CLUSTERING & PCA ASSIGNMENT

QUESTION 1:

Problem Statement: Categorize the countries using some socio-economic and health factors that determine the overall development of the country and suggest at least five countries which are in direst need of aid.

Followed the below steps to arrive at the final result.

1 .Read and understood the given data set.

2. Cleaned the data

a. Checked for null values

b. Checked for duplicate records

3. Visualized the data

a. Created pair plots for all the numeric variables in the data set

b. Created box plots for all the categorical variables in the data set

4. Data Preparation

a. Tried removing outliers(before PCA and after PCA),but as the outlier elimination is leading to loss of many countries and giving incorrect results ,went ahead without eliminating the outliers.

b. Scaled all the numeric variables using standard scaler.

5. Hopkins Statistics

As the Hopkins statistics value was good, went ahead with PCA

6. PCA

a. Performed PCA on the scaled data

b.Max varaiance was explained by the first 4 PCS(based on pca.explained_variance_ratio_ and the scree plot)

c.Performed incremental PCA for efficiency with 4 components - four components were selected based on the scree plot

d.Created correlation matrix and heatmap for the principal components

e.Made a scatter plot to visualize the PCs

7. Hopkins Statistics

As the Hopkins statistics for PCs was good, went ahead with K-Means Clustering

6. K-Means Clustering

a.Did silhouette score analysis and created elbow curve to identify the number of clusters

b.Performed K-Means clustering with four clusters

c.Analyzed the clusters formed

d.Plotted bar graphs between the clusters and the features and identified the clusters with countries in the direst need of aid(Clusters 1and 2)

I e.Identified top five countries from the cluster which are in direst need of the aid

7.Hierarchical Clustering

a.Performed Hierarchical clustering with two clusters

b.Analyzed the clusters formed

c.Plotted bar graphs between the clusters and the features and identified the clusters with countries in the direst need of aid(Clusters 0)

d.Identified top five countries from the cluster which are in direst need of the aid

8. The list of top five countries which are in the direst need of aid are:

1.Congo, Dem. Rep.

2.Burundi

3.Liberia

4.Niger

5.Central African Republic

QUESTION 2:

State at least three shortcomings of using Principal Component Analysis.

1.PCA is limited to linearity

2.PCA needs the components to be perpendicular. Though in some cases that may not be the best solution

3.PCA assumes that the columns with low variance are not useful, which may not be true in some cases(especially classification problems with class imbalance)

QUESTION 3:

Compare and contrast K-means Clustering and Hierarchical Clustering.

1. K-Means and Hierarchical clustering gave the same results for the given data set

2. The number of clusters created are different for K-Means(Number of clusters=4) and Hierarchical (Number of clusters=2)

3.K-Means is a non linear process where as hierarchical clustering is a linear process. So, hierarchical clustering takes more computational time compared to K-Means clustering.

4.We don’t need to know the number of clusters to do hierarchical clustering where as we need need to know the number of clusters in case of K- Means clustering.

5.Hierarchical clustering is more suitable for small data sets as the computational time is more.Whereas K- Means clustering can be used for big data sets