8-1-2024

Evaluating Feature Selection Methods in Machine Learning With Class Imbalance

Martha Asare
The University of Texas Rio Grande Valley

EVALUATING FEATURE SELECTION METHODS IN MACHINE LEARNING WITH CLASS IMBALANCE

A Thesis

by

MARTHA ASARE

Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: Applied Statistics and Data Science

The University of Texas Rio Grande Valley

August 2024

EVALUATING FEATURE SELECTION METHODS IN MACHINE LEARNING WITH CLASS IMBALANCE

A Thesis
by
MARTHA ASARE

COMMITTEE MEMBERS

Dr. Hansapani Rodrigo
Chair of Committee

Dr. SJ Kumar
Committee Member

Dr. Zhuanzhuan Ma
Committee Member

Dr. Tamer Oraby
Committee Member

August 2024

ABSTRACT

Asare, Martha, Evaluating Feature Selection Methods in Machine Learning with Class Imbalance. Master of Science (MS), August 2024, 100 pp., 46 tables, 8 figures, , 133 references.

Class imbalance is a common issue in various real-world machine learning applications such as medical diagnosis and fraud detection, where one class significantly predominates over the other(s). Conventional methods often lead to biased models that favor the majority class, which can negatively impact the performance of the minority class. To address this issue, techniques like Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and NearMiss have been employed to adjust the class distribution. However, these techniques may not effectively capture the nuances when the feature space is noisy, irrelevant, or high-dimensional. This thesis presents a novel approach that combines class rebalancing techniques with feature elimination strategies and then applies these techniques to each feature by passing them through a Random Forest (RF) and Artificial Neural Network (ANN) for in-depth analysis. The focus of this study is on feature elimination methods such as Chi-Square, Information Gain, Logistic Regression, Recursive Feature Elimination (RFE), LASSO, and Decision Tree-based importance to identify and discard non-informative features, thus streamlining the models and potentially reducing overfitting. The distinctive feature of this study is the use of a dual-modeling approach that combines the strengths of both random forests (RF) and artificial neural networks (ANN) to analyze feature importance rankings and complex pattern recognition abilities in the context of imbalanced datasets. By passing each selected feature through both models, we provide a deeper understanding of feature behavior and model performance. The study utilizes four datasets—Heart Disease, Fraud Detection, Breast Cancer and IT Customer Churn—each presenting its own unique challenges and class imbalance scenarios for a comprehensive evaluation of the proposed methods. Moreover, a thorough benchmarking analysis was conducted comparing the performance of conventional classifiers

on the original imbalanced datasets with those using our integrated approach of class rebalancing and feature elimination. This comparative assessment not only demonstrates the effectiveness of our method in various class imbalance scenarios but also evaluates the impact of each class rebalancing technique when combined with advanced predictive modeling. This study presents an integrated solution that addresses class imbalance through established resampling techniques and enhances predictive modeling using a unique feature elimination and dual-modeling approach. The findings of this study provide valuable insights and practical guidance for practitioners dealing with imbalanced datasets, aiming to improve model accuracy, interpretability, and generalization in real-world applications. *[Keywords:Class Imbalance, Feature Elimination, Artificial Neural Networks (ANN), Synthetic Minority Over-sampling Technique (SMOTE), Random Forest (RF), Predictive Modeling]*

DEDICATION

To my mom, my husband, my family, and support systems, who consistently supported me in my quest for education.

ACKNOWLEDGMENTS

First and foremost, I wish to express my deepest gratitude to God for His guidance and support throughout my master's program and this study. Without the help and support of numerous individuals, the successful completion of this research would not have been possible. I would like to extend my heartfelt thanks to my Supervisor, Professor Hansapani Rodrigo (Ph.D.), for her invaluable guidance and unwavering support during my period of study. I am grateful for her unwavering interest and constant provision of valuable suggestions throughout the course of the study. I also wish to express my appreciation to Professor SJ Kumar (Ph.D.),Professor Zhuanzhuan Ma (Ph.D), and Professor Tamer Oraby (Ph.D) for serving on my thesis committee and offering his expertise and insights.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Class imbalance is a phenomenon in machine learning that occurs when one class within a dataset has a significantly larger number of instances compared to the other classes (Narwane and Sawarkar 2021; Patel, Tailor, and Ganatra 2021). This disparity can result in biased predictive models that favor the majority class, leading to subpar classification performance for the minority class (Arputharaj, Datta, and Hasan 2019; Thölke et al. 2023). Class imbalance is widely recognized as a significant challenge in machine learning; however, its impact on model accuracy is not uniform across different performance metrics. While balancing classes may not significantly affect the accuracy of a model, it can improve other important metrics, such as sensitivity and the area under the receiver operating characteristic curve (AUC) (Patel, Tailor, and Ganatra 2021). Furthermore, the degree of class overlapping can have a greater effect on predictive performance than the global class imbalance ratio (Fernandes and Carvalho 2019).

Class imbalance is a prevalent issue in machine learning that can compromise the effectiveness of classification algorithms. It is crucial to address this imbalance in order to develop models that perform well across all classes, not just the majority (Cheng et al. 2021; Qu et al. 2020). Researchers have proposed various methods to mitigate the effects of class imbalance, including resampling techniques and specialized algorithms, to improve the robustness and fairness of machine learning models (Benkendorf et al. 2023; Dube and Verster 2023; Zheng et al. 2022). Standard machine learning algorithms, which are designed to optimize accuracy and minimize error, often fail under conditions of class imbalance, resulting in overfitting and inaccurate classification estimates (Sevastyanov and Shchetinin 2020). The techniques to address class imbalance in datasets, a prevalent issue in various domains, include a range of oversampling methods. Synthetic Minority Over-

sampling Technique (SMOTE) generates synthetic samples for the minority class by interpolating between existing minority instances (Cheah, Y. Yang, and Lee 2023; Y. Li et al. 2021; Sharma and Gosain 2022; Shoohi and Saud 2020). Additionally, (Davagdorj et al. 2020; Medha et al. 2022) have also proposed various oversampling methods. Adaptive Synthetic Sampling (ADASYN) is an approach that adapts the number of synthetic samples based on the local distribution of the minority class with the aim of achieving a more nuanced balance between classes. This method has been explored in various studies, including those by (Davagdorj et al. 2020; Haddadi et al. 2024; Medha et al. 2022; Sharma and Gosain 2022; Shoohi and Saud 2020; Tahfim and Chen 2024).

Nearness-based methods, which are not explicitly mentioned in the provided papers but generally involve considering the proximity of samples within the feature space to guide oversampling, are also available. Hybrid methods that combine oversampling with other techniques to enhance performance have also been developed. For example, SMOTE Tomek links use SMOTE in conjunction with Tomek links, which are pairs of nearest neighbor samples from different classes, to clean overlapping samples and improve the classification boundary (Sharma and Gosain 2022; Tahfim and Chen 2024). Other hybrid approaches include combining SMOTE with GANs (Generative Adversarial Networks) to generate more realistic synthetic samples (Cheah, Y. Yang, and Lee 2023) and integrating oversampling with clustering and ensemble techniques (Haddadi et al. 2024). Various strategies and techniques have been proposed for addressing class imbalance by creating synthetic samples to bolster the minority class. Synthetic samples generated by SMOTE and its derivatives, such as ADASYN, have proven to be effective in creating representative minority class samples. Hybrid methods, like SMOTE Tomek and those that combine SMOTE with GANs, aim to further enhance the quality of the synthetic samples and subsequent classification performance. These techniques play a vital role in improving model accuracy and reducing bias towards the majority class in imbalanced datasets ((Anusha, Visalakshi, and Srinivas 2023; Cheah, Y. Yang, and Lee 2023; Davagdorj et al. 2020; Haddadi et al. 2024; Medha et al. 2022; Y. Liu et al. 2021; Sharma and Gosain 2022; Shoohi and Saud 2020; Tahfim and Chen 2024).

However, these methods may not be effective in complex scenarios in which the feature space is noisy, irrelevant, or high-dimensional. This study proposes a novel approach that integrates class rebalancing techniques with feature elimination strategies to enhance the model performance on imbalanced datasets. This is achieved by passing each feature through both a Random Forest (RF) and Artificial Neural Network (ANN) for comprehensive analysis. This study delves into feature elimination techniques, such as chi-square, information gain, logistic regression, Recursive Feature Elimination (RFE), LASSO, and Decision Tree-based importance, to identify and remove non-informative features. By streamlining the models, these methods can help prevent overfitting. A unique aspect of this study is the dual-modelling approach, which combines the strengths of random forests (RF) and artificial neural networks (ANNs) to provide deeper insights into the feature behavior and model performance in the context of imbalanced datasets. Effectively addressing class imbalance requires tailored techniques for dataset recalibration. Ashraf (Ashraf et al. 2020) emphasized the importance of undersampling and oversampling techniques in rebalancing datasets. Oversampling, in particular, has been shown to significantly enhance classifier performance (Ashraf et al. 2020). Rekha further discussed this topic by introducing a novel cluster-based oversampling method combined with boosting algorithms, which demonstrated substantial improvements in classifier performance on highly imbalanced datasets (Rekha et al. 2021).

Combining supervised and unsupervised machine-learning approaches is another promising strategy for managing imbalanced datasets. Ugarković and Oreški used decision tree algorithms with cluster analysis to manage data imbalances effectively (Ugarković and Oreški 2022). Narwane and Sawarkar also explore the impacts of class imbalance on machine learning algorithms, emphasizing the need for targeted interventions to enhance system performance(Narwane and Sawarkar 2021). Qu et al. investigated and proposed approaches for mitigating the impact of class imbalance in machine learning applications, particularly in medical imaging. Their research demonstrated the effectiveness of both oversampling and undersampling techniques in addressing class imbalances (Qu et al. 2020). Additionally, (Alahmari 2020) examined class imbalance within the context of autism spectrum disorder (ASD) screening, identifying optimal data resampling techniques to stabi-

lize classification performance. In the domain of machine learning, class imbalance poses a significant challenge that impairs the performance of the standard classification algorithms. This thesis presents a hybrid methodology that integrates oversampling techniques with feature selection to address this issue. The proposed approach is underpinned by the synthetic minority oversampling technique (SMOTE) for augmenting the minority class, a strategy that was validated by (Medha et al. 2022) for its effectiveness in improving accuracy and enhancing model performance. Feature selection was initiated using the Chi-Squared Test as a filter method, aligning with the work of (Mirzaei, Nikpour, and Nezamabadi-pour 2021), who demonstrated the efficacy of clustering and density-based techniques for imbalanced data classification.

Further refinement is achieved through Recursive Feature Elimination (RFE) with a random forest classifier, mirroring the two-stage hybrid strategy described by (Mao et al. 2017), which emphasizes the combination of data- and algorithm-based strategies for online sequential prediction of imbalanced data. The selection is finalized with the application of the Least Absolute Shrinkage and Selection Operator (LASSO), which resonates with the findings of (Mao et al. 2017), who developed a hybrid optimal ensemble classifier framework that overcomes the limitations of traditional imbalance learning methods through multi-objective optimization. The desired outcome was generated using only American English, adhering strictly to its spelling, specific terms, and phrases. These studies underscore the critical need to develop and apply specialized techniques to address the challenges posed by class imbalance in machine learning. By integrating the insights from these research efforts with the novel approach presented in this thesis, a comprehensive foundation was established to explore innovative strategies and evaluate their effectiveness in addressing class imbalance across various machine learning applications. The hybrid approach employed in this thesis is expected to enhance the performance of classifiers by effectively addressing the intricate challenges posed by imbalanced datasets, as evidenced by recent studies in the field. The findings of this study provide valuable insights and practical guidance for practitioners dealing with imbalanced datasets with the aim of enhancing model accuracy, interpretability, and generalization in real-world applications.

1.1 Problem Statement

One of the major challenges in the field of machine learning is class imbalance, which poses a significant problem in critical applications, such as medical diagnostics and fraud detection. This issue arises when the number of instances in one class outnumbers those in other classes, resulting in models that perform well on the majority class but poorly on the minority class. This phenomenon is evident in several key datasets. The 2022 heart disease data from the Behavioral Risk Factor Surveillance System (BRFSS) demonstrates a stark imbalance, with 95% of instances representing non-heart attack cases and only 5% accounting for heart attack cases. Similarly, the credit fraud dataset exhibits an even more severe imbalance, with a disparity of 99% to 1% between the non-fraud and fraud instances. Such disproportionate class distributions can undermine the reliability and accuracy of predictive models, particularly in detecting critical outcomes, such as heart attacks and fraudulent transactions. Additionally, the Breast Cancer dataset, although less extreme, still presents a significant imbalance with a ratio of 61% not healed (malignant) cases to 39% healed (benign) cases. Furthermore the IT Customer Churn also presented an imbalance ratio of 74% cases of Churn and 26% cases of No Churn.This imbalance can result in models that are not sufficiently sensitive to malignant cases, which are of primary concern in medical diagnostics. Although traditional methods for addressing class imbalance, such as oversampling the minority class or undersampling the majority class, have been employed, they have certain limitations. Techniques such as the synthetic minority oversampling technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and NearMiss have been used to adjust class distributions; however, these methods alone may not fully address the complexities associated with noisy, irrelevant, or high-dimensional feature spaces. This study presents a novel hybrid machine learning approach that integrates class rebalancing techniques with advanced feature elimination strategies to improve model performance on imbalanced datasets. By combining oversampling methods and feature selection techniques, the study aims to recalibrate the datasets and enhance the performance of classification algorithms for more accurate and equitable detection of critical outcomes. When dealing with imbalanced datasets, the performance of machine learning models can be further compromised.

1.2 Related Work

Class imbalance poses a significant challenge in the realm of machine learning, particularly in applications, such as medical diagnosis, fraud detection, and network intrusion detection. This issue can significantly impact the performance of models because traditional algorithms tend to exhibit a preference for the majority class, resulting in sub-optimal detection of minority class instances. In this section, we review recent research that has proposed various strategies and methodologies for addressing class imbalance, while also examining their contributions and limitations. Techniques like Synthetic Minority Oversampling Technique(SMOTE) and Adaptive Synthetic Sampling (ADASYN) have been extensively explored for balancing class distributions by generating synthetic instances of the minority class. For instance,(Barkah et al. 2023) and (W. Zhang, Ramezani, and Naeim 2019) demonstrated the effectiveness of SMOTE and ADASYN in various contexts including network intrusion detection and boosting algorithms for imbalanced learning. These studies highlight the potential of synthetic data generation for mitigating class imbalances, albeit with considerations for data quality and model complexity. The significance of feature selection in enhancing the model performance on imbalanced datasets has also been emphasized. Techniques such as Recursive Feature Elimination (RFE) and the use of LASSO for feature importance ranking are critical for identifying and discarding non-informative features, thereby streamlining models and potentially mitigating overfitting. Barkah (Barkah et al. 2023) implemented RFE to select important features, indicating that while feature selection may slightly reduce model accuracy, it notably improves the training speed and model interpretability.

Recent studies have suggested hybrid approaches that integrate class-rebalancing techniques with advanced modeling strategies. For example,(Davagdorj et al. 2020) and (Al-Bahrani et al. 2021) have explored the integration of synthetic oversampling techniques with machine learning classifiers such as Gradient Boosting Trees, Random Forest, and deep learning models for smoking cessation intervention and generating synthetic minority class instances using recurrent neural networks.These studies highlight the benefits of leveraging the strengths of both synthetic oversampling and advanced modeling techniques to address class imbalance. Comprehensive bench-

marking and comparative analyses, such as those conducted by (Anusha, Visalakshi, and Srinivas 2023) and (Kotipalli and Suthaharan 2014), demonstrated the effectiveness of various class balancing and feature selection methods across different datasets and classifiers. These analyses provide valuable insights into the practical applications and impacts of different techniques in real-world scenarios. The development of new oversampling methods to address class imbalances continues to be a significant area of research. Xu (Xu et al. 2022) introduced a synthetic minority oversampling technique based on Gaussian Mixture Model Filtering (GMF-SMOTE), which effectively synthesizes majority and minority samples with dynamic oversampling ratios, showing superior performance in terms of sensitivity and specificity indices compared to traditional oversampling algorithms. The utilization of deep-learning models to address class imbalances has generated considerable interest. In a study published in 2018, (Y. Zhang 2018) proposed a deep generative model for multi-class imbalanced learning that employs a Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) as data generators to create high-dimensional image data. This method demonstrates the effectiveness and robustness of deep generative models for balancing data distributions and improving the classification performance. Researchers are actively exploring the creation of hybrid models combining various techniques to enhance the performance of imbalanced datasets. In a 2018 paper, (Susan and Amitesh Kumar 2020) introduced a novel approach that intelligently oversampled the minority class followed by intelligent undersampling of the majority class, achieving higher classification accuracies on benchmark datasets from the UCI repository. It is crucial to evaluate different class-balancing strategies and their impact on model performance. (Jeatrakul, K. W. Wong, and Fung 2010) combined SMOTE with Complementary Neural Network (CMTNN) to improve classification accuracy in imbalanced data, demonstrating that such combination techniques can significantly enhance performance for the class imbalance problem across various classification algorithms. Exploration of the preprocessing role in managing imbalanced datasets has led to the development of new methodologies aimed at improving data quality before model training. (Hussein et al. 2019) presented an advanced SMOTE algorithm that adjusts newly introduced minority class examples based on their distance to the original minority class samples,

thereby enhancing classifiers' performance by focusing on the minority class data structure. The pursuit of more sophisticated oversampling techniques has led to the development of methods designed to address the specific drawbacks of traditional approaches. For example, (Hao, Yanli Wang, and Bryant 2014) developed an efficient algorithm, GLMBoost, coupled with the synthetic minority oversampling technique (SMOTE) to enhance the predictive accuracy for minority classes in imbalanced datasets from PubChem BioAssay. This method demonstrates the potential of combining algorithmic enhancements with oversampling to more effectively detect rare samples. The integration of deep-learning techniques to correct class imbalances has yielded promising results. (Al-Bahrani et al. 2021) propose SIGRNN, a novel approach using sequence-to-sequence recurrent neural networks to synthesize minority class instances. This method underscores the versatility of deep learning models in learning data distributions and generating synthetic instances to augment imbalanced datasets effectively.

The integration of various machine-learning methods into hybrid models presents a versatile solution for addressing class imbalance issues. In a recent study, (Priyadharshini et al. 2023) introduced ASDMLC, a technique that combines Adaptive Synthetic (ADASYN) sampling with multi-label classification algorithms, resulting in a significant improvement in classification accuracy for medical applications. This research demonstrates the benefits of adaptive synthetic sampling in multilabel scenarios and emphasizes the effectiveness of hybrid approaches in handling complex imbalanced datasets. It is essential to evaluate and compare class-balancing strategies through comprehensive assessment to determine their effectiveness. (Bansal et al. 2021) analyzed the performance of a modified SMOTE algorithm (SMOTE-M) across various imbalanced datasets and compared it with traditional oversampling techniques. Their findings indicate the adaptability and efficiency of modified oversampling methods in improving classification performance, particularly in the context of the Internet of Things (IoT) environment.

Addressing the effects of noise and data quality on imbalanced datasets is critical in developing robust models.(C. Zhang et al. 2016) proposed an algorithm that combines SMOTE with stacked denoising auto-encoder (SDAE) neural networks to enhance the classification accuracy of minority classes in imbalanced datasets. This approach not only balances the dataset but also improves data quality by reducing noise, highlighting the importance of preprocessing in handling class imbalance.

1.3 Research Gap

Despite the considerable progress made in addressing class imbalance through the use of synthetic data generation, feature elimination, and advanced modeling techniques, a significant gap remains in the comprehensive evaluation and integration of these methodologies. In particular, the efficacy of class-rebalancing techniques combined with feature elimination strategies in high-dimensional, noisy, or irrelevant feature spaces has not been fully explored. Furthermore, although numerous individual methods have been assessed, the potential for synergistic improvement in model performance, interpretability, and generalization by integrating these methods in imbalanced datasets remains largely untapped. In addition, a systematic comparative analysis of the performance of these integrated approaches versus traditional classifiers in imbalanced datasets is lacking, leaving a void in the literature regarding the practical implications and effectiveness of these combined strategies in real-world applications.

With the need of class-rebalancing, the study will highlight these objectives;

- To evaluate the effectiveness of various feature selection methods, including Chi-Square, Information Gain, Logistic Regression, Recursive Feature Elimination (RFE), LASSO, and Decision Tree-based importance, in identifying and discarding non-informative features.
- To assess the impact of combining feature selection methods with oversampling techniques such as SMOTE, ADASYN, and NearMiss on the performance of Random Forest (RF) and Artificial Neural Network (ANN) models.

- To conduct a comprehensive benchmarking analysis by comparing the performance of traditional classifiers on the original imbalanced datasets with those subjected to the integrated approach of class rebalancing and feature elimination.

The organization of this research paper is as follows: Chapter 2 outlines the materials and methods employed throughout the investigation, providing a foundation for the study's approach. Chapter 3 details the findings and presents and analyzes the results obtained from the models. The concluding chapter (Chapter 4) synthesizes the study's insights, discusses its implications, and acknowledges its limitations.

CHAPTER II

METHODOLOGY

This chapter delves into the subject of machine learning and provides a comprehensive discussion of the class rebalancing and feature selection techniques implemented to balance the datasets for the Heart Disease, Fraud Detection, and Breast Cancer problems. The application of these methods was instrumental in achieving the objectives outlined in Chapter One.

2.1 Data Description

This section provides a detailed description of the three datasets used in this study: Heart Disease, Fraud Detection, and Breast Cancer.

2.1.1 Heart Disease Dataset

The first dataset for the study is selected from the Behavioral Risk Factor Surveillance System (BRFSS) and contains details of 245,979 records spanning 37 variables after cleaning the data and removing duplicates (Centers for Disease Control and Prevention 2022). The attributes/features of the data include; Sex, GeneralHealth, LastCheckupTime, PhysicalActivities, RemovedTeeth, HadHeartAttack, HadAngina, HadStroke, HadAsthma, HadSkinCancer, HadCOPD, HadDepressiveDisorder, HadKidneyDisease, HadArthritis, HadDiabetes, DeafOrHardOfHearing, BlindOrVisionDifficulty, DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands, SmokerStatus, ECigaretteUsage, ChestScan, RaceEthnicityCategory, AlcoholDrinkers, HIVTesting, FluVaxLast12, PneumoVaxEver, TetanusLast10Tdap, HighRiskLastYear, CovidPos, and Age.

In the analysis, whether an individual had a heart attack or not is the endpoint of interest. From the dataset, out of the 245,979 individuals, 13,435 (5%) had a heart attack, and 232,544 (95%) did not have a heart attack. Table 2.1 and 2.1 below provides a summary of the percentages for each class of the heart disease dataset.

Table 2.1: Percentages of Heart Disease Dataset Target Variable

Variables	Description	Percentages
HadHeartAttack	**Yes (1)** **No (0)**	**5%** **95%**

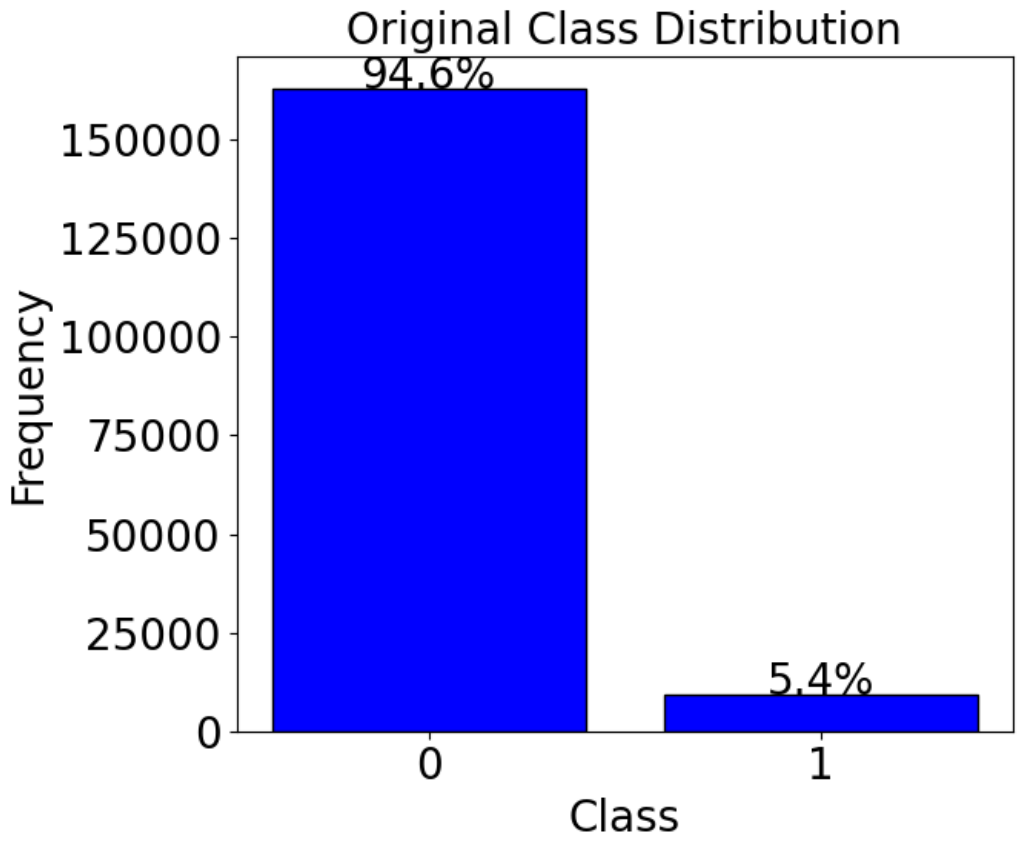

Figure 2.1: Bar Chart of target variable (HadHeartAttack) : 0 = No Heart Attack , 1 = Heart Attack

2.1.2 Fraud Detection Dataset

The widespread usage of credit and debit cards has significantly transformed the way payment processing is done. Nowadays, cash transactions only account for 20% of all in-person purchases, while plastic and digital wallets have become the preferred methods for everyday transactions, even beyond e-commerce. With roughly 2.8 billion credit cards in circulation globally, the potential for fraud has never been higher (Consulting 2024). It is crucial for merchants to understand the scope of credit card fraud, both in the US and worldwide, in order to safeguard their businesses and ensure secure transactions. Noteworthy statistics indicate that 46% of international credit card fraud occurs in the US, and the predicted global fraud is projected to reach $43 billion by 2026 2.2. Credit card fraud losses in the US are predicted to exceed $12.5 billion by 2025. Furthermore, 48% of consumers believe it is the responsibility of the merchant to protect them from fraud, and 55% of fraudulent credit and debit card transactions are worth less than $100. Every 14 seconds, a person in the US falls victim to identity theft, with an estimated 150 million Americans expected to be affected by credit card fraud this year(Consulting 2024).

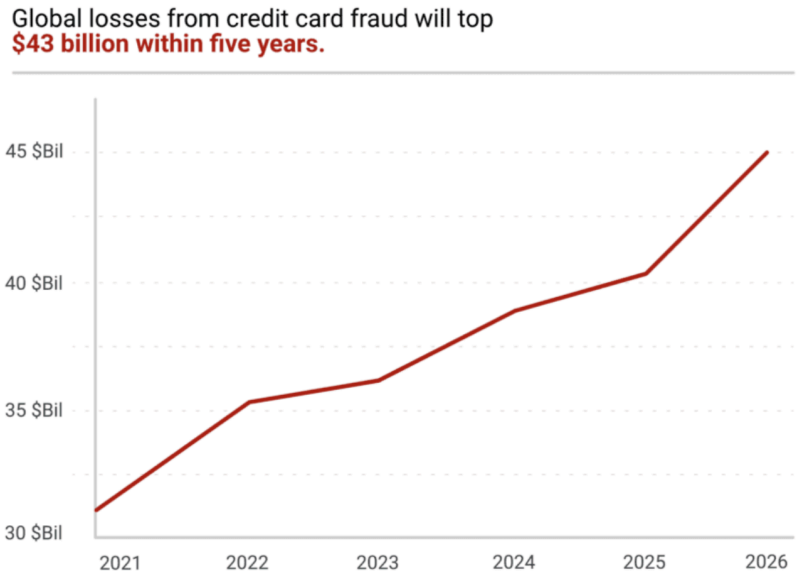

Figure 2.2: Credit Card Fraud Statistics (2024), (Consulting 2024)

The second dataset for the study is selected from the Credit Card Fraud Detection database. This dataset presents transactions that occurred over two days, where we have 473 frauds out of 283,726 transactions. The dataset is highly unbalanced, with the positive class (frauds) accounting for 0.17% of all transactions 2.2. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. The Features are V1, V2, … V28 , 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction amount. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise (Bruxelles 2016).

Table 2.2: Percentages of Fraud Dataset Target Variable

Variables	Description	Percentages
Fraud	**Yes (1)** **No (0)**	**0.17%** **99.83%**

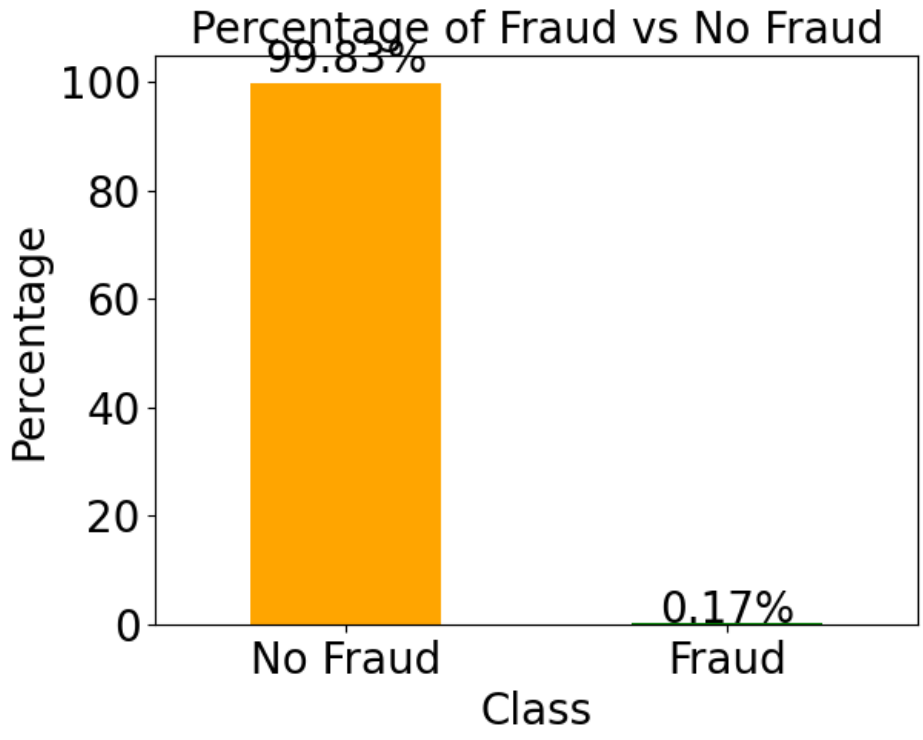

Figure 2.3: Bar Chart of target variable (Fraud) : 0 = No Fraud , 1 = Fraud

2.1.3 Breast Cancer Dataset

This dataset of breast cancer patients was obtained from the November 2017 update of the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI), which provides information on population-based cancer statistics (National Cancer Institute 2017). The attributes of the dataset include Age, which refers to the patient's age at diagnosis; Sequence Number, which indicates the order of all primary tumors; Year of Diagnosis, which refers to the year when the cancer was diagnosed; Primary Site, which indicates the specific location of the primary cancer; Laterality, which denotes the side of the body where the tumor is located; Reason No Cancer-Directed Surgery, which provides the reasons for not performing surgery; Histology Recode, which provides broad groupings of histology; ER Status Recode Breast Cancer (1990+), which indicates Estrogen Receptor status; PR Status Recode Breast Cancer (1990+), which indicates Progesterone Receptor status; Survival Months, which refers to the number of months a patient survived after diagnosis; and Breast - Adjusted AJCC 6th T (1988-2015), which classifies tumor size based on the AJCC 6th edition (National Cancer Institute 2017). Breast - Adjusted AJCC 6th N (1988-2015) classifies lymph node involvement based on the AJCC 6th edition. Breast - Adjusted AJCC 6th M (1988-2015) classifies metastasis based on the AJCC 6th edition. Breast - Adjusted AJCC 6th Stage (1988-2015) provides overall stage classification based on the AJCC 6th edition. Age at Diagnosis refers to the age of the patient at the time of diagnosis. Laterality Recoded provides recoded laterality information. Healing Status indicates whether the patient was healed or not. Surgery Performed indicates whether surgery was performed or not (National Cancer Institute 2017). In the analysis, the primary endpoint of interest is whether an individual was healed or not. From the dataset, out of the 156,124 females, 60,811 (38.95%) were healed, and 95,313 (61.05%) were not healed.

Table 2.3: Percentages of Breast Cancer Survival Dataset Target Variable

Variables	Description	Percentages
Healed	**Yes**	**38.95%**
	No	**61.05%**

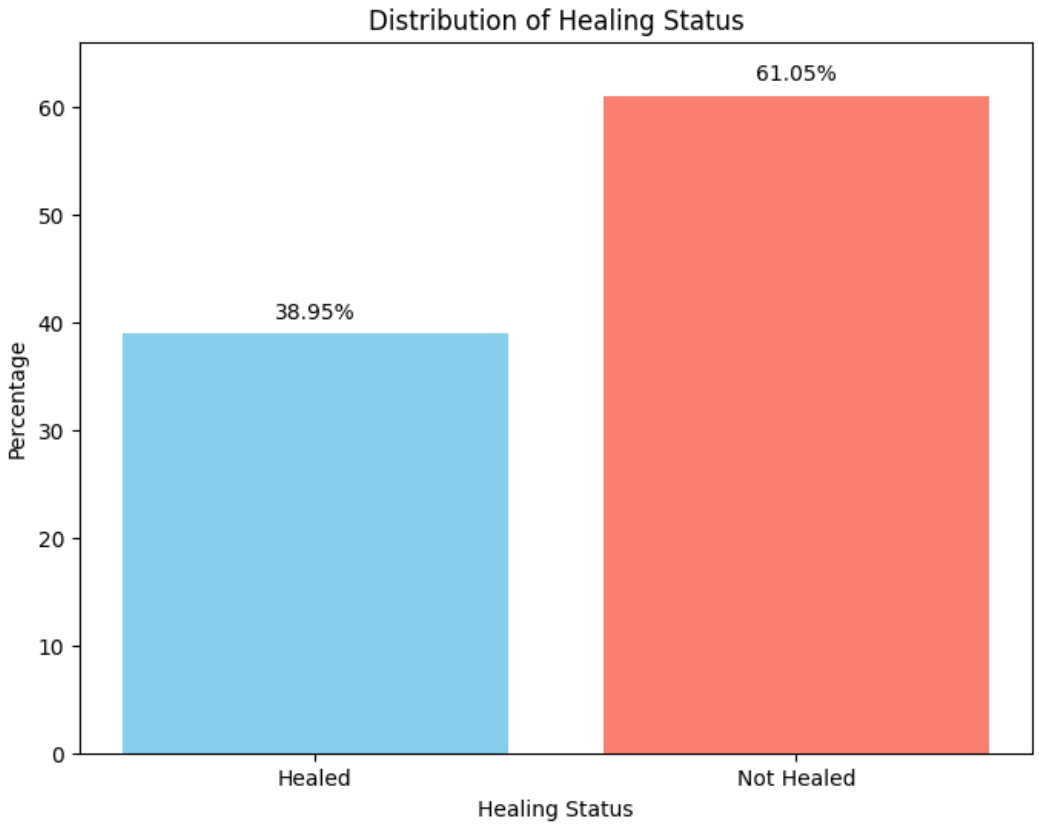

Figure 2.4: Bar Chart of target variable (Fraud) : 0 = Not Healed , 1 = Healed

2.1.4 IT Customer Churn

The IT Customer Churn dataset, is intended to facilitate analysis of customer churn within the information technology industry. This dataset incorporates a multitude of features that capture customer demographics, account information, and interaction details, making it a valuable resource for constructing predictive models that aim to identify customers who are likely to churn and comprehend the elements contributing to customer attrition (Tehranipour 2021). Some of the essential elements of the dataset comprise CustomerID, a unique identifier for every customer; Gender, indicating the gender of the customer; SeniorCitizen, a binary feature that indicates if the customer is a senior citizen; and Tenure, which records the number of months the customer has been with the company. The dataset also includes details on the type of internet service utilized by the customer, such as DSL or Fiber optic, and the MonthlyCharges and TotalCharges incurred by the customer. The Churn feature is particularly significant since it indicates whether the customer has churned or

not. This dataset is particularly beneficial for data scientists and analysts who focus on customer retention strategies. It offers a wealth of information that can be utilized for employing various machine learning techniques to predict churn behavior, enabling businesses to proactively address customer attrition.

Table 2.4: Percentages of Breast Cancer Survival Dataset Target Variable

Variables	Description	Percentages
Churn	**Yes**	**74%**
	No	**26%**

2.2 Data Pre-processing

Pre-processing is an essential step in preparing datasets for analysis and modeling. This entails several tasks, including handling missing values, normalizing features, and splitting the data into training and testing sets. The pre-processing steps for the Heart Disease, Credit Card Fraud Detection, and Breast Cancer datasets are detailed below.

2.2.1 Heart Disease Dataset

During the analysis of data, it is customary to investigate the unique values within a particular column of the dataset. In the case of the 37 variables or features, the data comprised 4,366 distinct values. The dataset included both categorical and numerical (continuous) variables. The numerical variables, such as height and weight, were excluded after the cleaning process and replaced by the Body Mass Index (BMI).The dataset's integrity was ensured by removing missing values during the data cleaning process. In the realm of machine learning, encoding signifies the transformation of categorical data into a format that is comprehensible to machine learning algorithms. As numerous algorithms require numerical input, encoding becomes a crucial step in preparing the dataset for model training. In this context, the Label Encoding method was employed to assign a unique numerical identifier to each column(as shown in 2.5) (Indonesia 2020).

Figure 2.5: Encoding Methods

To evaluate the effectiveness of our machine learning model, it is essential to divide the dataset into training and testing subsets. The training set is employed for the initial model fitting, where the model acquires knowledge from known data characteristics. Conversely, the testing set, which remains distinct from the training process, is utilized exclusively to assess the model's predictive capability on unseen data. This approach ensures that we gauge the model's performance and its ability to generalize beyond the data it was trained on. The original data was separated into two separate groups, with 30% (73,794) of the data allocated to the Test set and 70% (1,721,185) retained for the Training set. Additionally, the feature scaling method was implemented to standardize the range of the independent variables or features of the data. This technique adjusts the scale of the data, preventing any single feature from having a disproportionate influence on the final

results due to its larger or smaller numerical range. Standardizing the data is critical for maintaining fairness and balance among the various variables, ensuring that each one makes an appropriate contribution to the outcome. The formula for standardization is given by:

$$Z = \frac{X - \mu}{\sigma} \tag{2.1}$$

where X is the raw data point, μ is the mean of the feature , σ is the standard deviation of the feature and Z is the standardized value of X.

2.2.2 Fraud Detection Dataset

The initial step in the pre-processing of the fraud dataset was to eliminate duplicates. The majority of the data had already been scaled, with the exception of two columns that required scaling: 'Amount' and 'Time'. To address this, RobustScaler was utilized as it is particularly resistant to outliers.

$$X_{\text{scaled}} = \frac{X - Q_1}{Q_3 - Q_1} \tag{2.2}$$

where Q_1 is the first quartile (25th percentile) and Q_3 is the third quartile (75th percentile).

The categorical data was transformed into a machine-learning-compatible format through the application of one-hot encoding, resulting in the creation of binary columns for each category (see fig. 2.5. The data was subsequently partitioned into a testing set comprising 30% (85,118) of the total and a training set comprising 70% (198,608) of the total.

2.2.3 Breast Cancer Dataset

The pre-processing of the Breast Cancer dataset involved several steps to ensure that the data was clean, consistent, and ready for analysis. The first step was to remove all cases in which the Estrogen Receptor (ER) and Progesterone Receptor (PR) status was unknown (Not 1990+ Breast). The column "Laterality" was then recoded into three categories: left, right, and two-sided. Cases diagnosed later than 2011 were also filtered out. Patients were classified as "healed" or "not healed,"

and missing values were checked. New values in the "Primary Site – Labeled" column were replaced, and missing values in respective columns were replaced with the mode of each column. Hybrid encoding was applied, combining label encoding for the target variable (healed or not healed) and one-hot encoding for other categorical variables. This approach ensures that categorical data is in a format suitable for machine learning algorithms.

Table 2.5: Healed Status Encoding

Healed Status	**Encoded Value**
Healed	1
Not Healed	0

Table 2.6: Laterality Encoding

Laterality	**Left**	**Right**	**Two-Sided**
Left	1	0	0
Right	0	1	0
Two-Sided	0	0	1

The features were normalized using MinMaxScaler, which scales the data to a range of [0, 1]:

$$X_{\text{scaled}} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2.3}$$

where X is the raw data point, X_min is the minimum value of the feature and X_max is the maximum value of the feature.

The Breast Cancer dataset was then thoroughly pre-processed by splitting it into a 30% testing dataset comprising 46,837 instances and a 70% training set consisting of 109,287 instances. These steps were taken to ensure that the dataset was adequately prepared for the subsequent analysis and modeling stages.

2.2.4 IT Customer Churn

The initial step in the pre-processing of the IT churn dataset was to eliminate duplicates. To address scaling issues, RobustScaler was utilized as it is particularly resistant to outliers.

$$X_{\text{scaled}} = \frac{X - Q_1}{Q_3 - Q_1} \tag{2.4}$$

where Q_1 is the first quartile (25th percentile) and Q_3 is the third quartile (75th percentile).

2.3 Statistical Analyses

This section provides an overview of the statistical methods used to address class imbalance and perform feature selection on the datasets. The steps involved in the statistical analyses were as follows.

2.3.1 Handling Class Imbalance

Imbalanced classes arise when the number of instances in one class significantly exceeds the number of instances in another class. This disproportion can lead to biased model predictions, where the model demonstrates heightened proficiency in predicting the more prevalent class, while simultaneously displaying subpar performance in the less prevalent class. Various techniques are widely employed to tackle the class imbalance problem (CIP) in diverse datasets. These methods include the Synthetic Minority Over-sampling Technique (SMOTE) and its derivatives, such as Adaptive Synthetic Sampling (ADASYN) and SMOTE-Tomek, as well as under-sampling strategies like NearMiss. SMOTE generates synthetic samples by interpolating between instances of the minority class, while ADASYN focuses on generating samples along the boundary of the minority class. SMOTE-Tomek combines over-sampling with cleaning algorithms to remove Tomek links, which are pairs of closely positioned instances belonging to different classes. NearMiss, on the other hand, is an under-sampling technique that selects instances based on their proximity to the majority class (Elreedy and Atiya 2019; Sharma and Gosain 2022; Shoohi and Saud 2020; Soltanzadeh and Hashemzadeh 2021; Tahfim and Chen 2024). Several studies have demonstrated that var-

ious methods can have a considerable impact on the performance of classification models (Sharma and Gosain 2022; Shoohi and Saud 2020). Research has shown that SMOTE-Tomek links can outperform other techniques in specific datasets(Sharma and Gosain 2022), while ADASYN has been effective in creating a balance between minority and majority classes(Shoohi and Saud 2020). However, the efficacy of these methods can vary depending on the dataset and the classification algorithm employed. In the case of Female Daily's imbalanced data, the combination of SMOTE-Tomek with SVM demonstrated positive improvements for the minority class(Jonathan, P. H. Putra, and Ruldeviyani 2020). Additionally, the novel Range-Controlled SMOTE (RCSMOTE) method addresses some of the limitations of SMOTE by controlling the synthetic sample generation process(Soltanzadeh and Hashemzadeh 2021). In the field of crash data analysis, a combination of cluster-based undersampling with SMOTE Tomek and ADASYN has been suggested to improve model performance(Tahfim and Chen 2024). While SMOTE and its variants, along with NearMiss, have demonstrated effectiveness in addressing class imbalance, their performance is influenced by various factors, including the nature of the dataset and the classifiers used. Empirical evidence suggests that SMOTE Tomek links can be particularly effective, but novel approaches such as RC-SMOTE and cluster-based under-sampling with SMOTE Tomek also show promise in improving classification outcomes(Sharma and Gosain 2022; Soltanzadeh and Hashemzadeh 2021; Tahfim and Chen 2024). In our study, we utilized SMOTE, ADASYN, NearMiss, and SMOTE Tomek techniques to balance the class distribution and address this issue in our datasets.

2.3.1.1 Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a synthetic minority over-sampling technique that generates additional synthetic samples for the minority class by interpolating between existing minority instances. This method is effective in balancing class distribution without duplicating minority class samples.

$$\text{Synthetic Sample} = X_i + \lambda (X_j - X_i) \tag{2.5}$$

where X_i and X_j are two minority class samples and λ is a random number between 0 and 1.

Synthetic samples generation is a widely used approach to address the issue of class imbalance in datasets, which is a common problem in machine learning (Mohd et al. 2019; Soltanzadeh and Hashemzadeh 2021). The Synthetic Minority Over-sampling Technique (SMOTE) is a popular method for this purpose, which works by creating new examples that are interpolations between existing minority class instances and their nearest neighbors (Elreedy and Atiya 2019; Elreedy, Atiya, and Kamalov 2023). SMOTE aims to balance the class distribution and enhance the performance of classifiers on imbalanced datasets, making it a widely recognized and respected method in the field of machine learning (Mohd et al. 2019; Soltanzadeh and Hashemzadeh 2021).

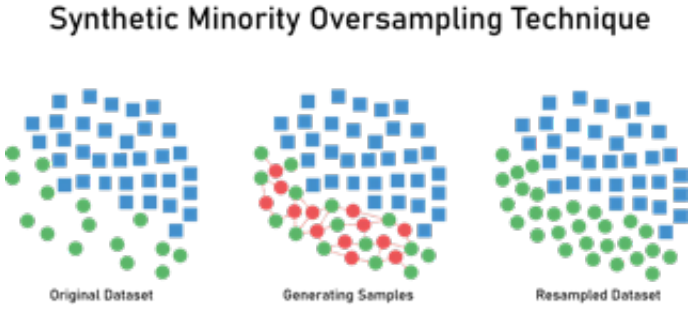

Figure 2.6: Synthetic Minority Oversampling Technique (Dholakiya 2020).

However, it should be noted that SMOTE is not without its limitations. It has been known to cause over-generalization, over-sampling of irrelevant or uninformative samples, and increased class overlap, which may adversely affect the performance of classifiers (Soltanzadeh and Hashemzadeh 2021). To address these challenges, modifications to SMOTE have been suggested, such as Range-Controlled SMOTE (RCSMOTE), which incorporate mechanisms to regulate the generation of synthetic samples and mitigate the aforementioned issues(Soltanzadeh and Hashemzadeh 2021). Additionally, SMOTE has been employed in various domains, including clinical disease classification, where it has demonstrated improved accuracy in machine learning models(Mohd et al. 2019).

It has also been integrated with other techniques, such as principal component analysis (PCA) for authorship attribution in Arabic text data, achieving high accuracy(Puri and Kumar Gupta 2022). SMOTE is a crucial technique for managing class imbalance by creating synthetic samples for the minority class, thereby facilitating the development of more balanced and effective machine learning models. Despite these challenges, ongoing research and hybrid approaches continue to refine its application across diverse fields, enhancing its usefulness and performance in practical application(Puri and Kumar Gupta 2022; Soltanzadeh and Hashemzadeh 2021).

2.3.1.2 Adaptive Synthetic Sampling (ADASYN). ADASYN is an extension of SMOTE that generates synthetic samples for the minority class, focusing more on difficult-to-learn examples. It adjusts the weights of minority class examples based on their learning difficulty.

$$G_i = \left(\frac{\Delta_i}{k}\right) G \tag{2.6}$$

where Δ_i is the number of majority class neighbors of minority class instance X_i and k is the total number of neighbors. G_i determines the number of synthetic samples for each minority instances and G is the total number of synthetic samples required.

ADASYN, a technique designed to address class imbalance by generating synthetic samples for the minority class, is particularly effective when the learning difficulty varies among minority class examples. This approach focuses on harder-to-learn instances and adaptively shifts the decision boundary towards these challenging areas(He et al. 2008). The effectiveness of ADASYN has been demonstrated in various domains, such as improving tornado prediction by combining it with the local outlier factor (LOF) algorithm to enhance model performance and noise immunity (Qing et al. 2022), and in bioinformatics, where it is coupled with ensemble multi-filter techniques to address the dual challenges of high dimensionality and class imbalance in DNA microarray datasets (Sharifai, Muraina, and Abdurrahman 2022).

Although ADASYN has shown effectiveness in numerous scenarios, it has limitations that should be taken into account. Originally designed for low-dimensional binary feature spaces, it

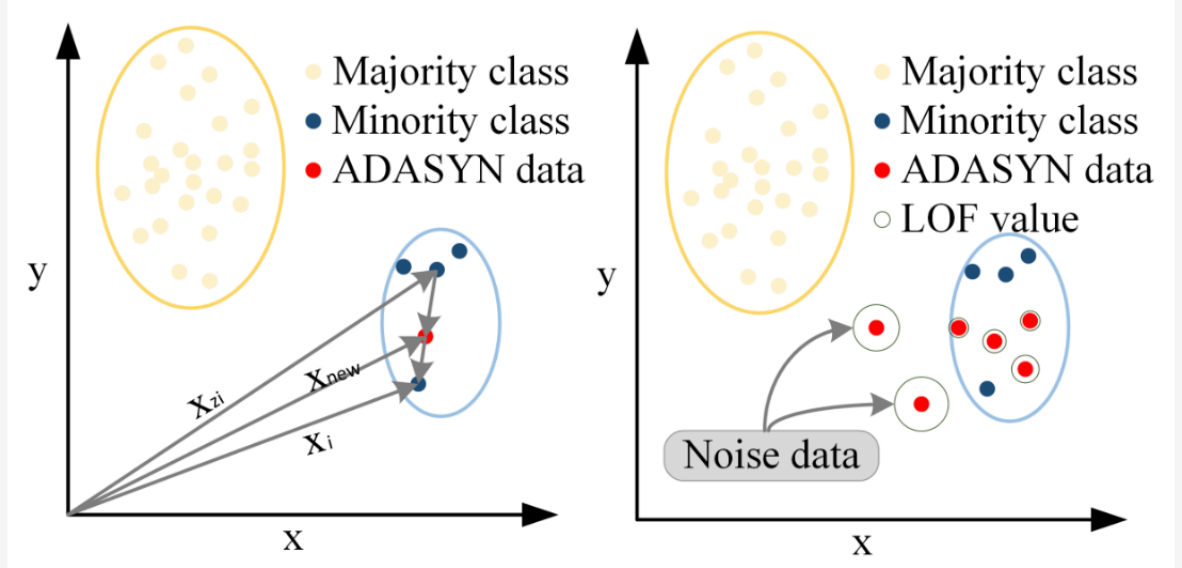

Figure 2.7: ADASYN : (Qing et al. 2022)

may not perform optimally in high-dimensional multi-class data without appropriate modifications (Xie 2024). Furthermore, ADASYN can be susceptible to challenges such as 'within class imbalance' and the 'small disjunct problem', which can be addressed by more advanced algorithms like KNNOR that consider the relative density and location of minority class samples (Islam et al. 2022). Despite these limitations, ADASYN has been successfully applied across various fields and datasets, making it a valuable tool in the machine learning toolkit for addressing class imbalance. Its ability to adapt to the difficulty of learning minority class examples makes it a robust choice, although it may require enhancements or complementary techniques to tackle specific challenges such as high dimensionality or within-class imbalances (Qing et al. 2022; He et al. 2008; Islam et al. 2022; Sharifai, Muraina, and Abdurrahman 2022; Xie 2024).

2.3.1.3 NearMiss(1). NearMiss is a method that employs under-sampling to balance the class distribution in datasets with an imbalance. By selecting majority class examples that are closest to the minority class examples, this technique aims to focus on examples near the decision boundary, potentially enhancing model performance (Mathew and Gunasundari 2023).

NearMiss-1 adopts a strategy of selecting majority class samples that are the closest to the minority class samples in terms of the average distance to the three nearest minority class sam-

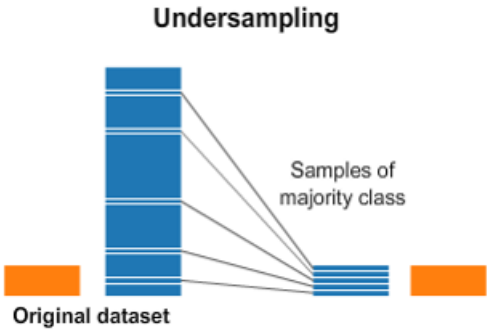

Figure 2.8: Undersampling Techniques; (Rutecki 2022)

ples. This approach guarantees that the majority class samples that exhibit the greatest similarity to the minority class samples are retained, ultimately refining the decision boundary (Mathew and Gunasundari 2023).

$$\text{NearMiss-1}(i) = \frac{1}{k}\sum_{j=1}^{k} d(X_i, X_j) \tag{2.7}$$

where $d(X_i, X_j)$ is the distance between majority class instance X_i and X_j, k is the number of nearest minority class neighbors.

The NearMiss1 under sampling technique is a method for addressing class imbalance in datasets by reducing the number of instances in the majority class. This technique specifically focuses on selecting majority class samples that are closest to the minority class samples, with the aim of preserving the majority class instances that are nearest to the minority class boundary (Mathew and Gunasundari 2023). However, it is essential to recognize that while NearMiss1 aims to balance the class distribution, it may inadvertently discard potentially useful or informative majority class instances, which could result in a loss of valuable information and potentially impact the performance of the classifier(Mathew and Gunasundari 2023). The design of NearMiss1 is intended to mitigate class imbalance by selecting the closest majority class instances to the minority class. Nevertheless, this approach may result in the loss of valuable information, which is a limitation that should be taken into account when applying this technique to imbalanced datasets(Mathew and Gunasundari 2023).

2.3.1.4 NearMiss(3). The NearMiss-3 methodology identifies majority class samples that are closest to the minority class samples based on the average distance to the three farthest minority class samples. By concentrating on retaining majority class samples that are farthest from the minority class samples, this approach ensures that the majority class samples that are most challenging to differentiate are preserved(Qing et al. 2022; Mathew and Gunasundari 2023).

$$\text{NearMiss-3}(i) = \frac{1}{k}\sum_{j=1}^{k} d(X_i, X_j) \tag{2.8}$$

where $d(X_i, X_j)$ is the distance between majority class instance X_i and the farthest X_j, k is the number of nearest minority class neighbors. NearMiss-3 determines the distances for each majority class sample to all minority class samples and designates the three farthest minority class samples for each majority class sample. The average distance from each majority class sample to its three farthest minority class samples is then computed. Lastly, the majority class samples with the smallest average distance to the three farthest minority class samples are chosen until the desired class balance is attained(Qing et al. 2022).

2.3.1.5 SMOTE Tomek. SMOTE Tomek is a hybrid data resampling method utilized to address the class imbalance issue in machine learning datasets. It integrates the Synthetic Minority Over-sampling Technique (SMOTE) with Tomek to balance the classes. SMOTE generates artificial samples for the minority class, while Tomek Links identifies and eliminates instances that are nearest neighbors but belong to opposite classes, which can be thought of as noise or borderline cases. Several studies have supported the effectiveness of SMOTE Tomek. Yu et al.(Yu et al. 2023) have demonstrated its utility in improving classification accuracy. Additionally,(Huisa et al. 2023; Jonathan, P. H. Putra, and Ruldeviyani 2020; Kotb and Ming 2021) have also highlighted the benefits of this approach.

$$\text{Tomek Links} = \{(x_i, x_j) \mid d(x_i, x_j) < d(x_i, x_k), \forall x_k \neq x_i\} \tag{2.9}$$

where $d(X_i, X_j)$ is the distance between X_i and the farthest X_j

The SMOTE Tomek method has demonstrated its utility in improving the performance of various machine learning algorithms in diverse domains. This technique has been used to enhance the precision-recall for minority classes in beauty product reviews (Jonathan, P. H. Putra, and Ruldeviyani 2020), significantly improve sentiment classification accuracy in user reviews (Switrayana et al. 2023), and increase the quality of cardiac PCG signal classification (Huisa et al. 2023). Additionally, it has been effective in predicting insurance premium defaults(Kotb and Ming 2021), classifying failure modes of reinforced concrete columns (Yu et al. 2023) , and improving software quality prediction (Jonathan, P. H. Putra, and Ruldeviyani 2020). The technique has also shown promise in reducing misclassification rates in extremely imbalanced datasets (Switrayana et al. 2023), addressing cyberbullying detection (Khairy, Mahmoud, and Abd-El-Hafeez 2024), and enhancing intrusion detection systems. SMOTE Tomek is a valuable resampling method that mitigates the challenges posed by imbalanced datasets, thereby enhancing the predictive accuracy and generalizability of machine learning models. Its effectiveness is evidenced by improved classification metrics across various studies, making it a recommended approach for dealing with class imbalance issues (Switrayana et al. 2023; Huisa et al. 2023; Khairy, Mahmoud, and Abd-El-Hafeez 2024; Kotb and Ming 2021; Yu et al. 2023) .

2.3.2 Feature Selection Methods

Feature selection is the process of identifying a subset of relevant features for use in constructing a model. Its purpose is to simplify the model, reduce overfitting, improve accuracy, and decrease training time. By eliminating redundant or irrelevant data, feature selection aims to enhance the model's performance by preventing increased complexity and overfitting(Hamdi et al. 2022). In this study, three categories of feature selection methods were employed: Filter Methods, Wrapper Methods, and Embedded Methods, as well as hybrid methods that integrate these techniques.

2.3.2.1 Filter Methods. Filter methods are statistical measures used to evaluate and rank features based on their score. These scores are then utilized to determine which features should be retained or eliminated prior to the application of the machine learning algorithm (Bharti and P. k. Singh 2014) . These methods are favored for their simplicity, computational efficiency, and independence from the learning process(Bharti and P. k. Singh 2014; Prastyo, Ardiyanto, and Hidayat 2020).

1. Chi-Square Test

   The Chi-square test is a statistical method used to assess the relationship between categorical variables and the target variable. It is employed to evaluate the independence between features and class labels with the aim of identifying the most informative features for predicting class membership (Anamisa, Mufarroha, and Jauhari 2023; A. E. Putra, Wardhani, et al. 2019; Qiu, W. Wang, and D. Y. Liu 2013).

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2.10}$$

   where (O_i) is the observed frequency and (E_i) is the expected frequency. The importance of the chi-square test in feature selection is demonstrated by its capacity to improve classification accuracy across a range of datasets and algorithms, including support vector machines, K-nearest neighbors, and Naive Bayes (A. E. Putra, Wardhani, et al. 2019; Yuxian Wang and Zhou 2021).

2. Information Gain

   Information Gain (IG) is a widely utilized criterion for feature selection in machine learning, as it assesses the reduction in entropy or uncertainty about a class due to the presence of a feature. Shi et al.(Shi et al. 2014) propose an enhanced IG method that incorporates word frequency and sentiment, demonstrating improved performance in Chinese text sentiment categorization. This method measures the entropy reduction when a feature is known. Similarly, B and V (BN 2022) apply IG in conjunction with other filter-based techniques

to identify an optimal feature subset for cervical cancer diagnosis, resulting in improved classifier performance. Moreover, (Habib and Khursheed 2022) employ IG along with other statistical techniques to rank features for detecting DDoS attacks, contributing to the superior performance of various machine learning models.

$$IG(T,X) = H(T) - H(T \mid X) \tag{2.11}$$

where $(H(T))$ is the entropy of the target variable and $(H(T \mid X))$ is the conditional entropy.

IG is an effective tool for selecting features in machine learning, which often results in increased classification accuracy and simplified models. However, their usefulness can vary depending on the context, and alternative methods may be more suitable in specific situations. Thus, it is essential to consider the characteristics of the problem domain and the learning algorithm when choosing a feature scoring measure (BN 2022; Habib and Khursheed 2022; Shi et al. 2014).

3. ANOVA F-value

   The ANOVA F-value is a statistical index used to gauge the significance of differences between groups across multiple samples, and it is commonly utilized in feature selection to identify features that have a strong relationship with the outcome variable. Several studies have demonstrated the usefulness of the ANOVA F-value in feature selection across a range of domains. For example, Shaharum et al. (Shaharum, Sundaraj, and Helmy 2015) and (Quek et al. 2023) have both highlighted the effectiveness of feature selection methods, such as one-way ANOVA, in improving classification performance in high-dimensional datasets when combined with machine learning algorithms like Artificial Neural Networks (ANN) and Vector Machines (SVM) (Quek et al. 2023; Shaharum, Sundaraj, and Helmy 2015). The ANOVA F-value measures the linear relationship between continuous features and the target variable.

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

where:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$$

$$SS_{\text{between}} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2$$

$$SS_{\text{within}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

$$df_{\text{between}} = k - 1$$

$$df_{\text{within}} = N - k$$

where:

- SS_{between} is the sum of squares between the groups
- SS_{within} is the sum of squares within the groups
- df_{between} is the degrees of freedom between the groups
- df_{within} is the degrees of freedom within the groups
- k is the number of groups
- n_i is the number of observations in group i
- $\bar{X}_i$ is the mean of group i
- $\bar{X}$ is the overall mean of all groups
- X_{ij} is the observation j in group i

The analysis of variance (ANOVA) F-value is a crucial metric for selecting features, as it enables the identification of the most influential features that contribute to a model's predictive capabilities.

The literature supports its application in various contexts, ranging from enhancing classification accuracy to reducing computational complexity and feature dimensions. Nevertheless, alternative or enhanced methods may offer benefits in specific situations, indicating that the choice of feature selection technique should be customized to the unique characteristics of the data and the goals of the analysis (Jeong et al. 2022; Quek et al. 2023; Shaharum, Sundaraj, and Helmy 2015; Zambom and Akritas 2015).

2.3.2.2 Wrapper Methods. Wrapper methods involve the use of a predictive model to evaluate the efficacy of feature combinations and determine their usefulness. These techniques are a subset of feature selection methods that select subsets of features based on their predictive power using a specific machine learning algorithm (Sumi and Narayanan 2019). They evaluate the usefulness of feature subsets by training a model on them and utilizing the model's performance to guide the selection process(Wald, Khoshgoftaar, and Napolitano 2013). Wrapper methods use a predictive model to evaluate the combination of features and determine their effectiveness.

1. Recursive Feature Elimination (RFE)

   Recursive Feature Elimination (RFE) is a highly regarded approach for feature selection in the field of machine learning, particularly beneficial when dealing with high-dimensional data or limited sample sizes. It operates by repeatedly eliminating the least relevant features based on specific criteria to improve the performance of a predictive model (Jeon and Oh 2020; Priyatno, Widiyaningtyas, et al. 2024). RFE is an invaluable tool for feature selection, and its effectiveness can be further augmented through various adjustments and hybrid approaches. These enhancements are intended to retain vital features, reduce bias, and increase computational efficiency, thereby improving the overall predictive performance of machine learning models (Brzezinski 2020; Ou et al. 2017; Yan and D. Zhang 2015).

$$\text{RFE}(X, y) = \min_{\theta} \sum_{i=1}^{n} L(y_i, f(X_i, \theta)) \tag{2.12}$$

where L is the loss function, y_i is the true value, X_i is the feature set, and θ are the model parameters.

2. Logistic Regression

Logistic Regression (LR) is widely recognized as a valuable tool for feature selection in various domains, including big data, hyperspectral imaging, and healthcare. Studies reviewed demonstrate a range of approaches to enhance the efficiency and accuracy of feature selection using LR models. Wichitaksorn et al.(Wichitaksorn, Kang, and F. Zhang 2023) propose a random subspace logistic regression method that offers a computationally less expensive alternative to traditional methods while maintaining classification accuracy. Pal, (Pal 2012) highlights the effectiveness of multinomial logistic regression-based feature selection, particularly the Cawley and Talbot approach, which requires no user-defined parameters and outperforms other methods in terms of computational efficiency and classification accuracy. However, Zhang et al. (S. Zhang et al. 2015) suggest that the LASSO method may not be satisfactory when the number of predictors significantly exceeds the number of observations, proposing the logistic elastic net as a superior alternative. Gunasekaran and Dhandayudam introduce the Multi filter union (MFU) feature selection method, which combines random forest and logistic regression algorithms, showing high performance in breast cancer datasets(Morkonda Gunasekaran and Dhandayudam 2021). Tsou et al.'s EDLRT algorithm emphasizes the use of entropy and dummy variables in logistic regression for decision tree processes, offering tolerance to missing values and effective outlier detection(Tsou, Chi, and Huang 2010).

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{p} \beta_j X_j)}} \tag{2.13}$$

$$\min_{\theta} \sum_{i=1}^{n} L(y_i, \hat{y}_i) \tag{2.14}$$

where L is the Log-Loss function:

$$L(y_i, \hat{y}_i) = -\left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right] \tag{2.15}$$

and

$$\hat{y}_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}} \tag{2.16}$$

where: y_i is the true label for the i-th sample, X_{ij} is the j-th feature for the i-th sample, β_0 is the intercept, β_j are the coefficients for the features, θ represents all the model parameters.

3. Random Forest: Random Forest (RF), a well-established method in the field, has demonstrated its reliability and effectiveness through a range of applications, such as predictive modeling for housing prices, solar radiation prediction, and cyberattack detection (Bojarajulu, Tanwar, and A. Rana 2021; Chaibi et al. 2022; Rai 2019). One of its key advantages is its ability to handle a significant number of variables, missing data, outliers, and noisy data (Jaiswal and Samikannu 2017). By employing an ensemble of decision trees, RF evaluates feature importance. As an ensemble learning method, RF creates multiple decision trees during training and provides the average prediction (regression) or majority vote (classification) of the individual trees. Additionally, it can be utilized as a wrapper method to assess feature importance by measuring the contribution of each feature to the model's predictive performance. For a given input X, the Random Forest prediction $\hat{y}$ is the average of the predictions from each individual decision tree:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(X) \tag{2.17}$$

where:

- T is the total number of trees in the forest.
- $f_t(X)$ is the prediction of the t-th decision tree for the input X.

The importance of a feature j can be evaluated by measuring the average decrease in impurity (such as, Gini impurity or entropy) across all trees in the forest. For each tree, calculate the total decrease in impurity from splitting on the feature and average this value over all trees. The feature importance score for feature j is given by:

$$\text{Importance}(X_j) = \frac{1}{T} \sum_{t=1}^{T} \sum_{n \in \text{nodes}} \frac{I_n(\text{split} = X_j)}{N_n} \tag{2.18}$$

where:

- $I_n(\text{split} = X_j)$ is the decrease in impurity from splitting on feature X_j at node n in tree t.
- N_n is the total number of samples that pass through node n.
- The inner sum is over all nodes n in tree t where feature X_j is used for splitting.

2.3.2.3 Embedded Methods. Embedded methods for feature selection are integrated within the learning algorithm itself and perform feature selection during the model training process. These methods often combine the qualities of filter and wrapper methods, aiming to take advantage of their strengths while mitigating their weaknesses(Molla et al. 2022; Zakharov and Dupont 2011). Interestingly, while embedded methods are designed to improve model performance, their effectiveness can vary depending on the context. For instance, the novel feature selection algorithm embedded in logistic regression described in (Zakharov and Dupont 2011) is particularly suited for high-dimensional biomedical data and outperforms other methods like Elastic Net and Random Forests in terms of stability and predictive performance. Similarly, the use of Lasso and Elastic-net as embedded feature selection techniques in (Jomthanachai, W. P. Wong, and Khaw 2022) and (Al Tawil et al. 2024) demonstrates their utility in identifying relevant features for predicting logistics performance and breast cancer, respectively. However, the effectiveness of these methods can be influenced by the specific characteristics of the dataset and the problem at hand.

1. LASSO (Least Absolute Shrinkage and Selection Operator)

   LASSO adds a penalty term to the regression model to perform feature selection. The penalty term is the sum of the absolute values of the coefficients, which forces some of the coefficients to be exactly zero, effectively selecting a subset of the features(Kumarage, Yogarajah, and Ratnarajah 2019). LASSO's ability to handle both continuous and discrete variables makes it versatile, as demonstrated in credit scoring models(Choi, Koo, and Park 2015), and its interpretability is enhanced when combined with other techniques such as Partial Least Squares(C. Li and W. Li 2010). The objective function for LASSO is given by:

$$\text{LASSO} : \min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.19}$$

   where:

   - y_i is the true value for the i-th sample.
   - X_i is the feature set for the i-th sample.
   - β are the model coefficients.
   - λ is the regularization parameter.
   - p is the number of features.

2. Ridge Regression (L2 Regularization)

   Ridge regression, also known as L2 regularization, is a technique used to address multicollinearity regression problems by penalizing the size of the coefficients (Wu 2021). However, it has been criticized for not being able to perform variable selection since it does not set coefficients exactly to zero(Wu 2021). Despite this, various methods have been proposed to enhance feature selection capabilities in the context of ridge regression. Ridge Regression adds a penalty term to the regression model to perform feature selection. The penalty term is the sum of the squared values of the coefficients, which helps in shrinking the coefficients but does not force them to be exactly zero(Paul and Drineas 2016; Lan, Hou, and Yi 2016).

The objective function for Ridge Regression is given by:

$$\text{Ridge Regression} : \min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.20}$$

where:

- y_i is the true value for the i-th sample.
- X_i is the feature set for the i-th sample.
- β are the model coefficients.
- λ is the regularization parameter.
- p is the number of features.

3. Decision Tree-based Feature Importance

   Decision tree-based feature importance measures feature importance based on the tree's structure and splits. The importance of a feature is calculated as the total reduction of the criterion (e.g., Gini impurity or entropy) brought by that feature. Decision Tree-based Feature Importance is a key technique for feature selection that can lead to improved model accuracy and efficiency. Its utility is evidenced across various domains, from healthcare to network traffic analysis, and it is particularly effective when combined with other feature selection methods (Agraz 2023; Aouedi, Piamrat, and Parrein 2021). However, the optimal application of this method requires careful consideration of the dataset and the objectives of the machine learning task. The importance score for a feature j is given by:

$$\text{Importance}(X_j) = \sum_{t=1}^{T} \sum_{n \in \text{nodes}} I_n(\text{split} = X_j) \tag{2.21}$$

   where:

   - T is the total number of trees in the forest.
   - n are the nodes in the decision tree.

- $I_n(\text{split} = X_j)$ is the reduction in impurity brought by the feature X_j at node n.

This method evaluates how much each feature contributes to the model by looking at the improvement in the splitting criterion (like Gini impurity or entropy) brought by each feature.

2.3.2.4 Hybrid Methods. Hybrid methods combine Filter, Wrapper, and Embedded methods to select the most informative features. This approach leverages the strengths of each method to improve feature selection. These robust methods were implemented in this study to ensure the selection of the most relevant features, which were then passed through Random Forest (RF) and Artificial Neural Network (ANN) models for further analysis.

1. Method 1: Chi-Squared Test (Filter) $\rightarrow$ RFE (Wrapper) $\rightarrow$ LASSO (Embedded)

   - The Chi-Squared test(Filter) evaluates the independence of features with respect to the target variable. It calculates the Chi-Squared statistic for each feature and selects the ones with the highest scores.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2.22}$$

     where O_i is the observed frequency and E_i is the expected frequency.

   - The Recursive Feature Elimination (RFE)(Wrapper) recursively removes the least important features and builds the model on the remaining features. It ranks features based on their importance to the model's performance.

$$\min_{\theta} \sum_{i=1}^{n} L(y_i, f(X_i, \theta)) \tag{2.23}$$

     where L is the loss function.

   - The LASSO (Embedded) adds a penalty term to the regression model, shrinking some coefficients to zero, effectively selecting a subset of features.

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.24}$$

2. Method 2: Information Gain (Filter) $\rightarrow$ Logistic Regression (Wrapper) $\rightarrow$ Ridge Regression (L2 Regularization) (Embedded)

   - Information Gain (Filter): Information gain measures the reduction in entropy when a feature is known, selecting features that provide the most information about the target variable.

$$IG(T,X) = H(T) - H(T \mid X) \tag{2.25}$$

     where $H(T)$ is the entropy of the target variable and $H(T \mid X)$ is the conditional entropy.

   - Logistic Regression (Wrapper): Logistic regression is used to evaluate feature subsets by fitting the model and minimizing the Log-Loss function.

$$\min_{\theta} \sum_{i=1}^{n} -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{2.26}$$

     where $\hat{y}_i = \frac{1}{1+e^{-(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}}$.

   - Ridge Regression (Embedded): Ridge regression adds a penalty term to the regression model, shrinking the coefficients without setting them to zero, helping to handle multi-collinearity.

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.27}$$

3. Method 3: ANOVA F-value (Filter) $\rightarrow$ Random Forest (Wrapper) $\rightarrow$ Decision Tree-based Feature Importance (Embedded)

- ANOVA F-value (Filter): ANOVA F-value measures the linear relationship between continuous features and the target variable, selecting features with the highest F-values.

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{2.28}$$

- Random Forest (Wrapper): Random Forest uses an ensemble of decision trees to evaluate feature importance by measuring the average decrease in impurity across all trees.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \tag{2.29}$$

  where $f_t(X)$ is the prediction of the t-th tree.

- Decision Tree-based Feature Importance (Embedded): Decision tree-based feature importance calculates the total reduction in impurity brought by each feature, averaging this value over all trees in the forest.

$$\text{Importance}(X_j) = \sum_{t=1}^{T}\sum_{n\in\text{nodes}} I_n(\text{split} = X_j) \tag{2.30}$$

These hybrid methods were used to select the most relevant features from each dataset. The selected features were then passed through Random Forest (RF) and Artificial Neural Network (ANN) models for further analysis. This comprehensive approach ensures that the most informative features are selected, enhancing the predictive performance and robustness of the models.

2.3.3 Model Training and Evaluation

Model training and evaluation are critical processes in the development of machine learning (ML) models, as they determine the model's ability to generalize to new data. The training phase involves adjusting the model's parameters to fit the training data, while evaluation assesses the model's performance on unseen data, often using a separate test set (Sharma and Gosain 2022). Hyperparameter tuning is an essential aspect of model training, as it can significantly enhance

model performance(Quan 2024) . An interesting fact is that while hyperparameter tuning is crucial, the method of tuning can vary in effectiveness. For instance, Quan (Quan 2024) highlights that random search can be particularly effective for hyperparameter tuning in urban building energy models. In contrast, Yao et al.(Yao et al. 2017) introduces a layer-by-layer strategy for hyperparameter optimization in deep generative models, which differs from traditional Bayesian optimization methods. Additionally, Zhang et al. (2022) demonstrates the use of high-performance computing to expedite hyperparameter tuning for imbalanced data, which is a common challenge in ML applications. Moreover, the use of advanced techniques like high-performance computing can reduce computational time and enhance the efficiency of the tuning process(Y. Zhang 2018). It is evident that the choice of hyperparameter tuning and evaluation methods can have a substantial impact on the success of ML models in various applications (Quan 2024; Yao et al. 2017). This section describes the process of selecting, training, and evaluating machine learning models used in the study. The selected models are trained on the pre-processed and balanced datasets using the chosen feature subsets. Evaluation metrics are then used to assess the performance of the models.

2.3.3.1 Model Selection. Machine learning model selection is a crucial process that entails evaluating and selecting the most appropriate model for a specific task based on performance metrics and other factors. According to Liu et al. (2021), feature selection plays a significant role in enhancing the performance of machine learning models for landslide susceptibility assessment (LSA), with recursive feature elimination (RFE) optimized random forest (RF) emerging as the best feature selection-based machine learning (FS-ML) model for this task (X. Liu 2024). Liu (2024) compares the performance and complexity of different models, noting that Forest and Support Vector Machine (SVM) models generally outperform the simpler Linear Model, although they are more complex and less interpretable (X. Liu 2024). Hananya and Katz introduce Adaptive machine learning for Dynamic ENvironments (ADEN), a method for selecting the most suitable machine learning model for time-series data without additional training(Hananya and Katz 2024). Banda et al. emphasize the need for model selection based on dataset types, recommending logistic regression for categorical datasets(Banda, Ngassam, and Mnkandla 2022; Hananya and Katz 2024).

Ribeiro and Reynoso-Meza presented a multi-criteria decision-making process for the selection of Pareto-optimal machine learning models, which could provide better solutions than single-criterion optimization(Ribeiro and Reynoso-Meza 2024). Liu and Chen found that the SVM was superior in predicting judicial decisions, emphasizing the role of semantic information in feature selection(L.-L. Liu, C. Yang, and X.-M. Wang 2021). Aderibigbe et al. (2023) conducted a review of the impact of AI and ML on enhancing energy efficiency in electricity demand forecasting, highlighting the importance of selecting appropriate models based on various criteria (Bouktif et al. 2018). The machine learning models selected for this study are:

- Random Forest (RF)
- Artificial Neural Network (ANN)

These models are chosen for their ability to handle complex relationships in the data and their robustness in various scenarios.

2.3.3.2 Model Training. The selected models are trained on the training dataset. The training process involves feeding the models with input data (features) and corresponding output data (target variable) to learn the underlying patterns and relationships.

1. Random Forest (RF)

   RF is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction (regression) or majority vote (classification) of the individual trees. RF model training involves constructing decision trees using bagging and feature selection to optimize performance. While the number of trees is a factor in model accuracy, advancements in RF algorithms aim to balance performance with computational efficiency. These models have been successfully applied in various domains, including evaporation prediction, myoelectric control, and medical diagnosis, demonstrating their versatility and effectiveness (Javeed et al. 2019; Jiang, Ma, and Nazarpour 2024; R. Wang, K. Li, and Su 2022; A. Singh, Mittal, Amrender Kumar, et al. 2020).

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(X) \tag{2.31}$$

where T is the total number of trees in the forest and $f_t(X)$ is the prediction of the t-th decision tree for the input X.

2. Artificial Neural Network (ANN)

   ANNs is a computational model inspired by the way neural net- works in the human brain process information. It consists of layers of interconnected nodes (neurons) that process input data and generate predictions (Qamar and Zardari 2023). ANNs have been applied in diverse domains, demonstrating their versatility. For instance, they have been used to emulate the electroacoustic wave behavior in high-Q piezoelectric resonators(Almalkawi and Caron 2021), predict chemical properties in food technology(Baykal and Yildirim 2013), and optimize process parameters in manufacturing techniques like friction stir welding(Mubiayi and Rao 2020). Moreover, ANNs have shown improvements in robustness under adversarial attacks when trained using novel gradient computation methods(Patel, Tailor, and Ganatra 2021). The output of an ANN with one hidden layer can be represented as:

$$\hat{y} = \sigma\left(W_2 \cdot \sigma(W_1 \cdot X + b_1) + b_2\right) \tag{2.32}$$

   where:

   - σ is the activation function.
   - W_1 and W_2 are the weight matrices for the input and hidden layers, respectively.
   - b_1 and b_2 are the bias vectors for the input and hidden layers, respectively.
   - X is the input feature set.

2.3.3.3 Evaluation Metrics. Evaluation metrics are essential for assessing the performance of machine learning (ML) models across various tasks, such as classification, regression, and object detection(Rainio, Teuho, and Klén 2024). These metrics provide a quantitative basis for comparing models and are crucial for model selection and validation. For instance, accuracy and AUC-ROC are commonly used metrics for classification tasks, while MAPE and RMSE are employed for evaluating time series forecasting models (Acharya and Das 2024; Nazmi et al. 2023). Interestingly, while evaluation metrics are widely used in predictive analytics, their role extends beyond mere performance measurement. In sociology, for example, ML-derived classifications and predictive performance metrics can highlight theoretical gaps and stimulate the development of new theories(Manzo and Baldassarri 2015)). Moreover, in healthcare analytics, evaluation metrics not only gauge model performance but also contribute to enhancing patient care by informing clinical decisions(Acharya and Das 2024) . The performance of a models' is evaluated using the following metrics:

1. Accuracy

   Accuracy is a fundamental metric for classifier evaluation (Y. Liu et al. 2021; Sawadogo et al. 2022; Shao et al. 2019). Accuracy measures the proportion of true results among the total number of cases examined, it does not differentiate between the types of errors by the classifier. This can lead to misleading conclusions, particularly when the class distribution is skewed(Sawadogo et al. 2022). Proportion of correctly predicted instances can be measured by:

   $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.33}$$

   where:

   - TP is the number of true positives.
   - TN is the number of true negatives.
   - FP is the number of false positives.

- FN is the number of false negatives.

For imbalanced classes problems, accuracy is usually not a reliable performance metric, and therefore precision, recall, f1 score, and the Area under the curve (AUC) must be taken into consideration for this study(Sawadogo et al. 2022).

2. Precision

   Precision evaluation metrics are crucial for assessing the effectiveness of various systems, from context-aware recommender systems to clinical analyses. Champiri et al. (2019) underscores the importance of precision, among other metrics, in evaluating context-aware scholarly recommender systems, noting that these metrics are grouped and applied differently depending on the evaluation method (Dehghani Champiri, Asemi, and Siti Salwah Binti 2019). That is, it is the proportion of true positive instances among the instances predicted as positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.34}$$

3. Recall

   Recall, also known as the true positive rate or sensitivity, measures the proportion of actual positives that are correctly identified by the classifier (Khan and Z. A. Rana 2019).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.35}$$

4. F1 Score

   The F1 score is the harmonic mean of precision and recall, providing a balance between the two by considering both false positives and false negatives (Movahedi, Padman, and Antaki 2023).

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.36}$$

5. ROC-AUC

The ROC-AUC (Receiver Operating Characteristic - Area Under Curve) is a performance measurement for classification problems at various threshold settings. The ROC is a probability curve, and AUC represents the degree or measure of separability, indicating how well the model is capable distinguishing between classes(Khan and Z. A. Rana 2019; Smithson 2023) .

These metrics provide a comprehensive evaluation of the model's performance, considering both the accuracy of predictions and the balance between precision and recall. The models' performances are compared using *F1 Score and ROC-AUC* metrics to identify the best feature selection method and class imbalance handling technique.

CHAPTER III

RESULTS OF FEATURE SELECTION METHODS AND MACHINE LEARNING TECHNIQUES

In this section, the analysis of the benchmark results across the four datasets—Heart Disease, Fraud Detection, Breast Cancer, and Churn—reveals informative trends and variations in model performance. These results were obtained from 5 iterations of random states (123, 3030, 500, 126, and 2021).

3.1 Benchmark Analysis

Table 3.1: Benchmark Results for Heart Disease Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.93 ± 0.00	0.83 ± 0.00	0.93 ± 0.00	0.87 ± 0.00	10
Information Gain	0.94 ± 0.00	0.84 ± 0.01	0.94 ± 0.01	0.88 ± 0.00	10
ANOVA F-value	0.93 ± 0.00	0.83 ± 0.00	0.93 ± 0.01	0.88 ± 0.00	10
Wrapper Methods					
RFE	**0.94 ± 0.00**	**0.88 ± 0.00**	0.92 ± 0.00	0.83 ± 0.00	10
Logistic Regression	0.94 ± 0.00	0.84 ± 0.00	**0.94 ± 0.00**	**0.88 ± 0.00**	10
Random Forest	0.93 ± 0.00	0.87 ± 0.00	**0.94 ± 0.00**	**0.88 ± 0.00**	10
Embedded Methods					
LASSO Regression	0.93 ± 0.00	0.87 ± 0.00	0.93 ± 0.01	0.83 ± 0.12	10
Ridge Regression	0.93 ± 0.00	0.84 ± 0.00	0.93 ± 0.00	0.85 ± 0.01	10
Decision Tree	0.93 ± 0.00	0.81 ± 0.00	0.93 ± 0.00	0.85 ± 0.00	10

Heart Attack: 13,435; No Heart Attack: 232,544; Shape (245,979, 37)

Table 3.1 illustrates that RFE exhibits exceptional performance, with F1 scores with respect to the ROC-AUC while Logistic Regression and RF also performed well with values 0.94 for both methods. This indicates strong predictive capabilities. For the Fraud Detection dataset, Logistic Regression achieves noteworthy results, attaining perfect scores with ANN.

However, the fluctuation in standard deviations, particularly in LASSO Regression for ANN, suggests potential overfitting or model instability.

Table 3.2: Benchmark Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.99 ± 0.02	0.94 ± 0.03	0.99 ± 0.01	0.98 ± 0.01	10
Information Gain	0.99 ± 0.02	0.94 ± 0.03	0.99 ± 0.01	0.96 ± 0.05	10
ANOVA F-value	0.99 ± 0.02	0.94 ± 0.03	**0.99 ± 0.01**	**0.98 ± 0.01**	10
Wrapper Methods					
RFE	**0.99 ± 0.01**	**0.97 ± 0.02**	0.95 ± 0.10	0.83 ± 0.10	10
Logistic Regression	0.99 ± 0.02	0.95 ± 0.03	1.00 ± 0.01	0.98 ± 0.01	10
Random Forest	0.98 ± 0.02	0.96 ± 0.04	0.99 ± 0.01	0.97 ± 0.02	10
Embedded Methods					
LASSO Regression	0.99 ± 0.10	0.93 ± 0.04	0.99 ± 0.32	0.82 ± 0.34	10
Ridge Regression	0.99 ± 0.03	0.94 ± 0.04	0.97 ± 0.02	0.97 ± 0.02	10
Decision Tree	0.99 ± 0.04	0.94 ± 0.03	0.99 ± 0.01	0.96 ± 0.02	10

Fraud: 473; No Fraud: 283,253; Shape (283726, 30)

Table 3.2 of the fraud dataset demonstrates a noteworthy pattern, displaying a high F1-Score of 0.99 under the Random Forest (RF) algorithm with a feature selection rate of 0.97 and a receiver operating characteristic-area under the curve (ROC-AUC) of 0.99. Additionally, the ANOVA F-value exhibited strong performance under the artificial neural network (ANN) model, with an F1-Score of 0.99 and a ROC-AUC of 0.98.

Table 3.3 below demonstrates a remarkable trend, as almost all methods achieve high scores, with the exception of LASSO Regression in ANN, which significantly underperforms with an F1 score of 0.61 and ROC-AUC of 0.66. This outlier suggests possible inefficiencies in LASSO's ability to conduct feature selection for this specific type of data. Conversely, the Churn dataset exhibits greater variability among techniques, with Random Forest slightly outperforming the rest with an F1 score of 0.80 and ROC-AUC of 0.81, indicating a more complex relationship between model features and techniques.

Table 3.3: Benchmark Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	10
Information Gain	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	10
ANOVA F-value	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	10
Wrapper Methods					
RFE	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	10
Logistic Regression	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	10
Random Forest	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	10
Embedded Methods					
LASSO Regression	1.00 ± 0.00	1.00 ± 0.00	0.61 ± 0.01	0.66 ± 0.01	10
Ridge Regression	0.60 ± 0.01	0.66 ± 0.01	0.99 ± 0.00	0.99 ± 0.00	10
Decision Tree	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	10

Healed: 60,811; Not Healed: 95,313; Shape (156124, 19)

Table 3.4: Benchmark Results for Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.76 ± 0.01	0.79 ± 0.02	0.79 ± 0.01	0.84 ± 0.01	10
Information Gain	0.76 ± 0.01	0.80 ± 0.01	0.79 ± 0.01	0.85 ± 0.01	10
ANOVA F-value	0.76 ± 0.02	0.81 ± 0.01	0.80 ± 0.01	0.85 ± 0.01	10
Wrapper Methods					
RFE	0.78 ± 0.02	0.83 ± 0.01	0.78 ± 0.02	0.83 ± 0.01	10
Logistic Regression	0.76 ± 0.02	0.79 ± 0.01	0.79 ± 0.02	0.84 ± 0.01	10
Random Forest	0.80 ± 0.02	0.81 ± 0.03	0.79 ± 0.01	0.84 ± 0.01	10
Embedded Methods					
LASSO Regression	0.76 ± 0.02	0.79 ± 0.01	0.79 ± 0.01	0.84 ± 0.01	10
Ridge Regression	0.69 ± 0.07	0.68 ± 0.15	0.73 ± 0.10	0.71 ± 0.23	10
Decision Tree	0.73 ± 0.02	0.71 ± 0.01	0.73 ± 0.01	0.73 ± 0.01	10

Churn: 5,164; No Churn:1,857; Shape (7021, 20)

The findings presented in table 3.4, however, did not prove to be satisfactory. These disparities underscore the varying resilience and responsiveness of machine learning techniques to data set characteristics. Notably, Random Forest and Logistic Regression generally offer promising results, suggesting their value as dependable starting points for preliminary model training.

Nevertheless, the detected variances, particularly the potential overfitting in data sets with considerable imbalance, such as Breast Cancer, necessitate meticulous model validation and fine-tuning to guarantee reliability and precision.

3.2 SMOTE Analysis

The Synthetic Minority Over-sampling Technique (SMOTE) analysis conducted on the Heart Attack, Fraud Detection, Breast Cancer, and Churn datasets demonstrates notable progress in addressing imbalanced data. However, the efficacy of this approach varies depending on the specific models and techniques employed.

Table 3.5: SMOTE Results for Heart Attack Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
textbfFilter Methods					
Chi-Squared Test	0.94 ± 0.00	0.81 ± 0.00	0.92 ± 0.01	0.86 ± 0.02	10
Information Gain	0.93 ± 0.00	0.82 ± 0.00	0.91 ± 0.02	0.84 ± 0.03	10
ANOVA F-value	0.94 ± 0.00	0.81 ± 0.00	0.92 ± 0.03	0.85 ± 0.01	10
Wrapper Methods					
RFE	0.94 ± 0.01	0.82 ± 0.01	0.92 ± 0.02	0.86 ± 0.02	10
Logistic Regression	0.93 ± 0.01	0.80 ± 0.00	0.92 ± 0.01	0.86 ± 0.02	10
Random Forest	0.94 ± 0.00	0.87 ± 0.00	0.89 ± 0.01	0.83 ± 0.02	10
Embedded Methods					
LASSO Regression	0.83 ± 0.00	0.78 ± 0.01	0.84 ± 0.02	0.79 ± 0.02	10
Ridge Regression	0.85 ± 0.01	0.77 ± 0.01	0.85 ± 0.01	0.79 ± 0.02	10
Decision Tree	0.88 ± 0.00	0.80 ± 0.00	0.87 ± 0.02	0.82 ± 0.01	10

Heart Attack: 162,778; No Heart Attack: 162,778

For the Heart Attack dataset (3.5), various techniques, including the Chi-Squared Test, Information Gain, and ANOVA F-value, demonstrate exceptional effectiveness, achieving F1 scores approximately 0.94 and ROC-AUC values in the low 0.80 range for both Random Forest (RF) and Artificial Neural Networks (ANN). The results indicate that key features in this dataset are highly predictive. However, LASSO and Ridge Regression show subpar performance, particularly in terms of ROC-AUC, suggesting potential limitations in these models' compatibility with the dataset or the SMOTE application itself.

Table 3.6: SMOTE Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.99 ± 0.01	0.96 ± 0.02	0.99 ± 0.35	0.94 ± 0.06	10
Information Gain	1.00 ± 0.00	0.97 ± 0.01	1.00 ± 0.07	0.95 ± 0.01	10
ANOVA F-value	1.00 ± 0.00	0.97 ± 0.01	1.00 ± 0.13	0.96 ± 0.01	10
Wrapper Methods					
RFE	1.00 ± 0.00	0.97 ± 0.01	0.99 ± 0.03	0.94 ± 0.02	10
Logistic Regression	1.00 ± 0.00	0.97 ± 0.01	0.99 ± 0.13	0.97 ± 0.02	10
Random Forest	1.00 ± 0.00	0.97 ± 0.01	1.00 ± 0.05	0.96 ± 0.01	10
Embedded Methods					
LASSO Regression	1.00 ± 0.02	0.97 ± 0.01	1.00 ± 0.09	0.96 ± 0.02	10
Ridge Regression	1.00 ± 0.01	0.96 ± 0.02	1.00 ± 0.11	0.96 ± 0.02	10
Decision Tree	1.00 ± 0.02	0.93 ± 0.02	1.00 ± 0.07	0.96 ± 0.02	10

Fraud: 198,274; No Fraud: 198,274

In the Fraud Detection dataset(3.6), there is a remarkable consistency in high performance, with many techniques achieving perfect or nearly perfect scores. This suggests a well-defined feature set that responds exceptionally well to SMOTE. Notably, there is some variability in the ANN model's performance, as seen in the standard deviations, particularly with the RFE and Logistic Regression methods, indicating possible overfitting.

Table 3.7: SMOTE Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	1.00 ± 0.00	10
Information Gain	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
ANOVA F-value	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Wrapper Methods					
RFE	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Logistic Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Random Forest	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Embedded Methods					
LASSO Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Ridge Regression	0.43 ± 0.00	0.56 ± 0.00	0.46 ± 0.04	0.57 ± 0.02	10
Decision Tree	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10

Healed: 66,693; Not Healed: 66,693

The Breast Cancer dataset(3.7) stands out with nearly universal perfect scores across all methods and both models, except for Ridge Regression, which notably dips in performance. This outlier highlights a potential mismatch between the method and the dataset characteristics or the SMOTE methodology. The general high performance across other methods suggests strong feature relationships and an effective application of SMOTE in balancing class distribution.

Table 3.8: SMOTE Results for Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Square	0.75 ± 0.00	0.78 ± 0.01	0.76 ± 0.01	0.83 ± 0.01	10
Info-Gain	0.75 ± 0.01	0.80 ± 0.01	0.76 ± 0.02	0.83 ± 0.01	10
ANOVA	0.74 ± 0.01	0.79 ± 0.01	0.75 ± 0.01	0.83 ± 0.01	10
Wrapper Methods					
RFE	0.77 ± 0.01	0.83 ± 0.01	0.76 ± 0.01	0.83 ± 0.01	10
Logistic	0.75 ± 0.01	0.80 ± 0.01	0.75 ± 0.01	0.83 ± 0.01	10
Random Forest	0.76 ± 0.00	0.80 ± 0.00	0.75 ± 0.01	0.83 ± 0.01	10
Embedded Methods					
Decision Tree	0.77 ± 0.01	0.81 ± 0.01	0.79 ± 0.01	0.83 ± 0.01	10
Ridge Regression	0.71 ± 0.00	0.74 ± 0.01	0.73 ± 0.01	0.81 ± 0.01	10
Lasso	0.76 ± 0.02	0.79 ± 0.01	0.79 ± 0.01	0.84 ± 0.01	10

Churn: 4,122; No Churn: 4122

The Churn dataset(3.8) above presents a more intricate picture of the influence of SMOTE, with generally good but not outstanding performance across various techniques. While Lasso in the ANN model achieves the best F1 and ROC-AUC scores, the overall results do not reach the exceptional levels observed in other datasets. This suggests a potentially more complex or noisy dataset where the impact of SMOTE is discernible but more restrained. These findings emphasize the utility of SMOTE in enhancing model performance across a variety of datasets, particularly in those with strong predictive features. However, the technique's effectiveness can be influenced by the specific characteristics of the dataset and the compatibility with the chosen analytical methods.

3.3 ADASYN Analysis

The ADASYN (Adaptive Synthetic Sampling Approach) analysis across the four datasets—Heart Attack, Fraud Detection, Breast Cancer, and Churn—reveals distinct patterns in effectiveness, demonstrating how this technique influences model performance in varied contexts.

Table 3.9: ADASYN Results for Heart Attack Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.84 ± 0.00	0.76 ± 0.01	0.83 ± 0.01	0.81 ± 0.01	10
Information Gain	0.86 ± 0.00	0.78 ± 0.00	0.84 ± 0.00	0.84 ± 0.00	10
ANOVA F-value	0.85 ± 0.00	0.77 ± 0.00	0.84 ± 0.01	0.82 ± 0.01	10
Wrapper Methods					
RFE	0.85 ± 0.00	0.83 ± 0.00	0.85 ± 0.01	0.82 ± 0.01	10
Logistic Regression	0.80 ± 0.00	0.75 ± 0.00	0.80 ± 0.01	0.74 ± 0.01	10
Random Forest	0.92 ± 0.00	0.83 ± 0.00	0.87 ± 0.02	0.79 ± 0.01	10
Embedded Methods					
LASSO Regression	0.83 ± 0.00	0.77 ± 0.00	0.83 ± 0.02	0.79 ± 0.01	10
Ridge Regression	0.92 ± 0.00	0.83 ± 0.00	0.87 ± 0.01	0.79 ± 0.01	10
Decision Tree	0.88 ± 0.00	0.79 ± 0.00	0.83 ± 0.02	0.82 ± 0.01	10

Heart Attack: 164,902; No Heart Attack: 162,778

The performance of the heart attack dataset (3.9) displays moderate results with F1 scores ranging from 0.80 to 0.92 and ROC-AUCs ranging from 0.75 to 0.87, across Random Forest (RF) and artificial neural network (ANN) models. Notably, the Random Forest model achieves a high F1 score of 0.92, indicating strong compatibility with the oversampling method. However, methods like Logistic Regression and LASSO Regression exhibit lower metrics, suggesting variability in how different models leverage the synthetic data generated by ADASYN.

The Fraud Detection Dataset (3.10) above demonstrates nearly perfect metrics across most methods, showcasing ADASYN's effectiveness in balancing highly skewed datasets. The exception is the Random Forest model in the RF configuration, which exhibits significant variability (SD=0.30), possibly indicating model overfitting or instability due to synthetic sample integration.

Table 3.10: ADASYN Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	1.00 ± 0.00	0.94 ± 0.03	1.00 ± 0.00	0.94 ± 0.05	10
Information Gain	1.00 ± 0.00	0.96 ± 0.01	1.00 ± 0.00	0.92 ± 0.06	10
ANOVA F-value	1.00 ± 0.00	0.96 ± 0.01	1.00 ± 0.00	0.94 ± 0.05	10
Wrapper Methods					
RFE	0.99 ± 0.01	0.98 ± 0.01	1.00 ± 0.00	0.90 ± 0.08	10
Logistic Regression	1.00 ± 0.00	0.96 ± 0.01	1.00 ± 0.00	0.95 ± 0.05	10
Random Forest	0.86 ± 0.30	0.97 ± 0.01	1.00 ± 0.00	0.91 ± 0.09	10
Embedded Methods					
LASSO Regression	1.00 ± 0.00	0.96 ± 0.01	0.97 ± 0.06	0.90 ± 0.08	10
Ridge Regression	1.00 ± 0.00	0.96 ± 0.02	0.98 ± 0.03	0.96 ± 0.01	10
Decision Tree	1.00 ± 0.00	0.96 ± 0.01	1.00 ± 0.00	0.92 ± 0.10	10

Fraud: 198,330; No Fraud: 198,274

Table 3.11: ADASYN Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.99 ± 0.00	10
Information Gain	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
ANOVA F-value	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Wrapper Methods					
RFE	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Logistic Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Random Forest	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Embedded Methods					
LASSO Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Ridge Regression	0.47 ± 0.05	0.56 ± 0.01	0.47 ± 0.05	0.56 ± 0.01	10
Decision Tree	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10

Healed: 66,692; Not Healed: 66,698

The breast cancer dataset (3.11) displays near-perfect or perfect performance for the majority of techniques, highlighting the dataset's clear, discriminative features that remain effective after the application of ADASYN. The exception is Ridge Regression, which significantly underperforms, suggesting a potential misalignment with the dataset's characteristics or the synthetic sampling method.

Table 3.12: ADASYN Results for Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.78 ± 0.01	0.75 ± 0.00	0.83 ± 0.01	0.74 ± 0.02	10
Information Gain	0.79 ± 0.01	0.74 ± 0.01	0.83 ± 0.01	0.74 ± 0.02	10
ANOVA F-value	0.77 ± 0.01	0.73 ± 0.00	0.82 ± 0.01	0.74 ± 0.02	10
Wrapper Methods					
RFE	0.83 ± 0.01	0.76 ± 0.01	0.83 ± 0.01	0.74 ± 0.03	10
Logistic Regression	0.78 ± 0.01	0.74 ± 0.01	0.83 ± 0.01	0.74 ± 0.01	10
Random Forest	0.79 ± 0.01	0.75 ± 0.01	0.83 ± 0.01	0.74 ± 0.02	10
Embedded Methods					
LASSO Regression	0.79 ± 0.01	0.74 ± 0.01	0.84 ± 0.01	0.75 ± 0.01	10
Ridge Regression	0.75 ± 0.02	0.71 ± 0.02	0.81 ± 0.01	0.71 ± 0.02	10
Decision Tree	0.79 ± 0.00	0.74 ± 0.00	0.83 ± 0.01	0.74 ± 0.01	10

Churn: 4,003; No Churn: 4,122

The churn dataset (3.12) exhibits relatively subdued performance compared to the other datasets, with F1 scores and ROC-AUCs typically ranging between 0.73 and 0.83. This moderate effectiveness implies that while ADASYN helps alleviate class imbalance for churn prediction, the intricate complexities or less discernible patterns in customer churn data may restrict the improvement of predictive performance.

These assessment highlights ADASYN's diverse impact across various datasets, demonstrating significant enhancements in datasets such as Fraud Detection and Breast Cancer, but a less pronounced influence in Heart Attack and Churn datasets. These disparities emphasize the necessity of employing dataset-specific strategies when utilizing synthetic oversampling techniques. ADASYN's performance emphasizes the importance of selecting tailored models and conducting rigorous validation processes to optimize outcomes in the presence of class imbalance. The method's effectiveness and the resulting model stability are evidently influenced by the interaction between the dataset features, the model employed, and the characteristics of the synthetic samples generated by ADASYN.

3.4 Nearmiss-1 Analysis

The analysis of Nearmiss-1 across four distinct datasets, namely Heart Attack, Fraud Detection, Breast Cancer, and Churn, unveils the intricate influence of this undersampling technique on the performance of predictive models. It is noteworthy that each dataset exhibits a unique response to Nearmiss-1, highlighting the inherent complexities and specificities of each dataset.

For the Heart Attack dataset (3.13) below, performance is moderate to low, evidencing lower F1 scores and ROC-AUC values than observed in other datasets. Notably, LASSO and Ridge Regression attain the highest scores, suggesting that some models can still effectively capture essential patterns despite significant undersampling. However, the overall lower performance across most techniques may indicate the loss of crucial information, which can be particularly detrimental in datasets where minor class signals are indispensable for accurate predictions.

Table 3.13: Nearmiss-1 Results for Heart Attack Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.56 ± 0.00	0.64 ± 0.00	0.56 ± 0.03	0.66 ± 0.01	10
Information Gain	0.45 ± 0.01	0.66 ± 0.01	0.49 ± 0.04	0.64 ± 0.01	10
ANOVA F-value	0.53 ± 0.01	0.65 ± 0.01	0.54 ± 0.02	0.66 ± 0.00	10
Wrapper Methods					
RFE	0.43 ± 0.02	0.64 ± 0.01	0.44 ± 0.02	0.64 ± 0.01	10
Logistic Regression	0.61 ± 0.03	0.63 ± 0.01	0.63 ± 0.03	0.62 ± 0.04	10
Random Forest	0.38 ± 0.01	0.68 ± 0.01	0.40 ± 0.12	0.56 ± 0.13	10
Embedded Methods					
LASSO Regression	0.67 ± 0.01	0.75 ± 0.01	0.67 ± 0.02	0.75 ± 0.01	10
Ridge Regression	0.67 ± 0.02	0.74 ± 0.02	0.36 ± 0.02	0.62 ± 0.01	10
Decision Tree	0.40 ± 0.01	0.64 ± 0.01	0.43 ± 0.01	0.63 ± 0.01	10

Heart Attack: 9,407; No Heart Attack: 9,407

The Fraud Detection dataset (3.14) presents an interesting contrast to the previous datasets. In this case, certain techniques have achieved remarkable results, particularly in ANN models, which have achieved perfect or near-perfect scores. This suggests a high degree of class separability, even after the aggressive reduction of classes. However, the substantial variability in performance, particularly in standard deviations, suggests that some models may be unstable or overfitting. This

Table 3.14: Nearmiss-1 Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.68 ± 0.09	0.92 ± 0.01	1.00 ± 0.11	0.96 ± 0.01	10
Information Gain	0.16 ± 0.17	0.87 ± 0.05	0.86 ± 0.08	0.93 ± 0.01	10
ANOVA F-value	0.15 ± 0.13	0.72 ± 0.36	0.61 ± 0.29	0.87 ± 0.11	10
Wrapper Methods					
RFE	0.17 ± 0.15	0.89 ± 0.04	0.42 ± 0.31	0.85 ± 0.04	10
Logistic Regression	0.26 ± 0.22	0.89 ± 0.04	0.73 ± 0.09	0.88 ± 0.05	10
Random Forest	0.26 ± 0.22	0.89 ± 0.04	0.71 ± 0.11	0.83 ± 0.11	10
Embedded Methods					
LASSO Regression	0.26 ± 0.22	0.89 ± 0.04	0.70 ± 0.13	0.87 ± 0.05	10
Ridge Regression	0.48 ± 0.34	0.93 ± 0.04	0.60 ± 0.34	0.93 ± 0.05	10
Decision Tree	0.41 ± 0.34	0.89 ± 0.05	0.78 ± 0.18	0.91 ± 0.06	10

Fraud: 324; No Fraud: 324

may be due to the dramatic reduction in majority class examples, which could create an unstable training environment for some models.

Table 3.15: Nearmiss-1 Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	1.00 ± 0.00	10
Information Gain	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
ANOVA F-value	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Wrapper Methods					
RFE	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Logistic Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Random Forest	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Embedded Methods					
LASSO Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Ridge Regression	0.54 ± 0.00	0.64 ± 0.02	0.54 ± 0.00	0.64 ± 0.02	10
Decision Tree	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10

Healed:42,588; Not Healed: 42,588

The Breast Cancer dataset (3.15)has demonstrated exceptional performance metrics across all methods, with the majority of techniques achieving perfect or near-perfect scores. This indicates that the dataset's features are sufficiently robust, as even significant undersampling does not impair the models' ability to classify effectively.

Hence, this consistent high performance may potentially conceal overfitting issues, particularly in real-world applications where class distributions may not be as optimal.

Table 3.16: Nearmiss-1 Results for Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.56 ± 0.01	0.54 ± 0.01	0.64 ± 0.04	0.57 ± 0.03	10
Information Gain	0.61 ± 0.02	0.57 ± 0.01	0.66 ± 0.03	0.61 ± 0.03	10
ANOVA F-value	0.56 ± 0.01	0.55 ± 0.01	0.59 ± 0.03	0.55 ± 0.03	10
Wrapper Methods					
RFE	0.63 ± 0.02	0.59 ± 0.02	0.63 ± 0.02	0.59 ± 0.03	10
Logistic Regression	0.62 ± 0.01	0.58 ± 0.01	0.69 ± 0.02	0.64 ± 0.03	10
Random Forest	0.67 ± 0.02	0.62 ± 0.01	0.73 ± 0.02	0.67 ± 0.03	10
Embedded Methods					
LASSO Regression	0.62 ± 0.01	0.57 ± 0.00	0.67 ± 0.02	0.62 ± 0.01	10
Ridge Regression	0.62 ± 0.01	0.59 ± 0.01	0.69 ± 0.02	0.63 ± 0.02	10
Decision Tree	0.67 ± 0.01	0.62 ± 0.01	0.73 ± 0.02	0.65 ± 0.03	10

Churn: 1,300; No Churn: 1,300

The Churn dataset (3.16)has achieved moderate outcomes, with the Random Forest model demonstrating the most exceptional performance. This indicates that certain algorithms might be more resistant to the data loss caused by Nearmiss-1. Nevertheless, the overall performance remains modest, highlighting the difficulties in employing Nearmiss-1 in datasets where intricate or delicate feature interactions are essential for precise forecasts. The results indicate that Nearmiss-1 can effectively identify important features in datasets with distinct class distinctions, but its effectiveness varies significantly across various scenarios. The technique's impact is heavily influenced by specific dataset characteristics, such as the number of features, the nature of the imbalance, and the complexity of feature relationships. Additionally, the performance variability across different models within the same dataset suggests that some models may be more sensitive to training data reduction, leading to potential issues of overfitting or underfitting. Thus, applying Nearmiss-1 requires careful consideration of both dataset characteristics and model selection to optimize performance in handling imbalanced datasets, ensuring that the reduction in sample size does not compromise the model's ability to generalize to new data.

3.5 Nearmiss-3 Analysis

The analysis of Nearmiss-3 across the four datasets—Heart Attack, Fraud Detection, Breast Cancer, and Churn—reveals the intricate nature and varying effectiveness of this undersampling technique in various analytical contexts. Each dataset exhibits unique responses to Nearmiss-3, which emphasizes the impact of data characteristics on the performance of undersampling techniques.

Table 3.17: Nearmiss-3 Results for Heart Attack Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.68 ± 0.03	0.73 ± 0.01	0.72 ± 0.04	0.77 ± 0.02	10
Information Gain	0.67 ± 0.01	0.68 ± 0.01	0.68 ± 0.05	0.74 ± 0.02	10
ANOVA F-value	0.67 ± 0.01	0.73 ± 0.01	0.69 ± 0.04	0.75 ± 0.02	10
Wrapper Methods					
RFE	0.65 ± 0.01	0.68 ± 0.01	0.65 ± 0.04	0.71 ± 0.02	10
Logistic Regression	**0.86 ± 0.03**	**0.86 ± 0.01**	**0.87 ± 0.03**	**0.85 ± 0.03**	10
Random Forest	0.63 ± 0.01	0.69 ± 0.01	0.61 ± 0.04	0.66 ± 0.02	10
Embedded Methods					
LASSO Regression	0.83 ± 0.03	0.82 ± 0.01	0.84 ± 0.04	0.84 ± 0.03	10
Ridge Regression	**0.86 ± 0.01**	**0.84 ± 0.01**	0.63 ± 0.03	0.67 ± 0.01	10
Decision Tree	0.58 ± 0.01	0.61 ± 0.01	0.60 ± 0.02	0.63 ± 0.02	10

Heart Attack: 9,407; No Heart Attack: 9,407

Nearmiss-3 produces moderate results in the Heart Attack dataset (3.17), with F1 scores ranging from 0.58 to 0.86 and ROC-AUC scores ranging from 0.61 to 0.86. Notably, Logistic Regression and Ridge Regression demonstrate greater resilience, achieving the highest scores, which indicates their robustness in the face of reduced majority class instances. This implies that while Nearmiss-3 simplifies the class structure, certain models are better equipped to utilize the remaining information for accurate predictions.

Table 3.18: Nearmiss-3 Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	0.99 ± 0.00	0.92 ± 0.01	1.00 ± 0.00	0.96 ± 0.02	10
Information Gain	1.00 ± 0.00	0.93 ± 0.01	1.00 ± 0.00	0.95 ± 0.02	10
ANOVA F-value	1.00 ± 0.00	0.93 ± 0.01	1.00 ± 0.00	0.95 ± 0.01	10
Wrapper Methods					
RFE	1.00 ± 0.00	0.93 ± 0.02	1.00 ± 0.00	0.94 ± 0.02	10
Logistic Regression	0.99 ± 0.00	0.94 ± 0.01	0.99 ± 0.01	0.94 ± 0.01	10
Random Forest	0.99 ± 0.00	0.94 ± 0.01	0.99 ± 0.00	0.94 ± 0.01	10
Embedded Methods					
LASSO Regression	0.99 ± 0.01	0.93 ± 0.02	0.99 ± 0.00	0.94 ± 0.01	10
Ridge Regression	0.98 ± 0.00	0.92 ± 0.02	0.99 ± 0.01	0.93 ± 0.00	10
Decision Tree	0.86 ± 0.30	0.92 ± 0.02	0.95 ± 0.07	0.93 ± 0.03	10

Fraud: 324; No Fraud: 291

In contrast, the Fraud Detection dataset(3.18) exhibits exceptional performance across all techniques, with F1 scores and ROC-AUCs approaching or achieving perfect scores. This suggests that the Fraud Detection dataset possesses intrinsic properties, such as clear separability and minimal noise, which enable even a reduced sample set to effectively represent the underlying patterns necessary for high model accuracy.

Table 3.19: Nearmiss-3 Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
Chi-Squared Test	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.02	1.00 ± 0.00	10
Information Gain	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	10
ANOVA F-value	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	1.00 ± 0.00	10
Wrapper Methods					
RFE	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Logistic Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10
Random Forest	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	10
Embedded Methods					
LASSO Regression	0.69 ± 0.24	0.84 ± 0.22	0.69 ± 0.24	0.84 ± 0.22	10
Ridge Regression	0.22 ± 0.00	0.60 ± 0.01	0.22 ± 0.00	0.60 ± 0.01	10
Decision Tree	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	10

Healed:42,588; Not Healed: 1,483

The Breast Cancer dataset (3.19)displays impressive results with Nearmiss-3, achieving high scores across various methods. The consistency in outstanding performance suggests that critical classification information is preserved despite aggressive undersampling. However, Ridge Regression stands out as a notable exception, underperforming significantly, which may indicate suboptimal interactions between this method and the data processing technique.

Table 3.20: Nearmiss-3 Results for Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Filter Methods					
0.72 ± 0.01	0.72 ± 0.01	0.79 ± 0.02	0.76 ± 0.01	10	
Information Gain	0.72 ± 0.01	0.72 ± 0.01	0.80 ± 0.02	0.76 ± 0.02	10
ANOVA F-value	0.56 ± 0.01	0.55 ± 0.01	0.60 ± 0.03	0.56 ± 0.02	10
Wrapper Methods					
RFE	0.63 ± 0.02	0.59 ± 0.02	0.62 ± 0.02	0.60 ± 0.03	10
Logistic Regression	0.62 ± 0.01	0.58 ± 0.01	0.68 ± 0.02	0.61 ± 0.01	10
Random Forest	0.67 ± 0.02	0.62 ± 0.01	0.72 ± 0.03	0.64 ± 0.03	10
Embedded Methods					
LASSO Regression	0.62 ± 0.01	0.57 ± 0.00	0.68 ± 0.03	0.62 ± 0.02	10
Ridge Regression	0.62 ± 0.01	0.59 ± 0.01	0.69 ± 0.02	0.63 ± 0.01	10
Decision Tree	0.67 ± 0.01	0.62 ± 0.01	0.73 ± 0.01	0.66 ± 0.02	10

Churn: 1,300; No Churn: 1,300

In contrast, the Churn dataset (3.20)shows more modest outcomes, with F1 scores and ROC-AUCs mostly falling between 0.60 and 0.73. This dataset likely contains intricate patterns that are somewhat disrupted by the substantial data reduction caused by Nearmiss-3. As a result, models do not perform as well here as in other datasets, indicating heightened sensitivity to reduced training instances. Across the datasets, Nearmiss-3 tends to excel in situations where class boundaries are clear and the necessary information for classification is resilient to data point loss. In datasets like Heart Attack and Churn, where subtle nuances and intricate relationships define the predictive patterns, the reduction in data can lead to a substantial decline in model performance.

This underscores the essential balance required in choosing undersampling techniques like Nearmiss-3, which must be tailored to the dataset's characteristics to prevent the loss of crucial information while still effectively addressing class imbalance.

3.6 Hybrid - SMOTE Analysis

The Hybrid-SMOTE technique, which integrates SMOTE with various feature selection methods, has been evaluated across four datasets—Heart Attack, Fraud Detection, Breast Cancer, and Churn. This approach aims to enhance model performance by addressing class imbalance and optimizing the feature space to improve prediction accuracy.

Table 3.21: Hybrid-SMOTE Results for Heart Attack Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE+ LASSO Regression	0.82 ± 0.00	0.83 ± 0.00	0.82 ± 0.00	0.84 ± 0.00	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.67 ± 0.00	0.78 ± 0.01	0.67 ± 0.00	0.78 ± 0.01	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.85 ± 0.01	0.82 ± 0.01	0.85 ± 0.01	0.83 ± 0.01	5

Heart Attack: 162,778; No Heart Attack: 162,778

The hybrid-SMOTE outcomes for the heart attack dataset (**??**)display varying performance based on the employed techniques. The approach that incorporates Chi-Squared Test, RFE, and LASSO Regression attains an F1 score and ROC-AUC of approximately 0.82, demonstrating a reasonable level of effectiveness. Conversely, the method that merges ANOVA F-value with Random Forest and Decision Tree achieves an F1 score and ROC-AUC of roughly 0.85, slightly surpassing the other techniques. In contrast, the combination of Information Gain with Logistic Regression and Ridge Regression exhibits reduced effectiveness, with scores around 0.67.

Table 3.22: Hybrid-SMOTE Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.99 ± 0.00	0.96 ± 0.02	0.99 ± 0.39	0.95 ± 0.02	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.97 ± 0.01	0.97 ± 0.02	0.97 ± 0.44	0.98 ± 0.01	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.99 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	5

Fraud: 198,274; No Fraud: 198,274

The Fraud Detection Dataset (Table 3.22) demonstrates an extraordinary responsiveness to Hybrid-SMOTE, particularly when it combines the Chi-Squared Test with RFE and LASSO Regression. This combination achieves an impressive level of accuracy with an F1 score of 0.99. Similarly, the method involving ANOVA F-value, Random Forest, and Decision Tree exhibits comparable high performance. The Information Gain, when combined with Logistic Regression and Ridge Regression, also performs well, although the performance of ANN is somewhat more variable. The performance of the Hybrid-SMOTE technique on the Breast Cancer Dataset in table 3.24 below is impressive, as it achieves high scores in most combinations due to the dataset's clear class separability. The Chi-Squared Test combined with RFE and LASSO Regression, and ANOVA F-value with Random Forest and Decision Tree achieve perfect scores across both models. However, the Information Gain with Logistic Regression and Ridge Regression method shows a significant drop in performance, suggesting a misalignment of the method with the dataset's characteristics.

Table 3.23: Hybrid-SMOTE Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	1 ± 0.00	1 ± 0.00	1 ± 0.00	1 ± 0.00	1
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.52 ± 0.09	0.61 ± 0.03	0.52 ± 0.09	0.61 ± 0.03	3
Method 3 ANOVA F-value + Random Forest + Decision Tree	1 ± 0.00	1 ± 0.00	0.99 ± 0.00	1 ± 0.00	5

Healed: 66,693; Not Healed: 66,693

Table 3.24: Hybrid-SMOTE Results for Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.76 ± 0.01	0.79 ± 0.01	0.75 ± 0.01	0.83 ± 0.01	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.73 ± 0.00	0.77 ± 0.01	0.76 ± 0.01	0.83 ± 0.01	5
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.73 ± 0.00	0.77 ± 0.01	0.76 ± 0.01	0.83 ± 0.01	5

Churn: 4,122; No Churn: 4122

In the Churn dataset, the Hybrid-SMOTE method demonstrates moderate performance improvements, with F1 scores and ROC-AUC ranging from 0.73 to 0.76. Among the techniques used, the combination of the Chi-Squared Test with RFE and LASSO Regression slightly outperforms the others, indicating better synergy between these techniques in handling the dataset's complexities. The Hybrid-SMOTE method highlights the significance of selecting the appropriate combination of feature selection and oversampling techniques. It is apparent that specific combinations, such as

Chi-Squared Test with RFE and LASSO Regression, consistently perform well on various datasets. Each dataset demonstrates a unique response to different hybrid combinations, indicating that the effectiveness of Hybrid-SMOTE heavily relies on the nature of the dataset and the compatibility of the feature selection method with the oversampling technique. There is a noticeable variability in model performance, particularly in ANN models, suggesting that while Hybrid-SMOTE can enhance performance, it may also introduce instability depending on the dataset and the specific models used. Overall, Hybrid-SMOTE presents a promising approach to addressing class imbalance by integrating robust feature selection with SMOTE, designed to significantly improve model performance across diverse datasets. However, careful consideration should be given to the choice of feature selection and oversampling methods to ensure optimal outcomes.

3.7 Hybrid - ADASYN Analysis

The Hybrid-ADASYN technique, which involves integrating ADASYN with different feature selection methods, has been evaluated across four datasets—Heart Attack, Fraud Detection, Breast Cancer, and Churn. This combined approach aims to not only correct class imbalances but also to refine the feature set in order to improve predictive accuracy.

Table 3.25: Hybrid-ADASYN Results for Heart Attack Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.82 ± 0.00	0.83 ± 0.00	0.82 ± 0.00	0.84 ± 0.00	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.67 ± 0.00	0.77 ± 0.00	0.67 ± 0.00	0.78 ± 0.00	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.84 ± 0.01	0.81 ± 0.01	0.83 ± 0.02	0.81 ± 0.01	5

Heart Attack: 164,902; No Heart Attack: 162,778

In the Heart Attack dataset(3.25), the hybrid approach demonstrates a noteworthy level of effectiveness, as evidenced by the Chi-Squared Test combined with RFE and LASSO Regression, which achieves F1 scores and ROC-AUC values in the range of 0.82 to 0.84. On the other hand, the combination of Information Gain, Logistic Regression, and Ridge Regression exhibits somewhat reduced performance, potentially due to mismatches between feature selection methods and modeling techniques. The integration of ANOVA F-value with Random Forest and Decision Tree yields competitive results, marginally surpassing the second method.

Table 3.26: Hybrid-ADASYN Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.98 ± 0.02	0.96 ± 0.02	0.96 ± 0.04	0.96 ± 0.01	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.94 ± 0.06	0.94 ± 0.03	0.95 ± 0.01	0.96 ± 0.02	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.99 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	5

Fraud: 198,330; No Fraud: 198,274

In the Fraud Detection Dataset, the Hybrid-ADASYN approach demonstrates exceptional performance, particularly when the Chi-Squared Test is combined with RFE and LASSO Regression, achieving near-perfect scores. The robustness of the dataset to ADASYN is evident, as even the less effective combinations still achieve high metrics. The method that combines ANOVA F-value with Random Forest and Decision Tree exhibits the highest stability and effectiveness.

The Breast Cancer Dataset exhibits exceptional performance, with the highest scores achieved by methods that combine Chi-Squared Test and ANOVA F-value with Random Forest. In contrast, the method combining Information Gain with Logistic Regression and Ridge Regression significantly underperforms other methods, which may indicate a particular sensitivity of this dataset to

Table 3.27: Hybrid-ADASYN Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0	1.00 ± 0.00	2
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.47 ± 0.01	0.59 ± 0.01	0.47 ± 0.02	0.59 ± 0.02	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	1.00 ± 0.00	1.00 ± 0	1.00 ± 0.00	1.00 ± 0.00	5

Healed: 66,692; Not Healed: 66,698

the selected features and synthetic sampling applied.

Table 3.28: Hybrid-ADASYN Results for Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.76 ± 0.01	0.79 ± 0.01	0.75 ± 0.01	0.83 ± 0.01	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.73 ± 0.01	0.77 ± 0.01	0.73 ± 0.01	0.82 ± 0.01	5
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.74 ± 0.02	0.77 ± 0.01	0.72 ± 0.01	0.83 ± 0.01	5

Churn: 4,003; No Churn: 4,122

In the Churn dataset, performance is moderate, with the best results obtained from the method that combines Chi-Squared Test with RFE and LASSO Regression. Although improvements are observed, they are not as substantial as in other datasets, suggesting that the churn dataset may have more intricate or less distinct classes that are less responsive to hybrid methods.

The success of the Hybrid-ADASYN method in these datasets emphasizes the importance of se-

lecting appropriate feature selection techniques and effective oversampling strategies. The performance depends significantly on how well these methods align with the characteristics of each dataset. The variation in method effectiveness across datasets highlights that the nature of the dataset, including the type of features and the degree of class imbalance, heavily influences the usefulness of hybrid techniques. Datasets with clearer class definitions and less noise, such as Breast Cancer and Fraud Detection, respond better to these methods. The results also underscore the susceptibility of different models to the changes in feature space and class distribution introduced by Hybrid-ADASYN. Some combinations lead to high variability in performance, especially in datasets with more intricate relationships.

3.8 Hybrid - Nearmiss-1 Analysis

The Hybrid-NearMiss-1 method, which combines NearMiss-1 undersampling with a range of feature selection strategies, has been assessed on four datasets—Heart Attack, Fraud Detection, Breast Cancer, and Churn—providing insights into its effectiveness in various contexts. The Heart

Table 3.29: Hybrid-NearMiss-1 Results for Heart Attack Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.55 ± 0.01	0.62 ± 0.00	0.55 ± 0.02	0.63 ± 0.00	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.84 ± 0.05	0.77 ± 0.01	0.84 ± 0.05	0.77 ± 0.01	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.53 ± 0.02	0.63 ± 0.00	0.55 ± 0.02	0.64 ± 0.01	5

Heart Attack: 9,407; No Heart Attack: 9,407

Attack Dataset exhibits a combination of outcomes with the hybrid approach. The most effective configuration features Information Gain in conjunction with Logistic Regression and Ridge Regression, achieving F1 and ROC-AUC scores of approximately 0.84, which attests to its efficacy

in selecting the most predictive features even with a limited amount of data. This configuration employs only 2 features, emphasizing the method's capacity to distill the most pertinent information from a potentially complex dataset.

Table 3.30: Hybrid-NearMiss-1 Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.28 ± 0.24	0.86 ± 0.07	0.77 ± 0.32	0.87 ± 0.10	4
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.73 ± 0.29	0.93 ± 0.05	0.92 ± 0.04	0.94 ± 0.01	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.96 ± 0.01	0.99 ± 0.00	0.94 ± 0.02	0.97 ± 0.00	5

Fraud: 324; No Fraud: 324

In the case of the Fraud Detection Dataset, the performance varies considerably, ranging from low to exceptional scores. The most notable method combines ANOVA F-value with Random Forest and Decision Tree, yielding nearly perfect scores. This suggests that even with a reduced sample size, the selected features are sufficient to capture the essential characteristics of the data. The models generally incorporate 4 to 5 features, indicating a balanced approach to preserving essential information while addressing class imbalance.

The Breast Cancer Dataset has shown exceptional responsiveness to all hybrid configurations, with various methods achieving perfect scores in multiple areas. These methods often utilize a minimal number of features, as few as one, which highlights the dataset's strong predictive signals and the effectiveness of isolating them even with a significant reduction in data volume.

The Churn Dataset's performance is generally moderate, with the best results coming from a configuration that incorporates Information Gain with Logistic Regression and Ridge Regression. This method outperforms others slightly, indicating that it more effectively aligns feature selection

Table 3.31: Hybrid-NearMiss-1 Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.63 ± 0.02	0.67 ± 0.02	0.63 ± 0.02	0.67 ± 0.02	3
Method 3 ANOVA F-value + Random Forest + Decision Tree	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	5

Healed:42,588; Not Healed: 42588

Table 3.32: Hybrid-NearMiss-1 Results for Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.60 ± 0.04	0.60 ± 0.04	0.58 ± 0.06	0.62 ± 0.02	4
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.61 ± 0.01	0.65 ± 0.01	0.69 ± 0.03	0.76 ± 0.02	5
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.57 ± 0.02	0.56 ± 0.02	0.54 ± 0.04	0.63 ± 0.05	5

Churn: 1,300; No Churn: 1,300

with the challenges of predicting churn. Typically, 4 to 5 features are used, suggesting the need to capture a broader range of data characteristics to manage the inherent complexity and variability of churn predictors.

The Hybrid-NearMiss-1 technique emphasizes the crucial importance of matching the appropriate feature selection methods with suitable undersampling techniques to optimize predictive performance. The success of these hybrid methods varies significantly depending on the dataset's charac-

teristics, which highlights the need for a nuanced, tailored approach when dealing with imbalanced data across different domains. The number of features selected plays a pivotal role, with the optimal count varying depending on the dataset and the specific interaction of features and response variables.

3.9 Hybrid - Nearmiss-3 Analysis

The Hybrid-NearMiss-3 method, which integrates NearMiss-3 undersampling and selective feature selection, exhibits varying levels of efficacy across datasets like Heart Attack, Fraud Detection, Breast Cancer, and Churn. The distinctive characteristics of each dataset contribute to the outcomes, highlighting the importance of customized hybrid approaches.

The Heart Attack Dataset below (3.33) demonstrates superior results with Information Gain, Logistic Regression, and Ridge Regression, achieving an F1 score and ROC-AUC of 0.94. This approach effectively utilizes only one feature, showcasing its ability to identify crucial predictive attributes while significantly reducing data. Other methods display moderate performance, with F1 scores and ROC-AUCs ranging from 0.62 to 0.65, indicating a variable compatibility between the undersampling method and the remaining features.

Table 3.33: Hybrid-NearMiss-3 Results for Heart Attack Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.63 ± 0.00	0.65 ± 0.00	0.64 ± 0.01	0.70 ± 0.01	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.94 ± 0.00	0.74 ± 0.00	0.94 ± 0.00	0.74 ± 0.00	1
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.62 ± 0.02	0.71 ± 0.01	0.62 ± 0.02	0.71 ± 0.01	5

Heart Attack: 9,407; No Heart Attack: 9,407

Table 3.34: Hybrid-NearMiss-3 Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.98 ± 0.01	0.92 ± 0.01	0.99 ± 0.01	0.93 ± 0.02	4
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.76 ± 0.19	0.78 ± 0.15	0.59 ± 0.34	0.69 ± 0.205	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.85 ± 0.02	0.88 ± 0.02	0.80 ± 0.03	0.85 ± 0.02	5

Fraud: 324; No Fraud: 291

The Fraud Detection Dataset exhibits considerable performance variability. The Chi-Squared Test combined with RFE and LASSO Regression yields nearly perfect scores, while other methods, particularly Information Gain with Logistic Regression and Ridge Regression, demonstrate more inconsistent results. This suggests that while some feature combinations maintain predictive strength after undersampling, others may lack sufficient complexity, leading to unreliable performances.

Table 3.35: Hybrid-NearMiss-3 Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	3
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.22 ± 0.00	0.59 ± 0.05	0.22 ± 0.00	0.59 ± 0.05	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	1

Healed:42,588; Not Healed: 1,483

The Breast Cancer Dataset shows that almost all methods perform exceptionally well, with many achieving perfect or nearly perfect metrics. However, the combination of Information Gain with Logistic Regression and Ridge Regression significantly underperforms, highlighting potential issues with overly simplistic feature selection in complex datasets.

Table 3.36: Hybrid-NearMiss-3 Results Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.68 ± 0.01	0.69 ± 0.01	0.76 ± 0.01	0.80 ± 0.01	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.69 ± 0.01	0.71 ± 0.01	0.76 ± 0.01	0.80 ± 0.01	5
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.69 ± 0.01	0.69 ± 0.01	0.76 ± 0.02	0.80 ± 0.02	5

Churn: 1,300; No Churn: 1,300

According to the Churn Dataset, the hybrid methods demonstrate fairly uniform performance, with F1 scores and ROC-AUCs typically falling within the range of 0.68 to 0.76. This uniformity suggests that while NearMiss-3 may provide valuable insights into churn behaviors, it does not significantly improve predictive accuracy due to the intricate nature of churn patterns. The Hybrid-NearMiss-3 method highlights the importance of selecting the appropriate feature selection and undersampling techniques for each dataset. It demonstrates that while some datasets can be effectively modeled with a reduced feature set, others require a more nuanced approach to maintain predictive accuracy.

The chosen number of features, ranging from 1 to 5 in these studies, plays a crucial role in balancing the reduction in data volume with the need to capture essential information for accurate model predictions. This methodological approach offers a strategic way to address class imbalances and refine input data, although its success varies significantly depending on the specific characteristics and complexities of each dataset.

3.10 SMOTE-TOMEK Analysis

The SMOTE-TOMEK hybrid approach, which seeks to address the challenges of imbalanced datasets by integrating Synthetic Minority Over-sampling Technique (SMOTE) with Tomek links for the purpose of cleaning overlapping samples, has been employed on datasets pertaining to Heart Attack, Fraud Detection, Breast Cancer, and Churn. The application of this approach to these datasets has yielded varying levels of improvement, demonstrating the nuanced impact of this method on disparate types of data.

Table 3.37: SMOTE-TOMEK Results for Heart Attack Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.82 ± 0.00	0.83 ± 0.01	0.83 ± 0.00	0.84 ± 0.00	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.67 ± 0.00	0.77 ± 0.00	0.67 ± 0.00	0.77 ± 0.00	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.86 ± 0.01	0.83 ± 0.00	0.85 ± 0.01	0.83 ± 0.00	5

Heart Attack: 162,808; No Heart Attack: 162,808

For the Heart Attack Dataset, the SMOTE-TOMEK method demonstrates considerable success, particularly when combined with ANOVA F-value and Random Forest, achieving an F1 score of 0.86 and an ROC-AUC of 0.83. This suggests strong compatibility between the selected features and the model's ability to generalize from the oversampled and cleaned data. Additionally, Method

1, which utilizes Chi-Squared Test combined with RFE and LASSO Regression, also performs well, indicating robustness across different feature selection techniques.

Table 3.38: SMOTE-TOMEK Results for Fraud Detection Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.99 ± 0.00	0.96 ± 0.02	0.99 ± 0.00	0.96 ± 0.02	4
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.97 ± 0.01	0.97 ± 0.02	0.96 ± 0.05	0.98 ± 0.01	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.99 ± 0.02	0.99 ± 0.01	0.97 ± 0.02	0.99 ± 0.02	5

Fraud: 198,284; No Fraud: 198,284

For the Fraud Detection Dataset, almost all configurations achieve near-perfect scores, with the standout configuration involving ANOVA F-value and Random Forest, highlighting the effectiveness of SMOTE-TOMEK in handling highly imbalanced fraud data when combined with powerful feature selection and ensemble methods.

Table 3.39: SMOTE-TOMEK Results for Breast Cancer Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	4
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.63 ± 0.02	0.66 ± 0.02	0.63 ± 0.02	0.67 ± 0.01	2
Method 3 ANOVA F-value + Random Forest + Decision Tree	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	5

Healed:66,637; Not Healed: 66,637

For the Breast Cancer Dataset, methods integrating Chi-Squared Test with RFE and LASSO Regression, and ANOVA F-value with Random Forest, achieve a perfect score. This reflects the dataset's responsiveness to oversampling in conjunction with effective noise reduction through Tomek links. However, the Information Gain with Logistic Regression and Ridge Regression method underperforms compared to others, possibly indicating an oversimplification or loss of crucial information despite the hybrid technique.

Table 3.40: SMOTE-TOMEK Results for Churn Dataset

	RF		ANN		
Technique	**Overall F1 Score**	**ROC-AUC**	**Overall F1 Score**	**ROC-AUC**	**Features**
Method 1 Chi-Squared Test + RFE + LASSO Regression	0.79 ± 0.01	0.76 ± 0.01	0.83 ± 0.01	0.76 ± 0.02	5
Method 2 Information Gain + Logistic Regression + Ridge Regression	0.78 ± 0.01	0.74 ± 0.01	0.83 ± 0.01	0.76 ± 0.01	5
Method 3 ANOVA F-value + Random Forest + Decision Tree	0.78 ± 0.01	0.74 ± 0.01	0.83 ± 0.01	0.75 ± 0.02	5

Churn: 3,429; No Churn: 3,429

The results achieved on the Churn Dataset can be considered moderately successful, with F1 and ROC-AUC scores hovering around 0.78 to 0.83. The SMOTE-TOMEK method proves to be particularly effective in enhancing the classifier's ability to predict churn when the class distribution is skewed. In this case, five features are consistently used, striking a balance between maintaining sufficient information for prediction accuracy and managing the increased data volume resulting from SMOTE.

Across various datasets, SMOTE-TOMEK emerges as a valuable strategy for improving classification performance in imbalanced situations. The selection of features, ranging from one to five, plays a crucial role in determining the efficacy of the approach, highlighting the importance of integrating feature selection with oversampling and cleaning techniques. This approach underscores

the significance of customized solutions in data science, where the choice and configuration of methods must closely align with the specific data characteristics and desired outcomes.

CHAPTER IV

CONCLUSION AND DISCUSSION

The study examines in depth the relationship between various feature selection techniques and class rebalancing strategies, with an aim to enhance the performance of Random Forest (RF) and Artificial Neural Network (ANN) models. This examination directly supports the research's goals, yielding vital insights into how these methods affect model accuracy and dependability across different datasets. The study's results indicate that the effectiveness of combining feature selection methods like Chi-Squared Test, Information Gain, and ANOVA F-value with class rebalance techniques such as SMOTE, ADASYN, and NearMiss varies considerably across different datasets. For instance, the NearMiss technique typically performed poorly due to potential data loss, while methods like SMOTE-TOMEK and Hybrid-ADASYN frequently improved both accuracy and ROC-AUC scores. This variability highlights the importance of selecting feature selection and rebalancing methods based on the dataset's specific characteristics and the severity of its class imbalance. The study's analysis suggests that there is no single combination of feature selection and class rebalancing methods that consistently perform best across all datasets or prediction scenarios. For instance, the Breast Cancer dataset showed strong performance with nearly all combinations, indicating robust underlying predictive signals that withstand various data manipulations. In contrast, the more complex pattern recognition required for the Churn dataset necessitated more nuanced method configurations, demonstrating moderate performance improvements. The study's results emphasize the importance of selectively choosing feature selection and rebalancing techniques based on the dataset's specific features and the severity of its class imbalance. The findings indicate that no single approach is suitable for all situations, and practitioners must carefully evaluate and configure these methods to achieve optimal results. This study proposes innovative

hybrid approaches that combine multiple feature selection and class-rebalancing techniques. These methods aim to enhance the quality of the dataset by utilizing the strengths of each technique. For example, Hybrid-SMOTE methods improve data balance and reduce noise through Tomek links, thereby enhancing the quality of the training set and, subsequently, the performance of the model. This approach is particularly effective, demonstrating how strategic combinations can overcome the limitations of individual methods. The varied impacts on RF and ANN models highlight that while RF may respond acutely to feature selection due to its mechanism of random subspace selection, ANNs often benefit more significantly from balanced datasets provided by sophisticated oversampling methods. The differential responses underscore the need for tailored strategies in predictive modeling, contingent upon the model type and specific data dynamics at play. This study achieves its objectives by emphasizing how the strategic integration of feature selection with class rebalancing can profoundly affect machine learning outcomes. The findings advocate for a context-driven approach in method selection, promoting the use of hybrid techniques as a forward-thinking solution for addressing class imbalance in predictive modeling. These insights encourage practitioners to consider both the intrinsic data characteristics and the specific requirements of their applications to maximize the efficacy and performance of their models.

4.1 Study Limitations and Further Research

This groundbreaking research, which demonstrates the intricate use of various data-driven techniques for feature selection and class rebalancing, has limitations that warrant further exploration in future studies:

- The study was hindered by the restricted size of the GPU, which may have confined the intricacy and scalability of the models. Although the study managed to complete a significant number of iterations (4,000), the utilization of more potent computational tools such as a supercomputer could potentially improve the model's capacity to learn from a more extensive and intricate characteristic set.

  This could facilitate more robust training, particularly when applying computationally de-

manding techniques like deep learning in ANNs or extensive hyperparameter adjustment in RF models.

- The constraint of 4,000 iterations might have limited the extent of learning, especially in situations involving artificial neural networks or intricate ensemble techniques that profit from lengthier training sessions. Elongated iterations may aid in attaining a more robust convergence of learning algorithms, particularly in hybrid models that incorporate multiple layers of data manipulation, such as SMOTE-TOMEK and Hybrid-ADASYN.

- Although the study used F1 Score and ROC-AUC to assess the model's performance, it would be beneficial for future studies to evaluate the precision and recall for each feature in greater detail. This approach would enable researchers to determine not only the overall performance of the model, but also how well it performs for each class specifically. This is especially important in imbalanced datasets, where the minority class may be of greater interest, and precision and recall can provide valuable insights into the model's performance regarding false positives and false negatives.

- Additional research could investigate the individual influence of each feature when combined with various class rebalancing strategies. This would entail extensive statistical analyses or machine learning feature importance assessments to determine which features are most predictive and how their representation affects overall model accuracy. This approach could be particularly beneficial in refining feature engineering techniques and enhancing model interpretability and performance.

REFERENCES

Acharya, Samik and Sima Das (2024). “Statistical Modeling for Predictive Healthcare Analytics”. In: *Revolutionizing Healthcare Treatment With Sensor Technology*. IGI Global, pp. 265–293.

Agraz, Melih (2023). “Comparison of Feature Selection Methods in Breast Cancer Microarray Data”. In: *Medical Records* 5.2, pp. 284–9.

Al Tawil, Arar et al. (2024). “Predictive modeling for breast cancer based on machine learning algorithms and features selection methods.” In: *International Journal of Electrical & Computer Engineering (2088-8708)* 14.2.

Alahmari, Fahad (2020). “A comparison of resampling techniques for medical data using machine learning”. In: *Journal of Information & Knowledge Management* 19.01, p. 2040016.

Almalkawi, Mohammad and Josh Caron (2021). “Domain decomposition based artificial neural networks (ANNs) modeling of acoustic wave resonators and filters”. In: *2021 IEEE 21st Annual Wireless and Microwave Technology Conference (WAMICON)*. IEEE, pp. 1–4.

Anamisa, Devie Rosa, Fifin Ayu Mufarroha, and Achmad Jauhari (2023). “Feature selection to increase the attractiveness of visitors in Bangkalan tourism, Madura based on chi-square method”. In: *AIP Conference Proceedings*. Vol. 2679. 1. AIP Publishing.

Anusha, Yamijala, R Visalakshi, and Konda Srinivas (2023). “Imbalanced data classification using improved synthetic minority over-sampling technique”. In: *Multiagent and Grid Systems* 19.2, pp. 117–131.

Aouedi, Ons, Kandaraj Piamrat, and Benoît Parrein (2021). “Performance evaluation of feature selection and tree-based algorithms for traffic classification”. In: *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, pp. 1–6.

Arputharaj, Anuprabha, Soma Datta, and Khondker Shajadul Hasan (2019). “Impact of Distance Measures on Imbalanced Classes for Rule Extraction”. In: *2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, pp. 29–34.

Ashraf, Shahzad et al. (2020). “Conversion of adverse data corpus to shrewd output using sampling metrics”. In: *Visual computing for industry, biomedicine, and art* 3, pp. 1–13.

Al-Bahrani, Reda et al. (2021). "SIGRNN: Synthetic Minority Instances Generation in Imbalanced Datasets using a Recurrent Neural Network." In: *ICPRAM*, pp. 349–356.

Banda, Misheck, Ernest Ketcha Ngassam, and Ernest Mnkandla (2022). "Preliminary Experiments on the Performance of Machine Learning Models". In: *2022 IST-Africa Conference (IST-Africa)*, pp. 1–11. URL: https://api.semanticscholar.org/CorpusID:251762173.

Bansal, Ankita et al. (2021). "Analysis of smote: Modified for diverse imbalanced datasets under the IoT environment". In: *International Journal of Information Retrieval Research (IJIRR)* 11.2, pp. 15–37.

Barkah, Azhari Shouni et al. (2023). "Impact of Data Balancing and Feature Selection on Machine Learning-based Network Intrusion Detection". In: *JOIV: International Journal on Informatics Visualization* 7.1, pp. 241–248.

Baykal, Halil and Hatice Kalkan Yildirim (2013). "Application of artificial neural networks (ANNs) in wine technology". In: *Critical reviews in food science and nutrition* 53.5, pp. 415–421.

Benkendorf, Donald J et al. (2023). "Correcting for the effects of class imbalance improves the performance of machine-learning based species distribution models". In: *Ecological Modelling* 483, p. 110414.

Bharti, Kusum Kumari and Pramod kumar Singh (2014). "A survey on filter techniques for feature selection in text mining". In: *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*. Springer, pp. 1545–1559.

BN, VI (2022). "Enhanced machine learning based feature subset through FFS enabled classification for cervical cancer diagnosis". In: *Int J Knowledge-based Intell Eng Syst* 26, pp. 79–89.

Bojarajulu, Balaganesh, Sarvesh Tanwar, and Ajay Rana (2021). "A Synoptic Review on Feature Selection and Machine Learning models used for Detecting Cyber Attacks in IoT". In: *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE, pp. 1–7.

Bouktif, Salah et al. (2018). "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches". In: *Energies* 11.7, p. 1636.

Bruxelles, MLG - Université Libre de (2016). *Credit Card Fraud Detection*. Accessed: 2024-06-24. URL: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

Brzezinski, Dariusz (2020). "Fibonacci and k-Subsecting Recursive Feature Elimination". In: *arXiv preprint arXiv:2007.14920*.

Centers for Disease Control and Prevention (2022). *BRFSS 2022 Codebook*. Behavioral Risk Factor Surveillance System. URL: https://www.cdc.gov/brfss/annual_data/2020/pdf/codebook20_llcp-v2-508.pdf.

Chaibi, Mohamed et al. (2022). "Machine learning models based on random forest feature selection and Bayesian optimization for predicting daily global solar radiation". In: *International Journal of Renewable Energy Development* 11.1, p. 309.

Cheah, Patience Chew Yee, Yue Yang, and Boon Giin Lee (2023). "Enhancing financial fraud detection through addressing class imbalance using hybrid SMOTE-GAN techniques". In: *International Journal of Financial Studies* 11.3, p. 110.

Cheng, Ruihan et al. (2021). "Probability density machine: A new solution of class imbalance learning". In: *Scientific Programming* 2021.1, p. 7555587.

Choi, Hosik, Ja-Yong Koo, and Changyi Park (2015). "Fused least absolute shrinkage and selection operator for credit scoring". In: *Journal of Statistical Computation and Simulation* 85.11, pp. 2135–2147.

Consulting, Merchant Cost (2024). *Credit Card Fraud Statistics (2024)*. Accessed: 2024-06-24. URL: https://merchantcostconsulting.com/lower-credit-card-processing-fees/credit-card-fraud-statistics/.

Davagdorj, Khishigsuren et al. (2020). "A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention". In: *Applied Sciences* 10.9, p. 3307.

Dehghani Champiri, Zohreh, Adeleh Asemi, and Salim Siti Salwah Binti (2019). "Meta-analysis of evaluation methods and metrics used in context-aware scholarly recommender systems". In: *Knowledge and Information Systems* 61, pp. 1147–1178.

Dholakiya, Parth (2020). *SMOTE: Synthetic Minority Over-sampling Technique*. URL: https://medium.com/@parthdholakiya180/smote-synthetic-minority-over-sampling-technique-4d5a5d69d720.

Dube, Lindani and Tanja Verster (2023). "Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models". In: *Data Science in Finance and Economics* 3.4, pp. 354–379.

Elreedy, Dina and Amir F Atiya (2019). "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance". In: *Information Sciences* 505, pp. 32–64.

Elreedy, Dina, Amir F Atiya, and Firuz Kamalov (2023). "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning". In: *Machine Learning*, pp. 1–21.

Fernandes, Everlandio RQ and Andre CPLF de Carvalho (2019). “Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning”. In: *Information Sciences* 494, pp. 141–154.

Habib, Beenish and Farida Khursheed (2022). “Performance evaluation of machine learning models for distributed denial of service attack detection using improved feature selection and hyperparameter optimization techniques”. In: *Concurrency and Computation: Practice and Experience* 34.26, e7299.

Haddadi, Seyed Jamal et al. (2024). “Customer churn prediction in imbalanced datasets with resampling methods: A comparative study”. In: *Expert Systems with Applications* 246, p. 123086.

Hamdi, Monia et al. (2022). “Chicken swarm-based feature subset selection with optimal machine learning enabled data mining approach”. In: *Applied Sciences* 12.13, p. 6787.

Hananya, Rotem and Gilad Katz (2024). “Dynamic selection of machine learning models for time-series data”. In: *Information Sciences* 665, p. 120360.

Hao, Ming, Yanli Wang, and Stephen H Bryant (2014). “An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data”. In: *Analytica chimica acta* 806, pp. 117–127.

He, Haibo et al. (2008). “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee, pp. 1322–1328.

Huisa, Carlos M et al. (2023). “PCG Heart Sounds Quality Classification Using Neural Networks and SMOTE Tomek Links for the Think Health Project”. In: *Proceedings of Data Analytics and Management: ICDAM 2022*. Springer, pp. 803–811.

Hussein, Ahmed Saad et al. (2019). “A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE”. In: *International Journal of Computational Intelligence Systems* 12.2, pp. 1412–1422.

Indonesia, Data Folks (2020). *Powering Up Your Pandas Part II: Label Encoding and One-Hot Encoding*. Accessed: 2024-06-27. URL: `https://medium.com/data-folks-indonesia/powering-up-your-pandas-part-ii-label-encoding-and-one-hot-encoding-dac0fce045da`.

Islam, Ashhadul et al. (2022). “KNNOR: An oversampling technique for imbalanced datasets”. In: *Applied soft computing* 115, p. 108288.

Jaiswal, Jitendra Kumar and Rita Samikannu (2017). “Application of random forest algorithm on feature subset selection and classification and regression”. In: *2017 world congress on computing and communication technologies (WCCCT)*. Ieee, pp. 65–68.

Javeed, Ashir et al. (2019). “An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection”. In: *IEEE access* 7, pp. 180235–180243.

Jeatrakul, Piyasak, Kok Wai Wong, and Chun Che Fung (2010). “Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm”. In: *Neural Information Processing. Models and Applications: 17th International Conference, ICONIP 2010, Sydney, Australia, November 22-25, 2010, Proceedings, Part II 17*. Springer, pp. 152–159.

Jeon, Hyelynn and Sejong Oh (2020). “Hybrid-recursive feature elimination for efficient feature selection”. In: *Applied Sciences* 10.9, p. 3211.

Jeong, Dong Hyun et al. (2022). “Designing a supervised feature selection technique for mixed attribute data analysis”. In: *Machine Learning with Applications* 10, p. 100431.

Jiang, Xinyu, Chenfei Ma, and Kianoush Nazarpour (2024). “One-shot random forest model calibration for hand gesture decoding”. In: *Journal of Neural Engineering* 21.1, p. 016006.

Jomthanachai, Suriyan, Wai Peng Wong, and Khai Wah Khaw (2022). “An application of machine learning regression to feature selection: a study of logistics performance and economic attribute”. In: *Neural Computing and Applications* 34.18, pp. 15781–15805.

Jonathan, Bern, Panca Hadi Putra, and Yova Ruldeviyani (2020). “Observation imbalanced data text to predict users selling products on female daily with smote, tomek, and smote-tomek”. In: *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*. IEEE, pp. 81–85.

Khairy, Marwa, Tarek M Mahmoud, and Tarek Abd-El-Hafeez (2024). “The effect of rebalancing techniques on the classification performance in cyberbullying datasets”. In: *Neural Computing and Applications* 36.3, pp. 1049–1065.

Khan, Shahzad Ali and Zeeshan Ali Rana (2019). “Evaluating performance of software defect prediction models using area under precision-Recall curve (AUC-PR)”. In: *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)*. IEEE, pp. 1–6.

Kotb, Mohamed Hanafy and Ruixing Ming (2021). “Comparing SMOTE family techniques in predicting insurance premium defaulting using machine learning models”. In: *International Journal of Advanced Computer Science and Applications* 12.9.

Kotipalli, Kiranmayi and Shan Suthaharan (2014). “Modeling of class imbalance using an empirical approach with spambase dataset and random forest classification”. In: *Proceedings of the 3rd annual conference on Research in information technology*, pp. 75–80.

Kumarage, Prabha M, B Yogarajah, and Nagulan Ratnarajah (2019). “Efficient feature selection for prediction of diabetic using LASSO”. In: *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*. Vol. 250. IEEE, pp. 1–7.

Lan, Gongmin, Chenping Hou, and Dongyun Yi (2016). “Robust feature selection via simultaneous capped ℓ2-norm and ℓ2, 1-norm minimization”. In: *2016 IEEE international conference on big data analysis (ICBDA)*. IEEE, pp. 1–5.

Li, Cuiying and Weiguo Li (2010). “Partial least squares method based on least absolute shrinkage and selection operator”. In: *2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*. Vol. 4. IEEE, pp. V4–591.

Li, Yihong et al. (2021). “SP-SMOTE: A novel space partitioning based synthetic minority oversampling technique”. In: *Knowledge-Based Systems* 228, p. 107269.

Liu, Lei-Lei, Can Yang, and Xiao-Mi Wang (2021). “Landslide susceptibility assessment using feature selection-based machine learning models”. In: *Geomech. Eng* 25.1, pp. 1–16.

Liu, Xinran (2024). “Comparison of different machine learning models: Linear model, forest and SVM”. In: *ACE* 51, pp. 225–230. DOI: 10.54254/2755-2721/51/20241467.

Liu, Yingxin et al. (2021). “A comparative study of forest methods for time-to-event data: variable selection and predictive performance”. In: *BMC Medical Research Methodology* 21, pp. 1–16.

Manzo, Gianluca and Delia Baldassarri (2015). “Heuristics, interactions, and status hierarchies: An agent-based model of deference exchange”. In: *Sociological Methods & Research* 44.2, pp. 329–387.

Mao, Wentao et al. (2017). “Online sequential prediction of imbalance data with two-stage hybrid strategy by extreme learning machine”. In: *Neurocomputing* 261, pp. 94–105.

Mathew, Rose Mary and Ranganathan Gunasundari (2023). “A Cluster-based Undersampling Technique for Multiclass Skewed Datasets”. In: *Engineering, Technology & Applied Science Research* 13.3, pp. 10785–10790.

Medha, Ashmita Roy et al. (2022). “A Synthetic Hybrid Approach for Class Imbalance”. In: *2022 IEEE Silchar Subsection Conference (SILCON)*. IEEE, pp. 1–7.

Mirzaei, Behzad, Bahareh Nikpour, and Hossein Nezamabadi-pour (2021). “CDBH: A clustering and density-based hybrid approach for imbalanced data classification”. In: *Expert Systems with Applications* 164, p. 114035.

Mohd, Fatihah et al. (2019). “Improving accuracy of imbalanced clinical data classification using synthetic minority over-sampling technique”. In: Springer, pp. 99–110.

Molla, MM Imran et al. (2022). "Feature Selection and Prediction of Heart Disease Using Machine Learning Approaches". In: *Proceedings of the 6th International Conference on Electrical, Control and Computer Engineering: InECCE2021, Kuantan, Pahang, Malaysia, 23rd August*. Springer, pp. 951–963.

Morkonda Gunasekaran, Dinesh and Prabha Dhandayudam (2021). "Design of novel multi filter union feature selection framework for breast cancer dataset". In: *Concurrent Engineering* 29.3, pp. 285–290.

Movahedi, Faezeh, Rema Padman, and James F Antaki (2023). "Limitations of receiver operating characteristic curve on imbalanced data: assist device mortality risk scores". In: *The Journal of thoracic and cardiovascular surgery* 165.4, pp. 1433–1442.

Mubiayi, Mukuna Patrick and Veeredhi Vasudeva Rao (2020). "Artificial Neural Networks (ANNs) for prediction and optimization in friction stir welding process: an overview and future trends". In: *Nature-Inspired Optimization in Advanced Manufacturing Processes and Systems*, pp. 37–50.

Narwane, Swati V and Sudhir D Sawarkar (2021). "Effects of class imbalance using machine learning algorithms: case study approach". In: *International Journal of Applied Evolutionary Computation (IJAEC)* 12.1, pp. 1–17.

National Cancer Institute (2017). *Surveillance, Epidemiology, and End Results (SEER) Program*. URL: `https://seer.cancer.gov/data-software/documentation/seerstat/nov2017/`.

Nazmi, Haziq et al. (2023). "Predictive modeling of marine fish production in brunei darussalam's aquaculture sector: A comparative analysis of machine learning and statistical techniques". In: *International Journal of Advanced and Applied Sciences* 10.7, pp. 109–126.

Ou, Ge et al. (2017). "Large margin distribution machine recursive feature elimination". In: *2017 4th International Conference on Systems and Informatics (ICSAI)*. IEEE, pp. 1518–1523.

Pal, Mahesh (2012). "Multinomial logistic regression-based feature selection for hyperspectral data". In: *International Journal of Applied Earth Observation and Geoinformation* 14.1, pp. 214–220.

Patel, Vibha, Jaishree Tailor, and Amit Ganatra (2021). "Handling Class Imbalance in Electroencephalography Data Using Synthetic Minority Oversampling Technique". In: *Advances in Computing and Data Sciences: 5th International Conference, ICACDS 2021, Nashik, India, April 23–24, 2021, Revised Selected Papers, Part II 5*. Springer, pp. 12–21.

Paul, Saurabh and Petros Drineas (2016). "Feature selection for ridge regression with provable guarantees". In: *Neural computation* 28.4, pp. 716–742.

Prastyo, Pulung Hendro, Igi Ardiyanto, and Risanuri Hidayat (2020). "A Review of Feature Selection Techniques in Sentiment Analysis Using Filter, Wrapper, or Hybrid Methods". In: *2020 6th International Conference on Science and Technology (ICST)*. Vol. 1. IEEE, pp. 1–6.

Priyadharshini, M et al. (2023). "Hybrid multi-label classification model for medical applications based on adaptive synthetic data and ensemble learning". In: *Sensors* 23.15, p. 6836.

Priyatno, Arif Mudi, Triyanna Widiyaningtyas, et al. (2024). "A Systematic Literature Review: Recursive Feature Elimination Algorithms". In: *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)* 9.2, pp. 196–207.

Puri, Arjun and Manoj Kumar Gupta (2022). "Improved hybrid bag-boost ensemble with K-means-SMOTE–ENN technique for handling noisy class imbalanced data". In: *The Computer Journal* 65.1, pp. 124–138.

Putra, Armanda Eka, Luh Kesuma Wardhani, et al. (2019). "Chi-Square Feature Selection Effect On Naive Bayes Classifier Algorithm Performance For Sentiment Analysis Document". In: *2019 7th International Conference on Cyber and IT Service Management (CITSM)*. Vol. 7. IEEE, pp. 1–7.

Qamar, Roheen and Zardari (2023). "Artificial neural networks: An overview". In: *Mesopotamian Journal of Computer Science* 2023, pp. 124–133.

Qing, Zhipeng et al. (2022). "ADASYN-LOF Algorithm for Imbalanced Tornado Samples". In: *Atmosphere* 13.4, p. 544. DOI: 10.3390/atmos13040544. URL: https://www.mdpi.com/2073-4433/13/4/544.

Qiu, Yun Fei, Wei Wang, and Da You Liu (2013). "Research on an improved CHI feature selection method". In: *Applied Mechanics and Materials* 241, pp. 2841–2844.

Qu, Wendi et al. (2020). "Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging". In: *International journal of computer assisted radiology and surgery* 15, pp. 2041–2048.

Quan, Steven Jige (2024). "Comparing hyperparameter tuning methods in machine learning based urban building energy modeling: A study in Chicago". In: *Energy and Buildings*, p. 114353.

Quek, Jia Yi Vivian et al. (2023). "Customer churn prediction through attribute selection analysis and support vector machine". In: *Journal of Telecommunications and the Digital Economy* 11.3, pp. 180–194.

Rai, Bharatendra (2019). "Supervised Machine Learning: Application Example Using Random Forest in R". In: *Mathematics Applied to Engineering and Management*. CRC Press, pp. 25–37.

Rainio, Oona, Jarmo Teuho, and Riku Klén (2024). “Evaluation metrics and statistical tests for machine learning”. In: *Scientific Reports* 14.1, p. 6086.

Rekha, G et al. (2021). “Class imbalanced data: Open issues and future research directions”. In: *2021 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, pp. 1–6.

Ribeiro, V Henrique Alves and Gilberto Reynoso-Meza (2024). “Multi-criteria Decision-Making Techniques for the Selection of Pareto-optimal Machine Learning Models in a Drinking-Water Quality Monitoring Problem”. In: *International Journal of Information Technology & Decision Making* 23.01, pp. 447–474.

Rutecki, Marcin (2022). *SMOTE and Tomek Links for Imbalanced Data*. Accessed: 2024-06-29. URL: `https://www.kaggle.com/code/marcinrutecki/smote-and-tomek-links-for-imbalanced-data`.

Sawadogo, Zakaria et al. (2022). “Android malware detection: Investigating the impact of imbalanced data-sets on the performance of machine learning models”. In: *2022 24th International Conference on Advanced Communication Technology (ICACT)*. IEEE, pp. 435–441.

Sevastyanov, Leonid A and Eugene Yu Shchetinin (2020). “On methods for improving the accuracy of multi-class classification on imbalanced data.” In: *ITTMM* 20, pp. 70–82.

Shaharum, Syamimi Mardiah, Kenneth Sundaraj, and Khaled Helmy (2015). “Performance analysis of feature selection method using ANOVA for automatic wheeze detection”. In: *Jurnal Teknologi* 77.7.

Shao, Guofan et al. (2019). *Strengthening Machine Learning Reproducibility for Image Classification. Advances in Artificial Intelligence and Machine Learning. 2022; 2 (4): 32*.

Sharifai, Abdulrauf Garba, Ishola Dada Muraina, and Usman Alhaji Abdurrahman (2022). “An adaptive synthetic sample coupled with ensemble multi-filter approaches for the high dimensional imbalanced dataset”. In: *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*. IEEE, pp. 1–7.

Sharma, Harsh and Anushika Gosain (2022). “Oversampling Methods to Handle the Class Imbalance Problem: A Review”. In: *International Conference on Soft Computing and its Engineering Applications*. Springer, pp. 96–110.

Shi, Jin Tao et al. (2014). “Chinese sentiment classifier machine learning based on optimized information gain feature selection”. In: *Advanced Materials Research* 988, pp. 511–516.

Shoohi, Liqaa M and Jamila H Saud (2020). “Adaptation Proposed Methods for Handling Imbalanced Datasets based on Over-Sampling Technique”. In: *Al-Mustansiriyah Journal of Science* 31.2, pp. 25–29.

Singh, Archana, Mamta Mittal, Amrender Kumar, et al. (2020). "Predictive modeling of pan evaporation using random forest algorithm along with features selection". In: *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, pp. 380–384.

Smithson, Michael (2023). "The receiver operating characteristic area under the curve (or mean ridit) as an effect size." In: *Psychological Methods*.

Soltanzadeh, Paria and Mahdi Hashemzadeh (2021). "RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem". In: *Information Sciences* 542, pp. 92–111.

Sumi, MS Suresh and Athi Narayanan (2019). "Improving classification accuracy using combined filter+ wrapper feature selection technique". In: *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, pp. 1–6.

Susan, Seba and Amitesh Kumar (2020). "Hybrid of intelligent minority oversampling and PSO-based intelligent majority undersampling for learning from imbalanced datasets". In: *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2*. Springer, pp. 760–769.

Switrayana, I Nyoman et al. (2023). "Sentiment Analysis and Topic Modeling of Kitabisa Applications using Support Vector Machine (SVM) and Smote-Tomek Links Methods". In: *International Journal of Engineering and Computer Science Applications (IJECSA)* 2.2, pp. 87–98.

Tahfim, Syed As-Sadeq and Yan Chen (2024). "Comparison of cluster-based sampling approaches for imbalanced data of crashes involving large trucks". In: *Information* 15.3, p. 145.

Tehranipour, Soheil (2021). *IT Customer Churn*. Accessed: 2024-07-22. URL: `https://www.kaggle.com/datasets/soheiltehranipour/it-customer-churn/data`.

Thölke, Philipp et al. (2023). "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data". In: *NeuroImage* 277, p. 120253.

Tsou, Chi-Ming, Shyue-Ping Chi, and Deng-Yuan Huang (2010). "EDLRT: Entropy-based dummy variables logistic regression tree". In: *Intelligent Data Analysis* 14.6, pp. 683–700.

Ugarković, Alen and Diiana Oreški (2022). "Supervised and Unsupervised Machine Learning Approaches on Class Imbalanced Data". In: *2022 International Conference on Smart Systems and Technologies (SST)*. IEEE, pp. 159–162.

Wald, Randall, Taghi M Khoshgoftaar, and Amri Napolitano (2013). "Stability of filter-and wrapper-based feature subset selection". In: *2013 IEEE 25th International conference on tools with artificial intelligence*. IEEE, pp. 374–380.

Wang, R, KY Li, and Y-x Su (2022). "Prediction of ameloblastoma recurrence using random forest—a machine learning algorithm". In: *International Journal of Oral and Maxillofacial Surgery* 51.7, pp. 886–891.

Wang, Yuxian and Changyin Zhou (2021). "Feature selection method based on chi-square test and minimum redundancy". In: *Emerging Trends in Intelligent and Interactive Systems and Applications: Proceedings of the 5th International Conference on Intelligent, Interactive Systems and Applications (IISA2020)*. Springer, pp. 171–178.

Wichitaksorn, Nuttanan, Yingyue Kang, and Faqiang Zhang (2023). "Random feature selection using random subspace logistic regression". In: *Expert Systems with Applications* 217, p. 119535.

Wu, Yichao (2021). "Can't ridge regression perform variable selection?" In: *Technometrics* 63.2, pp. 263–271.

Xie, Yazhou (2024). "Deep Learning in Earthquake Engineering: A Comprehensive Review". In: *arXiv preprint arXiv:2405.09021*.

Xu, Zhaozhao et al. (2022). "A synthetic minority oversampling technique based on Gaussian mixture model filtering for imbalanced data classification". In: *IEEE Transactions on Neural Networks and Learning Systems*.

Yan, Ke and David Zhang (2015). "Feature selection and analysis on correlated gas sensor data with recursive feature elimination". In: *Sensors and Actuators B: Chemical* 212, pp. 353–363.

Yao, Chengwei et al. (2017). "Pre-training the deep generative models with adaptive hyperparameter optimization". In: *Neurocomputing* 247, pp. 144–155.

Yu, Bo et al. (2023). "Classification method for failure modes of RC columns based on class-imbalanced datasets". In: *Structures*. Vol. 48. Elsevier, pp. 694–705.

Zakharov, Roman and Pierre Dupont (2011). "Ensemble logistic regression for feature selection". In: *Pattern Recognition in Bioinformatics: 6th IAPR International Conference, PRIB 2011, Delft, The Netherlands, November 2-4, 2011. Proceedings 6*. Springer, pp. 133–144.

Zambom, Adriano Zanin and Michael G Akritas (2015). "Nonparametric significance testing and group variable selection". In: *Journal of Multivariate Analysis* 133, pp. 51–60.

Zhang, Chenggang et al. (2016). "An imbalanced data classification algorithm of de-noising auto-encoder neural network based on SMOTE". In: *MATEC Web of Conferences*. Vol. 56. EDP Sciences, p. 01014.

Zhang, Shangli et al. (2015). “Variable selection in logistic regression model”. In: *Chinese Journal of Electronics* 24.4, pp. 813–817.

Zhang, Wenhao, Ramin Ramezani, and Arash Naeim (2019). “WOTBoost: Weighted oversampling technique in boosting for imbalanced learning”. In: *2019 IEEE international conference on big data (Big data)*. IEEE, pp. 2523–2531.

Zhang, Yazhou (2018). “Deep generative model for multi-class imbalanced learning”. In.

Zheng, Ming et al. (2022). “A method for analyzing the performance impact of imbalanced binary data on machine learning models”. In: *Axioms* 11.11, p. 607.

APPENDIX A

APPENDIX A

SAMPLE MACHINE LEARNING CODES

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn import linear_model
from sklearn import metrics
import plotly.express as px
from plotly.subplots import make_subplots
import plotly.graph_objects as go
plt.rc('font', size=20)
plt.rc('axes', titlesize=20)
plt.rc('axes', labelsize=20)
plt.rc('xtick', labelsize=20)
plt.rc('ytick', labelsize=20)
plt.rc('legend', fontsize=20)
plt.rc('figure', titlesize=20)
import warnings
warnings.filterwarnings('ignore')
%config Completer.use_jedi = False
```

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import ADASYN
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, roc_auc_score, f1_score, confusion_matrix, accuracy_score
from sklearn.preprocessing import StandardScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from tensorflow.keras.utils import to_categorical

# Apply ADASYN
adasyn = ADASYN(random_state=500)
X_train_adasyn, y_train_adasyn = adasyn.fit_resample(X_train, y_train)

# Feature Selection using Correlation Coefficient
correlation_matrix = X_train_adasyn.corr().abs()
upper = correlation_matrix.where(np.triu(np.ones(correlation_matrix.shape), k=1).astype(np.bool_))
to_drop = [column for column in upper.columns if any(upper[column] > 0.9)]
X_train_adasyn_filtered = X_train_adasyn.drop(columns=to_drop)
X_test_filtered = X_test.drop(columns=to_drop)

# Random Forest Classifier
rf_classifier = RandomForestClassifier(random_state=500)
rf_classifier.fit(X_train_adasyn_filtered, y_train_adasyn)
rf_pred = rf_classifier.predict(X_test_filtered)
rf_proba = rf_classifier.predict_proba(X_test_filtered)[:, 1]

# Random Forest Evaluation
rf_roc_auc = roc_auc_score(y_test, rf_proba)
rf_f1_score = f1_score(y_test, rf_pred, average='weighted')
print("Random Forest Classifier:")
print(classification_report(y_test, rf_pred))
print("ROC-AUC score:", rf_roc_auc)
print("F1 score:", rf_f1_score)
print("Confusion Matrix:\n", confusion_matrix(y_test, rf_pred))
print("Accuracy:", accuracy_score(y_test, rf_pred))
```

```
# Preprocessing for DNN
scaler = StandardScaler()
X_train_adasyn_scaled = scaler.fit_transform(X_train_adasyn_filtered)
X_test_scaled = scaler.transform(X_test_filtered)

# Convert the target variable to categorical
y_train_adasyn_cat = to_categorical(y_train_adasyn)
y_test_cat = to_categorical(y_test)

# DNN Model
dnn_model = Sequential()
dnn_model.add(Dense(64, input_shape=(X_train_adasyn_scaled.shape[1],), activation='relu'))
dnn_model.add(Dense(32, activation='relu'))
dnn_model.add(Dense(y_train_adasyn_cat.shape[1], activation='softmax'))
dnn_model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

# Train the DNN
dnn_model.fit(X_train_adasyn_scaled, y_train_adasyn_cat, epochs=10, batch_size=32, verbose=0)

# DNN Evaluation
dnn_pred = dnn_model.predict(X_test_scaled)
dnn_pred_classes = np.argmax(dnn_pred, axis=1)
dnn_proba = dnn_pred[:, 1]

dnn_roc_auc = roc_auc_score(y_test, dnn_proba)
dnn_f1_score = f1_score(y_test, dnn_pred_classes, average='weighted')
print("\nDeep Neural Network:")
print(classification_report(y_test, dnn_pred_classes))
print("ROC-AUC score:", dnn_roc_auc)
print("F1 score:", dnn_f1_score)
print("Confusion Matrix:\n", confusion_matrix(y_test, dnn_pred_classes))
print("Accuracy:", accuracy_score(y_test, dnn_pred_classes))
```

```
# Apply ADASYN for handling class imbalance
adasyn = ADASYN(random_state=3030)
X_train_adasyn, y_train_adasyn = adasyn.fit_resample(X_train, y_train)

# Ensure all features are non-negative for chi-square test
scaler = MinMaxScaler()
X_train_adasyn_scaled = scaler.fit_transform(X_train_adasyn)
X_test_scaled = scaler.transform(X_test)

# Feature Selection using Chi-Square
# Select the top k features
feature_names = X.columns
k = 10  # Example, select top 10 features
chi2_selector = SelectKBest(chi2, k=k)
X_train_adasyn_chi2 = chi2_selector.fit_transform(X_train_adasyn, y_train_adasyn)
X_test_chi2 = chi2_selector.transform(X_test)

# Getting the names of the selected features
selected_features_bool = chi2_selector.get_support()
selected_features_names = feature_names[selected_features_bool]
print("Selected features:", selected_features_names)


# Random Forest Classifier
```

```python
# Preliminary imports and setup
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import SelectKBest, chi2, RFE
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LassoCV
from sklearn.metrics import classification_report, f1_score, roc_auc_score
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from tensorflow.keras.utils import to_categorical
from sklearn.preprocessing import StandardScaler
from imblearn.combine import SMOTETomek

# Feature-target split
X = data.drop('HadHeartAttack', axis=1)
y = data['HadHeartAttack']

# Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=500

# Apply SMOTE-Tomek
smote_tomek = SMOTETomek(random_state=500)
X_train_smote_tomek, y_train_smote_tomek = smote_tomek.fit_resample(X_train, y_train)

# Step 1: Apply Filter Method (Chi-Squared)
chi2_selector = SelectKBest(chi2, k=10)
X_train_chi2 = chi2_selector.fit_transform(X_train_smote_tomek, y_train_smote_tomek)
selected_features_chi2 = X_train.columns[chi2_selector.get_support()]

selected_feature_names_chi2 = selected_features_chi2
chi2_pvalues = chi2_selector.pvalues_
selected_pvalues_chi2 = chi2_pvalues[chi2_selector.get_support()]

# Create a DataFrame to display feature names and their corresponding p-values
df_selected_features_chi2 = pd.DataFrame({
    'Feature Name': selected_feature_names_chi2,
    'P-value': selected_pvalues_chi2
})
print("Selected features from Chi-Squared:\n", df_selected_features_chi2)
```

```
# Train and Evaluate DNN Model
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_final)
X_test_scaled = scaler.transform(X_test_final)
y_train_categorical = to_categorical(y_train_smote_tomek)
y_test_categorical = to_categorical(y_test)

dnn_model = Sequential([
    Dense(64, input_dim=X_train_scaled.shape[1], activation='relu'),
    Dense(32, activation='relu'),
    Dense(y_train_categorical.shape[1], activation='softmax')
])

dnn_model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
dnn_model.fit(X_train_scaled, y_train_categorical, epochs=60, batch_size=32, verbose=1)

y_pred_dnn = dnn_model.predict(X_test_scaled)
y_pred_dnn_classes = np.argmax(y_pred_dnn, axis=1)
dnn_f1_score = f1_score(y_test, y_pred_dnn_classes, average='weighted')
dnn_roc_auc = roc_auc_score(y_test_categorical[:, 1], y_pred_dnn[:, 1])

# Corrected Print Statements for Metrics
print(f'DNN Model F1 Score: {dnn_f1_score:.4f}')
print(f'DNN Model ROC AUC: {dnn_roc_auc:.4f}')

# Evaluation Metrics for RF Model
print("Random Forest Classification Report:\n", classification_report(y_test, y_pred_rf))
print("Random Forest ROC AUC Score:", roc_auc_score(y_test, rf_classifier.predict_proba(X_test_final)[:, 1]))

# Correctly process DNN predictions for binary classification
if y_train_categorical.shape[1] == 2:
    y_pred_dnn_binary = (y_pred_dnn[:, 1] > 0.5).astype(int)
else:
    y_pred_dnn_binary = np.argmax(y_pred_dnn, axis=1)

# Evaluation Metrics for DNN Model
print("\nDeep Neural Network Classification Report:\n", classification_report(y_test, y_pred_dnn_binary))
print("Deep Neural Network F1 Score:", f1_score(y_test, y_pred_dnn_binary, average='weighted'))
dnn_roc_auc_score = roc_auc_score(y_test, y_pred_dnn[:, 1]) if y_train_categorical.shape[1] == 2 else "N/A for multiclass"
print("Deep Neural Network ROC AUC Score:", dnn_roc_auc_score)
```

VITA

Martha Asare began her academic journey at the Kwame Nkrumah University of Science and Technology in Ghana, where she earned a bachelor's degree in Statistics. She was a member of the American Statistical Association and the American Mathematics Association. In August 2024, she obtained a master's degree in Applied Statistics and Data Science from the University of Texas Rio Grande Valley (UTRGV), showcasing her exceptional aptitude and passion for utilizing machine learning techniques to address imbalanced datasets. Her academic interests included employing machine learning models to handle big datasets and enhance predictive modeling. Her work concentrated on implementing machine learning and deep learning techniques to examine Big Data. Martha had experience in teaching and was enthusiastic about data analysis. She believed that extensive training in data analytics and its applications would equip her with the ability to tackle complex and challenging projects, making a significant impact on the realm of data science. Martha accepted a position at Lawrence Berkeley National Laboratory as a research assistant in the summer of 2024. With a strong foundation in machine learning & access to the Perlmutter supercomputer, she envisioned a future where data-driven approaches could revolutionize our understanding and response to highly imbalanced data, improving data accuracy and implementation. Her long-term objective was to become a leading researcher and consultant in machine learning and data analytics, offering insights into the balancing, evaluation, and implementation of various datasets. During her time at UTRGV, Martha demonstrated significant leadership and community involvement. She was the co-founder of the African Students Association at UTRGV & the Vice President for the SPIE-UTRGV.In recognition of her academic excellence and research contributions, Martha received the Best Master's Research Student award & was on the honor roll for academic performance at UTRGV. To get in touch with Martha, please email her at marthaasare62@gmail.com.