III Year II Semester

L	T	P	C
3	1	0	3

Code: 17CS603

DATA WARE HOUSING AND MINING

OBJECTIVES:

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

Course outcomes:

- Understand stages in building a Data Warehouse
- Understand the need and importance of preprocessing techniques
- Understand the need and importance of Similarity and dissimilarity techniques
- Analyze and evaluate performance of algorithms for Association Rules.
- Analyze Classification and Clustering algorithms

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.