

# House Price Prediction Model Using Random Forest in Surabaya City

Rinabi Tanamal, Nathalia Minoque, Trianggoro Wiradinata,  
Yosua Soekamto, Theresia Ratih

*Universitas Ciputra Surabaya, UC Boulevard Citraland, Surabaya, Indonesia*

**Abstract** – A home is one of many fundamental human needs. Therefore, it is essential to arrange so that each family has a separate dwelling. Several prediction algorithms are presented in this study to forecast future property values. By interviewing real estate agents, combining many interviews with multiple agents engaged in the purchasing and selling of homes. Consequently, this study investigates Surabaya Real estate price forecasting models employing Random Forest machine learning algorithms and adopting seventeen regularly used characteristics from real estate agents, which are the most influential factor in determining house prices. The final model may assist in determining the appropriate price for the house. Several research trials have been conducted to achieve a high predictive value; however, the highest predictive value was achieved by using 80% of the data set for training and 20% of the data set for testing to provide output values with an 88% accuracy rate.

**Keywords** – housing price prediction, machine learning, classification, random forest, house sales, Surabaya city.

## 1. Introduction

Home is one of the social needs that cannot be separated from our daily needs.

---

DOI: 10.18421/TEM121-17

<https://doi.org/10.18421/TEM121-17>

**Corresponding author:** Rinabi Tanamal,  
*Universitas Ciputra Surabaya, Surabaya, Indonesia.*


**Email:** [r.tanamal@ciputra.ac.id](mailto:r.tanamal@ciputra.ac.id)

*Received: 17 September 2022.*

*Revised: 25 October 2022.*

*Accepted: 31 October 2022.*

*Published: 27 February 2023.*

 © 2023 Rinabi Tanamal et al; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

Because the house is a basic need, a place to live, a place to rest after tired of daily activities. Not inferior to other investment, even the house can be used as a tool to invest in the future due to price fluctuations. Over time, more and more people have housing needs, especially near workplaces, office centers, shopping centers, close to transportation, and so on, and it will certainly affect house prices quickly. In common practice people use HPI (House Pricing Index) to predict House price. The HPI is an approximate indication derived from all transactions; but it is ineffective for estimating the price of a particular house. Instead, than focusing only on repeat sales from earlier decades, many factors need to be taken into account, including district, age, and the number of floors, etc [10]. Hedonic pricing models are also used to estimate housing values. In these models, the price is influenced by both internal (bedrooms, baths, living space, etc.) and exterior (location) factors (neighboring houses, ZIP code, etc) [4].

Along with its development, Machine learning has recently become a crucial prediction approach since it can forecast property prices more precisely based on their qualities, regardless of the data from past years, because of the growing trend toward Big Data. Numerous studies by Fan et al [5] and Phan [9] investigated this issue and demonstrated the effectiveness of the machine learning strategy. This decade has seen a more rapid expansion of machine learning. Using a machine learning method can also be used for the selection procedure. Machine learning techniques and applications are constantly evolving. House price forecasting is one such use that can be seen in periodicals [13]. House Price prediction is used as a tool for prediction, and it has also been used as an important factor in purchasing decision making. Other researchers used The Support Vector Machine (SVM) method to forecast if a person will be accepted into the Developer Academy [15]. Many prediction model also researched by many researcher such as Shinde and Gawande in “Valuation of house prices using Predictive Techniques” compared the accuracy of their predictions of the sale price of the

homes using lasso, SVR, logistic regression, and decision tree machine learning [12]. Another house price prediction researched by Gerek [6], to acquire the best prediction accuracy, the initial house price prediction is difficult and calls for the finest methodology. Fuzzy logic is one of the methods that can be used to resolve the issue of estimating the sale price of a home with an uncertainty parameter. Other researcher compares fuzzy logic, artificial neural networks, and K-Nearest Neighbours to identify the most suitable approach that can be used as a guide for sellers to determine the price [8]. Predicting House Value with a Memristor-based Artificial Neural Network was done by Wang JJ et al. In order to determine a multivariable regression model with back-propagation formula, they had designed a synthetic neural network supported by memristors [14].

While other algorithm the KNN algorithm, the input data are mapped into predetermined groups or classes using the supervised machine learning process of classification. The primary requirement for using a classification technique is that all data items be classified, and that each data object be classified only once. To categorize data items, distance-based classification algorithms compute the distance between all training samples and the test sample using a distance function. However, distance-based algorithms were initially designed to deal with a certain sort of data, determining the similarity between data items using distance-based measures. Since real-world data sets are frequently diverse in terms of type, format, content, and quality, especially when they are collected from various sources, these algorithms were later created to enable handling of heterogeneous data. There are generally two types of techniques for classifying heterogeneous data when utilizing distance-based algorithms. The first category transforms values from one data type to another (such as binning, interpolating, or projecting data), after which the data can be classified using distance-based algorithms and the proper measurement [3]. To determine its closest neighbours, the primary idea behind k-NN focuses on measuring the distances between the tested and training data samples. The examined sample is then only placed in its closest neighbour's class [7]. K-nearest neighbours (k-NN) is a common and efficient classification technique. When it comes to implementation, it has two main issues: (1) it is a sluggish learning approach, and (2) it depends on the choice of the value of k. This approach has additional drawbacks that are related to the high memory usage that restricts its use [11]. In practice, K typically has an odd value, such as K=1, K=3, K=5, K=7., which is clear to prevent tie. The nearest neighbor categorization rule is most referred to as the K= 1

rule. Using a test set to calculate the classification error rate, one can discover the value of K (Number of Neighbors) by experimenting with various values of K, starting with K=1. The K-value with the lowest error rate is then chosen. To say that the value of K increases with the size of the training tuple. K-NN performs better when using a single instance, regardless of how big the training set is, but only if there is no noise and all characteristics are treated identically [1].

Random Forest is a common supervised machine learning classification and regression algorithm. It leverages ensemble learning to handle complicated issues by combining multiple classifiers to improve the accuracy of the model. Random Forest is a classifier that improves the predicted accuracy of a dataset by averaging the results of many decision trees applied to different subsets of the dataset. Instead of depending on a single decision tree, the random forest uses the projections from each tree to decide the ultimate performance according to the majority of votes [2]. So, to do Real estate price forecasting model using machine learning methods and algorithms. Random Forest Algorithm is the aim of this study. The benefits of the study are to learn and know how to apply machine learning using Random Forest algorithm for predicting house prices. Since this research aim is how to make a prediction model for property price in Surabaya to help determine either house price is under price, over price or in appropriate/normal price.

The organization of this article was structured as follows, in section 1, the Introduction were discussed. This section also explored the current research on machine learning algorithm, particularly the Random Forest. In Section 2, Research Method explain the research processes from data collection to Implementing Random Forest. Meanwhile, the research findings and its discussion were elaborated in section 3. In section 4, finally this paper was concluded.

## 2. Research Method

The price established by the homeowner is classified using a classification model. In this study specially to determine whether it is too under, normal, or over price. The procedures used in this study start with data collection to gather data, followed by data preprocessing to tidy up and get the dataset ready for modeling, model development and comparison to find the best classification algorithm, and model testing to validate the model.

To predict house prices, the research method uses quantitative method, then at the stage of the research method, data collection with combined variables, process the data initially, then apply Random Forest algorithms to data and search results, the processes in figure 1:

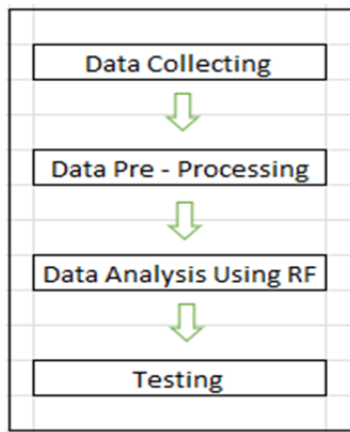


Figure 1. Research processes

### 2.1. Data Collection

In fact, this research uses data collection to predict Real estate pricing is done by interviewing six real estate agencies. Data is collected in July 2022, and collected data was 490 raw data. The dataset includes 17 features that include the factors required to consider home sales. To obtain the class target, interviews were conducted with three property agents who served as experts in categorizing the set price categories, which included underprice, standard or normal price, and overprice. The price category predictions are stored in this Pricing category, which is used as the target class. Price categorization decisions will be made based on the property agent's most numerous categorizations. If each of the three property agents specifies a different category, the data will be dropped, leaving the current dataset with 451 rows, 17 features, and 1 target class. The dataset will be divided into two parts: 80% for training and 20% for testing.

### 2.2. Data Pre-processing

Pre-processing is a step to help the method used produce better output values, this is a step that cannot be skipped in data processing. Proper data cleaning is necessary in data preprocessing to reduce bias. Pre-processing of data is carried out including cleaning data to repair or delete corrupted data or data recommend. Data preparation requires familiarity with the previous dataset. The dataset has been structured and explained in the following section as in table 1.

Table 1. Data pre-processing

ID	Name	Type	Description
1	cluster_name	Feature	The name of the area of a house is located
2	surface_area	Feature	Land area is calculated in square meters

3	building_area	Feature	The building area is calculated in square meters
4	bedrooms	Feature	Number of bedrooms
5	bathrooms	Feature	Number of bathrooms
6	storey	Feature	Number of house levels
7	community price	Feature	House prices usually offered by property agents in rupiah currency
8	price	Feature	The price of the house is set by the owner of the house in rupiah currency
9	ownership_status	Feature	Home ownership status, such as Green Letter (Surat Hijau), Binding Sale and Purchase Agreement (PPJB), Building Use Rights (HGB), Ownership Rights (SHM)
10	facing	Feature	The direction of facing the house, such as north, west, south, and east
11	house_position	Feature	The position of the house against the road, such as skewers, bends / hooks, back skewers / back to the road, back pockets, standard, cul de sac / the end of a dead-end alley.
12	road_width	Feature	The width of the road in front of the house is based on the size of the cars that can pass, such as < 1 car, 1-2 cars, and > 2 cars
13	urgent	Feature	Owner needs money
14	building_age	Feature	The age of the building is measured using an ordinal scale of 1 - 4 years, 5 - 10 years, and > 10 years
15	ready_to_use	Feature	The house is ready for occupancy
16	furnished	Feature	The house is filled with furniture
17	area_category	Feature	The house is in a standard, premium, or very premium area
18	pricing_category	Class target	Categorization of house prices based on 17 existing features, such as under-price, normal price, and Over price

The next step in preprocessing involves encoding certain required attributes. Encoding is done to turn data into numeric data for data processing applications. The features of the dataset include nominal and ordinal data; hence they must be encoded differently. As there is no order or level in nominal data types, encoding is performed by immediately converting data to numeric, such that all data are equivalent. The ordinal data type, on the other hand, contains a sequence or level, requiring encoding that additionally considers the level, beginning with 1 for the lowest level, 2 for one level above, and so on.

Continuing with preprocessing, the correlation between dataset features is examined. Using a heatmap is one technique to observe the relationship. A heatmap was built to examine the correlation between variables, where 0 denotes a negative correlation and  $> 0$  a positive one. The correlation value falls between -1 and 1, as in Figure 2.

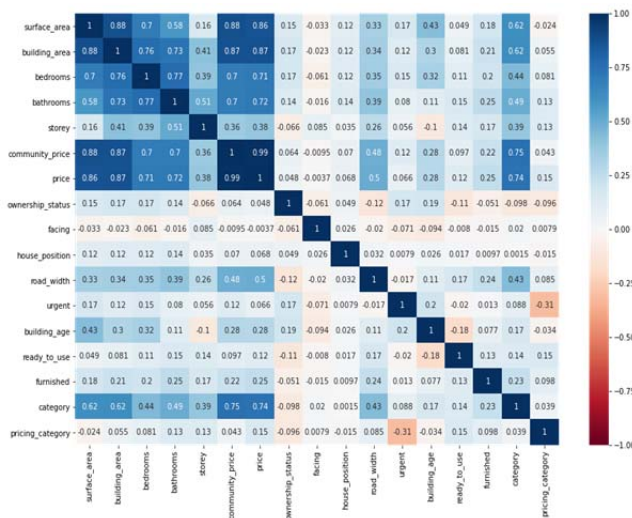


Figure 2. Heatmap relationship

Due to the extremely high correlation between community price and price and community price and category, it would be preferable to eliminate one of them to reduce the variance. After eliminating community prices and categories, the following step is to search the dataset to be processed for missing values. The following table 2 provides an explanation of the missing values within the dataset.

Table 2. Missing values within dataset

Feature Name	Missing Values	Percent of Total Values
building_age	42	8.6%
pricing_category	39	8.0%
urgent	10	2.0%

In the dataset, the building age feature contains 42 missing values (8.6%), the pricing category feature contains 39 missing values (8%), and the urgent feature contains 10 missing values (2%). When

handling missing values for each feature, the value that appears most frequently for each feature is used. This procedure completes the value for each characteristic so that it can be used in subsequent data processing.

### 3. Result

The result will be carried out using both the exploratory data analysis, and model development, evaluation, testing. On the first data analysis stage, the visualization on various house dataset will be shown, such as urgent need to sell house, mostly unfurnished house, mostly house ready to use, the categorization of building age, and pricing category imbalance. On second analysis, explain accuracy and F1 formulation and accuracy comparison on different algorithms.

#### 3.1. Exploratory Data Analysis

This study used exploratory data analysis on existing datasets in order to get additional information. As indicated in Figure 3, up to 66.25 percent of homeowners do not have a pressing need to sell their properties. Figure 4 demonstrates that the majority of residences for sale are unfurnished. The percentage of houses that are unfurnished is relatively high, at 79.39%. On the other hand, the proportion of residences that are ready for occupancy is extremely high, at 92.04% at figure 5.

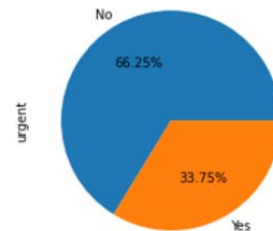


Figure 3. Urgent need to sell house

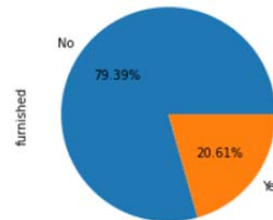


Figure 4. Mostly unfurnished house

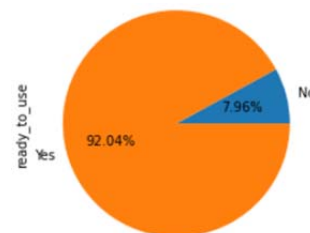


Figure 5. Mostly house ready to use

Figure 6 demonstrates that the average price given is inflated proportionally to the age of the structure. In addition, the number of buildings older than 10 years in Surabaya is significantly lower than the number of buildings younger than 10 years.

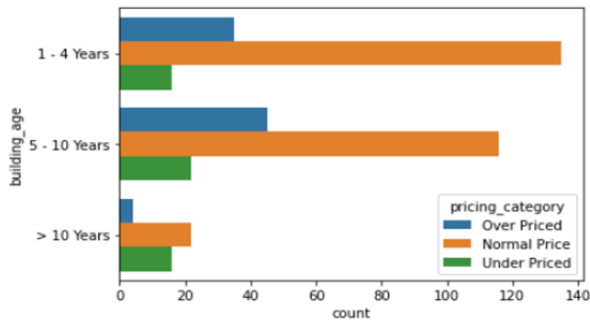


Figure 6. Categorization of building age

Exploratory data analysis was used to do the data analysis. As depicted in figure 7, pricing category, which is the target class in the dataset, contains the imbalance data.



Figure 7. Pricing category imbalance

Normal price has far more data than over price and underprice. Therefore, it is required to use the oversampling approach with SMOTE (Synthetic Minority Oversampling Technique) in order to balance the total amount of data in each category according to the proportion of the majority class. SMOTE generates synthetic data using the K-Nearest Neighbor technique. As seen in Figure 8, the implementation of SMOTE results in a more balanced distribution of data.

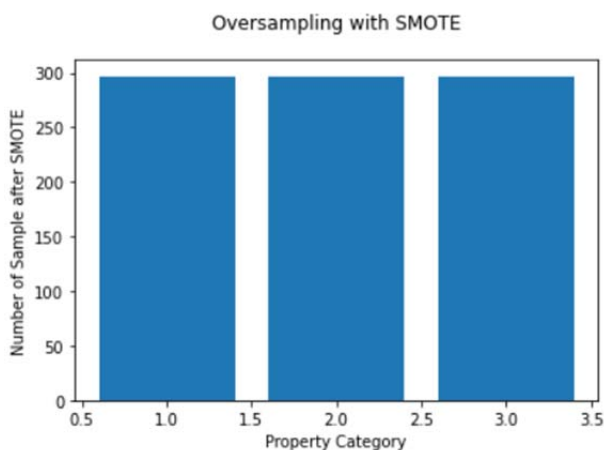


Figure 8. SMOTE implementation

### 3.2. Model Development, Evaluation, and Testing

Eighty percent of the dataset is training data and twenty percent is testing data. The researcher employs five classifiers, K-Nearest Neighbor, Logistics, Support Vector Model, Decision Tree, and Random Forest, to compare accuracy and the greatest F1 score when creating the model. F1 score is the combined average of precision and recall, whereas accuracy is the ratio of true positive to true negative total data. Figure 9. depicts the Accuracy and F1 score formula.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

$$F1 - Score = 2 \cdot \frac{(Precision+Recall)}{(Precision+Recall)}$$

Figure 9. Accuracy and F1 formulation

The random forest classifier has the highest F1 score and accuracy, as shown in the table 3. 88% for F1 score and 0.88 for accuracy. The next stage is to utilize a random forest to validate the model that has been developed.

Table 3. Accuracy comparison on different algorithms

Classifier	F1 Score	Accuracy
K-Nearest Neighbour	0,70	0,70
Logistic	0,65	0,65
Support Vector Model	0,65	0,65
Decision Tree	0,75	0,75
Random Forest	0,88	0,88

When comparing real data with expected data, the Confusion Matrix Heatmap is a useful tool for making the comparison. There was a total of 157 data that were accurately predicted based on the actual data, while there were 22 data that were wrongly anticipated as in figure 10.

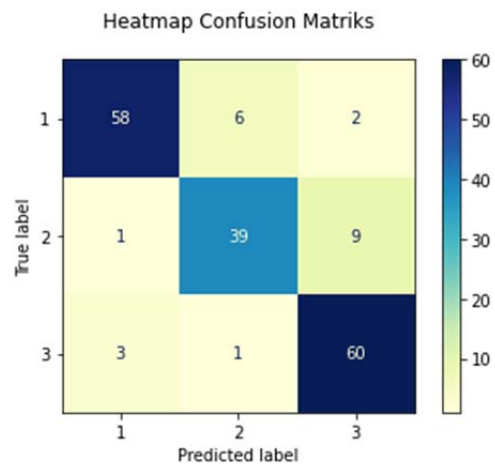


Figure 10. Heatmap matrix

The researcher created a comparison table comparing the actual data and the anticipated data, using 10 different samples of data. In the result column, you will see the value true if the actual data and the projected data are the same, and you will see the value false if they are not the same. This demonstrates that the random forest algorithm reveals accuracy, as the researcher had anticipated it would.

Table 4. Actual vs predicted sample data

No	Actual	Predicted	Result
1.	1	1	True
2.	1	1	True
3.	3	3	True
4.	3	3	True
5.	3	3	True
6.	2	2	True
7.	3	3	True
8.	3	3	True
9.	3	1	False
10.	1	1	True

#### 4. Conclusion

Several steps, including data collection, interviews with real estate agents, and modeling, were required to produce the most accurate machine learning models for assisting homeowners in determining house values. In the exploratory data analysis phase, the dataset reveals that the obtained data is unbalanced. As a result, the researcher employs SMOTE as an oversampling method to achieve data parity. In this modeling, sixteen features related to home sales are utilized, and one pricing group is the target class. Using five classifiers, researchers compared accuracy and the maximum F1 score. The random forest classifier yielded the greatest accuracy and F1 score, 88%, among all other classification methods. With this score, it is envisaged that homeowners will be able to decide whether the previously determined price is too low, average, or high, allowing for optimal property sales.

#### Acknowledgements

The articles of this study are derived from research funded by the Indonesian Ministry of Education, Culture, Research, and Technology.

#### References

- [1]. Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, 12(1), 90-108. <https://doi.org/10.1016/j.aci.2014.10.001>
- [2]. Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science*, 199, 806-813. <https://doi.org/10.1016/j.procs.2022.01.100>
- [3]. Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences*, 1(12), 1-15. <https://doi.org/10.1007/s42452-019-1356-9>
- [4]. Babb, O. (2019). A Comparison of Machine Learning Approaches to Housing Value Estimation. *SIAM Undergraduate Research Online*, 12, 367-380. <https://doi.org/10.1137/18s017296>
- [5]. Fan, C., Cui, Z., & Zhong, X. (2018, February). House prices prediction with machine learning algorithms. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing* (pp. 6-10). <https://doi.org/10.1145/3195106.3195133>
- [6]. Gerek, I. H. (2014). House selling price assessment using two different adaptive neuro-fuzzy techniques. *Automation in Construction*, 41, 33-39.
- [7]. Larose, D. T. (2015). *Data Mining and Predictive Analytics*. John Wiley & Sons.
- [8]. Mukhlisin, M. F., Saputra, R., & Wibowo, A. (2017, November). Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor. In *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 171-176). IEEE. <https://doi.org/10.1109/ICICOS.2017.8276357>
- [9]. Phan, T. D. (2018, December). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In *2018 International conference on machine learning and data engineering (iCMLDE)* (pp. 35-42). IEEE. <https://doi.org/10.1109/iCMLDE.2018.00017>
- [10]. Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174, 433-442. <https://doi.org/10.1016/j.procs.2020.06.111>
- [11]. Salvador-Meneses, J., Ruiz-Chavez, Z., & Garcia-Rodriguez, J. (2019). Compressed k NN: K-Nearest Neighbors with Data Compression. *Entropy*, 21(3), 1-20. <https://doi.org/10.3390/e21030234>
- [12]. Shinde, N., & Gawande, K. (2018). Valuation of house prices using predictive techniques. *Journal of Advances in Electronics Computer Science*, 5(6), 34-40.

- [13]. Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering and Electronic Business*, 11(2), 15-20. <https://doi.org/10.5815/ijieeb.2020.02.03>
- [14]. Wang, J. J., Hu, S. G., Zhan, X. T., Luo, Q., Yu, Q., Liu, Z., ... & Liu, Y. (2018). Predicting house price with a memristor-based artificial neural network. *IEEE Access*, 6, 16523-16528. <https://doi.org/10.1109/ACCESS.2018.2814065>
- [15]. Wiradinata, T., Tanamal, R., Saputri, T. R., & Soekamto, Y. S. (2021, September). An Implementation of Support Vector Machine Classification for Developer Academy Acceptance Prediction Model. In *2021 2nd International Conference on Innovative and Creative Information Technology (ICITech)* (pp. 110-116). IEEE. <https://doi.org/10.1109/ICITech50181.2021.9590146>