Clustering Assignment

Part II - Subjective Questions

ASSIGNMENT SUMMARY

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.

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.

We need to categorise the countries using some socio-economic and health factors that determine the overall development of the country and suggest the countries on which the CEO needs to focus on the most.

Solution Methodology

The method used to identify the countries that had low development was Clustering.

Clustering is a technique that involves grouping of data points, we have classified the countries using some socio-economic and health factors.

The 3 major Factors that we have used are:

- Income
- Child Mortality Rate
- GDPP

Steps Involved

1. **Data Understanding:**

   After reading and understanding the data, we noticed that values in the Columns Export, Imports and Health were given as GDP Percentage per capita. So, we converted the values into actual values.

2. **Data Preparation:**

   We had performed EDA to check the flow of the data and check how the data was distributed.

3. **Outlier Analysis**

   The outliers of the data were visualised using Box Plots and the columns that had many outliers were treated by capping the outliers at 0.99 quantile, as doing a harsh outlier treatment would affect the analysis.

4. **Clustering - K Means and Hierarchical**
   a. **K-Means Clustering:**

      K Means Clustering is a method that aims to partition 'n' observations of a data into 'k' clusters.

      Each cluster belongs to the nearest cluster centroid.

   b. **Hierarchical Clustering:**

      Hierarchical Clustering is another algorithm that groups similar objects into clusters.

      The endpoint is a set of clusters, where each cluster is distinct from each other, and the objects within each cluster are broadly similar to each other.

      We clustered the data using both Single and Complete Linkage.

The countries in the given dataset were divided into 3 clusters, based on which we have attained the list of countries which required the most attention from HELP International.

5. **Cluster Visualisation:**

    The clusters were visualised using scatter plots, bar plots and box plots based on Factors like Income, Child Mortality Rate and GDPP.

6. **Country Identification:**

    Looking at these graphs we were able to get a list of around 50 countries that had Low Development.

We also saw that the list of countries attained from both K-Means Clustering and Hierarchical Clustering were similar.

CLUSTERING QUESTIONS

Compare and contrast K-means Clustering and Hierarchical Clustering.

- ➔ Performing Hierarchical clustering on large data will not give us precise results, as the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).
- ➔ The results produced by performing the K-Means algorithm multiple times might differ, since we start with random choice of clusters. While results are more concrete and reproducible in Hierarchical clustering.
- ➔ K Means clustering requires prior knowledge of that is the number of clusters we want to divide our data into. But, we can stop at whatever number of clusters we find appropriate in hierarchical clustering by interpreting the dendrogram.

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

1. **Data Understanding:**

   We must read and understand the data. Check if they have the right data types or if they have any missing values or unusual data.

2. **Data Preparation**

   We need to handle the missing values and correct the data types of values if incase they are not in the proper format.

   a. Outlier Analysis:

      Checking for outliers is one of the most important steps for Clustering as not treating the outliers would give us unwanted clusters. We have two types of Outliers, Statistical and Domain Specific. We treat the Statistical Outliers based on statistical interpretation but for Domain Specific outliers we have to understand the domain and the client for whom we are performing the analysis and based on their required we chose to remove or keep the outliers.

   b. Data Scaling:

      We Scale the data to eliminate redundant data and ensure that good quality clusters are generated which can improve the efficiency of clustering algorithms.

   c. Hopkins Check:

      We check the Hopkins Statistic to check the cluster tendency.

3. **Perform Clustering:**

   We first choose the optimum value for K, using the Elbow Curve and the Silhouette Score and then run the K-Means algorithm on the original data for clustering.

4. **Cluster Visualisation**

   We visualise the clustered data using Scatter Plots and Bar Plots to analyse and classify the data based on the given problem statement.

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

The value of K is chosen by using the Elbow Curve and Silhouette Score.

1. **Elbow Curve:**

   We compute the Sum of Squared Distances in the data and plot a curve for the SSD. We chose the value of K where the curve looks like it is significantly dropping.

2. **Silhouette Score:**

   We compute the Silhouette Score for a range of clusters and the value corresponding to the cluster that has the highest Silhouette Score is chosen as the K Value.

Explain the necessity for scaling/standardisation before performing Clustering.

Scaling/Standardisation is used to eliminate redundant data to ensure that good quality clusters are generated which can improve the efficiency of clustering. So it becomes necessary to scale the data before clustering as Euclidean distance is very sensitive to the changes in the data points.

Explain the different linkages used in Hierarchical Clustering.

1. **Single Linkage:**

   In Single Linkage Clustering, the data points are merged in each step which have the smallest distance.

2. **Complete Linkage:**

   In Complete Linkage, the data points are merged in clusters which have maximum distance between any 2 points in the clusters.

3. **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.