BCS-328 Data Warehousing and Data Mining

Course Category: Program Elective (PE1)
Pre-requisite Subject: NIL
Contact Hours/Week Lecture: 3, **Tutorial:** 1, **Practical:** 0
Number of Credits: 4

Course Assessment Methods: Continuous Assessment (Assignments, Quizzes, Tutorials, Minor Test), and One Major Theory Examination.

Course Objectives:

1. Study data warehouse principles and its working.
2. Learn Data mining concepts and understand Association Rule Mining.
3. Study Classification Algorithms.
4. Gain knowledge of how data is grouped using clustering techniques.

Course Outcomes:

1. Comparison of functional differences between data warehouse and database systems.
2. Ability to perform the pre-processing of data and apply mining techniques on it.
3. Capability to identify the association rules, classification and clusters in large data sets.
4. Skills to solve real world problems in business and scientific information using data mining.

UNIT-I

Introduction to data warehousing: characteristics, difference between operational databases and data warehouses, and data warehouse architecture and its components. ETL (Extraction, Transformation, and Loading) process. Schema design techniques: star schema, snowflake schema, and fact constellation. Concepts of fact table, dimension table, fact-less facts, and types of measures (fully additive, semi-additive, non-additive). Overview of OLAP: OLAP cube, OLAP operations, and OLAP server architectures (ROLAP, MOLAP, HOLAP).

UNIT-II

Introduction to data mining: functionalities, classification of data mining systems, task primitives, and integration with databases or data warehouses. Major issues in data mining. Data preprocessing techniques: need for preprocessing, data cleaning, data integration and transformation, data reduction, discretization, and concept hierarchy generation.

UNIT-III

Association rule mining: problem definition, support and confidence measures, Apriori principle and Apriori algorithm, partition algorithm, and FP-growth algorithm. Compact representations of frequent item sets: maximal and closed frequent item sets. Classification: problem definition, general approaches, evaluation of classifiers, decision tree construction, attribute test conditions, and measures for selecting the best split. Overview of Naive Bayes and Bayesian belief networks. Introduction to K-nearest neighbor classification: algorithm and characteristics.

UNIT-IV

Prediction: accuracy and error measures, evaluating classifiers and predictors, and ensemble methods. Clustering techniques: overview and categorization of clustering methods. Partitioning methods (K-means, PAM) and hierarchical methods (agglomerative and divisive). Basic agglomerative hierarchical clustering algorithm, key issues in clustering, strengths and weaknesses of clustering techniques, and an introduction to outlier detection.

Textbooks:

1. ***Data Mining: Concepts and Techniques*** – Jiawei Han, Micheline Kamber, Jian Pei, Morgan Kaufmann Publishers, Elsevier, 3rd Edition, 2011.
2. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Educatior.

Reference books:

1. Data Mining Techniques, Arun KPujari, 3rd Edition, Universities Press.
2. Data Warehousing Fundament's, Pualraj Ponnaiah, Wiley Student Edition.
3. The Data Warehouse Life CycleToolkit — Ralph Kimball, Wiley Student Edition.
4. Data Mining, Vikaram Pudi, P Rddha Krishna, Oxford University Press