DengAI : Predicting Disease Spread

Bhavish Khanna Narayanan , Jayashree Sankar Kumar , Preethika Anand

Introduction & Problem Description

Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world. When the disease level is mild , the symptoms are similar to the flu: fever, rash, and muscle and joint pain. When it is severe, dengue fever can cause severe bleeding, low blood pressure, and even death.

Since it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation. Scientists believe that climate change is likely to produce distributional shifts that will have significant impact on public health worldwide.

The U.S. Federal Government agencies collected data from the Centers for Disease Control and Prevention the National Oceanic and Atmospheric Administration in the U.S. Department of Commerce to predict the number of dengue fever cases reported each week in San Juan, Puerto Rico and Iquitos, Peru.

Related Work :

We came across a IEEE paper titled "Developing Disease Risk Prediction Model Based on Environmental Factors" . According to this paper , analyzing the effects of various environmental factors on human diseases is one of the important issues in recent bioinformatics studies. Several environmental factors regarding Type-2 diabetes are investigated and some of them are selected to develop an analytical model of disease risk prediction.

Data was obtained from the Korean National Institute of Health (KNIH). Initially the environmental factors were preprocessed into categorical values and then max/min odds ratios was calculated for all the categorized environmental factors. Top ranked factors were chosen as input features for the prediction model. The disease risk prediction model was developed with SVM classifiers.

Dataset Description:

The dengue fever dataset is collected by various U.S. Federal Government agencies and used to predict the number of dengue fever cases reported each week in the following two cities : San Juan, Puerto Rico and Iquitos, Peru . The prediction is based on environmental variables describing changes in temperature, precipitation, vegetation, and more.

There are a total of 22 attributes & 1457 rows. The a single row in the dataset is indexed by (city, year, week of year). The different categories of attributes include :

1. City and date indicators
2. Daily climate data weather station measurement
3. Satellite precipitation measurements
4. Climate Forecast System Reanalysis measurements
5. Normalized Difference Vegetation Index measurements

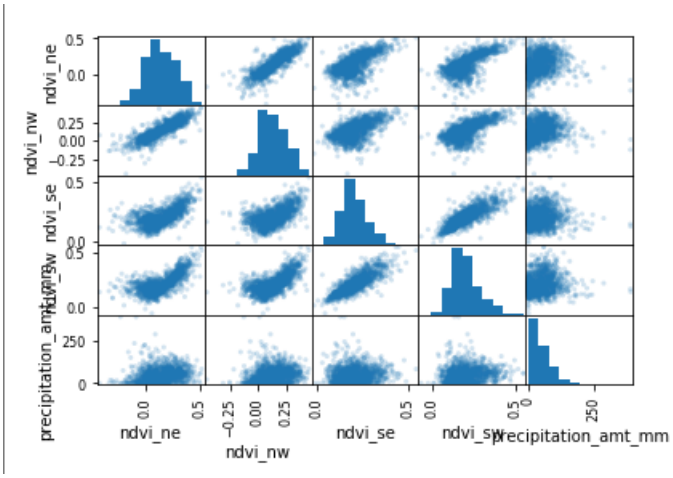
looks like copied!
Instead, we can show the range of data in a tabular form for each column.

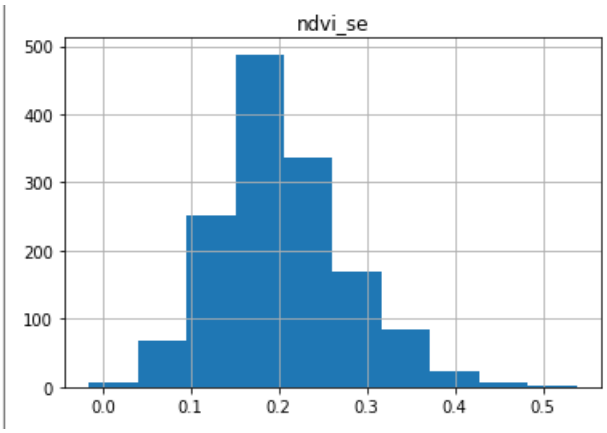
Feature data example

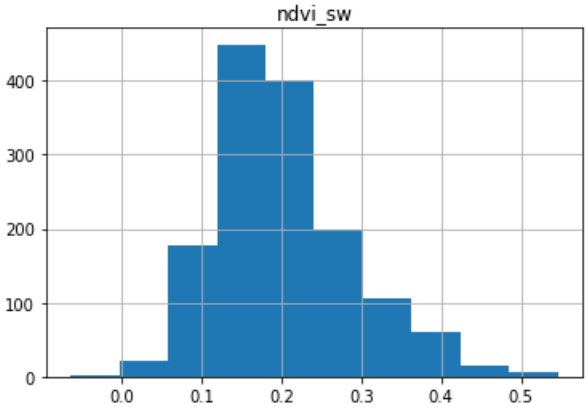
For example, a single row in the dataset, indexed by (city, year, weekofyear): (sj, 1994, 18), has these values:

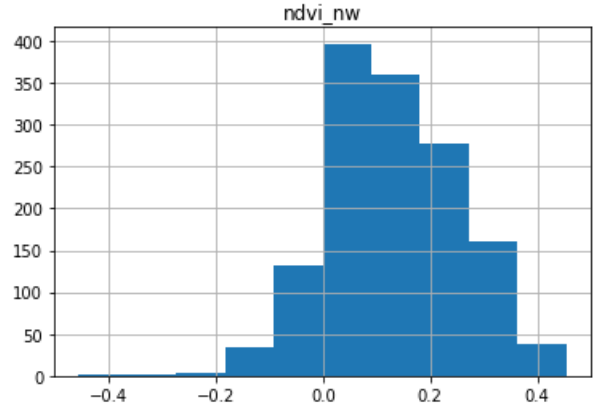
week_start_date	1994-05-07
total_cases	22
station_max_temp_c	33.3
station_avg_temp_c	27.7571428571
station_precip_mm	10.5
station_min_temp_c	22.8
station_diur_temp_rng_c	7.7
precipitation_amt_mm	68.0
reanalysis_sat_precip_amt_mm	68.0
reanalysis_dew_point_temp_k	295.235714286
reanalysis_air_temp_k	298.927142857
reanalysis_relative_humidity_percent	80.3528571429
reanalysis_specific_humidity_g_per_kg	16.6214285714
reanalysis_precip_amt_kg_per_m2	14.1
reanalysis_max_air_temp_k	301.1
reanalysis_min_air_temp_k	297.0
reanalysis_avg_temp_k	299.092857143
reanalysis_tdtr_k	2.67142857143
ndvi_location_1	0.1644143
ndvi_location_2	0.0652
ndvi_location_3	0.1321429
ndvi_location_4	0.08175

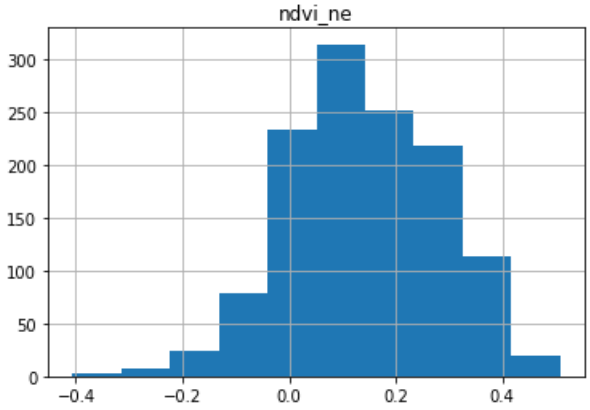
Data distribution of few features :

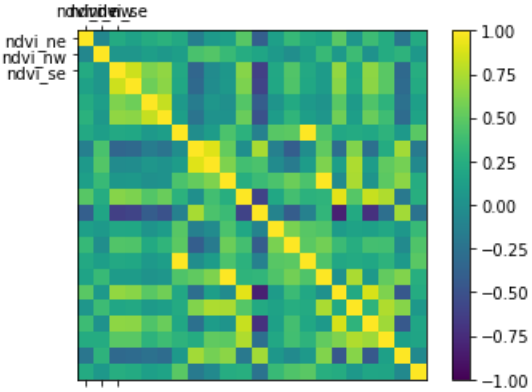












Preprocessing Techniques used :

1. **Removing the null values** from the dataset.
2. **Correlation matrix** : We will be setting the threshold value to 90% and any column value ~~below~~ 90% will be dropped. It is done by using corr() from pandas.
3. **Rescaling** of data using MinMaxScalar from Scikit. ~~This technique would be used on columns to rescale the data in 0-1 range.~~ when the columns with data range will be rescaled into the 0-1 range.

It is more like, we determine the max & min corr. & decide accordingly. change according to it above.

4. **Standardization** of data using StandardScalar from Scikit. This technique is used on columns to replace values with mean or standard deviation.
5. **Categorization** of data by LabelEncoder from Scikit. This technique will be applied on the columns with string/date as values.

Goal :

To predict the total_cases label for each (city, year, weekofyear) in the test set. There are two cities, San Juan and Iquitos, with test data for each city spanning 5 and 3 years respectively.

Proposed solution & methods :

The plan to achieve our goal is as follows:

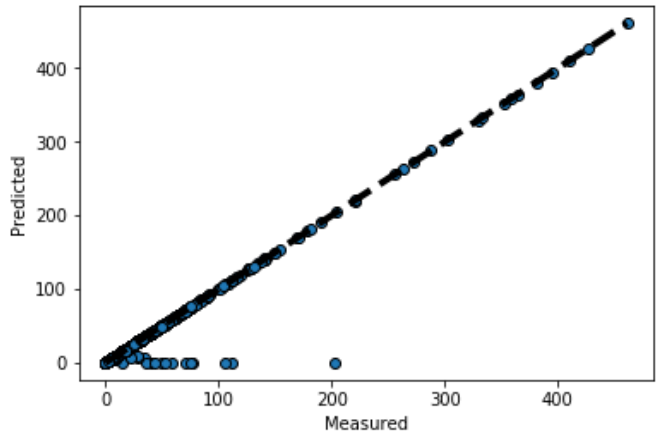
- Extract the information from dengue_features_train.csv about the features of dataset like week_start_date , precipitation , etc. and divided it into training & test data.
- Preprocess the dataset and retrieve only the attributes which gives better results in the training model.
- Create a model using the dengue_features_train.csv and dengue_labels_train.csv .
- Decision Tree, SVM & Random Forest methods are used to train and build the model.
- Coding Language is Python and we used libraries like scikit-learn, and data processing libraries like Pandas, Numpy.

Experimental results and analysis :

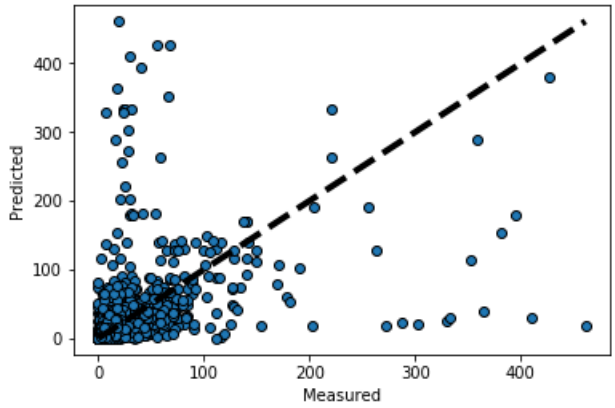
	Precision	Recall	F1-scrore
Decision Tree	0.9302606619570905	0.9271978021978022	0.9268067119805842
SVM	0.9993248277146582	0.9993131868131868	0.9993121747100526
Random Forst	0.9302606619570905	0.9271978021978022	0.9268067119805842

Plot between measured & predicted values :

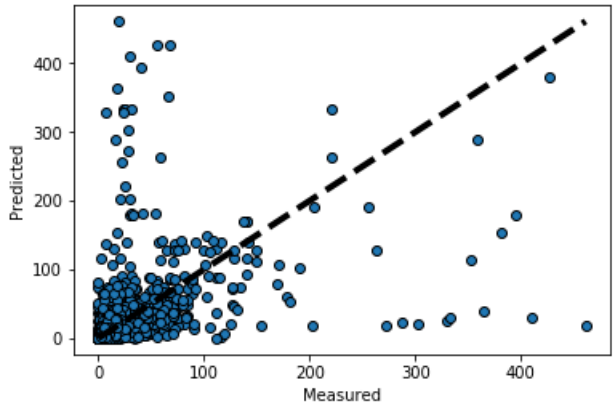
SVM :



Decision Tree :



Random Forest :



Conclusion : from the above result, it can be concluded that with SVM, a better prediction accuracy can be achieved.

~~Based on the above results,~~ we conclude that SVM gives ~~better results for our DengAI da~~taset.

Contribution of team members :

Bhavish – Preprocessing of data & training the dataset using decision tree classifier.

Jayashree – Data distribution graphs & SVM classifier implementation.

Preethika - Report & Random Forest Classifier.

References :

[1] Put the paper here.

1. https://towardsdatascience.com/random-forest-in-python-24d0893d51c0
2. https://scikit-learn.org/stable/modules/ensemble.html
3. https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/
4. https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/