HELP INTERNATIONAL IS AN INTERNATIONAL HUMANITARIAN NGO

Clustering Assignment

- Mohammed Anas Javeed

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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
- After the recent funding programmes, 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.

BUSINESS GOAL

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

ASSIGNMENT SUMMARY

I have used PCA method to reduce the variables involved and then did the clustering of countries based on those Principal components and later I identified factors like

1. child mortality
2. income etc…

these play an important role in deciding the development status of the country and built clusters of countries.

Clusters based on which have been identified the final list of countries which are in real need of aid.

The list of countries may change as it is based on the factors like Number of components taken, Number of Clusters chosen, Clustering method used …. which I have used to build the model.

K-MEANS CLUSTERING AND HIERARCHICAL CLUSTERING

K-means and Hierarchical clustering both were analysed and found clusters formed are not identical. The clusters formed in both the cases are not good but its great in K-means as compared to Hierarchical.

So, we proceeded with the clusters formed by K-means and based on the information provided by the final clusters the final list of countries which are in need of aid has been decided.

K-MEANS CLUSTERING ALGORITHM

- K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
- The algorithm works as follows:
  - First we initialize k points, called means, randomly.
  - We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
  - We repeat the process for a given number of iterations and at the end, we have our clusters.

VALUE OF ‘K’ CHOSEN IN K-MEANS CLUSTERING? EXPLAIN BOTH THE STATISTICAL AS WELL AS THE BUSINESS ASPECT OF IT.

- **Elbow Curve to get the right number of Clusters**
  - A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.

DIFFERENT LINKAGES USED IN HIERARCHICAL CLUSTERING

- **Single Linkage:**
  - In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.

- **Complete Linkage**
  - In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.

NECESSITY FOR SCALING/STANDARDISATION BEFORE PERFORMING CLUSTERING

- **Rescaling the Features**
  - Most software packages use SVD to compute the principal components and assume that the data is scaled and centered, so it is important to do standardization/normalisation. There are two common ways of rescaling:
    - Min-Max scaling
    - Standardization (mean-0, sigma-1)