



Comparative Analysis of Multiple – Linear Regression Algorithm with Random Forest Regression for Prediction of House Plot Prices

Sigit Auliana¹, Basuki Rakhim Setya Permana^{2*}, Gagah Dwiki Putra Aryono¹

¹Program Studi Sistem Informasi, Universitas Bina Bangsa, Indonesia

²Program Studi Ilmu Komputer, Universitas Bina Bangsa, Indonesia

Corresponding author email: basukirakhim@gmail.com

Article Info

Article history:

Received May 15, 2024

Approved June 17, 2024

Keywords:

Multiple - Linear
Regression, Random
Forest Regression
Prediction,
House Prices

ABSTRACT

Humans basically have a basic need to have a place to live, which can be a house or shelter. Along with the rapid population growth in Indonesia, which continues to increase every year, many people do not have or have a decent place to live. Therefore, careful planning is needed so that every family can have a decent home. One very important aspect in planning investment in the form of property is predicting future house prices. One approach that can be used is to use a Random Forest and Multiple Linear Regression algorithm, which is an algorithm from Machine Learning. There are several factors that can influence the price of a house, including land area, building area, number of bedrooms, bathrooms and garage. In this research, multiple linear regression and random forest regression methods were chosen. The aim of this research is to find the best prediction results between the two methods. To achieve accurate predictions, research was carried out repeatedly by dividing the dataset into 80% for training and 20% for testing. The research results show that the random forest regression algorithm provides the best results, with an accuracy of 81.6%.

Copyright © 2024, The Author(s).

This is an open access article under the CC-BY-SA license



How to cite: Auliana, S., Permana, B. R. S., & Aryono, G. D. P. (2024). Comparative Analysis of Multiple – Linear Regression Algorithm with Random Forest Regression for Prediction of House Plot Prices. *Jurnal Ilmiah Global Education*, 5(2), 1740–1750. <https://doi.org/10.55681/jige.v5i2.2794>

INTRODUCTION

Home, as one of our basic needs, has an important role as a place to live, a place to relax, to gather with relatives, a place to take shelter, and a place to rest after daily activities. Like gold investment, a house has potential as an investment in the future because its value can change and increase along with demand growth. Especially if the location of the house is strategic, such as a school, office building, shopping center and transportation facilities, this will significantly influence the price of the housing unit (Athiyah, at all (2021)).

Apart from being a place to live and settle, the house also acts as a place to rest, stopover for relatives, guests or visitors who travel from far away (Amalia, at all (2021)). In some cases, tourist trips take days, requiring accommodation in the form of temporary housing.

With the evolution of people's needs, especially in terms of housing, property developers are also competing to build houses or housing units or buy houses as a form of investment (Azhar, 2021). This growth raises questions for potential buyers regarding the potential profits from this investment, considering the upward trend in house prices over time (Ridho, et al. (2022)).

Prediction technique is a method used to estimate what the value will be in the future by considering current and future historical data (Hanifi et al., 2020) (Saputri, EA, & Ekojono, E. (2018)). The ability to predict with a fairly high level of accuracy is very valuable for companies or agencies in decision making. Another study regarding house price predictions around the district and city of Bandung using the Moving-Average algorithm model, shows that the model has an accuracy rate of between 70 and 90 percent, with an error rate of around 10 to 20 percent (Sanusi, et al. (2020)).

House prices continue to increase every year, and this increase can be monitored through various aspects or features or facilities owned (Mu'tashim, et al. (2021)). Because house prices can change quickly and are difficult to predict accurately, buyers are considered to need a system to predict prices based on the facilities they have (Saiful, A. (2021)). This research uses regression algorithms, namely multiple linear regression and random forest regression, to calculate house prices by analyzing data and building machine learning models for the purpose of predicting house prices

METHODS

In this design, quantitative descriptive methods are used. The quantitative approach in this research involves systematically exploring conditions or situations with the aim of solving problems based on available data and facts. The data used is quantitative data, which consists of numbers and can be manipulated using mathematical operations. This data was obtained from available factual sources, especially from the website <http://www.kaggle.com>.

1. Data collection

a) Observation

It is a research process that involves retrieving information related to research problems originating from *public data* obtained from the site platform <http://www.kaggle.com>.

b) Literature review

This method involves the use of reliable references as they are important for the preparation of a particular research report.

c) Documentation

The use of literature reviews has become common practice as a source of information in research, which is not only used to explore data, but also to interpret results and make predictions.

2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the first step in research which aims to find patterns, identify outliers, test hypotheses, and verify assumptions. EDA has significant value in detecting errors early on, as it allows users to find anomalies in data, understand relationships between variables, and extract key factors from datasets. The EDA process has an important role in statistical analysis.

Figure 1. Exploratory Data Analysis Research Flowmap



In this case, there are several techniques that can be used for data processing:

3. Pre-processing

Pre-processing is the stage where data is processed to correct or clean errors, so that the data can be used properly. At the pre-processing stage, the initial data used is still in raw form, then treated to suit the desired analysis needs.

4. Exploratory Data Analysis

Exploratory Data Analysis is a method for exploring and examining existing data and determining the necessary data processing steps. At this stage, checking for empty data is carried out, identifying and deleting duplicate data, and converting to categorical data types.

5. Representation

In this step, data undergoes transformation from raw form to visual representation. Data can be translated into various visualizations such as boxplots, histograms, and other visual formats.

6. Modelling

Modeling is the implementation of an algorithm model, where data will be processed further to draw conclusions. This process can produce a variety of results depending on the characteristics of the data used.

7. Deployment/Evaluation

In this stage, a conclusion is drawn which is the result of data *mining mining*. Final conclusions are drawn based on various hypotheses generated from the data *mining process*.

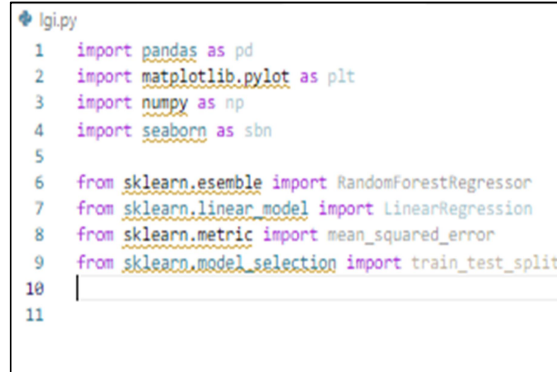
RESULTS AND DISCUSSION

In the research on house price prediction that is being researched, the *Exploratory Data Analysis* (EDA) method is used, which is a process of analyzing a set or several data to shorten the main characteristics in order to understand the condition of the dataset. In general, *Exploratory Data Analysis* (EDA) uses visualization methods such as Histogram, Box Plot, Violin Plot. The following is the process of this research using *Exploratory Data Analysis* (EDA).

1. Import Libraries

At this stage you are required to *import the library* so you can call and use the modules available in the Python language. *The library* used can be seen in Figure 2.

Figure 2. Import Library



```

1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
4 import seaborn as sns
5
6 from sklearn.ensemble import RandomForestRegressor
7 from sklearn.linear_model import LinearRegression
8 from sklearn.metrics import mean_squared_error
9 from sklearn.model_selection import train_test_split
10
11

```

The imported library function in Figure 4.1 is as follows:

- Pandas: An open-source library in the Python programming language used to process data, perform data manipulation, and perform data analysis.
- Numpy: A library in the Python programming language that functions to perform numerical computations.
- Matplotlib: A library used to visualize data in 2D and 3D, as well as to create high-quality images that can be saved in various image formats such as JPEG and PNG.
- Seaborn: A library used to create graphs and perform statistical analysis. Seaborn is built on Matplotlib and integrated with data structures from Pandas.
- Sklearn: Library in the Python programming language used to build machine learning models, such as regression, classification, and others.

2. Exploratory Data Analysis

Exploratory Data Analysis is a process for exploring and examining existing data with the aim of understanding its characteristics. This stage involves checking for empty data, deleting duplicate data, and checking the distribution and basic statistics of the existing variables.

Table 4. Data_type information

#	Column	Non-Null Count	D_type
0	HOME_NAME	1010. non-null	Object
1	HOUSE_PRICE	1010. non-null	Int. 64
2	LUAS_BUILD	1010. non-null	Int. 64
3	SURFACE AREA	1010. non-null	Int. 64
4	KT	1010. non-null	Int. 64
5	KM	1010. non-null	Int. 64
6	GRS	1010. non-null	Int. 64

In Table 3, a data description is presented which includes information such as the amount of data (count), average (mean), standard deviation (std), minimum value (min), quartile value, and maximum value (max) for each observed variable. , namely PRICE, L_B, L_T, K_T, K_M, and G_R_S. This helps in initial understanding of the data to be processed further

Table 5. Check missing_value

HOUSE NAME	.0
PRICE	.0
LB	.0
LT	.0
KT	.0
KM	.0
G. RS	.0
D_type : i	nt64

In table 5 it can be seen that when *checking missing value*, it appears that there is no empty data so that in this condition this stage can be continued to the next stage.

3. Representation

In this stage, visualization of the house data can be carried out, so that you can then see the distribution of the data as in Figures 3, 4 and 5.

Figure 3. Data_distribution (displot)

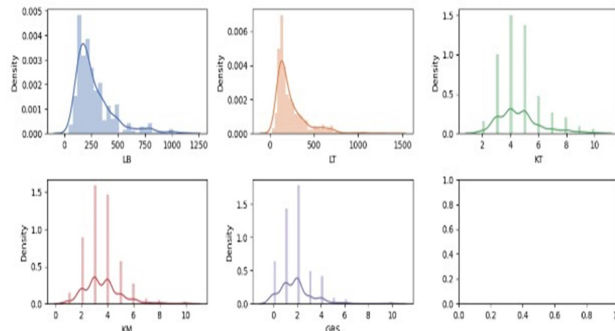


Figure 4. Data_distribution (boxplot)

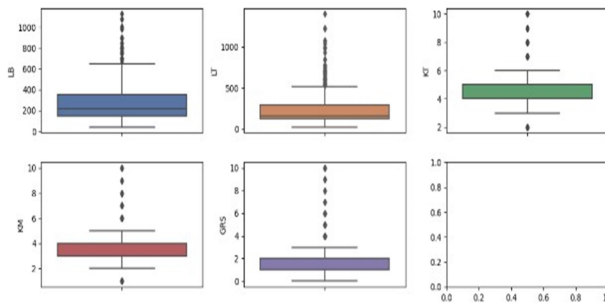
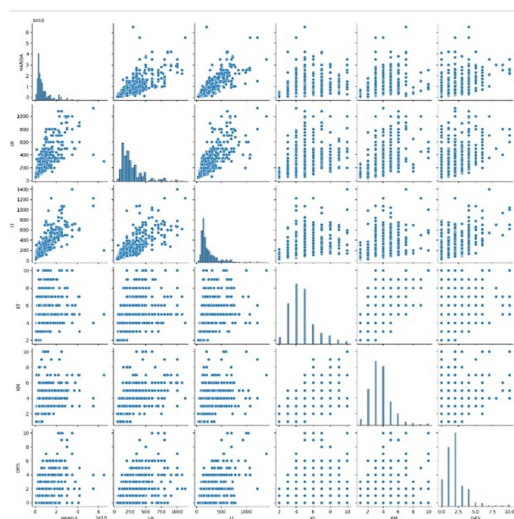
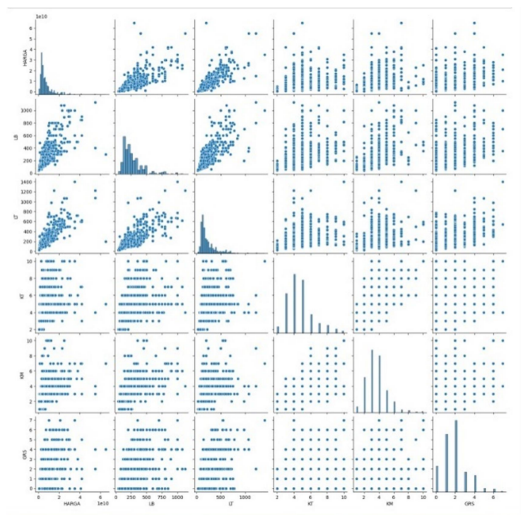


Figure 5. Distribution of house_data (pair_plot)



In Figures 3, 4, and 5, you can see the distribution of data for house attributes such as building area, land area, number of bedrooms, number of bathrooms, and number of garages. It can be observed that there are some outliers in the data which need to be cleaned to improve data cleanliness. Outliers that will be removed are only 1% of the total house data, so that the initial data percentage of 100% will be reduced to 99%. Data that has been cleaned from outliers can be seen in Figure 6 below.

Figure 6. Distribution of data after cleaning



4. Correlation Matrix

A correlation matrix is a table that visualizes the linear relationship between two or more variables. In the correlation matrix, the Pearson correlation coefficient is used as a measure, which has a value range between -1 to +1. When the value is positive, it indicates that there is a positive relationship between the variables, which means that if the value of one variable increases, it is likely that the value of the other variable also increases. Conversely, when the value is negative, it indicates that there is a negative relationship between the variables, which means that if the value of one variable increases, it is likely that

the value of the other variable will decrease. Meanwhile, a value of 0 indicates that there is no linear relationship detected between these variables.

Figure_7. Correlation matrix

	HARGA	LB	LT	KT	KM	GRS
HARGA	1.00	0.75	0.81	0.32	0.40	0.48
LB	0.75	1.00	0.74	0.44	0.53	0.49
LT	0.81	0.74	1.00	0.43	0.39	0.48
KT	0.32	0.44	0.43	1.00	0.67	0.28
KM	0.40	0.53	0.39	0.67	1.00	0.35
GRS	0.48	0.49	0.48	0.28	0.35	1.00

From the visualization in Figure 7, it can be seen that the purple area in the correlation coefficient indicates the existence of a significant and positive linear relationship between the variables. In contrast, white areas indicate the absence of a significant linear relationship, indicating that no correlation can be identified between the two variables.

5. Multiple Linear Regression

Multiple linear regression calculations were carried out using the Sklearn library. The dataset consists of 1010 entries which are partitioned into 80% training data (808 entries) and 20% test data (202 entries). The next step is to determine the column that acts as the dependent variable (y), namely the price column, and the independent variable column (x), namely LB, LT, KT, KM, and GRS. The multiple linear regression process was carried out as described in Table 6.

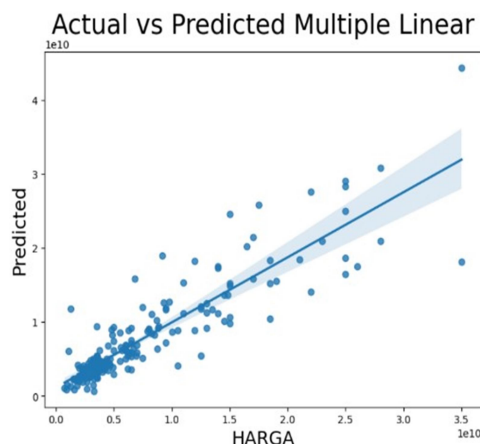
Table 6. Python commands in multiple linear regression

1	<code>x=data.drop(columns=['PRICE','HOUSE NAME']) y = data['PRICE']</code>	Set the columns that will be the dependent variable (y) and the independent variable (x)
2	<code>x_train, x_test, y_train, y_test = train_test_split(x, y,test_size=0.2, random_state=2)</code>	Divide the data into <i>training data</i> and <i>test data</i> . <i>Training data</i> is 80% and <i>test data</i> is 20%
3	<code>lin_reg = LinearRegression()</code>	Make models <i>linear regression</i>
4	<code>lin_reg.fit(x_train, y_train)</code>	
5	<code>print("Coefficient of determination :",r2_score(y_test,y_pred)) print("MSE:",mean_squared_error(y_test,y_p red)) print("RMSE:",np.sqrt(mean_squared_error(y_ test,y_pred)))</code>	Coefficient of determination: 0.7851252699343726 MSE: 9.633962972304015e +18 RMSE : 3103862589.146629

6	<code>y_pred=lin_reg.predict(x_test)</code>	Calculate predicted values
7	<code>plt.figure(figsize=(8,6)) plt.title("Actual vs. predicted",fontsize=25)</code> <code>plt.xlabel("Actual", fontsize=18)</code> <code>plt.ylabel("Predicted", fontsize=18)</code> <code>plt.scatter(x=test_y,y=test_predi ct)</code> <code>sns.regplot(x=y_test, y=ypredict) plt.show()</code>	Visualize initial y values and predicted y values

The following is a visualization of the linear regression process.

Figure 8. Multiple linear data visualization



6. Random Forest (RF) Regression

Random forest calculations were carried out using the Sklearn library. The dataset consists of 1010 entries which are partitioned into 80% training data (808 entries) and 20% test data (202 entries). The next step involves determining the column that acts as the dependent variable (y), namely the price column, as well as the independent variable column (x), namely LB, LT, KT, KM, and GRS. The random forest process was carried out as described in Table 7.

Table 7. Python commands in random forest

1	<code>x = data.drop(columns=['PRICE','HOUSE NAME'])</code> <code>y = data['PRICE']</code>	Set the columns that will be the dependent variable (y) and the independent variable (x)
2	<code>x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=2)</code>	Divide the data into <i>training data</i> and <i>test data</i> . <i>Training data</i> is 80% and <i>test data</i> is 20%
3	<code>random_forest = RandomForestRegressor(n_estimator=100 , random_state=0)</code>	Create a <i>random model forest</i> with 100 <i>trees</i> or trees.
4	<code>random_forest.fit(x_train, y_train)</code>	

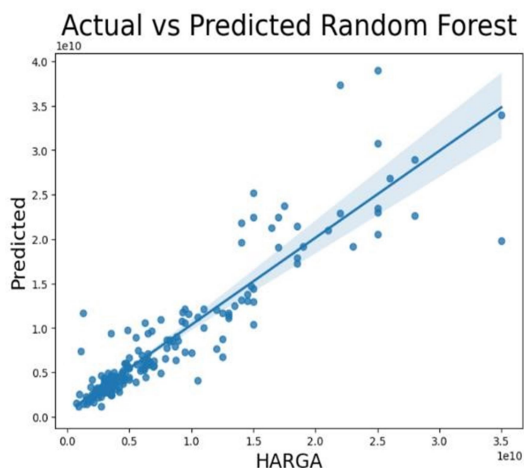
```

5 print("Coefficient of determination      Coefficient of determination
   :",r2_score(y_test,y_pred))           : 0.8162183421218491 MSE:
   print("MSE:",mean_squared_error(y_te  8.239897201713704e+18
   st,y_pred))                           RMSE: 2870522113.08565
   print("RMSE: ",np.sqrt(mean_squared_er
   ror(y_test,y_pred)))
6 y_pred=random_forest.predict(x_test)   Calculate predicted values
7 plt.figure(figsize=(8,6))              Visualize initial y values and predicted y
   plt.title("Actual vs. predicted",fontsi values
   ze=25)
   plt.xlabel("Actual",fontsize=18)
   plt.ylabel("Predicted",fontsize=18)
   #plt.scatter(x=test_y,y=test_predict)
   sns.regplot(x=y_test, y=ypredict)
   plt.show()

```

The following is a visualization of the *random forest process*.

Figure 9. Visualization of random forest data



7. Deployment/Evaluation

The accuracy value for the multiple linear regression algorithm is 78.5%, while for the random forest regression algorithm it is 81.6%.

Multiple Linear Regression

The development of a simple regression model is known as multiple linear regression (Puteri, K., & Silvanie, A. (2020)). There are only two variables in a simple regression model: one independent variable and one dependent variable. However, there are more independent variables than dependent variables in multiple linear regression analysis. After adjusting the independent variables, the general form of a regression line that includes two or more independent variables is as follows. Y is equal to α plus $\beta_1 X_1$ plus $\beta_2 X_2$ plus $\beta_n X_n$ plus e .

Information:

Y = variable- *response*

X = variable- *predictor*

α =Constant.

β = Coefficient- *estimate*

Random Forest Regression

Random forest is a random model consisting of several models used in the implementation of bootstrap regression and classification methods - combining and selecting features (Purwa, T. (2019)). A decision tree, also known as a decision tree, is a flow map designed to be used as an information gathering tool. Decision trees categorize unknown data samples into the correct categories. The goal of a decision tree is to reduce overfitting and achieve maximum accuracy. Random forest is a collection of trees that are then combined into one model (Hegelich, 2016). Random forests are based on a number of random vectors that have the same distribution across trees, with each decision tree having a maximum value. Random forest classification is based on the formula $\{h(x, \theta_k), k = 1, \dots\}$, where θ_k is a random vector (Dong et al., 2014).

Jupyter Notebooks

Jupyter Notebook is an application named after its three main programming languages, namely Ju(Java), Py(Python), and R. *Tools* This is a popular tool among data scientists to perform data processing in an interactive way, allowing users to integrate code and results directly into one document (Perkel, 2018). In Jupyter Notebook, users can write and run code, display mathematical equations, create data visualizations, and add rich narrative text. The ease of writing and sharing *text* and *code* makes it a very useful tool for collaboration between users.

CONCLUSION

From testing the two algorithm models, it was found that the multiple linear regression model had an accuracy level of 78.5%, while the random forest regression model had an accuracy level of 81.6%. Therefore, it can be concluded that the random forest regression model is more accurate for this research. For future researchers who are interested in researching the same topic, it is recommended to carry out research using a different algorithm.

REFERENCES

- Amalia, A., Radhi, M., Sinurat, SH, Sitompul, DRH, & Indra, E. (2021). Car price prediction using a regression algorithm with hyper-parameter tuning. *Prima Journal of Information Systems and Computer Science (JUSIKOM PRIMA)*, 4(2), 28-32.
- Athiyah, U., Hananta, A., Maulidi, T., Putra, V. M. E., Purba, T. F. H., & Bakowatun, E. A. W. (2021). Sistem Pendukung Keputusan Prediksi Harga Rumah Kost untuk Mahasiswa IT Telkom Purwokerto Menggunakan Metode Fuzzy Tsukamoto Berbasis Web. *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, 1(2), 77-81.
- Azhar, and I. (2021). Article, "House Price Prediction Using General Regression Neural Network," *J. Inform.*, 8(1).
- Dong, L., Li, X., & Xie, G. (2014). Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive Bayes classification. In *Abstract and applied analysis* (Vol. 2014, pp. 1-8). Hindawi Limited.

- Fahlepi, MR, & Widjaja, A. (2019). Application of the Multiple Linear Regression Method to Predict Boarding Room Rental Prices. *STRATEGY Journal-Maranatha Journal*, 1(2), 615-629.
- Hanifi, S., Liu, X., Lin, Z., & Lotfian, S. (2020). A critical review of wind power forecasting methods—past, present and future. *Energies*, 13(15), 3764.
- Hanifi, S., Liu, X., Lin, Z., & Lotfian, S. (2020). A critical review of wind power forecasting methods—past, present and future. *Energies*, 13(15), 3764.
- Hegelich, S. (2016). Decision trees and random forests: Machine learning techniques to classify rare events. *European Policy Analysis*, 2(1), 98–120.
- Mu'tashim, ML, Muhayat, T., Damayanti, SA, Zaki, HN, & Wirawan, R. (2021). Analysis of house price predictions according to specifications using multiple linear regression. *Informatics: Journal of Computer Science*, 17(3), 238-245.
- Purwa, T. (2019). Comparison of Logistic Regression and Random Forest Methods for Classifying Imbalanced Data (Case Study: Classification of Poor Households in Karangasem Regency, Bali, 2017). *Journal of Mathematics, Statistics and Computing*, 16(1), 58-73.
- Puteri, K., & Silvanie, A. (2020). Machine learning for basic food price prediction models using multiple linear regression methods. *National Journal of Informatics (JUNIF)*, 1(2), 82-94.
- Ridho, II, Mahalisa, G., Sari, DR, & Fikri, I. (2022). Neural Network Method for Determining the Accuracy of House Price Predictions. *Technologia: Scientific Journal*, 13(1), 56-58.
- Saiful, A. (2021). House price prediction using web scrapping and machine learning with a linear regression algorithm. *JATISI (Journal of Informatics Engineering and Information Systems)*, 8(1), 41-50.
- Sanusi, RM, Ansori, ASR, & Wijaya, R. (2020). Prediction of House Prices in the Eastern City of Bandung Using the Regression Method. *eProceedings of Engineering*, 7(3).
- Saputri, E. A., & Ekojono, E. (2018). Prediksi Volume Impor Beras Nasional Menggunakan Jaringan Saraf Tiruan Metode ELM (Extreme Learning Machine). *SENTIA 2018*, 10(1).
- Using the Web-Based Fuzzy Tsukamoto Method. *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, 1(2), 77-81.
- Wahyuni, I., Nafi'iyah, N., & Masruroh, M. (2019, September). Sistem Peramalan Penjualan Perumahan di Kabupaten Lamongan Dengan Menggunakan Metode Regresi Linier Berganda. In *Seminar Nasional Sistem Informasi (SENASIF)* (Vol. 3, pp. 1969-1973).