National College of Ireland

Feature Based Selection Technique For Credit Card Fraud Detection System Using Machine Learning Algorithms

MSc Research Project
Olaitan Olanlokun

Forename Surname
Student ID: x20113897

School of Computing
National College of Ireland

Supervisor: Ross Spelman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Olaitan Olanlokun……

Student ID: ……20113897……..……

Programme: ……Cyber Security…… **Year:** ……2021……..

Module: …MSc ……..……

Supervisor: ……..……

Submission Due Date: ……16/08/2021……..……

Project Title: … Feature Based Selection Technique for Credit Card Fraud Detection System Using Machine Learning Algorithms

……..……

Word Count: ……7703…… **Page Count**……21……..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……

Date: ……15/08/2021……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

ABSTRACT
One of the three most common forms of online fraud has been recognized as credit card fraud. Every year, these costs financial institutions billions of dollars. While the field of credit card fraud detection using machine learning algorithms has been extensively researched, numerous studies have revealed that there are still a considerable amount of false positives or false alarms. These false positives indicate people who may be denied access to products and/or services because their transactions are incorrectly classified as credit card fraud by the detection model, even though they are lawful. This ultimately leads to the objective of this proposed research. To compare and achieve the lowest possible false positive rate, the research proposes the use of different feature selection techniques such as recursive feature elimination and correlation-based feature selection, combined with machine learning algorithms such as random forest and logistic regression.
Keywords— **Credit card fraud, machine learning algorithms, feature selection techniques, false positive rate**

1. Introduction
Credit Card is a card made with plastic legally issued by a bank or a credit card service provider with correspondence on the credit limit. Credit cards can be used for the purchase of goods and services established on an agreement between the cardholder and the company issuing the card. A line of credit constitutes the amount the maximum amount that can be borrowed by the card user, which is usually issued to the cardholder with regards to the user's credit income. The cardholder accepts to repay for the goods and services received from the lender and make overtime payments on the purchased card. In return, the credit card company bills the cardholder an interest on his balance [1] When the credit card is used physically for transactions at a payment station, it is referred to as a physical credit card. When the card is used digitally to make a purchase online, it is referred to as a Virtual credit card.
The detection of fraud has been quite challenging because fraudsters have devised new ways to make their behavior seem legitimate. Some technologies that have been deployed that can be used to keep fraud from happening include chip and pin verification, Card Verification Code (CVV), and Address Verification System, even these technologies could fail and that's why the implementation of a credit card fraud detection system is of utmost importance. [2].
The major limitation with developing a credit card fraud detection system is that the number of legitimate transactions outnumbers the fraudulent ones which lead to an imbalanced dataset. The use of machine learning algorithms is now employed to analyze and report all authorized transactions that are suspicious. The reports are investigated by the right professionals who reach out to the card owners to confirm if the transactions were fraudulent or legitimate. The information gotten from the feedback is trained and used to update the algorithm which helps to improve the performance of the system over time for fraud detection. [3]
To stay one step ahead of these criminals, financial institutions deploy sophisticated systems that monitor consumer behavior and spending habits and alert them when an odd or suspicious activity/purchase occurs on their account. For example, if a person is known to make a series of transactions in each geographic place and then makes an odd purchase in a location other than their regular location, this transaction may be identified as suspicious behavior, and the user may be contacted to confirm the purchase's validity. Financial institutions have also put a lot of money on systems that use machine learning algorithms to analyze these patterns automatically and potentially detect and classify legitimate and fraudulent transactions. This research paper focuses on the detection of credit card fraud in financial institutions and how it

can be resolved. Random Forest, Logistic Regression, and XGBoost are the algorithms used in this research work. These algorithms were used with different feature selection technique to enhance the performance of the model after which evaluation metrics was used to determine which of the algorithms had the best result. Therefore, answering the research question; **How do feature selection techniques affect the false-positive rates in credit card fraud detection?**
Chapters that follow in this document are outlined as follows: Section 2 (Literature Review) details past research in credit card fraud prediction. Section 3 (Methodology) outlines the chosen methodology and its application to this research. Section 4 (Design Specification) details the programming languages, libraries, datasets, and tools used to conduct the research experiment. Section 5 (Model Implementation) Discusses how the experiment was carried out. Section 6 (Evaluation) Discusses the evaluation metrics used for the research and outcome of the experiment after the evaluation metrics. Section 7 (Conclusion and Future Work) discusses the outcome of the research paper and areas relating to the research topic that should be elaborated on.

2. Literature Review

With the increase in trade volume and growth in the advancement of digital commerce, the spread of fraud has resulted in major losses for financial institutions which also affects individual clients. Detection of credit card fraud has become a task for financial institutions. This paper aims to evaluate some models and feature selections that can detect credit card fraud. Fraud is the unethical act of getting money, a product, or a service. Credit card fraud is the conduct of using another individual's credit card to obtain money without authentication or acknowledgment from the individual. With the increase in the use of the internet and online transaction, Credit card theft activities are also on the rise, which has left banks with irrecoverable costs. Credit card fraud is categorized into application fraud and Behavioral fraud.

- Application fraud

This is when an individual illegally for a new card from issuing companies or from the bank using another person's information or false information. It could be a set of user details presented by a user which is referred to as duplication of fraud or identifying information with a different user which is also referred to as identity fraud.

- Behavioral Fraud

This fraud could be in the form of mail theft where a person receives the credit card or information of the credit card from the bank before it reaches the original owner, credit card counterfeit, absence of credit card owner, and lost/stolen credit card.
Common methods through which the information of the card owners can be obtained are as follows:

- Illegal Swipers: This could happen when employees at hotels, pubs, restaurants, and other establishments swipe the cards of unwitting card owners, downloading the encoded information on the cards, which can then be used to construct a fake card or make online payments without the need for a real card.

- Phishing Scams: When a hacker sends phishing emails that appear to come from the card owner's financial institution but contain codes that enter the owner's device and download financial information that is later exploited by the hackers, this is known as spear phishing.

- Network Intrusion: When a hacker intercepts communication on a company's network and obtains vital information about a customer's finances, this is known as a data theft. [4]

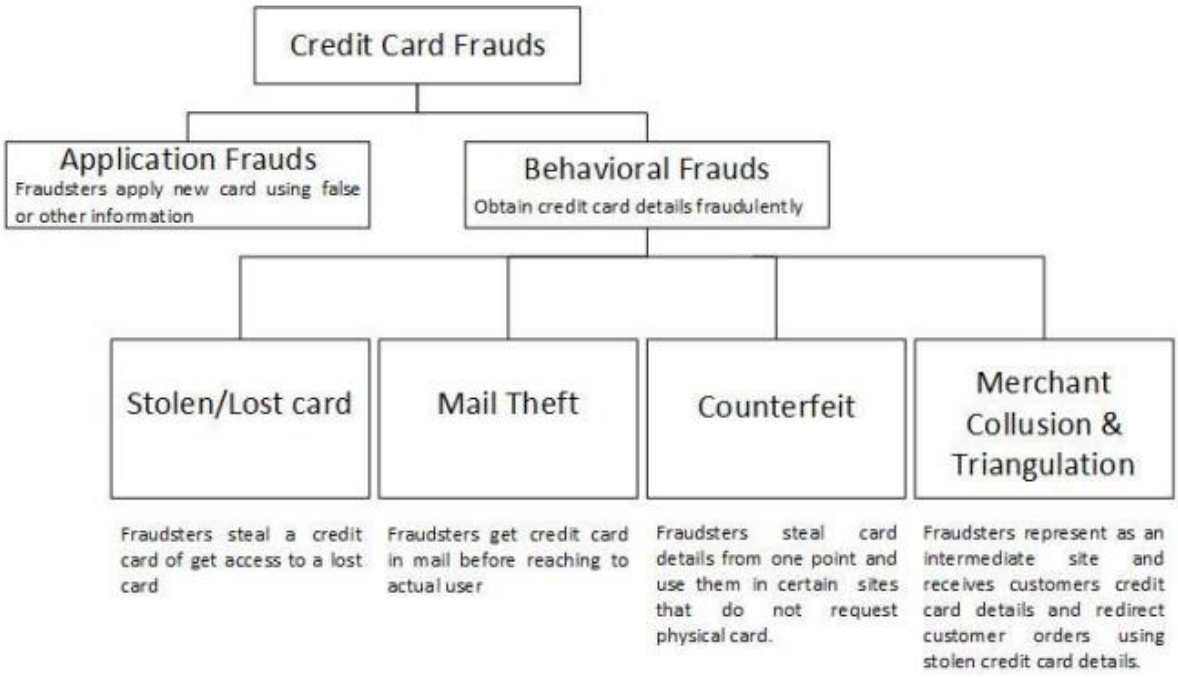


Figure 1: Credit Card Fraud Classification

2.1 Feature Selection

Legitimate and fraudulent card transactions are similarly characterized as fraudsters always attempting to mimic the spending habit of the original card owners to avoid detection, thereby make the transactions of fraudulent and legal transactions unstable and constantly changing. It is therefore of utmost importance to focus on features that can differentiate fraudulent transactions from legitimate transactions. The evaluation of the card owner's spending behavior, which can be observed using data mining techniques, is the foundation of credit card fraud detection. The optimal selection of the features that determine where each transaction falls is crucial for building a successful credit card fraud detection system.

The characteristics utilized in a credit card fraud detection system have a significant impact on the system's performance. When a set of characteristics is employed incorrectly, it might result in low accuracy, a high rate of false positives, and a low rate of false negatives as a result, different feature selection approaches must be used to locate the relevant columns (features) in the dataset that separate fraudulent from valid transactions. The main goal of feature selection is to improve the learning algorithm's performance and thus the system's accuracy. The variables in a dataset are frequently derived from actual financial institution transactions. However, due to the sensitivity of financial information, these datasets are only available in limited quantities. [5]

2.2 Data Mining

Data mining can be acknowledged as a key trend when it comes to disclosing fraud initiatives. It is also a major when it comes to assessing risk. Detection of credit card fraud can therefore be used with data mining.

The most successful way of revealing credit card fraud is by studying the pattern used on each card and deciphering if there is any distinctness concerning the typical spending pattern. The spending behavior accustomed a person tends to follow a specific routine, which can be translated into a set of patterns of the credit cardholder. A variation in the usual pattern could imply the transaction system is under threat.

Data mining can be presented as the procedure of taking over a set of analytical data tools and procedures to recognize patterns and behaviors in data to break them down into information that will be useful. Variations in patterns could be how often a purchase is made, types, and pattern of expenditure. [6]

Due to the ever nature of financial transactions, there are instances where legitimate transactions are mistakenly detected as fraudulent (false positives), or fraudulent transactions are mistakenly flagged as legitimate (false negatives) (false negatives). Any of these scenarios could be harmful to a financial institution, resulting in financial losses (in the case of false negatives) or customer dissatisfaction (in the case of false positives) (in the case of false positives). Different data mining approaches would be used as part of the research aims to reduce the number of false positives and false negatives obtained from each model.

Logistic regression, random forest, and decision tree were used in the detection of credit card fraud while for evaluation; error rate, accuracy, sensitivity, and specificity were used for the performance evaluation of the experiment carried out by [7], the following results were gotten respectively for Random Forest, Decision Tree and logistic regression, 95.5,94.3 and 90.0. His conclusion after comparison of all three methods was that Random Forest was better than decision tree and logistic regression.

The research carried out by [8], concluded that random forest was a better classifier for determining whether transactions were fraudulent or nonfraudulent. Accuracy, precision, and recall were used as the metrics. The performance of the model was evaluated following metrics. Logistic Regression, naïve bayes, random forest model, and Multilayer perception were the algorithms used for the experiment.

Different classifiers were applied to the different groups out of which scores were rated and generated for each classifier. [9] grouped customers according to their transactions and extracted their behavioural patterns to formulate a profile for each cardholder. Synthetic Minority Oversampling Technique (SMOTE) was used to balance the dataset, where it was noted that some classifiers performed better as compared to when Mathews Correlation Coefficient (MCC) was used to treat the imbalanced dataset. He however concluded his experiments with Logistic Regression, Random Forest, and Logistic regression giving better results than Support vector Machine, isolation forest, and local outlier.

The experiment conducted by [4] chose Random Forest as the best model out of the nine models he used for the experiment. The nine models include random forests, naïve Bayes, K-nearest neighbor, decision trees, neural network, AdaBoost, gradient boosted tree, and support vector machine out of which random forest resulted in the best performance. Random forest is also the model

With the least standard deviation with test scores making, it the most stable and robust model, it also outdoes the others in terms of testing and OOT scores. He also observed that complex mod\els such as boosted trees, Ada boost, and random forest which is nonlinear had a better performance than the simpler ones such as logistic regression.

The three models which are, logistic regression, random forest, and decision tree were evaluated by [10], using the confusion matrix which analyses how the training and testing are

correctly classified. He evaluated the models based on F1 score, precision, accuracy, and recall. The confusion matrix and its accuracy in predicting if a forthcoming transaction is a fraud or not, for logistic regression is 70%. While for the confusion matrix and its accuracy in predicting if a forthcoming transaction is a fraud or not for decision tree is 72%. This however makes logistic regression better when compared with Decision Tree. Using other parameters such as precision, likelihood ratio, prevalence, and false omission rate, he observed that the decision tree model was a better model as compared to the logistic regression model. To improve the performance of detecting fraud, he made use of a random forest model which was tested, and it was noted that random forest had a better performance than decision tree and logistic regression in terms of recall parameters, precision, and accuracy. He also observed that the major limitation with Random Forest is "overfitting of tree in memory with an increase in data". Overfitting of tree in memory lane is when a model perfectly fits its training data. The major disadvantage of overfitting is that the algorithm cannot accurately perform in predicting the results of untrained data.

According to the experiment conducted by [11], random forest algorithms will have a better performance if the training data has a larger number, however, speed during application and testing will be negatively affected. It was also observed that the Support vector machine algorithm will be affected due to the unbalanced dataset problem which requires more preprocessing to achieve better results.

To validate the respective performances of the effectiveness and efficiency of Support vector machine (SVM) as regards to (Random Forest Classifier) RFC, it is likened with Isolation Forest (IForest) and Local Outlier Factor (LOF). The evaluation metric used by [12] was accuracy, area under curve, and recall. For the recall parameter SVM and RFC had a better performance than LOF, Decision tree, and IForest. While SVM shows better accuracy than RFC. For the area under curve, SVM has a faster and more precise way of identifying credit card fraud than iForest and LOF. He concluded that SVM model has good accuracy and reduces the number of positive transactions that are false by improving the rate of sensitivity in an imbalanced dataset whose fraud rate is very low. In determining credit card transactions that were fraudulent, 70% of data was trained while the remaining 30% was tested. [13]used nine different models and evaluated their performance base on different metrics and performances. The nine models used include, K nearest neighbour, Naïve Bayes, Logistic Regression, Ada Boost, Random Forest, Quadrant Discriminant Analysis, Multilayer Perceptron, Ensemble learning, and Pipelining. He examined the performance of the classifiers using Precision, F1 score, accuracy, Mathew's correlation Coefficient, recall, and Balanced Classification Rate. The result of his research found the performance of Pipelining to be the best method. After implementing Naïve Bayes, Logistic Regression J48 and Ada Boost algorithm, [14]found Ada Boost Algorithm to be the best when compared with the other algorithms using time and accuracy. Logistic Regression and Adaboost had the highest accuracy of 100%, with Adaboost taking a very low time, thereby making Adaboost a better algorithm. The result of the test conducted by [15]showed that Deep Neutral Network performed more effectively in detecting delinquency, having the highest value in Area Under Curve (AUC) which is 0.9246 with recall and precision being the second highest with a value of 0.6053 and 0.9009. Naïve Bayes has the highest recall value which is 0.877. putting Recall and Precision and recall into consideration. Deep Neutral Network outperforms other algorithms by having the highest F1 score of 0.7241. comparing Artificial Neutral Network, logistic Regression, naïve bBaye,sand decision tree, Deep Neutral Network performs better at detecting credit card fraud. The research conducted by [5], Compare and analyzed five machine language models using different feature selection techniques to detect fraud in credit cards. He used MCCaccuracyy, sensitivity, accuracy, recall, and precision to evaluate the performance. Using the information gain method, the prediction accuracy of the PART

classifier and j48 classifier had an increment of 70.5% to 70.2%. while for wrapper method and information gain method the prediction accuracy increased from 70.2% to 71.9%. the precision of random forest classifier, Adaboost classifier, and J48 classifier improved from 78.7&, 77.6% and 78.2 % respectively. The study of machine learning models by [16], such as Random forest, logistic regression, and Naïve Bayes proves its accuracy in detecting fraudulent transactions and reducing the rate of false alerts. [16], used accuracy, f1 score, support recall, and precision to evaluate the performance of the models by comparing all three models. He found Random forest to be a better algorithm than logistic regression and Naïve Bayes. Bagging Ensemble classifier was used by [17] with other classifiers such as Decision tree, Naïve Bayes, K Nearest Neighbor, and Support Vector Machine. Bagging classifier was able to detect fraudulent transactions while still keeping the rate of false alarm very low which is contrary to other methods whose false alarm rate increases when detecting fraudulent transactions. He concluded his experiment with a Bagging classifier based on Support Vector Machine being more stable and less time taking than other methods used. Random Forest and Support Vector Machine was used by [18] as the prediction model. Feature selection techniques were used to improve the accuracy and performance of the training data. The results gotten showed that Support Vector Machine and Random Forest would have a better result when used with Feature selection techniques.

3. Research Methodology

Nowadays, credit card fraud has improved such that a compromise can occur in multiple ways with the cardholder not aware. Most times, fraudulent use of the card is usually unnoticed until the cardholder receives a statement. This research aims to streamline the machine language algorithms relevant to credit card fraud detection, using Knowledge Discovery in Databases (KDD) as the methodology. Knowledge Discovery in Databases (KDD) is the process of obtaining previously unidentified and possibly relevant information or data behavior stored in a database. Knowledge Discovery in Databases consists of the following processes, Data Selection, data clearing, data transformation, pattern searching which are also called data mining, presentations, findings, and evaluations.

The iterative process of KDD has enabled the integration of data, more refining of mined data, enhancing evaluation measures, and the transformation of data to get a variety of appropriate results. [17]

The KDD process has introduced the adoption of data analytic tools and procedures to identify patterns and relationships in data to be summarized into useful information.

The KDD methodology is represented in the figure below, followed by explanations of each of the stages involved in the methodology.

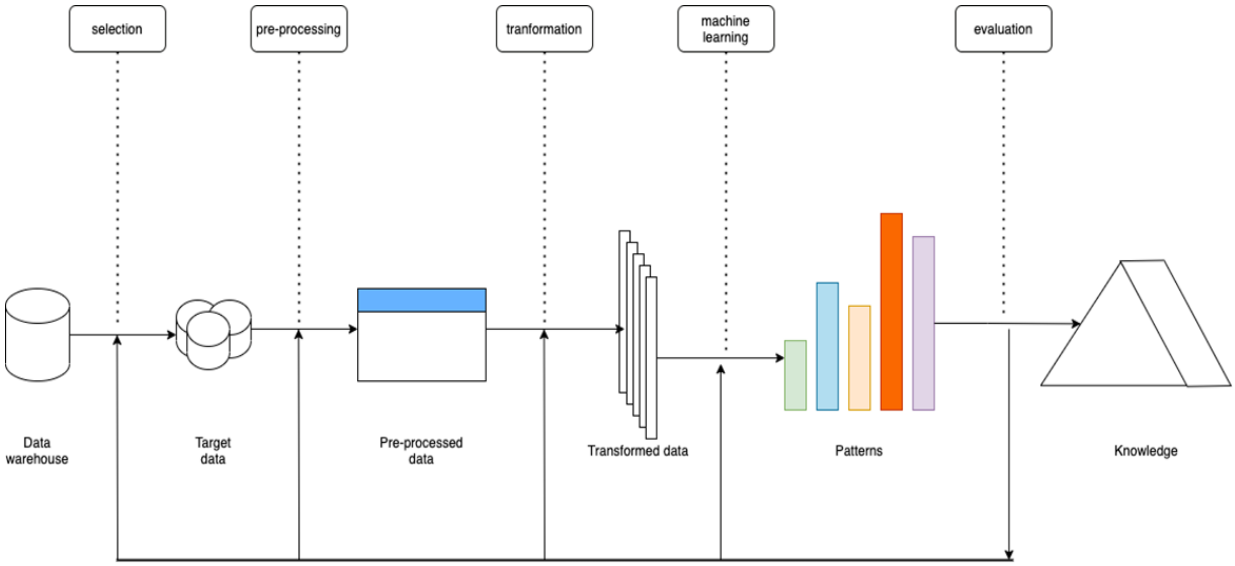


Figure 2: Kdd Methodology for Credit Card Fraud Detection

3.1 Data Selection

This is the first stage in KDD which is the retrieval of data relevant to the analysis from the data collection, which in this case is credit card fraud. The privacy of an individual may be compromised due to unethical access to an individual's personal information, the unauthorized use of personal data which has been retrieved for intent different from the original purpose, the exposure of an individual's data which may cause a form of embarrassment to the original owner of the data, etc. It is therefore of paramount importance to be ethical and follow due process of confidentiality and get the necessary approvals before accessing such information. The retrieval of a dataset should include a variety of attributes and behavior relating to the credit card to help the model recognize and identify certain patterns that can be used by the algorithm for prediction. such as customer profiles and spending behavior. The dataset should also be able to reflect real-world scenarios in capturing credit card transaction behavior [18]. Since the database of a credit card system is application-oriented rather than subject-oriented, the first step is to select the data relevant to the subject, [19], which is credit card fraud. The different information that should be found in the credit card database, includes the history of all accounts with trading log, credit card information, personal data, data of credit card transaction flow, business data, etc. The most relevant of the data set as related to this research includes a record of card log history which contains information like, card ID, card type, expiration date, trade company, liquidation date and amount, account ID, balance and date, etc.

3.2 Pre-process

Databases are sensitive to inconsistent, missing, and noisy datasets. This is because of their typically big size which is usually more than several gigabytes and is usually originated from heterogeneous and multiple sources. The poor quality of a result that has been mined is due to the poor quality of data. Statistical records propose that it requires about 60percent of time in preprocessing data to complete a data mining process [17]. Data must be preprocessed to help improve the data quality and its mining results

There are several data preprocessing techniques.

- Data cleaning: the essence of data cleaning is to fill all voids and nulls, eliminate noise discard inconsistent data, and erase isolated data.

- Data integration: Data integration brings together data from different sources into a datastore which can be a data warehouse. The data sources could be ordinary files, data cubes, or several databases.
- Data reduction: Data reduction minimizes the size of data through clustering, aggregating, or removal of redundant features. This should be done with the integrity of the data still in place.
- Data transformations: Data transformation (e.g., normalization) is applied to data that is required to fall within a smaller range like 0.0 to 1.0. This helps to enhance the efficiency and accuracy of mining algorithms with distant measurements.

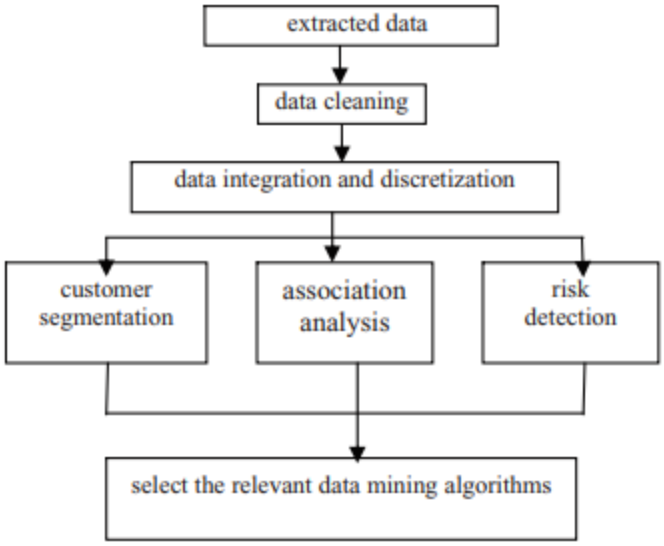


Figure 3: Data preprocessing flow

These processes in preprocessing are not mutually exclusive, they might depend on each other to work together. For instance, data cleaning could involve the use of data transformation to correct data that is labeled as wrong, an instance could be the transformation of a data field entry to a normal format which can be tested with each machine language algorithm to determine which feature is the best in the detection of legitimate and fraudulent credit card transactions.

3.3 Data Transformation

This stage of the KDD methodology entails altering the type, structure, or format of data so that the proposed machine learning algorithms can understand it. Processes like one-hot encoding and data normalization are used in this stage. After processing numerical data, many machine learning algorithms perform better. The term "Data normalization" refers to the process of converting numerical values in a dataset to a single scale without distorting the data's various ranges of values. Min-max normalization and z-score normalization are two typical data normalization steps.

3.4 Machine Learning

Machine learning tracks how the performance of computers can be improved with regards to the data made available. The ability to automatically learn to make an intelligent

decision with regards to the provided data and to identify complex becomes a major research project for computer programmers.
In this stage of the process, machine learning classifier algorithms are used to train and test the models to see how they perform in predicting legitimate and fraudulent credit card transactions. Machine learning is a subpart of the Artificial Intelligence framework which allows us to build machine learning models using historic datasets to imitate human thoughts and later to use the trained models to perform tasks automatically without being explicitly programmed. [20].
There are five (5) types of machine learning which evolved from supervised learning to deep machine learning, which is classified into different methods, each of which has its different algorithm. Find below a taxonomy of different machine language algorithms.
Algorithms such as support vector machine (SVM), decision tree, random forest, logistic regression, and K Nearest Neighbor (KNN) are common machine learning algorithms used in credit card fraud prediction.
The system architecture below shows the proposed techniques used to detect credit card fraud.

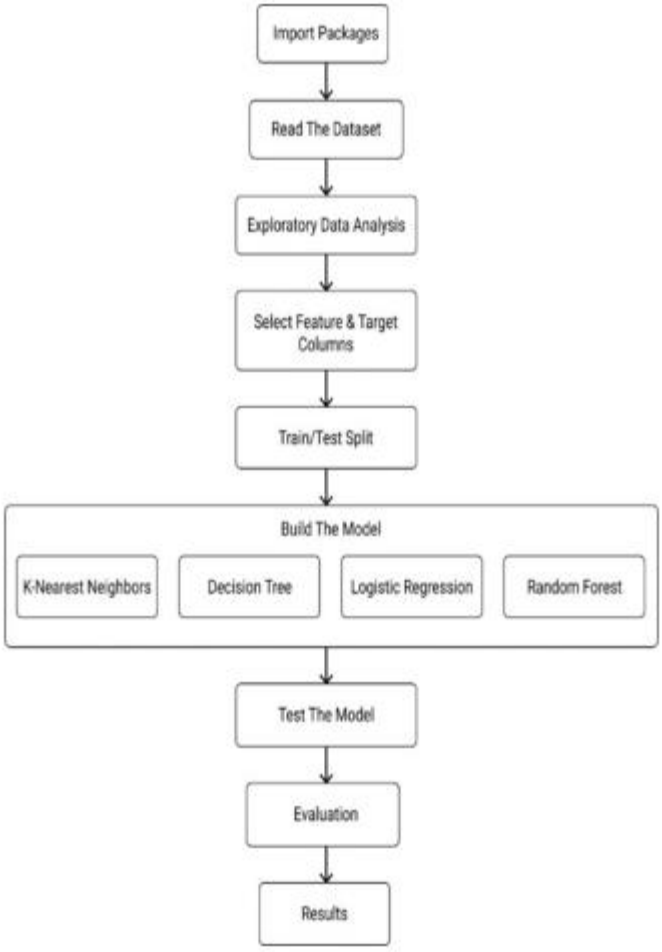


Figure 4: System Architecture to detect credit card fraud

- Support Vector Machine

With SVM, transactions are labeled as either nonfraudulent (legitimate) or fraudulent, which makes it a binary classification. SVM helps to identify if the behavior of a user is fraudulent.

- Decision Tree

The decision tree is a structure with a leaf node, branch node, and root node. Each of the internal nodes represents a test on an attribute, the conditional results of the test denote each branch & the class label holds by each leaf node, each leaf node is also called a terminal node represents the outcome and which does not split further. The highest node in a tree is the root node. The leaf nodes denote a decision, while the branches in a decision tree denote several alternatives. The disadvantage of this method is that implementation is easy, to display and comprehend. However, a disadvantage is the requirement to check each transaction one by one [21].

- Random Forest

Random forest is an ensemble-based machine learning algorithm. Ensemble learning is a type of learning where the same algorithms or a variety of algorithms are merged to create a prediction model that will be powerful. Random Forest algorithm combines different algorithms of the same category i.e., multiple decision trees that leads to a forest of trees, which is why it is called a “Random Forest". Random Forest can be used for classification tasks and regression. Random forest is built from multiple decision trees. [22]

- Logistic Regression

Logistic Regression is classifying algorithms into different categorical values. In Logistic regression, the use of more than one variable (x) that is independent to predict an outcome that is a dependent variable (y). Logistic Regression, which is like Linear regression, tends to predict a discrete target field or label rather than a numeric one. Such as, true/false, successful/unsuccessful yes/no etc. [22].

Logistic regression tends to predict the outcome of a class which is expressed in terms of probability. It tends to predict whether the transactions of a credit card are fraudulent or legitimate.

- K-Nearest Neighbor

A k-nearest-neighbor is a data classification algorithm that helps to determine what group a data point is in by looking at the data points around it

K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most like the available categories.

K-nearest neighbor algorithm is a classification algorithm that takes a group of labeled points which is used to train a system on how the other points are labeled. With this algorithm, the classification of data points is based on how similar they are to other data points. [22].

3.5 Feature Selection Techniques

Feature selection, an important aspect of this research, involves using data mining techniques to select the best set of features which can prove useful in building a predictive model, it’s a technique that involves selecting relevant features and removing redundant features.

Feature selection picks a subset of the most important features (columns) that apply to a dataset. Few features tend to allow machine learning algorithms to run more effectively and more

efficiently with complexity in time with lesser space. Some inputted features that are irrelevant can mislead some machine learning algorithms which can worsen their predictive performance thereby affecting the overall result.

Recursive Feature Selection, Sequential Feature Selection, Correlation-based Feature Selection, and Univariate Feature Selection are examples of feature selection approaches. These feature selection strategies will be compared to each machine learning algorithm to see which one correctly identifies the best attributes for detecting valid and fraudulent credit card transactions

Recursive Feature Selection

This is a feature selection technique that is commonly used. Recursive Feature Selection enables features and their subset to be ranked with their corresponding accuracy. The final subset is usually the subset with a preset number of features (PreNum) and the highest accuracy (HA).

Sequential Feature Selection

Sequential Feature Selector increases (forward selection) or decreases (backward selection) features to produce a feature subset greedily. At each process, the estimator selects the best feature to add or remove depending on the cross-validation score of an estimator.

Correlation-based Feature

The Correlation Based Feature method presumes that irrelevant features show a poor correlation with the class and should therefore be disregarded by the algorithm.

Univariate Feature Selection

Univariate feature selection selects the most appropriate features based on univariate statistical tests. Univariate feature selection can be a preprocessing step for an estimator. Features are individually examined to determine if the response variable has a relationship with the features.

3.6 Evaluation

Evaluation of a model is the process of assessing the accuracy of a model or algorithm of a dataset. Data set are usually divided into subsets; the training set which is used to build models that can be used to predict outcomes and the test subset which is also referred to as unseen data and it is used to assess the future performance of the model. The evaluation of a model requires the use of metrics to quantify the performance of a model. The choice of metrics used is dependent on the machine learning task. Some of the machine learning metrics include accuracy, precision, recall, confusion matrix, etc. [23]

4. Design Specification

The diagram below showcases the process flow that best describes the objectives of this research. It illustrates the process flow from data collection to implementation of results. The figure represents steps taken to implement the model. The steps start with the collection of the dataset which includes details and information of credit cardholders with different transaction behaviors. Data preparation and analysis were carried out to get rid of incomplete and inconsistent values. Feature selection was carried out next to extract the necessary features needed to carry out the testing and training phase. Four classifiers were used to train the model after which performance evaluation was done for each of the classifiers. The test was carried out next to which algorithm was better at predicting fraudulent and legitimate credit card transactions.

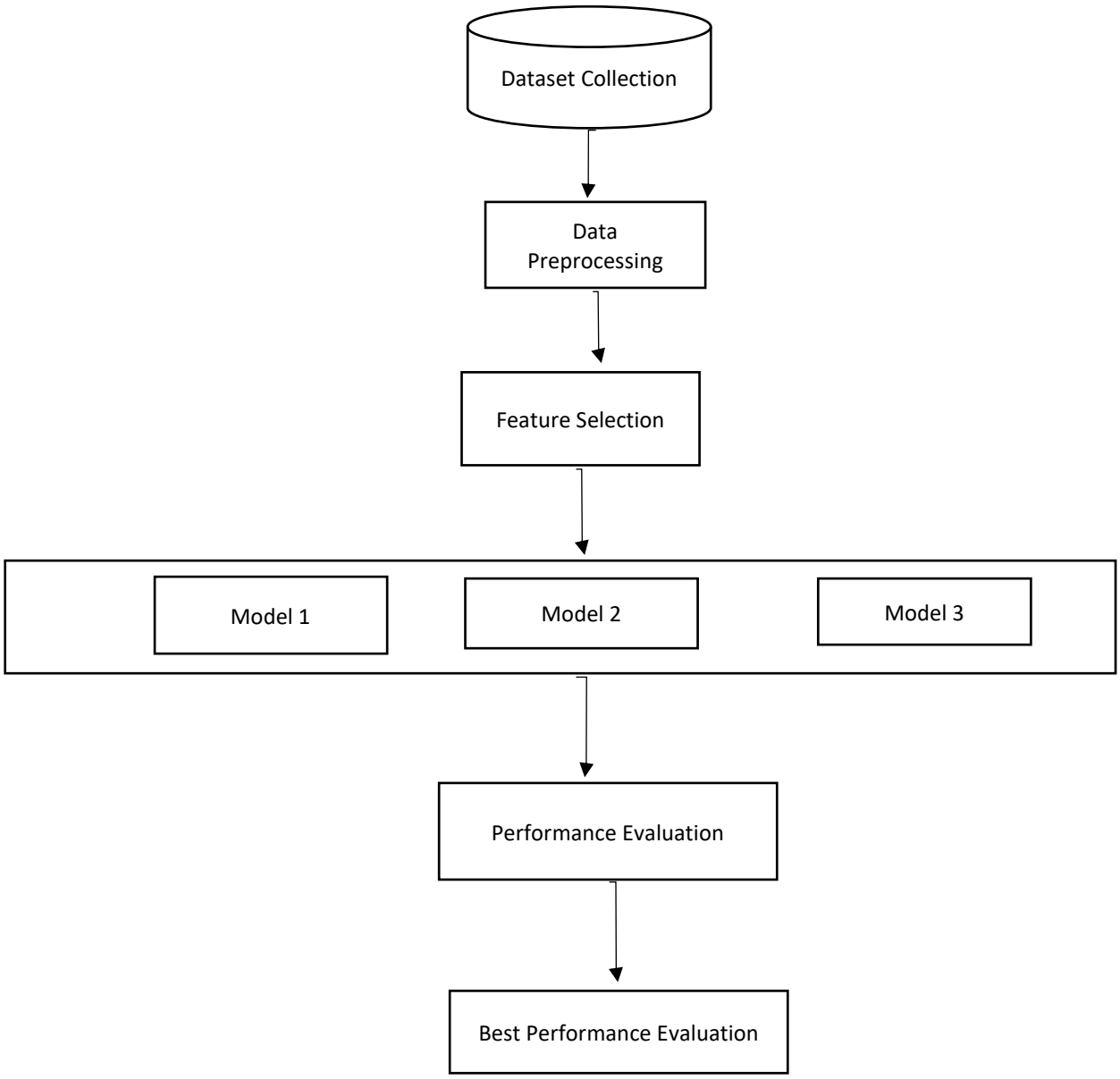


Figure 5: Architectural Design of the Proposed Model

4.1 Data Collection

This research paper made use of a dataset gotten from Data World compiled and collated by Vlad Marisca. The dataset consists of information with about 284,807 entries with a total of 31 columns. Features V1, V2, ... V28 are the principal components. Feature 'Time' is the seconds that go by in between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for dependent cost-sensitive earning. Feature 'Class' is the response variable, and it takes value 1 in case of fraud and 0 otherwise.

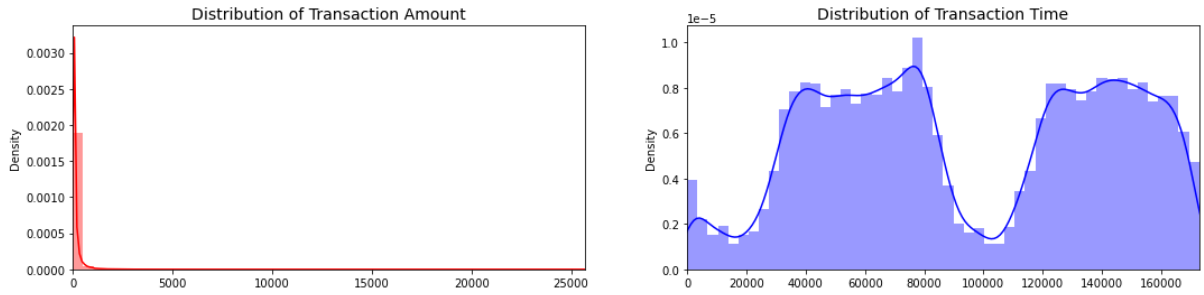


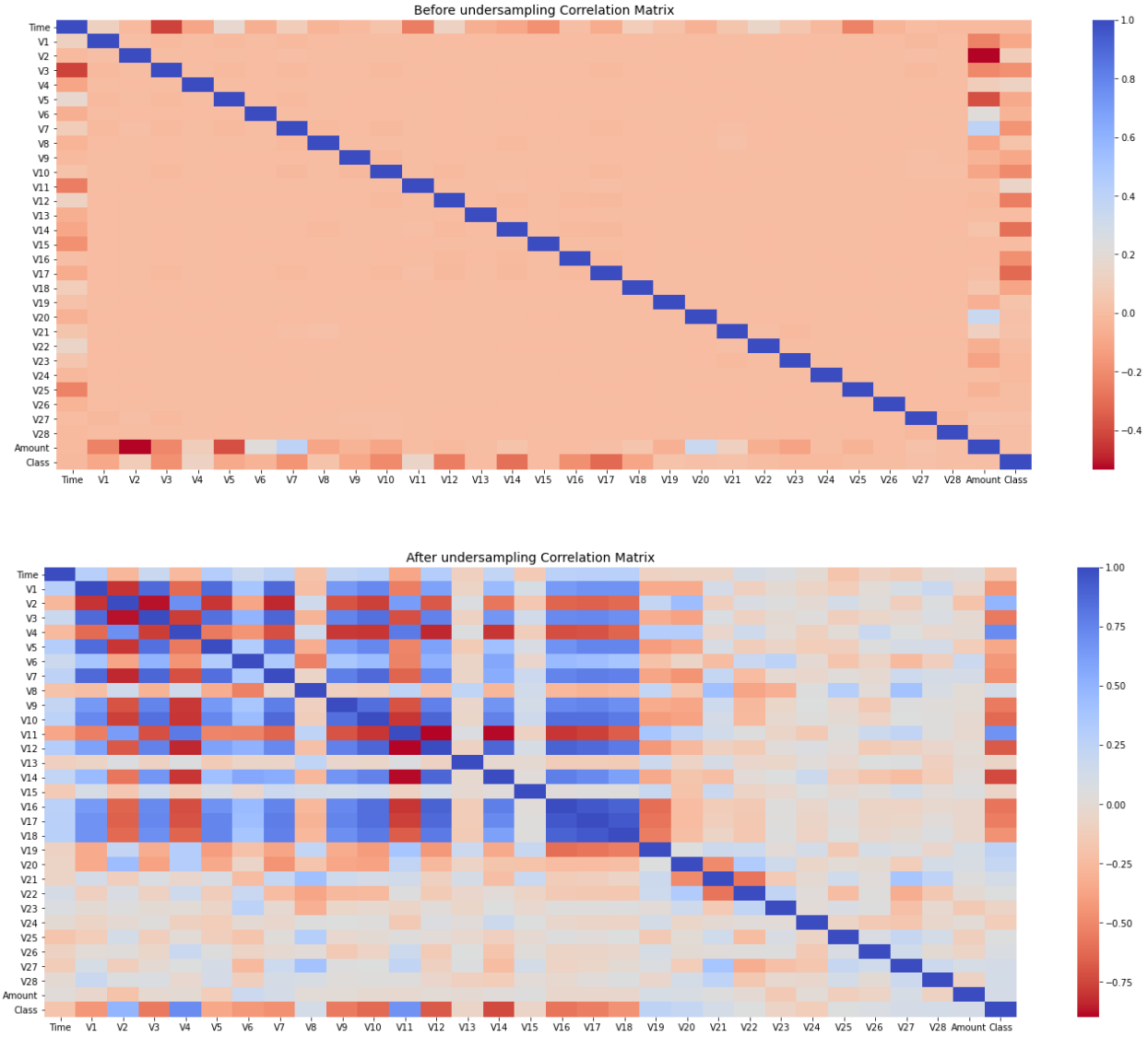
Figure 5: Presents the distribution of transaction time and transaction amount.

Figure 6: Presents the distribution of Transactions

Data Preprocessing

Data preprocessing is the preparation and cleaning of a dataset for training. It involves formatting, arranging, and cleansing, the dataset by removal of unwanted data. It ensures that the information to be trained is without error, complete, correct, and appropriate. Incomplete or unprocessed data can affect the overall accuracy of an algorithm due to different errors

contained in the dataset. In this stage of Data Preprocessing, data is cleansed and transformed in the very process of machine learning to make it easy for the machine in analyzing it, thereby enabling easy interpretation of the dataset by the algorithm.



Data Correlation before and after Undersampling

Feature Selection

Feature selection is the process of selecting a small subset of the best and valuable features from a given dataset with the original features. For this research Three feature selection techniques were used: Univariant Feature Selection, Select from Model Feature Selection, and Hybrid feature Selection.

Univariant Feature Selection: This feature selection method selects the best features of a statistical test.

Select from Model Feature Selection: This feature selection reduces the variables by using specific criteria to select variables that can be used for a predictive model.

Hybrid Feature Selection: This feature combines the various approaches to possibly get the best feature subset.

Modeling

The dataset was divided into training and testing data after which the data was brought together and prepared for analysis and evaluation so that it can be assessed on the test dataset to get the output of the model after it has been designed. The trained data consist of 70% of the dataset while the test data is 30% of the dataset. This helps the model to obtain the utmost accurate outcome by accepting likely inputs to enable the model to identify similar behavior and underlying patterns to check its recognition ability. The algorithms used for this research are, Random Forest, Logistic Regression, XGBoost, and Artificial Neural Network (ANN), using Anaconda, Jupyter Notebook, and Python language, then tensoflow, keras and sklearn, matplotlib, pandas, and NumPy libraries.

Evaluation

Evaluation is the process of assessing the quality of a model's prediction. It involves using machine learning to detect if a credit card transaction is legitimate or fraudulent. Evaluation analyses the performance and attributes of this research paper to show its, efficiency, effectiveness, and impact. For this research paper, the metrics reported when evaluating the models are; Accuracy, recall, precision, F1score, and AUC-ROC.

Accuracy: This refers to the percentage of predictions that were correctly made to a total of all the predictions made.

Recall: this is when the correct positive results are divided by the number of all samples that are relevant which is also a total of the actual positives.

Precision: this is when the number of correct positives is divided by the total predicted observations that are positive.

F1Score: This assesses the test accuracy by considering both the recall and precision of the test to work out the score

AOC-ROC: Area under ROC curve assesses the performance of a binary classifier to differentiate between positive and negative classes.

5. Implementation

Python

Python is a high-level, object-oriented programming language that is easy to use, simple, and can be easily interpreted. Python Syntax which illustrates clarity and readability aims to assist programmers to write logical and clear code for large- and small-scale projects and also minimizes the maintenance cost of the software. Python provides packages and modules which enhance the reuse of codes. Python is however perfect for Artificial Intelligence Projects and Machine learning due to its consistency, simplicity, and its access to outstanding libraries and frameworks for Artificial Intelligence and Machine Learning.

Libraries

The flexibility of Python has motivated a lot of developers to develop new libraries, which has made Python quite popular amongst most Machine learning professionals. Some of the Machine learning libraries used in the implementation of this model include, TensorFlow, keras and sklearn, matplotlib, pandas, numpy libraries.

Tensoflow: Tensoflow is an open-source library that acquires data training models, refines future results, and serves predictions. It uses Python as front end but efficiently runs in C++. Tensorflow brings together machine learning, deep learning, and algorithms. It gives developers the ability to create computational graphs. It has flexible and comprehensive

libraries, tools, and resources that help machine learning and developers conveniently build and deploy machine language-powered applications.
Keras: Keras provides an interface for Artificial Neural networks. It also acts as an interface for Tensoflow. Keras is written in Python, user friendly, and enables back-end computation of Neural networks.
SKlearn: Also known as Scikit-learn is an open-source python machine learning which is built on Matplotlib, NumPy, and Scipy by introducing algorithms for classification, clustering, and regression. Sklearn has a lot of tools that can be used for data analysis.
Matplotlib: Matplotlib is an interactive and very common python library that can be used across different platforms. It can be used for two-dimensional plotting and visualization. It helps with the visualization of patterns In a dataset, it offers various forms of visualization in form of plots and graphs such as bar charts, histograms, error charts, pie charts, etc.
Pandas: Pandas is an easy-to-use high-performance open-source python library with a variety of data analysis, data structure, and data manipulation tools for python language. Pandas can be used to data from different sources such as SQL Database, CSV, Excel, and JSON files.
NumPy: Also known as Numerical Python is one of the best in terms of science and mathematical programming library. It uses matrix processing and multi-dimensional array with high-level mathematical models. It is majorly used for computational analysis which makes it one of the most used Machines learning Python packages.

Anaconda

Anaconda is a desktop Graphical User Interface included in Anaconda distribution. It allows the launching of applications, and easy management of conda packages, channels, and environments without the use of command lines. It is Windows, macOS, and Linux compatible. Anaconda is a distribution of python programming languages for scientific computations such a data science, large-scale data processing, machine learning applications, predictive analytics, etc. It tends to simplify package deployment and management.

Jupyter Notebook

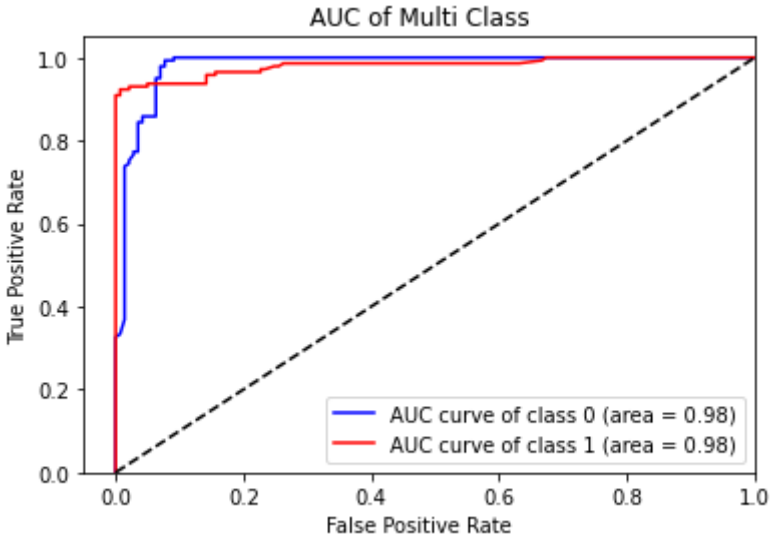
Jupyter Notebook is an open-source web application that enables the sharing and creating of documents that contain live code, visualizations, equations, and narrative text. It is used to transform and clean data, statistical modeling, numerical simulation, and machine learning. It supports over 40 programming languages such as Python. Scala, R, and Julia.

6. Evaluation

This chapter focuses on how effective the evaluation of the proposed model is. The experiment was carried out on a dataset with credit card transaction information. the aim is to analyze the performance of the proposed model in detecting fraudulent and legitimate credit card transactions. The evaluation was done for each algorithm.

Experiment 1: Random Forest Model

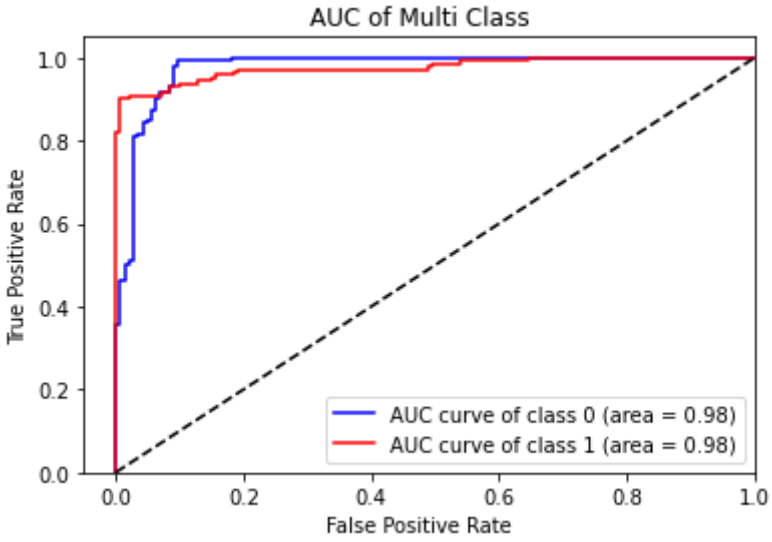
The result of the Random Forest Algorithm is shown In the diagram below, the algorithm got an accuracy, precision, recall, F1 score,e, and ROC of 0.954,0.95,0.92, 0.95, and 0.98.



ROC Curve of Random Forest for Fraud and Non-Fraud

Experiment 2: Logistic Regression

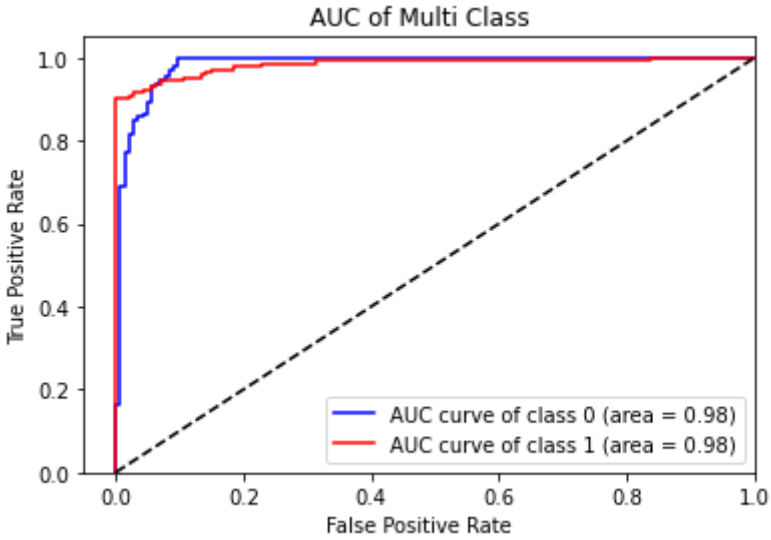
The logistic Regression Model is the second classifier that was evaluated and gave an accuracy of 0.954, precision of 0.93, recall of 0.91, and ROC of 0.98.



ROC Curve of Logistic Regression for Fraud and Non-Fraud

Experiment 3: XGBoost Model

XGBoost is the 3rd model to be evaluated which gave an accuracy of 0.937, precision of 0.94, recall of 0.92, F1 score of 0.94, and ROC curve of 0.98.



ROC Curve of XGboost for Fraud and Non-Fraud

Table 1: Comparison of all Model Results

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC-ROC Curve
Random Forest	0.954	0.95	0.92	0.95	0.98
Logistic Regression	0.954	0.93	0.91	0.93	0.98
XGBoost	0.937	0.94	0.92	0.94	0.98

6.2 Discussion

In the process of answering the research question: "How do feature selection techniques affect the false-positive rates in credit card fraud detection?". The use of feature selection techniques to extract the most useful and relevant variables from the dataset to improve the performance of the algorithm thereby also enhancing each model's ability in detecting and distinguishing true positive fraudulent and legitimate transactions and from false-positive fraudulent and non-fraudulent credit card transaction [26]. From the graphical representation of the ROC Curve and its value, it was deduced that as the True positive rate increases there is a decrease in the false positive rate for both fraudulent and non-fraudulent transactions for all three models used. This implies that the rate of true positive fraudulent credit card transactions and the true positive nonfraudulent transaction is higher than the rate of false-positive fraudulent and non-fraudulent transactions, which tends to confirm the accuracy of the test. The curve in the AUC-ROC shows that with feature selection, the system was able to detect credit card transactions that were fraudulent with a true positive rate of 98% and detect credit card transactions that were legitimate with a true positive rate of 98%. It was also observed from the experiment that out of the three machine learning algorithms used, Random Forest and Logistic Regression had the best Accuracy. However, Random Forest achieved the best result when compared with the other algorithm in terms of Accuracy, Precision, Recall, F1 score, and AUC-ROC Curve.

6.3 Conclusion

Credit card fraud has been identified as one of the leading frauds perpetrated online, which can lead to severe financial repercussions for institutions. In this research paper, the concept of a credit card fraud detection system was discussed. It was observed that feature-based selection techniques can be implemented to achieve the lowest possible false-positive rate. This research attempted to streamline machine learning algorithms using Knowledge Discovery in Databases (KDD) as the methodology. The use of some machine learning algorithms was employed in the detection of fraudulent and legitimate credit card transactions which included Random Forest, logistic regression and XGBoost. Some feature selection techniques, including recursive feature selection and correlation-based feature selection were used to extract the most relevant data from the dataset to improve the performance of the algorithms. Some evaluation metrics were finally used to bring out the best result among the three models after comparing them, thereby answering the research question; "How do feature selection techniques affect the false positive rates in credit card fraud detection?". Furthermore, it can be argued from an ethical perspective that credit card companies, banks, and other related financial institutions should be able to detect credit card frauds. However, the perpetrator is unlikely to operate on the premise of a professional criminal, which suggests that the direct cost of detection from the bank may be too expensive. This would no doubt lead to an ethical dilemma for banks, regarding whether they should attempt to detect credit card frauds despite the financial costs or should they act on behalf of the shareholder and minimize untoward financial costs.

Future Work

For future work, more in-depth research that focuses on genetic machine learning algorithms with more complex feature selection, which would give a better result, even with a robust dataset should be considered.

References

[1] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, and P. Beling, "Deep learning detecting fraud in credit card transactions," *2018 Systems and Information Engineering Design Symposium (SIEDS)*, 2018.

[2] N. Carneiro, G. Figueira, and M. Costa, "A data mining based system for credit-card fraud detection in e-tail," *Decision Support Systems*, vol. 95, pp. 91–101, 2017.

[3] S P Maniraj, Aditya Saini, Shadab Ahmed, and Swarna Deep Sarkar, "Credit card fraud detection using machine learning and data science," *International Journal of Engineering Research and*, vol. 08, no. 09, 2019.

[4] Y. Han, S. Yao, T. Wen, Z. Tian, C. Wang, and Z. Gu, "Detection and analysis of credit card application fraud using machine learning algorithms," *Journal of Physics: Conference Series*, vol. 1693, p. 012064, 2020.

[5] A. Singh and A. Jain, "Adaptive credit card fraud detection techniques based on feature selection method," *Advances in Intelligent Systems and Computing*, pp. 167–178, 2019.

[6] B. Al Smadi and M. Min, “A critical review of credit card fraud detection techniques,” *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2020.

[7] S. V. Lakshmi and D. K. Selvani, “Machine Learning For Credit Card Fraud Detection System,” *International Journal of Applied Engineering Research*, vol. 13, no. 24, pp. 16819–16824, 2018.

[8] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, “Credit card fraud detection - machine learning methods,” *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2019.

[9] V. N. Dornadula and S. Geetha, “Credit card fraud detection using machine learning algorithms,” *Procedia Computer Science*, vol. 165, pp. 631–641, 2019.

[10] S. Patil, V. Nemade, and P. K. Soni, “Predictive modelling for credit card fraud detection using data analytics,” *Procedia Computer Science*, vol. 132, pp. 385–395, 2018.

[11] G. Niveditha, K. Abarna, and G. V. Akshaya, “Credit card fraud detection using random forest algorithm,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 301–306, 2019.

[12] N. Rtayli and N. Enneya, “Selection features and support vector machine for credit card risk identification,” *Procedia Manufacturing*, vol. 46, pp. 941–948, 2020.

[13] S. Bagga, A. Goyal, N. Gupta, and A. Goyal, “Credit card fraud detection using Pipeling and ensemble learning,” *Procedia Computer Science*, vol. 173, pp. 104–112, 2020.

[14] H. Naik and P. Kanikar, “Credit card fraud detection based on machine learning algorithms,” *International Journal of Computer Applications*, vol. 182, no. 44, pp. 8–12, 2019.

[15] T. Sun and M. A. Vasarhelyi, “Predicting credit card delinquencies: An application of deep neural networks,” *Intelligent Systems in Accounting, Finance and Management*, vol. 25, no. 4, pp. 174–189, 2018.

[16] A. Bhanusri, K. R. S. Valli , P. Jyothi , G. V. Sai, and R. R. S. Subash, “Credit card fraud detection using Machine learning algorithms,” *Journal of Research in Humanities and Social Science*, vol. 8, no. 2, pp. 04–11, 2020.

[17] M. Zareapoor and P. Shamsolmoali, “Application of credit card fraud detection: Based on bagging ensemble classifier,” *Procedia Computer Science*, vol. 48, pp. 679–685, 2015.

[18] T. Surthi, “Credit Card Fraud Detection Using Supervised Learning Techniques,” *Science, Technology and Development*, vol. VIII, no. XI, pp. 203–210, Nov. 2019.

[19] J. Han, M. Kamber, and J. Pei, *Data Minning Concepts and Technology*, Third. Waltham: Morgan Kufmann, 2012.

[20] S. Sorournejad, Z. Zojaji, R. E. Atani, and A. H. Monadjemi, "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective," *Cornell University*, 2016.

[21] Z. Yan-li and Z. Jia, "Research on data preprocessing in credit Card consuming Behavior Mining," *Energy Procedia*, vol. 17, pp. 638–643, 2012.

[22] P. K. Sadineni, "Detection of fraudulent transactions in credit card using machine learning algorithms," *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2020.

[23] V. Shah, P. Shah, H. Shetty, and K. Mistry, "Review of credit card fraud detection techniques," *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2019.

[24] V. Vijayakumar, "Recent trends and challenges in real time face recognition system in video," *2014 Sixth International Conference on Advanced Computing (ICoAC)*, 2014.

[25] K. Ramasubramanian and A. Singh, "Machine learning model evaluation," *Machine Learning Using R*, pp. 483–531, 2018.

[26] S. Mittal and S. Tyagi, "Performance evaluation of machine learning algorithms for credit card fraud detection," *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2019.