E-Komtek Journal
Vol. 9, No. 1. (2025) pp. 111-122
https://jurnal.politeknik-kebumen.ac.id/index.php/E-KOMTEK
p-ISSN: 2580-3719 e-ISSN: 2622-3066

Comparative Analysis of the Accuracy of Multiple Linear Regression Method and Ridge Regression Method in Predicting Dengue Fever Cases in South Tangerang City

Dina Aulia[1], Herman Bedi Agtriadi[2], Luqman[3]

[1-3] Master of Computer Science, Institut Teknologi PLN, Indonesia, 12430

dina2330002@itpln.ac.id

https://doi.org/10.37339/e-komtek.v9i1.2292

Published by Piksi Ganesha Polytechnic Indonesia

Info Articles
Submitted:
23-01-2025
Revised:
04-06-2025
Accepted:
26-06-2025
Online first :
30-06-2025

Abstract

One of the main health issues in South Tangerang City is dengue fever (DBD). This study aims to compare the accuracy of Multiple Linear Regression and Ridge Regression methods in predicting the number of DBD cases using weather data such as temperature, humidity, and average rainfall. The analysis process includes preprocessing, splitting the dataset into training and testing data, and applying both regression methods. To determine the prediction error rate, model accuracy is evaluated using the Mean Absolute Percentage Error (MAPE) metric. The results indicate that Ridge Regression performs better for datasets with high multicollinearity, yielding a MAPE value of 20.12%, while Multiple Linear Regression is more effective for datasets with low feature correlation, showing a MAPE value of 44.6%. It is hoped that this research can improve mitigation and planning for DHF cases in South Tangerang City by choosing the appropriate approach.

Keywords *: Ridge Regression; Multiple Linear Regression; Dengue Fever Prediction; MAPE*

Abstrak

Salah satu masalah kesehatan utama di Kota Tangerang Selatan adalah demam berdarah dengue (DBD). Penelitian ini bertujuan untuk membandingkan keakuratan metode Regresi Linier Berganda dan Regresi Ridge dalam memprediksi jumlah kasus DBD dengan menggunakan data cuaca seperti suhu, kelembapan, dan rata-rata curah hujan. Proses analisis meliputi preprocessing, membagi dataset menjadi data training dan data testing, dan menerapkan kedua metode regresi. Untuk mengetahui tingkat kesalahan prediksi, akurasi model dievaluasi dengan menggunakan metrik Mean Absolute Percentage Error (MAPE). Hasil penelitian menunjukkan bahwa Ridge Regression berkinerja lebih baik untuk dataset dengan multikolinieritas tinggi, menghasilkan nilai MAPE sebesar 20,12%, sedangkan Regresi Linier Berganda lebih efektif untuk dataset dengan korelasi fitur yang rendah, menunjukkan nilai MAPE sebesar 44,6. Diharapkan penelitian ini dapat meningkatkan mitigasi dan perencanaan kasus DBD di Kota Tangerang Selatan dengan memilih pendekatan yang tepat.

Kata kunci :

1. Introduction

The South Tangerang City Health Office is a government agency at the regional level, precisely in the South Tangerang City, which is responsible for planning, developing, implementing and evaluating health policies and programs in South Tangerang. One of the responsibilities of the Health Office is epidemiological supervision and monitoring to control outbreaks and infectious diseases (Health Service, 2019) which are the topic of this research are cases of Dengue Hemorrhagic Fever (DHF) that occurred in South Tangerang City **[1-3]**.

Dengue Hemorrhagic Fever (DHF) is a disease caused by infection by the dengue virus transmitted through the bite of the Aedes aegypti mosquito which is a disease with the fastest spreading rate in the world for the past 50 years. According to *the World Health Association* (WHO), the Southeast Asia and Western Pacific region is the region with the most cases of dengue fever in the world. Countries in this region represent almost 75% of the total cases of dengue fever worldwide, with a population of around 1.3 billion people at risk of contracting dengue fever (who.int) **[4]**.

The increase and spread of dengue cases are most likely caused by high population mobility, metropolitan area development, environmental changes, changes in population density and distribution, and epidemiological elements that need to be further explored. Environmental changes affect rainfall, temperature and humidity which affect terrestrial and marine biological systems, including the proliferation of dengue vectors, especially the Aedes aegypti mosquito. **[5]**.

South Tangerang City is one of the endemic cities for dengue fever in Banten Province. In 2020, the number of dengue fever cases in this city increased to 2,183 with 23 people dying (South Tangerang City Health Office, 2020). Data from 2020 also shows that South Tangerang City has the highest number of dengue fever sufferers in Banten Province, namely 2,183 sufferers (Banten Provincial Health Office, 2020). Therefore, forecasting is needed *to* estimate the value of the data on the number of dengue fever cases in the future and to determine the tendency of the value of the predictor variables towards the increase in dengue fever cases in South Tangerang City.

Multiple Linear Regression is one of the commonly used statistical methods to model the relationship between a dependent variable and several independent variables. This method is quite simple and easy to interpret, but it has limitations in capturing non-linear patterns and complex interactions between variables **[6]**. On the other hand, ridge regression is a regression method that adds L2 regularization to overcome the problems of overfitting and multicollinearity. By using a penalty on the coefficients, ridge helps create a more stable, simple, and reliable model for making predictions, especially when there are many variables or high correlations among predictors **[7-9]**.

Prediction of the spread of dengue fever in South Tangerang City is very important because dengue fever is a disease caused by the dengue virus transmitted through the bite of the Aedes mosquito, which can cause outbreaks in various regions, especially in tropical countries. Dengue fever cases often increase significantly during the rainy season, and if left untreated, can cause many deaths. By utilizing data and predictive models, we can reduce the impact of this disease and protect the overall health of the community in South Tangerang City **[10]**.

This study aims to compare the performance of the DHF case forecasting model in South Tangerang City using the Multiple Linear Regression method and the Ridge Regression method. By comparing these two methods by exploring predictor variables that have a significant effect on the increase in DHF cases, such as rainfall, temperature and humidity. By understanding the dynamics of these variables, it is expected to determine which method has the best accuracy and which method is more effective in predicting the increase in DHF cases in South Tangerang City. The results of previous studies obtained a MAPE value of 44.6% using the multiple linear regression method, it is expected that by using the ridge regression method, a lower MAPE value can be obtained than previous studies in order to obtain the best accuracy in predicting DHF cases in South Tangerang City and can contribute to efforts to prevent and control DHF disease better, as well as assist the government and related agencies in planning more effective mitigation strategies **[11-13}**.

2. Method

This research method uses multiple linear regression method and ridge regression method by comparing the two methods to find out which method is more effective in predicting the increase in DHF cases in South Tangerang City.

1. Linear regression with multiple variables

This regression model attempts to model the relationship between two or more explanatory variables and the response variable by fitting a linear equation to the observed data. Each value of the independent variable x is associated with a value of the dependent variable y. The following is a multiple linear regression equation **[4]**.

$$\boldsymbol{Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n} \quad \textbf{(2.1)}$$

Where :

Y = Dependent variable (value to be predicted)

b_0 = Coefficient that describes the relationship between variable Y

$b_1, b_2, .., b_n$ = Regression coefficient

$X_1, X_2, .., X_n$ = Independent Variables

2. Ridge Regression

Ridge Regression is a linear regression method that addresses multicollinearity or overparameterization (too many features) by adding a penalty to the regression coefficients. The goal is to reduce model variance and improve predictive stability. By adding a regularization penalty λ to the square of the coefficients, Ridge Regression shrinks the coefficients, thereby reducing the model's sensitivity to changes in the data and avoiding overfitting. The larger the value of λ, the smaller the regression coefficients, which increases bias but reduces variance () **[14-15]**.

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \quad \textbf{(2.2)}$$

or in matrix form:

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y \quad \textbf{(2.3)}$$

Where :

X = feature matrix (independent variables).

Y = target vector (dependent variable).

λ= regularization parameters (hyperparameters to be specified).

I = identity matrix.

$\hat{\beta}_{ridge}$= the resulting Ridge regression coefficient.

This section explains the stages of research design starting from *identification problems* , then continued with *Data Collection* or data collection based on needs analysis, the next stages are *Preprocessing* , *Feature Classification* , *Evaluation,* to the last stage, namely *Forecasting Result* .

The research design has been designed in such a way that the data research runs well can be seen in Figure 1.

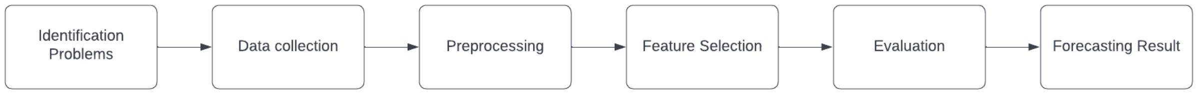

Figure 1. Research Design

There are several stages in this research. The stages start from problem identification to evaluation. The diagram of the research stages can be seen in **Figure 2**.

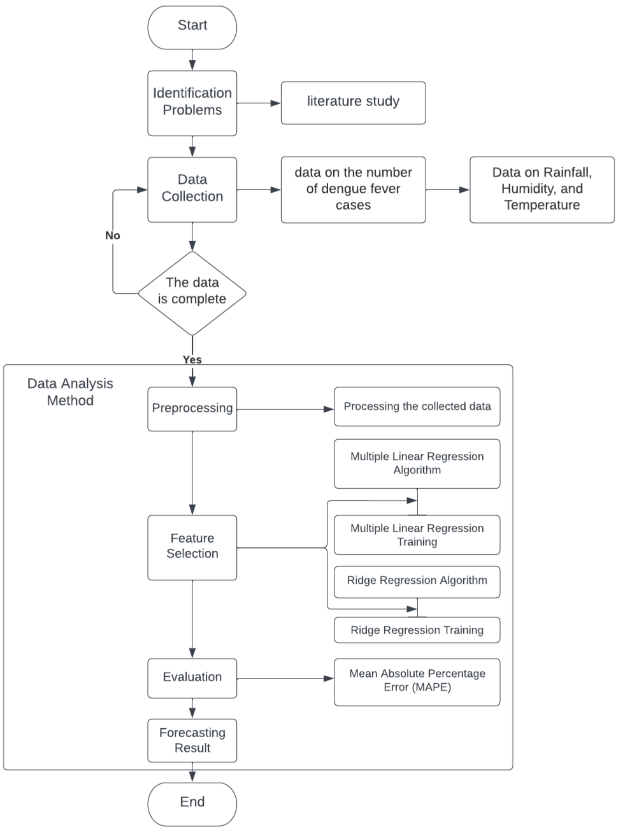

Figure 1. Research Workflow Diagram

Figure 3.1 explains the stages of research design in the application of Multiple Linear Regression Algorithm for the creation of a DHF disease forecasting model starting from Identification Problems, then continued with *Data Collection* or data collection based on needs analysis, the next stages are Preprocessing, Feature Classification, Evaluation, to the last stage, namely Forecasting Result. The research design has been designed in such a way that the data research runs well.

1. Identification Problems

At this stage, observations and identification of problems in the research object are carried out. Analyze the objectives and benefits of the research so as to obtain analysis results that are in accordance with the background and needs of the existing problems. Identification of problems in this study refers to the substance of making this *forecast* , where the forecasting results *produced* can help in predicting dengue fever outbreaks.

2. Data Collection

It is a stage in collecting data on the number of DHF cases, rainfall data, humidity data and temperature data as a dataset. The dataset used was obtained from the South Tangerang City Health Office and the Center for Meteorology, Climatology and Geophysics Region II South Tangerang.

3. Preprocessing

In this stage, the data that has been obtained is preprocessed before being used as input in the Multiple Linear Regression and Ridge Regression architecture.

4. Feature Selection

In this stage, the data on the number of DHF cases, rainfall data, humidity data and temperature data that have been preprocessed will be used as input for the next stage, namely selecting a subset of Multiple Linear Regression and Ridge Regression features. The subset of features that contributed and were tested in this study were Multiple Linear Regression and Ridge Regression. In this stage, a model will also be created.

5. Evaluation

error results generated from the model that has been obtained will be evaluated with the aim of ensuring whether the model used and tested has a low error .

$$\mathrm{MAPE} = \frac{\sum \frac{|X_i - F_i|}{X_i} \times 100\%}{n} \tag{2.4}$$

Where :

X_i = Dependent Variable

F_i = Prediction

n = Number of datasets

6. Forecasting Results

After the model goes through the evaluation stage , the Forecasting stage is carried out. The results of the model being tested with predicted data aim to ensure that the adjustment of the output results obtained is the same as the input.

3. Results and Discussion

3.1. Results

1. The results of the calculation *of the Mean Absolute Percentage Error* (MAPE) from the Multiple Linear Regression method

No	Tanggal	Jumlah Kasus (X_i)	Prediksi F_i	$\Sigma\frac{X_i - F_i}{X_i}$
1	Jan-16	79	58,156351	0,263844
2	Feb-16	85	58,56418	0,31101
3	Mar-16	71	59,647366	0,159896
4	Apr-16	63	57,377237	0,08925
5	May-16	78	55,757721	0,285157
6	Jun-16	69	45,577219	0,339461
7	Jul-16	71	42,974303	0,394728
8	Aug-16	42	41,693552	0,007296
9	Sep-16	38	44,255487	0,164618
10	Oct-16	20	49,164393	1,45822
11	Nov-16	23	51,385606	1,234157
12	Dec-16	16	48,749688	2,046855
13	Jan-17	27	48,204834	0,785364
14	Feb-17	26	51,144327	0,96709
15	Mar-17	40	46,05072	0,151268
16	Apr-17	16	50,182114	2,136382
17	May-17	20	46,556048	1,327802
18	Jun-17	19	43,504656	1,289719
19	Jul-17	14	35,706552	1,550468
20	Aug-17	19	19,534551	0,028134
21	Sep-17	12	24,944534	1,078711
22	Oct-17	12	42,399419	2,533285
23	Nov-17	22	47,748469	1,170385
24	Dec-17	19	44,571532	1,34587
25	Jan-18	21	45,514816	1,167372
26	Feb-18	27	50,620971	0,874851
27	Mar-18	30	48,850687	0,628356
28	Apr-18	34	51,798408	0,523483
29	May-18	27	43,435825	0,608734
30	Jun-18	25	36,374415	0,454977
31	Jul-18	76	20,28712	0,733064
32	Aug-18	20	13,625653	0,318717
33	Sep-18	17	10,504526	0,382087
34	Oct-18	9	21,941731	1,43797
35	Nov-18	50	45,236316	0,095274
36	Dec-18	148	44,840464	0,697024
37	Jan-19	86	48,41954	0,436982
38	Feb-19	49	53,204712	0,08581
39	Mar-19	58	50,462339	0,12996

Figure 3. Data on Number of Cases and Prediction Results

$$MAPE = \frac{\sum \frac{X_i - F_i}{X_i} \times 100\%}{n}$$

$$MAPE = \frac{37,457}{84} \times 100\%$$

$$MAPE = 0,445917$$

The predicted MAPE value obtained using Multiple Linear Regression was 0.445917% or 44.6%.

2. *the Mean Absolute Percentage Error (MAPE)* calculation from the Ridge Regression method

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import Ridge
from sklearn.preprocessing import StandardScaler, PolynomialFeatures
from sklearn.metrics import mean_absolute_percentage_error
from google.colab import files

# Fungsi untuk membersihkan outlier menggunakan metode IQR
def remove_outliers_iqr(data, column):
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return data[(data[column] >= lower_bound) & (data[column] <= upper_bound)]

# Fungsi untuk menambahkan fitur interaksi
def create_interactions(X):
    poly = PolynomialFeatures(degree=2, interaction_only=True, include_bias=False)
    return poly.fit_transform(X)

# Unggah dataset
print("Silakan unggah file dataset Anda (.csv):")
uploaded = files.upload()

# Membaca dataset yang diunggah
file_name = list(uploaded.keys())[0]
df = pd.read_csv(file_name)
```

Figure 4. Ridge Regression MAPE Calculation

In the initial step, the MAPE obtained was 44.6%. This could be caused by several factors that limit the model's ability to provide more accurate predictions.

2.1. Target Transformation Optimization

```
[ ] from scipy.stats import boxcox

    # Transformasi Box-Cox pada target
    y_transformed, fitted_lambda = boxcox(df['Total_Cases'] + 1)  # Menambahkan 1 untuk menghindari log(0)
    y_train, y_test = train_test_split(y_transformed, test_size=0.2, random_state=42)

    # Gunakan fitted_lambda untuk membalik transformasi nanti
```

Figure 5. Target Transformation Optimization

The logarithmic transformation of the target that has been used may not be entirely optimal. Experimenting with other transformations, such as *square root* or *box-cox* , can help make the distribution more normal, which makes it easier for the model to learn patterns.

2.2. Non-Linear Model Testing

```
[ ] from sklearn.ensemble import GradientBoostingRegressor

    # Membuat model Gradient Boosting
    gbr_model = GradientBoostingRegressor(n_estimators=500, learning_rate=0.01, max_depth=5, random_state=42)
    gbr_model.fit(X_train_scaled, y_train)

    # Melakukan prediksi
    y_pred_gbr_log = gbr_model.predict(X_test_scaled)

    # Transformasi kembali prediksi ke skala asli
    y_pred_gbr = np.expm1(y_pred_gbr_log)
    mape_gbr = mean_absolute_percentage_error(y_test_actual, y_pred_gbr) * 100

    print(f"\nMAPE Gradient Boosting Regressor: {mape_gbr:.2f}%")

    MAPE Gradient Boosting Regressor: 506.40%
```

Figure 6. Non-Linear Model Test

Use non-linear predictive models such as *Gradient Boosting Regressor* or *Random Forest Regressor* , which can capture more complex relationships between variables.

2.3. Advanced Feature Engineering

```
[ ] # Menambahkan fitur lag
    df['Total_Cases_Lag1'] = df['Total_Cases'].shift(1).fillna(method='bfill')
    df['Average_Rainfall_Lag1'] = df['Average_Rainfall'].shift(1).fillna(method='bfill')

    <ipython-input-16-8e7a86dfc7fa>:2: FutureWarning: Series.fillna with 'method' is deprecated and will raise in a future version. Use obj.ffill() or obj.bfill() instead.
      df['Total_Cases_Lag1'] = df['Total_Cases'].shift(1).fillna(method='bfill')
    <ipython-input-16-8e7a86dfc7fa>:3: FutureWarning: Series.fillna with 'method' is deprecated and will raise in a future version. Use obj.ffill() or obj.bfill() instead.
      df['Average_Rainfall_Lag1'] = df['Average_Rainfall'].shift(1).fillna(method='bfill')
```

Figure 7. Advanced Engineering Features

Add new features like:

a. Lag Features : Difference in dengue cases from the previous month.

b. Seasonal Average : Average temperature, humidity, or rainfall for the last few months.

c. Seasonal Effects : Month or season (rainy season).

2.4. Model validation with *cross-validation*

```
[ ] from sklearn.model_selection import cross_val_score

    # Evaluasi model dengan cross-validation
    scores = cross_val_score(ridge_model, X_train_scaled, y_train, cv=10, scoring='neg_mean_absolute_percentage_error')
    average_mape = -np.mean(scores) * 100
    print(f"\nAverage MAPE (CV): {average_mape:.2f}%")

    Average MAPE (CV): 20.12%
```

Figure 8. Model validation with *cross-validation*

Model validation using *cross-validation* on the dataset as a whole can provide a better picture of accuracy than a single data *split* , with this process the MAPE value changes to 20.12%.

2.5. Advanced outlier checking

```
import matplotlib.pyplot as plt
import seaborn as sns

sns.boxplot(df['Total_Cases'])
plt.title('Distribusi Target (Total Cases)')
plt.show()
```

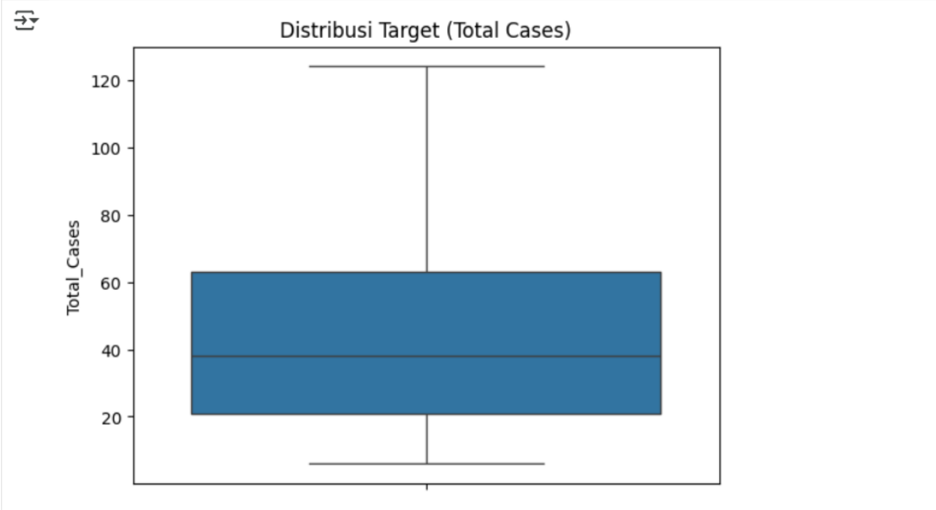

Figure 9. Advanced outlier check

Plot the target distribution to identify whether there are any outliers that have not been handled.

3.2. Discussion

This study was conducted based on data collection conducted at the South Tangerang City Health Office and the Meteorology, Climatology and Geophysics Center for Region II of South Tangerang City. This study was conducted to determine which method has higher accuracy and which method is more effective in predicting the increase in DHF cases in South Tangerang City.

In this study, the first thing to do is to prepare data on the number of DHF cases, rainfall, humidity and temperature which are used as parameters for predictive calculations. Here, researchers use data from 2016-2022 which are then made into a dataset in CSV format, to test predictions from this calculation using *the Mean Absolute Percentage Error* (MAPE). Then to find out the *Mean Absolute Percentage Error* (MAPE) value from the multiple linear regression method, the calculation was obtained in previous research using excel calculations and obtained a MAPE of 44.6% and to find out the MAPE value from the ridge regression method using google colab which goes through several stages in testing the MAPE value starting from increasing non-linear features, more accurate data cleaning, more comprehensive cross-validation, more extensive *Hyperparameter tuning and* Normalization and *Scaling* More Precise so that MAPE is obtained at 20.12%.

4. Conclusion

Based on the results of the research that has been conducted, several conclusions can be drawn, including:

a. In this study, the results show that Ridge Regression is more effective in handling datasets that have high multicollinearity compared to the Multiple Linear Regression method.
b. The results of *the Mean Absolute Percentage Error* (MAPE) obtained from the *forecasting results* of these two methods obtained a MAPE of 44.6% in the Multiple Linear Regression method, the MAPE value shows that the forecast is quite good, adequate and suitable for use and in the Ridge Regression method, a MAPE of 20.12% was obtained, the MAPE value shows that the forecast is good for use.

5. *Acknowledgement*

Acknowledgements to Dr. Herman Bedi Agtriadi, S.Si., M.Kom. As a supervisor of Master of Computer Science at the PLN Institute of Technology, so that with his guidance and assistance I as a student can complete this journal, where there are still many shortcomings that need to be corrected in this writing.

References

[1] AR Muhajir, E. Sutoyo, and I. Darmawan, "Forecasting Model of Dengue Hemorrhagic Fever Disease in DKI Jakarta Province Using Linear Regression Algorithm to Determine the Tendency of Predictor Variable Values Against the Increase in Cases," *Fountain of Informatics Journal* , vol. 4, no. 2, p. 33, Nov. 2019, doi: 10.21111/fij.v4i2.3199.

[2] M. Sarstedt and E. Mooi, "Regression Analysis," 2019, pp. 209–256. doi: 10.1007/978-3-662-56707-4_7.

[3] T. Dupré la Tour, M. Eickenberg, A. O. Nunez-Elizalde, and J. L. Gallant, "Feature-space selection with banded ridge regression," *Neuroimage* , vol. 264, Dec. 2022, doi: 10.1016/j.neuroimage.2022.119728.

[4] S. Rath, A. Tripathy, and AR Tripathy, “Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression models,” *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* , vol. 14, no. 5, pp. 1467–1474, Sept. 2020, doi: 10.1016/j.dsx.2020.07.045.

[5] E. Gothai, P. Natesan, RR Rajalaxmi, T. Vignesh, K. Srinithy, and TV Balaji, “Predictive analysis in determining the dissemination of infectious disease and its severity,” in *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021* , Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 1556–1562. doi: 10.1109/ICCMC51019.2021.9418228.

[6] SO Olukanmi, FV Nelwamondo, and NI Nwulu, “Utilizing Google Search Data with Deep Learning, Machine Learning and Time Series Modeling to Forecast Influenza-Like Illnesses in South Africa,” *IEEE Access* , vol. 9, pp. 126822–126836, 2021, doi: 10.1109/ACCESS.2021.3110972.

[7] MI Ullah, M. Aslam, and S. Altaf, “lmridge: A Comprehensive R Package for Ridge Regression.”

[8] A. Bhattacharyya, T. Chakraborty, and S.N. Rai, “Stochastic forecasting of COVID-19 daily new cases across countries with a novel hybrid time series model,” *Nonlinear Dyn* , vol. 107, no. 3, pp. 3025–3040, Feb. 2022, doi: 10.1007/s11071-021-07099-3.

[9] T. McAndrew *et al.* , “Chimeric forecasting: combining probabilistic predictions from computational models and human judgment,” *BMC Infect Dis* , vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12879-022-07794-5.

[10] M. Arashi, M. Roozbeh, N.A. Hamzah, and M. Gasparini, “Ridge regression and its applications in genetic studies,” *PLoS One* , vol. 16, no. April 4, Apr. 2021, doi: 10.1371/journal.pone.0245376.

[11] J.T. Lim, E.L.W. Choo, A. Janhavi, K.B. Tan, J. Abisheganaden, and B. Dickens, “Density prediction of conjunctivitis burden using high-dimensional environmental time series data,” *Epidemics* , vol. 44, Sept. 2023, doi: 10.1016/j.epidem.2023.100694.

[12] CB Aditya Satrio, W. Darmawan, BU Nadia, and N. Hanafiah, "Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET," in *Procedia Computer Science* , Elsevier BV, 2021, pp. 524–532. doi: 10.1016/j.procs.2021.01.036.

[13] MHDM Ribeiro, RG da Silva, VC Mariani, and L. dos S. Coelho, “Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil,” *Chaos Solitons Fractals* , vol. 135, Jun. 2020, doi: 10.1016/j.chaos.2020.109853.

[14] IN Tanawi, V. Vito, D. Sarwinda, H. Tasman, and GF Hertono, "Support Vector Regression for Predicting the Number of Dengue Incidents in DKI Jakarta," in *Procedia Computer Science* , Elsevier BV, 2021, pp. 747–753. doi: 10.1016/j.procs.2021.01.063.

[15] A. Dina, “WEB-Based Dengue Fever Outbreak Forecasting Model Using Multiple Linear Regression Method (Case Study of South Tangerang City Health Service),” Bachelor's Thesis, PLN Institute of Technology, 2023.