Assignment Part -II

Q1. 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.

Ans: Here given data consist of 167 distinct countries with their child mortality rate, income, gdp, life expectancy, inflation, imports, exports, health and total fertility columns are available. Here after as there are no null values in the dataset then need to find the outliers from the dataset. Removing outliers using lower and upper quartiles are in result removing 38 countries from the dataset. So continued the analysis without removing those outliers. Here income is in thousands and GDP also in thousands but rest of the columns are in hundreds. First GDP is not in the percentage, first need to convert it in to the percentage per 1000 value count. Then make sure we need to standardise the dataset to bring all columns are in the same scale. Once the scaling is done then implemented principal components on the data and by using scree plot identified that 4.2 components are able to explain 95% of variance from the dataset. So, taking 4.2 components is not possible instead take 5 components which will be explain more than 95% of the variance from the dataset.

Now we need to implement clustering on the newly formed principal component dataset. To implement clustering we need to do sanity check with the help of Hopkins test whether the data is suitable for clustering or not. According to Hopkins test if the value is greater than 0.5 then the data is suitable for the clustering else not suitable for the clustering. Here we got the value of 0.68, so it is suitable for the clustering. Implemented K-Means clustering and draw the elbow and silhouette plots. From the elbow plot and silhouette could see that till the number of 4 there is a variation in the graph but later it becomes constant. So intuitively taking K=4 is better from the graph. Implemented K-Means clustering with 4 clusters and identified the labels and appended them to the main dataset. Now the dataset consists of clusters as well. Profiling the clusters with their performance using the various factors like child mortality, life expectancy, imports, exports, income and GDP. Not only K- Means clustering, implemented Agglomerative clustering of single linkage and complete linkage also but out of all K- Means clustering is giving the good results.

Q2.

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

Ans:

K -Means	**Hierarchical Clustering**
No. of clusters is required before implementation.	No need of K.
Randomly initialize the centroid	There is no initialization here
Forms the cluster based on the Euclidean distance	Every data point forms cluster individual cluster and nearest datapoints will get combine to form single cluster based on less Euclidean distance

Repeat the iteration until there is no change in the centroid	Repeat the process until no point is left.
Works fast on large datasets	Suggestable for small datasets only. Else it is computationally very expensive.
Cluster selection is not accurate here	It gives us the better clusters

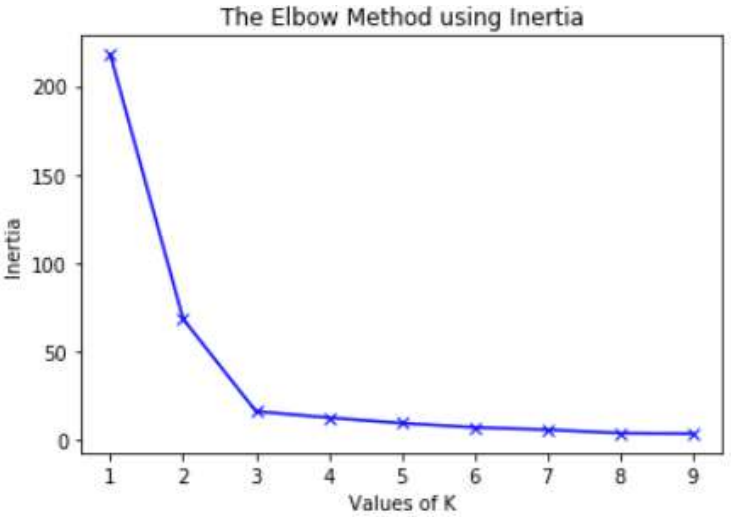
b) Briefly explain the steps of the K-means clustering algorithm.

Ans: Data should be in scaled before implementing K- Means algorithm since K- Means is works based on Euclidean distance. We need to mention the number of clusters value K to the algorithm before implementing it. We can select K by using statistical techniques of silhouette score or elbow graph to identify the number of clusters. Once the cluster number is finalized then we need to pass k value to the algorithm to implement k-means. Once it is implemented then we need to get the labels of those clusters and append them to the original dataset to profile the original data based on clusters. Analyse the clusters and interpret them to find the insights out of it.

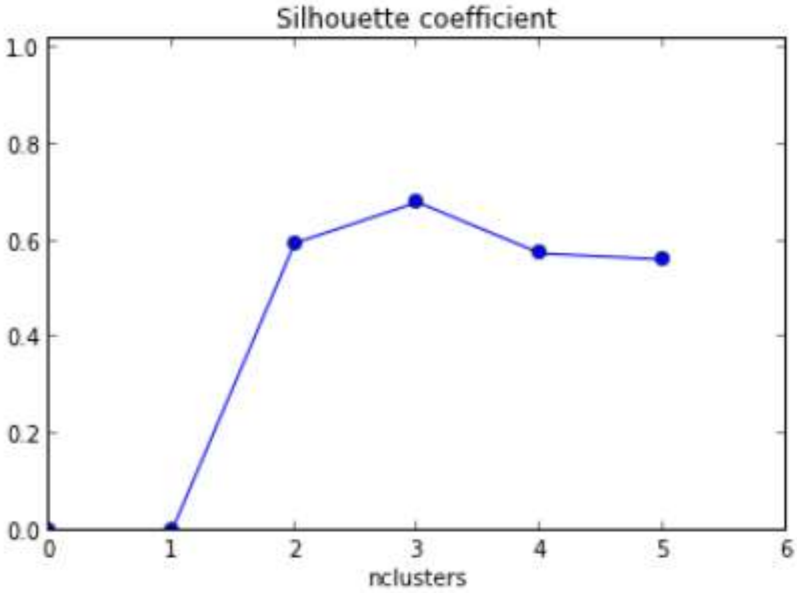
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Ans: The value of K can be chosen based on the elbow plot or silhouette plot.

Elbow Method: From the below graph, the magnitude variation is high from 1->2 and it decreases from 2->3 but still it has quite good magnitude. But from 3 onwards the length of the magnitude is same it means there is no variation in that. So in this case we can fix the K value is 3.



Silhouette Method: In Silhouette method we have to identify the peak from the graph. From the below plot 1->2 there is a peak and from 2->3 also there is a peak and it is high also. But from 3->4 it is decreasing and 4->5 also it is decreasing. So we can say that the highest peak from the graphs is at 3. So number of clusters K =3 by using the silhouette method.



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

Ans: Clustering is a distance based algorithm and it uses Euclidean distance as a measure to calculate the distance between the points. Usually clustering algorithm calculates the distance between two points and club them under one cluster based on the lowest euclidean distance. If there is no standardisation in the data then datapoints are in different scales and while calculating the distance between those points will be very high, so the algorithm thought that the point is very far and doesn't belongs to the nearby cluster even though it is near to it. So to overcome this problem it is always suggestable to standardise the data for better formation of clusters.

e) Explain the different linkages used in Hierarchical Clustering.

Ans: There are 3 types of linkages in the Hierarchical Clustering

i. Single Linkage
ii. Complete Linkage
iii. Average Linkage

Single Linkage: The distance between two clusters is defined as the shortest distance between the points in the two clusters.

Complete Linkage: The distance between any two clusters is defined as the maximum distance between any two points in the two clusters.

Average Linkage: The distance between two clusters is defined as the average distance between every point of one cluster to every other point of another cluster.

Q3. Give at least three applications of using PCA.

Ans: PCA is extensively used in Image Segmentation, Recommendation systems, banking industry where customers to attract with loan offers, etc.. where there is a need to keep the information without dropping the variables. So pca helps by clubbing the correlated variables under one component by this way it reduces the dimension.

b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

Ans: Basis is a unit which is used to express the vectors of the matrix. Vectors in any dimensional space or matrix can be represented as a linear combination of basis vector. So basis transformation is nothing but transforming the value of one vector in to linear combination of another vector without loosing its information is called Basis transformation. By this way it helps in dimensionality reduction.

Variance as information refers to the column which as more variance ultimately it possesses high variance. Since the more the variance the more the information it preserves.

c) State at least three shortcomings of using Principal Component Analysis.

Ans: 1. PCA is mainly suitable for linear combination of data and not suitable for non linear transformation methods.

2. PCA always considers the components having high variance as a priority and results in neglecting the low variance components.

3. PCA requires data to be highly correlated for it to create reasonable results and it creates components orthogonal to each other which means they are un correlated but sometimes correlated variables also give better results in few cases.