Project Report

A Machine Learning Approach To Extract Dengue Severity Pattern By Using K-Means Clustering

Student: **Sudipta Kumar Das**

Student ID: **20-43658-2**

Course: Data Science

May, 2023 | Spring

Abstract

Dengue fever is a significant public health concern worldwide, and its incidence is influenced by various environmental factors. This paper presents the results of an analysis of a dataset using K-means clustering with 10 clusters to gain insights into the relationship between environmental conditions and Dengue cases. The severity levels assigned to each cluster indicate that higher values of air temperature, humidity, and rainfall are associated with a higher probability of Dengue being reported. However, it is important to note that a high number of Dengue cases does not necessarily indicate a high severity level. These findings can inform public health strategies aimed at preventing and controlling the spread of Dengue fever. This paper discusses the implications of these results for public health interventions.

Keywords: Unsupervised learning, Clustering algorithms, Data Science, K-Means Cluster

Contents

List of Figures

Code Excerpts Index

1. Introduction

Based on an analysis of a dataset using K-means clustering with 10 clusters, insights have been gained into the relationship between environmental conditions and Dengue cases. The severity levels assigned to each cluster based on the features indicate that high values of air temperature, humidity, and rainfall are associated with a higher probability of Dengue being reported. However, it is important to note that a high number of Dengue cases does not necessarily indicate a high severity level. These findings can inform public health strategies aimed at preventing and controlling the spread of Dengue fever. This paper presents the results of this analysis and discusses their implications for public health interventions.

2. Background

Dengue fever is a vector-borne viral disease that is transmitted to humans by the Aedes mosquito. The incidence of dengue in Bangladesh has been on the rise in recent years, and climate change is believed to be a contributing factor [1]. Bangladesh is highly vulnerable to the impacts of climate change due to its location and geography.

Several studies have investigated the relationship between dengue and climate in Bangladesh. One such study conducted by Ahmed et al. (2019) analyzed the relationship between dengue incidence and climatic variables such as temperature, rainfall, and humidity in Dhaka, Bangladesh. The study found that there was a significant positive correlation between dengue incidence and temperature and rainfall [1].

Another study conducted by Hossain et al. (2020) examined the association between dengue fever and weather variables in Chittagong, Bangladesh. The study found that temperature and relative humidity were positively associated with dengue incidence, while wind speed was negatively associated [2].

Furthermore, a study conducted by Nusrat et al. (2018) investigated the impact of climate change on dengue transmission in Bangladesh. The study used a mathematical model to project future dengue transmission under different climate change scenarios. The study found that with increasing temperatures, the risk of dengue transmission would increase significantly [2].

In conclusion, there is strong evidence to suggest that climate change is contributing to the increasing incidence of dengue fever in Bangladesh. Continued research is needed to further understand this relationship and develop effective strategies for reducing the incidence of dengue in the country.

3. Data Description

The dataset titled ”Dengue Incidents & Weather of Bangladesh” contains information on Dengue cases and weather conditions in Bangladesh between 2008 and 2019. The dataset has a total of 134 instances and includes the following attributes:

3.1 Attribute : Year

An integer attribute representing the month in which data was recorded. The range of values in this attribute is from January to December.

3.2 Attribute : Month

numerical attribute representing the minimum temperature (in Celsius) in the given place during the recorded time period.

3.3 Attribute : MIN

The age of the customer is recorded in years. It is a numerical attribute and not categorical.The data range for this attribute can vary depending on the age range of the customers in the dataset.

3.4 Attribute : MAX

A numerical attribute representing the maximum temperature (in Celsius) in the given place during the recorded time period.

3.5 Attribute : Humidity

A numerical attribute representing the humidity level (in percentage) in the given place during the recorded time period.

3.6 Attribute : Rainfall

A numerical attribute representing the rainfall (in millimeters) in the given place during the recorded time period.

3.7 Attribute : Dengue

A numerical attribute representing the total number of Dengue cases reported during the recorded time period in the given place.

These attributes provide valuable information about the incidence of Dengue fever in Bangladesh and how it is related to environmental conditions like temperature, humidity, and rainfall. The dataset can be utilized for further analysis and modeling to gain insights into the factors that contribute to the spread of Dengue fever and inform public health strategies to mitigate its impact.

4. Approach

We will conduct data exploration for the dataset, and then we will conduct Multivariate exploration also for the dataset. We will also train the K-Means Clustering Algorithm .And to conduct the whole procedure We will use the R programming language to perform data analysis.

4.1 Data Preparation

Data preparation involves cleaning, transforming, and organizing raw data to make it suitable for analysis

4.1.1 Data Exploration

Data exploration involves examining and understanding the characteristics and patterns within a dataset, often using statistical methods and visualizations

4.1.1.1 Dataset Dimension

We need to know the dimensions of the dataset, to know how many records are available in the dataset as well as how many attributes are there in the dataset.

Code extraction 1: Dataset Dimension

```
dim(df)
```

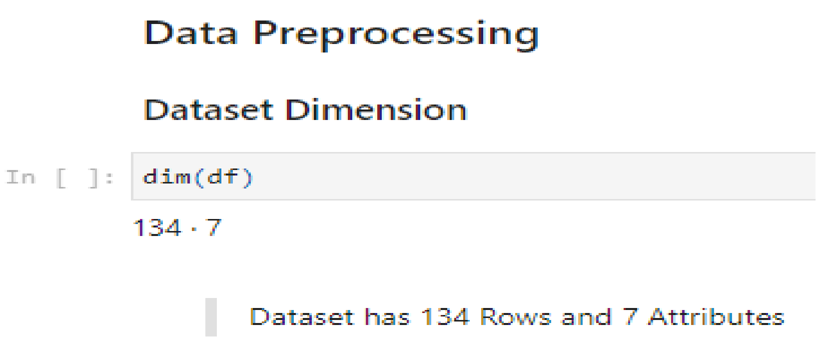

Figure 4.1.1.1: Dataset Dimension

After executing and analysis the dataset through the code, we have found that the dataset has 134 records and 7 attributes.

4.1.1.2 Attributes Data Types

We also need to know the data types of the eac attribute in the dataset. Because we know that different algorithms supports different kinds of data types. That is the reason we need to know the data types of the attributes in the dataset. And to achieve this, we use structure function of R

Code extraction 2: Attribute Datatype

```
str(df)
```

Dataset Structure (Data Types & Value-Glimpse)

```
In [ ]: str(df)
```

```
'data.frame':	134 obs. of  7 variables:
 $ YEAR    : int  2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...
 $ MONTH   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ MIN     : num  13 13.7 20.4 22.8 23.9 ...
 $ MAX     : num  25.1 26.4 31.4 34 34.2 ...
 $ HUMIDITY: num  78.8 72.6 76.9 73.9 77.4 ...
 $ RAINFALL: num  1.287 0.688 0.974 0.981 7.021 ...
 $ DENGUE  : int  0 0 0 0 0 0 160 473 334 184 ...
```

> Everything is Good

Figure 4.1.1.2: Attributes Data Type

After executing and analysis the dataset through the code, we have found that Only weight attribute is numeric blood attribute is character, rest of all are integer type.

4.1.1.3 Data Summary

Data Summery is a very important step in data exploration. Because it gives us a brief idea about the dataset. It gives us the idea about the mean, median, mode, standard deviation, minimum value, maximum value, and quartiles of the dataset. And to achieve this, we use summery function of R

Code extraction 3: Data Summary

```r
summary(df)
cat("Total Missing Values = ",sum(is.na(df)))
```

Dataset Summery(Mean, Median, Min, 1st-Quadrant, 3rd-Quadrant)

```
In [ ]: summary(df)
        cat("Total Missing Values = ",sum(is.na(df)))
```

```
      YEAR          MONTH             MIN             MAX
 Min.   :2008   Min.   : 1.000   Min.   :10.60   Min.   :23.52
 1st Qu.:2010   1st Qu.: 3.000   1st Qu.:16.40   1st Qu.:29.28
 Median :2013   Median : 6.000   Median :22.94   Median :31.99
 Mean   :2013   Mean   : 6.425   Mean   :20.94   Mean   :30.85
 3rd Qu.:2016   3rd Qu.: 9.000   3rd Qu.:25.28   3rd Qu.:32.68
 Max.   :2019   Max.   :12.000   Max.   :26.49   Max.   :35.77
    HUMIDITY        RAINFALL             DENGUE
 Min.   :67.55   Min.   :  0.0000   Min.   :   0.0
 1st Qu.:77.19   1st Qu.:  0.5478   1st Qu.:   0.0
 Median :80.09   Median :  6.0081   Median :  36.0
 Mean   :80.12   Mean   : 27.8406   Mean   : 211.6
 3rd Qu.:84.78   3rd Qu.: 12.7740   3rd Qu.: 187.0
 Max.   :88.38   Max.   :689.1613   Max.   :3087.0
Total Missing Values =  0
```

> No Missing values found

Figure 4.1.1.3: Data Summary

After executing and analysis the dataset through the code, we have found that there is no missing values in the dataset. Everything was good.

4.2 Feature Selection

We do not need un-necessary attributes. It can lessened the understanding of the K-Means Clustering model. We used co-relationship of the attributes of the dataset. And select the attributes who are both strong & positively and strong & negatively co-related.

Code extraction 4: Feature Selection

```
df_corr_pos <- cor(df[, -1])
highly_correlated_pos <- findCorrelation(df_corr_pos,
    cutoff = 0.3)

# Select features with correlation coefficient <= -0.9
df_corr_neg <- cor(df[, -1])
highly_correlated_neg <- findCorrelation(abs(df_corr_neg)
    >= 0.9, cutoff = 0)

# Combine selected features
dengue_features <- df[, c(highly_correlated_pos+1, highly_
    correlated_neg+1)]
dengue_std <- scale(dengue_features)
```

Feature Selection

```
In [ ]: # Select features with correlation coefficient >= 0.3
df_corr_pos <- cor(df[, -1])
highly_correlated_pos <- findCorrelation(df_corr_pos, cutoff = 0.3)

# Select features with correlation coefficient <= -0.9
df_corr_neg <- cor(df[, -1])
highly_correlated_neg <- findCorrelation(abs(df_corr_neg) >= 0.9, cutoff = 0)

# Combine selected features
dengue_features <- df[, c(highly_correlated_pos+1, highly_correlated_neg+1)]
dengue_std <- scale(dengue_features) # Standardize the features to ensure equal weight in clustering # nolint
```

> Those attributes who has the co-relationship = Highly and Moderately co-related both positive and negative (0.3 <= x <= -0.3)

Figure 4.2.0.0: Feature Selection

4.2.1 Optimal K by Elbow Method

In the K-Means Clustering we need to use a number of clusters on which the dataset pattern will be learn by the model. Here it is called ‘K’. To find out the perfect number of cluster, we can use Elbow method. It shows the lowest number of cluster could be perfect for the dataset to be learn by the model.

Code extraction 5: Optimal K

```
# Calculate WSS for k=1 to 10
wss <- sapply(1:10, function(k){
kmeans(dengue_std, centers=k)$tot.withinss
})

# Plot the elbow curve
plot(1:10, wss, type="b", xlab="Number of clusters (K)", ylab="WSS")

# Determine the optimal number of clusters (K)
k.optimal <- which.min(wss)
cat("Optimal value of K = ",k.optimal)
```

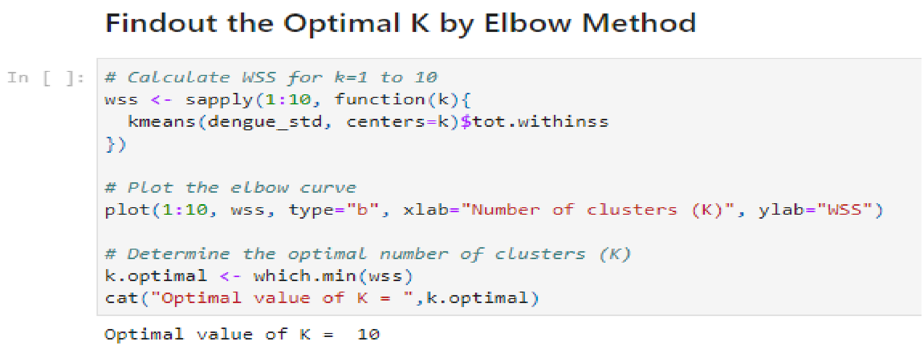

Figure 4.2.1.0: Optimal K Value

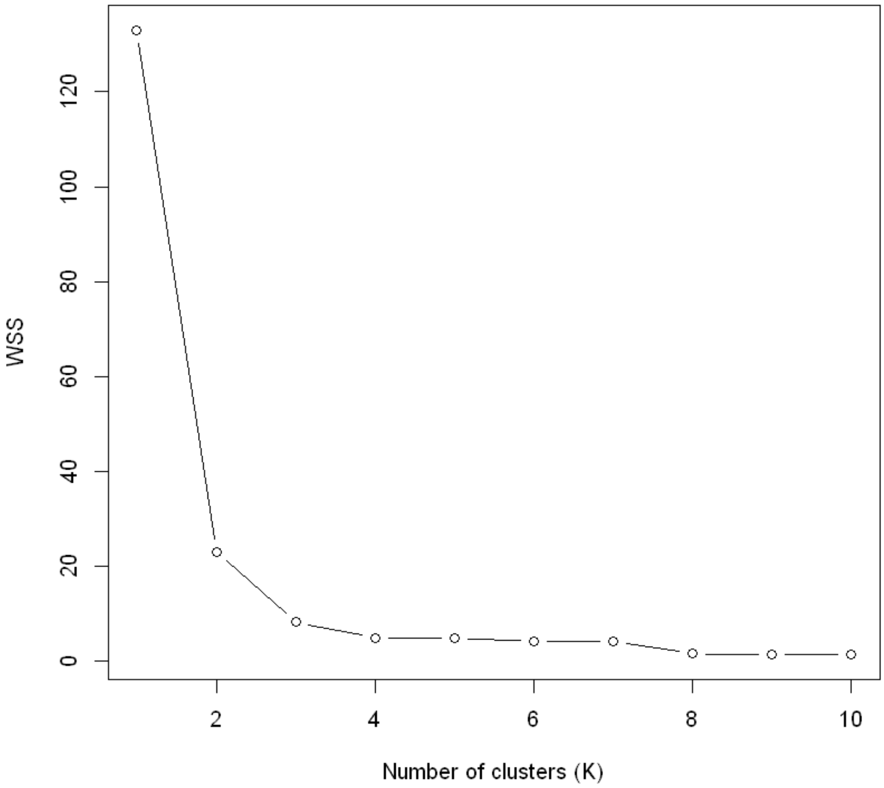

Figure 4.2.1.0: Elbow Method

4.3 Model Train : K-Means Clustering

K-means clustering is an unsupervised learning technique that segments data into K distinct clusters based on their similarity. The process of training a K Means Clustering model involves randomly initializing K cluster centroids and iteratively updating their positions until convergence.

Code extraction 6: K Means Cluster

```r
set.seed(123) # set a seed for reproducibility
kmeans_model <- kmeans(dengue_std, centers=k.optimal)
```

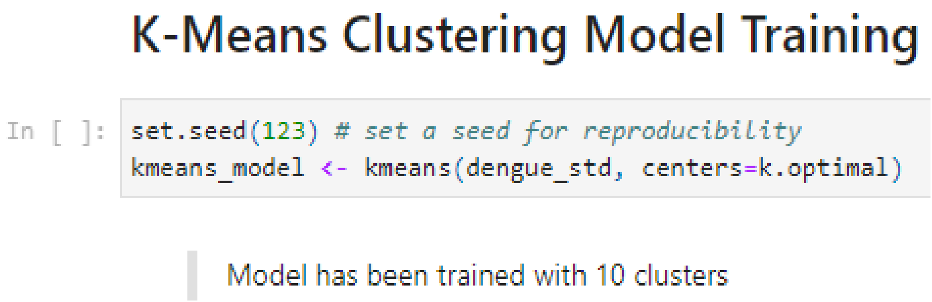

Figure 4.3.0.0: K-Means Clustering Model Training

4.3.1 Assigning Severity Levels to Incidents (Clusters)

Depending on the severity of the Dengue incident, we have assigned a name based on the severity level, that will project how the situation is based on the pattern of the environment.

Code extraction 7: Assigning Severity Levels

```
set.seed(123) # set a seed for reproducibility
kmeans_model <- kmeans(dengue_std, centers=k.optimal)
```

Assigning Severity Levels to Incidents (Clusters)

```
In [ ]: severity_levels <- case_when(
  kmeans_model$cluster == 1 ~ "Very Low",
  kmeans_model$cluster == 2 ~ "Low",
  kmeans_model$cluster == 3 ~ "Lower-Medium",
  kmeans_model$cluster == 4 ~ "Medium",
  kmeans_model$cluster == 5 ~ "Upper-Medium",
  kmeans_model$cluster == 6 ~ "High",
  kmeans_model$cluster == 7 ~ "Higher-High",
  kmeans_model$cluster == 8 ~ "Very High",
  kmeans_model$cluster == 9 ~ "Critical",
  kmeans_model$cluster == 10 ~ "Life-Threatening",
  TRUE ~ NA_character_
)
```

Severity level has been assigned successfully

Figure 4.3.1.0: K-Means Clustering Model Training

4.4 Visualize the Clusters

In this K-Means Clustering model, we have used center=10 that means there will be 10 clusters. And 10 different pattern or scenarios.

Code extraction 8: Cluster Visualization

```
clusplot(dengue_std, kmeans_model$cluster, color = TRUE,
    labels = 2, lines = 0, main = "Cluster plot of dengue
    dataset (K-means clustering)")
autoplot(kmeans_model,dengue_std, frame = TRUE)
```

Plot the Clustering Result

```
In [ ]: clusplot(dengue_std, kmeans_model$cluster, color = TRUE, labels = 2,
                 lines = 0, main = "Cluster plot of dengue dataset (K-means clustering)")
```

Figure 4.4.0.0: Clusplot Visualization

```
In [ ]: autoplot(kmeans_model,dengue_std, frame = TRUE)
```

Figure 4.4.0.0: Autoplot Visualization

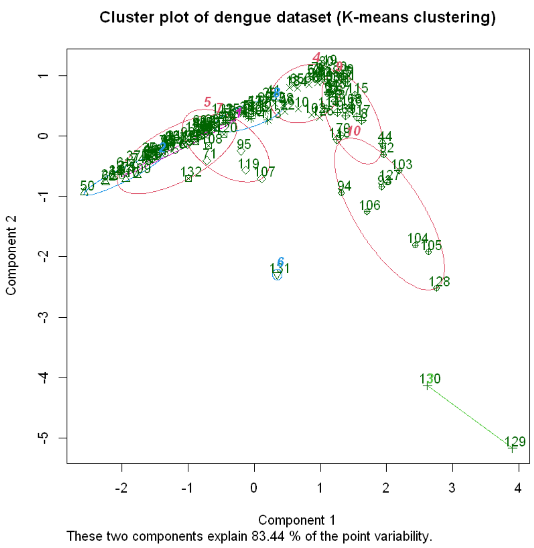

Figure 4.4.0.0: Cluster Visualization

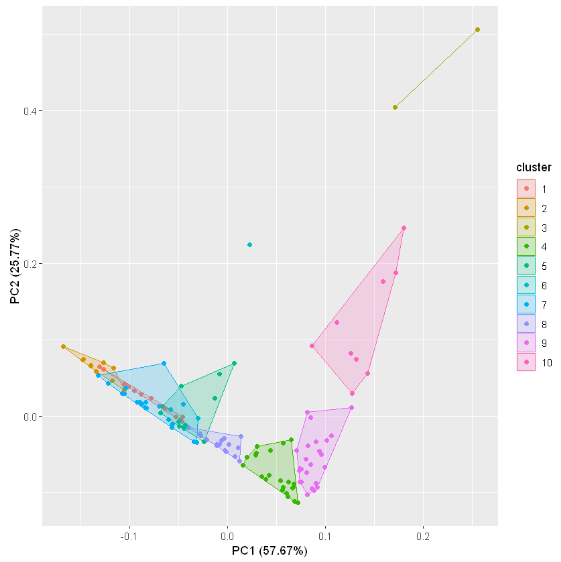

Figure 4.4.0.0: Cluster Visualization with Color

Figure 4.4.0.0: Cluster Visualization

5. Extracted Results

Based on the analysis of a dataset using K-means clustering with 10 clusters, severity levels have been assigned to each cluster based on the level of severity indicated by the features. The naming convention used is based on the severity of situations. The findings suggest that if the minimum and maximum values of air temperature, humidity, and rainfall are high, there is a higher probability of Dengue being reported. However, it is important to note that a high number of Dengue cases does not necessarily indicate a high severity level. These results provide insight into the relationship between environmental conditions and the occurrence of Dengue cases.

Result

Show Extracted Informations (As like severity level situations based on Humidy, Rainfall etc.)

Add the severity levels to the original dataset

In [ ]:
```
dengue_data$Severity <- severity_levels
tail(dengue_data)
```

A data.frame: 6 × 8

	YEAR	MONTH	MIN	MAX	HUMIDITY	RAINFALL	DENGUE	Severity
	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<chr>
129	2018	9	26.05279	33.24902	81.18693	263.16667	3087	Lower-Medium
130	2018	10	22.64687	31.37843	80.26309	181.22581	2406	Lower-Medium
131	2018	11	17.92814	29.99778	77.34087	6.50000	1192	High
132	2018	12	13.58079	26.10402	79.00390	16.22581	293	Higher-High
133	2019	1	12.06204	26.74104	77.99025	0.00000	38	Higher-High
134	2019	2	15.06038	27.98712	77.54676	56.17857	18	Higher-High

> The information that we have extracted the severity level based on situations. The pattern that we have found is,
> if Min, Max of Air Temperature, Humidity Rainfall is high then the Dengue will be high but that doesn't mean that if the Dengue case is high then the Severity is high

Figure 5.0.0.0: View Learned Patterns(Extracted Results)

Here we can see that when the Min Max of the Air Temperature and the Humidity and Rainfall in 130th instance is quite high so the dengue case goes high because this environment is suitable for dengue virus growth and spread to maximum. But the severity level is lower medium, that means that situation is not so critical, it is quite normal on that time and can also be handled by pre-preparations. But sometimes unusual time of dengue spread can be fatal. Because it not only carries Dengue but also carries other viruses also.

6. Conclusion

In conclusion, the K-means clustering analysis carried out on a dataset with 10 clusters has provided valuable insights into the relationship between environmental factors and Dengue cases. The naming convention used to assign severity levels to each cluster based on situations has proven useful in identifying patterns that indicate a higher probability of Dengue cases being reported when air temperature, humidity, and rainfall are high. However, it is important to note that the severity level should not be solely determined by the number of Dengue cases reported. Overall, these findings can help inform public health strategies aimed at preventing and controlling the spread of Dengue fever.

Bibliography

[1] FU Ahmad, SK Paul, MS Aung, R Mazid, M Alam, S Ahmed, N Haque, MA Hossain, S Paul, R Sharmin, et al. Co-circulation of dengue virus type 3-genotype i and type 2-cosmopolitan genotype in 2018 outbreak in dhaka, bangladesh. *New Microbes and New Infections*, 33:100629, 2020.

[2] Sabrina Islam, C Emdad Haque, Shakhawat Hossain, and John Hanesiak. Climate variability, dengue vector abundance and dengue fever cases in dhaka, bangladesh: a time-series study. *Atmosphere*, 12(7):905, 2021.