Vector Borne Diseases

Abstract

Climatic changes can have a dominant effect on the health of each individual. Changes in the environment like high precipitation value and land temperatures directly affect health care. If the precipitation value is higher there are chances of disease outbreak such as Dengue, Chikungunya and many diseases caused by vectors. Extreme climate and weather changes are the causes for droughts, floods and heat waves. These are favorable climatic conditions for Mosquitoes to breed which causes diseases like Dengue or Chikungunya. We are exploring the areas where the disease outbreak is highest and analyzing the precipitation values in that particular area.

Health care facilities play a very important role during this disease outbreak. We can't stop or eradicate the disease completely at once, but we can at least have some preventive measures and prepare for ourselves and health care facilities with necessary equipment for the treatment. For predicting the future outbreak to take preventive measures we are using the past data for the disease outbreak and the climate variables data. In recent times we all know about Floods in Florida the rising temperature and drastic climatic changes are causing these things for which we are focusing on the diseases caused by Mosquitoes due the drastic climatic changes.

Introduction

Global climatic changes are increasing nowadays due to human activities. The issues caused due to climatic changes are floods, global warming, heat waves and many other changes. Vector Borne is one of the causes due to vectors which are Mosquitoes. The favorable climate for the Mosquitoes to breed is when the precipitation values are high. The diseases like Dengue and Chikungunya are the major diseases caused due to these vectors. The effect of these diseases are worst on human life, for this predicting the future outbreak and analyzing the patterns across each region is necessary. This would help health care workers to take proper care before the outbreak occurs.

The dataset required to perform this analysis are climate variable and disease outbreak data.

The outbreak data from ProMed Emails extracted using python scripts. The extracted data is stored in CSV format and used for the analysis. The second dataset would be climate variables dataset from NASA GPM data, which is a HDF file format. I am converting that file to CSV format using python scripts and extracting the precipitation These two dataset are joined based on date and latitude and longitude of the location. From the final dataset we can get the combination of outbreak and precipitation information. This dataset helps us to predict the future outbreak and analyze the current outbreak based on the climatic conditions.

Exploratory Data Analytics (EDA) and Uncertainties:

Analyzing the number of cases based on the region would help us to understand the disease outbreak areas. The data represents the cases across the regions.

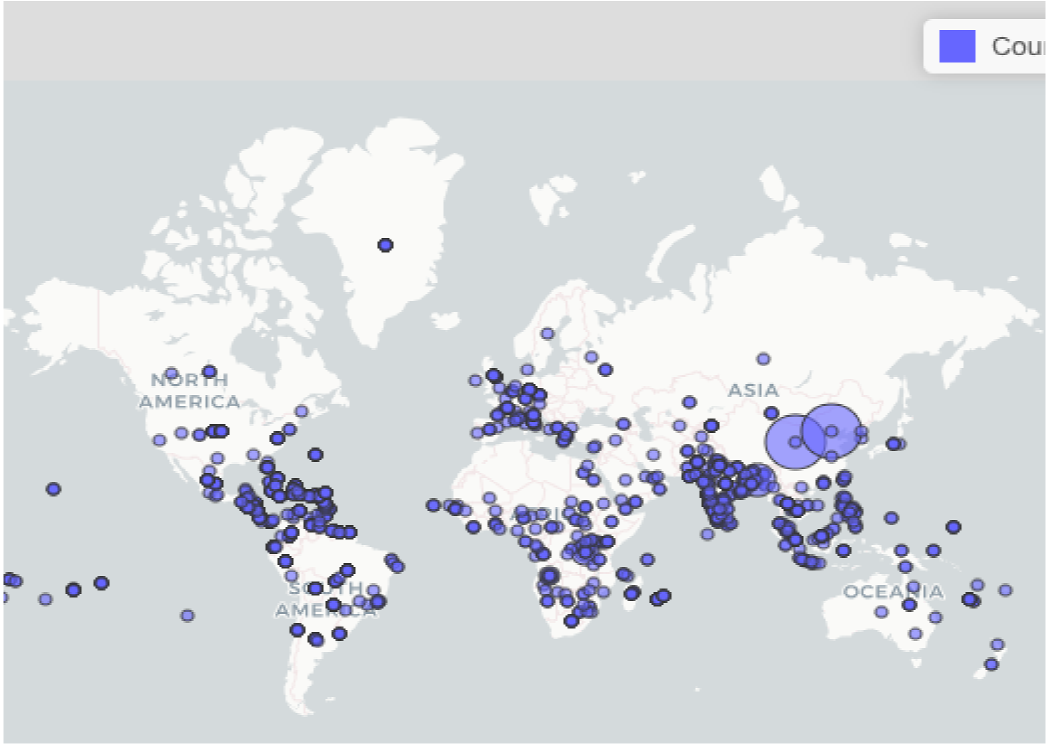


Figure 1 Vector borne cases across all countries

Figure 1 helps us understand the spread of vector borne diseases across all countries from 2000. A bigger circle size represents more cases in that particular area. The disease spread is across South America, Asia, Africa, Australia and the cases in North America are low when compared to other areas.

As there are outliers in the data, we can filter out the cases which are greater than 10,000 to get rid of outliers. Figure 2 helps understand the data after filtering out the outliers.

Filtering out the cases which are less than 3 and cases which are greater than 10,000. This would help us understand the spread and predict the future outbreak easily. Also filtered out the data before 2000.

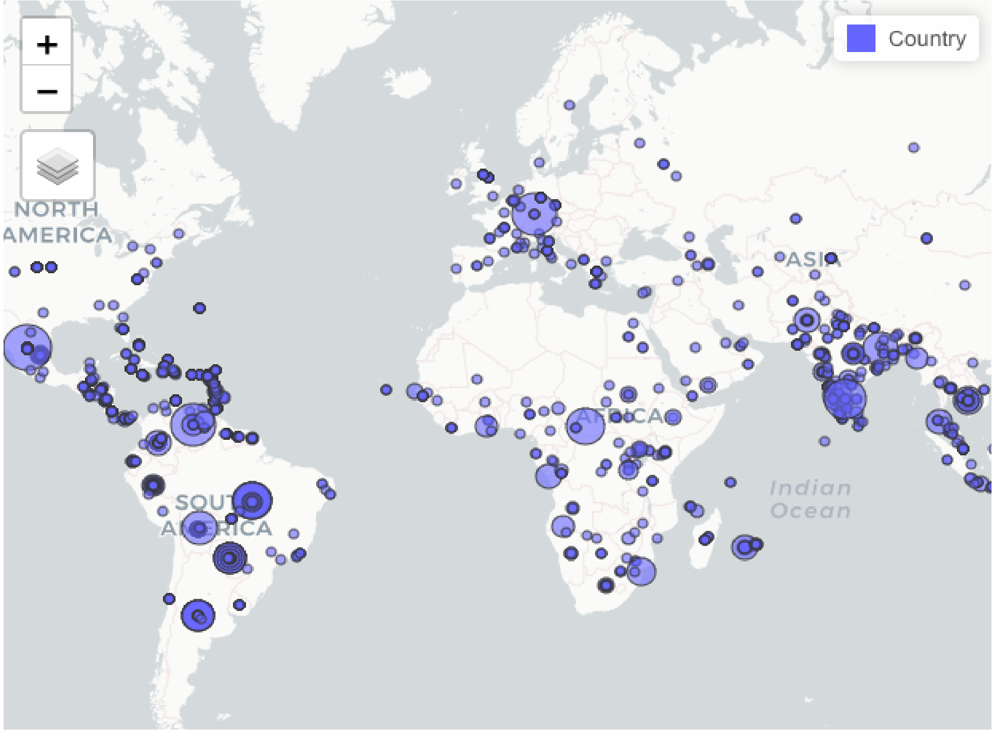


Figure 2 Vector borne cases after filtering out the outliers

In Figure 2 we are representing the data after filtering the outliers and for a particular period. There is a lot of difference between the Figure 1 and Figure 2 as outliers are the data the cases are not properly understood, now we can clearly understand how many cases there are and how the spread is in different parts of the world.

As I am concentrating on Dengue outbreak my point of focus is on Dengue outbreak across the world filtering to Dengue outbreak data.

Figure 3 shows the Dengue outbreak information across the world; the outbreak is mainly in South America and other parts of the world. The disease outbreak is high in some parts of South America according to Figure 3.

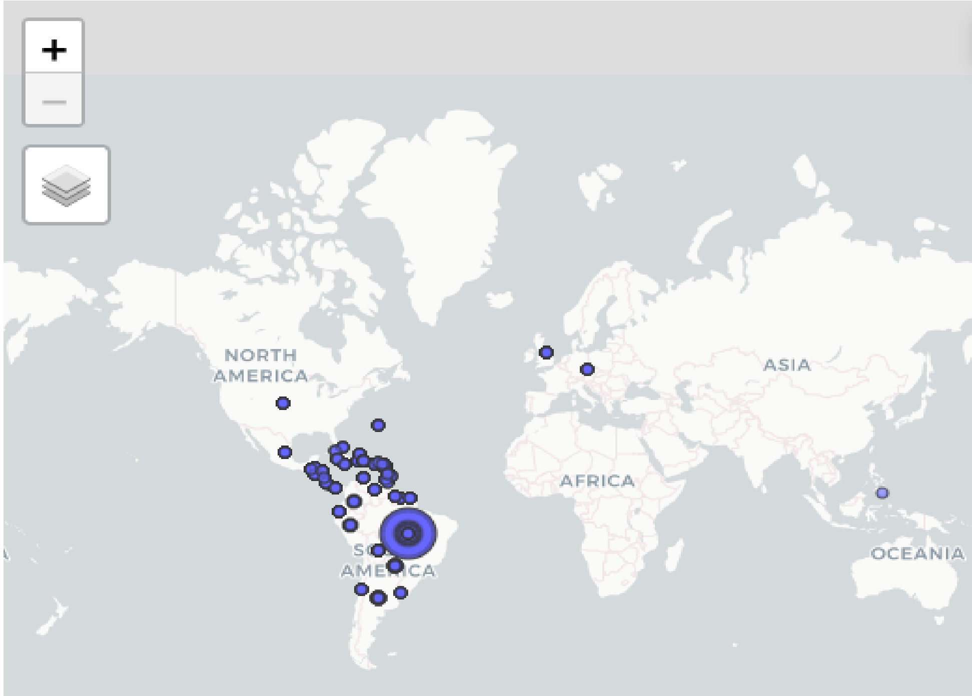


Figure 3 Dengue outbreak across different parts of the world

The month wise analysis helps us understand in which month the cases are high and which season of the year the cases are higher Figure 4 helps us understand the month wise dengue disease outbreak across all countries.

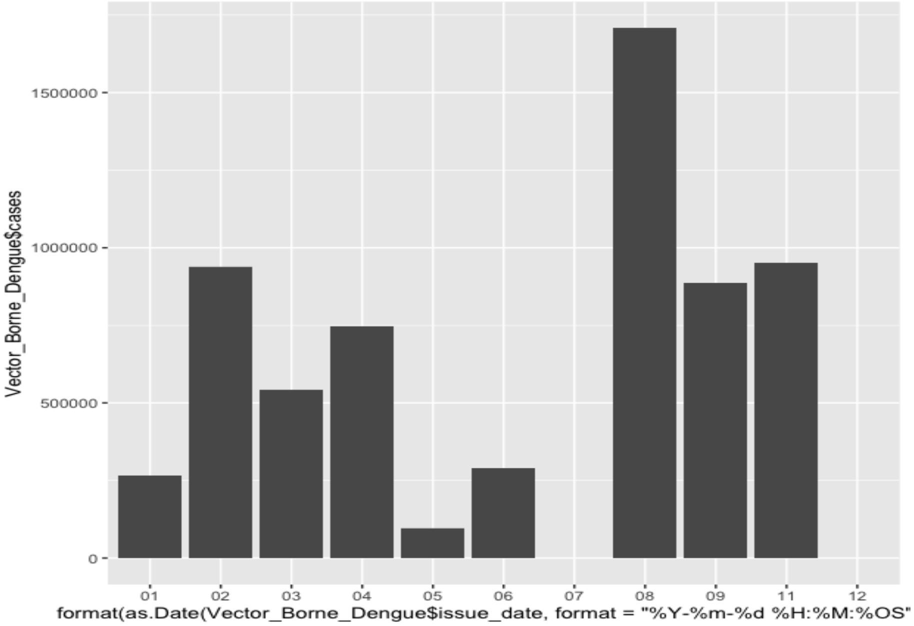


Figure 4 Month wise distribution of cases across different countries

Models

K Means:

K Means clustering helps to classify the cases based on the region. Classify the cases based on geo locations.

For classifying the data we would need the k value. K value can be derived from the number of records we are having in the data. Here our k value is 19. For we deriving the number of clusters we can do with our data we are using WSS function.

```
wssplot <- function(data, nc=19, seed=1234){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
      ylab="Within groups sum of squares")
  wss
}
```

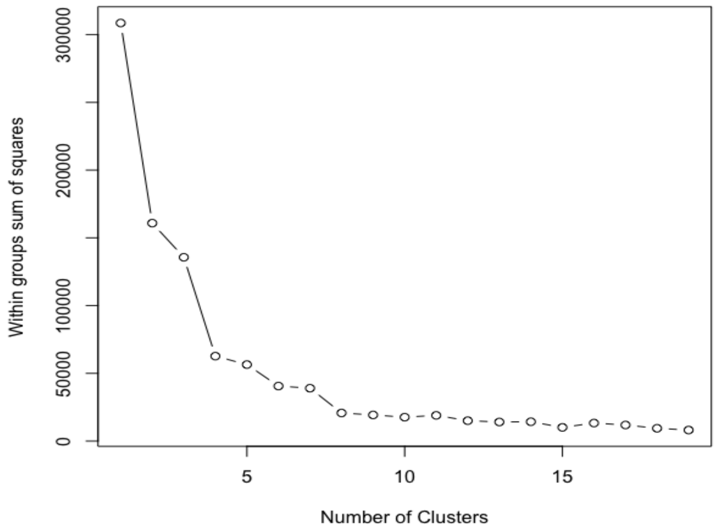


Figure 5 WSS plot for getting number of clusters

From Figure 5 we can conclude that the number of clusters we can use for classification is 5. After applying the k means algorithm for classifying the data based on the regions. The disease outbreak across the world is shown as below in Figure 6

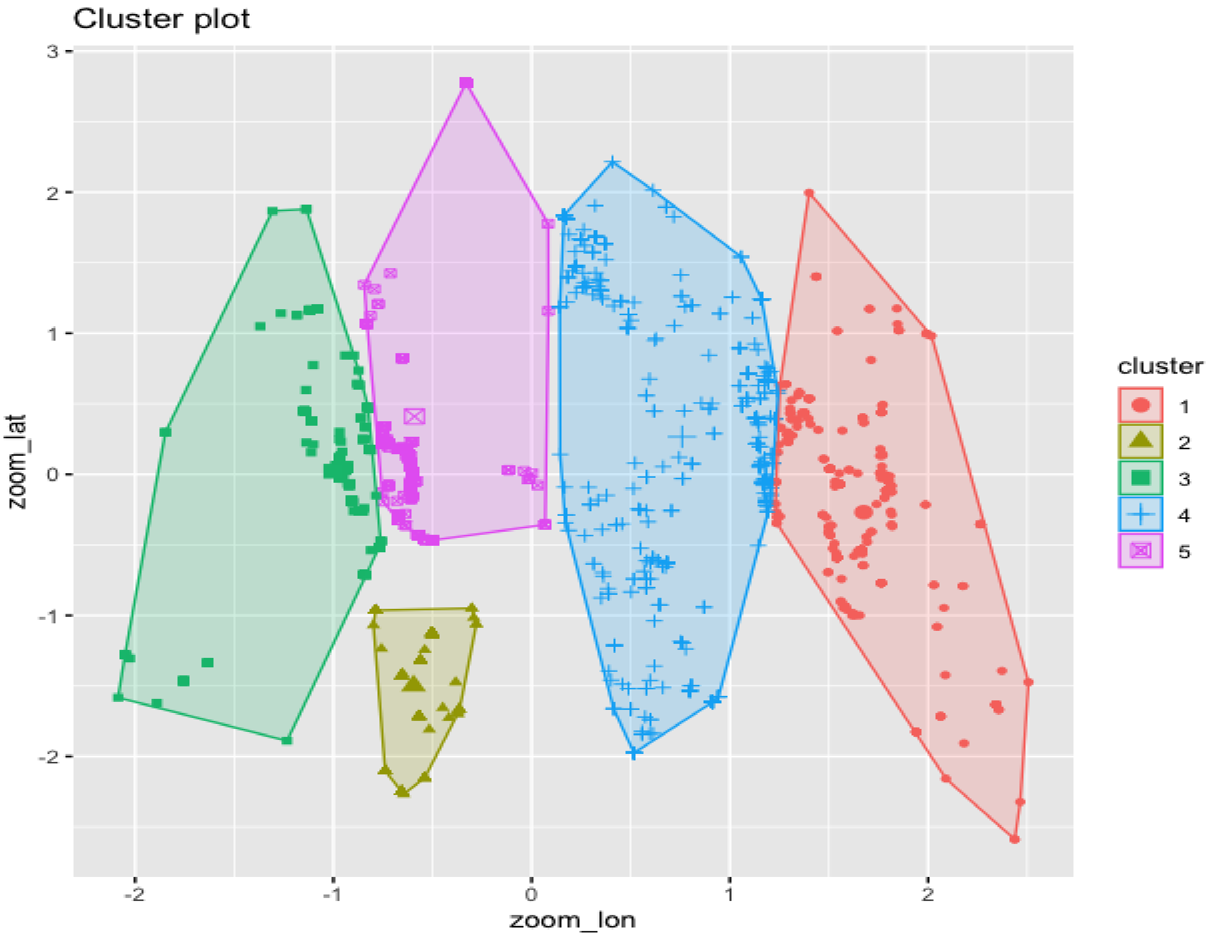


Figure 6 Clusters formed after applying K Means

Figure 6 is the classification of the diseases spread across different countries and the intensity of the spread. The clusters are made based on the similarity in the data and based on latitude and longitudes of the data. The 5 different clusters help us to understand the disease spread can be classified into 5 groups based on latitude and longitude.

DB Scan:

DB Scan algorithm helps us to classify the data based on disease spread and it is more accurate than K means algorithm. It is density based clustering hence it helps us to group the nearest neighbor into one group and classification would be more accurate. The disease cases are grouped into nearest neighbors. The below are the clusters formed after applying the Density Based Clustering. Each cluster has the number of cases at each location .

```
DBSCAN clustering for 2103 objects.
Parameters: eps = 8, minPts = 5
Using euclidean distances and borderpoints = TRUE
The clustering contains 29 cluster(s) and 34 noise points.
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24
 34  35 833  69 137  81  38  44   9 453 141   5  68   6  16   7   7   8   5   5   5   5   5   8  14
 25  26  27  28  29
  8  17   6  27   7
```

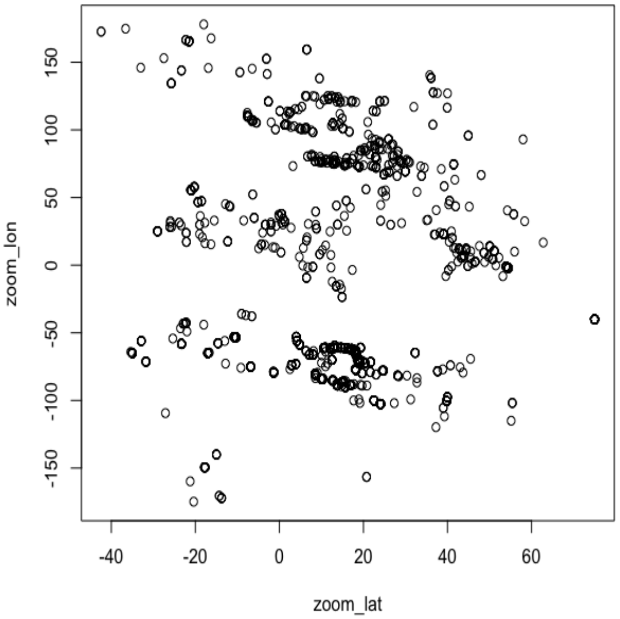


Figure 7 Before applying DB Scan algorithm

The Figure 7 is the plot before applying the DB Scan algorithm to identify the difference between before and after applying DB Scan.

The clusters formed are plotted on graphs in figure 8. Which signifies the classification of the cases based on longitude and latitude. The distance specified between each neighbor is 8 the cases reported around the nearest neighbor into one cluster. The number of clusters formed after applying the DB Scan algorithm are 29. These clusters have all the nearest neighbors in them. The black dots from the below figure 8 are outliers; they don't belong to any nearest neighbor so they are not classified into any of the groups.

The cluster 2 which is yellow in Figure 8 is having more number of cases in that cluster. The cluster 5 has more cases in the cluster which means the nearest elements to the center point of the cluster are more.

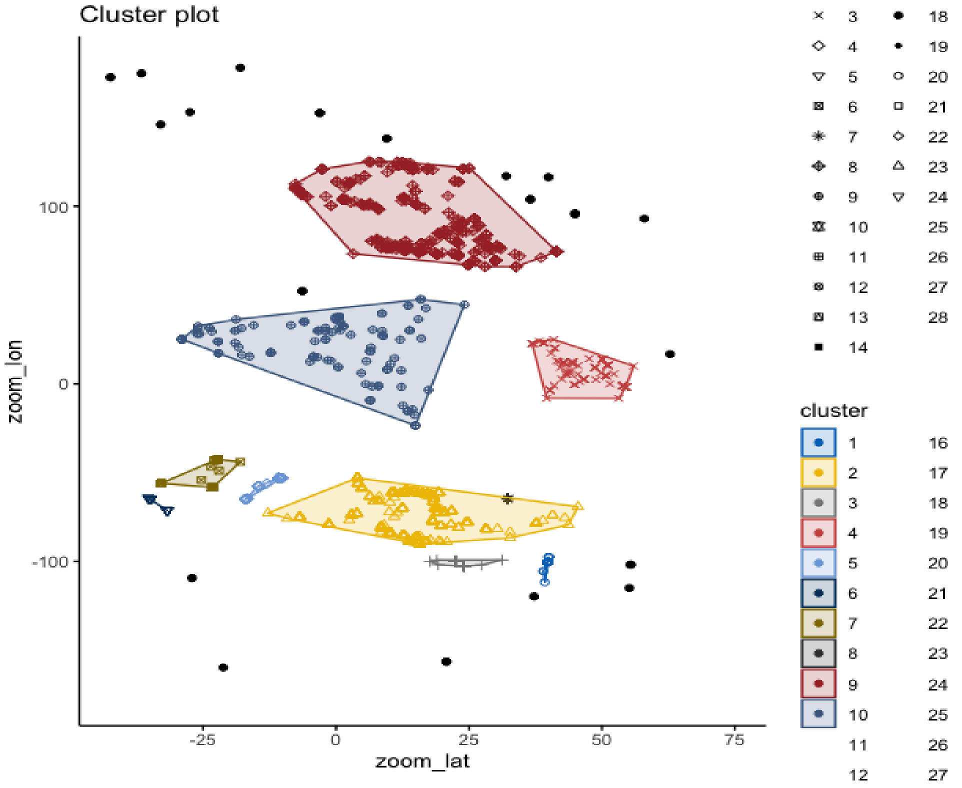


Figure 7 Clustering based on DB Scan algorithm

Decision Trees

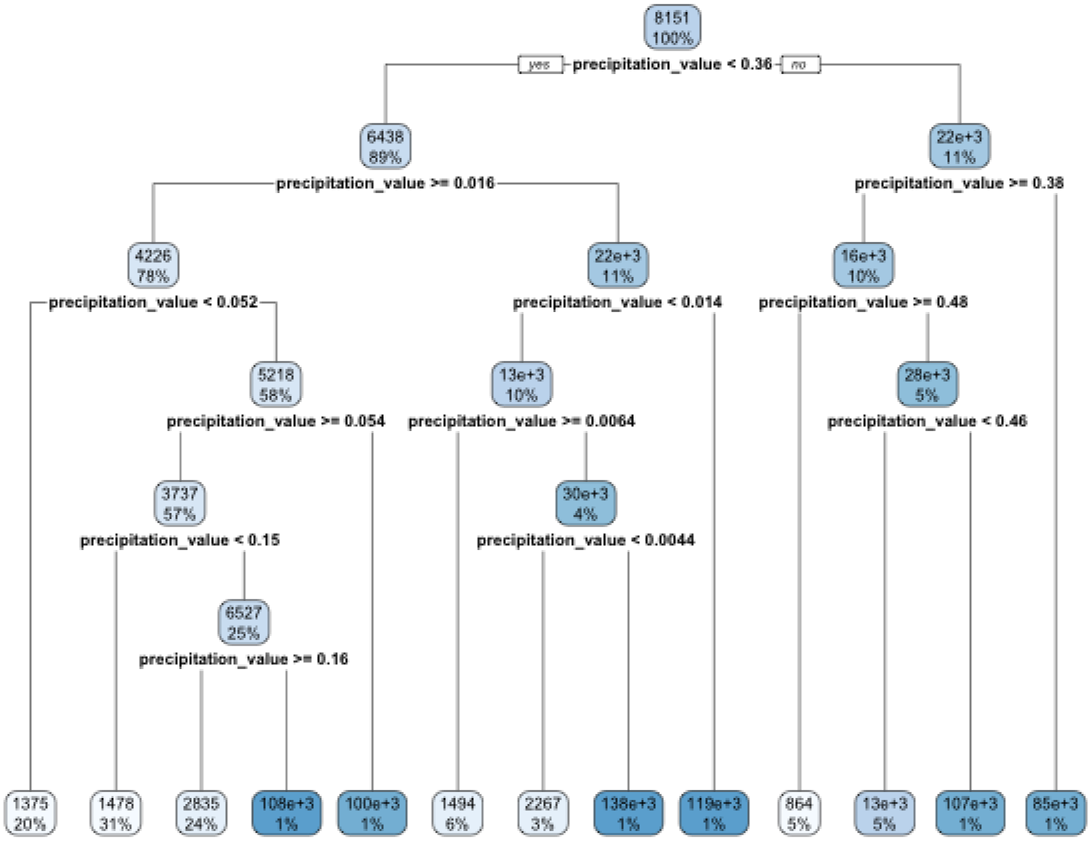

8151
100%
yes
precipitation_value < 0.36
no
6438
89%
22e+3
11%
precipitation_value >= 0.016
precipitation_value >= 0.38
4226
78%
22e+3
11%
16e+3
10%
precipitation_value < 0.052
precipitation_value < 0.014
precipitation_value >= 0.48
5218
58%
13e+3
10%
28e+3
5%
precipitation_value >= 0.054
precipitation_value >= 0.0064
precipitation_value < 0.46
3737
57%
30e+3
4%
precipitation_value < 0.15
precipitation_value < 0.0044
6527
25%
precipitation_value >= 0.16
1375
20%
1478
31%
2835
24%
108e+3
1%
100e+3
1%
1494
6%
2267
3%
138e+3
1%
119e+3
1%
864
5%
13e+3
5%
107e+3
1%
85e+3
1%

Figure 9 Decision tree classification between the cases and precipitation

The above Decision Tree helps us understand the classification of precipitation data vs the cases according to the precipitation value. The Decision Trees help to understand the tree traversal, the tree top node has precipitation value less than 0.36 the number of cases less than 0.36. Cases less than precipitation value 0.36 are 6438 cases, cases above precipitation value 0.36 traversed towards right. The cases are more if the precipitation value is greater than 0.36, the maximum number of cases are greater than 0.38. If the precipitation value is greater than 0.38 then the number of cases are high.

With the above model we can conclude that the greater the precipitation value the more the number of cases. Which means the rainy seasons where the precipitation is high are favorable conditions for mosquitoes to breed so the cases are high in this period.

Regression Model

```
Call:
lm(formula = cases ~ precipitation_value, data = Vector_Borne_Dengue_Col)
Residuals:
   Min     1Q Median     3Q    Max
-15424  -8372  -7361  -6701 823674
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)            6548       3191    2.05    0.041 *
precipitation_value   10422      14872    0.70    0.484
Signif. codes:  0 ‘***’0.001 ‘**’0.01 ‘*’0.05 ‘.’0.1 ‘ ’1
Residual standard error: 62500 on 786 degrees of freedom
Multiple R-squared:  0.000624,      Adjusted R-squared:  -0.000647
F-statistic: 0.491 on 1 and 786 DF,  p-value: 0.484
```

The above are the summary of the regression model generated for predicting the future cases using all other variables in the data frame. The summary for the regression model is generated using the precipitation value factor. For predicting the future outbreak the number of cases and number of deaths using the past factors. Using the intercept and error we can predict the future outbreak.

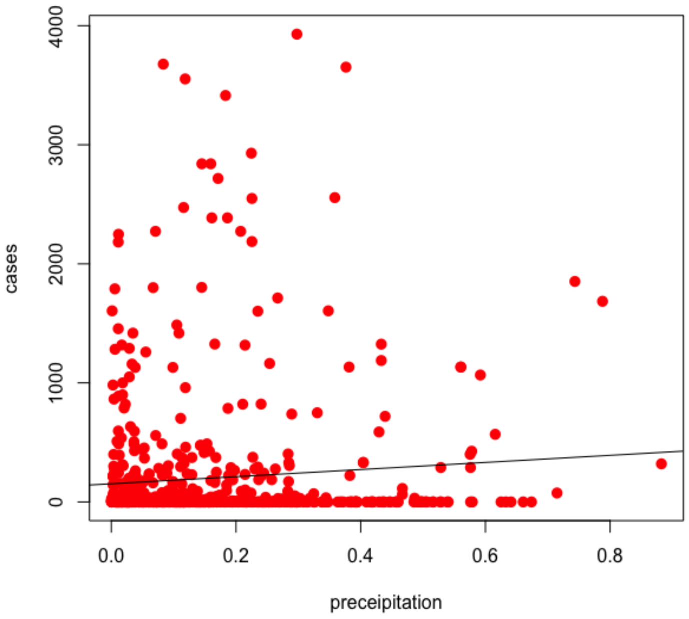
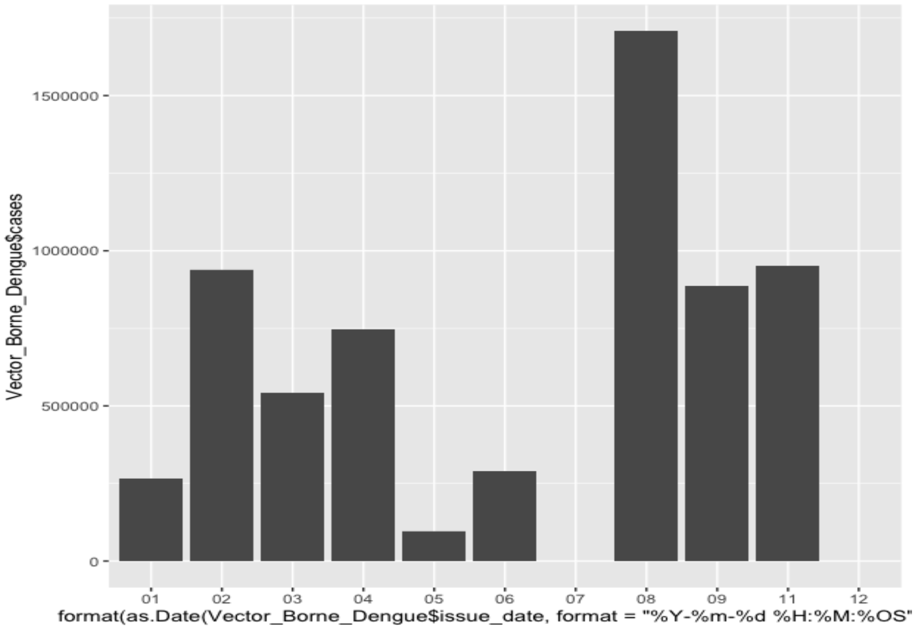


Figure 10 Regression Model plot

Conclusion

Models worked well after applying different algorithms, we can conclude by cluster models and regression models. By looking into cluster models we can conclude that in South America the cases are high. In the future we can predict the cases using the precipitation values. The decision tree helps us understand the cases which have the greater precipitation values. By looking into the plot in the figure 11 we can conclude that the cases are more in the month of August which is the Rainy season in most parts of the world where the chances of Mosquitoes breeding increase and the disease outbreak may increase in this period of the year. We can conclude that in future we can expect the disease outbreak may occur in this period of the year.



Gituhub Link for code:
https://github.com/sindhoora8/DataAnalyticsFall2022_Sindhoora_Mandadi/tree/main/Term%20Project

References

https://www.r-bloggers.com/2021/04/cluster-analysis-in-r/

https://www.mdpi.com/2076-0817/9/11/914

https://www.nature.com/articles/s41598-018-38034-z