# GUJARAT TECHNOLOGICAL UNIVERSITY

**With effective from academic year 2018-19**

**Syllabus for Integrated Master of Computer Applications, 6th Semester**

**Subject Name: Data Mining**

**Subject Code: 2668603**

## 1) Learning Objectives:

i. To understand concept of Business Intelligence, Business Analytics and Data Warehouse

ii. To understand the concept of Analytical Processing (OLAP) and its similarities & differences with respect to Transaction Processing (OLTP).

iii. To understand the need for data pre-processing.

iv. To get a clear idea of various classes of Data Mining techniques, their need, scenarios situations) and scope of their applicability.

v. To learn the algorithms used for various types of Data Mining problems.

## 2) Prerequisites: Knowledge of RDBMS and OLTP

## 3) Contents:

| Unit | Chapter Details | Weightage Percentage |
|---|---|---|
| **Unit I** | **Introduction to Data Warehouse, Business Intelligence and Business Analytics**<br>**Data Warehouse**: What is a Data Warehouse? Differences between operational database systems and Data Warehouses, Why have a separate data warehouse? , Data Warehousing : A multi-tiered architecture, Data Warehouse Models, Enterprise Warehouse, Datamart and Virtual Warehouse, Extraction, Transformation and Loading, Meta data Repository<br>**Data Warehousing Modeling:**<br>Data Cube, Stars, Snowflakes and Fact constellation schemas for Multi-dimensional data models. Dimensions: The role of Concept Hierarchies, Measures: Categorization and computation, Typical OLAP Operations, Starnet Query model for Querying multidimensional databases. | **15%** |
| **Unit II** | **Introduction: Data Mining**<br>What is Data Mining; Why Data Mining? What kind of data can be mined? What Kind of Patterns Can be Mined?<br>Which technologies are used? (Statistics, Machine learning, Database systems and data warehouses, information retrieval), Which kinds of applications are targeted ? ( Business Intelligence, Web Search Engines) , Major Issues in Data Mining<br>Data Pre-processing: An Overview, Data Cleaning, Data Integration, Data Reduction and Data Transformation and Data Discretization | **15%** |
| **Unit III** | **Mining Frequent Patterns, Associations, and Correlations**<br>**Basic Concepts**: Market-Basket Analysis, Frequent Item sets, Closed Item sets and Association rules. | **20%-** |

| | | |
|---|---|---|
| | **Frequent Itemset Mining methods**: Apriori algorithm, Generating association rules from frequent itemsets, improving efficiency of apriori.Pattern Evaluation Methods. | |
| **Unit IV** | **Classification & Prediction: Basic concepts and Methods**<br>**Classification:** Basic Concepts<br>**Decision Tree Induction:** Decision Tree Induction, Attribute Selection Measures; Tree Pruning; Scalability and Decision Tree Induction, Visual Mining for Decision tree induction<br>**Bayesian Classification:** Bayes' Theorem, Naïve Bayesian Classification<br>**Rule-based Classification:** Using IF-THEN Rules for Classification; Rule Extraction from a Decision Trees; Rule Induction Using a Sequential Covering Algorithm<br>**Model Evaluation and Selection:** Metrics for Evaluating Classifier Performance, Holdout Methods and Random Sub sampling, Cross-validation, Bootstrap,<br>**Techniques to improve Classification Accuracy:** Introducing Ensemble Methods, Bagging, Boosting and Adaboost, Random Forests, Improving Classification Accuracy of Class-imbalanced data | **25%** |
| **Unit V** | **Cluster Analysis: Basic concepts and Methods**<br>Cluster Analysis (What is cluster analysis? Requirement for Cluster analysis, Overview of Basic Clustering Methods)<br>Partitioning Methods (K-Means, K-Medoids)<br>Hierarchical Methods (Agglomerative versus Divisive) Hierarchical Clustering, Distance Measures in Algorithmic Methods, BIRCH Multiphase Hierarchical Clustering using clustering feature tree, Chameleon Multiphase Hierarchical Clustering using Dynamic Modeling, Probablistic Hierarchical Clustering, Density Based Methods (DBSCAN and OPTICS)<br>Outlier Analysis: What are outliers? Types of outliers, Challenges of outlier Detection. | **25%** |

## 4. Text Book:

1) Han, J., Kamber, M., Pei, J. Data mining concepts and techniques. Morgan Kaufmann, 3rd Edition, 2011

5. **Reference Books:**

**1)** Vikram Pudi & P. Radhakrishnan,Data Mining, Oxford University Press (2009).
**2)** Pieter Adriaans & Dolf Zentinge¸ Data Mining, Addison-Wesley, Pearson (2000).
**3)** Daniel T. Larose , Data Mining Methods & Models, , Wiley-India (2007).
**4)** Michael J. A. Berry & Gordon S. Linoff, Data Mining Techniques, Wiley-India (2008).
**5)** Richard J. Roiger & Michael W. Geatz, Data Mining – a Tutorial-based Primer, Pearson Education (2005).
**6)** Margaret H. Dunham & S. Sridhar, Data Mining: Introductory and Advanced Topics, Pearson Education (2008).
**7)** G. K. Gupta, Introduction to Data Mining with Case Studies, EEE, PHI (2006).
**8)** K.P.Soman, Shyam Diwakar and V. Ajay , Insight of Data Mining- theory and Practice, , PHI Publication
**9)** by K.P.Soman, Shyam Diwakar and V. Ajay, Insight of Data Mining- theory and Practice, PHI Publication

**Webliography:**

https://docs.oracle.com/database/121/DWHSG/E41670-08.pdf

1. **Chapter wise Coverage from Main Reference Book(s):**

| **Unit No.** | **Text Books** | **Topics/Subtopics** |
|---|---|---|
| I | Book 1 | **Chapter 4**(4.1 , 4.2) |
| II | Book 1 | **Chapter 1 ( 1.1 to 1.7)**<br>**Chapter 3 ( 3.1 to 3.5)** |
| III | Book 1 | **Chapter 6 (** 6.1, 6.2 (6.2.1 to 6.2.3) , 6.3) |
| IV | Book 1 | **Chapter 8 (8.1 to 8.6)** |
| V | Book 1 | **Chapter 10** ( 10.1, 10.2, 10.3, 10.4.1, 10.4.2<br>**Chapter 12** (12.1) |

7. **Accomplishments of the student after completing the course:**

1. **Laboratory Exercises**
   - **Part I:** Data Warehousing
     - Identify application for which
       - Write Dimension and Fact tables.
       - Draw Star, Snowflake and fact constellation schema
       - Identify OLAP Operations

| | |
|---|---|
| 1 | Create Dimension tables as follows :<br>PRODUCT(Product key, Product name, subcategory, category, product line, department)<br>STORE(store key, store name, territory, region)<br>TIME(time key, day, month, quarter, year) |

| | |
|---|---|
| | Note:<br>Product line can be group of related product manufactured by single company. E.g. Makeup product line includes eye shadow, lipsticks, eyeliner, blush, powder etc.<br><br>Create Fact table as follow:<br>SALESFACT(Product key, Time key, Store key, fixed cost, variable cost, indirect sales, direct sales, profit margin)<br><br>Write Data Warehouse SQL query:<br>a. Display total sales of all products for past five years in all stores.<br>b. List total sales for all stores, products by products between years 2015 and 2010.<br>c. List total sales for all stores, products by products between years 2015 and 2000 only for those products with reduced sales.<br>d. List total sales for “Alpha store” between years 2014 and 2015 for all products.<br>e. Show indirect sales for all stores of last two years of Indian Territory.<br>f. List profit margin for “Lakme” store for Indian Territory with region wise for previous month.<br>g. Show total sales across all products at increasing level of geography dimension from territory to region for year 2014 and 2015.<br>h. List top 10 stores name in Asia for automobile product last year whose profit margin is high.<br>i. Count total number of stores available in North region of Asia territory.<br>j. Count total number of stores and total sales for east region for Indian Territory in descending order.<br>k. List product category along with subcategory for “Beta” store between years 2015 to 2010 region wise and territory wise.<br>l. Give product category by store by date<br>m. List all products store wise and date wise in descending order.<br>n. Give total product category for last two months for “Beta” store.<br>o. List count of all stores quarter wise and product wise.<br>p. List department of product for year 2015 for Asia territory in north region.<br>q. List all products along with its sub category for all stores in 2012 year.<br>r. Show direct sales and indirect sales month wise, region wise.<br>s. Show lowest margin sales for each month for all stores, product wise.<br>t. Give year wise total sales of each product line. |
| 2 | Create Dimension tables as follows :<br>CUSTOMER(customer key, customer name, age, gender, city, region, state)<br>FURNITURE( furniture key, furniture type, category, material)<br>TIME(time key, day, month, quarter, year)<br>Note: Furniture type can be chair, table, cabinet etc..<br>Furniture category can be kitchen, living room, office room, etc.<br>Furniture material can be wood, marble, glass, lime, plastic, steel, etc,.<br><br>Create Fact table as follow:<br>SALES(customer key, furniture key, time key, quantity, discount, income)<br>Write Data Warehouse SQL query: |

| | |
|---|---|
| | a. Find discount for each furniture type for month “July” and year 2016.<br>b. Show average income of sales for “kitchen” material category last year.<br>c. Find average income and discount for each city monthwise and year wise.<br>d. Determine top 5 most sold material during the month of “November”.<br>e. Count total quantity sold for year between 2015 and 2010.<br>f. List name of customer with its gender who order steel material for his office room last quarter.<br>g. Count total number of customers who got maximum discount last month for all furniture type. |
| 3 | Create Dimension tables as follows :<br>DOCTOR(doctor key, doctor name, city, phone number, gender, experience years, expertise )<br>PATIENT( patient key, patient name, phone number, gender, treatment, address, age)<br>TIME(time key, day, month, quarter, year)<br><br>Note:<br>gender can be “male” or “female” only.<br><br>Create Fact table as follow:<br>CLINIC ( doctor key, patient key, time key, charge, count)<br>Write Data Warehouse SQL query:<br>a. Count total number of patient who appoint for “Monday” , “June -2016”.<br>b. Show charge of doctor on each day, each month, year wise.<br>c. List all doctors who take low charge for “Diabetic” treatment.<br>d. List citywise doctors, who has highest experienced for all treatments.<br>e. List all female doctors who have “cancer” expertise for year 2015 to 2010.<br>f. Count total number of doctors’ monthwise, quarter wise and yearwise.<br>g. List average charge the doctor takes on “Sunday” to treat children of below 10 years.<br>Q |
| 4 | Create Dimension tables as follows :<br>COURSE(course key, course name, department)<br>SEMESTER(semester key, semester name, year)<br>STUDENT(student key, student name, area key, major, status, university)<br>INSTRUCTOR(instructor key, department, instructor name)<br>AREA ( area key, city, state, country)<br><br>Note:<br>major student is ‘yes’ or ‘no’ Status of student can be part time student or full time student.<br><br>Create Fact table as follow:<br>UNIV( student key, course key, semester key, instructor key, count, average grade)<br><br>Write Data Warehouse SQL query:<br>a. List average grade of computer science course for GU university students.<br>b. List department wise course key. |

| | |
|---|---|
| | c. Count total number of students for MBA course university wise, city wise.<br>d. Show average grade of students for MCA course last year.<br>e. Give top 5 average grade students who study Information Technology course for all university in Gujarat state of India.<br>f. Count total number of students for all university, course wise, area wise, state wise.<br>g. Give average grade of students for all courses between year 2014 and 2015.<br>h. Give instructor name that is in IT department for more than five years. |
| | |

- **Part II:** Data Preprocessing
  - i. Implementation of various data transformation techniques (normalization, aggregation and generalization)
  - ii. Implementation of data cleaning methods ( incomplete(missing), noisy, inconsistent data)
  - iii. iii. Data smoothing techniques.

- **Part III : Data Mining**

  Following all the Algorithms can be implemented using **C /C++ /Java/ Python/R**

  - i. Implementation of Apriori Algorithm
  - ii. Implementation of Decision Tree Algorithm
  - iii. Implementation of K-Nearest Neighbor (K-NN) Algorithm
  - iv. Implementation of K-means Clustering Algorithm
  - v. Implementation of Naïve Bayesian Classification Algorithm
  - vi. Implementation of DBSCAN algorithm (prefer R programming)
  - vii. Implementation of CART Algorithm (prefer R Programming )
  - viii. Implementation of Bin Packing Algorithm (data smoothing by bining)

**Desirable:**

Study of Rapid Miner Tool and Comparative analysis of WEKA and Rapid Miner Tool