The data for this project was downloaded from DrivenData.org as part of their “DengAI: Predicting Disease Spread” challenge. In addition to the number of cases at each location, the data includes information on temperature, precipitation, humidity, vegetation, and what time of the year the data was obtained. Using these features a table was created with features chosen as predictors and number of dengue cases as response. The table is a 1456x24 matrix where 1456 is the number of training examples and 24 is the number of features.

Then this table is imported to Regression Learner app in MATLAB. A 5 fold cross validation scheme is chosen for training. To observe the correlation between total number of cases (y variable) and different features (x variables) several scatter plots were created using blue color. This plot is called response plot. Then the training was done using all the available regression models, e.g., Linear Regression, Step Wise Linear Regression, Fine Tree, Linear SVM, Quadratic SVM etc. The performance chosen for evaluating this model is RMSE (Root Mean Square Error). Comparing RMSE of different models we observed Step Wise Linear Regression performed best with lowest RMSE of 26.079. So, Step Wise Linear Regression was chosen for future prediction of test data set.

Predicted outputs (total number of cases) were plotted in scatter plot against different features using yellow color. Now we can compare these response plots with previous response plots. Then predicted vs actual plot and residual plotss were also observed to measure performance. Then using global data from NASA satellites as test data, dengue predictions for different regions of earth were calculated and a heat map was generated. So, by this way, a very basic framework was implemented to forcast dengue outbreak.