GUJARAT TECHNOLOGICAL UNIVERSITY, AHMEDABAD, GUJARAT

COURSE CURRICULUM
COURSE TITLE: DATA MINING AND WAREHOUSING
(COURSE CODE: 3361604)

Diploma Program in which this course is offered	Semester in which offered
Information Technology	SIXTH

1. RATIONALE

Data mining and warehousing are the essential components of decision support systems for the modern day industry and business. These techniques enable the knowledge worker to make better and faster decisions. The objective of this course is to introduce the student to various Data Mining and Data Warehousing concepts and techniques. A database perspective of an open source application is used throughout the course to 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 course content should be taught and implemented with the aim to develop required skills in the students so that they are able to acquire following competencies:

- **Apply techniques, data pre-processing, OLAP of data mining and warehousing using open source tools.**

3. COURSE OUTCOMES:

The theory should be taught and practical should be carried out in such a manner that students are able to acquire different learning out comes in cognitive, psychomotor and affective domain to demonstrate following course outcomes.

- Describe the concept of Data Mining & its attributes
- Apply the concept of data mining components and techniques in designing data mining systems.
- Solve basic Statistical calculations on Data
- Describe the aspect of data pre-processing
- Explain the concept of Data Cleaning & Integration
- Explain decision Trees and clustering
- Install and Configure WEKA Tool
- Demonstrate WEKA Explorer, Mining techniques and Attribute Relation File Format (ARFF).
- Compare various Data Mining techniques available in WEKA

4. TEACHING AND EXAMINATION SCHEME

Teaching Scheme (In Hours)			Total Credits (L+T+P)	Examination Scheme				
				Theory Marks		Practical Marks		Total Marks
L	**T**	**P**	**C**	**ESE**	**PA**	**ESE**	**PA**	**200**
3	0	4	7	70	30	40	60	

Legends: L - Lecture; **T -** Tutorial/Teacher Guided Student Activity; **P -** Practical; **C -** Credit; **ESE** - End Semester Examination; **PA** - Progressive Assessment

5. COURSE CONTENT DETAILS

Unit	Major Learning Outcomes (in cognitive domain)	Topics and Sub-topics
Unit – I **Fundamentals of data mining**	1a. Describe the concept of Data Mining	1.1 Data mining: History, strategies, techniques, applications, challenges of data mining, Future of data mining
	1b. Describe types of Data	1.2 Types of Data 1.2.1 Database Data 1.2.2 Data Warehouses 1.2.3 Transactional Data 1.2.4 Other Kinds of Data
Unit – II **Objects, Attributes, & Statistical Description of Data**	2a. Explain Mining techniques and Attribute Relation File Format (ARFF).	2.1 Data Attribute 2.1.1 Nominal Attributes 2.1.2 Binary Attributes 2.1.3 Ordinal Attributes 2.1.4 Numeric Attributes 2.1.5 Discrete versus Continuous Attributes
	2b. Solve basic Statistical calculations on Data	2.2 Mean, Median, and Mode 2.3 Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range using WEKA
Unit – III **Data Preprocessing**	3a. Describe the aspect of data preprocessing	3.1 Preprocess the Data 3.2 Major Tasks in Data Preprocessing
	3b. Explain the concept of Data Cleaning & Integration	3.2 Data Cleaning 3.2.1 Missing Values 3.2.2 Noisy Data 3.2.3 Data Cleaning as a Process 3.3 Data Integration 3.3.1 Entity Identification Problem 3.3.2 Redundancy and Correlation Analysis 3.3.3 Tuple Duplication 3.3.4 Data Value Conflict Detection and

Unit	Major Learning Outcomes (in cognitive domain)	Topics and Sub-topics
		Resolution 3.3.5 Use WEKA for cleaning and integration
Unit – IV **Classification**	4. Explain decision Trees and clustering	4.1 Decision tree: ID3 4.2 Probability based solving 4.3 Concepts of Clustering 4.4 Using WEKA for classification and clustering
Unit - V **Data Warehouse & OLAP Technology**	5a. Apply the concept of Data Ware housing using WEKA solution	5.1 Data Warehouse 5.2 Differences between Operational Database Systems and Data Warehouses 5.3 Enterprise Warehouse, Data Mart, and Virtual Warehouse
Unit - VI **Data Mining Tool: WEKA**	6. Install and Configure WEKA Tool	6.1 Basic of WEKA 6.1 Installing WEKA 6.2 WEKA data file format 6.3 Data visualization in WEKA 6.4 Data filtering 6.5 Using the concepts of data mining with WEKA

6. SUGGESTED SPECIFICATION TABLE WITH HOURS & MARKS (THEORY)

Unit No.	Unit Title	Teaching Hours	Distribution of Theory Marks			
			R Level	U Level	A Level	Total Marks
I	Fundamentals of data mining	4	4	4	2	10
II	Objects, Attributes, & Statistical Description of Data	8	4	6	4	14
III	Data Preprocessing	9	4	6	4	14
IV	Classification	8	2	4	4	10
V	Data Warehouse & OLAP Technology	8	4	4	4	12
VI	Data Mining Tool: WEKA	5	2	3	5	10
	Total	**42**	**20**	**27**	**23**	**70**

Legends: R = Remember; U = Understand; A = Apply and above levels (Bloom's revised 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.

7. SUGGESTED LIST OF EXERCISES/PRACTICAL

The practical/exercises should be properly designed and implemented with an attempt to develop different types of skills **(outcomes in psychomotor and affective domain)** so that students are able to acquire the competencies/programme outcomes. Following is the list of practical exercises for guidance.

***Note**: Here only outcomes in psychomotor domain are listed as practical/exercises. However, if these practical/exercises are completed appropriately, they would also lead to development of certain outcomes in affective domain which would in turn lead to development of **Course Outcomes** related to affective domain. Thus over all development of **Programme Outcomes (**as given in a common list at the beginning of curriculum document for this programme) would be assured.*

Faculty should refer to that common list and should ensure that students also acquire outcomes in affective domain which are required for overall achievement of Programme Outcomes/Course Outcomes.

S. No.	**UNIT**	**Practical Exercises** (Outcomes in Psychomotor Domain)	**Approx Hours. Required**
1.	II	Demonstrate the use of ARFF files taking input and diplay the output of the files.	2
2.	II	Create your own excel file. Convert the excel file to .csv format and prepare it as ARFF files.	2
3.	III	Preprocess and classify Customer dataset. http://archive.ics.uci.edu/ml/	4
4.	III	Perform Preprocessing, Classification techniques on Agriculture dataset. (http://archive.ics.uci.edu/ml/)	4
5.	III	Preprocess and classify Weather dataset. http://archive.ics.uci.edu/ml/	4
6.	III	Perform data Cleansing of customer dataset. http://archive.ics.uci.edu/ml/ , www.kdnuggets.com/**datasets/**	4
7.	IV	Perform Clustering technique on Customer dataset. http://archive.ics.uci.edu/ml/	2
8.	IV	Perform Clustering technique on Agriculture dataset. http://archive.ics.uci.edu/ml/	2
9.	IV	Perform Clustering technique on Weather dataset. http://archive.ics.uci.edu/ml/	2
10.	IV	Classify the dataset using decision tree. www.kdnuggets.com/**datasets/**	6
11.	V	Perform Association technique on Customer dataset. http://archive.ics.uci.edu/ml/, www.kdnuggets.com/**datasets/**	2
12.	V	Perform Association technique on Agriculture dataset. http://archive.ics.uci.edu/ml/, www.kdnuggets.com/**datasets/**	2
13.	V	Perform Association technique on Weather dataset.	2

S. No.	UNIT	**Practical Exercises** (Outcomes in Psychomotor Domain)	**Approx Hours. Required**
14.	VI	Compare various Data Mining techniques available in WEKA	6
15.	VI	Apply filters on the customer dataset using WEKA.	2
16.	VI	Install and Configure WEKA Tool	6
17.	VI	Demonstration of Weka Explorer, Mining techniques and Attribute Relation File Format (ARFF). http://archive.ics.uci.edu/ml/	4
		Total Practical Hours	56

Practical Examination can be conducted based on one of the Data mining dataset given at http://archive.ics.uci.edu/ml/, www.kdnuggets.com/**datasets**/. Viva can be conducted based on the understanding of various classification, clustering, warehousing and data mining techniques

8. SUGGESTED LIST OF STUDENT ACTIVITIES

i. Student should do as much practice as possible on related software to develop the mastery.

ii. Students in groups should visit different business organisation where data mining and warehousing is done and should study the methods and software in use. Moreover each group should study that for what purpose data mining is carried out and how mined data is used. All groups should prepare reports on their study and present in class. These presentations should generate group discussions.

iii. Search the net and find out different data mining and warehousing techniques and software being used.

9. SPECIAL INSTRUCTIONAL STRATEGIES (if any)

i. Concepts should be introduced in classroom input sessions and by giving demonstration through projector.

ii. Arrange expert lectures by IT experts working professionally in the area of data mining and warehousing.

iii. More focus should be given on practical work which will be carried out in laboratory sessions. If possible some theory sessions may be conducted in labs so that theory and practice can go hand in hand.

iv. Faculty should allow students to use their creativity and let them struggle to learn on their own during practical sessions. However, faculty should remain around the students and should help them when they are stuck.

v. Custom excel dataset can be created which can be used for data mining.

10. SUGGESTED LEARNING RESOURCES

A) List of Books

Sr. No.	Title of Book	Author	Publication
1	Data Mining Concepts and Techniques	Jiawei Han and Micheline Kamber	Kaufmann Publishers, 2011
2	Data Mining Techniques	Arun K Pujari	Orient Longman Publishers
3	Fundamentals of Data Warehouses	M.Jarke, M Lenzerni	
4	Principles of Data Mining	David Hand, Heikki Mannila, Padhraic Smyth,	PHI
5	Data Mining:Methods and Techniques	A B M Shawkat Ali, Saleh A, Wasimi	CENGAGE Learning

B) List of Major Equipment/ Instrument with Broad Specifications

Latest computers in sufficient numbers

C) List of Software/Learning Websites

1. **WEKA**: WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation.

   WEKA is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection.

2. **XLMiner**: XLMiner is a comprehensive data mining add-in for Excel. XLMiner can be used to mine data available in Excel worksheets. It includes capabilities that allow a miner to work with partitioning, neural networks, classification and regression trees, association rules, nearest neighbors, etc. With is ease of use and learning, XLMiner serves to be the perfect candidate tool to wet your feet in Data Mining as a novice miner. http://dataminingtools.net

   XLMiner can work with large data sets which may exceed the limits in Excel. A standard procedure is to sample data from a larger database, bring it into Excel to fit a model, and, in the case of supervised learning routines, score output back out to the database. In the standard edition of XLMiner, this feature is supported for Oracle, SQL Server and Access databases.

3. Data Mining Tutorial http://www.tutorialspoint.com/data_mining/

11. COURSE CURRICULUM DEVELOPMENT COMMITTEE

Faculty Members from Polytechnics

- **Prof. Parvez Faruki,** I/C Head, Information Technology, Sir BPTI, Bhavnagar.
- **Prof. Darshan M. Tank**, In-charge Head of Department, Information Technology, Lukhdhirji Engineering College (Diploma), Morbi
- **Prof. Hardik Patel,** Lecturer, Information Technology Dept, BPTI, Bhavnagar.

Coordinator and Faculty Members from NITTTR Bhopal

- **Dr. K. James Mathai,** Associate Professor, Dept. of Computer Engineering and Applications.
- **Prof. Priyanka Tripathi,** Associate Professor, Dept. of Computer Engineering and Applications.