Assignment: Part II

Question 1: Assignment Summary

Problem Statement

After the recent funding programs, Help foundation 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.

Your job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most

solution methodology

1. Reading and understanding the data
2. Data transformation
3. Outliers treatment
4. Data standardization
5. Modelling with Kmeans and Hierarchical Clustering. Kmeans clustering produced better result.
6. Analysis of the clusters and identify the ones which are in dire need of aid
7. Resulted clusters analysed against **gdpp**, **child_mort** and **income**
8. Differentiation of the clusters of developed countries from the clusters of under-developed countries.
9. visualizations on the clusters that have been formed
10. Final listing of 5 countries which are in direst need of aid

Question 2: Clustering

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

Ans :

The k-means algorithm is require the value *k initially* , which is the number of clusters that you want to create. The algorithm begins by creating *k* centroids. After this ,It then iterates between an assign step and an update step where each centroid is *updated* to become the mean of all the samples that are assigned to it. This iteration keep on until some stopping criteria is met and no change of centroid happens.

Hierarchical clustering, forms clusters incrementally.The algorithm begins by assigning each sample to its own cluster. At each step, the two clusters that are the most similar are fused; the algorithm continues until all of the clusters have been merged. Unlike k-means, we don't need to specify a *k* parameter: once the dendogram has been produced, we can navigate the layers of the tree to see which number of clusters makes the most sense to business requirement.

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

Ans :

The steps of the Kmeans algorithm are:

1. Start by choosing K random points the initial cluster centres.
2. Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
3. For each cluster, compute the new cluster centre which will be the mean of all cluster members.
4. Now re-assign all the data points to the different clusters by taking into account the new cluster centres.
5. Keep iterating through the step 3 & 4 until there are no further changes possible.

At this point, you arrive at the optimal clusters.

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

Ans :

There are two methods to choose the value of K.These methods are:

1.The Elbow 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 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.

2.The 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.

Ans :

- Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale helps in this process.
- The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.

e) Explain the different linkages used in Hierarchical Clustering.

Ans :

- **Single Linkage**

Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters

- **Complete Linkage**

Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

- **Average Linkage**

Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

THANK YOU