DENGUE FEVER ANALYSIS

DATA MINING AND MACHINE LEARNING PROJECT

Tian Xu Shuying Zhao Ami Doshi

IST707 | Professor Ying Lin | iSchool Syracuse University

ABSTRACT

Dengue fever is an epidemic carried by mosquitos. The disease can cause fever, rash, pain, even worse, it can lead severe blood-bleeding, low blood pressure, and death. In the tropical and subtropical area, the disease strongly harmed people's health in the 90s. Nowadays, along with the improving living quality and medical level, the dengue fever is under control to some extent. However, in the human society, there is not a vaccine to prevent the disease. A serious scenario is that as long as mosquitos exist, dengue fever exists.

Facing the serious scenario, we hope that using data mining and machine learning technology can support medical workers to control the disease. Owing to the advancement of data mining tools, controlling dengue fever has become increasingly available. The dataset of dengue fever includes 1456 rows and 24 variables with detailed information of two targeted cities, San Juan and Iquitos. The dataset tells us the number of total cases, weather information, humidity information, etc. The data was appropriately cleaned and analyzed to gain insights into the factors influencing dengue fever spreading. In the report, we describe the data, the method we preprocess the data, our analysis and implementation. We also conclude by determining the factors and stating how the factors influence dengue fever to help medical workers control the disease.

INTRODUCTION

This project was undertaken for the course IST 707: Data Mining and Machine Learning at the School of Information Studies, Syracuse University. The team had the dataset from *DRIVENDATA.com*. It is also a data science competition, *DengAI: Predicting Disease Spread*. The dataset includes the dengue fever case information from year 1990 to year 2010. Since the disease cannot be eliminated now, the aim of the project is to gain an insight into controlling it by carrying out data analysis.

OBJECTIVE

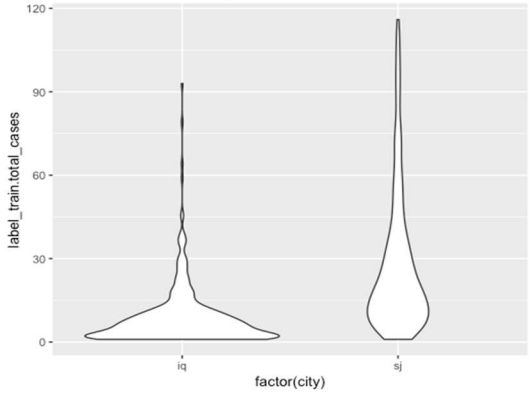
The objective of the project is to determine key factors that influence dengue fever spread in the typical cities, San Juan and Iquitos. In this case, we aim to gain an insight into how climate influence dengue fever disease spread in tropical and subtropical area by leveraging data mining and machine learning.

UNDERSTAND THE DATA

Before data analysis, we should understand the dataset in data type, data distribution, etc. to get the big picture of the data for future analysis and model selection. In this part, we leveraged visualization to help us have a direct understanding about the dataset.

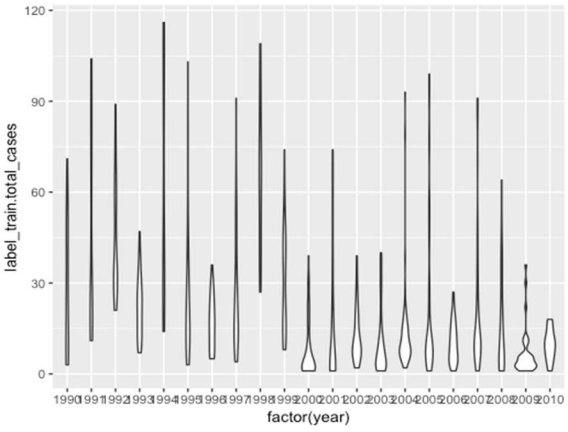
By structuring and summarizing the data, we learned that the data types are consisted of *Factor, Int, and Numerical.* Except for column city, year, weekofyear, week_start_date, ndvi_ne, ndvi_nw, ndvi_se, and ndvi_sw, the remaining columns are all about climate. Next, we summarized the data to understand the distribution and missing values. We learn that there are a number of missing values in the different columns. After checking the distribution of missing values, we had a plan to deal with the missing values, which is replacing NAs with latest values. In the following part, we will present the step. After having the first impression of the data, we visualized the detailed part to gain a better understanding. The below visualizations were created with R.

First, we visualized the total cases in the two cities to check the dengue fever case distribution of San Juan and Iquitos separately. Below is the visualization using violin shape. We learned that extreme cases happened in San Juan and many small numbers of cases happened in Iquitos.

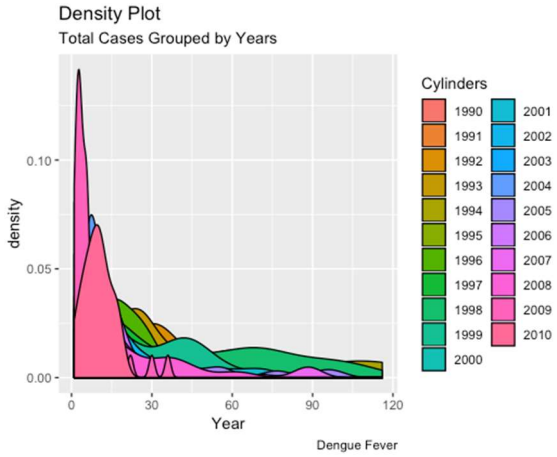


Second, we visualized the relations of different factors in San Juan and Iquitos from 1990-2010.

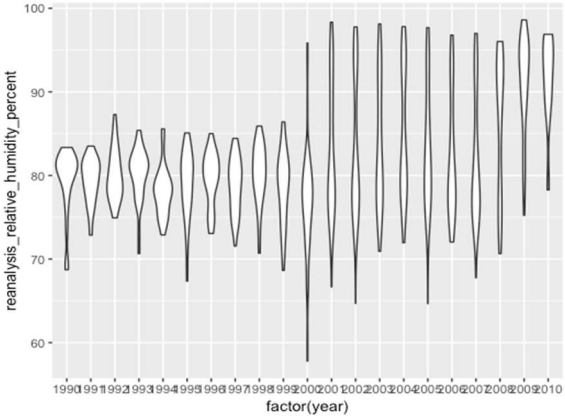
The total cases in each year



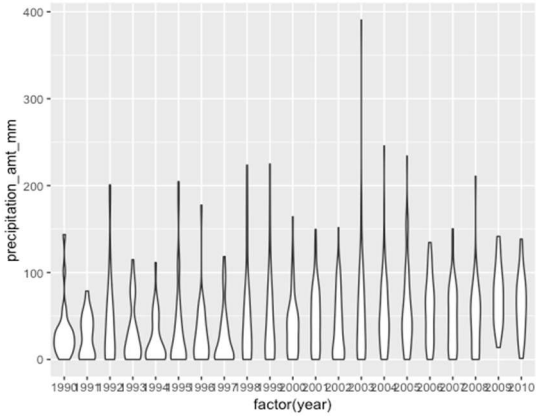
Density of total cases in each year



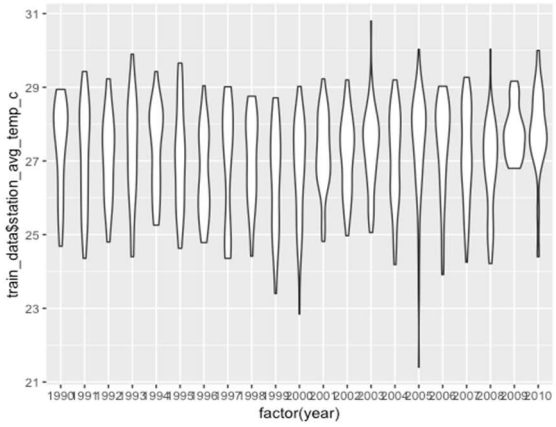
The humidity level in each year



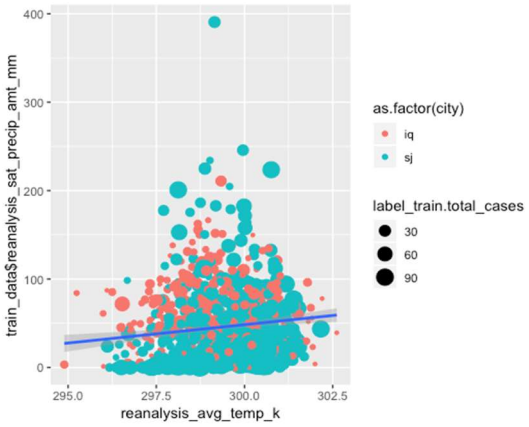
The precipitation in each year



The temperature level in each year



Multiple factors and total cases

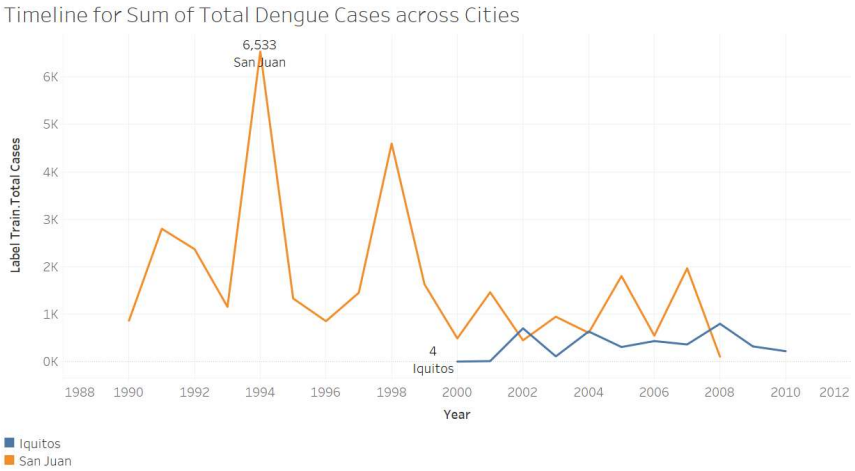


DATA PREPARATION

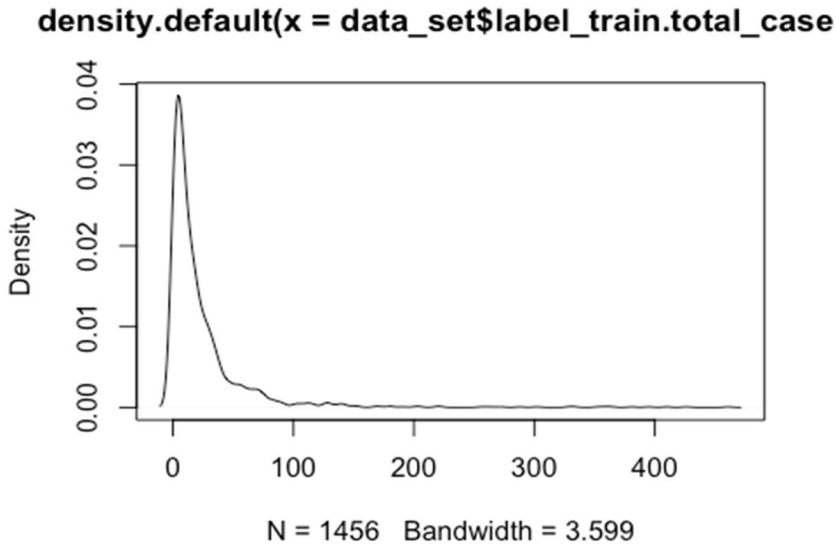
For the dataset that we have, we combined the total cases variables with other features variables into the whole dataset to build models. The test dataset would be the one to predict with the no labels of total cases.

We found there are 548 missing values which is the large amount of data that we cannot directly delete them directly. According to the kind of the data, we decided to use the latest values to replace the missing values.

While understanding the present columns in our dataset, we find that there was a spike in the outbreak of dengue during 1994 for San Juan and 2002 for Iquitos. Otherwise, we find a random trend with its total number oscillating, thus we remove the column 'week_start_date' which isn't adding much of a value.



For the dependent variables 'total cases', we need to familiar with its distribution.



We found that there are lots of the low value of total cases. According to the total cases distribution, we would category the numeric variables into categoric variables with fix breaks. With quantile of the total cases, it shows that the 40% of the total cases are equal or below 8 cases and 60% of the total cases are equal or below 17 cases.

For the further data processing, we did it under different algorithms sections separately. In the data preparing section, we split the train data into training dataset and validation dataset by using the Hold-Out method to sample the dataset. We chose to put 80% of the whole train dataset into the training dataset. Then the rest of the dataset went into the validation dataset.

MODEL SELECTION

In the model selection part, we not only tried association rule and decision tree with categorical independent variables, but also built models by using K Nearest Neighbor, four ensemble learning methods and Support Vector Machine with numeric independent variables. For the dependence variable 'total cases', we chose to categorize it to find the reason why total cases would be low, high or middle values.

Algorithm 1: Association Rule

We use the association rule to find out the co-occur relationship between feather variables and the high or the low total cases. We also discretized the independent variables into three levels with "low", "middle" and "high". Then we did the association rules with supervised learning.

- Data pre-processing for Association Rule

First, we prepare dataset by converting numeric variables into the categoric variables by using cluster method to discretize it into three bins: low, medium and high. Since the distribution of total cases isn't uniform, we break it into two bins: less than 12 total cases to be 'low' and more than 12 being high.

- Model

To find the reason that lead to low total cases, we listed the top 20 rules with higher lift values and found that date, city in iq, low minimum air temperature, low total precipitation and high diurnal temperature range in Forecast system reanalysis would lead to low total cases.

	lhs	lift
2	Week of year=middle,reanalysis_min_air_temp_k=low	1.977
3	City=iq, Week of year=middle, reanalysis_min_air_temp_k=low	1.977
4	Week of year=middle, reanalysis_air_temp_k=low	1.974
5	City=iq, Week of year=middle	1.971
6	city=iq, reanalysis_precip_amt_kg_per_m2=low, station_min_temp_c=low	1.929
7	city=iq, reanalysis_min_air_temp_k=low, reanalysis_precip_amt_kg_per_m2=low	1.879
8	reanalysis_min_air_temp_k=low, station_avg_temp_c=middle	1.878
9	city=iq, reanalysis_min_air_temp_k=low, station_avg_temp_c=middle	1.877
10	reanalysis_min_air_temp_k=low, reanalysis_precip_amt_kg_per_m2=low	1.877
11	city=iq, ndvi_nw=high, reanalysis_precip_amt_kg_per_m2=low	1.854
12	reanalysis_min_air_temp_k=low, reanalysis_tdtr_k=high	1.851
13	city=iq, reanalysis_min_air_temp_k=low, reanalysis_tdtr_k=high	1.837
14	city=iq, reanalysis_min_air_temp_k=low, station_min_temp_c=low	1.837
15	ndvi_nw=high, reanalysis_tdtr_k=high	1.835
16	city=iq, ndvi_nw=high, reanalysis_tdtr_k=high	1.82
17	city=iq, reanalysis_air_temp_k=low, reanalysis_min_air_temp_k=low	1.82
18	reanalysis_min_air_temp_k=low, station_min_temp_c=low	1.814
19	city=iq, ndvi_nw=high, station_min_temp_c=low	1.811
20	city=iq, reanalysis_air_temp_k=low, station_avg_temp_c=middle	1.811
21	reanalysis_precip_amt_kg_per_m2=low, reanalysis_tdtr_k=high	1.807

We also did the same step for the high total cases and found that city in sj, date and low pixel southwest of city centroid, low diurnal temperature range, low max temperature and middle mean relative humidity would lead to high total cases.

	lhs	life
2	city=sj, weekofyear=high, ndvi_nw=middle	1.881
3	weekofyear=high, ndvi_nw=middle, reanalysis_tdtr_k=low	1.881
4	city=sj, weekofyear=high, ndvi_nw=middle, reanalysis_tdtr_k=low	1.881
5	city=sj, weekofyear=high, ndvi_sw=low	1.846
6	weekofyear=high, ndvi_sw=low, reanalysis_tdtr_k=low	1.846
7	city=sj, weekofyear=high, ndvi_sw=low, reanalysis_tdtr_k=low	1.846
8	weekofyear=high, ndvi_sw=low, station_diur_temp_rng_c=low	1.841
9	city=sj, weekofyear=high, ndvi_sw=low, station_diur_temp_rng_c=low	1.841
10	weekofyear=high, ndvi_sw=low, reanalysis_tdtr_k=low, station_diur_temp_rng_c=low	1.841
11	city=sj, weekofyear=high, ndvi_sw=low, reanalysis_tdtr_k=low, station_diur_temp_rng_c=low	1.841
12	city=sj, weekofyear=high, reanalysis_relative_humidity_percent=middle	1.821
13	weekofyear=high, reanalysis_relative_humidity_percent=middle, reanalysis_tdtr_k=low	1.821
14	city=sj, weekofyear=high,reanalysis_relative_humidity_percent=middle, reanalysis_tdtr_k=low	1.821
15	weekofyear=high, reanalysis_max_air_temp_k=low	1.785
16	city=sj, weekofyear=high, reanalysis_max_air_temp_k=low	1.875
17	weekofyear=high, reanalysis_max_air_temp_k=low, reanalysis_tdtr_k=low	1.875
18	city=sj, weekofyear=high, reanalysis_max_air_temp_k=low, reanalysis_tdtr_k=low	1.875
19	weekofyear=high, reanalysis_max_air_temp_k=low, station_diur_temp_rng_c=low	1.773
20	city=sj, weekofyear=high, reanalysis_max_air_temp_k=low, station_diur_temp_rng_c=low	1.773
21	weekofyear=high, reanalysis_max_air_temp_k=low, reanalysis_tdtr_k=low, station_diur_temp_rng_c=low	1.773

Compare with the variables that lead to the low total cases, we found that different attributions would lead to the high total cases. Despite year and week of year, the only thing that we found that value of reanalysis_tdte_k (Diurnal temperature range) affects the total cases.

Additionally, we use association rules to find relationships between the climate parameters and total dengue cases. We found some interesting relations stating:

1. The San Juan city has higher total dengue fever cases over Iquitos.
2. High temperature leads higher dengue fever cases.
3. High precipitation leads higher dengue fever cases.

Our claims were given more weightage, with our findings through research that mosquitos are more active when the degree is more than 80 degree, and lay eggs on the surface of water leading to more diseases.

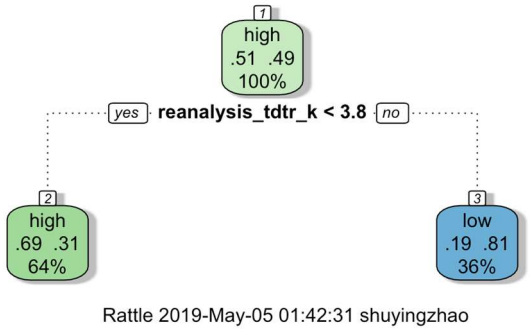
Algorithm 2: Decision Tree

- Data pre-processing for Decision Tree

We only categorized the total cases values into less than 12 total cases to be 'low' and more than 12 being high to build the decision tree.

- Model

After model tuning and post-pruning of the model, the decision tree shows below.



The result shows that the value of reanalysis_tdte_k (Diurnal temperature range) smaller than 3.8 would lead to higher value of the total cases.

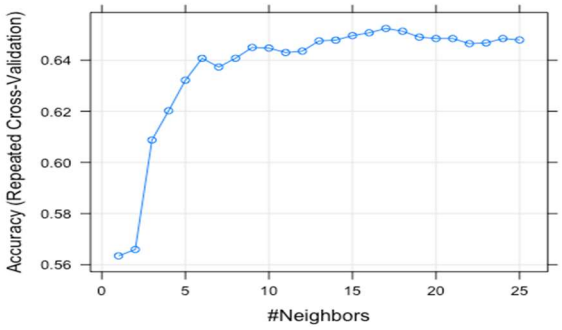
Algorithm 3: K Nearest Neighbor

- Data pre-processing for KNN

First, we prepare dataset by converting numeric variables into the categoric variables by using fixed method to discretize it into three bins: low, medium and high. Since the distribution of total cases isn't uniform, we break it into three bins of low, medium and high. Now that we have the categorized data, we center scale the data.

- Model

We use N Nearest Neighbor which is instance-based learning to make data points closer to each other tend to behave similarly. We tuned the After tuning the KNN model, the prediction accuracy increased to 64% with 6 neighbors which is low for us to applied into the testing dataset.



Algorithm 4: Ensemble learning methods

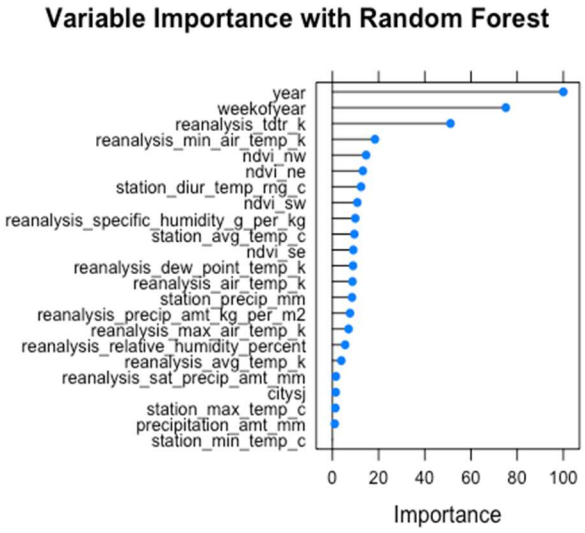
There are four ensemble learning methods that we tried to find one with the best performance.

- Bagging

We use bagging method which is bootstrap aggregation to create multiple samples to train multiple classifiers for reducing variance. The accuracy of model with bagging method is 69.1%.

- Random Forest

We trained large number of decision stumps to reduce bias. Finally, we used 12 features to random select at each split and bootstrapped 25 times. The accuracy of model with the random forest method is 70%.
The following plot show the important variables by using committee majority voting. We found that year, week of year and reanalysis_tdte_k (Diurnal temperature range) are the top three important variables that affect the total cases.



- Gradient Boosting Machine (EGB)

We first use gradient boosting machine to make the weak learners to strong learners and then minimize loss function by using gradient descent with new weights. The accuracy rate of model which using gradient boosting machine method is 67%.

- Extreme Gradient Boosting (XGB)

We used method which is similar with gradient boosting machine but faster using parallel computing. The accuracy of model with extreme gradient boosting is 70%.

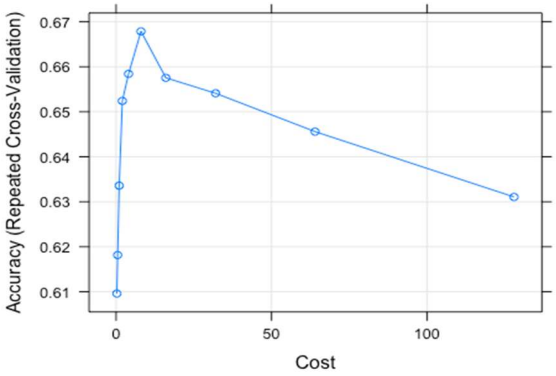
Algorithm 5: Support Vector Machine (SVM)

- Data pre-processing for SVM

First, we prepare dataset by scaling the data to center.

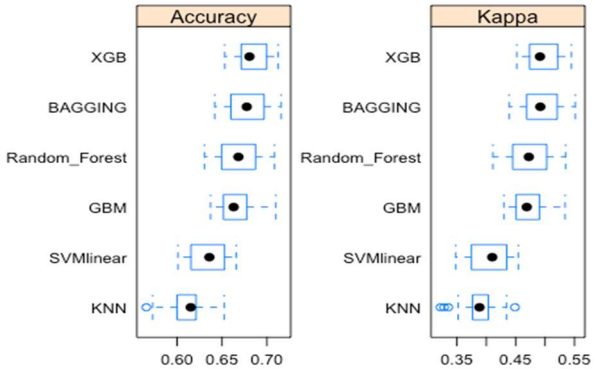
- Model

With linear kernel and three levels of total cases, we got the 61% accuracy rate with choosing 0.3 cost value. With non-linear kernel radial basis function, we tuned the model and finally chose 0.0117 sigma value and 8 cost value. Finally, we got 65% accuracy rate of the SVM model with non-linear kernel.



MODEL COMPARISON

In the last part of the modeling, we compared the models that we tried in previous. With high accuracy rate and kappa value, we finally choose Gradient Boosting Machine method to predict the total cases with higher accuracy rate and kappa value.



CONCLUSION

After working on different algorithms, we finally chose the model with Gradient Boosting Machine method to predict and found that date and reanalysis_tdte_k (Diurnal temperature range) really would affect the total cases. Our claims are that San Juan is more prone to Dengue, with cases of dengue linear correlation with temperature and precipitation with frequency majority dengue cases being lower than 100. We refer that the specific climate condition promotes mosquitos' activity because mosquitos are active when temperature is higher than 80 degree. What's more, San Juan is more prone to Dengue because it's an island, which means it's surrounded by water, of which the surface mosquitos lay eggs. Thus, we can gain insights into how climate influences Dengue Fever spread.