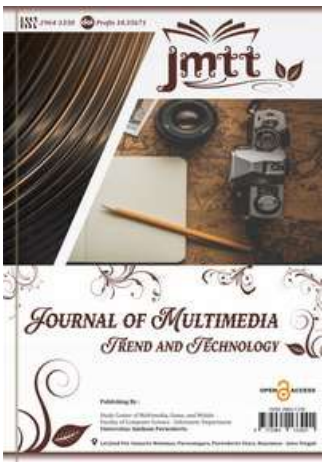
Journal of Multimedia Trend and Technology - JMTT
Volume 4, Issue 2, August 2025, ISSN 2964-1330
https://journal.educollabs.org/index.php/jmtt/

Predicting Greater Jakarta Area House Prices Using Random Forest and Linear Regression

Firli Firmansyah Agustin [1*], Fariz Nur Fikri Zaki [2]

[1,2] Department of Information System, Faculty of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia.

ARTICLE INFO



History :

Submit on 1 May 2025
Review on 12 June 2025
Accepted on 6 July 2025

Keyword :

Price Prediction;
Machine Learning;
Random Forest;
Multiple Linear;

ABSTRACT

This study focuses on analyzing and predicting house prices in the Greater Jakarta Area using a machine learning approach, specifically comparing the performance of random forest regression and multiple linear regression. the increasing demand for adequate housing in Greater Jakarta Area, coupled with fluctuating house prices influenced by factors like land size, building size, number of bedrooms, bathrooms, and other facilities, necessitates an accurate price prediction system to assist both the public and businesses in decision-making. data was collected from Rumah123.com via Kaggle, followed by pre-processing and exploratory data analysis (EDA). the models were built using both algorithms and evaluated through 10-fold cross-validation, with an 80% training and 20% testing data split. the results demonstrate that random forest regression outperforms multiple linear regression, achieving a correlation coefficient of 0.5043 and a mean absolute error of 157,698,532. in contrast, multiple linear regression (m5p) yielded a correlation coefficient of 0.4895 and a mean absolute error of 209,890,933. therefore, random forest regression is recommended as a superior model for house price prediction in the Greater Jakarta Area region.

***Corresponding Author:**

Firli Firmansyah Agustin
Department of Information Systems, Universitas Amikom Purwokerto, Purwokerto, Indonesia.
Email: firmansyahfirli93@gmail.com

INTRODUCTION

The property sector in the Greater Jakarta (Jakarta, Bogor, Depok, Tangerang, and Bekasi) area continues to experience rapid growth in line with increasing urbanization and its status as a national economic center. However, this growth is accompanied by complexities in determining house prices, which are influenced by various multivariate factors, ranging from strategic location and access to public transportation such as the LRT and MRT, to nearby supporting facilities. This price uncertainty often creates information asymmetry between sellers and buyers, necessitating the need for more accurate and objective valuation instruments.

Traditional property valuation methods are often considered inefficient in handling large volumes of data and non-linear relationships between variables. In recent years, literature has shown that data-driven approaches using machine learning algorithms can provide more precise predictions than manual estimation. The use of historical transaction data and physical building specifications allows for modeling that can capture hidden patterns behind the dynamic fluctuations in property market prices in the capital's buffer zone.

In this study, two popular algorithms were used as comparative methods: Linear Regression and Random Forest. Linear Regression was chosen because of its ability to provide clear interpretations of linear relationships between variables. On the other hand, Random Forest, an ensemble learning-based algorithm, is used to handle the data complexity and non-linear interactions often found in property data, while simultaneously mitigating the risk of overfitting that can occur with a single model.

The main objective of this study is to evaluate the performance of the two models in predicting house prices in the Greater Jakarta area. By comparing evaluation metrics such as Mean Absolute Error (MAE) or R-squared, this study is expected to provide recommendations on which model is most reliable for investors, developers, and the general public. The results of these predictions are expected to guide smarter financial decision-making in the competitive property market.

Over time, people's needs have continued to evolve, one of which is the need for housing. Property developers compete to build or purchase homes as an investment asset. This situation encourages potential buyers to carefully consider whether a house provides sufficient financial value, as property prices tend to increase continually [1].

Prediction is a technique used to estimate future values based on past or current data. Accurate prediction capabilities enable organizations and institutions to make informed decisions. Previous research titled "Prediction of House Prices in East Bandung Using the Moving Average Algorithm" demonstrated that the moving average model achieved an accuracy rate of 70–90% with an error rate between 10–30% [2][3].

The rising trend of house prices each year can be analyzed based on property attributes. Because prices are volatile and difficult to predict accurately, potential buyers need systems capable of estimating house prices based on these features. This study applies regression algorithms Multiple Linear Regression and Random Forest Regression to build predictive machine learning models that estimate property prices in Greater Jakarta Area [4][5].

Based on the market complexity described above, this study attempts to address the key question of how property variables, such as location, land area, and supporting facilities, influence house price formation in the Greater Jakarta area. The primary focus of the problem lies in finding the most accurate prediction model amidst dynamic price fluctuations, by questioning whether a simple linear approach can outperform more complex ensemble-based algorithms. Therefore, this study aims to identify key price determinants while empirically testing the performance of the Linear Regression and Random Forest algorithms. By comparing evaluation metrics such as R-Squared and Mean Absolute Error (MAE) of the two models, this study aims to determine the most reliable predictive model for stakeholders in decision-making in the property sector.

METHOD

A. Data Mining

A data mining approach was used to predict property prices across various geographic areas by considering variables such as location, size, and facilities. Data mining– based regression and classification techniques identify hidden patterns among property attributes that significantly affect prices. The results show that combining spatial and non- spatial variables improves model performance [6].

B. Software: Weka 3.9.6

The Waikato Environment for Knowledge Analysis (Weka) is a software tool that applies machine learning algorithms for data exploration and analysis.

C. Multiple Linear Regression

Multiple Linear Regression (MLR) extends the simple linear regression model [7]. While simple regression involves one independent variable, MLR includes multiple predictors, expressed as (1):

$$Y = \alpha + \beta 1 X 1 + \beta 2 X 2 + ... + \beta n X n + e \quad (1)$$

where:
Y = dependent variable (response)
X = independent variable (predictor) α = constant
β = slope or coefficient estimate

D. Random Forest Regression

A Random Forest consists of multiple decision tree models, applying the methods of bootstrap aggregating (bagging) and random feature selection for regression and classification. A decision tree is a flowchart-like structure designed as a tree, with a root node used for collecting information. The decision tree classifies data samples with unknown classes into predefined categories [8]. The objective of a decision tree is to avoid overfitting the dataset while achieving maximum accuracy.

A Random Forest is an ensemble of decision trees that are combined into a single model. Its performance relies on a random vector value that is uniformly distributed across all trees, with each decision tree built to a maximum depth. A Random Forest is a tree-based classifier {h(x, θk), k = 1, . . .}, where {θk} are independently distributed random vectors, and each tree casts a vote for the most popular class for a given input x [9].

E. Hybrid Machine Learning for House Price Prediction

Hybrid machine learning combines several algorithms such as Random Forest, Gradient Boosting, and Support Vector Regression to enhance predictive accuracy. This approach captures complex nonlinear relationships between features, yielding more accurate house price predictions [10].

F. Evaluation of Random Forest and XGBoost Algorithms

The Random Forest and XGBoost algorithms are widely used in house price prediction due to their capability to handle complex and non-linear data [11]. Random Forest operates using an ensemble technique that constructs multiple decision trees through the bootstrap aggregating (bagging) process. Meanwhile, XGBoost implements a more efficient and faster boosting method. Research indicates that ensemble models like XGBoost exhibit superior accuracy compared to simple linear regression [9].

G. Comparison of Regression Models of Property Prediction

Several regression models, such as Linear Regression, Decision Tree, and K-Nearest Neighbors, are comparatively evaluated in the context of property price prediction. Each algorithm possesses distinct advantages and disadvantages depending on the data's structure and distribution. This research concludes that no single model is superior for all conditions; therefore, the choice of algorithm must be tailored to the specific characteristics of the dataset and the final objective [12].

H. *Research Stages*

1. Preliminary Steps

a) *Observation:* Observation is a research activity involving the collection of information related to the research problem, sourced from public data obtained from Kaggle.com.
b) *Literature Review:* This method involves gathering credible references required for the research report.
c) *Documentation:* Literature has long been utilized as an information source in research for analysis, interpretation, and even prediction.

2. *Exploratory Data Analysis (EDA):* Exploratory Data Analysis (EDA) is the initial investigation process used to identify patterns, discover outliers, test hypotheses, and verify assumptions. EDA is highly beneficial for early error detection, as it allows users to find anomalies, understand relationships within the data, and extract significant factors. This process is instrumental in statistical analysis. In exploratory data analysis, several techniques can be employed for data processing [13].
3. *Pre-processing:* Pre-processing involves the treatment of data to correct or clean incorrect entries, thereby making the data usable. In this stage, the initial raw data is processed to fit the requirements of the analysis [14].
4. *Exploratory Data Analysis:* Exploratory Data Analysis is the method of examining the available data to determine how to process it. This stage includes checking for null values, removing duplicate data, and converting data categories.
5. *Representation:* In this stage, the raw data is transformed into a representation or visualization. The data can be visualized in forms such as boxplots, histograms, and other graphical representations.
6. *Modeling:* Modeling is the application of an algorithmic model where data is further processed to derive conclusions. This process will yield different results depending on the characteristics of the data [15].
7. *Deployment/Evaluation:* In this stage, conclusions are drawn from the data mining results. The final conclusion is based on the various hypotheses generated from the data mining process [6].

RESULT AND DISCUSSION

This house price prediction study employs the Exploratory Data Analysis (EDA) method. Exploratory Data Analysis (EDA) is a process of analyzing a dataset to summarize its primary characteristics and gain an understanding of its overall condition. Typically, EDA utilizes visualization methods such as Histograms, Box Plots, and Violin Plots. The research process, utilizing Exploratory Data Analysis (EDA), is outlined below [8].

A. Data Pre-processing

In this stage, the house price dataset is filtered. This selection process serves to remove irrelevant or unused data. The data to be discarded includes invalid, inconsistent, and null values.

Table 1. Sample of the dataset.

url	price_in_r	title	address	district
https://ww	12700000	Rumah Ke	Summarec	Summarec
https://ww	19500000	Rumah Ca	Summarec	Summarec
https://ww	33000000	Rumah Me	Summarec	Summarec
https://ww	45000000	Rumah Ho	Summarec	Summarec

ads_id	bedrooms	bathrooms	land_size_m	building_size
hos106803	03:00	02:00	55:00:00	69:00:00
hos106858	03:00	03:00	119:00:00	131:00:00

hos109277	03:00	03:00	180:00:00	174:00:00
hos107855	04:00	03:00	328:00:00	196:00:00

maid_bath	floors	building_a	year_built	property_c
00:00	02:00			bagus
01:00	02:00			bagus
01:00	02:00	06:00	2016:00:00	bagus seka
01:00	02:00	09:00	2013:00:00	bagus

[a] The table above presents the dataset after it has been parsed into distinct columns; however, it still contains several null values and unnecessary columns.

B. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the process of examining the available data to determine how it should be processed. This stage involves checking for null values and removing superfluous columns. Once the data has been cleansed of erroneous entries, descriptive statistics are subsequently applied to summarize and describe the data as follows.

Table 2. Sample Dataset.

price_in_r	city	bedroom	bathroom	land_size	building_s
45000000	Bekasi	4.0	3.0	328.0	196.0
27000000	Bekasi	3.0	3.0	136.0	200.0
27000000	Bekasi	3.0	3.0	136.0	200.0

t	maid_bed	maid_bat	floors	garages	furnishing
2200 mah	1.0	1.0	2.0	1.0	unfurnish
3500 mah	1.0	1.0	2.0	2.0	unfurnish
3500 mah	1.0	1.0	2.0	1.0	unfurnish

[b] The table above presents the cleaned dataset, which is ready for conversion into the .arff format readable by the Weka software.

C. Representation

In this stage, a visualization of the house price data is performed, allowing for the observation of its distribution.

Figure 1. Pilot Visualitation

Figure 1 illustrates the relationships between features within the property dataset, such as price (price_id_), land size (land_size_m2), building size (building_size_m2), number of rooms (bedrooms, bathrooms), and other attributes like city, certificate, and furnishing. In general, a pattern emerges indicating a positive correlation between building and land size with property price. This implies that properties with larger land and building sizes tend to have higher prices, although this relationship is not entirely linear due to significant price variations within similar size categories.

The relationship between land size and building size also appears to be quite strong. The majority of properties with large building sizes also have large land areas, though there are a few cases where large buildings are situated on relatively small plots of land, which could indicate townhouses or homes with extensive renovations. Furthermore, the 'city' variable appears to form distinct clusters in relation to price and property size, signifying that geographical location is a critical factor in determining property value.

Other features—such as the number of bedrooms, bathrooms, maid rooms, floors, carports, garages, and electricity capacity—show an influence on price, but their impact is relatively weaker compared to land size, building size, and location. Specifically, the influence of 'certificate' and 'furnishing' on price appears to be very limited and does not form a strong pattern. This suggests that aspects of legality and the condition of the home's contents are not primary drivers of price variation in this dataset.

Overall, the factors that most strongly determine property prices in this dataset are building size, land size, and location (city). Factors such as the number of rooms and electricity capacity have a moderate level of influence, while other factors like carports, garages, maid rooms, and furnishing exhibit a weak influence. Additionally, it should be noted that this dataset contains a significant number of outliers, such as small houses with very high prices or large houses with low prices. During the Data Preprocessing stage, extreme outliers and invalid data were identified and removed to ensure data quality. Nevertheless, for future model accuracy improvements, more advanced outlier handling processes (e.g., using statistical methods or data transformations like log-transformation) and data normalization will be crucial steps..

D. RandomForest

At this stage, an analysis is conducted using RandomForest with 10-fold cross-validation.

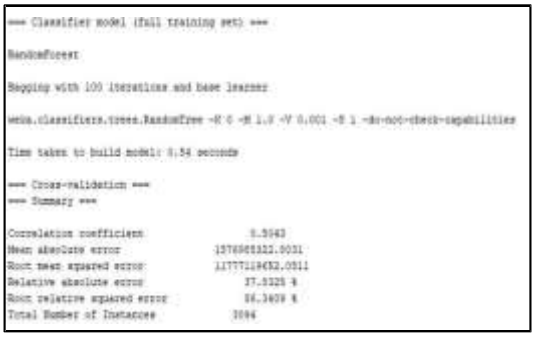


Figure 2. Result of the RandomForest Algorithm

From the 10-fold cross-validation, the Random Forest algorithm yields a correlation coefficient of 0.5043. This value indicates a moderate positive correlation between the predicted and actual prices, suggesting that the model is reasonably capable of capturing some of the relationship patterns between the features and the target price. The resulting mean absolute error (MAE) is approximately IDR 157,698,532. This MAE figure represents the average absolute difference between the predicted and actual prices. In the context of property prices in Greater Jakarta Area, which can reach billions of rupiah, an MAE of IDR 157 million indicates a fairly good level of accuracy, yet leaves significant room for improvement, especially for high-value investment decisions. The relative absolute error was recorded at 37.83%. The model's training time was also efficient, at only 0.54 seconds.

E. M5P

At this stage, an analysis is conducted using M5P.

```
=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)
LM1 (3094/85.391%)

LM num: 1
price_in_rp =
        1787135724.5189 * city=Tangerang,Jakarta_Barat,Jakarta_Utara,Jakarta_Selatan,Jakarta_Pusat
        + 2332434053.6246 * city=Jakarta_Selatan,Jakarta_Pusat
        + 4592110877.2373 * city=Jakarta_Pusat
        - 397845253.4576 * bedrooms
        + 2380964.5224 * land_size_m2
        + 19170068.7466 * building_size_m2
        + 346917956.8739 * carports
        + 2960836524.6365 * electricity=47500_mah,lainnya_mah,17600_mah,5500_mah,7700_mah,8000_mah,66
        - 2655300549.1093 * electricity=7700_mah,8000_mah,6600_mah,12700_mah,9500_mah,13300_mah,7600_
        - 3911298553.6885 * electricity=12700_mah,9500_mah,13300_mah,7600_mah,10600_mah,13200_mah,100
        + 5384156086.3874 * electricity=9500_mah,13300_mah,7600_mah,10600_mah,13200_mah,10000_mah,110
        + 2916139640.3799 * electricity=13200_mah,10000_mah,11000_mah,16500_mah,13900_mah,41500_mah,5
        + 13338487185.5584 * electricity=16500_mah,13900_mah,41500_mah,53000_mah,23000_mah,24000_mah,
        + 630820738.6096 * maid_bedrooms
        + 1072578289.2814 * maid_bathrooms
        - 1138561170.6691 * floors
        - 943447125.5449 * furnishing=semi_furnished,furnished
        + 997923122.0183

Number of Rules : 1

Time taken to build model: 0.49 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.4895
Mean absolute error               2098099331.2118
Root mean squared error           11907407492.4667
Relative absolute error                50.3342 %
Root relative squared error            87.296  %
Total Number of Instances            3094
```

Figure 3. Result of the M5P Algorithm

From the analysis using M5P, a correlation coefficient of 0.4895 was obtained, slightly lower than that of the Random Forest. The mean absolute error (MAE) was also larger, at approximately IDR 209,890,933, with a relative absolute error of 50.34%. This indicates that, in general, the predictive performance of the M5P model is less accurate than that of the Random Forest for this dataset, with a larger average prediction deviation. Nevertheless, the training time for M5P was also fast, at 0.49 seconds.

CONCLUTIONS

This research aims to analyze and predict house prices in the Greater Jakarta Area region using a machine learning approach with two regression algorithms: Multiple Linear Regression (M5P) and Random Forest Regression. Data was sourced from the website Rumah123.com, followed by data cleaning, exploration, and visualization to understand the dataset's characteristics. The primary factors influencing house prices include land area, building size, number of bedrooms and bathrooms, and location.

From the results of testing using 10-fold cross-validation, it was found that the Random Forest Regression algorithm demonstrated the best performance, with a correlation coefficient of 0.5043 and a mean absolute error of approximately 157 million. Meanwhile, the M5P algorithm produced a correlation coefficient of 0.4895 and a mean absolute error of 209 million. These results indicate that Random Forest is better able to capture the non-linear relationships between features in the property data compared to M5P, which forms a primary linear model.

Overall, Random Forest Regression is recommended as a more effective and efficient algorithm for house price prediction in the Greater Jakarta Area region, although further development is needed to improve accuracy, particularly in handling outliers and other complex variables.

REFERENCE

[1] Dhiwa Aqsha, "A Comparative Analysis of Extreme Gradient Boosting and Random Forest Algorithms for House Price Prediction in the Greater Jakarta Area," J. Ilmu Komput. dan Sist. Inf., vol. 13, no. 1, Jan. 2025, doi: 10.24912/jiksi.v13i1.32863.

[2] L. El Mouna, H. Silkan, Y. Haynf, M. F. Nann, and S. C. K. Tekouabou, "A Comparative Study of Urban House Price Prediction using Machine Learning Algorithms," E3S Web Conf., vol. 418, p. 03001, Aug. 2023, doi: 10.1051/e3sconf/202341803001.

[3] M. N. Hibatulloh, G. D. Prakoso, A. D. Putri Yunus, and T. D. Putra, "Predicting House Prices in Bandung 2024 Using Ensemble Learning: A Comparative and Interpretability Analysis," J. Inform. J. Pengemb. IT, vol. 10, no. 2, pp. 484–493, Apr. 2025, doi: 10.30591/jpit.v10i2.8200.

[4] L. Matic and Z. Kalinić, "HOUSING PRICE PREDICTION USING XGBOOST AND RANDOM FOREST METHODS," in ZBORNIK RADOVA, Faculty of Economics, Kragujevac, 2025, pp. 417–424. doi: 10.46793/EBM24.417M.

[5] C. Zou, "The House Price Prediction Using Machine Learning Algorithm: The Case of Jinan, China," Highlights Sci. Eng. Technol., vol. 39, pp. 327–333, Apr. 2023, doi: 10.54097/hset.v39i.6549.

[6] P. Mahajan, M. Gawade, A. Patel, S. Barhanpurkar, and O. Deshmukh, "House Price Prediction and Recommendation," Int. J. Res. Appl. Sci. Eng. Technol., vol. 11, no. 4, pp. 2009–2013, Apr. 2023, doi: 10.22214/ijraset.2023.49350.

[7] I. R. Ningsih, A. Faqih, and A. R. Rinaldi, "House Price Prediction Analysis Using a Comparison of Machine Learning Algorithms in the Greater Jakarta Area Area," J. Artif. Intell. Eng. Appl., vol. 4, no. 2, pp. 687–694, Feb. 2025, doi: 10.59934/jaiea.v4i2.733.

[8] D. J. C. Sihombing, "Application of Feature Engineering Techniques and Machine Learning Algorithms for Property Price Prediction," JITSI J. Ilm. Teknol. Sist. Inf., vol. 5, no. 2, pp. 72–76, Jun. 2024, doi: 10.62527/jitsi.5.2.241.

[9] C. Li, "House price prediction using machine learning," Appl. Comput. Eng., vol. 53, no. 1, pp. 225–237, Mar. 2024, doi: 10.54254/2755-2721/53/20241426.

[10] L. Fang, "Machine learning models for house price prediction," Appl. Comput. Eng., vol. 4, no. 1, pp. 409–415, May 2023, doi: 10.54254/2755-2721/4/20230505.

[11] H. Sharma, H. Harsora, and B. Ogunleye, "An Optimal House Price Prediction Algorithm: XGBoost," Analytics, vol. 3, no. 1, pp. 30–45, Jan.

[12] 2024, doi: 10.3390/analytics3010003.

[13] Nadia Putri Ariyanti, Agung Triayudi, and Ratih Titi Komala Sari, "Analysis of K-NN Algorithm and Linear Regression to Predict House Prices in Greater Jakarta Area," SaNa J. Blockchain, NFTs Metaverse Technol., vol. 2, no. 1, pp. 65–71, Feb. 2024, doi: 10.58905/sana.v2i1.265.

[14] J. Hao, "Housing Price Prediction Model and Impact Factors Analysis," Highlights Sci. Eng. Technol., vol. 39, pp. 1017–1023, Apr. 2023, doi:10.54097/hset.v39i.6696.

[15] H. Li, "House Price Prediction and Analysis Based on Random Forest and XGBoost Models," Highlights Business, Econ. Manag., vol. 21, pp. 934– 938, Dec. 2023, doi: 10.54097/hbem.v21i.14837.

[16] A. Deaconu, A. Buiga, and H. Tothăzan, "REAL ESTATE VALUATION MODELS PERFORMANCE IN PRICE PREDICTION," Int. J. Strateg. Prop. Manag., vol. 26, no. 2, pp. 86–105, Feb. 2022, doi: 10.3846/ijspm.2022.15962.