DMBI Overview

Here's a **summarized overview of Module 4: Classification** based on the content in your PDF:

---

1. Basic Concepts

- Classification is a **supervised learning** method that assigns a **discrete label** to input data.
- It involves two main phases:
  - **Model construction** (training with labeled data)
  - **Model usage** (predicting new/unseen data)
- Example: Identifying an email as spam or not spam.

---

2. Classification Methods

- **Eager Learners**: Build model in advance (e.g., Decision Tree, SVM, Logistic Regression)
- **Lazy Learners**: Memorize training data and classify by similarity (e.g., K-NN)

---

3. Decision Tree Induction

- Builds a tree where:
  - **Internal nodes** test features.
  - **Leaf nodes** give class labels.
- **Best Split Selection**:
  - Use impurity measures like **Information Gain**, **Gini Index**
  - Prefer splits that make classes purer (lower entropy).

- **Tree Pruning**:
  - Removes overfitting by eliminating nodes that add noise or outliers.
- **Attribute Selection Measures**:
  - Choose attribute with highest info gain or lowest impurity for the root.

---

4. Bayesian Classification

- Based on **Bayes' Theorem**.
- **Naïve Bayes** assumes features are independent.
- Computes posterior probabilities for each class and picks the highest.
- Example: Predicting if a person buys a computer based on age, income, etc.

---

5. Rule-Based Classification

- Uses **IF-THEN rules** derived from data:
  - Example: `IF age=young AND income=high THEN buys_computer=yes`
- Rules can be derived from decision trees or association rule mining.

---

6. Accuracy and Error Measures

- **Accuracy** = (Correct Predictions) / (Total Predictions)
- **Error Rate** = 1 – Accuracy
- **Precision** = TP / (TP + FP) → Focus on positive predictions
- **Recall** = TP / (TP + FN) → Focus on actual positives
- **F1 Score** = Harmonic mean of precision and recall

---

7. Model Evaluation Techniques

- **Holdout**: Split into train/test sets once
- **Random Sampling**: Repeat holdout with random splits
- **Cross Validation (k-fold)**:
  - Split data into k parts, train on k-1, test on 1
  - Rotate and average performance

Let me know if you want **Stepwise DSA Format** for decision trees or Bayes classifier, or if you want the **Naïve Bayes** calculations broken down.

Here's a **summarized overview of Module 5: Clustering** based on all your uploaded PDFs:

---

1. Cluster Analysis – Basic Concepts

- **Clustering** is an **unsupervised learning** technique used to group similar data points.
- It does **not use class labels**, unlike classification.
- Purpose: **maximize intra-cluster similarity** and **minimize inter-cluster similarity**.
- Applications: Market segmentation, document classification, image processing.

---

2. Partitioning Methods

a. K-Means Clustering

- Select **k** centroids randomly.
- Assign each point to the **nearest centroid**.
- Update centroids as the **mean of the cluster**.
- Repeat until centroids don't change.

- Works well on **large datasets**, but sensitive to outliers.

b. K-Medoids Clustering

- Similar to K-means, but uses **actual data points (medoids)** as cluster centers.
- More robust to **outliers and noise**.
- Cost is calculated as the **sum of distances to medoid**.
- PAM (Partitioning Around Medoids) is a common algorithm.

---

3. Hierarchical Methods

a. Agglomerative (Bottom-up)

- Start with **each data point as a separate cluster**.
- Merge closest clusters iteratively using distance measures like:
  - Single Link (min distance)
  - Complete Link (max distance)
  - Average Link

b. Divisive (Top-down)

- Start with **one large cluster**.
- Recursively split clusters until each data point is isolated or a stopping condition is met.
- **Dendrogram**: Tree-like diagram used to visualize clustering.

---

4. Density-Based Clustering: DBSCAN

- Groups data based on **region density**.
- Requires two parameters:
  - **Eps** (neighborhood radius)
  - **MinPts** (minimum points to form dense region)
- Can find **arbitrary-shaped clusters** and detect **noise**.

- Classifies points as:
  - **Core**: has enough points in its neighborhood
  - **Border**: near core points
  - **Noise**: doesn't belong to any cluster

---

5. Evaluation of Clustering

- No true labels → need **internal and external** evaluation measures:
  - **Silhouette coefficient**
  - **Dunn index**
  - **Rand index** (if ground truth is known)
- Evaluate **compactness** (intra-cluster similarity) and **separation** (inter-cluster dissimilarity).

---

6. Outliers

- Data points that **deviate significantly** from others in the dataset.

Types:

- **Global**: lies far from all clusters
- **Contextual**: unusual in a specific context (e.g., time)
- **Collective**: group of data points behaving anomalously together

Challenges:

- Defining a threshold, choosing the right method, scalability on large data.

---

7. Outlier Detection Methods

a. Supervised

- Use labeled normal and outlier data.
- Apply classification (e.g., decision tree).

b. Semi-Supervised

- Train on **only normal data** to detect deviations.

c. Unsupervised

- Detect outliers without labels by assuming outliers are **few and different**.
- Common in real-world applications.

d. Statistical Methods

- Assume data follows a known distribution (e.g., Gaussian).
- Points far from the mean are flagged.

e. Proximity-Based

- Use **distance or density**:
  - Outliers have **few neighbors** or lie **far away**.
  - Examples: k-NN-based outlier score

f. Clustering-Based

- Points **not assigned to any cluster** or belonging to **small, distant clusters** are treated as outliers.

Let me know if you want worked examples, derivations, or stepwise format for any clustering algorithm!

Here's a **summarized overview of Module 6: Business Intelligence** based on your uploaded PDF:

---

1. What is Business Intelligence (BI)?

- **BI** is a set of **analytical methods and models** that transform raw data into **useful information and knowledge** for **decision-making**.
- Helps in making **better, faster, and more effective decisions** by revealing patterns, trends, and insights.

---

2. BI Architecture

- **Three Major Components**:
  - **Data Sources**: Internal & external, structured & unstructured data.
  - **Data Warehouses & Data Marts**:
    - Use **ETL (Extract, Transform, Load)** processes to consolidate data.
  - **BI Methodologies**:
    - Apply **mathematical models**, **data mining**, and **analytics** for decision support.
- **Pyramid Architecture** (bottom to top):
  1. **Raw data**
  2. **Processed data** (warehouses/marts)
  3. **Exploration & reporting tools**
  4. **Data mining & analysis models**
  5. **Optimization models**
  6. **Final decisions**

---

3. Definition of Decision Support System (DSS)

- A **DSS** is an **interactive, computer-based system** that combines **data**, **models**, and **interfaces** to support decision-making in complex scenarios.
- Transforms data into actionable **insights** via analysis and simulation tools.

---

4. Development of a BI System

Phases:

- **Analysis**:
  - Identify organizational needs, objectives, and user requirements via interviews.
- **Design**:
  - Draft system architecture based on decision-making processes.
- **Planning**:
  - Detail out system functions, development steps, timelines, and costs.
- **Implementation & Control**:
  - Build data warehouse, marts, and BI tools.
  - Evaluate performance and adjust as needed.

---

5. Data Retrieval for Business Applications

a. Fraud Detection

- Analyze **patterns in transactions** to flag anomalies.
- BI tools detect **unusual activities** in banking or e-commerce.

b. Clickstream Mining

- Track and analyze user navigation behavior on websites.
- Helps optimize **user experience** and **advertisement placement**.

c. Market Segmentation

- Group customers based on behaviors, preferences, and demographics.
- Enables **targeted marketing** and **product customization**.

d. Retail Industry

- Analyze sales, customer behavior, inventory patterns.
- BI helps in **stock management**, **trend forecasting**, and **personalized promotions**.

e. Telecommunication Industry

- Manage and analyze **call records**, **network usage**, and **customer churn**.
- BI improves **service delivery**, **fraud detection**, and **customer retention**.

f. Banking & Finance

- Risk assessment, loan defaults, investment predictions.
- BI helps in **portfolio management**, **credit scoring**, and **fraud detection**.

g. Customer Relationship Management (CRM)

- Use data to **understand customer behavior**, preferences, and loyalty patterns.
- Supports **customer acquisition, retention**, and **value maximization**.

Let me know if you want **example workflows** or **system diagrams** for BI/DSS processes!