VOL. 9. NO. 1 AUGUST 2023
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v9i1.4203

# RAINFALL PREDICTION USING MULTIPLE LINEAR REGRESSION ALGORITHM

**Mulia Sulistiyono[1]; Acihmah Sidauruk[2]; Budy Satria[3*]; Raditya Wardhana[4]**

Department of Computer Science[1,2,4]
Universitas Amikom Yogyakarta
https://home.amikom.ac.id
muliasulistiyono@amikom.ac.id[1], acihmah@amikom.ac.id[2], raditic@amikom.ac.id[4]

Department of Computer Engineering[3*]
Institut Teknologi Mitra Gama
https://www.amikmitragama.ac.id/
budysatriadeveloper@gmail.com[3*]

(*) Corresponding Author

***Abstract***— *Indonesia is a tropical region with ever-changing weather changes. It is necessary to conduct a research on weather prediction as a decision making regarding weather information that will occur in the future. Rainfall is one of the factors that cause changes in weather in an area. This research was conducted on the climate in the Yogyakarta region in the form of mountains and lowlands causing differences in rainfall. The variables that are used to make predictions are several parameters that affect rainfall, namely temperature, humidity, wind speed and duration of solar radiation. These 5 variables are processed through the data obtained then carried out research and comparisons with the previous data. Multiple linear regression is the algorithm used. This algorithm is one of the machine learning techniques by making rainfall data as the dependent variable and other parameters as independent variables. This study uses Yogyakarta City, Central Java climate data for 2010-2020. The results obtained are an R2 score of 12.99%. Prediction of rainfall is obtained at 14.41778516. Then the RMSE evaluation resulted in a deviation between predicted rainfall and actual rainfall of 14.78316110508722. Based on these results, it shows that there is light rain because it is in the intensity category of 5 mm – 20 mm/day.*

***Keywords***: *data mining, prediction, multiple linear regression, rain, weather*

**Intisari**— *Negara Indonesia merupakan wilayah tropis dengan perubahan cuaca yang selalu berubah. Perlu dilakukan sebuah penelitian tentang prediksi cuaca sebagai pengambilan keputusan terhadap informasi cuaca yang akan terjadi dikemudian hari. Curah hujan merupakan salah satu faktor yang menyebabkan perubahan cuaca di suatu wilayah. Penelitian ini dilakukan terhadap iklim di wilayah Yogyakarta yang berupa pegunungan dan dataran rendah menyebabkan terjadinya perbedaan curah hujan. Variabel yang dijadikan untuk melakukan prediksi adalah beberapa parameter yang berpengaruh terhadap curah hujan yaitu suhu, kelembaban, kecepatan angina dan lama penyinaran matahari. 5 variabel tersebut diolah melalui data yang diperoleh kemudian dilakukan penelitian dan perbandingan terhadap data yang sebelumnya. Penelitian dilakukan menggunakan algoritma regresi linear berganda dengan menjadikan data curah hujan sebagai variabel dependen serta parameter lain sebagai variabel independen. Penelitian ini menggunakan data iklim yogyakarta tahun 2010-2020 Hasil yang diperoleh yaitu R2 score sebesar 12,99%. Prediksi curah hujan pada diperoleh sebesar 14.41778516. Kemudian evaluasi RMSE menghasilkan penyimpangan antara prediksi curah hujan dengan curah hujan aktual sebesar 14.78316110508722. Berdasarkan hasil tersebut menunjukkan bahwa terjadi hujan ringan karena terdapat pada kategori intensitas 5 mm – 20 mm/hari.*

***Kata Kunci***: *data mining, prediksi, regresi linier berganda, hujan, cuaca*

## INTRODUCTION

The Special Region of Yogyakarta has four physiographic units. The first is the physiography of Mount Merapi with an altitude of 80-2911 m in Sleman Regency and parts of Bantul Regency. Second, the physiography of the Southern Mountains (altitude 150-700 m) is located in the Gunungkidul Regency area. Third, the physiography of the Kulonprogo Mountains, located in the northern part of Kulonprogo and being a landscape with hilly topography. Fourth, the physiography of the lowlands (altitude 0-80 m) stretches in the southern part of the Special Region of Yogyakarta from Kulonprogo to the Bantul region which is bordered by the 1000 mountains[1].

The location of Indonesia is in a tropical area, causing climate change in Indonesia to be erratic. With Yogyakarta's landscape consisting of mountains and lowlands, excessive rainfall can result in natural disasters. The problems of flooding and landslides that occur are often associated with high rainfall and have become a recurring problem.

Some aspects that can affect rainfall include air pressure, air temperature, number of cloud layers, humidity, wind speed and duration of sunlight [2]. The greater the measured value, the heavier the rain that has taken place in the measurement area. Rainfall is the height of rainwater that collects on a flat place, does not absorb, does not seep, does not flow, and does not evaporate, then is measured on a rain gauge[3].

Rainfall is the height of rainwater that collects in a flat place, does not evaporate, does not seep, and does not flow [4]. One way that must be done to anticipate extreme weather conditions or irregular weather requires a study, one of which is weather prediction[5]. One of the methods used in prediction is multiple linear regression. This method is used to determine the influence of the independent variables on the dependent variable[6].

In this research, prediction of rainfall using multiple linear regression algorithms was carried out. Multiple Linear Regression Algorithm. Multiple Linear Regression Algorithm is a data analysis technique that is often used to examine the relationship between several variables and predict one variable [7].

Prediction is an activity that predicts what will happen in the future [8], so that the results of predictions are closely related to making a decision [9]. Rainfall prediction is done using predictors of temperature, humidity, sunshine duration and wind speed [10].

Rainfall in the tropics greatly influences the intensity of rainfall so that it has the potential for disasters that occur for human life [11]. then this research was conducted to predict rainfall in the Yogyakarta region.

Several studies have been carried out, namely predicting rainfall in the city of Sorong using the multiple regression method to obtain a correlation coefficient value of 0.8175, an MAE (Mean Absolute Error) value of 78.8695 and an RMSE (Root Mean Squared Error) value of 95. 1982 [12].

Research was also conducted to estimate the potential for rainfall in the city of Medan using multiple linear regression methods to solve the problem of potential future rainfall [13].

With the use of various parameters in this study it is expected to find more accurate results. Multiple linear regression is an algorithm that is used to explore the relationship pattern between the dependent variable and two or more independent variables [14]. The evaluation method used in this study is the Root Mean Square Error (RMSE). RMSE is an evaluation method used to determine the difference between the predicted values between models and the observed truth. Through RMSE it can be seen how much deviation there is between the predicted value of rainfall and the actual rainfall value [15].

## MATERIALS AND METHODS

The research flow used to predict rainfall using the multiple linear regression algorithm can be seen in Figure 1 below:

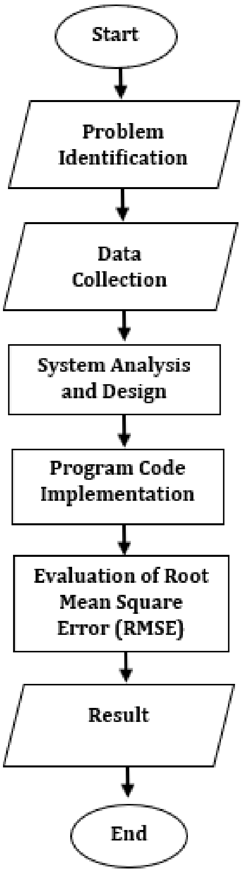

Start
Problem Identification
Data Collection
System Analysis and Design
Program Code Implementation
Evaluation of Root Mean Square Error (RMSE)
Result
End

Figure 1. Research flow

In Figure 1 it can be explained that the several stages of the research flow carried out are:

1. Problem Identification
   This stage is the beginning of conducting research. Find a problem to solve.
2. Data Collection
   The data used are data from 2010 to 2020 data obtained from the Sleman geophysical station.

After being re-elected, 4945 climate data were obtained in the city of Yogyakarta.

3. System analysis and design
   The stages of analysis and design are based on the analysis and problems that have been made before. This stage is the most optimal writing process, data, process flow and relationships between data.
4. Program Code.
   Is the process of doing the coding, and the system is ready to operate on actual data. At this stage the multiple linear regression method is translated into code using the Google Collab editor and the programming language used is Python.
5. Evaluation of Root Mean Square Error
   After processing the data and obtaining the equation of the results, it is necessary to do an evaluation. This evaluation is used to test the feasibility and relevance of the relationship between the independent and dependent variables.
6. Result
   Drawing a final conclusion on all data that has been processed into information.

## RESULTS AND DISCUSSION

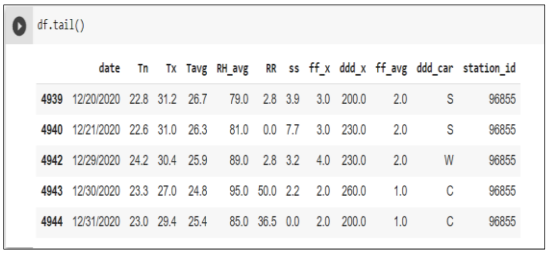
### A. Data Preparation

Data preparation where the process of taking raw data is carried out and preparing it to be absorbed into the analytical platform. One of the main functions of data preparation is to ensure the accuracy and consistency of the original data prepared for processing and analysis.

### B. Data Preprocessing

This stage is carried out to clean the data, where the raw data obtained needs to be re-selected, then deleted, incomplete or relevant, and inaccurate data are removed.

1. Calculating the number of missing values.
   To check whether there is a missing value, use the isna() function. This study obtained 284 missing values in temperature data, 287 in humidity data, 679 data for rainfall, 125 solar irradiation duration data and 175 wind speed data.
2. Delete NaN values
   Delete the "NaN" data contained in the raw dataset using the dropna function.
3. View Details of All Columns and Data
   This stage is carried out to see the columns and data in the dataset using the df.tail () function which can be seen in Figure 2 below:

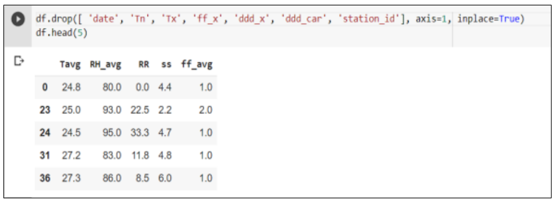


```
df.tail()
```

```
            date    Tn    Tx  Tavg  RH_avg    RR   ss  ff_x  ddd_x  ff_avg  ddd_car  station_id
4939  12/20/2020  22.8  31.2  26.7    79.0   2.8  3.9   3.0  200.0     2.0        S       96855
4940  12/21/2020  22.6  31.0  26.3    81.0   0.0  7.7   3.0  230.0     2.0        S       96855
4942  12/29/2020  24.2  30.4  25.9    89.0   2.8  3.2   4.0  230.0     2.0        W       96855
4943  12/30/2020  23.3  27.0  24.8    95.0  50.0  2.2   2.0  260.0     1.0        C       96855
4944  12/31/2020  23.0  29.4  25.4    85.0  36.5  0.0   2.0  200.0     1.0        C       96855
```

Figure 2. Column and data detail codes

4. Recalculating the missing value
   In this section there are no more missing values contained in the dataset because they were deleted in the previous stage.
5. Remove unused variable columns
   The df.drop() function can be used to remove some unnecessary column variables such as date, Tn, Tx, ff_x, ddd_x, ddd_car, and station_id. After the unused variable columns are deleted, only 5 variable columns remain which will be used to predict daily rainfall. The program code for deleting unused data columns can be seen in Figure 3 below:



```
df.drop([ 'date', 'Tn', 'Tx', 'ff_x', 'ddd_x', 'ddd_car', 'station_id'], axis=1, inplace=True)
df.head(5)
```

```
    Tavg  RH_avg    RR   ss  ff_avg
0   24.8    80.0   0.0  4.4     1.0
23  25.0    93.0  22.5  2.2     2.0
24  24.5    95.0  33.3  4.7     1.0
31  27.2    83.0  11.8  4.8     1.0
36  27.3    86.0   8.5  6.0     1.0
```

Figure 3. Delete Column

### C. Data Representation

This stage is a set of data used to build a multiple linear model

1. Average temperature (Tavg).
   Tavg data representation it can be concluded that the highest average temperature distribution is in the numbers 26°. Visualization of the average temperature can be seen in Figure 4 below:

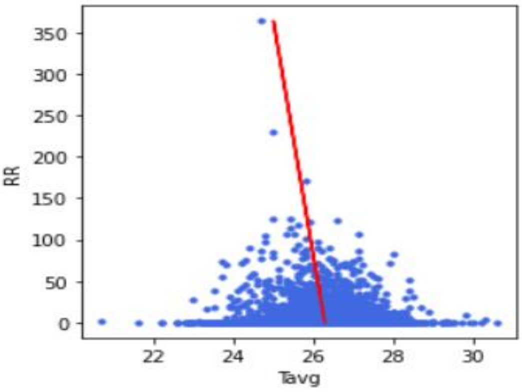


Figure 4. Average Temperature

Based on Figure 4, The graph above shows that the average temperature data with rainfall data have a weak correlation and are negative.

2. Average humidity (RH_avg).
   It can be concluded that the highest average moisture distribution data is at 80-85%. The following figure 4 shows the visualization of average humidity data. The graph shows that the average humidity with rainfall has a weak correlation and is positive. Visualization of Average Humidity Variable Data can be seen in Figure 5 below:

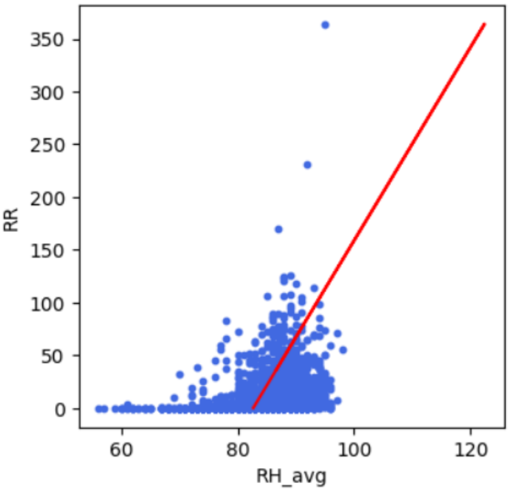


Figure 5. Average Humidity Variable Data Visualization

3. Data analysis for the longest irradiation was found at 4-7 hours. The correlation relationship that is formed is weak and has a negative value.
4. In the favg representation, the average wind speed distribution data is obtained at 1 - 3 m/s. The correlation is weak and negative
5. Analysis of the relationship and the level of correlation between the independent and dependent variables is shown in Figure 6.

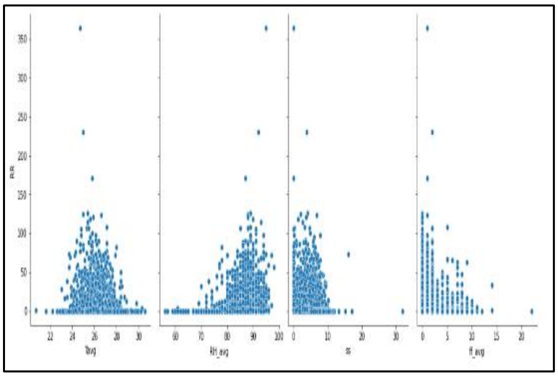
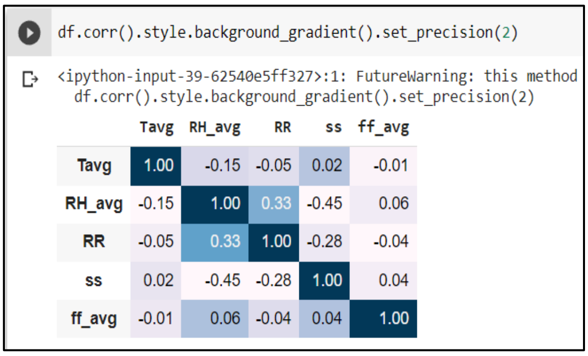


Figure 6. Distribution of all variables

The correlation between variables can be seen in Figure 7.



```
df.corr().style.background_gradient().set_precision(2)
```

```
<ipython-input-39-62540e5ff327>:1: FutureWarning: this method
  df.corr().style.background_gradient().set_precision(2)
```

| | Tavg | RH_avg | RR | ss | ff_avg |
|---|---|---|---|---|---|
| Tavg | 1.00 | -0.15 | -0.05 | 0.02 | -0.01 |
| RH_avg | -0.15 | 1.00 | 0.33 | -0.45 | 0.06 |
| RR | -0.05 | 0.33 | 1.00 | -0.28 | -0.04 |
| ss | 0.02 | -0.45 | -0.28 | 1.00 | 0.04 |
| ff_avg | -0.01 | 0.06 | -0.04 | 0.04 | 1.00 |

Figure 7. Correlation Between Variables

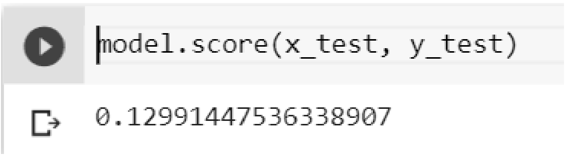
The results of calculating the correlation test using an average temperature predictor that is related to rainfall produce a correlation value of -0.05. In calculating the correlation test with the average humidity predictor which has a relationship with rainfall produces a correlation value of 0.33. Furthermore, in the calculation of the correlation test with the predictor of sunshine duration which is related to rainfall, it produces a correlation value of -0.28. In calculating the correlation test with the predictor of average wind speed which has a relationship with rainfall, it produces a correlation value of -0.04.

### D. Dataset Modeling

Original data set that will be further processed to obtain research results.

1. Data Splitting
   Process of creating variables X and Y variables and dividing the data in a 90:10 portion on training and testing data fortunately, testing and training is carried out.
2. Linear Regression Objects
   Create a linear regression object and train the model using split training data.
3. Looking for coefficient and intercept
   Carried out to find the value of the coefficient and intercept. A value of -0.26769828 was obtained for the average temperature variable (X1), for 0.79744447 for the average humidity variable (X2), for -0.9766293 for the long sunshine variable (X3), and for -0.40954049 for the average wind speed variable (X4).
4. Assess the size of the independent variable on the dependent variable
   This stage is carried out to find out how much the influence value of certain independent variables has on the dependent variable. The results obtained are the R2 score of 12.99%.
5. Looking for accuracy value
   Find the value of the accuracy of the prediction of rainfall. The resulting accuracy value is 0.12991447536338907. Accuracy value can be seen in Figure 8.

Accredited Rank 2 (Sinta 2) based on the Decree of the Directorate General of Higher Education, Research and Technology. Ministry of Education, Culture, Research and Technology of the Republic of Indonesia No.225/E/KPT/2022, 07 December 2022. Published by LPPM Universitas Nusa Mandiri



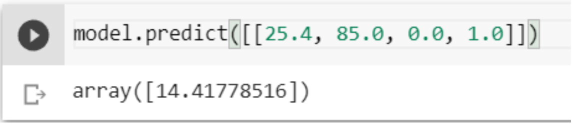
```
model.score(x_test, y_test)

0.12991447536338907
```

Figure 8. Accuracy value

### E. Rainfall Prediction

The results of the prediction of rainfall through the program code can be seen in Figure 9



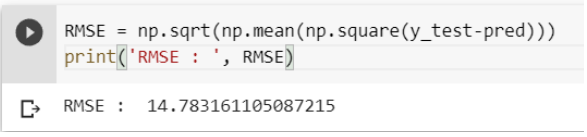
```
model.predict([[25.4, 85.0, 0.0, 1.0]])

array([14.41778516])
```

Figure 9. Rainfall Prediction Code Program

### F. Evaluation of RMSE

Evaluation is the final stage used to find out the difference between the predicted values between models and the observed truth values. Based on the evaluation that has been carried out, it produces an RMSE value of 14.783161105087215. Code Program Evaluasi RMSE can be seen in Figure 10



```
RMSE = np.sqrt(np.mean(np.square(y_test-pred)))
print('RMSE : ', RMSE)

RMSE :  14.783161105087215
```

Figure 10. RMSE Code Program

The calculation method is based on Table 1 data.

Table 1. Prediction Calculations

| Y Predicted |
|---|
| Y = bo+b1X1+b2X2+b3X3+b4X4 |
| b0 = -47,1403871 |
| b1 = -0,19140552 |
| b2 = 0,78621676 |
| b3 = -0,98725121 |
| b4 = -0,39532824 |

Y = -47.1403871 + (-0.19140552*25.4)
+ (0.78621676*85) + (-0.98725121*0)
+ (-0.39532824*1)
Y = -47.1403871 + -4,861700208 + 66,8284246
+ 0 + -0.39532824
Y = -52.002087308 + 66,8284246 + 0
+ -0.39532824
Y = 14.826337292 + 0 + -0.39532824
Y = 14,431009

Based on the calculation, it is obtained the predicted value of daily rainfall using the predictors of average temperature, average humidity, length of sunlight, and average wind speed of 14.41778516. Whereas the results of manual calculations that have been carried out using Microsoft Excel show the predicted results of daily rainfall of 14.431009. The results of these values indicate that on Thursday, December 31 2020, light rain occurred because it was in the intensity category of 5 mm – 20 mm/day. Therefore, the community is expected to anticipate the light rain weather. The RMSE evaluation calculation uses the formula below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)2} \text{.........................................} (1)$$

RMSE= √ (892067,864 / 3542)
RMSE= √251,85428119706380575945793337098
RMSE= 15,869917491816515608618901660547

The error rate that occurs due to deviations between the daily rainfall prediction data and the actual rainfall data predictor is 14.783161105087215, which means it is very good for the rainfall prediction process because of the small deviation (error) between the predicted rainfall value and the actual rainfall value. Meanwhile, manual calculations using Microsoft Excel show that the value of the deviation between the predicted daily rainfall data and the predictor of actual rainfall data is greater, namely 15.8699175. The smaller the RMSE value, the smaller the difference between the predicted value and the actual value and it can be said to have a good closeness between the two, meaning that the regression model fits the data better.

## CONCLUSION

Prediction of daily rainfall on Thursday, December 31, 2020 gets a prediction result of 14.41778516. Whereas manual calculations using Microsoft Excel produce a predicted value of 14.431009. Based on these results, it shows that there is light rain because it is in the intensity category of 5 mm – 20 mm/day. The error rate of the results of this rainfall prediction that occurs is 14.783161105087215 which means good because of the small number of deviations (errors) between the predicted rainfall value and the actual rainfall value. Whereas manual calculations using Microsoft Excel show a greater deviation value of 15.8699175. As an improvement to increase the prediction results of rainfall there are suggestions in the form of adding other variables that have a close relationship with the occurrence of rainfall.

## REFERENCES

[1] A. Apriani, “Statistik Non Parametrik Untuk Membandingkan Pembagian Fungsi Kawasan Dengan Penggunaan Lahan,” *Semin. Nas. Mat. Geom. Stat. dan Komputasi*, pp. 602–611, 2022.

[2] A. M. Sajiah, *et al*., "Aplikasi Perkiraan Curah Hujan Kota Kendari Menggunakan Metode Interval Type - 2 Fuzzy Logic System," *J. Fokus Elektroda* , vol. 08, no. 02, pp. 86–91, 2023.

[3] A. A. Al Badri, *et al*., "Optimasi Model Arima Dalam Prakiraan Curah Hujan di Jambi," *J. Perndidikan dan Peneliyian Geogr.*, vol. 4, no. 1, pp. 39–43, 2023, doi: 10.53682/gjppg.v4i1.7041.

[4] E. Q. Ajr and F. Dwirani, "Menentukan Stasiun Hujan Dan Curah Hujan Dengan Metode Polygon Thiessen Daerah Kabupaten Lebak," *Jurnalis*, vol. 2, no. 2, pp. 139–146, 2019.

[5] S. Agustian and S. Ramadhani, "Rancang bangun sistem monitoring curah hujan berbasis internet of things," *J. Comput. Sci. Inf. Technol.*, vol. 4, no. 1, pp. 42–49, 2023.

[6] A. Sari, Susanto, and H. O. L. W. Wijaya, "Rainfall prediction system with predictors of temperature humidity, wind speed and air temperature using r programming language," *Int. J. Cist.*, vol. 2, no. 01, pp. 12–16, May 2023, doi: 10.56481/cister.v2i01.157.

[7] E. Triyanto, *et al*., "Implementasi Algoritma Regresi Linear Berganda Untuk Memprediksi Produksi Padi Di Kabupaten Bantul," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 4, no. 2, pp. 66–75, 2019, doi: 10.36341/rabit.v4i2.666.

[8] B. Satria, *et al*., "Implementasi Metode Fuzzy Sugeno Untuk Prediksi Penentuan Porsi Dana Pembangunan Perumahan," *JSAI (Journal Sci. Appl. Informatics)*, vol. 4, no. 1, pp. 75–84, 2021, doi: 10.36085/jsai.v4i1.1330.

[9] B. Satria, "Prediksi Volume Penggunaan Air PDAM Menggunakan Metode Jaringan Syaraf Tiruan Backpropagation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 3, pp. 674–684, 2018, doi: 10.29207/resti.v2i3.575.

[10] E. D. S. Mulyani, *et al*., "Prediksi Curah Hujan Di Kabupaten Majalengka Dengan Menggunakan Algoritma Regresi," *Sist. Inf. dan Teknol. Inf.*, vol. 8, no. 1, pp. 67–77, 2019.

[11] M. A. Hasanah, *et al*., "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *J. Appl. Informatics Comput.*, vol. 5, no. 2, pp. 103–108, 2021, doi: 10.30871/jaic.v5i2.3200.

[12] M. Yusuf, *et al*., "Analisis Prediksi Curah Hujan Bulanan Wilayah Kota Sorong Menggunakan Metode Multiple Regression," *J. Sains Komput. Inform.*, vol. 6, no. 1, pp. 405–417, 2022.

[13] A. Marbun, *et al*., "Analisa Data Mining Untuk Mengestimasi Potensi Curah Hujan Dengan Menggunakan Metode Regresi Linear Berganda," *J. CyberTech*, vol. 4, no. 2, pp. 1–7, 2021, [Online]. Available: https://ojs.trigunadharma.ac.id/

[14] T. N. Padilah and R. I. Adam, "Analisis Regresi Linier Berganda Dalam Estimasi Produktivitas Tanaman Padi Di Kabupaten Karawang," *FIBONACCI J. Pendidik. Mat. dan Mat.*, vol. 5, no. 2, pp. 117–128, 2019, doi: 10.24853/fbc.5.2.117-128.

[15] I. J. A. Saragih, *et al*., "Prediksi Curah Hujan Bulanan Di Deli Serdang Menggunakan Persamaan Regresi dengan Prediktor Data Suhu dan Kelembapan Udara," *J. Meteorol. Klimatologi dan Geofis.*, vol. 7, no. 2, pp. 6–14, 2020.