SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 2, Issue 1, February 2024, Pages 65-71
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v2i1.265

Analysis of K-NN Algorithm and Linear Regression to Predict House Prices in Jabodetabek

Nadia Putri Ariyanti[1], Agung Triayudi[2*], Ratih Titi Komala Sari[3]

Faculty of Communication and Information Technology, Universitas Nasional, Jakarta, Indonesia

Author Email: nadiaaputrii19@gmail.com[1], agungtriayudi@civitas.unas.ac.id[2*], ratih.titi@civitas.unas.ac.id[3]

Abstract. Jabodetabek is now the region with the highest average level of citizen satisfaction, so many people migrate to this region in the hope of getting better living conditions, this will make people who want to buy a house question whether the house they want to buy is good value or not. The purpose of this study is to evaluate the effectiveness of multiple linear regression and K-Nearest Neighbors (KNN) algorithm on a dataset of house prices in Jabodetabek. Better results are obtained by using the Multiple Linear Regression model which has lower Mean Absolute Error (MAE) and Mean Squared Error (MSE) values and a fairly good R-squared of around 48.72%. However, the very high MAE and MSE values of the KNN model indicate inaccuracy and significant prediction variance. Although KNN has a relatively high R-squared value, more research is needed to see if the model can adequately explain data fluctuations. Based on the performance evaluation, multiple linear regression is ultimately a better choice.

Keywords: Jabodetabek. K-Nearest Neighbors (KNN), Multiple Linear Regression, House Prices, Accuracy

1 Introduction

Jabodetabek, also known as Jakarta, Bogor, Depok, Tangerang, and Bekasi, is an Indonesian metropolitan area covering several cities and regions in the provinces of West Java, Banten, and DKI Jakarta. Jabodetabek is now the region with the highest average level of citizen satisfaction in Indonesia. This is due to the weakening economy of Jabodetabek and weakening infrastructure, so many people migrate to this region in the hope of getting better living conditions. [1]

A house is a basic necessity for human life without a place to live, which means experiencing existence without a permanent residence. Not inferior to gold, houses can also be used as a tool for investing in the future due to price movements that change at any time and more and more people who need a house. [2]Then humans are difficult to live their lives well, can be creative, work, gather, raise children, rest and are a shelter from the weather (heat, rain, wind). A dwelling is not only used as a place to live, but many middle and upper class people utilize the dwelling or accommodation as an investment to taste with the benefits of income. Property entrepreneurs will compete to build or buy houses as a means of investment. [3]

This will make people who want to buy a house question whether the house they want to buy is good value or not, seen with the large number of new housing developments with competitive prices. [4] because house prices are increasing day by day

Buying a home is not an easy and straightforward task. Customers will actively seek information, both internal and external, up to the point of final purchase. Customers will always compare the price of the house offered based on their needs and abilities, so they will always adjust the price of the house and the method of payment, whether cash or credit, as well as any discounts or incentives offered by the seller. The price set by the manufacturer will have a negative impact on consumer purchasing decisions; Prices that are too high for consumers will ultimately discourage them from buying the product." [5]

The amount of different information about housing prices makes residents confused in determining the type, type and location of the house that suits their abilities. Some of these factors encourage the need to build a model that is able to simulate housing prices based on people's desires and/or abilities. So that everyone can estimate property prices based on land area, building area and desired location. [6]

Therefore, the use of multiple linear regression analysis and KNN can be an effective method for predicting house prices. The regression line method is a statistical technique used to maximize the correlation between one or more independent variables (factors affecting house prices) and one or more dependent variables (house prices).

Variables that have a significant influence on house prices include electrical power, land area, building area, and the number of bedrooms, bathrooms, and bedrooms. By using regression analysis, we can find out the relationship and correlation between the variables mentioned above and house prices in Jabodetabek. [7]

One method used in simple and effective classification is called the K-NN method. The basic idea behind the KNN algorithm is to find the nearest neighbor among the evaluated data with a number of K neighbors. Test data. KNN operates by comparing training/template data and test data. KNN has a special relationship with very noisy training data sets and is effective when it comes to small/large training data sets. However, to achieve optimal attribute selection, the shortcomings of KNN still require consideration of the value of K. The process of selecting a subset of the original features while excluding features is called irrelevant feature selection. The addition of features will improve efficiency. The main purpose of fitness selection is focus seeking. [8]

2 Methodology

Then at the research method stage that will be carried out is by collecting data [2]. In this study, the dataset used is sourced from the Kaggle site. Jabodetabek House Price List, then the stage carried out is data preprocessing then the data will be classified with the K-NN model and Linear Regression. After being classified by evaluating the model with the confusion Matrix tester method. Next, the two algorithms are compared by comparing accuracy, precision, recall, and f1-score to determine which algorithm or method can achieve the best classification. The process flow is shown in the following figure.

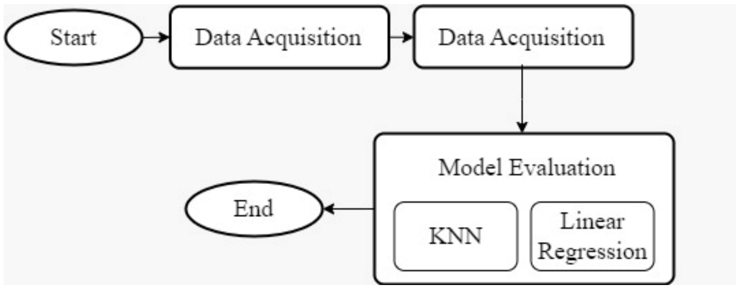


Figure 1. Process Design Flow

2.1 Prediction

Prediction is estimating data values of any type and at any time (past, present, future). There is one term that is similar to prediction, namely forecasting, which estimates the future values of time series data. [9]

2.2 Data Acquisition

In this study, the data used comes from the Kaggle.com site. the dataset used is the House Price List data in the Jabodetabek area which was updated 9 months ago by Nafis Barizki with a total data of ± 3553 data. In the data there are several attributes to identify house price predictions including: City, building area, land area, number of rooms, number of bathrooms, number of floors and year of construction.

2.3 Preprocessing

This step involves preparing the data so that it can be used in the classification process. The three steps are data cleaning, data transformation, and data normalization. prediction by dividing the dataset used is divided into two categories: training data and testing data, with 70% of the dataset used for training and 30% for testing.[10]

2.4 K-Nearest Neighbor

K-NN is a machine learning method in which data is classified based on the majority of its K nearest neighbors. The K value is the number of neighbors considered during the classification process. The K-NN algorithm uses the neighborhood classification as the predictive value of the new test sample.

The operation of the K-NN algorithm requires selecting the training dataset, testing dataset, andx value of K. Then the training dataset is sorted based on the calculation of the closest distance between the testing dataset and the training data. And finally, take the smallest average value of the training dataset according to the number k to determine the regression category. [11]

$$D = \sqrt{(x1 - y1)^2 + (x2 - y2)^2}$$

2.5 Multiple Linear Regression

Multiple Linear Regression is a type of regression model that includes more variables than in one model. Multiple linear regression is also used for forecasting with more than two factors, which can produce the best results between the independent variable and the dependent variable. [12]This is based on the idea that by using data with interval or ratio scales, linear regression models produce more accurate predictions than single models. In addition, this method analyzes what variables are independent and dependent. [13]

Linear regression is able to provide a deeper understanding of the variables that influence an event and can be utilized to gain insight. As for the multiple linear regression calculations used. [6]

$$Y = a + \beta 1\, X1 + \ \beta 2\, Xn + e$$

Description:
y = dependent variable
b = regression coefficient
x = independent variable
ε = error or deviation from predicted value

2.6 Evaluation

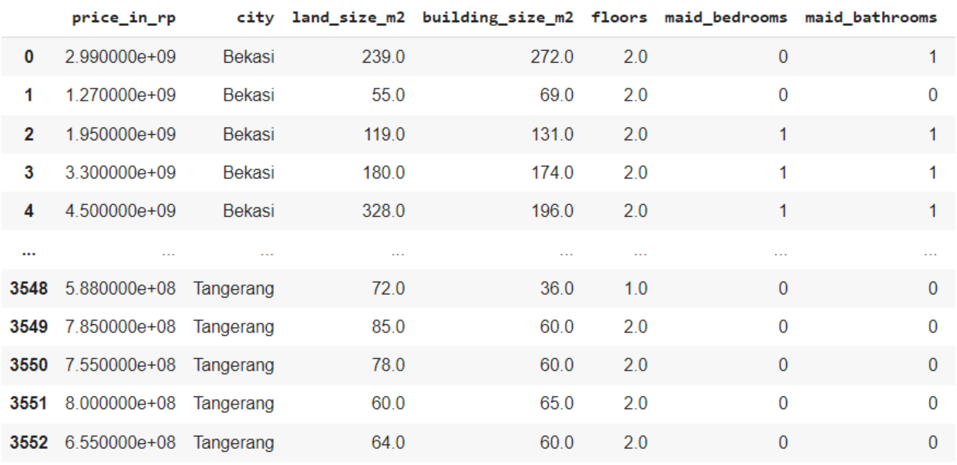
In the evaluation stage, the prediction model is carried out using a testing metric, namely MSE (Mean Squared Error), which measures the average of the squared difference between the predicted value and the actual value.

$$\text{MSE} = \frac{1}{n}\, \Sigma_{i}^{n}\, (y_i - \grave{y}_i)^2$$

n = is the number of observations or samples
y_i = is the true value of the response variable for the i-th observation
$\grave{y}_i$ = is the value predicted by the model for the i-th observation

3 Results and Discussion

The research conducted uses a dataset of house prices in Jabodetabek sourced from Kaggle. At this stage of the research, the data is divided into training data and testing data, with 70% training data and 30% testing data. The dataset used in this study is 5332 data with 8 variables. The variables in this dataset involve important information about house prices, cities, building area, resistant area, number of bedrooms, number of bathrooms and number of floors. As shown in the figure



	price_in_rp	city	land_size_m2	building_size_m2	floors	maid_bedrooms	maid_bathrooms
0	2.990000e+09	Bekasi	239.0	272.0	2.0	0	1
1	1.270000e+09	Bekasi	55.0	69.0	2.0	0	0
2	1.950000e+09	Bekasi	119.0	131.0	2.0	1	1
3	3.300000e+09	Bekasi	180.0	174.0	2.0	1	1
4	4.500000e+09	Bekasi	328.0	196.0	2.0	1	1
...	...	...	...	...	...	...	...
3548	5.880000e+08	Tangerang	72.0	36.0	1.0	0	0
3549	7.850000e+08	Tangerang	85.0	60.0	2.0	0	0
3550	7.550000e+08	Tangerang	78.0	60.0	2.0	0	0
3551	8.000000e+08	Tangerang	60.0	65.0	2.0	0	0
3552	6.550000e+08	Tangerang	64.0	60.0	2.0	0	0

Figure 2. Dataset

3.1 Preprocessing

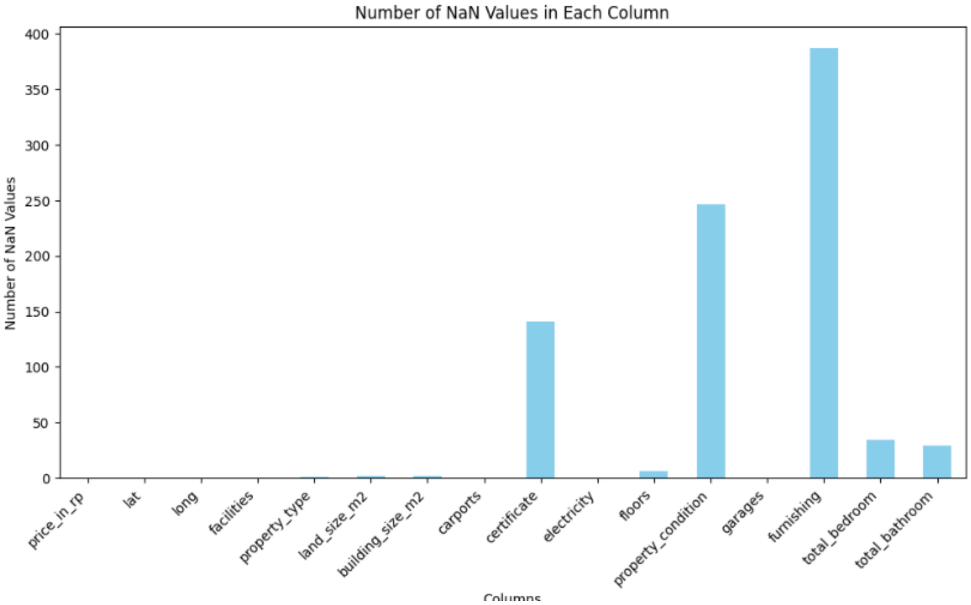


Figure 3. Null Bar Chart

At this stage, the following figure is the process to calculate and visualize the number of NaN (missing values) in each column. In the dataset, there are some data with null values. Figure 2 shows that land area, building area, certificate, floor, property, furniture, number of bedrooms, number of bathrooms.

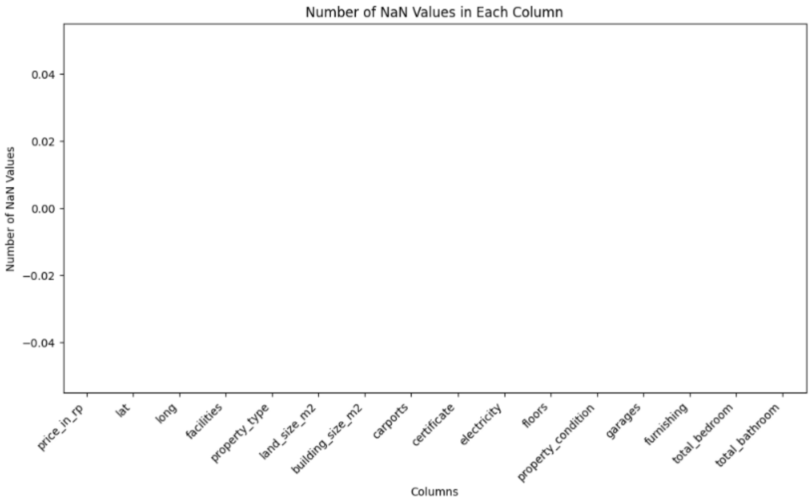


Figure 4. Graph After Missing Value Management

In Figure 3 the dataset has gone through a cleaning process, where missing or invalid values have been resolved. In handling missing values, it can be done by filling in the missing variable values using the average of these variables.

Table 1. feature selection

#	Column	Non-Null Count	Dtype
0	Price	3514 non-null	Float64
1	City	3514 non-null	Int64
2	Land area	3514 non-null	Float64
3	Building area	3514 non-null	Float64
4	Floor	3514 non-null	Float64
5	Bedroom	3514 non-null	Float64
6	Bathroom	3514 non-null	Float64

In table 1 the dataset has gone through a cleaning process, where missing or invalid values have been replaced with valid and consistent values in each column.

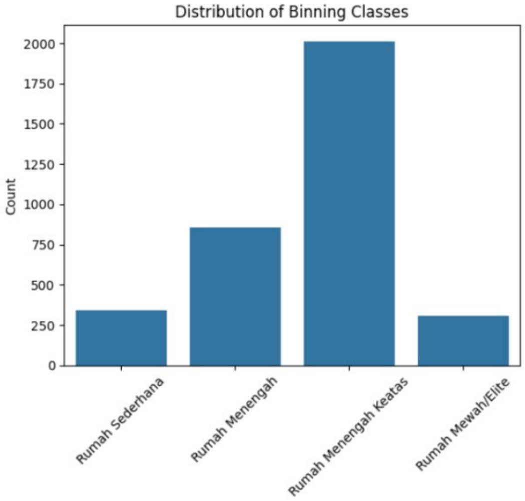


Figure 5. Class Distribution

The data visualization in Figure 4 uses a countplot to illustrate the class distribution. This graph provides a clear picture of the frequency distribution of each class, with each class given a more meaningful label such as "Simple House", "Middle House", Upper Middle House", and "Luxury/Elite House". It can be seen in Figure 4 that the class of "Middle to Upper Middle Houses" is more numerous than the class of "Luxury/Elite Houses". Thus, this visualization is an effective tool to understand the class distribution in the context of house price categories in the dataset.

3.2 Accuracy

Table 2. KNN Accuracy

Mean Absolute Error	Mean Squared Error	R - squared
1436846514.9359887	3.2204762939217633e+19	0.4914562765507773

Based on the table, the Mean Absolute Error (MAE) value is 1436846514.9359887 and the Mean Squared Error value is 3.2204762939217633 and the R - Squared value is 0.4914562765507773. MAE and MSE have very large values and R-squared has a value of around 0.49. large MAE and MSE values can indicate that the model has a very significant prediction error. The R-squared of about 0.49 indicates that the model can explain about 49% of the variability in the data.

Table 3. Linear Regression Accuracy

Mean Absolute Error	Mean Squared Error	R - squared
0.4250157118019093	0.29803757335087777	0.48723472390735667

Based on the table above, the Mean Absolute Error (MAE) value is 0.4250157118019093 and the Mean Squared Error value is 0.29803757335087777 and the R - Squared value is 0.48723472390735667. This model provides the best prediction accuracy relatively well with low MAE and MSE. Despite the improvement from the previous evaluation, there is still a large portion of variation in the target data that cannot be explained by the model.

4 Conclusion

Based on this research, it compares the KNN algorithm and the Regression linear multiple algorithm on the dataset of house prices in Jabodetabek. Based on the evaluation results of performance testing of the K-Nearest Neighbor and Regression Linear Multiple methods. The regression linear multiple model shows better performance with lower mean absolute error (MAE) and Mean Squared Error (MSE) values and R-Squared shows that this model can explain about 48.72% of the variation in the data, which is a relatively good level of explanation. Meanwhile, the KNN model has very high MAE and MSE values, indicating significant inaccuracy and deviation between the predicted and actual values. Although the R-Squared values seem quite high, it needs to be analyzed further to understand how well this model can explain the variations in the dataset. Multiple Linear Regression is a better choice based on the performance evaluation.

References

[1] I. Maula, L. U. Hasanah, and A. Tholib, “ANALISIS PREDIKSI HARGA RUMAH DI JABODETABEK MENGGUNAKAN MULTIPLE LINEAR REGRESSION,” *Jurnal Informatika Kaputama (JIK)*, vol. 7, no. 2, pp. 216–224, Jul. 2023, doi: 10.59697/jik.v7i2.135.

[2] A. Saiful, “Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning Dengan Algoritma Linear Regression,” *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 8, no. 1, pp. 41–50, Mar. 2021, doi: 10.35957/jatisi.v8i1.701.

[3] C. Haryanto, N. Rahaningsih, and F. Muhammad Basysyar, “KOMPARASI ALGORITMA MACHINE LEARNING DALAM MEMPREDIKSI HARGA RUMAH,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 533–539, Mar. 2023, doi: 10.36040/jati.v7i1.6343.

[4] U. Hayadi Umar and R. D. Putra, “Analisis Kenaikan Harga Properti Residensial Sederhana Untuk Wilayah Kelurahan Belian,” 2020.

[5] ST. H. Kadir, "Analisis Perbandingan Metode Artificial Neural Network, Regresi Linear Berganda Dan Regresi Polinomial Untuk Memprediksi Harga Jual Rumah," 2020.

[6] K. Muhammad *et al.*, "Analisis Tren Kepemilikan Rumah di Kota Palembang dan Prediksi Harga Rumah memanfaatkan Machine Learning Analysis of Home Ownership Trends in Palembang City and House Price Prediction with Machine Learning," vol. 8, no. 2, 2023, doi: 10.33772/jpw.v8i2.377.

[7] A. Vermaysha and U. Duta Bangsa Surakarta, "Prediksi Harga Rumah di Kabupaten Karanganyar Menggunakan Metode Regresi Linear Sistem Informasi," 2023.

[8] A. Bode, "K-NEAREST NEIGHBOR DENGAN FEATURE SELECTION MENGGUNAKAN BACKWARD ELIMINATION UNTUK PREDIKSI HARGA KOMODITI KOPI ARABIKA," *ILKOM Jurnal Ilmiah*, vol. 9, no. 2, pp. 188–195, Aug. 2017, doi: 10.33096/ilkom.v9i2.139.188-195.

[9] K. Puteri and A. Silvanie, "MACHINE LEARNING UNTUK MODEL PREDIKSI HARGA SEMBAKO DENGAN METODE REGRESI LINIER BERGANDA 1)," 2020. [Online]. Available: www.data.jakarta.go.id.

[10] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *Journal of Applied Computer Science and Technology*, vol. 4, no. 1, pp. 58–64, Jul. 2023, doi: 10.52158/jacost.v4i1.491.

[11] F. Rozi, M. Bagoes, and S. Junianto, "Penerapan Machine Learning Untuk Prediksi Harga Saham PT.Telekomunikasi Indonesia Tbk Menggunakan Algoritma K-Nearest Neighbors," *Jurnal Informatika MULTI*, vol. 1, no. 1, 2023.

[12] Miftahuljannah, Aswan Supriyadi Sunge, and Ahmad Turmudi Zy, "ANALISIS PREDIKSI PENJUALAN DENGAN METODE REGRESI LINEAR DI PT. EAGLE INDUSTRY INDONESIA," *Jurnal Informatika Teknologi dan Sains (Jinteks)*, vol. 5, no. 3, pp. 398–403, Aug. 2023, doi: 10.51401/jinteks.v5i3.3325.

[13] M. L. Mu'tashim, T. Muhayat, S. A. Damayanti, H. N. Zaki, and R. Wirawan, "Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression," *Informatik : Jurnal Ilmu Komputer*, vol. 17, no. 3, p. 238, Dec. 2021, doi: 10.52958/iftk.v17i3.3635.