Published *online* on the journal's web page : http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

Prediction of Dengue Fever Cases Using the Linear Regression Method Based on Open Data from West Java Province

Muhammad Khysam Firdaus[1], *Herman Yuliansyah*[2]

[1,2]*Department of Informatics, Universitas Ahmad Dahlan, Yogyakarta, 55191, Indonesia*
[1]*mkhysamf@gmail.com*, [2]*herman.yuliansyah@tif.uad.ac.id*

ARTICLE INFORMATION

Article History:
Received: June 26, 2025
Last Revision: October 17, 2025
Published *Online*: October 30, 2025

KEYWORDS

Dengue Hemorrhagic Fever,
Evaluation,
Linear Regression,
Machine Learing,
Prediction Model

CORRESPONDENCE

Phone: +6285138592043
E-mail: *herman.yuliansyah@tif.uad.ac.id*

ABSTRACT

Dengue Hemorrhagic Fever (DHF) is a widespread disease in tropical regions, including Indonesia. West Java Province reports the highest number of cases, influenced by factors such as rainfall, population density, and total population. Accurate prediction of DHF cases is essential for effective prevention and control strategies. This study aims to propose a predictive model for DHF cases in West Java using the Linear Regression method and to evaluate its performance using Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) metrics. The research utilizes secondary data from 2019 to 2023 on DHF cases, population density, and total population from the Open Data Jabar platform. Rainfall data were collected from Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) and Badan Pusat Statistik Indonesia (BPS). The research process includes data collection, preprocessing, time series splitting, model training and iteration, prediction, and performance evaluation. The results show that among the five focus regions, Bandung City achieved the best prediction performance, with a MAPE of 45.82% and an RMSE of 1216.105. These findings indicate that Linear Regression method is reasonably effective for predicting DHF cases, particularly in Bandung. Despite limitations in data availability especially rainfall data the model provides informative insights. Future work could improve prediction accuracy by incorporating additional independent variables and more advanced modeling techniques, such as machine learning.

1. INTRODUCTION

Dengue Hemorrhagic Fever (DHF) is a vector-borne infectious disease caused by the dengue virus carried by Aedes aegypti mosquitoes [1]. DHF is one of the most widespread tropical diseases and has the fastest-growing number of cases worldwide [2]. More than 3.9 billion people in over 129 countries are at risk of contracting dengue fever, with an estimated 96 million symptomatic cases and approximately 40,000 deaths annually. Mosquito-borne diseases are endemic in many regions, including Southeast Asia. This disease is a major cause of death and illness in many developing countries [3].

Indonesia is among the countries with the highest number of DHF cases in Southeast Asia and even globally. According to the 2021 Indonesian Health Profile, there were 73,518 reported cases of DHF with 705 resulting deaths. West Java was the province with the highest number of DHF cases in Indonesia that year, reaching 23,959 cases, consisting of 12,332 male patients and 11,627 female patients. This high number indicates the need for greater attention to the factors that may contribute to the increase in DHF cases in West Java, such as population size, population density, and climatic conditions [4].

The dengue prevention and control program in Indonesia has been in place for a long time, encompassing activities such as fogging, larviciding, and mosquito larvae inspections by *Jumantik* (community health volunteers) through house-to-house visits [5], [6]. In addition, the implementation of the 3M Movement (covering and cleaning water storage containers, and managing waste properly) is also an important part of the prevention efforts [7], [8]. However, DHF remains a significant public health issue in Indonesia. The spread of this disease continues to expand due to high mobility and population density, as well as weather factors such as rainfall [9]. Therefore, it is essential to predict the number of DHF cases in order to provide accurate and timely information that can support more effective and targeted planning, prevention, and disease control efforts.

Previous studies on dengue case prediction have applied methods such as Linear Regression using climate variables (temperature, rainfall, humidity) and Fuzzy Linear Regression to address uncertainty in medical data. However, these studies were generally limited in scope—either focusing on specific climatic factors or using datasets from regions outside West Java. Additionally, few studies have incorporated demographic variables such as population density and total population, which are highly relevant to dengue transmission patterns. Therefore, there remains a gap in developing a predictive model for dengue cases in West Java Province that integrates both climatic and demographic factors using a multi-year dataset. This research addresses that gap by applying a Linear Regression approach tailored to regional characteristics to produce more accurate and reliable predictions.

The prediction of dengue fever cases in West Java Province employs the Linear Regression method as it aligns with the numerical characteristics of the dataset and the linear relationship between demographic variables and dengue incidence. This method was chosen for its simplicity, ease of interpretation, and computational efficiency, allowing the development of a clear mathematical equation to explain the influence of population size and density on case numbers. In this study, dengue cases are set as the target variable, while population and population density serve as predictors.

The data used in this study includes the number of dengue fever cases, population density, and total population from 2019 to 2023, obtained from the official website of West Java Open Data (https://opendata.jabarprov.go.id), which provides secondary data from relevant government agencies over a specific period.. Primary rainfall data was collected directly from the Meteorology, Climatology, and Geophysics Agency (BMKG), while secondary rainfall data was obtained through the Central Statistics Agency (BPS) of the relevant region.

Based on the identified research gap, this study formulates the following research objectives. (1) To develop a predictive model for dengue fever cases in West Java Province using the Linear Regression method. (2) To evaluate the performance of the proposed model by calculating the Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) in predicting dengue fever cases in West Java Province.

The key contributions of this study to the existing body of knowledge are summarized as follows. (1) To produce a predictive model that can support authorities in planning preventive and control measures for future dengue outbreaks. (2) To enhance the preparedness of the healthcare system in anticipating potential surges in dengue cases, thereby reducing the burden on health facilities during outbreaks. (3) To improve the responsiveness of local authorities to public health needs in West Java Province through more informed decision-making.

2. RELATED WORK

Various previous studies have discussed the prediction of Dengue Hemorrhagic Fever (DHF) cases using a wide range of methods. A review of these studies was conducted to gain more comprehensive insights into previously identified issues.

In a previous study, Ichwani and Wibawa [10] conducted a study on predicting DHF incidence using the Extreme Learning Machine (ELM) method based on weather factors in Tembalang District. The data used included weather data and DHF case data in Tembalang District, Semarang City. The study involved neural network architecture construction, data normalization, training and testing using the ELM model. The results showed that the best ELM model had an MSE of 0.0116 with 13 splits and 6 hidden neurons, and achieved a training time of less than one second.

Furthermore, Aditya et al. [11] predicted the number of DHF cases at BRSUD Tabanan Regency using the Linear Regression method. The study utilized historical data on DHF cases, rainfall, air temperature, and humidity from 2017 to 2022. The process included data normalization using the Min-Max method, and splitting the data into training and test sets. The results showed that testing on 19 test data points yielded a MAPE of 35%. Cholil et al. [12] conducted a study on predicting the number of DHF cases at the Ngemplak Simongan Community Health Center using the C4.5 algorithm. The data used consisted of patient records from that health center. The research steps included data selection, transformation, and calculation using the C4.5 algorithm. The conclusion was that the implementation of the model in the form of a testing application achieved an accuracy of 94.44%.

Lestari and Witanti [13] predicted DHF cases based on weather factors in Sleman Regency using the Multivariate Autoregressive Integrated Moving Average (ARIMA) method. The data used included DHF case data and weather factors. The study involved data preprocessing, model implementation, and model evaluation using Mean Absolute Error (MAE), resulting in an MAE value of 18.12. Ahmad et al. [14] forecasted DHF cases using the Seasonal-ARIMA method in Bulukumba Regency. The data used were monthly DHF cases from January 2014 to December 2020. The study involved data stationarization to obtain a Seasonal-ARIMA model, identification of stationarity using the Box-Cox transformation, model identification, and diagnostic checking. The results showed that the model had a MAPE of 30.62%, making it a feasible option for forecasting.

In addition to Seasonal-ARIMA, Yulanda et al. [15] conducted a study on predicting DHF cases at RSU Haji Medan using the Multiplicative Decomposition method. The data covered monthly DHF cases from January 2017 to December 2021. The study calculated seasonal, trend (using linear trend), cyclical, and random components. The results showed that the model achieved an MSE of 0.0012 and a MAPE of 0.25%, indicating excellent accuracy and suitability for forecasting. Ichwanul Karo Muslimin [16] studied the forecasting of DHF spread in Bandung Regency using a hybrid model combining ARIMA and Average Neural Network (ANN). The study used two datasets: the Incident Rate (IR) and weather factor data.

The steps included determining ARIMA parameters using the Maximum Likelihood Estimation (MLE) function, ADFuller test, applying ANN with linear activation functions, and evaluating predictions using RMSE. Results showed both models performed very well in predicting DHF cases.

Bain and Eka [17] predicted DHF cases in Bangkalan Regency using the Fuzzy Linear Regression method. The study used three types of data: high temperature, high platelet count, and monthly DHF cases during 2020. The steps included parameter determination, constructing upper and lower boundary models, and prediction value calculation. The forecasting was based on predefined intervals, and results were evaluated using MAPE, which yielded a value of 2.87%, indicating high prediction accuracy. Aji et al. [18] forecasted DHF in DKI Jakarta Province using the Linear Regression algorithm. The indicator variables used included case counts, temperature, humidity, and rainfall. The process involved data preprocessing, model construction and analysis, testing, and performance evaluation using MSE, MAPE, and RMSE. The results indicated that Linear Regression produced a low error percentage and was thus deemed a viable prediction model.

Lastly, Lutfianawati et al. [19] applied the ARIMA algorithm to predict DHF occurrences in Kediri Regency. The study used only one data type: DHF case data from Papar Health Center from 2016 to 2021. The analysis included data plotting, stationarity testing, building a preliminary ARIMA model, diagnostic checking, forecasting, and model evaluation. The ARIMA (1,0,0) model resulted in an MSE of 28.41, indicating that the model was sufficiently accurate for forecasting use.

TABLE 1. LITERATURE REVIEW

Name, Year	Methods	Dataset	MAPE	RMSE	MSE	Accuracy	MAE	Reference
Lestari and Witanti, 2023	Multivariat Autoregressive Integrated Moving Average (ARIMA)	Dengue fever cases and weather factors	-	-	-	-	18.12	[13]
Aditya et al., 2023	Linear Regression	Dengue fever cases, rainfall, air temperature, and air humidity	35%	-	-	-	-	[11]
Ahmad et al., 2023	Seasonal-ARIMA	Number of dengue fever patients	30.62%	-	-	-	-	[14]
Yulanda et al., 2023	Multiplicative Decomposition	Number of dengue fever patients	0.25%	-	0.0012	-	-	[15]
Lutfianawati, 2023	Autoregressive Integrated Moving Average (ARIMA)	Number of cases	-	-	28.41	-	-	[19]
Ichwanul Karo Muslimin, 2022	Hybrid Autoregressive Integrated Moving Average (ARIMA) and Average Neural Network (ANN)	Number of cases, Incident Rate (IR), weather factor data	-	0.0087	-	-	-	[16]
Bain and Eka, 2021	Fuzzy Linear Regression	Number of cases, high fever, high platelet count	2.87%	-	-	-	-	[17]
Cholil et al., 2020	Algorithm C4.5	Dengue fever patients	-	-	-	94.44%	-	[12]
Ichwani and Wibawa, 2019	Extreme Learning Machine (ELM)	Weather data and dengue fever patient data	-	-	0.0116	-	-	[10]
Aji et al., 2019	Linear Regression	Number of cases, temperature, humidity, rainfall	0.681	94.001	74.795	-	-	[18]

Table 1 summarizes various methods that have been used in previous studies to predict dengue fever cases, such as Linear Regression, ARIMA, ELM, ANN, C4.5, Multiplicative Decomposition, and Fuzzy Linear Regression. Each study shows varying levels of accuracy depending on the methods and data used. In general, all methods demonstrate good predictive performance, making them valuable references for selecting appropriate models for forecasting dengue fever cases.

While these approaches have achieved notable accuracy, they largely overlook demographic determinants such as population size and density, which are critical to understanding transmission dynamics. Moreover, most investigations have been limited to localized districts or regencies, leaving broader provincial-level analyses underexplored. Addressing these gaps, this study examines DHF prediction in West Java Province, a region with high population density and significant public health challenges, thereby extending the scope of predictive modeling to new variables and a larger geographic context.

Although more complex models have demonstrated strong performance, their limited interpretability and higher computational demands can hinder practical application in public health decision-making. Linear Regression remains highly relevant as it offers transparency, efficiency, and the ability to clearly quantify relationships between demographic factors and disease incidence. By integrating population-based variables with standardized preprocessing and time series splitting, this study aims to produce a forecasting model that is both robust and accessible, supporting evidence-based strategies for DHF prevention and control at the provincial level.

3. METHODOLOGY

The following stages describe the main workflow of the research process, including data collection, data preprocessing, time series splitting, model training, data prediction, and model evaluation.

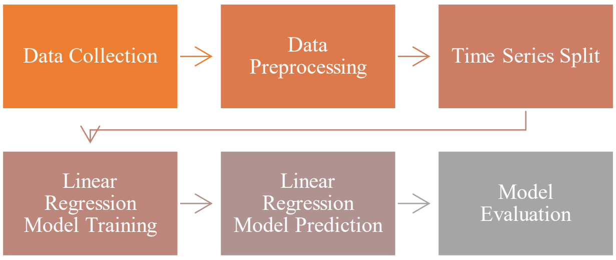

FIGURE 1. RESEARCH METHODOLOGY

3.1 Data Collection

The data used in this study includes the number of dengue fever cases, population density, and total population from 2019 to 2023, obtained from the official website of West Java Open Data (https://opendata.jabarprov.go.id), which provides secondary data from relevant government agencies over a specific period.. Primary rainfall data was collected directly from the Meteorology, Climatology, and Geophysics Agency (BMKG), while secondary rainfall data was obtained through the Central Statistics Agency (BPS) of the relevant region. In this study, the X variables (dengue fever cases, population density, and total population) serve as independent predictors representing health and demographic factors, while the Y variable (rainfall data) functions as the dependent outcome used to analyze their relationship with climatic conditions. Table 2 presents data on the number of dengue fever cases in West Java Province, showing the number of cases by gender across various regions within the province each year.

TABLE 2. DATA ON DENGUE FEVER CASES IN WEST JAVA PROVINCE

Regions	Gender	Number of Cases	Unit	Years
Kab. Bogor	Male	915	Person	2014
Kab. Bogor	Female	919	Person	2014
...	...	...	...	...
Kota Banjar	Male	24	Person	2023
Kota Banjar	Female	29	Person	2023

The data in Table 2 was then processed by first creating a pivot table based on the dataset. The result of the pivot table, derived from the number of cases in Table 2, is presented in Table 3, which shows the total number of dengue fever cases in West Java Province as the sum of cases from both genders, namely male and female.

TABLE 3. PIVOT TABLE DENGUE FEVER CASES IN WEST JAVA PROVINCE

Regions	2014	2015	…	2022	2023
Kota Bandung	3132	3640	…	5205	1856
Kab. Bandung	995	0	…	4191	1005
…	…	…	…	…	…
Kab. Pangandaran	13	16	…	502	253
Kota Banjar	129	75	…	113	53

The next step is to define the prediction time range, which spans from 2019 to 2023, and to identify the five regions with the highest number of cases as the primary focus of the predictive analysis namely Bandung City, Bandung Regency, Depok City, Bekasi City, and Bogor Regency. Table 4 presents the filtered data on the number of dengue fever cases from Table 3, showing the total cases limited to the prediction period from 2019 to 2023.

TABLE 4. RESULTS OF THE CASE DATA SELECTION

Regions	2019	2020	2021	2022	2023
Kota Bandung	4424	4424	3743	5205	1856
Kab. Bandung	2635	2303	2002	4191	1005
Kota Depok	2200	1276	3155	2234	1031
Kota Bekasi	2373	1646	2006	2442	1220
Kab. Bogor	1210	1296	2220	113	53

Besides the data on the number of cases, there is data on the population density of West Java Province. Table 5 presents the population density data of West Java Province, which shows the number of people per square kilometer in various regions within the province.

TABLE 5. POPULATION DENSITY DATA OF WEST JAVA PROVINCE

Regions	Population Density	Unit	Years
Kab. Bogor	1740	Person/KM2	2019
Kab. Sukabumi	620	Person/KM2	2019
...	...	...	...
Kota Tasikmalaya	4120	Person/KM2	2023
Kota Banjar	1590	Person/KM2	2023

Next, the data selection in Table 5 was also carried out using the Pivot Table method, as previously applied in Table 2. The result of the selected population density data is presented in Table 6, which shows the filtered population density data based on the prediction time range from 2019 to 2023.

TABLE 6. RESULTS OF POPULATION DENSITY DATA SELECTION

Regions	2019	2020	2021	2022	2023
Kota Bandung	14974	14916	15076	15277	15421
Kab. Bandung	2015	2027	2055	2130	2154
Kota Depok	9275	9351	9453	9605	9711
Kota Bekasi	11852	11929	11947	11671	11799
Kab. Bogor	1740	1893	1965	1830	1858

In addition to population density data, data on the total population in West Java Province is also available. Table 7 presents information on the total population in West Java Province, showing the distribution of the population by gender across each administrative region (regency/city) within the province.

TABLE 7. WEST JAVA PROVINCE POPULATION DATA

Regions	Gender	Number of Populations	Unit	Years
Kab. Bogor	Male	2419341	Person	2014
Kab. Bogor	Female	2296583	Person	2014
...	...	...	...	...
Kota Banjar	Male	104999	Person	2023
Kota Banjar	Female	103310	Person	2023

Subsequently, the data selection in Table 7 was also carried out using the Pivot Table method, as applied in Tables 2 and 5. The result of the selected total population data is presented in Table 8, which shows the total population in West Java Province as the sum of the male and female populations, filtered based on the prediction time range from 2020 to 2023.

TABLE 8. RESULTS OF TOTAL POPULATION DATA SELECTION

Regions	2020	2021	2022	2023
Kota Bandung	3583056	3633437	3708344	3749172
Kab. Bandung	5132355	5327131	5473476	5558885
Kota Depok	2500967	2527854	2545005	2569107
Kota Bekasi	2464719	2468448	2486251	2513669
Kab. Bogor	1872996	1893321	1920182	1941360

In addition, rainfall data was used based on the five regions with the highest number of cases identified in Table 4, which serve as the focus of the study for predicting dengue fever cases. Rainfall data for Bandung City and Bogor Regency was obtained from the local BPS websites, while data for Depok City, Bekasi City, and Bandung Regency was obtained directly from the BMKG agency through an online data request. The complete set of data used in this study is presented in Table 9, which contains the average monthly rainfall data for each year, calculated by averaging the monthly rainfall within the annual range from 2020 to 2023.

TABLE 9. RAINFALL DATA OF WEST JAVA PROVINCE

Regions	2020	2021	2022	2023
Kota Bandung	201.55	180.8917	192,6167	145,9197
Kab. Bandung	233.9	243.44	170.6	130.8
Kota Depok	259.5	278.2727	451.25	205
Kota Bekasi	127.25	149.2727	154.6667	117.6
Kab. Bogor	286.3333	292.5833	283.775	203.6667

3.2 Data Preprocessing

a. Data Filtering

Data selection is a crucial step in data analysis aimed at choosing the most relevant and significant attributes from all available features to be effectively used in the model development process [20]. In this study, data selection involved identifying relevant data needed to predict the number of dengue fever cases, determining the prediction time range (from 2020 to 2023), and focusing on the number of cases, population, and population density as the key variables for predictive analysis. Specifically number of dengue cases: Serves as the dependent variable and the primary target for prediction. Larger populations increase the absolute number of susceptible individuals, which can amplify outbreak magnitude. High-density areas facilitate mosquito–human contact, thereby increasing transmission risk. These variables were chosen based on prior research showing significant correlations between demographic factors and dengue incidence. Population and density are widely recognized as key determinants of vector-borne disease spread, while case counts provide the baseline for predictive modeling.

b. Data Merging

After data selection, the datasets for each variable were merged to ensure compatibility and consistency across variables. This merging process was conducted based on the target regions of prediction, ensuring that each region had a complete and integrated set of variables. This step was essential to avoid missing data and to maintain uniformity across the dataset, enabling robust comparative and predictive analysis.

c. Data Standardization

Before conducting further analysis, a standardization process was applied to adjust the scale of selected features, especially for modeling and forecasting purposes. The StandardScaler function from the Scikit-learn library was used for this purpose. This function plays a critical role in rescaling the dataset values to a defined scale. This step is important to prevent variables with larger scales from dominating the calculations, which could affect the results. Standardization is essential to meet algorithm assumptions and improve model performance [21].

3.3 Time Series Split

Time Series Split is a cross-validation method specifically designed for time series data, where observations are recorded at consistent time intervals. Unlike conventional cross-validation, Time Series Split does not shuffle the data, as the chronological order must be preserved. In each split, the training set consists of data from the beginning up to a certain point, while the test set includes subsequent data [22], [23]. The method works as a variation of k-fold cross-validation, where in each iteration, the test indices shift forward in time. This ensures that the test data always comes from a later time period than the training data, maintaining the temporal structure of the dataset, as illustrated in Figure 2 [23].

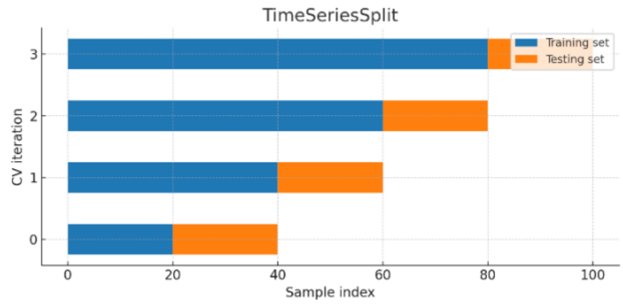

FIGURE 2. TIME SERIES SPLIT ILLUSTRATION

Figure 2 illustrates the concept of TimeSeriesSplit, a cross-validation strategy specifically designed for time series data. Unlike random splitting, this method respects the temporal order of observations by progressively expanding the training set and shifting the testing set forward in each iteration. As shown in the chart, each CV iteration (from 0 to 3) includes a larger portion of earlier data for training (blue) while reserving a later, non-overlapping segment for testing (orange). This approach ensures that future data is never used to predict past outcomes, preserving the integrity of time-dependent modeling. The x-axis represents the sample index (0 to 100), and the y-axis shows the iteration number, clearly demonstrating how the splits evolve over time.

3.4 Model Training

Linear Regression is a model used to analyze and estimate the value of a dependent variable based on a range of values from independent variables [24]. In this method, an observed outcome variable (dependent variable) is likely influenced by one or more other variables (independent variables), so the linear relationship between these variables can be analyzed to understand and predict existing patterns [25]. Regression analysis is widely used in big data statistical analysis because regression models are popular in data processing aimed at estimating the value of a dependent variable based on independent variable values [26].

The linear regression algorithm has several advantages, one of which is its simple structure. Based on the number of input variables, linear regression analysis is divided into two types: simple linear regression, which uses one independent variable, and multiple linear regression, which involves two or more independent variables [27]. Multiple linear regression is an extension of simple linear regression that involves more than one independent variable. Although the number of variables increases, the method is still referred to as 'linear' because it is assumed that the response variable has a direct relationship with the linear combination of the independent variables [28]. The multiple linear regression model has the following equation :

$$Y = a + b_1.X_1 + b_2.X_2 + \cdots + b_n.X_n \quad (1)$$

Where Y is the predicted value, a is the constant, X_1 and X_2 are the independent variables, while b_1 and b_2 are the coefficients of X_1 and X_2, respectively. The coefficient b_1 indicates the change in the value of Y for every one-unit increase in X_1 while keeping X_2 constant, whereas b_2 indicates the change in Y for every one-unit increase in X_2 while keeping X_1 constant.

To find the values of a, b_1, and b_2, you can use equations 2, 3, and 4 respectively.

$$a = \overline{Y} - b_1\overline{X_1} - b_2\overline{X_2} \quad (2)$$

$$b_1 = \frac{(\sum X_2^2)(\sum X_1Y) - (\sum X_1X_2)(\sum X_2Y)}{(\sum X_1^2)(\sum X_2^2) - (\sum X_1X_2)^2} \quad (3)$$

$$b_2 = \frac{(\sum X_2^2)(\sum X_2Y) - (\sum X_1X_2)(\sum X_2Y)}{(\sum X_1^2)(\sum X_2^2) - (\sum X_1X_2)^2} \quad (4)$$

Linear Regression is suitable for this dataset because it captures linear relationships between variables, offers simple yet interpretable results, and ensures efficient training. Combined with time series cross-validation like TimeSeriesSplit, it preserves temporal integrity, making it reliable for forecasting and trend analysis.

3.5 Model Evaluation

a. Mean Absolute Percentage Error (MAPE)

MAPE is used to assess the accuracy of prediction results by calculating the average percentage error (difference) between actual values and predicted values. MAPE serves as an evaluation tool in the prediction process, with accuracy adjusted to time series data and expressed as a percentage. The smaller the resulting MAPE value, the closer it is to the actual value. A prediction model is considered highly accurate if the MAPE value is less than 10%, and is categorized as good if it falls within the range of 10% to 20%. Mathematically, MAPE is expressed in Equation (5) [29].

$$\left(\frac{1}{n}\right)\sum_{t=1}^{n}\left|\frac{X_t - F_t}{X_t}\right| \quad (5)$$

Where X_t represents the actual data at period t, F_t is the predicted value at period t, and n denotes the total number of data points used. The interpretation of MAPE values is shown in Table 1 as follows.

TABLE 10. MAPE INTERPRETATION

MAPE (%)	Interpretation
<10	Very accurate
10 – 20	Accurate
20 – 50	Reasonable
> 50	Inaccurate

b. Root Mean Squared Error (RMSE)

In addition to using MAPE, model evaluation can also be performed using RMSE (Root Mean Squared Error) as an assessment metric. RMSE is the square root of the average of the squared differences between actual values and predicted values [30]. The RMSE formula is presented in Equation (6) as follows:

$$\sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}} \quad (6)$$

MAPE was chosen because it expresses prediction errors as percentages, making accuracy easy to interpret across different scales of data. RMSE was selected because it penalizes larger errors more strongly, providing a clear measure of overall deviation in the original units. Together, they offer complementary insights into both relative accuracy and absolute error magnitude.

4. RESULT AND DISCUSSION

This section presents the results of the conducted research, covering the data collection process, data splitting using the time series split method, model training and the evaluation of the model's performance in predicting the number of dengue fever cases. The performance analysis of the model is based on key evaluation metrics, namely RMSE and MAPE.

4.1 Implementation of the Research Phase

a. Preprocessing Data

The data preprocessing stage involves a series of processes carried out to prepare raw data before it is used in model development. Preprocessing is crucial because it ensures that the data is clean, consistent, and properly scaled, reducing biases or distortions that could compromise the model's accuracy. This stage includes data selection filtering, data merging, and data standardization. The results of the data preprocessing for each region can be seen in Tables 11 through 15, which illustrate the combined data from the five regions that are the main focus of the predictive analysis: Bandung City, Bandung Regency, Bogor Regency, Bekasi City, and Depok City.

TABLE 11. PREPROCESSING RESULTS FOR BANDUNG CITY DATASET

Year	Number of Cases	Population Density	Total Population	Rainfall
2019	4424	14794	2480464	169.2917
2020	4424	14916	2500967	201.55
2021	3743	15076	2527854	180.8917
2022	5205	15277	2545005	192.6167

TABLE 12. PREPROCESSING RESULTS FOR BANDUNG REGENCY DATASET

Year	Number of Cases	Population Density	Total Population	Rainfall
2019	2635	2015	3561679	268.38
2020	2303	2027	3583056	233.9
2021	2002	2055	3633437	243.44
2022	4191	2130	3708344	170.6

TABLE 13. PREPROCESSING RESULTS FOR BOGOR REGENCY DATASET

Year	Number of Cases	Population Density	Total Population	Rainfall
2019	1210	1740	4715924	203.6667
2020	1296	1893	5132355	283.775
2021	2220	1965	5327131	292.5833
2022	1953	1830	5473476	286.333

TABLE 14. PREPROCESSING RESULTS FOR BEKASI CITY DATASET

Year	Number of Cases	Population Density	Total Population	Rainfall
2019	2373	11852	2448830	107.912
2020	1646	11929	2464719	127.25
2021	2006	11947	2468448	149.2727
2022	2442	11671	2486251	154.6667

TABLE 15. PREPROCESSING RESULTS FOR DEPOK CITY DATASET

Year	Number of Cases	Population Density	Total Population	Rainfall
2019	2200	9275	1857734	207
2020	1276	9351	1872996	259.5
2021	3155	9453	1893321	278.2727
2022	2234	9605	1920182	451.25

The code used to implement this data standardization step is presented in Program Code 1.

Program Code 1. Data Standaridzation

```
#Import the feature scaling library to include
StandardScaler from sklearn.preprocessing
importStandardScaler.
#Feature standardization
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
#Show data standardization results
X_scaled_df=pd.DataFrame(X_scaled,
columns=X.columns)
print("\nHasil Standarisasi Data:")
print(X_scaled_df.head())
```

Program Code 1 is used to standardize feature data using the StandardScaler from the sklearn.preprocessing library. This standardization transforms each feature to have a mean value of 0 and a standard deviation of 1, ensuring uniform scaling across all features. The process is performed using the fit_transform() method on the data X, and the result is stored in X_scaled. The scaled data is then converted into a DataFrame with the same column names as X for easier readability. Finally, the first five rows of the standardized DataFrame are displayed on the screen. Tables 16 through 20 present the results of the data standardization, showing that each variable has been transformed to the same scale.

b. Time Series Split

After obtaining the datasets for each of the five regions, the Time Series Split method with four splits (n_splits = 4) was applied to each dataset. This step was carried out to ensure that the model validation process takes into account the sequential patterns based on the chronological order of the data. The code is presented in Program Code 2.

Program Code 2. Time Series Split

```
#Import the TimeSeriesSplit library for cross
validation on time series data from
sklearn.model_selectionimport TimeSeriesSplit.
#Time Series Split
tscv = TimeSeriesSplit(n_splits=4)
# Saving evaluation results
rmse_scores = []
mape_scores = []
results = []
#Perform iteration for each split
for i, (train_index, test_index) in
enumerate(tscv.split(X_scaled)):
#Split the data into train and test
  X_train, X_test = X_scaled[train_index],
   X_scaled[test_index]
  y_train, y_test = y.iloc[train_index],
   y.iloc[test_index]
```

Program Code 2 implements TimeSeriesSplit with 4 splits to perform cross-validation on time series data, dividing the data sequentially according to its chronology. Three lists are prepared to store the evaluation results: rmse_scores, mape_scores, and results. In each iteration, the data is split into training and testing sets based on the indices generated by tscv.split(X_scaled) using a for loop. A visualization can be seen in Figure 2.

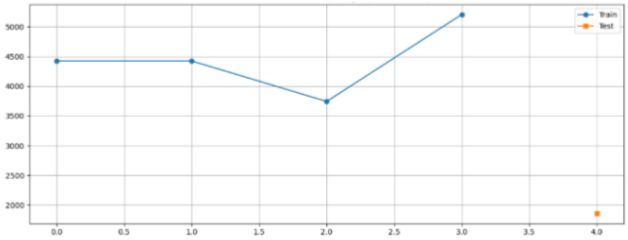

FIGURE 3. TIMESERIES SPLIT VISUALIZATION

c. Model Training

Next, the Linear Regression model was trained using the data splits from the Time Series Split. In each iteration, the model was trained on the training data and tested on the validation data to evaluate prediction performance. Program Code 3 presents the code used to carry out the model training process.

Program Code 3 Linear Regression Model Training

```
#Import LinearRegression to build a linear
regression model from sklearn.linear_model
import LinearRegression.
#Initialize and train a linear regression model
model = LinearRegression() model.fit(X_train,
y_train).
#Prediction y_pred = model.predict(X_test)
```

Program Code 3 is used to build and train a linear regression model using LinearRegression from sklearn.linear_model. The model object is first initialized, then trained using the training data X_train and y_train with the fit() method. After the model is trained, it is used to predict target values on the test data X_test using the predict() method. The prediction results are stored in the variable y_pred. Time Series Split is more appropriate for forecasting dengue hemorrhagic fever (DHF) cases because it preserves the chronological order of observations, ensuring that training data always precedes test data. This prevents information leakage from future values into past predictions, which can occur in conventional cross-validation methods that shuffle data. By shifting the test set forward in time at each iteration, Time Series Split provides a realistic evaluation of how the model performs when predicting unseen future cases. This makes it particularly suitable for forecasting, where maintaining temporal dependencies is critical. The results of the Linear Regression model training can be seen in Table 16.

TABLE 16. MODEL TRAINING RESULTS

Region	Intercept (b0)	Rainfall Coefficient (b1)	Chief Population Coefficient (b2)	Population Coefficient (b3)
Kota Bandung	4419,1072	209,738	3767,0755	-3601,3372
Kab. Bandung	3084,4264	-478,1170	3259,5397	-2862,2902
Kab. Bogor	2451,8395	-1422,1434	530,8860	1044,4530
Kota Bekasi	5513,6805	-3373,0223	1179,0801	6515,2874
Kota Depok	4485,0401	-3547,6021	41807,3516	-37316,417

Table 16 shows that the results of the Linear Regression model training produce a specific regression equation for each region. Based on the training results presented in Table 16, distinct Linear Regression equations were obtained for each region in West Java Province. These differences are reflected in the variation of coefficient values across the independent variables, namely rainfall, population density, and total population.

For instance, Bandung City has a positive rainfall coefficient of 209.738, indicating that an increase in rainfall tends to raise the number of dengue cases. In contrast, other regions such as Bandung Regency, Bogor Regency, Bekasi City, and Depok City exhibit negative rainfall coefficients, suggesting an opposite relationship. Moreover, substantial variations are observed in the coefficients of population density and total population. For example, Depok City shows a population density coefficient of 41,807.3516 and a total population coefficient of -37,316.417, which are considerably larger compared to other regions. These differences arise from the diverse characteristics of the data, including scale, annual fluctuations, and the patterns of relationships among variables.

d. Model Evaluation

After training the linear regression model, evaluation was conducted using two metrics: MAPE to measure prediction accuracy in percentage, and RMSE to assess the overall prediction error. These evaluation results are used to compare model performance and determine whether improvements are necessary. Program Code 4 is used to perform the model evaluation step.

Program Code 4. Model Evaluation

```
#Import functions to evaluate the predictive
accuracy of regression models from
sklearn.metrics import mean_squared_error,
mean_absolute_error
#Calculate RMSE and MAPE
  rmse=np.sqrt(metrics.mean_squared_error(y_te
    st, y_pred))
  mape = np.mean(np.abs((y_test - y_pred) /
    y_test)) * 100
  rmse_scores.append(rmse)
  mape_scores.append(mape)
```

Program Code 4 is used to evaluate the prediction accuracy of the regression model. The sklearn.metrics library is utilized to calculate the mean squared error and mean absolute error. The Root Mean Squared Error (RMSE) is computed as the square root of the mean squared error, while the Mean Absolute Percentage Error (MAPE) is calculated from the relative difference between the actual and predicted values. The resulting RMSE and MAPE values are stored in the lists rmse_scores and mape_scores for further analysis.

4.2 Experimental Results

After conducting a series of experiments using the Linear Regression method, a total of 20 iterations were performed, with each region analyzed through 4 iterations. The MAPE and RMSE results from each iteration were summed and then averaged to obtain the final evaluation score of the model for each region. The detailed results of these experiments are presented in Table 17.

TABLE 17. MODEL EVALUATION RESULTS

Regions	Average MAPE	Average RMSE
Kota Bandung	45.82	1216.105
Kab. Bogor	71.465	1376.2275
Kota Bekasi	81.85	1326.625
Kab. Bandung	103.41	1376.2275
Kota Depok	537.415	6763.015

The results in Table 17 reveal significant variation in model performance across different regions, as measured by Average MAPE and Average RMSE. Kota Bandung demonstrates the most accurate predictions, with the lowest Average MAPE of 45.82 and RMSE of 1216.105, indicating relatively reliable model outputs. In contrast, Kota Depok stands out with the highest error rates an Average MAPE of 537.415 and RMSE of 6763.015 suggesting substantial discrepancies between predicted and actual values in that region. Kabupaten Bandung also shows notably poor performance, with a high MAPE of 103.41 and RMSE matching Kabupaten Bogor at 1376.2275. Interestingly, while Kota Bekasi has a higher MAPE (81.85) than Kabupaten Bogor (71.465), its RMSE is slightly lower, implying that while percentage errors are greater, the absolute prediction errors may be less severe. Overall, the model appears to perform best in urban areas like Kota Bandung and worst in regions like Kota Depok, possibly due to data variability, population density, or other local factors affecting prediction accuracy.

4.3 Discussion

Overall, the model provides reasonably good predictions in certain areas but requires further development to improve accuracy, particularly in regions with high data variability. The model performance for each region is visualized in Figure 3 below.

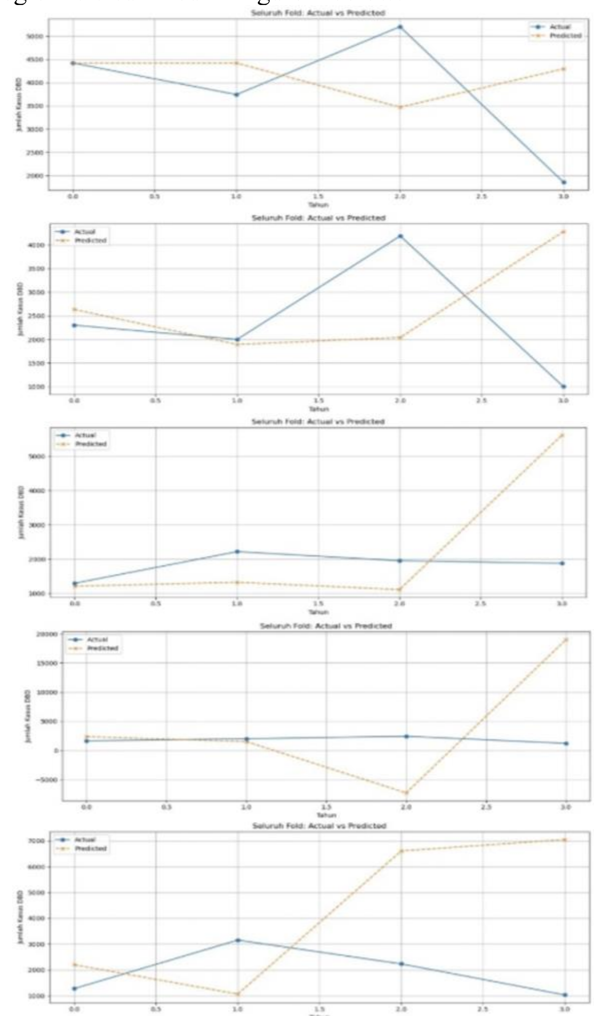

FIGURE 4. TIME MODEL PERFORMANCE VISUALIZATION FROM TOP TO BOTTOM: BANDUNG CITY, BANDUNG REGENCY, BOGOR REGENCY, BEKASI CITY, AND DEPOK CITY.

Figure 3 shows a comparison between the actual and predicted values of dengue fever cases in the five analyzed regions Bandung City, Bandung Regency, Bogor Regency, Bekasi City, and Depok City based on several folds from the Time Series Split cross-validation process. It can be observed that predictions from some folds are more accurate than others. This variation is caused by differences in data patterns within each fold, which affect the model's prediction results.

5. CONCLUSIONS

Among the five regions analyzed in this study, Bandung City yielded the most accurate prediction results, with an average Mean Absolute Percentage Error (MAPE) of 45.82% and a Root Mean Square Error (RMSE) of 1216.105. These figures indicate that the Linear Regression method is reasonably suitable for predicting dengue fever cases in this area compared to the others. The research strengths lie in its systematic data preprocessing, including filtering, merging, and standardization, which ensured consistency and comparability across datasets. The selection of demographic variables population size, population density, and case counts was grounded in their established correlation with dengue incidence, providing a solid basis for predictive modeling. Despite these strengths, several limitations were identified. The most critical challenge was the incomplete availability of rainfall data, with irregularities in annual records hindering the integration of environmental factors into the model. This limitation restricted the ability to fully capture mosquito breeding dynamics and transmission risks, thereby affecting prediction accuracy. Future research should focus on expanding the range of independent variables to include climatic and environmental indicators such as temperature and humidity, which are known to play a significant role in mosquito breeding and disease transmission. In addition, the integration of advanced machine learning approaches such as Random Forest, Gradient Boosting, or Neural Networks could enhance predictive accuracy by capturing complex, non-linear relationships that traditional regression methods may overlook. The effectiveness of these models would be further strengthened by the availability of real-time and comprehensive datasets, which would allow for more dynamic monitoring and forecasting of dengue incidence. Ultimately, these improvements would support the scalability of predictive models, enabling their application across larger regions and contributing to more effective public health planning and dengue control strategies.

REFERENCES

[1] S. Sulistyawati, H. Yuliansyah, T. Sukesi, A. Khusna, S. Mulasari, F. Tentama, B. Sudarsono, and F. Ghozali, "Rapid Appraisals of the Transformation Strategy Required to Sustain Dengue Vector Control During and After the COVID-19 Pandemic in Indonesia," *Risk Manag Healthc Policy*, vol. 16, pp. 93–100, 2023, doi: 10.2147/RMHP.S391933.

[2] A. Saputra, Y. Ariyani, and P. Dewi, "Faktor yang Berhubungan dengan Lingkungan Fisik dan Kebiasaan Keluarga terhadap Penyakit Demam Berdarah Dengue (DBD)," *Jurnal 'Aisyiyah Medika*, vol. 2, no. 2, Aug. 2023, doi: https://doi.org/10.36729/jam.v8i1.

[3] S. Sulistyawati, H. Yuliansyah, T. Sukesi, A. Khusna, S. Mulasari, F. Tentama, B. Sudarsono, and F. Ghozali, "Knowledge, Attitudes, and Behavior of Students in Combating Mosquito Nest: A Case Study in High School," *Asian Journal of Research in Infectious Diseases*, vol. 15, no. 2, pp. 24–30, Feb. 2024, doi: 10.9734/ajrid/2024/v15i2328.

[4] A. Ayuningtyas, "Analisis Hubungan Kepadatan Penduduk dengan Kejadian Demam Berdarah Dengue (DBD) di Provinsi Jawa Barat," *Jurnal Ilmiah Permas: Jurnal Ilmiah STIKES Kendal* , vol. 13, Apr. 2023, doi: https://doi.org/10.32583/pskm.v13i2.772.

[5] S. Sulistyawati, T. Sukesi, H. Yuliansyah, A. Khusna, and S. Mulasari, "Individual attentiveness in vector control should be strengthened during and after the COVID-19 pandemic," *Front Public Health*, vol. 10, Nov. 2022, doi: https://doi.org/10.3389/fpubh.2022.1055509.

[6] T. Sukesi, S. Mulasari, H. Yuliansyah, F. Tentama, and L. Nafiati, "Pengendalian demam berdarah dengue (DBD) 'Di Rumah Aja' di Wilayah Kerja Puskesmas Gamping 1 Sleman," *Prosiding Seminar Nasional Hasil Pengabdian Kepada Masyarakat Universitas Ahmad Dahlan*, pp. 586–591, Oct. 2021.

[7] T. Sukesi, H. Yuliansyah, S. Sulistyawati, A. Khusna, S. Mulasari, F. Tentama, B. Sudarsono, and F. Ghozali, "Home Environment and Larva Indices: A Cross-Sectional Study in the Indonesian Transition to Endemic COVID-19," *Jurnal Kesehatan Masyarakat*, vol. 19, no. 1, pp. 160–166, Jul. 2023, doi: 10.15294/kemas.v19i1.42605.

[8] T. Sukesi, H. Yuliansyah, S. Sulistyawati, A. Khusna, S. Mulasari, F. Tentama, B. Sudarsono, and F. Ghozali, "Association Between Home Environment and Larva Indices: A Cross-Sectional Study in the Indonesian Transition to Endemic COVID-19," *Jurnal Kesehatan Masyarakat*, vol. 19, no. 1, pp. 160–166, Jul. 2023, doi: http://dx.doi.org/10.15294/kemas.v19i1.42605.

[9] A. M. Ningtyas, I. K. Lubis, and G. B. Herwanto, "Monitoring Persebaran Penyakit Demam Berdarah Dengue dengan Memanfaatkan Data Berita Online," *Jurnal Kesehatan Vokasional*, vol. 4, no. 2, p. 105, May 2019, doi: 10.22146/jkesvo.44691.

[10] A. Ichwani and H. Wibawa, “Prediksi Angka Kejadian Demam Berdarah Dengue (DBD) Berdasarkan Faktor Cuaca Menggunakan Metode Extreme Learning Machine (Studi Kasus Kecamatan Tembalang),” *Jurnal IPTEK*, vol. 23, no. 1, 2019, doi: 10.31284/j.iptek.2019.v23i1.

[11] M. Aditya, I. N. Sukajaya, and I. Gunadi, “Forecasting Jumlah Pasien DBD di BRSUD Kabupaten Tabanan Menggunakan Metode Regresi Linier,” *Bali Medika Jurnal*, vol. 10, no. 1, pp. 1–12, Jun. 2023, doi: 10.36376/bmj.v9i3.

[12] S. R. Cholil, A. F. Dwijayanto, and T. Ardianita, “Prediksi Penyakit Demam Berdarah di Puskesmas Ngemplak Simongan menggunakan Algoritma C4.5,” *SISTEMASI : Jurnal Sistem Informasi*, vol. 9, no. 3, Jul. 2020, doi: https://doi.org/10.32520/stmsi.v9i3.898.

[13] N. T. Lestari and A. Witanti, “Analisis Prediksi Kasus DBD Berdasarkan Faktor Cuaca Dengan Multivariat ARIMA,” *PETIR : Jurnal Pengkajian dan Penerapan Teknik Informatika*, vol. 16, Sep. 2023, doi: https://doi.org/10.33322/petir.v16i2.2117.

[14] A. Zaki, M. S. Wahyuni, I. Irwan, and A. Rahman, “Peramalan Jumlah Penderita Demam Berdarah Dengue Menggunakan Metode Seasonal-ARIMA,” *ARRUS Journal of Mathematics and Applied Science*, vol. 3, no. 2, 2023, doi: 10.35877/mathscience2143.

[15] Y. Zebua, S. Dur, and R. Lubis, “Penerapan Metode Multiplicative Decomposition dalam Memprediksi Jumlah Kasus DBD di RSU. Haji Medan,” *Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 4, no. 1, 2023, doi: 10.46306/lb.v4i1.

[16] I. M. Karo, “Prediksi Penyebaran Demam Berdarah Dangue dengan Algoritma Hybrid Autoregressive Integrated Moving Average dan Artificial Neural Network: Studi Kasus di Kabupaten Bandung,” *Journal of Software Engineering, Information and Communication Technology (SEICT)*, vol. 2, no. 1, pp. 27–36, Feb. 2022, doi: 10.17509/seict.v2i2.40222.

[17] B. Khotimah and E. Rochman, “Model Peramalan Jumlah Penyakit Demam Berdarah dengan Pendekatan Metode Fuzzy Linear Regression (FLR),” *Jurnal Ilmiah NERO*, vol. 6, no. 1, 2021, doi: https://doi.org/10.21107/nero.v6i1.215.

[18] A. R. Muhajir, E. Sutoyo, and I. Darmawan, “Forecasting Model Penyakit Demam Berdarah Dengue Di Provinsi DKI Jakarta Menggunakan Algoritma Regresi Linier Untuk Mengetahui Kecenderungan Nilai Variabel Prediktor Terhadap Peningkatan Kasus,” *Fountain of Informatics Journal*, vol. 4, no. 2, p. 33, Nov. 2019, doi: 10.21111/fij.v4i2.3199.

[19] R. Lutfianawati, Ngadino, and Marlik, “Prediksi Kejadian Demam Berdarah Dengue di Kecamatan Papar Kabupaten Kediri Tahun 2016-2021,” *ASPIRATOR - Journal of Vector-borne Disease Studies*, vol. 14, no. 1, pp. 57–68, Dec. 2022, doi: 10.22435/asp.v14i1.5892.

[20] T. Gori, A. Sunyoto, and H. Al Fatta, “Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 215–224, Feb. 2024, doi: 10.25126/jtiik.20241118074.

[21] A. Reza and M. Rohman, “Prediction Stunting Analysis Using Random Forest Algorithm and Random Search Optimization,” *Journal of Informatics and Telecommunication Engineering*, vol. 7, no. 2, p. 539, Jan. 2024, doi: 10.31289/jite.v7i2.10628.

[22] scikit-learn, “TimeSeriesSplit.” Accessed: Apr. 28, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html

[23] S. Aisyah, N. Ulinnuha, and A. Hamid, “Penerapan Extreme Learning Machine dalam Meramalkan Harga Minyak Sawit Mentah,” *KUBIK: Jurnal Publikasi Ilmiah Matematika*, vol. 7, no. 2, p. 101, Nov. 2022, doi: https://doi.org/10.15575/kubik.v7i2.20460.

[24] B. Al Haddad, A. Bahtiar, and G. Dwilestari, “Implementasi Algoritma Regresi Linear Berganda untuk Memprediksi Biaya Asuransi Kesehatan,” *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 6, Mar. 2024.

[25] Z. Muttaqin and E. Srihartini, “Penerapan Metode Linier Regresi Sederhana untuk Prediksi Persediaan Obat Jenis Tablet,” *Sistem Informasi* |, vol. 9, no. 1, pp. 12–16, doi: http://dx.doi.org/10.30656/jsii.v9i1.4426.

[26] F. H. Hamdanah and D. Fitrianah, “Analisis Performansi Algoritma Linear Regression dengan Generalized Linear Model untuk Prediksi Penjualan pada Usaha Mikra, Kecil, dan Menengah,” *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 10, no. 1, p. 23, Apr. 2021, doi: 10.23887/janapati.v10i1.31035.

[27] F. Riandari, H. T. Sihotang, and H. Husain, “Forecasting the Number of Students in Multiple Linear Regressions,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, p. 251, Mar. 2022, doi: 10.30812/matrik.v21i2.1348.

[28] R. Tyasnurita, R. R. Luthfiansyah, and M. R. Brameswara, “Gold Price Forecasting using

Multiple Linear Regression Method," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 9, no. 3, pp. 339–344, Jun. 2023, doi: 10.33330/jurteksi.v9i3.1748.

[29] A. Sumari, M. Musthafa, and D. Putra, "Perbandingan Kinerja Metode-Metode Prediksi pada Transaksi Dompet Digital di Masa Pandemi," *Jurnal RESTI*, vol. 4, no. 3, p. 644, 2020, doi: https://doi.org/10.29207/resti.v4i4.2024.

[30] L. S. Ihzaniah, A. Setiawan, and R. W. N. Wijaya, "Perbandingan Kinerja Metode Regresi K-Nearest Neighbor dan Metode Regresi Linear Berganda pada Data Boston Housing," *Jambura Journal of Probability and Statistics*, vol. 4, no. 1, pp. 17–29, May 2023, doi: 10.34312/jjps.v4i1.18948.

AUTHORS

Muhammad Khysam Firdaus

He is an undergraduate student in the Informatics Study Program at Universitas Ahmad Dahlan, Yogyakarta, Indonesia. His research interests include data analysis and artificial intelligence. He is actively involved in academic activities and various information technology projects. He also enjoys analytical work as a way to enhance critical thinking and broaden his knowledge.

Herman Yuliansyah

He is a Associate Professor (Lektor Kepala) in the Department of Informatics at Universitas Ahmad Dahlan. He earned his Bachelor of Engineering (S.T.) in Electrical Engineering from Universitas Muhammadiyah Yogyakarta, his Master of Engineering (M.Eng.) from the Faculty of Engineering at Universitas Gadjah Mada, and his Doctor of Philosophy (Ph.D.) from the Faculty of Information Science and Technology at Universiti Kebangsaan Malaysia.