Data Mining Modules – Detailed Explanation

Module 1 [9L] – Introduction to Data Mining, Preprocessing, and Rule-based Classification

1. Basics of Data Mining

Data Mining is the process of discovering useful patterns, trends, or knowledge from large sets of data. It involves methods from statistics, machine learning, and database systems.

2. Why Do We Need Data Mining?

To extract hidden patterns and relationships in big data. Helps in decision-making, market analysis, fraud detection, medical diagnosis, etc.

3. Data Mining Architecture

Data Sources → Data Warehouse → Data Mining Engine → Pattern Evaluation Module → Knowledge Base → User Interface.

4. Goals of Data Mining

Prediction and Description.

5. Techniques of Data Mining

Classification, Clustering, Regression, Association Rule Mining, Outlier Detection.

6. Challenges in Data Mining

Handling noisy/incomplete data, scalability, high dimensionality, real-time processing.

7. Data Pre-processing

Data Cleaning: Fix missing, inconsistent or noisy data.
Data Transformation: Normalize, scale, or aggregate data.
Data Reduction: Reduce data size while preserving information.

8. Applications of Data Mining

Retail, Banking, Healthcare, Web.

9. Rule-Based Classification

Uses IF-THEN rules to classify data.

Rule-ordering: Ordered and Unordered.
Building Classifier: Direct and Indirect methods.

Module 2 [9L] – Data Mining Algorithms (Supervised Learning)

1. Bayesian Network

Uses Bayes' Theorem. Naïve Bayes and Gaussian Naive Bayes classifiers.

2. K-Nearest Neighbor (K-NN)

Instance-based classifier based on majority class of nearest neighbors.

3. Decision Trees

Splits data based on features.
Gini Index and Information Gain.

4. Support Vector Machines (SVM)

Finds best decision boundary.
Linear (separable/non-separable) and Non-linear SVM using kernels.

Module 3 [9L] – Ensemble Methods & Association Rule Mining

1. Ensemble Methods

Bagging, Boosting, Random Forests.

2. Association Rule Mining

Finds relationships between variables.
Support, Confidence.
Apriori and FP-Growth Algorithms.
Correlation Analysis and Subgraph Mining.

Module 4 [9L] – Cluster Analysis

1. Introduction

Clustering groups similar data points. Applications: marketing, social networks, etc.

2. Types of Clustering

Partitional, Hierarchical, Density-based.

3. Partitional Clustering

K-Means and K-Means++.

4. Hierarchical Clustering

Agglomerative (Bottom-Up), Divisive (Top-Down).
MIN, MAX linkage, Dendrograms.

5. Density-Based Clustering

DBSCAN detects clusters based on density and identifies outliers.

6. Cluster Evaluation

Internal and External evaluation methods.

7. Further Reading Algorithms

OPTICS, DENCLUE, CHAMELEON, BIRCH, CURE, ROCK.