III B.Tech – II Semester
(17CS603) DATA WARE HOUSING AND MINING

Int. Marks	Ext. Marks	Total Marks	L	T	P	C
40	60	100	3	1	-	3

Pre-Requisites: Data Base Management System, C Programming

Course Objectives:

- Students will be enabled to understand and implement classical models and algorithms in data warehousing and data mining.
- They will learn how to analyze the data, identify the problems, and choose the relevant models and algorithms to apply.
- They will further be able to assess the strengths and weaknesses of various methods and algorithms and to analyze their behavior.

UNIT-I: Introduction to Data Mining: What is data mining, motivating challenges, origins of data mining, data mining tasks , Types of Data-attributes and measurements, types of data sets, Data Quality (Tan)

UNIT-II: Data pre-processing, Measures of Similarity and Dissimilarity: Basics, similarity and dissimilarity between simple attributes, dissimilarities between data objects, similarities between data objects, examples of proximity measures: similarity measures for binary data, Jaccard coefficient, Cosine similarity, Extended Jaccard coefficient, Correlation, Exploring Data : Data Set, Summary Statistics (Tan)

UNIT-III: Data Warehouse: basic concepts:, Data Warehousing Modeling: Data Cube and OLAP, Data Warehouse implementation : efficient data cube computation, partial materialization, indexing OLAP data, efficient processing of OLAP queries. (H & C)

UNIT-IV: Classification: Basic Concepts, General approach to solving a classification problem, Decision Tree induction: working of decision tree, building a decision tree, methods for expressing attribute test conditions, measures for selecting the best split, Algorithm for decision tree induction.

Model over fitting: Due to presence of noise, due to lack of representation samples, evaluating the performance of classifier: holdout method, random sub sampling, cross-validation, bootstrap. (Tan)

UNIT-V: Association Analysis: Problem Definition, Frequent Item-set generation- The Apriori principle , Frequent Item set generation in the Apriori algorithm, candidate generation and pruning, support counting (eluding support counting using a Hash tree) , Rule generation, compact representation of frequent item sets, FP-Growth Algorithms. (Tan)

UNIT-VI:

Overview- types of clustering, Basic K-means, K –means –additional issues, Bisecting k-means,k-means and different types of clusters, strengths and weaknesses, k-means as an optimization problem. Agglomerative Hierarchical clustering, basic agglomerative hierarchical clustering algorithm, specific techniques, DBSCAN: Traditional density: centre-based approach, strengths and weaknesses (Tan)

Correlation of COs with POs & PSOs:

	PO-1	PO-2	PO-3	PO-4	PO-5	PO-6	PO-7	PO-8	PO-9	PO-10	PO-11	PO-12	PSO-1	PSO-2	PSO-3
CO-1	3	3	2	2	1	-	-	-	-	-	-	2	-	2	2
CO-2	3	3	2	3	2	-	-	-	-	-	-	2	1	-	2
CO-3	3	2	3	2	1	-	-	-	-	-	-	3	1	-	1
CO-4	3	2	2	3	2	-	-	-	-	-	-	-	-	-	2
CO-5	3	1	2	2	1	-	-	-	-	-	-	-	-	2	3
CO-6	3	3	2	3	1	-	-	-	-	-	-	-	-	-	3

Text Books:

1. Introduction to Data Mining : Pang-Ning tan, Michael Steinbach, Vipin Kumar, Pearson
2. Data Mining ,Concepts and Techniques, 3/e, Jiawei Han , Micheline Kamber , Elsevier

Reference Books:

1. Introduction to Data Mining with Case Studies 2nd ed: GK Gupta; PHI.
2. Data Mining : Introductory and Advanced Topics : Dunham, Sridhar, Pearson.
3. Data Warehousing, Data Mining & OLAP, Alex Berson, Stephen J Smith, TMH
4. Data Mining Theory and Practice, Soman, Diwakar, Ajay, PHI, 2006.