IJAIS International Journal of Artificial Intelligence and Science

Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

BBCA Stock Price Prediction Using Linear Regression Method

Shannon Dominique Saputra[1], **Albertus Dwiyoga Widiantoro**[2]

1[2]Faculty of Information System, Soegijapranata Catholic University, Semarang, Indonesia
Email: 22g40007@student.unika.ac.id[1], yoga@unika.ac.id[2]

Received:
August 20, 2024
Revised:
August 31, 2024
Accepted:
September 28, 2024
Published:
September 29, 2024

Corresponding Author:
Author Name*:
Shannon Dominique Saputra
Email*:
dominiqueshannon24@gmail.com

DOI: 10.63158/IJAIS.v1.i1.7

Abstract. This study focuses on predicting the stock price of Bank Central Asia (BBCA) using linear regression techniques, a widely utilized statistical method in financial forecasting. Stock price prediction is critical for investors, particularly in volatile markets like Indonesia. This research analyzes the relationship between key variables, such as adjusted closing prices and trading volume, based on historical data. The methodology includes data collection, preprocessing, model construction, and evaluation using metrics like Root Mean Square Error (RMSE) to assess the model's accuracy. The results indicate that linear regression can effectively predict BBCA stock prices with reasonable accuracy, providing a practical and interpretable tool for investors. These findings contribute to financial forecasting by demonstrating the utility of linear regression in stock price prediction, particularly in emerging markets.

Keywords: linear regression, stock price prediction, bbca, financial forecasting, volatile markets

1. INTRODUCTION

Predicting stock prices is a critical yet challenging task for investors and analysts, particularly in volatile markets like Indonesia's. Stock prices can fluctuate rapidly due to various unpredictable factors such as macroeconomic events, market sentiment, and trading volume, making it difficult for investors to accurately forecast price movements and make informed decisions. Bank Central Asia (BBCA), one of the most prominent and stable-performing stocks in Indonesia, is frequently targeted by investors. However, despite its solid reputation, accurately predicting BBCA's stock price remains a significant challenge because of the market's inherent complexity and volatility [1], [2]. This study seeks to address these prediction challenges by applying predictive modeling techniques to forecast BBCA stock prices.

A significant gap in the existing literature and practice involves the trade-off between model complexity and accessibility. Advanced predictive models, such as machine learning algorithms, often provide high levels of accuracy but come with substantial computational requirements and the need for specialized expertise. These methods are typically inaccessible to many investors or analysts with limited technical backgrounds [3]. On the other hand, traditional techniques like technical and fundamental analysis are easier to apply but frequently lack the predictive power required in volatile market conditions [4]. Additionally, while linear regression has been used extensively in financial forecasting, its specific utility in predicting stock prices in volatile, emerging markets like Indonesia has not been thoroughly investigated. Addressing this gap, there is a need for an accessible yet effective model that combines simplicity with the ability to handle stock market volatility.

The aim of this study is to bridge this gap by employing linear regression, a statistical method known for its simplicity and interpretability, to predict BBCA stock prices in the volatile Indonesian market. Linear regression is chosen for its straightforwardness and ability to reveal the relationships between critical variables—such as adjusted closing prices and trading volume—and stock price movements. While more advanced methods exist, linear regression remains a practical and efficient choice, particularly in situations where computational resources are limited, and market conditions are highly dynamic [5], [6]. Given these considerations, the study will assess how well linear regression can manage the complexities of predicting stock prices in emerging markets, such as Indonesia.

To further enhance the accessibility of predictive stock models, this research utilizes RapidMiner**,** a widely used data science platform that allows users to implement and test predictive models without requiring extensive programming skills. RapidMiner offers an intuitive interface and robust tools for statistical analysis, making it an ideal platform for implementing linear regression models [7]. By combining the simplicity of linear regression with the ease of use provided by RapidMiner, this study offers a practical solution for both novice and experienced investors. This approach addresses the need for an effective, user-friendly tool for predicting stock prices, especially in highly volatile markets like Indonesia [7], [8].

The primary objectives of this research are twofold: first, to evaluate the effectiveness of linear regression in predicting BBCA stock prices using historical data, and second, to demonstrate the usability and advantages of RapidMiner as a tool for implementing such models. By fulfilling these objectives, this study will contribute to the growing body of literature on financial forecasting by showing that simple, accessible methods can still provide reliable predictive insights. It will also offer practical solutions for investors and analysts seeking to navigate the challenges of stock price prediction in volatile markets. The findings are expected to demonstrate that variables like adjusted closing prices and trading volume are significant predictors of BBCA stock prices, affirming the relevance and practicality of this approach [6], [9].

2. METHODS

The flow in this study is structured to systematically apply a predictive model for Bank Central Asia (BBCA) stock prices using linear regression within the RapidMiner platform. This structured process ensures that each phase builds upon the previous one to yield accurate and interpretable results. Figure 1 is integrating all stages, from data collection and preprocessing to model building, evaluation, and optimization, ensuring a clear, repeatable methodology.

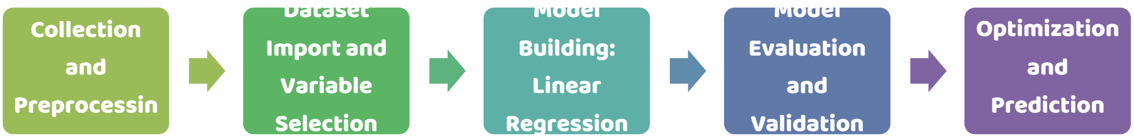

Collection and Preprocessin
Dataset Import and Variable Selection
Model Building: Linear Regression
Model Evaluation and Validation
Optimization and Prediction

Figure 1. Research Flow

2.1 Data Collection and Preprocessing

The first step in the research process involves gathering historical stock price data for BBCA, including daily values such as the opening price, highest price, lowest price, adjusted closing price, and trading volume. The dataset is sourced from the Indonesia Stock Exchange (IDX) and reliable financial platforms [10]. [11]. The collected data is stored in CSV format to facilitate easy import into RapidMiner. Preprocessing steps include addressing missing values, handling outliers, and ensuring consistency to prepare the data for accurate model training and testing [12].

2.2 Dataset Import and Variable Selection

After data collection and cleaning, the dataset is imported into RapidMiner. The key variables selected for the analysis include adjusted closing price as the dependent variable and trading volume as the independent variable. These variables are chosen for their relevance to predicting BBCA stock price behavior, while other variables like the opening, highest, and lowest prices provide additional context but are excluded from the regression model to maintain focus [13].

2.3 Model Building: Linear Regression

Linear regression is applied to model the relationship between the adjusted closing price (dependent variable) and trading volume (independent variable). The linear regression model uses the following Equation 1. The model is built using RapidMiner's linear regression operator, which aims to minimize the squared error between the predicted and actual stock prices. This model allows for stock price prediction based on trading volume and provides interpretable results for investors.

$$Y=\beta 0+\beta 1X+\epsilon \quad (1)$$

Where:

Y is the adjusted closing price, X is the trading volume, $\beta 0$ is the intercept, and $\beta 1$ is the regression coefficient representing the effect of changes in trading volume on stock price [14].

2.4 Model Evaluation and Validation

To assess the model's performance, the dataset is split into training and testing subsets using the Split Data operator in RapidMiner. The model is trained on a portion of the data (training set) and tested on unseen data (testing set). Key performance metrics such as R-square d and mean squared error (MSE) are used to evaluate the model's predictive power. These metrics offer insights into how well the model captures the relationships in the data and ensure it generalizes well to new, unseen data [15].

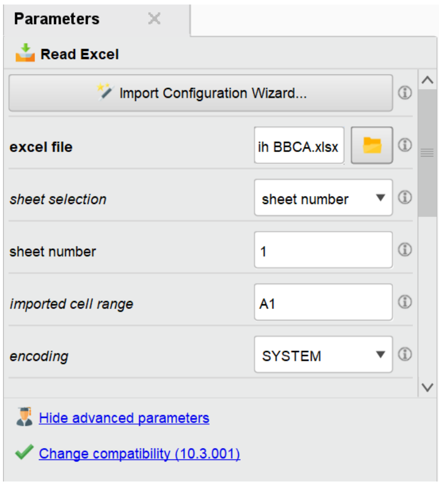
2.5 Optimization and Prediction

Once the model is validated, the Apply Model operator in RapidMiner is used to apply the trained model to new data. This allows for generating stock price predictions based on updated trading volumes. The Performance operator is employed to monitor prediction accuracy and optimize the model where necessary, ensuring it remains effective for real-world applications [10], [11].

3. RESULT AND DISCUSSION

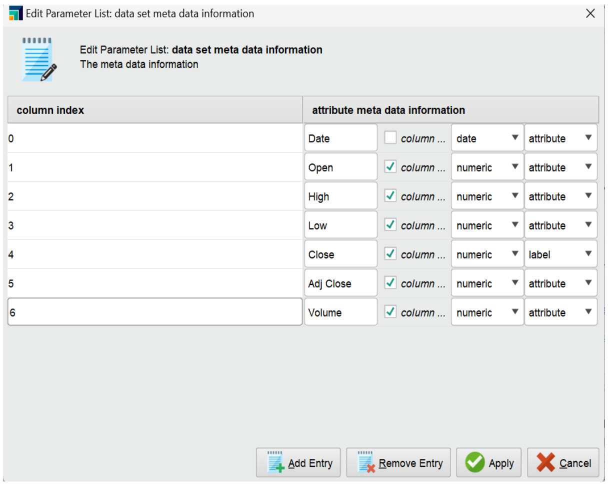
3.1 Experimental Performance

The experimental process began with converting the original dataset from CSV format into an Excel file for easier manipulation. Once imported, several key preprocessing steps were carried out to prepare the data for modeling. Specifically, the "Date" column, which did not offer any predictive value, was excluded. The "Close" column was redefined as the label, making it the target variable for predictive analysis. These changes ensured that the data was properly formatted for the linear regression model. Figure 2 illustrates the initial import of the dataset, while Figure 3 demonstrates the modification of the relevant attributes.


Parameters
Read Excel
Import Configuration Wizard...
Hide advanced parameters
Change compatibility (10.3.001)

excel file	ih BBCA.xlsx
sheet selection	sheet number
sheet number	1
imported cell range	A1
encoding	SYSTEM

Figure 2. Import file

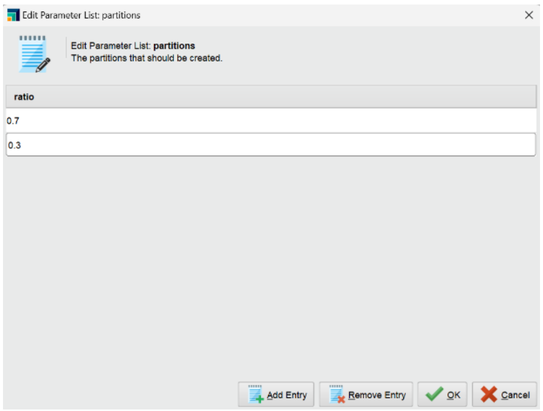

Edit Parameter List: data set meta data information
Edit Parameter List: data set meta data information
The meta data information
Add Entry
Remove Entry
Apply
Cancel

column index	attribute meta data information			
0	Date	□ column ...	date	attribute
1	Open	☑ column ...	numeric	attribute
2	High	☑ column ...	numeric	attribute
3	Low	☑ column ...	numeric	attribute
4	Close	☑ column ...	numeric	label
5	Adj Close	☑ column ...	numeric	attribute
6	Volume	☑ column ...	numeric	attribute

Figure 3. Modifying attributes

Next, the dataset was split into training and testing subsets using the Split Data operator, a crucial step for building and validating the model. This operator divided the data into three different configurations: 90%-10%, 80%-20%, and 70%-30%, respectively. The purpose of these variations was to evaluate how different proportions of training data affected the model's performance. The training and testing data for these splits are depicted in Figure 4, providing a visual representation of how the data was partitioned.

Once the data was split, a linear regression model was applied to each configuration. For the 90% training and 10% testing split, the model used the lowest price (Low) and volume as predictor variables. The coefficients for these variables were 1.020 for Low and 0.007 for Volume, with a constant of -2,106,148.974. The resulting Root Mean Square Error (RMSE) for this model was 8,805,774.938 ± 0.000**,** as shown in Table 2**.** This relatively low RMSE suggests that the model performed well in predicting values, with a small average error in the predictions.



Edit Parameter List: partitions

Edit Parameter List: partitions
The partitions that should be created.

ratio
0.7
0.3

Figure 4. Training Data and Testing Data

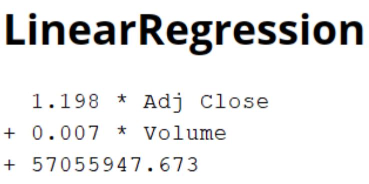
For the second configuration, with 80% training data and 20% testing data, the model maintained the same predictor variables: lowest price (Low) and volume. However, the coefficients shifted slightly to 1.020 and 0.006, and the constant changed to -2,375,493.414. This model yielded an RMSE of 9,217,478.177 ± 0.000**,** which is higher than that of the 90%-10% model, indicating a slight decline in accuracy. Nevertheless, the model still performed reasonably well, demonstrating that it could still predict the target variable with a relatively small margin of error, despite the reduced training data.

In the final configuration, where 70% of the data was used for training and 30% for testing, the model changed its predictor variables to the adjusted closing price (Adj Close) and volume. The coefficients for these variables were 1.198 and 0.007, respectively, and the constant was much larger at 57,055,947.673. The RMSE for this model was 11,089,148.439 ± 0.000**,** making it the least accurate of the three configurations. The larger RMSE value signifies a higher average prediction error, suggesting that reducing the amount of training data negatively impacted the model's performance. The detailed RMSE values for all three models are presented in Table 2**.**

Table 2. RSME Value

Training	Testing	Value	RMSE Value
90%	10%	1.020 * Low + 0.007 * Volume - 2,106,148.974	8,805,774.938 ± 0.000
80%	20%	1.020 * Low + 0.006 * Volume - 2,375,493.414	9,217,478.177 ± 0.000
70%	30%	1.198 * Adj Close + 0.007 * Volume + 57,055,947.673	11,089,148.439 ± 0.000

RMSE is a widely accepted metric in regression analysis as it provides a clear measure of the model's accuracy by quantifying the average magnitude of error between predicted and actual values. A lower RMSE indicates better predictive performance. As the results in Table 2 show, the model trained with 90% of the data achieved the best performance with the lowest RMSE of 8,805,774.938, while the model with 70% training data had the highest RMSE of 11,089,148.439**,** demonstrating poorer predictive ability. The model with 80% training data had an intermediate RMSE of 9,217,478.177, indicating that while it was less accurate than the 90% model, it was still more reliable than the 70% model.



```
LinearRegression

  1.198 * Adj Close
+ 0.007 * Volume
+ 57055947.673
```

Figure 5. Linear Regression

The experimental results illustrate the importance of having a sufficiently large training dataset. The model with 90% training data consistently outperformed the other configurations, producing the smallest prediction errors. This highlights the trade-off between the size of the training set and the model's accuracy, where more training data generally leads to better model performance. Figure 5 presents the linear regression outputs for these experiments. Moving forward, further refinement techniques such as cross-validation or parameter tuning could be employed to optimize model performance across various data splits.

3.2 Discussion

The results of the experimental performance, as presented in the previous section, provide valuable insights into the relationship between training data size and the predictive accuracy of linear regression models. The variation in RMSE values across the three different training-testing configurations—90%-10%, 80%-20%, and 70%-30%—highlights the impact of training data on model performance and the trade-offs between accuracy and the proportion of data reserved for testing.

The model trained on 90% of the data produced the best performance with the lowest RMSE value of 8,805,774.938. This suggests that when a larger portion of the dataset is allocated to training, the model has more data to learn from, allowing it to capture the underlying relationships between the predictor variables (Low price and Volume) and the target variable (Close price) more effectively. As a result, the predictions made by the model are more accurate, with smaller deviations from the actual values. The relatively low RMSE in this case indicates that the errors between predicted and actual values are minimal, which is a key characteristic of a well-performing model. This model's ability to minimize prediction errors demonstrates that it generalized well to the unseen test data, reinforcing the importance of using a sufficient amount of data for training in order to enhance model performance.

In contrast, the model trained on 80% of the data, while still reasonably accurate, exhibited a higher RMSE of 9,217,478.177**.** The increase in RMSE, although not drastic, indicates that the reduction in training data had a noticeable effect on the model's predictive ability. The decrease in training data likely limited the model's capacity to fully capture the patterns within the data, leading to slightly larger prediction errors. Nonetheless, the performance of this model remains acceptable, suggesting that using 80% of the dataset for training still provides the model with enough information to make relatively accurate predictions. However, this result also implies that there is a tipping point in the amount of training data required for the model to maintain optimal performance, and reducing the training set too much can negatively impact predictive accuracy.

The model trained on 70% of the data, with an RMSE of 11,089,148.439**,** had the poorest performance among the three configurations. The significantly higher RMSE indicates that this model struggled to generalize from the training data to the test data. The change in predictor variables, from the lowest price (Low) to the adjusted closing price (Adj Close), and the large shift in the constant value suggest that this model may have overfitted the smaller training dataset, leading to less reliable predictions when applied to the test data. Overfitting occurs when the model becomes too closely aligned with the training data, capturing noise or specific patterns that do not generalize well to new data. This, coupled with the reduced training size, likely contributed to the larger prediction errors observed in this model. The result demonstrates that the amount of training data used is crucial and reducing it below a certain threshold can severely degrade the model's performance.

The trend observed across the three models highlights the strong relationship between training data size and prediction accuracy. As expected, a larger training dataset enables the model to learn more effectively, thereby reducing the RMSE. This finding aligns with established machine learning principles, where models trained on larger datasets tend to perform better as they are exposed to more patterns and variability within the data. However, the experiment also raises important questions about the point of diminishing returns. While the 90%-10% split resulted in the best model, further experimentation could explore whether increasing the training data beyond 90% would continue to improve performance, or whether a plateau would be reached. Additionally, there may be practical considerations, such as computational resources or model training time, that need to be balanced against the marginal improvements in accuracy gained from using larger training sets.

These results underscore the importance of using an adequate amount of training data to achieve optimal model performance. The 90%-10% split provided the best balance between training and testing, yielding the lowest RMSE and the most accurate predictions. The performance degradation in the 80%-20% and 70%-30% configurations further emphasize the need for careful consideration of training data size when building predictive models. These findings also suggest that more sophisticated modeling techniques, such as cross-validation or regularization methods, could be employed in future work to improve model robustness and prevent overfitting, especially when working with smaller training datasets.

4. CONCLUSION

This research aimed to develop and assess a predictive model for BBCA stock prices, with RMSE as the key performance metric. The model produced an RMSE of 8,805,774.938, which, though lower than other models (9,217,478.177 and 11,089,148.439), still indicates a significant prediction error given the typical range of BBCA stock prices. While relatively better, this RMSE is not sufficient in absolute terms for reliable stock price prediction. To fully gauge the model's effectiveness, it should be compared against industry standards, including historical price trends, market indices, and company fundamentals like revenue growth and economic conditions. Further refinements, such as advanced modeling techniques and broader evaluation metrics, will be necessary to improve accuracy and ensure the model's applicability for real-world financial forecasting.

REFERENCES

[1] M. Q. Shobri et al., "Model Analisis Harga Saham Sektor Finansial PT. Bank Central Asia Tbk. (BBCA)," J. Keilmuan dan Keislam., pp. 260–269, 2023, doi: 10.23917/jkk.v2i4.170.

[2] R. A. Fahrezi, M. Y. Wijaya, and N. Fitriyati, "Prediksi Harga Penutupan Saham Bank Central Asia: Implementasi Algoritma Long Short-Term Memory Dan Perbandingannya Dengan Support Vector Machine," J. Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat., vol. 5, no. 1, pp. 452–464, 2024, doi: 10.46306/lb.v5i1.582.

[3] A. Fitri Boy, "Implementasi Data Mining Dalam Memprediksi Harga Crude Palm Oil (CPO) Pasar Domestik Menggunakan Algoritma Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi Sumatera Utara)," J. Sci. Soc. Res., vol. 4307, no. 2, pp. 78–85, 2020.

[4] S. A. L. Satriyo, A. Rizky Pratama, and Rahmat, “Perbandingan metode linear regresi dan polynomial regresi untuk memprediksi harga saham studi kasus Bank BCA,” INFOTECH J. Inform. Teknol., vol. 4, no. 1, pp. 59–70, 2023, doi: 10.37373/infotech.v4i1.602.

[5] I. M. Sinaga, A. Lubis, and A. Prayudi, “Pengaruh Internet Financial Reporting (Ifr) Dan Tingkat Pengungkapan Informasi Website Terhadap Frekuensi Perdagangan Saham Pada Perusahaan Pertambangan Yang Terdaftar Di Bei,” J. Ilm. Manaj. dan Bisnis, vol. 1, no. 2, pp. 106–111, 2020, doi: 10.31289/jimbi.v1i2.394.

[6] P. Wi and D. Anggraeni, “Faktor-Faktor Yang Mempengaruhi Minat Karyawan Perusahaan Untuk Berinvestasi Di Pasar Modal Pada Masa Pandemi Covid 19,” J. Revenue J. Ilm. Akunt., vol. 1, no. 1, pp. 81–89, 2020, doi: 10.46306/rev.v1i1.15.

[7] A. Riyandi, A. Aripin, I. N. Ardiansyah, R. Dany, and Y. Yusrizal, “Analisis Data Mining untuk Prediksi Harga Saham: Perbandingan Metode Regresi Linier dan Pola Historis,” J. Teknol. Sist. Inf., vol. 4, no. 2, pp. 278–288, 2023, doi: 10.35957/jtsi.v4i2.5158.

[8] E. P. Ariesanto Akhmad, “Data Mining Menggunakan Regresi Linear untuk Prediksi Harga Saham Perusahaan Pelayaran,” J. Apl. Pelayaran dan Kepelabuhanan, vol. 10, no. 2, p. 120, 2020, doi: 10.30649/japk.v10i2.83.

[9] P. Chang Hartono and A. Dwiyoga Widiantoro, “Analisis Prediksi Harga Saham Unilever Menggunakan Regresi Linier dengan RapidMiner,” J. Comput. Inf. Syst. Ampera, vol. 5, no. 3, pp. 2775–2496, 2024.

[10] B. Panwar, G. Dhuriya, P. Johri, S. S. Yadav, and N. Gaur, "Stock market prediction using linear regression and SVM," in *2021 Int. Conf. on Advance Comput. and Innovative Technol. in Eng. (ICACITE)*, Uttar Pradesh, India, Mar. 2021, pp. 629-631. DOI: 10.1109/ICACITE51222.2021.9404785.

[11] C. C. Emioma and S. O. Edeki, "Stock price prediction using machine learning on least-squares linear regression basis," in *J. Phys.: Conf. Ser.*, vol. 1734, no. 1, p. 012058, 2021. DOI: 10.1088/1742-6596/1734/1/012058.

[12] G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat, "Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications," *Int. J. Financ. Stud.*, vol. 11, no. 3, p. 94, July 2023. DOI: 10.3390/ijfs11030094.

[13] N. Manasa, D. W. Praveenraj, and L. SR, "Predictive analytics for stock market trends using machine learning," in *2023 4th Int. Conf. Comput., Autom. and Knowl. Manag. (ICCAKM)*, Dubai, UAE, Dec. 2023, pp. 1-8. DOI: 10.1109/ACCESS.2020.3004351.

[14] M. C. Chan, C. C. Wong, and C. C. Lam, "Financial time series forecasting by neural network using conjugate gradient learning algorithm and multiple linear regression weight initialization," in *Comput. Econ. Finance*, Kowloon, Hong Kong, vol. 61, pp. 326-342, 2000.

[15] M. Göçken, M. Özçalıcı, A. Boru, and A. T. Dosdoğru, "Stock price prediction using hybrid soft computing models incorporating parameter tuning and input variable selection," *Neural Comput. Appl.*, vol. 31, pp. 577-592, 2019. DOI: 10.1007/s00521-018-3523-2.