# Literature Review:

Dengue Fever (DF) is caused by dengue virus (DINV), quiet often occurs in the tropical and subtropical zones around the world. Just like Covid-19 it has flue like symptoms and might cause fever, rash, muscle pain and joint pain. If DF is serious it might cause severe bleeding, pressure, low blood and even death.

The virus is transmitted to humans through the bites of infected female mosquitoes, primarily the ***Aedes aegypti mosquito***. Other species within the Aedes genus can also act as vectors, but their contribution is secondary to Aedes aegypti. The Aedes genus are the type of mosquitoes are liable for spreading and transmitting Dengue Virus, and some other arboviruses such as yellow fever virus (YFV), chikungunya virus (CHIKV), and the infamous Zika virus (ZIKV) around the world. During the trans-Atlantic voyages between 1500s and 1700s, the Aedes aegypti arrived in the New World. The Dengue Virus (DV) has become a real issue in South America as the mosquito born disease in the continent.

**Article #1 DengAI Predicting Disease Spread ML CS 6375.501**

Based on an estimation of a model, it indicates that about 390 million cases per year for dengue virus infection (95% credible interval 284–528 million) [Bhatt, S., et al Nature, 2013]*,* of which 96 million (67–136 million) manifest clinically (with any severity of disease). 3.9 billion people are at risk of infection with Dengue Virus per research on DF prevention [J. E Cogan, 23, June 2020]

Per this Article they have focused on gradient boosting from all the other ML techniques they have tested. Using gradient boosting an ensemble technique which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Per scope of the project, the target was to achieve the best Mean Absolute Error and their model received a value of 4.27 for city Iquitos and 11.41 for San Juan. The Model was able to predict a better result for Iquitos. They have other techniques such as KNN, but they were getting higher MAE for both the cities.

**Article #2 Using LASSO to predict Dengue occurrence by Preethi Subramanian**

Although it is a complex relationship between Dengue Fever and climate, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have [significant public health implications worldwide](http://rstb.royalsocietypublishing.org/content/370/1665/20140135.full)[[Lindsay P. Campbell](https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0135), 2015].

This article focused on predicting the dengue cases using the (LASSO) Least Absolute Shrinkage Selector Operator. “LASSO is a type of penalized regression which can perform feature selection to prevent over-fitting”. It overcome the weaknesses of the Artificial Neural network (ANN) methodologies using time series forecasting models and other models. “Multiple LASSO models were created for each disease in different countries for 1-26 weeks

ahead. The optimal constraint parameters were determined using ten-fold cross validation in R. The performance of the other models waned beyond 4 weeks’ horizon” [Dr.Preethi Subramanian, 2019]. The meteorological variables were found to add more value in temperate region where cyclic patterns were observed for the diseases and climatic data, in comparison to the more stable climate in equatorial region.

**Article #3** **Development of a mechanistic dengue simulation model for Guangzhou**

This Article is on Chinese Lab data, I wanted to compare and get insights on a completely different data set and gain more knowledge on the models applied for prediction. Dengue infection in China has increased dramatically in recent years. Guangdong province

(main city Guangzhou) accounted for more than 94% of all dengue cases in the 2014 outbreak. In China, dengue cases have been recorded each year for the past 25 years. [Cheng Q et al. (2016)].

The data was taken from 4 different serotype of Dengue disease. All tests were done in between 2004 and 2010 with controlled temperature and supervised but mimicking real world climate. The Correlation coefficient (*r*) values were the following (0.28, 0.37. 0.32 and 0.75). “The validity of the dengue simulation model was analyzed by using cross-correlation analysis to compare the simulated case data with the reported dengue cases”. The best model fit that was considered for this study had the highest positive predictive value (PPV) of 0.83. Simple linear regression was used to test the probability of outbreaks occurring each year in the same season in Guangzhou.

After evaluating the results, it was concluded the simulation model did not work using the multiple serotypes. “First, there is a possible existing bias in the reporting of actual dengue cases which can lead to under-reporting of the magnitude of an outbreak. Another reason is the set simulated population size which differs from the actual population size, due to computer simulation capacities of the dengue model used” [G. Mincham 2019].

**Article #4** **DengAI: Predicting Disease Spread - Adharsh Rajendran, Mithil Gotarne**

Recently dengue fever is spreading all around the world, historically the disease has been most prevalent in Southeast Asia and Pacific regions. These days nearly billion Dengue fever cases has been reported in Latin America [Luis Villar, M.D, 2015]

1. Random Forest

A useful tool to learn about feature analysis and comparison.

2. XGBoost

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning

3. Neural Networks

This basic tool uses classification and clustering tools. As this tool is a developed pattern recognition device, we decided practicing with this would provide a great learning experience. On this article three different ML models were used to see the best performance. “Each of the models had very similar scores ranging from 25 to 32. The only major difference created was separating the cities ‘datasets during the preprocess. The initial assumption was XGBoost would overwhelm the other two methods chosen. However, it only slightly outperformed.” [**Github - Adharsh Rajendran, 2019]**

Through previous observations and researches, it is found out that dengue outbreaks occur mostly in some weeks of the year. If we can predict the cases during the peak season along with the factors that are most common in this season can help the health officials take proper measurements before the season arrives to control the situation.

# Dataset

I have downloaded the data set from Using environmental data collected by various **U.S. Federal Government agencies**—from the [**Centers for Disease Control and Prevention**](http://www.cdc.gov/)**to the**[**National Oceanic and Atmospheric Administration**](http://www.noaa.gov/)**in the**[**U.S. Department of Commerce**](https://www.commerce.gov/). This project is part of a Data Science competition on drivendata.org webpage

**Iquitos, Peru**

Torres, Orduna, Piña-Pozas, Vázquez-Vega and Sarti [Torres, J.R..2017] distinguished that Iquitos has equatorial climate with and dengue outbreak in rainy season around March. The Andean area experienced heavy downpour induced by the La Niña phenomenon in early 2011 as reported by the International Federation of Red Cross and Red Crescent Societies (IFRC) [Peru and Bolivia, 2011]. Consequently, Loreto (the region where Iquitos was located) was the worst hit area in the country’s 2011 most serious dengue crisis IFRC [Peru and Bolivia, 2011].

The city of Iquitos sits at the Amazon Rivers of northeast Peru and the confluence of [Nanay, Itaya. Stoddard et. al (2014)] research data reported from a laboratory confirmed dengue dynamics between 2000 and 2010. Their studies are split within 3 seasons: trimester I, II, and III. The warmest temperature recorded with Maximum and mean temperature were between November and April. Over all the years, rainfall was highest between 2003 and 2008 with significantly less rainfall subsequent years.

![A picture containing chart

Description automatically generated]()

**San Juan, Puerto Rico**

Sougata, Acebedo and Chua [Sougata, D. 2017] mentioned that San Juan reported that due to higher tropical monsoon climate and populations density there was an increase of Dengue cases which stood at 2.5 times more compared to Iquitos, Peru. Puerto Rico went through an epidemic during the period of 2007 and 2010 but systematic review conducted

[Torres, J.R..2017] did not support local weather as a crucial factor in explaining the changes in the annual case in Puerto Rico.

San Juan in Puerto Rico annually reports to have an average air surface temp of 24-29 °C, with average precipitation of about 1800 mm. During the dry season (0-50 mm occurs 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 AI networks (ANNs) to predict dengue fever cases between 1994-2012

**Variables (parameters):**

Below is the list of meteorological parameters and an example of assigned values

The following set of information has been provided on a (year, weekofyear) timescale:

(Where appropriate, units are provided as a \_unit suffix on the feature name.)

**City and date indicators**

* city – City abbreviations: sj for San Juan and iq for Iquitos
* week\_start\_date – Date given in yyyy-mm-dd format

**NOAA's GHCN**[daily climate data](https://www.ncdc.noaa.gov/oa/climate/ghcn-daily/)**weather station measurements**

* station\_max\_temp\_c – Maximum temperature in °C
* station\_min\_temp\_c – Minimum temperature in °C
* station\_avg\_temp\_c – Average temperature in °C
* station\_precip\_mm – Total precipitation
* station\_diur\_temp\_rng\_c – Diurnal temperature range in °C

**PERSIANN**[satellite precipitation measurements](http://www.ncdc.noaa.gov/cdr/operationalcdrs.html)**(0.25x0.25 degree scale)**

* precipitation\_amt\_mm – Total precipitation

**NOAA's NCEP**[Climate Forecast System Reanalysis](http://rda.ucar.edu/datasets/ds093.0/#metadata/detailed.html?_do=y)**measurements (0.5x0.5 degree scale)**

* reanalysis\_sat\_precip\_amt\_mm – Total precipitation
* reanalysis\_dew\_point\_temp\_k – Mean dew point temperature in °C
* reanalysis\_air\_temp\_k – Mean air temperature in °C
* 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 in
* reanalysis\_min\_air\_temp\_k – Minimum air temperature in
* reanalysis\_avg\_temp\_k – Average air temperature in
* reanalysis\_tdtr\_k – Diurnal temperature range

**Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's**[CDR Normalized Difference Vegetation Index](https://www.ncdc.noaa.gov/cdr)**(0.5x0.5 degree scale) measurements**

* 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

**Feature data example**

For example, a single row in the dataset, indexed by (city, year, weekofyear):

(sj, 1994, 18), has these values:

* week\_start\_date 5/7/1994
* total\_cases 22
* station\_max\_temp\_c 33.3 °C
* station\_avg\_temp\_c 27.75714286 °C
* station\_precip\_mm 10.5
* station\_min\_temp\_c 22.8 °C
* station\_diur\_temp\_rng\_c 7.7 °C
* precipitation\_amt\_mm 68
* reanalysis\_sat\_precip\_amt\_mm 68
* reanalysis\_dew\_point\_temp\_k 295.2357143
* reanalysis\_air\_temp\_k 298.9271429
* reanalysis\_relative\_humidity\_percent 80.35285714
* reanalysis\_specific\_humidity\_g\_per\_kg 16.62142857
* reanalysis\_precip\_amt\_kg\_per\_m2 14.1
* reanalysis\_max\_air\_temp\_k 301.1
* reanalysis\_min\_air\_temp\_k 297
* reanalysis\_avg\_temp\_k 299.0928571
* reanalysis\_tdtr\_k 2.671428571
* ndvi\_location\_1 0.1644143
* ndvi\_location\_2 0.0652
* ndvi\_location\_3 0.1321429
* ndvi\_location\_4 0.08175

# Approach

Specific goal for this project is to predict dengue cases for San Juan, Puerto Rico and Iquitos, Peru. Hence, a model needs to be developed by analyzing the relationship between dengue cases and climate data to predict peak time of dengue outbreak and maximum weekly incidence. Understanding these patterns in will help to better develop effective public health strategies to combat the potential increases in dengue outbreak.

**Multi Linear regression** is a standard statistical approach that allows questions that consider the role(s) of multiple independent variables in a single dependent variable to be answered (Nathans, 2012). Three types of work can be done by MLR: (1) defining relationships between dependent variables and independent variables, (2) Estimating the values of the dependent variables based on independent variables' observed values, (3) identifying independent variables affecting dependent variables (Schneider 2010; Jeon, 2015*).*

**The K-nearest neighbor (KNN)** algorithm is one of the simplest and earliest classification algorithms [29]. It can be thought a simpler version of a Naïve Bayes (NB) classifier. Unlike the NB technique, the KNN algorithm does not require to consider probability values.

**Support vector machine (SVM)** algorithm can classify both linear and non-linear data. It first maps each data item into an n-dimensional feature space where n is the number of features. It then identifies the hyperplane that separates the data items into two classes while maximizing the marginal distance for both classes and minimizing the classification errors.

Read dataset

Data Profiling

Split the Dataset

Between cities

Data Quality

Cleansing data by removing NULL values

Train the model with Training Data

**Iquitos, Peru / San Juan, Puerto Rico**

The K-nearest neighbor

Support Vector Machine

Multiple Linear Regression

Test Data

Compare Models

Predicted Cases

Results and Conclusion

Reference:

* (J. E Cogan, 23, June 2020) [*https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue#.YCBhJOTGLmI.gmail*](https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue#.YCBhJOTGLmI.gmail)
* Bhatt, S., et al., The global distribution and burden of dengue. Nature, 2013. 496(7446): p. 504–507.
* Brady, O.J., et al., Refining the global spatial limits of dengue virus transmission by evidence-based consensus. PLOS Neglected Tropical Diseases, 2012. 6(8): p. e1760.
* [Lindsay P. Campbell](https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0135), [Caylor Luther](https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0135), [David Moo-Llanes](https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0135), [Janine M. Ramsey](https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0135), [Rogelio Danis-Lozano](https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0135) and [A. Townsend Peterson](https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0135) Published:05 April 2015. [*https://doi.org/10.1098/rstb.2014.0135*](https://doi.org/10.1098/rstb.2014.0135)
* G. Mincham 2019 <https://www.cambridge.org/core/journals/epidemiology-and-infection/article/development-of-a-mechanistic-dengue-simulation-model-for-guangzhou/7AA81F08BA5C6F9FA74F9F50E09B9388>
* Luis Villar, M.D., Gustavo Horacio Dayan, M.D., José Luis Arredondo-García, M.D., Doris Maribel Rivera, M.D., Rivaldo Cunha, M.D., Carmen Deseda, ...N Engl J Med https://www.nejm.org/doi/10.1056/NEJMoa1411037
* Nathans, L. L., Oswald, F. L. and Nimon, K. (2012) ‘Interpreting multiple linear regression: A guidebook of variable importance’, Practical Assessment, Research & Evaluation, 17(9).
* Dr.Preethi Subramanian, 2019 ..Periodicals of Engineering and Natural Sciences <http://pen.ius.edu.ba/index.php/pen/article/viewFile/442/328>
* Schneider, A., Hommel, G. and Blettner, M. (2010) ‘Linear regression analysis: part 14 of a series on evaluation of scientific publications’, Deutsches Ärzteblatt International, 107(44), p. 776.
* <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226841>
* <https://medium.com/python-in-plain-english/how-i-solved-the-dengue-fever-competition-question-using-adaboostingregressor-a9e82cb25f95>
* Torres,J.R., Orduna,T.A., Piña-Pozas,M., Vázquez-Vega,D. and Sarti,E. Epidemiological Characteristics of Dengue Disease in Latin America and in the Caribbean: A Systematic Review of the Literature. Journal of Tropical Medicine, Vol. 2017, March, 2017. [Online serial]. Available at: *https://doi.org/10.1155/2017/8045435*. [Accessed Dec. 22, 2018].
* Peru and Bolivia: Dengue outbreak - DREF operation n° MDR46001 21 September 2011. The International Federation of Red Cross and Red Crescent Societies, 2011. [Online]. Available at: https://reliefweb.int/sites/reliefweb.int/files/resources/Full\_Report\_2391.pdf. [Accessed Dec. 5, 2018].
* Peru and Bolivia: Dengue outbreak - DREF operation n° MDR46001 18 February 2011. The International Federation of Red Cross and Red Crescent Societies, 2011. [Online]. Available at: *https://reliefweb.int/sites/reliefweb.int/files/resources/1507AAFA08C7F39E8525783B00764030-Full\_Report.pdf*. [Accessed Dec. 5, 2018].
* Sougata,D., Acebedo,C.M.L. and Chua,M.C.H. An ensemble prediction approach to weekly Dengue cases forecasting based on climatic and terrain conditions. J. Health. Soc. Sci., Vol. 2, no. 3, pp. 257-272, 2017.
* Laureano-Rosario, A., Duncan, A., Mendez-Lazaro, P., Garcia-Rejon, J., Gomez-Carro, S.,Farfan-Ale, J., . . . Muller-Karger, F. (2018, May). Application of arti\_cial 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
* 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
* Guzman MG and Harris E (2015) Dengue. The Lancet 385, 453–465.
* Cheng Q et al. (2016) Climate and the timing of imported cases as determinants of the dengue outbreak in Guangzhou, 2014: evidence from a mathematical model. PLoS Neglected Tropical Diseases 10, p. e0004417
* <https://github.com/mithilgotarne/DengAI-Predicting-Disease-Spread>
* Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7.