Name: Badavath Srikanth
Registration number:12019441

Assignment type: Subjective question

Question 1: Assignment Summary

Briefly describe the 'Clustering of Countries' assignment that you just completed in 200–300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly (what EDA you performed, which type of clustering produced a better result and so on).

Note: You do not have to include any images, equations or graphs for this question. Just text should be enough.

Answer:

The aim of the assignment titled "Clustering of Countries" was to identify countries that require aid urgently based on three key features - child mortality, income, and gdpp. Initially, we analyzed the data set and noted that it comprised 166 countries, each with nine features. Since we knew our primary focus features, we began the assignment with ease. We started with data understanding, cleaning, and preparation, and examined the dataframe structure upon importing the data from the CSV file. We searched for null values and outliers in the dataset.

After data standardization, we performed Principal Component Analysis (PCA), as directed in the pre-assignment session. Outlier analysis was conducted after PCA, which resulted in the removal of two countries from the dataset - Myanmar due to low import value and Nigeria due to inflation. We ran PCA again and determined that five clusters were adequate at PCA(0.94).

We initiated the clustering process by calculating the Hopkins Statistics, which showed that the data had a high tendency to cluster. Next, we used silhouette score analysis to determine the optimal number of clusters and performed a sum of squared error test. We plotted dendrograms for single and complete hierarchical clustering and determined that the clusters could be either three or four. We opted for four clusters initially but were not satisfied with the results. When we switched to three clusters, we noticed that cluster 1 consistently exhibited poor performance in all three fronts - child mortality, income, and gdpp. We also conducted line and bar plot analyses to examine the cluster closely.

Despite our analysis with four clusters, we were not convinced, so we performed a three-cluster analysis, which produced similar results for cluster 1. We compiled a list of the top n countries that were suffering in terms of child mortality, income, and gdpp and recommended common countries from the three lists.

Question 2: Clustering

This question is further divided into multiple questions as stated below:

1.Compare and contrast K-means clustering and hierarchical clustering.

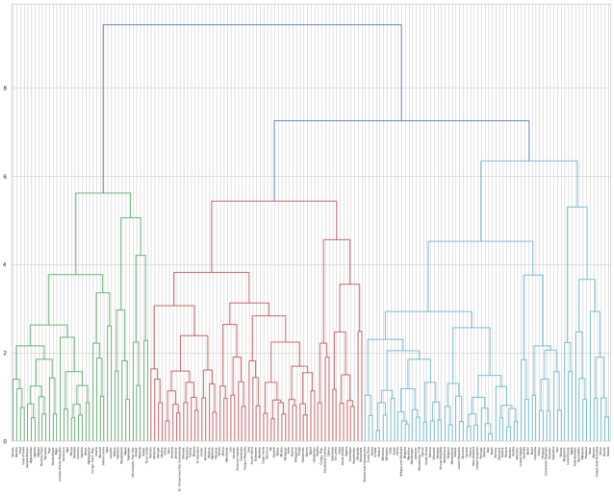
Answer:

Hierarchical clustering is an algorithm that constructs a hierarchy of clusters. It begins by assigning all data points to individual clusters. Then, the algorithm merges the two closest clusters into a single cluster. This process continues until only one cluster remains.

The output of hierarchical clustering can be visualized using a dendrogram. A dendrogram displays the relationships between clusters and can be interpreted as follows:

Name: Badavath Srikanth
Registration number:12019441

Assignment type: Subjective question



K-means clustering is an unsupervised learning method that is utilized when working with unlabelled data. This algorithm aims to identify groups within the data, with the number of groups specified by the variable K. The algorithm operates iteratively to allocate each data point to one of K clusters based on their provided features. Data points are clustered based on similarity in their features. The K-means clustering algorithm yields two outputs:

The centroids of the K clusters, which can be utilized to label new data.

1.Labels for the training data, assigning each data point to a single cluster.

2.Rather than defining groups prior to data analysis, clustering allows for the identification and

examination of organic groups.

The "Choosing K" section below outlines how the number of groups can be determined. The centroid of each cluster is a collection of feature values that define the resulting groups. Analyzing the centroid feature weights can provide a qualitative interpretation of the characteristics of each cluster.

K-means clustering and hierarchical clustering differ in several ways:

- Hierarchical clustering struggles to handle large datasets, while K-means clustering is better equipped for such scenarios. This is because the time complexity of K-means clustering is linear (O(n)), whereas hierarchical clustering has a quadratic time complexity (O(n^2)).

- The results obtained from running K-means clustering multiple times can vary due to the random initialization of cluster choices. In contrast, the results obtained from hierarchical clustering are reproducible.
- K-means clustering works best when the shape of the clusters is hyper-spherical (e.g., circular in 2D or spherical in 3D).
- K-means clustering requires prior knowledge of the number of clusters (K) into which the data should be divided. In contrast, hierarchical clustering allows you to stop at any appropriate number of clusters by interpreting the dendrogram.

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

Answer:

The K-means clustering algorithm is an iterative algorithm that partitions a given dataset into K clusters. The steps of the K-means algorithm are as follows:

1. Specify the number of clusters K.
2. Initialize K centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Assign each data point to the nearest centroid, which forms K clusters.
4. Calculate the mean of all data points in each cluster to determine the new centroid of each cluster.
5. Repeat steps 3-4 until the centroids no longer move or a maximum number of iterations is reached.
6. The final K clusters represent the grouping of the data based on feature similarity.
7. The centroids of the K clusters can be used to label new data.
8. Labels for the training data can be assigned based on which cluster the data point belongs to.

Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

Name: Badavath Srikanth
Registration number:12019441
Assignment type: Subjective question

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

Answer:

The elbow method is a useful tool for determining the optimal number of clusters in K-means clustering, as it provides a visual representation of the trade-off between the number of clusters and the distortion of the data. It is important to note, however, that the elbow method is not always straightforward to interpret, as the elbow point may not always be clear or may not exist at all. In such cases, other methods such as silhouette analysis or gap statistic can be used to determine the optimal number of clusters. Additionally, it is important to note that the optimal number of clusters may depend on the specific problem and context in which the clustering algorithm is being applied. Therefore, the elbow method should be used as a guide rather than a definitive answer.

Example:

4.Explain the necessity of scaling/standardisation before performing clustering.

Answer:

One of the critical issues in clustering is determining a good measure of distance between cases. If you have features with vastly different scales, you may end up with one feature driving most of the distance between cases. For instance, if you cluster people based on their weights in kilograms and heights in meters, a 1kg

difference might not be as significant as a 1m difference in height. Additionally, you may get different clustering results if you use weights in kilograms and heights in centimetres. If you answered "no" to the former and "yes" to the latter, you should consider scaling your data.

If you plan to use k-nearest neighbors with a Euclidean distance measure, it is essential to scale all features to weigh in equally. This is because the Euclidean distance measure is sensitive to magnitudes.

Scaling is also critical when performing Principal Component Analysis (PCA). PCA aims to identify the features with maximum variance, but if some features have higher magnitudes, they may skew the PCA towards those features. Therefore, it is crucial to scale the data before performing PCA to ensure that each feature contributes equally to the analysis.

In summary, scaling or standardization is necessary before performing clustering because it helps to ensure that each feature contributes equally to the distance calculation and analysis. Without scaling, the clustering results may be biased towards features with larger magnitudes, leading to inaccurate or irrelevant results. Additionally, scaling is crucial when using k-nearest neighbors with a Euclidean distance measure or performing PCA, as these methods are sensitive to feature magnitudes and can be skewed towards features with higher magnitudes if not properly scaled. Therefore, scaling the data helps to ensure that the analysis is fair and accurate, and produces meaningful results.

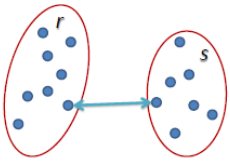
5.Explain the different linkages used in hierarchical clustering.

Answer:

Single linkage hierarchical clustering defines the distance between two clusters as the shortest distance between two points in each cluster. In other words, the distance between clusters "r" and "s" to the left can be determined by measuring the length of the arrow connecting their two closest points. This type of clustering is called "single linkage" because it focuses on the single shortest distance between two clusters.

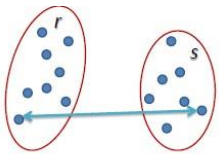
Name: Badavath Srikanth
Registration number:12019441

Assignment type: Subjective question



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

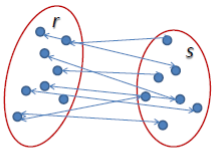
The distance between two clusters in complete linkage hierarchical clustering is determined by the longest distance between two points in each cluster. For instance, the distance between clusters "r" and "s" on the left is calculated by measuring the length of the arrow between their two furthest points.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average linkage hierarchical clustering calculates the distance between two clusters based on the average distance between every point in one cluster to every point in the other cluster. For instance, when determining the distance between clusters "r" and "s" in the image to the left, the average length of each arrow connecting the points of one cluster to the other is used.

This clustering method considers the average distance between all pairs of points in different clusters, instead of focusing on the distance between the closest or farthest points. As a result, it can be less sensitive to outliers or noise in the data than single or complete linkage clustering. However, it can also be more computationally intensive, especially for large datasets.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$