CS 6375 – MACHINE LEARNING

Project Status Report

April 09th

Shiva Podugu (sxp170130)

Manohar Katam (mxk164930)

Sravani Lingam (sxl170330)

Venkata Kartheek Madhavarapu (vxm153830)

Our Project is on **predicting the occurrences of** "**total_cases"** of Dengue based on "**DengAI: Predicting Disease Spread**" dataset which is in active driven data competitions. Here we have a set of weather information (precipitation, temperature, vegetation) from the two cities: San Juan (sj) and Iquitos (iq) with total cases of dengue by year and week of the year. We aim at making a complete analysis of the DengAI dataset to find the total number of Dengue affected cases in the given two cities with respect to set of climate variables as mentioned above.

The Dataset - DengAI: Predicting Disease Spread

The *DengAI* dataset is taken from an active DrivenData Competition and the link of which is given below:

https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

This dataset has two cities: San Juan (sj) and Iquitos (iq). Since we assume that the spread of dengue may follow different patterns between the two different cities, we will divide the dataset, train separate models for each city, and then join our predictions finally.

Number of attributes = 24

Number of instances = 1456

The attributes are the following:

1. city – we will divide all city = San Juan (sj) data into one dataset and
   all city = Iquitos (iq) data into other dataset
2. year

3. week_of_year
4. week_start_date
5. station_max_temp_c – Maximum temperature
6. station_min_temp_c – Minimum temperature
7. station_avg_temp_c – Average temperature
8. station_precip_mm – Total precipitation
9. station_diur_temp_rng_c – Diurnal temperature range
10. precipitation_amt_mm – Total precipitation
11. reanalysis_sat_precip_amt_mm – Total precipitation
12. reanalysis_dew_point_temp_k – Mean dew point temperature
13. reanalysis_air_temp_k – Mean air temperature
14. reanalysis_relative_humidity_percent – Mean relative humidity
15. reanalysis_specific_humidity_g_per_kg – Mean specific humidity
16. reanalysis_precip_amt_kg_per_m2 – Total precipitation
17. reanalysis_max_air_temp_k – Maximum air temperature
18. reanalysis_min_air_temp_k – Minimum air temperature
19. reanalysis_avg_temp_k – Average air temperature
20. ndvi_se – Pixel southeast of city centroid
21. ndvi_sw – Pixel southwest of city centroid
22. ndvi_ne – Pixel northeast of city centroid
23. ndvi_nw – Pixel northwest of city centroid
24. reanalysis_tdtr_k – Diurnal temperature range

Here is the snapshot of the data:

dengue_features_train <- read_csv ("~/Downloads/dengue_features_train.csv ”)

head(dengue_features_train)

dengue_features_train

	city	year	weekofyear	week_start_date	ndvi_ne	ndvi_nw	ndvi_se	ndvi_sw	precipitation_amt_mm	reanalysis_air_temp_k	reanalysis_avg_temp_k	reanalysis_dew_point_tem
1	sj	1990	18	1990-04-30	0.12260000	0.10372500	0.19848330	0.17761670	12.42	297.5729	297.7429	292.4
2	sj	1990	19	1990-05-07	0.16990000	0.14217500	0.16235710	0.15548570	22.82	298.2114	298.4429	293.9
3	sj	1990	20	1990-05-14	0.03225000	0.17296670	0.15720000	0.17084290	34.54	298.7814	298.8786	295.4
4	sj	1990	21	1990-05-21	0.12863330	0.24506670	0.22755710	0.23588570	15.36	298.9871	299.2286	295.3
5	sj	1990	22	1990-05-28	0.19620000	0.26220000	0.25120000	0.24734000	7.52	299.5186	299.6643	295.8
6	sj	1990	23	1990-06-04	NA	0.17485000	0.25431430	0.18174290	9.58	299.6300	299.7643	295.8

reanalysis_tdtr_k	station_avg_temp_c	station_diur_temp_rng_c	station_max_temp_c	station_min_temp_c	station_precip_mm
2.628571	25.44286	6.900000	29.4	20.0	16.0
2.371429	26.71429	6.371429	31.7	22.2	8.6
2.300000	26.71429	6.485714	32.2	22.8	41.4
2.428571	27.47143	6.771429	33.3	23.3	4.0
3.014286	28.94286	9.371429	35.0	23.9	5.8
2.100000	28.11429	6.942857	34.4	23.9	39.1

dengue_labels_train <- read_csv("~/Downloads/dengue_labels_train.csv")

head(dengue_labels_train)

dengue_features_train | dengue_labels_train

	city	year	weekofyear	total_cases
1	sj	1990	18	4
2	sj	1990	19	5
3	sj	1990	20	4
4	sj	1990	21	3
5	sj	1990	22	6
6	sj	1990	23	2

We are dividing dengue_features_train and dengue_labels_train into two datasets based on the city value by using the following commands:

sj_dengue_train_labels<-subset(dengue_labels_train,city=="sj")
iq_dengue_train_labels<-subset(dengue_labels_train,city=="iq")

sj_dengue_train_features<-subset(dengue_train_features,city=="sj")
iq_dengue_train_features<-subset(dengue_train_features,city=="iq")

Techniques we planned to use

We planned to apply the following techniques on the data to complete the required analysis –

- k-Nearest Neighbors (kNN)
- Random Forests
- Bagging
- Gradient Boosting

Experimental Methodology

We employ the following procedure in our project –

1. Pre-processing of the dataset
   - This step involves dealing with the NA values,
   - Scaling the required attributes,
   - Removing the uncorrelated attributes.
2. On the dataset
   - We perform each of the aforementioned techniques,
   - Also, vary the parameters and find the best set of parameters for the technique.
3. We evaluate the techniques using the following metrics –
   - Accuracy
   - Precision
   - Recall

- F-measure

4. We plot the results that aid in comparing the performance of the classifiers.

Programming Language

We plan to use **R** programming for the project.

Preliminary Results

We now present the results of the work we've done so far.

- ➔ Removing "week_start_date" attribute since it is not a feature for our model and removing it will not make any difference in the result.

- ➔ We plotted the CORRPLOT which aids in identifying the correlation of the attribute with the class attribute (total_cases).

The plot is as follows for each of the cities Iquitos (iq) and San Juan (sj) :-

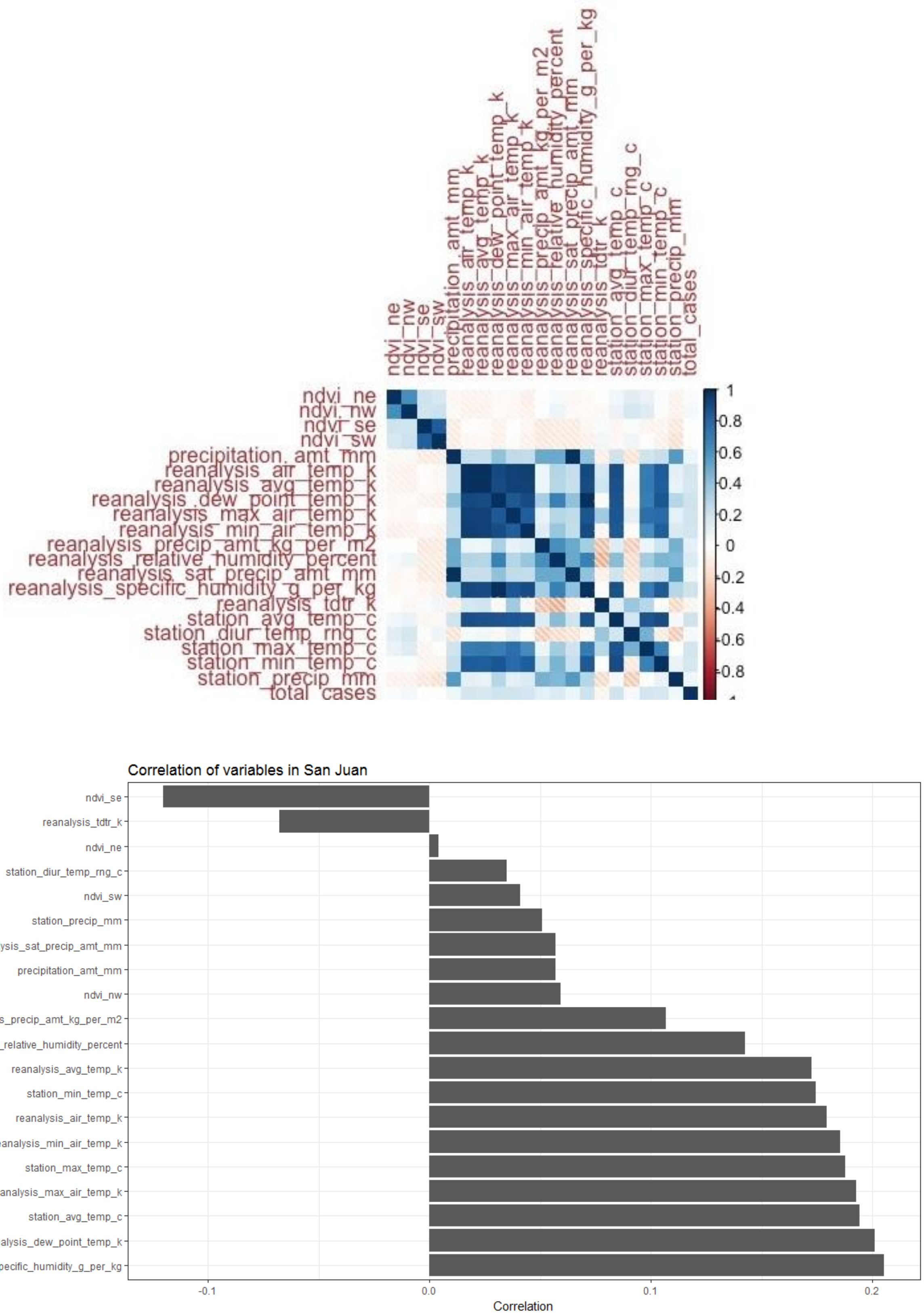

Fig 1. Correlation heat plot and bar plot of city San Juan (sj)

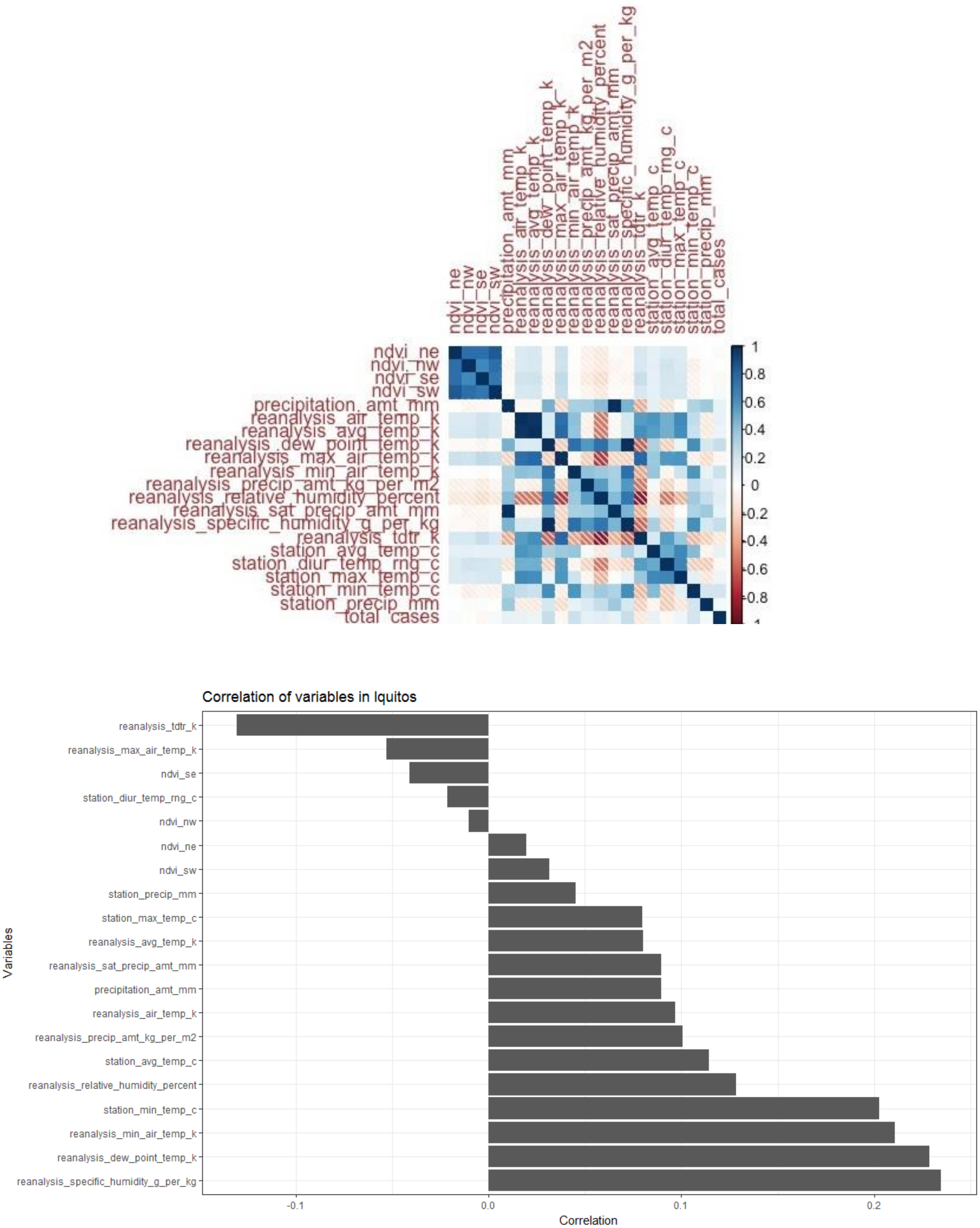

Fig.2 Correlation heat plot and bar plot of city Iquitos

Observations

1. We see that many of the temperature data are strongly correlated.
2. We noted that total_cases has weak correlations with the other attributes and vegetation_index attributes also has weak correlations with other attributes.
3. We also see that correlation strengths differ for each city and reanalysis_specific_humidity_g_per_kg and reanalysis_dew_point_temp_k are the most strongly correlated with total_cases. As we know that mosquitos thrive wet climates, we can infer why they are strongly correlated.
4. As minimum temperatures, maximum temperatures, and average temperatures rise, the total_cases of dengue fever tend to rise as well.
5. We also note that the precipitation measurements bear little to no correlation to total_cases, despite strong correlations to the humidity measurements.

Based on the correlation observations above, these are the attributes which have the strong correlations with the total_cases in each city:

- `reanalysis_specific_humidity_g_per_kg`
- `reanalysis_dew_point_temp_k`
- `station_avg_temp_c`
- `station_min_temp_c`

R_CODE:

```
#Read the dataset
dengue_labels_train <- read_csv("~/Downloads/dengue_labels_train.csv")
dengue_features_train <- read_csv("~/Downloads/dengue_features_train.csv")

View(dengue_labels_train)
View(dengue_features_train)

#Loading data by City
sj_dengue_train_labels<-subset(dengue_labels_train,city=="sj")
iq_dengue_train_labels<-subset(dengue_labels_train,city=="iq")
```

```
sj_dengue_train_features<-subset(dengue_features_train,city=="sj")
iq_dengue_train_features<-subset(dengue_features_train,city=="iq")
```

#Merging features and labels

```
merged_sj_train_features_instances <-
merge(sj_dengue_train_features,sj_dengue_train_labels,by=c('city','year','weekofyear'))
merged_iq_train_features_instances <-
merge(iq_dengue_train_features,iq_dengue_train_labels,by=c('city','year','weekofyear'))
```

#Pre-processing

When the data set is loaded in R, the null values are replaced by NA.

We have used "gam" package in which NAs are replaced by the mean of the non-missing entries.

```
library(gam)
merged_sj_train_features_instances <- na.gam.replace(merged_sj_train_features_instances)
merged_iq_train_features_instances <- na.gam.replace(merged_iq_train_features_instances)

View(merged_sj_train_features_instances)
View(merged_iq_train_features_instances)
```

Removing 'week_start_date' column

As it doesn’t impact the result significantly.

We have used "dplyr" package is used to remove unnecessary columns.

```
merged_sj_train_features_instances <- dplyr::select(merged_sj_train_features_instances, -week_start_date)
merged_iq_train_features_instances <- dplyr::select(merged_iq_train_features_instances, -week_start_date)

View(merged_sj_train_features_instances)
View(merged_iq_train_features_instances)
```

#Finding the correlation plot

```
library(corrplot)

sj_corrplot<-cor(merged_sj_train_features_instances[,4:24])
```

```
corrplot(sj_corrplot,type = 'full', tl.col = 'brown', method="shade")

iq_corrplot<-cor(merged_iq_train_features_instances[,4:24])
corrplot(iq_corrplot,type = 'full', tl.col = 'brown', method="shade")
```