III Year – II SEMESTER

T	P	C
4	0	3

Data Ware housing and Mining

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.

Course Outcomes:

1. understand why there is a need for data warehouse in addition to traditional operational database systems;
2. identify components in typical data warehouse architectures;
3. design a data warehouse and understand the process required to construct one;
4. understand why there is a need for data mining and in what ways it is different from traditional statistical techniques;
5. understand the details of different algorithms made available by popular commercial data mining software;
6. solve real data mining problems by using the right tools to find interesting patterns

Syllabus:

UNIT –I:

Introduction : What Motivated Data Mining? Why Is It Important, Data Mining—On What Kind of Data, Data Mining Functionalities—What Kinds of Patterns Can Be Mined? Are All of the Patterns Interesting? 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. **(Han & Kamber)**

UNIT –II:

Data Pre-processing : Why Pre-process the Data? Descriptive Data Summarization, Data Cleaning, Data Integration and Transformation, Data Reduction, Data Discretization and Concept Hierarchy Generation. **(Han & Kamber)**

UNIT –III:

Data Warehouse and OLAP Technology: An Overview : What Is a Data Warehouse? A Multidimensional Data Model, Data Warehouse Architecture, Data Warehouse Implementation, From Data Warehousing to Data Mining. **(Han & Kamber)**

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 an 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 & Vipin)**

UNIT –V

Association Analysis: Basic Concepts and Algorithms : Introduction, Frequent Item Set generation, Rule generation, compact representation of frequent item sets, FP-Growth Algorithm. **(Tan & Vipin)**

UNIT –VI

Cluster Analysis: Basic Concepts and Algorithms : What Is Cluster Analysis? Different Types of Clustering, Different Types of Clusters, K-means, The Basic K-means Algorithm, 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: Center-Based Approach, The DBSCAN Algorithm, Strengths and Weaknesses. **(Tan & Vipin)**

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, Michel Kamber, Elsevier.

Reference Books :

1. Data Mining Techniques and Applications: An Introduction, Hongbo Du, Cengage Learning.
2. Data Mining : Introductory and Advanced topics : Dunham, Pearson.
3. Data Warehousing Data Mining & OLAP, Alex Berson, Stephen Smith, TMH.
4. Data Mining Techniques, Arun K Pujari, Universities Press.