1a) Compare Supervised and Unsupervised Machine Learning models with example. (4 marks)

Feature	Supervised Learning	Unsupervised Learning
Definition	Learns a function from labeled data (input-output pairs).	Learns patterns or groupings from unlabeled data.
Goal	Predict output (classification/regression).	Find hidden patterns (clustering/dimensionality reduction).
Examples	Linear Regression, Decision Trees, SVM	K-Means, DBSCAN, PCA
Example Use Case	Predict house price from features.	Group customers by purchasing behavior.

1b) Explain the steps for cluster formation through K-means clustering. (4 marks)

1. **Initialization**: Randomly select *k* initial centroids.
2. **Assignment Step**: Assign each data point to the nearest centroid based on Euclidean distance.
3. **Update Step**: Recompute centroids by calculating the mean of all points assigned to each cluster.
4. **Repeat**: Iterate assignment and update steps until convergence (no change in centroids or minimal movement).
5. **Result**: Final cluster assignments and centroid positions.

1c) What is the need of linkage methods? Compare different linkage methods. (4 marks)

Need of Linkage Methods:

Linkage methods are used in hierarchical clustering to decide the distance between clusters when merging them.

Types of Linkage Methods:

Method	Description	Behavior
Single	Minimum distance between points in two clusters.	Tends to form elongated clusters (chaining effect).
Complete	Maximum distance between points in two clusters.	Tends to form compact clusters.
Average	Average distance between all points of two clusters.	Balanced approach between single and complete.
Ward	Minimizes total within-cluster vari	Produces more equally-sized clusters.

1d) The points in cluster1 at the final iteration are (2,4),(3,4),(1,3) and (2,5). Compute the cluster inertia. (4 marks)

Step 1: Compute centroid

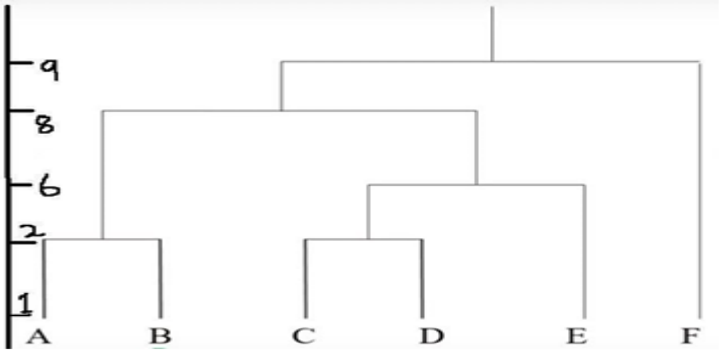
$$\text{Centroid } C = \left(\frac{2+3+1+2}{4}, \frac{4+4+3+5}{4}\right) = (2,4)$$

Step 2: Compute squared distances from each point to centroid

- (2,4): $(0)^2 + (0)^2 = 0$
- (3,4): $(1)^2 + (0)^2 = 1$
- (1,3): $(-1)^2 + (-1)^2 = 2$
- (2,5): $(0)^2 + (1)^2 = 1$

Cluster Inertia = Sum of squared distances

$$\text{Inertia} = 0 + 1 + 2 + 1 = \boxed{4}$$



1e) Compute the optimal number of clusters for the above dendrogram. What are all the observations (samples) present in each cluster? (4 marks)

4 - (A,B), (C,D), E, F

1a) List any three distance measures used for clustering with their mathematical expression. (4 marks)

1. **Euclidean Distance:**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2. **Manhattan Distance:**

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

3. **Cosine Distance:**

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|}$$

These are used to measure similarity or dissimilarity between data points in clustering algorithms.

1b) Explain the role of Silhouette score in measuring the quality of the clusters. (4 marks)

- The **Silhouette Score** measures how similar an object is to its own cluster compared to other clusters.
- It is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

  where:

  - $a(i)$: average distance to other points in the **same** cluster.
  - $b(i)$: average distance to points in the **nearest neighboring** cluster.
- The score ranges from -1 to 1:
  - **+1**: Point is well-matched to its cluster and poorly matched to others.
  - **0**: Point is on or very close to the decision boundary between two clusters.
  - **-1**: Point might be in the wrong cluster.
- **Purpose**: Helps in selecting the optimal number of clusters and validating clustering quality.

1c) Say the distance between cluster B and A is 4, and the distance between C and A is 6. What is the distance between the cluster (BC) and A using complete and single link method? (4 marks)

- **Single Link**: Minimum distance between elements of A and BC

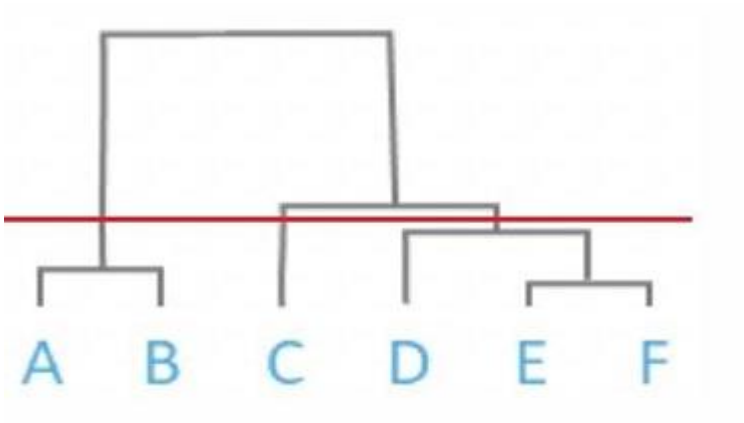
  Single Link: min(4,6)=4

- **Complete Link**: Maximum distance between elements of A and BC

  Complete Link: max(4,6)=6

1d) Compare PCA and SVD. (4 marks)

Feature	PCA (Principal Component Analysis)	SVD (Singular Value Decomposition)
Purpose	Dimensionality reduction, feature extraction	Matrix factorization, used in PCA, Latent Semantic Analysis, etc.
Decomposition	Covariance matrix is decomposed	Original matrix A is decomposed: $A = U\Sigma V^T$
Components	Eigenvalues & eigenvectors of covariance matrix	Singular values and singular vectors
Input Requirement	Requires mean-centered data	Does not require mean-centering (but often used with it)
Relation	PCA can be derived using SVD	SVD is more general and not limited to symmetric matrices



3 clusters – (A,B) , C, (D,E,F)

1c) Write any two characteristics of Principal Components derived from the data. (4 marks)

1. **Orthogonality**:
   - Principal Components (PCs) are **orthogonal** (uncorrelated) to each other, ensuring no redundancy in captured variance.
2. **Variance Maximization**:
   - The first PC captures the **maximum variance** in the data, with each succeeding PC capturing the next highest variance under orthogonality constraints.

- **What types of problems can be solved using unsupervised learning?**

  - Clustering (e.g., customer segmentation)
  - Dimensionality reduction (e.g., data visualization)
  - Anomaly detection (e.g., fraud detection)

- **What is clustering in unsupervised learning?**

Clustering is the process of grouping similar data points together based on similarity or distance.

- **What is the role of dimensionality reduction in unsupervised learning?**

It reduces the number of input variables to simplify models, remove noise, and help in visualization.

- **What is the elbow method in K-Means clustering?**

A technique to choose the optimal number of clusters by plotting inertia vs. number of clusters and identifying the "elbow" point.

- **What is hierarchical clustering? Name its types.**

It builds a hierarchy of clusters either **agglomeratively (bottom-up)** or **divisively (top-down)**.

- **Compare single, complete, and average linkage in hierarchical clustering.**

  - **Single Linkage**: Minimum distance between points.
  - **Complete Linkage**: Maximum distance between points.
  - **Average Linkage**: Average distance between all points.

- **What is DBSCAN and how is it different from K-Means?**

DBSCAN groups points based on **density** and can detect **arbitrary-shaped clusters** and noise, unlike K-Means which assumes spherical clusters.

- **What is the silhouette score and how is it interpreted?**

It measures how well a data point fits within its cluster vs. other clusters. Ranges from -1 to +1, where a higher value is better.

- **How do you choose the number of clusters in clustering algorithms?**

Use methods like the **elbow method**, **silhouette score**, or **gap statistic**.

- **Why is cluster evaluation more difficult in unsupervised learning than in supervised learning?**

Because there are no ground truth labels to compare with, making it hard to directly measure accuracy.

- **What is PCA and what is its objective?**

PCA transforms data into new axes (principal components) to maximize variance and reduce dimensionality.

- **What are principal components and how are they selected in PCA?**

Principal components are the **eigenvectors** of the covariance matrix. They're selected based on the **highest eigenvalues** (variance explained).

- **Compare PCA and SVD.**

PCA uses the **covariance matrix** to extract components, while SVD directly decomposes the data matrix. PCA can be derived from SVD.

- **Give two real-world applications of clustering.**

  - Customer segmentation in marketing
  - Document/topic grouping in NLP

- **How can unsupervised learning be used in anomaly detection?**

It can identify outliers that do not fit the patterns of the majority of the data (e.g., fraud detection).

- **Why is the 'label' column usually excluded in clustering tasks?**

Because unsupervised learning does not require or use labeled data — including labels biases the results.

- **How can unsupervised learning help in customer segmentation?**

It can group customers based on purchasing behavior or demographics, enabling targeted marketing strategies.

- **What is the curse of dimensionality and how does PCA help overcome it?**

As the number of features increases, data becomes sparse, reducing model performance. PCA reduces dimensions, preserving variance.

- **What is the role of distance metrics in clustering?**

They determine how similarity is measured, impacting cluster formation (e.g., Euclidean, cosine, Manhattan).

- **When should you use hierarchical clustering over K-Means?**

When you want to understand data structure at multiple levels or when the number of clusters is unknown.