GUJARAT TECHNOLOGICAL UNIVERSITY

Bachelor of Engineering

Subject Code: 3161610

DATAWAREHOUSING AND DATA MINING

B.E. 6th SEMESTER

Type of course: Under graduate (Elective)

Prerequisite: NA

Rationale: NA

Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks				Total Marks
L	T	P	C	Theory Marks		Practical Marks		
				ESE (E)	PA (M)	ESE (V)	PA (I)	
3	0	2	4	70	30	30	20	150

Content:

Sr. No.	Content	Total Hrs.	% Weightage
1	**Data Warehousing:** OLAP & OLTP, Data warehouse & Data mart, OLAM architecture, Extraction, Transform & Loading (ETL) concept for generic, two-tier, three - tier architecture, Data warehousing schema - Star, Snowflake, Fact Constellation (Galaxy) - Data Cube , Operations on Data cube (slicing, roll up, roll down, drill up etc)	5	10
1	**Introduction to data mining (DM):** Motivation for Data Mining - Data Mining-Definition and Functionalities – Classification of DM Systems - DM task primitives - Integration of a Data Mining system with a Database or a Data Warehouse - Issues in DM – KDD Process	3	10
2	**Data Pre-processing:** Data summarization, data cleaning, data integration and transformation, data reduction, data discretization and concept hierarchy generation, feature extraction , feature transformation, feature selection, introduction to Dimensionality Reduction, CUR decomposition	4	15
3	**Mining Frequent Patterns, Associations and Correlations:** Efficient and scalable frequent item-set mining methods, mining various kind of association rules, from association mining to correlation analysis, Advanced Association Rule Techniques, Measuring the Quality of Rules.	7	20
4	**Classification and Prediction:** Classification vs. prediction, Issues regarding classification and prediction, Statistical-Based Algorithms, Distance-Based Algorithms, Decision Tree-Based Algorithms, Neural Network-Based Algorithms, Rule-Based Algorithms, Combining Techniques, accuracy and error measures, evaluation of the accuracy of a classifier or predictor. Neural Network Prediction methods: Linear and nonlinear regression, Logistic Regression Introduction of tools such as DB Miner / WEKA / DTREG DM Tools	10	20
5	**Cluster Analysis:**	10	20

	Clustering: Problem Definition, Clustering Overview, Evaluation of Clustering Algorithms, Partitioning Clustering -K-Means Algorithm, K-Means Additional issues, PAM Algorithm; Hierarchical Clustering – Agglomerative Methods and divisive methods, Basic Agglomerative Hierarchical Clustering, Strengths and Weakness; Outlier Detection, Clustering high dimensional data, clustering Graph and Network data.		
8	**Advance topics:** Introduction to Web Mining, Spatial Data Mining, Temporal Mining, Text Mining and Multimedia Mining.	**3**	**10**

Suggested Specification table with Marks (Theory):

Distribution of Theory Marks					
R Level	U Level	A Level	N Level	E Level	C Level
10	**20**	**15**	**15**	**5**	**5**

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create and above Levels (Revised Bloom's Taxonomy)

Note: This specification table shall be treated as a general guideline for students and teachers. The actual distribution of marks in the question paper may vary slightly from above table.

Reference Books:

1. J. Han, M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann
2. M. Kantardzic, "Data mining: Concepts, models, methods and algorithms, John Wiley &Sons Inc.
3. Paulraj Ponnian, "Data Warehousing Fundamentals", John Willey.
4. M. Dunham, "Data Mining: Introductory and Advanced Topics", Pearson Education.
5. Ning Tan, Vipin Kumar, Michael Steinbanch Pang, "Introduction to Data Mining", Pearson Education

Course Outcome: After learning the course the students will be able

1. Understand why the data warehousing is important in addition to database systems.
2. Perform the preprocessing of data and apply mining techniques on it.
3. Identify the association rules, classification and clusters in large data sets.
4. Solve real world problems in business and scientific information using data mining.
5. Use data analysis tools for scientific applications.
6. Implement various supervised machine learning algorithms.

List of Experiments:
Laboratory work will be based on the above syllabus with minimum 10 experiments to be incorporated.