**VI Semester**

| DATA MINING AND DATA WAREHOUSING | | | |
|---|---|---|---|
| Course Code | 21IS643 | CIE Marks | 50 |
| Teaching Hours/Week (L:T:P: S) | 3:0:0:0 | SEE Marks | 50 |
| Total Hours of Pedagogy | 40 | Total Marks | 100 |
| Credits | 3 | Exam Hours | 3 Hrs |

**Course Learning Objectives:**

CLO 1. Introduction to general issues of Data Warehouse and Data Mining.
CLO 2. Understanding of the different architectures and mining techniques
CLO 3. The role and functions of Data Warehouse and Data Mining
CLO 4. Explain the stages and process different data mining techniques.
CLO 5. Learn mining and warehouse techniques through the use of different tools

**Teaching-Learning Process (General Instructions)**

These are sample Strategies, which teachers can use to accelerate the attainment of the various course outcomes.

1. Lecturer method (L) need not to be only a traditional lecture method, but alternative effective teaching methods could be adopted to attain the outcomes.
2. Use of Video/Animation to explain functioning of various concepts.
3. Encourage collaborative (Group Learning) Learning in the class.
4. Ask at least three HOT (Higher order Thinking) questions in the class, which promotes critical thinking.
5. Adopt Problem Based Learning (PBL), which fosters students' Analytical skills, develop design thinking skills such as the ability to design, evaluate, generalize, and analyze information rather than simply recall it.
6. Introduce Topics in manifold representations.
7. Show the different ways to solve the same problem with different circuits/logic and encourage the students to come up with their own creative ways to solve them.
8. Discuss how every concept can be applied to the real world - and when that's possible, it helps improve the students' understanding.

**Module-1**

**Data Warehouse**: Introduction to Data Ware House, Differences between operational database systems and data Ware House, Data Ware House characteristics, Data Ware House Architecture and its components, Extraction-Transformation-Loading, Logical (Multi- Dimensional), Data Modeling, Schema Design, star and snow-Flake Schema, Fact Constellation, Fact Table, Fully Addictive, Semi-Addictive, Non-Addictive Measures; Fact Less-Facts, Dimension Table characteristics; Fact-Less-Facts, Dimension Table characteristics; OLAP cube, OLAP Operations, OLAP Server Architecture-ROLAP, MOLAP and HOLAP.

**Textbook 2: Ch.4.1,4.2**

| **Teaching-Learning Process** | Chalk and talk method, PowerPoint Presentation, Demonstration |
|---|---|

**Module-2**

**Introduction to Data Mining**: Introduction, what is Data Mining, Definition, KDD, Challenges, Data Mining Tasks, Data Preprocessing- Data Cleaning, Missing Data, Dimensionality Reduction, Feature Subset Selection, Discretization and Binarization, Data Transformation; Measures of similarity and Dissimilarity-Basics.

**Textbook 2: Ch.4.4**
**Textbook 1: Ch.1.1,1.2,1.4, 2.1 to 2.4**

| **Pedagogy:** | Chalk and talk method, PowerPoint Presentation, Demonstration |
|---|---|

**Module-3**

**Association Analysis:** Association Analysis: Problem Definition, Frequent Item set Generation, Rule generation. Alternative Methods for Generating Frequent Item sets, FPGrowth Algorithm, Evaluation of Association Patterns.

**Textbook 1: Ch 6.1,6.2,6.3, 6.5, 6.6 and 6.7**

| | |
|---|---|
| **Teaching-Learning Process** | Chalk and talk method, PowerPoint Presentation, Demonstration, Problem based learning |

**Module-4**

**Classification:** Decision Trees Induction, Method for Comparing Classifiers, Rule Based Classifiers, Nearest Neighbor Classifiers, Bayesian Classifiers.

**Textbook 1: Ch 4.3,4.6,5.1,5.2,5.3**

| | |
|---|---|
| **Teaching-Learning Process** | Chalk and talk method, Demonstration, Problem based learning |

**Module-5**

**Clustering Analysis:** Overview, K-Means, Agglomerative Hierarchical Clustering, DBSCAN, Cluster Evaluation, Density-Based Clustering, Graph-Based Clustering, Scalable Clustering Algorithms.

**Textbook 1: Ch 8.1 to 8.5, 9.3 to 9.5**

| | |
|---|---|
| **Teaching-Learning Process** | Chalk and talk method, Demonstration, Problem based learning |

**Course Outcomes:** At the end of the course students should be able to:

CO 1. Understand warehousing architectures and tools for systematically organizing large database and use their data to make strategic decisions.
CO 2. Apply KDD process for finding interesting pattern from warehouse.
CO 3. Analyze the kinds of patterns that can be discovered by association rule mining.
CO 4. Evaluate interesting patterns from large amounts of data to analyze for predictions and classification.
CO 5. Design select suitable methods for data mining and analysis.

**Assessment Details (both CIE and SEE)**

The weightage of Continuous Internal Evaluation (CIE) is 50% and for Semester End Exam (SEE) is 50%. The minimum passing mark for the CIE is 40% of the maximum marks (20 marks). A student shall be deemed to have satisfied the academic requirements and earned the credits allotted to each subject/ course if the student secures not less than 35% (18 Marks out of 50) in the semester-end examination (SEE), and a minimum of 40% (40 marks out of 100) in the sum total of the CIE (Continuous Internal Evaluation) and SEE (Semester End Examination) taken together

**Continuous Internal Evaluation:**

Three Unit Tests each of **20 Marks (duration 01 hour**)

1. First test at the end of 5th week of the semester
2. Second test at the end of the 10th week of the semester
3. Third test at the end of the 15th week of the semester

Two assignments each of **10 Marks**

4. First assignment at the end of 4th week of the semester
5. Second assignment at the end of 9th week of the semester

Group discussion/Seminar/quiz any one of three suitably planned to attain the COs and POs for **20 Marks** (**duration 01 hours**)

6. At the end of the 13th week of the semester

The sum of three tests, two assignments, and quiz/seminar/group discussion will be out of 100 marks and will be **scaled down to 50 marks**

(to have less stressed CIE, the portion of the syllabus should not be common /repeated for any of the methods of the CIE. Each method of CIE should have a different syllabus portion of the course).

**CIE methods /question paper has to be designed to attain the different levels of Bloom's taxonomy as per the outcome defined for the course.**

**Semester End Examination:**

Theory SEE will be conducted by University as per the scheduled timetable, with common question papers for the subject (**duration 03 hours)**

1. The question paper will have ten questions. Each question is set for 20 marks. Marks scored shall be proportionally reduced to 50 marks
2. There will be 2 questions from each module. Each of the two questions under a module (with a maximum

of 3 sub-questions), **should have a mix of topics** under that module.

The students have to answer 5 full questions, selecting one full question from each module

**Suggested Learning Resources:**

**Textbooks**

1. Data Mining-Concepts and Techniques- Jiawei Han, Micheline Kamber, Morgan Kaufmann Publishers, Elsevier, 2 Edition, 2006.
2. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Education.

**Reference Books:**

1. Data Mining Techniques, Arun K Pujari, 3rd Edition, Universities Press.
2. Data Ware Housing Fundamentals, Pualraj Ponnaiah, Wiley Student Edition.
3. The Data Ware House Life Cycle Toolkit- Ralph Kimball, Wiley Student Edition.
4. Data Mining, Vikaram Pudi, P Radha Krishna, Oxford University

**Web links and Video Lectures (e-Resources):**

1. https://nptel.ac.in/courses/106/106/106106093/
2. https://nptel.ac.in/courses/110/107/110107092/
3. https://nptel.ac.in/courses/106/105/106105174/
4. VTU e-Shikshana Program
5. VTU EDUSAT Program

**Activity-Based Learning (Suggested Activities in Class)/ Practical Based learning**

- Flip Class
- Seminar/Poster Presentation
- Role play/Team Demonstration/Collaborative Activity
- Mini Project
- Case study
- Learn by Doing