**Unsupervised Learning**

## **What is Unsupervised Learning?**

Unsupervised learning is a type of machine learning where the model is **not given labeled data.**  
 This means:

* There are no target/output values (like “spam” or “not spam”).
* The algorithm tries to find hidden patterns, structures, or relationships in the data on its own.

Simply: The model learns from raw data without being told what the correct output should be.

**Why Use Unsupervised Learning?**

It’s useful when:

* You don’t have labelled data (labelling is expensive or impossible).
* You want to discover natural groupings or reduce data complexity.

**Key Goals**

* **Clustering:** Group similar data points together.
* **Dimensionality Reduction:** Simplify data by reducing the number of features while keeping important information (like PCA).
* **Association:** Find rules or relationships between variables (like market basket analysis).

**Types of Unsupervised Learning**

1. **Clustering** → Finding groups in data (K-Means, Hierarchical, DBSCAN)
2. **Dimensionality Reduction** → Summarizing data (PCA, t-SNE)
3. **Association Rule Learning** → Finding rules (Apriori, FP-Growth)

**Clustering Introduction**

**What is Clustering?**

Clustering is the process of **grouping data points** so that:

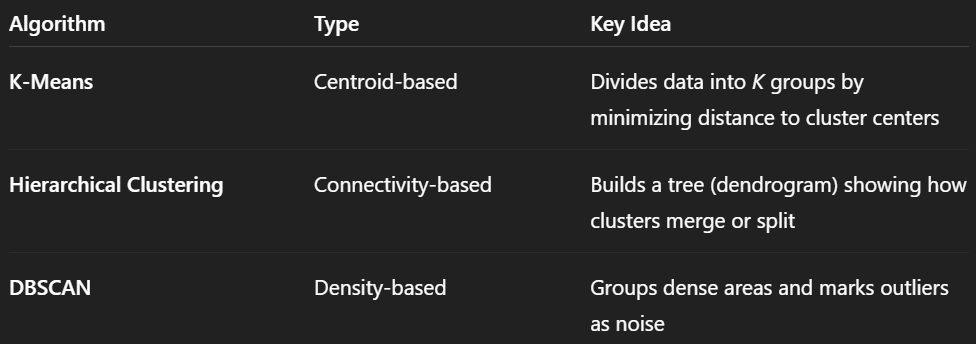
* Points **in the same cluster** are **more** **similar** to each other.
* Points **in different clusters** are **more dissimilar**.

Example: In a dataset of customers, clustering can automatically group customers with similar purchasing habits.

**Goal of Clustering**

To find hidden structure or natural groupings in data without any labels.

**Common Clustering Algorithms**



**How Clustering Works (General Idea)**

1. Choose the number of clusters (sometimes automatically found).
2. Assign data points to clusters based on similarity (e.g., Euclidean distance).
3. Update clusters until they stabilize (no big changes).
4. Evaluate results (using metrics like silhouette score or visual plots).

**K-Mean Algorithm**

**Introduction**

K-Means is one of the most popular unsupervised learning algorithms used for clustering — i.e., grouping similar data points together without using labels.

The goal is simple:

Divide the dataset into K clusters, such that points within the same cluster are more like each other than to those in other clusters.

### **Working of K-Means**

K-Means works in an **iterative** way and follows these main steps:

### **Step 1: Choose K**

Decide how many clusters (K) you want to divide your data into.

### **Step 2: Initialize Centroids**

Randomly pick K points from the dataset as the **initial centroids** (these act like the centres of clusters).

### **Step 3: Assign Points to Nearest Centroid**

For each data point, calculate its distance (usually **Euclidean distance**) to each centroid, and assign it to the nearest one.

### **Step 4: Update Centroids**

For each cluster, compute the **mean** of all points assigned to it.  
 This mean becomes the new centroid.

**Step 5: Repeat**

Repeat steps 3 and 4 until:

* The centroids stop changing significantly, or
* You reach a maximum number of iterations.

### **Example (Simple Intuition)**

Let’s say we have 10 data points representing customers’ annual income.

We choose K = 2 (we want 2 groups).

1. Randomly select 2 points as initial centroids.
2. Assign each customer to the nearest centroid → Group A or Group B.
3. Compute the average income of each group → update centroids.
4. Reassign customers using the new centroids.
5. Repeat until centroids don’t move much.

Result: two clusters, maybe low-income customers and high-income customers.

### **Key Points**

* K must be chosen manually (using the **Elbow Method** helps).
* Works well when clusters are spherical and similar in size.
* Sensitive to initialization and outliers.
* Uses distance-based learning, so scaling features (normalization) is important.