PCA AND CLUSTERING ASSINGMENT

Question 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(why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

Proccess of work flow

Problem statement – THE project is about a NGO name HELP INTERNATIONAL . it is a NGO which help countries to fight with the socio economic development.after a recent funding of the NGO they want to donate money to those countries which is on dire need of aid

We have got a data of 167 countries for the analysis and also some socio economic features values associated with them.we prepare the data for analysis

ANALYSIS PROCESS-

We have done principal component analysis on the first phase of process , and choose 4 principal component which describe about 90% variance in data .

After the first process we check silhouette analysis and sum of square distance method for k-means clustering and chose 5 cluster and categories the country on the basis of that

After that we do hierarchical clustering method and find that 5 cluster is best defining the data. So we categories the countries on basis of that

For final list of countries we choose 4 features i.e gdpp, child_mort,income , inflation for the county which describe country worst condition and cluster 4 and cluster 2 countries list from k-means and cluster 4 countries from hierarchial and merge into a dataframe and got the final list of countries.

PCA AND CLUSTERING ASSINGMENT

Question 2

State at least three shortcomings of using Principal Component Analysis.

1. PCA is limited to linearity,though we can use non-linear technique such as t-SNE as well

2. PCA need the components to be perpendicular . though in some case that may not be the best.

3. PCA assumes that columns with low variance are not useful which might not be true in prediction setup.

Question 3

Compare and contrast K-means Clustering and Hierarchical Clustering.

In k-means algorithm,you divide the data in first step itself.in the subsequent steps,you refined our clusters to get most optimal grouping

In hierarchical clustering the data is not partitioned into a particular cluster in a single step

PCA AND CLUSTERING ASSINGMENT

In k-means the relationship between cluster is undetermined

Hierarchical clustering repeatedly links pairs of cluster until every data object is included in the hierarchy.