Program Name : **Computer Engineering Program Group**

Program Code : **CO/CM/CW**

Semester : **Sixth**

Course Title : **Data Warehousing with Mining Techniques**

Course Code : **22621**

1. RATIONALE

Data mining and warehousing are the essential components of decision support systems for the modern days in industry and business. These techniques enable students to take better and faster decisions. The objective of this course is to introduce students to various Data Mining and Data Warehousing concepts and techniques. This course introduce principles, algorithm, architecture, design and implementation of data mining and data warehousing techniques. Learning this course would improve the employment potential of students in the information management sector.

2. COMPETENCY

The aim of this course is to help the student develop required skills so that they are able to acquire following competency:

- **Use Data mining techniques for data analysis to maintain Data warehouse.**

3. COURSE OUTCOMES (COs)

The theory, practical experiences and relevant soft skills associated with this course are to be taught and implemented, so that the student demonstrates the following ***industry oriented*** COs associated with the above mentioned competency:

a) Establish scope and necessity of Data Mining for various applications.
b) Establish scope and necessity of Data warehouse for various applications.
c) Use concept of data mining components and techniques in designing data mining systems.
d) Use data mining tools for different applications.
e) Apply basic Statistical calculations on Data.

4. TEACHING AND EXAMINATION SCHEME

Teaching Scheme			Credit (L+T+P)	Examination Scheme											
				Theory							Practical				
L	T	P		Paper Hrs.	ESE		PA		Total		ESE		PA		Total	
					Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min
3	-	2	5	3	70	28	30*	00	100	40	25@	10	25	10	50	20

(*): Under the theory PA, Out of 30 marks, 10 marks are for micro-project assessment to facilitate integration of COs and the remaining 20 marks is the average of 2 tests to be taken during the semester for the assessment of the UOs required for the attainment of the COs.
***Legends: L**-Lecture; **T** – Tutorial/Teacher Guided Theory Practice; **P** - Practical; **C** – Credit, **ESE** - End Semester Examination; **PA** - Progressive Assessment*

5. COURSE MAP (with sample COs, PrOs, UOs, ADOs and topics)

This course map illustrates an overview of the flow and linkages of the topics at various levels of outcomes (details in subsequent sections) to be attained by the student by the end of the

course, in all domains of learning in terms of the industry/employer identified competency depicted at the centre of this map.

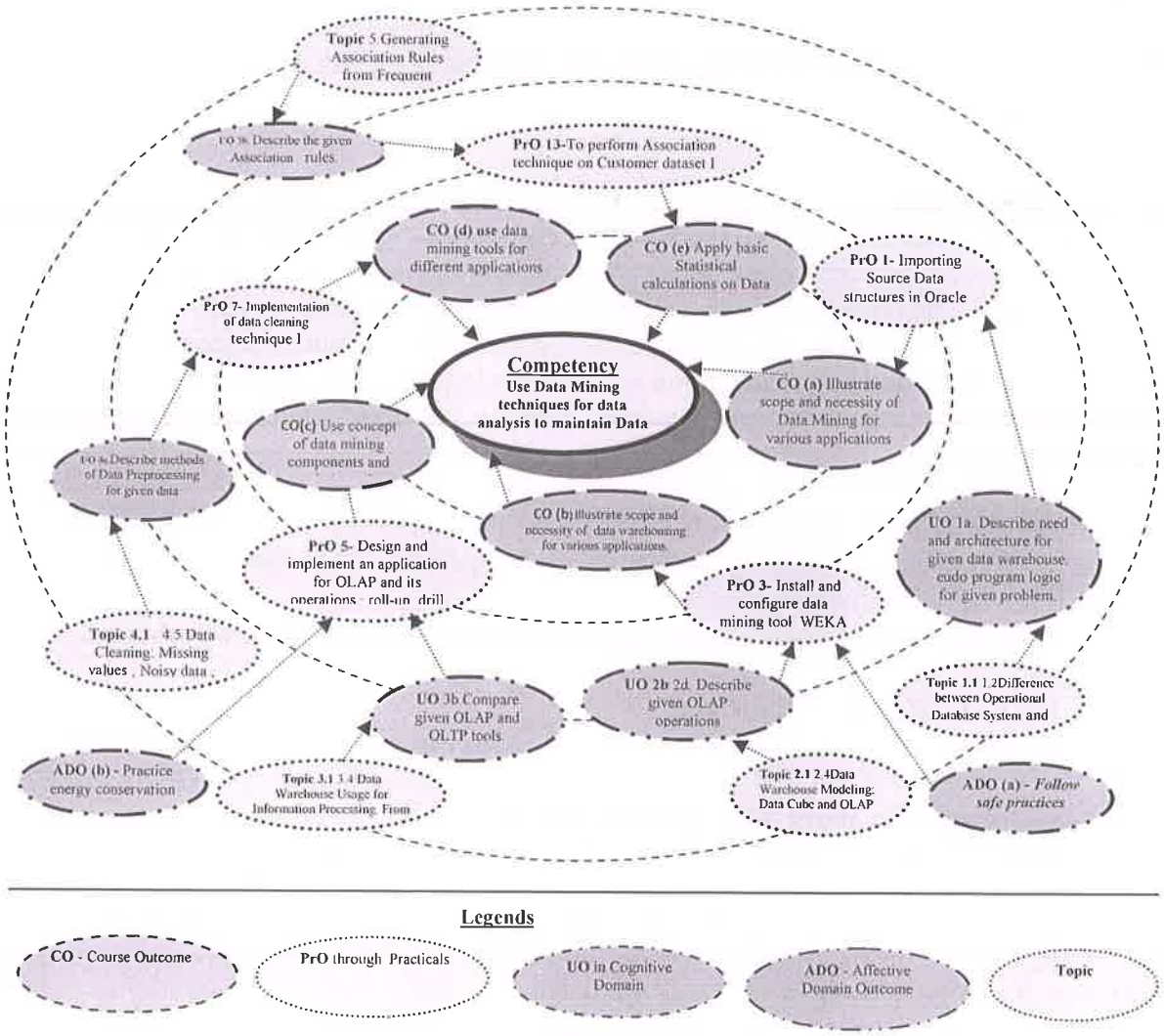

Figure 1 - Course Map

6. SUGGESTED PRACTICALS/ EXERCISES

The practicals in this section are PrOs (i.e. sub-components of the COs) to be developed and assessed in the student for the attainment of the competency.

S. No.	Practical Outcomes (PrOs)	Unit No.	Approx. Hrs. Required
1	Install Oracle Database Server and client.	I	02
2	Import Source Data structures in Oracle	I	02
3	Develop Target Data structures in Oracle	II	02
4	Install data mining tool WEKA. Study the GUI explorer on WEKA	II	02
5	Develop an application for OLAP and its operations: roll-up, drill down.	III	02
6	Develop an application for OLAP and its operations: Slice and dice.	III	02
7	Implement data cleaning technique I (Data Preprocessing --Finding and replacing Missing value in sample Dataset.)	IV	02

S. No.	Practical Outcomes (PrOs)	Unit No.	Approx. Hrs. Required
8	Implement data cleaning technique II (Data Transformation - Transforming data from one format to another format on sample data set)	IV	02
9	Preprocess dataset WEATHER.arff including creating an ARFF file and reading it into WEKA, and using the WEKA Explorer. Part - I	IV	02
10	Preprocess dataset WEATHER.arff including creating an ARFF file and reading it into WEKA, and using the WEKA Explorer. Part - II	IV	02
11	Demonstration of preprocessing on dataset Customer.arff ~~includes creating an ARFF file and reading it into WEKA, and using the WEKA Explorer.~~ Attributes Selection and Normalization.	IV	02
12	Demonstration of preprocessing on dataset Customer.arff ~~includes creating an ARFF file and reading it into WEKA, and using the WEKA Explorer.~~ Draw various graphs using WEKA	IV	02
13	Perform Association technique on Customer dataset I. (Implementing Apiori algorithm on customer dataset.)	V	02
14	Perform Association technique on Customer dataset II. (Using classification algorithm of KNN on sample dataset)	V	02
15	Apply clustering technique on Customer dataset I. (Using K-means clustering on sample customer dataset.)	V	02
16	Apply clustering technique on Customer dataset II. (Using K-means clustering on sample weather dataset)	V	02
	Total		**32**

Note

i. *A suggestive list of PrOs is given in the above table. More such PrOs can be added to attain the COs and competency. All the above listed practical need to be performed compulsorily, so that the student reaches the 'Applying Level' of Blooms's 'Cognitive Domain Taxonomy' as generally required by the industry.*

ii. *The 'Process' and 'Product' related skills associated with each PrO are to be assessed according to a suggested sample given below:*

S. No.	Performance Indicators	Weightage in %
1	Correctness of implementation of algorithm	40
2	Analysis and implementation ability	20
3	Quality of input and output displayed (messaging and formatting)	10
4	Answer to sample questions	20
5	Submit report in time	10
Total		**100**

The above PrOs also comprise of the following social skills/attitudes which are Affective Domain Outcomes (ADOs) that are best developed through the laboratory/field based experiences:

a) Work collaboratively in team
b) Follow ethical practices.

The ADOs are not specific to any one PrO, but are embedded in many PrOs. Hence, the acquisition of the ADOs takes place gradually in the student when s/he undertakes a series of practical experiences over a period of time. Moreover, the level of achievement of the ADOs according to Krathwohl's 'Affective Domain Taxonomy' should gradually increase as planned below:

- 'Valuing Level' in 1st year.
- 'Organization Level' in 2nd year.
- 'Characterization Level' in 3rd year.

7. MAJOR EQUIPMENT/ INSTRUMENTS REQUIRED

The major equipment with broad specification mentioned here will usher in uniformity in conduct of practicals, as well as aid to procure equipment by authorities concerned.

S. No.	Equipment Name with Broad Specifications	PrO. S. No.
	Computer system (Any computer system with basic configuration)	All
	Oracle Client and server	
	Data Mining tool : WEKA	

8. UNDERPINNING THEORY COMPONENTS

The following topics/subtopics should be taught and assessed to develop UOs in cognitive domain for achieving the COs to attain the identified competency. More UOs could be added.

Unit	Unit Outcomes (UOs) (in cognitive domain)	Topics and Sub-topics
Unit – I Introduction to Data Warehousing	1a. Describe need and architecture for the given data warehouse. 1b. Explain the benefits of data warehousing of the given application. 1c. Describe the given Data warehouse Models. 1d. Describe Extraction, Transformation and Loading for the given data warehouse 1e. Describe Metadata Repository for the given data warehouse.	1.1 Data warehousing, Difference between Operational Database System and Data warehouse. 1.2 Need for data warehousing. 1.3 A Multi tiered Architecture of data warehousing. 1.4 Data Warehouse Models: Enterprise Warehouse, Data Mart, and Virtual Warehouse. 1.5 Extraction, Transformation, and Loading. 1.6 Metadata Repository. 1.7 Benefits of Data warehousing.

Unit	Unit Outcomes (UOs) (in cognitive domain)	Topics and Sub-topics
Unit– II Data Warehouse Modeling and Online Analytical Processing I	2a. Describe Data Cube and OLAP for the given data warehouse. 2b. Explain Schemas for Multidimensional data models for the given data warehouse. 2c. Compare Stars, Snowflakes and Schema models for the given data warehouse on the basis of the given criteria. 2d. Describe the given OLAP operations 2e. Explain the benefits of the given OLAP tool.	2.1 Data Warehouse Modeling: Data Cube and OLAP, Data Cube: A Multidimensional Data Model. 2.2 Stars, Snowflakes, and Fact Constellations. 2.3 OLAP : Need of OLAP, OLAP Guidelines 2.4 Typical OLAP Operations
Unit– III Data Warehouse Designing and Online Analytical Processing II	3a. Describe design Process for the given data warehouse. 3b. Compare the given OLAP and OLTP tools, based on the given criteria. 3c. Design the given Data warehouse. 3d. Explain Bitmap and Join Index for the given OLAP. 3e. Compare OLAP server Architectures for the given data warehouse.	3.1 Data Warehouse Design and Usage. 3.2 A Business Analysis Framework for Data Warehouse Design. 3.3 Data Warehouse Design Process 3.4 Data Warehouse Usage for Information Processing. From Online Analytical Processing to Multi-dimensional Data Mining 3.5 Data Warehouse Implementation- Efficient Data Cube Computation: An Overview. 3.6 Indexing OLAP Data: Bitmap Index and Join Index, Efficient Processing of OLAP Queries 3.7 OLAP Server Architectures: ROLAP Versus MOLAP versus HOLAP
Unit-IV Introduction to Data Mining	4a. Explain concept of Data Mining. 4b. Describe the given data mining steps 4c. Explain Major issues for the given data. 4d. Explain the given data objects and attributes types. 4e. Describe methods of Data Preprocessing for the given data. 4f. Explain data cleaning process for the given data.	4.1 Introduction to Data Mining: Mining Steps in the process of knowledge discovery of Database (KDD) . 4.2 What Kind of data can be mined? Major issues in data mining. 4.3 Data Objects and Attributes types. 4.4 Data Preprocessing: Why Preprocess the data? Major Tasks in Data Preprocessing. 4.5 Data Cleaning: Missing values , Noisy data , Data cleaning as a process.

Unit	Unit Outcomes (UOs) (in cognitive domain)	Topics and Sub-topics
Unit –V Mining Frequent Patterns and Cluster Analysis	5a. Define the Itemsets for the given data. 5b. Describe the given Association Rules. 5c. Explain clustering methods for the given data 5d. Analyze Apriori Algorithm for the given data.	5.1 Mining Frequent Patterns: Basic Concepts: Market Basket Analysis, Frequent Itemsets, Closed Itemsets, and Association Rules 5.2 Frequent Itemsets Mining Methods: The Apriori Algorithm, Finding Frequent Itemsets Using Candidate Generation. 5.3 Generating Association Rules from Frequent Itemsets. 5.4 What is Cluster Analysis? Requirements for Cluster Analysis 5.5 Overview of Basic Clustering Methods. 5.6 General Applications of Clustering.

Note: To attain the COs and competency, above listed UOs need to be undertaken to achieve the 'Application Level' of Bloom's 'Cognitive Domain Taxonomy'

9. SUGGESTED SPECIFICATION TABLE FOR QUESTION PAPER DESIGN

Unit No.	Unit Title	Teaching Hours	Distribution of Theory Marks			
			R Level	U Level	A Level	Total Marks
I	Introduction to Data Warehousing	06	02	02	04	08
II	Data Warehouse Modeling and Online Analytical Processing	10	02	04	06	12
III	Data Warehouse Designing and Online Analytical Processing	10	04	06	08	18
IV	Introduction to Data Mining	12	02	08	08	18
V	Mining Frequent Patterns and Cluster Analysis	10	02	04	08	14
Total		**48**	**12**	**24**	**34**	**70**

Legends: *R=Remember, U=Understand, A=Apply and above (Bloom's Revised taxonomy)*
Note*: This specification table provides general guidelines to assist student for their learning and to teachers to teach and assess students with respect to attainment of UOs. The actual distribution of marks at different taxonomy levels (of R, U and A) in the question paper may vary from above table.*

10. SUGGESTED STUDENT ACTIVITIES

Other than the classroom and laboratory learning, following are the suggested student-related ***co-curricular*** activities which can be undertaken to accelerate the attainment of the various outcomes in this course: Students should conduct following activities in group and prepare reports of about 5 pages for each activity, also collect/record physical evidences for their (student's) portfolio which will be useful for their placement interviews:

a) Prepare journal of practicals.
b) Undertake micro-projects.

11. SUGGESTED SPECIAL INSTRUCTIONAL STRATEGIES (if any)

These are sample strategies, which the teacher can use to accelerate the attainment of the various learning outcomes in this course:

a) Massive open online courses (***MOOCs***) may be used to teach various topics/sub topics.

b) ***'L' in item No. 4*** does not mean only the traditional lecture method, but different types of teaching methods and media that are to be employed to develop the outcomes.

c) About ***15-20% of the topics/sub-topics*** which is relatively simpler or descriptive in nature is to be given to the students for ***self-directed learning*** and assess the development of the COs through classroom presentations (see implementation guideline for details).

d) With respect to item No.10, teachers need to ensure to create opportunities and provisions for ***co-curricular activities***.

e) Guide student(s) in undertaking micro-projects.

f) Demonstrate students thoroughly before they start doing the practice.

g) Encourage students to refer different websites to have deeper understanding of the subject.

h) Observe continuously and monitor the performance of students in Lab.

12. SUGGESTED MICRO-PROJECTS

Only one micro-project is planned to be undertaken by a student that needs to be assigned to him/her in the beginning of the semester. In the first four semesters, the micro-project are group-based. However, in the fifth and sixth semesters, it should be preferably be ***individually*** undertaken to build up the skill and confidence in every student to become problem solver so that s/he contributes to the projects of the industry. In special situations where groups have to be formed for micro-projects, the number of students in the group should ***not exceed three.***

The micro-project could be industry application based, internet-based, workshop-based, laboratory-based or field-based. Each micro-project should encompass two or more COs which are in fact, an integration of PrOs, UOs and ADOs. Each student will have to maintain dated work diary consisting of individual contribution in the project work and give a seminar presentation of it before submission. The total duration of the micro-project should not be less than ***16 (sixteen) student engagement hours*** during the course. The student ought to submit micro-project by the end of the semester to develop the industry oriented COs.

A suggestive list of micro-projects is given here. Similar micro-projects could be added by the concerned faculty:

a) Perform Association technique on Customer dataset /Agriculture dataset /

b) Weather dataset.

c) Create the data warehouse for any medical shop having 2 or more branches.

d) Predict traffic conditions for allocating more buses on various routes by bus controller.

e) Predict Job opportunities in Computer /IT field looking into the work generated last year.

f) Design a data mart or data warehouse for any organization.

13. SUGGESTED LEARNING RESOURCES

S. No.	Title of Book	Author	Publication
1	Data mining concepts and techniques	Han, Jiawei and Micheline Kamber.	Morgan Kaufmann Publications, Elsevier, 2012, ISBN: 978-0123814791
2	Data warehousing, data mining and OLAP	Berson, Alex	McGraw Hill New Delhi 2008. ISBN-13: 978-0070062726

S. No.	Title of Book	Author	Publication
3	The Data warehouse life cycle tool Kit	Kimball, .Ralph	John Wiley Third Edition ISBN: 978-0-471-20024-6
4	Data Based Management	Dr. Rajedra Kawle	Devraj Publication, ISBN- 978-93-86492-00-5

14. SOFTWARE/LEARNING WEBSITES

a) https://docs.oracle.com/
b) https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/
c) https://www.guru99.com/online-analytical-processing.html
d) https://www.tutorialspoint.com/dwh/dwh_relational_olap.htm
e) https://www.tutorialride.com/big-data-analytics/stream-cluster-analysis.htm