Final Report for DengAI-Predicting Dengue Disease Spread in Iquitos, Peru & San Juan, Puerto Rico

Paul Ycay
CKME136 Data Analytics Capstone
Ryerson University

April 6, 2020

1 Introduction

Dengue is the fastest spreading mosquito controlled disease worldwide, flourishing in poor urban areas & subtropical, tropical climates. The *Aedes Aegypti* species of mosquitoes are largely responsible for transmitting the virus, which causes symptoms of joint pain and high fever. Up to 50-100 million cases have been estimated in 100 endemic countries, spreading heavily within Latin America and Southeast Asia as of 2019 (Hosangadi, 2019; *What is dengue?*, 2017). Climatic variables play a large influence in the development of the Dengue mosquitoes. In light of the ongoing COVID-19 outbreak, studying the relationships between meteorological factors and reported cases during an epidemic will help warn the general public in taking necessary precautions of future outbreaks. Data science and predictive analytics can play an important role in alleviating the spread of disease. This project will use Dengue data taken from the competition DengAI: Predicting Disease Spread, hosted by DrivenData. The data includes climatic information on San Juan, Peru & Iquitos, Puerto Rico between 1990-2010 (training set). Using varying meteorological data provided by the National Centers for Environmental Information (NOAA), the goal of this project was to predict the number of cases in the test set, which spans between 2008-2013 in San Juan and 2010-2013 in Iquitos.

2 Literature Review

2.1 History and context of the Dengue fever

Dengue is a viral disease generated by dengue virus serotypes 1 to 4 (DENV-1 to DENV-4). The four serotypes (groups of microorganisms within a species sharing the equivalent number and type of antigens) originated about 1000 years back, with reports of endemic transmission in humans dating just a few hundred years ago. The earliest record of the virus dates back to 992 BC in a Chinese medical encyclopedia. There has been recurring epidemics of a disease with a strong homogeneity to dengue occurring in Asia and the Americas, just before the end

of the 18th century. Thus, scientists have hypothesized the spread of the virus originating from the tropic and subtropic regions (Salles et al., 2018).

The *Aedes* genus of mosquitoes are responsible for transmitting DENV, as well as other arboviruses that circulate around the world; these being the yellow fever virus (YFV), chikungunya virus (CHIKV), and the infamous Zika virus (ZIKV). These invasive species coexist in rural, suburban, and urban settlements in tropical and subtropical regions around the globe. This review will focus mostly on the *Aedes Aegypti*, the main vector involved in the transmission cycle of DENV. The *Ae. aegypti* mosquito originated from Africa as a more "zoophilic" form, evolving and expanding from tropical forests to urban areas. *Aedes aegypti* arrived to the New World during the transatlantic shipping traffic that occurred between the 1500s and 1700s; their migration brought along the first reports of Dengue in the region (Salgueiro et al., 2019). Due to the impact caused by the *Ae. aegypti* in recolonizing tropical and subtropical regions in the New World, the dengue virus in South America has become the most notorious case of mosquito borne diseases in that continent. Annually, over 1 million clinical cases are reported (at most 718,000 are laboratory confirmed) since their reintroduction in the 1980s after the 1960s eradication campaign was achieved in North America (Ferreira-De-Brito et al., 2016).

Aedes aegypti has been the key contributor to large dengue epidemics throughout the Americas in the last few decades, particularly from the influence climate factors play within the mosquito's life cycle. Globally, increasing average temperatures can lead to a higher rate of contact with prey, humans. Moreover, sea levels can influence the density of vector mosquitoes along the coast; rising sea levels could lead to more breeding sites for freshwater vectors (mosquitoes) within saline waters. These shifts in climate and connected extreme events have a high linkage in the epidemiology of mosquito-borne disease; meteorological factors influence the epidemiology of Dengue by increasing virus replication, mosquito development, population growth and human-mosquito interactions . It is important in studying environmental patterns in areas of high infection vector-borne diseases as it is associated with economic, demographic, and social factors (Méndez-Lázaro, Muller-Karger, Otis, Mccarthy, & Peña-Orellana, 2014).

San Juan in Puerto Rico (17.92°N-18.52°N, 65.62°W-67.28°W) is reported annually to have average air surface temperatures of 24-29°C, with average precipitation of about 1800 mm. During the dry season (0-50 mm occurring between March and June), air temperatures fluctuate between 36-40°C, while the rainy seasons report 30-35°C. Laureano-Rosario et.al (2018) used artificial neural networks (ANNs) to predict dengue fever occurrences between 1994-2012. Variables such as population size, date, maximum air temperature, and previous dengue cases were the most influential factors in predicting occurrences in San Juan. An increase in air temperature has high correlation with the number of mosquito bites due to the energy demands of the species, leading to higher probabilities of infection. Warmer climates reduce the time development of mosquitoes, and therefore increase their densities. Although an increase in minimum air temperature promotes mosquito succession, their development is inhibited as it becomes warmer than normal. Vulnerable human populations of the Dengue transmission and diagnosis are those at risk because of social mobility (school, work, etc...)

and public services (drainage cleaning, trash pick-up); the most susceptible are populations with biological conditions (< 5 years old) and chronic diseases (> 65 years old) (Laureano-Rosario et al., 2018).

The city of Iquitos (3.7437°S, 73.2516°W) sits at the Amazon Rivers of northeast Peru and the confluence of Nanay, Itaya. Stoddard et. al (2014) studied laboratory confirmed data related to dengue dynamics between the years 2000 and 2010. Their studies is split within 3 seasons: trimester I, II, and III. Maximum and mean temperatures per week were warmest between November and April, which coincide with the detection of most Dengue cases. Over all the years, rainfall was highest between 2003 and 2008 with significantly less rainfall subsequent years. Temperatures in trimester I were warm, Amazon river levels were increasing, Dengue cases were moderate, and rainfall is elevated. In trimester II (July-August), climate is relatively drier and cooler, fewer Dengue cases are reported, and river levels are moderate. In trimester III (late Summer-Fall), river levels are at their lowest and begins to rise again, precipitation increases, temperatures are their warmest, and transmission of Dengue picks up. The researchers have concluded that all climatic variables displayed modest seasonality, with variables correlating weakly with the amount of Dengue cases throughout a range of time lags. Intensive vector control efforts can reduce the transmission of the virus if placed and timed properly (Stoddard et al., 2014).

The prevention of Dengue largely depends on effective measures of controlling disease vectors and promote health. In suppressing the spread of Dengue without a current effective vaccine, it is important to focus on *Aedes* elimination. Communities are encouraged to survey mosquito activity within their households. Health education regarding the vector behavior is carried out as well. Lectures in schools address mosquito breeding sites. Residents should check their water storage tanks for breeding activity and report it. Overall, the importance of home inspection should be practiced by locals; containers that house water are typical areas in which mosquito eggs are laid. Screens on windows and doors and regular cleaning of water apparatus are important in mosquito control (Donateli, Einloft, Junior, Cotta, & Costa, 2019; *How to Prevent the Spread of the Mosquito that Causes Dengue.*, n.d.).

2.2 Useful sources on predictive analytics

This subsection will detail papers/sources on predictive analytics and the methods used in their respective work that may be of interest on the dataset we will be working on. Not all techniques will be performed, but the ones discussed below are of particular interest.

Principal component analysis is beneficial in large datasets as it reduces their dimensionality, while improving interpretability at a cost of minimizing information loss. This method creates new uncorrelated variables which maximizes variance in succession (Jolliffe & Cadima, 2016). Ahmed & Siddiqui (2014) perform PCA on Dengue cases with 5 climatic variables: wind speed (W), precipitation (P), maximum temperature (Mx), minimum temperature (Mn), and relative humidity (H). Their experiment is conducted on data reported during 2011-2012 in Lahore, Pakistan. The purpose of applying PCA was to analyze general environmental structure/factor which could have an affect in the wake of Dengue fever cases in

Pakistan's climate. The first principal component (PC1) is a linear combination of minimum temperature (Mn), maximum temperature (Mx) and wind (W); it is interpreted as "Windy and Hot". PC2 is a combination of precipitation (P) and relative humidity (H); it is interpreted as "Wetness". PC3 is a contrast of variables precipitation (P) and wind speed (W), labeled "Windy and Dry". PC4 is also a contrast of relative humidity (H) and precipitation (P), which translates to "Humid but no Rain". PC1 indicated the variation in temperature with high wind speed as the major climatic factor. PC2 concludes that wetness and low wind speed as the major factor of spreading Dengue fever. In PC3, low wind and wetness influence the breeding of mosquito vectors (Ahmed & Siddiqui, 2014).

Multiple linear regression extends on linear regression used to predict y, the outcome variable, on the basis of multiple distinct explanatory variables x_i, $i = 1 \dots n$. MLR can be useful in examining the correlation of Dengue incidences with respect to various climate factors (James, Witten, Hastie, & Tibshirani, 2017). Anggraeni et. al (2017) used regression and clustering methods to create a Dengue incidence prediction model based on several climatic and vector population factors. Regression methods in separate studies often obtain 92% prediction accuracy with weather data as independent variables. Data from previous cases and weather incidents are clustered and within each cluster, a regression model is built (Anggraeni et al., 2017)

The `caret` package in R (short for Classification and REgression Training) contains several functions to train the model for complex classification and regression problems. The package utilizes various other R packages, such as training random forests, Multivariate Adaptive Regression Splines (MARS), Adaboost Classification Trees, among others. We can play with various algorithms and test to see which is the most accurate in correlating incidents of Dengue with strong climatic features (Prabhakaran, 2018; Kuhn, 2020).

Random forests derive from the same techniques as decision trees and bagging. Bagging trees improves predictive performance by introducing a random component in the tree building process, thus reducing a single tree's prediction variance. The randomizing of individual trees in the process can be repeated hundreds of times, with results being averaged. Averaging results can reduce variance without increasing bias. Random forests perform typically well, with very little tuning required. They are also strong to outliers and do not require much pre-processing (*Random Forests*, n.d.; Lesmeister, 2015).

Support Vector Machine (SVM) is a supervised machine learning algorithm that classifies data into different classes. The ideology of SVM is behind the hyperplane, which acts as a decision boundary between numerous classes. Support Vector Machine is mainly known for classification problems, but can also be used in regression (Support Vector Regressor). Support Vector Machine uses labeled data to train on; this is what's known as a supervised learning algorithm. Moreover, the kernel trick is used in SVM to classify non-linear data. The kernel trick transforms data into another dimension of the hyperplane which has a clear margin divider between data classes (Lateef, 2019; Lesmeister, 2015).

3 Overview of the DengAI: Predicting Disease Spread Prediction Dataset

The dataset to be used for the project can be found at https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/. This data is part of an ongoing competition hosted by DrivenData titled DengAI: Predicting Disease Spread. The sources of the data are taken from the National Oceanic and Atmospheric Administration and Centers for Disease Control and Prevention (*Dengue*, 2020; *National Oceanic and Atmospheric Administration*, 2020). There are 4 `.csv` files titled `dengue_features_train.csv`, `dengue_features_test.csv`, `dengue_labels_train.csv`, & `submission_format.csv`. There are 25 variables in the training set, with 1457 observations. The original question posed by DrivenData as part of their competition was to predict the `total_cases` feature in the test set for each city. A `total_cases.csv` variable found in the training labels set can be merged into actual training data set. Two cities are used as the main classes for climate study: San Juan, Peru & Iquitos, Puerto Rico. Variables on minimum, maximum, & mean of precipitation (mm) are given. Mean relative & mean specific humidity, in % & kg, respectively, are described. Minimum, maximum, & mean of air temperatures in Kelvin (K) are provided for each city. An interesting set of variables are the Normalized Difference Vegetation Indices (NDVI), derived from remote sensing data closely linked to drought conditions. NDVI is a measure of plant health based on how the plant reflects light at certain frequencies; i.e. a calculation of vegetation health. This value ranges from -1 to 1. Negative values correspond to dead plants or inanimate objects, healthy plants have positive indices. The following figure illustrates the relationship between the indices & vegetation:

The next page will provide a quick description of the variables.

VARIABLE NAME (dengue_labels_train)	DESCRIPTION
city	Iquitos, Peru & San Juan, Puerto Rico
year	Year
weekofyear	Week of the corresponding year
week_start_date	Timeframe in DD-MM-YYYY
station_max_temp_c	Maximum temperature (°C): taken from National Centers for Environmental Information (NOAA) Global Historical Climatology Network (GHCN)
station_min_temp_c	Minimum temperature (°C): taken from National Centers for Environmental Information (NOAA) Global Historical Climatology Network (GHCN)
station_avg_temp_c	Average temperature (°C): taken from National Centers for Environmental Information (NOAA) Global Historical Climatology Network (GHCN)
station_precip_mm	Total precipitation (mm): taken from National Centers for Environmental Information (NOAA) Global Historical Climatology Network (GHCN)
station_diur_temp_rng_c	Diurnal temperature range (°C): taken from National Centers for Environmental Information (NOAA) Global Historical Climatology Network (GHCN)
precipitation_amt_mm	Total precipitation (mm)
reanalysis_sat_precip_amt_mm	Total precipitation (mm) : NOAA's National Centers for Environmental Prediction
reanalysis_dew_point_temp_k	Mean dew point temperature in Kelvin (K)
reanalysis_air_temp_k	Mean air temperature in Kelvin (K)
reanalysis_relative_humidity_percent	Mean relative humidity (ratio of the amount of water vapor actually present in the air to the greatest amount possible at the same temperature)
reanalysis_specific_humidity_g_per_kg	Mean specific humidity (mass g of water vapour in a unit mass kg of moist air)
reanalysis_precip_amt_kg_per_m2	Total precipitation (in kg /square meter)
reanalysis_max_air_temp_k	Max air temp in Kelvin (K)
reanalysis_min_air_temp_k	Min air temp in Kelvin (K)
reanalysis_avg_temp_k	Average air temp in Kelvin (K)
reanalysis_tdtr_k	Diurnal temperature range in Kelvin (K)
ndvi_se	NOAA's CDR Normalized Difference Vegetation Index. 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
total_cases	Total # of cases in timeframe

4 Data Approach

R was the programming language used for this project.

Step 1: Importing datasets, loading packages, and data preparation:

- Import the 4 datasets. Merge the following datasets together: dengue_features_train & dengue_labels_train
- Common libraries used to explore the data was `dplyr`, `tidyr`, `readr`, as well as pre-loaded R functions.
- Check the format of variables and convert it appropriately (such as the date format)

- Identify dimensions of the data & determine NAs. We subset the data by city.
- Impute missing climate data (either median or most recent non-NA prior to it)

Step 2: Exploratory Analysis

- We treat the training data as two seperate datasets based on city & make statistical assumptions on each dataset
- Using the `pastecs` library, we can easily generate our univariate & bivariate analysis in the form of a data-frame based on each city.
- The function stat_desc() will provide us basic statistics, such as the mean, median, mode, and any outliers of the data. It will also provide us advanced stats in a single data-frame
- We provide plots of our response variable `total_cases`, and how it functions overtime, as well as Time Series plots of relevant climate features.
- Correlation matrices can be used in determining which features have low influence on the target variable, `total_cases`
- Finally, we plot the response variable against the features in our data, which can tell us the appropriate machine learning algorithms to proceed with

Step 4: Modeling:

- From the hypotheses made about the data in the initial steps (correlation, p-values, plots), deploy techniques and algorithms to test the data.
- Test our data with multiple linear regression first, predicting `total_cases`.
- Apply Random Forest algorithm: used in both classification and regression.
- Random forest uses an ensemble of decision trees (randomized), where each tree determines a vote for prediction among the target variable; the algorithm picks the prediction with the most votes.
- Apply Support Vector Machine last; if regression is not suitable for this data due to non-linear relationships, SVM will be able to treat this

Step 5: Validation:

- Each algorithm will be approved on the testing set
- We pick the algorithm with the most accuracy and the least Mean Square Error
- This step will come towards the end of the project, after all tests have been made.

- We use the algorithm and its respective model to predict the total cases in the Test Set; we submit our predictions on the website. The website's scoring metric is based on Mean Absolute Error; used to calculate the amount of error in the predictions, and averages all of the absolute errors

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

5 Analysis & Methodology

5.1 Initial Analysis

The resources used to apply the Machine Learning on this dataset came various posts on StackExchange, Udemy's Machine Learning A-Z™: Hands-On Python & R In Data Science (https://www.udemy.com/course/machinelearning/), and Cory Lesmeister's *Mastering Machine Learning with R* (2015).
Much of the comments in the analysis & results can be found in the following link:
https://rpubs.com/pycay

We first import, merge data, and convert variables to their proper factor classes (such as converting the week_start_date variable to date). We impute the missing climate data with the most recent non-NA prior to it; we apply this method of imputation since our data is based on climatic variables, attributes are known to follow seasonal trends. Another thing we manipulated were features measured in kelvin (K); we converted these features in the celsius (°C) unit of measure because we discovered other features in our dataset, when importing, that used this unit of measure. We wanted to be consistent with the data, so we used a simple function and for loop to convert the variables to the right unit of temperature measure.

Regarding our uni-variate & bi-variate analysis, we are able to produce statistical results based on Iquitos, Peru and San Juan, Puerto Rico using the stat_desc() function. We split the training data based on cities since not all features share relationships based on location. To read the data-frame, nbr.val indicates the number of observations, nbr.null indicates the number of null values, nbr.na is the number of missing values (it's 0 in all features since we have imputed beforehand), SE.mean calculates the standard error of mean, while CI.mean calculates the mean values with confidence intervals.

Here is some information on Iquitos subsetted from the training data

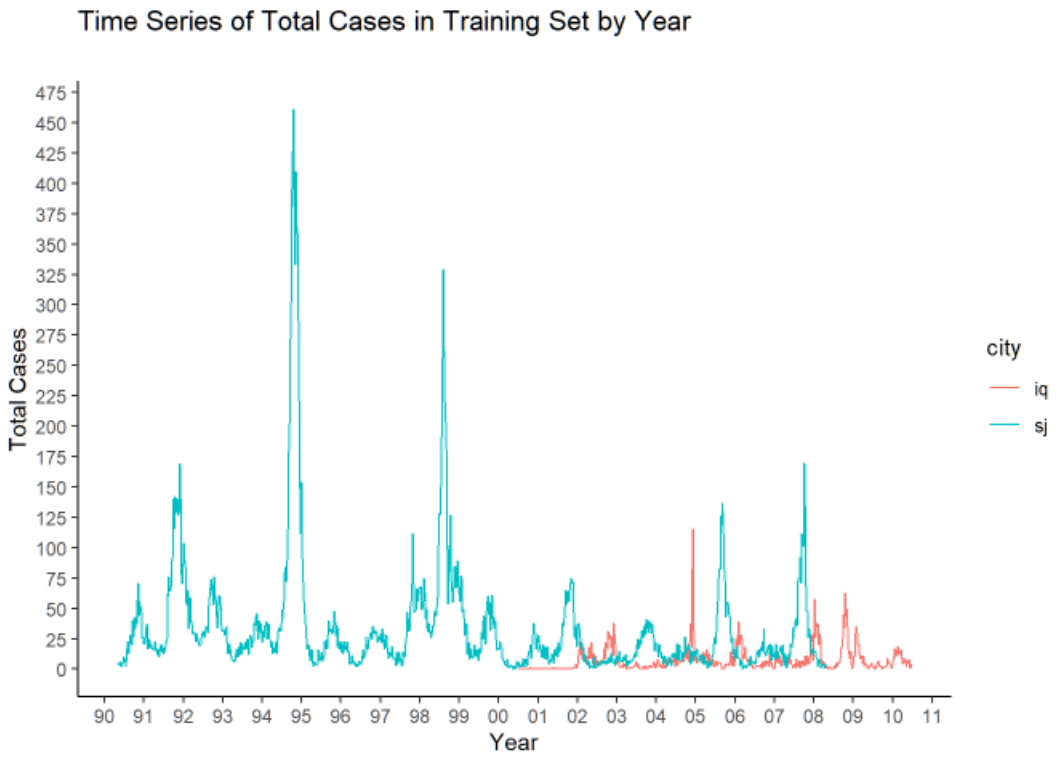
	nbr.val	nbr.null	nbr.na	min	max	range	sum	median	mean	SE.mean	CI.mean	var	std.dev	coef.var
ndvi_ne	516	0	0	0.06172857	0.5083571	0.4466285	136.3675	0.2653786	0.2642780	0.003577638	0.007028559	6.604538e-03	0.08126831	0.30751065
ndvi_nw	516	0	0	0.03586000	0.4544286	0.4185686	123.3429	0.2335928	0.2390366	0.003377975	0.006636305	5.887928e-03	0.07673283	0.32100869
ndvi_se	516	0	0	0.02988000	0.5383143	0.5084343	129.2624	0.2501286	0.2505086	0.003408345	0.006695970	5.994277e-03	0.07742272	0.30906215
ndvi_sw	516	0	0	0.06418333	0.5460167	0.4818334	137.9656	0.2624285	0.2673752	0.003794193	0.007453999	7.428284e-03	0.08618749	0.32234668
precipitation_amt_mm	516	4	0	0.00000000	210.8300000	210.8300000	33150.8000	60.4700000	64.2457364	1.550429800	3.045944900	1.240378e+03	35.21899490	0.54819190
reanalysis_air_temp_c	516	0	0	21.48571429	28.4871429	7.0014286	12755.2814	24.6728571	24.7195376	0.051550270	0.101274670	1.371234e+00	1.17099695	0.04737131
reanalysis_avg_temp_c	516	0	0	21.74285714	29.7785714	8.0357143	13407.2500	25.9714286	25.9830426	0.058641250	0.115205480	1.774419e+00	1.33207303	0.05126701
reanalysis_dew_point_temp_c	516	0	0	16.93857143	25.3000000	8.3614286	11528.9786	22.7021429	22.3429817	0.062390040	0.122570290	2.008539e+00	1.41722931	0.06343063
reanalysis_max_air_temp_c	516	0	0	26.85000000	40.8500000	14.0000000	17509.3000	33.9000000	33.9327519	0.104904860	0.206094090	5.678595e+00	2.38298030	0.07022656
reanalysis_min_air_temp_c	516	0	0	13.75000000	22.8500000	9.1000000	10173.8000	19.9000000	19.7166667	0.073212540	0.143831970	2.765799e+00	1.66306926	0.08434840
reanalysis_precip_amt_kg_per_m2	516	1	0	0.00000000	362.0300000	362.0300000	29726.6900	46.4400000	57.6098640	2.213742000	4.349076000	2.528738e+03	50.28655500	0.87288100
reanalysis_relative_humidity_percent	516	0	0	57.78714286	98.6100000	40.8228571	45737.7843	90.9171429	88.6391168	0.333862100	0.655899140	5.751537e+01	7.58388911	0.08555917
reanalysis_sat_precip_amt_mm	516	4	0	0.00000000	210.8300000	210.8300000	33150.8000	60.4700000	64.2457364	1.550429800	3.045944900	1.240378e+03	35.21899490	0.54819190
reanalysis_specific_humidity_g_per_kg	516	0	0	12.11142857	20.4614286	8.3500000	8821.5929	17.4285714	17.0961102	0.063646430	0.125038560	2.090248e+00	1.44576888	0.08456712
reanalysis_tdtr_c	516	0	0	-269.43571429	-257.1214286	12.3142857	-136194.7000	-264.1857143	-263.9432170	0.107790290	0.211762750	5.995273e+00	2.44852460	-0.00927671
station_avg_temp_c	516	0	0	21.40000000	30.8000000	9.4000000	14196.6905	27.6000000	27.5129660	0.040295890	0.079164530	8.378593e-01	0.91534655	0.03326964
station_diur_temp_rng_c	516	0	0	5.20000000	15.8000000	10.6000000	5417.3317	10.5266667	10.4987048	0.068831980	0.135226000	2.444726e+00	1.56356205	0.14892904
station_max_temp_c	516	0	0	30.10000000	42.2000000	12.1000000	17530.2000	34.0000000	33.9732558	0.059783450	0.117449430	1.844215e+00	1.35801892	0.03997318
station_min_temp_c	516	0	0	14.70000000	24.2000000	9.5000000	10939.0000	21.3500000	21.1996124	0.055332080	0.108704350	1.579806e+00	1.25690321	0.05928897
station_precip_mm	516	19	0	0.00000000	543.3000000	543.3000000	31972.9000	44.7500000	61.9629840	2.768783000	5.439498000	3.955737e+03	62.89465200	1.01503600
total_cases	516	95	0	0.00000000	116.0000000	116.0000000	3920.0000	5.0000000	7.5968992	0.475303200	0.933771600	1.165712e+02	10.79681320	1.42121320

Here is some information on San Juan subsetted from the training data

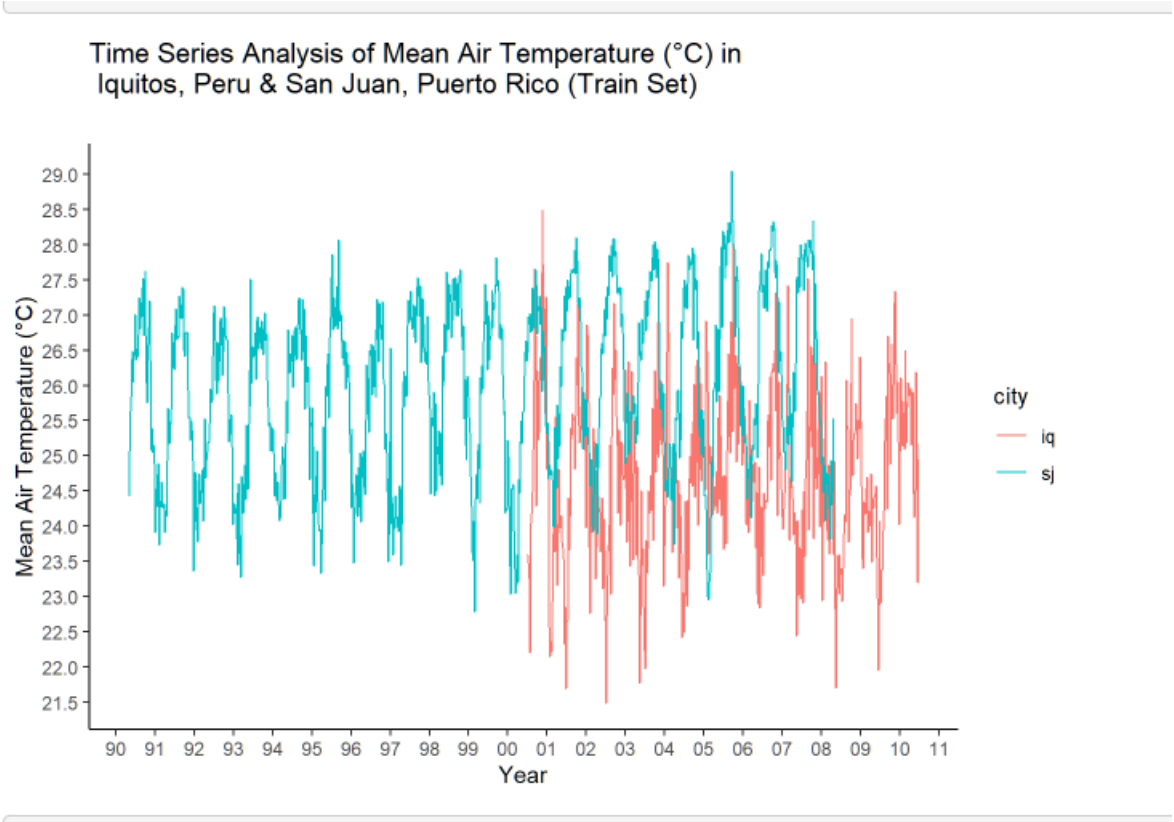
	nbr.val	nbr.null	nbr.na	min	max	range	sum	median	mean	SE.mean	CI.mean	var	std.dev	coef.var
ndvi_ne	930	0	0	-0.40625000	0.4934000	0.8996500	54.40049	0.0587750	0.05849515	0.003465400	0.006800921	1.116837e-02	0.10568051	1.80665429
ndvi_nw	930	0	0	-0.45610000	0.4371000	0.8932000	60.86540	0.0673875	0.06544667	0.003071183	0.006027260	8.771912e-03	0.09365849	1.43106568
ndvi_se	930	0	0	-0.01553333	0.3931286	0.4086619	165.14261	0.1767012	0.17757269	0.001869493	0.003668918	3.250353e-03	0.05701186	0.32106210
ndvi_sw	930	0	0	-0.06345714	0.3814200	0.4448771	155.00670	0.1677584	0.16667387	0.001831719	0.003594786	3.120329e-03	0.05585991	0.33514496
precipitation_amt_mm	930	238	0	0.00000000	390.6000000	390.6000000	32881.44000	20.6050000	35.35638700	1.461819000	2.868851000	1.987332e+03	44.57950200	1.26086100
reanalysis_air_temp_c	930	0	0	22.78857143	29.0500000	6.2614286	24192.69714	26.1042857	26.01365284	0.040544110	0.079568660	1.528757e+00	1.23642919	0.04753001
reanalysis_avg_temp_c	930	0	0	22.96428571	29.0142857	6.0500000	24298.03571	26.2285714	26.12692012	0.039960700	0.078423700	1.485077e+00	1.21863747	0.04664298
reanalysis_dew_point_temp_c	930	0	0	16.49285714	24.6457143	8.1528571	20422.35286	22.3142857	21.95951920	0.051480470	0.101031490	2.464722e+00	1.56994332	0.07149261
reanalysis_max_air_temp_c	930	0	0	24.65000000	31.1500000	6.5000000	26271.40000	28.3500000	28.24881720	0.041281830	0.081016450	1.584896e+00	1.25892666	0.04456564
reanalysis_min_air_temp_c	930	0	0	19.45000000	26.7500000	7.3000000	22461.20000	24.3500000	24.15182796	0.042455050	0.083318930	1.676261e+00	1.29470516	0.05360692
reanalysis_precip_amt_kg_per_m2	930	2	0	0.00000000	570.5000000	570.5000000	28332.84000	21.3000000	30.46541900	1.168290000	2.292793000	1.269358e+03	35.62805500	1.16945900
reanalysis_relative_humidity_percent	930	0	0	66.73571429	87.5757143	20.8400000	73068.40857	78.6678571	78.56818126	0.111145680	0.218125710	1.148863e+01	3.38948769	0.04314072
reanalysis_sat_precip_amt_mm	930	238	0	0.00000000	390.6000000	390.6000000	32881.44000	20.6050000	35.35638700	1.461819000	2.868851000	1.987332e+03	44.57950200	1.26086100
reanalysis_specific_humidity_g_per_kg	930	0	0	11.71571429	19.4400000	7.7242857	15393.74000	16.8457143	16.55240860	0.051184680	0.100451010	2.436481e+00	1.56092305	0.09430187
reanalysis_tdtr_c	930	0	0	-271.79285714	-268.7214286	3.0714286	-251689.37143	-270.6928571	-270.63373272	0.016359300	0.032105470	2.488928e-01	0.49889161	-0.00184342
station_avg_temp_c	930	0	0	22.84285714	30.0714286	7.2285714	25116.07143	27.2285714	27.00652842	0.046415200	0.091090800	2.003565e+00	1.41547346	0.05241227
station_diur_temp_rng_c	930	0	0	4.52857143	9.9142857	5.3857143	6284.35714	6.7571429	6.75737327	0.027413280	0.053799130	6.988838e-01	0.83599268	0.12371563
station_max_temp_c	930	0	0	26.70000000	35.6000000	8.9000000	29395.40000	31.7000000	31.60795699	0.056312380	0.110514210	2.949108e+00	1.71729665	0.05433115
station_min_temp_c	930	0	0	17.80000000	25.6000000	7.8000000	21018.60000	22.8000000	22.60064516	0.049392760	0.096934310	2.268869e+00	1.50627665	0.06664751
station_precip_mm	930	24	0	0.00000000	305.9000000	305.9000000	24910.50000	17.7500000	26.78548390	0.961631200	1.887221300	8.600032e+02	29.32581080	1.09483970
total_cases	930	4	0	0.00000000	461.0000000	461.0000000	31734.00000	19.0000000	34.12258100	1.688694000	3.314097000	2.652069e+03	51.49824200	1.50921300

5.2 Correlation, plots, & relationships of features against response variable

We create a time series plot of the `total_cases` variable in our training set. The test set does not have a `total_cases` column, so we omit the plot.



Upon viewing the figure above, we can make some assumptions. There has been a spike of cases around Q3 of 1994, with the highest number of cases reported around the 460 mark. Around this time, there are somewhat "jumps" in data. Between 2000-2010, there has been a surge of reported cases between the two cities, with majority of the numbers in double digits. We see an increase total cases between 2005-2006 and 2007-2008.
In Iquitos, case numbers are steady in the early 2000s. The number of cases in the city fall under 120. There have been a steady number of cases around 2008-2009. In San Juan, case numbers are much higher. This correlates with the fact that the disease carrier mosquitoes favor freshwater areas and saline waters for breeding sites, as mentioned in the literature review; San Juan is located in coastal Puerto Rico. The highest reported number of cases in the city were between 1994-1995, with one day reporting 460 cases. Between 1995-1998, a steady number of low cases are reported. Around 1998-1999, a spike in case numbers on a given day rise. In the early to mid 2000s, the frequency of cases on any given day increase, but numbers per day are not as high as previously. 2005-2008 see case numbers moderately spike.



This figure shows us how mean air temperature has fluctuated between both cities. The time series analysis of mean air temperature (°C) in the training set shows us that San Juan achieves higher air temperature compared to Iquitos. From the high number of cases between San Juan & Iquitos, our literatures suggest that high temperatures do in fact influence the number of cases between both regions.

The figure on the top left indicates how the response variable is paired with the first 6 climate features of our data; a linear regression line is fitted. The data is very dispersed, with the predictor variable showing weak relationships among the features. This is also indicated in the correlation matrix, where the `total_cases` variable show light heat maps.

Our analysis so far suggests that Multiple Linear Regression is not the best technique for this data. However, we try it anyways since that was the original model to use.

5.3 Multiple Linear Regression

To prepare our data for multiple linear regression, we split `train_df` into training & testing splits (80% & 20%, respectively). We fit the linear model using all possible climate features as the independent variables. Upon running, the most significant predictors in the model were the vegetation indices; multiple R^2 yielded a value of 0.1657 and an adjusted R^2 of 0.1519. These values are very low. We run the same regression using the subsetted data based on city. We get similar results on San Juan, but very low performance on Iquitos.

We then try backward elimination on `train_df` to eliminate candidate variables step by step, and choose the best model. The following variables were chosen as the final model: `ndvi_ne`, `ndvi_nw`, `ndvi_se`, `ndvi_sw`, `reanalysis_avg_temp_c`, `reanalysis_relative_humidity_percent`, `reanalysis_specific_humidity_g_per_kg`, `station_diur_temp_rng_c`, & `station_max_temp_c`. We achieved similar results from the first model.

5.4 Random Forest

Random forest creates multiple decision trees at a a time by taking select variables at random. It simultaneously develops multiple trees in combination and finally averages the error to bring out the best possible results. We use the package randomForest for our analysis.

We've created 9 models in total using this algorithm: 3 models on the enire train set, and 3 models for each city sets. First, we applied random forest on the entire train_df dataset, using the same test and train splits as before. The basic random forest model created 500 trees & selected 6 independent variables at random.

The second model applies is the same as the first, but uses the optimal number of trees which has the least Mean Squared Error. We then use this optimal tree number to prune the model. Pruning is a technique in machine learning that removes sections of the tree that provides low power in classifying instances. This in turn reduces the size of the decision trees. By reduction of over-fitting and reducing the final classifier's complexity, pruning improves on the accuracy of prediction.

The third model uses feature selection based on Node Purity. Node purity is defined as the ease of identifying the variables to a particular class or variable. The higher the Node Purity of that variable, the more useful it is in the model. We return the 6 variables with the highest node purity, and call it into the third model (Lesmeister, 2015).

The same three models are repeated in the Iquitos and San Juan split. Overall, the third regressor model in the entire train set was the best performing Random Forest model.

Refer to section 6.1 for an overview, description, and scores of each model (Random Forest & SVM).

5.5 Support Vector Machine

As SVM is non-parametric, it wont actually train the network. The SVM algorithm tries to plot all of the data in an n-dimensional hyper plane and applies the same logic on the test set, based on the reference created by the training set. The algorithm then tries to draw the boundary between the classes based on Support Vector machines. We use the package e1071 to perform our analysis. We try the grid approach to build multiple models at a time, so that we can pick the best model from all the developed models. We create 3 models: 1 for the entire train set, and 1 for each city subsets.

Out of the 3 models, Iquitos yielded the lowest MSE (Mean Square Error).

Refer to section 6.1 for performance statistics on the SVM models.

6 Results & Conclusion

6.1 Overview of Random Forest & SVM Models

Model Name	Algorithm	Description	MSE on train	% Variance covered on Train dataset	MSE on test
rf_df_regressor	Random Forest	Built on df_training dataset. We have used a basic Random forest model which has created 500 trees from the selection of 6 random independent variables.	1304.461	41.91	666.8971
rf_df_regressor2	Random Forest	After building the Random Forest algorithm, the algortithm tries to fit the model to the data, which may cause overfitting. As the model has generated 500 trees, it may generate a higher MSE. We developed the model based on the number of trees generating the least MSE	1260.36	43.87	620.1547
rf_df_regressor3	Random Forest	We have built this model with top 6 variables based on their Node Purity rating. We find it using the function VarImportance() . Node Purity indicates the ease of identifying the variable to a class or value. It's better to have high Node Purity.	1092.494	51.35	538.5882
rf_sj_regressor	Random Forest	Built on sj_training dataset. Similar to rf_df_regressor	1867.153	40.14	753.0785
rf_sj_regressor2	Random Forest	Built with trees pruned for minimum MSE	1938.588	37.85	737.931
rf_sj_regressor3	Random Forest	Built with the top 6 variables based on Node Purity	1464.458	53.05	790.4614
rf_iq_regressor	Random Forest	Built on iq_training dataset. Similar to rf_df_regressor	135.0008	-3.23	52.23506
rf_iq_regressor2	Random Forest	Built with trees pruned for minimum MSE	137.1541	-4.87	52.31801
rf_iq_regressor3	Random Forest	Built with top 5 Variables based on Node Purity	127.5276	2.49	60.39508
tunemodel_df	SVM	Built on df_training dataset with different Epsilon and Cost values; this has chosen the best possible model. Once the training completes, we can bring out the best tuned model and the parameters for it, such as Epsilon & Cost Once we pickup best model, we can use that for training	na	na	800.2505
tunemodel_sj	SVM	Similar to one above, but built on sj_training	na	na	856.6288
tunemodel_iq	SVM	Similar to one above, but built on iq_training	na	na	61.12607

We use Mean Square Error as a performance metric for our models:

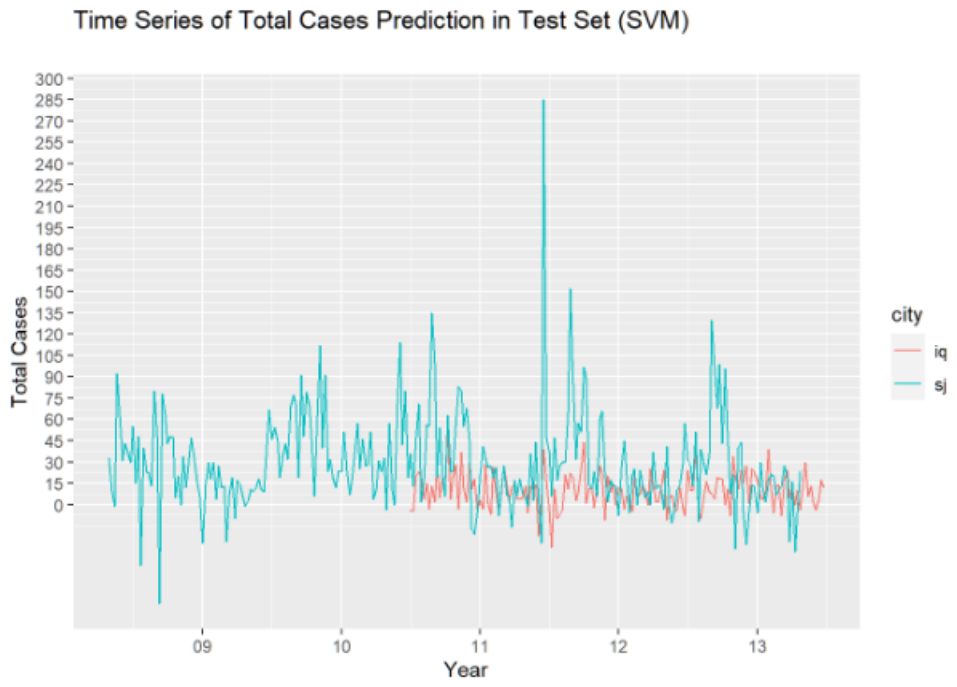
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where N is the number of data points, f_i is the value returned by the model, and y_i is the actual value for data point i. We calculate the scores of each model above using the test set sample split. The lower the MSE, the better the score.

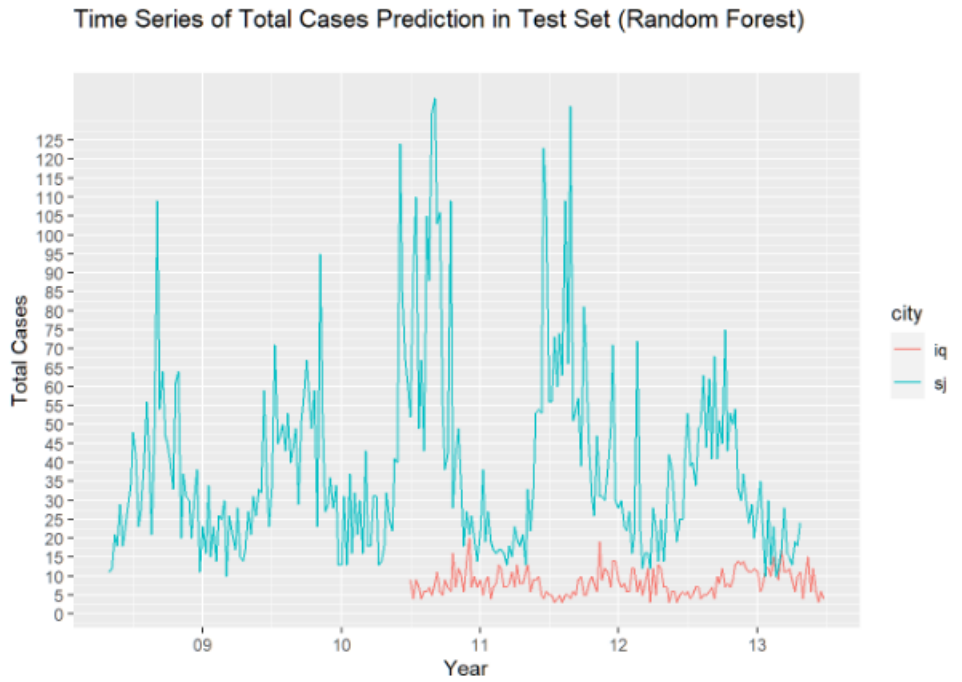
From above, rf_iq_regressor (Random Forest Regressor 1 for Iquitos) is the best model for predicting in the Iquitos test set. rf_sj_regressor2 (Random Forest Regressor 2 for San Juan) is the best model for predicting in the San Juan test set. rf_df_regressor3 (Random Forest Regressor 3 for Entire set) is the best model for predicting the entire test set.

6.2 Plotting predicted total cases

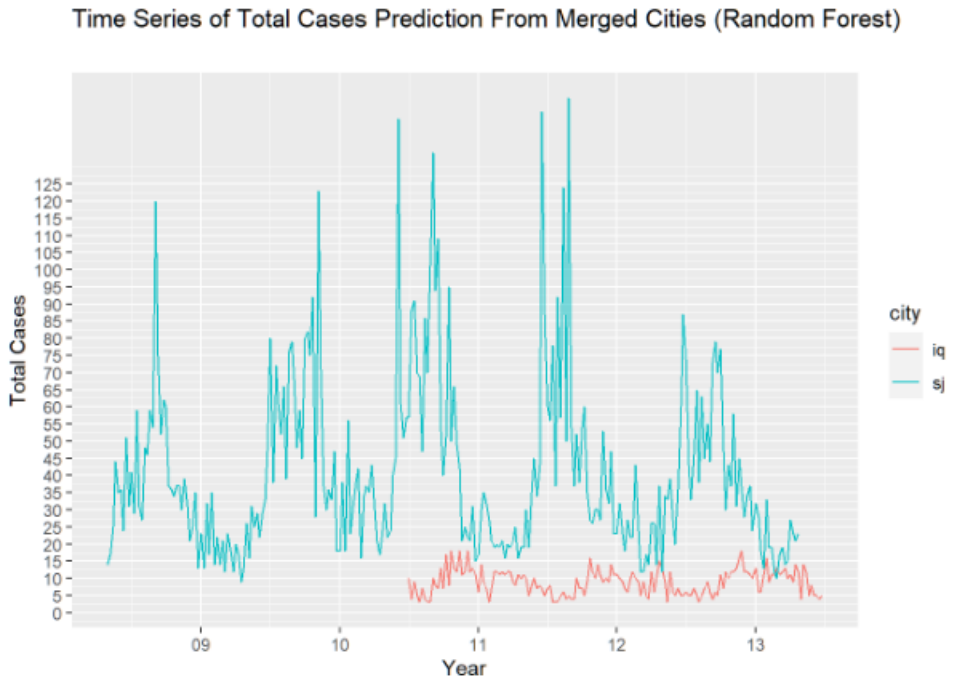
Plotting the prediction from the SVM model (not the optimal model):



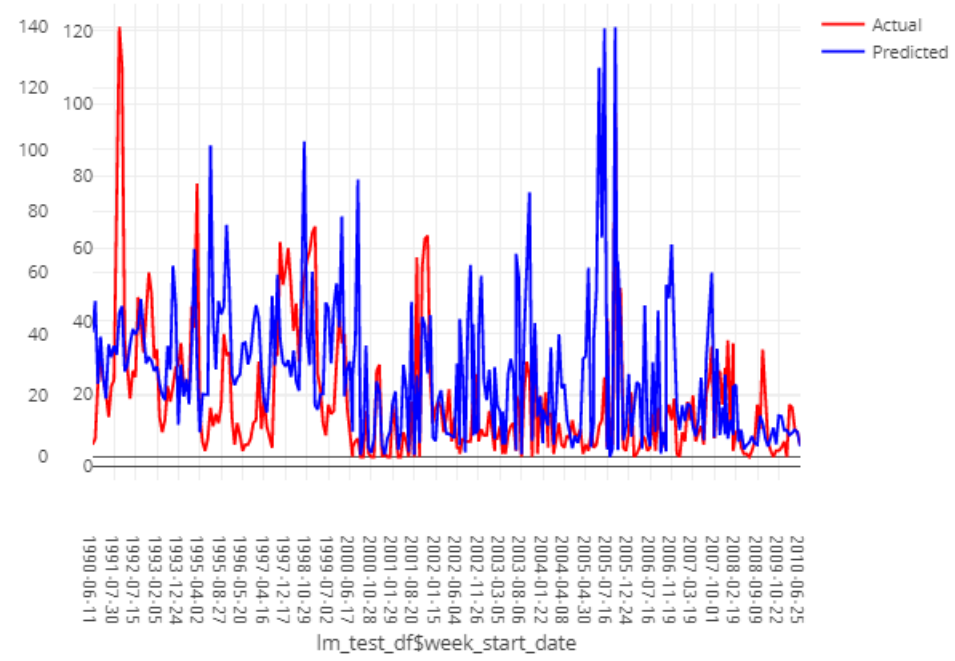
Plotting the predictions in the test set using the best model overall (Random Forest):



Plotting the predictions from merging the best models in both San Juan & Iquitos:



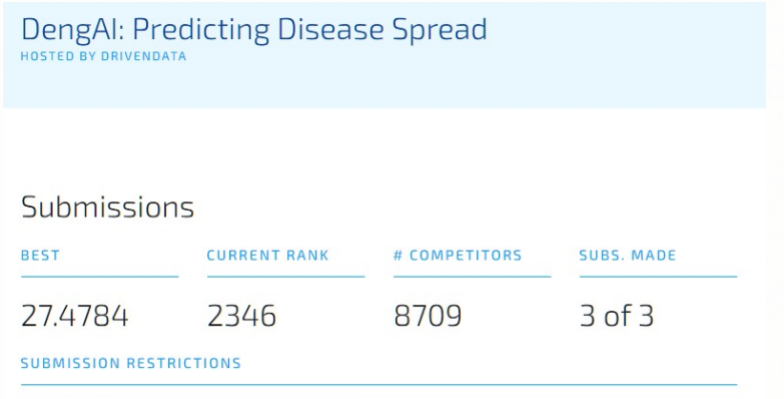
Actual vs. Predicted cases using the best model overall (Random Forest):



6.3 Final Statements & Further Results

Initially, our multiple linear regression was not the best fit for the data. Instead, we tried Random Forest and Support Vector Machine as they can detect any non-linear relationships present in the data. We were able to produce moderate results with these models.

The 3rd Random Forest regressor model on the entire training set yielded us the best results out of the 12 models trained. The winning model took 6 variables with the highest node purity. We have submitted our results to the website, and the following score was outputted:



We were able to place in the top 27% among 8709 competitors. Of course, we can improve this score by going back and refining our approach to the dataset, and the algorithms used.

In the beginning, we have imputed the missing variables of the climate features using the "Last Observation Carried Foward" technique. In an updated version of this project, we can inspect each attribute closely, and determine how to impute their missing values. Moreover, we can try normalizing the data since each feature has varying range.

Although we have written about Principal Component Analysis, we did not perform it in our data. In order to implement this technique, we would first have to normalize the data. When running our initial analysis, we had very weak correlations among the response variable. Implementing PCA would find the best characteristics and determine the best linear combinations among variables.

Artificial Neural Networks (ANN) may benefit this project as neural networks can identify the hidden patterns within the data. As the data is time series, we can also explore ARIMA (Auto Regressive Integrated Moving Average) models as well.

There were a lot of steep learning curves in regards to this project. We did not produce the best code and/or results related to the goal of the project. However, we were able to explore patterns in our data that derive from the literatures studied. In general, analyzing climate patterns in areas with high infection rates will help us determine how these vector borne diseases and carriers behave in the future. By determining these patterns, humans are more equipped to prevent such outbreaks. Likewise, the idea of analyzing social and phys-

ical patterns among humans during the COVID-19 pandemic (whether they are traveling, isolating, social distancing, etc...) will influence how the outbreak behaves in the future.

References

Ahmed, S. A., & Siddiqui, J. S. (2014, Dec). Principal component analysis to explore climatic variability and dengue outbreak in lahore. *Pakistan Journal of Statistics and Operation Research*, *10*(2), 247. doi: 10.18187/pjsor.v10i2.686

Anggraeni, W., Nurmasari, R., Riksakomara, E., Samopa, F., Wibowo, R. P., T., L. C., & Pujiadi. (2017). Modified regression approach for predicting number of dengue fever incidents in malang indonesia. *Procedia Computer Science*, *124*, 142–150. doi: 10.1016/j.procs.2017.12.140

Dengue. (2020, Feb). Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/dengue/

Donateli, C. P., Einloft, A. B. D. N., Junior, A. L. C., Cotta, R. M. M., & Costa, G. D. D. (2019, Apr). Endemic disease control agents' perception on the fight against aedes aegypti and the prevention of arbovirus infections in brazil. *PLOS Neglected Tropical Diseases*, *13*(10). doi: 10.1371/journal.pntd.0007741

Ferreira-De-Brito, A., Ribeiro, I. P., Miranda, R. M. D., Fernandes, R. S., Campos, S. S., Silva, K. A. B. D., ... et al. (2016, Mar). First detection of natural infection of aedes aegypti with zika virus in brazil and throughout south america. *Memórias do Instituto Oswaldo Cruz*, *111*(10), 655–658. doi: 10.1590/0074-02760160332

Hosangadi, D. (2019, Mar). *The global rise of dengue infections.* Outbreak Observatory. Retrieved from https://www.outbreakobservatory.org/outbreakthursday-1/3/21/2019/the-global-rise-of-dengue-infections

How to prevent the spread of the mosquito that causes dengue. (n.d.). Retrieved from https://www.cdc.gov/dengue/resources/vectorcontrolsheetdengue.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: with applications in r.* Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*(2065). doi: 10.1098/rsta.2015.0202

Kuhn, M. (2020). caret: Classification and regression training [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=caret (R package version 6.0-86)

Lateef, Z. (2019, May). *Support vector machine in r: Using svm to predict heart diseases.* Retrieved from https://www.edureka.co/blog/support-vector-machine-in-r/

Laureano-Rosario, A., Duncan, A., Mendez-Lazaro, P., Garcia-Rejon, J., Gomez-Carro, S., Farfan-Ale, J., ... Muller-Karger, F. (2018, May). Application of artificial neural networks for dengue fever outbreak predictions in the northwest coast of yucatan, mexico and san juan, puerto rico. *Tropical Medicine and Infectious Disease*, *3*(1), 5. doi: 10.3390/tropicalmed3010005

Lesmeister, C. (2015). *Mastering machine learning with r: master machine learning techniques with r to deliver complex and robust projects*. Packt Publishing.

Méndez-Lázaro, P., Muller-Karger, F., Otis, D., Mccarthy, M., & Peña-Orellana, M. (2014, Nov). Assessing climate variability effects on dengue incidence in san juan, puerto rico. *International Journal of Environmental Research and Public Health*, *11*(9), 9409–9428. doi: 10.3390/ijerph110909409

National oceanic and atmospheric administration. (2020, Mar). Wikimedia Foundation. Retrieved from https://en.wikipedia.org/wiki/National_Oceanic_and_Atmospheric_Administration

Prabhakaran, S. (2018, May). *Caret package - a complete guide to build machine learning in r.* Retrieved from https://www.machinelearningplus.com/machine-learning/caret-package/

Random forests. (n.d.). Retrieved from https://uc-r.github.io/random_forests

Salgueiro, P., Serrano, C., Gomes, B., Alves, J., Sousa, C. A., Abecasis, A., & Pinto, J. (2019, Mar). Phylogeography and invasion history of aedes aegypti , the dengue and zika mosquito vector in cape verde islands (west africa). *Evolutionary Applications*, *12*(9), 1797–1811. doi: 10.1111/eva.12834

Salles, T. S., Sá-Guimarães, T. D. E., Alvarenga, E. S. L. D., Guimarães-Ribeiro, V., Meneses, M. D. F. D., Castro-Salles, P. F. D., ... et al. (2018). History, epidemiology and diagnostics of dengue in the american and brazilian contexts: a review. *Parasites Vectors*, *11*(1). doi: 10.1186/s13071-018-2830-8

Stoddard, S. T., Wearing, H. J., Reiner, R. C., Morrison, A. C., Astete, H., Vilcarromero, S., ... et al. (2014). Long-term and seasonal dynamics of dengue in iquitos, peru. *PLoS Neglected Tropical Diseases*, *8*(7). doi: 10.1371/journal.pntd.0003003

What is dengue? (2017, Jan). World Health Organization. Retrieved from https://www.who.int/denguecontrol/disease/en/