Evaluation of Machine Learning classification techniques for handling class imbalance in medical datasets

Olga Margarita Minguett Pirela
20179766

Faculty of Science and Engineering
Department of Computer Science and Information Systems
University of Limerick

Submitted to the University of Limerick for the degree of
Master in Artificial Intelligence (MSc)

May 2022

1. Supervisor (s): Tiziana Margaria, Prof. Dr.
University of Limerick
Ireland

Evaluation of Machine Learning classification techniques for handling class imbalance in medical datasets

Abstract

Imbalanced class distribution in datasets occurs when one of the classes, usually the most interesting class in the dataset, also described as the positive or minority class, is not accurately characterized and likely considered noise or outlier. Having said that, in the artificial intelligence field, What type of Machine Learning classification techniques are most suitable for dealing with imbalanced classes in medical datasets? What are the appropriate data augmentation techniques that can help to emulate real-world medical data? How can we determine whether a medical dataset is imbalanced versus unequal or unknown?

The methodology included a literature review that compiles the analysis of the relevant information provided by other researchers in this area. Secondly, we tested classification algorithms such as Support Vector Machine (SVM), Decision Tree, Gaussian Naïve Bayes (GNB), K-Nearest Neighbourhood (KNN), and Logistic Regression (LR), handling imbalanced medical datasets by simulating oversampling and undersampling use cases. Finally, the implementation had minimum hyper-parameterization, and performance classification metrics measured results.

The results displayed that generally, oversampling techniques performed the best for any percentage of imbalance property in the dataset, where Random OverSampler tends to go along with Decision Trees and Gaussian Naive Bayes algorithms. On the other hand, the hybrid technique SMOTE Tomek performed the greatest accompanied by Logistic Regression and Decision Tree algorithms.

Keywords: Imbalanced Class Distribution, Classification Machine Learning Algorithms, Medical Dataset, resampling techniques.

Acknowledgements

I would like to express my deepest appreciation to my research supervisor, Prof. Dr Tiziana Margaria, for providing guidance and feedback throughout this project. Our meetings and conversations we had were vital in inspiring me to finish my dissertation by working on multiple perspectives to form a comprehensive and objective result.

Special thanks to the University of Limerick for providing such an incredible platform for learning through great lecturers and material.

This endeavour would not have been possible without my classmates; I appreciate your advice and comments in every activity, and support during these two years.

I am also grateful to Optum, which involves my managers, director of my team, and colleagues, for assisting in the realization of this course.

To conclude, I cannot forget to thank my husband, parents, siblings and friends for all the unconditional support in this very intense academic undertaking; I could not have made this journey without their support and words of encouragement.

Dedication

To my dearest and extraordinary husband,
Nessan J. O'Donovan.
We did it!

Glossary

ACC: Accuracy
AUC: Area Under the Curve
DT: Decision Tree
EHT: Electronic Health Record
EMR: Electronic Medical Record
FP: False Positive
FN: False Negative
GNB: Gaussian Naïve Bayes
KNN: K-Nearest Neighbourhood
LR: Logistic Regression
ML: Machine Learning
NN: Nearest Neighbour
PBR: Paper-Based Record
ROC: Receiver Operating Characteristic
SMOTE: Synthetic Minority Over-sampling Technique
SVM: Support Vector Machine
TP: True Positive
TN: True Negative

Table of Contents

List of Tables

List of Figures

Introduction

"With data collection, 'the sooner the better' is always the best answer."

– Marissa Mayer

Background / Problem Statement

Imbalanced class distribution in datasets takes place when one of the classes, usually the most interesting class in the dataset, also described as the positive or minority class, is not accurately characterized. That means the number of classes is not balanced (Ganganwar, V., 2012). As the positive class is not well represented, they are likely considered noise or outlier. As a result, they are ignored, unexplored and misclassified. Imbalanced classifications pose a challenge for predictive modelling as most machine learning algorithms used for variety were designed to assume an equal number of examples for each class. (Brownlee, 2019)

Positive or smaller classes hold more importance and interest, bringing more attention and demanding recognition. For instance, the medical diagnosis of a rare disease is critical in so many ways (Thabtah et al.,2020). Starting from its identification within the normal population, delineate the evolution of that disease in sick patients, define proper medication and treatment, and Population health management. Resulting in massive assistance and help to doctors and healthcare professionals that can provide better care for their patients.

The considerable demands of the class imbalance problem and its recurrent frequency in practical applications of pattern recognition and data mining have engrossed many researchers lists of studies (Liu,2021). As a result, some areas or fields have started to understand much better how imbalance distribution takes place within their domains and why it is common, growing research with more emphasis in recent years. Including ranges like anomaly detection (Omar et al., 2013), email spam detection (Dada et al.,2019), credit card fraud detection (Awoyemi et al., 2017) and early detection of some diseases like cancer and diabetes (Liu et al., 2020. pp.171263-171280).

Research Goals / Research Question

The three inter-related goals of this research are:

- *To evaluate Machine Learning classification techniques to deal with imbalanced class distributions in medical datasets.*
- *To assess class rebalances strategies' appropriateness depending on the level/percentage of imbalanced class distributions in medical datasets.*

- *To create a reference guide to determine how to use Machine Learning classification techniques and class rebalances strategies depending on the level/percentage of imbalanced class distributions in medical datasets.*

Furthermore, the development of the objectives mentioned above is accompanied by the research questions that will guide us to achieve those goals.

- *What type of Machine Learning classification techniques are most suitable for dealing with imbalanced classes in medical datasets?*
- *What are the appropriate data augmentation techniques that can help to emulate real-world medical data?*
- *How can we determine a medical dataset is imbalanced versus unequal or unknown?*

In contrast, a result will create a classification framework for imbalanced medical datasets, objectives, and rationale.

- *Abled to train machine learning classification algorithms with imbalanced medical datasets to determine performance as a baseline.*
- *Used the output obtained for the previous model training and asses class rebalance strategies to improve performance*
- *Expected an information sheet with explanations about the performance of the different machine learning models concerning class rebalance strategies to determine preferably and recommended level depending on the problem to solve.*

Research Structure

To initially reflect feedback elicit and research planning, to begin with, this dissertation will be structured as follows.

The first section, Introduction, already defined the background and context of the problem intended to be researched, outlined the problem statement and delineated the research goals and questions.

The second section, Literature Review, is divided into:
Chapter I: Classification in Imbalanced Datasets: This chapter describes imbalanced classification, definition, causes of class imbalance, why is it a problem, description of

strategies that could help in tackling the problem, including definition, characteristics, differences, and usability of the various techniques that could be used when rebalancing datasets.

Chapter II: Classification on Imbalanced Medical Data: This chapter elaborates on the definition, description, and characteristics of medical data, the importance of medical data, the impact of imbalanced medical data in classification algorithms.

Chapter III: Classification Algorithms: This chapter is about what is classification in machine learning, definition, algorithms used for classification, tasks, classification metrics, and comparison between algorithms previously defined

Literature Review

"Predicting the future isn't magic, it's artificial intelligence."

– Dave Waters

Chapter I: Classification in Imbalanced Datasets

Class Imbalance: the definition

The definition of class imbalance is, to some extent, straightforward, and according to Luque et al. (2019) is the total number of positive classes found in the data is smaller than the total number of negative classes in a dataset with an unequal proportion of observations. For example, a dataset is imbalanced when one type is represented by a few training examples described as the minority class while other classes compensate the majority (Ganganwar, V., 2012). Essentially, if the example refers to a binary classification problem, we would have more elements from one group and a few from another. However, in multi-classification situations, the observations would have diminished or clustered compared to the others (Ferreiro Volpi, G., 2019).

Furthermore, Zhao et al. in 2018 said that we could initially determine a class imbalance in our dataset when we respond to the following questions, Can we determine if the data collection process is imbalanced? If so, what is the imbalance rate? Furthermore, How are the classes divided between minority/majority? (Wang, 2014). As a matter of fact, when we describe imbalanced classification instead of "unbalanced classification", we refer to a class distribution that is fundamentally not balanced as unbalance class distribution states that at some point it was balanced and is no longer distributed in the same proportion. (Brownlee, 2020).

The most well-known cases where the class distribution is inherently imbalanced are

- credit card fraud detection, where profiles of normal and fraudulent behaviours constantly change (Awoyemi et al., 2017).
- Email spam detection, which with the increased volume of unwanted emails, filtering services represent the solution to the destructive effects on memory, communication bandwidth, CPU power, or user time that this menace could have (Dada et al.,2019).
- Anomaly detection in which regular behaviour disturbance denotes intended or unintended provoked attacks, faults, defects, and others (Omar et al., 2013).
- The early detection of diabetes or cancer accurately represents a practical approach to improving healthcare and long-term survival (Liu et al., 2018. pp.171263-171280).

The list above illustrates some examples of imbalanced classification problems that can signify entire fields of study where the intricacies of every topic can describe the complexity of this problem from diverse points of view. Therefore, they should be solved using different approaches tailored to the particular domain. (Liu,2021).

Modern machine learning techniques place massive effort to deal with the class imbalanced present in datasets, where attempts are focused on improving the threshold between accuracy and overall error for the majority class while avoiding overlooking the minority class (Thabtah et al.,2020). Moreover, Ferreiro (2019) mentioned that as the algorithms are designed to assume the same number of instances per label, no method recognises when we are under the class imbalance effects, which seriously hampers the detection of rare events. Datasets with a skewed class distribution are a common problem when solving real-world classification tasks, and they are also not clean and can be noisy (Sadawi, 2021). So being that the case, why is class imbalance a problem? Moreover, why is imbalanced classification difficult?

Class Imbalance: the problem

Due to its widespread in real-world applications, the class imbalance problem has a high amount of coverage and discussion from researchers and academics. Desuky and Hussain in 2021 said that the also known "curse of imbalanced datasets" begins when the learned classifier obtains favouritism towards the majority samples while ignoring the minority samples. According to Lemaitre et al. (2017), means the difficulty of learning a concept from the class with the small number of pieces also translates to model performance dropping significantly.

Concurring with Google Developers (2021), the imbalance properties in a dataset can be divided into 3 degrees— mild (20–40%), moderate (1–20%), and extreme (<1%). When examples are uncommonly existent, they could be classified as rare occurrences, noise, or outliers, resulting in their misclassification as a positive class compared to the negative class. Ironically, the smaller class is regularly more interested and critical; consequently, it demands identification urgency as part of the problem (Aada and Tiwari, 2019).

Following Brown and Mues (2012), besides the sample misclassification issue, another concern is that standard classifiers assumed that the domain application datasets are

equal in distributions. As such, they often fail or result in misleading outcomes. When the class distribution is not balanced, most machine learning algorithms will perform inadequately or poorly, and this behaviour involves adjustments that elude the prediction of the majority class during training time (Brownlee, 2019). In addition, evaluation metrics for classification tasks can lose their power when facing this problem, implying that alternative use methods are necessary to evaluate the predictions made (Ganganwar, 2012). Therefore, classifying the minority records appropriately with the proper techniques is of the uttermost significance while dealing with imbalance distributions. (Desuky and Hussain, 2021).

In addition to misclassification and equal distribution assumptions by the algorithms, the class imbalance is a problem characteristic on some domains because data is hard and expensive to gather, and most of the time, work needs to be done with less information than the one provided. As Badr (2019) stated, this radically influences our ability to secure representatives samples for real-world problems, so we end up with ambiguous observations on the class margin or similarly with errors in the data collection process that could influence observations somewhere in the feature space. (Brownlee, 2019) (Ali et al., 2015). On account of this, what are the possible reasons or causes of classes that may be imbalanced?

Class Imbalance: the cause

The imbalance classes in the data distribution affecting classification tasks may have various causes; these could be generated during the data collection process in the data sampling or driven by the domain issue as mentioned above. However, they are mainly grouped in measurement errors and biased sampling. Defined as:

Measurement error is the difference between a recorded value and the actual value of something. This error can be systematic or random and may generate bias and extra variability in statistical outputs. Identifying the causes of errors can help lessen their influence ad record accurate and exact measurements (Bhandari, 2021) (CROSS European Commission, 2019).

Firstly, in consonance with Bhandari in 2021 that defines random errors are mistakes made by chance that drive the difference between the observed values and the actual value of something; for example, when a data scientist misreads the results of a survey

recording incorrect measurements. This type of error does not influence the average of the data and could be considered noise as it blurs the actual value of the measurements, having no steady effect on the samples. Trochim (2021) says that common sources of random errors involve unreliable measurement instruments, low controlled experimental techniques, and discrepancies in the practical context that can be avoided by taking measurements repeatedly or increasing the sample size. (The BMJ, 2021)

Secondly, in agreement with Bhandari in 2021 that describes systematic errors as consistent or proportional mistakes that can vary between the observed and actual value; for instance, a miscalibrated scale registering weights lower than they are. This type of error does tend to affect either positive or negative samples, and they are referred to as bias, skewing the true values, leading to inaccurate conclusions. Common sources of systematic errors could come from the research material leading to false behaviour by the participants, known as response bias or experimenter drift that occurs when the observers become super users in the process deviating from the standardised procedures (Trochim, 2021). These can be reduced using methods such as triangulation where multiple techniques to record observations are being used, regular calibration of the instruments to determine the standard quantity or randomisation of the sampling method to ensure samples do not differ from the population. (The BMJ, 2021).

Last but not least, Trochim (2021) mentioned that in comparison, systematic errors are a bigger problem in the research field, as random errors tend to cluster around the actual value when taken multiple measurements, as large samples help to lower them down. However, systematic errors can consistently skew the data leading to false conclusions.

Biased sampling is one of the sources of systematic errors within the measurement errors range. Sampling bias happens when systematically a cohort of people is more likely to get selected than others, limiting the generalizability of findings and potential discoveries over the population in observation, as applicability only belongs to that population in particular, as mentioned by Schildcrout et al. (2015). This cause of data imbalance could be driven either by research design or the data collection procedures in place. Furthermore, as we know that distinguishing the size of the focus population could be a challenging task, the careful selection of sampling procedures

can correct under-coverage bias, or defining a target population in a better fashion will result in enhanced sampling practices. (Bhandari, 2021) (Trochim, 2021).

Leimaitre et al. (2017) reported that imbalance problems might occur in the data collection process. When selecting the appropriate sample size, impracticalities come to pass, and depending on the population magnitude, techniques like surveys are not necessarily suited for the purpose. If applying methods like oversampling, and some groups are underrepresented in the dataset, it leads to selecting some of the respondents from that group to be weighted in the pool of answers to add to the share, helping to remove the biased sample. (Trochim, 2021) (Bhandari, 2021). Biased sampling hinders external validity and limits the findings' generalizability because it is impossible to apply them in broader groups. (Brownlee, 2019)

In conclusion, the class imbalance in a dataset could be caused by various reasons, including the property of the problem domain or incorrect cohort selection. However, determining the motive to correct it could generate extra computation cost, time, and resources expenditures, in addition to possibly needing more sample gathering. Having a plan that takes us from the sample or target population, problem scope or domain, selection criteria of the observations, and sample size determination in our process could help to have in mind the possible ways to improve class distribution to remedy the situation at the right time, in consonance with Brownlee (2019). Nevertheless, in the case that we are facing the class imbalance problem, how could we solve it? Or perhaps fix it in our datasets?

Class Imbalance: the resampling techniques

Resampling techniques assist in the imbalance class distribution problem, providing suitable and efficient strategies to continue examining a given instance or population while using the standard machine learning classifiers as it modifies the training set instead of the algorithm used. (Kaldy and Kambhampati, 2018) (Cao et al., 2017). Furthermore, resampling deals with the class unevenness issue, owing to the fact that these approaches handle the dataset in such a way that allows building a dissemination process of the data points per class. In other words, it adjusts the datasets by altering the informational index that compounds the minor and significant parts of the categories present in the data, as mentioned by Kashyap and Gulati (2020) (Aada and Tiwari, 2019).

Resampling is an effective solution to generate a moderately balanced class distribution in line with Wang et al. (2015); the process could be described as a system that guarantees to reach the exact sizes of classes per each kind of class label in the dataset. Along these lines, resampling is a practice that, while reasonably using the data sample it improves the accuracy of measuring the ambiguity of a population parameter (Brownlee, 2019). However, just as Montero (2021) mentioned, the accountability given to the usage of resampling relies on the fact that it compresses basically by random chance. Meaning that after a single sample is drawn from a population, statistics are calculated, and the parameters are projected, and as a result, they can be either closed or far from the real population value provided. Therefore, with these techniques, we can draw various samples, try different combinations, compute more statistics, and offer a range of values that could take us closer to the proper population parameters.

Consequently, in the standard model training process, it is well-known that achieving continuous improvements in the model's accuracy is challenging, and contributing factors includes handling imbalanced class distributions (Ghorbani, R. and Ghousi, R., 2021) (He et al., 2008). Just as Japkowicz, N. and Stephen, S. (2002) and Liu, Wu and Zhou (2006) agreed on its impact that made us understand the importance of the class imbalance problem, defining a relationship between complexity, the size of the training dataset, and the class imbalance level. With great attention or afflux devoted to combating the class imbalance issue, various solutions have been proposed with time and research, solutions that can be separated into three groups: (a) data-driven methods, (b) algorithm-driven methods, and (c) hybrid systems (Liu, Wu and Zhou, 2006) (Desuky, A. and Hussain, S., 2021). Explain as follows:

Data-driven methods involve stabilising the training dataset by either reducing the number of samples in more classes or increasing the minority class. Therefore, the dataset is balanced out either randomly or deterministically (Liu et al., 2008) (Wang et al., 2015). During data processing, the process takes place to redistribute the training data before the model development stage (Longadge and Dongre, 2013) (Santiso et al., 2018). In this technique, the data produced is handled to disseminate the balance in every data point, thus allowing ML algorithms to learn from the balanced data with no bias as any standard algorithm (he). Some techniques in the data-driven methods are undersampling, oversampling, and hybrid methods (He et al., 2008) (Wang et al., 2015). In the following, methods are described as:

As reported by Kashyap and Gulati in 2020, Undersampling methods handle the imbalanced data problem by adjusting the data points in the majority class by removing them compared to the minority class. The undersampling method can be applied when enough data is collected (Ghorbani and Ghousi, 2021). The advantage of using this method involves runtime improvement of the models, as it reduces memory consumption lowering the training data samples (Sayak, 2018) (Kushi et al., 2021). However, the disadvantages of this method include discarding helpful information from the data when dropping data points and not correctly representing the population, resulting in poor performance in test data (Telenti and Jiang, 2020) (Guo et al., 2008).

There are different undersampling models, such as:

- Random-Undersampling (RUS): Removes instances from the more prominent class in relation to the size of the minority class in the dataset(Taghizadeh-Mehrjardi, 2020).
- Cluster Centroids: Replaces the cluster with most samples by the cluster centroid of the KMeans algorithm (Lemaitre et al., 2017).
- Near Miss: Eliminates instances from the larger class with the shortest distance with the smaller class (Madhukar, 2020).
- Instance Hardness Threshold (IHT): Removes majority class samples that overlap with the minority class sample in the space considered complex samples (Smith et al., 2013).
- Tomek Links (TL): Removes from the majority class selected pairs of instances belonging to distinctive classes and that are each other's Nearest Neighbours (Sasada et al., 2020).
- Condensed Nearest Neighbour (CNN): Collects samples in the training dataset part of the minimum consistent set (Rupak, 2021).
- All K-Nearest Neighbours (All KNN): Discards instances in the majority class if their neighbour classifies a sample incorrectly, applying KNN to every sample (Lemaitre et al., 2017).
- Edited Nearest Neighbours (ENN): Removes misclassified data points, applying the KNN rule k=3 before a k=1 classification rule is applied (Rupak, 2021).
- Repeated ENN: A repeated process of ENN (Rupak, 2021).
- One-Sided Selection (OSS): Combination of Tomek Links and CNN techniques removes data points from the majority class that are either on the class boundary or are redundant and far from the decision boundary (Rupak, 2021).

- Neighbourhood Cleaning Rule (NCR): Combining CNN rule and ENN techniques, removing redundant, noisy and ambiguous data. (Rupak, 2021).

Oversampling methods based on Ghorbani and Ghousi (2021) described them as strategies where the diffusion of the samples in the training set is adjusted with new data points from the minority class per their replication. Oversampled data is introduced in the covariate space to generate new examples of that specific class. This process reaches a balance in the class distribution, creating stronger class boundaries (Kushi et al., 2021). The advantages of oversampling leads to no loss of information. The disadvantages of oversampling include increased training time and possible overfitting due to the replication of minority classes (Sayak, 2018).

Oversampling methods include:

- Random Over-Sampling (ROS): Increases the size of repeating the original data points. Random Oversampling does not create either new samples or add variety (Li et al., 2010).
- Synthetic Minority Over-Sampling Technique (SMOTE): Creates synthetic data points from the minority class, taking samples of the feature space and using NN to combine them. New samples are not copies of the existing ones (Ghorbani and Ghousi, 2021).
- Synthetic Minority Over-Sampling Technique - Nominal Continuous (SMOTE-NC): nominal features are encoded as numeric values, and the difference between them reflects the amount of change of association with the minority class (Mukherjee and Khushi, 2021).
- Borderline SMOTE: Creates new data points in the minority class, identifying the ones near the borderline, which are the most misclassified ones (Guo et al., 2008).
- Support Vector Machine (SVM) SMOTE: Combination SVM and SMOTE, where augments the minority of the class concerning their NN. This technique creates new classes to facilitate setting boundaries between classes while using the SVM model (Kelly, 2020).
- KMeans SMOTE: Combination of K-Means and SMOTE, where after clustering the minority class, clusters are oversampled about their density (Liu, 2021)
- Adaptive Synthetic (ADASYN): Creates instances from the minority class in the dimensions where are low (Brownlee, 2018).

Hybrid methods blend undersampling and oversampling methods to overcome the possible difficulties of dealing with imbalanced datasets. Considering that we can use undersampling when we have lots of data available, and oversampling in the opposite scenario and trying various ratios of class-labels targets (Sayak, 2018) (Khushi et al., 2021). The methods include SMOTE-ENN and SMOTE-Tomek, which combine the oversampling technique SMOTE, and the undersampling techniques ENN and Tomek Links. These methods aim to balance the training dataset and eliminate the noisy points from the mistaken margin of the decision boundary (Khushi et al., 2021).

Algorithm-driven methods are another way to manage the imbalanced data problem. This approach aims to augment the algorithm's performance (Wasikowski and Chen, 2010) by either modifying the algorithm itself, associating a cost variable or weight that makes it more precise with the minority class, or building a new implementation that includes algorithms that are not affected by the skew distribution. As reported by (Khushi et al., 2021), this technique, also called a cost-sensitive framework, is characterised by an intended understanding of the minority sample, encouraging correct classification. Consequently, it minimises cost errors while comparing misclassification of positive class vs negative class examples. Some of the models included in this resampling method are imbalance-sensitive and cost-sensitive learning. (Khaldy and Kambhampati, 2018).

Hybrid systems are ensemble methods created based on the two previous techniques described. Ensembles are intended to boost the accuracy of a single classifier while training multiple and diverse classifiers are then joining their results to output a single class label. (Krawczyk et al., 2014) The results modify the learning process because the generalisation capabilities are more robust at the collective level than the individual. (Liu et al., 2020) Most ensemble methods are binary class problems drive, and hybrid systems can categorise them into parallel and sequential ensemble methods.

Some examples included in this resampling method are Bagging-style, boosting-based, and hybrid ensemble methods. The report of Krawczyk et al. (2014) defines bagging, or bootstrap aggregation is an ensemble method that generates data samples, randomly selecting data points to create a subset that will be replaced within the dataset reiteratively. Boosting is an ensemble method that sequentially adjusts the observation weight depending on its performance during classification. If incorrectly classified, more weight and vice versa (Vadapalli, 2020).

The standard ensemble methods are as follows:

- Balanced Bagging: Implementation of bagging and random undersampling, resampling subset of the data ser before integrating its estimators. (Alam et al., 2020)
- Balanced Random Forest: Develop samples from the minority class present in the dataset, layering its result within the trees (Khushi et al., 2021).
- Easy Ensemble: Selects a subset of the samples from both classes to create a dataset through selective construction (Brownlee, 2019)
- Random Under-Sampling Boost (RUSBoost): Per every round of boosting, random undersampling is performed (Seiffert et al., 2018)
- Balance Cascade: Incorporates bagging and boosting, selecting a subset of both classes, minority and majority, that can be easily classified (Leimaitre et al. 2017)

In conclusion, resampling techniques are a key tool in machine learning, as data is often scarce; these methods can help us solve the class imbalance problem; of course, a more in-depth research is needed to determine which one suits our problem better and level of imbalance distribution. Having said that, the problem to solve in this case relates to medical data and medical settings, therefore how vital is medical data? How the imbalanced class problem affects medical data? Is it possible to do an accurate classification of diseases?

Chapter II: Classification of Imbalanced Medical Data

Medical Data: the information

Medical data, also known as health data, according to Tzourakis and Wang in 2016, was defined as any information that reflects health conditions, reproductive outcomes, causes of death, or describes a quality of life from an individual or an entire population. In addition to clinical metrics, health and wellness are part of the medical record, including environmental, socioeconomic and behavioural data. Furthermore, Raghupathi and Raghupathi (2014) added that an extensive amount of the health data collected and consumed originates from the interaction of individuals with health care systems, such as hospitals in outpatient and inpatient settings, clinics, nursing homes, insurance companies. As a result, health care providers' gathered data will usually include records of services received clinical outcomes or information regarding those services (Reid et al., 2015)

Moreover, claimed by Hogarth in 2021, medical data is divided into two categories, described as Source Health Data and Derived Health Data. The Source Health Data is the dataset originating from the patient. In this case, regardless of how insignificant or intricate the derivations might be, the source medical data is collected. That includes any information concerning the health, care, and treatment of the patients, which results in reports that are used in treatment or monitoring of a patient, that could generate a claim or bill for services provided for the insurance companies or used for operations, financial management, population health activities, or quality metrics. Derived Health Data is characterised as some transformation made over the original source data.

Health care organisations aim to re-establish patients' health, which requires operative and competent medical systems that manage data for evidence-based intercession. The healthcare industry traditionally has produced significant volumes of data, driven by record keeping, regulatory requirements, and patient supervision (Tierney et al., 2015). The enormous amounts of data collected hold the potential of assisting a wide variety of medical and healthcare functions, plus other clinical decisions such as disease supervision and population health management (Raghupathi and Raghupathi, 2014). Furthermore, the majority of this data is kept in hard copy form; the existing tendency is in the direction of rapid digitisation of these

enormous volumes of data, going from a paper-based record (PBR) to an electronic health record (EHR) system to manage the full spectrum of patient's data. (Adane et al., 2019)

As part of the computerisation of the medical data in the healthcare data management systems, the data can be structured. Structured data describes standardised and effortlessly transferable information between health systems, for instance, patients' names, date of birth, results from medical tests (Kudyba, 2010). On the other hand, unstructured data refers to information that is not normalised and homogeneous, such as diagnosis in medical coding format, physician notes, x-rays, emails or voice notes or recordings about a patient. The improvement of EHR systems standardising policies and procedures has also enhanced the conversion of most unstructured data to its structured format, which allows the gathering of great insights from the patients and health facilities, physicians, types of medications, and medical services (Raghupathi and Raghupathi, 2014).

As already stated, Tierney et al. (2015) mentioned that health data could be employed to promote individuals' health and public health and medical research and development. The same way that can be classified as structured and unstructured, plus source and derived, according to its usage, could be either primary or secondary. The primary type associates the importance of the medical data with the benefits provided to the individual from whom it was gathered. The secondary is about medical usage beyond that individual (Safran et al., 2017).

Even though the main source of healthcare data are individuals as it is reflected on patient demographic information, progress notes, vital signs, medications diagnoses, and biometric data gathering, not all of the data is produced in the same manner. The data could also come from clinical trials, labs, pharmacies, payers/claims, nursing homes, health agencies, as individuals interact with these touchpoints. Medical informatics will gather the necessary information and use it in accordance. (Adane et al., 2019)

Therefore, as specified by Adane et al. (2019), the cumulative variety and use of health data are crucial for eHealth or electronic health improvement, but also the arrival and advancements of technology in medical settings has produced new privacy, security, and ethical concerns (Safran et al., 2017). Furthermore, although the medical industry is an integrated platform, its data documentation practices and data protection laws

diverge significantly amongst hospitals, providers and countries (Lujic et al., 2014). Consequently, the comprehensive healthcare data management policies need to determine how to handle privacy, retention policy, confidentiality and avoid restoration after the destruction has been in place. Furthermore, the codes of these procedures should be categorised concerning the benefits and inconveniences of wiping out or preserving medical data (Yang et al., 2014).

Medicine is an incidental research asset since its principally focused is on patient-care tasks; which means that gathering medical data only used is to help the individual patient, and its applicability to all people moves to the background with no sense of urgency, or moral obligation, merely to be used for beneficial reasons. Consequently, researchers in various fields will not necessarily be aware of the specific restrictions and complications of the privacy-sensitive, assorted, but extensive data captured in medicine. Hence, medical data's ethical and legal attributes must be discussed, including data ownership, expected benefits, and administrative management issues. (Cios et al., 2002)

Finally, with all of the technological improvements in the medical setting, where we can track patients over time, identify patients in need of preventive screenings and check-ups, monitor and improve quality of care, we wonder how the imbalanced data problem affects health data?

Medical Data: the imbalanced data problem

In medical datasets, data are mainly constituted of normal or negative samples, which are the majority of data points, with a small proportion of abnormal ones or positive, that precedes the so-called class imbalance problems. As previously mentioned in class imbalance problems, recording data into the classifier to shape the learning model will regularly direct learning bias to the majority class. Li D et al. (2010) declared that advanced machine learning methods have focused on clinical practices in quest of interpretability of the diagnostic/prognostic causes that could bring confidence to doctors when looking after their patients applying this type of technology. However, when the data gathered is imbalanced in these diagnostic groupings, we observe that the standard ML techniques might generate results that overpowered most classes, fading expected accuracy (Zhou and Wong, 2021).

Zhang et al. (2010) said that a balanced dataset is crucial for generating an excellent predictive model; their purpose is to give similar consideration to both classes without discrimination while augmenting the global precision without respecting the relative distribution of each class. Nevertheless, real-world data is characteristically imbalanced, which naturally causes the reduction of generalisation in machine learning algorithms, meaning that suitable sampling techniques for medical datasets are needed to avoid the cost implied in the non-prediction of minority classes when patients at great risk are part of the circumstances. (Rahman and Davis, 2013)

Logically as Rahman and Davis (2013) said, within the data, occurrences of patients suffering the particular disease of interest would cause a loss of function in the model and skew it towards this negative classification as it does not have the registration of enough patients with the disease or lack of a gold standard of patients and associated symptoms. In addition, some possible difficulties of presenting a trustworthy diagnosis and dispensing appropriate treatment, so the data generated by these procedures, having numerous issues are originated from the fact that a disease is rare.

By definition, rare events occur with an obvious and considerably reduced regularity than more ordinary events, and they are the ones that have even greater importance when correctly classified, hence the effort required for proper classification (Maalouf and Trafalis, 2011). That means that some healthcare-related datasets with imbalanced classification, perfectly uncovering minority class utterances, always turn out to be of immense significance as they relate to the high-impact occurrences. According to Legay et al (2016) that define rare events or rare properties as extremely important to the performance of the system, and the situation develops since they are difficult to detect which can represent a problem for statistical model checking, due to the fact that small probabilities estimation can result in large relative errors, but having in mind that expanding number of observations confidence of the estimate grows as well.

For example, more efforts have been made to distinguish medical incident reports, rarer disorders like restrictive cardiomyopathy or targeting events recognised as frequent sources of medication errors that may result in adverse or harmful patient outcomes (Wong, 2014). Because considering the opposite case where a multi-speciality hospital naturally will have sufficient records for common illnesses like colds/flu, allergies, lower respiratory infections to be used for ML classification tasks (Sundararaman, 2021). The future simply belongs to those healthcare service

providers who overcome this bias. Henceforward, it is required to generate precise, transparent, and interpretable results in decision-making without sacrificing accuracy for these imbalanced groups (Zhou and Wong, 2021).

In line with Emanet et al. (2014) that described the main goal is the development of tools that are programmed for high accuracy while being low-cost. Consequently, those tools will address a wide diversity of problem types in many fields of healthcare and medicine and ultimately operate as a decision assistance tool for physicians and healthcare professionals. This data-rich industry can undoubtedly take advantage of what modern decision analytics offers, defining models and techniques that can handle the imbalanced distributions of classes as per the nature of this field. The initial priority is to focus on complicated classification problems, as they promoted the fast growth in the field due to the gathering and storing of the data produced by it, which leads to faster and better decision-making in the healthcare/medicine area when done correctly. Finally, how the implementation of imbalanced classes in classification tasks looks?

Medical Data: the machine learning classification task

Advanced healthcare projects can be described as proof-driven and model-assisted. In regular conditions, resolutions taken in the clinical ecosystem should be strengthened by statistical models that calculate potential risks and positive outcomes. As Razzaghi et al. mentioned in 2015, in any case, this is established on an enquiry of related clinical information and functional history of a patient because healthcare counts on how adequately state-of-the-art algorithms of machine learning accommodate clinical data. As medical diagnosis is part of that ecosystem, it is an essential yet complex task that needs to be performed precisely and efficiently due to the impactful significance in people's lives can have the incorrect disease prognosis description (Soni et al., 2011).

Considering that some healthcare problems affect pattern recognition as regularly implemented, where even with large-scale data, the cohort selected can contain missing or flawed features that skew the distribution of classes (Razzaghi et al., 2015). Therefore, most of the practical machine learning techniques need supervised learning. A supervised learning method can forecast iteratively on the labelled training data through classification methods. Classification is a technique that needs machine learning algorithms and labelled training data to understand how to

designate class labels to sample from the domain, such as health data (Johnson and Khoshgoftaar, 2019).

As aforementioned, the most important results in classification problems relate to imbalanced or incorrectly labelled problems, because as a precedent, healthcare analytics is inspired by rare events such as healthcare emergencies, severe chronic conditions, inconsistencies and restricted access to care. On account of Zhu et al. in 2018 that confirmed that classification in imbalanced class data drags considerable attention because the methods in place are predisposed to catalogue the samples into the majority class, which results in bias and insufficient detection of the minority class affecting medical applications. Therefore, if the problem gets fixed, it will englobe an equivalent depiction of all the classes in the dataset to drive accurate classification and effective performance (Johnson and Khoshgoftaar, 2019).

Finally, as we are trying to overcome the classification problems in imbalanced datasets issue and give the correct use to the available data, various state-of-the-art learning techniques have suggested particular and individual ways, depending on how it should be addressed. Methods like algorithmic level, data-level or hybrid level have demonstrated their benefits in medical data space, enhancing the overall performance of the classifier or producing high accuracy in identifying both classes, in some cases without altering the training set. Moreover, these models have implemented a form of a simplified risk-assessment formula on a sophisticated machine learning tool. (Kumar P, et al. 2021). On account of this, what machine learning techniques could help us take the most advantage from medical data?

Chapter III: Machine Learning Classification

Over the past years, machine learning has gained a central position in developing algorithms and models to interpret medical data and improve clinical diagnosis and prognosis. Machine learning approaches have successfully simplified the construction of prognostic models for health assessment based on available historical labelled data issued from similar systems or specific physical models. (Berghout et al., 2021) (Livieris et al., 2018) Using supervised learning, the goal is to learn a model from labelled data points, where the final model is applied to an unobserved test set, and the method is corroborated based on how effective it was in allocating test data to diverse classes. One standard formulation of the supervised learning task is the classification problem, described as follow: (Peikari et al, 2018) (Osisanwo et al., 2017)

Classification: the machine learning classification tasks

According to Brownlee (2019), the classification task is part of the supervised learning techniques from the machine learning algorithms that basically gain information from the class labelled to implement it in the examples from the problem domain or test data. Apart from the definition mentioned above, classification is the method that predicts the class of given data points, and those classes are frequently called targets/ labels or categories. The classification predictive modelling is the task of impending a mapping function (f) from input variables (X) to discrete output variables (y) while using a training dataset that contains many instances of inputs and outputs to learn from (Asiri, 2018).

On top of that, Novaković et al., in 2017, said that classification is a popular task in machine learning technique but struggles with unidentified cases present in one of the groupings classes. The classification of an object is based on finding correspondences with encoded objects that are associates of other classes, with the comparison of those two objects resolved by evaluating their features. Being the target functions its most critical observation, a discrete variable countable in a finite amount of time, meaning that the class label cannot have allocated numerical or some other values with those characteristics, plus the class attributes, whose value should be determined, and can be categorical attributes.

According to Osisanwo et al. (2017), one typical definition of the classification problem refers to the learner compelling to acquire or approach a function's behaviour, which maps a vector into one of several classes by observing at numerous input and output instances of the function. In classifying all objects are classified into one of the classes with specific accuracy. The assignment is that on the features of objects whose classification is known beforehand, create a model that will be achieved the classification of new objects. More generally speaking, generating a classifier can be employed to derive from new instances (Novaković et al., 2017).

In addition, there are three main types of classification tasks:

- First, binary classification: refers to classifying elements of a given dataset into one of two classes (Brownlee, 2020).
- Multi-Class classification: refers to classifying one of more than two classes. Multi-Class classification assumes that each sample is assigned to one and only one label, meaning that class labels are mutually exclusive. (Nabi, 2018)
- Multi-Label classification: describes predicting one or more classes per example, meaning that class labels are non-mutually exclusive (Brownlee, 2020).

Some examples of classification problems in the medical field include:

- Given a medical record, classify if it is diabetes or not. This example will belong to binary classification.
- Given a medical code, classifying it as one type of cancer represents a multi-class classification.
- This example represents multi-label classification, given lab results, classified as a blood donor with Hepatitis C or not.

Finally, researchers have endeavoured to employ varied methods to correct the accuracy of classification methods for the provided dataset in the medical field. Classification techniques whose improvement suffices when producing information to recognise the potential patient's diagnosis/prognosis and thereby advance inpatient care were also data mining comes to play and is applied to classify medical data in recent studies and obtained significant results. (Babu and Suresh, 2013). So, what are the primary classification techniques?

Classification: the machine learning classification algorithms

The primary purpose of a classification problem is to determine the category/class to which a new data point will fall, categorising the data into a given number of classes. A classification model attempts to deduce the input values specified for training to consequently calculate the class labels/categories for the additional data (Garg, 2018). An accurate classification model could identify and detect errors earlier, guiding rapid control outcomes. Classification can be implemented on structured or unstructured data. (Ardakani et al., 2016). There are many diverse classification models that we may come across in machine learning and specialised methods to modelling that can be employed for each, such as:

Support Vector Machine (SVM) utilises the class of kernel/margin-based methods, achieving great generalisation and described as a helpful approach for solving machine learning problems. In classification, SVM achieves a linear partition (hyper-plane/ hyper line) between the data points fitting two classes in a multidimensional area. A collection of occurrences closest to the ideal hyperplane is recognised as a support vector or the margin, and determining the optimal hyperplane is nothing but a linear classification. (Ardakani et al., 2016) (Tarle, 2016)

SVMs, as Tarle (2016) mentioned, is employed to determine associations between features and redundant features, which means there is a trade-off between increasing the margin and reducing the number of misclassified instances. When the dimensionality of the feature space increases, a nonlinear SVM can be generated. It is achievable to find a distinguishing hyperplane in higher dimensions. The kernel type can be Gaussian Kernel, polynomial kernel and linear kernel or dot-product kernel. (Ardakani et al., 2016)

Decision Tree (DT), as Mohamed stated in 2017, is amongst the most extensively utilised classifiers in statistics and machine learning. A decision tree is a hierarchical pattern model that applies the divide-and-conquer method, being a non-parametric technique employed in classification tasks. Decision trees use recursive data partitioning, resulting from its hierarchical design. Its uncomplicated description causes the readers to decode and understand the outcome, which can be modified into a set of simple if-then rules. Hence, DT is applicable and appropriate for large training sets. Also, the DT group of algorithms does not need any supplementary

information; moreover, it was previously confined in the training data, and as a rule-based classifier, it used its accuracy as an objective function.

Each non-leaf node denotes a test on the data item under significance in decision trees. The result of the test determines the path or specific branch be chosen. We can categorise a data item by initiating at the root node and following the path until we arrive at a leaf. When a terminal node is approached, the decision is made. (Zekić-Sušac, et al, 2014)

Gaussian Naïve Bayes (GNB) is the technique based on Bayesian theory that accepts that all attributes do not depend on one another. Its focus is on prior, posterior and discrete probability distributions of sample data, showing promising results in reliable circumstances, regardless of the simplifying hypotheses behind the GNB. Bayes classifiers are typically less precise than other more complex learning algorithms, but it is as competitive when benchmarked on standard datasets, in special ones with substantial data feature dependencies. It works with a gradient descent algorithm that enhances any differentiable loss function. These are straightforward Bayesian networks that create directed acyclic graphs with only one parent representing the unobserved node and numerous children that relate to observed nodes with an explicit acceptance of individuality between child nodes in the setting of their parent. (Bozkurt, 2021) (Tarle, 2016)

K-Nearest Neighbourhood (KNN), according to Ardakani et al. in 2016, is a distance-based and linear supervised classification technique. KNN determines the location of the data points in relation to others by considering the k-nearest patterns in the training data to the test pattern. First, the KNN classifier finds unseen examples of unidentified data points using the previously known as the nearest neighbour and classified data points. Then, it classifies the data points using more than one nearest neighbour; as a result, the findings are displayed in a feature space-defining its qualities or characteristics.

Every point gets classified using distance measures based on its affinity to other data points trained by the model. The new point class is determined by picking the K closest points and the average of the majority of the class the new point is surrounded by to be the class of the new point. (Ardakani et al., 2016). In estimating the model, it is vital to pick the suitable estimate of k since it can influence the predictive capacity as such a small value k will direct to a significant variance in predictions, while a

greater value of k may lead to a significant model bias. (Zekić-Sušac, Pfeifer and Šarlija, 2014)

Logistic Regression (LR) is amongst the most commonly used tools for applied statistics and discrete data analysis that predicts the probability of a categorical dependent variable. In logistic regression, the dependent variable is binary coded as 1 or 0, referring to yes/success or no/failure, respectively (Bozkurt, 2021). In other words, the logistic regression model predicts P(Y=1) as a function of X, as it tries to find the best line that separates the two classes. This classification function uses a class for building and uses a single multinomial logistic regression model with a single estimator. (Mohamed, 2017)

Furthermore, LR usually defines where the boundary among the classes is, also states the class likelihoods differ on distance from the boundary in a specific approach. It moves towards the extremes (0 and 1) more rapidly when the larger dataset. These statements about probabilities make logistic regression more than just a classifier. It makes more robust, more detailed predictions and can be fit differently, but those robust predictions could be wrong. However, prediction results in a dichotomous outcome. (Mohamed, 2017) (Bozkurt, 2021)

Having described these algorithms, their properties, their features, and the various ways that can help us to solve problems; It is also valuable to determine their purpose and possible applications for specific machine learning tasks such as the classification task, and these can be described through their strengths and weaknesses, defined as follows:

	Advantages	Disadvantages
Support Vector Machine	• It extends reasonably well to high dimensional data and works with continuous and categorical data. • The data structure is not inferred due to its distribution-free statistic method, and its robustness makes it capable of managing data that contains errors. • The trade-off between model intricacies and the error can be contained to avoid overfitting problems.	• It can be resource-intensive. • It needs to provide a good window space to transform the data. • It is designed to solve binary and multi-class classification problems by dividing the classes into pairs. • It lacks clarity in the results due to its distribution-free statistic method unless the features are interpretable.

Decision Tree	• Domain knowledge is not required, and it is easy to understand, as it has a simple schematically representation that follows a set of rules comprehensible for the reader. • The data with high dimensions can be quickly processed, and numerical and categorical, outliers and missing values data can be handled. • It is a non-parametric tool; therefore, it does not need any statistical model specification.	• It is commonly used for classification tasks in comparison to regression tasks. • It can be resource-intensive, leaning into data overfitting. • A categorical output attribute is generated as a result. • The performance and complexity of the classifier depend upon the type of dataset.
Gaussian Naïve Bayes	• It performs in comparison to other models when the assumption of independent predictors holds true. • The computation process is more straightforward as it requires a small amount of data to estimate over the test data. • It is suitable for continuous-valued attributes.	• It assumes that predictors are mutually independent variables. • It does not predict using categorical data. • It requires several parameters like the topology or structure of the network.
K-Nearest Neighbourhood	• It is easy to implement as it only requires two parameters, the value of K and the distance function. • The algorithm's accuracy will not be impacted after adding new data with now training before the predictions feature. • No Training Period is needed; the algorithm only learns from the training datasets to make accurate predictions.	• The algorithm does not work well in high dimensional data as it becomes challenging to calculate the distance in each dimension. • It needs standardisation and normalisation before algorithm implementation, as well as imputation of values and remove outliers manually. • It can be resource-intensive and needs large memory allocated to manage the training examples.

Logistic Regression	• It performs better when the dataset is linearly separable. • It is straightforward to implement, understand and very effective to train. • It calculates how significant a predictor is and its associated positive or negative class.	• The real-world data is hardly ever linearly divisible, so LR limits itself following the hypothesis of linearity amongst the dependent and the independent variables. • The dependent variable is limited to the discrete number set and its prediction. • It has a predisposition to over-fitting in high dimensional datasets.

Table 1: Advantages vs Disadvantages of Machine Learning Algorithms
(Kumar, 2019)((Mohamed, 2017)

The analysis of classification methods in statistics is vast, and there are more types of classification algorithms than the ones aforementioned, that applicability will depend on the dataset. Consequently, for every model, we need to ask ourselves. What do we want to optimise? What are the performance metrics for classification tasks? In particular, for imbalanced problems?

Classification: the machine learning classification evaluation metrics

The assessment methods, also known as evaluation metrics as reported by Tharwat in 2020, are critical in evaluating the classifier implementation and conducting the modelling process. Selecting wrong metrics could lead to developing a flawed model or misleading results. Keeping in mind that the training error obtained through the metrics is usually lower than the testing and the validation error since the trained model fits the same data, the training error will measure precisely the model matched the data.

Furthermore, the impact of balanced and imbalanced data on each valuation technique must be determined. Different evaluation metrics could be responsive to the imbalanced data when the instances of one class in a dataset exceed the instances of the other classes, which means that the strength of each approach against imbalanced data needs to be verified (Tharwat, 2020). Knowing that they help to quantify or measure the model's performance, being an important part of the machine learning model pipeline process, the quintessential evaluation metrics that are widely used for evaluating machine learning classification models, are

Confusion Matrix

To begin with, there are for 2-class classification problems, and four possible results that describe the components or sections of a 2×2 confusion matrix or a contingency table, described by Rácz et al. (2019) as tabular visualisation represent correct predictions or ground-truth labels vs the indicated incorrect predictions. That means that

- If the sample is positive and classified as positive, it is calculated as a true positive (TP),
- If the sample is positive and classified as negative, it is reflected a false negative (FN) or Type II error,
- If the sample is negative and classified as negative, it is considered true negative (TN),
- If the sample is negative, it is classified as positive; it is considered false positive (FP), false alarm, or type I error.

The confusion matrix is not a classification metric but a building block utilised to compute numerous classification methods. (Tharwat, 2020). An illustrative example of a confusion matrix, with two true classes (2x2) and for multiple classes (3x3), P as Positive and N as Negative, describes as follows

Predicted Class	True/Actual Class: Positive (P)	Negative (N)
True (T)	True Positive (TP)	False Positive (FP)
False (F)	False Negative (FN)	True Negative (TN)
	P = TP + FN	N = FP + TN

(Tharwat, 2020)

Predicted Class	True Class: A	B	C
A	TP_A	E_{BA}	E_{CA}
B	E_{AB}	TP_B	E_{CB}
C	E_{AC}	E_{BC}	TP_C

(Tharwat, 2020)

Accuracy

It defines the number of ground-truth labels made by the model divided by the number of predictions made. Consider a great metric only if the target variable classes in the dataset are almost balanced and should under no circumstances be employed as a metric when the target variable classes in the data are greater in one class. (Nighania, 2018) (Sunasra, 2017)

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Precision

It is the percentage of positive classes out of the predicted positive classes. It is giving us information about its performance with respect to false positives. This translates to How accurate is the model when it says it is accurate. (Nighania, 2018) (Sunasra, 2017)

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

Recall / Sensitivity / True Positive Rate

It is the percentage of positive classes out of the actual positive classes. It is giving us information about its performance with respect to false negatives. This translates to How many additional true data points we caught and how many we missed (Nighania, 2018) (Sunasra, 2017).

$$\frac{TP}{TP + FN}$$

Specificity

It is the percentage of negative classes out of the *total actual negative* classes, being the exact opposite of recall; it is a measure to determine how dispersed the classes are. Specificity is the exact opposite of Recall (Nighania, 2018) (Sunasra, 2017).

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

F1 score

It is defined as the harmonic mean of precision and recall that captures the pair's contribution, so the bigger the F1 score, the better. The harmonic mean is a type of an average when x and y are equivalent. Due to the result in the numerator, the final F1 score goes down significantly if one goes down. However, when x and y are

different, it is closer to the smaller number than the larger number (Nighania, 2018) (Sunasra, 2017).

$$\frac{2}{1/\text{Precision} + 1/\text{Recall}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision-Recall or PR curve

It is the area under the curve. The curve relates precision and recall for various threshold estimations. The higher its numerical value, the better (Nighania, 2018).

ROC (Receiver Operating Characteristics) curve

ROC or Receiver Operating Characteristic, and the diagram is plotted alongside True Positive Rate (TPR) and False Positive Rate (FPR) for several threshold values. As TPR increases, FPR also increases. The higher its numerical value, the better (Nighania, 2018).

$$\text{True Positive Rate} = \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate} = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$

Area under the curve (AUC)

Area under the curve, is a performance measurement diagram plotted alongside ROC (Receiver Operating Characteristics) or Precision-Recall in relation to all possible threshold values. For ROC values are from 0 to 1, The higher its numerical value, the better. For RlP average of precision scores calculated for each recall threshold. (Nighania, 2018).

Besides the most common metrics mentioned earlier, exists another range of lesser-known evaluation metrics that lets you assess the machine learning model's performance. For instance,

- F-beta score where the beta parameter determines the recall weight in the combined score.
- F2 score where the emphasis is put in recall with a beta = 2
- Cohen Kappa Metric calculates how better your model is than the random classifier that predicts based on class frequencies.
- Matthews Correlation Coefficient (MCC) is a correlation between predicted classes and ground truth.

- Log Loss / Cross-Entropy is the score that condenses the average variance amongst two probability distributions. An ideal classifier has a log loss of 0.0, with poor values being positive up to infinity.
- Brier Score / Brier Skill Score is the mean squared error between the expected probabilities for the positive class vs the predicted probabilities or observed values.
- Cumulative gains chart is the gain for every percentile you calculate the fraction of true positive observations up to that percentile.
- Lift curve/lift chart is the ratio of the fractions for every percentile of true positive observations up to that percentile for our model and for a random model.
- Kolmogorov-Smirnov plot and statistic assess the separation between prediction distributions for positive and negative classes.

As noticed, there are many evaluation metrics that we select from, so how to choose an evaluation metric. As stated by Brownlee in 2019, there are many evaluation metrics to choose from that can serve our particular problem, and selecting an assessment method is out of the most utter importance. Instead, selection should outline and consider what is indispensable about the model or the set of predictions given as a result. Then we can define several metrics that appear to secure it and test the metric with various set-ups. For instance, simulate a set of predictions for a test dataset with a skewed class distribution that matches the problem domain; or what occurs to the metric when the model predicts all the majority or minority classes individually if it does well or does poorly.

Alternative another tactic might be to carry out a literature review and discern what metrics are most used by other researchers and academics working on the same or similar problem, having in mind that fields of study might fall into adopting a metric that compares models at scale but is awful for model selection, so discernment is needed as well. Moreover, according to Emanet et al. (2014), the main idea to have in mind is the following:

- To use the Area under the ROC score, the positive class is the majority, and your focus class is the negative class.
- To use Precision, Recall and F1-score, the negative class is the majority, and your focus class is the positive class.

- To use TPR defines how the classifier/Model predicted a high number of True Positives instances.
- The classes should be balanced to use the Accuracy score, as it does not help much in Imbalanced Classification.

The final approach is to select the correct evaluation metric that perhaps involves a good understanding of the business problem and requirement that could impact and generate a low recall vs. low precision, and decide what metric to optimize for. Starting determining the class distribution or the ratio among the positive and negative classes and how the interaction between them is characterized. When the valuation metric uses values from both classes in relation to the confusion matrix, it will be sensitive to the imbalanced data, as the data allocations change, the selected metric will change, even if the classifier performance does not, which give us a more compelling result. (Brownlee, 2019) (Tharwat, 2020).

Methodology

"I am among those who think that science has a great beauty."
– Marie Curie

Research Design/Approach

Methodology

In order to determine to what extent various the classification algorithms for supervised learning in Machine Learning were able to handle imbalance classes in medical datasets, a set of experiments were set up, and several approaches have been considered. In the first place, a literature review that included a compilation and analysis of the relevant information provided by additional research made in this area. This approach simplified the study of this problem since it does not require executing and verifying every element that the documentation describes. However, this brings also a disadvantage, as the results are very experimental and do the actual implementation.

Furthermore, the second place, the method chosen has been experimenting with algorithms such as Support Vector Machine (SVM), Decision Tree, Gaussian Naïve Bayes (GNB), K-Nearest Neighbourhood (KNN), and Logistic Regression (LR) by simulating oversampling and down-sampling use cases and comparing the actual results with their concerning performance according to the classification metrics, and implementing minimum hyper parameterization. This method permits us to proceed into verifying models, methods and metrics and define what can be improved and tailored to the various levels of imbalances, as well as helps us to outline the steps to reproduce and replicate the results. Unfortunately, experiments set up can be convoluted and complicated, requiring familiarity with the programming language, concepts and business problems to solve.

Data Pre-processing & Exploratory Data Analysis

The main objective of creating a reference guide is to determine how to use Machine Learning classification techniques, and class rebalances strategies regarding the level/percentage of imbalanced class distributions in medical datasets. Extracting and preparing the raw data is an integral part of the process, as it enhances its quality and extraction of meaningful insights, transforming real-world data into an understandable and readable format. Therefore, the first step involved data gathering as part of the data pre-processing. In this case, the data was sourced from various locations like Kaggle, Physionet, Centres for Disease and Control Prevention, and Mendeley. The datasets are described as follows:

Cerebral Stroke Dataset: The files associated with this dataset are licensed under an Attribution-NonCommercial 3.0 Unported license. This dataset was created by HealthData.gov, utilised initially as the benchmark dataset in a Kaggle competition. The dataset is a typical primary dataset of stroke prediction with an imbalanced class type with the subject of Medicine, health and Life Sciences. It contains 11 features, where 783 occurrences of stroke were included in 43,400 recorded samples. Therefore, the problem becomes a binary classification for stroke prediction, which means the prognosis results are divided into stroke and non-stroke.

Features	values	Features	values
Patient ID	1-43400	Hypertension(hyp)	Yes/No
Gender(gen)	Male/Female	Married(mar)	Yes/No
Residence type	Urban/Rural	Age	0.08-82
Avg-glucose(glu)	55-291	Heart disease(htd)	Yes/No
Work type(work)	Private/Employed	BMI	10.1-97.6
Smoking status	Smoked/Formerly/Never		

Figure 1: Description of Cerebral Stroke dataset

According to the Google Developers classification, the cerebral stroke dataset contains a 1.84% percentage of imbalance property that represents an extreme degree (2021). The label or predictive feature is binary defined by stroke, with 0 representing No stroke and 1 representing Stroke.

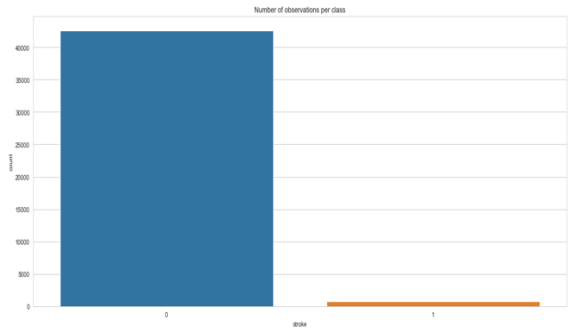


Figure 2: Cerebral Stroke Dataset label distribution

The tools for Explainable AI like SHAP (Shapley Additive Explanations) are implemented to improve model transparency and explainability. This tool helps us determine the average contributions of every feature present in the dataset, represented by the average impact on model output magnitude. In the case of the cerebral stroke dataset, based on 11 features, the most important ones concerning both classes are age with the highest, followed by BMI and average glucose level.

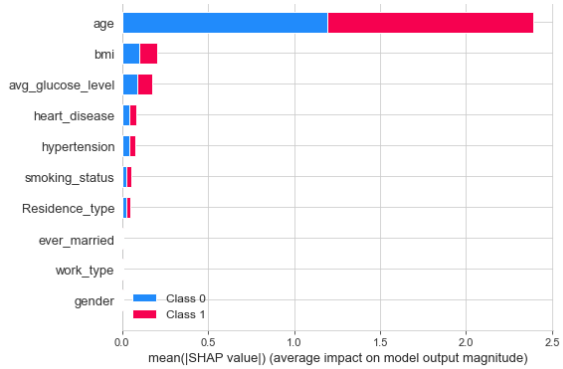


Figure 3: Cerebral Stroke Dataset Feature Importance

Diabetes Dataset: The files associated with this dataset are licensed under a Public Domain Dedication No Copyright CC0 1.0 Universal. This dataset is based on the Behavioural Risk Factor Surveillance System (BRFSS), a health-related telephone survey collected annually by the CDC. The survey collected responses from over 400,000 Americans on health-related risk behaviours, chronic health conditions, and the use of preventative services, conducted since 1984. For this project, a csv of the dataset available on Kaggle for 2015 was used.

This original dataset contains responses from 441,455 individuals and has 330 features. Initially, the data was divided into three files containing various cleaning versions and splits of the different health indicators. Finally, the file selected was diabetes_binary_health_indicators_BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable diabetes_binary has two classes. 0 is for no diabetes, and one is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced. Reiterating that the cleaned and consolidated dataset created was accessed directly from Kaggle and created with the inspiration of Zidian Xie et al. (2014) for Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques using the 2014 BRFSS.

According to the Google Developers classification, the diabetes dataset contains a 16.19% percentage of imbalance property that represents a mild degree (2021). The label or predictive feature is binary defined by Diabetes_binary, with 0 representing No diabetes and 1 representing Diabetes.

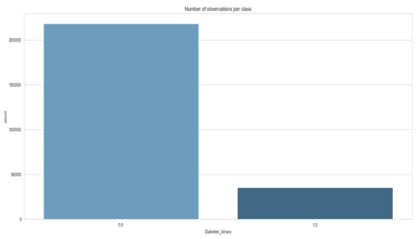


Figure 4: Diabetes Dataset label distribution

The SHAP (Shapley Additive Explanations) values that represented the average impact on model output magnitude in the case of the diabetes dataset that is based on 20 features, the most important ones concerning both classes are General Health, with the highest, followed by High Blood Pressure, Age, BMI and High Cholesterol. The values are calculated based on 20 features.

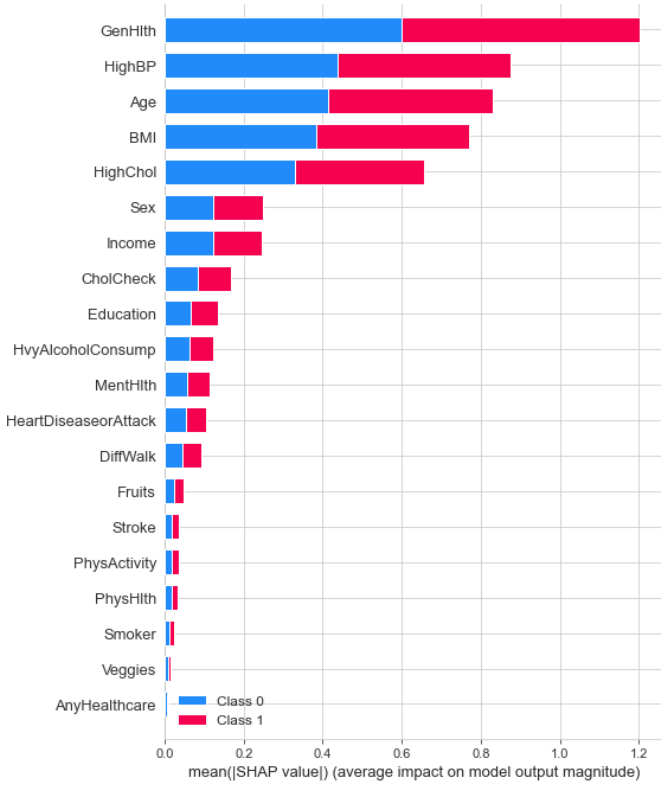


Figure 5: Diabetes Dataset Feature Importance

Sepsis Dataset: The files associated with this dataset are licensed under a Creative Commons Attribution-Share Alike 4.0 International Public License. This dataset is based on the Computing in Cardiology Challenge from Physionet 2019. The goal is the early detection of sepsis using physiological data. Sepsis is defined according to the Sepsis-3 guidelines in the patient's Sequential Organ Failure Assessment (SOFA) score and clinical suspicion of infection by ordering blood cultures or IV antibiotics.

Furthermore, the data was sourced from ICU patients in three separate hospital systems. Available patient co-variates consist of Demographics, Vital Signs, and Laboratory values.

According to the Google Developers classification, the sepsis dataset contains a 7.87% percentage of imbalance property that represents a moderate degree (2021). The label or predictive feature is binary defined by sepsis, with 0 representing No sepsis and 1 representing Sepsis.

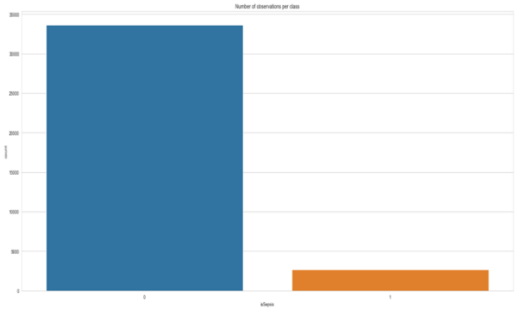


Figure 6: Sepsis Dataset label distribution

In the sepsis dataset case, the SHAP (Shapley Additive Explanations) values that represented the average impact on the model output magnitude were ICULOS (ICU Length of Stay) with the most significant influence by far, followed by WBC (Leukocyte count), Temperature and Hospital Admission time (Hours between hospital admit and ICU admit). Therefore, after removing features with the high missing values, the values are calculated based on 29 features.

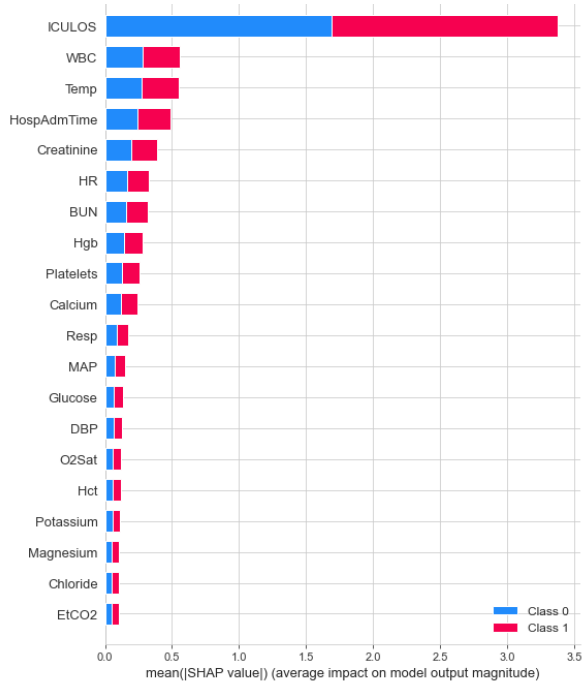
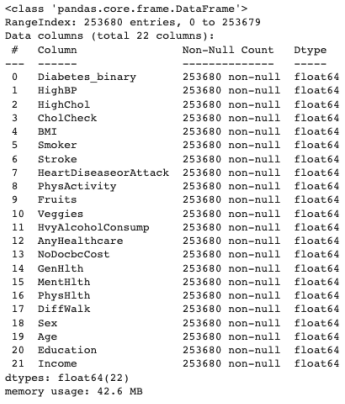


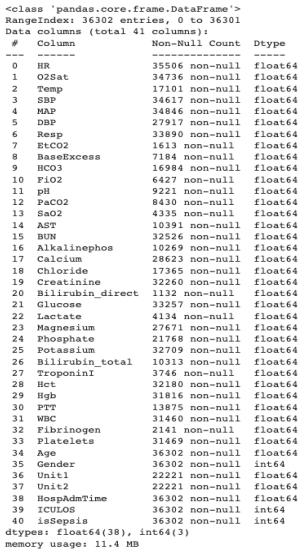
Figure 7: Sepsis Dataset Feature Importance

When it comes to identifying and handling the missing values, as well as Encoding categorical data, the three datasets have different states, for instance:

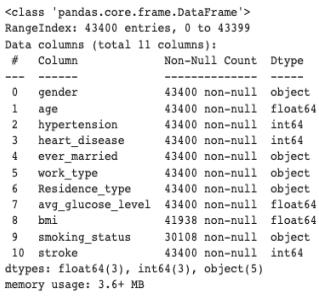
- The diabetes dataset has no missing values and no categorical variables to be modified.
- In the sepsis dataset, from 41 features, only 5 had no missing values. The features BaseExcess, pH, PaCO2, SaO2, AST, Alkalinephos, Bilirubin_direct, Lactate, Bilirubin_total, TroponinI, Fibrinogen were dropped because they had over 70% of missing values. The other features were imputed using the SimpleImputer and median as a strategy. There is no presence of categorical variables.
- Finally, the cerebral stroke dataset contains 11 features, with two missing values. First, the BMI was imputed using the mode and secondly, for the smoking status created a new category represented by unknown. The categorical variables reached 5 of them and were transformed to numerical, replacing their values with mappings that go from 1 to 5 as descriptors.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36302 entries, 0 to 36301
Data columns (total 41 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   HR                 35506 non-null  float64
 1   O2Sat              34736 non-null  float64
 2   Temp               17101 non-null  float64
 3   SBP                34617 non-null  float64
 4   MAP                34846 non-null  float64
 5   DBP                27917 non-null  float64
 6   Resp               33890 non-null  float64
 7   EtCO2              1613 non-null   float64
 8   BaseExcess         7184 non-null   float64
 9   HCO3               16984 non-null  float64
 10  FiO2               6427 non-null   float64
 11  pH                 9221 non-null   float64
 12  PaCO2              8430 non-null   float64
 13  SaO2               4335 non-null   float64
 14  AST                10391 non-null  float64
 15  BUN                32526 non-null  float64
 16  Alkalinephos       10269 non-null  float64
 17  Calcium            28623 non-null  float64
 18  Chloride           17365 non-null  float64
 19  Creatinine         32260 non-null  float64
 20  Bilirubin_direct   1132 non-null   float64
 21  Glucose            33257 non-null  float64
 22  Lactate            4134 non-null   float64
 23  Magnesium          27671 non-null  float64
 24  Phosphate          21768 non-null  float64
 25  Potassium          32709 non-null  float64
 26  Bilirubin_total    10313 non-null  float64
 27  TroponinI          3746 non-null   float64
 28  Hct                32180 non-null  float64
 29  Hgb                31816 non-null  float64
 30  PTT                13875 non-null  float64
 31  WBC                31460 non-null  float64
 32  Fibrinogen         2141 non-null   float64
 33  Platelets          31469 non-null  float64
 34  Age                36302 non-null  float64
 35  Gender             36302 non-null  int64
 36  Unit1              22221 non-null  float64
 37  Unit2              22221 non-null  float64
 38  HospAdmTime        36302 non-null  float64
 39  ICULOS             36302 non-null  int64
 40  isSepsis           36302 non-null  int64
dtypes: float64(38), int64(3)
memory usage: 11.4 MB
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43400 entries, 0 to 43399
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   gender             43400 non-null  object
 1   age                43400 non-null  float64
 2   hypertension       43400 non-null  int64
 3   heart_disease      43400 non-null  int64
 4   ever_married       43400 non-null  object
 5   work_type          43400 non-null  object
 6   Residence_type     43400 non-null  object
 7   avg_glucose_level  43400 non-null  float64
 8   bmi                41938 non-null  float64
 9   smoking_status     30108 non-null  object
 10  stroke             43400 non-null  int64
dtypes: float64(3), int64(3), object(5)
memory usage: 3.6+ MB
```

Diabetes Dataset Sepsis Dataset Cerebral Stroke Dataset

Table 2: Summary information of the Dataframes

Concerning the procedure of splitting of the dataset, the subsets train/test were divided 80/20 percentage, with a random state set up for reproducibility purposes, and the stratification implemented, feature for classification problems only, that in the presence of imbalance classes, the train set and the test set will preserve similar proportions that the ones observed in the original dataset. The Feature scaling was performed using the RobustScaler() transform function over the datasets, removing the median and scaling the data based on the quantile range, used in cases where the features are robust to outliers.

```
# split the dataset into a training and test sets.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, shuffle=True, stratify=y, random_state=13)
print('Train:', X_train.shape, y_train.shape)
print('Test:', X_test.shape, y_test.shape)
```

```
# perform a robust scaler transform of the dataset
rs = RobustScaler()
X_train = rs.fit_transform(X_train)
X_test = rs.transform(X_test)
```

Figure 8: Train/Test Dataset Split - Dataset Normalization

Experiments and Results

Experiment Setup

We set up the experiments based on the different percentages of imbalance properties present in the datasets, such as diabetes, sepsis, and cerebral stroke. That can be described as the only differentiator between them. The three of them are created for binary classification type of machine learning algorithm and having medical descriptor of the diseases.

After exploratory data analysis, we applied various sampling techniques to the datasets, missing value evaluation, feature importance, and feature normalization. Finally, we imported packages from the sklearn model library for:

- Support Vector Machine Classifier
- Decision Tree Classifier
- Gaussian Naïve Bayes
- K-Nearest Neighbours Classifier
- Logistic Regression

As well as Imbalanced-learn open-source library for the over-sampling, under-sampling, combination/hybrid, and ensemble techniques, like:

- Oversampling: RandomOverSampler, SMOTE, SMOTENC, BorderlineSMOTE, SVMSMOTE, KMeansSMOTE, ADASYN
- Undersampling: RandomUnderSampler, ClusterCentroids, NearMiss, InstanceHardnessThreshold, TomekLinks, CondensedNearestNeighbour, AllKNN, EditedNearestNeighbours, RepeatedEditedNearestNeighbours, OneSidedSelection, NeighbourhoodCleaningRule
- Combine/Hybrid: SMOTEENN, SMOTETomek
- Ensemble: EasyEnsembleClassifier, RUSBoostClassifier, BalancedBaggingClassifier, BalancedRandomForestClassifier

The experiments are processed through four functions that include:

- Evaluation of the trained classification model (evaluate_classificationmodel) returning results on the following metrics accuracy, precision, recall, f1 score, number of occurrences, predictions count, confusion matrix (true positive, true negative, false positive, false negative), and area under the curve.

- Resampling techniques pipeline (resampling_techniques_pipeline) takes the split dataset and the trained classification model. Then, it returns the before and after results after applying the different oversampling, undersampling, hybrid and ensemble techniques over the dataset. This function inherits the results of the function mentioned above (evaluate_classificationmodel).
- Evaluation of the Resampling Method, that the results obtained after resampling_techniques_pipeline function execution, plot the metrics selected, in this case, precision, recall and fscore, followed by the area under the curve concerning the adjusted class weights.
- The metrics dataframe (metrics_dataframe) extracts the results obtained from the various resampling techniques and creates a dataframe that displays them as a table.

Results

The performance metrics used during the experiments for classification machine learning algorithms were:

- Precision, what proportion of positive identifications was correct?
- Recall, what proportion of actual positives were identified correctly?
- F1-score, the harmonic mean of the combination of both.

As per the results, descriptions of the insights will be described using the F1-score metrics to measure the effectiveness of a model with respect to the sampling techniques, as it gives a larger weight to lower numbers.

Furthermore, as a reminder, the diabetes dataset is considered mild, the sepsis dataset is deemed to be moderate and the cerebral stroke is extreme on the imbalance properties scale.

Undersampling Techniques

Support Vector Machine

In the diabetes dataset for the undersampling techniques, the CondensedNearestNeighbour scored the highest with 0.856 with the lowest false positive count with 1025 data points. Followed by RandomUnderSampler with 0.791 and OneSidedSelection with 0.789.

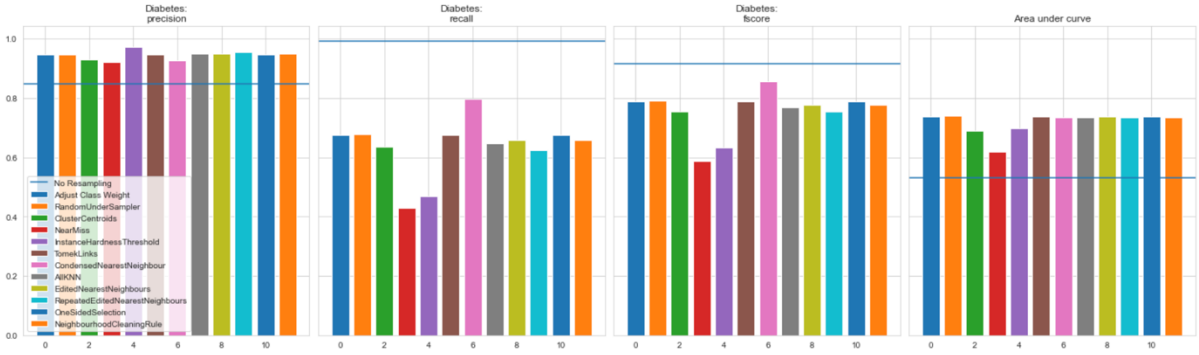


Figure 9: Support Vector Machine Algorithm - Diabetes Dataset - Undersampling Technique

In the sepsis dataset for the undersampling techniques, CondensedNearestNeighbour performed the best with 0.95 as a score. Along those lines, TomekLinks and OneSidedSelection got over 0.926. It is worth noticing the lowest score was from NearMiss with 0.45.

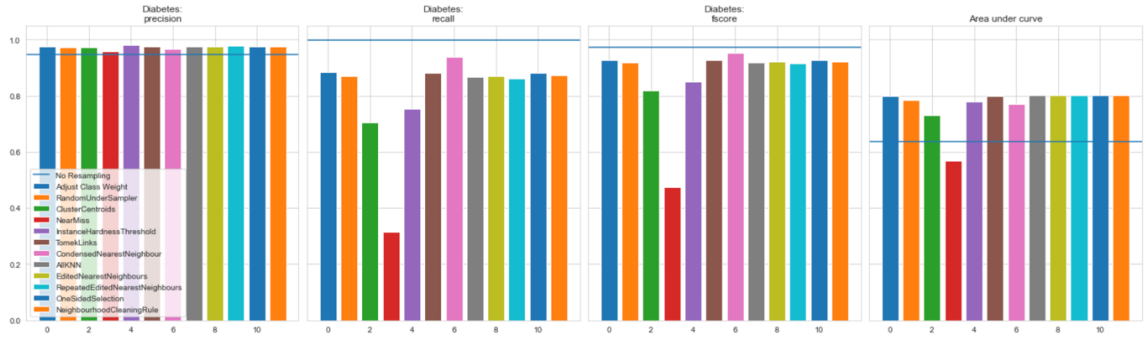


Figure 10: Support Vector Machine Algorithm - Sepsis Dataset - Undersampling Technique

In the cerebral stroke dataset for the undersampling techniques, CondensedNearestNeighbour has got 0.90, next OneSidedSelection with 0.866 as score and RepeatedEditedNearestNeighbours with 0.863. It is worth noticing the lowest score was from NearMiss with 0.35.

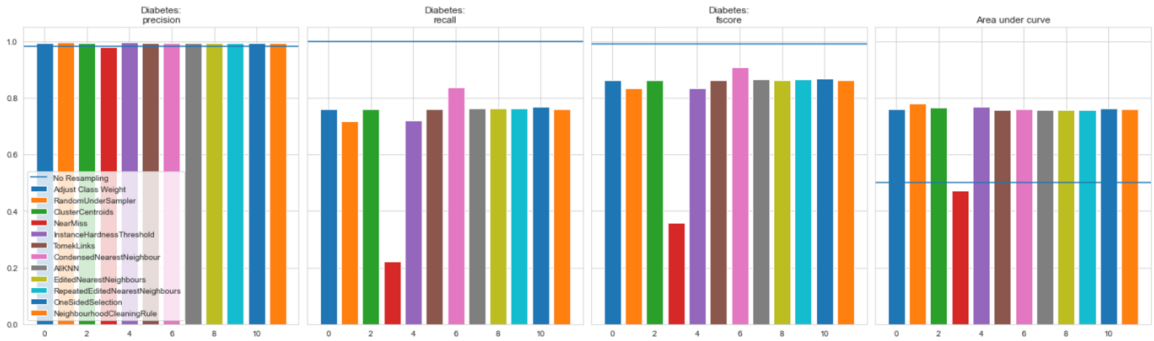


Figure 11: Support Vector Machine Algorithm - Cerebral Stroke Dataset - Undersampling Technique

Decision Tree

In the diabetes dataset for the undersampling techniques, the highest scores techniques were OneSidedSelection with 0.8685 and TomekLinks with 0.8644. The lower score techniques were ClusterCentroids with 0.411 and NearMiss with 0.555 as a score.

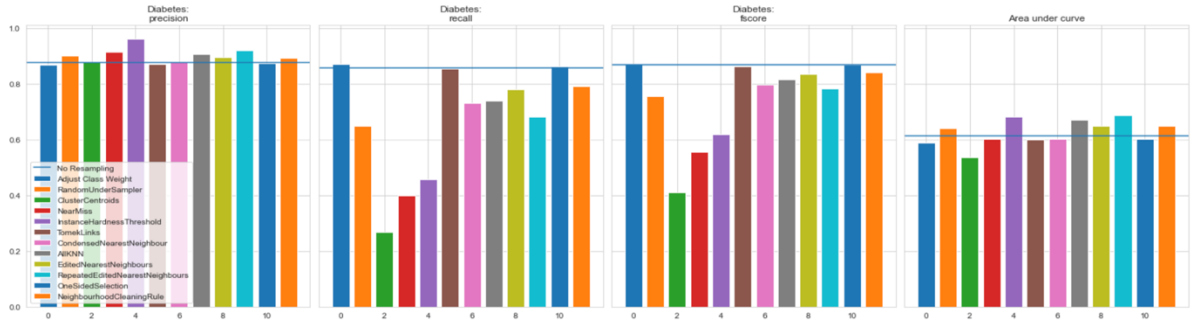


Figure 12: Decision Tree Algorithm - Diabetes Dataset - Undersampling Technique

In the sepsis dataset for the undersampling techniques, OneSidedSelection reached the highest score with 0.9719, TomekLinks with 0.9716, and finally EditedNearestNeighbours with 0.9713. The other methods performed as great as the top ones, with scores around 0.97 as well. It is worth noticing the lowest score was from ClusterCentroids with 0.58.

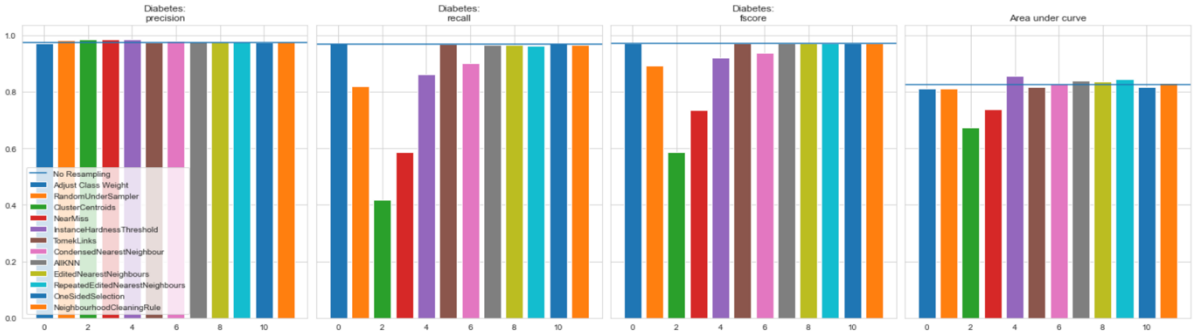


Figure 13: Decision Tree Algorithm - Sepsis Dataset - Undersampling Technique

In the cerebral stroke dataset for the oversampling techniques, RandomOverSampler has got 0.98, and BorderlineSMOTE with 0.978 as a score. It is worth noticing the lowest score was from NearMiss with 0.22.

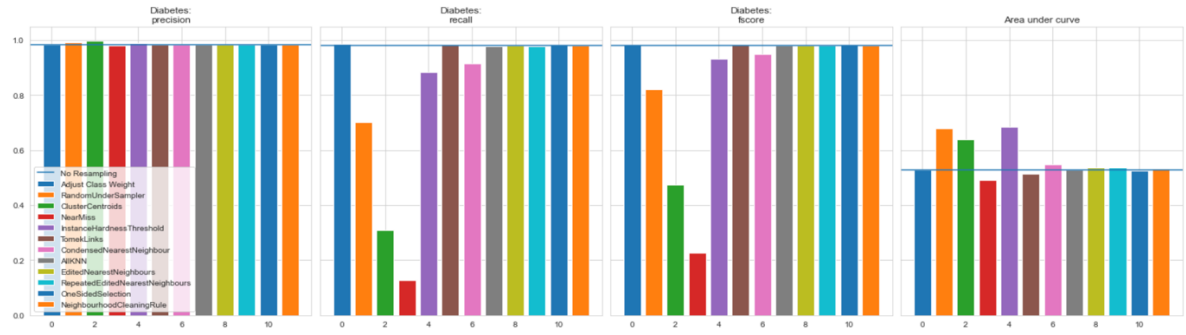


Figure 14: Decision Tree Algorithm - Cerebral Stroke Dataset - Undersampling Technique

Gaussian Naïve Bayes (GNB)

In the diabetes dataset for the undersampling techniques, the highest scores techniques both were OneSidedSelection and TomekLinks with 0.7648 as a score. The lower score techniques were ClusterCentroids with 0.59 and NearMiss with 0.51 as a score.

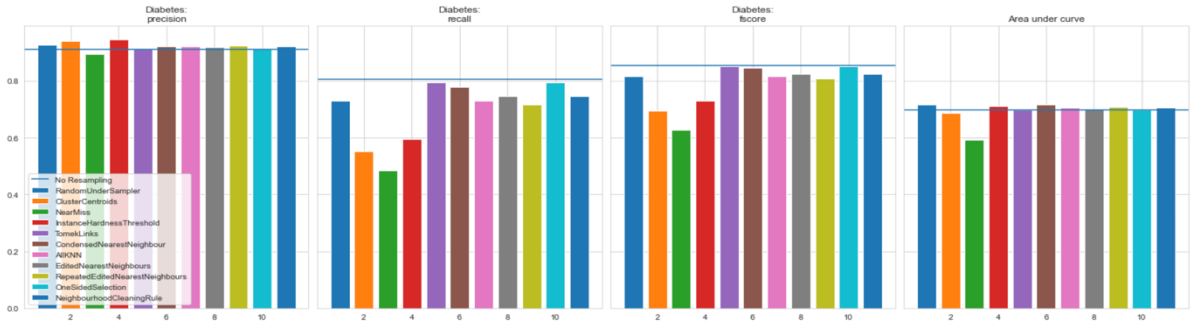


Figure 15: GNB Algorithm - Diabetes Dataset - Undersampling Technique

In the sepsis dataset for the undersampling techniques, OneSidedSelection reached the highest score with 0.948, TomekLinks with 0.9475, and finally EditedNearestNeighbours with 0.9472. The other methods performed as great as the top ones, with scores around 0.94 as well. It is worth noticing the lowest score was from ClusterCentroids with 0.45.

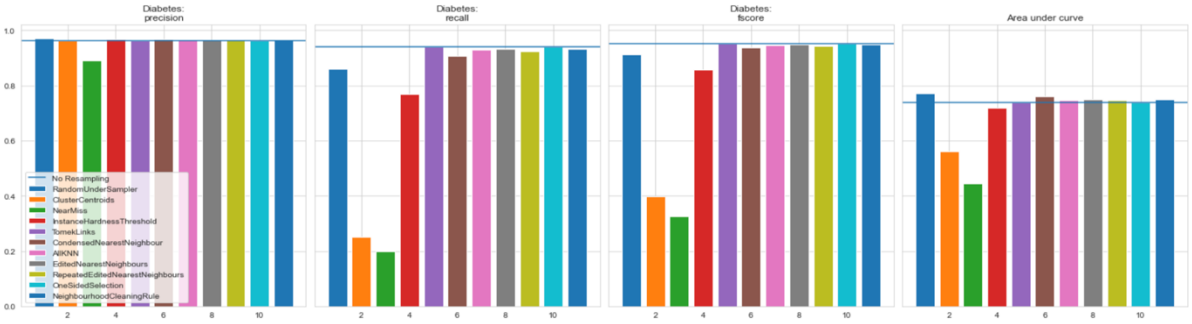


Figure 16: GNB Algorithm - Sepsis Dataset - Undersampling Technique

In the sepsis dataset for the undersampling techniques, OneSidedSelection reached the highest score with 0.9660 and TomekLinks with 0.9653. The other methods performed as great as the top ones, with scores around 0.96 as well. It is worth noticing the lowest score was from ClusterCentroids with 0.14.

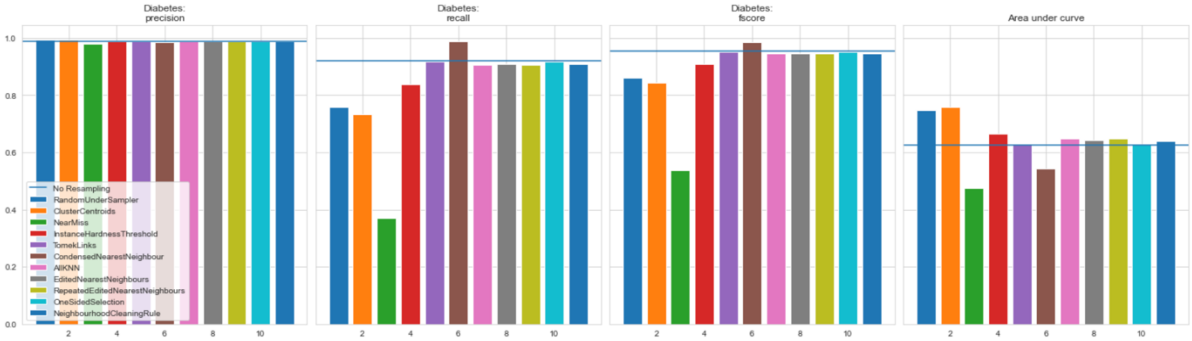


Figure 17: GNB Algorithm - Cerebral Stroke Dataset - Undersampling Technique

K-Nearest Neighbourhood (KNN)

In the diabetes dataset for the undersampling techniques, the highest scores techniques both were OneSidedSelection and TomekLinks with 0.82 as a score. The other methods performed as great as the top ones, with scores around ~0.70 as well. It is worth noticing the lowest score position was shared by NearMiss and InstanceHardnessThreshold 0.594.

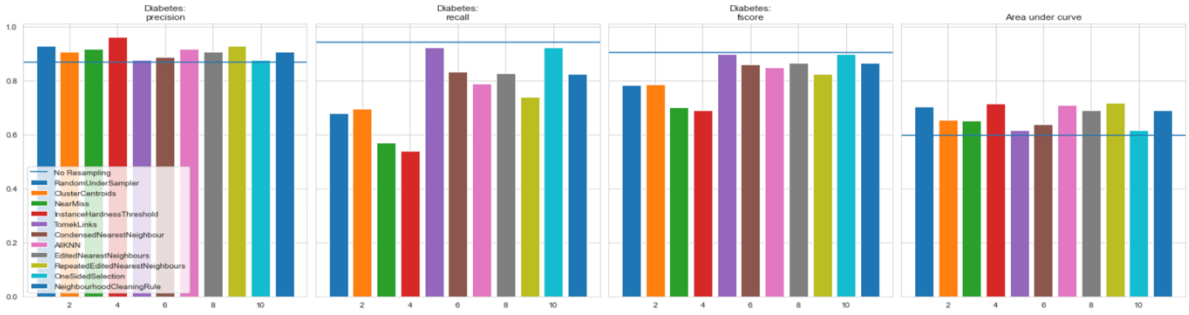


Figure 18: KNN Algorithm - Diabetes Dataset - Undersampling Technique

In the sepsis dataset for the undersampling techniques, OneSidedSelection and TomekLinks reached the highest score with 0.95, and finally EditedNearestNeighbours with 0.9420. The other methods performed as great as the top ones, with scores around ~0.93 as well. It is worth noticing the lowest score was from ClusterCentroids with 0.46.

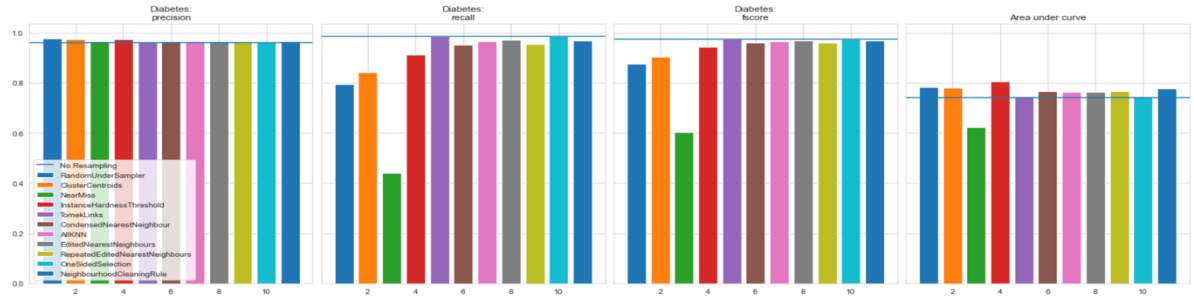


Figure 19: KNN Algorithm - Sepsis Dataset - Undersampling Technique

In the sepsis dataset for the undersampling techniques, OneSidedSelection and TomeLinks reached the highest score with 0.9981. The other methods performed as great as the top ones, with scores around 0.97 as well. It is worth noticing the lowest score was from NearMiss with 0.29.

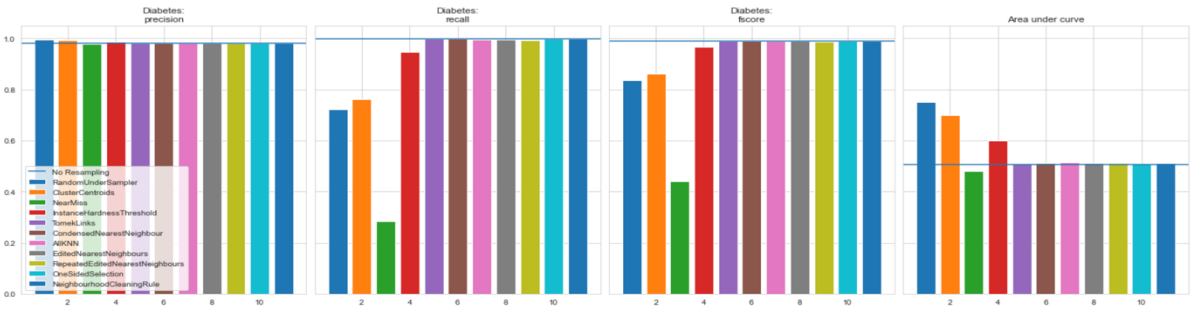


Figure 20: KNN Algorithm - Cerebral Stroke Dataset - Undersampling Technique

Logistic Regression (LR)

In the diabetes dataset for the undersampling techniques, the highest scores techniques both were OneSidedSelection with 0.78 and TomekLinks with 0.77 as a score. The lower score techniques were ClusterCentroids 0.35. The other sampling techniques got inconsistent performance.

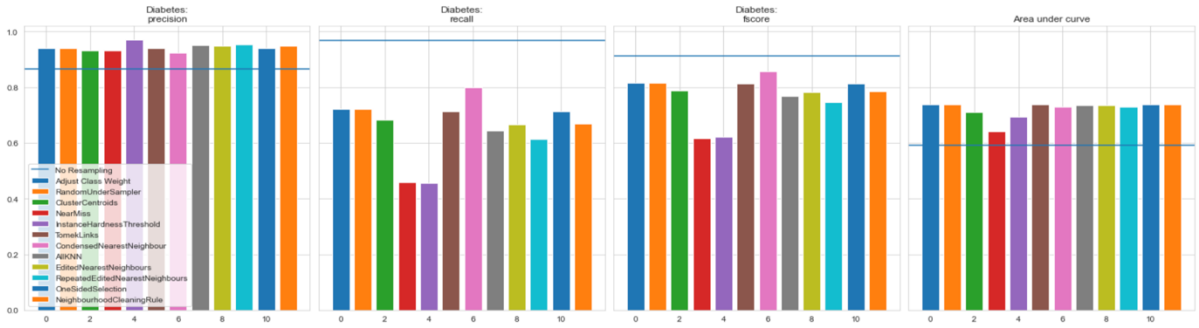


Figure 21: Logistic Regression Algorithm - Diabetes Dataset - Undersampling Technique

In the sepsis dataset for the undersampling techniques, most of the sampling techniques got over 0.94 as a score, such as OneSidedSelection, EditedNearestNeighbours and TomekLinks. It is worth noticing the lowest score was from ClusterCentroids with 0.45.

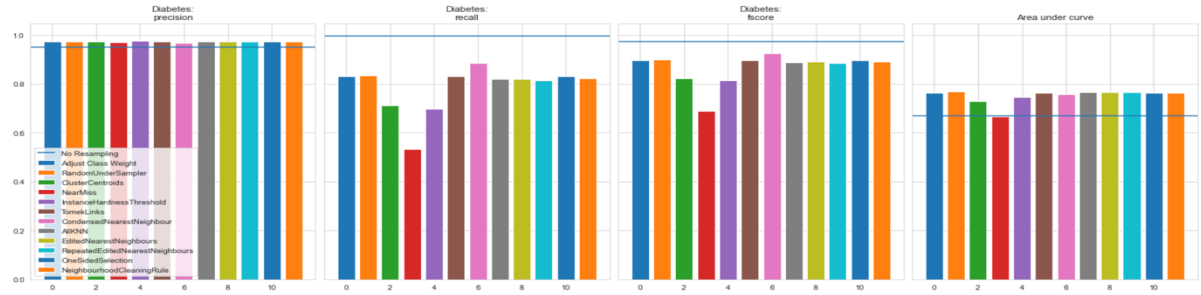


Figure 22: Logistic Regression Algorithm - Sepsis Dataset - Undersampling Technique

In the sepsis dataset for the undersampling techniques, OneSidedSelection with 0.9660 and TomeLinks with 0.9653 reached the highest score. The other methods performed as great as the top ones, with scores around ~0.94 as well. The lowest score was from NearMiss with 0.14.

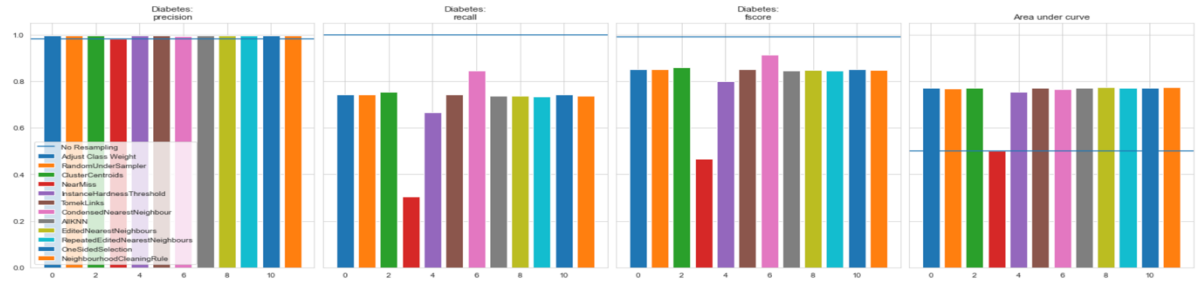


Figure 23: Logistic Regression Algorithm - Cerebral Stroke Dataset - Undersampling Technique

Oversampling Techniques

Support Vector Machine

In the diabetes dataset for the oversampling techniques, the SVMSMOTE scored the highest with 0.85 with the lowest false positive count with 1047 data points. Followed by KMeansSMOTE with 0.84 and SMOTENC with 0.81.

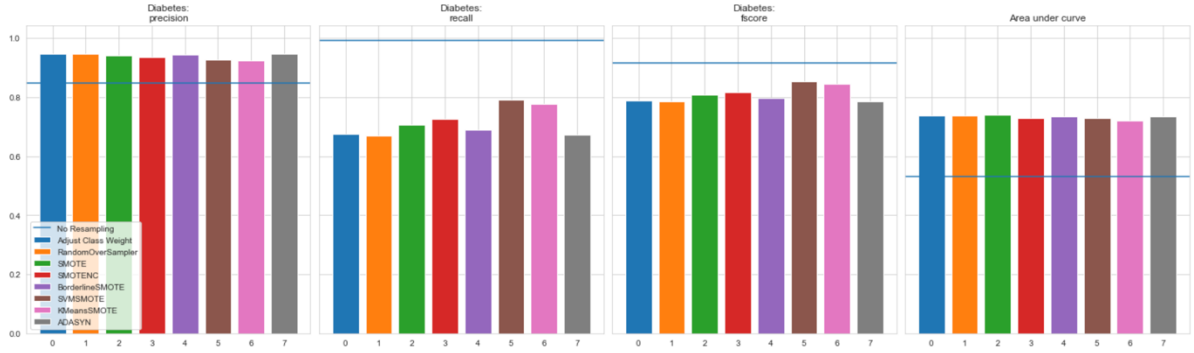


Figure 24: Support Vector Machine Algorithm - Diabetes Dataset - Oversampling Technique

In the sepsis dataset for the oversampling techniques, RandomOverSampler reached the highest score with 0.932, KMeansSMOTE with 0.9314, and finally SMOTE and SMOTENC with 0.9311. The other methods performed as great as the top ones, with a score bigger than 0.88.

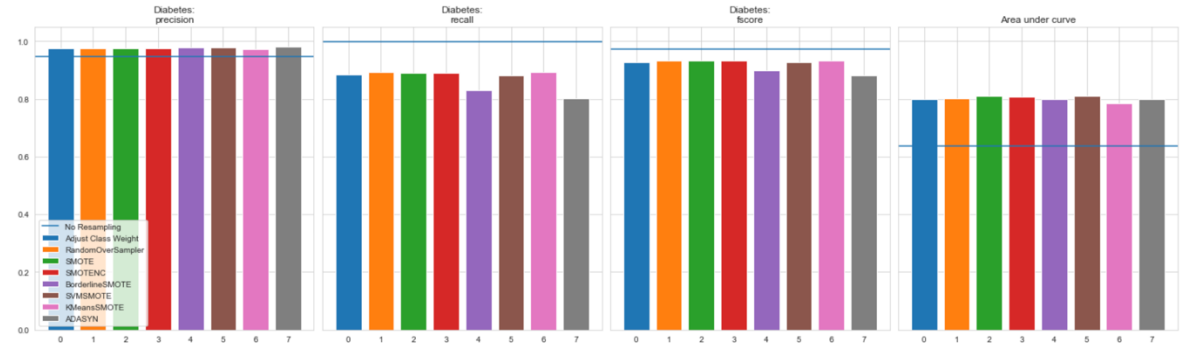


Figure 25: Support Vector Machine Algorithm - Sepsis Dataset - Oversampling Technique

In the cerebral stroke dataset for the oversampling techniques, BorderlineSMOTE scored 0.923 vs SVMSMOTE with 0.920. Followed by KMeansSMOTE with 0.89. The other method's lowest score had a minimum of 0.86, reached by the RandomOverSampler.

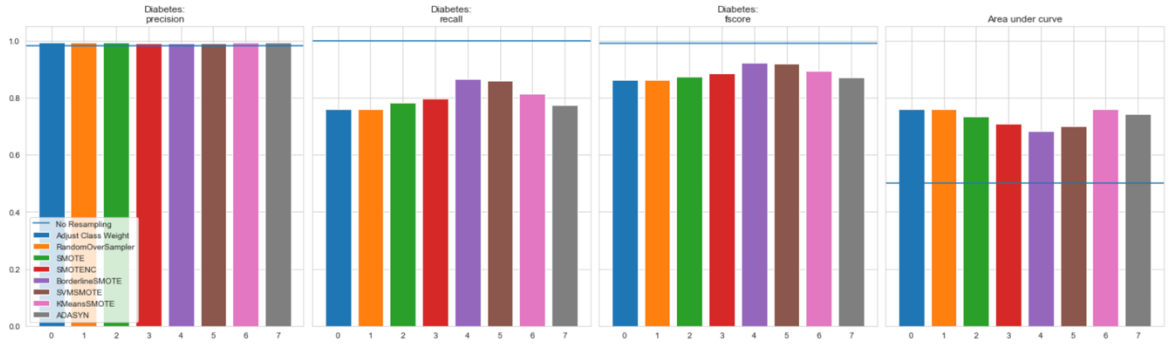


Figure 26: Support Vector Machine Algorithm - Cerebral Stroke Dataset - Oversampling Technique

Decision Tree

In the diabetes dataset for the oversampling techniques, RandomOverSampler performed the best with 0.868 as a score, along those lines SVMSMOTE and SMOTENC got over 0.8638 and 0.8606, respectively. Moreover, the lowest score was 0.8598, not much of a difference in comparison.

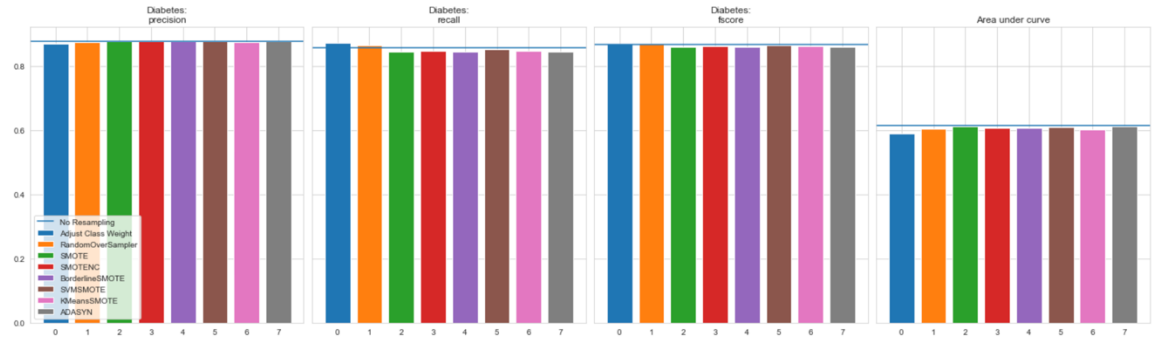


Figure 27: Decision Tree Algorithm - Diabetes Dataset - Oversampling Technique

In the sepsis dataset for the oversampling techniques, RandomOverSampler performed the best with 0.97 as a score, along those lines KMeansSMOTE and SVMSMOTE got 0.9696 and 0.9648 respectively.

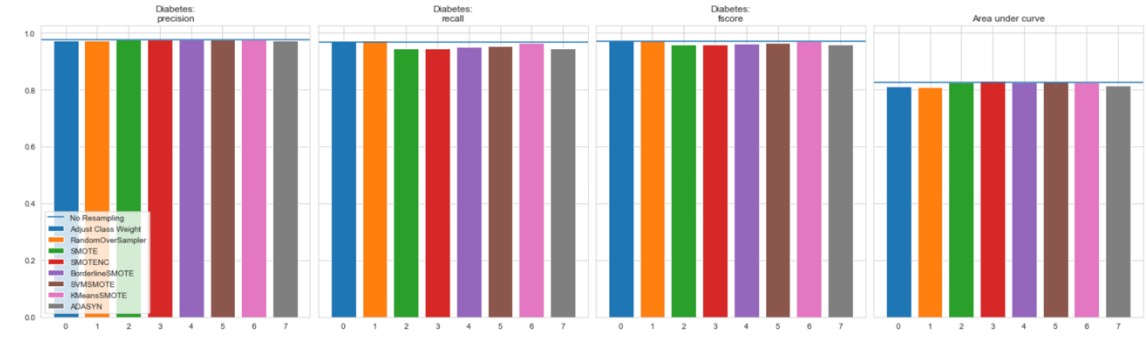


Figure 28: Decision Tree Algorithm - Sepsis Dataset - Oversampling Technique

In the cerebral stroke dataset for the oversampling techniques, RandomOverSampler has got 0.98, and BorderlineSMOTE with 0.978 as a score. It is worth noticing the lowest score was from NearMiss with 0.22.

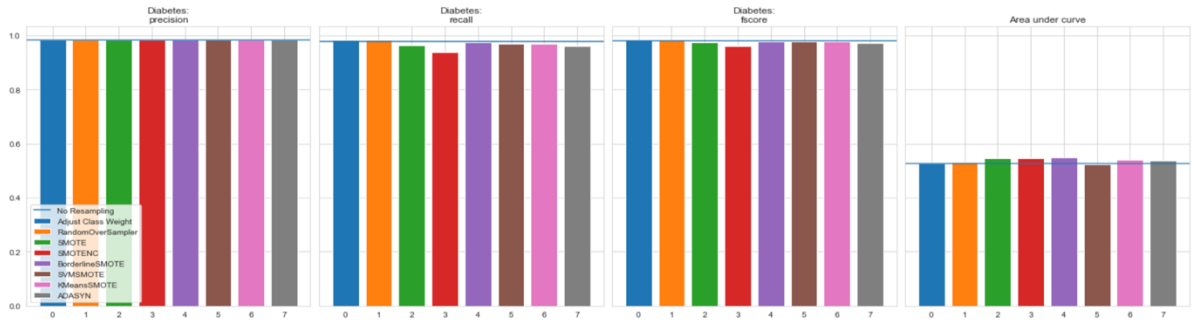


Figure 29: Decision Tree Algorithm - Cerebral Stroke Dataset - Oversampling Technique

Gaussian Naïve Bayes (GNB)

In the diabetes dataset for the oversampling techniques, KMeansSMOTE performed the best with 0.76 as a score, along those lines SVMSMOTE and RandomOverSampler got over 0.74 and 0.72, respectively. Moreover, the other sampling techniques got around 0.65 as a score.

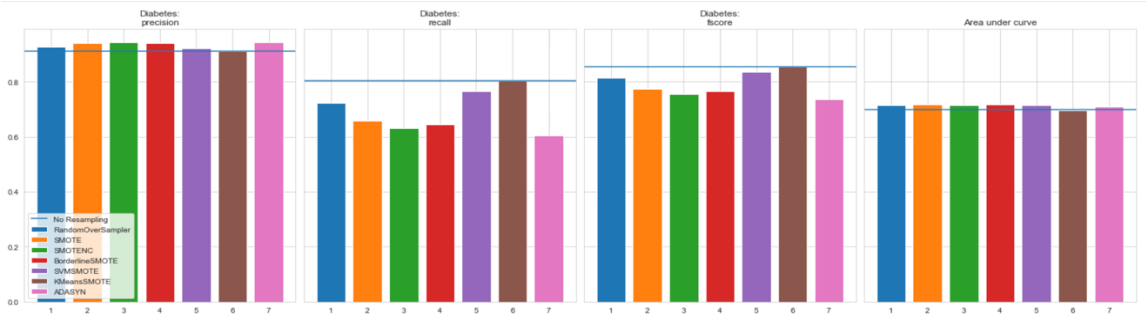


Figure 30: GNB Algorithm - Diabetes Dataset - Oversampling Technique

In the sepsis dataset for the oversampling techniques, RandomOverSampler performed the best with 0.95 as a score, along those lines KMeansSMOTE got 0.94. The rest of the sampling techniques got consistent performance with the lowest score being 0.91.

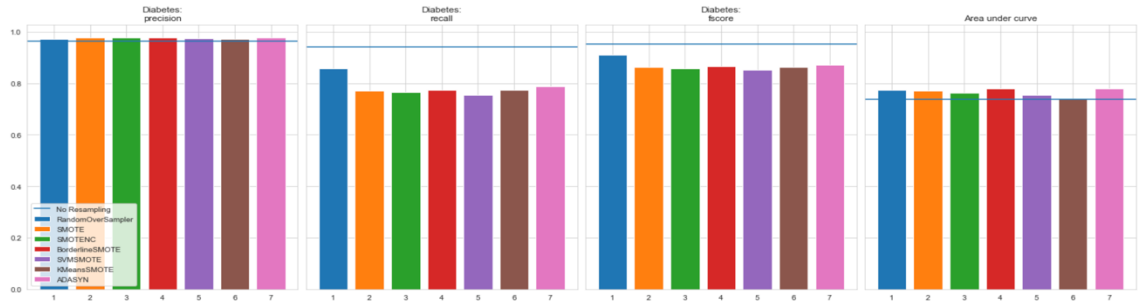


Figure 31: GNB Algorithm - Sepsis Dataset - Oversampling Technique

In the cerebral stroke dataset for the oversampling techniques, RandomOverSampler has got 0.98, and BorderlineSMOTE with 0.978 as a score. It is worth noticing the lowest score was from NearMiss with 0.22.

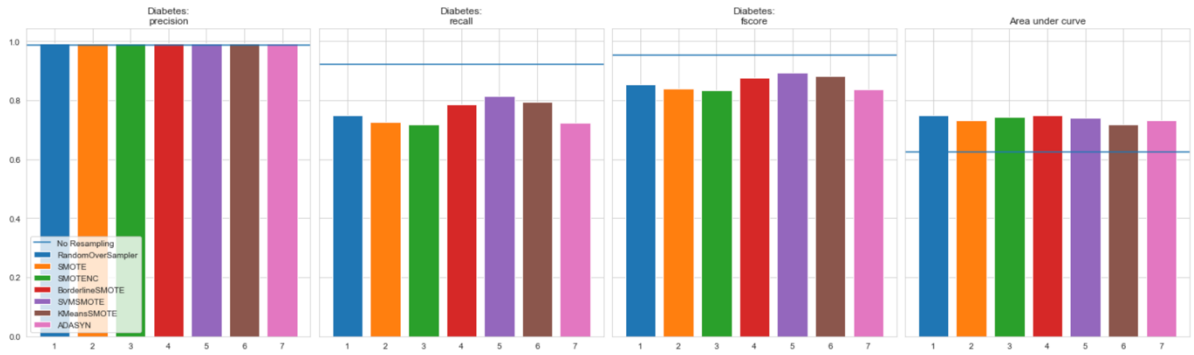


Figure 32: GNB Algorithm - Cerebral Stroke Dataset - Oversampling Technique

K-Nearest Neighbourhood (KNN)

In the diabetes dataset for the oversampling techniques, KMeansSMOTE performed the best with 0.77 as a score, along those lines SVMSMOTE got over 0.73. Moreover, SMOTE technique got around 0.69, the lowest score.

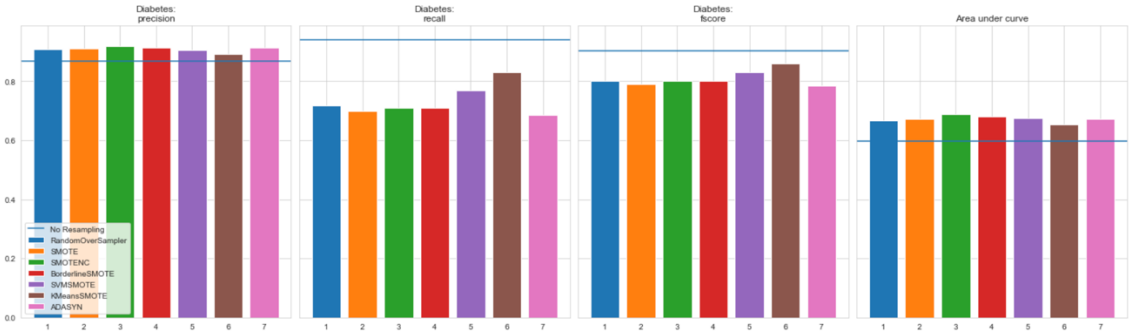


Figure 33: KNN Algorithm - Diabetes Dataset - Oversampling Technique

In the sepsis dataset for the oversampling techniques, RandomOverSampler performed the best with 0.88 as a score, followed by SVMSMOTE and KMeansSMOTE got 0.87 and 0.86, respectively. The rest of the sampling techniques got consistent performance with the lowest score being 0.78.

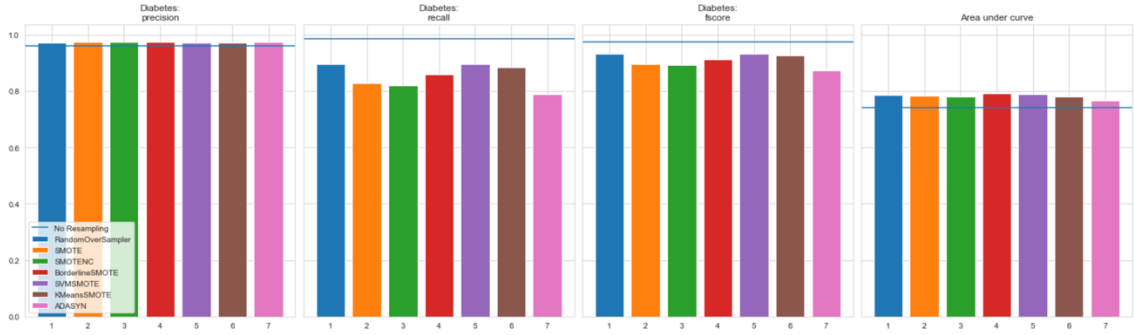


Figure 34: KNN Algorithm - Sepsis Dataset - Oversampling Technique

In the cerebral stroke dataset for the oversampling techniques, RandomOverSampler has got 0.941, and SVMSMOTE with 0.94 as a score. The rest of the sampling techniques got consistent performance with the lowest score being ~0.86.

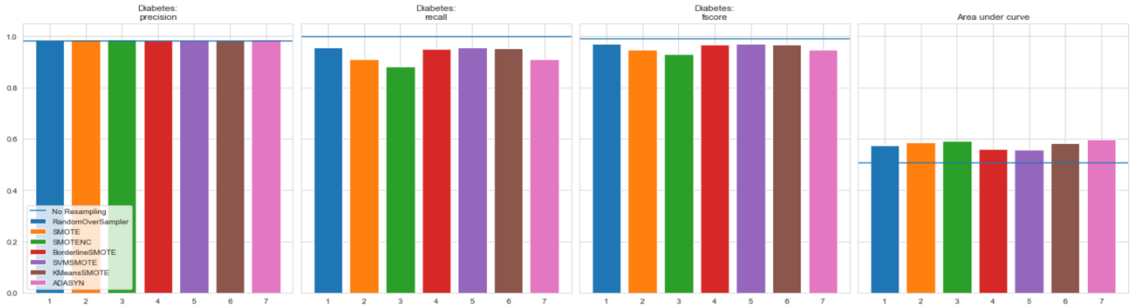


Figure 35: KNN Algorithm - Cerebral Stroke Dataset - Oversampling Technique

Logistic Regression (LR)

In the diabetes dataset for the oversampling techniques, RandomOverSampler performed the best with 0.78 as a score, followed by SVMSMOTE 0.77. Moreover, the other sampling techniques got around ~0.76 as a score.

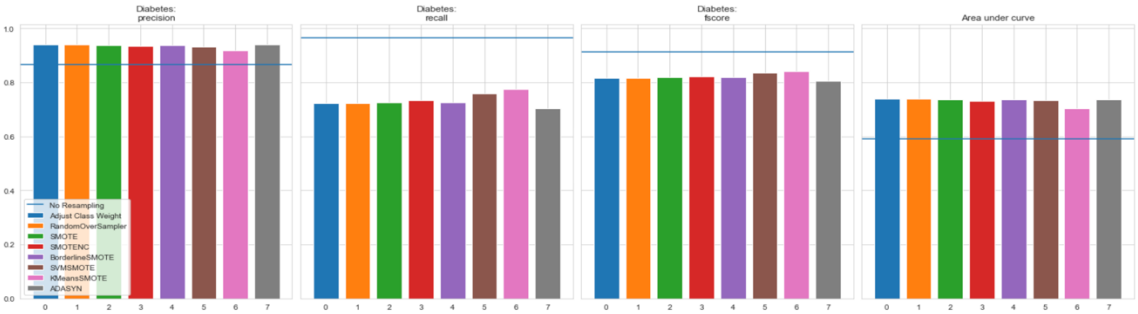


Figure 36: Logistic Regression Algorithm - Diabetes Dataset - Oversampling Technique

In the sepsis dataset for the oversampling techniques, RandomOverSampler performed the best with 0.95 as a score, followed by KMeansSMOTE got 0.94. The other sampling techniques got consistent performance with the lowest score being 0.92.

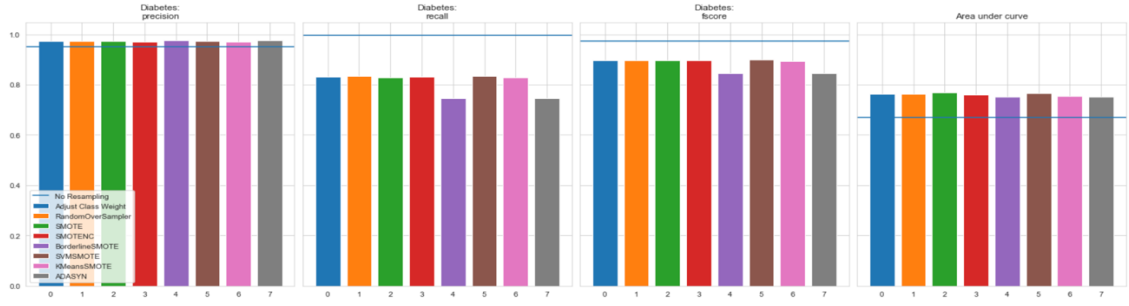


Figure 37: Logistic Regression Algorithm - Sepsis Dataset - Oversampling Technique

In the cerebral stroke dataset for the oversampling techniques, RandomOverSampler has got 0.96, and KMeansSMOTE, SVMSMOTE and SMOTE with 0.95 as a score. The rest of the sampling techniques got consistent performance with the lowest score being ~0.92.

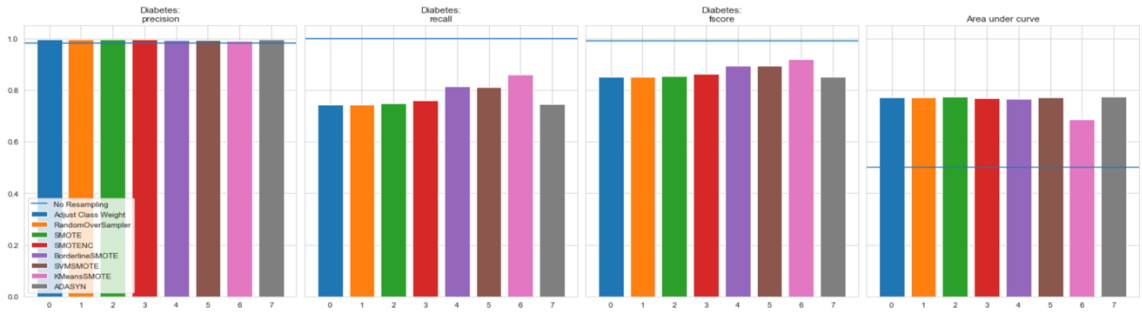


Figure 38: Logistic Regression Algorithm - Cerebral Stroke Dataset - Oversampling Technique

Hybrid Techniques

Support Vector Machine

In the diabetes dataset for the hybrid techniques, SMOTETomek scored the highest with 0.80. Followed by SMOTEENN with 0.78.

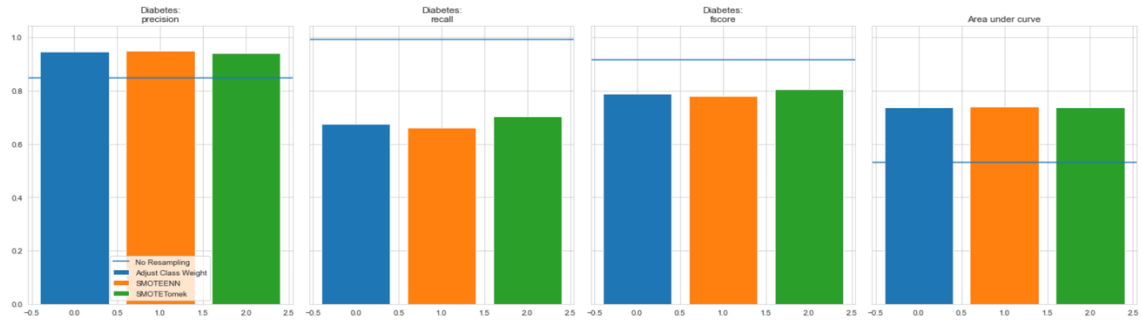


Figure 39: Support Vector Machine Algorithm - Diabetes Dataset - Hybrid Technique

In the sepsis dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.92 and SMOTEENN with 0.90.

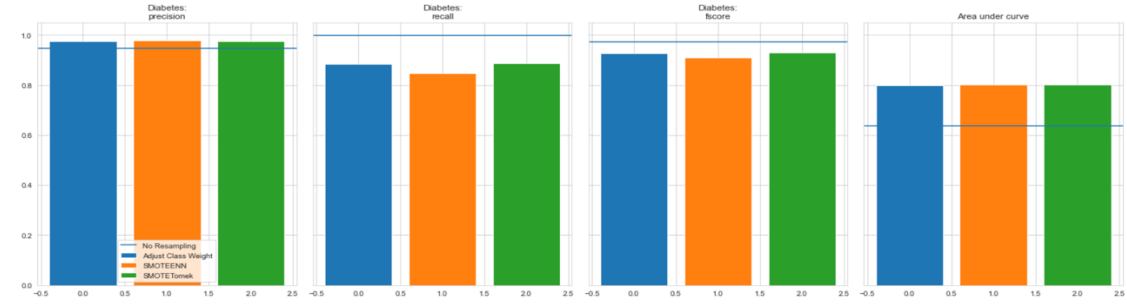


Figure 40: Support Vector Machine Algorithm - Sepsis Dataset - Hybrid Technique

In the cerebral stroke dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.873 and SMOTEENN with 0.872.

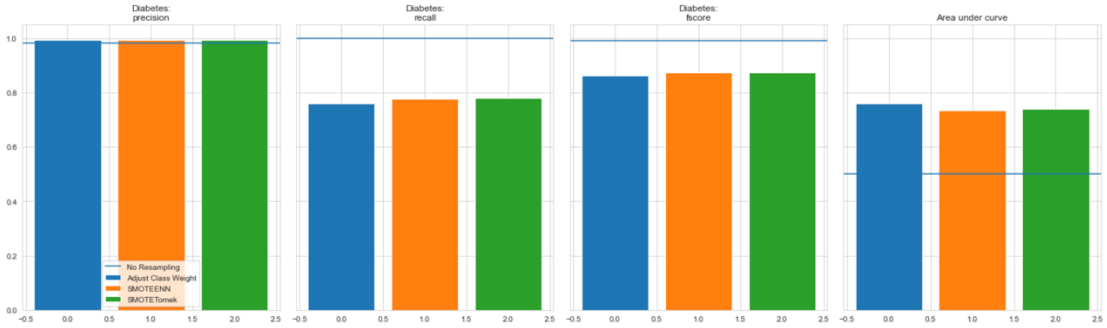


Figure 41: Support Vector Machine Algorithm - Cerebral Stroke Dataset - Hybrid Technique

Decision Tree

In the diabetes dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.86. Followed by SMOTEENN with 0.81.

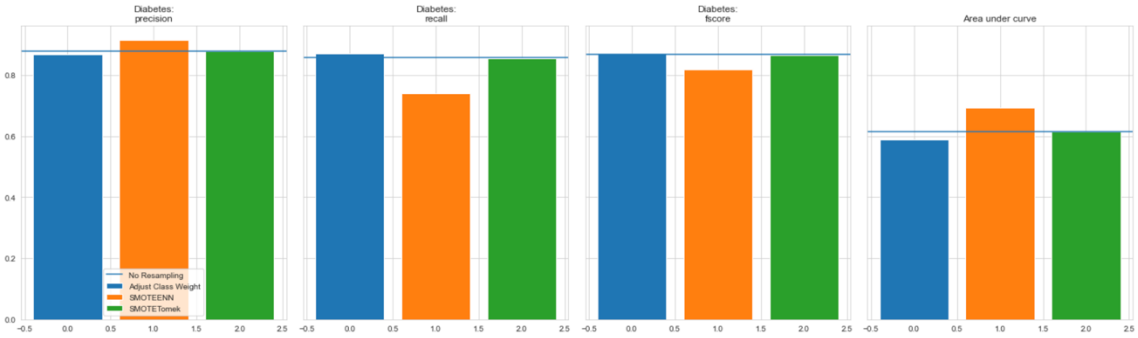


Figure 42: Decision Tree Algorithm - Diabetes Dataset - Hybrid Technique

In the sepsis dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.96, and SMOTEENN with 0.95.

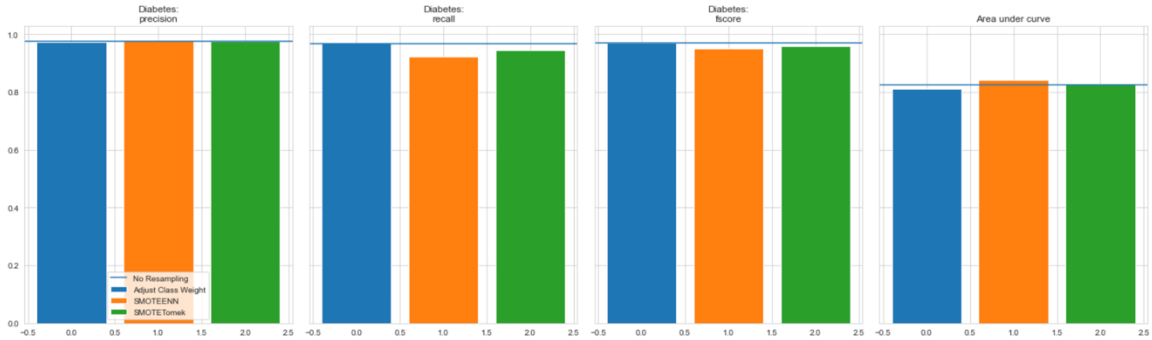


Figure 43: Decision Tree Algorithm - Sepsis Dataset - Hybrid Technique

In the cerebral stroke dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.97 and SMOTEENN with 0.96.

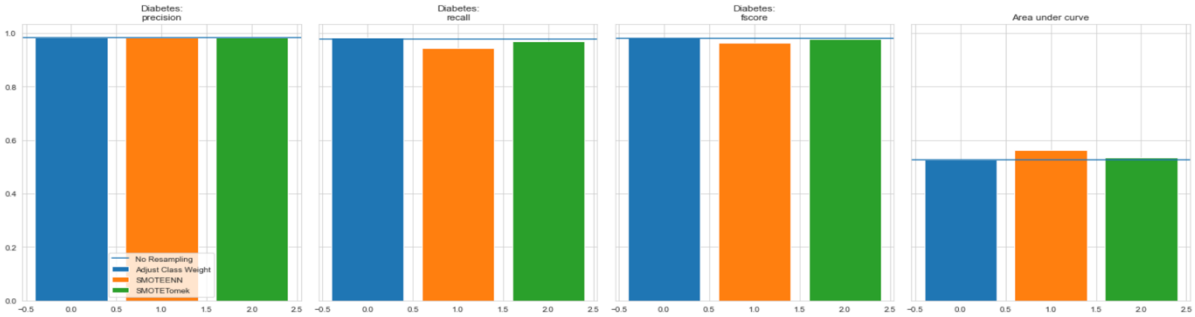


Figure 44: Decision Tree Algorithm - Cerebral Stroke Dataset - Hybrid Technique

Gaussian Naïve Bayes (GNB)

In the diabetes dataset for the hybrid techniques, the SMOTETomek and SMOTEENN scored 0.67.

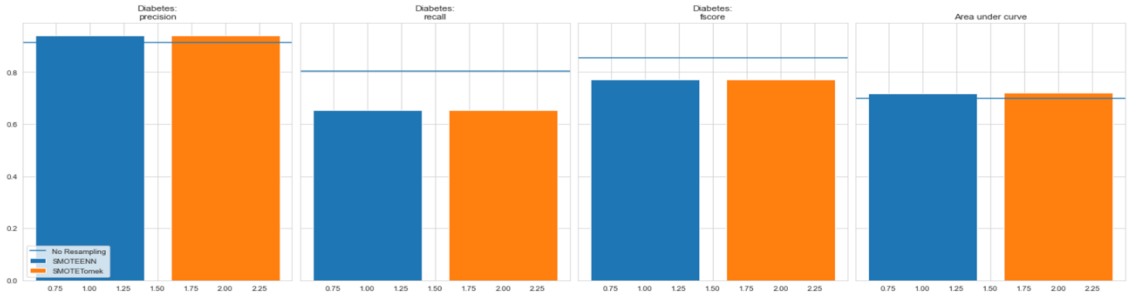


Figure 45: GNB Algorithm - Diabetes Dataset - Hybrid Technique

In the sepsis dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.92, and SMOTEENN with 0.91.

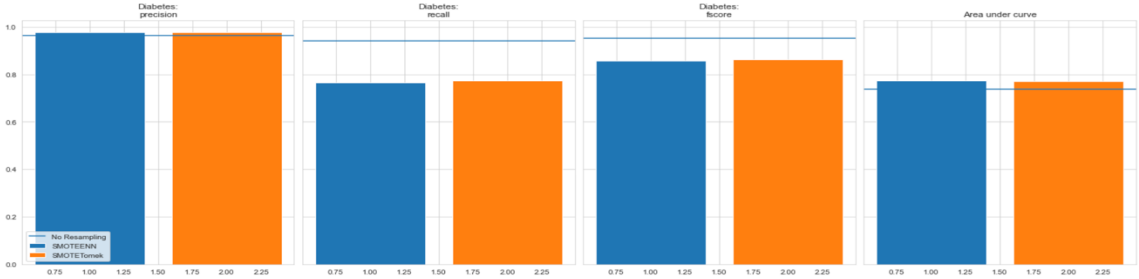


Figure 46: GNB Algorithm - Sepsis Dataset - Hybrid Technique

In the cerebral stroke dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.95 and SMOTEENN with 0.92.

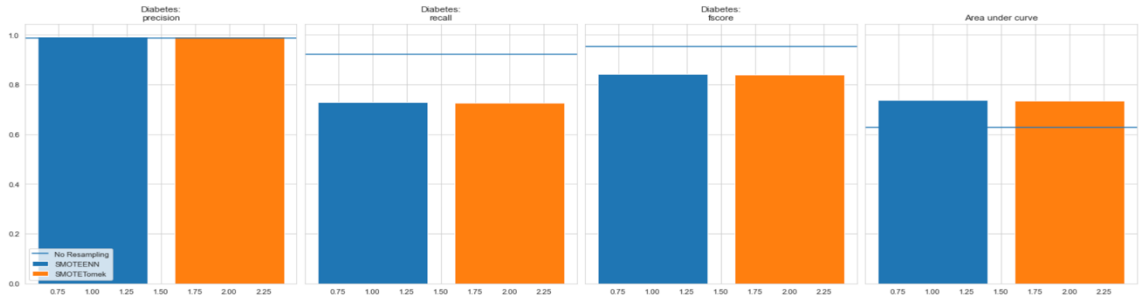


Figure 47: GNB Algorithm - Cerebral Stroke Dataset - Hybrid Technique

K-Nearest Neighbourhood (KNN)

In the diabetes dataset for the hybrid techniques, the SMOTETomek and scored the highest with 0.69. Followed by SMOTEENN with 0.64.

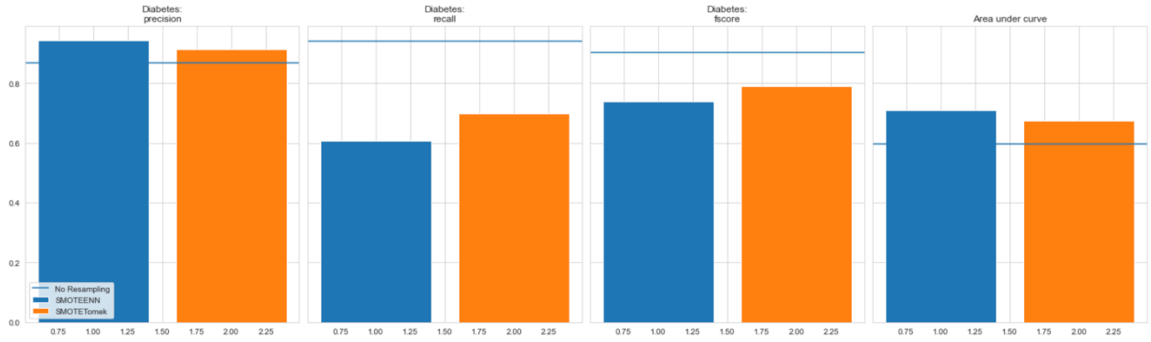


Figure 48: KNN Algorithm - Diabetes Dataset - Hybrid Technique

In the sepsis dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.81, and SMOTEENN with 0.75.

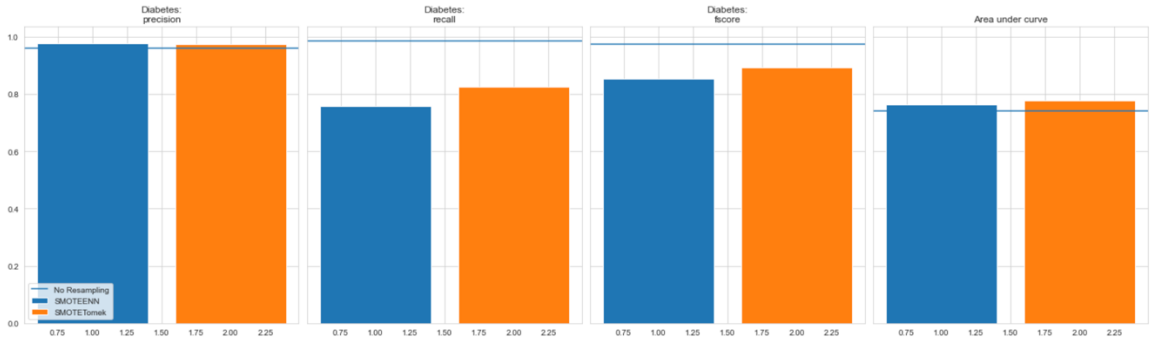


Figure 49: KNN Algorithm - Sepsis Dataset - Hybrid Technique

In the cerebral stroke dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.89 and SMOTEENN with 0.86.

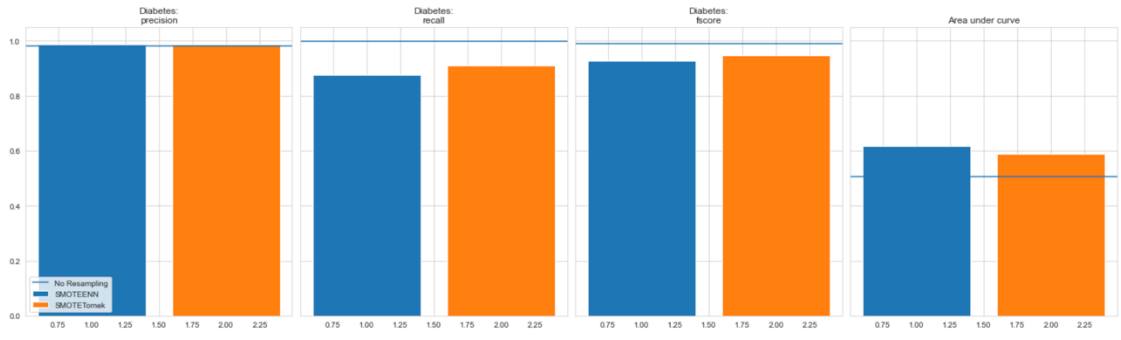


Figure 50: KNN Algorithm - Cerebral Stroke Dataset - Hybrid Technique

Logistic Regression (LR)

In the diabetes dataset for the hybrid techniques, the SMOTETomek and scored the highest with 0.77. Followed by SMOTEENN with 0.72.

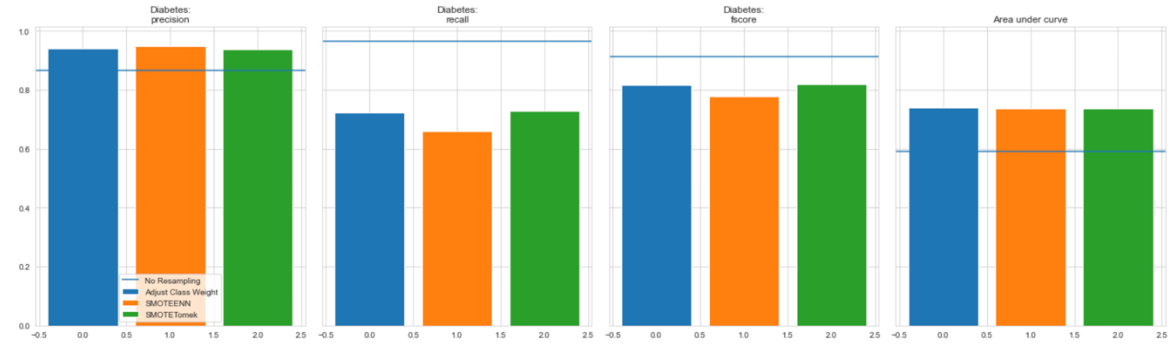


Figure 51: Logistic Regression Algorithm - Diabetes Dataset - Hybrid Technique

In the sepsis dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.92, and SMOTEENN with 0.91.

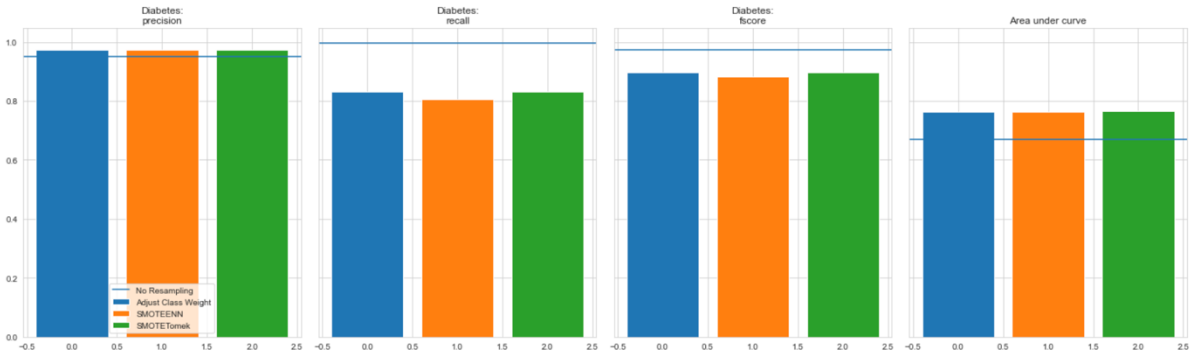


Figure 52: Logistic Regression Algorithm - Sepsis Dataset - Hybrid Technique

In the cerebral stroke dataset for the hybrid techniques, the SMOTETomek scored the highest with 0.95 and SMOTEENN with 0.92.

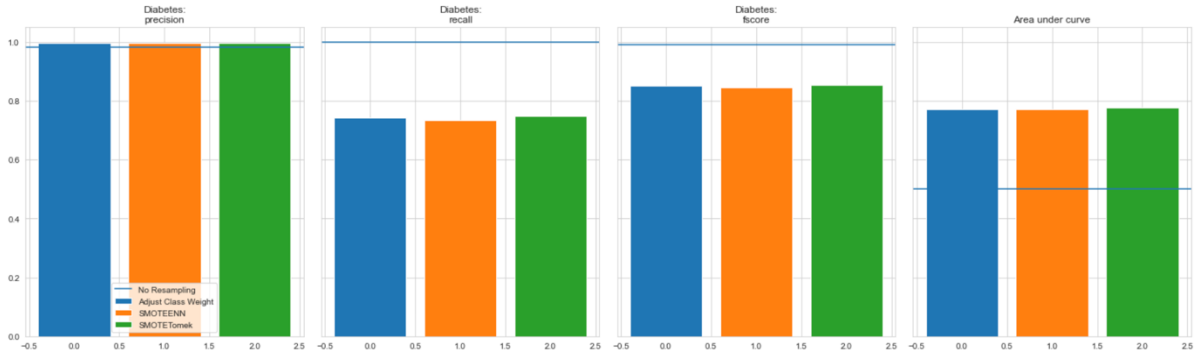


Figure 53: Logistic Regression Algorithm - Cerebral Stroke Dataset - Hybrid Technique

Discussion

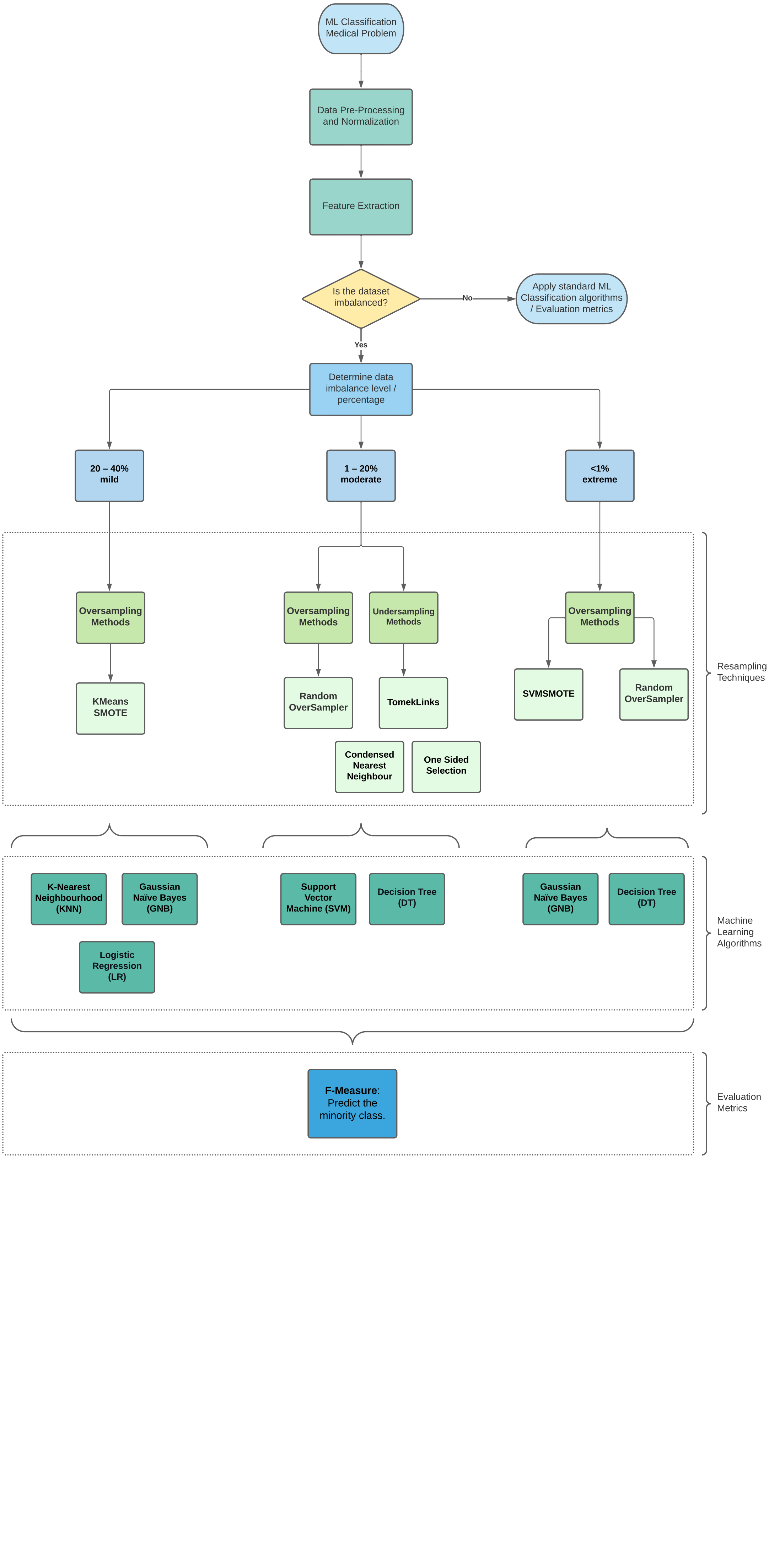
For the Diabetes dataset, categorised in the imbalance properties range as mild, the oversampling techniques seem to be the best option, as they got the highest scores from the tested selection of those types of sampling techniques. In the case of GNB, undersampling scores were around ~0.80, with the exceptions of Cluster Centroids and Near Miss. The best scores were from KMeansSMOTE, TomekLinks,

OneSidedSelection, and CondensedNearestNeighbour. On the other hand, ClusterCentroids, and NearMiss generally got the lowest score of the techniques. Finally, according to the results, KNN, GNB and Logistic Regression are the best algorithms for this type of data.

The Sepsis dataset categorised in the imbalance properties ranges as moderate. The undersampling and oversampling techniques seem to be the best option either way, as they closely followed each other in the results, with the exceptions of ClusterCentroids, and NearMiss generally got the lowest score of the techniques. The best scores were from RandomOverSampler, CondensedNearestNeighbour, TomekLinks, and OneSidedSelection, firmly right behind by AllKNN, EditedNearest Neighbours, RepeatedEditedNearestNeighbours, NeighbourhoodCleaningRule in the SVM and Decision Tree algorithms. In conclusion, according to the results, GNB and Logistic Regression are the best algorithms for this type of data.

The Stroke dataset categorised in the imbalance properties ranges as extreme. The oversampling techniques seem to be the best option getting consistently the highest scores. It's worth noticing that ClusterCentroids and NearMiss generally got the lowest score of the techniques. The best scores were from RandomOverSampler, SVMSMOTE, CondensedNearestNeighbour, TomekLinks, and OneSidedSelection, firmly right behind by AllKNN, EditedNearestNeighbours, RepeatedEditedNearest Neighbours, NeighbourhoodCleaningRule in the Logistic Regression and Decision Tree algorithms. Lastly, according to the results, GNB and Decision Tree are the best algorithms for this type of data.

In the hybrid sampling techniques, SMOTETomek consistently scored over SMOTEENN. The differences varied from 2 points to a maximum of 10. The Decision Tree and Logistic Regression algorithms performed the best. The diabetes datasets unfailingly kept getting the poorer scores, usually under 0.78, contra part of the sepsis and stroke datasets where the lowest was 0.75.


ML Classification Medical Problem
Data Pre-Processing and Normalization
Feature Extraction
Is the dataset imbalanced?
No
Apply standard ML Classification algorithms / Evaluation metrics
Yes
Determine data imbalance level / percentage
20 – 40% mild
1 – 20% moderate
<1% extreme
Oversampling Methods
KMeans SMOTE
Oversampling Methods
Random OverSampler
Undersampling Methods
TomekLinks
Condensed Nearest Neighbour
One Sided Selection
Oversampling Methods
SVMSMOTE
Random OverSampler
Resampling Techniques
K-Nearest Neighbourhood (KNN)
Gaussian Naïve Bayes (GNB)
Logistic Regression (LR)
Support Vector Machine (SVM)
Decision Tree (DT)
Gaussian Naïve Bayes (GNB)
Decision Tree (DT)
Machine Learning Algorithms
F-Measure: Predict the minority class.
Evaluation Metrics

ML Classification Medical Problem

Data Pre-Processing and Normalization

Feature Extraction

Is the dataset imbalanced?

No

Apply standard ML Classification algorithms / Evaluation metrics

Yes

Determine data imbalance level / percentage

20 – 40% mild

1 – 20% moderate

<1% extreme

Hybrid Techniques

SMOTE Tomek

Resampling Techniques

Logistic Regression (LR)

Decision Tree (DT)

Machine Learning Algorithms

F-Measure: Predict the minority class.

Evaluation Metrics

Conclusions

Limitations

After completing the analysis of the research findings in the previous section, we can define the potential study limitations described hereafter that describe the considered impact of research weakness thoroughly. The methodological limitations of the study included:

- Methods/instruments/techniques used to collect the data were not executed by the researcher. Instead, the data was acquired by web search and selected according to the field and features.
- Abundance of previous research studies on the topic made the citing and referencing extensive and challenging to define fundamental foundations for the research question investigation. Assumption based that the preliminary research information collected suffice and answered the research question appropriately.

On the other hand, the researcher's limitations involved:

- Limited access to data, we usually think that data is everywhere and anywhere, but open-source datasets are a real treasure that needs to be discovered. For reproducibility purposes, the data used in this research was required to be freely available for access, use, modification and sharing. However, these properties are not all present in the datasets. For example, some of the datasets are required to accredit the researchers or developers that initially created them, but in all cases are open to being utilized for commercial purposes.
- Limited access to computing resources, the models had to be run at different times due to the lack of a powerful enough system to provide results promptly. As a result, various environments that comprised google collab pro and Azure public cloud were tested, getting the same results as using the resources from local settings, with nothing that could affect the results.
- Time constraints, the execution of this research was done in parallel with other commitments that unfortunately superseded and had to be prioritized concerning extending the research and making it more comprehensive and extensive. It could also include another type of classification method or the ensemble techniques, for instance, that we dropped due to the additional work for proper implementation.

A proposed direction for future studies to overcome these limitations could include but is not limited to either creating the dataset or connecting directly with the dataset creator to include and describe the data collection process in the research. This option possibly extends the research period as this process is quite long and challenging but ensures more long-lasting results. Next, narrow down the research question to avoid overwhelming and wide-ranging results when investigating the topic. Followed by engaging with the research body or any institute that can provide powerful computing resources that help train models that use datasets of different sizes so that we can add more complexity to the research. Finally, define a better work execution schedule that can compellingly distribute the workload and help spread other lines of examination within the research.

Future Research & Recommendations

The original research proposal aimed to evaluate Machine Learning classification techniques for imbalanced class distributions in medical datasets while applying various resampling techniques to such datasets. As a result, it will allow us to create a reference guide to determine which algorithms and techniques are the most suitable depending on the level/percentage of imbalanced class distributions.

So, as part of the suggestion for future research and recommendations given as an outcome of this body of work, two types of proposals can be defined

Re-evaluate the framework; this research was tied to a specific context that comprises resampling techniques over medical datasets. In this case, we generate synthetic data that gets added to the dataset. Although these techniques, at their core, mirror the same statistical properties of the original dataset, this means that it does not necessarily uncover information about the patients that the dataset intended to describe and removes potential helpful information. Furthermore, outlining what entitles an imbalanced class or missing data could have countless sides in healthcare and medical datasets—for example, starting from the fact that an imbalance distribution could be due to the field being inherently imbalanced, as it does not have enough data to be collected, like cases of rare diseases. Moreover, the missing data, for instance, could be owed that the patient's doctor determined a laboratory study was not required/needed, which in the result seems to be an absent data point, but in reality, describes the real-world patient care journey.

As part of the recommendations, let's start with creating a built-for-purpose dataset and, in the case of rare diseases, gathering data intentionally and filling up every necessary study, beginning at admission and concluding with hospital discharge if it is the circumstance. Remember, if implementing that even though data comes from the medical setting, it is biased by default because it does not represent the patient's real experience throughout an episode of care but can help us answer some questions.

Expanding the framework, as aforementioned, this study consist of testing resampling strategies' behaviour about machine learning classification algorithms. Escalating the study could include the implementation of hyper-parameters, and tailored performance classification metrics. Furthermore, adding other classification models, like tree-based algorithms (Random Forest, XGBoost) and deep learning algorithms (Neural Networks, Convolutional Neural Networks). Implementing supplementary models translates to defining all possible solutions that complement the result in an extensive mode, which means that we confidently can describe all of the nuances of the problem. Additionally, multiclass/multilabel classification tasks are a new construct of the original relationship between the classification algorithm and the resampling technique. We move a binary class to multiple classes that need to be assigned to each sample, and to a data point not having a mutually-exclusive predicted target label. The final prospect of this body of work is to select one algorithm per task to test the behaviour of the resampling technique and determine the level of complexity that comprises such a process.

References

- Aada, A. and Tiwari, S., 2019. Predicting Diabetes in Medical Datasets Using Machine Learning Techniques. International Journal of Scientific Research & Engineering Trends, [online] 5(2), ISSN (Online): 2395-566X). Available at: <https://ijsret.com/wp-content/uploads/2019/03/IJSRET_V5_issue2_154.pdf> [Accessed 22 November 2021].
- Adane, K., Gizachew, M. and Kendie, S., 2019. The role of medical data in efficient patient care delivery: a review. Risk Management and Healthcare Policy, [online] Volume 12, pp.67-73. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6486797/> [Accessed 11 December 2021].
- Alam, T., Shaukat, K., Hameed, I., Luo, S., Sarwar, M., Shabbir, S., Li, J. and Khushi, M., 2020. An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. IEEE Access, [online] 8, pp.201173-201198. Available at: <https://ieeexplore.ieee.org/document/9239944> [Accessed 29 November 2021].
- Ali, A., Shamsuddin, S. and Ralescu, A., 2015. Classification with class imbalance problem: A review. Int. J. Advance Soft Compu. Appl, [online] 5(3), pp.176-204. Available at: <https://www.researchgate.net/publication/288228469_Classification_with_class_imbalance_problem_A_review> [Accessed 16 November 2021].
- Ardakani, M., Askarian, M., Shokry, A., Escudero, G., Graells, M. and Espuña, A., 2016. Optimal Features Selection for Designing a Fault Diagnosis System. Computer Aided Chemical Engineering, [online] 38, pp.1111-1116. Available at: <https://www.sciencedirect.com/science/article/abs/pii/B9780444634283501909> [Accessed 7 December 2021].
- Asiri, S., 2018. Machine Learning Classifiers. [online] Medium. Available at: <https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623> [Accessed 6 December 2021].
- Awoyemi, J., Adetunmbi, A. and Oluwadare, S., 2017. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI), [online] pp.1-9. Available at: <https://ieeexplore.ieee.org/abstract/document/8123782> [Accessed 15 November 2021].
- Babu, G. and Suresh, S., 2013. Meta-cognitive RBF Network and its Projection Based Learning algorithm for classification problems. Applied Soft Computing, [online] 13(1), pp.654-666. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S1568494612004206> [Accessed 6 December 2021].
- Badr, W., 2019. Having an Imbalanced Dataset? Here Is How You Can Fix It. [online] Medium. Available at: <https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb> [Accessed 10 November 2021].
- Berghout, T., Mouss, L., Bentrcia, T., Elbouchikhi, E. and Benbouzid, M., 2021. A deep supervised learning approach for condition-based maintenance of naval propulsion systems. Ocean Engineering, [online] 221. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0029801820314323> [Accessed 8 December 2021].
- Bhandari, P., 2021. Random vs systematic error. [online] Scribbr. Available at: <https://www.scribbr.com/methodology/random-vs-systematic-error/> [Accessed 16 November 2021].
- Bozkurt, E., 2021. Machine Learning Classification Algorithms with Codes. [online] Medium. Available at: <https://medium.com/analytics-vidhya/machine-learning-classification-algorithms-with-codes-5a8af4491fcb> [Accessed 8 December 2021].

- Brown, I. and Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, [online] 39(3), pp.3446-3453. Available at: <https://www.sciencedirect.com/science/article/pii/S095741741101342X> [Accessed 23 November 2021].
- Brownlee, J., 2019. A Gentle Introduction to Imbalanced Classification. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/what-is-imbalanced-classification/> [Accessed 15 November 2021].
- Brownlee, J., 2019. A Gentle Introduction to Statistical Sampling and Resampling. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/statistical-sampling-and-resampling/> [Accessed 23 November 2021].
- Brownlee, J., 2020. 4 Types of Classification Tasks in Machine Learning. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/ https://machinelearningmastery.com/types-of-classification-in-machine-learning/> [Accessed 10 November 2021].
- Brownlee, J., 2020. A Gentle Introduction to Imbalanced Classification. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/what-is-imbalanced-classification/> [Accessed 10 November 2021].
- Brownlee, J., 2020. Bagging and Random Forest for Imbalanced Classification. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/> [Accessed 10 November 2021].
- Cao, P., Liu, X., Zhang, J., Zhao, D., Huang, M. and Zaiane, O., 2017. ℓ2,1 norm regularized multi-kernel based joint nonlinear feature selection and over-sampling for imbalanced data classification. Neurocomputing, [online] 234, pp.38-57. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S092523121631551X> [Accessed 22 November 2021].
- Cios, K. and William Moore, G., 2002. Uniqueness of medical data mining. Artificial Intelligence in Medicine, [online] 26(1-2), pp.1-24. Available at: <https://www.sciencedirect.com/science/article/pii/S0933365702000490> [Accessed 2 December 2021].
- CROS - European Commission. 2019. Measurement error - CROS - European Commission. [online] Available at: <https://ec.europa.eu/eurostat/cros/content/measurement-error_en> [Accessed 16 November 2021].
- Dada, E., Bassi, J., Chiroma, H., Abdulhamid, S., Adetunmbi, A. and Ajibuwa, O., 2019. Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, [online] 5(6), p.e01802. Available at: <https://www.sciencedirect.com/science/article/pii/S2405844018353404> [Accessed 15 November 2021].
- Dembla, G., 2020. Intuition behind Log-loss Score. [online] Medium. Available at: <https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a> [Accessed 10 December 2021].
- Desuky, A. and Hussain, S., 2021. An Improved Hybrid Approach for Handling Class Imbalance Problem. Arabian Journal for Science and Engineering, [online] 46(4), pp.3853-3864. Available at: <https://link.springer.com/article/10.1007/s13369-021-05347-7#citeas> [Accessed 16 November 2021].
- Emanet, N., Öz, H., Bayram, N. and Delen, D., 2014. A comparative analysis of machine learning methods for classification type decision problems in healthcare. Decision Analytics, [online] 1(1). Available at: <https://decisionanalyticsjournal.springeropen.com/articles/10.1186/2193-8636-1-6#citeas> [Accessed 3 December 2021].
- Ferreiro Volpi, G., 2019. Class Imbalance: a classification headache. [online] Medium. Available at: <https://towardsdatascience.com/class-imbalance-a-classification-headache-1939297ff4a4> [Accessed 15 November 2021].

- Ganganwar, V., 2012. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, [online] 2(4). Available at: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.413.3344&rep=rep1&type=pdf> [Accessed 15 November 2021].
- Garg, R., 2018. 7 Types of Classification Algorithms. [online] Analytics India Magazine. Available at: <https://analyticsindiamag.com/7-types-classification-algorithms/> [Accessed 7 December 2021].
- Ghorbani, R. and Ghousi, R., 2021. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. IEEE Access, [online] 8, pp.67899 - 67911. Available at: <https://ieeexplore.ieee.org/abstract/document/9062549> [Accessed 23 November 2021].
- Goldstein-Greenwood, J., 2021. A Brief on Brier Scores. [online] Data.library.virginia.edu. Available at: <https://data.library.virginia.edu/a-brief-on-brier-scores/> [Accessed 10 December 2021].
- Google Developers. 2021. Imbalanced Data. [online] Available at: <https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data> [Accessed 2 December 2021].
- Guo, X., Yin, Y., Dong, C., Yang, G. and Zhou, G., 2008. On the Class Imbalance Problem. 2008 Fourth International Conference on Natural Computation, [online] Available at: <https://sci2s.ugr.es/keel/pdf/specific/congreso/guo_on_2008.pdf> [Accessed 28 November 2021].
- He, H., Bai, Y., Garcia, E. and Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), [online] pp.1322-1328. Available at: <https://ieeexplore.ieee.org/abstract/document/4633969> [Accessed 22 November 2021].
- Japkowicz, N. and Stephen, S., 2002. The class imbalance problem: A systematic study1. Intelligent Data Analysis, [online] 6(5), pp.429-449. Available at: <https://content.iospress.com/articles/intelligent-data-analysis/ida00103> [Accessed 16 November 2021].
- Johnson, J. and Khoshgoftaar, T., 2019. Survey on deep learning with class imbalance. Journal of Big Data, [online] 6(1). Available at: <https://www.researchgate.net/publication/332165523_Survey_on_deep_learning_with_class_imbalance> [Accessed 4 December 2021].
- Kashyap, J. and Gulati, P., 2020. Hybrid Resampling Technique to Tackle the Imbalanced Classification Problem. J.C. Bose University of Science and Technology, YMCA, [online] Available at: <https://assets.researchsquare.com/files/rs-36578/v1_covered.pdf?c=1631836348> [Accessed 22 November 2021].
- Khaldy, M. and Kambhampati, C., 2018. Resampling Imbalanced Class and the Effectiveness of Feature Selection Methods for Heart Failure Dataset. International Robotics & Automation Journal, [online] 4(1), pp.37-45. Available at: <https://medcraveonline.com/IRATJ/resampling-imbalanced-class-and-the-effectiveness-of-feature-selection-methods-for-heart-failure-dataset.html#ref3> [Accessed 22 November 2021].
- Khushi, M., Shaukat, K., Alam, T., Hameed, I., Uddin, S., Luo, S., Yang, X. and Reyes, M., 2021. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access, [online] 9, pp.109960-109975. Available at: <https://ieeexplore.ieee.org/abstract/document/9505667> [Accessed 24 November 2021].
- Krawczyk, B., Woźniak, M. and Schaefer, G., 2014. Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Computing, [online] 14 C, pp.554-562. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S1568494613002895?via%3Dihub> [Accessed 25 November 2021].

- Kumar, P., Bhatnagar, R., Gaur, K. and Bhatnagar, A., 2021. Classification of Imbalanced Data: Review of Methods and Applications. IOP Conference Series: Materials Science and Engineering, [online] 1099(1), p.012077. Available at: <https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012077/pdf> [Accessed 4 December 2021].
- Lemaitre, G., Nogueira, F. and Aridas, C., 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. [online] arXiv.org. Available at: <http://arxiv.org/abs/1609.06570> [Accessed 16 November 2021].
- Legay A., Sedwards S., Traonouez LM. (2016) Rare Events for Statistical Model Checking an Overview. In: Larsen K., Potapov I., Srba J. (eds) Reachability Problems. RP 2016. Lecture Notes in Computer Science, vol 9899. Springer, Cham. https://doi.org/10.1007/978-3-319-45994-3_2
- Li, D., Liu, C. and Hu, S., 2010. A learning method for the class imbalance problem with medical data sets. Computers in Biology and Medicine, [online] 40(5), pp.509-518. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0010482510000405> [Accessed 3 December 2021].
- Liu, J., 2021. Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data. Soft Computing, [online] Available at: <https://www.researchgate.net/publication/356424067_Importance-SMOTE_a_synthetic_minority_oversampling_method_for_noisy_imbalanced_data> [Accessed 29 November 2021].
- Liu, N., Li, X., Qi, E., Xu, M., Li, L. and Gao, B., 2020. A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data. IEEE Access, [online] 8, pp.171263-171280. Available at: <https://ieeexplore.ieee.org/document/9159642> [Accessed 15 November 2021].
- Liu, X. and Zhou, Z., 2013. Ensemble Methods for Class Imbalance Learning. Imbalanced Learning, [online] (1), pp.61-82. Available at: <https://ieeexplore.ieee.org/document/6542372> [Accessed 25 November 2021].
- Liu, X., Wu, J. and Zhou, Z., 2006. Exploratory Under-Sampling for Class-Imbalance Learning. *Sixth International Conference on Data Mining (ICDM'06)*, [online] Available at: <https://www.researchgate.net/publication/220766726_Exploratory_Under-Sampling_for_Class-Imbalance_Learning> [Accessed 10 December 2021].
- Livieris, I., Drakopoulou, K., Tampakas, V., Mikropoulos, T. and Pintelas, P., 2018. Predicting Secondary School Students' Performance Utilizing a Semi-supervised Learning Approach. Journal of Educational Computing Research, [online] 57(2), pp.448-470. Available at: <https://journals.sagepub.com/doi/abs/10.1177/0735633117752614> [Accessed 8 December 2021].
- Longadge, R. and Dongre, S., 2013. Class Imbalance Problem in Data Mining Review. [online] arXiv.org. Available at: <https://arxiv.org/abs/1305.1707> [Accessed 24 November 2021].
- Lujic, S., Watson, D., Randall, D., Simpson, J. and Jorm, L., 2014. Variation in the recording of common health conditions in routine hospital data: study using linked survey and administrative data in New South Wales, Australia. BMJ Open, [online] 4(9), pp.e005768-e005768. Available at: <https://pubmed.ncbi.nlm.nih.gov/25186157/> [Accessed 3 December 2021].
- Luque, A., Carrasco, A., Martín, A. and de las Heras, A., 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, [online] 91, pp.216-231. Available at: <https://www.sciencedirect.com/science/article/pii/S0031320319300950> [Accessed 15 November 2021].
- Maalouf, M. and Trafalis, T., 2011. Robust weighted kernel logistic regression in imbalanced and rare events data. Computational Statistics & Data Analysis, [online] 55(1), pp.168-183. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0167947310002598> [Accessed 3 December 2021].

- Madhukar, B., 2020. Using Near-Miss Algorithm For Imbalanced Datasets. [online] Analytics India Magazine. Available at: <https://analyticsindiamag.com/using-near-miss-algorithm-for-imbalanced-datasets/> [Accessed 28 November 2021].
- Mohamed, A., 2017. Comparative Study of Four Supervised Machine Learning Techniques for Classification. International Journal of Applied Science and Technology, [online] 7(2). Available at: <https://www.researchgate.net/publication/319313534_Comparative_Study_of_Four_Supervised_Machine_Learning_Techniques_for_Classification> [Accessed 7 December 2021].
- Montero, M., 2021. Resampling Methods for Machine Learning modeling. [online] Medium. Available at: <https://medium.com/geekculture/resampling-methods-for-machine-learning-modeling-d2cdc1d3640f> [Accessed 23 November 2021].
- Nabi, J., 2018. Machine Learning—Multiclass Classification with Imbalanced Data-set. [online] Medium. Available at: <https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a> [Accessed 7 December 2021].
- Nighania, K., 2018. Various ways to evaluate a machine learning models performance. [online] Medium. Available at: <https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15> [Accessed 10 December 2021].
- Novaković, J., Veljović, A., Ilić, S., Papić, Ž. and Milica, T., 2017. Evaluation of Classification Models in Machine Learning. Theory and Applications of Mathematics & Computer Science, [online] 7(1), pp.39-46. Available at: <https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158/126> [Accessed 6 December 2021].
- Omar, S., Ngadi, A. and H. Jebur, H., 2013. Machine Learning Techniques for Anomaly Detection: An Overview. International Journal of Computer Applications, [online] 79(2), pp.33-41. Available at: <https://www.researchgate.net/profile/Salima-Benqdara/publication/325049804_Machine_Learning_Techniques_for_Anomaly_Detection_An_Overview/links/5af3569b4585157136c919d8/Machine-Learning-Techniques-for-Anomaly-Detection-An-Overview.pdf> [Accessed 6 December 2021].
- Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J. O, Olakanmi, O., and Akinjobi J., 2017. Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology, [online] 48(3), pp.128-138. Available at: <https://www.researchgate.net/profile/J-E-T-Akinsola/publication/318338750_Supervised_Machine_Learning_Algorithms_Classification_and_Comparison/links/596481dd0f7e9b819497e265/Supervised-Machine-Learning-Algorithms-Classification-and-Comparison.pdf> [Accessed 6 December 2021].
- Peikari, M., Salama, S., Nofech-Mozes, S. and Martel, A., 2018. A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification. Scientific Reports, [online] 8(1). Available at: <https://www.nature.com/articles/s41598-018-24876-0#citeas> [Accessed 8 December 2021].
- Rácz, A., Bajusz, D. and Héberger, K., 2019. Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics. Molecules, [online] 24(15), p.2811. Available at: <https://www.mdpi.com/1420-3049/24/15/2811> [Accessed 10 December 2021].
- Rahman, M. and Davis, D., 2013. Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing, [online] 3(2), pp.224-228. Available at: <http://www.ijmlc.org/index.php?m=content&c=index&a=show&catid=36&id=306> [Accessed 23 November 2021].
- Razzaghi, T., Roderick, O., Safro, I. and Marko, N., 2015. Fast imbalanced classification of healthcare data with missing values. 18th International Conference on Information Fusion, [online] pp.774-781. Available at: <https://ieeexplore.ieee.org/abstract/document/7266639/authors#authors> [Accessed 4 December 2021].

- Rupak Roy - II, B., 2021. Condensed Nearest Neighbor Rule Undersampling (CNN). [online] Medium. Available at: < https://bobrupakroy.medium.com/condensed-nearest-neighbor-rule-undersampling-cnn-380c0d84ca88> [Accessed 28 November 2021].
- Rupak Roy - II, B., 2021. Edited Nearest Neighbors ENN. [online] Medium. Available at: <https://bobrupakroy.medium.com/ edited-nearest-neighbors-enn-c446a15e4bbe> [Accessed 28 November 2021].
- Rupak Roy - II, B., 2021. OSS & NCR. [online] Medium. Available at: <https://bobrupakroy.medium.com/oss-ncr-c16f69627715> [Accessed 28 November 2021].
- Sadawi, N., 2021. Advanced Machine Learning: How to Effectively Work with Imbalanced Data. [online] Oreilly.com. Available at: <https://www.oreilly.com/live-events/advanced-machine-learning-how-to-effectively-work-with-imbalanced-data/0636920388593/0636920062047/> [Accessed 16 November 2021].
- Safran, C., Bloomrosen, M., Hammond, W., Labkoff, S., Markel-Fox, S., Tang, P. and Detmer, D., 2007. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association, [online] 14(1), pp.1-9. Available at: <https://link.springer.com/article/10.1186%2F2047-2501-2-3> [Accessed 3 December 2021].
- Santiso, S., Casillas, A. and Pérez, A., 2018. The class imbalance problem detecting adverse drug reactions in electronic health records. Health Informatics Journal, 25(4), pp.1768-1778. Available at: < https://pubmed.ncbi.nlm.nih.gov/30230408/> [Accessed 6 December 2021].
- Sasada, T., Liu, Z., Baba, T., Hatano, K. and Kimura, Y., 2020. A Resampling Method for Imbalanced Datasets Considering Noise and Overlap. Procedia Computer Science, [online] 176, pp.420-429. Available at: <https://www.researchgate.net/publication/346066569_A_Resampling_Method_for_Imbalanced_Datasets_Considering_Noise_and_Overlap> [Accessed 29 November 2021].
- Schildcrout, J., Rathouz, P., Zelnick, L., Garbett, S. and Heagerty, P., 2015. Biased sampling designs to improve research efficiency: Factors influencing pulmonary function over time in children with asthma. The Annals of Applied Statistics, [online] 9(2). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4551501/> [Accessed 17 November 2021].
- Seiffert, C., Khoshgoftaar, T., Van Hulse, J. and Napolitano, A., 2008. RUSBoost: Improving classification performance when training data is skewed. 19th International Conference on Pattern Recognition, [online] Available at: <https://www.researchgate.net/publication/220928945_RUSBoost_Improving_Classification_Performance_when_Training_Data_is_Skewed> [Accessed 29 November 2021].
- Smith, M., Martinez, T. and Giraud-Carrier, C., 2013. An instance level analysis of data complexity. Machine Learning, [online] 95(2), pp.225-256. Available at: <https://link.springer.com/article/10.1007%2Fs10994-013-5422-z> [Accessed 28 November 2021].
- Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications (0975 – 8887), [online] 17(8). Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.3899&rep=rep1&type=pdf> [Accessed 2 December 2021].
- Sunasra, M., 2017. Performance Metrics for Classification problems in Machine Learning. [online] Medium. Available at: <https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b> [Accessed 10 December 2021].
- Sundararaman, K., 2021. Addressing the Data Imbalance Problem in Healthcare - Ideas2IT. [online] Ideas2IT. Available at: <https://www.ideas2it.com/blogs/addressing-the-data-imbalance-problem-in-healthcare/> [Accessed 3 December 2021].

- Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M., Davatgar, N., Toomanian, N. and Scholten, T., 2020. Synthetic resampling strategies and machine learning for digital soil mapping in Iran. European Journal of Soil Science, [online] 71(3), pp.352-368. Available at: <https://onlinelibrary.wiley.com/doi/full/10.1111/ejss.12893> [Accessed 22 November 2021].
- Tarle, B., 2016. Medical data classification using different optimization techniques: a survey. International Journal of Research in Engineering and Technology, [online] 05(17), pp.101-108. Available at: <https://www.researchgate.net/publication/323027991_medical_data_classification_using_different_optimization_techniques_a_survey>[Accessed 2 December 2021].
- Telenti, A. and Jiang, X., 2020. Treating medical data as a durable asset. Nature Genetics, [online] 52(10), pp.1005-1010. Available at: <https://www.nature.com/articles/s41588-020-0698-y> [Accessed 2 December 2021].
- Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A., 2020. Data imbalance in classification: Experimental evaluation. Information Sciences, [online] 513, pp.429-441. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0020025519310497> [Accessed 16 November 2021].
- Tharwat, A., 2020. Classification assessment methods. Applied Computing and Informatics, [online] 17(1), pp.168-192. Available at: <https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.003/full/html> [Accessed 10 December 2021].
- The BMJ: leading general medical journal. Research. Education. 2021. Chapter 4. Measurement error and bias. [online] Available at: <https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/4-measurement-error-and-bias> [Accessed 16 November 2021].
- Trochim, W., 2021. Measurement Error. [online] Conjointly.com. Available at: <https://conjointly.com/kb/measurement-error/> [Accessed 16 November 2021].
- Wang, Q., 2014. A Hybrid Sampling SVM Approach to Imbalanced Data Classification. Abstract and Applied Analysis, [online] 2014, pp.1-7. Available at: <https://www.hindawi.com/journals/aaa/2014/972786/> [Accessed 23 November 2021].
- Wang, S., Minku, L. and Yao, X., 2015. Resampling-Based Ensemble Methods for Online Class Imbalance Learning. IEEE Transactions on Knowledge and Data Engineering, [online] 27(5), pp.1356-1368. Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6871400> [Accessed 24 November 2021].
- Yang, P., Cao, Y., Liu, D., Bai, Y., Pan, F. and Xu, Y., 2014. The effect of electronic medical record application on the length of stay in a Chinese general hospital: a department- and disease-focused interrupted time-series study. Journal of Medical Systems, [online] 38(5). Available at: <https://pubmed.ncbi.nlm.nih.gov/24760225/> [Accessed 3 December 2021].
- Zekić-Sušac, M., Pfeifer, S. and Šarlija, N., 2014. A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem. Business Systems Research Journal, [online] 5(3), pp.82-96. Available at: <https://hrcak.srce.hr/ojs/index.php/bsr/article/view/12540/6157> [Accessed 8 December 2021].
- Zhang, Y., Zhang, L. and Wang, Y., 2010. Cluster-based majority under-sampling approaches for class imbalance learning. 2010 2nd IEEE International Conference on Information and Financial Engineering, [online] pp.400-404. Available at: <https://ieeexplore.ieee.org/abstract/document/5609385> [Accessed 3 December 2021].
- Zhao, Y., Wong, Z. and Tsui, K., 2018. A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection. Journal of Healthcare Engineering, [online] 2018(2018), pp.1-11. Available at: <https://www.hindawi.com/journals/jhe/2018/6275435/> [Accessed 23 November 2021].
- Zhou, P. and Wong, A., 2021. Explanation and prediction of clinical data with imbalanced class distribution based on pattern discovery and disentanglement. BMC Medical Informatics and

Decision Making, [online] 21(1). Available at: <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01356-y> [Accessed 3 December 2021].

- Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J. and Ning, G., 2018. Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data. IEEE Access, [online] 6, pp.4641-4652. Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8246503> [Accessed 3 December 2021].