόνο περιορισμένη αύξηση στο Recall. Η καλύτερη συμπεριφορά του παρατηρήθηκε με το Borderline-SMOTE, όμως η συνολική του απόδοση ως προς τη θετική κλάση παρέμεινε χαμηλή. Αυτό δείχνει ότι, στο συγκεκριμένο dataset, το Random Forest δεν κατάφερε να αξιοποιήσει αποτελεσματικά την ενίσχυση της μειοψηφικής κλάσης.

Το XGBoost παρουσίασε πιο ισορροπημένη συμπεριφορά σε σχέση με το Random Forest, αλλά η καλύτερη τεχνική για αυτό ήταν η εσωτερική στάθμιση της θετικής κλάσης μέσω `scale_pos_weight`. Οι SMOTE-based τεχνικές βελτίωσαν την απόδοση σε σχέση με το baseline, αλλά όχι στον ίδιο βαθμό. Το CTGAN δεν φάνηκε να προσφέρει ουσιαστικό πλεονέκτημα στο XGBoost, καθώς η απόδοση με συνθετικά δεδομένα παρέμεινε χαμηλή.

Ένα από τα βασικά συμπεράσματα που προκύπτει από την ανάλυση είναι ότι δεν υπάρχει μία τεχνική που να είναι καθολικά καλύτερη για όλα τα μοντέλα. Για τη Logistic Regression, οι τεχνικές `class weights` και Borderline-SMOTE ήταν ιδιαίτερα αποτελεσματικές. Για το Random Forest, το Borderline-SMOTE ήταν η καλύτερη από τις τεχνικές που δοκιμάστηκαν, αλλά με περιορισμένη βελτίωση. Για το XGBoost, η πιο αποτελεσματική προσέγγιση ήταν το `scale_pos_weight`. Επομένως, η επιλογή της κατάλληλης τεχνικής εξισορρόπησης πρέπει να γίνεται εμπειρικά και σε συνδυασμό με τον αλγόριθμο που χρησιμοποιείται.

Επιπλέον, τα αποτελέσματα έδειξαν ότι το CTGAN, παρόλο που αποτελεί πιο σύγχρονη και σύνθετη μέθοδο παραγωγής συνθετικών δεδομένων, δεν ξεπέρασε σταθερά τις πιο κλασικές τεχνικές εξισορρόπησης. Αυτό δεν σημαίνει ότι η μέθοδος δεν έχει αξία, αλλά δείχνει ότι η παραγωγή συνθετικών δεδομένων είναι ιδιαίτερα απαιτητική διαδικασία. Τα *synthetic samples* πρέπει να αποτυπώνουν με ρεαλιστικό τρόπο την κατανομή της μειοψηφικής κλάσης, διαφορετικά μπορεί να μη βελτιώσουν ουσιαστικά την απόδοση των μοντέλων στο πραγματικό test set.

Το σημαντικότερο εύρημα της ενότητας αυτής είναι ότι η απόδοση των μοντέλων δεν πρέπει να αξιολογείται μόνο με βάση το πόσο σύνθετος είναι ο αλγόριθμος ή πόσο προηγμένη είναι η τεχνική εξισορρόπησης. Στο συγκεκριμένο πρόβλημα, πιο απλές προσεγγίσεις, όπως η στάθμιση κλάσεων και το SMOTE, αποδείχθηκαν σε αρκετές περιπτώσεις πιο χρήσιμες από πιο σύνθετες μεθόδους. Αυτό δείχνει ότι σε ιατρικά προβλήματα με έντονη ανισορροπία, η πειραματική αξιολόγηση είναι απαραίτητη, επειδή η θεωρητική πολυπλοκότητα μιας μεθόδου δεν εγγυάται απαραίτητα καλύτερη πρακτική απόδοση.

Συνολικά, η ανάλυση της συμπεριφοράς των μοντέλων επιβεβαιώνει ότι η πρόβλεψη εγκεφαλικού επεισοδίου σε ανισόρροπα δεδομένα αποτελεί ένα απαιτητικό πρόβλημα. Τα baseline μοντέλα εμφάνισαν περιορισμένη ικανότητα εντοπισμού της θετικής κλάσης, ενώ οι τεχνικές εξισορρόπησης βελτίωσαν την απόδοση σε διαφορετικό βαθμό ανάλογα με το μοντέλο. Η συστηματική σύγκριση όλων των τεχνικών επέτρεψε να φανεί ότι η βελτίωση του Recall είναι εφικτή, αλλά συνήθως συνοδεύεται από αύξηση των false positives και μείωση του Precision. Ως αποτέλεσμα, η αξιολόγηση των μοντέλων πρέπει να γίνεται με βάση τη συνολική ισορροπία των μετρικών και όχι μεμονωμένα με βάση μία μόνο τιμή.

Πίνακας 4.10: Σύγκριση βέλτιστων επιδόσεων ανά μοντέλο (Real Test Set).

Μοντέλο	Καλύτερη Τεχνική	Recall	F1 - Score	PR-AUC	Accuracy
Logistic Regression	Borderline SMOTE	0.78	0.29	0.27	0.81
Random Forest	Borderline SMOTE	0.16	0.18	0.14	0.93
XGBoost	scale_pos_weight	0.50	0.26	0.20	0.86

Όπως φαίνεται στον Πίνακα 4.10, η Logistic Regression με Borderline-SMOTE παρουσίασε την καλύτερη συνολική συμπεριφορά στο πραγματικό test set. Παρόλο που η τεχνική των class weights πέτυχε οριακά υψηλότερο Recall, το Borderline-SMOTE πέτυχε πιο ισορροπημένη απόδοση, καθώς διατήρησε υψηλή ανάκληση και ταυτόχρονα παρουσίασε το υψηλότερο F1-score. Αυτό δείχνει ότι το μοντέλο κατάφερε να εντοπίσει μεγάλο μέρος των πραγματικών περιστατικών εγκεφαλικού, χωρίς να αυξήσει υπερβολικά τις λανθασμένες θετικές προβλέψεις. Αντίθετα, το Random Forest παρουσίασε περιορισμένη ικανότητα εντοπισμού της θετικής κλάσης, ακόμη και μετά την εφαρμογή τεχνικών εξισορρόπησης. Το καλύτερο αποτέλεσμά του εμφανίστηκε με το Borderline-SMOTE, όμως το Recall παρέμεινε χαμηλό. Το XGBoost είχε καλύτερη συμπεριφορά από το Random Forest, κυρίως με τη χρήση της class weight, η οποία βοήθησε το μοντέλο να αναγνωρίσει περισσότερα περιστατικά της μειοψηφικής κλάσης.

Η δυσκολία του CTGAN να ξεπεράσει τις SMOTE-based τεχνικές ενδεχομένως οφείλεται στο περιορισμένο μέγεθος του αρχικού dataset, το οποίο περιλαμβάνει 5.110 εγγραφές. Επιπλέον, τα δείγματα της θετικής κλάσης είναι πολύ λίγα σε σχέση με την αρνητική κλάση, γεγονός που πιθανόν δεν επέτρεψε την πλήρη εκπαίδευση ενός παραγωγικού μοντέλου, όπως το GAN, ώστε να μάθει με ακρίβεια τις λεπτές στατιστικές σχέσεις της μειοψηφικής κλάσης. Για τον λόγο αυτό, τα συνθετικά δείγματα που παράχθηκαν από το CTGAN δεν φαίνεται να βελτίωσαν την απόδοση των ταξινομητών στον ίδιο βαθμό με τις τεχνικές SMOTE.

Από ιατρική σκοπιά, η επιλογή της Logistic Regression με Borderline-SMOTE κρίνεται ως η πιο ενδεδειγμένη για το συγκεκριμένο πειραματικό πλαίσιο. Σε ένα πρόβλημα προληπτικού ελέγχου, όπως η πρόβλεψη εγκεφαλικού επεισοδίου, προτεραιότητα έχει η ελαχιστοποίηση των ψευδώς αρνητικών προβλέψεων, ώστε να μη διαφεύγουν πραγματικά περιστατικά υψηλού κινδύνου. Επομένως, ένα μοντέλο με υψηλό Recall και παράλληλα αποδεκτή συνολική ισορροπία, όπως η Logistic Regression με Borderline-SMOTE, είναι πιο κατάλληλο από ένα μοντέλο που παρουσιάζει υψηλό Accuracy αλλά αποτυγχάνει να εντοπίσει τη θετική κλάση.

4.4 Πειραματική Ανάλυση με Εξισορρόπηση του Test Set

Εκτός από την κύρια αξιολόγηση των μοντέλων στο πραγματικό και ανισόρροπο test set, στην παρούσα εργασία πραγματοποιήθηκε και ένα επιπλέον πειραματικό σενάριο, με στόχο να εξεταστεί πώς μεταβάλλεται η συμπεριφορά των μοντέλων όταν το test set εξισορροπείται τεχνητά. Το συγκεκριμένο πείραμα δεν αποτελεί την κύρια αξιολόγηση της εργασίας, αλλά χρησιμοποιείται ως συμπληρωματική ανάλυση, ώστε να γίνει πιο κατανοητή η επίδραση της κατανομής των κλάσεων στις τελικές μετρικές απόδοσης.

Στο βασικό πειραματικό μέρος της εργασίας, το test set διατηρήθηκε στην αρχική του ανισόρροπη μορφή, καθώς αυτό θεωρήθηκε πιο ρεαλιστικό για το πρόβλημα της πρόβλεψης εγκεφαλικού επεισοδίου. Πιο συγκεκριμένα, το test set περιλάμβανε 972 περιπτώσεις χωρίς εγκεφαλικό επεισόδιο και 50 περιπτώσεις με εγκεφαλικό επεισόδιο. Η κατανομή αυτή είναι σημαντική, επειδή αντικατοπτρίζει τη φύση του προβλήματος, όπου τα θετικά περιστατικά είναι πολύ λιγότερα από τα αρνητικά. Με βάση τα παραπάνω, όλα τα βασικά αποτελέσματα της εργασίας βασίζονται στην αξιολόγηση των μοντέλων πάνω σε αυτό το πραγματικό, ανισόρροπο test set.

Ωστόσο, στο πλαίσιο της πειραματικής διερεύνησης, θεωρήθηκε χρήσιμο να εξεταστεί και ένα δεύτερο σενάριο, στο οποίο εφαρμόστηκε SMOTE και στο test set. Η χρήση αυτής της μεθόδου, έγινε όχι για να παρουσιαστεί ως ρεαλιστική μέθοδος αξιολόγησης, αλλά για να μελετηθεί πειραματικά πώς επηρεάζονται οι μετρικές όταν η θετική κλάση ενισχύεται και στο σύνολο ελέγχου. Με άλλα λόγια, το συγκεκριμένο πείραμα δεν απαντά στο ερώτημα «πώς θα αποδώσει το μοντέλο σε πραγματικά δεδομένα», αλλά στο ερώτημα «πώς αλλάζει η αριθμητική συμπεριφορά των μοντέλων όταν η αξιολόγηση γίνεται σε πιο ισορροπημένη κατανομή».

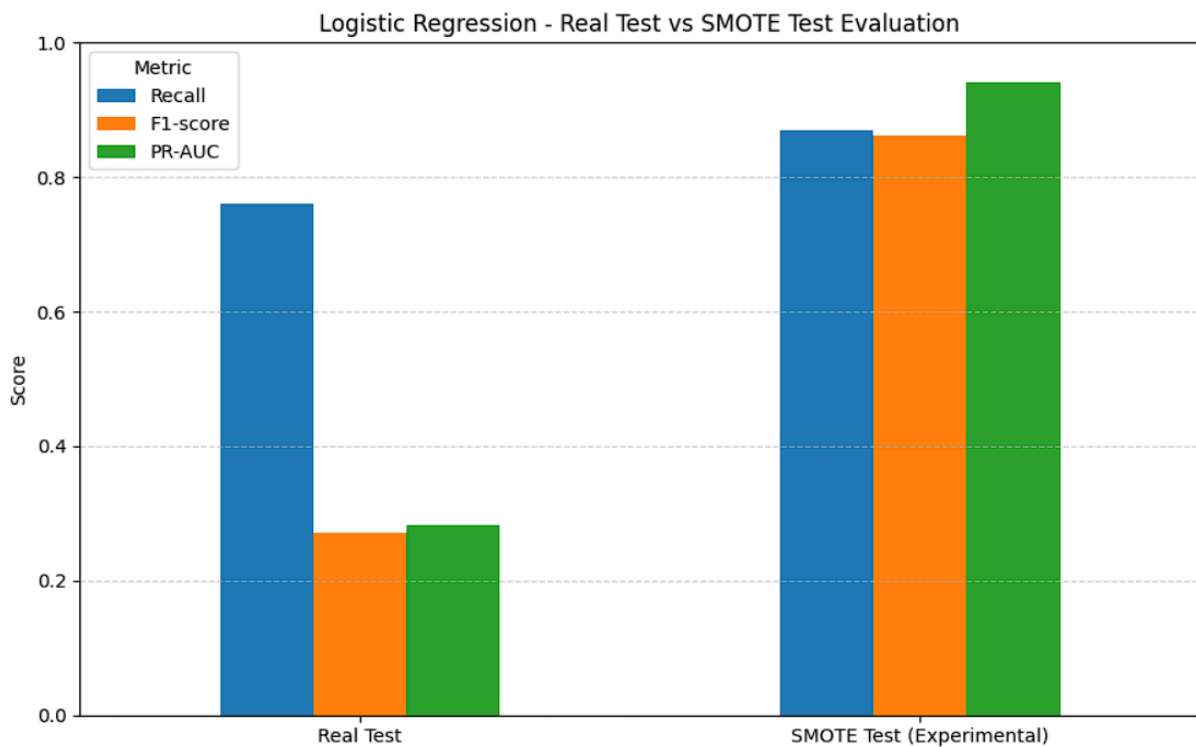
Η διάκριση αυτή είναι ιδιαίτερα σημαντική. Σε πραγματικές συνθήκες, δεν είναι σωστό να εξισορροπείται το test set πριν από την κύρια αξιολόγηση, επειδή το test set πρέπει να αντιπροσωπεύει όσο το δυνατόν καλύτερα την πραγματική κατανομή των δεδομένων. Αν το test set τροποποιηθεί τεχνητά, τότε η αξιολόγηση μπορεί να δώσει πολύ πιο αισιόδοξα αποτελέσματα από αυτά που θα παρατηρούνταν σε πραγματική εφαρμογή. Για αυτόν τον λόγο, στην παρούσα εργασία το πείραμα αυτό παρουσιάζεται ξεκάθαρα ως συμπληρωματική και πειραματική ανάλυση, και όχι ως βασική μέθοδος αξιολόγησης των μοντέλων.

Η διαδικασία που ακολουθήθηκε αναλύεται ως εξής. Αρχικά επιλέχθηκε μία αντιπροσωπευτική τεχνική από κάθε μοντέλο, με βάση τα προηγούμενα πειράματα στο πραγματικό test set. Η επιλογή της συγκεκριμένης τεχνικής, δεν σημαίνει ότι ορίστηκε ένας τελικός ή μοναδικός καλύτερος εκπρόσωπος για κάθε μοντέλο στο σύνολο της εργασίας. Αντίθετα, έγινε μόνο για τις ανάγκες του συγκεκριμένου πειραματικού σεναρίου, ώστε να μπορεί να πραγματοποιηθεί μια πιο καθαρή σύγκριση μεταξύ της αξιολόγησης στο πραγματικό ανισόρροπο test set και της αξιολόγησης στο εξισορροπημένο test set.

Συγκεκριμένα, για τη Logistic Regression χρησιμοποιήθηκε το μοντέλο που είχε εκπαιδευτεί με SMOTE, για το Random Forest χρησιμοποιήθηκε το αντίστοιχο μοντέλο με τεχνική υπερδειγματοληψίας, ενώ για το XGBoost χρησιμοποιήθηκε το μοντέλο που είχε παρουσιάσει την πιο αντιπροσωπευτική συμπεριφορά στο προηγούμενο στάδιο των πειραμάτων. Στη συνέχεια, εφαρμόστηκε SMOTE στο test set, ώστε να αυξηθεί ο αριθμός των θετικών περιπτώσεων και να δημιουργηθεί ένα πιο ισορροπημένο σύνολο αξιολόγησης. Τα μοντέλα αξιολογήθηκαν εκ νέου πάνω σε αυτό το τροποποιημένο test set και καταγράφηκαν οι ίδιες μετρικές αξιολόγησης, δηλαδή Accuracy, Precision, Recall, F1-score, ROC-AUC και PR-AUC.

Όπως παρουσιάζεται και στο Σχήμα 4.7, τα αποτελέσματα της πειραματικής αυτής αξιολόγησης παρουσίασαν σημαντική διαφοροποίηση σε σχέση με την κύρια αξιολόγηση στο πραγματικό test set. Για τη Logistic Regression, το Recall αυξήθηκε σε περίπου 0,8691, το F1-score σε περίπου 0,8609 και το PR-AUC σε περίπου 0,9415. Οι τιμές αυτές είναι αισθητά υψηλότερες από εκείνες που παρατηρήθηκαν στο πραγματικό ανισόρροπο test set. Η διαφορά αυτή φαίνεται καθαρά και στο διάγραμμα, όπου οι μπάρες του SMOTE Test είναι σημαντικά υψηλότερες σε όλες τις βασικές μετρικές σε σύγκριση με το Real Test.

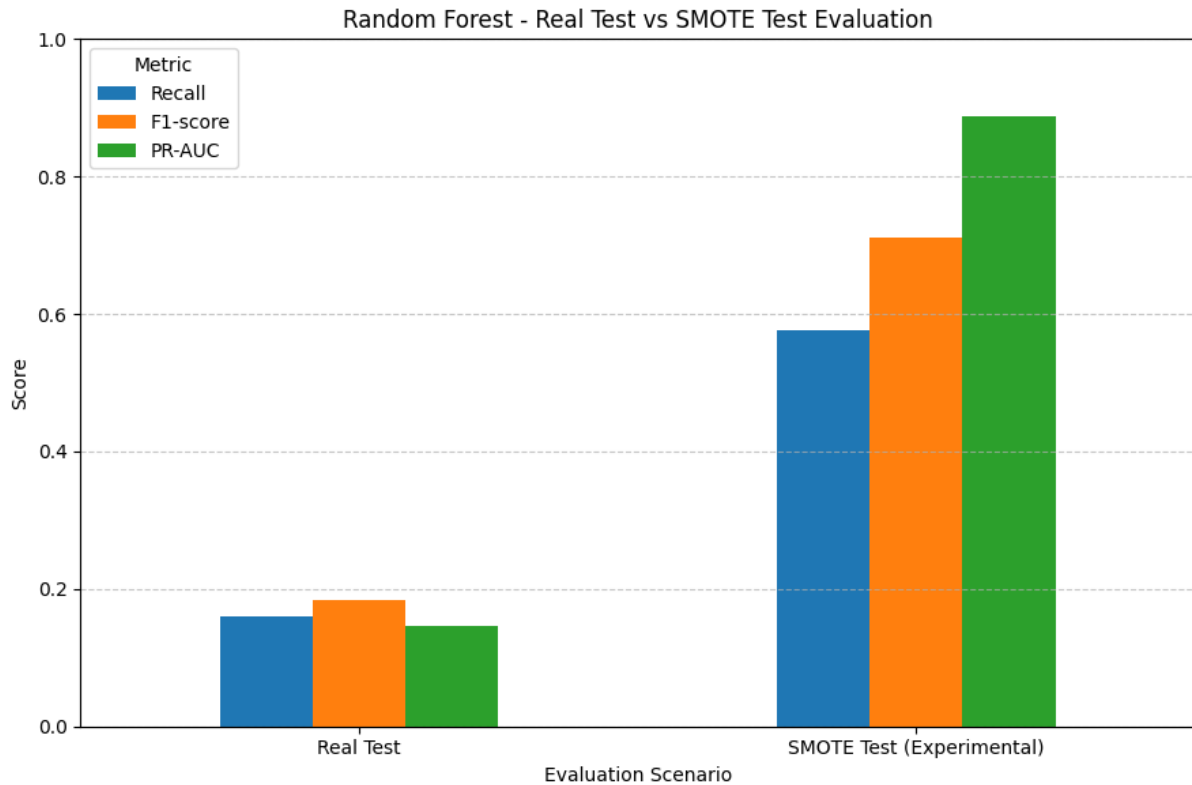
Η βελτίωση αυτή δείχνει ότι, όταν η θετική κλάση ενισχύεται και στο test set, το μοντέλο εμφανίζει πολύ καλύτερη αριθμητική συμπεριφορά, επειδή αξιολογείται σε μια κατανομή όπου τα θετικά περιστατικά είναι περισσότερα και επομένως η κλάση του εγκεφαλικού εκπροσωπείται πιο ισορροπημένα. Ωστόσο, το αποτέλεσμα αυτό πρέπει να ερμηνευθεί με προσοχή, καθώς το SMOTE-balanced test set δεν αντικατοπτρίζει την πραγματική ανισορροπία του προβλήματος. Για τον λόγο αυτό, η συγκεκριμένη αξιολόγηση χρησιμοποιείται μόνο ως συμπληρωματικό πειραματικό σενάριο και όχι ως βασικό κριτήριο τελικής απόδοσης του μοντέλου.



Σχήμα 4.7: Real test set – SMOTE test set Logistic Regression

Όπως φαίνεται στο Σχήμα 4.8, αντίστοιχη αύξηση παρατηρήθηκε και στην περίπτωση του Random Forest κατά την αξιολόγηση στο SMOTE-balanced test set. Συγκεκριμένα, το Recall έφτασε περίπου στο 0,5765, το F1-score στο 0,7121 και το PR-AUC στο 0,8886. Οι τιμές αυτές είναι αισθητά υψηλότερες σε σχέση με την κύρια αξιολόγηση στο πραγματικό ανισόρροπο test set, όπου το Random Forest παρουσίαζε περιορισμένη ικανότητα εντοπισμού της θετικής κλάσης.

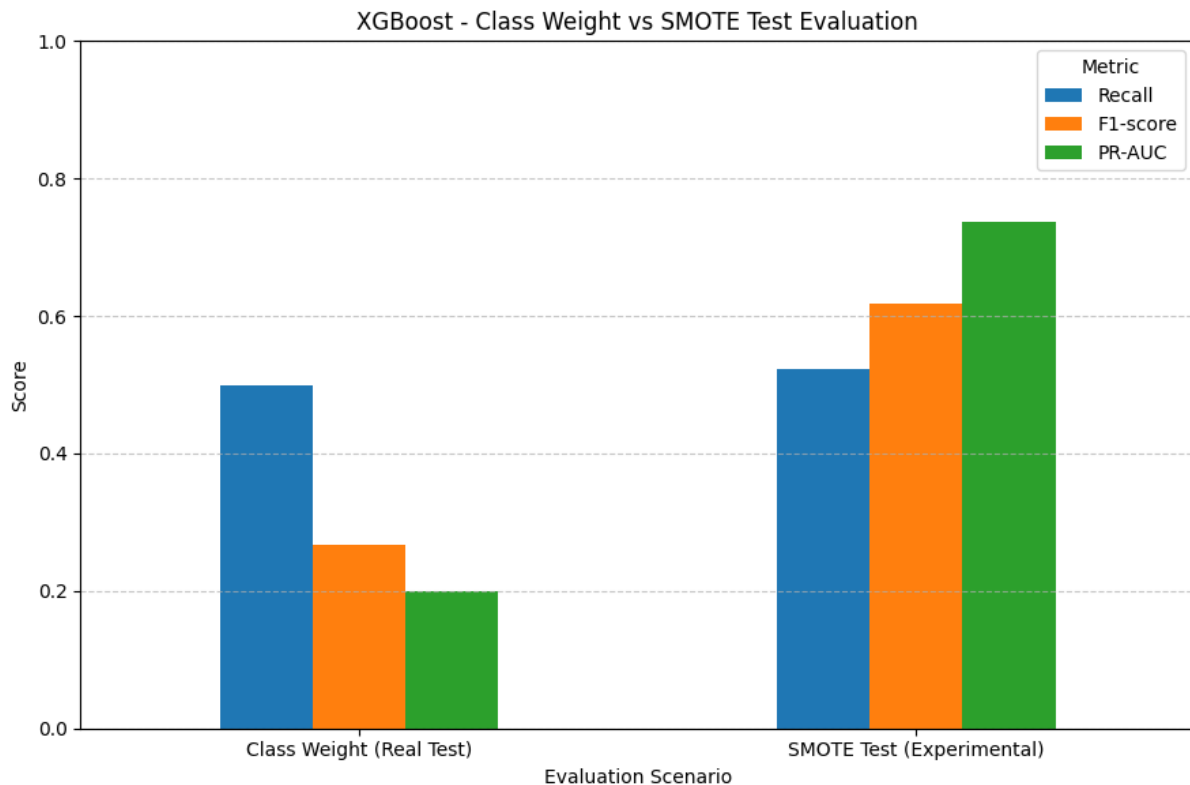
Η διαφοροποίηση αυτή αποτυπώνεται καθαρά και στο διάγραμμα, καθώς οι μπάρες του SMOTE Test είναι σημαντικά υψηλότερες από τις αντίστοιχες του Real Test και στις τρεις βασικές μετρικές. Αυτό δείχνει ότι η ενίσχυση της θετικής κλάσης επηρεάζει έντονα την εικόνα της απόδοσης του μοντέλου.



Σχήμα 4.8: Real test set – SMOTE test set Random Forest

Όπως παρουσιάζεται στο Σχήμα 4.9, και στην περίπτωση του XGBoost η αξιολόγηση στο SMOTE-balanced test set οδήγησε σε υψηλότερες τιμές σε σχέση με την κύρια αξιολόγηση στο πραγματικό test set. Πιο συγκεκριμένα, στο πειραματικό σενάριο το Recall έφτασε περίπου στο 0,5221, το F1-score στο 0,6190 και το PR-AUC στο 0,7378. Οι τιμές αυτές είναι αυξημένες σε σύγκριση με τις αντίστοιχες τιμές του πραγματικού test set, όπου το XGBoost με χρήση `scale_pos_weight` είχε Recall 0,50, F1-score 0,26 και PR-AUC 0,20.

Η διαφορά αυτή αποτυπώνεται καθαρά και στο διάγραμμα, καθώς οι μπάρες του SMOTE Test είναι υψηλότερες κυρίως ως προς το F1-score και το PR-AUC. Αυτό δείχνει ότι, όπως και στα υπόλοιπα μοντέλα, η εξισορρόπηση του test set οδηγεί σε πιο ευνοϊκή εικόνα απόδοσης. Παρόλο που στο XGBoost η αύξηση του Recall είναι σχετικά μικρότερη σε σχέση με τα άλλα μοντέλα, η συνολική βελτίωση των μετρικών παραμένει εμφανής, όχι ως κύρια αξιολόγηση του μοντέλου.



Σχήμα 4.9: Real test set – SMOTE test set XGBoost

Η σημαντική αύξηση των μετρικών στο πείραμα αυτό δεν πρέπει να ερμηνευθεί ως πραγματική βελτίωση της ικανότητας γενίκευσης των μοντέλων. Τα μοντέλα δεν άλλαξαν ως προς τον τρόπο εκπαίδευσης στο στάδιο αυτό, αλλά αυτό που άλλαξε ήταν η κατανομή των δεδομένων αξιολόγησης. Συνεπώς, οι αυξημένες τιμές των μετρικών οφείλονται σε μεγάλο βαθμό στο γεγονός ότι το test set έγινε πιο ισορροπημένο και όχι απαραίτητα στο ότι τα μοντέλα έγιναν πιο κατάλληλα για πραγματική κλινική εφαρμογή. Ως εκ τούτου, τα αποτελέσματα του συγκεκριμένου πειράματος πρέπει να διαβαστούν με προσοχή.

Έπειτα από τα αποτελέσματα αυτά διαπιστώνουμε ότι, το πείραμα αυτό ήταν χρήσιμο γιατί ανέδειξε με πρακτικό τρόπο πόσο μεγάλη επίδραση έχει η κατανομή των κλάσεων στις τελικές μετρικές. Στην κύρια αξιολόγηση, τα μοντέλα έπρεπε να εντοπίσουν λίγες θετικές περιπτώσεις μέσα σε έναν μεγάλο αριθμό αρνητικών περιπτώσεων, κάτι που είναι πολύ πιο δύσκολο και πιο κοντά στην πραγματικότητα. Αντίθετα, στο εξισορροπημένο test set, η θετική κλάση είχε πολύ μεγαλύτερη παρουσία, με αποτέλεσμα οι μετρικές όπως το Recall, το F1-score και το PR-AUC να βελτιώνονται σημαντικά. Αυτό δείχνει ότι οι μετρικές δεν πρέπει ποτέ να ερμηνεύονται ανεξάρτητα από την κατανομή των δεδομένων πάνω στα οποία υπολογίζονται.

Ιδιαίτερα το PR-AUC επηρεάστηκε έντονα από την εξισορρόπηση του test set. Αυτό είναι αναμενόμενο, καθώς η καμπύλη Precision-Recall είναι πολύ ευαίσθητη στη συχνότητα της θετικής κλάσης. Όταν η θετική κλάση είναι σπάνια, όπως στο πραγματικό test set, το Precision μειώνεται εύκολα λόγω των false positives και η συνολική τιμή του PR-AUC παραμένει χαμηλότερη. Όταν όμως η θετική κλάση αυξάνεται τεχνητά, η σχέση μεταξύ Precision και Recall αλλάζει και οι τιμές του PR-AUC μπορούν να γίνουν πολύ υψηλότερες. Για αυτόν τον λόγο, η σύγκριση PR-AUC μεταξύ ανισόρροπου και εξισορροπημένου test set πρέπει να γίνεται μόνο ως πειραματική παρατήρηση και όχι ως ισοδύναμη αξιολόγηση.

Το ίδιο ισχύει και για το F1-score. Σε ένα πραγματικό ανισόρροπο test set, ακόμη και αν ένα μοντέλο εντοπίσει αρκετά θετικά περιστατικά, μπορεί να έχει χαμηλό F1-score λόγω αυξημένων false positives ή χαμηλού Precision. Αντίθετα, σε ένα εξισορροπημένο test set, η ύπαρξη περισσότερων θετικών περιπτώσεων μπορεί να οδηγήσει σε καλύτερη ισορροπία μεταξύ Precision και Recall. Αυτό εξηγεί γιατί στο πειραματικό σενάριο οι τιμές του F1-score αυξήθηκαν σημαντικά για όλα τα μοντέλα.

Συνεπώς, το βασικό συμπέρασμα από την πειραματική αυτή ανάλυση δεν είναι ότι τα μοντέλα είναι απαραίτητα πολύ πιο αποτελεσματικά σε πραγματικές συνθήκες, αλλά ότι η κατανομή του test set επηρεάζει σε μεγάλο βαθμό την εικόνα της απόδοσης. Η κύρια και πιο αξιόπιστη αξιολόγηση παραμένει αυτή που έγινε στο πραγματικό ανισόρροπο test set, επειδή αυτό αντιπροσωπεύει καλύτερα το πρόβλημα της πρόβλεψης εγκεφαλικού επεισοδίου. Το πείραμα με SMOTE στο test set λειτουργεί συμπληρωματικά, δείχνοντας πώς οι μετρικές μεταβάλλονται όταν αφαιρεθεί, σε έναν βαθμό, η δυσκολία της σπάνιας θετικής κλάσης.

Από το Σχήμα 4.8 φαίνεται η διαφορά ανάμεσα στην αξιολόγηση στο πραγματικό ανισόρροπο test set και στην πειραματική αξιολόγηση με SMOTE-balanced test set. Στο πραγματικό test set, το Recall παραμένει υψηλό, όμως το F1-score και το PR-AUC είναι αρκετά χαμηλότερα, λόγω της έντονης ανισορροπίας και των πολλών λανθασμένων θετικών προβλέψεων. Αντίθετα, στο SMOTE test set οι τιμές των μετρικών αυξάνονται σημαντικά, κάτι που δείχνει ότι η εξισορρόπηση του test set δημιουργεί πιο ευνοϊκές συνθήκες αξιολόγησης.

Συμπερασματικά, η πειραματική ανάλυση με εξισορροπημένο test set επιβεβαιώνει ότι η αξιολόγηση μοντέλων σε ανισόρροπα ιατρικά δεδομένα πρέπει να γίνεται με ιδιαίτερη προσοχή. Τα αποτελέσματα μπορούν να διαφέρουν σημαντικά ανάλογα με την κατανομή των κλάσεων στο σύνολο αξιολόγησης. Έτσι, τα αποτελέσματα του SMOTE-balanced test set δεν χρησιμοποιούνται ως κύρια βάση συμπερασμάτων, αλλά ως πρόσθετη ανάλυση που βοηθά στην κατανόηση της επίδρασης της ανισορροπίας στις μετρικές απόδοσης. Η προσέγγιση αυτή βοηθά, να γίνει πιο ξεκάθαρο ότι η πραγματική πρόκληση δεν είναι απλώς να αυξηθούν οι αριθμητικές τιμές των μετρικών, αλλά να αναπτυχθούν μοντέλα που μπορούν να εντοπίζουν αξιόπιστα τη θετική κλάση σε πραγματικά ανισόρροπες συνθήκες.

Ο Πίνακας 4.11 παρουσιάζει τη σύγκριση των βασικών μετρικών Recall, F1-score και PR-AUC ανάμεσα στο πραγματικό test set και στο SMOTE-balanced test set. Όπως φαίνεται, η αξιολόγηση στο εξισορροπημένο test set οδηγεί σε σημαντικά υψηλότερες τιμές, ιδιαίτερα ως προς το F1-score και το PR-AUC. Η αύξηση αυτή είναι αναμενόμενη, καθώς το SMOTE-balanced test set περιέχει πιο ισορροπημένη κατανομή των δύο κλάσεων και επομένως δεν αντικατοπτρίζει πλήρως τη δυσκολία του αρχικού προβλήματος.

Παρόλα αυτά, τα αποτελέσματα αυτά δεν χρησιμοποιούνται ως κύρια αξιολόγηση της εργασίας, αλλά μόνο ως συμπληρωματική πειραματική ανάλυση. Το πραγματικό test set παραμένει πιο αξιόπιστο για την τελική ερμηνεία, επειδή διατηρεί την αρχική ανισορροπία των δεδομένων και προσεγγίζει καλύτερα τις συνθήκες ενός πραγματικού προβλήματος πρόβλεψης εγκεφαλικού επεισοδίου.

Πίνακας 4.11: Σύγκριση Real Test Set – SMOTE Test Set.

Model	Recall		F1 - Score		PR - AUC	
	Real Test Set	SMOTE Test Set	Real Test Set	SMOTE Test Set	Real Test Set	SMOTE Test Set
Logistic Regression	0.76	0.87	0.27	0.86	0.28	0.94
Random Forest	0.16	0.57	0.18	0.71	0.14	0.88
XGBoost	0.50	0.52	0.26	0.62	0.20	0.73

Κεφάλαιο 5ο: Federated Learning

5.1 Σχεδιασμός Προσομοίωσης

Στο προηγούμενο κεφάλαιο η αξιολόγηση των μοντέλων πραγματοποιήθηκε με centralized τρόπο, δηλαδή με τη συγκέντρωση όλων των δεδομένων εκπαίδευσης σε ένα ενιαίο σύνολο. Η προσέγγιση αυτή είναι η πιο συνηθισμένη σε κλασικά πειράματα μηχανικής μάθησης, καθώς επιτρέπει στο μοντέλο να εκπαιδευτεί πάνω σε ολόκληρο το διαθέσιμο training set. Ωστόσο, στον χώρο της υγείας, η συγκέντρωση δεδομένων από διαφορετικούς φορείς δεν είναι πάντα εύκολη ή εφικτή. Τα ιατρικά δεδομένα είναι ευαίσθητα, καθώς συνδέονται άμεσα ή έμμεσα με την κατάσταση υγείας των ασθενών, και η ανταλλαγή τους μεταξύ νοσοκομείων ή οργανισμών μπορεί να περιορίζεται από ζητήματα ιδιωτικότητας και ασφάλειας. Για τον λόγο αυτό, στο παρόν κεφάλαιο εξετάζεται η χρήση του Federated Learning ως εναλλακτική προσέγγιση εκπαίδευσης. Η βασική ιδέα του Federated Learning είναι ότι τα δεδομένα δεν μεταφέρονται σε έναν κεντρικό server. Αντίθετα, παραμένουν τοπικά στους φορείς που τα κατέχουν, ενώ αυτό που ανταλλάσσεται είναι οι παράμετροι ή τα βάρη των μοντέλων. Με τη λογική αυτή, κάθε client εκπαιδεύει το μοντέλο στα δικά του δεδομένα και στη συνέχεια ο server συνδυάζει τα τοπικά μοντέλα, δημιουργώντας ένα νέο καθολικό μοντέλο. Η διαδικασία αυτή επιτρέπει τη συνεργατική εκπαίδευση ενός μοντέλου χωρίς να απαιτείται η άμεση συγκέντρωση των αρχικών δεδομένων [18].

Στην παρούσα διπλωματική εργασία δεν χρησιμοποιήθηκαν πραγματικά δεδομένα από διαφορετικά νοσοκομεία. Αντίθετα, πραγματοποιήθηκε προσομοίωση Federated Learning με βάση το ήδη διαθέσιμο training set του Stroke Prediction Dataset. Συγκεκριμένα, το αρχικό training set χωρίστηκε τεχνητά σε τρία υποσύνολα, τα οποία θεωρήθηκαν ως τρεις ανεξάρτητοι clients. Κάθε client προσομοιώνει έναν διαφορετικό φορέα υγείας ή νοσοκομείο, το οποίο διαθέτει το δικό του τοπικό σύνολο δεδομένων. Με τη συγκεκριμένη προσέγγιση, δημιουργήθηκε ένα ελεγχόμενο federated περιβάλλον, μέσα στο οποίο μπορούσε να εξεταστεί η συμπεριφορά της κατανεμημένης εκπαίδευσης. Ο σχεδιασμός της προσομοίωσης βασίστηκε σε τρία βασικά στοιχεία. Πρώτον, πραγματοποιήθηκε ο ίδιος διαχωρισμός στα train και test set όπως είχε χρησιμοποιηθεί και στα centralized πειράματα. Αυτό ήταν απαραίτητο ώστε τα αποτελέσματα του federated learning να μπορούν να συγκριθούν με δίκαιο τρόπο με τα προηγούμενα αποτελέσματα. Το test set δεν τροποποιήθηκε σε κανένα federated πείραμα και παρέμεινε ανισόρροπο, όπως ήταν και στην αρχική μορφή του. Έτσι, η αξιολόγηση έγινε σε πραγματικές συνθήκες ανισορροπίας και όχι σε τεχνητά εξισορροπημένο test set. Δεύτερον, το σύνολο εκπαίδευσης χωρίστηκε σε τρεις clients. Η επιλογή των τριών clients έγινε ώστε να υπάρχει μια απλή αλλά επαρκής προσομοίωση ενός federated συστήματος. Ένας πολύ μεγάλος αριθμός clients θα αύξανε την πολυπλοκότητα της υλοποίησης χωρίς να είναι απαραίτητος για τον σκοπό της εργασίας. Αντίθετα, οι τρεις clients επιτρέπουν να παρουσιαστεί καθαρά η λογική του Federated Learning, να εφαρμοστεί ο αλγόριθμος FedAvg και να εξεταστεί η επίδραση διαφορετικών κατανομών δεδομένων. Τρίτον, διατηρήθηκαν οι ίδιες βασικές μετρικές αξιολόγησης που χρησιμοποιήθηκαν και στα centralized πειράματα. Συγκεκριμένα, για κάθε federated πείραμα υπολογίστηκαν Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC και Confusion Matrix. Η διατήρηση των ίδιων μετρικών ήταν σημαντική, καθώς επέτρεψε την άμεση σύγκριση ανάμεσα στις centralized και federated προσεγγίσεις. Ιδιαίτερη έμφαση δόθηκε και εδώ στις μετρικές Recall, F1-score και PR-AUC, καθώς το πρόβλημα χαρακτηρίζεται από έντονη ανισορροπία κλάσεων.

Για την υλοποίηση του federated learning χρησιμοποιήθηκε μοντέλο Logistic Regression. Η επιλογή αυτή έγινε επειδή η λογιστική παλινδρόμηση είναι απλό, ερμηνεύσιμο και κατάλληλο μοντέλο για

δυναμική ταξινόμηση. Επιπλέον, σε αντίθεση με μοντέλα όπως το Random Forest ή το XGBoost, η Logistic Regression μπορεί να ενσωματωθεί πιο άμεσα σε διαδικασία Federated Averaging, καθώς το μοντέλο αποτελείται από βάρη και bias που μπορούν να συνδυαστούν αριθμητικά μεταξύ των clients. Με αυτόν τον τρόπο, η υλοποίηση του FedAvg έγινε καθαρή και κατανοητή, χωρίς να απαιτείται πιο σύνθετο federated framework. Στην προσομοίωση, κάθε round της federated εκπαίδευσης ακολουθούσε συγκεκριμένη διαδικασία. Αρχικά, ο server δημιουργούσε ή διατηρούσε ένα καθολικό μοντέλο (global model). Στη συνέχεια, το ίδιο ενιαίο μοντέλο αποστέλλόταν στους τρεις clients. Κάθε client εκπαιδευε τοπικά ένα αντίγραφο του μοντέλου πάνω στα δικά του δεδομένα για προκαθορισμένο αριθμό εποχών. Μετά την τοπική εκπαίδευση, οι clients επέστρεφαν μόνο τα ενημερωμένα βάρη του μοντέλου στον server. Ο server συνδυάζε τα βάρη αυτά μέσω του αλγορίθμου FedAvg και δημιουργούσε ένα νέο ενιαίο μοντέλο. Η διαδικασία αυτή επαναλήφθηκε για 20 γύρους επικοινωνίας. Σε κάθε γύρο επικοινωνίας, κάθε client εκπαιδευε το τοπικό του μοντέλο για 5 local epochs. Μετά την ολοκλήρωση της τοπικής εκπαίδευσης, οι ενημερωμένες παράμετροι κάθε client αποστέλλονταν στον server, ώστε να υπολογιστεί το νέο Καθολικό Μοντέλο μέσω FedAvg. Η διαδικασία αυτή επαναλήφθηκε για 20 συνολικούς γύρους επικοινωνίας. Είναι σημαντικό να διευκρινιστεί ότι η συγκεκριμένη υλοποίηση αποτελεί προσομοίωση Federated Learning και όχι πλήρες σύστημα παραγωγής. Δηλαδή, δεν υλοποιήθηκαν μηχανισμοί όπως secure aggregation, differential privacy ή πραγματική επικοινωνία μεταξύ απομακρυσμένων συστημάτων. Παρ' όλα αυτά, η προσομοίωση αποτυπώνει τη βασική λογική της federated εκπαίδευσης, καθώς τα δεδομένα χωρίζονται σε ανεξάρτητους clients και ο server λαμβάνει μόνο τις παραμέτρους των τοπικών μοντέλων. Για τον σκοπό της παρούσας εργασίας, αυτό ήταν επαρκές, καθώς το ζητούμενο ήταν να αξιολογηθεί αν μια federated προσέγγιση μπορεί να παρουσιάσει συγκρίσιμη συμπεριφορά με την centralized εκπαίδευση στο πρόβλημα πρόβλεψης εγκεφαλικού επεισοδίου.

Συνολικά, ο σχεδιασμός της προσομοίωσης είχε ως στόχο να δημιουργήσει ένα ελεγχόμενο πειραματικό περιβάλλον, μέσα στο οποίο μπορούσε να εξεταστεί η επίδραση του Federated Learning σε συνθήκες ανισόρροπων ιατρικών δεδομένων. Η προσέγγιση αυτή επέτρεψε τη σύγκριση ανάμεσα σε centralized και federated εκπαίδευση, αλλά και την αξιολόγηση τεχνικών αντιμετώπισης της ανισορροπίας, όπως τα class weights και το local SMOTE, μέσα σε federated περιβάλλον.

5.2 Σενάρια Κατανομής

Ένα από τα σημαντικότερα ζητήματα στο Federated Learning είναι ο τρόπος με τον οποίο κατανέμονται τα δεδομένα στους clients. Σε ένα κλασικό centralized πείραμα, όλα τα δεδομένα εκπαίδευσης βρίσκονται συγκεντρωμένα σε ένα ενιαίο σύνολο δεδομένων εκπαίδευσης. Αντίθετα, στο Federated Learning τα δεδομένα είναι κατανεμημένα σε διαφορετικούς clients, οι οποίοι μπορεί να έχουν διαφορετικά μεγέθη dataset, διαφορετικές αναλογίες κλάσεων και διαφορετικά χαρακτηριστικά πληθυσμού. Η παρατήρηση αυτή οδήγησε στο ότι, η κατανομή των δεδομένων μεταξύ των clients μπορεί να επηρεάσει σημαντικά τη συμπεριφορά του τελικού global model. Στην παρούσα εργασία εξετάστηκαν δύο βασικά σενάρια κατανομής τα οποία είναι το IID και το Non-IID. Τα δύο αυτά σενάρια χρησιμοποιήθηκαν ώστε να αξιολογηθεί η συμπεριφορά του Federated Learning τόσο σε ένα σχετικά ομοιόμορφο περιβάλλον όσο και σε ένα πιο ρεαλιστικό και ανομοιογενές περιβάλλον. Αυτό έχει ως αποτέλεσμα, να εξετάζεται αν το FedAvg μπορεί να εκπαιδεύσει ένα μοντέλο σε κατανεμημένα δεδομένα, αλλά και αν η απόδοσή του επηρεάζεται από το πόσο παρόμοια ή διαφορετικά είναι τα δεδομένα των clients.

Στο πρώτο σενάριο, δηλαδή στο IID σενάριο, το training set χωρίστηκε σε τρεις clients με τέτοιο τρόπο ώστε κάθε client να διατηρεί περίπου την ίδια αναλογία κλάσεων με το αρχικό training set. Η κατανομή

αυτή έγινε με τέτοιο τρόπο, ώστε κάθε client να περιέχει τόσο δείγματα της πλειοψηφικής κλάσης όσο και δείγματα της μειοψηφικής κλάσης σε παρόμοιο ποσοστό. Εφόσον το αρχικό training set ήταν έντονα ανισόρροπο, με πολύ περισσότερα δείγματα χωρίς εγκεφαλικό επεισόδιο σε σχέση με τα δείγματα με εγκεφαλικό επεισόδιο, η ίδια ανισορροπία διατηρήθηκε περίπου και σε κάθε client. Συγκεκριμένα, το training set περιείχε 4.088 δείγματα, από τα οποία 3.889 ανήκαν στην κλάση no stroke και 199 στην κλάση stroke. Στο IID σενάριο, τα δεδομένα μοιράστηκαν σε τρεις clients περίπου ίσου μεγέθους. Κάθε client είχε περίπου 1.360 δείγματα και διατηρούσε ποσοστό θετικής κλάσης κοντά στο 5%. Αυτό σημαίνει ότι οι clients είχαν παρόμοια στατιστική συμπεριφορά με το συνολικό training set. Η συγκεκριμένη κατανομή χρησιμοποιήθηκε ως πιο ελεγχόμενο σενάριο, στο οποίο οι διαφορές μεταξύ των clients είναι περιορισμένες. Η IID κατανομή είναι χρήσιμη γιατί επιτρέπει να ελεγχθεί αν η federated διαδικασία μπορεί να προσεγγίσει τη συμπεριφορά της centralized εκπαίδευσης όταν οι clients έχουν παρόμοια δεδομένα. Σε αυτό το σενάριο, κάθε τοπικό μοντέλο εκπαιδεύεται σε δείγματα που ακολουθούν περίπου την ίδια κατανομή με τα υπόλοιπα. Γι' αυτό, τα βάρη που επιστρέφουν οι clients στον server αναμένεται να είναι σχετικά συμβατά μεταξύ τους. Για αυτόν τον λόγο, σε IID περιβάλλοντα το FedAvg συνήθως έχει πιο σταθερή συμπεριφορά, καθώς ο server συνδυάζει τοπικές ενημερώσεις που προέρχονται από παρόμοιες κατανομές δεδομένων. Ωστόσο, στην πράξη τα δεδομένα που βρίσκονται σε διαφορετικούς φορείς υγείας σπάνια είναι πλήρως IID. Διαφορετικά νοσοκομεία μπορεί να εξυπηρετούν διαφορετικούς πληθυσμούς, ηλικιακές ομάδες ή ομάδες ασθενών με διαφορετικό επίπεδο κινδύνου. Για παράδειγμα, ένα νοσοκομείο μπορεί να έχει περισσότερα περιστατικά υψηλού κινδύνου, ενώ ένα άλλο μπορεί να διαθέτει κυρίως γενικό πληθυσμό με λιγότερα περιστατικά εγκεφαλικού. Επομένως, ένα πιο ρεαλιστικό federated σενάριο πρέπει να λαμβάνει υπόψη την ανομοιογένεια των δεδομένων. Για αυτόν τον λόγο, δημιουργήθηκε και δεύτερο σενάριο, το Non-IID.

Στο Non-IID σενάριο, οι τρεις clients δεν είχαν την ίδια αναλογία θετικών περιστατικών. Αντίθετα, τα δείγματα της θετικής κλάσης κατανεμήθηκαν άνισα μεταξύ των clients. Πιο συγκεκριμένα, ο πρώτος client έλαβε μικρό ποσοστό των περιστατικών stroke, ο δεύτερος client έλαβε μεσαίο ποσοστό, ενώ ο τρίτος client έλαβε το μεγαλύτερο ποσοστό της θετικής κλάσης. Έτσι, προσομοιώθηκε μια κατάσταση όπου οι διαφορετικοί φορείς δεν έχουν την ίδια συχνότητα εμφάνισης της ασθένειας στα τοπικά τους δεδομένα.

Στην υλοποίηση, η κατανομή των θετικών δειγμάτων στο Non-IID σενάριο έγινε περίπου με αναλογία 10%, 30% και 60% των συνολικών stroke εγγραφών στους τρεις clients. Με δεδομένο ότι το training set περιείχε 199 θετικά δείγματα, αυτό οδήγησε περίπου σε 20 θετικά δείγματα στον πρώτο client, 60 στον δεύτερο και 119 στον τρίτο. Τα δείγματα της αρνητικής κλάσης κατανεμήθηκαν έτσι ώστε οι clients να έχουν σχετικά παρόμοιο συνολικό μέγεθος. Έτσι, η βασική διαφορά μεταξύ των clients δεν ήταν τόσο το συνολικό πλήθος δειγμάτων, αλλά η αναλογία της θετικής κλάσης. Η Non-IID κατανομή θεωρείται πιο απαιτητική για το Federated Learning, επειδή κάθε client εκπαιδεύει τοπικά το μοντέλο σε διαφορετική κατανομή δεδομένων. Αυτό μπορεί να οδηγήσει σε τοπικά μοντέλα που κινούνται προς διαφορετικές κατευθύνσεις κατά την εκπαίδευση. Για παράδειγμα, ένας client με πολύ λίγα θετικά δείγματα μπορεί να μάθει πιο έντονα τη συμπεριφορά της πλειοψηφικής κλάσης, ενώ ένας client με περισσότερα θετικά δείγματα μπορεί να δώσει μεγαλύτερη σημασία στη μειοψηφική κλάση [19]. Όταν ο server συνδυάζει αυτά τα τοπικά μοντέλα, η διαδικασία FedAvg μπορεί να γίνει λιγότερο σταθερή σε σχέση με το IID σενάριο.

Όπως φαίνεται στον Πίνακα 5.X, στο Non-IID σενάριο οι τρεις clients δεν διαθέτουν την ίδια αναλογία περιστατικών stroke και no-stroke. Ο πρώτος client περιλαμβάνει πολύ μικρό αριθμό θετικών περιστατικών, ενώ ο τρίτος client περιλαμβάνει σημαντικά μεγαλύτερο αριθμό. Με αυτόν τον τρόπο

προσομοιώνεται ένα πιο ρεαλιστικό περιβάλλον Federated Learning, όπου οι τοπικοί κόμβοι δεν έχουν απαραίτητα την ίδια κατανομή δεδομένων.

Πίνακας 5.1: Κατανομή δειγμάτων ανά client στο Non-IID σενάριο.

Client	No Stroke δείγματα	Stroke δείγματα	Σύνολο δειγμάτων	Ποσοστό Stroke δειγμάτων
1	1343	20	1363	1.47%
2	1303	60	1363	4.4%
3	1243	119	1362	8.74%

Η σύγκριση ανάμεσα σε IID και Non-IID κατανομή ήταν σημαντική για την παρούσα εργασία, καθώς επέτρεψε να εξεταστεί αν η federated προσέγγιση παραμένει αποτελεσματική όταν τα δεδομένα των clients δεν είναι ομοιόμορφα κατανομημένα. Στο πρόβλημα της πρόβλεψης εγκεφαλικού επεισοδίου, η ανομοιογένεια αυτή έχει ιδιαίτερη σημασία, επειδή η θετική κλάση είναι ήδη πολύ περιορισμένη. Με βάση τα παραπάνω, ακόμα και μικρές αλλαγές στην κατανομή των stroke samples ανά client μπορούν να επηρεάσουν την τοπική εκπαίδευση και, τελικά, την απόδοση του καθολικού μοντέλου. Σε όλα τα σενάρια, το test set παρέμεινε κοινό και αμετάβλητο. Αυτό σημαίνει ότι τόσο τα IID όσο και τα Non-IID federated πειράματα αξιολογήθηκαν στο ίδιο πραγματικό, ανισόρροπο test set. Η συγκεκριμένη πρακτική ήταν απαραίτητη για να υπάρχει δίκαιη σύγκριση μεταξύ των πειραμάτων. Αν το test set άλλαζε ή εξισορροπούταν, τα αποτελέσματα δεν θα ήταν άμεσα συγκρίσιμα με τα centralized πειράματα του προηγούμενου κεφαλαίου. Επομένως, όλες οι διαφοροποιήσεις έγιναν μόνο στο training set και συγκεκριμένα στον τρόπο που αυτό μοιράστηκε στους clients.

Τα αποτελέσματα συνοψίζονται στο ότι, τα δύο σενάρια κατανομής σχεδιάστηκαν ώστε να καλύψουν δύο διαφορετικές πλευρές της federated εκπαίδευσης. Το IID σενάριο χρησιμοποιήθηκε ως ελεγχόμενη περίπτωση, όπου οι clients έχουν παρόμοια δεδομένα και το FedAvg αναμένεται να λειτουργεί πιο ομαλά. Το Non-IID σενάριο χρησιμοποιήθηκε ως πιο ρεαλιστική και απαιτητική περίπτωση, όπου οι clients έχουν διαφορετική αναλογία θετικών περιστατικών. Με αυτόν τον τρόπο, η αξιολόγηση του Federated Learning δεν περιορίστηκε μόνο σε ιδανικές συνθήκες, αλλά επεκτάθηκε και σε συνθήκες ανομοιογένειας, οι οποίες είναι πιο κοντά σε ένα πραγματικό περιβάλλον υγειονομικών δεδομένων.

5.3 Υλοποίηση FedAvg

Για την υλοποίηση του Federated Learning χρησιμοποιήθηκε ο αλγόριθμος Federated Averaging, γνωστός και ως FedAvg. Ο FedAvg αποτελεί μία από τις βασικές και πιο διαδεδομένες προσεγγίσεις στο Federated Learning, καθώς επιτρέπει την εκπαίδευση ενός global model μέσω της συνεργασίας πολλών clients, χωρίς να απαιτείται η μεταφορά των τοπικών δεδομένων τους σε έναν κεντρικό server. Στην παρούσα εργασία, ο FedAvg χρησιμοποιήθηκε ώστε να συνδυαστούν τα τοπικά μοντέλα των τριών clients και να δημιουργηθεί ένα κοινό μοντέλο για την πρόβλεψη εγκεφαλικού επεισοδίου.

Η βασική λογική της υλοποίησης ήταν ότι κάθε client εκπαιδεύει ένα αντίγραφο του ίδιου μοντέλου στα δικά του τοπικά δεδομένα. Στη συνέχεια, αντί να στείλει τα δεδομένα του στον server, στέλνει μόνο τις ενημερωμένες παραμέτρους του μοντέλου. Ο server λαμβάνει τις παραμέτρους από όλους τους clients και τις συνδυάζει μέσω σταθμισμένου μέσου όρου. Το αποτέλεσμα αυτής της διαδικασίας είναι ένα νέο ενιαίο μοντέλο, το οποίο στη συνέχεια αποστέλλεται ξανά στους clients για τον επόμενο γύρο εκπαίδευσης. Η διαδικασία αυτή επαναλαμβάνεται για συγκεκριμένο αριθμό γύρων επικοινωνίας [20].

Στην παρούσα υλοποίηση, η διαδικασία του Federated Learning υλοποιήθηκε με χρήση της βιβλιοθήκης PyTorch . Η βιβλιοθήκη αυτή χρησιμοποιήθηκε για τον ορισμό του μοντέλου Logistic Regression, τη μετατροπή των δεδομένων σε tensors, την τοπική εκπαίδευση κάθε client και τη διαχείριση των παραμέτρων του μοντέλου που συνδυάζονται μέσω FedAvg. Το μοντέλο αποτελούνταν από ένα γραμμικό επίπεδο, το οποίο δεχόταν ως είσοδο τα 16 χαρακτηριστικά του dataset και παρήγαγε μία τιμή εξόδου για τη δυαδική ταξινόμηση. Η έξοδος του μοντέλου ήταν ένα logit, το οποίο στη συνέχεια μετατρεπόταν σε πιθανότητα μέσω της συνάρτησης sigmoid κατά την αξιολόγηση. Για την εκπαίδευση χρησιμοποιήθηκε η συνάρτηση απώλειας BCEWithLogitsLoss, η οποία είναι κατάλληλη για προβλήματα δυαδικής ταξινόμησης και συνδυάζει εσωτερικά τη sigmoid με το binary cross-entropy loss.

Η επιλογή της Logistic Regression έγινε για δύο βασικούς λόγους. Πρώτον, αποτελεί ένα απλό και ερμηνεύσιμο μοντέλο, το οποίο είχε ήδη χρησιμοποιηθεί στα centralized πειράματα του προηγούμενου κεφαλαίου. Έτσι, ήταν δυνατή η πιο άμεση σύγκριση ανάμεσα σε centralized και federated εκπαίδευση. Δεύτερον, η Logistic Regression είναι κατάλληλη για εφαρμογή του FedAvg, καθώς οι παράμετροί της αποτελούνται από βάρη και bias, τα οποία μπορούν να συνδυαστούν αριθμητικά μέσω του μέσου όρου. Αντίθετα, μοντέλα όπως το Random Forest και το XGBoost δεν είναι τόσο άμεσα συμβατά με την κλασική μορφή του FedAvg, καθώς δεν βασίζονται στην ίδια λογική ενημέρωσης συνεχών παραμέτρων.

Η διαδικασία εκπαίδευσης οργανώθηκε σε 20 rounds. Σε κάθε round, ο server διατηρούσε το τρέχον καθολικό μοντέλο και το έστειλε στους τρεις clients. Κάθε client δημιουργούσε τοπικό αντίγραφο του global model και το εκπαιδευε τοπικά στα δικά του δεδομένα για 5 εποχές. Μετά την ολοκλήρωση της τοπικής εκπαίδευσης, κάθε client επέστρεφε στον server μόνο τα ενημερωμένα βάρη του μοντέλου. Ο server, με τη σειρά του, υπολόγιζε τον σταθμισμένο μέσο όρο των βαρών αυτών και δημιουργούσε το νέο global model.

Είναι σημαντικό να τονιστεί ότι τα rounds και οι εποχές δεν είναι το ίδιο πράγμα. Ένα round επικοινωνίας αντιστοιχεί σε έναν πλήρη κύκλο επικοινωνίας μεταξύ server και clients. Δηλαδή, σε κάθε round το μοντέλο αποστέλλεται στους clients, εκπαιδεύεται τοπικά και στη συνέχεια επιστρέφονται τα ενημερωμένα βάρη στον server. Αντίθετα, οι εποχές αφορούν το πόσες φορές κάθε client περνάει από τα δικά του τοπικά δεδομένα κατά τη διάρκεια ενός round. Γι' αυτό, στην παρούσα εργασία κάθε client εκπαιδευόταν για 5 εποχές μέσα σε κάθε round, και η συνολική federated διαδικασία επαναλαμβανόταν για 20 rounds.

Ο συνδυασμός των τοπικών μοντέλων πραγματοποιήθηκε με σταθμισμένο μέσο όρο. Αυτό σημαίνει ότι το μοντέλο κάθε client δεν συμμετείχε απαραίτητα ισότιμα στον υπολογισμό του καθολικού μοντέλου, αλλά η συμβολή του εξαρτιόταν από το πλήθος των δειγμάτων που είχε ο συγκεκριμένος client. Αν ένας client είχε περισσότερα δεδομένα, τότε τα βάρη του είχαν μεγαλύτερη επίδραση στο τελικό μοντέλο. Αντίστοιχα, αν ένας client είχε λιγότερα δεδομένα, η επίδρασή του ήταν μικρότερη. Η λογική αυτή χρησιμοποιήθηκε ώστε το μοντέλο να αντιπροσωπεύει καλύτερα το συνολικό σύνολο εκπαίδευσης.

Ο υπολογισμός του FedAvg περιγράφεται γενικά ως εξής. Αν υπάρχουν K clients, όπου κάθε client διαθέτει n_i δείγματα και έχει εκπαιδεύσει τοπικά βάρη w_i , τότε τα νέα βάρη του συνολικού μοντέλου στον server υπολογίζονται από τη σχέση [20]:

$$W_{server} = \frac{\sum_{i=1}^K n_i * w_i}{\sum_{i=1}^K n_i} \quad (5,1)$$

Σε αυτή τη ΔΕ, η παραπάνω σχέση εφαρμόστηκε για $K = 3$, καθώς η προσομοίωση του Federated Learning πραγματοποιήθηκε με τρεις clients. Η ίδια διαδικασία εφαρμόστηκε για όλες τις παραμέτρους του μοντέλου, δηλαδή τόσο για τα βάρη όσο και για το bias. Με αυτόν τον τρόπο, ο server δεν επιλέγει απλώς ένα από τα τοπικά μοντέλα, αλλά δημιουργεί ένα νέο συνολικό μοντέλο που συνδυάζει τη γνώση που προκύπτει από όλους τους clients. Η στάθμιση με βάση το n , δηλαδή το πλήθος των τοπικών δειγμάτων κάθε client, διασφαλίζει ότι το συνολικό μοντέλο δεν επηρεάζεται δυσανάλογα από clients που διαθέτουν πολύ μικρό αριθμό δεδομένων. Αντίθετα, οι clients με περισσότερα δεδομένα έχουν μεγαλύτερη επίδραση στο τελικό μοντέλο, καθώς τα τοπικά τους βάρη προκύπτουν από μεγαλύτερο μέρος του συνολικού συνόλου εκπαίδευσης.

Στην υλοποίηση, κάθε client εκπαιδευόταν ανεξάρτητα πάνω στα δικά του δεδομένα. Το μοντέλο του Client 1 δεν συνέχιζε την εκπαίδευση πάνω στα δεδομένα του Client 2 και στη συνέχεια του Client 3. Αντίθετα, σε κάθε round οι τρεις clients ξεκινούσαν από το ίδιο καθολικό μοντέλο και παρήγαγαν τρία διαφορετικά τοπικά μοντέλα. Αυτή η λεπτομέρεια είναι σημαντική, γιατί διαχωρίζει τη federated εκπαίδευση από μια απλή σειριακή εκπαίδευση πάνω σε διαφορετικά τμήματα του dataset. Στο Federated Learning, οι clients λειτουργούν ως ανεξάρτητες οντότητες και ο server είναι υπεύθυνος μόνο για τη σύνθεση των τοπικών ενημερώσεων. Μετά από κάθε γύρο επικοινωνίας, το global model αξιολογούνταν στο κοινό test set. Η αξιολόγηση αυτή γινόταν κυρίως για να παρακολουθηθεί η πορεία της εκπαίδευσης και η σταθερότητα του μοντέλου από round σε round. Στο τέλος των 20 rounds κρατήθηκαν τα τελικά αποτελέσματα του μοντέλου. Οι μετρικές που υπολογίστηκαν ήταν οι ίδιες με αυτές των centralized πειραμάτων.

Η χρήση κοινού test set ήταν απαραίτητη για τη συγκρισιμότητα των αποτελεσμάτων. Παρόλο που το training set χωρίστηκε σε clients, το test set παρέμεινε ενιαίο και αμετάβλητο σε όλα τα federated πειράματα. Έτσι, η αξιολόγηση του global model έγινε με τον ίδιο τρόπο που είχε γίνει και στα centralized πειράματα. Αυτό επέτρεψε να εξεταστεί αν η federated εκπαίδευση μπορεί να προσεγγίσει την απόδοση της centralized εκπαίδευσης, χωρίς να αλλάξει το πλαίσιο αξιολόγησης.

Συνολικά, η υλοποίηση του FedAvg στην παρούσα εργασία ακολούθησε μια απλή αλλά σαφή δομή. Το αρχικό training set χωρίστηκε σε τρεις clients, κάθε client εκπαιδευε τοπικά ένα αντίγραφο του κοινού μοντέλου, ο server συνδύαζε τα βάρη μέσω σταθμισμένου μέσου όρου και το τελικό μοντέλο αξιολογούνταν στο κοινό test set. Η υλοποίηση αυτή επέτρεψε να εξεταστεί η συμπεριφορά του Federated Learning τόσο σε IID όσο και σε Non-IID κατανομή δεδομένων, καθώς και να αξιολογηθεί η επίδραση τεχνικών αντιμετώπισης της ανισορροπίας μέσα σε federated περιβάλλον.

5.4 Federated Learning με Τοπική Εξισορρόπηση

Ένα από τα βασικά προβλήματα που έπρεπε να αντιμετωπιστούν και στο federated περιβάλλον ήταν η έντονη ανισορροπία των κλάσεων. Όπως είχε φανεί ήδη από την ανάλυση του dataset, τα δείγματα της κλάσης stroke είναι πολύ λιγότερα σε σχέση με τα δείγματα της κλάσης no stroke. Το πρόβλημα αυτό δεν εξαφανίζεται όταν τα δεδομένα χωρίζονται σε clients. Αντίθετα, σε ένα federated περιβάλλον μπορεί να γίνει ακόμα πιο έντονο, επειδή κάθε client διαθέτει μόνο ένα μέρος του συνολικού training set. Επομένως, ορισμένοι clients μπορεί να έχουν πολύ μικρό αριθμό θετικών περιστατικών, γεγονός που δυσκολεύει την τοπική εκπαίδευση του μοντέλου.

Συνεπώς, εκτός από το απλό federated baseline, εξετάστηκαν και τεχνικές αντιμετώπισης της ανισορροπίας μέσα στο federated περιβάλλον. Συγκεκριμένα, για κάθε σενάριο κατανομής, δηλαδή τόσο για το IID όσο και για το Non-IID, πραγματοποιήθηκαν τρία πειράματα. Το πρώτο πείραμα ήταν το απλό FedAvg χωρίς καμία παρέμβαση στην ανισορροπία. Το δεύτερο πείραμα χρησιμοποίησε class

weights κατά την τοπική εκπαίδευση κάθε client. Το τρίτο πείραμα χρησιμοποίησε Local SMOTE, δηλαδή εφαρμογή SMOTE ξεχωριστά στα τοπικά δεδομένα κάθε client.

Το πρώτο πείραμα, δηλαδή το federated baseline, χρησιμοποιήθηκε ως σημείο αναφοράς. Σε αυτό το σενάριο, οι clients εκπαιδεύτηκαν στα αρχικά τοπικά τους δεδομένα χωρίς καμία εξισορρόπηση και χωρίς τροποποίηση της συνάρτησης απώλειας. Η προσέγγιση αυτή ήταν απαραίτητη, ώστε να φανεί η αρχική συμπεριφορά του FedAvg πάνω σε έντονα ανισορροπα δεδομένα. Με τον τρόπο αυτό, μπορούσε να αξιολογηθεί αν το federated μοντέλο, χωρίς καμία βοήθεια, ήταν ικανό να εντοπίσει τη μειοψηφική κλάση.

Στην πράξη, το federated baseline αντιστοιχεί στην πιο απλή μορφή της εκπαίδευσης. Ο server αποστέλλει το καθολικό μοντέλο στους clients, κάθε client εκπαιδεύει τοπικά το μοντέλο στα δικά του δεδομένα και στη συνέχεια ο server συνδυάζει τα βάρη μέσω FedAvg. Σε αυτό το σενάριο, όμως, η συνάρτηση απώλειας αντιμετωπίζει και τις δύο κλάσεις με τον ίδιο τρόπο. Επειδή η αρνητική κλάση είναι πολύ μεγαλύτερη, το μοντέλο έχει την τάση να μαθαίνει κυρίως τη συμπεριφορά της πλειοψηφικής κλάσης. Αυτό είναι ιδιαίτερα προβληματικό στο συγκεκριμένο πρόβλημα, καθώς ένα μοντέλο που προβλέπει σχεδόν πάντα no stroke μπορεί να εμφανίζει υψηλή Accuracy, αλλά να είναι μη παραγωγικό από ιατρική άποψη.

Το δεύτερο federated πείραμα χρησιμοποίησε class weights. Η λογική των class weights είναι ότι η μειοψηφική κλάση λαμβάνει μεγαλύτερη βαρύτητα στη συνάρτηση απώλειας. Έτσι, τα λάθη στα δείγματα της κλάσης stroke τιμωρούνται περισσότερο σε σχέση με τα λάθη στην πλειοψηφική κλάση. Στην παρούσα υλοποίηση, το βάρος της θετικής κλάσης υπολογίστηκε τοπικά για κάθε client με βάση τη δική του κατανομή δεδομένων. Συγκεκριμένα, χρησιμοποιήθηκε η αναλογία των αρνητικών προς τα θετικά δείγματα του κάθε client.

Η τοπική χρήση class weights είναι σημαντική, επειδή στο Federated Learning κάθε client μπορεί να έχει διαφορετική κατανομή κλάσεων. Στο IID σενάριο, οι clients έχουν περίπου παρόμοια αναλογία θετικών και αρνητικών δειγμάτων, επομένως τα class weights είναι σχετικά παρόμοια μεταξύ τους. Αντίθετα, στο Non-IID σενάριο, οι clients έχουν διαφορετικό αριθμό θετικών περιστατικών. Αυτό σημαίνει ότι το βάρος της θετικής κλάσης μπορεί να διαφέρει από client σε client. Ένας client με πολύ λίγα stroke samples θα δώσει πολύ μεγαλύτερο βάρος στη θετική κλάση, ενώ ένας client με περισσότερα stroke samples θα έχει μικρότερο class weight.

Η προσέγγιση αυτή επιλέχθηκε γιατί προσαρμόζει την τοπική εκπαίδευση στις ανάγκες κάθε client. Αντί να χρησιμοποιηθεί ένα ενιαίο weight για όλους τους clients, κάθε client υπολόγισε το δικό του βάρος με βάση τα δικά του δεδομένα. Μέσω αυτής της διαδικασίας, η τοπική συνάρτηση απώλειας αντανακλούσε καλύτερα την ανισορροπία που υπήρχε σε κάθε επιμέρους σύνολο δεδομένων. Ωστόσο, η προσέγγιση αυτή μπορεί να αυξήσει την ευαισθησία του μοντέλου στη θετική κλάση και να οδηγήσει σε περισσότερα false positives, ειδικά σε clients με πολύ λίγα θετικά δείγματα.

Το τρίτο federated πείραμα χρησιμοποίησε Local SMOTE. Σε αντίθεση με το centralized SMOTE, όπου η υπερδειγματοληψία εφαρμόζεται στο ενιαίο training set, στο federated περιβάλλον το SMOTE εφαρμόστηκε ξεχωριστά σε κάθε client. Δηλαδή, κάθε client δημιούργησε συνθετικά δείγματα της μειοψηφικής κλάσης μόνο με βάση τα δικά του τοπικά δεδομένα. Τα δεδομένα των άλλων clients δεν χρησιμοποιήθηκαν στη διαδικασία αυτή. Αυτό έχει ως αποτέλεσμα να διατηρείται η βασική λογική του Federated Learning, σύμφωνα με την οποία τα δεδομένα δεν μεταφέρονται σε κεντρικό σημείο.

Η εφαρμογή του Local SMOTE είχε ως στόχο να βελτιώσει την τοπική εκπαίδευση κάθε client. Όταν ένας client έχει πολύ λίγα θετικά δείγματα, το μοντέλο που εκπαιδεύεται τοπικά σε αυτόν τον client

δυσκολεύεται να μάθει τα χαρακτηριστικά της θετικής κλάσης. Με τη δημιουργία επιπλέον συνθετικών θετικών δειγμάτων, ο client αποκτά ένα πιο ισορροπημένο τοπικό training set. Έτσι, το τοπικό μοντέλο έχει περισσότερες ευκαιρίες να μάθει πρότυπα που σχετίζονται με την εμφάνιση εγκεφαλικού επεισοδίου.

Στην υλοποίηση χρησιμοποιήθηκε $\text{sampling_strategy} = 0.7$. Αυτό σημαίνει ότι μετά την εφαρμογή του SMOTE, ο αριθμός των δειγμάτων της μειοψηφικής κλάσης αντιστοιχούσε περίπου στο 70% του αριθμού των δειγμάτων της πλειοψηφικής κλάσης. Η επιλογή αυτή έγινε ώστε να ενισχυθεί σημαντικά η θετική κλάση, χωρίς όμως να δημιουργηθεί πλήρης ισορροπία. Η πλήρης εξισορρόπηση μπορεί σε ορισμένες περιπτώσεις να οδηγήσει σε υπερβολική επίδραση των συνθετικών δειγμάτων, ειδικά όταν ο αρχικός αριθμός πραγματικών θετικών δειγμάτων είναι μικρός. Συνεπώς, η επιλογή αναλογίας 0.7 θεωρήθηκε πιο ελεγχόμενη και συνεπής με τα προηγούμενα centralized πειράματα.

Κατά την εφαρμογή του Local SMOTE δόθηκε ιδιαίτερη προσοχή στον αριθμό των θετικών δειγμάτων κάθε client. Η μέθοδος SMOTE απαιτεί επαρκή αριθμό δειγμάτων της μειοψηφικής κλάσης, επειδή δημιουργεί νέα συνθετικά δείγματα με βάση κοντινούς γείτονες. Γι' αυτό, στην υλοποίηση ελέγχθηκε ο αριθμός των θετικών δειγμάτων κάθε client και το πλήθος των γειτόνων προσαρμόστηκε ανάλογα. Συγκεκριμένα, το $k_neighbors$ ορίστηκε ώστε να μην υπερβαίνει τον διαθέσιμο αριθμό θετικών δειγμάτων. Αυτό ήταν απαραίτητο για να αποφευχθούν σφάλματα κατά την εφαρμογή του SMOTE, ειδικά στο Non-IID σενάριο όπου ο πρώτος client είχε μικρότερο αριθμό stroke samples.

Ένα σημαντικό στοιχείο της μεθοδολογίας είναι ότι το SMOTE εφαρμόστηκε μόνο στα training δεδομένα των clients. Το test set δεν τροποποιήθηκε σε κανένα federated πείραμα. Αυτό σημαίνει ότι η αξιολόγηση έγινε πάντοτε στο αρχικό, πραγματικό και ανισόρροπο test set. Η επιλογή αυτή είναι κρίσιμη, γιατί το ζητούμενο ήταν να εξεταστεί αν το μοντέλο μπορεί να αναγνωρίσει τη θετική κλάση σε πραγματικές συνθήκες ανισορροπίας. Αν το test set εξισορροπούνταν, τα αποτελέσματα δεν θα αντανάκλαζαν την πραγματική δυσκολία του προβλήματος και δεν θα ήταν άμεσα συγκρίσιμα με τα centralized πειράματα.

Στα federated πειράματα δεν εφαρμόστηκε CTGAN. Αυτή η προσέγγιση έγινε για λόγους μεθοδολογικής καθαρότητας και περιορισμού της πολυπλοκότητας. Το CTGAN είχε ήδη αξιολογηθεί στο centralized μέρος της εργασίας, όπου χρησιμοποιήθηκε για τη δημιουργία συνθετικών δειγμάτων της μειοψηφικής κλάσης. Η μεταφορά του CTGAN στο federated περιβάλλον θα απαιτούσε την εκπαίδευση ξεχωριστού generative model σε κάθε client. Αυτό θα αύξανε σημαντικά την πολυπλοκότητα της υλοποίησης και θα δημιουργούσε πρόσθετα ζητήματα, ειδικά σε clients με πολύ λίγα θετικά δείγματα. Σε τέτοιες περιπτώσεις, η τοπική εκπαίδευση ενός CTGAN θα μπορούσε να είναι ασταθής και τα συνθετικά δεδομένα να μην είναι αξιόπιστα.

Επιπλέον, δεν εφαρμόστηκαν στο federated μέρος οι παραλλαγές Borderline-SMOTE και SMOTE-Tomek. Οι τεχνικές αυτές είχαν ήδη εξεταστεί στο centralized περιβάλλον και προσέφεραν χρήσιμη σύγκριση με το απλό SMOTE. Στο federated κομμάτι, όμως, ο στόχος ήταν διαφορετικός. Το βασικό ερώτημα δεν ήταν να συγκριθούν όλες οι πιθανές παραλλαγές υπερδειγματοληψίας, αλλά να εξεταστεί αν μια απλή και ελεγχόμενη μορφή τοπικής εξισορρόπησης μπορεί να λειτουργήσει σε κατανομημένο περιβάλλον. Για τον λόγο αυτό, επιλέχθηκε το Local SMOTE ως αντιπροσωπευτική και πιο καθαρή προσέγγιση.

Η απόφαση αυτή βοήθησε ώστε το federated κομμάτι της εργασίας να παραμείνει εστιασμένο. Αν είχαν προστεθεί CTGAN, Borderline-SMOTE και SMOTE-Tomek σε IID και Non-IID σενάρια, ο αριθμός των πειραμάτων θα αυξανόταν σημαντικά και η ανάλυση θα γινόταν πιο δύσκολη. Αντίθετα, με την

Κεφάλαιο 5ο:

επιλογή τριών βασικών πειραμάτων ανά κατανομή, δηλαδή baseline, class weights και Local SMOTE, η σύγκριση παρέμεινε καθαρή και άμεσα ερμηνεύσιμη.

Γενικά, η federated μεθοδολογία με τοπική εξισορρόπηση σχεδιάστηκε ώστε να απαντήσει σε δύο βασικά ερωτήματα. Πρώτον, αν το FedAvg μπορεί να εκπαιδεύσει ένα μοντέλο πρόβλεψης εγκεφαλικού σε κατανεμημένα και ανισόρροπα δεδομένα. Δεύτερον, αν τεχνικές όπως τα class weights και το Local SMOTE μπορούν να βελτιώσουν την ανίχνευση της μειοψηφικής κλάσης σε federated περιβάλλον. Με αυτόν τον τρόπο, το federated μέρος της εργασίας δεν περιορίστηκε μόνο στην απλή εφαρμογή του FedAvg, αλλά επεκτάθηκε στην αξιολόγηση πρακτικών τεχνικών που είναι σημαντικές για προβλήματα υγείας με έντονη ανισορροπία κλάσεων.

Πίνακας 5.2: Συγκεντρωτική παρουσίαση των σεναρίων Federated Learning.

ID	Σενάριο	Κατανομή δεδομένων	Τεχνική εξισορρόπησης	Στόχος
1	IID Baseline	IID	Καμία	Αξιολόγηση της βασικής συμπεριφοράς του FedAvg όταν οι clients έχουν παρόμοια κατανομή δεδομένων.
2	IID Class Weight	IID	Class weights	Εξέταση της επίδρασης της στάθμισης των κλάσεων στην τοπική εκπαίδευση των clients.
3	IID Local SMOTE	IID	Local SMOTE	Αξιολόγηση της τοπικής υπερδειγματοληψίας της μειοψηφικής κλάσης σε συνθήκες IID.
4	Non-IID Baseline	Non-IID	Καμία	Αξιολόγηση της βασικής συμπεριφοράς του FedAvg όταν οι clients έχουν διαφορετική κατανομή.
5	Non-IID Class Weight	Non-IID	Class weights	Διερεύνηση της επίδρασης των class weights σε πιο ρεαλιστικό και ανομοιογενές περιβάλλον clients.
6	Non-IID Local SMOTE	Non-IID	Local SMOTE	Εξέταση του αν η τοπική υπερδειγματοληψία μπορεί να βελτιώσει τον εντοπισμό της θετικής κλάσης στο Non-IID σενάριο.

Όπως φαίνεται στον Πίνακα 5.2, η πειραματική διαδικασία του Federated Learning οργανώθηκε σε έξι βασικά σενάρια. Τα τρία πρώτα αφορούν την περίπτωση IID, όπου οι clients έχουν παρόμοια κατανομή δεδομένων, ενώ τα τρία επόμενα αφορούν την περίπτωση Non-IID, όπου η αναλογία των stroke και no-stroke δειγμάτων διαφέρει μεταξύ των clients. Για κάθε τύπο κατανομής εξετάστηκε αρχικά το baseline

σενάριο χωρίς εξισορρόπηση και στη συνέχεια εφαρμόστηκαν δύο τεχνικές αντιμετώπισης της ανισορροπίας: class weights και Local SMOTE. Με αυτόν τον τρόπο, ήταν δυνατή η σύγκριση της συμπεριφοράς του FedAvg τόσο σε ισορροπημένη όσο και σε πιο ανομοιογενή κατανομή δεδομένων.

Κεφάλαιο 6ο: Συγκριτική Αξιολόγηση και Συζήτηση

6.1 Συνολικά Αποτελέσματα

Στο παρόν κεφάλαιο παρουσιάζονται και συζητούνται συγκεντρωτικά τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν στην εργασία. Η αξιολόγηση περιλαμβάνει τόσο τα centralized πειράματα μηχανικής μάθησης όσο και τα federated πειράματα που βασίστηκαν στον αλγόριθμο FedAvg. Στόχος της ενότητας δεν είναι απλώς η παράθεση των αριθμητικών αποτελεσμάτων, αλλά η ερμηνεία της συμπεριφοράς των μοντέλων σε σχέση με το βασικό πρόβλημα της εργασίας, δηλαδή την πρόβλεψη της μειοψηφικής κλάσης stroke σε ένα έντονα ανισόρροπο ιατρικό dataset.

Όπως αναδείχθηκε από την αρχική ανάλυση των δεδομένων, το dataset χαρακτηρίζεται από πολύ μεγάλη ανισορροπία μεταξύ των δύο κλάσεων. Η κλάση no stroke αποτελεί τη συντριπτική πλειοψηφία των δειγμάτων, ενώ η κλάση stroke αντιστοιχεί σε πολύ μικρό ποσοστό του συνόλου. Η κατανομή αυτή επηρεάζει έντονα τη συμπεριφορά των μοντέλων. Με βάση αυτό, η αξιολόγηση των μοντέλων δεν βασίστηκε αποκλειστικά στην μετρική Accuracy. Αντίθετα, δόθηκε ιδιαίτερη έμφαση σε μετρικές όπως το Recall, το F1-score και το PR-AUC. Το Recall είναι ιδιαίτερα σημαντικό στο συγκεκριμένο πρόβλημα, επειδή δείχνει πόσα από τα πραγματικά περιστατικά εγκεφαλικού κατάφερε να εντοπίσει το μοντέλο. Σε ιατρικά προβλήματα πρόβλεψης, η αποτυχία εντοπισμού θετικών περιστατικών είναι κρίσιμη, καθώς μπορεί να σημαίνει ότι άτομα υψηλού κινδύνου δεν αναγνωρίζονται από το σύστημα. Παράλληλα, το F1-score και το PR-AUC χρησιμοποιήθηκαν για να αξιολογηθεί πιο ισορροπημένα η συμπεριφορά των μοντέλων, λαμβάνοντας υπόψη τόσο την ανίχνευση της θετικής κλάσης όσο και τα false positives.

Στα centralized πειράματα εξετάστηκαν τρία βασικά μοντέλα, τα οποία είναι η Logistic Regression, το Random Forest και το XGBoost. Τα μοντέλα αυτά αξιολογήθηκαν αρχικά χωρίς κάποια τεχνική αντιμετώπισης της ανισορροπίας και στη συνέχεια με διαφορετικές παρεμβάσεις, όπως class weights, SMOTE, Borderline-SMOTE, SMOTE-Tomek και CTGAN. Τα αποτελέσματα έδειξαν ότι τα baseline μοντέλα παρουσίαζαν αδυναμία να εντοπίσουν τη θετική κλάση. Η Logistic Regression χωρίς παρέμβαση πέτυχε πολύ υψηλό Accuracy, όμως εντόπισε μόνο 1 από τα 50 πραγματικά stroke cases στο test set, οδηγώντας σε Recall ίσο με 0.02. Αντίστοιχα, το Random Forest baseline απέτυχε πλήρως να εντοπίσει τη θετική κλάση, ενώ και το XGBoost baseline είχε πολύ χαμηλό Recall.

Η συμπεριφορά αυτή ήταν αναμενόμενη λόγω της έντονης ανισορροπίας. Τα μοντέλα εκπαιδεύτηκαν σε ένα training set όπου τα περισσότερα δείγματα ανήκαν στην αρνητική κλάση. Κατά συνέπεια, χωρίς κάποια ειδική παρέμβαση, είχαν την τάση να προβλέπουν κυρίως no stroke. Αυτό οδήγησε σε υψηλές τιμές Accuracy, αλλά σε πολύ χαμηλές τιμές Recall και F1-score. Το αποτέλεσμα αυτό επιβεβαιώνει ότι η Accuracy δεν είναι κατάλληλη ως βασική μετρική αξιολόγησης για το συγκεκριμένο πρόβλημα, καθώς μπορεί να δώσει παραπλανητική εικόνα για την πραγματική χρησιμότητα του μοντέλου.

Η χρήση class weights βελτίωσε σημαντικά τη συμπεριφορά ορισμένων μοντέλων, κυρίως της Logistic Regression και του XGBoost. Στην περίπτωση της Logistic Regression, η εφαρμογή class weights αύξησε το Recall στο 0.80, γεγονός που σημαίνει ότι το μοντέλο κατάφερε να εντοπίσει 40 από τα 50 πραγματικά περιστατικά εγκεφαλικού στο test set. Αντίστοιχα, το XGBoost με scale_pos_weight πέτυχε Recall ίσο με 0.50, εντοπίζοντας 25 από τα 50 θετικά περιστατικά. Ωστόσο, η αύξηση του Recall συνοδεύτηκε από αύξηση των false positives, κάτι που αποτυπώθηκε στις σχετικά χαμηλές τιμές Precision. Αυτό δείχνει ότι τα class weights κάνουν το μοντέλο πιο ευαίσθητο στη μειοψηφική κλάση,

αλλά ταυτόχρονα αυξάνουν την πιθανότητα να ταξινομηθούν λανθασμένα αρκετά αρνητικά δείγματα ως θετικά.

Οι τεχνικές SMOTE παρουσίασαν επίσης σημαντική βελτίωση σε σχέση με τα baseline πειράματα. Ειδικά στη Logistic Regression, το SMOTE 0.7 και το Borderline-SMOTE έδωσαν από τα πιο ισορροπημένα αποτελέσματα. Το SMOTE 0.7 πέτυχε Recall ίσο με 0.76, F1-score περίπου 0.2705 και PR-AUC περίπου 0.2896, ενώ το Borderline-SMOTE πέτυχε Recall 0.78 και F1-score 0.2910. Τα αποτελέσματα αυτά δείχνουν ότι η υπερδειγματοληψία της μειοψηφικής κλάσης βοήθησε το μοντέλο να μάθει καλύτερα τα πρότυπα που σχετίζονται με τα περιστατικά εγκεφαλικού. Σε σχέση με τα class weights, το SMOTE είχε ελαφρώς χαμηλότερο Recall σε ορισμένες περιπτώσεις, αλλά καλύτερο F1-score και καλύτερη ισορροπία μεταξύ Precision και Recall.

Αντίθετα, το CTGAN δεν παρουσίασε την ίδια αποτελεσματικότητα. Παρότι η χρήση συνθετικών δεδομένων μέσω CTGAN βελτίωσε τη συμπεριφορά σε σχέση με το απλό baseline, τα αποτελέσματα ήταν γενικά χαμηλότερα από αυτά των SMOTE-based τεχνικών. Για παράδειγμα, στην περίπτωση της Logistic Regression με CTGAN 0.7, το Recall έφτασε στο 0.36, τιμή σαφώς υψηλότερη από το baseline, αλλά αρκετά χαμηλότερη από το SMOTE και τα class weights. Το αποτέλεσμα αυτό δείχνει ότι, στο συγκεκριμένο dataset και με τις συγκεκριμένες συνθήκες εκπαίδευσης, τα συνθετικά δείγματα που δημιουργήθηκαν από το CTGAN δεν ήταν εξίσου χρήσιμα για την εκπαίδευση του classifier όσο τα δείγματα που δημιουργήθηκαν μέσω SMOTE.

Στο federated μέρος της εργασίας, η αξιολόγηση πραγματοποιήθηκε με Logistic Regression και FedAvg. Τα δεδομένα χωρίστηκαν σε τρεις clients και εξετάστηκαν δύο σενάρια κατανομής, τα IID και Non-IID. Σε κάθε σενάριο πραγματοποιήθηκαν τρία πειράματα που βασίστηκαν στις τεχνικές baseline, class weights και Local SMOTE. Τα αποτελέσματα των federated πειραμάτων ήταν ιδιαίτερα ενδιαφέροντα, καθώς σε αρκετές περιπτώσεις παρουσίασαν παρόμοια συμπεριφορά με τα centralized πειράματα της Logistic Regression.

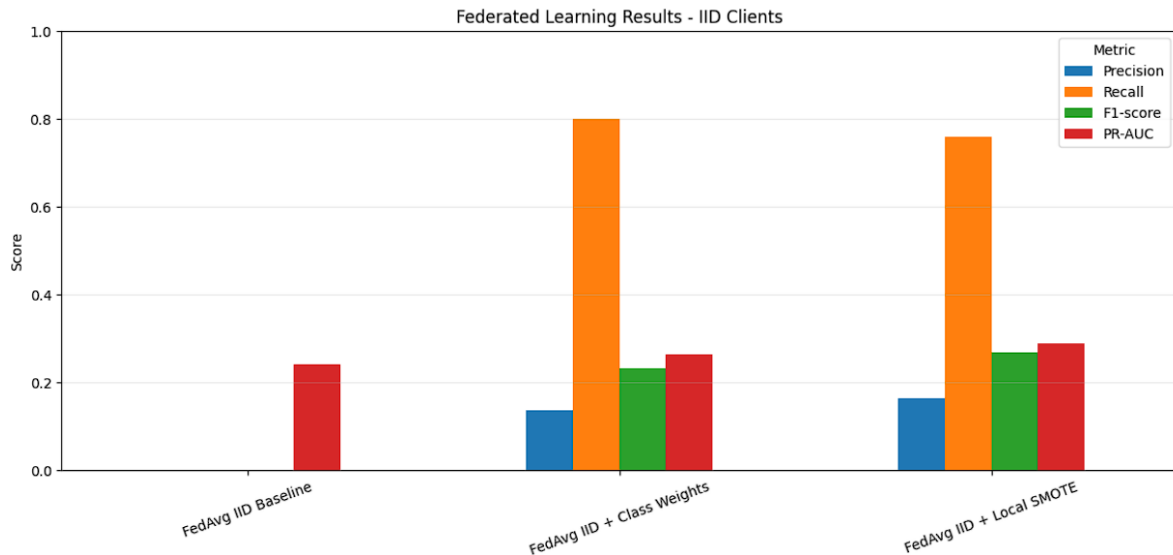
Στα federated baseline πειράματα, τόσο στο IID όσο και στο Non-IID σενάριο, το μοντέλο απέτυχε να εντοπίσει τη θετική κλάση. Και στις δύο περιπτώσεις το confusion matrix έδειξε ότι όλα τα δείγματα του test set ταξινομήθηκαν ως no stroke, με αποτέλεσμα το Recall και το F1-score να είναι μηδενικά. Παράλληλα, το Accuracy παρέμεινε περίπου στο 0.9511, γεγονός που οφείλεται αποκλειστικά στην κυριαρχία της αρνητικής κλάσης στο test set. Το αποτέλεσμα αυτό είναι συνεπές με τα centralized baseline πειράματα και επιβεβαιώνει ότι το πρόβλημα της ανισορροπίας παραμένει εξίσου έντονο και στο federated περιβάλλον.

Η εφαρμογή class weights στο federated περιβάλλον βελτίωσε σημαντικά το Recall. Στο IID σενάριο, το FedAvg με class weights πέτυχε Recall ίσο με 0.80, εντοπίζοντας 40 από τα 50 stroke cases. Στο Non-IID σενάριο, το αντίστοιχο Recall ήταν 0.78, με 39 σωστά εντοπισμένα θετικά περιστατικά. Η μικρή μείωση στο Non-IID σενάριο είναι λογική, καθώς οι clients δεν είχαν την ίδια κατανομή θετικών δειγμάτων. Παρ' όλα αυτά, η συνολική συμπεριφορά παρέμεινε αρκετά κοντά στο IID σενάριο, γεγονός που δείχνει ότι τα class weights μπορούν να λειτουργήσουν αποτελεσματικά και σε federated περιβάλλον.

Το Local SMOTE παρουσίασε επίσης πολύ καλή συμπεριφορά. Στο IID federated σενάριο, το FedAvg με Local SMOTE πέτυχε Recall 0.76, F1-score 0.2695 και PR-AUC 0.2897. Οι τιμές αυτές είναι σχεδόν ίδιες με το centralized Logistic Regression σε συνδιασμό με την τεχνική SMOTE 0.7, το οποίο πέτυχε Recall 0.76, F1-score 0.2705 και PR-AUC 0.2896. Αυτό αποτελεί ένα από τα πιο σημαντικά ευρήματα

της εργασίας, καθώς δείχνει ότι η τοπική εφαρμογή SMOTE σε federated περιβάλλον μπορεί να προσεγγίσει την απόδοση της centralized εφαρμογής SMOTE, τουλάχιστον στο IID σενάριο.

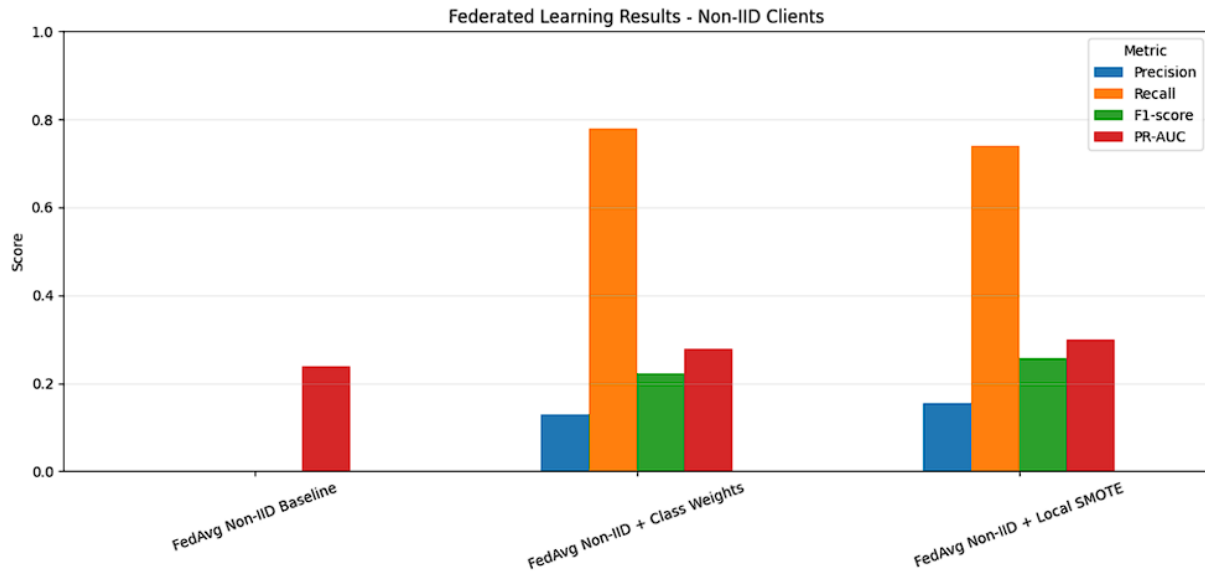
Όπως φαίνεται στο Σχήμα 6.1, το baseline μοντέλο δεν κατάφερε να εντοπίσει τη θετική κλάση, καθώς το Recall και το F1-score παρέμειναν μηδενικά. Αντίθετα, η χρήση class weights και Local SMOTE βελτίωσε σημαντικά την ικανότητα του μοντέλου να αναγνωρίζει περιστατικά εγκεφαλικού. Το class weights πέτυχε ελαφρώς υψηλότερο Recall, ενώ το Local SMOTE παρουσίασε καλύτερο F1-score και PR-AUC. Αυτό δείχνει ότι η τοπική εφαρμογή SMOTE σε federated περιβάλλον μπορεί να προσεγγίσει την απόδοση της centralized εφαρμογής SMOTE, τουλάχιστον στο IID σενάριο.



Σχήμα 6.1: Αποτελέσματα Federated Learning στα IID σενάρια.

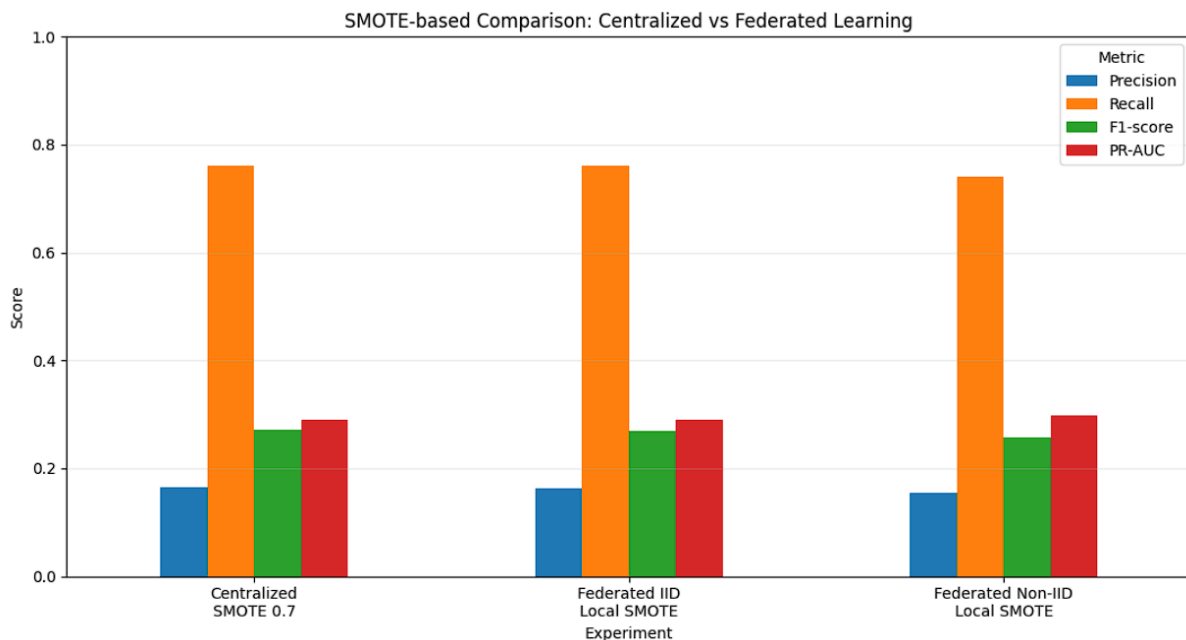
Στο Non-IID σενάριο, το Local SMOTE παρουσίασε ελαφρώς χαμηλότερο Recall, ίσο με 0.74, και F1-score 0.2569. Παρ' όλα αυτά, το PR-AUC ήταν 0.2988, δηλαδή ελαφρώς υψηλότερο από το αντίστοιχο IID πείραμα. Η μείωση του Recall δείχνει ότι η ανομοιογενής κατανομή των δεδομένων μεταξύ των clients επηρεάζει την απόδοση του μοντέλου, αλλά όχι σε βαθμό που να καθιστά τη μέθοδο μη αποτελεσματική. Αντίθετα, το Local SMOTE παρέμεινε ανταγωνιστικό και στο Non-IID σενάριο, παρουσιάζοντας πιο ισορροπημένη συμπεριφορά σε σχέση με τα class weights, κυρίως λόγω μικρότερου αριθμού false positives.

Η συμπεριφορά των τριών Non-IID σεναρίων παρουσιάζεται στο Σχήμα 6.2. Όπως φαίνεται, το baseline μοντέλο δεν κατάφερε να εντοπίσει τη θετική κλάση, καθώς το Recall και το F1-score παρέμειναν μηδενικά. Αντίθετα, τόσο η χρήση class weights όσο και η εφαρμογή Local SMOTE βελτίωσαν σημαντικά την απόδοση του FedAvg. Το class weights πέτυχε ελαφρώς υψηλότερο Recall, ενώ το Local SMOTE παρουσίασε καλύτερο PR-AUC, γεγονός που δείχνει ότι η τοπική εξισορρόπηση μπορεί να βοηθήσει ακόμη και όταν οι clients έχουν διαφορετική κατανομή δεδομένων.



Σχήμα 6.2: Αποτελέσματα Federated Learning Non IID

Λαμβάνοντας υπόψη τα παραπάνω, τα αποτελέσματα δείχνουν ότι οι τεχνικές αντιμετώπισης της ανισορροπίας είναι απαραίτητες για το συγκεκριμένο πρόβλημα. Τα baseline μοντέλα, είτε centralized είτε federated, δεν είναι επαρκή για την ανίχνευση της θετικής κλάσης. Αντίθετα, τα class weights και οι SMOTE-based τεχνικές βελτιώνουν σημαντικά το Recall και καθιστούν τα μοντέλα πιο χρήσιμα για το πρόβλημα της πρόβλεψης εγκεφαλικού. Από τις τεχνικές που εξετάστηκαν, το SMOTE και ειδικά το Local SMOTE στο federated περιβάλλον παρουσίασαν πολύ καλή ισορροπία μεταξύ Recall, F1-score και PR-AUC.



Σχήμα 6.3: Σύγκριση SMOTE Centralized – Federated Learning

Για να αξιολογηθεί πιο καθαρά η σχέση ανάμεσα στην κεντριοποιημένη και την κατανεμημένη εκπαίδευση, πραγματοποιήθηκε σύγκριση μεταξύ του centralized Logistic Regression με SMOTE 0.7, του FedAvg IID με Local SMOTE και του FedAvg Non-IID με Local SMOTE. Η σύγκριση αυτή

παρουσιάζεται στο Σχήμα 6.3 και εστιάζει στις βασικές μετρικές Precision, Recall, F1-score και PR-AUC.

Τα συνολικά ευρήματα υποστηρίζουν ότι το Federated Learning μπορεί να αποτελέσει μια ρεαλιστική εναλλακτική της centralized εκπαίδευσης σε προβλήματα υγείας, ειδικά όταν η συγκέντρωση των δεδομένων δεν είναι επιθυμητή ή δεν είναι εύκολα εφικτή. Στην έρευνα που πραγματοποιήθηκε, το federated μοντέλο με Local SMOTE κατάφερε να προσεγγίσει πολύ κοντά την απόδοση του centralized SMOTE, διατηρώντας παράλληλα τη βασική λογική ότι τα δεδομένα παραμένουν τοπικά στους clients. Αυτό το εύρημα είναι ιδιαίτερα σημαντικό, καθώς δείχνει ότι η κατανεμημένη εκπαίδευση μπορεί να συνδυαστεί με τεχνικές αντιμετώπισης της ανισορροπίας χωρίς μεγάλη απώλεια απόδοσης.

6.2 SMOTE vs CTGAN

Ένα από τα βασικά σημεία της πειραματικής αξιολόγησης ήταν η σύγκριση ανάμεσα σε κλασικές τεχνικές υπερδειγματοληψίας, όπως το SMOTE, και σε πιο σύνθετες generative προσεγγίσεις, όπως το CTGAN. Και οι δύο μέθοδοι είχαν τον ίδιο γενικό στόχο, να αντιμετωπίσουν την έντονη ανισορροπία των κλάσεων μέσω της ενίσχυσης της μειοψηφικής κλάσης. Ωστόσο, τα αποτελέσματα έδειξαν ότι οι δύο προσεγγίσεις δεν είχαν την ίδια αποτελεσματικότητα στο συγκεκριμένο πρόβλημα.

Το SMOTE βασίζεται στη δημιουργία νέων συνθετικών δειγμάτων της μειοψηφικής κλάσης μέσω παρεμβολής μεταξύ υπαρχόντων minority samples. Με απλά λόγια, η μέθοδος εντοπίζει κοντινά δείγματα της θετικής κλάσης και δημιουργεί νέα σημεία στον χώρο των χαρακτηριστικών, τα οποία βρίσκονται ανάμεσα σε αυτά. Η προσέγγιση αυτή είναι σχετικά απλή, αλλά μπορεί να είναι πολύ αποτελεσματική όταν τα υπάρχοντα minority samples είναι λίγα, αλλά αρκετά ώστε να δώσουν μια βασική εικόνα της κατανομής της θετικής κλάσης.

Στο συγκεκριμένο dataset, η χρήση SMOTE αποδείχθηκε ιδιαίτερα χρήσιμη για τη Logistic Regression. Το centralized Logistic Regression με SMOTE 0.7 πέτυχε Recall ίσο με 0.76, F1-score περίπου 0.2705 και PR-AUC περίπου 0.2896. Τα αποτελέσματα αυτά ήταν σαφώς καλύτερα από το baseline Logistic Regression, το οποίο εντόπισε μόνο 1 από τα 50 πραγματικά stroke cases. Η βελτίωση αυτή δείχνει ότι η προσθήκη συνθετικών minority samples βοήθησε το μοντέλο να μετατοπίσει το decision boundary και να αναγνωρίζει πιο συχνά τη θετική κλάση.

Παρόμοια συμπεριφορά παρατηρήθηκε και στις παραλλαγές του SMOTE. Το Borderline-SMOTE παρουσίασε ακόμα καλύτερο F1-score, περίπου 0.2910, και Recall ίσο με 0.78. Η τεχνική αυτή εστιάζει περισσότερο στα minority samples που βρίσκονται κοντά στα όρια απόφασης μεταξύ των δύο κλάσεων. Αυτό φαίνεται ότι βοήθησε τη Logistic Regression, καθώς το μοντέλο κατάφερε να βελτιώσει την ικανότητά του να διαχωρίζει τα δύσκολα περιστατικά. Το αποτέλεσμα αυτό δείχνει ότι, για το συγκεκριμένο πρόβλημα, η ενίσχυση των minority samples κοντά στα όρια των κλάσεων είχε πρακτική αξία.

Αντίθετα, το CTGAN παρουσίασε πιο περιορισμένη βελτίωση. Στην περίπτωση της Logistic Regression με CTGAN 0.7, το Recall έφτασε στο 0.36, το F1-score στο 0.2 και το PR-AUC στο 0.15. Παρότι αυτά τα αποτελέσματα ήταν καλύτερα από το baseline, παρέμειναν αρκετά χαμηλότερα από τα αντίστοιχα αποτελέσματα του SMOTE. Δηλαδή, το CTGAN βοήθησε το μοντέλο να εντοπίζει περισσότερα stroke cases από ό,τι το απλό baseline, αλλά δεν κατάφερε να πλησιάσει την απόδοση των SMOTE-based τεχνικών.

Η διαφορά αυτή μπορεί να εξηγηθεί από τη φύση του dataset και από τον μικρό αριθμό διαθέσιμων θετικών δειγμάτων. Το CTGAN είναι ένα πιο σύνθετο generative μοντέλο, το οποίο προσπαθεί να μάθει

την κατανομή των δεδομένων και να παράγει νέα συνθετικά δείγματα που μοιάζουν με τα πραγματικά. Για να λειτουργήσει αποτελεσματικά, όμως, χρειάζεται επαρκή ποσότητα και ποιότητα δεδομένων, ειδικά για την κλάση που θέλουμε να μοντελοποιήσουμε. Εν προκειμένω, η θετική κλάση περιείχε μόνο 199 δείγματα στο training set. Αυτός ο αριθμός είναι αρκετά περιορισμένος για την εκπαίδευση ενός generative μοντέλου που πρέπει να μάθει αξιόπιστα τη δομή της μειοψηφικής κλάσης.

Επιπλέον, το CTGAN εκπαιδεύτηκε για περιορισμένο αριθμό epochs, συγκεκριμένα για 30 epochs. Η απόφαση αυτή έγινε για λόγους υπολογιστικού κόστους και χρόνου εκτέλεσης. Ωστόσο, είναι πιθανό ότι ο αριθμός αυτός δεν ήταν αρκετός ώστε το generative μοντέλο να μάθει πλήρως την κατανομή της θετικής κλάσης. Από την άλλη πλευρά, η αύξηση των epochs δεν εγγυάται απαραίτητα καλύτερα αποτελέσματα, καθώς σε τόσο μικρή minority class υπάρχει και ο κίνδυνος το CTGAN να παράγει δείγματα που δεν γενικεύουν καλά ή που απλώς αναπαράγουν θορυβώδη πρότυπα των αρχικών δεδομένων.

Ένας ακόμη παράγοντας είναι ότι το Stroke Prediction Dataset περιέχει tabular δεδομένα με συνδυασμό αριθμητικών και κατηγορικών χαρακτηριστικών. Τα tabular datasets συχνά παρουσιάζουν δυσκολίες για generative μοντέλα, ειδικά όταν υπάρχουν έντονα άνισες κατηγορίες και η μειοψηφική κλάση είναι μικρή. Το CTGAN είναι σχεδιασμένο για tabular δεδομένα, όμως αυτό δεν σημαίνει ότι θα υπερτερεί πάντα έναντι απλούστερων μεθόδων. Στην παρούσα περίπτωση, φαίνεται ότι η απλούστερη γραμμική λογική του SMOTE ήταν πιο σταθερή και πιο αποτελεσματική για την εκπαίδευση των classifiers.

Η σύγκριση αυτή είναι σημαντική, γιατί δείχνει ότι πιο σύνθετη μέθοδος δεν σημαίνει απαραίτητα καλύτερη μέθοδος. Το CTGAN είναι θεωρητικά πιο ισχυρό, επειδή μπορεί να μάθει σύνθετες σχέσεις μεταξύ χαρακτηριστικών και να παράγει νέα συνθετικά δείγματα. Ωστόσο, η αποτελεσματικότητα ενός τέτοιου μοντέλου εξαρτάται σε μεγάλο βαθμό από τα διαθέσιμα δεδομένα, τις παραμέτρους εκπαίδευσης και τη σταθερότητα του. Στο συγκεκριμένο πείραμα, το CTGAN δεν κατάφερε να παραγάγει συνθετικά samples που να βοηθήσουν τους classifiers περισσότερο από το SMOTE.

Αντίθετα, το SMOTE λειτουργήσε πιο άμεσα και σταθερά. Δεν προσπάθησε να μάθει ολόκληρη την κατανομή της θετικής κλάσης, αλλά δημιούργησε νέα δείγματα με βάση τις υπάρχουσες τοπικές σχέσεις μεταξύ minority samples. Αυτή η απλότητα φαίνεται ότι ήταν πλεονέκτημα στο συγκεκριμένο dataset. Επειδή τα θετικά δείγματα ήταν λίγα, μια μέθοδος που βασίζεται σε τοπική παρεμβολή αποδείχθηκε πιο πρακτική από ένα generative μοντέλο που απαιτεί περισσότερα δεδομένα για να εκπαιδευτεί αξιόπιστα.

Στα Random Forest και XGBoost, τα αποτελέσματα των τεχνικών υπερδειγματοληψίας ήταν πιο περιορισμένα σε σχέση με τη Logistic Regression. Το Random Forest παρέμεινε γενικά αδύναμο στον εντοπισμό της θετικής κλάσης, ακόμα και μετά την εφαρμογή SMOTE ή CTGAN. Το XGBoost παρουσίασε καλύτερη συμπεριφορά με class weights σε σχέση με τις τεχνικές υπερδειγματοληψίας. Αυτό δείχνει ότι η αποτελεσματικότητα κάθε τεχνικής εξισορρόπησης εξαρτάται και από το μοντέλο που χρησιμοποιείται. Συνεπώς, η Logistic Regression φάνηκε να αξιοποιεί καλύτερα τα συνθετικά δείγματα που δημιουργήθηκαν με SMOTE.

Ιδιαίτερο ενδιαφέρον έχει και η σύγκριση του centralized SMOTE με το federated Local SMOTE. Το centralized Logistic Regression με SMOTE 0.7 και το Federated IID με Local SMOTE παρουσίασαν σχεδόν ταυτόσημες τιμές σε βασικές μετρικές. Συγκεκριμένα, και τα δύο πέτυχαν Recall 0.76, ενώ οι τιμές F1-score και PR-AUC ήταν σχεδόν ίδιες. Αυτό δείχνει ότι η εφαρμογή SMOTE τοπικά στους clients μπορεί να προσεγγίσει πολύ καλά την απόδοση της centralized εφαρμογής SMOTE, τουλάχιστον όταν οι clients έχουν IID κατανομή.

Στο Non-IID federated σενάριο, το Local SMOTE είχε ελαφρώς χαμηλότερο Recall, ίσο με 0.74, αλλά παρέμεινε αρκετά κοντά στην centralized προσέγγιση. Αυτό είναι σημαντικό, γιατί δείχνει ότι ακόμα και όταν οι clients έχουν διαφορετική κατανομή θετικών περιστατικών, η τοπική εξισορρόπηση μπορεί να βοηθήσει το ενιαίο μοντέλο να διατηρήσει ανταγωνιστική απόδοση. Αν και η ανομοιογένεια των δεδομένων επηρεάζει ελαφρώς την απόδοση, δεν αναιρεί τη χρησιμότητα του Local SMOTE.

Με βάση τα συνολικά αποτελέσματα, το SMOTE μπορεί να θεωρηθεί η πιο επιτυχημένη τεχνική δημιουργίας συνθετικών minority samples στην παρούσα εργασία. Δεν έδωσε πάντα το υψηλότερο Recall, καθώς τα class weights σε ορισμένες περιπτώσεις πέτυχαν ελαφρώς μεγαλύτερη ευαισθησία στη θετική κλάση. Ωστόσο, το SMOTE και ειδικά το Local SMOTE παρουσίασαν καλύτερη ισορροπία μεταξύ Recall, Precision, F1-score και PR-AUC. Αυτό σημαίνει ότι δεν αύξησαν απλώς τον αριθμό των εντοπισμένων stroke cases, αλλά διατήρησαν και πιο ελεγχόμενο αριθμό false positives.

Συμπερασματικά, η σύγκριση SMOTE και CTGAN έδειξε ότι για το συγκεκριμένο dataset και το συγκεκριμένο πρόβλημα, οι κλασικές SMOTE-based τεχνικές ήταν πιο αποτελεσματικές από το CTGAN. Το CTGAN αποτελεί μια ενδιαφέρουσα και πιο σύνθετη προσέγγιση, αλλά στην περίπτωση αυτή δεν κατάφερε να αξιοποιηθεί στον ίδιο βαθμό. Η περιορισμένη διαθεσιμότητα θετικών δειγμάτων, ο μικρός αριθμός εποχών εκπαίδευσης και η πολυπλοκότητα της μοντελοποίησης tabular δεδομένων πιθανώς συνέβαλαν στη χαμηλότερη απόδοσή του. Αντίθετα, το SMOTE αποδείχθηκε πιο σταθερό, πιο ερμηνεύσιμο και πιο αποδοτικό τόσο στο centralized όσο και στο federated περιβάλλον.

6.3 Ερμηνεία Αποτελεσμάτων

Η συνολική ερμηνεία των αποτελεσμάτων δείχνει ότι το βασικότερο πρόβλημα της παρούσας εργασίας δεν ήταν μόνο η επιλογή του κατάλληλου μοντέλου, αλλά κυρίως η διαχείριση της έντονης ανισορροπίας των δεδομένων. Το dataset που χρησιμοποιήθηκε περιείχε πολύ μικρό αριθμό περιστατικών εγκεφαλικού σε σχέση με τα δείγματα χωρίς εγκεφαλικό. Αυτό είχε ως αποτέλεσμα τα μοντέλα να επηρεάζονται έντονα από την πλειοψηφική κλάση και, όταν δεν εφαρμόστηκε κάποια τεχνική εξισορρόπησης, να αποτυγχάνουν σχεδόν πλήρως στην αναγνώριση της θετικής κλάσης.

Το φαινόμενο αυτό φάνηκε καθαρά στα baseline πειράματα. Τόσο στο centralized όσο και στο federated περιβάλλον, τα μοντέλα χωρίς κάποια παρέμβαση στην ανισορροπία εμφάνισαν υψηλό Accuracy, αλλά πολύ χαμηλό ή μηδενικό Recall. Αυτό σημαίνει ότι η υψηλή ακρίβεια δεν αντανακλούσε πραγματική ικανότητα πρόβλεψης εγκεφαλικού επεισοδίου. Αντίθετα, οφειλόταν κυρίως στο γεγονός ότι η πλειοψηφία των δειγμάτων ανήκε στην κλάση no stroke. Επομένως, ένα μοντέλο μπορούσε να πετυχαίνει φαινομενικά καλή απόδοση προβλέποντας σχεδόν πάντα την αρνητική κλάση.

Αυτό είναι ιδιαίτερα σημαντικό για προβλήματα υγείας. Γενικότερα σε ένα τέτοιο πλαίσιο, καθίσταται πολύ πιο κρίσιμη η αποτυχία εντοπισμού ενός θετικού περιστατικού σε σύγκριση με ένα μη ιατρικό πρόβλημα μηχανικής μάθησης. Ειδικότερα, θεωρείται πιο σοβαρό σφάλμα ένα πραγματικό περιστατικό εγκεφαλικού το οποίο, το μοντέλο το ταξινομεί λανθασμένα, σε αντίθεση με ένα περιστατικό που ταξινομείται στην κλάση stroke ενώ στην πραγματικότητα δεν ανήκει στην κλάση αυτή. Για αυτόν τον λόγο, το Recall αποτέλεσε κεντρική μετρική στην αξιολόγηση. Η εργασία δεν είχε ως στόχο απλώς να μεγιστοποιήσει τη συνολική ακρίβεια, αλλά να εξετάσει αν τα μοντέλα μπορούν να εντοπίσουν όσο το δυνατόν περισσότερα πραγματικά περιστατικά της μειοψηφικής κλάσης.

Η χρήση class weights έδειξε ότι η τροποποίηση της συνάρτησης απώλειας μπορεί να αλλάξει σημαντικά τη συμπεριφορά των μοντέλων. Όταν η θετική κλάση έλαβε μεγαλύτερο βάρος κατά την εκπαίδευση, τα μοντέλα έγιναν πολύ πιο ευαίσθητα στα περιστατικά εγκεφαλικού. Αυτό φάνηκε κυρίως στη Logistic Regression, τόσο στο centralized όσο και στο federated περιβάλλον. Στα αντίστοιχα

πειράματα, το Recall αυξήθηκε σημαντικά, φτάνοντας περίπου στο 0.80. Με άλλα λόγια, το μοντέλο κατάφερε να εντοπίσει περίπου 8 στα 10 πραγματικά stroke cases του test set.

Ωστόσο, παρά την αύξηση του Recall, παρατηρήθηκε μείωση απόδοσης σε ορισμένους σημαντικούς δείκτες. Ο κυριότερος από αυτούς είναι τα false positives, όπου η χρήση της τεχνικής class weight οδήγησε σε σημαντική αύξηση τους. Αυτό με τη σειρά του, φέρνει ως αποτέλεσμα τη μείωση του Precision. Η συμπεριφορά αυτή είναι αναμενόμενη, επειδή το μοντέλο, δίνοντας μεγαλύτερη σημασία στη θετική κλάση, γίνεται πιο πρόθυμο να προβλέψει stroke. Αφενός, αυτό μειώνει τα false negatives, αφετέρου αυξάνει τα false positives.

Η παραπάνω παρατήρηση δείχνει ότι δεν υπάρχει μία απόλυτα ιδανική λύση. Στα ιατρικά προβλήματα, ένα υψηλό Recall είναι ιδιαίτερα σημαντικό, αλλά δεν μπορεί να αγνοηθεί πλήρως η Precision. Ένα μοντέλο που προβλέπει υπερβολικά πολλά περιστατικά ως θετικά μπορεί να δημιουργήσει μεγάλο αριθμό ψευδών συναγερμών. Στην πράξη, αυτό θα μπορούσε να οδηγήσει σε περιττές εξετάσεις ή επιπλέον φόρτο για τους επαγγελματίες υγείας. Το γεγονός αυτό οδηγεί στο συμπέρασμα ότι, το ζητούμενο δεν είναι μόνο να αυξηθεί το Recall, αλλά να επιτευχθεί μια πιο ισορροπημένη συμπεριφορά μεταξύ Recall και Precision.

Σε αυτό το σημείο, οι SMOTE-based τεχνικές παρουσίασαν ιδιαίτερο ενδιαφέρον. Το SMOTE δεν τροποποιεί απλώς το βάρος της θετικής κλάσης στη συνάρτηση απώλειας, αλλά αλλάζει το ίδιο το training set μέσω της δημιουργίας συνθετικών minority samples. Κατ' αυτόν τον τρόπο, το μοντέλο εκπαιδεύεται σε ένα πιο ισορροπημένο σύνολο δεδομένων και αποκτά περισσότερες πληροφορίες για τη μειοψηφική κλάση. Συγκεκριμένα, το SMOTE και ειδικά το Borderline-SMOTE οδήγησαν σε πολύ καλύτερη ισορροπία μεταξύ Recall, F1-score και PR-AUC σε σχέση με τα baseline μοντέλα.

Η Logistic Regression φάνηκε να αξιοποιεί ιδιαίτερα καλά το SMOTE. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι η Logistic Regression είναι ένα γραμμικό μοντέλο και επηρεάζεται αρκετά από την κατανομή των δειγμάτων στον χώρο των χαρακτηριστικών. Όταν η μειοψηφική κλάση είναι πολύ μικρή, το decision boundary του μοντέλου μετατοπίζεται προς την πλειοψηφική κλάση και δεν εντοπίζει αρκετά θετικά περιστατικά. Με τη χρήση SMOTE, δημιουργούνται επιπλέον σημεία της θετικής κλάσης, τα οποία βοηθούν το μοντέλο να διαμορφώσει ένα πιο κατάλληλο όριο απόφασης. Αντίθετα, το Random Forest δεν παρουσίασε την ίδια βελτίωση. Ακόμα και με τεχνικές εξισορρόπησης, η ικανότητά του να εντοπίζει τη θετική κλάση παρέμεινε περιορισμένη. Αυτό δείχνει ότι η απλή εφαρμογή μιας τεχνικής υπερδειγματοληψίας δεν εγγυάται βελτίωση σε όλα τα μοντέλα. Η αποτελεσματικότητα κάθε τεχνικής εξαρτάται από τον τρόπο με τον οποίο το εκάστοτε μοντέλο μαθαίνει από τα δεδομένα. Στο συγκεκριμένο dataset, η Logistic Regression είχε πιο σταθερή και προβλέψιμη συμπεριφορά σε σχέση με τα συνθετικά δείγματα του SMOTE.

Το XGBoost παρουσίασε ενδιαμέση συμπεριφορά. Το baseline μοντέλο είχε χαμηλό Recall, αλλά η χρήση scale_pos_weight βελτίωσε σημαντικά την ανίχνευση της θετικής κλάσης. Αυτό δείχνει ότι το XGBoost μπορεί να επωφεληθεί από την εσωτερική στάθμιση της μειοψηφικής κλάσης. Ωστόσο, τα SMOTE-based πειράματα με XGBoost δεν ξεπέρασαν τη συμπεριφορά της Logistic Regression. Επομένως, παρότι το XGBoost θεωρείται γενικά ισχυρό μοντέλο, στο συγκεκριμένο πρόβλημα δεν ήταν απαραίτητα η καλύτερη επιλογή με βάση το Recall, το F1-score και το PR-AUC.

Το CTGAN, παρότι αποτελεί πιο σύνθετη μέθοδο παραγωγής συνθετικών δεδομένων, δεν κατάφερε να ξεπεράσει το SMOTE. Αυτό είναι ένα σημαντικό εύρημα, γιατί δείχνει ότι η πολυπλοκότητα μιας μεθόδου δεν οδηγεί πάντα σε καλύτερα αποτελέσματα. Το CTGAN χρειάζεται επαρκή δεδομένα για να μάθει την κατανομή της κλάσης που προσπαθεί να αναπαραγάγει. Ωστόσο, η θετική κλάση ήταν πολύ

μικρή. Ως αποτέλεσμα, το generative μοντέλο πιθανώς δεν είχε αρκετή πληροφορία ώστε να δημιουργήσει συνθετικά δείγματα που να βελτιώνουν ουσιαστικά την εκπαίδευση των classifiers.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η ερμηνεία των federated αποτελεσμάτων. Στο IID σενάριο, τα αποτελέσματα του Federated Learning ήταν πολύ κοντά στα centralized αποτελέσματα της Logistic Regression. Αυτό είναι λογικό, επειδή οι τρεις clients δημιουργήθηκαν ώστε να έχουν παρόμοια κατανομή κλάσεων με το αρχικό training set. Κατά συνέπεια, κάθε client εκπαιδεύτηκε σε δεδομένα που έμοιαζαν αρκετά με τα δεδομένα των υπόλοιπων clients. Σε τέτοιες συνθήκες, το FedAvg μπορεί να λειτουργήσει σταθερά, καθώς οι τοπικές ενημερώσεις των clients δεν αποκλίνουν πολύ μεταξύ τους.

Το σημαντικότερο εύρημα στο federated κομμάτι ήταν ότι το Local SMOTE στο IID σενάριο παρουσίασε σχεδόν ίδια απόδοση με το centralized SMOTE. Συγκεκριμένα, οι τιμές Recall, F1-score και PR-AUC ήταν πρακτικά ίδιες. Αυτό δείχνει ότι, τουλάχιστον σε συνθήκες IID κατανομής, η τοπική εφαρμογή SMOTE στους clients μπορεί να αντικαταστήσει σε μεγάλο βαθμό την centralized εφαρμογή SMOTE, χωρίς σημαντική απώλεια απόδοσης. Το συμπέρασμα αυτό, δείχνει ότι η εκπαίδευση μπορεί να γίνει με κατανεμημένο τρόπο, χωρίς να συγκεντρωθούν όλα τα δεδομένα σε ένα σημείο.

Στο Non-IID σενάριο, η απόδοση μειώθηκε ελαφρώς, αλλά όχι δραματικά. Το γεγονός ότι οι clients είχαν διαφορετική κατανομή θετικών περιστατικών έκανε την εκπαίδευση πιο απαιτητική. Κάθε client έβλεπε διαφορετική εικόνα του προβλήματος, με αποτέλεσμα τα τοπικά μοντέλα να μην ενημερώνονται με απόλυτα ομοιόμορφο τρόπο. Παρ' όλα αυτά, τόσο τα class weights όσο και το Local SMOTE διατήρησαν ικανοποιητική απόδοση. Αυτό δείχνει ότι το FedAvg μπορεί να παραμείνει χρήσιμο ακόμα και όταν οι clients δεν έχουν πλήρως ομοιόμορφη κατανομή δεδομένων.

Η σύγκριση μεταξύ class weights και Local SMOTE στο federated περιβάλλον δείχνει μια επαναλαμβανόμενη τάση. Τα class weights πέτυχαν ελαφρώς υψηλότερο Recall, αλλά με περισσότερα false positives. Αντίθετα, το Local SMOTE είχε λίγο χαμηλότερο Recall, αλλά καλύτερο F1-score, καλύτερη Precision και υψηλότερο PR-AUC. Επομένως, αν το μοναδικό ζητούμενο ήταν η μεγιστοποίηση του Recall, τα class weights θα μπορούσαν να θεωρηθούν πιο επιθετική επιλογή. Αν όμως ζητείται πιο ισορροπημένη συνολική συμπεριφορά, το Local SMOTE φαίνεται πιο κατάλληλο.

Αυτή η διαφορά είναι σημαντική για την πρακτική ερμηνεία των αποτελεσμάτων. Σε ένα σύστημα υποβοήθησης ιατρικής απόφασης, η επιλογή μεταξύ class weights και SMOTE δεν θα ήταν καθαρά τεχνική, αλλά θα εξαρτιόταν από το επιθυμητό επίπεδο ευαισθησίας. Αν προτεραιότητα είναι να μη χαθεί σχεδόν κανένα πιθανό περιστατικό, τότε ένα μοντέλο με υψηλότερο Recall μπορεί να είναι προτιμότερο, ακόμα και αν παράγει περισσότερα false positives. Αν όμως χρειάζεται καλύτερη ισορροπία μεταξύ εντοπισμού θετικών περιστατικών και περιορισμού ψευδών συναγερμών, τότε το SMOTE μπορεί να είναι πιο κατάλληλη επιλογή.

Τα αποτελέσματα δείχνουν επίσης ότι το πρόβλημα της πρόβλεψης εγκεφαλικού δεν μπορεί να αντιμετωπιστεί αποτελεσματικά μόνο με την επιλογή ενός πιο σύνθετου μοντέλου. Η διαχείριση της ανισορροπίας αποδείχθηκε πιο σημαντική από την πολυπλοκότητα του classifier. Για παράδειγμα, η απλή Logistic Regression με κατάλληλη εξισορρόπηση ξεπέρασε σε αρκετές περιπτώσεις πιο σύνθετα μοντέλα. Αυτό δείχνει ότι, σε μικρά και ανισόρροπα ιατρικά datasets, η σωστή προεπεξεργασία και η σωστή πειραματική μεθοδολογία μπορεί να έχουν μεγαλύτερη σημασία από την επιλογή ενός πολύπλοκου αλγορίθμου.

Παρά τα θετικά αποτελέσματα, πρέπει να αναγνωριστούν και ορισμένοι περιορισμοί στην ερμηνεία. Πρώτον, το dataset είναι σχετικά μικρό ως προς τον αριθμό των θετικών περιστατικών. Το test set περιέχει μόνο 50 stroke cases, επομένως μικρές αλλαγές στον αριθμό των σωστών προβλέψεων

μπορούν να επηρεάσουν αισθητά το Recall και το F1-score. Για παράδειγμα, η διαφορά μεταξύ 37 και 40 σωστά εντοπισμένων περιστατικών φαίνεται αριθμητικά ως διαφορά 0.06 στο Recall. Συνεπώς, τα αποτελέσματα πρέπει να ερμηνευθούν με προσοχή και όχι ως απόλυτη απόδειξη υπεροχής μιας μεθόδου.

Δεύτερον, το Federated Learning υλοποιήθηκε ως προσομοίωση. Οι clients δημιουργήθηκαν τεχνητά από το ίδιο αρχικό training set και δεν αντιστοιχούν σε πραγματικά νοσοκομεία με ανεξάρτητα δεδομένα. Αυτό σημαίνει ότι, παρότι η προσομοίωση είναι χρήσιμη για την αξιολόγηση της μεθόδου, δεν μπορεί να αναπαραστήσει πλήρως την πολυπλοκότητα ενός πραγματικού federated περιβάλλοντος. Σε πραγματικές συνθήκες, οι διαφορές μεταξύ των clients μπορεί να είναι μεγαλύτερες, τόσο ως προς την κατανομή των κλάσεων όσο και ως προς τα χαρακτηριστικά των ασθενών.

Τρίτον, δεν εφαρμόστηκαν πρόσθετοι μηχανισμοί ιδιωτικότητας, όπως differential privacy ή secure aggregation. Η εργασία αυτή, εστίασε στη βασική λογική του Federated Learning και στη σύγκριση της απόδοσης με centralized μοντέλα. Έτσι, παρότι τα δεδομένα στην προσομοίωση παραμένουν τοπικά στους clients, δεν μπορεί να υποστηριχθεί ότι η υλοποίηση καλύπτει πλήρως όλες τις απαιτήσεις ενός πραγματικού συστήματος. Αυτό αποτελεί πιθανή κατεύθυνση για μελλοντική επέκταση.

Συμπερασματικά, τα αποτελέσματα της εργασίας δείχνουν ότι η αντιμετώπιση της ανισορροπίας είναι καθοριστικός παράγοντας για την πρόβλεψη εγκεφαλικού επεισοδίου. Τα baseline μοντέλα δεν ήταν επαρκή, ενώ οι τεχνικές class weights και SMOTE βελτίωσαν σημαντικά την ανίχνευση της θετικής κλάσης. Επιπλέον, το Federated Learning με Local SMOTE κατάφερε να προσεγγίσει πολύ κοντά την απόδοση της centralized εκπαίδευσης, ειδικά στο IID σενάριο. Αυτό υποστηρίζει την άποψη ότι η federated προσέγγιση μπορεί να αποτελέσει μια πρακτική και αξιολογη εναλλακτική σε περιβάλλοντα όπου η συγκέντρωση ιατρικών δεδομένων δεν είναι επιθυμητή ή δεν είναι εύκολα εφικτή.

Κεφάλαιο 7ο: Συμπεράσματα και Μελλοντική Έρευνα

7.1 Σύνοψη Ευρημάτων

Η παρούσα διπλωματική εργασία είχε ως αντικείμενο την αξιολόγηση τεχνικών μηχανικής μάθησης για την πρόβλεψη εγκεφαλικού επεισοδίου, με ιδιαίτερη έμφαση στην αντιμετώπιση της ανισορροπίας των δεδομένων και στη διερεύνηση της δυνατότητας χρήσης Federated Learning. Το πρόβλημα που εξετάστηκε είναι ιδιαίτερα απαιτητικό, καθώς τα περιστατικά εγκεφαλικού αποτελούν τη μειοψηφική κλάση του dataset, ενώ τα δείγματα χωρίς εγκεφαλικό είναι πολύ περισσότερα. Αυτή η ανισορροπία επηρέασε σημαντικά τη συμπεριφορά των μοντέλων και αποτέλεσε βασικό σημείο της πειραματικής αξιολόγησης.

Αρχικά, εξετάστηκε η συμπεριφορά μοντέλων μηχανικής μάθησης σε centralized περιβάλλον. Τα μοντέλα που χρησιμοποιήθηκαν ήταν η Logistic Regression, το Random Forest και το XGBoost. Η αξιολόγηση έδειξε ότι τα baseline μοντέλα, δηλαδή τα μοντέλα χωρίς κάποια τεχνική αντιμετώπισης της ανισορροπίας, δεν ήταν επαρκή για το συγκεκριμένο πρόβλημα. Παρότι σε ορισμένες περιπτώσεις εμφάνισαν υψηλό Accuracy, η ικανότητά τους να εντοπίζουν τα πραγματικά περιστατικά εγκεφαλικού ήταν πολύ περιορισμένη. Αυτό επιβεβαίωσε ότι το Accuracy δεν μπορεί να χρησιμοποιηθεί ως κύρια μετρική αξιολόγησης σε τόσο ανισόρροπα ιατρικά δεδομένα.

Το βασικό συμπέρασμα από τα centralized πειράματα είναι ότι η διαχείριση της ανισορροπίας είναι καθοριστική για την απόδοση των μοντέλων. Η χρήση class weights βελτίωσε σημαντικά το Recall, ιδιαίτερα στη Logistic Regression και στο XGBoost. Στην περίπτωση της Logistic Regression, τα class weights οδήγησαν σε Recall ίσο με 0.80, γεγονός που σημαίνει ότι το μοντέλο κατάφερε να εντοπίσει τα περισσότερα πραγματικά stroke cases του test set. Ωστόσο, η βελτίωση αυτή συνοδεύτηκε από αυξημένο αριθμό false positives, κάτι που αποτυπώθηκε στο χαμηλό Precision.

Οι τεχνικές SMOTE παρουσίασαν επίσης πολύ σημαντική βελτίωση σε σχέση με τα baseline μοντέλα. Ιδιαίτερα στη Logistic Regression, το SMOTE 0.7 και το Borderline-SMOTE έδωσαν από τα πιο ισορροπημένα αποτελέσματα. Το SMOTE 0.7 πέτυχε Recall 0.76 και F1-score περίπου 0.2705, ενώ το Borderline-SMOTE πέτυχε Recall 0.78 και F1-score περίπου 0.2910. Τα αποτελέσματα αυτά δείχνουν ότι η δημιουργία συνθετικών δειγμάτων της μειοψηφικής κλάσης βοήθησε το μοντέλο να μάθει καλύτερα τη συμπεριφορά των θετικών περιστατικών.

Αντίθετα, το CTGAN δεν κατάφερε να ξεπεράσει τις SMOTE-based τεχνικές. Παρότι βελτίωσε την απόδοση σε σχέση με το απλό baseline, τα αποτελέσματα παρέμειναν χαμηλότερα από αυτά του SMOTE. Στην περίπτωση της Logistic Regression με CTGAN 0.7, το Recall έφτασε στο 0.36, τιμή η οποία δείχνει βελτίωση, αλλά όχι στον βαθμό που επιτεύχθηκε με το SMOTE ή τα class weights. Το εύρημα αυτό δείχνει ότι οι πιο σύγχρονες τεχνικές δεν είναι απαραίτητο ότι οδηγούν σε καλύτερα αποτελέσματα, σε αντίθεση με τις λιγότερο σύνθετες τεχνικές.

Ένα ακόμη σημαντικό εύρημα ήταν ότι η Logistic Regression, παρότι είναι απλούστερο μοντέλο σε σχέση με το Random Forest και το XGBoost, παρουσίασε ιδιαίτερα καλή συμπεριφορά όταν συνδυάστηκε με κατάλληλες τεχνικές εξισορρόπησης. Αυτό δείχνει ότι σε μικρά και ανισόρροπα ιατρικά datasets η πολυπλοκότητα του μοντέλου δεν είναι πάντα ο σημαντικότερος παράγοντας. Αντίθετα, η σωστή προεπεξεργασία, η επιλογή κατάλληλης τεχνικής εξισορρόπησης και η χρήση κατάλληλων μετρικών αξιολόγησης μπορούν να επηρεάσουν περισσότερο το τελικό αποτέλεσμα.

Στο δεύτερο μέρος της εργασίας εξετάστηκε η χρήση Federated Learning. Το αρχικό training set χωρίστηκε σε τρεις clients, οι οποίοι προσομοίωσαν διαφορετικούς φορείς υγείας. Η federated

εκπαίδευση πραγματοποιήθηκε με τον αλγόριθμο FedAvg και μοντέλο Logistic Regression. Στη προσέγγιση αυτή, εξετάστηκαν δύο σενάρια κατανομής δεδομένων, IID και Non-IID, ώστε να αξιολογηθεί η συμπεριφορά του federated μοντέλου τόσο σε ομοιόμορφη όσο και σε ανομοιογενή κατανομή μεταξύ των clients.

Τα federated baseline πειράματα έδειξαν ότι το πρόβλημα της ανισορροπίας παραμένει έντονο και στο federated περιβάλλον. Τόσο στο IID όσο και στο Non-IID σενάριο, το baseline FedAvg μοντέλο απέτυχε να εντοπίσει τη θετική κλάση, παρουσιάζοντας μηδενικό Recall και F1-score. Παρά το υψηλό Accuracy, το μοντέλο ουσιαστικά ταξινόμησε όλα τα θετικά δείγματα ως no stroke. Το αποτέλεσμα αυτό ήταν αναμενόμενο και επιβεβαίωσε ότι η federated εκπαίδευση από μόνη της δεν αρκεί για την αντιμετώπιση της ανισορροπίας.

Η χρήση class weights στο federated περιβάλλον βελτίωσε σημαντικά το Recall. Στο IID σενάριο, το FedAvg με class weights πέτυχε Recall 0.80, ενώ στο Non-IID σενάριο πέτυχε Recall 0.78. Αυτό δείχνει ότι η στάθμιση της θετικής κλάσης μπορεί να λειτουργήσει αποτελεσματικά και σε κατανεμημένο περιβάλλον. Ωστόσο, όπως και στα centralized πειράματα, η αύξηση του Recall συνοδεύτηκε από μεγάλο αριθμό false positives, γεγονός που περιορίσε την Precision και το F1-score.

Το Local SMOTE αποτέλεσε την πιο ισορροπημένη federated προσέγγιση. Στο IID σενάριο, το FedAvg με Local SMOTE πέτυχε αξιοπρεπείς τιμές στις μετρικές Recall, F1-score και PR-AUC. Οι τιμές αυτές ήταν σχεδόν ίδιες με το centralized Logistic Regression με SMOTE 0.7. Αυτό αποτελεί ένα από τα σημαντικότερα ευρήματα της εργασίας, καθώς δείχνει ότι η τοπική εφαρμογή SMOTE στους clients μπορεί να προσεγγίσει την απόδοση της centralized εφαρμογής SMOTE χωρίς να απαιτείται η συγκέντρωση όλων των δεδομένων σε ένα ενιαίο training set.

Στο Non-IID σενάριο, το Local SMOTE παρουσίασε ελαφρώς χαμηλότερο Recall, αλλά διατήρησε καλή συνολική συμπεριφορά. Σε σύγκριση με τα class weights, το Local SMOTE είχε μικρότερο αριθμό false positives και καλύτερες τιμές Precision, F1-score και PR-AUC. Γι' αυτό, παρότι τα class weights πέτυχαν ελαφρώς υψηλότερο Recall, το Local SMOTE μπορεί να θεωρηθεί πιο ισορροπημένη επιλογή, ιδιαίτερα όταν δεν ενδιαφέρει μόνο η ανίχνευση όσο το δυνατόν περισσότερων θετικών περιστατικών, αλλά και ο περιορισμός των ψευδών θετικών προβλέψεων.

Σε γενικές γραμμές, τα αποτελέσματα της εργασίας δείχνουν ότι το Federated Learning μπορεί να προσεγγίσει την απόδοση της centralized εκπαίδευσης, ιδιαίτερα όταν οι clients έχουν IID κατανομή και εφαρμόζεται τοπική εξισορρόπηση. Η σύγκριση μεταξύ centralized SMOTE και federated Local SMOTE έδειξε σχεδόν ταυτόσημη συμπεριφορά στο IID σενάριο, γεγονός που υποστηρίζει τη δυνατότητα χρήσης federated προσεγγίσεων σε ιατρικά προβλήματα όπου η συγκέντρωση δεδομένων δεν είναι πάντα επιθυμητή ή εφικτή.

Το τελικό συμπέρασμα της εργασίας είναι ότι η πρόβλεψη εγκεφαλικού επεισοδίου σε ανισόρροπα δεδομένα απαιτεί συνδυασμό κατάλληλου μοντέλου, σωστής τεχνικής εξισορρόπησης και προσεκτικής επιλογής των μετρικών αξιολόγησης. Η Logistic Regression με SMOTE ή class weights παρουσίασε ισχυρή συμπεριφορά στο centralized περιβάλλον, ενώ το FedAvg με Local SMOTE έδειξε ότι παρόμοια απόδοση μπορεί να επιτευχθεί και σε federated περιβάλλον. Επομένως, η εργασία αναδεικνύει τόσο τη σημασία της αντιμετώπισης της ανισορροπίας όσο και τη δυνατότητα εφαρμογής κατανεμημένης εκπαίδευσης σε προβλήματα πρόβλεψης υγείας.

7.2 Περιορισμοί της Έρευνας

Παρότι τα αποτελέσματα της έρευνας που υλοποιήθηκε, ήταν χρήσιμα και ανέδειξαν σημαντικά συμπεράσματα σχετικά με την αντιμετώπιση της ανισορροπίας και τη χρήση Federated Learning, υπάρχουν ορισμένοι περιορισμοί που πρέπει να ληφθούν υπόψη. Οι περιορισμοί αυτοί δεν αναιρούν την αξία των πειραμάτων, αλλά καθορίζουν το πλαίσιο μέσα στο οποίο πρέπει να ερμηνευθούν τα αποτελέσματα.

Ο πρώτος και σημαντικότερος περιορισμός αφορά το ίδιο το dataset. Το Stroke Prediction Dataset είναι σχετικά μικρό, ιδιαίτερα ως προς τον αριθμό των θετικών περιστατικών. Παρότι το συνολικό dataset περιλαμβάνει 5.110 εγγραφές, τα περιστατικά εγκεφαλικού είναι πολύ λιγότερα σε σχέση με τα δείγματα χωρίς εγκεφαλικό. Μετά τον διαχωρισμό σε training και test set, στο training set υπήρχαν 199 δείγματα της θετικής κλάσης, ενώ στο test set υπήρχαν μόνο 50 πραγματικά stroke cases. Αυτό σημαίνει ότι η αξιολόγηση των μοντέλων βασίστηκε σε μικρό αριθμό θετικών δειγμάτων.

Ο περιορισμένος αριθμός θετικών περιστατικών επηρεάζει άμεσα τη σταθερότητα των μετρικών αξιολόγησης. Για παράδειγμα, στο test set κάθε σωστή ή λανθασμένη πρόβλεψη ενός stroke case αλλάζει το Recall κατά 0.02, καθώς υπάρχουν μόνο 50 θετικά δείγματα. Αυτό σημαίνει ότι, μικρές διαφορές στον αριθμό των true positives μπορούν να εμφανίζονται ως σχετικά σημαντικές διαφορές στις μετρικές. Για τον λόγο αυτό, τα αποτελέσματα πρέπει να ερμηνευθούν προσεκτικά και όχι ως απόλυτη απόδειξη υπεροχής μιας τεχνικής έναντι μιας άλλης.

Ένας δεύτερος περιορισμός αφορά τη γενικευσιμότητα των αποτελεσμάτων. Το dataset που χρησιμοποιήθηκε είναι δημόσιο και δεν προέρχεται από πραγματική κλινική εφαρμογή στο πλαίσιο της εργασίας. Άρα, τα αποτελέσματα δείχνουν τη συμπεριφορά των μοντέλων πάνω στο συγκεκριμένο dataset, αλλά δεν μπορούν να θεωρηθούν άμεσα γενικεύσιμα σε όλους τους πραγματικούς πληθυσμούς ασθενών. Σε πραγματικές συνθήκες, τα δεδομένα μπορεί να περιέχουν περισσότερα χαρακτηριστικά, διαφορετική ποιότητα καταγραφής, ελλειπείς τιμές, θόρυβο, διαφορετικές ηλικιακές ομάδες και γενικά μεγαλύτερη ετερογένεια.

Επιπλέον, τα χαρακτηριστικά του dataset είναι σχετικά περιορισμένα. Το μοντέλο βασίστηκε σε διαθέσιμες μεταβλητές όπως ηλικία, υπέρταση, καρδιακή νόσος, μέσο επίπεδο γλυκόζης, BMI, κάπνισμα και ορισμένα δημογραφικά χαρακτηριστικά. Παρότι αυτά τα χαρακτηριστικά σχετίζονται με τον κίνδυνο εγκεφαλικού, δεν αποτυπώνουν πλήρως την πολυπλοκότητα της κλινικής πραγματικότητας. Σε ένα πραγματικό σύστημα πρόβλεψης θα μπορούσαν να χρησιμοποιηθούν περισσότερα ιατρικά δεδομένα, όπως ιστορικό φαρμακευτικής αγωγής, εργαστηριακές εξετάσεις, απεικονιστικά ευρήματα, οικογενειακό ιστορικό και χρονική εξέλιξη της κατάστασης του ασθενούς.

Ένας ακόμη περιορισμός αφορά τη χρήση συνθετικών δεδομένων. Οι τεχνικές SMOTE και CTGAN χρησιμοποιήθηκαν για την ενίσχυση της μειοψηφικής κλάσης, όμως τα συνθετικά δείγματα δεν είναι πραγματικά περιστατικά ασθενών. Ειδικά σε ιατρικά προβλήματα, η δημιουργία συνθετικών δεδομένων πρέπει να αντιμετωπίζεται με προσοχή. Παρότι το SMOTE βελτίωσε σημαντικά την απόδοση των μοντέλων, τα συνθετικά δείγματα που δημιουργεί βασίζονται σε μαθηματική παρεμβολή μεταξύ υπαρχόντων minority samples. Αυτό σημαίνει ότι μπορεί να ενισχύσει χρήσιμα πρότυπα, αλλά μπορεί και να δημιουργήσει δείγματα που δεν αντιστοιχούν πλήρως σε πραγματικές κλινικές περιπτώσεις.

Αντίστοιχα, το CTGAN αποτελεί πιο σύνθετη generative τεχνική, αλλά και αυτό έχει περιορισμούς. Αναλυτικότερα, το CTGAN εκπαιδεύτηκε για 30 εποχές, κυρίως λόγω χρόνου εκτέλεσης και υπολογιστικού κόστους. Ο αριθμός αυτός ενδέχεται να μην ήταν αρκετός για να μάθει πλήρως την κατανομή της θετικής κλάσης. Από την άλλη πλευρά, η αύξηση των εποχών δεν θα εξασφάλιζε απαραίτητα καλύτερα αποτελέσματα, καθώς ο μικρός αριθμός πραγματικών stroke περιστατικών θα

μπορούσε να οδηγήσει σε αστάθεια ή υπερπροσαρμογή. Επομένως, τα αποτελέσματα του CTGAN πρέπει να ερμηνευθούν με βάση αυτούς τους πρακτικούς περιορισμούς.

Περιορισμός υπάρχει επίσης ως προς την επιλογή και ρύθμιση των μοντέλων. Στην εργασία εξετάστηκαν τρία βασικά μοντέλα στο centralized περιβάλλον, η Logistic Regression, το Random Forest και το XGBoost. Παρότι τα μοντέλα αυτά καλύπτουν διαφορετικές κατηγορίες αλγορίθμων, δεν πραγματοποιήθηκε εξαντλητική αναζήτηση υπερπαραμέτρων. Η εργασία εστίασε περισσότερο στη σύγκριση τεχνικών αντιμετώπισης της ανισορροπίας και λιγότερο στη βελτιστοποίηση κάθε μοντέλου. Κατά συνέπεια, είναι πιθανό ότι με πιο εκτεταμένο hyperparameter tuning ορισμένα μοντέλα, όπως το Random Forest ή το XGBoost, θα μπορούσαν να παρουσιάσουν διαφορετική ή βελτιωμένη συμπεριφορά.

Παράλληλα, ένα σημαντικό εμπόδιο αφορά και το Federated Learning. Η federated προσέγγιση που χρησιμοποιήθηκε ήταν προσομοίωση και όχι πραγματική εφαρμογή σε ανεξάρτητα νοσοκομεία ή οργανισμούς. Οι τρεις clients δημιουργήθηκαν τεχνητά από το ίδιο αρχικό training set. Αυτό επέτρεψε τον ελεγχόμενο πειραματισμό με IID και Non-IID κατανομές, αλλά δεν αναπαριστά πλήρως τις συνθήκες ενός πραγματικού federated περιβάλλοντος. Σε πραγματικά νοσοκομεία, οι διαφορές μεταξύ των τοπικών datasets μπορεί να είναι μεγαλύτερες και πιο σύνθετες, τόσο ως προς την κατανομή των κλάσεων όσο και ως προς τα χαρακτηριστικά των ασθενών.

Επιπλέον, στο federated κομμάτι χρησιμοποιήθηκε μόνο Logistic Regression. Η συγκεκριμένη επιλογή έγινε γιατί η Logistic Regression είναι απλή, ερμηνεύσιμη και κατάλληλη για υλοποίηση FedAvg, καθώς τα βάρη της μπορούν να συνδυαστούν άμεσα μέσω μέσου όρου. Ωστόσο, αυτό σημαίνει ότι δεν εξετάστηκε federated εκδοχή πιο σύνθετων μοντέλων, όπως Random Forest ή XGBoost. Αυτό σημαίνει ότι, τα συμπεράσματα του Federated Learning αφορούν κυρίως τη συμπεριφορά της Logistic Regression σε federated περιβάλλον και όχι κάθε πιθανό μοντέλο μηχανικής μάθησης.

Σοβαρό περιορισμό αποτελεί επίσης το γεγονός ότι, δεν εφαρμόστηκαν πρόσθετοι μηχανισμοί προστασίας ιδιωτικότητας, όπως differential privacy ή secure aggregation. Στην προσομοίωση, τα δεδομένα παρέμειναν τοπικά στους clients και ο server συνδύαζε μόνο τα βάρη των μοντέλων. Ωστόσο, σε ένα πραγματικό privacy-preserving federated σύστημα, θα έπρεπε να ληφθούν επιπλέον μέτρα ώστε να μειωθεί ο κίνδυνος διαρροής πληροφορίας από τις παραμέτρους του μοντέλου. Συνεπώς, η παρούσα υλοποίηση αποτυπώνει τη βασική λογική του Federated Learning, αλλά δεν καλύπτει πλήρως όλες τις απαιτήσεις ασφάλειας και ιδιωτικότητας ενός πραγματικού ιατρικού συστήματος.

Επιπλέον, η Non-IID κατανομή που χρησιμοποιήθηκε δημιουργήθηκε τεχνητά με βάση την αναλογία των θετικών δειγμάτων ανά client. Παρότι αυτό βοήθησε να προσομοιωθεί μια πιο απαιτητική κατάσταση, η ανομοιογένεια περιορίστηκε κυρίως στην κατανομή της κλάσης stroke. Σε πραγματικές εφαρμογές, το Non-IID φαινόμενο μπορεί να είναι πιο σύνθετο. Οι clients μπορεί να διαφέρουν όχι μόνο ως προς το ποσοστό των θετικών περιστατικών, αλλά και ως προς την ηλικιακή κατανομή, τον τρόπο καταγραφής των δεδομένων και την ποιότητα των μετρήσεων. Επομένως, η προσομοίωση Non-IID της εργασίας αποτελεί απλοποιημένη εκδοχή ενός πραγματικού σεναρίου.

Τέλος, η εργασία δεν έχει χαρακτήρα κλινικής εφαρμογής. Τα μοντέλα που αναπτύχθηκαν αξιολογήθηκαν πειραματικά με βάση μετρικές μηχανικής μάθησης, αλλά δεν μπορούν να χρησιμοποιηθούν απευθείας για ιατρική διάγνωση ή λήψη αποφάσεων. Η πρόβλεψη εγκεφαλικού επεισοδίου είναι ένα σύνθετο ιατρικό πρόβλημα και οποιοδήποτε υπολογιστικό μοντέλο θα έπρεπε να αξιολογηθεί περαιτέρω από ειδικούς, να δοκιμαστεί σε ανεξάρτητα κλινικά datasets και να ενταχθεί σε κατάλληλο πλαίσιο ιατρικής αξιολόγησης πριν θεωρηθεί πρακτικά χρήσιμο.

Συνολικά, οι παραπάνω περιορισμοί δείχνουν ότι τα αποτελέσματα της εργασίας πρέπει να ερμηνευθούν ως πειραματική αξιολόγηση τεχνικών μηχανικής μάθησης και Federated Learning πάνω σε δημόσιο, ανισόρροπο dataset. Η εργασία ανέδειξε χρήσιμες τάσεις και συγκρίσεις, αλλά δεν αποτελεί τελική λύση για το πρόβλημα της πρόβλεψης εγκεφαλικού. Παρόλα αυτά, τα ευρήματα προσφέρουν μια σαφή βάση για μελλοντική έρευνα, ιδιαίτερα προς την κατεύθυνση της χρήσης πιο πλούσιων ιατρικών δεδομένων, πραγματικών federated σεναρίων και πιο προηγμένων μηχανισμών προστασίας ιδιωτικότητας.

7.3 Προτάσεις για Μελλοντική Επέκταση

Με βάση τα αποτελέσματα και τους περιορισμούς της εργασίας, υπάρχουν αρκετές κατευθύνσεις στις οποίες θα μπορούσε να επεκταθεί η έρευνα στο μέλλον. Η εργασία έδειξε ότι οι τεχνικές αντιμετώπισης της ανισορροπίας, όπως τα class weights και το SMOTE, μπορούν να βελτιώσουν σημαντικά την ανίχνευση της θετικής κλάσης. Παράλληλα, η προσομοίωση Federated Learning έδειξε ότι η καταναεμημένη εκπαίδευση μπορεί να προσεγγίσει την απόδοση της centralized προσέγγισης, ειδικά όταν συνδυάζεται με τοπική εξισορρόπηση των δεδομένων. Ωστόσο, υπάρχουν αρκετά σημεία που θα μπορούσαν να εξεταστούν πιο αναλυτικά σε μελλοντική εργασία.

Μία πρώτη κατεύθυνση αφορά τη χρήση μεγαλύτερων και πιο αντιπροσωπευτικών datasets. Το dataset που χρησιμοποιήθηκε ήταν χρήσιμο για πειραματική αξιολόγηση, όμως περιείχε περιορισμένο αριθμό θετικών περιστατικών εγκεφαλικού. Σε μελλοντική έρευνα θα ήταν ιδιαίτερα χρήσιμο να χρησιμοποιηθούν μεγαλύτερα ιατρικά datasets, με περισσότερα πραγματικά περιστατικά εγκεφαλικού και περισσότερα χαρακτηριστικά ανά ασθενή. Μέσω αυτής της διαδικασίας, τα μοντέλα θα μπορούσαν να εκπαιδευτούν σε πιο πλούσια πληροφορία και τα αποτελέσματα θα ήταν πιο σταθερά και πιο αξιόπιστα.

Επιπλέον, θα μπορούσε να εξεταστεί η χρήση δεδομένων από περισσότερες από μία πηγές. Στην παρούσα εργασία, το Federated Learning προσομοιώθηκε με τεχνητό διαχωρισμό του ίδιου training set σε τρεις clients. Σε ένα μελλοντικό στάδιο, θα είχε ιδιαίτερο ενδιαφέρον να αξιολογηθεί η ίδια μεθοδολογία σε πραγματικά δεδομένα από διαφορετικά νοσοκομεία ή φορείς υγείας. Ένα τέτοιο σενάριο θα ήταν πιο κοντά στην πραγματική χρήση του Federated Learning, καθώς κάθε φορέας θα διέθετε πραγματικά ανεξάρτητο τοπικό dataset, με διαφορετικά χαρακτηριστικά πληθυσμού και πιθανώς διαφορετική ποιότητα καταγραφής.

Μία δεύτερη πιθανή επέκταση είναι η διερεύνηση πιο σύνθετων Non-IID σεναρίων. Ειδικότερα, το Non-IID σενάριο δημιουργήθηκε κυρίως με άνιση κατανομή των θετικών δειγμάτων στους clients. Ωστόσο, στην πράξη η ανομοιογένεια των δεδομένων μπορεί να εμφανιστεί με πολλούς διαφορετικούς τρόπους. Για παράδειγμα, οι clients μπορεί να διαφέρουν ως προς την ηλικιακή κατανομή, το φύλο, τα επίπεδα γλυκόζης, τη συχνότητα υπέρτασης ή άλλους παράγοντες κινδύνου. Κατά συνέπεια, σε μελλοντική έρευνα θα μπορούσαν να σχεδιαστούν πιο σύνθετα Non-IID σενάρια, ώστε να εξεταστεί πιο αναλυτικά η ανθεκτικότητα του FedAvg σε ανομοιογενή ιατρικά δεδομένα.

Μία ακόμη κατεύθυνση αφορά τη χρήση περισσότερων clients. Εν προκειμένω, χρησιμοποιήθηκαν τρεις clients, ώστε η προσομοίωση να παραμείνει απλή, καθαρή και ερμηνεύσιμη. Μελλοντικά, θα μπορούσε να εξεταστεί η επίδραση μεγαλύτερου αριθμού clients, όπως πέντε, δέκα ή περισσότερων. Αυτό θα επέτρεπε να μελετηθεί αν η απόδοση του federated μοντέλου παραμένει σταθερή όταν το training set διαμοιράζεται σε περισσότερα και μικρότερα τοπικά σύνολα. Παράλληλα, θα μπορούσε να εξεταστεί και η περίπτωση όπου δεν συμμετέχουν όλοι οι clients σε κάθε γύρο εκπαίδευσης, αλλά

επιλέγεται κάθε φορά ένα υποσύνολο clients, όπως συμβαίνει συχνά σε πιο ρεαλιστικά federated περιβάλλοντα.

Σημαντική μελλοντική επέκταση θα ήταν επίσης η ενσωμάτωση πρόσθετων μηχανισμών ιδιωτικότητας και ασφάλειας καθώς υλοποιήθηκε η βασική λογική του Federated Learning, όπου τα δεδομένα παραμένουν τοπικά στους clients και ο server συνδυάζει μόνο τα βάρη των μοντέλων. Ωστόσο, σε πραγματικές εφαρμογές υγείας αυτό δεν είναι πάντα αρκετό. Οι παράμετροι ενός μοντέλου μπορεί, σε ορισμένες περιπτώσεις, να περιέχουν έμμεση πληροφορία για τα δεδομένα εκπαίδευσης. Κατά συνέπεια, θα μπορούσαν να ενσωματωθούν τεχνικές όπως differential privacy και secure aggregation σε δεύτερο χρόνο, ώστε να ενισχυθεί περισσότερο η προστασία της ιδιωτικότητας.

Μία ακόμη ενδιαφέρουσα επέκταση αφορά τη χρήση διαφορετικών federated αλγορίθμων πέρα από το FedAvg. Ο FedAvg αποτελεί βασική και απλή μέθοδο, όμως δεν είναι πάντα η πιο αποδοτική σε έντονα Non-IID δεδομένα. Μελλοντικά θα μπορούσαν να εξεταστούν πιο προχωρημένες προσεγγίσεις, όπως FedProx ή άλλες παραλλαγές federated optimization, οι οποίες έχουν σχεδιαστεί ώστε να αντιμετωπίζουν καλύτερα την ανομοιογένεια μεταξύ των clients. Η σύγκριση αυτών των μεθόδων με τον FedAvg θα μπορούσε να δείξει αν υπάρχουν πιο κατάλληλοι αλγόριθμοι για προβλήματα υγείας με ανισόρροπα και κατανεμημένα δεδομένα.

Επιπλέον, στο federated μέρος της παρούσας εργασίας χρησιμοποιήθηκε μόνο Logistic Regression. Η συγκεκριμένη επιλογή, ήταν σκόπιμη, επειδή το μοντέλο είναι απλό, ερμηνεύσιμο και συμβατό με τον μηχανισμό FedAvg. Παρ' όλα αυτά, σε μελλοντική εργασία θα μπορούσαν να εξεταστούν και πιο σύνθετα μοντέλα σε federated περιβάλλον. Για παράδειγμα, θα μπορούσαν να υλοποιηθούν νευρωνικά δίκτυα με περισσότερα επίπεδα ή άλλες προσεγγίσεις που μπορούν να εκπαιδευτούν με gradient-based optimization. Έτσι, θα ήταν δυνατό να μελετηθεί αν πιο σύνθετα μοντέλα μπορούν να αξιοποιήσουν καλύτερα τα κατανεμημένα δεδομένα χωρίς να θυσιάζεται η δυνατότητα federated εκπαίδευσης.

Μια ακόμα πιθανή επέκταση αφορά τη βελτιστοποίηση των υπερπαραμέτρων. Στην έρευνα που πραγματοποιήθηκε, χρησιμοποιήθηκαν συγκεκριμένες τιμές για τις παραμέτρους του SMOTE, ο αριθμός των γύρων επικοινωνίας, οι τοπικές εποχές εκπαίδευσής και ο ρυθμός εκμάθησης. Παρότι οι τιμές αυτές έδωσαν ικανοποιητικά αποτελέσματα, δεν πραγματοποιήθηκε εξαντλητική αναζήτηση υπερπαραμέτρων. Μελλοντικά, θα μπορούσε να γίνει πιο συστηματική διερεύνηση διαφορετικών τιμών, ώστε να εξεταστεί αν η απόδοση μπορεί να βελτιωθεί περαιτέρω. Για παράδειγμα, θα μπορούσε να μελετηθεί η επίδραση διαφορετικών τιμών του SMOTE ratio, διαφορετικού αριθμού rounds ή διαφορετικού αριθμού εποχών.

Ιδιαίτερο ενδιαφέρον θα είχε επίσης η περαιτέρω μελέτη του CTGAN και γενικότερα των generative μοντέλων για tabular ιατρικά δεδομένα. Αξίζει να σημειωθεί ότι, τα αποτελέσματα έδειξαν πως το CTGAN δεν κατάφερε να ξεπεράσει τις SMOTE-based τεχνικές. Ωστόσο, αυτό δεν σημαίνει ότι οι generative μέθοδοι δεν είναι χρήσιμες γενικά. Σε μεγαλύτερα datasets, με περισσότερα θετικά περιστατικά και καλύτερη δυνατότητα εκπαίδευσης, το CTGAN ή άλλες generative προσεγγίσεις θα μπορούσαν να αποδώσουν καλύτερα. Μελλοντική έρευνα θα μπορούσε να εξετάσει μεγαλύτερο αριθμό εποχών εκπαίδευσης, διαφορετικές ρυθμίσεις του μοντέλου ή εναλλακτικές generative τεχνικές για τη δημιουργία πιο ρεαλιστικών συνθετικών δειγμάτων της μειοψηφικής κλάσης.

Μία ακόμη χρήσιμη κατεύθυνση θα ήταν η αξιολόγηση διαφορετικών thresholds ταξινόμησης. Στο πλαίσιο της έρευνας αυτής, οι προβλέψεις βασίστηκαν στο κλασικό threshold 0.5. Ωστόσο, σε ιατρικά προβλήματα αυτό το threshold δεν είναι πάντα το καταλληλότερο. Αν ο στόχος είναι η αύξηση του Recall και η μείωση των false negatives, μπορεί να χρειάζεται χαμηλότερο threshold. Από την άλλη

πλευρά, αν ο στόχος είναι ο περιορισμός των false positives, μπορεί να χρειάζεται υψηλότερο threshold. Επομένως, μελλοντική εργασία θα μπορούσε να εξετάσει συστηματικά τη ρύθμιση του threshold με βάση τις ανάγκες του προβλήματος και το αποδεκτό trade-off μεταξύ Recall και Precision.

Επιπλέον, θα μπορούσε να προστεθεί ερμηνευσιμότητα των μοντέλων. Σε ιατρικά προβλήματα, δεν αρκεί ένα μοντέλο να δίνει μια πρόβλεψη. Είναι επίσης σημαντικό να μπορεί να εξηγηθεί ποια χαρακτηριστικά επηρέασαν περισσότερο την απόφαση του μοντέλου. Μελλοντικά θα μπορούσαν να χρησιμοποιηθούν τεχνικές explainable AI, όπως feature importance, SHAP values ή άλλες μέθοδοι ερμηνείας, ώστε να εξεταστεί ποιοι παράγοντες συμβάλλουν περισσότερο στην πρόβλεψη εγκεφαλικού. Αυτό θα έκανε τα αποτελέσματα πιο κατανοητά και πιο χρήσιμα σε ένα πιθανό ιατρικό περιβάλλον.

Τέλος, μία σημαντική μελλοντική κατεύθυνση είναι η αξιολόγηση της μεθοδολογίας σε πραγματικό κλινικό πλαίσιο. Η συγκεκριμένη εργασία αποτελεί πειραματική μελέτη μηχανικής μάθησης και δεν έχει χαρακτήρα κλινικής εφαρμογής. Για να χρησιμοποιηθεί ένα τέτοιο σύστημα στην πράξη, θα χρειαζόταν αξιολόγηση από ειδικούς υγείας, δοκιμή σε ανεξάρτητα datasets και έλεγχος της αξιοπιστίας του σε πραγματικές συνθήκες. Ένα μοντέλο πρόβλεψης εγκεφαλικού θα μπορούσε ενδεχομένως να λειτουργήσει υποστηρικτικά, ως εργαλείο εκτίμησης κινδύνου, αλλά όχι ως αντικατάσταση της ιατρικής κρίσης.

Συνολικά, η μελλοντική επέκταση της εργασίας μπορεί να κινηθεί σε τρεις βασικούς άξονες, αρχικά στη χρήση πιο πλούσιων και πραγματικών ιατρικών δεδομένων, έπειτα στη βελτίωση της federated μεθοδολογίας και τέλος στην ενίσχυση της αξιοπιστίας και ερμηνευσιμότητας των μοντέλων. Η εργασία αυτή έδειξε ότι ο συνδυασμός τεχνικών εξισορρόπησης και Federated Learning μπορεί να δώσει υποσχόμενα αποτελέσματα στην πρόβλεψη εγκεφαλικού επεισοδίου. Μελλοντικές εργασίες μπορούν να βασιστούν σε αυτά τα ευρήματα και να εξετάσουν πιο ρεαλιστικά, ασφαλή και κλινικά χρήσιμα συστήματα πρόβλεψης.

BIBΛΙΟΓΡΑΦΙΑ

- [1] S. K. Satapathy, A. Patel, P. Yadav, Y. Thacker, D. Vaniya and D. Parmar, "Machine Learning Approach for Estimation and Novel Design of Stroke Disease Predictions using Numerical and Categorical Features," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-6, doi: 10.1109/ICONAT57137.2023.10080722.
- [2] T. N. Deepthi, S. Sharmila, M. Swarna, M. Gouthami and C. L. Akshaya, "Prediction of Brain Stroke in Human Beings using Machine Learning," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 1314-1320, doi: 10.1109/ICEARS56392.2023.10085128.
- [3] World, "Stroke," *Who.int*, Dec. 19, 2025.
- [4] P. Premisha, S. Prasanth, M. Kanagarathnam and K. Banujan, "An Ensemble Machine Learning Approach for Stroke Prediction," 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE), Colombo, Sri Lanka, 2022, pp. 165-170, doi: 10.1109/SCSE56529.2022.9905215.
- [5] S. Mushtaq and K. S. Saini, "A Review on Predicting Brain Stroke using Machine Learning," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 667-673.
- [6] T. Aref, "Predicting Stroke Risk Using SMOTE and ROS Machine Learning Techniques," 2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Male, Maldives, 2024, pp. 1-5, doi: 10.1109/ICECCME62383.2024.10796090.
- [7] R. Kumar, J. P. Gujjar, M. S. G. Prasad, R. M. Devadas, Preethi and U. Kumar, "Balancing the Imbalanced Datasets for Machine Learning Classification Problems," 2025 *International Conference on Next Generation Communication & Information Processing (INCIP)*, Bangalore, India, 2025, pp. 1-4, doi: 10.1109/INCIP64058.2025.11019446.
- [8] A. A. El-Sayed, M. A. M. Mahmood, N. A. Meguid and H. A. Hefny, "Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE)," 2015 Third World Conference on Complex Systems (WCCS), Marrakech, 2015, pp. 1-5, doi: 10.1109/ICoCS.2015.7483267.
- [9] I. -M. Chiu, W. -H. Zeng and C. -H. R. Lin, "Using multiclass machine learning model to improve outcome prediction of acute ischemic stroke patients after reperfusion therapy," 2020 International Computer Symposium (ICS), Tainan, Taiwan, 2020, pp. 225-231, doi: 10.1109/ICS51289.2020.00053.
- [10] B. Jonathan, P. H. Putra and Y. Ruldeviyani, "Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek," 2020 *IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, Bali, Indonesia, 2020, pp. 81-85, doi: 10.1109/IAICT50021.2020.9172033.
- [11] H. S. Khan and J. Panda, "Enhancing Medical Research with Synthetic Data Generation via CTGAN on Cleveland Heart Dataset," 2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), Coimbatore, India, 2025, pp. 1-5, doi: 10.1109/ICC-ROBINS64345.2025.11086164.

- [12] S. Sangve, K. Oswal, T. Sarade, O. Tongare, N. Sangode and Q. Shergardwala, "Synthetic Data Generation for Enhanced Model Efficiency: A GAN- Based Approach to Tabular Data," 2025 International Conference on Computing Technologies & Data Communication (ICCTDC), HASSAN, India, 2025, pp. 1-7, doi: 10.1109/ICCTDC64446.2025.11157963.
- [13] M. A. Moh'd Al-Oudat and E. Tawfiq Abusamra, "Stroke Prediction with Explainable AI and PR-AUC-Driven Model Selection," 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA), Amman, Jordan, 2026, pp. 1-6, doi: 10.1109/ICCIAA68481.2026.11543849.
- [14] D. L. Pereira Gomes, A. Grégio, M. A. Zanata Alves and P. R. Lisboa de Almeida, "Efficient Prequential AUC-PR Computation," 2023 International Conference on Machine Learning and Applications (ICMLA), Jacksonville, FL, USA, 2023, pp. 2222-2227, doi: 10.1109/ICMLA58977.2023.00335.
- [15] J. Zhang, "A Study of Heart Disease Prediction Based on Logistic Regression and CART Classification," 2024 5th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Shenzhen, China, 2024, pp. 720-724, doi: 10.1109/ICBAIE63306.2024.11117142.
- [16] S. Dhanka and S. Maini, "Random Forest for Heart Disease Detection: A Classification Approach," 2021 IEEE 2nd International Conference On Electrical Power and Energy Systems (ICEPES), Bhopal, India, 2021, pp. 1-3, doi: 10.1109/ICEPES52894.2021.9699506.
- [17] Y. Zhou, X. Song and M. Zhou, "Supply Chain Fraud Prediction Based On XGBoost Method," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 2021, pp. 539-542, doi: 10.1109/ICBAIE52039.2021.9389949.
- [18] J. Lin, J. Chen, J. Xiong, D. Jiao, W. Zhao and Y. Xiang, "A Lightweight Privacy-Preserving Federated Learning Framework for Heterogeneity-Resilient Skin Cancer Diagnosis," in *IEEE Journal of Biomedical and Health Informatics*, vol. 30, no. 8, pp. 6378-6390, Aug. 2026, doi: 10.1109/JBHI.2025.3578125.
- [19] H. Lee, M. Jiang and Q. Zhao, "FedAssist: Federated Learning in AI-Powered Prosthetics for Sustainable and Collaborative Learning," 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 2024, pp. 1-5, doi: 10.1109/EMBC53108.2024.10781961.
- [20] N. Y. Kim, H. U. Rashid and S. H. Jeong, "Scaled Entropy Guided FedAvg: Improving Federated Learning on Imbalanced Data," 2025 International Conference on Mobile, Military, Maritime IT Convergence (ICMIC), Cebu, Philippines, 2025, pp. 270-273, doi: 10.1109/ICMIC66299.2025.11257820.
- [21] O. Segura et al., "Enhancing Stroke Severity Evaluation: A Machine Learning Approach to Mortality Prediction Versus Traditional Scales," 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Copenhagen, Denmark, 2025, pp. 1-5, doi: 10.1109/EMBC58623.2025.11251883.
- [22] R. Das, S. K. Biswas, D. Devi and B. Sarma, "An Oversampling Technique by Integrating Reverse Nearest Neighbor in SMOTE: Reverse-SMOTE," 2020 International Conference on Smart Electronics

and Communication (ICOSEC), Trichy, India, 2020, pp. 1239-1244, doi: 10.1109/ICOSEC49089.2020.9215387.

[23] N. Thakur and N. Gupta, "Machine Learning-Driven Stroke Prediction: Improving Patient Outcomes Through Early Diagnosis," 2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI), Erode, India, 2025, pp. 1199-1203, doi: 10.1109/ICMSCI62561.2025.10894443.

[24] K. Sundaram, L. B. K. K. and A. K. Ramamoorthy, "Enhanced Brain Stroke Prediction: An Ensemble of Random Forest, Logistic Regression and XGBoost," 2024 International Conference on Emerging Research in Computational Science (ICERCS), Coimbatore, India, 2024, pp. 1-5, doi: 10.1109/ICERCS63125.2024.10895046.

[25] R. Suryawanshi, V. Kulkarni, P. Ghule, K. Patil, H. Patil and Y. Manala, "Brain Stroke Prediction Using Logistic Regression with Logarithmic Transform," 2024 4th International Conference on Sustainable Expert Systems (ICSES), Kaski, Nepal, 2024, pp. 873-877, doi: 10.1109/ICSES63445.2024.10763034.

[26] S. Kumar, "Adaptive Logistic Regression: Leveraging Arctan in Active Learning," 2025 10th International Conference on Signal Processing and Communication (ICSC), Noida, India, 2025, pp. 608-612, doi: 10.1109/ICSC64553.2025.10968235.

[27] O. Elnakib, E. Shaaban, M. Mahmoud and K. Emara, "Evaluation of Centralized, Distributed and Federated Learning for IoT Intrusion Detection Systems," 2023 *Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, Cairo, Egypt, 2023, pp. 315-320, doi: 10.1109/ICICIS58388.2023.10391151.