Year: B. Tech III (Semester V)

Subject Name: Data Mining and Data Warehousing **Subject Code:** BTCO14505
Type of course: Professional Elective I
Prerequisite: Database Management System

List of Courses where this course will be prerequisite: Big data analytics

Rationale: This course is of prime importance in Computer science and engineering. The knowledge inferred from huge amounts of historical data can be used for betterment of human lives in many ways. This course introduces the data pre-processing, various data mining methods and various applications of data mining.

Teaching and Examination Scheme:

Teaching Scheme				Theory Marks			Practical Marks		Total
L	T	P	C	TEE	CA1	CA2	TEP	CA3	
3	0	2	4	60	25	15	30	20	150

CA1: Continuous Assessment (assignments/projects/open book tests/closed book tests CA2: Sincerity in attending classes/class tests/ timely submissions of assignments/self-learning attitude/solving advanced problems TEE: Term End Examination TEP: Term End Practical Exam (Performance and viva on practical skills learned in course) CA3: Regular submission of Lab work/Quality of work submitted/Active participation in lab sessions/viva on practical skills learned in course

Contents:

Unit No	Contents	Total Hrs
1	**Data warehousing :** Basic Concepts – Data warehouse basics, Differences between Operational Database Systems and Data Warehouses – Online Transaction Processing, Online Analytical Processing, OLAP vs OLTP, Need for separate Data Warehouse, A Multi-tiered Architecture of Data Warehouse, Data warehouse models – Enterprise warehouse, Data Mart, Virtual Warehouse, Data Warehouse vs Data Mart Data Warehouse Modeling - Data Cube: A Multidimensional Data Model, fact table, dimension table, schemas for Multidimensional Data Models - stars, snowflakes and fact constellations schemas, Concept hierarchies, Categorization and computation of measure, OLAP operations - drill-down, roll-up, slice, dice and pivoting operations	08
2	**Introduction to Data Mining :** Basic Concepts -Data Mining basics, Motivation for Data Mining – Knowledge Discovery from Data (KDD) process, Kinds of data that can be mined, Data Mining functionalities - Kinds of patterns that can be mined, Pattern interestingness, Major issues in data mining	04

3	**Data Pre-processing :** An overview of Data Pre-processing, Need of pre-processing, Major task in Data pre-processing, Data cleaning – missing values, noisy data, data integration – entity identification, redundancy and correlation analysis, tuple duplication, data value conflict detection, data reduction – attribute subset selection, regression models, clustering, sampling, data transformation – smoothing, attribute construction, aggregation, normalization, discretization, concept hierarchy generation for nominal data	06
4	**Frequent Pattern and Association Rule Mining :** Basic concepts of frequent pattern mining, market basket analysis, concepts of support, confidence, minimum support threshold, minimum confidence threshold, frequent itemsets, closed frequent itemsets, maximal frequent itemsets, Apriori algorithm – Finding frequent itemsets, generating strong association rules, Improving efficiency of apriori algorithm, FP-Growth algorithm	08
5	**Classification and Prediction :** Basic concepts of classification, Classification as a two step process, Decision tree induction – attribute selection measures, tree pruning, Bayes classification methods - Bayes' theorem, Naive Bayesian classification, Rule based classification – IF-THEN rules for classification, Rule extraction from a decision tree, Classification by backpropagation, K-nearest neighbor classification, Classification model evaluation – confusion matrix, accuracy, error rate, sensitivity (recall), specificity, precision, F-score, Prediction: Linear and nonlinear regression	09
6	**Clustering :** Basic Concepts : Cluster and clustering basics, Requirement for clustering, Partitioning methods - K-means clustering, k-medoids clustering, Hierarchical methods - Agglomerative clustering method and divisive clustering method;	06
7	**Advance topics :** Text mining, Web mining – web content mining, web structure mining, web usage mining, privacy preserving data mining	04

Suggested Specification table with Marks (Theory): (For B.Tech only)

Distribution of Theory Marks					
R Level	U Level	A Level	N Level	E Level	C Level
5	20	20	10	5	0

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create and above Levels (Revised Bloom's Taxonomy)

Reference Books:

Sr No	Title of book /article	Author(s)	Publisher and details like ISBN	Year of publication / Publication Edition

1	Data Mining : Concepts and Techniques	Jiawei Han, Micheline Kamber, Jian Pei	Morgan Kaufmann	Latest Edition
2	Data mining : Concepts, Models, Methods and Algorithms	Mehmed Kantardzic	Wiley	
3	Introduction to Data Mining",	Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar	Pearson Education	
4	Web Data Mining : Exploring hyperlinks, contents and usage data	Bing Liu	Springer	

Course Outcomes (CO):

Sr.No.	CO statement	Marks % weightage
1	Attribute the basics of Data Mining and Data Warehousing	12%
2	Compare OLAP and OLTP techniques and generate various warehouse schemas	15%
3	Select appropriate pre-processing techniques to produce task relevant data	15%
4	Infer frequent patterns and association rules from given transactional data	18%
5	Construct classification and prediction model using classification and prediction methods respectively and verify the accuracy of these models	20%
6	Construct clustering model using clustering methods	10%
7	Utilize the concepts of various techniques such as text mining, web mining, privacy preserving data mining etc for various applications	10%

List of Open learning website:

1. NPTEL free course on Data Mining https://nptel.ac.in/courses/106/105/106105174/
2. https://www.javatpoint.com/data-mining
3. https://www.tutorialspoint.com/data_mining/index.htm
4. https://www.guru99.com/data-mining-tutorial.html

5. https://www.tutorialride.com/data-mining/data-mining-tutorial.htm

List of Open Source Software:

1. WEKA- Waikato Environment for Knowledge Analysis
2. Orange
3. Knime- KoNstanz Information MinEr
4. Rattle- R Analytical Tool To Learn Easily
5. ELKI- Environment for DeveLoping KDD-Applications Supported by Index-Structures

List of Experiments:

Sr.No	Practical
1	Build a Data Warehouse and explore the WEKA Tool.
2	**Pre-processing of missing values :** a) Replace the missing values for given automobile dataset "imports-85.data" with user specified global constant. b) Replace the missing values for given automobile dataset "imports-85.data" with mean, median and mode value of numeric attribute. c) Replace the missing values for the given automobile dataset "imports-85.data" with the mean value of each attribute class. (Consider no. of doors as the class attribute - 6th attribute) **Download Dataset From:** https://github.com/nyuvis/datasets/blob/master/auto/imports-85.data **Dataset Information :** https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names
3	**Preprocessing of Noisy Data :** Consider the following values for age attribute of total 21 records : 13, 52, 15, 16, 45, 19, 20, 21, 22, 25, 30, 33, 35, 36, 40, 46, 70, 16, 25, 22, 33 a) Implement smoothing by bin means to smooth these data, using a suitable bin depth. b) Implement smoothing by bin medians to smooth these data, using a suitable bin depth.

<table>
<tr><td></td><td>c) Implement smoothing by bin boundaries to smooth these data, using a suitable bin depth</td></tr>
<tr><td>4</td><td>Apply the Min-Max Normalization technique to following result data and normalize them in the range [0.0 to 100.0] . Prepare the merit list according to normalization result and assign merit numbers.
<table>
<tr><th colspan="2">University - 1</th><th colspan="2">University - 2</th></tr>
<tr><th>Candidate Name</th><th>Performance Index (out of 10)</th><th>Candidate Name</th><th>Performance Index (out of 4)</th></tr>
<tr><td>A</td><td>9.1</td><td>B</td><td>3.7</td></tr>
<tr><td>C</td><td>4.0</td><td>D</td><td>2.1</td></tr>
<tr><td>E</td><td>5.4</td><td>F</td><td>3.8</td></tr>
<tr><td>G</td><td>7.2</td><td>H</td><td>3.1</td></tr>
</table></td></tr>
<tr><td>5</td><td>Implement the z - score Normalization and Decimal Scaling Normalization technique for attribute age data. The data for tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
Normalize any one of above age values.</td></tr>
<tr><td>6</td><td>Perform data preprocessing tasks and demonstrate association rule mining on data sets using WEKA tool.</td></tr>
<tr><td>7</td><td>Implement Apriori algorithm for frequent itemset mining on small dataset.</td></tr>
<tr><td>8</td><td>Using real dataset, implement the Apriori algorithm to generate the frequent patterns and also to generate strong association rules.
Instructions:
1. The program should be generic and executable with 3 parameters: the name of the input dataset file, the threshold of minimum support count, and the name of the output file.
2. The program should generate an output file that contains all the frequent itemsets</td></tr>
</table>

	together with their support. The output file (sample output) should have the following format: each line contains a single frequent itemset as a list of items separated by whitespace. At the end of the line, its support is displayed between a pair of parenthesis. For example: 2 3 8 (5) represents an itemset containing items 2, 3 and 8 with a support count of 5. 3. Test your implementation on the dataset http://fimi.uantwerpen.be/data/retail.dat and measure execution time as well as number of frequent itemsets with various minimum support values. Detailed descriptions about the dataset can be found at http://fimi.uantwerpen.be/data/retail.pdf. Also try your program with various other frequent itemset mining datasets like: http://fimi.uantwerpen.be/data/
9	Demonstrate classification on data sets using WEKA tool.
10	Implement a decision tree algorithm for classification.
11	Demonstrate clustering on data sets using WEKA tool.
12	Demonstrate regression on data sets using WEKA tool.
13	Implement Data Aggregation operations listed below on the data of your choice on the Oracle Database using SQL queries: 1. ROLLUP EXTENSION to GROUP BY 2. ROLLDOWN EXTENSION to GROUP BY 3. CUBE EXTENSION to GROUP BY 4. GROUPING SETS.