ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΔΠΜΣ
ΕΠΙΣΤΗΜΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

«Ταξινόμηση δεδομένων καρκίνου του μαστού με χρήση στατιστικών μεθόδων και μεθόδων μηχανικής μάθησης»

Διπλωματική Εργασία
της
Χριστίνας Γιαννακουδάκη
Α.Μ. 03400126

Επιβλέπουσα: Δρ. Χρυσηίς Καρώνη-Ρίτσαρντσον, Καθηγήτρια

Τριμελής Εξεταστική Επιτροπή:

Χ. Καρώνη-Ρίτσαρντσον	Β. Παπανικολάου	Κ. Χρυσαφίνος
Ομ. Καθηγήτρια	Ομ. Καθηγητής	Καθηγητής

Αθήνα, Ιούνιος 2023

Στους γονείς μου, Γιάννη και Ελένη
και στον αδερφό μου, Γιώργη

....

Χριστίνα Γιαννακουδάκη

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. ΄Εχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

Ευχαριστίες

Η εκπόνηση της παρούσας μεταπτυχιακής εργασίας, η οποία υλοποιήθηκε κατά το ακαδημαϊκό έτος 2022-2023 στο ΔΠΜΣ Επιστήμης Δεδομένων και Μηχανικής Μάθησης της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, στηρίχθηκε στην συμβολή πολλών και σημαντικών ανθρώπων. Αρχικά, θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια μου, την Καθηγήτρια Δρ. Χρυσηίς Καρώνη, η οποία με καθοδήγησε συστημάτικά κατά την διάρκεια εκπόνησης της παρούσας εργασίας. Επιπλέον, θα ήθελα να ευχαριστήσω τους γονείς μου και τον αδερφό μου για την συνεχή στήριξη καθώς επίσης και τους φίλους μου για την διαρκή υποστήριξη και τις συμβουλές που μου παρείχαν καθ'όλη την περίοδο της μελέτης και της υλοποίησης της διπλωματικής εργασίας.

Contents

List of Figures

List of Tables

Περίληψη

Μια από τις πιο κοινές και θανατηφόρες ασθένειες είναι ο καρκίνος που πλήττει εκατομμύρια άτομα παγκοσμίως. Η ανίχνευση και η θεραπεία του καρκίνου έχουν προχωρήσει σημαντικά χάρη στην ιατρική έρευνα, αλλά το θέμα εξακολουθεί να είναι περίπλοκο και δύσκολο και απαιτεί συνεχείς βελτιώσεις στη μεθοδολογία και την τεχνολογία. Οι αλγόριθμοι μηχανικής μάθησης έχουν πρόσφατα επιδείξει σημαντικές δυνατότητες στον τομέα της ιατρικής έρευνας, ιδίως στην ανάλυση τεράστιου όγκου δεδομένων για τη διάγνωση κακοήθων όγκων. Ωστόσο, η ποιότητα και η ισορροπία των δεδομένων εκπαίδευσης έχουν σημαντικό αντίκτυπο στο πόσο καλά αποδίδουν αυτοί οι αλγόριθμοι. Οι αλγόριθμοι μηχανικής μάθησης μπορεί να αποδίδουν ανεπαρκώς σε μη ισορροπημένα σύνολα δεδομένων όπου ο ένας τύπος όγκου υπερισχύει έναντι του άλλου, οδηγώντας σε μοντέλα που δεν αναγνωρίζουν τη μειοψηφική κατηγορία. Η συγκεκριμένη διπλωματική εργασία εστιάζει στην εφαρμογή μεθόδων δειγματοληψίας, όπως η τυχαία υπερδειγματοληψία, η τυχαία υποδειγματοληψία, και οι τεχνικές SMOTE και ADASYN, για την εξισορρόπηση της κλάσης μειοψηφίας, με στόχο την παραγωγή μοντέλων που αναγνωρίζουν επαρκώς τόσο τους καρκινικούς όσο και τους καλοήθεις όγκους.Η απόδοση τριών ταξινομητών, συγκεκριμένα των Decision Trees, των Random Forests και του XGBoost, αξιολογήθηκε με τη χρήση αυτών των τεχνικών δειγματοληψίας και συγκρίθηκε με τους ίδιους ταξινομητές χωρίς δειγματοληψία. Επιπλέον, για να εκτιμηθεί ο αντίκτυπος της μη αντιμετώπισης του προβλήματος της ανισορροπίας των κλάσεων, δύο μοντέλα, συγκεκριμένα το Multilayer Perceptron και η λογιστική παλινδρόμηση LASSO με επιλογή χαρακτηριστικών, εφαρμόστηκαν στο σύνολο δεδομένων χωρίς δειγματοληψία και εξετάστηκε η απόδοσή τους. Η καλύτερη επιτευχθείσα ακρίβεια τόσο με όσο και χωρίς τεχνικές δειγματοληψίας ξεπέρασε το 96 % στο σύνολο δοκιμών.

Λέξεις-κλειδιά: καρκίνος του μαστού, μηχανική μάθηση, ταξινόμηση, ανισορροπία συνόλου δεδομένων, δειγματοληψία, νευρωνικά δίκτυα.

Abstract

One of the most common and deadly diseases is cancer, which affects millions of people worldwide. The detection and treatment of cancer have advanced significantly thanks to medical research, but the subject is still complex and challenging and requires continuous improvements in methodology and technology. Machine learning algorithms have recently demonstrated significant potential in the field of medical research, in particular in analysing huge amounts of data for the diagnosis of malignant tumours. However, the quality and balance of training data have a significant impact on how well these algorithms perform. Machine learning algorithms may perform poorly on imbalanced datasets where one tumor type predominates over another, leading to models that do not recognize the minority category. This thesis focuses on the application of sampling methods for balancing the minority class, namely Random Oversampling and Undersampling, SMOTE and ADASYN, with the goal of producing models that adequately identify both malignant and benign tumors. The performance of three classifiers, specifically Decision Trees, Random Forests, and XGBoost, was evaluated using these sampling techniques and compared to the same classifiers without sampling. Additionally, to assess the impact of not addressing the class imbalance problem, two models, namely the Multilayer Perceptron and the LASSO logistic regression with feature selection, were applied to the utilized dataset without sampling, and their performance was examined. The best achieved accuracy both with and without sampling techniques surpassed 96 % on the test set.

Keywords: breast cancer, machine learning, classification, dataset imbalance, sampling, neural networks.

1 Introduction

1.1 Breast Cancer

Breast cancer is a prevalent form of cancer affecting women, accounting for more than 10% of new cancer cases annually. It is the second leading cause of cancer-related deaths among women worldwide. Anatomically, the breast's milk-producing glands are located in front of the chest wall, supported by ligaments connecting them to the chest wall and resting on the pectoralis major muscle. The breast is composed of 15-20 lobes arranged in a circular pattern, and their size and shape are determined by the fat surrounding the lobes (Alkabban and Ferguson (2022)).

Each lobe consists of lobules, which contain glands responsible for milk production when stimulated by hormones. Breast cancer typically develops silently, and many individuals become aware of the disease through routine screenings. However, it can also present as a breast lump discovered accidentally, changes in breast size or contour, or nipple discharge. Mastalgia, or breast pain, is a common condition but is not necessarily indicative of breast cancer (Alkabban and Ferguson (2022)).

Diagnosing breast cancer involves a physical examination, imaging techniques such as mammography, and a tissue biopsy. Early detection plays a crucial role in improving survival rates. The aggressive spread of breast cancer through the lymphatic and hematological systems can lead to poor prognosis and distant metastasis. Therefore, the importance of breast cancer screening initiatives is underscored by these factors (Alkabban and Ferguson (2022)).

1.1.1 Pathophysiology

Breast cancer is caused by genetic mutations and DNA damage, both of which can be impacted by estrogen exposure. Sometimes, DNA flaws or cancer-causing genes like BRCA1 and BRCA2 are inherited. Therefore, having ovarian or breast cancer in the family raises the risk of developing breast cancer. In a healthy person, cells with aberrant DNA or abnormal development are attacked by the immune system. When breast cancer patients experience this failure, tumors develop and spread (Alkabban and Ferguson (2022)).

1.1.2 Etiology

In general health screening for women, determining characteristics linked to a higher risk of breast cancer development is crucial. Seven major categories can be used to classify breast cancer risk factors (Alkabban and Ferguson (2022)):

1. **Age**: The age-adjusted incidence of breast cancer continues to increase with the advancing age of the female population.
2. **Gender**: Most breast cancers occur in women.
3. **Personal history of breast cancer**: A history of cancer in one breast increases the likelihood of a second primary cancer in the contralateral breast.
4. **Histologic risk factors**: Breast biopsy histologic abnormalities are a significant group of breast cancer risk factors. These abnormalities include proliferative alterations with atypia and lobular carcinoma in situ (LCIS).
5. **The family history of breast cancer and genetic risk factors**: First-degree relatives of breast cancer patients have a 2- to 3-fold increased risk of getting the illness. Genetic

factors may be the cause of 5% to 10% of all breast cancer occurrences, but they may also be the cause of 25% of instances in women under the age of 30. The two most significant genes linked to an elevated risk of breast cancer are BRCA1 and BRCA2.

6. **Reproductive risk factors**: Women are assumed to have an increased chance of developing breast cancer after reproductive milestones that raise their lifetime estrogen exposure. Menarche starting before the age of 12, the first live birth occurring after the age of 30, nulliparity, and menopause occurring after the age of 55 are some of these.

7. **Exogenous hormone use**: The two most frequent uses of therapeutic or supplementary estrogen and progesterone are contraception in premenopausal women and hormone replacement treatment in postmenopausal women.

1.1.3 Epidemiology

Breast cancer holds the position as the most prevalent malignant tumor affecting women worldwide. It accounts for a significant portion, approximately 36%, of all cancer cases. In 2018, an estimated 2.089 million women received a breast cancer diagnosis. The incidence of this type of cancer is on the rise across all regions globally, with the highest rates observed in industrialized nations. Developed countries contribute to almost half of all reported cases. This increase can be attributed primarily to the adoption of a Western lifestyle characterized by unhealthy eating habits, tobacco use, high stress levels, and sedentary behavior.Mammography has emerged as the recognized screening method for breast cancer. It offers substantial benefits, particularly for women aged 50 to 69. Classical mammography demonstrates a sensitivity and specificity ranging from 75% to 95% and an accuracy level of 80% to 95%. In cases where there is suspicion of hereditary breast cancer, magnetic resonance mammography is employed as a screening tool. If a mammogram reveals a suspicious lesion, an ultrasound examination is conducted, followed by a thick needle biopsy if necessary. The tumor is then subjected to a histopathological examination for further evaluation (Smolarz et al. (2022)).

For a specific tumour in a given population, crude rates are calculated simply by dividing the number of new cancers or cancer deaths observed during a given time period by the corresponding number of individuals in the population at risk. For cancer, the result is commonly expressed as an annual rate per 100 000 individuals at risk (Smolarz et al. (2022)). In 2018, the United States recorded 234,087 cases of breast cancer (crude rate: 85/105), followed by 55,439 cases in the United Kingdom (crude rate: 94/105), 56,162 cases in France (crude rate: 99/105), 71,888 cases in Germany (crude rate: 85.4/105), and 66,101 cases in Japan (crude rate: 58/105).

Belgium has the highest incidence rate worldwide (crude rate: 113/105), with Australia leading among continents (crude rate: 94/105). Poland also experiences breast cancer as the most commonly diagnosed malignant tumor in women, showing a consistent increase in cases from 8,000 new cases in 1990 to 20,203 new cases in 2018. The average incidence rate in Europe stands at 84/105. In Southeast Asian and African countries, breast cancer has the lowest incidence, with standardized rates not exceeding 25/105. Bhutan (crude rate: 5/105) and the Republic of The Gambia (crude rate: 6.5/105) recorded the lowest incidence rates in 2018. Despite advances in diagnostics and pharmacotherapy, breast cancer remains the leading cause of death from malignant tumors among women worldwide, claiming the lives of 626,679 individuals in 2018. The highest mortality rates are observed in developing countries, such as Fiji (crude rate: 36/105), Somalia (crude rate: 29/105), Ethiopia (crude rate: 23/105), Egypt (crude rate: 21/105), Indonesia (crude rate: 17/105), and Papua New Guinea (crude rate: 25/105), where 60% of all breast cancer deaths occur. This trend primarily stems from limited screening opportunities, and lack of access to diagnostics, and modern treatment methods (Smolarz et al. (2022)).

In contrast, Belgium reports a standardized death crude rate of 16.3/105, the United States at 13/105, and Japan at 9.3/105. Poland exhibits significantly lower breast cancer incidence compared to EU countries, with a standardized incidence rate of 51.8 for Poland compared to 106.6 for the EU in 2013. The incidence of breast cancer among adult premenopausal women (20-49 years) has nearly doubled over the past three decades. Unfortunately, Polish women show lower sensitivity towards prevention, often neglecting their breast health and underestimating the importance of regular check-ups. In comparison to other European countries, Polish women have lower rates of preventive care, with only 44% reporting free mammogram prevention programs, while the Netherlands reports 80% and England reports 71%. The 5-year survival rate for breast cancer in Poland stands at 78.5%, significantly lower than the 90% achieved in the United States (Smolarz et al. (2022)).

Based on the research conducted by the Global Cancer Observatory, breast cancer accounted for 27.5% of cancer cases among the female population of Greece in 2020, as depicted in Figure 1. Additionally, the corresponding death rate for these cases was 7.0%. The age-standardized incidence rate per 100,000 females was 71.9%, while the mortality rate was 14.5%. An age-standardized rate (ASR) is a comprehensive measure of the rate that would be observed if the population had a standard age structure. Standardization becomes necessary when comparing multiple populations that differ in terms of age because age strongly influences the risk of cancer. The ASR is calculated as a weighted average of the age-specific rates, with the weights determined by the population distribution of a standard population. The World (W) Standard Population is the most commonly used standard population for this purpose.

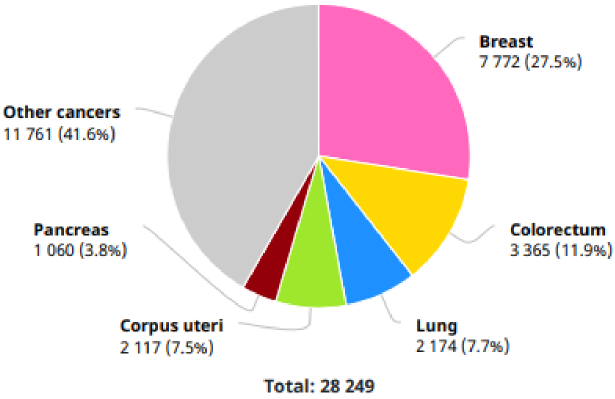


Figure 1: Pie chart for cancer cases in Greece during 2020 (Global Cancer Observatory)

Roughly 13% of women in United States of America, which is equivalent to 1 in 8, are expected to receive a diagnosis of invasive breast cancer during their lifetime. Furthermore, approximately 3% of women, or 1 in 39, will unfortunately succumb to the disease as shown in Table 1. Lifetime risk takes into account the possibility of deaths from other causes that may occur before a breast cancer diagnosis. While the risk of being diagnosed with breast cancer reaches its highest point among women aged 70-79 years (4.1%) and decreases afterward, the risk of mortality due to the disease continues to rise as age advances (Giaquinto et al. (2022)).

Current age, years	**Diagnosed with invasive breast cancer**	**Dying from breast cancer**
20	0.1% (1 in 1439)	<0.1% (1 in 18,029)
30	0.5% (1 in 204)	<0.1% (1 in 2945)
40	1.6% (1 in 63)	0.1% (1 in 674)
50	2.4% (1 in 41)	0.3% (1 in 324)
60	3.5% (1 in 28)	0.5% (1 in 203)
70	4.1% (1 in 24)	0.7% (1 in 137)
80	3.0% (1 in 33)	1.0% (1 in 100)
Lifetime risk	**12.9% (1 in 8)**	**2.5% (1 in 39)**

Table 1: Breast cancer risk by age. Probability is among those who have not been previously diagnosed with cancer and reflects the likelihood of diagnosis/death within 10 years of current age. Percentages and “1 in” numbers may not be numerically equivalent because of rounding (Giaquinto et al. (2022)).

1.2 Literature Review

Countless people worldwide are impacted by the common and deadly disease known as breast cancer. Breast cancer survival rates and patient outcomes can be greatly enhanced by early and precise identification. The use of machine learning and data mining approaches for the classification of breast cancer has attracted increasing interest in recent years. These methods have demonstrated substantial potential for helping doctors make accurate diagnoses and treatment decisions. Breast cancer is categorized by dividing tumor samples into groups such as malignant (cancerous) and benign (non-cancerous). Mammography and histological examination are two examples of traditional diagnostic techniques that have limits in terms of precision and dependability. On the other hand, machine learning algorithms have the capacity to examine complex patterns and correlations inside huge datasets, enabling more precise and effective breast cancer classification.

This literature review seeks to present a summary of the current research on machine learning algorithms for breast cancer classification. It will examine numerous techniques used in research, including support vector machines, random forests, artificial neural networks, and ensemble methods. The evaluation will also point out the field's advances, strengths, and limits as well as possible future directions. This study aims to add to the knowledge of the present state-of-the-art in breast cancer classification using machine learning approaches by synthesizing and assessing the available research. It attempts to find the best techniques, and characteristics for precise and trustworthy categorization. Researchers, physicians, and other healthcare professionals involved in the detection and treatment of breast cancer can benefit greatly from the conclusions of this analysis.

Ara et al. (2021) utilized the Wisconsin Breast Cancer Dataset (WBCD) from the FNA biopsy system to apply various machine learning (ML) classifiers and determine the type of breast cancer in a suspected patient. Six classification models were employed, including Random Forest, Logistic Regression, Decision Tree, Naive Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). To identify the most suitable model for breast cancer prediction, the obtained results were evaluated to compare the algorithms. The best models, based on testing accuracy, were identified as Random Forest and SVM, both achieving an accuracy of 96.5 %. According to Abed et al. (2016), a hybrid classification algorithm combining the Genetic Algorithm (GA) and KNN is suggested. The GA algorithm is utilized for its primary purpose as an optimization technique for KNN, involving feature selection and optimization of the k value. On the other hand, kNN is employed for classification purposes. The effectiveness of the proposed algorithm is evaluated by applying it to the WBCD. The algorithm is compared to different classifier algorithms using the same database. The evaluation results of the proposed algorithm demonstrate a remarkable accuracy of 99%.

The classification accuracy, sensitivity, specificity, and other characteristics of four machine learning algorithms—Logistic Regression, SVM, KNN, and Naive Bayes—are calculated and compared in the work of Kumar et al. (2020). The various hyperparameters utilized by various ML algorithms were chosen manually. SVM outperformed all other methods, with an accuracy of roughly 98.24%. Moreover, an efficient hybridized classifier for diagnosing breast cancer is suggested in the paper of Mittal et al. (2015). Self organizing maps (SOM), an unsupervised artificial neural network (ANN) technique, and stochastic gradient descent (SGD), a supervised classifier, are used to create the classifier. Additionally, a comparison is made between the suggested method and three cutting-edge supervised machine learning techniques: decision trees (DTs), random forests (RF), and SVM. The SGD approach is initially employed independently for the classification task, then after being hybridized with the unsupervised ANN methodology on the WBCD, it is made to conduct the classification. The findings of the classification experiment employing the hybridization of SOM and SGD are much better than SGD alone.

The utilization of deep learning technology for breast cancer diagnosis is demonstrated in the paper of Khuriwal and Mishra (2018a). Despite being commonly employed in high-task objective areas such as Computer Vision, Image Processing, Medical Diagnosis, and Natural Language Processing, the application of deep learning techniques on the WBCD is explored. The results reveal the significant benefits of utilizing deep learning technology, achieving an impressive accuracy of 99.67% for breast cancer diagnosis. The paper is structured into three parts, beginning with data collection and the application of preprocessing algorithms to scale and filter the data. Subsequently, the dataset is split into training and testing subsets, with visualizations generated to aid data comprehension. Finally, the model is implemented on the training dataset, leading to the accuracy of 99.67%.

In the paper of Algarni et al. (2021), a deep learning architecture is proposed to support the detection of breast tumors using structured features. First, the performance of multiple state-of-the-art machine learning approaches, including SVM, DT, Logistic Regression, and Convolutional Neural Networks (CNN), was evaluated and compared. Additionally, a combined ensemble model with three base models was constructed to induce better generalization performance. These approaches were evaluated for automatically classifying tumors using the publicly available breast cancer dataset, the WBCD. Experimental results indicate that the highest classification accuracy (98%) is achieved by the proposed CNN deep model classifier.

RF and Extreme Gradient Boosting (XGBoost), two ensemble machine learning classifiers, are compared in terms of performance on the WBCD in the research of Abdulkareem and Abdulkareem (2021). The major goal of this study is to evaluate the accuracy of the classifiers in terms of their effectiveness and efficiency in classifying the dataset. This was accomplished by using both the dataset's full set of features as well as its reduced set, which was produced using the Recursive Feature Elimination (RFE) feature selection approach. Accuracy, Precision, Recall, and F1-Score were the four metrics utilized in the study to assess the classifiers. The results of the experiment demonstrate that the XGBoost algorithm with 5 reduced features and the RFE feature selection approach provides the best accuracy (99.02%) and lowest error rate.

In the study of Basunia et al. (2020), an ensemble method named "stacking classifier" was proposed, which combines multiple classification techniques and effectively classifies the benign and malignant tumor. The WBCD was used for the experiment. Different classification techniques were applied over the dataset, and their parameters were tuned to improve accuracy. The three best classifiers, namely KNN, SVM, and RF, were chosen for the proposed method. Generally, the proposed stacking classifier combined the results of those best classifiers using a meta classifier, specifically Logistic Regression, and achieved a breast cancer prediction accuracy of 97.20%. Furthermore, in the paper of Khuriwal and Mishra (2018b), an adaptive ensemble voting method was proposed for diagnosing breast cancer using the WBCD. The aim of this work is to compare and explain how a better solution is provided by the Artificial Neural Networks (ANN) and Logistic Regression algorithm when working with ensemble machine learning algorithms for breast cancer diagnosis, even with reduced variables. When compared to related work from the literature, it is demonstrated that an accuracy of 98.50% is achieved by the ANN approach with the Logistic algorithm, surpassing that of other machine learning algorithms.

The paper of Telsang and Hegde (2020) introduces a breast cancer prediction utilizing various machine learning algorithms, comparing their prediction accuracy, area under the receiver operating characteristic curve (AUC), and performance parameters. The WBCD is employed for simulation purposes. Through analysis, the SVM model achieves an accuracy of 96.25% with an AUC of 99.4. Additionally, there is potential to enhance the breast cancer prediction by modifying the mathematical models of these algorithms. Similarly, the research of Sinha (2020) utilized the WBCD, renowned as the benchmark database for result comparison across various algorithms. The classification of benign and malignant tumors was performed using the following machine learning classification techniques: SVM, KNN, RF, Adaboost Classifier, and XGboost Classifier. The accuracy achieved in scaled features for this classifiers is 96%, 57%, 75%, 94% and 98% respectively. In the same research pattern, the paper of Singh and Thakral (2018) utilized the WBCD by implementing a classification analysis using DT classifier (J4.8, Simple CART) and Bayes classifier (Naive Bayes, Bayesian Logistic Regression). The experimental result shows that among all the classifiers, DT classifier i.e. Simple CART (98.13%) gives higher accuracy.

The research of Agarap (2018) presents a comparative analysis of six machine learning (ML) algorithms: GRU-SVM, Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbor (NN) search, Softmax Regression, and SVM. The performance of these algorithms is evaluated using the WBCD, which contains features computed from digitized images on breast masses. The dataset is divided into a 70% training set and a 30% testing set. The classification test accuracy, sensitivity, and specificity values are measured for each algorithm. The results indicate that all the ML algorithms perform well, with test accuracies exceeding 90%. Notably, the MLP algorithm demonstrates outstanding performance, achieving a test accuracy of approximately 99.04%.

The paper of Sidey-Gibbons and Sidey-Gibbons (2019) addresses the growing interest in machine learning techniques for medical research and clinical applications. It provides both a conceptual introduction and a practical guide to developing and evaluating predictive algorithms using freely-available open source software and public domain data. The study focuses on cancer diagnosis and demonstrates the use of machine learning techniques by developing three predictive models using WBCD. Algorithms including General Linear Model Regression (GLMs) and specifically LASSO Logistic Regression, SVM, and ANN are trained on the evaluation sample and used to predict diagnostic outcomes. The results show that the trained algorithms achieve high accuracy (0.94 - 0.96), sensitivity (0.97 - 0.99), and specificity (0.85 - 0.94) in classifying cell nuclei. The SVM algorithm achieves the highest accuracy (0.96) and area under the curve (0.97). The performance slightly improves when the algorithms are combined into a voting ensemble (accuracy = 0.97, sensitivity = 0.99, specificity = 0.95).

The paper of Gosain and Sardana (2017) addresses the Class Imbalance Problem (CIP), which refers to the situation where the distribution of classes in a dataset is significantly skewed, with one class being heavily represented compared to the others. Four oversampling techniques, namely Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling approach (ADASYN), Borderline-SMOTE, and Safe-Level SMOTE, were utilized to handle this problem. The performance of these oversampling techniques in dealing with the CIP is compared using four publicly available datasets, with one of them being the WBCD. Popular classification models, including Naïve Bayes, SVM, and KNN, are employed for the comparison. The evaluation metrics used to assess performance include Overall Accuracy, Sensitivity, Specificity, Precision, F-measure, G-mean, and Area under the curve (AUC) value. Based on the results, Safe-Level SMOTE demonstrates superior performance, particularly in terms of F-measure and G-mean across most datasets. The generation of minority instances around larger safe levels contributes to its higher accuracy performance compared to SMOTE, ADASYN, and Borderline-SMOTE.

The paper of Wang et al. (2021) proposes an improved version of the SMOTE algorithm for the expansion and classification of imbalanced data. The authors compare the performance of the proposed algorithm to the original SMOTE algorithm on several imbalanced datasets, including WDBC. The experimental results show that the proposed algorithm achieves better classification performance than the original SMOTE algorithm on all datasets, including WDBC. Specifically for WDBC dataset, the proposed algorithm achieved an Area under the ROC of 96.49%, which is slightly higher than the AUC achieved by the original SMOTE algorithm (95.61%).

This research of Cahyana et al. (2019) explores the use of oversampling techniques (SMOTE, Borderline-SMOTE, and ADASYN) to address imbalanced data classification challenges. The study evaluates their impact on classification accuracy using the XGBoost algorithm and seven datasets. Results show that oversampling improves accuracy by 2% to 11% in most datasets, with Borderline-SMOTE yielding the highest improvements. Interestingly, the WBCD exhibits steady accuracy regardless of oversampling. However, the effectiveness of oversampling depends on the dataset and algorithm sensitivity, highlighting the need for careful consideration when applying these techniques.

The study of Cai (2018) presents a model that combines an ensemble method and an imbalanced learning technique for the classification of breast cancer data, specifically utilizing the WBCD. The model consists of two main steps. Firstly, the Synthetic Minority Over-Sampling Technique (SMOTE) is applied to the selected dataset. Secondly, multiple baseline classifiers are tuned using Bayesian Optimization. Finally, a stacking ensemble method is employed to combine the optimized classifiers for the final decision. Comparative analysis demonstrates that the proposed model outperforms conventional methods in terms of classification accuracy, specificity, and AUC. The best baseline accuracy arose for XGBoost classifier with SMOTE approximately in 97%. The ensemble method with SMOTE achieved an accuracy of 97.5%.

The research of S. A. Mohammed et al. (2020) focuses on improving the accuracy and performance of three classifiers (Decision Tree (J48), Naïve Bayes, and Sequential Minimal Optimization (SMO)) for predicting early-stage breast cancer. The classifiers are validated and compared using two benchmark datasets: WBCD and Breast Cancer. The paper addresses the challenge of imbalanced classes in the data and proposes a data-level approach of resampling to mitigate the impact of class imbalance. The evaluation is done using 10-fold cross-validation, and the classifiers' efficiency is assessed based on various metrics such as true positive rate, false positive rate, ROC curve, and accuracy. The experimental results demonstrate that using a resample filter enhances the classifiers' performance, with SMO performing best on the WBCD and J48 outperforming others on the Breast Cancer dataset.

The study of Solanki et al. (2021) focuses on improving the accuracy of machine learning models for breast cancer prognosis. It explores wrapper-based feature selection methods and uses SVM, J48 DT, and Multilayer Perceptron (MLP) classifiers. The research emphasizes handling imbalanced datasets and evaluates different sampling techniques, such as SMOTE. The results show that the J48 DT classifier, combined with genetic search for feature selection, achieves high accuracy (98.83%) and other performance metrics. The study highlights the importance of sampling techniques in addressing class imbalance for accurate breast cancer prognosis.

Comparing different models using the same dataset is essential because model performance can vary widely when applied to different datasets. By using the same dataset, researchers ensure a fair and unbiased evaluation. This approach allows for a direct comparison of the advantages and disadvantages of each model, considering how well they perform with the specific data features and patterns in the dataset. Using a consistent dataset helps establish a reliable benchmark for comparison, enabling researchers to select the best model for their specific problem. It also highlights the importance of robustness and generalizability, as models that work well on one dataset may not be as effective on new, unseen data. Ultimately, using the same dataset for evaluation improves the reliability of the study and leads to more accurate findings.

2 Theoretical Background

2.1 Logistic Regression

Multivariable problems are frequently encountered in medical research. A typical question of researchers is to what extent a variable or a set of them affects a disease outcome. The disease outcome considered dichotomous with 0 representing **not diseased** and 1 representing **diseased**. To evaluate the extent to which the variables are associated with the outcome, the multivariable problem considers the variables independent and the outcome is set as a dependent binary variable (Kleinbaum and Klein (2006)). Logistic Regression is a modeling approach that can be used to describe the relationship of several variables to a dichotomous dependent variable.

The equation 1 presents the function on which the Logistic Regression is based. The plot of this function is presented in figure 2. When z approaches $-\infty$, the logistic function $f(z)$ equals 0. On the other side, when z approaches ∞, the logistic function $f(z)$ equals 1. Thus, as the graph describes, the range of $f(z)$ is between 0 and 1, regardless of the value of z.

$$f(z) = \frac{1}{1+e^{-z}} \tag{1}$$

The fact that the logistic function $f(z)$ varies between 0 and 1 is the fundamental reason the logistic model is so popular. The model is designed to describe a probability, which is always a value between 0 and 1. In medical terms, such a probability conveys the risk of an individual contracting a disease. So, the logistic model is constructed to guarantee that whatever estimate of risk is received, the risk will always be a value between 0 and 1.

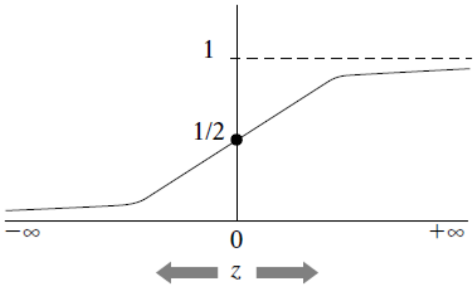


Figure 2: Logistic Function (Kleinbaum and Klein (2006))

The logistic function's shape is another factor contributing to the logistic model's applicability. As shown in the figure 2, if it started at $z = -\infty$ and go to the right, then as z increases, the value of $f(z)$ hovers close to zero for a while, then starts to increase dramatically toward 1, and then levels off around 1 as z increases approaching ∞. This S-shape image of $f(z)$ implies that the effect of z on an individual's risk is modest for low z's until some threshold is achieved. The danger then increases quickly over a specific range of intermediate z values, and once z is large enough, it stays extraordinarily high around 1.

2.1.1 Logistic Model

The logistic function can be utilized to model the probability of the disease outcome D. More specifically, considering the observation of the independent variables $X_1, X_2, ..., X_p$ for a group of subjects for whom the outcome $D = 1$ or 0 is determined, the probability of a new observation to get the disease can be modeled. To obtain the logistic model from the logistic function, z is expressed as a linear combination of the independent variables and $f(z)$ is linked with the probability of the disease outcome $D = 1$ given the values of $X_1, X_2, ..., X_p$. The logistic function (Kleinbaum and Klein (2006)) then models that probability through the equation 1:

$$P(D = 1 \mid X_1, X_2, ..., X_p) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{p} \beta_i X_i)}} \tag{2}$$

The terms a and b_i are unknown parameters that must be calculated using the information of the independent variables and their linked outcome on the group of subjects. Using the equation 3, the probability of the negative outcome $D = 0$ can be calculated (Kleinbaum and Klein (2006)) by the rule of subtraction:

$$\begin{aligned} P(D = 0 \mid X_1, X_2, ..., X_p) &= 1 - P(D = 1 \mid X_1, X_2, ..., X_p) \\ P(D = 0 \mid X_1, X_2, ..., X_p) &= \frac{e^{-(\beta_0 + \sum_{i=1}^{p} \beta_i X_i)}}{1 + e^{-(\beta_0 + \sum_{i=1}^{p} \beta_i X_i)}} \end{aligned} \tag{3}$$

2.1.2 Fitting of the Logistic model

Fitting the logistic model is equivalent to determining the unknown parameters - coefficients of the equation 3 by the maximization of the log-likelihood of the N observations that correspond to the known group of subjects. The log-likelihood of the model is given by:

$$\ell(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; \beta) \tag{4}$$

where $p_k(x_i; \beta) = P(D = k \mid x_i; \beta)$ for $k = 0, 1$. β represents the vector of all the unknown parameters and $x_i = [1, X_1, X_2, ..., X_p]^T$. The two classes are coded through g_i via 0/1 response y_i, where $y_i = 1$ when $g_i = 1$ and $y_i = 0$ when $g_i = 0$. Because the problem consists of two classes $p_1(x_i; \beta) = p(x_i; \beta)$ and $p_0(x_i; \beta) = 1 - p(x_i; \beta)$. Therefore, the log-likelihood can be written (Hastie et al. (2009)):

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^{N} \{y_i \log p(x_i; \beta) + (1 - y_i) \log (1 - p(x_i; \beta))\} \\ &= \sum_{i=1}^{N} \left\{y_i \beta^T x_i - \log \left(1 + e^{\beta^T x_i}\right)\right\}. \end{aligned} \tag{5}$$

To maximize the log-likelihood, the derivatives are set to zero:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i; \beta)) = 0 \tag{6}$$

which are p+1 equations nonlinear in β. To solve these equations, Newton-Raphson algorithm is utilized (Hastie et al. (2009)):

$$\beta^{\text{new}} = \beta^{\text{old}} - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T}\right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta} \tag{7}$$

Let $\mathbf{y}$ denote the vector of y_i values, $\mathbf{X}$ the $N \times (p + 1)$ matrix of x_i values, $\mathbf{p}$ the vector of fitted probabilities with ith element $p(x_i; \beta^{old})$ and $\mathbf{W}$ a $N \times N$ diagonal matrix of weights with ith diagonal element $p(x_i; \beta^{old})(1 - p(x_i; \beta^{old}))$. Then (Hastie et al. (2009)):

$$\begin{aligned}\frac{\partial \ell(\beta)}{\partial \beta} &= \mathbf{X}^T(\mathbf{y}-\mathbf{p}) \\ \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} &= -\mathbf{X}^T\mathbf{W}\mathbf{X}\end{aligned} \tag{8}$$

Thus, the Newton step is:

$$\begin{aligned}\beta^{\text{new}} &= \beta^{\text{old}} + \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{p}) \\ &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\left(\mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})\right) \\ &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}\end{aligned} \tag{9}$$

In the second and third line the Newton step has been re-expressed as a weighted least squares step, with the response:

$$\mathbf{z} = \mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p}) \tag{10}$$

These equations are solved repeatedly, since at each iteration $\mathbf{p}$ changes, and hence so does $\mathbf{W}$ and $\mathbf{z}$. This algorithm is referred to as Iteratively Reweighted Least Squares or IRLS (Hastie et al. (2009)), since each iteration solves the weighted least squares problem:

$$\beta^{\text{new}} \leftarrow \arg\min_{\beta}(\mathbf{z}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{z}-\mathbf{X}\beta) \tag{11}$$

2.1.3 Wald test

By applying the Newton-Raphson method using the equation 11, the logistic regression coefficients $\hat{\beta}$ are estimated. However, the corresponding variables might not be statistically significant for the model. To determine this information, Wald test must be applied under the H_0 hypothesis of $\beta_j = 0$ and the alternate hypothesis H_1 of $\beta_j \neq 0$, as it is considered in Likelihood Ratio test in equation 14. The estimator of the Maximum Likelihood of the model follows asymptotically the Normal distribution (Karoni and Oikonomou (2017)) and hence it is approximately true that:

$$\frac{\hat{\beta}_j - \beta_j}{(\mathbf{I}^{-1}(\hat{\beta})_{jj})^{\frac{1}{2}}} \dot{\sim} \mathbf{N}(0,1),\ j = 0,1,...,p \tag{12}$$

where p is the number of coefficients and $\mathbf{I}(\hat{\beta})$ is the observed information matrix. The information matrix is defined as the matrix product $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ where $\hat{\mathbf{W}}$ is the diagonal matrix $\hat{\mathbf{W}} = diag(e^{x_i^T\hat{\beta}})$. Based on theory of Maximum Likelihood it is shown that the variance of the coefficient $\hat{\beta}_j$, $\hat{\mathbf{V}}(\hat{\beta}_j)$ is the j-th element of the matrix $\mathbf{I^{-1}}(\hat{\beta})$ with corresponding standard error $se(\hat{\beta}_j) = (\hat{\mathbf{V}}(\hat{\beta}_j))^{\frac{1}{2}} = (\mathbf{I}^{-1}(\hat{\beta})_{jj})^{\frac{1}{2}}$ (Karoni and Oikonomou (2017)).

2.1.4 Deviance and Goodness of Fit

When it comes to the selection of a model for a specific dataset, it is very important to evaluate the eligibility of the model under specific requirements. A very significant factor is the quality of the data description that the model offers. This information could be measured using the scaled *Deviance* function for the model M_0:

$$\mathbf{D_0} = -2(\hat{\ell}_0 - \hat{\ell}_S) \dot{\sim} \chi^2_d \tag{13}$$

where $\hat{\ell}_0$ is the maximized log-likelihood of the model M_0. $\hat{\ell}_S$ is the corresponding maximized log-likelihood of the saturated model, which is described by the same number of coefficients as the number of samples. d is equal to the difference between the number of coefficients of M_0 model and the saturated model (Karoni and Oikonomou (2017)). The saturated model has perfect fit on the data and hence, the deviance function measures how much the model M_0 deviates from the saturated one. If the deviance is large, then the model adaptation to the data is poor and another model with different variables must be evaluated.

Nevertheless, the purpose in data analysis is not only to find one model that fits well enough to the data, but to determine the best model out of many possible ones, which are comprised of different combinations of the independent variables. Deviance function is a suitable indicator of the goodness of fit to the data. Therefore, the difference of deviances of two models could measure the comparison of the adaptation of these models to the data. More specifically, if the model M_0 contains p_0 variables that are a subset of the p_1 variables of a model M_1, then these models are considered nested. M_0 model arises from M_1, if $d = p_1 - p_0$ constraints like $\beta_j = 0$ are set for d independent variables of M_1. Equation 13 is valid for nested models asymptotically and hence, if D_1 and D_0 are the deviance values of the models M_1 and M_0 respectively (Karoni and Oikonomou (2017)), then the Likelihood Ratio test is:

$$\mathbf{D_0} - \mathbf{D_1} = -2(\hat{\ell}_0 - \hat{\ell}_S) + 2(\hat{\ell}_1 - \hat{\ell}_S) = -2(\hat{\ell}_0 - \hat{\ell}_1) \dot{\sim} \chi^2_d \tag{14}$$

2.1.5 L_1 Regularization

L1 regularization is a technique used in Logistic Regression to prevent overfitting of the model. In L1 regularization, a penalty term is added to the log-likelihood function 5 of the Logistic Regression model, which encourages the model to have small weights for some of the features. Mathematically, the L1 regularization penalty is defined as the sum of the absolute values of the weights except the intercept (Hastie et al. (2009)):

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i \left(\beta_0 + \beta^T x_i \right) - \log \left(1 + e^{\beta_0 + \beta^T x_i} \right) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{15}$$

The regularization term contains the L_1 norm of the vector of the coefficients and when the expression of 15 is maximized then some of the coefficients might be pushed towards zero or even become exactly zero. Equation 11 results in a solution $\hat{\beta}$, as it is presented in figure 3. L_1 regularization finds the first point where the elliptical contours hit the blue constraint region. For L_1 regularization this constraint region is given by the equation (Hastie et al. (2009)):

$$\sum_{i=1}^{p} |\beta_i| \leq t \tag{16}$$

If the solution occurs at a corner, then it has one parameter β_i equal to zero. When $p > 2$, the blue region becomes a rhomboid with numerous corners, flat edges and faces. Hence, the odds of the estimated coefficients being zero are much higher.

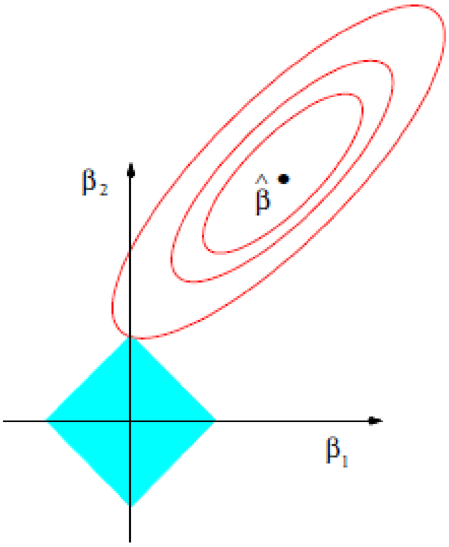


Figure 3: Estimation picture for L1 regularization.The solid blue areas are the constraint regions of the coefficients, while the red ellipses are the contours of the IRLS error function (Hastie et al. (2009))

2.1.6 Coordinate Descent

The Newton algorithm for maximizing log-likelihood of equation 5 amounts to IRLS. Hence if the current estimates of the parameters are $(\tilde{\beta}_0, \tilde{\beta})$, a quadratic approximation is formed to the log-likelihood (Friedman et al. (2010)) which is:

$$\ell_Q\left(\beta_0, \beta\right) = -\frac{1}{2N} \sum_{i=1}^{N} w_i \left(z_i - \beta_0 - x_i^{\top} \beta\right)^2 + C\left(\tilde{\beta}_0, \tilde{\beta}\right)^2 \tag{17}$$

where:

$$\begin{aligned} z_i &= \tilde{\beta}_0 + x_i^{\top} \tilde{\beta} + \frac{y_i - \tilde{p}\left(x_i\right)}{\tilde{p}\left(x_i\right)\left(1 - \tilde{p}\left(x_i\right)\right)} \\ w_i &= \tilde{p}\left(x_i\right)\left(1 - \tilde{p}\left(x_i\right)\right) \end{aligned} \tag{18}$$

where $\tilde{p}\left(x_i\right) = \tilde{p}\left(x_i; \beta\right)$ is evaluated at the current parameters. For each value of λ, an outer loop is created that computes the quadratic approximation 17 about the current parameters $(\tilde{\beta}_0, \tilde{\beta})$. Then, the coordinate descent is utilized to solve the penalized weighted least-squares problem (Friedman et al. (2010)):

$$\min_{\left(\beta_0, \beta\right) \in R^{p+1}} \left\{ -\ell_Q\left(\beta_0, \beta\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right\} \tag{19}$$

This amounts to a sequence of nested loops:

1. **outer loop:** Decrement of λ
2. **middle loop:** Update the quadratic approximation ℓ_Q using the current parameters $(\tilde{\beta}_0, \tilde{\beta})$
3. **inner loop:** Run the coordinate descent algorithm on the penalized weighted-least-squares problem

2.2 Decision Trees

In supervised learning, the value of one or more response variables that describe an outcome can be predicted using a collection of independent variables (predictors). A prediction model is utilized to match the independent predictors with a specific outcome. The main objective is to create a model that is capable of predicting the value of the response variable by learning rules that arise through the independent predictors. The data used to assess the prediction model consists of a set of observations that include the independent features and the response variable. When the response variable is unknown, the fitted model is employed to predict its value. (Saxena (2022)).

A structure called a *Decision Tree* can group data by applying a set of straightforward rules on the independent variables. Each observation is classified into a group based on the applied rules on its independent variables. The outcome of the successive rules on the tree's structure constitutes a hierarchy of groups inside other groups. The groups in each level of the hierarchy are called *nodes*, and the first group that belongs to the root node is formed by the whole set of data. A node and its successors create a *branch* from this node and the last nodes of the hierarchy are the *leaves* of the tree structure. A selection that is applied in each of the leaves' instances is made by employing the last rules in each of the leaves. In supervised learning, this selection concerns the predicted value of the response variable (Saxena (2022)).

This non-parametric supervised learning technique is utilized for both classification and regression applications (Saxena (2022)). A regression tree represents a continuous response, whereas a classification tree models a categorical response as can be visualized in figure 4. Because the model is written as a series of if-then statements, both forms of trees are referred to as decision trees.

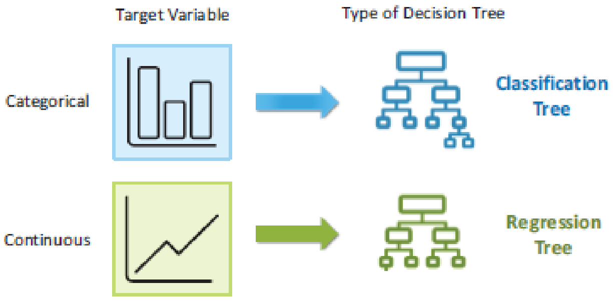


Figure 4: Classification and Regression Trees

Tree models can utilize both categorical and numerical features, and the space that represents them includes all their possible combinations. This space is split into non-overlapping parts that are depicted by the leaves of the tree. The root node, which contains the whole set of data, is repeatedly separated until a stopping criterion is met. By selecting an independent variable and a splitting value for that variable that minimizes the variability according to a defined measure in the outcome variable for all child nodes, the parent node is then split into child nodes at each stage. Different measures like the Gini index, entropy, and residual sum of squares can be used to assess the candidate splits for each node. The selected independent variable and its splitting value are called the primary splitting rule. (Saxena (2022)).

The instances are classified with guidance from the root node to the leaves, according to the results of the rule application along the tree path. Specifically, the root node and corresponding feature are evaluated to determine which branch of the tree the observed value corresponds to. Then, the next node of the specific branch is evaluated, and the procedure is repeated until a leaf is reached. It should be noted that both categorical and numerical features are incorporated into the tree structure (Rokach and Maimon (2008)). In case of numeric attributes, decision trees can be geometrically interpreted as a collection of hyperplanes, each orthogonal to one of the axes, as it is visualized in figure 5.

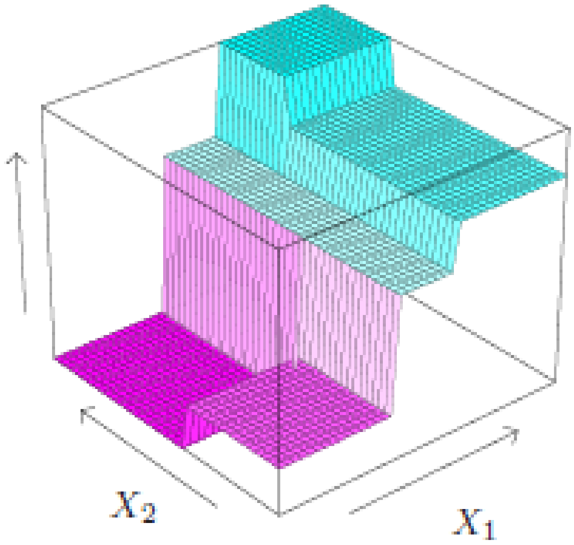


Figure 5: Perspective plot of the prediction surface of 3-dimensional input variable decision tree (Hastie et al. (2009))

2.2.1 Splitting Criteria

The splitting criteria used at the root node of a decision tree is of significant importance as it determines how the initial split is made, which subsequently affects the entire structure and predictive accuracy of the decision tree. The splitting criteria in the root node is typically chosen based on the features of the input data, and it plays a critical role in dividing the data into distinct subsets or branches.

The importance of the splitting criteria at the root node of a decision tree can be summarized as follows (Rokach and Maimon (2008)):

1. Decision-making: The primary separation criterion in the root node determines the feature or property used to make the initial decision for separating the data into different branches. This initial separation sets the foundation for subsequent divisions in the tree, leads to the creation of decision rules, and determines the final predicted results for unseen data in the model.

2. Predictive accuracy: The selection of the split criterion in the root node determines the accuracy of the tree model. A well-selected split criterion could lead to more homogeneous data subsets in the branches and, subsequently, better predicted accuracy. On the other hand, a poorly chosen separation criterion could lead to imbalanced or insufficiently separated subsets and, therefore, reduced accuracy and lower quality decision rules.

3. Interpretability: The root node separation criterion could affect the interpretability of the tree model. A decision tree with a clear and meaningful separation criterion in the root node could be easily understood and interpreted by human users, making it useful for explaining the decision-making process and gaining insights from the model's predictions.

4. Computational efficiency: The selection of the separation criterion in the root node could affect the computational efficiency of the tree model. Some separation criteria may require more computational resources for calculation or evaluation, while others may be less costly in terms of computing resources. Determining a suitable separation criterion could contribute to the optimization of the decision tree construction efficiency.

In the ID3 algorithm, the information gain criterion is employed for split selection. Given a training set $\mathbf{D}$, the entropy of $\mathbf{D}$ is defined as (Zhou (2012)):

$$\operatorname{Ent}(D) = -\sum_{y \in \mathcal{Y}} P(y \mid D) \log P(y \mid D) \tag{20}$$

where $P(y \mid D)$ is the probability of randomly selecting an example in class y from possible classes that are included in $\mathcal{Y}$. Entropy is the degree of uncertainty, impurity or disorder of a random variable. It characterizes the impurity of an arbitrary class of examples in a node. If the training set D is divided into subsets $D_1, D_2, ..., D_k$, the entropy may be reduced, and the amount of the reduction is the information gain (Zhou (2012)):

$$G\left(D; D_1, \ldots, D_k\right) = \operatorname{Ent}(D) - \sum_{i=1}^{k} \frac{\left|D_k\right|}{|D|} \operatorname{Ent}\left(D_k\right) \tag{21}$$

The attribute that is used for each split is the one that maximizes the information gain. CART is another famous decision tree algorithm, which uses Gini index (Zhou (2012)) for selecting the split maximizing the gain equation:

$$\begin{aligned} G_{gini}\left(D; D_1, \ldots, D_k\right) = I(D) - \sum_{i=1}^{k} \frac{|D_k|}{|D|} I\left(D_k\right) \\ where \;\; I(D) = 1 - \sum_{y \in \mathcal{Y}} P(y \mid D)^2 \end{aligned} \tag{22}$$

The Gini Index or Impurity $I(D)$ measures the probability for a random instance being misclassified when chosen randomly. The lower the Gini Index, the lower the likelihood of misclassification.

2.2.2 Stopping Criteria

Decision trees created from a training data sample are not ideal classifiers for the entire population of data objects for a variety of reasons. It frequently occurs that the descriptions of data objects are noisy, and this noise may originate from erroneous measurements. Also, there are a lot of local minima that ”conceal” the global minimum, and learning machines’ decision functions are ineffective at explaining hidden dependencies. When overfitting happens close to the root node, the tree model is utterly wrong and frequently cannot be fixed. But when overfitting is caused by splits close to the leaves, *pruning* certain tree branches can be a successful remedy that simplifies the models. Pruning is a technique that removes the parts of the Decision Tree which prevent it from growing to its full depth (Krzysztof (2014)).

Most of the commonly used pruning techniques belong to one of two groups (Krzysztof (2014)):

1. pre-pruning: the methods acting within the process of decision tree construction, which can block splitting particular nodes

2. post-pruning: the methods that act after complete trees are built and prune them afterwards by removing the nodes estimated as not generalizing well

2.3 Random Forests

2.3.1 Bagging

The abbreviation 'Bagging' has arisen from the combination of 'Bootstrap' and 'Aggregating'. As this name suggests, the process of Bagging consists of two main elements: bootstrapping and aggregation. Its aim is to achieve highly effective reduction of errors through the combination of base learners with simple assumptions. One method to achieve independence among the base learners is to train each of them on a non-overlapping data subset derived from a large training set. However, due to the lack of training data, this procedure could produce insignificant and unrepresentative samples that might negatively affect the performance of the base learner (Zhou (2012)).

Bootstrapping is used by Bagging to construct many base learners. More specifically, a bootstrap sample from a learning sample L_n of size n is obtained by randomly drawing n observations from L_n with replacement. Each observation (X_i, Y_i) in L_n has a probability of $1/n$ of being selected in each draw (Genuer and Poggi (2020)). While some instances may be absent from the sample, other samples may appear more than once. T samples of n training instances are obtained by repeating the algorithm T times. Then, the algorithm can be used to train one base learner for each sample.

Voting for classification and mean value for regression are the most common processes used by Bagging to combine the results of the base learners. During the prediction of a testing observation, Bagging feeds the instance to its base learners, collects the outputs, votes for the targets, and uses the label with the majority of votes as the prediction, with ties arbitrarily broken (Zhou (2012)). Bagging can handle both multiclass and binary classification.

2.3.2 Forests

Examples of modern ensemble approaches are the Random Forests. It is a development of bagging, with the addition of randomized feature selection being the main distinction from simple bagging. The base learners of the random forests are simple decision trees, that their rules are based on the random selected subset of the features. In each split selection step, during the development of a component decision tree, random forests randomly picks a subset of features before performing the standard separation selection technique inside the chosen feature subset (Zhou (2012)).

2.4 Extreme Gradient Boosting

2.4.1 Boosting

The Boosting technique is based on the existence of a weak or simple learning algorithm, which, when given labeled training data, produces a weak or simple classifier. By presenting the weak learning algorithm as a 'black box' that can be used repeatedly as a subroutine, but whose internal operations cannot be altered, the Boosting technique aims to enhance its effectiveness (Schapire and Freund (2014)).

To the extent that the error rates are slightly better than a classifier where each prediction is a random guess, the weak learners could be sketchy and somewhat inaccurate, but they are not completely simple and uninformative. The assumption of weak learning, which is fundamental for the boosting technique, states that the principal model produces a weak hypothesis that is slightly better than a random guess on the data that has been trained (Schapire and Freund (2014)).

The main principle that regualtes the Boosting technique refers to the selection of training samples in a way that forces the base learner to extract a new conclusion every time it is called. This can be achieved by selecting training sets in which it is expected that the performance of the base learner will be very poor, even worse than its typical weak performance. If successful, it can be predicted that the Boosting model will produce a new classifier that significantly differs from its predecessors. This is because although the basic classifier is considered to be a weak and mediocre learning algorithm, it is expected to produce classifiers that can make complex predictions (Schapire and Freund (2014)).

2.4.2 Key Mathematical features of XGBoost

The base learners of the XGBoost algorithm are the so called Classification and Regression Trees (CART). Each tree contains a continuous score on each of its leaves. This score is represented by $\mathbf{w_i}$ and it corresponds to the i-th leaf. For a given observation, the rules of the decision trees will be used to classify it into the leaves. The final prediction will be given by the sum of the scores of the corresponding leaves in each of the trees (given by $\mathbf{w}$). In order to determine the specific tree rules that create these leaves, the objective function of the loss with regularization should be minimized (Chen and Guestrin (2016)):

$$\begin{gathered}\mathcal{L}(\phi)=\sum_{i} l\left(\hat{y}_{i}, y_{i}\right)+\sum_{k} \Omega\left(f_{k}\right) \\ \text{where } \Omega(f)=\gamma T+\frac{1}{2} \lambda\|w\|^{2}\end{gathered} \tag{23}$$

where T is the number of leaves in each of the k trees that are take part in the classifier and λ is the regularization coefficient. The f_k functions represent the subset of the feature rules in each of the k trees that correspond to the score in the specific leaf. The loss function l is differentiable, convex and calculates the difference between the prediction $\hat{y}_i$ and the target y_i. The second term penalizes the model's complexity. In order to prevent over-fitting, the extra regularization term Ω helps to smooth the learned weights. It makes sense that the regularized objective would favor models with straightforward and predictive functions (Chen and Guestrin (2016)).

The f_k functions are parameters in the tree ensemble model in 23, which makes it impossible to optimize it using conventional Euclidean-space techniques (Chen and Guestrin (2016)). The model is instead trained in an additive way. Formally, if $\hat{y}_t^{(t)}$ is the prediction of i-th instance on the t-th iteration tree, f_t needs to be added in order to minimize the objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(\mathrm{x}_i\right)\right) + \Omega\left(f_t\right) \tag{24}$$

The f_t that is chosen greedily is the one that most improves the model based on the minimization of the objective function 23. Second-order Taylor approximation is applied to speed up the optimization of the objective (Chen and Guestrin (2016)):

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t\left(\mathrm{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\mathrm{x}_i\right)\right] + \Omega\left(f_t\right) \tag{25}$$

where g_i and h_i are the first and second order gradients of the loss function with respect to $\hat{y}^{(t-1)}$. By removing the constant terms and by expanding Ω from equation 23, the equation 25 can be rewritten:

$$\begin{aligned}\tilde{\mathcal{L}}^{(t)} &= \sum_{i=1}^{n}\left[g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\mathbf{x}_i\right)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T\end{aligned} \tag{26}$$

This is a second order formula of w_j which has its minimum value at (Chen and Guestrin (2016)):

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{27}$$

where $I_j = \{i | q(x_i) = j\}$ the instance set of leaf j for the specific tree structure q. The corresponding optimal value of the objective function is (Chen and Guestrin (2016)):

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{28}$$

2.5 Neural Networks

Neural networks have become a powerful and adaptable family of machine learning algorithms that are capable of tackling challenging issues in a variety of fields, including speech recognition, image recognition, natural language processing, and game playing. The Multi Layer Perceptron (MLP), a feedforward neural network with several hidden layers, is one common form of neural network design (Mehrotra et al. (1999)).

Neural networks, such as Multi Layer Perceptrons (MLPs), work by simulating the structure and operation of the human brain in order to analyze information and generate predictions. Neural networks are made up of connected nodes or *neurons* that are arranged in layers. Weights are used to represent the connections between neurons and are learnt from training data throughout the training phase. Each neuron in a neural network receives inputs, applies an activation function, and produces an output that is sent to the next layer (Graupe (2013)).

An input layer, one or more hidden layers, and an output layer are the three components of MLPs which is a feedforward neural network. In the hidden and output layers, each neuron calculates the weighted total of its inputs, applies an activation function, and generates an output that is sent to the next layer. By introducing non-linearity, the activation function enables MLPs to describe non-linear interactions between input and output data (Graupe (2013)).

MLP training typically consists of two main steps: the forward pass, which propagates input data through the network to compute predicted output, and the backward pass (also known as backpropagation), which computes the gradients of the loss function with respect to the weights and biases and uses them to update the weights and biases in order to reduce prediction error. Up until the model converges to a desirable extent of accuracy, this procedure is done repeatedly (Graupe (2013)).

MLPs provide a number of benefits, including the capacity to learn complex patterns from massive volumes of data, the ability to approximate any continuous function, and the adaptability in processing a broad range of data formats. They have been extensively employed in many different applications, including financial forecasting, natural language processing, picture and audio recognition, and recommendation systems (Graupe (2013)).

MLPs do, however, have certain drawbacks, such as their susceptibility to overfitting, the requirement for a substantial quantity of training data, and the lack of interpretability of their predictions. Nevertheless, MLPs may be extremely successful in achieving state-of-the-art performance with the right tuning, regularization strategies, and careful consideration of model design (Graupe (2013)).

It is very important to distinguish the different steps that are included in the training of the MLPs. Backpropagation algorithm constitutes an effective training algorithm of the weights of the neural network and it will be presented in the next steps. Some useful notation is:

1. X represents the input data.

2. Y represents the target output.

3. $W^{(l)}$ represents the weight matrix for layer l.

4. $b^{(l)}$ represents the bias vector for layer l.

5. $\sigma(z)$ represents the activation function, where z is the input to the function.

Here are the steps involved in training an MLP with logarithmic loss and backpropagation (Goodfellow et al. (2018)):

1. Initialize the weights and biases for all the layers. For a network with L layers, the following parameters should be initialized:
$$W^{(1)}, \ldots, W^{(L)}$$
$$b^{(1)}, \ldots, b^{(L)}$$

2. Compute the output of the network for a given input X using forward propagation. For each layer l:

    (a) Compute the pre-activation values: $z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$

    (b) Compute the activation values: $a^{(l)} = \sigma(z^{(l)})$ where $a^{(0)} = X$ and σ the activation function.

3. Compute the loss function for the predicted output and the actual output using the logarithmic loss function:
$$J(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{29}$$
where n is the number of training examples, y_i is the true label for the i-th training example, and $\hat{y}_i$ is the predicted label for the i-th training example.

4. Compute the error in the output layer:
$$\delta^{(L)} = J(\hat{y}, y)$$
where $\hat{y}$ is the predicted output and y is the true output.

5. Compute the error for the previous layers using backpropagation. For each layer l, compute the error:
$$\delta^{(l)} = (W^{(l+1)})^T \delta^{(l+1)} \odot \sigma'(z^{(l)}) \tag{30}$$
where $\odot$ denotes element-wise multiplication and $\sigma'(z^{(l)})$ is the derivative of the activation function $\sigma(z^{(l)})$.

6. Compute the gradients for the weights and biases for each layer. For each layer l:

    (a) Compute the weight gradient: $\nabla_{W^{(l)}} J(y, \hat{y}) = \delta^{(l)} (a^{(l-1)})^T$

    (b) Compute the bias gradient: $\nabla_{b^{(l)}} J(y, \hat{y}) = \delta^{(l)}$

7. Update the weights and biases for each layer using gradient descent:

    (a) Update the weights: $W^{(l)} = W^{(l)} - \alpha \nabla_{W^{(l)}} J(y, \hat{y})$

    (b) Update the biases: $b^{(l)} = b^{(l)} - \alpha \nabla_{b^{(l)}} J(y, \hat{y})$

    where α is the learning rate.

8. Repeat steps

2.6 Metrics and Statistics

Metrics are very informative tools for evaluating the performance of a machine learning model. By using different metrics, the model becomes easier to interpret and evaluate, as insights such as patterns or anomalies in different classes become more apparent. Measuring key performance metrics such as accuracy, precision, recall and specificity, and calculating the corresponding confusion matrix, the model's parameters can be tuned, and the best models can eventually be selected.

The Confusion Matrix indicates the number of correct and incorrect predictions for each class. In particular, each row corresponds to the class predicted by the model and each column to the actual class. Therefore, each element of the matrix indicates the number of predictions made by the model for the class of that particular row while the observation belongs to the class of that particular column. In this way, the following items (Larose and Larose (2015)) can be calculated for each class:

1. True Positive (TP): These are the cases where the model correctly predicts the positive class. In other words, the model predicts a positive outcome, and the actual outcome is also positive.
2. True Negative (TN): These are the cases where the model correctly predicts the negative class. In other words, the model predicts a negative outcome, and the actual outcome is also negative.
3. False Positive (FP): These are the cases where the model predicts a positive outcome, but the actual outcome is negative. In other words, the model incorrectly predicts a positive outcome.
4. False Negative (FN): These are the cases where the model predicts a negative outcome, but the actual outcome is positive. In other words, the model incorrectly predicts a negative outcome.

Based on this elements, the following metrics are computed for each class, which the higher they are, the more efficient the model is evaluated (Larose and Larose (2015)):

1. Accuracy is the proportion of all correctly classified instances (both positive and negative) out of the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{31}$$

2. Precision is the proportion of true positives out of all predicted positive instances. It measures how precise the model is when it predicts a positive outcome.

$$Precision = \frac{TP}{TP + FP} \tag{32}$$

3. Sensitivity (also known as recall or true positive rate) is the proportion of true positives out of all actual positive instances. It measures how well the model can identify positive instances.

$$Sensitivity = \frac{TP}{TP + FN} \tag{33}$$

4. Specificity (also known as true negative rate) is the proportion of true negatives out of all actual negative instances. It measures how well the model can identify negative instances.

$$Specificity = \frac{TN}{TN + FP} \tag{34}$$

The Youden index, also known as the Youden's J statistic, is a performance metric that combines the sensitivity and specificity of a machine learning model into a single value (Schisterman and Perkins (2007)). The Youden index ranges from 0 to 1, with higher values indicating better performance. The importance of the Youden index is that it provides a useful summary of the overall performance of a model in identifying both positive and negative instances. The Youden index is calculated as follows:

$$J = Sensitivity + Specificity - 1 = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \tag{35}$$

The Youden index can be used to determine the optimal cut-off point for a binary classification model. The cut-off point is the threshold probability value above which an instance is classified as positive, and below which it is classified as negative (Schisterman and Perkins (2007)). The optimal cut-off point is the one that maximizes the Youden index, which corresponds to the point on the ROC curve where sensitivity and specificity are balanced.

2.7 Handling imbalance

2.7.1 Random Oversampling and Undersampling

Random Oversampling is a data augmentation technique that is commonly used to handle class imbalance problems in machine learning. In this technique, samples from the minority class are randomly duplicated to create an equal number of samples for both the minority and majority classes, as it is visualized in figure 6. This technique is easy to implement and does not require significant computational power. However, random oversampling has its limitations, the most severe being overfitting on the training data, as the model sees the same examples multiple times during training. Moreover, it may not provide a significant improvement in the model's performance, particularly when dealing with highly imbalanced data. Nonetheless, it is a promising starting point to address class imbalance, and it can be combined with other techniques to enhance the model's performance (Ganganwar (2012)).

Random undersampling is a popular data reduction technique commonly used to handle imbalance problems that often arise during the training of machine learning algorithms. In this method, samples from the majority class are removed from the training set to achieve a balanced distribution of samples between the two classes. Random undersampling is easy to implement and computationally inexpensive, but it can lead to a loss of useful information, especially if the majority class contains informative samples. Therefore, it is crucial to carefully select the appropriate ratio of minority and majority samples for undersampling to ensure that the model's performance is not adversely affected (Ganganwar (2012)).

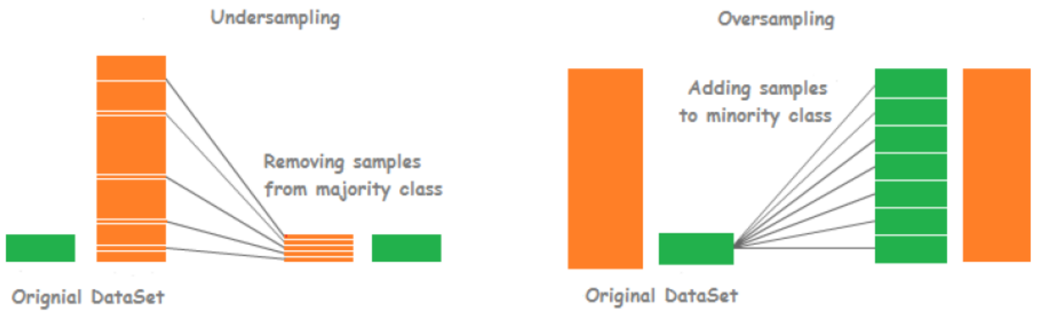


Figure 6: Random Oversampling and Undersampling techniques (R. Mohammed et al. (2020))

2.7.2 SMOTE

Synthetic Minority Over-sampling Technique or SMOTE is an oversampling approach where the minority class is oversampled using synthetic observations and not by simple oversampling with replacement (Chawla et al. (2002)). The idea for this approach came from a successful technique in handwritten character recognition. Instead of using application-specific features, such as rotation and warping, synthetic samples were generated in feature space. In oversampling the minority class, synthetic examples are introduced along the line segments that join any or all of the k nearest neighbors. Each instance of the minority class is taken for this purpose. Depending on the amount of oversampling required, five nearest neighbors are currently used in the implementation, from which the neighbors are randomly selected. The visualization of this procedure is presented in figure 7.

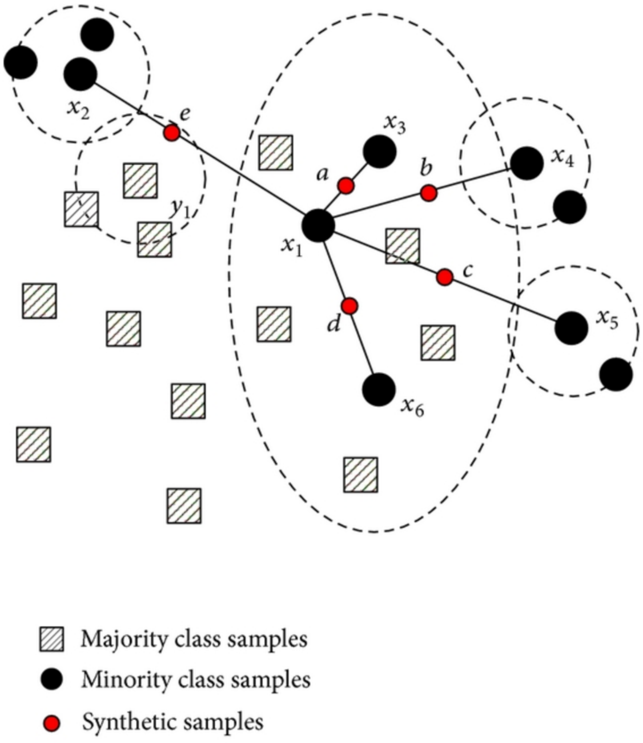


Figure 7: Smote visualization (Rida (2019))

The synthetic examples are generated by taking the difference between the feature vector (sample) under consideration and its nearest neighbor, multiplying this difference by a random number between 0 and 1, and then adding it to the feature vector. This creates a random point along the line segment between two specific features, making the decision area of the minority class more general (Chawla et al. (2002)). This approach effectively eliminates the bias in the model caused by class imbalance, resulting in better model performance.The algorithm 1 shows the step-by-step procedure that is implemented to balance the minority class with the majority class.

Algorithm 1 SMOTE Algorithm (Chawla et al. (2002))

```
1: Initialize the synthetic sample set $S$ to be empty.
2: For each minority sample $x_i$ in the dataset, find its $k$ nearest neighbors.
       $N_i = find\_k\_nearest\_neighbors(x_i, k)$
3: For each minority sample $x_i$, select $n$ samples from its $k$ nearest neighbors and generate $n$
   synthetic samples between $x_i$ and each of the $n$ selected neighbors. The synthetic sample is
   created as follows:
       $x_{i,new} = x_i + \lambda(x_{nn} - x_i)\ S \quad = S \cup x_{i,new}$
   where $x_{nn}$ is one of the $n$ nearest neighbors of $x_i$, $\lambda$ is a random number between 0 and 1,
   and $x_{i,new}$ is the newly generated synthetic sample.
4: If the desired balance between classes is achieved or if the maximum number of iterations is
   reached, terminate the algorithm. Otherwise, go to step 2.
       if $\frac{|\text{minority class}|}{|\text{majority class}|} \geq$ desired balance or max iterations reached: return
```

2.7.3 ADASYN

The key idea of the Adaptive synthetic sampling or ADASYN algorithm is to utilize a density distribution $\hat{r}_i$ as a criterion to automatically determine the number of synthetic samples that should be generated for each minority data instance. $\hat{r}_i$ is a measure of the distribution of weights for different minority class instances based on their level of difficulty in learning. The resulting dataset after applying ADASYN not only provides a balanced representation of the data distribution, according to the desired balance level defined by the β coefficient, but also forces the learning algorithm to focus on those instances that are difficult to learn. This is a significant difference from the SMOTE algorithm, where the same number of synthetic samples are generated for each minority data instance (He et al. (2008)).

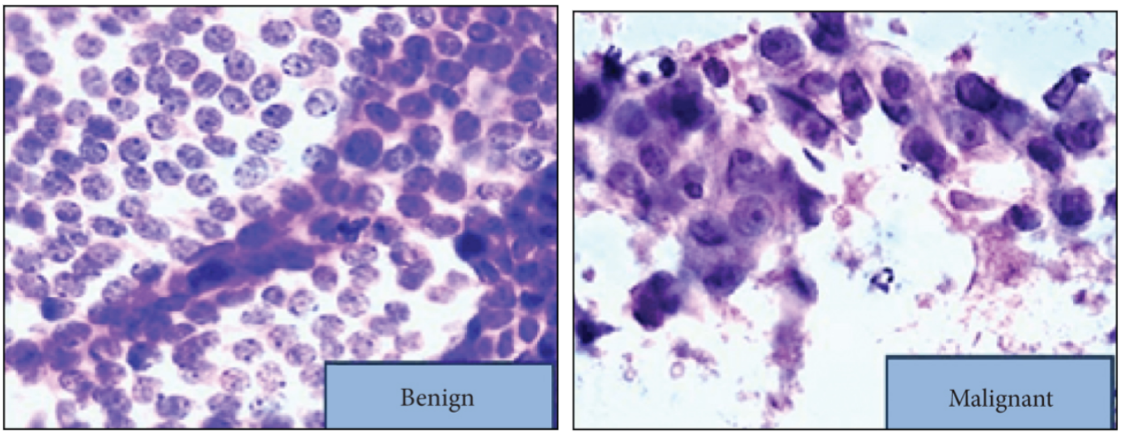
Algorithm 2 ADASYN Algorithm (He et al. (2008))

```
Initialize the synthetic sample set $S$ to be empty.
Compute the density distribution function $g(x)$ for each sample $x$ in the minority class.
Compute the relative importance of each sample $x$ based on its density distribution $g(x)$ as:
    $$\omega_x = \frac{g(x)}{\sum_{x' \in X_{min}} g(x')}$$
Compute the number of synthetic samples to generate for each minority sample $x_i$ as:
    $$G_i = \lfloor \omega_i \times N_{maj} \rfloor$$
    where $N_{maj}$ is the number of samples in the majority class.
For each minority sample $x_i$, select $G_i$ samples from its $k$ nearest neighbors and generate $G_i$ synthetic samples between $x_i$ and each of the $G_i$ selected neighbors. The synthetic sample is created as follows:
    $$x_{i,new} = x_i + \lambda(x_{nn} - x_i) \; S \quad = S \cup x_{i,new}$$
    where $x_{nn}$ is one of the $G_i$ nearest neighbors of $x_i$, $\lambda$ is a random number between 0 and 1, and $x_{i,new}$ is the newly generated synthetic sample.
Combine the original minority class and the synthetic samples to form the new minority class.
```

3 Application

3.1 Exploratory Analysis

In this study, the Breast Cancer Wisconsin (Diagnostic) Data Set will be examined in order to utilize and evaluate different machine learning algorithms. With 569 observations and 30 features, this dataset presents a rich and vast source of information that will be thoroughly investigated using a diverse range of techniques and visualizations to uncover hidden trends and relationships that may not be readily apparent. As it is presented in figure 8, the features are related to the image of the benign or malignant tumor cells. The primary objective of this analysis is to determine the basic statistics of these features such as mean values and standard deviations, first and third quantiles and extreme values to gain deeper insights into the underlying characteristics of the data that are inherent to breast cancer tumors.



(a) (b)

Figure 8: Images of benign and malignant tumor cells (Mohammad et al. (2022))

In the forthcoming analysis, ten out of thirty features will be chosen for examination. These features are represented by the mean values of various characteristics exhibited by cancerous tumors. All variables are numeric, and there are no missing values. The dataset was partitioned into two distinct subsets, namely the training set and the test set, in accordance with a predetermined ratio of 0.8 to 0.2, respectively. The specific variables concerning the cancerous tumors for the train set that are identified for the analysis, are as follows:

1. Radius
The variable radius_mean exhibits a distribution with a mean value of 14.127 and a standard deviation of 3.524. The distributions of this variable for each class do not significantly overlap, suggesting that it can be considered a reliable indicator for discriminating between the two classes. The first quartile of the data accumulates at a value of 11.7, while the third quartile accumulates at a value of 15.7.

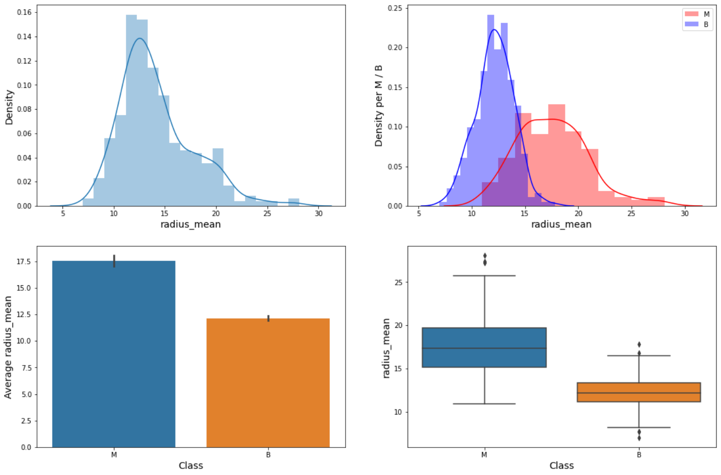


Figure 9: Top left: Distribution of the variable radius_mean, Top right: Distribution of the variable radius_mean per class, Bottom left: Average value per class for the variable radius_mean, Bottom right: boxplots per class for variable radius_mean

2. Area

The variable area_mean has a distribution with a mean value of 654.88 and a standard deviation of 351.91 that characterizes it. This variable may be a reliable indicator for differentiating between the two classes because the distributions for each class show little overlap. The data accumulates at a value of 420.3 for the first quartile and 782.7 for the third quartile.

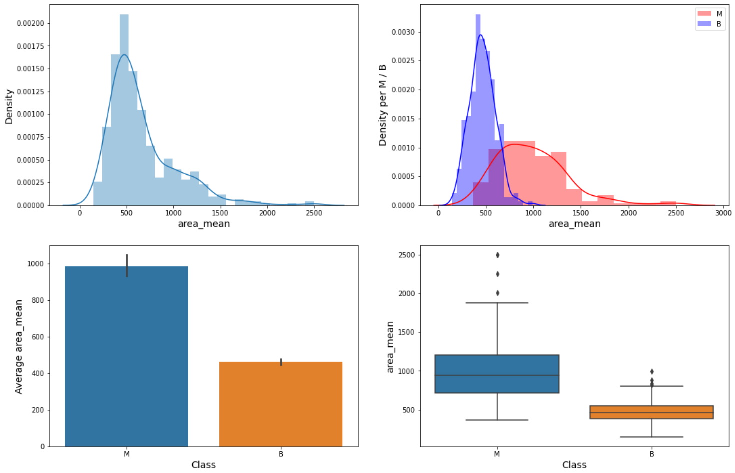


Figure 10: Top left: Distribution of the variable area_mean, Top right: Distribution of the variable area_mean per class, Bottom left: Average value per class for the variable area_mean, Bottom right: boxplots per class for variable area_mean

3. Compactness

The variable compactness_mean displays a distribution with a mean of 0.104 and a standard deviation of 0.053. The distributions of this variable for each class show mediocre overlap, suggesting that it might not be a reliable indicator for distinguishing between the two classes. The first quartile of the data concentrates at a value of 0.065, while the third quartile concentrates at a value of 0.130.

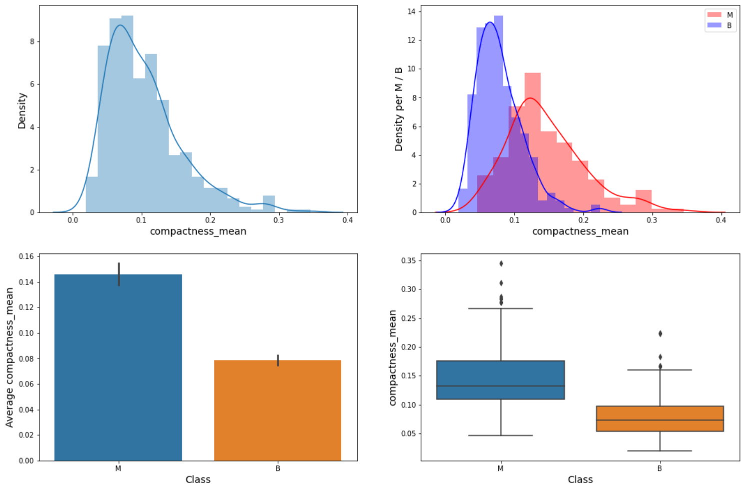


Figure 11: Top left: Distribution of the variable compactness_mean, Top right: Distribution of the variable compactness_mean per class, Bottom left: Average value per class for the variable compactness_mean, Bottom right: boxplots per class for variable compactness_mean

4. Concave points

The variable concave_points_mean exhibits a distribution characterized by a mean value of 0.049 and a standard deviation of 0.039. This distribution suggests that concave_points_mean may serve as a dependable indicator for distinguishing between the two classes, as there is minimal overlap in the distributions for each class. The first quartile of the data concentrates at a value of 0.02, while the third quartile concentrates at a value of 0.074.

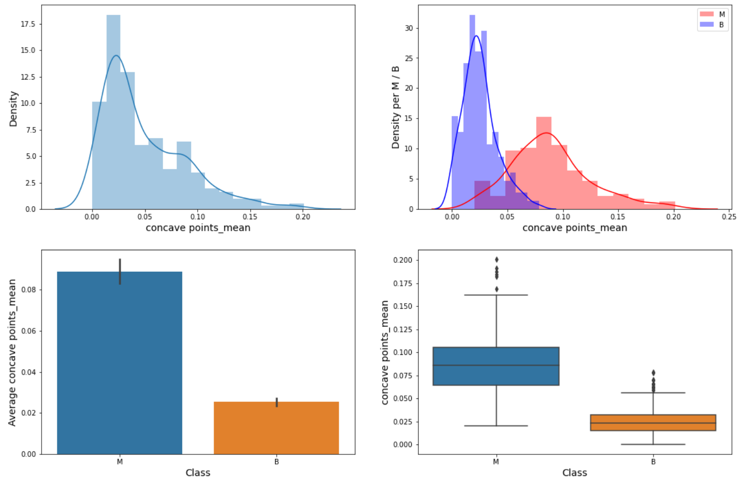


Figure 12: Top left: Distribution of the variable concave_points_mean, Top right: Distribution of the variable concave_points_mean per class, Bottom left: Average value per class for the variable concave_points_mean, Bottom right: boxplots per class for variable concave_points_mean

5. Concavity

The variable concavity_mean demonstrates a distribution characterized by an average value of 0.089 and a standard deviation of 0.079. The distributions per class suggest that concavity_mean may be a reliable indicator for distinguishing between the two classes, as there is minimal overlap between them. The first quartile of the data is accumulated by the value of 0.029, while the third quartile is centered around a value of 0.131.

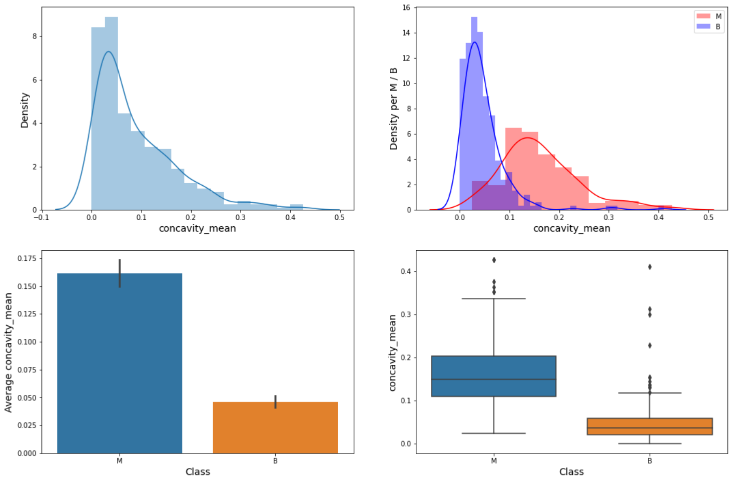


Figure 13: Top left: Distribution of the variable concavity_mean, Top right: Distribution of the variable concavity_mean per class, Bottom left: Average value per class for the variable concavity_mean, Bottom right: boxplots per class for variable concavity_mean

6. Fractal Dimension

The distribution for the variable fractal_dimension_mean has a mean of 0.063 and a standard deviation of 0.007. Significant overlap between the distributions of this variable for each class indicates that it might not be a valid indicator for differentiating between the two classes. The data's first quartile is concentrated at 0.058, while the third quartile is concentrated at 0.066.

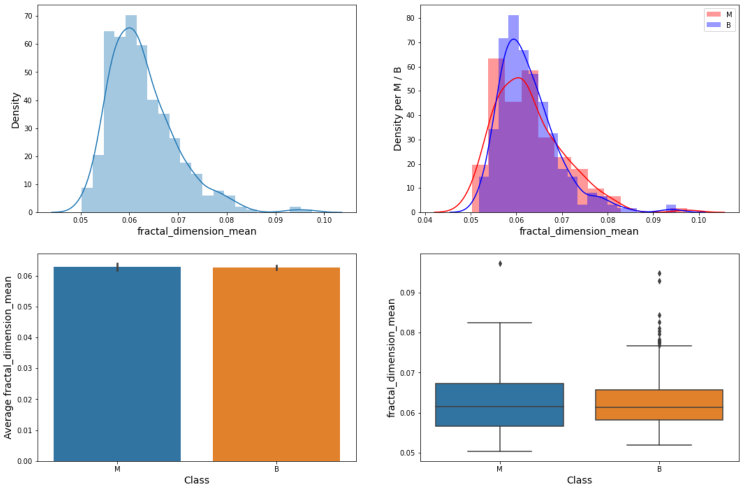


Figure 14: Top left: Distribution of the variable fractal_dimension_mean, Top right: Distribution of the variable fractal_dimension_mean per class, Bottom left: Average value per class for the variable fractal_dimension_mean, Bottom right: boxplots per class for variable fractal_dimension_mean

7. Perimeter

The variable perimeter_mean demonstrates a distribution with a mean value of 91.96 and a standard deviation of 24.29. The distributions per class suggest that perimeter_mean may be a reliable indicator for distinguishing between the two classes, as there is minimal overlap between them. The first quartile of the data is accumulated by the value of 75.17, while the third quartile is centered around a value of 104.1.

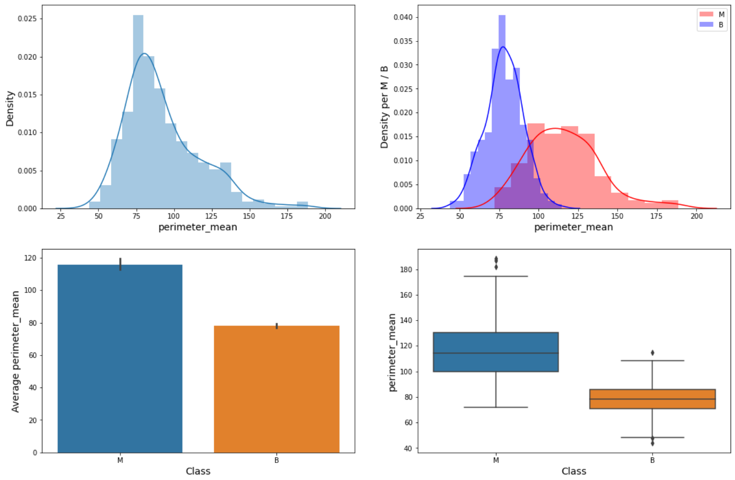


Figure 15: Top left: Distribution of the variable perimeter_mean, Top right: Distribution of the variable perimeter_mean per class, Bottom left: Average value per class for the variable perimeter_mean, Bottom right: boxplots per class for variable perimeter_mean

8.Smoothness

The distribution for the variable fractal_dimension_mean has a mean of 0.096 and a standard deviation of 0.014. Significant overlap between the distributions of this variable for each class indicates that it might not be a valid indicator for differentiating between the two classes. The data's first quartile is concentrated at 0.086, while the third quartile is concentrated at 0.105.

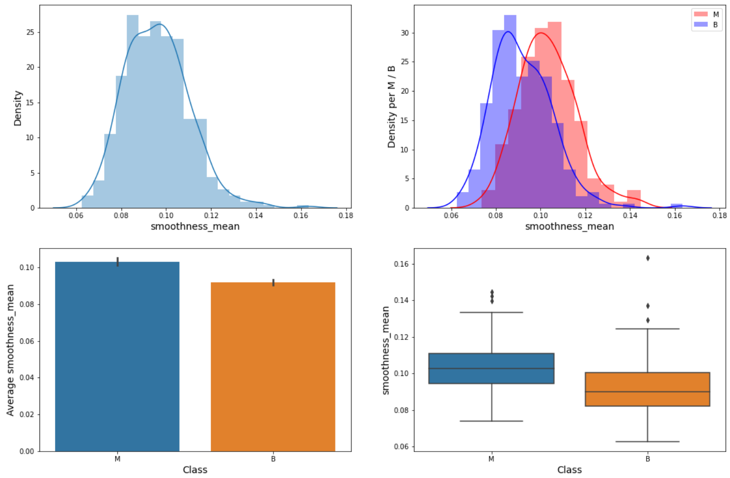


Figure 16: Top left: Distribution of the variable smoothness_mean, Top right: Distribution of the variable smoothness_mean per class, Bottom left: Average value per class for the variable smoothness_mean, Bottom right: boxplots per class for variable smoothness_mean

9. Symmetry

The mean and standard deviation of the distribution for the variable symmetry_mean are 0.181 and 0.027, respectively. This variable's distributions for the two groups significantly overlap, which suggests that it might not be a reliable indicator for discriminating between the two classes. The first and third quartiles of the data are concentrated at 0.161 and 0.196, respectively.

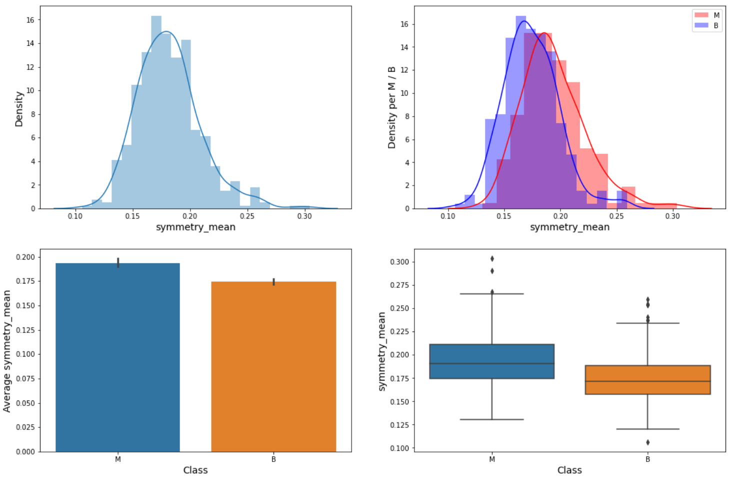


Figure 17: Top left: Distribution of the variable symmetry_mean, Top right: Distribution of the variable symmetry_mean per class, Bottom left: Average value per class for the variable symmetry_mean, Bottom right: boxplots per class for variable symmetry_mean

10. Texture

For the variable texture_mean, the distribution's mean and standard deviation are, respectively, 19.28 and 4.3. The distributions of this variable for the two groups greatly overlap, which raises the possibility that it may not be an accurate marker for differentiating between the two classes. The data are concentrated in the first and third quartiles at 16.17 and 18.84, respectively.

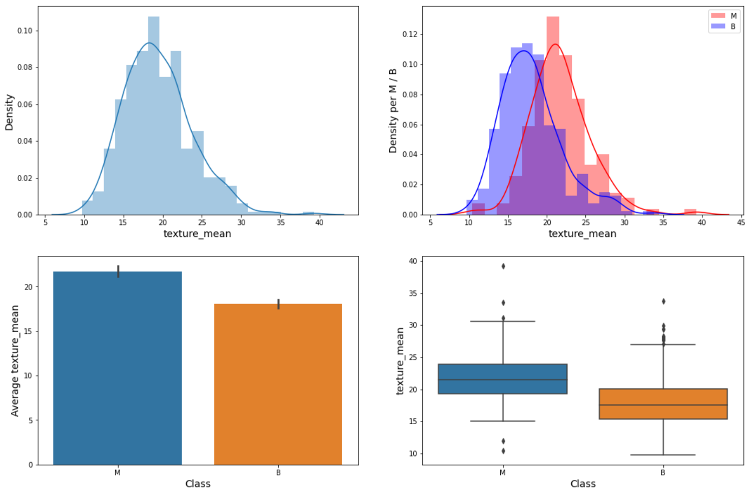


Figure 18: Top left: Distribution of the variable texture_mean, Top right: Distribution of the variable texture_mean per class, Bottom left: Average value per class for the variable texture_mean, Bottom right: boxplots per class for variable texture_mean

In Figure 19, the countplot displays the distribution of the target variable in the dataset, revealing that the negative class (benign tumors) comprises a significant majority, accounting for 62.74% of the samples, while the positive class (malignant tumors) represents a smaller proportion of 37.26%. The heights of the bars on the countplot depict the frequency of occurrences for each class, with the vertical axis indicating the counts and the horizontal axis denoting the class labels. It can be observed that the negative class dominates the dataset, surpassing the positive class. The countplot provides a visual representation of this class imbalance, which should be taken into consideration during model development and evaluation to account for potential biases and ensure optimal model performance.

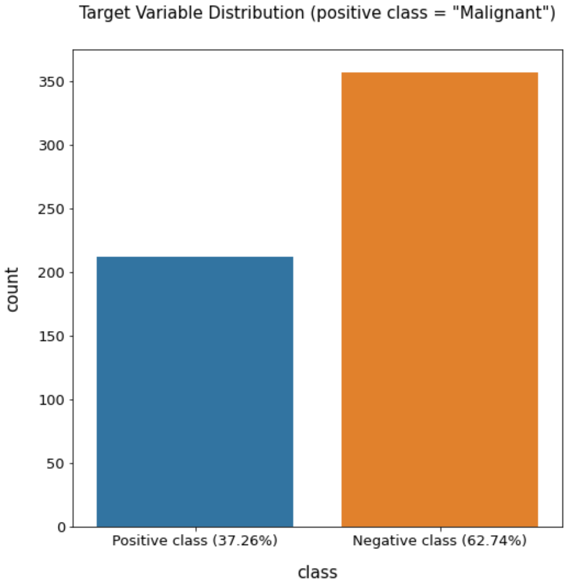


Figure 19: Target Variable Distribution

3.2 Methodology

One of the most common and deadly diseases is cancer that affects millions of individuals globally. Cancer detection and treatment have advanced significantly thanks to medical research, but the subject is still complicated and challenging and calls for ongoing improvements in methodology and technology. Machine learning algorithms have recently shown considerable potential in the field of medical research, particularly in the analysis of vast amounts of data for the diagnosis of malignant tumors. However, the quality and balance of the training data have a significant impact on how well these algorithms perform. Machine learning algorithms may perform poorly in unbalanced datasets where one type of tumor predominates over the other, leading to biased models that do not identify the minority class (Mohammad et al. (2022)).

In medical research on cancer tumors, handling unbalanced datasets is crucial to ensuring that the models produce accurate and trustworthy predictions. Missed diagnosis, postponed treatments, and more severe mortality rates could result from biased algorithms that are unable to identify the minority class. Falsely positive tests may also result in therapies that are not essential and may cause patients' adverse side effects and psychological discomfort. Hence, researchers should develop effective strategies for handling unbalanced datasets and improve the efficiency of machine learning algorithms in cancer tumor detection. Resampling, ensemble learning, and cost-sensitive learning are examples of specialized techniques that can be used to reduce the effects of unbalanced datasets and provide objective and precise predictions for medical diagnosis and treatment (Mohammad et al. (2022)).

Random undersampling, random oversampling, SMOTE (Synthetic Minority Over-sampling Technique), and ADASYN (Adaptive Synthetic Sampling) are techniques used to handle imbalanced datasets in machine learning algorithms.

3.2.1 Cancer Tumor Classification Problem

The Wisconsin Breast Cancer Dataset (WBCD) has been widely used for classification studies due to its characteristics, which include a relatively small number of input variables and binary classification. In this study, a classification problem on the BCW dataset will be tackled using several machine learning algorithms, namely Lasso Logistic Regression, Decision Trees, Random Forests, XGBoost, and Multilayer Perceptrons (MLP).

Lasso Logistic Regression is a variant of Logistic Regression that uses L1 regularization to prevent overfitting and improve the model's interpretability. It is useful for datasets with many features, as it can automatically perform feature selection by reducing the coefficients of irrelevant variables to zero. Decision Trees, on the other hand, are simple and powerful models that can capture nonlinear relationships between variables. Random Forests are an extension of Decision Trees that use multiple trees and bagging to reduce overfitting and improve the model's generalizability.

XGBoost is a gradient boosting algorithm that has achieved state-of-the-art performance in various machine learning competitions. It uses an ensemble of decision trees and gradient descent to minimize the loss function, resulting in a highly accurate and interpretable model. Finally, MLPs are a type of artificial feedforward neural network that can learn complex relationships between variables. They consist of multiple layers of interconnected neurons that can capture nonlinearities and interactions in the data.

In this study, the performance of these algorithms will be compared in terms of accuracy, precision, recall, area under the ROC curve. Furthermore, sampling techniques such as Random OverSampling, SMOTE and ADASYN will be applied in tree based algorithms namely Decision Trees, Random Forests and XGBoost. The BCW dataset is preprocessed to standardize the input variables. Gridsearch will be used to tune the hyperparameters of each algorithm and avoid overfitting. The results will be analyzed to determine which algorithm performs best and provide insights into the factors that influence breast cancer diagnosis.

3.2.2 Experiment Design and Architecture

In this study, five different classifiers will be examined to classify observations of cancer tumors as either benign or malignant. Each classifier will be assessed based on its own parameters, and their results will be compared in terms of accuracy, precision, and recall. Tree-based methods will use sampling techniques to handle the imbalance problem, and the models will be evaluated based on the area under the ROC curve. The implementation of Lasso Logistic Regression is based on the glmnet package in R, while the Multilayer Perceptron and tree-based methods are implemented in Python using the PyTorch and scikit-learn libraries.

Lasso Logistic Regression

Lasso Logistic Regression will be performed using the *cv.glmnet* function of *glmnet* package in R. *glmnet* is a package that uses penalized maximum likelihood to fit generalized linear models. For the regularization parameter lambda, the regularization route is calculated for the lasso or elastic net penalty at a grid of values (on the log scale). The algorithm is very quick and takes use of the input matrix's sparsity. *glmnet* solves the problem:

$$\min_{(\beta_0,\beta)\in R^{p+1}} \left[-\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + x_i^\top \beta) - \log\left(1+\exp(\beta_0 + x_i^\top \beta)\right) \right] + \lambda \left[(1-\alpha)\frac{||\beta||_2^2}{2} + \alpha||\beta||_1 \right] \quad (36)$$

over a grid of values of λ covering a wide range of possible solutions. β is a vector of model coefficients, λ is the regularization parameter, and α controls the weighting between the Lasso (L1) and Ridge (L2) penalties in the Elastic Net regularization. In this study, the goal is to perform feature selection, and hence, the Lasso penalty will be utilized by setting the parameter α to 1. The *cv.glmnet* function uses k-fold cross-validation to evaluate the model on different values of λ. Two values along the λ sequence are of major importance, as well as their corresponding β coefficients. These values are λ_{min}, which returns the minimum mean cross-validated error, while λ_{1se} is the value of λ that gives the most regularized model, such that the cross-validation error is within one standard error of the minimum value.

Due to the high imbalance of the data, it is important to determine the proper cut-off point for the probability of the class threshold. This can be calculated using the *coords* function from the *pROC* R package. The associated metrics will be evaluated for both thresholds: the default threshold, which is 0.5, and the best threshold determined using the Youden's statistic.

Multi Layer Perceptron

In this study, Multi Layer Perceptron will be used to implement a feedforwrd architecture of neural networks. The input layer will consist of 10 neurons, which corresponds to the number of variables to be analyzed. The hidden layer will contain 8 neurons that are fully connected with the input neurons. The output layer will have 2 neurons since it needs to predict two classes for the binary classification problem. The learning rate of the gradient descent algorithm is set to 10^{-3} and the loss function applied to the logits of the output layer is the Cross Entropy Loss Function:

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{37}$$

Here, $\hat{y}$ is the predicted probability vector, y is the true label vector, and N is the number of samples. Furthermore, due to the imbalance of the dataset, it is necessary to calculate the proper cut-off threshold based on the Youden's index.

Tree Based Methods with sampling techniques

The imbalance of the dataset forces the analysis to move towards sampling techniques in order to produce models that are not biased on the majority class. Three classifiers will be examined using sampling techniques: Decision Trees, Random Forests, and XGBoost. Their performance will be compared to the same classifiers without sampling. Moreover, Youden's index will be calculated in order determine the best cut-off point to balance between sensitivity and specificity. To achieve the best possible prediction performance, GridSearch will be applied to each classifier. These classifiers are from the *scikit learn* Python library, and the parameters to be examined come from their respective classes in the library. The tables 2, 3 and 4 summarize the grid of parameters for each classifier.

Parameter	**Values**	**Description**
criterion	gini, entropy	Splitting criterion
min_samples_leaf	25, 50, 100	Minimum number of samples required to be at a leaf node
min_impurity_decrease	0.0, 0.05, 0.1	Minimum impurity decrease required for a split

Table 2: GridSearch Parameters for Decision Tree

Parameter	**Values**	**Description**
n_estimators	50, 100, 200	Number of trees in the forest
max_depth	3, 4, 5, 6	Maximum depth of the tree
min_samples_leaf	10, 20, 30	Minimum number of samples required to be at a leaf node

Table 3: GridSearch Parameters for Random Forest

Parameter	**Values**	**Description**
n_estimators	50, 100, 200	Number of boosting trees
max_depth	3, 4, 5, 6	Maximum depth of each tree
learning_rate	0.1, 0.01	Learning rate for the boosting process

Table 4: GridSearch Parameters for XGBoost

3.3 Results

3.3.1 Logistic Regression

Lasso Logistic Regression is implemented using the function *cv.glmnet* from *glmnet* package in R. This technique acts like a feature selection method because some of the coefficients are pushed towards zero. The *cv.glmnet* function is used to perform cross-validation for a range of lambda values in the context of fitting a generalized linear model. When *cv.glmnet* runs for a logistic regression model with a binary response variable, the function returns a plot of the cross-validation error versus the log of the lambda sequence.

The plot, as it is presented in figure 20, shows how the cross-validation error changes as the penalty parameter lambda is varied. The x-axis of the plot represents the log of the lambda sequence, and the y-axis represents the cross-validation error for misclassification, which is a measure of how well the model predicts new data that was not used in fitting the model. The plot indicates two specific values of lambda regularization parameter that are important for the model implementation. These values along the λ sequence are of major importance, as well as their corresponding β coefficients. These values are λ_{min}, which returns the minimum mean cross-validation error, while $\lambda_{1\text{se}}$ is the value of λ that gives the most regularized model, such that the cross-validation error is within one standard error of the minimum value. For this dataset, $\lambda_{\text{min}} = 0.00827531$ and $\lambda_{1\text{se}} = 0.03340757$.

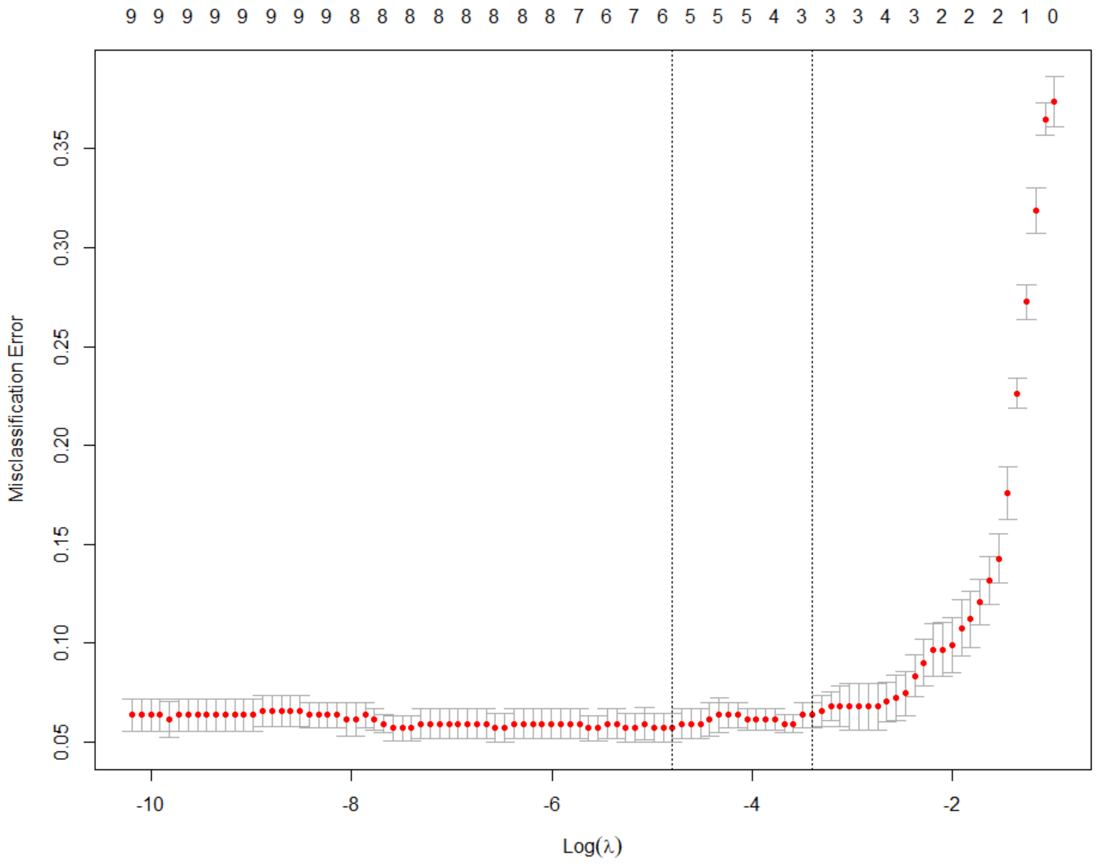


Figure 20: Cross Validation lambda sequence

The corresponding β coefficients for λ_{min} and λ_{1se} are presented in Table 5. In general, the lambda value with the minimum cross-validation error (λ_{min}) will provide the model with the best predictive performance. However, this lambda value may result in a model that is too complex and overfits the training data. The lambda value with the largest lambda within one standard error of the minimum (λ_{1se}) can be used to obtain a simpler model that has slightly worse predictive performance, but may generalize better to new data. The idea behind using λ_{1se} is that it is less likely to overfit the training data compared to λ_{min}. As can be observed from Table 5, seven out of the eleven coefficients of the λ_{1se} model are set to zero. As a result, a simpler model than the one corresponding to λ_{min} is found with *cv.glmnet*. Therefore, predictions will be made using the simpler model of λ_{1se}.

Variables	λ_{1se}	λ_{min}
intercept	-0.577	-0.579
radius_mean	0.979	1.826
texture_mean	0.532	1.135
perimeter_mean	0	0
area_mean	0	0.335
smoothness_mean	0	0.427
compactness_mean	0	0
concavity_mean	0	0
concave.points_mean	2.018	2.353
symmetry_mean	0	0.189
fractal.dimension_mean	0	0

Table 5: β coefficients for λ_{min} and λ_{1se} for Lasso Logistic Regression optimization problem

In order to calculate the prediction metrics on the test set, the dataset's imbalance must be taken into account. Youden index will be calculated for different cut-off points to determine the one that maximizes this statistic. The maximization of Youden index indicates the balance between the true positive rate and true negative rate, and therefore optimizes the model's resolution between the two discreet classes. For this specific dataset, Youden index is maximized at a cut-off point of 0.2523426 for the positive class. This means that if an observation corresponds to a probability beyond this threshold, it will be classified as a malignant tumor.

The test set used to evaluate the model has high performance metrics, with an area under the receiver operating characteristic curve (AUC) of 0.99. This means that the predictive power of the model is very high using different cut-off point probabilities. The model also achieves high recall of 0.875, indicating that it can accurately identify the majority of positive cases. Moreover, the precision metric of 1.00 suggests that when the model predicts a positive case, it is almost always correct. The model also has a high accuracy score of 0.921, which indicates that it is able to correctly classify the majority of both positive and negative cases. Overall, the performance metrics of the model on the test set suggest that it is highly effective in accurately identifying positive cases while maintaining a low rate of false positives.

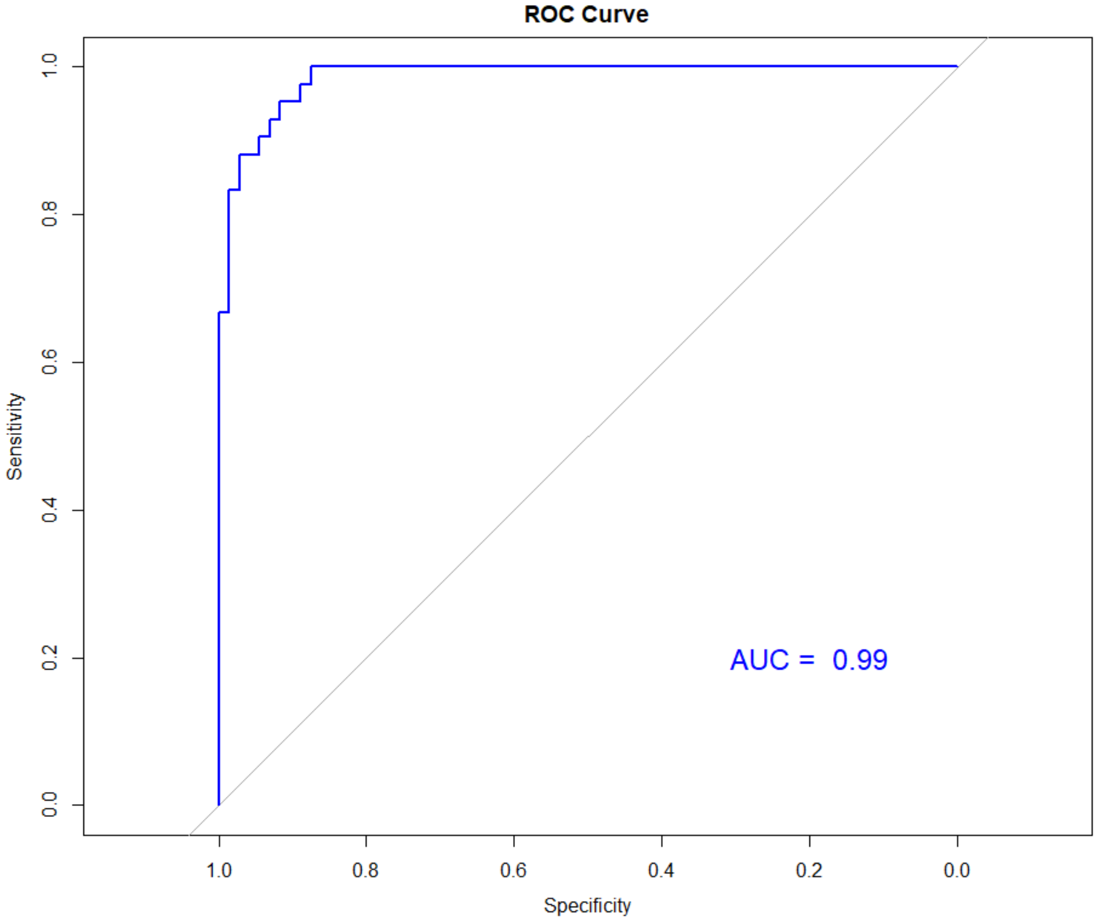


Figure 21: Reciever Operator Curve for Lasso Logistic Regression

3.3.2 Multi Layer Perceptron

The multilayer perceptron is trained for 50 epochs and the training and test loss in each epoch is visualized in figure 22. As it is presented in this figure, the main conclusions that are arisen are:

1. **The model is overfitting after the twentieth epoch:** Overfitting occurs when a model becomes too complex and starts to memorize the training data instead of learning generalizable patterns. As a result, the model's performance on new, unseen data (testing data) starts to degrade. If the training loss is consistently lower than the testing loss after the twentieth epoch, it suggests that the model is well-regularized and has learned generalizable patterns during the initial training epochs. However, after the twentieth epoch, the model starts to overfit to the training data, leading to a higher testing loss.

2. **The model may benefit from early stopping:** Early stopping is a regularization technique that stops the training process when the performance on the validation set starts to degrade. If the model is overfitting after the twentieth epoch, it suggests that the training process could benefit from early stopping to prevent the model from memorizing the training data and to improve its generalization performance.

3. **The model may benefit from more regularization**: If the training loss is consistently higher than the testing loss during the first twenty epochs and lower after the twentieth epoch, it suggests that the model is not fully capturing the underlying patterns in the data and is overfitting to the training set. In this case, adding more regularization techniques such as dropout, weight decay, or early stopping may help the model generalize better to unseen data. Increasing the amount of regularization can help the model learn more robust and generalized features by reducing the model's sensitivity to the specific examples in the training data.

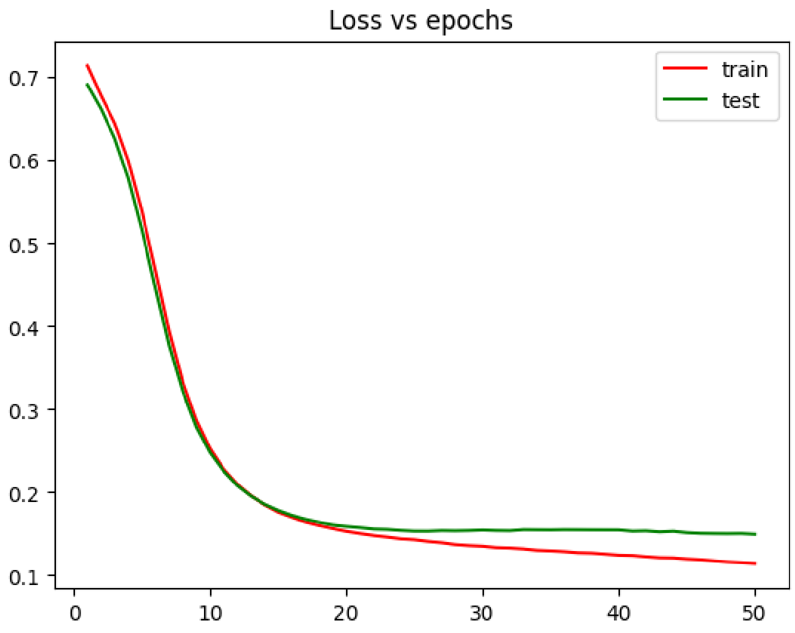


Figure 22: Training and Testing loss for the multilayer perceptron

Youden's index must be calculated for this specific dataset due to its inherent imbalance. The cut-off point that balances the true positive rate and the true negative rate, and maximizes the Youden's index statistic, is determined to be 0.546. This means that if an observation corresponds to a probability beyond this threshold, it will be classified as a malignant tumor.

The testing set used to evaluate the performance of a multilayer perceptron (MLP) model has yielded impressive results. The area under the receiver operating characteristic curve (AUC) is 0.987, indicating that the model's ability to distinguish between positive and negative classes is excellent. The recall score of 0.95 suggests that the model can correctly identify 95% of the actual positive cases in the test set. The precision score of 0.93 indicates that the model is precise in its predictions, as it correctly identifies 93% of the positive predictions. Finally, the accuracy score of 0.96 shows that the model is able to make correct predictions for 96% of the total cases in the test set. These metrics indicate that the MLP model has performed well on the testing set and can be considered as a reliable predictor for the specific task.

The confusion matrix visualized in Figure 23 depicts the overall performance of the MLP model. A True Positive rate of 0.95 suggests that the model has a high predictive power for positive outcomes, as does the True Negative rate of 0.96 for negative outcomes. However, the False Positive rate of 0.042 indicates that around 4% of positive predictions are falsely classified as malignant tumors, when they are actually benign. Similarly, the False Negative rate of 0.048 suggests that around 4.8% of negative predictions are falsely classified as benign tumors, when they are actually malignant.

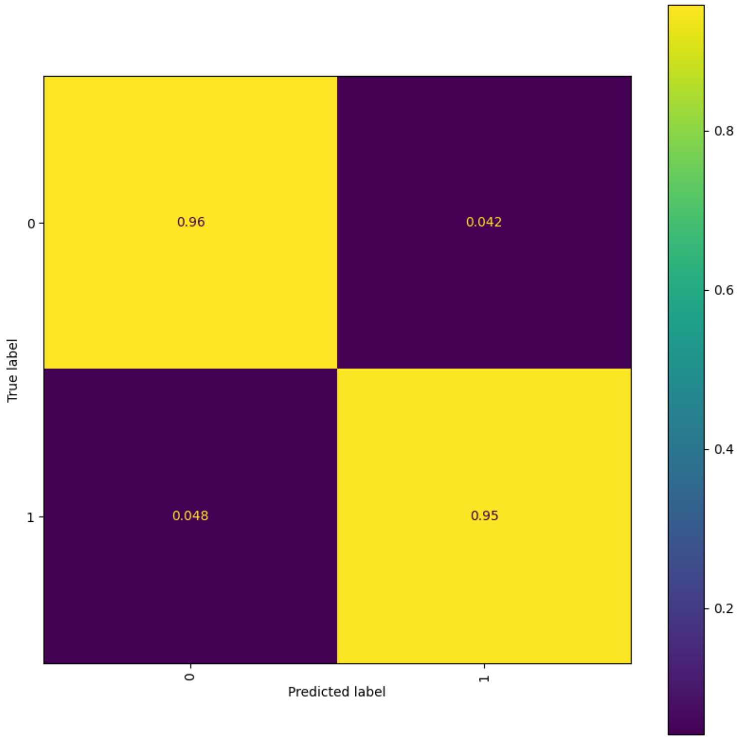


Figure 23: Confusion Matrix on the test set for multilayer perceptron

3.3.3 Tree based methods with sampling techniques

Decision Trees

In the field of machine learning and data analysis, decision tree models are widely used for classification tasks. Decision trees provide a transparent and interpretable framework for making predictions based on a set of rules derived from the training data. They partition the feature space into regions that correspond to different class labels, allowing for accurate classification of new instances.

However, when dealing with imbalanced datasets, where one class significantly outnumbers the other, decision tree models can face challenges in accurately representing and classifying the minority class. In the context of classifying benign and malignant tumors, the presence of class imbalance poses a significant problem. The minority class (malignant tumors) may be underrepresented, leading to biased predictions and reduced performance.

To address this issue, various sampling techniques have been developed to rebalance the class distribution and improve the performance of decision tree models. Random Oversampling, Random Undersampling, SMOTE (Synthetic Minority Over-sampling Technique), and ADASYN (Adaptive Synthetic Sampling) are among the commonly used sampling techniques. For each sampling technique, a Gridsearch is performed to find the optimal parameters that maximize the Area under the ROC Curve. The results on the test set for decision trees with the corresponding sampling techniques are presented in Table 6.

Sampling Technique	**Cut-off**	**Accuracy**	**AUC**	**Recall**	**Precision**
SMOTE	0.30	0.868	0.886	0.952	0.944
Random Oversampling	0.55	0.903	0.909	0.927	0.830
Random Undersampling	0.64	0.903	0.904	0.905	0.844
ADASYN	0.78	0.903	0.909	0.929	0.830
No Sampling	0.42	0.903	0.904	0.904	0.844

Table 6: Decision Tree evaluation metrics on the test set

The summarization of these results is presented below:

1. SMOTE: This technique achieved an accuracy of 0.868 and an AUC of 0.886. It performed well in terms of recall with a score of 0.952, indicating its ability to correctly identify positive cases. The precision score of 0.944 also suggests a low rate of false positives. Overall, SMOTE showed a balanced performance in terms of accuracy, AUC, recall, and precision.

2. Random Oversampling: With an accuracy of 0.903 and an AUC of 0.909, this technique improved upon the performance of SMOTE. It achieved a good recall score of 0.927, indicating its ability to correctly identify positive cases. However, the precision score of 0.830 suggests a higher rate of false positives compared to SMOTE.

3. Random Undersampling: This technique also achieved an accuracy of 0.903, similar to Random Oversampling. It had a slightly lower AUC of 0.904 but maintained a good recall score of 0.905. The precision score of 0.844 indicates a relatively low rate of false positives. Random Undersampling provided a balanced performance in terms of accuracy, AUC, recall, and precision.

4. ADASYN: Similar to Random Oversampling, ADASYN achieved an accuracy of 0.903 and an AUC of 0.909. It showed a good recall score of 0.929, indicating its ability to correctly identify positive cases. However, the precision score of 0.830 suggests a higher rate of false positives compared to other techniques.

5. No Sampling: The decision tree without sampling achieved the same accuracy as Random Undersampling and ADASYN. The AUC score of this technique reached the level of 0.904. It had a precision score of 0.844, similar to Random Undersampling. This technique provided a balanced performance with respect to accuracy, AUC, recall, and precision.

Based on the above analysis, the best model in terms of overall performance, particularly based on the AUC score, is the decision tree with ADASYN. This specific model will be compared to the biased no-sampling model based on the importance of the features, the ROC curves, and the confusion matrices. Figures 24 and 25 show the difference in the importance of the features in each classifier with the corresponding sampler.

The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. As shown below, six out of ten features are participating in the creation of the decision tree rules, particularly in the classifier with the ADASYN sampler. The most important feature is the *concave.points_ mean* and the less important one but with contribution to the total reduction of the criterion is the *concavity*. The biased model with no sampling is constructed with even fewer features, specifically four out of ten. The decision rules of this classifier consists of *concave.points_ mean*, *perimeter_ mean*, *texture_ mean* and *compactness_ mean*.

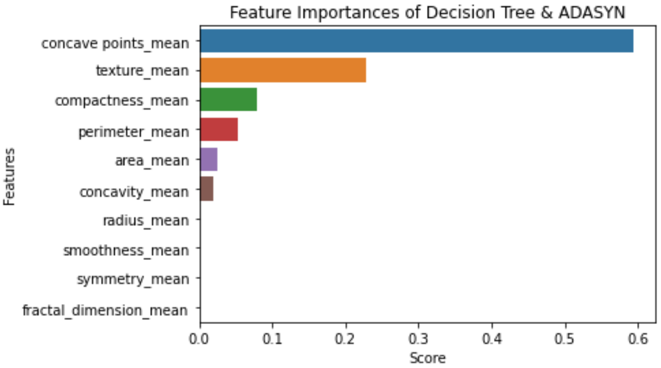


Figure 24: Feature Importance for Decision Tree and ADASYN

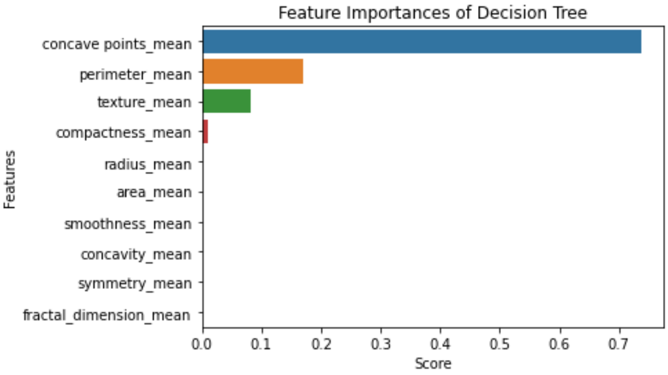


Figure 25: Feature Importance for Decision Tree and no sampling

The performance of classification models is often evaluated using the Receiver Operating Characteristic (ROC) curve. The ROC curve demonstrates its effectiveness in differentiating between benign and malignant tumors in the setting of the model using ADASYN sampling approach, as it is visualized in Figure 26. The decision tree with ADASYN model exhibits strong discriminatory power with an AUC of 0.909, suggesting its capacity to precisely categorize examples from both groups. By maximizing the Youden's index, the 0.78 cut-off point was set to further improve the model's performance by balancing the trade-off between true positive rate and false positive rate. This indicates that the model may provide a high true positive rate while maintaining a reasonably low false positive rate, leading to a more accurate classification of malignant tumors.

The model without sampling, on the other hand, has a somewhat lower AUC of 0.904, as it is visualized in Figure 27. Although it still shows a respectable level of discriminating ability, it falls slightly short of the ADASYN model. Based on the Youden's index, a new threshold for categorizing cases is indicated by the chosen cut-off point of 0.42. The model without sampling can still successfully identify between benign and malignant tumors despite having a lower AUC, but it can have a different ratio of true positive to false positive predictions than the ADASYN model.

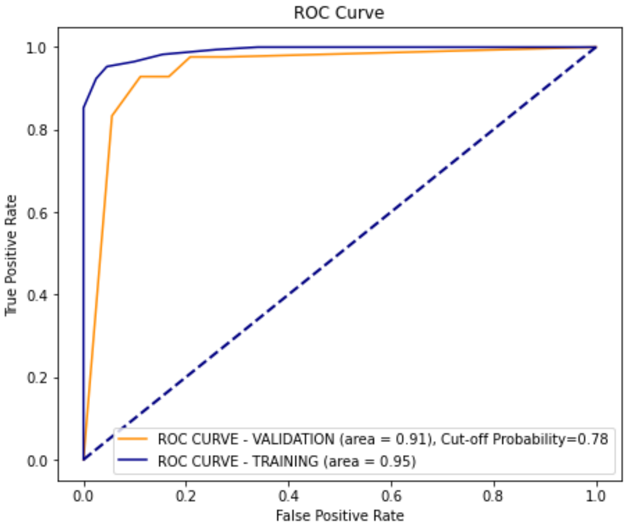


Figure 26: ROC Curve for Decision Tree and ADASYN

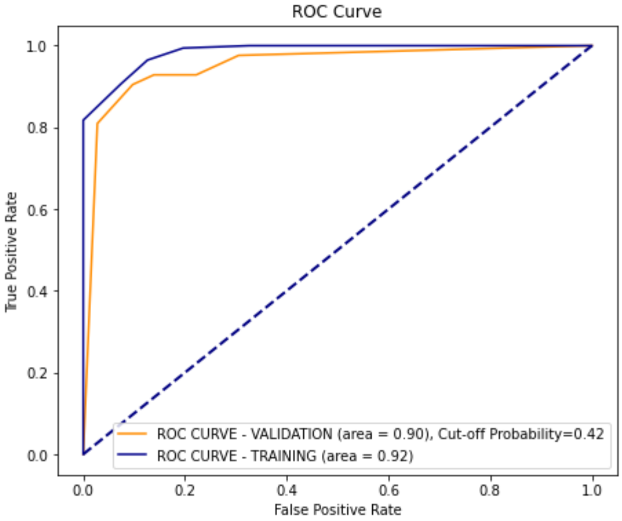


Figure 27: ROC Curve for Decision Tree and no sampling

The confusion matrix visualized in Figure 28 depicts the overall performance of the decision tree with ADASYN model. A True Positive rate of 0.93 suggests that the model has a high predictive power for positive outcomes, as does the True Negative rate of 0.89 for negative outcomes. However, the False Positive rate of 0.071 indicates that around 7% of positive predictions are falsely classified as malignant tumors, when they are actually benign. Similarly, the False Negative rate of 0.11 suggests that around 11% of negative predictions are falsely classified as benign tumors, when they are actually malignant.

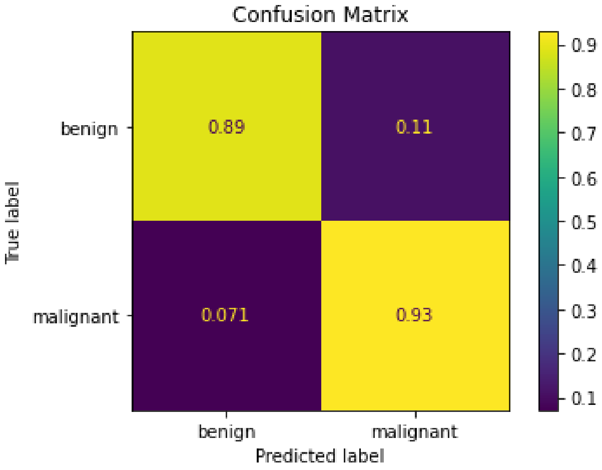


Figure 28: Confusion Matrix on the test set for Decision Tree and ADASYN

The decision tree with no sampling overall effectiveness is shown by the confusion matrix in Figure 29. A true positive rate of 0.9 and a True Negative rate of 0.9 both point to the model's strong prediction ability for malignant and benign outcomes. The False Positive rate of 0.097, however, shows that almost 9.7% of positive predictions are mistakenly labeled as malignant tumors when they are benign. Similarly, the False Negative rate of 0.095 indicates that around 9.5% of negative predictions are mistakenly labeled as benign tumors when they are actually malignant.

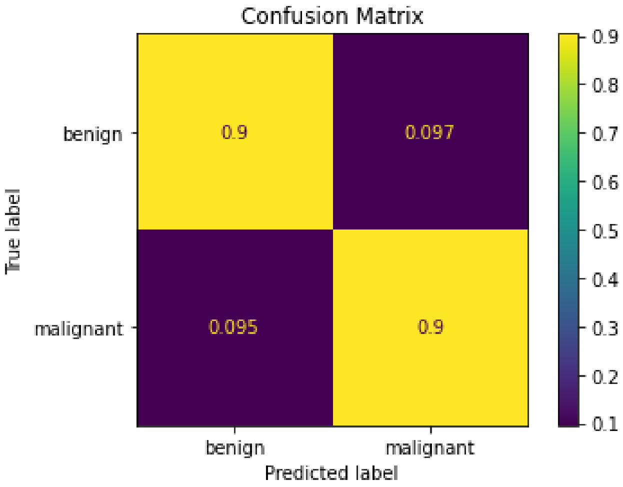


Figure 29: Confusion Matrix on the test set for Decision Tree and no sampling

Random Forests

Balancing the minority class is of utmost importance in random forests to ensure accurate and robust predictions. Random forests are ensemble models composed of multiple decision trees. Each decision tree is trained on a bootstrapped sample of the data, which introduces randomness and variability into the model. However, when the dataset is imbalanced, with the minority class being significantly underrepresented, the random forest tends to prioritize the majority class, resulting in biased predictions and poor performance on the minority class.

Different sampling methods have been devised to solve this problem, equalize the class distribution, and enhance the functionality of decision tree models. Among the often employed sampling techniques are Random Oversampling, Random Undersampling, SMOTE (Synthetic Minority Oversampling Technique), and ADASYN (Adaptive Synthetic Sampling). A Gridsearch is conducted for each sample strategy to identify the ideal parameters that maximize the Area under the ROC Curve. Table 7 displays the findings from the test set using the aforementioned sampling methods.

Sampling Technique	Cut-off	Accuracy	AUC	Recall	Precision
SMOTE	0.62	0.939	0.932	0.905	0.927
Random Oversampling	0.63	0.927	0.932	0.905	0.927
Random Undersampling	0.45	0.921	0.928	0.952	0.851
ADASYN	0.61	0.956	0.960	0.976	0.911
No Sampling	0.20	0.903	0.919	0.976	0.804

Table 7: Random Forest evaluation metrics on the test set

The summarization of the results are:

1. SMOTE: This technique achieved an accuracy of 0.939 and an AUC of 0.932. It performed well in terms of recall with a score of 0.905, indicating its ability to correctly identify positive cases. The precision score of 0.927 suggests a low rate of false positives. Overall, SMOTE showed a balanced performance in accuracy, AUC, recall, and precision.

2. Random Oversampling: This method underperformed SMOTE with an accuracy of 0.927 and

an AUC of 0.932. It successfully attained a recall score of 0.905, demonstrating its capacity to recognize positive instances. The precision score of 0.927, however, points to a comparable number of false positives as compared to SMOTE.

3. Random Undersampling: This method demonstrated comparable performance to Random Oversampling, achieving an accuracy of 0.921 and an AUC of 0.928. Notably, it achieved a higher recall score of 0.952, indicating its effectiveness in accurately identifying positive cases. Moreover, with a precision score of 0.851, Random Undersampling exhibited a lower rate of false positives when compared to Random Oversampling. These results highlight the balanced performance of Random Undersampling in terms of accuracy, AUC, recall, and precision.

4. ADASYN: This technique attained an accuracy of 0.956 and an AUC of 0.960. It demonstrated a notable recall score of 0.976, signifying its efficiency in accurately detecting positive cases. With a precision score of 0.911, ADASYN exhibited a reduced occurrence of false positives compared to Random Oversampling. However, it should be noted that the precision score of ADASYN is slightly lower than that of Random Oversampling.

5. No Sampling: The random forest without any sampling achieved an accuracy of 0.903 and an AUC of 0.919. It had a high recall score of 0.976 similar to ADASYN, and a precision score of 0.804. This technique provided a balanced performance in terms of accuracy, AUC, recall, and precision.

The random forest with ADASYN is the model with the highest overall performance, especially when considering the AUC score, according to the aforementioned research. The significance of the features, the ROC curves, and the confusion matrices of this particular model will be compared to the biased no-sampling model. Figures 30 and 31 illustrate how the importance scores are distributed to various features in each classifier and accompanying sampler.

Random forests have a large number of decision trees, which means that more features are used to create the decision rules compared to single decision trees. This can be seen in Figures 30 and 31, where all ten features contribute to the creation of decision rules in both ADASYN and no-sampling techniques. The feature *concave.points_ mean* is the most important feature in both ADASYN and no-sampling. In the ADASYN sampler, the least significant feature is *fractal_ dimension_ mean*, while in the no-sampling technique, the least important feature is *symmetry_ mean*.

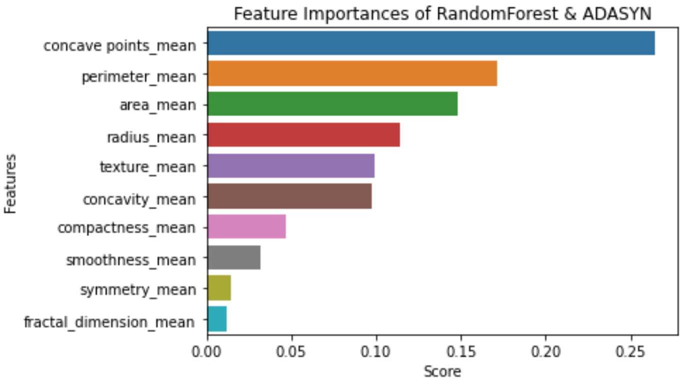


Figure 30: Feature Importance for Random Forest and ADASYN

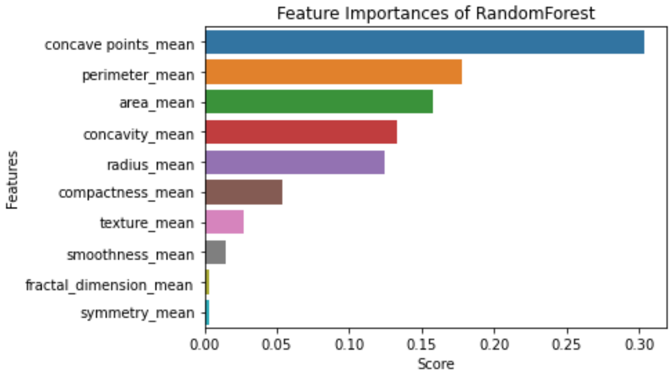


Figure 31: Feature Importance for Random Forest and no sampling

The random forest model using ADASYN shows strong discriminatory power with an impressive AUC of 0.960, indicating its ability to accurately classify examples from both groups, as it is visualized in Figure 32. By carefully selecting a cut-off point of 0.61 based on maximizing the Youden's index, the model strikes a balance between correctly identifying positive cases and minimizing false positives. This means that the ADASYN model can achieve a high rate of correctly identifying malignant tumors while keeping false positives to a minimum.

On the other hand, the model without sampling has a lower AUC of 0.919, as shown in Figure 33. Despite this, it still demonstrates a decent ability to distinguish between benign and malignant tumors. Using the Youden's index, a new threshold for classifying cases is determined with a cut-off point of 0.20. Although the model without sampling may not perform as well as the ADASYN model in terms of AUC, it can still effectively differentiate between benign and malignant tumors. However, it's important to note that the balance between correctly identifying true positives and false positives may differ from the ADASYN model.

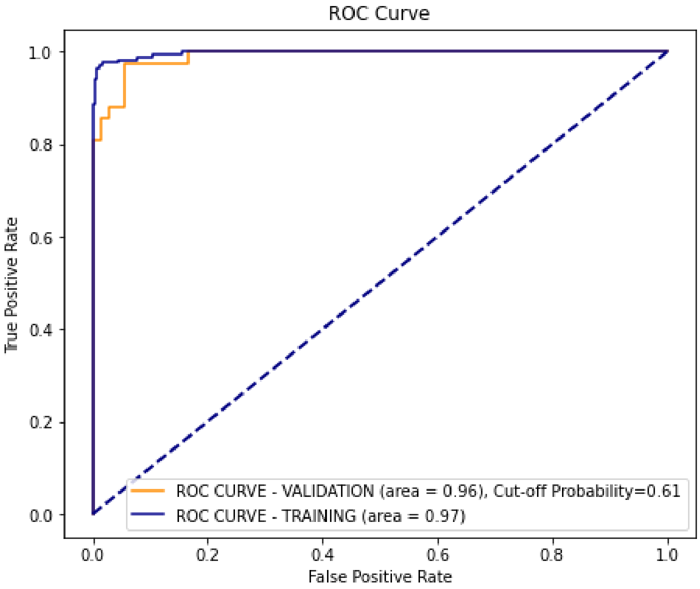


Figure 32: ROC Curve for Random Forest and ADASYN

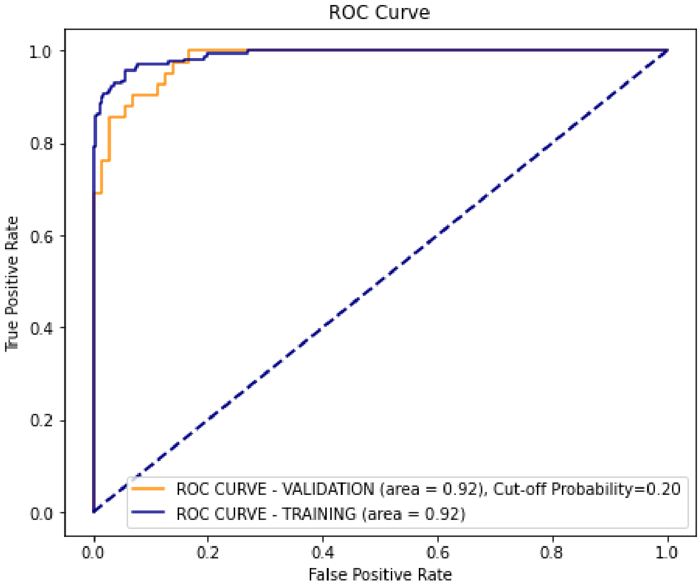


Figure 33: ROC Curve for Random Forest and no sampling

The confusion matrix, as shown in Figure 34, provides an overview of the overall performance of the random forest model with ADASYN. A True Positive rate of 0.98 indicates that the model is highly accurate in predicting positive outcomes, while a True Negative rate of 0.94 demonstrates its proficiency in predicting negative outcomes. However, the False Positive rate of 0.056 reveals that approximately 5.6% of positive predictions are incorrectly classified as malignant tumors when they are actually benign. Likewise, the False Negative rate of 0.024 suggests that around 2.4% of negative predictions are falsely identified as benign tumors when they are actually malignant.

On the other hand, the effectiveness of the random forest model without sampling is illustrated by the confusion matrix displayed in Figure 29. A true positive rate of 0.98 and a true negative rate of 0.86 both indicate the model's strong predictive capability for identifying malignant and benign outcomes. However, the false positive rate of 0.14 reveals that approximately 14% of positive predictions are incorrectly classified as malignant tumors when they are actually benign. Similarly, the false negative rate of 0.024 suggests that around 2.4% of negative predictions are mistakenly labeled as benign tumors when they are, in fact, malignant.

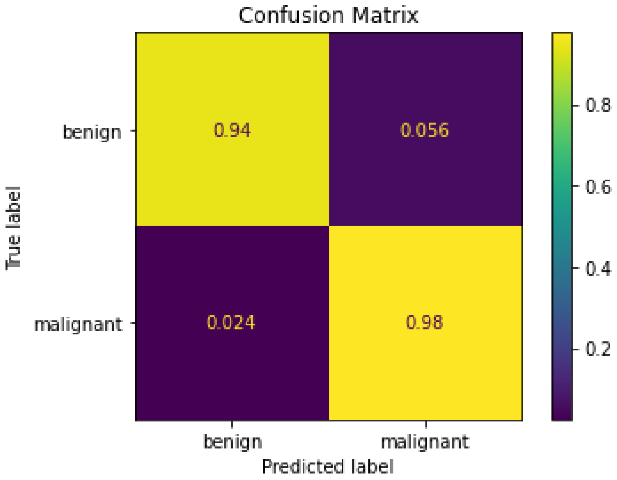


Figure 34: Confusion Matrix on the test set for Random Forest and ADASYN

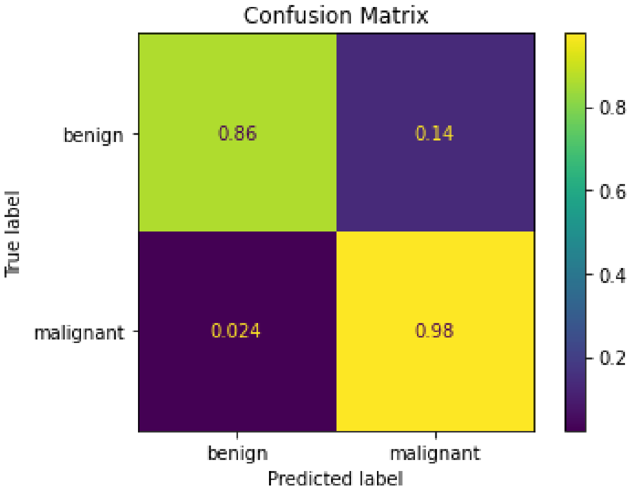


Figure 35: Confusion Matrix on the test set for Random Forest and no sampling

XGBoost

Having a balanced representation of different classes is very important in XGBoost to deal with class imbalance and improve the model's performance. XGBoost is a powerful algorithm that is good at handling complex data. However, when there is an imbalance between classes, with one class having very few examples, XGBoost tends to focus more on the majority class during training. This can result in biased predictions and less accurate results for the minority class.

Sampling techniques play a key role in addressing class imbalance, specifically in XGBoost. For example, oversampling techniques like SMOTE or random oversampling increase the number of examples in the smaller class by creating additional synthetic examples. This helps make the representation of both classes more balanced in the training data. On the other hand, undersampling techniques like Random Undersampling reduce the number of examples in the larger class, ensuring a more fair representation of both classes. By creating a balanced training dataset, sampling techniques help XGBoost learn from a more fair distribution of examples and make accurate predictions for both the larger and smaller classes. This prevents bias towards the larger class and allows XGBoost to better understand the unique patterns and characteristics of the smaller class, resulting in improved performance and more reliable predictions in real-world situations. To determine the optimal parameters that maximize the Area under the ROC Curve, a Gridsearch is performed for each sampling strategy. The results obtained from the test set using these sampling methods are summarized in Table 8.

Sampling Technique	**Cut-off**	**Accuracy**	**AUC**	**Recall**	**Precision**
SMOTE	0.59	0.956	0.955	0.952	0.930
Random Oversampling	0.42	0.947	0.948	0.952	0.909
Random Undersampling	0.86	0.956	0.955	0.952	0.930
ADASYN	0.46	0.960	0.960	0.976	0.911
No Sampling	0.23	0.947	0.953	0.976	0.891

Table 8: XGBoost evaluation metrics on the test set

The results indicate the following conclusions:

1. SMOTE: The accuracy and AUC of the SMOTE approach are 0.956 and 0.955 respectively, showing high overall performance. With a recall score of 0.952, it demonstrates a high level of reliability in identifying positive cases. Furthermore, a reasonably low proportion of false positives is shown by the precision score of 0.930. SMOTE displays balanced performance in terms of recall, accuracy, AUC, and precision.

2. Random Oversampling: This technique achieves an accuracy of 0.947 and an AUC of 0.948. It demonstrates a similar recall score of 0.952 compared to SMOTE, suggesting effective identification of positive cases. However, the precision score of 0.909 indicates a higher rate of false positives compared to SMOTE. Random Oversampling performs well but falls slightly short of SMOTE in terms of accuracy, AUC, and precision.

3. Random Undersampling: Similar to SMOTE, Random Undersampling achieves an accuracy of 0.956 and an AUC of 0.955. It exhibits the same recall score of 0.952, indicating its ability to correctly identify positive cases. The precision score of 0.930 suggests a relatively low rate of false positives, matching SMOTE's performance. Random Undersampling demonstrates balanced performance across all evaluation metrics, comparable to SMOTE.

4. ADASYN: With an accuracy of 0.960 and an AUC of 0.960, ADASYN showcases strong overall performance. It achieves an exceptional recall score of 0.976, indicating its ability to correctly identify positive cases. The precision score of 0.911 suggests a relatively low rate of false positives. ADASYN performs competitively across accuracy, AUC, recall, and precision, showcasing its effectiveness.

5. No Sampling: The XGBoost model without any sampling achieves an accuracy of 0.947, similar to Random Oversampling, and an AUC of 0.953. It achieves a high recall score of 0.976 and a precision score of 0.891. Although it falls slightly short in precision compared to other techniques, it still demonstrates balanced performance in terms of accuracy, AUC, recall, and precision.

The XGBoost model with ADASYN is the one with the highest overall performance, especially when considering the AUC score, according to the aforementioned research. The significance of the features, the ROC curves, and the confusion matrices of this particular model will be compared to the biased no-sampling model. Figures 36 and 37 illustrate how the importance scores are distributed to various features in each classifier and the corresponding sampler.

XGBoost consist of numerous decision trees, allowing for the utilization of more features in the creation of decision rules compared to individual decision trees. This characteristic is evident in Figures 36 and 37, where all ten features play a role in constructing decision rules for both the ADASYN and no-sampling techniques. Notably, the feature *concave.points_ mean* emerges as the most influential in both the ADASYN and no-sampling approaches. In the ADASYN sampler, the *symmetry_ mean* exhibits the least significance, while in the no-sampling technique, the feature *fractal_ dimension_ mean* holds the least importance.

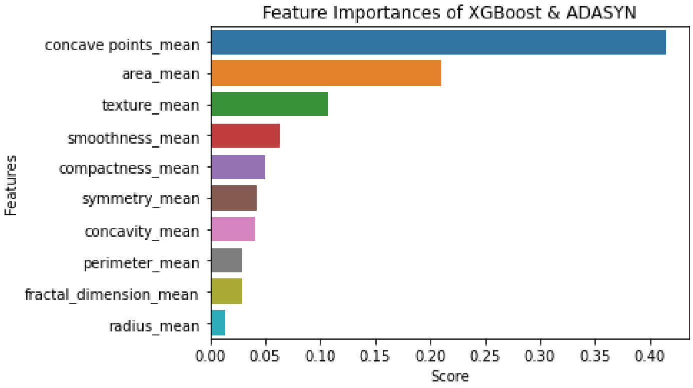


Figure 36: Feature Importance for XGBoost and ADASYN

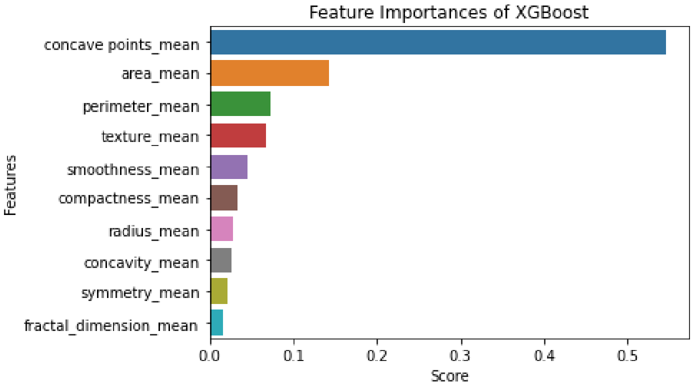


Figure 37: Feature Importance for XGBoost and no sampling

The XGBoost model with ADASYN showcases exceptional discriminatory strength, as indicated by its AUC of 0.960 in Figure 38. This suggests its ability to accurately classify examples from both groups. In order to further enhance the model's performance and strike a balance between the true positive rate and false positive rate, a cut-off point of 0.46 is selected by maximizing the Youden's index. By employing this threshold, the model achieves a high true positive rate while keeping the false positive rate low, resulting in more precise identification of malignant tumors.

On the contrary, the model without sampling exhibits a slightly lower AUC of 0.953, as observed in Figure 39. Nevertheless, it still demonstrates a commendable level of discriminatory capability, although not as strong as the ADASYN model. In line with the maximization of Youden's index, a new threshold is determined using a cut-off point of 0.23 to classify cases. Despite the lower AUC, the model without sampling effectively differentiates between benign and malignant tumors.

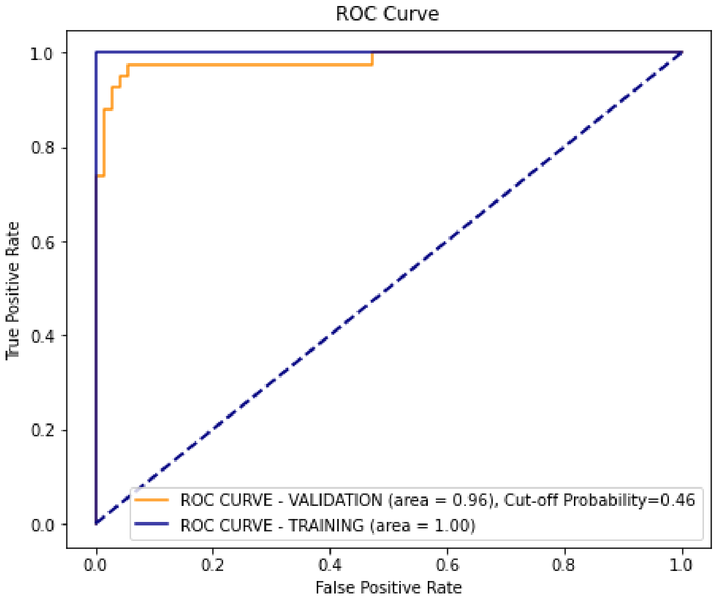


Figure 38: ROC Curve for XGBoost and ADASYN

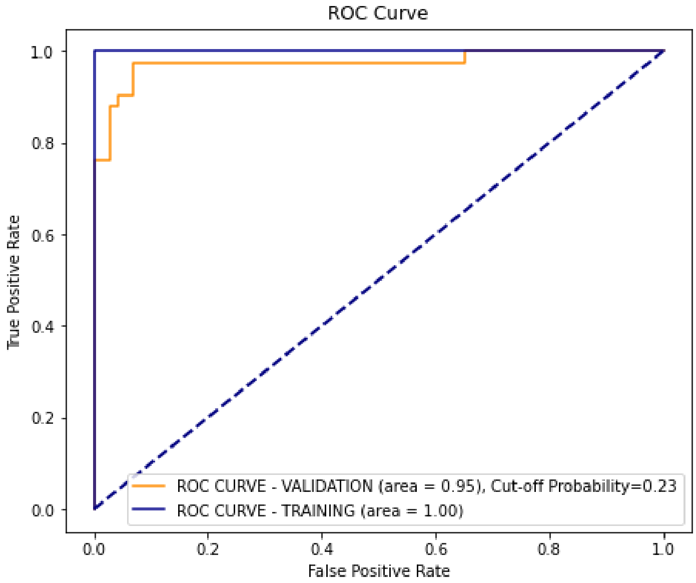


Figure 39: ROC Curve for XGBoost and no sampling

The confusion matrix, which is shown graphically in Figure 40, offers a thorough evaluation of the overall effectiveness of the XGBoost model with ADASYN. The model predicts malignant tumors with a very high True positive rate of 0.98, while it predicts negative outcomes with an impressively high True Negative rate of 0.94. The False Positive rate of 0.056 suggests that 5.6% of positive predictions, however, are incorrectly labeled as malignant tumors when they are benign tumors. Similar to the False Positive rate, the False Negative rate of 0.024 indicates that 2.4% of negative predictions are mistakenly classified as benign tumors when they are, in practice, malignant tumors.

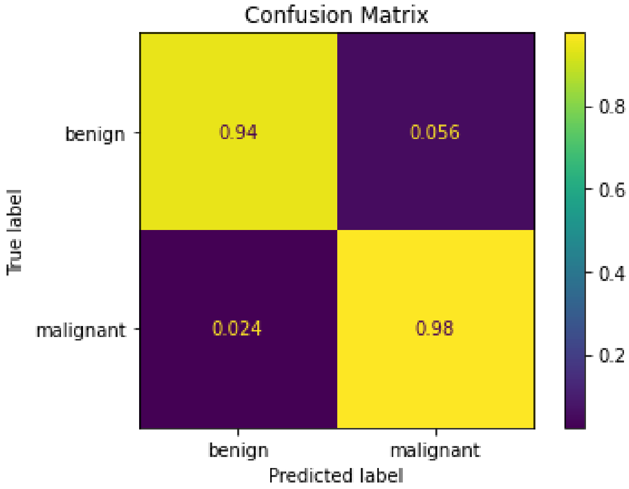


Figure 40: Confusion Matrix on the test set for XGBoost and ADASYN

On the contrary, the effectiveness of the XGBoost model without sampling can be observed through the confusion matrix presented in Figure 41. With a true positive rate of 0.98 and a true negative rate of 0.93, the model exhibits a strong predictive capability in identifying both malignant and benign outcomes. However, the false positive rate of 0.069 reveals that approximately 6.9% of positive predictions are incorrectly classified as malignant tumors when they are actually benign. Likewise, the false negative rate of 0.024 suggests that around 2.4% of negative predictions are mistakenly categorized as benign tumors.

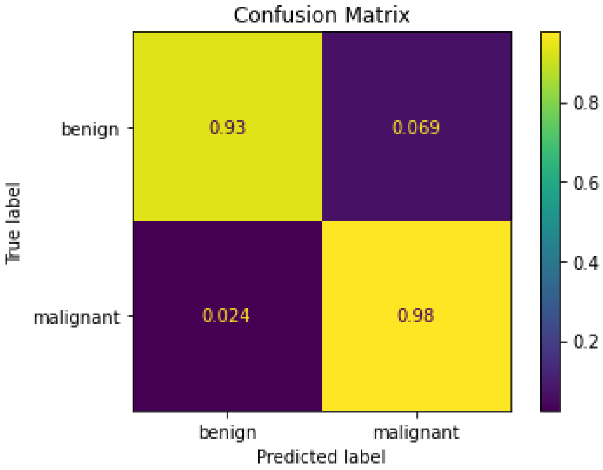


Figure 41: Confusion Matrix on the test set for XGBoost and no sampling

3.4 Discussion

The scope of this research was to establish a methodology for handling imbalanced datasets in order to produce unbiased models. LASSO Logistic Regression and Multilayer Perceptrons were utilized without the technique of sampling, resulting in biased models with respect to the majority class. To address this bias, three different classifiers were applied to the dataset, and various sampling methods were used to balance the minority class. The models that achieved the highest AUC scores utilized the ADASYN sampling method and are therefore considered the ones that could generalize the best. The accuracy of these methods is presented in Table 9, and the corresponding AUC scores are presented in Table 10.

Model	**Accuracy on Test Set**
LASSO	0.921
MLP	0.960
Decision Tree with ADASYN	0.903
Random Forest with ADASYN	0.956
XGBoost with ADASYN	0.960

Table 9: Accuracy on the test set for all models

Model	**AUC on Test Set**
LASSO	0.990
MLP	0.987
Decision Tree with ADASYN	0.909
Random Forest with ADASYN	0.960
XGBoost with ADASYN	0.960

Table 10: AUC on the test set for all models

Based on the provided tables of model performance on the test set, the following conclusions can be arisen:

1. Accuracy Comparison:
   (a) The MLP (Multilayer Perceptron) model achieved the highest accuracy on the test set, with an accuracy of 96.0%.
   (b) The LASSO model performed slightly lower than the MLP, with an accuracy of 92.1%.
   (c) The Decision Tree model with ADASYN sampling achieved an accuracy of 90.3%, which is lower than both the MLP and LASSO models.
   (d) The Random Forest and XGBoost models with ADASYN sampling achieved comparable high accuracies of 95.6% and 96.0%, respectively.
2. AUC Comparison:
   (a) The LASSO model achieved the highest AUC (Area Under the Curve) score on the test set, with a value of 0.990.
   (b) The MLP model closely followed with an AUC of 0.987.
   (c) The Decision Tree model with ADASYN sampling achieved an AUC of 0.909, indicating moderate performance compared to the other models.
   (d) Both the Random Forest and XGBoost models with ADASYN sampling achieved identical AUC scores of 0.960.

Based on these results, it can be concluded that the MLP model demonstrates the highest accuracy among all the models, suggesting its effectiveness in correctly classifying instances in the test set. The LASSO model achieves the highest AUC score, indicating its ability to distinguish between positive and negative instances with high accuracy. Models utilizing ADASYN sampling, such as Random Forest and XGBoost, perform consistently well, achieving comparable accuracies and AUC scores.

As it could be seen from this tables, the models without sampling perform the best in respect to AUC score, but this result could be misleading because of the imbalanced nature of the dataset. When working with unbalanced datasets, biased models might be created since the distribution of classes is frequently defined by the dominance of the majority class over one or more minority classes. Without being addressed, this class imbalance can lead to biased model performance, favoring the majority class while potentially ignoring important patterns within the minority class. The dataset may be balanced and a more representative training set can be given to the model by using the right sampling procedures, such as oversampling the minority class or undersampling the dominant class. This makes it possible to reduce bias and increase the model's capacity to identify and predict trends in both the majority and minority groups.

As a consequence, a trade-off exists between the increased accuracy or AUC score and the production of non-biased models, which must be considered when selecting an appropriate model for a specific application. Particularly, when the minority class pertains to the cancer tumor class, where the patient's survival is at stake, it becomes crucial to choose a model capable of identifying patterns within the minority class to accurately classify a malignant tumor. Thus, selecting a non-biased model is preferred when the minority class is associated with the patient's survival.

The LASSO logistic regression model trained for this research demonstrates lower performance compared to the model trained in the paper by Sidey-Gibbons and Sidey-Gibbons (2019). Specifically, the accuracy achieved in the paper is 95%, whereas the LASSO model trained for this thesis achieves 92% accuracy. Regarding the MLP model, the predicted accuracy matches the 96% accuracy achieved in the paper by Agarap (2018). In the study conducted by Cahyana et al. (2019), the ADASYN sampling technique is utilized alongside the XGBoost classifier, resulting in an accuracy of 97%, which surpasses the corresponding accuracy in this thesis of 96%.

4 Conclusions

According to data from throughout the world, breast cancer is one of the most prevalent diseases in women and accounts for the majority of new cancer cases and cancer-related deaths, making it a serious public health issue in today's society. Because it can encourage prompt clinical care for patients, an early diagnosis of breast cancer can considerably enhance the prognosis and likelihood of survival. A more precise categorization of benign tumors might spare people from receiving unneeded medical care. As a result, there is a lot of study on the proper diagnosis of breast cancer and the classification of individuals into benign or malignant categories. Machine learning is widely acknowledged as the preferred approach in breast cancer pattern classification and forecast modeling due to its distinct advantages in essential features discovery from complicated breast cancer datasets.

In this thesis, five different machine learning models were applied in order to classify instances that correspond to benign or malignant tumor features. The utilized dataset is the Breast Cancer Wisconsin (Diagnostic) Data Set which contains 569 instances of features that describe tumor cells. However, the Class Imbalance Problem characterizes this dataset because the malignant class is underrepresented. The problem with class imbalance in machine learning models arises when there is a significant disparity in the number of observations between different classes. This can lead to biased model performance, as the model tends to favor the majority class, ignoring or misclassifying instances from the minority class.

The imbalanced distribution of classes can result in models with poor predictive accuracy, low sensitivity, and high false positive rates for the minority class. In such cases, the model may incorrectly classify most instances as belonging to the majority class, thus compromising the overall performance and effectiveness of the model. Addressing class imbalance is crucial to ensure that the model can accurately capture patterns and make informed predictions for both the majority and minority classes.

Due to the imbalance in the dataset, sampling techniques were employed in the analysis to develop unbiased models that address the majority class. The performance of three classifiers, specifically Decision Trees, Random Forests, and XGBoost, was evaluated using these sampling techniques and compared to the same classifiers without sampling. Furthermore, Youden's index was utilized to determine the optimal threshold for achieving a balance between sensitivity and specificity. Additionally, to assess the impact of not addressing the class imbalance problem, two models, namely the Multilayer Perceptron and the LASSO logistic regression with feature selection, were applied to the dataset without sampling, and their performance was examined.

In summary, although models without sampling achieved the highest AUC score, this result can be misleading due to the imbalanced dataset. Class imbalance can lead to biased model performance, favoring the majority class and overlooking patterns in the minority class. Balancing the dataset through appropriate sampling techniques reduces bias, especially when the minority class is critical, such as in cancer tumor classification with patient survival at stake. Therefore, it is important to consider a non-biased model when dealing with imbalanced data and crucial outcomes like patient survival.

References

Abdulkareem, S., & Abdulkareem, Z. (2021). An evaluation of the wisconsin breast cancer dataset using ensemble classifiers and RFE feature selection technique. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, *55*, 67–80.

Abed, B. M., Shaker, K., Jalab, H. A., Shaker, H., Mansoor, A. M., Alwan, A. F., & Al-Gburi, I. S. (2016). A hybrid classification algorithm approach for breast cancer diagnosis, 269–274. https://doi.org/10.1109/IEACON.2016.8067390

Agarap, A. F. M. (2018). On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset. *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing - ICMLSC '18*. https://doi.org/10.1145/3184066.3184080

Algarni, A., Aldahri, B. A., & Alghamdi, H. S. (2021). Convolutional neural networks for breast tumor classification using structured features. *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, 1–5. https://doi.org/10.1109/WiDSTaif52235.2021.9430225

Alkabban, F. M., & Ferguson, T. (2022). *Breast cancer*. StatPearls Publishing.

Ara, S., Das, A., & Dey, A. (2021). Malignant and benign breast cancer classification using machine learning algorithms. *2021 International Conference on Artificial Intelligence (ICAI)*, 97–101. https://doi.org/10.1109/ICAI52203.2021.9445249

Basunia, M. R., Pervin, I. A., Al Mahmud, M., Saha, S., & Arifuzzaman, M. (2020). On predicting and analyzing breast cancer using data mining approach. *2020 IEEE Region 10 Symposium (TENSYMP)*, 1257–1260. https://doi.org/10.1109/TENSYMP50017.2020.9230871

Cahyana, N., Khomsah, S., & Aribowo, A. S. (2019). Improving imbalanced dataset classification using oversampling and gradient boosting, 217–222. https://doi.org/10.1109/ICSITech46713.2019.8987499

Cai, T. (2018). Breast cancer diagnosis using imbalanced learning and ensemble method. *Applied and Computational Mathematics*, *7*, 146. https://doi.org/10.11648/j.acm.20180703.20

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*. https://doi.org/10.18637/jss.v033.i01

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets.

Genuer, R., & Poggi, J.-M. (2020). *Random forests with R*. Springer.

Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., Jemal, A., & Siegel, R. L. (2022). Breast cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(6), 524–541. https://doi.org/10.3322/caac.21754

Goodfellow, I., Bengio, Y., & Courville, A. (2018). *Deep learning*. MITP.

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review, 79–85. https://doi.org/10.1109/ICACCI.2017.8125820

Graupe, D. (2013). *Principles of artificial neural networks*. World Scientific.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning, second edition : Data mining, inference, and prediction*. Springer.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. https://doi.org/10.1109/ijcnn.2008.4633969

Karoni, C., & Oikonomou, P. (2017). *Statistical regression models with the use of Minitab and R*. Symeon Publications.

Khuriwal, N., & Mishra, N. (2018a). Breast cancer diagnosis using deep learning algorithm. *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. https://doi.org/10.1109/icacccn.2018.8748777

Khuriwal, N., & Mishra, N. (2018b). Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. *2018 IEEMA Engineer Infinite Conference (eTechNxT)*. https://doi.org/10.1109/etechnxt.2018.8385355

Kleinbaum, D. G., & Klein, M. (2006). *Logistic regression, a self-learning text*. Springer Science Business Media.

Krzysztof, G. (2014). *Meta-learning in decision tree induction*. Springer International Publishing Switzerland.

Kumar, N., Sharma, G., & Bhargava, L. (2020). The machine learning based optimized prediction method for breast cancer detection. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. https://doi.org/10.1109/iceca49313.2020.9297479

Larose, D., & Larose, C. (2015). *Data mining and predictive analytics* (2nd ed.). Wiley.

Mehrotra, K., Mohan, C. K., & Ranka, S. (1999). *Elements of artificial neural networks*. NetLibrary, Inc.

Mittal, D., Gaurav, D., & Sekhar Roy, S. (2015). An effective hybridized classifier for breast cancer diagnosis. *2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. https://doi.org/10.1109/aim.2015.7222674

Mohammad, W. T., Teete, R., Al-Aaraj, H., Rubbai, Y. S., & Arabyat, M. M. (2022). Diagnosis of breast cancer pathology on the wisconsin dataset with the help of data mining classification and clustering techniques. *Applied Bionics and Biomechanics*, *2022*, 1–9. https://doi.org/10.1155/2022/6187275

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. *2020 11th International Conference on Information and Communication Systems (ICICS)*. https://doi.org/10.1109/icics49469.2020.239556

Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. *Data Mining and Big Data*, 108–117. https://doi.org/10.1007/978-981-15-7205-0_10

Rida, A. (2019). Machine and deep learning for credit scoring: A compliant approach. *Master Thesis in École Polytechnique*.

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: Theory and applications*. World Scientific.

Saxena, S. (2022). *Tree-based machine learning methods in SAS® Viya®*. SAS Institute.

Schapire, R. E., & Freund, Y. (2014). *Boosting: Foundations and algorithms*. MIT Press.

Schisterman, E. F., & Perkins, N. (2007). Confidence intervals for the youden index and corresponding optimal cut-point. *Communications in Statistics - Simulation and Computation*, *36*(3), 549–563. https://doi.org/10.1080/03610910701212181

Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*, *19*, 64. https://doi.org/10.1186/s12874-019-0681-4

Singh, S. N., & Thakral, S. (2018). Using data mining tools for breast cancer prediction and analysis. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. https://doi.org/10.1109/ccaa.2018.8777713

Sinha, N. K. (2020). Developing a web based system for breast cancer prediction using XGboost classifier. *International Journal of Engineering Research*, *V9*(06). https://doi.org/10.17577/ijertv9is060612

Smolarz, B., Nowak, A. Z., & Romanowicz, H. (2022). Breast cancer—epidemiology, classification, pathogenesis and treatment (review of literature). *Cancers*, *14*(10), 2569. https://doi.org/10.3390/cancers14102569

Solanki, Y. S., Chakrabarti, P., Jasinski, M., Mitolo, M., Bolshev, V., Vinogradov, A., Jasińska, E., Gono, R., & Nami, M. (2021). A hybrid supervised machine learning classifier system for breast cancer prognosis using feature selection and data imbalance handling approaches. *Electronics*, *10*, 699–699. https://doi.org/10.3390/electronics10060699

Telsang, V. A., & Hegde, K. (2020). Breast cancer prediction analysis using machine learning algorithms. *2020 International Conference on Communication, Computing and Industry 4.0 (C2I4)*. https://doi.org/10.1109/c2i451079.2020.9368911

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, *11*, 24039. https://doi.org/10.1038/s41598-021-03430-5

Zhou, Z.-H. (2012). *Ensemble methods foundations and algorithms*. Taylor & Francis.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΔΠΜΣ
ΕΠΙΣΤΗΜΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

«Ταξινόμηση δεδομένων καρκίνου του μαστού με χρήση στατιστικών μεθόδων και μεθόδων μηχανικής μάθησης»

Διπλωματική Εργασία
της
Χριστίνας Γιαννακουδάκη
Α.Μ. 03400126

Επιβλέπουσα: Δρ. Χρυσηΐς Καρώνη-Ρίτσαρντσον, Καθηγήτρια

Τριμελής Εξεταστική Επιτροπή:

Χ. Καρώνη-Ρίτσαρντσον	Β. Παπανικολάου	Κ. Χρυσαφίνος
Ομ. Καθηγήτρια	Ομ. Καθηγητής	Καθηγητής

Αθήνα, Ιούνιος 2023

Στους γονείς μου, Γιάννη και Ελένη
και στον αδερφό μου, Γιώργη

....

Χριστίνα Γιαννακουδάκη

Το περιεχόμενο αυτής της εργασίας δεν απηχεί απαραίτητα τις απόψεις του Τμήματος, του Επιβλέποντα, ή της επιτροπής που την ενέκρινε.

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. ΄Εχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

Περίληψη

Μια από τις πιο κοινές και θανατηφόρες ασθένειες είναι ο καρκίνος που πλήττει εκατομμύρια άτομα παγκοσμίως. Η ανίχνευση και η θεραπεία του καρκίνου έχουν προχωρήσει σημαντικά χάρη στην ιατρική έρευνα, αλλά το θέμα εξακολουθεί να είναι περίπλοκο και δύσκολο και απαιτεί συνεχείς βελτιώσεις στη μεθοδολογία και την τεχνολογία. Οι αλγόριθμοι μηχανικής μάθησης έχουν πρόσφατα επιδείξει σημαντικές δυνατότητες στον τομέα της ιατρικής έρευνας, ιδίως στην ανάλυση τεράστιου όγκου δεδομένων για τη διάγνωση κακοήθων όγκων. Ωστόσο, η ποιότητα και η ισορροπία των δεδομένων εκπαίδευσης έχουν σημαντικό αντίκτυπο στο πόσο καλά αποδίδουν αυτοί οι αλγόριθμοι. Οι αλγόριθμοι μηχανικής μάθησης μπορεί να αποδίδουν ανεπαρκώς σε μη ισορροπημένα σύνολα δεδομένων όπου ο ένας τύπος όγκου υπερισχύει έναντι του άλλου, οδηγώντας σε μοντέλα που δεν αναγνωρίζουν τη μειοψηφική κατηγορία. Η συγκεκριμένη διπλωματική εργασία εστιάζει στην εφαρμογή μεθόδων δειγματοληψίας, όπως η τυχαία υπερδειγματοληψία, η τυχαία υποδειγματοληψία, και οι τεχνικές SMOTE και ADASYN, για την εξισορρόπηση της κλάσης μειοψηφίας, με στόχο την παραγωγή μοντέλων που αναγνωρίζουν επαρκώς τόσο τους καρκινικούς όσο και τους καλοήθεις όγκους.Η απόδοση τριών ταξινομητών, συγκεκριμένα των Decision Trees, των Random Forests και του XGBoost, αξιολογήθηκε με τη χρήση αυτών των τεχνικών δειγματοληψίας και συγκρίθηκε με τους ίδιους ταξινομητές χωρίς δειγματοληψία. Επιπλέον, για να εκτιμηθεί ο αντίκτυπος της μη αντιμετώπισης του προβλήματος της ανισορροπίας των κλάσεων, δύο μοντέλα, συγκεκριμένα το Multilayer Perceptron και η λογιστική παλινδρόμηση LASSO με επιλογή χαρακτηριστικών, εφαρμόστηκαν στο σύνολο δεδομένων χωρίς δειγματοληψία και εξετάστηκε η απόδοσή τους. Η καλύτερη επιτευχθείσα ακρίβεια τόσο με όσο και χωρίς τεχνικές δειγματοληψίας ξεπέρασε το 96 % στο σύνολο δοκιμών.

Λέξεις-κλειδιά: καρκίνος του μαστού, μηχανική μάθηση, ταξινόμηση, ανισορροπία συνόλου δεδομένων, δειγματοληψία, νευρωνικά δίκτυα.

A Εκτεταμένη Ελληνική Περίληψη

A.1 Κεφάλαιο 1: Εισαγωγή

Ο καρκίνος του μαστού είναι μια διαδεδομένη μορφή καρκίνου που επηρεάζει τις γυναίκες, αποτελώντας περισσότερο από το 10% των νέων κρουσμάτων καρκίνου ετησίως. Αποτελεί τη δεύτερη κύρια αιτία θανάτου από καρκίνο στις γυναίκες παγκοσμίως. Ανατομικά, οι αδένες παραγωγής γάλακτος του μαστού βρίσκονται μπροστά από το τοίχωμα του στήθους, υποστηρίζονται από δεσμίδες και τοποθετούνται στον μυ του μεγάλου κρεμαστήρα. Ο μαστός αποτελείται από 15-20 λόβια που διάταξή τους είναι κυκλική, και το μέγεθος και η μορφή τους καθορίζονται από το λίπος που περιβάλλει τους λοβούς (Alkabban και Ferguson (2022)).

Κάθε λοβός αποτελείται από λόβια, τα οποία περιέχουν αδένες που είναι υπεύθυνοι για την παραγωγή γάλακτος όταν διεγείρονται από ορμόνες. Ο καρκίνος του μαστού αναπτύσσεται συνήθως αθόρυβα, και πολλοί ανακαλύπτουν την ασθένεια μέσω τακτικών ελέγχων. Ωστόσο, μπορεί επίσης να εμφανιστεί ως ένα ακούσιο εύρημα όγκου στο μαστό, αλλαγές στο μέγεθος ή τη μορφή του μαστού, ή διάχυση από το θηλία. Η μασταλγία, δηλαδή ο πόνος στο μαστό, είναι μια συνηθισμένη κατάσταση, αλλά δεν είναι απαραίτητα ενδεικτική του καρκίνου του μαστού (Alkabban και Ferguson (2022)).

Η διάγνωση του καρκίνου του μαστού περιλαμβάνει μια φυσική εξέταση, τεχνικές απεικόνισης όπως η μαστογραφία και μια βιοψία ιστού. Η έγκαιρη ανίχνευση παίζει κρίσιμο ρόλο στη βελτίωση των ποσοστών επιβίωσης. Η επεκτατική εξάπλωση του καρκίνου του μαστού μέσω του λεμφατικού και του αιματολογικού συστήματος μπορεί να οδηγήσει σε δυσμενή πρόγνωση και απομακρυσμένες μεταστάσεις. Επομένως, η σημασία των πρωτοβουλιών για τον έλεγχο του καρκίνου του μαστού υπογραμμίζεται από αυτούς τους παράγοντες (Alkabban και Ferguson (2022)).

A′.1.1 Παθοφυσιολογία

Ο καρκίνος του μαστού προκαλείται από γενετικές μεταλλάξεις και φθορές στο DNA, οι οποίες μπορούν να επηρεαστούν από την έκθεση σε οιστρογόνο. Μερικές φορές, ελαττώματα στο DNA ή γονίδια που προκαλούν καρκίνο, όπως τα BRCA1 και BRCA2, κληρονομούνται. Επομένως, η ύπαρξη καρκίνου των ωοθηκών ή του μαστού στην οικογένεια αυξάνει τον κίνδυνο εμφάνισης καρκίνου του μαστού. Σε έναν υγιή άνθρωπο, τα κύτταρα με ανωμαλίες στο DNA ή ανορθόδοξη ανάπτυξη καταπολεμώνται από το ανοσοποιητικό σύστημα. Όταν οι ασθενείς με καρκίνο του μαστού αντιμετωπίζουν αυτήν την αποτυχία, αναπτύσσονται όγκοι και εξαπλώνονται (Alkabban και Ferguson (2022)).

A′.1.2 Στατιστικά Στοιχεία

Βάσει της έρευνας που διεξήχθη από το Global Cancer Observatory, ο καρκίνος του μαστού αντιπροσώπευε το 27,5% των περιπτώσεων καρκίνου μεταξύ του θηλυκού πληθυσμού της Ελλάδας το 2020. Επιπλέον, ο αντίστοιχος ρυθμός θανάτου για αυτές τις περιπτώσεις ήταν 7,0%. Ο προσαρμοσμένος σε ηλικία ρυθμός προσβολής ανά 100.000 γυναίκες ήταν 71,9%, ενώ ο ρυθμός θνητότητας ήταν 14,5%. Ο προσαρμοσμένος στην ηλικία ρυθμός (ASR) είναι ένα ολοκληρωμένο μέτρο του ρυθμού που θα παρατηρούνταν αν ο πληθυσμός είχε μια τυπική ηλικιακή δομή. Η προσαρμογή γίνεται απαραίτητη κατά τη σύγκριση πολλαπλών πληθυσμών που διαφέρουν ως προς την ηλικία, καθώς η ηλικία επηρεάζει έντονα τον κίνδυνο του καρκίνου. Ο προσαρμοσμένος σε ηλικία ρυθμός υπολογίζεται ως ένας κατανεμημένος μέσος όρος των ηλικιακών ρυθμών, με τα βάρη που καθορίζονται από την ηλικιακή κατανομή ενός τυπικού πληθυσμού. Ο Παγκόσμιος (W) Τυπικός Πληθυσμός είναι ο συχνα χρησιμοποιούμενος για αυτόν τον σκοπό.

Α΄.2 Κεφάλαιο 2: Θεωρητικό Υπόβαθρο

Αμέτρητοι άνθρωποι παγκοσμίως πλήττονται από την κοινή και θανατηφόρα ασθένεια που είναι γνωστή ως καρκίνος του μαστού. Τα ποσοστά επιβίωσης του καρκίνου του μαστού και τα αποτελέσματα των ασθενών μπορούν να βελτιωθούν σημαντικά με την έγκαιρη και ακριβή αναγνώριση. Η χρήση προσεγγίσεων μηχανικής μάθησης και εξόρυξης δεδομένων για την ταξινόμηση του καρκίνου του μαστού έχει προσελκύσει αυξανόμενο ενδιαφέρον τα τελευταία χρόνια. Οι μέθοδοι αυτές έχουν επιδείξει σημαντικές δυνατότητες για να βοηθήσουν τους γιατρούς να κάνουν ακριβείς διαγνώσεις και να λαμβάνουν αποφάσεις για τη θεραπεία. Ο καρκίνος του μαστού κατηγοριοποιείται με τη διαίρεση των δειγμάτων όγκων σε ομάδες όπως κακοήθεις (καρκινικές) και καλοήθεις (μη καρκινικές). Η μαστογραφία και η ιστολογική εξέταση είναι δύο παραδείγματα παραδοσιακών διαγνωστικών τεχνικών που έχουν όρια όσον αφορά την ακρίβεια και την αξιοπιστία. Από την άλλη πλευρά, οι αλγόριθμοι μηχανικής μάθησης έχουν την ικανότητα να εξετάζουν πολύπλοκα μοτίβα και συσχετίσεις μέσα σε τεράστια σύνολα δεδομένων, επιτρέποντας ακριβέστερη και αποτελεσματικότερη ταξινόμηση του καρκίνου του μαστού. Για τον σκοπό αυτό παρουσιάζονται ενδεικτικά πέντε αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν στην συγκεκριμένη εφαρμογή καθώς επίσης και δύο αλγόριθμοι διαχείρησης της ανισορροπίας των δεδομένων, πρόβλημα το οποίο εμφανίζεται συχνά στην ιατρική έρευνα.

Α΄.2.1 Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση είναι μια ισχυρή τεχνική στο πεδίο της μηχανικής μάθησης που χρησιμοποιείται για την πρόβλεψη κατηγορικών μεταβλητών. Αντίθετα από τη γραμμική παλινδρόμηση που ασχολείται με την πρόβλεψη συνεχών μεταβλητών, η λογιστική παλινδρόμηση είναι κατάλληλη για προβλέψεις πιθανοτήτων και ταξινόμησης σε διάφορες κατηγορίες. Κατά την εκπαίδευση ενός μοντέλου λογιστικής παλινδρόμησης, χρησιμοποιούνται δεδομένα εκπαίδευσης που περιλαμβάνουν τόσο τις ανεξάρτητες μεταβλητές όσο και την εξαρτημένη μεταβλητή που επιθυμούμε να προβλέψουμε. Το μοντέλο προσαρμόζεται στα δεδομένα εκπαίδευσης με σκοπό να μάθει τη σχέση μεταξύ των μεταβλητών και να εκτιμήσει την πιθανότητα της εξαρτημένης μεταβλητής να ανήκει σε μία από τις κατηγορίες.

Η λογιστική παλινδρόμηση βασίζεται στη χρήση της λογιστικής συνάρτησης, γνωστής και ως σιγμοειδής συνάρτηση. Αυτή η συνάρτηση μετασχηματίζει το γραμμικό μοντέλο σε μια συνάρτηση πιθανοφάνειας, καθιστώντας δυνατή την εκτίμηση των πιθανοτήτων της κάθε κατηγορίας. Αυτό επιτρέπει στο μοντέλο να προβλέπει την πιθανότητα ενός δείγματος να ανήκει σε μία συγκεκριμένη κατηγορία, βασιζόμενο στις τιμές των ανεξάρτητων μεταβλητών. Η λογιστική παλινδρόμηση χρησιμοποιείται ευρέως σε πολλούς τομείς, όπως η αναγνώριση προτύπων, η κατηγοριοποίηση εικόνων και το φιλτράρισμα ανεπιθύμητων μηνυμάτων στην ηλεκτρονική αλληλογραφία. Επίσης, η λογιστική παλινδρόμηση μπορεί να επεκταθεί για την αντιμετώπιση πολυκατηγορικών προβλημάτων, όπου η εξαρτημένη μεταβλητή μπορεί να ανήκει σε περισσότερες από μία κατηγορίες. Με τη βοήθεια της λογιστικής παλινδρόμησης, είναι δυνατή η κατηγοριοποίηση και η πρόβλεψη σε πολύπλοκα προβλήματα ταξινόμησης.

Α΄.2.2 Δέντρα Απόφασης

Τα Δέντρα Απόφασης είναι ένας από τους πιο δημοφιλείς αλγορίθμους μηχανικής μάθησης που χρησιμοποιούνται για την ανάλυση και την πρόβλεψη δεδομένων. Αυτοί οι αλγόριθμοι βασίζονται σε μια δομή δέντρου, όπου κάθε κόμβος αντιπροσωπεύει μια απόφαση και κάθε φύλλο αντιπροσωπεύει μια τελική κατηγορία ή μια πρόβλεψη. Η δημιουργία ενός Δέντρου Απόφασης γίνεται με την ανάλυση των δεδομένων εκπαίδευσης και την εύρεση των βέλτιστων κανόνων απόφασης για τη διαίρεση των δεδομένων σε διακλαδώσεις. Κατά την ανάπτυξη του δέντρου, γίνεται επίσης χρήση κριτηρίων, όπως η εντροπία και το Gini, για την επιλογή των βέλτιστων διακλαδώσεων και τη δημιουργία ενός ισορροπημένου και αποδοτικού δέντρου.

Τα Δέντρα Απόφασης είναι εύκολα ερμηνεύσιμα και μπορούν να παράγουν ακριβείς προβλέψεις. Επιπλέον, μπορούν να χρησιμοποιηθούν για την αντιμετώπιση προβλημάτων ταξινόμησης και παλινδρόμησης, καθώς μπορούν να προβλέψουν την κατηγορία ενός δείγματος ή την τιμή μιας συνεχούς μεταβλητής. Ένα από τα πλεονεκτήματα των Δέντρων Απόφασης είναι η δυνατότητα αντιμετώπισης αποφάσεων με πολλαπλά κριτήρια και πολλαπλές μεταβλητές. Επιπλέον, μπορούν να αντιμετωπίσουν και αυξημένο αριθμό δεδομένων, καθώς η απόδοσή τους δεν εξαρτάται από το μέγεθος του συνόλου δεδομένων. Ωστόσο, ένα από τα πιθανά μειονεκτήματα των Δέντρων Απόφασης είναι η τάση προς υπερπροσαρμογή στα δεδομένα εκπαίδευσης, προκαλώντας πιθανή χαμηλή απόδοση σε νέα δεδομένα.

Α'.2.3 Τυχαία Δάση

Τα Τυχαία Δάση είναι ένας από τους πιο ισχυρούς και ευέλικτους αλγορίθμους μηχανικής μάθησης που χρησιμοποιούνται για την ανάλυση και την πρόβλεψη δεδομένων. Βασίζονται στην ιδέα της συνδυασμένης πρόβλεψης από πολλά ανεξάρτητα δέντρα απόφασης, γνωστά και ως Δέντρα Απόφασης. Κάθε δέντρο απόφασης δημιουργείται με τυχαία επιλογή δειγμάτων και χαρακτηριστικών από το σύνολο εκπαίδευσης.

Η κύρια ιδέα πίσω από τα Τυχαία Δάση είναι η συνδυαστική δύναμη των πολλαπλών δέντρων απόφασης. Κάθε δέντρο παρέχει μια πρόβλεψη και η τελική πρόβλεψη γίνεται με βάση την πλειοψηφία των προβλέψεων από τα δέντρα. Αυτή η συνδυαστική προσέγγιση βοηθά στη μείωση της υπερπροσαρμογής και της διακύμανσης των προβλέψεων, προσφέροντας πιο σταθερά και αξιόπιστα αποτελέσματα. Τα Τυχαία Δάση μπορούν να χρησιμοποιηθούν τόσο για προβλήματα ταξινόμησης όσο και παλινδρόμησης. Επιπλέον, μπορούν να αντιμετωπίσουν μεγάλα σύνολα δεδομένων και υψηλές διαστάσεις χαρακτηριστικών, καθώς η απόδοσή τους δεν επηρεάζεται αρνητικά από την αύξηση των διαστάσεων.

Α'.2.4 Extreme Gradient Boosting

Ο XGBoost (eXtreme Gradient Boosting) είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που ανήκει στην κατηγορία των ensemble methods. Έχει γίνει δημοφιλής λόγω της αποτελεσματικότητας του στην ανάλυση δεδομένων και την πρόβλεψη. Ο XGBoost χρησιμοποιεί μια τεχνική που ονομάζεται gradient boosting, η οποία είναι ένας συνδυασμός από δέντρα απόφασης και gradient descent. Κάθε δέντρο που προστίθεται στο μοντέλο εστιάζει στην ελαχιστοποίηση του σφάλματος του προηγούμενου δέντρου, δημιουργώντας ένα σύνολο δέντρων που συνεργάζονται για να βελτιώσουν την προβλεπτική ικανότητα του μοντέλου. Ένα από τα κύρια χαρακτηριστικά του XGBoost είναι η δυνατότητα του να χειρίζεται μεγάλα σύνολα δεδομένων και υψηλές διαστάσεις χαρακτηριστικών, ενώ παραμένει γρήγορος και αποδοτικός. Αυτό επιτυγχάνεται μέσω της χρήσης βελτιστοποιημένων αλγορίθμων και τεχνικών όπως η συμπίεση δεδομένων και η παράλληλη επεξεργασία.

Επιπλέον, ο XGBoost προσφέρει και άλλα χαρακτηριστικά που συμβάλλουν στην αποτελεσματική ανάλυση δεδομένων. Μερικά από αυτά περιλαμβάνουν την αυτόματη ρύθμιση των υπερπαραμέτρων του μοντέλου, την ανίχνευση και εξάλειψη των ανεπιθύμητων χαρακτηριστικών και την αντιμετώπιση της ανισορροπίας κλάσεων. Τέλος, ο XGBoost έχει εφαρμογές σε πολλούς τομείς, συμπεριλαμβανομένων της αναγνώρισης προτύπων, της πρόβλεψης και της κατάταξης. Η ικανότητά του να αντιμετωπίζει πολύπλοκα προβλήματα και να παράγει ακριβείς προβλέψεις τον καθιστά ένα από τα πιο αξιόπιστα εργαλεία στον τομέα της μηχανικής μάθησης.

Α′.2.5 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα αποτελούν μια ισχυρή και αποδοτική μέθοδο μηχανικής μάθησης, που βασίζεται στην απομίμηση του ανθρώπινου εγκεφάλου. Η δομή τους είναι κρίσιμη για την απόδοση και την ακρίβεια των αποτελεσμάτων που επιτυγχάνουν. Τα νευρωνικά δίκτυα αποτελούνται από διάφορα επίπεδα νευρώνων. Κάθε επίπεδο αποτελείται από μια ομάδα νευρώνων που λαμβάνουν εισόδους, εκτελούν υπολογισμούς και παράγουν εξόδους. Τα επίπεδα συνδέονται μεταξύ τους με βάση τον τρόπο που μεταδίδονται οι πληροφορίες.

Η αρχιτεκτονική του νευρωνικού δικτύου αναφέρεται στον τρόπο με τον οποίο τα επίπεδα νευρώνων είναι οργανωμένα και συνδεδεμένα. Υπάρχουν διάφορες αρχιτεκτονικές, όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs), τα συνελικτικά νευρωνικά δίκτυα (CNNs) και τα πλήρως συνδεδεμένα νευρωνικά δίκτυα (FNNs). Η συνάρτηση ενεργοποίησης καθορίζει τον τρόπο με τον οποίο ένας νευρώνας ανταποκρίνεται σε μια είσοδο και παράγει μια έξοδο. Κοινές συναρτήσεις ενεργοποίησης είναι η σιγμοειδής (sigmoid), η υπερβολική εφαπτομένη (tanh) και η συνάρτηση ενεργοποίησης ReLU (Rectified Linear Unit). Η εκπαίδευση των νευρωνικών δικτύων απαιτεί τη ρύθμιση των παραμέτρων τους για την επίτευξη επιθυμητών αποτελεσμάτων. Η διαδικασία εκπαίδευσης συνήθως περιλαμβάνει την προσαρμογή των βαρών των νευρώνων με τη χρήση αλγορίθμων βελτιστοποίησης, όπως το backpropagation και οι μέθοδοι κατάβασης της κλίσης (gradient descent).

Η δομή των νευρωνικών δικτύων επηρεάζει την ικανότητά τους να μάθουν από τα δεδομένα και να παράγουν ακριβείς προβλέψεις. Η επιλογή της κατάλληλης δομής είναι σημαντική για την επίτευξη των επιθυμητών αποτελεσμάτων σε μια εφαρμογή μηχανικής μάθησης.

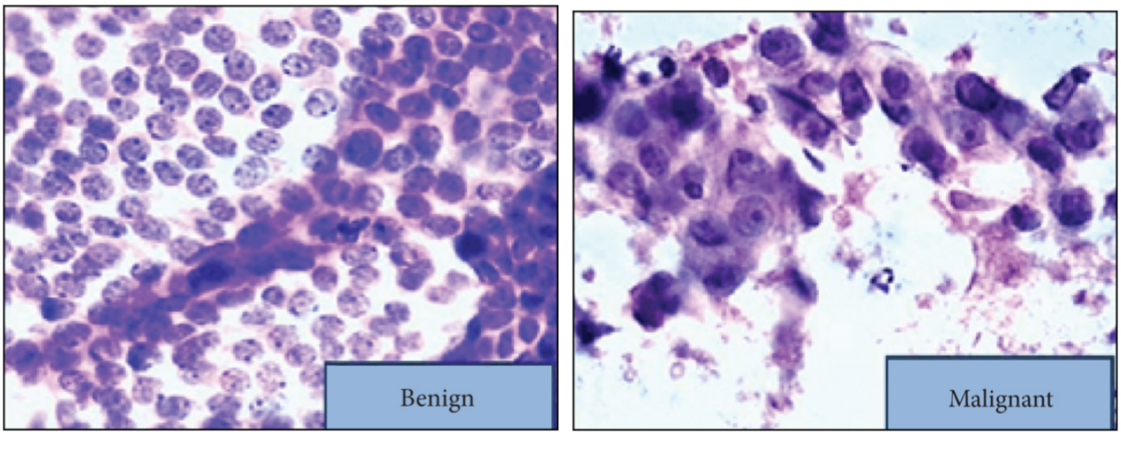
Α′.2.6 Διαχείρηση Ανισορροπίας Δεδομένων

Η τυχαία υπερδειγματοληψία είναι μια τεχνική επαύξησης δεδομένων που χρησιμοποιείται συνήθως για την αντιμετώπιση προβλημάτων ανισορροπίας κλάσεων στη μηχανική μάθηση. Σε αυτή την τεχνική, τα δείγματα από την κλάση της μειονότητας διπλασιάζονται τυχαία για να δημιουργηθεί ίσος αριθμός δειγμάτων τόσο για την κλάση της μειονότητας όσο και για την κλάση της πλειοψηφίας. Η τεχνική αυτή είναι εύκολη στην εφαρμογή και δεν απαιτεί σημαντική υπολογιστική ισχύ. Ωστόσο, η τυχαία υπερδειγματοληψία έχει τους περιορισμούς της, με τον σοβαρότερο να είναι η υπερπροσαρμογή στα δεδομένα εκπαίδευσης, καθώς το μοντέλο βλέπει τα ίδια παραδείγματα πολλές φορές κατά τη διάρκεια της εκπαίδευσης. Επιπλέον, ενδέχεται να μην παρέχει σημαντική βελτίωση στην απόδοση του μοντέλου, ιδίως όταν πρόκειται για δεδομένα με μεγάλη ανισορροπία. Παρ' όλα αυτά, αποτελεί ένα πολλά υποσχόμενο σημείο εκκίνησης για την αντιμετώπιση της ανισορροπίας των κλάσεων και μπορεί να συνδυαστεί με άλλες τεχνικές για τη βελτίωση της απόδοσης του μοντέλου. Επιπρόσθετα, μπορούν να χρησιμοποιηθούν τεχνικές παραγωγής συνθετικών δειγμάτων όπως είναι η SMOTE και η ADASYN.

Η τυχαία υποδειγματοληψία είναι μια δημοφιλής τεχνική μείωσης δεδομένων που χρησιμοποιείται συνήθως για την αντιμετώπιση προβλημάτων ανισορροπίας που συχνά προκύπτουν κατά την εκπαίδευση αλγορίθμων μηχανικής μάθησης. Σε αυτή τη μέθοδο, δείγματα από την πλειοψηφούσα κλάση αφαιρούνται από το σύνολο εκπαίδευσης για να επιτευχθεί μια ισορροπημένη κατανομή των δειγμάτων μεταξύ των δύο κλάσεων. Η τυχαία υποδειγματοληψία είναι εύκολη στην εφαρμογή και υπολογιστικά ανέξοδη, αλλά μπορεί να οδηγήσει σε απώλεια χρήσιμων πληροφοριών, ειδικά αν η πλειοψηφική κλάση περιέχει πληροφοριακά δείγματα. Επομένως, είναι ζωτικής σημασίας να επιλέγεται προσεκτικά η κατάλληλη αναλογία δειγμάτων μειονότητας και πλειοψηφίας για την υποδειγματοληψία, ώστε να διασφαλίζεται ότι η απόδοση του μοντέλου δεν επηρεάζεται αρνητικά.

Α΄.3 Κεφάλαιο 3: Εφαρμογή και Αποτελέσματα

Στην παρούσα μελέτη, θα εξεταστεί το Breast Cancer Wisconsin (Diagnostic) Data Set προκειμένου να χρησιμοποιηθούν και να αξιολογηθούν διάφοροι αλγόριθμοι μηχανικής μάθησης. Με 569 παρατηρήσεις και 30 χαρακτηριστικά, αυτό το σύνολο δεδομένων αποτελεί μια πλούσια και τεράστια πηγή πληροφοριών που θα διερευνηθεί διεξοδικά χρησιμοποιώντας ένα ευρύ φάσμα τεχνικών και οπτικοποιήσεων για την αποκάλυψη κρυφών τάσεων και σχέσεων που μπορεί να μην είναι άμεσα εμφανείς. Όπως παρουσιάζεται στην εικόνα 42, τα χαρακτηριστικά σχετίζονται με την εικόνα των καλοήθων ή κακοήθων καρκινικών κυττάρων.



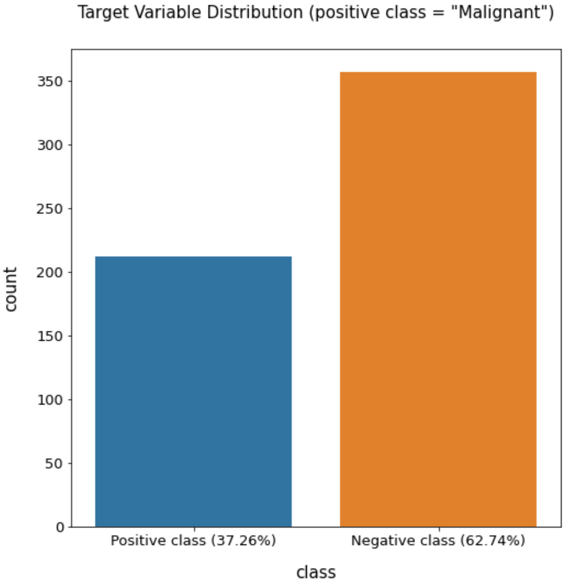
(α΄) (β΄)

Σχήμα 42: Εικόνες καλοηθών και κακοηθών καρκινικών κυττάρων (Mohammad κ.ά. (2022))

Στην επικείμενη ανάλυση, δέκα από τα τριάντα χαρακτηριστικά θα επιλεγούν για εξέταση. Τα χαρακτηριστικά αυτά αντιπροσωπεύονται από τις μέσες τιμές διαφόρων χαρακτηριστικών που παρουσιάζουν οι καρκινικοί όγκοι. Όλες οι μεταβλητές είναι αριθμητικές και δεν υπάρχουν ελλείπουσες τιμές. Το σύνολο δεδομένων χωρίστηκε σε δύο διακριτά υποσύνολα, δηλαδή το σύνολο εκπαίδευσης και το σύνολο δοκιμής, σύμφωνα με μια προκαθορισμένη αναλογία 0.8 προς 0.2, αντίστοιχα. Οι συγκεκριμένες μεταβλητές που αφορούν τους μελετούμενους όγκους είναι οι ακόλουθες:

1. Radius
2. Area
3. Compactness
4. Concave Points
5. Concavity
6. Fractal Dimension
7. Perimeter
8. Smoothness
9. Symmetry
10. Texture

Ώστόσο, στο συγκεκριμένο dataset, οι κλάσεις που περιγράφουν τα δεδομένα δεν έχουν ίση αντιπροσώπευση. Ειδικότερα, η κλάση των καρκινικών όγκων καταλαμβάνει το 37.26% των συνολικών δεδομένων ενώ η κλάση των καλοηθών όγκων αντιπροσωπεύεται από το 62.74%, όπως παρουσιάζεται στο ραβδόγραμμα της εικόνας 43. Αυτό μπορεί να οδηγήσει σε δημιουργία μοντέλων που είναι θετικά προσκείμενα στην κλάση πλειοψηφίας. Για τον σκοπό αυτό θα χρησιμοποιηθούν αλγόριθμοι δειγματοληψίας για να εξισορρροπηθεί η κλάση μειοψηφίας. Ειδικότερα, θα χρησιμοποιηθούν οι αλγόριθμοι Random Oversampling, Random Undersampling, SMOTE και ADASYN.



Σχήμα 43: Κατανομή μεταβλητής στόχου

Το αντικείμενο της παρούσας έρευνας ήταν η δημιουργία μιας μεθοδολογίας για τον χειρισμό συνόλων δεδομένων με ανισορροπία κλάσεων, προκειμένου να παραχθούν αμερόληπτα μοντέλα. Χρησιμοποιήθηκαν η λογιστική παλινδρόμηση LASSO και τα Multilayer Perceptrons χωρίς την τεχνική της δειγματοληψίας, με αποτέλεσμα να προκύψουν μεροληπτικά μοντέλα σε σχέση με την πλειοψηφούσα κλάση. Για να αντιμετωπιστεί αυτή η μεροληψία, εφαρμόστηκαν τρεις διαφορετικοί ταξινομητές στο σύνολο δεδομένων και χρησιμοποιήθηκαν διάφορες μέθοδοι δειγματοληψίας για την εξισορρόπηση της κλάσης μειονότητας. Τα μοντέλα που πέτυχαν τα υψηλότερα AUC χρησιμοποίησαν τη μέθοδο δειγματοληψίας ADASYN και, επομένως, θεωρούνται αυτά που θα μπορούσαν να γενικεύσουν καλύτερα. Η ακρίβεια αυτών των μεθόδων παρουσιάζεται στον Πίνακα 11, και τα αντίστοιχα AUC παρουσιάζονται στον Πίνακα 12.

Model	**Accuracy on Test Set**
LASSO	0.921
MLP	0.960
Decision Tree with ADASYN	0.903
Random Forest with ADASYN	0.956
XGBoost with ADASYN	0.960

Πίνακας 11: Ακρίβεια στο σύνολο δοκιμής για όλα τα μοντέλα

Model	**AUC on Test Set**
LASSO	0.990
MLP	0.987
Decision Tree with ADASYN	0.909
Random Forest with ADASYN	0.960
XGBoost with ADASYN	0.960

Πίνακας 12: AUC στο σύνολο δοκιμής για όλα τα μοντέλα

Με βάση αυτά τα αποτελέσματα, μπορεί να συναχθεί το συμπέρασμα ότι το μοντέλο MLP παρουσιάζει την υψηλότερη ακρίβεια μεταξύ όλων των μοντέλων, γεγονός που υποδηλώνει την αποτελεσματικότητά του στην ορθή ταξινόμηση των περιπτώσεων στο σύνολο δοκιμών. Το μοντέλο LASSO επιτυγχάνει την υψηλότερη βαθμολογία AUC, υποδηλώνοντας την ικανότητά του να διακρίνει μεταξύ θετικών και αρνητικών περιπτώσεων με υψηλή ακρίβεια. Τα μοντέλα που χρησιμοποιούν τη δειγματοληψία ADASYN, όπως το Random Forest και το XGBoost, αποδίδουν σταθερά καλά, επιτυγχάνοντας συγκρίσιμες ακρίβειες και AUC.

Όπως φαίνεται από τους πίνακες αυτούς, τα μοντέλα χωρίς δειγματοληψία έχουν τις καλύτερες επιδόσεις όσον αφορά ως προς το AUC, αλλά το αποτέλεσμα αυτό θα μπορούσε να είναι παραπλανητικό λόγω της ανισορροπίας του συνόλου δεδομένων. Κατά την εργασία με μη ισορροπημένα σύνολα δεδομένων, ενδέχεται να δημιουργηθούν μεροληπτικά μοντέλα, καθώς η κατανομή των κλάσεων ορίζεται συχνά από την κυριαρχία της πλειοψηφικής κλάσης έναντι μιας ή περισσότερων μειοψηφικών κλάσεων. Χωρίς να αντιμετωπιστεί, αυτή η ανισορροπία των κλάσεων μπορεί να οδηγήσει σε μεροληπτική απόδοση του μοντέλου, ευνοώντας την κλάση της πλειοψηφίας, ενώ δυνητικά αγνοεί σημαντικά πρότυπα στην κλάση της μειονότητας. Το σύνολο δεδομένων μπορεί να εξισορροπηθεί και να δοθεί στο μοντέλο ένα πιο αντιπροσωπευτικό σύνολο εκπαίδευσης χρησιμοποιώντας τις σωστές διαδικασίες δειγματοληψίας, όπως η υπερδειγματοληψία της τάξης της μειονότητας ή η υποδειγματοληψία της κυρίαρχη; τάξη. Αυτό καθιστά δυνατή τη μείωση της μεροληψίας και την αύξηση της ικανότητας του μοντέλου να εντοπίζει και να προβλέπει τάσεις τόσο στις πλειοψηφικές όσο και στις μειονοτικές ομάδες.

Κατά συνέπεια, υπάρχει μια αντιστάθμιση μεταξύ της αυξημένης ακρίβειας ή του AUC και της παραγωγής αμερόληπτων μοντέλων, το οποίο πρέπει να λαμβάνεται υπόψη κατά την επιλογή ενός κατάλληλου μοντέλου για μια συγκεκριμένη εφαρμογή. Ιδιαίτερα, όταν η μειονοτική κατηγορία αφορά την κατηγορία των καρκινικών όγκων, όπου διακυβεύεται η επιβίωση του ασθενούς, καθίσταται ζωτικής σημασίας η επιλογή ενός μοντέλου ικανού να εντοπίζει μοτίβα εντός της μειονοτικής κλάσης για την ακριβή ταξινόμηση ενός κακοήθους όγκου. Έτσι, η επιλογή ενός αμερόληπτου μοντέλου προτιμάται όταν η κλάση μειονότητας σχετίζεται με την επιβίωση του ασθενούς.