DURBAN UNIVERSITY OF TECHNOLOGY

Machine Learning: A Data-Point Approach to Solving Misclassifications in the Imbalanced Credit Card Datasets

By

Nhlakanipho Michael Mqadi

21349052

A dissertation submitted in fulfilment of the requirement for the degree of Master of Information and Communications Technology

Faculty of Accounting and Informatics,

Department of Information Technology,

Postgraduate Studies

Supervisor: Dr N. Naicker

Co-Supervisor: Dr T. Adeliyi

2021

DECLARATION

I, Nhlakanipho Michael Mqadi, declare that:

(i) The research reported in this dissertation, except where otherwise indicated, and is my original research.

(ii) This dissertation has not been submitted for any degree or examination at any other university.

(iii) This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

- Their words have been re-written but the general information attributed to them has been referenced.
- Where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the Reference Section of this dissertation.

Signature: *Nhlakanipho Mqadi*

Date: 23 May 2021

Approvea: Supervisor
30 October 2021

ACKNOWLEDGEMENTS

Kind acknowledgement goes to the Durban University of Technology for making funding opportunities and materials for experiments available for this research project.

A special acknowledgement goes to Dr N. Naicker for the supervision and to Dr T. Adeliyi for the co-supervision and contributions towards the success of this work.

A special thanks to my mother Nomusa Mqadi for all the support, and the financial contribution towards the successful completion of my studies.

A special thanks to Delisile Khwela for all the moral and emotional support, and always providing me with encouragement that has contributed towards the successful completion of my studies.

A special thanks Dr Michael Olusanya for providing guidance, words of wisdom every time we interacted, and the willingness to always contribute towards the successful completion of my research.

A special thanks to a friend and mentor Logan Govender for offering to be my sponsor and assisting me financially to ensure that I successfully complete of my studies.

"There is no passion to be found playing small – in settling for a life that is less than the one you are capable of living." – Nelson Mandela

ABSTRACT

Machine learning (ML) uses algorithms with the complexity to iterate over massive datasets to analyse the data for past behaviour with the aim to predict future outcomes. Financial institutions are using ML to detect Credit Card Fraud (CCF) by learning the patterns that distinguish between legitimate and fraudulent actions from historic data of credit card transactions to combat CCF. The market economic order has been negatively affected by CCF, which has contributed to low consumer confidence in financial institutions, and loss of interest from investors. The CCF loses continue increasing every year despite existing efforts to prevent fraud, which amount to billions of dollars lost annually. ML techniques consume large volumes of historical credit card transaction data as examples for learning. In ordinary credit card datasets, there are far fewer fraudulent transactions than legitimate transactions. In dealing with the credit card data imbalance problem, the ideal solution must have low bias, low variance, and high accuracy. The aim of this study was to provide an in-depth experimental investigation of the effect of using the data-point approach to resolve the class misclassification problem in imbalanced credit card datasets.-The study focused on finding a novel way to handle imbalanced data, to improve the performance of ML algorithms in identifying fraud or anomaly patterns in massive amounts of financial transaction records, where the class distribution was imbalanced. The experiment led to the introduction of two unique multi-level hybrid data-point approach solutions, namely, Feature Selection with Near Miss Undersampling; and Feature Selection with SMOTe based Oversampling. The results were obtained using four widely used ML algorithms, namely, Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression to build the classifiers. These algorithms were implemented for classification of credit card datasets and the performance was assessed using selected performance metrics. The findings show that using the data-point approach improved the predictive accuracy of the ML fraud detection solution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

The following are the most important acronyms in the study:

ACID – Association Correction for Imbalanced Data

ANN – Artificial Neural Network

AUC – Area-Under the ROC Curve

BRNN – Bidirectional Recurrent Neural Network

CBCE – Class-Based ensemble for Class Evolution

CCRE – Clustering, Classification, and Range Estimation

CGAN-MBL – Conditional Generative Adversarial Network-Mean-Covariance Balancing Labelling

DBE – Distance-based Balancing Ensemble

DCR – Distance-based Combination Rule

DGC – Data Gravitation Classification

DNN – Deep Neural Network

DT – Decision Tree

FN – False Negative

FP – False Positive

GWO – Grey Wolf Optimizer

IGT – Information Gain Technique

LFR – Linear Four Rates

LR – Logistic Regression

LSH – Locality Sensitive Hashing

MBL – Mean-Covariance Balancing Labelling

ML – Machine Learning

MSE – Mean Squared Error

MTS – Mahalanobis-Taguchi System

OOB – Oversampling-based Online Bagging

RF – Random Forest

ROS – Random Oversampling

RUS – Random Undersampling

SBE – Splitting Balancing Ensemble

SDS – Selective Down-Sampling

SMOTe – Synthetic Minority Oversampling Technique

SOS – Spatio-temporal Oversampling

SVM – Support Vector Machine

T-Link – Tomek Link

TN – True Negative

TP – True Positive

TPC – The Pearson Correlation

UCI – University of California Irvine

UOB – Undersampling-based Online Bagging

VR – Variance Ranking

LIST OF TERMINOLOGY

The following terminology as related to the study are explained:

Accuracy – The percentage of correctly classified instances in a dataset.

Card Fraud – Unauthorised usage of another individual's credit or debit card without the issuer and card owner's consent.

Data Imbalance – A situation whereby the number of legitimate transactions recorded highly outweighs the fraudulent transactions in a dataset.

Data-Point Approach – An approach that consists of techniques for re-sampling the data in order to deal with imbalanced classes.

Dataset – The collection of instances of data of the same type.

Feature Selection – The process where the features that contribute most to the prediction variable or output of interest are automatically or manually selected.

Machine Learning – The science of designing and applying algorithms that are able to learn patterns from past cases.

Misclassification – The incorrect classification of classes in a dataset.

ML Classifiers – Logic models that are trained by continually feeding input data and assessing the performance.

Overfitting – A modelling error that occurs when a model learns the detail and noise in the training data to the extent that it negatively influences the performance of the model on new data.

Oversampling – Refers to increasing the count of the minorities to balance with the majority class.

Predictive Accuracy – The ability to classify legitimate and fraudulent transactions successfully.

Testing dataset – A subset dataset used to validate the model built.

Training dataset – A subset dataset used to build up a model.

Undersampling – Refers to a group of techniques designed to balance the class distribution for a classification dataset that has a skewed class distribution.

CHAPTER ONE: INTRODUCTION TO STUDY

1.1 Introduction

Machine Learning (ML) is the science of designing and building algorithms that use past cases to learn and identify existing patterns (Wei *et al.* 2020). According to Wang and Japkowicz (2010), the ML classifiers are logic models that are trained by continually feeding input data and assessing their performances. A ML classification solution uses algorithms with the complexity to iterate over massive datasets to analyse the data for existing patterns (Saad, Omar, and Maghraby 2019). One of the features of ML systems is the capability to recognise and categorize existing classes in the data (Fang *et al.* 2019). The instances of related data stored together is known as a dataset, and a ML solution will use these datasets for training and testing a model (Kotsiantis, Kanellopoulos, and Pintelas 2006). The dataset used to build and train a model is known as a training dataset.

The unseen instances of data used to validate a model is referred to as a testing dataset. A dataset can be divided into training and testing subset, where the data instances found in the training set are omitted from the testing dataset (Ge, Yue, and Chen 2017). In ML, the aim is to create models that are able to accurately identify and classify patterns found on the testing dataset based on pattern shown on the training dataset. Therefore, the training dataset is used for teaching a model and the testing dataset is used to evaluate it (Hukerikar *et al.* 2011). According to Galar *et al.* (2012), the models generated are to predict the hidden results, namely, the test data. Financial institutions are using ML to detect Credit Card Fraud (CCF) by learning patterns that distinguish between legitimate and fraudulent actions from historic data of credit card transactions (Adewumi and Akinyelu 2017). However, the performances of most existing fraud detection methods are still less than ideal in actual situations.

In the real world, the number of fraudulent transactions recorded account for a very small percentage of the total transactions (Batuwita and Palade 2010). This data distribution inequality is formally referred to as data imbalance. Data imbalance is a ML learning problem that causes most fraud detecting methods to be unable to achieve ideal fraud detection capabilities (Bian *et al.* 2016). In most credit card datasets, there are more legitimate over fraudulent transactions, as a result, CCF detection continues to be a significant challenge facing business intelligence technologies (Wasikowski and Chen 2010). Card fraud in the form of a credit or debit card transaction is the unauthorised usage of another individual's credit or debit card without the consent of the issuer and card owner (Kumar *et al.* 2019). Card fraud occurs in different forms, namely, lost card fraud, Card-Not-Present (CNP) fraud, Not-Received-Issued-Card (NRIC) fraud, and stolen card fraud. Lost card fraud describes the fraud that occurs when the legal card owner loses a legitimately issued card, which then gets used to transact fraudulently. CNP fraud is the leading contributor to gross card fraud losses, which occurs when fraudster transacts fraudulently using the card details of another individual's card, without the consent of the owner and while the physical card is in the owner's possession.

NRIC fraud can be described as the phenomenon that occurs when legitimately issued cards are intercepted by impostors before reaching the intended customer and are used fraudulently (South African Banking Risk Information Centre [SABRIC] 2019). Stolen card fraud is similar to lost card fraud but instead of the owner losing the card, a fraudster would intentionally remove a legitimately issued card from the owner's possession, to transact fraudulently. In fraudulent credit card transactions, the data that is collected and stored by financial issuers is very small compared to legitimate transactions, which results in a highly imbalanced dataset (Shen, Tong, and Deng 2007). The imbalance problems cause the ML classification solutions to be biased towards legitimate transactions and to have a high misclassification rate (Williams, Myers, and Silvious 2009).

1.2 Statement of the Problem

The market economic order has been negatively affected by CCF, which has contributed to low consumer confidence in financial institutions, and loss of interest from investors. According to Xu-Yin, Jianxin, and Zhi-Hua (2009), the risk of financial losses due to fraudulent activities can be mitigated by using ML to improve fraud detection systems. ML measures accuracy by analysing the rate of True Negatives (TN), True Positives (TP), False Positives (FP), and False Negatives (FN). According to Mathew *et al.* (2018), accuracy is described as the percentage of instances correctly classified as follows, (TP + TN) / (TP + TN + FP + FN). This study introduces the term Predictive Accuracy (PA), which defines the ability to successfully flag fraudulent and legitimate transactions.

The South African banking industry's annual crime statistics for 2019 released by SABRIC, revealed that online banking fraud incidents increased by 20% between 2018 and 2019

(SABRIC 2020). Card fraud crimes have had substantial negative economic and social effects worldwide. In 2015, card fraud losses approximately reached US$21.84 billion worldwide (Robertson 2016). The latest statistics showed that card fraud losses continued rising every year as they reached US$28.65 billion in 2019 worldwide, and they are projected to exceed US$32 billion by the end of 2021 (Nilson Report 2020).

In ML, the ideal solution must have low bias and should accurately model the true relationship of positive and negative classes (Ghorbani and Ghousi 2020). Most ML classifiers tend to perform well with a dataset that has a balanced class distribution (Wang and Yao 2013). The use of one dataset tends to produce a biased conclusion, as the data is often pre-processed to support the classification solution, which prevents the conclusion from being generalised (Jia, Li, and Zhang 2020). The ideal binary classification solutions should have low variability, by producing consistent predictions across different datasets (de la Fuente, Urrutia, and Chávez 2019). The inability to handle imbalanced data negatively affect the integrity and predictive capabilities of ML systems, which can result in a high financial impact (Lin *et al.* 2017).

To alleviate the effect of imbalanced data, the data-point level approach consists of re-sampling techniques to handle imbalanced classes (Hasanin *et al.* 2019). These techniques consist of Oversampling, Undersampling, and Feature Selection (Sotiris, Dimitris, and Panayiotis 2006). An experimental study was conducted on two credit card datasets, surveying the techniques of the data-point level approach to determine the best method of alleviating the imbalanced class distribution problem.

1.3 Aim and Objectives

The aim of the study was to model a data-point technique that improves the ML fraud detection model's predictive accuracy and produces accurate results with imbalanced credit card datasets. To expedite the achievement of the aim; the research objectives of the study are:

[RO1] - To examine the machine learning fraud identification algorithms with imbalanced credit card fraud datasets.

[RO2] - To investigate in-depth, the data-point level methods of handling imbalanced datasets.

[RO3] - To determine whether the data handling methods improve the predictive accuracy of the ML credit card fraud identification models through extensive comparative and statistical analysis of the results.

The details of how the research objectives were achieved are conferred in the research methodology section.

1.4 Significance of the Study

The study is significant because billions of dollars are lost by the financial industry every year due to fraudulent activities, which results to loss of faith in financial institutions by investors and consumers, and disrupts the economy. Therefore, this gives impetus to the development of robust checking mechanisms to reduce the occurrence of such crimes. The data-point level approach was extensively studied to address the misclassification problem created by imbalanced data in identifying CCF using ML. The study further proposed an effective method that was more efficient in handling imbalanced credit card datasets. Financial institutions and researchers can use the findings to develop and improve the accuracy of fraud detection solutions with imbalanced credit card datasets.

1.5 Scope and Delimitations of Study

The study focused on the misclassification and bias that occurs in the identification and classification of credit card fraud due to imbalanced data. The study was limited to researching the data-point approach to solve the data imbalanced problem in credit card datasets. The scope of this was to only address imbalanced credit card datasets, therefore does not cover any other datasets outside the scope. The solution presented to solve the data imbalance problem is only limited to credit card datasets and cannot be generalised to cater for imbalanced datasets across various areas.

1.6 Research Output

The research output through a scholarly or professional publication was the findings and knowledge of handling imbalanced data to reduce misclassification and bias in identifying fraud or anomaly patterns in massive amounts of financial transaction records.

The details of the Department of Higher Education and Education (DHET) approved journal publications are as follows:

Mqadi, N., Naicker, N., and Adeliyi, T. 2021. A SMOTe based Oversampling Data-Point Approach to Solving the Credit Card Data Imbalance Problem in Financial Fraud Detection. *International Journal of Computing and Digital Systems*, 10(1): 277:286. DOI: http://dx.doi.org/10.12785/ijcds/100128

Mqadi, N., Naicker, N., and Adeliyi, T. 2021. Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss. *Mathematical Problems in Engineering*, 2021(1): 1:16. DOI: https://doi.org/10.1155/2021/7194728

1.7 Structure of the Dissertation

This presentation of the study was organised in five chapters, which are arranged in the following manner:

Chapter One: Introduction to Study – This chapter introduces the study, outlining the research problem, the aim and objectives, and states the main research question of the study.

Chapter Two: Literature Review – This chapter unpacks and reviews the work that has been done on fraud detection and identifies the ML algorithms and other methods that have been experimented with and implemented to detect and prevent credit card fraud.

Chapter Three: Research Methodology – This chapter explains the research methodology and how the study was conducted to achieve the objectives. The chapter also gives an in-depth explanation of the dataset, tools, and actions carried out in the study.

Chapter Four: Results, Analysis and Discussion – The findings from the experiment and quantitative analysis are herein presented on this chapter. The chapter also provides graphical visualisation of the results and a detailed discussion based on the findings.

Chapter Five: Summary, Conclusions and Implications of Study – This is the last chapter which provides a summary, conclusion and presents the study contributions. This chapter explains how the aims and objectives introduced in Chapter One were achieved to answer the research question. The chapter also provides insights into areas for future research.

1.8 Chapter Summary

The performance of ML techniques is highly dependent on the data because they learn from past cases. In real life, the credit card datasets are highly imbalanced; therefore, these techniques fail to perform as expected in actual situations. This chapter introduced the credit card imbalance problem and the potential of the data-point level approach to solve the misclassification of imbalanced credit card datasets. It further highlighted the aim and objectives of the study. One journal publication was generated as a research output. Lastly, the presentation of the structure of the dissertation. The next chapter reviews the literature related to this study.

CHAPTER TWO:LITERATURE REVIEW

2.1 Introduction

The objective of this literature study is to unpack and critically review the work that has been done on class imbalance, fraud detection, ML algorithms and alternative methods used for imbalanced datasets and classification problems. The chapter covers literature on ML algorithms that have been used in classification problems, and provides an in-depth look at the potential role that the data-point approach methods have in mitigating class imbalance in fraud detection and other domains.

2.2 Machine Learning Algorithms

Lejon, Kyosti, and Lindstrom (2018) used an in-depth qualitative research approach to explore ML methods within the manufacturing industry. The data utilised in the study was from industrial control systems. The finding revealed that ML algorithms are appropriate for identifying anomalies in the complex processes forming different automotive parts and components. Sadgali, Sael, and Benabbou (2018) conducted a study with the aim to review and classify the best practices and approaches that were implemented, which yielded the most admirable results.

The study carried out a comparative investigation of fraud detection approaches used to detect financial fraud, including ML. According to Sadgali, Sael, and Benabbou (2018), techniques such as ML play a significant role in fraud detection, since they permit for the extraction and discovery of hidden meanings in large data. In the study, the results revealed that the most used approaches in detecting fraud are those that are hybrid in nature as they use a combination of strengths from multiple traditional methods to detect anomalies.

In addition, Sadgali, Sael, and Benabbou (2018) discovered that, no single solution caters for all the different variations of fraud and there are unique constraints specific to each distinction of fraud. Randhawa *et al.* (2018) conducted an experimental study to identify Credit Card Fraud (CCF) using ML algorithms. Twelve algorithms in total were implemented to carry out this experimental study in collaboration with the AdaBoost and majority voting methods. The study applied both standard models and hybrid methods that applied majority voting and AdaBoost techniques. During the evaluation process for the efficiency of the model, a credit card dataset was used which was publicly available for use. Thereafter, a real-world credit card dataset was consumed that was obtained from an undisclosed financial institution. The strength of the algorithms was demonstrated by conducting a noise test on the data samples. The study implemented the Matthews Correlation Coefficient (MCC) metric as a measure of performance. According to Randhawa *et al.* (2018), the MCC metric considers the predicted TN, FN, TP, and FP as an outcome. The result of the experiment indicated a positive outcome, which was that the majority voting technique yielded a good rate of accuracy in detecting CCF. The majority voting technique after adding 30% noise to the dataset achieved the best MCC score of 0.942.

2.2.1 Support Vector Machine

The Support Vector Machine (SVM) is an algorithm built for classification problems, which uses supervised learning to differentiate between fraud and normal classes (Minku, Wang, and Yao 2016). A hyperplane is constructed to isolate the transactions on either side of the hyperplane and grouping them as fraud or normal. Zareapoor and Shamsolmoali (2015) conducted a study to examine the performance of well-known state of the art practices used to foresee credit and debit card fraud, namely, Naïve Bayes (NB), SVM, K-Nearest Neighbours (KNN) and the Bagging ensemble classifier.

The performance of these advanced data mining methods was studied on a credit card dataset containing real-life transactions. The validation of the fraud detection technique for both credit and debit cards was performed using the 10-fold cross validation technique. The findings showed that based on the Decision Tree (DT); the Bagging classifier produced favourable results with this type of data. The technique displayed the ability to manage class imbalances in detecting fraudulent credit card transactions. Subudhi and Panigrahi (2015) presented a different technique that made use of One-Class SVM created in the form of Quarter Sphere-SVM.

The aim of this study was to tackle the issues of detecting and flagging users' fraudulent calls on mobile devices by undertaking a comparison of the past usage patterns with the pattern of recent calls. The proposed approach's performance was tested using a reality mining dataset. A comparative analysis was conducted using Quarter Sphere-SVM. The findings revealed Quarter Sphere-SVM usage produced better outcomes compared to the ordinary SVM.

2.2.2 Logistic Regression

A Logistic Regression (LR) classifier uses statistical functions to give interpretations to individual classes and organises them into clusters (Jeon and Lim 2020). The transformation of the classification is achieved by utilising the logistic sigmoid method to return a probability value that is mapped, which could either be a fraudulent or legitimate transaction in the context of CCF (Bauder and Mkoshgoftaar 2018). Only specific categories or values are allowed on the LR predictions. Albashrawi (2016) conducted a review of the studies that have been conducted on data mining techniques as an approach to combating financial fraud during a period of a decade.

A total of 41 techniques were reviewed that had been applied across a spectrum of financial solutions that are vulnerable to fraudulent attacks. The study found that, based on the comparison between the data mining methods used from 2004 – 2015, the LR method was the leading model with 13% usage. The second position was tied between the DT and the Neural Network (NN) algorithms, each achieving 11%. Albashrawi (2016) supports the findings by Zareapoor and Shamsolmoali (2015) in which the decision tree was among the best performing techniques for detecting credit card fraud.

2.2.3 Decision Tree

The DT is an ensemble classifier that constructs decision structures from the classification with training data. A tree-like structure is created by the classifier, where the classifications are symbolised by leaves, the features are symbolised by non-leaf nodes, and the combinations of features that lead to the classifications are symbolised by branches (Liu *et al.* 2020). West and Bhattacharya (2016) analysed the issues that were relevant for experiments in detecting credit card-based fraud. Their study tackled general problems in data mining that have been investigated by previous researchers in different domains.

They recognised some aspects that were not thoroughly investigated in the financial fraud detection setting; including the presentation of a problem, selection of features, and the performance metrics. Furthermore, their study explored these issues with simulations in a controlled environment, focused mainly on algorithms that are used for detection, selection of features, and metrics to measure performance for credit card fraud. Findings revealed that most detection algorithms have strengths in only a limited number of areas and recommended that different areas be assigned corresponding priorities to determine the suitable algorithm.

2.2.4 Neural Network Algorithms

A study by Khoshgoftaar, Hulse, and Napolitano (2010) presented a broad experiential investigation that used imbalanced data with labelling errors to learn using NN algorithms. The initial phases of their study examined the effect of class imbalance and class noise on two NN learning algorithms. The next phase studied the impact of addressing the data imbalance problem by improving the performance of the NN algorithms using data sampling. They trained and tested over two million classifiers. The study then concluded that the NN responded differently to resampling data from other commonly used C4.5 classifiers and observed that the performance of NN improved after resampling the data.

According to Su and Hsiao (2007), multivariate data is diagnosed and forecasted using the Mahalanobis-Taguchi System (MTS) method. Instead of learning directly from the dataset for training, the MTS classifier is created by assembling an incessant measurement scale. Consequently, Su and Hsiao (2007) conducted a study to compare numerous common classification methods with MTS. The finding showed superiority of MTS as a vigorous method to address the imbalanced data classification problems. Furthermore, to define the MTS classification threshold, the study developed a probabilistic threshold technique. Lastly, the MTS method was used in mobile phone manufacturing to analyse the radio frequency inspection process.

Implementation results on the imbalanced data gathered from the process of radio frequency inspection revealed that the attributes of the inspection were significantly reduced. Furthermore, the high inspection accuracy can be maintained by the radio frequency inspection process. Chouiekha, Hassane, and Haj (2018) examined how the Convolution Neural Network (CNN) performance compared to the old-style ML methods, which recognise events that were fraudulent in a network of mobile communication. Deep learning techniques were used on a fraud dataset obtained from a real mobile carrier containing records of customer details. The extraction and classification of features used in learning were categorised into events of fraud and non-fraud. Experiments of multiple types were executed to assess the quality of the model that was proposed. The study discovered that the performance was better in terms of training duration and accuracy of the Deep-CNN method versus traditional algorithms of ML. These algorithms were the Gradient Boosting Classifier, Random Forest (RF), and SVM. Hi and Jiang (2019) investigated a fault diagnosis model that is data-driven to deal with imbalanced chemical data streams. The faults that occurred were diverse with various frequencies that arrive constantly in chemical plants. The study proposed a novel technique to support the imbalanced data stream of fault diagnosis called incremental Imbalance Modified Deep Neural Network (incremental-IMDNN).

The combinations of active learning and employment of an imbalance-modified technique for the collection and creation of the most valuable information was the initial phase in designing the incremental-IMDNN. To mine potential information, a Deep Neural Network (DNN) diagnostic model was utilised. Thereafter, the DNN model was passed on to each following increment in a hierarchical manner to cater for fault modes that were new and continuously arriving. According to Hi and Jiang (2019), their model used existing knowledge (inherited and hierarchically) to train and develop the diagnosis model derived on the faults' similarity. The technique uses dynamic data, which differs from traditional models that use static copies of the data to train.

A superclass is formed by merging the same class of faults that have been identified by Fuzzy Clustering, and all the sub-models have a uniform architecture and allow parallel training, which was established in previous research. The performance was validated on a Tennessee Eastman dataset, and the experiment results revealed that incremental-IMDNN was superior to existing methods and had substantial adaptability and robustness. An empirical study by Zhou and Liu (2006) looked at the consequence of using threshold moving and sampling for NN cost-sensitive based training. The training data distribution is modified by both methods, such that the appearances of the cases is explicitly used to convey the costs of the cases.

Threshold moving is used to align the classes that are inexpensive with the threshold output so that it becomes easier to classify the cases with classes that have a higher cost. The study also tests the use of soft-ensemble and hard-ensemble, i.e., the hybrid of sampling and threshold moving based on the soft or hard voting schemes. The empirical study used 21 University of California Irvine (UCI) datasets that contained three variations of cost metrics and multiple real-world cost-sensitive datasets. The findings revealed that the learning of cost-sensitive with two-class tasks was less complex than with multiclass tasks, and the complexity may increase with a higher level of class imbalance. Additionally, the study showed that majority of the methods were effective with two-class tasks, whereas the majority struggled and influenced the results negatively with multiclass tasks. The overwhelming conclusion was that soft-ensemble and threshold moving are suitable choices for NN cost-sensitive training.

2.2.5 Other Algorithms

The widely used fraud detection algorithms are the SVM, DT, LR, and RF. Many researchers (Chouiekha, Hassane, and Haj 2018; Jiang *et al.* 2018; Wang and Jin 2020) proposed their own unique models using different algorithms. The performance of the models was compared against existing and most used techniques through conducting numerous experiments. Mubalik and Adali (2017) suggested using an Artificial Neural Network (ANN) based on Multi-Layer Perceptron (ANN-MPL), built on top of the well-known implementation of ML strategies. According to Mubalik and Adali (2017), it provides the ability to expect and swiftly identify fraud.

Zanin *et al.* (2017) introduced two strategies that had the ability to detect a real-world card transaction dataset. These strategies included complex and hybrid data mining networks, and the algorithms for classification. Malini and Pushpa (2017) applied the KNN algorithm and detection outlier techniques to find the most effective solution for the fraud identification problem. Askari and Hussain (2017) proposed a fraud identification solution built on Fuzzy-ID3. The study revealed that the solution was effective in identifying fraud.

Some improvements on these methods were observed and contributed to the existing fraud detection knowledge of the strengths and weaknesses of the models. For instance, Chouiekha *et al.* (2018) showed that the Deep-CNN method displayed better accuracy when compared to traditional ML algorithms. However, these solutions fail to address the shortfalls and the persistent issues currently existing with fraud detection models such as eliminating false positives and pro-actively identifying fraudulent financial transaction in real-time, with better accuracy. Other researchers (Shevchik *et al.* 2017; Randhawa *et al.* 2018) have proposed using a hybrid approach.

Randhawa *et al.* (2018) experimented using ML to detect fraud with a mixed approach that applied the majority voting and AdaBoost techniques, while Shevchik *et al.* (2017) examined the detection of lubricated surfaces failures using the RF algorithm and Acoustic Time–Frequency Features. To identify the best performing method, many other studies compared and reviewed existing financial fraud detection models against each other (Albashrawi 2016; Patil, Nemade, and Kumar 2018; Sadgali *et al.* 2018). A study by Patil *et al.* (2018) utilized the confusion matrix to review the models and discovered that the RF model performance was superior to that of the DT and LR after measuring the performance using well-known metrics, i.e., recall, precision, and accuracy. In contrast, a data mining study by Albashrawi (2016) found that the most effective technique for financial fraud detection was the LR model.

2.3 Data Sampling Approach

The data-point level approach comprises of mediations to relieve the class imbalance effects on the data, and it offers the versatility to support most of the latest algorithms, which include but are not limited to SVM, DT, and LR (Scornet 2016).

2.3.1 Undersampling, Balance-Cascade, and Easy-Ensemble

An experiment conducted by Liu, Wu, and Zhou (2009) proposed two algorithms to deal with the deficiency of applying Undersampling to control the influence of imbalanced classes. The Undersampling deficiency was the disregarding of numerous majority classes during resampling. Therefore, Easy-Ensemble and Balance-Cascade was proposed. The majority class is divided by Easy-Ensemble into numerous smaller portions, then the learner is trained per portion per iteration, and finally, each output from the iterations get bundled together.

Balance-Cascade utilizes a sequential training-based approach, wherein each sequence, examples belonging to the majority class that have been rightfully classified are discarded to avoid duplication in the following sequence. During a comparison with other approaches, both the Easy-Ensemble and Balance-Cascade showed a higher Area under the ROC Curve (AUC), F-measure, and G-mean values. According to Liu Wu, and Zhou (2009), the training time closely matched to Undersampling, and compared to other approaches, it was considerably faster.

2.3.2 Logical Graph of Behaviour Profiles

Zheng et al. (2018) performed an experimental study on detecting transaction fraud-based behaviour diversity and total order relation, which was used to identify fraudulent transactions from online-generated transactions. The study proposed the utilisation of a Logical Graph of Behaviour Profiles (LGBP) that was a complete set of order-based models to denote the logical relation between the attributes and features of transaction records. A path-based shift likelihood from one attribute to another attribute can be computed based on transaction records of users' and LGBP (Zheng et al. 2018). The study also defined a Diversity Coefficient based on information entropy to outline users' transactional behaviour diversity. Furthermore, the study defined a State Transition Probability matrix for the user's temporal transaction features to be captured. The study collected a sample of actual data from 70 users. The dataset contained 70 records of online shopping transactions for each user on TaoBao.

The study selected five columns to use for the transaction records: transaction_location, shipping_address, category_of_goods, transaction_time, and amount. According to Zheng et al. (2018), when detecting transaction fraud, those were the five most effective attributes. The study compared the performance of the model against Srivastava's method that employed Markov Chain as the Behaviour Profile model and ran trials using data generated for the simulation. In addition, a comparison was made against two other methods for detecting anomalies namely, self-organising maps-based and Bayesian learning-based fraud detection methods. The comparison was made using six popular evaluation metrics. The results of the experiment indicated that the introduced model incapacitates the limitation of models based on the Markov Chain because of the diversity characterisation of the behaviours of users.

2.3.3 Oversampling- and Undersampling-Based Online Bagging

Wang, Minku, and Yao (2015) suggested using two learning algorithms that constructed an ensemble classifier capable of dealing with class imbalances in a live environment using time-decayed metrics and resampling techniques called Undersampling-based Online Bagging (UOB) and Oversampling-based Online Bagging (OOB). The study further attempted to create an advanced strategy for resampling by amending UOB and OOB, and to improve their performance with data streams that are dynamic or static. The paper firstly provided a broad analysis of data streams with class imbalance, which covered the distributions of data, rate of imbalance, and the status alterations in class imbalance.

The results showed that UOB was more effective at identifying examples of the minority-class in data streams that were static while OOB was stronger with changes that were dynamic in class imbalance status. It was found that distributions of the data were a major factor towards their performance. Wang *et al.* (2015) further proposed two novel ensemble techniques that preserved both UOB and OOB called Weighted Ensemble-based Online Bagging (WEOB) version 1 and 2. Both the ensemble techniques had adaptive weights for the ultimate prediction and were found to display robustness and good accuracy while maintaining the strengths of UOB and OOB.

2.3.4 Locality Sensitive Hashing based on SMOTe

Hassib *et al.* (2019) presented a framework comprising three stages that improved the performance of optimisation algorithms for mining imbalanced Big Data by eliminating problems with the local optima. Stage one was a pre-processing stage that used Locality Sensitive Hashing (LSH) based on a modified version of the Synthetic Minority Oversampling Technique (SMOTe) method for addressing the problem of class imbalance. The instances of the dataset were hashed into buckets of data using LSH solution. In stage two, they implemented a bucket search process that used Grey Wolf Optimizer (GWO) to allow searching within each bucket for the global optimum for training Bidirectional Recurrent Neural Network (BRNN).

The final third stage introduced the combined GWO-BRNN approach, which scanned all global optimums from all the buckets to determine the global optimum of the entire dataset. Hassib *et al.* (2019) proposed a framework called LSH-GWO-BRNN, which was evaluated on nine datasets against seven ML algorithms using AUC and Mean Squared Error (MSE) metrics. The algorithm was tested over four well-known datasets and measured by the accuracy of the classification and MSE. The findings substantiated that their introduced framework delivered high avoidance of local optima, showed fewer complexities, greater accuracy, and minimum overheads. Pes (2020) investigated a combined approach to improve performance that was made up of techniques for addressing the issue of imbalanced classes mixed with techniques for reducing dimensionality.

The study combined multiple variations of feature selection with different data sampling strategies and cost-sensitive grouping techniques. An experiential study was executed to evaluate the resulting learning systems on six benchmark datasets using leading classification algorithms. According to Pes (2020), the superiority of the combined learning approach, which aims to solve both the issue of imbalanced classes and high dimensionality, was proven by the study with using the commonly adopted paired *t*-test, at a confidence level of 0.05. Their findings revealed that the combined approaches are significantly better than single techniques such as feature selection for datasets that are mostly imbalanced.

2.3.5 Spatio-temporal Oversampling and Selective Down-Sampling

Zhang *et al.* (2017) stated that the problem of class imbalance was with background subtraction in video data, where the minority class was the foreground and the majority class was the background. The authors introduced a framework for imbalance compensation to deal with background subtraction by exploring temporal and spatial correlation, characteristic in video data. The framework comprised of imbalance-compensated Bayesian classification and imbalance-compensated bilayer modelling sequential components. The first component operated at the data level, which contained Selective Down-Sampling (SDS) and Spatio-temporal Oversampling (SOS) to fix the imbalanced data.

SDS works by deleting the majority class samples that are overlapping in the data selectively, whereas SOS rebalances the data by increasing samples of the minority class (Zhang *et al.* 2017). Consequently, after the first component, a bilayer model was trained using the resulting data. The second component operated at the algorithm level and introduced new cost functions to address the effect of imbalanced classes. The imbalance measurement built the cost functions, which were consumed in the Bayesian classification scheme to create the prior term. The findings based on the trials executed using public datasets proved that the introduced approach was more effective than the benchmark.

2.3.6 Association Correction for Imbalanced Data

A study by Bao, Huang, and Bao (2018) found that the issue of imbalanced data causes highly biased classifications, which leads to inaccurate flagging of the true class relationships. To address this issue, Bao *et al.* (2018) introduced a computational framework that was designed to enhance the performance of the Genome-Wide Association Study (GWAS) with imbalanced data, called the Association Correction for Imbalanced Data (ACID). The ACID framework, built on the theory of imbalanced learning, was enhanced to perform the function of association discovery from sequential genomic data. After conducting the experiments, they observed that the efficiency of GWAS on a highly imbalanced dataset was improved dramatically using

ACID. They tested the effectiveness of ACID by performing a GWAS analysis with two types of cancer datasets that were highly imbalanced, namely, Bladder and Gastric Cancers. The finding showed that, compared to regression methods, ACID was more sophisticated in discovering suspicious loci and was found to be consistent with existing discoveries.

2.3.7 Conditional Generative Adversarial Network-Mean-Covariance Balancing Labelling

Li, He, and Li (2019) proposed a method to enhance the reliability evaluation of transmission gears with data that is imbalanced and insufficient, called Conditional Generative Adversarial Network-Mean-Covariance Balancing Labelling (CGAN-MBL). A model was built based on a Conditional Generative Adversarial Network (CGAN) to produce reliable instances matching the original distributions of the data. During the study, they found that when adding a network weight initialisation scheme to the model, the capability to escape from local optimum was improved and the CGAN feature of universal search was strengthened. They designed a mini-batch discriminator with the ability to iteratively separate generated and original batches of data.

Secondly, to alleviate the weaknesses of CGAN with generated and unlabelled data, they introduced Mean-Covariance Balancing Labelling (MCBL) to label the data that was generated. MCBL works by discovering the closest distance amongst original class centres and generated instances. Lastly, the classifiers were trained using artificial data created by combining the original and generated data. The experiment outputs revealed that their approach was very effective for a transmission-gear imbalanced dataset that was captured from the real world. The study tested different models and they found that the RF model performed effectively on the artificial data; the ability to classify the minority class improved significantly.

2.3.8 Dynamic AdaBoost classifier and SMOTe

A study by Lin *et al*. (2019) stated that a major problem facing companies was lack of support for off-the-shelf offline classifiers. They found that support for ensemble learning has focused mostly on online classifiers in past studies. To provide offline support, Lin *et al*. (2019) introduced a new ensemble-learning algorithm with a three-stage Condition-Based Maintenance (CBM) to handle imbalanced data with concept drifts. The first stage involved training an ensemble classifier.

The second stage attempted to detect concept drifts in imbalanced data using an enhanced Linear Four Rates (LFR) technique. The third stage handled the imbalanced data using SMOTe and used the early learnings to create an improved Dynamic AdaBoost classifier. A dataset containing various levels of imbalance was used to conduct the experiments. The findings revealed that all the concept drifts were detected successfully with their proposed solution, and there was a 94% accuracy rate in minority class detection, which was the highest they recorded.

2.3.9 Novel Class Imbalance and Classification Methods

Gomaa and Jalloul (2014) proposed a new method for data-aided approximation of imbalance phases and mixer gain in the analogue frontend of direct-conversion receivers. Their study estimated the In-phase and Quadrature Imbalance (IQI) isolated parameters from the failing channel projection to reimburse the time-domain using projected IQI parameters. According to Gomaa and Jalloul (2014), the method was inspired by the situation in which the antenna ports used for the transmission of data resided in different locations from the IQI projection-training symbols that were transmitted. In this situation, they found that the channel projection and conventional IQI composite was not suitable for detecting the signal, since different channels were used by the data and training symbols.

Statistical results revealed that their method obtained up to 10-decibel (dB) enhancement with just two training symbols. Jiang *et al.* (2018) introduced a unique technique for fraud detection made up of four stages: pre-processing data; clustering behavioural patterns; classifying the patterns of behaviour and assignment; and updating the behavioural profile of cardholders via a feedback mechanism. To deepen the behavioural patterns of the cardholder, in the initial phases the study utilised the historic transaction records of the cardholder. In experimenting, Jiang *et al.* (2018) proposed a window-sliding technique to provide a cumulative total of the transactions in each group. Next, for each cardholder, they performed an extraction for a pool of precise behavioural patterns founded on the transactions of the cardholder that were grouped and historical.

Thereafter, for each cluster, a set of classifiers was trained on the foundation of all behavioural patterns. Lastly, in the final phase, the classifiers were used to identify online fraud and if the incoming transaction received a fraudulent flagging, a mechanism for feedback was taken in the process of detection to unravel the issue of notion drift. To choose the best fraud detection model, the study used a framework by Shen, Tong, and Deng (2007) for performance evaluation of the technique that was proposed against LR and RF. The experiment findings demonstrated that the introduced technique had a performance that was better than the two methods. According to Cano, Zafra, and Ventura (2013), the purpose of Data Gravitation Classification (DGC) was to categorise the samples of data by evaluating the gravitation among distinct classes. However, they discovered that due to the distinct relevance of the attribute of data for computation of distance, the issue of class imbalance, and attributes that are noisy and lack relevance, gravitation calculation was a more complex problem.

Therefore, Cano *et al.* (2013) presented an improved classification algorithm that was gravitation-based but featured enhancements to alleviate the identified problems with existing gravitation models. The improved algorithm was called Data Gravitation Classification Plus (DGC+), in which per class classification, a matrix of weights was introduced to define each attribute's importance that was used to assess the distance among the samples of data. The performance for the classification was improved by considering the local and global information of data, particularly within boundaries of the decision. They experimented using numerous statistical nonparametric tests and measured DGC+ against well-known classification algorithms based on 44 imbalanced datasets and 35 standards. The findings showed that the DGC+ algorithm out-performed the majority of the well-known solutions

A study by Pozzolo *et al.* (2018) was carried out to elucidate the measure of performance that was most suitable to be used for the purpose of detecting fraud. A new strategy was developed for learning that addressed class imbalance effectively, concept drift, and latency for verification was designed and evaluated. Two large datasets of European Credit Card owners' transactions from online e-commerce were used. Pozzolo *et al.* (2018) also performed an investigation to illustrate the consequence of concept drift and imbalanced classes within the real-world context with a data stream consisting of over 75 million transactions, which were authorised over a three-year period. The findings show that in learning, it is obligatory to convey higher significance to the feedback of the issues to induce clear-cut alerts. It was also shown that in the learning process, the solutions with inferior emphasis on feedback were more likely to provide less clear-cut alerts.

2.3.10 Novel Undersampling and Class-Based ensemble for Class Evolution

Minku, Wang, and Yao (2016) investigated the gradual disappearance and emergence of classes' phenomenon. Their study introduced the Class-Based ensemble for Class Evolution (CBCE) approach. This was designed to quickly adapt to class evolution by retaining a base learner per class and updating them with fresh data. Thereafter, the paper proposed a new Undersampling technique to address the vibrant imbalanced class problem resulting from the classes' gradual evolution. According to Minku *et al.* (2016), the CBCE effectiveness was demonstrated by empirical studies in several class evolution situations, which compared findings to class evolution adaptation methods that already exist.

The study revealed that CBCE adapted better than existing techniques with all three of the class evolution cases, namely, disappearance, emergence, and reoccurrence. The empirical studies verified that CBCE was more reliable and performed better than the well-known class evolution algorithms regarding the ability to adapt and the performance of the overall classification. However, according to Minku *et al.* (2016), they discovered some disadvantages of CBCE, which included the rating of class importance. For example, emerging or non-evolved classes might be more important than a disappearing class in some real-world applications. They found that performance with non-evolved classes might decay since the evolved classes are given more prominence. Therefore, the study concluded that mining data streams was still a complex task.

2.3.11 Pattern Recognition Model

Adedoyin (2016) investigated a novel mobile money predictive model for pattern recognition of transfer transactions. The study replicated various scenarios of mobile money fraud to create an artificial dataset containing transfer transactions. The model's applicability was examined using the artificial dataset and the findings revealed that the model produced promising performance with recognising fraudulent patterns. They discovered that experts can use the rankings of transactions into groups provided by the findings to gain insight into transactions that are suspicious for proactive investigating. Weng, Chang, and Chiong (2009) presented an active balun with low imbalance, dc-21 gigahertz (GHz) using a two µm InGaP/GaAs HBT process for fast data communications.

The 3dB bandwidth was enhanced for the active balun proposed by adopting a Darlington cell. The method addressed the phase errors that occurred among the differential output ports due to diverse numbers of stages, by designing a feedback capacitor. The experimental results showed that there was a 1.2 dB imbalance amplitude maximum, a 2.5dB small signal-gain average, a below 5 phase error, and a 21GHz broad bandwidth improvement. The balun group delay measured was less than 30ps with minimum disparity. Furthermore, they presented an eye diagram containing up to 12.5Gbps of bit-stream pseudorandom. They found that based on the group delay and low imbalance, the active balun was more suitable for fast data communication.

2.3.12 Range Estimation, Classification, and Clustering

To address the challenge of insufficient data, Fang *et al.* (2019) proposed the novel and customised Clustering, Classification, and Range Estimation (CCRE) statistical method. The CCRE method identifies a match amongst the rich data network having clusters with current phase time series and the scarce data network having only the annual average current phases. A range estimation was then performed by CCRE for the energy loss that was imbalance-induced, which the network with scarce data become resembled by the network clusters that were data rich.

The range was narrowed down by applying the Chebyshev’s inequality theory that represented the energy loss, which was imbalance-induced, interval of confidence for the networks with scarce data. The study showed that over 80% of the networks with scarce data that were clustered correctly due to having small amount of data. Furthermore, the energy loss that was imbalance-induced achieved 89% confidence estimation.

2.3.13 NSE-SMOTe and GSMOTe-NFM

Ditzler and Polikar (2013) defined a hybrid method for concept drift that was based on ensemble learning for imbalanced data. In the first method, to address the data imbalance problem they used SMOTe and logically combined it with the NSE algorithm for concept drift to create an integration method called NSE-SMOTe. The mechanism for weighing errors that were class-independent from NSE with a penalty constraint provided the ability for the algorithm to be forced to balance the accuracy on all classes. The novelty of their method was provided by the capability to apply the accuracy of modified class imbalance and the time of each classifier to past and current environments, in order to determine the weights of voting for members of ensembles to combine.

They found that the proposed method was more favourable than the benchmark method, which indicated the ability for NSE-SMOTe to tackle the challenge of concept drift and the problem of data imbalance. Cheng *et al.* (2019) found that the local distribution characteristics tend to be neglected since SMOTe assigns only the global neighbourhood K parameter and during Oversampling, the noisy information is usually propagated. To address these two problems synchronously, they introduced a Grouped SMOTe with Noise Filtering Mechanism (GSMOTe-NFM) approach (Cheng *et al.* 2019). To gain insight on the real distributions of the minority and majority classes, the algorithm adopted the Gaussian-Mixture Model (GMM).

The comparing of the probability densities of similar instances belonging to two distinct classes allowed the removal of instances with the most noise. The remaining instances of the minority and majority classes were used to construct two new GMMs. The corresponding information of the probability density formed the basis for dividing all instances of the minority class into three distinct groups, namely, outlier, safety, and boundary. Lastly, a single K parameter was assigned to instances that belong specifically to each group, for new instances to be generated. Cheng *et al.* (2019) performed experiments using the three leading classification models, 24 binary-class benchmark datasets, and evaluated GSMOTe-NFM against various Oversampling algorithms. The study revealed that the introduced GSMOTe-NFM performed better than SMOTe variations.

2.3.14 Hybrid Sampling Algorithm Called RUSBoost

Seiffert *et al.* (2010) introduced the RUSBoost solution, which was a combination of Random Undersampling and AdaBoost for learning from a training dataset that was severely skewed. The solution provided enhancements to offer a faster and simpler SMOTe Boost alternative, which combined data sampling and boosting. The RUSBoost and SMOTe Boost solutions with their component performances were assessed individually. Experimental research was carried out to examine 15 datasets sourced from different application domains, four metrics for evaluation, and four classifiers. The findings supported the introduction of RUSBoost, which was faster, simpler, and performed better than other solutions, including SMOTe Boost.

2.3.15 Map Reduce Implementations of Naïve Bayes Classifier and Random Forest Classifier

Somasundaran and Reddy's (2016) paper focused on one problem that exists prominently in real life applications, data imbalance. This problem arises because of the uneven nature of particular real-time applications. The proposed approach analyses imbalanced data and its impact on the classification process. The Map Reduce applications of RF and NB Classifiers were used on Big Data to identify the reliability and accuracy exhibited by these methods when the training given to the classifier model was biased. The sampled data with several intensities of imbalance were used for investigation. Experiments were conducted and the results exhibited were analysed to identify the threshold levels of imbalance, best sampling technique for Big Data and the comportment of algorithms towards imbalanced data.

2.3.16 Distance-based Balancing Ensemble

According to Chen *et al*. (2019), the ensemble learning method can address the imbalance problem more effectively than other methods. A popular method was the Splitting Balancing Ensemble (SBE) that learns from imbalanced datasets by transforming them into various subsets that are balanced and on which are used to build sub-classifiers. Chen *et al.* (2019) further state that the subsets that are generated by SBE have very limited samples, which leads to under-fitting when learning from a dataset that is highly imbalanced.

Therefore, Chen *et al.* (2019) proposed a method with two tasks called the Distance-based Balancing Ensemble (DBE) to address the problem of under-fitting and enhance the classification algorithms' generalisation performance. The first task of DBE involved creating various unbalanced subsets with a ratio of imbalance that was much lower, from dividing an extremely unbalanced learning dataset. The second task obtained balanced subsets using an Oversampling method that was a weighted semi-unsupervised, adaptive modified method, and applied it to each subset.

The ensemble results were combined using a method that was more effective, called Distance-based Combination Rule (DCR) which was introduced by the study. To show that the DBE model with the DCR was more effective, they assessed the method using 48 public datasets that were extremely unbalanced from publicly available repositories. The DBE-DCR model showed superior performance to other ensemble models.

2.3.17 Deep Fuzzy C-mean Clustering for Imbalanced Multi-class Classification

Rashad, Riaz, and Jiao (2019) introduced the Deep Fuzzy C-mean Clustering for imbalanced Multi-class Classification (DFCM-MC) semi-supervised algorithm. In their study, the depth of how the strategy for decomposition was applied was represented by the word "Deep". Firstly, the semi-supervised original data was decomposed into unlabelled and labelled data. The classification model was trained using both unlabelled and labelled data to extract information that was descriptive and useful. Secondly, multi-intra clusters were formed by breaking down the labelled and unlabelled data to tackle the issue of imbalance in multi-class data that tends to maximise the features of intra-cluster and intra-cluster classes.

The study proposed the DFCM-MC approach that uses multi-intra clusters to extract new features for redundancy control in the classification of multi-class data with high imbalance that is associated with the features with high similarities among multi-intra clusters. Moreover, they addressed the issue of data imbalance by applying the technique for data re-sampling to improve the performance of DFCM-MC classification. Rashad *et al.* (2019) conducted their trials on 18 imbalanced multi-class benchmark datasets, four leading algorithms for learning, and measured the performance with three mean metrics, namely, the AUC, accuracy, and f-measure. The output through the metrics showed that the ability to identify and merge vital information from unlabelled data enabled DFCM-MC to have an edge and perform at a higher level of accuracy.

2.3.18 Beneish M-Score

Tarjo and Herawati (2015) conducted a study to examine the capability of Beneish M-Score in identifying fraud on financial data. The study used the data of the corporations that were found to have committed fraud on the Fraud Database of Sanctions of Issuer Cases Public Companies that was made available by the Financial Services Authority during the period 2001 – 2014. The financial data of companies that committed no fraud for the same period and of the same industries were used for comparison. There was a total of 70 companies in the database, equally classified into a group of two categories; fraud and non-fraud. The classification accuracy to identify fraudulent transactions was 77.1% in the fraud group. The classification accuracy, from the non-fraud group made up of 35 companies, was 80% which meant that 28 were correctly flagged as not committing fraud. This also meant there was a 20% false positive rate, which were companies inaccurately flagged to have committed fraud. Based on the results, the

Beneish M-Score model demonstrated that it does have the ability to identify financial fraud but does not resolve the issue of false positives.

2.3.19 Process Mining Technique and Anomaly Feature Detection

Baader and Krcmar (2018) stated that the red flag technique introduces indications for behaviours that are unusual, while process mining restructures and envisages the as-is process of the business from the core dataset. Baader and Krcmar (2018) explored the enabling of recognition and ensuing visualisation of potential process instances that could be fraudulent by using a mixture of the red flag, process mining technique, and by assigning equivalent red flags to those processes. There were 226 031 instances of processes of the IDES dataset and of the GBI dataset there was a total of 161 101 process instances that formed part of the overall dataset that was analysed.

Eight cases of fraud were identified out of 13 by the prototype within the IDES dataset, whereas the prototype was only able to identify 7 out of 18 implemented cases of fraud in the GBI dataset. The study revealed that the total number of cases of fraud successfully identified as true positive was 15 out of 31 cases, and even though available, the prototype was unable to detect 16 cases of fraud (FN) in the GBI and IDES datasets. There were 1 399 legitimate cases that were identified as fraud and a total of 377 101 legitimate instances that were identified accurately (TN) by the prototype. The rate of FP was shown using the confusion matrix.

However, the study also found that a thorough confusion matrix was difficult to compute with a live dataset, since the number of cases of fraud present in the dataset was undefined, but for both real and synthetic datasets, the rate of true and FP can be determined. According to Huang *et al.* (2018), most of the methods that already exist focused individually on information of features and networks and never consumed both types of information as a unity. Thus, Huang *et al.* (2018) conducted an experimental study using Anomaly Feature Detection to detect financial fraud and proposed a framework for detecting fraud called CoDetect that can consume both the feature information and network information for detecting financial fraud. The study found that after numerous trials with both real world and artificial data, the proposed framework was effective and efficient in preventing financial fraud. They discovered that CoDetect excelled mostly in detecting money laundering fraud.

2.3.20 Feature Selection Based on Random Forest

Yao *et al.* (2020) conducted a study to address the required complicated classification model and the issue of higher feature dimension for recognition of insulation defect in partial discharge. The paper introduced a method combining Random Forest with Variance Analysis (RF-VA) for serialised forward selection in selecting the best subset. Two aspects of the method were improved. The first aspect uses a Variance Analysis based method that measures feature dissimilarities among classes, and attains a revised arrangement displacement scheme to direct the values order rearrangement from a bag released data sample.

The second aspect implemented the method for Arrangement Forward Search used to perform Feature Selection could get the results for iteration evaluation. The method addresses randomness to define the feature subset size and unpredictability of the results found in the original algorithm. The finding revealed that RF-VA achieved a better features subset, reduced the set of characteristics of partial discharge dimensions and improved the defect type of partial discharge identification rate.

2.3.21 Undersampling using Easy-Ensemble and Balance-Cascade

According to Liu *et al.* (2009), an efficient method of addressing the issue of class imbalance was Undersampling because it extracts majority classes to create subsets that match the minority class. However, Liu *et al.* (2009) stated that most examples of the majority class are neglected, which was the main deficiency with Undersampling. The study proposed the Easy-Ensemble and Balance-Cascade algorithms to deal with this deficiency. Numerous subsets of the majority class are sampled using Easy-Ensemble, which uses each subset to train a classifier, and the outputs of those classifiers are combined at the end.

Balance-Cascade performs sequential training of the classifiers, where in each step, the examples of the majority class that were correctly classified get deleted to prevent them from being considered again. The results of the experiment revealed that the G-mean, AUC, F-measure performance measures of Easy-Ensemble and Balance-Cascade were higher than the leading methods that were widely used to combat class imbalance. Furthermore, they found that the duration of training of the proposed methods was approximately similar to Undersampling alone, under the same conditions, which was very fast compared to other Undersampling-based methods.

2.4 Chapter Summary

Many studies have been conducted that focused on fraud detection algorithms and class imbalance problems. This chapter has presented classification algorithms from previous literature that can be applied to build CCF detection systems. Standard models and hybrid methods using the ML algorithms were reviewed, namely, AdaBoost and majority voting, SVM, Naïve Bayes, K-Nearest Neighbours and the Bagging ensemble classifier, Decision Tree, Logistic Regression, Neural Network algorithm, Convolution Neural Network, Deep-CNN, and Artificial Neural Network. More than 22 data-point approached were discussed that were used to solved class imbalance issues across various domains including CCF.

A highlight of some the approached include Undersampling, Balance-Cascade, and Easy-Ensemble, a Feature Selection based Logical Graph of Behaviour Profiles, Oversampling and Undersampling-Based Online Bagging, Locality Sensitive Hashing based on SMOTe, Association Correction for Imbalanced Data, Conditional Generative Adversarial Network-Mean-Covariance Balancing Labelling, Dynamic AdaBoost classifier and SMOTe, Novel Undersampling and Class-Based ensemble for Class Evolution. The performance was measured using well-known metrics, i.e., recall, precision, Matthews Correlation Coefficient (MCC) metric, confusion matrix, G-mean, AUC, F-measure and accuracy. Literature reveals that there is a uniform solution for the fraud detection and class imbalance problem.

The overwhelming conclusion is that if the dataset is imbalanced, the problem of distinguishing between the positive and negative cases continues to persist in detecting CCF. Most studies introduced novel approached that have their own advantages and disadvantages but there has been little research to discover the data-point approach combination that is best to use with the classification algorithms to deal with class imbalance in CCF detection. Hence, the study was conducted to investigate in-depth the methods of handling class imbalance in the credit card datasets to determine the best approach to address the misclassification problem in CCF detection. Our study provides an intensive statistical and comparative analysis of the prediction results before and after using the data-point approach to solve the problem of imbalanced CCF datasets.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Introduction

The research methodology defines the specific procedures and techniques used to select the suitable design and datasets, explains the process for carrying out the experiment, and how the information was analysed to answer the research question. This chapter is structured as follows: research design; scientific research modelling; architecture of the study; tools; and datasets, as well as cross validation, performance metrics, classification algorithms, the data-point approach, and finally, the chapter summary. The data-point approach sub-sections describe the novelty of how the Feature Selection, Undersampling, and Oversampling techniques were utilized to solve the problems created by credit card datasets with an imbalanced class distribution.

3.2 Research Design

A quantitative study was executed using an experimental research design. This design was selected because it is more appropriate when the objective is to test the effects of manipulating the independent variable on the dependent variable (Feldt 1958). An experimental study was considered most suitable for studying the ML model's ability to identify fraud before and after manipulation, through methods of handling imbalanced data in the credit card datasets. The experiment was carried out using ML and the data-point level approach. According to Xiaolong, Wen, and Yanfei (2019) the data-point level approach comprises of data interventions to relieve the effects of imbalanced classes, and it is built with the elasticity to support the latest algorithms such as LR, DT, and SVM.

The study applied a supervised learning strategy that used the classification technique. The powerful tools that come with supervised learning enable data processing and classification using ML (Zhang, Zulkernine, and Haque 2008). Supervised learning was used with labelled data. Labelled data refers to a dataset with classified and structured data. The classification algorithms that were used for the experiments focused on distinguishing binary classes by examining labelled data and identifying related patterns (Wu *et al.* 2012). During the experiments, the classification algorithms were used to distinguish between fraudulent and legitimate transaction.

3.3 Scientific Research Modelling

In scientific research, an idea is simplified and presented using a model, which can be an object, a process, or a system that can clarify and describe a phenomenon that is difficult to directly experience (Krstic and Bjelica 2015). Figure 3.1 below shows the high-level flow of activities to complete the binary classification experiment.

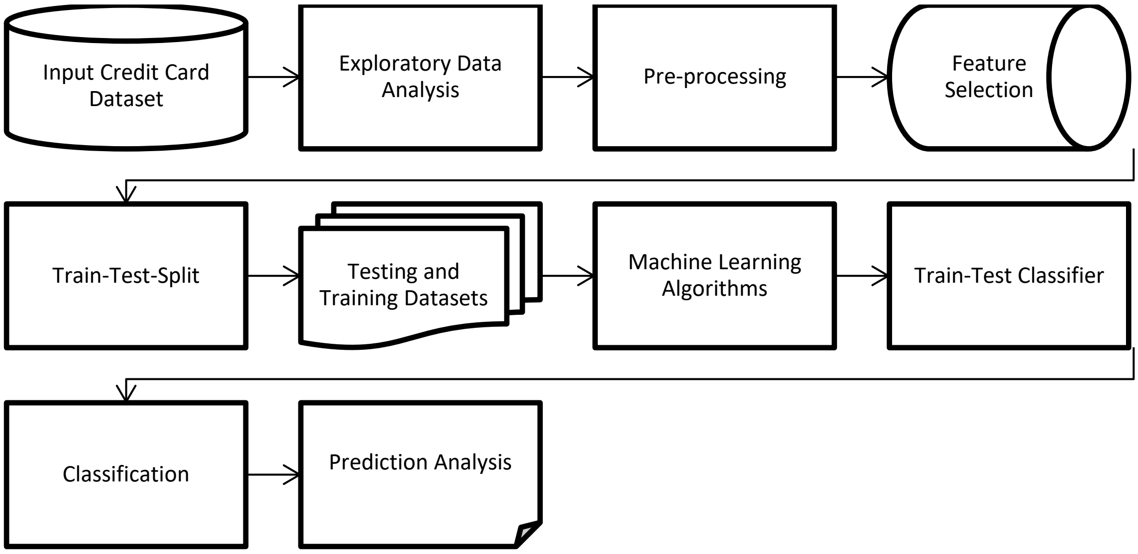

Figure 3.1: High-level experiment process model for binary classification experiment

Figure 3.1 shows that the experiment began by inputting the credit card dataset into the Python notebook. The next step was to perform exploratory data analysis to gain insight on the dataset. Pre-processing followed thereafter to correct any data issues and ensure there are no null values. Feature selection is a data-point technique, defined as the process of selecting features that have an influence on the variable that must be predicted using manual or automated methods. The study explored three Feature selection techniques, namely, Correlation Matrix with Heatmap, Univariate Selection, and Feature Importance.

Train-Test-Split is the process of splitting the main dataset into training and testing subsets. The next step was to build the ML classifiers using classification algorithms, namely, SVM, RF, DT, and LR. The subset datasets were then fed to the classifiers for training and evaluation. The classifiers were used to perform a binary classification that was measured using performance metrics to analyse the accuracy of the prediction. The prediction analysis focused on determining the rates of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) to gain insight on predictive accuracy and the rate of misclassification.

Figure 3.2 below illustrates the overview of the ML classification process.

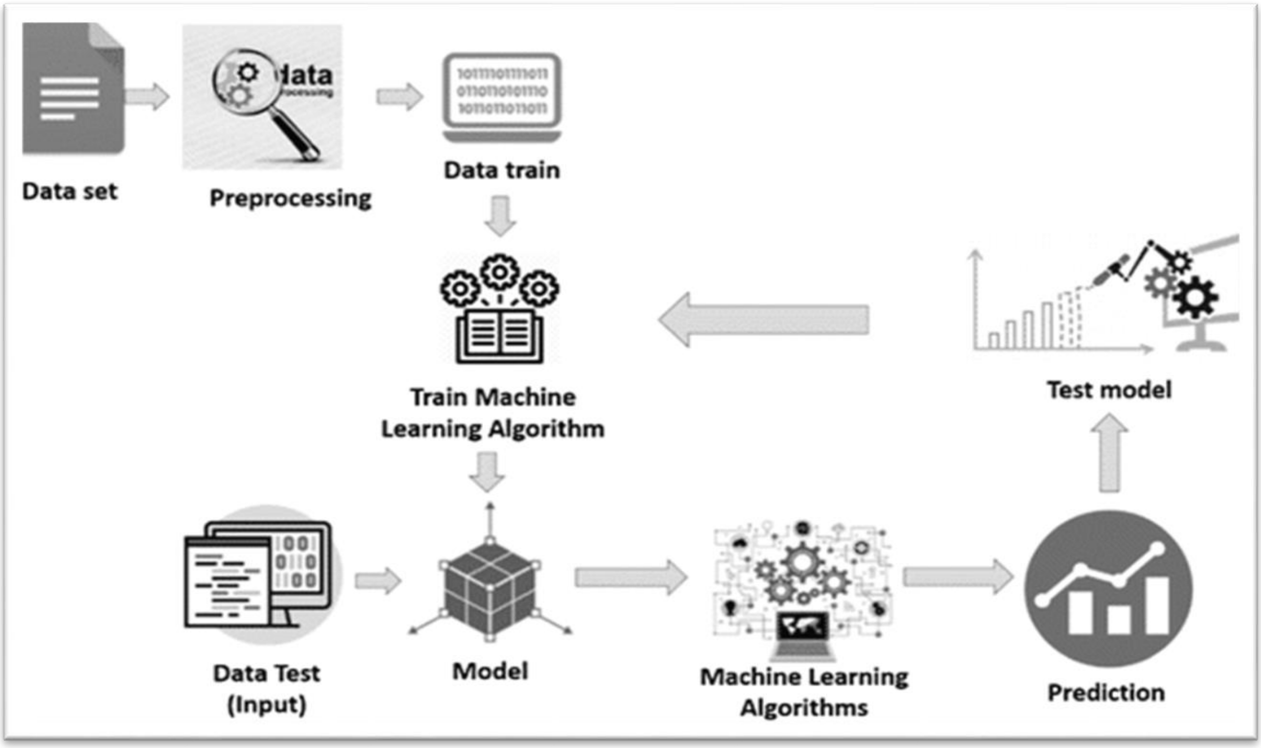

Figure 3.2: Machine Learning Classification Process (Candanedo et al. 2018)

Figure 3.2 describes the flow of activities in a ML classification process (Candanedo *et al.* 2018). Pre-processing was done to clean and validate the condition of the credit card dataset. 'Data train' was the training dataset that was fed into the classification model, which was built using ML algorithms. 'Data Test' was a testing dataset that was fed into the trained Model, which used the ML algorithms to perform a prediction and produce a classification report. The prediction was used to assess the model's predictive accuracy. The process can be iterative and the 'Model' is continuously evaluated and fine-tuned after each iteration to improve the performance.

Figure 3.3 below is a novel model showing the four logical stages used to carry out this experimental research.

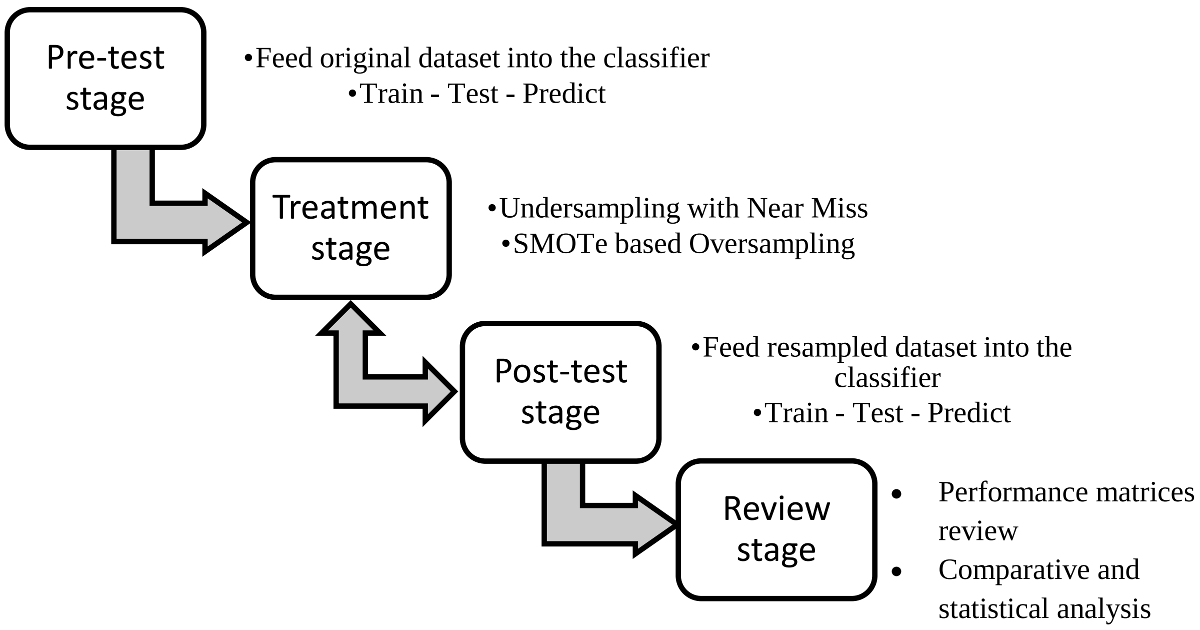

Figure 3.3: Four logical stages for conducting the experiment (Author)

The experiment was executed in four stages: pre-test stage; treatment stage; post-test stage; and review stage. During the pre-test, the original datasets were fed into the ML classifier and each of the four selected algorithms mentioned above were used to train and assess the predictive accuracy of the classifier. The datasets were fed into the ML model using the 3-step loop: training; testing; and prediction. During the treatment stage of the experiment, the data-point level approach methods were applied to the dataset to ease the consequences of class imbalance. The study investigated the Undersampling, Oversampling, and Feature selection techniques to identify the best solution, to yield the most accurate results on the credit card dataset.

The study also investigated the adequate amount of application of each of the three techniques. The output of this stage was a resultant dataset from each method. In the post-test stage, the resultant dataset was taken and again fed into the classifiers. Stages two and three were iterative processes; the aim was to overcome the imbalanced data problem, therefore an in-depth review and analysis of accuracy for each result was conducted after each iteration, to optimise the process for better accuracy. Lastly, the review stage carried out a comprehensive review of the performance of each algorithm for both the pre-test and post-test results. Then, a cross comparison of all the data-point level methods was done to determine the best performing method for each algorithm. Lastly, the method that had the best post-test results across all the algorithms was compared to the pre-test results of each algorithm.

3.4 Architecture of the Study

Figure 3.4 below shows the binary classification without the data-point approach.

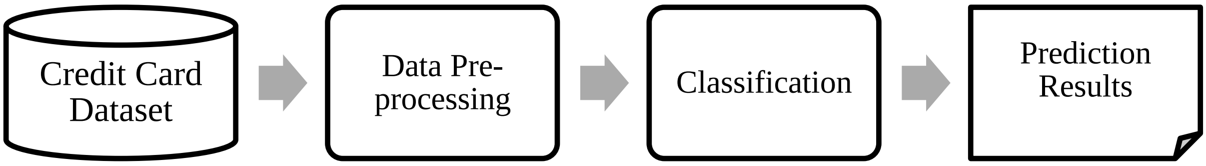

Figure 3.4: Binary classification without the data-point approach

The experimental study was used to examine the ML algorithms for the identification of fraud with an imbalanced CCF dataset (Figure 3.4). The initial pre-test stage experiment was conducted without applying the data-point approach techniques.

Figure 3.5 below shows the binary classification with the application of the data-point approach.

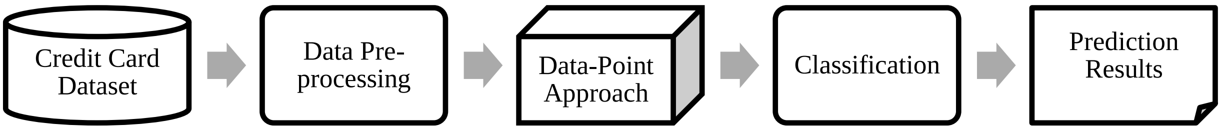

Figure 3.5: Binary classification with the data-point approach

Figure 3.5 includes interventions through an in-depth investigation of the data-point level methods of handling imbalanced datasets. The study proposed two multi-level hybrid data-point approaches combining Feature Selection and Undersampling with Near Miss, and the other combining Feature Selection with the SMOTe methods, which are explained later in the chapter and were examined during the experiment. Both approaches were compared to determine the best data handling method that improved the predictive accuracy of ML credit card fraud identification classifiers through extensive statistical and comparative analysis of the findings.

3.5 Tools

The study made use of the Google Colab notebook and Python programming language. Python provides concise and human readable code, a large repository of frameworks and libraries for implementing ML algorithms and reducing the time for development; hence, selected because it was the most appropriate for the resources and timeframe of this study. The Google Colab notebook runs on cloud servers belonging to Google and was used to write and compile the Python code. Google Colab was chosen because it runs on a Google web browser and to leverage on Google's hardware advancement, which included the Tensor Processing Units (TPUs), and Graphics Processing Units (GPUs).

3.6 Datasets

The study was executed using two credit card datasets from Kaggle, obtained from https://www.kaggle.com/mlg-ulb/creditcardfraud/home. The datasets were chosen because they are excessively imbalanced, labelled, and convenient to the researcher because they are publicly available, and easily accessible, which made them fit for this study. The first dataset is the European Credit Card dataset, which comprised of the transactions belonging to European cardholders. The dataset contained 492 fraud cases out of a sample size of 284 807 transactions. The minority class only made up for 0.173% of all transactions, which were instances recognized as actual cases of fraud in the dataset.

$$\text{Fraud}_{\text{cases}} = \frac{fraud}{instance\ size} * 100 = \frac{492}{284807} * 100 = 0.173\% \quad (3.1)$$

Figure 3.6 below illustrates the graphical representation of the class distribution for the European Credit Card dataset.

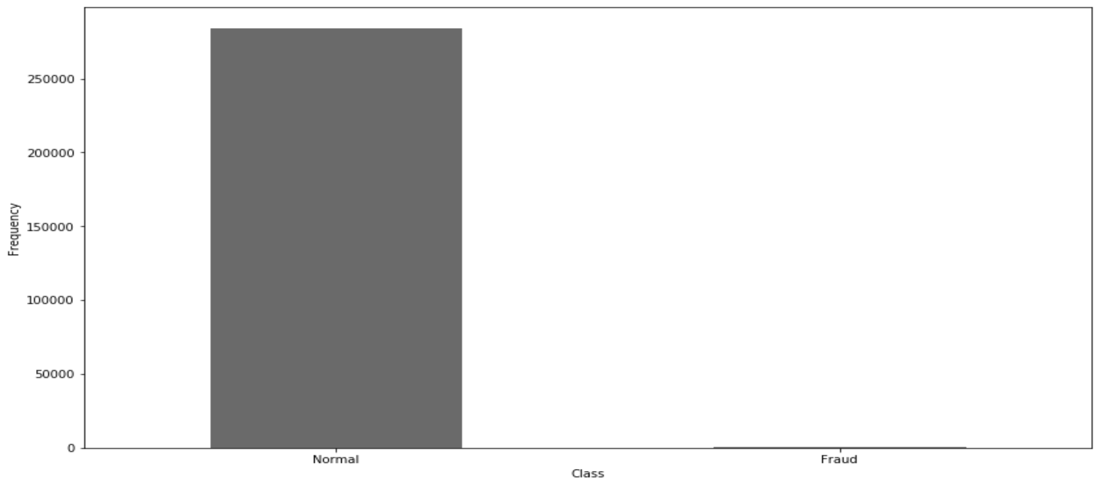

Figure 3.6: European Credit Card dataset class distribution

Figure 3.6 elucidates the bar graph representation of the two classes and the imbalanced class distribution found in the European Credit Card dataset. The x-axes is for the 'Class' and the y-axes is for the ‘Frequency’ of occurrence of each Class. The short bar, which is hardly visible, shows the 0.173% fraudulent transactions, which is the minority class. The data contains 31 features. Features V1, V2, up to V28 were the principal components gathered from the conversion known as Principal Component Analysis (PCA), which was done because of confidentiality legislation. The features that were left out of the PCA conversion were “Time”, “Amount”, and “Class”. The feature “Class” stores a numeric value, which can either be 0 to indicate a transactions that is normal (legitimate) or 1 to indicate fraudulent transactions.

Figure 3.7 below shows the graphical representation of the time of transaction versus US$ amount by Class for the European Credit Card dataset.

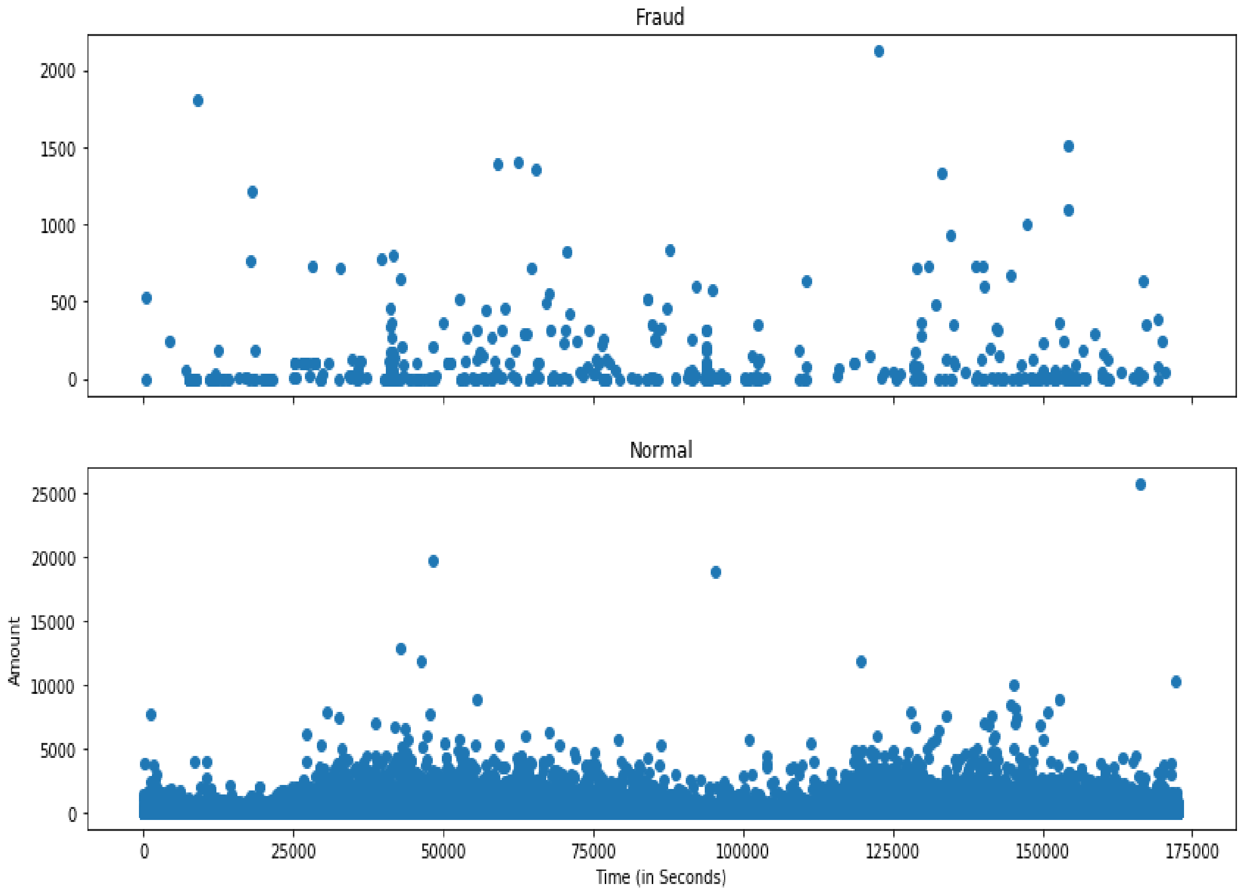

Figure 3.7: Time of transaction versus US$ amount by Class

Figure 3.7 illustrates the graphical depiction of the amount in US Dollars that the transactions found to fraudulent accounted versus those that were legitimate in the dataset. The amounts were within the range of the transaction that are legitimate (normal), which causes difficulties when considering amount as an indicator to distinguish between the classes.

Figure 3.8 below illustrates the graphical representation of the amount per transaction by class for the European Credit Card dataset.

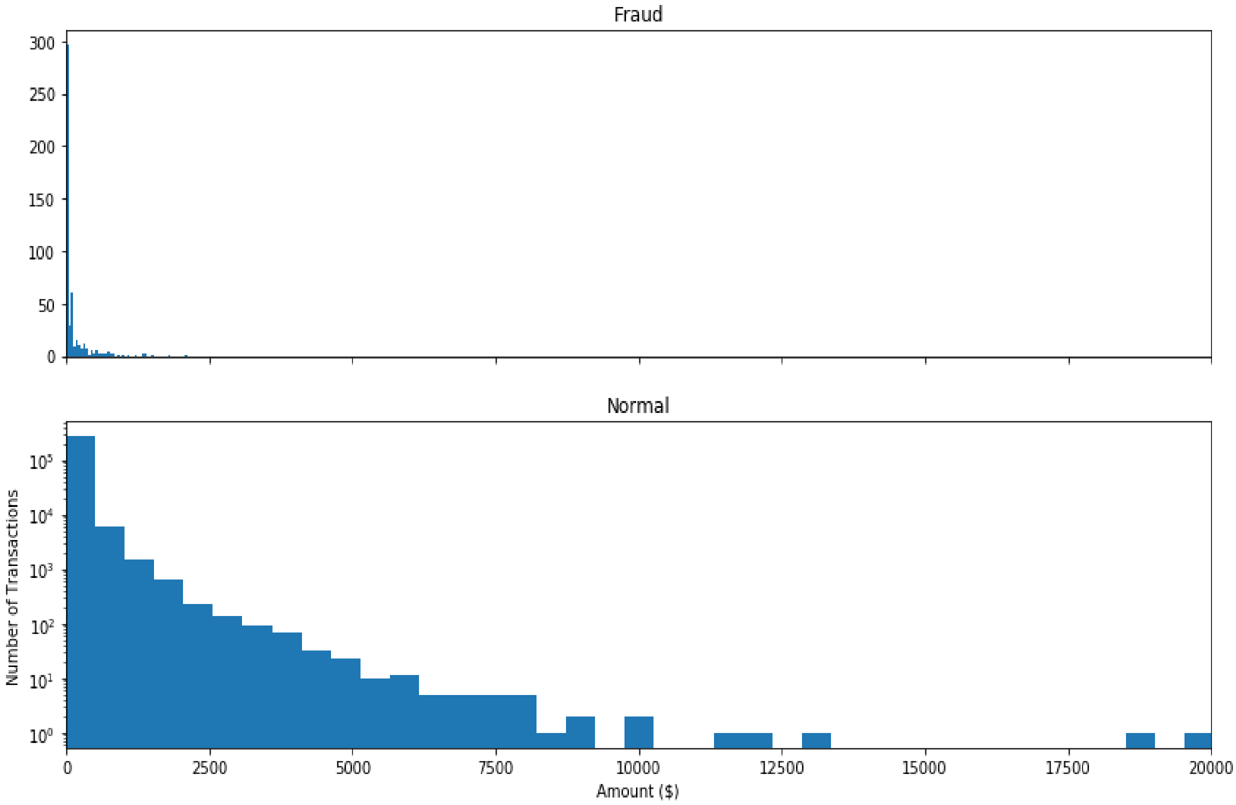

Figure 3.8: Amount per transaction by Class

The plot in Figure 3.8 illustrates a graphical depiction to compare the number of transaction that occur per specific amount for both fraud and normal class on the European Credit Card dataset. The fraudulent transaction amounts are within the range of the legitimate transactions and account for a very small number. The figure 3.8 shows that majority of transaction are within the amount of US$(0 to 2500) for both classes, which makes them difficult to distinguish from the normal transactions. An interesting observation is that there are no fraudulent transactions that exceed this amount of US$2500, therefore increasing the complexity to distinguish the two classes.

The second dataset is called the UCI Credit Card dataset, which comprises of data about Taiwan credit card clients. The data includes information about payment history, credit data, bill statements, and demographic factors of credit card clients from April to September in 2005. The dataset was imbalanced and contained 30 000 instances, of which 6 636 were positive credit card cases.

Figure 3.9 below shows the graphical representation of the class distribution for the UCI Credit Card dataset.

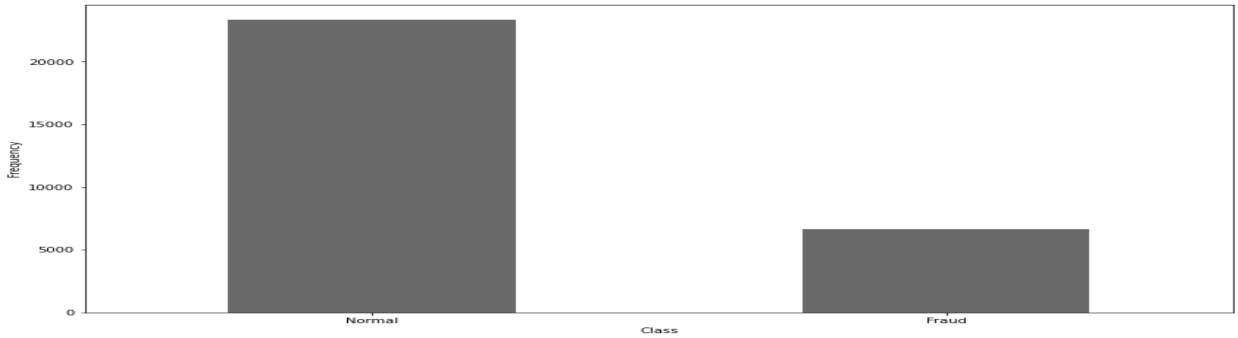

Figure 3.9: UCI Credit Card Dataset Class Distribution

Figure 3.9 shows the UCI Credit Card dataset class distribution. The short bar is the minority class that caters for 22.12% of the dataset and represents the positive credit card cases.

$$\mathbf{Positive_{cases}} = \frac{\boldsymbol{fraud}}{\boldsymbol{instance\ size}} * \mathbf{100} = \frac{\mathbf{6636}}{\mathbf{30000}} * \mathbf{100} = \mathbf{22.12}\% \qquad (3.2)$$

The longer bar shows the normal transactions, which is the majority class on both datasets. The imbalance ratio of both dataset was maintained throughout all the subsets of the original dataset that were taken for the experiments conducted before any of the data-point techniques were applied. The UCI Credit Card dataset has 24 numeric attributes, which makes the dataset suitable for a classification problem. An attribute called '*default.payment.next.month*' contained the values of either 0 or 1. The '0' represents a normal (legitimate) case and the value of '1' represents the fraudulent case. Validation was performed on each dataset to make sure that missing columns or null values are found and fixed.

Thereafter, to gain insight and a visual illustration of the data to make it more information, an exploratory data analysis was conducted. During the data preparation step, the first step of the multi-level hybrid data-point approached was to use Feature Selection to determine the dependent and independent features. The goal was to find those independent features that influenced the prediction outcome. The X variable was used to store the independent features. The Y variable was used to store the dependent feature, which was the column that indicated the class of the transaction. The Y variable is used by the ML classifier to predict the X variable. Once the X and Y variables were created, the train-test-split method, which was inherited from the sklearn library was called to split the dataset into training and testing subsets. The classifier was trained using the training subset and then evaluated using the testing subset. The method call accepts three mandatory parameters, namely the X and Y variable, and the test-size, which indicates the percentage of samples to be allocated to the testing subset. A test-size of (0.30) was used on both datasets, which means that from the original dataset, 70% of the total sample size was allocated to the training subset and 30% was allocated to the testing subset. This is referred to as a 70:30 split ratio, which was chosen because it is a recommended benchmark ratio that is widely adopted to divide the dataset to achieve optimal results (Favieiro and Balbinot 2019).

3.7 Cross Validation

K-Fold Cross Validation (CV) is the practice of partitioning a particular dataset into a ***K*** number of folds, where at some point, each fold will be used as a testing subset (Shao 1993). K-Fold CV is commonly used to achieve an unbiased classification with a small balanced dataset that does not have enough examples to train and test the model (Kubat and Matwin 1997). K-Fold CV was not used for this study because our study used big datasets that are highly imbalanced. According to Zhang (1992), the K-Fold CV is not suitable for evaluating classifiers that are imbalanced because the dataset is divided into K-Fold's distribution with the same data allocation probability.

Therefore, the performance of CV is fine for a dataset with a class distribution that is balanced. However, in the case of a highly imbalanced class distribution, there is a risk of having folds with minimal or zero instances of the minority class. Consequently, there will be certain evaluations of the model that will be misleading, since the model will be exposed and predicting only one class.

3.8 Performance Metrics

The measurement metrics chosen to assess the performance was the standard accuracy score and the following sensitivity metrics; the F1-score, precision, Average Precision (AP), recall, and confusion matrix. An in-depth review and analysis of the accuracy was conducted using these performance evaluation metrics. The measurement of the ability to classify classes that are positive is called precision (Rekha, Tyagi, and Krishna 2019). Precision can be represented mathematically as (TP / (TP + FP)). Recall refer to the model's ability to correctly classify the

TP (Feng, Huang, and Bao 2019). Recall = TP / (TP + FN). According to Askari and Hussain (2017), evaluating both precision and recall is useful to determine the model's predictive accuracy in a binary classification. According to Abdulsalam, Skillicorn, and Martin (2011), a Precision-Recall (P-R) curve is a plot for different thresholds of the precision (y-axis) and the recall (x-axis). The F1 score is the measurement of the test accuracy, which is computed by incorporating both the precision and the recall scores. A confusion matrix is a table used to map the outcome of the classification for the positive and negative class (Khosravi and Jouybari-Moghaddam 2019). The confusion matrix table provides a mapping of the rate of TN, TP, FN, and FP (Guo *et al.* 2019). The confusion matrix table is useful to quantify the number of misclassification instances for both the negative and positive classes.

To determine whether there was any improvement in the ability to detect the positive class, the following formulas were used to calculate the Average Improvement (AI) for the Precision, Recall, and F_1 score for both datasets. For the purpose of the formula, D1 was used to represent the European Credit Card dataset and D2 for the UCI Credit Card dataset:

$$\boldsymbol{Sum_{differences}} = (\boldsymbol{D1\_score_{after}} - \boldsymbol{D1\_score_{before}}) + (\boldsymbol{D2\_score_{after}} - \boldsymbol{D2\_score_{before}}) \tag{3.3}$$

$$\mathbf{AI_{score}} = \frac{\boldsymbol{Sum_{differences}}}{\mathbf{2}} * \mathbf{100} = \%\, \boldsymbol{value} \tag{3.4}$$

Sensitivity metrics are the best way to evaluate an imbalanced dataset, whereas the standard accuracy score is rely with a balanced dataset. In this study, both the standard and sensitivity metrics were used to measure the performance with imbalanced and balanced datasets. The goal was to generate a detailed classification report of the classifiers from both the pre-test with an imbalanced datasets and the post-test with a balanced datasets. The classification report was used to conduct a fair comparison and an in-depth analysis of the performance.

3.9 Classification Algorithms

Machine Learning (ML) algorithms are widely used for data classification. This study proposes the use of the RF algorithm to build a credit card fraud detection solution. The RF algorithm is designed to builds a classifier that uses supervised learning for classification and regression. The RF is categorised as an ensemble technique and it was chosen because it is made up of many decision trees that improve the classification; during the experiment, the forest consisted of 600 trees. Jiang, Lu, and Xia (2016) stated that, a class prediction is produced by each trees and the class containing more occurrences becomes the classifier's final prediction. The data-point approach was combined with RF to achieve the ideal fraud detection capabilities. The use of RF is supported by literature; Bader-El-Den, Teitei, and Perry (2019) introduced a new method based on ensemble learning for handling the problem of class imbalance. They proposed using a RF algorithm that uses the nearest neighbour algorithm to categorise the areas that are critical in a particular dataset.

Their motivation of the study came from the idea that instead of performing Oversampling on the data level, it will be done on the algorithm level. This meant that the minority class instances were not going to be increased on the datasets as per usual. Oversampling would take place in the classification ensemble, where the classifiers that belong to the minority class would be increased in numbers. The findings illustrate that the introduced method was effective in handling the issue of class imbalance. Patil, Nemade, and Kumar (2018) discussed an analytical framework based on Big Data to process huge amounts of data and employed numerous ML fraud identification algorithms and monitored how they performed in real-time detecting detection of fraud against a benchmark dataset.

In the construction of the analytical model, a dataset with almost 1 000 German credit card transactions containing fraudulent records was used, which was had a mixture of numeric and categorical attributes. Their study ran three models on the dataset and evaluated the precision of the analytical model using the confusion matrix. The study discovered that the RF model performance was better than the performance that of the DT and LR, in terms of precision, recall, and accuracy. The following ML algorithms were selected in the study: The SVM, LR, DT and RF. The classifiers were built using the four selected ML algorithms. A supervised learning approach was selected to explore and examine the predictive abilities of the classifiers.

Literature suggested the LR, DT, and SVMs to be the leading detection algorithms, hence the study used these three mentioned methods to build, which were compared to the proposed RF (Albashrawi 2016; Patil *et al.* 2018). The train-test-predict approach was used to train and evaluate the classifiers for detecting fraud, which were built using the selected ML algorithms. The four ML algorithms were chosen because they are suitable for binary classification problems and align with the proposed strategy to deal with imbalanced classes (Sadgali *et al.* 2018).

Table 3.1 shows the instantiation and technical specifications of each classification algorithm used for the experiments.

Table 3.1: List of the four algorithms and classifiers used during the experiment

ALGORITHM	CLASSIFIER
SVM	SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)
LR	LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=100, multi_class='auto', n_jobs=None, penalty='l2', random_state=None, solver='lbfgs', tol=0.0001, verbose=0, warm_start=False)
DT	DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort='deprecated', random_state=None, splitter='best')
RF	RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=600, n_jobs=None, oob_score=False, random_state=None, verbose=0, warm_start=False)

Table 3.1 presents classifies that were build using SVM, DT, LR, and RF. The classifier are configured by passing attributes such as random_state, which accepts an integer value to controls the randomness of the estimator. This study proposed the RF classifier, which is a meta estimator made up of DT nodes split into multiple levels and use averaging to controls over-fitting and improve the predictive accuracy (Li, He, and Li 2019). In each split, random permutation of the features is conducted always. To decrease the consumption of the memory,

the size and complexity of the trees have to be controlled, which is achieved by manipulating the parameters.

3.10 The Data-Point Approach

An experimental study exploring the data-point approach was conducted to solve the problem of class imbalance. The study investigated the two proposed multi-level hybrid data-point approaches. The first approach was the combination of Feature Selection and Undersampling with Near Miss (Elhassan *et al.* 2017). The second approach was the combination of Feature Selection and SMOTe (Abdoh, Abo, and Maghraby 2018). A detailed comparison of these two approaches was performed to determine the method that would yield the best results on the credit card datasets.

3.10.1 Feature Selection

Feature Selection is a procedure that can be done manually or automatically, which consists of identifying and selecting the independent features that influence the dependent variable to be predicted (Zhang *et al.* 2014). The study used Feature Selection as a step following the data pre-processing stage before the learning occurred. The snippet of code below was used to generate the Correlation Heat Map.

```
## Correlation
import seaborn as sns
#get correlations of each features in dataset
corrmat = data1.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(20,20))
#plot heat map
g=sns.heatmap(data[top_corr_features].corr(),annot=True,cmap="RdYlGn")
```

A heat map is a graphical visualisation of data, where the values of data are denoted using colours. The colours are used to group features that belong within a similar rank in terms of importance. The different colours signify different levels and inform the reader about the significance of each value. The heat map is a useful tool to gain insights on the key attributes through the use of colours to highlight important data-points of large voluminous data.

Figure 3.10 shows the Correlation Matrix with Heat Map of the 32 features of the European Credit Card dataset.

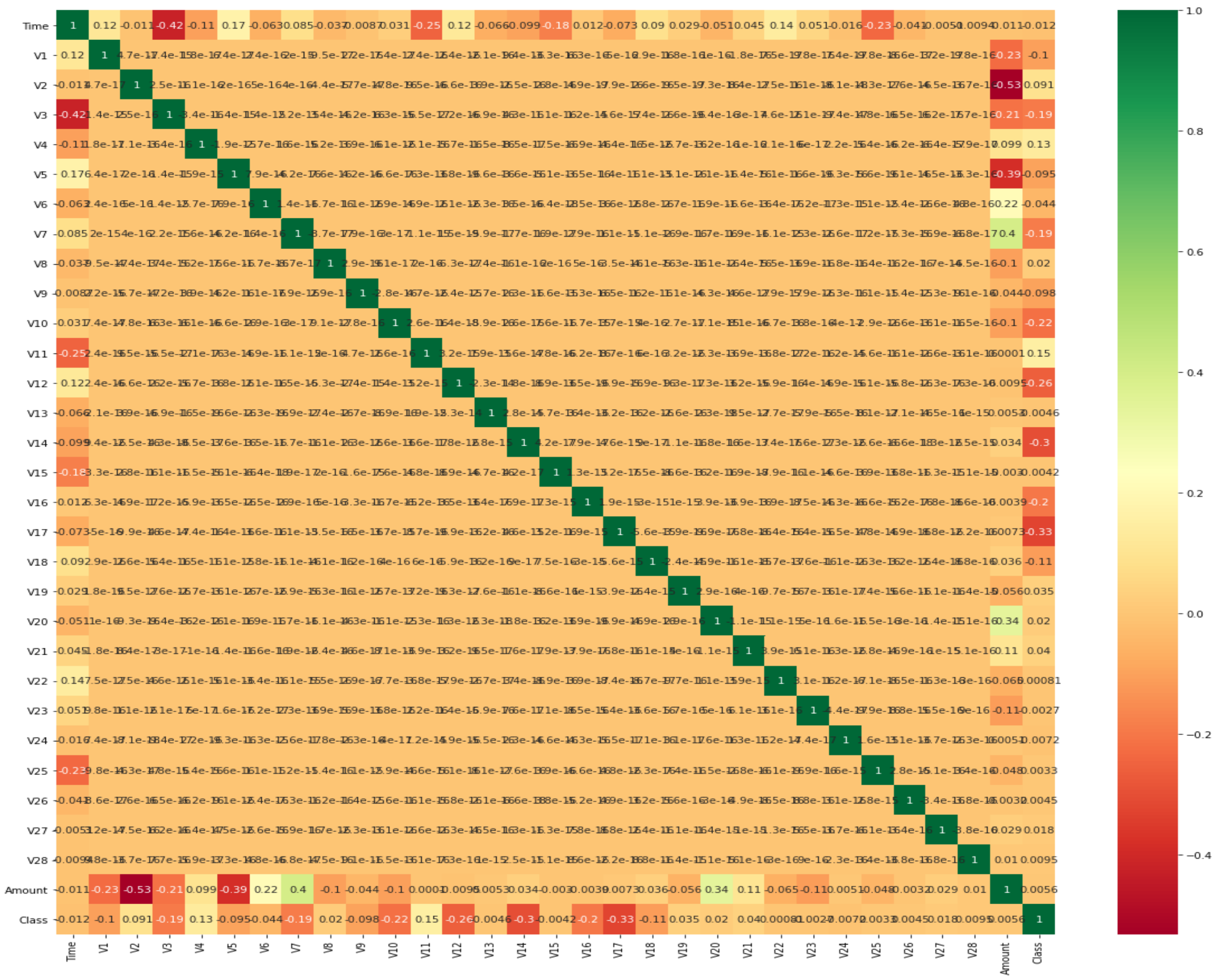

Figure 3.10: European Credit Card Dataset Correlation Matrix with Heat Map

Figure 3.10 shows the Correlation Heat Map that was created and analysed during Feature Selection to determine the important features. The last part of the Feature Selection was to create the independent and dependent features, and selecting the features that contribute most to the prediction variable. Then, the transformed output was used to train the classifiers with only the features that are relevant.

Figure 3.11 shows the Correlation Matrix with Heat Map of the 28 features of the UCI Credit Card dataset.

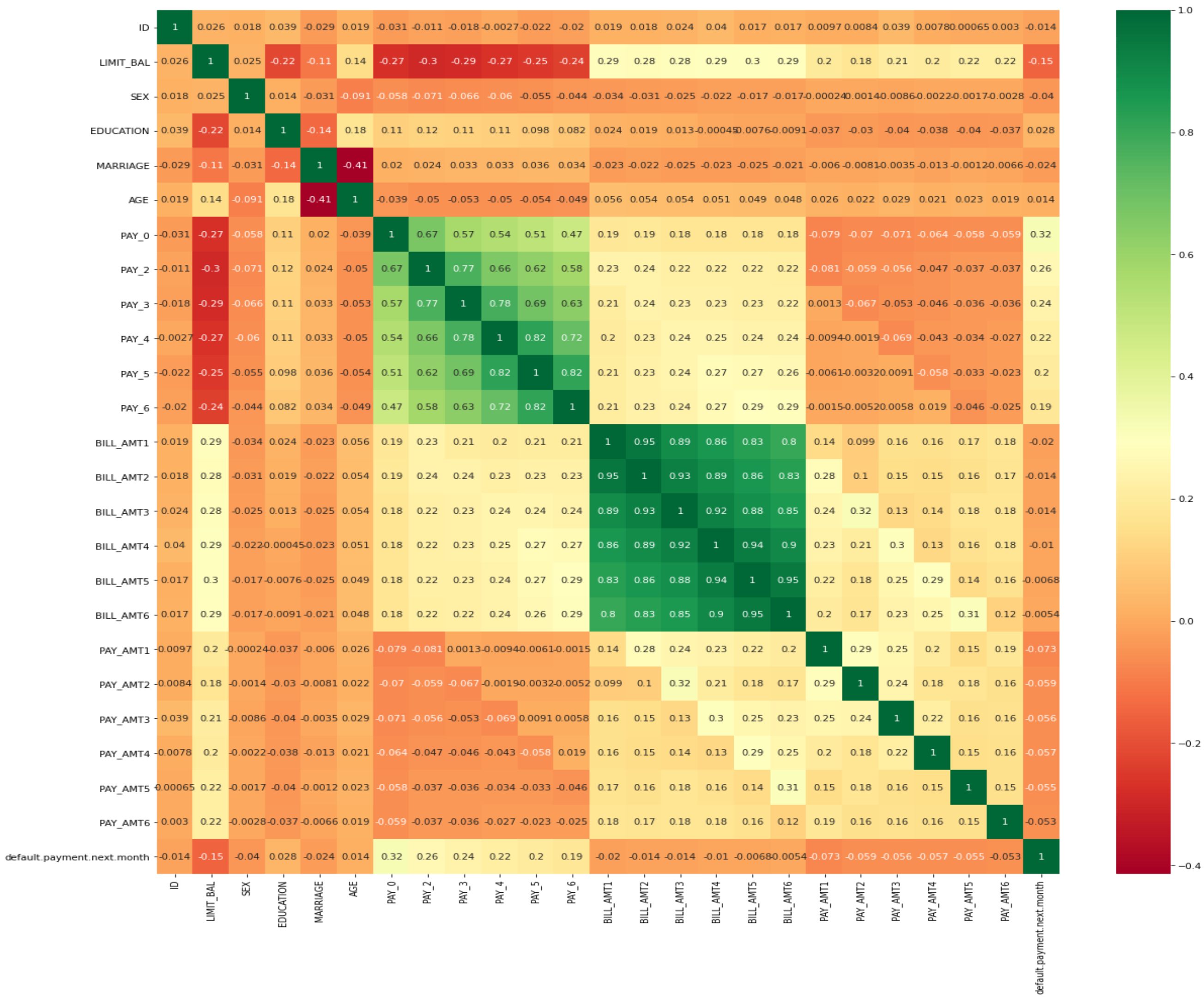

Figure 3.11: European Credit Card Dataset Correlation Matrix with Heat Map

The green denotes positive, red means negative. The more saturated the colour, the larger the correlation magnitude. The Heat Map is used to evaluate the importance of the feature and to reveal the most relevant features. While the weak correlations visually disappear, the most visible areas represent high correlation.

3.10.2 Undersampling

Undersampling refers to the process of balancing the distribution of the classes for a classification dataset that has an unbalanced distribution (Ng *et al.*, 2015). Undersampling gets rid of examples that belong to the dominating class, to create an even class distribution. The technique can decrease the skew from a 1:100 to a 1:10, 1:2, or even a 1:1 class distribution. The experiment used an imbalanced-learn library, to call a class to implement Undersampling based on the Near Miss technique. The Near Miss technique was manipulated by passing parameters that are in order to meet the desired results.

There are three versions of the Near Miss method, namely:

1. NearMiss-1 where the samples of the positive class are selected when the closest samples' average distance of the classes that are negative is the smallest.
2. NearMiss-2 where the samples of the positive are selected when the farthest samples' average distance of the classes that are negative is the smallest.
3. NearMiss-3 which is a 2-step algorithm. In step-1, for each sample that is negative, the algorithm identifies and stores their nearest neighbours. In step-2, the only samples which are selected that are positive are those that have the largest average distance to the nearest neighbours.

When Undersampling a particular class, the presence of noise can alter NearMiss-1. In fact, it will obscure targeted class sample since NearMiss-1 focuses on the nearest samples. Nonetheless, under normal circumstances, the samples that will be selected are those next to the boundaries (Dong, Du, and Zhang 2015). NearMiss-2 focuses on farthest samples, which helps prevent the effect of noise. The existence of noise can be rehabilitated by mainly sampling in the marginal outliers presence. NearMiss-3 is also less impacted by noise because of the first step sample selection.

The chosen variation for this study was the NearMiss-2 version after multiple iterations were executed using all three different versions to select the most suitable version for the credit card dataset. A uniform experiment was conducted on both datasets to ensure a fair cross comparison.

Table 3.2 is a snippet of parameters that were used to instantiate the Near Miss technique.

Table 3.2: Near Miss method call parameters

Parameter	Value
Sampling Strategy	Auto
Return Indices	False
Random State	None
Version	2
N Neighbours	3
N Neighbours ver3	3
N Jobs	1
Ratio	None

A list of all the parameters and their associated values, which were passed when instantiating the Near Miss method using an API call on the imbalanced-learn library is shown in Table 3.2. Performance of the Near Miss method was optimised by changing the version parameter to determine the best version to use during the experiment.

Figure 3.12 below shows the graphical representation of re-sampling with Undersampling.

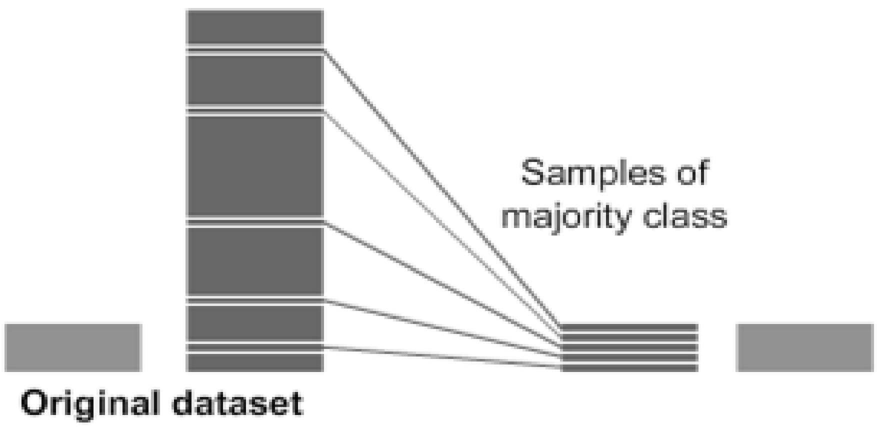

Figure 3.12: Resampling with Undersampling (Raj 2019)

A general view of how an original dataset that was highly imbalanced and was then re-sampled using Undersampling is shown in Figure 3.8. The majority class was reduced by removing instances to balance the dataset by 1:1 ratio.

3.10.3 Oversampling

Oversampling is a data-point technique designed to fix the data imbalance problem, where the minority samples numbers are increased. Oversampling works by replicating existing samples to generate new instances that will balance the data (Yan *et al.* 2019). The intention with Oversampling is to preserve valuable data of the majority class by sampling up the minority class and evade losing samples. Repetition, bootstrapping, and SMOTe are some of the most popular methods used to sample up the minority class (Tarawneh *et al.* 2020). This study adopted the SMOTe method based on the nearest neighbours that are arbitrated by the Euclidean Distance between data-points within a feature space because of it robustness and easy application.

According to Tarawneh *et al.* (2020), a percentage to indicate the amount of synthetic samples to be generated is passed as a parameter, which is a multiple of 100 always. To generate a copy for each instance of the minority samples, the percentage must be set to 100, thereafter the total count of the minority class will double. For example, a dataset with 492 minority cases will be doubled to 984 cases. To triple the number of minority samples, the percentage must be set to 200. Therefore, the percentage parameter helps control the replication of samples.

In SMOTe,

- The KNN for each minority class are identified that are belonging to the same class group.
- The KNN feature vectors are compared with the considered instance's feature vector to find dissimilarity. So, the K number of divergent vectors are attained.
- A random number within the range of 0 and 1 is used to multiply each of the K divergent vectors.
- Finally, the divergent vectors and the product of the random numbers, are added to the original minority instance feature vector at each repetition.

The below equation is used to determine the value of K:

$$(\mathrm{SMOTe}\ \%)\ /100 = \mathrm{K} \tag{3.5}$$

A library from *imblearn* was imported to inherit the SMOTe implementation to reduce the development time (Imbalanced-learn 2020). The SMOTe method was called and configured by passing parameters. The researcher used inheritance to reduce programming time by reusing existing code, which allowed more flexibility and increased existing code.

Table 3.3 shows the parameters and associated values selected for the experiments.

Table 3.3: SMOTe method call parameters

Parameter	Value
Sampling Strategy	Auto
Random State	None
K Neighbours	5
M Neighbours	Deprecated
Out Step	Deprecated
Kind	Deprecated
SVM Estimator	Deprecated
N Jobs	1
Ratio	None

Table 3.3 provides a list of all the parameters and their associated values, which were passed when instantiating the SMOTe method using an API call on the imbalanced-learn library. During the experiment, numerous combinations of the parameters were explored to discover a combination shown in Table 3.3 of parameters identified to have produced the desired outcome.

Figure 3.13 shows the graphical representation of re-sampling with Oversampling.

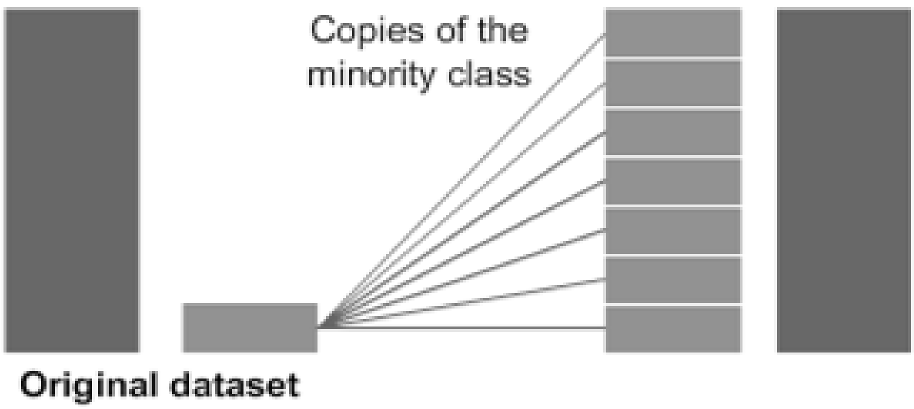

Figure 3.13: Re-sampling with Oversampling (Raj 2019)

Figure 3.13 shows an overview of how an original dataset that was highly imbalanced was re-sampled using Oversampling. The minority class was increased by copying and making the duplicating minority class instances balance the dataset by a 1:1 ratio. SMOTe was introduced as an advancement to Random Oversampling (ROS), which optimised the oversampling of minority classes, resulting in sub-optimal performance (Agrawal, Viktor, and Paquet 2015; Jiang, Lu, and Xia 2016). However, Douzas, Bacao, and Last (2018) indicated that excessive Oversampling can cause overfitting with a dataset that is severely imbalanced. To prevent this problem, the data-point approach was applied using the SMOTe technique to render new instances by interpolating existing samples in the dataset. The approach was chosen because it avoids the creation of random samples and overfitting by eliminating the within-class imbalances and between-class imbalances.

3.11 Chapter Summary

Performance of ML techniques are highly dependent on the data because they learn from past cases. In real life, the credit card dataset is highly imbalanced; therefore, these techniques fail to perform as expected in real-life situations. The main contribution is that the study conducted an in-depth investigation of the data-point level approach to solve the misclassification problem in fraud detection using ML and determined a novel method that is more efficient than other options in handling imbalanced credit card dataset.

The study introduced two multi-level hybrid approaches that were proposed to be used with the RF classifier, which were novel to the area of CCF detection. The first approach was the combination of Feature Selection and Undersampling with Near Miss. The second approach was the combination of Feature Selection and SMOTe. These approaches addressed the issues of class imbalance and misclassification by identifying the key feature, based on feature importance and then used the Near Miss and SMOTe technique to resample the dataset respectively.

CHAPTER FOUR: PRESENTATION OF RESULTS AND DISCUSSION

4.1 Introduction

This chapter is structured as follows; the pre-treatment-test results presentation, presentation of post-treatment-test results, and chapter summary. The pre-treatment-test results section presents the outcome before using the data-point approach, following which is a detailed classification report, the confusion matrix, and the precision-recall curves without the data-point approach. The post-treatment-test results section is organised into Undersampling and Oversampling subsections. Undersampling subsection presents the classification report, the confusion matrix with Near Miss, and the precision-recall curves with Near Miss. Oversampling subsection presents the classification report, the confusion matrix with SMOTe, and the precision-recall curves with SMOTe.

4.2 Pre-Treatment-Test Results

The objective for the pre-treatment-test was to examine the performance of ML algorithms for the identification of fraud with a credit card dataset that was imbalanced. The purpose of this test was to gather evidence to prove that imbalanced datasets produce biased and inaccurate predictions, which have a high misclassification rate (Wang *et al.* 2012). The original datasets were split using a 70:30 ratio to create a training sample and a testing sample. The 70:30 split ratio was selected because it is a widely used benchmark split ratio in ML studies (Zhu *et al.* 2018). The testing dataset was fed into the ML classifiers using each of the four algorithms, namely, SVM, LR, DT, and RF to train and test the predictive performance of the classifiers.

The performance metrics used to assess the accuracy of the performance are recall, F1-score, precision, Average Precision (AP) and confusion matrix. According to Tarawneh *et al.* (2020), the most appropriate performance metrics to use with highly imbalanced data are precision, recall, and the F1-score.

4.2.1 Classification Report

Following the conclusion of the initial experiment, the y-test variable was compared to the prediction results to generate a classification report. A classification report is a presentation of the scores achieved for each performance metric on the prediction of the negative class and positive class. The tables that present results on this chapter use P for Precision, C for Class, ALG for algorithm, R for recall, AC for accuracy, AP for the Average Precision, and F1 for F1-score. The Negative column represents legitimate cases and the Positive column represents the fraudulent cases. The 0.00 for 0% represents the lowest possible value and the maximum value attainable is represented as 1.00 for 100%. Table 4.1 shows the European Credit Card dataset results of the classification for the four algorithms before using the data-point approach.

Table 4.1: Classification of the imbalanced European Credit Card dataset

Classifier	Metric	Negative	Positive
SVM	Precision	1.00	0.00
	Recall	1.00	0.00
	F1-score	1.00	0.00
	Accuracy	1.00	
	Average Precision	0.00	
LR	Precision	1.00	0.57
	Recall	1.00	0.47
	F1-score	1.00	0.52
	Accuracy	1.00	
	Average Precision	0.48	
DT	Precision	1.00	0.50
	Recall	1.00	0.47
	F1-score	1.00	0.48
	Accuracy	1.00	
	Average Precision	0.24	
RF	Precision	1.00	0.90
	Recall	1.00	0.53
	F1-score	1.00	0.67
	Accuracy	1.00	
	Average Precision	0.66	

The testing dataset for the European Credit Card dataset had a sample size of 8 545 cases and there were 17 positive cases out of the total sample size of 8 545 cases. According to the classification report, there was 100% accuracy from all the classifiers with the European Credit Card dataset, which is highly misleading. Looking only at the accuracy score with imbalanced datasets does not reflect the true outcome of the classification. On the European Credit Card dataset classification, we can observe that for the SVM classifier, there was high bias towards the negative classes. All 8 545 cases were flagged as legitimate transactions; this is due to the fact that there were only 17 fraudulent transactions in the testing dataset.

The LR performed better than the SVM, the classifier was biased but observing the recall, precision, and F1 score shows that some positive classes were able to be classified. The F1 score verified that the test was not accurate. The report does not tell us if the positive classes identified were TP or FP, even though the Recall score gives an indication that there was a great deal of misclassification. The most common misclassification problems are FP and FN, which means even though the classifier has 100% accuracy and has the ability to predict both positive and negatives classes; it fails to produce a successful prediction. A similar observation is seen in the DT and RF, although the RF performed much better compared to the other three classifiers.

Table 4.2 shows the UCI Credit Card dataset results of the classification results for the four algorithms before using the data-point approach.

Table 4.2: Classification of the imbalanced UCI Credit Card dataset

Classifier	Metric	Negative	Positive
SVM	Precision	0.78	0.00
	Recall	1.00	0.00
	F1-score	0.87	0.00
	Accuracy	0.78	
	Average Precision	0.22	
LR	Precision	0.78	1.00
	Recall	1.00	0.01
	F1-score	0.87	0.02
	Accuracy	0.78	
	Average Precision	0.36	
DT	Precision	0.82	0.37
	Recall	0.82	0.37
	F1-score	0.82	0.37
	Accuracy	0.72	
	Average Precision	0.28	
RF	Precision	0.84	0.63
	Recall	0.94	0.36
	F1-score	0.88	0.46
	Accuracy	0.81	
	Average Precision	0.37	

The UCI Credit Card testing dataset contained a sample size of 900 cases. The classification report on the UCI Credit Card dataset shows similar results seen on the European Credit Card dataset. The SVM classifier was biased towards the negative class. The UCI Credit Card testing dataset has a slightly lower imbalance ratio; there were 204 positive cases out of the total sample size of 900 cases. The accuracy recorded was 78%, which is far less than ideal for a binary classification solution. Therefore, without even considering the bias and misclassification problem, the accuracy score alone shows that the SVM classifier is not consistent across multiple datasets. The LR had an accuracy score of 78%, which is the same as the SVM classifier. The major difference is the positive class precision score, which was 100% for the LR, implying that the classifier was able to predict all the positive classes. Therefore, we look at the Recall score of 1% and based on this figure, we can conclude that the classifier was poor when the actual outcome was positive, which means the number of FP and FN was high. Based on the precision score, we can conclude that the classifier is biased and the prediction was unable to eliminate FP and FN. The DT was the least effective in terms of the accuracy score, which was 72%. The precision, recall, and F1 score were all 37% for the positive class. The RF continued to lead with an accuracy score of 81%. The positive class precision was 63%. The Recall and F1 score show that nearly half of the predictions were FP. The initial finding reveals that there was a bias towards predicting the majority class, representing normal/ legitimate transactions.

4.2.2 The Confusion Matrix

The confusion matric table was used to provide the results for each algorithm on the original dataset, and the results after using Undersampling and Oversampling. The total sample size used during testing is the sum of TN, FN, TP, and FP as per the blueprint of the confusion matrix. The confusion matrix also helps us to understand if the classification was biased.

Table 4.3 below shows the confusion matrix table(s) blueprint. The blueprint was used to present the classification results. The zero (0) represents the negative class and the one (1) represents the positive class.

Table 4.3: Confusion Matrix Table(s) Blueprint

	Predicted 0	**Predicted 1**
Actual 0	TN	FP
Actual 1	FN	TP

The top heading of Table 4.3 represents the outcome of the cases from the classification and the left side heading represents the actual class of the cases. The four possible classifications for cases are TN, FN, TP, and FP.

4.2.3 The Confusion Matrix without the data-point approach

Table 4.4 below contains the SVM confusion matrix results before applying the data-point approach.

Table 4.4: Confusion Matrix of the SVM classifier

European Credit Card Dataset			UCI Credit Card Dataset		
SVM	**Predicted 0**	**Predicted 1**	**SVM**	**Predicted 0**	**Predicted 1**
Actual 0	8 528	0	**Actual 0**	696	0
Actual 1	17	0	**Actual 1**	204	0

The confusion matrix revealed a poor performance by the SVM classifier. The European Credit Card dataset contained a sample size of 8 545 transactions. The positive class only accounted for 17 fraudulent cases, which were all falsely classified as legitimate cases. According to the confusion matrix, there were zero FP, zero TP, 17 FN out of the 8 545 cases predicted as legitimate, and 8 528 TN. The predictive accuracy was 100% and a 0% misclassification rate for the negative class. The predictive accuracy was 0% and a misclassification rate of 100% for the positive class. The overall calculated predictive accuracy was 50% and a misclassification rate of 50% for the European Credit Card dataset. The UCI Credit Card dataset contained a testing sample of 900 cases.

There was a class distribution of 698 negatives cases and 202 positive cases. Similarly, the SVM classifier failed to predict any positive cases accurately with the UCI Credit Card dataset. All the cases were predicted to be negative even though there were 202 positive cases in the testing samples. The confusion matrix revealed that there were 696 TN, 202 FN, zero TP, and zero FP. Similarly, the UCI Credit Card dataset predictive accuracy was 100% with a 0% misclassification rate for the negative class. A misclassification rate of 0% indicates that all the negative cases were correctly classified. The predictive accuracy was 0%, with a misclassification rate of 100% for the positive class.

A misclassification rate of 100% indicates that there were no TP but instead, the prediction produced more FN. Therefore, in a fraud detection system, all the fraudulent transactions would be flagged as legitimate transactions. The misclassification rate is a great indicator of the impact of the imbalanced dataset on the FN and FP. The overall calculated predictive accuracy was 50% and a misclassification rate of 50% for the European Credit Card dataset. Therefore, the average predictive accuracy and misclassification rate across both datasets was 50% for the SVM classifier. According to these results, the performance was very poor and the classification was 100% biased towards the majority class for both datasets.

Table 4.5 contains the LR confusion matrix results with imbalanced datasets.

Table 4.5: Confusion matrix of the Logistic Regression classifier

European Credit Card Dataset			**UCI Credit Card Dataset**		
LR	**Predicted 0**	**Predicted 1**	**LR**	**Predicted 0**	**Predicted 1**
Actual 0	8 526	2	**Actual 0**	696	0
Actual 1	9	8	**Actual 1**	204	0

The results in Table 4.5 show that, the classifier was both biased and highly inaccurate. The results for the UCI Credit Card dataset show that, out of a testing sample of 900 cases, the LR classifier was able predict all of negative cases correctly but failed to correctly predict any of the positive cases. The testing set contained 696 negative cases, whereby the total number of cases classified as negative was 900, which sets the predictive accuracy at 100% for negative cases, biased by 204 cases. However, the confusion matrix revealed that 204 out of the 900 predicted cases were FN, which makes the true predictive accuracy of negative cases 100% but also results in a misclassification rate of 100% for the positive class.

The LR failed to classify any positive cases. The confusion matrix revealed that there were zero FP and zero TP. The European Credit Card dataset had a testing sample of 8 545 transactions, 99.8% of negative and 47% of positive cases were correctly classified respectively. The LR predicted 8 535 negative cases even though there were 8 528 instances of the negative class. However, out of the 8 535, there were 8 526 TP, which means that there was some confusion in classifying negative and positive cases. The confusion matrix revealed that there were 9 FN and 8 FP. The confusion matrix showed that a total of 10 positive cases were predicted, but there was a lot of misclassification for both classes. There were only 8 TP out of a possible of 17 instances of the positive class. The true predictive accuracy for positive cases was 47%, instead of the 58.8% if we consider the total number of instances classified as positive. The LR was slightly better than the SVM classifier, but both were highly biased on the two datasets.

Table 4.6 contains the DT confusion matrix results before Undersampling.

Table 4.6: Confusion matrix of the Decision Tree classifier

European Credit Card Dataset			**UCI Credit Card Dataset**		
DT	**Predicted 0**	**Predicted 1**	**DT**	**Predicted 0**	**Predicted 1**
Actual 0	8 520	8	**Actual 0**	557	139
Actual 1	9	8	**Actual 1**	124	80

The confusion matrix of the DT revealed that the classifier was less biased but highly inaccurate. The UCI dataset testing sample contained 696 negative cases and 204 positive cases. The total number of cases predicted as negative were 681 out of 696 for the negative class. Looking at the prediction, we can assume the model was accurate. However, 124 out of the 681 cases classified as negative were FN. Therefore, the true predictive accuracy was 80% and the misclassification rate for the negative class was 20%. The classifier predicted 219 positive cases, which meant the classifier was biased by 15 cases. However, only 80 out of the 219 cases were TP and the other 139 were FP.

Based on these findings, the predictive accuracy was 39%, with a misclassification rate of 61% for the positive class even though the classifier was biased towards the minority class. The overall predictive accuracy was 59.5% and a 40.5% misclassification rate. The European Credit Card dataset contained a testing sample of 8 545 transactions and the highly imbalanced distribution was 8 528 negatives cases and 17 positive cases. The confusion matrix revealed that the DT classifier had 8 520 TN, 9 FN, 8 TP and 8 FP. Therefore, the predictive accuracy was 99.91% and a misclassification rate of 0.09% was recorded for the negative class whereas a predictive accuracy of 47% and a misclassification rate of 53% were recorded for the positive class. The overall predictive accuracy was 73.5% and the misclassification rate was 26.5%. Based on the results, the average predictive accuracy was 66.5% and there was a 33.5% misclassification rate.

Table 4.7 contains the RF confusion matrix results with the data-point approach.

Table 4.7: Confusion matrix of the RF classifier

European Credit Card Dataset			**UCI Credit Card Dataset**		
RF	**Predicted 0**	**Predicted 1**	**RF**	**Predicted 0**	**Predicted 1**
Actual 0	8 527	1	**Actual 0**	658	38
Actual 1	8	9	**Actual 1**	140	64

The confusion matrix for the RF was both biased and highly inaccurate. Out of a testing sample of 900 cases for the UCI Credit Card dataset, 94.5% and 31% of negative and positive cases were correctly classified, respectively. There were 658 TN out of 696 negative cases and only 64 TP out of 204 positives. There was a 5.5% misclassification rate for the negative class and massive 69% misclassification rate for the positive class. An overall predictive accuracy of 62.75% and a 37.25% misclassification rate were reported, which are reflective of very poor performance for a fraud detection solution.

The European Credit Card dataset had a testing sample of 8 545 transactions, 99.99% of negative cases were correctly classified versus the 53% of positive cases which were correctly classified. The negative class accounted for the majority of the transactions with 8 528 negatives and only 17 positives cases. The confusion matrix revealed that there were 8 527 TP, 8 FN, 9 TP, and only 1 FP. There was a misclassification rate of 0.01% for the negative class and a misclassification rate of 47% for the positive class. The overall predictive accuracy was 76.5% and a misclassification rate of 23.5%, which also reflects a very poor performance, as seen with the UCI Credit Card dataset. The average predictive accuracy was 69.63% and the misclassification rate was 30.37% across both datasets.

4.2.4 The Precision-Recall Curve

At each threshold, a weighted mean of the achieved precisions and recall scores is summarised to create a prediction score called the Average Precision (AP) that is used to create a Precision-Recall (P-R) curve (Subudhi and Panigrahi 2015).

$$\mathbf{AP} = \sum_{n} * \,(\boldsymbol{R_n} - \boldsymbol{R_{n-1}})\boldsymbol{P_n} \tag{4.1}$$

AP was computed using equations (4.1), where R_n and P_n are the respective recall and precision scores at the *n*th threshold. Both scores are always between the range of 0 and 1, which makes the AP to sit within that same range. The AP score is a useful indicator of the classifier's overall accuracy and the closer the value is to 1, the better the accuracy. To observe graphically, a P-R curve is plotted and if the accuracy is high, the chart will lean towards the upper right corner.

4.2.5 Precision-recall curve(s) without the data-point approach

This subsection presents the precision-recall curves before using the data-point approach. Figure 4.1 below shows the European Credit Card dataset precision-recall curve for SVM before Undersampling.

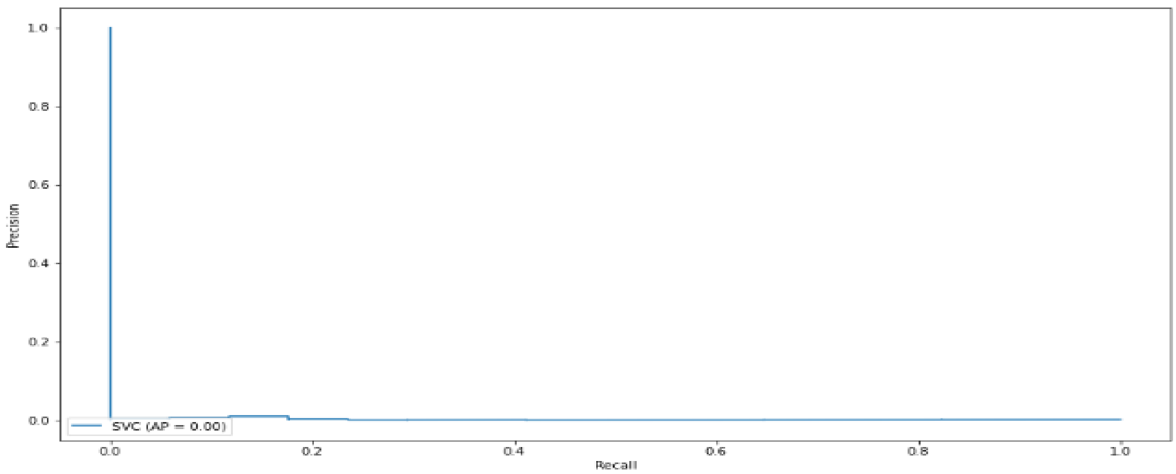

Figure 4.1: European Credit Card Dataset SVM P-R curve

The SVM Precision-Recall curve for the European Credit Card dataset has an average precision of 0.00 (Figure 4.1). The P-R curve of the SVM classifier shows that the graph is leaning sharply towards the lower left corner, which indicates that the classifier was unable to predict any of the positive classes with a straight line across the 0.00 x and y-axes.

Figure 4.2 below shows the UCI Credit Card dataset P-R curve for SVM before Undersampling.

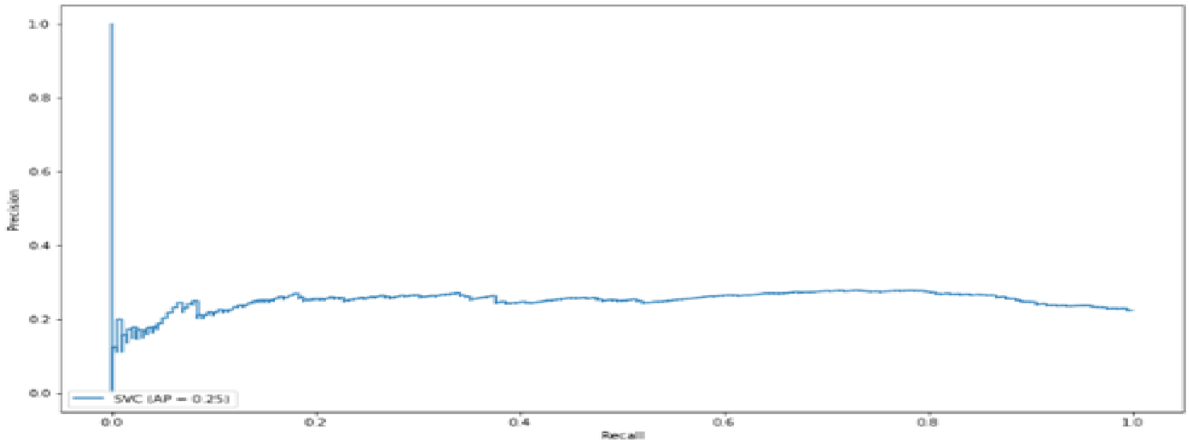

Figure 4.2: UCI Credit Card Dataset SVM P-R curve

Figure 4.2 plots the SVM P-R curve when the Average Precision = 0.22. Figure 4.2 shows the P-R curve of the SVM classifier on the UCI Credit Card dataset, the graph is leaning towards the lower left corner and floating steady at around the 0.22 mark, which indicates that the classifier was struggling to predict any of the positive classes.

Figure 4.3 below shows the European Credit Card dataset P-R curve for LR before Undersampling.

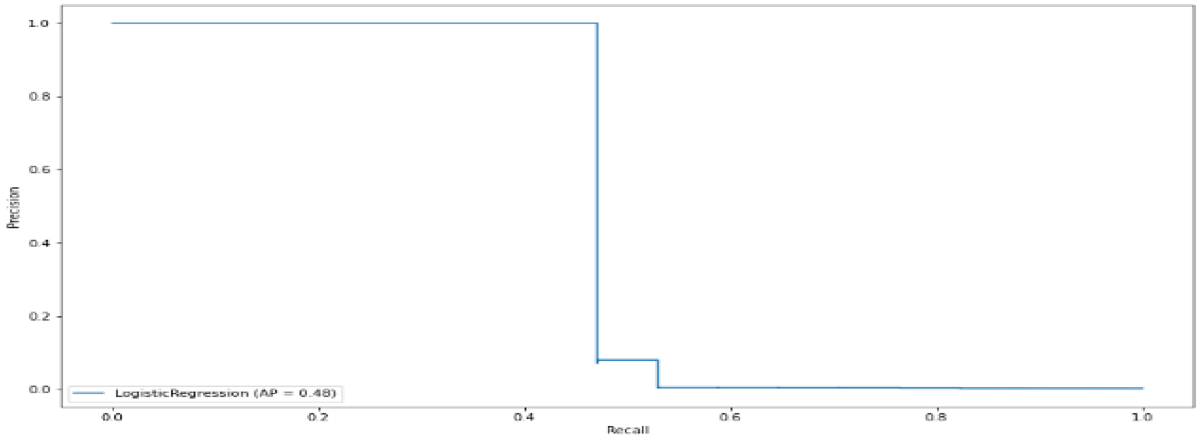

Figure 4.3: European Credit Card Dataset LR P-R Curve

The LR performed slightly better than SVM, the curve starts sharply on the upper left corner and halfway through, falls sharply on the lower right corner. The curve justifies the 0.48 average precision (Figure 4.3).

Figure 4.4 below shows the UCI Credit Card dataset P-R curve for LR before Undersampling.

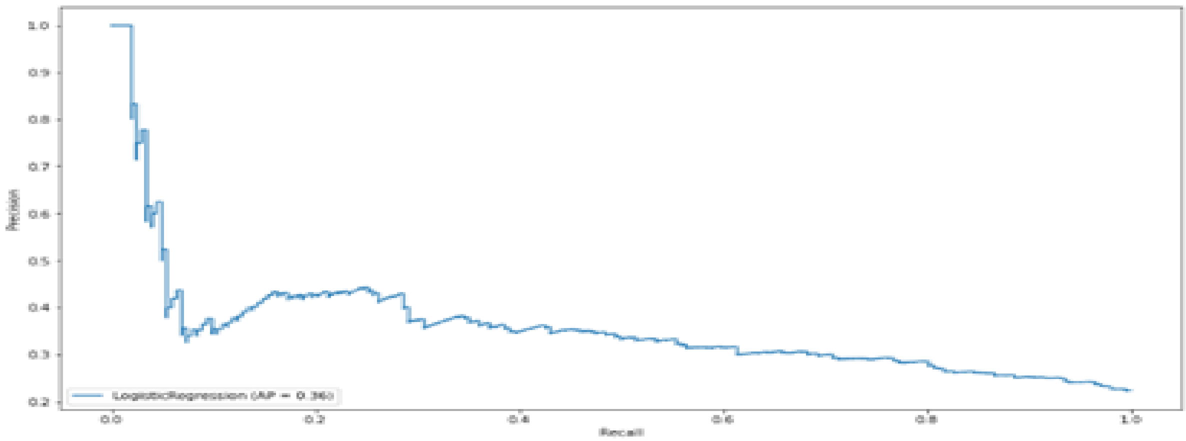

Figure 4.4: UCI Credit Card Dataset LR P-R Curve

Figure 4.4 shows the LR P-R curve of the UCI Credit Card dataset with an average precision of 0.36. The classifier performed poorly as the curve is gradually leaning against the lower left corner towards the lower right corner, showing that quality fading over time. The LR was poor but performed slightly better compared to the SVM.

Figure 4.5 below shows the European Credit Card dataset P-R curve for DT before Undersampling.

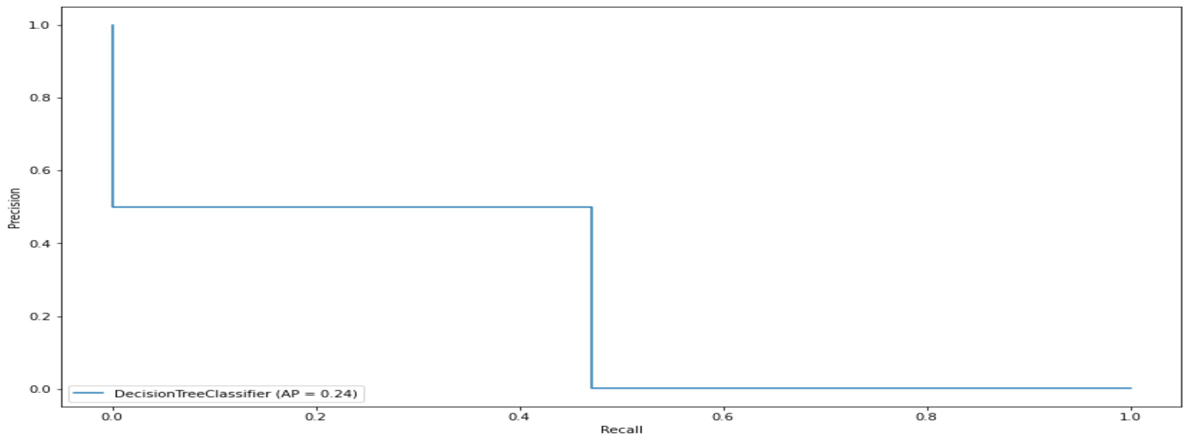

Figure 4.5: European Credit Card Dataset DT P-R Curve

Figure 4.5 shows the DT P-R curve with the European Credit Card dataset. The average precision of 0.24 can be seen through the curve. The DT classifier's P-R curve is leaning towards the lower right corner to indicate the low quality of the classifier.

Figure 4.6 below shows the UCI Credit Card dataset P-R curve for the DT before Undersampling.

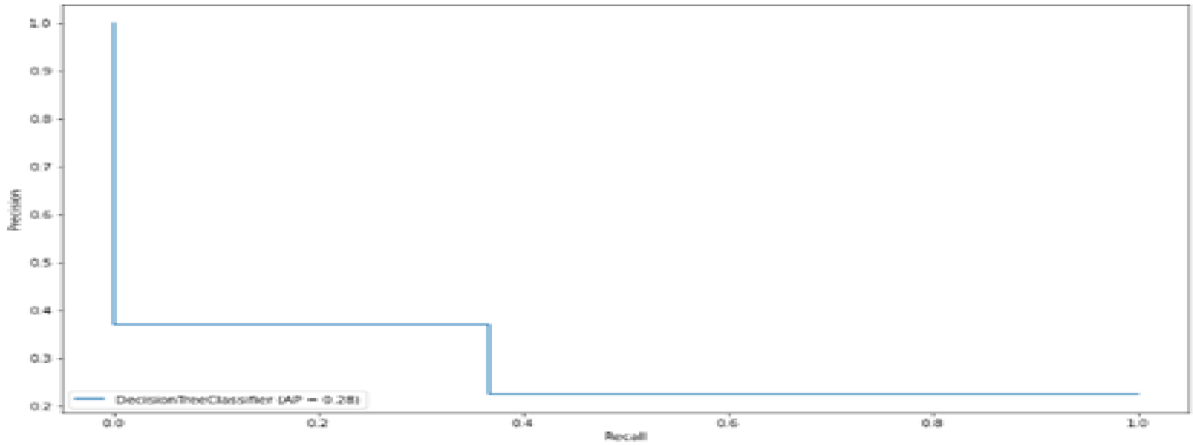

Figure 4.6: UCI Credit Card Dataset DT P-R Curve

Figure 4.6 plots the P-R curves of the UCI Credit Card dataset that is also leaning against the lower left corner toward the lower right corner. Both the P-R curves for the DT look similar; these curves represent poor quality that is confirmed by the low average precision. The UCI Credit Card dataset has a slightly higher average precision of 0.28 compared to the European Credit Card dataset. The lower the average precision, the poorer the quality.

Figure 4.7 below shows the European Credit Card dataset P-R curve for RF before Undersampling.

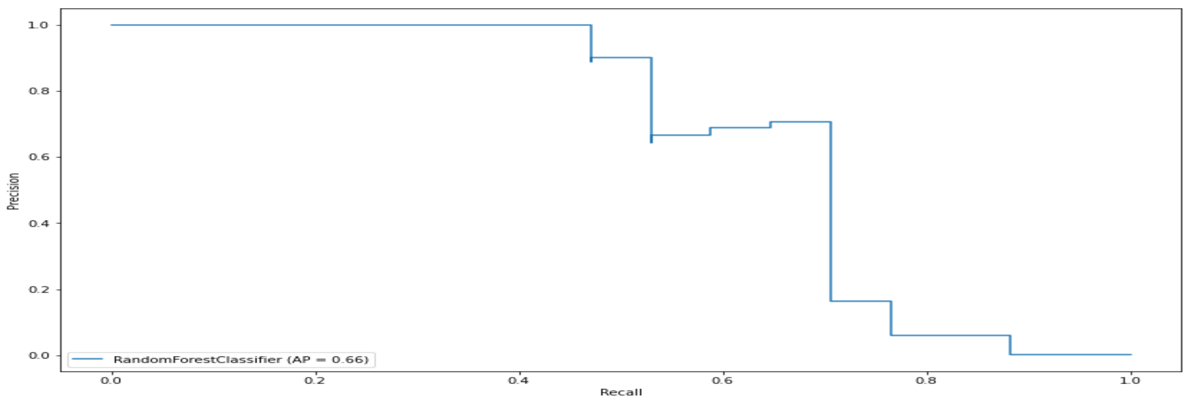

Figure 4.7: European Credit Card Dataset RF P-R Curve

The RF P-R curve for the European Credit Card dataset starts off straight across the highest point and halfway through gradually starts to lean towards the lower right corner. The average precision was 0.66.

Figure 4.8 below shows the UCI Credit Card dataset P-R curve for LR before Undersampling.

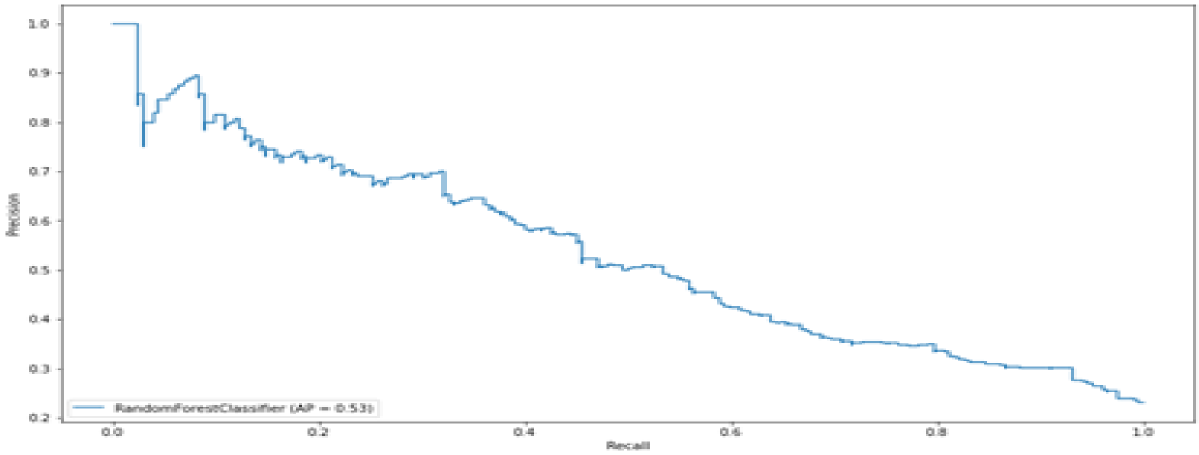

Figure 4.8: UCI Credit Card Dataset (RF) P-R Curve

The RF P-R curve for the UCI Credit Card dataset gradually leaned towards the lower right corner from the beginning. The average precision was 0.37 and the difference can be observed on P-R curves of the RF. The performance was good on the European dataset but was not consistent across both datasets. Figure 4.7 and Figure 4.8 show that the RF performed better than all the other classifies. However, all the above results show poor quality in the ability to predict positive classes for all classifiers. The P-R curve is a simple way to analyse the quality of a classifier without having to perform complex analysis. The next step was to apply the data-point approach and observe the change in quality.

4.3 Post-Treatment-Test results

The next phase of the experiment was to apply the last step of the multi-level hybrid data-point level approach methods, which was to resample the datasets, whereby Undersampling and Oversampling were used to mitigate the effect caused by the class imbalance.

4.3.1 Undersampling

Undersampling was used based on the Near Miss technique to under-sample the majority instances and made it equal to the minority class. Here, the majority class was reduced to the total number of the minority class, so that both classes had an equal number of records. The treatment stage was an iterative process; the aim was to solve the problem of imbalanced data, therefore an in-depth review and analysis were conducted after each iteration to optimise the process.

Table 4.8 shows the results for the classification of the European Credit Card dataset after application of the Undersampling with Near Miss technique.

Table 4.8: Classification of the European Credit Card dataset with Near Miss

Classifier	Metric	Negative	Positive
SVM	Precision	0.65	1.00
	Recall	1.00	0.47
	F1-score	0.79	0.64
	Accuracy	0.73	
	Average Precision	0.73	
LR	Precision	0.88	0.93
	Recall	0.93	0.87
	F1-score	0.90	0.90
	Accuracy	0.90	
	Average Precision	0.87	
DT	Precision	1.00	1.00
	Recall	1.00	1.00
	F1-score	1.00	1.00
	Accuracy	1.00	
	Average Precision	1.00	
RF	Precision	0.83	1.00
	Recall	1.00	0.80
	F1-score	0.91	0.89
	Accuracy	0.90	
	Average Precision	1.00	

The classification report of the European Credit Card dataset after Undersampling with Near Miss to solve the imbalance problem shows a significant improvement on the ability to predict fraudulent transactions. The dataset was balanced with a testing subset containing 30 cases out of a sample size of 98 instances evenly distributed between the two classes, namely, normal and fraudulent transactions. The accuracy score for the SVM classifier decreased from 1.00 to 0.73. However, the ability to predict positive classes improved, the precision score for the positive class increased from 0.00 to 1.00, a 100% improvement.

The recall score increased from 0.00 to 0.47, an improvement of 47%, which means that the SVM classifier had the ability to predict TP after Undersampling with Near Miss even though the percentage achieved is not ideal. The F1-score also increased from 0.00 to 0.64; the improvement verifies the accuracy of the test. The LR reported an accuracy score of 90%, which is a decrease of 10% compared to the results achieved before Undersampling. However, there was a 39% average precision increase from 0.48 to 0.87. The increase in average precision shows that even though accuracy decreased, the overall predictive accuracy increased.

The increase in predictive accuracy is observed by the increase in precision, recall, and F1-score for positive classes. The DT precision increased from 0.57 to a decent 0.93, recall increased from 0.47 to 0.87, and the F1-score increased from 0.52 to 0.90 for the positive class. The negative class performed fairly well too, even though the initial 100% accuracy was not achieved, the classifier was not biased in either class. The precision was 0.88, recall was 0.93, and the F1-score was 0.90 for the negative class. The RF classification was similar to the LR, which also reported an accuracy of 90%.

The precision was 0.83 for the negative class and 1.00 for the positive class. Recall was 1.00 for the negative class and 0.80 for the positive class. The F1-score was 0.91 for the negative class and 0.89 for the positive class. The RF performed better than all other classifiers before using Undersampling but was closely followed by the DT. However, the DT surpassed the RF and gave better results after Undersampling with Near Miss. The DT maintained an accuracy score of 100% and the average precision increased from 28% to 100%. The precision, recall, and F1-score for both the negative and positive classes was an impressive 100%.

Table 4.9 shows the results for the classification of the UCI Credit Card dataset after application of the Undersampling with Near Miss technique.

Table 4.9: Classification of the UCI Credit Card dataset with Near Miss

Classifier	Metric	Negative	Positive
SVM	Precision	0.77	0.96
	Recall	0.97	0.73
	F1-score	0.86	0.83
	Accuracy	0.85	
	Average Precision	0.84	
LR	Precision	0.70	0.76
	Recall	0.79	0.66
	F1-score	0.74	0.71
	Accuracy	0.73	
	Average Precision	0.79	
DT	Precision	0.85	0.86
	Recall	0.85	0.86
	F1-score	0.85	0.86
	Accuracy	0.85	
	Average Precision	0.81	
RF	Precision	0.86	0.92
	Recall	0.92	0.86
	F1-score	0.89	0.89
	Accuracy	0.89	
	Average Precision	0.86	

The classification report for the UCI Credit Card dataset revealed that there was an overall improvement in the ability to predict positive classes. The SVM reported an accuracy score of 85%, which is an increase of 7% compared to the accuracy achieved before Undersampling. The ability to predict the positive class improved as the average precision increased from 0.22 to 0.84, an improvement of 62%. The LR accuracy decreased from 0.78 to 0.73. However, the average precision improved from 0.36 to 0.79. These results show that the LR improved its ability to predict positive classes.

The DT reported an increase in accuracy from 0.72 to 0.85; indicating an improved accuracy of 85%. The average precision also increased from 0.28 to 0.81, an improvement of 53%. The RF reported an accuracy score of 89%, which was the highest reported. The average precision also increased from 0.37 to 0.86, an improvement of 49%. All the classifiers reported improved precision, recall, and F1-scores after using Undersampling.

4.3.1.1 The Confusion Matrix with Near Miss

Table 4.10 below contains the SVM confusion matrix after Undersampling with Near Miss.

Table 4.10: Confusion matrix of the SVM classifier with Near Miss

European Credit Card Dataset			**UCI Credit Card Dataset**		
SVM	**Predicted 0**	**Predicted 1**	**SVM**	**Predicted 0**	**Predicted 1**
Actual 0	15	0	**Actual 0**	208	4
Actual 1	8	7	**Actual 1**	40	154

The SVM confusion matrix showed improvement in the ability to predict positive classes. Even though some confusion level still exists, the effect of Near Miss was observed on both datasets. The European Credit Card testing dataset contained 30 samples that were evenly distributed. A total of 7 out of 15 positive cases were correctly predicted and a total of 8 positive cases were misclassified as negative cases. The misclassification rate of positive cases was 55%. The predictive accuracy of true positives improved when compared to classification results before using the data-point approach. There were more FN than TP, which shows that the SVM classifier was still biased towards the negative class. To further support these findings, there were no FP in the European Credit Card dataset, which means that all the negative cases were correctly classified.

The confusion matrix shows that a total of 23 instances were predicted as negative cases, even though there were only 15 instances of the negative class on the testing dataset. There was 100% predictive accuracy for TN, 15 out 15 instances of the negative class were correctly classified. The confusion matrix of the UCI Credit Card dataset revealed that the SVM classifier was less biased, but there was still a significant amount of misclassification. The testing dataset contained a total of 406 instances, wherein the negative class accounted for 212 instances and the positive class accounted for 194 instances. Interestingly, the SVM classifier was still more biased towards the negative class even though there was a small difference between the class distributions on the testing dataset.

The confusion matrix showed that a total of 248 cases were classified as the negative class and a total of 158 cases were classified as the positive class. There were 40 FN, which means that only 168 out of the 212 negative cases were correctly classified. There were 4 FP, which means only 154 out of 194 positive cases were correctly classified.

Table 4.11 below shows the LR confusion matrix after Undersampling with Near Miss.

Table 4.11: Confusion matrix of the Logistic Regression classifier with Near Miss

European Credit Card Dataset			**UCI Credit Card Dataset**		
LR	**Predicted 0**	**Predicted 1**	**LR**	**Predicted 0**	**Predicted 1**
Actual 0	15	0	**Actual 0**	168	44
Actual 1	2	13	**Actual 1**	61	133

The confusion matrix for the LR model also shows that the Near Miss technique worked well for both datasets. There was a 100% predictive accuracy for negative cases and 87% predictive accuracy for positive cases on the European Credit Card dataset. There was no misclassification for the negatives, 15 out of 15 negative cases were accurately predicted. However, even though there was no confusion on the negative cases, some confusion and bias was observed on the positive cases. The classifier predicted a total of 17 instead of 15 negatives cases, which means that there were 2 FN. The misclassification rate for positive cases was 13%, since 2 out of the 15 positive cases on the testing dataset were misclassified. Since the original dataset contained 492 fraudulent transactions, a misclassification rate of 15% would mean that nearly 74 in every 492 fraudulent cases would still be processed.

The UCI Credit Card dataset had an accuracy of 80% for negative cases and 66% for positive cases. The classifier predicted 229 negative cases even though there were only 212 instances of the negative class on the testing dataset. The classifier was biased towards the negative class by 17 instances. However, there was confusion on both classes; out of the 229 instances classified as negative cases, there were only 168 TN predicted and the misclassification rate for negative cases was 21%. The classifier predicted 177 positives cases, of which 44 of those cases were FP. Therefore, only 133 TP were classified out of 194 positive cases that were on the testing dataset. The misclassification rate for positive cases was 31%, there were 61 FN misclassified out of the 194 positive cases. The ability to predict the positive class improved after using Near Miss with the LR classifier, but some bias towards the negative class was still observed on both datasets, along with some misclassifications.

Table 4.12 below shows the DT confusion matrix after Undersampling with Near Miss.

Table 4.12: Confusion matrix of the Decision Tree classifier after with Near Miss

European Credit Card Dataset			**UCI Credit Card Dataset**		
DT	**Predicted 0**	**Predicted 1**	**DT**	**Predicted 0**	**Predicted 1**
Actual 0	15	0	**Actual 0**	176	36
Actual 1	0	15	**Actual 1**	27	167

Using the Near Miss technique, the DT produced results with high predictive accuracy. There was no confusion (100% accuracy) for both classes on the European Cardholders' transactions dataset. That means the ability to predict positive classes improved by 47% after Undersampling with Near Miss. There was a total of 30 instances in the testing dataset that were evenly distributed for the negative and positive classes. The confusion matrix revealed that there was a total of 15 TN and 15 TP, zero FN and zero FP. Therefore, there were no misclassifications and no bias reported on the European Credit Card dataset. The DT after Undersampling with Near Miss on the UCI Credit Card dataset also performed well but the results were not as good as those achieved on the European Credit Card dataset.

The UCI Credit Card testing dataset contained 406 instances. The negative and positive classes recorded 212 and 194 cases respectively. The confusion matrix revealed that the classifier predicted 203 out of the 212 negative cases. However, there were only 176 TN, which means that the predictive accuracy was 83% for the negative class and a misclassification rate of 17%. The classifier predicted 36 FP and 167 TP out of the 203 cases classified as the positive class. The predictive accuracy for the positive class was 86% and a misclassification rate of 14%. The overall predictive accuracy for the DT classifier with Near Miss was 84.5% and an average rate of 15.5% misclassification rate. The average predictive accuracy across both datasets for the DT classifier was 92%, which was a significant improvement compared to the results obtained with the imbalanced dataset.

The confusion matrix revealed the classifier to be biased towards the positive class even though the testing dataset contained more negative cases. To explain this phenomenon, the classes were distributed evenly after Undersampling the dataset, a random split of 70:30 meant that if the testing dataset contained more negative cases, then, the training dataset had more positive cases. However, the same distribution was used on the LR but the results show that the classification was biased towards the negative class. Therefore, the difference was too little to influence the performance and accuracy of the test and to cause the classifier to be biased towards either class.

Table 4.13 below contains the RF confusion matrix after Undersampling with Near Miss.

Table 4.13: Confusion matrix of the RF classifier with Near Miss

European Credit Card Dataset			**UCI Credit Card Dataset**		
RF	**Predicted 0**	**Predicted 1**	**RF**	**Predicted 0**	**Predicted 1**
Actual 0	15	0	**Actual 0**	193	19
Actual 1	3	12	**Actual 1**	21	173

The RF classifier performed well with Near Miss. The European Credit Card dataset contained 30 testing samples. There was an equal class distribution of 15 negative cases and 15 positive cases. The confusion matrix revealed that the classifier predicted 18 negative cases. There were 15 TN, which meant all the negative cases were correctly predicted. The classifier was biased towards the negative class by three FN. There was a predictive accuracy of 100% on the European Credit Card dataset and a zero misclassification for the negative class. The RF with Near Miss classifier predicted 12 out 15 positive cases. There were zero FP and a misclassification of 20% for the positive cases.

The predictive accuracy of the positive class was 80%. The overall predictive accuracy was 90%, with a misclassification rate of 10%. The UCI Credit Card dataset contained a testing sample of 406 instances. There were 212 negatives cases and 194 positive cases. The confusion matrix revealed that the RF classifier predicted 214 negative cases, wherein there were 193 TP and 21 FN. The classifier was biased by only two cases. The predictive accuracy was 91%, with a misclassification of 9% for negative cases. The confusion matrix also showed that the classifier predicted 192 out of 194 positive cases.

However, from the 192 predicted cases, there were only 173 TP and 19 FP. There was a predictive accuracy of 89% and a misclassification rate of 11% for the positive class. The overall predictive accuracy was 90%, with a 10% misclassification rate. These results show that the RF classifier with Near Miss was consistent across both datasets. The overall predictive accuracy and misclassification rates were the same, which meant that the average predictive accuracy remained unchanged at 90%. The confusion matrix revealed that there was low bias and low variety. The misclassification rate for the RF classifier with Near Miss was very low at 10% compared to the misclassification rate of 30.37% that was obtained by the RF before using any data-point on both datasets.

Based on these findings, the misclassification rate decreased by 20.37% and the predictive accuracy improved by the same 20.37% (from 69.63% to 90%). The RF classifier was more consistent by achieving 90% on both datasets with 0% performance difference. The average predictive accuracy of the DT was higher at 92%. However, the DT only performed better since the 100% accuracy linked to the European Credit Card dataset boosted the average accuracy; there was only 84.5% accuracy with the UCI Credit Card dataset, resulting in a 15.5% performance difference. Therefore, our conclusion was that based on performance difference, the RF classifier was the best performing solution after Undersampling with Near Miss.

4.3.1.2 Precision-Recall curve(s) with Near Miss

The figures below show the P-R curve after treatment using Undersampling with the Near Miss technique applied. A P-R curve is an effective way to provide a graphical visualisation of the quality of a classifier. The P-R curves below show the improvement in the quality of the classifiers after using the data-point approach.

Figure 4.9 below shows the European Credit Card dataset P-R curve for SVM after Undersampling.

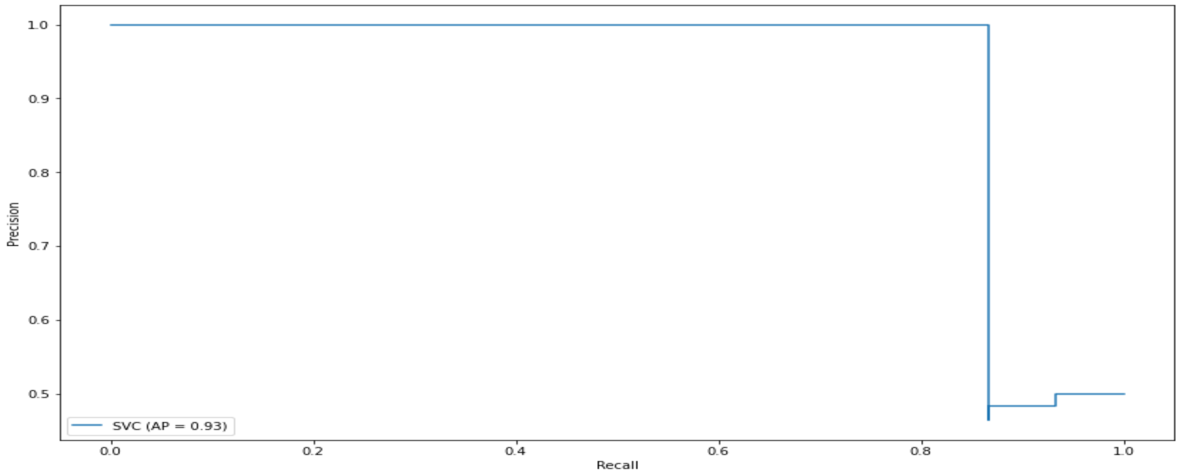

Figure 4.9: European Credit Card Dataset (SVM with Near Miss) Precision-Recall Curve

The European Credit Card dataset SVM Precision-Recall curve is a straight line leaning towards the upper right corner and loses momentum at the end with a sharp decline towards the end of the curve but slightly increase thereafter, hence the average precision was 0.93. The SVM classifier improved significantly after Undersampling compared to the performance achieved before any data treatment was applied. The classifier improved by 93% on the European Credit Card dataset.

Figure 4.10 below shows the UCI Credit Card dataset P-R curve for SVM after Undersampling.

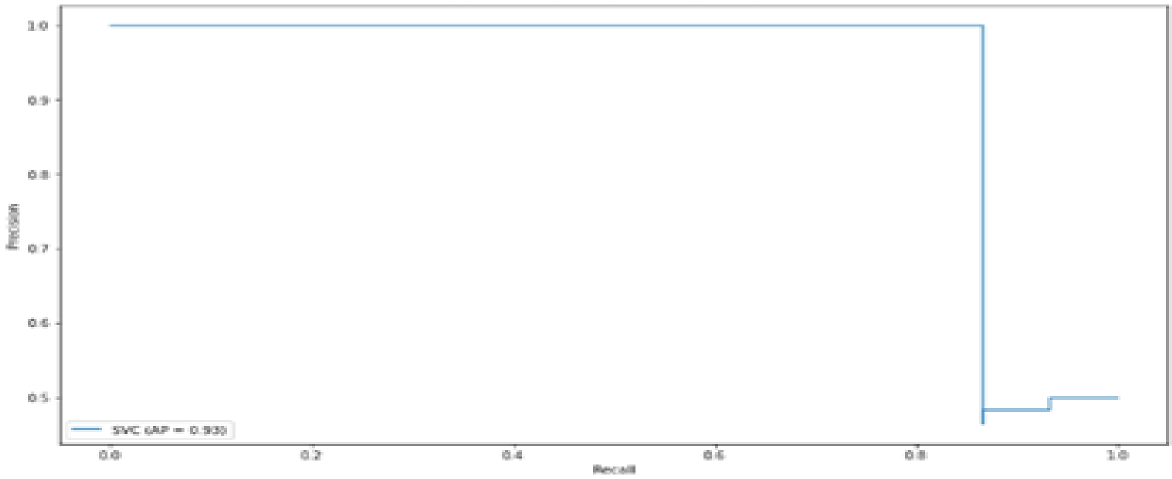

Figure 4.10: UCI Credit Card Dataset (SVM with Near Miss) P-R Curve

Figure 4.10 shows the SVM P-R curve on the UCI Credit Card dataset. The curve is similar for both datasets. The average precision increased from 0.22 to 0.84, an improvement of 62%. The SVM classifier showed a great improvement in quality.

Figure 4.11 below shows the European Credit Card dataset Precision-Recall curve for LR after Undersampling.

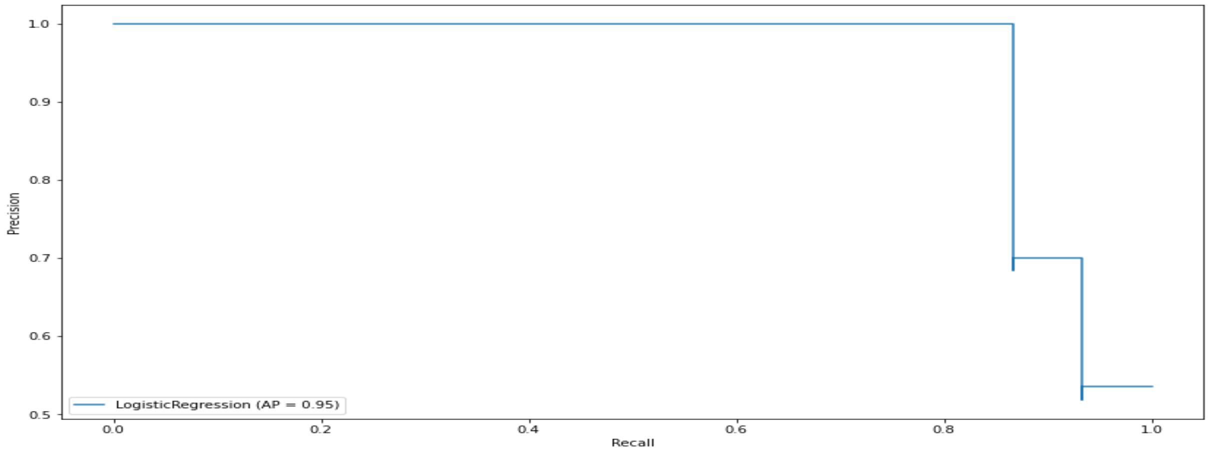

Figure 4.11: European Credit Card Dataset (LR with Near Miss) P-R Curve

Figure 4.11 shows the LR P-R curve for the European Credit Card dataset. The average precision was 0.87 and curve is positioned across the upper right corner which indicates that the quality of the classifier was good.

Figure 4.12 below shows the UCI Credit Card dataset P-R curve for LR after Undersampling.

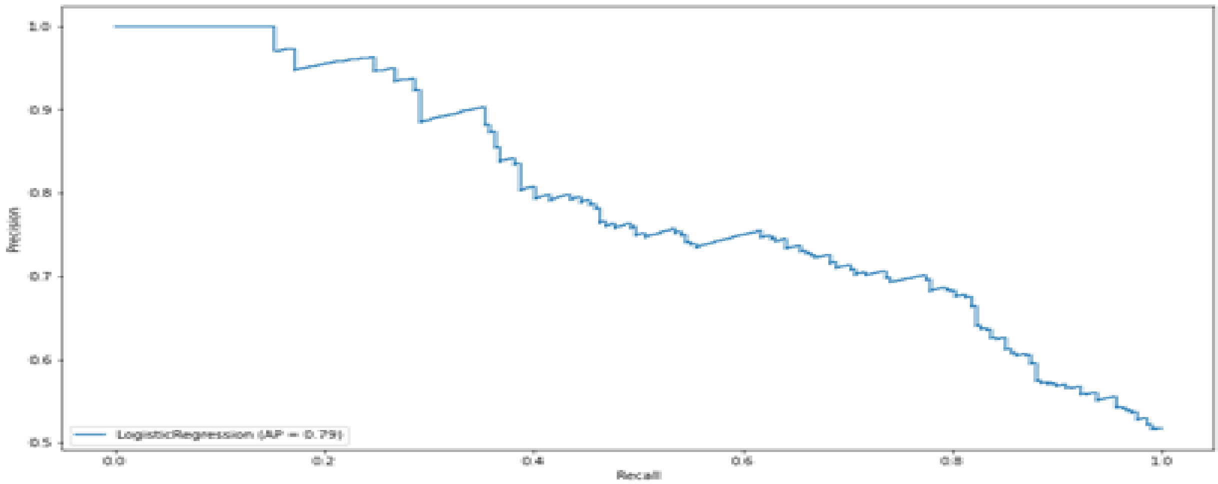

Figure 4.12: UCI Credit Card Dataset (LR with Near Miss) P-R Curve

Figure 4.12 shows the LR P-R curve of the UCI Credit Card dataset. The curve gradually leans towards the lower right corner. The average precision increased from 0.36 to 0.79, an improvement of 43%.

Figure 4.13 below shows the European Credit Card dataset P-R curve for DT after Undersampling.

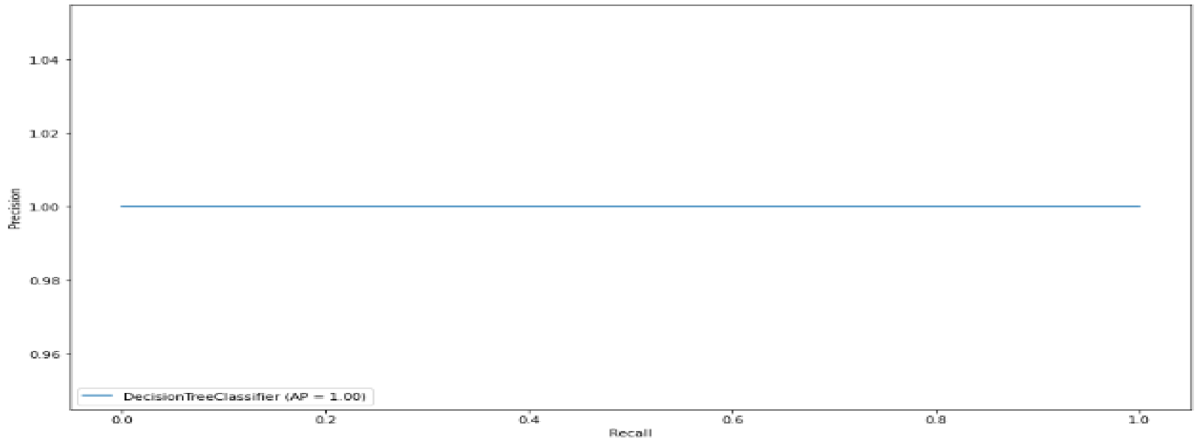

Figure 4.13: European Credit Card Dataset (DT with Near Miss) P-R Curve

Figure 4.13 shows the DT P-R curve on the European Credit Card dataset. The classifier improved by 76% as the average precision increased from 0.24 to 1.00, indicated by the straight line on the value of 1 across the y-axis. A straight-line graph represents the best quality.

Figure 4.14 below shows the UCI Credit Card dataset P-R curve for the DT after Undersampling.

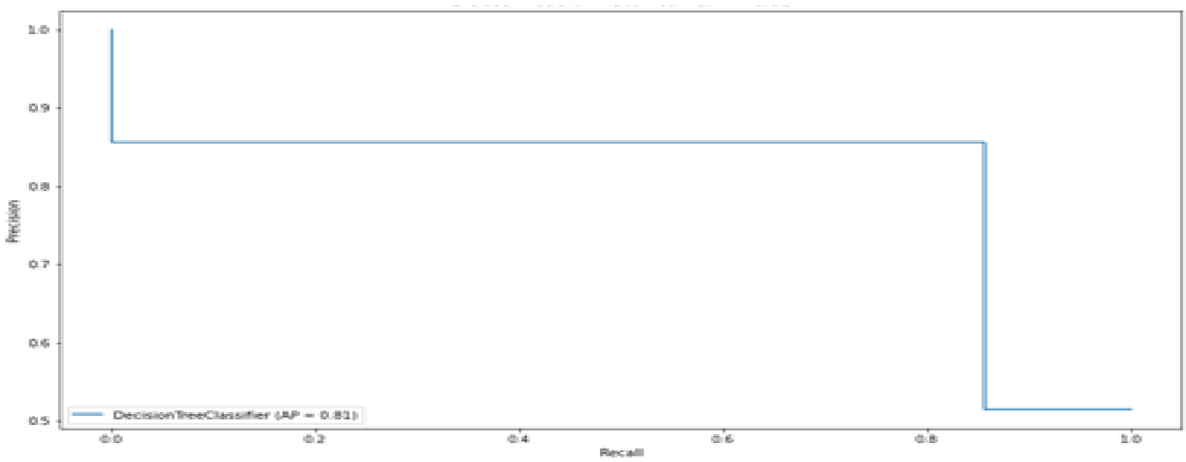

Figure 4.14: UCI Credit Card Dataset (DT with Near Miss) P-R Curve

\

Figure 4.14 shows the DT P-R curve of the UCI Credit Card dataset. The curve is sharply leaning towards the upper right corner. The average precision increased from 0.28 to 0.81, a quality improvement of 53% for the DT classifier.

Figure 4.15 below shows the European Credit Card dataset Precision-Recall curve for RF after Undersampling.

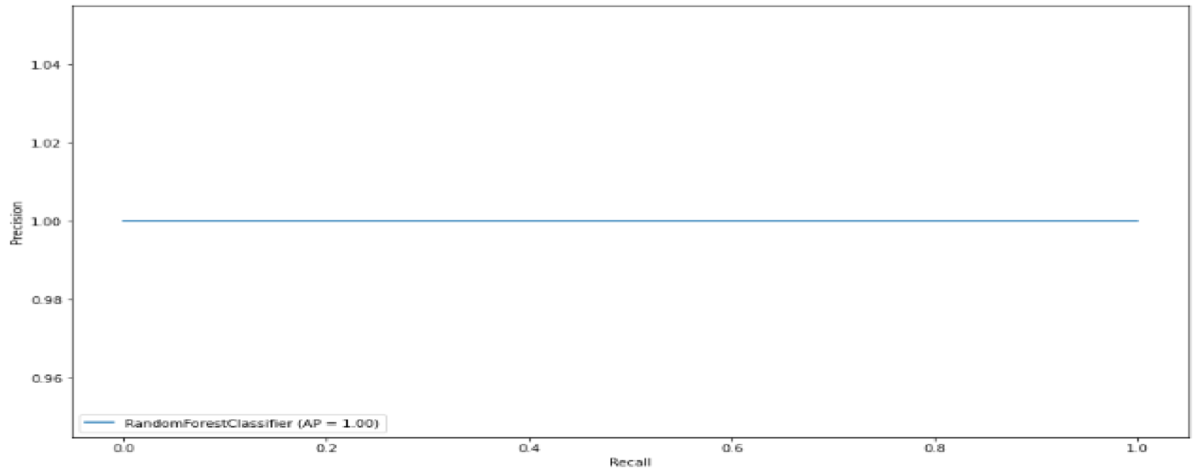

Figure 4.15: European Credit Card Dataset (RF with Near Miss) P-R Curve

Figure 4.15 shows the RF P-R curve on the European Credit Card dataset. The classifier improved by 33% as the average precision increased from 0.66 to 1.00, indicated by the straight line on the value of 1 across the y-axis, as observed on the DT P-R curve.

Figure 4.16 below shows the UCI Credit Card dataset P-R curve for RF after Undersampling.

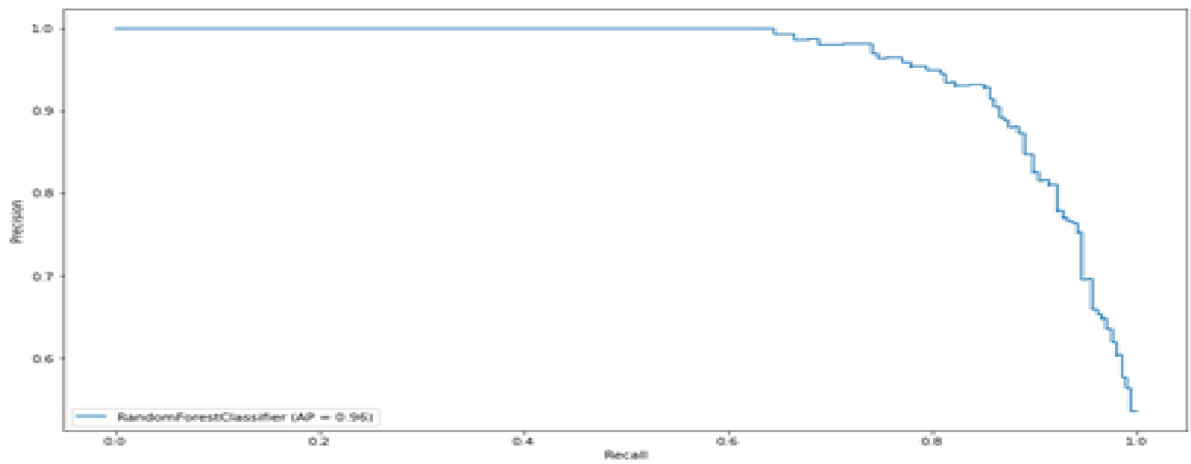

Figure 4.16: UCI Credit Card Dataset (RF with Near Miss) P-R Curve

Figure 4.16 shows the RF P-R curve of the UCI Credit Card dataset. The curve starts straight on the value of 1 on the y-axis, moving across the x-axis and ends by a gradual decline, while leaning towards the upper right corner. The average precision increased from 0.28 to 0.81, a quality improvement of 53% for the RF classifier. A P-R curve that is a straight line on the y-axis value of 1 across the x-axis, such as in Figure 4.13 of the DT and Figure 4.15 of the RF with the European Credit Card dataset represents the best possible quality.

A P-R curve that leans more towards the upper right corner is also a sign that the classifier has good quality, such as in Figure 4.10, Figure 4.14, and Figure 4.16 of the various classifiers on the UCI Credit Card datasets are good quality classifiers. Figure 4.12 of the LR on the UCI Credit Card dataset is the only one to lean more towards the lower right corner after Undersampling with Near Miss, which indicates that the LR classifier was not consistent across different datasets and there was still high variance.

4.3.2 Oversampling

The original dataset was resample using a SMOTe based Oversampling technique. The randomness of the bootstrapping of the samples was controlled using a parameter called random-state with the value of 42. The value was selected after multiple iterations of experimenting were conducted to find the most effective random state.

Figure 4.17 below shows the distribution of the classes after applying SMOTe.

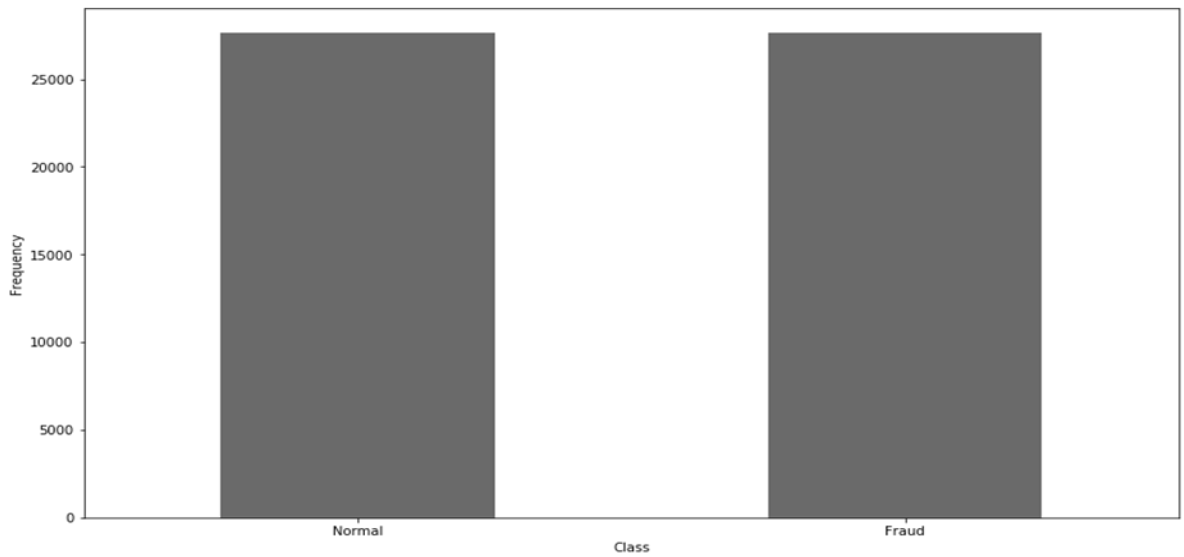

Figure 4.17: Transaction Class Distribution after Oversampling with SMOTe

Figure 4.17 shows the new class distribution of the normal and fraud cases. The dataset is evenly balanced after increasing the minority cases.

Table 4.14 below shows the classification report for the European Credit Card dataset after Oversampling was applied to mitigate the effects caused by class imbalance.

Table 4.14: Classification of the European Credit Card dataset with SMOTe

Classifier	Measure	Negative	Positive
SVM	Precision	0.60	0.55
	Recall	0.37	0.76
	F1-score	0.46	0.64
	Accuracy	0.57	
	Average Precision	0.53	
LR	Precision	0.97	0.97
	Recall	0.97	0.97
	F1-score	0.97	0.97
	Accuracy	0.97	
	Average Precision	0.96	
DT	Precision	1.00	1.00
	Recall	1.00	1.00
	F1-score	1.00	1.00
	Accuracy	1.00	
	Average Precision	1.00	
RF	Precision	1.00	1.00
	Recall	1.00	1.00
	F1-score	1.00	1.00
	Accuracy	1.00	
	Average Precision	1.00	

The testing set for the European Credit Card dataset contained a sample size of 17 060 cases. The distribution was 8 468 for negative cases and 8 592 for the positive class. The accuracy score for the SVM classifier decreased from 1.00 to 0.57. Though, the ability to predict positive classes improved, the precision score for the positive class increased from 0.00 to 0.55, an improvement of 55%. The recall score increased from 0.00 to 0.76, an improvement of 76%, which means that after applying SMOTe, the SVM classifier had the ability to predict TP. The F1-score also increased from 0.00 to 0.64; the improvement verifies the accuracy of the test increased. The LR reported a score of 97% for all the metrics except the average precision, which achieved a score of 0.96.

These scores show that the overall predictive accuracy increased for the LR. The increase in predictive accuracy is observed by the increase in precision, recall, and F1-score for positive classes while maintaining a good score for the negative class. The DT saw the precision increase from 0.57 to a decent 1.00, recall increased from 0.47 to 1.00, and the F1-score increased from 0.52 to 1.00 for the positive class. The negative class also achieved the score of 100% for all the metrics. The RF classification was similar to the DT, which also reported a perfect score of 1.00 for all the metrics for both negative and positive cases.

Table 4.15 shows the classification report for the UCI Credit Card dataset after Oversampling.

Table 4.15: Classification of the UCI Credit Card dataset with SMOTe

Classifier	Measure	Negative	Positive
SVM	Precision	0.65	0.59
	Recall	0.52	0.70
	F1-score	0.58	0.64
	Accuracy	0.61	
	Average Precision	0.56	
LR	Precision	0.65	0.61
	Recall	0.59	0.68
	F1-score	0.62	0.64
	Accuracy	0.63	
	Average Precision	0.57	
DT	Precision	0.81	0.76
	Recall	0.75	0.82
	F1-score	0.78	0.79
	Accuracy	0.78	
	Average Precision	0.71	
RF	Precision	0.88	0.88
	Recall	0.89	0.87
	F1-score	0.88	0.88
	Accuracy	0.88	
	Average Precision	0.83	

The testing set for the UCI Credit Card dataset contained a sample size of 1 395 cases. The distribution was 711 cases for negative cases and 684 cases for the positive class. The classification algorithm that produced the best results on both datasets was the RF with an average accuracy of 94% after recording an accuracy of 100% on the European Credit Card dataset and 88% on the UCI Credit Card dataset. In second position was the DT with an average accuracy of 85.5%, followed by the LR with an average accuracy of 76.5%. The least effective method was the SVM classifier with an average accuracy of 59%. Although the classification was consistent across both datasets, SMOTe performed very well on the European Credit Card dataset compared to the UCI Credit Card dataset.

4.3.2.1 The Confusion Matrix with SMOTe

Table 4.16 below shows the SVM confusion matrix after applying the SMOTe based Oversampling technique.

Table 4.16: Confusion matrix of the SVM classifier after Oversampling with SMOTe

European Credit Card Dataset			**UCI Credit Card Dataset**		
SVM	**Predicted 0**	**Predicted 1**	**SVM**	**Predicted 0**	**Predicted 1**
Actual 0	3 107	5 361	**Actual 0**	370	341
Actual 1	2 056	6 536	**Actual 1**	202	482

Compared to the SVM confusion matrix before the data-point approach, there was significant improvement in the ability to predict positive cases. Even though some confusion level still exists, the effect of SMOTe was observed on both datasets. The total number of positive cases predicted was 11 897 whereas the actual number of positive cases in the testing dataset was 8 592 cases. This time, the SVM classifier was more biased towards the positive class. However, the prediction was still unsuccessful, 5 361 cases out of the 11 897 that were classified as positive cases were FP. Therefore, the FP misclassification rate was 45%, which is too high for a fraud detection solution.

The predictive accuracy of TP improved significantly by 76%; only 2 056 out of the 8 592 positive cases were incorrectly classified. The total number of negative cases predicted was 5 163 out of 8 468 testing cases in the dataset. However, the confusion matrix shows that only 3 107 cases were TN, which meant that 63% of the negative cases were misclassified and 39% of prediction was FN. The SVM classification of the UCI Credit Card dataset was similar to the European Credit Card dataset prediction. The classifier was also biased towards the positive class, which meant that the performance was consistent across both datasets. There were 823 positive cases predicted even though there were only 684 positive cases in the testing dataset.

Out of the 823 cases, the misclassification rate for FP was 41%, which makes the average misclassification rate for FP 43% across both datasets. The predictive accuracy for TP also improved by 70% after using SMOTe compared to the results obtained before any data-point treatment was applied; 482 out of 684 TP were correctly classified. There were 202 FN and a success rate of 52% for TN, which meant the average misclassification rate of TP was 56% for an SVM with SMOTe based Oversampling solutions.

Table 4.17 below shows the LR confusion matrix after applying the SMOTe based Oversampling technique.

Table 4.17: Confusion matrix of the Logistic Regression classifier with SMOTe

European Credit Card Dataset			**UCI Credit Card Dataset**		
LR	**Predicted 0**	**Predicted 1**	**LR**	**Predicted 0**	**Predicted 1**
Actual 0	8 231	252	**Actual 0**	421	290
Actual 1	357	8 220	**Actual 1**	222	462

The confusion matrix shows that the LR classifier performed much better than the SVM classifier. The results for the European Credit Card dataset show that out of a testing dataset of 17 060 cases, there were 8 588 negative cases and 8 472 positive cases predicted. There was a total of 8 577 TP in the testing dataset. Based on these findings, the classifier was 99% accurate when looking only at the sample size. Examination of the confusion matrix reveals that there were 252 FP, which made the predictive accuracy of TP to be 96% with a misclassification rate of 4%. The number of FN was 357 cases, which makes the total misclassification rate only 3% and the predictive accuracy, 97%. The overall performance with the European Credit Card dataset was acceptable.

To validate the consistency of these results, we further examined the UCI Credit Card dataset classification. Firstly, the findings reveal that there was low bias, 752 positive cases were predicted but there was only a total of 684 TP in the testing dataset. The difference with this classification is that the classifier is now biased towards the positive class. However, there was minimum bias as there were 643 negative cases predicted out of a total of 711 TN existing in the testing dataset. Secondly, to validate the consistency, we considered the misclassification rate and the predictive accuracy. There was a total of 290 FP, a misclassification rate of 32% for positive cases, and 68% predictive accuracy.

The total number of FN was 222 cases, the misclassification rate was 41%, and the predictive accuracy was 59% for the negative class. The average predictive accuracy across both datasets was 78% for the negative class and 82% for the positive class. The difference between the positive class misclassification rate of the European Credit Card dataset and UCI Credit Card dataset was 28%, which is a sizeable difference. The difference for the negative class misclassification rate was 38%, and based on these findings, a LR with SMOTe based Oversampling solution is less biased but has high variance.

Table 4.18 below shows the DT confusion matrix after applying the SMOTe based Oversampling technique.

Table 4.18: Confusion matrix of the Decision Tree classifier with SMOTe

European Credit Card Dataset			UCI Credit Card Dataset		
DT	**Predicted 0**	**Predicted 1**	**DT**	**Predicted 0**	**Predicted 1**
Actual 0	8 445	23	**Actual 0**	532	179
Actual 1	1	8 591	**Actual 1**	122	562

The confusion matrix of the DT after Oversampling with SMOTe showed an improved performance but some misclassification was observed in both datasets. The European Credit Card testing dataset contained 17 060 instances. The negative class accounted for 8 468 cases and the positive class accounted for 8 592 cases. The confusion matrix revealed that the classifier predicted 8 446 of the 8 468 negative cases, whereby only one case was a FN. The calculated predictive accuracy was 99.7% for the negative class, with a misclassification rate of 0.27% for the 23 misclassified negative cases. The predictive accuracy for the positive class was calculated at 99.9%, the classifier predicted 8 614 positive cases even though there were only 8 592 positive cases in the testing dataset.

Out of the 8 614 cases, there were 8 591 TP and 23 FP. Only one positive case was misclassified as a FN. The overall predictive accuracy for the classes was 99.8%, which shows that the DT classifier with SMOTe was able to classify both classes accurately with minimum bias and a misclassification rate of less than 1%.The UCI Credit Card testing dataset contained a sample size of 1 395 instances. There were 711 negative cases and 684 positive cases. The confusion matrix revealed that the classifier predicted 654 out of 711 negative cases. Out of the 654, there were 532 TP and the other 122 were FN. The calculated predictive accuracy was 75% and a misclassification rate of 25% for the negative class. The confusion matrix shows that the classifier was biased by 57 cases as the predicted number of the positives was 741 instead of the 684 cases. However, the confusion matrix revealed that there was both bias and confusion; out of the 741 cases there were only 562 TP. The calculated predictive accuracy was 82% and a misclassification rate of 18% was shown for the positive class.

The average predictive accuracy for both classes was 78.5% and the average misclassification rate was 9.5% for the UCI Credit Card dataset. The average predictive accuracy across both datasets was 89.15%, which meant that the DT with SMOTe classifier was able to produce consistent results across the two distinct datasets. Even though some confusion was still observed, the average misclassification rate was just below 10% due to the slightly higher misclassification rate of the UCI Credit Card dataset. The solution performed very well on the European Credit Card dataset but the confusion matrix revealed that further research can be conducted to understand the factors that contribute to the overall performance of the classifier.

More research can also be done to find additional optimisation that can be applied to achieve similar results on other datasets. The overall conclusion is that the DT with SMOTe was able to address the issues caused by class imbalance in the credit card dataset. The predictive accuracy and the ability to predict positive cases improved significantly when compared to the results achieved without the data-point approach.

Table 4.19 below shows the RF confusion matrix after applying the SMOTe based Oversampling technique.

Table 4.19: Confusion matrix of the RF classifier with SMOTe

European Credit Card Dataset			**UCI Credit Card Dataset**		
RF	**Predicted 0**	**Predicted 1**	**RF**	**Predicted 0**	**Predicted 1**
Actual 0	8 467	1	**Actual 0**	631	80
Actual 1	0	8 592	**Actual 1**	89	595

The confusion matrix revealed that the RF classifier with SMOTe based Oversampling produced the best classification results compared to the other three algorithms. The European Credit Card testing dataset contained a sample size of 17 060 instances. The total number of negative cases was 8 468 and the total number of positive cases was 8 592. The confusion matrix revealed that the RF classifier with SMOTe predicted 8 467 out of the 8 648 negative cases. All the 8 467 predicted cases were TN and there were zero FN. The classification only failed to predict one negative case. The predictive accuracy was 99.99%, with a misclassification rate of 0.01% for the negative class.

The confusion matrix revealed that the classifier predicted 8 593 positive cases, which was biased by one FP. All the positive cases were correctly classified, there were 8 592 TP. Therefore, there was a 100% predictive accuracy and zero misclassification of the positive class. The overall predictive accuracy was 99.995%, with a 0.005% misclassification rate. The UCI Credit Card testing dataset contained a sample size of 1 395 instances. There were 711 negatives cases and 684 positive cases. The confusion matrix revealed that there were 631 TN, 89 FN, 595 TP, and 80 FP. The predictive accuracy was 89%, with a misclassification rate of 11% for the negative class.

The predictive accuracy was 87%, with a misclassification rate of 13% for the positive class. The overall predictive accuracy was 88%, with a 12% misclassification rate. The average predictive accuracy was 94%, with a 6% misclassification rate.

4.3.2.2 Precision-Recall curve(s) with SMOTe

Figure 4.18 to Figure 4.21 plot the respective P-R curves of the classification results after using SMOTe to resample the datasets. The goal was to validate if there has been improvements on the accuracy by observing whether the P-R curve is moving towards the upper right corner of the chart. The quality of a P-R curve is good when the line is closer to the value of 1 on the y-axes.

Figure 4.18 illustrates the P-R curve of the SVM classification where there was a 0.53 average precision.

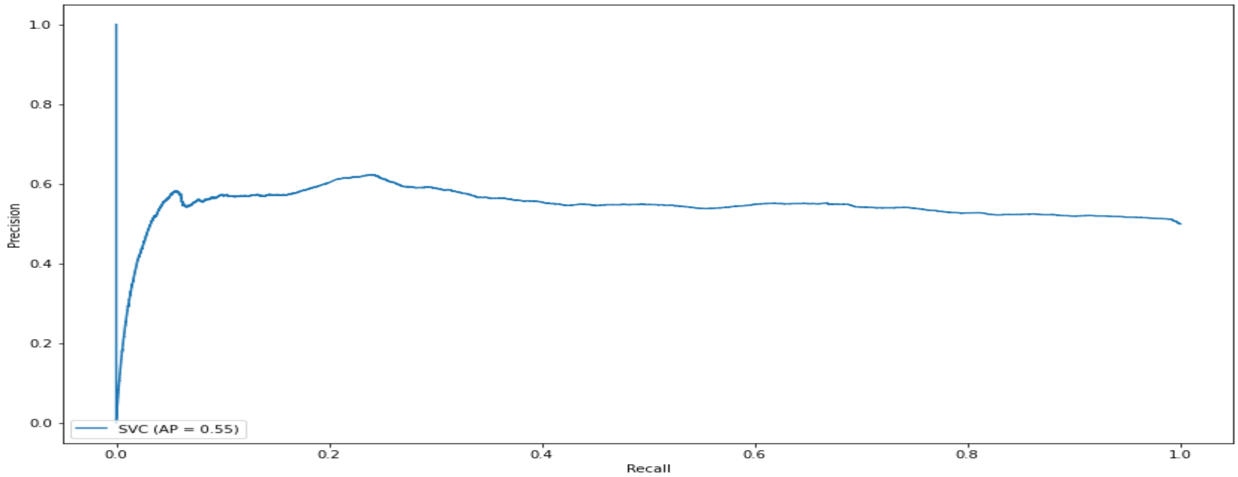

Figure 4.18: SVM with SMOTe P-R curve, AP = 0.53

Figure 4.19 shows the P-R curve of the LR classification where the average precision computed was 0.96.

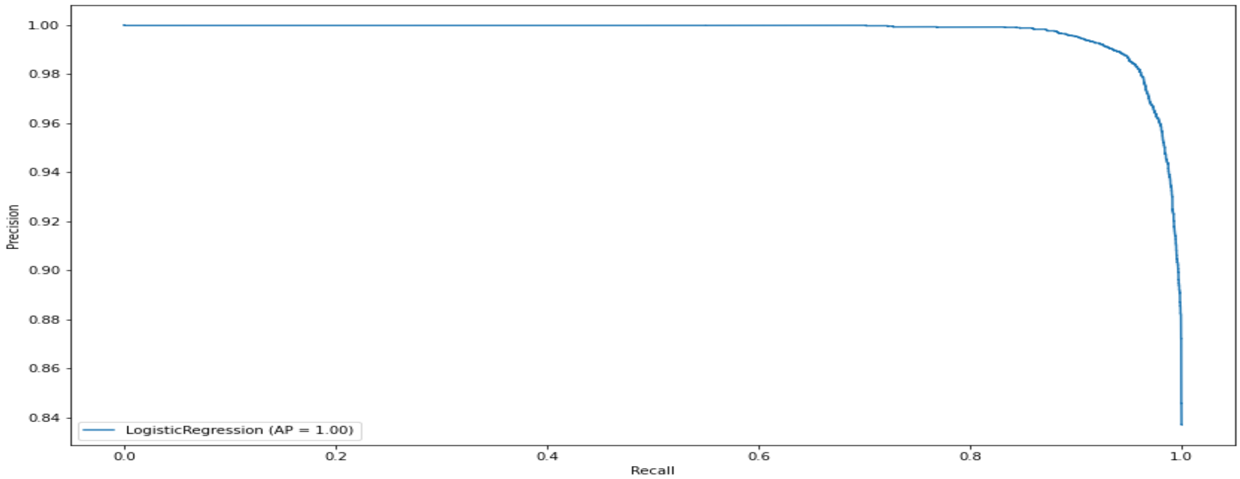

Figure 4.19: LR with SMOTe P-R curve, AP = 0.96

Figure 4.20 shows the P-R curve of the DT classification where the average precision computed was 1.00.

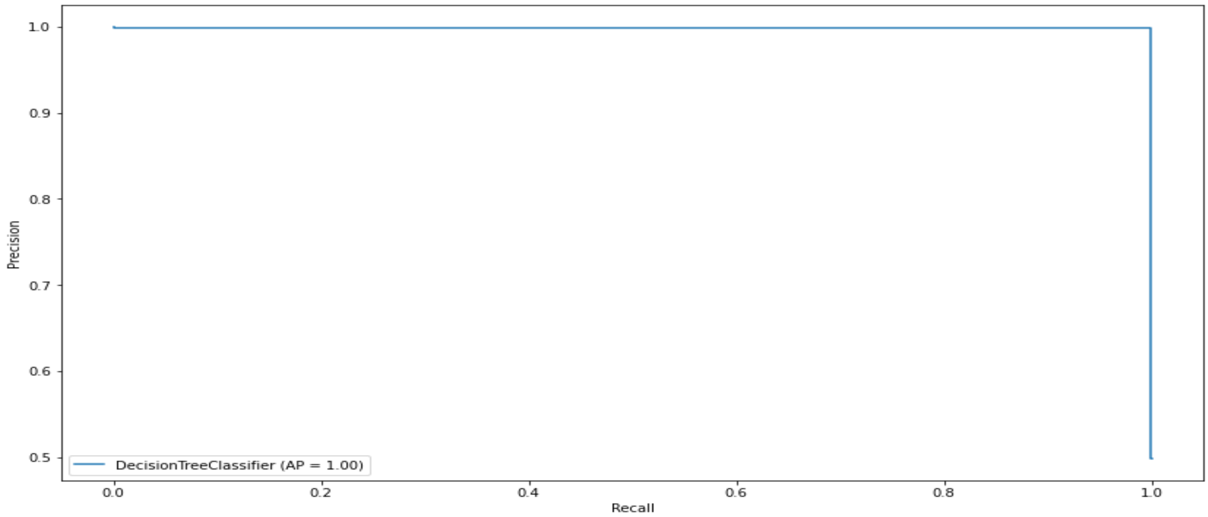

Figure 4.20: DT with SMOTe P-R curve, AP = 1.00

Figure 4.21 shows the P-R curve of the RF classification where the average precision computed was 1.00.

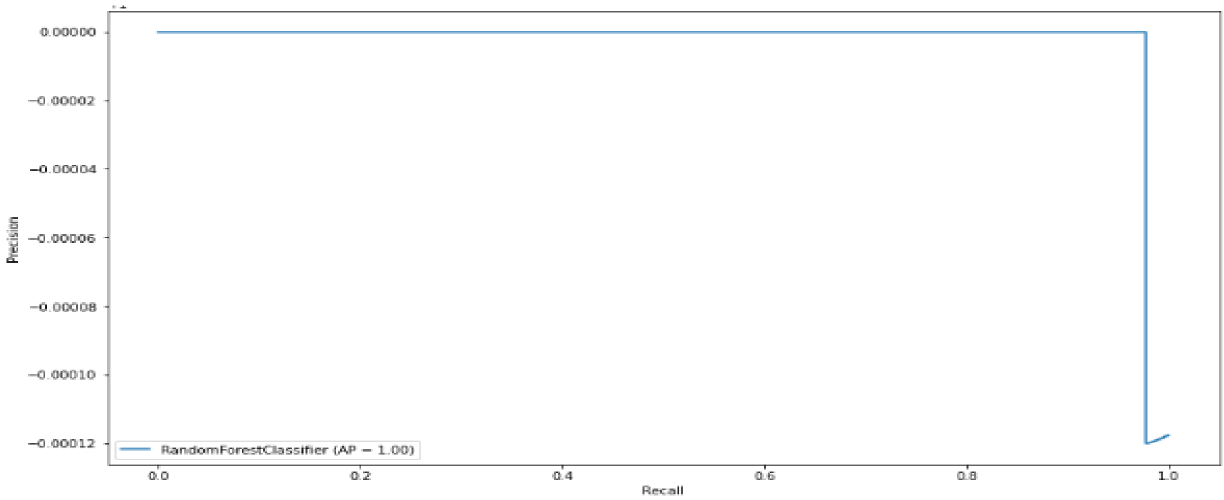

Figure 4.21: RF with SMOTe P-R curve, AP = 1.00

As previously mentioned, a P-R curve is an effective means of providing a graphical visualisation of the quality of a classifier. A P-R curve that is a straight line towards the upper right corner, such as the one of Figure 4.20 and Figure 4.21, represents the best possible quality. These two P-R curves tell us that the classifiers were able to predict the positive classes at a 100% accuracy.

4.4 Results Comparison with Literature

This section presents the existing related results in the literature, which are compared and contrasted with the results obtained from this study. The results presented below were extracted from a study by Elhassan *et al*. (2017) that measured the performance of different sampling techniques on the Ecoli2 dataset. The study proposed a hybrid approach combining Tomek Link (T-Link) with Random Undersampling (RUS) to improve the classification of imbalance data. To provide a fair comparison, all our finding discussed in this section were achieved using our proposed multi-level hybrid data-point approach combining Feature Selection with SMOTe based Oversampling.

Table 4.20 Performance measures of different sampling techniques using SVM (Elhassan et al. 2017)

Support Vector Machine	Original	T-Link	SMOTE	SMOTE/T-Link	Over-sampling	Over/T-Link	Under-sampling	under/T-Link
Recall	0.577	0.673	0.929	0.91	0.931	0.883	0.942	0.923
Accuracy	0.893	0.918	0.923	0.904	0.913	0.882	0.913	0.923
Precision	0.682	0.778	0.895	0.871	0.899	0.883	0.891	0.923
F-statistics	0.625	0.722	0.912	0.89	0.915	0.883	0.916	0.923
G-mean	0.741	0.805	0.924	0.905	0.913	0.882	0.913	0.923
AUC	0.913	0.922	0.945	0.953	0.932	0.934	0.96	0.95

Table 4.20 presents the results obtained by Elhassan *et al*. (2017) using Support Vector Machine (SVM). The best results were achieved by the combination of Undersampling and T-Link. The ability to predict positive classes improved, the precision score increased by 0.241, moving from 0.682 to 0.923. The improvement rate was 26.11%, which is less than our improvement rate of 55% achieved using SVM after applying hybrid solution. The recall score increased from 0.577 to 0.923, an improvement of 37.49% versus the 76% we achieved with our method, which means that our solution produced better results in terms of improvement.

Table 4.21 Performance measures of different sampling techniques using Logistic Regression (Elhassan et al. 2017)

Logistic Regression	Original	TLink	SMOTE	SMOTE/TLink	Over-samplig	Over/TLink	Under-sampling	Under-/TLink
Recall	0.615	0.615	0.923	0.91	0.924	0.906	0.92	0.923
Accuracy	0.911	0.905	0.909	0.904	0.908	0.893	0.903	0.923
Precision	0.762	0.744	0.873	0.871	0.895	0.883	0.888	0.923
F-statistics	0.681	0.674	0.897	0.89	0.909	0.894	0.905	0.923
G-mean	0.771	0.769	0.911	0.905	0.907	0.893	0.903	0.923
AUC	0.934	0.92	0.949	0.95	0.935	0.936	0.954	0.95

Table 4.21 presents the results obtained by Elhassan *et al*. (2017) using Logistic Regression (LR). Our finding reported a score of 0.97 for all the metrics except the average precision, which achieved a score of 0.96 when using LR. Whereas, Elhassan *et al*. (2017) reported a best score of 0.923 for all metrics except the AUC, which achieved 0.95. Based on this figures, our hybrid solution performed better than their best method of using the combination of Undersampling and T-Link.

Table 4.22 Performance measures of different sampling techniques using Random Forest (Elhassan et al. 2017)

Random Forest	Original	TLink	SMOTE	SMOTE/TLink	Over-samplig	Over/TLink	Under-sampling	Under-/TLink
Recall	0.808	0.827	0.91	0.923	0.906	0.899	0.904	0.923
Accuracy	0.955	0.966	0.94	0.962	0.91	0.893	0.933	0.933
Precision	0.894	0.956	0.947	0.986	0.913	0.889	0.959	0.941
F-statistics	0.848	0.887	0.928	0.954	0.909	0.894	0.931	0.932
G-mean	0.891	0.906	0.936	0.956	0.91	0.893	0.932	0.933
AUC	0.96	0.97	0.98	0.99	0.97	0.972	0.96	0.983

Table 4.21 presents the results obtained by Elhassan *et al*. (2017) using Random Forest (RF). According to table 4.22, the best method was a combination of SMOTe and T-Link, which reported 0.923 for Recall, 0.986 for precision, and 0.962 for accuracy. The RF classification after applying the combination of Feature Selection and SMOTe reported a perfect score of 1.00 for all the metrics. Based on these results, our conclusion is that our solution has a better performance across all the different algorithms used in both studies.

Our finding comes along with the finding by Wing *et al.* (2015), where they showed that using a combine sampling method has an improved performance in terms of the performance measures that have been mentioned in this study using classification algorithm. By looking at (Table 4.22), the data showed that using SMOTE and T-Link as a combined sampling method has a better performance over standalone sampling methods and the original data. However, compared to all the methods in (Table 4.22), our study showed the best results with our proposed novel multi-level hybrid data-point approaches.

4.5 Chapter Summary

The comparison of Near Miss and SMOTe was done based on the average predictive accuracy score because the predictive accuracy is based on both the TN and TP rate and it is inversely proportional to the misclassification rate. The performance difference measures the consistency of the classifier of both datasets. To ensure that the findings are easy to read and understand, a simple comparison format was developed and used for all four of the algorithms. The average predictive accuracy was 81% for the SVM classifier with Near Miss and was 58.75% for the SVM classifier with SMOTe solution. The performance difference was 15% for Near Miss and 4.5% for SMOTe. Based on these results, the Near Miss technique performed better, with a higher predictive accuracy and low performance difference.

The average predictive accuracy was 83.63% for the LR classifier with Near Miss was and 80% for the LR classifier with SMOTe solution. The performance difference was 19.75% for Near Miss and 33% for SMOTe. Based on these results, the Near Miss technique performed better with a higher predictive accuracy and low performance difference. The average predictive accuracy was 92% for the DT classifier with Near Miss and 89.15% for the DT classifier with SMOTe solution. The performance difference was 15.5% for Near Miss and 21.5% for SMOTe. Based on these results, the Near Miss technique performed better with a higher predictive accuracy and low performance difference.

The average predictive accuracy was 90% for the RF classifier with Near Miss and was 94% for the RF classifier with SMOTe solution. The performance difference was 0% for Near Miss and 12% for SMOTe. Based on these results, the Near Miss technique performed better with a higher predictive accuracy and low performance difference. These findings suggest that the proposed RF solution with the hybrid data-point approach of Feature Selection and Undersampling with Near Miss produced the best classification results. The RF classifier with SMOTe based Oversampling also produced impressive results but the Near Miss technique was more consistent across the two credit card datasets.

The classification results and analysis of the SVM, LR, DT, and RF before and after applying the data-point approach have been presented. The findings revealed that before any data-point approach was applied, all the classifiers struggled to predict the positive class and the classification was highly biased towards the majority class. There was a high misclassification rate and low predictive accuracy. After the data-point approach was used, the ability to predict the positive class improved and there was minimum bias. The misclassification rate was reduced and the predictive accuracy increased.

CHAPTER FIVE: SUMMARY, CONCLUSIONS AND IMPLICATIONS OF STUDY

5.1 Introduction

Chapter Five is the final chapter of the study that has the following sections: a summary of how the research objectives were met, a discussion of future works, and the conclusion of the study. Furthermore, the summary section compares the Near Miss and SMOTe technique and discusses the results to identify the most effective technique.

5.2 Summary

This section provides a detailed discussion of the research outcome and how the objectives of the study were met. The detailed discussion compares the classification performance of each algorithm for both the pre-test and post-test results to assert whether the accuracy has improved. Then, a cross comparison of the two datasets was conducted to validate the two proposed data-point approaches and to determine the best performing algorithm across both datasets.

The first research objective was to examine the ML algorithms for fraud identification with datasets for Credit Card Fraud (CCF) that were imbalanced. To facilitate the achievement of this objective, a pre-treatment-test was conducted to assess the performance of four ML classifiers on two imbalanced credit card datasets. The ML classifiers were each constructed using the SVM, LR, DT, and RF algorithms respectively. All four algorithms scored an average accuracy score of 1.00 for legitimate cases with the European Credit Card dataset and an average accuracy score of 0.87 with the UCI Credit Card dataset.

The weighted average for the results from the classification with both the imbalanced datasets showed that the RF model was the best performer for detecting minority classes, with an AP score of 0.77 and a 0.45 average recall score. Therefore, comparing both precision and recall scores reveals that the model did not perform well. Furthermore, the combined computed average precision across both datasets was 0.43, which reveals that the model was not performing ideally and further treatment was necessary. The worst performing model was the SVM classifier, having the positive class precision and recall scores of 0.00 for both datasets. The score of 0.00 means that the SVM model was 100% biased and completely failed to identify any minority classes due to the imbalanced class distribution. The discussed precision and recall scores show that due to the imbalance level, the majority class was dominant over the minority class.

The second research objective was to investigate in-depth, the data-point approach methods for handling imbalanced datasets. To facilitate the fulfilment of this objective, a review of three data-point approach methods was carried out, namely, Feature Selection, Undersampling, and Oversampling. The post-treatment-test experiment led to the introduction of two novel multi-level hybrid approaches for the Credit Card Fraud domain. The first approach was the combination of Feature Selection and Undersampling with Near Miss. The second approach was the combination of Feature Selection and SMOTe based Oversampling. These two approaches addressed the issues of class imbalance and reduced misclassification by identifying the key feature based on feature importance and then used the Near Miss and SMOTe technique to resample the dataset, respectively.

The third and last research objective was to determine whether the data handling methods improved the predictive accuracy of ML Credit Card Fraud identification models through extensive comparative and statistical analysis of the results. After using the Feature Selection and Undersampling with Near Miss hybrid approach, the average precision score improved by 98% for SVM, 49.5% for DT, 19.5% for RF, and 5.5% for LR for the positive class. The Recall score shows that the strength of identifying TP, which refers the actual fraudulent cases, improved by 60% for SVM, 51.5% for LR, 51% for DT, and 38.5% for RF for the positive class. The results revealed that the F1-Score improved by 73.5% for SVM, 52.5% for LR, 50.5% for the DT, and 32.5% for RF for the positive class. Comparing the F1 scores shows that the ability to detect positive classes was improved.

The findings also suggested that predicting accuracy improved for all the algorithms on both the datasets. Based on these results, the RF algorithm was the leading algorithm, followed by DT, LR and SVM, ranked from best to worst using a calculated average score of the precision, recall, and F1 score for each classifier. After using the Feature Selection and SMOTe based Oversampling hybrid approach, the precision score improved by 55% for SVM, 55% for LR, 42% for the DT, and 10% for RF for the positive class. The Recall score shows that the strength of identifying TP improved by 76% for SVM, 50% for LR, 47% for RF, and 39% for the DT for the positive class. The results revealed that the F1-Score improved by 64% for SVM, 53%

for LR, 35% for the DT, and 33% for RF for the positive class. Comparing the F1 scores shows that the ability to detect positive classes was improved.

An interesting observation was that the classification of negative class for the LR, DT and RF algorithms remained good and consistent throughout the experiment. SVM performed well initially with the overall accuracy score of 100%; however, after using Feature Selection and SMOTe based Oversampling hybrid approach, the score was down to 47%, meaning that even though the ability to recognise positive classes improved, the ability to recognise negative classes degraded. Therefore, SVM is not an ideal solution for Credit Card Fraud Detection. Based on the results, the RF algorithm is the leading algorithm. The ranking from best to worst was in the following order: RF, DT, LR, and SVM, which was identical to the findings obtained after using the Feature Selection and Undersampling with Near Miss hybrid approach.

The findings are also supported by literature that used RF and Oversampling in other domains outside of Credit Card Fraud (Tarawneh *et al.* 2020; Saad, Omar, and Maghraby 2019; Xiaolong, Wen, and Yanfei 2019). To improve the performance of the training, Zhang *et al.* (2019) presented a classification solution constructed with an Oversampling scheme combined with RF. An Oversampling scheme with capability to recreate the data of hotspots was used to resample and balance the specified training dataset. The RF algorithm was then applied to produce a group of forest trees for the oversampled training dataset. The final prediction performance can be computed recursively after the Oversampling and training process.

The proposed method had the ability to select features randomly and build a robust RF algorithm to prevent overfitting the training dataset. Experimental results from three datasets indicate that the performance of hotspot predictions was significantly improved compared to existing classification methods.

5.3 Conclusion

To facilitate the achievement of the aim, an experimental study was carried out that answered the research question defined in Chapter One and to ensure that all the research objectives were met. The output of the first research objective revealed that, when the dataset is highly imbalanced, the classifier struggles to detect the minority class. In association with the second research objective to investigate in-depth the data-point level methods of handling imbalanced datasets, the experiment resulted in the introduction of two hybrid solutions as output, namely, the Feature Selection with Near Miss Undersampling and Feature Selection with the SMOTe based Oversampling hybrid approaches. Using these two approaches to solve the imbalance problem, the third research objective was met through statistical and extensive analysis of results from the pre-treatment-test and post-treatment-test. The output of the third research objective revealed that there was a significant improvement in the ability to predict positive classes while maintaining the ability to detect negative classes.

The misclassification rate was also reduced. Based on the findings, the data-point approach works well with the DT and RF based classifiers. Both the algorithms produced the best performance with the two credit card datasets. After using the hybrid data-point approach, the ability to predict positive classes improved. Furthermore, based on the analysis of the results, the conclusion was that Near Miss out-performed SMOTe on all four of the classification algorithms.

5.4 Future Work

The significance of this research was that the study investigated the data-point level approach to address the misclassification problem in identifying credit card fraud using ML and proposed an effective method that is more effective in handling imbalanced credit card datasets. Two multi-level hybrid data-point approach solutions have been introduced, which were Feature Selection and Undersampling with Near Miss, and Feature Selection and SMOTe based Oversampling. This study focused on the detection of fraud; however, the next step is the prevention of fraud that requires a time-critical and proactive solution, which can accurately flag fraudulent and legitimate classes in a live environment. Therefore, further studies can use the findings of this study to investigate building and deploying a real-time solution that can detect fraud as and when the transactions are occurring, to improve its prevention.

5.5 Limitations of Study

The credit card transactions data stored by financial institutions is confidential and is protected by data and privacy laws and regulations, which makes them inaccessible to the public. The study had to rely on publicly available credit card datasets to carry out the experiment. The number of credit card datasets available is very limited, which consequently limited the number of credit card datasets used in the study. Another limitation was the time available to conduct the research, which had to be within the time allowed to complete a Master's degree. Therefore, the scope of study was also limited by the time available.

5.6 Contributions and Implications of Study

Associated with the aim to identify the data-point technique that improves the predictive accuracy of ML solutions for fraud detection, the contributions of the study were extending knowledge regarding the use of ML and the data-point approach in Credit Card Fraud Detection. The findings of this study contributed to the knowledge base by providing an in-depth understanding of the impact of imbalanced datasets on ML classification solutions for Credit Card Fraud Detection. The study also provided knowledge of how the data-point approach could be used to address the class imbalance problem.

Another contribution was that based on the findings, an effective method that is more effective in handling imbalanced credit card datasets was proposed. The implications of the study involved having two solutions as research output and knowledge that constitute the base for effective and efficient methods of solving the imbalanced credit card dataset problem. As a result, financial institutions and researchers can use the findings to develop and improve fraud detection systems that are unbiased and more accurate when dealing with imbalanced credit card datasets.

5.7 Chapter Summary

The statistical analysis of results has herein been presented and discussed in this chapter in relation to the research objectives. The findings show that, the strength of using the hybrid data-point approach to improve the algorithm performance was observed with both Near Miss and SMOTe. This study has met all the research objectives and proven that using the most appropriate data handling technique for the imbalanced credit card datasets makes a significant different towards combating the class imbalance problem, minimising bias, and reducing the number of FP and FN. After using the hybrid data-point approach, the ability to predict positive classes improved. The ranking from best to worst was in the following order: RF, DT, LR and SVM for both the introduced hybrid approaches. Future research can perform an in-depth cross comparison across multiple credit card datasets to verify the consistency of the data-point approaches in handling imbalanced datasets. Further studies can use the findings of this study to investigate the building and deployment of a real-time solution that can detect fraud as and when the transaction is occurring.

REFERENCES

Abdoh, S. F., Abo M. and Maghraby, F. A. 2018. Cervical cancer diagnosis using random forest classifier with SMOTe and feature reduction techniques. *IEEE Access*, 6(1): 59475-59485. Available: https://doi.org/10.1109/ACCESS.2018.2874063 (Accessed 02 May 2019).

Abdulsalam, H., Skillicorn, D. B. and Martin, P. 2011. Classification using streaming random forests. *IEEE Transactions on Knowledge and Data Engineering*, 23(1): 22-36. Available: https://doi.org/10.1109/TKDE.2010.36 (Accessed 24 October 2020).

Adedoyin, A. 2016. Predicting fraud in mobile money transfer. *IEEE Transactions on Knowledge and Data Engineering*, 28(6): 1- 203.

Adewumi, A. O. and Akinyelu, A. A. 2017. A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of System Assurance Engineering and Management*, 8(2): 937-953. Available: https://doi.org/10.1007/s13198-016-0551-y (Accessed 03 July 2019).

Agrawal, A., Viktor, H. L. and Paquet, E. 2015. November. SCUT: Multi-class imbalanced data classification using SMOTe and cluster-based Undersampling. In: *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Lisbon, 12-14 November 2015. Lisbon: IC3K, 226-234.

Albashrawi, M. 2016. Detecting financial fraud using data mining techniques: a decade review from 2004 to 2015. *Journal of Data Science*, 14(1): 553-570.

Askari, S. and Hussain, A. 2017. Credit card fraud detection using Fuzzy ID3. *Computing, Communication and Automation (ICCCA)*, 40(1): 446-452. Available: https://doi.org/10.1109/ccaa.2017.8229897 (Accessed 03 July 2019).

Baader, G. and Krcmar, H. 2018. Reducing false positives in fraud detection: combining the red flag approach with process mining. *International Journal of Accounting Information Systems*, 31(1): 1–16. Available: https://doi.org/10.1016/j.accinf.2018.03.004 (Accessed 03 July 2019).

Bader-El-Den, M., Teitei, E. and Perry, T. 2019. Biased random forest for dealing with the class imbalance problem. *IEEE Transactions on Neural Networks and Learning Systems*, 30(7): 2163-2172. Available: https://doi.org/10.1109/TNNLS.2018.2878400 (Accessed 24 July 2020).

Bao, F., Deng, Y. and Dai, Q. 2018. ACID: Association correction for imbalanced data in GWAS. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(1): 316-322.

Batuwita R. and Palade, V. 2010. FSVM-CIL: Fuzzy support vector machines for class imbalance learning. *IEEE Transactions on Fuzzy Systems*, 18(3): 558-571. Available: https://doi.org/10.1109/TFUZZ.2010.2042721 (Accessed 04 April 2020).

Bauder, R. A. and Khoshgoftaar, T. M. 2018. The effects of varying class distribution on learner behaviour for Medicare fraud detection with imbalanced Big Data. *Health Information Science and Systems*, 6(9): 1-14. Available: https://doi.org/10.1007/s13755-018-0051-3 (Accessed 03 July 2019).

Bian, Y., Cheng, M., Yang, C., Yuan, Y., Li, Q., Zhao J. L. and Liang, L. 2016. Financial fraud detection: a new ensemble learning approach for imbalanced data. *20th Pacific Asia Conference on Information Systems (PACIS 2016)*. Nice Prince Hotel, Chiayi, 27 June - 1 July 2016. Taiwan: Pacific Asia Conference on Information Systems, 1-11.

Candanedo, I. S., Nieves, E. H., González, S. R., Martín, M. T. S. and Briones, A. G. 2018. Machine learning predictive model for industry 4.0. *Knowledge Management in Organizations*, 1(877): 501-510.

Cano, A., Zafra, A. and Ventura, S. 2013. Weighted data gravitation classification for standard and imbalanced Data. *IEEE Transactions on Cybernetics*, 43(6): 1672-1687.

Chen, D., Wang, X., Zhou, C. and Wang, B. 2019. The distance-based balancing ensemble method for data with a high imbalance ratio. *IEEE Access*, 7(1): 68940- 68956.

Cheng, K., Zhang, C., Yu, H., Yang, X., Zou, H. and Gao, S. 2019. Grouped SMOTe with noise filtering mechanism for classifying imbalanced data. *IEEE Access*, 7(1): 170668-170681. Available: https://doi.org/10.1109/access.2019.2955086 (Accessed 02 May 2020).

Chouiekha, A., Hassane, E., and Haj, E. 2018. ConvNets for Fraud Detection analysis. *Procedia Computer Science*, 127(1): 133–138. Available: https://doi.org/10.1016/j.procs.2018.01.107 (Accessed 24 July 2020).

de la Fuente, C., Urrutia, A. and Chávez, E. 2019. Using the random forest algorithm for searching behaviour patterns in electronic health records. *IEEE Latin America Transactions*, 17(05): 875-881. Available: https://doi.org/10.1109/TLA.2019.8891957 (Accessed 24 July 2020).

Ditzler, G. and Polikar, R. 2013. Incremental learning of concept drift from streaming imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 25(10): 2283-2301.

Dong, Y., Du, B. and Zhang, L. 2015. Target detection based on random forest metric learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4): 1830-1838. Available: https://doi.org/10.1109/JSTARS.2015.2416255. (Accessed 14 May 2020).

Douzas, G., Bacao, F. and Last, F. 2018. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTe. *Information Sciences*, 465(1): 1-20.

Ebenuwa, S. H., Sharif, S. and Alazab, M. 2019. Variance ranking attributes selection techniques for binary classification problem in imbalance data. *IEEE Access*, 7(1): 24649-24666.

Elhassan, A. T., Aljourf, M., Al-Mohanna, F., and Shoukri, M. 2017. Classification of imbalance data using Tomek Link (T-Link) combined with Random Under-sampling (RUS) as a data reduction method. *Global Journal of Technology and Optimization*, 1(1): 1-11. Available: https://doi.org/10.4172/2229-8711.S1111 (Accessed 03 July 2019).

Fang, L., Ma, K., Li, R., Wang, Z. and Shi, H. 2019. A statistical approach to estimate imbalance-induced energy losses for data-scarce low voltage networks. *IEEE Transactions on Power Systems*, 34(4): 2825- 2835.

Favieiro, G. W. and Balbinot, A. 2019. Paraconsistent Random Forest: an alternative approach for dealing with uncertain data. *IEEE Access*, 7: 147914-147927. Available: https://doi.org/10.1109/ACCESS.2019.2946256 (Accessed 24 July 2020).

Feldt, L. S. 1958. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, 23(1): 335-353. Available: https://doi.org/10.1007/BF02289783 (Accessed 03 July 2019).

Feng, W., Huang, W., and Bao, W. 2019. Imbalanced hyperspectral image classification with an adaptive ensemble method based on SMOTe and rotation forest with differentiated sampling rates. IEEE Geoscience and Remote Sensing Letters, 16(12): 1879-1883. Available: https://doi.org/10.1109/LGRS.2019.2913387 (Accessed 02 May 2020).

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F. 2012. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4): 463-484, Available: https://doi.org/10.1109/TSMCC.2011.2161285 (Accessed 04 April 2020).

Ge, Y., Yue, D. and Chen, L. 2017. Prediction of wind turbine blades icing based on MBK-SMOTe and random forest in imbalanced data set. *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, Beijing, 2017): 1-6, Available: https://doi.org/10.1109/EI2.2017.8245530 (Accessed 04 August 2020).

Ghorbani, R. and Ghousi, R. 2020. Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access*, 8(1): 67899-67911. Available: https://doi.org/10.1109/ACCESS.2020.2986809 (Accessed 02 July 2020).

Gomaa, A. and Jalloul, L. M. A. 2014. Data-Aided I/Q Imbalance Estimation and Compensation in OFDM Systems. *IEEE Communications Letters*, 18(3): 459- 462.

Guo, Y., Zhou, Y., Hu, X. and Cheng, W. 2019. Research on recommendation of insurance products based on random forest. *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2019*: 308-311, Available: https://doi.org/10.1109/MLBDBI48998.2019.00069 (Accessed 24 July 2020).

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. and Seliya, N. 2019. Investigating random Undersampling and feature selection on bioinformatics big data. *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (Big Data Service), Newark, CA, USA, 2019*: 346-356, Available: https://doi.org/10.1109/BigDataService.2019.00063 (Accessed 02 May 2020).

Hassib, E. M., El-Desouky, A. I., El-Kenawy E. M. and Ghamrawy, S. M. 2019. Imbalanced big data mining framework for improving optimization algorithms performance. *IEEE Access*, 7(1): 170774-170795.

Hi, Z. and Jiang, P. 2019. Imbalance modified deep neural network with dynamical incremental learning for chemical fault diagnosis. *IEEE Transactions on Industrial Electronics*, 66(1): 540-550.

Huang, D., Mu, D., Yang, L. and Cai, X. 2018. CoDetect: financial fraud detection with anomaly feature detection. *National Natural Science Foundation of China*, 6(2): 19161-19174. Available: https://doi.org/10.1109/access.2018.2816564 (Accessed 15 June 2019).

Hukerikar, S., Tumma, A., Nikam, A. and Attar, V. 2011. SkewBoost: an algorithm for classifying imbalanced datasets. *2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011).* Allahabad, India, 15-17 September 2011. IEEE, *46-52.* Available: https://doi.org/10.1109/iccct.2011.6075185 (Accessed 10 January 2020).

Imbalanced-learn. 2020. *Imbalanced-learn documentation*. Available: https://imbalanced-learn.readthedocs.io/en/stable (Accessed 15 June 2019).

Jeon, Y. and Lim, D. 2020. PSU: Particle Stacking Undersampling method for highly imbalanced big data. *IEEE Access*, 8: 131920-131927. Available: https://doi.org/10.1109/ACCESS.2020.3009753 (Accessed 10 November 2020).

Jia, D., Li, Z. and Zhang, C. 2020. A parametric optimization oriented, AFSA based random forest algorithm: application to the detection of cervical epithelial cells. *IEEE Access*, 8: 64891-64905. Available: https://doi.org/10.1109/ACCESS.2020.2984657 (Accessed 03 October 2020).

Jiang, C., Song, J., Liu, G., Zheng, L. and Luan, W. 2018. Credit card fraud detection: a novel approach using aggregation strategy and feedback mechanism. *IEEE Internet of Things Journal*, 5(5): 3637-3647. Available: https://doi.org/10.1109/jiot.2018.2816007 (Accessed 23 March 2019).

Jiang, K., Lu, J. and Xia, K. 2016. A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTe. *Arabian Journal for Science and Engineering*, 41(8): 3255-3266. Available: https://doi.org/10.1007/s13369-016-2179-2 (Accessed 23 March 2019).

Khoshgoftaar, T. M., Hulse, J. V. and Napolitano, A. 2010. Supervised neural network modelling: an empirical investigation into learning from imbalanced data with labelling errors. *IEEE Transactions on Neural Networks*, 21(5): 813-830.

Khosravi, I. and Jouybari-Moghaddam, Y. 2019. Hyperspectral imbalanced datasets classification using filter-based forest methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(12): 4766-4772. Available: https://doi.org/10.1109/JSTARS.2019.2914668 (Accessed 03 October 2020).

Kotsiantis, S., Kanellopoulos, D. and Pintelas P. 2006. Handling imbalanced datasets: a review. *GESTS International Transactions on Computer Science and Engineering*, 30(1): 25-36. Available: https://doi.org/10.1007/s13369-016-2179-2 (Accessed 23 March 2019).

Krstic, M. and Bjelica, M. 2015. Impact of class imbalance on personalized program guide performance. *IEEE Transactions on Consumer Electronics*, 61(1): 90-95. Available: https://doi.org/10.1109/TCE.2015.7064115 (Accessed 24 October 2020).

Kubat, M. and Matwin, S. 1997. Addressing the curse of imbalanced training sets: one-sided selection. *ICML*, 97(1): 179-186.

Kumar, M. S., Soundarya, V., Kavitha, S., Keerthika, E. S. and Aswini, E. 2019. Credit card fraud detection using random forest algorithm. *2019 3rd International Conference on Computing and Communications Technologies (ICCCT), Chennai, India, 2019*: 149-153, Available: https://doi.org/10.1109/ICCCT2.2019.8824930 (Accessed 24 July 2020).

Lejon, E., Kyosti, P. and Lindstrom, J. 2018. Machine learning for detection of anomalies in press hardening: selection of efficient methods. *Process IT Innovations R&D Centre*, 1(1): 1079-1083. Available: https://doi.org/10.1016/j.procir.2018.03.221 (Accessed 23 March 2019).

Li, J., He, H, and Li, L. 2019. CGAN-MBL for reliability assessment with imbalanced transmission gear data. *IEEE Transactions on Instrumentation and Measurement*, 68(9): 3173-3183.

Lin, C., Deng, D., Kuo, C. and Chen, L. 2019. Concept drift detection and adaption in big imbalance industrial IoT data using an ensemble learning method of offline classifiers. *IEEE Access*, 7(1): 56198-56207. Available: https://doi.org/10.1.1.309.1465 (Accessed 23 January 2020).

Lin, W., Wu, Z., Lin, L., Wen, A. and Li, J. 2017. An ensemble random forest algorithm for insurance big data analysis. *IEEE Access*, 5: 16568-16575. Available: https://doi.org/10.1109/ACCESS.2017.2738069 (Accessed 24 October 2020).

Liu, X., Wu, J. and Zhou, Z. 2009. Exploratory Undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2): 539-550. Available: https://doi.org/10.1109/TSMCB.2008.2007853 (Accessed 24 November 2019).

Liu, Y., Liu, L., Gao, Y. and Yang, L. 2019. An improved random forest algorithm based on attribute compatibility. *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 2019*: 2558-2561, Available: https://doi.org/10.1109/ITNEC.2019.8729146 (Accessed 24 July 2020).

Liu, Z., Wen, T., Sun, W. and Zhang, Q. 2020. Semi-supervised self-training feature weighted clustering decision tree and random forest. *IEEE Access*, 8: 128337-128348. Available: https://doi.org/10.1109/ACCESS.2020.3008951 (Accessed 24 October 2020).

Malini, N. and Pushpa, M. 2017. Analysis on credit card fraud identification techniques based on KNN and outlier detection. *Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 1(1): 1-12. Available: https://doi.org/10.1109/aeeicb.2017.7972424 (Accessed 23 March 2019).

Mathew, J., Pang, C. K., Luo, M. and Leong, W. H. 2018. Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 29(9): 4065-4076. Available: https://doi.org/10.1109/TNNLS.2017.2751612 (Accessed 12 April 2019).

Minku, L. L., Wang, S. and Yao, X. 2016. Online ensemble learning of data streams with gradually evolved classes. *IEEE Transactions on Knowledge and Data Engineering*, 28(6): 1532- 1545.

Mqadi, N., Naicker, N., and Adeliyi, T. 2021. A SMOTe based Oversampling Data-Point Approach to Solving the Credit Card Data Imbalance Problem in Financial Fraud Detection. International Journal of Computing and Digital Systems, 10(1): 277:286. DOI: http://dx.doi.org/10.12785/ijcds/100128

Mqadi, N., Naicker, N., and Adeliyi, T. 2021. Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss. Mathematical Problems in Engineering, 2021(1): 1:16. DOI: https://doi.org/10.1155/2021/7194728

Mubalik, A. and Adali, E. 2017. Multilayer perception neural network technique for fraud detection, *Computer Science and Engineering (UBMK)*, 1(1): 383-387.

Nilson Report. 2020. Card fraud losses reach $28.65 billion. Available: https://nilsonreport.com/mention/1313/1link/ (Accessed 12 March 2021).

Ng, W. W. Y., Hu, J., Yeung, D. S., Yin, S. and Roli, F. 2015. Diversified sensitivity-based Undersampling for imbalance classification problems. *IEEE Transactions on Cybernetics*, 45(11): 2402-2412, Nov. 2015, Available: https://doi.org/10.1109/TCYB.2014.2372060 (Accessed 14 April 2015).

Patil, S., Nemade, V. and Kumar, P. 2018. Predictive modelling for credit card fraud detection using data analytics. *Computational Intelligence and Data Science*, 132(1): 385–395. Available: https://doi.org/10.1016/j.procs.2018.05.199 (Accessed 23 March 2019).

Pes, B. 2020. Learning from high-dimensional biomedical datasets: the issue of class imbalance. *IEEE Access*, 8(1): 13527-13540.

Pozzolo, A. D., Boracchi, G., Caelen, O., Alippi, C. and Bontempi, G. 2018. Credit card fraud detection: a realistic modelling and a novel learning strategy. *Transactions on Neural Networks and Learning Systems,* 29(8): 3784-3797. Available: https://doi.org/10.1109/tnnls.2017.2736643 (Accessed 23 March 2019).

Raj, J. T. 2019. *What to do when your classification data is imbalanced*. Available online: https://towardsdatascience.com/what-to-do-when-your-classification-dataset-is-imbalanced-6af031b12a36 (Accessed on 10 September 2020).

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P. and Nandi, A. K. 2018. Credit card fraud detection using AdaBoost and majority voting. *IEEE Access,* 6(1): 14277-14284.

Rashad, A., Riaz, S. and Jiao, L. 2019. Semi-supervised deep fuzzy C-Mean clustering for imbalanced multi-class classification. *IEEE Access*, 7(1): 28100- 28112.

Rekha, G., Tyagi, A. K. and Krishna, R. V. 2019. Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method. *International Journal of Hybrid Intelligent Systems*, 15(2): 67–76. Available: https://doi.org/10.3233/his-190261 (Accessed 21 March 2019).

Robertson, D. 2016. *The Nelson Report*. October. Available online: http://www.nelsonreport.com/upload/content_promo/The_Nelson_Report_10-17-2016.pdf (Accessed 03 February 2019).

Saad, A. I., Omar, Y. M. K. and Maghraby, F. A. 2019. Predicting drug interaction with adenosine receptors using machine learning and SMOTe techniques. *IEEE Access*, 7(1): 146953-146963. Available: https://doi.org/10.1109/ACCESS.2019.2946314 (Accessed 12 March 2020).

South African Banking Risk Information Centre (Sabric). 2020. Annual crime stats. 2020. Available: https://www.sabric.co.za/media-and-news/press-releases/sabric-annual-crime-stats-2019/ (Accessed 22 July 2020).

Sadgali, I., Sael, N. and Benabbou, F. 2018. Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science,* 148(1): 45-54. Available: https://doi.org/10.1016/j.procs.2019.01.007 (Accessed 12 July 2019).

Scornet, E. 2016. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3): 1485-1500. Available: https://doi.org/10.1109/TIT.2016.2514489 (Accessed 24 October 2020).

Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V. and Napolitano, A. 2010. RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems*, 40(1): 185-197.

Shao, J. 1993. Linear model selection via cross-validation. *Journal of the American Statistical Association*, 88(422): 486-494.

Shen, A., Tong, R. and Deng, Y. 2007. Application of classification models on credit card fraud detection. *IEEE International Conference on Service-Oriented Systems*, 1(1): 1-4.

Shevchik, S. A., Saeidi, F., Meylan, B. and Wasmer, K. 2017. Prediction of failure in lubricated surfaces using acoustic time–frequency features and random forest algorithm. *IEEE Transactions on Industrial Informatics*, 13(4): 1541-1553. Available: https://doi.org/10.1109/TII.2016.2635082 (Accessed 24 July 2020).

Somasundaran, A. and Reddy, U. S. 2016. Data imbalance: effects and solutions for classification of large and highly imbalanced data. *Proc. of 1st. International Conference on Research in Engineering, Computers and Technology*, 25(10): 28-34.

Sotiris, K., Dimitris, K. and Panayiotis, P. 2016. Handling imbalanced datasets: a review. *GESTS International Transactions on Computer Science and Engineering*, 30(1): 1-12.

Su, C. and Hsiao, Y. 2007. An Evaluation of the Robustness of MTS for Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 19(10): 1321-1332.

Subudhi, S. and Panigrahi, S. 2015. Quarter-sphere support vector machine for fraud detection in mobile telecommunication networks. *Procedia Computer Science*, 48(1): 353-359.

Tarawneh, A. B., Hassanat, A., Almohammadi, K., Chetverikov, D. and Bellinger, C. 2020. SMOTeFUNA: Synthetic Minority Over-Sampling Technique Based on Furthest Neighbour Algorithm. *IEEE Access*, 8(1): 59069-59082, 2020, Available: https://doi.org/10.1109/ACCESS.2020.2983003 (Accessed 07 August 2020).

Tarjo, A. and Herawati, N. 2015. Application of Beneish M-Score models and data mining to detect financial fraud. *Social and Behavioural Sciences*, 211(1): 924 – 930.

Wang, B. X. and Japkowicz, N. 2010. Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, 20(1): 1-20. Available: https://doi.org/10.1007/s10115-009-0198-y (Accessed 04 April 2020).

Wang, H. and Jin, Y. 2020. A random forest-assisted evolutionary algorithm for data-driven constrained multi-objective combinatorial optimization of trauma systems. *IEEE Transactions on Cybernetics*, 50(2): 536-549. Available: https://doi.org/10.1109/TCYB.2018.2869674 (Accessed 03 October 2020).

Wang, S. and Yao, X. 2013. Using class imbalance learning for software defect prediction. *IEEE Transactions on Reliability*, 62(2): 434-443. Available: https://doi.org/10.1109/TR.2013.2259203 (Accessed 24 October 2020).

Wang, S., Minku, L. L. and Yao, X. 2015. Resampling-based ensemble methods for online class imbalance learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(5): 1356-1368.

Wang, Y., Simon, M., Bonde, P., Harris, B. U., Teuteberg, J. J., Kormos, R. L., and Antaki J. F. 2012. Prognosis of right ventricular failure in patients with left ventricular assist device based on decision tree with SMOTe. *IEEE Transactions on Information Technology in Biomedicine*, 16(3): 383-390. Available: https://doi.org/10.1109/titb.2012.2187458 (Accessed 02 May 2019).

Wasikowski, M. and Chen, X. 2010. Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 22(10): 1388-1400. Available: https://doi.org/10.1109/TKDE.2009.187 (Accessed 04 August 2020).

Wei, J., Lu, Z., Qiu, K., Li, P. and Sun, H. 2020. Predicting drug risk level from adverse drug reactions using SMOTe and machine learning approaches. *IEEE Access*, 8(1): 185761-185775. Available: https://doi.org/10.1109/ACCESS.2020.3029446 (Accessed 02 July 2020).

Weng, S., Chang, H. and Chiong, C. 2009. A DC-21 GHz low imbalance active Balun using Darlington cell technique for high speed data communications. *IEEE Microwave and Wireless Components Letters*, 19(11): 728-730.

West, J. and Bhattacharya, M. 2016. Some experimental issues in financial fraud mining. *Procedia Computer Science*, 80(1): 1734-1744.

Williams, D. P., Myers, V. and Silvious, M. S. 2009. Mine classification with imbalanced data. *IEEE Geoscience and Remote Sensing Letters*, 6(3): 528-532. Available: https://doi.org/10.1109/LGRS.2009.2021964 (Accessed 19 April 2019).

Wing, W. Y., Hu, J., Yeung, D. S., Yin, S., and Roli, F. 2015. Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems. *IEEE Transactions on Cybernetics*, 45(11): 2402-2412. Available: https://doi.org/10.1109/TCYB.2014.2372060

Wu, Q., Ye, Y., Liu, Y. and Ng, M. K. 2012. SNP selection and classification of genome-wide SNP data using stratified sampling random forests. *IEEE Transactions on Nano Bioscience*, 11(3): 216-227. Available: https://doi.org/10.1109/TNB.2012.2214232 (Accessed 14 May 2019).

Xiaolong, X., Wen, C. and Yanfei, S. 2019. Over-sampling algorithm for imbalanced data classification. *Journal of Systems Engineering and Electron*ics, 30(6): 1182-1191, Dec. 2019, Available: https://doi.org/10.21629/JSEE.2019.06.12 (Accessed 04 August 2020).

Xu-Yin, L., Jianxin, W. and Zhi-Hua, Z. 2009. Exploratory Undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 39(2): 539-550.

Yan, Y., Liu, R., Ding, Z., Du, X., Chen, J. and Zhang, Y. 2019. A parameter-free cleaning method for SMOTe in imbalanced Classification. *IEEE Access*, 7(1): 23537-23548. Available: https://doi.org/10.1109/access.2019.2899467 (Accessed 02 May 2020).

Yao, R., Li, J., Hui, M., Bai, L. and Wu, Q. 2020. Feature selection based on random forest for partial discharges characteristic set. *IEEE Access*, 8: 159151-159161. Available: https://doi.org/10.1109/ACCESS.2020.3019377 (Accessed 24 July 2020).

Zanin, M., Romance, M., Moral, S. and Criado, R. 2017. Credit card fraud detection through parenclitic network analysis. *IEEE Access*, 1(1): 1-8, 2017. Available: https://doi.org/10.1155/2018/5764370 (Accessed 03 April 2019).

Zareapoor, M. and Shamsolmoali, P. 2015. Application of credit fraud detection: based on bagging ensemble classifier. *Inter-science Institute of Management and Technology*, 48(1): 679-685.

Zhang, G., Wang, L., Duffy, A. P., Sasse, H., Di Febo, D., Orlandi, A. and Aniserowicz, K. 2014. Downsampled and Undersampled datasets in Feature Selective Validation (FSV). *IEEE Transactions on Electromagnetic Compatibility*, 56(4): 817-824. Available: https://doi.org/10.1109/TEMC.2014.2307355 (Accessed 24 November 2019).

Zhang, J., Zulkernine, M. and Haque, A. 2008. Random-forests-based network intrusion detection systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(5): 649-659. Available: https://doi.org/10.1109/TSMCC.2008.923876 (Accessed 24 July 2020).

Zhang, P. 1992. On the distributional properties of model selection criteria. *Journal of the American Statistical Association*, 87(419): 732–737.

Zhang, X., Lin, X., Zhao, J., Huang, Q. and Xu, X. 2019. Efficiently predicting hotspots in PPIs by combining random forest and synthetic minority over-sampling technique. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(3): 774-781. Available: https://doi.org/10.1109/TCBB.2018.2871674 (Accessed 03 October 2020).

Zhang, X., Zhu, C., Wu, Z., Liu, Z. and Xu, Y. 2017. An imbalance compensation framework for background subtraction. *IEEE Transactions on Multimedia*, 19(11): 2425-2438.

Zheng, L., Liu, G., Yan, C. and Jiang, C. 2018. Transaction fraud detection based on total order relation and behaviour diversity. *IEEE Transactions on Computational Social Systems,* 3(5): 796-806.

Zhou, Z. and Liu, X. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1): 63-77.

Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., Ning, G. 2018. Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access*, 6): 4641-4652. Available: https://doi.org/10.1109/ACCESS.2018.2789428 (Accessed 24 July 2020).

ANNEXURE A: LANGUAGE EDITING CERTIFICATE

Geraldine Coertze

Independent Communications Consultant

Language Practitioner -

English Language Academic Editor

Certificate of English Language Academic Editing

Author: Nhlakanipho Michael Mqadi (21349052) **Date:** 04 May 2021

Document type: Masters dissertation

Discipline: Information and Communication Technology

Academic supervisor: Dr N. Naicker

Higher Education Institution: Durban University of Technology

Dissertation title: Machine Learning: A Data-Point Approach to Solving Misclassifications in the Imbalanced Credit Card Datasets

This serves to certify that the above dissertation was language edited and that assistance was provided with checking the citations and the format of the reference list.

The document was returned to the author with tracked changes, comments and a comprehensive academic editing report to facilitate the correction process. It was the responsibility of the author to accept or reject changes and to attend to all issues raised in the comments and in the report. The final, corrected version of the document was not proofread, although assistance was provided with the Table of Contents, List of Figures, List of Tables, cross-referencing and document layout.

Geraldine Coertze

Communications Consultant and Language Practitioner

Address: Pelham, Pietermaritzburg Email: gericoe@gmail.com Cell: +2782 620 2266

Individual member: ***South African Translators' Association***

Associate member: ***Professional Editors' Guild (SA)***

Entry-level member: ***Chartered Institute of Editing and Proofreading (UK)***

Full member: ***South African Communication Association***

Highest qualification: ***MSocSc (CCMS), UKZN***

ANNEXURE B: SUPERVISOR ENDORSEMENT

DUT
DURBAN UNIVERSITY OF TECHNOLOGY
INYUVESI YASETHEKWINI YEZOBUCHWEPHESHE

FACULTY OF ACCOUNTING & INFORMATICS

Mr NM Mqadi **(Student no: 21349052)**

Dear Mr NM Mqadi

MASTER OF INFORMATION AND COMMUNICATIONS TECHNOLOGY

NM Mqadi (Student no: 21349052)

The Faculty Research Committee considered the examiners' reports on your dissertation and requires that certain corrections be effected before submission of **two hard bound copies (maroon cover with gold writing on cover) and one electronic copy in a single PDF sent on email to nitashas@dut.ac.za** of final versions of your dissertation, and the publication of your result.

You are required to contact your supervisor, **Dr N Naicker** who has been sent the examiner's comments. He has also been requested by the committee to confirm your proper compliance with all outstanding requirements by signing the respective title pages of all requisite copies of your dissertation as follows:

APPROVED FOR FINAL SUBMISSION

____________________ 02 November 2021

PhD, M.Sc, Hons B.Sc, B.Sc, HED

NAME DATE

- The Supervisor will sign only when he agrees to the final submission. Please therefore ensure that such a signed endorsement appears in all final copies of your dissertation submitted to this office.
- Please also ensure that your Supervisor's abbreviated academic qualifications are inserted after his name.

Yours faithfully

____

Ms Nitasha Singh-Sakichand

FACULTY OFFICER: ACCOUNTING AND INFORMATICS

ANNEXURE C: TURNITIN REPORT

Dissertation 23 May 2021

ORIGINALITY REPORT

13%	8%	10%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journal.uob.edu.bh Internet Source	3%
2	Submitted to Birkbeck College Student Paper	1%
3	Nhlakanipho Mqadi, Nalindren Naicker, Timothy Adeliyi. "A SMOTe based Oversampling Data-Point Approach to Solving the Credit Card Data Imbalance Problem in Financial Fraud Detection", International Journal of Computing and Digital Systems, 2021 Publication	<1%
4	"Machine Learning and Information Processing", Springer Science and Business Media LLC, 2020 Publication	<1%
5	ugspace.ug.edu.gh Internet Source	<1%
6	erepository.uonbi.ac.ke Internet Source	<1%