



A T M E
College of Engineering



MODULE 4

Unsupervised Learning



A T M E
College of Engineering



Characteristics of Unsupervised Learning

- Works with **unlabeled datasets**
- No predefined output classes
- Automatically identifies patterns
- Discovers hidden relationships
- Useful for exploratory data analysis
- Suitable for large datasets



A T M E
College of Engineering



1. Introduction

Machine Learning enables computers to learn patterns from data without being explicitly programmed. Based on the availability of labeled data, machine learning is broadly classified into:

1. Supervised Learning
2. Unsupervised Learning
3. Semi-Supervised Learning
4. Reinforcement Learning

In **Unsupervised Learning**, the training data does **not contain predefined output labels**. The objective is to discover hidden structures, relationships, or patterns in the data.

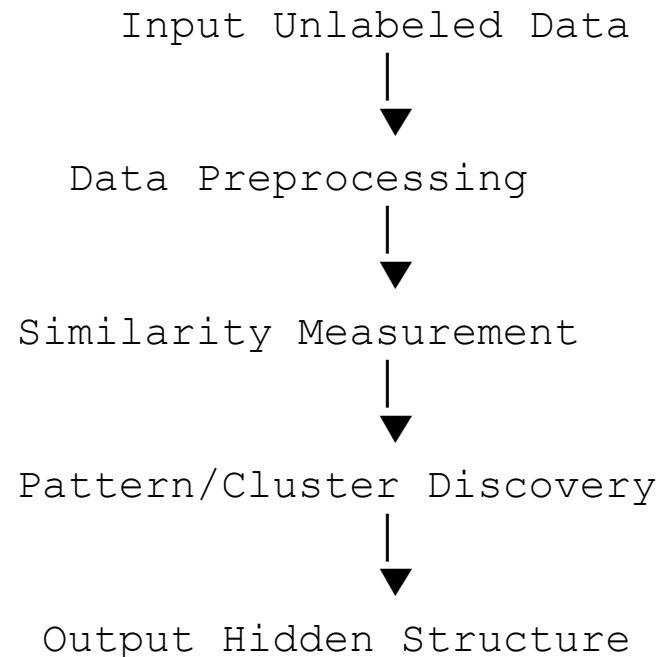
Unlike supervised learning, the algorithm is not told what the correct output is. Instead, it independently analyzes the input data and groups similar data points or discovers meaningful relationships.



Working Principle

The algorithm examines similarities and differences among data points and groups them based on common characteristics.

Flowchart





A T M E
College of Engineering



Types of Unsupervised Learning

Unsupervised learning is broadly classified into:

1. **Clustering**
2. **Association Rule Learning**
3. **Dimensionality Reduction**



A T M E
College of Engineering



Clustering

- Clustering is the process of grouping similar data points into clusters such that:
 - Data points within the same cluster are highly similar.
 - Data points belonging to different clusters are significantly different.



A T M E
College of Engineering



Example

Suppose customer data contains:

- Age
- Income
- Spending Score

The algorithm may automatically identify:

- High-income customers
- Budget-conscious customers
- Frequent shoppers

without being told these categories in advance.



A T M E
College of Engineering



Types of Clustering

1. Partition-Based Clustering

Example:

- K-Means
- K-Medoids

2. Hierarchical Clustering

Creates a hierarchy of clusters.

Two approaches:

- Agglomerative (Bottom-Up)
- Divisive (Top-Down)



A T M E
College of Engineering



3. Density-Based Clustering

Groups densely connected points.

Example:

- DBSCAN

4. Model-Based Clustering

Assumes data follows a probability distribution.

Example:

- Gaussian Mixture Models (GMM)



A T M E
College of Engineering



K-Means Clustering

Definition

K-Means is one of the simplest and most widely used clustering algorithms.

It partitions the dataset into **K clusters**, where each cluster is represented by its centroid.

Algorithm

1. Select the number of clusters KKK.
2. Initialize KKK centroids randomly.
3. Assign each data point to the nearest centroid.
4. Recalculate the centroid of each cluster.
5. Repeat Steps 3–4 until centroids no longer change or a maximum number of iterations is reached.



A T M E
College of Engineering



Advantages

- Simple to implement
- Fast for large datasets
- Easy to interpret

Limitations

- Requires predefined KKK
- Sensitive to outliers
- May converge to a local optimum



A T M E
College of Engineering



Hierarchical Clustering

Hierarchical clustering builds a tree-like structure called a **dendrogram**.

Types

Agglomerative (Bottom-Up)

- Start with each data point as a separate cluster.
- Merge the closest clusters iteratively.

Divisive (Top-Down)

- Start with one large cluster.
- Split it recursively into smaller clusters.



A T M E
College of Engineering



Advantages

- No need to specify KKK
- Produces a hierarchy of clusters

Limitations

- Computationally expensive for large datasets



A T M E
College of Engineering



DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN groups points that are closely packed together and identifies outliers as noise.

Key Parameters

- **Epsilon** (ϵ): Neighborhood radius.
- **MinPts**: Minimum number of points required to form a dense region.



A T M E
College of Engineering



Advantages

- Detects clusters of arbitrary shape.
- Handles noise effectively.
- No need to specify the number of clusters.

Limitations

- Sensitive to parameter selection.
- Less effective when cluster densities vary significantly.



A T M E
College of Engineering



Principal Component Analysis (PCA)

PCA transforms correlated variables into a smaller number of uncorrelated variables called **principal components**.

Steps

1. Standardize the data.
2. Compute the covariance matrix.
3. Calculate eigenvalues and eigenvectors.
4. Select the principal components with the largest eigenvalues.
5. Transform the original data.



A T M E
College of Engineering



Advantages

- Reduces dimensionality
- Removes redundancy
- Speeds up learning algorithms

Limitations

- Reduced interpretability
- Assumes linear relationships



A T M E
College of Engineering



Applications of Unsupervised Learning

- Customer segmentation
- Recommendation systems
- Image segmentation
- Social network analysis
- Fraud detection
- Medical diagnosis
- Document clustering
- Topic modeling
- Gene sequence analysis
- Anomaly detection



A T M E
College of Engineering



Advantages

- Does not require labeled data
- Identifies hidden structures
- Useful for exploratory analysis
- Discovers unknown patterns
- Can process large datasets



A T M E
College of Engineering



Limitations

- Difficult to evaluate results
- Sensitive to algorithm parameters
- Interpretation of clusters may be subjective
- Some algorithms are computationally expensive



A T M E
College of Engineering



Sampling

Introduction

Machine learning algorithms are trained using data. In real-world applications, datasets may contain **millions of records**, making it computationally expensive and time-consuming to process the entire dataset.

Instead of using the complete dataset, a **representative subset**, called a **sample**, is selected. The process of selecting this subset is known as **sampling**.

A well-chosen sample preserves the characteristics of the original dataset, enabling machine learning models to learn effectively while reducing computational cost.



A T M E
College of Engineering



Why is Sampling Important?

Sampling is essential because:

- Large datasets require high computational resources.
- It reduces memory usage.
- It decreases training time.
- It allows faster experimentation.
- It improves efficiency in model development.
- It is useful when collecting the entire dataset is impractical.



A T M E
College of Engineering



Characteristics of a Good Sample

A good sample should:

- Represent the population accurately.
- Be free from bias.
- Include sufficient observations.
- Capture the variability of the population.



A T M E
College of Engineering



Types of Sampling

Sampling techniques are broadly classified into:

A. Probability Sampling

Each observation has a known probability of being selected.

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling
4. Cluster Sampling

B. Non-Probability Sampling

The probability of selection is unknown.

1. Convenience Sampling
2. Judgment Sampling
3. Snowball Sampling
4. Quota Sampling



A T M E
College of Engineering



Simple Random Sampling

Every observation has an equal probability of being selected.

Procedure

1. Assign a unique number to each observation.
2. Randomly select the required number of observations.
3. Use the selected observations as the sample.

Example

Suppose a dataset contains **1,000 customer records**.

To select a sample of **100 customers**, every customer has an equal chance of being selected.



A T M E
College of Engineering



Advantages

- Easy to implement
- Reduces selection bias
- Produces representative samples (when the population is homogeneous)

Limitations

- May not represent minority classes well
- Requires access to the complete dataset



Systematic Sampling

Systematic sampling selects every **k-th observation** after a random starting point.

Formula

$$k = \frac{N}{n}$$

where:

- N = Population size
- n = Sample size

Example

Population = 1,000

Sample size = 100

$$k = \frac{1000}{100} = 10$$

If the random start is 4, the selected observations are:

4, 14, 24, 34, ...



A T M E
College of Engineering



Advantages

- Simple
- Faster than random sampling

Limitations

- Can introduce bias if the data has periodic patterns



A T M E
College of Engineering



Applications of Sampling

- Big data analytics
- Medical diagnosis
- Financial forecasting
- Image processing
- Customer behavior analysis
- Survey analysis
- Recommendation systems
- Fraud detection
- Quality control



A T M E
College of Engineering



Advantages of Sampling

- Reduces computational cost
- Saves memory
- Faster model training
- Enables efficient experimentation
- Makes large datasets manageable
- Reduces data collection costs



A T M E
College of Engineering



Limitations of Sampling

- Poor sampling may introduce bias.
- Small samples may not represent the population.
- Important information may be omitted.
- Incorrect sampling affects model accuracy.