Clean Energy and Smart Technology
Volume 04, Number 01, October 2025
DOI : https://doi.org/10.58641/cest.v4i1.173

e-ISSN : 2964-2647

GOLD PRICE PREDICTION SYSTEM USING THE RANDOM FOREST METHOD

Kurnia Agung Prastyo[1)*], Hidayatus Sibyan[2)], Nur Hasanah[3)]
[1)2)3)] Universitas Sains Al-Qur'an, Indonesia
[1)]katsuragi.k4gur0@gmail.com, [2)]hsibyan@unsiq.ac.id, [3)]nurhasanah@unsiq.ac.id

*katsuragi.k4gur0@gmail.com

Submitted :20 Oktober 2025 | **Accepted** : 28 Oktober 2025| **Published** : 31 Oktober 2025

Abstract: Gold is one of the most important commodities, serving as an investment instrument and a hedge against inflation. The high volatility of gold prices demands accurate predictions to support investment decision-making. This study aims to develop a gold price prediction system using the Random Forest method based on machine learning. The dataset used consists of daily gold prices from Yahoo Finance covering the period from 2020 to 2024. The research stages include data collection, preprocessing, model training, evaluation, and implementation into an interactive website. Evaluation results show a MAE of 329.31, MSE of 148,599.40, RMSE of 385.49, and a negative R² value (-1.97), indicating the model is not yet accurate. However, the system can provide a general overview of gold price trends and can be further improved to enhance prediction accuracy.

Keywords: gold price prediction, Random Forest, machine learning, investment

1. INTRODUCTION

The rapid development of digital technology in recent decades has had a significant impact on various sectors of life, including finance and investment. One result of this progress is the emergence of the concept of Financial Technology (Fintech), which integrates information technology with financial services to create faster, more efficient, and more accessible solutions. In the investment context, Fintech enables various financial instruments to be analyzed and managed more sophisticatedly, including in terms of prediction and data-driven decision-making (Kandregula, 2018).

One investment instrument that remains a primary choice amidst market dynamics is gold. Gold is viewed as a commodity with relatively stable long-term value and functions as a safe haven, especially during economic crises or global uncertainty. These advantages make gold a frequent component of investors' portfolio diversification strategies. However, despite its long-term stability, gold prices are also heavily influenced by various external factors such as inflation rates, exchange rate fluctuations, interest rates, and global political and economic conditions (Changani, 2024). As a result, gold prices experience significant fluctuations, which in turn pose challenges to price prediction for market players and investors.

To address these challenges, data-driven approaches and predictive technology are increasingly necessary. One rapidly developing method is machine learning-based forecasting. Machine learning is a branch of artificial intelligence that enables computers or systems to learn patterns from historical data without the need for explicit programming. In the context of predicting commodity prices, such as gold, this approach is considered capable of addressing the complexity of dynamic and non-linear data.

Of the various algorithms used in machine learning, Random Forest is one of the most widely adopted methods due to its reliability in processing large and complex data sets. Random Forest works by forming a number of decision trees, which are then combined to produce a more accurate final prediction. The main advantages of this algorithm are its ability to avoid overfitting, maintain prediction stability, and produce good performance on data with many variables or noise. Furthermore, this algorithm is also able to handle time series data effectively, which is particularly relevant in the case of daily gold price predictions (Landge, 2024).

Several previous studies support the effectiveness of the Random Forest algorithm in predicting commodity prices. Research conducted by Wahyuningsih (2024) and Hutagalung (2023), for example, demonstrated that a Random Forest-based predictive model can produce precise results with relatively low prediction error, as measured by Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These results strengthen the rationale for using Random Forest as the primary algorithm for gold price forecasting.

Based on this background and challenges, this research aimed to develop a machine learning-based gold price prediction system using the Random Forest algorithm. This research utilized historical gold price data as model input and evaluated model performance using several statistical metrics, namely MAE, Mean Squared Error (MSE), RMSE, and the coefficient of determination (R-squared/R²). The developed predictive system is expected to provide useful information for investors in developing more precise, measurable, and data-driven gold investment strategies.

Furthermore, from an academic and scientific and technological perspective, this research is expected to serve as a reference in the application of machine learning algorithms for commodity price prediction, as well as encourage the use of the Random Forest method in similar problems in the fields of economics and finance.

2. METHOD

This research used a quantitative approach to predict gold prices using machine learning algorithms, specifically the Random Forest method. The data used was secondary and obtained from Yahoo Finance in the form of historical daily gold closing prices for a five-year period, from January 2020 to December 2024. The dataset consisted of 1,257 daily data points, with the independent variable being the time index (Day) and the dependent variable being the daily gold closing price in USD per troy ounce.

Before building the prediction model, a data pre-processing stage was performed to ensure data integrity and readiness. The first step was to check for missing values, and the results showed that the data was clean and complete, allowing for direct use without imputation. Next, the time (date) variable was transformed into a new numeric feature, "Day," which represents the number of days since the first date in the dataset. This transformation is necessary because the Random Forest algorithm cannot directly interpret time-type data. Normalization was not performed because Random Forest is insensitive to variable scale, unlike other algorithms such as Support Vector Regression, which requires a uniform scale across features.

After the transformation process was complete, the data was divided into two main groups: 80% as the training set and 20% as the testing set. The division was done chronologically to maintain consistency in the time series data, where time sequence is important in reflecting real-world prediction scenarios. Specifically, the training data covered the period from January 2020 to the end of September 2023, while the testing data covered the period from early October 2023 to December 2024.

The prediction model was built using the Random Forest Regressor algorithm, a decision tree-based ensemble learning method. This algorithm works by constructing multiple decision trees from a randomly selected subset of data and feature subsets (bootstrap sampling), then combining the predictions from all trees to produce a more stable and accurate final output. The implementation was carried out in the Python programming language, with the assistance of the Scikit-learn library. The model parameters used were the number of estimators (n_estimators = 100) and random_state = 42 to maintain the stability and reproducibility of the experimental results.

The model creation process consists of two main stages: training using training data and predicting gold prices using test data. Furthermore, the model is tested to predict gold prices several days into the future, excluding historical data, by incorporating additional time indices as input. The predicted results are then directly compared with the actual values in the test data to assess the model's accuracy.

A comprehensive model performance evaluation is performed using regression evaluation metrics, namely (Fadly, 2025):

Mean Absolute Error (MAE):

MAE measures the average absolute error between the predicted value and the actual value. This metric assigns a uniform weight to errors regardless of the direction of the error.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} f_t + y_t$$

Mean Squared Error (MSE):

MSE calculates the average of the squared differences between predicted and actual values. This metric is more sensitive to large errors because the squared difference magnifies the weight of outliers.

$$RMSE = \frac{\sum_{i=1}^{n}(Y_i - \mathrm{Y}_i)^2}{n}$$

Root Mean Squared Error (RMSE):

RMSE is the square root of the MSE, and its units are the same as the original data. RMSE is suitable for evaluating the magnitude of the prediction error in the context of the original values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \mathrm{Y}_i)^2}{n}}$$

Coefficient of Determination (R²):

R² indicates the proportion of actual data variance successfully explained by the model. An R² value close to 1 indicates a very good model fit, while a negative value indicates the model's failure to capture data patterns.

$$R^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y}_i)^2}$$

After the evaluation process is complete and the results are analyzed, the prediction system is implemented as an interactive web application utilizing the Streamlit framework. This implementation aims to provide an easily accessible interface for users, allowing them to view gold price predictions based on specific dates and to visualize gold price trends in graphical form. This application can serve as an informational tool for investors and the general public in making investment decisions based on historical data automatically predicted by the system.

3. RESULT AND DISCUSSION

This study uses historical daily gold price data collected from Yahoo Finance over a five-year period (January 2020 – December 2024). This dataset consists of 1,257 daily data points, focusing on two key attributes: the date (Date) and the gold closing price (Close). Based on descriptive statistical analysis, the average gold price during this period was USD 1,823.26/troy ounce, with a standard deviation of USD 119.73. The highest price was recorded on March 8, 2022, while the lowest price was recorded on March 16, 2020. During the observation period, gold prices exhibited significant fluctuations, influenced by various global factors such as the COVID-19 pandemic, central bank monetary policies, the energy crisis, and global geopolitical tensions.

Pre-processing and Random Forest Modeling

The initial stage of the study involved data pre-processing to ensure the dataset was ready for use in model training. The results showed no missing values, so all data could be used directly without the need for imputation. Because the Random Forest algorithm does not recognize date formats, the date variable was transformed into a new numeric feature, the day index (Day), representing the chronological order of the data since the first observation.

The normalization process was not performed in this study, as Random Forest is not sensitive to feature scale. After the feature transformation was completed, the data was divided into two main groups: training data (80%) with 1,005 entries and testing data (20%) with 252 entries. The division was performed chronologically to maintain the integrity of the time series structure and realistically simulate future prediction scenarios.

The prediction model was developed using the Random Forest Regressor algorithm, which works by constructing a set of decision trees from a bootstrapped subset of data and random features, then combining the predictions from all these trees. This algorithm was implemented in Python, using the Scikit-learn library with default parameter settings of n_estimators = 100 and random_state = 42 for consistency of results.

3.2. Model Evaluation

The performance of the Random Forest model was evaluated using four standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²). The evaluation results showed the following model performance:

MAE = 329.31 USD

This value indicates that the average absolute error between the predicted results and the actual data is 329.31 USD. This figure is relatively large for the scale of gold prices and indicates that the model is not yet optimal in capturing accurate patterns.

MSE = 148,599.40 USD²

A high MSE indicates that there are large prediction errors in some data points, likely influenced by outliers or extreme price spikes.

RMSE = 385.49 USD

A relatively large RMSE indicates that the predicted fluctuations deviate significantly from the actual price. Because the unit is the same as the gold price (USD), this value provides a more intuitive indication that the model's predictions are not yet practically accurate.

$R^2 = -1.97$

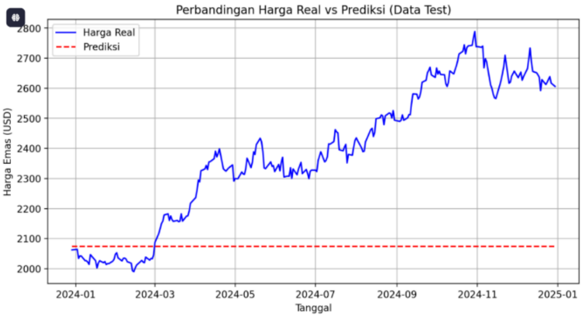

Figure 1. Gold price time series graph

A negative R-squared value indicates that the model fails to explain data variability. This means that the resulting predictions are worse than those made with simple assumptions such as using average prices. This indicates that the model is unable to effectively map the relationship between time and gold prices.

Overall, these four metrics indicate that the Random Forest model in its current configuration does not provide good predictive performance. This may be due to the lack of additional features in the dataset (e.g., global economic indicators, currency exchange rates, or interest rates) and the lack of optimization of the model parameters (hyperparameter tuning).

3.3. Prediction Results and System Implementation

Visualization of the prediction results compared to actual gold price data shows that the model is able to follow the general trend but has difficulty capturing sharp fluctuations. The prediction results graph shows that Random Forest tends to smooth out price movements, making predictions appear smooth, but is less responsive to sudden changes that are common in commodity markets.

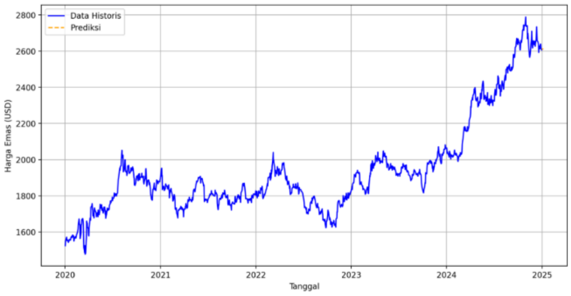

Figure 2: Gold Price Line Series Graph

A visualization of the Random Forest model's predicted results compared to the actual gold price demonstrates the model's ability to identify basic price movement patterns, albeit with several limitations:

The actual gold price graph shows fluctuating market dynamics, with a consistent upward trend from early 2024 to reach the USD 2,700-2,800 range at the end of the year. This movement reflects the characteristics of the gold market, which is influenced by various global macroeconomic factors, including inflation, monetary policy, and geopolitical uncertainty.

The comparison graph between the actual gold price (shown by the blue line) and the predictions from the Random Forest model (shown by the dashed orange line) shows a striking difference in patterns. During 2024, the actual gold price experienced a significant upward trend with sharp spikes in certain months. In contrast, the model's predictions remained flat, hovering around USD 2,080, and failed to reflect the actual price dynamics.

The model showed some accuracy midway through the year, but failed to anticipate the price spike at the end of the year. This reflects the characteristics of Random Forest, which is better suited to intermediate trends and less responsive to daily dynamics.

The negative R^2 value of -1.97 indicates that the model is not adequately capturing data variance. Therefore, method and parameter refinements are needed to improve prediction accuracy.

Nevertheless, the model is integrated into an interactive web application based on Streamlit to make it easier for users to access and understand the prediction results.

4. CONCLUSION

The Random Forest model's performance in predicting gold prices remains below expectations. The MAE of USD 329.31 and RMSE of USD 385.49 indicate a high error rate, while the R^2 of -1.97 indicates that the model is unable to outperform the simple baseline approach.

The graphical visualization results reveal that although the model is able to follow the general trend direction, the prediction results are not sufficiently responsive to market volatility, particularly in the final years of the observation period.

The Streamlit-based prediction system has been successfully implemented and provides features such as interactive graphs, comparisons of actual and predicted prices, data export to Excel, and an easy-to-use user interface.

The model's limitations lie in its limited ability to interpret short-term dynamics, its lack of adaptability to sudden changes, and its reliance on historical data without considering external factors such as global economic data.

Based on the research findings, several recommendations for further development are as follows:

First, regarding model development, improvements are recommended by adding more complex predictor variables, such as inflation rates, interest rates, and currency exchange rates. Furthermore, model parameter optimization can be performed using approaches such as Grid Search or Random Search to improve prediction accuracy. Implementing more sophisticated ensemble algorithms, such as XGBoost or LightGBM, can also be a promising alternative for improving predictive model performance.

Second, in terms of system development, consideration should be given to integrating real-time data from trusted sources to align predictions with actual market conditions. The addition of automatic notification features for significant price changes and the development of a mobile application version are also recommended to make the system more accessible to users across various platforms.

Third, for future research, it is recommended to explore deep learning-based algorithmic approaches, such as LSTM (Long Short-Term Memory), which are considered more adaptive in processing time series data. Further research could also include analyzing the influence of external factors, including global geopolitical conditions, on gold prices. Furthermore, comparative studies with other safe haven commodities such as silver or the US dollar could enrich the research context.

Fourth, in the context of practical implementation, collaboration with financial institutions or digital gold investment platforms should be considered to test the system in real-world conditions. Furthermore, the development of a system-based educational module could be useful as a financial literacy tool for novice investors.

By implementing these suggestions, it is hoped that the machine learning-based gold price prediction system can be developed into a more accurate, adaptive, and applicable tool, both in academic and industrial environments.

5. REFERENCES

Changani, J. G. (2024). Factors influencing gold price movements: a time series analysis perspective. Available at SSRN 4815102.

Fadly, H. D., & Arifin, F. (2025). Indonesian Gold Price Prediction: A Machine Learning Approach Using Random Forest Regressor.

Hutagalung, S. V., Yennimar, Y., Rumapea, E. R., Hia, M. J. G., Sembiring, T., & Manday, D. R. (2023). Comparison of support vector regression and random forest regression algorithms on gold price predictions. Jurnal Sistem Informasi dan Ilmu Komputer, 7(1), 255-262.

Kandregula, N. (2018). AI-Driven Financial Forecasting in Fintech: Enhancing Predictive Accuracy through Machine Learning and Deep Learning Models.

Landge, U., Phokmare, O., Borane, N., & Shelke, P. (2024, June). Gold price prediction using random forest algorithm. In 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 1287-1292). IEEE.

Wahyuningsih, T., Manongga, D., Sembiring, I., & Wijono, S. (2024). Comparison of effectiveness of logistic regression, naive bayes, and random forest algorithms in predicting student arguments. Procedia Computer Science, 234, 349-356.