Terbit online pada laman web jurnal: **https://journal.upp.ac.id/index.php/rjti/index**

RJTI
Riau Jurnal Teknik Informatika

Vol. 4 No. 3 (2025) 436 – 440 | ISSN: 2829-2529(Media Online)

Application of Multiple Linear Regression Algorithm for House Price Estimation Based on Building Location and Area to Improve Predictive Accuracy in Real Estate Valuation

Kecitaan Harefa[1]
[1]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang
[1]dosen00842@unpam.ac.id

Abstract

Significant differences in home prices, even on properties with similar building sizes and locations, pose a major challenge in accurately determining property valuations. The discrepancy between the actual market price and the estimated value makes it difficult for potential buyers, sellers, and developers to make the right decision. To overcome these problems, this study applied the Multiple Linear Regression (MLR) algorithm in the Decision Support System (DSS) to estimate house prices based on the location and area of the building. The dataset used consists of 545 housing data points with variables such as house prices, locations, and building areas. The research stages include data collection, pre-processing (data cleaning and normalization), model development using MLR, and model performance evaluation. The evaluation was carried out using the division of trained data and test data with an 80:20 ratio, so that the model was tested using data that was not previously trained. The results showed that the model produced a Mean Absolute Error (MAE) of 1,474,748.13, a Root Mean Squared Error (RMSE) of 1,917,103.70, and a coefficient of determination (R^2) of 0.273. A relatively low R^2 value indicates that the location and area variables of the building are not sufficient to explain the overall variation in house prices, so the addition of other variables—such as the number of rooms, facilities, and environmental conditions—is needed to improve the accuracy of the prediction and produce a more representative price estimate.

Keywords: *House Price Estimation, Multiple Linear Regression, Location, Building Area.*

Abstrak

Perbedaan harga rumah yang signifikan, bahkan pada properti dengan luas bangunan dan lokasi yang serupa, menimbulkan tantangan besar dalam menentukan penilaian properti secara akurat. Ketidaksesuaian antara harga pasar aktual dan nilai estimasi menyebabkan kesulitan bagi calon pembeli, penjual, maupun pengembang dalam mengambil keputusan yang tepat. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan algoritma *Multiple Linear Regression (MLR)* dalam *Decision Support System (DSS)* untuk memperkirakan harga rumah berdasarkan lokasi dan luas bangunan. Dataset yang digunakan terdiri atas 545 data perumahan dengan variabel harga rumah, lokasi, dan luas bangunan. Tahapan penelitian meliputi pengumpulan data, pra-pengolahan (pembersihan dan normalisasi data), pembangunan model menggunakan *MLR*, serta evaluasi kinerja model. Evaluasi dilakukan menggunakan pembagian data latih dan data uji dengan rasio 80:20, sehingga model diuji menggunakan data yang tidak dilatih sebelumnya. Hasil penelitian menunjukkan bahwa model menghasilkan *Mean Absolute Error (MAE)* sebesar 1.474.748,13, *Root Mean Squared Error (RMSE)* sebesar 1.917.103,70, dan koefisien determinasi (R^2) sebesar 0,273. Nilai R^2 yang relatif rendah mengindikasikan bahwa variabel lokasi dan luas bangunan belum cukup untuk menjelaskan variasi harga rumah secara keseluruhan, sehingga diperlukan penambahan variabel lain—seperti jumlah kamar, fasilitas, dan kondisi lingkungan—untuk meningkatkan akurasi prediksi dan menghasilkan estimasi harga yang lebih representatif.

Kata kunci: Estimasi Harga Rumah, Regresi Linear Berganda, Lokasi, Luas Bangunan.

1. Introduction

The development of the property sector in Indonesia has increased rapidly in recent years. The community's need for decent and strategic housing continues to drive fluctuations in house prices, especially in urban areas such as Greater Jakarta [1]. A major challenge that arises is determining the selling or buying price of a house that accurately reflects market conditions, as prices are influenced by factors such as location, building area, land area, and available facilities. These

discrepancies often lead to inefficient and less objective decision-making for buyers, sellers, and developers [2].

To date, house price determination is still often conducted based on subjective estimates or by merely referring to local market listings without deeper analytical consideration [3]. This practice frequently results in significant gaps between estimated and actual prices. Such mismatches can cause financial losses for sellers who undervalue their property and potential overspending for buyers. Therefore, a data-driven approach capable of quantitatively identifying how various determinants influence house prices is urgently needed [4].

Several previous studies have investigated house price prediction using statistical and machine learning approaches. For example, [5] employed linear regression to estimate house prices in Jakarta with promising results on homogeneous datasets. Study [6] applied the Random Forest Regression method and demonstrated improved accuracy compared to simple linear models. Meanwhile, [7] combined multiple regression with feature selection to analyze the influence of location and public facilities. Another study by [8] utilized Support Vector Regression (SVR) and found that higher model complexity does not always guarantee better predictive performance. Furthermore, [9] confirmed that multiple linear regression remains relevant when independent variables exhibit strong correlations with house prices.

Based on this literature review, it can be concluded that many studies incorporate numerous variables, often overlooking model simplicity and ease of implementation. Conversely, research that focuses solely on core variables such as location and building area is still limited, even though these two factors significantly influence property value. This highlights the need to examine how well these core predictors alone can explain variations in house prices.

The choice of the Multiple Linear Regression (MLR) algorithm in this study is motivated by its ability to analyze linear relationships between one dependent variable and multiple independent variables. MLR is easy to interpret, computationally efficient, and suitable for datasets of moderate size. It also enables the examination of partial effects of each factor on the dependent variable, making it useful for assisting both developers and home buyers in decision-making processes [10].

Compared to more complex methods such as Random Forest or Neural Networks, MLR offers advantages in terms of model transparency and straightforward implementation in decision support systems (DSS). Its outputs take the form of clear mathematical equations, allowing users to easily understand and apply the price estimation process. A web-based integration of this approach also enhances public accessibility to objective house price information [11].

The novelty of this study lies in its focus on evaluating the predictive capability of only two core variables—location and building area—using a standardized dataset of 545 housing records. Unlike previous studies that rely on numerous predictors, this research intentionally isolates the fundamental determinants to examine how much variance in house prices can be explained by these essential features alone. Additionally, the development of a DSS framework enhances the practical contribution of this work by enabling real-time price estimation [12].

Thus, this study aims to apply the Multiple Linear Regression algorithm to build a decision support system for estimating house prices based on building location and area. The results of this research are expected to contribute to the fields of information systems and property data analysis, and to serve as a reference for property developers, agents, and buyers in setting more accurate and rational prices [13]. This study also aims to strengthen the existing literature on the use of regression-based methods within decision support systems in the property sector.

2. Research Methods

The research method was prepared to provide a systematic overview of the stages carried out, starting from data collection, data processing, algorithm application, to evaluation of model results [14].

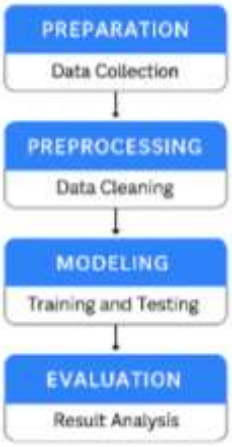

Figure 1. Research Methods

The following is an explanation of the steps of the method above [15], [16]:

a. Preparation
This stage involves collecting the dataset that forms the basis of the analysis. The data used in this research consists of 545 housing records, containing information such as house price, location, and

DOI : https://doi.org/10.52158/rjti.v4i3.3993

building area. The dataset was obtained from publicly available property data sources. The purpose of this stage is to ensure that the dataset is representative of real market conditions and suitable for statistical analysis.

b. Preprocessing
The collected dataset is not always ready for direct use due to issues such as missing values, inconsistent formatting, or outliers. Therefore, preprocessing steps were conducted, including:
1. Data cleaning, such as removing duplicates and handling missing values
2. Converting price formats into numeric values
3. Normalizing numerical variables, especially building area
4. Encoding categorical variables, such as location, using one-hot encoding

These preprocessing procedures were carried out using the Python programming language with supporting libraries such as pandas, NumPy, and scikit-learn. Preprocessing is essential to improve data quality and ensure reliable model performance.

c. Modeling
At this stage, the Multiple Linear Regression (MLR) algorithm was applied to develop the house price prediction model. The dataset was divided using a train–test split of 80:20, where 80% (436 data points) were used for training and 20% (109 data points) were used for model testing.
The modeling process included:
1. Training the MLR model using the training data
2. Testing the model using the testing data
3. Generating a mathematical regression equation that describes the relationship between the independent variables (location and building area) and the dependent variable (house price)

The modeling process was implemented using scikit-learn's LinearRegression library. As an additional validation measure, the model can also be strengthened through k-fold cross-validation, ensuring that model performance remains consistent across different subsets of the data.

d. Evaluation
After the model was constructed, its performance was evaluated using recognized regression metrics, including:
1. Mean Absolute Error (MAE)
2. Root Mean Squared Error (RMSE)
3. Coefficient of Determination (R^2)

These metrics were calculated using the scikit-learn metrics module. The R^2 value was used to measure how much variance in house prices can be explained by the building area and location variables. If the model achieves satisfactory accuracy, it is considered suitable for integration into the Decision Support System (DSS).

3. Results and Discussion

1. Preprocessing

The dataset used in this study consists of 545 housing data points obtained from Kaggle. Each record represents a single housing unit with a variety of numerical and categorical attributes that potentially influence its market price. Numerical variables include price, area (building area), bedrooms, bathrooms, and stories, while categorical variables include mainroad, guestroom, basement, airconditioning, parking, prefarea, and furnishingstatus. These attributes indicate diverse property characteristics; for example, one of the entries describes a house with a building area of 7,420 m², four bedrooms, two bathrooms, main road access, and a furnished interior.

Before modeling, the dataset underwent preprocessing steps including handling missing values, normalizing numerical attributes, and encoding categorical variables to ensure their compatibility with the Multiple Linear Regression model. This stage is essential for enhancing data quality and preparing a reliable foundation for subsequent analysis, particularly to identify the relationship between house characteristics and selling price.

Amount of Data: 545

5 Top Data:

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	\
0	13300000	7420	4	2	3	yes	no	no	
1	12250000	8960	4	4	4	yes	no	no	
2	12250000	9960	3	2	2	yes	no	yes	
3	12215000	7500	4	2	2	yes	no	yes	
4	11410000	7420	4	1	2	yes	yes	yes	

	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	no	yes	2	yes	furnished
1	no	yes	3	no	furnished
2	no	no	2	yes	semi-furnished
3	no	yes	3	yes	furnished
4	no	yes	2	no	furnished

Figure 2. Preprocessing Data

2. Modeling

The predictive model was developed using the Multiple Linear Regression (MLR) algorithm. Following preprocessing, the dataset was divided into training and testing sets using an 80:20 ratio, enabling the model to be evaluated on data not seen during training. The evaluation metrics indicate that the MLR model yielded a Mean Absolute Error (MAE) of 1,474,748.13, a Root Mean Squared Error (RMSE) of 1,917,103.70, and a Coefficient of Determination (R^2) of 0.273.

The MAE and RMSE values suggest that the model retains a relatively high prediction error, demonstrating that the predicted prices deviate considerably from the actual values. The R^2 value of 0.273 indicates that only 27.3% of the variation in house prices can be explained by the predictor variables—primarily building area and location—while the remaining 72.7% is influenced by additional factors such as number of rooms, facilities, environmental conditions, and accessibility.

DOI : https://doi.org/10.52158/rjti.v4i3.3993

Although the model is able to capture the general linear trend between variables, its explanatory power remains limited. These results imply that incorporating more relevant features or experimenting with alternative algorithms such as Random Forest Regression or Gradient Boosting may yield better predictive performance. Nonetheless, the current model provides a strong foundational baseline and demonstrates the feasibility of using simple linear models for initial property valuation.

```
=== RESULTS OF MODEL EVALUATION ===
Mean Absolute Error (MAE)  : 1,474,748.13
Root Mean Squared Error (RMSE): 1,917,103.70
R² Score                   : 0.273
```

Figure 3. Result of Model Evaluation

3. Evaluation

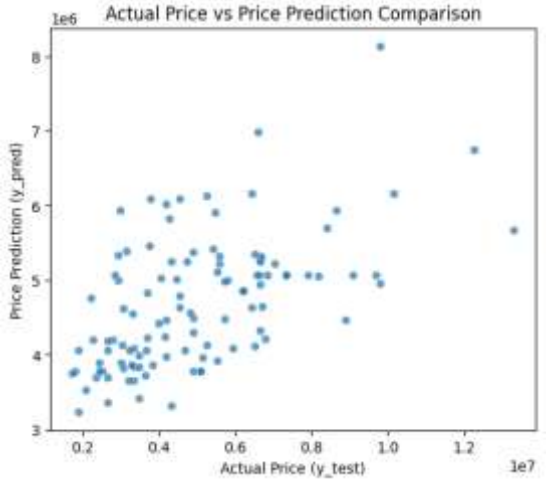

Figure 4. Actual Price vs Price Prediction Comparison

The visualization above compares the actual house prices with the predicted values generated by the MLR model. Each point represents one data sample, where the X-axis denotes the actual price and the Y-axis denotes the predicted price. The scatterplot shows a general upward trend, indicating that the model successfully identifies the basic pattern between house characteristics and market prices.

However, many data points appear widely dispersed from the ideal diagonal line—representing perfect predictions—signifying notable prediction errors. This dispersion suggests that the model does not fully capture the complexity of the factors influencing house prices. The wide variability reflected in the plot aligns with the relatively low R^2 value and supports the conclusion that location and building area alone cannot comprehensively predict property values.

Thus, while the MLR model provides a preliminary understanding of linear relationships within the dataset, the findings highlight opportunities for future work. Enhancing the model with additional features or leveraging more sophisticated algorithms such as Random Forest, Gradient Boosting, or XGBoost could improve predictive accuracy and generate more reliable house price estimates for practical use.

4. Conclusion

The application of the Multiple Linear Regression algorithm is able to provide a quantitative overview of the relationship between building location and building area toward house prices. The model developed produced a Mean Absolute Error (MAE) of 1,474,748.13, a Root Mean Squared Error (RMSE) of 1,917,103.70, and a Coefficient of Determination (R^2) of 0.273, indicating that approximately 27.3% of the variation in house prices can be explained by these two variables. These results show that although the model successfully captures the fundamental linear pattern between location and building area and their influence on house prices, the predictive accuracy remains limited due to the absence of other influential factors—such as building condition, number of rooms, public facilities, environmental accessibility, and socio-economic indicators of the surrounding area.

Thus, this study demonstrates that applying multiple linear regression can address the problem of estimating house prices in a data-driven and objective manner. However, the findings also highlight the need for improvement. Future research could incorporate additional variables such as proximity to public services, construction quality, neighborhood characteristics, or socio-economic profiles to enhance the model's explanatory power. Furthermore, experimenting with more advanced or ensemble-based learning methods—such as Random Forest, Gradient Boosting, or hybrid regression approaches—may provide substantially higher predictive accuracy and strengthen the resulting decision support system for real estate valuation.

Reference List

[1] E. Alenany, L. A. Lekham, and S. Lu, "Integrated Clustering Regression for Real Estate Valuation," Real Estate Finance, 2021.

[2] A. Deaconu, A. Buiga, and H. Tothăzan, "Real estate valuation models performance in price prediction," International Journal of Strategic Property Management, vol. 26, no. 2, pp. 86–105, Feb. 2022, doi: 10.3846/ijspm 2022.15962.

[3] N. T. Yousir, S. M. Abdulameer, and S. A. Mostafa, "Data Mining Approach in Predicting House Price for Automated Property Appraiser Systems," in Lecture Notes in Networks and Systems, 2023. doi: 10.1007/978-981-99-3010-4_45.

[4] O. Saraswat and N. Arunachalam, "Efficient Rent Price Prediction Model for the Development of a House Marketplace Website by Analyzing Various Regression-Based Machine Learning Algorithms," in Lecture Notes in Networks and Systems, 2023. doi: 10.1007/978-981-99-5166-6_72.

[5] E. Palupi, "House Price Prediction Using Data Mining with Linear Regression and Neural Network Algorithms," Jurnal

Riset Informatika, vol. 6, no. 1, pp. 15–20, Dec. 2023, doi: 10.34288/jri.v6i1.262.

[6] E. A. F. Elmuna, T. Chamidy, and F. Nugroho, "Optimization of the Random Forest Method Using Principal Component Analysis to Predict House Prices," International Journal of Advances in Data and Information Systems, vol. 4, no. 2, 2023, doi: 10.25008/ijadis.v4i2.1290.

[7] C. D. Whitmire, J. M. Vance, H. K. Rasheed, A. Missaoui, K. M. Rasheed, and F. W. Maier, "Using Machine Learning and Feature Selection for Alfalfa Yield Prediction," AI (Switzerland), vol. 2, no. 1, 2021, doi: 10.3390/ai2010006.

[8] M. P. Kusuma and A. Kudus, "Penerapan Metode Support Vector Regression (SVR) pada Data Survival KPR PT. Bank ABC, Tbk.," Bandung Conference Series: Statistics, vol. 2, no. 2, 2022, doi: 10.29313/bcss.v2i2.3614.

[9] A. N. Safira, B. Warsito, and A. Rusgiyono, "ANALISIS SUPPORT VECTOR REGRESSION (SVR) DENGAN ALGORITMA GRID SEARCH TIME SERIES CROSS VALIDATION UNTUK PREDIKSI JUMLAH KASUS TERKONFIRMASI COVID-19 DI INDONESIA," Jurnal Gaussian, vol. 11, no. 4, 2023, doi: 10.14710/j.gauss.11.4.512-521.

[10] L. Rampini and F. Re Cecconi, "Artificial intelligence algorithms to predict Italian real estate market prices," Journal of Property Investment and Finance, vol. 40, no. 6, pp. 588–611, Sep. 2022, doi: 10.1108/JPIF-08-2021-0073.

[11] P. Y. Wang, C. T. Chen, J. W. Su, T. Y. Wang, and S. H. Huang, "Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism," IEEE Access, vol. 9, pp. 55244–55259, 2021, doi: 10.1109/ACCESS.2021.3071306.

[12] W. K. O. Ho, B. S. Tang, and S. W. Wong, "Predicting property prices with machine learning algorithms," Journal of Property Research, vol. 38, no. 1, pp. 48–70, 2021, doi: 10.1080/09599916.2020.1832558.

[13] J. Jerry, Y. Christian, and H. Herman, "Rental Price Prediction of Boarding Houses in Batam City Using Linear Regression and Random Forest Algorithms," Journal of Applied Informatics and Computing, vol. 7, no. 2, 2023, doi: 10.30871/jaic.v7i2.6732.

[14] Y. Shi, "Application of Improved Linear Regression Algorithm in Business Behavior Analysis," in Procedia Computer Science, 2023. doi: 10.1016/j.procs.2023.11.144.

[15] Willis Puspita Sari, Nurhasanah, Andin Eka Safitri, and Sartika Lina Mulani Sitio, "Analisis Pengaruh Gaya Kepemimpinan, Motivasi, dan Disiplin Kerja Terhadap Kinerja Karyawan Menggunakan Regresi Linier Berganda," Riau Jurnal Teknik Informatika, vol. 3, no. 3, pp. 80–84, Nov. 2024, doi: https://doi.org/10.30606/rjti.v3i3.3446.

[16] I. P. Putra and I. K. G. Suhardana, "Perbandingan Akurasi Algoritma Regresi Linier, Regresi Polinomial, dan Support Vector Regression Pada Model Sistem Prediksi Harga Rumah," Jurnal Nasional Teknologi Informasi dan Aplikasinya, vol. 1, no. 1, 2022.

DOI : https://doi.org/10.52158/rjti.v4i3.3993