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. 4 No. 2 (2024) 98 - 104 E-ISSN: 2809-8064

Comparative Analysis of Linear Regression, Decision Tree and Gradient Boosting for Predicting Stock Price of Bank Rakyat Indonesia

Rahma Dwi Ningsih[1*], Sarwido[2], Gentur Wahyu Nyipto Wibowo[3]
[1*,2,3]Informatics Engineering, Sains and Technology, Universitas Islam Nahdlatul Ulama Jepara
[1*]211240001140@unisnu.ac.id, [2]sarwido@unisnu.ac.id, [3]gentur@unisnu.ac.id

Abstract

Investing funds with the expectation of future profit is a common financial strategy. Among various investment types, stocks are particularly attractive due to their potential for high returns. However, stock prices can fluctuate rapidly, influenced by factors such as company performance, interest rates, economic conditions, and government policies. In Indonesia, PT Bank Rakyat Indonesia Tbk (BBRI) reported the highest profit among the top 10 banks by the end of March 2024, with a profit of IDR 13.8 trillion. This study compares the effectiveness of linear regression, decision tree, and gradient boosting in predicting the stock price of Bank Rakyat Indonesia. The study aims to identify the most effective method among the three models by evaluating their accuracy and precision using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2). The results show that linear regression provides the best performance with an MAE of 65.724, RMSE of 86.746, MAPE of 9.951%, and the highest R^2 value of 0.967, suggesting excellent predictability. Gradient Boosting also shows good performance with an MAE of 176.546, RMSE of 252.198, MAPE of 8.506%, and an R^2 value of 0.721. The Decision Tree has the lowest performance among the three models, with an MAE of 199.322, RMSE of 278.501, MAPE of 8.596%, and an R^2 value of 0.659. Based on the analysis, linear regression is recommended as the most reliable prediction model for practical use in this context, while gradient boosting can be considered for more accurate long-term predictions.

Keywords: *Stock Price Prediction, Linear Regression, Decision Tree, Gradient Boosting, Bank Rakyat Indonesia*

1. Introduction

An investment is the placement of a current amount of funds in the hope of generating a profit in the future [1]. There are several kinds of investments, generally including savings, lending, gold, deposits, bonds, stocks, and property [2].

A share is proof of ownership of a person or legal entity in a company and is usually in paper form [3]. Investing in stocks is a very exciting thing because it can bring great returns to investors. However, investors often face risks because the price of stocks always fluctuates very quickly [3]. Factors influencing stock price changes include company conditions and performance, dividends, interest rates, risks, economic conditions, government policies, current issues, inflation rates, and supply and demand[4]. In the stock market, there are nine stock sectors: agricultural, mining, basic and chemical industries, machinery, goods and consumer industries, property and building construction, infrastructure and transportation, trade and investment, and finance and banking [5].

In Indonesia, there are 10 banks with the highest net profits. By the end of March 2024, *PT Bank Rayat Indonesia Tb (BBRI)* had become the largest profitable bank in Indonesia. This is demonstrated by the profit recapitulation of the top 10 RI banks per quarter 1 in 2024, where BRI earned a profit of IDR 13.8 trillion. Higher returns attract more investors to purchase equities, increasing demand and potentially lowering stock prices. Therefore, accurate stock price prediction is crucial for investors to make informed decisions.

Predicting stocks is an attempt to forecast future stock prices in order to maximize potential returns for investors while making investment decisions [3]. Accurate predictions help investors make sound investment decisions. Various methods have been

Received: 19-07-2024 | Accepted: 02-08-2024 | Published: 05-08-2024

developed for stock price prediction, from traditional statistical models to advanced machine learning algorithms. This study focuses on comparing Linear Regression, Decision Tree, and Gradient Boosting methods for predicting the stock price of Bank Rakyat Indonesia.

Understanding the most effective method for stock price prediction is essential for investors and financial analysts. Accurate predictions can lead to better investment strategies and higher returns.

Linear regression makes it possible to model relationships between independent variables (inputs) and dependent variables in the form of linear equations [6]. Previous research on Bri's stock price prediction uses a linear regression algorithm as a strategy to sell and buy stocks found an optimal train and test ratio of 80:20 and the error result of the prediction calculated using MAPE produced a percentage of 13,773% for the test data [4].

Decision Tree is a classification algorithm expressed as a recursive partition of a sample space. A decision tree consists of a node that forms a root tree, which means a tree is directed by a knot called a root. A node with an outward edge is called an internal or test node. All other knots are called leaves. In the decision tree, each internal node divides spaces, for example, into two or more subspaces according to a particular discrete function of the value attribute [7].

Gradient Boosting algorithm combines a weak learner into a single strong learns iteratively[8]. In a previous study entitled “BCA Bank Share Price Prediction Using XGBoost” it was found that XGBoost good forecast accuracy with a MAPE of 4.01 percent using hyperparameter settings. However, due to the COVID-19 epidemic, the predictions were somewhat less accurate in March 2020 [8].

The inclusion of these models allows for a comprehensive comparison to identify the most accurate method for stock price prediction, focusing on minimizing errors using metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

The selection of Linear Regression, Decision Tree, and Gradient Boosting is based on their distinct approaches and proven effectiveness in previous studies. This study aims to compare the performance of Linear Regression, a basic yet powerful statistical method. Decision Tree, known for its interpretability and Gradient Boosting, recognized for its high accuracy and ability to handle complex relationships.

Based on the reasons and the research, the question that arises in this study is How to Compare Linear Regression, Decision Tree, and Gradient Boosting to Predict the Stock Price of Bank Rakyat Indonesia?

2. Research Methods

This research method starts with collecting data, processing data, training data, modelling, testing data and evaluation. Here's a diagram that shows in Figure 1.

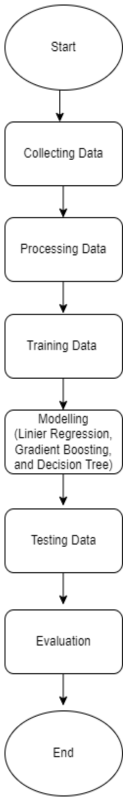


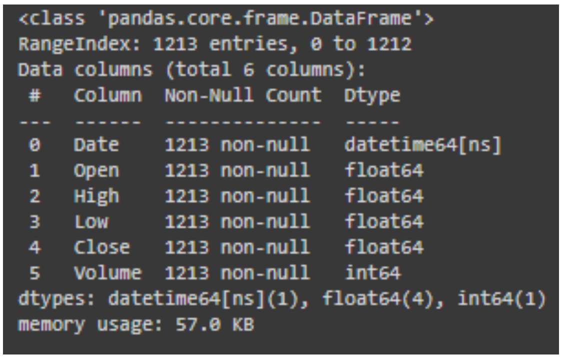
Figure 1. Flowchart Research

2.1 *Collecting Data*

At this stage, the collection uses a data set obtained from the Yahoo Finance website, accessible via the link https://finance.yahoo.com/quote/BBRI.JK/history/ . This data set is a data set of the stock price of Bank Rakyat Indonesia for the period from 01 July 2019 to 01 July 2024, and has seven attributes: Date, Open, High, Low, Close, Adj Close and Volume.

2.2 *Processing Data*

The preprocessing phase of this study is utilized to process the obtained datasets, thereby facilitating data processing on the model. The data set processing procedures will be performed based on the data sets that have already been obtained in order to expedite future data processing. This data is processed in two stages: first, the typed data date is changed to datetime, and then the adj close column is removed because it is unnecessary.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1213 entries, 0 to 1212
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    1213 non-null   datetime64[ns]
 1   Open    1213 non-null   float64
 2   High    1213 non-null   float64
 3   Low     1213 non-null   float64
 4   Close   1213 non-null   float64
 5   Volume  1213 non-null   int64
dtypes: datetime64[ns](1), float64(4), int64(1)
memory usage: 57.0 KB
```

Figure 2. Result Processing Data

2.3 *Data Training Splitting*

At this stage, the data that has been collected will be divided into two subsets: training data and testing data. This process is known as data splitting. This data division is very important in machine learning because it allows us to train models on specific data and test the performance of models. Where the ratio of each training data and testing data is 80:20.

2.4 *Modelling*

In this phase there are three methods of algorithms used to model the prediction of the stock price, and these are the following:

2.4.1 *Linier Regression*

A statistical procedure called a linear regression algorithm is used to determine the impact of one or more variables on a single variable. A variable that is subject to change is called a variable. The variables that affect are called independent or free variables. A regression equation where Y is the predicted value can be seen in equations 1 and 2 [9].

$$Y = a + b_1X_1$$

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n \qquad (1)$$

The values of a and b can be calculated using equations 2 and 3.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x^2)} \qquad (2)$$

$$b = \frac{(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x^2)} \qquad (3)$$

where, Y is the responsive variable or the non-free variable (dependent), X is the predictor or independent variable, a is the constant, and b is the regression coefficient.

2.4.2 *Decision Tree*

The decision tree consists of a set of nodes that are connected by branches. There are three types of nodes in the decision tree [10]:

1. A root node is a node that does not have an input but has more than one output.
2. An internal node has one input and more than two outputs.
3. A leaf or terminal node has one input and no output.

On every decision tree, each leaf has a class name. The root node and the internal node contain the rules used to distinguish data that has different characters. To calculate it, use the following formula:

$$Entropy(S) = \sum_{i=1}^{n} - pi.\,log_2\,pi \qquad (4)$$

Where S is the set of cases, n is the number of partitions of S, and pi is the ratio of Si to S.

To calculate it, use the following formula:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \qquad (5)$$

Where S is the set of cases, A is the feature, n is the number of partitions of the attribute A, |Si | is the ratio of Si to S, |S | is the number of cases in S.

2.4.3 *Gradient Boosting*

The gradient boosting algorithm works sequentially by adding previous predictors that do not match the prediction to the ensemble, ensuring that previously made errors are corrected [11]. Gradient Boosting starts with creating an initial model, then gradually adjusting new models by reducing prediction errors. This process is done continuously until the most accurate model is obtained [12].

$$-logL1 = - \sum_{i=1}^{N} y_i \log(odds)\ \log(1 + e^{\log(odds)}) \qquad (6)$$

2.5 *Testing Data*

After training, the model is tested on the testing data. The model makes predictions on this data, and these predictions are compared against the actual values to evaluate performance. This process is crucial because it helps us understand how well the model can generalize the knowledge it gained during training to new, unseen data. In other words, we want to ensure that the model is not just memorizing the training data but is also capable of making accurate predictions when given new data.

2.6 *Evaluation*

The evaluation metrics used to measure the performance of the model are Mean Absolute Error (MAE), Root Mean squared error (RMSE), Mean Absolute Percentage Error (MAPE) and R-squared (R^2).

2.6.1 *MAE*

Mean Absolute Error (MAE) is one of the evaluation methods commonly used in data science. MAE calculates the average of the absolute difference between predicted values and actual values[13].

In other words, the MAE calculates how much the absolute error averages in predictions. The smaller the MAE value, the better the quality of the model.

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |y_i - \hat{y}i|^2 \quad (7)$$

Where n is the number of samples in the data, yi is the actual value, ŷi is the prediction value

2.6.2 *RMSE*

Root Mean Square Error (RMSE) is a derivative of Mean Squared Error (MSE). The way to calculate it is to sum up all the squares of the prediction error, then divide the amount by the amount of prediction data, and finally take the root of the result. To calculate RMSE on equation 8 [14].

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (yi - \hat{y}i)^2} \quad (8)$$

Where n is the number of samples in yi data is the actual value ŷi is the prediction value

2.6.3 *MAPE*

Mean Absolute Percentage Error (MAPE) is another evaluation method used in data science. MAPE calculates the average of the percentage difference between the predicted value and the actual value [5].

In other words, MAPE calculates how much the average error in the prediction is as a percentage of the actual value. The smaller the value, the better the quality of the model. Next on Equation 9 is the calculation formula MAPE.

$$MAPE = \frac{1}{n} \sum \frac{|actual - forecast|}{|actual|} \times 100 \quad (9)$$

where n is the amount of data or record, forecast is the estimated or predicted value, and actual is the actual value.

To determine the accuracy of the MAPE method, the error rate is calculated in percentages. This percentage is the result of the formula MAPE, which can be calculated using equation 10 [9].

$$Level\ of\ accuracy = 100\% - MAPE \quad (10)$$

In the case of forecasting, a high level of accuracy serves as an indication that the predicted value is also very accurate. On the contrary, if the level of precision is low, then the predicted value is less accurate as well. In simpler terms, the value of the forecasted level is directly correlated with the value's accurate level. The value criteria MAPE show in table 1 [9].

Table 1. Table MAPE value criteria

MAPE value	Criteria
<10%	Very good
10% - 20%	Good
20% - 50%	Enough
>50%	Bad

2.6.4 *R^2 (determination coefficient)*

The value of the determination coefficient is $0 < R^2 < 1$. The higher the value of R^2, the better the model because of the greater diversity of dependent variables that can be described by independent variables. The calculation of R^2 can be done with equation 11[15].

$$R^2 = 1 - \frac{SS\ error}{SS\ total} = 1 - \frac{\sum_{i=1}^{n} (yi - \hat{y}i)^2}{\sum_{i=1}^{n} (yi - \bar{y})^2} \quad (11)$$

Where R^2 determination coefficient yi is the observation bound variable i ŷi is the value approximated with the regression model for observation i, $\bar{y}$ is the average of the observed variable in all observations.

3. Results and Discussion

3.1 *Description Dataset*

In this study, the data set obtained from the Yahoo Finance website, this data set is a data set of the stock price of Bank Rakyat Indonesia for the period beginning from July 1, 2019 to July 1, 2024 amounting to 1213 lines. Quotes are the history data of the Close Prize as shown in Figure 4.

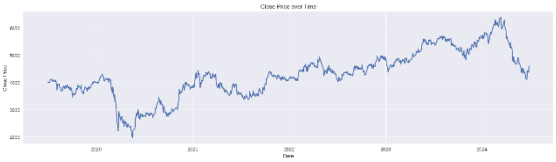


Figure 3. Close Price Over Time

3.2 *Split Dataset*

The datasets are divided into training and testing sets with a 80:20 ratio to test the performance of the predictive model. Split datasets based on figure 4

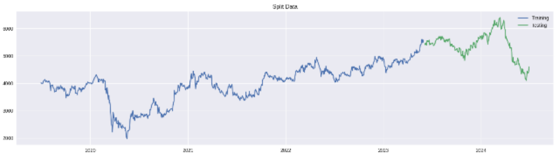
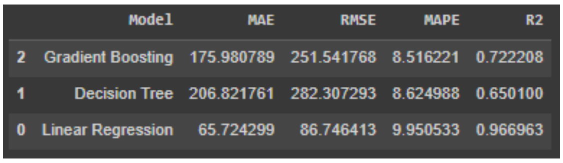


Figure 4. Split Dataset

3.3 *Result Evaluation Model*

Results from a linear regression, decision tree, and gradient boosting algorithm comparison like in figure 5.



	Model	MAE	RMSE	MAPE	R2
2	Gradient Boosting	175.980789	251.541768	8.516221	0.722208
1	Decision Tree	206.821761	282.307293	8.624988	0.650100
0	Linear Regression	65.724299	86.746413	9.950533	0.966963

Figure 5. Result Evaluation Model

Figure 5 shows that the Gradient Boosting model has good performance, with the lowest MAPE score of 8.51, indicating that this model is capable of predicting with relatively smaller errors than other models. Here is a comparison graph of the prediction data with the testing data on the Gradient Boosting model found in figure 6.

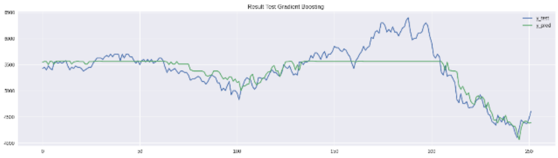
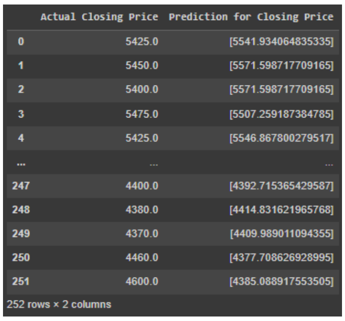


Figure 6. Result Testing Gradian Boosting

Here is the result comparison of the actual closing price and the prediction closing price using the gradient boosting model found in figure 7.



	Actual Closing Price	Prediction for Closing Price
0	5425.0	[5541.934064835335]
1	5450.0	[5571.598717709165]
2	5400.0	[5571.598717709165]
3	5475.0	[5507.259187384785]
4	5425.0	[5546.867800279517]
...	...	...
247	4400.0	[4392.715365429587]
248	4380.0	[4414.831621965768]
249	4370.0	[4409.989011094355]
250	4460.0	[4377.708626928995]
251	4600.0	[4385.088917553505]

252 rows × 2 columns

Figure 7. Comparation of Actual and Prediction Price of Stock using Gradient Boosting Model

Figure 8 is a comparative graph of the prediction data with the testing data on the Decision Tree model.

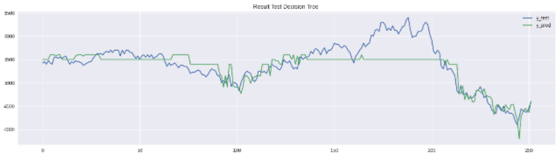
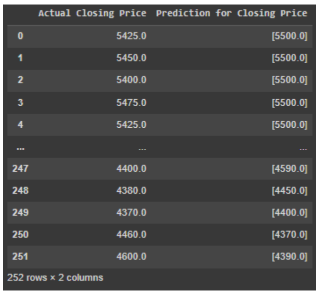


Figure 8. Result Testing Decision Tree

The result of the decision tree is an MAE of 206.82, with an RMSE of 282.30, a MAPE of 8.62, and an R^2 of 0.65. This model shows the least optimum performance among the three models tested. Although it has a fairly low MAPE, high MAE and RMSE values indicate that the model is less accurate in predicting stock prices than the other two models.



	Actual Closing Price	Prediction for Closing Price
0	5425.0	[5500.0]
1	5450.0	[5500.0]
2	5400.0	[5500.0]
3	5475.0	[5500.0]
4	5425.0	[5500.0]
...	...	...
247	4400.0	[4590.0]
248	4380.0	[4450.0]
249	4370.0	[4400.0]
250	4460.0	[4370.0]
251	4600.0	[4390.0]

252 rows × 2 columns

Figure 9. Comparation of Actual and Prediction Price of Stock using Decision Tree Model

Figure 9 show that the result comparison of the actual closing price and the prediction closing price using the decision tree model.

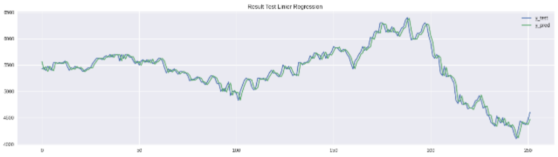
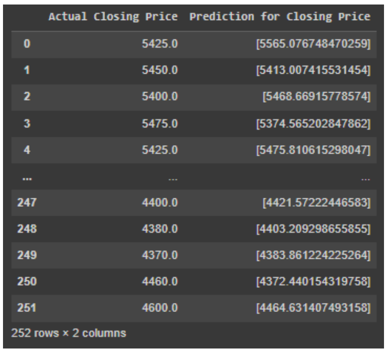


Figure 10. Result Test Linier Regression

On figure 10 is a comparison graph of the prediction data with the testing data of the Linear Regression model. The Linear Regression model showed the best performance with the lowest MAE and RMSE values, as well as the highest R^2 value of 0.96. It showed that linear regression was able to predict the stock price with a high degree of accuracy and relatively low error.



	Actual Closing Price	Prediction for Closing Price
0	5425.0	[5565.076748470259]
1	5450.0	[5413.007415531454]
2	5400.0	[5468.66915778574]
3	5475.0	[5374.565202847862]
4	5425.0	[5475.810615298047]
...	...	...
247	4400.0	[4421.57222446583]
248	4380.0	[4403.209298655855]
249	4370.0	[4383.861224225264]
250	4460.0	[4372.440154319758]
251	4600.0	[4464.631407493158]

252 rows × 2 columns

Figure 11.Comparation of Actual and Prediction Price of Stock using Linier Regression Model

Figure 11 show that the result comparison of the actual closing price and the prediction closing price using the linier regression model.

4. Conclusion

In this comparative analysis, three predictive models have been tested to predict the stock price of Bank Rakyat Indonesia: Linear Regression, Decision Tree, and Gradient Boosting. The results of the analysis show that each model has different performance in terms of accuracy and prediction precision.

Linear regression provides the best performance with a Mean Absolute Error (MAE) of 65,724, Root Mean Square Errors (RMSE) of 86.746, and Mean Absolute Percentage Error (MAPE) of 9.951%. The model also has the highest R-squared (R^2) value of 0.967, suggesting that linear regression has excellent predictability and can explain about 96.7% of variability in the data.

Gradient boosting showed good performance with an MAE of 176.546, an RMSE of 252.198, and a MAPE of 8.506%. The model has an R^2 value of 0.721, suggesting that gradient boosting is able to explain about 72.1% of the variability in the data. Although not as good as linear regression, this model still gives quite accurate results.

The decision tree has the lowest performance among the three models, with an MAE of 199.322, an RMSE of 278.501, and a MAPE of 8.596%. The R^2 value obtained at 0.659 indicates that the model is only able to explain about 65.9% of variability in the data, which means its predictive performance is less accurate compared to the other two models.

Based on the results of this analysis, it can be concluded that Linear Regression is the most effective model for predicting the stock price of Bank Rakyat Indonesia in this study. Although Gradient Boosting shows good performance in some metrics, Linear Regression remains superior in terms of overall accuracy. Therefore, Linear Regression is recommended as a reliable prediction model for practical use in this context. However, for a more accurate prediction in the long term, Gradient Boosting can be considered.

References

[1] R. R. Fitriani, Ernastuti, and E. R. Swedia, "Algoritma Learning Vector Quantization Dan Fuzzy K-NN Untuk Prediksi Saham Berdasarkan Pesaing," *J. Ilm. Teknol. dan Rekayasa*, vol. 24, no. 1, pp. 1–9, 2019.

[2] A. Putri Ramadhani, I. Afifah Septyasari, F. Nur Hasannah, and D. Kustiawati, "Investasi ditinjau dari Perspektif Ekonomi dan Ekonomi Islam," *J. Indones. Sos. Sains*, vol. 3, no. 12, pp. 1579–1589, 2022.

[3] M. Saham, D. I. Bank, B. R. I. Di, and B. Saham, "Analisis machine learning algoritma regresi linear untuk memprediksi saham di bank bri di bursa saham indonesia," vol. 6, no. 8, pp. 81–87, 2023.

[4] J. S. Putra, R. D. Ramadhani, and A. Burhanuddin, "Prediksi Harga Saham Bank Bri Menggunakan Algoritma Linear Regresion Sebagai Strategi Jual Beli Saham," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 2, no. 1, pp. 1–10, 2022.

[5] I. Prima and D. Ahmad, "Analisis Conditional Restricted Boltzman Machine Untuk Memprediksi Harga Saham Bank Syariah Indonesia," *J. Lebesgue J. Ilm. Pendidik. Mat.*

Mat. dan Stat., vol. 4, no. 1, pp. 409–416, 2023.

[6] P. Triya, N. Suarna, and N. D. Nuris, "Penerapan Machine Learning Dalam Melakukan Prediksi Harga Saham PT . Bank Mandiri (Persero) Tbk Dengan Algoritma Linear Regression," vol. 8, no. 1, pp. 1207–1214, 2024.

[7] N. Nazifah, C. Prianto, and C. A. Id, "Decision Tree Algoritma C4.5 dengan algoritma lainnya: Sistematic Literature Review," *J. Inform. dan Teknol. Komput.*, vol. 04, no. https://ejurnalunsam.id/index.php/jicom/, pp. 57–64, 2023, [Online]. Available: https://ejurnalunsam.id/index.php/jicom/

[8] B. Jange, "Prediksi Harga Saham Bank BCA Menggunakan XGBoost," *Arbitr. J. Econ. Account.*, vol. 3, no. 2, pp. 231–237, 2022.

[9] L. Alpianto, A. Hermawan, and Junaedi, "Moving Average untuk Prediksi Harga Saham dengan Linear Regression," *J. Buana Inform.*, vol. 14, no. 02, pp. 117–126, 2023,.

[10] Dwita Elisa Sinaga, Agus Perdana Windarto, and Rizki Alfadillah Nasution, "Analisis Data Mining Algoritma Decision Tree Pada Prediksi Persediaan Obat (Studi Kasus : Apotek Franch Farma)," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 2, no. 4, pp. 123–131, 2022.

[11] S. E. Suryana, B. Warsito, and S. Suparti, "Penerapan Gradient Boosting Dengan Hyperopt Untuk Memprediksi Keberhasilan Telemarketing Bank," *J. Gaussian*, vol. 10, no. 4, pp. 617–623, 2021.

[12] V. Atlantic, E. Sulistianingsih, and H. Perdana, "Gradient Boosting Machine Pada Klasifikasi Kelulusan Mahasiswa," *Bul. Ilm. Math. Stat. dan Ter.*, vol. 13, no. 2, pp. 165–174, 2024.

[13] S. Sugiyanto, "Predict high school students' final grades using basic machine learning," *J. Appl. Data Sci.*, vol. 2, no. 1, Jan. 2021.

[14] N. Puspithasari, N. Fadhilah, and N. Hayati, "Analisis Dan Evaluasi Implementasi Sistem Erp (Enterprise Resources Planning) Pada Pt Petrokopindo Cipta Selaras," *Darmabakti J. Pengabdi. dan Pemberdaya. Masy.*, vol. 4, no. 2, pp. 255–264, 2023.

[15] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, Jul. 2021.