TASK 3.2(PRATHAM KITAWAT)

Introduction to Clustering:-

Clustering is a technique used in machine learning to group similar data points together. It is an unsupervised learning method that does not require predefined classes or prior information.

Clustering helps to **identify patterns and relationships** in data that might be difficult to detect through other methods.

Clustering is important because it can help to simplify and summarize complex data sets, making it easier to analyze and understand.

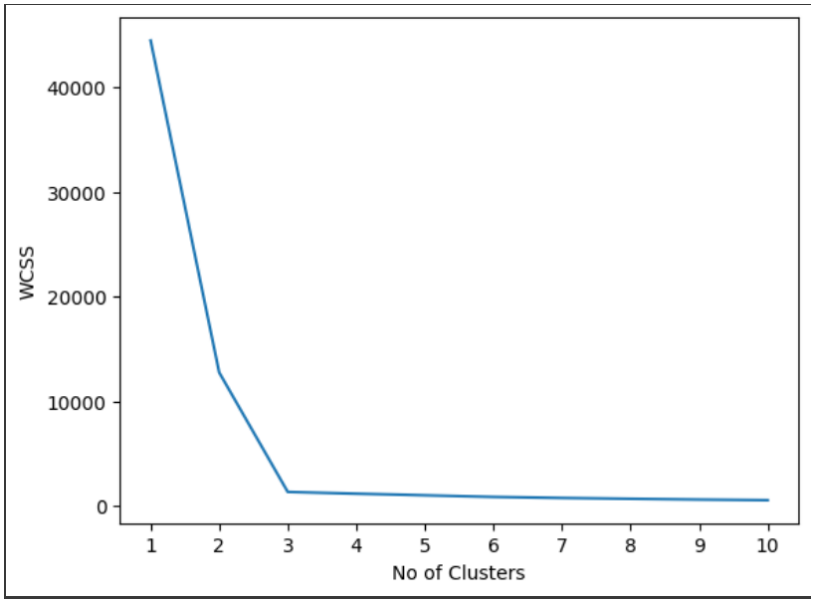
1. K-Means Clustering :-

What is it: K-means is a popular clustering algorithm that aims to partition n observations into k clusters. It's widely used due to its simplicity and efficiency in handling large datasets.

How K-means Works: 1. Choose the number of clusters you want (let's call it K).

2. Randomly place K points in your data space. These are your initial cluster centroids.
3. Assign each data point to the nearest centroids.
4. Move each centroids to the average position of all points assigned to it.
5. Repeat steps 3-4 until the centroids don't move much, or you've done it a set number of times.

We find the number of clusters we need with the help of WCSS (Within Cluster Sum Of Squares)method and then plot WCSS against the total no of clusters on a graph and the point at which the graph becomes stable is the no of clusters(Elbow method) .



For example , in the above graph K will be initialized to 3 .

Implementation: In this, we first generate a sample dataset of 1000 samples and 2 features with the 3 centres. Then, we apply K-Means Clustering on our sample dataset and set the clusters to 3. The init parameter "k-means++" is used to make sure the centroids are not spawned too close to our data points. Finally, we then plot the data points, colored according to their assigned clusters.

Advantages:

1. Simple to understand and implement
2. Efficient for large datasets

Disadvantages:

1. Requires specifying the number of clusters (k) in advance
2. Assumes spherical cluster shapes
3. Can be affected by outliers

3. Hierarchical Clustering :-

What is it: Hierarchial Clustering is a type of clustering algorithm that builds a hierarchy of clusters by recursively merging or splitting clusters based on their similarity.

The result is a tree-like structure called a dendrogram that shows the relationships between data points and clusters.

There are two main approaches to hierarchical clustering:

1. Agglomerative (bottom-up): Start with each data point as a separate cluster and merge clusters iteratively.
2. Divisive (top-down): Start with all data points in one cluster and split clusters iteratively.

We'll focus on the more common agglomerative approach.

How It Works: 1. Treat each data point as a single cluster.

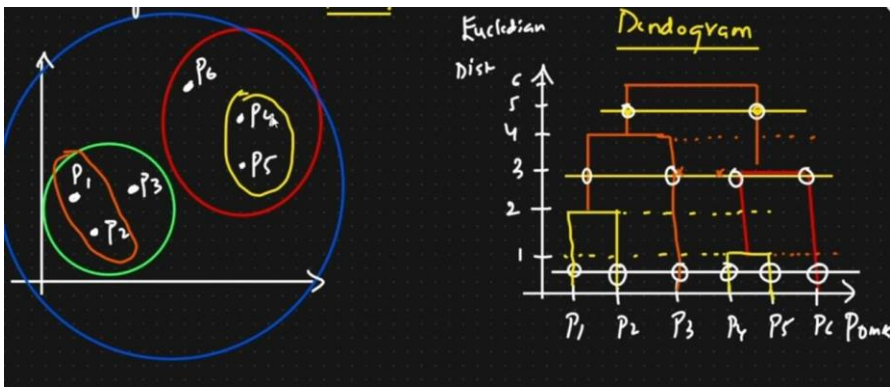
2. Calculate the distance between all pairs of clusters.

3. Merge the two closest clusters.

4. Repeat steps 2 and 3 until all points are in a single cluster.

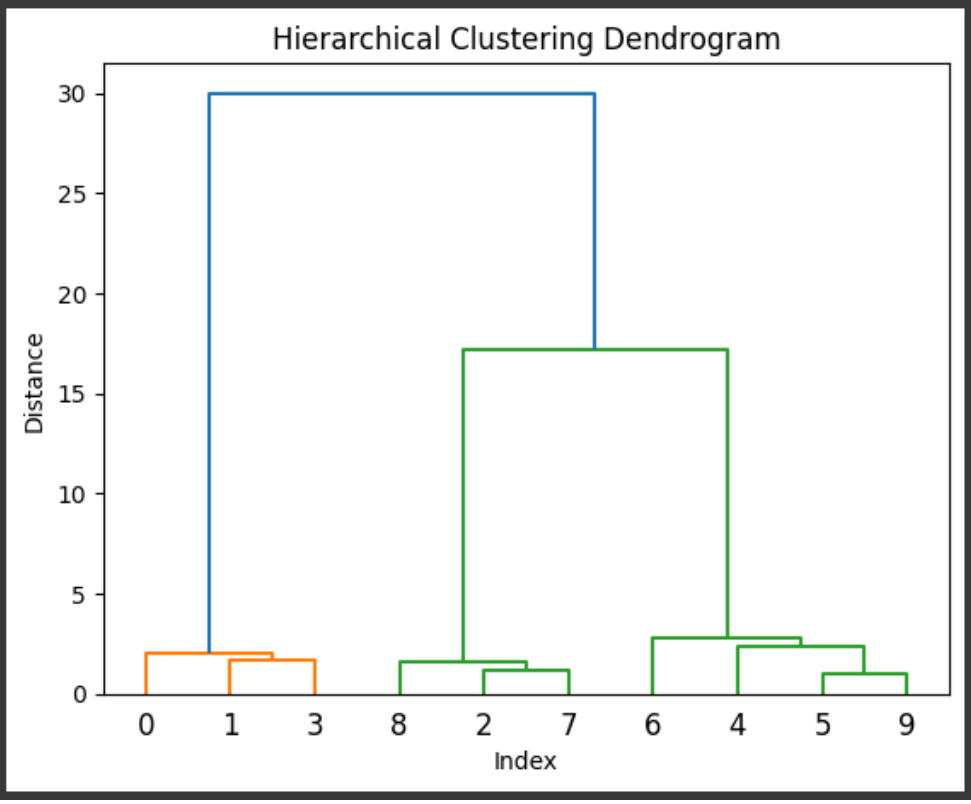
The result is a dendrogram showing the entire clustering process, allowing users to choose the number of clusters by cutting the dendrogram at different levels.

Example -



Implementation: The initial steps are same as before . This code performs hierarchical clustering using Ward's method and visualizes the result as a dendrogram. The dendrogram shows how data points are grouped into clusters.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage
from sklearn.datasets import make_blobs
x, y = make_blobs(n_samples=10, n_features=2,centers=3,random_state=42)
# Perform hierarchical clustering
linkage_matrix = linkage(x, method='ward')
# Plot the dendrogram
dendrogram(linkage_matrix)
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('Index')
plt.ylabel('Distance')
plt.show()
```



Advantages:

- You don't need to specify the number of clusters in advance.
- Provides a detailed hierarchy of how clusters form.
- Can find clusters of various shapes and sizes.

Disadvantages:

- Slower than K-means, especially for large datasets.
- Can be sensitive to outliers.

Conclusion:

Both K-means and Hierarchical Clustering are valuable tools for finding patterns in data, but they have different strengths:

- K-means is faster and works well for large datasets, especially when you have an idea of how many clusters you're looking for and expect them to be roughly round-shaped.
- Hierarchical Clustering provides a more detailed view of how data groups together at different levels. It's great when you're not sure how many clusters there are and want to explore the data's structure.

References : [Complete Unsupervised Machine Learning Tutorials In Hindi- K Means,DBSCAN, Hierarchical Clustering (youtube.com)]