Detecting Fraudulent Transactions in Ethereum Blockchain via Machine Learning Classification

MSc Research Project
Data Analytics

Camila da Silva Weber
Student ID: x20166371

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Camila da Silva Weber……

Student ID: x20166371……

Programme: Data Analytics…… **Programme:** Data Analytics……

Module: MSc Research Project……

Supervisor: Vladimir Milosavljevic……

Submission Due Date: 14/08/2023……

Project Title: Detecting Fraudulent Transactions in Ethereum Blockchain via Machine Learning Classification ……

Word Count: 5868☐ …… **Page Count**…20……

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

Signature: ……

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

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

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

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

Detecting Fraudulent Transactions in Ethereum Blockchain via Machine Learning Classification

Camila da Silva Weber
x20166371

Abstract

Blockchain is a network that has grown exponentially in recent years, it is a decentralized technology and for this reason it requires the development of secure applications and a better understanding of this networking. The project focuses on creating a classification model with a secure approach towards the Ethereum network, as it is the second most important cryptocurrency in the Blockchain domain. Implementing Machine Learning methods, the aim of the project is to investigate and detect possible fraud within Ethereum. Exploring the concepts of classification methods in Machine Learning, and developing important models that reinforce security in these decentralized networks. By analysing the data and answering the research question of the project, Machine Learning models were applied such as the Support Vector Machine, Logistic Regression and Random Forest for effective detection of fraud in transactions on the Ethereum network. This research still presents significant studies and methodologies that lead insights for future projects.

1 Introduction

The Blockchain technology emerged in 2009, that would represent a great technological advance with its distributed and decentralized approach, eliminating the need for a central authority in financial transactions. The popularity of this technology has grown over ther years and specially after the introduction of the cryptocurrencies, that is a type of decentralized digital currency that uses cryptography to ensure the security of transactions, control the creation of new units and verify the transfer of assets, currently the best known cryptocurrency is Bitcoin, resulting in significant investments from various sectors of the financial market. Thus, the reliability and quality of operations are fundamental since Blockchain technology is decentralized.

The project will be focused on the Blockchain technology approach, especially in the context of the Ethereum network. It was founded by Vitalik Buterin and Gavin Wood in 2015, and its network is known for the use of smart contracts and its the second most popular cryptocurrency technology after Bitcoin. The project will be concentrating in the fraud detection in transactions for the Ethereum network, since there are not many studies based on fraud in this network. The objective is to develop analytical models to detect fraud in the Ethereum network, seeking important insights to achieve good results. Thus, analytical methods such as Business Intelligence and Data Science are important, since these analytical

methods will be providing a comprehensive understanding of the business problem and facilitating the identification of suitable solutions. Business Intelligence will be concentrating on collecting data, to be able to understand the context and business problem, in addition to identifying trends and optimizing the performance of different analytical methods, while Data Science employs statistical analysis and Machine Learning techniques to discover relevant insights. Both approaches complement each other, with Business Intelligence integrating information for a complete view of data and Data Science ensuring data quality and accuracy through pre-processing, thus increasing project efficiency.

Although fraud detection and prevention techniques have been extensively studied in traditional centralized systems, the application of machine learning methods adapted to the unique characteristics of the Ethereum network is considered a relatively new approach.

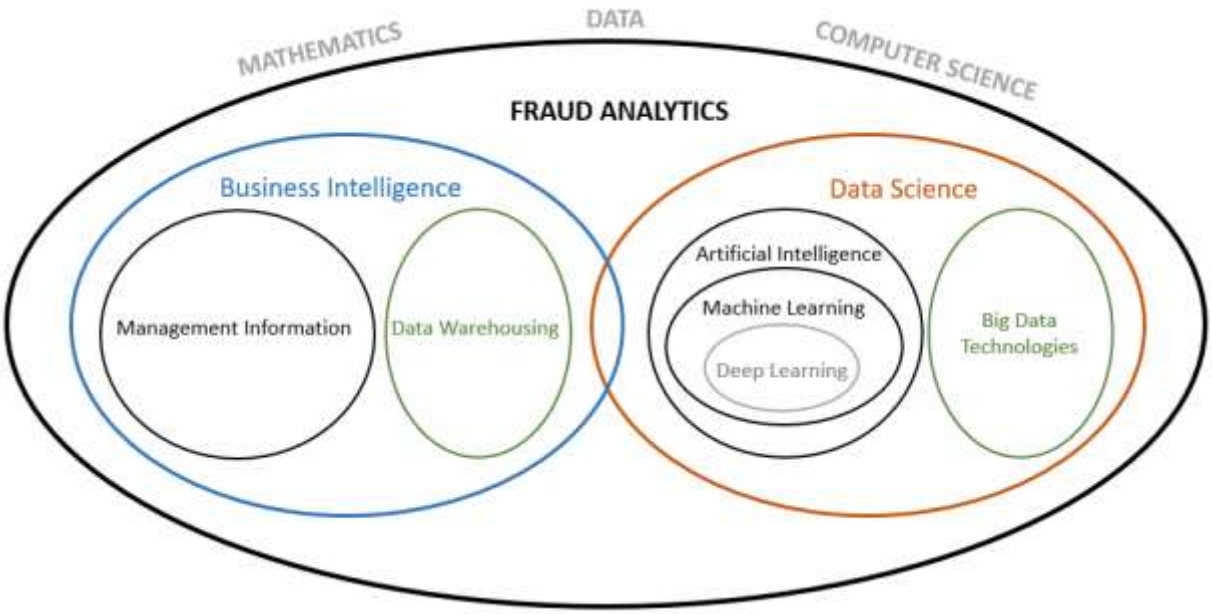


Figure 1: Fraud Analytics

The aim of this project is to explore and develop innovative machine learning algorithms and models specially designed to detect and prevent fraudulent transactions on the Ethereum network. It is important to explore the data applying analytical and Machine Learning methods to effectively analyse transactions and token patterns based on standardized operating parameters of ERC-20 tokens in order to identify potential fraudulent activities. The Ethereum Request for Comment 20 (ERC-20) protocol is especially relevant in the context of the Ethereum network, as it allows the creation of fungible tokens with specific operational standards, enabling the exchange of tokens through smart contracts. The Data collection for the project is from publicly available sources on the Kaggle website.[1] After data analysis and pre-processing, Machine Learning classification methods will be applied, to

[1] Kaggle Dataset: https://www.kaggle.com/datasets/vagifa/ethereum-frauddetection-dataset

gain relevant insights and provide appropriate solutions for transaction analysis, enabling anomaly detection and valid transactions.

RQ: How Machine Learning techniques can increase security for decentralized Blockchain networks such as cryptocurrency Ethereum, preventing fraud in transactions using classification models?

The project is important for focusing on blockchain networks and machine learning models specific to the Ethereum network. This analysis being relevant, which can be applied in future projects, aiming to identify effective solutions for detecting fraudulent transactions on the Ethereum network, thus using models to reduce possible fraud that cause revenue losses for users and organizations that use the Blockchain platform. The project seeks to obtain valuable information and solutions following the methodology (CRISP-DM). By effectively organizing the information and evaluating the results, this project aims to find important insights and solutions to the problem of fraud within the Ethereum Network. The project uses Classification Machine Learning models to search for features capable of identifying correlations and improving the accuracy of training models. The objective is the more efficient detection of fraud in transactions, providing greater security and confidence to users of the Ethereum network.

2 Related Work

The initial section of this research proposal provides an overview of the background and definition of Blockchain technology, along with its key implications. The Literature Review section is divided into two subsections. Subsection 2.1 presents significant studies focused on Malicious Transactions within the Blockchain. Subsection 2.2 encompasses studies that explore the application of Machine Learning methods for Fraud detection in the Blockchain network, as well as other investigations involving the Ethereum cryptocurrency.

2.1 Malicious Transactions in Blockchain Technology

Blockchain technology is a technology that facilitates transactions but also there is some concerns about its security, and like other financial systems, cryptocurrencies have become vulnerable to fraud in their transactions, requiring research aimed at detecting these fraudulent activities. The analysis and investigation of transactions within the Blockchain network present significant challenges due to the dynamic nature of the system and the volume of data involved. In this context, some related research will be presented to provide support and background to the project.

Jung et al. (2019) explained that cryptocurrencies operate in a decentralized structure, that means its a technology that is characterized by the absence of a central authority such as banks and conventional financial entities. Thus, this characteristic of decentralization makes it more difficult to identify in cases of anomalies and fraudulent activities, making it difficult to track those responsible for possible fraud in transactions and illicit activities. Bartoletti et al. (2018) analyzed a large number of smart contracts within the Ponzi scheme which was

operated on the Ethereum network, this research used similarities between contract bytecodes in the identification and location of 191 Ponzi schemes and these contracts collectively accumulated almost half a million dollars from more than 2,000 individual users. In this study, the authors highlighted some distinct characteristics that were identified in the Ponzi scheme, with emphasis on a high Gini coefficient as an important indicator in the analysis. Since the Gini coefficient is a measure used to assess the inequality of distribution of financial capital made by contracts for investors, that provides information about smart contracts within the Ethereum network and possible fraudulent behavior. The authors also drew on previous research with the high Gini coefficient and furthermore applied other data mining techniques to aid in the detection of Ponzi schemes on the Bitcoin network. Using resources based on the lifetime of a contract, their classifier successfully detected 31 out of 32 Ponzi schemes with a low 1% false positive rate. In this study, it is possible to notice that there are limitations regarding the analysis based on contract characteristics in the detection of Ponzi schemes, which indicates limitations regarding safer measures against these schemes. As tokens are not refunded or flagged as fraudulent within Blockchain transactions, other more efficient analysis methods are needed to mitigate the impact of Ponzi schemes effectively. Thus, emphasizing the importance of new research, innovations and strategies to combat fraudulent activities on the network.

Pham and Lee (2016) conducted an approach to the detection of anomalies in their study, identifying suspicious users and transactions in the Bitcoin network. Since the purpose of this study is to evaluate transactions that align with a group of previously identified malicious transactions. Using such suspicious activities as proxies, the research sought to establish a possible connection between such actions and activities on the Ethereum network that are potentially fraudulent. The evaluation of the approach occurs through the comparison of the results obtained with a set of already known malicious transactions, verifying if the standard method can effectively identify these transactions and, thus, verify its effectiveness in detecting malicious activities. Muller et al. (2015) emphasize that Blockchain-based cryptocurrencies have certain vulnerabilities in terms of their security, which can be identified as illicit activities carried out on the network and facilitating access by potential hackers. Unlike more conventional frauds, such as credit card fraud, cryptocurrencies and the Blockchain network in general do not have an official standard registration process, as it is a decentralized technology and does not have the supervision of a financial entity. Consequently, when fraud takes place, obtaining a refund becomes difficult. Since the manual analysis of possible transactions in search of irregular characteristics is challenging and can lead to failures in mitigating fraudulent activities.

Thus, it is necessary to look for other alternatives such as analysis using other methods such as Machine Learning in the prevention of these fraudulent transactions. In the work by Harlev et al. (2018), techniques known as machine learning are employed in a supervised manner with the aim of reducing the level of anonymity within the Bitcoin Blockchain network. The study was based on pre-processed data in which addresses were manually grouped and labelled based on their behavior patterns. Even though it is advantageous to use pre-processed data in some scenarios, this data can also be a challenge in another approach in the Blockchain network, since a quality dataset and precisely labeled is necessary for the success of machine learning projects applied to analysis.

2.2 Malicious Transactions in Blockchain Technology

Initially Blockchain technology was more focused on cryptocurrencies and electronic payments, this technology was gaining more space and creating new applications, one of the new applications being smart contracts, which are self-executing programs operating within the Blockchain network, being a fundamental part of the Ethereum network. Since contracts operate without intermediaries, it is necessary to seek new measures and solutions to protect the integrity and reliability of the network. In their research, Ibrahim et al. (2021) used supervised and ensemble learning methods to detect illicit transactions on the Ethereum network. Using Machine Learning models such as the random forest and decision trees looking for good insights in your result. In the construction of the dataset, features based on the correlation coefficient were used to train the model and predict the accuracy of transactions. This combination of supervised and joint learning techniques allowed the creation of an effective model in identifying illicit transactions in the analyzed dataset. Using feature extraction driven by correlation coefficients has proven to be a good strategy in identifying significant factors that influence illicit activities within transactions. This study provides promising information for future projects to increase the effectiveness of methods to detect and prevent future fraud within the Blockchain network.

In their research Ajay et al. (2020) used Machine Learning methods to detect anomalies in the Ethereum cryptocurrency network. Employing techniques such as Decision Trees, achieving an accuracy of 83.7%, and implementing Random Forest algorithms, resulting in an accuracy of 98.9%. Being classification methods in the area of Machine Learning and presenting excellent performance and great potential in identifying fraudulent activities in the Ethereum network. Cryptcurrencies are known to be good investments, but being their decentralized nature, many users venture into the crypto market without the necessary knowledge, thus being able to have revenue losses and many frauds are still being applied, in this way new technologies must be employed to mitigate these possible security problems. Machine Learning models are an effective way to look for models and methods to protect users and the network itself from possible fraud.

Chen et al. (2019) was inspired by the work of Bartoletti et al. (2018) extending their search and incorporating additional features of account data and Opcode. Introducing the features of Opcode, which are contract codes that are stored on the Blockchain network. The Machine Learning used in the research was XGBoost being a classification algorithm, thus three distinct models were created, which were based on account, on opcode and a combined model of account + opcode. The Account+Opcode model outperformed the other two, achieving impressive results with an accuracy rate of 94% and a recall rate of 81%. Where high accuracy and recovery values indicate the model's ability to effectively detect Ponzi schemes implemented in the Bitcoin network. Since the incorporation of Account and Opcode resources allowed a more comprehensive analysis where the analysis with Machine Learning had a good performance and good accuracy in identifying fraud associated with the Ponzi scheme. A new supervised learning approach was presented by Pham and Lee (2016) for identifying anomalies within the Bitcoin network. This method aims to identify which transactions are suspicious and label which users have some involvement in these transactions. The main technique used in this study was the K-means clustering algorithm,

and the authors analyzed 30 cases in the Bitcoin network known to include cases of fraud and revenue losses. However, even analyzing it in a complete way, the accuracy of the proposed model was only 10%. Being a preliminary model, since only the K-means method was used in the study, the performance was limited in detecting possible fraudulent transactions. Being a complex task, other researches are necessary for the improvement of new analysis models and new more precise approaches in the detection of anomalies and also of labels to users of the Bitcoin network. Brown et al (2018) introduced the application of recurrent neural networks (RNNs) for detecting anomalies in system logs which are presented in the form of sequential data, thus unstructured sequential log messages were analyzed and the result was promising compared to other previously applied methodologies. Furthermore, LSTM (Long Short-Term Memory) neural networks were used in malware detection using opcode sequences. This study is efficient in malware detection, especially in scenarios where behavior-based predictions require a safe runtime environment.

Tan et al. (2021) proposed in their research a different approach based on the detection of ambiguous transactions using a convolutional graph network model (GCN). Being a very advanced neural network architecture, the objective when using this model was the identification and differentiation of ambiguous transactions in the analyzed data set. The model performed excellent results with an accuracy rate of 95% in detecting ambiguous transactions. The authors sought the necessary data for the study using web trackers and capturing addresses associated with fraudulent transactions. This data set was used as a resource in the training and validation of the GCN model, which effectively distinguished between ambiguous and legitimate transactions. Highlighting the potential of these models and seeking to improve such mechanisms in the detection of anomalies and fraud in transactions. Vulnerability detection is crucial to prevent potential attacks and ensure the integrity of operations within the Blockchain network.

3 Research Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was applied in the project, as it is a widely recognized and effective for conducting studies in the area of data analytics, providing a structure that guides decision for the project by applying the steps of this methodology that includes data understanding, data preparation, modelling, evaluation and implementation.

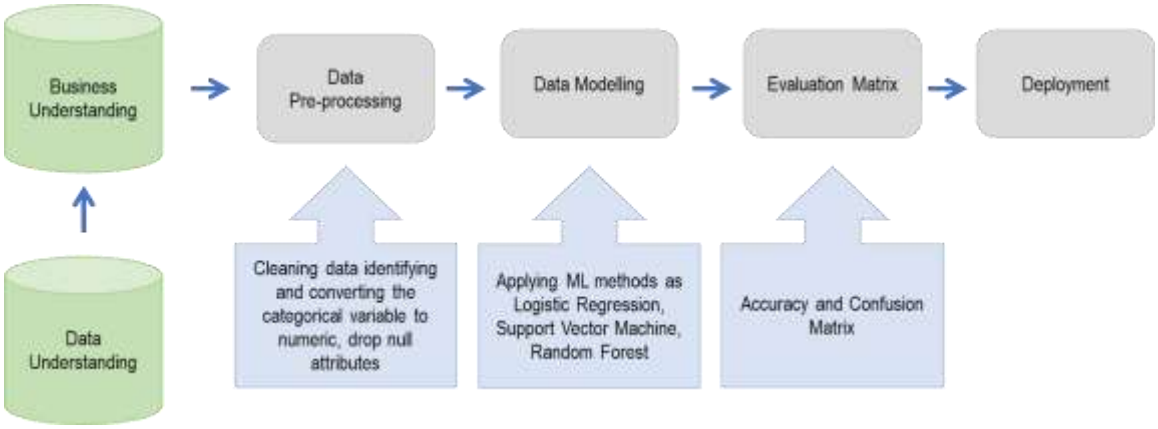


Figure 2: Methodology Approach

Implementing a systematic approach such as CRISP-DM is important since it is an organized methodology that brings better insights and ensures the reliability of the results, applying steps within the methodologies and building the result and effectiveness of the model. The understanding of the business is crucial referring to the security of decentralized networks and mainly in the detection of fraudulent activities in the Ethereum network since there are not many studies based on this network. Therefore, this study is necessary to maintain trust and integrity within the Ethereum network. The CRISP-DM methodology has an important approach that supports the business context with its structure, organizing and formulating other objectives that are focused regarding data analysis and fraud detection. The business understanding phase is essential to solve the proposed problems, applying data mining is an advantage in the efficiency of the project results. It also provides other important perspectives such as new opportunities within the project, with the understanding of the problem and the context of the business, seeking effective solutions with a focus on the protection and reliability of the Ethereum network.

The Data Understanding phase within the CRISP-DM methodology is a crucial preparatory stage that establishes a comprehensive understanding of the dataset on the Ethereum. With the support of the methodology and the exploration of the data, it is possible to understand the information and establish relevant resources regarding the construction of effective models in the detection of frauds that contribute to the security and reliability of the Ethereum network. In this step, the dataset is collected and investigated, the dataset used in the project is public and was collected from the Kaggle website, the dataset consists of 9841 rows and 51 columns, with data from fraudulent transactions and valid transactions.

For the preparation of the data, the ETL (Extract, Transform, Load) will be applied, as the cleaning of the data is essential for a better understanding of the data, analysing and understanding what information are valid or not for the study, and also preparing the data that are important to receive the machine learning techniques that will be proposed. After preparing the data the machine learning methods will be used, it is important to apply the conversion of variables where the model can be executed without major problems or interruptions. When the data is pre-processed, it is possible to have a better detection of frauds and the data becomes more reliable. After cleaning and pre-processing, this data set will be divided into other subsets, defined as training, validation and testing so that a better accuracy in the performance of the model is possible. As for data modelling, binary classification will be used, where it will be focused on identifying fraud in the Ethereum network. Machine Learning Classification models such as Logistic Regression, Support Vector Machine (SVM) and Random Forest were implemented. These classification models will be used in the analysis to understand the patterns in the Ethereum network data according to their fraudulent or non-fraudulent transactions. Training and testing the data that were previously labelled, where it is possible to make other predictions of fraud that are based on the characteristics extracted from the data. With the Machine Learning methods results, the confusion matrix will be analysed where the models performance information will be conducted and the results identified. The model within the application of the models will be compared by evaluating the evaluation metrics such as the accuracy of the analysed data, F1 and recall. The deployment phase in the CRISP-DM methodology is the final stage in the process of detecting fraud on the Ethereum network.

4 Design Specification

The project design specification provides a clear view of the structure and processes involved from start to finish of data analysis. This architecture was divided into three layers between the Pre-processing and Data Transformation Layer, Modeling and Training Layer and Evaluation and Results Layer, having an approach that are important for each step of the project until its completion. Each layer has other processes involved focusing in achieving the project objectives.

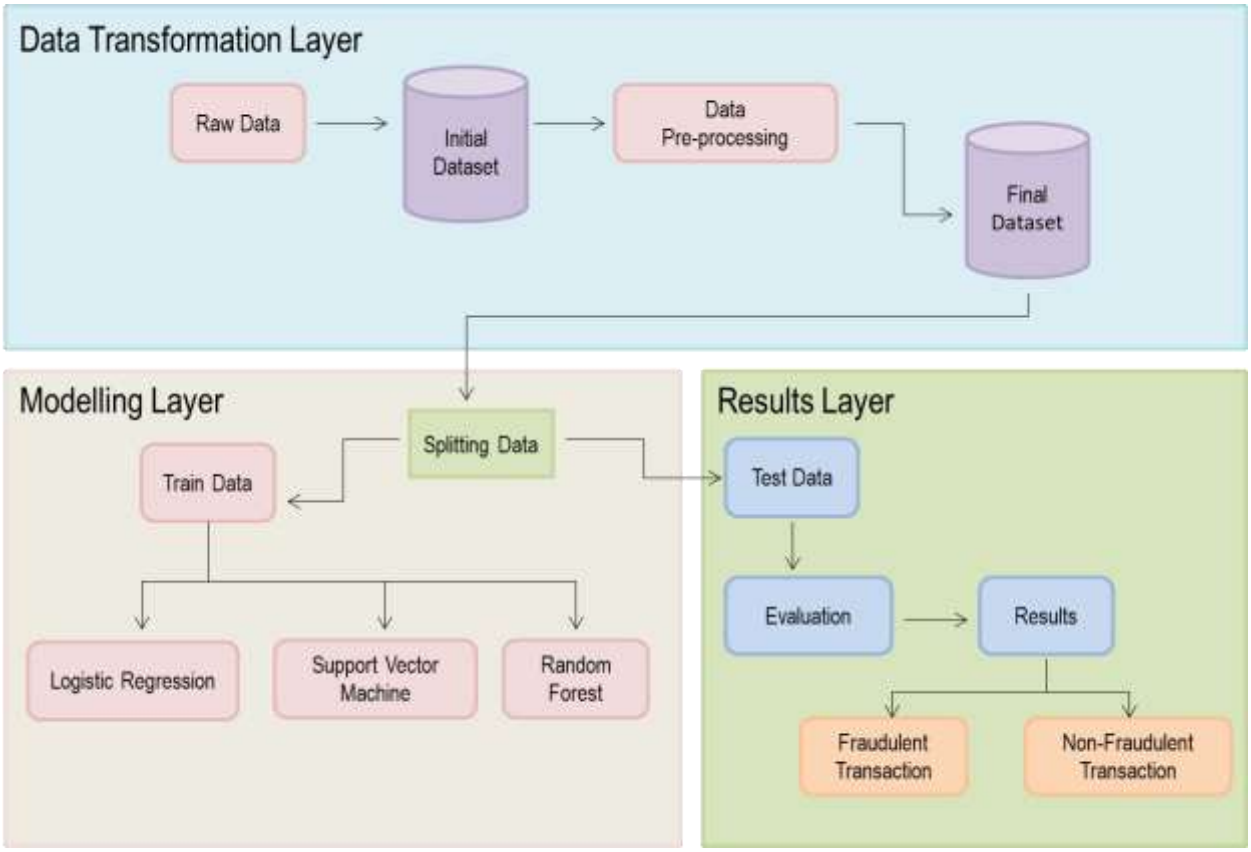


Figure 3: Flowchart Design Specification

4.1 Pre-processing and Data Transformation Layer

In this layer, the relevant data is collected and prepared for analysis, this phase includes data cleaning, data processing and required transformations, ensuring the quality of the data that will be analyzed using machine learning models. In the pre-processing stage, data relevant to the project is collected, this involves cleaning the data and processing the information, applying the necessary transformations to guarantee a good result when the Machine Learning models are applied.

4.2 Modelling and Training Layer

In this layer, Machine Learning models such as Logistic Regression, Support Vector Machine and Random Forest are implemented and trained with the pre-processed data to optimize model performance.

4.2.1 Modelling Techniques

Logistic regression will be one of the models in Machine Learning that will be applied in the project, it is a statistical method that serves to model the relationship between a binary or categorical dependent variable and one or more independent variables. Introducing the binary response that is, two categories as positive or negative. It is an important technique in statistical analysis as one of the fundamental tools in data analysis for solving classification problems and making reliable predictions with categorical responses.

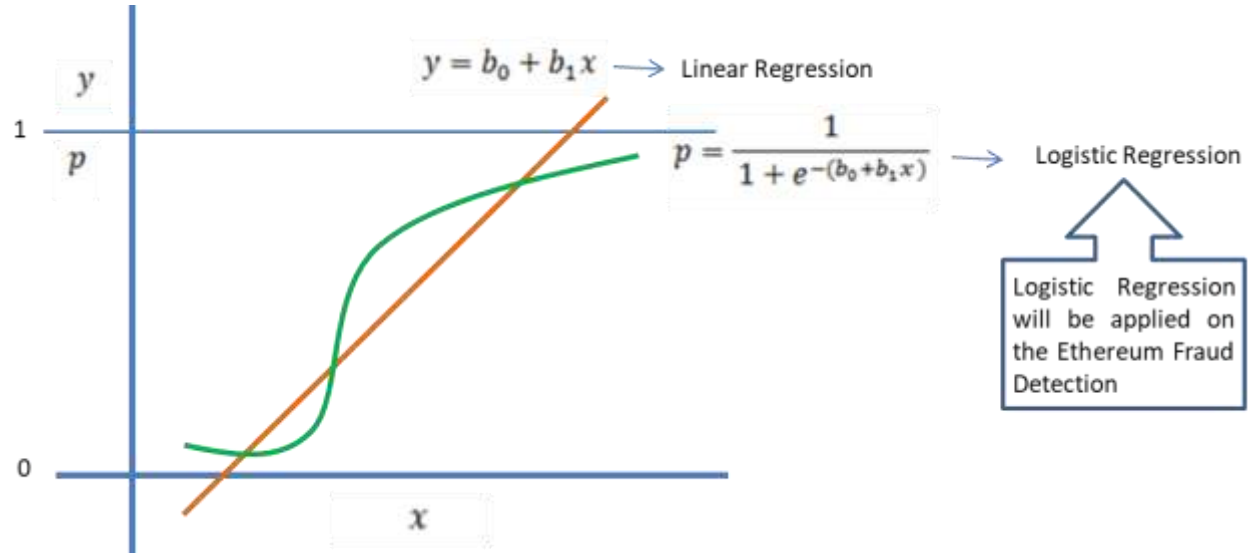


Figure 4: Logistic Regression

Random Forest is characterized as an algorithm that is used for classification tasks, regression and other forms of data analysis. Combining predictions from multiple models to improve the accuracy. Random Forest creates multiple decision trees during the training process, where each tree is built from a random sample of data and independent variables, that ensures that they are different from each other, avoiding over fitting the training data. When a prediction is needed, Random Forest makes each decision tree vote for its class or regression value. The class or value that receives the most votes is considered the final prediction.

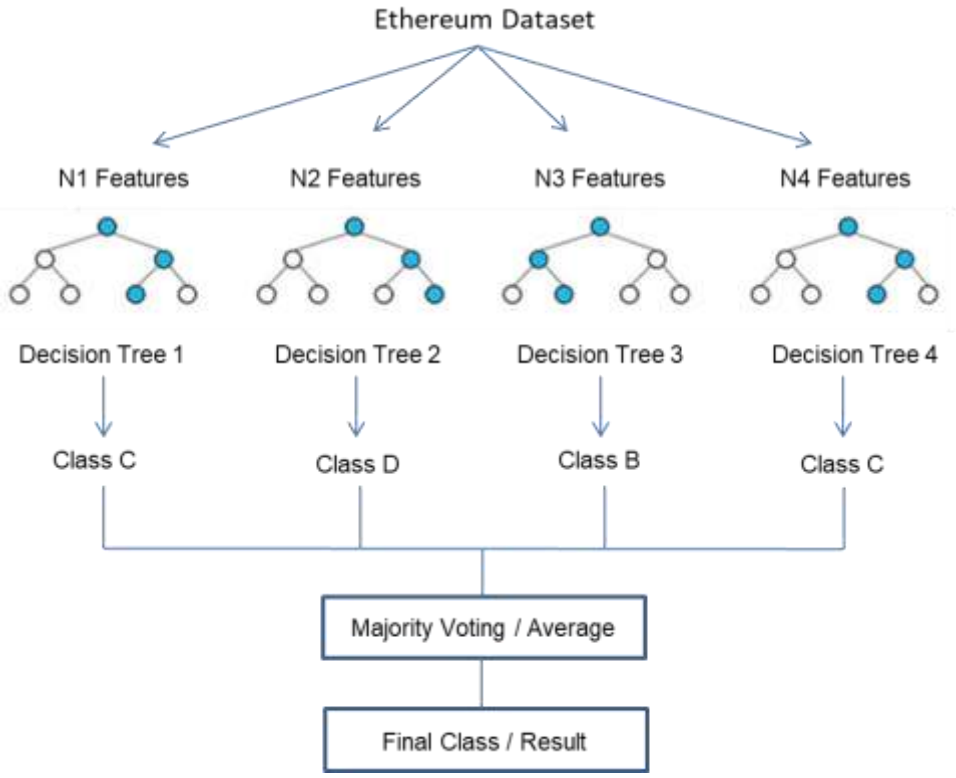


Figure 5: Random Forest

SVM (Support Vector Machine) is a machine learning algorithm used for classification and regression tasks. It is mainly used in classification problems where the data are separated linearly and the classes are separated by a line or hyperplane, that serves to increase the margin between the closest samples of each class and it is called vectors. SVM will be inserted in the project for data analysis, complementing the other models in the analysis of fraud in the Ethereum network data.

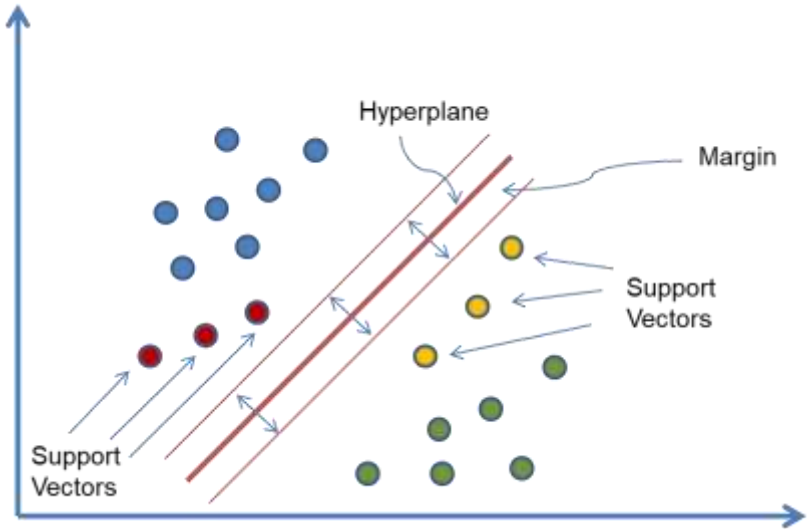


Figure 6: Support Vector Machine

4.3 Evaluation and Results Layer

After the pre-processing of the data, it is necessary to apply an evaluation of the effectiveness of the machine learning methods applied to the dataset, using specific metrics in the detection of fraudulent transactions and analysis of the best results.

Precision: It can be characterized as the proportion of fraudulent transactions on the Ethereum network that were correctly identified in relation to all transactions that are fraud. Being a measure of how reliable it can be the model in detecting fraud.

Recall: It is based on identifying the proportion of fraudulent transactions correctly based on the test set, being a measure indicating how well the model can find fraudulent transactions in the network.

F1-Score: Combining precision and recall, this metric provides a balance with the other metrics and the classes in the dataset.

Confusion Matrix: It is a table that summarizes the predictions of the model in relation to the true labels of the test set. It shows true positives, true negatives, false positives, and false negatives.

These metrics are necessary for evaluating the performance of the models in detecting fraud on the network, and in correctly identifying which transactions are fraudulent and non-fraudulent. Since the objective is to find the best model with the best performance, having a balanced precision and have an effective model in the detection of frauds in the Ethereum network.

5 Implementation

In the project implementation, the relevant resources are identified to process the data sets in detecting frauds in the Ethereum network. The dataset used in the project is publicly available in CSV format including information about transactions, ERC-20 tokens, characteristics of the Ethereum network, among other data relevant to the analysis. The language used in the project was Python, since it is a simple and readable programming, so it is ideal for the development of the project, with specific libraries that are applicable to the model.

For the next step when analyzing the data, it was possible to identify variables correlated to the target variable, so Python codes were implemented to transform the data and its variables, applying Machine Learning models to improve the accuracy of the data and their insights. After pre-processing the data, the data set was divided into training, validation and testing. This division is necessary to ensure that the models are trained, and adjusted, allowing the evaluation of the capacity of the models on previously unobserved data. In the project, Python libraries such as sklearn were used, applying machine learning models such as Logistic Regression, Random Forest and SVM to identify fraudulent transactions in the analyzed data. In addition to the analysis of the accuracy of the results, the F1 metric and the confusion Matrix will also be used to analyze the performance of the project, after the analysis of the confusion Matrix, it is important in the accuracy of the data where false

positive and false negative information can be analyzed. The standard confusion matrix is exemplified in Fig. 7 with the definitions of the different labels.

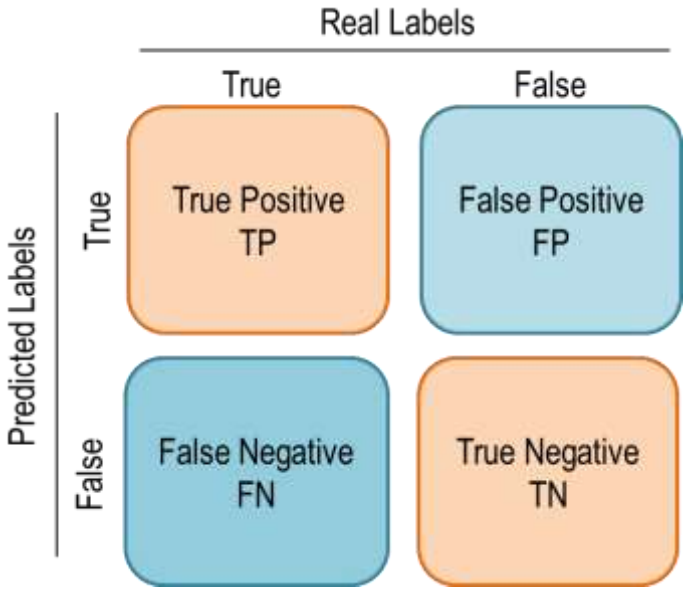


Figure 7: Confusion Matrix

True Positive (VP): Represents the cases in which the model correctly classified the samples as positive and which are actually positive. Being successful in correctly identifying positive samples.
False Positive (FP): Where the model mistakenly classified the samples as positive, when in fact they are negative, being errors in the identification of negative samples as positive.
False Negative (FN): Indicates cases in which the model erroneously classified the samples as negative, when they are actually positive, these are errors in identifying positive samples as negative.
True Negative (TN): Denotes cases where the model correctly classified the samples as negative being actually negative. Being successful in correctly identifying negative samples.

6 Evaluation

The objective of this research is to apply supervised machine learning techniques in the analysis of the fraud detection project in the Ethereum network. The aim is to obtain valuable insights through the use of logistic regression, SVM and random Forest methods and to evaluate which of the analyzed methods present the best performance for the model. With the results obtained, it will be possible to improve the fraud detection project, contributing to support the security and reliability of the Ethereum network and benefiting users and organizations that use this platform.

Atribute Name	Description	Data Type
FLAG	Whether the transaction is fraud or not	int64
Avg min between sent tnx	Average time between sent transactions for account in minutes	float64
Avg min between received tnx	Average time between received transactions for account in minutes	float64
Time Diff between first and_last (Mins)	Time difference between the first and last transaction	float64
Sent_tnx	Total number of sent normal transactions	int64
Received_tnx	Total number of received normal transactions	int64
NumberofCreated_Contracts	Total Number of created contract transactions	int64
MaxValueReceived	Maximum value in Ether ever received	float64
AvgValueReceived	Average value in Ether ever received	float64
AvgValSent	Average value of Ether ever sent	float64
MinValueSentToContract	Minimum value of Ether sent to a contract	float64
TotalEtherSent	Total Ether sent for account address	float64
TotalEtherBalance	Total Ether Balance following enacted transactions	float64
ERC20TotalEther_Received	Total ERC20 token received transactions in Ether	float64
ERC20TotalEther_Sent	Total ERC20token sent transactions in Ether	float64
ERC20TotalEtherSentContract	Total ERC20 token transfer to other contracts in Ether	float64
ERC20UniqSent_Addr	Number of ERC20 token transactions sent to Unique account addresses	float64
ERC20UniqRecTokenName	Number of Unique ERC20 tokens received	float64

Figure 8: Research Dataset Explanation

In the table, the data reveals that in the Ethereum network, the ERC20 standard represents an Exchange Rule that makes it possible to convert other currencies (eg DOGECoin) into Ethereum (ETH). ERC20 are the platforms main Tokens and represent the native currencies of the Ethereum network that circulate alongside ETH. Tokens can have different functions within the Ethereum network, promoting diversification and facilitating the integration of digital assets within that network. The ETL process is essential to guarantee the quality and integrity of the data, as it facilitates the analysis and decision-making when looking at data in search of more reliable information, analyzing and structuring the dataset for the application of Machine Learning models. For the analysis of real data we have target distribution of being Fraud or not. With a percentage of non-fraudulent instances of 77.9% and dealing with 22% fraudulent instances. Then Machine Learning models will be applied in the project to better analyse the real data, using Machine Learning techniques focused on the analysis of the models and seeking to apply training and testing of the data to see which techniques are more promising in the analysis of the data and better results.

Figure 9: Fraud and non-Fraud Chart

The dataset exhibited an imbalance, with 22% of accounts flagged as fraudulent and 78% as non-fraudulent. To assess the outcomes, the evaluation metrics employed were the Confusion Matrix, F1 Score, and Recall, aiming for enhanced optimization. Emphasis was placed on False Positives for accuracy and on False Negatives for recall. The F1-Score metric, combining accuracy and recall, provided a comprehensive measure of model efficacy. Throughout the training and testing phases, accuracy and recall results were achieved for non-fraudulent transactions, leading to an increase in the F1 score. Consequently, the Logistic Regression and SVM models fell short in performance, exhibiting a lower F1 score in comparison to the Random Forest model. This underscores the Random Forest model as the preferred choice, especially for addressing imbalanced data scenarios.

Data cleaning was necessary after carrying out the first analyzes where it was possible to observe the existence of variables correlated with each other. Thus, it was necessary to discard some features to avoid multicollinearity and possible redundancy problems in later analyses. By eliminating some of the variables that had the highest correlation, the dataset became more suitable for applying machine learning techniques, providing more accurate results. The data cleaning and pre-processing step was essential to ensure that the fraud detection model could be reliably trained and evaluated, contributing to a more accurate analysis of the project.

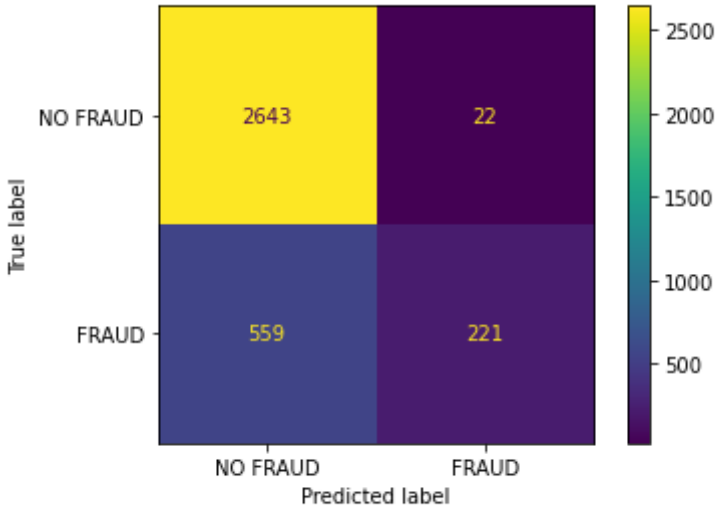
```
drop = ['total transactions (including tnx to create contract'
        'total ether sent contracts',
        'max val sent to contract',
        ' ERC20 avg val rec',
        ' ERC20 avg val rec',
        ' ERC20 max val rec',
        ' ERC20 min val rec',
        ' ERC20 uniq rec contract addr',
        'max val sent',
        ' ERC20 avg val sent',
        ' ERC20 min val sent',
        ' ERC20 max val sent',
        ' Total ERC20 tnxs',
        'avg value sent to contract',
        'Unique Sent To Addresses',
        'Unique Received From Addresses',
        'total ether received',
        ' ERC20 uniq sent token name',
        'min value received',
        'min val sent',
        ' ERC20 uniq rec addr' ]

drop_new = []

for index, text in enumerate(drop):
    drop_new.append(text.strip().replace(" ", "_"))

df_transaction.drop(drop_new, axis=1, inplace=True)
```

Figure 10: Correlated features

6.1 Experiment 1: An Exploratory Analysis of Ethereum Fraud Detection through the Implementation of the Logistic Regression Machine Learning Model

Logistic regression is the first technique applied in the project, it is focused on binary variables, and thus allows estimating the probability regarding the identification of transactions. When detecting fraud on the Ethereum network, Logistic Regression can be used to estimate the probability that a transaction is fraudulent or genuine, based on the known characteristics of the transactions. This approach helps in identifying suspicious transactions contributing to the security and reliability of the Ethereum network.

	precision	recall	f1-score	support
0	0.83	0.99	0.90	2665
1	0.91	0.28	0.43	780
accuracy			0.83	3445
macro avg	0.87	0.64	0.67	3445
weighted avg	0.84	0.83	0.79	3445




```
<Figure size 720x720 with 0 Axes>
```

```
lgr.score(features_test, labels_test)
```

```
0.8313497822931786
```

Figure 11: Confusion Matrix

Considering the confusion matrix:

Using 0.65 for training of the real data we got a better result. LR model, correctly identified 559 (TP) true positive of FRAUD cases, out of 780 (P) positive cases. LR model flagged as

FRAUD 22 (FP) false positive out of 2665, when these cases were actually NON-FRAUD. With an accuracy of 83% in identifying fraud, it proved to be a good model in the analysis, but not ideal.

6.2 Experiment 2: An Exploratory Analysis of Ethereum Fraud Detection through the Implementation of the Support Vector Machine (SVM) Machine Learning Model

The SVM was also applied in the project as it is a supervised machine learning algorithm and handles classification or regression challenges in a variety of applications, including fraud detection on the Ethereum network. The focus of this model is on the training and classification of a data set, seeking to find separation hyperplanes between non-fraudulent and fraudulent transactions.

	precision	recall	f1-score	support
0	0.78	1.00	0.88	5367
1	0.00	0.00	0.00	1522
accuracy			0.78	6889
macro avg	0.39	0.50	0.44	6889
weighted avg	0.61	0.78	0.68	6889

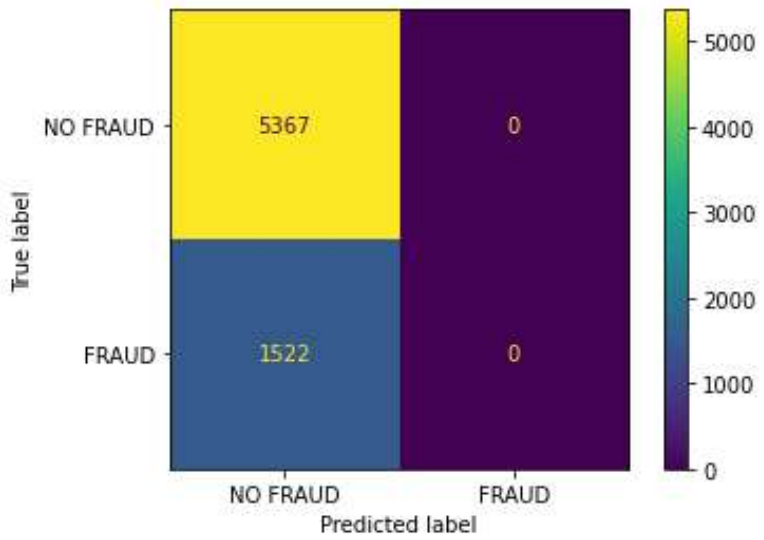


Figure 12: Confusion Matrix

Using 0.7 for training of real data but it was not satisfactory, probably overfitting considering the confusion matrix, SVM could not identify frauds in the model, even with an accuracy of 78%, it is not a good model to be implemented in this case. It does not perform well when we have a large dataset because the required training time is large.

6.3 Experiment 3: An Exploratory Analysis of Ethereum Fraud Detection through the Implementation of the Random Forest Machine Learning Model

Random Forest uses decision trees to combine its classification results, where trees are built with probability, decision and termination nodes, with possible different decisions. Thus, this model has shown promise in detecting fraud on the Ethereum network.

	precision	recall	f1-score	support
0	0.97	0.99	0.98	5367
1	0.95	0.91	0.93	1522
accuracy			0.97	6889
macro avg	0.96	0.95	0.95	6889
weighted avg	0.97	0.97	0.97	6889

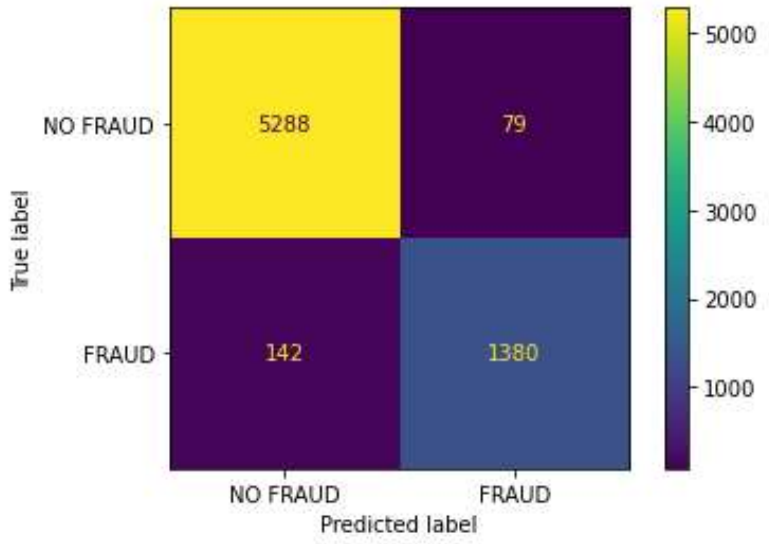


Figure 13: Confusion Matrix

Considering the confusion matrix:
Using 0.7 for training of the real data we got a better result.
Random Forest model, correctly identified 142 (TP) true positive of FRAUD cases, out of 1520 (P) positive cases.
Random Forest model flagged as FRAUD 79 (FP) false positive out of 5367, when theses cases were actually NON-FRAUD.
With an accuracy of 97% in identifying fraud, it proved to be the best model for identifying fraud. The most important features in the analysis of fraud or non-fraud were the following, with ERC20_uniq_rec_token being the most important for the definition of fraud in transactions.

When comparing its results with other classification models, such as Logistic Regression and the Support Vector Machine, Random Forest was the best model, presenting effectiveness in identifying fraudulent transactions. This range of combined trees allowed to improve the overall accuracy of the model, making it a more suitable option to support the security and integrity of the Ethereum network.

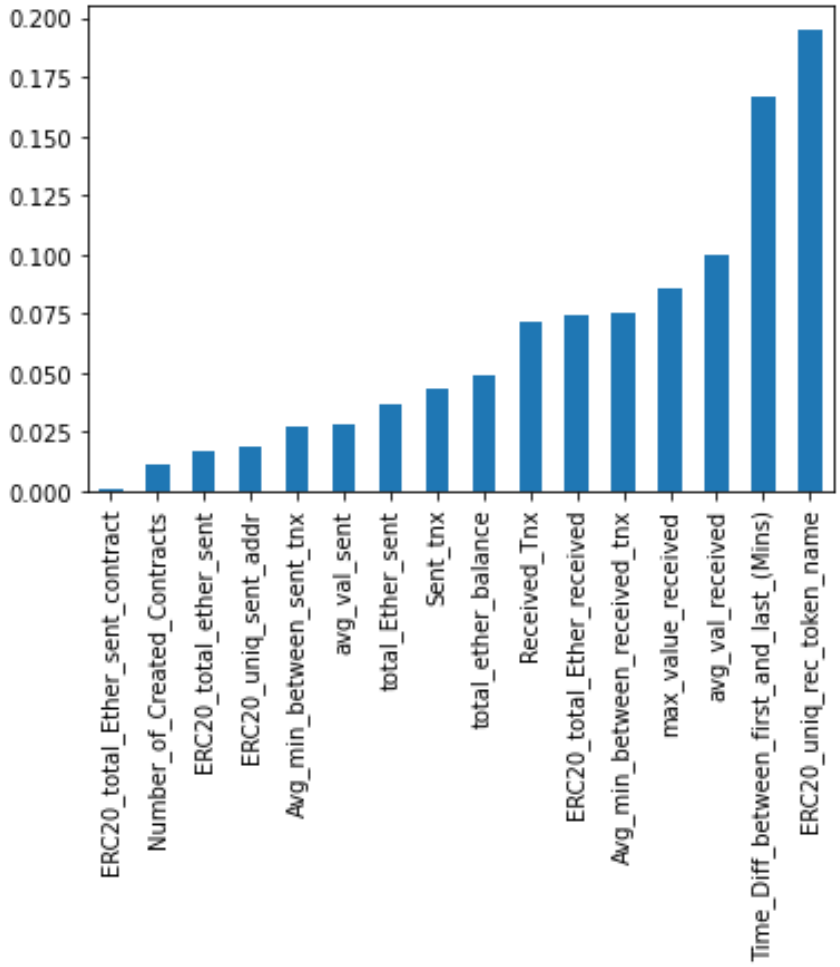


Figure 14: Important Features

When examining the Ethereum dataset, important factors that emerged as crucial for detecting fraud included token-related attributes, average values derived from smart contracts, and time frame between initial and concluding transactions. Researching into the analysis of smart contract actions linked to a transaction proved crucial, as it enabled the identification of suspicious interactions and potentially malicious contracts, both serving as red flags for fraudulent activity. In regard to tokens, monitoring balances and token transfers are important to finding irregularities. Furthermore, analyzing transaction timestamps presented valuable insights, particularly when transactions occurred during unconventional hours. Employing Machine Learning models played an indispensable role in extracting these critical features and accurately find fraudulent transactions within the dataset.

7 Discussion

Important metrics were used in detecting fraud on the Ethereum dataset, through the utilization of a confusion matrix analysis. True Negatives, representing instances where the model correctly identified legitimate transactions, were crucial in ensuring the model's accuracy in identifying non-fraudulent activities. A substantial count of True Negatives was highly required as it indicated the models ability in identifying genuine transactions. False Positives represented cases where the model erroneously categorized legitimate transactions as fraudulent, potentially leading to misinformation. Reducing the number of False Positives was of highest importance to maintain data integrity. Precision metrics were employed to determine the proportion of transactions identified as fraudulent by the model that really exhibited fraudulent characteristics. A high precision value indicated that a significant portion of transactions labeled as fraudulent were indeed fraudulent, a crucial factor in mitigating False Positives. In terms of Recall, it evaluated the models ability to identify fraudulent transactions among the fraudulent transactions. A high Recall, as exemplified by the Random Forest model, indicated that the model was good in identifying most fraudulent transactions, thus reducing False Negatives.

The F1 score emerged as a valuable metric finding a balance between accuracy and recall. In the Ethereum network fraud project, achieving this balance was essential to minimize both false positives and false negatives. A high F1 score signaled that the model accurately balanced accuracy and recall. Overall, when evaluating fraud-related metrics within the Ethereum network context, is important to prioritize high accuracy to control False Positives and look for high Recall to ensure the detection of most fraudulent transactions.

Analyzing the research question, it can be clarified that the classification models are adequate for the analysis. Three machine learning models were applied in the project to investigate its effectiveness in detecting fraud on the Ethereum network, as Logistic Regression, Support Vector Machine and Random Forest. Logistic regression showed an accuracy of 83% so it was a good result but it is not the best method to detect fraud on the network. Even with an accuracy of 78%, the SVM proved to be unreliable due to its inability to correctly identify fraud in transactions, as evidenced by the confusion matrix.

SVM model showed unsatisfactory results during the study implementation stages. Comparing the results of the three models, Random Forest presented promising results in detecting fraud on the Ethereum network, this model was the more effective and suitable approach for this project. Comparing this project to studies previously discussed in the literature review, it can be define as a necessary project for future new studies in this field, with two of the models showing important results and can be implemented in different datasets for possible new analyses.

8 Conclusion and Future Work

The present project investigated fraudulent transactions on the Ethereum network, using machine learning classification models such as Logistic Regression, Support Vector Machine and Random Forest, that were evaluated for their ability to identify fraudulent transactions.

After the analysis using the models, the results showed that Logistic Regression proving to be a viable alternative for fraud detection. However, the Support Vector Machine was not reliable in identifying fraud, not demonstrating good effectiveness for this project, as indicated by the confusion matrix.

Both models faced difficulties in achieving satisfactory results in the study implementation stages. However, when comparing the overall results of the three models, Random Forest is the most promising approach in detecting fraud on the Ethereum network. Therefore, based on the findings of this study, it is recommended that future research consider the use of Random Forest as a more efficient and adequate option to detect fraud in the Ethereum network, providing valuable insights to improve the security and integrity of transactions in this network.

References

Ajay K., Abhishek K., Nerurkar P., Ghalib M.R., Shankar A., Cheng X., 2020. Secure Smart Contracts for Cloud Based Manufacturing Using Ethereum Blockchain, Emerging Telecommunications Technologies, 4129.

Brown A., Tuor A., Hutchinson B., and Nichols N., "Recurrent neural network attention mechanisms for interpretable system log anomaly detection," in Proceedings of the First Workshop on Machine Learning for Computing Systems, 2018, pp. 1–8.

Bartoletti M., Carta S., Cimoli T., and Saia R., "Dissecting ponzi schemes on ethereum: identification, analysis, and impact," http://arxiv. org/abs/1703.03779, 2017.

Bartoletti M., Pes B., and Serusi S., "Data mining for detecting bitcoin ponzi schemes," http://arxiv.org/abs/1803.00646, 2018.

Harlev M. A., Sun Yin H., Langenheldt K. C., Mukkamala R., and Vatrapu R., "Breaking bad: De-anonymising entity types on the bitcoin blockchain using supervised machine learning," in Proceedings of the 51st Hawaii International Conference on System Sciences, 2018.

Ibrahim R.F., Mohammad Elian A., Ababneh M., Illicit account detection in the ethereum blockchain using machine learning, in: IEEE intl. Conf. On information technology, ICIT, 2021, pp. 488e493. https://doi: 10.1109/ICIT52682.2021. 9491653.

Jung E., Tilly M.L., Gehani A., Ge Y., 2019. Data Mining Based Ethereum Fraud Detection. In Conference of Blockchain, IEEE, pp. 266-273.

Muller G., Accorsi R., and Brenig C., 2015. Economic Analysis of Cryptocurrency Backed Money Laundering. European Conference on Information Systems.

Pham T. and Lee S., "Anomaly detection in the bitcoin system-a network perspective," preprint arXiv:1611.03942, 2016.

Tan R., Tan Q., Zhang P., Li Z., Graph neural network for ethereum fraud detection, in: IEEE intl. Conf. On big knowledge, ICBK, 2021, pp. 78e85. https://doi: 10.1109/ ICKG52313.2021.00020.

Weber M., Domeniconi G., Chen J., 2019. Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics, IEEE, pp. 4-7.