Clustering Assignment – HELP International

By Saranga Veeramangal Hebbar

Question 1: Assignment Summary

Problem Statement: HELP International is an NGO that is committed to fight against poverty and it has recently got $10,000,000.00 funding from different sources. The NGO's CEO intends to spend this money for the countries in real need 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.

Solution Expected: As a Data analyst, will have to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then will have to suggest the countries which the CEO needs to focus on the most.

Solution: The steps are broadly:

1. Read and understand the data:

   The dataset contains socio-economic attributes of 167 countries like GDPP, import & export percentage, health expenses, child mortality, life expectancy etc. The import, export and health expenses are provided in the percentage values against the GDPP and hence to know the right amount, it has to be calculated by taking percentage value against the GDPP.

2. Clean the data

   The dataset is cleaned in this step and EDA is performed to analyze the data. In brief, I have performed Univariate and Bivariate analysis to understand the dependency between the variables and outliers within the variables. We can conclude with the analysis that there are more than 75% of countries whose income > $10000 and there are less than 20 countries whose child mortality is > 100.

3. Prepare the data for modelling
   For the purpose of building the modelling, GDPP, income and child mortality attributes are picked and the model is fit for the dataset to build the clusters. For the outlier analysis, it is evident that there are quite a few countries that are having income range lying outside the outlier.
4. Modelling
   a. K-Means Modelling

- First create the elbow curve by using the range of clusters from 2 to 8
- Then Use the Sillhoutte method to identify the ideal number of clusters
- Based on the 2 methods finalize the number of clusters to be used for the K-Means clustering. Here we find that cluster count =3 is an ideal number based on the above 2 methods
- Use the number of clusters based on the above step and build the K-Means clustering algorith m on the dataset. The 3 clusters formed will have GDPP, Income are inversely related to child mortality and the cluster would be following the same trend

   b. Hierarchical modelling

The hierarchical modelling is done with single linkage and complete linkage. By drawing the dendrogram and cutting the dendrogram at a right cluster count will help forming the clusters for the modelling. In the single linkage, though the cluster count = 3 will form the 3 clusters, it will have only 1 value each in 2 out of the 3 clusters where as the 3rd one will have the entire population, and hence it cannot be the right clustering. The complete linkages will build the clusters in the right need and have the clusters with GDPP/income inversely related to child mortality

Here too, the number of clusters is chosen at 3 to get the right bucketing of the data

5. Cluster Profiling

The cluster_id = 0 has very low income and GDPP value whereas it has very high child mortality rate. Hence, we need to pick the countries in the cluster #1. Identify the top 5 countries in the c luster that has lowest GDPP & income and very high child mortality rate.

6. Final analysis and recommendation

Based on the profiling, the cluster=0 is chosen and the bottom most countries with very low GDPP, Income but very high child mortality is chosen and would be shared with the NGO to provide the assistance. Finally, countries : Burundi, Liberia, Congo, Niger and Sierra Leone are the countries that would need the assistance from the NGO to help reducing the higher child mortality rate

Question 2: Clustering

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

Ans:

#	Characteristics	K-Means	Hierarchical
1	Modelling Approach	Top-down	Bottom-Up
2	Pre-requisite	Number of clusters needs to be provided	Not needed to limit any specific clusters. Based on dendrogram, the number of clusters can be limited
3	Processing	Median or mean can be used as cluster center initially and then the cluster center is fine-tuned based on the other data points	In this approach, every data points are initially cluster centers (n clusters) and then sequentially combined to merge the similar clusters until there is only one cluster is formed
4	Computation process	Computationally less intensive and better suited for larger dataset	Computationally intensive process and better suited for not so large dataset
5	Reproducibility	Since the modelling is based on the initial set of centroids that are chosen, the results produced by the algorithm might differ	Since this is bottom up approach of modelling, the results would be consistent and reproducible
6	Advantages	✓ Convergence is guaranteed	✓ Ease of handling of any forms of similarity or distance ✓ Consistency and reproduceable

		✓ The algorithm can produce the clusters of different size and shapes	
7	Disadvantages	× K-value is difficult to predict × Doesn't work well with global cluster	× High computation and storage intensive × For large datasets, the algorithm can be expensive and slow

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

Ans:

The Objective of the K-Means algorithm is to minimize the Euclidean distance that each datapoints has from the centroid of the cluster. This is also known as intra cluster variance. The result of the algorithm ensures that the inter-cluster distance is maximum.

Step 1 – Number of clusters (K) needs to be provided to the algorithm

Step 2 – Randomly initialize and select the k-points (mean)

Step 3 – Use Euclidean distance to identify the centroid of each cluster

Step 4 – Steps 2 & 3 are iterated over until an optimal centroid which is the assignment of data points to the clusters that are not changing any more

4.1 – sum of squared distance between data points and centroids would be computed. (Euclidean distance)

4.2 – Assign each data point to the cluster that is closer than other cluster (centroid).

4.3 – Compute centroids for the clusters by taking the average of all data points of that cluster.

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: Since the K-Means clustering needs the pre-requisite of having the number of clusters to be provided, it is very important to identify how many clusters must be there in the algorithm.

The 2 statistical methods that can be used to identify the ideal cluster count are

i) Elbow curve : The K-means clustering is run for a range of cluster count and the graph is plotted with k against the distortions. The k value where we find "elbow" in the graph is taken as the k clusters

ii) Silhouette analysis: The silhouette score is provided for each cluster values in the cluster range and right cluster size is picked where the count is optimal

The Business methods to identify the ideal cluster count is basically to understand the business requirements (like for retail : RFM attributes being analyzed and the high value, low value of these attributes being clearly bucketed into clusters) and the cluster counts are identified based on these business needs

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

Ans: The attributes that are to be used for the modelling might not be in the same range of values. If these values are belonging to a different range of values, then the comparison or bringing them together in a graph would not be possible. Hence to ensure these values are representable within the same range of values, all the attributes are scaled or standardized. We use StandardScalar function to initialize and fit and transform the data into standardized values

e) Explain the different linkages used in Hierarchical Clustering

ANS:

In Hierarchical clustering, there are 2 types of linkages

i) **Single linkages** : In this approach, the clustering is done in the bottom-up way and at each step, 2 clusters that are closest pairs of elements and are not part of the same clusters are combined together to form a bigger cluster. This process is repeated until the final 1 single cluster is formed. Drawback of this linkages is that though we can cut the dendrogram at any cluster numbers, there will be many clusters that are thin clusters (with only 1 datapoints in it).

ii) **Complete linkages:** At the beginning, each datapoints are assigned as clusters of its own. Then these clusters are sequentially combined together to form a bigger cluster until all the elements end up in the same cluster. The result of clustering can be visualized well in the dendrogram which shows the sequence of cluster fusion. This linkage would produce a better clustering when the dendrogram is cut at any point as the data points are uniformly distributed at that level and would not end up having many thin clusters like in Single linkages