Predicting Spread of Dengue Disease

Submitted to
Prof Dr. Muhammad ShakhaWat Hossain
University of Barishal
Barishal, Bangladesh

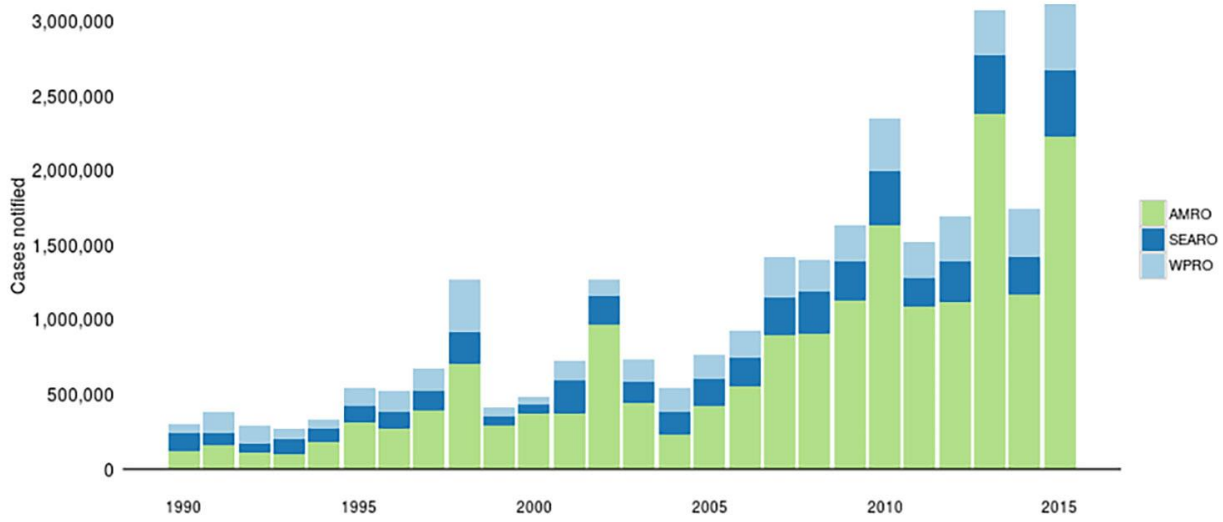
18 September, 2021

by
Md Rashidunnabi
University of Barishal
Barishal, Bangladesh

Abstract — Since a few years now, dengue fever has been on the rise. This is a worrisome indicator for today's human civilization. It spreads quite rapidly in India, with numerous countries reporting multiple infections and fatalities. In order to understand the reasons behind the sickness, a case study had been carried out and the dengue data set was studied in different Indian states. The aim of this project is to employ machine learning techniques to estimate the number of fatalities in the near future. This information will allow the governments to take the actions they need to reduce the threat posed by this disease and save the

1. Introduction

Dengue is carried by Aedes aegypti and Aedes albopictus, two kinds of mosquitoes. Four serotypes and three kinds of illness are connected with mosquitos such as dengue fever and dengue hemorrhagic fever (DHF), and DSS. One serotype cannot be infected and the build-up of multiple serotypes results in a greater likelihood of more severe Dengue symptoms. In the region where there is a danger of dengue transmission, 2,5 billion people, almost 40 percent of the world's population now live. According to the World Health Organization (WHO), 500,000 cases and 22,000 fatalities are estimated to be approximately 50 to 100 million infects every year.

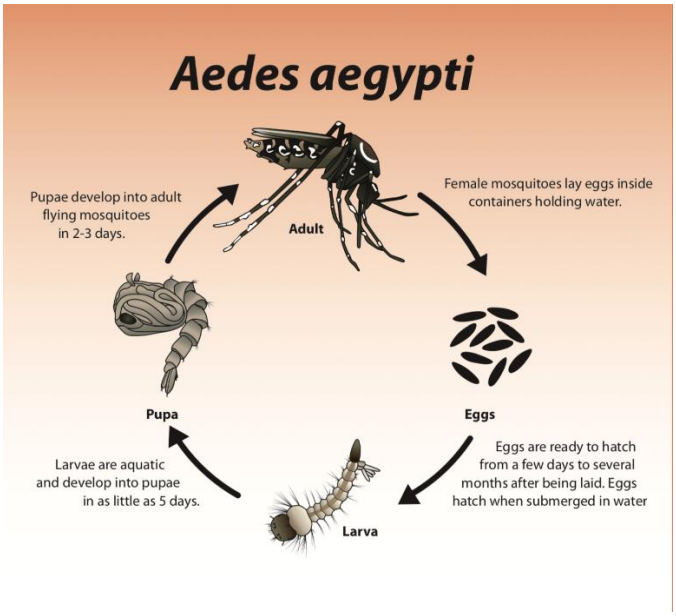


Number of dengue cases reported to WHO, probable or laboratory confirmed 1990–2015

The virus may generally only be transmitted from people to mosquitoes or mosquitoes to people. Dengue virus may also, in rare circumstances, be passed to the fetus through blood transfusions and organ transplants from infected donors, or through her infected pregnant mum. The mosque must feed a person during a five-day period when high quantities of virus appear in the blood for the transfer from a mosquito to a human being. Infection symptoms start four - seven days after the mosquito bite and persist three - ten days. In the case of dengue virus in a blood meal, it takes 8 to 12 days for vi to be incubated.

Vectors for Dengue Virus

Because of the steady habitat, female mosquitos that are carriers of the dengue virus prefer stagnant water surroundings. Around 100-200 eggs per lot can be laid and up to 5 lots in the lifecycle can be produced. They are typically placed on moist, flooding areas. The larvae have their eggs when the water level increases owing to rain or other factors. But these eggs may thrive for several months, even without moisture, when eggs are inundated by water. This mosquito's life cycle comprises an aquatic and terrestrial phases. When mosquito achieves adulthood during the terrestrial phase the life cycle may vary, depending of environmental factors, from 2 weeks to a month.



Life Cycle of Aedes aegypti, the main vector for the dengue virus

These mosquitos are very vulnerable, including temperature, precipitation and humidity, to environmental factors. The climate is often linked to dengue events. Additional studies have utilized several templates such as temperature, precipitation and humidity to prevent additional malaria and zika illnesses. Higher temperature settings can shorten the time needed to reproduce the virus in a mosquito, and make the insect more likely to infect a human before it dies. Precipitation leads to flooded water and the quantity of dengue virus vectors available is increased. Moisture delivers moisture and is better linked to mosquito survival. In research to forecast a virus and may offer data on urban structure and environmental variables in any particular area, the Normalized Vegetation Difference Index (NDVI) has been previously employed. These are the characteristics that we were given in the competition. In forecasting instances of dengue virus, they are crucial to us.

2. Background

With extended fields and recent increases in incidence, Dengue is still a significant public health problem. There are still no precise and accurate projections of the incidence of dengue in China. We aim towards the development of an exact prediction model of dengue using the state-of-the-art machine learning techniques. Methodology/main results: Weekly dengue incidences, search questions from Baidu and climatic conditions (mean temperature, relative humidity and rainfall). In conjunction with climatic variables, a dengue search index was developed for the development of prediction models. The year and week recorded were also incorporated in the long-term trend and seasonality control models . As candidate models, several learners utilized dengue incidence predicting methods, including SVR, step-down linear regression model, GBM gradient boosted regression tree algorithm and linear regression model LASSO (Linear Gradient Boast), Extra Tree and Random Forest). The performance and fitness of the models were evaluated using the RMSE and R-squared metrics. In order to evaluate the validity of the models, autocorrelation and partial analysis of the autocorrelation functions were studied for the remaining models. The models have been verified with five ot dengue monitoring data . Moreover, in order to monitor dengue and anticipate disease outbreaks in other sectors the ExtraTree model had constantly lowest predictive error rates. Conclusion and importance: in contrast with other forecasting approaches tested in this study,

the suggested ExtraTree model demonstrated higher performance. The results can assist government and community respond to dengue epidemics at an early stage.

3. System Diagram

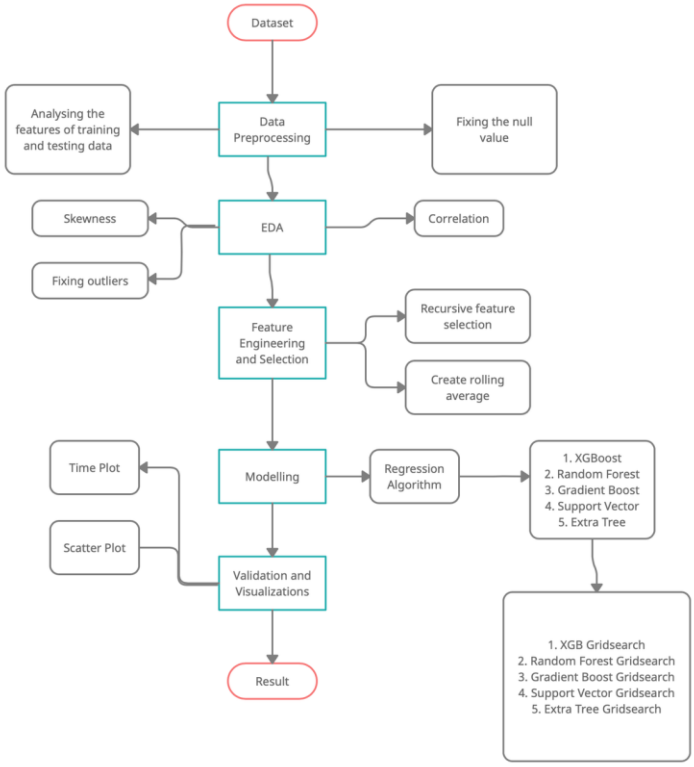


Figure : System diagram

4.Objective

Moreover, in order to monitor dengue and anticipate disease outbreaks in other sectors the ExtraTree model had constantly lowest predictive error rates. Conclusion and importance: in contrast with other forecasting approaches tested in this study, the suggested ExtraTree model demonstrated higher performance. The results can assist government and community respond to dengue epidemics at an early stage.

5. Experiment

5.1 Corpus

The characteristics in this data set

Our dataset has been gathered from https://www.kaggle.com/qianyigang129/dengai-dataset

The following package (year, week of year) information is supplied to us:

(Included as a _unit suffix on the function name if applicable.)

City and date indicators

- city – City abbreviations: sj for San Juan and iq for Iquitos
- week_start_date – Date given in yyyy-mm-dd format

NOAA daily observations of the NOAA GHCN climate data stations

- station_max_temp_c – Maximum temperature
- station_min_temp_c – Minimum temperature
- station_avg_temp_c – Average temperature
- station_precip_mm – Total precipitation
- station_diur_temp_rng_c – Diurnal temperature range

Measuring PERSIANN satellite rainfall (0.25x0.25 degree scale)

- precipitation_amt_mm – Total precipitation

Climate Climate Climate Forecast System Measurements for NOAA (0.5x0.5 degree scale)

- reanalysis_sat_precip_amt_mm – Total precipitation
- reanalysis_dew_point_temp_k – Mean dew point temperature
- reanalysis_air_temp_k – Mean air temperature
- reanalysis_relative_humidity_percent – Mean relative humidity
- reanalysis_specific_humidity_g_per_kg – Mean specific humidity
- reanalysis_precip_amt_kg_per_m2 – Total precipitation
- reanalysis_max_air_temp_k – Maximum air temperature
- reanalysis_min_air_temp_k – Minimum air temperature
- reanalysis_avg_temp_k – Average air temperature
- reanalysis_tdtr_k – Diurnal temperature range

NOAA's Normalized Vegetation Difference Index (0.5X0.5 degree scale) observations Vegetation satellite - NDVI - Normalized differential vegetation index

- ndvi_se – Pixel southeast of city centroid
- ndvi_sw – Pixel southwest of city centroid
- ndvi_ne – Pixel northeast of city centroid
- ndvi_nw – Pixel northwest of city centroid

Training Data and Testing data :

Figure 1 : Shape of Training and Testing data

There are 1456 rows and 25 columns in the training dataset. And in the testing dataset there are 415 rows and 24 columns.

5.2 Experiments Steps :

1. **Data Preparation**

We will have to examine whether the provided data sets include missing values.
Training Features Checking NA number:

Training Labels Check Number of NA:

```
Out[10]: city                                    0
         year                                    0
         weekofyear                              0
         week_start_date                         0
         ndvi_ne                               194
         ndvi_nw                                52
         ndvi_se                                22
         ndvi_sw                                22
         precipitation_amt_mm                   13
         reanalysis_air_temp_k                  10
         reanalysis_avg_temp_k                  10
         reanalysis_dew_point_temp_k            10
         reanalysis_max_air_temp_k              10
         reanalysis_min_air_temp_k              10
         reanalysis_precip_amt_kg_per_m2        10
         reanalysis_relative_humidity_percent   10
         reanalysis_sat_precip_amt_mm           13
         reanalysis_specific_humidity_g_per_kg  10
         reanalysis_tdtr_k                      10
         station_avg_temp_c                     43
         station_diur_temp_rng_c                43
         station_max_temp_c                     20
         station_min_temp_c                     14
         station_precip_mm                      22
         total_cases                             0
         dtype: int64
```

In the specified training functions, NaN values exist. For all columns of trainings with numeric values, the NaN values appear to be present. NaN is present mostly in satellite vegetation data for nvdi_* and GHCN daily cli mate data weather station_* reported climatic conditions. When doing an analysis, NaN values will be troubl esome. Since we know null values are sequence-ordered, we opted to fill the nulls using forward-fill in order t o spread the preceding value forward instead of discarding the rows. We may opt out of using variables with frequent NaN values since they may have an effect on the findings As previously mentioned in the problem d escription, several variables in the dataset are likewise identical but come from other sources. There may be r elatively similar findings and variations in certain factors. By eliminating these factors, we hope of preventing multi-linearity.

We have chosen to construct shifted total-cases because the cases are reported late in comparison with the real mosquito bitings and people's illnesses. We'll find out what variables lead to larger shift correlations. Because a mosquito takes approximately 8-10 days to enter his adult stage, the virus will be contagious during an incubation duration of around 8-12 days and for Dengue virus symptoms there will be time delays of 7-10 days. There is a difference of around 3-5 weeks, and we shall investigate 4 weeks' shift outcomes. Some characteristics such as the NDVI are currently being considered, so as to make some preliminary decisions about our data.

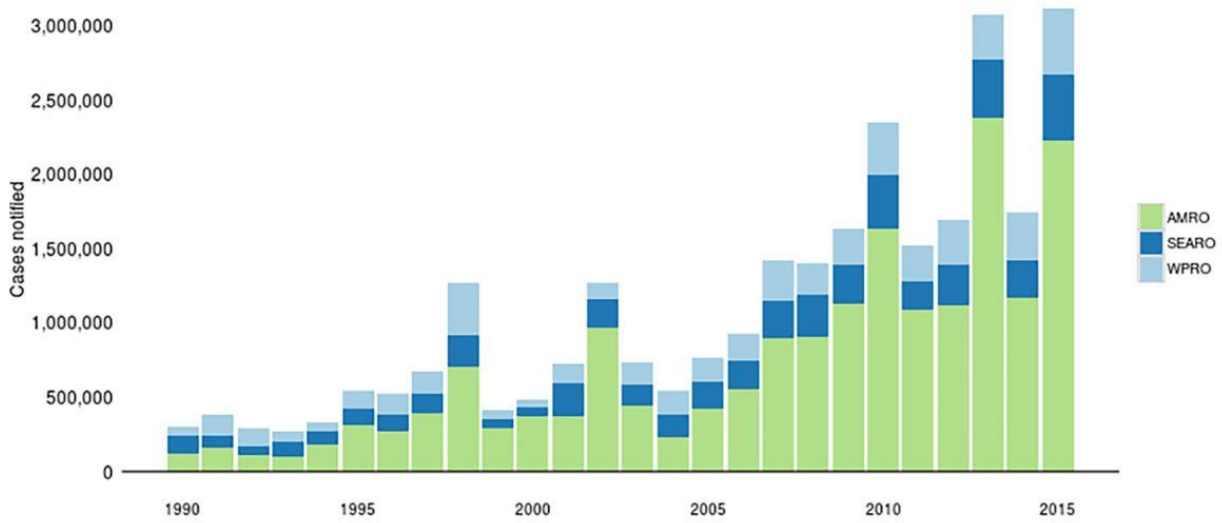


Figure 2 : Decision making for our data

For NDVI across the several sites, the data appears to be quite comparable with barely changed values. We took the average to prevent further noise and multicolinearity.
Then we will search for which precipitation source to employ. It is an essential climatic variable, as mosquitoes of interest need water in wet circumstances and have higher survival.

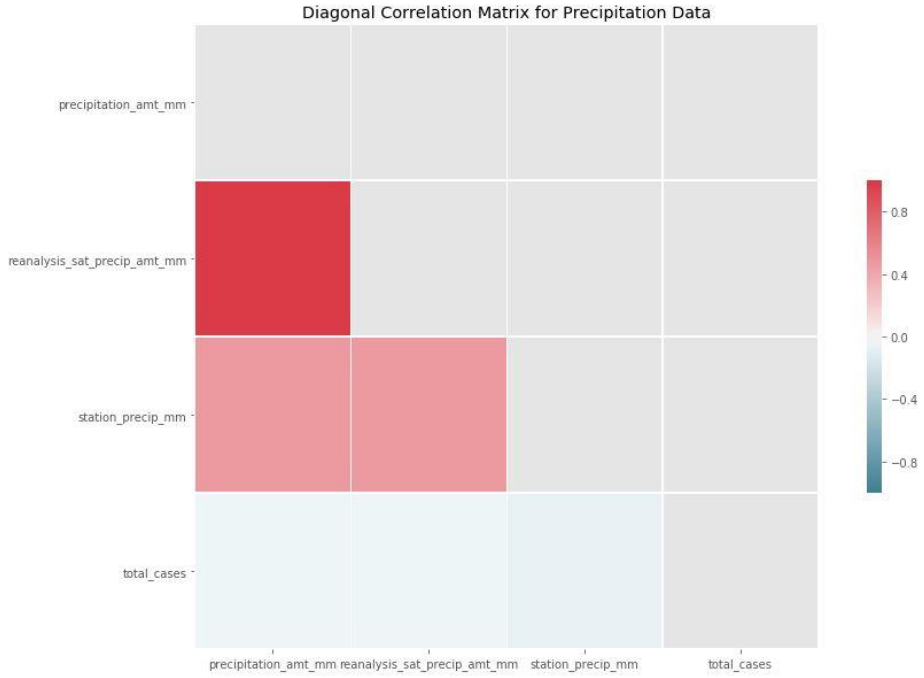


Figure 3 : Diagonal Correlation Matrix for Preparation Data

The variables precipitation amt mm and reanalysis sat precip amt mm are the same, such that one of them is used. We obtain quite different findings from station precip mm, but we are going to droPlacement Variable

since reanalysis appears to yield somewhat better correlational values. There were also more missing values in station data, which warrants your decision.

Another important point is that the association of our total cases with precipitation increases with the shifting values which we are going to investigate later.
Since we believe it is crucial for various environmental circumstances, we chose to develop some of our own variables to help enhance the model further.
The temperature range was an essential feature we wanted to incorporate in our model. However, we may predict an opposite connection in this temperature range. We want smaller temperature ranges with bigger numbers to show that mosquitoes are more likely to survive. We've chosen to construct a variable which divides 100 into kelvins by the temperature range. This temp r we'll call.
TempR=100ReanalysisTdtrK
TempR=100ReanalysisTdtrK
We'll additionally collect another component we'll call air dew. This aims to capture the air temperature variation from the droplet temperature. More gout equals more water that may be beneficial for mosquitoes.
AirDew=100|ReanalysisDewPointTempK−ReanalysisAirTempK|AirDew=100|ReanalysisDewPointTempK−ReanalysisAirTempK|
We opted to divide data amongst towns since we assume that the environmental circumstances will be different. The data collection also appears to last for several years. In contrast, modelling them may be beneficial in figuring out how various values in the variable at various places may impact dengue instances.

Exploratory Data Analaysis

We will first do a single analysis of the dengue virus total cases.

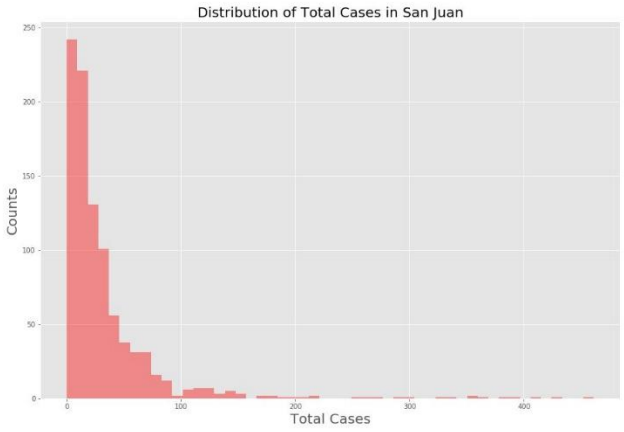


Figure 4 : Distribution of total case in San Juan

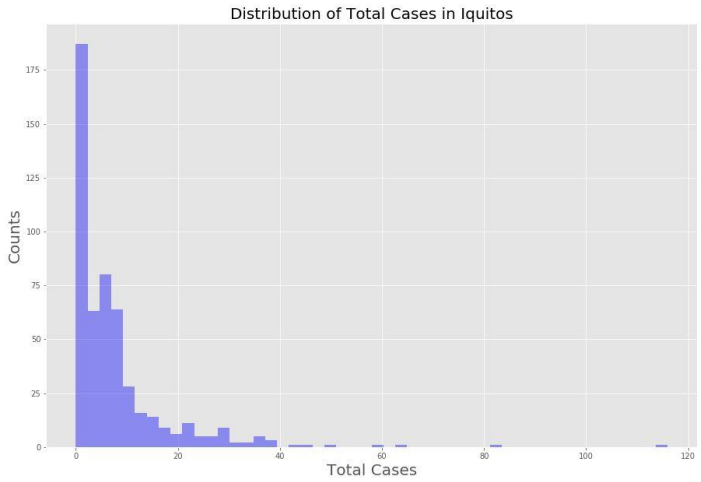


Figure 4 : Distribution of total case in Iquitos

The histograms for the town of San Juan and Iquitos are shown below. The most evident distinction between the two is the range. San Juan has recorded total cases for several weeks at or above 400. The maximum value of Iquitos is not more than 120 instances in total. This demonstrates that in overall, outbreaks in San Juan were considerably more serious than in Iquitos. The change may be linked to the pattern in the seasonal climate with the dengue virus. We also find similar forms to their distributions, with a higher number of low level observations sloping away when we increase the number of cases. We will examine a number of our case data summary statistics next.

	precipitation_amt_mm	reanalysis_sat_precip_amt_mm	station_precip_mm	total_cases
precipitation_amt_mm	1.000000	1.000000	0.482086	-0.042134
reanalysis_sat_precip_amt_mm	1.000000	1.000000	0.482086	-0.042134
station_precip_mm	0.482086	0.482086	1.000000	-0.074663
total_cases	-0.042134	-0.042134	-0.074663	1.000000

Figure 5 : Distribution of the training data

The histograms for the town of San Juan and Iquitos are shown below. The most evident distinction between the two is the range. San Juan has recorded total cases for several weeks at or above 400. The maximum value of Iquitos is not more than 120 instances in total. This demonstrates that in overall, outbreaks in San Juan were considerably more serious than in Iquitos. The change may be linked to the pattern in the seasonal climate with the dengue virus. We also find similar forms to their distributions, with a higher number of low level observations sloping away when we increase the number of cases. We will examine a number of our case data summary statistics next.

The mean and variances are significantly too far, which indicates that we cannot apply our model with a regre ssion of fish. We will instead use a negative binomial regression, if the mean and variance are not equal. Let's look a little more at the data.

Weeks in San Juan with null cases: 4

Number of weeks in Iquitos with zero cases: 96

We see that there is a large number of weeks where there are no cases of dengue virus in Iquitos. We may need to investigate why this is the case.

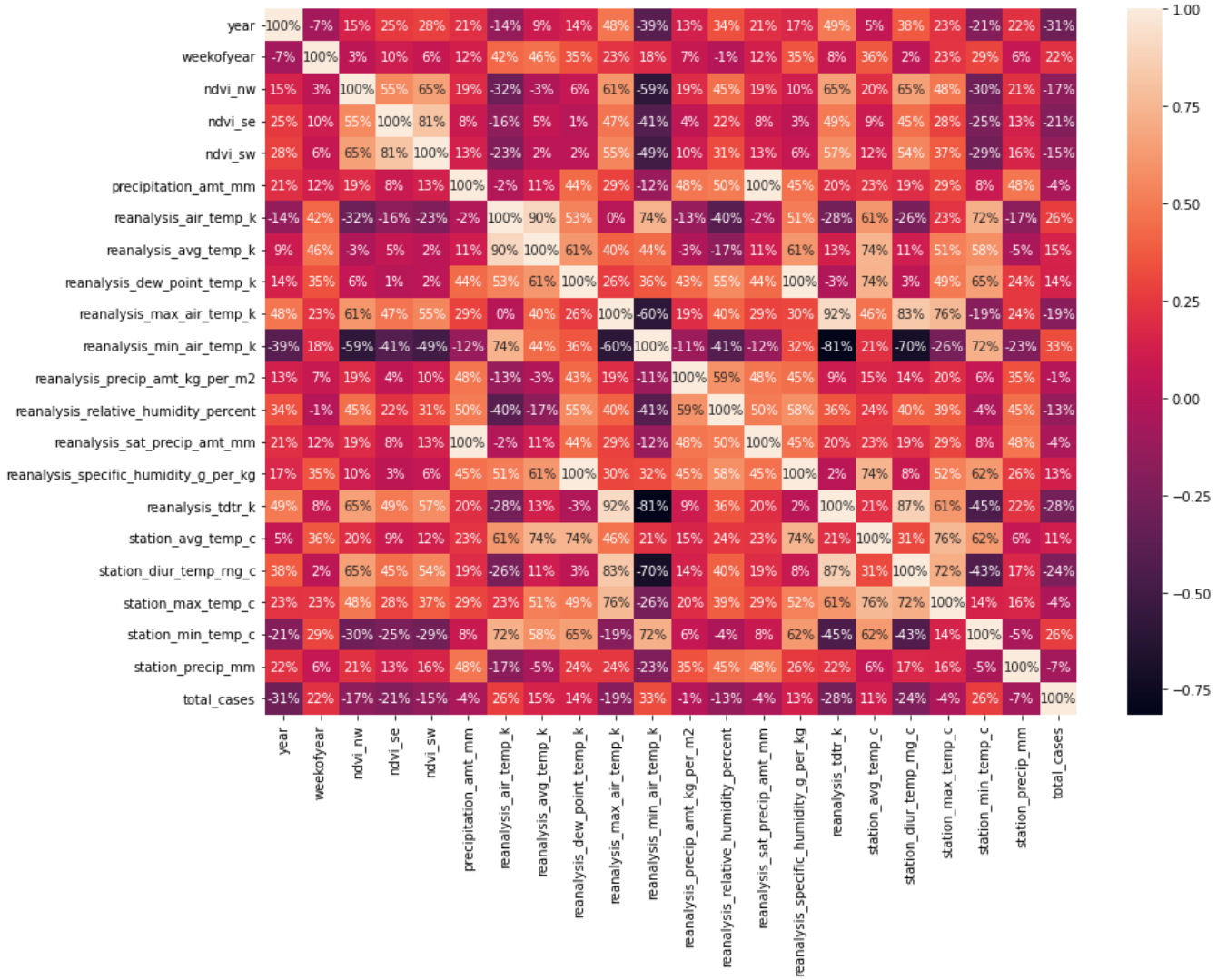


Figure 6 : Correlation of training data (SJ)

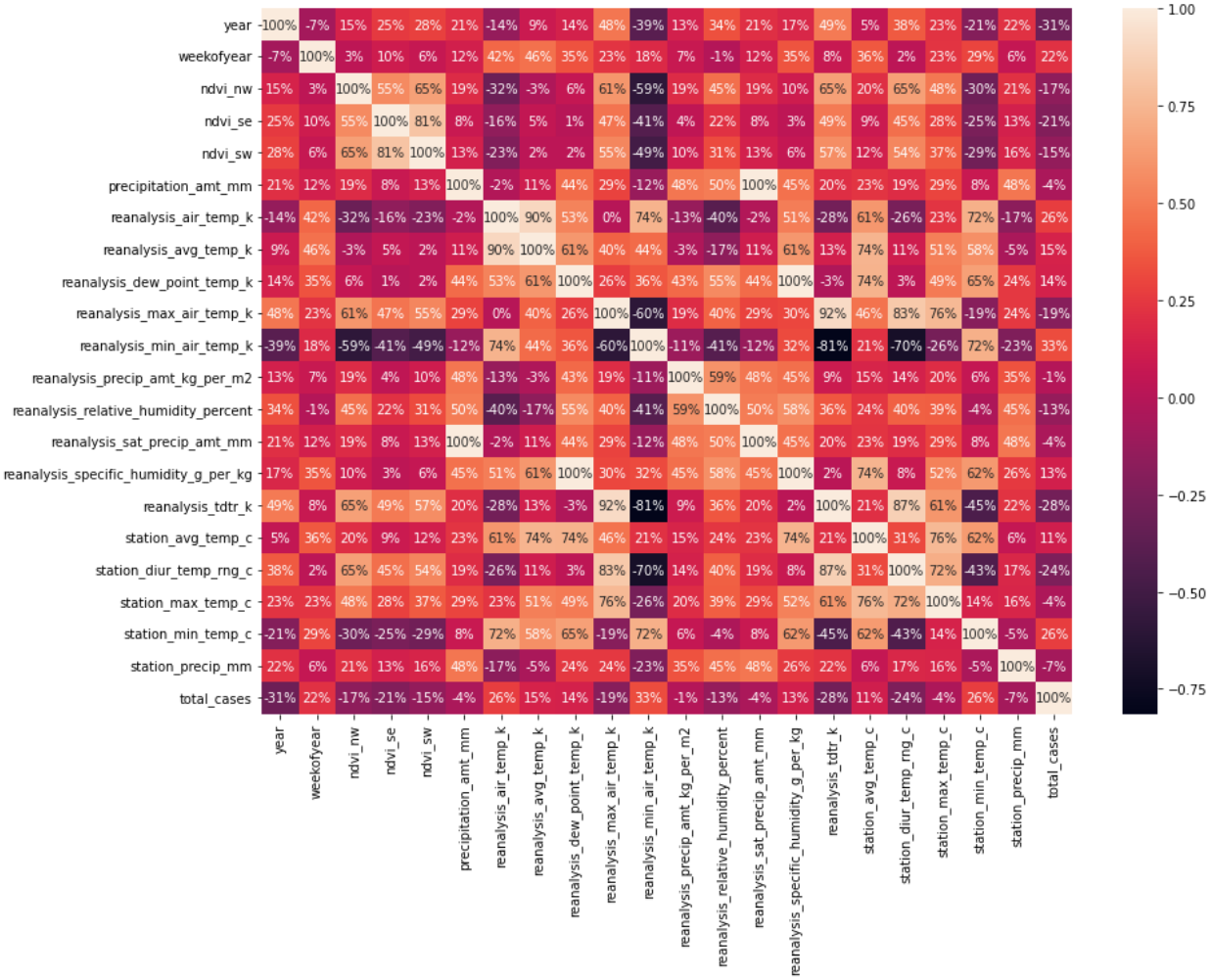


Figure 7 : Correlation of training data (IQ)

It appears that all variables are slightly linked, but in no way. The correlation matrix reveals that the total cases shifted improves Iquitos values. In comparison with the different city examples, we can observe that after 3-4 weeks of shifts, this offers more coherent findings for San Juan and Iquitos correlation signals. This demonstrates that some of the time delays in the reported cases might be attributed to shifts.

Many of the factors are the same. We chose to remove the variables from the station and leave the rest from the same satellite source in an attempt to remain consistent.

None of the variables are strongly linked to the total cases' Many variables reflect same values but distinct satellite or station sources. From this bar diagram, we can observe that in some situations, the station has somewhat better correlation values than the refreshment temperature. In Iquitos, the min temperature seems to be the variable most linked. This may be because Iquitos is cooler than San Juan. The Dengue virus may not benefit from such chilly temperatures.

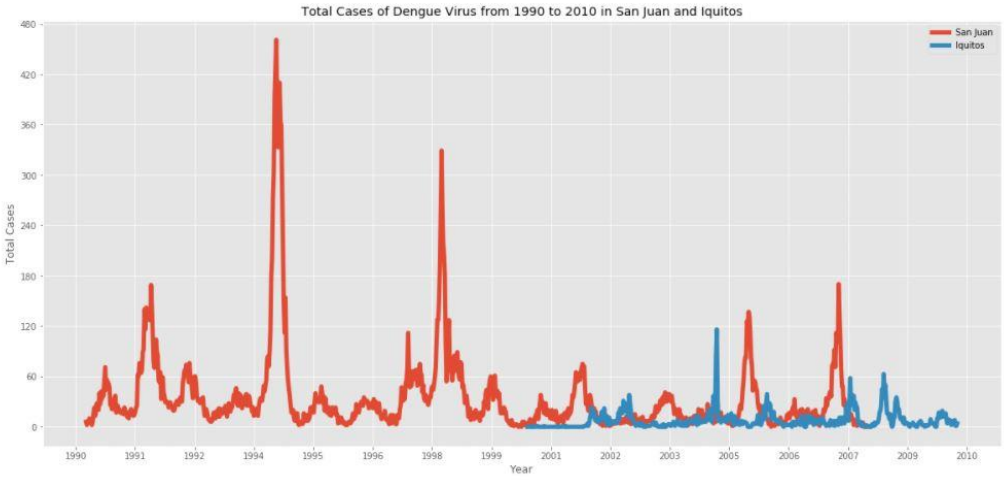


Figure 8 : Total Cases of Dengu virus from 1990 to 2010 (SJ)

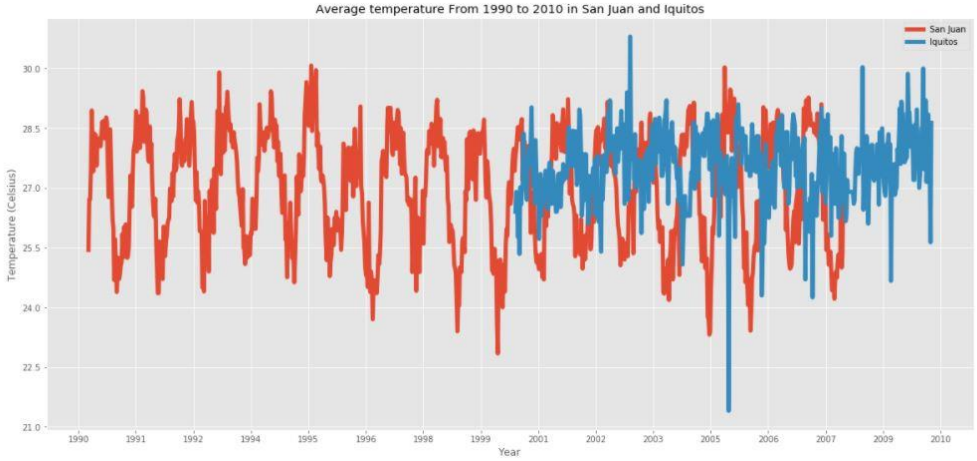


Figure 9 : Total Cases of Dengu virus from 1990 to 2010 (IQ)

We observe that the variations of the virus, when they occur during specific periods of the year, have a seasonal influence. In cases of dengue virus, however, there are certain peaks. In 1994 and 1998, during the period of dengue epidemics this appears to be the most significant. In both towns temporal fluctuations do not appear to be constant, which may be beneficial for the dengue virus infections between Iquitos and San Juan owing to differing climatic circumstances. The total cases in Iquitos appear to be less patterned, with seasonality being unpredictable. In addition to San Juan, we observe that the Iquitos temperature is significantly brighter.

As city, year, station dio temp ng c, reanalysis tdtr k we will move certain variables of little relevance since they are categorical or not linked with total case. The shifting total instances other than week 4 will also be removed: reanalysis avg temp k, reanalysis relative humidity percent and eanalysis precip amt kg per m2 since we chose to utilize the medium mm midrange temperature, specific humidity and rainfall to improve correlation values. No repeating variables are desired.

Feature Engineering and Selection

We will now try to select the data functionality to be used.
One way is just to look at the value of total cases and utilize them for our characteristics. You may accomplish this using a linear kernel with the recursive function selection.

```
sj feature ranks
1 ndvi_nw
1 ndvi_se
1 ndvi_sw
13 precipitation_amt_mm
1 reanalysis_air_temp_k
10 reanalysis_avg_temp_k
11 reanalysis_dew_point_temp_k
5 reanalysis_max_air_temp_k
4 reanalysis_min_air_temp_k
12 reanalysis_precip_amt_kg_per_m2
3 reanalysis_relative_humidity_percent
15 reanalysis_sat_precip_amt_mm
2 reanalysis_specific_humidity_g_per_kg
1 reanalysis_tdtr_k
9 station_avg_temp_c
7 station_diur_temp_rng_c
8 station_max_temp_c
6 station_min_temp_c
14 station_precip_mm
```

```
Ranking of iq
1 ndvi_nw
1 ndvi_se
1 ndvi_sw
13 precipitation_amt_mm
1 reanalysis_air_temp_k
10 reanalysis_avg_temp_k
11 reanalysis_dew_point_temp_k
5 reanalysis_max_air_temp_k
4 reanalysis_min_air_temp_k
12 reanalysis_precip_amt_kg_per_m2
3 reanalysis_relative_humidity_percent
15 reanalysis_sat_precip_amt_mm
2 reanalysis_specific_humidity_g_per_kg
1 reanalysis_tdtr_k
9 station_avg_temp_c
7 station_diur_temp_rng_c
8 station_max_temp_c
6 station_min_temp_c
14 station_precip_mm
```

Figure 10 : Feature Ranking for SJ and IQ

In fact, in our model predictions we have tried RFE before. When using them, the outcomes were unsatisfactory, therefore we relied our feature choices on the study, instead of using a feature selection method. We are not going to include the efforts we have made for RFE in our report, given the long-term training of the ML models.

We will try to test the performance of the models in some characteristics we believe are essential. We picked time measurements for stations because we think measurements for stations are more precise than satellite measures like temperature. We discovered that a satellite would be preferable for other information like precipit. Our prior study findings have shown that for the survival of dengue virus vectors the following characteristics are important:

```
Out[31]: ['reanalysis_dew_point_temp_k',
          'reanalysis_precip_amt_kg_per_m2',
          'reanalysis_specific_humidity_g_per_kg',
          'station_avg_temp_c',
          'station_max_temp_c',
          'station_min_temp_c']
```

Figure 11 : Climate Columns

We will try to test the performance of the models in some characteristics we believe are essential. We picked time measurements for stations because we think measurements for stations are more precise than satellite measures like temperature. We discovered that a satellite would be preferable for other information like pre cipit. Our prior study findings have shown that for the survival of dengue virus vectors the following character istics are important:

```
rolling_column
Out[33]: ['roll_mean_reanalysis_dew_point_temp_k',
          'roll_sum_reanalysis_precip_amt_kg_per_m2',
          'roll_mean_reanalysis_specific_humidity_g_per_kg',
          'roll_mean_station_avg_temp_c',
          'roll_mean_station_max_temp_c',
          'roll_mean_station_min_temp_c']
```

Figure 12 : Rolling Columns

We tried normalization but it seemed to make our predictions worse

5.3 Experimental Result

Modeling

We choose to employ a number of modelling approaches. We search for grid to identify the optimum parameters, based on the average absolute error score of the various models. We can also compare the adjustment of our baseline to the gridsearch results that show us how much better they do than the baseline. We can also obtain a fairly clear overview of how the different models perform on the baseline tests versus each other. Most of this portion is code. By adjusting them I have tried various parameters and made the steps smaller as the results have shown which values are better for better values.

XGBoost Baseline

SJ

```
R2 score of XGBRegressor : 0.8587512645367884
Training r2_score is : 99.98940314317552
Testing r2_score is : 85.87512645367885
Mean Absolute Eror :  8.867901079384376
Mean Squared Eror :  304.49878117959474
Root Mean Squared Eror :  17.44989344321606
```

IQ

```
R2 score of XGBRegressor:0.8371963914419132
Training r2_score is : 99.98834992208987
Testing r2_score is : 83.71963914419132
Mean Absolute Eror :  9.583463905208683
Mean Squared Eror :  320.12845186121694
Root Mean Squared Eror :  17.892133798438266
```

Random Forest Baseline

SJ

```
R2 score of RandomForestRegressor() is : 0.8537057405322466
Training r2_score is : 97.4346473609381
Testing r2_score is : 85.37057405322466
Mean Absolute Eror :  9.234691075514874
Mean Squared Eror :  315.3757345537757
Root Mean Squared Eror :  17.758821316567598
```

IQ:

```
R2 score of RandomForestRegressor() is:0.81103355631689
Training r2_score is : 97.6237846292954
Testing r2_score is : 81.103355631689
Mean Absolute Eror :  10.24837528604119
Mean Squared Eror :  371.57367459954236
Root Mean Squared Eror :  19.276246382518107
```

Gradient Boost Baseline

SJ:

```
R2 score of GradientBoostingRegressor() is :
0.8087750523658406
Training r2_score is : 94.07588970374492
Testing r2_score is : 80.87750523658406
Mean Absolute Eror :  11.040223824784654
Mean Squared Eror :  412.2356444097077
Root Mean Squared Eror :  20.303586983824008
```

IQ:

```
R2 score of GradientBoostingRegressor() is : 0.7823327063244246
Training r2_score is : 94.140687048106
Testing r2_score is : 78.23327063244247
Mean Absolute Eror :  11.32870672721775
Mean Squared Eror :  428.0095162652436
Root Mean Squared Eror :  20.688390857320044
```

ExtraTreesRegressor

SJ :

```
R2 score of ExtraTreesRegressor() is :
0.9169168082985576
Training r2_score is : 100.0
Testing r2_score is : 91.69168082985576
Mean Absolute Eror :  7.450572082379863
Mean Squared Eror :  179.10766086956522
Root Mean Squared Eror :  13.383111031055718
```

IQ:

```
R2 score of ExtraTreesRegressor() is : 0.8869663662146222
Training r2_score is : 100.0
```

```
Testing r2_score is : 88.69663662146222
Mean Absolute Eror :  8.254988558352403
Mean Squared Eror :  222.2633915331808
Root Mean Squared Eror :  14.908500646717657
```

From all baseline models, we can observe that in terms of the basic model, the gradient boost model works well. Now we create our Gridsearch settings to automatically adjust the optimum parameters. Then we will again check for the greatest performance. As indicated in the lecture, we chose to make a 10-split KFold for our GridSearchCV. In order to discover which one fared better on average I completed many hundred baseline tests. The gradient boost and xgboost seem to be the most efficient.

Side note: The results for models were selected since several predicted negative values. We merely choose to cut 0.

XGB Gridsearch

```
MAE: 8.867901079384376
-8.115957044953687
{'gamma': 0.2, 'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 500}
MAE: 9.014476234566429
MAE: 9.583463905208683
-8.474947780817912
{'n_estimators': 50}
MAE: 9.550864353094003
```

Random Forest Gridsearch

```
MAE: 9.325171624713958
-8.95664492331586
{'n_estimators': 25}
MAE: 8.906819221967964
MAE: 10.071601830663615
-8.943363521646281
{'n_estimators': 200}
MAE: 9.890068649885583
```

Gradient Boost Gridsearch

```
MAE: 11.037702193030869
-8.549663660088601
{'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 100}
MAE: 9.165402811422988
MAE: 11.337963491626915
-10.037931603404491
{'n_estimators': 250}
MAE: 10.646030145373816
```

ExtraTree Gridsearch

```
MAE: 7.330251716247139
-7.813404193360512
{}
MAE: 7.505652173913044
MAE: 7.256910755148741
-7.533114540865851
```

```
{}
MAE: 7.30812356979405
```

We definitely do better than our basic models on average in all situations except ExtraTree. From our testing, we discovered that our MAE was about 7-9 SJ and 7-9 IQ. ExtraTree looks awfully at the other regressors.

5.4 Validation and Visualization

Time Plots

XGB Predictions Overtime for SJ and IQ

The forecasts probably overestimate since we cannot appropriately adjust the parameters to avoid overfitting. We may try to increase our score later on by altering our settings. It does in most of our testing pretty much like the Random Forest. The forecasts also seemed very much the same as in the next figure.

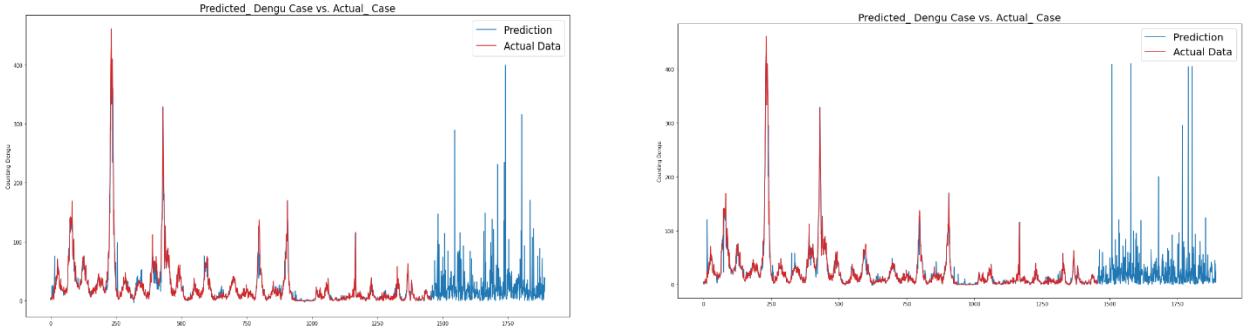


Figure 13 : XGB Predictions Overtime for SJ and IQ

Random Forest Pedcitions for SJ and IQ

Random forest was the second best for us in our submissions, out of the average of four algorithms at about 23.5 and 26. The diagrams appear far more plausible, following a pattern in which peaks and lows are present.

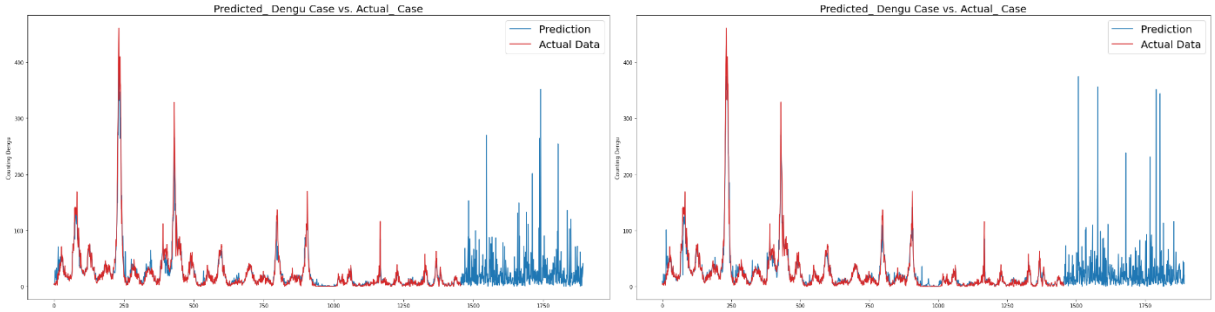


Figure 14 : Random Forest Pedcitions for SJ and IQ

Gradient Boost Pedcitions for SJ and IQ

At the time of writing, the gradient boost seemed to work best. We can't notice overfitting as with the other models. The number of errors, however, is considerably different from that of Iquitos. We may also try to re-send normalized values of functions. It is also seemed extremely noisy. At the time we wrote it the best score was 23.19.

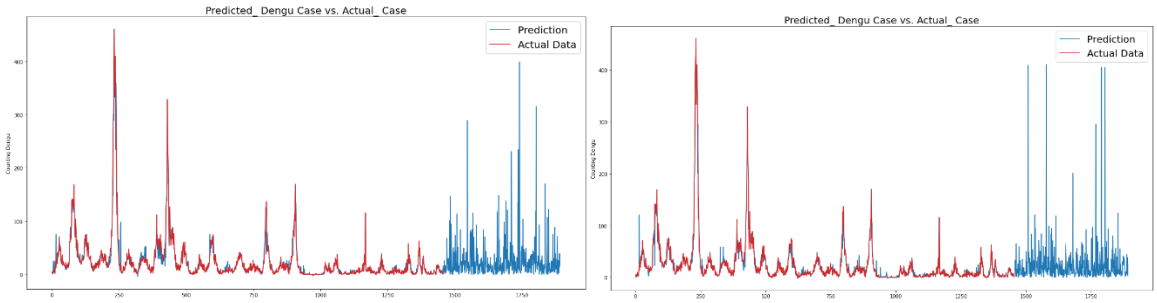


Figure 15 : **Gradient Boost Pedcitions for SJ and IQ**

Extra Tree

At the time of writing, the gradient boost seemed to work best. We can't notice overfitting as with the other models. The number of errors, however, is considerably different from that of Iquitos. We may also try to re-send normalized values of functions. It is also seemed extremely noisy. At the time we wrote it the best score was 23.19.

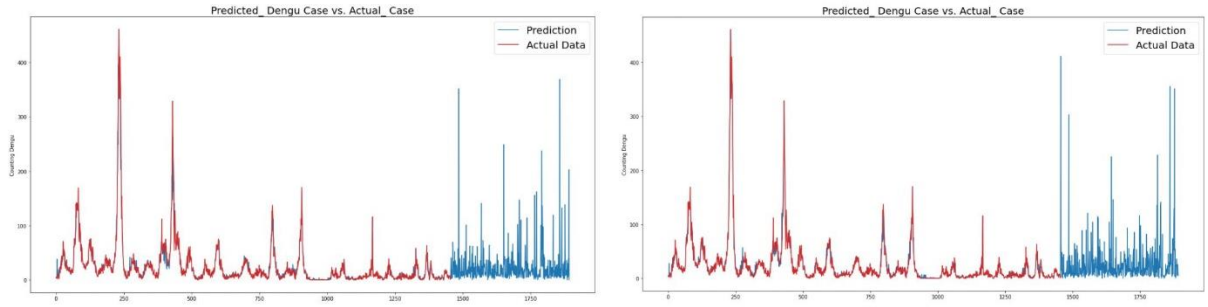


Figure 16 : Extra Tree Pedcitions for SJ and IQ

As shown by the time series analysis of our data, None of our predictions were overfitting the data.

Scatter Plots of Train Predicted vs Actual Total Cases

We have developed scatterplots to assess the predictions of our models versus actual values. These diagrams substantially represent the MAE and temporal trends we find. ExtraTree all worked well

with most real and predictive effects. XGBoost is Random Forest, Gradient Boost and ExtraTree. However, ExtraTree performs well among all models.

XGBoost scatter plot:

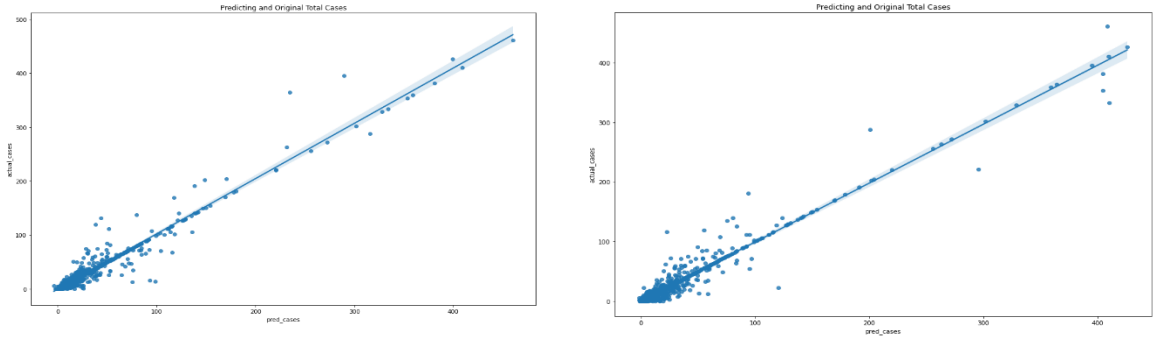


Figure 17 : XGBoost scatter plot for SJ and IQ

Gradient descent scatter plot :

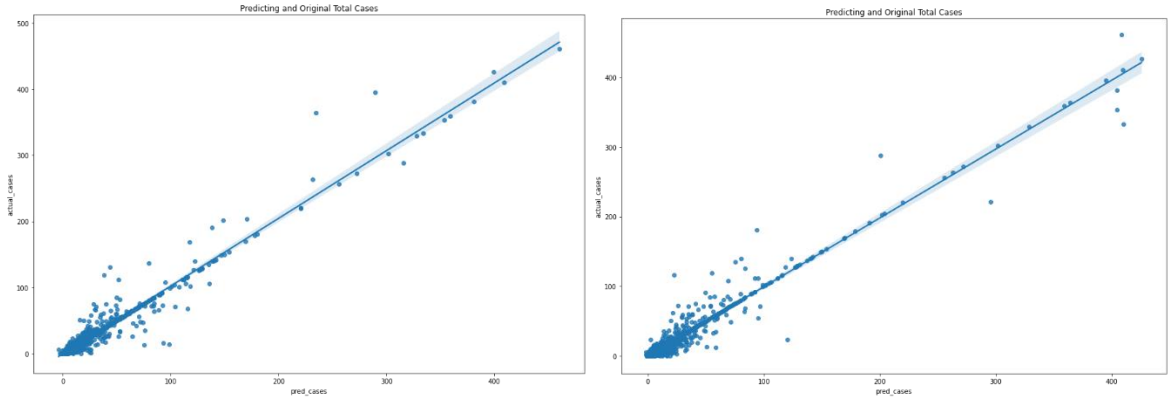


Figure 18 : Gradient descent scatter plot for SJ and IQ

Random Forest scatter plot:

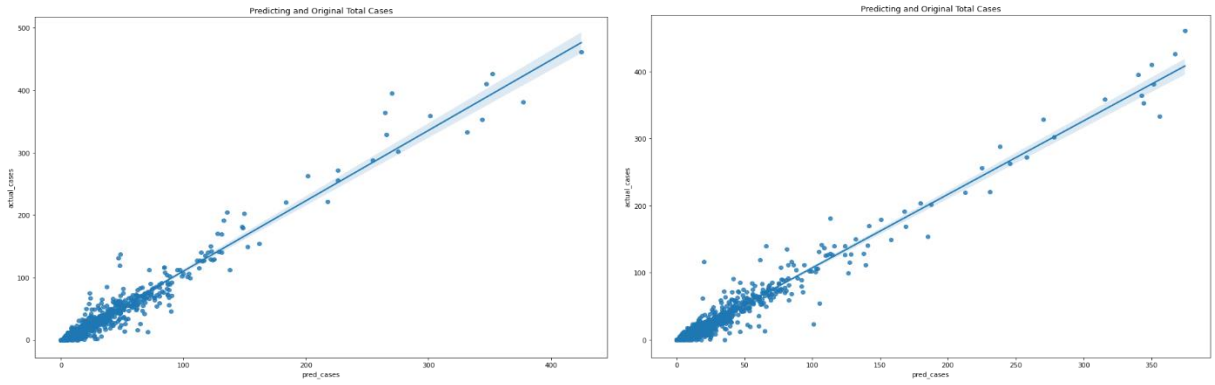


Figure 19 : Random Forest scatter plot for SJ and IQ

Extra Tree scatter plot :

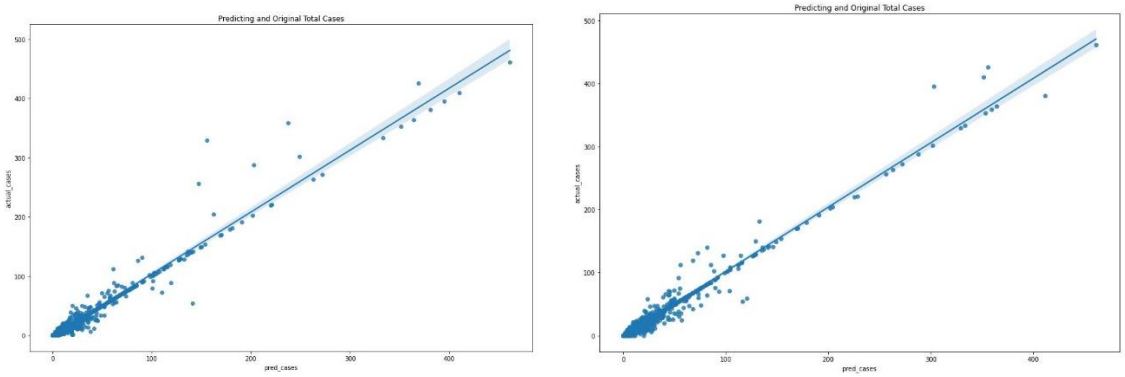


Figure 20 : Extra Tree scatter plot for SJ and IQ

6. Conclusion

The most satisfactory, troubling human therapeutic data is rapidly, since every normal data should be identified and broken down. Isolating useful information, detecting patterns and categorizing data entails making sensible choices for examining and treating an illness in which the database is essentially undiscovered. Machine learning processes will lead to a response to this problem. A notable goal is to use machine learning algorithms to make accurate judgments in medical and health applications . Machine learning here plays a key part by using different algorithms to provide results faster and precisely. Computer methods like this are fascinating; since they can be deployed cheaply and pragmatically by using individual devices such as portable workstations, mobile phones or tablets; and factual models and machine learning can forecast them once created in seconds. Based on an experimental examination of the DecisionTree, ExtraTree, RandomForest,XGBoost and other algorithms, the Simple CART results are superior for all performance measurement metricities such as r-2, MAE, MSE and RMSE.

7. Future Work

As future research, we will hunt for new data sets related to dengue characteristics and start to use real-world machine training and profound learning algorithms to develop a good model that can help health professionals spread the development of the illness. We also want to use the "Coronavirus" sickness to machine learning and deep learning algorithms and to develop strategies to halt this disease.

8. Reference

[1] Dave Kaveri Atulbhai and Shilpa Serasiya, “A Survey: Prediction & Detection of Dengue – Mining Methods & Techniques, IJARIIE–ISSN (O)-2395-4396, Vol-3, Issue-2, 2O17.

[2] Xavier-CarvalhoC, Cezar RDDS, Freire NM, Vasconcelos CMM, Solorzano VEF, de Toledo-Pinto TG, Fialho LG, do Carmo RF, Vasconcelos LRS, Cordeiro MT,

[3] Dr.Arun Kumar.P.M, Chitra Devi.B, Karthick.P, Ganesan.M and Madhan.A.S, “Dengue Disease Prediction Using Decision Tree and Support Vector Machine”, www.internationaljournalssrg.org, ISSN: 2348 – 8387, 2017, pp: 60-63.

[4] Rohit Morlawar, V.A. Kothiwale, “A study of clinical profile in different serological diagnostic parameters of dengue fever”, 2017, Volume: 10, Issue: 2, pp: 178-182

[5] William Caicedo-Torres, Angel Paternina, and Hernando Pinzon, “Machine Learning Models for Early Dengue Severity Prediction”, @Springer International Publishing AG 2016

[6] M. Montes-y-Gomez et al. (Eds.): IBERAMIA 2016, LNAI 10022, pp. 247–258, 2016, DOI: 10.1007/978-3-319-47955-2 21 [7] M.Bhavani and S.Vinod kumar, “A Data Mining Approach For Precise Diagnosis of Dengue Fever”, International Journal of Latest Trends in Engineering and Technology, Vol

(7), Issue(4), pp.352- 359, nov 2016, DOI: http://dx.doi.org/10.21172/1.74.048

[8] Purushottam Kumar, Rajendra T. Ankushe, Bina M. Kuril1, Mohan K. Doibale, Syed J. Hashmi, Sandeep B. Pund, “An Epidemiological Study of Fever Outbreak in Aurangabad, Maharashtra, India”, International Journal of Community Medicine and Public Health Kumar P et al. Int J Community Med Public Health. 2016 May; 3(5), pp: 1107-1111

[9] Phong D. Tong, Vinh D. Le, Hieu N. Duong, Hien T. Nguyen, Vaclav Snasel, “Decision trees for diagnosis of dengue fever”, International Conference on Information and Convergence Technology for Smart Society, Jan 2016, 19-21, in Ho Chi Minh, Vietnam

[10] Danilo Bretas de Oliveira, Guilherme Machado, Gabriel Magno de Freitas Almeida, Paulo Cesar Peregrino Ferreira, Claudio Antonio Bonjardim, Giliane de Souza Trindade, Jonatas Santos Abrahao and Erna Geessien Kroon, “Infection of the central nervous system with dengue virus 3 genotype I causing neurological manifestations in Brazil”, Revista da Sociedade Brasileira de Medicina Tropical 49(1):125-129, Jan-Feb, 2016, http://dx.doi.org/10.1590/0037- 8682-0208-2015

[11] Kashish Ara Shakil, Shadma Anis and Mansaf Alam, “Dengue Disease Prediction using Weka Data Mining Tool”, 18 Feb 2015.

[12] Kamran Shaukat, Nayyer Masood, Sundas Mehreen and Ulya Azmeen, “Dengue Fever Prediction: A Data Mining Problem”, Dar et al., J Data Mining Genomics Proteomics 2015, 6:3 http://dx.doi.org/10.4172/2153-0602.1000181

[13] P.H.M.Nishanthi, Herath, A.A.I., Perera, H.P.Wijekoon, “Prediction of Dengue Outbreaks in Sri Lanka using Artificial Neural Networks”, International Journal of Computer Applications (0975 – 8887) Volume 101– No.15, September 2014

[14] N K Kameswara Rao, Dr. G P Saradhi Varma, Dr.M.Nagabhushana Rao, “Classification Rules Using Decision Tree for Dengue Disease”, International Journal of Research in Computer and Communication Technology, Vol 3, Issue 3, March- 2014.

[15] Bhaswati Bandyopadhyay, Indrani Bhattacharyya, Jayshree Konar, Srima Adhikary, Nidhi Dawar, Jayeeta Sarkar,1 SaiantaniMondal, Mayank Singh Chauhan, Nemai Bhattacharya, Asit Biswas, Anita Chakravarty, and Bibhuti Saha, “A Comprehensive Study on the 2012 Dengue Fever Outbreak in Kolkata, India”, Hindawi Publishing Corporation, ISRN Virology, Article ID 207580, Volume 2013, 5.pages, http://dx.doi.org/10.5402/2013/207580.

[16] Ragini Singh, S. P. Singh, Niaz Ahmad. “A study of clinical and laboratory profile of dengue fever in a tertiary care centre of Uttarakhand, India”, International Journal of Research in Medical Sciences , Singh R et al. Int J Res Med Sci., 2014 Feb; 2(1), pp: 160- 163

[17] Anna L Buczak, Phillip T Koshute, Brian H Feighner, Steven M Babin and Sheryl H Lewis, “A data-driven epidemiological prediction method for dengue outbreak using local and remote sensing data”, Buczak et al. BMC Medical Informatics and Decision Making 2012, 12:124, http://www.biomedcentral.com/1472-6947/12/124

[18] Shih-Jie lo, Shih_Chun Yang, Da_Jeng Yao, Jianm_Hwa Chen, Chao_Min Cheng. “Molecular-Level Dengue Fever Diagnostics”, December 2012, IEEE nanotechnology magazine, pp: 26-30.

[19] Daranee Thitiprayoonwongse, Pratap Suriyaphol and Nuanwan Soonthornphisaj. “Data Mining of Dengue Infection Using Decision Tree”, ISBN: 978-1-61804-092-3, 2012, pp: 154-159.

[20] Ashwini Kumar, Chythra R Rao, Vinay Pandit, Seema Shetty, Chanaveerappa Bammigatti, Charmaine Minoli Samarasinghe, “Clinical Manifestations and Trend of Dengue Cases Admitted in a Tertiary Care Hospital, Udupi District, Karnataka”, Indian Journal of Community Medicine, Issue 3, Vol 35, July 2010, pp: 386-390.

[21] Paul V. Effler, Lorrin Pang, Paul Kitsutani, Vance Vorndam, Michele Nakata, Tracy Ayers, Joe Elm, Tammy Tom, Paul Reiter, José G. Rigau-Perez, John M. Hayes, Kristin Mills, Mike Napier, Gary G. Clark, and Duane J. Gubler. “Dengue Fever, Hawaii, 2001–2002”, Emerging Infectious Diseases, www.cdc.gov/eid, Vol. 11, No. 5, May 2005, pp: 742-749.

[22] Fatimah Ibrahim, Mohd Nasir Taib, Senior Member, IEEE, Wan Abu Bakar Wan Abas, Chan Chong Guan, and Saadiah Sulaiman. 218 International Journal of Engineering &Technology “A Novel Approach to Classify Risk in Dengue Hemorrhagic Fever (DHF) Using Bioelectrical Impedance Analysis (BIA)”, IEEE Transactions on Instrumentation and Measurement, VOL. 54, NO.1, Feb 2005, pp: 237-244.

[23] Fatimah Ibrahim, Mohd Nasir Taib, Senior Member, IEEE, Wan Abu Bakar Wan Abas, Chan Chong Guan, and Saadiah Sulaiman, “A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN)”, Computer Methods and Programs in Biomedicine (2005) 79, pp: 273-281.

[24] Ole Wichmann, Annekathrin Lauschke, Christina Frank, Pei-Yun Shu, Matthias Niedrig, Jyh-Hsiung Huang, Klaus Stark, and Tomas Jelinek, “Dengue Antibody Prevalence in German Travelers”, Emerging Infectious Diseases www.cdc.gov/eid, Vol. 11, No.5, May 2005, pp: 762-766.

[25] Nivedita Gupta, Sakshi Srivastava, Amita Jain & Umesh C. Chaturvedi, “Dengue in India”, Indian J Med Res 136, September 2012, pp: 373-390 [26] http://www.denguevirusnet.com/aedes-aegypti.html [27] J. Jain, S.K. Dubey, J. Shrinet, S. Sunil, “Dengue Chikungunya coinfection: A live-in relationship”, Biochemical and Biophysical Research Communications, [28] Rasana Patil, Tina Makhija, H. P. Suryawanshi, S.P.Pawar.A Review on – Dengue.Research J. Pharm. and Tech. 6(9): September 2013.