Question-1: Summary of Assignment

Answer – We have to identify countries that are in dire need of financial aid, when disaster of calamity comes, so that they have basic needs being made. We have data of 167 countries. Data is about various socio-economic indicators,

Dataset was clean and had no null values. Also, the data has no absurd values in any features. We can directly proceed with clustering.

First, we did EDA for the data, and plotted correlation plot, we found the data has 2 kinds of indicator – positive indicator (income, gdpp, health, life expectancy) and negative indicator (child_mort , total_fer, inflation) . A negative correlation between positive and negative indicator is seen in correlation plot.

Given, our objective to find under-developed nations, and only 167 data points we can hardly remove any outliers based on negative features so, we only removed datapoints based on GDPp, a positive feature. So, we removed most developed countries in such a way.

Then we scaled the data by using MinMax Scaler, Scaling is important in clustering, since the cluster distances should be impartial towards features attributes.

After scaling, we made a base k-means clustering model, with k (number of clusters = 3), we also, found optimal k by elbow method – again k we got 3. So, without need to train other model, we used this. So, in this model where we had 167 datapoints we obtained 46 underdeveloped countries.

Then we also, did same for data with outliers removed and obtained 42 underdeveloped countries. All of which where already in our first model, without removing outliers.

We proceed with hierarchical clustering model, used the outliers removed data. We did agglomerative hierarchical clustering with single and complete linkage. Single linkage solution was skewed and so not used. While complete linkage solution, was sensible. So, we analysed our result for it. We obtained 39 underdeveloped countries in this model.

We found in our clustering, that underdeveloped countries have very low GDPp , income and health parameters, also negative features like inflation, child mortality and total fertility were socio-economic problems of this countries.

Question–2:

Sub-questions

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

Kmeans clustering –

- We cluster similar data points into a same cluster and dissimilar datapoints in different clusters.
- Cluster similarity is based on distance from cluster centre. Distance measures – Euclidean, Manhattan, etc
- Number of clusters are not known and optimal cluster to be obtained by elbow method (SSD) or Silhouette analysis.

Hierarchical clustering –

- In Agglomerative hierarchal cluster, we create a tree like structure called dendrogram based on the least distance between initial cluster. Initial cluster are equal to total data points and clusters are fused according to linkage rule, ex- simple linkage and complete linkage.
- Here we don't need to specifiy number of cluster. Instead, we cut dendrogram at different threshold values to obtain various number of cluster.
- It is more interpretable than kmeans clustering, it takes more time to run than kmeans. Because, number of computations are more as only one cluster is fused in each iteration of growing dendrogram.

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

Step-1: Scale the data

Step-2: Initialize cluster centre position of k-clusters randomly from datapoints.

Step-3: Assign cluster numbers to datapoints, based on rule of minimum distance from cluster centre.

Step-4: Recompute cluster centre based all cluster datapoints.

Step-5: Run step 2 to 4 until, the cluster center positions do not move significantly in consecutive iterations.

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

Mathematically, k is chosen by elbow method (SSD) or Silhouette Analysis. However, from business point of view, the objective is that we have to use cluster to interpret the segments of data space. So, the number of cluster should be meaningful in business terms. If cluster number is high, we might have to put similar data points in different clusters, with very little difference. Also, when number of cluster is very less, we are not solving the clustering problem objectively. Example – it is useful to cluster countries as under-developed, developing and developed this makes more sense than say doing clustering in 5 buckets.

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

Scaling is useful because, the kind of distance measure – Euclidean we use, is sensitive to variance in the features. Features for which range is more the weight of those features in calculating distance will be more, and vice versa. So, by scaling we are giving equal importance to all features.

e) Explain the different linkages used in Hierarchical Clustering.

Idea of linkage is used when we have to fuse 2 clusters, in at least 1 out of which has more than 1 points in them.

Single Linkage – Two clusters are fused by rule of single linkage when, the minimum distance between any two pair of points out of the two clusters is least.

Complete Linkage – Two clusters are fused by rule of complete linkage when, the maximum distance between any two pair of points out of the two clusters is least.

Average Linkage – Average distance between all possible pair of two points (one from each cluster) is least.