DATA MINING

Paper Code	**CEN-705**
Course Credits	**4**
Lectures / week	**3**
Tutorial / week	**1**

Course Description

UNIT – I

Introduction to Data Mining: KDD, Process and Data Mining; KDD Steps; Types of Data for Data Mining, Data Mining Functionalities: Data Characterization, Data Discrimination, Mining Frequent Patterns, Association, Correlation, Classification, Prediction, Cluster Analysis, Outlier Analysis, and Evolution Analysis; Classification of Data Mining Systems; Data Mining Task Primitives; Major Issues in Data Mining.

Data Preprocessing: Introduction to Data Preprocessing; Descriptive Data Summarization: Measuring and Central Tendency and Dispersion of Data; Visualization of Descriptive Data Summaries; Data Cleaning: Handling Missing Values, Filtering Noisy Data – Binning Method; Data Integration; Data Transformation: Smoothing, Aggregation, Generalization, Normalization and Feature Selection; Data Reduction; Data Discretization and Concept Hierarchy Generation.

UNIT- II

Data Warehouse and OLAP Technology: Introduction to Data Warehouse; Features of Data Warehouse; Operational Database Systems vs. Data Warehouses; Difference Between OLTP and OLAP; Multidimensional Data Models: Data cubes, Star Schema, Snowflake Schema; Concept Hierarchies; Various OLAP Operations; Three-Tier Data Warehouse Architecture; Types of OLAP Servers: ROLAP, MOLAP, and HOLAP.

Association Rule Mining: Market basket Analysis; Frequent Itemsets, Closed Itemsets, and Association Rules; Support and Confidence; Apriori Algorithm for Mining Frequent Itemsets Using

Candidate Generation; Generating Association Rules from Frequent Itemsets; Improving the Efficiency of Apriori Algorithm; FP-Growth Algorithm for Mining Frequent Itemsets without Candidate Generation; Mining Closed Frequent Itemsets; Correlation Analysis.

UNIT- III

Classification Rule Mining: Introduction to Classification and Prediction; Classification by Decision Induction; Attribute Selection Measures: Information Gain, Gain Ratio, and Gini Index; Tree Pruning; Bayesian Classification: Bayes' Theorem, Naïve Bayesian Classification, Bayesian Belief Networks; Classifier Accuracy Measures: Sensitivity, Specificity, Precision, and Accuracy; Predictor Error Measures; Accuracy Evaluation Methods: Holdout, Random Subsampling, Cross-validation, and Bootstrap; Accuracy Enhancement Methods: Bagging and Boosting; Lazy Learners: K-Nearest-Neighbour Classifier; Prediction : Introduction to Linear and Non-Linear Regression

UNIT- IV

Cluster Analysis: Introduction to Cluster and Clustering; Features Required for Clustering Algorithms; Data Types and Dissimilarity Measures in Cluster Analysis; Categorization of Clustering Methods; Partitioning-Based Clustering: k-means Algorithms, k-medoids algorithms (PAM, CLARA, CLARANS); Hierarchical Clustering: Agglomerative and Divisive Methods (e.g.: AGNES, DIANA, BIRCH); Density-Based Clustering: DBSCAN. Methods of Outlier Analysis.

UNIT – V

Introduction to Web Mining and Text mining. Complexities involved in Web data and Text data. Introduction to Natural Language Processing techniques.   Problem discussions and Case study.

References / Text Books:

1. Jiawei Han, Micheline Kamber: Data Mining Concepts and Techniques, 3rd Edition, Morgan Kaufman Publishers.
2. Tan, Steinbach and Kumar: Introduction to Data Mining – Pearson Publication.

3. H. Witten and E. Frank: Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations, 2nd Edition, Morgan Kaufmann, Publishers.
4. Web Data Mining – Bing Lui, Springer Publication.

Computer Usage / Software Requires: Practical implementation can be done on any tool like WEKA, Rapid Miner, GATE for problems of Association rule Mining, Classification and Clustering.