Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 5 No. 2 (2025) 220 - 229 E-ISSN: 2809-8064

Comparison of Accuracy of Linear Regression and Random Forest Models in Predicting Bitcoin Prices

Ahmad Habib Awwaluddin[1*], Teguh Tamrin[2] , Nur Aeni Widiastuti[3]

Informatics Engineering, Faculty of Science and Technology, Nahdlatul Ulama Islamic University of Jepara

[1*]211240001122@unisnu.ac.id, [2]teguh@unisnu.ac.id, [3]nuraeniwidiastuti@unisnu.ac.id

Abstract

Bitcoin is one of the digital assets with high price volatility, making its price movement prediction a crucial challenge for market participants. This study aims to compare the accuracy of two predictive algorithms—Linear Regression and Random Forest—in forecasting Bitcoin's closing price based on daily historical data from 2018 to 2025, obtained from the Kaggle platform. The research process includes data preprocessing, construction of predictive features (Open, High, Low, Volume), and normalization. The models were evaluated using two validation schemes: a 70:30 data split and 10-fold Cross-Validation, along with three main evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Additionally, a simple baseline model (naive prediction) was used as an initial benchmark, and the Wilcoxon Signed-Rank test was applied to assess the statistical significance of performance differences. The results show that Linear Regression outperformed both Random Forest and the baseline model, achieving the lowest RMSE of 1314.47 under cross-validation. However, the Wilcoxon test indicated that the performance difference was not statistically significant. This study recommends exploring other models such as Ridge and Lasso Regression, as well as enhancing features and applying time-aware validation in future research. These findings are expected to serve as a reference for developing efficient and accurate machine-learning-based cryptocurrency price prediction systems.

Keywords: Bitcoin, price prediction, Linear Regression, Random Forest, model validation, baseline, Wilcoxon test

1. Introduction

Bitcoin is one of the most popular digital assets in the world and a pioneer of the blockchain-based cryptocurrency system. Since it was introduced by Satoshi Nakamoto in 2009, Bitcoin has evolved from an experimental system into one of the most valuable investment assets with the largest market capitalization globally [1].

However, despite Bitcoin's long-term value growth, its high volatility makes it a risky asset. Bitcoin's price can change drastically in a short period, influenced by various factors such as government regulations, public opinion, system hacks, and even tweets from influential figures on social media. This condition presents a unique challenge for investors and analysts to make accurate predictions [2].

Bitcoin price prediction plays a crucial role in supporting financial decision-making, especially in an unstable crypto market. Therefore, an approach that not only relies on manual technical or fundamental analysis but also utilizes technological advancements in machine learning is needed to model patterns and trends from historical data.

In the field of data science, predictive algorithms such as Linear Regression and Random Forest are commonly used to model relationships between variables. Linear Regression is favored for its ease of interpretation and computational efficiency, while Random Forest is known for its strength in handling data complexity and its resistance to overfitting [3].

This study compares the performance of Linear Regression and Random Forest in the context of Bitcoin price prediction. The choice of these two algorithms is

Received: 28-07-2025 | Accepted: 13-08-2025 | Published: 15-08-2025

based on their contrasting approaches—linear versus non-linear, simple model versus ensemble learning. By comparing the two, it is hoped that the most suitable model for predicting highly fluctuating and non-deterministic crypto data can be identified.

The data used in this study was obtained from the open platform Kaggle, consisting of daily Bitcoin prices from 2018 to 2025. This dataset was selected for its comprehensiveness and its ability to represent market dynamics over the medium to long term. Before modeling, the data underwent several preprocessing stages, including data cleaning, normalization, and the construction of predictive features based on historical values such as Open, High, Low, and Volume.

The models were evaluated using two validation approaches: 70:30 data split (train-test split) and 10-fold cross-validation. However, it is important to note that standard cross-validation methods like K-Fold may lead to data leakage in time series contexts, as they do not preserve the chronological order of the data. Therefore, this study also recommends the adoption of time-aware validation methods (Time Series Cross-Validation) in future research to improve the reliability of model evaluation [4].

The performance of both models was assessed using three primary evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics are widely used in regression-based predictive modeling [5]. In addition, a baseline model using naive prediction—which assumes the closing price of the current day as the prediction for the next day—was used as an initial benchmark. This approach serves as a fundamental comparison to determine whether the tested models offer significant performance improvements. To further strengthen the comparative analysis, this study also employed the non-parametric Wilcoxon Signed-Rank Test to assess the statistical significance of performance differences across each cross-validation fold, ensuring that the evaluation results are not only descriptive but also statistically supported.

Previous studies have shown that Linear Regression tends to produce stable results on linear data but is less responsive to dynamic changes. On the other hand, Random Forest often yields more precise predictions on complex patterns, albeit at a higher computational cost. Therefore, an empirical analysis is needed to determine which model is more appropriate in the context of Bitcoin price prediction[6].

The results of this study are expected to provide meaningful contributions to the development of data-driven predictive systems, and serve as a reference for investors or digital financial system developers in selecting the most appropriate predictive approach for assets with high volatility characteristics like Bitcoin.

2. Research Methods

In preparing this research report, several stages were systematically carried out. The research began with data collection, followed by pre-processing and data processing, the application of two predictive models: Linear Regression and Random Forest, model performance evaluation using two validation schemes, and finally, analysis of the results and drawing conclusions. The stages in this research process are depicted in Figure 1 below.

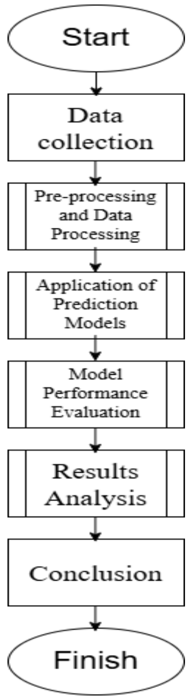

Figure 1. Process Flow

2.1. Dataset and Data Source

This study uses a historical dataset of Bitcoin prices obtained from the Kaggle platform. The dataset includes daily Bitcoin price data from 2018 to 2025. It consists of several columns, including: Open Time, Open, High, Low, Close, Volume, and Quote Asset Volume. In this research, the Close column is selected as the target for prediction because it represents the daily closing price of Bitcoin. The data was then cleaned and formatted to meet the requirements of the modeling process.

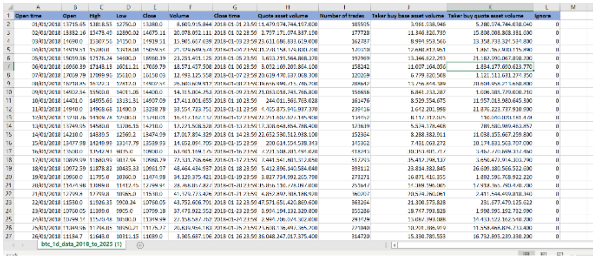

Figure 2. Dataset Bitcoin

2.2. *Data Preprocessing and Processing*

Before the data is used to build a prediction model, the preprocessing stage is an essential step to ensure data quality and readiness for analysis. Data preprocessing is crucial because raw data often contains issues such as missing values, duplicate data, format inconsistencies, and varying numerical scales[7]. The preprocessing and data handling steps in this study are as follows:

2.2.1 Data Cleaning

In this stage, invalid or incomplete data is identified and handled. For example, if there are null values in the price column, those rows are removed to maintain consistency and data integrity. Cleaning also includes removing duplicate records that could affect analysis results, and ensuring the correct format for date and numeric values. This process is important to prevent bias or errors in model training caused by invalid data.

2.2.2 Data Normalization

Normalization is carried out to standardize the scale of numerical features. This is particularly important when using algorithms sensitive to feature scale, such as Linear Regression. For instance, Bitcoin prices can reach tens of thousands of dollars, while other features such as trading volume may have vastly different scales. Therefore, value transformation is performed to bring all features into a comparable scale, using methods such as Min-Max Scaling or Standard Scaling[8].

2.2.3 Data Splitting

After cleaning and normalization, the next step is to split the data into training and testing sets. In this study, 70% of the data is used to train the model, and the remaining 30% is used to test model performance. The purpose of this split is to evaluate how well the trained model can predict data it has never seen before (generalization test).

2.2.4 Pemilihan Fitur

In addition to the three steps above, the data processing stage also includes selecting the most relevant features for the prediction target, namely the closing price (Close). The features used in this study include opening price (Open), highest price (High), lowest price (Low), and transaction volume. These features are selected based on the assumption that they have a direct impact on the final closing price of Bitcoin.

By conducting thorough preprocessing and data processing, the model will be built upon clean, representative data that is ready for training and evaluation. This stage also helps to minimize prediction errors caused by bad data quality.

2.3. Predictive Model Architecture

This study employs two machine learning models, namely Linear Regression and Random Forest Regressor, to predict the Bitcoin closing price based on historical data. The selection of these two models is based on their different approaches to handling data: Linear Regression is suitable for linear relationships, while Random Forest is designed to handle complex non-linear relationships.

2.3.1. *Regresi Linier*

Linear Regression is one of the most commonly used statistical methods in predictive analysis. This model works by finding the best fit line that represents the relationship between independent and dependent variables. In this context, the independent variables consist of features such as Open, High, Low, and Volume, while the dependent variable is the Close (closing price)[9].

Mathematically, simple linear regression can be expressed with the following formula:

$$\hat{y} = \beta_0 + \beta_{1x1} + \beta_{2x2} + \cdots + \beta_{nxn}$$

Explanation:

- $\hat{y}$= predicted result
- β_0 = intercept (constant)
- β_1, β_2,…, β_n= regression coefficients for each feature
- x_1, x_2,…, x_n= independent variables (input features)

Example of Linear Regression implementation in Python:

Program Jurnal

```
from sklearn.linear_model import
LinearRegression

  model_lr = LinearRegression()
  model_lr.fit(X_train, y_train)
y_pred_lr = model_lr.predict(X_test)
```

This model is used as a baseline for comparison with more complex models.

2.3.2 *Random Forest*

Random Forest is an ensemble learning algorithm composed of a collection of decision trees. Each tree in the Random Forest is trained on a different subset of the data, and the final prediction result is obtained by averaging the predictions of all trees. This approach improves the model's robustness against overfitting and enhances its generalization ability[10].

Random Forest is capable of handling non-linear relationships that cannot be captured by Linear Regression. It also offers flexibility in setting key parameters, such as[11]:

- n_estimators: he number of trees in the forest,
- max_depth: the maximum depth of each tree,
- min_samples_split: the minimum number of samples required to split an internal node,
- min_samples_leaf: the minimum number of samples required to be at a leaf node.

The final prediction in Random Forest is obtained by averaging the outputs of all decision trees, as follows:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

Explanation:

- T = the number of trees in the forest
- $h_t\ (x)$ = the prediction result of the t-th tree for input x
- $\hat{y}$ = the average of all tree predictions

In this study, optimal parameter selection was carried out using GridSearchCV, which tests various combinations of parameters and selects the best one based on evaluation scores.

Example of Random Forest implementation with GridSearchCV:

Program Jurnal

```
from sklearn.ensemble import
RandomForestRegressor
from sklearn.model_selection import
GridSearchCV

  param_grid = {
      'n_estimators': [50, 100],
      'max_depth': [None, 10, 20],
      'min_samples_split': [2, 5],
      'min_samples_leaf': [1, 2]
  }

  rf = RandomForestRegressor()
  grid_search = GridSearchCV(estimator=rf,
param_grid=param_grid, cv=5)
  grid_search.fit(X_train, y_train)
y_pred_rf = grid_search.predict(X_test)
```

The results of this GridSearch are then used to evaluate the model against the test data and also in the K-Fold scheme.

2.4. *Model Validation and Evaluation*

After the model is built, the next important step is to perform validation and evaluation to measure the model's performance in predicting previously unseen data. Validation aims to assess how well the model can generalize to new data, while evaluation is used to determine the quality of predictions using specific metrics.

2.4.1 *Validation Schemes*

This study uses two validation approaches:

- 70:30 Data Split

The dataset is split into 70% for training and 30% for testing. This approach is commonly used in initial experiments due to its simplicity and speed.

- K-Fold Cross Validation (K-Fold CV)

In this method, the data is divided into k equal-sized folds. In each iteration, one fold is used as the test set, while the remaining folds are used for training. This process is repeated k times, and the average of all evaluation results is used as the model's final score. In this study, K = 10 (10-fold CV) is applied.

2.4.2 *Evaluation Metrics*

This study uses three main evaluation metrics commonly applied in regression tasks:

- Mean Absolute Error (MAE)

MAE measures the average of the absolute differences between the predicted values and the actual values. The smaller the MAE value, the better the prediction

$$\text{MAE } = \frac{1}{n}\sum_{i=1}^{n}|y_{i-}\ \hat{y}_i|$$

Explanation:

- y_i= actual value of the i-th value
- $\hat{y}_i$ = predicted value of the i-th value
- n = total number of data

MAE shows the average of the absolute differences between actual and predicted values[12].

- Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i + \hat{y}_i)^2$$

MSE measures the average of the squared differences between the actual and predicted values. This metric penalizes larger errors more heavily due to the squaring of the differences[13].

- Root Mean Squared Error (RMSE)

$$\text{RMSE } \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i + \hat{y}_i)^2}$$

RMSE is the square root of the MSE. This metric presents the error in the same unit as the original data, providing a more interpretable measure of prediction accuracy[12].

These evaluations are applied to both models: Linear Regression and Random Forest**,** using both the 70:30 data split and the 10-fold cross-validation schemes. The results of the evaluations will be compared to determine which model provides the best prediction performance.

2.5 Baseline Model (Naive Forecast)

The baseline model is used as an initial benchmark in evaluating prediction performance. This model forecasts the closing price of Bitcoin for the next day based on the previous day's closing price (persistence model). This approach does not involve model training but assumes that tomorrow's price will be the same as today's price [14].

The prediction is performed by shifting the historical data one day forward. Evaluation metrics such as MAE, MSE, and RMSE are used to measure how far the baseline prediction deviates from the actual values.

3. Results and Discussion

This study aims to find the best performance by comparing two regression algorithms Linear Regression and Random Forest in predicting Bitcoin prices based on historical data. Evaluation was conducted using two validation schemes: 70:30 data split and 10-Fold Cross Validation. The evaluation metrics used include Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

3.1. *Linear Regression Model Evaluation Results*

After training and testing the Linear Regression model, the evaluation results are as follows:

Table 1. Table Linear Regression Model Evaluation

Skema Validasi	MAE	MSE	RMSE
Split 70:30	739.82	1,759,573.89	1,326.49
K-Fold Cross Validation (10 Fold)	758.31	1,733,762.22	1,314.47

From the table above, it can be seen that the Linear Regression model yields relatively low error values, both in the 70:30 split validation and 10-fold cross-validation. Although this model is simple and assumes linearity, it performs quite well in capturing Bitcoin price trends.

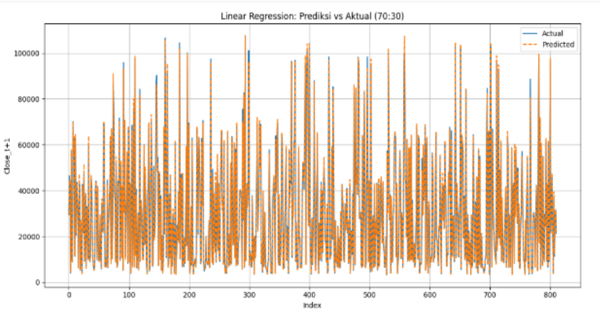

Figure 3. Visualization of Bitcoin price prediction results using a Linear Regression model

The graph below illustrates the Bitcoin price prediction using the Linear Regression model with the 70:30 data split scheme. The blue line represents the actual values, while the orange dashed line shows the predicted values. The graph shows that the model is able to follow the general trend of the actual data, although deviations occur during periods of high fluctuation.

3.2. *Random Forest Model Evaluation Results*

The following are the evaluation results of the Random Forest model, both before and after optimization using GridSearchCV:

Table 2. Random Forest Model Evaluation

Model	Validation Shcemes	MAE	MSE	RMSE
Random Forest (default)	Split 70:30	869.26	2,144,269.89	1,464.33
Random Forest (default)	K-Fold Cross Validation (10 Fold)	870.16	2,146,352.29	1,461.25
Random Forest + GridSearchCV	Split 70:30	847.22	2,077,061.91	1,441.20
Random Forest + GridSearchCV	K-Fold Cross Validation (10 Fold)	857.34	2,102,343.98	1,446.35

Based on the results, it can be observed that parameter optimization using GridSearchCV improves the performance of the Random Forest model. However, even though Random Forest is a more complex non-linear model, the resulting error values are still higher compared to those of the Linear Regression model.

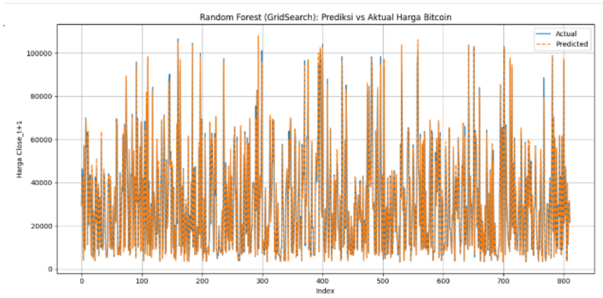

Figure 4. Visualization of Bitcoin price prediction results using a Random Forest model

The Bitcoin price prediction results using the Random Forest model optimized with GridSearchCV show that the blue line represents the actual values, while the orange dashed line indicates the predicted values. Overall, the Random Forest model is able to follow the general trend of price movements; however, it exhibits greater variability compared to the Linear Regression model, resulting in less stable predictions. To better understand the contribution of each feature to the prediction outcomes, Figure 5 presents a visualization of the feature importance scores in the Random Forest model.

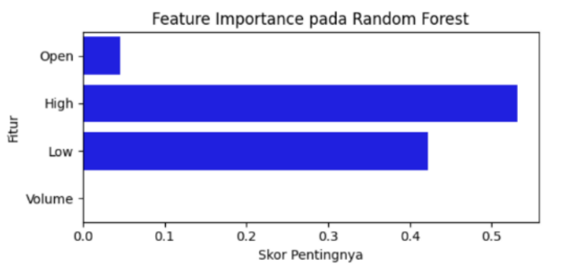

Figure 5. Visualization of feature importance scores in the Random Forest model

The feature importance visualization from the Random Forest model indicates that the "High" feature contributes the most to the prediction of the next day's Bitcoin closing price, followed by "Low" and then "Open," while "Volume" shows an insignificant contribution. This suggests that the daily high and low prices have a stronger correlation with the next closing price compared to trading volume.

3.3 *Compared to Baseline Model*

Table 3. Comparison Table with Baseline Model

Model	MAE	MSE	RMSE
Naive (Baseline)	708.55	1531097.59	1237.38

The baseline model yielded an MAE of 708.55, MSE of 1,531,097.59, and RMSE of 1,237.38. These results serve as a benchmark for evaluating the performance of other predictive models. Both Linear Regression and Random Forest achieved lower error metrics than the baseline, indicating that these algorithms are more effective in capturing the patterns of Bitcoin price movements compared to the simple prediction method.

3.4 *Statistical Evaluation of Model Performance*

To determine whether the performance differences between Linear Regression and Random Forest are statistically significant, the Wilcoxon Signed-Rank test was conducted using MAE, MSE, and RMSE values from each fold of the 10-Fold Cross Validation scheme. This test was chosen due to its non-parametric nature and suitability for comparing two models on paired data.

The Wilcoxon test results yielded a p-value of 1.0000 across all three evaluation metrics, indicating that there is no statistically significant difference between the two models. Although Linear Regression demonstrated better average performance than Random Forest, the statistical test suggests that the observed advantage is not strong enough to conclude a consistent difference across all folds.

These findings highlight the importance of not solely relying on average performance metrics, but also considering model stability and consistency through appropriate statistical evaluation.

The table below presents the statistical performance evaluation data for each fold:

Table 4. Linear Regression Model Evaluation Results (10-Fold Cross Validation)

Fold	MAE	MSE	RMSE
1	808.35	1,904,578.43	1,380.06
2	664.03	1,375,331.73	1,172.75
3	741.76	1,974,544.39	1,405.18
4	726.66	1,692,153.54	1,300.83
5	792.39	1,678,398.93	1,295.53
6	752.37	1,403,541.68	1,184.71
7	743.01	1,755,269.64	1,324.87
8	733.45	1,737,785.72	1,318.25
9	815.31	1,882,068.11	1,371.88
10	805.77	1,933,949.98	1,390.67
Rata-rata	758.31	1,733,762.22	1,314.47

Table 5. Random Forest Model Evaluation Results (10-Fold Cross Validation, GridSearch)

Fold	MAE	MSE	RMSE
1	958.08	2,668,713.98	1,633.62
2	775.77	1,756,690.72	1,325.40
3	810.10	2,406,437.06	1,551.27
4	809.91	1,910,410.38	1,382.18
5	926.71	2,233,473.88	1,494.48
6	831.69	1,702,443.91	1,304.78
7	861.49	2,270,689.25	1,506.88
8	804.08	1,789,017.32	1,337.54
9	900.79	2,082,978.27	1,443.25
10	894.76	2,202,585.05	1,484.11
Rata-rata	857.34	2,102,343.98	1,446.35

3.5 Discussion

Based on the evaluation results, the Linear Regression model outperformed Random Forest in both the 70:30 data split and K-Fold cross-validation schemes. Linear Regression produced lower MAE and RMSE values, indicating that this simple model is effective at capturing short-term trends in Bitcoin price movements. Its performance even surpassed that of the baseline model (naive prediction), which assumes that the closing price today will be the same as tomorrow's. This reinforces the notion that linear models, while simple, can still be accurate in the context of daily Bitcoin price data.

Conversely, the Random Forest model, despite being optimized using GridSearchCV, resulted in higher error values than both Linear Regression and the baseline. Although Random Forest is a complex ensemble model capable of capturing non-linear relationships, it appeared less optimal in this study. Several factors may have contributed to this outcome:

- Daily Bitcoin price patterns tend to exhibit linear behavior in the short term, making linear models more appropriate for such data.
- The input features used (Open, High, Low, Volume) may not have been complex enough to fully leverage the power of Random Forest.
- Overfitting may have occurred in the Random Forest model due to the lack of diverse and representative features.

To determine whether the performance difference between Linear Regression and Random Forest was statistically significant, the Wilcoxon Signed-Rank test was conducted using MAE, MSE, and RMSE values from each fold of the cross-validation. The test produced a p-value of 1.0000 across all three metrics, indicating no statistically significant difference between the two models. Although Linear Regression showed better average performance, the statistical test suggests that the difference was not strong enough to be considered consistent across all folds.

As an alternative for evaluating linear models, regularization approaches such as Ridge and Lasso Regression are worth considering, as they are effective in addressing overfitting and multicollinearity. While not applied in this study, both models are recommended for future research to assess prediction stability in the volatile Bitcoin market.

Given the sensitivity of cryptocurrency data to outliers and non-stationarity, preprocessing techniques such as log transformation, differencing, or filtering can enhance model robustness. Evaluations during high-volatility periods—such as the Bitcoin price spike in 2021—also emphasized the need for more adaptive models: Linear Regression exhibited increased error, while Random Forest tended to overfit.

In practical applications, the models developed in this study have potential to be integrated into real-time predictive systems via reusable machine learning pipelines that can be adapted to other cryptocurrencies. In the future, hybrid approaches that combine the trend-capturing strength of Linear Regression with non-linear models such as Random Forest or XGBoost are recommended to more effectively handle both long-term trends and short-term fluctuations in dynamic digital asset markets.

4. Conclusion

This study compares the accuracy of Linear Regression and Random Forest models in predicting Bitcoin's daily closing price using historical data. Based on the evaluation using MAE, MSE, and RMSE, Linear Regression yielded lower error rates than Random Forest across both the 70:30 data split and K-Fold cross-validation schemes. Furthermore, Linear Regression also outperformed the naive baseline model, which simply uses today's price to predict tomorrow's.

However, the Wilcoxon Signed-Rank test applied to the K-Fold results indicated no statistically significant difference between the two models, with a p-value of 1.0000. This suggests that although Linear Regression showed superior average performance, the advantage was not consistent enough across all folds to be considered statistically significant. The Random Forest model may require more complex or diverse input features to achieve optimal performance and avoid overfitting.

This research recommends further exploration of alternative linear models such as Ridge and Lasso Regression, as well as hybrid approaches that combine the strengths of both linear and non-linear models. Additionally, time-aware validation methods like Time Series Cross-Validation are important to avoid data leakage and improve evaluation realism in time-series prediction tasks. These findings are expected to support the development of more accurate and adaptive predictive systems in the context of dynamic cryptocurrency markets..

Acknowledgments

The author expresses his deepest respect and gratitude to all those who have provided support, assistance, and prayers during the preparation and completion of this research. Special thanks go to:

1. Mr. Nur Kholis, the author's father, who has provided sincere and continuous moral and material support.

2. Mrs. Siti Rismiati, the author's mother, who has always prayed for him, given him invaluable love, and provided invaluable support.

3. All members of the author's family, for their motivation, attention, and facilities to ensure the smooth running of this research.

4. The author's friends and colleagues, who have provided assistance, input, and encouragement throughout the research process.

5. Ms. Lutfiyatul Malikhah, who has provided invaluable moral support, attention, and encouragement, and has been a source of inspiration in completing this research.

6. The Supervisors, who have patiently guided, provided direction, constructive input, and motivation to the author throughout the process of preparing this research.

The author realizes that without the assistance, support, and prayers of all the parties mentioned, this research would not have been completed successfully.

References

[1] M. R. Maulana, "Bitcoin dan Konsep Uang Digital: Tinjauan Historis dan Teoritis," *Waralaba J. Econ. Bus.*, vol. 1, no. 2, Dec. 2024, doi: 10.61590/waralaba.v1i2.144.

[2] Syarif Aminul Khoiri and Abdul Wahid, "Analisis Kinerja Algoritma Machine Learning dalam Prediksi Harga Cryptocurrency," *JUSTINDO (Jurnal Sist. dan Teknol. Inf. Indones.*, vol. 9, no. 2, pp. 133–141, Aug. 2024, doi: 10.32528/justindo.v9i2.1965.

[3] S. Wijaya and Fauziah, "A COMPERATIVE ANALYSIS OF LINEAR REGRESSION AND RANDOM FOREST METHODS FOR PREDICTING PALM OIL PRICE SALES," *J. Ilm. Teknol. dan Rekayasa*, vol. 28, no. 3, pp. 265–274, Dec. 2023, doi: 10.35760/tr.2023.v28i3.10111.

[4] W. Wijiyanto, A. I. Pradana, S. Sopingi, and V. Atina, "Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa," *J. Algoritm.*, vol. 21, no. 1, pp. 239–248, 2024, doi: 10.33364/algoritma/v.21-1.1618.

[5] I. Amansyah, J. Indra, E. Nurlaelasari, and A. R. Juwita, "Prediksi Penjualan Kendaraan Menggunakan Regresi Linear:Studi Kasus pada Industri Otomotif di Indonesia," *J. Soc. Sci. Res.*, vol. 4, pp. 1199–1216, 2024, [Online]. Available: https://j-innovative.org/index.php/Innovative

[6] S. Fachid and A. Triayudi, "Perbandingan Algoritma Regresi Linier dan Regresi Random Forest Dalam Memprediksi Kasus Positif Covid-19," *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 1, p. 68, Jan. 2022, doi: 10.30865/mib.v6i1.3492.

[7] M. Istaltofa, S. Sarwido, and A. Sucipto, "Comparison of Linear Regression and LSTM (Long Short-Term Memory) in Cryptocurrency Prediction," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 4, no. 2, pp. 141–148, Aug. 2024, doi: 10.20895/dinda.v4i2.1575.

[8] L. Santoso and Priyadi, "Mengoptimalkan Proses Pembersihan Data dalam Analisis Big Data Menggunakan Pipeline Berbasis AI," *J. Elektron. Dan Komput.*, vol. 17, no. 2, pp. 657–666, 2024.

[9] I. D. Hartarti, I. A. Septiyani, D. A. Gultom, Y. Hendrian, and S. L. Kinanti, "Prediksi Harga Rumah di Boston Dengan Model Regresi Linear Menggunakan Python," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 2, pp. 4250–4256, Jun. 2025, doi: 10.31004/riggs.v4i2.1210.

[10] M. A. Lubis and S. Samsudin, "Using the Random Forest Method in Predicting Stock Price Movements," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 5, no. 1, pp. 28–35, Feb. 2025, doi: 10.20895/dinda.v5i1.1765.

[11] A. H. Mubarok, P. Pujiono, D. Setiawan, D. F. Wicaksono, and E. Rimawati, "Parameter Testing on Random Forest Algorithm for Stunting Prediction," *sinkron*, vol. 9, no. 1, pp. 107–116, Jan. 2025, doi: 10.33395/sinkron.v9i1.14264.

[12] Nurdin, N. Suarna, and W. Prihartono, "ALGORITMA REGRESI LINIER SEDERHANA UNTUK PREDIKSI PENGGUNAAN VOLUME AIR BERDASARKAN JENIS PELANGGAN PDAM," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 4, no. 1, pp. 43–52, Jan. 2025, doi: 10.69916/jkbti.v4i1.187.

[13] D. Arifuddin, K. Kusrini, and K. Kusnawi, "Perbandingan Performansi Algoritma Multiple Linear Regression dan Multi Layer Perceptron Neural Network dalam Memprediksi Penjualan Obat," *MALCOM Indones. J. Mach. Learn.*

Comput. Sci., vol. 5, no. 2, pp. 722–737, Apr. 2025, doi: 10.57152/malcom.v5i2.1952.

[14] M. Frohmann, M. Karner, S. Khudoyan, R. Wagner, and M. Schedl, "Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting," *Big Data Cogn. Comput.*, vol. 7, no. 3, 2023, doi: 10.3390/bdcc7030137.