Assignment Summary

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes. We have the dataset of all the countries with various factors such as health, import, export, child mortality, GDP and so on. My objective is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then I must suggest the countries which is most need of the Financial aid.

My main task is to cluster the countries by the factors mentioned above and then present the solution. The following approach was used:

I first started with necessary data inspections/data cleaning after which I performed Exploratory Data Analysis on the overall dataset through which I found the basic idea of the countries which need the financial aid, this was done using bar charts from matplotlib library.

After that I performed Outlier Analysis by plotting box plots to check if any attributes have outliers which can hamper the overall clustering algorithms. I found that GDP, income, health, import and exports had few outliers which I removed by performing outlier capping.

After that I performed Hopkins test to check if the data is good enough to get better clusters in which the score was more than 0.80 which denoted better formation of clusters.

Post that I did rescaling of dataset using Standardization technique.

To begin with the Model building process, I found out the optimal number of clusters required using techniques like Silhouette Analysis and Elbow Curve. Both analyses suggested the optimal cluster number = 3. So, I performed K-means clustering with cluster size of 3. Through which the data was divided into 3 labels. By visualization of those label, I found out that 0 label had data with lowest socio-economic countries, so I sorted that dataset by ['gdpp', 'child_mort','income'] which was our target variables.

Similarly, I performed Hierarchical Clustering using Single Linkage and Complete Linkage method and followed the same steps to get such countries.

After using both clustering techniques, the top countries which were in most need of financial aid are **Burundi, Liberia, Congo, Dem. Rep., Niger, Sierra Leone.**

Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

K means clustering is the process of partitioning a set of data objects into subsets. Each subset is a cluster such that the similarity of data points between two cluster is greater and the similarity of data points within the clusters is less.

Whereas, while partitioning methods meet basic clustering requirements of organizing the set of objects into several exclusive groups, in some situations we may want to partition our data into groups at different levels such as in a hierarchy. This method is called as Hierarchical clustering. This clustering method works by grouping data objects into a hierarchy or "tree" of clusters.

b) Briefly explain the steps of the K-means clustering algorithm.

Step 1: Initialize data centres

Step 2: Assign observations to the closest data centres based on minimum distance to the cluster centre.

Step 3: Move the centroid.

Step 4: Repeating step 3 and 4 until the centroids stop moving.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

There are two methods: Elbow Curve and Silhouette method

Elbow Curve:

Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.

For each k, calculate the total within-cluster sum of square (wss).

Plot the curve of wss according to the number of clusters k.

The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Silhouette Method:

Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.

For each k, calculate the average silhouette of observations (avg.sil).

Plot the curve of avg.sil according to the number of clusters k.

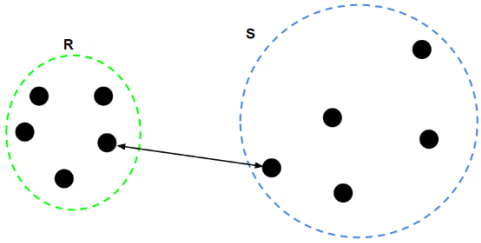
The location of the maximum is considered as the appropriate number of clusters.

d) Explain the necessity for scaling/standardisation before performing Clustering.

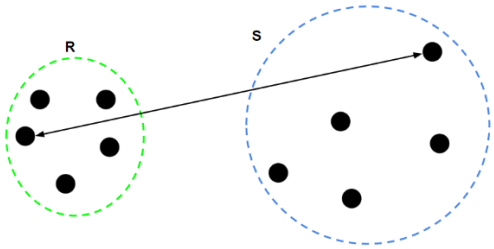
Scaling is important before clustering because it controls the variability of the dataset, it converts data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms. All distance-based algorithms are affected by the scale of the variables. Consider your data has an age variable which tells about the age of a person in years and an income variable which tells the monthly income of the person in rupees. Here the Age of the person ranges from 25 to 40 whereas the income variable ranges from 50,000 to 110,000. Thus, it is important to scale the features to get them into same range of values in order to apply clustering on it.

e) Explain the different linkages used in Hierarchical Clustering.

1. **Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.



2. **Complete Linkage:** For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S



3. 3. **Average Linkage:** For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.