Capstone 1 Project Proposal:

Predicting the spread of dengue fever using climate reports

**The goal**: To predict the total\_cases label for each city, year, weekofday in the test set.

Using data of two cities, San Juan and Iquitos, with test data for each city, spanning 5 and 3 years respectively, we will make predictions for the two cities whether they will report cases of dengue fever in the test set. The test set is a pure future hold-out, meaning the test data are sequential and non-overlapping with any of the training data.

**Usefulness of the proposal**: This project will be used to develop a classification model to predict whether these cities, in this scope, or future cities depending on the performance of the model, will be likely to report cases of dengue fever based on past reports and the climate reports of the location. If the model performs well, we can expand the scope of the project to include more problem areas of the world to develop an accurate predictor of dengue outbreak to better prepare World Health Organizations, WHO, to be more proactive instead of reactive to outbreaks to be able to provide more support to locations predicted to have a possible outbreak.

**The Data:**

The training set includes the following features on a year, week of year timescale:

Daily climate data measurementsof the region from the National Oceanic and Atmospheric Administration’s Global Historical Climatology Network, GHCN:

|  |  |
| --- | --- |
| 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 |

PERSIANN satellite precipitation measurements:

|  |  |
| --- | --- |
| precipitation\_amt\_mm | Total Precipitation |

NOAA’s National Centers for Enviornmental Predications, NCEP, climate forecast system reanalysis measurements:

|  |  |
| --- | --- |
| reanalysis\_sat\_precip\_amt\_mm | Total precipitation |
| reanalysis\_dew\_point\_temp\_k | Mean dew point temp |
| reanalysis\_air\_temp\_k | Mean air temp |
| reanalysis\_relative\_humidity\_precent | 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 | Max air temp |
| reanalysis\_min\_air\_temp\_k | Min air temp |
| reanalysis\_avg\_temp\_k | Avg air temp |
| reanalysis\_tdtr\_k | Dinural temp range |

NOAA’s CDR Normalized difference vegetation index, NDVI, 0.5x0.5 degree scale:

|  |  |
| --- | --- |
| nvdi\_se | Pixel southeast of city centroid |
| nvdi\_sw | Pixel southwest of city centroid |
| nvdi\_ne | Pixel northeast of city centroid |
| nvdi\_nw | Pixel northwest of city centroid |

The training set also includes labels for each city in that week of year indicating whether there were reported cases of dengue or not.

**The approach**:

The first thing to look for in determining the key features is the correlation between these features and the result, reposted cases or no reported cases. The outcome of a given week is likely not a result of one single feature, so we will explore the combinations of features as well as looking at the time of year to see if there is a time pattern the corresponds to climate readings, vegetation reports, and mosquito breeding cycles to determine a pattern. Examining the reports of a city in a given span of time, we can see how recent reports are indicative of future outbreaks, one city may be better at public outreach, protecting and warning citizens of outbreaks versus the other city in the set.

**Deliverables:**

The product of this project will be the code used to predict the labels in the training set, as well as the reports and charts with the findings of correlations amongst the features in this dataset. Possible recommendations to organizations like the WHO on predictive best-practices for preventing and being ready for future outbreaks may be provided contingent on the performance of the model.