NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

A Comparative Analysis of Synthetic Data Generation Techniques in the Context of Medical Applications

Susana Teresa Dias

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

A Comparative Analysis of Synthetic Data Generation Techniques in the Context of Medical Applications

by

Susana Teresa Dias

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Professor Fernando José Ferreira Lucas Bação

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

DEDICATION

This section wouldn't make sense without first mentioning the person who made this and all my achievements tangible. Mom, you are my biggest inspiration and strength. I am so grateful for all your support, love, and understanding, especially over the past few years. Although words and writing are not my strong suit, even if they were, there wouldn't be enough to express my feelings.

Thank you to the person whose "mimi" ends this cycle today. You are the one in whom I find the necessary peace of mind I need and a place of great love.

"Siblings by chance, friends by choice": to the person whose words always manage to comfort me during my tense moments, even without me mentioning it. Your words and your wisdom are precious to me.

To my three "borrowed" sisters, Beatriz, Sofia and Mariana. Thank you for your tireless support and unique understanding throughout my life. Always and forever.

To the person who witnessed first-hand not only this journey but also the one I have been on for the past seven years ago. *"Um bocadinho ".*

To Gina, Alexandre, and Nuno*: "Cousins are the friends who feel like family and the family who feels like friends".*

To my friend Sara Quintal, who has experienced the path I've been on in parallel, and who has always been there in times of inner crisis.

Despite being geographically separated and having months between conversations, my friends' encouraging words and support meant a lot to me. Thank you, Mário and Gonçalo.

To my "home" family, who shared many of my headaches and always made sure I had everything I needed to finish this journey.

ACKNOWLEDGEMENTS

Firstly, I want to express my gratitude to my supervisor Professor Fernando Bação for his constant support and guidance throughout this academic year. Without his persistence, time spent providing me with advice and his commitment to my studies, it would not have been possible for me to complete this work.

I also want to thank Professor Georgios Douzas whose advice was invaluable, especially during the experimental stage of this thesis.

Furthermore, I must express my appreciation to the NOVA IMS teachers who assisted me during this journey. They have been extremely supportive and helpful, demonstrating a constant willingness to answer all questions without hesitation even beyond class hours. I truly appreciate their dedication and commitment to all students' education.

ABSTRACT

In an era marked by the presence of data and technological advancements, numerous machine learning applications across various industries are still hindered by the reliance on small datasets. This results in suboptimal prediction performance in supervised learning tasks and, as a result, poor decision-making. The medical field is no exception to this issue. This research aims to address the small dataset problem by employing artificial data generation techniques on both minority and majority classes to attempt to recreate the original dataset. The objective is to compare and assess the performance of these techniques to determine which one most accurately replicates the original data. Five artificial data generation methods were applied to medical datasets and evaluated using four different classifiers. Experimental results demonstrate favorable f-scores compared to those achieved with the original small datasets. The results indicate that generating artificial data is a potential approach to address the issue of small data in classification tasks.

KEYWORDS

Data Generation Techniques; Synthetic Data; Class Imbalance; Oversampling; Healthcare

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADASYN	Adaptive Synthetic Sampling
AUC	Area Under the ROC Curve
BP	Boostrapping Procedure
BPNN	Back Propagation Neural Network
B-SMOTE	Borderline Synthetic Minority Over-Sampling Technique
DNN	Diffusion Neural Network
FMS	Flexible Manufacturing Systems
FN	False Negative
FP	False Positive
FVP	Functional Virtual Population
G-SMOTE	Geometric Synthetic Minority Over-Sampling Technique
GB	Gradient Boosting
KNN	K-Nearest Neighbor
LR	Logistic Regression
ML	Machine Learning
MTD	Mega-Trend-Diffusion
RF	Random Forest
SMOTE	Synthetic Minority Over-Sampling Technique
TN	True Negative
TP	True Positive
VSG	Virtual Sample Generation

1. INTRODUCTION

In an era marked by the presence of data and technological advancements, Machine Learning (ML) and Artificial Intelligence (AI) have found applicability in various sectors. The ability of ML algorithms to discern patterns, interpret data, and make predictions has been harnessed in diverse fields, from finance and manufacturing to transportation and entertainment (Tonge Buradkar & More, 2020). Nowadays, ML serves as an invaluable tool, streamlining operations, optimising processes, and facilitating data-driven decision-making in these realms.

ML has emerged as a powerful tool in the healthcare industry, as Rubinger et al. (2023) explains. The conventional methodologies of case-control studies and randomized controlled trials, while fundamental, have limitations: they are time-consuming, expensive, and prone to biases. ML provides an alternative to these methods (Javaid et al., 2022). As an example of its applicability, models have been trained using labelled medical imaging data to effectively identify cancerous tumours, and forecast patient outcomes using records, genetic information, and other healthcare data. This allows healthcare professionals to enhance patient care and emphasizes the relevance of supervised learning in the healthcare sector. The subcategory's capacity to forecast and monitor disease outbreaks, streamline discovery procedures, and enhance operational efficiency has established its status as a revolutionary influence in the healthcare field (Javaid et al., 2022).

One thing that should be taken into consideration when developing ML models is that the quality of input data significantly influences the assessment and predictions formulated by machine learning algorithms (Rubinger et al., 2023): *"To obtain a precise prediction model, a large sample set of data is required for the learning process."* (Abdul Lateh et al., 2017).

In the medical field, despite the technological advances that ML has brought to this industry, the prevalence of small and imbalanced datasets constitutes two common challenges.

As highlighted by Abdul Lateh et al. (2017), when working with small datasets, more difficulties in building accurate prediction models are faced, due to the limited amount of data available: they might lack the necessary information to build accurate prediction

models, and information gaps – Figure 1 - that may exist between samples lead to limited learning capabilities: *"(...) which cause most of the learning tools are difficult to predict."* (Abdul Lateh et al., 2017).

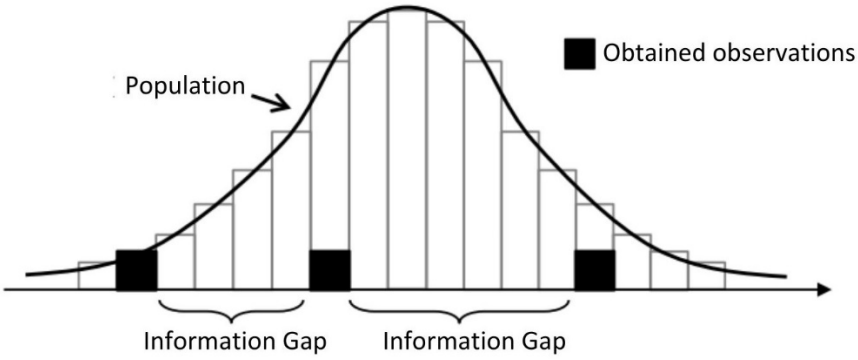

Figure 1 - Distribution of a small dataset relative to its population (C. H. Tsai & Li, 2015)

The imbalanced problem – another type of data scarcity problem - arises when there is an insufficient amount of data belonging to one class in a binary or multi-class classification task. In Fernández, García, Galar, et al. (2018), the authors note that this occurs frequently in disease diagnosis, since more benign records are available than malign, creating an imbalanced classification problem. The majority class tends to be prioritized over the minority class, resulting in a biased model that performs well in majority class instances but poorly in minority class instances, leading to inaccurate predictions for the less represented class.

The type of data scarcity addressed in this study is the small data problem, which differs from the imbalanced learning problem primarily in that both classes are affected by data insufficiency, not just the minority class.

From the literature, different methods have been explored to address this problem, one of them being the integration of artificial data into the system. The accuracy of the learning algorithm is improved when the amount of training data examples increases. Consequently, this will encompass the approach of generating artificial data, to effectively address the challenges associated with a small dataset problem in predictive modelling, specifically within medical datasets (Abdul Lateh et al., 2017b).

In subsequent sections, we will undertake a comprehensive review of the literature on artificial data generation techniques to overcome the small dataset problem. This will be followed by an explanation of the research methodology used in this work. Ultimately, we will provide the empirical findings, and analyze and evaluate these data to conclude. Finally, we will be discussing the results and culminating in the formulation of conclusions.

2. LITERATURE REVIEW

In this section, we will explain what has been done to overcome the small data problem. The research community has proposed various techniques to increase the size of datasets through pre-processing. This section covers the three most significant methods: first, we explore fuzzy theories, a historically prevalent technique for easing issues with small datasets. Secondly, resampling techniques like bootstrapping, and, finally, we will discuss oversampling methods which can significantly increase the sample size for all classes in a small dataset.

2.1. FUZZY THEORIES

Artificial data generation methods employ fuzzy theories and neural networks to estimate or approximate functions, enhancing the learning process and thereby facilitating the creation of additional training set samples (Abdul Lateh et al., 2017b).

A strategy initially created to address the problem of small datasets is to utilize existing knowledge to create virtual data, as Niyogi et al. (1998) suggested, thereby increasing the size of the training dataset. In a study by Niyogi et al. (1998), the Virtual Sample Generation (VSG) concept was originally introduced, where they demonstrated that *"the process of creating artificial samples is mathematically equivalent to incorporating prior knowledge"* (Niyogi et al., 1998). The method they conducted involves mathematically transforming a given 3-D view of an object to create new perspectives, referred to as virtual samples. These newly generated views expand the available information of the training set. Based on the findings thus far, incorporating prior knowledge into an example-based learning framework could be a viable approach. This would result in systems that can generalize effectively, even when dealing with limited amounts of data commonly encountered in real-world scenarios.

Within the domain of manufacturing, D. C. Li et al. (2003) introduced the Functional Virtual Population (FVP) approach, intending to enhance the precision of acquiring scheduling knowledge in dynamic manufacturing environments by employing an artificial neural network. Despite being one of the first proposed methods for small dataset learning in scheduling problems, the findings revealed that even with as few as 40 samples, the FVP approach significantly enhanced the accuracy of learning. Years later, another solution was

proposed by D.-C. Li et al. (2006) to tackle the challenge of acquiring management knowledge in the initial stages of Flexible Manufacturing Systems (FMS). They introduced a novel methodology for constructing scheduling knowledge in an FMS by leveraging statistical learning theories, density estimation, and the generation of virtual samples. To evaluate this approach, an experiment was conducted, involving the creation of virtual samples using intervalized kernel density estimators. The results demonstrated a significant improvement in testing accuracies when utilizing the generated virtual data. This highlights the effectiveness of the approach in enhancing learning outcomes from limited datasets. The study illustrated that this devised strategy is capable of overcoming the limitations imposed by small data sizes during the training of a learning system (D.-C. Li et al., 2006).

In order to expand the size of a dataset, fuzzifying information was employed in the development of the Diffusion Neural Network (DNN), which was introduced by Huang & Moraga (2004). The main objective of the DNN model is to enhance the estimation of non-linear functions by utilizing neural networks. By incorporating the principle of information diffusion, which partially fills the information gaps, introduced by Chongfu (1997), the DNN model effectively addresses the limitations associated with small datasets, such as reduced ability to generalize and increased errors between real and estimated functions. Through computer simulation experiments comparing DNN with Back Propagation Neural Network (BPNN), the study demonstrates that the first surpasses the latter in accurately estimating non-linear functions, particularly when dealing with small sample sizes (Huang & Moraga, 2004).

Back in the realm of early FMS environments, D.-C. Li et al. (2006) proposed the mega-fuzzification technique. This method, distinct from conventional individual data point fuzzification, entails the collective fuzzification of a group of data through a shared membership function. Within this study, mega-fuzzification intertwines with data trend estimation and the application of the Adaptive-Network-Based Fuzzy Inference System (ANFIS) to augment the precision of FMS scheduling. By integrating these components, the authors aim to demonstrate an effective approach for improving learning accuracy in FMS scheduling (D.-C. Li et al., 2006).

To fully fill the information gaps, as well as to enhance the accuracy of machine learning for early FMS scheduling knowledge, D. C. Li et al. (2007) proposed the Mega-Trend-Diffusion (MTD). This approach combines the mega diffusion and the data trend estimation, by diffusing the sample set as a whole, prior to diffusing each sample (D. C. Li et al., 2007).

Taking into consideration Figure 2, and given two datasets m and n: these are diffused simultaneously into one function defined by the boundaries a and b. With the minimum and maximum values of the dataset, the algorithm determines the values of a and b, by assessing the number of data points that fall below or above the average value, therefore taking into consideration the domain range and the data skewness of the dataset. Figure 2 illustrates the membership function's triangular shape, which indicates the similarity among samples. Once the domain range between a and b has been estimated, artificial samples are generated randomly within this range utilizing a shared diffusion function. Subsequently, and as proposed by Huang & Moraga (2004), these samples are then trained with a BPNN. The primary aim of this diffusion process explained by D. C. Li et al. (2007) is to assess the potential coverage of the data sets collectively, considering them as a cohesive group. The study demonstrates significant improvements in learning accuracy for FMS scheduling with a small dataset (D. C. Li et al., 2007).

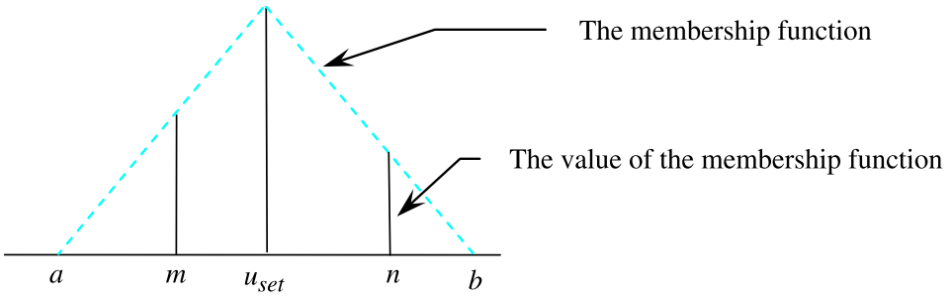

Figure 2 – Membership Function (D. C. Li et al., 2007)

In subsequent studies, the MTD technique has gained significant traction as a method for generating synthetic samples.

A new VSG method named Genetic Algorithm-based Virtual Sample Generation (GABVSG) was developed by D. C. Li & Wen (2014), which differs from other VSG approaches concerning attribute analysis: unlike common methods, the new approach explores into the combined impacts of attributes. This approach involves a three-step process: the initial step

entails randomly selecting samples of various sizes, to serve as training data sets. The second step involves utilizing a genetic algorithm to identify a set of virtual samples that are highly feasible. Finally, in the last step, average errors are calculated and compared with the actual values (D. C. Li & Wen, 2014). The comparison findings demonstrate that the GABVSG method outperforms the utilization of original training data without any virtual samples and the previous approach (D. C. Li & Wen, 2014).

The fuzzy-based algorithms mentioned above have limitations when dealing with numerical attributes. However, if a dataset includes categorical features, one can consider using a resampling mechanism like Bootstrap (Lechleitner, 2020). In the following section, we will discuss the fundamental concepts of the Bootstrap method and its application to address the small data problem.

2.2. BOOTSTRAPPING PROCEDURE

Bootstrapping Procedure (BP) is a method that was first introduced by Efron & Tibshirani (1994): BP constructs new training sets by iteratively resampling instances from the observed data, allowing for replacement (Efron & Tibshirani, 1994). The main distinction between fuzzy theories and BP is that BP generates new training sets by randomly resampling instances from the observed data, allowing for replacement. The main difference to the fuzzy theories is that BP creates new training sets by randomly resampling instances from the measured data with replacement (Efron & Tibshirani, 1994). This iterative process allows algorithms to revisit the same samples, gradually refining identified patterns to enhance predictive accuracy.

In the article of Ivănescu et al. (2006), the authors investigated the impact of limited historical data on the performance of regression-based order acceptance procedures in batch chemical plants, employing BP to generate additional job sets. The study demonstrated that the performance of the BP regression policy notably improved, reaching levels comparable to those achieved by a regression policy utilizing an extensive historical data set.

T. I. Tsai & Li (2008) also conducted a study investigating the application of the bootstrap methodology in small data set learning for pilot-run modelling of manufacturing systems. The research addressed the difficulties of building robust management models and

accurately forecasting production parameters. Through a case study using real data from a Taiwanese manufacturer of multi-layer ceramic capacitors, the results demonstrated a significant decrease in prediction error rates, indicating the effectiveness of the BP.

In the medical field, Chao et al. (2011) applied the BP algorithm to generate virtual samples to fill the information gaps (Lechleitner, 2020). In this study, measuring 13 proteins in individual cell lines post-irradiation with cobalt-60, revealed a stable increase in learning accuracy - from 55% to 85% - with an augmented number of training data. Applying this method to analyze the connection between radiotherapy outcomes and protein expression profiles aids patients in informed decision-making for treatment choices.

While BP effectively doubles the number of observations in a training set, its application to small datasets may lead to two problems:

- **Unstable data structure:** The use of BP sets may lead to an unstable data structure in small datasets, which can impact the effectiveness of learning algorithms.
- **Overfitting:** Bootstrapping in small datasets can lead to overfitting, where the identified patterns represent the behaviours of only a few observations, rather than capturing the overall characteristics of the dataset (D. C. Li et al., 2018).

Consequently, BP is not seen as an optimal solution for the issue of limited data sets. Its main function is to expand the number of training samples rather than enhance the quantity of artificially generated information, as explained by (C. H. Tsai & Li, 2015).

2.3. OVERSAMPLING TECHNIQUES

The fundamental idea of oversampling methods consists of the generation of artificial data for the minority class, in order to mitigate the impact of class imbalance in datasets, as explained in Chapter 1.

2.3.1. SMOTE

The Synthetic Minority Over-Sampling Technique (SMOTE) algorithm is widely used for oversampling to address class imbalance. It works by creating synthetic examples for the minority class and adding them to the training set. These artificial examples are generated along line segments that connect the minority class samples with their nearest neighbours

from the *k*-nearest neighbors (Chawla et al., 2002). In other words, it suggests forming a line segment between neighbouring instances of the minority class and generating synthetic data between them. The selection of these neighbours is random and depends on the desired level of oversampling. Figure 3 presents an illustrative scenario frequently addressed with SMOTE. It involves two distinguishable clusters, where one cluster represents the minority class and the other forms the majority class. During the sample generation process, SMOTE targets instances from the minority class to alleviate the issue of class imbalance.

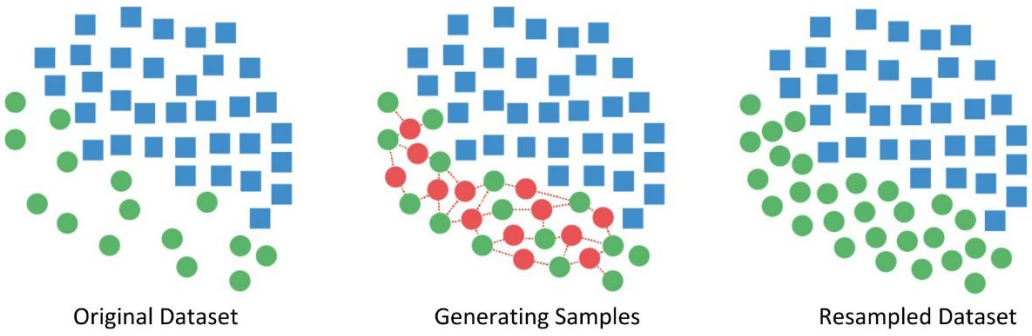

Figure 3 – SMOTE algorithm (Dholakiya, 2023)

SMOTE does have its limitations. In real-world scenarios, SMOTE encounters challenges due to the ambiguity or overlap between regions occupied by majority and minority classes. This ambiguity often renders the distinction between the two classes burdensome or even impossible. Consequently, when a minority class sample is situated within the vicinity of the majority class regions, SMOTE may inadvertently produce noisy or less reliable synthetic samples (Douzas et al. (2022) and Fernández, García, Herrera, et al. (2018)). Figure 4 demonstrates a situation in which a minority occurrence is produced within the majority region dominated by the majority - a noisy sample (Douzas & Bacao, 2019; Lechleitner, 2020). Additionally, SMOTE may generate minority class samples on a minority cluster, resulting in generating redundant information - Figure 5 presents this situation (Douzas & Bacao, 2019; Lechleitner, 2020).

Even though this method increases dataset size and enhances the generalization of a model, it should be carefully done since it may lead to overfitting which causes the classifier to rely too much on similar data points.

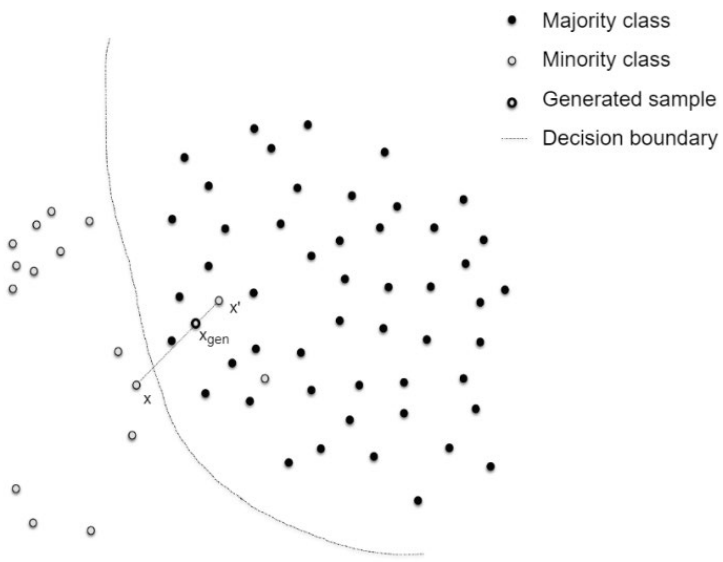

Figure 4 – Generation of noisy samples (Douzas & Bacao, 2019)

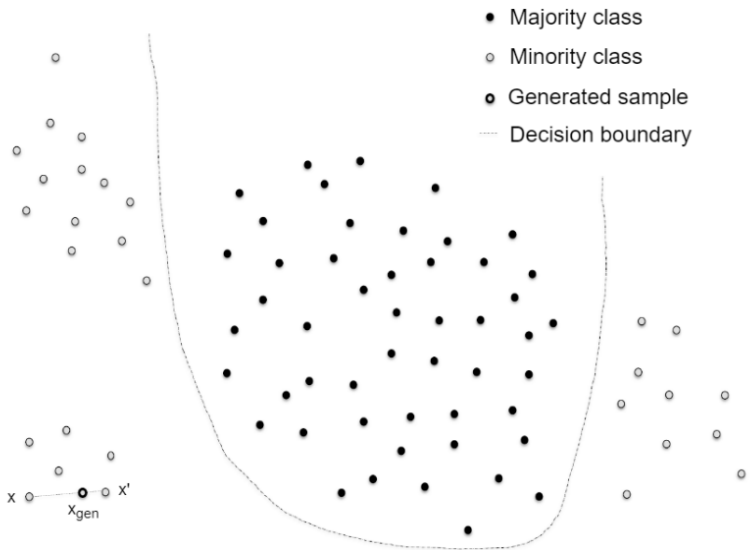

Figure 5 – Generation of redundant samples (Douzas & Bacao, 2019)

SMOTE has become very popular in the field of imbalanced learning, and, over time, numerous variations and extensions have been proposed for this algorithm, thus making it more applicable and efficient when dealing with imbalanced and small datasets (Fernández, García, Herrera, et al., 2018). These extensions have the consideration of SMOTE drawbacks, which include the inability to generate noisy examples. They achieve this by identifying the borderline instances for both the majority and minority classes, which are then used to find the informative instances of the minority class (Fernández, García, Herrera, et al., 2018). This basic framework has over time been tailored to individual domain problems or improved

with additional approaches for better results hence leading to many variations. The forthcoming sections will discuss three extensions of SMOTE.

Despite being primarily known as an algorithm widely used to address class imbalance, SMOTE can also be used to address the issue of small datasets. In a study conducted by D. C. Li et al. (2018), it was demonstrated that SMOTE can bridge the information gaps by generating synthetic samples. However, the algorithm did not yield the most optimal results due to certain limitations.

2.3.2. Borderline SMOTE

Borderline SMOTE (B-SMOTE) is an extension of SMOTE that aims to address the issue of synthesizing samples near the decision boundary between the minority and majority classes. In standard SMOTE, synthetic samples are generated indiscriminately from all minority class instances, including those that are well-separated from the majority class as well as those that are near the decision boundary (Goswami, 2020). This approach may lead to the generation of noisy or less informative synthetic samples, especially for minority class instances that are already well-represented or far from the decision boundary (Han et al., 2005).

B-SMOTE on the other hand, focuses on synthesizing samples specifically for minority class instances that are near the decision boundary. These instances are considered more crucial because they are likely to be harder to classify correctly by machine learning models due to their proximity to the majority class (Fernández, García, Herrera, et al., 2018).

The algorithm begins by classifying the minority class observations. Then, it identifies any minority observation as a noise point if all its neighbours belong to the majority class, and such observations are ignored when creating synthetic data (Goswami, 2020). It identifies certain points as border points if their neighbourhood includes both majority and minority class observations, and it resamples exclusively from these points (extreme observations that are typically the focus of support vectors) (Fernández, García, Herrera, et al., 2018; Goswami, 2020).

In summary, B-SMOTE disregards noisy points and standard minority points, focusing exclusively on border points for synthetic data generation. The primary limitation of this algorithm is its tendency to disproportionately emphasize these border points (Goswami, 2020).

2.3.3. Geometric SMOTE

Geometric SMOTE (G-SMOTE) is an improvement of the SMOTE algorithm designed to generate synthetic samples in a geometric region of the input space surrounding each selected instance (Douzas & Bacao, 2019). It begins by selecting a minority class instance from the dataset. Unlike SMOTE, which generates synthetic instances along a line segment, G-SMOTE defines a geometric region around the selected minority instance. By default, this region is a hyper-sphere, but it can be adjusted to a hyper-spheroid for greater flexibility. Synthetic occurrences are generated within this geometric region, with the algorithm controlling the region's expansion and shape to produce diverse and meaningful synthetic samples for the minority class (Douzas & Bacao, 2019; Lechleitner, 2020).

This algorithm addresses some inefficiencies of the SMOTE algorithm, such as the generation of noisy and duplicated instances due to the flexible geometric region around each minority class instance. Additionally, G-SMOTE's approach of defining geometric regions around minority instances reduces the sensitivity to the *k* value seen in SMOTE and provides a more controlled way of generating synthetic data (Douzas & Bacao, 2019).

In the study of Douzas & Bacao (2019), G-SMOTE has been evaluated against SMOTE and other techniques, demonstrating a substantial enhancement in the quality of the generated data when G-SMOTE is employed.

2.3.4. ADASYN

Similarly to the previous extensions of SMOTE, the Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) was also created to enhance the SMOTE method. The main idea is to generate synthetic samples for the minority class, focusing on creating more samples in regions where the minority class is underrepresented and is harder to classify (Goswami, 2020; He et al., 2008b).

The primary challenges that ADASYN seeks to address include generating synthetic data points specifically in regions where the minority class is underrepresented; and adaptively altering the distribution of synthetic samples based on learning difficulty, thereby producing more samples for minority class instances. It also takes into consideration the density distribution of minority class instances, leading to targeted sampling in sparse areas (Brandt & Lanzén, 2020). By focusing on these more challenging instances, ADASYN helps the classifier better understand complex decision boundaries, thus minimizing the risk of overfitting to simpler ones. The adaptive nature of ADASYN helps to reduce the introduction of noise by generating synthetic samples according to the difficulty level of each minority instance (Goswami, 2021; He et al., 2008b).

The main purpose of this approach is to specifically target the issue of class imbalance by creating artificial samples for the minority class. While it can help improve the performance of ML models on imbalanced datasets, it does have some relevance to the small data problem but with limitations.

ADASYN generates synthetic samples, thereby increasing the number of training instances. However, these synthetic samples are based on the existing data distribution, meaning that if the original dataset is very small, the synthetic samples may lack the diversity needed to significantly improve the learning process (Guggenheim, 2022). In small datasets with class imbalance, ADASYN ensures a better representation of the minority class, potentially enhancing model performance. However, if both classes are equally small, ADASYN's focus on the minority class does not address the overall data scarcity across both classes (Brandt & Lanzén, 2020).

By concentrating on minority class observations, ADASYN can help create more robust decision boundaries, which is beneficial when data is limited. Yet this improvement is constrained by the ability of the synthetic samples to accurately represent the underlying data distribution.

Like other synthetic sampling methods, ADASYN increases the quantity of data but not necessarily its quality. The fundamental challenge of small datasets — limited real-world examples — might remain. For very small datasets, generating numerous synthetic samples can lead to overfitting, as the model may learn noise rather than the true underlying

patterns (Guggenheim, 2022). The effectiveness of ADASYN heavily depends on the quality and distribution of the initial dataset. If the initial data is not representative, the synthetic samples will also lack representativeness (Gomede PhD, 2024).

While ADASYN can help mitigate some aspects of the small data problem by increasing the number of training instances and improving minority class representation, it is not a comprehensive solution. Additional approaches such as data augmentation, transfer learning, and using pre-trained models may be necessary to more effectively address the challenges posed by small datasets.

By using a weighted distribution to generate synthetic data for minority class examples based on their level of difficulty in learning, ADASYN aims to reduce bias introduced by class imbalance and enhance learning performance (Fernández, García, Herrera, et al., 2018; Peláez-Rodríguez et al., 2022). The algorithm dynamically adjusts weights and employs an adaptive learning procedure to better handle imbalanced data sets, ultimately improving the accuracy and robustness of machine learning models in scenarios where minority class samples are critical (He et al., 2008b). It adaptively decides the number of synthetic samples to create for each minority instance based on the density of the majority class instances in its neighbourhood. This helps to balance the dataset and improve the performance of classifiers on imbalanced data (He et al., 2008a).

In conclusion, the small dataset problem is a persistent challenge in the field of data science. Many researchers have attempted to address this problem through various approaches, including fuzzy theories, bootstrapping procedures, and oversampling techniques. The subsequent section will discuss the methodology employed in this research to tackle the small dataset problem.

3. METHODOLOGY

The main goal of this study is to evaluate the effectiveness of different artificial data generation methods in addressing the challenges posed by small datasets, by recreating the original data. To accomplish this, a diverse set of medical datasets, evaluation metrics, and a selection of classifiers were employed.

3.1. EXPERIMENTAL DATA

The experimental data for this research includes thirteen distinct medical datasets, each representing a medical issue or treatment with a binary target. The datasets were sourced from the Open ML Repository (*UC Irvine Machine Learning Repository*, n.d.), comprising real-world and artificial data, with imbalance ratios varying from 2 to 19. This selection ensures a comprehensive evaluation of performance metrics across various medical contexts. Detailed descriptions of the datasets and the experimental data can be found in Table 1.

Dataset	Description	Nr. of Instances	Nr. of Attributes	Imbalance Ratio
Stroke	Predict whether a patient is likely to get a stroke based on characteristics	5110	12	19,5
Heart Disease	Possible heart disease	303	14	3,1
Hepatitis C	Laboratory values of blood donors and Hepatitis C patients	615	14	7,6
Breast Cancer	Prediction models to identify potential biomarkers for breast cancer	116	10	2,46
Indian Liver Patient	Liver disease based on biochemical markers and demographic information	416	10	2,49
Cirrhosis	Used to study primary biliary cirrhosis of the liver	418	20	2,12
Thoracic Surgery	Classification problem related to the post-operative life expectancy in lung cancer patients	470	17	5,71
Autism	The dataset is used for screening and analyzing Autistic Spectrum Disorder in adults	94	19	6,23
Cryotherapy	Analyze the results of wart treatment using cryotherapy in patients	76	7	2,29

Immunotherapy	Document the outcomes of wart treatment in 90 patients employing immunotherapy	90	8	3,74
Caesarian	Information about c-section results of pregnant women with important characteristics of delivery problems	67	6	2,19
Jugoslavia Breast Cancer	Dataset from the University Medical Centre, Institute of Oncology, Ljubljana, focusing on breast cancer prediction	285	10	2,35
Heart Failure	This dataset contains the medical records of 299 patients	299	13	2,11

Table 1 – Description of the Experimental Data Used

We performed some preprocessing steps on the datasets, firstly with the division of the dataset into train (80%) and test (20%) subsets. The stratified *k-fold* cross-validation method was used to perform this splitting, where $k = 5$. The rationale behind employing Stratified K-Fold is rooted in its ability to mitigate potential biases in classification tasks arising from disparate class proportions (Olamendy, 2024). It ensures that the distribution of classes remains consistent across the training and test sets, thereby enhancing the reliability and generalizability of the model's performance metrics.

Furthermore, this also ensures the isolation of the test data from the training data, precluding inadvertent leakage of information that could compromise the integrity of the model evaluation process. For instance, utilizing information from the test dataset to impute missing values in the training dataset may introduce biases, thereby skewing the outcomes.

In addition to the partitioning of the dataset, some data cleaning procedures were undertaken: the normalization of numerical data, encoding of categorical features, and handling of missing values and outliers. Appendix A – Table 7 describes the problem encountered by each dataset, and Appendix A - Table 8 provides the data preprocessing steps undertaken for each dataset.

3.2. EVALUATION METRICS

To evaluate the efficacy of the algorithms employed in the experiments, the F-score and AUC were considered as the main metrics, even though the accuracy and the geometric mean were also computed in the analysis (Bekkar et al., 2013; Japkowicz, 2013; Luque et al., 2019).

3.2.1. Confusion Matrix

A confusion matrix is a tabular representation that is used to assess the performance of a classification algorithm (Ohsaki et al., 2017). The confusion matrix is a representation where each row corresponds to the examples in a specific real class, and each column corresponds to the instances in a specific predicted class. The confusion matrix consists of four primary components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

Using the confusion matrix, we can derive various metrics to evaluate the performance of a model more comprehensively. Sensitivity or Recall quantifies the model's capacity to accurately detect instances that are truly positive (Powers, 2020).

$$Sensivity = Recall = \frac{TP}{TP + FN}$$

Specificity assesses the model's ability to correctly identify true negative cases. It answers the question: "Of all the actual negative cases, how many were correctly identified by the model?" (Powers, 2020).

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

Finally, precision evaluates the model's ability to accurately identify true positives among all positive predictions: "Of all the cases predicted as positive, how many are actually positive?" (Powers, 2020).

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

These metrics provide valuable insights into the model's performance and are essential for calculating other important metrics, such as the F1 score and accuracy, which will be discussed further (Jeni et al., 2013; Thabtah et al., 2020).

While in scenarios like spam email detection or sentiment analysis, a false positive or false negative may result in inconvenience or annoyance, in the medical domain, the consequences of wrong-classified instances can be more severe and even cost a life (Kault, 2014; Kumaravel & Vijayan, 2023). False negative instances occurs when a disease is incorrectly classified as absent when it is present (Chubak et al., 2010). This can lead to delayed or missed treatment, potentially allowing the condition to worsen or progress unchecked. In some cases, such as cancer screenings, a false negative diagnosis can mean a missed opportunity for early intervention, which can significantly impact prognosis and survival rates.

Conversely, a false positive occurs when a condition is wrongly predicted as present when it is absent. While false positives may lead to unnecessary follow-up tests, procedures, or treatments, they are generally considered less harmful than false negatives in the medical context.

As such, minimizing errors and optimizing the balance between sensitivity and specificity in medical diagnostics is crucial for ensuring patient safety and optimal healthcare outcomes (Lafreniere et al., 2021).

3.2.2. F-Score

The F-score - or F1 score - considers both precision and recall, providing a balanced measure that considers both false positives and false negatives (Powers, 2020). It is particularly useful when there is a class imbalance as it gives equal weight to both precision and recall, making it a more reliable metric for assessing model performance in such scenarios (Wardhani et al., 2019).

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

In the context of evaluating classification performance on imbalanced data, using the F1 score may be more appropriate than accuracy. This is because accuracy can be misleading

when dealing with imbalanced datasets where one class significantly outnumbers the other (Gu et al., 2009; Japkowicz, 2013).

The F1 score is a comprehensive metric that takes into account both precision and recall, offering a well-balanced evaluation that considers both false positives and false negatives. It is especially advantageous in situations where there is an unequal distribution of classes, as it assigns equal importance to both precision and recall, making it a more reliable metric for assessing model performance in such scenarios (Thabtah et al., 2020).

Therefore, in situations where imbalanced data is present, it is generally recommended to prioritize the F1 score over accuracy for a more accurate evaluation of classification performance (Jeni et al., 2013).

3.2.3. Area Under the ROC Curve

The Area Under the Curve (AUC) measures the trade-off between true positive rate and false positive rate, offering insights into the classifier's performance at various operating points. This metric is considered robust to class imbalance and is not affected by changes in class distribution, making it a reliable metric for evaluating classifiers on imbalanced datasets (Bekkar et al., 2013; Jeni et al., 2013).

3.3. CLASSIFIERS

The experimental procedure employed different models to ensure that the results were not affected by the ML algorithm. The classifiers used were the Logistic Regression (LR) (McCullagh & Nelder, 1989), K-Nearest Neighbor (KNN) (Cover & Hart, 1967), Random Forest (RF) (Ho, 1995) and Gradient Boosting (GB) (Friedman, 2001).

3.4. EXPERIMENTAL PROCEDURE

Similar to the methodology used in Lechleitner (2020) and in Douzas & Bacao (2019), the experimental procedure used in this study starts by random undersampling each dataset with different undersampling ratios: 50%, 75%, 90% and 95%. Meaning, at this point, that each dataset will have 5 variants of itself – see Table 2.

Dataset	Undersampling Ratio	Percentage of data Kept	Meaning
Variant 1	0%	100%	No oversampling will be applied to this dataset variant.
Variant 2	50%	50%	Oversampling will be applied to 50% of the original data.
Variant 3	75%	25%	Oversampling will be applied to 25% of the original data.
Variant 4	90%	10%	Oversampling will be applied to 10% of the original data.
Variant 5	95%	5%	Oversampling will be applied to 5% of the original data.

Table 2 – Variants of the datasets after performing undersample

Secondly, different oversampling techniques are used to oversample the datasets back to their original size (Douzas et al., 2022; Lechleitner, 2020). It is important to highlight the particularity of this research: the oversampling will be applied to both minority and majority classes, to try to replicate the original dataset and assess the quality of the artificially generated data. This means that, in the end, all datasets must have the same number of total observations the original dataset had before applying the undersampling. More specifically, this also means that each dataset must have the same number of majority and minority instances as before applying the undersampling. Taking Table 2 as a reference, Variant 5 of the dataset will only take 5% of the original data as a basis for trying to recreate the original dataset, meaning that it will have to generate 95% artificial data to try to recreate the original dataset. On the other hand, Variant 2 of the dataset will take 50% of the original data as a reference to build the original data, having to generate 50% of the data.

To evaluate the performance of each combination of undersampling ratio, a resampler method and classifier, n-fold cross-validation with $n = 5$ was applied to aim to maximize the F-Score. It is important to note that a range of hyper-parameters was used for the classifiers during the training phase. Hyperparameter tuning was performed using grid search, and, despite the limited search space for this component of the experimental procedure, the

hyperparameters for each classifier were identical, ensuring consistency among ratios and oversampling. After completing this process and applying the classifiers mentioned in section 3.3 Classifiers to each dataset, the results were computed to compare them. Each resampling approach was assigned a ranking score ranging from 1 to 6, with the best method earning a score of 1 and the worst method obtaining a score of 6.

The Friedman test was performed to determine the statistical significance of the observed variations in performance, followed by Holm's correction to modify the *p-values*. These methods aid in evaluating the magnitude of performance disparities and guaranteeing reliable conclusions regarding the efficacy of each resampling technique in enhancing classifier performance on imbalanced datasets (Hoffman, 2015). Figure 6 presents a visualization of a high-level explanation of the experimental procedure.

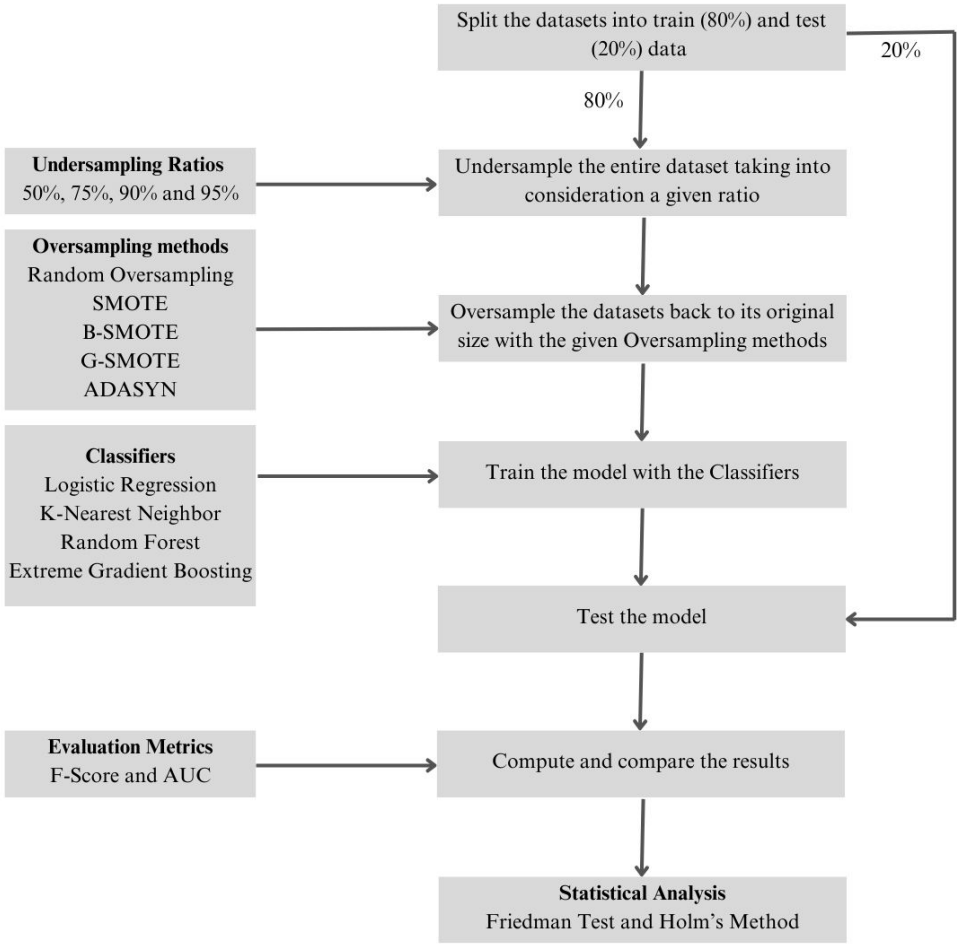

Figure 6 – Visualization of the experimental procedure (Douzas et al., 2022; Lechleitner, 2020)

As discussed in section 3.1 Experimental Data, before splitting the datasets and initiating the undersampling process, the data was preprocessed to ensure it was suitable for ML algorithms and models.

This method allows a direct comparison between the quality of the artificial generated sample set and the observed data, concerning the F-Score and AUC for each combination of undersample ratio, oversampling method, and classifier.

The oversampling methods to be compared to the original datasets (no oversampling) are random oversample, SMOTE, B-SMOTE, G-SMOTE and ADASYN.

3.5. SOFTWARE IMPLEMENTATION

The experimental methodology is implemented using the Python programming language, specifically in Jupyter Notebooks. Our primary tool for data analysis and general machine learning tasks was the Scikit-Learn library. In addition, data preprocessing was performed using the numpy and pandas packages. The imbalanced-learn library was utilized for the tested methods. Scipy and statsmodels were used for the statistical analysis. The visualization was conducted with Plotly.

4. RESULTS AND DISCUSSION

This section presents the performance of the different oversampling methods. The results are examined by employing diverse statistical tests to determine their relevance.

4.1. COMPARATIVE PRESENTATION

The mean cross-validation scores for each undersampling ratio, classifier and metric are shown across all datasets in Table 3. The scores are presented for each undersampling ratio to understand how the methods perform as the dataset size gets smaller (Lechleitner, 2020). The highest scores for each row are highlighted.

Ratio	Classifier	Metric	NONE	ROS	SMOTE	B-SMOTE	G-SMOTE	ADASYN
50	LR	F-Score	0.458	0.484	0.567	0.558	0.532	**0.585**
50	LR	AUC	0.594	0.586	**0.660**	0.632	0.619	0.652
50	KNN	F-Score	0.533	0.557	**0.687**	0.597	0.567	0.611
50	KNN	AUC	0.624	0.633	0.678	0.669	0.670	**0.690**
50	RF	F-Score	0.611	0.587	0.633	0.627	0.629	**0.652**
50	RF	AUC	0.687	0.656	0.724	0.697	0.702	**0.725**
50	GB	F-Score	0.574	0.563	**0.628**	0.615	0.581	0.622
50	GB	AUC	0.658	0.648	0.696	**0.712**	0.677	0.682
75	LR	F-Score	.458	0.489	0.568	0.558	0.532	**0.577**
75	LR	AUC	0.594	0.587	**0.640**	0.632	0.619	0.639
75	KNN	F-Score	0.533	0.562	0.581	0.598	0.567	**0.619**
75	KNN	AUC	0.624	0.636	0.670	0.670	0.669	**0.698**
75	RF	F-Score	0.607	0.594	0.629	**0.673**	0.593	0.655
75	RF	AUC	0.672	0.663	0.721	**0.741**	0.707	0.724
75	GB	F-Score	0.589	0.583	0.640	0.627	0.586	**0.642**
75	GB	AUC	0.677	0.640	0.712	**0.718**	0.674	0.696
90	LR	F-Score	0.458	0.492	**0.582**	0.558	0.532	0.582
90	LR	AUC	0.594	0.590	**0.644**	0.632	0.619	0.644
90	KNN	F-Score	0.521	0.560	0.597	0.591	0.559	**0.634**
90	KNN	AUC	0.624	0.635	0.672	0.664	0.662	**0.701**
90	RF	F-Score	0.625	0.598	0.621	0.614	0.589	**0.647**
90	RF	AUC	0.682	0.683	0.694	0.679	0.686	**0.710**
90	GB	F-Score	0.562	0.580	0.596	**0.642**	0.584	0.626

90	GB	AUC	0.684	0.645	0.700	**0.730**	0.679	0.687
95	LR	F-Score	0.458	0.487	0.571	0.558	0.532	**0.577**
95	LR	AUC	0.594	0.587	0.633	0.632	0.619	**0.640**
95	KNN	F-Score	0.525	0.558	**0.610**	0.585	0.563	0.609
95	KNN	AUC	0.625	0.634	0.678	0.660	0.665	**0.688**
95	RF	F-Score	0.619	0.594	**0.639**	0.634	0.609	0.637
95	RF	AUC	0.678	0.692	0.703	**0.734**	0.706	0.723
95	GB	F-Score	0.589	0.582	0.632	0.630	0.587	**0.643**
95	GB	AUC	0.675	0.639	0.707	0.716	0.681	**0.716**

Table 3 - Results for mean cross validation scores of oversamplers

The F-score values are consistently high, particularly with the ADASYN approach, but also with SMOTE. This is generally the case with ADASYN in the majority of instances. The greatest F-score achieved is 68.7% using the KNN algorithm with a 50% ratio and the SMOTE technique.

In terms of AUC scores, the ADASYN method and the B-SMOTE method both show competitive results. The B-SMOTE achieves the highest AUC score of 74.1% at a 75% ratio. ADASYN demonstrates strong performance across several classifiers and metrics, frequently achieving the highest values in F-score and AUC.

The highest values of each combination of ratio, classifier and metric significantly improved the performance of the algorithm in the NONE datasets - where no oversampling was applied to either class. This indicates that, although not achieve the highest scores, the algorithms effectively improved the performance of the classifier.

Even though this is not a definitive rule, the performance of oversampling methods tends to improve with the increase of the dataset size (i.e., as the undersampling ratio declines) in the majority of cases (Lechleitner, 2020). This is evident from the generally higher scores at 95% ratios compared to 50%.

To each oversampling method, including the special case where no oversampling (NONE) was applied, a ranking score ranging from 1 to 6 was assigned. Table 4 displays the mean ranking of these methods across all datasets, grouped by classifier and metric.

Classifier	Metric	NONE	ROS	SMOTE	B-SMOTE	G-SMOTE	ADASYN
LR	F-Score	6	5	1.75	3	4	1.25
LR	AUC	5	5.75	2.5	3	3.5	2.25
KNN	F-Score	6	4.75	1.75	2.75	4.25	1.5
KNN	AUC	5.25	5.25	2.75	3	3.25	1.5
RF	F-Score	3.75	5.5	2.25	3	5	1.5
RF	AUC	5.25	5.25	2.75	3	3.25	1.5
GB	F-Score	4.5	5.5	1.75	2.5	5	1.75
GB	AUC	4.75	6	2	1.75	4.25	2.25

Table 4 – Results for mean rankings per classifier

ADASYN is the highest-ranking technique when comparing the rest of the evaluated methods, except for one case: the AUC metric with the GB classifier. In this case, B-SMOTE method outperforms ADASYN for the GB classifier.

Additionally, Figure 7 visually plots the mean ranks for the F-Score metric, grouped by classifier and ratio, allowing us to observe how classifier behavior changes as the ratio varies. By analyzing this figure, we can see that ADASYN notably outperforms the LR and KNN classifiers in almost every ratio, except where it is surpassed by the SMOTE algorithm at the 90% ratio for LR and the 50% and 90% ratios for KNN. In these two classifiers, the undersampled datasets and the randomly oversampled datasets rank in the last two spots. Regarding the RF classifier, ADASYN outperforms at ratios of 50% and 90%, while B-SMOTE and SMOTE outperform at ratios of 75% and 95%, respectively. For the GB classifier, ADASYN outperforms at a 95% ratio, B-SMOTE at a 90% ratio, while SMOTE outperforms at ratios of 50% and 75%.

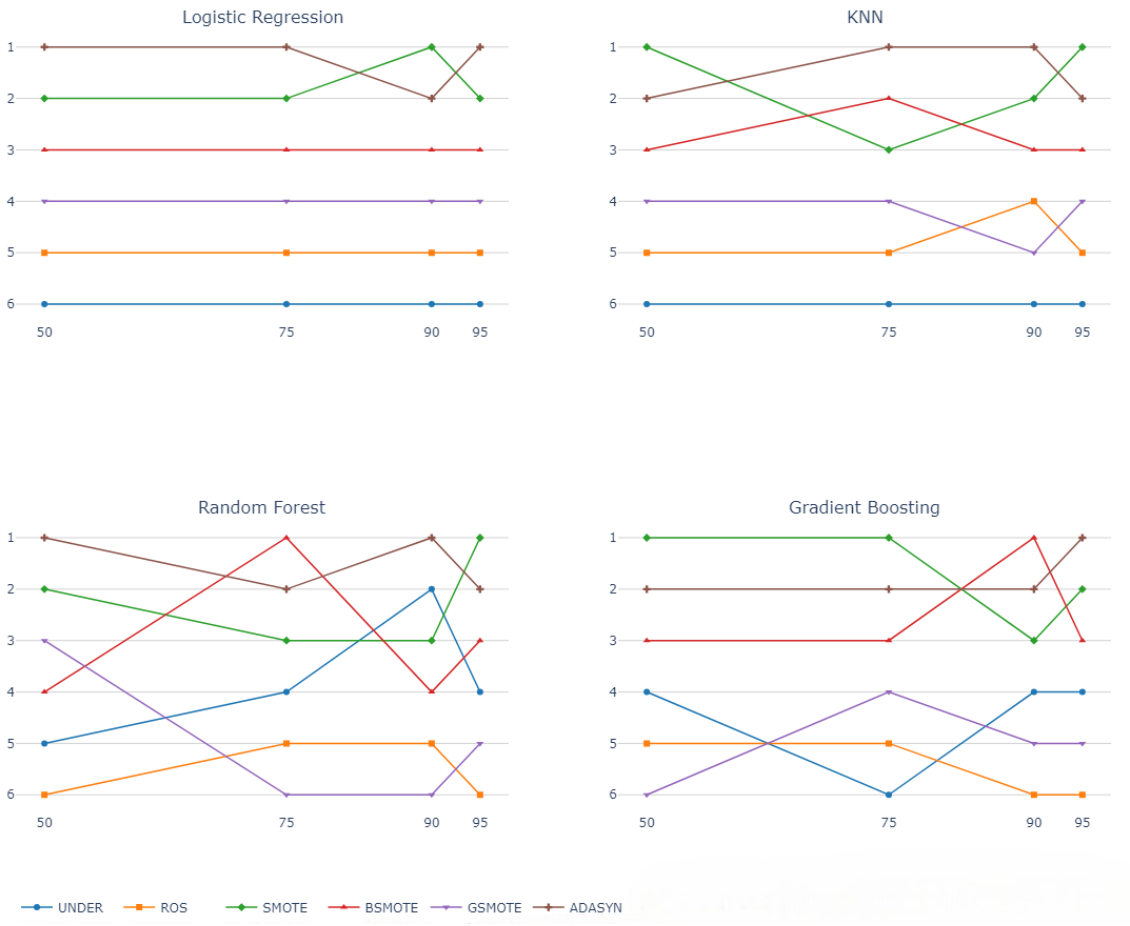

Figure 7 – Mean ranking per classifier (F-Score)

Figure 8 plots the mean ranking for the AUC, also grouped by classifier and ratio. From this plot, we can infer that ADASYN outperformed for the KNN and RF classifiers at ratios of 50%, 90%, and 95% for the former, and 50% and 90% for the latter. For KNN, the second-best algorithm was SMOTE, which outperformed at the 75% ratio. Regarding RF, the second-best classifier was B-SMOTE. For LR, SMOTE outperformed in half of the ratios, while ADASYN and G-SMOTE outperformed at the 75% and 90% ratios. For the GB classifier, B-SMOTE outperformed at the 50% and 90% ratios, while SMOTE and ADASYN outperformed at the remaining ratios. Notably, the random oversampling and undersampled methods consistently ranked in the last two places.

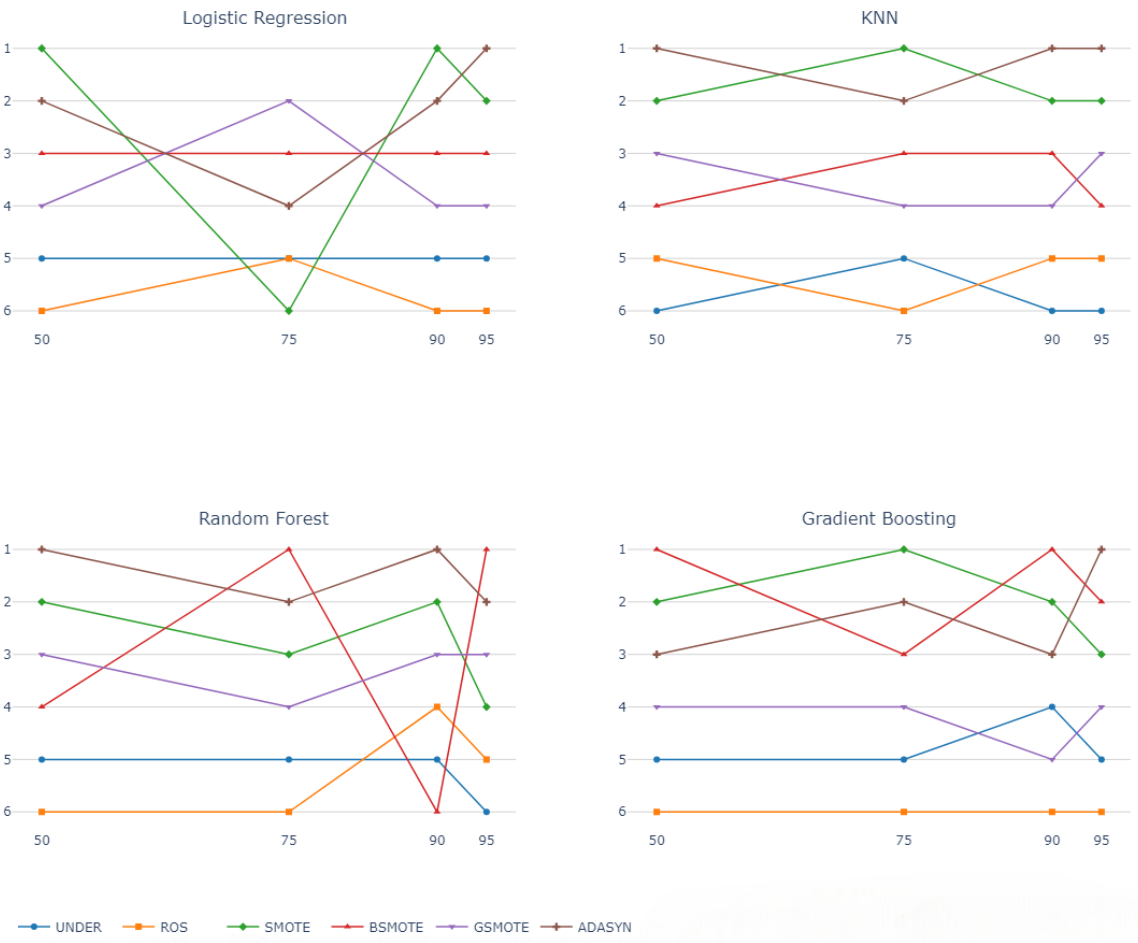

Figure 8 - Mean ranking per classifier (AUC)

4.2. STATISTICAL ANALYSIS

To confirm the significance of the presented results, both the Friedman and the Holm Tests were employed. The application of the Friedman test is presented on Table 5.

Classifier	Metric	*p-value*	Significance
LR	F-Score	2.1e-06	True
LR	AUC	6.0e-11	True
KNN	F-Score	3.9e-05	True
KNN	AUC	3.9e-10	True
RF	F-Score	0.008	True
RF	AUC	1.1e-06	True
GB	F-Score	0.0004	True
GB	AUC	2.6e-06	True

Table 5 - Results for the Friedman test

The null hypothesis of the Friedman test is rejected with a significance level of $\alpha = 0.05$. In terms of the mean ranking across all classifiers and assessment measures, this indicates that the synthetic data generation approaches do not perform similarly.

Holm's method is implemented to modify the *p-values* of the paired difference test, with the ADASYN algorithm serving as the control method, as this was the method that presented the best overall results (Dubitzky et al., 2013). The results are shown in Table 6.

Classifier	Metric	NONE	ROS	SMOTE	B-SMOTE	G-SMOTE
LR	F-Score	**1.5e-03**	**3.5e-03**	1.4e-01	**5.4e-04**	**1.3e-03**
LR	AUC	**1.4e-03**	**3.1 e-04**	9.5e-01	**2.1e-03**	**2.0e-03**
KNN	F-Score	**2.03e-04**	**1.09e-04**	6.2e-02	**1.2e-02**	**2.2e-03**
KNN	AUC	**7.2e-06**	**1.1e-07**	9.953-01	**6.8e-04**	**1.2e-02**
RF	F-Score	3.3e-01	**2.5e-04**	2.03-01	4.2e-01	**2.9e-02**
RF	AUC	**1.2e-03**	**2.3e-05**	4.1-01	4.1e-01	7.9e-02
GB	F-Score	**8.8e-04**	**5.5e-04**	1	1	**1.3e-03**
GB	AUC	5.1e-02	**8.1e-04**	4.7-01	2.4e-01	2.2e-01

Table 6 – Adjusted *p-values* using Holm's method

At a significance level of $\alpha = 0.05$, the null hypothesis of Holm's test is rejected for 24 of the 40 combinations. These results demonstrate that the ADASYN approach outperforms the other methods in most cases.

5. CONCLUSIONS

Many sectors and applications, including the medical industry, still face limitations in utilizing small datasets. The inadequate quantity of training data often leads to inferior performance of ML algorithms.

This study compares different techniques to overcome the small data problem in medicine-related binary classification tasks. As seen in the previous section, both the ADASYN and SMOTE algorithms possess the capability to generate artificial data of average quality. The quality of the generated data was suboptimal but not extremely poor, with an F-Score reaching up to 68.7% and an AUC score of 74.1%. Achieving high scores is advantageous for addressing health and medical issues, as it plays a significant role in the healthcare sector.

The highest values for each combination of ratio, classifier, and metric resulted in a significant improvement in the algorithm's performance compared to the NONE datasets. More specifically, this demonstrates that the ADASYN algorithm was successful in improving the classifier's performance, despite not achieving the highest possible scores during tests. This method overall performed the best.

In some cases, it may be appropriate to apply the ADASYN algorithm to generate artificial samples and improve classifier performance. Specifically in this study, this approach demonstrated its efficacy in scenarios where the dataset was particularly small and the minority class was significantly underrepresented of medial domain. ADASYN was effective when the classifier struggled with decision boundaries due to sparse data regions, as it adaptively altered the distribution of synthetic samples based on the learning difficulty of minority class instances.

6. LIMITATIONS AND FUTURE WORK

While this study demonstrates that ADASYN can enhance the performance of classifiers in binary classification tasks related to the medical field, it is important to note that the generated artificial data did not always yield the highest possible results. Additionally, this study primarily focused on the general efficacy of the ADASYN algorithm without extensively exploring the specific contexts or medical conditions where its application would be most beneficial.

Future work should aim to identify specific cases in the medical field where the application of the ADASYN algorithm to generate artificial samples would be particularly advantageous. Research could focus on the following areas:

- Clinical Data Scenarios: Exploring various clinical scenarios where data is scarce, such as early-stage clinical trials or niche patient demographics, to determine if ADASYN can effectively enhance predictive modelling in these contexts.
- Comparative Studies: Conducting comparative studies to benchmark ADASYN against other data generation techniques across different medical datasets to better understand its relative strengths and weaknesses.
- Quality *versus* Quantity Trade-Off: Further examining the trade-off between the diversity of synthetic samples generated by ADASYN and the overall quality of these samples, particularly in high-stakes medical applications where precision is crucial.

By focusing on these areas, future research can better delineate the scenarios where ADASYN is most effective, thereby providing more targeted recommendations for its application in the healthcare sector. This will enable practitioners to make informed decisions about when to use ADASYN for artificial data generation, recognizing that while it may not always produce the highest possible results, it can still offer meaningful performance improvements in specific contexts.

BIBLIOGRAPHICAL REFERENCES

Abdul Lateh, M., Muda, A. K., Izzah Mohd Yusof, Z., Azilah Muda, N., & Sanusi Azmi, M. (2017a). Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review. *Journal of Physics: Conference Series*, *892*(1). https://doi.org/10.1088/1742-6596/892/1/012016

Abdul Lateh, M., Muda, A. K., Izzah Mohd Yusof, Z., Azilah Muda, N., & Sanusi Azmi, M. (2017b). Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review. *Journal of Physics: Conference Series*, *892*(1). https://doi.org/10.1088/1742-6596/892/1/012016

Bekkar, M., Djema, H., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, *3*, 27–38.

Brandt, J., & Lanzén, E. (2020). *A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification*.

Chao, G. Y., Tsai, T. I., Lu, T. J., Hsu, H. C., Bao, B. Y., Wu, W. Y., Lin, M. T., & Lu, T. L. (2011). A new approach to prediction of radiotherapy of bladder cancer cells in small dataset analysis. *Expert Systems with Applications*, *38*(7), 7963–7969. https://doi.org/10.1016/j.eswa.2010.12.035

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chongfu, H. (1997). Principle of information diffusion. In *Fuzzy Sets and Systems* (Vol. 91).

Chubak, J., Boudreau, D. M., Fishman, P. A., & Elmore, J. G. (2010). Cost of breast-related care in the year following false positive screening mammograms. *Medical Care*, *48*(9), 815–820.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Dholakiya, P. (2023, April 26). *SMOTE(Synthetic Minority Over-sampling Technique)*. https://medium.com/@parthdholakiya180/smote-synthetic-minority-over-sampling-technique-4d5a5d69d720

Douzas, G., & Bacao, F. (2019). Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. *Information Sciences*, *501*, 118–135. https://doi.org/10.1016/j.ins.2019.06.007

Douzas, G., Lechleitner, M., & Bacao, F. (2022). Improving the quality of predictive models in small data GSDOT: A new algorithm for generating synthetic data. *PLoS ONE*, *17*(4 April). https://doi.org/10.1371/journal.pone.0265626

Dubitzky, W., Wolkenhauer, O., Yokota, H., & Cho, K.-H. (2013). *Encyclopedia of Systems Biology*. https://doi.org/10.1007/978-1-4419-9863-7_708

Efron, Bradley., & Tibshirani, Robert. (1994). *An introduction to the bootstrap*. Chapman & Hall.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. In *Learning from Imbalanced Data Sets*. Springer International Publishing. https://doi.org/10.1007/978-3-319-98074-4

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gomede PhD, E. (2024). *ADASYN algorithm for unbalanced classification problems*. https://pub.aimind.so/adasyn-algorithm-for-unbalanced-classification-problems-4e0b08e83bd7

Goswami, S. (2020, November 2). *Class Imbalance, SMOTE, borderline SMOTE, ADASYN*. https://towardsdatascience.com/class-imbalance-smote-borderline-smote-adasyn-6e36c78d804

Goswami, S. (2021). *Class Imbalance, SMOTE, borderline SMOTE, ADASYN - Towards Data Science*. https://towardsdatascience.com/class-imbalance-smote-borderline-smote-adasyn-6e36c78d804

Gu, Q., Zhu, L., & Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. *Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings 4*, 461–471.

Guggenheim, D. (2022, June 15). *The Mystery of ADASYN is Revealed - Towards Data Science*. https://towardsdatascience.com/the-mystery-of-adasyn-is-revealed-73bcba57c3fe

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *LNCS*, *3644*, 878–887.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008a). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks*, 1322–1328.

He, H., Bai, Y., Garcia, E., & Li, S. (2008b). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *Proceedings of the International Joint Conference on Neural Networks*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Hoffman, J. I. E. (2015). Chapter 26 - Analysis of Variance II. More Complex Forms. In J. I. E. Hoffman (Ed.), *Biostatistics for Medical and Biomedical Practitioners* (pp. 421–447). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-802387-7.00026-3

Huang, C., & Moraga, C. (2004). A diffusion-neural-network for learning from small samples. *International Journal of Approximate Reasoning*, *35*(2), 137–161. https://doi.org/10.1016/j.ijar.2003.06.001

Ivănescu, V. C., Bertrand, J. W. M., Fransoo, J. C., & Kleijnen, J. P. C. (2006). Bootstrapping to solve the limited data problem in production control: An application in batch process industries. *Journal of the Operational Research Society*, *57*(1), 2–9. https://doi.org/10.1057/palgrave.jors.2601966

Japkowicz, N. (2013). Assessment metrics for imbalanced learning. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 187–206.

Javaid, M., Haleem, A., Pratap Singh, R., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, *3*, 58–73. https://doi.org/10.1016/j.ijin.2022.05.002

Jeni, L., Cohn, J., & De la Torre, F. (2013). Facing Imbalanced Data - Recommendations for the Use of Performance Metrics. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013* (Vol. 2013). https://doi.org/10.1109/ACII.2013.47

Kault, D. (2014). False negatives in evidence based medicine. *Journal of Medical Statistics and Informatics*, *2*.

Kumaravel, A., & Vijayan, T. (2023). Comparing cost sensitive classifiers by the false-positive to false-negative ratio in diagnostic studies. *Expert Systems with Applications*, *227*, 120303.

Lafreniere, B., R. Jonker, T., Santosa, S., Parent, M., Glueck, M., Grossman, T., Benko, H., & Wigdor, D. (2021). False positives vs. false negatives: The effects of recovery time and

cognitive costs on input error preference. *The 34th Annual ACM Symposium on User Interface Software and Technology*, 54–68.

Lechleitner, M. (2020). *Small Data Oversampling Improving small data prediction accuracy using the Geometric SMOTE algorithm*. Universidade NOVA de Lisboa.

Li, D. C., Chen, L. S., & Lin, Y. S. (2003). Using functional virtual population as assistance to learn scheduling knowledge in dynamic manufacturing environments. *International Journal of Production Research*, *41*(17), 4011–4024. https://doi.org/10.1080/0020754031000149211

Li, D. C., Lin, W. K., Chen, C. C., Chen, H. Y., & Lin, L. S. (2018). Rebuilding sample distributions for small dataset learning. *Decision Support Systems*, *105*, 66–76. https://doi.org/10.1016/j.dss.2017.10.013

Li, D. C., & Wen, I. H. (2014). A genetic algorithm-based virtual sample generation technique to improve small data set learning. *Neurocomputing*, *143*, 222–230. https://doi.org/10.1016/j.neucom.2014.06.004

Li, D. C., Wu, C. Sen, Tsai, T. I., & Lina, Y. S. (2007). Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Computers and Operations Research*, *34*(4), 966–982. https://doi.org/10.1016/j.cor.2005.05.019

Li, D.-C., Wu, C.-S., Tsai, T.-I., & Chang, F. M. (2006). Using mega-fuzzification and data trend estimation in small data set learning for early FMS scheduling knowledge. *Computers & Operations Research*, *33*, 1857–1869. https://doi.org/10.1016/S0305-0548(04)00324-7

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/https://doi.org/10.1016/j.patcog.2019.02.023

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed., Vol. 37). Chapman and Hall.

Niyogi, P., Girosi, F., & Poggio, T. (1998). *Incorporating Prior Information in Machine Learning by Creating Virtual Examples*.

Ohsaki, M., Wang, P., Matsuda, K., Katagiri, S., Watanabe, H., & Ralescu, A. (2017). Confusion-Matrix-Based Kernel Logistic Regression for Imbalanced Data Classification. *IEEE Transactions on Knowledge and Data Engineering*, *29*(9), 1806–1819. https://doi.org/10.1109/TKDE.2017.2682249

Olamendy, J. (2024, April 2). *A Comprehensive Guide to Stratified K-Fold Cross-validation for Unbalanced Data*.

Peláez-Rodríguez, C., Pérez-Aracil, J., Fister, D., Prieto-Godino, L., Deo, R. C., & Salcedo-Sanz, S. (2022). A hierarchical classification/regression algorithm for improving extreme wind speed events prediction. *Renewable Energy*, *201*, 157–178. https://doi.org/10.1016/j.renene.2022.11.042

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint ArXiv:2010.16061*.

Rubinger, L., Gazendam, A., Ekhtiari, S., & Bhandari, M. (2023). Machine learning and artificial intelligence in research and healthcare. *Injury*, *54*, S69–S73. https://doi.org/10.1016/j.injury.2022.01.046

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429–441.

Tonge Buradkar, V., & More, M. (2020). *Introduction to Machine Learning and Its Applications: A Survey*. https://doi.org/10.5281/zenodo.3775092

Tsai, C. H., & Li, D. C. (2015). Improving knowledge acquisition capability of M5' model tree on small datasets. *Proceedings - 3rd International Conference on Applied Computing and Information Technology and 2nd International Conference on Computational Science and Intelligence, ACIT-CSI 2015*, 379–386. https://doi.org/10.1109/ACIT-CSI.2015.72

Tsai, T. I., & Li, D. C. (2008). Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems. *Expert Systems with Applications*, *35*(3), 1293–1300. https://doi.org/10.1016/j.eswa.2007.08.043

UC Irvine Machine Learning Repository. (n.d.).

Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., & Lestantyo, P. (2019). Cross-validation metrics for evaluating classification performance on imbalanced data. *2019 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, 14–18.

APPENDIX A

Dataset	Categorical Features	Binary Features	Numerical Features	Duplicates	Missing Values
Stroke	Yes	Yes	Yes	No	Yes
Heart Disease	Yes	Yes	Yes	No	No
Hepatitis C	Yes	No	Yes	No	Yes
Breast Cancer	No	No	Yes	No	No
Indian Liver Patient	Yes	No	Yes	Yes	Yes
Cirrhosis	Yes	No	Yes	No	Yes
Thoracic Surgery	Yes	No	Yes	No	No
Autism	Yes	Yes	Yes	Yes	Yes
Cryotherapy	No	Yes	Yes	No	No
Immunotherapy	No	Yes	Yes	No	No
Caesarian	No	Yes	Yes	Yes	No
Jugoslavia Breast Cancer	Yes	No	Yes	No	No
Heart Failure	No	Yes	Yes	No	No

Table 7 – Data Quality Assessment

Dataset	Target	Normalization	Encoding	Missing Values	Outliers
Stroke	1 if the patient had a stroke and 0 if not	MM[1]	Label and One-hot Encoder	Median Imputation	IQR
Heart Disease	1 if the patient has a heart disease and 0 if not	MM	Label and One-hot Encoder	N/A	Manual Filtering
Hepatitis C	Grouped "Cirrhosis" and "Fibrosis" under the "Hepatitis C" category; and "Suspect Blood Doner" into the "Blood Doner" category	MM	Label Encoder	Median Imputation	Manual Filtering
Breast Cancer	1 if it's a healthy control and 2 if it's patients	MM	N/A	N/A	Manual Filtering
Indian Liver Patient	1 if the patient has liver disease and 2 if it doesn't	MM	Label Encoder	Median Imputation	Manual Filtering
Cirrhosis	Grouped the "C" (Censored) and the "CL" (Censored due to liver transplant) into class 1	MM	Label and One-hot Encoder	Median Imputation	Kept
Thoracic Surgery	1-year survival period: T if the patient died; F otherwise	MM	Label and One-hot Encoder	N/A	Manual Filtering
Autism	Class Autism Spectrum Disorder: Yes, No	MM	Label and One-hot Encoder	Median Imputation	N/A
Cryotherapy	Result of treatment: 1 or 0	MM	Label and One-hot Encoder	N/A	N/A
Immunotherapy	Result of treatment: 1 or 0	MM	Label and One-hot Encoder	N/A	N/A
Caesarian	Caesarean: 1 or 0	MM	N/A	N/A	N/A
Jugoslavia Breast Cancer	Class: no recurrence events and recurrence events	MM	Label and One-hot Encoder	Median Imputation	N/A
Heart Failure	1 if the patient died, 0 otherwise	MM	Label Encoder	N/A	N/A

Table 8 – Data Preprocessing Applied

[1] MinMax Scaler

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa