# Predicting Dengue Spread in San Juan and Iquitos

An OpenAI challenge

Anuranjan M B, Chandana.Divya Vani, Chanpreet Singh, Samarjeet Barman, Dr. Kuldeep Chaurasia

***Abstract*—Dengue fever is one of the well renowned mosquito-borne diseases that occurs in tropical and sub-tropical parts of the world. The transmission of dengue can be related to climatic variables since it is spread by mosquitoes. Using the environmental data collected by various Government agencies we try to predict the number of dengue fever cases reported each week in two cities- San Juan, Puerto Rico and Iquitos, Peru. This study aims to design two time series based Nonlinear Regression Models (NLRM) and a data manipulation technique using different parameters such as temperature, vegetation and rainfall data and incorporating time series, dimension reduction for better prediction of dengue outbreak. This study considers three different modelling techniques namely Interpolation, Gradient Boosting Regression and Random Forest Regression. Parameters were tuned and adjusted for optimal performance. Comparisons of results are made based on prediction accuracy and mean absolute error (MAE). The performance was analyzed and the result points out that the Gradient Boosting Regression performs significantly better than the other models and is therefore considered to be a better approach. Future improvements to result can be made by obtaining large amounts of meaningful data and implementing better models associated with time series predicting**

***Index Terms*—Machine Learning, Dengue, Predictive models, Pattern analysis, etc.**

## I. Introduction

DENGUE is a exponentially emerging commonly-prone viral disease in many parts of the world. Dengue flourishes in urban poor areas, suburbs and the countryside but also affects more affluent neighbourhoods in tropical and subtropical countries. Apart from causing a severe flu-like illness, sometimes it's prone to cause a lethal complication called severe dengue. According to WHO (World Health Organisation) 50-100 million infections are now estimated to occur annually [1]. The Aedes aegypti mosquito transmits the viruses that cause dengue. Because dengue is carried by mosquitoes, the transmission of dengue is related to meteorological and environmental data such as temperature, precipitation and vegetation. A vaccine to prevent dengue is licensed and available in some countries. Surprisingly, the vaccine manufacturer has announced in 2017 that people who have not been previously infected and received the vaccine may be at risk of developing severe dengue if they get infected following the vaccination [2]. The backbone for treatment of dengue is supportive care [8]. It becomes vital to recognize the 3 phases of dengue [9]: febrile, critical and recovery. Hence, serious efforts are required to control and prevent this disease. This makes predictions on dengue outbreaks very important. With the help of this prediction, health departments across the world can take preventive measures to combat dengue fever before the outbreak begins, saving millions of lives. The two cities considered for this experiment are San Juan, Puerto Rico and Iquitos, Peru. San Juan is the capital city of Puerto Rico and also happens to be the most populous municipality. Iquitos is the capital city of the Mayan Province and Loreto Region and is claimed to be the ninth most populous city of Peru.

## II. Related Work

Several approaches have been used to predict dengue outbreak. Rachata et al researched on the use of ANNs with an entropy technique to build a predictive model for Dengue outbreaks in Thailand [10]. N. Aditya Sundar et al proposed a system which contributes to the detection of dengue disease using blood pressure, viral infection, sex and age factors. It used Naïve Bayesian classification to train the model on existing data. Even patient and nurses can use this model to supply features and get the prediction on disease occurrence. A study in Sri Lanka, using past weather patterns and past dengue cases as inputs, introduced an ANN(Artificial Neural Network) to predict the dengue outbreak in Kandy district [11]. In 2014, [3] showed that land use factors other than human settlements, like different types of agricultural land , water bodies and forest can be associated with reported dengue cases in the state of Selangor, Malaysia and used boosted regression to account for non-linearities and interactions between these factors. In 2017, [4] proposed that human mobility has significant effect on the importation of dengue to an immunologically dengue 'naive' regions.

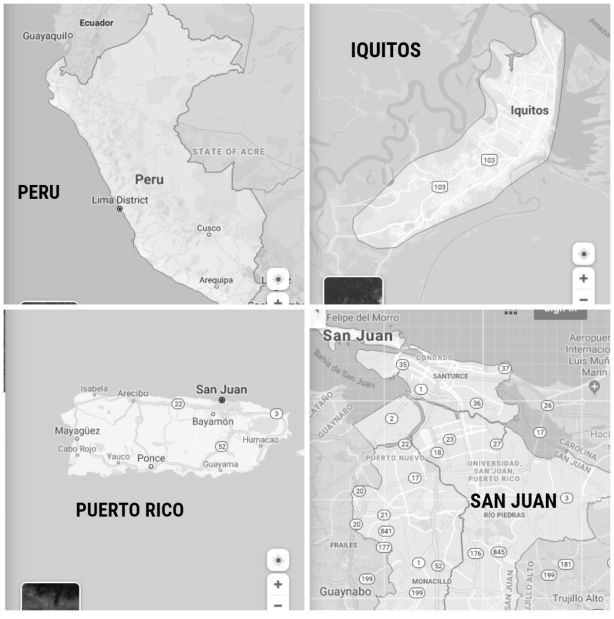


Fig. 1. Source: Google Maps

The data set was derived using Mobile Network Big Data in Sri Lanka. The predictions were made using ANN and XGBoost. Recent work of P.Muhilthini et al [6] in 2018 proposed a dengue possibility forecasting model, the data set contains information about number of dengue cases observed every week for several years in many number of countries. It contains details about the weather conditions like temperature, precipitation amount, humidity and so on using GBR to find the pattern and dependencies in the given training data set and predict number of dengue cases for the given week and year of a country in the test data set. In 2017 [5] proposed a model for the prediction of dengue, diabetes and swine flu using random forest. The main aim of this model is to predict the disease by using the symptoms taken from patients and to recommend the specialized doctor, from this the risky cases of that particular disease in a week of that particular area was also calculated. Ong et al used random forest regression to predict the risk rank of dengue transmission in 1km grids, with dengue, population, entomological and environment data in Singapore. More than 80% of the observed risk ranks fell within the 80% prediction interval [7].

## III. Validation

Performance metrics used: MAE (Mean Absolute Error), Predicted vs Actual Plots

The Mean Absolute Error (MAE) measures the closeness of forecasts predictions to the actual outcomes. It's given by (fig. 2):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|\,.$$

$$AE = |e_i| = |y_i - \hat{y}_i|$$

$$Actual = y_i$$

$$Predicted = \hat{y}_i$$

Fig. 2. Mean Absolute Error

## IV. Methodology

The methodology involved seven steps- Acquiring the data set, Pre-processing the data, selecting a relevant predictive model, building the model, fitting and training the model with the training data, tuning the parameters for optimal performance, predicting the values (Fig 3).

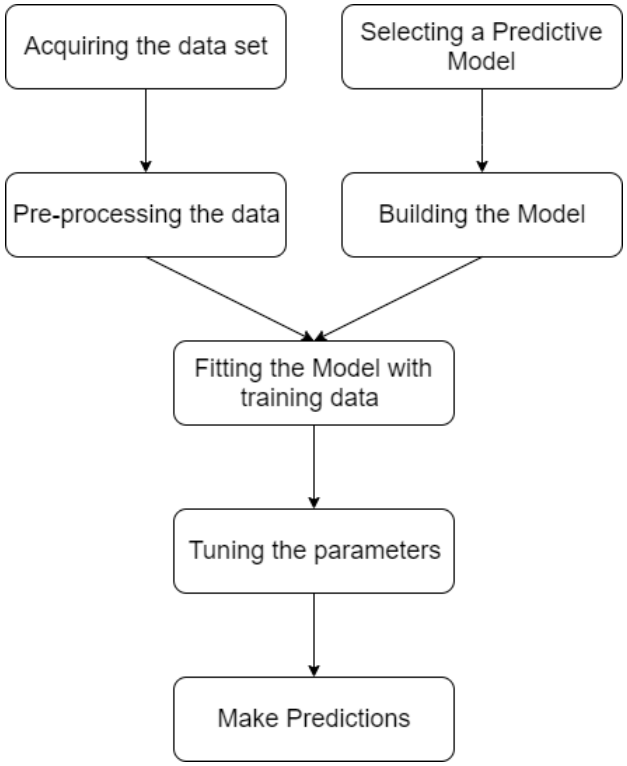


Fig. 3. Methodology

### A. Data set and Data Pre-processing

The data used in this study was collected by various U.S Federal Government Agencies (including Centers for Disease Control and Prevention, National Oceanic and Atmospheric Administration and the U.S. Department of Commerce). The consolidated data set was acquired from the "Data Download" section of the openAI competition "DengAI: Predicting Disease Spread" hosted by drivendata.org.

The data set was a combination of meteorological data including the features: City abbreviations: sj for San Juan and iq for Iquitos, week_start_date, Maximum temperature , Minimum temperature, Average temperature, Total precipitation, Diurnal temperature range, Total precipitation, Mean dew point temperature, Mean air temperature, Mean relative humidity, Mean specific humidity, Maximum air temperature, Minimum air temperature, Average air temperature, Pixel southeast of city centroid, Pixel southwest of city centroid, Pixel northeast of city centroid, Pixel northwest of city centroid.

The acquired data had to be pre-processed for maximum performance of the models. Various data pre-processing techniques were used in this study. To begin with, the data-set had to be divided based on the two cities. A new feature called "month" had to be extracted from the "week_start_date" since, the latter is a little harder to work with. The NaN (Not a Number) values were replaced with the mean of their respective columns. It can be noticed that few features

associated temperature were recorded in Kelvin and a few others in Celsius. The features were all converted to Celsius.

### B. Interpolation

Interpolation is the process of acquiring a function from a set of data points such that the function passes through all the given data points and can be used to appraise data points in-between the given ones.

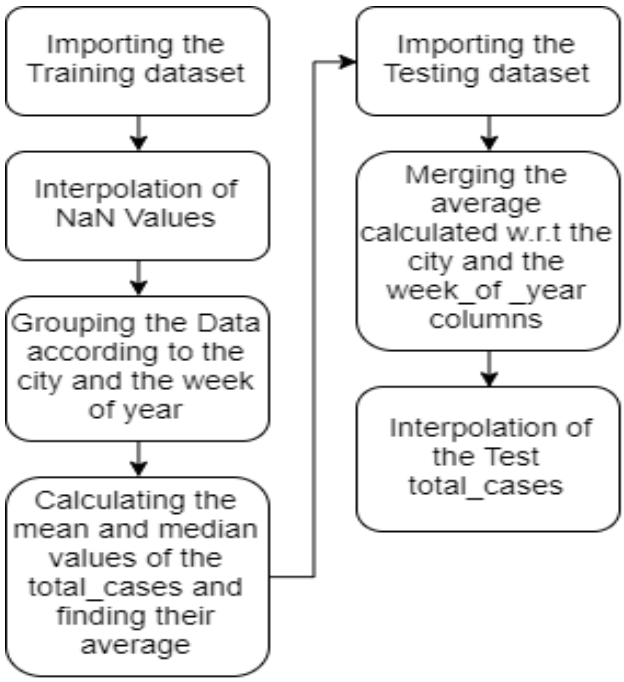


Fig. 4. Interpolation

Using interpolation (fig. 4) and the weekly pattern of the recorded cases, it's possible to make a plausible prediction. In order to attain a weekly pattern, the data from the training set can be grouped together in accordance with the week_of_year column. The mean and median of this grouped data is averaged to get an appreciable week pattern. The pattern is then later used to predict the total_cases reported in the testing data. Interpolation of the predicted total_cases is done for accurate results.

### C. Time Series forecasting with Random Forest

Random forest is a Supervised Learning Algorithm which uses ensemble learning method. It can be used for both classification and regression. The trees in Random Forests run in parallel meaning there's no interaction between these these trees. A random forest combines many decision trees. The differences lie in the number of features that can be split at each node and the randomness added by random selection of sample data from the original data by the decision trees to avoid over-fitting. A time series is a vector of values that are indexed by time. Sometimes it's necessary to perform some pre-processing to make it comprehensible.

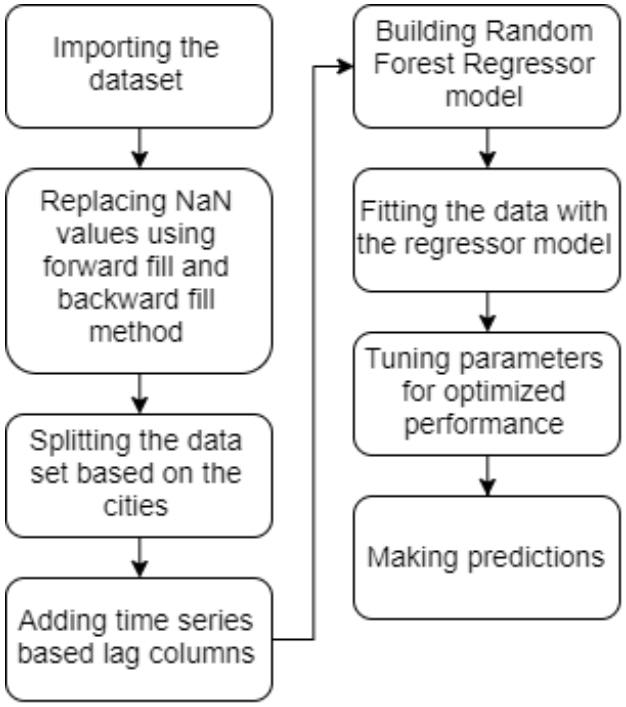


Fig. 5. Random Forest Regression

This algorithm involved the prediction of total_cases using Time Series forecasting with Random Forest. The data set after being imported was pre-processed. The data pre-processing involved replacing the NaN values using the forward and backward fill method. It was necessary to split the data set based on the cities for accurate predictions. The next step involved introducing time-series based lag which served as the base for the time-series forecasting of the model. A random forest regressor with the necessary parameters was built. The next step involved the fitting of the data set to the model. Two parameters were tuned for better performance.

| Parameter Tuning | | | |
|---|---|---|---|
| Parameter | Values | San Juan | Iquitos |
| n_estimator | 10,20,50,100,200,300,500 | 20 | 100 |
| max_depth | 10,15,20,40,50,100 | 10 | 10 |

Table:1 Parameters tuned: Random Forest

The two parameters considered are n_estimators and max_depth. After trying out various values, the above tabulated values (table 1) were considered for the cities as using this parameters gave the best result. Following which the model was trained again with the tuned parameters and used to predict the total_cases.

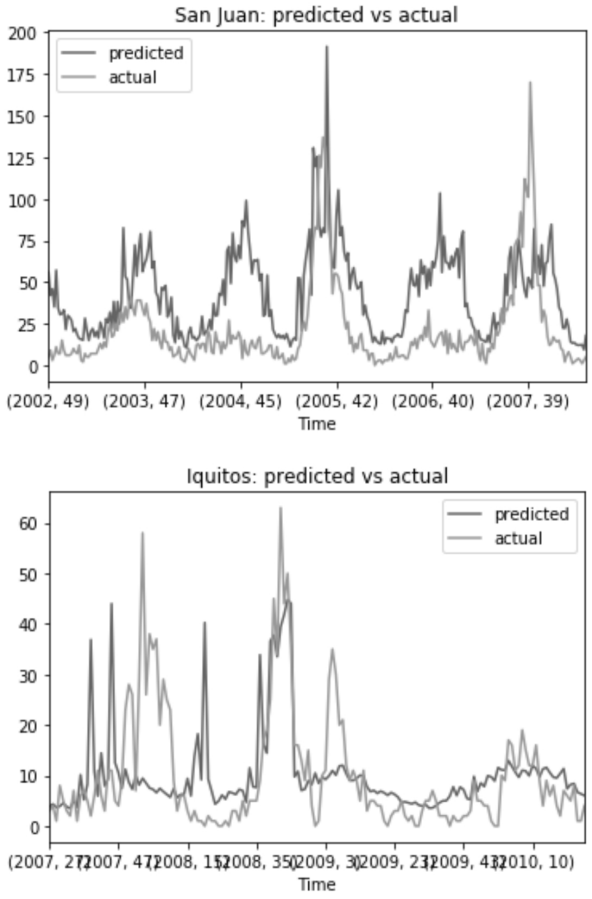


Fig. 6. Plotting result - Random Forest

The Mean Absolute Error computed (using time-series based algorithm) for both the cities was better than the MAE calculated using normal Random Forest algorithm

### *D. Gradient boosting for Time Series Prediction*

Gradient boosting is one of the most powerful techniques for building predictive models. It is a machine learning technique widely used for regression and classification problems. Gradient boosting aims at producing an ensemble of weak prediction models. The algorithm repetitively leverages the patterns in residuals and bolster a model with weak prediction to improve it.

This algorithm involved the prediction of total_cases using Time Series forecasting with Gradient boosting. The data set after being imported needed to be pre-processed. The data pre-processing involved replacing the NaN values using the forward and backward fill method. It was necessary to split the data set based on the cities for more accurate predictions. The next step involved introducing time-series based lag which served as the base for the time-series forecasting of the model. A gradient boosting regressor with the necessary parameters was built.

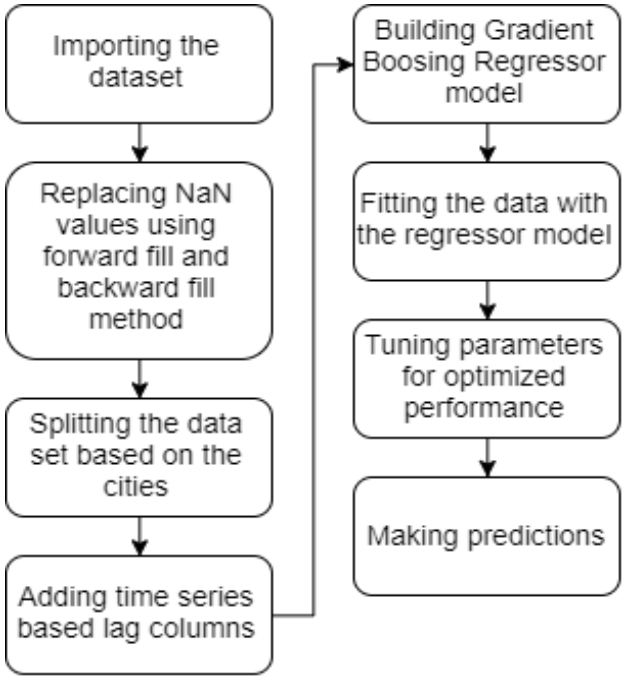


Fig. 7. Gradient Boosting Regression

The next step involved the fitting of the data set to the model. Four parameters were tuned for better performance.

| Parameter Tuning | | | |
|---|---|---|---|
| Parameter | Values | San Juan | Iquitos |
| max_iter | 10,20,50,100,200,300,500 | 50 | 10 |
| max_leaf_nodes | 10,16,32,64,128,256 | 10 | 16 |
| max_depth | 4,8,16,32 | 8 | 10 |
| max_bins | 10,16,32,64,128,256 | 10 | 32 |

Table:2 Parameters tuned: Gradient Boosting

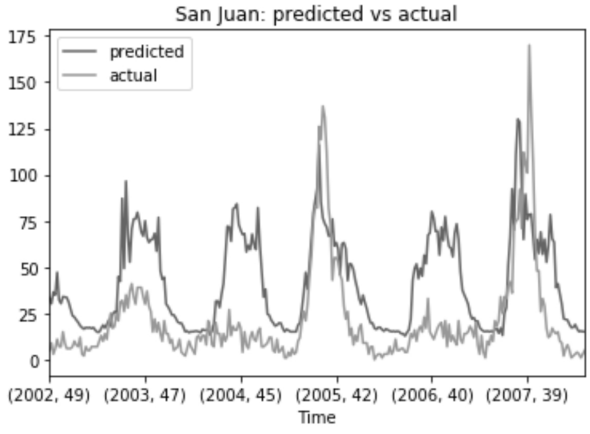


Fig. 8. Plotting result - Gradient boosting

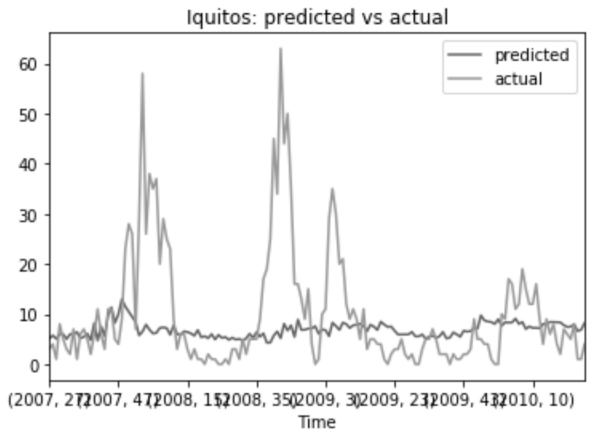


Fig. 9. Plotting result - Gradient boosting

## V. Results and Discussion

Time Series based prediction is a good alternative for ordinary predictive models. The three algorithm used were Manipulation using Interpolation, Times Series based Random Forest and Gradient boosting. Although these algorithms perform well, the time series based gradient boosting algorithm seems have the best of results. The MAE of this algorithm is better than those of others. The respective MAE values have been tabulated (table 3).

The following limitations made the process of achieving better results harder and also undermined the scope of using deep neural networks to make predictions: Data set given data is limited or less, Two cities are located in different geographical areas, Dengue occurs at particular months of the year. These result can further be developed by using models that are known to work better for Time Series based data like the ARIMA model.

| Comparing Results | | |
|---|---|---|
| Algortihm | San Juan | Iquitos |
| Random Forest | 26.66 | 6.7 |
| Gradient Boosting | 24.11 | 7.36 |
| Interpolation | 25.98 | 25.98 |

Table:3 Result Comparison

## Acknowledgment

The authors would like to thank Bennett University for supporting the project and providing all necessary facilities. The authors would also like to thank the following U.S Federal Government Agencies for providing the data- CDC (Centres for Disease Control and Prevention), National Oceanic and Atmospheric Administration and the U.S. Department of Commerce. The project was proposed through a series of competitions held by DrivenData.Org. The authors would like to acknowledge drivendata.org for this opportunity. This project was promoted by LeadingIndia.AI. The authors would also like to extend their regards to the program coordinator Dr. Madhushi Verma and the director Dr. Deepak Garg.

## References

[1] Dengue and severe dengue. [Online]. Available: https://www.who.int/health-topics/dengue-and-severe-dengue

[2] Dengue Vaccine. [Online]. Available: https://www.who.int/health-topics/dengue-and-severe-dengue

[3] Cheong, Y. L., Leitão, P. J., Lakes, T. (2014). Assessment of land use factors associated with dengue cases in Malaysia using Boosted Regression Trees. Spatial and spatio-temporal epidemiology.

[4] K.G.S. Dharmawardana, J.N. Lokuge, P.S.B. Dassanayake, M.L. Sirisena, M.L. Fernando, A.S. Perera, and S. Lokanathan. (2017). Predictive Model for the Dengue Incidences in Sri Lanka Using Mobile Network Big Data.

[5] Tate, A., Gavhane, U., Pawar, J., Rajpurohit, B., Deshmukh, G. B. (2017). Prediction of Dengue, Diabetes and Swine Flu Using Random Forest Classification Algorithm.

[6] Muhilthini, P., Meenakshi, B. S., Lekha, S. L., Santhanalakshmi, S. T. (2018). Dengue Possibility Forecasting Model using Machine Learning Algorithms.

[7] Ong, J., Liu, X., Rajarethinam, J., Kok, S. Y., Liang, S., Tang, C. S., ... Yap, G. (2018). Mapping dengue risk in Singapore using Random Forest. PLoS neglected tropical diseases.

[8] Wallace D, Canouet V, Garbes P, Wartel TA (2013). Challenges in the clinical development of a dengue vaccine. Curr Opin Virol. 2013 Jun; 3(3):352-6.

[9] Stages of Dengue. [Online]. Available: https://www.cdc.gov/dengue/training/cme/ccm/page86100.html

[10] N. Rachata, P. Charoenkwan, T. Yooyativong, K. Chamnongthai, C. Lursinsap, and K. Higuchi. “Automatic prediction system of dengue haemorrhagic-fever outbreak risk by using entropy and artificial neural network,” no. Iscit, 2008, pp. 210–214.

[11] P. H. M. N. Herath, A. A. I. Perera, and H. P. Wijekoon, “Prediction of dengue outbreaks in Sri Lanka using artificial neural networks.” International Journal of Computer Applications, vol. 101, no. 15, pp. 1–5, 2014.