**International Institute of Information Technology Bangalore**



Subject: - Machine Learning

Topic: - Clustering Assignment

Submitted By:-

Ajinkya Shakkarwar

1. **Briefly describe the "Clustering of Countries" assignment that you just completed within 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)**

Ans: **- Problem Statement:**

To 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.

1. **Objective:**

• To categorise the countries using some socio-economic and health factors that determine the overall development of the country.

• To suggest the countries which the CEO needs to focus on the most.

1. **Method followed**: -

• It was found that there were no null values

• There were also no duplicate values for country

• There were a few outliers and they were treated later on during PCA

• The data was standardized for Principal Component Analysis

1. **Screeplot:**

* 4 components are good enough to get a 95% of variance in the data. So PC is selected to be 4.

1. **Clustering:**

• Both the methods K means and Hierarchical Clustering was used on the 4 PCA components

• For K means , K= 3 was taken using the elbow dip and silhouette analysis .

• While doing the Hopkins Statistics a value of 0.77 was attained.

• If the Hopkins Statistics values are:

- 0.3 : Low chase of clustering - around 0.5 :

- Random - 0.7 - 0.99 : High chance of clustering

-Finally using all these values clusters of 3 were formed and the countries are split into clusters.

**Question 2: Clustering**

1. **Compare and contrast K-means Clustering and Hierarchical Clustering.**

**ANS: -** • K Means needs a prior knowledge of number of centroid (K) whereas hierarchical cluster do not need these kinds of parameters. Cut tree () function is used to create the number of clusters of any choice.

. • In K Means clustering the algorithm will calculate the centroid each time.

• K Means is fast compare to hierarchical clustering.

• Hierarchical clusters need more ram to run.

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

**ANS: -** Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

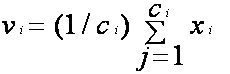
1) Select *‘c’* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

where, *‘ci’* represents the number of data points in *ith* cluster.



5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

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

**ANS:** The basic idea behind k-means consists of defining k clusters such that total**within-cluster variation (or error) is minimum.**

A cluster center is the representative of its cluster. The squared distance between each point and its cluster center is the required variation. The aim of k-means clustering is to find these k clusters and their centers while reducing the total error.

  There are two methods that can be useful to find this mysterious k in k-Means.

These methods are:

1. **The Elbow Method**
2. **The Silhouette Method**
3. **Explain the necessity for scaling/standardisation before performing Clustering.**

**ANS: -** When we standardize the data prior to performing cluster analysis, the clusters change. We find that with more equal scales, the Percent Native American variable more significantly contributes to defining the clusters. Standardization prevents variables with larger scales from dominating how clusters are defined.  
If you have two features, one where the differences between cases is large and the other small, are you prepared to have the former as almost the only driver of distance?

So for example if you clustered people on their weights in kilograms and heights in metres, is a 1kg difference as significant as a 1m difference in height? Does it matter that you would get different clustering’s on weights in kilograms and heights in centimetres? If your answers are "no" and "yes" respectively then you should probably scale.

On the other hand, if you were clustering Canadian cities based on distances east/west and distances north/south then, although there will typically be much bigger differences east/west, you may be happy just to use unscaled distances in either kilometres or miles (though you might want to adjust degrees of longitude and latitude for the curvature of the earth).

**e) Explain the different linkages used in Hierarchical Clustering.**

The process of Hierarchical Clustering involves either clustering sub-clusters(data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster into smaller sub-clusters in a top-down manner. During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed. The different types of linkages describe the different approaches to measure the distance between two sub-clusters of data points. The different types of linkages are:-

1. **Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.

2. **Complete Linkage:** For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.