GUJARAT TECHNOLOGICAL UNIVERSITY
MASTER OF COMPUTER APPLICATIONS (MCA)
Semester: IV

Subject Name**: Elective I – Data Warehousing & Data Mining (DWDM)**
Subject Code**: 2640005**

Learning Objectives:

- ✓ To understand the need of Data Warehouses over Databases, and the difference between usage of operational and historical data repositories.
- ✓ To be able to differentiate between RDBMS schemas & Data Warehouse Schemas.
- ✓ To understand the concept of Analytical Processing (OLAP) and its similarities & differences with respect to Transaction Processing (OLTP).
- ✓ To conceptualize the architecture of a Data Warehouse and the need for pre-processing.
- ✓ To understand the need for Data Mining and advantages to the business world. The validating criteria for an outcome to be categorized as Data Mining result will be understood.
- ✓ To get a clear idea of various classes of Data Mining techniques, their need, scenarios (situations) and scope of their applicability.
- ✓ To learn the algorithms used for various type of Data Mining problems.

Pre-requisites: Knowledge of RDBMS and OLTP

Contents:

1. **Introduction to Data Warehousing, A Multi-dimensional Data Model & Schemas, OLAP Operations & Servers (6 Lect.)**
   - An overview and definition along with clear understanding of the four key-words appearing in the definition.
   - Differences between Operational Database Systems and Data Warehouses; Difference between OLTP & OLAP
   - Overview of Multi-dimensional Data Model, and the basic differentiation between "Fact" and "Dimension"; Multi-dimensional Cube
   - Concept Hierarchies of "Dimensions" Parameters: Examples and the advantages
   - Star, Snowflakes, and Fact Constellations Schemas for Multi-dimensional Databases
   - Measures: Their Categorization and Computation
   - Pre-computation of Cubes, Constraint on Storage Space, Possible Solutions
   - OLAP Operations in Multi-dimensional Data Model: Roll-up, Drill-down, Slice & Dice, Pivot (Rotate)
   - Indexing OLAP Data; Efficient Processing of OLAP Queries
   - Type of OLAP Servers: ROLAP versus MOLAP versus HOLAP
   - Metadata Repository

2. **Data Warehouse Architecture; Further Development of Data Cube & OLAP Technology (3 Lect.)**
   - The Design of A Data Warehouse: A Business Analysis Framework; The Process of Data Warehouse Design
   - A 3-Tier Data Warehouse Architecture; Enterprise Warehouse, Data mart, Virtual Warehouse
   - Discovery-Driven Exploration of Data Cubes; Complex Aggregation at Multiple Granularity: Multi-feature Cubes
   - Constrained Gradient Analysis of Data Cubes

3. **P re-processing (7 Lect.)**
   - The need for Pre-processing, Descriptive Data Summarization
   - Data Cleaning: Missing Values, Noisy Data, Data Cleaning as a Process
   - Data Integration & Transformation
   - Data Cube Aggregation; Attribute Subset Selection
   - Dimesionality Reduction: Basic Concepts only
   - Numerosity Reduction: Regression & Log-linear Models, Histograms, Clustering, Sampling
   - Data Dicretization & Concept Hierarchy Generation
   - For Numerical Data: Binning, Histogram Analysis, Entropy-based Discretization, Interval Merging by χ^2 Analysis, Cluster Analysis, Discretization by Intuitive Partitioning
   - For Categorical Data

4. **Data Mining: Introduction (4 Lect.)**
   - An Overview; What is Data Mining; Data Mining – on What Kind of Data
   - Data Mining Functionalities – What Kind of Patterns Can be Mined; Concept/Class Description: Characterization & Discrimination; Mining Frequent Patterns, Associations, and Correlations; Classification & Prediction; Cluster Analysis; Outlier Analysis
   - 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

5. **Attribute-Oriented Induction: An Alternate Method for Data Generalization & Concept Description (4 Lect.)**
   - Attribute-Oriented Induction for Data Characterization, and Its Efficient Implementation; Presentation of the Derived Generalization
   - Mining Class Comparisons: Discrimination between Different Classes
   - Class Descriptions: Presentation of both Characterization & Comparison

6. **Mining Frequent Patterns, Associations, and Correlations (4 Lect.)**
   - Basic Concepts: Market Basket Analysis; Frequent Itemsets, Closed Itemsets, and Association Rules; Frequent Pattern Mining: A Roadmap
   - Apriori Algorithm: Finding Frequent Itemsets Using Candidate Generation; Generating Association Rules from Frequent Itemsets; Improving the Efficiency of Apriori
   - From Association Mining to Correlation Analysis; Strong Rules Are Not Necessarily Interesting: An Example; From Association Analysis to Correlation Analysis

7. Classification & Prediction (9+2 Lect.)

- Introduction to Classification and Prediction; Basics of Supervised & Unsupervised Learning; Preparing the Data for Classification and Prediction; Comparing Classification and Prediction Methods
- Classification by Decision Tree Induction, Attribute Selection Measures; Tree Pruning; Scalability and Decision Tree Induction
- Rule-based Classification: Using IF-THEN Rules for Classification; Rule Extraction from a Decision Trees; Rule Induction Using a Sequential Covering Algorithm
- Bayesian Classification: Bayes' Theorem, Naïve Bayesian Classification; Bayesian Belief Networks
- An Overview of Other Classification Methods (2 Lectures)
- Prediction: Linear Regression; Non-linear Regression; Other Regression Models
- Classifier Accuracy and Error Measures: Classifier Accuracy Measures; Predictor Error Measures
- Evaluating the Accuracy of a Classifier or Predictor: Holdout Method and Random Sub-sampling; Cross Validation; Bootstrap
- Ensemble Methods – Increasing the Accuracy: Bagging; Boosting

8. Cluster Analysis (6+2 Lect.)

- Introduction to Cluster Analysis; Types of Data in Cluster Analysis; A Categorization of major Clustering Methods
- Partitioning Methods; Centroid-Based Technique: K-Means Method; Overview of Other Clustering Methods
- An Overview of Other Clustering Methods (2 Lectures)
- Outlier Analysis; Statistical Distribution-based Outlier Detection; Distance-based Outlier Detection; Density-based Outlier Detection; Deviation-based Outlier Detection

9. Data Mining Applications (3 Lect.)

- Data Mining for: (a) Financial Data Analysis; (b) The Retail Industry; (c) The Telecommunication Industry; (d) Biological Data Analysis; (e) Other Scientific Applications; (f) Intrusion detection
- Data Mining Systems: (a) How to Choose; (b) Examples of Commercial Data Mining Systems

Text Book:

1. Jiawei Han & Micheline Kamber, "Data Mining: Concepts & Techniques", Morgan Kaufmann Publishers (2002)

Other Reference Books:

1. W. H. Inmon, "Building the Data Warehouse", Wiley Dreamtech India Pvt. Ltd.
2. Mohanty, Soumendra, "Data Warehousing: Design, Development and Best Practices", Tata McGraw Hill (2006)
3. Pieter Adriaans & Dolf Zentinge, "Data Mining", Addison-Wesley, Pearson (2000) Rs. 195/-
4. Daniel T. Larose, "Data Mining Methods & Models", Wiley-India (2007)
5. Vikram Pudi & P. Radhakrishnan, "Data Mining", Oxford University Press (2009)
6. Alex Berson & Stephen J. Smith, "Data Warehousing, Data Mining & OLAP", Tata McGraw-Hill (2004)
7. Michael J. A. Berry & Gordon S. Linoff, "Data Mining Techniques", Wiley-India (2008)
8. Richard J. Roiger & Michael W. Geatz, "Data Mining – a Tutorial-based Primer", Pearson Education (2005)
9. Margaret H. Dunham & S. Sridhar, "Data Mining: Introductory and Advanced Topics", Pearson Education (2008) Rs. 235/-
10. G. K. Gupta, "Introduction to Data Mining with Case Studies", EEE, PHI (2006) Rs. 325/-

Chapter wise Coverage from the Text Books:

Unit-1: 3.1, 3.1.1, 3.2, 3.2.1 to 3.2.6, 3.4.1 to 3.4.3, 3.3.4, 3.3.5
Unit-2: 3.3, 3.3.1, 3.3.2, 4.2.1 to 4.2.3
Unit-3: 2.1, 2.2, 2.2.1 to 2.2.3, 2.3.1 to 2.3.3, 2.4.1, 2.4.2, 2.5.1, 2.5.2, (Introductory Portion of 2.5.3), 2.5.4, 2.6, 2.6.1, 2.6.2
Unit-4: 1.1 to 1.3: 1.3.1 to 1.3.4, 1.4, 1.4.1 to 1.4.5, 1.5 to 1.9
Unit-5: 4.3.1 to 4.3.5
Unit-6: 5.1.1 to 5.1.3, 5.2.1 to 5.2.3, 5.4, 5.4.1, 5.4.2
Unit-7: 6.1, 6.2, 6.2.1, 6.2.2, 6.3, 6.3.1 to 6.3.4, 6.5, 6.5.1 to 6.5.3, 6.4, 6.4.1 to 6.4.3, 6.11, 6.11.1 to 6.11.3, 6.12, 6.12.1, 6.12.2, 6.13, 6.13.1 to 6.13.3, 6.14, 6.14.1, 6.14.2
Unit-8: 7.1, 7.2, 7.2.1 to 7.2.5, 7.3, 7.4, 7.4.1, 7.11, 7.11.1 to 7.11.4
Unit-9: 11.1, 11.1.1 to 11.1.6, 11.2, 11.2.1, 11.2.2

Accomplishment of the students after completing the course:

- Ability to create a Star Schema for a given Data Warehousing requirements
- Ability to decide the number & levels of pre-computed Data Cubes, the corresponding Metadata, and the appropriate OLAP operation
- Ability to apply pre-processing on existing operational & historical data for creation of Data Warehouse
- Ability to apply Apriori algorithm for Association Mining
- Ability to apply Decision Tree and Bayesian algorithms for Classification
- Ability to mine Statistical Measures in large databases
- Ability to differentiate between Classification & Clustering, and similarly between Supervised Learning & Unsupervised Learning

Suggested Continuous Evaluation Components (CEC):

Case study

1. Data Warehouse Applications: CRM; SCM; Banking sector; Insurance sector; Retail banking Industry case study, Hospital application.
2. Design a data mart from scratch to store the credit history of customers of a bank. Use this credit profiling to process future loan applications.

3. Design and build a Data Warehouse using bottom up approach titled 'Citizen Information System'. This should be able to serve the analytical needs of the various government departments and also provide a global integrated view.

Group Project

Based on their collective work experience, each group should identify, and to the extent possible, execute a business intelligence project that relies on the data mining techniques we will cover in the class. The key tasks here are:

- To identify a business problem or a series of interesting questions that deal with either classification, prediction or clustering
- Identify sources of data that could potentially be useful in addressing your questions
- Pre-process – clean, validate, visualize your data
- Develop your model considering alternative techniques, selecting the most appropriate one in the process.
- Interpret your results, and write a final report including an executive summary of your findings. This will be due during the finals week.
- Prepare a 10-15 minute presentation for the last class meeting

Laboratory Exercise

The objective of the lab exercises is to use data mining techniques to identify customer segments and understand their buying behavior and to use standard databases available to understand DM processes using WEKA (or any other DM tool)

1. Gain insight for running pre- defined decision trees and explore results using MS OLAP Analytics.
2. Using IBM OLAP Miner – Understand the use of data mining for evaluating the content of multidimensional cubes.
3. Using Teradata Warehouse Miner – Create mining models that are executed in SQL.

BI Portal Lab: The objective of the lab exercises is to integrate pre-built reports into a portal application

4. Publish cognos cubes to a business intelligence portal.

Metadata & ETL Lab: The objective of the lab exercises is to implement metadata import agents to pull metadata from leading business intelligence tools and populate a metadata repository. To understand ETL processes

5. Import metadata from specific business intelligence tools and populate a metadata repository.
6. Publish metadata stored in the repository.
7. Load data from heterogeneous sources including text files into a pre-defined warehouse schema.

Major Tools for Lab Exercise

1. Weka (an Open Source) by The University of Waikato
2. IBM Intelligent Miner
3. MS OLAP Analytics
4. XLMiner
5. Programming in "R"