

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/348125890>

K-Nearest Neighbors Algorithm (KNN): An Approach to Detect Illicit Transaction in the Bitcoin Network

Chapter · January 2021

DOI: 10.4018/978-1-7998-5781-5.ch008

CITATIONS

7

READS

491

3 authors:



Abdelaziz Elbaghdadi

Ecole de Sciences Appliqués de Tanger

4 PUBLICATIONS 15 CITATIONS

SEE PROFILE



Mezroui Soufiane

Ecole de Sciences Appliqués de Tanger

20 PUBLICATIONS 33 CITATIONS

SEE PROFILE



Ahmed El Oualkadi

Abdelmalek Essaâdi University

156 PUBLICATIONS 801 CITATIONS

SEE PROFILE

Chapter 8

K-Nearest Neighbors Algorithm (KNN): An Approach to Detect Illicit Transaction in the Bitcoin Network


Abdelaziz Elbaghdadi

Abdelmalek Essaadi University, Morocco

Soufiane Mezroui

Abdelmalek Essaadi University, Morocco

Ahmed El Oualkadi

 <https://orcid.org/0000-0002-4953-1000>

Abdelmalek Essaadi University, Morocco

ABSTRACT

The cryptocurrency is the first implementation of blockchain technology. This technology provides a set of tracks and innovation in scientific research, such as use of data either to detect anomalies either to predict price in the Bitcoin and the Ethereum. Furthermore, the blockchain technology provide a set of technique to automate the business process. This chapter presents a review of some research works related to cryptocurrency. A model with a KNN algorithm is proposed to detect illicit transaction. The proposed model uses both the elliptic dataset and KNN algorithm to detect illicit transaction. Furthermore, the elliptic dataset contains 203,769 nodes and 234,355 edges; it allows to classify the data into three classes: illicit, licit, or unknown. Each node has associated 166 features. The first 94 features represent local information about the transaction. The remaining 72 features are called aggregated features. The accuracy exceeded 90% with $k=2$ and $k=4$, the recall reaches 56% with $k=3$, and the precision reaches 78% with $k=4$.

DOI: 10.4018/978-1-7998-5781-5.ch008

I. INTRODUCTION

New Technologies are created to spur financial innovation and improve the financial inclusion. These technologies deviate to their main goals and give a new opportunity for criminals and terrorists to launder their proceeds or their illicit activities. The Financial Action Task Force international standards combating money laundering and the financing of terrorism members in 2012 adopted their standards to monitor the risks relating to new technology. In 2014, the Financial Action Task Force (FATF) published virtual currency key definitions and potential AML /CFT Risks regarding the revolution of the cryptocurrency and their mechanisms associated with payment for giving a new method for transmitting values over the internet.

The FATF defines cryptocurrency as a decentralized convertible virtual currency protected by cryptography. The FATF discover and analyze the concrete action taken by criminals to launder incriminated funds through cryptocurrencies, they offer recommendations for compliance officers and companies that deal with cryptocurrencies.

Today, data analysis can be used to detect the anomalies or predict the future results with the help of data in the different fields. Furthermore, others issues should be solved with data analysis in the cryptocurrencies such as the influence of the distance used in the performance of k-Nearest Neighbors (KNN) model and how to use deep learning methods to evaluate precision, recall, F1 and accuracy for this task. The aim of this study is the detection of illicit transaction with KNN algorithm using Elliptic dataset. Furthermore, the blockchain technology uses it for a set field such as Business Intelligence, this technology gives a set technique such as smart contract to automate the processes in the enterprise without a central authority.

The rest of this chapter is organized as follows. The related works in cryptocurrency are presented in section II. In the third section, the Bitcoin and Ethereum network overview are described. In the fourth section, the machine learning technique is described. The fifth section presents the proposed methodology in this study. In the sixth section, the obtained results are discussed. Finally, the conclusion is given.

II. RELATED WORKS

The blockchain technology creates potential innovations in the processing of the business activities in various sectors, which makes this technology face to a set of attacks and illicit activity. This section reviews some related work which target cryptocurrency, such as anomaly detection, data analysis and business intelligence

1. Detection of Fraud and Anomalies in Cryptocurrency

The openness is one of the main characteristics of the Bitcoin network. This technology is open to public at any time. In addition, the public key of the bitcoin is a 160 bits hash generated by the secp256k1 curve (Antonopoulos M. Andreas, 2014) (Joppe W. Bos and al, 2014). This address can change often (Antonopoulos M. Andreas, 2014), and this propriety removes the possibility to identify the Bitcoin users via the public key. This still removes the possibility of tracking identity by analyzing the use of public keys on the network. However the address bitcoin is the main identifier to make a transaction and any one can stipulate that if two addresses (public key) are used as entries in the same transaction, then

K-Nearest Neighbors Algorithm (KNN)

the same user controls these public keys (Sarah Meiklejohn and al. 2013), For that reason the structure, form a Directed Acyclic Graph (DAG) can be used as technical analysis of transactions. Reid and Harrigan propose a method for breaking the bitcoin system in to DAGs, as part of their research on Bitcoin anonymity, they first developed a user graph that represents the flow of Bitcoin between users over time ((Reid and Harrigan 2013). Furthermore, K-means and Role Extraction are used to identify users who make the transactions typically associated with money laundering. Fifteen features are extracted using all transactions having a minimum of 650 transactions in the network. (Reid Fergal, and Martin Harrigan, 2012).

The researchers and the blockchain enthusiasts notice that, since their launch, blockchain has been implemented in the different fields such as cryptocurrency, IOT (Internet of things) and finance. The utility of this technology can be deviate to the original goal such as money laundering and illicit activity. That is why many of research papers target this technology such frauds and anomalies detection and identification of Bitcoin users etc. Zambre and al attempted to distinguish between normal users and malicious users based on real reported cases (Zambre, Deepak, and Ajey Shah 2013). Monamo and al use trimmed k-means clustering for anomaly detection (Monamo, Patrick, Vukosi Marivate, and Bheki Twala, 2016). To analyze the behavior patterns of user and transactions Pham and al use two types of graphs. Consequently, they detected three of the 30 known cases (Pham Thai, and Steven Lee, 2016). In addition, these authors have obtained similar results using k-means clustering Local Outlier Factor and power degree and densification in their subsequent study (Pham Thai, and Steven Lee, 2016). Yining Hu and al have created a transaction graph using data differentiate between laundering and regular transaction lies in their output value and neighborhood information (Yining Hu and al, 2019). Furthermore, Mark Weber and al use the elliptic data set to classify illicit and licit transaction (M. Weber and al, 2019). They used the variations of Logistic Regression (LR), Random Forest, Multilayer perceptron, and Graph convolutional networks (GCN) which is an emerging new method of capture relational information. Mark Weber and al share experimental results using a set of methods, from standard classification techniques (Logistic Regression, Random Forest, and Multilayer Perceptron) to the more sophisticated Graph Convolutional Networks. Important conclusions can be drawn from this work, including that: the Random Forest is figured as the best classification model for this problem. In addition, GCN are not the best performing models but is an interesting finding.

Moreover, a lot of the research works have been done based on others cryptocurrency such as Ethereum and Monero etc. A deep learning approach is proposed to detect DDos (Distributed Denial of Service) attack in the Bitcoin ecosystem (Ui-Jun Baek and al, 2019). In 2020, Abdelaziz and al have discussed the untraceability and unlinkability and describe how this cryptocurrency works and have presented some attacks against this technology (Abdelaziz, Soufiane, and Ahmed, 2020). Furthermore, the illicit activity has targeted as well the Ethereum blockchain, through the identification of Ponzi schemas deployed as a smart contract. In 2019, Chen and al classify the account with the help of the operation code using unsupervised learning, the interesting finding is the Random Forest has possibility to classify 305 out of 394 identified smart-Ponzi schemes (Chen Weili and al, 2019). To detect illicit the fraudulent activity on the blockchain Ethereum network. OKane examines the transactions enacted by scams available on Etherscambd database along with token and exchange addresses using a set of algorithms as the following: principal component analysis, random forest and K-means (O’Kane 2018). Steven Farrugia and al propose an approach to identify the illicit accounts on the Ethereum blockchain, this study uses an XGBoost classifier which 42 features as input, which XGBoost is an open source software allowing to implement Gradient boosting methods in both R and Python programming languages (Farrugia, Steven,

Joshua Ellul, and George Azzopardi, 2020. The XGBoost was attained a mean accuracy of 0.963 (\pm 0.006) with a mean AUC (Area Under the ROC curve) of 0.994 (\pm 0.0007).

2. Data Analysis in Cryptocurrency and Business Intelligence Using Blockchain

Several related work uses the data either to predict the price or to trade digital currencies. A supervise learning algorithms such as logistic regression, Naive Bayes Theory and SVM have used to identify cryptocurrency market (Colianni Stuart, Stephanie Rosales, and Michael Signorotti, 2015). A linear model and sentiment analysis have been used to predict price both in the Ethereum and bitcoin cryptocurrencies (Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan 2018). Israa and al have analyzed the evolution of the entire bitcoin transaction graph with the help of data analysis (Israa Alqassem, Iyad Rahwan, and Davor Svetinovic, December 2018). Jingming and al. have achieved a study using data of Monero cryptocurrency and discussed the potentials of the blockchain in the energy industry. The obtained result of this study have shown the electricity consumption of the Monero mining (Jingming Li Nianping Li Jinqing Peng Haijiao Cui Zhibin Wu November, 2018). Salim Lahmiri and al have used deep learning algorithm to predict the price of bitcoin, Digital cash and Ripple cryptocurrencies (Salim Lahmiri, Stelios Bekiros, 2019).

In 2016, Ingo Weber and al have developed a technique to integrate blockchain into the choreography of processes without central authority (Ingo Weber, Xiwei Xu, Régis Riveret, Guido Governatori, Alexander Ponomarev, Jan Mendling, 2016). A block chain-based framework is proposed as well as the use of smart contract to derive the possible advantage of the supply chain process design. The authors have provided a workable use case for business process disintermediation (Shuchih Ernest Chang, Yi-Chian Chen, and Ming-Fang Lu, 2018). Melanie Swan have discussed Blockchain distributed ledgers in the context of public and private Blockchains (Melanie Swan, 2018). Daniel E and O'Leary have investigated some of the implications and strategies that include the use of that open information. (Daniel E. O'Leary, 2018).

3. Summary

The blockchain technology offers a set of tracks in the scientific research as mentioned previously. The table 1 shows a review of some related work which target this topic.

III. THE BITCOIN AND ETHEREUM NETWORK OVERVIEW

The Merkle Trees was incorporated into the design of the concept of the cryptography to verify the integrity of a set of data without necessarily having all of them at the time (Vujičić, Jagodić, and Randjić 2018). Furthermore, a system to secure the chain of blocks with the help of the cryptography technique that signify time-stamp a digital document are developed (Bralić, Stančić, and Stengard, 2020). The Proof of Work (PoW) has been proposed as a prototype for digital cash to solve the double-spending problem (A. Meneghetti, M. Sala, D Taufer, 2020). Moreover, in 2008 Satoshi Nakamoto conceptualized the theory of distributed blockchains, he has developed a design in a unique way to add blocks to the initial chain without requiring them to be signed by trusted parties (S. Nakamoto, 2008). The Bitcoin

K-Nearest Neighbors Algorithm (KNN)

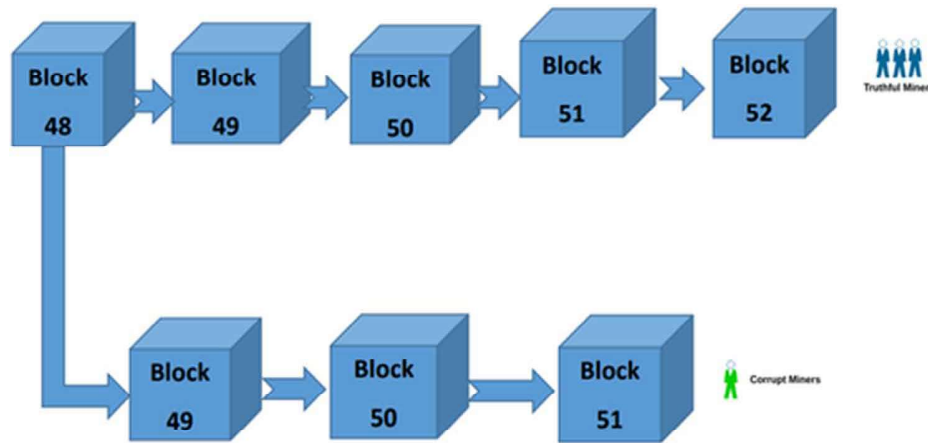
Table 1. A review of some related works.

Cryptocurrencies		
Bitcoin	Monero	Ethereum
<ul style="list-style-type: none"> ✓ DDoS attack detection in bitcoin Ecosystem with help of data analysis ✓ Anomalies Detection (Identify the bitcoin users, Illicit Transaction detection in the bitcoin Network, Money laundering etc.) ✓ Predict the Price of the Bitcoin cryptocurrency with help of data and sentiment analysis 	<ul style="list-style-type: none"> ✓ Untraceability and linkability in Monero Cryptocurrency using Data analysis ✓ Electricity consumption of the Monero mining 	<ul style="list-style-type: none"> ✓ Detection illicit Activity in the blockchain Ethereum ✓ Predict the Price of the Ethereum ✓ Trading in the cryptocurrency ✓ identify cryptocurrency market
Business intelligence using the Blockchain		
<ul style="list-style-type: none"> ✓ use the blockchain Technology for business process (smart contract as example) 		

is a set of the concept and technologies forms the base of the digital currency ecosystem, it is used to retain and transmit value between the participants in the Bitcoin network (Reid and Harrigan, 2013).

The Bitcoin network users communicate between each other using the Bitcoin protocol mainly via the internet. The Bitcoin users can transfer the Bitcoin in the network to do everything that can be done with traditional currency, buy or sell goods and services, send money to individuals or organization, or provide credit. To ensure the security in the network, the Bitcoin technologies includes functionalities that are based on encryption and digital signatures. The Bitcoins can be bought, sold and exchanged for other currencies on specialized exchanges. The Bitcoin is in a sense the perfect form of currency for the internet as it is fast, secure (Antonopoulos M. Andreas, 2014).

Contrary to traditional currencies, the Bitcoins are virtual. There is no physical coin or even a digital coin. The coins are included in the transactions transmitting the value from sender to receiver. The Bitcoin users have keys that prove ownership of transactions on the Bitcoin network and unlock the value to spend and transfer it to another recipient. These keys are often stored in a digital wallet on each user's computer. The possession of a key to unlock a transaction is the only prerequisite for spending Bitcoins, so this system gives users complete control. The Bitcoin is a fully distributed pair-to-pair system. Then, there is no server or point of control. The Bitcoins are created through a process called "mining", which involves finding the solution to a problem that is difficult to solve. Any participant in the Bitcoin network (i.e., any computer operating the complete Bitcoin stack) can act as a minor, using the computing power it has at its disposal to solve the problem. Every 10 minutes on average, a new solution is found by someone who is then able to validate the transactions in the last 10 minutes. In summary, the Bitcoin mining decentralizes the issuance of money and reconciliation procedures, making the intervention of an agency similar to central banks unnecessary. The Bitcoin protocol includes predefined algorithms that regulate the mining function on the network. The difficulty of the miner's task execution to record a block of transaction on the Bitcoin network is adjusted so that on average someone gets there every 10 minutes, regardless of the number of miners (and CPUs) working on this task at time t. The protocol halves the number of Bitcoins created every four years and limits the total number of Bitcoins issued to a total of 21 million pieces. Seen as its speed of issuance declines, over the long term, the Bitcoin currency is deflationary. The Bitcoin cannot be artificially inflated by generating money beyond the rate of emission expected. The Bitcoin is also the name of a protocol, a network, and an innovation in distributed computing. Bitcoin as a currency is really only the first application of this invention (Antonopoulos M. Andreas, 2014) (Du Mingxiao, Ma Xiaofeng, Zhang Zhe, Wang Xiangwei, Chen Qijun, 2017).

Figure 1. The 51% attack scheme

As mentioned previously, the Bitcoin has been invented in 2008 with the publication entitled “Bitcoin: A peer to peer electronic cash system” written under the pseudonym Satoshi Nakamoto which has combined a set of previous inventions such as b-money and hashcash system. To create a cryptocurrency not relying to any central authority for the transmission of money or the regulation. Furthermore, the main innovation has been the use of a distributed computing system (called algorithm « Proof of work ») realizing “election” every 10 minutes, allowing the decentralized network to arrive at a consensus on the status of transactions, that solves the problem of double spending or a unit of currency cannot be spent twice. Previously, the problem of double spending was a weakness of digital currency and was resolved by having all transactions verified by a clearing organization (Antonopoulos M. Andreas, 2014).

The Bitcoin network started in 2009 based on the paper published by Satoshi Nakamoto. The distributed computing which provides the security and resistance to Bitcoin has grown exponentially. The identity of the person or group of people behind the creation of Bitcoin is not known to this day. However, either Satoshi Nakamoto or anyone else has any control over the Bitcoin system, which operates only according to fully transparent mathematical principles (Antonopoulos M. Andreas, 2014) (Eyal and Sirer, 2014). In the literature, a set of attack target the blockchain technology has been proposed. The 51% or double spending attack is one of the famous attacks in the Bitcoin network. The miner or a group of the miner testing to spend their transactions on the blockchain twice. In this scenario of attack, the attacker can prevent new transactions from taking place or being confirmed. The attacker has the possibility to block a new transaction from taking place or to be confirmed. Furthermore the attacker able to reverse the transactions that have already validated (Martijn Bastiaan, 2015). When anyone (miner) valid a block of the transactions, the block will broadcast to another miner on the network. The block can only be accepted if all transaction in the block are valid based on the available recording on a blockchain. However, the attacker with more than 50 Percent of a network’s hash rate does not broadcast solutions to the rest of the network. The first version of the block is the public version of the blockchain which is being followed by legal miners. The second is used by the attacker who are not broadcasting it to the rest of the network. Figure 1 shows the 51% attack operation. The 51 Percent attack is one of theory attack because the attacker needs more than 50 percent of the computing power of the entire network.

K-Nearest Neighbors Algorithm (KNN)

A Sybil attack occurs when a large number of nodes on a network belonging to the same party try to cause disruption to the business (Eyal and Sirer 2014) . In order to be able to create major disruptions, they will, for example, try to create a large number of fake transactions or simply manipulate the relay handling the real transactions. This type of attack occurs mainly in reputation systems. It generally disrupts them by creating false identities. This most often happens in a P2P type computer network. The Sybil attack targets the peer-to-peer network (Eyal and Sirer 2014). A hacker wants to make this attack on the Bitcoin network. In this case, the several identities at the same time and undermines the authority of the reputation system are exploited in the network by a node. Its primary objective is to obtain the majority of the influence in the network to carry out illegal actions in the system. Furthermore, the Sybil attack is not easy to detect and prevent, but the following measures may be helpful:

- Increasing the cost of creating a new identity.
- Joining the network requires validation of identities or trust.
- Making different power to different members.

The second famous blockchain is Ethereum which are invented by Vitalik Buterin in 2014, this cryptocurrency described as “the world computer”. Moreover, Ethereum is an open source, decentralized computing infrastructure that executes programs called smart contracts. A blockchain is used to synchronize and store the system state changes, as well as a cryptocurrency called “ether” to measure and limit the cost of execution resources (Wood Gavin, 2014).

The Ethereum platform allows developers to build strong decentralized applications with integrated economic functions. This Technology provides high availability, transparency, auditability and neutrality. Furthermore, it reduces or eliminates supervision and minimizes certain counterparty risks. The Ethereum has a lot of common features with other blockchains such as:

- A peer-to-peer network between the participants.
- To synchronization of state updates a Byzantine fault–tolerant consensus algorithm is used.
- The cryptographic primitives such as digital signatures and hashes, and a digital currency are used

The blockchain Ethereum uses the proof of stake (PoS) to validate each block, however, the Bitcoin uses the PoW. These two methods allow the creation of new blocks of a blockchain. These blocks contain the transactions performed. There are many differences between PoS and PoW. The PoS is a method that allows to record, validate transactions in blocks. For this, a person must have a certain number of tokens of a crypto-money. This method also allows the creation of new blocks but in a different way. To do this, users will make available the computing power of their machines. The latter will have to solve

Table 2. Different features between PoW and PoS

Type of methods	Use of resources	Degree of decentralization	Speed of transactions	Transaction fees
PoW	Very strong	low	slow	Quite high
PoS	Low	High Fairly	high	Low

more and more difficult mathematical calculations (Antonopoulos M. Andreas and Gavin Wood, 2018). Table 2 shows the difference between Ethereum and blockchain.

Both methods allow the same thing, i.e. the creation and validation of blocks in a blockchain. Each method has advantages and drawbacks. One of the big advantages of PoS is the energy gain. There is no need to spend astronomical amounts of electricity to validate transactions and enter them into the block chain. One of the big advantages of PoW is the security of the transactions as they all have to be validated by the miners.

The disadvantages of PoW are as the following:

- Requires very powerful equipment that is only owned by a minority of companies, which favours centralization
- Huge energy costs
- High transaction costs if the network begins to saturate

The disadvantages of PoS are as the follows (Saleh Fahad, 2020):

- Less secure than PoW
- Early investors have a big advantage
- Little use of token

IV. MACHINE LEARNING TECHNIQUE

Machine learning is a field of the artificial intelligence which is based on mathematical and statistical approaches to give computers the ability to “learn” from the data, i.e. to improve their performance in solving tasks without being explicitly programmed for each one. The machine learning algorithms can be categorized according to the learning mode used. The first category if the classes are predetermined and the examples are labelled, the system learns to classify according to a classification or grading model; this is called supervised learning (Mohamed Alloghani and al. 2020). However, when the system or operator has only sampled, but no label, and the number of classes and their nature has not been predetermined, this is called unsupervised learning or clustering. No experts are required. The algorithm must discover by itself the more or less hidden structure of the data. Data clustering is an unsupervised learning algorithm (Mohamed Alloghani and al, 2020). The Semi-supervised learning is a class of machine learning techniques that uses a set of labelled and unlabelled data. It is thus situated between supervised learning, which uses only labelled data, and unsupervised learning, which uses only unlabelled data. The use of untagged data, in combination with tagged data, has been shown to significantly improve the quality of learning (V. Engelen, E. Jesper, and H. Hoos, 2020). Furthermore, Machine learning solves two problems such as regression problem or classification problem which the first is used in order to predict a value and the second is used in order to predict a class (Mohamed Alloghani and al. 2020).

In the literature, a set of algorithms used to solve the different problems of machine learning has been proposed (J Qiu, Q Wu, G Ding, Y Xu, S Feng, 2016). The K-means is an algorithm for clustering the data. Furthermore, this algorithm belongs to unsupervised learning. This clustering algorithm is used

K-Nearest Neighbors Algorithm (KNN)

to identify groups of observations with similar characteristics. Mathematically, given a set of points (x_1, \dots, x_n) , one must try to divide the n points into k sets such as:

$$S = (S_1, \dots, S_k) (k \leq n) \quad (1)$$

by minimizing the distance between the points and the center of the group or class, $\arg_s \min \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$ where μ_i is the centroid of the points in S_i and x_j is a point in S_i (Liu, Zhiguo, Changqing Ren, and Wenzhu Cai, 2020).

The Random Forest is a classification algorithm that reduces the variance of forecasts in a decision tree alone, thereby improving performance. It does this by combining multiple decision trees in a bagging approach. The random forest algorithm is proposed by L. Breiman. In its most classical formula, it performs parallel learning on multiple randomly constructed decision trees trained on different subsets of data. The ideal number of trees, which can be several hundred or more, is an important parameter: it is highly variable and depends on the problem (M. Schonlau, and R. Yuyan Zou, 2020). Logistic regression is a predictive technique. It aims to build a model to predict/explain the values taken by a qualitative target variable (most often binary, this is called binary logistic regression; if it has more than two modalities, it is called polytomous logistic regression) from a set of quantitative or qualitative explanatory variables (coding is necessary in this case) (Kuha Jouni, and Colin Mills, 2020). In machine learning, a convolutional neural network (CNN) is a type of acyclic artificial neural network, in which the connection pattern between neurons is inspired by the visual cortex of animals. The neurons in this region of the brain are arranged so that they correspond to overlapping regions when the visual field is paved. Their functioning is inspired by biological processes, they consist of a multi-layered stack of perceptrons, the purpose of which is to pre-process small amounts of information (Irfan Aziz, 2020). Furthermore, one of the machine learning technique for regression and classification problems is Gradient boosting which is used to calculate the weights of individuals when the construction of each new model. The gradient boosting can be considered as an optimization algorithm (Haihao Lu, Sai Praneeth Karimireddy, Natalia Ponomareva and Vahab Mirrokni, 2020).

A set of technique and metrics are used to evaluate the quality of prediction model (Qiu et al. 2016). Confusion matrix or contingency table is used to evaluate the quality of a classification. It is obtained by comparing the classified data with reference data, which must be different from those used for classification. This matrix is used to calculate a set of metrics such as precision, Recall and accuracy. To evaluate the proposed model, four evaluation metrics are used: Precision, Recall, accuracy and F1 Score. To see if the predictions are right or wrong the metrics using are defined in terms of true and false positives. These metrics are defined by:

- TN / True Negative: The case was negative but predicted negative.
- TP / True Positive: The case was positive but predicted positive.
- FN / False Negative: The case was positive and predicted negative.
- FP / False Positive: The case was negative and predicted positive.

K-Nearest Neighbors Algorithm (KNN)

Various metrics are used for this evaluation test. The first metric is precision which is defined as the report of correctly predicted positive observations to the total predicted positive observations. This metric is calculated by

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

The second metric is Recall which is defined as the report gives the proportion of positive identifications was corrected and calculated by

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

The third metric is the F-score which is defined as a compromise of accuracy and recall giving the performance of the system. This compromise is given in a simple way by the harmonic mean of accuracy and recall. The formula is calculated by

$$F1 = \frac{2 * precision * Recall}{precision + Recall} \quad (4)$$

The accuracy is the fourth metric which is defined as one of the criteria for evaluating classification models. Informally, the accuracy refers to the proportion of correct predictions made up by the model. Formally, accuracy is defined by

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (5)$$

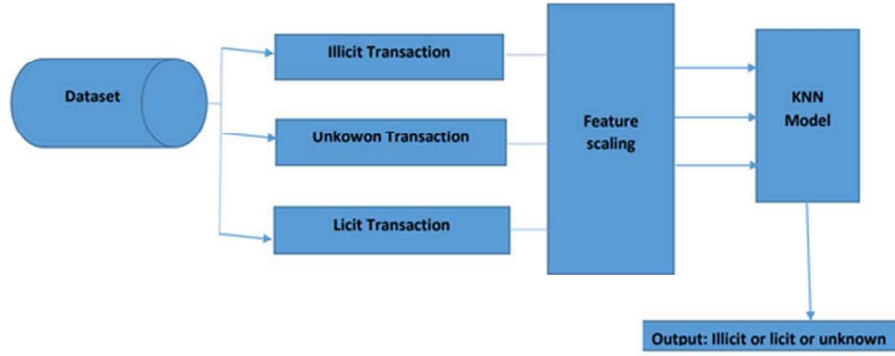
V. METHODOLOGY

The proposed approach uses both the Elliptic dataset and the KNN algorithm to detect illicit transaction. In this study, a model is developed by using the KNN algorithm and implemented in python. The dataset Elliptic is used to test the proposed model as shown in figure 2.

To test the performance of the proposed model, 25% of data is used and 75% for the training module. The Elliptic dataset contains 203,769 nodes and 234,355 edges as detailed in the next section. The feature scaling phase is used to normalize the range of independent variables or features of data.

K-Nearest Neighbors Algorithm (KNN)

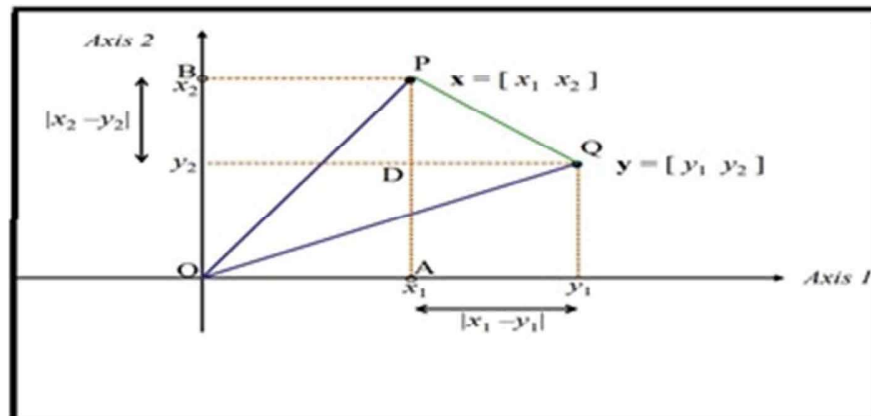
Figure 2. Detection Model using KNN algorithm and Elliptic dataset.



VI. ELLIPTIC DATASET OVERVIEW

The Elliptic dataset is a sup-graph of the Bitcoin data that contains 203.769 nodes and 234.355 edges, their use is to classify the data into three classes: illicit, licit or unknown. The transaction is licit if has been created by one entity belongs to exchanges, wallet providers, miners, financial service providers, etc. However, the transaction is illicit if has been created by one of the entities belong to scams, malware, terrorist organizations, ransom ware, Ponzi schemes, etc. No indication is provided on the other nodes which are classified as “unknown”. Each node has associated to 166 features including number of inputs, number of the outputs, transaction fee, average number of incoming transactions etc. The first 94 features represent local information about the transaction. The remaining 72 features, called aggregated features, are obtained by aggregating transaction (Mark Weber and al, 2019). In this study 165 features are used (time step feature is excluded)

Figure 3. Distance between two points



1. Euclidian Distance

Several functions are used to distance calculation such as Manhattan distance, Minkowski distance, Jaccard distance, Hamming distance, etc. The distance function is chosen according to the types of data being manipulated. In this study, the Euclidian distance between two points or two vectors is used (Figure 3).

To calculate the distance between two points $P=(x_1, x_2)$ and $Q=(y_1, y_2)$, the following equation is used:

$$d(Q, P) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (6)$$

For n-dimensional space, the distance Euclidian is calculated by:

$$d(Q, P) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (7)$$

Such that P and Q has n input (parameters).

2. Principle of KNN Algorithm

The KNN algorithm is a supervised learning method which can be used in the regression and classification problems. These operations can be likened to the following analogy “tell me who your neighbors are, I’ll tell you who you are”. Therefore, to make a prediction, the KNN algorithm will base itself on the entire dataset. Indeed, for an observation, which is not part of the dataset that we want to predict, the algorithm will look for the K instances of the dataset closest to our observation. Then for these K neighbors, the algorithm will use their output variable y. Then, for K neighbors, the algorithm will use their output to calculate the value of the variable y of the observation you wish to predict. In addition, if KNN is used for the regression, it is the mean (or median) of the y variables of the K closest observations that will be used for the prediction. However, if KNN is used for classification, the mode of the variables y of the closest K observations will be used for prediction.

The steps of the KNN algorithm are described in Figure 4. This algorithm has a data set D, a distance function d, and an integer K (Number of neighbors) as an input. For a new observation X and to predict the output variable y, the algorithm should follow steps in the figure 4.

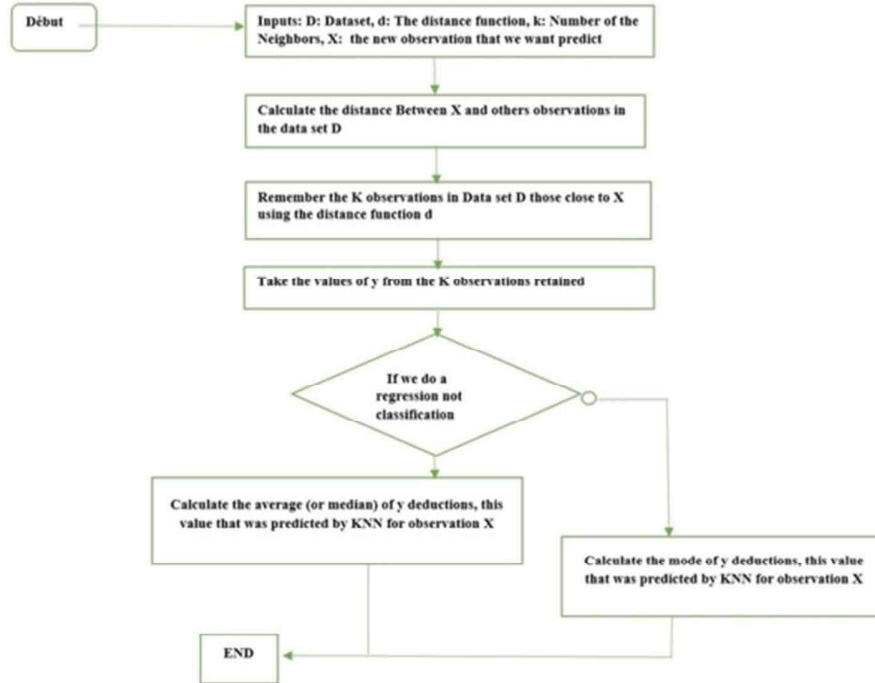
In relation to illicit transaction in the Bitcoin network problem, a transaction is considered as a vector with n parameters. Furthermore, each transaction is characterized by 165 features such that $T = (x_1, \dots, x_{165})$, and the main aim is to predict to which class the transaction belongs: Illicit or licit or unknown.

VII. RESULTS AND DISCUSSION

As mentioned previously, a confusion matrix is used to test the performance of the classification model. Table 3 shows a confusion matrix for various values of k (k=2, k=3, K=4, K=5, and K=20).

K-Nearest Neighbors Algorithm (KNN)

Figure 4. Steps of KNN algorithm



For $k=2$, 573 cases belong to illicit class and predict as unknown class, 546 cases belong to illicit and predict by the KNN model as illicit transaction and 21 case belong to illicit but predict as licit transaction. For $k=3$, 455 cases belong to illicit class and predict as unknown class, 645 cases belong to illicit and predict by the KNN model as illicit transaction and 40 case belong to illicit but predict as licit transaction. For $k=4$, 542 cases belong to illicit class and predict as unknown class, 564 cases belong to illicit and predict by the KNN model as illicit transaction and 34 case belong to illicit but predict as licit transaction. For $k=5$, 469 cases belong to illicit class and predict as unknown class, 631 cases belong to illicit and predict by the KNN model as illicit transaction and 40 case belong to illicit but predict as licit transaction. For $k=20$, 583 cases belong to illicit class and predict as unknown class, 524 cases belong to illicit and predict by the KNN model as illicit transaction and 33 case belong to illicit but predict as licit transaction.

In order to evaluate the proposed model and to find the best k value for the proposed system prediction, various k ($k=2$, $k=3$, $k=4$, $k=5$, and $k=20$) are used to calculate some metrics for each case. Table 4 shows the obtained results for various values of K and Euclidean distances.

The precision measures the percentage of the transaction identified as illicit that have been classified correctly. The precision reaches (74%, 29%, 78%, 71%, and 70%) with various k ($k=2$, $k=3$, $k=4$, $k=5$, $k=20$) respectively. The recall measures the percentage of actual illicit transactions that were classified correctly. The recall reaches (48%, 56%, 49%, 55%, and 46%) with various k ($k=2$, $k=3$, $k=4$, $k=5$, $k=20$) respectively. The accuracy is one of the criteria for evaluating classification models. The accuracy refers to the proportion of correct predictions made by the model. The accuracy measure reaches (89.62%, 89.91%, 90.08%, 90.09%, and 89.38%) with various k ($k=2$, $k=3$, $k=4$, $k=5$, $k=20$) respectively.

K-Nearest Neighbors Algorithm (KNN)

Table 3. Confusion matrix for each case

K=2			
	Unknown	Illicit	Licit
Unknown	38471	136	622
Illicit	573	546	21
Licit	3874	58	6642
	K=3		
	Unknown	Illicit	Licit
Unknown	37710	257	1262
Illicit	455	645	40
Licit	3104	18	7452
	K=4		
	Unknown	Illicit	Licit
Unknown	38395	154	680
Illicit	542	564	34
Licit	3619	24	6931
	K=5		
	Unknown	Illicit	Licit
Unknown	37981	240	1008
Illicit	469	631	40
Licit	3272	18	7284
	K=20		
	Unknown	Illicit	Licit
Unknown	38330	212	687
Illicit	583	524	33
Licit	3901	8	6665

Table 4. Obtained results for various k and Euclidean distances.

Metrics Value of K	Recall	Precision	F1	Accuracy
K=2	0.48	0.74	0.58	0.8962
K=3	0.56	0.29	0.38	0.8991
K=4	0.49	0.78	0.6	0.9008
K=5	0.55	0.71	0.62	0.9009
K=20	0.46	0.7	0.55	0.8938

VIII. CONCLUSION

Data analysis is important for detecting and predicting outcomes, for example, it can be used to detect anomalies or predict the future prices of cryptocurrencies such as Bitcoin, Ethereum, etc. This chapter reviews some issues in the cryptocurrency such as data analysis and business intelligence in the blockchain technology and proposes some solutions related to these issues. Moreover, the chapter proposes a model to detect an illicit transaction in Bitcoin cryptocurrency. This detection model uses Elliptic Dataset as reference and divides this dataset to both test and training models. The model is tested with same random value of neighborhood and Euclidian distance. The evaluation of the proposed model is based on calculation of some metrics to evaluate KNN model such as precision, recall and F1 score.

The obtained results are very interesting. The accuracy exceeded the 90% with $k=2$ and $k=4$, the recall reaches 56% with $k=3$ and the precision reaches the 78% with $k=4$. In the next work, others issues should be solved such as: the influence of the distance used in the performance of KNN model and the comparison of this KNN approach with machine learning algorithms (SVM, Logistic Regression, Linear regression, Random Forest...) to shows how to use deep learning methods to evaluate precision, recall, F1 and accuracy for this task.

REFERENCES

- Abdelaziz, E., Soufiane, M., & Ahmed, E. (2020). Survey of Monero Security. *Dynamical and Control Systems*, 12, 88-93.
- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Supervised and Unsupervised Learning for Data Science*. Springer. doi:10.1007/978-3-030-22475-2_1
- Andreas, A. M. (2014). *Mastering Bitcoin: Unlocking Digital Cryptocurrencies*. O'Reilly Media, Inc.
- Andreas, A. M., & Wood, G. (2018). *Mastering Ethereum: Building Smart Contracts and Dapps*. O'Reilly Media.
- Aziz, I. (2020). *Deep Learning: An Overview of Convolutional Neural Network*. CNN.
- Baek, U.-J., Ji, S.-H., Park, J. T., Lee, M.-S., Park, J.-S., & Kim, M.-S. (2019). *DDoS Attack Detection on Bitcoin Ecosystem Using Deep-Learning*. IEEE Conference Publication. <https://ieeexplore.ieee.org/abstract/document/8892837>
- Bastiaan, M. (2015). *Preventing the 51%-Attack: A Stochastic Analysis of Two Phase Proof of Work in Bitcoin*. [Http://Refraat. Cs. Utwente. Nl/Conference/22/Paper/7473/Preventingthe-51-Attack-a-Stochasticanalysis-Of-two-Phase-Proof-of-Work-in-Bitcoin.Pdf](http://Refraat.Cs.Utwente.Nl/Conference/22/Paper/7473/Preventingthe-51-Attack-a-Stochasticanalysis-Of-two-Phase-Proof-of-Work-in-Bitcoin.Pdf)
- Bos, Halderman, Heninger, Moore, Naehrig, & Wustrow. (2014). Elliptic Curve Cryptography in Practice. In *International Conference on Financial Cryptography and Data Security*. Springer.
- Bralić, V., Stančić, H., & Stengard, M. (2020). A Blockchain Approach to Digital Archiving: Digital Signature Certification Chain Preservation. *Records Management Journal, ahead-of-print*(ahead-of-print). Advance online publication. doi:10.1108/RMJ-08-2019-0043

- Chang, S. E., Chen, Y. C., & Lu, M. F. (2019). Supply Chain Re-Engineering Using Blockchain Technology: A Case of Smart Contract Based Tracking Process. *Technological Forecasting and Social Change*, 144, 1–11. doi:10.1016/j.techfore.2019.03.015
- Chen, W., Zheng, Z., Ngai, E. C.-H., Zheng, P., & Zhou, Y. (2019). Exploiting Blockchain Data to Detect Smart Ponzi Schemes on Ethereum. *IEEE Access: Practical Innovations, Open Solutions*, 7, 37575–37586. doi:10.1109/ACCESS.2019.2905769
- Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient KNN Classification Algorithm for Big Data. *Neurocomputing*, 195, 143–148. doi:10.1016/j.neucom.2015.08.112
- Documents - Financial Action Task Force (FATF). (2020). <http://www.fatf-gafi.org/publications/financialinclusion/documents/bcbs-meeting-2-october-2014.html>
- Du Mingxiao, M. X., Zhe, Z., Wang, X., & Chen, Q. (2017). A Review on Consensus Algorithm of Blockchain. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 10.1109/SMC.2017.8123011
- Fahad. (2020). *Blockchain without Waste: Proof-of-Stake*. Academic Press.
- Farrugia, S., Ellul, J., & Azzopardi, G. (2020, April 13). Detection of Illicit Accounts over the Ethereum Blockchain. *Expert Systems with Applications*, 150, 113318. doi:10.1016/j.eswa.2020.113318
- Fergal, R., & Harrigan, M. (2013). An Analysis of Anonymity in the Bitcoin System. In *Security and Privacy in Social Networks* (pp. 197–223). Springer. doi:10.1007/978-1-4614-4139-7_10
- Gavin. (2014). *Ethereum: A Secure Decentralised Generalised Transaction Ledger*. Ethereum project yellow paper 151(2014): 1–32.
- Henderson. (2012). Rolx: Structural Role Extraction & Mining in Large Graphs. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1231–1239.
- Hu, Y., Seneviratne, S., Thilakarathna, K., Fukuda, K., & Seneviratne, A. (2019). *Characterizing and Detecting Money Laundering Activities on the Bitcoin Network*. arXiv preprint arXiv:1912.12060
- Israa, A., Rahwan, I., & Svetinovic, D. (2018). The Anti-Social System Properties: Bitcoin Network Data Analysis. *IEEE Transactions on Systems, Man, and Cybernetics. Systems*.
- Ittay, E., & Sirer, E. G. (2014). Majority Is Not Enough: Bitcoin Mining Is Vulnerable. In *International Conference on Financial Cryptography and Data Security*. Springer.
- Jethin, A., Higdon, D., Nelson, J., & Ibarra, J. (2018). *Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis*. SMU Data Science Review.
- Jouni, K., & Mills, C. (2020). On Group Comparisons with Logistic Regression Models. *Sociological Methods & Research*, 49(2), 498–525. doi:10.1177/0049124117747306
- Lahmiri, S., & Bekiros, S. (2019). Cryptocurrency Forecasting with Deep Learning Chaotic Neural Networks. *Chaos, Solitons, and Fractals*, 118, 35–40. doi:10.1016/j.chaos.2018.11.014

K-Nearest Neighbors Algorithm (KNN)

- Li, J., Li, N., Peng, J., Cui, H., & Wu, Z. (2019). Energy Consumption of Cryptocurrency Mining: A Study of Electricity Consumption in Mining Cryptocurrencies. *Energy*, 168, 160–168. doi:10.1016/j.energy.2018.11.046
- Liu, Z., Ren, C., & Cai, W. (2020). Overview of Clustering Analysis Algorithms in Unknown Protocol Recognition. In *MATEC Web of Conferences*. EDP Sciences. 10.1051/mateconf/202030903008
- Lu, H., Karimireddy, S. P., Ponomareva, N., & Mirrokni, V. (2020). Accelerating Gradient Boosting Machines. *International Conference on Artificial Intelligence and Statistics*, 516–526.
- Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., & Voelker, G. M. (2013). A Fistful of Bitcoins: Characterizing Payments among Men with No Names. *Proceedings of the 2013 Conference on Internet Measurement Conference*, 127–140. 10.1145/2504730.2504747
- Meneghetti, Sala, & Taufer. (2020). A Survey on PoW-Based Consensus. *Annals of Emerging Technologies in Computing*.
- Monamo, P., Marivate, V., & Twala, B. (2016). Unsupervised Learning for Robust Bitcoin Fraud Detection. 2016 Information Security for South Africa (ISSA), 129–134. doi:10.1109/ISSA.2016.7802939
- Nakamoto, S., & Bitcoin, A. (2008). *A Peer-to-Peer Electronic Cash System*. <https://bitcoin.org/bitcoin.pdf>
- O’Kane, E. (2018). *Detecting Patterns in the Ethereum Transactional Data Using Unsupervised Learning*. Academic Press.
- O’Leary, D. E. (2018). Open Information Enterprise Transactions: Business Intelligence and Wash and Spoof Transactions in Blockchain and Social Commerce. *Intelligent Systems in Accounting, Finance & Management*, 25(3), 148–158. doi:10.1002/isaf.1438
- Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A Survey of Machine Learning for Big Data Processing. *EURASIP Journal on Advances in Signal Processing*, 2016(1), 67. doi:10.1186/13634-016-0355-x
- Schonlau, M., & Yuyan Zou, R. (2020). The Random Forest Algorithm for Statistical Learning. *The Stata Journal*, 20(1), 3–29. doi:10.1177/1536867X20909688
- Stuart, Rosales, & Signorotti. (2015). Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis. *CS229 Project*, 1–5.
- Swan, M. Chapter Five - Blockchain for Business: Next-Generation Enterprise Artificial Intelligence Systems. In P. Raj & G. C. Deka (Eds.), *Advances in Computers, Blockchain Technology: Platforms, Tools and Use Cases* (pp. 121–162). Elsevier. <https://www.sciencedirect.com/science/article/pii/S0065245818300287>.2018
- Thai, P., & Lee, S. (2016). *Anomaly Detection in the Bitcoin System-a Network Perspective*. arXiv pre-print arXiv:1611.03942
- Van Engelen, J. E., & Hoos, H. H. (2020). A Survey on Semi-Supervised Learning. *Machine Learning*, 109(2), 373–440. doi:10.1007/10994-019-05855-6

Vujičić, D., Jagodić, D., & Randjić, S. (2018). Blockchain Technology, Bitcoin, and Ethereum: A Brief Overview. In *2018 17th International Symposium Infoteh-Jahorina (Infoteh)*. IEEE. 10.1109/INFOTEH.2018.8345547

Weber, I., Xu, X., Riveret, R., Governatori, G., Ponomarev, A., & Mendling, J. (2016). Untrusted Business Process Monitoring and Execution Using Blockchain. In M. La Rosa, P. Loos, & O. Pastor (Eds.), *Business Process Management* (pp. 329–347). Lecture Notes in Computer Science. Springer International Publishing. doi:10.1007/978-3-319-45348-4_19

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). *Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics*. KDD '19 Workshop on Anomaly Detection in Finance, Anchorage, AK.

Zambre, D., & Shah, A. (2013). *Analysis of Bitcoin Network Dataset for Fraud*. Unpublished Report 27.

ADDITIONAL READING

Alexander, C., & Dakos, M. (2020). A Critical Investigation of Cryptocurrency Data and Analysis. *Quantitative Finance*, 20(2), 173–188. doi:10.1080/14697688.2019.1641347

Patel, M. M., Tanwar, S., Gupta, R., & Kumar, N. (2020). A Deep Learning-Based Cryptocurrency Price Prediction Scheme for Financial Institutions. *Journal of Information Security and Applications*, 55, 102583. doi:10.1016/j.jisa.2020.102583

KEY TERMS AND DEFINITIONS

AML/CFT: Anti-money laundering and countering the financing of terrorism.

Blockchain: Is a technology for the storage and transmission of information, transparent, secure, and operating without a central control body.

Business Intelligence: Business intelligence, or BI, is a tool that allows the generation of reports in an automated way and in real time. This methodology is based on professional software solutions. The collection of data and their aggregation in readable documents give the management and operational functions the keys to guide the company's strategy.

Cryptocurrency: Is a currency issued on a peer-to-peer basis, without the need for a central bank, that can be used by means of a decentralized computer network.

Data Analysis: Is a family of statistical methods whose main characteristics are that they are multidimensional and descriptive.

FATF: Financial Action Task Force is an intergovernmental organism for the fight against money laundering and the financing terrorist.

Money Laundering: Money laundering is the action of concealing the origin of money acquired in illegal ways (illegal speculation, mafia activities, drug and arms trafficking, extortion, corruption, tax evasion, etc.) by reinvesting it in legal activities (trade, real estate construction, casinos, etc.).