Course Title: Data Warehousing and Data Mining

Course no: CSC-451 **Full Marks: 60+20+20**

Credit hours: 3 **Pass Marks: 24+8+8**

Nature of course: Theory (3 Hrs.) + Lab (3 Hrs.)

Course Synopsis: Analysis of advanced aspect of data warehousing and data mining.

Goals: This course introduces advanced aspects of data warehousing and data mining, encompassing the principles, research results and commercial application of the current technologies

Unit	Course content-breakdown	Lecture Hours	Remarks
1	**Introduction** ▪ What motivated Data mining? What is Data Mining? ▪ Types of databases (Relational database, Data Warehouses, Transactional Database) ▪ Functionalities of data mining – What kinds of Pattern can be mined? ▪ Association Analysis, Cluster Analysis, Outlier Analysis, Evolution Analysis ▪ Stages of Knowledge discovery in database(KDD) ▪ Setting up a KDD environment ▪ Issues in Data Warehouse and Data Mining ▪ Application of Data Warehouse and Data Mining	5	
2	**Data Warehouse for Data mining** ▪ Differences between operational database systems and data warehouses ▪ Data Warehouse Architecture ▪ Distributed and Virtual Data Warehouse ▪ Data Warehouse Manager ▪ Data marts, Metadata, Multidimensional data model ▪ From Tables and Spread Sheets to Data Cubes	4	

	▪ Star schema, Snowflake schema and Fact constellation schema		
3	**OLAP technology for Data Mining** ▪ On-line analytical processing models and operations (drill down, drill up, slice, dice, pivot) ▪ Types of OLAP Servers: ROLAP versus MOLAP versus HOLAP ▪ OLTP	6	
4	**Tuning for data warehouse** ▪ Computation of Data Cubes, modeling ▪ OLAP data, OLAP queries ▪ Data Warehouse back end tools ▪ Tuning and testing of Data Warehouse	4	
5	**Data Mining techniques** ▪ Data Mining definition and Task ▪ KDD versus Data Mining ▪ Data Mining techniques, tools and application	4	
6	**Data mining query languages** ▪ Data mining query languages ▪ Data specification, specifying knowledge, hierarchy specification, pattern presentation & visualization specification ▪ Data mining languages and standardization of data mining	5	
7	**Association analysis** ▪ Association Rule Mining (Market basket analysis) ▪ Why Association Mining is necessary? ▪ Pros and Cons of Association Rules ▪ Apriori Algorithm	6	
8	**Cluster analysis, Classification and Predication** ▪ What is classification? What is predication?	7	

	▪ Issues regarding classification and prediction (Preparing the data for classification and prediction, Comparing classification methods) ▪ Classification by decision tree induction (Extracting classification rules from decision trees) ▪ Bayesian Classification ▪ Classification by back propagation ▪ Introduction to Regression (Types of Regression) ▪ Clustering Algorithm (K-mean and K-Mediod Algorithms)		
9	**Advanced concepts in data mining** ▪ Mining Text Databases ▪ Mining the World Wide Web ▪ Mining Multimedia and Spatial Databases	4	

Laboratory:

1. Creating a simple data warehouse

2. Concepts of data cleaning and preparing for operation

3. Implementing classification and clustering algorithms in any programming language

4. Association rule mining though data mining tools

5. Data Classification through data mining tools

6. Clustering through data mining tools

7. Data visualization through data mining tools

Text Books:

1. Data Mining Concepts and Techniques, Morgan Kaufmann J. Han, M Kamber Second Edition ISBN: 978-1-55860-901-3

2. Data Warehousing in the Real World – Sam Anahory and Dennis Murray, Pearson Edition Asia.

References:

1. Data Mining Techniques – Arun K Pujari, University Press.

2. Data Mining- Pieter Adriaans, Dolf Zantinge

3. Data Mining, Alex Berson,Stephen Smith,Korth Theorling,TMH.

4. Data Mining, Adriaans, Addison-Wesley Longman.

Model Question

Full marks: 60
Pass marks: 24
Time: 3 hours.

Bachelor Level/ Fourth Year/Eight Semester/Science

Data Warehousing and Data Mining (CSC-451**)**

Candidates are required to give their answers in their own words as far as practicable. The figures in the margin indicate full marks.

Group-A

Long Answer Questions (Attempt any Two questions) [2x10=20]

1. Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg-grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg-grade measure stores the actual course grade of the student. At higher conceptual levels, avg-grade stores the average grade for the given combination.
   a) Draw a snowflake schema diagram for the data warehouse.
   b) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University Student.
   c) If each dimension has five levels (including all), such as “student < major < status < university < all”, how many cuboids will this cube contain (including the base and apex cuboids)?
2. A= {A1, A2, A3, A4, A5, A6}, Assume $\sigma = 35\%$. Use A priori algorithm to get the desired solution.

A1	**A2**	**A3**	**A4**	**A5**	**A6**
0	0	0	1	1	1
0	1	1	1	0	0
1	0	0	1	1	1
1	1	0	1	0	0
1	0	1	0	1	1
0	1	1	1	0	1
0	0	0	1	1	0

0	1	0	1	0	1
1	0	0	1	0	0
1	1	1	1	1	1

3. What kind of data preprocessing do we need before applying data mining algorithm to any data set. Explain binning method to handle noisy data with example.

Group- B

Short Answer Questions (Attempt any Eight questions) [8x5=40]
Question number 13 is compulsory.

4. Explain the use of frequent item set generation process. [5]
5. Differentiate between data marts and data cubes. [5]
6. Explain OLAP operations with example? [5]
7. List the drawbacks of ID3 algorithm with over-fitting and its remedy techniques [5]
8. Write the algorithm for K-means clustering. Compare it with k-nearest neighbor algorithm. [5]
9. What is text mining? Explain the text indexing techniques. [5]
10. Describe genetic algorithm using as problem solving technique in data mining. [5]
11. What do you mean by WWW mining? Explain WWW mining techniques. [5]
12. What is DMQL? How do you define Star Schema using DMQL? [5]
13. Write short notes (Any Two) [2x2.5=5]

a) Text Database Mining
b) Back propagation Algorithm
c) Regression
d) HOLAP

*****