International Journal of Advances in Data and Information Systems
Vol. 4, No. 2, October 2023, pp. 155~166
ISSN: 2721-3056, DOI: 10.25008/ijadis.v4i2.1290

Optimization of the Random Forest Method Using Principal Component Analysis to Predict House Prices: A Case Study of House Prices in Malang City

Emha Ahdan Fahmi Elmuna[1], Totok Chamidy[2], Fresy Nugroho[3]
[1,2,3] Faculty of Science and Technology, Program Specification for Master Study in Computer Science,
Universitas Islam Negeri Maulana Malik Ibrahim, Indonesia

Article Info

Article history:

Received May 06, 2023
Revised Jun 12, 2023
Accepted Oct 12, 2023

Keywords:

House Price Prediction
Random Forest Method
Principal Component Analysis
Malang

ABSTRACT

Investment is an interesting thing, especially property investment. The developer must also be careful in determining the price of the property. It should be noted that every year, both short-term and long-term, property prices increase and rarely go down. In determining the price, it is often also based on the features of the house such as the concept, location, bedrooms, etc. To predict house prices based on their features, the random forest has a good performance for predicting house prices. However, the random forest method has the disadvantage that if you use too many variables, the training process will take longer and feature selection tends to select features that are not informative. One way to reduce features without removing other features is to use Principal Component Analysis. In this research, the method used is Principal Component Analysis (PCA) and Random Forest. From the results of model training, it can be concluded that the use of model evaluation results using PCA has a smaller error rate and more consistent values, with an average of 0.018. While the results of the evaluation without PCA and using only Random Forest have a higher error value with an average of 0.03125. The training time using the PCA model has a faster time, with an average of 7918 milliseconds, while those using only random forest without PCA have an average time of 8975 milliseconds.

Corresponding Author:

Emha Ahdan Fahmi Elmuna,
Faculty of Science and Technology,
Program Specification for Master Study in Computer Science (Magister Informatika)
Universitas Islam Negeri Maulana Malik Ibrahim, Indonesia
Jl. Gajayana No.50, Dinoyo, Kec. Lowokwaru, Kota Malang, Jawa Timur 65144
Email: elmuna17@gmail.com

1. INTRODUCTION

Investment is an interesting thing, especially property investment. Since 2011, this property investment has increased both in demand and sales and not only that, according to the Indonesian Statistics Center, in East Java, 50% of the population is classified as young and this young generation will need a home in the future. [1]. It should be noted that every year, both short-term and long-term, property prices increase and rarely even go down [2]. Some countries use the HPI or House Price Index to calculate the increase in house prices [3], but the physical condition, concept,

location, number of bedrooms, bathrooms, and building size also influence determining house prices [1][4].

The random forest method was created by Leo Breimen [5]. This random forest method can be used to solve classification problems [6] or regression problems[7]. Random Forest performs well and has a lower error rate than others in house price prediction [8]–[12]. Shahhosseini et.al. [9] conducted a research prediction on house prices using two types of datasets, namely the Boston and Ames house datasets. His research uses seven algorithms used in making machine learning models. The random forest method obtained has a low error rate of 0.0183. In another study conducted by Ja'far et al. [8] regarding literature review research related to house price prediction. It was found that the best method found was the Random Forest method. Random Forest performs better than others in terms of house prediction.

Random forest has the disadvantage of making a large number of decision trees, this method can provide a high level of accuracy and can avoid overfitting problems, but it has the disadvantage of a long training time [11] because random forest is classified as an ensemble learning concept, namely the concept which averages the results of multiple decision trees applied to the data set to improve accuracy. In addition, randomization on bagging samples and feature selection on random forests tend to select features that are not informative for node separation [13]. This makes the random forest method have poor accuracy when using high-dimensional data.

Feature selection is a step to get good performance [13]. In addition, reducing the number of features in the random forest method can speed up the performance of the model. One method used to reduce features without removing other features is using Principal Component Analysis [14], [15]. The use of PCA and random forest can improve performance to be more effective, and efficient and can provide high accuracy values and low errors [16]–[20]. In another study related to random forest and PCA, Waskle [21] conducted a study related to IDS, or intrusion detection system, to help find attacks on systems and intruder detection, it was found that random forest and PCA have more efficient performance in terms of accuracy compared to other techniques such as SVM, Decision Tree, and Naïve Bayes, with a value performance time of 194400 milliseconds, and has an accuracy of 96.78%, and has an error value of 0.21%. Not only with random forests, but using PCA with other methods can also improve performance to be more effective and proven to improve accuracy [22]–[24].

In another study conducted by Čeh regarding predicting apartment prices using the random forest method combined with the principle component analysis method. Later the performance results from the random forest method will be compared with the commonly used hedonic model based on multiple regression to predict apartment prices. The data set includes 7407 apartment transaction records referring to real estate sales from 2008-2013 in the city of Ljubljana, the capital of Slovenia. From the research results, it was found that random forest and PCA showed much better results for predicting with a MAPE value or an error of 7.27% [20].

From the previous explanation, it is necessary to conduct research that is used to develop a system that can later help property business people to predict property prices that are used to maximize profits. A method is needed to make price predictions based on the features of the property. In this research, the method used is Principal Component Analysis and Random Forest

2. RESEARCH METHOD

This research design describes how the system to be built works. Beginning with collecting data on housing sales in the city of Malang, taken from the results of scraping the rumah.com website. After data collection, the next stage is the preprocessing stage where principal component analysis is applied. Then for system planning use an experimental model with the random forest method. Next is to look at the evaluation results and analyze them and draw conclusions. The flow of the research process can be seen in Figure 1.

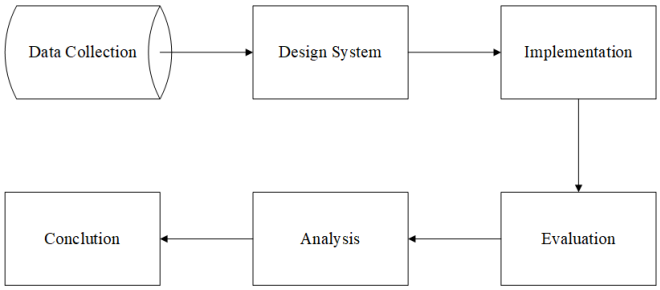

Figure 1. Research Design

In this case, the input used is a property specification, wherein the property specification has several feature variables owned by the property and the price of the property. The method used is the random forest method. Then for optimization, this study used Principal component analysis. Then the output of the system to be built is a prediction of house prices. The system design process is depicted in Figure 2

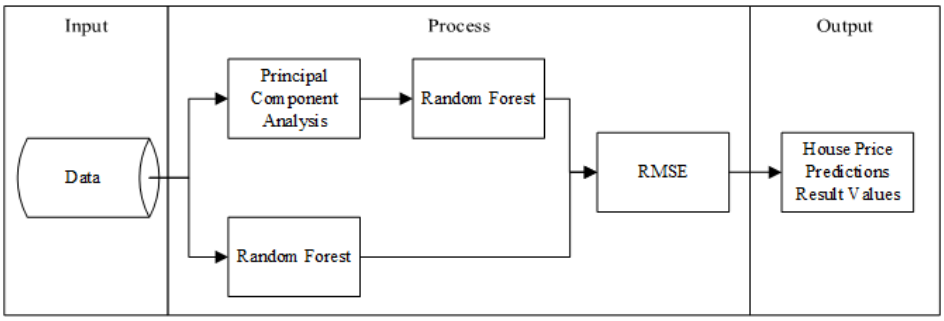

Figure 2. Design System

Before the data enters the modeling stage, the data is first processed to obtain the information contained in the data and processes the data according to modeling standards. There are two stages, namely the first stage is exploratory data analysis and the second stage is the data preparation stage.

The first stage is exploratory data analysis or often called EDA. Exploratory data analysis can provide insights and check for data errors without a direct impact on the model[25]. Exploratory data analysis can also detect errors, find appropriate data, and can also find correlations between variables [26]. In this exploratory data analysis stage, several things will be done, namely variable descriptions, handling outliers and missing values, univariate analysis, and multivariate analysis. Description of variables to identify and describe the properties of the data received by these variables. Then Handle Outliers and Missing Values. Outliers are values that are very different from most of the data in a variable. The missing value is data that is not available or not recorded in a variable. There are several ways to detect outliers, one of which is using the IQR method[27]. IQR stands for Inter Quartile Range. IQR method (Interquartile Range method) is a technique to determine whether a value is an outlier or not. IQR is based on the concept that values that are outside the interquartile range (Q3-Q1) can be categorized as outliers. After detecting outliers, the next step in the EDA stage is univariate analysis, multivariate analysis[28]. Univariate analysis is a data analysis technique that focuses on only one variable. The goal is to understand the characteristics of a variable, such as distribution, central tendency, and variability. After univariate analysis, the next step is multivariate analysis. Multivariate analysis is a data analysis technique that focuses on more than one variable at a time. The goal is to understand the relationships between variables and how one variable affects another.

The second stage is the data preparation stage. The data preparation stage is carried out to carry out the data pre-processing process before the data enters the modeling stage. This stage is important because at this stage the data will be transformed into data suitable for the modeling process. In the data preparation stage, there are several things to do, namely data encoding, then dimension reduction using principal component analysis, data split, and finally normalization. Data

encoding is the stage where categorical data will be converted into numeric data using a one-hot encoding technique. One-Hot encoding is one way to convert categorical data into numeric data and represent categorical values as bit vectors [29]. In this stage, each categorical data will be coded into a dummy feature which is converted to zero or one. Then the next step is dimension reduction with PCA. This PCA optimization is in the data preparation stage. Principe Component Analysis or PCA is a data analysis technique used to reduce the dimensions of a dataset while maintaining information in the dataset. Then split the data to divide the training data and testing data. Then the next stage is normalization. This normalization stage is the stage where the values contained in the data will be changed to a value range of 0-1. this feature transformation process results in an enhanced and continuous data set that machine learning can easily understand [30].

To ensure the method used is good or not, namely by testing it. To evaluate the regression model technically, it only calculates the difference between the actual value and the predicted value, which in this case can be called an error. To test, the matrix that will be used is the RMSE matrix. RMSE measures the square root of the mean squared error of the actual value and the estimated value. If the RMSE value is getting smaller, the performance of the model is getting better [31]. For more details described in the flowchart in Figure 3

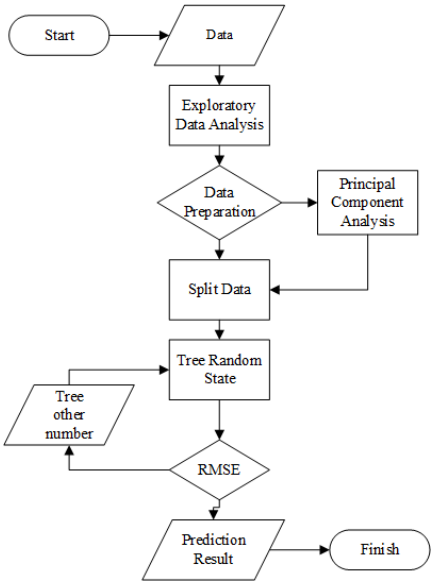

Figure 3. Flowchart Principal Component Analysis and Random Forest Research

3. RESULTS AND DISCUSSION

This section presents the results of the research as well as a comprehensive discussion. Explanation of the steps and results of the analysis. Here is an explanation.

3.1. Exploratory Data Analysis - Variable Description

Data collection in this study uses public data, taken from scraping results from the Malang City housing sales website, namely from the rumah.com website. In the dataset, there are 6130 rows and 11 columns

Table 1. Dataset Type and Description

Variable	Type	Data Type	Description
Number of Bedrooms	Features	Float	Number of bedrooms in the house for sale
Total Bathrooms	Features	Float	Number of bathrooms in the house to be sold
Surface Area	Features	Float	The area of the house in square meters
Price per Meter	Features	Float	The price value of a house is calculated per square meter
Address	Features	Object	Location of the house being sold
Building Area	Features	Float	Building area in square meters
Certificate	Features	Object	Certificate of sale of a house
Interiors	Features	Object	The type of interior that is in the house
Parking	Features	Float	The number of parking spaces in the house
Electricity	Features	Float	The electrical voltage at the house is sold in watts
price	Target	Float	Value of the price of the house being sold

3.2. Exploratory Data Analysis - Handle Missing Values and Outliers

This stage will see which data has a missing value. Following are the results of missing value detection

Bedrooms	0
Bathrooms	0
Surface Area	0
Price per Meter	0
Address	0
Building Area	0
certificate	0
Interiors	2599
Parking	2104
Electricity	909
price	0

Figure 4. Missing Value

From Figure 4 it can be seen that the parking and interior data have a very high missing value, so this variable will be omitted because it will affect the performance of the model. While the electricity variable has a missing value of 909 which will be filled with the average value of the electricity variable.

After checking the missing value, the next step is to detect outliers. Then after checking whether there are outliers or not, to handle outliers here using the IQR method, after the function is run the row that has this outlier value will be deleted.

3.3. Exploratory Data Analysis - Univariate Analysis

This stage is the stage for analyzing the dataset used. In the dataset used, 2 types of categorical data are still used, namely address data and certificates. For numerical features, the target feature will be analyzed, namely the price variable

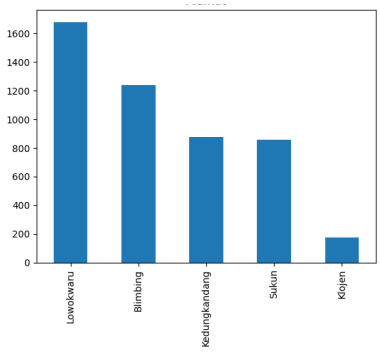

Figure 5. Address Value

From the visualization in Figure 5, it can be concluded that Lowokwaru District has the highest number of house sales, namely 1677 or 34.8% and Klojen District has the least number of house sales, 174 or 3.6%.

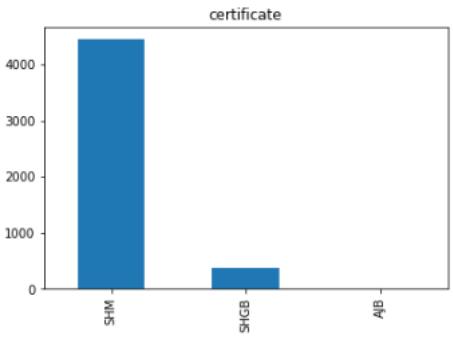

Figure 6. Certificates Value

From the visualization in Figure 6, it can be concluded that houses that have SHM certificates - Freehold Certificates have the highest number, namely 4440 or 92.1%, and AJB - Sale and Purchase Deeds have the smallest number, namely 17 or 0.4%.

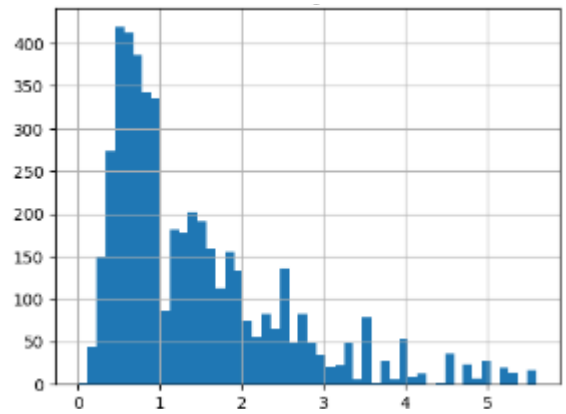

Figure 7. Price Value

From the visualization of Figure 7, some information can be obtained

- The price increase is proportional to the decrease in the number of samples/number of houses. It can be seen from the graph that it has decreased with the number of samples
- The distribution of prices is skewed to the right, indicating that this will influence the model.

3.4. Exploratory Data Analysis - Multivariate Analysis

This stage is the stage to find out the relationship between two or more variables used in the data.

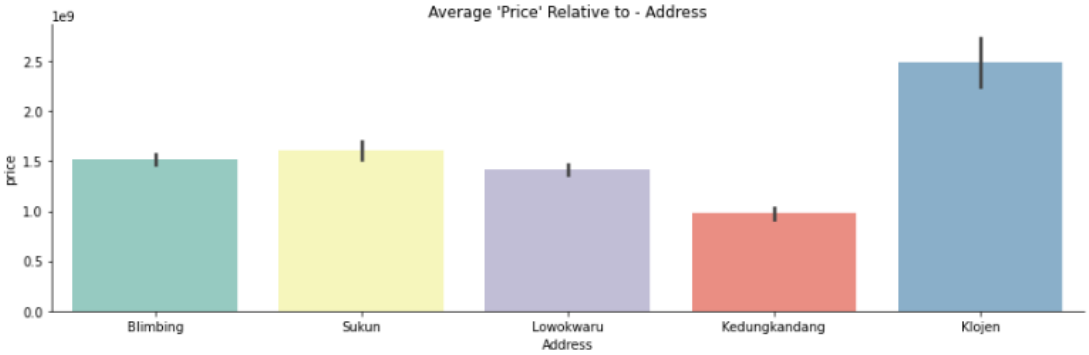

Figure 8. Average "Price" Relative to Address

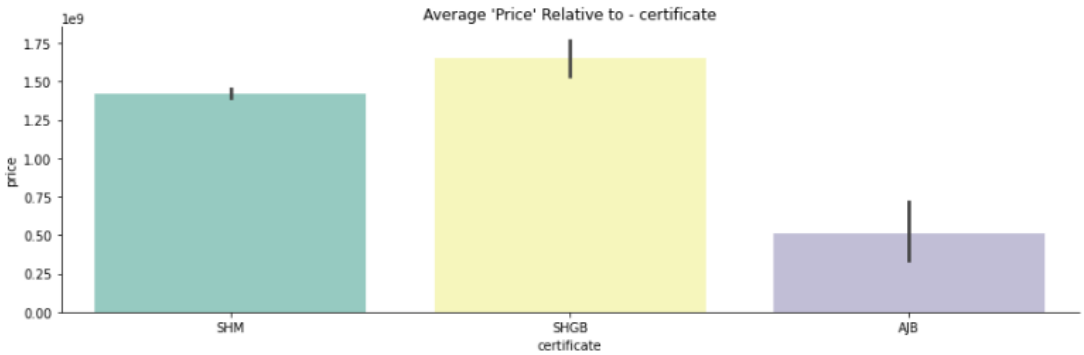

Figure 9. Average "Price" Relative to Certificate

From the visualization of Figures 8 and 9, some information is obtained, namely:

- For the address variable, the average price tends to be similar between the Sukun, Blimbing, and Lowokwaru sub-districts, around IDR 1.5 billion. Whereas those with addresses at Koljen have an average price that tends to be high, around 2.5 billion
- In the certificate variable, the average price for houses that have SHGB and SHM certificates tends to be higher than for houses that have AJB certificates.

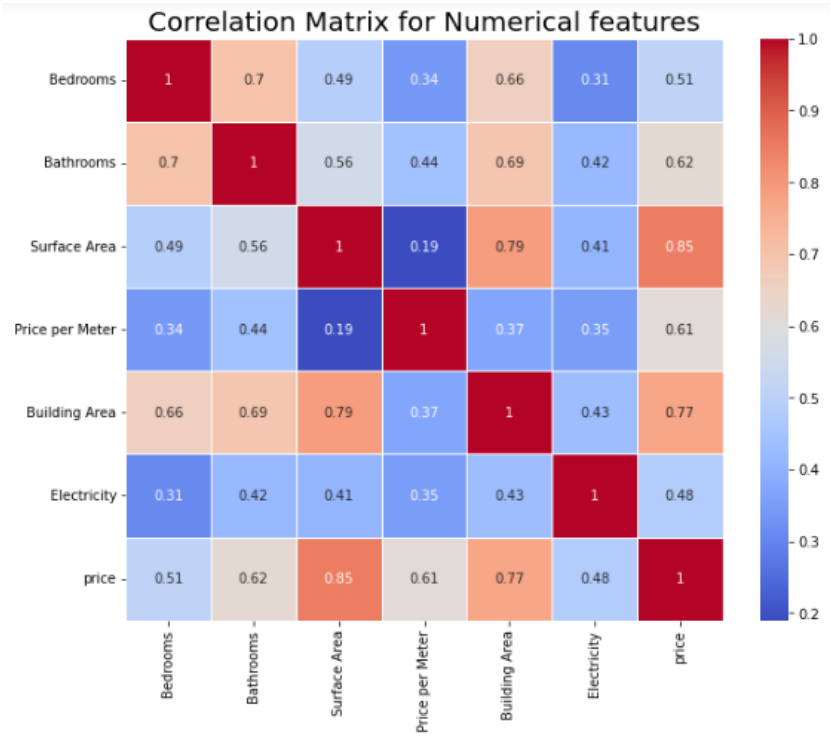

Figure 10. Correlation Matrix for Numerical Features

From the visualization of Figure 10, it can be concluded that variables or features of land area and building area have a high influence on prices, while the electricity variable has a low correlation.

3.5. Data Preparation - One Hot Encoding

This stage is the stage where categorical data will be converted into numeric data. Here it has 2 categorical data namely address and certificate. At this stage, new variables will be created based on the many types of categorical data. In the address variable, there are 5 types and in the certificate variable there are 3 types, so later there will be 8 new variables whose values are 0 and 1 based on the value of these variables.

Table 2. Dataset After One Hot Encoding Process

Variable	Type	Data Type
Number of Bedrooms	Features	Float
Total Bathrooms	Features	Float
Surface Area	Features	Float
Price per Meter	Features	Float
Cut_sukun	Features	Float
Cut_kedungkandang	Features	Float
Cut_klojen	Features	Float
Cut_lowokwaru	Features	Float
Cut_Blimbing	Features	Float
Building Area	Features	Float
Cut_SHGB	Features	Float
Cut_AJB	Features	Float
Cut_SHM	Features	Float
Interiors	Features	Float
Parking	Features	Float
Electricity	Features	Float
price	Target	Float

3.6. Data Preparation - Dimension Reduction with PCA

Dimension reduction techniques are procedures that reduce the number of features while retaining the information in the data. The most popular dimensionality reduction technique is principal component analysis or PCA. It is a technique for reducing size, extracting features, and transforming data from an "n-dimensional space" into a new coordinate system with m dimensions, where m is less than n.

PCA works with linear algebra methods. The data set is assumed to be the most important (primary) with the greatest variance direction. PCA is usually used when the variables in the data are highly correlated. This high correlation indicates that the data is repeatable. Therefore, the PCA technique is used to reduce the original variables into a small number of new variables that are not linearly correlated, which are called principal components (PC). This principal component can capture most of the variance of the original variable. So that when the PCA technique is applied to the data, it only uses the main components and ignores the others.

Below is an explanation of each major component (PC):

- the first principal component represents the direction of greatest variance in the data. It collects the most information from all data characteristics.
- the second principal component captures most of the data left after the first principal component.
- the third principal component collects most of the data left by the first principal component, the second computer, etc.

From the one-hot encoding process, more and more variables are created, which makes the training process longer and less effective. There are 8 new variables, namely Cut_Blimbing, Cut_kenedukandang, Cut_klojen, Cut_lowokwaru, and Cut_sukun which have the same information, namely address, while Cut_SHGB, Cut_AJB, Cut_SHM have the same information, namely certificates. PCA will reduce dimensions, into a new coordinate system. In this case, the new variables will be reduced based on the original variables, namely addresses, and certificates.

3.7. Data Preparation - Normalization

Machine learning will have better performance and will be faster when modeled with data that has the same or close scale. This normalization process will make processing by this machine learning algorithm easier. Normalization is used to scale values to fit within a certain range. Adjusting the range of values is especially important when dealing with different Attribute units and scales. This normalization stage is the stage where the values contained in the data will be changed to a value range of 0-1.

3.8. Modeling

The random forest algorithm is one of the algorithms included in supervised learning. This algorithm can solve classification or regression problems. The Random Forest algorithm is included in ensemble learning. This ensemble learning works with several methods that work together to carry out its performance.

Two results will be described in this study based on the system design made previously, namely the predicted results using only random forests and the results using a combination of PCA and random forests. A comparison of the results between the combination of PCA and random forest and the model without PCA was carried out to see whether the use of PCA and random forest, especially in terms of predicting house prices, could work optimally or not. All stages before entering the modeling stage are the same, but the difference is the application of PCA. PCA is applied at the data preprocessing stage, namely reducing address and certificate features which are broken down into several new variables.

3.9. Evaluation

To ensure the method used is good or not, namely by testing it. To evaluate the regression model technically, it only calculates the difference between the actual value and the predicted

value, which in this case can be called an error. To test, the matrix that will be used is the RMSE matrix, the following is the equation

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(At - Ft)^2}{n}}$$

Where At is the actual value / actual value, Ft is the predicted value, and N is the number of datasets.

Then tested the model with several iterations, this study used 4 iterations namely 100, 500, 1000, and 2000. The following are the results of testing the evaluation of the PCA and random forest models which will be presented in Table 3

Table 3. Result Evaluation

Iterations	PCA + Random Forest		Random Forest	
	Error	Time	Error	Time
100	0,018	755	0,032	875
500	0,018	4095	0,031	4590
1000	0,018	9089	0,031	10779
2000	0,018	17732	0,031	19657
Average	0,0180	7918	0,03125	8975

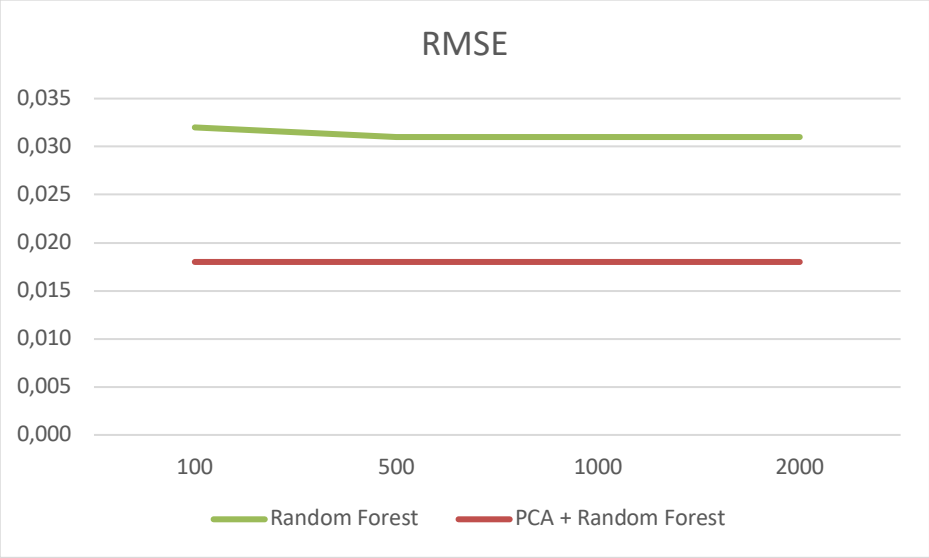

Figure 11. RMSE Value of Testing Model

From Figure 11 it can be seen that the evaluation results use the error value, namely using the RMSE. From the visualization results, there are 4 iterations, namely 100, 500, 1000, and 2000. The error value on the random forest graph at iteration 100 has an error value of 0.032, and iterations 500, 1000, and 2000 have an error value of 0.031. The PCA and random forest graphs at iterations 100, 500, 1000, and 2000 have the same error value of 0.018.

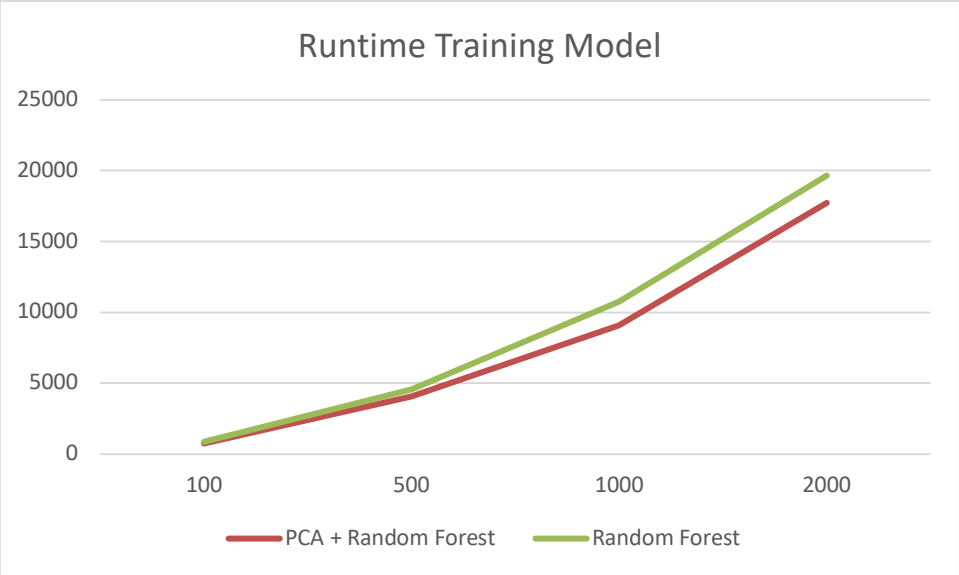

Figure 12. Runtime Training Model

From Figure 12 it can be seen the results of the model training time. From the visualization results, there are 4 iterations, namely 100, 500, 1000, and 2000. The model training time on the random forest graph at iteration 100 takes 875 milliseconds, iteration 500 takes 4590 milliseconds, iteration 1000 takes 10779 milliseconds, and iteration 2000 takes 19657 milliseconds. On the PCA and random forest graphs, 100 iterations take 755 milliseconds, 500 iterations take 4095 milliseconds, 1000 iterations take 9089 milliseconds and 2000 iterations take 17732 milliseconds.

From the results of the training which can be seen in Table 3 and the visualization of Figures 11 and 12, it can be concluded that the use of the evaluation results of models using PCA has a smaller error rate and more consistent values, with an average of 0.018. While the results of the evaluation without PCA and using only Random Forest have a higher error value with an average of 0.03125. The training time using the PCA model has a faster time, with an average of 7918 milliseconds, while those using only random forest without PCA have an average time of 8975 milliseconds.

4. CONCLUSION

The results of the analysis can be concluded that the most sales of houses are in the Lowok Waru sub-district area, and houses that have SHM certificates - Freehold Certificates have high sales. Of the several variables, the variables or features of land area and building area have a high influence on prices, while the electricity variable has a low correlation. Then for model training results it can be concluded that the use of model evaluation results using PCA has a smaller error rate and the value is more consistent with an average of 0.018. While the results of the evaluation without PCA and using only Random Forest have a higher error value with an average of 0.03125. The training time using the PCA model has a faster time, with an average of 7918 milliseconds, while those using only random forest without PCA have an average time of 8975 milliseconds.

ACKNOWLEDGEMENTS

The author would like to thank the lecturers of the Master of Informatics UIN Malang who have guided or provided support regarding this research, both scientific support, information, and facilities.

REFERENCES

[1] A. Nur, R. Ema, H. Taufiq, and W. Firdaus, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study : Malang, East Java, Indonesia," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017, doi: 10.14569/ijacsa.2017.081042.

[2] Y. Feng and K. Jones, "Comparing Multilevel Modelling and Artificial Neural Networks in House Price Prediction," 2015.
[3] C. Garriga, A. Hedlund, Y. Tang, and P. Wang, "Rural-urban migration and house prices in China," *Reg Sci Urban Econ*, vol. 91, Nov. 2021, doi: 10.1016/j.regsciurbeco.2020.103613.
[4] Y. Kang *et al.*, "Understanding house price appreciation using multi-source big geo-data and machine learning," *Land use policy*, vol. 111, Dec. 2021, doi: 10.1016/j.landusepol.2020.104919.
[5] L. Breiman, "Random Forests," 2001.
[6] P. Oskar Gislason, J. Atli Benediktsson, and J. R. Sveinsson, "Random Forest Classification of Multisource Remote Sensing and Geographic Data," 2004. [Online]. Available: http://www.r-project.org
[7] P. Jiang, X. Sun, and Z. Lu, "Quantitative Estimation of siRNAs Gene Silencing Capability by Random Forest Regression Model," 2007.
[8] N. Shahirah, J. ' Afar, J. Mohamad, and S. Ismail, "MACHINE LEARNING FOR PROPERTY PRICE PREDICTION AND PRICE VALUATION: A SYSTEMATIC LITERATURE REVIEW," 2021.
[9] M. Shahhosseini, G. Hu, and H. Pham, "Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction," in *Springer Proceedings in Business and Economics*, Springer Science and Business Media B.V., 2020, pp. 87–97. doi: 10.1007/978-3-030-30967-1_9.
[10] T. Wiradinata, F. Graciella, R. Tanamal, Y. S. Soekamto, T. Ratih, and D. Saputri, "POST-PANDEMIC ANALYSIS OF HOUSE PRICE PREDICTION IN SURABAYA: A MACHINE LEARNING APPROACH," *Xinan Jiaotong Daxue Xuebao/Journal of Southwest Jiaotong University*, vol. 57, no. 5, pp. 562–573, Oct. 2022, doi: 10.35741/issn.0258-2724.57.5.45.
[11] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 806–813. doi: 10.1016/j.procs.2022.01.100.
[12] N. Sharma, Y. Arora, P. Makkar, V. Sharma, and H. Gupta, "Real Estate Price's Forecasting Through Predictive Modelling," in *Lecture Notes in Networks and Systems*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 589–597. doi: 10.1007/978-981-15-7106-0_58.
[13] T. T. Nguyen, J. Z. Huang, and T. T. Nguyen, "Unbiased feature selection in learning random forests for high-dimensional data," *Scientific World Journal*, vol. 2015, 2015, doi: 10.1155/2015/471371.
[14] C. Gardner and D. C. T. Lo, "PCA embedded random forest," in *Conference Proceedings - IEEE SOUTHEASTCON*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021. doi: 10.1109/SoutheastCon45413.2021.9401949.
[15] D. Festa *et al.*, "Unsupervised detection of InSAR time series patterns based on PCA and K-means clustering," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, Apr. 2023, doi: 10.1016/j.jag.2023.103276.
[16] S. Lu, Q. Li, H. Yu, and X. Wang, "Damage Evaluation Method of CFRP Structures Based on PCA and Random Forest Algorithm," in *Proceedings - 2020 Chinese Automation Congress, CAC 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 3804–3807. doi: 10.1109/CAC51589.2020.9327009.
[17] Q. Song and Y. Huang, "A Solution for Liquor Recognition Based on PCA-RF and Laser Induced Fluorescence," *IEEE Access*, vol. 9, pp. 35101–35108, 2021, doi: 10.1109/ACCESS.2021.3049941.
[18] Subhash Waskle, Lokesh Parashar, and Upendra Singh, *Intrusion Detection System Using PCA with Random Forest Approach*. Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)IEEE Xplore, 2020.
[19] C. Gardner and D. C. T. Lo, "PCA embedded random forest," in *Conference Proceedings - IEEE SOUTHEASTCON*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021. doi: 10.1109/SoutheastCon45413.2021.9401949.
[20] M. Čeh, M. Kilibarda, A. Lisec, and B. Bajat, "Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments," *ISPRS Int J Geoinf*, vol. 7, no. 5, May 2018, doi: 10.3390/ijgi7050168.
[21] Institute of Electrical and Electronics Engineers and Hindusthan Institute of Technology, *Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) : 02-04, July 2020*.
[22] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Inform Med Unlocked*, vol. 19, Jan. 2020, doi: 10.1016/j.imu.2020.100330.

[23] M. Mrówczyńska, J. Sztubecki, and A. Greinert, "Compression of results of geodetic displacement measurements using the PCA method and neural networks," *Measurement (Lond)*, vol. 158, Jul. 2020, doi: 10.1016/j.measurement.2020.107693.
[24] Q. Xiong, H. Xiong, Q. Kong, X. Ni, Y. Li, and C. Yuan, "Machine learning-driven seismic failure mode identification of reinforced concrete shear walls based on PCA feature extraction," *Structures*, vol. 44, pp. 1429–1442, Oct. 2022, doi: 10.1016/J.ISTRUC.2022.08.089.
[25] D. J. Butts, N. E. Thompson, S. A. Christensen, D. M. Williams, and M. S. Murillo, "Data-driven agent-based model building for animal movement through Exploratory Data Analysis," *Ecol Modell*, vol. 470, p. 110001, Aug. 2022, doi: 10.1016/J.ECOLMODEL.2022.110001.
[26] R. Indrakumari, T. Poongodi, and S. R. Jena, "Heart Disease Prediction using Exploratory Data Analysis," *Procedia Comput Sci*, vol. 173, pp. 130–139, Jan. 2020, doi: 10.1016/J.PROCS.2020.06.017.
[27] P. Chakri, S. Pratap, Lakshay, and S. K. Gouda, "An exploratory data analysis approach for analyzing financial accounting data using machine learning," *Decision Analytics Journal*, vol. 7, p. 100212, Jun. 2023, doi: 10.1016/J.DAJOUR.2023.100212.
[28] D. Chowdhury, S. Hovda, and B. Lund, "Analysis of cuttings concentration experimental data using exploratory data analysis," *Geoenergy Science and Engineering*, vol. 221, p. 111254, Feb. 2023, doi: 10.1016/J.PETROL.2022.111254.
[29] I. Erjavac, D. Kalafatovic, and G. Mauša, "Coupled encoding methods for antimicrobial peptide prediction: How sensitive is a highly accurate model?," *Artificial Intelligence in the Life Sciences*, vol. 2, p. 100034, Dec. 2022, doi: 10.1016/J.AILSCI.2022.100034.
[30] K. Ogunsina, I. Bilionis, and D. DeLaurentis, "Exploratory data analysis for airline disruption management," *Machine Learning with Applications*, vol. 6, p. 100102, Dec. 2021, doi: 10.1016/J.MLWA.2021.100102.
[31] H. N. Bhandari, B. Rimal, N. R. Pokhrel, R. Rimal, K. R. Dahal, and R. K. C. Khatri, "Predicting stock market index using LSTM," *Machine Learning with Applications*, vol. 9, p. 100320, Sep. 2022, doi: 10.1016/J.MLWA.2022.100320.