DETERMINING THE CORRELATION BETWEEN WEATHER AND DENGUE FEVER (DBD) RATES IN JAKARTA

Fundamentals of Data Science Final Project

Lecturer: D6211 - Nunung Nurul Qomariyah, S.Kom., M.T.I., Ph.D

By:
Tiffany Widjaja - 2802503791
Allenxavinzky Adjiewibowo - 2802467880
Farrell Raffelino Sunarman - 2802503476

Computer Science Major
Class : L3AC

December 2025
BINUS INTERNATIONAL UNIVERSITY SENAYAN
BATCH 2028

Table of Contents

Chapter 1 - Problem Analysis

Dengue fever is not unheard of in Indonesia, especially in a humid city like Jakarta where every month there will always be a case of Dengue [3]. According to an UNAIR news Article, there is a correlation between weather patterns and the increase of Dengue fever rates [1], and thus using that information, we want to see if there really is a correlation between weather and Dengue fever rates by making a predictive model using reputable data from Weather Spark, the Ministry of Health (Kemenkes), and Open Dengue which combined consists of the Jakartan data for average temperature, average Humidity, average rainfall, and monthly as well as yearly Dengue fever rates in Jakarta.

From those data we could determine whether there is any correlation between weather and Dengue fever rates, which would make our problem statement “Effect of rain and humidity on dengue (DBD) rates in Jakarta”. However, it should be known that the effect of rain and humidity on Dengue rates is a very relevant problem for Indonesians, especially Jakartans, as while Dengue fever are sometimes mild, they can also prove fatal in the worse case scenario, which is why we think that Jakartans should care about how rain and humidity effects dengue fever considering that Jakarta is a humid city with frequent rains during rainy season [4].

Chapter 2 - Related Works

A number of studies and research have been made and conducted about the number of dengue fever cases and its possibility of direct correlation to the weather here in Indonesia. Often intended to help with dengue fever cases around Indonesia by determining how climate factors, such as temperature, rainfall frequency, and humidity affect the number of dengue fever cases. Other methods like the impact of climate change on mosquito growth is also conducted due to the belief that its development correlates to the rise in dengue cases [14]. It corresponds to helping in identifying whether the changes in weather patterns contribute to the number of cases in Indonesia.

The study focuses on primarily the Jakarta region of Indonesia, where there were quite a number of cases that are regularly occurring. The areas were chosen based on the reliability of health, and data related to the weather and its changes. To ensure that variations and long-term patterns in data can be assessed properly, the studies were conducted in spans of years and analyzed on a weekly and some monthly basis [5][6]. The result that it is able to produce varies due to the different methods and research goals that other related works used.

Looking through the method that similar studies use, while they all differ, the goal and question is still quite similar. Some use data from hospitals and conduct their research through patients. From IJID Regions, the research was conducted on patients that are admitted with 3 or more days of fever without any proven cause of bacterial infection or autoimmune disease [7][13]. Research by Telkom University and Poltekkes uses a broader number in population patterns and correlates them with the factors of the environment, such as temperature, humidity, and rainfall [8][9]. Ecological Complexity uses a similar method, but uses incidence data from BMKG and uses it through a regression model to predict dengue fever rates [10][14].

While most research uses fever cases and population numbers, some research about dengue fever focuses more on the weather factor. Research from KnE Life Sciences Journal not only found normal factors such as rainfall, humidity, and temperature, but they also found solar radiation could have a key part in the fever [5]. Other notable research by PLOS and MDPI, have great data that both use graphs to visualize their data to easily compare between the dengue fever and weather patterns [11][12].

Though the related works have differ data that it produces, it all leads back to the correlation between the weather and the dengue fever cases. How the data differs due to the methods has become quite unreliable and possibly outdated. This project will further expand from those initial ideas used in their methods and improve it to have much more linear data, though limitations are acknowledged when interpreting results.

Chapter 3 - Methodology

A. Data Research and Cleaning

The first step in our project would be to find datasets that will be used in our research and cleaning the data for better reading. We found various datasets that we could use but the dataset we chose to train our data model comes from a study by Universitas Islam Negeri Syarif Hidayatullah, which gave us a great amount of data on our topic of dengue fever, such as its rates per province and the monthly national rate, to allow us to get a good approximation for the monthly dengue fever rates in Jakarta [15].

B. Data Model and Techniques

In our project we tested 7 models for our machine learning, each of them having different yet similar uses that helps us determine the factors and conclusion to our research.

1. Multiple Linear Regression

As the name suggests, this data model is used between more than input variables on a continuous output via a straight line. It works best when relationships between input and output are simple and linear, as close as possible to the actual data points [18]. In our research, it helps to estimate the number of dengue fever cases from weather factors. It assumes using:

$$\hat{y} = \beta_o + \beta_1 x_1 + \cdots + \beta_n x_n$$

Where:

- y^ is the predictive value
- Xn are the independent variables
- βn are the coefficients corresponding to each predictor
  [18]

2. **Polynomial Regression**

Similar to Linear Regression, but it is able to capture non-linear data due to its polynomial transformations feature. It allows the Linear Regression model to find coefficients on a curved relationship, instead of a straight linear line. This type of model is mainly useful when data shows curved trends [20]. In our type of research, it helps to answer statements like, “The Dengue Fever case increases when temperature rises from 20°C -> 32°C, but decreases when temperatures exceed over 35°C, which then kills the mosquitoes”. Using:

$$\hat{y} = \beta_o + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n$$

Where:

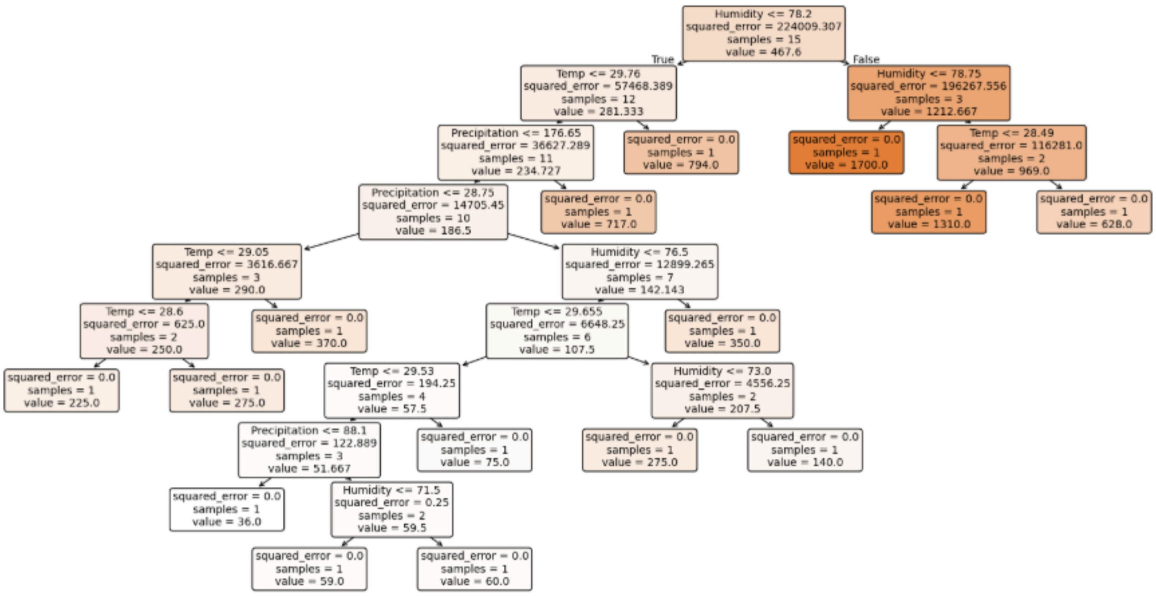
- y^ is the predictive value
- X^n are the input features
- βn are the coefficients the model learns

  [20]

3. **Decision Tree**

When data is inserted into a decision tree, it splits into increasing numbers of groups forming a branch-like look through a series of yes or no questions on each of its features [21]. A decision tree helps to make decisions by mapping out various choices and each of their possible outcomes, which is why it is used in tasks like classification and prediction [21]. In our research, it helps in finding thresholds in weather variables that could be affecting the number of dengue cases and explaining the complicated interactions between weather factors. Examples being, if humidity > 80%, dengue cases spike.

Our data model visualizes the following in our research:



4. Random Forest

It's an algorithm that uses many decision trees to make better predictions [22]. It builds trees that each look at random parts of data, results, and a random subset of the features (columns) combined, giving each tree their own answer or prediction based on what it has learned from its part of the data[22]. It is very good at capturing complex non-linear relationships and measuring variable importance [22]. In our research, this data model tells us which weather factors affect the dengue fever cases the most as it is able to handle noisy real-world weather data well.

5. XGBoost

Extreme Gradient Boosting builds trees in sequences, where every new tree made tries to fix errors that previous trees have [23]. The key idea of this model is in its name "Gradient Boosting", where it starts with a simple prediction (a tree), computes the residuals, and grows a small decision tree to predict the residuals, adding it into the model and repeats the same sequences many times until a stopping criterion is met [23]. In our research, this helps with managing missing niche weather data that we don't have and capturing delayed weather effects. The model can be represented mathematically as:

$$\hat{y}_n = \sum_{k=1}^{K} f_k\ (x_n)$$

Where:

- yn is the final predicted value for the n-th data point
- K is the number of trees in the ensemble
- fk(xn) represents the prediction of the K-th tree for the n-th data point
  [23]

6. **Poisson Regression**

A statistical technique used to model and analyze count data, where the outcome variable represents the number of times an event occurs in a fixed interval of time, space, or any other dimension [24]. It is useful when prediction is needed in forms of counts, like “how much” or “how many” [24]. In our research, this model is used to estimate the effect of each weather variable on the dengue fever cases. Mathematically, Poisson Regression can be formulated as:

The output of Y is assumed to follow a predicted count

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

Where:

- P is the count variable
- y is a particular count
- λ is the expected rate of occurrence
- e is Euler’s number (approximately 2.718)
  [24]

Instead of modeling Y directly, models the log of the expected value:

$$\log(\lambda) = \beta_o + \beta_1 x_1 + \cdots + \beta_n x_n$$

Where:

- λ is the expected count
- Xn are independent variables
- βn are the coefficients to be learned

7. Negative Binomial Regression

Negative binomial regression is a generalization of Poisson regression that is specifically useful for over-dispersed data [25]. It is useful when Poisson Regression is unable to be used well due to the extra noise that the data includes, and even though it is slightly more complex, it is much more flexible [25]. In our type of research, it helps to study and find long-term dengue fever and weather correlations, where data usually has overdispersion where in many weeks with a few cases or a few weeks with larger outbreaks.

C. Evaluation Method

1. Evaluation Technique

In our project, we would be determining the metrics that will allow us to evaluate how well our predictive model is performing, which would help us to determine which machine learning models are worth using in our research. We decided on 2 factors, which includes:

- **R-Squared / Coefficient of Determinant.**

Coefficient of Determinant measures the quality of a regression machine learning model. This method evaluates how well a model fits its data. How the coefficient of Determinant evaluates the model is by accessing the proportion of variance in the dependent variable that is defined by the independent variable. Its metric ranges from 0 to 1, where 1 indicates that the model had perfectly fit the data, while 0 indicates that the model didn’t explain any of the variability present in the dependent variable [29].

- **Root Mean Square Error (RMSE) / Mean Absolute Error (MAE).**

Root Mean Square Error (RMSE) measures the average difference between values predicted by a model and the actual values [26]. It provides an estimation of how well the model is able to predict the target value [26]. It has the advantage of representing the amount of error in the same unit as the predicted column making it easy to interpret [26].

The lower the value of the Root Mean Squared Error, the better the model is. A perfect model would have a Root Mean Squared Error value of 0 [26].

Mean Absolute Error (MAE) is used to evaluate the accuracy of regression models [27]. It measures the average absolute difference between the predicted values and the actual target values [27]. Mean Absolute Error doesn't square the errors, which means it gives equal weight to all errors, regardless of their direction [27]. This makes it useful when you want to understand the magnitude of errors without considering whether they are overestimations or underestimations [27].

- **Poisson Deviance**

There are two deviances in Poisson Deviance:

The Null Deviance that shows how well the response variable is predicted by the model that includes only the intercept (the grand mean) [28].

The Residual Deviance is -2 times the difference between the log likelihood evaluated at the maximum likelihood estimate (MLE) and the log likelihood for a "saturated model" [28].

2. Splitting Dataset

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 1)
```

```
X_train
```

```
array([[ 29.2   ,  71.9  ,   5.    ],
       [ 29.61  ,  74.   ,  63.3   ],
       [ 28.3   ,  68.6  ,   0.    ],
       [ 29.34  ,  70.   ,  77.9   ],
       [ 29.7   ,  71.   ,  80.    ],
       [ 29.8   ,  73.8  ,  24.7   ],
       [ 29.72  ,  75.   ,  52.5   ],
       [ 29.1   ,  78.4  , 132.5   ],
       [ 29.15  ,  73.   ,  98.3   ],
       [ 29.45  ,  70.   , 131.9   ],
       [ 28.28  ,  82.   , 607.2   ],
       [ 28.7   ,  79.1  , 509.3   ],
       [ 29.48  ,  78.   , 142.2   ],
```

```
       [ 28.75  ,  78.    , 211.1  ],
       [ 28.9   ,  68.7   ,   0.    ]])
```

```
y_train
```

```
array([ 370,   75,  225,   36,  275,  794,  140, 1700,   60,   59, 1310,
        628,  350,  717,  275])
```

```
X_test
```

```
array([[ 28.7 ,  68.5 ,   0.   ],
       [ 29.06,  71.  ,  99.9  ],
       [ 29.5 ,  69.  ,   1.   ],
       [ 27.92,  83.  , 784.5 ]])
```

```
y_test
```

```
array([ 244,   50,  351, 1028])
```

Chapter 4 - Results and Analysis

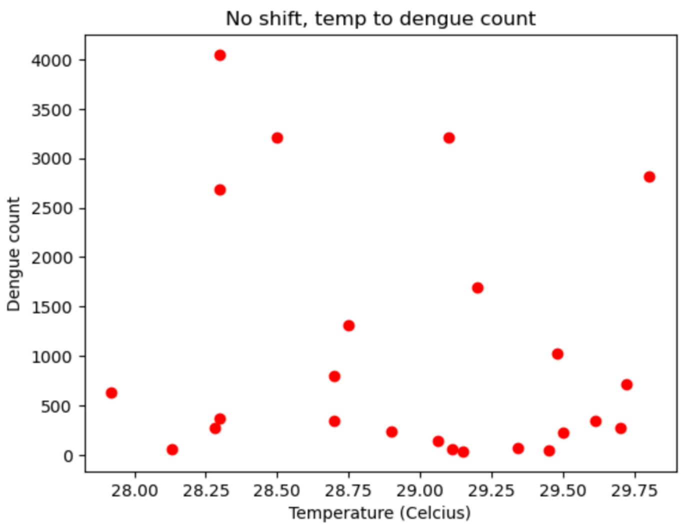
A. Correlation Testing Methodology

The way we tested for correlation of the resulting graphs is to calculate the Pearson Correlation Coefficient (PCC) where the values range from -1 to 1, where -1 shows a perfect negative correlation, 1 shows a perfect positive correlation, and 0 shows no correlation at all [16]. By using PCC, we can determine whether the resulting graph shows a strong, weak, or no correlation. Along with PCC, we also pair it with calculating the Probability Value of the correlation being real, with a significance (α) of 0.05, where if the Probability Value is above 0.05, that means that the correlation is not real while if the Probability Value is below 0.05, that means that the correlation is real [17].

B. No Shift

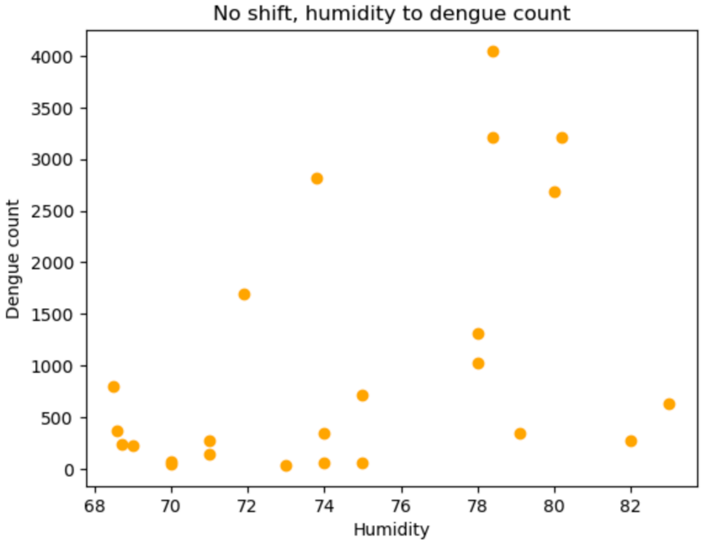
During the testing of the Machine Learning, we used data we gathered on Dengue Fever rates per month in Jakarta and comparing it to Jakartan data for Temperature, Humidity, and Precipitation to see if there was a correlation.

```
plt.figure()
plt.scatter(dataset_no_shift["Temp"], dataset_no_shift["Dengue_count"], c="red")
plt.xlabel("Temperature (Celcius)")
plt.ylabel("Dengue count")
plt.title("No shift, temp to dengue count")
plt.show()
```



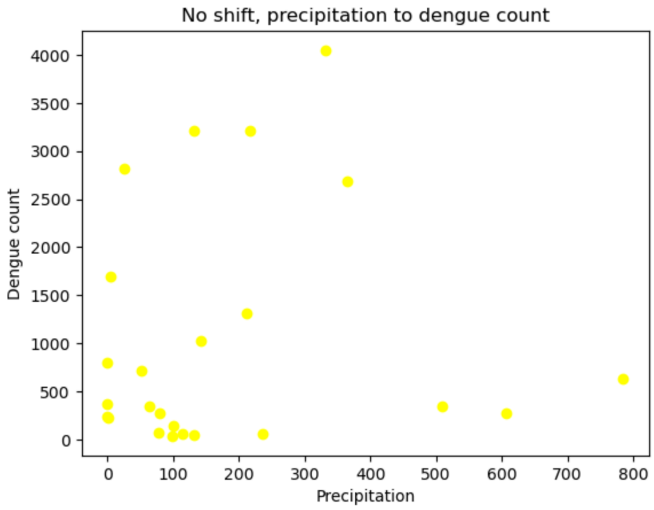
For the temperature, we didn't detect any correlation due to the PCC value being -0.167 which is close to zero, while the Probability Value was above 0.05, which meant that the correlation was not real.

```
plt.figure()
plt.scatter(dataset_no_shift["Humidity"], dataset_no_shift["Dengue_count"], c="orange")
plt.xlabel("Humidity")
plt.ylabel("Dengue count")
plt.title("No shift, humidity to dengue count")
plt.show()
```



For the humidity, we detected a weak positive correlation with the PCC value being -0.451, and with the Probability Value being below 0.05, we can conclude that for humidity, the correlation is real, albeit weak.

```
plt.figure()
plt.scatter(dataset_no_shift["Precipitation"], dataset_no_shift["Dengue_count"], c="yellow")
plt.xlabel("Precipitation")
plt.ylabel("Dengue count")
plt.title("No shift, precipitation to dengue count")
plt.show()
```

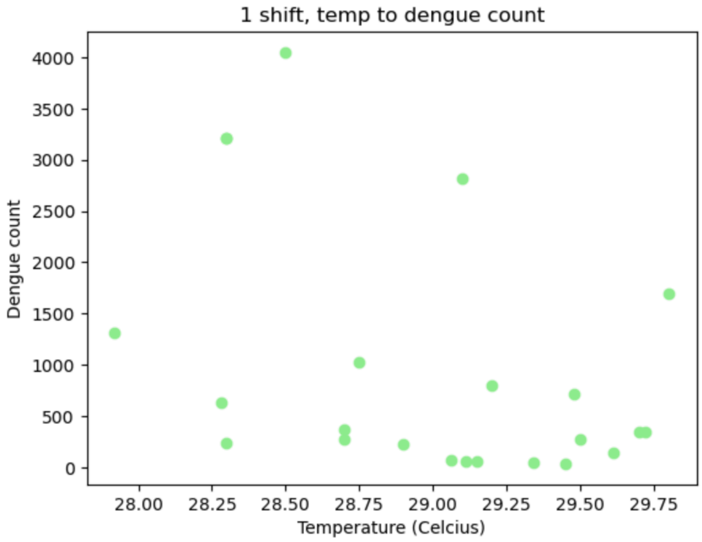


For the precipitation, we didn’t detect any correlation due to the PCC value being -0.088 which is close to zero, while the Probability Value was above 0.05, which meant that we can conclude that there is no correlation for precipitation as of now.

C. 1 Month Shift

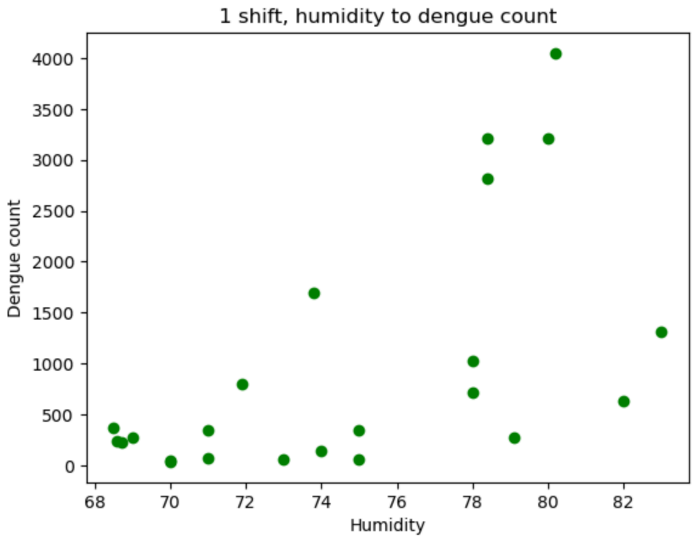
To accommodate for the life cycle of a Mosquito as well as the Virus incubation period for it to become transmissible, we tried shifting the months by 1 to see if there were any difference, and the result showed that there were improving correlations.

```
plt.figure()
plt.scatter(dataset_1_shift["Temp"], dataset_1_shift["Dengue_count"], c="lightgreen")
plt.xlabel("Temperature (Celcius)")
plt.ylabel("Dengue count")
plt.title("1 shift, temp to dengue count")
plt.show()
```



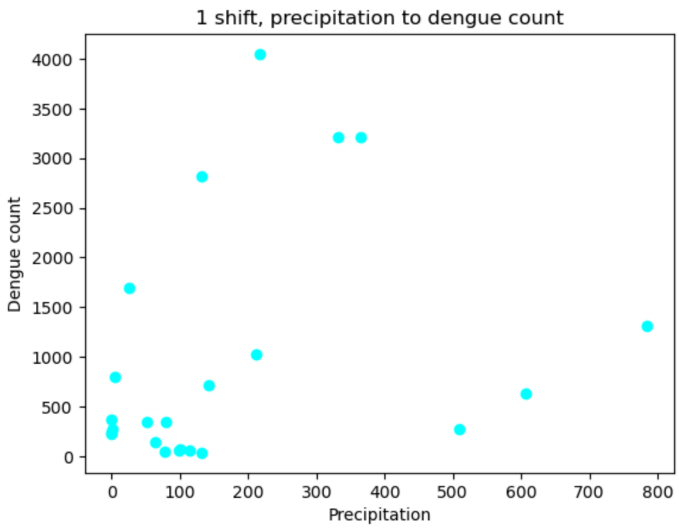
For the temperature, after we shifted by 1 month, we detected a weak negative correlation with the PCC value being -0.414, and with a Probability Value that is below 0.05, which meant that the correlation is real, though only barely.

```
plt.figure()
plt.scatter(dataset_1_shift["Humidity"], dataset_1_shift["Dengue_count"], c="green")
plt.xlabel("Humidity")
plt.ylabel("Dengue count")
plt.title("1 shift, humidity to dengue count")
plt.show()
```



For the humidity, after we shifted by 1 month, we detected a strong positive correlation with the PCC value being 0.592, and with a Probability Value that is below 0.01, far below the maximum of 0.05, which means that the correlation is very real and strong.

```
plt.figure()
plt.scatter(dataset_1_shift["Precipitation"], dataset_1_shift["Dengue_count"], c="cyan")
plt.xlabel("Precipitation")
plt.ylabel("Dengue count")
plt.title("1 shift, precipitation to dengue count")
plt.show()
```

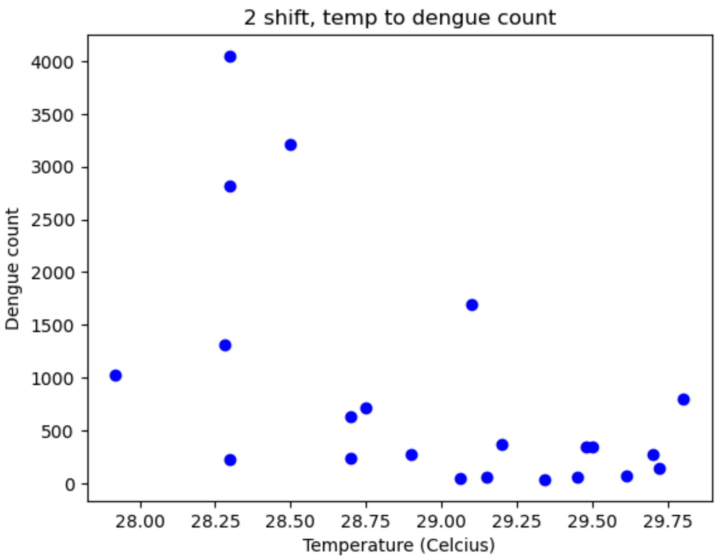


For the precipitation, after we shifted by 1 month, we detected a weak positive correlation with the PCC value being 0.291, and with a Probability Value that is below 0.05, which meant that the correlation is real, though not very strong.

D. 2 Months Shift

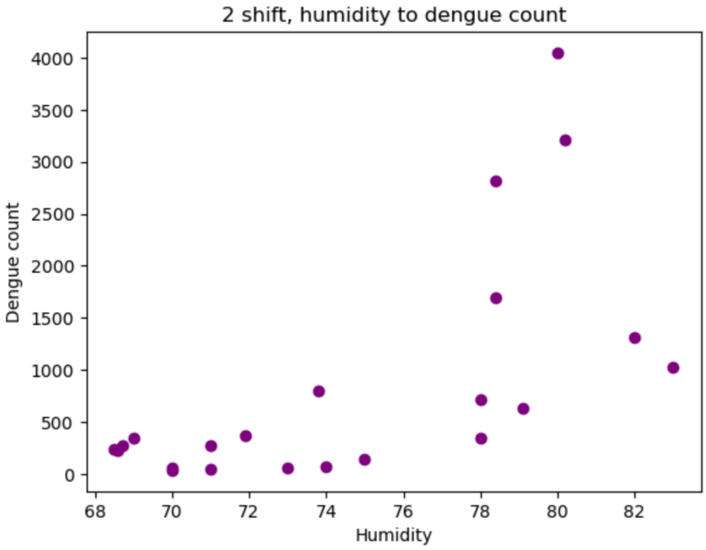
We decided to shift the month by 2 to see if the correlation improves or decreases as the previous correlation was lackluster apart from humidity, which is still below the standard to be used for Machine Learning.

```
plt.figure()
plt.scatter(dataset_2_shift["Temp"], dataset_2_shift["Dengue_count"], c="blue")
plt.xlabel("Temperature (Celcius)")
plt.ylabel("Dengue count")
plt.title("2 shift, temp to dengue count")
plt.show()
```



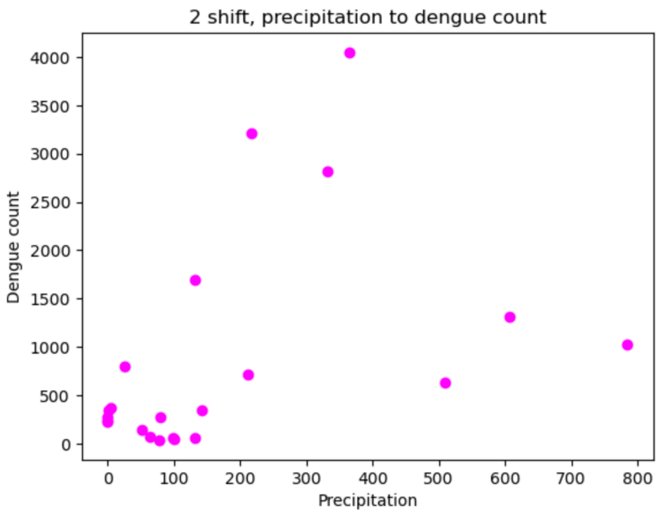
For the temperature, after we shifted by 2 months, we detected a strong negative correlation with the PCC value being -0.552, and with a Probability Value that is below 0.01, we can conclude that the correlation is very real and strong.

```
plt.figure()
plt.scatter(dataset_2_shift["Humidity"], dataset_2_shift["Dengue_count"], c="purple")
plt.xlabel("Humidity")
plt.ylabel("Dengue count")
plt.title("2 shift, humidity to dengue count")
plt.show()
```



For the humidity, after we shifted by 2 months, we detected a strong positive correlation with the PCC value being 0.638, while the Probability Value that is below 0.05, it is also nearing 0.01, which means we can conclude that the correlation is very real and strong.

```
plt.figure()
plt.scatter(dataset_2_shift["Precipitation"], dataset_2_shift["Dengue_count"], c="magenta")
plt.xlabel("Precipitation")
plt.ylabel("Dengue count")
plt.title("2 shift, precipitation to dengue count")
plt.show()
```



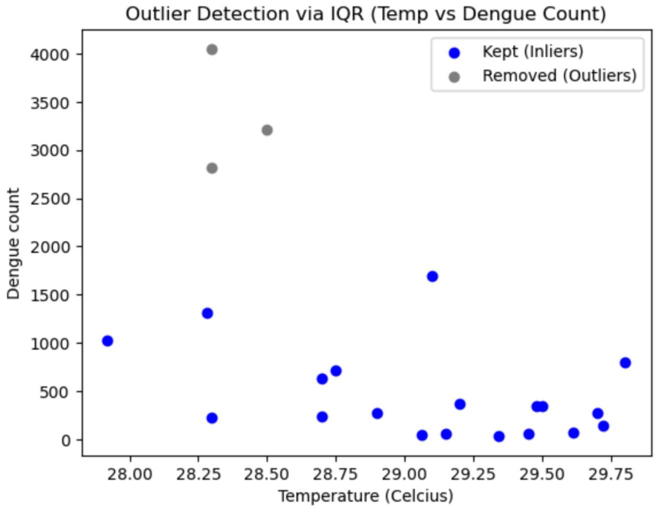
For the Precipitation, after we shifted by 2 months, we detected a positive correlation, albeit not too strong, with the PCC value being 0.0430, with a Probability Value being below 0.05, which means that the correlation is real.

E. Outlier Removal

To improve our graph for Machine Learning, we decided to remove outliers that are present in the data for the Monthly dengue count when compared to the data for humidity, temperature, and precipitation, the result was a more defining graph ready to be used for Machine Learning. The Removal Process are as follows:

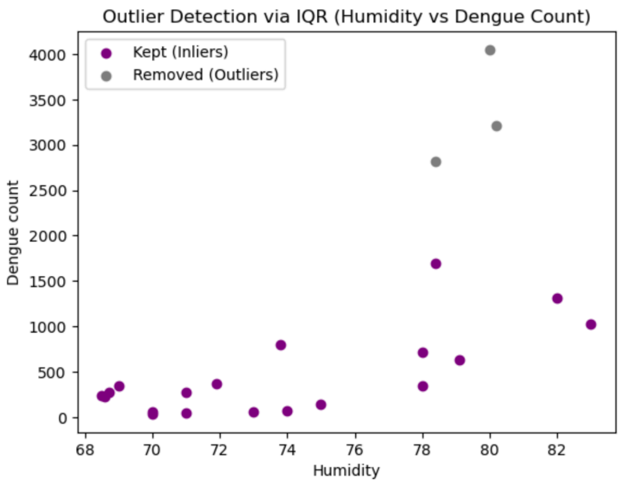
```
# values kept (inliers)
inliers = dataset_2_shift[
    (dataset_2_shift["Dengue_count"] >= lower_dengue) &
    (dataset_2_shift["Dengue_count"] <= upper_dengue)
#    (dataset_2_shift["Temp"] >= lower_temp) &
#    (dataset_2_shift["Temp"] <= upper_temp) &
#    (dataset_2_shift["Humidity"] >= lower_humidity) &
#    (dataset_2_shift["Humidity"] <= upper_humidity) &
#    (dataset_2_shift["Precipitation"] >= lower_precipitation) &
#    (dataset_2_shift["Precipitation"] <= upper_precipitation) &
#    (dataset_2_shift["Dengue_count"] != forced_outlier_value)
]

# values removed (outliers)
outliers = dataset_2_shift[
    (dataset_2_shift["Dengue_count"] < lower_dengue) |
    (dataset_2_shift["Dengue_count"] > upper_dengue)
#    (dataset_2_shift["Temp"] < lower_temp) |
#    (dataset_2_shift["Temp"] > upper_temp) |
#    (dataset_2_shift["Humidity"] < lower_humidity) |
#    (dataset_2_shift["Humidity"] > upper_humidity) |
#    (dataset_2_shift["Precipitation"] < lower_precipitation) |
#    (dataset_2_shift["Precipitation"] > upper_precipitation) |
#   (dataset_2_shift["Dengue_count"] == forced_outlier_value)
]
```

```
plt.figure()
plt.scatter(inliers["Temp"], inliers["Dengue_count"], c="blue", label="Kept (Inliers)")
plt.scatter(outliers["Temp"], outliers["Dengue_count"], c="grey", label="Removed (Outliers)")
plt.xlabel("Temperature (Celcius)")
plt.ylabel("Dengue count")
plt.title("Outlier Detection via IQR (Temp vs Dengue Count)")
plt.legend()
plt.show()
```



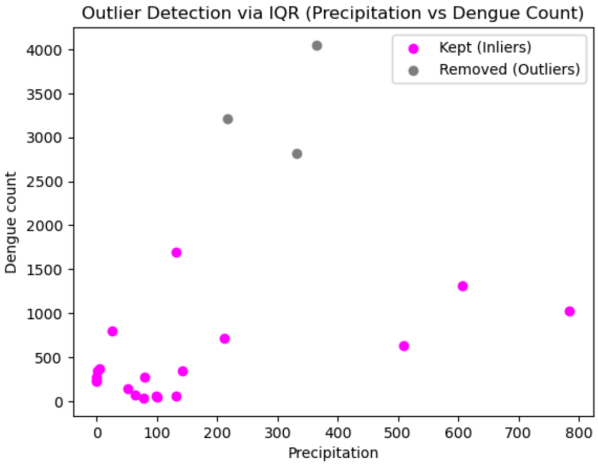
For the precipitation based on the one that was shifted by 2 months, after we removed 3 outliers, we detected a negative correlation, albeit not too strong, with the PCC value being -0.406, though with a Probability Value being above 0.05, which means that the correlation is not real.

```
plt.figure()
plt.scatter(inliers["Humidity"], inliers["Dengue_count"], c="purple", label="Kept (Inliers)")
plt.scatter(outliers["Humidity"], outliers["Dengue_count"], c="grey", label="Removed (Outliers)")
plt.xlabel("Humidity")
plt.ylabel("Dengue count")
plt.title("Outlier Detection via IQR (Humidity vs Dengue Count)")
plt.legend()
plt.show()
```



For the humidity based on the one that was shifted by 2 months, after we removed 3 outliers, we detected a strong positive correlation, with the PCC value being 0.718, with a Probability Value below 0.01, it is safe to say that the correlation is very real and strong.

```
plt.figure()
plt.scatter(inliers["Precipitation"], inliers["Dengue_count"], c="magenta", label="Kept (Inliers)")
plt.scatter(outliers["Precipitation"], outliers["Dengue_count"], c="grey", label="Removed (Outliers)")
plt.xlabel("Precipitation")
plt.ylabel("Dengue count")
plt.title("Outlier Detection via IQR (Precipitation vs Dengue Count)")
plt.legend()
plt.show()
```



For the precipitation based on the one that was shifted by 2 months, after we removed 3 outliers, we detected a somewhat strong positive correlation, with the PCC value being 0.556, with a Probability Value below 0.05 nearing 0.01, it is safe to say that the correlation is very real and strong.

F. Machine Learning Evaluation

Model	R^2	RMSE	MAE	Poisson deviance	Notes
Linear Regression	0.62	225.9	191	N/A	Okay
Polynomial Regression	-12.9	1373.4	813	N/A	Not fitting at all
Decision Tree	0.85	142.2	86	18	Best Poisson deviance
Random Forest	0.92	103.4	95	44	Best overall
XGBoost	0.85	143.5	93	22	Good
Poisson Regression	0.80	152.7	152	131	Okay, but better than linear reg.

1. Negative Binomial Regression

Negative Binomial Regression was the 7th Machine Learning that we tried testing, we decided to abandon using this model because our dataset was too small for this model to use, which is why no test results appeared on the table above.

Chapter 5 - Conclusion and Future Work

A. Dengue to Weather Correlation

Our main hypothesis was that the increase or decrease in dengue fever cases in a month is based on environmental factors, we pinpointed 3, which were: temperature, humidity, precipitation. After comparing the data between Dengue and weather related factors like humidity, temperature, and precipitation, it was proven that there a correlation, with the increase in humidity and precipitation, there would be a higher chance of cases of dengue fever, while the higher the temperature, the less likely it is for cases of dengue fever to occur. We can conclude that humidity, temperature, and precipitation does play a role in how many dengue cases there are in a month, and that it takes a perfect combination of the three environmental factors to have dengue fever cases peak in certain months, like the rainy season which generally happens between October and April [19].

B. Future Work

While the data that were compared showed a correlation, it was not optimal because the data being used only contained 2 years worth of data, instead of the general standard of 5 to 10 years of data, this was caused by data collection by the BMKG which isn't generally as complete as other nations with some years being missing, and sometimes the years that are available are not consistent as some data might start at 2015 and end in 2023, but other data might start at 2018, and end in 2025,

References

[1] unairnews, “Hubungan Faktor Iklim dengan Demam Berdarah Dengue - Universitas Airlangga Official Website,” *Universitas Airlangga Official Website*, Dec. 27, 2021. https://unair.ac.id/hubungan-faktor-iklim-dengan-demam-berdarah-dengue/ (accessed Oct. 12, 2025).

[2] Kemenkes, “Selamat Datang di Situs Resmi Kementerian Kesehatan Republik Indonesia,” Kemkes.go.id, 2024. https://kemkes.go.id/id/

[3] Open Dengue, “OpenDengue: Data,” *OpenDengue*. https://opendengue.org/data.html

[4] Weather Spark, “Iklim, Cuaca Menurut Bulan, Suhu Rata-Rata Jakarta (Indonesia) - Weather Spark,” *id.weatherspark.com*. https://id.weatherspark.com/y/116847/Cuaca-Rata-rata-pada-bulan-in-Jakarta-Indonesia-Sepanjang-Tahun

[5] Hasanah and D. Susanna, “Weather Implication for Dengue Fever in Jakarta, Indonesia 2008-2016,” *KnE Life Sciences*, vol. 4, no. 10, p. 184, Feb. 2019, doi: https://doi.org/10.18502/kls.v4i10.3719.

[6] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, and J. Rocklöv, “Forecast of Dengue Incidence Using Temperature and Rainfall,” *PLoS Neglected Tropical Diseases*, vol. 6, no. 11, p. e1908, Nov. 2012, doi: https://doi.org/10.1371/journal.pntd.0001908.

[7] S. Ali *et al.*, “Clinical and Laboratory Profiles of Dengue Infection in the Hospitals in North Jakarta, Indonesia,” *IJID Regions*, vol. 14, pp. 100612–100612, Feb. 2025, doi: https://doi.org/10.1016/j.ijregi.2025.100612.

[8] 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: https://doi.org/10.21111/fij.v4i2.3199.

[9] Rojali Rojali, Indah Restiaty, D. Lisa, and Muhammad Dimas Setyadi, “HUBUNGAN PERUBAHAN IKLIM DENGAN KEJADIAN DEMAM BERDARAH DENGUE (DBD) DI KOTA ADMINISTRASI JAKARTA TIMUR,” *Sulolipu*, vol. 23, no. 1, pp. 172–186, Jan. 2024, doi: https://doi.org/10.32382/sulo.v23i1.427.

[10] Muhammad Fakhruddin *et al.*, “Assessing the interplay between dengue incidence and weather in Jakarta via a clustering integrated multiple regression model,” *Ecological Complexity*, vol. 39, pp. 100768–100768, Jun. 2019, doi: https://doi.org/10.1016/j.ecocom.2019.100768.

[11] Mamenun, Yonny Koesmaryono, Ardhasena Sopaheluwakan, Rini Hidayati, Bambang Dwi Dasanto, and R. Aryati, “Spatiotemporal Characterization of Dengue Incidence and Its Correlation to Climate Parameters in Indonesia,” *Insects*, vol. 15, no. 5, pp. 366–366, May 2024, doi: https://doi.org/10.3390/insects15050366.

[12] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, and J. Rocklöv, “Forecast of Dengue Incidence Using Temperature and Rainfall,” *PLoS Neglected Tropical Diseases*, vol. 6, no. 11, p. e1908, Nov. 2012, doi: https://doi.org/10.1371/journal.pntd.0001908.

[13] S. Ali *et al.*, “Clinical and Laboratory Profiles of Dengue Infection in the Hospitals in North Jakarta, Indonesia,” *IJID Regions*, vol. 14, pp. 100612–100612, Feb. 2025, doi: https://doi.org/10.1016/j.ijregi.2025.100612.

[14] BMKG, “BMKG | Badan Meteorologi, Klimatologi, dan Geofisika,” *Bmkg.go.id*, 2020. https://www.bmkg.go.id/

[15] R. A. Permaisuri, “Skripsi Visualisasi Dashboard Tableau dan Peramalan Jumlah Kasus Demam Berdarah Dengue di DKI Jakarta Menggunakan Metode Arima.” *Universitas Islam Negeri*. https://repository.uinjkt.ac.id/dspace/bitstream/123456789/64985/1/RENDA%20AUDI%20PERMAISURI-FST.pdf

[16] Geeks for Geeks, “Pearson Correlation Coefficient,” *GeeksforGeeks*, Feb. 14, 2022. https://www.geeksforgeeks.org/maths/pearson-correlation-coefficient/

[17] Geeks for Geeks, “P-value in Machine Learning,” *GeeksforGeeks*, Jul. 09, 2020. https://www.geeksforgeeks.org/machine-learning/p-value-in-machine-learning/

[18] Geeks for Geeks, “Linear Regression in Machine Learning,” *GeeksforGeeks,* Nov. 22, 2025. https://www.geeksforgeeks.org/machine-learning/ml-linear-regression/

[19]Selective Asia, “Best time to visit Indonesia - weather by month - climate - seasons,” Selectiveasia.com, 2019. https://www.selectiveasia.com/indonesia-holidays/weather

[20] Geeks for Geeks, “Implementation of Polynomial Regression,” *GeeksforGeeks*, Jul. 11, 2025. https://www.geeksforgeeks.org/machine-learning/python-implementation-of-polynomial-regression/

[21] Geeks for Geeks, “Decision Tree,” *GeeksforGeeks*, Jun. 30, 2025.
https://www.geeksforgeeks.org/machine-learning/decision-tree/

[22] Geeks for Geeks, “Random Forest Algorithm in Machine Learning,” *GeeksforGeeks*, Oct. 31, 2025.
https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

[23] Geeks for Geeks, “XGBoost,” *GeeksforGeeks*, Oct. 24, 2025.
https://www.geeksforgeeks.org/machine-learning/xgboost/

[24] Geeks for Geeks, “Poisson Regression,” *GeeksforGeeks*, Jul. 23, 2025.
https://www.geeksforgeeks.org/machine-learning/poisson-regression/

[25]“Negative Binomial Regression | Stata Data Analysis Examples,” *stats.oarc.ucla.edu*.
https://stats.oarc.ucla.edu/stata/dae/negative-binomial-regression/

[26]“Root Mean Squared Error (RMSE),” *help.sap.com*.
https://help.sap.com/docs/SAP_PREDICTIVE_ANALYTICS/41d1a6d4e7574e32b815f1cc87c00f42/5e5198fd4afe4ae5b48fefe0d3161810.html

[27] M. W. Ahmed, “Understanding Mean Absolute Error (MAE) in Regression: A Practical Guide,” *Medium*, Aug. 24, 2023.
https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df

[28] E. Liu, “Deviance of Poisson Regression,” HAL open science, 2019. [Online].
https://hal.science/hal-02335439v1/document#:~:text=Page%202,β0%20will%20be

[29]E. Onose, “R Squared: Understanding the Coefficient of Determination,” *Arize AI*, Aug. 08, 2023.
https://arize.com/blog-course/r-squared-understanding-the-coefficient-of-determination/

Miscellaneous

Work Distribution

- Report : Allen & Farrell
- Machine Learning : Tiffany
- Data Research and Prep : Allen & Tiffany
- Similar Works Research : Farrell
- Machine Learning Testing : Tiffany

.

Link to Github: https://github.com/frost-drago/Jakarta-Dengue-DS/tree/main