Predicting Bitcoin Volatility Using Machine Learning Algorithms and Blockchain Technology

Mimoza Mjoska [1], Blagoj Ristevski [1], Snezana Savoska [1] and Vladimir Trajkovik [2]

[1] *Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", ul. Partizanska bb, Bitola, 7000 RN Macedonia*
[2] *Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, RN Macedonia*

Abstract

Blockchain technology has the potential to be applied in a variety of areas of our daily life. Blockchain is the foundation of cryptocurrency, but the applications of blockchain technology are much more expansive. This technology is considered to be a revolutionary solution for the financial industry. Also, it can be successfully applied in scenarios involving data validation, auditing, and sharing. On the other hand, machine learning is one of the most noticeable technologies in recent years. Both technologies are data-driven, and thus there are rapidly growing interests in integrating them for more secure and efficient data sharing and analysis. This paper shows how these two technologies, blockchain and machine learning, can be combined in predicting bitcoin volatility. To analyze and predict bitcoin volatility, bitcoin data from real-time series and random forests as a machine learning algorithm were used. When predicting bitcoin volatility, low statistical errors were obtained in the training and test set. This confirms that the forecasting model is well designed.

Keywords:

blockchain technology, machine learning algorithms, random forests, bitcoin, time-series data.

1. Introduction

In 2008, powerful US financial institutions and insurance companies were on the edge of bankruptcy. These circumstances called for immediate intervention by the federal government to avoid domestic and possibly global financial col-

Information Systems & Grid Technologies: Fifteenth International Conference ISGT'2022, May 27–28, 2022, Sofia, Bulgaria
EMAIL: mijoska.mimoza@uklo.edu.mk (M. Mjoska); blagoj.ristevski@uklo.edu.mk (B. Ristevski); snezana.savoska@uklo.edu.mk (S. Savoska); vladimir.trajkovik@finki.ukim.mk (V. Trajkovik)
ORCID: 0000-0002-4248-2760 (M. Mjoska); 0000-0002-8356-1203 (B. Ristevski); 0000-0002-0539-1771 (S. Savoska); 0000-0001-8103-8059 (V. Trajkovik)

lapse. Moreover, these events illustrated the dangers of living in a digital, interconnected world that depends on transactional intermediaries and leaves people vulnerable to digital exploitation, greed, and crime.

Blockchain technology is a fast-growing part of the Internet and cloud computing. Similar data structures existed long before the famous bitcoin was conceived, but the main theories about blockchain architectures used today were originally defined in the original article on bitcoin written and published by a person under the pseudonym Satoshi Nakamoto in 2008 [1].

Nowadays, we live in a world where an abundance of data surrounds us. Therefore, it is interesting to develop software systems that can use this information, learn from it, and offer valuable behaviours based on it. This is possible by using machine learning algorithms.

Realized volatility measures the actual performance of an asset in the past and helps to understand the stability of the asset based on its past performance. It indicates how an asset's price has changed in the past and the period in which it has changed. Higher the volatility, the higher the price risk associated with the stock and, therefore, the higher the premium attached to the stock. The realized volatility of the asset may be used to forecast future volatility, i.e., implied volatility of the asset. While entering into transactions with complex financial products such as derivatives, options, etc., the premiums are determined based on the underlying volatility and influence the prices of these products [26].

This paper describes an algorithm that combines blockchain technology and machine learning algorithms to predict bitcoin volatility. The remainder of the paper is organized as follows. Section 2 highlights the principles of blockchain technology and its application. Section 3 elaborates on the machine learning algorithms focusing on the random forest algorithm. Next, an application in the programming language R is used to predict the bitcoin volatility, as described in section 4. Finally, concluding remarks and directions for further works are given in the last section.

2. Blockchain technology

"Blockchain" is a coined word composed of the words "block" and "chain". Blockchain is a distributed replicated database organized in the form of a single linked list – chain, where nodes are blocks of data for transactions. These blocks, after grouping, are protected by using cryptographic methods [3][4]. Blockchain technology enables the accomplishment of digital transactions without intermediaries.

Blockchain (distributed ledger technology) is a network software protocol that enables the secure transfer of money, assets, and information through the Internet without a third-party organization as an intermediary [3]. It can safe-

ly store transactions such as digital cryptocurrencies or data/information about debt, copyrights, equity, and digital assets. Furthermore, the stored data cannot be easily forged and tampered with because it requires individual approval of all distributed nodes. This significantly reduces the cost of trust and accounting in non-digital economies and other social activities [2].

Blockchain is a technology that is constantly evolving. The most common types of blockchains are:

- public Blockchain [6],
- private Blockchain and
- hybrid Blockchain [10].

Researchers pointed out that applying blockchain technology to cross-border payment has a high potential effect. Holotiuk et al. in [14] stated that Blockchain technology will improve the payment system by providing a solid structure for cross-border transactions, removing expensive intermediary costs, and gradually weakening or altering the business model of the existing payment industries [11].

Because modifying transactions or whole blocks is almost impossible in blockchain, blockchain technology makes it easy to prove the integrity of the electronic files used in accounting and auditing[13]. This technology is considered to be a revolutionary solution for the financial industry. Also, this technology is the basis of the digital currency bitcoin whose volatility is analyzed in this paper.

3. Machine learning algorithms

Among the numerous applications of machine learning in healthcare, science, industry, etc., is the timely detection of diseases such as cancer, glaucoma and other conditions that take human lives at high speed. Another application is the visualization of smart cars, efficient web browsing that facilitates Internet searches, language translations that help immensely in world communications and limit the significant language barrier between countries, and the implementation of fraud detection and face-recognition systems. Well-known machine learning algorithms are the k-nearest neighbors (k-NN) algorithm, artificial neural networks, random forests, support vector machines, the Naïve Bayesian classifier, etc. Machine learning systems can be classified according to the amount and type of supervision they receive during the training. There are three main categories: supervised learning [18], unsupervised learning [16], and reinforcement learning [20] [16].

3.1. Random forests

The Random Forest is a supervised learning algorithm used for regression and classification. Technically, it is an ensemble algorithm. The algorithm generates individual decision trees through an attribute selection indication. Each tree relies

on an independent random sample [19]. In a classification problem, each tree votes, and the most popular class is the end result. On the other hand, in a regression problem, the average of all trees' outputs is calculated, which will be the result [19]. This algorithm is used to predict the realized volatility of Bitcoin, in this paper.

4. Discussion and analysis of the results obtained in creating a model that overlooks the time series of realized volatility of the market price of bitcoin

4.1. Realized volatility

Volatility forecasting plays a critical role in financial modelling and financial decision-making. Realized volatility assesses variation in returns for an investment product by analyzing its historical returns within a defined period [26]. Assessment of the degree of uncertainty and/or potential financial loss/gain from investing in a company may be measured using variability/volatility in the stock prices of the entity. In statistics, the most common measure to determine variability is by measuring the standard deviation, i.e., the variability of returns from the mean. It is an indicator of the actual price risk [26].

The realized volatility or actual volatility in the market is caused by two components – a continuous volatility component and a jump component, which influence the stock prices. Continuous volatility in a stock market is affected by intra-day trading volumes. For example, a single high-volume trade transaction can introduce a significant variation in the price of an instrument [26].

This paper predicts the realized volatility of bitcoin. Analysts use High-variance daily data to estimate hourly/daily/weekly or monthly frequency levels. The data can then be used to estimate volatile sales movement. During the analysis, data whose frequency is 1 hour from the Gemini platform were taken [21], and then using that data, the realized volatility is calculated with a daily frequency. There is OHLC (Open / High / Low / Close) pricing data in each file that is updated daily. For this paper, granular hourly data are taken back to the 2015 year, for the pair of bitcoin/dollar. The estimation of variability is calculated by measuring the standard deviation from the average price of the monitored object in a given period. Because volatility is nonlinear, the realized variance is first measured by translating the values taken from the stock market into logarithmic values and then calculating the standard deviation. The realized variance is calculated by calculating the sum of the squares of the standard deviation. The next step is to calculate the realized volatility, which is the square root of the realized variance:

$$\text{realized\_volatility}_t = \sqrt{\sum_{i=1}^{n} r_t^2} \tag{1}$$

To calculate the realized volatility of bitcoin, an application was created in the programming language R. For the necessary analysis, the package Rstudio Version 1.4.1717 for ubuntu 20.04 and the necessary packages *dplyr, forecastML, ggplot2, glmnet, DB, randomForest* and *caret* were used. The application first loads the hourly market price of bitcoin downloaded from the Gemini platform using the read.csv() function.

A date sequence is then added using the seq function, defining the start and end time points with a frequency of 1 hour. In the time series "2015-10-08 13:00:00" is taken as the starting date, and "2022-01-12 12:00:00" is taken as the end date. The price of bitcoin was taken at the close of the calculations, and other data were omitted. The logarithmic values of the bitcoin price are calculated to make the results more accurate. Using the library (*dplyr*) are created a new column with a unique date for each day. It is an identification column, so later, using the *group_by*, *sum* and *arrange* functions, the squares of the standard deviation corresponding to one date for one day are summed so that we get the realized variance. Using the *sqrt()* function, the realized volatility in daily frequency (1) is calculated.

For the needs of the research, free data is downloaded from the website https://www.blockchain.com/charts in .csv format for the last 3 years for the properties shown in Table 1.

Table 1
Properties of bitcoin are downloaded from https://www.blockchain.com/charts

Variable	Description
cost_per_transaction	Miners' income is divided by the number of transactions.
cost_per_transaction_percent	Miners' income as a percentage of transaction volume.
estimated_transaction_volume_usd	The total estimated value in US dollars of blockchain transactions.
miners_revenue	The total US value of block rewards and transaction fees paid to miners.
n_transactions	The total number of confirmed transactions per day.
n_transactions_per_block	The average number of transactions per block in the past 24 hours.
output_volume	The total value of all outgoing transactions per day.
trade_volume	The total value in US dollars on the trading volume of major bitcoin exchanges.
transaction_fees_usd	The total value in US dollars of all transaction fees paid to the miners.

All the data for the properties with the *read.csv()* command are loaded into separate variables in the R programming language. Then all the data is loaded into a data table with the command *data.frame()*. Finally, the volatility of bitcoin is added to that data, which was previously calculated with data downloaded from the cryptocurrency exchange platform www.gemini.com.

Once the database is complete, some transformations need to be made to balance the data to begin data analysis. Each vector x is replaced by its natural logarithm with function in R, log(x). This transformation is done so that the results of statistical analysis can become more accurate.

To prepare the data for machine learning, the data is normalized. The purpose of the normalization is to adjust the values of the numeric columns in the database to a standard scale, without distorting the variations in the range of values. The min-max normalization method is selected. The data is normalized to have values between 0 and 1.

A machine learning algorithm randomForest is used to predict the realized bitcoin volatility. The random forest algorithm creates a large number of decision trees during training. From the created set of decision trees, the random forest method gets a prediction from each tree individually and the average of all trees' outputs is calculated, which will be the end result.

The algorithm for predicting bitcoin volatility uses the forecastML package in the programming language R. When machine learning algorithms are used, the model is first generated using training data, and then the values for the test data are predicted. Once the data is ready, the next step in the research is to divide the data into a training and test set. The data used to predict the volatility of bitcoin contain 1095 observations, starting from 14.01.2019 to 12.01.2022. The first 915 observations are taken for the training set, and the remaining 180 observations are used as a test set.

The forecasting method uses four different forecasting horizons. These different horizons are used to predict in the short and long term to combine the predictions in the final forecast and thus minimize the error. The randomForest function is then defined with its arguments.

The first step in the prediction process is to create some validation windows to perform nested cross-validation. After training the model the predictions, residuals, and some error metrics are presented. The test set is then predicted using the validation windows, and the actual versus predicted values are displayed. Initially, the size of each forecast horizon is defined. The first horizon is seven steps ahead, the second is 30 steps forward, the third is 90 steps forward, and the last is 180 steps forward. The horizons to be predicted are chosen, and the view through certain time steps in the past is also chosen.

Ten validation windows are created. This means that ten models will need to be trained for each direct forecast horizon, each theoretically selecting differ-

ent optimal hyperparameters and having different coefficients from the internal cross-validation process [27]. Estimating variations between these models is a good way to estimate stability under the dynamics of different time series in a given modeling method.

In the given research the elbow method is used to select which number of trees will be used in the given test. After approximately 100 trees, there is almost no difference in error reduction. That is why it is chosen to use 100 trees in this research.

The importance of the variables is used so that we can understand which variables affect the volatility of bitcoin.

The importance of the variables used in the forecast is presented in Figure 1.

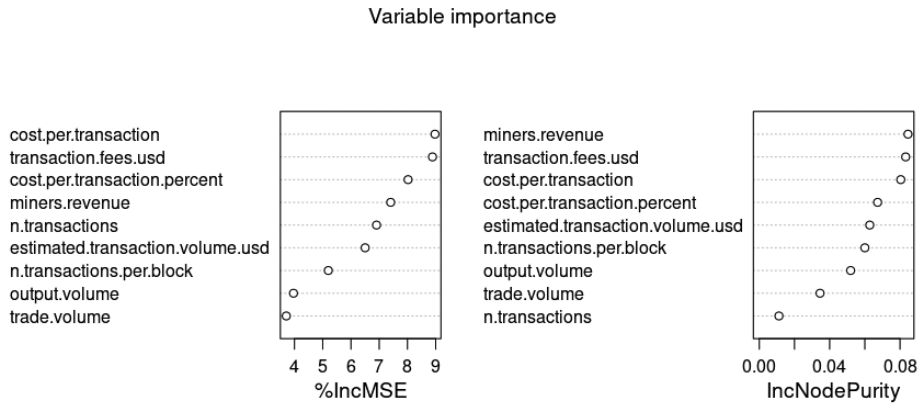

Figure 1: Importance of the properties of bitcoin

From Figure 1, it can be seen that of the selected features, cost.per.transaction is the most important in increasing the percentage of the mean square error (% IncMSE), and miners.revenue is the most important in increasing the node purity (IncNodePurity).

The following Figure 2 shows the standard errors in the training set:

Show 10 entries Search:

	model	window_start	window_stop	mae	mape	smape
1	RandomForest	2019-07-13	2019-07-13	0.019	74.008	52.231

Figure 2: Standard set errors in the training set

The next step is to predict the test set. The error metric for the test set by horizons, the predicted values versus the actual values, is presented in Figure 3.

Show 10 entries Search:

	model	model_forecast_horizon	mae	mape	mdape	smape
1	RandomForest	7	0.015	62.141	72.31	44.586
2	RandomForest	30	0.014	47.969	43.177	37.13
3	RandomForest	90	0.017	61.183	57.476	43.778
4	RandomForest	180	0.015	57.589	46.174	40.957

Figure 3: Standard errors in the test set

Compared to the training set error metric, the test set error metric is smaller. The error values are satisfactory, indicating that the model is good.

In the second part of the analysis, it is necessary to train the model throughout the entire training database without nested cross-validation. Without nested cross-validations and retention windows, the forecast graph basically fits the model.

The next step is to predict the test set and re-display the real versus predicted values and error metrics for the test set.

The predictions on each horizon of the predicted values of the test set are then combined. The final part is to predict the off-sample for each horizon using the training and test set and re-combine the out-of-sample forecasts for each horizon to display the final combined forecast.

A new model is now being re-trained using only one validation window.

The standard prediction errors in the training set with a single validation window are shown in Figure 4.

Show 10 entries Search:

	model	window_start	window_stop	mae	mape	mdape	smape
1	RandomForest	2019-07-13	2019-07-13	0.006	21.007	13.638	18.084

Figure 4: Standard errors in the training set generated by R

The table from Figure 5 shows the error metrics of the predicted values in the test set without validation windows.

Show 10 entries Search:

	model	mae	mape	mdape	smape
1	RandomForest	0.015	55.515	44.946	40.634

Figure 5: Standard errors in the test set

If we compare the data in Figure 3 and Figure 5, it is noticed that the test set is error is more significant than the training set. That result is expected because test data are not used when the model is being trained. The fact that the error in the test set does not become extremely large indicates that our model predicts well. The mean absolute error in the training set is 0.006, and the test set is 0.015, proving that the model has a good prediction. According to the random forest algorithm, the next step is to combine the predictions of each direct forecast horizon using the *combine_forecasts ()* function.

The next step according to the random forest algorithm is to make an off-sample prediction, which is presented in Figure 6.

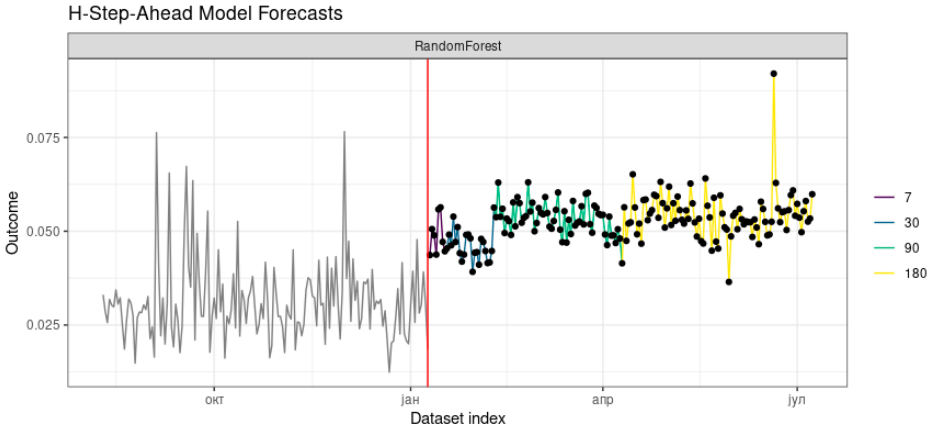

Figure 6: Prediction outside the sample Predict the horizons of each out-of-sample horizon again using the combine_forecasts() function

The obtained combined forecast scheme outside the sample for 180 steps forward is presented in Figure 6. The command *combine_forecasts()* predicts 180 steps forward, where different predictions are combined to get a better result. Using the standard *type = "horizon",* forecasts are combined so that short-term forecasts are produced from short-term models and long-term forecasts from the long-term model.

5. Conclusion

Bitcoin is the most popular cryptocurrency, nowadays. It's widely researched in economics and computer science. This paper uses a random forest machine-learning algorithm to predict the time series of realized volatility of the bitcoin. This algorithm is one of the best machine learning algorithms and predicts very well with high accuracy, gives better and more accurate results than traditional

predictions, such as autoregressive and vector autoregressive methods. This paper also examines whether the information obtained from blockchain can be used to predict the volatility and price of bitcoin. Many people in the world use bitcoin as an investment due to its high volatility, and in this way, they can get enormous profits for a certain period. The R package uses the forecastML library, which contains many time-series modeling visualizations. It is concluded that blockchain information may be included as a variable in further research into the price and volatility of Bitcoin.

According to the literature the smaller the mean absolute error, the better the prediction model. In our model mean absolute error (mae) in the training set using 10 validation windows is 0.019, and in the test set, the mean absolute error from the different horizons is 0.015. The mae in the training set without validation windows is 0.006, and in the test set is 0.015. The paper uses data on variables only from blockchain information and concludes that the obtained empirical results correspond to the literature that blockchain information can be used to examine the price and volatility of bitcoin. From this research, we can conclude that if in the real data the volatility of bitcoin is big, then the prediction error will be bigger. From the example shown, it can be concluded that the error in predicting the data is smaller when the forecast horizon is shorter and when volatility is smaller.

For further research different numbers of trees in the random forest, and different forecast horizons can be used. In addition, different machine learning algorithms can be used for further work. In this way, the accuracy of different time series algorithms can be tested and the one with the best results and accuracy can be selected for prediction. To determine the error with which they predict and compare that error to choose which machine learning algorithm is most suitable to predict the relevant data.

6. References

[1] Satoshi Nakamoto (2008). "Bitcoin: A Peer-to-Peer Electronic Cash System".

[2] Dragana Tadić Živković **(**2018). "Blockchain technology: opportunity or a threat to the future development of banking", Proceedings of *Ekonbiz*.

[3] Swan M. (2015). "Blockchain: Blueprint for a new economy", https://www.goodreads.com/work/best_book/44338116-blockchain-blueprint-for-a-new-economy (last accessed 06/10/2021).

[4] Mijoska M. and Ristevski B. (2020). "Blockchain Technology and its Application in the Finance and Economics", 10th International Conference on Applied Information and Internet Technologies – AIIT 2020, October, Zrenjanin, Serbia.

[5] Li Zhang, Yongping Xie, Yang Zheng, Wei Xue, Xianrong Zheng, Xiao-

bo Xu (2020). “The challenges and countermeasures of blockchain in finance and economics”, John Wiley & Sons, Ltd. https://doi.org/10.1002/sres.2710.
[6] Dejan Vujicic, Dijana Jagodic, Siniša Ranđić (2018). “Blockchain technology, bitcoin, and Ethereum: A brief overview”, Available online: https://doi.org/10.1109/INFOTEH.2018.8345547.
[7] Julija Basheska, Vladimir Trajkovik (2018). “Blockchain based Transformation in government: review of case studies”, ETAI 2018.
[8] Nick Szabo (1997). “The idea of smart contracts”. Nick Szabo’s Papers and Concise Tutorials. https://fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/idea.html (last accessed 06.10.2021).
[9] Aleksandar Matanović, “Osnove kriptovaluta i blokčein tehnologije”, http://fzp.singidunum.ac.rs/demo/wp-content/uploads/Osnove-kriptovaluta-i-blok%C4%8Dein-tehnologije.pdf, (last accessed 06.10.2021).
[10] Gu, J.; Sun, B.; Du, X.; Wang, J.; Zhuang, Y.; Wang, Z. (2018). “Consortium blockchain-based malware detection inmobile devices”, IEEE.
[11] https://doi.org/10.1088/1742-6596/1693/1/012025.
[12] Androulaki, E.; Barger, A.; Bortnikov, V.; Cachin, C.; Christidis, K.; De Caro, A.; Enyeart, D.; Ferris, C.; Laventman, G.; Manevich, Y.; et al. (2018). “Hyperledger fabric: A distributed operating system for permissioned blockchains”. In Proceedings of the Thirteenth EuroSys Conference ACM, Porto, Portugal, 23–26 April; pp. 1–15. https://doi.org/10.1145/3190508.3190538.
[13] Hyperledger Burrow – Hyperledger. Available online: https://www.hyperledger.org/projects/hyperledger-burrow (last accessed 06.10.2021).
[14] Quora (2017). “How will blockchain impact accounting, auditing & finance?”, https://www.quora.com/How-will-blockchain-impact-accounting-auditing-finance, (last accessed 06.10.2021).
[15] Holotiuk, F., Pisani, F., & Moormann, J. (2017). The impact of blockchain technology on business models in the payments industry. Wirtschaftsinformatik 2017 Proceedings. Retrieved https://wi2017.ch/images/wi2017-0263.pdf (last accessed 06.10.2021).
[16] Crosman, P. (2017). “R3 to take on Ripple with cross-border payments blockchain”. American Banker; New York, N.Y. Retrieved from https://www.bitcoinisle.com/2017/10/31/r3-to-take-on-ripple-with-cross-border-payments-blockchain/ (last accessed 06.10.2021).
[17] Aurélien Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, Published by O’Reilly Media, 2017.
[18] Molly Galetto, Machine learning and big data analytics: the perfect marriage http://www.ngdata.com/machine-learning-and-big-data-analytics-the-perfect-marriage.

[19] Sotiris B Kotsiantis, Ioannis D Zaharakis, and P.E.Pintelas. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3): 159–190, 2006.
[20] Pavan Vadapalli, Random Forest Classifier: Overview, How Does it Work, Pros & Cons, https://www.upgrad.com/blog/random-forest-classifier (last accessed 08.07.2021).
[21] Marko ˇCupić, Umjetna inteligencija, Uvod u strojno uˇcenje, 2020, http://java.zemris.fer.hr/nastava/ui/ml/ml-20200410.pdf.
[22] Gemini, https://www.cryptodatadownload.com/data/gemini.
[23] Ross Jacobucci, Random Forests, University of Notre Dame, https://statisticalhorizons.com/wp-content/uploads/2021/11/Advanced-Machine-Learning.pdf.
[24] Daniel Johnson, R Random Forest Tutorial with Example, 2022, https://www.guru99.com/r-random-forest-tutorial.html (last accessed 11.04.2022).
[25] Jeffrey Craig, What is Transactions Per Second (TPS): A Comparative Look at Networs,2021, https://phemex.com/blogs/what-is-transactions-per-second-tps (last accessed 11.04.2022).
[26] Miroslav Minovic, Blockchain technology: usage beside crypto currencies, (2017). Available online: https://www.researchgate.net/publication/318722738.
[27] Madhuri Thakur, Realized volatility, https://www.wallstreetmojo.com/realized-volatility/ (last accessed 08.04.2022).
[28] Aristeidou Christoforos**,** Study of the volatility of Bitcoin cryptocurrency using Machine Learning methods: An implementation in R, 2020.