Viva Questions & Answers on K-Means Clustering Practical

Q1. What is clustering?

A: Clustering is an unsupervised machine learning technique where the goal is to group similar data points into clusters based on their features.

Q2. Which clustering algorithm did you use in your practical?

A: I used the K-Means clustering algorithm.

Q3. Is K-Means a supervised or unsupervised algorithm?

A: K-Means is an unsupervised algorithm.

Q4. Explain the K-Means algorithm steps.

A: Steps are:

1. Choose the number of clusters (k).

2. Initialize k cluster centers randomly.

3. Assign each data point to the nearest cluster center.

4. Update the cluster centers as the mean of assigned points.

5. Repeat steps 3 and 4 until convergence.

Q5. How did you decide the number of clusters (k)?

A: In this practical, k was set to 4 manually. In real scenarios, methods like the Elbow Method can be used to determine the best k.

Q6. Why did you standardize the data before clustering?

A: Standardization ensures that all features contribute equally to the distance calculations, preventing bias due to different scales.

Q7. What function did you use to standardize the data?

A: I used StandardScaler() from the sklearn.preprocessing module.

Q8. Can K-Means be applied to non-numeric data?

A: No, K-Means requires numeric data because it relies on distance calculations.

Q9. How did you visualize the clusters?

A: I used a scatter plot with matplotlib for 2D data, coloring points by their assigned cluster.

Q10. What library did you use for clustering?

A: I used KMeans from the sklearn.cluster module.

Q11. What does the fit() method do in KMeans?

A: It trains the KMeans model by finding cluster centers and assigning labels to data points.

Q12. How can missing values affect clustering?

A: Missing values can distort distance calculations, so I used dropna() to remove rows with missing values.

Q13. What is the role of random_state in KMeans?

A: It sets a seed to make the clustering results reproducible.

Q14. What is inertia in KMeans?

A: Inertia is the sum of squared distances between data points and their cluster centers. Lower inertia means better clustering.

Q15. What is the main drawback of K-Means?

A: K-Means assumes spherical clusters and is sensitive to the initial placement of cluster centers.

Objective of Practical:

To perform data clustering using the K-Means algorithm and understand clustering concepts.