Erdős Institute Data Science May-Summer 2024 Executive Summary

Climate-Based Forecasting of Dengue Epidemic Months: A Case Study of Bangladesh

Haridas Kumar Das and Abdullah Al Helal

Overview

Over the past two decades, there has been a notable rise in dengue cases worldwide, with significant impacts observed in countries like Brazil and Bangladesh. Numerous studies have demonstrated the correlation between climate factors—such as temperature and rainfall—and the transmission of dengue, Zika, chikungunya, and yellow fever.

Objective

In this project, we develop machine learning algorithms to analyze climate and epidemiological data in order to

- classify and
- forecast

dengue epidemic months, focusing on the analysis of Bangladesh.

Evaluation Methodology

Data Preprocessing

The Open-Mateo climate data, the UN population data and the Bangladesh dengue epidemiological data were preprocessed using the following steps:

- The daily climate data was converted into monthly data by taking the maximum.
- The yearly population data was repurposed as monthly data
- The number of cases from the daily epidemiological data was accumulated into monthly data.

Labeling

The number of cases was used to generate the epidemic label using the following formula:

- label = epidemic if (total number of cases per month / total population) * 10^6 >= 10, otherwise not epidemic.

Exploratory Data Analysis

- There are 18 features, all numeric and 1 label, categorical. There are 192 observations during the years 2008 to 2023.
- The Pearson, Spearman and Kendall’s tau correlation matrices show high correlation between temperature variables, between moisture variables and between precipitation

and rain, prompting to keep only 8 features for classification and forecasting. These sets of variables show similar distributions as well.

- Time series plots reveal epidemics showing up only after 2018. Some variables tend to show higher values during epidemic months.

Evaluation

The classification problem was evaluated quantitatively using the accuracy score of the machine learning methods on the test dataset. We have used three error metrics for model testing: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). Additionally, to calculate the prediction accuracy using True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), we calculated the accuracy using the following formula:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN).$$

Results

Classification

- Ten machine-learning methods were used. The dataset was split using stratified sampling into 70%-10%-20% training-validation-test samples.
- Accuracy scores varied from 82.05% to 100% across all methods.

Forecasting

- CNN was used to forecast the epidemic label for the year 2024 using the same split of 70%, 10%, and 20% training-validation-test samples.
- Ten machine-learning models were implemented on the CNN prediction data for future epidemic months predictions.

Modeling limitation

- One area for improvement in modeling lies in the potential for the model to overlook complex relationships within climate data, particularly evident in the CNN model, which can result in less accurate predictions during dengue months by machine learning algorithms.

Conclusion and Future Directions

Conclusion

- Classified and forecasted dengue epidemic months from climate data.

Future Directions

- Classification can be improved by using a bigger dataset.
- Future work will involve implementing sophisticated probabilistic time series forecasting algorithms.