DEMOCRATIC AND POPULAR
REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION
AND SCIENTIFIC RESEARCH

LARBI TEBESSI UNIVERSITY - TEBESSA

FACULTY OF NATURAL SCIENCES AND LIFE SCIENCES

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

THESIS

FOR THE COMPLETION OF THE MASTER'S DEGREE IN

COMPUTER SCIENCE

Specialty:

<u>THEME</u>

Machine Learning For The Detection of Money Fraud

Presented by:

Sadaani Ahmed Oualid

Zitari Marouane

In front on the following committee members:

Nait-Hamoud Mohamed Cherif	MCA	Tebessa University	President
Khediri Abderrazak.	MCA	Tebessa University	Examiner
Guerieb Nawel	MCB	Tebessa University	Supervisor

Academic Year 2023/ 2024

بسم الله الرحمن الرحيم

Acknowledgment

We would first like to thank the Almighty and Merciful God, who gave us the

strength and patience to do this modest work, which allowed us to be what we are today.

Our sincere thanks go to Dr. Gerieb Nawel for accepting to supervise this work, and for her advice and valuable guidance, her patience, and her support throughout our research.

In particular, we would like to thank the jury members for accepting.

to evaluate this modest work: Nait-Hamoud Mohamed Cherif and Dr. Khediri Abderrazak.

We thank all the people at Sheikh Larbi Tebessi University.

Finally, we also extend our thanks to all those who contributed.

from near or far to the realization of this work

Dedicated

we dedicate this memory to our dear families who have always been by our side and have always supported us throughout these long years of studies.

To our dear friends and colleagues, who have been our source of comfort, laughter and of encouragement in times of doubt and fatigue. Your friendship has been the refuge precious in this adventure.

And to all those who, near or far, believed in us, supported us and encouraged to pursue our dreams. This achievement is as much yours as our.

Zitari Marouane And Sadaani Ahmed Oualid

Abstract: This manuscript focuses on the detection of **credit card fraud** using machine learning techniques. The rapid increase in digital transactions, especially during the COVID-19 pandemic, has heightened the need for robust fraud detection mechanisms. This study explores the various types of **bank fraud**, particularly credit card fraud, and provides an overview of the evolution of payment cards and electronic payment systems. The research delves into different **machine learning algorithms** and tools used for fraud detection, including data preprocessing, feature selection, and handling imbalanced data. It also outlines the system architecture for implementing these techniques in real-world applications. The findings underscore the effectiveness of **machine learning** in enhancing fraud detection and suggest future research directions for improving security measures

Key words: credit card fraud, bank fraud, machine learning.

ملخص يركز هذا المخطوط على الكشف عن الاحتيال ببطاقات الائتمان باستخدام تقنيات التعلم الآلي. وقد أدت الزيادة السريعة في المعاملات الرقمية، خاصة خلال جائحة كوفيد-19، إلى زيادة الحاجة إلى آليات قوية للكشف عن الاحتيال. تستكشف هذه الدراسة الأنواع المختلفة من الاحتيال البنكي، لا سيما الاحتيال ببطاقات الائتمان، وتقدم نظرة عامة على تطور بطاقات الدفع وأنظمة الدفع الإلكتروني. يتناول البحث مختلف الخوارزميات والأدوات المستخدمة في الكشف عن الاحتيال، بما في ذلك معالجة البيانات المسبقة، واختيار الميزات، والتعامل مع البيانات غير المتوازنة. كما يوضح هيكل النظام لتطبيق هذه التقنيات في التطبيقات الواقعية. تؤكد النتائج فعالية التعلم الآلي في تعزيز الكشف عن الاحتيال وتقترح اتجاهات البحث المستقبلية لتحسين تدابير الأمان.

الكلمات الدالة: الكشف عن الاحتيال في بطاقات الائتمان،الاحتيال البنكي ، والتعلم الآلي.

Résumé : Ce manuscript se concentre sur **la détection de la fraude par carte de crédit** à l'aide des techniques **d'apprentissage automatique**. L'augmentation rapide des transactions numériques, en particulier pendant la pandémie de COVID-19, a accentué la nécessité de mécanismes robustes de détection de la fraude. Cette étude explore les différents types **de fraude bancaire**, en particulier la fraude par carte de crédit, et fournit un aperçu de l'évolution des cartes de paiement et des systèmes de paiement électronique. La recherche examine les différents algorithmes et **outils d'apprentissage automatique** utilisés pour la détection de la fraude, y compris le prétraitement des données, la sélection des caractéristiques et la gestion des données déséquilibrées. Elle décrit également l'architecture système pour la mise en œuvre de ces techniques dans des applications réelles. Les résultats soulignent l'efficacité de l'apprentissage automatique pour améliorer la détection de la fraude et suggèrent des pistes de recherche futures pour améliorer les mesures de sécurité.

Mots-clés : Détection de fraude par carte de crédit, fraude bancaire, apprentissage automatique

Table of Contents

List of Figures

List of Tables

General Introduction

General Introduction

Credit card usage in the Arab world, including countries such as the UAE, Saudi Arabia, Algeria, and Tunisia, has shown significant expansion in recent years, reflecting global trends towards digital and financial integration. The UAE and Saudi Arabia have reported higher adoption rates, indicative of increasing reliance on electronic payments among their populations. However, precise statistics on credit card fraud across the region, including in countries like Algeria and Tunisia, can be challenging to ascertain due to varying reporting standards and data availability. Nonetheless, credit card fraud remains a universal concern, prompting financial institutions across the Arab world to invest in advanced security technologies.

In this work, we tested six machine learning techniques to select the best model for credit card fraud detection. The primary objective was to prepare the dataset for modeling, addressing challenges such as imbalance and outliers.This manuscript is organized as:

The first chapter aims to offer a clear understanding of what monetary fraud entails and different varieties of it, of which credit card fraud is a part. It explains the historical development and current organization of electronic payment systems and considers the effects of fraud worldwide. Different techniques of credit card fraud, including CNP fraud, skimming, phishing scams, and account hijacking, are explained. Fingerprints and signatures are also discussed as traditional and advanced methods of fraud detection.

The second chapter seeks to provide an understanding of what artificial intelligence and machine learning entail and how they can be useful in card fraud detection. In this chapter, supervised learning, unsupervised learning as well as semi-supervised learning followed by some of the well known classification algorithms are presented.

The last chapter is dedicated to present our credit card fraud detection model. First, it introduces the development tools and languages including Python, scikit-learn Anaconda and Jupyter. Then, data preprocessing like data balancing, feature selection and partitioning of the dataset into training and testing set has been explained. Finally, this chapter compares the accuracy of different machine learning models and selects the best model based on the best performance measures.

Chapter I: State Of The Art on Monetary Fraud

1. Introduction

Banks play a significant role in the economy and are one of the most essential components of the state. Their role is not only to provide liquidity but also to facilitate the development and advancement of the economy. However, the majority of banking systems have recently been exposed to piracy and theft, particularly during the Corona era, due to a lack of sufficient means to deter fraud and theft, and in this section, we will become acquainted with banking fraud, its types and effects, and particularly the means of combating and attempting to eliminate it.

2. Definition of bank fraud

Bank fraud may be defined as unethical and/or criminal conduct committed by an individual or organization in order to obtain or receive funds from a bank or financial institution. In general, bank fraud can include any illegal activity aimed at defrauding a financial institution. It can be an intentional activity to receive assets—money, mobile values, credits, or financial institution assets—by using false or misleading information. The law provides a broad definition of bank fraud, and several facets of this offense must be considered. The informational cost is due expanding and continual diversity, and in order to turn them into relevant data in numerous domains (education, trade, scientific research, etc.), the researchers engaged in data mining [1].

3. Electronic payment system structure

Monetization is an ecosystem that relies on a large number of actors, whether they are institutions, businesses, or individuals. As we can see in Figure 1.1, the operation of the monetary system is sometimes described as a "four coins" system [2].

The terminal's primary function in the monetary system is to serve as an interface between the card and the acquisition and management network. It is still an active element, in charge of:

- ☞ Receiving acquisition parameters;
- ☞ Selecting payment applications;

☞Evaluating and accepting/refusing a transaction;

☞Driving and operating the transaction.

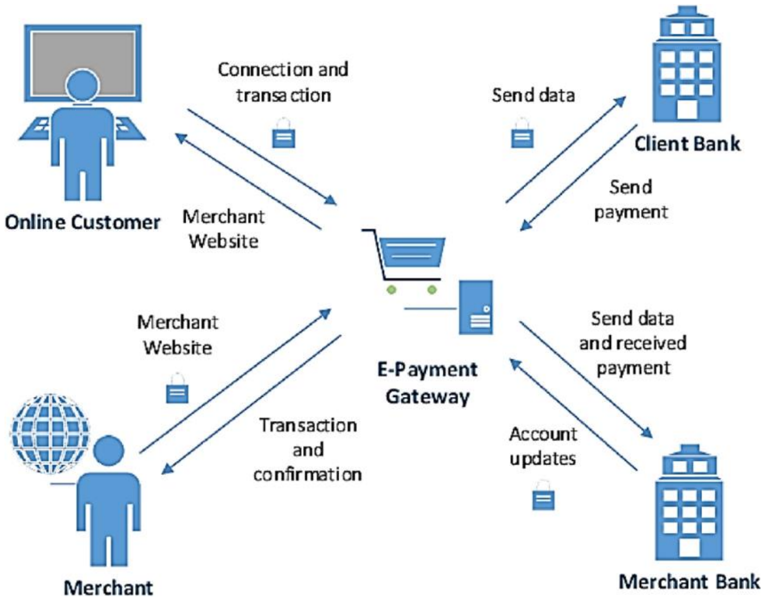


Figure 4.1: Electronic payment system structure

4. History of the payment card

Payment cards date back to the 1950s, as shown in figure 1.2, with the presentation of a Diners' Club card, it was possible to pay for meals, travel, and business expenses in partner stores, first in the United States and then across the world [3].

The payment card did not replace the paper support until 1959, when American Express introduced its first plastic card. However, these cards remain private, and it takes until the 1960s to see the emergence of inter-bank transactions, with the establishment of Bank America card in 1958 (later renamed Visa) and the Inter Bank Card Association in 1966 (later renamed MasterCard Worldwide). Card embossing has been effective since 1959, allowing merchants to create American Express card impressions, facilitating replication and avoiding errors [4] .

Figure 1.5: payment card style example

5. Visual interface of the payment card

The form and visual of the payment card adhere to international standards, including ISO 7810, which defines its size; ISO 7811-1, which defines the position of the number line, the name and address zone, and the characteristics of readable characters; ISO 7811-2, which defines the position of magnetic tracks; ISO 7812-1, which defines the format of the number; and ISO 7816-2, which defines the position of contacts [4]. Thus, the visual identity of the payment card can be distinguished by the information contained in the payment card as shown in Figure1.3.

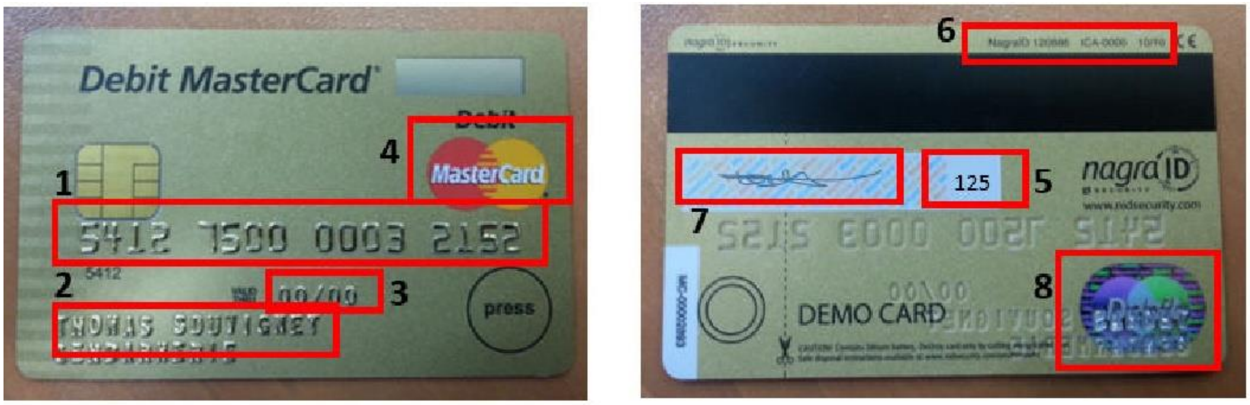


Figure 1.6: Visual interface of the payment card

The principal information contained in the payment card is:

1. Pin Account Number (PAN);

2. The name of the holder;

3. Expiration date;

4. Payment system logo (s);

5. Verification code (CVV2/CVC2) for remote payment.

6. Manufacturer's serial number (or mask);

7. Signature of the holder;

8. Holograms and UV security.

6. The Types of Bank Fraud

Fraud in banks can take several forms, including internal (committed by bank employees) and external (committed by clients, individuals, or entities outside the bank). Among the most common of its sorts are the following three forms:

6.1. Credit Card Fraud

It is a fraudulent attempt by an individual or organization to use a credit card or debit card without proper authorization for financial benefit. One of the most common types of card-based fraud occurs after the theft or loss of a debit or credit card. In this case, an unauthorized party can gain access to another person's credit card or debit card numbers. However, if the PIN code is not known, automatic withdrawal of funds is nearly impossible and The criminals may possess different illegal means of gaining the cardholders' personal data for illegal purposes so that the fraudsters can do illegal card transactions through online access. This is a synopsis of the ways that card fraud can happen online, derived from the sources mentioned: this is a synopsis of the ways that card fraud can happen online, derived from the sources mentioned [6,7]:

6.1.1. Skimming Devices

The fraudsters use these magnetic devices to get card information and card number. This type of equipment facilitates fraudsters to forge cards used for unlawful payouts at the point of sale of credit cards or ATMs. They may be installed on valid card readers or ATMs that have been compromised. Account Takeover is one of this types of fraud, criminals possess the power to contact card providers in order to order a new card on the account without the victim's knowledge; they will also be able to change the card's address which will also alert the bank. They get access to the victim's account through a card affixed to it or that is used to obtain the replacement card [8,9].

6.1.2. Phishing schemes

Phishing is the name of a worldwide fraudulent scheme in which the fraudsters create counterfeit emails, text messages, and advertising immediately or anyway to create a sense of legitimacy for the organization to get the targeted information or infiltrate malware.

- These particular scams are the emails that appear to be from trusted sources like banks, internet service providers, or government agencies, with the request for some information such as account numbers, passwords, or Social Security numbers.
- There are manifold phishing attacks that include traditional email attacks, malware phishing, spear fishing, atomic variation phishing (smishing), voice phishing (vishing), pharming, cloned phishing, man-in-the-middle phishing, business email compromise (BEC), and phishing marketing for malware.
- These attacks, known as phishing, are designed to make people believe that the message is legitimate and from another person. This can be done through emails, text messages, and social media messages by mainly tricking individuals into providing sensitive data or clicking on malicious links that can cause financial loss, identity theft, or the installation of malware on their devices.
- Impersonation is usually used by people doing phishing scams; therefore, being vigilant and not clicking on any questionable links or even attachments, confirming the authenticity of the sender, following up with a call directly to the company, and, of course, also reporting any phishing attempts to the authorities is something people should do.

6.1.3. Account Takeover

Account takeover (ATO) is one kind of cybercrime that comes into play when cybercriminals use their knowledge and experience to access users' online accounts illegally and commit various malicious activities like stealing data, delivering malware, disgraceful use of account permissions, etc.

This is one specific form of attack that has developed rapidly since its origin, and now cybercriminals use advanced techniques and automated tools to attack multiple accounts at a time.

ATO attacks represent an unpleasant trend, featuring financial services as the main target, and BEC losses exceeded $1.8 billion in 2020. The rising incidences of these attacks pose a great danger to institutions, thus demanding proactive actions such as establishing awareness, monitoring, and fighting them to protect digital data and maintain cyber security.

In order to prevent account takeover fraud, companies apply a "defense in-depth" strategy, improve account security practices, and use technologies like Zero Trust Network Access (ZTNA) for identity verification and managing the risk of ATO attacks correctly.

6.2. Check Fraud

Check fraud is becoming one of the most serious problems that businesses and financial institutions face. With the advancements of information technology, it is becoming increasingly easy for criminals, either alone or in organized gangs, to use the criminal justice system in order to defraud innocent victims who are waiting for their money.

Fraud by Check refers to the unauthorized use of a Check for a financial gain, and it is committed by the use of computer-assisted publication and copying to create or copy a financial document, which entails deleting all or part of the information and manipulating it for the benefit of the criminal. The victims include financial institutions, businesses that receive and issue Checks, and consumers. In the majority of cases, these crimes begin with the theft of a financial document [10].

Check fraud can take several forms, some of which are listed below:

- Placing a check on an account without proper authorization.
- Modifying a check involves changing bank information, such as account numbers.
- Use a check to make a payment while knowing there aren't enough funds in the account.
- Modification of payment amount on a check.
- The use of check for bogus bills.

7. Global impact of bank fraud

The impact of fraud in the foreign banking sector is felt by everyone, whether as a customer or a citizen. Fraud has several negative effects on society, because this industry plays an essential role in our society and economy, particularly in the banking sector, which is prone to fraud and His success or failure is an extremely crucial factor in determining the success of the society.

Fraud is a major cause of bank failures. Indeed, the number of frauds that occur in banks is always increasing, which has prompted a wave of investigation since it has a complete impact on the banks' poor performance. High balances on bank accounts are considered a loss for him since they generate no revenue for the bank. As a result, the bank faces a challenging issue, a lack of liquidity.

8. Detection Methods

It is not feasible to completely eliminate bank fraud. There are methods to improve and increase the likelihood of this happening. There exist two more effective methods, which are the following:

8.1. Fingerprint

The digital footprint represents an important fraud detection technique, particularly in the banking sector, because it contains a characteristic that distinguishes one person from the next, providing a unique means of identification for each person in the world. She is used in the field of online banking services due to the difficulties that banks face in identifying sources based only on IP addresses, which might change over time [6].

To address this, an appropriate solution has been proposed. The access device is determined by a component that must be downloaded and installed on the client device. This component generates a digital footprint of the access device and sends it to the bank's website as part of the transaction details. The digital signature is then calculated by applying a cryptographic function to hardware and software information such as the processor, operating system serial numbers, MAC addresses, and certain configuration details [6].

- To implement the component, we need the following three basic requirements:
- The device generates a unique digital fingerprint for each unique user account.
- It provides a certain characteristic while generating digital fingerprints due to the difficulty of identity theft by other devices.
- The device notifies the user of the new digital fingerprint every time the device's configuration changes.
- In fact, the proposed system is based on the component that is now used by the online banking system.

8.2. Signature

The most common method used by banks and financial institutions, as well as their clients, to authenticate a person's identity is signature verification, but it is a time-consuming method that requires practice and diligence and does not require a valid signature for comparison.

To guarantee that this operation goes successfully, financial organizations back offices are the best venues to check signatures. However, they are presented with various challenges, of which we shall highlight the two most significant: The bank cannot check the legitimacy of signatures on each payment item because there are too many check presented for payment. Because of the widespread availability of numeric peripherals, copying non-authorized signatures is an easy technique to create fraudulent papers.

9. Advancements in Credit Card Fraud Detection

In literature, there are different approaches based on artificial intelligence to successfully solve credit card fraud problems. These include supervised and unsupervised methods of analysis. The mentioned researchers have introduced new ideas for enhancing the performance of models for fraud analysis.

The following is a brief summary of the existent researches, which show how to detect credit card fraud using various machine learning techniques.

9.1. Credit Card Fraud Detection Approaches

Authors in [12] undertook research on the application of machine learning techniques for the identification of fraudulent credit cards. Researchers applied the Logistic Regression algorithm, the Decision Tree algorithm, the Random Forest algorithm, and the Naïve Bayes algorithm to examine and recognize fraudulent transactions in online transactions. This work assessed their performance on a real-world dataset, and it is evident that these techniques can help in detecting fraudulent transactions effectively and accurately. Citing feature selection and dimensionality reduction methods like PCA, the authors optimized the performance of the algorithms. Among the classifiers, Random Forest has the highest prediction capability, by mere tenths, trailed by Decision Trees and Naïve Base classifiers. In particular, the study revealed that ensemble methods and feature selection techniques are extremely important when it comes to the accuracy of credit card fraud detection models.

A credit card fraud detection model utilizing machine learning strategies have been proposed in [13] . The intended purpose was to give the model the ability to cope with the large number of transactions and the dynamic pattern of fraud. As a result of this, methods like feature selection techniques, oversampling techniques, and ensemble methods for dealing with the imbalance problem were used in order to enhance the performance of the model. To deal with these problems, feature selection was used to combat the high-dimensionality of the data, and oversampling was used to combat the skewed class distribution. Expanding on that further, concepts learned under the topics of bagging as well as boosting involve learning more than one machine learning algorithm with the same data with the intention of improving the performance. The model resulted in high levels of accuracy and precision with regard to fraudulent transactions and can work more efficiently than models that use only one algorithm. It is also possible to focus on other domains, often involving high-dimensional datasets and significant class imbalances, including medical diagnosis or credit risk models.

In the study presented in [14], authors adopted a machine learning approach to take the transaction history of an account, represent it as a sequence, and identify fraudulent and legitimate transactions. It exposed the common issues of fraud detection using data, including the skewed ratio between the fraud cases and the genuine ones, dependency in the samples, and concept evolution. They used both feature engineering and machine learning algorithms to reduce the rate of wrong classifications. Specifically, the study revealed that including temporal dependencies in data through LSTM networks had proven highly effective for fraud detection, more so for offline transactions. The authors also focused their discussion on the aspect of feature engineering in improving the performance of the fraud detection models.

Authors in [14] describe a new approach to identifying credit card fraud with the help of CNN. The researchers apply the CNN approach to extract features from the sequence of transactions and classify fraudulent ones while utilizing the capability of CNN to live up to spatial-temporal patterns. To sum up, it is worth noting that the applied methodology is suitable for dealing with the high dimensionality of data and data imbalance and, therefore, can effectively detect credit card fraud. It also highlights the use-case implementation of the methodology in practice in areas like real-time fraud scheme identification and prevention.

10. Conclusion

In short, the chapter highlights the utmost significance of the implementation of more advanced security measures, the never-ending vigilance and technological advancements in the fight against bank fraud, the protection of customer data, and the stability of financial systems. The facts here underline the role of those measures that are done in advance, sound detection techniques and continuous designing of secure technology tools in protection of banking transactions against fraudulent operations.

Chapter II: Introduction to Machine Learning Techniques

1. Introduction

Artificial intelligence (AI) is a fast expanding science that seeks to create intelligent computers capable of doing activities that would traditionally need human intelligence. AI has several subfields, including machine learning, deep learning, and natural language processing. Machine learning, in particular, has received a lot of interest because of its capacity to learn from data and anticipate outcomes without being explicitly programmed.

Machine learning algorithms may be divided into three categories: supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning trains models using labeled data, whereas unsupervised learning uses unlabeled data to identify hidden patterns and structures. Semi-supervised learning uses labeled and unlabeled data to improve model performance.

Several machine learning algorithms have been created, each with unique strengths and disadvantages. The most often utilized algorithms are K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Naive Bayes, and Logistic Regression.

2. Artificial Intelligence (AI)

Artificial intelligence (AI) is a new discipline that studies and develops theories, methods, techniques, and application systems to simulate and expand human intellect. In 1956, John McCarthy proposed the notion of AI for the first time, defining it as "the science and engineering of the fabrication of intelligent machines, particularly intelligent software." Artificial intelligence is responsible for making machines function intelligently, similar to how the human mind functions. Currently, AI has evolved into an interdisciplinary course encompassing several fields as shown in figure 2.1 [15].

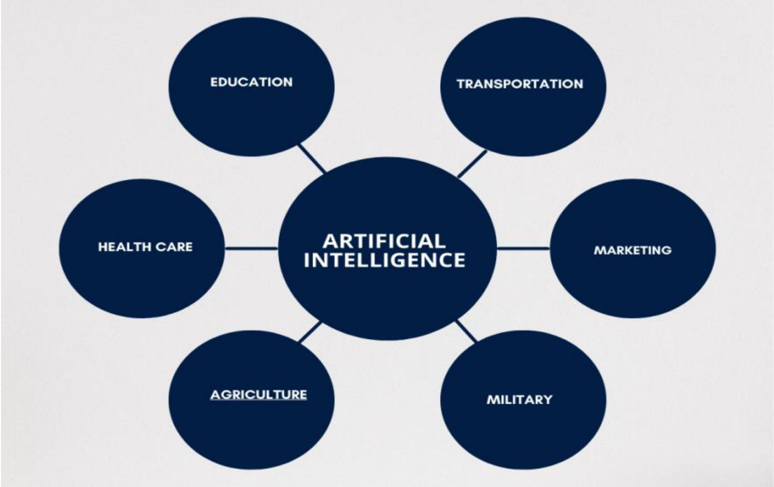


Figure 2.1: Artificial intelligence fields

3. Machine Learning

Automated learning is not a new technology. The first artificial neural network, known as the "perceptron," was developed in 1958 by American psychologist Frank Rosenblatt. Automated learning is a technology that allows computers to learn without explicit programming. However, for learning and development, computers require data for analysis and training. Automated learning uses several methods to generate models from data. These methods are essentially algorithms. The goal is to enable machines or computers to solve complicated problems by handling large amounts of information [16].

3.1. Types of Machine Learning

Within the field of machine learning, there are two primary categories of learning: supervised and unsupervised as shown in figure 2.2. The primary distinction between these two kinds is that supervised learning is done based on a fact. Said another way, we have a prior knowledge of what the sample value should be before it leaves. The goal of supervised learning is to learn a function using expected results and given sample sizes that approximate the relationship between the input and output observables in the data as closely as possible. Conversely, unsupervised learning does not yield labeled results. Its goal is to extract the natural structure present in a collection of data points [17].

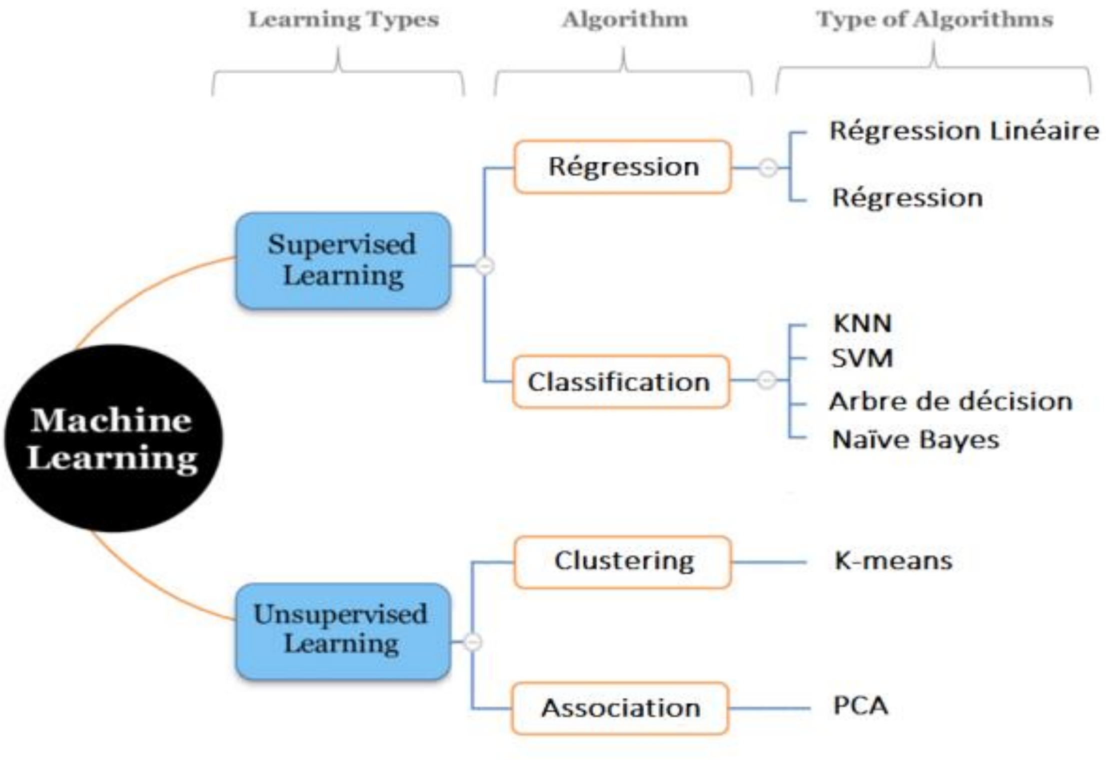


Figure 2.2: Types of Machine Learning

3.1.1. Supervised learning

Supervised learning is a type of machine learning where a model is trained on a labeled dataset. In this approach, the dataset contains input-output pairs, where the input data is accompanied by the correct output (label). The model learns to map the input data to the correct output by finding patterns and relationships within the data. The main goal of supervised learning is to make accurate predictions or classifications when presented with new, unseen data [16].

☞ **Key Components**

1. *Labeled Data*: Training data consists of input-output pairs where each example is associated with a label or target variable.
2. *Learning Algorithm:* The algorithm uses the labeled data to learn a mapping from input variables (features) to the output variable (target).
3. *Prediction*: After training, the model can make predictions on new, unseen data by generalizing from the training set.

☞ **Types of Supervised Learning**

1. *Regression:* When the target variable is continuous, such as predicting house prices based on features like size, location, etc.
2. *Classification:* When the target variable is categorical, such as predicting whether an email is spam or not based on its content.

☞ **Process:**

1. *Data Collection:* Gather a dataset with labeled examples.
2. *Training:* Feed the labeled data into the algorithm to adjust its parameters.
3. ***Evaluation****:* Assess the model's performance on a separate validation or test set to ensure it generalizes well.
4. *Prediction:* Deploy the trained model to make predictions on new data.

☞ **Applications**

1. Supervised learning is widely used in various fields:
2. Image and speech recognition
3. Medical diagnosis
4. Financial forecasting
5. Natural language processing (NLP)
6. Recommendation systems

3.1.2. Unsupervised learning

Unsupervised learning involves training algorithms on data sets without labeled responses. Instead of being told the correct answer, the algorithm explores the data and identifies patterns or hidden structures on its own.

☞ **Key Concepts**

1. *Unlabeled Data**:*** The input data consists of features without corresponding target variables or labels.
2. *Learning Structure**:*** Algorithms aim to discover inherent patterns, relationships, or groupings within the data.
3. *Clustering**:*** A common task in unsupervised learning where the algorithm groups similar instances together based on features.

4. *Dimensionality Reduction:* Another task where algorithms reduce the number of variables under consideration, aiming to retain essential information while simplifying the dataset.

☞ **Types of Unsupervised Learning**

1. *Clustering:* Algorithms group similar instances together into clusters. Examples include k-means clustering, hierarchical clustering, and DBSCAN.
2. *Dimensionality Reduction:* Techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor embedding (t-SNE) reduce the number of variables or features in the dataset while preserving essential information.
3. *Association Rule Learning:* Finding interesting relationships or associations among variables in large datasets, often used in market basket analysis or recommendation systems.

☞ **Applications**

Unsupervised learning finds applications in various domains:

- *Customer segmentation:* Identifying distinct groups of customers based on purchasing behavior.
- *Anomaly detection:* Finding unusual patterns that do not conform to expected behavior, such as fraud detection.
- *Data compression:* Representing data in a more compact form while retaining essential information.
- *Feature learning:* Automatically discovering features or representations that are useful for subsequent tasks.

4. Machine learning algorithm

In this section we describe briefly the well known machine learning algorithm that we have used in this work.

4.1. K-Nearest Neighbors (KNN)

The K-Nearest Neighbor (KNN) approach is supervised which stores all available labeled examples and their classes. For a new sample, the algorithm finds the 'k' nearest neighbors from the training dataset based on a distance metric (e.g., Euclidean distance) as shown in Figure2.3.

It then assigns the class label by majority vote (for classification) or averages the labels (for regression) of its k nearest neighbors [18].

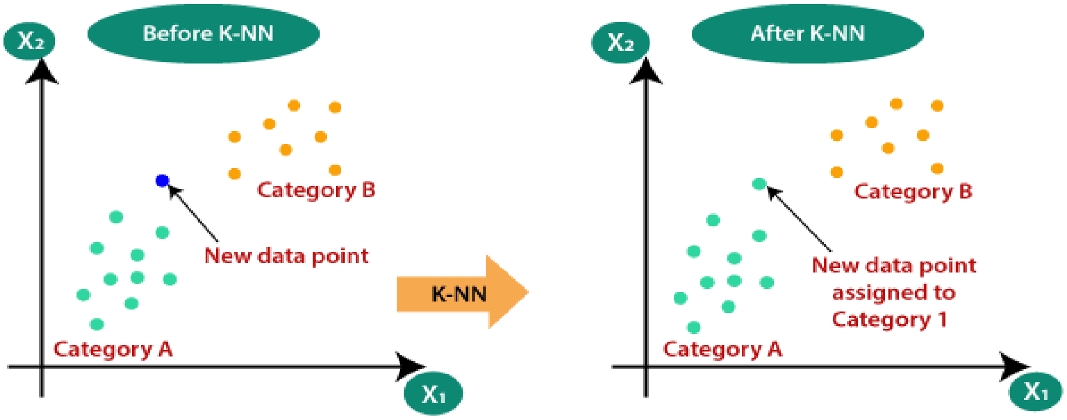


Figure 2.3: K-Nearest Neighbors algorithm

4.2. Support Vector Machine (SVM)

SVM stands for Support Vector Machine. It's a type of machine learning algorithm used for classification (and also regression tasks, though less commonly). SVM finds a line (or hyper plane) that best separates different classes in the data. The SVM aims to find the best possible separation (hyper plane) between classes as shown in Figure 2.4.

. It maximizes the distance between the hyper plane and the nearest points (support vectors) of each class. For non-linear data, SVM uses a kernel function to transform the data into a higher-dimensional space where classes become separable [19].

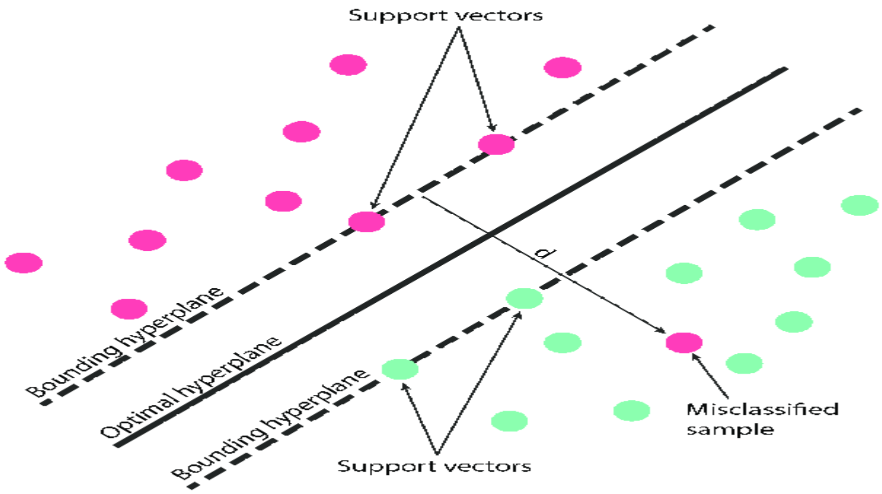


Figure 2.4: Support Vector Machine

4.3. Decision Trees

Decision Tree is a tree-like structure where each internal node represents a "decision" based on a feature attribute. It partitions the data into subsets that contain instances with similar characteristics. At the leaf nodes, the final output is a class label (for classification) or a continuous value (for regression) [20] as shown in Figure 2.5.

The main decision trees operations are:

- ☞ Begins with the entire dataset at the root node.
- ☞ Features are selected based on criteria like information gain (for classification) or variance reduction (for regression).
- ☞ Each internal node represents a decision based on a feature, splitting the dataset into smaller subsets.
- ☞ This process continues recursively until the leaf nodes are pure (contain only instances from one class) or a stopping criterion is met.
- ☞ Each path from the root to a leaf represents a decision rule.
- ☞ Rules are simple to understand and interpret; making decision trees a popular choice for explanatory modeling.

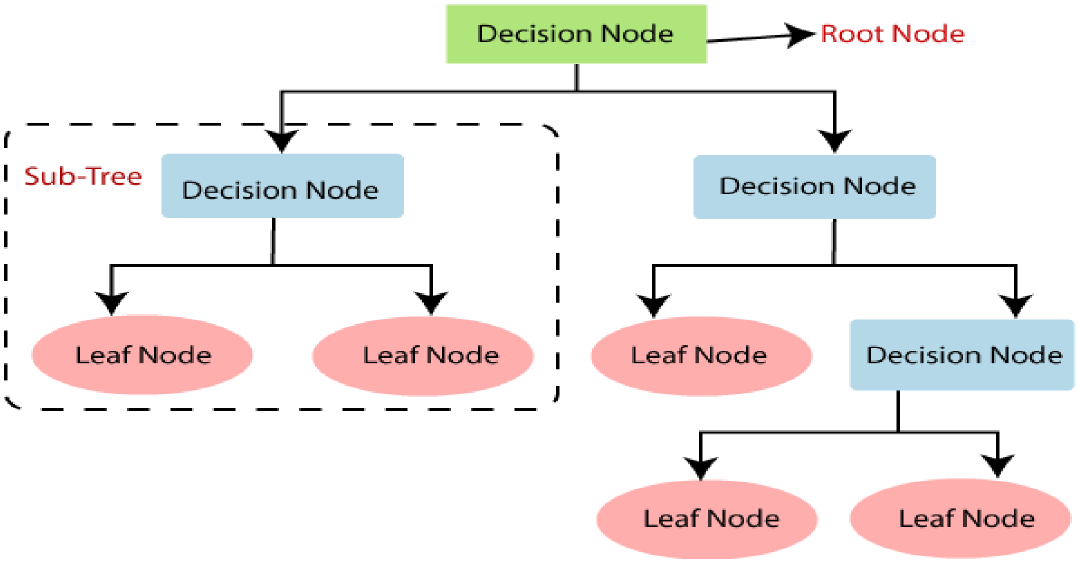


Figure 2.5: Decision Trees

4.4. Naïve Bayes

The Bayes naïf classifier is based on the Bayes theory. The Naive Bayes algorithm generates a hypothesis for a given set of classes, calculates the posterior probability of each class for a new unlabeled observation as indicated in formula 2, and selects the class with the highest posterior probability as a prediction. One of the major advantages of Naive Bayes is its simplicity and speed of calculation [21].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

where:

- P(A) is the probability of occurring the event A.
- P(B) is the probability of occurring the event B.
- P(A|B) is the conditional probability of event A occurring given that event B has already occurred.
- P(B|A) is the conditional probability of event B occurring given that event A has already occurred.
- P(A∩B) is the probability of both event A and event B occurring at the same time.

4.5. Logistic regression

Logistic regression is an automatic learning technique used for binary classification, predicting two separate classes. In this model, the example's characteristics are multiplied by coefficients to provide a continuous value known as score. The logistic function is used to the score to provide a probability value between 0 and 1 as we can see in Figure 2.6. The logistic regression has several advantages, including its simplicity, interpretability, and calculation speed. It is resilient to large datasets and can handle multi-class classification problems [22].

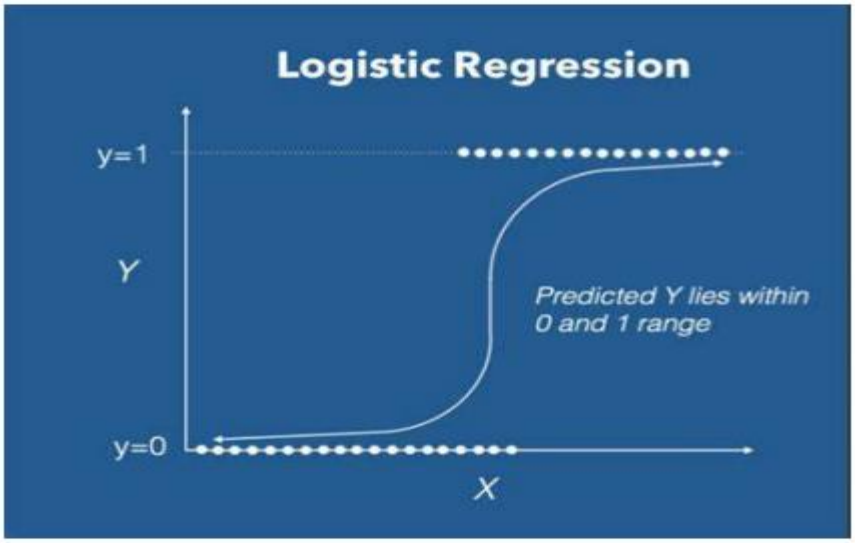


Figure 2.6: Logistic regression

4.6. Random Forest

Random Forest is an algorithm that combines the decision-making nature of multiple decision trees into one as we can see on Figure 2.7, allowing for high accuracy forecasting. It is a simple algorithm that can handle both categorical and numeric data without pre-processing tasks. It has functions like prediction mode (PM), a predictive mean-matching task that ranks key features, and can overcome bias in imbalanced data sets through oversampling or under sampling methods. Random Forest is widely used in finance, healthcare, marketing, manufacturing, credit, fraud detection, disease diagnosis, and customer segmentation [23].

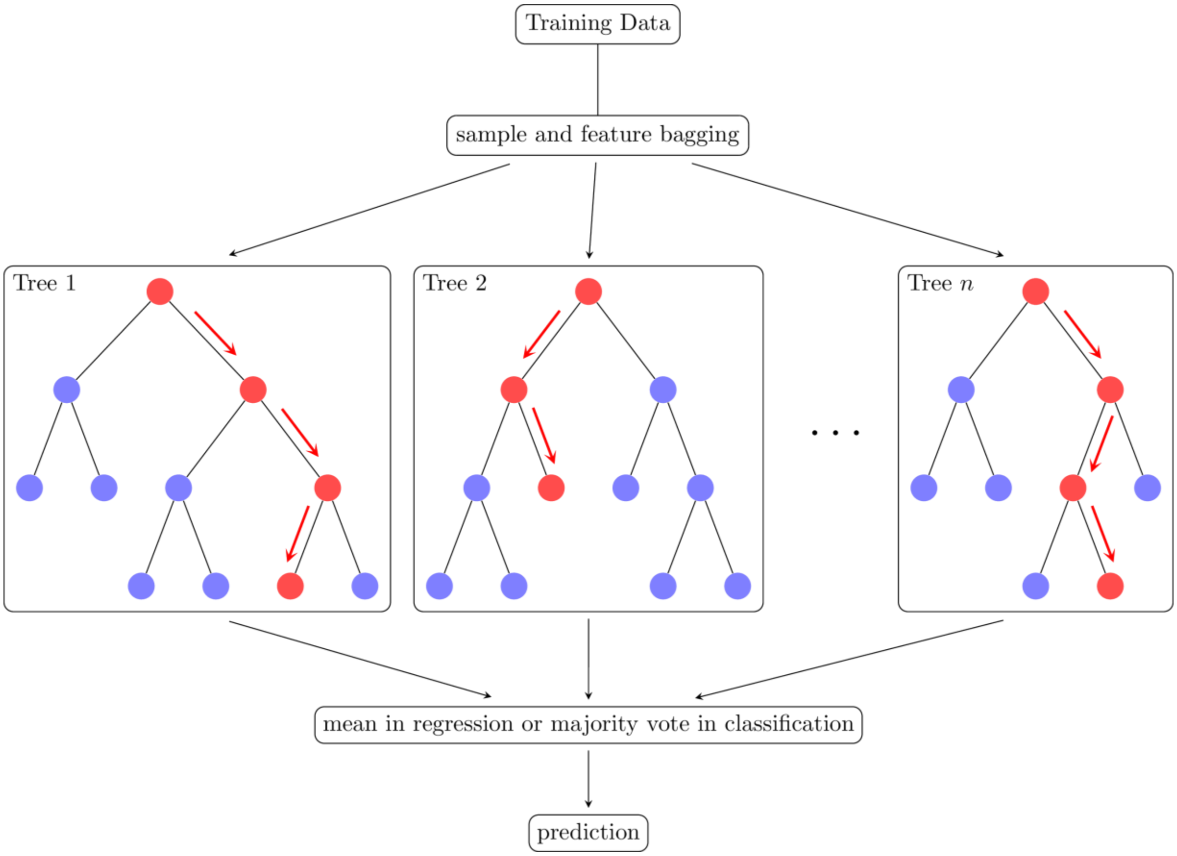

Training Data
sample and feature bagging
Tree 1
Tree 2
Tree n
. . .
mean in regression or majority vote in classification
prediction

Figure 2.7: Random Forest

5. Evaluation of Model Performance

Each model is being compared to the others, as table 2.1 illustrates, in order to determine which model is most effective in spotting fraudulent credit card transactions.

The total number of properly predicted occurrences is known as accuracy. Accuracy is displayed as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) in a confusion matrix. The term "True Positive" refers to transactions that the model properly identified as fraudulent.

The term "True Negative" denotes transactions that the model properly anticipated to be not fraudulent. False positive, the third rating, denotes transactions that are fraudulent but were mistakenly identified as non-fraudulent. Lastly, there is False Negative, which refers to the detected non-fraudulent transactions. The table below displays the confusion matrix, indicating fraud.

	Predicted Fraud	Predicted Non-Fraud
Actual Fraud	True Positive (TP)	False Negative (FN)
Actual Non-Fraud	False Positive (FP)	True Negative (TN)

Table 4.6.1:Model metrics evaluation

- True Positive (TP): Transactions that were recognized as fraud by the model to be correct.
- True Negative (TN): Transactions which were accurately flagged by the model not to be fraud.
- False Positive (FP): Fraud cases that had a score lower than the threshold and were recorded as non-fraudulent by the model but in reality were actually fraudulent.
- False Negative (FN): These include situations where the model had labeled specific transactions as fraudulent while they were actually non-fraudulent.

The overall accuracy of a model is calculated as:

$$\text{Accuracy} = \text{TP} + \text{TN} / \text{TP} + \text{TN} + \text{FP} + \text{FN}$$

Total accuracy of the classification is given by this formula, which shows the total number of correctly classified transactions divided by total number of transactions which include both fraudulent and non fraudulent transactions. By looking at the confusion matrices and accuracy of different models we identify which model is more effective in detecting credit card fraud. In other words, the highest accuracy level and the lowest False Positive and False Negative scores would be considered best. [24]

6. Conclusion:

To summarize, artificial intelligence and machine learning have transformed several sectors by allowing intelligent decision-making, pattern detection, and prediction. The development of sophisticated algorithms, as well as the availability of big datasets, has accelerated advancement in this sector.

As AI advances, it is critical to assess the effectiveness of machine learning models using proper metrics and methods. This involves evaluating the models' correctness, dependability, and resilience to ensure their usefulness in real-world applications.

While AI has enormous promise, it is critical to solve issues such as data bias, interpretability, and ethical concerns. Ongoing study and cooperation among specialists from diverse fields will be critical in propelling artificial intelligence forward and realizing its full potential for human benefit.

Chapter III: Credit Card Fraud Detection based on Machine Learning

1. Introduction

After defining the theoretical concepts of banking fraud and card fraud detection methods, we present in this chapter the project's goal, structure, and overall design. We will highlight each step by citing the main algorithms and approaches used in that step.

The aim of this study is to propose a method that enables banks to prevent losses resulting from theft and bank fraud based on K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Naive Bayes, and Logistic Regression classifiers.

2. Development tools and languages

Automatic learning is a field of artificial intelligence that allows computers to learn and improve without being explicitly programmed. It is used in a variety of applications, including image recognition, natural language processing, system recommendations, and fraud detection. To get familiar with automatic learning, it is necessary to master common tools and libraries. In this part, we will introduce two essential libraries for Python automatic learning: Pandas and scikit-learn.

2.1. Python

Python is a high-level object-oriented programming language used for web and application development. Python is a basic, easy-to-learn language that handles the use of modules and packages. Programs can be designed in a modular form, with code that can be reused in many projects. Once a module or package is developed, it may be scaled for use in other projects by utilizing import and export capabilities [25].

2.2. Scikit-learn

Scikit-learn is a Python package for automated and statistical learning. It offers a variety of supervised and unsupervised learning methods, as well as tools for data preparation, feature selection, and model validation. Scikit-learn is extensively used to create and test machine learning models in a range of domains [26]. The principal component of scikit-learn are:

- ☞ *Data preparation*: involves data normalization, scalability, and category codage.
- ☞ *Feature Selection*: Identify the key features for prediction.

☞ *Introducing Automatic Learning Models*: Linear regression, K-nearest neighbor classification, support vector machines, decision trees, and artificial neural networks.

☞ *Evaluation of Automatic Learning Models*: Calculate the accuracy, recall, ROC curve, and other metrics.

2.3. Anaconda

Anaconda is an open source Python and R distribution used for data research, automated learning, and deep learning with over 300 libraries. These libraries enable easy data collection from many sources using automated learning algorithms and AI. It helps to have an easily manageable environment setup that can deploy any project by clicking on a single button.

2.3.1. Pandas

Pandas is a Python package that facilitates data manipulation and analysis. She offers quick and simple data structures, such as Data Frames and Series, for storing and analyzing tabular data. Pandas is widely used for data cleaning and preprocessing, exploratory data analysis, and data visualization. Data may be read from many file formats, including CSV, Excel, and JSON and then can be used for:

- *Exploratory data analysis*: Generate descriptive statistics, graphs, and visualizations.
- *Data manipulation*: includes triage, filtering, and data aggregation. [27]

3. Proposed credit card fraud detection method

The proposed method, shown in figure 3.1, is composed from two main steps including: data analysis and classification steps. The first step involves specific operations on the dataset such as cleaning, normalization and filtering. After obtaining a new processed dataset, the second step consists of using, separately, six machine learning techniques including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Naive Bayes, and Logistic Regression to choose the best model.

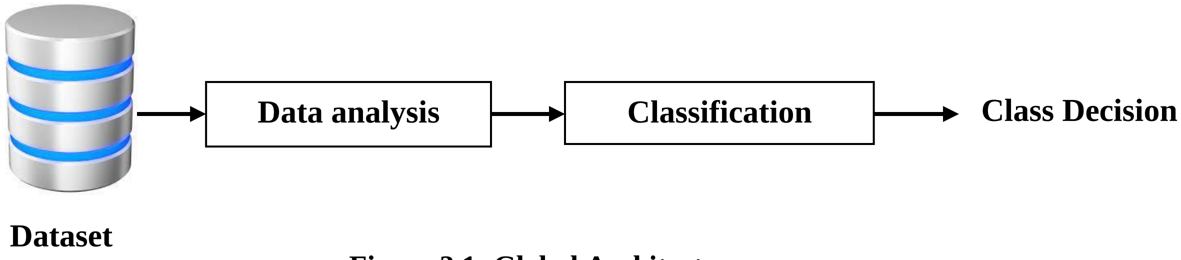
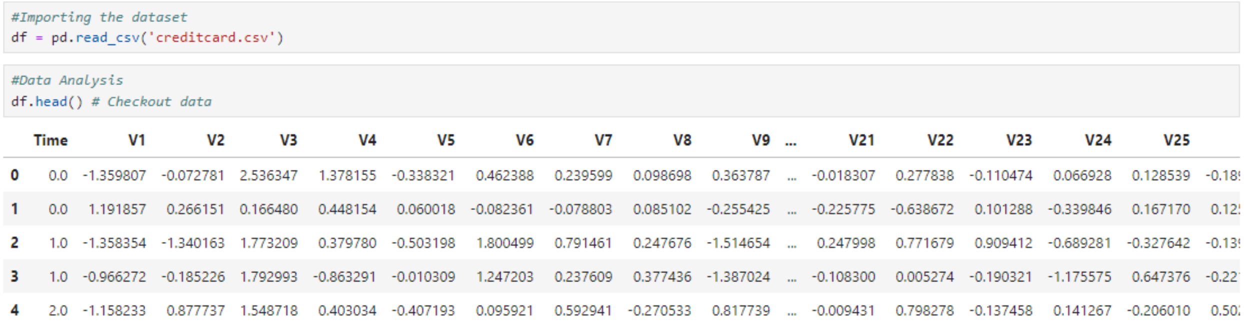


Figure 3.1: Global Architecture

3.1. Dataset Description

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions, where 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, the original features and more background information about the data was replaced. Features V_1, V_2, …, V_{28} are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value **1** in case of fraud and **0** otherwise.



```
#Importing the dataset
df = pd.read_csv('creditcard.csv')
```

```
#Data Analysis
df.head() # Checkout data
```

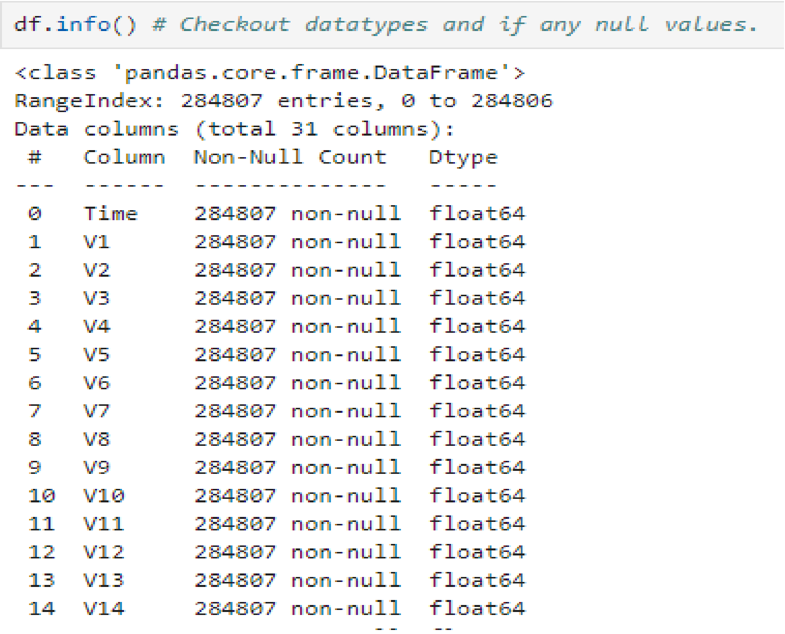
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.18
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.12
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.13
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.22
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.50

Figure 3.2: Dataset Description Code importing and checkout the data

3.2. Data Analysis

Missing data

In this step, the fist operation that we have processed is the verification of the missing data. As shown in figure 3.3, the code checks the data types of the columns in the dataset using the df.info () command. The output shows that the dataset has 31 columns, with 30 columns of data type float64 and 1 column of data type int64. Also, the output confirms that there are no null values in the dataset, as it states "284807 non-null" for each column.



```
df.info() # Checkout datatypes and if any null values.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
```

Figure 3.3: Data Analysis checkout Types of the Data

3.3. Histogram of Genuine and Fraudulent Transactions

In order to verify the distribution of transactions, histogram of genuine and fraudulent transactions are presented in Figure3.4 which we are categorizing as Class = 0, and the fraudulent transactions, which are categorized as Class = 1

.

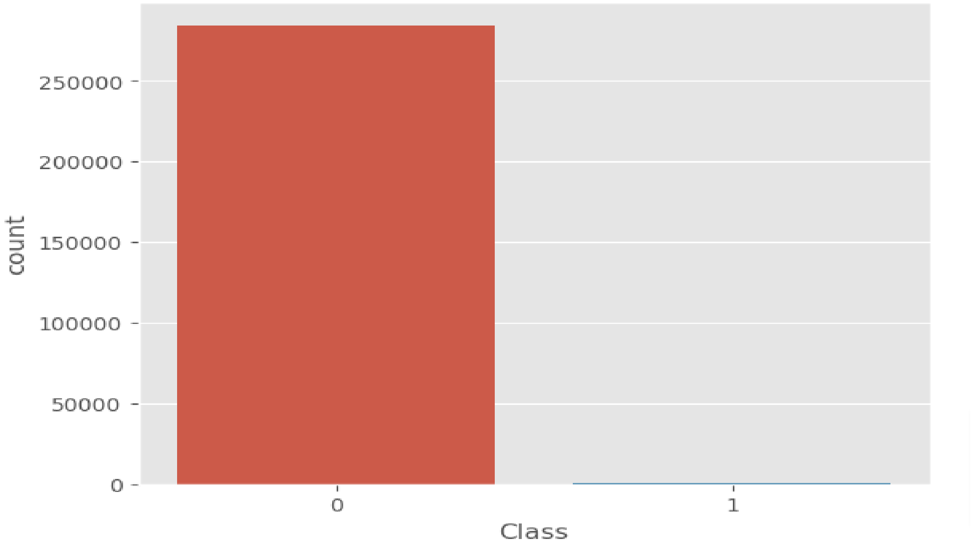
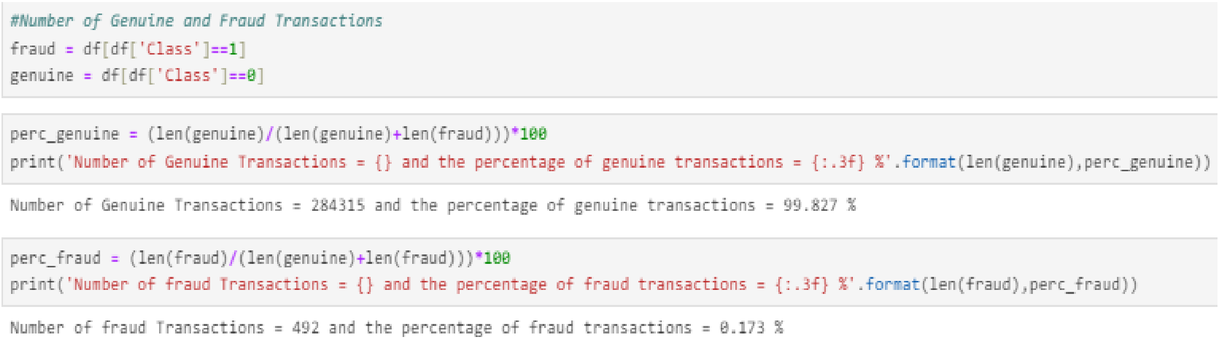


Figure 3.4: Histogram for Genuine and Fraudulent Transactions

From this figure3.4, we observe that there is a significant imbalance data, with a majority of transactions being non-fraudulent.



```
#Number of Genuine and Fraud Transactions
fraud = df[df['Class']==1]
genuine = df[df['Class']==0]
```

```
perc_genuine = (len(genuine)/(len(genuine)+len(fraud)))*100
print('Number of Genuine Transactions = {} and the percentage of genuine transactions = {:.3f} %'.format(len(genuine),perc_genuine))
```

```
Number of Genuine Transactions = 284315 and the percentage of genuine transactions = 99.827 %
```

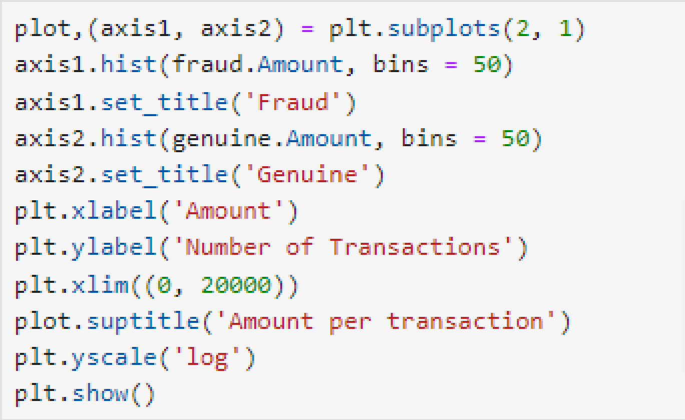
```
perc_fraud = (len(fraud)/(len(genuine)+len(fraud)))*100
print('Number of fraud Transactions = {} and the percentage of fraud transactions = {:.3f} %'.format(len(fraud),perc_fraud))
```

```
Number of fraud Transactions = 492 and the percentage of fraud transactions = 0.173 %
```

Figure 3.5: Number of genuine transactions and the percentage of genuine transactions

3.4. Scatter Plots for Time and Amount Features

For the purpose of visual analysis, scatter plots code was created to illustrate the role of ‘Time’ and ‘Amount’ of a particular dataset as it showing in Figure 3.6 This can be useful in showing whether there are regular patterns or time dependencies in the frequency or magnitudes of transactions, which might reflect corresponding patterns in other noticeable characteristics of those two intervals.



```
plot,(axis1, axis2) = plt.subplots(2, 1)
axis1.hist(fraud.Amount, bins = 50)
axis1.set_title('Fraud')
axis2.hist(genuine.Amount, bins = 50)
axis2.set_title('Genuine')
plt.xlabel('Amount')
plt.ylabel('Number of Transactions')
plt.xlim((0, 20000))
plot.suptitle('Amount per transaction')
plt.yscale('log')
plt.show()
```

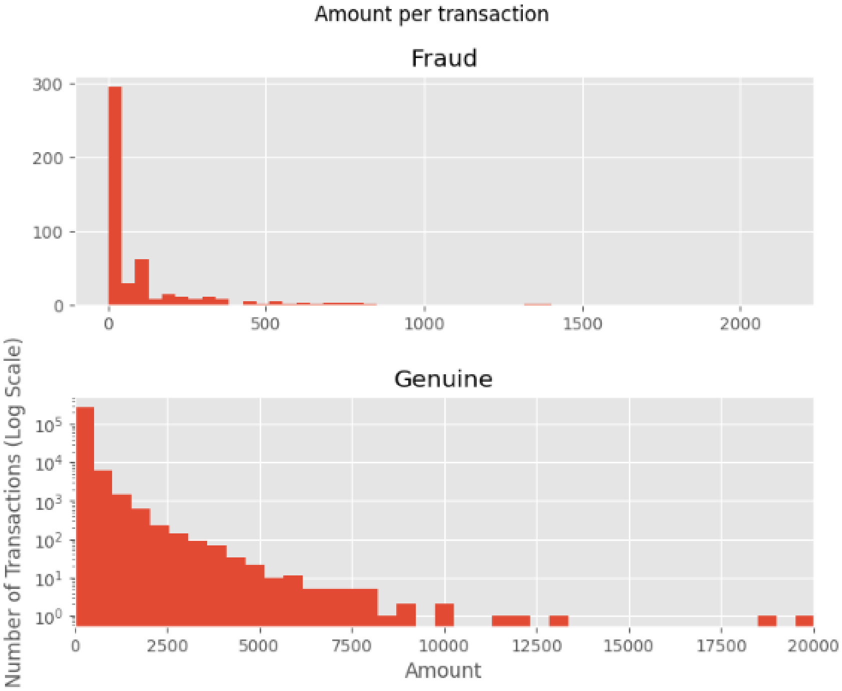


Figure 3.6: Fraud and genuine (with log scale) transactions histogram

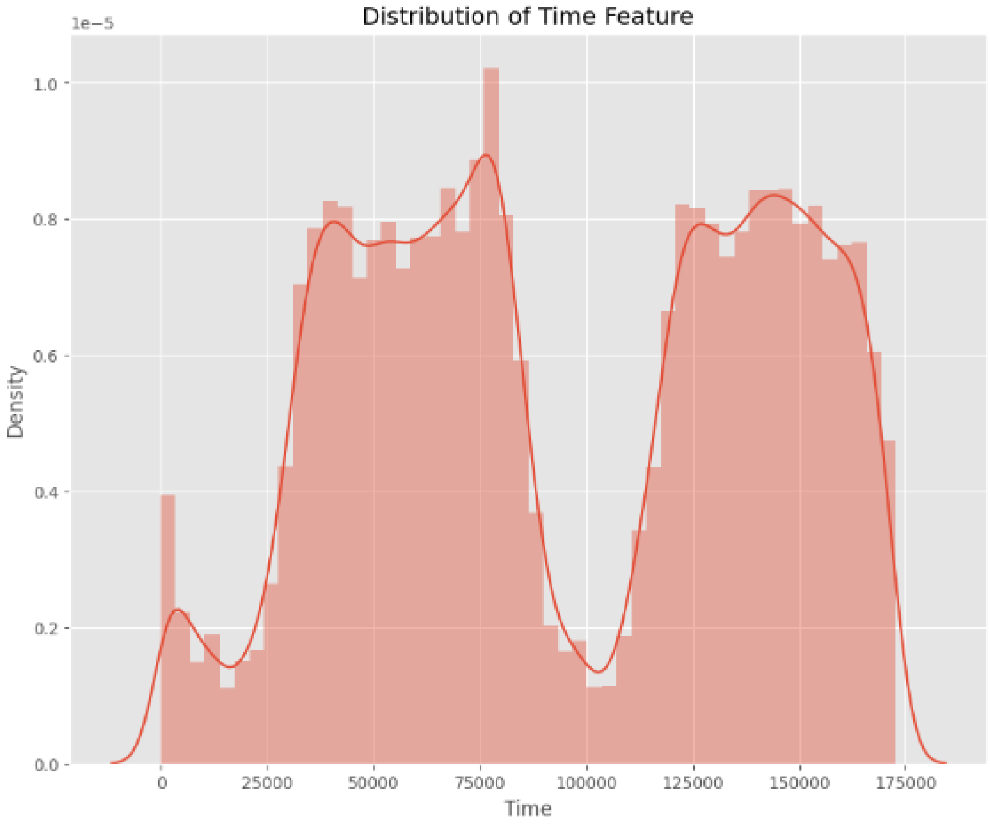
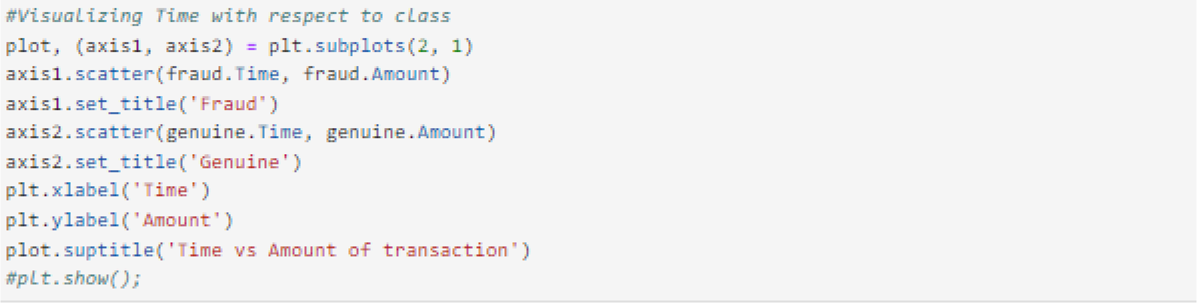


Figure 3.7: Histogram of Time of feature using Density Plot

Visualization of fraudulent transactions plot was generation code using the following code.



```
#Visualizing Time with respect to class
plot, (axis1, axis2) = plt.subplots(2, 1)
axis1.scatter(fraud.Time, fraud.Amount)
axis1.set_title('Fraud')
axis2.scatter(genuine.Time, genuine.Amount)
axis2.set_title('Genuine')
plt.xlabel('Time')
plt.ylabel('Amount')
plot.suptitle('Time vs Amount of transaction')
#plt.show();
```

Figure 3.8: Code of visulazing Time with respect to Class

Figure 3.8 shown the fraudulent transactions (at the top) and the legitimate transactions (at the bottom), fraudulent transactions typically involve larger sums of money. From this feature, we observe that, the data on fraudulent transactions is more dispersed, indicating that fraudulent transactions can happen for any amount of money and at any time of day. On the other hand, legitimate transactions often involve smaller sums of money and tend to occur at specific times of the day.

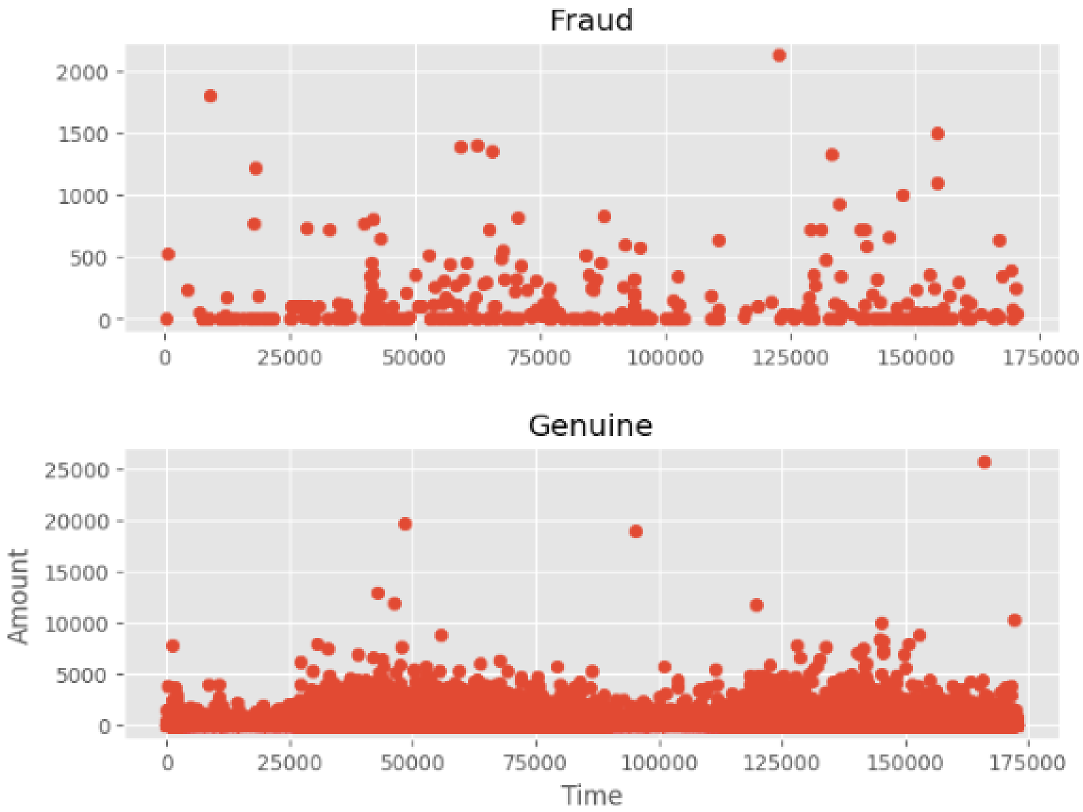


Figure 3.9: Time vs Amount of transaction histogram

3.5. Correlation Matrices for Feature Relationships

Figures 3.10 present the correlation hetamap which will help us to understand the correlation and the relationship that exists between the various features. It is important for the feature selection and engineering steps, as it determines areas where more modeling can be done in order to identify important relationships in features.

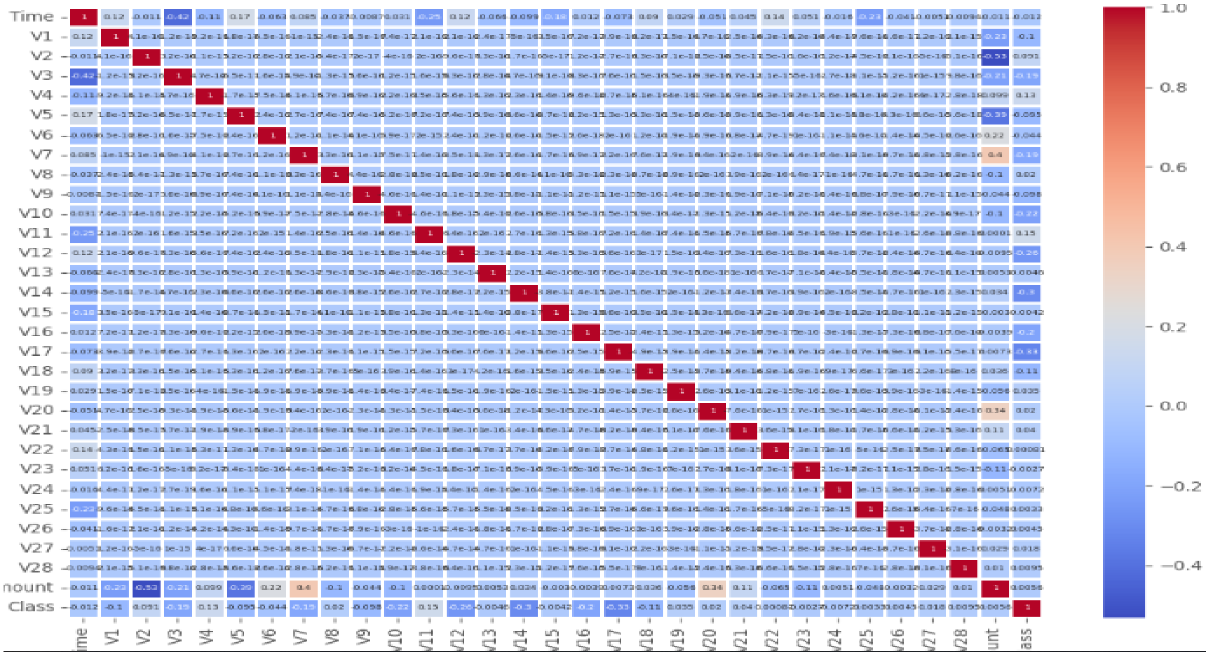
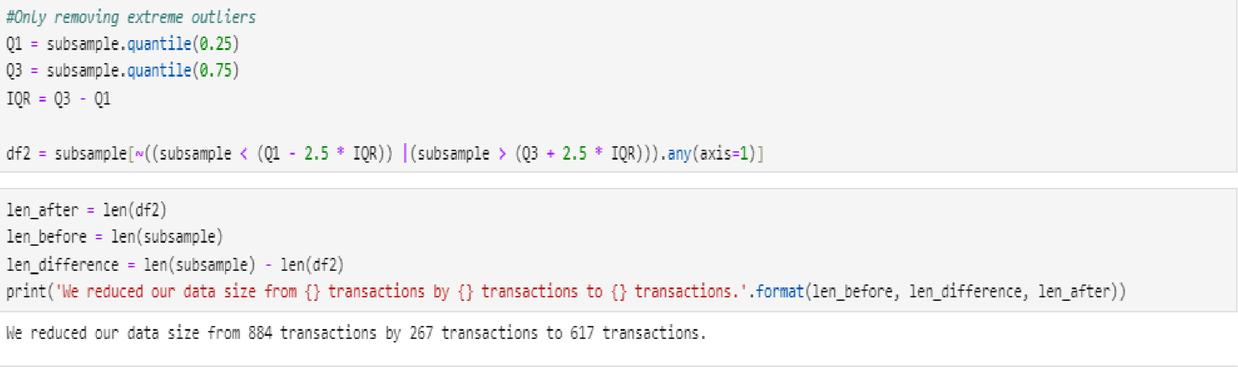


Figure 3.10: Correlation heatmap

3.6. Outliers values identification and elimination

In order to identify the outliers, the following was generated to identify and eliminate extreme values that might skew the data.



```
#Only removing extreme outliers
Q1 = subsample.quantile(0.25)
Q3 = subsample.quantile(0.75)
IQR = Q3 - Q1

df2 = subsample[~((subsample < (Q1 - 2.5 * IQR)) |(subsample > (Q3 + 2.5 * IQR))).any(axis=1)]
```

```
len_after = len(df2)
len_before = len(subsample)
len_difference = len(subsample) - len(df2)
print('We reduced our data size from {} transactions by {} transactions to {} transactions.'.format(len_before, len_difference, len_after))
```

```
We reduced our data size from 884 transactions by 267 transactions to 617 transactions.
```

Figure 3.11:Code of Outliers values identification and elimination

In our study, the outlier values identification and elimination can be summarized as follows:

1. *Finding Quartiles:* First, the code computes the first quartile (Q1) and the third quartile (Q3) of the encompassed data. These denote the upper and lower quartiles which are the 75th and 25th percentiles of the data, respectively.

2. *Calculating IQR*: The general formula for the inter-quartile range (IQR) is the difference between the third quartile (Q3) and the first quartile (Q1). This can basically be summarized as the range, between the middle 50% of the data.
3. *Identifying Outliers*: Data points that are two or more standard deviations away from the overall mean are considered outliers. Identifying lower outliers and upper outliers, the values five times the IQR below Q1 or above Q3 are included for removal. The threshold is 2. The most frequent multiplier used for calculating IQR is 1.5, but this multiplier can be varied depending on the type of application.
4. *Filtering and Counting*: The code deletes rows that have values that are too high or too low relative to other columns and writes cleaned data to a new Data Frame. It also computes the number of outliers removed in case it may be useful for any purpose for explanation of the values.

3.7. Balancing the Dataset

In this work, we have employed the SMOTE method (Synthetic Minority Over-sampling Technique) to address class imbalance in our dataset. SMOTE enhances the representation of the minority class by generating synthetic samples. The primary steps involved in balancing the dataset using SMOTE are as follows:

1. *Identify Neighbors*: For each sample in the minority class, SMOTE identifies the k-nearest neighbors (typically k=5).
2. *Select a Neighbor*: A random neighbor is selected from these k-nearest neighbors.
3. *Generate Synthetic Samples*: A synthetic sample is created by interpolating between the feature values of the original sample and the selected neighbor.
4. *Repeat the Process:* This process is repeated until the minority class is sufficiently oversampled to achieve the desired balance.

3.8. Classification step

In the literature review, various algorithms have been applied to fraud detection, each possessing unique strengths and weaknesses. In this study, we have utilized six machine learning techniques: Logistic Regression, Decision Tree, KNN, SVM, Gaussian Naive Bayes, and Random Forest.

3.9. Experiments results

After preparing the data for modeling, six models were created using the scikit-learn package. The data was split into 80% for training and 20% for testing. Table 3 displays the performance of each classifier.

Model	Accuracy
Logistic Regression	**97,57%**
KNN	**95,74%**
Decision Tree	**89,04%**
SVM	**97,36%**
Naive Bayes	**95,98%**
Random Forest	**97,34%**

Table 3.9.1:classifier performances

Based on Table 3.1, the most accurate models are Logistic Regression with an accuracy of 97.57%, SVM with an accuracy of 97.36%, and Random Forest Classifier with an accuracy of 97.34%. These models excel at detecting fraudulent transactions in this dataset.

In comparison, KNN Classifier and Naive Bayes show slightly lower performance, with accuracies of 95.74% and 95.98%, respectively. The Decision Tree Classifier has the lowest performance, with an accuracy of 89.04%, indicating it is less suitable for this binary classification problem.

4. Conclusion

This chapter underscores the importance of data analysis in achieving optimal results. It covers various data visualizations, including histograms and scatter plots, to examine data distribution and detect anomalies that may impact classifier performance.

Additionally, the study demonstrates that the chosen model effectively identifies fraudulent transactions, offering a robust solution for banks to mitigate losses from theft and fraud.

General Conclusion

General Conclusion

The study conducted in this manuscript is supportive of the fact that machine learning holds great promise as well as immediate applicability in credit card fraud detection. In this study, several machine learning algorithms, the Logistic Regression algorithm, Decision Trees, Random Forest, and Naïve Bayes were used to assess the performance of the model in detecting fraudulent transactions.

In order to make a comprehensive assessment of the models, metric such as accuracy were used. The findings showed that the models such as the Logistic Regression model as well as the support vector machine model and the Random forest model had the highest outcome of accuracy for fraudulent transaction detection.

However, it was not without a problem, and the research encountered some challenges, specifically, the class unbalance in fraud detection datasets raised serious challenges regarding the model performance. Although it was possible to have partial solutions using oversampling and ensemble methods, we also require future investigation to propose more complex solutions for imbalanced data. Moreover, the use of historical data may not be enough to incorporate the new emerging fraud types. More studies should be aimed at the development of self-evolving programs that can take into consideration new fraud patterns as they emerge in an effort to enhance the efficacy of fraud fighting tools.

There are several approaches to enhance the model, including applying it to other datasets of varying sizes and data kinds, adjusting the data splitting ratio, and examining it from different algorithm perspectives.

Bibliography

[1] study.com(https://study.com/academy/lesson/what-is-bank-fraud-definition-prevention.html) : 04/05/2021.

[2] Bank Cards: Membership of the credit card payment system cb "merchant contract" version of self-service payment machine, February 2007

[3] Bank Cards: Card proximity payment acceptance contract
"cb" or approved "cb", November 2009.

[4] Bank Cards: Interbank commissions: the competition authority accepts the commitments proposed by the credit card group cb. Press release, July 2011.

[5] https://www.avg.com/fr/signal/identity-theft, 04/05/2021.

[6] Berry law (https://jsberrylaw.com/blog/bank-fraud-definition-penalties/) : 04/05/2021.

[7] Difs (https://www.michigan.gov/difs/0,5269,7-303–458212–,00.html). 04/05/2021.

[8] Investopedia (https://www.investopedia.com/terms/m/moneylaundering.asp). 04/05/2021.

[9] National check fraud center. 1995/2011.

[10] Sequential pattern mining (https://www.cc.gatech.edu/hic/cs7616/pdf/lecture13.pdf). 2013.

[11] Shambhavi. (2023, July 1). SONY PLAYSTATION NETWORK CASE STUDY- DIGITAL FORENSIC.Medium.https://medium.com/@shambhavi1408/sony-playstation-network-case-study-digital-forensic-b6367451e6d1

[12] S. Ravi • J. Thanga Kumar• Dr. Linda Joseph • Sumanth Raju Kunjeti•(2021/02/08),9.4. An Unique Methodology for Credit Card Fraud Detection based on Convolutional Neural Network

[13] https://web.archive.org/web/20210716142243id_/http:/alinteridergisi.com/wp-content/uploads/2021/05/AJAS21041-2.pdf

[14] Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. https://www.semanticscholar.org/paper/Sequence-classification-for-credit-card-fraud-Jurgovsky-Granitzer/63055b52bdbca928cace9874f540a5bfaab5d5c1

[15] Fighting Credit Card Fraud Using Machine Learning. (2021). Retrieved from https://www.researchgate.net/publication/380534765_Fighting_Credit_Card_fraud_using_machine_learning

[16] Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. Procedia Computer Science, 165, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

[17] Dr. Abdelhamid Djeffal. Introduction aux données séquentielles (Master 2 Informatique de l'optimisation et de la décision). 2020/2021.

[18] International Conference on Computer Networks and Computing (ICCNI). https://doi.org/10.1109/iccni.2017.8123782

[19] Dheepa, V. and Dhanapal, R. (2012). Behavior based credit card fraud detection using support vector machines. ICTACT Journal on Soft Computing, 02(04), 391-397.

[20] Decision Tree Analysis. Mind Tools – Essential skills for excellent career. [consulté le 10 janvier 2012]. Disponible sur : http://www.mindtools.com/dectree.html.

[21] Credit card statistics. Change credit card processing. (August 30, 2021). Retrieved https://shiftprocessing.com/credit-card/

[22] Alenzi, H. Z. and Aljehane, N. O. (2020). Fraud detection in credit cards using logistic regression. International Journal of Advanced Computing and Applications,

[23] Ho, T. K. Random Decision Forests. In Proceedings of the 12th International Conference on Machine Learning, Morgan Kaufmann, 1995.

[24] D. Abdelhamid. Cours de classification. Université de Biskra, 2019.

[25] Le langage de programmation Python. 2019.

[26] Jupyter Notebook (Project Jupyter | Home). 24/06/2021.

[27] Apriori algorithm in data mining: Implementation with examples. May 30, 2021.