CSPE3009 DATA MINING AND DATA WAREHOUSING (3-0-0)

Course Objectives:

- To understand the need of Data Warehouses.
- To conceptualize the architecture of a Data Warehouse.
- To understand the role and functions of Data Warehouse and Data Mining.
- To explain the stages and process different data mining techniques.
- To learn mining and warehouse techniques through the use of different tools

Module – I: (07 Hours)

Fundamentals of Data Mining

Introduction, Data Mining Functionalities, Classification of Data Mining systems, knowledge discovery in databases (KDD), challenges, Data Mining Task Primitives, Integration of a Data Mining System with a Database or Data Warehouse System, Major issues in Data Mining-Data Preprocessing: Need for Preprocessing the Data, Data Cleaning, Data Integration &Transformation, Data Reduction, Discretization and Concept Hierarchy Generation.

Module –II: (08 Hours)

Data Warehouse

Introduction to Data warehouse, Difference between operational database systems and data warehouses, Data warehouse Characteristics, Data warehouse Architecture and its Components, Extraction-Transformation-Loading, Logical(Multi-Dimensional), Data Modeling, Schema Design, Star and Snow-Flake Schema, Fact Constellation, Fact Table, Fully Addictive, Semi-Addictive, Non Addictive Measures; Fact-Less-Facts, Dimension Table Characteristics; OLAP Cube, OLAP Operations, OLAP Server Architecture-ROLAP, MOLAP and HOLAP.

Module –III: (08 Hours)

Association Rules

Problem Definition, Frequent Item Set Generation, The APRIORI Principle, Support and Confidence Measures, Association Rule Generation; APRIOIRI Algorithm, The Partition Algorithms, FP-Growth Algorithms, Compact Representation of Frequent Item Set- Maximal Frequent Item Set, Closed Frequent Item Set.

Module –IV: (09 Hours)

Classification And Prediction

Problem Definition, General Approaches for solving a classification problem, Evaluation of Classifiers, Classification techniques, Decision Trees-Decision tree Construction, Methods for Expressing attribute test conditions, Measures for Selecting the Best Split, Algorithm for Decision tree Induction; Naive-Bayes Classifier, Bayesian Belief Networks; K- Nearest neighbor classification-Algorithm and Characteristics.

Prediction: Accuracy and Error measures, evaluating the accuracy of classifier or a predictor, Ensemble methods

Module –V: (08 Hours)

Clustering

Clustering Overview, A Categorization of Major Clustering Methods, Partitioning Methods, Hierarchical Methods, Partitioning Clustering-K-Means Algorithm, PAM Algorithm; Hierarchical Clustering-Agglomerative Methods and divisive methods, Basic Agglomerative Hierarchical Clustering Algorithm, Key Issues in Hierarchical Clustering, Strengths and Weakness, Outlier Detection.

Course Outcomes:

After the completion of the course, students will be able to:

- Understand warehousing architectures and tools for systematically organizing large database and use their data to make strategic decisions.
- Apply knowledge discovery in databases (KDD) process for finding interesting pattern from warehouse.
- Analyze the kinds of patterns that can be discovered by association rule mining.
- Evaluate interesting patterns from large amounts of data to analyze for predictions and classification.
- Design suitable methods for data mining and analysis.

Text Books:

1. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Education.
2. Data Mining-Concepts and Techniques- Jiawei Han, Micheline Kamber, Morgan Kaufmann Publishers, Elsevier, 2 Edition, 2006.

Reference Books:

1. George M. Marakas, “Modern Data Warehousing, Mining and Visualization: Core Concepts”, Pearson Education.
2. Alex Berson & Stephen J. Smith, “Data Warehousing, Data Mining & OLAP”, Tata McGraw-Hill.

Weblinks & Video Lectures (E-Resources):

1. https://onlinecourses.nptel.ac.in/noc20_cs12/preview
2. https://onlinecourses.swayam2.ac.in/cec19_cs01/preview