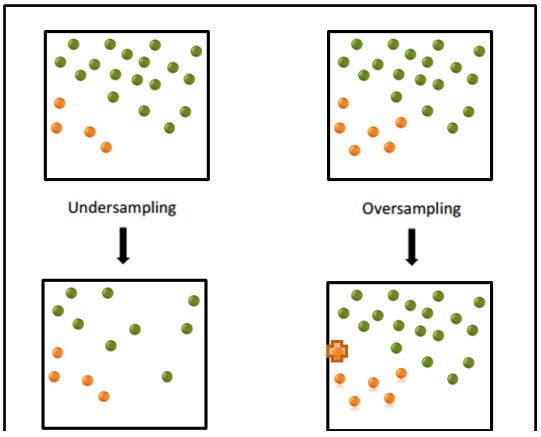
ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΑΝΙΣΟΚΑΤΑΝΟΜΗ ΜΕΣΩ ΤΕΧΝΙΚΩΝ ΥΠΕΡΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΚΑΙ ΥΠΟΔΕΙΓΜΑΤΟΛΗΨΙΑΣ»



Του φοιτητή:
ΔΗΜΗΤΡΙΟΣ ΛΑΪΝΑΣ

Αρ. Μητρώου: 09/2023

Επιβλέπων
Ονοματεπώνυμο: ΣΤΕΦΑΝΟΣ ΟΥΓΙΑΡΟΓΛΟΥ
ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ

ΤΙΤΛΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΑΝΙΣΟΚΑΤΑΝΟΜΗ ΜΕΣΩ ΤΕΧΝΙΚΩΝ ΥΠΕΡΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΚΑΙ ΥΠΟΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

Ονοματεπώνυμο φοιτητή :ΔΗΜΗΤΡΙΟΣ ΛΑΪΝΑΣ

Ονοματεπώνυμο εισηγητή :ΣΤΕΦΑΝΟΣ ΟΥΓΙΑΡΟΓΛΟΥ

Ημερομηνία ανάληψης Δ.Ε. 30/10/24

Ημερομηνία περάτωσης Δ.Ε. 14/02/26

Ημερομηνία:2/1/2026

Το σύγγραμμα αυτό είναι αφιερωμένο στην πολυαγαπημένη μου αδερφούλα.

Πρόλογος

Η επιλογή του θέματος της παρούσας διπλωματικής εργασίας προέκυψε από το ενδιαφέρον μου για τη μηχανική μάθηση και, ειδικότερα, για τα προβλήματα που εμφανίζονται όταν τα δεδομένα παρουσιάζουν έντονη ανισορροπία μεταξύ των κλάσεων. Η ανισοκατανομή των δεδομένων αποτελεί ένα συχνό φαινόμενο σε πλήθος πραγματικών εφαρμογών, όπως η ιατρική διάγνωση, η ανίχνευση απάτης και η κυβερνοασφάλεια, και επηρεάζει σημαντικά την αξιοπιστία των μοντέλων πρόβλεψης.

Μέσα από την εκπόνηση της εργασίας αυτής είχα την ευκαιρία να μελετήσω σε βάθος τόσο τις θεωρητικές αρχές της ανισοκατανεμημένης ταξινόμησης όσο και τις πρακτικές τεχνικές αντιμετώπισής της. Η πειραματική διαδικασία συνέβαλε ουσιαστικά στην εξοικείωσή μου με σύγχρονες μεθόδους δειγματοληψίας, την αξιολόγηση μοντέλων με κατάλληλες μετρικές και την υλοποίηση αναπαραγώγιμων πειραμάτων σε περιβάλλον Python.

Το βασικό όφελος που αποκόμισα από την παρούσα διπλωματική εργασία είναι η απόκτηση σφαιρικής κατανόησης των προκλήσεων που συνοδεύουν την ανάλυση ανισοκατανεμημένων δεδομένων, καθώς και η ενίσχυση των δεξιοτήτων μου στον σχεδιασμό και την κριτική αξιολόγηση πειραματικών μελετών στον χώρο της μηχανικής μάθησης.

Περίληψη

Η ανισορροπία κλάσεων αποτελεί ένα από τα σημαντικότερα προβλήματα στη μηχανική μάθηση και επηρεάζει αρνητικά την απόδοση των αλγορίθμων ταξινόμησης, ιδιαίτερα ως προς την αναγνώριση σπάνιων αλλά κρίσιμων περιπτώσεων. Στην παρούσα διπλωματική εργασία μελετάται συστηματικά η επίδραση τεχνικών υπερδειγματοληψίας (over-sampling) και υποδειγματοληψίας (under-sampling) στην απόδοση μοντέλων δυαδικής ταξινόμησης σε ανισοκατανεμημένα σύνολα δεδομένων.

Η πειραματική αξιολόγηση πραγματοποιήθηκε σε δεκατέσσερα σύνολα δεδομένων από το KEELDatasetRepository, τα οποία χαρακτηρίζονται από διαφορετικό τύπο χαρακτηριστικών (αριθμητικά, κατηγορικά και μικτά) και υψηλό δείκτη ανισορροπίας. Για την αντιμετώπιση της ανισορροπίας εφαρμόστηκαν διάφορες τεχνικές δειγματοληψίας, συμπεριλαμβανομένων παραλλαγών της μεθόδου SMOTE, καθώς και μεθόδων υποδειγματοληψίας. Ως κοινός ταξινομητής χρησιμοποιήθηκε ο αλγόριθμος k-NearestNeighbors, ώστε να διασφαλιστεί η δίκαιη σύγκριση των μεθόδων. Η αξιολόγηση της απόδοσης βασίστηκε σε μετρικές όπως η ακρίβεια (Accuracy), η ανάκληση (Recall), η ακρίβεια θετικών προβλέψεων (Precision) και ο δείκτης F1, με έμφαση στη μειοψηφική κλάση.

Τα αποτελέσματα δείχνουν ότι οι τεχνικές υπερδειγματοληψίας μπορούν να βελτιώσουν σημαντικά την απόδοση ως προς τη μειοψηφική κλάση, ιδιαίτερα σε περιπτώσεις όπου η αρχική απόδοση είναι χαμηλή, ενώ οι τεχνικές υποδειγματοληψίας εμφανίζουν μεγαλύτερη αστάθεια και απαιτούν προσεκτική εφαρμογή. Συμπερασματικά, η αποτελεσματική αντιμετώπιση της ανισορροπίας κλάσεων εξαρτάται από τον τύπο των δεδομένων, τον βαθμό ανισορροπίας και τους στόχους της εκάστοτε εφαρμογής.

«CLASSIFICATION OF IMBALANCED DATA USING OVERSAMPLING AND UNDERSAMPLING TECHNIQUES»

Name & Surname Graduate Student: DIMITRIOS LAINAS

Abstract

Class imbalance is one of the most significant challenges in machine learning and has a detrimental impact on the performance of classification algorithms, particularly with respect to the detection of rare but critical instances. This thesis systematically investigates the effect of over-sampling and under-sampling techniques on the performance of binary classification models applied to imbalanced datasets.

The experimental evaluation was conducted on fourteen datasets obtained from the KEEL Dataset Repository, which differ in terms of feature types (numerical, categorical, and mixed) and exhibit high imbalance ratios. To address class imbalance, a variety of sampling techniques were applied, including several variants of the SMOTE method as well as under-sampling approaches. A k-Nearest Neighbors classifier was employed as a common base model across all experiments in order to ensure a fair and consistent comparison of the sampling methods. Model performance was evaluated using standard classification metrics, namely Accuracy, Recall, Precision, and F1-score, with particular emphasis on the minority class.

The experimental results demonstrate that over-sampling techniques can significantly improve the performance with respect to the minority class, especially in cases where the baseline performance is low. In contrast, under-sampling techniques exhibit greater variability and may lead to performance degradation if applied without caution, particularly in datasets with extreme imbalance. Overall, the findings indicate that effective handling of class imbalance depends on multiple factors, including the type of features, the degree of imbalance, and the specific objectives of the application.

Ευχαριστίες

Πριν ξεκινήσω την παρουσίαση της εργασίας μου, κρίνω απαραίτητο να ευχαριστήσω θερμά όσους με βοήθησαν να ολοκληρώσω αυτή τη διπλωματική εργασία. Θα ήθελα να ευχαριστήσω ιδιαιτέρως τον επιβλέποντα Καθηγητή μου κ. Στέφανο Ουγιάρογλου για την εμπιστοσύνη, την ακούραστη επιστημονική του καθοδήγηση, τη συνεχή ενθάρρυνση και τις πολύτιμες συμβουλές του. Τον ευχαριστώ πολύ, για την υπομονή του, για την τιμή που μου έκανε να μου δώσει την ευκαιρία όλο αυτό το διάστημα που συνεργαστήκαμε να αποκτήσω γνώση και αγάπη για την έρευνα μέσω του ξεχωριστού διδακτικού του ταλέντου.

Περιεχόμενα

ΚατάλογοςΠινάκων

Συντομογραφίες

Δ.Ε.	Διπλωματική Εργασία
ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
Π.Ε.	ΠτυχιακήΕργασία
CNN	CondensedNearestNeighbors
ENN	Edited Nearest Neighbors
kNN	k Nearest Neighbors
MLP	Multi-Layer Perceptron
NN	NearestNeighbors
RENN	Repeated Edited Nearest Neighbors
SVM	Support Vector Machines

Κεφάλαιο 1ο: Εισαγωγή

1.1 Κατηγοριοποίηση δεδομένων

Η κατηγοριοποίηση δεδομένων (data classification) αποτελεί ένα από τα θεμελιώδη προβλήματα της εξόρυξης δεδομένων (datamining) και της μηχανικής μάθησης (machinelearning) και αναφέρεται στη διαδικασία ανάθεσης ενός στιγμιότυπου δεδομένων σε μία από ένα προκαθορισμένο σύνολο κλάσεων. Η διαδικασία αυτή βασίζεται στη μάθηση ενός μοντέλου από ιστορικά δεδομένα, των οποίων οι κλάσεις είναι γνωστές εκ των προτέρων, και στη συνέχεια στη χρήση του μοντέλου για την πρόβλεψη της κλάσης νέων, άγνωστων δεδομένων [1].

Η κατηγοριοποίηση εντάσσεται στην κατηγορία της επιβλεπόμενης μάθησης (supervised learning), όπου το σύνολο εκπαίδευσης αποτελείται από ζεύγη εισόδου–εξόδου, με την έξοδο να αντιστοιχεί στην ετικέτα κλάσης. Στόχος είναι η εκμάθηση μιας συνάρτησης που να προσεγγίζει όσο το δυνατόν καλύτερα τη σχέση μεταξύ των χαρακτηριστικών εισόδου και των αντίστοιχων κλάσεων, ελαχιστοποιώντας το σφάλμα πρόβλεψης σε άγνωστα δεδομένα [2].

Οι τεχνικές κατηγοριοποίησης χρησιμοποιούνται ευρέως σε πληθώρα εφαρμογών, όπως η ιατρική διάγνωση, η ανίχνευση απάτης, η αναγνώριση προτύπων, η ανάλυση βιολογικών δεδομένων, η ασφάλεια πληροφοριακών συστημάτων και η εξόρυξη γνώσης από μεγάλες βάσεις δεδομένων [3]. Η αποτελεσματικότητα ενός συστήματος κατηγοριοποίησης εξαρτάται σε μεγάλο βαθμό τόσο από την ποιότητα των δεδομένων όσο και από τα χαρακτηριστικά του εκάστοτε προβλήματος, όπως ο αριθμός των κλάσεων, ο τύπος των χαρακτηριστικών και η κατανομή των δεδομένων ανά κλάση.

Στη βιβλιογραφία έχουν προταθεί πολυάριθμοι αλγόριθμοι κατηγοριοποίησης, οι οποίοι διαφέρουν ως προς τη θεωρητική τους βάση, την υπολογιστική πολυπλοκότητα και τις υποθέσεις που κάνουν για τη δομή των δεδομένων. Ενδεικτικά παραδείγματα αποτελούν οι αλγόριθμοι k-Nearest Neighbors (kNN), τα δέντρα αποφάσεων, οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines), οι πιθανοτικοί ταξινομητές (probabilisticclassifiers) και τα νευρωνικά δίκτυα [4]. Κάθε αλγόριθμος παρουσιάζει πλεονεκτήματα και περιορισμούς, καθιστώντας την επιλογή του κατάλληλου ταξινομητή άρρηκτα συνδεδεμένη με τη φύση του προβλήματος.

Ένα από τα σημαντικότερα ζητήματα που επηρεάζουν την απόδοση των αλγορίθμων κατηγοριοποίησης είναι η κατανομή των κλάσεων στα δεδομένα εκπαίδευσης. Στην πράξη, πολλές εφαρμογές χαρακτηρίζονται από έντονη ανισοκατανομή μεταξύ των κλάσεων, όπου ο αριθμός των δειγμάτων της πλειοψηφικής κλάσης υπερτερεί σημαντικά έναντι της μειοψηφικής. Το φαινόμενο αυτό μπορεί να οδηγήσει σε μεροληπτικά μοντέλα, τα οποία επιτυγχάνουν υψηλή συνολική ακρίβεια αλλά αποτυγχάνουν να αναγνωρίσουν σωστά τα στιγμιότυπα της μειοψηφικής κλάσης [5]. Το ζήτημα της ανισοκατανομής των κλάσεων και οι επιπτώσεις του στην κατηγοριοποίηση δεδομένων αναλύονται εκτενέστερα στην επόμενη ενότητα.

1.2 Πρόβλημα της ανισοκατανομής των κλάσεων

Η ανισοκατανομή των κλάσεων (classimbalance) αποτελεί ένα από τα σημαντικότερα και πιο μελετημένα προβλήματα στη μηχανική μάθηση και την κατηγοριοποίηση δεδομένων. Ένα σύνολο δεδομένων χαρακτηρίζεται ως ανισοκατανεμημένο όταν ο αριθμός των στιγμιότυπων που ανήκουν στη μία κλάση υπερτερεί σημαντικά έναντι των στιγμιότυπων της άλλης ή των άλλων κλάσεων. Συνήθως, η κλάση με τα λιγότερα δείγματα αναφέρεται ως μειοψηφική κλάση, ενώ η κλάση με τα περισσότερα δείγματα ως πλειοψηφική [5 - 6].

Το πρόβλημα της ανισοκατανομής εμφανίζεται σε πληθώρα πραγματικών εφαρμογών, όπως η ανίχνευση απάτης σε οικονομικές συναλλαγές, η διάγνωση σπάνιων ασθενειών, η ανίχνευση εισβολών σε συστήματα υπολογιστών και η ανάλυση σφαλμάτων σε βιομηχανικά συστήματα. Σε τέτοιες περιπτώσεις, τα στιγμιότυπα της μειοψηφικής κλάσης είναι συχνά τα πλέον κρίσιμα, καθώς αντιστοιχούν σε ανεπιθύμητα ή επικίνδυνα γεγονότα, παρότι εμφανίζονται με πολύ μικρότερη συχνότητα [7].

Η ανισοκατανομή των κλάσεων επηρεάζει σημαντικά τη διαδικασία εκπαίδευσης των ταξινομητών. Οι περισσότεροι κλασικοί αλγόριθμοι κατηγοριοποίησης έχουν σχεδιαστεί με την υπόθεση ότι οι κλάσεις είναι περίπου ισοκατανεμημένες. Ως αποτέλεσμα, κατά την εκπαίδευση σε έντονα ανισοκατανεμημένα δεδομένα, οι αλγόριθμοι τείνουν να ευνοούν την πλειοψηφική κλάση, επιτυγχάνοντας υψηλή συνολική ακρίβεια, αλλά χαμηλή απόδοση ως προς τη μειοψηφική κλάση [6], [8].

Ένα χαρακτηριστικό πρόβλημα που προκύπτει είναι ότι ένας ταξινομητής μπορεί να επιτυγχάνει υψηλές τιμές accuracy ακόμη και όταν αποτυγχάνει πλήρως να αναγνωρίσει τα δείγματα της μειοψηφικής κλάσης. Για παράδειγμα, σε ένα σύνολο δεδομένων όπου το 99% των δειγμάτων ανήκουν στην πλειοψηφική κλάση, ένας ταξινομητής που προβλέπει πάντοτε την πλειοψηφική κλάση θα επιτύχει accuracy 99%, χωρίς ωστόσο να έχει πρακτική χρησιμότητα [9].

Για την ποσοτικοποίηση του βαθμού ανισοκατανομής χρησιμοποιείται συχνά ο δείκτης ανισορροπίας (imbalanceratio), ο οποίος ορίζεται ως ο λόγος του πλήθους των δειγμάτων της πλειοψηφικής κλάσης προς το πλήθος των δειγμάτων της μειοψηφικής. Όσο μεγαλύτερη είναι η τιμή του δείκτη αυτού, τόσο δυσκολότερη καθίσταται η εκμάθηση ενός αξιόπιστου μοντέλου για τη μειοψηφική κλάση [5].

Η παρουσία ανισοκατανομής δεν επηρεάζει μόνο την απόδοση των ταξινομητών, αλλά και την αξιολόγησή τους. Μετρικές όπως η accuracy ενδέχεται να δώσουν παραπλανητική εικόνα της πραγματικής απόδοσης, καθιστώντας αναγκαία τη χρήση μετρικών που λαμβάνουν υπόψη την απόδοση ανά κλάση, όπως η precision, η recall και το F1-score. Το ζήτημα αυτό καθιστά επιτακτική την ανάγκη ανάπτυξης και εφαρμογής ειδικών τεχνικών αντιμετώπισης της ανισοκατανομής, οι οποίες παρουσιάζονται στην επόμενη ενότητα.

1.3 Αντιμετώπιση του προβλήματος της ανισοκατανομής των κλάσεων

Η ανισοκατανομή των κλάσεων αποτελεί ένα διαδεδομένο και ιδιαίτερα απαιτητικό πρόβλημα στην κατηγοριοποίηση δεδομένων, γεγονός που έχει οδηγήσει στην ανάπτυξη ποικίλων τεχνικών για την αποτελεσματική αντιμετώπισή του. Οι τεχνικές αυτές στοχεύουν κυρίως στη βελτίωση της ικανότητας των ταξινομητών να αναγνωρίζουν σωστά τα στιγμιότυπα της μειοψηφικής κλάσης, χωρίς να θυσιάζεται σημαντικά η συνολική απόδοση του μοντέλου [9].

Μία από τις πιο διαδεδομένες προσεγγίσεις για την αντιμετώπιση της ανισοκατανομής είναι η παρέμβαση σε επίπεδο δεδομένων, μέσω της τροποποίησης της κατανομής των κλάσεων πριν από την εκπαίδευση του ταξινομητή. Στο πλαίσιο αυτό, οι τεχνικές διακρίνονται κυρίως σε μεθόδους υπερδειγματοληψίας (oversampling) και υποδειγματοληψίας (undersampling) [6], [10].

Η υπερδειγματοληψία αποσκοπεί στην αύξηση του πλήθους των δειγμάτων της μειοψηφικής κλάσης, με στόχο την εξισορρόπηση της κατανομής των δεδομένων. Αυτό μπορεί να επιτευχθεί είτε μέσω της απλής αναπαραγωγής υπαρχόντων δειγμάτων είτε μέσω της δημιουργίας νέων, συνθετικών στιγμιότυπων που προσεγγίζουν τη δομή της μειοψηφικής κλάσης στον χώρο των χαρακτηριστικών.

Με τον τρόπο αυτό, ο ταξινομητής εκτίθεται σε περισσότερα παραδείγματα της μειοψηφικής κλάσης κατά την εκπαίδευση, γεγονός που μπορεί να οδηγήσει σε βελτιωμένη απόδοση ως προς την αναγνώρισή της [11].

Αντίστοιχα, η υποδειγματοληψία στοχεύει στη μείωση του πλήθους των δειγμάτων της πλειοψηφικής κλάσης, αφαιρώντας επιλεγμένα στιγμιότυπα από το σύνολο εκπαίδευσης. Η διαδικασία αυτή έχει ως αποτέλεσμα την εξισορρόπηση της κατανομής των κλάσεων, μειώνοντας την κυριαρχία της πλειοψηφικής κλάσης κατά την εκπαίδευση του μοντέλου. Παρότι η υποδειγματοληψία μπορεί να οδηγήσει σε απώλεια πληροφορίας, παρουσιάζει το πλεονέκτημα της μείωσης του υπολογιστικού κόστους και της απλοποίησης της διαδικασίας εκπαίδευσης [6].

Η ύπαρξη πλήθους διαφορετικών τεχνικών υπερδειγματοληψίας και υποδειγματοληψίας, καθεμία με διαφορετικά χαρακτηριστικά, πλεονεκτήματα και περιορισμούς, καθιστά σαφές ότι δεν υπάρχει μία καθολικά βέλτιστη λύση για όλα τα προβλήματα ανισοκατανομής. Η απόδοση των τεχνικών αυτών εξαρτάται σε μεγάλο βαθμό από τη φύση των δεδομένων, τον βαθμό ανισοκατανομής και τον χρησιμοποιούμενο ταξινομητή [9]. Το γεγονός αυτό αποτελεί και το βασικό κίνητρο για τη συστηματική σύγκριση διαφορετικών τεχνικών oversampling και undersampling, όπως επιχειρείται στην παρούσα εργασία.

1.4 Κίνητρο

Παρά την εκτενή ερευνητική δραστηριότητα γύρω από το πρόβλημα της ανισοκατανομής των κλάσεων, η επιλογή της κατάλληλης τεχνικής αντιμετώπισής του παραμένει ένα ανοικτό και μη τετριμμένο ζήτημα. Στη βιβλιογραφία έχει προταθεί πλήθος μεθόδων υπερδειγματοληψίας και

υποδειγματοληψίας, καθεμία από τις οποίες βασίζεται σε διαφορετικές υποθέσεις και παρουσιάζει διακριτά πλεονεκτήματα και περιορισμούς [9], [10].

Ένα από τα βασικά προβλήματα που ανακύπτουν στην πράξη είναι ότι δεν υπάρχει μία καθολικά βέλτιστη τεχνική oversampling ή undersampling που να αποδίδει ικανοποιητικά σε όλα τα σύνολα δεδομένων και για όλους τους ταξινομητές. Η αποτελεσματικότητα των μεθόδων αυτών εξαρτάται σε μεγάλο βαθμό από παράγοντες όπως ο βαθμός ανισοκατανομής, ο τύπος των χαρακτηριστικών (αριθμητικά, κατηγορικά ή μικτά), η γεωμετρία των δεδομένων στον χώρο χαρακτηριστικών και ο χρησιμοποιούμενος αλγόριθμος ταξινόμησης [9], [12].

Επιπλέον, η βιβλιογραφία συχνά επικεντρώνεται στην αξιολόγηση μεμονωμένων τεχνικών ή σε περιορισμένο αριθμό συνόλων δεδομένων, γεγονός που δυσχεραίνει τη γενίκευση των συμπερασμάτων. Παράλληλα, παρατηρείται ότι διαφορετικές μελέτες καταλήγουν σε αντικρουόμενα αποτελέσματα αναφορικά με το αν οι τεχνικές υπερδειγματοληψίας υπερτερούν των τεχνικών υποδειγματοληψίας ή το αντίστροφο, ιδίως σε περιπτώσεις έντονης ανισοκατανομής [6], [13].

Το παραπάνω πλαίσιο αποτέλεσε το βασικό κίνητρο για την παρούσα εργασία. Συγκεκριμένα, στόχος είναι η συστηματική και συγκριτική αξιολόγηση διαφορετικών τεχνικών oversampling και undersampling σε πληθώρα ανισοκατανεμημένων συνόλων δεδομένων, με κοινό ταξινομητή και ενιαία μεθοδολογία αξιολόγησης. Μέσω της πειραματικής αυτής σύγκρισης επιδιώκεται η διερεύνηση του πώς επηρεάζουν οι διάφορες τεχνικές δειγματοληψίας την απόδοση ενός ταξινομητή ως προς τη μειοψηφική κλάση, καθώς και η ανάδειξη προτύπων συμπεριφοράς που μπορούν να φανούν χρήσιμα σε αντίστοιχα προβλήματα της πράξης.

1.5 Συνεισφορά

Η παρούσα εργασία εστιάζει στη συστηματική μελέτη και πειραματική σύγκριση τεχνικών υπερδειγματοληψίας και υποδειγματοληψίας στο πλαίσιο προβλημάτων δυαδικής κατηγοριοποίησης με έντονη ανισοκατανομή κλάσεων. Η κύρια συνεισφορά της εργασίας έγκειται στη διεξαγωγή ενός εκτεταμένου και συνεκτικού benchmarking, με κοινή πειραματική βάση, που επιτρέπει την αξιόπιστη σύγκριση της επίδοσης διαφορετικών τεχνικών δειγματοληψίας.

Συγκεκριμένα, χρησιμοποιούνται δεκατέσσερα (14) ανισοκατανεμημένα σύνολα δεδομένων, όλα με υψηλές τιμές δείκτη ανισορροπίας (Imbalance Ratio), τα οποία προέρχονται από το αποθετήριο KEEL Dataset Repository, μια ευρέως αναγνωρισμένη πηγή για πειραματική αξιολόγηση αλγορίθμων μηχανικής μάθησης σε προβλήματα ανισοκατανομής [14]. Η επιλογή συνόλων δεδομένων με διαφορετικά χαρακτηριστικά (αριθμητικά, κατηγορικά και μικτά) επιτρέπει τη διερεύνηση της συμπεριφοράς των τεχνικών oversampling και undersampling σε ποικίλα σενάρια.

Επιπλέον, η εργασία υλοποιεί και συγκρίνει πολλαπλές τεχνικές υπερδειγματοληψίας και υποδειγματοληψίας, οι οποίες εφαρμόζονται με ενιαία μεθοδολογία αξιολόγησης και με τη χρήση

κοινού ταξινομητή. Η επιλογή αυτή διασφαλίζει ότι οι παρατηρούμενες διαφορές στην απόδοση οφείλονται στις ίδιες τις τεχνικές δειγματοληψίας και όχι σε διαφοροποιήσεις του πειραματικού πλαισίου ή των παραμέτρων του μοντέλου.

Ιδιαίτερη έμφαση δίνεται στη σωστή πειραματική διαδικασία, με την εφαρμογή των τεχνικών δειγματοληψίας αποκλειστικά στο σύνολο εκπαίδευσης, αποφεύγοντας φαινόμενα διαρροής πληροφορίας (data leakage). Η αξιολόγηση πραγματοποιείται με τη χρήση καθιερωμένων μετρικών απόδοσης, όπως η accuracy, η precision, η recall και το F1-score, επιτρέποντας την ανάλυση της συμπεριφοράς των μεθόδων τόσο σε συνολικό επίπεδο όσο και ειδικά ως προς τη μειοψηφική κλάση.

Μέσω της εκτεταμένης αυτής πειραματικής μελέτης, η εργασία φιλοδοξεί να προσφέρει εμπειρικά τεκμηριωμένα συμπεράσματα σχετικά με την αποτελεσματικότητα των τεχνικών oversampling και undersampling σε διαφορετικούς βαθμούς ανισοκατανομής και τύπους δεδομένων, συμβάλλοντας στην καλύτερη κατανόηση των πλεονεκτημάτων και περιορισμών τους σε πρακτικά προβλήματα κατηγοριοποίησης.

1.6 Οργάνωση της εργασίας

Η παρούσα εργασία οργανώνεται σε επτά κεφάλαια, τα οποία καλύπτουν τόσο το θεωρητικό υπόβαθρο όσο και την πειραματική μελέτη της κατηγοριοποίησης ανισοκατανεμημένων δεδομένων μέσω τεχνικών υπερδειγματοληψίας και υποδειγματοληψίας.

Στο Κεφάλαιο 1 παρουσιάζεται η εισαγωγή στο αντικείμενο της εργασίας. Αρχικά ορίζεται η έννοια της κατηγοριοποίησης δεδομένων, αναλύεται το πρόβλημα της ανισοκατανομής των κλάσεων και παρουσιάζονται συνοπτικά οι βασικές προσεγγίσεις αντιμετώπισής του. Επιπλέον, τεκμηριώνεται το κίνητρο της παρούσας μελέτης και συνοψίζονται οι βασικές της συνεισφορές.

Στο Κεφάλαιο 2 παρουσιάζονται οι τεχνικές υπερδειγματοληψίας (oversampling), με έμφαση στις βασικές αρχές λειτουργίας τους και στις σημαντικότερες μεθόδους που έχουν προταθεί στη βιβλιογραφία. Αναλύονται τόσο κλασικές όσο και πιο σύγχρονες προσεγγίσεις δημιουργίας συνθετικών δειγμάτων της μειοψηφικής κλάσης.

Στο Κεφάλαιο 3 εξετάζονται οι τεχνικές υποδειγματοληψίας (undersampling), οι οποίες στοχεύουν στη μείωση του πλήθους των δειγμάτων της πλειοψηφικής κλάσης. Παρουσιάζονται οι κυριότερες μέθοδοι και συζητούνται τα πλεονεκτήματα και οι περιορισμοί τους στο πλαίσιο ανισοκατανεμημένων συνόλων δεδομένων.

Στο Κεφάλαιο 4 παρουσιάζονται οι λεπτομέρειες υλοποίησης της πειραματικής διαδικασίας σε περιβάλλον Python. Περιγράφονται τα εργαλεία, οι βιβλιοθήκες και οι βασικές επιλογές σχεδιασμού που ακολουθήθηκαν για την υλοποίηση των τεχνικών δειγματοληψίας και της διαδικασίας αξιολόγησης.

Στο Κεφάλαιο 5 παρουσιάζεται η πειραματική μελέτη της εργασίας. Αναλύονται τα χρησιμοποιηθέντα σύνολα δεδομένων, η πειραματική μεθοδολογία και τα αποτελέσματα της σύγκρισης των τεχνικών oversampling και undersampling, με βάση καθιερωμένες μετρικές απόδοσης.

Τέλος, στο Κεφάλαιο 6 συνοψίζονται τα κύρια συμπεράσματα της εργασίας και συζητούνται πιθανοί άξονες μελλοντικής έρευνας που προκύπτουν από τα αποτελέσματα της πειραματικής ανάλυσης.

Κεφάλαιο 2ο: Τεχνικές Υπερδειγματοληψίας

2.1 Εισαγωγή

Οι τεχνικές υπερδειγματοληψίας (oversampling) αποτελούν μία βασική κατηγορία μεθόδων για την αντιμετώπιση της ανισοκατανομής των κλάσεων στην κατηγοριοποίηση δεδομένων. Κύριος στόχος τους είναι η ενίσχυση της μειοψηφικής κλάσης στο σύνολο εκπαίδευσης, με σκοπό τη βελτίωση της ικανότητας των ταξινομητών να αναγνωρίζουν σωστά τα αντίστοιχα στιγμιότυπα.

Σε αντίθεση με τις απλές τεχνικές αναπαραγωγής δεδομένων, οι σύγχρονες μέθοδοι υπερδειγματοληψίας βασίζονται στη δημιουργία συνθετικών δειγμάτων, αξιοποιώντας τη δομή και τις τοπικές σχέσεις των δεδομένων στον χώρο των χαρακτηριστικών. Με τον τρόπο αυτό επιχειρείται η αποφυγή φαινομένων υπερπροσαρμογής και η αποτελεσματικότερη αναπαράσταση της μειοψηφικής κλάσης.

Στο παρόν κεφάλαιο παρουσιάζονται οι κυριότερες τεχνικές υπερδειγματοληψίας που χρησιμοποιούνται στην εργασία, με έμφαση στις βασικές αρχές λειτουργίας τους. Στις επόμενες ενότητες αναλύονται οι μέθοδοι SMOTE, ADASYN, BorderlineSMOTE, KMeansSMOTE και SVMSMOTE.

2.2 SMOTE

Η τεχνική SMOTE (Synthetic Minority Over-sampling Technique) αποτελεί μία από τις πιο διαδεδομένες και καθιερωμένες μεθόδους υπερδειγματοληψίας για την αντιμετώπιση της ανισοκατανομής των κλάσεων. Σε αντίθεση με την απλή αναπαραγωγή υπαρχόντων δειγμάτων της μειοψηφικής κλάσης, η SMOTE δημιουργεί νέα, συνθετικά δείγματα, με στόχο τη βελτίωση της γενικευσιμότητας των ταξινομητών και τη μείωση του κινδύνου υπερπροσαρμογής [11].

Η βασική αρχή λειτουργίας της SMOTE βασίζεται στη γεωμετρική δομή των δεδομένων στον χώρο των χαρακτηριστικών. Για κάθε δείγμα της μειοψηφικής κλάσης, επιλέγονται οι k πλησιέστεροι γείτονές του, οι οποίοι ανήκουν επίσης στη μειοψηφική κλάση. Στη συνέχεια, νέα συνθετικά δείγματα δημιουργούνται μέσω γραμμικής παρεμβολής μεταξύ του αρχικού δείγματος και ενός τυχαία επιλεγμένου γείτονα. Με τον τρόπο αυτό, τα νέα δείγματα τοποθετούνται σε ενδιάμεσες θέσεις στον χώρο χαρακτηριστικών, επεκτείνοντας τις περιοχές που καταλαμβάνει η μειοψηφική κλάση.

Η κλασική υλοποίηση της SMOTE προϋποθέτει ότι όλα τα χαρακτηριστικά είναι αριθμητικά, καθώς η διαδικασία δημιουργίας των συνθετικών δειγμάτων βασίζεται σε υπολογισμό αποστάσεων και αριθμητικές πράξεις παρεμβολής. Ως αποτέλεσμα, η απευθείας εφαρμογή της σε σύνολα δεδομένων που περιέχουν κατηγορικά χαρακτηριστικά δεν είναι εφικτή χωρίς τροποποιήσεις [15].

Για την αντιμετώπιση αυτού του περιορισμού έχουν προταθεί επεκτάσεις της SMOTE, προσαρμοσμένες στη φύση των δεδομένων. Η μέθοδος SMOTEN (SMOTE for Nominal features)

εφαρμόζεται σε σύνολα δεδομένων που αποτελούνται αποκλειστικά από κατηγορικά χαρακτηριστικά. Στην περίπτωση αυτή, η δημιουργία νέων δειγμάτων δεν βασίζεται σε αριθμητική παρεμβολή, αλλά σε συνδυασμούς κατηγορικών τιμών από γειτονικά δείγματα της μειοψηφικής κλάσης, επιλέγοντας τις πιο συχνές τιμές ανά χαρακτηριστικό [15].

Αντίστοιχα, η SMOTENC (SMOTE for Nominal and Continuous features) έχει σχεδιαστεί για σύνολα δεδομένων με μικτά χαρακτηριστικά, δηλαδή συνδυασμό αριθμητικών και κατηγορικών μεταβλητών. Στη μέθοδο αυτή, η παρεμβολή εφαρμόζεται μόνο στα αριθμητικά χαρακτηριστικά, ενώ τα κατηγορικά χαρακτηριστικά αντιμετωπίζονται με ειδικό μηχανισμό επιλογής τιμών, ώστε να διατηρείται η λογική και σημασιολογική συνοχή των παραγόμενων συνθετικών δειγμάτων [15].

Η SMOTE και οι παραλλαγές της χρησιμοποιούνται ευρέως στη βιβλιογραφία και συχνά αποτελούν σημείο αναφοράς για τη σύγκριση νεότερων τεχνικών υπερδειγματοληψίας. Παρότι παρουσιάζουν σημαντικά πλεονεκτήματα, ενδέχεται να δημιουργήσουν συνθετικά δείγματα σε περιοχές όπου οι κλάσεις επικαλύπτονται, γεγονός που μπορεί να επηρεάσει αρνητικά την απόδοση του ταξινομητή. Οι περιορισμοί αυτοί οδήγησαν στην ανάπτυξη πιο εξειδικευμένων μεθόδων, οι οποίες παρουσιάζονται στις επόμενες ενότητες.

2.3 ADASYN

Η μέθοδος ADASYN (Adaptive Synthetic Sampling) αποτελεί τεχνική υπερδειγματοληψίας που επεκτείνει τη βασική φιλοσοφία της SMOTE, εισάγοντας έναν προσαρμοστικό μηχανισμό δημιουργίας συνθετικών δειγμάτων της μειοψηφικής κλάσης. Κύριος στόχος της ADASYN είναι η εστίαση της διαδικασίας υπερδειγματοληψίας σε περιοχές του χώρου χαρακτηριστικών όπου η ταξινόμηση των δειγμάτων της μειοψηφικής κλάσης είναι δυσκολότερη [16].

Σε αντίθεση με τη SMOTE, η οποία δημιουργεί συνθετικά δείγματα με σχετικά ομοιόμορφο τρόπο για όλα τα δείγματα της μειοψηφικής κλάσης, η ADASYN προσαρμόζει τον αριθμό των παραγόμενων συνθετικών δειγμάτων ανάλογα με την τοπική δυσκολία ταξινόμησης. Η δυσκολία αυτή εκτιμάται βάσει της αναλογίας των γειτόνων της πλειοψηφικής κλάσης στο σύνολο των πλησιέστερων γειτόνων κάθε δείγματος της μειοψηφικής κλάσης.

Πιο συγκεκριμένα, για κάθε δείγμα της μειοψηφικής κλάσης υπολογίζεται ένας δείκτης δυσκολίας, ο οποίος αντανακλά τον βαθμό στον οποίο το δείγμα περιβάλλεται από δείγματα της πλειοψηφικής κλάσης. Δείγματα που βρίσκονται κοντά στα όρια απόφασης ή σε περιοχές έντονης επικάλυψης των κλάσεων λαμβάνουν υψηλότερη τιμή δείκτη και, συνεπώς, συνεισφέρουν στη δημιουργία περισσότερων συνθετικών δειγμάτων. Αντιθέτως, δείγματα που βρίσκονται σε «ασφαλείς» περιοχές του χώρου χαρακτηριστικών ενισχύονται σε μικρότερο βαθμό [16].

Η δημιουργία των συνθετικών δειγμάτων στην ADASYN πραγματοποιείται μέσω παρεμβολής μεταξύ ενός δείγματος της μειοψηφικής κλάσης και ενός από τους πλησιέστερους γείτονές του που

ανήκουν επίσης στη μειοψηφική κλάση, με τρόπο αντίστοιχο της SMOTE. Ωστόσο, η κατανομή των παραγόμενων δειγμάτων δεν είναι ομοιόμορφη, αλλά καθορίζεται από τον προσαρμοστικό μηχανισμό της μεθόδου.

Ένα από τα βασικά πλεονεκτήματα της ADASYN είναι ότι ενισχύει κυρίως τις «δύσκολες» περιοχές του χώρου χαρακτηριστικών, γεγονός που μπορεί να οδηγήσει σε καλύτερη προσαρμογή των ταξινομητών στα όρια μεταξύ των κλάσεων. Παράλληλα, η μέθοδος ενδέχεται να αυξήσει την ευαισθησία του μοντέλου σε θόρυβο, καθώς η υπερενίσχυση περιοχών με έντονη επικάλυψη μπορεί να οδηγήσει στη δημιουργία λιγότερο αντιπροσωπευτικών συνθετικών δειγμάτων. Οι ιδιότητες αυτές καθιστούν την ADASYN ιδιαίτερα ενδιαφέρουσα για συγκριτική αξιολόγηση με άλλες τεχνικές υπερδειγματοληψίας.

2.4 BorderlineSMOTE

Η μέθοδος BorderlineSMOTE αποτελεί επέκταση της κλασικής SMOTE και έχει σχεδιαστεί με στόχο τη βελτίωση της απόδοσης των ταξινομητών σε περιπτώσεις όπου οι κλάσεις παρουσιάζουν έντονη επικάλυψη. Η βασική της ιδέα είναι ότι τα δείγματα της μειοψηφικής κλάσης που βρίσκονται κοντά στα όρια απόφασης είναι πιο κρίσιμα για τη διαδικασία μάθησης και, συνεπώς, θα πρέπει να ενισχύονται περισσότερο κατά την υπερδειγματοληψία [17].

Σε αντίθεση με τη SMOTE, η οποία δημιουργεί συνθετικά δείγματα για όλα τα στιγμιότυπα της μειοψηφικής κλάσης, η BorderlineSMOTE επικεντρώνεται αποκλειστικά σε εκείνα τα δείγματα που βρίσκονται σε «επικίνδυνες» περιοχές του χώρου χαρακτηριστικών. Τα δείγματα αυτά εντοπίζονται με βάση τη σύνθεση των πλησιέστερων γειτόνων τους. Συγκεκριμένα, ένα δείγμα της μειοψηφικής κλάσης θεωρείται ότι βρίσκεται κοντά στο όριο απόφασης όταν σημαντικό ποσοστό των k πλησιέστερων γειτόνων του ανήκει στην πλειοψηφική κλάση.

Η μέθοδος διαχωρίζει τα δείγματα της μειοψηφικής κλάσης σε τρεις κατηγορίες:

- «ασφαλή» δείγματα, τα οποία περιβάλλονται κυρίως από δείγματα της ίδιας κλάσης,
- «επικίνδυνα» δείγματα (borderline), τα οποία βρίσκονται κοντά στα όρια μεταξύ των κλάσεων,
- «θορυβώδη» δείγματα, τα οποία περιβάλλονται σχεδόν αποκλειστικά από δείγματα της πλειοψηφικής κλάσης.

Η διαδικασία υπερδειγματοληψίας εφαρμόζεται κυρίως στα «επικίνδυνα» δείγματα, καθώς θεωρούνται τα πλέον καθοριστικά για τη διαμόρφωση των ορίων απόφασης του ταξινομητή [17].

Η δημιουργία των συνθετικών δειγμάτων πραγματοποιείται μέσω παρεμβολής μεταξύ των επιλεγμένων «borderline» δειγμάτων και των πλησιέστερων γειτόνων τους που ανήκουν στη μειοψηφική κλάση, ακολουθώντας τη βασική φιλοσοφία της SMOTE. Με τον τρόπο αυτό, η μέθοδος

επιδιώκει να ενισχύσει τη διακριτική ικανότητα του ταξινομητή στις περιοχές όπου παρατηρείται η μεγαλύτερη αβεβαιότητα.

Η BorderlineSMOTE παρουσιάζει το πλεονέκτημα ότι αποφεύγει τη δημιουργία συνθετικών δειγμάτων σε «ασφαλείς» περιοχές, όπου η πληροφορία είναι ήδη επαρκής, μειώνοντας έτσι τον κίνδυνο υπερπροσαρμογής. Ωστόσο, όπως και άλλες τεχνικές που εστιάζουν στα όρια απόφασης, μπορεί να είναι ευαίσθητη σε θόρυβο και λανθασμένα επισημασμένα δείγματα. Οι ιδιότητες αυτές καθιστούν τη BorderlineSMOTE ιδιαίτερα ενδιαφέρουσα για συγκριτική αξιολόγηση με μεθόδους όπως η SMOTE και η ADASYN.

2.5 KMeansSMOTE

Η μέθοδος KMeansSMOTE αποτελεί μια πιο πρόσφατη και εξελιγμένη τεχνική υπερδειγματοληψίας, η οποία συνδυάζει τη βασική φιλοσοφία της SMOTE με τεχνικές ομαδοποίησης (clustering), με στόχο τη βελτίωση της ποιότητας των παραγόμενων συνθετικών δειγμάτων. Η κύρια ιδέα της μεθόδου είναι ότι η δημιουργία συνθετικών δειγμάτων θα πρέπει να λαμβάνει υπόψη τη συνολική δομή των δεδομένων και όχι μόνο τις τοπικές σχέσεις μεταξύ γειτονικών δειγμάτων [18].

Στην KMeansSMOTE, το σύνολο εκπαίδευσης αρχικά ομαδοποιείται με τη χρήση του αλγορίθμου k-means. Στη συνέχεια, επιλέγονται εκείνες οι συστάδες (clusters) που περιέχουν σημαντικό ποσοστό δειγμάτων της μειοψηφικής κλάσης. Η υπερδειγματοληψία εφαρμόζεται μόνο σε αυτές τις συστάδες, καθώς θεωρείται ότι αντιπροσωπεύουν περιοχές του χώρου χαρακτηριστικών όπου η ενίσχυση της μειοψηφικής κλάσης είναι περισσότερο αναγκαία.

Αφού προσδιοριστούν οι κατάλληλες συστάδες, η δημιουργία των συνθετικών δειγμάτων πραγματοποιείται μέσω της κλασικής διαδικασίας της SMOTE, δηλαδή μέσω παρεμβολής μεταξύ δειγμάτων της μειοψηφικής κλάσης εντός της ίδιας συστάδας. Με τον τρόπο αυτό, η μέθοδος επιδιώκει να διατηρήσει τη συνοχή των δεδομένων και να αποτρέψει τη δημιουργία συνθετικών δειγμάτων σε περιοχές του χώρου χαρακτηριστικών που δεν είναι αντιπροσωπευτικές της μειοψηφικής κλάσης.

Ένα βασικό πλεονέκτημα της KMeansSMOTE είναι ότι μειώνει τον κίνδυνο δημιουργίας συνθετικών δειγμάτων σε περιοχές με έντονη επικάλυψη των κλάσεων ή σε αραιά τμήματα του χώρου χαρακτηριστικών. Επιπλέον, η ενσωμάτωση της πληροφορίας της ομαδοποίησης επιτρέπει καλύτερο έλεγχο της διαδικασίας υπερδειγματοληψίας, ιδιαίτερα σε σύνολα δεδομένων με πολύπλοκη γεωμετρική δομή.

Ωστόσο, η απόδοση της KMeansSMOTE εξαρτάται από την ποιότητα της ομαδοποίησης και από την επιλογή του αριθμού των συστάδων. Μη κατάλληλες επιλογές παραμέτρων ενδέχεται να οδηγήσουν σε ανεπαρκή ή υπερβολική ενίσχυση της μειοψηφικής κλάσης. Παρά τους περιορισμούς

αυτούς, η KMeansSMOTE αποτελεί μια ιδιαίτερα ενδιαφέρουσα προσέγγιση, καθώς επιχειρεί να συνδυάσει τοπικές και καθολικές πληροφορίες για τη βελτίωση της διαδικασίας υπερδειγματοληψίας.

2.6 SVMSMOTE

Η μέθοδος SVMSMOTE αποτελεί μία εξειδικευμένη τεχνική υπερδειγματοληψίας που συνδυάζει τη βασική ιδέα της SMOTE με πληροφορία που προέρχεται από τη χρήση μηχανών διανυσμάτων υποστήριξης (Support Vector Machines – SVM). Στόχος της μεθόδου είναι η δημιουργία συνθετικών δειγμάτων της μειοψηφικής κλάσης σε περιοχές του χώρου χαρακτηριστικών που βρίσκονται κοντά στα όρια απόφασης μεταξύ των κλάσεων, τα οποία θεωρούνται ιδιαίτερα κρίσιμα για τη διαδικασία ταξινόμησης [19].

Η SVMSMOTE αξιοποιεί έναν SVM ταξινομητή για τον εντοπισμό των διανυσμάτων υποστήριξης (support vectors) της μειοψηφικής κλάσης. Τα δείγματα αυτά βρίσκονται πλησίον του υπερεπιπέδου απόφασης και, συνεπώς, αντιπροσωπεύουν περιοχές όπου η διάκριση μεταξύ των κλάσεων είναι πιο δύσκολη. Η διαδικασία υπερδειγματοληψίας επικεντρώνεται κυρίως στα συγκεκριμένα δείγματα, σε αντίθεση με την κλασική SMOTE που αντιμετωπίζει όλα τα δείγματα της μειοψηφικής κλάσης με παρόμοιο τρόπο.

Η δημιουργία των συνθετικών δειγμάτων πραγματοποιείται μέσω παρεμβολής μεταξύ των επιλεγμένων support vectors της μειοψηφικής κλάσης και των πλησιέστερων γειτόνων τους, οι οποίοι ανήκουν επίσης στη μειοψηφική κλάση. Με τον τρόπο αυτό, τα νέα δείγματα τοποθετούνται κοντά στα όρια απόφασης, ενισχύοντας τη δυνατότητα του ταξινομητή να μάθει πιο ακριβή και σταθερά όρια διαχωρισμού.

Ένα βασικό πλεονέκτημα της SVMSMOTE είναι ότι αποφεύγει την υπερδειγματοληψία «ασφαλών» περιοχών του χώρου χαρακτηριστικών, όπου η πληροφορία είναι ήδη επαρκής, και εστιάζει στις περιοχές που επηρεάζουν περισσότερο τη διαδικασία μάθησης. Ωστόσο, η μέθοδος προϋποθέτει την εκπαίδευση ενός επιπλέον μοντέλου SVM, γεγονός που αυξάνει το υπολογιστικό κόστος. Επιπλέον, η απόδοσή της εξαρτάται από την επιλογή των παραμέτρων του SVM και από την ποιότητα του διαχωρισμού που επιτυγχάνεται.

Παρά τους περιορισμούς αυτούς, η SVMSMOTE θεωρείται ιδιαίτερα αποτελεσματική σε προβλήματα με έντονη επικάλυψη των κλάσεων και αποτελεί σημαντικό σημείο αναφοράς στη συγκριτική αξιολόγηση τεχνικών υπερδειγματοληψίας που εστιάζουν στα όρια απόφασης.

2.7 Επίλογος

Στο παρόν κεφάλαιο παρουσιάστηκαν οι βασικές τεχνικές υπερδειγματοληψίας που χρησιμοποιούνται για την αντιμετώπιση του προβλήματος της ανισοκατανομής των κλάσεων στην κατηγοριοποίηση δεδομένων. Αρχικά, δόθηκε μια συνοπτική εισαγωγή στις μεθόδους oversampling και στη φιλοσοφία δημιουργίας συνθετικών δειγμάτων της μειοψηφικής κλάσης.

Στη συνέχεια, αναλύθηκε η κλασική μέθοδος SMOTE, η οποία αποτελεί τη βάση για πολλές σύγχρονες τεχνικές υπερδειγματοληψίας, καθώς και οι επεκτάσεις της SMOTEN και SMOTENC, που επιτρέπουν την εφαρμογή της σε σύνολα δεδομένων με κατηγορικά ή μικτά χαρακτηριστικά. Ακολούθως, παρουσιάστηκε η ADASYN, η οποία εισάγει έναν προσαρμοστικό μηχανισμό δημιουργίας συνθετικών δειγμάτων, εστιάζοντας σε περιοχές του χώρου χαρακτηριστικών με αυξημένη δυσκολία ταξινόμησης.

Επιπλέον, εξετάστηκε η BorderlineSMOTE, η οποία επικεντρώνεται στα δείγματα της μειοψηφικής κλάσης που βρίσκονται κοντά στα όρια απόφασης, καθώς και η KMeansSMOTE, η οποία ενσωματώνει τεχνικές ομαδοποίησης με στόχο τη βελτίωση της ποιότητας και της τοποθέτησης των παραγόμενων συνθετικών δειγμάτων. Τέλος, παρουσιάστηκε η SVMSMOTE, μια πιο εξειδικευμένη μέθοδος που αξιοποιεί πληροφορία από μηχανές διανυσμάτων υποστήριξης για την ενίσχυση κρίσιμων περιοχών του χώρου χαρακτηριστικών.

Η ποικιλία των τεχνικών υπερδειγματοληψίας που παρουσιάστηκαν καταδεικνύει ότι η αποτελεσματική αντιμετώπιση της ανισοκατανομής των κλάσεων δεν μπορεί να βασιστεί σε μία και μοναδική προσέγγιση. Κάθε μέθοδος παρουσιάζει διαφορετικά πλεονεκτήματα και περιορισμούς, γεγονός που καθιστά αναγκαία τη συγκριτική αξιολόγησή τους υπό κοινές πειραματικές συνθήκες. Στο επόμενο κεφάλαιο παρουσιάζονται οι τεχνικές υποδειγματοληψίας, οι οποίες αποτελούν μια εναλλακτική και συμπληρωματική προσέγγιση για την αντιμετώπιση του ίδιου προβλήματος.

Κεφάλαιο 3ο: Τεχνικές Υποδειγματοληψίας

3.1 Εισαγωγή

Οι τεχνικές υποδειγματοληψίας (undersampling) αποτελούν μια εναλλακτική προσέγγιση για την αντιμετώπιση της ανισοκατανομής των κλάσεων στην κατηγοριοποίηση δεδομένων. Σε αντίθεση με τις τεχνικές υπερδειγματοληψίας, οι μέθοδοι undersampling στοχεύουν στη μείωση του πλήθους των δειγμάτων της πλειοψηφικής κλάσης, με σκοπό την εξισορρόπηση της κατανομής των δεδομένων στο σύνολο εκπαίδευσης.

Η βασική φιλοσοφία των τεχνικών υποδειγματοληψίας είναι η απομάκρυνση πλεοναζόντων ή λιγότερο αντιπροσωπευτικών δειγμάτων της πλειοψηφικής κλάσης, ώστε ο ταξινομητής να μην επηρεάζεται δυσανάλογα από αυτήν κατά τη διαδικασία εκπαίδευσης. Με τον τρόπο αυτό, επιδιώκεται η βελτίωση της απόδοσης ως προς τη μειοψηφική κλάση, συχνά με μειωμένο υπολογιστικό κόστος.

Στο παρόν κεφάλαιο παρουσιάζονται οι βασικότερες τεχνικές υποδειγματοληψίας που χρησιμοποιούνται στην εργασία, από απλές τυχαίες μεθόδους έως πιο εξελιγμένες προσεγγίσεις που βασίζονται σε γειτνίαση και τοπική δομή των δεδομένων. Στιςεπόμενεςενότητεςαναλύονταιοιμέθοδοι Random Undersampling, Edited Nearest Neighbors, Repeated Edited Nearest Neighbors, All k Nearest Neighbors, Condensed Nearest Neighborsκαι Tomek Links.

3.2 Random Undersampler

Η μέθοδος RandomUndersampling αποτελεί την απλούστερη και πιο άμεσα εφαρμόσιμη τεχνική υποδειγματοληψίας για την αντιμετώπιση της ανισοκατανομής των κλάσεων. Η βασική της ιδέα συνίσταται στη τυχαία αφαίρεση δειγμάτων από την πλειοψηφική κλάση, έως ότου επιτευχθεί μια επιθυμητή αναλογία μεταξύ των κλάσεων στο σύνολο εκπαίδευσης [12].

Η διαδικασία αυτή χαρακτηρίζεται από «γνωστική ουδετερότητα» ως προς τα δεδομένα, καθώς δεν συνεκτιμά τη γεωμετρική κατανομή στον χώρο των χαρακτηριστικών ή τις σχέσεις γειτνίασης μεταξύ των σημείων. Κάθε δείγμα της πλειοψηφικής κλάσης κατέχει την ίδια στατιστική πιθανότητα διαγραφής, ανεξάρτητα από το αν βρίσκεται κοντά στα όρια απόφασης (decision boundaries) ή αν αποτελεί κεντρικό, αντιπροσωπευτικό στοιχείο της κλάσης του. Αυτή η έλλειψη πολυπλοκότητας καθιστά τη μέθοδο εξαιρετικά αποδοτική από υπολογιστική άποψη, απαιτώντας ελάχιστους πόρους και χρόνο επεξεργασίας, ακόμη και σε περιβάλλοντα με περιορισμένη μνήμη.

Ένα από τα κύρια πλεονεκτήματα του Random Undersampling είναι η δραστική μείωση του όγκου του συνόλου εκπαίδευσης. Η συρρίκνωση αυτή όχι μόνο επιταχύνει τους αλγορίθμους εκπαίδευσης, αλλά σε ορισμένες περιπτώσεις μειώνει και τον κίνδυνο υπερεκπαίδευσης (overfitting) που συχνά συνοδεύει την αντίστροφη μέθοδο της υπερδειγματοληψίας (oversampling).

Εξισορροπώντας το βάρος των κλάσεων, η μέθοδος αναγκάζει τον ταξινομητή να «δώσει προσοχή» στα σπάνια δείγματα της μειοψηφικής κλάσης, βελτιώνοντας σημαντικά δείκτες απόδοσης όπως η ανάκληση (recall) και το F1-score, ειδικά όταν η αρχική ανισοκατανομή είναι τόσο έντονη που το μοντέλο τείνει να αγνοεί πλήρως τη μικρή κλάση.

Ωστόσο, η τυχαία φύση της επιλογής αποτελεί ταυτόχρονα και τη μεγαλύτερη αδυναμία της τεχνικής. Η διαγραφή δειγμάτων ενδέχεται να οδηγήσει σε σημαντική απώλεια πληροφορίας (information loss), καθώς υπάρχει ο κίνδυνος να αφαιρεθούν κρίσιμα στιγμιότυπα που οριοθετούν την πλειοψηφική κλάση ή περιγράφουν την εσωτερική της ποικιλομορφία. Η απώλεια αυτή μπορεί να αλλοιώσει την περιγραφική ικανότητα του μοντέλου για την πλειοψηφική κλάση, οδηγώντας σε πτώση της γενικευτικής ισχύος και σε ασταθή αποτελέσματα, ιδιαίτερα όταν το διαθέσιμο σύνολο δεδομένων είναι ήδη μικρό ή όταν η δομή των κλάσεων παρουσιάζει μεγάλες επικαλύψεις.

Στην πράξη, λόγω της ευκολίας υλοποίησης και της ταχύτητάς της, η Random Undersampling χρησιμοποιείται συστηματικά ως η βασική γραμμή σύγκρισης (baseline) σε πειραματικές μελέτες. Επιτρέπει στους ερευνητές να αξιολογήσουν εάν η χρήση πιο εξελιγμένων, «ευφυών» τεχνικών — όπως εκείνων που βασίζονται σε γειτονικά δείγματα (π.χ. Tomek Links ή NearMiss)— προσφέρει ουσιαστική βελτίωση στην απόδοση που να δικαιολογεί το αυξημένο υπολογιστικό τους κόστος.

3.3 Edited Nearest Neighbors

Η μέθοδος Edited Nearest Neighbors (ENN) αποτελεί μια σαφώς πιο εξελιγμένη και «ευφυή» τεχνική υποδειγματοληψίας σε σχέση με την απλή τυχαία αφαίρεση δειγμάτων, καθώς βασίζεται στην ανάλυση της τοπικής γεωμετρικής δομής των δεδομένων. Ενώ το Random Undersampling εστιάζει αποκλειστικά στην ποσοτική εξισορρόπηση των κλάσεων, η ENN στοχεύει στην ποιοτική αναβάθμιση του συνόλου εκπαίδευσης. Η κεντρική της επιδίωξη είναι ο εντοπισμός και η απομάκρυνση δειγμάτων που χαρακτηρίζονται ως «θορυβώδη» (noisy) ή εμφανίζονται ως ακραίες τιμές (outliers), συμβάλλοντας έτσι στη δημιουργία ενός πιο καθαρού και διαχωρίσιμου συνόλου δεδομένων [20].

Η βασική αρχή λειτουργίας της ENN εδράζεται στον αλγόριθμο των k-πλησιέστερων γειτόνων (k-NearestNeighbors). Για κάθε δείγμα του συνόλου δεδομένων, εξετάζεται η γειτονιά του, η οποία ορίζεται από τους k πλησιέστερους γείτονες (συνήθως με τιμή k=3 ή k=5). Αν η ετικέτα κλάσης του εξεταζόμενου δείγματος έρχεται σε αντίθεση με την πλειοψηφική τάση των γειτόνων του —δηλαδή, αν το δείγμα φαίνεται να «περιβάλλεται» από δείγματα της άλλης κλάσης— τότε θεωρείται ασυνεπές ή λανθασμένα ταξινομημένο και αφαιρείται από το σύνολο. Παρόλο που η διαδικασία μπορεί να εφαρμοστεί σε ολόκληρο το dataset, συνηθίζεται να εστιάζει στα δείγματα της πλειοψηφικής κλάσης, ώστε να επιτυγχάνεται ταυτόχρονα η μείωση του όγκου της και η αποσαφήνιση των ορίων απόφασης.

Σε αντίθεση με τις τυφλές μεθόδους υποδειγματοληψίας, η ENN «σέβεται» τη γεωμετρική κατανομή του χώρου. Επιχειρεί να διατηρήσει μόνο εκείνα τα δείγματα που παρουσιάζουν υψηλή

τοπική συνέπεια, γεγονός που οδηγεί σε πιο ομαλά και ευδιάκριτα όρια απόφασης (decisionboundaries) για τον ταξινομητή. Ένα ιδιαίτερο πλεονέκτημα της μεθόδου είναι η ικανότητά της να καθαρίζει τις περιοχές επικάλυψης (overlapregions) μεταξύ των κλάσεων. Απομακρύνοντας τα πλειοψηφικά δείγματα που εισχωρούν βαθιά μέσα στα όρια της μειοψηφικής κλάσης, η ENN μειώνει το «θόρυβο» που συχνά οδηγεί τους αλγορίθμους σε εσφαλμένες γενικεύσεις.

Ωστόσο, η χρήση της ENN ενέχει ορισμένους περιορισμούς. Καθώς η μέθοδος είναι αυστηρά βασισμένη στην τοπική ομοιότητα, υπάρχει ο κίνδυνος να αφαιρεθούν δείγματα της μειοψηφικής κλάσης, εάν αυτά βρίσκονται σε εξαιρετικά αραιές περιοχές ή αν ο περιβάλλον θόρυβος είναι πολύ έντονος. Επιπλέον, από υπολογιστική άποψη, η ENN είναι απαιτητική, καθώς ο εντοπισμός των k-πλησιέστερων γειτόνων για κάθε σημείο του συνόλου δεδομένων απαιτεί σημαντικό χρόνο και μνήμη, ειδικά σε σύνολα δεδομένων μεγάλης κλίμακας (bigdata). Παρά το κόστος αυτό, η μέθοδος παραμένει ένα ισχυρό εργαλείο προεπεξεργασίας, αποτελώντας τη βάση για ακόμη πιο ισχυρές επαναληπτικές τεχνικές, όπως η RepeatedEditedNearestNeighbors (RENN), η οποία θα αναλυθεί στη συνέχεια.

3.4 Repeated Edited Nearest Neighbors

Η μέθοδος Repeated Edited Nearest Neighbors (RENN) αποτελεί μια εξελιγμένη επέκταση της τεχνικής ENN, η οποία στοχεύει στη μεγιστοποίηση της ποιότητας του συνόλου εκπαίδευσης μέσω μιας επαναληπτικής και σωρευτικής διαδικασίας φιλτραρίσματος. Η θεμελιώδης παραδοχή της RENN είναι ότι η εφάπαξ εφαρμογή του κανόνα των πλησιέστερων γειτόνων ενδέχεται να μην επαρκεί για τον πλήρη καθαρισμό ενός συνόλου δεδομένων, ιδιαίτερα όταν αυτό παρουσιάζει πολύπλοκη τοπική δομή, έντονη επικάλυψη (overlap) μεταξύ των κλάσεων ή υψηλά επίπεδα θορύβου που σχηματίζουν «συστάδες» λανθασμένων δειγμάτων [21].

Η λειτουργία της RENN βασίζεται στην αναδρομική εφαρμογή του αλγορίθμου ENN μέχρις ότου το σύνολο δεδομένων φτάσει σε μια κατάσταση σταθερότητας. Σε κάθε επανάληψη, εξετάζεται εκ νέου η συνέπεια κάθε δείγματος με βάση τουςk πλησιέστερους γείτονές του. Εάν ένα δείγμα κριθεί ασυνεπές με την πλειοψηφική ετικέτα της γειτονιάς του, αφαιρείται αμέσως. Η κρίσιμη διαφορά έγκειται στο ότι η αφαίρεση δειγμάτων σε μια επανάληψη μεταβάλλει τη σύνθεση των γειτονιών για τα εναπομείναντα δείγματα στην επόμενη. Έτσι, δείγματα που αρχικά φαινόταν να έχουν επαρκή υποστήριξη από τη γειτονιά τους, μπορεί στην πορεία να αποκαλυφθούν ως θορυβώδη καθώς απομακρύνονται οι «συνεργοί» τους. Η διαδικασία αυτή συνεχίζεται αυτόματα έως ότου δεν είναι πλέον δυνατή η αφαίρεση άλλων δειγμάτων ή μέχρι να ικανοποιηθεί ένας προκαθορισμένος αριθμός επαναλήψεων.

Σε σύγκριση με την απλή ENN, η RENN ακολουθεί μια σαφώς πιο επιθετική στρατηγική υποδειγματοληψίας. Το αποτέλεσμα είναι η δημιουργία ενός εξαιρετικά «καθαρού» συνόλου δεδομένων, όπου οι κλάσεις διαχωρίζονται από ευκρινή και απλοποιημένα όρια απόφασης. Αυτή η αυστηρή εξυγίανση του χώρου των χαρακτηριστικών επιτρέπει στους ταξινομητές να επικεντρωθούν

στη βασική δομή των δεδομένων χωρίς να παραπλανώνται από μεμονωμένες εξαιρέσεις ή τοπικές ανωμαλίες.

Ωστόσο, η αυξημένη αποτελεσματικότητα της RENN συνοδεύεται από ανάλογο υπολογιστικό κόστος. Η επαναληπτική φύση της απαιτεί πολλαπλούς υπολογισμούς αποστάσεων και γειτονιών, γεγονός που την καθιστά πιο χρονοβόρα από την απλή ENN. Επιπλέον, ελλοχεύει ο κίνδυνος της υπερβολικής αφαίρεσης δειγμάτων (over-cleaning), η οποία σε ορισμένες περιπτώσεις μπορεί να οδηγήσει σε απώλεια χρήσιμης ποικιλομορφίας, ακόμα και από τη μειοψηφική κλάση, αν αυτή είναι ιδιαίτερα αραιή. Παρά τις προκλήσεις αυτές, η RENN παραμένει ένα από τα πιο ισχυρά εργαλεία προεπεξεργασίας για την αντιμετώπιση του θορύβου σε ανισόρροπα δεδομένα, προσφέροντας μια σταθερή βάση για την ανάπτυξη αξιόπιστων μοντέλων μηχανικής μάθησης.

3.5 All k Nearest Neighbors

Η μέθοδος All k-Nearest Neighbors (All kNN) αποτελεί μια προηγμένη τεχνική υποδειγματοληψίας που βασίζεται στις αρχές της τοπικής γειτνίασης, λειτουργώντας ουσιαστικά ως μια ευρύτερη γενίκευση της μεθόδου Edited Nearest Neighbors (ENN). Ενώ στην κλασική ENN ο έλεγχος συνέπειας ενός δείγματος περιορίζεται σε μια στατική τιμή του k, η All kNN εισάγει μια πολυεπίπεδη προσέγγιση ελέγχου. Η κεντρική της φιλοσοφία εστιάζει στην εξέταση της γειτονιάς ενός δείγματος υπό διαφορετικές κλίμακες, επιτρέποντας μια πιο αξιόπιστη και σφαιρική αναγνώριση των θορυβωδών ή ασυνεπών δειγμάτων που ενδέχεται να αλλοιώνουν το σύνολο εκπαίδευσης [22].

Η βασική λειτουργία της μεθόδου συνίσταται στην επαναλαμβανόμενη εφαρμογή του κανόνα των k πλησιέστερων γειτόνων για ένα εύρος αυξανόμενων τιμών, από 1 έως μια μέγιστη τιμή K (για παράδειγμα, k={1, 2, 3}. Για κάθε δείγμα του συνόλου δεδομένων, ο αλγόριθμος εξετάζει αν η ετικέτα κλάσης του παραμένει συνεπής με την πλειοψηφική τάση των γειτόνων του σε κάθε ένα από αυτά τα βήματα. Εάν ένα δείγμα αποδειχθεί ασυνεπές για οποιαδήποτε από τις επιλεγμένες τιμές του k, χαρακτηρίζεται ως αναξιόπιστο και αφαιρείται οριστικά από το σύνολο. Αυτός ο αυστηρός, κλιμακωτός έλεγχος διασφαλίζει ότι ένα δείγμα θα διατηρηθεί μόνο αν η παρουσία του δικαιολογείται από τη δομή των δεδομένων τόσο σε πολύ στενό όσο και σε ευρύτερο τοπικό πλαίσιο.Η χρήση πολλαπλών τιμών του k προσφέρει στη μέθοδο All kNN μια ιδιαίτερη ανθεκτικότητα έναντι τοπικών ανωμαλιών που θα μπορούσαν να «ξεγελάσουν» την απλή ENN. Με την ανάλυση της γειτονιάς σε διαφορετικά επίπεδα ανάλυσης, επιτυγχάνεται μια πολύ πιο ακριβής εκτίμηση της πραγματικής γεωμετρικής κατανομής των δεδομένων. Το αποτέλεσμα είναι η δραστική απομάκρυνση του θορύβου και η δημιουργία εξαιρετικά σαφών και ομαλών ορίων απόφασης, τα οποία διευκολύνουν τον ταξινομητή να αναγνωρίσει τα θεμελιώδη πρότυπα (patterns) κάθε κλάσης χωρίς παρεμβολές.

Ωστόσο, αυτό το αυξημένο επίπεδο ελέγχου συνοδεύεται από ορισμένες προκλήσεις. Το υπολογιστικό κόστος είναι σαφώς υψηλότερο σε σχέση με τις προηγούμενες μεθόδους, καθώς η διαδικασία απαιτεί τον εντοπισμό και την επεξεργασία πλησιέστερων γειτόνων για πολλαπλά σενάρια.

Επιπλέον, η All kNN τείνει να είναι ιδιαίτερα επιθετική στην αφαίρεση δειγμάτων. Σε περιβάλλοντα με έντονη φυσική επικάλυψη μεταξύ των κλάσεων, η μέθοδος ενδέχεται να απομακρύνει μεγάλο όγκο πληροφορίας από την πλειοψηφική κλάση, γεγονός που απαιτεί προσεκτική παραμετροποίηση για να μην αλλοιωθεί η αντιπροσωπευτικότητα του δείγματος. Παρά τους περιορισμούς αυτούς, η All kNN θεωρείται μια από τις πλέον ισχυρές και αυστηρές εναλλακτικές λύσεις, ιδανική για περιπτώσεις όπου η καθαρότητα των δεδομένων και η εξομάλυνση των ορίων απόφασης αποτελούν προτεραιότητα.

3.6 Condensed Nearest Neighbors

Η μέθοδος Condensed Nearest Neighbors (CNN) αποτελεί μια επιθετική τεχνική υποδειγματοληψίας που εισάγει μια ριζικά διαφορετική προσέγγιση στη διαχείριση του όγκου των δεδομένων. Σε αντίθεση με τις μεθόδους ENN ή RENN, οι οποίες εστιάζουν στην απομάκρυνση του θορύβου και των ασυνεπών δειγμάτων, η CNN στοχεύει στη δραστική συμπύκνωση του συνόλου εκπαίδευσης. Η κεντρική της φιλοσοφία είναι η δημιουργία ενός «πυρηνικού» υποσυνόλου δεδομένων, το οποίο, παρά το μικρό του μέγεθος, διατηρεί ανέπαφη τη διαχωριστική ικανότητα και τα όρια απόφασης του πλήρους αρχικού συνόλου [23].

Η λειτουργία της CNN βασίζεται σε μια επαναληπτική διαδικασία «εκμάθησης μέσω σφαλμάτων». Η διαδικασία ξεκινά με την αρχικοποίηση ενός συνόλου αναφοράς (store set), το οποίο περιλαμβάνει συνήθως ένα τυχαίο δείγμα από κάθε κλάση. Στη συνέχεια, ο αλγόριθμος διατρέχει όλα τα υπόλοιπα δείγματα του αρχικού συνόλου και επιχειρεί να τα ταξινομήσει χρησιμοποιώντας τον κανόνα του πλησιέστερου γείτονα (1-NN) με βάση μόνο τα στοιχεία που υπάρχουν στο σύνολο αναφοράς. Εάν ένα δείγμα ταξινομηθεί σωστά, θεωρείται «περιττό» και παραλείπεται. Εάν, όμως, ταξινομηθεί λανθασμένα, σημαίνει ότι μεταφέρει κρίσιμη πληροφορία για τη γεωμετρία των κλάσεων που λείπει από το σύνολο αναφοράς, και έτσι προστίθεται σε αυτό. Η διαδικασία συνεχίζεται με διαδοχικά περάσματα έως ότου το συμπυκνωμένο σύνολο είναι σε θέση να ταξινομήσει σωστά όλα τα δείγματα του αρχικού συνόλου.

Με την εφαρμογή αυτής της στρατηγικής, η CNN διατηρεί κυρίως τα δείγματα που βρίσκονται πλησίον των ορίων απόφασης (decision boundaries), καθώς αυτά είναι τα πλέον επιρρεπή σε λανθασμένη ταξινόμηση και ταυτόχρονα τα πιο απαραίτητα για τον ορισμό των κλάσεων. Το αποτέλεσμα είναι ένα εξαιρετικά περιορισμένο σύνολο εκπαίδευσης, το οποίο επιταχύνει δραματικά τις διαδικασίες εκπαίδευσης και πρόβλεψης, καθιστώντας την τεχνική ιδιαίτερα ελκυστική για εφαρμογές μεγάλης κλίμακας με περιορισμένους υπολογιστικούς πόρους.

Ωστόσο, η επιθετική φύση της CNN ενέχει κινδύνους, ειδικά σε ανισοκατανεμημένα περιβάλλοντα. Η μέθοδος τείνει να διατηρεί σχεδόν όλα τα δείγματα της μειοψηφικής κλάσης (καθώς είναι σπάνια και συχνά ταξινομούνται λάθος από ένα μικρό σύνολο αναφοράς), ενώ μειώνει δραστικά την πλειοψηφική κλάση. Αυτό μπορεί να οδηγήσει σε μια «τεχνητή» μετατόπιση της μεροληψίας του ταξινομητή υπέρ της μειοψηφικής κλάσης. Επιπλέον, η CNN είναι εξαιρετικά ευαίσθητη στον

θόρυβο: αν ένα δείγμα έχει λανθασμένη ετικέτα, ο αλγόριθμος θα το συμπεριλάβει υποχρεωτικά στο συμπυκνωμένο σύνολο για να «διορθώσει» το σφάλμα ταξινόμησης, με αποτέλεσμα την παραμόρφωση των ορίων απόφασης. Παρά τις αδυναμίες αυτές, η CNN παραμένει ιστορικό σημείο αναφοράς, αναδεικνύοντας τη σημασία της επιλεκτικής διατήρησης δειγμάτων στη μηχανική μάθηση.

3.7 Tomek Links

Η μέθοδος Tomek Links αποτελεί μια εξειδικευμένη τεχνική υποδειγματοληψίας που επικεντρώνεται στη γεωμετρική αποσαφήνιση των ορίων μεταξύ των κλάσεων. Σε αντίθεση με τις μεθόδους που στοχεύουν στη μαζική μείωση του όγκου της πλειοψηφικής κλάσης, η προσέγγιση Tomek Links λειτουργεί με χειρουργική ακρίβεια. Στόχος της δεν είναι η ποσοτική εξισορρόπηση των δεδομένων, αλλά η ποιοτική βελτίωση του συνόλου εκπαίδευσης μέσω της απομάκρυνσης δειγμάτων που προκαλούν σύγχυση στους ταξινομητές, ιδιαίτερα σε περιοχές έντονης επικάλυψης (overlap) [21].

Ένα ζεύγος δειγμάτων (x_i, x_j) ορίζεται ως Tomek Link όταν πληρούνται τρεις συγκεκριμένες συνθήκες: (α) τα δύο δείγματα ανήκουν σε διαφορετικές κλάσεις, (β) το x_j είναι ο πλησιέστερος γείτονας του x_i, και (γ) το x_i είναι ο πλησιέστερος γείτονας του x_j. Η ύπαρξη ενός τέτοιου ζεύγους υποδηλώνει ότι τα δύο σημεία βρίσκονται σε εξαιρετικά κοντινή απόσταση μεταξύ τους, «εκατέρωθεν» του ιδεατού ορίου απόφασης. Τέτοια δείγματα είναι συχνά θορυβώδη ή αποτελούν οριακές περιπτώσεις που καθιστούν το διαχωρισμό των κλάσεων ασαφή και πολύπλοκο.

Στο πλαίσιο της αντιμετώπισης ανισοκατανεμημένων δεδομένων, η εφαρμογή της τεχνικής εστιάζει στην αφαίρεση του δείγματος που ανήκει στην πλειοψηφική κλάση από κάθε εντοπισμένο ζεύγος. Με τον τρόπο αυτό, η μειοψηφική κλάση παραμένει ανέπαφη, ενώ η πλειοψηφική υποχωρεί ελαφρώς, δημιουργώντας έναν «κενό χώρο» (margin) ανάμεσα στις δύο κατηγορίες. Αυτή η διαδικασία αποσαφήνισης των ορίων επιτρέπει στον ταξινομητή να ορίσει πιο απλά και αποτελεσματικά τη διαχωριστική γραμμή, μειώνοντας τις πιθανότητες εσφαλμένης ταξινόμησης λόγω γειτνίασης.

Το κύριο πλεονέκτημα της μεθόδου Tomek Links είναι ο συντηρητικός και ήπιος χαρακτήρας της. Καθώς αφαιρεί μόνο τα δείγματα που βρίσκονται σε «αμφισβητούμενες» ζώνες, δεν αλλοιώνει τη συνολική κατανομή ή τη στατιστική δομή της πλειοψηφικής κλάσης. Αυτό την καθιστά ιδανική για περιπτώσεις όπου η απώλεια πληροφορίας πρέπει να ελαχιστοποιηθεί. Ωστόσο, η αποτελεσματικότητά της είναι άρρηκτα συνδεδεμένη με τη χρησιμοποιούμενη μετρική απόστασης (π.χ. Ευκλείδεια), καθώς σε χώρους πολλών διαστάσεων η έννοια του «πλησιέστερου γείτονα» μπορεί να καταστεί ασταθής.

Στην πράξη, τα Tomek Links σπάνια χρησιμοποιούνται για την πλήρη εξισορρόπηση ενός dataset. Αντιθέτως, αξιοποιούνται συχνά ως ένα κρίσιμο στάδιο «μετα-επεξεργασίας» μετά την εφαρμογή τεχνικών υπερδειγματοληψίας (όπως η SMOTE). Σε αυτούς τους συνδυαστικούς

αλγορίθμους, η υπερδειγματοληψία δημιουργεί νέα δείγματα και η μέθοδος Tomek Links καθαρίζει τα τεχνητά όρια που προκύπτουν, οδηγώντας σε μοντέλα με εξαιρετική γενικευτική ικανότητα.

3.8 Επίλογος

Στο παρόν κεφάλαιο παρουσιάστηκαν οι βασικές τεχνικές υποδειγματοληψίας που χρησιμοποιούνται για την αντιμετώπιση της ανισοκατανομής των κλάσεων στην κατηγοριοποίηση δεδομένων. Αρχικά, δόθηκε μια συνοπτική εισαγωγή στη φιλοσοφία των μεθόδων undersampling και στη διαφορά τους από τις τεχνικές υπερδειγματοληψίας.

Στη συνέχεια, αναλύθηκε η μέθοδος Random Undersampling, η οποία αποτελεί την απλούστερη προσέγγιση υποδειγματοληψίας, καθώς και οι τεχνικές που βασίζονται στους πλησιέστερους γείτονες, όπως οι Edited Nearest Neighbors και Repeated Edited Nearest Neighbors, οι οποίες επιχειρούν την απομάκρυνση θορυβωδών ή ασυνεπών δειγμάτων με βάση την τοπική δομή των δεδομένων.

Ακολούθως, παρουσιάστηκε η μέθοδος All kNN, η οποία επεκτείνει την ιδέα της ENN εξετάζοντας πολλαπλές τιμές του k, καθώς και η μέθοδος Condensed Nearest Neighbors, η οποία στοχεύει στη συμπύκνωση του συνόλου εκπαίδευσης διατηρώντας μόνο τα πλέον κρίσιμα δείγματα. Τέλος, εξετάστηκε η τεχνική Tomek Links, η οποία επικεντρώνεται στη στοχευμένη απομάκρυνση δειγμάτων που βρίσκονται κοντά στα όρια απόφασης μεταξύ των κλάσεων.

Η ποικιλία των τεχνικών υποδειγματοληψίας που παρουσιάστηκαν αναδεικνύει ότι, όπως και στην περίπτωση των τεχνικών υπερδειγματοληψίας, δεν υπάρχει μία καθολικά βέλτιστη λύση για όλα τα προβλήματα ανισοκατανομής. Κάθε μέθοδος παρουσιάζει διαφορετικά πλεονεκτήματα και περιορισμούς, γεγονός που καθιστά αναγκαία τη συγκριτική αξιολόγησή τους υπό κοινές πειραματικές συνθήκες. Στο επόμενο κεφάλαιο παρουσιάζονται οι αλγόριθμοι κατηγοριοποίησης που χρησιμοποιούνται στην εργασία και ειδικότερα ο ταξινομητής k-Nearest Neighbors, ο οποίος αποτελεί τη βάση της πειραματικής μελέτης.

Κεφάλαιο 4ο: Υλοποίηση σε Python

4.1 Εισαγωγή

Στο παρόν κεφάλαιο παρουσιάζεται η υλοποίηση του πειραματικού μέρους της εργασίας σε γλώσσα προγραμματισμού Python. Στόχος είναι η επεξήγηση του βασικού κώδικα που χρησιμοποιήθηκε για την επεξεργασία των δεδομένων, την αντιμετώπιση της ανισορροπίας κλάσεων, την εκπαίδευση των μοντέλων και την αξιολόγηση της απόδοσής τους.

Η παρουσίαση οργανώνεται με βάση τις κύριες βιβλιοθήκες που αξιοποιήθηκαν, συγκεκριμένα τις pandas, imbalanced-learn (imblearn) και scikit-learn (sklearn), αναδεικνύοντας τον ρόλο καθεμίας στα επιμέρους στάδια του πειραματικού pipeline. Για κάθε βιβλιοθήκη περιγράφονται οι βασικές λειτουργίες που χρησιμοποιήθηκαν, συνοδευόμενες από στοχευμένα και αντιπροσωπευτικά αποσπάσματα κώδικα.

Η προσέγγιση αυτή αποσκοπεί στην καλύτερη κατανόηση της υλοποίησης, στη διασύνδεση του κώδικα με τη μεθοδολογία που αναλύθηκε στο προηγούμενο κεφάλαιο και στη διασφάλιση της αναπαραγωγιμότητας των πειραματικών αποτελεσμάτων.

4.2 Βιβλιοθήκη pandas — Διαχείριση και προετοιμασία δεδομένων

Η βιβλιοθήκη pandas χρησιμοποιήθηκε σε όλα τα στάδια που αφορούν τη φόρτωση, την αρχική επεξεργασία και την οργάνωση των δεδομένων, καθώς και για τη συγκέντρωση και αποθήκευση των πειραματικών αποτελεσμάτων. Ο ρόλος της είναι θεμελιώδης, καθώς παρέχει μια ενιαία και ευέλικτη αναπαράσταση των datasets σε μορφή πινάκων (DataFrame), πάνω στην οποία βασίζεται ολόκληρο το πειραματικό pipeline.

Στις επόμενες υποενότητες παρουσιάζονται τα βασικά στάδια στα οποία αξιοποιήθηκε η pandas.

4.2.1 Φόρτωση datasets και αρχική επεξεργασία

Κάθε dataset φορτώνεται από αρχείο CSV και μετατρέπεται σε αντικείμενο DataFrame. Για λόγους ομοιομορφίας, η στήλη που αντιστοιχεί στη μεταβλητή στόχο μετονομάζεται σε Class, ανεξάρτητα από την αρχική της ονομασία. Με αυτόν τον τρόπο διευκολύνεται η επαναχρησιμοποίηση του ίδιου κώδικα σε όλα τα datasets.

```
importpandasaspd


df=pd.read_csv(dataset_path)
df=df.rename(columns={df.columns[-1]: "Class"})
```

Στη συνέχεια πραγματοποιείται ένας αρχικός έλεγχος της δομής του dataset (πλήθος δειγμάτων, πλήθος χαρακτηριστικών), καθώς και της κατανομής των κλάσεων, ώστε να επιβεβαιωθεί η ύπαρξη έντονης ανισορροπίας.

```
df.shape
df["Class"].value_counts()
```

Η παραπάνω πληροφορία είναι κρίσιμη, καθώς καθορίζει τη μετέπειτα ανάγκη εφαρμογής τεχνικών over-sampling και under-sampling.

4.2.2 Διαχωρισμός χαρακτηριστικών και ετικετών στόχου

Αφού ολοκληρωθεί η αρχική επεξεργασία, το dataset διαχωρίζεται σε:

- πίνακα χαρακτηριστικών εισόδου (X)
- διάνυσμα ετικετών (y)

Ο διαχωρισμός αυτός πραγματοποιείται με απλές λειτουργίες της pandas, διατηρώντας όλα τα χαρακτηριστικά εκτός της στήλης στόχου.

```
X= df.drop(columns=["Class"])
y= df["Class"]
```

Η χρήση της pandas σε αυτό το στάδιο επιτρέπει εύκολο χειρισμό τόσο αριθμητικών όσο και κατηγορικών χαρακτηριστικών, ανεξάρτητα από τον τύπο του dataset (αριθμητικό, κατηγορικό ή μικτό).

4.2.3 Υποστήριξη δυαδικής ταξινόμησης και ανάλυση κατανομής κλάσεων

Για τις ανάγκες της παρούσας εργασίας, όλα τα προβλήματα αντιμετωπίζονται ως δυαδικά προβλήματα ταξινόμησης. Η pandas χρησιμοποιείται για την ανάλυση της κατανομής των κλάσεων και τον εντοπισμό της μειοψηφικής κλάσης, η οποία αποτελεί και το κύριο ενδιαφέρον κατά την αξιολόγηση.

```
class_counts= y.value_counts()
minority_class=class_counts.idxmin()
```

Η πληροφορία αυτή αξιοποιείται έμμεσα στα επόμενα στάδια, καθώς οι μετρικές απόδοσης (Precision, Recall, F1-score) υπολογίζονται με έμφαση στη μειοψηφική κλάση.

4.2.4 Συγκέντρωση και αποθήκευση πειραματικών αποτελεσμάτων

Τέλος, η pandas χρησιμοποιείται για τη συγκέντρωση των αποτελεσμάτων από τα πολλαπλά πειράματα που εκτελούνται (διαφορετικά datasets και διαφορετικές τεχνικές δειγματοληψίας). Τα

αποτελέσματα αποθηκεύονται αρχικά σε δομές τύπου λίστας και στη συνέχεια μετατρέπονται σε DataFrame, διευκολύνοντας την ανάλυση και την εξαγωγή τους.

```
results= []
results.append({
   "Dataset": dataset_name,
   "Sampler": sampler_name,
   "Accuracy": accuracy,
   "Precision": precision,
   "Recall": recall,
   "F1-score": f1
})

results_df= pd.DataFrame(results)
```

Η τελική μορφή των αποτελεσμάτων μπορεί να εξαχθεί σε αρχείο CSV, ώστε να χρησιμοποιηθεί άμεσα στο κεφάλαιο παρουσίασης και ανάλυσης των πειραματικών αποτελεσμάτων.

```
results_df.to_csv("results.csv", index=False)
```

Μετά τη φόρτωση και την αρχική προετοιμασία των δεδομένων μέσω της pandas, το επόμενο κρίσιμο στάδιο του pipeline αφορά την αντιμετώπιση της ανισορροπίας κλάσεων. Στην ενότητα που ακολουθεί παρουσιάζεται ο ρόλος της βιβλιοθήκης **imbalanced-learn (imblearn)** και οι τεχνικές δειγματοληψίας που εφαρμόστηκαν αποκλειστικά στο σύνολο εκπαίδευσης.

4.3 Βιβλιοθήκη imbalanced-learn (imblearn) — Αντιμετώπιση ανισορροπίας κλάσεων

Η βιβλιοθήκη imbalanced-learn (imblearn) χρησιμοποιήθηκε για την αντιμετώπιση του προβλήματος της έντονης ανισορροπίας κλάσεων που χαρακτηρίζει όλα τα datasets της παρούσας εργασίας. Ο ρόλος της imblearn εντάσσεται αποκλειστικά στο στάδιο της προετοιμασίας του συνόλου εκπαίδευσης και προηγείται της εκπαίδευσης του ταξινομητή, με στόχο τη βελτίωση της ικανότητας γενίκευσης του μοντέλου ως προς τη μειοψηφική κλάση.

Στις επόμενες υποενότητες παρουσιάζονται οι βασικές τεχνικές over-sampling και under-sampling που αξιοποιήθηκαν, καθώς και ο τρόπος ενσωμάτωσής τους στο πειραματικό pipeline.

4.3.1 Τεχνικές over-sampling

Οι τεχνικές over-sampling στοχεύουν στην αύξηση του πλήθους των δειγμάτων της μειοψηφικής κλάσης, μέσω της δημιουργίας συνθετικών παραδειγμάτων. Στην παρούσα εργασία χρησιμοποιήθηκαν τόσο η βασική μέθοδος SMOTE, όσο και εξειδικευμένες παραλλαγές της, ανάλογα με τον τύπο των χαρακτηριστικών του εκάστοτε dataset.

Για datasets που περιέχουν αποκλειστικά αριθμητικά χαρακτηριστικά, εφαρμόστηκε η κλασική εκδοχή του SMOTE:

```
from imblearn.over_sampling import SMOTE

sampler= SMOTE(random_state=42)
X_train_res, y_train_res=sampler.fit_resample(X_train, y_train)
```

Στην περίπτωση datasets με αποκλειστικά κατηγορικά χαρακτηριστικά, χρησιμοποιήθηκε η μέθοδος SMOTEN, η οποία έχει σχεδιαστεί ειδικά για τέτοιου είδους δεδομένα:

```
from imblearn.over_sampling import SMOTEN

sampler= SMOTEN(random_state=42)
X_train_res, y_train_res=sampler.fit_resample(X_train, y_train)
```

Τέλος, για datasets με μικτά (αριθμητικά και κατηγορικά) χαρακτηριστικά, εφαρμόστηκε η μέθοδος SMOTENC, όπου δηλώνονται ρητά οι δείκτες των κατηγορικών χαρακτηριστικών:

```
from imblearn.over_sampling import SMOTENC

sampler= SMOTENC(
   categorical_features=categorical_indices,
   random_state=42
)
X_train_res, y_train_res=sampler.fit_resample(X_train, y_train)
```

Η επιλογή της κατάλληλης παραλλαγής του SMOTE σε κάθε περίπτωση διασφαλίζει τη σωστή δημιουργία συνθετικών δειγμάτων και την αποφυγή αλλοίωσης της δομής των δεδομένων.

4.3.2 Τεχνικές under-sampling

Οι τεχνικές under-sampling μειώνουν το πλήθος των δειγμάτων της πλειοψηφικής κλάσης, αφαιρώντας παραδείγματα που θεωρούνται πλεονάζοντα ή λιγότερο αντιπροσωπευτικά. Στην εργασία εφαρμόστηκαν διάφορες μέθοδοι under-sampling, με στόχο τη συγκριτική αξιολόγηση της επίδρασής τους στην απόδοση του ταξινομητή.

Ενδεικτικά, η χρήση του RandomUnderSampler υλοποιείται ως εξής:

```
from imblearn.under_sampling import RandomUnderSampler

sampler= RandomUnderSampler(random_state=42)
X_train_res, y_train_res=sampler.fit_resample(X_train, y_train)
```

Αντίστοιχα, χρησιμοποιήθηκαν και πιο σύνθετες μέθοδοι, όπως οι Tomek Links, Edited Nearest Neighbours και Condensed Nearest Neighbour, οι οποίες βασίζονται σε τοπικές σχέσεις γειτνίασης μεταξύ των δειγμάτων.

Όλες οι μέθοδοι under-sampling αξιοποιούν το ίδιο interface της imblearn (fit_resample), γεγονός που επιτρέπει την εναλλακτική χρήση τους χωρίς μεταβολή της υπόλοιπης ροής του κώδικα.

4.3.3 Ενσωμάτωση των τεχνικών sampling στο πειραματικό pipeline

Για τη συστηματική εκτέλεση των πειραμάτων, οι τεχνικές δειγματοληψίας οργανώθηκαν σε δομές τύπου λεξικού, επιτρέποντας την επαναληπτική εφαρμογή τους στο ίδιο σύνολο εκπαίδευσης και τη σύγκριση των αποτελεσμάτων υπό κοινές συνθήκες.

```
samplers= {
   "SMOTE": SMOTE(random_state=42),
   "RandomUnderSampler": RandomUnderSampler(random_state=42),
   "TomekLinks": TomekLinks()
}
```

Κατά την εκτέλεση των πειραμάτων, κάθε sampler εφαρμόζεται αποκλειστικά στο training set, ενώ το test set παραμένει ανέπαφο. Η πρακτική αυτή είναι κρίσιμη για την αποφυγή διαρροής πληροφορίας (data leakage) και τη διασφάλιση έγκυρης αξιολόγησης.

```
forname, samplerin samplers.items():
   X_res, y_res=sampler.fit_resample(X_train, y_train)
```

Με αυτόν τον τρόπο διασφαλίζεται ότι όλες οι τεχνικές sampling συγκρίνονται δίκαια, καθώς επηρεάζουν μόνο τη φάση εκπαίδευσης του μοντέλου.

Αφού ολοκληρωθεί η εξισορρόπηση του συνόλου εκπαίδευσης με τη χρήση της βιβλιοθήκης imbalanced-learn, το επόμενο στάδιο του πειραματικού pipeline αφορά την προεπεξεργασία των χαρακτηριστικών, την εκπαίδευση του ταξινομητή και την αξιολόγηση της απόδοσής του. Στην επόμενη ενότητα παρουσιάζεται ο ρόλος της βιβλιοθήκης **scikit-learn (sklearn)**, η οποία χρησιμοποιείται για τα παραπάνω στάδια.

4.4 Βιβλιοθήκη scikit-learn — Προεπεξεργασία, εκπαίδευση και αξιολόγηση

Η βιβλιοθήκη scikit-learn χρησιμοποιήθηκε για όλα τα στάδια του πειραματικού pipeline που αφορούν τον διαχωρισμό των δεδομένων, την προεπεξεργασία των χαρακτηριστικών, την εκπαίδευση του ταξινομητή και την αξιολόγηση της απόδοσής του. Αποτελεί τον βασικό κορμό της υλοποίησης, καθώς ενοποιεί τις διαδικασίες μάθησης και αξιολόγησης σε ένα συνεκτικό και αναπαραγώγιμο πλαίσιο.

Στις επόμενες υποενότητες παρουσιάζονται τα επιμέρους στάδια στα οποία αξιοποιήθηκε η scikit-learn.

4.4.1 Διαχωρισμός δεδομένων σε σύνολα εκπαίδευσης και ελέγχου

Αρχικά, το dataset διαχωρίζεται σε σύνολο εκπαίδευσης (training set) και σύνολο ελέγχου (test set), με χρήση στρωματοποιημένης δειγματοληψίας. Η στρωμάτωση διασφαλίζει ότι η αναλογία των κλάσεων διατηρείται και στα δύο σύνολα, κάτι ιδιαίτερα σημαντικό σε προβλήματα με έντονη ανισορροπία.

```
fromsklearn.model_selectionimporttrain_test_split

X_train, X_test, y_train, y_test=train_test_split(
   X, y,
   test_size=0.3,
   stratify=y,
   random_state=42
)
```

Ο σταθερός ορισμός της παραμέτρου random_state επιτρέπει την αναπαραγωγιμότητα των πειραμάτων.

4.4.2 Προεπεξεργασία χαρακτηριστικών

Για datasets με αριθμητικά χαρακτηριστικά, εφαρμόστηκε κανονικοποίηση μέσω της κλάσης StandardScaler. Η εκπαίδευση του scaler πραγματοποιείται αποκλειστικά στο σύνολο εκπαίδευσης, ενώ οι ίδιες παράμετροι εφαρμόζονται στο σύνολο ελέγχου, ώστε να αποφευχθεί διαρροή πληροφορίας.

```
fromsklearn.preprocessingimportStandardScaler

scaler=StandardScaler()
X_train_scaled=scaler.fit_transform(X_train)
X_test_scaled=scaler.transform(X_test)
```

Η προεπεξεργασία αυτή είναι ιδιαίτερα σημαντική για ταξινομητές που βασίζονται σε αποστάσεις, όπως ο k-Nearest Neighbors, καθώς διασφαλίζει ότι όλα τα χαρακτηριστικά συνεισφέρουν ισότιμα στον υπολογισμό της απόστασης.

4.4.3 Εκπαίδευσηταξινομητή k-Nearest Neighbors

Ως ταξινομητής επιλέχθηκε ο αλγόριθμος k-Nearest Neighbors (kNN), ο οποίος χρησιμοποιήθηκε με σταθερό αριθμό γειτόνων k = 3 σε όλα τα πειράματα. Η χρήση ενός ενιαίου ταξινομητή επιτρέπει τη δίκαιη σύγκριση των διαφορετικών τεχνικών δειγματοληψίας.

```
fromsklearn.neighborsimportKNeighborsClassifier


model=KNeighborsClassifier(n_neighbors=3)
model.fit(X_train_resampled, y_train_resampled)
```

Η εκπαίδευση του μοντέλου πραγματοποιείται στο resampled σύνολο εκπαίδευσης, το οποίο έχει προκύψει από την εφαρμογή των τεχνικών over-sampling ή under-sampling που παρουσιάστηκαν στην προηγούμενη ενότητα.

4.4.4 Παραγωγή προβλέψεων

Μετά την εκπαίδευση του ταξινομητή, παράγονται προβλέψεις για το σύνολο ελέγχου, το οποίο δεν έχει υποστεί καμία μορφή δειγματοληψίας.

```
y_pred= model.predict(X_test_scaled)
```

Η διάκριση αυτή μεταξύ training και test set είναι κρίσιμη, καθώς διασφαλίζει ότι η αξιολόγηση της απόδοσης βασίζεται σε δεδομένα που δεν έχουν επηρεαστεί από τη διαδικασία εκπαίδευσης.

4.4.5 Αξιολόγηση απόδοσης και υπολογισμός μετρικών

Η αξιολόγηση της απόδοσης των μοντέλων πραγματοποιήθηκε με τη χρήση τυπικών μετρικών ταξινόμησης, όπως η ακρίβεια (Accuracy), καθώς και οι Precision, Recall και F1-score. Ιδιαίτερη έμφαση δίνεται στις μετρικές που αφορούν τη μειοψηφική κλάση, καθώς αυτή αποτελεί το κύριο αντικείμενο ενδιαφέροντος σε προβλήματα ανισόρροπης ταξινόμησης.

```
fromsklearn.metricsimportaccuracy_score, classification_report


accuracy=accuracy_score(y_test, y_pred)
report=classification_report(y_test, y_pred, output_dict=True)
```

Από το αναλυτικό report εξάγονται οι επιμέρους μετρικές που αντιστοιχούν στη μειοψηφική κλάση, ώστε να είναι δυνατή η συγκριτική αξιολόγηση των διαφορετικών τεχνικών δειγματοληψίας.

```
precision= report[str(minority_class)]["precision"]
recall= report[str(minority_class)]["recall"]
f1= report[str(minority_class)]["f1-score"]
```

Οι παραπάνω διαδικασίες (εκπαίδευση, πρόβλεψη και αξιολόγηση) ενσωματώνονται σε έναν επαναληπτικό βρόχο, ο οποίος επιτρέπει την εκτέλεση πολλαπλών πειραμάτων υπό κοινές συνθήκες. Με αυτόν τον τρόπο διασφαλίζεται η συνεπής και συστηματική σύγκριση των αποτελεσμάτων.

Η scikit-learn παρέχει ένα συνεκτικό και ενιαίο interface που καθιστά εφικτή την αυτοματοποίηση της πειραματικής διαδικασίας, χωρίς απώλεια ελέγχου ή διαφάνειας.

4.5 Επίλογος

Στο παρόν κεφάλαιο παρουσιάστηκε αναλυτικά η υλοποίηση του πειραματικού pipeline σε Python, με έμφαση στον ρόλο των βασικών βιβλιοθηκών που χρησιμοποιήθηκαν. Η βιβλιοθήκη pandas αξιοποιήθηκε για τη διαχείριση και προετοιμασία των δεδομένων, η imbalanced-learn για την αντιμετώπιση της ανισορροπίας κλάσεων, και η scikit-learn για την προεπεξεργασία, την εκπαίδευση των μοντέλων και την αξιολόγηση της απόδοσής τους.

Η δομημένη αυτή προσέγγιση διασφαλίζει τη σαφή αντιστοίχιση του κώδικα με τη μεθοδολογία που παρουσιάστηκε στα προηγούμενα κεφάλαια και υποστηρίζει την αναπαραγωγιμότητα των πειραματικών αποτελεσμάτων.

Κεφάλαιο 5ο: Πειραματική Μελέτη

5.1 Εισαγωγή

Στο παρόν κεφάλαιο παρουσιάζεται η πειραματική μελέτη που πραγματοποιήθηκε με στόχο την αξιολόγηση της αποτελεσματικότητας διαφόρων τεχνικών δειγματοληψίας σε προβλήματα δυαδικής ταξινόμησης με έντονη ανισορροπία κλάσεων. Η μελέτη επικεντρώνεται στη σύγκριση μεθόδων υπερδειγματοληψίας (over-sampling) και υποδειγματοληψίας (under-sampling), εξετάζοντας την επίδρασή τους στην απόδοση ενός ενιαίου ταξινομητή ως προς τη μειοψηφική κλάση.

Αρχικά παρουσιάζονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν, τα οποία προέρχονται από το KEEL Dataset Repository και χαρακτηρίζονται από υψηλό δείκτη ανισορροπίας. Στη συνέχεια περιγράφεται η διαδικασία εγκαθίδρυσης των πειραμάτων, συμπεριλαμβανομένης της προεπεξεργασίας των δεδομένων, της εφαρμογής των τεχνικών δειγματοληψίας και της εκπαίδευσης του μοντέλου. Ακολουθεί η παρουσίαση των πειραματικών μετρήσεων, με χρήση κατάλληλων μετρικών αξιολόγησης, και τέλος πραγματοποιείται συζήτηση των αποτελεσμάτων, με έμφαση στα πλεονεκτήματα και τους περιορισμούς κάθε προσέγγισης.

Η δομημένη αυτή προσέγγιση επιτρέπει τη συστηματική ανάλυση της συμπεριφοράς των μεθόδων δειγματοληψίας σε διαφορετικά είδη ανισοκατανεμημένων συνόλων δεδομένων και παρέχει τη βάση για την εξαγωγή τεκμηριωμένων συμπερασμάτων.

5.2 Σύνολα δεδομένων

Η πειραματική μελέτη βασίστηκε σε δεκατρία (13) ανισοκατανεμημένα σύνολα δεδομένων, τα οποία έχουν προσαρμοστεί ώστε να διαμορφώνουν προβλήματα δυαδικής ταξινόμησης με έντονη ανισορροπία μεταξύ των κλάσεων. Όλα τα σύνολα δεδομένων που χρησιμοποιήθηκαν παρουσιάζουν δείκτη ανισορροπίας (imbalance ratio) μεγαλύτερο του 9, γεγονός που καθιστά την πρόβλεψη της μειοψηφικής κλάσης ιδιαίτερα απαιτητική.

Τα σύνολα δεδομένων προέρχονται από το KEEL Dataset Repository, μια ευρέως αναγνωρισμένη πηγή δεδομένων για την αξιολόγηση αλγορίθμων μηχανικής μάθησης σε προβλήματα ταξινόμησης με ανισορροπία κλάσεων. Η επιλογή τους έγινε με στόχο την κάλυψη διαφορετικών τύπων δεδομένων, τόσο ως προς τη φύση των χαρακτηριστικών όσο και ως προς τον βαθμό ανισορροπίας.

5.2.1 Περιγραφή δεδομένων

Ο Πίνακας 1 συνοψίζει τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην πειραματική μελέτη, καθώς και τα βασικά χαρακτηριστικά τους. Η ποικιλία ως προς τον βαθμό ανισορροπίας

επιτρέπει τη διερεύνηση της συμπεριφοράς των τεχνικών δειγματοληψίας σε διαφορετικά και απαιτητικά σενάρια.

Σύνολο Δεδομένων	**Περιγραφή**	**Πλήθος Εγγραφών**	**Δείκτης Ανισορροπίας**
glass2	Ταξινόμηση τύπων γυαλιού βάσει φυσικοχημικών ιδιοτήτων. Η μειοψηφική κλάση αφορά συγκεκριμένο τύπο γυαλιού.	214	11.59
ecoli4	Βιολογικό dataset για εντόπιση πρωτεϊνών στο κύτταρο, με σπάνια κατηγορία εντόπισης.	336	15.8
dermatology-6	Ιατρικό dataset σχετικό με δερματολογικές παθήσεις· η μειοψηφική κλάση αντιστοιχεί σε συγκεκριμένη διάγνωση.	358	16.9
poker-8-9_vs_5	Συνδυασμοί φύλλων πόκερ, όπου η μειοψηφική κλάση αφορά σπάνιο τύπο συνδυασμού.	2075	82
led7digit-0-2-4-5-6-7-8-9_vs_1	Αναγνώριση ψηφίων επτά τμημάτων·το ψηφίο “1” αποτελεί τη μειοψηφική κλάση.	443	10.97
cleveland-0_vs_4	Ιατρικό dataset καρδιολογικών δεδομένων, με μειοψηφική κλάση σοβαρής καρδιοπάθειας.	177	12.67
winequality-red-4	Ποιότητα κόκκινου κρασιού βάσει φυσικοχημικών χαρακτηριστικών·σπάνιο επίπεδο ποιότητας.	1599	29.17

Σύνολο Δεδομένων	Περιγραφή	Πλήθος Εγγραφών	Δείκτης Ανισορροπίας
yeast-2_vs_4	Βιολογικό dataset πρωτεϊνών ζύμης με σπάνια κατηγορία εντόπισης.	514	9.08
car-good	Αξιολόγηση αυτοκινήτων βάσει κατηγορικών χαρακτηριστικών·μειοψηφική κλάση υψηλής ποιότητας.	1728	24.04
flare-F	Δεδομένα ηλιακών εκλάμψεων, όπου η μειοψηφική κλάση αφορά έντονα φαινόμενα.	1066	23.79
kr-vs-k-zero_vs_eight	Θέσεις τελικού παιχνιδιού στο σκάκι·μειοψηφική κλάση σπάνιας διάταξης.	1460	53.07
abalone9-18	Βιολογικό dataset οστρακοειδών·μειοψηφική κλάση συγκεκριμένου ηλικιακού εύρους.	731	16.4
kddcup-buffer_overflow_vs_back	Dataset κυβερνοασφάλειας·μειοψηφική κλάση επιθέσεων τύπου bufferoverflow.	2233	73.43

Πίνακας 1: Σύνολα δεδομένων που χρησιμοποιήθηκαν στην πειραματική μελέτη

5.2.2 Κατηγοριοποίηση συνόλων δεδομένων

Για λόγους συστηματικής ανάλυσης, τα σύνολα δεδομένων κατηγοριοποιήθηκαν σε τρεις βασικές ομάδες, ανάλογα με τον τύπο των χαρακτηριστικών που περιέχουν:

- Σύνολα δεδομένων με μόνο αριθμητικά χαρακτηριστικά
- Σύνολα δεδομένων με μόνο κατηγορικά χαρακτηριστικά
- Σύνολα δεδομένων με μικτά (αριθμητικά και κατηγορικά) χαρακτηριστικά

Η κατηγοριοποίηση αυτή είναι κρίσιμη, καθώς επηρεάζει άμεσα την επιλογή των κατάλληλων τεχνικών υπερδειγματοληψίας, όπως οι παραλλαγές της μεθόδου SMOTE (SMOTE, SMOTEN, SMOTENC).

5.2.3 Σύνολα δεδομένων με μόνο αριθμητικά χαρακτηριστικά

Στην πρώτη κατηγορία ανήκουν οκτώ (8) σύνολα δεδομένων, τα οποία περιέχουν αποκλειστικά αριθμητικά χαρακτηριστικά. Ενδεικτικά αναφέρονται τα datasets glass2, ecoli4, dermatology-6 και poker-8-9_vs_5. Ο δείκτης ανισορροπίας σε αυτή την κατηγορία κυμαίνεται από 9.08 (yeast-2_vs_4) έως και 82 (poker-8-9_vs_5), υποδηλώνοντας περιπτώσεις εξαιρετικά έντονης ανισορροπίας.

Η υψηλή τιμή του imbalance ratio σε ορισμένα από τα σύνολα αυτά, όπως στο poker-8-9_vs_5, έχει ως αποτέλεσμα οι αρχικές (μη εξισορροπημένες) ταξινομήσεις να αποτυγχάνουν πλήρως στην αναγνώριση της μειοψηφικής κλάσης, γεγονός που καθιστά αναγκαία την εφαρμογή τεχνικών δειγματοληψίας.

5.2.4 Σύνολα δεδομένων με μόνο κατηγορικά χαρακτηριστικά

Η δεύτερη κατηγορία περιλαμβάνει τρία (3) σύνολα δεδομένων με αποκλειστικά κατηγορικά χαρακτηριστικά, συγκεκριμένα τα car-good, flare-F και kr-vs-k-zero_vs_eight. Τα σύνολα αυτά παρουσιάζουν ιδιαίτερα υψηλό δείκτη ανισορροπίας, ο οποίος φτάνει έως και 53.07 στην περίπτωση του kr-vs-k-zero_vs_eight.

Λόγω της φύσης των χαρακτηριστικών τους, τα σύνολα αυτά απαιτούν τη χρήση εξειδικευμένων τεχνικών υπερδειγματοληψίας, όπως η μέθοδος SMOTEN, η οποία έχει σχεδιαστεί ειδικά για κατηγορικά δεδομένα.

5.2.5 Σύνολα δεδομένων με μικτά χαρακτηριστικά

Η τρίτη κατηγορία περιλαμβάνει δύο (2) σύνολα δεδομένων με μικτά χαρακτηριστικά, αριθμητικά και κατηγορικά, συγκεκριμένα τα abalone9-18 και kddcup-buffer_overflow_vs_back. Και στις δύο περιπτώσεις παρατηρείται σημαντική ανισορροπία μεταξύ των κλάσεων, με τον δείκτη ανισορροπίας να φτάνει έως και 73.43 στο kddcup-buffer_overflow_vs_back.

Ιδιαίτερα στο τελευταίο dataset, η μειοψηφική κλάση εμφανίζεται περίπου 75 φορές λιγότερο συχνά από την πλειοψηφική, γεγονός που καθιστά τις παραδοσιακές μεθόδους ταξινόμησης ανεπαρκείς χωρίς προηγούμενη εξισορρόπηση των δεδομένων. Για την κατηγορία αυτή χρησιμοποιήθηκε η μέθοδος SMOTENC, η οποία επιτρέπει τον ταυτόχρονο χειρισμό αριθμητικών και κατηγορικών χαρακτηριστικών.

Η επιλογή συνόλων δεδομένων με διαφορετικά χαρακτηριστικά και διαφορετικούς βαθμούς ανισορροπίας επιτρέπει τη σφαιρική αξιολόγηση των τεχνικών δειγματοληψίας υπό μελέτη. Με τον τρόπο αυτό, καθίσταται δυνατή η διερεύνηση τόσο της γενικής συμπεριφοράς των μεθόδων όσο και των περιορισμών τους σε ακραία σενάρια ανισορροπίας.

Αφού παρουσιάστηκαν τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην πειραματική μελέτη, στην επόμενη ενότητα περιγράφεται η διαδικασία εγκαθίδρυσης των πειραμάτων, συμπεριλαμβανομένου του πειραματικού πρωτοκόλλου, των ρυθμίσεων του ταξινομητή και των μετρικών αξιολόγησης.

5.3 Εγκαθίδρυση πειραμάτων

Η πειραματική διαδικασία σχεδιάστηκε με στόχο τη δίκαιη και αναπαραγώγιμη σύγκριση τεχνικών δειγματοληψίας (over-sampling / under-sampling) σε προβλήματα δυαδικής ταξινόμησης με έντονη ανισορροπία κλάσεων. Για να τεκμηριωθεί η επιστημονική ορθότητα της μεθοδολογίας, τηρήθηκαν αρχές όπως:

a) σαφής ορισμός πειραματικών συνθηκών
b) αποφυγή διαρροής πληροφορίας (data leakage)
c) κοινό μοντέλο βάσης για όλες τις συγκρίσεις
d) σταθερές ρυθμίσεις τυχαιότητας για αναπαραγωγιμότητα (reproducibility).

5.3.1 Πειραματικό σενάριο και μεταβλητές σύγκρισης

Για κάθε σύνολο δεδομένων, διαμορφώθηκε ένα πειραματικό σενάριο όπου συγκρίνονται πολλαπλές τεχνικές εξισορρόπησης της κλάσης μειονότητας:

- **Over-sampling:** SMOTE (καιπαραλλαγές SMOTEN/SMOTENC), ADASYN, BorderlineSMOTE, KMeansSMOTE, SVMSMOTE
- **Under-sampling:** RandomUnderSampler, Edited Nearest Neighbours, Repeated Edited NN, All kNN, Condensed NN, Tomek Links

Η βασική ανεξάρτητη μεταβλητή είναι η μέθοδος δειγματοληψίας, ενώ όλες οι υπόλοιπες παράμετροι (μοντέλο ταξινόμησης, διαδικασία split, μετρικές αξιολόγησης) κρατήθηκαν σταθερές ώστε η σύγκριση να αποδίδει τη διαφορά στα αποτελέσματα στη δειγματοληψία και όχι σε άλλους παράγοντες.

5.3.2 Διαχωρισμός σε σύνολο εκπαίδευσης/ελέγχου

Για κάθε dataset εφαρμόστηκε στρωματοποιημένος διαχωρισμός (stratifiedsplit) σε:

- training set: 70%
- test set: 30%

Η στρωμάτωση διατηρεί την αρχική αναλογία κλάσεων και στα δύο σύνολα, γεγονός που κρίνεται απαραίτητο σε προβλήματα που διακρίνονται από έντονη ανισοκατανομή. Επιπλέον, ορίστηκε σταθερή τιμή random_state = 42 ώστε η διαδικασία να είναι πλήρως αναπαραγώγιμη (reproducible).

5.3.3 Στάδιο εφαρμογής τεχνικών δειγματοληψίας

Η επιστημονική εγκυρότητα της αξιολόγησης εξαρτάται καθοριστικά από το να μην επηρεάζεται το testset από τη διαδικασία εξισορρόπησης. Για τον λόγο αυτό:

1. πρώτα γίνεται ο διαχωρισμός train/test,
2. στη συνέχεια εφαρμόζεται η εκάστοτε τεχνική sampling αποκλειστικά στο trainingset,
3. το testset παραμένει ανέπαφο και χρησιμοποιείται μόνο για τελική αξιολόγηση.

Αυτός ο σχεδιασμός αποτρέπει τη διαρροή πληροφορίας (π.χ. δημιουργία συνθετικών δειγμάτων που «μοιάζουν» με δείγματα του testset), η οποία θα οδηγούσε σε τεχνητά βελτιωμένες επιδόσεις και μη αξιόπιστα συμπεράσματα.

5.3.4 Προεπεξεργασία και επιλογή κατάλληλης παραλλαγής SMOTEανά τύπο δεδομένων

Η πειραματική διαδικασία προσαρμόστηκε στον τύπο χαρακτηριστικών κάθε dataset, ώστε οι τεχνικές να εφαρμόζονται ορθά:

- **Αριθμητικά δεδομένα:** εφαρμόστηκε κλιμάκωση (StandardScaler) πριν από την εκπαίδευση, ώστε ο ταξινομητής kNN (distance-based) να μην επηρεάζεται δυσανάλογα από χαρακτηριστικά με μεγαλύτερη κλίμακα.
- **Μόνο κατηγορικά δεδομένα:** εφαρμόστηκε κατάλληλη κωδικοποίηση και χρησιμοποιήθηκε SMOTEN, που είναι σχεδιασμένο για nominal/categoricalfeatures.
- **Μικτά δεδομένα:** χρησιμοποιήθηκε SMOTENC με ρητό ορισμό των κατηγορικών γνωρισμάτων, ώστε η δημιουργία συνθετικών δειγμάτων να λαμβάνει υπόψη τη φύση κάθε χαρακτηριστικού.

Η επιλογή SMOTE/SMOTEN/SMOTENC τεκμηριώνεται και θεωρητικά, καθώς οι κλασικές SMOTE τεχνικές βασίζονται σε αποστάσεις/παρεμβολή, που είναι νοηματικά ορθές κυρίως για αριθμητικά δεδομένα.

5.3.5 Μοντέλο βάσης και δίκαιη σύγκριση

Ως κοινός ταξινομητής για όλα τα πειράματα επιλέχθηκε ο k-Nearest Neighbors (kNN) με k = 3. Η χρήση ενός ενιαίου μοντέλου σε όλα τα datasets και για όλες τις τεχνικές sampling διασφαλίζει ότι οι διαφοροποιήσεις στις μετρικές απόδοσης οφείλονται κατά κύριο λόγο στη δειγματοληψία.

Το kNN είναι επίσης κατάλληλο για την παρούσα μελέτη, διότι είναι ευαίσθητο σε αλλαγές στη γεωμετρία/κατανομή του training set, άρα αναδεικνύει με σαφή τρόπο την επίδραση του over/under-sampling.

5.3.6 Ροή πειραμάτων

Για κάθε dataset και για κάθε μέθοδο δειγματοληψίας ακολουθήθηκε η ίδια ακολουθία βημάτων:

1. Φόρτωση δεδομένων και ορισμός μεταβλητών εισόδου και εξόδου
2. Stratified train/test split (70%/30%)
3. (Όπου απαιτείται) προεπεξεργασία/κωδικοποίηση και scaling
4. Εφαρμογή sampling μόνο στο σύνολο εκπαίδευσης
5. Εκπαίδευση kNN στο resampledσύνολο εκπαίδευσης
6. Πρόβλεψη στο αμετάβλητο σύνολο ελέγχου
7. Υπολογισμός μετρικών και αποθήκευση αποτελεσμάτων

Αυτή η αυστηρά επαναλαμβανόμενη ροή εξασφαλίζει ότι οι πειραματικές συνθήκες παραμένουν σταθερές, άρα τα αποτελέσματα είναι συγκρίσιμα.

5.3.7 Μετρικές αξιολόγησης

Η αξιολόγηση πραγματοποιήθηκε με τις μετρικές:

- Accuracy
- Precision
- Recall
- F1-score

Δεδομένης της ανισορροπίας, το κύριο ενδιαφέρον είναι η απόδοση ως προς τη μειοψηφική κλάση, καθώς ένα μοντέλο μπορεί να εμφανίζει υψηλή accuracy προβλέποντας σχεδόν πάντα την πλειοψηφία. Για τον λόγο αυτό, οι Precision/Recall/F1 χρησιμοποιούνται για να αποτιμηθεί ουσιαστικά η ικανότητα εντοπισμού των σπάνιων (και συνήθως κρίσιμων) περιπτώσεων.

5.4 Πειραματικές μετρήσεις

5.4.1 Αποτελέσματα σε σύνολα δεδομένων με μόνο αριθμητικά χαρακτηριστικά

Στην παρούσα υποενότητα παρουσιάζονται τα αποτελέσματα της πειραματικής αξιολόγησης για τα σύνολα δεδομένων που περιέχουν αποκλειστικά αριθμητικά χαρακτηριστικά. Οι Πίνακες 2 - 14 συνοψίζουν την απόδοση του ταξινομητή για κάθε τεχνική δειγματοληψίας, τόσο πριν όσο και μετά την εφαρμογή over-sampling και under-sampling.

5.4.1.1 glass2 (IR: 11.59)

Μέθοδος	Accuracy	Precision	Recall	F1-score
Original	0.9296	1.0000	0.1667	0.2857

Μέθοδος	Accuracy	Precision	Recall	F1-score
Over-sampling				
SMOTE	0.9014	0.4545	0.8333	0.5882
ADASYN	0.8592	0.3571	0.8333	0.5000
BORDERLINESMOTE	0.8732	0.3846	0.8333	0.5263
KMEANSSMOTE	0.9014	0.4545	0.8333	0.5882
Under-sampling				
Random Undersampler	0.5634	0.1212	0.667	0.2051
Edited NN	0.9014	0.3333	0.1667	0.2222
Repeated Edited NN	0.9014	0.3333	0.1667	0.2222
All kNN	0.9014	0.3333	0.1667	0.2222
Condensed NN	0.8592	0.2500	0.3333	0.2857
Tomek Links	0.9296	1.0000	0.1667	0.2857

Πίνακας 2: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: glass2).

5.4.1.2 ecoli4 (IR: 15.8)

Μέθοδος	Accuracy	Precision	Recall	F1-score
Original	0.9910	1.0000	0.8571	0.9231
Over-sampling				
SMOTE	0.9910	1.0000	0.8571	0.9231
ADASYN	0.9910	1.0000	0.8571	0.9231
BORDERLINESMOTE	0.9820	1.0000	0.7143	0.8333
KMEANSSMOTE	0.9910	1.0000	0.8571	0.9231
SVMSMOTE	0.9910	1.0000	0.8571	0.9231
Under-sampling				
Random Undersampler	0.9640	0.6364	1.0000	0.7778
Edited NN	0.9910	1.0000	0.8571	0.9231

Repeated Edited NN	0.9910	1.0000	0.8571	0.9231
All kNN	0.9910	1.0000	0.8571	0.9231
Condensed NN	0.9820	0.7778	1.0000	0.8750
Tomek Links	0.9910	1.0000	0.8571	0.9231

Πίνακας 3: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: ecoli4).

5.4.1.3 dermatology-6 (IR: 16.9)

Μέθοδος	**Accuracy**	**Precision**	**Recall**	**F1-score**
Original	1.0000	1.0000	1.0000	1.0000
Over-sampling				
SMOTE	0.9832	0.7778	1.0000	0.8750
ADASYN	1.0000	1.0000	1.0000	1.0000
BORDERLINESMOTE	1.0000	1.0000	1.0000	1.0000
KMEANSSMOTE	0.9916	0.8750	1.0000	0.9333
SVMSMOTE	0.9916	0.8750	1.0000	0.9333
Under-sampling				
Random Undersampler	0.8824	0.3333	1.0000	0.5000
Edited NN	1.0000	1.0000	1.0000	1.0000
Repeated Edited NN	1.0000	1.0000	1.0000	1.0000
All kNN	1.0000	1.0000	1.0000	1.0000
Condensed NN	0.9580	0.5833	1.0000	0.7368
Tomek Links	1.0000	1.0000	1.0000	1.0000

Πίνακας 4: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: dermatology-6).

5.4.1.4 poker-8-9_vs_5 (IR: 82)

Μέθοδος	**Accuracy**	**Precision**	**Recall**	**F1-score**

Μέθοδος	**Accuracy**	**Precision**	**Recall**	**F1-score**
Original	0.9883	0.0000[1]	0.0000	0.0000
Over-sampling				
SMOTE	0.9066	0.0758	0.6250	0.1351
ADASYN	0.9066	0.0758	0.6250	0.1351
BORDERLINESMOTE	0.9606	0.1481	0.5000	0.2286
KMEANSSMOTE	0.9358	0.0714	0.3750	0.1200
SVMSMOTE	0.9606	0.1481	0.5000	0.2286
Under-sampling				
Random Undersampler	0.4058	0.0416	0.7500	0.0286
Edited NN	0.9883	0.0000	0.0000	0.0000
Repeated Edited NN	0.9883	0.0000	0.0000	0.0000
All kNN	0.9883	0.0000	0.0000	0.0000
Condensed NN	0.9752	0.0000	0.0000	0.0000
Tomek Links	0.9883	0.0000	0.0000	0.0000

Πίνακας 5: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: poker-8_9_vs_5).

5.4.1.5 led7digit-0-2-4-5-6-7-8-9_vs_1 (IR: 10.97)

Μέθοδος	**Accuracy**	**Precision**	**Recall**	**F1-score**
Original	0.9456	0.6429	0.7500	0.6923
Over-sampling				
SMOTE	0.9320	0.5833	0.5833	0.5833
ADASYN	0.9252	0.5385	0.5833	0.5600
BORDERLINESMOTE	0.9388	0.5789	0.9167	0.7097
KMEANSSMOTE	0.9320	0.5833	0.5833	0.5833
SVMSMOTE	0.9252	0.5294	0.7500	0.6207

[1] Το 0 σημαίνει ότι το μοντέλο ταξινομεί όλα τα δείγματα στην κλάση πλειονότητας (TP = 0).

Μέθοδος	Accuracy	Precision	Recall	F1-score
Under-sampling				
Random Undersampler	0.9252	0.5263	0.8333	0.6452
Edited NN	0.9388	0.5882	0.8333	0.6897
Repeated Edited NN	0.9388	0.6000	0.7500	0.6667
All kNN	0.9456	0.6250	0.8333	0.7143
Condensed NN	0.9456	0.8333	0.4167	0.5556
Tomek Links	0.9456	0.6429	0.7500	0.6923

Πίνακας 6: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: led7digit-0-2-4-5-6-7-8-9_vs_1).

5.4.1.6 cleveland-0_vs_4 (IR: 12.67)

Μέθοδος	Accuracy	Precision	Recall	F1-score
Original	0.9483	1.0000	0.2500	0.4000
Over-sampling				
SMOTE	0.9483	0.6000	0.7500	0.6667
ADASYN	0.9310	0.5000	0.7500	0.6000
BORDERLINESMOTE	0.9310	0.5000	0.5000	0.5000
KMEANSSMOTE	0.9310	0.5000	0.5000	0.5000
SVMSMOTE	0.9483	0.6000	0.7500	0.6667
Under-sampling				
Random Undersampler	0.8448	0.2727	0.7500	0.4000
Edited NN	0.9655	1.0000	0.5000	0.6667
Repeated Edited NN	0.9655	1.0000	0.5000	0.6667
All kNN	0.9655	1.0000	0.5000	0.6667
Condensed NN	0.9655	0.7500	0.7500	0.7500
Tomek Links	0.9655	1.0000	0.5000	0.6667

Πίνακας 7: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων:).

5.4.1.7 winequality-red-4 (IR: 29.17)

Μέθοδος	**Accuracy**	**Precision**	**Recall**	**F1-score**
Original	0.9583	0.1667	0.0556	0.0833
Over-sampling				
SMOTE	0.8693	0.1077	0.3889	0.1687
ADASYN	0.8674	0.1061	0.3889	0.1667
BORDERLINESMOTE	0.9072	0.1220	0.2778	0.1695
KMEANSSMOTE	0.9072	0.1395	0.3333	0.1967
SVMSMOTE	0.9129	0.1500	0.3333	0.2069
Under-sampling				
Random Undersampler	0.6061	0.0642	0.7778	0.1186
Edited NN	0.9489	0.0909	0.0556	0.0690
Repeated Edited NN	0.9489	0.0909	0.0556	0.0690
All kNN	0.9489	0.0909	0.0556	0.0690
Condensed NN	0.9337	0.1304	0.1667	0.1463
Tomek Links	0.9545	0.1250	0.0556	0.0769

Πίνακας 8: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: winequality-red-4).

5.4.1.8 yeast-2_vs_4 (IR: 9.08)

Μέθοδος	**Accuracy**	**Precision**	**Recall**	**F1-score**
Original	0.9765	1.0000	0.7647	0.8667
Over-sampling				
SMOTE	0.9647	0.7895	0.8824	0.8333
ADASYN	0.9294	0.6000	0.8824	0.7143
BORDERLINESMOTE	0.9353	0.6250	0.8824	0.7317

Μέθοδος	Accuracy	Precision	Recall	F1-score
KMEANSSMOTE	0.9471	0.7222	0.7647	0.7429
SVMSMOTE	0.9471	0.6818	0.8824	0.7692
Under-sampling				
Random Undersampler	0.9000	0.5000	0.8235	0.6222
Edited NN	0.9706	0.9286	0.7647	0.8387
Repeated Edited NN	0.9647	0.8667	0.7647	0.8125
All kNN	0.9647	0.8667	0.7647	0.8125
Condensed NN	0.9353	0.6500	0.7647	0.7027
Tomek Links	0.9765	1.0000	0.7647	0.8667

Πίνακας 9: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: yeast-2_vs_4).

Τα αποτελέσματα αυτά παρέχουν μια συγκεντρωτική εικόνα της επίδρασης των τεχνικών δειγματοληψίας σε αριθμητικά δεδομένα και αποτελούν τη βάση για τη συγκριτική ανάλυση που ακολουθεί.

5.4.2 Αποτελέσματα σε σύνολα δεδομένων με μόνο κατηγορικά χαρακτηριστικά

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα για τα σύνολα δεδομένων με αποκλειστικά κατηγορικά χαρακτηριστικά. Η ανάλυση περιλαμβάνει τόσο μεθόδους υπερδειγματοληψίας όσο και υποδειγματοληψίας, με έμφαση στη χρήση τεχνικών κατάλληλων για κατηγορικά δεδομένα.

5.4.2.1 car-good (IR: 24.04)

Μέθοδος	Accuracy	Precision	Recall	F1-score
Original	0.9667	0.7000	0.3043	0.4242
Over-sampling				
SMOTEN	0.9737	0.7000	0.6087	0.6512
ADASYN	0.9632	0.5227	1.0000	0.6866
BORDERLINESMOTE	0.9615	0.5116	0.9565	0.6667
KMEANSSMOTE	0.9720	0.5946	0.9565	0.7333

Μέθοδος	**Accuracy**	**Precision**	**Recall**	**F1-score**
SVMSMOTE	0.9475	0.4340	1.0000	0.6053
Under-sampling				
Random Undersampler	0.5867	0.0856	0.9565	0.1571
Edited NN	0.9685	0.8571	0.2609	0.4000
Repeated Edited NN	0.9650	0.5882	0.4348	0.5000
All kNN	0.9667	0.6250	0.4348	0.5128
Condensed NN	0.9702	0.5789	0.9565	0.7213
Tomek Links	0.9702	0.8750	0.3043	0.4516

Πίνακας 10: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: car-good).

5.4.2.2 flare-F (IR: 23.79)

Μέθοδος	**Accuracy**	**Precision**	**Recall**	**F1-score**
Original	0.9631	0.5556	0.3571	0.4348
Over-sampling				
SMOTEN	0.9659	0.6250	0.3571	0.4545
ADASYN	0.9375	0.3462	0.6429	0.4500
BORDERLINESMOTE	0.9403	0.3600	0.6429	0.4615
KMEANSSMOTE	0.9403	0.3333	0.5000	0.4000
SVMSMOTE	0.9432	0.3846	0.7143	0.5000
Under-sampling				
Random Undersampler	0.7614	0.1354	0.9286	0.2364
Edited NN	0.9176	0.2727	0.6429	0.3830
Repeated Edited NN	0.9148	0.2500	0.5714	0.3478
All kNN	0.9545	0.4167	0.3571	0.3846
Condensed NN	0.9574	0.4545	0.3571	0.4000
Tomek Links	0.9602	0.5000	0.3571	0.4167

Πίνακας 11: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: flare-F).

5.4.2.3 kr-vs-k-zero_vs_eight (IR: 53.07)

Μέθοδος	**Accuracy**	**Precision**	**Recall**	**F1-score**
Original	0.9979	1.0000	0.8889	0.9412
Over-sampling				
SMOTEN	0.9979	1.0000	0.8889	0.9412
ADASYN	1.0000	1.0000	1.0000	1.0000
BORDERLINESMOTE	1.0000	1.0000	1.0000	1.0000
KMEANSSMOTE	1.0000	1.0000	1.0000	1.0000
SVMSMOTE	0.9979	0.9000	1.0000	0.9474
Under-sampling				
Random Undersampler	0.7697	0.0750	1.0000	0.1395
Edited NN	0.9979	1.0000	0.8889	0.9412
Repeated Edited NN	0.9979	1.0000	0.8889	0.9412
All kNN	0.9979	1.0000	0.8889	0.9412
Condensed NN	0.9855	0.5714	0.888	0.6957
Tomek Links	0.9979	1.0000	0.8889	0.9412

Πίνακας 12: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: kr-vs-k-zero_vs_eight).

Οι πίνακες αποτυπώνουν τη συμπεριφορά των μεθόδων σε δεδομένα μη αριθμητικής φύσης, επιτρέποντας τη σύγκριση με τα αποτελέσματα της προηγούμενης κατηγορίας.

5.4.3 Αποτελέσματα σε σύνολα δεδομένων με μικτά χαρακτηριστικά

Ακολουθεί η παρουσίαση των αποτελεσμάτων για τα σύνολα δεδομένων με μικτά χαρακτηριστικά, αριθμητικά και κατηγορικά. Η κατηγορία αυτή είναι ιδιαίτερα απαιτητική, καθώς συνδυάζει διαφορετικούς τύπους γνωρισμάτων και υψηλά επίπεδα ανισορροπίας.

5.4.3.1 abalone9-18 (IR: 16.4)

Μέθοδος	Accuracy	Precision	Recall	F1-score
Original	0.9504	0.7500	0.2143	0.3333
Over-sampling				
SMOTENC	0.8595	0.1667	0.3571	0.2273
ADASYN	0.8802	0.2414	0.5000	0.3256
BORDERLINESMOTE	0.9008	0.1875	0.2143	0.2000
KMEANSSMOTE	0.9008	0.2222	0.2857	0.2500
SVMSMOTE	0.9050	0.2353	0.2857	0.2581
Under-sampling				
Random Undersampler	0.7438	0.1364	0.6429	0.2250
Edited NN	0.9504	0.7500	0.2143	0.3333
Repeated Edited NN	0.9504	0.7500	0.2143	0.3333
All kNN	0.9504	0.7500	0.2143	0.3333
Condensed NN	0.8967	0.2105	0.2857	0.2424
Tomek Links	0.9504	0.7500	0.2143	0.3333

Πίνακας 13: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: abalone9-18).

5.4.3.2 kddcup-buffer_overflow_vs_back (IR: 73.43)

Μέθοδος	Accuracy	Precision	Recall	F1-score
Original	0.9959	1.0000	0.7000	0.8235
Over-sampling				
SMOTENC	0.9959	1.0000	0.7000	0.8235
ADASYN	0.9959	1.0000	0.7000	0.8235
BORDERLINESMOTE	0.9959	1.0000	0.7000	0.8235
KMEANSSMOTE	0.9959	1.0000	0.7000	0.8235
SVMSMOTE	0.9959	1.0000	0.7000	0.8235

Μέθοδος	Accuracy	Precision	Recall	F1-score
Under-sampling				
Random Undersampler	0.9864	0.5000	0.8000	0.6154
Edited NN	0.9959	1.0000	0.7000	0.8235
Repeated Edited NN	0.9959	1.0000	0.7000	0.8235
All kNN	0.9959	1.0000	0.7000	0.8235
Condensed NN	0.1058[2]	0.0149	1.0000	0.0295
Tomek Links	0.9959	1.000	0.7000	0.8235

Πίνακας 14: Αποτελέσματα δυαδικής ταξινόμησης με διάφορες μεθόδους δειγματοληψίας (σύνολο δεδομένων: kddcup-buffer_overflow_vs_back).

Τα παρουσιαζόμενα αποτελέσματα αναδεικνύουν τη συμπεριφορά των τεχνικών δειγματοληψίας σε πιο σύνθετα σενάρια δεδομένων.

5.5 Συζήτηση

Η παρούσα ενότητα συνοψίζει και ερμηνεύει τα αποτελέσματα της πειραματικής αξιολόγησης τεχνικών υπερδειγματοληψίας (over-sampling) και υποδειγματοληψίας (under-sampling) σε ανισοκατανεμημένα σύνολα δεδομένων. Η συζήτηση εστιάζει κυρίως στην απόδοση ως προς τη μειοψηφική κλάση, όπως αυτή αποτυπώνεται από τις μετρικές Precision, Recall και F1-score, λαμβάνοντας παράλληλα υπόψη τη συνολική ακρίβεια (Accuracy).

5.5.1 Γενικές τάσεις και συμβιβασμοί μεταξύ μετρικών

Ένα επαναλαμβανόμενο μοτίβο που προκύπτει από τους πίνακες αποτελεσμάτων είναι ότι οι τεχνικές over-sampling τείνουν να αυξάνουν την ανάκληση (Recall) της μειοψηφικής κλάσης, συχνά με κόστος μείωσης της ακρίβειας (Accuracy) ή/και της ακρίβειας θετικών προβλέψεων (Precision). Αυτό είναι αναμενόμενο, καθώς η αύξηση των δειγμάτων της μειοψηφικής κλάσης μεταβάλλει το decision boundary υπέρ της ανίχνευσης περισσότερων θετικών περιπτώσεων, οδηγώντας συνήθως σε περισσότερα false positives. Παράλληλα, αρκετές τεχνικές under-sampling εμφανίζουν το αντίστροφο φαινόμενο: η δραστική μείωση της πλειοψηφικής κλάσης μπορεί να υποβαθμίσει τη γενίκευση, προκαλώντας αισθητή πτώση στην accuracy, ενώ η επίδραση στις μετρικές της μειοψηφικής κλάσης εξαρτάται έντονα από το dataset.

[2] Η συγκεκριμένη πολύ χαμηλή τιμή accuracy οφείλεται στο ότι ο αλγόριθμος CondensedNN γίνεται πολύ “aggressive” σε έντονα ανισόρροπο (imbalanced) dataset, με αποτέλεσμα να διατηρεί σχεδόν όλα τα στιγμιότυπα της κλάσης μειονότητας και ελάχιστης της κλάσης πλειονότητας. Έτσι, ο αλγόριθμος kNN (με k=3) καταλήγει να προβλέπει συνέχεια την κλάση μειονότητας, οδηγώντας σε ένα overall χαμηλό accuracy.

Ιδιαίτερη σημασία έχει ότι η Accuracy από μόνη της δεν επαρκεί ως κριτήριο σε έντονα ανισόρροπες συνθήκες. Σε αρκετές περιπτώσεις η baseline λύση (Original) εμφανίζει υψηλή accuracy αλλά πολύ χαμηλό recall/F1 για τη μειοψηφική κλάση, γεγονός που καταδεικνύει ότι το μοντέλο τείνει να ευνοεί την πλειοψηφία. Αυτό το φαινόμενο γίνεται ιδιαίτερα εμφανές σε datasets με πολύ υψηλό imbalance ratio.

5.5.2 Σύνολα δεδομένων με μόνο αριθμητικά χαρακτηριστικά

5.5.2.1.1 glass2

Στην αρχική μορφή του συνόλου δεδομένων (Original), παρατηρείται έντονα το φαινόμενο του «παράδοξου της ορθότητας» (accuracy paradox), καθώς η υψηλή τιμή του Accuracy (0.9296) συνυπάρχει με εξαιρετικά χαμηλό Recall (0.1667) για την κλάση μειονότητας. Παρόλο που το Precision είναι μέγιστο (1.0000), ο ταξινομητής αποτυγχάνει να εντοπίσει το μεγαλύτερο μέρος των κρίσιμων δειγμάτων, οδηγώντας σε έναν ιδιαίτερα χαμηλό δείκτη F1-score (0.2857). Η κατάσταση αυτή επιβεβαιώνει ότι, χωρίς παρέμβαση, το μοντέλο παρουσιάζει ισχυρή μεροληψία υπέρ της πολυπληθούς κλάσης, καθιστώντας τα αποτελέσματα πρακτικά μη αξιοποιήσιμα για την αναγνώριση της μειονότητας.

Η εφαρμογή τεχνικών Over-sampling επιφέρει τη σημαντικότερη βελτίωση στην απόδοση, με τις μεθόδους SMOTE και KMeansSMOTE να αναδεικνύονται ως οι πλέον αποτελεσματικές. Παρατηρείται κατακόρυφη αύξηση του Recall στο 0.8333, γεγονός που υποδηλώνει ότι η δημιουργία συνθετικών δειγμάτων επέτρεψε στον αλγόριθμο να αναγνωρίσει επιτυχώς τα πρότυπα της κλάσης μειονότητας. Παρά την αναμενόμενη μείωση του Precision λόγω της αύξησης των ψευδώς θετικών αποτελεσμάτων (False Positives), ο δείκτη F1-score υπερδιπλασιάζεται (0.5882), υποδεικνύοντας μια ανώτερη συνολική ισορροπία στη διαδικασία της ταξινόμησης.

Αντιθέτως, οι τεχνικές Under-sampling κρίνονται ως αναποτελεσματικές ή και επιζήμιες για το συγκεκριμένο σύνολο δεδομένων. Η μέθοδος Random Undersampler οδηγεί σε κατάρρευση τόσο του Accuracy όσο και του Precision, λόγω της εκτεταμένης απώλειας πληροφορίας από την πλειονότητα, ενώ οι μέθοδοι καθαρισμού θορύβου (όπως οι Edited NN και Tomek Links) αποτυγχάνουν να βελτιώσουν το Recall της κλάσης μειονότητας. Συμπερασματικά, το Under-sampling στο glass2 δεν καταφέρνει να προσφέρει την απαραίτητη ενίσχυση στη μάθηση του μοντέλου, διατηρώντας το F1-score σε χαμηλά επίπεδα ή υποβαθμίζοντας τη γενικότερη αξιοπιστία της πρόβλεψης.

5.5.2.1.2 ecoli4

Στην αρχική του μορφή (Original), το σύνολο δεδομένων επιδεικνύει εξαιρετικά υψηλή επίδοση, με το Accuracy να αγγίζει το 0.9910. Σε αντίθεση με άλλα ανισοκατανεμημένα σύνολα, εδώ παρατηρείται ότι το μοντέλο επιτυγχάνει ήδη υψηλό Recall (0.8571) και τέλειο Precision (1.0000),

οδηγώντας σε ένα πολύ ικανοποιητικό F1-score (0.9231). Αυτό υποδηλώνει ότι η κλάση μειονότητας στο ecoli4 είναι διακριτή και τα χαρακτηριστικά της επιτρέπουν στον ταξινομητή να τη διαχωρίζει αποτελεσματικά από την πλειονότητα, παρά την αριθμητική υπεροχή της δεύτερης.

Η εφαρμογή τεχνικών Over-sampling (SMOTE, ADASYN, κ.λπ.) δεν επιφέρει περαιτέρω βελτίωση στις μετρικές, με τις περισσότερες μεθόδους να διατηρούν τις επιδόσεις στα επίπεδα του αρχικού δείγματος. Αξιοσημείωτη εξαίρεση αποτελεί η μέθοδος BorderlineSMOTE, η οποία προκαλεί υποβάθμιση του Recall (0.7143) και κατ' επέκταση του F1-score (0.8333). Η συμπεριφορά αυτή υποδεικνύει ότι η δημιουργία συνθετικών δειγμάτων στα όρια των κλάσεων (decision boundaries) εισήγαγε θόρυβο σε ένα ήδη σαφώς καθορισμένο πρόβλημα ταξινόμησης, δυσχεραίνοντας αντί να διευκολύνοντας τη μάθηση.

Όσον αφορά τις τεχνικές Under-sampling, παρατηρείται μια ενδιαφέρουσα διαφοροποίηση. Ενώ οι μέθοδοι καθαρισμού (π.χ. Edited NN, Tomek Links) διατηρούν το μοντέλο σταθερό, οι μέθοδοι Random Undersampler και Condensed NN καταφέρνουν να μεγιστοποιήσουν το Recall (1.0000), εξασφαλίζοντας τον πλήρη εντοπισμό όλων των δειγμάτων της μειονότητας. Ωστόσο, αυτό επιτυγχάνεται με σημαντική θυσία στο Precision (πτώση στο 0.6364 και 0.7778 αντίστοιχα), γεγονός που μειώνει το συνολικό F1-score. Συμπερασματικά, για το συγκεκριμένο dataset, η αρχική κατανομή ή η ήπια υποδειγματοληψία φαίνεται να υπερτερεί, καθώς οι επιθετικές τεχνικές δειγματοληψίας διαταράσσουν την ήδη επιτυχημένη ισορροπία μεταξύ των μετρικών.

5.5.2.1.3 dermatology-6

Στην αρχική του κατάσταση (Original), το σύνολο δεδομένων παρουσιάζει την ιδεατή περίπτωση ταξινόμησης, με όλες τις μετρικές (Accuracy, Precision, Recall, F1-score) να αγγίζουν τη μέγιστη τιμή (1.0000). Το αποτέλεσμα αυτό υποδηλώνει ότι η κλάση μειονότητας είναι γραμμικά διαχωρίσιμη και απόλυτα διακριτή από την πλειονότητα στο featurespace. Στην περίπτωση αυτή, η ανισοκατανομή των δεδομένων δεν λειτουργεί ως εμπόδιο για τον αλγόριθμο, καθώς τα διαθέσιμα δείγματα επαρκούν για τον πλήρη καθορισμό των ορίων απόφασης.

Η εφαρμογή τεχνικών Over-sampling σε ένα ήδη «τέλειο» μοντέλο αποδεικνύεται εν μέρει πως αντενδείκνυται. Ενώ μέθοδοι όπως οι ADASYN και BorderlineSMOTE διατηρούν το απόλυτο σκορ, οι SMOTE, KMeansSMOTE και SVMSMOTE προκαλούν ελαφρά μείωση στο Accuracy και αισθητή πτώση στο Precision (έως και 0.7778 στην περίπτωση της SMOTE). Η παραγωγή συνθετικών δειγμάτων σε ένα τόσο καθαρό σύνολο δεδομένων εισάγει τεχνητή επικάλυψη (overlap) μεταξύ των κλάσεων, οδηγώντας σε εσφαλμένες προβλέψεις για την κλάση μειονότητας (False Positives) και υποβαθμίζοντας το F1-score.

Ανάλογη εικόνα παρατηρείται και στις τεχνικές Under-sampling, όπου η αφαίρεση δειγμάτων επηρεάζει αρνητικά την απόδοση. Συγκεκριμένα, η μέθοδος Random Undersampler επιφέρει τη μεγαλύτερη πτώση στο Precision (0.3333) και στο Accuracy (0.8824), ενώ η Condensed NN επίσης

μειώνει τη συνολική αξιοπιστία. Παρόλο που το Recall παραμένει σταθερό στο 1.0000, η δραματική μείωση του F1-score καθιστά αυτές τις μεθόδους ακατάλληλες. Αντιθέτως, οι μέθοδοι καθαρισμού (Edited NN, Tomek Links) διατηρούν το άριστο αποτέλεσμα, επιβεβαιώνοντας ότι η βέλτιστη στρατηγική για το dermatology-6 είναι η διατήρηση της αρχικής δομής των δεδομένων.

5.5.2.1.4 poker-8-9_vs_5

Στην αρχική του μορφή (Original), το σύνολο δεδομένων αναδεικνύει την απόλυτη επικράτηση του «παράδοξου της ορθότητας». Ενώ το Accuracy εμφανίζεται υψηλό (0.9883), οι μετρικές Precision, Recall και F1-score είναι μηδενικές (0.0000). Αυτό υποδηλώνει ότι ο ταξινομητής αγνοεί πλήρως την κλάση μειονότητας, κατατάσσοντας το σύνολο των δειγμάτων στην πλειοψηφική κλάση. Λόγω του εξαιρετικά υψηλού IR, η μειονότητα είναι πρακτικά «αόρατη» για το μοντέλο χωρίς τη χρήση τεχνικών εξισορρόπησης.

Η εφαρμογή τεχνικών Over-sampling είναι η μόνη προσέγγιση που καταφέρνει να προσδώσει στο μοντέλο μια στοιχειώδη ικανότητα εντοπισμού της μειονότητας. Οι μέθοδοι SMOTE και ADASYN επιτυγχάνουν το υψηλότερο Recall (0.6250), θυσιάζοντας ωστόσο σημαντικά το Precision (0.0758) και το Accuracy (0.9066). Οι τεχνικές BorderlineSMOTE και SVMSMOTE παρουσιάζουν την καλύτερη ισορροπία για αυτό το δύσκολο dataset, επιτυγχάνοντας το υψηλότερο F1-score (0.2286) με Recall 0.5000 και συγκριτικά καλύτερο Precision (0.1481). Η δημιουργία συνθετικών δειγμάτων αποδεικνύεται κρίσιμη, καθώς επιτρέπει στο μοντέλο να "ξεφύγει" από τον μηδενισμό των προβλέψεων.

Αντιθέτως, οι τεχνικές Under-sampling αποτυγχάνουν σχεδόν καθολικά να διαχειριστούν την ακραία ανισορροπία. Η μέθοδος Random Undersampler παρέχει μεν το υψηλότερο Recall (0.7500), οδηγεί όμως σε πλήρη κατάρρευση του Accuracy (0.4058) και του Precision (0.0416), καθιστώντας το μοντέλο αναξιόπιστο λόγω του τεράστιου αριθμού ψευδώς θετικών αποτελεσμάτων. Όλες οι υπόλοιπες μέθοδοι υποδειγματοληψίας και καθαρισμού (π.χ. Edited NN, Tomek Links, Condensed NN) διατηρούν μηδενικές επιδόσεις στις μετρικές της μειονότητας, αποδεικνύοντας ότι η απλή αφαίρεση δειγμάτων από την πλειονότητα δεν επαρκεί όταν η μειονότητα είναι τόσο περιορισμένη αριθμητικά.

5.5.2.1.5 led7digit-0-2-4-5-6-7-8-9_vs_1

Στην αρχική του μορφή (Original), το σύνολο δεδομένων επιδεικνύει μια σχετικά ισορροπημένη συμπεριφορά παρά την ανισοκατανομή, με το Accuracy στο 0.9456. Το Recall (0.7500) και το Precision (0.6429) υποδηλώνουν ότι ο ταξινομητής διαθέτει μια εγγενή ικανότητα να αναγνωρίζει την κλάση μειονότητας, επιτυγχάνοντας ένα αρχικό F1-score της τάξης του 0.6923. Η κατάσταση αυτή υποδεικνύει ότι τα δεδομένα δεν παρουσιάζουν ακραία επικάλυψη, επιτρέποντας στο μοντέλο να διακρίνει τα βασικά χαρακτηριστικά της μειονότητας χωρίς προεπεξεργασία.

Η εφαρμογή τεχνικών Over-sampling παρουσιάζει ανομοιογενή αποτελέσματα, με τις περισσότερες μεθόδους να οδηγούν σε υποβάθμιση της συνολικής απόδοσης. Εξαίρεση αποτελεί η μέθοδος BorderlineSMOTE, η οποία επιτυγχάνει τη μέγιστη τιμή Recall (0.9167), αν και με ελαφρά μείωση του Precision (0.5789), οδηγώντας στο υψηλότερο F1-score (0.7097) της συγκεκριμένης κατηγορίας. Αντιθέτως, οι μέθοδοι SMOTE, ADASYN και KMeansSMOTE προκαλούν ταυτόχρονη μείωση σε όλες τις μετρικές της μειονότητας, υποδηλώνοντας ότι η αλόγιστη δημιουργία συνθετικών δειγμάτων εισάγει θόρυβο που συγχέει τα όρια απόφασης του ταξινομητή.

Όσον αφορά τις τεχνικές Under-sampling, η μέθοδος All kNN αναδεικνύεται ως η βέλτιστη στρατηγική για το συγκεκριμένο dataset, επιτυγχάνοντας το υψηλότερο συνολικό F1-score (0.7143) μέσω της βελτίωσης του Recall στο 0.8333. Η μέθοδος Condensed NN, αν και αυξάνει σημαντικά το Precision (0.8333), προκαλεί δραματική πτώση στο Recall (0.4167), καθιστώντας την ακατάλληλη για τον εντοπισμό της μειονότητας. Συμπερασματικά, στο led7digit, οι μέθοδοι που εστιάζουν στον καθαρισμό των ορίων (όπως η All kNN) ή στην επιλεκτική ενίσχυση των οριακών δειγμάτων (BorderlineSMOTE) υπερτερούν έναντι των πιο γενικών προσεγγίσεων δειγματοληψίας.

5.5.2.1.6 cleveland-0_vs_4

Στην αρχική του μορφή (Original), το μοντέλο εμφανίζει υψηλό Accuracy (0.9483), το οποίο όμως συνοδεύεται από χαμηλό Recall (0.2500). Παρά το γεγονός ότι το Precision είναι απόλυτο (1.0000), η αδυναμία εντοπισμού των περισσότερων θετικών δειγμάτων οδηγεί σε έναν μέτριο δείκτη F1-score (0.4000). Η συμπεριφορά αυτή υποδηλώνει ότι ο ταξινομητής είναι εξαιρετικά συντηρητικός στις προβλέψεις του για την κλάση μειονότητας, αποφεύγοντας τα σφάλματα τύπου I (False Positives) αλλά αποτυγχάνοντας να καλύψει το εύρος των πραγματικών περιπτώσεων.

Η εφαρμογή τεχνικών Over-sampling επιφέρει σημαντική βελτίωση στην ικανότητα αναγνώρισης της μειονότητας. Οι μέθοδοι SMOTE και SVMSMOTE αναδεικνύονται ως οι πλέον αποδοτικές, καθώς τριπλασιάζουν το Recall (0.7500) διατηρώντας παράλληλα το Accuracy σταθερό (0.9483). Αν και το Precision υποχωρεί στο 0.6000, η συνολική ισορροπία βελτιώνεται αισθητά, με το F1-score να ανέρχεται στο 0.6667. Αντιθέτως, οι μέθοδοι BorderlineSMOTE και KMeansSMOTE φαίνονται λιγότερο αποτελεσματικές, καθώς επιτυγχάνουν χαμηλότερο Recall (0.5000) και υποβαθμίζουν περαιτέρω το F1-score.

Στην κατηγορία του Under-sampling, παρατηρείται μια ενδιαφέρουσα τάση βελτίωσης μέσω του καθαρισμού των δεδομένων. Οι μέθοδοι Edited NN, Repeated Edited NN, All kNN και Tomek Links αυξάνουν το Accuracy στο 0.9655 και διπλασιάζουν το Recall στο 0.5000, διατηρώντας το Precision στο μέγιστο (1.0000). Ωστόσο, η μέθοδος Condensed NN επιτυγχάνει το βέλτιστο F1-score (0.7500) όλης της δοκιμής, συνδυάζοντας υψηλό Recall (0.7500) με ικανοποιητικό Precision (0.7500). Αντιθέτως, το Random Undersampler αποδεικνύεται αναποτελεσματικό, καθώς η μεγάλη πτώση στο Precision (0.2727) ακυρώνει τα κέρδη από την αύξηση της ανάκλησης.

Συνολικά για το cleveland-0_vs_4, οι τεχνικές επιλεκτικής υποδειγματοληψίας (Condensed NN) και η στοχευμένη υπερδειγματοληψία (SMOTE/SVMSMOTE) αποτελούν τις βέλτιστες στρατηγικές για την ενίσχυση της προβλεπτικής ικανότητας του μοντέλου.

5.5.2.1.7 winequality-red-4

Στην αρχική του μορφή (Original), το σύνολο δεδομένων εμφανίζει ένα υψηλό Accuracy (0.9583), το οποίο όμως είναι πλασματικό. Οι μετρικές Precision (0.1667) και Recall (0.0556) είναι εξαιρετικά χαμηλές, υποδηλώνοντας ότι το μοντέλο αδυνατεί σχεδόν ολοκληρωτικά να εντοπίσει την κλάση μειονότητας. Το αποτέλεσμα είναι ένας πολύ χαμηλός δείκτης F1-score (0.0833), γεγονός που καθιστά τον αρχικό ταξινομητή ανεπαρκή για τη συγκεκριμένη εφαρμογή λόγω της έντονης μεροληψίας του υπέρ της πλειοψηφικής κλάσης.

Η εφαρμογή τεχνικών Over-sampling επιφέρει μια γενικευμένη, αν και συγκρατημένη, βελτίωση στις μετρικές απόδοσης της μειονότητας. Οι μέθοδοι SMOTE και ADASYN καταφέρνουν να αυξήσουν σημαντικά το Recall (0.3889), θυσιάζοντας όμως ένα μέρος του Accuracy και του Precision. Η μέθοδος SVMSMOTE αναδεικνύεται ως η βέλτιστη στην κατηγορία αυτή, επιτυγχάνοντας το υψηλότερο F1-score (0.2069), καθώς καταφέρνει να διατηρήσει το Precision στο 0.1500 ενώ παράλληλα ενισχύει την ανάκληση. Συνολικά, η υπερδειγματοληψία βοηθά το μοντέλο να "μάθει" καλύτερα τα σπάνια δείγματα, παρά την υψηλή δυσκολία του συγκεκριμένου dataset.

Αντιθέτως, οι τεχνικές Under-sampling παρουσιάζουν απογοητευτική εικόνα, με εξαίρεση τη μέθοδο Random Undersampler. Η τελευταία επιτυγχάνει το υψηλότερο Recall (0.7778) όλης της δοκιμής, αλλά με καταστροφικό κόστος στο Precision (0.0642) και στο Accuracy (0.6061), οδηγώντας σε ένα μη λειτουργικό μοντέλο λόγω του υπερβολικού αριθμού ψευδώς θετικών αποτελεσμάτων. Οι υπόλοιπες μέθοδοι καθαρισμού (π.χ. Edited NN, Tomek Links) αποτυγχάνουν να βελτιώσουν ουσιαστικά την ανάκληση, διατηρώντας το F1-score σε πολύ χαμηλά επίπεδα. Συμπερασματικά, για το winequality-red-4, η υπερδειγματοληψία μέσω SVMSMOTE φαίνεται να προσφέρει την πιο ορθολογική προσέγγιση για τη διαχείριση της ανισορροπίας.

5.5.2.1.8 yeast-2_vs_4

Στην αρχική του μορφή (Original), το σύνολο δεδομένων επιδεικνύει υψηλή απόδοση, με το Accuracy στο 0.9765 και ένα πολύ ικανοποιητικό Recall στο 0.7647. Το γεγονός ότι το Precision είναι μέγιστο (1.0000) οδηγεί σε έναν ισχυρό δείκτη F1-score (0.8667), υποδηλώνοντας ότι ο ταξινομητής μπορεί να αναγνωρίσει με επιτυχία την κλάση μειονότητας χωρίς να παράγει ψευδώς θετικά αποτελέσματα (FalsePositives). Η κατανομή των δεδομένων επιτρέπει στον αλγόριθμο να διακρίνει σαφώς τα όρια των κλάσεων.

Η εφαρμογή τεχνικών Over-sampling καταφέρνει να ενισχύσει περαιτέρω το Recall, ανεβάζοντάς το στο 0.8824 στις περισσότερες μεθόδους (SMOTE, ADASYN, BorderlineSMOTE,

SVMSMOTE). Ωστόσο, αυτή η αύξηση της ευαισθησίας συνοδεύεται από αναπόφευκτη μείωση του Precision, με αποτέλεσμα το F1-score να υποχωρεί ελαφρώς σε σύγκριση με την αρχική κατάσταση (π.χ. 0.8333 για τη μέθοδο SMOTE). Η υπερδειγματοληψία σε αυτό το dataset φαίνεται να "πιέζει" το μοντέλο προς μια πιο επιθετική αναγνώριση της μειονότητας, η οποία όμως εισάγει μικρό ποσοστό σφάλματος στις προβλέψεις.

Στις τεχνικές Under-sampling, παρατηρείται παρόμοια τάση υποβάθμισης της συνολικής απόδοσης. Η μέθοδος RandomUndersampler αυξάνει το Recall (0.8235), αλλά προκαλεί σημαντική πτώση στο Precision (0.5000) και στο F1-score (0.6222). Αντιθέτως, οι μέθοδοι καθαρισμού (EditedNN, RepeatedEditedNN, AllkNN) διατηρούν υψηλά επίπεδα Accuracy και Precision, αλλά αποτυγχάνουν να βελτιώσουν το Recall, με αποτέλεσμα το F1-score να παραμένει χαμηλότερο από το αρχικό. Η μέθοδος TomekLinks είναι η μόνη που διατηρεί την απόλυτη επίδοση του Original δείγματος, επιβεβαιώνοντας ότι η παρέμβαση στα δεδομένα του yeast-2_vs_4 δεν κρίνεται απαραίτητη.

Συμπερασματικά, για το σύνολο δεδομένων yeast-2_vs_4, η αρχική κατανομή προσφέρει τη βέλτιστη ισορροπία μεταξύ των μετρικών, καθώς οι τεχνικές δειγματοληψίας, παρόλο που αυξάνουν την ανάκληση, διαταράσσουν την εξαιρετική ακρίβεια του μοντέλου.

5.5.3 Σύνολα δεδομένων με μόνο κατηγορικά χαρακτηριστικά

5.5.3.1.1 car-good

Στην αρχική του μορφή (Original), το σύνολο δεδομένων εμφανίζει υψηλό Accuracy (0.9667), το οποίο όμως συγκαλύπτει τη μέτρια απόδοση στην κλάση μειονότητας. Το Recall περιορίζεται στο 0.3043, υποδηλώνοντας ότι το μοντέλο αδυνατεί να εντοπίσει το μεγαλύτερο μέρος των θετικών δειγμάτων, ενώ το Precision (0.7000) οδηγεί σε ένα μέτριο F1-score (0.4242). Η συμπεριφορά αυτή επιβεβαιώνει ότι η ανισορροπία των κλάσεων εμποδίζει τον ταξινομητή από το να γενικεύσει αποτελεσματικά για τη σπάνια κλάση.

Η εφαρμογή τεχνικών Over-sampling επιφέρει δραστική και ουσιαστική βελτίωση σε όλες τις μετρικές της μειονότητας. Οι μέθοδοι ADASYN και SVMSMOTE επιτυγχάνουν το απόλυτο Recall (1.0000), εξασφαλίζοντας τον πλήρη εντοπισμό των δειγμάτων της κλάσης "good". Ωστόσο, η μέθοδος KMeansSMOTE αναδεικνύεται ως η πλέον ισορροπημένη, επιτυγχάνοντας το υψηλότερο F1-score (0.7333) με εξαιρετικό Recall (0.9565) και το υψηλότερο Precision (0.5946) της κατηγορίας. Η δημιουργία συνθετικών δειγμάτων στο συγκεκριμένο dataset αποδεικνύεται ιδιαίτερα αποδοτική, ενισχύοντας τη μάθηση χωρίς να προκαλεί υπερβολικό θόρυβο.

Στον αντίποδα, οι τεχνικές Under-sampling παρουσιάζουν ανομοιογενή αποτελέσματα. Η μέθοδος Random Undersampler οδηγεί σε κατάρρευση του Accuracy (0.5867) και του Precision (0.0856), καθιστώντας το μοντέλο πρακτικά άχρηστο λόγω των υπερβολικών ψευδώς θετικών

προβλέψεων. Αντιθέτως, η μέθοδος Condensed NN παρουσιάζει εξαιρετική επίδοση, αγγίζοντας ένα F1-score της τάξης του 0.7213, το οποίο ανταγωνίζεται άμεσα τις μεθόδους υπερδειγματοληψίας. Οι υπόλοιπες μέθοδοι καθαρισμού (π.χ. Edited NN, Tomek Links) αποτυγχάνουν να βελτιώσουν αισθητά το Recall, διατηρώντας την απόδοση σε χαμηλά επίπεδα.

Συνολικά για το car-good, η χρήση KMeansSMOTE ή Condensed NN αποτελεί τη βέλτιστη στρατηγική, καθώς οι τεχνικές αυτές καταφέρνουν να εξισορροπήσουν την ανάγκη για υψηλή ανάκληση χωρίς να θυσιάζουν υπερβολικά την ακρίβεια των προβλέψεων.

5.5.3.1.2 flare-F

Στην αρχική του μορφή (Original), το σύνολο δεδομένων παρουσιάζει υψηλό Accuracy (0.9631), το οποίο όμως οφείλεται στην κυριαρχία της πλειοψηφικής κλάσης. Το Recall περιορίζεται στο 0.3571, υποδηλώνοντας ότι το μοντέλο αδυνατεί να αναγνωρίσει δύο στα τρία δείγματα της κλάσης μειονότητας, ενώ το Precision (0.5556) και το F1-score (0.4348) επιβεβαιώνουν τη μέτρια συνολική απόδοση του ταξινομητή προ οποιασδήποτε παρέμβασης.

Η εφαρμογή τεχνικών Over-sampling επιφέρει θετική επίδραση στην αναγνώριση της μειονότητας, με τη μέθοδο SVMSMOTE να αναδεικνύεται ως η πλέον αποτελεσματική, επιτυγχάνοντας το υψηλότερο F1-score (0.5000) και σημαντική αύξηση του Recall στο 0.7143. Οι μέθοδοι ADASYN και BorderlineSMOTE ενισχύουν επίσης την ανάκληση (0.6429), αλλά με μεγαλύτερο κόστος στο Precision (περίπου 0.35), γεγονός που υποδεικνύει την εισαγωγή θορύβου στα όρια απόφασης. Η μέθοδος SMOTEN διατηρεί την αρχική ανάκληση αλλά βελτιώνει την ακρίβεια, δείχνοντας μια πιο συντηρητική προσέγγιση στην παραγωγή συνθετικών δεδομένων.

Στην κατηγορία του Under-sampling, η μέθοδος Random Undersampler μεγιστοποιεί το Recall (0.9286), οδηγώντας όμως σε σημαντική υποβάθμιση του Accuracy (0.7614) και του Precision (0.1354), καθιστώντας τις προβλέψεις μη αξιόπιστες λόγω των πολλών ψευδώς θετικών δειγμάτων. Οι τεχνικές καθαρισμού και επιλεκτικής υποδειγματοληψίας (π.χ. Edited NN, Condensed NN, Tomek Links) αποτυγχάνουν να προσφέρουν ουσιαστικό πλεονέκτημα, καθώς είτε διατηρούν το Recall σε χαμηλά επίπεδα είτε υποβαθμίζουν το F1-score σε σύγκριση με την αρχική κατάσταση.

Συμπερασματικά για το flare-F, η τεχνική SVMSMOTE αποτελεί τη βέλτιστη επιλογή, καθώς επιτυγχάνει τη δικαιότερη ισορροπία μεταξύ της ανάγκης για εντοπισμό της μειονότητας και της διατήρησης της ακρίβειας του μοντέλου.

5.5.3.1.3 kr-vs-k-zero_vs_eight

Στην αρχική του μορφή (Original), το σύνολο δεδομένων επιδεικνύει εξαιρετικά υψηλή απόδοση, με το Accuracy να ανέρχεται στο 0.9979. Παρά την έντονη ανισοκατανομή, το μοντέλο επιτυγχάνει ήδη πολύ υψηλό Recall (0.8889) και τέλειο Precision (1.0000), γεγονός που οδηγεί σε ένα ιδιαίτερα ισχυρό F1-score (0.9412). Τα αποτελέσματα αυτά υποδηλώνουν ότι η κλάση μειονότητας

είναι σαφώς διαχωρίσιμη στο featurespace, επιτρέποντας στον ταξινομητή να την εντοπίζει με μεγάλη ακρίβεια χωρίς προηγούμενη παρέμβαση.

Η εφαρμογή τεχνικών Over-sampling καταφέρνει να μεγιστοποιήσει τις επιδόσεις του μοντέλου. Συγκεκριμένα, οι μέθοδοι ADASYN, BorderlineSMOTE και KMeansSMOTE οδηγούν σε απόλυτες τιμές (1.0000) σε όλες τις μετρικές (Accuracy, Precision, Recall, F1-score), εξαλείφοντας πλήρως κάθε σφάλμα ταξινόμησης. Η συνθετική ενίσχυση της μειονότητας φαίνεται να προσφέρει την απαραίτητη πυκνότητα πληροφορίας ώστε ο αλγόριθμος να ορίσει με απόλυτη ακρίβεια τα όρια απόφασης. Ακόμη και η μέθοδος SVMSMOTE, παρόλο που αυξάνει το Recall στο 1.0000, προκαλεί μια ελαφρά πτώση στο Precision (0.9000), αναδεικνύοντας τη λεπτή ισορροπία μεταξύ ευαισθησίας και ακρίβειας.

Αντιθέτως, οι τεχνικές Under-sampling κρίνονται ως περιττές ή και επιζήμιες για το συγκεκριμένο dataset. Η μέθοδος Random Undersampler επιφέρει δραματική υποβάθμιση του Accuracy (0.7697) και του Precision (0.0750), παρά το γεγονός ότι επιτυγχάνει απόλυτο Recall. Η απώλεια πληροφορίας από την πλειονότητα οδηγεί σε έναν εξαιρετικά χαμηλό δείκτη F1-score (0.1395). Οι μέθοδοι καθαρισμού (π.χ. Edited NN, Tomek Links) διατηρούν τις επιδόσεις της αρχικής κατάστασης, επιβεβαιώνοντας ότι δεν υπάρχουν επικαλυπτόμενα δείγματα που να δυσχεραίνουν την ταξινόμηση.

Συμπερασματικά για το kr-vs-k-zero_vs_eight, ενώ η αρχική κατανομή είναι ήδη επαρκής, η υπερδειγματοληψία (κυρίως μέσω ADASYN και KMeansSMOTE) οδηγεί στην τέλεια μοντελοποίηση του προβλήματος, σε αντίθεση με την υποδειγματοληψία η οποία υποβαθμίζει σημαντικά την αξιοπιστία του μοντέλου.

5.5.4 Σύνολα δεδομένων με μικτά χαρακτηριστικά

Στα datasets με μικτά χαρακτηριστικά παρατηρείται εντονότερα η εξάρτηση των αποτελεσμάτων από το συνδυασμό: (α) τύπος χαρακτηριστικών, (β) βαθμός ανισορροπίας, και (γ) συμπεριφορά της επιλεγμένης μεθόδου sampling.

Στο abalone9-18 (IR 16.4) η baseline επίδοση στη μειοψηφική κλάση είναι χαμηλή (Recall 0.2143, F1 0.3333). Οι τεχνικές over-sampling (SMOTENC, ADASYN κ.λπ.) αυξάνουν γενικά το recall (έως 0.5000), ωστόσο η precision μειώνεται σημαντικά σε ορισμένες περιπτώσεις (π.χ. 0.1667 με SMOTENC), με αποτέλεσμα το F1 να μην βελτιώνεται ουσιαστικά ή να παραμένει σε παρόμοια επίπεδα (π.χ. 0.3256 με ADASYN). Η συμπεριφορά αυτή υποδεικνύει ότι η αύξηση των θετικών προβλέψεων δεν αρκεί από μόνη της· απαιτείται και επαρκής ακρίβεια στις θετικές προβλέψεις ώστε το F1 να ενισχυθεί.

Η πλέον χαρακτηριστική περίπτωση είναι το kddcup-buffer_overflow_vs_back (IR 73.43). Παρά την εξαιρετικά υψηλή ανισορροπία, η baseline απόδοση είναι ήδη υψηλή (F1 = 0.8235), και οι

περισσότερες τεχνικές over-sampling παρουσιάζουν πρακτικά ταυτόσημα αποτελέσματα. Αυτό μπορεί να υποδηλώνει ότι τα διαθέσιμα χαρακτηριστικά παρέχουν ισχυρό σήμα για τη μειοψηφική κλάση, ή/και ότι το μοντέλο έχει ήδη βρει μια ικανοποιητική διαχωριστική περιοχή.

Ωστόσο, το ίδιο dataset αναδεικνύει και τους κινδύνους του under-sampling. Η μέθοδος Condensed NN οδηγεί σε εξαιρετικά χαμηλή accuracy (0.10583) και πολύ χαμηλό F1 (0.0295), ενώ ταυτόχρονα εμφανίζει recall 1.0. Η συμπεριφορά αυτή συνδέεται με υπερβολικά "επιθετική" αφαίρεση δειγμάτων της πλειοψηφικής κλάσης, η οποία προκαλεί κατάρρευση της ισορροπίας της πληροφορίας στο training set και οδηγεί τον ταξινομητή να προβλέπει σχεδόν πάντα τη μειοψηφική κλάση. Η παρατήρηση αυτή τεκμηριώνεται και στο συνοδευτικό σχόλιο του πίνακα αποτελεσμάτων.

5.5.5 Συνολική αποτίμηση και πρακτικές επισημάνσεις

Συνολικά, τα αποτελέσματα καταδεικνύουν ότι:

- Η δειγματοληψία είναι ιδιαίτερα ωφέλιμη όταν η baseline αποτυγχάνει στη μειοψηφική κλάση, ακόμη κι αν η accuracy εμφανίζεται υψηλή (π.χ. poker-8-9_vs_5).
- Σε datasets με ήδη υψηλή επίδοση στη μειοψηφική κλάση, η επιπλέον δειγματοληψία συχνά δεν προσφέρει ουσιαστικό όφελος και ενδέχεται να εισάγει θόρυβο (π.χ. ecoli4, yeast-2_vs_4, kr-vs-k-zero_vs_eight).
- Οι μέθοδοι under-sampling απαιτούν μεγαλύτερη προσοχή, καθώς μπορούν να αφαιρέσουν κρίσιμη πληροφορία της πλειοψηφικής κλάσης και να οδηγήσουν σε αστάθεια ή και κατάρρευση της απόδοσης, ειδικά σε ακραίες ανισορροπίες (χαρακτηριστικό παράδειγμα το Condensed NN στο kddcup-buffer_overflow_vs_back).
- Η βελτίωση σε Recall δεν συνεπάγεται αυτομάτως βελτίωση σε F1-score, καθώς το F1 επηρεάζεται ταυτόχρονα από precision και recall. Αυτό εξηγεί γιατί σε ορισμένα datasets οι τεχνικές που "ανεβάζουν" το recall εμφανίζουν περιορισμένη συνολική πρόοδο στο F1.
- Με βάση τα παραπάνω, η επιλογή τεχνικής δειγματοληψίας δεν μπορεί να θεωρηθεί καθολική, αλλά εξαρτάται από: (α) το επίπεδο ανισορροπίας, (β) τον τύπο χαρακτηριστικών, (γ) το πόσο καλά αποδίδει ήδη το baseline, και (δ) το επιθυμητό trade-off μεταξύ precision και recall, ανάλογα με το εφαρμοστικό πλαίσιο.

Κεφάλαιο 6ο: Συμπεράσματα και Μελλοντική Έρευνα

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκε συστηματικά η επίδραση τεχνικών υπερδειγματοληψίας και υποδειγματοληψίας στην απόδοση μοντέλων δυαδικής ταξινόμησης σε ανισοκατανεμημένα σύνολα δεδομένων. Η πειραματική αξιολόγηση πραγματοποιήθηκε σε δεκατέσσερα σύνολα δεδομένων διαφορετικού τύπου και βαθμού ανισορροπίας, με στόχο την εξαγωγή γενικών και τεκμηριωμένων συμπερασμάτων σχετικά με τη χρησιμότητα και τους περιορισμούς των εξεταζόμενων μεθόδων.

Ένα από τα βασικά συμπεράσματα της μελέτης είναι ότι η συνολική ακρίβεια (Accuracy) δεν αποτελεί αξιόπιστο κριτήριο αξιολόγησης σε προβλήματα με έντονη ανισορροπία κλάσεων. Σε πολλές περιπτώσεις παρατηρήθηκε ότι μοντέλα με υψηλή accuracy αποτυγχάνουν να αναγνωρίσουν επαρκώς τη μειοψηφική κλάση, γεγονός που καθιστά αναγκαία τη χρήση μετρικών όπως η ανάκληση (Recall), η ακρίβεια θετικών προβλέψεων (Precision) και ο δείκτης F1-score.

Τα πειραματικά αποτελέσματα έδειξαν ότι οι τεχνικές υπερδειγματοληψίας μπορούν να βελτιώσουν σημαντικά την απόδοση ως προς τη μειοψηφική κλάση, ιδιαίτερα σε περιπτώσεις όπου η αρχική (baseline) απόδοση είναι χαμηλή. Σε datasets με υψηλό ή πολύ υψηλό δείκτη ανισορροπίας, η εφαρμογή τεχνικών όπως οι παραλλαγές της μεθόδου SMOTE οδήγησε σε αισθητή αύξηση της ανάκλησης και, σε αρκετές περιπτώσεις, σε βελτίωση του F1-score, έστω και με μερική θυσία της συνολικής ακρίβειας. Το εύρημα αυτό υπογραμμίζει τη σημασία της εξισορρόπησης όταν ο στόχος είναι η ανίχνευση σπάνιων αλλά κρίσιμων περιπτώσεων.

Αντίθετα, σε σύνολα δεδομένων όπου η baseline επίδοση στη μειοψηφική κλάση ήταν ήδη υψηλή, η εφαρμογή τεχνικών δειγματοληψίας δεν παρείχε ουσιαστικό όφελος και, σε ορισμένες περιπτώσεις, οδήγησε ακόμη και σε ελαφρά υποβάθμιση της απόδοσης. Το αποτέλεσμα αυτό καταδεικνύει ότι η δειγματοληψία δεν αποτελεί πανάκεια και ότι η ανάγκη εφαρμογής της εξαρτάται από τη δομή των δεδομένων και τον βαθμό διαχωρισιμότητας των κλάσεων.

Όσον αφορά τις τεχνικές υποδειγματοληψίας, τα αποτελέσματα ανέδειξαν τη μεγαλύτερη αστάθεια και τον αυξημένο κίνδυνο απώλειας κρίσιμης πληροφορίας της πλειοψηφικής κλάσης. Ιδιαίτερα σε περιπτώσεις έντονης ανισορροπίας, ορισμένες μέθοδοι under-sampling οδήγησαν σε σημαντική υποβάθμιση της συνολικής απόδοσης ή σε ακραία συμπεριφορά του ταξινομητή. Το εύρημα αυτό υποδεικνύει ότι οι τεχνικές αυτές απαιτούν προσεκτική εφαρμογή και δεν είναι πάντοτε κατάλληλες για datasets με πολύ υψηλό imbalance ratio.

Ένα ακόμη σημαντικό συμπέρασμα αφορά την εξάρτηση της απόδοσης των τεχνικών από τον τύπο των χαρακτηριστικών. Η χρήση εξειδικευμένων παραλλαγών της μεθόδου SMOTE για κατηγορικά και μικτά δεδομένα αποδείχθηκε αναγκαία για τη σωστή εφαρμογή της

υπερδειγματοληψίας, επιβεβαιώνοντας ότι η φύση των δεδομένων αποτελεί καθοριστικό παράγοντα στην επιλογή της κατάλληλης μεθόδου.

Συνοψίζοντας, τα αποτελέσματα της εργασίας καταδεικνύουν ότι η αποτελεσματική αντιμετώπιση της ανισορροπίας κλάσεων προϋποθέτει συνδυασμό κατάλληλων μετρικών αξιολόγησης, προσεκτικής επιλογής τεχνικής δειγματοληψίας και κατανόησης των ιδιαιτεροτήτων του εκάστοτε συνόλου δεδομένων. Η μελέτη επιβεβαιώνει ότι δεν υπάρχει μία καθολικά βέλτιστη λύση, αλλά ότι η επιλογή της κατάλληλης προσέγγισης εξαρτάται από το εφαρμοστικό πλαίσιο και τους στόχους της ανάλυσης.

6.2 Μελλοντική έρευνα

Τα αποτελέσματα της παρούσας διπλωματικής εργασίας αναδεικνύουν μια σειρά από ζητήματα και περιορισμούς που μπορούν να αποτελέσουν αντικείμενο περαιτέρω διερεύνησης στο πλαίσιο μελλοντικής έρευνας. Αν και η μελέτη παρείχε χρήσιμα συμπεράσματα σχετικά με τη συμπεριφορά τεχνικών δειγματοληψίας σε ανισοκατανεμημένα δεδομένα, υπάρχουν αρκετές κατευθύνσεις στις οποίες η ανάλυση θα μπορούσε να επεκταθεί.

Μια προφανής κατεύθυνση για μελλοντική εργασία αφορά τη χρήση και σύγκριση διαφορετικών ταξινομητών. Στην παρούσα μελέτη χρησιμοποιήθηκε ένας ενιαίος ταξινομητής (k-Nearest Neighbors), ώστε να διασφαλιστεί η δίκαιη σύγκριση των τεχνικών δειγματοληψίας. Ωστόσο, η διερεύνηση της συμπεριφοράς των ίδιων τεχνικών σε συνδυασμό με άλλους αλγορίθμους, όπως δέντρα αποφάσεων, τυχαία δάση, υποστηρικτές διανυσμάτων (SVM) ή νευρωνικά δίκτυα, θα μπορούσε να αποκαλύψει διαφορετικά μοτίβα και να οδηγήσει σε πιο γενικεύσιμα συμπεράσματα.

Επιπλέον, η παρούσα εργασία επικεντρώθηκε σε συγκεκριμένες μετρικές αξιολόγησης, με έμφαση στη μειοψηφική κλάση. Μελλοντικές μελέτες θα μπορούσαν να ενσωματώσουν επιπλέον μετρικές, όπως το ROC-AUC, το PR-AUC, το G-mean ή τη balanced accuracy, ώστε να εξεταστεί πληρέστερα η απόδοση των μοντέλων σε διαφορετικές πτυχές της ανισοκατανεμημένης ταξινόμησης.

Μια ακόμη κατεύθυνση αφορά τη χρήση εναλλακτικών ή υβριδικών προσεγγίσεων αντιμετώπισης της ανισορροπίας. Συνδυασμοί over-sampling και under-sampling, προσαρμοσμένες (cost-sensitive) μέθοδοι μάθησης ή τεχνικές ensemble που ενσωματώνουν μηχανισμούς εξισορρόπησης θα μπορούσαν να προσφέρουν καλύτερη ισορροπία μεταξύ precision και recall, ιδιαίτερα σε datasets με ακραίο imbalance ratio.

Επιπρόσθετα, η διερεύνηση της επίδρασης της επιλογής υπερπαραμέτρων τόσο των τεχνικών δειγματοληψίας όσο και των ταξινομητών αποτελεί ένα ακόμη πεδίο μελλοντικής έρευνας. Στην παρούσα μελέτη υιοθετήθηκαν σταθερές ρυθμίσεις, με στόχο τη συγκρισιμότητα των πειραμάτων. Ωστόσο, η βελτιστοποίηση των παραμέτρων μέσω διαδικασιών όπως cross-validation ή grid search

ενδέχεται να οδηγήσει σε διαφοροποιημένα αποτελέσματα και βαθύτερη κατανόηση της συμπεριφοράς των μεθόδων.

Τέλος, μια ενδιαφέρουσα προοπτική είναι η εφαρμογή της πειραματικής διαδικασίας σε μεγαλύτερα ή πιο σύνθετα σύνολα δεδομένων, καθώς και σε πραγματικά δεδομένα από συγκεκριμένους τομείς εφαρμογής, όπως η ιατρική διάγνωση, η ανίχνευση απάτης ή η κυβερνοασφάλεια. Η μελέτη τέτοιων περιπτώσεων θα μπορούσε να αναδείξει πρακτικές προκλήσεις που δεν εμφανίζονται σε ελεγχόμενα πειραματικά περιβάλλοντα και να ενισχύσει τη χρησιμότητα των συμπερασμάτων σε πραγματικές συνθήκες.

Συνολικά, η παρούσα εργασία θέτει ένα σταθερό πλαίσιο για τη μελέτη της ανισοκατανεμημένης ταξινόμησης και μπορεί να αποτελέσει αφετηρία για περαιτέρω έρευνα, με στόχο την ανάπτυξη πιο αποδοτικών και αξιόπιστων μεθόδων αντιμετώπισης της ανισορροπίας κλάσεων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[3] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[4] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.

[5] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[6] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[7] G. M. Weiss, "Mining with rarity: A unifying framework," *SIGKDD Explorations*, vol. 6, no. 1, pp. 7–19, Jun. 2004.

[8] S. Wang and X. Yao, "Relationships between diversity of classification ensembles and single-class performance measures," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 206–219, Jan. 2013.

[9] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends," *Information Sciences*, vol. 250, pp. 113–141, Nov. 2013.

[10] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," Journal of Artificial Intelligence Research, vol. 61, pp. 863–905, 2018.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[12] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explorations*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[13] A. Fernández, V. López, M. Galar, M. J. del Jesus, and F. Herrera, "Analyzing the classification of imbalanced data sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-Based Systems*, vol. 42, pp. 97–110, Apr. 2013.

[14] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Machine Learning Research*, vol. 11, pp. 2795–2800, 2011.

[15] G. Lemaitre, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning,” *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[16] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,”

[17] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning,” *Advances in Intelligent Computing*, vol. 3644, pp. 878–887, 2005.

[18] F. Last, G. Douzas, and F. Bacao, “K-Means SMOTE: A clustering approach for imbalanced classification,” *Neurocomputing*, vol. 275, pp. 173–190, Jan. 2018.

[19] H. M. Nguyen, E. W. Cooper, and K. Kamei, “Borderline over-sampling for imbalanced data classification,” *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, 2011.

[20] D. L. Wilson, “Asymptotic properties of nearest neighbor rules using edited data,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 3, pp. 408–421, Jul. 1972.

[21] I. Tomek, “Two modifications of CNN,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, Nov. 1976.

[22] I. Tomek, “An experiment with the edited nearest-neighbor rule,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 6, pp. 448–452, Jun. 1976.

[23] P. E. Hart, “The condensed nearest neighbor rule,” *IEEE Transactions on Information Theory*, vol. IT-14, no. 3, pp. 515–516, May 1968.

[24] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[25] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[26] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[27] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.

[28] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[30] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[31] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2009.