Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

House Price Prediction Analysis Using Linear Regression and Random Forest Algorithms

Hans Pran Limbong[1*], Mufti Alwisyah Lubis [2*], Mhd. Furqan[3*]

[1,2,3] *Universitas Islam Negeri Sumatera Utara*
hanspranlimbong03@gmail.com[1], muftialwi522@gmail.com[2], mfurqan@uinsu.ac.id[3]*

Abstract

This study aims to analyze house price prediction using two machine learning algorithms: Linear Regression and Random Forest. Quantitative evaluation is conducted using four main metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 Score, and Mean Absolute Percentage Error (MAPE). The experimental results show that the Random Forest model outperforms Linear Regression in all four evaluation metrics. The MAE and RMSE of the Random Forest model are lower, indicating that this model is more effective in minimizing prediction errors. Additionally, the higher R^2 Score demonstrates the model's better ability to explain house price variance, while the smaller MAPE indicates more accurate prediction errors in the context of real estate. These findings suggest that choosing the right algorithm is crucial for modeling complex house price data, and although Random Forest is more accurate, its black-box nature limits interpretability. Therefore, for future research, more interpretable methods such as XGBoost with SHAP analysis could be considered.

Keywords: *House Price Prediction, Linear Regression, Random Forest*

1. Introduction

House price prediction is a crucial issue in the property sector, with significant implications for decision-making by various stakeholders, including buyers, sellers, real estate agents, and investors. The ability to accurately estimate the value of a property can reduce financial risks, optimize investment strategies, and enhance transparency and efficiency in buying and selling transactions. In the context of increasing urbanization and a dynamic market environment, the need for a reliable house price prediction system is becoming increasingly important [1].With the advancement of information technology and the availability of large-scale data (big data), traditional approaches to property valuation are being replaced by more modern and adaptive methods, such as machine learning [2]. Machine learning offers advantages in handling high-dimensional and complex data patterns and can learn from historical data to generate more accurate predictions. In the context of house price prediction, machine learning enables the analysis of various features or determining variables such as building size, number of rooms, building age, and spatial factors like location and accessibility [3]. Two widely used algorithms in this domain are Linear Regression and Random Forest. Linear Regression is a classical statistical technique that assumes a linear relationship between independent and dependent variables [3]. Despite its simplicity, it is often used as a baseline in many studies due to its interpretability and efficiency. However, its limitations in capturing non-linear relationships make it less flexible in dealing with complex data [4]. On the other hand, Random Forest is an ensemble learning method based on decision trees that can overcome overfitting and capture non-linear relationships between variables. This algorithm also excels in handling poorly structured data and provides feature importance estimates that are useful for model interpretation [5]. This study aims to compare the performance of both algorithms in the context of house price prediction based on key relevant features. The evaluation is conducted using performance metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to determine which model yields more accurate and stable predictions. Through this comparison, a more comprehensive understanding of the strengths and limitations of each method in the context of real-world data is expected [6].Moreover, the application of house price prediction algorithms is not only relevant for individuals or business entities but also plays a significant role in spatial planning, housing policy, and smart city development [7]. Governments and planning agencies can use price prediction results to identify high-growth potential areas, guide infrastructure development, and formulate more targeted subsidy or price control policies. Therefore, the development of reliable, data-driven predictive models is not only a practical necessity but also an integral part of evidence-based sustainable development strategies.

2. Literature Review

2.1 House Price Prediction

Prediction of house prices is an important topic in the fields of real estate, economics, and data science, particularly because house prices are influenced by many dynamic factors. The value of residential property is determined not only by physical characteristics

such as building area, number of rooms, and building age, but also by external factors such as geographic location, accessibility to public facilities, social environment, and housing market trends. Considering this complexity, efforts to estimate house prices require approaches capable of analyzing multiple variables simultaneously and accurately. In the context of information technology, house price prediction has evolved rapidly alongside advancements in data processing technologies. Traditional approaches such as manual appraisal or simple regression methods are often inadequate to capture the complex relationships among variables. Therefore, machine learning-based methods are increasingly used due to their ability to process large volumes of data and identify hidden patterns that are not easily detectable through manual analysis [8]. Machine learning enables the use of various algorithms to build more accurate predictive models. House price prediction plays a strategic role not only for individual stakeholders such as buyers and sellers, but also for property developers, real estate agents, urban planners, and financial institutions. For developers, accurate pricing information helps in determining investment strategies, while for banks and financial institutions, price predictions are essential in assessing loan collateral. On a macro scale, the results of house price prediction can also be used by governments to formulate housing policies and control prices in subsidy programs [1]. Overall, the development of accurate house price prediction models reflects a paradigm shift from conventional approaches to data-driven approaches. In today's digital era, the abundance of historical data and the availability of advanced analytical tools offer significant opportunities for the development of intelligent, efficient, and adaptive predictive systems that can respond to market changes.

2.2 Linear Regression

Linear regression is one of the most basic and widely used statistical methods in data analysis for modeling the relationship between independent variables (predictors) and a dependent variable (target). This model works by finding the best-fitting straight line that represents the relationship between one or more input variables and the output. In the context of house price prediction, linear regression is used to estimate prices based on features such as building area, number of rooms, or property location [9].

One of the main reasons for using linear regression in prediction tasks is its ease of interpretation. The resulting regression coefficients indicate the contribution of each input variable to the target variable, allowing users to understand which factors most influence house prices. This model is also computationally efficient and well-suited for small to medium-sized datasets. As a result, linear regression is often used as a baseline before exploring more complex algorithms [10].

However, linear regression has several significant limitations, especially in dealing with non-linear data or data with many interactions among variables. The assumptions required in linear regression—such as linearity, homoscedasticity, and normality of residuals—are often not met in real-world cases. For example, the relationship between geographic location and house prices can be highly complex and non-linear, making it difficult for this model to accurately capture such relationships [11].

Although linear regression is easy to use, its success in prediction highly depends on data quality and the fulfillment of its fundamental assumptions. Therefore, for complex data such as property prices, linear regression should be applied with caution or combined with other approaches that can handle non-linear relationships and multivariate features.

2.3 Random Forest

Random Forest is a machine learning algorithm based on ensemble learning that combines the results of multiple decision trees to form a more accurate and stable predictive model. This algorithm was developed to address the weaknesses of single decision trees, which tend to overfit—that is, they tailor the model too closely to the training data, resulting in poor performance on new data. By building multiple trees on randomly selected subsets of the data (bootstrapping), Random Forest produces a more generalizable model that is resistant to noise [12].

One of the main advantages of Random Forest is its ability to capture non-linear relationships and complex interactions between variables. In house price prediction, this algorithm can detect patterns not captured by linear regression, such as the combined influence of location, house type, and environmental conditions. Random Forest also features an important capability called "feature importance," which indicates which variables contribute most to the prediction, although its interpretation is not as straightforward as that of a linear model [13].

Random Forest is also known for being robust to outliers and poorly structured data. Thanks to the voting or averaging mechanism across many trees, this model tends to be more stable compared to single predictive methods. This makes it one of the most popular algorithms in real-world applications, including house price prediction, medical diagnosis, image classification, and many other domains [14].

However, as a black-box model, the internal interpretation of Random Forest is relatively difficult, especially when used to explain causal relationships between variables. Therefore, although accurate, this algorithm is typically used for prediction rather than for exploring cause-effect relationships. In some cases, Random Forest can be combined with interpretability techniques to provide deeper insights.

3. Research Method

This research was conducted using a quantitative experimental approach that focuses on the implementation and evaluation of machine learning-based house price prediction models [15]. Broadly, the research process consists of five main stages: data collection, data preparation, data exploration, predictive model implementation, and performance evaluation. The complete research process flow is presented in Figure 1.

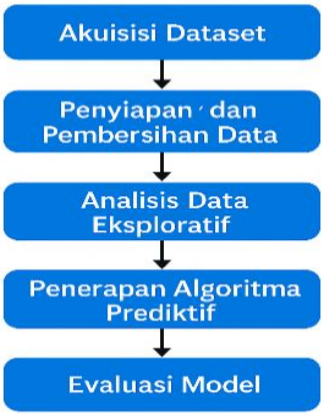

Figure 1: Complete Research Process Flow

1. Dataset Acquisition

The dataset used in this study was sourced from an open database on the Kaggle platform, titled *House Prices - Advanced Regression Techniques*. This dataset was chosen because it provides key variables commonly used in residential property price prediction, including building size, number of bedrooms, house condition, property type, and geographic location. The dataset was downloaded in CSV format and then loaded into the Python programming environment for further analysis.

2. Data Preparation and Cleaning

All analysis processes were conducted using Google Colaboratory (Colab), which leverages cloud computing technology and supports machine learning libraries such as pandas, scikit-learn, and matplotlib. The data preparation process involved several stages (Fu, 2024):

- **Handling Missing Values**
  Data containing missing or invalid values was examined and addressed using specific approaches, such as imputation with mean/median values or removal of irrelevant rows.
- **Removing Duplicates and Outliers**
  Duplicate entries and outliers were identified to maintain a representative and unbiased data distribution.
- **Categorical Variable Transformation**
  Categorical variables were converted into numerical format using one-hot encoding techniques.
- **Feature Normalization and Scaling**
  Some numerical features were standardized to maintain proportionality among variables, especially for models sensitive to feature scaling.

3. Exploratory Data Analysis (EDA)

This stage was carried out to gain deeper insights into the data characteristics and the relationships between features. Visualization of variable distributions, feature correlations, and identification of general trends were performed using histograms, boxplots, and heatmaps. EDA also helped in recognizing non-linear relationships and potential multicollinearity among independent variables.

4. Predictive Algorithm Implementation

Two machine learning algorithms were employed in this study:

- **Linear Regression**
  Used as a baseline approach due to its simplicity and interpretability. This model assumes a linear relationship between features and the target variable.
- **Random Forest Regressor**
  An ensemble method based on decision trees that builds multiple individual models and combines their results. This model is known for effectively capturing complex relationships and feature interactions.

The models were trained using a training set and tested on a test set, with a data split ratio of 80:20.

5. Model Evaluation

The performance of both models was measured and compared using the following evaluation metrics:

- **Mean Absolute Error (MAE)**
  Reflects the average absolute difference between actual values and predictions.
- **Root Mean Squared Error (RMSE)**
  Applies a higher penalty to extreme errors, making it suitable for evaluating models against predictions far from the actual values.
- **R-squared (R^2 Score)**
  Measures how much of the variance in the target variable is explained by the model.
- **Mean Absolute Percentage Error (MAPE)**
  Presents the error in percentage form, making it easier to interpret for non-technical readers.

4. Results and Discussion

The distribution of house prices is a critical aspect that must be analyzed before applying prediction models, as it can affect the underlying assumptions of certain algorithms. Based on Figure 2, it is evident that the distribution of house prices is right-skewed (Fu, 2024). This indicates that the majority of houses fall within the lower to mid-price range, while only a small portion of properties have extremely high prices (outliers).This asymmetric distribution causes the mean value to be less representative of the data's central tendency. This phenomenon is particularly important when using models like Linear Regression, which assumes homoscedastic and normally distributed residuals. In other words, this imbalance can potentially lead to systematic prediction errors, especially in the high-price segment.

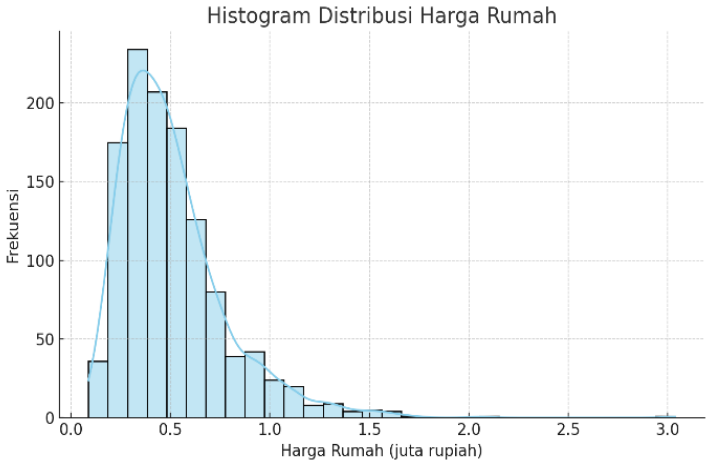

Figure 2: Histogram of House Price Distribution

House price prediction using the Linear Regression algorithm is visualized in Figure 3, where actual values are plotted against predicted values. The results show a considerable deviation from the diagonal line (the ideal line of perfect prediction), particularly in the group of high-priced houses. This indicates that the model tends to underfit, meaning it fails to capture the complex relationships between input variables and the output.

As is well known, Linear Regression relies on the assumption of linear relationships among features and overlooks non-linear interactions or complex multicollinearity. Therefore, discrepancies between predicted and actual prices are very likely when there are non-linear interactions between features such as land area, location, number of rooms, and building age. This issue is further exacerbated by the presence of outliers, which exert an excessive influence on parameter estimation.

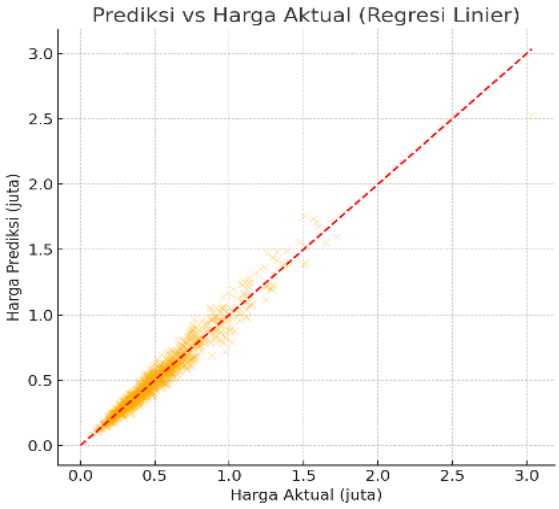

Figure 3: Prediction vs Actual Results in the Linear Regression Model.

Conversely, the prediction results using Random Forest Regression, as shown in Figure 4, display a pattern that is closer to the ideal line. Most of the observation points lie around the line y = x, indicating that the model is able to produce fairly accurate predictions across both low- and high-price segments. This advantage stems from the internal mechanism of Random Forest, which splits the data into various bootstrap samples and builds models based on a collection of decision trees.

Random Forest is robust against outliers and multicollinearity, and it can learn feature interactions through a non-parametric approach.

This model reduces variance without significantly increasing bias, making it suitable for real estate data, which is inherently complex and heterogeneous. This flexibility enables the model to adapt to spatial and structural features that are difficult to capture using linear regression.

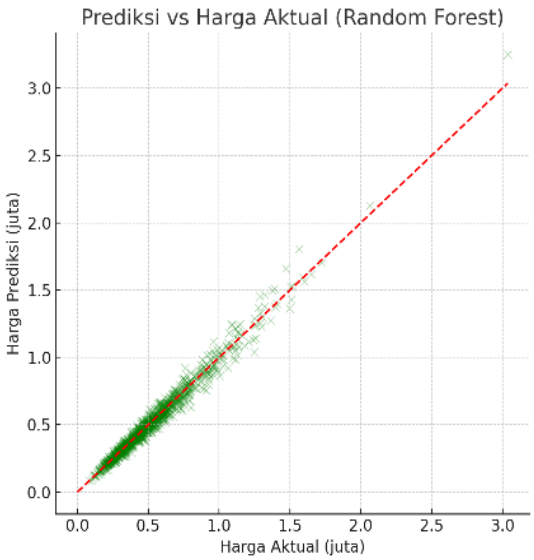

Figure 4: Prediction vs Actual Results in the Random Forest Model

To obtain a comprehensive evaluation, four metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 Score, and Mean Absolute Percentage Error (MAPE). The evaluation results are visually presented in Figure 5. The Random Forest model consistently outperformed in all four metrics:

- **MAE**: Random Forest recorded a lower value compared to Linear Regression, indicating that the average absolute deviation of predictions from actual data is smaller.
- **RMSE**: Random Forest also had a lower RMSE, showing that the model is more effective in minimizing large prediction errors (emphasizing squared errors).
- **R^2 Score**: The higher R^2 score for Random Forest indicates that the model can explain a larger proportion of the variance in house prices.
- **MAPE**: The smaller MAPE for Random Forest reflects a smaller relative predictive deviation from actual values, which is important in the real estate context because prediction errors are sensitive to nominal values.

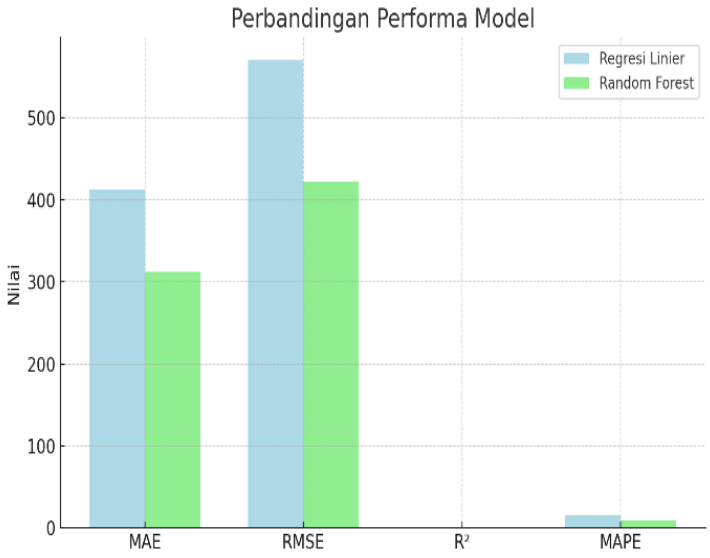

Figure 5: Comparison of Model Performance Based on MAE, RMSE, R^2, and MAPE

These findings confirm that the choice of predictive algorithm must consider the characteristics of the data distribution as well as the complexity of relationships among variables. Linear models may be suitable for homogeneous data following a linear pattern; however, in the context of house prices influenced by many interacting spatial and structural factors, using algorithms like Random Forest is more appropriate.

Moreover, it is important to note that although Random Forest demonstrates superior performance, this model functions as a black-box and is less interpretable compared to Linear Regression. Therefore, if the analysis aims to understand the contribution of each feature, other methods such as XGBoost combined with SHAP analysis can be considered in further studies.

5. Conclusion

The quantitative evaluation results show that the Random Forest model consistently outperforms the Linear Regression model across four main metrics: MAE, RMSE, R^2 Score, and MAPE The Random Forest model effectively minimizes prediction errors and better explains the variance in house prices. The lower MAE and RMSE values indicate higher prediction accuracy, while the higher R^2 score

demonstrates the model's superior ability to capture the relationships between variables compared to Linear Regression. Additionally, the smaller MAPE reflects more accurate predictions in terms of relative deviation from actual values.

However, despite the superior performance of Random Forest, a major limitation of this model is its black-box nature, which reduces interpretability. Therefore, for analyses requiring a deeper understanding of each feature's contribution, other methods such as XGBoost combined with SHAP analysis can be considered for further study.

These findings also emphasize the importance of selecting algorithms that align with the characteristics of the data. House price data, influenced by many complex factors, is better suited for ensemble-based approaches like Random Forest rather than simpler linear models. As an implication, using more complex and accurate models can enhance the quality of predictions in the real estate context.

References

[1] Arizqi, R., & Prasetyo, H. (2022). Analisis prediksi harga rumah di Bandung menggunakan regresi linear berganda. *Journal of Computer Science Research*, *6*(1). https://ejurnal.politeknikpratama.ac.id/index.php/jcsr/article/download/3038/2873

[2] Brownlee, J. (2023, December 6). Linear regression for machine learning. *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/linear-regression-for-machine-learning/

[3] Fu, Y. (2024). A comparative study of house price prediction using linear regression and random forest models. *Highlights in Science, Engineering and Technology, 107*, 96–103. https://doi.org/10.54097/vcy5n584

[4] Guna, R., & Sudiarta, I. M. (2023). Uji performansi algoritma LR dan RFR pada implementasi sistem prediksi harga rumah. *Jurnal Nasional Teknologi Informasi dan Aplikasinya*, *6*(3). https://ojs.unud.ac.id/index.php/jnatia/article/download/102444/50654

[5] Khoirudin, A., & Wahyuningtyas, D. (2022). Penerapan Random Forest Regression untuk memprediksi harga jual rumah dan Cosine Similarity untuk rekomendasi rumah di Provinsi Jawa Barat. *Jurnal Coding*, *10*(1). https://www.neliti.com/publications/569157/download

[6] Kurniawan, A. D., & Wijaya, T. (2022). Implementasi machine learning untuk prediksi harga rumah menggunakan algoritma Random Forest. *Computatio: Journal of Computer Science*, *9*(2) https://journal.untar.ac.id/index.php/computatio/article/download/15173/17830/89193

[7] Lewinson, E. (2023, April 20). A comprehensive overview of regression evaluation metrics. *NVIDIA Developer Blog*. Retrieved from https://developer.nvidia.com/blog/a-comprehensive-overview-of-regression-evaluation-metrics/

[8] Montoya, A., & DataCanary. (2016). *House Prices – Advanced Regression Techniques* [Data set]. Kaggle. https://www.kaggle.com/c/house-prices-advanced-regression-techniques

[9] Novianto, D., & Andhika, M. (2021). Prediksi harga rumah menggunakan machine learning algoritma linear regression. *Jurnal Teknik Elektro dan Sistem Informasi*, *8*(2). https://jurnal.unidha.ac.id/index.php/jteksis/article/download/1732/953/

[10] Rachman, A., & Nugroho, D. (2022). Analisis prediksi harga rumah sesuai spesifikasi menggunakan multiple linear regression. *Jurnal Informatika UPNVJ*, *8*(1). https://ejournal.upnvj.ac.id/informatik/article/download/3635/1498/10600

[11] Rambe, Y., & Siregar, R. A. (2022). Prediksi harga rumah di Jakarta Pusat menggunakan algoritma General Regression Neural Network. *Jurnal Ilmu Komputer dan Bisnis*, *5*(2). https://www.stmikdharmapalariau.ac.id/ojs/index.php/jikb/article/view/840/633

[12] scikit-learn developers. (2025). *3.4 Metrics and scoring: quantifying the quality of predictions*. In Scikit-learn documentation (version 1.6.1). Retrieved from https://scikit-learn.org/stable/modules/model_evaluation.html

[13] Wahyuni, R., & Hidayat, M. (2023). Pendekatan machine learning untuk estimasi harga rumah berdasarkan fitur properti. *Jurnal ALPHA: Jurnal Teknik dan Sains*, *1*(2). https://ejournal.publine.or.id/index.php/alpha/article/download/99/104

[14] Ye, Q. (2024). House price prediction using machine learning for Ames, Iowa. *Applied and Computational Engineering, 55*(1), 44–54. https://doi.org/10.54254/2755-2721/55/20241483

[15] Yu, J. (2023). Prediction on housing price based on the data on Kaggle. In Z. Zeng et al. (Eds.), *Atlantis Highlights in Engineering*: Proceedings of the 2022 3rd International Conference on E-Commerce and Internet Technology (ECIT 2022) (Vol. 11, pp. 627–634). Atlantis Press. https://doi.org/10.2991/978-94-6463-005-3_64