Data Mining & Warehousing Notes

Unit 1: Data Warehousing

1. Introduction to Data Warehousing

- Definition: A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data used to support decision-making.

- Characteristics:

  1. Subject-Oriented: Focuses on business subjects like sales, marketing, etc.
  2. Integrated: Combines data from various sources.
  3. Time-Variant: Stores historical data for trends.
  4. Non-Volatile: Data is read-only and cannot be modified.

2. Data Preprocessing

- Data Cleaning: Handling missing values, removing noise.

- Data Integration: Combining data from various formats.

- Data Reduction: PCA, aggregation, sampling.

3. Data Warehouse Design

- Star Schema: Central fact table linked to dimension tables.

- Snowflake Schema: Normalized dimensions.

Unit 2: OLAP Systems

1. Basic Concepts

- OLAP (Online Analytical Processing) performs multidimensional analysis of data.

- Operations:

  1. Slice: Extract data for a specific dimension.

  2. Dice: Apply multiple filters.

  3. Roll-up: Summarize data by climbing a hierarchy.

  4. Drill-down: Break data into finer details.

2. Types of OLAP Servers

- ROLAP: Relational OLAP, uses relational databases.

- MOLAP: Multidimensional OLAP, uses pre-computed cubes.

- HOLAP: Hybrid of ROLAP and MOLAP.

Unit 3: Introduction to Data & Data Mining

1. Data Types

- Structured: Rows and columns (e.g., SQL databases).

- Unstructured: Text, images, videos.

- Semi-Structured: XML, JSON.

2. Data Mining

- Definition: Data mining is the process of discovering patterns, correlations, and trends by analyzing large datasets.

3. KDD Process

- Selection -> Preprocessing -> Transformation -> Mining -> Evaluation.

Unit 4: Supervised Learning

1. Decision Trees

- Use tree-like structures to classify data.

- Example Algorithms: ID3, C4.5, CART.

2. k-NN (k-Nearest Neighbors)

- Classifies based on the nearest k points in feature space.

3. Naive Bayes

- Probabilistic algorithm using Bayes' Theorem.

Unit 5: Clustering & Association Rule Mining

1. Clustering

- Definition: Grouping data into clusters where intra-cluster similarity is high and inter-cluster similarity is low.

- Types:

  1. Hierarchical: Divisive and agglomerative methods.
  2. Partitioning: k-Means, k-Medoids.
  3. Density-Based: DBSCAN (identifies noise and clusters of arbitrary shapes).

2. Association Rule Mining

- Apriori Algorithm: Finds frequent itemsets using support and confidence.

- FP-Growth Algorithm: Builds a tree for frequent pattern mining.