Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Predicting House Prices in South Jakarta using Linear Regression and Decision Tree Regressor Methods

Frendy[1*], Neviasary[2], Nursiah[3], Laura Vidiani[4], Siti Nurdiani [5]

[1,2,3,4,5]*Program Studi Infirmatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika,Indonesia*
frendycoc862@gmail.com [1*], neviasary211004@gmail.com [2], nursiahn777@gmail.com [3], vidianiilauraa@gmail.com [4], siti.sxd@bsi.ac.id[5]

Abstract

The development of digital technology and data analytics has significantly improved predictive capabilities across various sectors, including the property market. House price prediction is a critical element in decision-making for buyers, sellers, investors, and developers, as prices are influenced by factors such as location, building size, land area, number of rooms, and available facilities. This study aims to build a house price prediction model in South Jakarta using Linear Regression and Decision Tree Regressor machine learning algorithms and compare their performance based on regression evaluation metrics. The dataset consists of 1010 entries, divided into 80% training data and 20% testing data. The experimental results show that Linear Regression produced the best performance with an R^2 score of 0.7713, meaning that the model can explain 77% of the variance in house prices. The model achieved a prediction error of approximately MAE Rp 1.98 billion and RMSE Rp 3.26 billion. Meanwhile, the Decision Tree Regressor obtained an R^2 score of 0.5560 with higher prediction errors, indicating a tendency toward overfitting and weaker generalization on testing data. Therefore, Linear Regression is recommended as the most effective approach for predicting property prices in South Jakarta and has the potential to be applied in decision-support systems for real estate market analysis.

Keywords: *Prediction, House Price, Linear Regression, Decision Tree Regressor, Machine Learning*

1. Introduction

The development of artificial intelligence is advancing rapidly across almost every field, especially in the property sector. In the era of artificial intelligence, it is used to predict changes in house prices over time and the impact of location, providing a broader understanding of the dynamics of the property market [1]. Predicting house prices is a crucial aspect of real estate and economics, as it helps decision-making for buyers, sellers, and developers [2]. Property prices are often influenced by several factors, including geographical location, land and structure size, number of rooms, available amenities, market conditions, ease of access, and surrounding environmental conditions [3], [4].

In-house price prediction has seen significant changes in the methods and models used to estimate price changes. This change is happening because of the rapid developments in information and communication technology, which are affecting how companies interact with customers and providing new access to crucial data [5]. House prices in South Jakarta are recognized as being among the fastest-growing in Indonesia[6].

Various approaches have been introduced to improve prediction accuracy, with some relying on traditional statistical models while others apply more modern machine learning techniques. Machine learning techniques such as linear regression and decision tree regressors have become crucial tools for sellers, buyers, developers, and investors in making data-driven decisions [7]. Linear regression is one of the models with advantages in terms of interpretability and computational efficiency. However, this model is often insufficient to capture nonlinear patterns or complex interactions that may exist in house price data [8]. On the other hand, a decision tree regressor, which can partition data to capture the complexity of feature relationships, can provide better predictive performance for nonlinear data[9].

This research aims to build a model for predicting house prices in South Jakarta using linear regression and decision tree regressor methods, analyze the performance of both methods in predicting property prices based on relevant evaluation metrics, and compare the accuracy of both methods to determine the best model for predicting house prices in South Jakarta.

2. Theoretical Foundation

2.1 Machine Learning

Computers can learn from data without explicit programming thanks to a subfield of artificial intelligence called machine learning. Machine learning is used in home price prediction to identify trends and relationships between property selling prices and their attributes. Supervised learning, unsupervised learning, and reinforcement learning are the three primary subcategories of machine learning. Because this study utilizes historical data with established price labels, it employs a supervised learning approach.

2.2 Supervised Learning

A machine learning technique called supervised learning employs data that already contains labels or intended outputs to train a model. The model learns the relationship between the desired output (home price) and the input parameters (land area, building area, number of rooms, etc.) to predict home prices. Minimizing the variance between the model's predictions and the actual values in the training data is how the learning process proceeds.

2.3 Linear Regression

Linear regression is a statistical method for modeling the relationship between independent and dependent variables with a straight line. Linear regression is easy to understand, quick to compute, and simple to use. However, it can be affected by changes in the data and struggles to show nonlinear relationships.

2.4 Decision tree regressor

A decision tree regressor is a machine learning method for predicting numerical values. It works by splitting data into groups based on set values for each feature. At each step, it picks the feature and value that lead to the largest decrease in data spread or the greatest increase in information gain. Decision tree regressors can find patterns that aren't linear, don't require data scaling, handle outliers well, and select useful features automatically. However, if not adjusted with the right settings, they often fit the training data too closely.

2.5 Metrik Evaluasi

To evaluate the performance of the regression model, several standard metrics are used:

Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values. Calculated using the following formula:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (1)$$

Mean Squared Error (MSE) is used to measure the average squared error value. It is calculated using the following formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (2)$$

Root Mean Squared Error (RMSE) is the square root of MSE. It is calculated using the following formula:

$$\text{RMSE} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (3)$$

R-squared (R^2) or the coefficient of determination measures the proportion of variance in the dependent variable that can be explained by the independent variables. Calculated using the following formula:

$$R^2 = 1 - \frac{\text{SS\_res}}{\text{SS\_tot}} \quad (4)$$

Dimana:
n = amount of data,
y_i = actual value,
$\hat{y}_i$ = predicted value,
SS_res = sum of squared residuals,
SS_tot = total sum of squares.

The R^2 value ranges from 0 to 1, with higher values indicating a better model in explaining data variability.

3. Research Methods

This research uses regression methods with linear regression and decision tree regressor algorithms in Google Colab, and a performance comparison is conducted to determine the best model based on regression metric results.

3.1 Research Stages

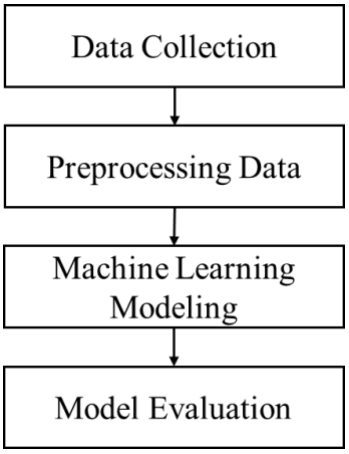

Fig. 1: The sequence of stages in conducting this research.

This is a discussion about the stages of the auction process:

1. Data collection : the data used is house price list data.
2. Preprocessing data : checking the data to see if there is any missing or inconsistent data.
3. Machine learning modeling : the process by which data will be grouped using the linear regression algorithm and the decision tree regressor.
4. Model evaluation : evaluate the results obtained from data clustering to assess the relevance and accuracy of the findings.

3.2. Dataset

The dataset used comes from Kaggle [10]. This dataset contains input variables consisting of house address, building area, land area, number of bedrooms, number of bathrooms, and total car capacity in the garage, as well as the house price, which is the target output.

3.3. Preprocessing Data

The data preprocessing stages performed include the following:

1. Data Cleaning
   Removing duplicate data and correcting column formats for consistency.
2. Handling Missing
   Values Missing data in numerical features is filled using the median method, while categorical features are imputed using the mode.
3. Outlier Treatment
   Outliers were detected using the Interquartile Range (IQR) method, and unrealistic extreme values were removed.
4. Feature Encoding
   Categorical attributes (e.g., location) are converted to numerical using Label Encoding.
5. Normalization (Min-Max Scaling)
   This is done for the Linear Regression model to make the variable distribution more stable.
6. Train-Test Split
   The dataset is divided into training and testing data using an 80%:20% ratio randomly, ensuring balanced stratification in the data.

4. Results and discussion

4.1. Choosing Data

NO	NAMA RUMAH	HARGA	LB	LT	KT	KM	GRS
1	Rumah Murah Hook Tebet Timur, Tel	3800000000	220	220	3	3	0
2	Rumah Modern di Tebet dekat Stasiu	4600000000	180	137	4	3	2
3	Rumah Mewah 2 Lantai Hanya 3 Me	3000000000	267	250	4	4	4
4	Rumah Baru Tebet, Tebet, Jakarta S	430000000	40	25	2	2	0
5	Rumah Bagus Tebet komp Gudang I	9000000000	400	355	6	5	3
6	Rumah Mewah Modern Murah 3 lant	4970000000	300	154	5	3	3
7	Rumah lama di Tebet, dekat MT Har	2600000000	120	150	3	2	1
8	RUMAH BAGUS KEREN JALAN LE	10500000000	350	247	4	4	0
9	Minimalis Baru Jalan 1 Mobil Akses I	3250000000	125	90	3	3	0
10	Minimalis Baru Jalan 2 Mobil Tebet T	4500000000	250	96	5	4	1

Fig. 2: View a dataset

In the initial stage, the first step was to enable Google Colab to read the dataset being used.

4.2. Prediction Model Testing

During the model testing process, data was separated, and linear regression and decision tree regression models were created. The dataset is processed using Google Colab. The models are trained using the training data to learn the relationship patterns between property features (LB, LT, KT, KM, GRS) and the target house price.

1. Model Linear Regression

```
===== MODEL LINEAR REGRESSION =====
Koefisien: [ 1.22852541e+07  2.32681420e+07 -6.38754626e+08  5.56618859e+08
  2.47983448e+08]
Intercept: -775923911.9377832
```

Fig. 3 : Model Linear Regression

The interpretation of the obtained coefficient values can be summarized as follows: LB (building area) and LT (land area) have positive values because the larger the building and land area, the greater the likelihood that the house price will increase. KM (bathroom) and GRS (garage) also make a significant positive contribution because homes with more bathrooms and garages are generally valued higher. The number of bedrooms (BR) actually has a negative value because in this dataset, an increase in the number of bedrooms is not always directly proportional to an increase in price. This can happen if adding a room disrupts the quality of other spaces (e.g., a cramped living room, poor layout), or due to the strong influence of other variables (land or building area and location). An intercept value of -7.76×10^{8} means that when all variables are zero, the base price prediction starts from that value.location).

This coefficient should not be interpreted literally one by one, but rather indicates that the Linear Regression model attempts to capture the linear pattern between features and price based on the available historical data.

2. Model Decision Tree Regressor

```
===== MODEL DECISION TREE REGRESSOR =====
Depth Tree : 22
Jumlah Leaf: 656
----------------------------------------
```

Fig. 4: Model Decision Tree Regressor

The training output shows that the extremely large depth of the tree (22 levels) and the high number of leaf nodes (656 terminal nodes) indicate that the model is attempting to partition the data in great detail. This makes the decision tree very flexible, but on the other hand, they are highly susceptible to overfitting, which means the model follows the training data too closely, making it less capable of generalizing well to test data.

3. Prediction Results on Test Data

```
===== Prediction Results =====
Nilai aktual    : [np.int64(8900000000), np.int64(6500000000), np.int64(6500000000), np.int64(3700000000), np.int64(18500000000)]
Prediksi LR     : [np.float64(9418764095.641926), np.float64(4391303398.928098), np.float64(7587916795.932466), np.float64(22205053388.947056), np.float64(14867874539.81147)]
Prediksi Tree   : [np.float64(7500000000.0), np.float64(2600000000.0), np.float64(6650000000.0), np.float64(10000000000.0), np.float64(22000000000.0)]
----------------------------------------
```

Fig. 5: Prediction Results

As an illustration, one of the outputs displayed in the first five data points is as follows:
Nilai aktual (y_test):
[8.9×10^{9}, 6.5×10^{9}, 6.5×10^{9}, 3.7×10^{9}, 1.85×10^{9}]

Prediksi Linear Regression (y_pred LR):

[9.42×10^9, 4.39×10^9, 7.59×10^9, 2.22×10^{10}, 1.49×10^{10}]

Prediksi Decision Tree (y_pred DT):
[7.50×10^9, 2.60×10^9, 6.65×10^9, 1.00×10^{10}, 2.20×10^9]

From the calculations, it's clear that the two models aren't always exactly the same as the actual price (which is reasonable in regression), but we need a quantitative metric to assess the overall average error. For some data, linear regression predicts much higher (overestimates), for example, for the 4th and 5th values, which indicates the presence of outliers or specific patterns that are difficult to capture accurately. Decision trees also make significant errors, even tho some values appear closer at first glance, the overall error is greater.

4.3. Evaluation

Model	MAE	MSE	RMSE	R2
Linear Regression	1.980346e+09	1.067573e+19	3.267373e+09	0.771313
Decision Tree	2.186994e+09	2.072894e+19	4.552904e+09	0.555962

Fig. 6 : Test results and scores

Based on the tests carried out, the results assess the effectiveness of two linear regression and decision tree regression algorithms. Four metrik were used in the evaluation: R2, RMSE, MSE, and MAE. The test findings are displayed as follows:

1. Mean Absolute Error (MAE)

Linear regression ≈ 1.98×10^9
Decision tree regressor ≈ 2.19×10^9
On average, linear regression predictions missed the actual price by about 1.98 billion rupiah, while decision trees missed by about 2.19 billion rupiah. This difference of hundreds of millions is quite significant in the context of house prices.

2. Mean Squared Error (MSE)

Linear regression: 1.0675×10^{19}
Decision tree regressor: 2.0729×10^{19}
Because MSE squares the errors, a large value indicates that some predictions are very far from the actual values. The MSE of the decision tree is almost double that of linear regression, which suggests that the decision tree makes large errors more frequently or with a higher magnitude.

3. Root Mean Squared Error (RMSE)

Linear regression: 3.267×10^9
Decision tree regressor: 4.553×10^9
RMSE has the same unit as the price of a house (rupiah). This means that, typically, linear regression makes errors of around 3.26 billion rupiah, while decision trees make errors of around 4.55 billion rupiah. LR: 0.7713
DT: 0.5560

4. R-Squared (R^2)

LR: 0.7713
DT: 0.5560

An R^2 value of 0.7713 means that approximately 77% of the variation in house prices in the test data can be explained by the linear regression model. Meanwhile, the decision tree was only able to explain about 56% of the data variation. This difference of over 20% indicates that linear regression is far more capable of capturing the global patterns in the relationship between features and house prices.

5. Conclusion

In this section you should present the conclusion of the paper. Conclusions must focus on the novelty and exceptional results you acquired. Allow a sufficient space in the article for conclusions. Do not repeat the contents of Introduction or the Abstract. Focus on the essential things of your article.

From the results of applying the house price prediction method in South Jakarta, it can be concluded that the data processing and train-test split process significantly influence the accuracy level of the prediction model. In this study, a dataset of 1010 data points was used, which was then divided into 80% training data and 20% testing data. The test results show that the Linear Regression algorithm is able to provide a better level of accuracy in predicting house prices compared to the Decision Tree Regressor.

The linear regression model achieved an R^2 accuracy value of 0.7713 or 77%, which means that the model is able to explain 77% of the variation in house prices based on the input features. However, the model still produces a considerable prediction error value, as indicated by an MAE of approximately 1.98×10^9 (Rp 1.98 billion) and an RMSE of around 3.26×10^9 (Rp 3.26 billion). Meanwhile, the decision tree regressor shows lower prediction performance with an R^2 value of only 0.5560, indicating a tendency for overfitting and a limited ability to generalize to the test data.

Thus, it can be concluded that linear regression is a recommended model for predicting house prices based on the physical characteristics of the property, and has the potential to be used in a property price appraisal decision support system in South Jakarta.

References

[1] H. Hakim, D. Kamil, and B. Alatas, "Pendekatan Machine Learning untuk Estimasi Harga Rumah dengan Regresi Linier," *ALPHA: Journal of Science and Technology*, vol. 1, no. 1, pp. 18–22, Jan. 2025, doi: 10.70716/alpha.v1i1.99.

[2] I. D. Hartarti, I. A. Septiyani, D. A. Gultom, Y. Hendrian, and S. L. Kinanti, "Prediksi Harga Rumah di Boston Dengan Model Regresi Linear Menggunakan Python," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 2, pp. 4250–4256, Jun. 2025, doi: 10.31004/riggs.v4i2.1210.

[3] A. Pratama, A. Maulana, R. A. Saputra, and U. Kalimantan Barat, "Implementasi Algoritma Linear Regression Untuk Prediksi Harga Rumah di Daerah Tebet," *JIFOTECH (JOURNAL OF INFORMATION TECHNOLOGY*, vol. 05, no. 01, p. 280, doi: https://doi.org/10.46229/jifotech.v5i1.986.

[4] A. Wang, "Factors affected housing prices: taking Boston as an example," *Theoretical and Natural Science*, vol. 42, no. 1, pp. 53–63, Nov. 2024, doi: 10.54254/2753-8818/42/2024CH0213.

[5] A. Wafda, "Integrasi Machine Learning dalam Ritel: Tinjauan Komprehensif tentang Prediksi Harga, Analisis Data Pelanggan, dan Pemanfaatan Media Sosial," *Journal Artificial: Informatika dan Sistem Informasi*, vol. 2, no. 2, pp. 90–106, Oct. 2024, doi: 10.54065/artificial.543.

[6] F. A. Rangkuti, Khairunnisa, and S. Sundari, "IMPLEMENTASI GRADIENT BOOSTING MACHINES UNTUK PREDIKSI HARGA RUMAH PADA JAKARTA SELATAN," *Jurnal Kecerdasan Buatan dan Teknologi Informasi*, vol. 4, no. 2, pp. 164–172, May 2025, doi: 10.69916/jkbti.v4i2.318.

[7] S. Nagula, "Real Estate Price Prediction Using Machine Learning Models," *Int J Res Appl Sci Eng Technol*, vol. 13, no. 7, pp. 157–163, Jul. 2025, doi: 10.22214/ijraset.2025.72962.

[8] Z. Li, "A Comparative Study of Regression Models for Housing Price Prediction," *Transactions on Computer Science and Intelligent Systems Research*, vol. 5, pp. 810–816, Aug. 2024, doi: 10.62051/qjs7y352.

[9] X. Ouyang, "House Price Prediction Based on Machine Learning Models," *Highlights in Science, Engineering and Technology*, vol. 85, pp. 870–878, Mar. 2024, doi: 10.54097/ftyf9665.

[10] "Daftar Harga Rumah." Accessed: Nov. 26, 2025. [Online]. Available: https://www.kaggle.com/datasets/wisnuanggara/daftar-harga-rumah