

## SEMESTER-V

### COURSE 11: DATA MINING

**Theory**

**Credits: 3**

**3 hrs/week**

---

#### **Course Objectives:**

- Understand the concepts, architecture, and applications of Data Warehousing and its role in business intelligence.
- Gain knowledge of data preprocessing techniques used in preparing data for analysis.
- Learn various data mining techniques including association rule mining, clustering, and classification using decision trees.
- Explore advanced algorithms for pattern discovery and knowledge extraction.
- Apply web mining and text mining techniques for extracting meaningful patterns from unstructured data sources.

#### **Course Outcomes:**

Upon successful completion of the course, the student will be able to:

1. Understand the architecture and components of Data Warehousing, including multidimensional data models, OLAP operations, and data preprocessing techniques for effective data storage and analysis.
2. Demonstrate knowledge of core data mining concepts, processes, and techniques, and distinguish them from traditional database systems.
3. Apply various association rule mining algorithms (such as Apriori, FP-Growth, and others) to discover hidden patterns and relationships in large datasets.
4. Implement and evaluate clustering and classification techniques (like k-Medoid, DBSCAN, Decision Trees) for data grouping and predictive modelling.
5. Analyze and apply web and text mining methods to extract meaningful insights from semi-structured and unstructured data sources for real-world applications.

#### **Unit-1: Data Warehousing:**

Introduction, What is Data Warehouse? Definition, Multidimensional Data Model, OLAP Operations, Warehouse Schema, Data Warehouse Architecture, Warehouse Server, Metadata, OLAP Engine, Data Warehouse Backend Process, Other Features Data Preprocessing, Descriptive Data Summarization, Data Cleaning, Data Integration and Transformation, Data Reduction, Data Discretization and Concept Hierarchy Generation

**Unit-2: Data Mining:**

What is Data Mining? Data Mining: Definitions, KDD vs Data Mining, DBMS vs DM, Other Related Areas, DM Techniques, Other Mining Techniques, Issues and Challenges in DM, DM Applications- Case Studies

**Association Rules:** What is an Association Rule?, Methods to Discover Association Rules, A Priori Algorithm, Partition Algorithm, Pincer-Search Algorithm, Dynamic Itemset Counting Algorithms, FP-Tree Growth Algorithm, Generalized Association Rule, Association Rules with Item Constraints

**Unit-3: Clustering Techniques:**

Clustering Paradigms, Partitioning Algorithms, k-Medoid Algorithms, CLARA, CLARANS, Hierarchical Clustering: DBSCAN, BIRCH, CURE, Categorical Clustering Algorithms: STIRR, ROCK, CACTUS

**Unit-4: Decision Trees:**

What is a Decision Tree?, Tree Construction Principle, Best Split, Splitting Indices, Splitting Criteria, Decision Tree Construction Algorithms: CART, ID3, C4.5, Approximate Methods: CLOUDS, BOAT, Pruning Techniques, Integration of Pruning and Construction, Ideal Algorithm

**Unit-5: Web Mining:**

Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, Text Mining, Unstructured Text, Episode Rule Discovery for Texts, Hierarchy of Categories, Text Clustering

**Textbooks:**

1. Data Mining Techniques, Arun K Pujari, 3<sup>rd</sup> Edition, Universities Press
2. Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei, 3<sup>rd</sup> Edition, Morgan Kaufmann Publishers

**Reference Books:**

1. K.P. Soman , Shyam Diwakar, V.Ajay ,2006, Insight into Data Mining Theory and Practice, Prentice Hall of India Pvt. Ltd - New Delhi.
2. Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, 2<sup>nd</sup> edition,

**Activities:****Outcome:**

**Understand the architecture and components of Data Warehousing, including multidimensional data models, OLAP operations, and data preprocessing techniques for effective data storage and analysis.**

## SEMESTER-V

### COURSE 11: DATA MINING

**Practical**

**Credits: 1**

**2 hrs/week**

---

#### List of Experiments:

Recommended datasets: **weather.arff**, **iris.arff**, **supermarket.arff**, **vote.arff**, **contact-lenses.arff**, or custom CSV datasets.

1. Load datasets in WEKA and explore data formats (ARFF/CSV)
2. Perform data cleaning and handle missing values using filters
3. Apply normalization and discretization on numeric attributes
4. Reduce data using attribute selection and PCA
5. Summarize and visualize data using statistical tools and class-wise comparison in WEKA.
6. Generate association rules using the Apriori algorithm
7. Apply multilevel association rule mining using hierarchical attributes
8. Apply K-means clustering and interpret the cluster outputs.
9. Perform hierarchical clustering and visualize results using dendrograms.
10. Apply Expectation-Maximization (EM) clustering and analyze cluster summaries.
11. Build a decision tree classifier using J48 and evaluate its performance.
12. Perform Naive Bayes classification and compare with decision tree results.
13. Apply rule-based classification using PART or JRip algorithms.
14. Compare classifiers using confusion matrix, accuracy, and ROC curves.
15. Perform basic text preprocessing and clustering using TF-IDF and K-means.