Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

COMPARATIVE ANALYSIS OF THE APPLICATION OF FEATURE SELECTION IN RANDOM FOREST REGRESSION FOR STOCK PRICE PREDICTION

Emil Agusalim Habi Talib[1*], Alvina Felicia Watratan[2], Saharuddin[3]

[1]Universitas Muhammadiyah Makassar
[2,3]STMIK Profesional Makassar

[1]Email: emil@unismuh.ac.id
[2]Email: vinawatratan@stmikprofesional.ac.id
[3]Email: Saharuddin@stmikprofesional.ac.id

Abstract
The rapid development of information technology and data mining has encouraged the use of machine learning algorithms in various fields, including the financial sector and capital markets. One of the main challenges in stock price prediction is the large number of available variables, not all relevant to the target variable, potentially reducing accuracy and causing overfitting. This study aims to analyze the benefits of applying feature selection in improving the performance of the Random Forest Regression algorithm for stock price prediction. The dataset used in this research consists of ten years of historical stock price data from PT Aneka Tambang Tbk (ANTM). The research was conducted using an experimental approach by developing two models: (1) Random Forest Regression without feature selection and (2) Random Forest Regression with feature selection using the Spearman Correlation method. Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE). The experimental results show that the model with feature selection achieved better performance, with improvements in all evaluation metrics, such as reduced error values (MAE: 26.22; RMSE: 51.82; MAPE: 1.32%) and increased R^2 (0.9895). These findings confirm that integrating feature selection with Random Forest Regression can improve prediction accuracy, reduce model complexity, and minimize overfitting risk. Therefore, feature selection plays a significant role in enhancing the effectiveness of machine learning models in stock price prediction.
Keywords: *Feature Selection, Random Forest Regression, Spearman Correlation, Stock Price Prediction, Machine Learning*

INTRODUCTION

The development of information technology and data mining has encouraged the increased use of machine learning algorithms in various fields, including the financial sector and the capital market. One of the main problems in the analysis of financial data, especially stock price predictions, is the number of variables available. However, not all of them are relevant to the predicted target. The presence of irrelevant features can reduce the model's accuracy, increase complexity, and cause overfitting in the machine learning process. (Muhamad Zulfani & Dapadeda, 2024).

One widely used approach to overcoming these problems is feature selection, which aims to select the most relevant features for prediction. Various methods can use this technique, one of which is correlation-based, such as

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

Spearman Correlation and Correlation-Based Feature Selection (CFS), which can effectively identify the relationship between variables and targets. (Priantama & Yoga Siswa, 2022).

On the other hand, recent studies have also emphasized the importance of Multi-Criteria Decision Making (MCDM) in complex data-driven decision-making (Faisal, Irmawati, et al., 2025; Mulyadi et al., 2024). In the financial context, MCDM can help investors and market analysts consider fundamental, technical, and external indicators before making investment decisions. This aligns with the principle that a single factor and the interaction of various complex variables influence stock price predictions. Integrating MCDM concepts with machine learning also provides a more systematic framework for selecting features that contribute to predicted outcomes (Faisal et al., 2024; Faisal, Abd Rahman, et al., 2025).

Feature selection is an important technique in machine learning that aims to select a subset of variables most relevant to the prediction target. This technique has several advantages, including improving model performance, reducing complexity, and providing better data interpretation. (Armaya, 2024). Conceptually, reducing dimensions through feature selection can reduce model variance, speed up training time, and retain important information that affects prediction results. (Budiman et al., 2021).

Various feature selection methods have been used in applied studies in the academic and industrial realms. Techniques such as correlation-based methods, *information gain*, *wrapper methods*, and optimization-based methods, such as Genetic Algorithms, have been proven to produce a subset of more relevant features and improve the performance of predictive models. (Armaya, 2024; Budiman et al., 2021; Priantama & Yoga Siswa, 2022; Somantri & Khambali, 2017). Among these methods, Correlation-Based Feature Selection (CFS) and Spearman correlation are widely used because they can identify monotonic relationships between features and targets without assuming complete linearity. This approach is considered adequate, especially in data with non-linear relationships with weak to moderate correlation strengths (Priantama & Siswa, 2022).

The results of empirical research show that the application of CFS can eliminate features with low correlation that only add noise to the model. For example, in the educational domain, the use of CFS has significantly improved the performance of the Random Forest Classifier because the model is trained only with relevant features. (Priantama & Yoga Siswa, 2022)These findings reinforce the argument that although ensemble algorithms such as Random Forest are relatively resistant to irrelevant features, feature selection remains important for improving the quality of input and prediction results. (Budiman et al., 2021).

Random Forest Regression is an ensemble-based algorithm that has proven effective in handling nonlinear and complex data, including in the field of prediction. (Lestari & Astuti, 2022). The main advantage of this algorithm lies in its ability to combine the results of many decision trees through *bagging* techniques and random feature selection. These characteristics make Random Forest more robust against overfitting and able to work well on large datasets and high dimensions. (Budiman et al., 2021; Lestari & Astuti, 2022; Priantama

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

& Yoga Siswa, 2022).

However, although Random Forest is often said not to require strict feature selection, the presence of redundant or noisy features can still reduce computational efficiency and model accuracy. Therefore, the implementation of feature selection remains relevant to ensure the model works more optimally. Several studies confirm that integrating Random Forest with feature selection methods, such as Spearman Correlation or CFS, can improve model performance while reducing the risk of overfitting. (Budiman et al., 2021; Lestari & Astuti, 2022; Priantama & Yoga Siswa, 2022).

In stock price prediction, combining pre-processing time-series data (e.g., lag features, rolling statistics, and non-stationariness handling) with feature selection has improved the model's ability to capture relevant signals and suppress market noise. Studies have shown that non-linear machine learning algorithms, such as Random Forest Regression and Support Vector Regression (SVR), as well as deep learning methods (LSTM/CNN), often outperform classic statistical models such as ARIMA, especially on stock data that is non-linear and highly fluctuating. (Budiprasetyo et al., 2023; Fathoni et al., 2025; Kurnia et al., 2025; Kurniawati & Arima, 2021).

Meanwhile, deep learning approaches and hybrid models (e.g., CNN-LSTM, Attention-CNN-LSTM + XGBoost, or LSTM + Random Forest) are also reported to be able to improve accuracy when architecture and pre-processing are optimized (Budiprasetyo et al., 2023; Fathoni et al., 2025). However, deep learning models require large amounts of data and careful feature selection, so integration with ensembles such as Random Forest becomes a more practical strategy on medium-scale datasets.

Then, in the evaluation process, several previous studies recommended the use of standard regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE). These metrics aim to provide a comprehensive picture of the model's performance. With the combination of these metrics, the evaluation of regression models can be carried out more comprehensively, both in terms of accuracy, stability, and good data generalization ability. (Budiprasetyo et al., 2023; Karmilasari, 2022; Lestari & Astuti, 2022).

MAE measures the average absolute difference between the predicted value ($\hat{y}$) and the actual value (y), where all errors are given equal weight. MAE values that are getting closer to 0 indicate a smaller level of prediction error. MSE calculates the square average of the difference between the predicted and actual values. The squaring process provides a greater penalty for significant errors, making it sensitive to outliers. A small MSE value indicates the model has good accuracy but is sensitive to extreme values. (Lestari & Astuti, 2022).

RMSE is the square root of MSE, which presents the magnitude of the prediction error in the same unit as the target variable. A low RMSE value indicates a small rate of prediction error. R^2 measures the proportion of variation in dependent variables that can be explained by independent variables in the model. The value is in the range 0–1. An R^2 value close to 1 indicates a good model's ability to explain data variability (Karmilasari, 2022). MAPE calculates the average absolute error percentage between the predicted and actual values.

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

This metric presents accuracy in percentage form, making it easy to interpret. The smaller the MAPE value, the higher the prediction accuracy rate (Budiprasetyo et al., 2023).

Based on this background and the literature review results, this study aims to conduct a comparative analysis to identify the benefits of feature selection in improving the performance of the Random Forest Regression model on stock price prediction. The Random Forest algorithm was chosen because it has advantages in stability against non-linear data and resistance to outliers. The dataset used is historical data on shares of PT Aneka Tambang Tbk (ANTM) for the last 10 years. (Karmilasari, 2022)It is expected to provide a comprehensive overview of the effectiveness of integrating feature selection methods in Random Forest Regression in the context of stock price prediction.

METHODOLOGY

This study uses a quantitative approach with a machine learning-based computational experiment method. The model used in this study is *Random Forest Regression* as a *supervised learning* algorithm for prediction, which is combined with *feature selection methods* to optimize the selection of predictor variables. This study aims to identify the benefits of applying *the feature selection* method in regression analysis using *the Random Forest Regression* algorithm to predict stock price movements. In this study, several stages are carried out, such as data collection, data analysis, data preprocessing, algorithm training, model testing, and the last one is model evaluation. (Fathoni et al., 2025). The procedure of this research can be seen in Figure 1 below.

Figure 1. Research Procedure

Data Collection

The dataset used in this study is historical data on the share price of PT Aneka Tambang Tbk (ANTM). (Investing, 2025) This includes stock price data for the last ten years, from August 1, 2015, to July 31, 2025. The dataset structure used in this study consists of 7 columns and 2,415 rows, which contain several main variables, such as the date variable that shows the date the stock price was recorded in the market. The Last variable represents the stock's closing price at the end of the trading session, and the Opening variable represents the opening price at the beginning of the trading session. The High and Lowest variables, respectively, describe the stock's highest and lowest prices reached during the trading session on that date. The Vol. (*Volume*) variable indicates the total number of shares traded, and the Change% variable indicates the percentage change in the stock price compared to the closing price on the previous trading day. Table 1 presents historical data on shares of PT Aneka Tambang Tbk (ANTM) used in this study.

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

Table 1. Stock Historical Data PT Aneka Tambang Tbk (ANTM)

No.	Date	Last	Opening	Highest	Lowest	Vol.	Change%
1	31/07/2025	2.850	2.950	2.970	2.770	235,21M	-4,68%
2	30/07/2025	2.990	3.010	3.050	2.970	58,85M	-0,33%
3	29/07/2025	3.000	2.970	3.020	2.970	53,89M	1,01%
…	…..	…..	…..	…..	…..	…..	…..
2414	04/08/2015	500	407	500	405	127,69M	24,38%
2415	03/08/2015	402	403	417	397	31,01M	0,75%

Exploratory Data Analysis

Before data *preprocessing*, exploratory data analysis *(EDA) is carried out* to understand the dataset's characteristics and structure. Through EDA, we can determine the *data preprocessing* method that best suits the dataset's conditions. In this study, the stages of EDA carried out include: (1) importing and displaying datasets, (2) identifying data types, (3) detecting missing *values*, (4) analyzing data distribution using descriptive statistics and visualization, (5) creating box *plots* to identify the existence of *outliers* and (6) visualizing the trend of the closing price of shares. This is done to ensure the quality of the data, identify potential problems that can affect the results of the analysis, and obtain an initial picture of the patterns, trends, and anomalies contained in the data.

Preprocessing Data

The next step is data preprocessing after the exploratory data analysis (EDA) stage. This stage aims to prepare a dataset that will be used in the machine learning algorithm training process. In this study, the *data preprocessing* method applied is dataset *encoding*, which changes the data type from the format of objects or characters to numerical data types such as *integers* or *floats*. This process is important because most *machine learning* algorithms can only process data in numerical form. This study did not apply other data preprocessing methods, such as dataset normalization. This is due to the characteristics of the algorithm used, namely *Random Forest*, which is robust to the data scale, so the normalization process is unnecessary.

Training Algorithm

After the data *preprocessing* stage, the next step is machine learning algorithm training or the prediction model development process. In this study, model development was carried out through two approaches: (1) model development without applying *the feature selection* method, and (2) model development by applying *the feature selection method*. The application of these two approaches aims to conduct a comparative analysis of the performance of the built model, so that it can be identified to what extent the use of *feature selection* can contribute to improving the performance of machine learning models.

The feature selection used in this study is the Spearman Correlation method. This method was chosen because it can measure the strength and direction of the relationship between two variables in a monotonic manner, both linear and non-linear, and does not depend on data distribution. Thus, this method is expected to help identify the most relevant features to the target variable while reducing model complexity and potential overfitting. The

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

Spearman Correlation formula can be seen in equation (1) (Bocianowski et al., 2023).

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (1)$$

Where,

ρ = Sparman rating correlation coefficient
d_i = Difference in rank between two variables for observation *i*
n = Number of observation pairs

The feature-free approach uses all the variables in the *preprocessing* dataset to make predictions. In contrast, the *feature-selected* approach only uses a subset of variables selected based on their relevance to the target variable. This comparison is important because *feature selection* not only has the potential to improve prediction accuracy, but it can also reduce model complexity, speed up training time, and minimize the risk of *overfitting*. Thus, the results of this analysis are expected to provide empirical justification regarding the effectiveness of *feature selection* in developing *machine learning models* in the context of the research conducted.

In the model training process, the dataset is divided into two parts, namely 80% training data and 20% testing data. This division aims to enable the model to learn complex patterns from training data optimally, then evaluate using never-before-seen data to measure the model's generalization capabilities. Thus, this approach is expected to provide a more accurate picture of the effectiveness of the application of feature selection in improving model performance. The flow of the model development process is shown in Figure 2.

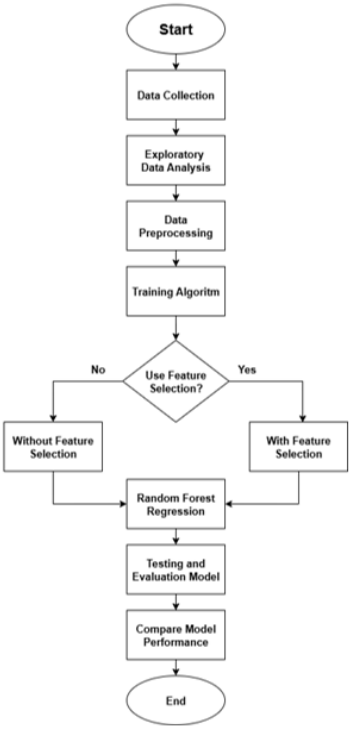

Figure 2. Model Development Process Flow

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

Testing Model

After the model training process is completed, the next stage is model testing using *data testing*. This test assesses how much the *resulting machine learning* model can predict new data. In other words, it measures the model's generalization ability against data that has never been seen before. In addition, this stage ensures that the model can recognize patterns from training data and produce consistent and accurate predictions on new data.

Model Evaluation

Finally, the evaluation process was carried out to measure the performance of the Random Forest Regressor model in predicting the target value. Five evaluation metrics were used to assess the overall performance of the model, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE). These five metrics aim to provide a comprehensive picture of the model's accuracy, consistency, and generalization capabilities on test data. The formula of each evaluation metric used in this study is presented in Equations (2), (3), (4), (5), and (6) (Budiprasetyo et al., 2023; Karmilasari, 2022; Lestari & Astuti, 2022).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

$$\text{RMSE} = \sqrt{\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{6}$$

Where.

y_i = actual value
$\hat{y}_i$ = predicted value
n = amount of data

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

$\bar{y}$ = average Score

RESULT AND DISCUSSION

This study was conducted to test the benefits of applying the feature selection method in regression analysis using the Random Forest Regression algorithm. Several important stages must be implemented, including data collection, *exploratory data analysis* (EDA), pre-data processing, algorithm training, model testing, and model performance evaluation using various evaluation metrics. Each stage is designed so that the analysis process runs structured and systematically, so that the results obtained can provide an accurate picture of the influence of *feature selection* on model performance. The following are the results of the research that has been conducted.

In this study, the first step after obtaining training data is conducting exploratory data analysis (EDA). The initial stage of EDA begins with importing and displaying the dataset to determine the initial state of the data obtained. Figure 3 shows the import process results, and the appearance of the initial data can be seen in Figure 3.

	Tanggal	Terakhir	Pembukaan	Tertinggi	Terendah	Vol.	Perubahan%
0	2025-07-31	2850	2950	2970	2770	235,21M	-4,68%
1	2025-07-30	2990	3010	3050	2970	58,85M	-0,33%
2	2025-07-29	3000	2970	3020	2970	53,89M	1,01%
3	2025-07-28	2970	2970	3030	2950	85,13M	0,00%
4	2025-07-25	2970	3040	3040	2950	61,09M	-1,98%
...	...	...	...	...	...	...	...
2409	2015-08-07	504	517	529	500	23,16M	-2,51%
2410	2015-08-06	517	512	546	491	85,39M	1,77%
2411	2015-08-05	508	512	550	504	135,09M	1,60%
2412	2015-08-04	500	407	500	405	127,69M	24,38%
2413	2015-08-03	402	403	417	397	31,01M	0,75%

2414 rows × 7 columns

Figure 3. ANTM Stock Dataset Import Results

The data import results show seven columns and 2414 rows of historical data of ANTM shares. Then, data types and missing values are identified to determine the data types and the number of missing values contained in the dataset. The results of the identification of data types and missing values can be seen in Figures 4 and 5.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2414 entries, 0 to 2413
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Tanggal     2414 non-null   datetime64[ns]
 1   Terakhir    2414 non-null   int64
 2   Pembukaan   2414 non-null   int64
 3   Tertinggi   2414 non-null   int64
 4   Terendah    2414 non-null   int64
 5   Vol.        2414 non-null   object
 6   Perubahan%  2414 non-null   object
dtypes: datetime64[ns](1), int64(4), object(2)
memory usage: 132.1+ KB
```

Figure 4. Identify Data Types

	0
Tanggal	0
Terakhir	0
Pembukaan	0
Tertinggi	0
Terendah	0
Vol.	0
Perubahan%	0

dtype: int64

Figure 5. Identifying Missing Value

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

The results of data type and *missing value* identification showed that the dataset had three types of data types, namely *the datetime* data type for the date column, *the integer* data type for the last, opening, highest, and lowest columns, and the object data type for the volume and change columns (%). Furthermore, the results of identifying missing values show no lost value in the historical data of ANTM shares. After that, data distribution analysis was carried out using descriptive statistics and visualization to understand the characteristics of the data, such as mean value, median, standard deviation, minimum value, maximum value, and data distribution for each variable. The results of data distribution identification with descriptive statistics and visualization can be seen in Figures 6 and 7.

	Tanggal	Terakhir	Pembukaan	Tertinggi	Terendah	Vol.	Perubahan%
count	2414	2414.000000	2414.000000	2414.000000	2414.000000	2414.0	2414.000000
mean	2020-07-22 14:59:33.156586752	1330.025684	1333.011599	1357.507042	1308.277133	11640247721.623861	0.128041
min	2015-08-03 00:00:00	287.000000	290.000000	294.000000	285.000000	358000000.0	-14.370000
25%	2018-01-25 06:00:00	735.000000	735.000000	750.000000	720.000000	3881250000.0	-1.467500
50%	2020-07-21 12:00:00	995.000000	997.500000	1022.500000	975.000000	6685500000.0	0.000000
75%	2023-01-05 18:00:00	1953.750000	1955.000000	1980.000000	1933.750000	12632000000.0	1.337500
max	2025-07-31 00:00:00	3550.000000	3590.000000	3660.000000	3470.000000	219000000000.0	24.840000
std	NaN	725.742924	726.815792	741.308188	711.591018	17101392197.247293	3.101100

Figure 6. Data Distribution with Descriptive Statistics

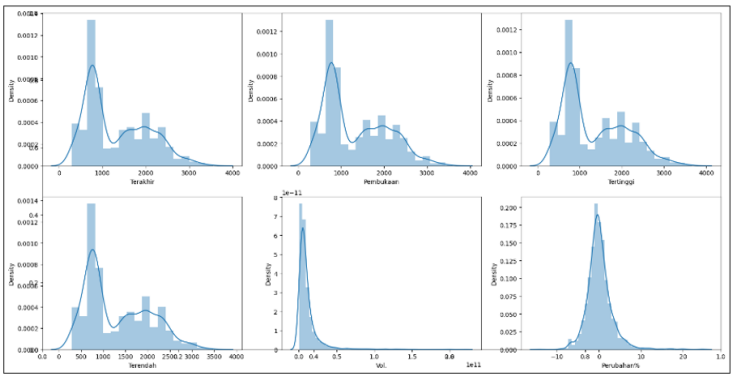

Figure 7. Data Distribution Visualization

Based on the table, it is known that the number of data (count) available for each variable is 2,414 entries, with no missing values. The average value (mean) of the closing price (Last) is 1,330, the opening price is 1,333, the high price is 1,357, and the low price is 1,308. The average trading volume was around 1.16 billion shares, while the average daily price change percentage was 0.12%. Each variable's minimum and maximum values show the range of stock price fluctuations from 2015 to 2025, with the lowest price being 287 and the highest price being 3,550. A considerable standard deviation in the price and volume variables indicates significant variation during the observation period.

Then, based on the visualization results, the data showed significant variation across all variables, with a distribution pattern that was generally right-skewed. Next, a boxplot visualization was carried out to identify outliers in each variable. The visualization of the boxplot can be seen in Figure 8.

Based on the boxplot visualization results, an *outlier* was identified in the Volume and Change variables. *Outliers* in the Volume variable occur due to

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

specific periods in which trading activity jumps sharply, which can be influenced by factors such as market sentiment, corporate action, or important economic events. Meanwhile, *the outlier* in the % Change variable reflects days with very extreme price movements, both up and down, which can be triggered by the company's internal factors and external market conditions.

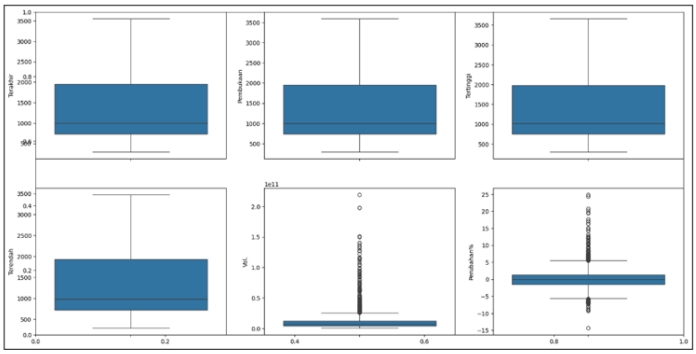

Figure 8. Visualization of the Boxplot.

Furthermore, the trend of price movements at the closing of the stock is visualized to identify the pattern of price fluctuations from time to time. This can provide an overview of the trend of price ups and downs and the period of significant changes. The results of the visualization of the closing price trend can be seen in Figure 9.

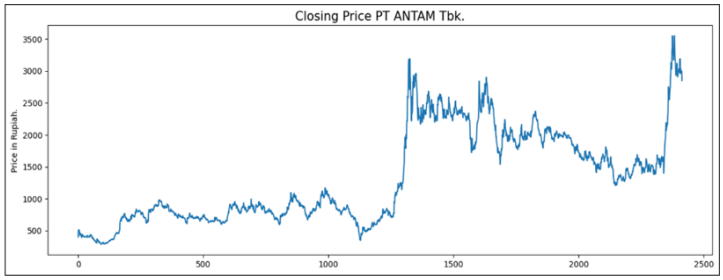

Figure 9. Visualization of the Trend of Stock Closing Price Movements

The visualization results show a fluctuating price trend with two significant spikes: the first around the middle of the period, where prices jump sharply from ±500–1,000 to more than 3,000, and the second at the end with similar increases. The surge indicates a period of very rapid price growth, likely influenced by the company's fundamental factors or strong market sentiment.

After exploratory data analysis, the next stage is data preprocessing. The preprocessing carried out in this study includes the process of encoding or converting data formats on the variables Volume and Change% so that the model can process them. *Data encoding* on the Volume variable is done by converting the values that were initially in the form of text with units "B" (billion), "M" (million), and "K" (thousand) to numerical values in units of shares. In this process, the comma (,) is removed from the initial value, then the unit letters are identified and converted according to their multiples, i.e., "B" is multiplied by 1,000,000,000, "M" is multiplied by 1,000,000, and "K" is multiplied by 1,000.

Then *the encoding* of the % Change variable is carried out by changing the value originally in the form of a percentage text to a decimal number. This

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

process begins by removing the percent (%) symbol from each value to avoid interfering with converting to a numerical format. Next, the comma (,) used as a decimal separator is replaced with a period (.) to match the numeric format in Python programming. After that, all values are converted into numeric data types. With these two methods, the data on the Volume and Change% variables, initially in text form, can be used directly in statistical analysis and machine learning modeling.

After data encoding, the algorithm training process or machine learning model development is carried out. Before the training process, the dataset is split into two parts, namely training data and test data, with a proportion of 80% for training data and 20% for test data. After that, the algorithm training process can be carried out. The algorithm training process is carried out in two stages: algorithm training without feature selection and algorithm training using feature selection.

The first stage of training was carried out without using feature selection, where the training process was carried out by utilizing all independent variables in the dataset, such as Date, opening, highest, lowest, vol., and change. Then the second stage of training was carried out using feature selection using the Spearman Correlation method. The results of applying the feature selection method were identified as four independent variables with the most influence on the dependent variables. The variables consist of Date, opening, highest, and lowest. Visualization of the results of the application of the Spearman method can be seen in Figure 10.

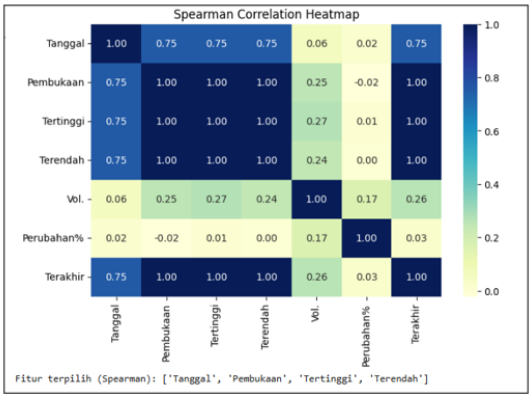

Figure 10. Visualization of Feature Selection Results with the Spearman Correlation Method

The two models' performance is shown in Table 2 based on the training results, while the graphs of the prediction results of each model can be seen in Figure 11 and Figure 12.

Table 2. Predictive Model Performance

Model	MAE	MSE	RMSE	R^2	MAPE
Random Forest Regression	30.4684	3206.9353	56.6298	0.9874	1.57%
Random Forest Regression + Feature Selection	26.2276	2685.3927	51.8208	0.9895	1.32%

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

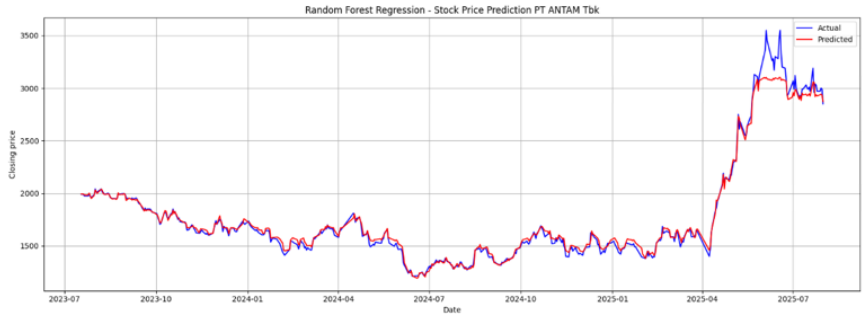

Figure 11. Random Forest Regression Algorithm Prediction Results Chart

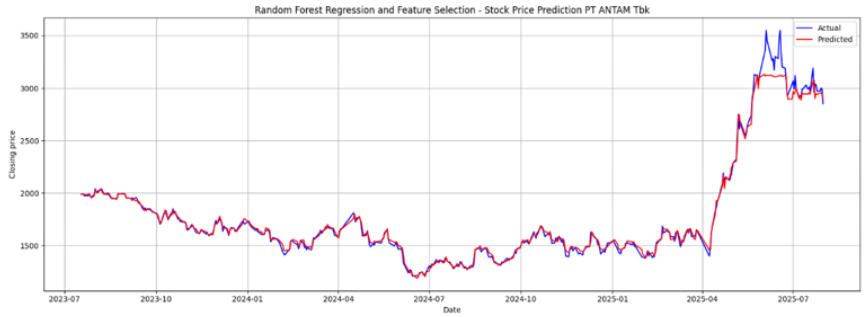

Figure 12. Random Forest Regression Algorithm Prediction Results Graph with Feature Selection

This study evaluates the performance of the Random Forest Regression algorithm in predicting the share price of PT ANTAM based on historical data over the last ten years. Two training scenarios were carried out: (1) algorithm training without the application of feature selection, and (2) algorithm training with the application of feature selection based on Spearman correlation. The model evaluation is based on five main metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE).

The test results in Table 2 show that the implementation of feature selection resulted in significant improvements in all evaluation metrics. In the model without feature selection, the MAE value was recorded at 30.4684, MSE 3206.9353, RMSE 56.6298, R^2 0.9874, and MAPE 1.57%. Meanwhile, errors in the model with feature selection Spearman correlation decreased with MAE 26.2276, MSE 2685.3927, RMSE 51.8208, and MAPE 1.32%. In addition, the R^2 value increased to 0.9895. In general, these results suggest that the elimination of highly correlated features can improve the accuracy of the model while improving generalization capabilities (Budiprasetyo et al., 2023; Karmilasari, 2022; Lestari & Astuti, 2022).

Visualizing the prediction graphs in Figures 11 and 12 supports these quantitative findings. The model with feature selection produces a prediction pattern closer to the actual value, while in the model without feature selection, there is a relatively larger deviation. This suggests that reducing redundancy between features through the correlation method helps the Random Forest model focus more on the genuinely relevant variables to the target. (Budiman et al., 2021; Priantama & Yoga Siswa, 2022).

These findings align with the latest literature that emphasizes the

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

importance of feature selection in improving machine learning algorithms' performance. According to (Priantama & Yoga Siswa, 2022; Pudjihartono et al., 2022)Selecting the right features reduces model complexity, improves prediction accuracy, and reduces the risk of overfitting, especially on high-dimensional data. In the context of regression, a study by (Hwang et al., 2023; Lestari & Astuti, 2022) Also, trimming excess features in Random Forest Regression can improve important variables' interpretability and R^2 values. In addition, correlation-based methods such as Spearman correlation effectively identify redundant features so that the model works more efficiently. (Hwang et al., 2023; Karmilasari, 2022; Lestari & Astuti, 2022; Rickert et al., 2023).

Practically, the application of feature selection in this study has a real impact in reducing prediction errors (MAE, MSE, RMSE, MAPE) while increasing goodness-of-fit (R^2). This confirms that integrating simple but systematic feature selection methods can be an important strategy in Random Forest-based regression analysis, especially in stock price prediction with complex data patterns and potential multicollinearity between variables.

CONCLUSION AND SUGGESTION

This study proves that applying the Spearman Correlation feature selection method in the Random Forest Regression algorithm can improve the performance of PT Aneka Tambang Tbk (ANTM) stock price predictions. The test results showed that models with feature selection had lower error values (MAE, MSE, RMSE, and MAPE) and increased coefficient of determination values (R^2) than those without feature selection. This indicates that eliminating less relevant features can improve accuracy, speed up the training process, and reduce model complexity. Thus, integrating feature selection methods is an important strategy in increasing the effectiveness of Random Forest Regression, especially in the case of stock price prediction with non-linear and complex data characteristics.

Based on the results of this study, several suggestions can be submitted for further research. First, advanced model development can be done by comparing other feature selection methods, such as Information Gain, Recursive Feature Elimination (RFE), and optimization-based methods, such as Genetic Algorithm, to identify the most effective approach in stock price prediction. Second, the model can be further developed by combining Random Forest with *deep learning* algorithms (e.g., LSTM or CNN-LSTM) to capture non-linear patterns and temporal dynamics in stock data. Third, it is also important to expand the dataset by using data from various sectors or other companies to test the consistency of the effectiveness of *the feature selection method*. In addition, adding fundamental variables and external factors, such as macroeconomic conditions and market sentiment, can potentially improve the quality of prediction results.

REFERENCES

Armaya, A. M. R. (2024). Pengaruh Feature Selection Dan Feature Extraction Dalam Peningkatan Akurasi Klasifikasi Kebakaran Hutan. *JuTI "Jurnal Teknologi Informasi,"* 3 (1), 13.

Bocianowski, J., Wrońska-Pilarek, D., Krysztofiak-Kaniewska, A., Matusiak,

K., & Wiatrowska, B. (2023). Comparison of Pearson's and Spearman's Correlation Coefficients Values for Selected Traits of Pinus sylvestris L. 17, 302.

Budiman, S., Sunyoto, A., & Nasiri, A. (2021). Analisa Performa Penggunaan Feature Selection untuk Mendeteksi Intrusion Detection Systems dengan Algoritma Random Forest Classifier. *Sistemasi*, 10 (3), 753.

Budiprasetyo, G., Hani'ah, M., & Aflah, D. Z. (2023). Prediksi Harga Saham Syariah Menggunakan Algoritma Long Short-Term Memory (LSTM). *Jurnal Nasional Teknologi Dan Sistem Informasi*, 8 (3), 164-172.

Faisal, M., Abd Rahman, T. K., Zainal, D., Mubarak, H., Shabir, F., Anwar, N., & Asrowardi, I. (2025). Utilizing Machine Learning-Based Decision-Making to Align Higher Education Curriculum with Industry Requirements. *International Journal of Modern Education and Computer Science*, 17 (4), 1-25.

Faisal, M., Irmawati, Rahman, T. K. A., Jufri, Sahabuddin, Herlinah, & Mulyadi, I. (2025). A Hybrid MOO, MCGDM, and Sentiment Analysis Methodologies for Enhancing Regional Expansion Planning: A Case Study, Luwu - Indonesia. *International Journal of Mathematical, Engineering and Management Sciences*, 10 (1), 163-188.

Faisal, M., Rahman, T. K. A., Mulyadi, I., Aryasa, K., Irmawati, & Thamrin, M. (2024). A Novelty Decision-Making Based on Hybrid Indexing, Clustering, and Classification Methodologies: An Application to Map the Relevant Experts Against the Rural Problem. *Decision Making: Applications in Management and Engineering*, 7 (2), 132-171.

Fathoni, F., Ibrahim, A., Septiana, R., Rielisa Putri, A., Ispahan, T., & Shifa Maharani, W. (2025). Analisis Prediksi Harga Saham Pada Perusahaan T Menggunakan Kombinasi Cnn-Lstm. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9 (4), 6669-6675.

Hwang, S. W., Chung, H., Lee, T., Kim, J., Kim, Y. J., Kim, J. C., Kwak, H. W., Choi, I. G., & Yeo, H. (2023). Feature importance measures from a random forest regressor using near-infrared spectra to predict kraft lignin-derived hydrochar's carbonization characteristics. *Journal of Wood Science*, 69 (1).

Investing. (2025). *Aneka Tambang Persero Tbk (ANTM)*. Investing.Com.

Karmilasari, S. D. K. (2022). Implementasi Long Short-Term Memory Pada Prediksi Harga Saham PT Aneka Tambang Tbk. *Jurnal Ilmiah Komputasi*, 21 (1).

Kurnia, F. A., Hardianti, M., Sinurat, M., & Cahyadi, L. (2025). Analisis Prediksi Harga Saham PT. BCA Dengan Menggunakan Metode ARIMA. *ECo-Fin*, 7 (2), 880-896.

Kurniawati, A., & Arima, A. (2021). Analisis Prediksi Harga Saham PT. Astra International Tbk Menggunakan Metode Autoregressive Integrated Moving Average (ARIMA) dan Support Vector Regression (SVR). *Jurnal Ilmiah Komputasi*, 20 (3), 417-423.

Lestari, E. S., & Astuti, I. (2022). Penerapan Random Forest Regression Untuk Memprediksi Harga Jual Rumah Dan Cosine Similarity Untuk Rekomendasi Rumah Pada Provinsi Jawa Barat. *Jurnal Ilmiah FIFO*, 14 (2), 131.

Nusantara Hasana Journal
Volume 5 No. 3 (Agustus 2025), Page: 334-348
E-ISSN : 2798-1428

Muhamad Zulfani, & Dapadeda, A. (2024). Prediksi Harga Saham Menggunakan Algoritma Neural Network. *Jurnal Teknologi Informasi: Jurnal Keilmuan Dan Aplikasi Bidang Teknik Informatika*, 18 (1), 1-6.

Mulyadi, I., Thamrin, M., Faisal, M., Yunarti, S., Saharuddin, Abd Djalil, A., & Mallu, S. (2024). A Hybrid Model for Palm Sugar Type Classification: Advancing Image-Based Analysis for Industry Applications. *Ingénierie Des Systèmes d Information*, 29 (5), 1937-1948.

Priantama, Y., & Yoga Siswa, T. A. (2022). Optimasi Correlation-Based Feature Selection Untuk Perbaikan Akurasi Random Forest Classifier Dalam Prediksi Performa Akademik Mahasiswa. *JIKO (Jurnal Informatika Dan Komputer)*, 6 (2), 251.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, 2.

Rickert, C. A., Henkel, M., & Lieleg, O. (2023). An efficiency-driven, correlation-based feature elimination strategy for small datasets. *APL Machine Learning*, 1 (1).

Somantri, O., & Khambali, M. (2017). Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, 6 (3), 301-306.