Semester: 6th			
Paper code: AIML304T	**L**	**T/P**	**Credits**
Subject: Introduction to Data Mining	**3**	**0**	**3**

Marking Scheme:

1. Teachers Continuous Evaluation: As per university examination norms from time to time
2. End Term Theory Examination: As per university examination norms from time to time

INSTRUCTIONS TO PAPER SETTERS: Maximum Marks: As per university norms

1. There should be 9 questions in the end term examination question paper.
2. Question No. 1 should be compulsory and cover the entire syllabus. This question should have objective or short answer type questions.
3. Apart from Question No. 1, the rest of the paper shall consist of four units as per the syllabus. Every unit should have two questions. However, students may be asked to attempt only 1 question from each unit.
4. The questions are to be framed keeping in view the learning outcomes of course/paper. The standard/ level of the questions to be asked should be at the level of the prescribed textbooks.
5. The requirement of (scientific) calculators/ log-tables/ data-tables may be specified if required.

Course Objectives:

1.	To identify the different types of data and using data pre-processing techniques applicable on the dataset.
2.	To evaluate various classification and clustering techniques on real world datasets.
3.	To apply data mining techniques on complex data types.
4.	To analyze different association rule mining and sequence mining techniques.

Course Outcomes:

CO1	Interpret the basic concepts of data mining techniques to identify interesting and relevant patterns.
CO2	Apply and perform pre-processing steps to prepare the data and get insights into the dataset.
CO3	Analyze different association rules identified using association rule mining or sequence mining on real life datasets.
CO4	Design and Develop models using classification and clustering techniques on complex data types.

Course Outcomes (CO) to Programme Outcomes (PO) Mapping

(Scale 1: Low, 2: Medium, 3: High)

CO/ PO	PO01	PO02	PO03	PO04	PO05	PO06	PO07	PO08	PO09	PO10	PO11	PO12
CO1	2	1	2	-	3	-	-	1	-	-	-	-
CO2	2	2	2	3	-	-	-	-	1	-	-	-
CO3	2	-		2	3	-	1	-	-	1	-	-
CO4	2	2		3	3	-	-	-	-		1	2

Course Overview:

The subject gives a detailed overview on data mining as a process starting from pre-processing the dataset to classification/clustering techniques on the data. The students are introduced to different techniques that can be applied to various types of complex data. Concepts like association rule mining and ensemble methods are also discussed in this subject.

UNIT I **[8]**

Data Mining Basics- What is Data Mining, Kinds of Patterns to be Mined, Tasks of Data Mining, Data Mining Applications- The Business Context of Data Mining, Data Mining as a Research Tool, Data Mining for Marketing, Benefits of data mining.

Data Pre-processing- Review of Data Pre-processing: Types of Data, Data Quality, Measurement and Data Collection Issues, Aggregation, Sampling, Dimensionality Reduction, Feature Subset Selection, Feature Creation, Data Discretization and Binarization, Variable Transformation, Measures of Similarity and Dissimilarity.

UNIT II **[8]**

Classification- Types of classifiers, Rule based classifiers, Model Selection, Model Evaluation, Artificial Neural Networks: Activation Functions (Sigmoid, Tanh, ReLU, Leaky ReLU, Selu), Perceptron, Multilayer Feed-Forward Neural Network, Backpropagation, Semi-supervised classification, Active Learning, Ensemble Methods: Methods for Constructing an Ensemble Classifier, Bias-Variance Decomposition, Bagging, Boosting, GBM, XGBoost, Stacking, Random Forest. Metrics for Evaluating Classification Performance: Holdout method, Cross Validation, Bootstrap

Handling Class Imbalance Problem: Evaluating Performance with Class Imbalance, Finding an Optimal Score Threshold, Multiclass Problem.

UNIT III **[8]**

Association Rule Mining- Mining Frequent Patterns, Associations and correlations, Market Basket Analysis, Apriori algorithm, Support Counting, Improving the efficiency of Apriori, Rule generation in Apriori algorithm, FP growth algorithm, Eclat algorithm, Mining Various kinds of Association Rules, Maximal Frequent Itemsets, Closed Itemsets, Evaluation of Association Patterns. Handling Categorical Attributes, Handling Continuous Attributes.

Sequential Patterns- Sequential Pattern Discovery, GSP algorithm, SPADE algorithm, Timing Constraints.

UNIT IV **[8]**

Cluster detection- Different Types of Clusters, Hierarchical Methods: Agglomerative and Divisive Clustering, Density based Clustering: DBSCAN algorithm, Comparing K-means and DBSCAN, Self-Organizing Maps (SOM), Cluster Evaluation. Outlier Analysis, Outlier Detection Methods. Mining Complex Data Types.

Avoiding False Discoveries- Significance Testing, Hypothesis Testing, Multiple Hypothesis Testing, Pitfalls in Statistical Testing

Text Books:

1. Tan Pang- Ning, Steinbach M., Viach, Kumar V., “Introduction to Data Mining”, Second Edition, Pearson, 2013.

2. Han J., Kamber M. and Pei J., “Data Mining Concepts and Techniques”, Second Edition, Hart Court India P. Ltd., Elsevier Publications, 2001.

Reference Books:

1. Zaki M.J., Meira W., “Data Mining and Machine Learning: Fundamental Concepts and Algorithms”, Second Edition, Cambridge University Press, 2020
2. Witten, E. Frank, M. Hall, “Data Mining: Practical Machine Learning Tools and Techniques”, Morgan Kaufmann Publishers, 2011.