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

- Ans :- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. $O(n^2)$.
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.

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

Ans :- K-Means algorithm is the process of dividing the N data points into K groups or clusters. Here the steps of the algorithm are:

- Start by choosing K random points the initial cluster centres.
- Assign each data point to their nearest cluster centre. The most common way of measuring the distance
- between the points is the Euclidean distance.
- For each cluster, compute the new cluster centre which will be the mean of all cluster members.
- Now re-assign all the data points to the different clusters by taking into account the new cluster centrers.
- Keep iterating through the step 3 & 4 until there are no further changes possible.

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

Ans :- 1.) **The choice of initial cluster centre has an impact on the final cluster composition.**

We see the impact of the initial cluster centres through the visualisation tool with a different set of initial cluster centres, we will get different clusters at the end.

2 . Choosing the number of clusters K in advance

There are a number of pointers that can help us decide the K for our K-means algorithm:-

1. Elbow method:-

• Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance,
by varying k from 1 to 10 clusters.

• For each k, calculate the total within-cluster sum of square (wss).

• Plot the curve of wss according to the number of clusters k.

• The location of a bend (knee) in the plot is generally considered as an indicator of the
appropriate number of clusters.

2. Average silhouette Method

• Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance,
by varying k from 1 to 10 clusters.

• For each k, calculate the average silhouette of observations (avg.sil).

• Plot the curve of avg. sil according to the number of clusters k.

• The location of the maximum is considered as the appropriate number of clusters.

Explain the necessity for scaling/standardisation before performing Clustering.

Ans:- Standardisation of data, that is, converting them into z-scores with mean 0 and standard deviation 1, is important for 2 reasons in K-Means algorithm:

• Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale helps in this process.
• The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.

Explain the different linkages used in Hierarchical Clustering.

Ans :- **Single Linkage**

Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters

Complete Linkage

Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

Average Linkage

Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

Assignment Summary:-

Overview:-

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. 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.

My job is:- To categorise the countries using some socio-economic and health factors that determine the overall development of the country.
We need to suggest the countries which the CEO needs to focus on the most.
The steps are broadly:
Read and understand the data:- In this step we have to read &understand the data . It is basically the interpretation which can be made out of the given data by preforming some analysis step and visualizing data through different graph .It also includes the method for handling the null values and dropping rows.
We go through the file .one major thing to realise is the conversation of export, health and import into actual number
Exploratory Data Analysis: In this step we draw different -2 graph Distplot, Heatmap, pairplot, boxplot .
Some of the finding EDA are: -
As the export increase the gdpp also increase.
Health and income are closely related.

i. Performing clustering:-This is the step which divide the data into groups which finally help in the analysis and find the countries
ii. Data for clustering should not have outlier and should be scaled.
iii. So the process done into two parts:-
Outlier treatment :-we used capping to avoid missing on the countries.
Hopkins check: Needed to authenticate the use of clustering

Clustering : -

i. K-means :- Start by choosing K random points the initial cluster centres.Assign each data point to their nearest

cluster centre. The most common way of measuring the distance between the points is the Euclidean distance
Using the Elbow and silhouette score to determine K.
Run K means with chosen K.
Visualize the cluster: gdpp VS income, gdpp vs child_mort, income vs child_mort.
Cluster Profiling using gdpp ,income, child_mort: low gdpp ,low income, high child_mort.

ii. Hierarchical clustering :Here we do not choose the number of cluster The major steps are:-
As the dataframe has already been scaled. We apply the clustering using the single and complete linkage.
Visualize the cluster: gdpp vs income, gdpp vs child_mort, income vs child_mort.
Cluster Profiling using gdpp , income, child_mort: low gdpp ,low income, high child_mort.

Finally we summarize our result and conclusion. The process of clustering easily identified the countries with the direst need.

.