Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 1 January 2026, pp: 186-192
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Predictive Analytics for Property Valuation Using Random Forest in Malang City

Sandrian Yulian Firmansyah Noorihsan[1*], Tintrim Dwi Ary Widhianingsih[2], Heri Kuswanto[3]

[1]Interdisciplinary School of Management and Technology,
Institut Teknologi Sepuluh Nopember, Indonesia
[2,3]Department of Statistics, Institut Teknologi Sepuluh Nopember, Indonesia

E-Mail: [1]6032241037@student.its.ac.id, [2]dwi.ary@its.ac.id, [3]heri.kuswanto@its.ac.id

Received Nov 02nd 2025; Revised Dec 12th 2025; Accepted Dec 28th 2025; Available Online Jan 17th 2026
Corresponding Author: Sandrian Yulian Firmansyah Noorihsan

Abstract

The property market in Malang City continues to expand alongside rising housing demand, yet limited price transparency still constrains informed decision-making for buyers, sellers, and developers. This study develops a data-driven property price prediction model using the Random Forest algorithm, selected for its robustness and ability to capture complex nonlinear relationships. An initial dataset of 4,358 property listings was collected through web scraping from Rumah123.com, and after thorough preprocessing including data cleaning, handling missing values, and feature refinement 1,573 valid observations remained for analysis. The model incorporates key property characteristics, covering temporal variables (month, year), physical attributes (land area, building area, number of bedrooms and bathrooms, electricity capacity, number of floors), property characteristics (certificate type, property type, property condition, furniture condition, hook position), and price information. Using optimally tuned hyperparameters, the final Random Forest model achieved an R^2 of 76.66% and a MAPE of 25.27%, indicating strong predictive performance relative to standard regression benchmarks. These findings offer managerial implications by providing objective, data-driven price estimates that can support developers, agents, and prospective buyers in pricing decisions, marketing strategies, and fair value assessments during negotiations.

Keywords: Malang City, Mean Decrease Impurity, Property Price Prediction, Random Forest, Web Scraping

1. INTRODUCTION

The property market is one of the most vital sectors of the Indonesian economy. In the first quarter of 2024, property contributed 2.43% to Indonesia's GDP and 2.33% to Indonesia's GDP in the second quarter of 2024 [1]. However, the property industry is often characterized by a lack of price transparency [2], which can hinder decision-making by buyers and investors.

The lack of price transparency underscores the urgency of increasing transparency in the property market, particularly in major Indonesian cities, such as Malang, a major migration destination. According to data from the Central Statistics Agency (BPS), from 2019 to 2023, the number of inbound migrants to Malang reached 20,000 per year [1]. This data indicates high demand for housing and property. Furthermore, the lack of transparency in property prices can influence decision-making by prospective buyers and investors. Therefore, accurate, objective, and transparent prediction models are essential to foster a transparent and fair property market for all parties.

Malang City is experiencing rapid urban development and a highly active property market, driven by its role as both an Education City and Tourism City. These conditions have resulted in strong housing demand and significant increases in land prices, where key locations such as Ijen Street and Kawi Street have reached Rp20–50 million per m² far above the official NJOP of Rp3–4 million per m² [3]. This escalation is supported by the development of new road networks, shopping centers, tourist destinations, and university expansions. Such dynamics illustrate Malang’s diverse social, geographic, and market characteristics, making the city an ideal setting for research on accurate, data-driven property price prediction.

Determining accurate property price predictions requires the use of artificial intelligence, particularly the Random Forest algorithm. Machine learning has been widely applied to support decision-making across various domains such as healthcare, manufacturing, education, finance, policy, and marketing [4]. Machine learning describes a system for solving a problem using automatically trained data [5], reinforcing its relevance for complex predictive tasks such as property valuation. Several studies have demonstrated the strong performance of this algorithm in estimating property prices. A study in 2023 compared Random Forest Regression with several other regression algorithms and showed that Random Forest achieved superior accuracy and stability in handling complex property price variables [6]. Another 2023 study also developed a Random Forest-based house price prediction model and found that it outperformed several baseline models, with location, size, and number of Bedrooms identified as the most influential variables [7]. Similar findings were reported in [8], where Random Forest was selected as the primary method for an AI-based property price prediction system due to its high accuracy and ability to capture complex data patterns.

In Indonesia, research applying the Random Forest algorithm in a local context includes a 2023 study that developed a house price prediction model for the Surabaya area using property characteristics commonly utilized by real estate agents. The findings showed that Random Forest achieved a prediction accuracy of up to 88%, making it the best-performing method compared to the other algorithms tested in the study [9]. Another study demonstrated that Random Forest was capable of predicting house prices effectively using the RapidMiner application, achieving an R^2 of 0.78. However, the high RMSE and MAE values indicated that the model still required refinement, particularly in optimizing feature selection and hyperparameters [10]. Further research compared Multiple Linear Regression and Random Forest for predicting land plot prices, revealing that Random Forest delivered superior performance with an accuracy of 81.6%. This result reinforces the advantage of Random Forest in capturing complex interactions among property variables that linear regression methods are unable to model effectively [11]. Additionally, a 2022 study compared Random Forest with Linear Regression, SVR, and Decision Trees, and found that Random Forest produced the lowest RMSE and offered the best goodness of fit. This advantage stems from its ability to manage non-linear relationships and generate more accurate property price predictions compared to the other algorithms evaluated [12].

Beyond these, other comparative studies have also validated the strong predictive capability of the Random Forest model. A study conducted a comparative analysis between Linear Regression and Random Forest for house price prediction using a large-scale Kaggle dataset, finding that Random Forest achieved a higher accuracy ($R^2 = 0.85$) than Linear Regression ($R^2 = 0.70$), emphasizing its ability to capture non-linear feature relationships [13]. Similarly, another study developed a Random Forest-based housing price prediction model for Shanghai, demonstrating superior performance compared to Decision Tree models, with lower mean absolute error (MAE) and better generalization in complex urban property markets [14]. These findings further validate the robustness and adaptability of Random Forest models for diverse real estate markets.

Taken together, these studies further confirm Random Forest as a leading algorithm for data-driven property price prediction due to its robustness, flexibility, and strong performance in handling complex variable relationships. This research aims to build an objective, data-driven artificial intelligence model to predict property prices in Malang City using property characteristics data. This research is also expected to produce an accurate and objective property price prediction model in Malang City, thereby increasing property price transparency and being used by potential buyers and investors in decision-making.

2. MATERIALS AND METHOD

2.1. Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees built using bootstrap sampling and random feature selection, enabling it to model complex and non-linear relationships with high stability [15]. In regression tasks, predictions from all trees are averaged to produce the final output. The method is known for its robustness against overfitting, tolerance to noise and outliers, and strong performance even with high-dimensional data.

Random Forest also provides feature importance metrics, including Mean Decrease Impurity (MDI), which measures how much each variable contributes to impurity reduction across the forest. This makes the algorithm particularly useful for identifying key predictors in property valuation. Its performance depends on several hyperparameters such as number of trees, maximum depth, and minimum samples per split which can be optimized using grid search and cross-validation [15]. Mathematically, the Random Forest regression prediction can be expressed as equation 1.

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

The final prediction in a Random Forest model is obtained by averaging the outputs of all decision trees within the ensemble. Each tree $h_t(x)$produces its own prediction for the input x, and these individual predictions are then averaged based on the total number of trees in the forest. Consequently, the final predicted value $\hat{y}$represents the aggregated result of all trees, a process that enhances the model's stability and accuracy compared to relying on a single decision tree.

2.2. Feature Importance

MDI is a standard metric used to measure the importance of input variables in tree-based models such as Random Forests [16]. MDI quantifies how much a feature contributes to reducing node impurity each time it is used for a split, weighted by the proportion of samples reaching that node. These contributions are then summed and averaged across all trees in the forest to obtain the final importance score. This method provides an effective way to identify the most influential predictors in a regression or classification task [16].

2.3. Model Evaluation

Model evaluation metrics are essential for assessing how closely a predictive model approximates actual values, and the choice of metric depends on research objectives and dataset characteristics [17]. In regression tasks, several standard metrics are commonly used. MAE measures the average absolute difference between predicted and actual values and is relatively robust to outliers. R-squared (R^2) indicates the proportion of variance in the target variable explained by the model, with higher values reflecting better explanatory power. Additionally, the Mean Absolute Percentage Error (MAPE) measures prediction error as a percentage, providing an intuitive interpretation of relative model accuracy [18]. The formula of MAE on equation 2, MAPE on equation 3, and R^2 on equation 4.

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

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \tag{4}$$

The MAE quantifies the average magnitude of the errors by calculating the mean of the absolute differences between the actual values and the predicted values. MAPE extends this idea by expressing the absolute error as a percentage of the actual value, making it useful for evaluating proportional accuracy across observations. Meanwhile, the coefficient of determination (R^2) assesses how well the model explains the variance in the actual data by comparing the total prediction error with the total variability of the observed values. In these definitions, y_i represents the actual value for observation i, $\hat{y}_i$ denotes the predicted value for observation i, $\bar{y}$ refers to the mean of all actual values, and nis the total number of observations.

2.4. Data Collection

This study used a web scraping method with the Selenium Web Driver framework for data collection. Based on the available data features, the data collection process was limited to the Rumah123.com website. The data collection process yielded 4,358 property data sets in Malang City. The data variables used in this study are shown in Table 1.

Table 1. Research Data

Variables Name	Data Types
month	Continuous
year	Categorical
Land Area	Continuous
Building Area	Continuous
Certificate	Categorical
Property Type	Categorical
Bedrooms	Continuous
Bathrooms	Continuous
Electricity Capacity	Categorical
Number of Floors	Continuous
Property Condition	Categorical
Furniture Condition	Categorical
Hook	Categorical
Price	Continuous

2.5. Data Preprocessing

Data preprocessing was conducted to resolve inconsistencies and enhance the overall quality of the dataset prior to model development. Several issues were identified in the raw data, including incorrect district labels and substantial missing information in key property attributes such as the number of floors, furnishing condition, hook position, certificate type, land area, building area, number of bedrooms, number of bathrooms, electrical capacity, and property condition. District-related inconsistencies were corrected through manual verification using Google Maps, ensuring that all retained entries were genuinely located within Malang City. Missing values were manually supplemented by reviewing the listing photos, while variables irrelevant to land-only properties were assigned appropriate categorical labels. Entries lacking essential attributes that could not be reliably verified were removed to preserve dataset validity. Before proceeding to model training, numerical outliers particularly in price and key physical features were identified and filtered using the Interquartile Range (IQR) method to produce a more balanced distribution and enhance model stability. The deletion method was deemed appropriate because it helps preserve the integrity of the remaining data, especially when the removed observations do not exhibit systematic missing patterns and the overall dataset size remains sufficient for analysis [19].

Following the cleaning and refinement steps, several data transformation procedures were applied. Ordinal Encoding was used for variables with inherent ordering, including electrical capacity, property condition, and furnishing condition, while One-Hot Encoding was applied to nominal variables such as district, certificate type, property type, and hook position. The dataset was then partitioned into training and test sets using three different proportions (90/10, 80/20, and 70/30) by adjusting the test_size parameter in train_test_split to evaluate model robustness across data splits. A logarithmic transformation was applied to the target variable to reduce skewness, and predictor scaling was performed after the data splitting process to avoid data leakage. The log transformation method was selected because it is widely recognized for reducing skewness in positively skewed distributions, thereby producing a distribution that more closely approximates normality, which in turn improves model stability and interpretability [20]. The resulting changes in skewness values across numerical variables are presented in Table 2, demonstrating the effectiveness of the preprocessing steps in producing a more suitable dataset for model development.

2.6. Training Model

The model development process consisted of several critical stages: hyperparameter tuning, feature importance assessment, feature selection, and construction of the final Random Forest model. Hyperparameter optimization was conducted using Grid Search Cross-Validation (GridSearchCV) with a 10-fold validation scheme to ensure robust generalization. The tuning procedure systematically evaluated multiple combinations of parameter values, including the number of estimators set to 180, 200, 220, and 240; maximum tree depth tested at 5, 10, 15, and 20; minimum samples split examined at values of 1, 5, 7, and 9; and minimum samples per leaf assessed at 1, 2, 3, and 4. Additionally, the model explored variations of maximum features at ratios of 0.6, 0.7, and 0.8, along with bootstrap conditions True. These configurations were evaluated using the Negative Mean Absolute Error metric to determine the most effective parameter set for accurate price prediction.

After hyperparameter tuning, feature importance was assessed using the MDI method to identify the predictors that contributed most significantly to the model. Variables with importance scores above 0.01 were retained as key features, while all district-related variables were preserved due to the one-hot encoding structure generated during preprocessing. The final Random Forest model was then constructed using the optimal hyperparameters identified through GridSearchCV and trained on the selected high-importance features. This approach ensured that the model relied only on the most influential attributes, enhancing predictive accuracy, model stability, and interpretability in capturing property price patterns.

Table 2. Example Of The Results Of Transforming Several Variables Using Log Transform

Variables Name	Skewness Before Transformation	Skewness After Transformation
month	-0.69	-1.02
year	0	0
Land Area	3.32	-1.27
Building Area	15.45	0.14
Bedrooms	2.49	-0.94
Bathrooms	3.65	-0.49
Number of Floors	1.03	-0.73
price	29.15	-0.95

3. RESULTS AND DISCUSSION

3.1. Prediciton Results

Based on the feature importance analysis using MDI, the variables that most influence property prices are land area (0.479) and building area (0.250), followed by other variables such as bathroom (0.048768), bedroom (0.047671), Sukun_district (0.026268), Blimbing_district (0.020165), month (0.019067), Kedungkandang_district (0.018946), number of floors (0.015095), Lowokwaru_district (0.014919), property condition (0.012602), and Klojen_district (0.005503). After all preprocessing steps, including handling missing values, manual correction of categorical fields, removal of invalid records, outlier filtering, encoding, scaling, and feature selection, a total of 1,573 cleaned observations were used for model training and evaluation. Although approximately 64% of the initial 4,358 records were removed due to incomplete or inconsistent entries, exploratory checks indicated that the missing data did not follow any specific pattern related to property characteristics such as price range, location, or size. Therefore, the data loss is considered random rather than systematic, and it is unlikely to introduce bias or distort the representativeness of the remaining dataset.

Several experimental configurations were evaluated to obtain the Random Forest model after feature selection with the best MAE, MAPE, and coefficient of determination (R^2). The optimal results were achieved through outlier removal and an 80:20 train–test split, producing an MAE of 367,907,294.81, a MAPE of 25.27%, and an R^2 of 76.66%.

The final model was trained using the best-performing hyperparameters, namely number of estimators = 200, minimum samples split = 5, minimum samples leaf = 1, maximum depth = 20, maximum features = 0.7, and bootstrap = False. An R^2 value of 76.66% indicates that the model explains more than three-quarters of the variation in property prices. This level of performance reflects the moderate influence of key physical attributes especially land area and building area along with structural, temporal, and spatial features represented by district indicators. These combined features enable the model to effectively capture price variation across Malang City. Table 3 summarizes the experimental scenarios and their results before feature selection.

Table 3. Evaluation Results of Preprocessing Combination on Prediction Accuracy Before Feature Selection

Outliers Handling	Train/Test Split	Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)	R^2
Deletion	70:30	407,526,756.48	0.2706	0.7197
	80:20	376,777,158.83	0.2594	0.7535
	90:10	368,592,464.51	0.2627	0.7216
Median Imputation	70:30	489,570,877.66	0.7074	0.6330
	80:20	437,282,169.02	1.2760	0.6778
	90:10	425,087,341.24	2.5766	0.6980

3.2. Prototype Implementation

The prototype aims to interactively visualize the system's workflow and provide an overview of how end users will interact with the prediction model. The implementation uses the Streamlit library to create the website's interface and then uses Streamlit Cloud to facilitate widespread use. The prototype can be accessed at https://bit.ly/MalangHousePricePredictionApp. The prototype's appearance can be seen in Figure 1.

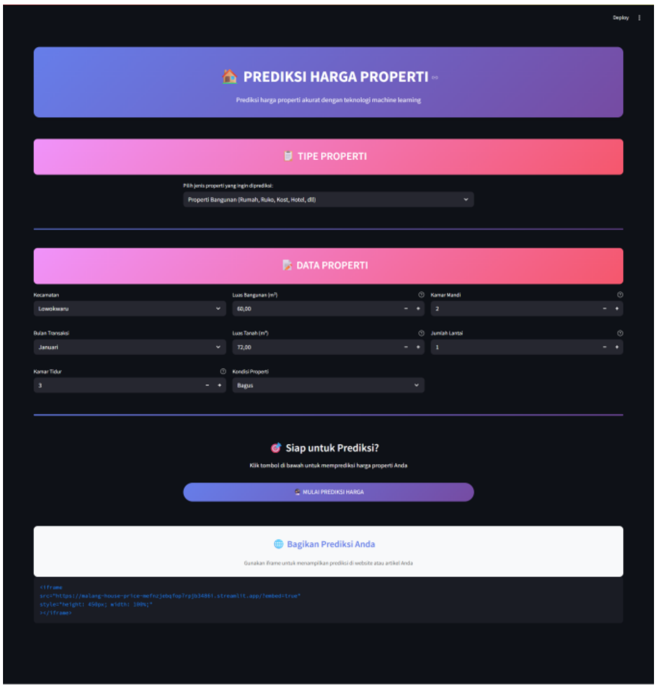

Figure 1. Prototype Implementation

The application provides several key features, including a price prediction tool where users begin by selecting the property type, land or building, followed by entering the relevant variables identified in this study, such as land area, building area, number of bedrooms, number of bathrooms, property condition, and district. The system then displays the predicted property price based on the trained Random Forest model. Additionally, the application offers an embed link feature, allowing users to copy and integrate the prediction widget into external websites for broader practical use.

3.3. Managerial Implications

The managerial implications of this study are based on the model's performance in predicting property prices in Malang City, the results of the MDI analysis showing the most influential variables, and a prediction prototype that can be practically used by users. These findings serve as a basis for decision-making for developers, individual sellers or property agents, and prospective buyers.

1. Implications for Property Developers: The prediction prototype serves as a data-driven reference for setting property prices more transparently. Based on the MDI results, land area (0.479) and building area (0.249) are identified as the strongest predictors, enabling developers to prioritize product planning around these high-impact characteristics.
2. Implications for Individual Sellers or Property Agents: For sellers and agents, the pricing tool supports more objective and market-aligned price decisions. Insights from MDI also help guide marketing strategies by highlighting key features such as land area, building area, number of Bathrooms, and number of Bedrooms.
3. Implications for Prospective Buyers: For buyers, the model's price predictions provide an objective benchmark to assess price fairness and strengthen negotiation positions. This information enables more informed and accurate purchasing decisions.

3.4. Discussion

The results of this study confirm the capability of the Random Forest algorithm in modeling property prices, particularly in complex urban environments such as Malang City. The feature importance analysis indicates that land area and building area are the most influential predictors, which aligns closely with findings from previous studies. Research conducted by Chhiller et al. [2] and Sri et al. [5] similarly identified physical property attributes, especially land size and building size, as dominant factors in determining housing prices. In addition, studies by Ratih et al. [7] and Fang et al. [10] also emphasized the significance of structural features such as the number of bedrooms and bathrooms, reinforcing the consistency of this study's findings with established empirical evidence. This alignment suggests that Random Forest reliably captures fundamental value drivers in residential property markets across different geographic contexts.

In terms of predictive performance, the Random Forest model developed in this study achieved an R^2 of 76.66% and a MAPE of 25.27%, indicating a moderate level of accuracy in explaining property price variation. These results are comparable to those reported in earlier studies, although variations can be observed due to differences in data sources, sample sizes, and market characteristics. For instance, Ratih et al. [7] reported higher predictive accuracy (up to 88%) in the Surabaya market, while Fang et al. [10] and Pebriadi and Fitria [8] reported R^2 values ranging from 0.78 to 0.85. The slightly lower performance observed in this study may be attributed to the heterogeneous nature of the Malang property market, which encompasses diverse district characteristics, land-use patterns, and price disparities. Nevertheless, the achieved performance remains within an acceptable range and demonstrates the robustness of Random Forest in handling non-linear relationships and high-dimensional feature spaces.

Furthermore, the successful implementation of feature selection using MDI strengthens the interpretability of the Random Forest model, addressing a common criticism of ensemble-based methods. Similar approaches have been adopted in previous research to identify key drivers of property prices while maintaining predictive performance [12]. The findings of this study validate that combining Random Forest with rigorous preprocessing, outlier handling, and feature selection can produce a stable and interpretable prediction model suitable for practical applications. Compared with linear regression approaches, which often struggle to capture complex interactions among property attributes [11], Random Forest demonstrates superior flexibility and generalization capability. Overall, this study extends prior research by providing empirical evidence from Malang City, reinforcing the applicability of Random Forest as a reliable and data-driven tool for improving price transparency and supporting decision-making in local property markets.

4. CONCLUSION

This study addressed the issue of limited price transparency in Malang's rapidly growing property market by developing a data-driven prediction model using the Random Forest algorithm. After performing extensive preprocessing including handling missing values, removing outliers, encoding categorical variables, and selecting the most relevant features the model was trained using the optimal hyperparameters identified through tuning. The analysis showed that land area, building area, and key structural and locational attributes were the most influential predictors of property prices. With this configuration, the Random Forest model achieved strong predictive performance, indicating its effectiveness in capturing the underlying patterns of property valuation in Malang City.

REFERENCES

[1] Badan Pusat Statistik, *Produk Domestik Bruto Indonesia Triwulanan: Quarterly Gross Domestic Product of Indonesia 2020–2024*, vol. 7. Jakarta, Indonesia: BPS, 2024.

[2] M. Chhiller, S. Shivam, and R. Kumar, "Propertie Price Prediction Using Machine Learning," *Int. J. Sci. Res. Eng. Manag.*, vol. 7, no. 11, pp. 1–11, 2023, doi: 10.55041/ijsrem26731.

[3] D. Widyanto, "The Highest and Best Use Analysis and Feasibility Study of Residential Housing in Malang," *Adv. Civ. Eng. Sustain. Archit.*, vol. 1, no. 1, 2018, doi: 10.9744/acesa.v1i1.6815.

[4] R. Mitchell, *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, n.d., p. 238.

[5] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Markets*, vol. 31, no. 3, 2021, doi: 10.1007/s12525-021-00475-2.

[6] O. Lohith, A. Jha, and S. C. Tamboli, "Comparative Analysis of Random Forest Regression for House Price Prediction," *Int. J. Creative Res. Thoughts (IJCRT)*, pp. 2320–2882, 2023. [Online]. Available: www.ijcrt.org.

[7] B. U. Sri, C. S. K. Reddy, C. R. Kumar, A. Vyshnavi, B. Vinod, and B. K. Reddy, "Random Forest-Based House Price Prediction," in *Proc. 2nd Int. Conf. Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, 2023, pp. 954–959, doi: 10.1109/ICAISS58487.2023.10250452.

[8] B. Lekkihal, V. Kumar, B. Shankar, and N. Kalyani, "Real Estate Price Prediction Using Artificial Intelligence," *Int. Res. J. Comput. Sci.*, 2023, doi: 10.26562/irjcs.2023.v1005.29.

[9] T. Ratih, R. Tanamal, T. Wiradinata, Y. Soekamto, and N. Minoque, "House Price Prediction Model Using Random Forest in Surabaya City," *TEM J.*, 2023, doi: 10.18421/tem121-17.

[10] M. Pebriadi and F. Fitria, "House Price Prediction Using the Random Forest Algorithm on the RapidMiner Application," *Formosa J. Sci. Technol.*, 2025, doi: 10.55927/fjst.v4i2.20.

[11] G. Aryono, S. Auliana, and B. Permana, "Comparative Analysis of Multiple–Linear Regression Algorithm with Random Forest Regression for Prediction of House Plot Prices," *J. Ilmiah Glob. Educ.*, 2024, doi: 10.55681/jige.v5i2.2794.

[12] Y. Fang, T. Li, and H. Zhao, "Random Forest Model for the House Price Forecasting," in *Proc. 14th Int. Conf. Computer Research and Development (ICCRD)*, Shenzhen, China, 2022, pp. 140–143, doi: 10.1109/ICCRD54409.2022.9730190.

[13] Y. Fu, "A Comparative Study of House Price Prediction Using Linear Regression and Random Forest Models," *Highlights in Science, Engineering and Technology*, vol. 107, AMMSAC 2024, pp. 96–103, 2024, doi: 10.54097/vcy5n584.

[14] S. Wu, "Shanghai House Price Prediction Using Random Forest," in *Proc. 3rd Int. Conf. Business and Policy Studies*, 2024, doi: 10.54254/2754-1169/66/20241234.

[15] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, pp. 69–79, 2024, doi: 10.58496/BJML/2024/007.

[16] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in Neural Information Processing Systems 26*, 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/e3796ae838835da0b6f6ea37bcf8bcb7-Paper.pdf

[17] P. Dumre, S. Bhattarai, and H. Shashikala, "Optimizing Linear Regression Models: A Comparative Study of Error Metrics," in *Proc. 4th Int. Conf. Technological Advancements in Computational Sciences (ICTACS)*, 2024, pp. 1856–1861, doi: 10.1109/ICTACS62700.2024.10840719.

[18] A. De Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for Regression Models," *Neurocomputing*, vol. 192, pp. 38–48, 2016, doi: 10.1016/j.neucom.2015.12.114.

[19] A. Atoum, "The Critical Role of Evaluation Metrics in Handling Missing Data: A Review of Modern Approaches," *Int. J. Adv. Appl. Sci.*, vol. 12, no. 1, pp. 1–14, 2025. [Online]. Available: https://www.science-gate.com/IJAAS/Articles/2025/2025-12-01/1021833ijaas202501011.pdf

[20] R. West, "Best Practice in Statistics: The Use of Log Transformation," *Ann. Clin. Biochem.*, vol. 59, pp. 162–165, 2021, doi: 10.1177/00045632211050531.