



A Comparative Analysis and Prediction over Bitcoin Price Using Machine Learning Technique

Meenu Gupta¹, Riya Srivastava²

¹ Chandigarh University, INDIA

² Chandigarh University, INDIA

Emails: meenu.e9406@cumail.in; cu.17bcs1711@gmail.com

* Correspondence: meenu.e9406@cumail.in

Abstract

Bitcoin is one of the primary computerized monetary forms to utilize peer innovation to work with moment installments. The free people and organizations who own the overseeing figuring control and take part in the bitcoin network—bitcoin "miners"—are accountable for preparing the exchanges on the blockchain and are persuaded by remunerations (the arrival of new bitcoin) and exchange charges paid in bitcoin. These excavators can be considered as the decentralized authority implementing the believability of the bitcoin network. New bitcoin is delivered to the excavators at a fixed yet occasionally declining rate. There is just 21 million bitcoin that can be mined altogether. As of January 30, 2021, there are around 18,614,806 bitcoin in presence and 2,385,193 bitcoin left to be mined. This paper will predict the nature of bitcoin price because according to the reports of the past few years. The year 2020-present appeared to be a good time for bitcoin because, during this time duration, bitcoin has seen huge ups and downs. This paper will use various Machine Learning Techniques for the predictive analysis of bitcoin to accurately predict the price's nature. As the price of bitcoin depends upon various factors, and these factors directly affect the price, i.e., multiple factors of bitcoin are dependent on each other. After analyzing the results from multiple research papers and review papers, we discovered each algorithm has its advantages and disadvantages when predicting the bitcoin value. Keeping in mind all the findings, we will find algorithms that predict the bitcoin price accurately and without fewer disadvantages. So, if we go as per assumptions, regression would be the best choice for predicting the bitcoin value, but there are others algorithms also. So, in this paper, we will see the results of the multiple algorithms and then choose the correct algorithm after analyzing the results of all the implemented algorithms. This paper also includes the implementation of the comparison charts with each algorithm so that it will be easy to analyze the findings of each algorithm.

Keywords: Regression; Machine Learning; Bitcoin; Algorithms; Predictive Analysis; Accuracy; Exploratory Data Analysis

1. Introduction

Bitcoin is one of the booming concepts of the present day because of its value's massive ups and downs. These ups and downs depending on various factors. One of the factors is the media attention and the continued investments in bitcoin. And now some the companies, such as Tesla, have announced that they will accept payment via bitcoin [1-5]. This is also a factor in the increment of the price of bitcoin. If we see the comparison in the price of bitcoin over the last 20 years, we will find from the last few years that only the bitcoin price is moving upwards at full pace. Bitcoin is cryptographic cash applied globally for advanced installment or hypothesis purposes[6-12]. Bitcoin is decentralized as an instance. It isn't possessed by each person [13]. Exchanges made through Bitcoins are easy as

they are not connected to any use Speculation should be viable through excellent commercial enterprise facilities known as "bitcoin trades." These permit individuals to sell/buy Bitcoins utilizing various monetary forms [14].

One of the reasons for the getting this popularity of bitcoin is its security because it is challenging to penetrate the bitcoin system. Bitcoin's technology in cryptography works on the node-to-node system, making it very difficult to track down the payment path [15]. Here in cryptography, we don't have any servers. The payment is directly transferred from one person to another without the involvement of a third party. Bitcoin is one of the revolutionary changes in terms of payment. If we predict the popularity of bitcoin, we will find it in the near future, and it can be one of the payment options available for every payment [16].

Bitcoin is a cryptocurrency that uses the concept of cryptography. Here, we use a pointer to move from one node to another, and the node is open until the use. After that, it is locked and will only be accessed by the original user. If we compare the price of cryptocurrency with the starting and the present, we will be able to find that with how much percentage the bitcoin has seen the price hike [15]. For example, there is a case study of the person who has bitcoins and has ordered a pizza with that bitcoins, but now we have to order a pizza then we don't even require a single bitcoin. It can be bought with just 1/100 part of the bitcoin. This is the price difference I was talking about, and this price difference did not just come within 100 years or something. This price difference came within just 20-30 years.

In this research work, we will focus on different machine learning algorithms to predict bitcoin prices. We can have insights into various machine learning algorithms easily. A few machine learning algorithms we will be using are Regression, Classification, K-Mean Clustering, J48, and a few more [18]. With the help of these algorithms, we will try to find out the best algorithm from these algorithms, and that algorithm must be able to predict the bitcoin price accurately and be time-efficient. The resultant algorithm must have more advantages and fewer or no disadvantages. There will be various factors that will affect the price of bitcoin. We will have to consider all the dependent attributes that may have a hand in the cryptocurrency price fluctuation. We have also neglected the attributes that show the minor dependency. All these things will be managed under the pre-processing section of the topic [19, 26-28].

This paper is further classified into sections where section 2 discusses the related work of various researchers' predictions and analysis of bitcoin. The material, dataset, and methodology followed are discussed in section 3. Further, the evaluation of results in terms of training and testing used is discussed in section 4. Next, this work is concluded in section 5 with its future scope.

2. Related Work

Currency plays a vital role in the development of any country. If it is misused, then the global value of that currency is dropped down, resulting in GDP drop down and problems, so to overcome such problems, some researchers have proposed the following explanations.

In [17], the blockchain information, the authors develop the exchange organization and the client organization. This addresses the progression of bitcoins between exchanges, where every vertex addresses an exchange and each coordinated edge demonstrates whether there is an info/yield address that connects the exchanges. The last addresses the stream of bitcoin clients throughout the time. To develop the client organization, creators bunch addresses of a similar client expecting that all information locations of exchange have a place with a similar client. At that point, outer data on bitcoin addresses are acquired from current Internet assets (like Twitter posts, discussions, particular bitcoin applications - like bitcoin spigot) to help the bunching cycle and recognize the clients behind such groups.

In [21], Androulaky changes moves into gathering addresses. Considering a comparative idea, where all data areas of a comparable trade are assembled, they added another heuristic using the yield areas of trade. Expecting that most trades have only two yield addresses, if one of the two has successfully displayed in the blockchain, the other will be a shadow address. It can be clustered with the information addresses. Moreover, they apply direct-based packing systems, K-Means, besides Hierarchical Agglomerative Clustering, to further develop the gathering creation. In solicitation to perform such examination, the makers produce designed data from a particular explanation bitcoin test framework that they made. Data from the proliferation appreciates the advantage furthermore to gives a ground truth to evaluating their grouping measures. With this multiplication environment and the proposed systems, makers show that the profiles of 40% of bitcoin customers can be uncovered.

In [22] author has played out an assessment of bitcoin customer lead from the blockchain data rather than endeavoring to deanonymize customer information. Furthermore, they use the doubt that various data conveys have a spot with a comparative customer to depict the customer directly. They induce that until May 13, 2012, most of the new-made coins stayed unexpended in the printed addresses and that there were endless tiny trades that moved portions of bitcoins. In addition, they mindfully separate the most significant trades of the association until that second and give an unmistakable graph development of their turns of events.

In [23], they have played out an assessment of bitcoin customers direct from the blockchain data rather than attempting to deanonymize customer information. They also use the doubt that various data conveys have a spot with a comparative customer to portray customer direct. They construe that until May 13, 2012, most of the new-made coins stayed unexpended in the printed addresses and that there were valuable infinitesimal trades that moved portions of bitcoins. Furthermore, they warily separate the most significant trades of the association until that second and give an unmistakable graph development of their turns of events.

The problems were solved using the different machine learning algorithms [24-25]. There was a need to find such an algorithm capable of generating better accuracy with less time and space complexity. The other problem which was related to the algorithm for that paper firstly worked on anaconda and found out the list of algorithms, including the accuracy of that, and then switched the software and started working on weka to take a few more algorithms under consideration and weka is a toll with limited space, so the problem of space was also taken in consideration.

This paper has analyzed the data set of different persons. It uses that data against various machine learning algorithms using different platforms, which helps us identify which methodology can give us accurate results by applying different cross-validation points. It is generally challenging to find an algorithm that gives the best accuracy for the given dataset as more than one algorithm has the same accuracy.

3. Material and Methodology

The dataset used and the methodology used are explained in the subsequent sections.

A. Dataset

The dataset used in this paper for bitcoin evaluation is taken up from Kaggle. The dataset includes specific attributes which have certain information embedded in it. The included attributes are: - Data, High, Low, Open, Close, Volume, and Market Cap. Here, the Date includes the time when the crypto was purchased. It also includes the details regarding the investment made by persons on that day and on which crypto. The other attribute was Open, which includes the price of the crypto on that day while opened for purchasing. The others are high, low, and close. These are the attributes that all are related to the price of the crypto here. The high includes the highest price of the crypto on that day, the low includes the price of the crypto price on that day, and the last close includes the information of the crypto price at the end of the day. At last, the market cap includes the information regarding the investment made by the investors in the crypto [16].

B. Proposed Work

Data obtained from the Kaggle repository contains some missing values as it is data from the real world. The reason could be many such as data entry errors or problems related to data collection. The dataset is first cleaned, and the missing values are replaced with 0.0 to make the results more accurate. To implement the machine learning algorithm EDA, this work follows the following approach depicted in the figure.



Figure 1: Methods used for Bitcoin Price Prediction

Figure 1 represents how the operations are carried out on the dataset. Then on that dataset proposed novel finishes the algorithms both from weka and anaconda then the outcomes generated were within the shape of accuracy. Here paper used unique algorithms of weka in addition to anaconda but in weka used a way of move validation and in that got divided the complete dataset into groups and were given the accuracy than compared the accuracy of the algorithms from both software programs and in comparison them to get the required accuracy. Paper has implemented the pass validation in weka, intending to get the real consequences even after the distinctive information units. If the given records are in companies of 1, 2, 3, and with exceptional groups, accuracy may vary concerning the distinct algorithms. The matching algorithm could range using different companies; then, it can discover the set of rules whose cost doesn't alternate with the trade-in cost of move validation. Whereas on anaconda, it was now not that clean to divide the information into groups and after dividing it had been now not capable of getting the accuracy as per the desires in order in keeping with a look at it changed into precise that the weka having algorithm is giving an awful lot better accuracies than the anaconda. These algorithms are, which include:-

Ordinary Least Square

Ordinary \sum Conventional least squares, or direct least squares, gauge the boundaries in a relapse model by limiting the amount of the squared residuals. This strategy defines a boundary through the information focuses that limit the squared contrasts between the noticed qualities and the comparing fitted qualities.

(1)

$$m = \frac{\sum(xi-\bar{x})(yi-\bar{y})}{\sum(xi-\bar{x})^2}$$

$$b = \bar{y} - m * \bar{x}$$

Logistic Regression

It is a supervised machine learning algorithm. It defines the relationship between the one dependent binary variable and one or more nominal. It is a statistical data set for analyzing a dataset with more independent variables that determine an outcome. The outcome which is observed is most probably only two outcomes, not more than two. There should not exceed the possibility of more than two outcomes. The goal of this algorithm is to find the relationship between the dataset and the two outcomes only.

(2)

$$P(A/B) = P(B/A)P(A) / P(B)$$

IBK

It is java based algorithm that helps in finding accuracy. Class IBK is a machine learning algorithm from the Lazy package based on weka. It stands for Instance-Based K-Nearest Neighbor. The extended version of the k-nearest neighbor can efficiently find out the accuracy with the best results by deducting the required space to be used by the dataset.

It is quite similar to the k-nearest neighbor, but it efficiently finds the value of k (an unknown value), which constitutes classes already classified between classes. It uses the same method to classify the data points as we were calculating in the k-nearest neighbor. Still, here we are doing it inefficiently by utilizing less space than the k-nearest neighbor.

(3)

distances[i] = distances[i]*distances[i];

distances[i] = Math.sqrt(distances[i]/m_NumAttributesUsed);

Random Sub Space

Random SubSpace is a machine learning algorithm from weka. This class belongs to the Meta package. This algorithm is also known as attribute bagging or, in other ways, feature bagging. It helps in reducing the correlation between the estimators by training that dataset on the random sample chosen from the dataset. It selects the features from the dataset randomly.

In this algorithm, we combine the different models produced by different learners. We do such a combination because it gives better accuracy than the other original output, and the combining of these different learners is also known as bagging.

(4)

$$fb(L)(x^*) = 1/L \sum_{l=1}^L g^l(x^*)$$

Stacking

Stacking is a gathering learning technique to combine distinctive portrayal models through a meta-classifier. The individual course of action models are arranged ward on the complete getting ready set; by then, the meta-classifier is fitted ward on the yields - meta-features - of the individual portrayal models in the gathering. The meta-classifier can either be ready on the expected class names or probabilities from the company.

Input Mapped Classifiers

Wrapper classifier tends to incongruent preparing and test information by building planning between the preparation information that a classifier has worked with and the approaching test cases' design.

Model credits that are not found in the approaching cases get missing qualities, and so do approaching ostensible characteristic qualities that the classifier has not seen previously. Another classifier can be prepared or a current one stacked from a document.

C. Performing EDA and making predictions

Information Analysis is the insights and likelihood to sort out patterns in the informational index. It is utilized to show verifiable information by utilizing some examination apparatuses. It helps in penetrating down the data, to change measurements, realities, and considers along with development drives [28]. Here, the paper is going to perform EDA to determine the objectives of this paper. Simply defining, EDA is what analysts do with a large dataset to look for patterns and investigate and summarize the data's main characteristics. EDA is used primarily to find out what data can reveal beyond the formal modeling and provides a better understanding of the various variables of the data set and the relationship between them. EDA is a supplement to inferential insights, which will generally be genuinely unbending with rules and recipes. At a high level, EDA includes reviewing and depicting the informational collection from various points and afterward summing up it.

The purpose of exploratory data analysis is to:

- Check for missing information and different missteps.
- Gain the most extreme knowledge about the informational index and its hidden design.
- Uncover a tightfisted model, one which clarifies the information with a base number of indicator factors.

- Check suspicions related to any model fitting or theory test.
- Create a rundown of exceptions or different peculiarities.
- Find boundary gauges and their related certainty stretches or safety buffers.
- Identify the most influential factors.

In this section, after performing feature engineering, we will perform the EDA to understand the dataset and visualize the predictions.

4. Experimental Results and Discussion

After performing Exploratory Data Analysis on the collected dataset, the results and predictions are visualized with the help of graphs. For that, using different models of graphical views. Given below are the results after implementing the algorithm:

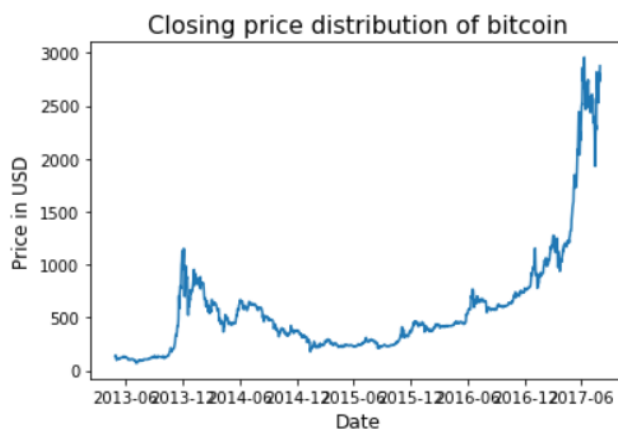


Figure 2: Closing price distribution to year

Figure 2 gives the price of bitcoin closed in the particular year from 2013-2017. The graph also signifies the sudden growth between 2016 and 2017, hitting the 3000 USD price; hence this paper focused on the increasing demands of cryptocurrency and sudden growth in the coming several years.

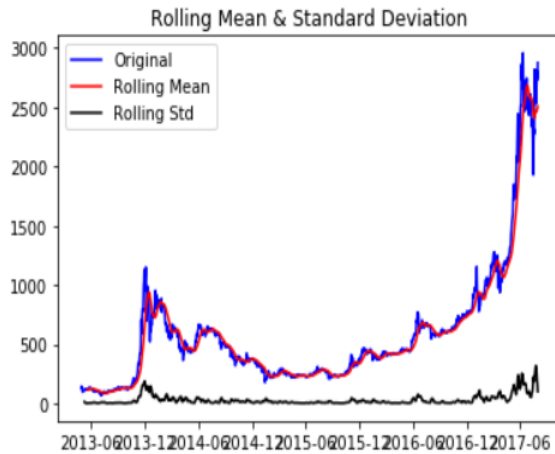


Figure 3: Representation and analysis of mean and standard deviation

Figure 3 describes the graphical representation of the closing prices of bitcoin with its mean and standard deviation. This graph shows how the original values and to mean coincide and tells the deviation for the year.

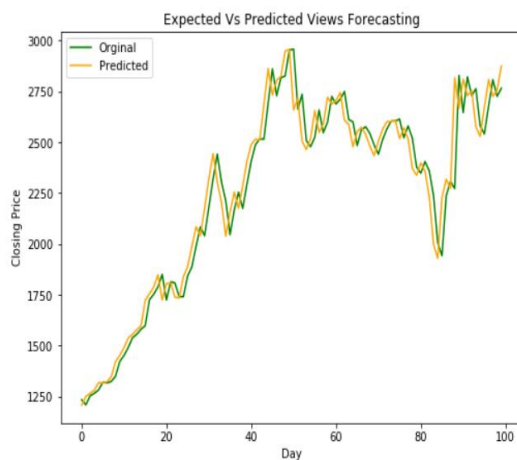


Figure 4: Predicted vs. expected closing price

Figure 4 represents the accuracy as the predicted values of bitcoin using several algorithms mostly coincide with the expected price of the dataset.

1. Accuracy comparison between Gretl and weka

In this paper, the different machine learning algorithms to get the best accuracy for which, we have used the below-mentioned algorithms.

The different algorithms used in this work are:-

- Logistic Regression.
- Ordinary Least Square.
- IBK.

- Random SubSpace.
- Stacking.
- Input Mapped Classifiers.

That Ordinary, Least Square, and Logistic Regression are the ones that give the highest accuracy in all of these algorithms. All of these are the best algorithm on there, but Ordinary, at least, is the one with the best accuracy for this particular dataset.

The values we got in the table show how the values will fluctuate as we change the algorithms. Here fluctuation can be termed as the change in accuracy. The below table shows the different algorithm and their accuracies with them.

To overcome the problem of fluctuation which we are facing and get the accuracy that doesn't change even if we change the value of clusters then in that case we are required to use the weka software to check for the different clusters used in the proposed algorithm and also to get them some new accuracies using some new algorithms to search for the algorithm having the better accuracy than we got in the Gretl. Thus, we have to use both the software to get the best accuracy as per the need, make the machine perfect for all the conditions, and check that the proposed algorithm gives the output as per the need.

Accuracy Table (Gretl): -

Table 1: Accuracy results of algorithms performed in Gretl

Accuracy	Methods Used
0.99	Ordinary Least Square
0.89	Logistic Regression

These algorithms are used, and Ordinary Least Square is the one that gives the highest accuracy in all of these algorithms.

All of these are the best algorithm on there, but Logistic Regression is the one with the best accuracy for this particular dataset. The values we got in the table show how the values will fluctuate as we change the algorithms. Here fluctuation can be termed as the change in accuracy. The below table shows the different algorithm and their accuracies with them.

Accuracy Table (Weka): -

Table 2: Accuracy results of algorithms performed in weka

Accuracy(10 Cross Folds)	Methods Used
30%	Stacking
27.7%	IBK
27%	Random Sub Space
30%	Input Mapped Classifier

The above table gives the accuracies from the weka software using different algorithms.

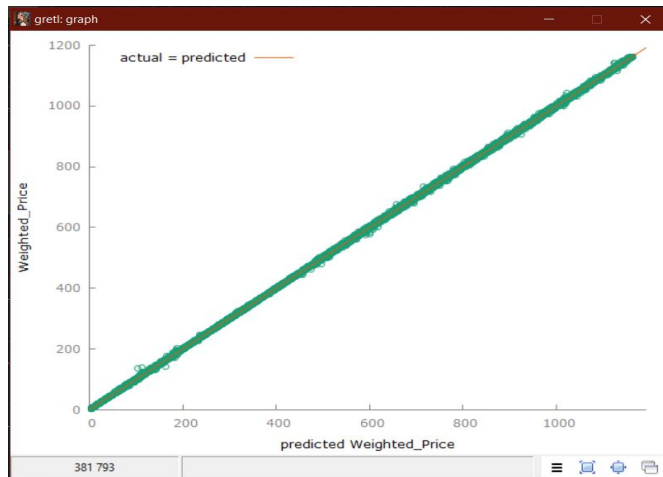


Figure 5: Actual vs. Predicted for Ordinary Least Square

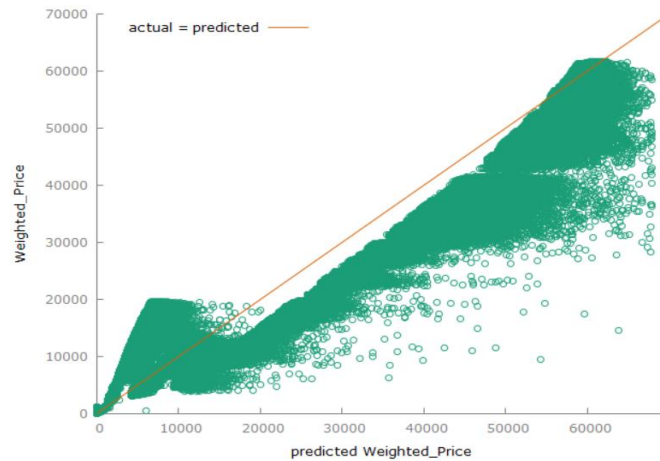


Figure 6: Actual vs. Predicted for logistic regression

Figure 3.4 and Figure 3.5 show the actual vs. predicted comparison, with an accuracy of 99% and 89%, respectively. Hence this data set gives the best accuracy with the Ordinary Least Square algorithm.

5. Conclusion and Future Scope

In this paper, we infer that the study report will simply present modules of Bitcoin value expectation and machine calculations. Hear the Comparison table of ML calculation model exactness which tells that the linear relapse model will have more precision than different calculations. In this paper, we infer that the straightforward relapse calculation is more productive than different calculations by taking assistance from that straightforward relapse calculation. The AI calculations will improve that include thought of digital currencies. That will improve the market cost of global ventures. In this paper, we proposed a new calculation to discover the component value precision. That aids the client's augmentations and benefits. The novel resides the algorithm which gives the maximum accuracy is ordinary least square, and after that, it is logistic regression. This paper concludes that the regression-based algorithm can be best suited to implement the bitcoin price prediction with the best accuracy by examining this entire algorithm.

REFERENCES

- [1] Velankar, S., Valecha, S., & Maji, S. (2018, February). Bitcoin price prediction using machine learning. In 2018 20th International Conference on Advanced Communication Technology (ICACT) (pp. 144-147). IEEE.
- [2] L. Tan, K. Yu, N. Shi, C. Yang, W. Wei, and H. Lu, "Towards Secure and Privacy-Preserving Data Sharing for COVID-19 Medical Records: A Blockchain-Empowered Approach," *IEEE Transactions on Network Science and Engineering*,
- [3] L. Tan, N. Shi, K. Yu, M. Aloqaily, Y. Jararweh, "A Blockchain-Empowered Access Control Framework for Smart Devices in Green Internet of Things", *ACM Transactions on Internet Technology*, vol. 21, no. 3, pp. 1-20, 2021,
- [4] K. Yu, L. Tan, M. Aloqaily, H. Yang, and Y. Jararweh, "Blockchain-Enhanced Data Sharing with Traceable and Direct Revocation in IIoT", *IEEE Transactions on Industrial Informatics*
- [5] K. Yu, L. Tan, X. Shang, J. Huang, G. Srivastava, and P. Chatterjee, "Efficient and Privacy-Preserving Medical Research Support Platform Against COVID-19: A Blockchain-Based Approach", *IEEE Consumer Electronics Magazine*,
- [6] L. Tan, H. Xiao, K. Yu, M. Aloqaily, Y. Jararweh, "A Blockchain-empowered Crowdsourcing System for 5G-enabled Smart Cities", *Computer Standards & Interfaces*,

- [7] C. Feng et al., "Efficient and Secure Data Sharing for 5G Flying Drones: A Blockchain-Enabled Approach," *IEEE Network*, vol. 35, no. 1, pp. 130-137, January/February 2021
- [8] N. Shi, L. Tan, W. Li, X. Qi, K. Yu, "A Blockchain-Empowered AAA Scheme in the Large-Scale HetNet", *Digital Communications and Networks*,
- [9] Z. Guo, L. Tang, T. Guo, K. Yu, M. Alazab, A. Shalaginov, "Deep Graph Neural Network-based Spammer Detection Under the Perspective of Heterogeneous Cyberspace", *Future Generation Computer Systems*, <https://doi.org/10.1016/j.future.2020.11.028>.
- [10] Z. Guo, Y. Shen, A. K. Bashir, M. Imran, N. Kumar, D. Zhang, and K. Yu, "Robust Spammer Detection Using Collaborative Neural Network in Internet of Thing Applications", *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9549-9558, June 15, 2021
- [11] K. Yu, L. Tan, X. Shang, J. Huang, G. Srivastava, and P. Chatterjee, "Efficient and Privacy-Preserving Medical Research Support Platform Against COVID-19: A Blockchain-Based Approach", *IEEE Consumer Electronics Magazine*,
- [12] Z. Guo, K. Yu, A. Jolfaei, A. K. Bashir, A. O. Almagrabi, and N. Kumar, "A Fuzzy Detection System for Rumors through Explainable Adaptive Learning", *IEEE Transactions on Fuzzy Systems*,
- [13] Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365, 112395.
- [14] Phaladisailoed, T., & Numnonda, T. (2018, July). Machine learning models comparison for bitcoin price prediction. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 506-511). IEEE.
- [15] <https://www.kaggle.com/prakhargupta231/bitcoin-price-arima/edit>
- [16] Ji, S., Kim, J., & Im, H. (2019). A comparative study of bitcoin price prediction using deep learning. *Mathematics*, 7(10), 898.
- [17] Azari, A. (2019). Bitcoin price prediction: An ARIMA approach. *arXiv preprint arXiv:1904.05315*.
- [18] M. Daniela and A. BUTOI, —Data mining on Romanian stock market using neural networks for price prediction. *Informatica Economica*, 17,2013.
- [19] Herrera-Joancomartí J. (2015) Research and Challenges on Bitcoin Anonymity. In: Garcia-Alfaro J. et al. (eds) *Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance. DPM 2014, QASA 2014, SETOP 2014. Lecture Notes in Computer Science*, vol 8872. Springer.
- [20] Androulaki, E., Karame, G., Roeschlin, M., Scherer, T., Capkun, S.: Evaluating user privacy in bitcoin. In Sadeghi, A.R., ed.: *Financial Cryptography and Data Security. Volume 7859 of Lecture Notes in Computer Science*. Springer Berlin Heidelberg (2013)
- [21] Ron, D., Shamir, A.: Quantitative analysis of the full bitcoin transaction graph. In Sadeghi, A.R., ed.: *Financial Cryptography and Data Security. Volume 7859 of Lecture Notes in Computer Science*. Springer Berlin Heidelberg (2013)
- [22] Spagnuolo, M., Maggi, F., Zanero, S.: Bitiodine: Extracting intelligence from the bitcoin network. In Christin, N., Safavi-Naini, R., eds.: *Financial Cryptography and Data Security. Volume 8437 of Lecture Notes in Computer Science*. Springer Berlin Heidelberg (2014)
- [23] Rane, P. V., & Dhage, S. N. (2019, March). Systematic erudition of bitcoin price prediction using machine learning techniques. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)* (pp. 594-598). IEEE.
- [24] Madan, I., Saluja, S., & Zhao, A. (2015). Automated bitcoin trading via machine learning algorithms. URL: [http://cs229.stanford.edu/proj2014/Isaac% 20Madan, 20](http://cs229.stanford.edu/proj2014/Isaac%20Madan,20).
- [25] <https://www.kaggle.com/mczielinski/bitcoin-historical-data>
- [26] Lamothe-Fernández, P., Alaminos, D., Lamothe-López, P., & Fernández-Gámez, M. A. (2020). Deep learning methods for modeling bitcoin price. *Mathematics*, 8(8), 1245.
- [27] Awoke, T., Rout, M., Mohanty, L., & Satapathy, S. C. (2021). Bitcoin price prediction and analysis using deep learning models. In *Communication Software and Networks* (pp. 631-640). Springer, Singapore.
- [28] Dutta, A., Kumar, S., & Basu, M. (2020). A gated recurrent unit approach to bitcoin price prediction. *Journal of Risk and Financial Management*, 13(2), 23.