Question 1: Assignment Summary

Ans : Topic – Clustering of Countries

Problem Statement

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.

After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Business Objective -

Identify top countries that are in direst need of aid.Objective is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then suggest the countries which the CEO needs to focus on the most.

Steps to Procedure in the Assignment

1. We will start out by Understanding and reading the dataset.
2. This will be followed by changing the columns import, export and health from percentage to actual values since they are given as percent of GDPP.
3. Than we will move on towards Clustering.
4. This would first involve Data preparation which includes to things the first being Outlier treatment and second being carrying out the Hopkins Check.
5. Finally we will go forward with actual clustering.
6. We will start with K-Means clustering

7. In this technique we will run K-means and choose K using both Elbow and Silhouette score.
8. Than run K-Means with the chosen K and visualise the clusters.
9. Clustering profiling using three columns gdpp, child_mort and income.
10. Perform Hierarchical Clustering.
11. This will include both single and complete linkage.
12. Choosing any one method based on results obtained.
13. Visualising the clusters.
14. Cluster profiling using the same three columns which were used earlier for K-Means.
15. Country Identification based on the above analysis
16. Identification parameters will be socio-economic and health factors

Question 2: Clustering

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

Ans) 1. Both K-means Clustering and Hierarchical Clustering are Unsupervised learning Techniques employed for Machine learning purposes

2. K-means clustering is carried out by using pre-specified number of clusters therefore value of K should be known in advance whereas in Hierarchical Clustering tends to build a hierarchy of clusters without actually having any definite value of K.

3. In K-means clustering mean or median is used as cluster centre to represent a cluster whereas in Hierarchical Clustering we begin with 'n' and begin combining clusters in a sequence until only one cluster is left.

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

Ans) Steps of the K-means clustering algorithm are :-

1. First step is to specify the number of clusters K.
2. Initialising the centroids by shuffling the dataset and random selection of K data points for centroids without replacement.
3. Iteration process is done until there is no change is observed to the centroids. This is done by calculating the sum of squared distance between data points and the centroids.
4. Each data point is than assigned to the nearest cluster (centroid).
5. By calculating the Average of all data points that belong to each cluster in order to find the Centroids for the clusters.

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

Ans) The value of 'k' is chosen in K-means clustering by two methods :-

1. Elbow Method – It is widely used for determining the number of clusters.

   In this method Within cluster sum of squared errors (WSS) for different values of k is calculated than that value of k is chosen for which WSS first starts to decrease. The plot of WSS vs k this can be seen as an Elbow.

2. Silhouette Method – It measures how similar a point is to its own cluster compared to other clusters. The range of the Silhouette value is between +1 and -1. A high value is preferred which is indicative of the point being placed into a correct cluster.

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

Ans.) The necessity for scaling/standardisation before performing Clustering :-

1. Standardisation is important on datasets where each variable has a different unit
2. In certain datasets where scales of different variables are very different from one another.
3. It helps to make the relative weight of each variable to equal by conversion into a unitless measure

e) Explain the different linkages used in Hierarchical Clustering.

Ans.) Different linkages used in Hierarchical Clustering are :-

1. Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters.
2. Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters.
3. Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance.