JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 6, No. 1. December 2023 DOI: https://doi.org/10.34288/jri.v6i1.262

Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

HOUSE PRICE PREDICTION USING DATA MINING WITH LINEAR REGRESSION AND NEURAL NETWORK ALGORITHMS

Endang Sri Palupi[1]

Teknologi Informasi
Universitas Bina Sarana Informatika
endang.epl@bsi.ac.id

Abstract

The need for housing in big cities is very high because most offices and economic centers are in big cities. Limited land and high demand cause house prices to rise. Many developers build housing on the outskirts of big cities with access to trains and toll roads to make transportation easier. Property developers compete by providing the best prices, various choices of house specifications, ease of the mortgage process, and attractive promotions such as no down payment. A house is a long-term investment whose price increases yearly, so proper analysis is needed to buy a place to live in. Several factors influence the price of a house, including location, land area, building area, building type, and so on. This research aims to create a house price prediction model using the Linear Regression Algorithm and Neural Network so that the results can be useful for property agents in predicting house sales or from the buyer's side in predicting house prices. The results of this research use the Linear Regression Algorithm RMSE 0.775, while the Neural Network Algorithm uses RMSE 0.645. From this research, modeling using the Linear Regression Algorithm has better results. Still, the Linear Regression Algorithm and Neural Network Algorithm have RMSE results that are close to accurate and have small errors.

Keywords: Linear Regression, Neural Network, Prediction.

Abstrak

Kebutuhan rumah untuk tempat tinggal di kota besar sangat tinggi, dikarenakan perkantoran dan pusat ekonomi paling banyak di kota – kota besar. Lahan yang terbatas dan kebutuhan yang tinggi menyebabkan harga rumah menjadi tinggi. Banyak developer yang membangun perumahan dipinggiran kota besar dengan akses kereta dan jalan tol untuk pemermudah transportasi.Para developer property bersaing dengan memberikan harga terbaik, berbagai pilihan spesifikasi rumah, kemudahan proses KPR, dan promo – promo menarik seperti tanpa DP dan lain sebagainya. Rumah merupakan investasi jangka Panjang yang setiap tahun harganya semakin naik, sehingga dibutuhkan analisa yang tepat untuk membeli semua rumah untuk tempat tinggal. Ada beberapa faktor – faktor yang mempengaruhi harga sebuah rumah dari lokasi, luas tanah, luas bangunan, tipe bangunan, dan sebagainya. Penelitian ini bertujuan untuk membuat model prediksi harga rumah menggunakan Algoritma Linear Regression dan Neural Network sehingga hasil penelitian ini dapat berguna bagi agent property dalam memprediksi penjualan rumah atau dari sisi pembeli untuk memprediksi harga rumah. Hasil penelitian ini menggunakan Algoritma Linear Regression RMSE 0.775, sedangkan menggunakan Algoritma Neural Network RMSE 0.645. Dari hasil penelitian tersebut pemodelan menggunakan Algoritma Linear Regression lebih baik hasilnya, tetapi Algoritma Linear Regression dan Algoritma Neural Network mempunyai hasil RMSE yang mendekati akurat dan memiliki kesalahan yang kecil.

Kata kunci: Linear Regression, Neural Network, Prediksi.

INTRODUCTION

Population growth in Indonesia causes the need for housing to increase, especially in large cities and their surroundings. A house is one of the primary needs that must be fulfilled so that humans can survive. The house is a place of refuge and a symbol of prosperity for a family. The homeowner's level of welfare and identity can be described by the shape of the residence or house they own.(Afika & Ariusni, 2019) Apart from being a place to live, a home is also a long-term investment, so many factors must be considered when buying a house. Research shows that house prices will increase by 2.5 percent in September 2023. Research results from 99 Group in October 2023 show the trend of annual house prices that experienced an increase of 2.5 percent in September 2023 compared to the same period last year. (Bayu Saputra, 2023) Factors that influence house prices are the location of the house, accessibility of the house, physical condition of the house, property prices around the house, and completeness of legal documents. (Admin Aesia, 2023)

Before conducting this research, the author also searched a lot of literature as reference material and references from existing journals to look for comparisons and similarities so that they relate to the research results. This is also to avoid duplication of research writing.

The journal is House Price Prediction Using Web Scraping and Machine Learning with Linear Regression Algorithms, written by Andi Saiful and colleagues in 2021. To get a high prediction value, the research was carried out repeatedly using 80% training data and 20% testing with prediction accuracy results of 88%. This research used web scraping to collect data from several house-buying and selling websites, while the author took data from https://www.kaggle.com/datasets/yasserh/housing-prices-dataset. In this research, the author carried out modeling using the Linear Regression Algorithm and Neural Network to compare the prediction results from each algorithm, and the results are better using the Linear Regression Algorithm. The dataset distribution is also the same as the author, namely 80% using training data and 20% test data and the results using the Linear Regression Algorithm RMSE 0.775, while using the Neural Network Algorithm RMSE 0.645. (Saiful, 2021)

In 2021, a journal entitled House Price Prediction Using General Regression Neural Network was written by Evi Febrion Rahayuningtyas and colleagues. The results of this research are in the form of actual data and predicted data, which are visualized using line plots, while the results of the author's study are in the form of curves, the same as the author's prediction data, pictured in the form of line and curve plots in the Rapid Miner Studio Framework. According to the research results of Evi Febrion and colleagues, there is no relationship between house price and house age; houses that are closer to the MRT station and have more shops around will be more expensive. In contrast, in the author's dataset, house prices are largely determined by the area of the house, land area, number of rooms, and neighborhood. The accuracy test and modeling performance test were also carried out using 3 types of evaluation with the results of the three types of assessment, namely the MSE score of 58.72, the RMSE score of 7.66, and the MAE score of 5.99, while the author only carried out tests to get the RMSE. The RMSE results using the General Regression Neural Network look bigger at 7.66.(Rahayuningtyas et al., 2021)

Furthermore, the journal entitled Analysis of House Price Predictions According to Specifications Using Multiple Linear Regression was written in 2021 by Moh Labib and colleagues. This research uses the Multiple Linear Regression Algorithm. It is calculated manually with an accuracy result of 66%, and in the data sample test, it uses 1001 rows of data and 7 columns containing house price data in South Jakarta. Meanwhile, the author uses two algorithms as a comparison, namely Linear Regression and Neural Network, and uses the Rapid Miner Framework by dividing the dataset into 80% training data and 20% test data from a total of 500 datasets with 8 variables, data taken at https://www.kaggle.com /datasets/yasserh/housing-prices-dataset. The results use the Linear Regression Algorithm RMSE 0.775 while using the Neural Network Algorithm RMSE 0.645. (Mu'tashim et al., 2021)

In 2022, Vania Ariyani and colleagues wrote research entitled Performance Comparison of Naive Bayes and K-Nearest Neighbor (KNN) Algorithms for House Price Prediction. This research uses a lazy learning model and has superior performance in accuracy and speed scores on training data with accuracy scores and time required of 0.5714 and 0.0839 seconds using the K-Nearest Neighbor algorithm. Meanwhile, the enthusiastic learning model of 0.4 obtained the highest accuracy score with the fastest data training time of 0.1615 seconds using the Naïve Bayes algorithm. The algorithms used, namely Naive Bayes and K-Nearest Neighbor, are not usually used for prediction. These two algorithms are typically used for data classification, so the results are not good.(Ariyani et al., 2022)

The next research in 2023, entitled Comparison of Machine Learning Algorithms in Predicting House Prices, was written by Cep Haryanto and colleagues. The methods used in this research are multiple linear regression and random forest regression. The testing is the same, with 80% training and 20% testing data. In the multiple linear regression algorithm, an accuracy value of 78.5% was obtained, while in the random forest regression algorithm, an accuracy value of 81.6% was obtained using Python. The difference is that the author uses modeling with the Rapid Miner Framework. Still, the attributes used are almost the same, namely the house area and number of bedrooms, bathrooms, and garages, with 1010 datasets, also taken from kaggle.com. (Haryanto et al., 2023)

Furthermore, research written by Nuraeni Septiani and colleagues in 2023 entitled Application of the K-Means Clustering Algorithm for House Prices in South Jakarta obtained results from 10 cluster groups with the best dbi value, namely 0.129. Meanwhile, the author researched predictions using the Linear Regression Algorithm and Neural Network with the highest RMSE results of 0.775 using the Linear Regression Algorithm. This research uses the Rapid Miner framework but uses different algorithms according to their function for prediction or clustering. The k-means clustering algorithm determines price differences between groups and appropriate prices for each group while using the linear regression and neural network algorithm to predict house prices. (Aji et al., 2023)

RESEARCH METHODS

This research is quantitative. Quantitative analysis carries out systematic investigations to examine a phenomenon by collecting data that can be measured using statistics, mathematics, and computing. Quantitative research aims to develop hypothetical theories that are related to natural wonders. This quantitative research has an important objective regarding measurement. In this research, measurement is the center of the study. (Qotrun A, 2021)

The author carries out data mining predictions using the Linear Regression Algorithm. The advantage of linear regression is that this method is simple and easy to understand but still produces powerful insights. Determining the Strength of a Predictor can identify how strongly a predictor variable (independent variable) influences other variables (dependent variables). (Agus Setiawan, 2023) The author also uses the Neural Network Algorithm to compare house price prediction results in this research. Neural networks try to imitate the structure/architecture and workings of the human brain, so it is hoped that they can and will be able to replace some human jobs. Neural networks are useful for solving problems related to pattern recognition, classification, prediction, and data mining. (Shukla et al., 2010)

The data in this study was taken from https://www.kaggle.com/datasets/yasserh/housing-prices-dataset. It consists of 500 datasets, and the data is divided into 80% of the dataset for training and 20% for testing. Dataset is an informal term that refers to a collection of data. Generally, a dataset contains more than one variable and concerns a particular topic. A dataset is also said to be a collection of data that comes from information from the past and is ready to be managed into new knowledge. (Mustakim, 2022)

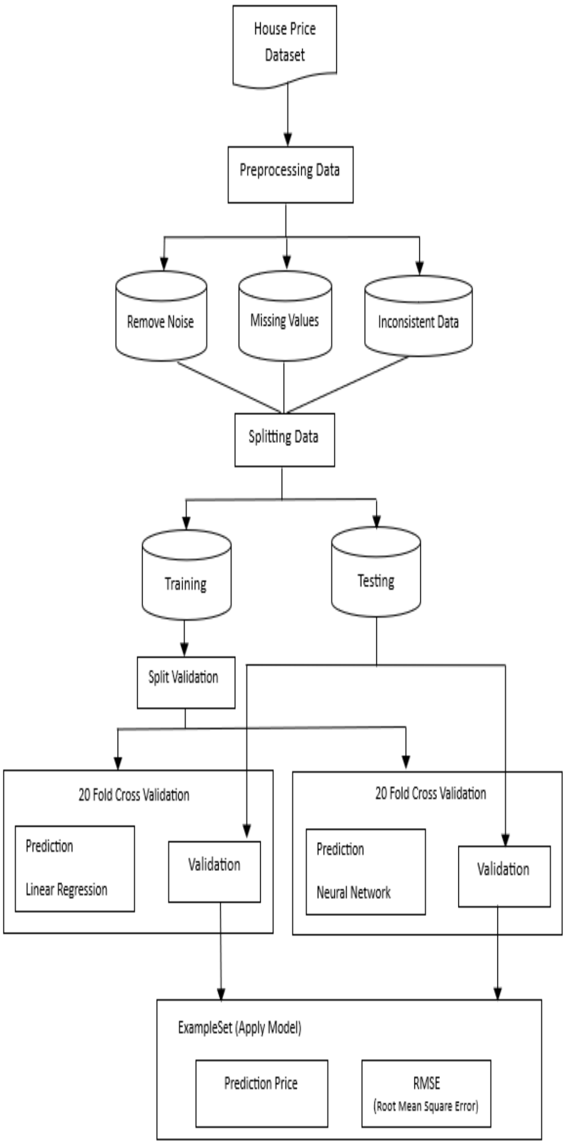

Figure 1. Research Stages

The research stages in Figure 1 are generally as follows:

1. **Data Preprocessing**

Data preprocessing is a set of techniques applied to a database to remove noise, missing values, and inconsistent data. (Galuh Nurvinda, 2021). In this process, we replace missing values to clean up empty data. Then, use Filter Example to select the data used, namely 500 data.

2. **Splitting Data**

Splitting data divides training data and testing data, namely 80% training data and 20% testing data.

3. **Building Model**

Model creation is the most important process, and the author creates modeling using the Rapid Miner framework using the Linear Regression and Neural Network Algorithms to calculate the RMSE (Root Mean Square Error).

4. **Model Evaluation**

At this stage, we compare the RMSE results for each algorithm model used.

RESULTS AND DISCUSSION

Weight by Correlation is a method used to increase the accuracy value of the machine learning model being developed. The preferred feature selection method is Weight by Correlation, which means weighting attributes by connecting (correlating) one attribute with another.(Santoso et al., 2021) The process for selecting weights uses the select-by-weight operator with weight selection parameters above 0.15. This limit was chosen considering that the limit of 0.15 is still close to the sufficient correlation limit (>0.25 – 0.5), and testing will be carried out to see whether the attribute below the limit is enough to influence the model. (Populix, 2023).

Table 1. Weight By Correlation Results

Attribute	**Weight**
furnishingstatus = semi-furnished	0.064
hotwaterheating = no	0.093
hotwaterheating = yes	0.093
basement = no	0.187
basement = yes	0.187
furnishingstatus = furnished	0.229
guestroom = no	0.256
guestroom = yes	0.256
furnishingstatus = unfurnished	0.281
mainroad = yes	0.297
mainroad = no	0.297
environmental safety	0.330
building type	0.330
accessibility	0.366
neighborhood	0.384
parking	0.421
bathrooms	0.433
bedrooms	0.463
land area	0.518
area	0.536

After the attribute weighting process is complete, the data is ready to be processed using a linear regression algorithm using attributes that have been selected into 20 attributes before the 25 attributes are shown in Table 1. There are 5 attributes that are not used: park, year of construction, square feed, basement, and geography. The highest weights are area, land area, bedrooms, bathrooms, parking, and environment. This research uses the 20 attributes above to calculate house price predictions.

Table 2. House price prediction using the Linear Regression Algorithm

Price	Prediction (Price)	Land Area
USD 129.450	USD 125.000	9400
USD 90.010	USD 90.100	8900
USD 130.100	USD 129.000	9350
USD 90.050	USD 90.100	8900
USD 130.000	USD 129.010	9200
USD 99.500	USD 95.040	8900
USD 120.000	USD 195.000	9200
USD 89.200	USD 90.000	8900
USD 99.000	USD 98.050	8900
USD 98.500	USD 98.000	8900
USD 129.450	USD 125.000	9400
USD 130.000	USD 129.010	9200
USD 99.500	USD 95.040	8900
USD 130.000	USD 128.150	9350
USD 98.500	USD 98.000	8900
USD 129.450	USD 125.000	9400
USD 130.000	USD 129.010	9200
USD 99.500	USD 95.040	8900
USD 130.000	USD 129.010	9200

Table 2 shows the results of house price predictions using the Linear Regression Algorithm after modeling using Rapid Miner. The attribute that determines the price of a house is the area of the house. Linear regression is a simple analysis model based on interval and ratio types. Linear regression finds information about independent variables that correlate with the dependent variable. Apart from that, to find out what variables can influence the dependent variable. (Konsultan Data Penelitian & ArcGIS, n.d.)

Table 3. House price prediction using the Neural Network Algorithm

Price	Prediction (Price)	Land Area
USD 129.450	USD 120.000	9400
USD 90.010	USD 95.100	8900
USD 130.100	USD 125.000	9350
USD 90.050	USD 95.100	8900
USD 130.000	USD 125.010	9200
USD 100.000	USD 110.025	9100
USD 99.500	USD 90.040	8900
USD 120.000	USD 198.000	9200
USD 89.200	USD 91.200	8900
USD 10.010	USD 10.900	9100
USD 99.000	USD 98.950	8900
USD 98.500	USD 99.000	8900
USD 129.450	USD 135.000	9400
USD 130.000	USD 133.010	9200
USD 100.000	USD 101.025	9100
USD 99.500	USD 93.040	8900
USD 130.000	USD 127.100	9350
USD 98.500	USD 98.400	8900
USD 129.450	USD 123.100	9400
USD 130.000	USD 127.010	9200
USD 99.500	USD 93.040	8900
USD 130.000	USD 127.210	9200

Table 3 results from house price prediction using the Neural Network Algorithm. With the same land area of the house compared with the house price prediction results using the Linear Regression Algorithm. It can be seen that the prediction results using the Linear Regression Algorithm predict prices that are closer to the original price. Neural networks are useful for decision support systems and data mining. One of the advantages of the neural network is that it is quite good at handling noise in the data it processes. However, the downside is that the performance of the neural network model is difficult for even experts to understand in its application. The algorithm used in the neural network model is very sensitive to the data's format. Most of the output is numerical, so it needs to be redefined in actual calculations. (Larose & Daniel T, 2005)

Table 4. Comparison of RMSE Results

Algoritma	RMSE
Linear Regression	0.775 +/- 0,000
Neural Network	0.645 +/- 0,000

It can be seen in Table 2 that the comparison of the RMSE results of the Linear Regression Algorithm is 0.775, and the Neural Network Algorithm is 0.645. The RMSE results are in a good category and are close to accurate, so they have small errors.(Pertiwi & Indrajit, 2017)

CONCLUSIONS AND SUGGESTIONS

Conclusion

This research aims to create a house price prediction model using the Linear Regression Algorithm and Neural Network so that the results can be useful for property agents in predicting house sales or from the buyer's side to predict house prices. House price predictions using Linear Regression Algorithm modeling and Neural Network Algorithms have good results, are close to accurate, and have small errors with RMSE values for Linear Regression 0.775 and Neural Network 0.645, respectively. Predictions are made using 500 datasets and divided into 80% training and 20% testing data after data preprocessing and the attribute weighting process with parameters below 0.15 to produce 20 attributes from the previous 25 attributes. The research results show that the Linear Regression Algorithm has a higher RMSE value, so it has a small error and is close to accurate.

Suggestion

Future research can use a more varied dataset and other algorithms to predict with optimization so that the accuracy results can be even better.

REFERENCES

Admin Aesia. (2023). *5 Faktor yang Mempengaruhi Harga Jual Rumah*. https://aesia.kemenkeu.go.id/berita-properti/properti/5-faktor-yang-mempengaruhi-harga-jual-rumah-98.html

Afika, Y. A., & Ariusni. (2019). Faktor - Faktor Yang Mempengaruhi Permintaan Rumah Di Indonesia. *Jurnal Kajian Ekonomi Dan Pembangunan*, *1*(Mei), 497–508. https://doi.org/10.1002/ejoc.201200111

Agus Setiawan, T. (2023). Penerapan Linear Regression Pada Estimasi Harga Sewa Alat Berat Pada PT FJB. *Jurnal TEKINKOM*, *6*(1), 135–142. https://doi.org/10.37600/tekinkom.v6i1.733

Aji, B. G., Sondawa, D. C. A., Gifari, M. R., & Wijayanto, S. (2023). Penerapan Algoritma K-Means Untuk Clustering Harga Rumah Di Bandung.

Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Jurnal Ilmiah Informatika Global, *14*(2), 17–23.

Ariyani, V., Putri, P., Prasetijo, A. B., & Eridani, D. (2022). Perbandingan Kinerja Algoritme Naïve Bayes Dan K-Nearest Neighbor (Knn) Untuk Prediksi Harga Rumah. *Jurnal Ilmiah Teknik Elektro*, *4*(4). https://ejournal.undip.ac.id/index.php/transmisi

Bayu Saputra. (2023). *Riset Menunjukkan Tren Harga Rumah Naik 2,5 Persen Pada September 2023*. https://www.antaranews.com/berita/3778617/riset-menunjukkan-tren-harga-rumah-naik-25-persen-pada-september-2023

Galuh Nurvinda. (2021). *Langkah Awal dalam Pemrosesan Data: Data Preprocessing dalam Data Mining*. DQLab. https://dqlab.id/langkah-awal-dalam-pemrosesan-data-dalam-data-mining#:~:text=Data preprocessing merupakan sekumpulan teknik,data transformation%2C dan data reduction.

Haryanto, C., Rahaningsih, N., & Muhammad Basysyar, F. (2023). Komparasi Algoritma Machine Learning Dalam Memprediksi Harga Rumah. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 533–539. https://doi.org/10.36040/jati.v7i1.6343

Konsultan Data Penelitian & ArcGIS. (n.d.). *Mengenal Analisis Regresi Linier dalam Penelitian*. https://patrastatistika.com/analisis-regresi-linear/

Larose, & Daniel T. (2005). *Discovering Knowledge in Data : An Introduction to Data Mining*. JohnWilley's & Sons, Inc.

Mu'tashim, M. L., Muhayat, T., Damayanti, S. A., Zaki, H. N., & Wirawan, R. (2021). Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression. *Informatik : Jurnal Ilmu Komputer*, *17*(3), 238. https://doi.org/10.52958/iftk.v17i3.3635

Mustakim. (2022). *Empat Sumber Dataset untuk Belajar dan Penelitian Bidang Data Mining*. Https://Mustakim.Irpi.or.Id/. https://mustakim.irpi.or.id/2022/05/18/empat-sumber-dataset-untuk-belajar-dan-penelitian-bidang-data-mining/

Pertiwi, M. W., & Indrajit, R. E. (2017). Metode Regresi Linier Untuk Prediksi Pengadaan Inventaris Barang. *Simposium Nasional Ilmu Pengetahuan Dan Teknologi (SIMNASIPTEK)*, 27–30.

Populix. (2023). *Koefisien Korelasi: Pengertian, Rumus, dan Cara Hitungnya*. https://info.populix.co/articles/koefisien-korelasi-adalah/#:~:text=%3E 0 – 0%2C25 %3A,0%2C99 %3A Korelasi sangat kuat

Qotrun A. (2021). *5 Jenis-Jenis Penelitian: Kuantitatif, Kualitatif sampai Campuran*. Gramedia Blog. https://www.gramedia.com/literasi/jenis-jenis-penelitian/

Rahayuningtyas, E. F., Rahayu, F. N., & Azhar, Y. (2021). Prediksi Harga Rumah Menggunakan General Regression Neural Network. *Jurnal Informatika*, *8*(1), 59–66. https://doi.org/10.31294/ji.v8i1.9036

Saiful, A. (2021). Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning Dengan Algoritma Linear Regression. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *8*(1), 41–50. https://doi.org/10.35957/jatisi.v8i1.701

Santoso, I., Gata, W., & Paryanti, A. B. (2021). Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum. *JURNAL RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *1*(10), 5–11.

Shukla, A., Tiwari, R., & Kala, R. (2010). *Real Life Applications of Soft Computing*. Taylor and Francis Group, LLC.