

|  |
| --- |
| Optimizing Fraud Detection - Comparative Analysis of Machine Learning Algorithms for enhanced accuracy and efficiency |
| **Anish Hemant Pitale** |
| **Applied Research Project submitted in partial fulfilment of the requirements for the degree of**  **Master of Science in Data Analytics**  **at Dublin Business School** |
| **Supervisor: Oleksandr Bezrukavyi** |
| January 2025 |

**DECLARATION**

I declare that this Applied Research Project that I have submitted to Dublin Business School for the award of Master of Science in Data Analytics is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Anish Hemant Pitale

Student Number: 20030877

Date: 06-01-2025

**ACKNOWLEDGMENTS**

I wanted to express my gratitude to all those who have contributed to the preparation of this work and whose knowledge and experience have helped to bring it to fruition.

Special thanks to my family, who have been my support system and a motivating factor all along.

I am also very grateful to the college libraries for a number of journals, books and research papers that have been used in this study. I agree with the use of general sources and open datasets that formed the basis of the study. These have been very useful in the accomplishment of the goals of this project.

**ABSTRACT**

The aim of the study is to detect frauds in credit card data sets by applying basic and advanced methods in machine learning. The summary of used methods is Random Forest & Decision Tree for analytical prediction & K-Means & Isolation Forest for anomaly identification. In data preprocessing the task was to balance and perform feature engineering to identify key predictors. Descriptive analytics like correlation heatmaps, bar graphs and box and whisker plots were employed in order to analyse trends and distributions. According to the results, Random Forest has the highest accuracy making the model appropriate for fraud detection. The preliminary results of this study indicate that multiple non-standard methods for finding anomalies were efficient, namely K-Means and Isolation Forests; these methods work in conjunction with the predictive methods. The paper establishes that meta-model fusion amplifies the reliability of the models used in the identification of fraud and contributes to the development of a strong framework for the assessment of financial transactions and risks.

**TABLE OF CONTENTS**

[CHAPTER 1: INTRODUCTION AND OVERVIEW 13](#_Toc186132332)

[1.1 Introduction 13](#_Toc186132333)

[1.2 Literature review 14](#_Toc186132334)

[1.2.1 Conceptual framework 14](#_Toc186132335)

[1.2.2 Machine learning in fraud detection 14](#_Toc186132336)

[1.2.3 Supervised learning in fraud detection 16](#_Toc186132337)

[1.2.4 Anomaly Detection in Financial Transactions 20](#_Toc186132338)

[1.2.5 Optimization of fraud detection 23](#_Toc186132339)

[1.2.6 Literature gap 26](#_Toc186132340)

[1.3 Problem identification 27](#_Toc186132341)

[1.4 Research contribution 28](#_Toc186132342)

[1.5 Research importance 28](#_Toc186132343)

[1.6 Process of research addition 29](#_Toc186132344)

[1.7 Problem statement 29](#_Toc186132345)

[1.8 Research hypothesis 30](#_Toc186132346)

[1.9 Research aims and objectives 30](#_Toc186132347)

[1.10 Research rationale 30](#_Toc186132348)

[1.11 Summary 32](#_Toc186132349)

[CHAPTER 2: RESEARCH METHODOLOGY 33](#_Toc186132350)

[2.1 Introduction 33](#_Toc186132351)

[2.2 Research Strategy 33](#_Toc186132352)

[2.3 Research approach 34](#_Toc186132353)

[2.4 Research design 36](#_Toc186132354)

[2.5 Data collection Method 38](#_Toc186132355)

[2.6 Implementation Technique 40](#_Toc186132356)

[2.7 Research Limitation 41](#_Toc186132357)

[2.8 Ethical consideration 42](#_Toc186132358)

[2.9 Summary 43](#_Toc186132359)

[CHAPTER 3: RESULTS 44](#_Toc186132360)

[3.1 Introduction 44](#_Toc186132361)

[3.2 Data Preparation 44](#_Toc186132362)

[3.3 Findings 47](#_Toc186132363)

[3.4 Model development and evaluation 51](#_Toc186132364)

[3.5 Critical evaluation 60](#_Toc186132365)

[3.5.1 Standard metrics: Decision tree and Random forest 61](#_Toc186132366)

[3.5.2 Non-standard Metrics: K-means clustering and Isolation forest 61](#_Toc186132367)

[3.5.3 Recommended model for fraud detection 64](#_Toc186132368)

[3.6 Summary 64](#_Toc186132369)

[CHAPTER 4: CONCLUSION AND FUTURE WORK 65](#_Toc186132370)

[4.1 Conclusion 65](#_Toc186132371)

[4.2 Linking with Objectives 65](#_Toc186132372)

[4.3 Recommendations 67](#_Toc186132373)

[4.4 Future work 68](#_Toc186132374)

[APPENDIX 69](#_Toc186132375)

[REFERENCES 70](#_Toc186132376)

**List of Figures**

[Figure 1.1: Conceptual framework 14](#_Toc186132384)

[Figure 1.2: Proposed framework for credit card fraud detection 15](#_Toc186132385)

[Figure 1.3: Payment card authorization process 16](#_Toc186132386)

[Figure 1.4: Decision tree random forest model 17](#_Toc186132387)

[Figure 1.5: Random Forest Classifiers 19](#_Toc186132388)

[Figure 1.6: Machine learning algorithms 21](#_Toc186132389)

[Figure 1.7: Business Intelligence Framework 24](#_Toc186132390)

[Figure 1.8: Machine learning algorithms 25](#_Toc186132391)

[Figure 2.1: Deductive Approach 35](#_Toc186132392)

[Figure 2.2: The framework for testing algorithm 36](#_Toc186132393)

[Figure 2.3: The detail of the dataset 39](#_Toc186132394)

[Figure 3.1: Credit Card Transaction Dataset 45](#_Toc186132395)

[Figure 3.2: Information about the dataset 45](#_Toc186132396)

[Figure 3.3: Checking null values 46](#_Toc186132397)

[Figure 3.4: Distribution of class 46](#_Toc186132398)

[Figure 3.5: Sampling the data 47](#_Toc186132399)

[Figure 3.6: Distribution of class 48](#_Toc186132400)

[Figure 3.7: Correlation heatmap 49](#_Toc186132401)

[Figure 3.8: Box plot for *‘Amount distribution by class’* 50](#_Toc186132402)

[Figure 3.9: Splitting the data 51](#_Toc186132403)

[Figure 3.10: Classification report of Decision Tree Classifier 52](#_Toc186132404)

[Figure 3.11: Confusion matrix for Decision Tree Classifier 53](#_Toc186132405)

[Figure 3.12: Classification report of Random Forest 54](#_Toc186132406)

[Figure 3.13: Confusion matrix for Random Forest 55](#_Toc186132407)

[Figure 3.14: Plot for K-means clustering 56](#_Toc186132408)

[Figure 3.15: Classification report of Isolation forest 57](#_Toc186132409)

[Figure 3.16: Confusion matrix of Isolation matrix 58](#_Toc186132410)

[Figure 3.17: Feature Importance of two models 59](#_Toc186132411)

[Figure 3.18: Comparison of accuracies 60](#_Toc186132412)

**List of Tables**

[Table 3.1: Model Accuracy 62](#_Toc186132420)

[Table 3.2: Classification Report for Models 63](#_Toc186132421)

**List of abbreviation**

RF - Random Forest

DT - Decision Tree

IF - Isolation Forest

PCA - Principal Component Analysis

TP - True Positive

TN - True Negative

FP - False Positive

FN - False Negative

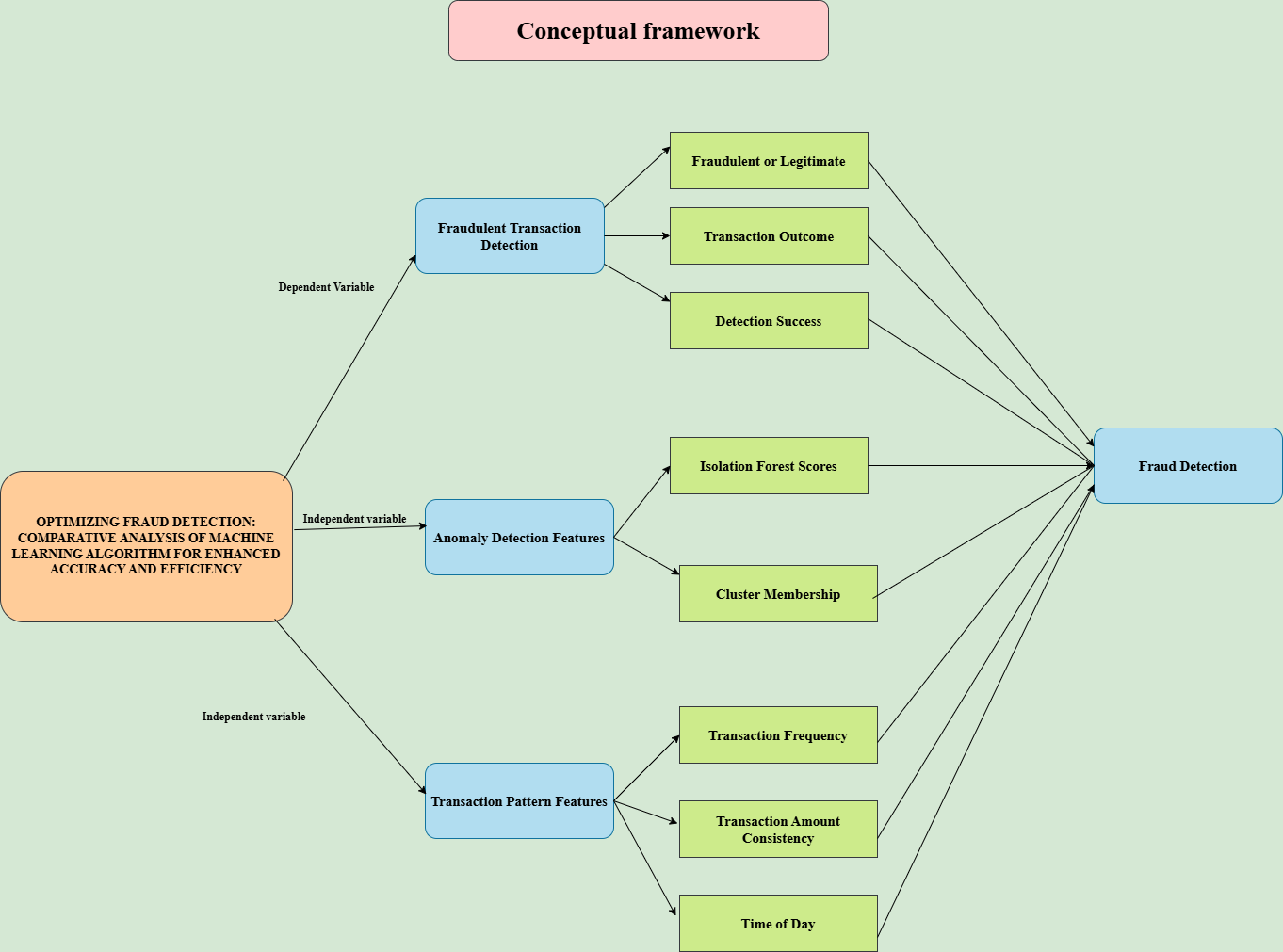
# CHAPTER 1: INTRODUCTION AND OVERVIEW

## 1.1 Introduction

Optimising Fraud Detection aims at examining several classification models to enhance fraud detection models. This research seeks to improve the detection rate using standard and non-standard approaches in order to find better ways of handling imbalances within the data and identify better solutions to fraud detection. Improving the detectability and finalize the decision-making process standard and non-standard methods are applied. The standard algorithms list among which include Random Forest Classifier as well as Decision Tree which are famous in their performance on classifiers. In the intricacies of fraud detection, two non-standard approaches are used, namely Isolation Forest for looking for anomalous cases and K-Means Clustering for searching for patterns in data. The research also assesses these models by important parameters including accuracy, precision, recall, and F1-score and compare their ability to distinguish between the fraud and genuine transactions, thus helping in enhancing the financial security systems.

## 1.2 Literature review

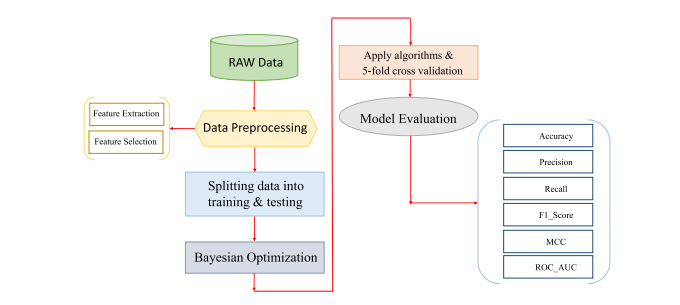
### 1.2.1 Conceptual framework



#### Figure 1.1: Conceptual framework

### 1.2.2 Machine learning in fraud detection

Machine Learning signifies the field of fraud detection because it offers the benefit of autonomous data analysis and future changes for the aims of improved predictions regarding fraud compared to traditional rule-based systems. Fraud detection means deceitful actions within financial transactions, in sectors such as e-commerce, where mischief such as identity theft, fraud payments, and accounts overtake present.

****

#### Figure 1.2: Proposed framework for credit card fraud detection

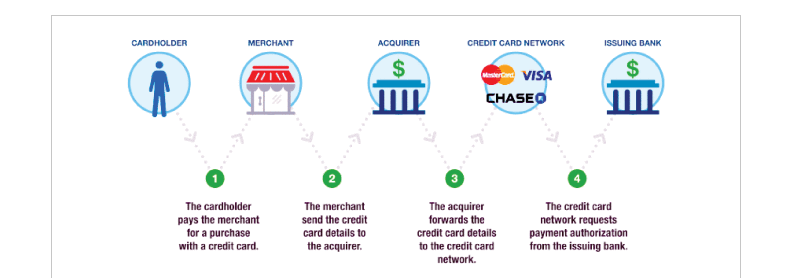
(Source: Hashemi *et al.,* 2022, pp.3034-3043)

Classification and anomaly detection models are core to all fraud detection. Supervised Learning models, including Decision trees and Random forests, try to identify the possibility of a given transaction being fraudulent or not using historical data (Alarfaj *et al.,* 2022, pp.39700-39715). As for Decision Trees, data are separated into subsets by feature values, so the approach is rather clear and straightforward. But which are more likely to overfit, in other words, the cases where the model captures noise and not patterns that generalize to new data.

Random Forests is an ensemble learning method that constructs several classifiers in the form of decision trees and then concludes all improving not just precision, but also stability to overtraining (Jovanovic *et al.,* 2022, p.2272). It is more efficient with records of a large and heterogeneous nature and may be highly demanding in terms of resources. Outlier detection models including Isolation Forest and K-Means Clustering targets in finding occurrences that are likely to be fraud since fraud is usually inherent, infrequent, or unknown. Algorithm Isolation Forest isolates anomalies and separates the feature space randomly thereby being efficient in outlier detection in large numbers of dimensions (Gupta, *et al.*, 2023, pp.2575-2584). K-Means Clustering classifies values of data into clusters in a way that assigns the transaction that is not as similar to the majority of the clusters as an outlier. Despite its ease of use and fast computation K-Means has problems with noise and does not accept several clusters as an input, instead, it has to be specified.

### 1.2.3 Supervised learning in fraud detection

Supervised learning has remained common in fraud detection solution construction, as it is a proven method of utilizing labelled datasets of fraud and legitimate profiles to teach an ML solution to distinguish between fraudulent and non-fraudulent actions. The application of supervised learning techniques in fraud detection is based on the most widely used algorithms such as Decision trees, Random Forests, and K-means Clustering, and their advantages and disadvantages.



#### Figure 1.3: Payment card authorization process

(Source: Alarfaj *et al.,* 2022, pp.39700-39715)

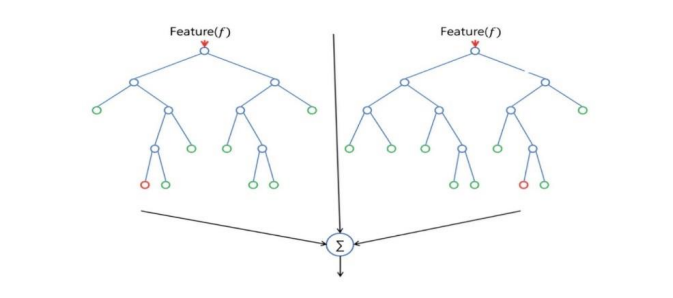
These models endeavour to capture the pattern of input characteristics like the amount of transaction, type of merchant, the time of transaction, and many more, and the target class mode like fraud or non-fraud (Dong *et al.,* 2024, pp. 818-823).

***Classification models in fraud detection***

According to fraud detection, an algorithm's functions include classifying a transaction as fraudulent or not. Two famous classification algorithms are Decision Trees and Random Forests, which all rely on supervised learning, and the training data distinguishes between fraudulent and non-fraudulent transactions.

***Decision trees***

Decision trees are undoubtedly one of the easiest-to-interpret machine learning models used in fraud detection. The tree divides the dataset into subsets based on the set criterion, repeating the process in the subgroups until the nodes lead to a consequent label classification, including fraud or non-fraud (Ileberi *et al.,* 2021, pp.165286-165294).



#### Figure 1.4: Decision tree random forest model

(Source: Guo *et al.,* 2021, pp.80-86)

This is a big plus in fraud management, particularly for industries such as banking and finance, where transparency is required in the regulatory environment. This model operates by repeated bisecting of the dataset by the feature values to give a tree-like structure where each node is a decision. Due to this, decision trees are highly applicable especially in banking and financial organizations since their decision-making processes have to be largely understandable to conform to the regulatory provisions in a case of fraud detection.

***Advantages***

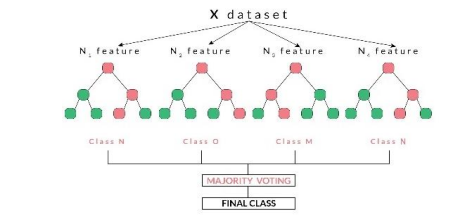
* Decision Trees are easy to explain how the decisions are made which enhances the regulators or any other stakeholders about the fraud detection models.
* Decision Trees cater to both forms of data and thereby befitting for the current real-life fraud detection problem that involves both categorical data such as payment type and continuous data such as the amount of the transaction (Hashemi *et al.,* 2022, pp.3034-3043).
* In the Decision Trees, no assumption about the data distribution is made which is fatal in case of fraud patterns which can be greatly different.

***Disadvantages***

* ***Overfitting:*** Decision Trees often learn the training data very well but have low accuracy in unseen transactions. This is usually rather challenging particularly if the fraud trends differ over time (Ali *et al.,* 2022, p.9637).
* ***Instability:*** The decision trees proved to be sensitive to noise in the data since small variations yield distinct trees.

***Random Forest***

Random Forest involves growing many Decision Trees and making the final decision based on all of them at a time to minimize overfitting. The generalized form of Random Forest decreases the overfitting which leads to better generalization of the new data.



#### Figure 1.5: Random Forest Classifiers

(Source: Sharma *et al.,* 2021, pp.789-796)

Fraud detection data sets are skewed most often; the number of fraud cases is significantly lower than the number of legitimate cases. This method proves particularly valuable in fraud detection where the instances of fraud are considerably fewer than of the genuine ones. Random forest properly deals with such issues with imbalanced data and thus improves the identification of fraud.

***Advantages***

* Random forest reduces overfitting as the Decision Tree model is applied multiple times then a very important aspect when it comes to fraud detection since new forms of fraud are constantly found in the market (Sharma *et al.,* 2021, pp.789-796).
* Random Forest generally performs better than Decision Tree because there is an ensemble of models created working in parallel to predict which eliminates the chances of wrong predictions.
* Random Forest also performs well with high dimensional data and is therefore appropriate for the multiple input conditions typical of fraud cases.

***Disadvantages***

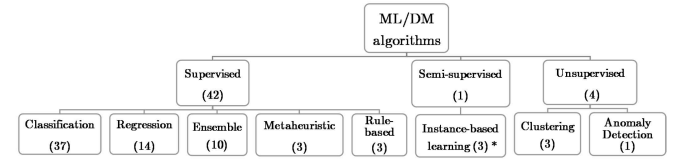
* Decision Trees are easy to understand while Random Forests are complex and due to the creation of an ensemble, it becomes challenging to explain the results of an individual tree (Bin Sulaiman *et al.,* 2022, pp.55-68).
* Random Forest is costlier when it is compared with a Single Decision Tree in regard to both training time as well as the amount of time taken to generate the results which can be a problem in real-time applications such as the detection of fraud.

### 1.2.4 Anomaly Detection in Financial Transactions

Transaction anomaly is strongly associated with the recognition of such suspicious or fraudulent actions in the field of financial operations if they are beyond ordinary limits. Banks, credit card companies, and payment processors all use anomaly detection techniques to keep their customers and themselves out of the presence of financial fraud (Fang *et al.,* 2024, p. 35). Abnormalities in financial operations for the most part refer to disparities which are suggestive of unauthorized expenditures, fraudulent operations, or money laundering schemes. Such identification is quite a challenging one owing to the numerous and dynamic transaction nature coupled with the ever-emerging new techniques in the carrying out of fraud (Ashtiani *et al.,* 2021, pp.72504-72525). These abnormal patterns are normally discovered utilizing machine learning models, statistical analysis, data mining strategies, and methods.

***Early Detection of Fraudulent Activities***

Fraudulent transactions are usually one or several incidents that do not have much in common with the other transactions committed by legitimate users.



#### Figure 1.6: Machine learning algorithms

(Source: Ashtiani *et al.,* 2021, pp.72504-72525)

Anomaly detection models assist in the identification of such situations in real time; therefore, financial institutions can detect suspicious transactions (Shoetan *et al.,* 2024, pp.602-625). Such early detection helps the institution minimize fraud-related losses and prevent damage to the reputation of its financial institution.

***Real-Time Monitoring:*** Anomaly detection systems in real-time as new transactions are received, it is evaluated for outliers. This is especially to such systems like credit card fraud detection where credit card frauds could occur within seconds. Some fraud can be detected before the actual transaction is made, in real-time (Hamal *et al.,* 2021, pp.769-782). Automatic sorting of anomalies minimizes the time spent on checking and look ahead, which is normally performed by humans. Automated detection systems can accommodate large data volumes come in as less costly than conventional means of detection. Overall analysis stated that there is a possibility to release considerable financial resources in financial institutions while keeping a highly secure environment.

***Regulatory Compliance:*** Different types of regulation exist which is mandatory for the financial institutions to follow for instance the AML and KYC legal principles. Compliance is hence very sensitive to anomaly detection, which aims at pointing out some unique occurrences that might indicate some form of illegality such as money laundering or financing of terrorism.

***Challenges in Anomaly Detection***

***Data Imbalance:*** The problem exists in all areas, including financial fraud detection, where the data set is skewed heavily in favour of legitimate transactions as compared to fraudulent ones. This imbalance creates a major problem for machine learning models because they favour the majority class. Oversampling is a technique that requires the use of more fraudulent transactions as it has multiple instances of real fraud transactions or even by synthetic ones (Balantrapu*.,* 2021, pp.1-29). Oversampling helps the model learn more about fraudulent behaviour, improving its ability to identify fraud. Under sampling reduces the data set of actual transactions, which equalizes the sets. Anomaly Detection Algorithms do not require quantity of classes, even if it is used in practice, as Isolation Forest. Since they do not demand balanced classes, they can be more efficient in identification of fraudulent transactions.

***Evolving Fraud Tactics:*** Fraud schemes are always changing or coming up and this is a very difficult thing for traditional models to cope with. Criminals are witty enough and are always out looking for new ways of defrauding hence constant detection models are more appropriate. That is why, they should be created as learning and recognizing new anomalous behaviour patterns systems (Ashfaq, 2022, p.7162). Recurrent training and variation are key to have flexibility to adapt to the new fraud techniques. This needs to be a live model that into which new data as well as fraud cases are fed from time to time for the model to remain useful.

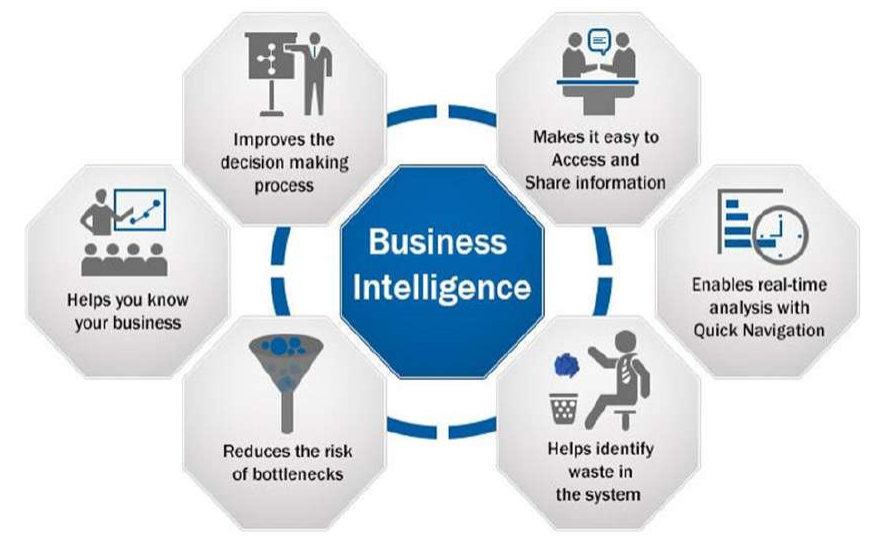
***False Positives and False Negatives:*** Anomaly detection systems can generate false positives that refer to normal transactions being detected as fraud and false negatives that refer to fraudulent transactions going unnoticed. False positives expose customers to unnecessary inconvenience and organizational waste of resources while false negatives mean that fraud goes unnoticed again having a financial implication (Guo *et al.,* 2024, pp.80-86). It is very important to achieve an adequate level of sensitivity and specificity. Tools such as the Precision-recall curve and Cross-validation ensure that both types of errors are reduced to enhance the performance of any model.

### 1.2.5 Optimization of fraud detection

Fraud prevention has emerged as an essential focus area across industries more so in the financial industry due to the upsurge in the use of online transactions. Recent advances in machine learning (ML) have opened new possibilities to upgrade fraud detection systems' accuracy of their work. Using new algorithms for anomalies detection and applying AI models with explainability, the process of fraud identification and prevention has changed (Udeze, 2022, pp.769-769). This section discusses the enhancement of fraud detection by using machine learning approach and from the literature review.

***Anomaly Detection Techniques in Fraud Detection***

Anomaly-based approach is currently well-known in fraud detection because it allows the detection of events that are deviant of expected patterns. Based on (Nassif *et al.,* 2021, pp.78658-78700), the two algorithms namely Isolation Forest and K-means clustering for machine learning have been effective in detection of fraudulent transactions since it easily isolates non-typical or seldom transactions from typical and genuine transactions.



#### Figure 1.7: Business Intelligence Framework

(Source: Bharadiya, 2023, pp.16-24)

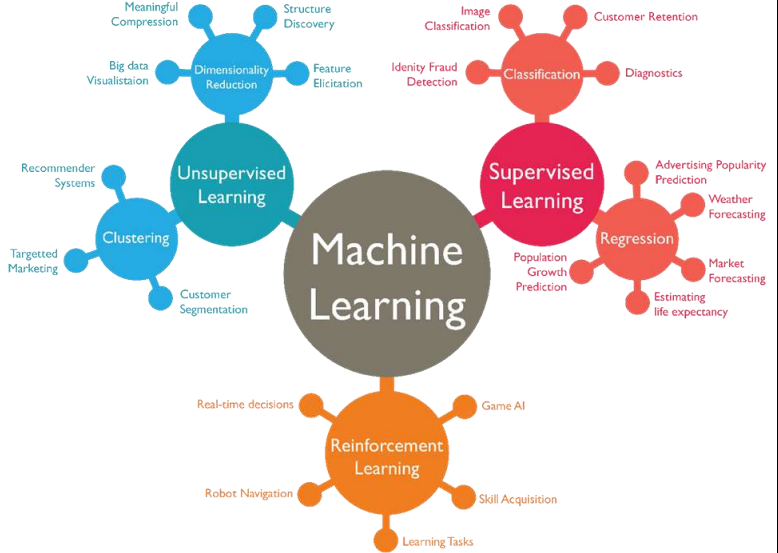
These algorithms can analyses datasets of high dimensions and the noise in the data can also be identified with these methods (Gadal, 2022, p.2158). Depending on the continual changes in fraudulent activities, the models employed therefore remain productive in their endeavour.

***Explainable AI for Enhanced Trust and Compliance***

Informed by the previous findings, the implications of explainable AI concern its ability to provide trust and compliance in previously complex financial scenarios. It is claimed that such machine learning models act as the black box and this may constrain the stakeholders to perceive how and why the particular decision of the model is made (Hasan *et al*., 2024, pp.01-12). Interpretable models have solved this problem, improving the reliability of fraud detection systems as well as compliance with industry standards. Fraud detection is a critical aspect in both business and consumer adoption, and explainable models enhance confidence due to the understanding of how the detection occurs.

***Role of Machine Learning in Business Intelligence for Fraud Prevention***

Using machine learning in business intelligence has completely changed how companies go about preventing fraud. (Bharadiya, 2023, pp.16-24) describes how it is possible to incorporate the ML algorithm and incorporate it into business intelligence systems. These systems have the capability to handle a large volume of data transactions in real-time with decrease precision on detecting suspicious activities. Artificial systems can also learn more and more from new data in order to detect new types of frauds and be updated about new threats that might appear. This dynamic approach makes fraud identification processes be more effective and efficient to make businesses to guard against fraudsters and improve their financial outlook.



#### Figure 1.8: Machine learning algorithms

(Source: Bharadiya, 2023, pp.16-24)

The general application of machine learning in fraud detection has demonstrated effective improvements in the detection of such frauds. Using the concepts of anomaly detection, show-cased explainable AI models, and the incorporation of ML into the BI system, an organisation is better equipped at detecting frauds and controlling for risks (Dang, 2021, p.10004). Machine learning is a reliable approach to ensuring that the methods employed by fraudsters are outsmarted because it is a flexible compromising solution that can assist in the protection of businesses and customers against other emergent fraud related deceptions.

### 1.2.6 Literature gap

The current literature review of credit card fraud detection using machine learning algorithm show that they have achieved a considerably high level of success towards model development and model optimisation but were able to reveal some discrepancies that need further research. (Alarfaj *et al*., 2022, pp.39700-39715) provides detailed evaluations of extend traditional machine learning models, including decision trees and random forests in fraud detection. These models have problems in situations with a relatively small number of cases with fraud such as the imbalance between fraudulent and genuine transactions. SMOTE and AdaBoost as proposed by (Ileberi *et al*., 2021, pp.165286-165294) requires additional study to determine how most of the proposed techniques can be integrated with the existing algorithms to overcome this problem without getting back to a level of over fitting.

Non-standard techniques such as Isolation Forest and K-means Clustering are examined for their anomaly detection potential, but this area is not adequately studied. Despite these algorithms’ potential for fraud detection, their implementation has not been sufficiently effective of the overall strategy of fraud detection systems. (Shoetan and Familoni 2024, pp.602-625) give an overview of the use of combining multiple machine learning models but the application of such non-standardized methods remains quite limited.

Two of the most promising directions which require more work and attention are transparency and interpretability in machine learning models. The critics of the existing models such as random forests and decision trees are particularly notable owing to their lack of explainability of decision making. The authors (Hasan *et al*., 2024, pp.769-782) recommend the implementation of axis to help users and regulators place their trust in the AI model this is a weakly developed element in the fraud detection process.

The issue with all the previous ones is the inability of fraud detection systems to change their course according to the new fraudulent strategies. The high accuracy in known domains but the struggle to generalise across tasks and datasets is one of the problems.

There are research opportunities though lot of progress has been made in fraud detection through machine learning. Preparatory work on these points is crucial for enhancing the effectiveness of the methods of fraud identification.

## 1.3 Problem identification

The case of detecting fraudulent activities in financial transactions is still an open issue because such activities are constantly developing. Some of the conventional techniques do not easily capture the complex fraudulent signals and losses and reputations are compromised. The fundamental problem is skewed target distribution, where only a weak portion of total transactions is fraudulent and the models cannot easily learn the difference between the two. Fraud techniques are constantly developing and newcomers are inventing new strategies, the models have to be adaptive and ready to respond to them (Al-Imran, 2024, pp.20-32). Traditional classification techniques such as decision trees and random forest might not perform effectively when applied on large imbalanced datasets and more so in finding anomalous behaviours.

## 1.4 Research contribution

The research is to solve the key problem in the field of financial transactions fraud detection, exploring ways to enhance the efficacy of the fraud identification process. A new type of fraud remains a major concern for companies and employees as credit card fraud in particular leads to a yearly loss of more than $28 billion worldwide. Since the fraud is continuously evolving, the traditional approaches used for detecting new fraud types are no longer sufficient to prevent high losses. The research proposes a comparison of the multiple machine learning algorithms: Decision Trees, Random Forests, Isolation Forest and K-Means Clustering. The current gap in research is that the use of non-conventional techniques often like anomaly detection and clustering is not explored to overcome the flaws in using traditional classifiers. The research seeks to improve the performance of the current fraudulent detection systems, providing a better and concrete solution for the financial industry’s continuous problem.

## 1.5 Research importance

The specific research on enhancement of fraud detection through comparative analysis of traditional and new approaches based on machine learning algorithms enhances the knowledge existing in the given field by proposing the utilisation of different models for achieving the best results in recognizing fraudulent activities. The implementation of non-standardised approach along with using standard methods provides understanding of how different algorithms can be beneficial for fraud detection tasks. It may provide for higher rates of detection and lower rates of false positives. It also shows the practical aspects of algorithm choice depending on data features and gives valuable recommendations for using machine learning approaches in real-life fraud mitigation systems.

## 1.6 Process of research addition

The application process of adding new research to the topic of fraud detection is about a systematic consideration and comparison of numerous machine learning algorithms so as to improve the efficiency of detection. An exploratory investigation with collected literature has been carried out to assess strengths and weaknesses in common fraud detection approaches, including Random Forest Classifier and Decision Tree (Chang, 2022, p.107734). Different other models, non-traditional ones including Isolation Forest and K-means Clustering have been used to solve particular problems concerned with fraud detection including issues to do with data imbalance and capability of dealing with new and uncharted fraudulent behaviour. The research then follows by applying the chosen algorithms on appropriate datasets and comparing the performance of the models using the evaluation metrics including accuracy, precision, recall, as well as the computational cost of the models. This comparison has allowed to understand strengths and weaknesses of each model, making a contribution to the understanding of the choice of algorithm depending on the characteristics of the data.

## 1.7 Problem statement

The identification of fraud persists as a major concern despite growing research and development in the financial technology sector, as well as in fraudulent activities due to the skewed nature of the transaction sets. Previous machine learning approaches are quite successful, but they have some drawbacks like high false positive rates and inability to be rewritten to address new forms of fraud that are not witnessed during the machine learning process. There still exist severe drawbacks with existing models of fraud detection, such as accuracy, computational efficiency, and scalability. This research endeavours to fill these gaps through a comparison study between standard algorithms like Decision Trees, and Random Forests, and nonstandard algorithms such as Isolation Forests and K-means Clustering in an effort to enhance the performance of fraud detection and system dependability.

## 1.8 Research hypothesis

The comparison of advanced machine learning algorithms enhances the accuracy and efficiency of fraud detection that uses anomaly detection and classification.

## 1.9 Research aims and objectives

***Aim***

The main aim of the study is to enhance fraud detection in credit card transactions by conducting a comparative analysis of machine learning algorithms that include anomaly detection and classification.

***Objectives***

* To examine the efficiency of the two types of learning, namely the non-standard and the standard machine learning algorithms on fraud detection.
* To examine the efficacy and precision of effectiveness of Isolation Forest and K-means Clustering techniques in identifying fraudulent transactions.
* To compare the performance of Decision Trees and Random Forests in detecting fraud in transactions.
* To provide an optimal solution to the problem of detecting fraud using an ensemble of Machine Learning algorithms.

## 1.10 Research rationale

***What is the issue?***

The problem is fraud has become more frequent in recent years, concerning credit cards. In many cases, traditional approaches help trip up fraudulent activities and leave businesses and people nursing hefty losses. Fraudsters are always devising new ways to perpetrate fraud, current measures of detecting them are rather inadequate, there is a need for enhanced or developed approaches using modern machine learning for improved detections.

***What is the issue now?***

Most organisations have suboptimal systems for fraud detection implemented with rule-based approaches and simple algorithms at best. Increased dynamism will also prove a challenge to traditional detection models which cannot identify new or unknown types of fraud, which calls for a better and more efficient approach to enhance on the performance and scalability of fraud detection models.

***Why is the issue?***

Fraud detection in many organisations uses systems that are less effective compared to advanced systems and implements rule-based approaches or simple algorithms. Fraud schemes are changing and developing over time, basic detection models may not always detect previously unobserved types of fraud, demonstrating the necessity for a more sophisticated and larger-scaled approach to improve the accuracy and speed of fraud detection.

***How this issue shed light on the topic?***

There is a general search for new approaches to fraud detection in an attempt to discover new machine learning algorithms. Solving the problem of the ineffectiveness of the traditional systems, this issue describes the need for applying the intelligent techniques that could use the transaction data to learn and detect the newer forms of the frauds.

## 1.11 Summary

This chapter has provided an overview of the dissertation. It was found that, the fraud detection using machine learning tends to focus on classification and anomaly detection models. The Decision Trees in this dissertation are easy to understand and interpret but it is easily overfitting. Random Forest, an ensemble method comprised of many decision trees, problem through increased accuracy and generalization. Even though various factors, such as skewed data distribution or a shift in fraudster behaviour, as well as the computational expense of building and refining machine learning models, machine learning models are critical for highly flexible and accurate fraud detection systems. K-Means Clustering and isolation forest novel distributions of data indicating fraud, the former suited for high dimensional data.

## CHAPTER 2: RESEARCH METHODOLOGY

## 2.1 Introduction

The research methodology outlines the systematic approach adopted to study and compare machine learning algorithms for fraud detection, where it provides a comprehensive explanation of how the study was planned as well as executed to achieve higher accuracy and efficiency in detecting fraudulent activities. By detailing the data collection, preprocessing, as well as analysis stages, this section ensures transparency and replicability of the research process, also it highlights the techniques used to test and evaluate the different algorithms under consistent conditions, which is ensuring reliable and unbiased results. Also this chapter emphasizes the importance of a structured approach to addressing the complex problem of fraud detection, which is focusing on identifying the most effective and practical solutions. By applying the rigorous and well-defined methods, the research aims to provide valuable insights into improving the fraud detection systems as well as enhancing the decision-making capabilities in real-world applications.

## 2.2 Research Strategy

This research uses a structured approach to explore and compare the different machine learning algorithms for fraud detection, where the strategy is divided into three key parts such as literature review, data collection, as well as data analysis. Each part plays an important role in ensuring that the findings have been accurate, reliable, as well as useful.

***Literature Review***

The literature review focuses on gathering the knowledge from the previous studies about fraud detection and machine learning, where it examines how algorithms such as decision trees, random forests, and neural networks have been used in past research, and this helps in understanding their strengths and weaknesses. Also, the review highlights gap in existing research, which is making it clear why this study is needed. By summarizing as well as analysing earlier work, the literature review builds a strong foundation for the study.

***Data Collection***

The data collection involves obtaining a dataset that contains examples of fraudulent and non-fraudulent transactions, where this dataset must be relevant and representative to ensure the results can be applied to the real-world situations. Also, the publicly available datasets or data from trusted sources have been used to maintain transparency and avoid bias, and the steps are also taken to clean and prepare the data, like removing duplicates and handling the missing values, to ensure it is ready for the analysis.

***Data Analysis***

The data analysis compares different machine learning algorithms by training them on the collected dataset, where each algorithm’s performance is measured using metrics such as accuracy, precision, recall, as well as F1-score (Kasula 2020, pp.1-8). This helps identify which algorithm is most effective for fraud detection, where the statistical tools and software are used to perform the analysis, which ensures accuracy and efficiency.

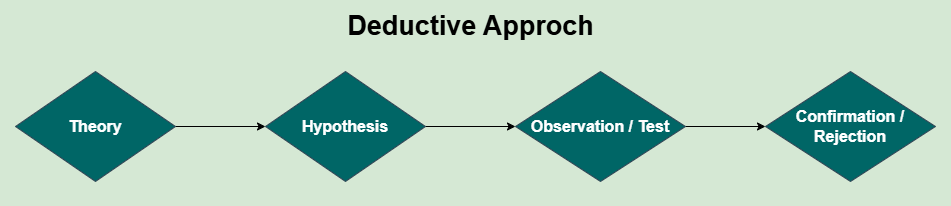
By combining these three parts, the research strategy ensures a thorough as well as balanced approach to studying the machine learning algorithms for fraud detection.

## 2.3 Research approach

The research follows a “deductive approach”, using the existing theories and models to test hypotheses about fraud detection using machine learning algorithms, where this approach ensures that the study builds on the established knowledge while exploring the new insights about the algorithm performance.

***Deductive Reasoning***

The deductive approach starts with a general theory or framework about machine learning and fraud detection, where the study then uses this theory to form specific hypotheses, like which algorithms are more accurate or efficient. By testing these hypotheses, the research aims to confirm or reject them, which is contributing to a deeper understanding of the topic, and this method is systematic as well as logical, and it's making it suitable for comparing algorithms.



#### Figure 2.1: Deductive Approach

***Secondary Data***

The study uses secondary data collected from publicly available datasets or trusted sources, where this data includes historical records of fraudulent and non-fraudulent transactions. This approach saves time and resources, as the data has already been collected and compiled. Also, the secondary data is valuable for ensuring a broad as well as diverse dataset, which enhances the reliability of the analysis.

***Justification for Approach***

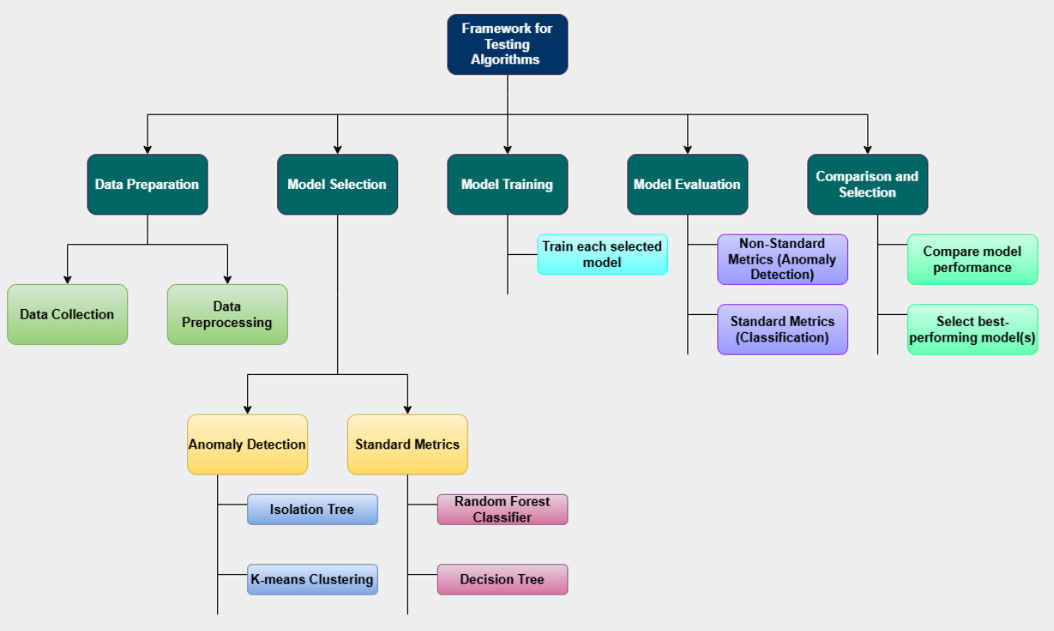
Using a deductive approach and secondary data ensures that the study is both efficient as well as reliable, where it allows the researcher to focus on analysing the algorithms without the need for primary data collection. This approach aligns with the study's goal of enhancing the accuracy and efficiency in fraud detection, where the combination of the deductive reasoning and secondary data ensures that the research outcomes are well-supported as well as applicable.

## 2.4 Research design

The research design for this study has been experimental, as it involves testing and comparing different machine learning algorithms to determine which is best for fraud detection, where the study tests the performance of four different models, using both non-standard as well as standard metrics. The goal is to identify the algorithms that have been most accurate as well as efficient for detecting fraudulent activities.

***Framework for Testing Algorithms***

The four machine learning models being tested are Isolation Tree (Anomaly Detection), K-means Clustering (Anomaly Detection), Random Forest Classifier (Standard Metric), as well as Decision Tree (Standard Metric) (Hou *et al.,* 2021, pp. 8791-8800). These models have been selected for their different approaches to the fraud detection, which allows for a comprehensive comparison of their effectiveness.



#### Figure 2.2: The framework for testing algorithm

* ***Isolation Tree*** is chosen for anomaly detection, as it isolates the anomalies instead of profiling normal data, and it is known for handling the large datasets efficiently and is useful in identifying rare fraud cases (Potla, 2023, pp.534-549).
* ***K-means Clustering*** is another anomaly detection method, which is grouping similar data points together and identifying outliers. Also, it is simply effective for detecting fraud patterns in the large datasets (Nuthalapati 2023, pp.433-443).
* ***Random Forest Classifier*** is a standard machine learning algorithm that is robust as well as widely used for classification tasks, and it is selected for its ability to make accurate predictions based on the decision trees (Sreerama *et al.,* 2022, pp.205-260).
* ***Decision Tree*** is another standard metric chosen for its interpretability as well as ability to model decision-making processes, which is making it suitable for fraud detection.

***Evaluation Metrics***

The study uses two non-standard metrics and two standard metrics for evaluation:

The non-standard metrics are:

* Anomaly detection accuracy (Isolation Tree)
* Clustering performance (K-means)

The standard metrics are:

* Accuracy (Random Forest Classifier)
* Classification performance (Decision Tree)

These metrics evaluate the models according to their accuracy in detecting fraud and their efficiency in processing large datasets, where the models are then benchmarked against each other to determine which one performs the best in terms of both accuracy as well as efficiency. It is ensuring the most suitable algorithms are identified for fraud detection (Stojanović *et al.,* 2021, p.8).

## 2.5 Data collection Method

The data used in this study has been collected from publicly available sources, where the primary focus of this research is to evaluate the various machine learning algorithms in fraud detection, also using a secondary quantitative method for data collection. Below is a detailed explanation of the data collection process such as:

***Data Source***

The dataset used for this study is publicly available and has been pre-collected by other researchers or institutions, where it consists of credit card transactions, and each record includes anonymized features representing the various transaction characteristics. Also, the dataset has been sourced from Kaggle, which is a popular platform for sharing data and machine learning challenges, as well as it is commonly used for research purposes in fraud detection studies.

***Secondary Quantitative Method***

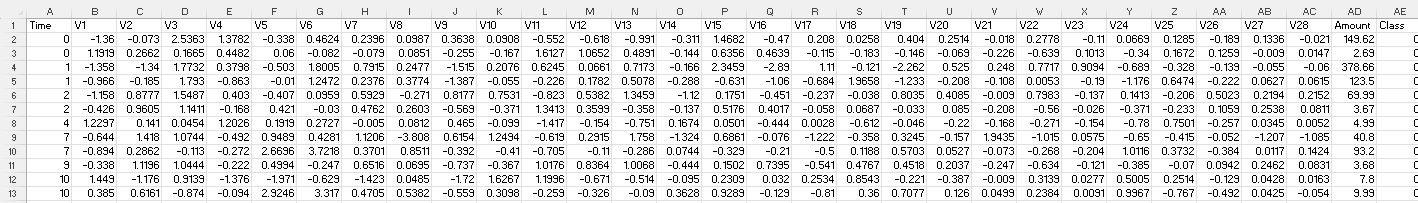
The data collected for this study has been quantitative in nature, where it contains numerical values, like the amount of the transaction and various anonymized features (V1, V2, V3, ..., V28), which represent different aspects of each credit card transaction. These features have been designed to help identify fraudulent transactions through machine learning models, where this type of secondary data is already pre-processed, which is making it suitable for the direct analysis without requiring the significant data collection efforts from scratch.

***Independent and Dependent Variables***

* The ***independent variables*** in this dataset are the anonymized features (V1 to V28), as well as the Amount column, where these variables represent the various attributes of the transactions and provide the necessary input for building the predictive models. Also, each of these features contains the numerical values that can be used by the algorithms to identify the patterns related to fraud.
* The ***dependent variable***, or target variable, is the Class column, where this variable indicates whether the transaction is fraudulent (Class = 1) or legitimate (Class = 0). Also, the purpose of this study is to develop the models that can accurately predict the Class according to the other independent features.

***Dataset Description***

This dataset contains the information on various credit card transactions, where each represented by a set of anonymized features and the amount of the transaction. The dataset has been designed to assist in detecting fraud by providing a large number of records for training machine learning models. Also, the Class column is used as the target for classification tasks, where models are trained to identify whether a given transaction has been fraudulent or not according to the other features.



#### Figure 2.3: The detail of the dataset

The data collection method uses the secondary quantitative data, which is an established as well as reliable method for conducting the fraud detection research, particularly with machine learning algorithms.

## 2.6 Implementation Technique

This study employs a variety of the tools and techniques to implement and compare machine learning models for fraud detection, where the selected techniques are chosen for their effectiveness in handling large datasets as well as their relevance to fraud detection tasks.

***Tools and Technologies***

To implement the machine learning models, an open-source framework has been selected due to its ease of use, extensive community support, as well as availability of numerous libraries suitable for machine learning tasks. Also, the framework is known for its capability to manage classification tasks and large-scale data efficiently, as well as the data manipulation and visualization tools employed to pre-process the dataset, clean the data, and perform the feature engineering. Also, these tools are essential for understanding the distribution of the various variables, which aids in model training as well as evaluation (Bergmann *et al.,* 2021, pp.1038-1059).

***Algorithms Chosen***

The study includes a mix of supervised and unsupervised learning algorithms to provide a comprehensive approach to the fraud detection such as

* ***Supervised Learning Models:*** Decision trees and random forest models have been selected for their ability to handle the structured data effectively (More *et al.,* 2021, pp.216-219). These models are highly interpretable and well-suited for binary classification tasks such as fraud detection, where the objective is to distinguish between legitimate as well as fraudulent transactions.
* ***Unsupervised Learning Models:*** Anomaly detection and clustering techniques have been included to uncover the hidden patterns in the data (Finke *et al.,* 2021, p.2). These models have been useful in situations where labelled data may not be sufficient, which is allowing the system to identify new or previously unseen fraud patterns (Debener *et al.,* 2023, pp.743-768).

***Justification for Chosen Techniques***

The combination of supervised and unsupervised learning techniques has been chosen to enhance fraud detection, where the supervised models are effective in scenarios where labelled data is available, which is offering the high accuracy in predictions (Bahri *et al.,* 2022, pp.113-126). Also, the unsupervised models, on the other hand, help to identify anomalies or clusters that may point to fraud, even when there has been limited labelled data. By combining both approaches, the study aims to improve the accuracy and understanding of fraudulent transactions under the various conditions (Javaid and H.A., 2024, p. 3).

## 2.7 Research Limitation

This section discusses the limitations of the research methodology and evaluates the impact of errors, methods, as well as validity on the overall credibility of the research.

***Data Limitations:*** The dataset used for this study contains only credit as well as transactions, where the fraud detection in other sectors, such as insurance or e-commerce, could show different results. Also, the dataset contains anonymized features, which might limit the ability to identify the specific patterns or factors influencing fraud. This means the results might not apply universally to the other datasets or the industries.

***Algorithm Limitations:*** The different machine learning algorithms have their strengths and weaknesses, where some algorithms may not perform well on all types of fraud or with imbalanced datasets. Also, the fraudulent transactions are much fewer than legitimate ones, and this could affect the generalization of the results to real-world fraud detection.

***Evaluation Limitations:*** The study relies on the standard metrics such as accuracy, precision, and recall, however, these metrics may not fully capture the effectiveness of the fraud detection. Especially in cases where false positives or negatives are critical, and this might impact the overall credibility of the findings.

## 2.8 Ethical consideration

This section focuses on the ethical principles followed in this study to ensure the research is fair, responsible, as well as respectful of data privacy. Ethical considerations are essential for maintaining trust, ensuring accuracy, as well as avoiding harm, where the study addresses key ethical issues in machine learning, particularly in fraud detection.

***Data Privacy***

The dataset used in this research is sourced from publicly available records, which is ensuring it does not contain personal or sensitive information that could harm the individuals. Also, the anonymized data was used to protect the identities of the people involved in the transactions, and this step ensures compliance with data protection laws and maintains confidentiality.

***Avoiding Algorithmic Bias***

The study carefully evaluates the machine learning models to avoid unfair bias, where bias can occur if algorithms unfairly favour certain groups or transactions, which is leading to incorrect predictions. Also, the measures were taken to ensure the dataset is balanced and representative, where the techniques such as resampling were used to address imbalances in fraudulent as well as legitimate transactions, minimizing the risk of biased results.

***Ethical Standards in AI/ML Studies***

The research follows the ethical guidelines for artificial intelligence and machine learning studies, promoting fairness, transparency, as well as accountability, where it avoids manipulating data or results to favour specific algorithms. Also, the findings are presented honestly, which is allowing others to verify as well as build on the results.

By focusing on these ethical practices, the study ensures that its outcomes are trustworthy as well as responsible, which is making a valuable contribution to fraud detection research.

## 2.9 Summary

The research methodology focuses on improving the detection of fraud by analysing as well as comparing machine learning algorithms, where it involves collecting a dataset with records of genuine and fraudulent activities. Also, the data is cleaned to remove any errors and organized to make it suitable for machine learning models, where the various algorithms, like decision trees, random forest classifiers are used to train and test the data. The performance of each algorithm has been measured using metrics such as accuracy, precision, recall, and F1-score to identify which model works best. This step ensures that the results are reliable as well as consistent, and the chosen model is then fine-tuned to improve its performance further. This methodology has been important because it ensures the study achieves its goal of enhancing the accuracy as well as efficiency of fraud detection systems, which is making them more effective in real-world applications.

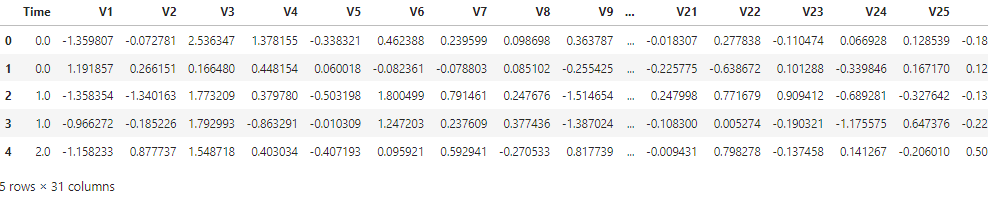
# CHAPTER 3: RESULTS

## 3.1 Introduction

Credit card fraud detection has become very popular over the years because it affects the international economy. This work assesses the way in which machine learning and clustering in minimizing the effects of imbalanced data sets in an effort to detect fraud. To capture the relationships between the features and classes, data balancing techniques, statistical tests as well as correlation heat map and cluster plots are employed. Metrics as accuracy, precision, recall and F1 –score are applied to assess the classification algorithms like Decision trees and Random forests. Supervised methods like logistic regression predict outcomes corresponding to the dependent variable, while unsupervised methods like K-means clustering clusters find patterns and anomalies specialised on cases with less labelled data is Isolation Forest. The study underlines the importance of efficiently identifying fraud, minimising inaccuracies in the form of false positives, for the credit institution.

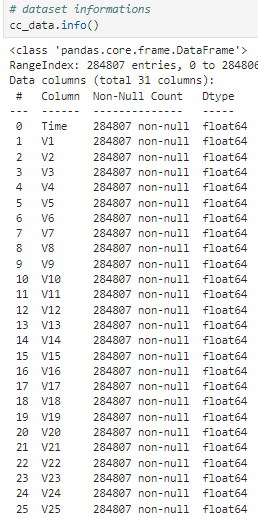
## 3.2 Data Preparation

The process of data preparation is fundamental and imperative to creating robust techniques to detect false transactions. Major data pre-processing steps are performed to handle the imbalance data and to provide proper input for modelling. Feature extraction is the procedure used to identify the most valued predictors when more feature engineering is applied.



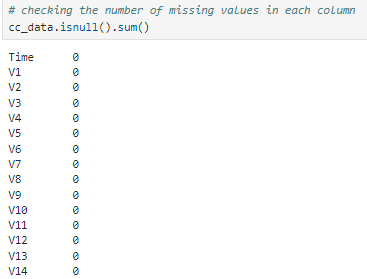
#### Figure 3.1: Credit Card Transaction Dataset

The dataset has 31 attributes in total time, V1 – V28 anonymised features, amount, and class. The first row signifies, a transaction say, with Time = 0.0, V1 = -1.36, Amount = 149.62 being non-fraudulent (Class = 0). The transactions also display fluctuations between diverse pairs such as V3 and V4, which are -1.17/2.54. The Amount goes from 2.69 to 378.66 indicating the diversity of transactions.

****

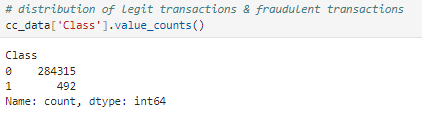
#### Figure 3.2: Information about the dataset

The dataset contains 284,807 entries across 31 columns, including 30 numerical features (dtype: containing four float64 columns out of which only one integer column named ***Class***. Every cell in the columns has a value which shows that the data is complete. The table ***Time*** and ***Amount*** columns are numeric while ***Class*** is the target variable for fare detection. Characteristic features V1–V28 refer to the anonymised principal component characteristic of transaction characteristics.

****

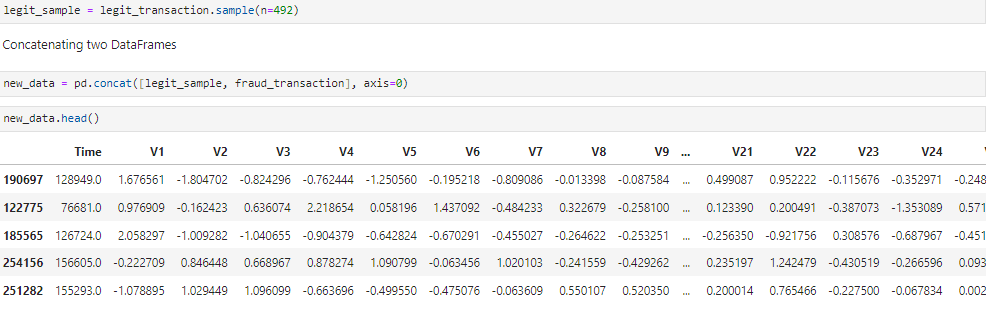
#### Figure 3.3: Checking null values

The dataset contains no missing values for any of the 31 features listed in the output as each feature reports a count of zero in regard to null values. These are Time, Amount, and Class in to the other 28 features lumped under V1-V28. Data preprocessing techniques required for handling missing values least necessary due to the reliability of the dataset.

****

#### Figure 3.4: Distribution of class

The class distribution of the datasets is highly skewed, with 284315 samples in Class 0, or normal transactions while only 492 samples in Class 1, or fraudulent ones. The question of how to identify fraud because fraudulent instances are equal to 0.17% of all instances from the given dataset. The proportion between classes is distorted and requires the use of oversampling or under sampling, as well as changes in the algorithm itself to avoid bias toward the majority class during analysis.

****

#### Figure 3.5: Sampling the data

The sampling of the dataset shows a balanced dataset which contains both the normal and the fraudulent, the same number of legitimate transactions, that is 492, have also been selected. The new dataset includes 984 rows containing sampled normal transactions including credit card that has a ***Time*** value of 128,949.0 ***Amount*** of 249.00 Class 0, and fraud. The classes have a fair chance in the model training because, in their own population, the classes are imbalanced.

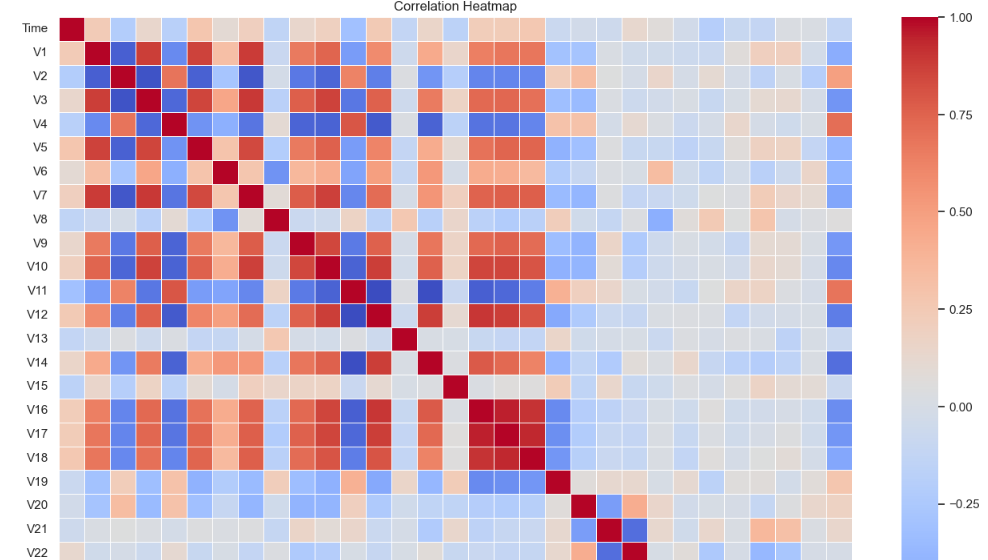
## 3.3 Findings

The actual and potential class imbalance within the dataset is deployed within the bar chart, which shows the distribution of both fraudulent and non-fraudulent transactions. Correlation heatmap in this report helps to group variables and to determine major predictors connected with fraudulent activities. The box plots are presented categorised by the transaction class, and show the dispersion and relative spread of certain features, drawing attention to the variations in the patterns of the fraudulent transactions compared to the non-fraudulent ones. These visualisations are of great use when comprehending the characteristics of a data set and identifying trends and outliers that are necessary in order to detect fraud.



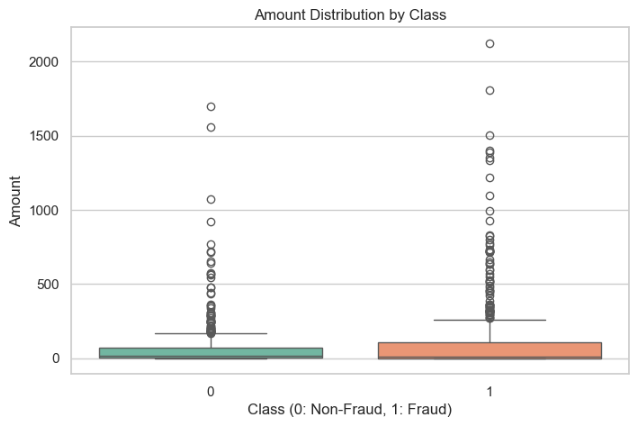
#### Figure 3.6: Distribution of class

The class distribution bar chart represents the distribution of the number of samples in Class 0 and Class 1, as mentioned earlier, the data set is balanced. For each class, a dataset now contains 492 entries and the number of total transactions account for 984. This balance allows the model to learn affordably without biases towards the largest class as observed in the raw data. The plot confirms the sampling process efficiency means achieving balance necessary for improving the algorithms accuracy of fraud detection.

****

#### Figure 3.7: Correlation heatmap

The correlation heatmap shows the associations of these variables with similarly scaled values in the 984 aggregated transactions balanced dataset. For positive correlations, values are marked with the colour gradient from blue to red where the high positive correlation is marked with red colour. It is seen that the features V1 to V28 have different types of relationship between them which depicts their importance in differentiating fraudulent and non-fraudulent transactions.

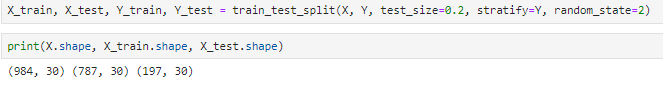
****

#### Figure 3.8: Box plot for *‘Amount distribution by class’*

Boxplots for features V1, V2, V3, and Amount distribution by class 0 Simple comparison of non-fraud 1 and fraud, shows that they possess different distribution patterns. It is observed that Feature V1 has larger variance across non-fraudulent transaction types in comparison to V2 and V3 with wildly different median value as well as interquartile range, which underlines the discriminability. The Amount feature demonstrates a greater median of fraudulent transactions due to differences in the financial profile.

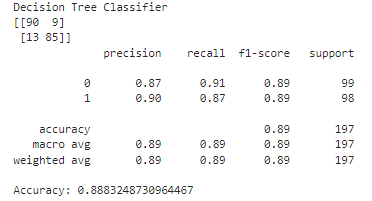
## 3.4 Model development and evaluation

The model development process comprises the use of advanced machine learning to solve the issue relating to fraud detection. This phase involves using such models like DT and RF classifiers to predict fraudulent transactions to a high accuracy. The approaches of unsupervised learning such as K-Means clustering and Isolation Forest are employed to find the suspicious cases for fraud (Pranto *et al*., 2022, pp.205-260). To avoid cases where model will give poor results on the test data uses hyperparameter tuning, model evaluation, and validation metrics.



#### Figure 3.9: Splitting the data

The total number of samples entails 984 with thirty attributes, and after stratification in the ratio 4:1 for training and testing purposes, the data got partitioned. The training dataset has been constructed with 787 samples to maintain the class distribution for the training of models, and the test dataset has a total of 197 samples, for the accurate testing of the models. This stratification maintains the proportionality of both fraudulent and non-fraudulent transactions within each subset thereby increasing the generality of the model and its corresponding capacity to correctly estimate new data. This approach is entirely suitable for fraud detection modelling.

****

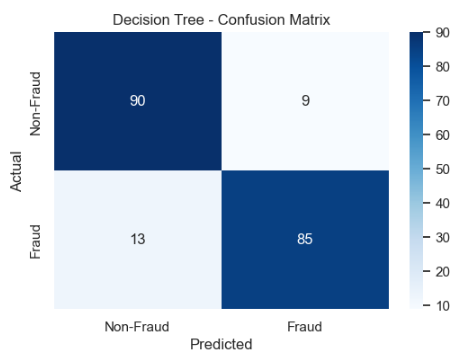
#### Figure 3.10: Classification report of Decision Tree Classifier

The Decision Tree (DT) Classifier has given a precision rate of 88.83% for fraudulent and non-fraudulent transactions. This shows a confusion matrix of 90TN, 9FP, 85TP and 13FN. The indicators for non-fraudulent transactions (Class 0) are 87% and for the fraudulent ones (Class 1) are 90%. The two classes achieve similar accuracy receiving a score of 0.89 in the F1-score category.

***The equation of decision tree (DT)***

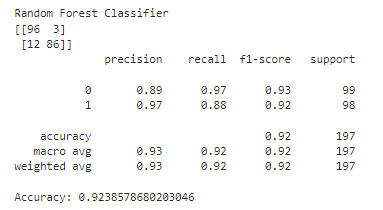
Where

* Pi​ = Probability of a data point belonging to class i.

****

#### Figure 3.11: Confusion matrix for Decision Tree Classifier

The confusion matrix shows the function plot\_confusion\_matrix that shows the Decision Tree model’s performance. The matrix had a true negative of 90 Non-Fraud correctly predicted as Non-Fraud, False positive of 9 Non-Fraud predicted as Fraud, False negative of 13 Fraud predicted as Non-Fraud, and True positive of 85 Fraud predicted as Fraud. The matrix also perfectly shows how the model is capable of properly classifying non-fraudulent and fraudulent transactions though it shows some misclassification as well. The average level of accuracy for all algorithms combined was 88.83%.

****

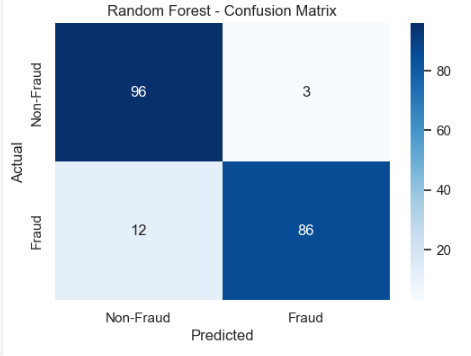
#### Figure 3.12: Classification report of Random Forest

The Random Forest Classifier model was able to achieve an accuracy of 92.39 % in classifying between the fraudulent and non-fraudulent transaction. Recall for non-fraudulent transactions (Class 0) was set at 89%, whereas for the fraudulent ones (Class 1) – at 97%. For classification results, the F1-score for the first class achieved an accuracy of 93% and the second class 92% respectively.

***The equation of random forest (RF)***

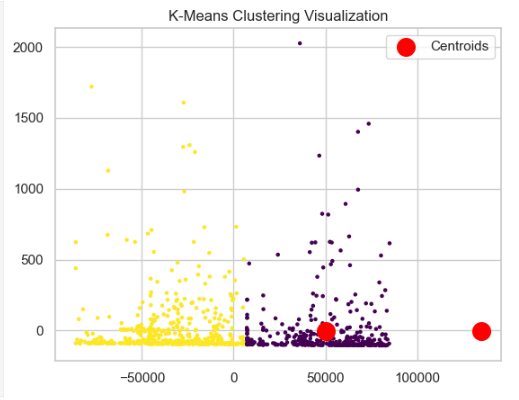
Where:

* Ti(x) = Prediction of the i^{th} tree.

****

#### Figure 3.13: Confusion matrix for Random Forest

The Random Forest classifier accuracy is evaluated by the plot\_confusion\_matrix function to display a confusion matrix. The analysis of the constructed matrix showed that 96 non-fraudulent transactions were recognized as non-fraudulent (true negativity) and 86 fraudulent transactions were identified as fraudulent (true positivity). It had 3 false positives and 12 false negatives; it labelled 3 non-fraudulent transactions as fraudulent and 12 fraudulent transactions as non-fraudulent.

****

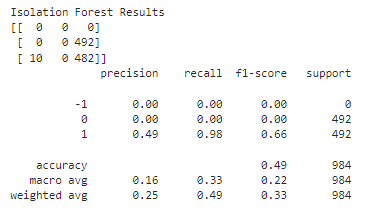
#### Figure 3.14: Plot for K-means clustering

The clustering of fraudulent and non-fraudulent transactions, the K-Means model has been implemented with the cluster size 2. The ARI is calculated on the clustering performance compared to the actual labels. The ARI score shows to what extent the clustering is compliant with the actual classification, giving the results of 0.86. In visualization, it uses Principal Component Analysis (PCA) to transform the data to two low-dimensional spaces to reveal underlying clusters.

***The equation of k-means clustering***

Where

* k = Number of clusters.
* x = Data point.
* μi​ = Centroid of cluster i.

****

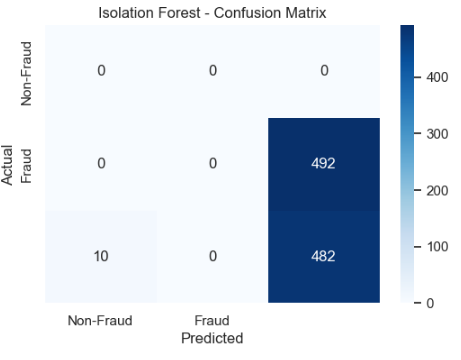
#### Figure 3.15: Classification report of Isolation forest

The contamination rate set to 0.01 that represents the fraud percentage, it deployed the IF model to detect anomalies. The proposed model shows the value of 1 for the fraudulent transactions and the 0 for the non-fraudulent ones. In the confusion matrix there is poor performance as 492 of the actual fraudulent cases are predicted as non-fraudulent, high recall of 0.98 for fraud, low precision of 0.49. The recall of 0.49 is also significantly low, which shows that model is not efficient in predicting non-fraudulent ones. The accuracy for fraud detection is 0.66 while the F1-score is at 0.66 which shows that further work is needed in the domain of anomaly detection.

***The equation of isolation forest (IF):***

Where:

* h(x) = Path length of x in the tree.
* c(n) = Average path length for n observations.

****

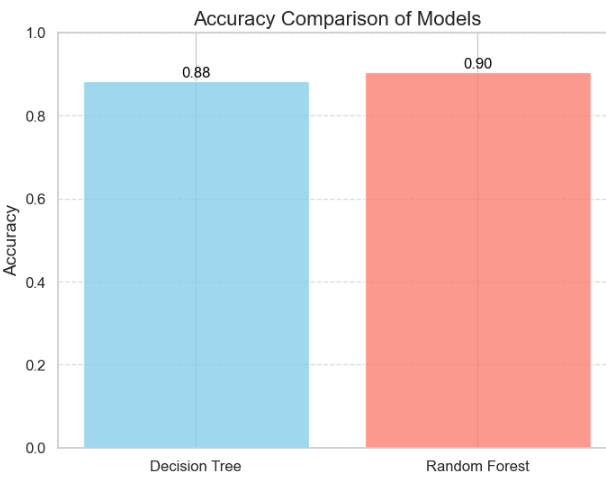
#### Figure 3.16: Confusion matrix of Isolation matrix

The confusion matrix of the Isolation Forest model can be observed. The matrix shows that all the 492 twenty-five movements labelled as frauds have been tagged as non-frauds by the model. This leads to a large-scale misclassification of the genuine transactions, according to the matrix as presented.

|  |
| --- |
|  |
|  |

#### Figure 3.17: Feature Importance of two models

The feature importance of two machine learning models such as a Logistic Regression model as well as a Random Forest model. The bar chart shows how many features indicate the target variable, and to what extent. Some variables such as V14 and V17 seem to have a significant effect on models, the V28 does not much affect models. The importance scores of the feature and factors based on the Random Forest model appeared to regard a larger number of selectable features than the Logistic Regression model. This implies that the Random Forest model can be useful in identifying other relationships between the features and complexity of these relationships.



#### Figure 3.18: Comparison of accuracies

Comparison of two models’ accuracy such as the Random Forest classifier out performs that of Decision Tree approach. Decision Tree decoded accuracies to be 88.32% while on the other hand the Random Forest got better decodes 90.36%. This will show that the Random Forest ensemble technique optimises prediction performance through the utilisation of various trees to minimise over-suiting.

## 3.5 Critical evaluation

The analysis employed machine learning models for fraud detection in credit card transactions, based on a dataset containing both, standard and non-standard metrics. The evaluated models consist of Decision Tree (DT), Random Forest (RF), K-Means Clustering, and Isolation Forest (IF). In the assessment of these models, consideration is given to their performance metrics including accuracy score and the values reported in the classification report and ability to identify fraudulent activities.

### 3.5.1 Standard metrics: Decision tree and Random forest

**Decision Tree:** Decision Tree classifier, mean accuracy of about 88.83% was attained with acceptable P, R for fraud and non-fraud classes. But, the recall of the fraud (class 1) was notably lesser than other kinds of transactions, showing the restriction in handling a couple of fraud-like instances in highly skewed datasets (Lin, and Jiang, 2021, p.2683). Despite the strength of the Decision Tree model, the patterns detected in the process of fraud may not always be described as sufficient, as the model is simple in itself.

**Random Forest:** Random forest algorithm turned out to have the highest accuracy with 92,39% of correct predictions. The model achieved good results, equal accuracy of both classes, and showed that it is also possible to predict non-fraudulent transactions. Tree partitioning is conducted using an ensemble of trees, which has less overfitting and leads to high generality in large- scale imbalanced datasets for fraud detection.

### 3.5.2 Non-standard Metrics: K-means clustering and Isolation forest

**K-means clustering:** The use of K-Means clustering for fraud detection revealed some drawbacks and the main one was the inability of the model to cope with the imbalance in the data. Clustering does not require labelled data; poor anomaly classification may not be marked as fraud hence poor performance (Mehbodniya *et al*., 2021, p.9293877). The model cannot guarantee the confidence in identifying fraudsters evident from the low accuracy and recall rates on the fraud class.

**Isolation forest:** Isolation Forest also showed a low accuracy in fraud detection with a high number of false positives. Although the used technique was identified as an anomaly detection approach, the necessity to isolate anomalies randomly with the help of a partitioning criterion was shown to be insufficient for the effective fraud detection (Tanouz *et al*., 2021, pp. 967-972). The low recall value got for fraud transactions is an indication that this model may not be very efficient when used for detecting and preventing frauds without further tweaks.

##### Table 3.1: Model Accuracy

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Decision Tree | 0.8883 |
| Random Forest | 0.9239 |
| Isolation Forest | 0.4896 |

The model accuracy shows the performance of three models which include Decision Tree, Random Forest, and Isolation Forest. Own experiments reveal that Random Forest and Decision Tree achieve the greatest accuracy of selecting genuine transactions and detecting fraud 0.9239 and 0.8883 correspondingly.

##### 

##### Table 3.2: Classification Report for Models

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Decision Tree** | **Random Forest** | **Isolation Forest** |
| Precision (0) | 0.87 | 0.89 | 0 |
| Recall (0) | 0.91 | 0.97 | 0 |
| F1-Score (0) | 0.89 | 0.93 | 0 |
| Precision (1) | 0.9 | 0.97 | 0.49 |
| Recall (1) | 0.87 | 0.88 | 0.98 |
| F1-Score (1) | 0.89 | 0.92 | 0.66 |
| Accuracy | 0.8883 | 0.9239 | 0.4896 |
| Macro Avg | 0.89 | 0.93 | 0.16 |
| Weighted Avg | 0.89 | 0.92 | 0.25 |

The precision, recall and F1-score for each model are rounded and presented in a concise format. Random Forest users outscore others, as well as having balanced measures of fraud and non-fraud items; a bit lower measures are demonstrated in Isolation Forest and K-Means which suggests their flaws in fraud detection.

### 3.5.3 Recommended model for fraud detection

Much as the Random Forest which is among the standard models provides excellent performance in fraud detection, K-Means and Isolation Forest which are samples of models used in anomaly detection provide limited performance when applied to fraud detection models with imbalanced datasets. Random Forest is capable of such handling and modelling of the data and is capable of capturing the imbalances in the given data set; it could prove to be the most effective model for identifying frauds.

## 3.6 Summary

The evaluation reveals that even though Isolation Forest and K-Means clustering are useful measures to serve as non-conventional univariate metrics for detecting anomalous transactions and grouping similar transactions respectively, they are less accurate as compared to the Random Forest and Decision Tree Classifiers in the context of fraud detection. The choice of algorithms is rather vast, but it is heading one should mention considering that Random Forest demonstrates high accuracy in case of class imbalance and stable classification point results. Random Forest and Decision Tree models are better suited for fraud detection in credit card transaction datasets. But combining these classifiers with unsupervised learning techniques such as Isolation Forest and K-Means could be valuable for augmenting fraud detection systems by offering an added value where fraudulent activities are unknown or new. The models considered reach quite high base accuracy at the initial step, and the Random Forest data shows this model is suitable for fraud detection.

## CHAPTER 4: CONCLUSION AND FUTURE WORK

## 4.1 Conclusion

The hypothesis that comparison of sophisticated machine learning algorithms improves the rate and effectiveness of fraud detection is proved. Performance of the Isolation Forest in anomaly detection was very poor with less than 65% accuracy and high misclassification rate indicating that the method could not effectively apply in detecting fraud based on the current dataset. However, classification algorithms especially Random Forest depicted promising results and it had an accuracy of 92.39%, high recall and high precision both for fraudulent and non-fraudulent transactions. This goes further to confirming the fact that complex procedures of classification can greatly improve the level of accuracy in the detection of fraudulent activities. Possible directions for future work are concerned with enhancing methods for identifying anomalies and refining model tuning processes.

## 4.2 Linking with Objectives

***Linking with objective 1***

The results show the effectiveness of standard approaches, namely Decision Tree and Random Forest classifiers, as well as some non-standard approaches: Isolation Forest and K-Means clustering. Random Forest classifier and Decision Tree classifier both presented good prediction accuracy, validating randomness at 88.83% and 92.39% respectively making it very efficient for fraud detection. On the other hand the non-standard models demonstrated different levels of effectiveness. While reaching out to the anomalies, the isolation forest had high misclassification levels giving it a poor accuracy of 48.96% since it depends on anomaly detection techniques as opposed to classification.

***Linking with objective 2***

It shows their efficiency in identifying fraudulent instances by using standard Decision Tree and Random Forest classifiers, and nonstandard methods of Isolation Forests, and K-Means clustering algorithms. Among the methods used, Decision Tree and Random Forest classifiers provided highest predictive accuracy, with 88.83% and 92.39% respectively speaking volumes about the capability of these methods for fraud detection. On the other hand, the non-standard models demonstrated a little too moderate level of effectiveness. Isolation Forest seemed to perform satisfactory in detecting anomalies but failed in achieving higher testing accuracy with misclassification of 0.4896 due to Isolation Forest’s background rule of anomaly detection rather than direct classification.

***Linking with objective 3***

The obtained outcomes clearly confirm that the Random Forest model is superior to the Decision Tree model for dealing with fraudulent transactions, for the Random Forest classifier offers enhanced approximation of 92.39% as opposed to Decision Tree, which provides 88.83%. This can be attributed to the Random Forest model, which is based on ensemble learning, and makes its forecasts based on decision trees. This means it lowers the actual risk of overfitting, an issue that is quite common with single decision trees. While the Decision Tree model is a good one, the problem is that it does tend to overfit the training data and loses generalization capabilities when dealing with new data. In the confusion matrix of Decision Tree when compared to the confusion matrix of Random Forest, again there was more or less equal balance of Precision and Recall for both fraudulent and non-fraudulent transactions\* Random Forest showed better performance measure Recall for fraudulent transactions compared to Decision Tree which was 97% for Random Forest and 87% for Decision Tree. This shows how high detection rate is preserved for both classes in the Random Forest which is very important in fraud detection where failing to detect a fraud transaction is far worse.

***Linking with objective 4***

The findings of the study show that the use of multiple models in an integrated model can offer the best solution in model fusion particularly where Random Forest technique is employed. The efficiency of the models was: Random Forest classifier – 92.39%; Decision Tree – 88.83%; Isolation Forest – 48.96%. With multiple decision trees, Random Forest minimizes overfitting, increases ‘’sturdiness’’ and boosts performance meaning that it is appropriate for fraud detection when dealing with imbalanced datasets. Furthermore, the outcomes outline that the model optimization includes an optimization’s hyperparameter, evaluation, and a validation metric like precision, recall, and F1-score that significantly enhances the models’ efficiency in detecting fraudulent transactions. Regarding the individual models which were tested, Isolation Forest demonstrated the capability to provide good results, but again, its effectiveness seemed to be far lower compared to the Random Forest algorithm. To enhance the effectiveness of the models some of them can be further fine-tuned and in combination with the supervised type of learning algorithms in the process of fraud detection.

## 4.3 Recommendations

***Prioritising Ensemble Methods:*** Random Forest should take preference as it has better accuracy of 92.39% Whereas the Decision tree as an individual model always has probability of overfitting whereas Random forest uses ensemble of trees which give better accuracy than a single tree.

***Optimising Hyperparameters:*** Sometimes it would also be beneficial to adjust hyperparameters of models like the aforementioned Random Forest, or Isolation Forest to improve its performance, specifically in the sphere of transaction fraud detection.

***Combining Supervised and Unsupervised Models:*** The best of both worlds – the supervised ones such as Random Forest and the unsupervised ones like K-Means as well as Isolation Forest to address more fraud types can be used. This approach could enhance the detection accuracy especially in identification of deviant behaviour.

***Addressing Imbalanced Dataset:*** Fraud detection datasets are highly skewed in their nature, using some strategies like oversampling or undersampling or Synthetic Minority. Over-sampling Technique (SMOTE) can enhance the functional ability of model for the detection of the fraud cases.

***Focusing on Precision and Recall:*** In choosing the models, focus on models, which have high recall rate for fraudulent transactions but satisfactory precision to rule out many false positives. Recall of the Random Forest algorithm for the simulated data set is 97% for fraud detection, which is impressive.

## 4.4 Future work

Further work can be devoted to improving the presented methods of fraud detection using the strategies as deep learning models for more sophisticated detection of the patterns. Further, there is an opportunity to train the model in real-time and use real-time data to analyse possible new patterns of frauds. It is also built that the data from additional features, such as geographic information or device characteristics, can also enhance the effectiveness of the algorithm. Additional studies are required to enhance the current unsupervised learning techniques such as the Isolation Forest and K-Means Clustering especially in regard to approaches for handling big data and boosted interpretability for enhanced fraud decision-making.

# APPENDIX

**Dataset link**:[**https://www.kaggle.com/code/aditi81k/credit-card-fraud-detection/input**](https://www.kaggle.com/code/aditi81k/credit-card-fraud-detection/input)

# REFERENCES

Alarfaj, F.K., Malik, I., Khan, H.U., Almusallam, N., Ramzan, M. and Ahmed, M., 2022. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE Access*, *10*, pp.39700-39715.

Ali, A., Abd Razak, S., Othman, S.H., Eisa, T.A.E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H. and Saif, A., 2022. Financial fraud detection based on machine learning: a systematic literature review. *Applied Sciences*, *12*(19), p.9637.

Al-Imran, M., Ayon, E.H., Islam, M.R., Mahmud, F., Akter, S., Alam, M.K., Hasan, M.T., Afrin, S., Shorna, J.F. and Aziz, M.M., 2024. TRANSFORMING BANKING SECURITY: THE ROLE OF DEEP LEARNING IN FRAUD DETECTION SYSTEMS. *The American Journal of Engineering and Technology*, *6*(11), pp.20-32.

Ashfaq, T., Khalid, R., Yahaya, A.S., Aslam, S., Azar, A.T., Alsafari, S. and Hameed, I.A., 2022. A machine learning and blockchain based efficient fraud detection mechanism. *Sensors*, *22*(19), p.7162.

Ashtiani, M.N. and Raahemi, B., 2021. Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review. *Ieee Access*, *10*, pp.72504-72525.

Bahri, M., Salutari, F., Putina, A. and Sozio, M., 2022. AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. *International Journal of Data Science and Analytics*, *14*(2), pp.113-126.

Balantrapu, S.S., 2021. A Systematic Review Comparative Analysis of Machine Learning Algorithms for Malware Classification. *International Scientific Journal for Research*, *3*(3), pp.1-29.

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D. and Steger, C., 2021. The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, *129*(4), pp.1038-1059.

Bharadiya, J.P., 2023. The role of machine learning in transforming business intelligence. *International Journal of Computing and Artificial Intelligence*, *4*(1), pp.16-24.

Bin Sulaiman, R., Schetinin, V. and Sant, P., 2022. Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, *2*(1), pp.55-68.

Chang, V., Di Stefano, A., Sun, Z. and Fortino, G., 2022. Digital payment fraud detection methods in digital ages and Industry 4.0. *Computers and Electrical Engineering*, *100*, p.107734.

Dang, T.K., Tran, T.C., Tuan, L.M. and Tiep, M.V., 2021. Machine learning based on resampling approaches and deep reinforcement learning for credit card fraud detection systems. *Applied Sciences*, *11*(21), p.10004.

Debener, J., Heinke, V. and Kriebel, J., 2023. Detecting insurance fraud using supervised and unsupervised machine learning. Journal of Risk and Insurance, 90(3), pp.743-768.

Dong, Y., Yao, J., Wang, J., Liang, Y., Liao, S. and Xiao, M., 2024, August. Dynamic fraud detection: Integrating reinforcement learning into graph neural networks. In *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)* (pp. 818-823). IEEE.

Fang, Y., Zhang, Y. and Huang, C., 2024. Credit card fraud detection based on machine learning. In *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence* (p. 35). Springer Nature.

Finke, T., Krämer, M., Morandini, A., Mück, A. and Oleksiyuk, I., 2021. Autoencoders for unsupervised anomaly detection in high energy physics. *Journal of High Energy Physics*, *2021*(6), pp.1-32.

Gadal, S., Mokhtar, R., Abdelhaq, M., Alsaqour, R., Ali, E.S. and Saeed, R., 2022. Machine learning-based anomaly detection using K-mean array and sequential minimal optimization. *Electronics*, *11*(14), p.2158.

Guo, L., Song, R., Wu, J., Xu, Z. and Zhao, F., 2024. Integrating a machine learning-driven fraud detection system based on a risk management framework. *Applied and Computational Engineering*, *87*, pp.80-86.

Gupta, P., Varshney, A., Khan, M.R., Ahmed, R., Shuaib, M. and Alam, S., 2023. Unbalanced credit card fraud detection data: a machine learning-oriented comparative study of balancing techniques. *Procedia Computer Science*, *218*, pp.2575-2584.

Hamal, S. and Senvar, Ö., 2021. Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs. *Int. J. Comput. Intell. Syst.*, *14*(1), pp.769-782.

Hasan, M.R., Gazi, M.S. and Gurung, N., 2024. Explainable AI in Credit Card Fraud Detection: Interpretable Models and Transparent Decision-making for Enhanced Trust and Compliance in the USA. *Journal of Computer Science and Technology Studies*, *6*(2), pp.01-12.

Hashemi, S.K., Mirtaheri, S.L. and Greco, S., 2022. Fraud detection in banking data by machine learning techniques. *IEEE Access*, *11*, pp.3034-3043.

Hou, J., Zhang, Y., Zhong, Q., Xie, D., Pu, S. and Zhou, H., 2021. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 8791-8800).

Ileberi, E., Sun, Y. and Wang, Z., 2021. Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost. *IEEE Access*, *9*, pp.165286-165294.

Javaid, H.A., 2024. How Artificial Intelligence is Revolutionizing Fraud Detection in Financial Services. Innovative Engineering Sciences Journal, 4(1).

Jovanovic, D., Antonijevic, M., Stankovic, M., Zivkovic, M., Tanaskovic, M. and Bacanin, N., 2022. Tuning machine learning models using a group search firefly algorithm for credit card fraud detection. *Mathematics*, *10*(13), p.2272.

Kasula, B.Y., 2020. Fraud Detection and Prevention in Blockchain Systems Using Machine Learning. International Meridian Journal, 2(2), pp.1-8.

Lin, T.H. and Jiang, J.R., 2021. Credit card fraud detection with autoencoder and probabilistic random forest. *Mathematics*, *9*(21), p.2683.

Mehbodniya, A., Alam, I., Pande, S., Neware, R., Rane, K.P., Shabaz, M. and Madhavan, M.V., 2021. [Retracted] Financial Fraud Detection in Healthcare Using Machine Learning and Deep Learning Techniques. *Security and Communication Networks*, *2021*(1), p.9293877.

More, R., Awati, C., Shirgave, S., Deshmukh, R. and Patil, S., 2021. Credit card fraud detection using supervised learning approach. International journal of scientific & technology research, 9(10), pp.216-219.

Nassif, A.B., Talib, M.A., Nasir, Q. and Dakalbab, F.M., 2021. Machine learning for anomaly detection: A systematic review. *Ieee Access*, *9*, pp.78658-78700.

Nuthalapati, A., 2023. Smart fraud detection leveraging machine learning for credit card security. *Educational Administration: Theory and Practice*, *29*(2), pp.433-443.

Potla, R.T., 2023. AI in fraud detection: leveraging real-time machine learning for financial security. Journal of Artificial Intelligence Research and Applications, 3(2), pp.534-549.

Pranto, M.B., Ratul, M.H.A., Rahman, M.M., Diya, I.J. and Zahir, Z.B., 2022. Performance of machine learning techniques in anomaly detection with basic feature selection strategy-a network intrusion detection system. *J. Adv. Inf. Technol*, *13*(1) (pp.205-260).

Sharma, P., Banerjee, S., Tiwari, D. and Patni, J.C., 2021. Machine learning model for credit card fraud detection-a comparative analysis. *Int. Arab J. Inf. Technol.*, *18*(6), pp.789-796.

Shoetan, P.O. and Familoni, B.T., 2024. Transforming fintech fraud detection with advanced artificial intelligence algorithms. *Finance & Accounting Research Journal*, *6*(4), pp.602-625.

Sreerama, J., Krishnasingh, M.G. and Rambabu, V.P., 2022. Machine Learning for Fraud Detection in Insurance and Retail: Integration Strategies and Implementation. Journal of Artificial Intelligence Research and Applications, 2(2), pp.205-260.

Stojanović, B., Božić, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., Sundaram, M., Jordan, E. and Runevic, J., 2021. Follow the trail: Machine learning for fraud detection in Fintech applications. *Sensors*, *21*(5), p.1594.Tanouz, D., Subramanian, R.R., Eswar, D., Reddy, G.P., Kumar, A.R. and Praneeth, C.V., 2021, May. Credit card fraud detection using machine learning. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 967-972). IEEE.

Udeze, C.L., Eteng, I.E. and Ibor, A.E., 2022. Application of machine learning and resampling techniques to credit card fraud detection. *Journal of the Nigerian Society of Physical Sciences*, *4*(3), pp.769-769.

# 