Exploration and Analysis of Data for Machine Learning: Dengue Fever Case Prediction

Cheng Yan
Department of Computer Science
University of Nottingham
Nottingham, UK
psxcy8@nottingham.ac.uk

Xuanhong Luo
School of Computer Science
University of Nottingham
Nottingham, UK
psxxl19@nottingham.ac.uk

Guihua Zou
School of Computer Science
University of Nottingham
Nottingham, UK
psxgz6@nottingham.ac.uk

***Abstract*—Dengue fever is a globally frequent endemic disease; however, with the development of climate change, this change has also significantly increased, becoming a major problem for global public health. In this study, we use the overall data of San Juan and Iquitos, the prevalence of dengue fever in the local area is predicted by the prediction model, including climate variables and the historical incidence rate of dengue fever. In case to improve the accuracy of the model, we choose a variety of data preprocessing techniques, such as handling missing values, normalization and feature engineering. At first, through the EDA we found a significant relationship between climate factors and dengue cases, which illustrated the impact of temperature and precipitation on dengue fever disease transmission. Then at the stage of constructing prediction models is carried out using advanced machine learning algorithms that are best suited to solve the problem of time series prediction random forests and gradient boosting machines. In total, the performance of multi machine learning models was compared on the current data set. These methods are selected in order to work with non-linear relationships and to obtain error tolerance against overfitting. The methodology correction of data anomalies, scaling of temperature indicators, and accounting for lag climate variables for the temporal characteristic of dengue transmission. In addition to model selection, performance is assessed using appropriate metrics to ensure both robustness and reliability.**

***Index Terms*—Machine Learning, Dengue Fever Prediction, Public Health, San Juan and Iquitos, Random Forest**

I. Data sets and research questions

In this study, we intend to use machine learning techniques to predict the number of dengue fever. Dengue fever is an infectious disease that is widely spread by Aedes mosquitoes in tropical and subtropical regions all over the globe(Sarma et al., 2020)[7]. Furthermore, the fluctuation in the number of dengue fever may be seasonal factors, including temperature, rainfall, humidity, among others. Understanding how they fluctuate the occurrence of dengue fever is also essential as far as public health is concerned. This will enable the concerned sector to take the necessary regulations and ensure dengue fever is prevented before it becomes a menace.

This study utilizes dengue weekly cases from the two cities and years and other environmental and climate factors. These data include but are not limited to city names, specific dates, weekly average temperatures, total rainfall, and relative humidity. Through conducting exploratory and descriptive data analysis to understand data distribution and whether there is any correlation between seasonal factors and dengue fever.

Research question:

1) What are the climate factors that affect the number of dengue fever cases significantly?

2) What is the most effective machine learning model to predict the number of dengue fever cases?

3) Can we use the prediction model predict dengue fever outbreaks accurately(low prediction error)?

Through the above investigation and analysis, this study designs and develops a good disease prediction model to predict the number of dengue fever cases accurately. Public health and early prediction of disease outbreaks are both crucial. This research helps determine a scientific basis for disease control and prevention, assists epidemic prevention departments in preparing for dengue prevention and control in advance, and reduces personnel and economic losses. Our focus is how to reduce our reaction time and prevention efficiency in the face of major public health matters, such as dengue fever.

II. Literature Review

Recent literature search: A review of recent research to gain deeper insight into the machine models used for predicting dengue fever cases through a search of recent literature. The results show that most current research progress has applied different machine learning models that pay attention to the application of appropriate data processing and feature engineering.

For example, Guo et al.[3] pointed out in a key study: Support Vector Machines will manage the complexity of the nonlinear dataset for predicting dengue fever in China. This suggests that SVM is a practical model for creating more complex epidemiological data.This also provides guidance for us to use SVM models for prediction and comparison of model performance and encourages us to adopt the SVM model of prediction to compare the model performance.

Salim et al.[7] and Sarma et al.[9] explored decision trees, random forests, and neural networks. Their findings emphasize that random forests perform exceptionally well in datasets rich in environmental and climate variables, demonstrating the advantage of the model in capturing complex relationships in the data. Through this literature, we have added decision trees, random forests, and neural networks to the prediction model for comparison.

Deb et al.[2]'s research further enriched this field by introducing integrated prediction methods. Their method combines multiple time series and regression models to improve the accuracy of weekly dengue fever case prediction based on climate and terrain conditions. The integration method adopted by Deb et al.[2] has been proven to significantly exceed traditional methods, achieving a lower average absolute error rate, and demonstrating a robust model for practical application in public health. This provides guidance for us to consider incorporating time factors in the special engineering stage, such as adding lag features and moving average features of key climate factors.

Besides, the systematic literature valuation by Hoyos et al.hoyos2021dengue. synthesized the machine learning activity for the benchmarking to illustrate the essential incorporation of pre-process, feature selecting, and evaluation. This study is essential as it showed the crucial evaluation to guarantee the model application in the real-world environment.

In conclusion, these papers show the potential of machine learning in enhancing predictive models in public health, more specifically in forecasting dengue fever cases. Moreover, the model not only applies the importance of meteorological factors, but it could also have a pronounced effect in improving public health planning and interventions.

III. Methodology

When training and predicting datasets with machine learning models, the accuracy of the prediction results of models depends on the data analysis results, data preprocessing cleanliness, feature engineering effects, and model selection. For more information, see the key steps outlined in Figure 1.

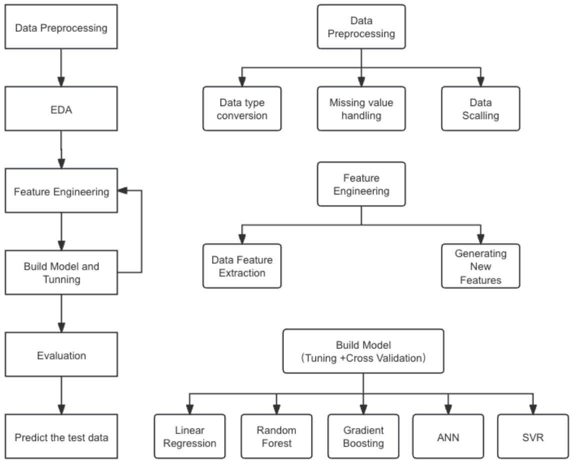


Fig. 1. Workflow

A. *Data Collection*

The dataset used in this analysis combines a large number of historical records of climate variables and health data delineating the incidence of dengue fever. Given the differing geographic distributions of the two locations and their climates, the dataset makes a significant contribution to the understanding of environmental factors affecting dengue fever spread.

B. *Data Processing*

The data processing stage was meticulously designed to enhance data quality and ensure compatibility with analytical methods.

- **Data Type Conversion:** Initial processing involved converting data into appropriate formats necessary for subsequent analyses, ensuring that all numerical and categorical data were accurately represented.
- **Missing Value Handling:** To address gaps in the dataset, missing values were imputed using median values, chosen for their robustness against data skewness and outliers, thereby preserving the underlying data distribution.
- **Data Scaling:** Numerical features were scaled to a uniform range to prevent any single variable from disproportionately influencing the model's performance. This scaling was essential for maintaining a balanced influence among all features in the predictive models.

C. *Feature Engineering*

Feature engineering aims to extract meaningful information from the dataset and generate new features to improve model performance.

- **Data Feature Extraction:** Critical date-related features were extracted, such as year, month, and day to capture the temporal patterns associated with dengue incidences, which are particularly pronounced during specific times of the year.
- **Generating New Features:** New features were created by calculating moving averages and lagged features for key climatic variables, such as temperature and humidity, to account for their delayed effects on the incidence of dengue fever. This approach allowed the models to incorporate historical context into their predictions.

D. *Model Development and Evaluation*

Multiple machine learning models were developed and evaluated to determine the most effective approach for predicting dengue fever outbreaks.

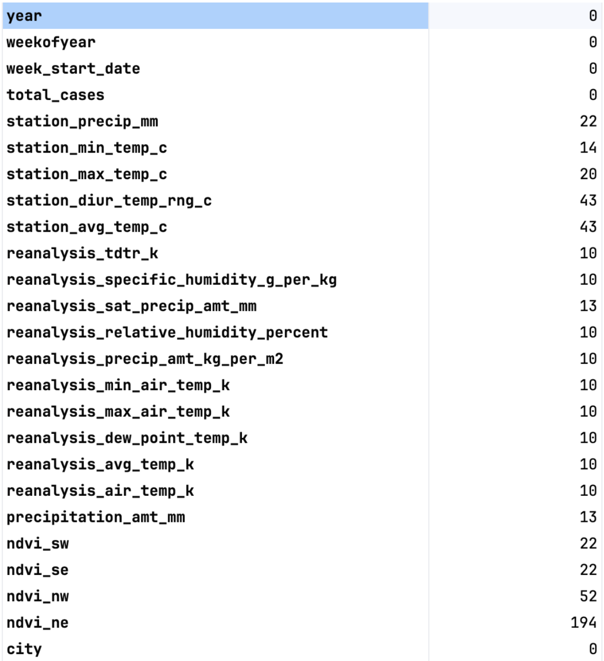
- **Linear Regression:** Chosen for its straightforward approach and ease of understanding, Linear Regression offers a clear standard against which we can measure the performance of more complex models.
- **Random Forest:** Chosen for its ability to handle large datasets with numerous features and for its robustness to overfitting.
- **Gradient Boosting:** Employed for its predictive power, particularly in scenarios where the relationship between predictors and outcome is complex.

- **Artificial Neural Network (ANN):** Used to model non-linear relationships deeply through its layered structure and ability to learn from large amounts of data.
- **Support Vector Regression (SVR):** Applied due to its effectiveness in high-dimensional spaces and its capability to model non-linear relationships through kernel functions.

All our models were subjected to rigorous testing using cross-validation techniques to guarantee that our findings were strong and replaceable. Mainly, our models' performance was based on which model had the least Mean Absolute Error and which other model had the greater accuracy. In the end, our models were subjected to another test set to ensure their reliability on our predictions.

IV. Result from each stage

A multi-stage methodology was followed to implement machine learning techniques to predict dengue fever cases in this study. Loading the data is a very critical stage, and we analyzed the nulls values in the dataset as represented in Figure 2 which is appropriate. Null values such as this must be handled to maintain the integrity and accuracy enhance the predictive modality.



```
year                                      0
weekofyear                                0
week_start_date                           0
total_cases                               0
station_precip_mm                        22
station_min_temp_c                       14
station_max_temp_c                       20
station_diur_temp_rng_c                  43
station_avg_temp_c                       43
reanalysis_tdtr_k                        10
reanalysis_specific_humidity_g_per_kg    10
reanalysis_sat_precip_amt_mm             13
reanalysis_relative_humidity_percent     10
reanalysis_precip_amt_kg_per_m2          10
reanalysis_min_air_temp_k                10
reanalysis_max_air_temp_k                10
reanalysis_dew_point_temp_k              10
reanalysis_avg_temp_k                    10
reanalysis_air_temp_k                    10
precipitation_amt_mm                     13
ndvi_sw                                  22
ndvi_se                                  22
ndvi_nw                                  52
ndvi_ne                                 194
city                                      0
```

Fig. 2. Check the missing values

In the data preprocessing stage, data were properly structured to prepare the dataset for analysis using different strategies to the load, merge and preprocess the features and labels of dengue fever cases from various CSV files. We segmented the data based on the city and the timing and normalized the data by clipping all values to observed in the test data. Missing values in the data were handled using the interpolation approach; the categorical values were encoded to make the data uniform for the model. This guarantees that the model is tested on real-world data hence can be generalized to the model the other random dataset.

During data exploration, we generated the scatter plot in Figure 3 showing that there was right-skewed distribution of the dengue cases in the various datasets in all the timing considered. The correlations with the dengue cases including the scatter plot in Figure 4 as well as the features relationships with feature relationship map in Figure 5 are essential to focus on feature selection. Some features can be removed since they overwhelmingly correlated with other features, and by their inclusion, it may make the model generalize to the other unique dataset.

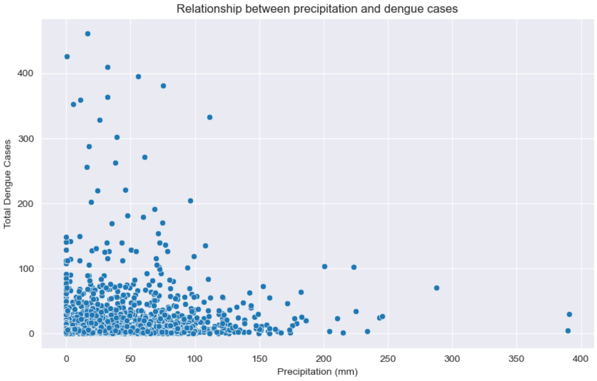


Fig. 3. Scatter plot of data distribution

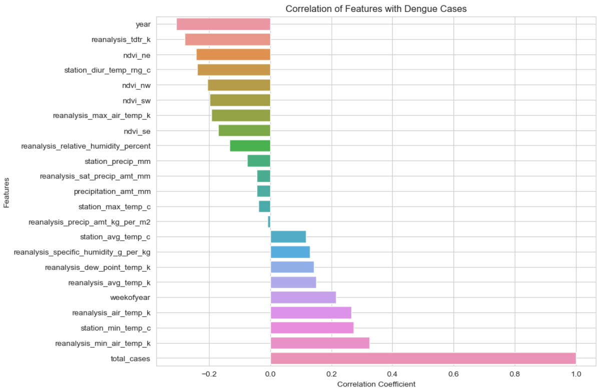


Fig. 4. Correlations with dengue cases

Furthermore, a correlation analysis in the stage of feature engineering, expressed through a heatmap matrix in Figure 5 and many features were discovered to have strong associations with one another that enable us to assess the importance of a priority of independent variables. From dates, we extracted month and day features in Figure 6 which added more columns to data file making them 26. Moreover, we also fed moving average characteristics, or focusing on climate-related factors of the highest interests in this study such as rainfall, air temperatures, and relative humidity. Newly generated features are expressed in Figure 7, further adding more columns to 35,

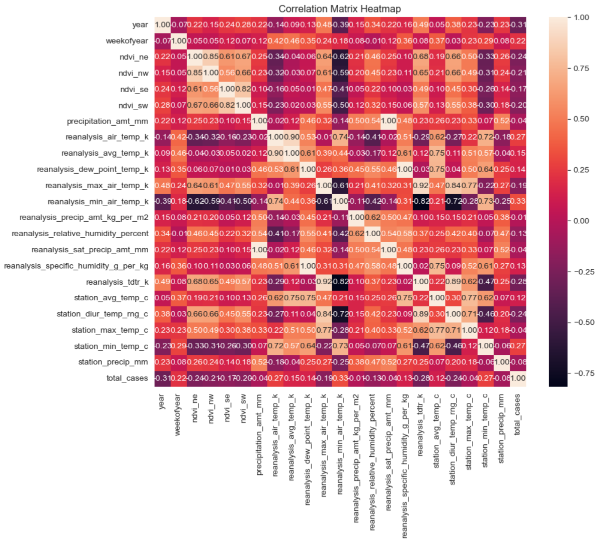
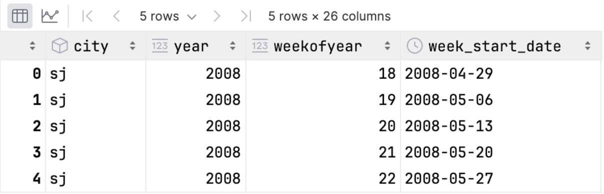


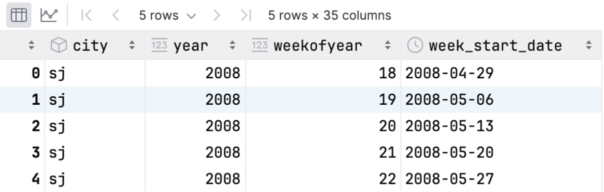
Fig. 5. Correlation matrix of features

representing richer dynamics within a time interaction with environmental factors. We also explored the use of different combinations of features to generate moving average features, including 'reanalys_precipt_kg_per_m2', 'station_max_temp_c', 'reanalys_specific_humidity_g_per_kg', etc., to evaluate their potential impact on model performance, which helps to identify factors that may not have been observed in the original data. In addition, we also attempted to remove low variance and high correlation features from the dataset to assess the impact of low variance and high correlation features on the predictive performance of the model.



	city	year	weekofyear	week_start_date
0	sj	2008	18	2008-04-29
1	sj	2008	19	2008-05-06
2	sj	2008	20	2008-05-13
3	sj	2008	21	2008-05-20
4	sj	2008	22	2008-05-27

Fig. 6. Extracting date features



	city	year	weekofyear	week_start_date
0	sj	2008	18	2008-04-29
1	sj	2008	19	2008-05-06
2	sj	2008	20	2008-05-13
3	sj	2008	21	2008-05-20
4	sj	2008	22	2008-05-27

Fig. 7. Generating new moving average features

In the modeling phase of our project, we applied various machine learning algorithms to forecast dengue fever cases effectively. The models tested included Linear Regression, Support Vector Machines, Gradient Boosting, Artificial Neural Networks, and Random Forest. Our primary evaluation metric was the Mean Absolute Error (MAE), which provides a clear measure of average prediction error in the same units as the data, making it particularly useful for interpreting the model's performance in practical terms.

TABLE I
OPTIMIZED PARAMETERS FROM GRID SEARCH

City	San Juan	Iquitos
max_depth	35	10
max_features	5	2
min_samples_leaf	3	2
min_samples_split	2	2
n_estimators	100	300
criterion	absolute error	absolute error

The Random Forest model demonstrated exceptional performance across all tested models, exhibiting the lowest Mean Absolute Error (MAE). This success is attributed to its ability to handle non-linear relationships and resist overfitting in complex epidemiological datasets. To optimize this model, TABLE I show that GridSearchCV was used to meticulously tune hyperparameters such as n_estimators and max_features, aiming to reduce the MAE.

TABLE II
CROSS-VALIDATION PARAMETERS FOR LINEAR REGRESSION MODELS

City	San Juan	Iquitos
Model Type	Linear Regression	Linear Regression
Cross-Validation Folds	5	5
Scoring Metric	MAE	MAE

TABLE III
CROSS-VALIDATION PARAMETERS FOR GRADIENT BOOSTING MODELS

City	San Juan	Iquitos
Model Type	Gradient Boosting	Gradient Boosting
Cross-Validation Folds	5	5
Scoring Metric	MAE	MAE

TABLE IV
CROSS-VALIDATION PARAMETERS FOR ANN MODELS

Parameter	San Juan	Iquitos
Model Type	ANN (MLP)	ANN (MLP)
Preprocessing	Standard Scaling	Standard Scaling
Max Iterations	500	500
Learning Rate Initial	0.001	0.001
Cross-Validation Folds	5	5
Scoring Metric	MAE	MAE

Tables II, III, IV, and V, respectively, detail the parameters for the remaining models, each optimized through rigorous cross-validation to ensure the best predictive performance.

TABLE V
CROSS-VALIDATION PARAMETERS FOR SVR MODELS

City	San Juan	Iquitos
Model Type	SVR	SVR
Cross-Validation Folds	5	5
Scoring Metric	MAE	MAE

V. DISCUSSIONS

In this study, different experimental methods for predicting dengue fever cases in different stages have been introduced through the machine learning method, such as data preprocessing, data excavation analysis, feature engineering, model and tunning. The results of various models were compared with existing literature research methods to enhance the prediction accuracy and prediction performance of dengue fever case on such big infectious disease problems like.

TABLE VI
MEAN ABSOLUTE ERROR (MAE) FOR VARIOUS PREDICTIVE MODELS

Model	MAE San Juan	MAE Iquitos
Random Forest	15.23	4.01
Linear Regression	25.93	6.58
Gradient Boosting Regressor	26.12	7.25
ANN (Artificial Neural Network)	24.35	6.30
SVR (Support Vector Regressor)	23.46	5.99

As shown in Table VI, the prediction model evaluations achieve considerable successful different models based on different cities that have been employed.

Moreover, the data cleaning phase of the research identified core issues in the dataset, such as missing values and right-skewness in the data distribution. Discoveries made from the excavation section and visual analysis ensured a truly accurate dataset conversion to clean and fill the dataset in the data cleaning stage. Secondly, the dataset was justified for integrity and comparability among them by trimming values to scale the data and data normalization. Guo et al. (2017)[3] also emphasized the importance of conducting preliminary data analysis in their research, which can ensure that the model can effectively capture key information. In addition, the exploration of data distribution and correlation provides valuable insights into the relationship between dengue fever cases and environmental factors, and provides reference for subsequent feature engineering decisions. Through exploration of the dataset, it can be seen that for the current case data, especially during low rainfall periods, the number of cases is relatively concentrated, which may indicate a complex nonlinear relationship between rainfall and the number of cases. This is consistent with the findings of Sarma et al. (2020) [9]. Similarly, Deb et al. (2017)[2] also emphasized the significant impact of climate conditions on the number of dengue fever cases.

The preprocessing strategies: median padding, forward-backward padding, and missing data were achieved. Median padding was more useful as this was evident from the MAE for the median padding closely resembled the ideal values, and the strategy was actually fits for our dataset.

Definitely, feature engineering affects the model performance, as Iqbal&Islam(2019)[6] emphasized the importance of features engineering in addressing this health problem in public. Thus, we tried other ways of extracting features: removing features with low variance and high correlation, and playing with new lag and interaction features. The goal was to solve overfitting and improve the accuracy. Finally, removing some had a marginal impact on MAE which showed an insignificant reduction, but adding new features make a meaningful difference in the outcome acquired.

During the model training phase, we found that the processing efficiency and time consumption of different models were basically the same. Random forests showed the best prediction results by comparing the models, according to the average absolute error and accuracy of outcomes. Similarly, Salim et al.[7] , and Sarma et al. [9] compared the prediction methods which upon the tree decision and random forests and then decided the model of random forests performed the best prediction. Additionally, Zhao et al.[12] compared random forests and artificial neural networks in prediction which showed that random forests did better in a few instances. This method of multi model comparison has been widely recommended in diverse literatures as it serves in testing different models to identify the most appropriate model to select for the specifics of the problem for higher accuracy and efficiency in prediction.

Our models showed varied performance across two distinct locales, as reflected in the MAE values from Table VI. The Random Forest model's superior performance in Iquitos, with an MAE of 4.01, contrasts sharply with the higher MAE values in San Juan, underscoring the need for localized model tuning. We choose the random forest model, and continuously carry out hyperparameter optimization, re-feature processing and other operations to make the accuracy and stability of the model optimal.

The efficacy of our models is visually supported by Figures 8 and 9, which display the predictive results using test data. These figures illustrate that our models are capable of closely replicating actual case numbers, thereby validating their practical utility in real-world scenarios.

Finally, by doing wide various methods experiments at various stage, we created good machine learning model to predict dengue fever case numbers using environmental factor data. Moreover, as shown in this paper the feature engineering helped optimization to choose a good model. This model is well suited to predict dengue fever case number. Furthermore, moving average was helpful to optimize ML model created with environmental factors. Thus, it could be done optimization various ways which is important to understand when modeling in the future. Our research provided important data and scientific basis for dengue fever outbreak, prevention and control prediction. It could be used for real-world application.

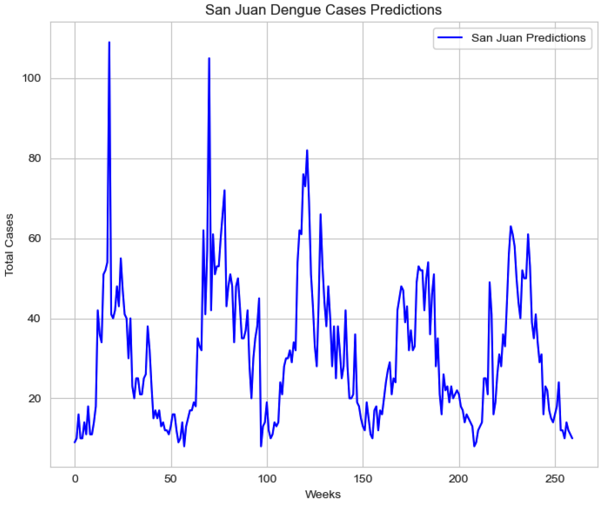


Fig. 8. SanJuan Prediction Using Test Data

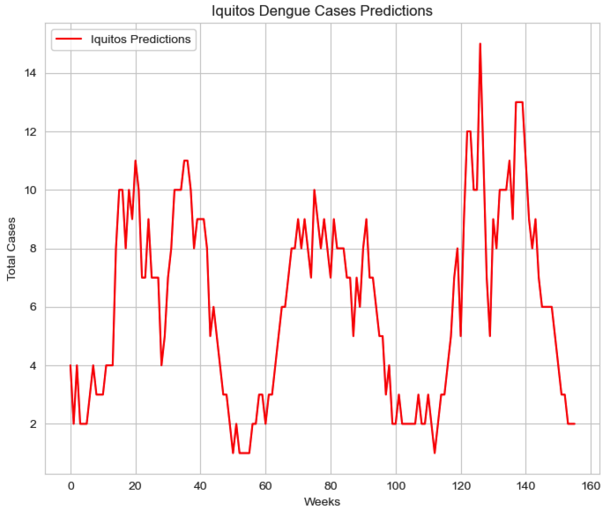
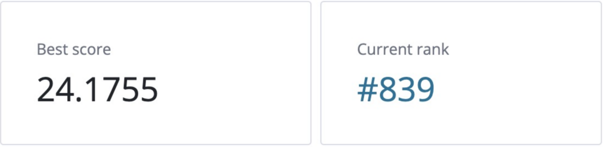


Fig. 9. Iquitos Prediction Using Data

VI. Conclusions and Recommendations for Future Research

Following the submission of predictions for test data on DataDriven, our study achieved a Mean Absolute Error (MAE) of 24.117, positioning us at rank 839, as the figure shows. This is a commendable result that highlights the effectiveness of our approach. Among the machine learning models evaluated, the Random Forest model emerged as the most effective, demonstrating superior capability in accurately forecasting the occurrence of dengue fever cases.

Addressing our research questions is crucial for refining the predictive models and enhancing their practical utility in public health scenarios. Thus, we now turn our attention to the critical research questions that emerged from our study, discussing each in the context of our findings and the broader implications for disease prediction and management.



Best score

24.1755

Current rank

#839

Fig. 10. The submission of prediction

1) **Key Climate Factors:** Our study reveals the climate factors associated with dengue fever cases in the given regions, such as temperature, humidity, and precipitation. Climate factors affect the population dynamics of Aedes mosquitoes, which are responsible for the dengue outbreaks.
2) **Effective Machine Learning Models:** The Random Forest model significantly performed well in dengue fever epidemic prediction for the given region using historical climatic and health data. The solves non-linear problems and feature interactions and, is, thus, effective in various conditions.
3) **Forecasting Accuracy:** The application of this predictive model, particularly the Random Forest, holds potential in accurately predicting the timing and intensity of dengue outbreaks. These could be useful in public health monitoring and preparedness. Additional data for the model and improved parameters may enhance the prediction to be more accurate and reliable.

Therefore, Future research should aim to enhance the model's performance by improving the quality and quantity of data, such as by collecting more comprehensive data sets or refining data collection methods. Alternatively, the utilization of advanced model architectures or novel algorithms may enhance the prediction accuracy as well.

In conclusion, the likelihood of future enhancements is high, and those enhancements will include improved accuracy in case prediction. As a result, more effective and globally endorsed tools will be available to the public health management sector. Such tools are expected to have an impact on society by helping to contain the spread of dengue fever and similar cases.

References

[1] Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D., & Watanabe, K. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC Infectious Diseases, 18*(1). https://doi.org/10.1186/s12879-018-3066-0

[2] Deb, S., Acebedo, C. M. L., Dhanapal, G., and Heng, C. M. C., "An ensemble prediction approach to weekly Dengue cases forecasting based on climatic and terrain conditions," *Journal of Health and Social Sciences*, vol. 2, no. 3, pp. 257-272, 2017. https://doi.org/10.19204/2017/nnsm3

[3] Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., Luo, G., Li, Z., He, J., and Zhang, Y., "Developing a dengue forecast model using machine learning: A case study in China," *PLoS Neglected Tropical Diseases*, vol. 11, no. 10, p. e0005973, 2017. https://doi.org/10.1371/journal.pntd.0005973

[4] Hoyos, W., Aguilar, J., and Toro, M., "Dengue models based on machine learning techniques: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 119, p. 102157, 2021. https://doi.org/10.1016/j.artmed.2021.102157

[5] Ho, T. S., Weng, T., Wang, J., Han, H., Cheng, H., Yang, C., Yu, C. H., Liu, Y. J., Hu, C. H., Huang, C. F., Chen, M. H., King, C. C., Oyang, Y. J., & Liu, C. C. (2020). Comparing machine learning with case-control models to identify confirmed dengue cases. *PLoS Neglected Tropical Diseases, 14*(11), e0008843. https://doi.org/10.1371/journal.pntd.0008843

[6] Iqbal, N., and Islam, M. M., “Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers,” *Informatica*, vol. 43, no. 3, 2021. https://doi.org/10.31449/inf.v43i3.1548

[7] N. A. Salim, Y. B. Wah, C. Reeves, M. Smith, W. F. W. Yaacob, R. N. Mudin, R. Dapari, N. N. F. F. Sapri, and U. Haque, ”Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques,” *Scientific Reports*, vol. 11, no. 1, 2021. https://doi.org/10.1038/s41598-020-79193-2

[8] Nejad, F. Y., & Varathan, K. D. (2021). Identification of significant climatic risk factors and machine learning models in dengue outbreak prediction. *BMC Medical Informatics and Decision Making, 21*(1). https://doi.org/10.1186/s12911-021-01493-y

[9] Sarma, D., Hossain, S., Mittra, T., Bhuiya, M. A. M., Saha, I., and Chakma, R., “Dengue prediction using machine learning algorithms,” *IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)*, 2020. https://doi.org/10.1109/R10-HTC49770.2020.9357035

[10] Sajana, T., Navya, M., Gayathri, Y., & Reshma, N. (2018b). Classification of Dengue using Machine Learning Techniques. *International Journal of Engineering & Technology, 7*(2.32), 212. https://doi.org/10.14419/ijet.v7i2.32.15570

[11] Siddiq, A., Shukla, N., & Pradhan, B. (2021). Predicting dengue fever transmission using machine learning methods. *2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. https://doi.org/10.1109/ieem50564.2021.9672977

[12] Zhao, N., Charland, K., Carabalí, M., Nsoesie, E. O., Maheu-Giroux, M., Rees, E. E., Yuan, M., Balaguera, C. G., Ramirez, G. J., & Zinszer, K. (2020). Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. *PLoS Neglected Tropical Diseases, 14*(9), e0008056. https://doi.org/10.1371/journal.pntd.0008056