REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

DEVELOPING AN ARTIFICIAL INTELLIGENCE BASED SYSTEM TO DETECT FRAUD IN CREDIT CARD TRANSACTIONS

Ali Abdulsalam Hussein HUSSEIN

Master's Thesis

Supervisor

Dr. Oğuz KARAN

Istanbul, 2023

DEVELOPING AN ARTIFICIAL INTELLIGENCE BASED SYSTEM TO DETECT FRAUD IN CREDIT CARD TRANSACTIONS

Ali Abdulsalam Hussein HUSSEIN

Electrical and Computer Engineering

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

The thesis titled Developing an Artificial Intelligence Based System to Detect Fraud In Credit Card Transactions prepared by (Ali Abdulsalam Hussein HUSSEIN) and submitted on (12/04/2023) has been **accepted unanimously/by majority of votes** for the degree of Master of Science in ECE

Dr. Öğr. Üyesi Oğuz KARAN

Thesis Defense Committee Members:

Dr. Öğr. Üyesi OĞUZ KARAN	Department, University	_____
Dr. Öğr. Üyesi SEFER KURNAZ	Department, University	_____
Dr. Öğr. Üyesi ABDULLAHI ABDU IBRAHIM	Department, Altinbas University	_____
Dr. Öğr. Üyesi Serdar Kargın	Department, University	_____
Dr. Öğr. Üyesi Zeynep Altan	Department, University	_____

I hereby declare that this thesis/dissertation meets all format and submission requirements of Electrical and Computer Engineering master's thesis.

Submission date of the thesis to Institute of Graduate Studies: 12/04/2023

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Ali HUSSEIN

Signature

ABSTRACT

DEVELOPING AN ARTIFICIAL INTELLIGENCE BASED SYSTEM TO DETECT FRAUD IN CREDIT CARD TRANSACTIONS

HUSSEIN, Ali

M.Sc., Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst.Prof. Dr. Oğuz KARAN

Date: April / 2023

Pages: 72

This thesis uses machine learning and deep learning to detect credit card fraud. The Worldline-ULB Machine Learning Group dataset was used for big data mining and fraud detection. 492 of 284,807 transactions are fraudulent. Just 0.172% of transactions are favourable, creating a severe imbalance. The prediction algorithm favours most samples when fed significantly skewed class distribution data. Hence, it often conceals fraudulent transactions. A data-level strategy that included multiple resampling methods is used, including random under sampling, Tomek links removal, random oversampling, Synthetic Minority Over-sampling Technique (SMOTE), and a hybrid SMOTE-Tomek links removal strategy to tackle this problem. To solve class disproportion, bagging and boosting algorithms is employed. Finding the optimal classifier required several experiments. The suggested system compares logistic regression, support vector machine, naïve bayes classifier, decision tree, random forest, XGBoost, modified XGBoost, KNN, and multi-layer perceptron. These studies employed many training features and class distributions. The class distribution design of training sets must be carefully considered due to inconsistencies between training and testing data in classifier evaluations created from distinct sets of trials. TN and FN rates were included in this evaluation. Random forest appears to be the strongest model, with 99.957% accuracy, an F1-score of 0.875, recall of 0.857, precision of 0.893, and ROC AUC of 0.98. The XGBoost model improves, but most measurements suffer. MLP model accuracy and ROC AUC are excellent regardless of F1 score, recall, or precision.

Keywords: Artificial Neural Network, Credit Card, Deep Learning, Fraud, Machine Learning.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

AI	:	Artificial Intelligence
CNP	:	Certified Nurse Practitioner
AVS	:	Address Verification Service
CVM	:	Card Verification Method
PIN	:	Personal Identification Number
FDS	:	Dynamic Fraudulent Conduct
AP	:	Average Precision
AUC	:	Area Under the Curve
SMOTE	:	Synthetic Minority Over-Sampling Method
XGBoost	:	Extreme Gradient Boosting
APATE	:	Anomaly Prevention Via Advanced Transaction Exploration
SCARFF	:	Scalable Realtime Fraud Finder
HMM	:	Hidden Markov Model
VISA	:	Visitors International Stay Admission
POS	:	Point-Of-Sale
Cis	:	Card Issuers
FI	:	Financial Institution
TX	:	Tax form
LR	:	Logistic Regression
SVM	:	Support Vector Machine
NB	:	Naive Bayes
DT	:	Decision Tree
RF	:	Random Forest
KNN	:	K-Nearest Neighbors
MLP	:	Multi-Layer Perceptron

1. INTRODUCTION

Since it was first made available, online shopping has seen phenomenal expansion. To increase output in international commerce, it has become regular practice for both private firms and government organizations to adopt this instrument in order to achieve their goals. One of the most important contributors to the growth of the e-commerce business has been the widespread adoption of credit cards as a method of making payments [1]. Every time there is a conversation about monetary transactions, someone always brings up the subject of financial fraud. When a con artist steals money from an unsuspecting victim with the intention of getting some form of benefit, whether monetary or otherwise, this is an example of financial fraud [2], sometimes known as financial embezzlement. The dramatic growth in popularity of online shopping over the past few years has made the use of plastic payment cards the standard. This shift has occurred concurrently with an increase in the number of credit card thefts. Because of the large amounts of money that are being lost due to fraud, businesses and government organizations are facing a severe problem. According to The Nilson Report [3], the total amount of money lost by individuals and companies due to fraudulent activity using credit, debit, and prepaid cards in 2015 was $16.31 billion. According to the most recent information that can be found in The Nilson Report [3], the amount of money that was lost due to fraud in 2018 was a total of $22.8 billion. The fact that it has increased by 4% since 2015 is indicative of the likelihood that this pattern will continue into the foreseeable future. In 2012, the entire amount of money lost due to fraud was around $5.6 billion, however in 2018, the total amount of money lost due to fraud was over $9.1 billion, which is equivalent to almost two-fifths of the total amount lost. Seventy percent of these instances are the result of card-not-present fraud, often known as fraud, conducted online or over the phone. This type of fraud is referred to as CNP fraud. Twenty percent of these incidents are caused by counterfeits, while ten percent are the consequence of cards being misplaced, stolen, or lost [2]. Post-hoc detection, also known as investigation, carried out after the fact, and proactive prevention are the two primary approaches that may be taken to address the problem (what happens beforehand). Two examples of verification services that are often used for the purpose of preventing fraud are the Address Verification Service (AVS) and the Card Verification Method (CV2) (CVM). Data mining and rule-based filtering are only two of the many tools that are utilized extensively in preventative measures

[4]. On the other hand, when fraud is unavoidable, it is of the utmost importance that it be uncovered as soon as possible so that necessary actions may be taken. Authenticity of a transaction can only be determined with the use of fraud detection [5]. Because it is impossible for humans to manually assess each transaction to decide whether or not it is fraudulent, automated fraud detection solutions are necessary.

1.1 BACKGROUND

Figure 1.1 depicts the terminal that is used for the initial step in determining whether or not the transactions are legitimate. At the terminal, checks are performed to confirm that the transaction is authentic. These checks include validating the personal identification number (PIN) and checking the balance. A predictive model's role is to give each legitimate transaction a score, and then use that score to decide whether or not the transaction should be approved. If the score is high enough, the transaction should be accepted. Each false alarm will cause an investigation to be performed, and the results of those investigations will be included into the prediction model so that it can become more accurate over time [6].

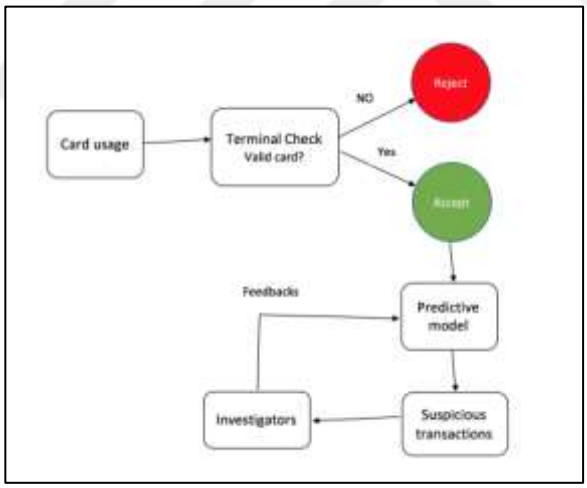

Figure 1.1: Card Fraud Detection.

The creation of a system that is able to detect fraudulent behaviour is a more difficult task than it may appear to be at first look. The knowledgeable individual is responsible for selecting a learning method (supervised or unsupervised), an algorithm (such as logistic regression, decision trees, etc.), features, and most crucially, a strategy for addressing the problem of class imbalance [6]. The divide that exists between different socioeconomic classes is yet another significant obstacle that has to be overcome in the fight against fraud.

The vast majority of machine learning algorithms are unsuccessful in this context [6] because the data pertaining to the transactions are not very detailed. This is due to the fact that the real and fraudulent classes share much too much overlap with one another.

A fraud detection model, when triggered, assesses a transaction to confirm its validity and informs the appropriate authorities to any transaction that raises the reddest flags for potentially fraudulent activity. After that, investigators delve further deeper, and then they incorporate the findings of that investigation into the system that detects fraudulent activity. However, this strategy can be arduous and time-consuming for investigators, resulting in the timely confirmation of just a tiny fraction of transactions overall. Because of this circumstance, the predictive model receives fewer feedbacks, which frequently leads to a model that is less accurate [7].

In response to the growth in the number of instances of identity theft using credit cards, several banks and other types of financial institutions have spent enormous sums of money and gathered teams of specialists in order to develop fraud detection systems. Academics have dedicated a significant amount of their time and resources to the study and development of improved fraud detection systems in order to solve difficulties such as class imbalance, class overlap, and dynamic fraudulent conduct (FDSs). In what comes next, we'll take a look at some of the older research that were done in this area, and we'll compare and contrast their findings [8]. For the purpose of their study, Pozzolo et al. [9] compared and contrasted two approaches to the detection of fraudulent activity: the static approach, in which a detection model is trained at regular intervals (for example, monthly or annually), and the online approach, in which the model is updated as soon as possible after new transaction data is received. In the static approach, a detection model is trained at regular intervals (for example, monthly or annually); in the online approach, a detection model is the authors came to the conclusion that the online method was more effective than other approaches in spotting instances of fraud. They argue that the most effective tactic is to get as much knowledge as possible on the avoidance of fraud through web research. Average Precision (AP), Area Under the Curve (AUC), and Precision Rank are the three measures that Pozzolo et al. [9] have determined to be the most effective ones for the task of fraud detection. Within the context of a fraud detection task research, Pozzolo et al. [6] discovered that random forest performed better than other techniques. Awoyemi et al. [1] utilized K-nearest neighbour, logistic regression, and Naive Bayes in their analysis of data obtained from credit card

transactions. After that, the data were resampled using a technique called the Synthetic Minority Over-Sampling Method (SMOTE). According to the findings, the option that uses the k-nearest neighbour is preferable to the other two options. The Matthews correlation coefficient, the balanced classification rate, and specificity were some of the other metrics that were utilized in the process of evaluating the outcomes.

1.2 PROBLEM STATEMENT

When it comes to disclosing customer information, financial organizations like banks are notoriously coy, and for good reason: protecting customers' privacy is a top priority. In the area of research that focuses on the identification of fraudulent activity, this represents a substantial challenge that has to be conquered. The purpose of this thesis is to implement machine learning strategies in order to carry out predictive analysis on a dataset that is made up of credit card transactions. The final objective of this analysis is to determine which of the transactions contained in the dataset are fraudulent.

The objective here is to determine, through the use of predictive models, whether a certain transaction in question is honest or dishonest. Before presenting the findings, the dataset will be processed through a number of machine learning procedures, such as logistic regression, random forest, and XGBoost. These procedures will be used to analyse the data. After this, the study will proceed to address the problem of class disparity by utilizing a number of different examples of different solutions.

1.3 CHALLENGES IN FRAUD DETECTION

Putting together a mechanism that can detect fraudulent activity is a more difficult task than it may appear to be at first look. The practitioner is the one who is responsible for making decisions regarding a learning strategy (such as supervised or unsupervised learning), algorithms (such as logistic regression, decision trees, etc.), features, and, most importantly, how to handle the class imbalance problem (fraudulent cases are extremely rare in comparison to legitimate cases) [6]. Inequality across different socioeconomic classes is another significant problem that has to be addressed in the area of fraud detection. There is also a problem with the classification task because there is overlap between the real and fraudulent groups as a result of the lack of detail in the transaction data [10], and the majority of machine learning algorithms are not successful in performing well in these kinds of situations [11].

An alert for the transaction that investigators should be the most concerned about is generated by a fraud detection model, which predicts the kind of class (genuine or fraudulent). Following this first stage, investigators will conduct follow-up inquiries and provide recommendations about how the effectiveness of the fraud detection system may be improved. Even while this is required, it can be a burden for investigators, which results in just a tiny fraction of transactions being validated in a timely manner. Because there are only a certain number of feedbacks that can be used, the predictive model will probably have less of an impact [9]. In conclusion, genuine financial statistics are hard to come by due to the fact that financial organizations seldom make customer data public due to the confidentiality concerns that are associated with such actions. In the subject of research pertaining to the identification of fraud, this is an important challenge that has to be conquered.

1.4 OBJECTIVE

The major objective of this thesis is to make use of machine learning techniques to carry out predictive analysis on a dataset consisting of credit card transactions, with the end aim of determining which of those transactions are fraudulent within the dataset. Utilizing predictive models to ascertain whether a specific transaction comes under the "regular" category or the "fraudulent" category is the objective here. The dataset is going to be analysed with a variety of machine learning techniques, such as logistic regression, random forest, and xgboost, and the findings are going to be reported after that.

1.5 OUTLINE

In the first chapter, the effects that fraud has on the financial market is discussed, as well as provided a brief overview of the process of detecting fraud, the challenges that are inherent in the process of designing an efficient fraud detection system, and a method that is recommended for accomplishing this task. Following the discussion of the evolution of research into methods for detecting credit card fraud in Chapter 2, more details in order to gain a fundamental comprehension of machine learning.In order to have a better grasp of the analysis, additional depth about the methodologies used will be presented in Chapter 3.

In Chapter 4, the primary focus is on conducting an investigation into the experimental comparisons of various models for asymmetrical data streams.

The conclusion, a summary of the findings, and some ideas for going forward will all be included in Chapter 5.

2. LITERATURE REVIEW

Before digging further into a particular application domain, it is recommended to first become familiar with some of the fundamental concepts behind machine learning. This chapter will discuss the recommended technique to detecting fraud using machine learning, as well as how that approach operates. Particularly, focusing on supervised learning, which is a form of machine learning in which the model is educated through the use of data that has already been labelled.

2.1 RELATED WORK

Because identity theft involving credit cards has become such a significant problem in the financial industry, a number of financial institutions have made significant investments in the research and development of fraud detection technologies and assembled teams of human experts to combat the issue. A number of scholars have been putting a lot of effort into developing better fraud detection systems in order to solve difficulties such as class imbalance, class overlaps, and shifting fraudulent conduct (FDSs). In this part, some of the most influential research that has been done is presented in this field throughout the years.

Pozzolo et al. [9] recently investigated two distinct approaches to the detection of fraudulent activity: the static approach, in which a detection model is trained at regular intervals (such as monthly or annually), and the online approach, in which the model is updated as soon as possible after new transaction data is received. In the static approach, a detection model is trained at regular intervals (such as monthly or annually); in the online approach, a detection model is trained at regular intervals (such as monthly They contend that training in this manner can be more successful than the conventional method of learning in a classroom setting due to the fact that fraudulent practices change over time. In addition, Pozzolo et al. [9] proposed that the Average Precision (AP), the Area Under the Curve (AUC), and the Precision Rank are the most useful metrics for the work of fraud detection. Pozzolo et al. [12] came to the conclusion that the random forest method is the most efficient approach to the task of detecting fraudulent activity.

After using K-nearest neighbor analysis, logistic regression, and the Naive Bayes method to a dataset consisting of credit card transactions, Awoyemi et al. [1] Synthetic 's Minority Over-sampling Technique was used to resample the data. This technique was developed by Awoyemi (SMOTE). The K-nearest neighbour algorithm was found to be the most

successful solution out of the three that were evaluated. The findings were evaluated using the recall rate, the accuracy rate, the balanced classification rate, the specificity rate, and the Mathews correlation coefficient.

In the paper "Lebichot et al. [13], the authors offer an approach that uses graphs to create a system that can detect fraudulent activities. Using a method called collective inference, the fraudulent influence was spread throughout a network with the use of particular fictitious transaction data that was deployed as part of this plan. A previous fraud detection system known as APATE (Anomaly Prevention via Advanced Transaction Exploration) was improved upon by Lebichot et al.[13] in a variety of ways in order to accomplish this goal.

Carcillo et al. [14] have proposed an innovative solution to the issue of detecting fraudulent activity in financial transactions. They studied the possibilities of big data technology in the field of fraud detection using big data technologies such as Cassandra, Spark, and Kafka. Based on their findings, they designed a detection system that they termed the Scalable Realtime Fraud Finder (SCARFF). They made a point to emphasize the fact that big data technology has advantages over traditional system design in terms of scalability, fault-tolerance, and accuracy. This is due to the fact that fraud detection needs to take place in a real-time setting and because of the enormous volume of transaction data.

Domingos [15] presented the idea for the cost-conscious Meta Cost model. His research has shown that not all instances of incorrect categorization result in the same level of financial damage. As a result, he has devised a method for including the cost of misclassification into the classifier's approach to minimizing costs. According to the findings of the research, he arrived at the conclusion that the utilization of Meta cost is connected to a systematic reduction in total cost when compared to the utilization of alternative classifiers that are insensitive.

Database mining with neural networks has been suggested as a method for identifying fraudulent activity by Aleskerov et al. [16]. Srivastava et al. [17] suggest a Hidden Markov Model (HMM) for spotting fraudulent behaviour by evaluating a customer's spending habits. This may be accomplished by using a customer's transaction history. According to the findings of Wheeler and Aitken's research, the adaptive approach has the potential to filter and order the fraud cases, which will result in a reduction in the total number of fraud investigations. In their paper [18], Baader and Krcmar propose an automated feature engineering technique with the goal of reducing the high proportion of false positives that is

typically observed in fraud prediction findings. There has been a lot of work done on various aspects of fraud detection, such as keeping tabs on cardholder behaviour, speeding up the fraud detection processing time, reducing the costs of misclassification, and using different techniques for data mining. However, there has been very little work done on how to deal with the class imbalance problem in the fraud detection task. As a result, the goal of this thesis is to perform predictive analysis on the task of detecting credit card fraud, with a major emphasis on resampling, ensemble, and hybrid techniques to handle the class imbalance issue.

2.2 CREDIT CARDS

2.2.1 Legacy

According to this view, there was credit long before there was ever monetary trade. On an Assyrian tablet dating back to the year 2000 B.C., there is a depiction of a credit transaction. It's likely that the signet rings that Teutonic knights used to wear in the Middle Ages were probably the thing that came the closest to a credit card in those days. The artisan who carved the coat of arms of the knight onto each ring also kept note of which rings were which. Following the registration of the rings, the lists were distributed to the various local businesses. When the knight conducted business, he would press the signet ring that was on his finger into molten wax, and that would serve as his signature [19]. The castle would receive the bills that needed to be paid from the various merchants and artisans. These bills would be paid at a later time.

Near the end of World War I, American oil and hotel firms were the first to present travel cards to a small set of customers as a means of payment for their services. These cards, in essence, served as an introduction to the individual as a valued client of the firm that issued them as well as a guarantee that the items and services acquired would be paid for once the consumer returned to their home. In 1950, an American entrepreneur named Frank X. MacNamara was the first to propose the idea of taking credit cards in establishments other than restaurants. This idea was the inspiration behind Diners Club, a credit card that can be used at restaurants as well as for other experiences. Companies providing financial services, such as Carte Blanche and American Express, began to compete in the market not long after. The Franklin National Bank of New York was the institution that in 1951 issued the first bank card, which is widely regarded as the beginning of the credit card business. Over the

following four years, about one hundred additional financial institutions (Fis) such as banks and credit unions will issue cards. Both the Bank Americard, which would eventually become VISA, and the Master Charge would make their debut in 1956 (Now MasterCard). Two instances of fraudulent activity were reported on the very first day that Chargex credit cards, which are now known as VISA cards, were available for use in 1968 [19].

2.2.2 Easy And Convenient Payment Options

A credit card is a one-of-a-kind product [20] owing to the fact that it enables millions of individuals all over the world to make purchases on credit for up to 51 days from the date of the transaction, with no interest accruing on the amount if it is paid in full by the day that the statement is due.

Credit cards are a common method of payment for purchases conducted across the continent of North America. The following are the most essential variables that have contributed to the success of this trend:

a. The existence of a fully developed point-of-sale (POS) infrastructure.
b. The decreased requirement to carry cash, as well as the advantages of interest-free credit for a period of a few weeks, in addition to a number of additional services and benefits, such as frequent flyer miles, insurance, and more.
c. The cardholder has a maximum responsibility of $50 in the case of card loss or theft, provided the cardholder notifies the financial institution of the loss or theft as soon as possible. This ensures the safety of the money.

Processing payments made by credit card streamlines company transactions, which is to the benefit of all parties involved. Card issuers (Cis) have a few different options available to them when it comes to how they might generate revenue: (1) by charging fees to merchants; (2) by charging fees to cardholders; and (3) by charging interest on outstanding balances.

A consumer uses a credit card to make payments on a line of credit that has been provided to them by the bank that issued the credit card when they make a purchase (CI). After the CI has used a credit card to pay the vendor, the buyer is responsible for paying the outstanding balance to the CI. Because the claim that is delivered as payment is considered to be the responsibility of the credit card issuer, the risk that a transaction will fail because there are insufficient funds is transferred from the seller to the credit card issuer. The amount of money that will never be recouped as a result of these losses is taken into account when calculating

the annual fees and interest rates that Cis charges to make up for these losses. It is essential to keep in mind that the overwhelming majority of Cis are in fact Fls, despite the fact that there are a great number of Cis who are not Fls.

2.3 CREDIT CARDS TRANSACTION PROCESS

2.3.1 Transaction

It takes the participation of four distinct parties—the cardholder, the merchant, the financial institution (FI), and VISA—in order to successfully conduct a transaction that involves a credit card.

If the whole amount of the cardholder's monthly bill is paid in full before the due date, the cardholder will not be subject to any interest charges for making payments with their credit card. The increased peace of mind that comes with making a payment with FIs. is a perk for customers of businesses who accept the card.

It is the duty of the FIs to issue the cards, settle cardholder and merchant transactions conducted by other FIs with VISA, process incoming transactions, and deliver monthly statements to cardholders.

Fourth, VISA ensures that the process of conducting transactions is standardized and offers a neutral platform for cardholders, retailers, and financial institutions to connect with one another. It also acts as a transaction log, in addition to recording purchases and being used for marketing purposes.

It is common practice for a transaction to be unable to take place without both the cardholder's and the retailer's FIs being present. Customers of a given FI have the option of shopping at establishments that are part of either the same FI or a different FI. Therefore, it is possible for the FI to operate as an agent for either the cardholder or the merchant; yet it is also possible for the FI to function as an agent for neither of them. Therefore, the entering transactions of a particular FI need to be separated from those of the other FIs, and then they need to be sent to the appropriate location within the transaction processing system so that they may be authorized and records kept.

When making a purchase with a credit card over the internet, the authorization step is the single most important component of the whole process of executing the transaction. The authorization step acts as the cardholder's first line of defence against illegal use of their account and as a method of keeping their available credit under tight control. It also serves

as a means to maintain their available credit under tight management. When the transaction is placed into the cardholder file, the authorization is temporarily saved for up to five days, and the cardholder's account balance is brought up to date at the same time. In addition to the commonplace practice of obtaining online authorizations, several businesses additionally employ stricter constraints. If the total amount of the transaction is less than the threshold amount, the merchant can avoid using the financial institution's system and instead authorize the transaction on their own if the total amount of the transaction is less than the threshold amount. In point of fact, many businesses in North America have zero' floor restrictions as a result of the pervasive POS network that exists there; virtually all transactions need to be allowed online by the appropriate FI.

An authorization attempt is carried out whenever a credit or debit card is used to make a purchase. A magnetic stripe on the back of the card contains encoded information about the cardholder's name, account number, credit limit, and expiration date. This information may be read by a point-of-sale (POS) system. The authorization is considered to be complete after the transaction has been approved and the cardholder has signed the associated receipt. After it has been finished, the TS can be delivered to the FI either manually or electronically, depending on whatever method the sender prefers. The POS device keeps a history of all authorizations and sends them to the Fl when an electronic transfer is being processed. Because the slips are not present, merchants are exempt from the requirement that they submit them. In the manual method, the store sends the slips to the FI through regular mail, and the workers at the FI key them in by hand. In any case, when the merchant has submitted their transactions, the financial institution will deposit the appropriate sum into the merchant's account.

In order to effectively handle the monthly billing for a very large number of cards, Fis has implemented numerous billing cycles for each month. Every cycle features its own one-of-a-kind billing date as well as its own individual collection of connected cards. A fresh statement is prepared and issued to the cardholder after the conclusion of the billing cycle. The cardholder also receives a new payment due date at this time. On the statement, you will find information on the transactions, such as the dates on which they took place, the purpose of the transactions, the amount of money involved, and the purpose of the transactions. Additional information, including the previous and current balances, the amount of the minimum payment that is required, and the total credit that is available, is also displayed.

The cardholder is the one who is responsible for paying off any outstanding balance, in full or in part. There will be no further charge for interest if the remaining balance of the obligation is paid in full. If the payment made by the cardholder is less than the minimum payment that is required, the payment may be considered late, and the impact on the cardholder's credit may be severe. Cash advances are another option available to those who utilize their credit cards. Even if the whole sum is paid off on the date that the statement is due, interest will still begin to accumulate on the day that the withdrawal was made. In order to effectively manage the enormous volume of daily transactions, both FIs and VISA have developed a non-stop, real-time computer system comprised of computer hardware and software. This technology not only processes data and keeps records, but it also makes communication between FIS and the VISA network easier. Figure 2.1 provides an overview of the VISA transaction processing system from a more high-level perspective.

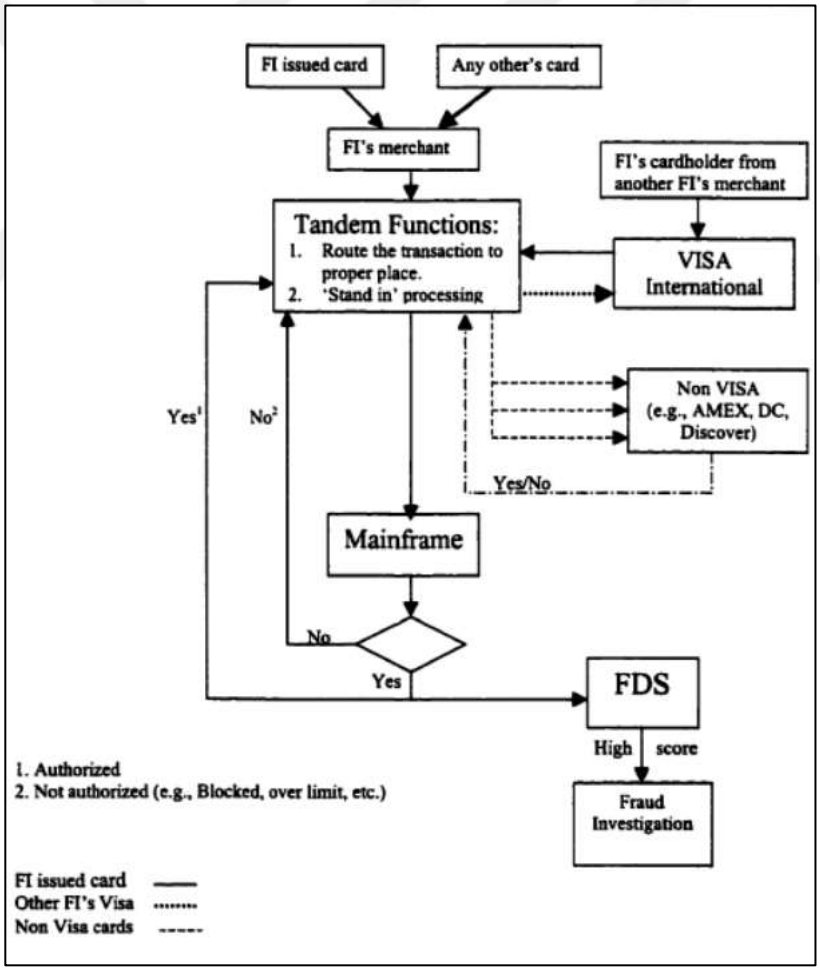

Figure 2.1: VISA Processing Transaction System.

2.3.2 Parties Involved

A magnetic stripe reader is available at the cash register for customers to use in order to process payments made with your credit card. On the face of the card, users will find information on the card, including the cardholder's name, the account number, the expiration date, and a hologram. Today, consumers have access to a wide variety of options for purchasing cards. There are three primary categories that may be applied to standard playing cards: (1) the classic, (2) the gold, and (3) the platinum. Traditional credit cards often do not come with any kind of rewards program or levy any kind of annual fee. In addition to having a substantially higher credit limit than normal cards, gold and platinum cards generally have annual fees associated with their use. Cards of this type generally come with a variety of services and perks, such as insurance for car rentals.

In North America, the authorization of online purchases is often performed by means of swipe machines, also referred to as point-of-sale terminals. When a consumer swipes their credit card at a point-of-sale terminal and enters the total of their purchase through the keypad, the terminal receives the information from the magnetic stripe on the card and begins a phone call to the server belonging to the merchant's financial institution. Through the use of a modem, this information, together with the merchant ID, is transmitted to the online authorisation system. These services are frequently processed by tandem computers throughout the clock.

Tandem is a computer that is always on and always functioning, and it processes all electronic transactions, independent of the institution or country from whence the transactions originated. Tandems are linked to the VISA network and are utilized by all of the world's financial institutions.

The "negative file" and a predefined minimum are used in the Tandem's approval process for business transactions. You may see all of the credit card numbers that have been reported as having a possibility of being used in fraudulent activity in the global negative file. This list is maintained by Visa globally, and it is updated on a daily basis as new incidents of fraud come to light. Before providing authorization, the Tandem checks to make sure that the card number is not on any prohibited or restricted lists. Tandem does not have access to the cardholder's account information; therefore, it is unable to check the cardholder's credit limit. However, it does verify and adhere to a predefined limit on incoming transactions. Tandem

will instantly or, if necessary, immediately provide a batch tile containing the approved transactions that were executed when the mainframe at the cardholder FI comes back up.
Transactions initiated by either cardholders or merchants affiliated with the FI are sorted by the Tandem before being forwarded to the mainframe computer of the FI for authorization. The mainframe of the system is responsible for processing any authorization requests that come in. The mainframe will put the buyer through a series of tests in order to determine whether or not they are eligible to make transactions [21].
When an authorization is authorized, the mainframe changes the authorization log database as well as the posted TX file with the updated information about the account's available credit and the amount or number of authorizations that have been authorized. At the end of the cycle date, the posted transactions from each account's posted TX file will be moved to the statement file processor. These transactions will be taken from the posted TX file. The monthly statements that cardholders get may be printed out with this device. When a consumer pays off their whole amount or makes at least the necessary minimum payment, the transaction is shown as a credit on the customer's account. Only the three most recent statements are kept in memory by the statement tile. When a new statement is made, the oldest statement is immediately archived so that there is room for the new statement.

2.4 CREDIT CARD FRAUD

Both consumers and companies have shown a marked growth in their use of payment methods that are based on the use of plastic cards in recent years. Due to the industry's fast development, there is now a significant risk of being misled.
Losses incurred as a result of fraudulent use of Visa and MasterCard cards were projected to reach $1.63 billion in 1995, representing a significant increase from 1980's figure of $110 million. The United States of America was responsible for the largest portion of these losses, which amounted to more than $875 million in 1995 alone. Because of this, the fact that 71% of all revolving credit cards now in circulation throughout the world are issued in that country should not come as a surprise to anybody. Over 124 million people out of a total population of 193 million in the United States had a credit card in 1994 [22]. There are no specific data available for the credit card industry as a whole; nevertheless, estimations based on the total amount charged to credit cards in 1995, which was $879 billion, indicate a fraud percentage of between 0.1 and 0.2 percent. The American Bankers Association commissioned research

in 1996 that estimated the entire gross fraud losses for bank cards (MasterCard and Visa) in 1995 to be $812 million. This was based on sales totalling $451 billion, which equates to a loss rate of 0.18 percent [23]. When looking at the years 1990 and 1997, the total amount of money lost due to fraudulent credit card use went from $440 million to an estimated $2 billion.

When a credit card is used to make a purchase without the authorization of the cardholder, this is an example of credit card fraud. Visa and MasterCard account for over 65% of all outstanding revolving credit in the globe, and the overwhelming majority of fraudulent transactions utilize either one of these cards or both [22].

The most typical forms of fraudulent activity with credit cards include the following [19]:

a. Lost / Stolen.
b. "Never received issued" is the abbreviation for "never received," which indicates that the item was never delivered (Mail theft).
c. The counterfeit term is the third erroneous definition.
d. telemarketing and direct mail marketing.
e. applications that do not follow proper procedures.

The majority of occurrences of credit card fraud involve cards that have been misplaced, stolen, or both. Cards that are either misplaced or stolen are responsible for 55% of all fraudulent activity on Visa and 49% of all fraudulent activity on MasterCard. On average, victims of this kind of con lose as much as $700 when they fall victim to it [24]. Loss or theft of a credit card can be the first step in the process of credit card fraud. The typical hotspots for credit card theft include places such as the office, the glove compartment of a vehicle, and sporting events. Quite frequently, these losses are the result of a friend or member of the cardholder's family making transactions with the card without the approval of the cardholder [19]. Cardholders may, in some circumstances, sell their card to criminals, then later deny any transactions made with the card and claim it as lost or stolen after selling it to criminals. Over 439,000 new, renewed, and replacement post cards are mailed out by the United States Postal Service every single day. Cards that are never received are those that are stolen from the mail while they are in transit from the card issuer to the intended receiver, either by an employee of the card issuer or by a third party who is not authorized to take possession of the cards. It is feasible to use the card, and then later sell it on the dark web [24]. The average loss from this kind of fraud is rather significant. This is due to the fact that the cardholder

does not often become aware of the theft until it is too late. The majority of people who have their credit card stolen through the mail don't realize it has been taken until they receive their monthly bill. 1997 was the year that witnessed a 68 percent rise in the number of Visa cards that were either lost or stolen. The average loss caused by a card that was lost, stolen, or never received is $750, whereas the average loss caused by a card that was never received is $1,500 [24]. A unique approach to combating this form of fraud is to provide the cardholder with a plastic card that is blank on the inside (electronically blocked). After receiving their cards, customers are required to contact the financial institution that issued them in order to activate their cards.

The area of bank card fraud that is expanding at the quickest rate is the forgery of Visa and MasterCard cards. Counterfeiters now have the ability to produce playing cards that are, in all outward appearances, identical to the originals, thanks to the utilization of cutting-edge technologies. According to some estimates, the typical loss incurred as a result of this type of fraud is around $4,500 [24], making it the most expensive form of fraud in terms of the amount of money that is claimed to have been lost. When a dishonest merchant or telemarketing company phones a cardholder to offer a product that does not exist over the phone, they may later charge the card without the cardholder's authorization after obtaining the cardholder's personal information. It is important to stress that increased customer knowledge has contributed to a decline in the occurrence of this type of fraud. The perpetrators of this type of identity theft give false personal information to financial institutions in order to get credit cards in the victims' names. This type of theft falls under the category of "identification fraud." Since these cards are not signed, it takes far more time to uncover any fraudulent activity on them than it would with stolen cards. As more individuals get educated about the Fls, a smaller percentage of people will be duped by this swindle.

The frequency and severity of card counterfeiting have both been on the rise recently. A stolen account number serves as the basis for this form of fraudulent activity, which can occur in a variety of settings (such hotels, restaurants, stores, abandoned checks, or computer systems).

Financial institutions are the ones that generate credit card numbers. A method known as "skipping" may be utilized to choose a subset of these numbers (let's say 500), which can then be used in the process of creating credit cards. A tool that creates accounts, such as

Credit Master or Credit Wizard, might assist fraudsters in locating the skipping code and exposing the genuine credit card account numbers that they require in order to commit fraud against the FLS. Sniffers are another type of malware that fraudsters deploy, and they may be utilized to steal credit card information while it is in transit. This piece of software will search public networks for 16-digit numbers, jot them down, and then send them on to the con artist [21].

Skimming, also known as bit copying, is a cutting-edge method that may be used to manufacture counterfeit credit cards. During this process, the information contained on the magnetic stripe of one card is copied onto the magnetic stripe of another card. This type of fraud, which is one of the most common methods by which credit cards are counterfeited and is on the increase in Canada, Particular establishments, such restaurants and gas stations, have been singled out as key nodes for the operation of this kind of con [25].

After that, the acquired number will be encoded or stamped on a unique piece of plastic that has been set aside for that purpose. White plastic fraud takes place when the number of a genuine card is placed on an otherwise blank plastic card. When a credit card's number is embossed and/or encoded onto a card that has already expired or been stolen, the card is seen as having been "altered" or "falsified," and its original data is deemed to have been removed. It is possible to copy or recreate the magnetic stripes seen on cards by making use of instruments that are commonly accessible at computer and electronics businesses. A credit card that has had its legitimate number placed to a card that is otherwise fraudulent is referred to as "pure counterfeit". Figure 2.2 depicts the process that must be followed in order to create a phony ID card.

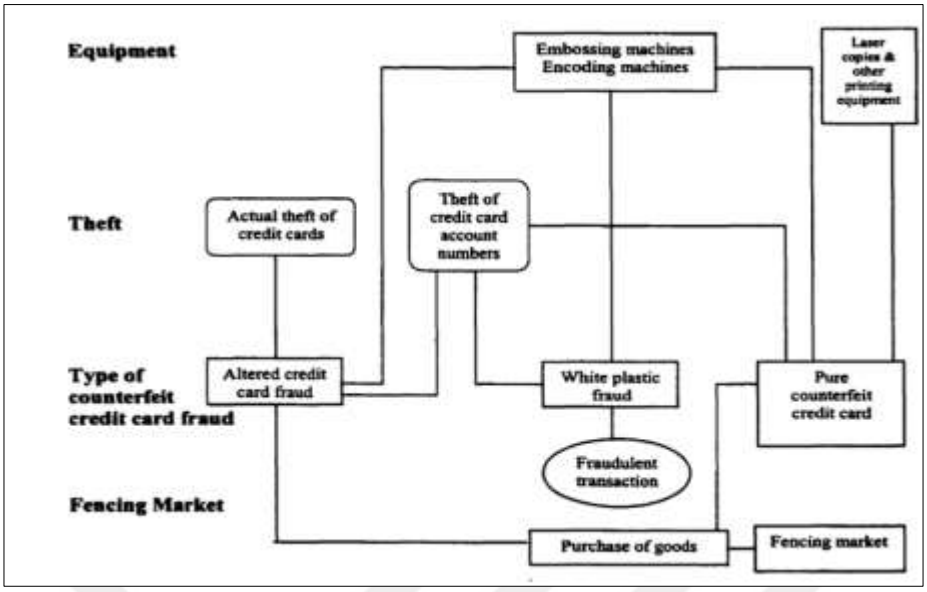

Figure 2.2: The Use of Counterfeit Cards in Fraudulent Transactions.

2.5 MACHINE LEARNING

In a broad sense, machine learning is a subfield of artificial intelligence that seeks to make accurate predictions of the future by teaching computers to learn from their past mistakes without any direct involvement from humans. This is accomplished through the process of teaching machines to learn from their own mistakes. Machine learning marks a significant divergence from the conventional approaches to computing, in which computers are purposefully programmed to calculate or solve a problem. In contrast, typical computing methods include programming computers to calculate or solve a problem. In machine learning, input data are used to train a model, which then uses its newly acquired knowledge of the data to create forecasts about the future. This process is called "training." The applications of machine learning are quite varied and extensive. It is utilized in a wide variety of applications, including spam filters, stock market projections, medical diagnostics, fraud detection systems, self-driving vehicle systems, home value estimation systems, and even facial recognition systems.

The three primary subfields that make up the discipline of machine learning are supervised learning, unsupervised learning, and reinforcement learning. The component of this thesis pertaining to supervised learning will be the topic of discussion in the following paragraphs. For the time being, supervised learning is described as a technique where the model is trained using both input and output labels. This is the definition of what is known as "supervised

learning." Unsupervised learning, on the other hand, is characterized by the absence of input labels in the dataset (i.e., a model is trained using unlabelled data), and the learner is instead presented with a number of different patterns and structures. Image recognition, robotics, language comprehension, and a host of other applications are common settings in which it is utilized. The purpose of reinforcement learning, which is to achieve increasingly complex objectives, is to step-by-step learn how to optimize along a particular dimension (e.g., maximizing the points won in every round [26]).

2.5.1 Supervised Learning

Supervised learning is a method of machine learning in which input and output labels are provided to the model as it is being trained. You may describe it as follows: The supervised model might potentially learn to detect patterns in unlabelled data if it is provided with labelled input and output data. The extracted patterns are used as support for judgments that will be made in the future. The following is an official definition of what is known as supervised learning: The mapping function, f(x), determines the value of the output variable, Y, and x stands for the variables that are used to determine the output. When given an unknown input, the goal is an estimate of the mapping function that is close enough to the true function such that Y can be successfully anticipated. In addition, two further forms of supervised learning are the classification and the regression methods. In most cases, the conclusion of a classification issue takes the form of a list of labels corresponding to a certain category (e.g., fraud or genuine, rainy or sunny, etc.). When an issue involving regression is solved, the solution that is obtained is a numerical number (e.g., the price of a house, temperature, etc.). The problem of categorization is the sole topic that will be discussed throughout this thesis.

2.5.2 Classification

When talking about machine learning, the process of assigning a predicted label to a data piece is referred to as the "classification issue." This might be construed as a difficult task. One probable explanation for the difficulties in recognizing fraudulent activity is that it is difficult to classify the behaviour. The major focus of this conversation is on determining whether or not a certain transaction constitutes fraudulent activity or legitimate business. The process of classification can be broken down into the following three broad categories: binary classification, in which there are only two possible target labels (such as "fraud" or "genuine"

for a given transaction); multi-class classification, in which there are multiple target labels (such as "Rose," "Lily," and "Sun" for a given set of flower images); and multi-label classification, in which the data samples are not exclusive and each data sample is assigned a set of target lab labels This thesis focuses mostly on binary classification, in which the final classification might either be "normal" or "fraud."

2.5.3 Class Imbalance Problem

In the vast majority of real-world scenarios, there is a large imbalance between the numbers of labels that belong to the various classes. The work of detecting fraud serves as a great example of the class imbalance problem. In this job, the number of fraud class labels is significantly low in proportion to the ordinary class label. In general, the performance of machine learning algorithms is not very good when the distribution of classes is unequal (i.e., the predictive model tends to classify the minority example as the majority example). As a consequence of this, a number of problems may emerge, such as how to approach the problem of class disparities. Which machine learning algorithms should be utilized when there is an imbalance between the classes? (iii) When there is a significant imbalance in the dataset, how is it possible to accurately evaluate the performance of a prediction model? The next sections will devote their attention to discussing the responses to these questions.

2.6 HANDLING CLASS IMBALANCE MACHINE LEARNING PROBLEM

In what comes next, a few of the possible solutions to the problem of social divisions based on economic status will be disposed. The methodologies that have been presented to us thus far can be categorized into one of three major categories: resampling, ensemble, or cost-sensitive learning. The only approaches that are going to be covered in this thesis are going to be resampling and ensemble methods, both of which are going to be covered in greater detail further down. In a nutshell, cost-sensitive learning takes into account the potential monetary losses that might result from incorrect categorization. For example, the cost of misclassification associated with missing a malignancy is far larger than the cost associated with projecting that an individual who appears to be healthy actually has cancer. One strategy to improve the model's true positive rate is to take use of the fact that the cost of incorrectly classifying members of the minority group is often higher than that of the majority group.

2.6.1 Resampling Approach

In general, the performance of predictive models is low if there is an imbalance between the classes that are being researched. Therefore, the data that are going to be utilized as input into the model need to go through some kind of pre-processing. In the case that there is an issue with class imbalance, such data preparation, which is also known as a resampling approach, is carried out on a data level. There are primarily three different approaches of resampling data: under sampling, oversampling, and a mix of the two. In order to attain statistical parity in the dataset, under sampling entails reducing the size of the majority class, as can be seen in figure 2.2. It is suggested to use this method when the dataset is very large since reducing the sample size can greatly speed up the processing and ease storage difficulties.

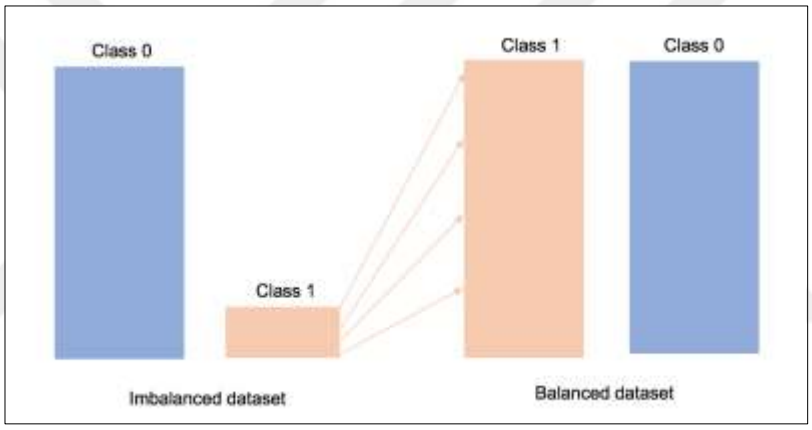

Figure 2.3: Method Based on Under Sampling.

In contrast to the under-sampling approach, the oversampling approach involves collecting an excessive number of samples. Successful with the group that is underrepresented. As can be seen in figure 2.4, it replicates the observations from the minority class in order to bring the percentage of the majority and minority samples into closer alignment with one another. Last but not least, a hybrid method accomplishes the same rebalancing goal by employing both an under- and an over-sampling strategy.

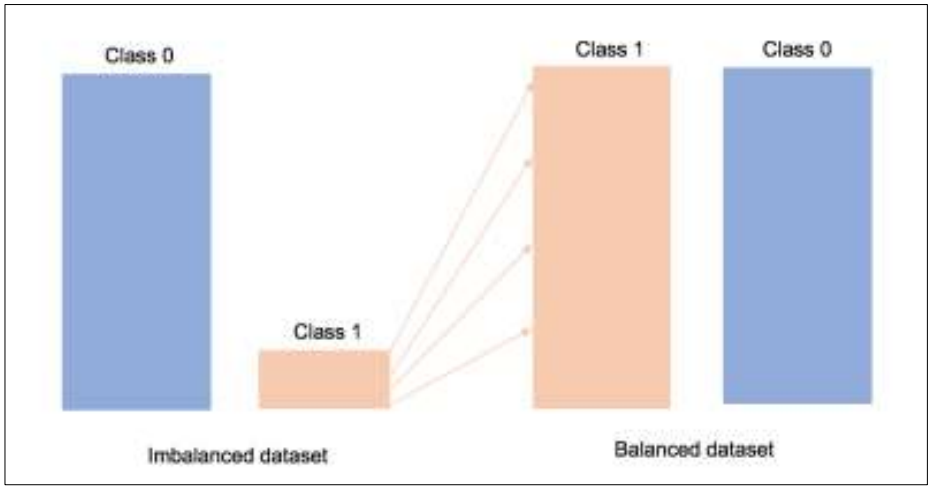

Figure 2.4: Method Based on Oversampling.

2.6.1.1 Random oversampling

In order to ensure that the dataset contains a diverse range of individuals, this approach includes randomly replicating minority sample groups. In contrast to the random under sampling method, this methodology does not result in any of the data that was collected being lost. In spite of this, there is a high probability that the data will be overfit because it is a replication of the minority samples. This feature contributes to the likelihood of this happening.

2.6.1.2 Random under sampling

This method will eliminate the more typical samples at random in order to produce a dataset that is more representative of the whole. This strategy shines when working with a vast quantity of training data since it efficiently handles the data. Because fewer majority samples are collected less frequently, runtime is improved, and storage concerns are alleviated; both of these benefits are due to the reduction in sample size. The issue with utilizing this tactic, on the other hand, is that there is a possibility that it may cause crucial data to be discarded in the course of eliminating the most well-liked candidates. This indicates that the prognosis for the classier might be incorrect.

2.6.1.3 Tomek link removals

The Tomek Link provides an illustration of the several classes that are depicted as being nearest neighbours to one another. A pair (E1, E2) is considered to be a Tomek Link if, and only if, none of the distances between E1 and E3 [27] nor the distance between E2 and E3 is

shorter than the distance that separates E1 and E2. The removal of Tomek connections has the potential to be seen as a tactic for under sampling because this removes the link that contains the overwhelming majority of the sample.

2.6.1.4 Synthetic minority over-sampling technique

A well-known method for reorganizing the data is called the SMOTE algorithm, and it was invented by Chawla [28]. Instead of oversampling with replacement, as is seen in Figure 2.5, the technique of interpolation is utilized in order to generate fresh minority class samples. The minority samples that are the most similar to one another are used as a basis for the interpolation process (synthetic instances). As a consequence of this, the issue of the training data being overfit is reduced in severity. When doing an oversample, it is standard practice to select at random the nearest neighbours of any minority cases that are being examined.

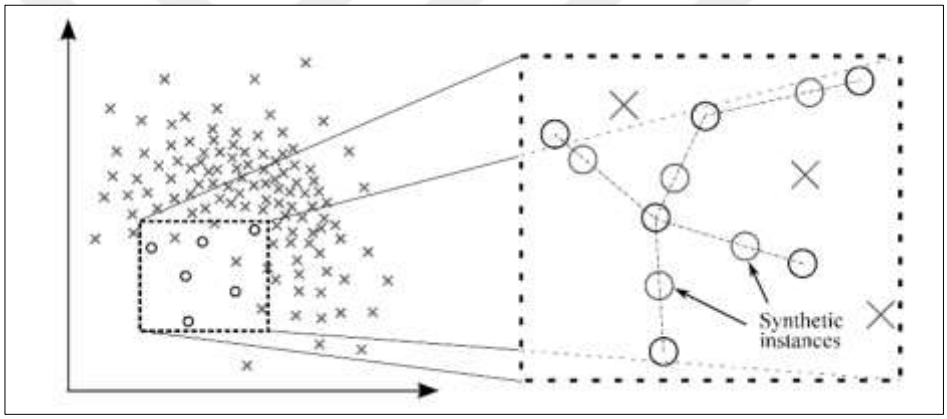

Figure 2.5: Smote [28].

2.6.1.5 Elimination of smote in addition to the tomek link simultaneously

The SMOTE approach is an effective strategy that may be used to equalize participation across classrooms. There is a possibility that new synthetic minorities may be produced, but at the same time, the class cluster of the minority may grow into the space that is traditionally allotted for the majority. Overfitting is possible if the model is provided with an excessive amount of information like this. As a result, the SMOTE approach and the Tomek Link elimination method both have the potential to be utilized in order to assist in equalizing the classes that are involved. The initial training dataset is first subjected to an oversampling procedure called SMOTE. Subsequently, the generated dataset is cleaned up by removing Tomek links, which ultimately leads to a balanced dataset.

2.6.2 Ensemble Approach

Previously, the data-level technique of applying resampling methods is discussed to establish more equal distributions across classes. This strategy can be found in the previous section. The ensemble strategy, which focuses on changing existent classification algorithms, is one approach that may be used to alleviate the unequal class distribution. An ensemble technique is a form of learning algorithm that, as can be seen in Figure 2.6 and is explained in [29], combines the outputs of many classifiers into a single classification decision. This is done in order to improve accuracy. The two sorts of ensemble approaches that are utilized most frequently are known as bagging and boosting.

Before moving on to the actual process of bagging, first go over the concept of bootstrapping, which is utilized by bagging algorithms. In the process of training a machine learning model, bootstrapping is utilized to get samples from the training data that are representative of the whole. The bootstrap sampling method assures that every sample will have its own set of one-of-a-kind characteristics, as can be shown in Figure 2.7. With the use of these data, the models might be taught to have a deeper comprehension of the data and provide more precise predictions.

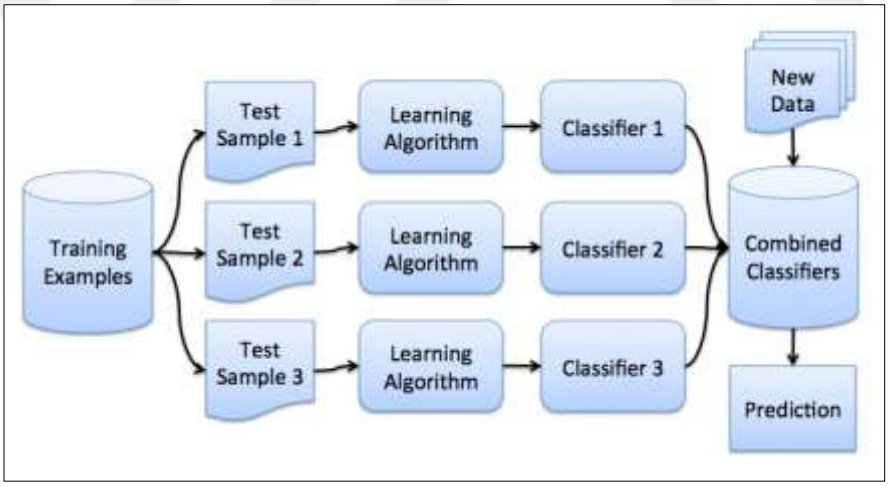

Figure 2.6: Ensemble Approach.

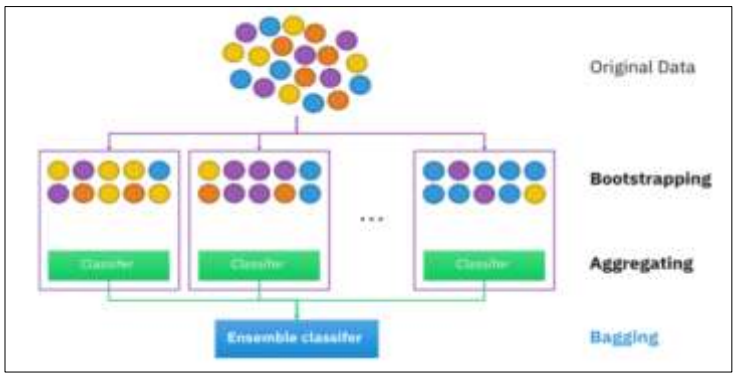

Figure 2.7: Bootstrapping.

2.6.2.1 Boosting

The technique known as "boosting" is yet another effective component of an ensemble. When coupled, weak learners, often referred to as basic learners, can be used to create a stronger learner that is capable of producing outcomes that are superior to those generated by a single student. In contrast to bagging, in which each model is run in parallel, boosting trains the weak learners in a sequential fashion, with each learner attempting to rectify its predecessor by adding additional weights to the samples that were previously misclassified. This is in contrast to bagging, in which the outputs of each model are merged at the end of the process. As a direct result of this, the incorrectly identified circumstances will become an increasingly crucial factor for the prospectively bad learner. Figure 2.8 illustrates boosting in a manner that is more understandable. A further benefit of using bootstrapping is that it eliminates the risk of over-fitting and variance. The phrase "boosting algorithm" refers to a category of algorithms that can take on a number of different forms.

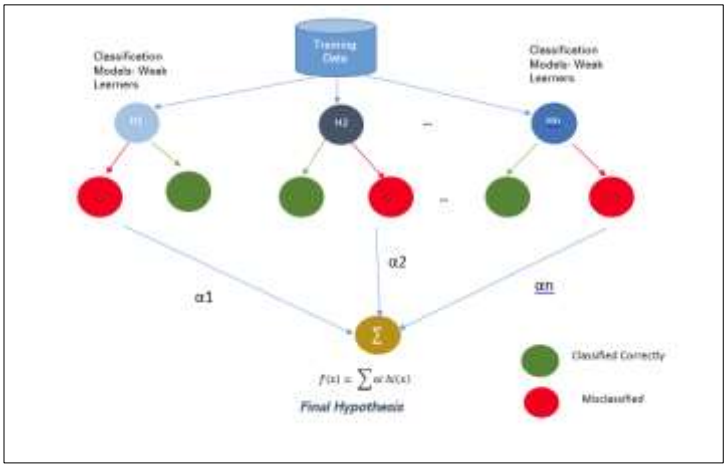

Figure 2.8: Overview of Boosting.

2.6.2.2 Bagging

The name "Bagging," which stands for "Bootstrap Aggregation," refers to a basic and efficient approach for assembling an ensemble. Bootstrapping is a method that may be utilized to generate new training samples by arbitrarily picking and then replacing nodes in the first training set. These freshly created sets of training data have been given the term bootstrap training samples since they were given a name. Every model that is utilized for prediction undergoes independent training with the use of bootstrap samples. In the end, in order to produce an aggregate forecast, the results of each bootstrapped model are either averaged (in the case of regression) or voted on (in the case of classification). Figure 2.9 illustrates bagging in a more clear and concise manner. As a direct consequence of this, over betting is reduced, and there is consequently less fluctuation. The process of bagging frequently starts with the construction of a decision tree as its basic model. Nevertheless, in addition to these, there are other possible avenues of attack that might be used.

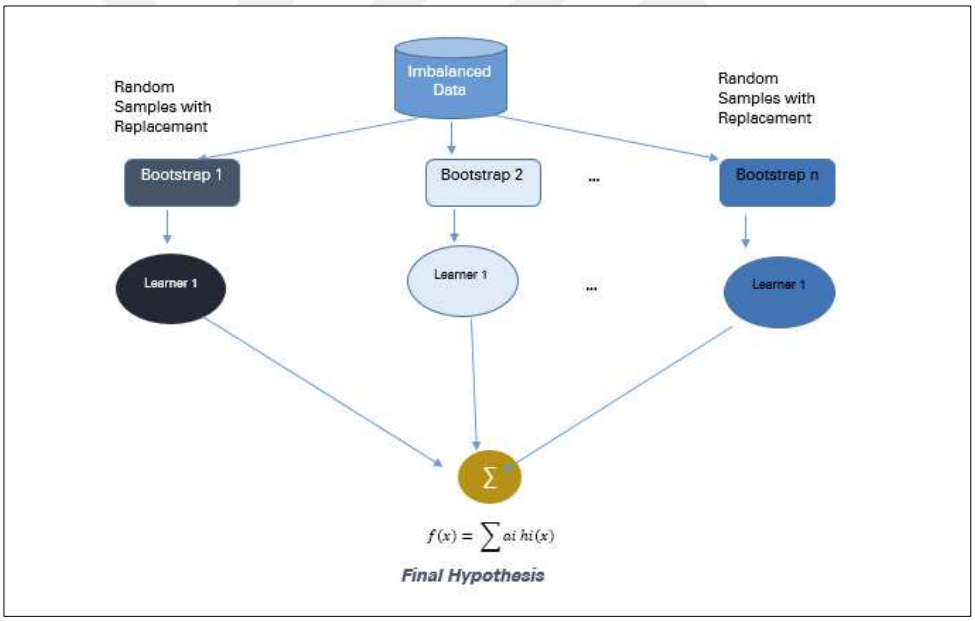

Figure 2.9: Overview of Bagging.

2.7 EVALUATION METRICS

In the field of machine learning, it is standard practice to first train a model by instructing it to learn from the data that is available for training, and then to evaluate the model's capacity to generalize. To put it more simply, the performance of the model is examined using test data that it has never been exposed to before. In such a scenario, how can determine whether

or not the model is effective? Evaluation metrics are utilized, regardless of whether attempting regression or classification, in order to determine how effective, the model is.

2.7.1 Confusion Matrix

Because it is easy to understand and can be used to the calculation of other important metrics such as accuracy, recall, precision, and so on, it is the statistic that is employed the most frequently when analysing predictive analytic models. When dealing with a problem involving classification, the performance of a model is typically expressed by an N-by-N matrix, where N is the number of class labels. A confusion matrix of dimension 22 is shown in figure 2.10, and it is utilized in the process of binary classification.

In order to build a confusion matrix, statistics such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are computed from the actual and predicted values. A True Positive (TP) is a result that happens when both the observed values and the predicted values are positive. This is what the term suggests (such as in the case of fraud). The term "false positive" (FP) refers to the situation in which the observed value is positive despite the anticipated value being negative (normal, for example). A True Negative, also known as a TN, happens when both the observed and predicted values end up having a negative value (often in the normal distribution). The term "false negative" (FN) refers to the situation in which the actual value is positive (such as in the case of fraud), but the expected value is negative.

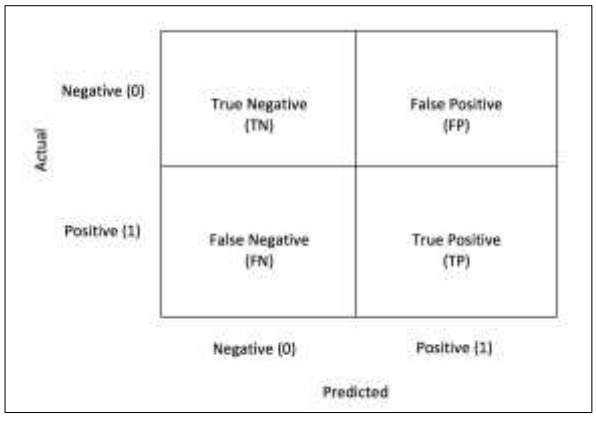

Figure 2.10: Confusion Matrix.

2.7.2 Precision

It is represented by equation 2.1, and it demonstrates the percentage of accurate predictions that were made in contrast to the total number of forecasts that were made (both correct and incorrect). When talk about how accurate something is, what mean by "accuracy" is how many of the examples that were found truly existed.

$$Precision = \frac{TP}{TP+FP} \tag{2.1}$$

2.7.3 Accuracy

TP stands for the number of attacks that have been properly categorized, TN for the number of rows that do not include attacks, FP for the number of assaults that have been wrongly classified, and FN for the number of records that do not contain attacks that have been mistakenly classified. The following definition of accuracy is derived from these:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{2.2}$$

2.7.4 Recall

Recall, which is also known as sensitivity, is represented in Equation 2.3 as the proportion of right diagnoses compared to the total number of occasions when a positive result was obtained. In actuality, what this implies is that recall represents the proportion of positive situations for which information was really recalled.

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

2.7.5 F1 Score

The mathematical arithmetic mean of the recall and accuracy scores is referred to as the F1 Score. This score is also known as the F score or the F-measure. It is conceivable for its value to be anything between 0 and 1, with 0 being the worst consequence possible and 1 representing the best. The formula may be written down as follows.

$$F1 = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{2.4}$$

2.8 TECHNIQUES BASED ON MACHINE LEARNING MODELS

2.8.1 Decision Tree Classifier

In the field of machine learning, decision trees are a sort of algorithm that belongs to the supervised learning category (meaning that they were given labelled data to train on) [29][30]. Whenever necessary, the training data is partitioned into two extra sub-nodes utilizing a single constant parameter. It is vital for comprehension that you pay attention to the leaves and branch tips, which are referred to as decision nodes. At the decision nodes, the data is broken down into the individual components that make up the whole. Depending on your point of view, the leaves might either symbolize the decisions that were made or the outcomes of those decisions. A decision tree is a structure used in computer programming that resembles a tree and has a number of "if statements" at each node, which finally leads to a "leaf node" (the final outcome).

2.8.2 K-Nearest-Neighbour Classifier

It would be difficult to simplify the explanation of the k-nearest-neighbour classifier any more than what has been presented here. This phrase has made its way into a great number of other cultures and languages over the course of human history. According to Proverbs 13:20, which states, "He who walks with clever men will be wise, but the friend of idiots will suffer damage," it seems reasonable to assume that associating with intelligent people is beneficial. This is because the verse states that "he who walks with clever men will be wise." The fundamental premise that underpins nearest neighbour classification is the idea that it is necessary to identify the k training samples that are the most comparable to a new sample that has to be categorized. The examples close by a new one will be considered when choosing a name for the new one. When using k-nearest neighbour classifiers, the user is responsible for setting a fixed value that determines the number of neighbours from which to select one neighbour’s status. An additional kind of learning algorithm is known as the neighbourhood-based radius learning approach. The point density in the area is taken into consideration by each approach in order to pick a unique number of neighbours. Any metric measure in general, including but not limited to the following: The choice determined by the Euclidean distance is the one that the vast majority of people go with, and this is by a

significant margin. The neighbours-based approach, in contrast to more generic kinds of machine learning, "remembers" each and every piece of training data. Using the votes of the individuals who live nearby the unidentified sample and applying the principle of majority rule, it is possible to arrive at a categorization [31][33].

2.8.3 Linear Regression Model

As a first-order approach of predictive analysis, linear regression is the method that is most commonly used in practice. The primary objective of regression analysis is to make a comparison between two variables, namely: How accurately can a set of predictor factors forecast the value of a target variable (also known as a dependent variable)? Which particular predictor factors are there, and how do these predictor variables influence the outcome variable (as evidenced by the magnitude and direction of the beta estimates)? These regression estimates are quite helpful whenever one is attempting to comprehend the relationship that exists between a dependent variable and a number of independent variables. In the equation for regression with one variable, y is the estimated score on the dependent variable, c is a constant, b is the regression coefficient, and x is the score on the independent variable. The equation for regression with one variable is represented as y = c + b*x. Providing a Characterization of the Variables There are a several names for the essential part of a regression analysis, but they all refer to the same thing. It is possible to refer to it as a regressor, a criterion variable, an endogenous variable, or an outcome variable. All of these names are applicable. There are other names for the independent variables, including exogenous factors, predictor variables, and regressor variables [34][35]. The three primary uses of regression analysis are as follows: (1) determining the accuracy of predictors, (2) forecasting the outcome of an event and (3) predicting the direction of a trend. The technique of linear regression is one of the most common ones used for forecasting. The primary objective of regression analysis is to make a comparison between two variables, namely: Is it feasible to make an accurate prediction of an outcome (dependent) variable by employing a predetermined set of variables that serve as predictors. To be more specific, which factors have a significant effect on the outcome variable, and how do those factors influence it (as evidenced by the magnitude and direction of the beta estimates)? These regression estimates are quite helpful whenever one is attempting to comprehend the relationship that exists between a dependent variable and a number of independent variables. The most

straightforward form of the regression equation with one dependent and one independent variable looks like this: y = c + b*x, where y is the estimated score on the dependent variable, c is a constant, b is the regression coefficient, and x is the score on the independent variable. This gives us the simplest version of the equation. Providing a Characterization of the Variables There are a several names for the essential part of a regression analysis, but they all refer to the same thing. A regressor, a criterion variable, an endogenous variable, or an outcome variable are all names that may be used to refer to it. There are other names for the independent variables, including exogenous factors, predictor variables, and regressor variables. The three primary uses of regression analysis are as follows: (1) determining the accuracy of predictors, (2) forecasting the outcome of an event and (3) predicting the direction of a trend.

2.8.4 Linear Support Vector Machine

Support Vector Machines are a type of learning approach known as supervised learning. Because of this, they require previously labelled data in order to appropriately categorize new input. The development of a function that can either (a) separate the data into its component labels with the maximum number of false negatives or (b) separate the data into its component labels with the lowest number of false positives is the initial stage in any technique for classifying data.

Having additional white space near to the splitting function helps minimize the number of mistakes made since it makes it simpler to differentiate between the various labels. It is possible for many functions to successfully and error-freely partition a data set, as seen in Figure 2.11. As a direct consequence of this, the region around the separation function is now being used as a metric in and of itself. It is recommended to choose Separation A since it provides a more defined demarcation between the two groups.

In mathematical parlance, support vector machines (SVMs) produce n-dimensional spaces that may contain one or more hyperplanes. The first phase in every kind of data analysis is known as data splitting, and it always begins with a linear separation of the data into its appropriate labels. The provided example makes use of a data set that contains n data points, with each data point consisting of a label y 2 that reads "purchase" or "no purchase" and an attribute vector x that stores the data values for that particular session. This is done in order to forecast the likelihood that a customer will make a purchase from an online retailer. Now,

the Support Vector Machine searches for a function that can separate the data into two distinct categories: those in which y is equivalent to "yes," and those in which y is equivalent to "no." If the data can be completely divided up in a linear fashion, then the function that was produced may be utilized in order to classify subsequent events. According to Steinwart and Christmann [36], this technique has two significant flaws that need to be addressed: To begin, it is important to note that it is feasible for the data to not be linearly separable, or to not be linearly separable at all; this is typically the case with data taken from the actual world. There is the potential for events in online businesses to unfold in a manner that is analogous to the one that was just described, such as when two customers display the same behaviours but only one completes a transaction. The data that resulted would be impossible to separate due to the fact that the same attribute vector would now have unique labels.

The second potential cause for concern is that the SVM has been overfit. The only way to stop something like this from happening is to conduct some preliminary processing on the data in order to get rid of the noise and leave some room for error in the classifications. If this is not avoided, the accuracy values produced by the SVM will be erroneous, which will result in worse mistakes being made when categorizing following occurrences. The kernel technique gives a solution to the first problem by projecting the n-dimensional input data onto a higher-dimensional space, which allows the data to be linearly separated. This allows the kernel approach to solve the problem.

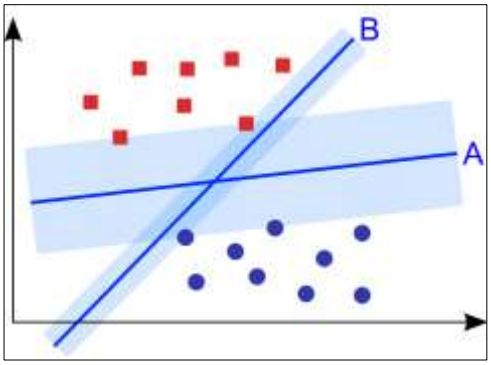

Figure 2.11: Visualization of a Support Vector Machine.

2.8.5 Logistic Regression Model

When it comes to clicking prediction, this is the tried-and-true method. The logistic regression model is a sort of parametric regression that is used for forecasting. This model is also referred to as logit regression and the logit model. Given a group of predictors, it computes the chance of a categorical result being one of those outcomes [37].

There are two primary categories that may be utilized to classify logistics models. When there are just two potential outcomes for the dependent variable, such as yes or no or success or failure, the independent variables can either be continuous or categorical [37]. The binary logistic regression model is used to analyse data in this situation. In order to generalize the result, it is common practice to substitute the two categories of the response variable with the values zero and one respectively. When there are more than two categories that can be assigned to the dependent variable, multinomial or nominal logistic regression models are utilized [37]. When the multilevel dependent variable is of the categorical kind, the ordinal logistic regression model is the one that is utilized (Light, Medium, or Dark). It is considered more successful than other methods since it takes into account the order in which responses are provided. The first stage in the process of building a logistic regression is to determine the odds. Assuming that the likelihood that outcome Y is 1 is p, then the probability that it is 0 is 1 minus p. This means that the likelihood that it is 1 cannot be determined. The odds that this is the case may be found by solving Equation 2.5.

$$odds = \frac{p}{1-p} = \frac{probability\ (y=1)}{probability\ (y=0)} \tag{2.5}$$

In order to accomplish this, it need to take the natural log of the probabilities and then regress them linearly on the independent variables using equation 2.6, where x is a predictor and beta is a coefficient. The associated coefficient in logistic regression may be thought of as the amount by which the log odds move in the opposite direction when one unit is added to a particular predictor variable. This is because logistic regression uses log odds. After doing linear regression, it is helpful to convert the log chances back to probability p by computing the antilog of equation (2.7). This is done in order to make the conversion.

Then, using the probability value p, extend the result to the dependent variable Y. If p is more than some predetermined threshold, anticipating that the value of Y will be 1, and if p is less than that threshold, anticipating that Y will have no value. This is the expectation.

The aim of the logistic regression approach is to forecast whether a result will be a click based on a set of predictors, and the best fitted model for doing so is obtained via a process of backwards elimination. In other words, the method predicts whether a result will be a click.

2.8.6 Random Forest Classifier

The classification and regression tree (CART) model is an example of a decision tree approach. This model may be used to predict either a regression or a classification depending on the input data. It means "forming a local model in each area of input space that arises from recursively dividing the input space" [38]. Each CART unit may be represented as a binary tree with one leaf for every region of the tree. Each leaf node, which corresponds to a subset of values for the independent variable, provides a prediction for the value of the dependent variable. This prediction is based on the values of the independent variable. The CART algorithm has some important differences when contrasted with the logistic regression method. The classifications that are made by the logistic regression algorithm are based on the relationship that exists between the response and the predictors. On the other hand, the decision tree is based on the idea that the dataset should be divided into smaller sections in accordance with certain rules until a small enough set is reached in which all of the data points in it fall under the same label. When using the technique of classification trees, contamination may be utilized to pick the split variables and their thresholds. After that, the expected label (category) of y can be derived based on the proportion of y that is present in each zone of x. Simply split the data in half, calculate the impurity, and choose a threshold value at random for x. After doing this procedure a number of times, select a split point that has the least quantity of impurities. The expression of impureness may be seen in Equation 2.4 by the use of a Gini index function.

$$G = \sum_{k=1}^{K}(p_{m,k} * (1 - p_{k,m})) \quad (2.4)$$

If region m is one of those into which x is divided, and pm,k is the proportion of training instances that fall into category k for that specific area, then the proportion is stated as follows: To put it another way, when G=0 occurs at a node, it means that all of the data points at that node are of the same type. When G is increased, the range of possible Y values in the node becomes increasingly extreme [39].

The method is finished when either all of the cases inside a node have the same value or when the size of the node has fallen to below a threshold. At this point, a value is assigned to the node as the projected solution.

The random forest technique is an improved version of the tree algorithm that produces a very large forest. It generates the forest from scratch. It has been shown that integrating extra

trees in a model result in improved accuracy of prediction. This is due to the fact that each tree acts as a relatively weak classifier with high variance. In the CART technique, as an additional step before the tree growth, repeatedly sample both the observations and the predictors utilizing replacement, and then generate a forest of trees from those samples. This phase comes before the tree development stage. The following is an exhaustive illustration of the procedure [40]. The first step is to select k features at random from a pool of features where k is less than m. The second step is to use the CART method to determine the best split point for the selected features. The third step is to perform the split. The fourth step is to repeat steps one and two until the size of each node reaches the threshold value or the values in each node become identical. If you follow these instructions, each tree will be able to serve as a minimum classifier using a subset of variables and data points that are totally determined by you. The accuracy of the combined classifier increases as a result of the consistent generation of new trees, and the overall variance of the forest finally reduces to a level that is significantly lower than that of a single tree. The random forest approach is more accurate and dependable than any individual decision tree because of this reason.

2.8.7 Gradient Boosting Machine (GBM)

There are some parallels can be seen between the CART algorithm and the boosting method, often known as the random forest algorithm. Both of these methods may be considered to be better versions of the CART algorithm. The boosting method achieves the same goal by reweighing the data, but the random forest method improves the performance of the model by resampling the data.

The practice of boosting has quickly emerged as one of the most powerful and widely used strategies in the field of machine learning. By employing a number of different fundamental classifiers, it is possible to turn a weak learner into a strong learner who has much increased performance [41]. A weak learner is one whose performance is only slightly but objectively better than random chance, whereas a strong learner provides classifications with a much-reduced error probability [41]. A bad learner is one whose performance is only marginally but objectively better than random chance. The objective of the boosting technique is to produce a strong classifier by merging numerous weak learners, such as decision trees, which each perform badly on their own. In the method, each observation is assigned a weight, with greater weights being assigned to more difficult records and lower weights being assigned to

records that are simpler [42]. Before attempting to make a forecast, the approach requires first that you give equal weight to all of the observations. It does this by giving an increased weight to the records for which the earlier learner made an incorrect prediction [42]. It continues to add learners and iterates with a new random sample each time until either the lower limit of the node size is reached, or the desired accuracy is attained [42]. Until one of these two milestones is reached, the process will not end. The ultimate prediction is arrived at by taking the average of the predictions made by all of the weak classifiers and voting with a weighted majority; the weight that is placed on each value is established through an iterative process. The name "Gradient Boosting Machine" refers to one of these generic boosting frameworks (GBM). Training is not done on a newly sampled distribution; rather, it is done on the mistakes that were left over from the strong learner. This results in higher performance from the model [43]. While doing so, it uses the residuals from the most recent classification to train a less capable learner. This process is repeated until the error is less than a predetermined threshold [43], at which point it moves on to the next iteration. The algorithm that is utilized by a gradient boosting machine is characterized by Equation 2.5.

$$f(x) = \hat{f}(x) + beta * h(x) \tag{2.5}$$

where $\hat{f}_x$ represents the prior guess, h_x represents the beginning learner and is the weight for the most recent learner that is added at each iteration of the process. In this discussion, the term "learner" refers to the tree that is generated by the algorithm at the end of each iteration. The step is denoted by the product h_x at each and every iteration of the process. The error term is represented by the notation $L\left(y_i, \hat{f}(x_i) + beta * h(x_i)\right)$, and the literature has a variety of other expressions of it. Here, It is important to use of the L2 regularization, which produces an error that has the least number of squares and is represented by the equation L = $\sum(\left(y_i, \hat{f}(x_i) + beta * h(x_i)\right)^2)$. The ultimate goal is to bring the amount of error that is brought on by the fact that the current fbx is not a good match down to a more manageable level. In order to do this, constantly update the function $\hat{f}_x$ by adding the step ter.

2.8.8 Xgboost

XGBoost may be thought of as a more advanced gradient boosting computer in one approach to think about it. In place of the more frequent practice of using gradient descent, the method proposed by Newton makes use of first and second derivatives in order to maximize the value

of the objective function, also known as the loss function. To begin, recall that equation 2.6 is the equation for the Second Taylor Expansion. This will get you off on the right foot.

$$f(x_0 + \Delta x) = f(x_0) + f'(x_0) * \Delta x + \frac{1}{2} * f''x_0 * \Delta x^2 \quad (2.6)$$

To cultivate trees in a manner that results in the lowest possible value for the objective function $g_i T(x_i) + a_i T(x_i)^2$, where $g_i = L'\left(y_i, \hat{f}(x_i)\right)$, $a_i = L''\left(y_i, \hat{f}(x_i)\right)$, and $T(xi) = \Delta f(xi)$. This made it possible to retrieve the optimal weight of each leaf, which is similar to how the beta in the GBM approach works.

2.9 TECHNIQUES BASED ON NEURAL NETWORK MODELS

2.9.1 Neural Network

One way to conceive of a neural network is as a network of logistic regression units that has many more layers than the traditional network. The model's predictions will be more accurate than those produced by the conventional one-layer parametric logistic regression since it has extra layers that have a more profound and intricate structure. In order to evaluate whether or not this is the case, compare the outcomes of the two models once this model is tested using actual data. A neural network is a type of nonlinear model that may be utilized for activities such as prediction and classification. The distinct but linked building pieces of each of the network's three layers—an input layer, a hidden layer (which may contain multiple sublayers), and an output layer—are called neurons, nodes, or units [44]. In order for a neural network to function, data must be sent from the input layer to the hidden layer. Once there, nodes will apply weights to each input unit, and the process will continue until the weighted inputs are multiplied by the total weights that have been assigned [44]. The output of a neural network is computed by applying a function known as the activation function to a weighted sum of the inputs [44]. This process is described in more detail here. There are many different kinds of unit activation functions, but some of the more common ones include linear, step, logistic (sigmoid), tanh, and rectified linear (ReLU) (Figure 2.12).

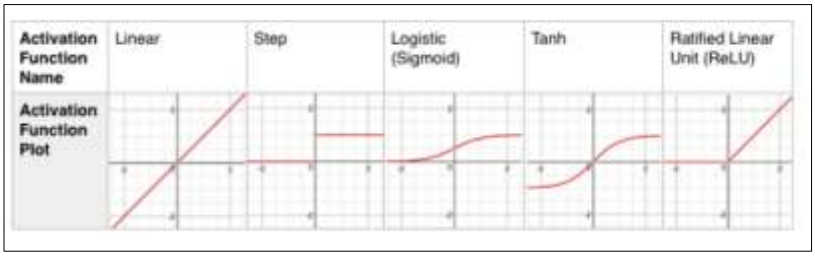

Figure 2.12: Activation function

The following two tables provide a summary of all activation functions for your convenience. The feedforward network, the convolutional network (CNN), the recurrent network (RNN), and the recurrent network (RN) are a few examples of neural networks.

2.9.2 Multi-Layer Perceptron Classifier

A neural network that may link multiple layers of a target graph is referred to as a multilayer perceptron, which is exactly what its name suggests it is. This suggests that there is a signal channel that only sends signals in one direction across the nodes. Every single one of the nodes, with the exception of the input nodes, has a non-linear activation function. Backward propagation is a method of directed learning that is used within an MLP. The Multi-Layer Perceptron (MLP) is a way to learning that is regarded to be a deep learning method since it incorporates several layers of neural connections. A multilayer perceptron, often known as an MLP, is a type of neural network that generates a set of outputs in response to a collection of inputs. Figure 2.13 shows a perceptron broken down into its two distinct layers. The most popular use of MLP is for the solution of supervised problems, but it is also used in the study of computational neuroscience and in parallel distributed processing. Automated translation, image identification, and voice recognition are just some of the helpful applications that may be created with this technology.

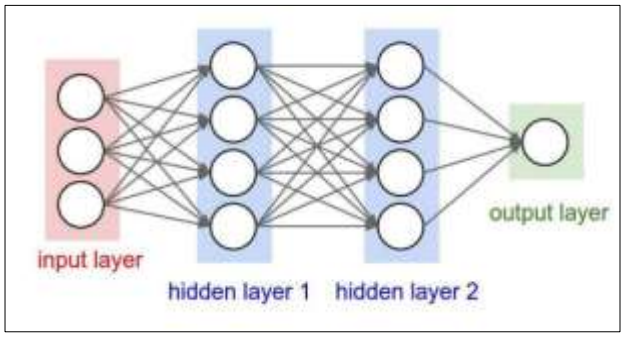

Figure 2.13: Multilayer perceptron with 2 hidden layers

3. METHODOLOGY

3.1 DATASET

Credit Card Fraud Detection dataset [45] is used in this study, which is a set of transactions made with credit cards that are either validated in secret or identified as fraudulent. During the course of a research project that focused on huge data mining and fraud detection, which Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) at ULB (Université libre de Bruxelles) collaborated on, the dataset was collected and processed. In order to prevent their customers from being charged for items that they did not order, businesses need the ability to detect fraudulent use of credit cards. The dataset includes purchases made with credit cards throughout Europe during September 2013. These purchases were made in 2013. This report reveals specifics regarding financial transactions that took place over the period of two days; out of a total of 284,807 transactions, 492.3% are fraudulent. This information was obtained from. The dataset is significantly skewed with just 0.172% of all transactions falling into the positive class (frauds). This percentage is really low. Only numeric variables that have been transformed using PCA are allowed in as inputs. However, due to privacy concerns, the original features and further context around the data are not available. Only the 'Time' and 'Amount' characteristics have not been altered by PCA, which has resulted in the generation of the primary components V1,V2,V28. When you look at the feature labelled "Time," you will notice the amount of milliseconds that have passed since the very first transaction that was included in the dataset. The 'Value' function displays the total amount of the transaction, and it may be used to facilitate cost-sensitive learning that is derived from specific cases. In instances where there was an attempt at deception, the "Class" response variable will have the value 1, while in all other instances, it will have the value 0. They suggest that the Area Under the Precision-Recall Curve (AUC) be used as the performance indicator, taking into account the discrepancy that exists across the classes (AUPRC). There is no point in discussing this topic because the unbalanced classification is not impacted in any way by the correctness of the confusion matrix. A simulator for transaction data can be found at the following URL: https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter 3 Getting Started/SimulatedDataset.html. This simulator is included in the hands-on guide to Machine Learning for Credit Card Fraud Detection, which can be found at the aforementioned URL. Reading the book is highly recommended for any

practitioner who has an interest in fraud detection datasets. The data simulator and the credit card fraud detection methods described in the book are both presented in the book.

3.2 PROPOSED METHODOLOGY

The general workflow methodology proposed in figure 3.1. The chosen dataset with be pre-processed by cleaning before exploratory data analysis and encoding. Encoded data will be clustered by feature engineering. After preparation, this data will be fitting by the chosen models.

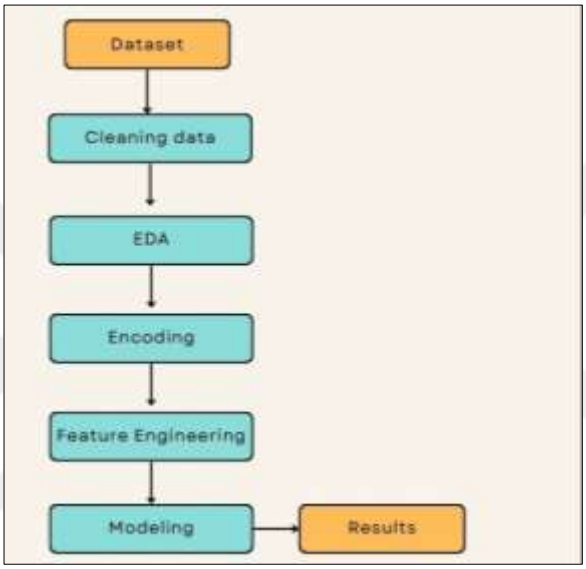

Figure 3.1: General Workflow of the Proposed Methodology.

3.2.1 DATASET

Data is described as in figure 3.2 before any processing.

	Time	V1	V2	V3	V4	V5
count	284807.000000	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05
mean	94813.859575	1.168375e-15	3.416908e-16	-1.379537e-15	2.074095e-15	9.604066e-16
std	47488.145955	1.958696e+00	1.651309e+00	1.516255e+00	1.415869e+00	1.380247e+00
min	0.000000	-5.640751e+01	-7.271573e+01	-4.832559e+01	-5.683171e+00	-1.137433e+02
25%	54201.500000	-9.203734e-01	-5.985499e-01	-8.903648e-01	-8.486401e-01	-6.915971e-01
50%	84692.000000	1.810880e-02	6.548556e-02	1.798463e-01	-1.984653e-02	-5.433583e-02
75%	139320.500000	1.315642e+00	8.037239e-01	1.027196e+00	7.433413e-01	6.119264e-01
max	172792.000000	2.454930e+00	2.205773e+01	9.382558e+00	1.687534e+01	3.480167e+01

Figure 3.2: Database example

3.2.2 Cleaning Data

The process of rectifying or removing any erroneous, corrupt, badly formatted, duplicate, or missing information from a dataset is referred to as "data cleaning," and the word "data cleaning" refers to the processes that are done to do so. When combining data from many sources, there is a significant chance that some of the data may be duplicated, and some of the labels will be incorrect. If the data that was used to create the results and algorithms was wrong, then the results and algorithms should not be trusted, even if they appear to be correct. Due to the fact that different datasets have varying degrees of complexity, there is no one approach that can reliably prescribe the many actions that constitute the data cleaning process. First things first, the process need to get rid of any material that is pointless or superfluous. Get rid of those items from dataset, whether find them to be duplicates or they are just not relevant. It is not uncommon to arrive at the same conclusion more than once while one is collecting information. When combine data from a variety of sources, search the internet for information, or obtain input from a number of different sources, it is probable that may wind up with duplicate information (such as clients or different departments). The process of de-duplication is one of the most important considerations to give attention to. It is important to come across observations that are irrelevant when locate data points that do not apply to the circumstance that is now taking place. Therefore, if the app is required to examine data pertaining to millennial consumers, but the dataset also contains information pertaining to customers from prior generations, it is imperative that the older customers be removed. The final dataset will be more manageable, will have improved functionality, and will make it possible to do analysis in a more time-effective manner. The next thing that has to be done is to fix any underlying structural issues. During the measurement or transfer of a dataset, issues with its structure become evident when unusual naming practices, typos, or incorrect capitalization are discovered. It's possible that groups will be misclassified as a result of these differences. For example, come across the notations "N/A" and "Not Applicable," but you should consider them to be interchangeable terms. Three, remove irrelevant outliers. When you're looking at a set of data, it always come across some outliers that, at first glance, don't appear to belong in that collection. The overall performance of the data set is improved if an outlier is eliminated for a reason that is considered to be legitimate, such as inaccurate data entry. On the other hand, the existence of an outlier can frequently

serve as the conclusive piece of evidence that is required to validate a hypothesis. In the fourth stage, the process will address any informational gaps that have been identified.

When there is insufficient data, many algorithms will not perform as intended. Researchers could use a variety of methods depending on the circumstances when they are faced with knowledge gaps. Both of these choices are not without flaws, but they are still workable alternatives. Always, the process have the option of excluding observations that are missing values as a first resort, but doing so might result in the loss of data, so give this option a lot of thought before deciding to go with it. Manually inserting missing figures is also an option based on other observations. However, this technique also runs the risk of diminishing the quality of the data because it may be depending on speculation rather than on concrete facts. Altering the mechanism in which the data is accessed is the third possible course of action to take in order to properly navigate through null values. In Step 5, check for problems and do quality assurance. In order to do some very fundamental validation, the conclusion of data cleansing operation should provide you the ability to answer the following questions:

In what ways are the results credible? Are the numbers in line with the norms that are typical for the area? Does it lend credence to or raise questions about the validity of existing theory? Are you able to make use of the data to recognize trends that can then utilize to inform your next hypothesis? If this is not the case, might it be due to inaccurate data? Incorrect or "dirty" data can result in false conclusions, which in turn can have a severe impact on the strategic planning and decision-making processes of a company. When data fails to hold up to scrutiny, it might be embarrassing to discuss it in a gathering with other people. The company must first build a culture of producing high-quality data in order to get to where you want it to be. The first stages are quite important, and they consist of defining data quality and documenting the techniques used to accomplish it. Figure 3.3 describe data before cleaning, Figure 3.4 shows information about data after cleaning.

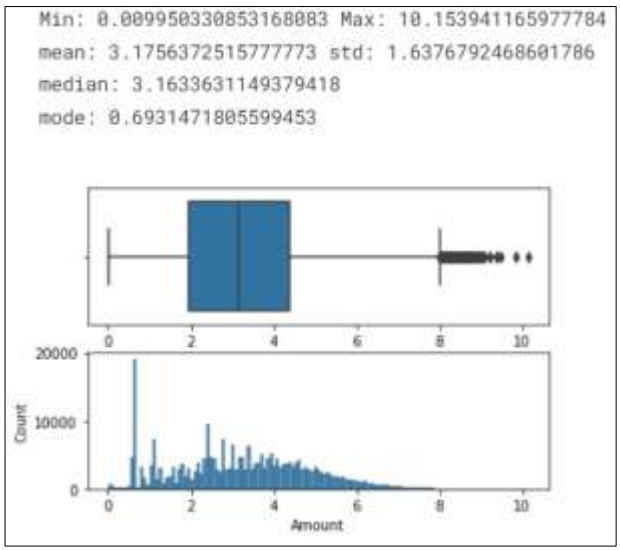

Figure 3.3: Data Before Cleaning.

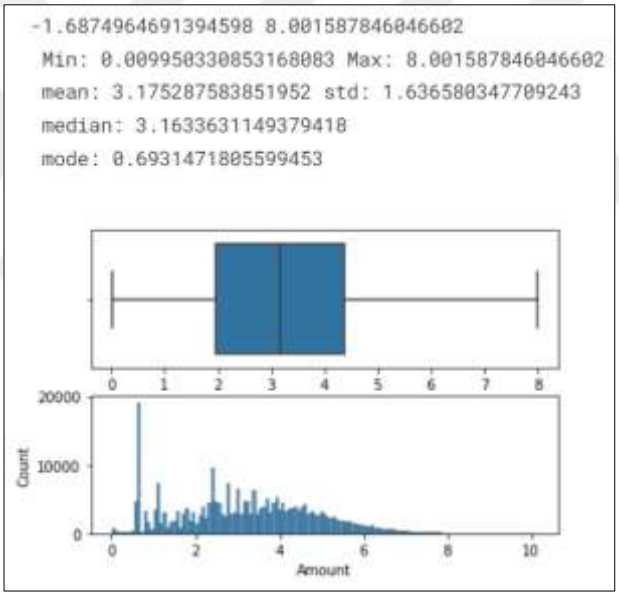

Figure 3.4: Data After Cleaning.

3.2.3 Exploratory Data Analysis (EDA)

A method known as exploratory data analysis may be utilized to study data sets and produce summaries of those analyses. Exploratory data analysis, often known as EDA, is a method that data scientists use to sift through big datasets, recognize patterns, and arrive at judgments on the importance of the data. If data scientists are aware of how to effectively modify data sources in order to acquire the findings they seek, it will be much simpler for them to discover trends, recognize outliers, propose hypotheses, and test assumptions. EDA is frequently

utilized with the purpose of gaining a more in-depth grasp of the variables in a data collection and the interactions between them. This is done with the intention of uncovering information that cannot be gathered by conventional modeling or the testing of hypotheses. In addition to this, it may be utilized to evaluate the reliability of possible statistical approaches to the data being analyzed. John Tukey, a mathematician from the United States, devised EDA methods in the 1970s; these approaches continue to be a prominent tool for investigating data even now. The basic responsibility of EDA is to investigate all of the facts before coming to any judgments. It is possible to utilize it to identify blatant errors, learn more about the structure of the data, discover fascinating relationships between variables, and do a lot more besides. Exploratory analysis is something that data scientists may do in order to ensure the validity of their results and ensure that they can be used. When stakeholders make use of EDA, they have the ability to make certain that they are concentrating on the appropriate queries. EDA may be helpful in a variety of contexts, including those involving categorical variables, standard deviations, and confidence intervals, to name just a few. After the EDA has been completed and the conclusions have been drawn, the features can be used for other forms of advanced data analysis or modeling, such as machine learning. The following is a list of some of the more specialized statistical procedures and methodologies that may be carried out with the help of EDA software: Methods of clustering and dimension reduction could be beneficial to visualizations of high-dimensional data, or of data with a great deal of variety in the aspects to take into consideration. A graphical depiction of the unprocessed data, with summaries of the statistics for each field shown along a single dimension. Bivariate visualizations and summary statistics provide you with the ability to investigate the relationship that exists between each variable in the dataset and the variable that you are now focusing on. Through the use of multivariate visualizations, it is possible to map and comprehend the interactions that occur between different data fields. K-means the clustering method divides the data points into K groups, where K is the total number of clusters, based on how far away they are from the center of the cluster. The data points that are located in close proximity to a certain centroid will be categorized into the same category. Clustering is utilized in the K-means analysis for the purposes of market segmentation, pattern recognition, and picture compression. For the purpose of forecasting future events, predictive models, such as linear regression, make use of statistical analysis and data. After EDA, correlation matrix was extracted shown in figure 3.5.

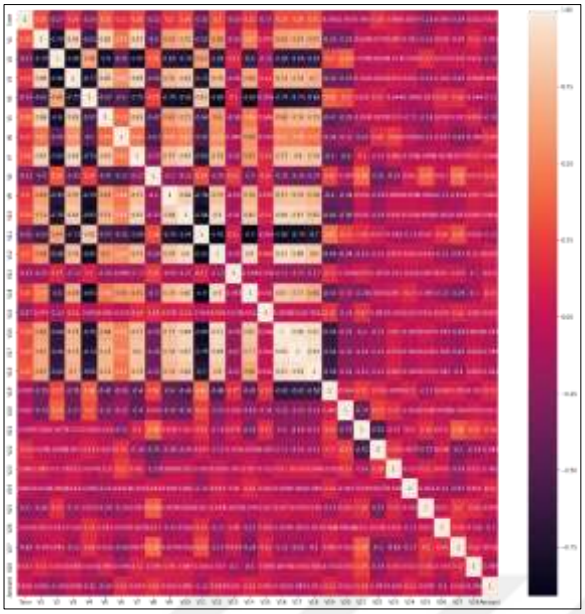

Figure 3.5: Data Correlation Matrix After EDA

3.2.4 Encoding

The preparation of data must always come first in the field of data science before any modeling can occur. The process of preparing the data encompasses a large number of individual processes. Encoding categorical data is one of these highly crucial operations that must be completed. Although the majority of machine learning models are able to process integer values and some are able to process other values that are understandable to the model, the vast bulk of data that comes from the actual world is in the form of categorical text values. Every model, at its heart, performs mathematical operations, which can be performed in a variety of different ways depending on the model. However, the unavoidable truth is that mathematics is entirely predicated on numerical information. Instead of text or other sorts of input, the bulk of the models require numerical values, namely numbers (either float or integer). The process of translating qualitative information into numerical values for the purpose of use in statistical analysis and the development of models is referred to as categorical data encoding.

3.2.5 Feature Engineering

Feature engineering prepares the groundwork for supervised learning by selecting, manipulating, and transforming raw data into features. Supervised learning can then proceed.

To see successful performance from machine learning when it is applied to fresh occupations, perhaps put more work into the design of features and training. One example of a characteristic that might be included in a predictive model is a person's voice tone or the color of an object. Another feature could be the shape of an object. To put it another way, feature engineering is the process of applying statistical or machine learning methods to raw data in order to change it into the qualities that are of interest. Feature engineering is a technique for machine learning that makes use of data that already exists in order to produce new variables that aren't a part of the training set. This method is called feature engineering. It is possible that new features for supervised and unsupervised learning will be generated by it with the goal of enhancing model accuracy, streamlining and speeding up data transformations, and streamlining and speeding up data transformations. When working with models for machine learning, feature engineering is an absolutely necessary step. A problematic feature will have an immediate and direct impact on the model, whatever the data or the architecture. The process of engineering features requires several phases to be completed. The process of establishing new variables that will be of the most use to the model is known as feature creation. It's possible that certain features will be added or taken away. The column that displays the price per square foot was included as an additional feature at no additional cost. Changes: A feature transformation is just a function that modifies the features of a representation in order to make them compatible with another representation. This exercise's goal is to visualize and plot the data in order to decide whether or not more features should be eliminated, whether or not training time can be reduced, and whether or not the accuracy of the model can be improved. The process of discovering useful information by utilizing characteristics that have been extracted from a dataset and then using those extracted characteristics is referred to as feature extraction. It lessens the amount of data that algorithms have to analyze while maintaining the authenticity of the fundamental linkages and specifics. Benchmark: The Benchmark Paradigm is the model that allows for comparisons in the most straightforward way possible, which is also the most trustworthy, open, and easy to grasp. You need to put your brand-new machine learning model through its paces on a selection of benchmark datasets to establish whether or not it is capable of outperforming the industry norm. These benchmarks are used to evaluate the relative merits of a wide variety of machine learning models, such as neural networks and support vector machines, linear and non-linear classifiers, and methods such as bagging and boosting. The

goal is to determine which type of machine learning model is the most effective. In the proposed methodology, K-Means Clustering algorithm is used as feature engineering algorithm. Unsupervised machine learning needs the models to instead learn aspects of the data in order to appropriately reflect the structure of the features. This is necessary because there is no clear objective to be attained in this type of machine learning. Clustering is the process of organizing data into subsets that have similarities. Finding clusters of customers who represent a market segment in order to identify customer churn is one example. Another example would be finding clusters of regions with similar weather patterns in order to split the world into climatically favorable and unfavorable zones. The variance that exists inside each cluster is decreased with each iteration of the clustering procedure, which continues until there are k clusters.

3.2.6 Modeling

As shown in figure 3.6, the proposed system aims to compare between logistic regression, SVM, Naive Bayes classifier, Decision Tree, Random Forest, XGBoost, KNN and Multi-layer perceptron classifier methods.

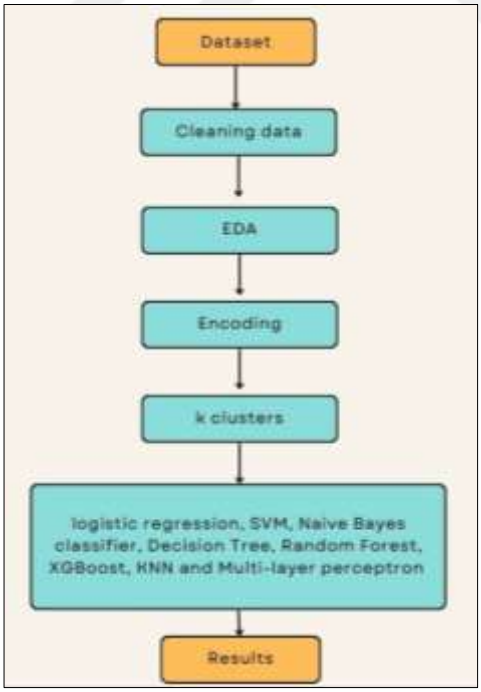

Figure 3.6: Proposed Method Flowchart.

4. RESULTS AND DISCUSION

4.1 ROC CURVE

To evaluate the efficacy of diagnostic tests, one might use the receiver operating characteristic (ROC) curve, which is a plot of the test's sensitivity (y-axis) against its 1-specificity (x-axis) or false positive rate (FPR). Some summary indices have connections to the ROC curve. One common measure is the AUC, or area under the receiver operating characteristic curve. Together, sensitivity and specificity are assessed by AUC. The area under the curve (AUC) is a measure of a diagnostic test's overall performance that may be thought of as the average of its sensitivity and specificity. Given the values on both the x and y axes are between 0 and 1, it may be anything between 0 and 1. Closeness of AUC to 1 indicates optimal diagnostic efficiency; a perfect test would have an AUC of 1. As seen in (Figure 4.1). In order to be effective, a diagnostic test has to have an AUC of 0.50 or above. A line whose length is exactly half an area, stretching from (0, 0) to (1, 1) (Figure 4.1). So, if a diagnostic test has an AUC greater than 0.5, it can accurately distinguish between healthy people and those with a disease (Figure 4.1). Similarly, to sensitivity and specificity, AUC is unrelated to disease prevalence.

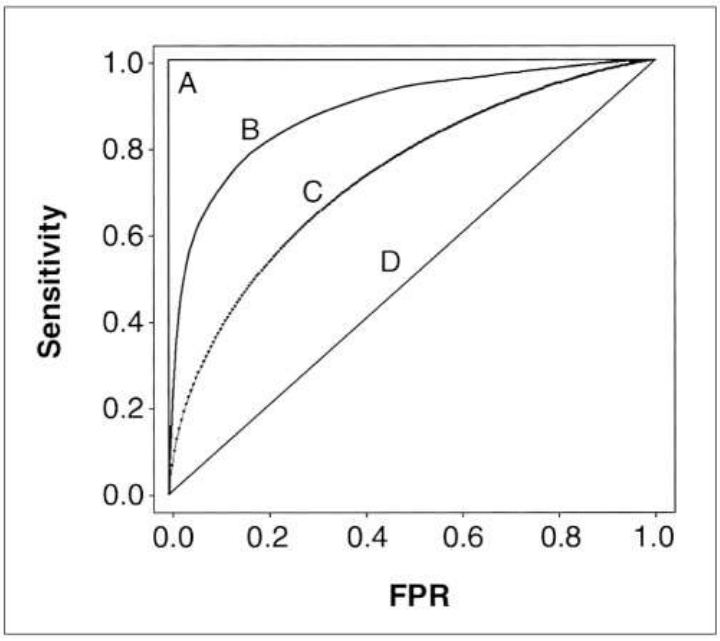

Figure 4.1: Four ROC Curves.

4.2 XGBOOST MODEL

After fitting of the XGBoost model, the model gives an accuracy of 99.947%, a F1-score of 0.851, a recall 0.877 and a precision of 0.827 (see Figure 4.2). The AUC of ROC is equal to 0.97 as shown in Figure 4.3.

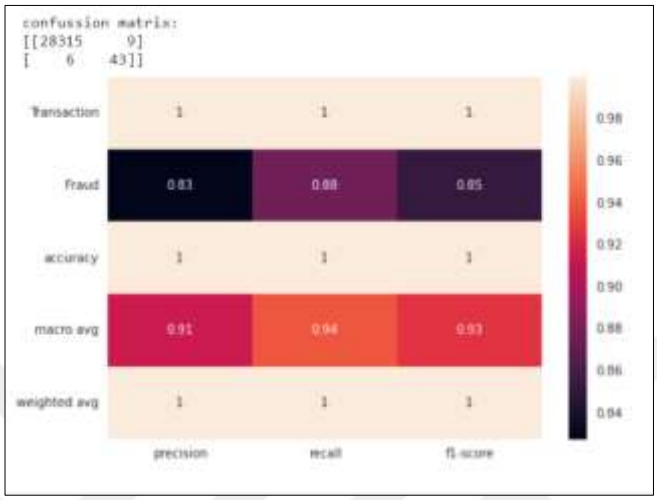

Figure 4.2: XGBoost Results.

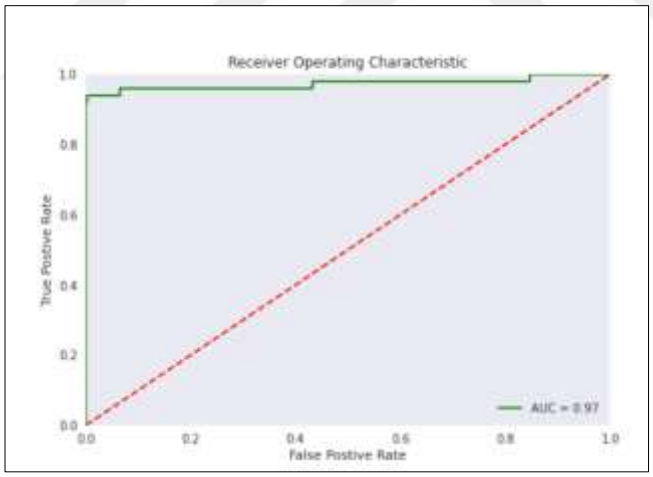

Figure 4.3: ROC of XGBoost.

4.3 KNN MODEL

After fitting of the KNN model, the model gives an accuracy of 99.806%, a F1-score of 0.598, a recall 0.836 and a precision of 0.466 (see Figure 4.4). The AUC of ROC is equal to 0.97 as shown in Figure 4.5.

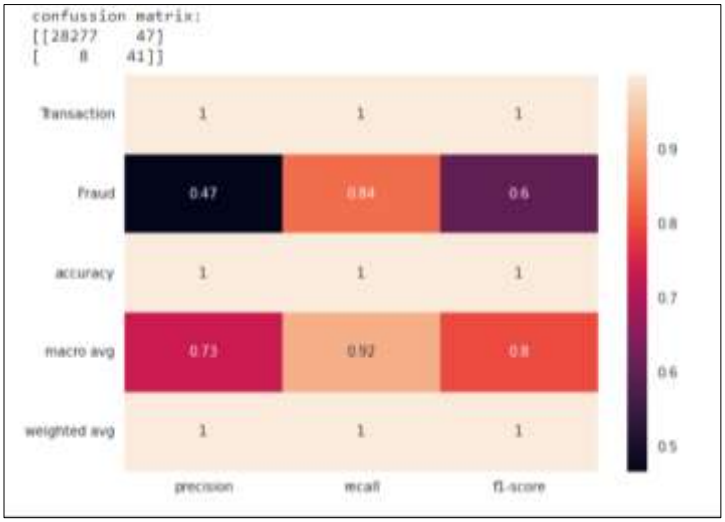

Figure 4.4: KNN Results.

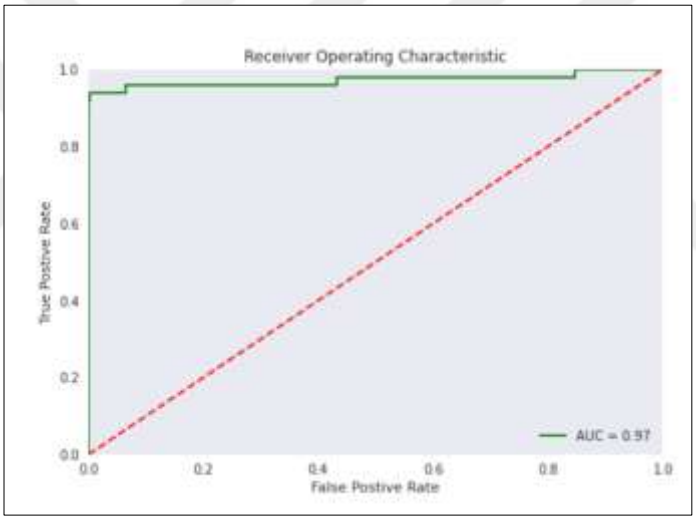

Figure 4.5: KNN Roc.

4.4 RANDOM FOREST MODEL

After fitting of the random forest model, the model gives an accuracy of 99.957%, a F1-score of 0.875, a recall 0.857 and a precision of 0.893 (see Figure 4.6). The AUC of ROC is equal to 0.98 as shown in Figure 4.7.

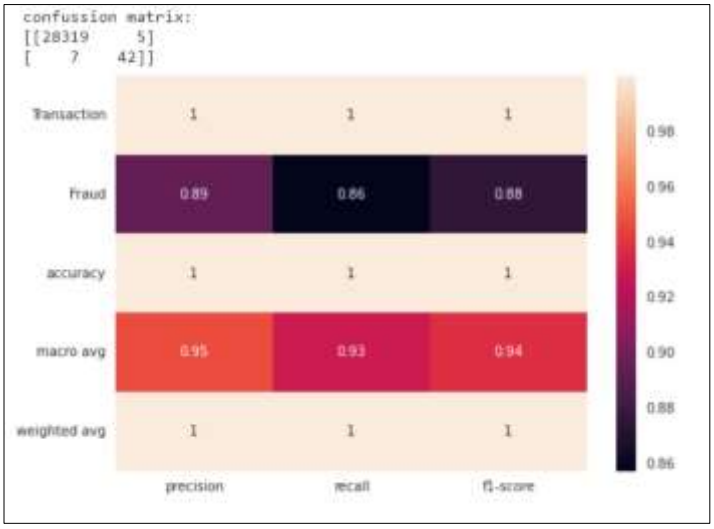

Figure 4.6: Random Forest Results.

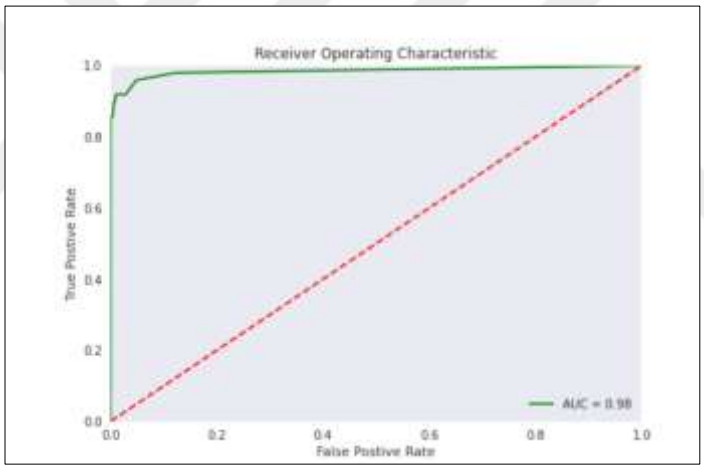

Figure 4.7: Random Forest ROC.

4.5 DECISION TREE MODEL

After fitting of the decision tree model, the model gives an accuracy of 99.788%, a F1-score of 0.565, a recall 0.795 and a precision of 0.438 (see Figure 4.8). The AUC of ROC is equal to 0.9 as shown in Figure 4.9.

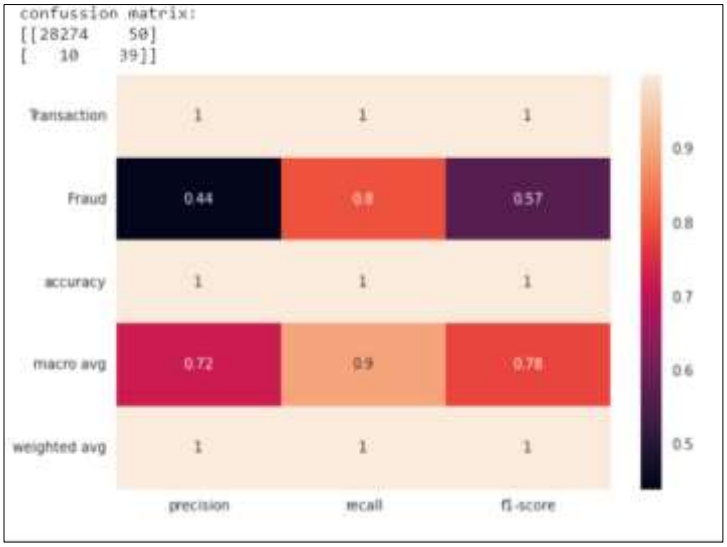

Figure 4.8: Decision Tree Results.

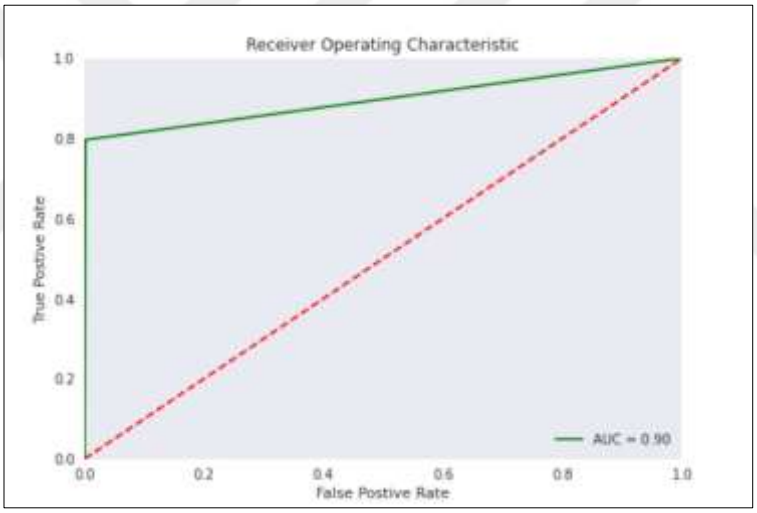

Figure 4.9: Decision Tree ROC.

4.6 NAIVE BAYES CLASSIFIER MODEL

After fitting of the naïve bayes model, the model gives an accuracy of 97.561%, a F1-score of 0.113, a recall 0.898 and a precision of 0.06 (see Figure 4.10). The AUC of ROC is equal to 0.98 as shown in Figure 4.11.

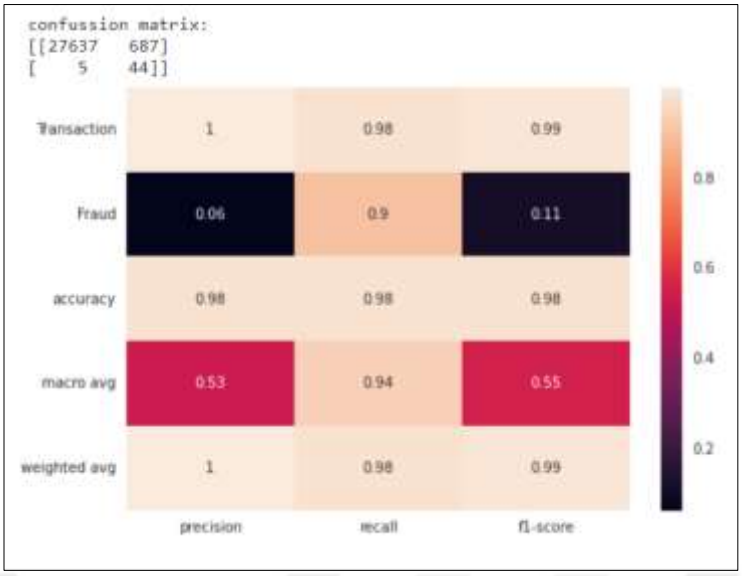

Figure 4.10: Naive Bayes Results.

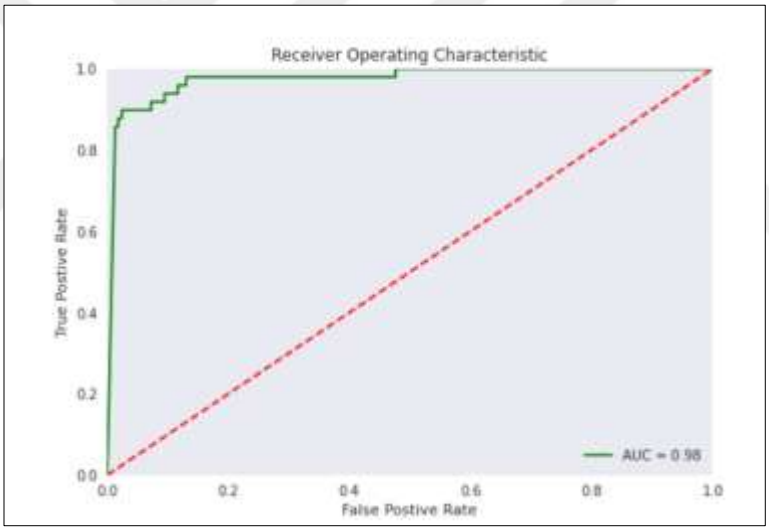

Figure 4.11: Naive Bayes ROC.

4.7 SUPPORT VECTOR MACHINE MODEL

After fitting of the SVM model, the model gives an accuracy of 99.474%, a F1-score of 0.381, a recall 0.939 and a precision of 0.239 (see Figure 4.12). The AUC of ROC is equal to 0.98 as shown in Figure 4.13.

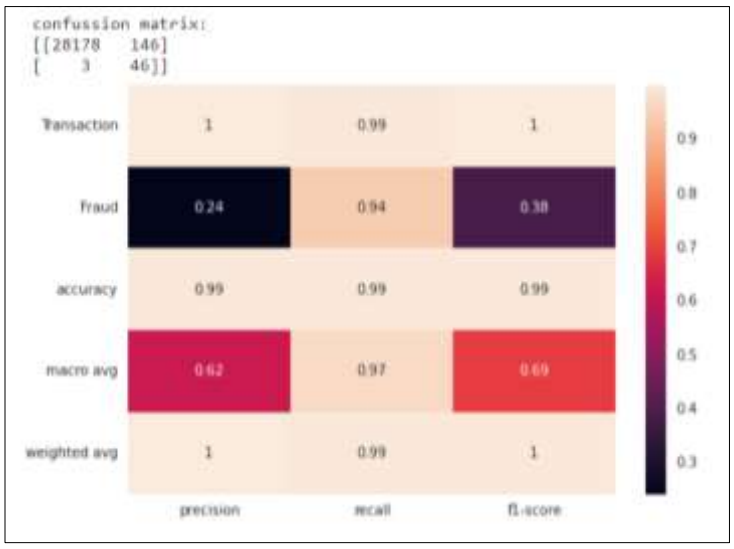

Figure 4.12: SVM Results.

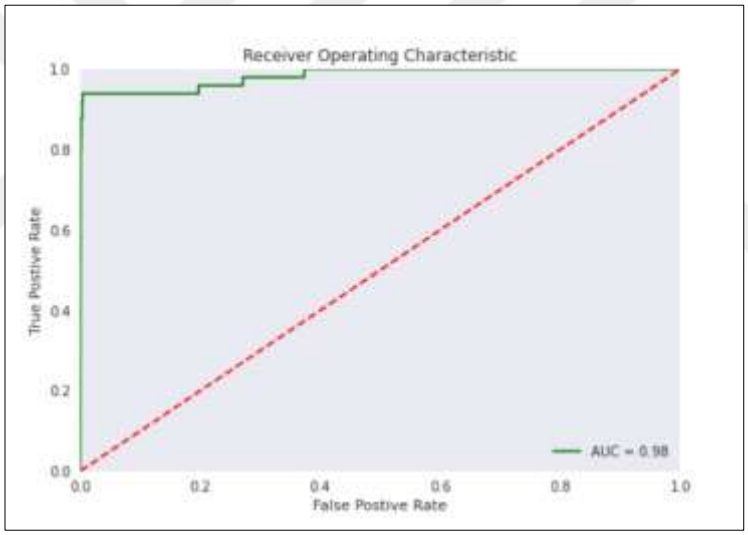

Figure 4.13: Svm Roc.

4.8 LOGISTIC REGRESSION MODEL

After fitting of the Logistic Regression model, the model gives an accuracy of 99.013%, a F1-score of 0.239, a recall 0.898 and a precision of 0.138 (see Figure 4.14). The AUC of ROC is equal to 0.99 as shown in Figure 4.15.

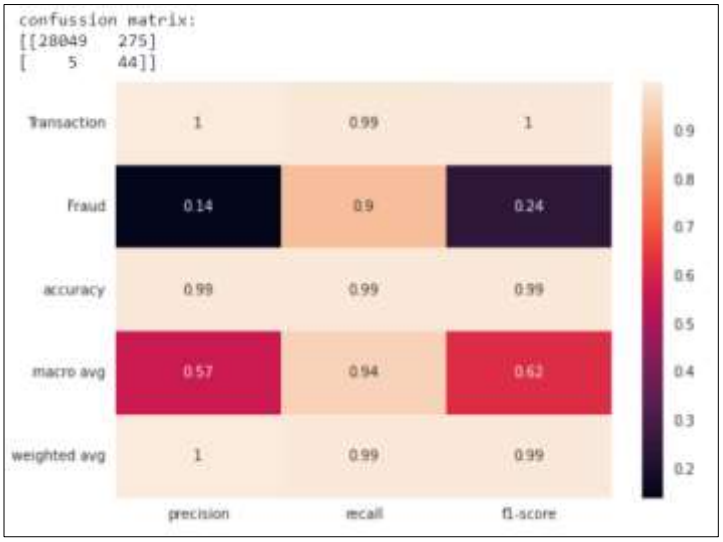

Figure 4.14: Logistic Regression Results**.**

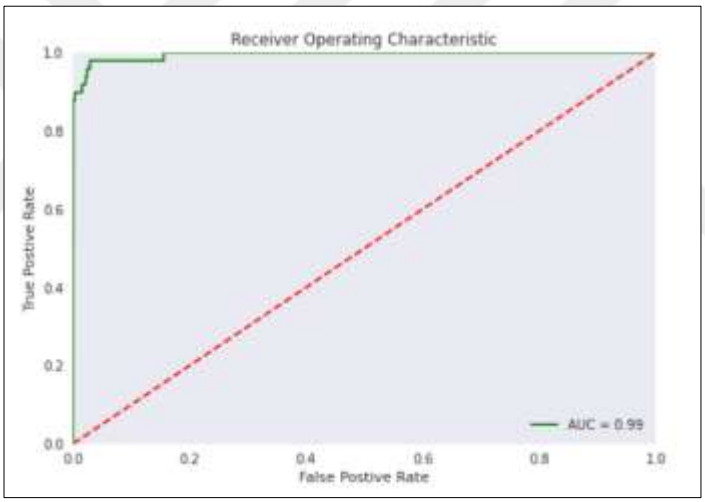

Figure 4.15: Logistic Regression Roc.

4.9 MULTILAYER PERCEPTRON (MLP) MODEL

After fitting of the MLP model, the model gives an accuracy of 99.926%, a F1-score of 0.796, a recall 0.837 and a precision of 0.759 (see Figure 4.16). The AUC of ROC is equal to 0.99 as shown in Figure 4.17.

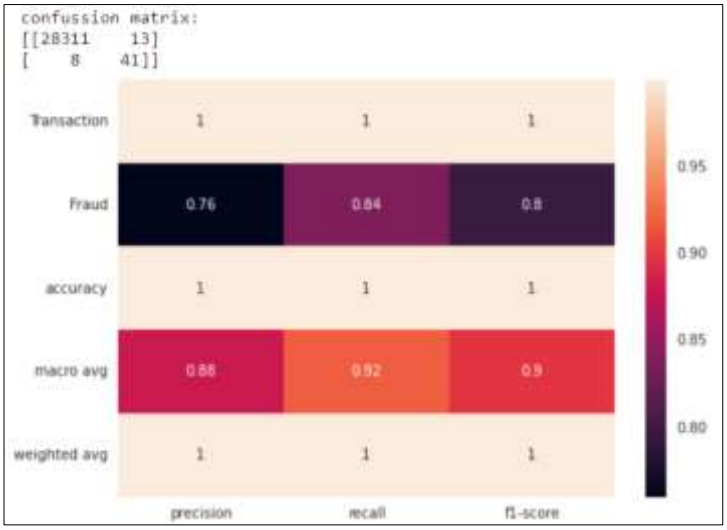

Figure 4.16: MLP Results.

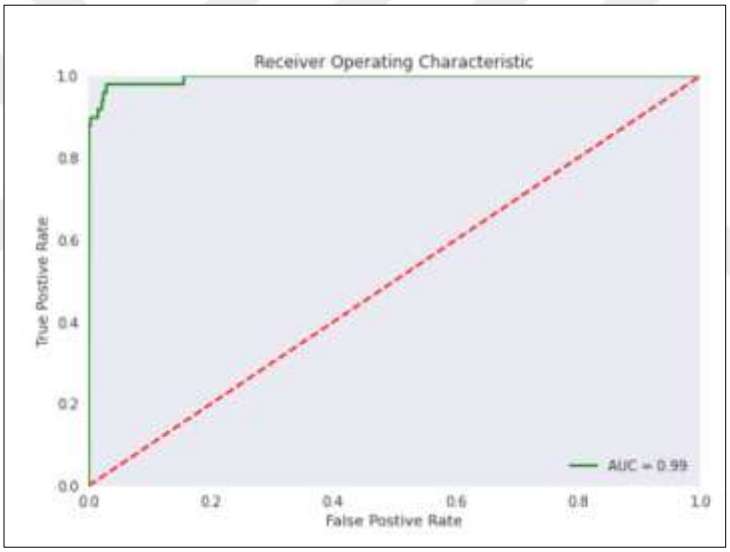

Figure 4.17: MLP ROC

4.10 ENHANCED XGBOOST MODEL

Adjustable values or weights are what are known as hyperparameters, and they are used in the process of defining how an algorithm learns. As was noted before, XGBoost provides a hyperparameter configuration that may be adjusted in a flexible manner. By making adjustments to XGBoost's hyperparameters, one may be able to release the program's full potential. To enhance the XGBoost model, hyperparameter is used. After fitting of the enhanced XGBoost model, the model gives an accuracy of 99.859%, a F1-score of 0.697, a

recall 0.836 and a precision of 0.597 (see Figure 4.18). The AUC of ROC is equal to 0.98 as shown in Figure 4.19.

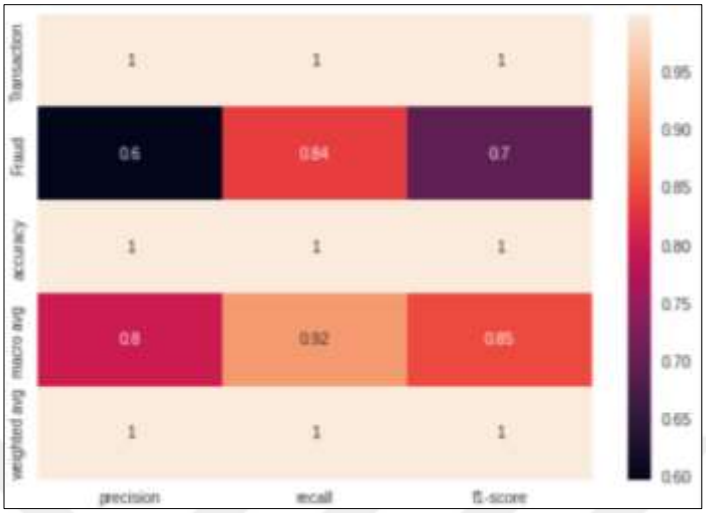

Figure 4.18: Enhanced XGBoost Results

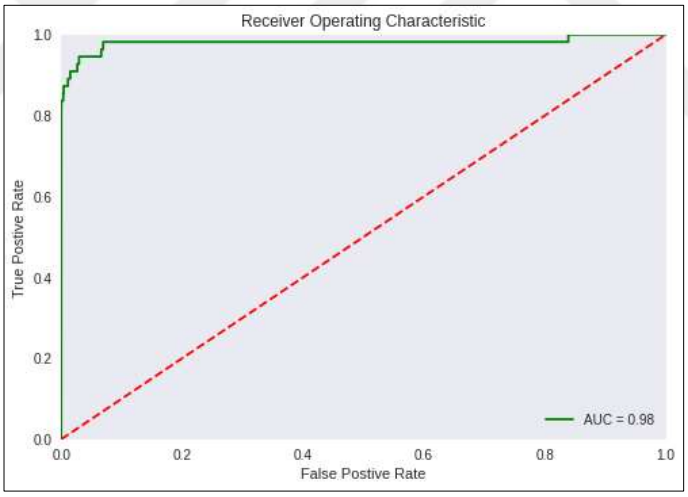

Figure 4.19: Enhanced XGBoost Roc.

4.11 COMPARISON

All results are compromised in Table 4.1 for better comparison. It seems that random forest works the best among all proposed models. XGBoost model performs well, however the enhancement affects most metrics negatively. Regardless F1 score, recall and precision, MLP model performs outstanding in term of accuracy and ROC AUC.

Table 4.1: Comparison Results.

Model	Accuracy	F1 score	Recall	Precision	ROC AUC
XGBoost	99.947	0.85	0.877	0.827	0.97
KNN	99.806	0.598	0.836	0.466	0.97
RF	99.957	0.875	0.857	0.893	0.98
DT	99.788	0.565	0.795	0.438	0.9
NB	97.561	0.113	0.898	0.06	0.98
SVM	99.474	0.381	0.939	0.239	0.98
LR	99.013	0.239	0.898	0.138	0.99
MLP	99.926	0.796	0.837	0.759	0.99
Enhanced XGBoost	99.859	0.697	0.836	0.597	0.98

5. CONCLUSION

In this thesis, credit card transactions is analyzed using machine learning and deep learning techniques to evaluate whether or not they involve fraudulent activity. The research made use of a dataset that was gathered and processed as a result of a collaborative effort between Worldline and the Machine Learning Group at ULB. The research focused primarily on big data mining and the identification of fraudulent activity. Fake transactions only account for 492 of the total 284,807 transactions. There is a significant imbalance in the dataset, with the positive class only accounting for 0.172% of all transactions in the database. When the predictive model is given data with a highly skewed class distribution, it tends to favor the samples that make up the bulk of the population. As a result, it commonly conceals fraudulent activity by making it appear as though the transaction was real. A data-level strategy is used that incorporates a number of different resampling methods, such as random under sampling, Tomek links removal, random oversampling, Synthetic Minority Over-sampling Technique (SMOTE), and a hybrid resampling strategy that consists of SMOTE and Tomek links removal, in order to solve this problem. In addition, in order to solve the problem of class disproportion, Algorithmic approaches is implemented such as bagging and boosting. In the search for the optimal classifier, a variety of trials were carried out. Evaluation and comparison of a number of different classification strategies is the objective of the system that has been proposed. These classification strategies include logistic regression, support vector machine, naive bayes classifier, decision tree, random forest, XGBoost, improved XGBoost, KNN, and multi-layer perceptron. These investigations were based on a wide variety of class distributions and features that were utilized during the training process. The fact that there are inconsistencies between the training data and the testing data during the evaluation of classifiers generated from distinct sets of trials demonstrates that careful thought has to be paid to the class distribution design of the training sets. Over the course of this analysis, both the True Negative (TN) and the False Negative (FN) rates were taken into consideration. According to the data, random forest appears to be the model that performs the best out of the ones that have been proposed. It has an accuracy rate of 99.957%, an F1-score of 0.875, recall of 0.857, precision of 0.893, and ROC AUC of 0.98. Even if the XGBoost model performs rather well, the enhancement has a detrimental impact on the great majority of the measurements. The MLP model performs extraordinarily well in terms of

accuracy and ROC AUC, independent of F1 score, recall, or precision. This is due to the fact that it has a large number of hidden layers.

REFERENCES

[1] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017, vol. 2017-January, pp. 1–9, Nov. 2017.

[2] “U.S. payment card fraud losses by type, Statista.” ,2022.

[3] “Card and Mobile Payment Industry News, Nilson Report Newsletter Archive.” ,2022.

[4] P. Juszczak, N. M. Adams, D. J. Hand, C. Whitrow, and D. J. Weston, “Off-the-peg and bespoke classifiers for fraud detection,” Comput Stat Data Ana*l*, vol. 52, no. 9, pp. 4521–4532, 2008.

[5] B. Manderick, “Credit card fraud detection using Bayesian and neural networks,” Proceedings of the 1st international naiso congress on neuro fuzzy technologies, Jan. 2002.

[6] A. D. Pozzolo and G. Bontempi, “Adaptive Machine Learning for Credit Card Fraud Detection,” 2015.

[7] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” Intelligent Data Analysis, vol. 6, no. 5, pp. 429–449, 2002.

[8] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, “Credit card fraud detection and concept-drift adaptation with delayed supervised information,” Proceedings of the International Joint Conference on Neural Networks, 2015.

[9] A. Dal Pozzolo, O. Caelen, Y. A. le Borgne, S. Waterschoot, and G. Bontempi, “Learned lessons in credit card fraud detection from a practitioner perspective,” Expert Syst Appl, vol. 41, no. 10, pp. 4915–4928, Aug. 2014.

[10] R. C. Holte, “Concept Learning and the problem of small disjuncts,”2020.

[11] JapkowiczNathalie and StephenShaju, “The class imbalance problem: A systematic study,” Intelligent Data Analysis, Oct. 2002.

[12] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating probability with under sampling for unbalanced classification,” Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015, pp. 159–166, 2015.

[13] B. Lebichot, F. Braun, O. Caelen, and M. Saerens, “A graph-based, semi-supervised, credit card fraud detection system,” Studies in Computational Intelligence, vol. 693, pp. 721–733, 2017.

[14] F. Carcillo, A. Dal Pozzolo, Y. A. le Borgne, O. Caelen, Y. Mazzer, and G. Bontempi, “SCARFF: A scalable framework for streaming credit card fraud detection with spark,” ArXiv, vol. 41, pp. 182–194, May 2017.

[15] P. M. Domingos, “Meta Cost: a general method for making classifiers cost-sensitive,” pp. 155–164, Aug. 1999.

[16] E. Aleskerov, B. Freisleben, and B. Rao, “CARDWATCH: A neural network-based database mining system for credit card fraud detection,” IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFEr), pp. 220–226, 1997.

[17] A. Srivastava, A. Kundu, S. Sural, and A. K. Majumdar, “Credit card fraud detection using Hidden Markov Model,” IEEE Trans Dependable Secure Computing, vol. 5, no. 1, pp. 37–48, 2008.

[18] G. Baader and H. Krcmar, “Reducing false positives in fraud detection: Combining the red flag approach with process mining,” International Journal of Accounting Information Systems, vol. 31, pp. 1–16, Dec. 2018.

[19] “Credit card fraud - Canada.ca.” 2023.

[20] C. C. Bertaut and M. Halisassos, “Credit Cards: Facts and Theories”.

[21] F. J. Rowe, L. R. Hepworth, K. L. Hanna, and C. Howard, “Visual Impairment Screening Assessment (VISA) tool: pilot validation,” BMJ Open, vol. 8, no. 3, pp. 20562, Mar. 2018.

[24] K. Pandey, P. Sachan, Shakti, and N. G. Ganpatrao, “A Review of Credit Card Fraud Detection Techniques,” Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, pp. 1645–1653, Apr. 2021.

[25] V. D. S, S. B. Basapur, V. Abhay, and N. Natesh, “Credit Card Fraud Detection Systems (CCFDS) using Machine Learning (Apache Spark),” International Research Journal of Engineering and Technology, 2023.

[26] “Beginners Guide to Deep Reinforcement Learning | by Abacus.AI | Abacus.AI Blog (Formerly RealityEngines.AI) , Medium.” 2023.

[27] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29, 2004.

[28] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2011.

[29] F. Roli, J. Kittler, and T. Windeatt, Eds., “Multiple Classifier Systems,” vol. 3077, 2004.

[30] L. Rokach and O. Maimon, “Decision Trees,” Data Mining and Knowledge Discovery Handbook, pp. 165–192, 2006.

[31] P. Cunningham and S. J. Delany, “k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples),” ACM Comput Surv, vol. 54, no. 6, 2020.

[32] P. Cunningham and S. J. Delany, “K-Nearest Neighbour Classifiers-A Tutorial,” ACM Comput Surv, vol. 54, no. 6, Jul. 2021.

[33] M. Singh, W. : Www, A. Kataria, and M. D. Singh, “A Review of Data Classification Using K-Nearest Neighbour Algorithm Design & Development of Wireless Health Monitoring System for Ambulatory Subjects View project Computer Aided System for Classification of mammograms View project Aman Kataria Central Scientific Instruments Organization International Journal of Emerging Technology and Advanced Engineering A Review of Data Classification Using K-Nearest Neighbour Algorithm,” Certified Journal, vol. 9001, no. 6, 2023.

[34] K. Kumari and S. Yadav, “Linear regression analysis study,” Journal of the Practice of Cardiovascular Sciences, vol. 4, no. 1, pp. 33, 2018.

[35] A. Schneider, G. Hommel, and M. Blettner, “Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications,” Dtsch Arztebl Int, vol. 107, no. 44, pp. 776, 2010.

[36] E. Martin *et al.*, “Support Vector Machines,” Encyclopaedia of Machine Learning, pp. 941–946, 2011.

[37] H. A. Park, “An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain,” J Korean Acad Nurs, vol. 43, no. 2, pp. 154–164, 2013.

[38] K. P. Murphy, “Machine Learning a Probabilistic Perspective”.

[39] “Classification and Regression Trees for Machine Learning - MachineLearningMastery.com.” 2023.

[40] “How Random Forest Algorithm Works in Machine Learning | by Synced | Medium.” 2023.

[41] A. J. Ferreira and M. A. T. Figueiredo, “Boosting Algorithms: A Review of Methods, Theory, and Applications,” Ensemble Machine Learning, pp. 35–85, 2012.

[42] “Boosting Algorithm, Boosting Algorithms in Machine Learning.” 2023.

[43] “A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning - MachineLearningMastery.com.” 2023.

[44] K. Gurney and K. N. Gurney, “An Introduction to Neural Networks (Google eBook),” pp. 234, 1997, 2023.

[45] “Credit Card Fraud Detection, Kaggle.” https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud ,2023.