Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2023, 11, 400-419
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Predicting Future Cryptocurrency Prices Using Machine Learning Algorithms

Vaibhav Saha

Grade 10, Calcutta International School, Kolkata, West Bengal
Email: vaibhavsaha.cis@gmail.com

How to cite this paper: Saha, V. (2023) Predicting Future Cryptocurrency Prices Using Machine Learning Algorithms. *Journal of Data Analysis and Information Processing*, **11**, 400-419.
https://doi.org/10.4236/jdaip.2023.114021

Received: September 8, 2023
Accepted: November 7, 2023
Published: November 10, 2023

Abstract

Cryptocurrency price prediction has garnered significant attention due to the growing importance of digital assets in the financial landscape. This paper presents a comprehensive study on predicting future cryptocurrency prices using machine learning algorithms. Open-source historical data from various cryptocurrency exchanges is utilized. Interpolation techniques are employed to handle missing data, ensuring the completeness and reliability of the dataset. Four technical indicators are selected as features for prediction. The study explores the application of five machine learning algorithms to capture the complex patterns in the highly volatile cryptocurrency market. The findings demonstrate the strengths and limitations of the different approaches, highlighting the significance of feature engineering and algorithm selection in achieving accurate cryptocurrency price predictions. The research contributes valuable insights into the dynamic and rapidly evolving field of cryptocurrency price prediction, assisting investors and traders in making informed decisions amidst the challenges posed by the cryptocurrency market.

Keywords

Cryptocurrency Price Prediction, Machine Learning Algorithms, Feature Engineering, Performance Metrics

1. Introduction

Cryptocurrencies are emerging as a disruptive force in the financial market, introducing decentralized digital assets that operate on blockchain technology. Bitcoin, the pioneering cryptocurrency, is initiating a global surge in digital currencies, leading to the creation of numerous alternative cryptocurrencies, commonly referred to as altcoins. The growing popularity of cryptocurrencies is attracting significant attention from investors, traders, and financial institutions

worldwide. Meanwhile, the decentralized nature, potential for substantial gains, and unique market dynamics of cryptocurrencies are making them a compelling and intriguing asset class. The growing adoption of cryptocurrencies by mainstream institutions and businesses is legitimizing the cryptocurrency market. Additionally, the development of central bank digital currencies is a significant step forward as governments are exploring digital alternatives to their fiat currencies. Interoperability between different blockchain networks, sustainability concerns and regulatory frameworks are also shaping the market's evolution but the extreme volatility and unpredictability of cryptocurrency prices pose considerable challenges for investors seeking to capitalize on market opportunities. Hence, there is a compelling need for robust and accurate predictive models that can assist investors in making informed decisions in this rapidly evolving financial landscape.

The inherent complexities and volatility of cryptocurrency markets are motivating the exploration of innovative approaches to forecast future price movements. Traditional financial models often struggle to capture the unique characteristics of cryptocurrencies, prompting researchers to turn to machine learning algorithms as a potential solution. Machine learning techniques show promise in handling nonlinear relationships and capturing patterns in vast and complex datasets, making them suitable candidates for predicting cryptocurrency prices. By leveraging historical price data and employing sophisticated machine learning algorithms, this research aims to develop predictive models that can discern meaningful trends and patterns.

This research is of practical significance as it can assist investors and traders in making informed decisions, managing risks and potentially increasing their returns in a highly volatile market while also helping policymakers and regulators develop appropriate guidelines and safeguards for the cryptocurrency market, promoting stability and protecting consumers. Moreover, businesses can benefit by incorporating accurate price forecasts into their financial planning and strategies. This study can advance our understanding of the underlying market dynamics and contributes to the broader field of financial analysis, fostering innovation and adaptability in the evolving digital economy.

The primary objective of this research is to predict future cryptocurrency prices using machine learning algorithms. By creating relevant features specific to cryptocurrency price data, such as simple moving average, relative strength index, moving average convergence divergence, and on-balance volume in the chosen dataset, this study seeks to train models that can provide accurate price predictions and valuable insights into potential trends in cryptocurrency prices. The significance of this research lies in its potential to enhance the decision-making process for cryptocurrency investors and traders, helping them strategize their trades, manage risks effectively, and identify lucrative market opportunities. In the field of cryptocurrency, an accuracy of about 60% is considered adequate, but this research aims to achieve over 90% accuracy on some

models. Additionally, this research contributes to the broader field of cryptocurrency price prediction, offering valuable insights into the strengths and limitations of different machine learning algorithms. Overall, this study seeks to bridge the gap between traditional financial models and the unique challenges posed by the cryptocurrency market, thereby contributing to the growing body of knowledge in this rapidly evolving domain.

2. Literature Review

In the realm of cryptocurrency price prediction, researchers are currently exploring various methodologies and approaches. Traditional time series analysis [1], statistical models, and machine learning algorithms [2] are frequently utilized, including support vector machines, random forests, and neural networks, to forecast cryptocurrency prices. These investigations often incorporate historical price data, trading volumes, technical indicators, and market sentiment as predictive features. Additionally, the integration of sentiment analysis from social media and news data helps assess the impact of public perception on price fluctuations [3]. Despite some promising outcomes, the intricate and volatile nature of cryptocurrency markets poses challenges for accurate predictions. As the field progresses, researchers continually seek innovative techniques and incorporate additional data sources to enhance prediction accuracy and account for the dynamic environment of the cryptocurrency market.

In the domain of cryptocurrency price prediction, a variety of machine learning algorithms are harnessed to leverage the predictive potential of data. Traditional statistical models like quadratic discriminant analysis (QDA) [4] and logistic regression [5] are commonly employed for binary classification tasks, attempting to predict whether prices will rise or fall. Decision trees [6] are employed to capture complex interactions among predictors and forecast price movements, serving as maps to understand the impact of these interactions on cryptocurrency prices. K-nearest neighbourhood (KNN) [7] is also utilized to identify similar patterns in historical data and extrapolate price trends. Moreover, neural networks [8], particularly deep learning architectures like long short-term memory (LSTM) networks, are extensively explored to capture sequential patterns in time-series cryptocurrency data. Although each algorithm demonstrates promise in cryptocurrency price prediction, their efficacy often hinges on feature quality and selection, as well as their ability to manage the inherent market volatility and noise.

Current strategies for predicting cryptocurrency prices exhibit distinct strengths and limitations. A notable advantage lies in the application of technical analysis, involving the examination of historical price charts to identify patterns, trends, and support/resistance levels. This approach yields valuable insights into market sentiment and investor behaviour. However, it has constraints, as technical analysis might overlook external factors such as regulatory changes or shifts in market sentiment (which is also a limitation of the chosen methodology for this project). Additionally, fundamental analysis, which assesses cryptocurrency intrinsic val-

ue based on adoption, technology, and utility, offers a long-term perspective on price shifts. Nevertheless, fundamental analysis remains subjective and challenging to quantify, resulting in diverse interpretations and predictions. A more comprehensive perspective on cryptocurrency price movements may emerge from combining multiple approaches that consider both technical and fundamental factors.

3. Data Collection and Preprocessing

Open source historical data is collected to construct the dataset used in this study. It contains Bitcoin price data recorded over 7 months in 2018.

One of the challenges encountered during data preprocessing is dealing with missing values in the historical price dataset. To address this issue, interpolation techniques are employed to estimate and fill in the missing data points. Linear interpolation is used to approximate missing values in the time series data, enabling the construction of a continuous and complete dataset. By employing interpolation, the impact of missing data on the machine learning models' performance is mitigated, ensuring a more robust and informative dataset for predicting cryptocurrency prices. Before training the machine learning algorithms, the collected cryptocurrency price data undergoes several preprocessing steps to ensure data quality and enhance the performance of the models. All numerical data is converted to a single data type (float64), making it easier to work with and manage the dataset. It also helps streamline the analysis process.

Additionally, feature selection is performed to identify the most relevant features for predicting prices, usually used for both cryptocurrency and stock markets. Among the various potential indicators, simple moving average (SMA) [9], relative strength index (RSI) [10], moving average convergence divergence (MACD) [11] and on-balance volume (OBV) [12] are selected as the most influential indicators based on previous literature and domain knowledge. SMA is a suitable feature for prediction because it provides a smoothed representation of historical data, reducing noise and highlighting underlying trends. This makes it valuable for identifying and understanding the direction of price movements over time, serving as a foundation for forecasting future trends in a more stable and interpretable manner. RSI quantifies the momentum of price changes, helping to identify overbought and oversold conditions in an asset. It offers insights into potential trend reversals and can assist in predicting price movements by indicating when an asset is likely to be due for a correction or a continuation of its current trend. MACD combines short-term and long-term moving averages to detect potential trend changes and the strength of price momentum. It provides timely signals for entry and exit points in the market, making it valuable for predicting price movements and identifying potential trading opportunities. OBV helps assess the volume of trading activity accompanying price movements. It reflects buying and selling pressure and can provide early indications of potential trend reversals or continuations. By incorporating volume data, it offers valuable insights for predicting price direction and market trends.

After feature selection, outlier removal is done to remove all infinite values from the dataset to prevent the disruption of calculations and allow a better representation of the underlying patterns due to the reduced influence of extreme values.

4. Methodology

Before carrying out feature engineering, the closing prices of consecutive rows are compared. Each row has a time difference of an hour with the next. If the closing price at a given time is greater than that an hour later, it is assigned a label of 1. Otherwise, it is assigned a label of 0. These labels are stored in a column called 'Label 1', indicating the time difference of 1 hour between the compared closing prices. This process is repeated using 3-hour, 7-hour and 14-hour intervals. These label values are then each assigned to a new column, added to the data frame and named "Label 3", "Label 7" and "Label 14" respectively, with the column name denoting the number of elapsed hours between the compared closing prices. Labelling the data in this manner leads to a more comprehensive understanding of the price fluctuations of cryprocurrency.

Feature engineering techniques specific to cryptocurrency data such as simple moving average, relative strength index, moving average convergence divergence and on balance volume are implemented after labelling the data. The necessary python libraries are imported and then feature engineering is done as described in the following subsections.

a) Simple Moving Average (SMA)

After setting the default style for the plots, the interactive plotting mode is enabled. The column of the data frame containing the closing prices of the cryptocurrency data is selected and a rolling window of size 30 is created. The mean value within each window is computed and the resulting values are assigned to a new column in the data frame. A line graph is made by plotting the SMA and closing price values against the elapsed time in hours as shown in Figure 1.

b) Relative Strength Index (RSI)

The difference between consecutive values in the closing price column is calculated and labelled as "gain" or "loss". The average gain and loss are computed by calculating a rolling mean over a window of 14 hours. Dividing the average gain by the average loss at every hour is done to calculate a new column, the relative strength (RS). Using the formula: 100 − (100/1 + RS), the RSI is calculated for each element and assigned to a new column. An RSI above 70 indicates that a stock is overbought and an RSI below 30 indicates that it is oversold so two red lines are plotted at y = 70 and y = 30 as shown in Figure 2. This indicates the relationship between the RSI and time for the cryptocurrency price data, by displaying the stability of the cryptocurrency.

c) Moving Average Convergence Divergence (MACD)

Two exponential moving averages (EMAs) are calculated, one using 12-hour intervals and the other using 26-hour intervals. The MACD line, the difference

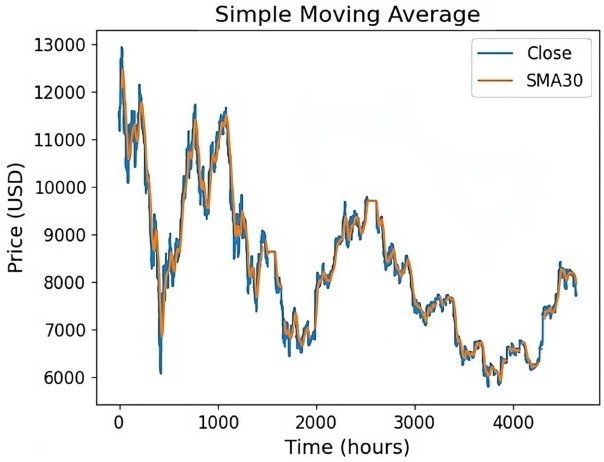

Figure 1. Simple moving average.

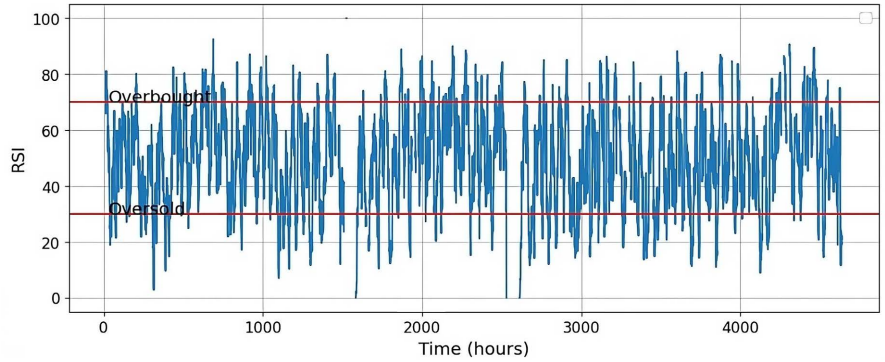

Figure 2. Relative strength index.

between the two EMAs, is plotted in Figure 3. The signal line is calculated by calculating the EMA of the MACD line using 9-hour intervals to provide buy and sell signals, indicating potential shifts in the trends of the cryptocurrency price movement. The two lines are combined into a single data frame to make the plotting easier.

d) On Balance Volume (OBV)

The initial OBV value for all rows is set to 0. The closing price of every row is compared to the row before it. The first row is skipped as it cannot be compared to any value. If the closing price of the row is greater than that of the previous row, its OBV value is updated by adding its volume value to the OBV value of the previous row. If the closing price of the row is lesser than that of the previous row, its OBV value is updated by subtracting its volume value from the OBV value of the previous row. If two consecutive rows have the same value for their closing prices, their OBV value is also equal. These values are then assigned to a new column and plotted in the form of a line plot as shown in Figure 4 below. Negative OBV values are consistently obtained, indicating persistent selling pressure and a pessimistic sentiment in the market.

The following machine learning algorithms are implemented after training them on the cryptocurrency price dataset.

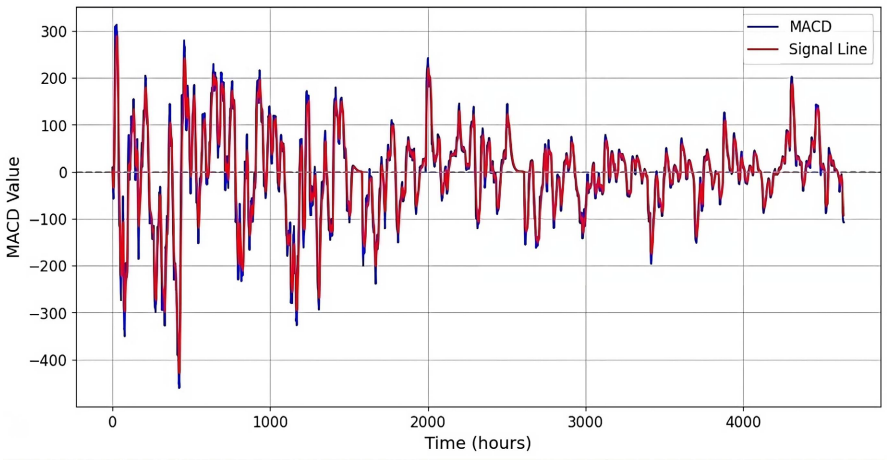

Figure 3. Moving average convergence divergence.

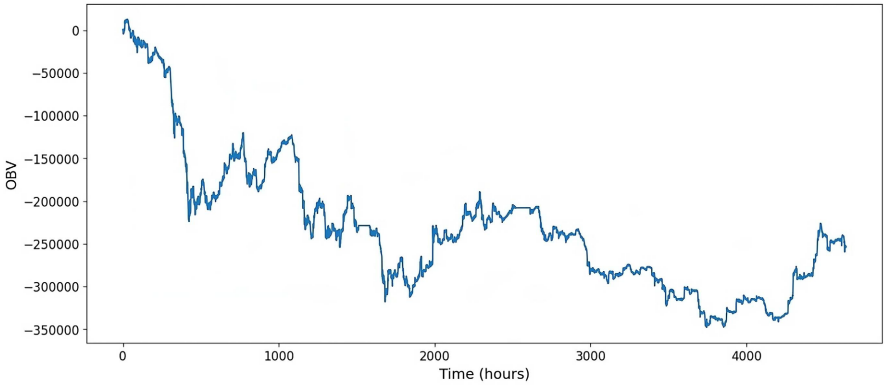

Figure 4. On balance volume.

The following machine learning algorithms are implemented after training them on the cryptocurrency price dataset:

a) Quadratic Discriminant Analysis (QDA)

QDA is a statistical classification technique used to model and predict the classification of data points into multiple classes. It assumes that each class follows a quadratic distribution, meaning that the relationships between features and classes are modeled using quadratic equations. QDA calculates the likelihood of a data point belonging to each class and uses Bayes' theorem to make predictions. In the context of predicting cryptocurrency prices, QDA could be suitable due to its ability to capture complex relationships between multiple features and price movements. Its ability to model non-linear relationships and account for interactions between features can potentially capture the intricate dynamics underlying cryptocurrency price changes. However, the effectiveness of QDA, like any model, would depend on the assumption that the underlying data distribution aligns with its quadratic distribution assumption.

b) K-Nearest Neighbourhood (KNN)

KNN is a simple yet effective machine learning algorithm used for classification tasks. It operates on the principle that similar data points are likely to have

similar outcomes. KNN classifies a new data point by identifying its "k" nearest neighbours in the training data and determining the majority class among them for classification. The choice of "k" impacts the algorithm's sensitivity to noise and generalizability. KNN's ability to capture local patterns makes it useful for short-term predictions, where recent price changes are likely to influence immediate future movements. However, KNN's effectiveness might be limited by sudden market shifts or changes in trading behaviour, as it doesn't inherently capture complex relationships or long-term trends.

c) Logit Model

Logit model, also known as logistic regression, is a statistical technique used for binary classification problems. It models the probability of an instance belonging to a particular class (e.g., 0 or 1) by employing the logistic function to map input features to a range of 0 to 1. This model is particularly suitable for situations where the dependent variable is categorical and involves two outcomes, such as whether an event will happen or not. Logistic regression estimates the coefficients for independent variables and predicts the probability of the positive class, which can then be thresholded for classification. It can be helpful in cryptocurrency price predictions due to its ability to model binary outcomes, which can be applied to certain aspects of cryptocurrency analysis. It can be adapted to predict price movements in cryptocurrencies by treating the problem as a classification task where the outcome is whether the price will increase or decrease within a certain timeframe. Logistic regression can learn patterns that indicate the likelihood of a price increase or decrease. However, it's worth noting that cryptocurrency price prediction is a complex task influenced by a multitude of factors, and while logistic regression can offer insights into certain aspects of this prediction, it may not capture all the nuances inherent in the market.

d) Decision Tree

A decision tree is a machine learning algorithm that models decisions and their possible consequences as a tree-like structure. The tree consists of nodes representing decisions based on input features and branches representing possible outcomes. The algorithm recursively partitions the data into subsets, making decisions based on feature values, until reaching terminal nodes where predictions are made. Decision trees are easy to understand, interpret, and visualize, making them valuable for gaining insights into complex decision-making processes. This model can be suitable for predicting cryptocurrency prices due to its ability to capture nonlinear relationships in data. Cryptocurrency markets are known for their volatility and complex dynamics influenced by various factors. Decision trees can handle such complexities by identifying patterns and nonlinear trends, helping to capture sudden price shifts, making them a suitable choice for cryptocurrency price prediction.

Overfitting is a concern but it can be addressed using techniques like pruning, regularization and ensemble methods. However, these were not used in this paper.

e) Neural Network

A neural network is a computational model inspired by the human brain's structure and functioning. It consists of interconnected nodes or "neurons" organized in layers—an input layer, one or more hidden layers, and an output layer. Neurons in each layer process inputs, apply weights to them, and pass the results through an activation function to produce an output. Through a process called training, the network adjusts its weights using data, enabling it to recognize patterns and relationships in complex datasets. Neural networks are capable of learning and generalizing from examples, making them powerful tools for various machine learning tasks, including prediction and classification. They are well-suited for predicting cryptocurrency prices due to their ability to capture complex nonlinear relationships within data. Neural networks can process and learn from the intricate interactions within cryptocurrency markets, making them effective in capturing both short-term fluctuations and long-term trends. Additionally, neural networks can adapt and learn from new market patterns, providing a dynamic approach to predicting cryptocurrency prices in a highly volatile and evolving landscape.

5. Model Fitting and Evaluation

The dataset is separated into five parts "X", "y", "y3", "y7" and "y14". "X" contains the original cryptocurrency price data along with the selected features while 'y' holds the label value (0 or 1), containing the 1-hour lookahead price direction prediction. Similarly, "y3" contains the 3-hour lookahead predictions, 'y7' contains the 7-hour lookahead predictions and 'y14' contains the 14-hour lookahead predictions. The machine learning algorithms are implemented on the dataset using each of the following combinations once: "X" & "y", "X" & "y3", "X" & "y7" and "X" & "y14".

The following metrics have been used to assess the performance of the machine learning algorithms that are implemented on the dataset:

a) Accuracy [13]: the proportion of correctly classified instances out of the total instances in the dataset.

b) MSE (Mean Squared Error) [14]: The average squared difference between predicted and actual values (lower is better).

c) AUC (Area under the Curve) [15]: The area under the receiver operating characteristic curve, used to evaluate binary classifiers (higher is better).

d) CV Score (Cross-Validation Score) [16]: The performance metric of the model after cross-validation, which helps to assess the generalizability (ability to perform well on unseen data) of the model.

The results obtained after implementing the machine learning algorithms across the four prediction horizons, as well as assessing their performance, are presented and analysed below:

a) Quadratic Discriminant Analysis

The results obtained after implementing QDA are shown in **Table 1**. For the 1-hour lookahead prediction, a very high accuracy of 0.998 is achieved, indicating

Table 1. Performance of quadratic discriminant analysis.

Performance Metric	Machine Learning Model			
	Quadratic Discriminant Analysis			
	1-Hour Lookahead Predictions	3-Hour Lookahead Predictions	7-Hour Lookahead Predictions	14-Hour Lookahead Predictions
Accuracy	0.998	0.711	0.768	0.941
MSE	0.005	0.289	0.232	0.059
AUC	0.317	0.421	0.447	0.529
CV Score	0.998	0.698	0.756	0.932

the model's extremely good performance. A low MSE of 0.005 suggests that the predictions barely deviate from the true data. A low AUC of 0.317 indicates that the model's ability to differentiate between price rises and falls is not even better than random chance. A high CV score of 0.998 implies excellent generalizability for this price forecast. For the 3-hour lookahead prediction, accuracy drops to 0.711. MSE increases to 0.289, indicating less accurate predictions. AUC increases to 0.421, which is better, but still not very high. The CV score drops to 0.698, suggesting slightly worse generalizability compared to the 1-hour forecast. Accuracy further decreases to 0.768 for the 7-hour lookahead prediction. MSE decreases to 0.232, which is better than the 3-hour forecast but still worse than the 1-hour forecast. AUC increases to 0.447, showing improvement in binary classification performance. CV score decreases to 0.756, indicating slightly less generalizability compared to the 3-hour forecast. Accuracy is 0.941, showing good performance on the data obtained by the 14-hour lookahead predictions. MSE is only 0.059, indicating highly accurate predictions due a lower margin of error. AUC is 0.529, which is higher than all previous price forecasts but still relatively low. The CV score is 0.932, showing good generalizability on the 14-hour price forecast data. In summary, the model performs best on the 1-hour lookahead predictions and its performance worsens as the length of the price forecasting periods increases. The model's binary classification performance can be improved for all the forecasting periods. Cross-validation indicates that the model's performance is consistent and capable of generalizing well on most prediction horizons. However, for the 7-hour lookahead predictions, it shows slightly less generalizability compared to the others. Further analysis and optimization can be done to improve the binary classification performance for all the prediction horizons.

b) K-Nearest Neighbourhood

The results obtained after implementing KNN are shown in Table 2. The model's accuracy is 0.497 for the 1-hour lookahead predictions, indicating that it performs moderately well. The MSE is 0.418, which suggests that the model's predictions have considerable error. The AUC is 0.5, which means the binary classification performance is not better than random guessing. The cross-validation score is 0.489, indicating moderate generalizability. The accuracy increases slightly

Table 2. Performance of K-Nearest neighbourhood.

Performance Metric	Machine Learning Model			
	K-Nearest Neighbourhood			
	1-Hour Lookahead Predictions	3-Hour Lookahead Predictions	7-Hour Lookahead Predictions	14-Hour Lookahead Predictions
Accuracy	0.497	0.537	0.591	0.674
MSE	0.418	0.369	0.274	0.059
AUC	0.5	0.5	0.5	0.5
CV Score	0.489	0.502	0.501	0.507

to 0.537 for the 3-hour lookahead predictions. The MSE decreases to 0.369, indicating slightly better performance compared to the 1-hour price forecast. The AUC remains at 0.5, suggesting the binary classification performance is still no better than random guessing. The cross-validation score is 0.502, showing limited generalizability. The accuracy increases to 0.591 for the 7-hour lookahead predictions. The MSE decreases to 0.274, showing improved performance compared to previous forecasting periods. The AUC remains at 0.5, indicating no improvement in binary classification performance. The cross-validation score is 0.501, suggesting limited generalizability. The accuracy increases to 0.674 for the 14-hour lookahead predictions. The MSE is 0.059, indicating relatively accurate predictions, which is an improvement over the previous prediction horizons. The AUC remains at 0.5, implying no improvement in binary classification performance. The cross-validation score is 0.507, indicating limited generalizability. In summary, the KNN algorithm does not perform well on any of the forecasting periods. Its accuracy and binary classification performance are all mediocre or worse. The model's generalizability to new data is limited, as indicated by the CV scores, which are close to 0.5 for all prediction horizons. This suggests that the model is not effectively learning patterns in the data and is not suitable for the given task. Further analysis and potentially using a different model or optimizing the KNN hyperparameters can be helpful in improving performance.

c) Logit Model

The results obtained after implementing logit model are shown in Table 3. The model's accuracy is high at 0.979 for the 1-hour lookahead predictions, indicating that it performs very well. The MSE is 0.418, which suggests the model's predictions contain some error. The AUC is 0.633, showing decent binary classification performance. The cross-validation score is even higher at 0.987, indicating excellent generalizability. The accuracy decreases to 0.737 for the 3-hour price forecasts. The MSE is 0.369, showing a similar performance compared to the 1-hour forecasts. The AUC increases to 0.718, indicating improved binary classification performance. The cross-validation score is 0.735, suggesting good generalizability. The accuracy increases to 0.787 for the 7-hour lookahead

Table 3. Performance of logit model.

Performance Metric	Machine Learning Model			
	Logit Model			
	1-Hour Lookahead Predictions	3-Hour Lookahead Predictions	7-Hour Lookahead Predictions	14-Hour Lookahead Predictions
Accuracy	0.979	0.737	0.787	0.899
MSE	0.418	0.369	0.274	0.059
AUC	0.633	0.718	0.779	0.931
CV Score	0.987	0.735	0.779	0.878

predictions. The MSE is 0.274, showing better performance than the previous forecasting periods. The AUC increases to 0.779, indicating better binary classification performance. The cross-validation score is 0.779, showing good generalizability. The accuracy further increases to 0.899 for the 14-hour lookahead predictions. The MSE is 0.059, indicating highly accurate predictions, an improvement over the previous prediction horizons. The AUC increases significantly to 0.931, indicating excellent binary classification performance. The cross-validation score is 0.878, suggesting good generalizability. In summary, the logit (logistic regression) model performs exceptionally well on all forecasting periods. Its accuracy and binary classification performance show that it is more suitable for long-term predictions. Additionally, the model's generalizability is remarkable, as indicated by the high cross-validation scores for all prediction horizons. Logistic regression is a well-suited model for the given task and it effectively captures patterns in the data to make accurate predictions and perform binary classification tasks.

d) Decision Tree

The results obtained after implementing decision tree are shown in Table 4. The model's accuracy using the 1-hour lookahead predictions is exceptionally high at 1, indicating it performs perfectly. The MSE is 0.418, which suggests the model's predictions have some error. The AUC is 0.5, indicating that the binary classification performance is no better than random guessing. The cross-validation score is 0.999, suggesting outstanding generalizability. The accuracy drops to 0.617 for the 3-hour lookahead predictions. The MSE is 0.369, showing a similar performance compared to the 1-hour price forecasts. The AUC remains at 0.5, indicating poor binary classification performance. The cross-validation score is 0.599, suggesting limited generalizability. The accuracy increases to 0.704 for the 7-hour price forecasts. The MSE falls to 0.274, showing good performance compared to the previous prediction horizons. The AUC remains at 0.5, indicating poor binary classification performance. The cross-validation score is 0.633, suggesting moderate generalizability. The accuracy is again exceptionally high at 1 for the 14-hour lookahead predictions. The MSE decreases significantly to 0.059, indicating highly accurate predictions, an improvement over the previous

Table 4. Performance of decision tree.

Performance Metric	Machine Learning Model			
	Decision Tree			
	1-Hour Lookahead Predictions	3-Hour Lookahead Predictions	7-Hour Lookahead Predictions	14-Hour Lookahead Predictions
Accuracy	1	0.617	0.704	1
MSE	0.418	0.369	0.274	0.059
AUC	0.5	0.5	0.5	0.5
CV Score	0.999	0.599	0.633	1

forecasting periods. The AUC remains at 0.5, implying poor binary classification performance. The cross-validation score is 1, suggesting excellent generalizability. In summary, the decision tree model performs remarkably well on 1 and 14-hour price forecasts, achieving perfect accuracy and excellent generalizability. However, its performance decreases significantly for 3 and 7-hour lookahead predictions, where accuracy and AUC are relatively low. The model achieved the lowest mean squared error for the 14-hour price forecasts. Further analysis and potentially using different tree-based algorithms or adjusting hyperparameters can help improve the model's performance for the 3 and 7-hour lookahead predictions.

e) Neural Network

The results obtained after implementing neural network for classification are shown in Table 5. The model's accuracy is high at 0.991 for the 1-hour price forecasts, indicating it performs very well. The MSE is 0.286, which suggests the model's predictions have low error. The AUC is 0.598, indicating decent binary classification performance. The cross-validation score is 0.553, suggesting moderate generalizability. The accuracy drops to 0.679 for the 3-hour lookahead predictions. The MSE is 0.278, showing better performance compared to the 1-hour forecasts. The AUC increases to 0.668, indicating improved binary classification performance. The cross-validation score is 0.572, suggesting moderate generalizability. The accuracy increases to 0.736 for the 7-hour lookahead predictions. The MSE is 0.281, showing similar performance compared to the 3-hour lookahead predictions. The AUC increases to 0.781, indicating good binary classification performance. The cross-validation score is 0.576, suggesting moderate generalizability. The accuracy is exceptionally high at 0.997 for the 14-hour price forecasts. The MSE is 0.266, indicating highly accurate predictions, an improvement over the previous prediction horizons. The AUC increases significantly to 0.999, indicating excellent binary classification performance. The cross-validation score is 0.584, suggesting moderate generalizability. In summary, the neural network model performs very well on most forecasting periods. Its accuracy and binary classification performance generally improved as we move from 3 to 14-hour lookahead predictions, indicating its helpfulness as a

Table 5. Performance of neural network.

Performance Metric	Machine Learning Model			
	Neural Network			
	1-Hour Lookahead Predictions	3-Hour Lookahead Predictions	7-Hour Lookahead Predictions	14-Hour Lookahead Predictions
Accuracy	0.991	0.679	0.736	0.997
MSE	0.286	0.278	0.281	0.266
AUC	0.598	0.668	0.781	0.999
CV Score	0.553	0.572	0.576	0.584

long-term price prediction algorithm. The model has moderate to good generalizability, as indicated by the cross-validation scores. Notably, the model achieves exceptional accuracy and binary classification performance for the 14-hour price forecasts, suggesting it is particularly well-suited for this forecasting period. However, there is some room for improvement in generalizability, especially for the 1-hour lookahead predictions. Fine-tuning the neural network's architecture and hyperparameters can potentially improve its performance for all prediction horizons.

Comparing the performances of the different models above, significant variations in their abilities to handle the diverse prediction horizons and the tasks associated with them are observed. The QDA model exhibits outstanding performance for the 1-hour lookahead predictions, achieving near-perfect accuracy and a high cross-validation score. However, its binary classification performance, as indicated by the AUC values, is relatively poor across all prediction horizons, suggesting limited discriminative ability. On the other hand, the logit model shows remarkable performance across all prediction horizons, with consistently high accuracy and low error. It excels in binary classification, particularly for the 14-hour forecasts. Additionally, the model demonstrates excellent generalizability, as seen in the high cross-validation scores. The decision tree model performs relatively well for the 1 and 14-hour lookahead predictions but struggles on the 3 and 7-hour predictions, indicating variations in performance across different prediction horizons. It demonstrates perfect accuracy for the 1 and 14-hour price forecasts but lacks the ability to distinguish between classes in the binary classification tasks, as evidenced by the consistently low AUC values. Moreover, the model's generalizability appears to be limited, as reflected in the relatively low cross-validation scores for the 3 and 7-hour forecasts. Lastly, the neural network model displays a mix of strengths and weaknesses across the forecasting periods. It achieves outstanding accuracy for the 1 and 14-hour forecasts, showcasing its potential for complex classification tasks. However, its performance for the 3 and 7-hour lookahead predictions is less impressive, where the accuracy drops and AUC values are modest. The model exhibits decent generalizability, but some variations in cross-validation scores suggest room for

improvement in handling certain classes.

The impact created by employing interpolation and feature engineering is shown in Table 6. One of the best performing models, the logit model, is implemented on the dataset thrice, each time with a distinct configuration: once employing interpolation and feature engineering, another time without interpolation and a third time without feature engineering. In terms of accuracy, the model achieves a commendable score of 0.899 on the dataset with interpolation and feature engineering, signifying that it accurately predicts approximately 89.9% of instances. Interestingly, the model without interpolation surpasses this performance, recording an accuracy of 0.973. However, the model without feature engineering demonstrates comparatively weaker predictive power, registering an accuracy of 0.546. Comparing the MSE of the model on the three occasions unveils distinctions in error magnitude. Notably, the model devoid of feature engineering displays the highest MSE at 0.458, indicative of larger prediction errors. Conversely, the model without interpolation exhibits the lowest MSE at 0.027, signifying enhanced predictive accuracy, beating the model with both interpolation and feature engineering as that obtained an MSE of 0.059. Moreover, the AUC values shed light on classification proficiency, with the model lacking interpolation yielding an impressive AUC of 0.998, signifying robust classification performance. Additionally, the model with interpolation and feature engineering demonstrates competitive AUC at 0.931, while the model without feature engineering lags with an AUC of 0.572. CV Scores serve to assess the generalizability, showcasing marginal differences across the models. Specifically, the model devoid of interpolation garners the highest CV score of 0.882, followed closely by the model with interpolation and feature engineering at 0.878, and the model without feature engineering displaying a lower CV score of 0.556. In comparing the three models, it's evident that the model without interpolation consistently showcases strong predictive prowess, excelling in accuracy and AUC metrics. On the other hand, the model with interpolation and feature engineering provides a balance between accuracy and precision, with competitive performance across various metric while the model without feature engineering falls short in multiple aspects, displaying relatively weaker accuracy and AUC scores as well as a high amount of error. The contrast between the models underscores the importance of strategic feature engineering and interpolation

Table 6. Effect of interpolation and feature engineering.

Performance Metric	Logit Model's 14-Hour Lookahead Predictions		
	With Interpolation and Feature Engineering	Without Interpolation	Without Feature Engineering
Accuracy	0.899	0.973	0.546
MSE	0.059	0.027	0.458
AUC	0.931	0.998	0.572
CV Score	0.878	0.882	0.556

techniques in enhancing predictive capabilities. While the model with both features and interpolation performs admirably, the choices surrounding feature engineering and interpolation are pivotal factors that influence the models' overall effectiveness in predicting cryptocurrency prices. Further experimentation and evaluation are crucial for the fine-tuning of these models and the identification of the optimal configuration for achieving accurate and reliable predictions in the volatile cryptocurrency market.

6. Results and Discussion

Among the models mentioned above, logit model is likely to be the best choice for predicting cryptocurrency prices. It consistently demonstrates remarkable performance across all prediction horizons, with high accuracy and low MSE. Its binary classification ability is particularly strong, as evidenced by the high AUC value for the 14-hour lookahead predictions. Additionally, the model exhibits outstanding generalizability, as indicated by the high cross-validation scores. Predicting cryptocurrency prices involves binary classification tasks (e.g., predicting price movements as "up" or "down"). Logit model's ability to handle classification tasks effectively makes it a suitable choice for cryptocurrency price prediction. However, it is crucial to note that the choice of the best model ultimately depends on the specific dataset, features generated and the complexity of the cryptocurrency price prediction problem. Careful experimentation and tuning of hyperparameters can help obtain the optimal performance for a given dataset.

The findings of this study are in line with several previous research works. Similar to previous studies, technical indicators such as SMA, RSI, MACD and OBV are used as features [17] for the predictive models, which have been considered influential in capturing market trends and price movements. This research confirms the importance of feature engineering in cryptocurrency price prediction as running a model without feature engineering caused key performance metrics like accuracy and AUC to plummet. The use of interpolation to handle missing data is also a widely adopted practice in cryptocurrency price prediction studies [18]. The findings of this study, however, show that more accurate results were obtained when the missing values are omitted instead of interpolating the dataset. This may have occurred as the missing values contribute to less than 1.5% of the entire datset used in this research. Moreover, the comparison of different machine learning algorithms' performance corroborates earlier studies' insights into the performance of various models in cryptocurrency price prediction. Logit model demonstrates outstanding performance with high accuracy and excellent classification capabilities, consistent with its widespread application in financial prediction tasks [19]. The neural network model's varying performance across different prediction horizons highlights the need for further exploration of complex architectures and hyperparameter tuning. However, its accuracy for longer forecasting periods is fairly high, making it a suitable predictor in many cases [20]. The study's results reinforce the notion that the choice of machine learning algorithm plays a critical role in achieving accurate and re-

liable cryptocurrency price predictions [21]. The use of open source data, technical indicators, and interpolation techniques, along with the comparison of machine learning algorithms, adds to the growing understanding of effective methodologies in this rapidly evolving field.

This study has some limitations and potential sources of bias that should be acknowledged. Firstly, the use of historical cryptocurrency price data might not fully represent the highly volatile and dynamic nature of the cryptocurrency market. Market sentiment, regulatory changes, and other external factors that can significantly impact cryptocurrency prices are not possible to capture solely through historical price data (external data sources are required). The choice of machine learning algorithms and hyperparameter tuning could introduce bias towards specific models and settings, potentially influencing the performance results. Moreover, the feature engineering techniques employed in the study may not capture all relevant information, and other meaningful features could be overlooked, affecting the predictive accuracy. Additionally, there may be overfitting issues as the same data was used for model selection and hyperparameter tuning. The evaluation metrics chosen might not fully capture all aspects of model performance, and using only a few metrics may not provide a comprehensive view of the models' strengths and weaknesses. Ensemble learning techniques [22] could have been used to combine the predictions of multiple machine learning models and add to the prediction accuracy and robustness. Applying regularization techniques such as lasso and ridge [23] could have helped combat overfitting issues. Experimentation with different regularization strengths could have helped find the optimal balance between model complexity and generalization. Furthermore, subjecting the models to stress tests [24] such as extreme market conditions or data anomalies could have helped assess their performance under adverse conditions. All of these factors have an effect on the predictive accuracy of the models so incorporating them could have led to a more comprehensive interpretation of the models' capability to predict future cryptocurrency prices.

Despite the limitations, this research provides valuable insights into predicting cryptocurrency prices using machine learning algorithms. To address the identified limitations and potential sources of bias, future research should focus on expanding the dataset to include data from multiple exchanges and time periods, which will increase the sample size and provide diverse market insights, as well as incorporating other relevant features such as relative vigor index (RVI) and commodity channel index (CCI). Employing more sophisticated techniques, such as sentiment analysis or incorporating external data, can enhance the model's performance and improve its generalizability. Exploring ensemble methods, such as combining the predictions of multiple models, may further enhance the predictive accuracy and robustness. By addressing these areas, future studies can contribute to the development of more reliable and accurate models for predicting cryptocurrency prices, offering valuable insights to cryptocurrency investors and the financial market.

7. Conclusions

This study aimed to predict future cryptocurrency prices using machine learning algorithms. Historical cryptocurrency price data was obtained and data preprocessing techniques to prepare the data for model training were employed. Feature engineering techniques specific to cryptocurrency data were used to create relevant features. Five machine learning algorithms were implemented and their performances were evaluated using various performance metrics. Through rigorous analysis, insights into the predictive capabilities of each model were gained.

This research has yielded several key findings and contributions to the field of cryptocurrency price prediction. Firstly, the logistic regression model demonstrated outstanding performance across all prediction horizons, with high accuracy and good binary classification performance. It proved to be a useful choice for handling binary classification tasks. Secondly, the neural network model showed strong predictive capabilities on certain prediction horizons, particularly the 14-hour lookahead predictions, but its performance varied across different classes. This highlights the importance of considering the diversity of the cryptocurrency market when selecting the appropriate model. Lastly, the study provided a comprehensive comparison of the different machine learning algorithms' performance, shedding light on their respective strengths and limitations.

The findings of this research have profound ramifications for the cryptocurrency market as accurate price predictions can help investors make informed decisions and better manage risks in their cryptocurrency trading strategies. By understanding the strengths and limitations of different machine learning algorithms, investors can select appropriate models to suit their trading preferences. Moreover, the research contributes to the growing body of knowledge in the field of cryptocurrency price prediction, paving the way for more advanced and reliable models to be developed in the future.

In conclusion, this study highlights the potential of machine learning algorithms in predicting future cryptocurrency prices. While the logit model stands out as a robust and versatile choice, the neural network model also shows promise with its strong performance on specific prediction horizons. Nevertheless, it is essential to be aware of the limitations and potential sources of bias in the study, such as data selection, algorithm choice and evaluation metrics. Future research should focus on addressing these limitations and exploring more sophisticated techniques to further enhance the accuracy and generalizability of the prediction models. Overall, this study provides significant insights to the realm of forecasting cryptocurrency prices, establishing a foundation for upcoming developments in this dynamic and swiftly progressing area.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Karasu, S., Altan, A., Saraç, Z. and Hacioğlu, R. (2018, May) Prediction of Bitcoin Prices with Machine Learning Methods Using Time Series Data. 2018 26*th Signal Processing and Communications Applications Conference* (*SIU*), Izmir, 2-5 May 2018, 1-4. https://doi.org/10.1109/SIU.2018.8404760

[2] Derbentsev, V., Matviychuk, A. and Soloviev, V.N. (2020) Forecasting of Cryptocurrency Prices Using Machine Learning. In: Pichl, L., Eom, C., Scalas, E. and Kaizoji, T., Eds., *Advanced Studies of Financial Technologies and Cryptocurrency Markets*, Springer, Berlin, 211-231. http://dx.doi.org/10.1007/978-981-15-4498-9_12

[3] Abraham, J., Higdon, D., Nelson, J. and Ibarra, J. (2018) Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Science Review*, **1**, Article No. 1.

[4] Rakotomarolahy, P. (2021) Predicting the Bitcoin Return Direction with Logistic, Discriminant Analysis and Machine Learning Classification Techniques. *Model Assisted Statistics and Applications*, **16**, 169-176. https://doi.org/10.3233/MAS-210530

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

[6] Rathan, K., Sai, S.V. and Manikanta, T.S. (2019, April) Crypto-Currency Price Prediction Using Decision Tree and Regression Techniques. 2019 3*rd International Conference on Trends in Electronics and Informatics* (*ICOEI*), Tirunelveli, 23-25 April 2019, 190-194. https://doi.org/10.1109/ICOEI.2019.8862585

[7] Fatah, H. and Subekti, A. (2018) Cryptocurrency Price Prediction Using the K-Nearest Neighbors Method. *Pilar Nusa Mandiri Journal*, **14**, 137-144. https://doi.org/10.33480/pilar.v14i2.30

[8] Jay, P., Kalariya, V., Parmar, P., Tanwar, S., Kumar, N. and Alazab, M. (2020) Stochastic Neural Networks for Cryptocurrency Price Prediction. *IEEE Access*, **8**, 82804-82818. https://doi.org/10.1109/ACCESS.2020.2990659

[9] Yudono, M.A.S., Sidik, A.D.W.M., Kusumah, I.H., Suryana, A., Junfithrana, A.P., Nugraha, A. and Imamulhak, Y. (2022) Bitcoin USD Closing Price (BTC-USD) Comparison Using Simple Moving Average and Radial Basis Function Neural Network Methods. *FIDELITY*: *Jurnal Teknik Elektro*, **4**, 29-34. https://doi.org/10.52005/fidelity.v4i2.74

[10] Zatwarnicki, M., Zatwarnicki, K. and Stolarski, P. (2023) Effectiveness of the Relative Strength Index Signals in Timing the Cryptocurrency Market. *Sensors*, **23**, Article No. 1664. https://doi.org/10.3390/s23031664

[11] Aguirre, A.A.A., Medina, R.A.R. and Méndez, N.D.D. (2020) Machine Learning Applied in the Stock Market through the Moving Average Convergence Divergence (MACD) Indicator. *Investment Management & Financial Innovations*, **17**, 44-60. http://dx.doi.org/10.21511/imfi.17(4).2020.05

[12] Bruno, K.W. (1984) An Objective Empirical and Statistical Analysis of Joseph Granville's On-Balance Volume Stock Analysis Technique. Lamar University, Beaumont.

[13] Koosha, E., Seighaly, M. and Abbasi, E. (2022) Measuring the Accuracy and Precision of Random Forest, Long Short-Term Memory, and Recurrent Neural Network Models in Predicting the Top and Bottom of Bitcoin Price. *Journal of Mathematics*

and Modeling in Finance, **2**, 107-128.
https://doi.org/10.22054/jmmf.2023.15183

[14] Thompson, P.A. (1990) An MSE Statistic for Comparing Forecast Accuracy across Series. *International Journal of Forecasting*, **6**, 219-227.
https://doi.org/10.1016/0169-2070(90)90007-X

[15] Carrington, A.M., Manuel, D.G., Fieguth, P.W., Ramsay, T., Osmani, V., Wernly, B. and Holzinger, A. (2021) Deep ROC Analysis and AUC as Balanced Average Accuracy to Improve Model Selection, Understanding and Interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 329-341.
https://doi.org/10.48550/arXiv.2103.11357

[16] Al Helal, M., Haydar, M.S. and Mostafa, S.A.M. (2016, December) Algorithms Efficiency Measurement on Imbalanced Data Using Geometric Mean and Cross Validation. 2016 *International Workshop on Computational Intelligence* (*IWCI*), Dhaka, 12-13 December 2016, 110-114.
https://doi.org/10.1109/IWCI.2016.7860349

[17] Fu, D. and Ismail, M.T. (2023) The Long Short-Term Memory (LSTM) Model Combines with Technical Analysis to Forecast Cryptocurrency Prices. *MATEMATIKA*: *Malaysian Journal of Industrial and Applied Mathematics*, **39**, 149-158.

[18] Tiwari, R.G., Agarwal, A.K., Kaushal, R.K. and Kumar, N. (2021, October) Prophetic Analysis of Bitcoin Price Using Machine Learning Approaches. 2021 6*th International Conference on Signal Processing*, *Computing and Control* (*ISPCC*), Solan, 7-9 October 2021, 428-432.

[19] Colianni, S., Rosales, S. and Signorotti, M. (2015) Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis. *CS*229 *Project*, **1**, 1-4.

[20] Charandabi, S.E. and Kamyar, K. (2021) Using a Feed Forward Neural Network Algorithm to Predict Prices of Multiple Cryptocurrencies. *European Journal of Business and Management Research*, **6**, 15-19.
https://doi.org/10.24018/ejbmr.2021.6.5.1056

[21] Awotunde, J.B., Ogundokun, R.O., Jimoh, R.G., Misra, S. and Aro, T.O. (2021) Machine Learning Algorithm for Cryptocurrencies Price Prediction. In: Misra, S. and Tyagi, A.K., Eds., *Artificial Intelligence for Cyber Security*: *Methods*, *Issues and Possible Horizons or Opportunities*, Springer International Publishing, Cham, 421-447.
https://doi.org/10.1007/978-3-030-72236-4_17

[22] Webb, G.I. and Zheng, Z. (2004) Multistrategy Ensemble Learning: Reducing Error by Combining Ensemble Learning Techniques. *IEEE Transactions on Knowledge and Data Engineering*, **16**, 980-991. https://doi.org/10.1109/TKDE.2004.29

[23] Melkumova, L.E. and Shatskikh, S.Y. (2017) Comparing Ridge and LASSO Estimators for Data Analysis. *Procedia Engineering*, **201**, 746-755.
https://doi.org/10.1016/j.proeng.2017.09.615

[24] Aluç, G., Hartig, O., Özsu, M.T. and Daudjee, K. (2014) Diversified Stress Testing of RDF Data Management Systems. *The Semantic Web-ISWC* 2014: 13*th International Semantic Web Conference*, Riva del Garda, 19-23 October 2014, 197-212.
https://doi.org/10.1007/978-3-319-11964-9_13