DATA MINING AND WAREHOUSING
(Common to CSE&IT)

III B. Tech. - II Semester	L	T	P	C
Course Code: A3CS26	3	1	-	3

COURSE OVERVIEW:

This course helps the students to understand the overall architecture of a data warehouse and methods for data gathering and data pre-processing using OLAP tools. The different data mining models and techniques will be discussed in this course. Data mining and data warehousing applications in bioinformatics will also be explored.

COURSE OBJECTIVES:

1. To teach the basic principles, concepts and applications of data warehousing and data mining
2. To introduce the task of data mining as an important phase of knowledge recovery process
3. To familiarize Conceptual, Logical, and Physical design of Data Warehouses OLAP applications and OLAP deployment
4. To impart knowledge of the fundamental concepts that provide the foundation of data mining

COURSE OUTCOMES:

After undergoing the course, Students will be able to understand

1. Design a data mart or data warehouse for any organization
2. Develop skills to write queries using DMQL
3. Extract knowledge using data mining techniques
4. Adapt to new data mining tools.
5. Apply the techniques of clustering, classification, association finding, feature selection and visualization to real world data

SYLLABUS

UNIT - I

INTRODUCTION TO DATA MINING: Motivation, Importance, Definition of Data Mining, Kind of Data,Data Mining Functionalities, Kinds of Patterns, Classification of Data Mining Systems, Data Mining Task Primitives, Integration of A Data Mining System With A Database or Data Warehouse System, Major Issues In Data Mining, Types of Data Sets and Attribute Values, Basic Statistical Descriptions of Data, Data Visualization, Measuring Data Similarity.

PREPROCESSING: Data Quality, Major Tasks in Data Preprocessing, Data Reduction, DataTransformation and Data Discretization, Data Cleaning and Data Integration.

UNIT - II

DATA WAREHOUSING AND ON-LINE ANALYTICAL PROCESSING: Data Warehouse basicconcepts, Data Warehouse Modeling - Data Cube and OLAP, Data Warehouse Design and Usage, Data Warehouse Implementation, Data Generalization by Attribute-Oriented Induction.

DATA CUBE TECHNOLOGY: Efficient Methods for Data Cube Computation, Exploration andDiscovery in Multidimensional Databases.

UNIT - III

MINING FREQUENT PATTERNS, ASSOCIATIONS AND CORRELATIONS: Basic Concepts,Efficient and Scalable Frequent Item set Mining Methods, Are All the Pattern Interesting, Pattern Evaluation Methods, Applications of frequent pattern and associations.

FREQUENT PATTERN AND ASSOCIATION MINING: A Road Map, Mining Various Kinds ofAssociation Rules, Constraint-Based Frequent Pattern Mining, Extended Applications of Frequent Patterns.

UNIT - IV

CLASSIFICATION: Basic Concepts, Decision Tree Induction, Bayesian Classification Methods, Rule-Based Classification, Model Evaluation and Selection, Techniques to Improve Classification Accuracy: Ensemble Methods, Handling Different Kinds of Cases in Classification, Bayesian Belief

Networks, Classification by Neural Networks, Support Vector Machines, Pattern-Based Classification, Lazy Learners (or Learning from Your Neighbors), Other Classification Methods.

UNIT - V

CLUSTER ANALYSIS: Basic Concepts of Cluster Analysis, Clustering structures, Major ClusteringApproaches, Partitioning Methods, Hierarchical Methods, Density-Based Methods, Model-Based Clustering - The Expectation-Maximization Method, Other Clustering Techniques, Clustering High-Dimensional Data, Constraint-Based and User-Guided Cluster Analysis, Link-Based Cluster Analysis, Semi-Supervised Clustering and Classification, Bi-Clustering, Collaborative Clustering.

OUTLIER ANALYSIS: Why outlier analysis, Identifying and handling of outliers, Distribution-BasedOutlier Detection: A Statistics-Based Approach, Classification-Based Outlier Detection, Clustering-Based Outlier Detection, Deviation-Based Outlier Detection, Isolation-Based Method: From Isolation Tree to Isolation Forest.

TEXT BOOKS:

1. Jiawei Han, MichelineKamber, Jian Pei (2012), Data Mining: Concepts and Techniques, 3rdedition, Elsevier, United States of America.

REFERENCE BOOKS:

1. Margaret H Dunham (2006), Data Mining Introductory and Advanced Topics, 2ndedition,Pearson Education, New Delhi, India.
2. Amitesh Sinha (2007), Data Warehousing, Thomson Learning, India.
3. Xingdong Wu, Vipin Kumar (2009), the Top Ten Algorithms in Data Mining, CRC Press, UK.