http://doi.org/10.35784/iapgos.5657 received: 08.12.2023 | revised: 10.05.2024 | accepted: 22.06.2024 | available online: 30.06.2024

EVALUATING THE PERFORMANCE OF BITCOIN PRICE FORECASTING USING MACHINE LEARNING TECHNIQUES ON HISTORICAL DATA

Mamun Ahmed[1], Sayma Alam Suha[2], Fahamida Hossain Mahi[1], Forhad Uddin Ahmed[1]

[1]Bangladesh Army International University of Science & Technology, Computer Science & Engineering, Cumilla, Bangladesh
[2]Bangladesh University of Professionals, Department of Computer Science & Engineering, Dhaka, Bangladesh

Abstract. *Since entering the market in 2009, Bitcoin has had a price that is extremely erratic. Its price is influenced by factors such as adoption rates, regulatory changes, geopolitical occurrences, and macroeconomic developments. Experts believe that Bitcoin's price will rise in the long run due to limited supply and rising demand. Therefore, the aim of this study is to propose an ensemble feature selection and machine learning-based approach to predict bitcoin price. For this research purpose, the cryptocurrency-based dataset has been used, visualized, and preprocessed. Five different feature selection approaches (Pearson, RFE, Embedded Random Forest, Tree-based and Light GBM) are followed by ensemble methodology, with the maximum voting approach to extract the most significant features and generate a dataset with reduced attributes. Then the dataset with or without feature selection is used for bitcoin price prediction by applying ten different machine learning regressing models, which includes six traditional, four bagging and boosting ensemble techniques. The comparative result analysis through multiple performance parameters reveals that the decreased number of features improves the performance for each of the models and the ensemble models outperform other types of models. Therefore, Random Forest regression ensemble ML model can get the best prediction accuracy with 0.036018 RMSE, 0.029470 MAE and 0.934512 R2 employing the dataset with reduced features for estimating the value of bitcoin.*

Keywords: machine learning, bitcoin, regression models, ensemble feature selection

OCENA SKUTECZNOŚCI PROGNOZOWANIA CEN BITCOINÓW PRZY UŻYCIU TECHNIK UCZENIA MASZYNOWEGO NA DANYCH HISTORYCZNYCH

Streszczenie. *Od momentu wejścia na rynek w 2009 roku, cena Bitcoina jest niezwykle nieregularna. Na jego cenę wpływają takie czynniki, jak wskaźniki popularności, zmiany regulacyjne, wydarzenia geopolityczne i zmiany makroekonomiczne. Eksperci uważają, że cena Bitcoina wzrośnie w dłuższej perspektywie ze względu na ograniczoną podaż i rosnący popyt. Dlatego też celem niniejszego badania jest zaproponowanie podejścia opartego na selekcji cech i uczeniu maszynowym do przewidywania ceny bitcoina. Do tego celu badawczego wykorzystano, zwizualizowano i wstępnie przetworzono zbiór danych oparty na kryptowalutach. Zastosowano pięć różnych podejść do wyboru cech (Pearson, RFE, Embedded Random Forest, Tree-based i Light GBM), a następnie metodologię ensemble, z podejściem maksymalnego głosowania w celu wyodrębnienia najważniejszych cech i wygenerowania zbioru danych ze zredukowanymi atrybutami. Następnie zbiór danych z lub bez selekcji cech jest wykorzystywany do przewidywania cen bitcoinów poprzez zastosowanie dziesięciu różnych modeli regresji uczenia maszynowego, w tym sześciu tradycyjnych, czterech technik baggingu i boostingu. Analiza porównawcza wyników za pomocą wielu parametrów wydajności pokazuje, że zmniejszona liczba cech poprawia wydajność każdego z modeli, a modele zespołowe przewyższają inne typy modeli. W związku z tym model Random Forest regression ensemble ML może uzyskać najlepszą dokładność przewidywania z 0,036018 RMSE, 0,029470 MAE i 0,934512 R2, wykorzystując zbiór danych ze zredukowanymi funkcjami do szacowania wartości bitcoinów.*

Słowa kluczowe: uczenie maszynowe, bitcoin, modele regresji, selekcja cech zespołu

Introduction

Bitcoin, the first and most prominent cryptocurrency, has revolutionized the financial landscape since Satoshi Nakamoto, who goes by a pseudonym, introduced it in 2009. Bitcoin's origins trace back to in October 2008, an individual or collective known as Satoshi Nakamoto published a whitepaper titled "Bitcoin: A Decentralized Digital Cash System" [40]. Bitcoin, as the leading cryptocurrency, has exhibited tremendous volatility, representing a high-risk high-reward investment. Its price oscillations have stimulated considerable intrigue and speculation among investors, economists, and researchers alike. One of the main challenges in financial markets, including the cryptocurrency market, is predicting future prices. The potential to predict Bitcoin prices accurately could lead to substantial economic gain and financial market stability.

Machine learning is an important tool for Bitcoin price prediction due to its potential to improve forecasting accuracy. Machine learning algorithms can process large datasets, identify intricate patterns and correlations, handle complex and non-linear relationships, capture and exploit intricate patterns in Bitcoin price data, and adapt quickly to changing market conditions. Accurate Bitcoin price predictions are essential for effective risk management and investment strategies, as investors and traders can utilize machine learning models to estimate potential risks and returns associated with different trading decisions. Additionally, machine learning-based trading algorithms can automate trading processes, taking advantage of price inefficiencies and maximizing profit potential. The application of machine learning techniques for Price forecast of bitcoin also contributes to advancing academic research in finance and economics.

In order to estimate the price of bitcoin, this work aims to offer a machine learning regression hypothesis and an ensemble feature selection method. The study uses five alternative feature selection algorithms after data pre-processing, each of which uses the dataset to extract the important properties, to accomplish this purpose. The dataset is based on cryptocurrencies. In order to examine the smaller set of features, the outcomes from various feature selection methodologies are combined through ensemble majority voting. The work then employs bagging and boosting type ensemble procedures to perform ML regression techniques with and without the dataset with decreased features for the model's development and testing. The usefulness of the ensemble techniques as well as the impact of features reduction are confirmed through a comparison analysis using several traditional and ensemble regression models utilizing multiple performance indicators. The research identifies a model's most effective results for predicting the price of bitcoin with the best features, where ensemble ML approach using Random Forest regression eventually outperforms other methods.

The subsequent section of the paper is organized as follows: Section 2 provides an overview of the relevant literature; Section 3 outlines the methodology adopted for this study; Section 4 examines the results, findings, and a comparative analysis of performance; Lastly, Section 5 concludes by summarizing the main findings, benefits, limitations, and suggesting future research directions.

1. Literature review

Numerous recent scholarly inquiries have delved into the methodologies employed to predict Bitcoin prices. As an illustration,

Zheshi Chen et al. [8], proposed features of elevated dimensions such as Exchange and consumers, concentrate, and gold, together with real estate and system as well as fundamental trading information from a cryptocurrency exchange, are utilized to anticipate Bitcoin values across a range of periods. Hari Krishnan Andi et al. [1], suggested the LSTM machine learning approach is the preferred endeavor for standardizing a dataset to achieve precise estimations of Bitcoin prices, with superior accuracy, recall, precision, and sensitivity. Neha Mangla et al. [22], Strive to deliver a precise evaluation of Bitcoin prices, considering daily variations and identifying the most impactful factors influencing the price. Moreover, Wei Chen et al. [7], determined if economic and technological factors can accurately estimate the exchange rate of bitcoin, in this work, a two-stage methodology was established. To decrease in the initial phase, two nonlinear feature selection techniques were employed to determine the potential predictors' count. Results revealed that in terms of predictive performance, LSTM demonstrated superior performance compared to other methods such as support vector regression, auto-regressive integrated moving average, adaptive network fuzzy inference system, and LSTM itself. Using a sizable dataset of 30,000 items. Kavitha H et al. [16], predicted the price of Bitcoin where different techniques such as recurrent neural network, long short-term memory, and linear regression were employed. Hakan Pabuçcu et al. [25], Utilizing SVM, ANN, NB, and RF for forecasting Bitcoin price trends, the outcomes indicated RF excelled in predictive capability, while NB exhibited the least favorable performance. Sahar Erfanian et al. [11], proposed a technique using macroeconomic, microeconomic, technical, and blockchain factors, a research employing ML approaches to estimating bitcoin's price. The effectiveness theories on long-term prediction involving demand, inventory, and cost-efficient pricing is thus established. Patrick Jaquart et al. [15], according to research, technological factors are more crucial for most strategies than asset-, sentiment-, and blockchain-based elements. Longer prediction horizons also lead to more predictability. Qinghe Li et al. [20], Employing data records capturing daily occurrences throughout history, the XGBoost classifier surpassed the performance of two logistic regressions, resulting in favorable returns and a maximum retracement. This study improves Bitcoin purchases and sales, increases returns, and could offer suggestions for future stock market research.

Reshma Sundari Gadey et al. [12], proposed on ML and AI utilized in trading to produce atypical returns from the bitcoin market. When predicting the value of the bitcoin price, LSTM Architecture produces more accurate results than any other machine learning algorithms and architecture. Poongodi M et al. [27], assessing the machine learning prediction algorithms and the LTMS network's underlying technology in this research report. We gathered data from the bitcoin blockchain and used the ARIMA model to forecast bitcoin's price. Eng Chuen Loh et al. [20], working with three neural networks – FNN, NARX, and NAR – are compared in this research. The most efficient model is selected by carefully examining each model's performance measurement and applying it to price estimates for Bitcoin. The findings showed that NARX surpassed FNN and NAR forecasting the value of Bitcoin, making it the top neural network for this task. This presents novel viewpoints on Bitcoin forecasting, which is helpful for investors and economists. Ehsan Sadeghi Pour et al. [28], using a hybrid artificial neural network model that combines Bayesian Optimization and Long Short-Term Memory, this paper presents a price forecast for Bitcoin prices. It improves on earlier techniques for forecasting bitcoin prices and culminates with graphs and tables summarizing the outcomes of the optimization. Keyue Yan et al. [42], the use of ensemble learning algorithms, the study seeks to forecast Bitcoin price patterns. Different labels are produced by variations in closing prices or moving average prices, and the anticipated close price can be used to guide future investing choices, such as whether to engage in longing or shorting activities.

Athanasia Dimitriadou et al. [10], using 24 possible explanatory factors, the machine-learning approach proposed in this study aims to forecast Bitcoin price changes. Based on the results, it appears that the conventional logistic regression model, which achieves an accuracy of 66%, performs better than the linear support vector machine and random forest technique. There is evidence to support the idea that the Bitcoin market rejects poor form efficiency. Sahi et al. [34], using several machine-learning algorithms, the current study seeks to predict the values of three cryptocurrencies. Five Machine Learning models were used to assess the data, which was gathered from well-known crypto monitoring websites. According to the findings, the Support Vector Machine method produced results with the best accuracy, with RMSE and MAPE values that were most similar to. Arumalla et al. [2], research tries to use machine learning techniques to forecast the movement of Bitcoin prices. Comparative analysis of six modules: gradient boosting, linear regression, decision tree regression, random forest regression, support vector regression, and lasso regression. Metrics including mean square error, root mean square error, mean absolute error, score function, and median absolute error will be used to gauge how accurate these models are. The most accurate model will be determined, and a user interface will be created to help users explore the market price of Bitcoin. Bhatt et al. [4], with an emphasis on incorporating multimodal fusion model and on-chain data, this study examines how social media attitudes affect propensity of machine learning models to anticipate Bitcoin prices. The historical data was used to train several models using the information from the crypto market, the blockchain, and pertinent social media that were gathered 2014 through 2022. With Twitter-based Roberta sentiment producing 0.79 on the F1 scale as a whole, the introduction of sentiment data regularly improved performance. An enhanced Multi Modal Fusion classifier integrating sentiment from Twitter Roberta generated the best results, with an F1 score of 0.85.

Chen et al. [6], develop a highly accurate prediction model using random forest regression for forecasting the next-day price of Bitcoin, while also identifying the key factors influencing its price. Random forest regression outperforms LSTM in terms of accuracy and can be employed to monitor fluctuations in the factors impacting Bitcoin's price. Pragadareddy et al. [29], using decision tree classification, this research proposes a model for predicting Bitcoin price. Open, high, low, and closing price information is captured in the dataset. This study's objective is to assess how reliable Bitcoin price forecasts are using several machine learning algorithms and to compare those accuracies. In order to forecast price values, a machine learning module is introduced. With accuracy of 97%, precision of 96.7%, and recall of 96.9%, Decision Tree surpasses other classifiers. Kiranashree et al. [19], by considering a number of variables that affect bitcoin's value, this research seeks to more precisely predict its price. The data collection contains daily data spanning a five-year period and many variables connected to the price of bitcoin. The data will be used in the investigation's second phase to project the daily price change's direction. Bhattad et al. [5], examined the machine learning algorithms used in 9 studies to determine the most effective model to forecast time series models' expenses. The outcomes demonstrated that the machine learning models could anticipate trends rather accurately. For a highly volatile asset like cryptocurrencies, however, generalizing long-term forecasts based on a limited numer of models produces results with low accuracy. To close this gap, we advocate the use of diverse models. Auti et al. [3], using 1-minute interval trade data from the Bitstamp Bitcoin exchange website, this study explores machine learning techniques to determine the most efficient and reliable model for forecasting Bitcoin prices. Iqbal et al. [13], with the use of methods like ARIMA, FBProphet, and XG Boosting, machine learning-based time series analysis can forecast the stability and price of the Bitcoin market. Models are assessed using RMSE, MAE, and R2 parameters. The strongest model for predicting the price of Bitcoin on the cryptocurrency market is ARIMA,

which has RMSE scores of 322.4 and MAE scores of 227.3. Samaddar et al. [35], approaches for predicting future price using data from the actual world. Convolutional, recurrent, and artificial neural networks are among the machine learning models that are compared, as well as supervised learning methods like Random Forest and k-nearest neighbours. Using time price prediction and epoch loss accuracy graphs, the study investigates the variations in results. Shahbazi and Byun et al. [37], based on prior price inflations discovered through research, bitcoin price predictions are made. Using the XGBoost algorithm for security and transparency, this paper provides a blockchain-based exchange rate prediction system. For accurate prediction, data mining techniques, a range of filters, and coefficient weights are used. In order to improve performance, cross-validation is used during training. Reddy et al. [31], proposed a method where maximize profits for investors, research the forecast of Bitcoin values using machine learning techniques. Kervanci and Akay et al. [17], ML approaches outperform other methods in comparative studies. This review investigates combining earlier research to forecast bitcoin values more accurately. It explores statistical approaches, ML and methods, frequency effects, social media impact, causality, and hyper-parameter optimization. Shankhdhar et al. [39], find the best accurate way to forecast the price of bitcoin, the study compares deep learning algorithms like LSTM and GRU with machine learning models like Multivariate Linear Regression, Theil-Sen Regression, and Huber Regression. An IoT alert system is created, and the dataset is saved in MongoDB.

Suha and Sanam et al. [41], proposed a method that utilizes the Random Forest Regression Model to analyze patient data and create a robust prediction model. Utilizing a hospital discharge dataset, the model employs an interquartile range-based outlier reduction strategy and features like PCA and Chi-square. Validating performance using 10 different regression models and deep learning methods, the Random Forest Regression model outperformed other models in performance indicators. Khedr et al. [18], analyses articles predicting cryptocurrency prices from 2010 to 2020 using statistical and machine learning methods, highlighting challenges and suggesting ML and DL approaches for improved precision. Ranjan et al. [30], uses machine learning techniques to predict Bitcoin values, focusing on daily and high-frequency pricing. Logistic Regression achieves 64.84% accuracy, while XGBoost estimates prices for 5-minute intervals with 59.4% accuracy. Chowdhury et al. [9], uses machine learning algorithms and models to predict the closing price of the cryptocurrency index 30 and its nine constituents, simplifying trading for users. Various approaches and algorithms are used, and models are compared for optimal results. Shakri et al. [38], compares the performance of alternating model tree, random forest (RF), multiple linear regression, multi-layer perceptron regression, and M5 Tree algorithms in predicting time series data of Bitcoin returns. Parvez et al. [26], proposed method Bitcoin valuation machines are prepared using Random Forest regression to fake fearlessness in machine learning computations, deconstructing AI models from top-performing ones. Ren et al. [32], proposed method on cryptocurrency price trends and income fluctuations is increasingly utilizing machine learning methods, despite concerns about overfitting and interpretability. The use of multiple approaches in cryptocurrency research is becoming more prevalent, indicating the need for further research. Mujlid et al. [23], proposed study emphasizes unanswered research concerns with the application of machine learning algorithms in cryptocurrencies and analyzes the literature on machine learning-based bitcoin price prediction. Nagamani et al. [24], uses machine learning architecture to predict bitcoin price using Support Vector Machine (SVM) and K Nearest Neighbor (KNN) algorithms, demonstrating superior accuracy compared to the current KNN approach. Senthilkumar et al. [36], proposed a method predict Bitcoin's predicted price direction using machine learning models. The model uses biases to forecast the price, learning from dataset patterns. The study has improved the model with expert input, indicating that it may not predict the future of cryptocurrency.

However, Zheshi Chen et al. [8], approach has certain constraints, notably the omission of certain machine learning algorithms from consideration, which is one of the research's limitations in terms of data sources and analyses. We need to gather pricing data with additional characteristics and granularity in order to conduct a more thorough study on Bitcoin price prediction. Hari Krishnan Andi et al. [1], research bears a limitation – attributed to overfitting and errors induced by extensive datasets – resulting in the ineffectiveness of the majority of models. Neha Mangla et al. [22], research encounters a constraint – precise classification through logistic regression-based models is contingent upon the existence of a differentiable hyperplane. Kavitha H et al. [16], method has the enormous amount of computation needed to train both models is a drawback. When the dataset is small, the RNN model does not train properly and produces poor predictions. Hakan Pabuçcu et al. [25], effectively predicting changes in Bitcoin prices necessitates empirical research, constituting a limitation in the scope of the work. Sahar Erfanian et al. [11], a limitation in this study's datasets lies in the utilization of a limited number of fundamental approaches for feature selection. Reshma Sundari Gadey et al. [12], research the likelihood of overfitting increases as the training sets grow in size. Keyue Yan et al. [42], shortage of different kinds of real-time data on bitcoin. B. K. Kiranashree et al. [19], cryptocurrencies are not secure due to their volatility and lack of supervision by the FCA, and can be used as a means of payment for fraud. Samaddar et al. [35], drawback of this research has amount of data being collected is a problem. The statistics predicted by many research turn out to be drastically off because they lack comprehensive, large data sets.

2. Methodology

A synopsis of the research procedure is presented in figure 1. A number of GitHub datasets were taken into consideration before moving forward with the implementation. One suitable dataset is chosen for our research. The dataset has then been carefully examined using numerous visualization approaches that have been used to achieve a comprehensive understanding of the data, which has undergone pre-processing to transform it into a well-organized and appropriate dataset suitable for machine learning purposes. The proposed method is referred to as a hybrid strategy since it includes two main phases, the first of which is feature engineering, which was carried out in a number of stages to find and choose the perfect set of features required for estimating bitcoin prices; and during the classification phase, a regression model based on machine learning will be a technique explored to anticipate Bitcoin price aforementioned attributes. The next section describes each stage's specifics.

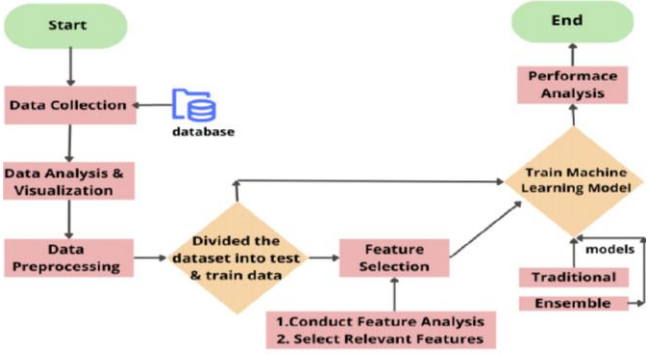

Fig. 1. Framework of methodology

2.1. Data collection

Collecting data is an essential element of any thesis research since it forms the basis for analysis and enables the derivation of significant conclusions. The process of data collection involves gathering relevant information or observations to address the research questions or hypotheses outlined in the thesis. Depending on the nature of the study, data can be collected through various methods [33].

Commencing with dataset compilation encompassing cryptocurrency-related information marked the initial phase. Leveraging the widely acknowledged open-source dataset bitcoin-price-prediction from GitHub, the study not only utilizes but also validates the proposed methodology. The study employed a dataset titled "btc_final.csv," encompassing data from November 13, 2017, to April 8, 2023. With a total of 1972 records, each containing 23 attributes, including a target attribute featuring a floating-point output denoting the average daily Bitcoin value. The dataset encapsulates numerical values for each attribute.

2.1.1. Data description

In this part, we provide an overview of the data utilized in the survey, specifically highlighting the features used to analyse and train the machine learning models. The data analysis section and the machine learning section rely on these features to explain and interpret the results. The dataset consists of 1972 records of data, each of which has 23 attributes, among them is the target attribute, constituting a floating-point output. Comprehensive attribute details are elucidated in table 1.

Table 1. Attribute delineation

Attribute Name	Attribute Description	Data Type
bt_close	Bitcoin close	Float
bt_volume	Bitoin volume	Float
bt_day_diff	Bitcoin_day difference	Float
bt_close_off_high	Bitcoin close off high	Float
bt_volatility	Bitcoin volatility	Float
eur_exch_rate	Euro exchange rate	Float
jpy_exch_rate	Japanese exchange rate	Float
cny_exch_rate	Chinese exchange rate	Float
google_trends_bitcoin	Google trends bitcoin	Float
avg_block_size	Average block size	Float
transactions	Transactions	Float
difficulty	Difficulty	Float
bchain_size	Block chain size	Float
mining_revenue	Mining revenue	Float
hash_rate	Hash rate	Float
cost_per_transaction	Cost per transaction	Float
sp_close	Whole sell price per day	Float
dj_close	Corporation price USD	Float
nasdap_close	Stock market index	Float
vix_cloase	Volatility index	Float
gold_am	Gold amount	Float
Silver	Silver	Float
old_price	Oil price	Float
date	Date	Date type

2.2. Data analysis and visualization

2.2.1. Data analysis

The dataset employed in this analysis was a sample of cryptocurrency that was collected from GitHub. It consisted of 1972 instance data and 23 attributes. There were 24 float types of data, where float meant numerical data. The focus of this essay did not extend to the scientific implications of those factors. To start the analysis, the data set was read and its dimensions and data types were checked. Descriptive statistics and the first few rows of the data set were printed out. Duplicate rows were also checked for and removed if necessary. Finally, the unique values of each column were checked to see if there were any issues or inconsistencies.

2.2.2. Data visualization

Data visualizations are graphical representations of data and information. They are used to communicate complex data sets in a visual and easily understandable format. Data visualizations can be an effective way to present and analyse data in a thesis. The selection of visualization methods relies on the inherent characteristics of the data and its particularities research inquiries being explored. There were some common visual representations shown figure 2 and figure 3.

Fig. 2. Histogram view using all attributes

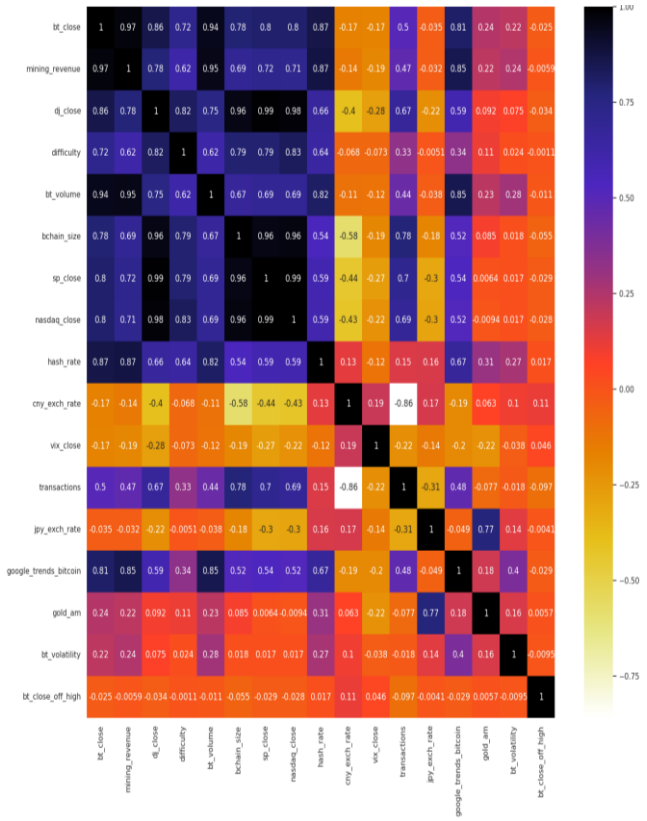

Fig. 3. After data pre-processing and feature selection the data correlations

2.3. Data pre-processing

Data pre-processing stands as a crucial initial phase in research that involves cleaning, transforming, and preparing raw data for analysis. The accuracy and dependability of research findings heavily rely on the quality of the data, and enhancing the data quality can be achieved through data pre-processing techniques. The choice of data pre-processing techniques depends on the research question, the nature of the data, and the specific analysis techniques that will be used. It is important to document all data pre-processing steps taken in a research project, including the reasons for selecting particular techniques and the outcomes of those techniques. Proper data pre-processing can help ensure that research findings are precise, dependable, and significant, and can lead to more efficient decision formulation and adept problem resolution. In order to prepare the data for our work, we performed several pre-processing tasks including removing redundant columns, identifying and addressing missing or null values, and validating numerical and categorical data.

2.4. Feature prioritization and selection

Five different feature selection strategies are used in this study with the goal of examining the 23 numerical attributes' most predominant qualities. The results are then combined using a majority vote procedure [14]. A synopsis of the majority voting ensemble model procedure is presented in figure 4. The best 16 traits for each technique have been chosen, and all the features have been ranked based on the votes. The remaining 6 features are then deleted from the data frame, leaving only the 16 attributes that garnered the most votes when employing feature selection procedures.

The feature selection methods are: Recursive Feature Elimination (RFE), Tree-based Feature Elimination, Embedded Random Forest, Pearson Correlation Coefficient Technique, and Light GBM.

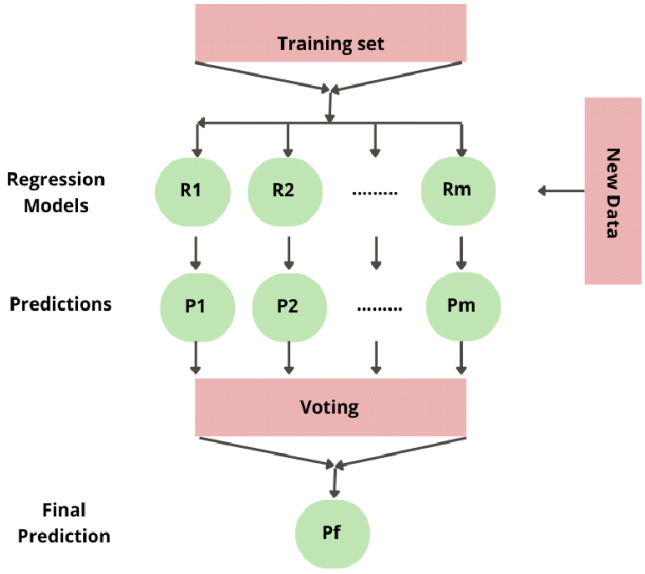

Fig. 4. Framework of majority voting ensemble model

The comparative study of the ML model's performances with and without the condensed set of features has been used to execute an ablation test in order to explain the relative contributions of this dominant feature selection phase utilizing ensemble technique to the effectiveness of predictive analytics.

2.5. Machine learning models

This study utilizes a diverse set of ten machine learning regression models. It divides into two discernible clusters: the initial set comprises conventional Machine Learning regression techniques like Support Vector Regression, Decision Tree Regressor, Linear Regression, Gaussian Process Regressor, Neural Network Regression and K-Nearest Neighbors Regressor. The second category encompasses ensemble Machine Learning regression models, specifically boosting or bagging ensemble models, including Hist Gradient Boosting Regressor, Gradient Boosting Regressor, eXtreme Gradient Boosting (XGBoost) Regressor, and Random Forest. The ensemble technique in machine learning merges multiple individual models to construct a single, optimized prediction model. The dataset with and without decreased feature sets was used to train and evaluate each of these models, and the strategy with the highest performance was then identified. Distinct performance indicators were employed to identify the most effective performance technique.

2.6. Performance analysis

A number of performance indicators, such as RMSE, MAE, and R2, are employed to assess the performance of the prediction models. RMSE offers a measure of the average prediction error, with a greater number indicating larger errors. These performance indicators are obtained from a contrast between true values and estimated data from the practise dataset. It is simpler to read because it is expressed in the same units as the dependent variable. Better model performance is shown by lower RMSE values. Additionally, MAE measures the average absolute prediction error and is expressed in the same units as the dependent variable. Lower values of MAE suggest higher model performance, similar to RMSE. R2 can be negative if the model performs worse than the basic average. Higher R2 values often indicate greater model performance, but to avoid potential problems, they should be assessed in conjunction with other metrics.

Root Mean Squared Error (RMSE): A regression model's key performance criterion is the RMSE. It determines the gap between actual and projected values and provides an estimation of the model's forecasting capabilities. The RMSE decreases with the enhancement of model excellence.

The following expression represents the formula for calculating the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{yi} - yi)^2}$$

Mean Absolute Error (MAE): is a measurement utilized to evaluate the average size of errors between predicted and actual values, offering a clear evaluation of the model's effectiveness. It focuses solely on the absolute values of errors and does not take into account their direction.

The following expression represents the formula used to calculate Mean Absolute Error:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|yi - \hat{yi}\right|$$

R squared (R2): When predicting future results or testing hypotheses, the coefficient of determination (R2) is a metric for extent to which a model aligns with the dataset. The greater the R-squared, the better the model.

The expression for the R2 Formula is outlined as follows:

$$R^2 = 1 - \frac{\sum(yi - \hat{yi})}{\sum(yi - \bar{yi})}$$

The result is the overall performance of the dataset used in the article.

3. Result analysis

3.1. Feature selection

In the context of this research, five different attribute selection techniques were used: Pearson correlation coefficient methodology, Recursive Feature Elimination (RFE), Embedded Random Forest, Tree-based Feature Selection and Light GBM. Following data pre-processing, each of them utilized a different approach to select the 16 features out of a total of 20 features that they felt were most crucial, leaving out the other 4 non-prioritized qualities. Votes for each attribute are then calculated using a variety of methods and outcomes are displayed in a table 2.

These five characteristics "mining revenue, dj close, difficulty, bt volume, and bchain size" have received the most votes across all five strategies, as can be shown in table 2. Following that, four approaches each gave votes to the attributes "sp close", "nasdaq close", "hash rate," and "cny exch rate" while three techniques gave votes to the attributes "vix close", "transactions", "silver," "jpy", "exch rate", "google trends", "gold am" and "bt volatility". The other four attributes, "oil price, euro exchange rate, cost per transaction, and bt day dif" can all be regarded as the least prioritized ones because they have all received only two votes from any one of the feature selection methods. As a result, the new dataset only comprises the 16 traits that the assessment decided to be the most important, leaving out the four qualities mentioned earlier.

3.2. ML regression model

This study employed a variety of ten ML regression models, categorized into conventional, bagging, and boosting types, to forecast Bitcoin prices. In order to investigate the results of the features reduction technique, each of these models has been trained and tested using the dataset with complete features as well as a reduced features set. Table 3 showcases outcomes from diverse machine learning models, presenting evaluations through RMSE, MAE, and R2 metrics. Table 3 makes it clear that, when compared to traditional models utilizing both types of feature sets, ML models with bagging and boosting ensemble models perform noticeably better. The results also figure 5, Comparative R-Squared analysis of different ML Models demonstrate that for all classifiers, accuracy increases when the feature set is decreased. The best accuracy in this case is **89.1391** (nearly 89.2%) when all of the features from the dataset are used, while the best accuracy when utilizing the smaller feature set is **93.4512 (almost 93.5%)**. figure 5 graphically shows the comparative R-Squared analysis of ML models. This means that, when using the bitcoin dataset for price prediction, ensemble forms of machine learning techniques often produce higher results. Additionally, a smaller dataset with the 16 most important features and a "Random Forest regression" ensemble ML model can get achieving optimal predictive accuracy with an **RMSE** of **0.036018, MAE** of **0.029470,** and **R2 of 0.934512.**

Table 2. Attribute voting based on feature selection techniques

SL.	Feature	Pearson	RFE	Random Forest	Tree-based	Light GBM	Total
1	mining_revenue	TRUE	TRUE	TRUE	TRUE	TRUE	5
2	dj_close	TRUE	TRUE	TRUE	TRUE	TRUE	5
3	difficulty	TRUE	TRUE	TRUE	TRUE	TRUE	5
4	bt_volume	TRUE	TRUE	TRUE	TRUE	TRUE	5
5	bchain_size	TRUE	TRUE	TRUE	TRUE	TRUE	5
6	sp_close	TRUE	TRUE	TRUE	TRUE	FALSE	4
7	nasdaq_close	TRUE	TRUE	TRUE	TRUE	FALSE	4
8	hash_rate	TRUE	TRUE	TRUE	FALSE	TRUE	4
9	cny_exch_rate	TRUE	TRUE	TRUE	FALSE	TRUE	4
10	vix_close	TRUE	TRUE	TRUE	FALSE	FALSE	3
11	transactions	TRUE	TRUE	TRUE	FALSE	FALSE	3
12	silver	TRUE	TRUE	TRUE	FALSE	FALSE	3
13	jpy_exch_rate	TRUE	TRUE	TRUE	FALSE	FALSE	3
14	google_trends_bitcoin	TRUE	TRUE	TRUE	FALSE	FALSE	3
15	gold_am	TRUE	TRUE	TRUE	FALSE	FALSE	3
16	bt_volatility	TRUE	TRUE	TRUE	FALSE	FALSE	3
17	oil_price	TRUE	FALSE	TRUE	FALSE	FALSE	2
18	eur_exch_rate	FALSE	TRUE	TRUE	FALSE	FALSE	2
19	cost_per_transaction	TRUE	TRUE	FALSE	FALSE	FALSE	2
20	bt_day_diff	FALSE	TRUE	TRUE	FALSE	FALSE	2

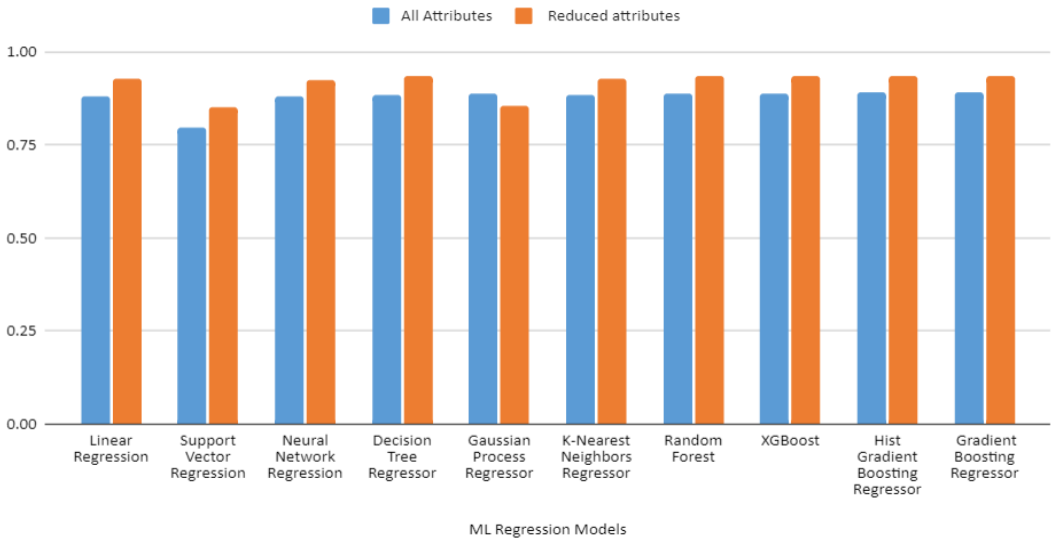

Fig. 5. Comparative R-Squared analysis of different ML Models

4. Discussion and conclusion

One popular subject of research in the domains of finance and artificial intelligence involves the application of ML techniques to estimate future Bitcoin price movements. As part of this study, examined machine learning approaches that predict Bitcoin values based on sample properties and dimensions. To extract the essential features, the proposed approach takes a novel approach to examining and pre-processing the commonly employed cryptocurrency dataset. A number of widely established approaches are used in the feature selection procedure as well. Different types of ML methodologies with diverse regression models are trained and tested employing both the dataset with and without reduced features to explore the best performing one the result analysis reveals that, using a dataset with fewer attributes, the ensemble model with Random Forest regression

surpasses others with an R2 value 93.5%, which is a much greater performance than that of another research that have been done in this domain with this dataset.

Nevertheless, the study's limitations should be acknowledged, including reliance on a single dataset and not exploring all machine learning algorithms. Additionally, potential issues like large dataset overfitting and the inherent volatility of cryptocurrencies.

Therefore, the research team plans to do additional research in this area in the future utilizing a larger dataset, especially records from the least developed countries in order to perform predictive analytics for another cryptocurrency price. In addition, to acquire a deeper understanding, the proposed method may be integrated with sophisticated ML technologies in the future, such as deep learning, AI federated learning, etc.

Table 3. Evaluating machine learning model performance with complete and simplified feature sets

Type	ML Models	with full feature set			With reduced feature set		
		RMSE	MAE	R2	RMSE	MAE	R2
Traditional Models	Linear Regression	0.065201	0.053346	0.881453	0.040728	0.033323	0.925949
	Support Vector Regression	0.112056	0.091682	0.796262	0.081122	0.066373	0.852505
	Neural Network Regression	0.066337	0.054275	0.879388	0.041240	0.033742	0.925017
	Decision Tree Regressor	0.064226	0.052548	0.883226	0.036649	0.029985	0.933367
	Gaussian Process Regressor	0.061329	0.050178	0.888493	0.079730	0.065233	0.855037
	K-Nearest Neighbors Regressor	0.064461	0.052741	0.882798	0.038945	0.031864	0.929191
Bagging Ensemble	Random Forest Regressor	0.060950	0.049869	0.889181	**0.036018**	**0.029470**	**0.934512**
Boosting Ensemble	XGBoost	0.061016	0.049922	0.889062	0.036466	0.029836	0.933698
	Hist Gradient Boosting Regressor	0.060296	0.049333	0.890371	0.036449	0.029822	0.933728
	Gradient Boosting Regressor	**0.059923**	**0.055545**	**0.891391**	0.036061	0.029504	0.934435

References

[1] Andi H. K.: An Accurate Bitcoin Price Prediction Using Logistic Regression with LSTM Machine Learning Model. Journal of Soft Computing Paradigm 3(3), 2021, 205–217 [https://doi.org/10.36548/jscp.2021.3.006].

[2] Arumalla G. S. et al.: Bitcoin price fluctuation analysis and prediction using machine learning. International Journal of Progressive Research in Engineering Management and Science – IJPREMS 03(03), 2022, 421–425.

[3] Auti A. et al.: Bitcoin Price Prediction Using Svm. International Journal of Engineering Applied Sciences and Technology 6(11), 2022, 226–229.

[4] Bhatt S. et al.: Machine Learning based Cryptocurrency Price Prediction using Historical Data and Social Media Sentiment. Computer Science & Information Technology – CS & IT 13, 2023, 1–11 [https://doi.org/10.5121/csit.2023.131001].

[5] Bhattad S. et al.: Review of Machine Learning Techniques for Cryptocurrency Price Prediction. EasyChair 10190, 2023.

[6] Chen J.: Analysis of Bitcoin Price Prediction Using Machine Learning. Journal of Risk and Financial Management 16(1), 2023, 51 [https://doi.org/10.3390/jrfm16010051].

[7] Chen W. et al.: Machine Learning Model for Bitcoin Exchange Rate Prediction Using Economic and Technology Determinants. International Journal of Forecasting 37(1), 2021, 28–43 [https://doi.org/10.1016/j.ijforecast.2020.02.008].

[8] Chen Z. et al.: Bitcoin Price Prediction Using Machine Learning: An Approach to Sample Dimension Engineering. Journal of Computational and Applied Mathematics 365, 2020, 112395 [https://doi.org/10.1016/j.cam.2019.112395].

[9] Chowdhury R. et al.: An Approach to Predict and Forecast the Price of Constituents and Index of Cryptocurrency Using Machine Learning. Physica. A 551, 2020, 124569 [https://doi.org/10.1016/j.physa.2020.124569].

[10] Dimitriadou A., Gregoriou A.: Predicting Bitcoin Prices Using Machine Learning. Entropy 25(5), 2023, 777 [https://doi.org/10.3390/e25050777].

[11] Erfanian S. et al.: Predicting Bitcoin (BTC) Price in the Context of Economic Theories: A Machine Learning Approach. Entropy 24(10), 2022, 1487 [https://doi.org/10.3390/e24101487].

[12] Gadey R. S. et al.: Price prediction of bitcoin using machine learning. International Journal of Engineering Applied Science and Technology 5(1), 2020, 502–506 [https://doi.org/10.33564/ijeast.2020.v05i01.089].

[13] Iqbal M. et al.: Time-Series Prediction of Cryptocurrency Market Using Machine Learning Techniques. EAI Endorsed Transactions on Creative Technologies 8(28), 2021, 170286 [https://doi.org/10.4108/eai.7-7-2021.170286].

[14] Islam M. R. et al.: Data-Driven Heart Disease Prediction by Ensemble Feature Selection and Machine Learning Techniques. 25th International Conference on Computer and Information Technology (ICCIT), 2022, 575–580 [https://doi.org/10.1109/iccit57492.2022.10054998].

[15] Jaquart P. et al.: Short-term Bitcoin Market Prediction via Machine Learning. Journal of Finance and Data Science 7, 2021, 45–66 [https://doi.org/10.1016/j.jfds.2021.03.001].

[16] Kavitha H. et al.: Performance Evaluation of Machine Learning Algorithms for Bitcoin Price Prediction. 2020 Fourth International Conference on Inventive Systems and Control (ICISC), 2020, [https://doi.org/10.1109/icisc47916.2020.9171147.

[17] Kervanci, I. S., Akay F.: Review on Bitcoin Price Prediction Using Machine Learning and Statistical Methods. Sakarya University Journal of Computer and Information Sciences 3(3), 2020, 272–282 [https://doi.org/10.35377/saucis.03.03.774276].

[18] Khedr A. M. et al.: Cryptocurrency Price Prediction Using Traditional Statistical and Machine-learning Techniques: A Survey. International Journal of Intelligent Systems in Accounting, Finance & Management 28(1), 2021, 3–34 [https://doi.org/10.1002/isaf.1488].

[19] Kiranashree B. K. et al.: Price Prediction of Bitcoins. 22 Mar. 2023, [https://journal.ijmdes.com/ijmdes/article/view/115].

[20] Li Q.: Predicting Trends of Bitcoin Prices Based on Machine Learning Methods. 4th International Conference on Software and e-Business, 2020, 49–52 [https://doi.org/10.1145/3446569.3446588].

[21] Loh E. C.: Emerging Trend of Transaction and Investment: Bitcoin Price Prediction Using Machine Learning. International Journal of Advanced Trends in Computer Science and Engineering 9(1.4), 2020, 100–104 [https://doi.org/10.30534/ijatcse/2020/1591.42020].

[22] Mangla N. et al.: Bitcoin price prediction using machine learning. International Journal of Information and Computing Science 6(5), 2019, 318–320.

[23] Mujlid H.: A Survey on Machine Learning Approaches in Cryptocurrency: Challenges and Opportunities. 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), IEEE, 2023.

[24] Nagamani P. et al.: Bitcoin Price Prediction Using Machine Learning Algorithms. Advances in engineering research/Advances in Engineering Research, 2023, 389–396 [https://doi.org/10.2991/978-94-6463-252-1_43].

[25] Pabuçcu H. et al.: Forecasting the Movements of Bitcoin Prices: An Application of Machine Learning Algorithms. Quantitative Finance and Economics 4(4), 2020, 679–692 [https://doi.org/10.3934/qfe.2020031].

[26] Parvez S. J. et al.: Bitcoin price prediction using Random Forest Regression. Journal of Positive School Psychology, 2022, 4352–4358.

[27] Poongodi M. et al.: Bitcoin Price Prediction Using ARIMA Model. International Journal of Internet Technology and Secured Transactions 10(4), 2020, 396 [https://doi.org/10.1504/ijitst.2020.108130].

[28] Pour E. S. et al.: Cryptocurrency Price Prediction with Neural Networks of LSTM and Bayesian Optimization. European Journal of Business and Management Research 7(2), 2022, 20–27 [https://doi.org/10.24018/ejbmr.2022.7.2.1307].

[29] Pragadareddy K. T. et al.: Price prediction model of bitcoin using decision tree classification. International Journal of Food and Nutritional Sciences (IJFANS) 11(1), 2022.

[30] Ranjan S. et al.: Bitcoin Price Prediction: A Machine Learning Sample Dimension Approach. Computational Economics 61(4), 2022, 1617–1636 [https://doi.org/10.1007/s10614-022-10262-6].

[31] Reddy K. R. et al.: Bitcoin Price Prediction and Forecasting. International Research Journal of Engineering and Technology (IRJET) 9(04), 2022, 2395–0056.

[32] Ren Y.-S. et al.: Past, Present, and Future of the Application of Machine Learning in Cryptocurrency Research. Research in International Business and Finance 63, 2022, 101799 [https://doi.org/10.1016/j.ribaf.2022.101799].

[33] Roh Y. et al.: A Survey on Data Collection for Machine Learning: A Big Data – AI Integration Perspective. IEEE Transactions on Knowledge and Data Engineering 33(4), 2021, 1328–1347 [https://doi.org/10.1109/tkde.2019.2946162].

[34] Sahi G. et al.: Predicting Cryptocurrency Price Using Machine Learning. European Economic Letters (EEL) 13(1), 2023, 11–16.

[35] Samaddar M. et al.: A Comparative Study of Different Machine Learning Algorithms on Bitcoin Value Prediction. International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2021 [https://doi.org/10.1109/icaect49130.2021.9392629].

[36] Senthilkumar S., Nivedha B.: Bitcoin Price Prediction Using Ml. Social Science Research Network, 2022 [https://doi.org/10.2139/ssrn.4128261].
[37] Shahbazi Z., Byun Y.-C.: Knowledge Discovery on Cryptocurrency Exchange Rate Prediction Using Machine Learning Pipelines. Sensors 22(5), 2022, 1740 [https://doi.org/10.3390/s22051740].
[38] Shakri I. H.: Time Series Prediction Using Machine Learning: A Case of Bitcoin Returns. Studies in Economics and Finance 39(3), 2021, 458–470 [https://doi.org/10.1108/sef-06-2021-0217].
[39] Shankhdhar A. et al.: Bitcoin Price Alert and Prediction System Using Various Models. IOP Conference Series. Materials Science and Engineering 1131(1), 2021, 012009 [https://doi.org/10.1088/1757-899x/1131/1/012009].
[40] Squarepants S.: Bitcoin: A Peer-to-Peer Electronic Cash System. Social Science Research Network, 2008 [https://doi.org/10.2139/ssrn.3977007].
[41] Suha S. A., Sanam T. F.: A Machine Learning Approach for Predicting Patient's Length of Hospital Stay With Random Forest Regression. IEEE Region 10 Symposium (TENSYMP), 2022 [https://doi.org/10.1109/tensymp54529.2022.9864447].
[42] Yan K., Wang Y.: Prediction of Bitcoin prices' trends with ensemble learning models. Fifth International Conference on Computer Information Science and Artificial Intelligence (CISAI 2022), 2023, 900–905 [https://doi.org/10.1117/12.2667793].

M.Sc. Mamun Ahmed
e-mail: mamun.cse@baiust.ac.bd

Mamun Ahmed holds an M.Sc. in Signal Processing from Blekinge Institute of Technology (BTH), Sweden, and a Bachelor's degree in CSE from CUET, Bangladesh. He served as an RF Engineer at Motorola, Bangladesh, from 2006 to 2009. Since 2013, he has been an esteemed faculty member (Associate professor) in the Department of Computer Science and Engineering at Bangladesh Army International University of Science and Technology (BAIUST), Cumilla, Bangladesh. Published about 25 scientific papers.

https://orcid.org/0000-0002-3980-3981

M.Sc. Sayma Alam Suha
e-mail: suha.mist@gmail.com

Sayma Alam Suha is currently pursuing her Ph.D. at Bangladesh University of Engineering and Technology (BUET), Dhaka. She holds an M.Sc. in Computer Science and Engineering from Military Institute of Science and Technology (MIST), another M.Sc. in Management of Technology from BUET and B.Sc. in CSE from MIST, Dhaka, Bangladesh. She has received national and international fellowships and scholarships in recognition of her contributions to healthcare analytics using Artificial Intelligence. She has been an esteemed faculty member (Lecturer) in the Department of Computer Science and Engineering at Bangladesh University of Professionals (BUP), Dhaka. She has published about 26 research articles.

https://orcid.org/0000-0002-7935-3698

B.Sc. Fahamida Hossain Mahi
e-mail: fahamidamahi@gmail.com

Fahamida Hossain Mahi earned her B.Sc. in Computer Science and Engineering from Bangladesh Army International University of Science and Technology (BAIUST). She is passionate about machine learning (ML), deep learning, and software testing. She has co-authored two conference papers in the field of machine learning. She is currently seeking opportunities for higher studies to further her knowledge and expertise in these areas.

https://orcid.org/0009-0006-2624-3993

B.Sc. Forhad Uddin Ahmed
e-mail: forhad.uddin@baiust.edu.bd

Forhad Uddin Ahmed is a Software Engineer at Sicunet Inc. He completed his B.Sc. in Computer Science and Engineering from Bangladesh Army International University of Science and Technology (BAIUST). His research interests lie in machine learning (ML) and artificial intelligence (AI). Forhad has authored two conference papers in the field of machine learning and is an expert in data structures and algorithms (DSA). He has participated in the ACM ICPC Regional competition three times. He is actively seeking PhD opportunities to further his research and contribute to advancements in ML and AI.

https://orcid.org/0009-0008-1513-5238