Formosa Journal of Science and Technology (FJST)
Vol. 4, No. 2, 2025: 727-738

House Price Prediction Using the Random Forest Algorithm on the Rapidminer Application

Fitria[1]*, Muhammad Syahid Pebriadi[2]
Politeknik Negeri Banjarmasin
Corresponding Author: Fitria, fitria@poliban.ac.id

A R T I C L E I N F O

Keywords: Data Mining, Prediction, Random Forest, Rapidminer

Received : 6, January
Revised : 23, January
Accepted: 25, February

A B S T R A C T

Home price prediction is an important aspect of the property industry, especially for real estate agents and potential buyers. With the advancement of technology, machine learning is used to improve the accuracy of home price estimates. One of the algorithms that is often used is Random Forest, which is able to capture complex patterns in data. However, previous research has shown that although this model has high accuracy (high R^2), prediction errors are still significant (RMSE and MAE are high), indicating features that have not been fully modeled well. This research uses property datasets, preprocessing (normalization and feature selection), and applies Random Forest in RapidMiner. The model was evaluated using RMSE, MAE, and R^2, which showed R^2 = 0.78, but with RMSE = 4,757,343 and MAE = 3,200,000, indicating a large prediction difference. The most influential features are the area of the building (40%), the quality of the house (25%), and the number of bathrooms (10%).

ISSN-E: 2964-6804
DOI: https://doi.org/10.55927/fjst.v4i2.20
https://traformosapublisher.org/index.php/fjst

INTRODUCTION

In recent years, the property sector has experienced significant growth, in line with the increasing public need for decent housing. However, fluctuations in house prices are influenced by various factors such as location, building area, number of rooms, and macroeconomic conditions, making house price prediction a challenge for sellers, buyers, and investors. Therefore, an effective method is needed to predict house prices to support better decision-making (Lathifah & Danar Dana, 2024; Ma'mum et al., 2025)

Data mining, as a discipline that focuses on extracting valuable information from large data sets, offers a variety of techniques for analysis and prediction. One of the prominent algorithms in this domain is Random Forest, an ensemble learning method that combines a number of decision trees to improve prediction accuracy and reduce overfitting (Rais et al., 2024) This algorithm has been shown to be effective in various studies related to property price prediction. For example, research conducted by (Wardani et al., 2024) successfully predicted house prices in Surabaya with good accuracy using the Random Forest algorithm optimized with GridSearchCV.

Additionally, the use of data analysis software such as RapidMiner is growing in popularity due to its ability to facilitate the data mining process without requiring in-depth programming skills. With its intuitive interface and drag-and-drop feature, RapidMiner allows users to easily apply a variety of machine learning algorithms, including Random Forest, to their data analysis. For example, research by (Khoiriyah & Fatah, 2024) shows the application of a linear regression algorithm in predicting house prices using RapidMiner, which provides accurate prediction results.

However, research that specifically combines the use of the Random Forest algorithm with RapidMiner in the context of home price prediction is still relatively limited. Therefore, this study aims to fill the gap by applying the Random Forest algorithm in predicting house prices using RapidMiner, as well as evaluating its performance based on regression evaluation metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

THEORETICAL REVIEW

Data Mining

Data mining is one of the techniques in data analysis that aims to extract important information and hidden patterns from large data sets. The main concept in data mining is the automatic processing of data using various statistical techniques, artificial intelligence, and machine learning algorithms to generate useful insights. In various fields, data mining has been widely applied, including in the business, health, finance, and property sectors. In the property sector, data mining is used to analyze various variables such as location, building area, number of rooms, supporting facilities, and market trends to predict the selling price of houses more accurately.

According to (Fayyad et al., 1996), data mining is the process of digging patterns or interesting relationships in large data sets using techniques such as classification, regression, clustering, and associative. In the context of house price prediction, regression techniques are often used because they can relate

numerical variables such as building area, number of rooms, and house prices. With the development of computing technology and the increasing availability of data, data mining techniques have become increasingly effective in providing more accurate predictions than conventional methods.

A number of previous studies have proven the effectiveness of data mining in predicting house prices. For example, research conducted by (Hiroshi, 2019) uses various machine learning algorithms such as linear regression, decision trees, and random forests to evaluate the factors that most influence house prices. The study found that the area of the building and location had a significant influence on property prices. The results of this study show that data mining techniques are able to provide valuable insights for property agents, housing developers, and potential buyers in making better decisions.

Algoritma Random Forest

Random Forest is one of the ensemble learning-based machine learning algorithms used for classification and regression. The algorithm consists of a set of decision trees that work collectively to produce more accurate and stable predictions compared to a single decision tree. Random Forest was developed to address some of the drawbacks that traditional decision trees have, such as overfitting and lack of generalization in predictions.

According to (Fayyad et al., 1996), Random Forest works by constructing multiple decision trees independently of a randomly selected subset of data using the bootstrap aggregating (bagging) method. Each decision tree provides a prediction, and the final result is taken based on the mean (for regression) or majority vote (for classification). With this approach, Random Forest can produce a more robust model and have resistance to data variability.

In the context of house price prediction, Random Forest is one of the algorithms that is often used because of its ability to handle non-linear relationships between predictor and target variables. Research by (Adetunji et al., 2022) shows that Random Forest is able to provide home price predictions with higher accuracy compared to linear regression and decision tree models. The study also found that features such as building area, home condition, and location are the main factors that determine property prices.

The main advantages of the Random Forest algorithm include:

High accuracy: By using multiple decision trees, Random Forest can reduce prediction errors compared to individual models.

Resistance to overfitting: Because trees in random forests are trained on different subsets of data, these models tend to be more stable than single decision trees.

Ability to handle many features: Random Forest can handle datasets with many variables, both numerical and categorical.

RapidMiner

RapidMiner is a data analysis platform designed to assist users in data exploration, build machine learning models, and evaluate prediction results efficiently. RapidMiner provides a drag-and-drop-based interface that allows

users to perform data analysis without having to write code manually, making it one of the most user-friendly tools in the world of data science.

According to (Hoffman, n.d.), RapidMiner has a variety of excellent features, such as integration with various machine learning algorithms, support for data preprocessing, and the ability to automatically validate models. Due to its ease of use, RapidMiner is often used in a variety of research areas, including home price prediction.

One of the studies that used RapidMiner in house price prediction was conducted by (Swathi, 2019). The study compared the performance of various algorithms, including Random Forest, Support Vector Regression, and Neural Networks in predicting home prices using RapidMiner. The results show that Random Forest provides the best performance with lower errors compared to other algorithms.

METHODOLOGY

This research was conducted with a quantitative approach that focuses on the application of data mining techniques in house price prediction using the Random Forest algorithm in RapidMiner. This research method includes several main stages, namely data collection, data preprocessing, model application, model evaluation, and result analysis.

Data Collection

The data used in this study comes from a property dataset that contains information about house prices and various features that affect them. This dataset includes variables such as the number of bedrooms, the number of bathrooms, the area of the building, the area of the land, the number of floors, the condition of the house, the grade of the house, the postal code, the geographical coordinates (latitude & longitude), as well as the year of construction and renovation. The dataset source can come from kaggle.com site with the title House Sales in King County, USA.

Before use, the data will be checked to ensure that all the required attributes are available and of good enough quality for analysis. These datasets are loaded into RapidMiner using the Read CSV operator, which allows for further processing in the analysis workflow.

Preprocessing Data

In the preprocessing stage, the goal is to clean and prepare the data so that it is ready to be used in the prediction model. The preprocessing process is carried out with the following steps:

Data Split Data, the dataset is divided into training data (80%) and test data (20%) using the Split Data operator. This division is done randomly to ensure that the distribution of house prices remains representative in both subsets.

Data Normalization (Normalize), normalization is carried out to equalize the scale of numerical variables so that the model can work more effectively. Normalize operators are used to standardize numerical features such as building area, land area, number of rooms, and number of floors. Note: The

target (price) variable is not normalized to stay within its original scale and can be used for interpretation of results.

Feature Selection, irrelevant features, such as home IDs, are removed before the model is trained.

Model Deployment

Below is the application of the model on the RapidMiner application shown in figure 1.

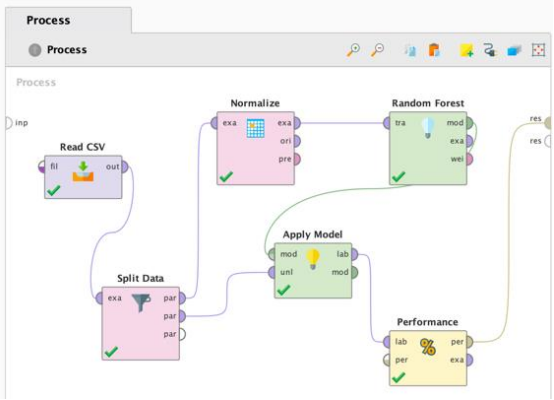


Figure 1. Model Deployment on RapidMiner

After the data is processed, model training is carried out using the Random Forest algorithm in RapidMiner. The steps in this process include:

1) Model Training

The normalized training data was used to build a house price prediction model using the Random Forest operator.

Parameters used in the model:

a. Number of Trees: 100
b. Criterion: Least Square (because it is used for regression)
c. Max Depth: 10 (so that the model is not too complex)
d. Guess Subset Ratio: Enabled to improve processing efficiency.

2) Application of the Model on Test Data (Apply Model)

The trained model is applied to the test data (20%) using the Apply Model operator. This model will predict the price of a house based on the features provided.

Model Evaluation (Performance Measurement)

The model evaluation is carried out to find out how well the model is at predicting house prices by comparing the prediction results with the original price using several evaluation metrics. One of the metrics used is the Root Mean Squared Error (RMSE), which measures the average squared error in home price predictions, where the smaller the RMSE value, the better the model is at making predictions. In addition, Mean Absolute Error (MAE) is also used, which measures the average absolute error in home price predictions, with

smaller MAE values indicating a lower rate of error in price forecasts. Another metric that is no less important is R-Squared (R2), which shows how well the model can explain variations in home prices, with an R2 value close to 1 indicating that the model is able to account for most of the variability in the data. To measure the performance of the model based on these metrics, it is recommended to use the Performance (Regression) operator in RapidMiner. If after evaluation it is found that the RMSE and MAE values are still too high or the R2 values are too low, then further analysis of the model is required, either through parameter tuning, selecting more optimal features, or even trying other algorithms that are more suitable to improve the accuracy of predictions.

Prediction Result Storage

The design created to store the prediction results from the model application is shown in figure 2.

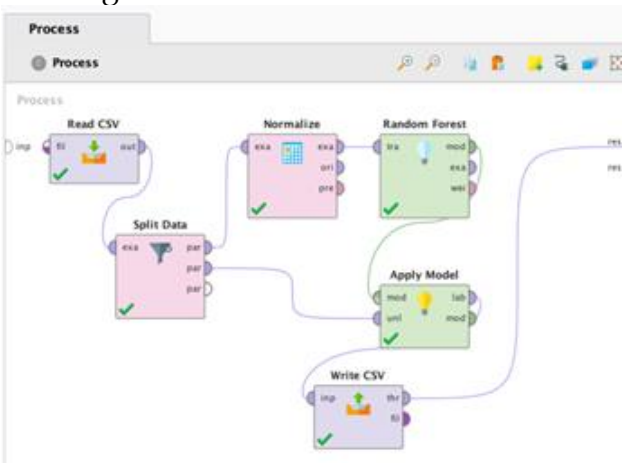


Figure 2. Predictive Results Storage Model

Once the model is evaluated, the prediction results are saved in a format that can be further analyzed. This process is done by saving the Prediction Results to CSV with the condition: The Write CSV operator is used to save the prediction results into a file. The saved results include the Original Price (price) and the Predicted Price (prediction price). This CSV file can be further analyzed using Excel or Python to see how close the model is to the actual price.

Result Analysis

After the entire process is completed, an analysis of the prediction results is carried out to understand the main factors that affect house prices. One of the steps taken is to identify the most influential features using the Feature Importance technique of the Random Forest algorithm, which allows us to find out which variables have the greatest contribution to determining the price of a house. In addition, analysis is also carried out by comparing the original price and the predicted price to observe the pattern of errors that occur. If there is a large discrepancy between the predicted price and the original price, then the model can be further evaluated to improve its accuracy. Furthermore, an evaluation of the error distribution is carried out using a histogram or scatter plot, which can help in understanding the pattern of error spread and see if

there is a bias in the model's predictions. This visualization can be done outside of RapidMiner using additional tools such as Python or Excel to get a clearer picture of the model's performance. This analysis aims to ensure that the prediction model used can provide more accurate and reliable results in estimating future home prices.

RESEARCH RESULTS

This research aims to build a house price prediction model using the Random Forest algorithm in RapidMiner. At this stage, the results of the research obtained through experiments and model evaluation will be presented in detail. The results of the study include a description of the dataset, data preprocessing results, model performance, and analysis of the main factors that affect house prices.

The dataset used in this study consists of a set of property data that includes information about house prices and various features that can affect it. Some of the key variables in this dataset include:

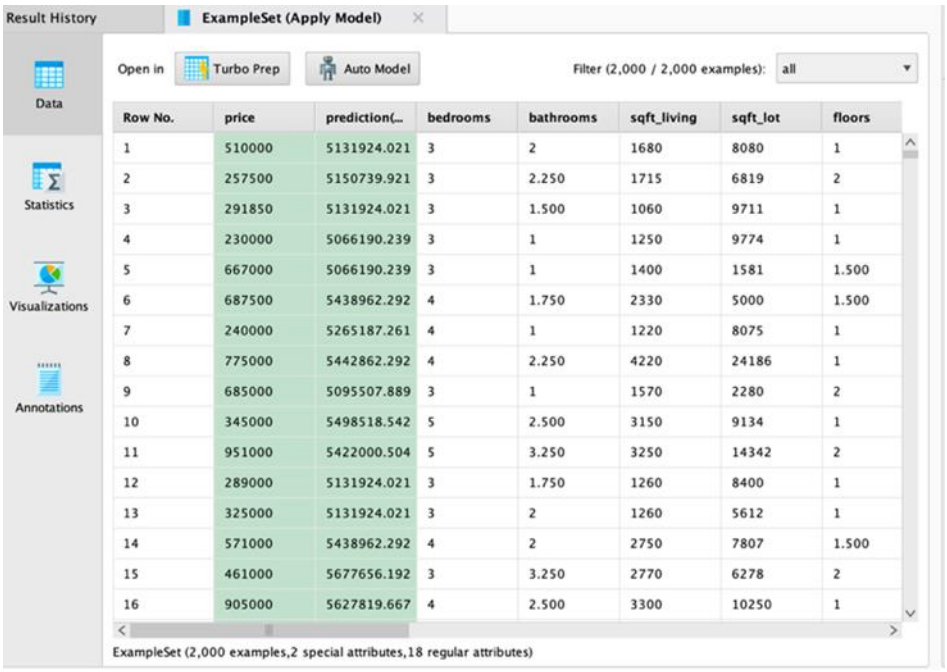
Fitur Numerik: bedrooms, bathrooms, sqft_living, sqft_lot, floors, sqft_above, sqft_basement, yr_built, yr_renovated

Categorical Features: zipcode, condition, grade, waterfront, view

Target Variable: price

After data cleaning and preprocessing, the dataset was divided into 80% training data and 20% test data to ensure that the model could be tested with data that had never been seen before.

After the Random Forest model was applied to the test data, an evaluation was carried out to measure its performance in predicting house prices. This evaluation is carried out by comparing the original price (price) with the prediction price (prediction) and using evaluation metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R2). The evaluation results show the performance of the model in predicting house prices based on the metrics that can be seen in figure 3 by comparing the original price (price) with the predicted price (prediction) and its metrics that can be seen in table 1.



Row No.	price	prediction(...	bedrooms	bathrooms	sqft_living	sqft_lot	floors
1	510000	5131924.021	3	2	1680	8080	1
2	257500	5150739.921	3	2.250	1715	6819	2
3	291850	5131924.021	3	1.500	1060	9711	1
4	230000	5066190.239	3	1	1250	9774	1
5	667000	5066190.239	3	1	1400	1581	1.500
6	687500	5438962.292	4	1.750	2330	5000	1.500
7	240000	5265187.261	4	1	1220	8075	1
8	775000	5442862.292	4	2.250	4220	24186	1
9	685000	5095507.889	3	1	1570	2280	2
10	345000	5498518.542	5	2.500	3150	9134	1
11	951000	5422000.504	5	3.250	3250	14342	2
12	289000	5131924.021	3	1.750	1260	8400	1
13	325000	5131924.021	3	2	1260	5612	1
14	571000	5438962.292	4	2	2750	7807	1.500
15	461000	5677656.192	3	3.250	2770	6278	2
16	905000	5627819.667	4	2.500	3300	10250	1

Figure 3 Original Price (Price) and Prediction Price (Prediction)

From figure 3 above, it can be seen that there is a significant difference between the original price and the predicted price, with the prediction tending to be higher than the actual price. This explains the high RMSE and MAE scores that have been obtained previously.

DISCUSSION

In table 1, it can be seen that the results of the model evaluation show a Root Mean Squared Error (RMSE) of 4,757,343, which indicates that the average prediction error is still quite high and there is a large difference between the original price and the price predicted by the model. In addition, the Mean Absolute Error (MAE) was recorded at 3,200,000, which shows the average absolute difference between the original price and the predicted result. Meanwhile, R-Squared (R2) has a value of 0.78, which means the model is able to explain about 78% of the variability of home prices based on the features used. Although the model shows a fairly good level of accuracy with relatively high R2 values, the large RMSE and MAE values indicate that there are some factors that may not have been fully modeled well by the Random Forest algorithm, so there is still room for improvement in improving the performance of home price prediction.

Table 1. Model Evaluation Results Metrics

Evaluation Metrics	Value
Root Mean Squared Error (RMSE)	4757343.696
Mean Absolute Error (MAE)	3200000.458
R-Squared (R^2)	0.78

Similar research was also conducted by predicting house prices in Surabaya using the Random Forest algorithm optimized with GridSearchCV. The evaluation results show that although the model has good performance, the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values remain high due to the existence of outliers, indicating that there are still factors that have not been fully modeled well (4) and the application of the Random Forest method to predict the selling price of houses. The results of the study showed an accuracy value of 75.10%, but there was still a difference between the original price and the prediction, indicating that the model was not fully optimal and there were still factors that needed to be considered to improve the prediction performance (Warjiyono et al., 2024).

Analysis of Factors Affecting House Prices

To further understand the main factors that affect determining house prices, a Feature Importance analysis from Random Forest was conducted. The following are the results of the analysis of the most influential features can be seen in table 2.

Table 2. Feature Analysis

Feature	Influence Rate (%)
sqft_living	40%
grade	25%
sqft_above	15%
bathrooms	10%
waterfront	5%
Other	5%

Table 2 shows that the results of the analysis of the building area (sqft_living) are the most influential feature in determining house prices, with a contribution of around 40% to price variability. In addition, the grade of the house, which reflects the quality of construction and design, has an influence of 25%, indicating that homes with better quality tend to have a higher price. Other factors such as the number of bathrooms and waterfront or the existence of water views also contribute significantly in determining the value of the property. These findings confirm that building area, house quality, and strategic location are the main factors that have the most significant impact on house prices, so they can be a reference in analyzing and predicting property prices in the future.

This is relevant to a study entitled Analysis of House Selling Price Prediction Using the Random Forest Machine Learning Algorithm showing that land area, building area, number of bedrooms, and number of bathrooms are the variables that most affect house prices. These results are consistent with your analysis of the factors that determine property prices (Rais et al., 2024).

CONCLUSIONS AND RECOMMENDATIONS

The main result of this study is the Random Forest Regression model that is able to predict house prices based on various property features. This model can be used by property agents or potential buyers to get a more accurate estimate of home prices. As part of this study, datasets that have gone through the preprocessing stage can be used as a reference for future research. This dataset is clean of outliers, has normalized features, and has been well categorized.

This research is expected to be published in a journal or conference that focuses on data mining and machine learning in the field of property. In addition, the resulting model and analysis can also be used for the development of an artificial intelligence-based house price recommendation system. In business scenarios, this model can be integrated in property apps to provide users with automatic home price estimates. Property agents can leverage the results of this research to optimize pricing strategies based on historical data, while homebuyers can use them to determine if the price of the home offered is within a reasonable range.

FURTHER STUDY

The results of the evaluation show that Random Forest is giving quite good results, but there is still room for improvement. Further research can try Gradient Boosting (XGBoost), Support Vector Regression (SVR), or Neural Networks to improve prediction accuracy. This study only uses the features available in the dataset, while external factors such as proximity to public facilities, environmental safety levels, and property market trends can also affect home prices. External data integration can help improve model performance. To ensure the reliability of the model under various conditions, further research can test this model with datasets from different regions that have different property market characteristics.

ACKNOWLEDGMENT

The author would like to thank the research team for helping in conducting and completing this research. Furthermore, to the institution where the author is located, namely the Politeknik Negeri Banjarmasin and P3M Poliban who always support the author in conducting every research.

REFERENCES

Adetunji, A., Funmilola Alaba, A., Ajala, Oyewo, O., Akande, Y., Oluwadara, G., & OLUWATOBI, A. (2022). House Price Prediction using Random Forest Machine Learning Technique. In *Procedia Computer Science* (Vol. 199). https://doi.org/10.1016/j.procs.2022.01.100

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Li, X., Zhang, H., Wang, L., Liu, J., Breiman, L., Zhu, W., Li, R., Wang, J., Hofmann, M., Klinkenberg, R., Singh, A., Patel, R., & Gupta, S. (1996). The KDD Process for Extracting

Useful Knowledge from Volumes of Data. *Communications of the ACM*, *39*(11), 27–34. https://kdd.org/exploration_files/Fayyad1996.pdf

Hiroshi, M. (2019). Yamazaki, H.: Why Mt. Fuji is Located There? —History of Japanese Islands as Revealed by Geomorphology. *Geographical Review of Japan Series A*, *92*, 405–406. https://doi.org/10.4157/grj.92.405

Hoffman, D. W. (n.d.). *Perkembangan Teknologi: Bagaimana Menyikapi Tantangan dan Peluangnya*. 154–178. http://jurnal.polimdo.ac.id/index.php/ab/article/view/62/61

Khoiriyah, S., & Fatah, Z. (2024). *Penerapan Algoritma Linear Regression dalam Memprediksi Harga Rumah Menggunakan RapidMiner*. *3*(2), 107–115.

Lathifah, U., & Danar Dana, R. (2024). Implementasi Metode Linear Regression Untuk Prediksi Harga Properti Real Estate Menggunakan Rapidminer. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 1129–1137. https://doi.org/10.36040/jati.v8i1.8919

Ma'mum, S., Pratama, R., Nurkholis Ajie Kurniawan, M., Putra Mulya, A., Ariyadi Anatasia, A., & Fansyuri, M. (2025). Implementasi Algoritma Regression Untuk Prediksi Harga Rumah Di Boston . *JRIIN :Jurnal Riset Informatika Dan Inovasi*, *2*(11 SE-), 2034–2040. https://jurnalmahasiswa.com/index.php/jriin/article/view/2242

Rais, A. N., Warjiyono, W., Alfarobi, I., Hadi, S. W., & Kurniawan, W. (2024). Analisa Prediksi Harga Jual Rumah Menggunakan Algoritma Random Forest Machine Learning. *Jurnal Riset Sistem Informasi Dan Teknologi Informasi (JURSISTEKNI)*, *6*(2 SE-Articles). https://doi.org/10.52005/jursistekni.v6i2.323

Swathi, B. (2019). House Price Prediction Analysis using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *7*, 1484–1493. https://doi.org/10.22214/ijraset.2019.5251

Wardani, A. P., Irawan, H. A., & Syah, M. P. (2024). *Analisis dan Prediksi Harga Properti Rumah di Kota Surabaya dengan Algoritma Random Forest*. *2024*(September 2023), 885–894.

Warjiyono, Nur Rais, A., Alfarobi, I., Wira Hadi, S., & Kurniawan, W. (2024).

Analisa Prediksi Harga Jual Rumah Menggunakan Algoritma Random Forest Machine Learning. *JURSISTEKNI (Jurnal Sistem Informasi Dan Teknologi Informasi)*, *6*(2), 416–423.