Project Report

Group 08

Sachini Weerasekara

Lai Peng

857-701-8317 (Sachini Weerasekara)

857-302-8247 (Lai Peng)

Weerasekara.s@husky.neu.edu

Peng.la@husky.neu.edu

Percentage of Effort Contributed by Sachini Weerasekara : 50%

Percentage of Effort Contributed by Lai Peng : 50%

Submission Date : 04/10/2020

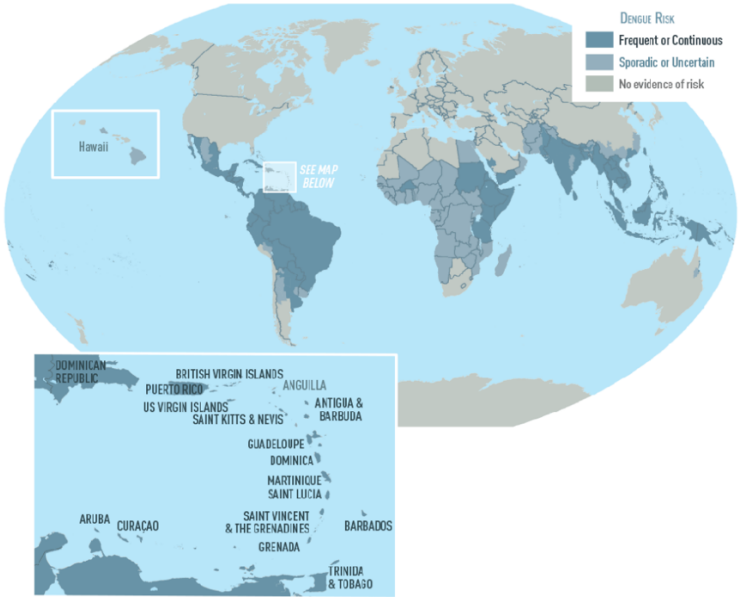
Contents

1. Problem Setting

Dengue fever is a class of mosquito-borne viruses which spreads to human through the bite of an infected Aedes species mosquito. This is common in more than 100 countries around the word and is mostly found in tropical and sub-tropical parts of the world. Forty percent of the world's population, about 3 billion people live in areas with a risk of dengue. Each year, up to 400 million people get infected with dengue. Approximately 100 million people get sick from infection, and 22,000 die from severe dengue. (CDC) In mild cases, symptoms are similar to the flu: fever, rash, and muscle and joint pain. In several cases, dengue fever can cause severe bleeding, low blood pressure, and even death.

Because it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation. Although the relationship to climate is complex, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide.

In recent years, dengue fever has been spreading. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands. These days many of the nearly half billion cases per year are occurring in Latin America:



Source: CDC Website

An understanding of the relationship between climate and dengue dynamics can improve research initiatives and resource allocation to help fight life-threatening pandemics.

2. Problem Definition

The dataset consists of climatic and non-climatic variables that are expected to be having some connection with the total number of Dengue cases in each week of year, for two different Latin American cities; San Juan and Iquitos. In addition, a new column was added with the class of total number of Dengue cases for each week of year. Hence, with the assumption that these associations differ between cities owing to their differences in geographical conditions, we further define the problem into four different subsets as follows.

Problem 1.1 : Classifying class of total number of Dengue cases for week of year for San Juan

Problem 1.2 : Prediction of total number of Dengue cases for week of year for San Juan

Problem 2.1 : Classifying class of total number of Dengue cases for week of year for Iquitos

Problem 2.2 : Prediction of total number of Dengue cases for week of year for Iquitos

In conclusion, based on climatic variables and non-climatic variables in our dataset, we aim to predict numbers of Dengue cases and class of Dengue cases for week of year for two cities San Juan and Iquitos.

Solutions to above problems are expected to assist the healthcare sector of a country to be better prepared to the Dengue epidemic to make sure necessary actions are taken at appropriate times. This would help in saving lives through resilient healthcare systems and also make favorable impacts to the economy of a country.

3. Data Sources

The source of the Dengue dataset is the website named Drivendata and specific website address is: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/

4. Data Description

There are two original datasets(Degue_Train, Dengue_Response) and then we merge two of them to one dataset called *Dengue*, which includes 25 columns and 1456 rows, out of which all predictors are numerical in nature. There are 520 rows for Iquitos, while there are 926 rows for San Juan. The 25 columns with different types are shown below:

City variables:

There are two city variables, which are sj and iq.

- Sj: sq represents San Juan
- Iq: iq means Iquitos

Date variable:

- week_start_date : using yyyy-mm-dd format to show dates

Daily climate data weather variables:

There are five variables, which are station_max_temp_c, station_min_temp_c, station_avg_temp_c, station_precip_mm and station_diur_temp_rng_c. These five variables are NOAA's GHCN daily climate data weather station measurements.

- 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

Precipitation variable (0.25x0.25 degree scale):

This variable is PERSIANN satellite precipitation measurements, which is 0.25x0.25 degree scale.

- precipitation_amt_mm : total precipitation

Climate Forecast System Reanalysis variables(0.5x0.5 degree scale):

There are ten variables, which are NOAA's NCEP Climate Forecast System Reanalysis measurements and their scale is 0.5x0.5 degree.

- 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

Satellite vegetation variables:

There are four variables, which are normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index measurements, and their scale are also 0.5x0.5 degree scale.

- 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

Case numbers variable:

- total_case: numbers of dengue cases

5. Data Exploration

5.1 Exploration of the response variable

Response variable is the total number of cases predicted for each city and following are the means and variances for total cases in each city.

	San Juan	Iquitos
Mean	34.21	2641.96
Variance	7.57	115.896

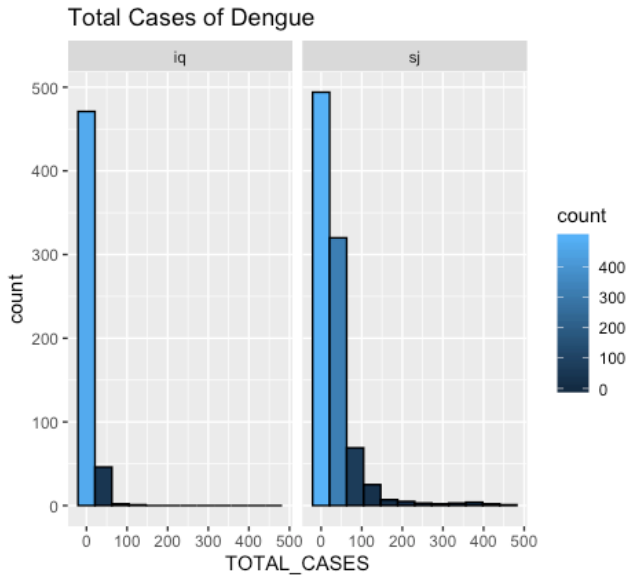


Fig. 1: Distribution of total cases for two cities

Histogram of the total number of cases in San Juan and Iquitos

Variance seems to be very large compared to the mean which depicts that the spread is wide and also, a skewed distribution is observed for total cases for both the cities.

Total cases is response variable, which's graph with week of year in San Juan in 2000 are shown below, and the explanation is as followed:

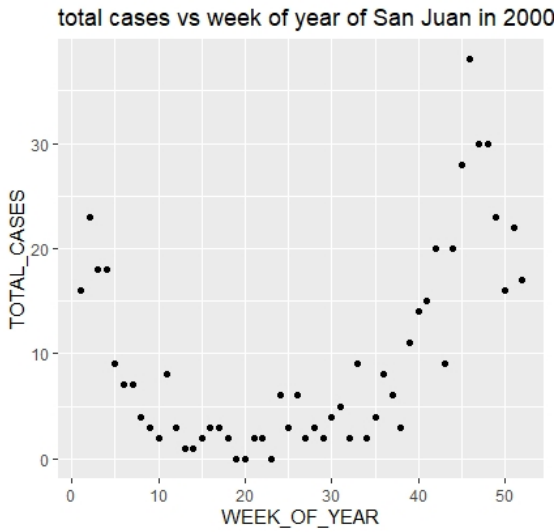


Fig 2: Total cases vs Week of year in San Juan for year 2000

From this scatterplot, we could get the following information:

1. It basically shows U-shaped association between these two variables---- weeks of year and total cases in San Juan in 2000;

2. From the first week to twentieth week, it presents negative association. The number of cases decreased from 1-20 weeks.

3. From the twentieth week to fifty-second week, it mainly shows positive association. The number of cases increased from 20-52 weeks.

4. This shows that the total number of cases are highly associated with the time of the year.

Total cases is response variable, which's graph with week of year in Iquitos in 2000 are shown below, and the explanation is as followed:

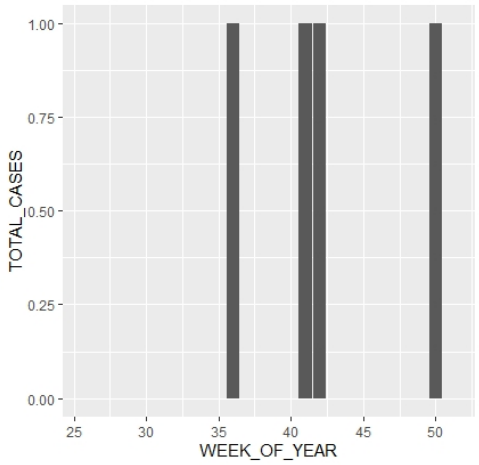


Fig 3: Total cases ratio vs weeks of year in Iquitos in 2000

From the bar chart, we could get the following information:

1. Only 36, 41, 42, 50 weeks existed only 1 case, there isn't any cases in other weeks in Iquitos in 2000.

2. It is almost no correlation between total case and week of year in Iquitos in 2000.

Total cases is response variable, which's graph with pixel southwest of city centroid in San Juan in 2000 is shown below, and the explanation is as followed:

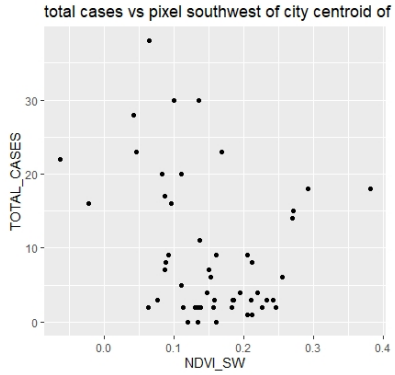


Fig 4: South West vegetation index vs total cases in San Juan for year 2000

From this scatterplot, we could the following information:

1. It is almost no correlation between total case and pixel southwest of city centroid of 2000 in San Juan.

2. There are relatively more infected cases concentrated on 0.05-0.25 of pixel southwest of city centroid in San Juan in 2000.

Total cases is response variable, which's graph with pixel northeast of city centroid in San Juan in 2000 is shown below, and the explanation is as followed:

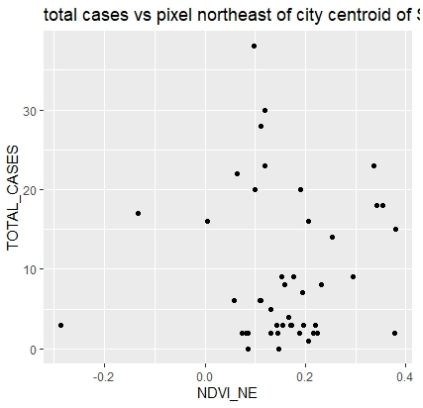


Fig 4: North East vegetation index vs total cases in San Juan for year 2000

From this scatterplot, we could the following information:

1. It is almost no correlation between total case and pixel northeast of city centroid of 2000 in San Juan.

2. There are relatively more infected cases concentrated on 0.05-0.21 of pixel northeast of city centroid in San Juan in 2000.

Total cases is response variable, which's graph with pixel southeast of city centroid in San Juan in 2000 is shown below, and the explanation is as followed:

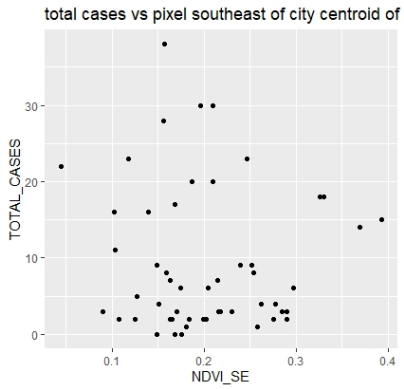


Fig 5: South East vegetation index vs total cases in San Juan for year 2000

From this scatterplot, we could the following information:

1. It is almost no correlation between total case and pixel southeast of city centroid of 2000 in San Juan.

2. There are relatively more infected cases concentrated on 0.1-0.3 of pixel southeast of city centroid in San Juan in 2000.

Total cases is response variable, which's graph with pixel northwest of city centroid in San Juan in 2000 is shown below, and the explanation is as followed:

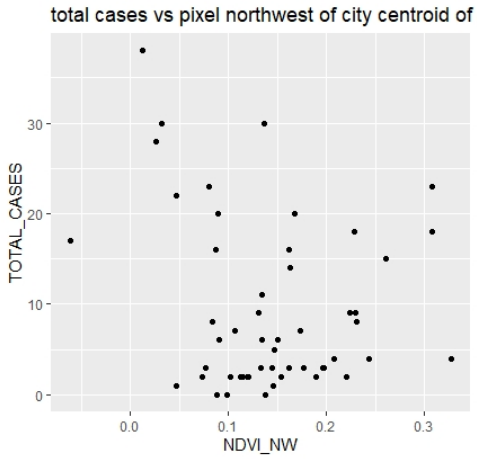


Fig 6: North West vegetation index vs total cases in San Juan for year 2000

From this scatterplot, we could the following information:

1. It is almost no correlation between total case and pixel northwest of city centroid of 2000 in San Juan.

2. There are relatively more infected cases concentrated on 0.1-0.2 of pixel northwest of city centroid in San Juan in 2000.

Total cases is response variable, which's graph with pixel southwest of city centroid in Iquitos in 2000 is shown below, and the explanation is as followed:

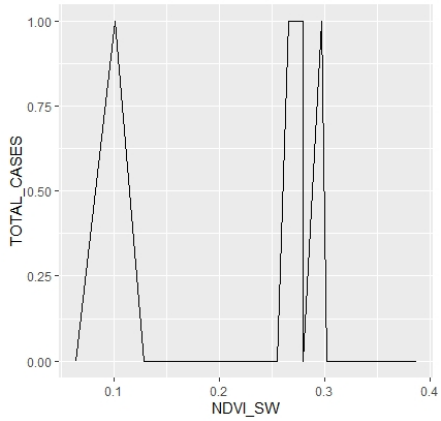


Fig 7: South West vegetation index vs total cases in Iquitos for year 2000

From this scatterplot, we could the following information:

1. It is almost no linear correlation between total case and pixel southwest of city centroid of 2000 in Iquitos. .

2. The peak value of total case in Iquitos in 2000 is 1, which's corresponding value of pixel northeast of city centroid are 0.1, 0.266, 0.28, 0.298.

Total cases is response variable, which's graph with pixel northeast of city centroid in Iquitos in 2000 is shown below, and the explanation is as followed:

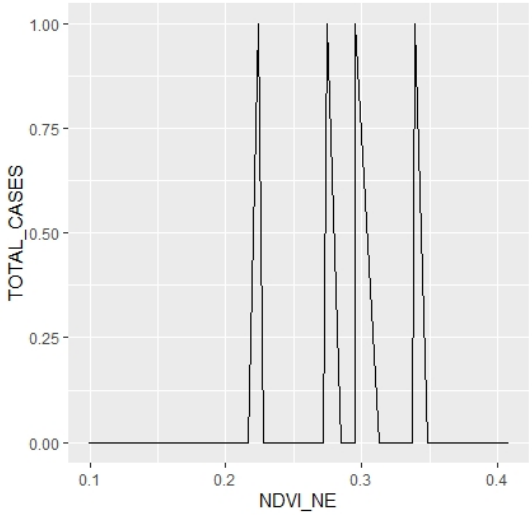


Fig 8: North East vegetation index vs total cases in Iquitos for year 2000

From this scatterplot, we could the following information:

1. It is almost no linear correlation between total case and pixel northeast of city centroid of 2000 in Iquitos. .

2. The peak value of total case in Iquitos in 2000 is 1, which's corresponding value of pixel northeast of city centroid are 0.231, 0.269, 0.3, 0.34.

Total cases is response variable, which's graph with pixel southeast of city centroid in Iquitos in 2000 is shown below, and the explanation is as followed:

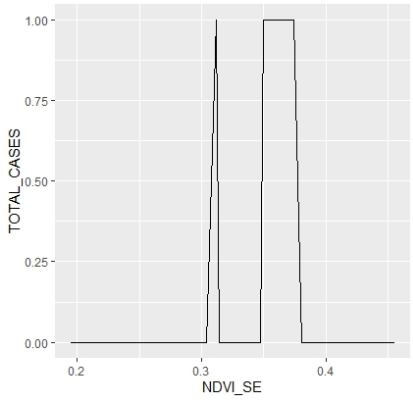


Fig 9: North East vegetation index vs total cases in Iquitos for year 2000

From this scatterplot, we could the following information:

1. It is almost no linear association between total case and pixel southeast of city centroid of 2000 in Iquitos. .

2. The peak value of total case in Iquitos in 2000 is 1, which's corresponding value of pixel northeast of city centroid are 0.32, 0.349, 0.351, 0.374.

Total cases is response variable, which's graph with pixel northwest of city centroid in Iquitos in 2000 is shown below, and the explanation is as followed:

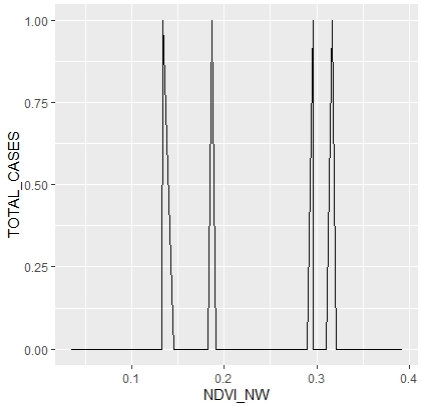


Fig 10: North West vegetation index vs total cases in Iquitos for year 2000

From this scatterplot, we could the following information:

1. It is almost no linear association between total case and pixel southwest of city centroid of 2000 in Iquitos. .

2. The peak value of total case in Iquitos in 2000 is 1, which's corresponding value of pixel northeast of city centroid are 0.133, 0.181, 0.296, 0.317.

Total cases is response variable, which's graph with total precipitation of San Juan in 2000 is shown below, and the explanation is as followed:

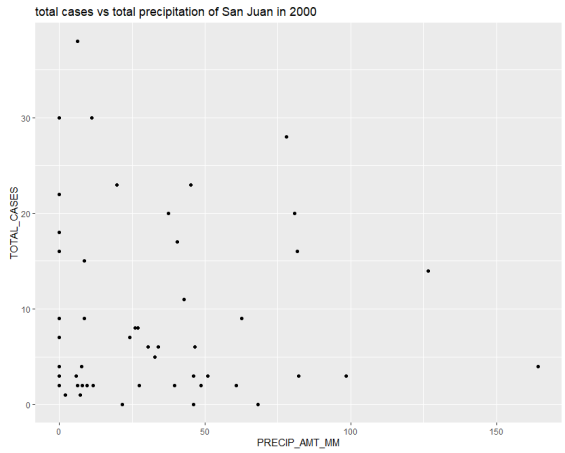


Fig 10: Precipitation vs total cases for San Juan in 2000

From this scatterplot, we could the following information:

1. It is almost no positive association between total case and total precipitation of San Juan in 2000.

2. There exists relatively weak and negative relationship between these two predictors, the numbers of cases more concentrated on 1-10, and total precipitation is more focused on 0-50 of San Juan in 2000.

Total cases is response variable, which's graph with total precipitation of Iquitos in 2000 is shown below, and the explanation is as followed:

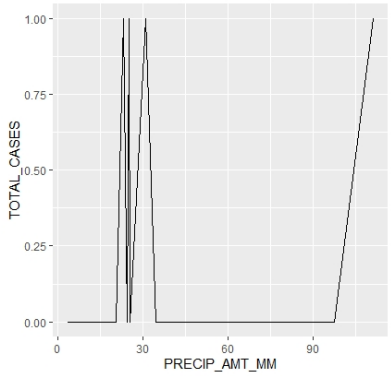


Fig 10: Precipitation vs total cases for Iquitos in 2000

From this line graph, we could the following information:

1. It is almost no linear association between total case and total precipitation of Iquitos in 2000.

2. The peak value of total case in Iquitos in 2000 is 1, which's corresponding value of total precipitation are 23.12, 31.10, 25.12, 111.06 of Iquitos in 2000.

The 3d plot graph of total cases vs average air temperature vs mean relative humidity of San Juan in 2000 is shown below, and the explanation is as followed:

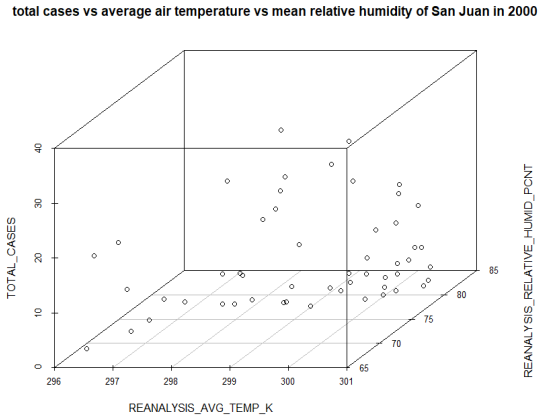


Fig 11: Total cases vs Reanalysis average temperature vs humidity percentage for San Juan in 2000

From this 3d plot, we could the following information:

1. It could be initially thought there exists certain positive association between total case and mean relative humidity of San Juan in 2000.

2. It could be initially thought there exists certain positive association between total case and average air temperature of San Juan in 2000.

3. There exists positive association between mean relative humidity and average air temperature.

The 3d plot graph of total cases vs average air temperature vs mean relative humidity of Iquitos in 2000 is shown below, and the explanation is as followed:

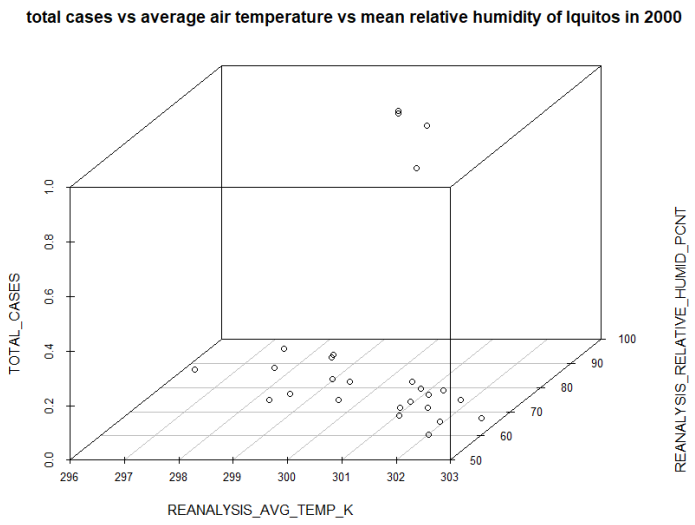


Fig 12: Total cases vs Reanalysis average temperature vs humidity percentage for Iquitos in 2000

From this 3d plot, we could the following information:

1. It could be initially thought there does not exist linear association among in total cases, mean relatively humid and average air temperature.

2. It seems Dengue had been controlled well in Iquitos in 2000, because there is only 4 cases of Dengue around the whole year in Iquitos.

3. There exists weak and negative association between mean relative humidity and average air temperature.

5.2 Correlation of predictors with response variable

San Juan

Iquitos

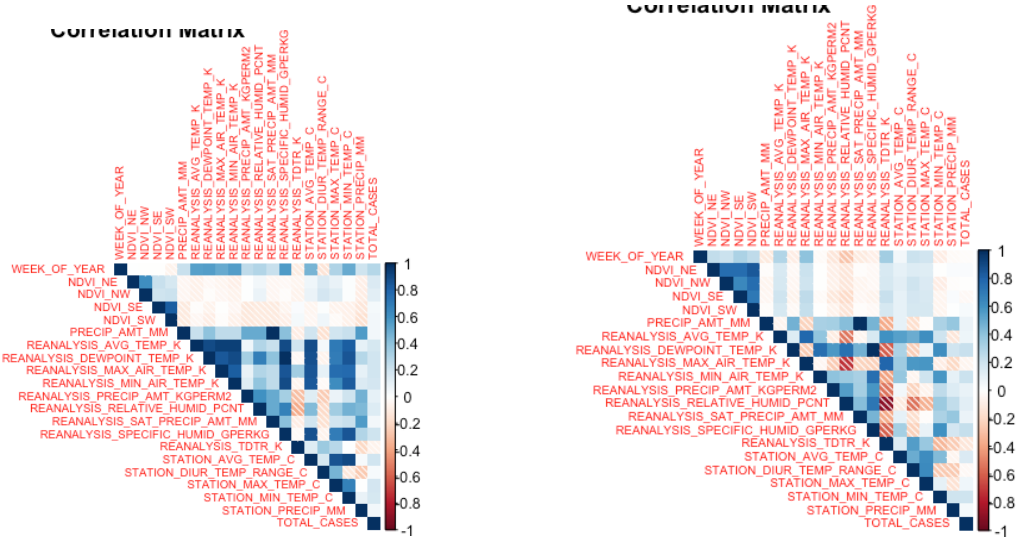


Fig 13: Correlation Matrix for San Juan and Iquitos

Some predictors show strong correlations with each other which highlights the need for dimension reduction. In addition, there seems to be no direct correlation between the predictors and response variable. However, above plots show strong associations between total number of cases and some predictors. Reason behind this could be the time series nature of features and existence of complex relationships between predictors and response variables.

6. Data Mining Tasks

6.1 Treating noise in data

Out of the 24 variables that are dealt with, 20 are numerical variables associated with climate and vegetation of the cities. Minimum and maximum values of each variable were observed in order to check for unreasonable values, which are far from reality. Observation was that all variables are within desired minimum and maximum values, hence no treatment was done in that aspect.

6.2 Missing data imputation

It is observed that 548 datapoints are missing in the overall dataset, out of which majority, 194 are vegetation data in North East regions of the two cities. All the 21 predictors consist of missing values but number of missing values in each column can be regarded as insignificant. However, missing values are imputed with the most recent value as values which are in similar time zone are expected to be similar, given the nature of data, which are climate and vegetation data (time series data).

6.3 Data transformation

6.3.1 Logistic transformation of response variable

Because of the asymmetric nature of total cases, which is the response variable, log transformation was done to it with the intention of improving the performance of models, specially multiple linear regression model.

Log transferred values for total cases results in a rough normal distribution as below.

San Juan Iquitos

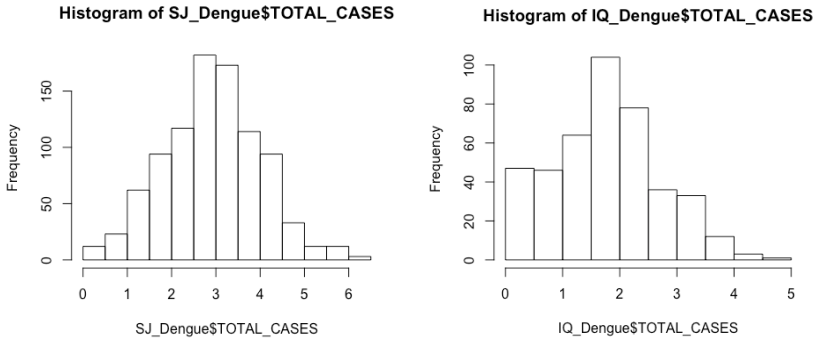


Fig 14: Histogram of log transformed total cases

Log transformed total number of cases seem to be following a normal distribution, which would be better to be used in modelling rather than the initial data.

6.4 Feature Engineering

Several factors contribute to dengue infection. Those can be categorized into two main sections as,

1. Climatic factors
2. Non-climatic factors

Transmission of dengue depends on several climatic factors. Climatic factors include temperature, humidity, precipitation etc. Climatic changes have a major impact on transmission of dengue[1]. These changes have an impact on the dynamic pattern of climate variables especially the temperature, rainfall, precipitation and humidity. Non-climatic factors include human growth, human movement, and socioeconomic constraints. When training the prediction model, we used climatic and non-climatic factors provided in the dataset. Mentioned in the problem description are several sensory measurements that we used as features.

The first feature added to this existing feature set is the number of dengue cases in the past week. This feature improved the prediction by a large margin. The reason for this behavior is that the number of dengue cases tends to increase and decrease gradually over several weeks for most of the time. Still, there can be contradictory cases where dengue cases increase suddenly but it is very rare.

As mentioned previously, climatic variables affect the number of dengue cases. This relation is indirect. The weather affects the life circle and survival of the mosquitos carrying dengue. Due to this, number of dengue cases could either increase or decrease. For example, temperature increase not only affects the reproduction and mosquito activity but also decreases the incubation time of larvae [2]. It is observed in past research that it takes different lag times for the larva incubation period ranging from 4 to 16 weeks [3]. We included a feature lag after reading above articles. We performed it by shifting the features by four weeks.

The increase or decrease in temperature and other features affect the number of dengue cases as well. Another feature was used to represent these changes, which show the difference of values with the previous one.

Accordingly, 6 new climatic and non-climatic features were introduced to the dataset as follows with the intension of reaching better accuracies of modelling.

i)Number of Dengue cases reported previous week (PRE_WEEK_ CASES)

ii)Humidity of previous month (PRE_MONTH_HUMIDITY)

iv)Dew point temperature of previous month (PRE_MONTH_DEWPOINT_TEMP)

v)Station average temperature of previous month (PRE_MONTH_STATION_AVG_TEMP)

vi)Maximum air temperature of previous month (PRE_MONTH_MAX_AIR_TEMP)

6.5 Dimension Reduction

Correlation matrix drawn under data exploration section shows there exists high correlation between certain predictors. Hence, it would be a better approach to reduce dimensions using Principal Component Analysis before modelling.

A scree plot was created to decide the number of principle components to be used. The plot displays the results of the scree test based on the observed eigenvalues (as straight line segments and x's),the mean eigen values derived from 100 random data matrices (as dashed lines) and the eigenvalues greater than 1 criteria.

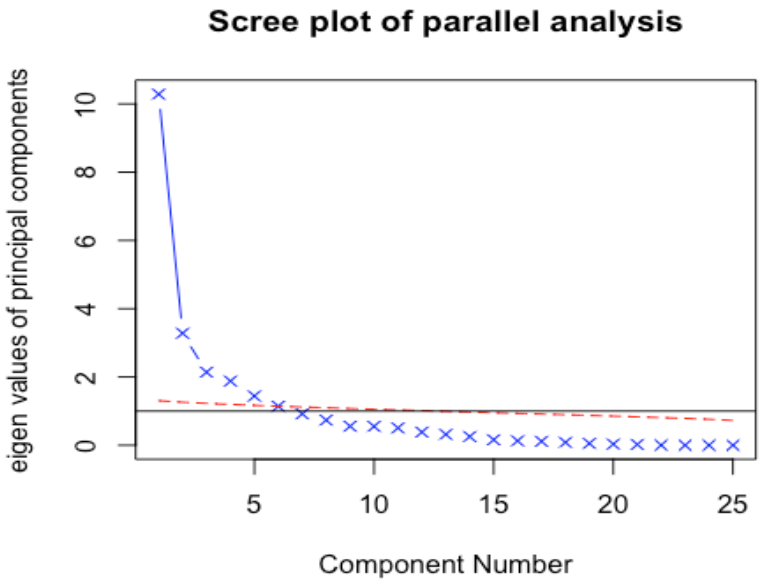


Figure 15: Scree plot

The scree plot suggests to have the dimensions reduced to 6 components. Thus, principal component scores were calculated under 6 principal components for each city.

Adhering to all the above data mining techniques, modelling was done to following variations of the data sets for each city; San Juan and Iquitos.

i. Original data
ii. Feature engineered data (Supported by literature review)
iii. Principal components
iv. Heuristically selected subset of variables through literature review and simple correlation coefficients

6.6 Classification of number of total Dengue cases

A new response variable was added as class of total number of Dengue classes, to support the classification aspect of the problem. Total number of cases were split in to 5 classes as Very Low, Low, Medium, High and Very High as per below criteria.

Condition	Class
0 <= TOTAL CASES <= 7	Very Low
8 <= TOTAL CASES <= 15	Low
16 <= TOTAL CASES <= 25	Medium
26 <= TOTAL CASES <= 45	High
TOTAL CASES >= 45	Very High

Whilst data related to San Juan showed no domination between classes, data related to Iquitos showed high domination of certain classes for above criteria. Over sampling turned out to be impossible as the least represented class contained only 4 cases. Once classification model; Random Forest Classifier was exercised on Iquitos and the results further solidified this situation.

7 Data Mining Models/Methods

Total number of Dengue cases, which is the response variable is given as a numerical value in the original variable. This response variable was modified in such a way to result in two separate classification and prediction problems as mentioned in chapters before.

7.1 Classification of Dengue cases

7.1.1 Random Forest Classification

San Juan Iquitos

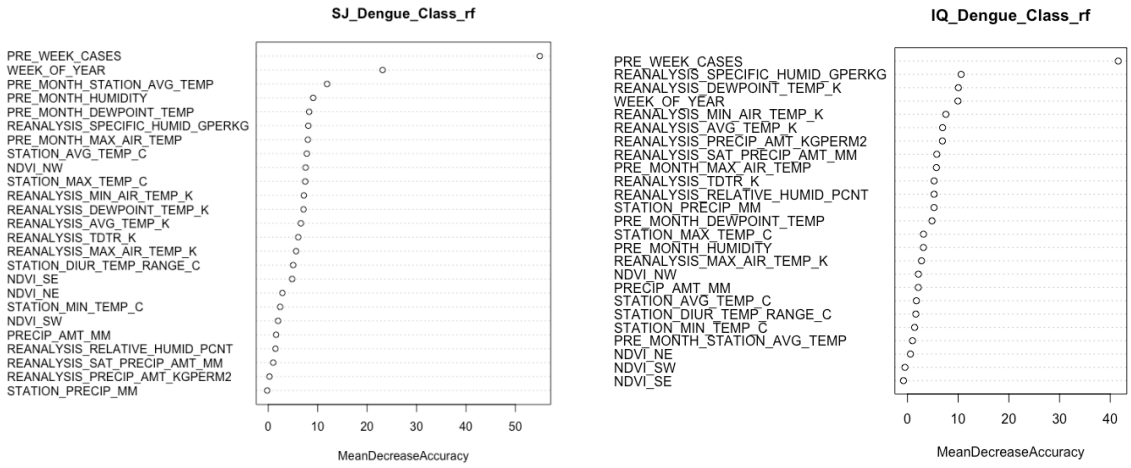


Figure 16: Feature importance table

Random Forest Classification was done with 600 trees, 7 variables randomly sampled as candidates at each split and keeping 5 as minimum size of terminal nodes. Following is the variable importance plot which depicts that the 5 most prominent features are PRE_WEEK_CASES, WEEK_OF_YEAR , PRE_MONTH_STATION_AVG_TEMP , PRE_MONTH_HUMIDITY and PRE_MONTH_DEWPOINT_TEMP. Out of these five, almost four

are features that were added later under feature engineering, which once again justifies the addition of those new features.

7.1.2 K-Nearest Neighbors

Number of neighbors from 1 to 14 were tested to arrive at the ideal number of neighbors for the model for each city. K was selected in such a way that it does not overfit the model as well as does not ignore local structure and the ideal number of neighbors turned out to be 8.

7.1.3 Neural Networks

A neural network was build using principal components as predictors and Total Cases class as response variable. Input layer comprised of 6 nodes (the six principal components), one hidden layer was exercised with 6 nodes and the output layer comprised of 4 nodes which represent the 4 classes. Algorithm “backpropagation” was used with logistic activation function in the architecture of the neural network and 5 fold cross validation was used throughout in building the model as neural networks are so much prone to overfitting.

7.1.4 Discriminant Analysis

Discriminant Analysis was identified as a possible model to evaluate because of the nature of data, where all predictors are numerical and response variable in the considered problem is categorical. Hence, most relevant features were selected using stepwise regression and correlation analysis, which was then followed by a Discriminant analysis done to classify the of total number of Dengue cases for classes for each of the city. Features used for the model are WEEK_OF_YEAR, REANALYSIS_DEW_POINT_TEMP_K, REANALYSIS_SCPECIFIC_HUMID_GPERKG, STATION_MAX_TEMP_C, PRE_WEEK_CASES, PRE_MONTH_HUMIDITY, which resulted in following functions for each of the city.

<u>San Juan</u>

	High	Low	Medium	Very High
constant	-5.146981e+06	-5.148011e+06	-5.14763e+06	-5.14698e+06
WEEK_OF_YEAR	3.222564e+01	3.218842e+01	3.219229e+01	3.222305e+01
REANALYSIS_DEW_POINT_TEMP_K	3.696052e+04	3.696421e+04	3.696286e+04	3.696051e+04
REANALYSIS_SCPECIFIC_HUMID_GPERKG	-3.798538e+04	-3.798922e+04	-3.79879e+04	-3.79852e+04
STATION_MAX_TEMP_C	1.705876e+02	1.707535e+02	1.706001e+02	1.706360e+02
PRE_WEEK_CASES	-3.360654	-3.378060	-3.372086	-3.314213
PRE_MONTH_HUMIDITY	6.033990	6.033440	6.035363	6.034671

Iquitos

	High	Low	Medium	Very High
constant	-5.058709e+06	-5.058842e+06	-5.05796e+06	-5.06117e+06
WEEK_OF_YEAR	2.400404e+01	2.401973e+01	2.399564e+01	2.391499e+01
REANALYSIS_DEW_POINT_TEMP_K	3.620189e+04	3.620242e+04	3.619923e+04	3.621072e+04
REANALYSIS_SCPECIFIC_HUMID_GPERKG	-3.554736e+04	-3.554816e+04	3.554465e+04	-3.55563e+04
STATION_MAX_TEMP_C	6.013114e+02	6.011402e+02	6.011983e+02	6.000015e+02
PRE_WEEK_CASES	2.530108e+01	2.491341e+01	2.501301e+01	2.530520e+01
PRE_MONTH_HUMIDITY	3.803757e+02	3.806310e+02	3.805293e+02	3.800513e+02

7.2 Prediction of total number of Dengue cases

The number of Dengue cases are predicted for the two cities separately with multiple linear regression, K-Nearest Neighbors, Regression Trees and Neural Network and with different variations of datasets such as principal component scores, heuristically selected features and full feature engineered dataset.

7.2.1 Multiple Linear Regression

Multiple linear regression was performed assuming following characteristics, which are believed to be mandatory if to be fitted to multiple linear regression model.

i) Noise and Response variable follow normal distribution. As shown above, it was observed that Total Cases follow an asymmetric distribution. Therefore, the response variable was replaced by their logarithmic values to ensure the normality.

ii)Predictors show linear relationship with the response variable.

iii)Homoscedasticity is maintained.

iv)Predictors are deterministic.

v)Perfect multicollinearity is not present between predictors.

Stepwise and Exhaustive search techniques are performed to select a subset of features from the full feature set. Full feature set included of the original features and added features after feature engineering. Following table shows the features suggested by the two approaches.

San Juan

Stepwise Feature Search	Exhaustive Search
WEEK_OF_YEAR	WEEK_OF_YEAR
REANALYSIS_DEWPOINT_TEMP_K	REANALYSIS_DEWPOINT_TEMP_K
REANALYSIS_SPECIFIC_HUMID_GPERKG	REANALYSIS_SPECIFIC_HUMID_GPERKG
STATION_MAX_TEMP_C	STATION_MAX_TEMP_C
PRE_WEEK_CASES	PRE_WEEK_CASES
PRE_MONTH_HUMIDITY	PRE_MONTH_HUMIDITY
	PRE_MONTH_STATION_AVG_TEMP

Iquitos

Stepwise Feature Search	Exhaustive Search
WEEK_OF_YEAR	WEEK_OF_YEAR
NDVI_SW	NDVI_SW
REANALYSIS_SPECIFIC_HUMID_GPERKG	REANALYSIS_SPECIFIC_HUMID_GPERKG
STATION_MAX_TEMP_C	STATION_MAX_TEMP_C
PRE_WEEK_CASES	PRE_WEEK_CASES
PRE_MONTH_HUMIDITY	PRE_MONTH_HUMIDITY
PRE_MONTH_DEWPOINT_TEMP	PRE_MONTH_DEWPOINT_TEMP
REANALYSIS_RELATIVE_HUMID_PCNT	REANALYSIS_MAX_AIR_TEMP_K
	REANALYSIS_RELATIVE_HUMID_PCNT
	REANALYSIS_TDTR_K

Both the approaches suggest to use 3 features which were added later during feature engineering, in the model, which further justifies the addition of these features to the dataset.

The two multiple linear regression models resulted are quite similar to each other as they share many predictors in common for both the cities.

Model 1 (Using stepwise feature selection) :

San Juan

TOTAL_CASES = 33.5042 + 1.8107 * WEEK_OF_YEAR - 24.9751 * REANALYSIS_DEWPOINT_TEMP_K + 26.5415 * REANALYSIS_SPECIFIC_HUMID_GPERKG + 1.8695 * STATION_MAX_TEMP_C + 45.0145 * PRE_WEEK_CASES - 2.4380 * PRE_MONTH_HUMIDITY

Iquitos

TOTAL_CASES = 7.1571 - 0.8156 * WEEK_OF_YEAR + 1.0217 * NDVI_SW - 1.9507 * REANALYSIS_MAX_AIR_TEMP_K - 2.2997 * REANALYSIS_RELATIVE_HUMID_PCNT + 1.7450 *

REANALYSIS_SPECIFIC_HUMID_GPERKG + 0.5936 * STATION_MAX_TEMP_C + 5.7477 * PRE_WEEK_CASES - 8.1030 * PRE_MONTH_HUMIDITY + 7.6790 * PRE_MONTH_DEWPOINT_TEMP

Model 2 (Using exhaustive search) :

San Juan

TOTAL_CASES = 33.4947 + 1.7758 * WEEK_OF_YEAR - 25.4482 * REANALYSIS_DEWPOINT_TEMP_K + 26.7668 * REANALYSIS_SPECIFIC_HUMID_GPERKG + 1.6345 * STATION_MAX_TEMP_C + 44.9644 * PRE_WEEK_CASES - 3.8788 * PRE_MONTH_HUMIDITY + 1.9315 * PRE_MONTH_STATION_AVG_TEMP

Iquitos

TOTAL_CASES = 7.1621 - 0.7678 * WEEK_OF_YEAR + 0.9751 * NDVI_SW - 1.5752 * REANALYSIS_MAX_AIR_TEMP_K - 2.9440 * REANALYSIS_RELATIVE_HUMID_PCNT + 1.6984 * REANALYSIS_SPECIFIC_HUMID_GPERKG - 1.0600 * REANALYSIS_TDTR_K + 0.5291 * STATION_MAX_TEMP_C + 5.7574 * PRE_WEEK_CASES - 8.0868 * PRE_MONTH_HUMIDITY + 7.5815 * PRE_MONTH_DEWPOINT_TEMP

7.2.2 K-Nearest Neighbors

As discussed above, most of the predictors are related to weather and this gives an insight on the range of number of neighbors to evaluate in the model. Hence, mean absolute errors of predictions when K ranges from 1 to 14 were analyzed, leaving climate variables to develop meaningful relations and at the same time ensuring to capture local structure without overfitting.

Both dataset with all the features (including added features) and principal component scores were modeled using this technique.

When all the features were used, the desirable number of neighbors were 6 for both the cities. This figure was selected in such a way that it minimizes error and at the same time causes no overfit.

When using principal components, the desirable number of neighbors were 2, and 3 respectively for the two cities which is justifiable given the lesser number of features.

7.2.3 Regression trees

Regression trees were drawn to the two cities under following scenarios.

Scenario 1 :- Dataset with heuristically selected features, based on literature review and correlation coefficient analysis.

Fully grown tree was pruned in such a way that the error is minimized.

Initial set of features used for San Juan are PRE_WEEK_CASES , WEEK_OF_YEAR , REANALYSIS_DEW_POINT_TEMP_K, REANALYSIS_SPECIFIC_HUMID_GPERKG , PRE_MONTH_HUMIDITY, PRE_MONTH_DEWPOINT_TEMP , PRE_MONTH_STATION_AVG_TEMP , PRE_MONTH_MAX_AIR_TEMP out of which, most prominent features depicted for San Juan after pruning the tree are; PRE_WEEK_CASES , WEEK_OF_YEAR , REANALYSIS_DEW_POINT_TEMP_K, REANALYSIS_SPECIFIC_HUMID_GPERKG.

Similarly, out of the initial set of features used for Iquitos, PRE_ WEEK_CASES , WEEK_OF_YEAR , REANALYSIS_DEW_POINT_TEMP_K, REANALYSIS_MIN_AIR_TEMP_K , REANALYSIS_SPECIFIC_HUMID_GPERKG, REANALYSIS_SPECIFIC_HUMID_PCT turned out to be the most prominent ones.

Dataset 2:- Principal component scores

Full regression tree done to principal component scores were pruned to attain at the most prominent principal components; PC5, PC3 and PC1 for San Juan and PC3, PC2, PC1, PC6, PC4 for Iquitos

7.2.4 Neural Networks

Two neural networks were done for each city, San Juan and Iquitos. The first was done using heuristically selected features through literature review and correlation analysis and the second was done using principal components. Architectures comprised of one hidden layer with 6 nodes and one output layer which corresponds to the response variable and logistic activation function and backpropagation weight update strategy was used.

8 Performance Evaluation

All the models except neural networks use 70% : 30% training validation split for training and performance evaluation purposes. five fold cross validation was used in neural networks with the purpose of avoiding overfitting the model to the training dataset.

8.1 Classification performance

8.1.1 Random Forest Classification

San Juan

Accuracy = 0.74 with 95% CI

P-Value < 2.2e-16

Kappa = 0.63

	Low	Medium	High	Very High
Sensitivity	0.5424	0.8509	0.6774	0.8
PPV	0.8649	0.6879	0.6774	0.9

Receiver Operating Curves are drawn for each class by converting performance evaluation in a series of binary problems and getting probabilities of each observation to belong to each of the

classes when the validation test is predicted. Following ROCs and Area Under the Curve values were resulted from this process.

Legend : Orange = Low , Blue = Medium , High = Red , Brown = Very High

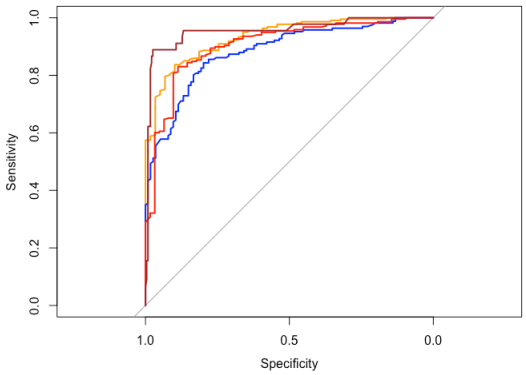


Fig 17: ROC for Random Forest Classification in San Juan

	Low	Medium	High	Very High
Area Under the Curve	0.934	0.883	0.9057	0.9561

Iquitos

Accuracy = 0.73

P-Value = 0.05

Kappa = 0.37

	Low	Medium	High	Very High
Sensitivity	0.97	0.33	0	0
PPV	0.8	0.48	0	NA

Accuracy being favorable, P-Value and Kappa being unfavorable, sensitivity being favorable only for certain classes hint of an unbalanced dataset. When exploring, only 4 cases belonging to class Very High and 29 cases belonging to class High were found, which is not enough data

points even to do oversampling. Hence, a good classification model which isn't bias could not be built because of shortage of cases belonging to some classes.

8.1.2 K-Nearest Neighbors

Accuracy = 0.54 with 95% CI

P-Value < 9.54e-06

Kappa = 0.32

	Low	Medium	High	Very High
Sensitivity	0.31	0.73	0.40	0.53
PPV	0.47	0.53	0.45	0.83

Following ROCs and Area Under the Curve values were resulted from this process.

Legend : Orange = Low , Blue = Medium , High = Red , Brown = Very High

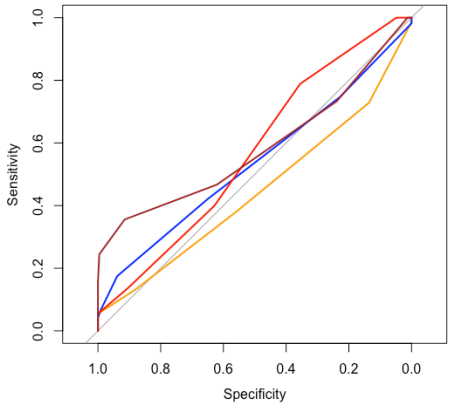


Fig 18: ROC for K Nearest Neighbors in San Juan

	Low	Medium	High	Very High
Area Under the Curve	0.45	0.54	0.57	0.58

8.1.3 Discriminant Analysis

San Juan

Accuracy = 0.54 with 95% CI

P-Value < 9.54e-06

Kappa = 0.32

	Low	Medium	High	Very High
Sensitivity	0.31	0.73	0.40	0.53
PPV	0.47	0.53	0.45	0.83

8.2 Prediction performance

8.2.1 Multiple Linear Regression

The multiple linear regression models developed which incorporated stepwise feature search and exhaustive feature search are observed to be equally performing when considering their accuracies and lift charts.

San Juan

	Stepwise	Exhaustive Search
Mean Absolute Error	13.286	13.281
Root Mean Square Error	27.739	27.847

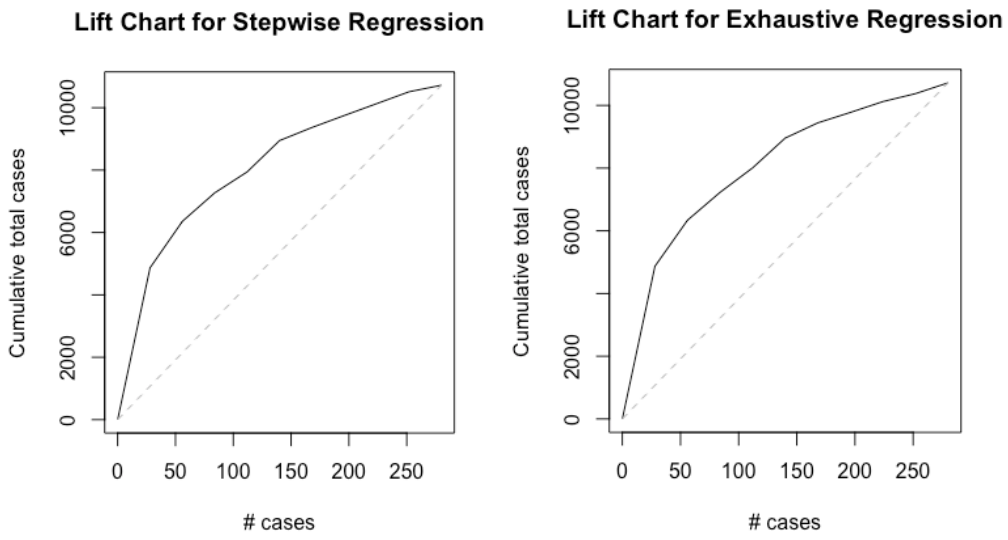


Fig 19: Lift charts for Multiple Linear Regression in San Juan

Iquitos

	Stepwise	Exhaustive Search
Mean Absolute Error	5.27	5.27
Root Mean Square Error	10.58	10.56

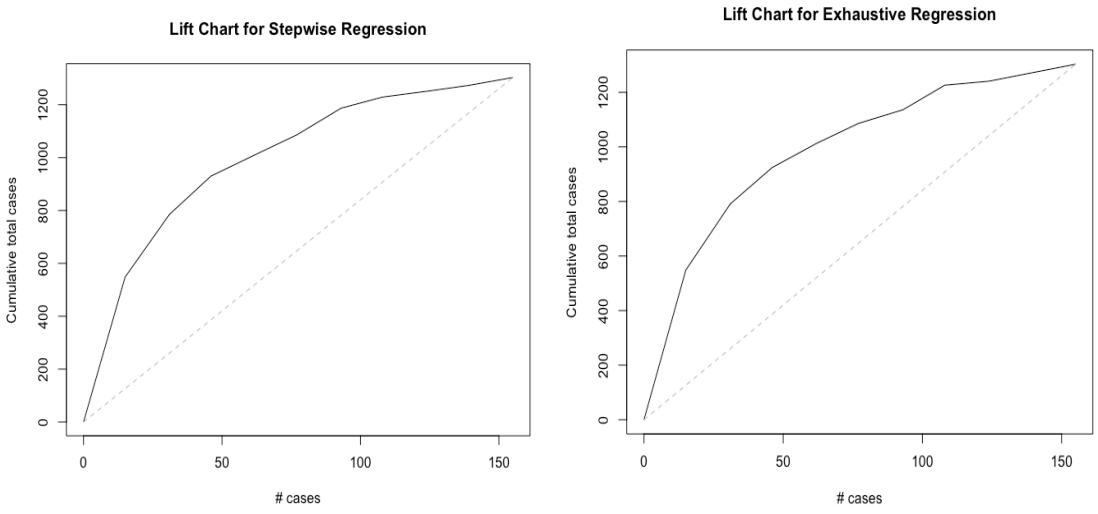


Fig 20: Lift charts for Multiple Linear Regression for Iquitos

8.2.2 K-Nearest Neighbors

It was decided that the optimal number of neighbors to be evaluated is 6, which gives the following accuracy.

San Juan

	All features	Principal Components
Selected K value	6	2
Mean Absolute Error	6.63	18.76
Root Mean Square Error	13.25	32.74

Iquitos

	All features	Principal Components
Selected K value	6	2
Mean Absolute Error	2.9	8.74
Root Mean Square Error	14.24	16.24

8.2.3 Regression trees

San Juan

	Situation 1 (Heuristically selected features)	Situation 2 (Principal components)
Mean Absolute Error	11.73	15.9

Iquitos

	Situation 1 (Heuristically selected features)	Situation 2 (Principal components)

Mean Absolute Error	5.3	8.85

9 Project Results

9.1 Performance comparison and final model selection for Problem 1.1

Problem 1.1 which is defined in chapter 2 under problem definition revolved around building a classification model to determine the class of total number of Dengue cases for San Juan. A comparison of performance of the models developed is as follows.

Max AUC = Maximum area under the curve among 4 classes; Low, Medium, High and Very High
Min AUC = Minimum area under the curve among 4 classes; Low, Medium, High and Very High

	Accuracy	Kappa	Max AUC	Min AUC
Random Forest Classification	0.74	0.63	0.95	0.88
K-Nearest Neighbors	0.54	0.32	0.45	0.58
Neural Networks	0.7			
Discriminant Analysis	0.53	0.30	0.55	0.44

The model with highest accuracy, kappa and max and min values is selected as the final model.

Selected model = Random Forest Classification

9.2 Performance comparison and final model selection for Problem 1.2

Problem 1.2 refers to prediction of total number of Dengue cases for the week of year for the city of San Juan. Mean Absolute Error is used as the key performance matrix for selecting this model as it is expected to be the most ideal measure in predicting epidemics. Because, predicting lesser than actual can lead to adverse effect of unexpected burden on healthcare system and predicting more than actual can cause in unnecessary expenditure in preparing for worse situations. Moreover, these adverse effects are expected to have a linear relationship

with errors rather than exponential. If that association is exponential, higher errors would have increased effects and at such situations Root Mean Square Error would be a better measure. But, at this situation where it is a linear association between errors and adverse effects that is expected, Mean Absolute Error would be the ideal measure to go for. Hence, following performance comparison.

	Mean Absolute Error
Multiple Linear Regression - Stepwise	13.286
Multiple Linear Regression – Exhaustive Search	13.281
K Nearest Neighbors – All Features	6.63
K Nearest Neighbors – Principal components	18.76
Pruned Regression Tree	11.79
Regression Tree – Principal components	15.9
Neural Network – Selected features	27.87

Even though K nearest Neighbors model seems to have yielded the minimum error, it would be tricky with large datasets when putting it into use because of its high computational time. Hence selected model = Pruned Regression Tree.

But KNN with all the features including added features during feature engineering process could also be used with small datasets.

9.3 Performance comparison and final model selection for Problem 2.1

Problem 2.1 refers to determining the class of total cases for week of year for the city of Iquitos. However, it was found out that only 4 data points are available for the class “Very High”, which was not enough to build any model. Since oversampling would also not be possible with such a small number.

9.4 Performance comparison and final model selection for Problem 2.2

Problem 2.2 revolves around building a prediction model for the number of total Dengue cases reported for week of year for the city of Iquitos and following is the comparison between the models built.

	Mean Absolute Error
Multiple Linear Regression - Stepwise	5.27
Multiple Linear Regression – Exhaustive Search	5.27
K Nearest Neighbors – All Features	3
K Nearest Neighbors – Principal components	8.74
Pruned Regression Tree	5.3
Regression Tree – Principal components	8.85
Neural Network – Selected features	5.27

All errors here seem to be low and this could also hint an overfit of the models to the available dataset. This scenario might also be explained as a result of the limited availability of datapoints for the city of Iquitos.

However, the best performer from above is K Nearest Neighbors applied to all features. But as discussed in chapter 9.2 above, there might be computational difficulties with large datasets. Hence, selected model = Pruned Tree.

10 Insights for Decision Making

- The data mining task of feature engineering done has proven to have significant impacts on the models built in terms of performance increments. Climate factors such as the humidity a month ago, temperature and non-climatic factors such as number of cases in previous week make considerable impact on the number of cases in the target week. This is a very valuable insight which helps the decision makers to get an idea on the expected number of cases in coming weeks by observing existing climate conditions and case numbers.

- The two cities San Juan and Iquitos perform differently even though the basic structure is the same. Model parameters for each city varies from one another by considerable amounts. Moreover, number of cases in Iquitos seem to have higher association with vegetation, which is not a relationship seen in San Juan. Hence, an insight can be drawn that the cities behave differently, owing to geographical differences.
- KNN model performing better hints that distance matrix of predictor parameters have a considerable impact on total number of cases and are of course closely connected with the response variable.

11 Impact of the Project Outcomes

Through the above insights, healthcare sector and other related sectors of any city can be better prepared for the surges of the Dengue epidemic, which results in

> ensuring lesser mortality rates
> preventive actions which helps to keep the numbers down
> increasing quality of healthcare service
> increasing drug availability and reducing deficiencies of the same
> avoiding supply chain disturbances
> reducing adverse economic impacts through prior availability of information

12 References

[1] Aziz, T. a. Lukose, D. a. b. A. Bakar, S. a. Sattar and Abdul, "A Literature Review of Methods for Dengue Outbreak Prediction".

[2] Kuno and Goro, "Review of the factors modulating dengue transmission," Epidemiologic reviews, vol. 17, pp. 321-335, 1995.

[3] Y.-H. Hsieh and C. W. S. Chen, "Turning points, reproduction number, and impact of climatological events for multi-wave dengue outbreaks," Tropical Medicine \& International Health, vol. 14, pp. 628-638, 2009.