22DS1T4: DATA MINING TECHNIQUES

<table>
<tr><td><b>Course Name</b></td><td colspan="2">Data Mining Techniques</td><td><b>L</b></td><td><b>T</b></td><td><b>P</b></td><td><b>C</b></td><td><b>CIA</b></td><td><b>SEE</b></td><td><b>TM</b></td></tr>
<tr><td><b>Course Code</b></td><td colspan="2">22DS1T4</td><td>4</td><td>0</td><td>0</td><td>4</td><td>30</td><td>70</td><td>100</td></tr>
<tr><td><b>Year of Introduction:</b>
2020</td><td><b>Year of Offering:</b> 2022</td><td colspan="4"><b>Year of Revision</b>:
No Revision</td><td colspan="4"><b>Percentage of Revision:</b>
Nil</td></tr>
<tr><td colspan="10"><b>L</b>-Lecture, T-Tutorial, <b>P</b>-Practical, <b>C</b>-Credits, <b>CIA</b>-Internal Marks, <b>SEE</b>-External Marks, <b>TM</b>-Total Marks</td></tr>
</table>

Course Description and Purpose: Python Programming is a course that illustrates to *Data Mining Concepts*, *Data Preprocessing*, *Data Warehousing and Online Analytical Processing*, *Mining Frequent Patterns*, *Association and Correlation*, *Basic Concepts and Methods*, *Advanced Pattern Mining*, *Classification Basic and Advanced Methods*, *Clustering Analysis* and *Outlier Detection*.

Course Objectives:
This course will help enable the students to understand and learn Data Mining Techniques like *Data Preprocessing*, *Data Warehousing and Online Analytical Processing*, *Mining Frequent Patterns*, *Association and Correlations*, *Pattern Mining Techniques*, *Classification* and *Clustering Techniques.*

Specific objectives include:

- ✓ To understand *Fundamentals of Data Mining* & *Data Preprocessing*.
- ✓ To learn Data *Warehousing and Online Analytical Processing* concepts.
- ✓ To understand various *Mining Frequent Patterns Methods* & *Various Association Rules*.
- ✓ To lean different *Classification* & *Prediction* Methods.
- ✓ To *understand* & *apply* various Clustering Algorithms.

Course Learning Outcomes:
Upon successful completion of the course, the student will be able to:
CO1: Understand *Fundamentals of Data Mining* & *Data Preprocessing*.
CO2: Learn Data *Warehousing and Online Analytical Processing* concepts.
CO3: Understand various *Mining Frequent Patterns Methods* & *Various Association Rules*.
CO4: Lean different Classification & Prediction Methods.
CO5: U*nderstand* & *apply* various Clustering Algorithms.

UNIT I (12 Hours)

Introduction: What is Data mining **-** *What Kind of Data can be Mined* (Database Data, Data Warehouses Transactional Data, Other Kinds of Data) **-** *What kinds of Patterns can be Mined* (Class/Concept Description: Characterization and Discrimination, Mining Frequent Patterns, Associations and Correlations, Classification and Regression for Predictive Analysis , Cluster Analysis , Outlier Analysis, Are All Patterns Interesting?) **-** *Which Technologies are Used?* (Statistics, Machine Learning, Database Systems and Data Warehouses, Information Retrieval) **-** *Major Issues in Data Mining* (Mining Methodology User Interaction, Efficiency and Scalability, Diversity of Database Types, Data Mining and Society)
Data Preprocessing: *An Overview of Data Preprocessing* (Why Preprocess the Data?, Major Tasks in Data Preprocessing) **-** *Data Cleaning* (Missing Values, Noisy Data, Data Cleaning as a Process) **-** *Data Integration* (Entity Identification Problem, Redundancy and Correlation Analysis,Tuple Duplication, Data Value Conflict Detection and Resolution) **-** *Data Reduction* (Overview of Data Reduction Strategies, Attribute Subset Selection, Regression and Log Linear Models, Histograms, Sampling and Datacube Aggregation) **-** *Data Transformation* (Data Transformation strategies Overview, Data Transformation by Normalisation, Discretization by Binning).

UNIT II (12 Hours)

Data Warehousing and Online Analytical Processing: *Data Warehouse Basic Concepts* (What Is a Data Warehouse?, Difference between Operational Database Systems and Data Warehouses, Why have a separate Data warehouse?, Data Warehousing:A Multiered Architecture, Data Warehouse Models: Enterprise Warehouse, Data Mart and Virtual Warehouse, Extraction, Transformation and Loading, Metadata Repository, Datawarehouse Modeling:Datacube and OLAP, Data Cube: A Multidimensional Data Model, Stars, Snowflakes, and Fact Constellations Schemas for Multidimensional Data Models, Dimensions: The Role of Concept Hierarchies, Measures:Their Categorisation and Computation, Typical OLAP Operations, A Starnet Query Model for Querying Multidimensional Databases) **-** *Data Warehouse Implementation* (Efficient Data Cube Computation: An Overview Indexing OLAP, Data: Bitmap Index and Join Index, OLAP Server Architectures: ROLAP versus MOLAP versus HOLAP).

UNIT III (12 Hours)

Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods:
Basic Concept (Market Basket Analysis: A Motivational Example, Frequent Itemsets, Closed Itemsets and Association Rules) **-** *Frequent itemset Mining Methods* (Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate Generation, Generating Association Rules from Frequent Itemsets, Improving the Efficiency of Apriori, A Pattern Growth Approach for Mining Frequent Itemsets, Mining Frequent Itemsets Using Vertical Data Format)
Advanced Pattern Mining: *Pattern Mining: A Road Map* **-** *Pattern Mining in Multilevel*, *Multidimensional Space* (Mining Multilevel Association Rules, Mining Multi Dimensional Associations, Mining Quantitative Association Rules).

UNIT IV (12 Hours)

Classification:Basic Concepts: *Basic Concepts* (What Is Classification?, General Approaches to Classification) **-** *Decision Tree Induction* (Decision Tree Induction, Attribute Selection Measures, Tree Pruning, Scalability and Decision Tree Induction) **-** *Bayes Classification Methods* (Bayes Theorem, Naïve Bayesian Classification) **-** *Model Evaluation and Selection* (Metrics for Evaluating Classifier Performance, Holdout Method and Random Subsampling, Cross **-** Validation and Bootstrap).

Classification:Advanced Methods: *Bayesian Belief Networks* (Concepts and Mechanisms, Training Bayesian Belief Networks) **-** *Classification by Back Propagatio*n (A Multilayer Feed Forward Neural Network, Defining a Network Topology, Backpropagation).

UNIT V (12 Hours)

Cluster Analysis: Basic Concepts and Methods: *Cluster Analysis* (What is Cluster Analysis? Requirements for Cluster Analysis) **-** *A Partitioning Methods (k-M*eans and K-Medoid) **-** *Hierarchical Methods* (Agglomerative versus Divisive Hierarchical Clustering, Distance Measures in Algorithmic Methods, BRICH:Multiphase Hierarchical Clustering using Clustering Feature Trees, Chameleon: Multiphase Hierarchical Clustering Using Dynamic Modeling Hierarchical Clustering) **-** *Density Based Method* (DBSCAN).
Outlier Detection: *Outliers and Outlier Analysis* (What are Outliers Analysis?, Types of Outliers) **-** *Statistical Approaches* (Parametric Methods, Nonparametric Methods).

Reference Text Books:

1. Jiawei Han, Micheline Kamber, Data Mining: Concepts & Techniques, 2012.
2. Ralph Kimball, The Data Warehousing Toolkit, Wiley, Thomson, July 2013.
3. S.N.Sivanandam and S.Sumathi, Data Mining Concepts, Tasks and Techniques, Springer, October 2006.

PARVATHANENI BRAHMAYYA SIDDHARTHA COLLEGE OF ARTS & SCIENCE

(An Autonomous College in the jurisdiction of Krishna University)

M.Sc.(Computer Science), First Semester

Course Name: Data Mining Techniques

Course Code: 22DS1T4

(w.e.f admitted batch 2022-23)

Time: 3 Hours **Max Marks: 70**

SECTION-A

Answer all questions **5*4 = 20 Marks**

1. a) What are major issues of Data Mining?(CO1,L1)

(or)

b) Define *Data Preprocessing* and its steps (CO1,L1)

2. a) What is a *Data Warehouse* and OLTP? (CO2,L1)

(or)

b) What is difference between *OLAP Server* and *RLAP Server* (CO2,L1)

3. a) What is Pattern Mining? Lst out different methods for *Pattern Mining*. (CO3,L1)

(or)

b) What is *Market Basket Analysis* with example. (CO3L1)

4. a) Explain *Classification*? (CO4,L2)

(or)

b) Explain is *Bayes Theorem*. (CO4,L2)

5. a) What is *Cluster Analysis*? State different types *Cluster Analysis*? (CO5,L1)

(or)

b) What is *Outliers Analysis* and its method? (CO5,L1)

Answer all questions. All question carry equal marks. **5 × 10 = 50 Marks**

6. a) Define *Data Mining*. Describe the functionalities of Data Mining. (CO1,L1) 5 Marks

b) What is *Noisy Data*? Explain the *Binning Methods* for Data Smoothing. (CO1,L1) 5 Marks

(or)

c) What are different methods used in Data *Cleaning* and *Data Transformation* in *Data Preprocessing*? (CO1,L1) 10 Marks

7. a) Define *Data Warehouse*. Differentiate *Operational Databases* and *Data Warehouses*. (CO2,L1) 10Marks

(or)

b) List different schemas used in *Multidimensional Data Models* with diagrams. (CO2,L1) 5 Marks

c) What are the different OLAP operations in *Multidimensional Data Models*? (CO2,L1) 5 Marks

8. a) Explain the *Frequent Itemset Generation* in the *Apriori Algorithm.* (CO3,L2) 5 Marks

b) Explain different types of *Association Rules* (CO3,L2) 5 Marks

(or)

c) Explain *FP-Growth Algorithm* with example. (CO3,L2) 10 Marks

9. a) Explain how classification is done using *Decision Tree*. (CO4,L5) 5 Marks

b) Explain algorithm for *Decision Tree Induction.* (CO4,L5) 5 Marks

(or)

c) Explain *Bayes Theorem* in detail. (CO4,L5) 5 Marks

d) Explain *Bayesian Belief Network*. (CO4,L5) 5 Marks

10. a) Explain *Partitioning Methods* in *Cluster Analysis* with examples. (CO5,L5) 10 Marks

(or)

b) Explain *Chameleon* & *BIRCH Hierarchical* Clustering. (CO5,L5) 5 Marks

c) Explain different types of Outliers. (CO5,L5) 5 Marks