GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Bachelor of Engineering
Level: UG
Subject Code: BE05000181
Subject Name: Data Mining Techniques

w.e.f. Academic Year:	2024-25
Semester:	5
Category of the Course:	Professional Elective Course - 2

Prerequisite:	Basic understanding of programming, data structures, databases, and fundamentals of statistics.
Rationale:	This subject enables students to understand techniques and methodologies for extracting meaningful patterns and knowledge from large volumes of data. It covers essential concepts such as data preprocessing, classification, clustering, association rule mining, and evaluation of models. Students learn how to transform raw data into useful insights that support decision-making in various domains. With the rapid growth of data in industries, the knowledge of data mining equips students with the ability to analyze complex datasets, identify trends, and build predictive models. This understanding is crucial for developing intelligent systems and solving real-world problems, making it an essential component for aspiring computer engineers and data professionals.

Course Outcome:

After Completion of the Course, Student will be able to:

No	Course Outcomes	Marks% weightage
1	Apply data pre-processing techniques to prepare datasets for effective data mining.	20
2	Analyze concept description methods and frequent pattern mining techniques to extract meaningful patterns from large datasets.	20
3	Evaluate and apply classification and prediction algorithms along with appropriate accuracy and error measures.	20
4	Apply clustering techniques to group similar data and evaluate the quality of the groups.	20
5	Explore advanced data mining topics, hyperparameter tuning and explainable AI and interpret their applications.	20

**Revised Bloom's Taxonomy (RBT)*

Teaching and Examination Scheme:

Teaching - Learning Scheme (in Hours per Semester)					Total Credits = TH/30	Assessment Pattern and Marks					Total Marks
						Theory		Tutorial / Practical			
L	T	P	PBL*	TH		ESE (E)	PA (M)	PA (I)	PBL (I)	ESE (V)	
45	0	30	15	90	03	70	30	20	30	50	200

* Problem Based Learning (PBL) aims to accommodate learning beyond syllabus as per clause 9.4 of NBA manual.

Course Content:

Unit No.	Content	No. of Hours
1.	Data Mining, Knowledge Discovery in Databases (KDD) Process, Data Mining Functionalities, Applications of Data Mining, Data Mining Task Primitives, Major Issues in Data Mining, Introduction to Statistical Measures: measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and basic data distribution concepts.	3
2.	Data Pre-processing: Data Cleaning, Data Integration, Data Reduction, Data Transformation and Data Discretization. Introduction of data mining tool such as orange.	5
3.	Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods: Market Basket Analysis, Frequent Itemsets, Closed Itemsets, and Association Rule, Frequent Itemset Mining Methods, mining various kind of association rules, from association mining to correlation analysis, Which Patterns Are Interesting? —Pattern Evaluation Methods.	7
4.	Classification and Prediction: Classification vs. prediction, Issues regarding classification and prediction, Statistical-Based Algorithms, Distance-Based Algorithms, Decision Tree Induction, Bayes Classification Methods, Rule-Based Classification, Model Evaluation and Selection, Techniques to Improve Classification Accuracy, Bayesian Belief Networks, Classification by Backpropagation, Support Vector Machines, Linear and nonlinear regression, Logistic Regression, Introduction of data mining tool such as WEKA.	10
5.	Cluster Analysis: What Is Cluster Analysis? Requirements for Cluster Analysis, Partitioning Methods: k-Means: A Centroid-Based Technique, k-Medoids: A Representative Object-Based Technique, Hierarchical Methods: Agglomerative and Divisive Hierarchical Clustering, Density-Based Methods: DBSCAN.	10
6.	Introduction to hyperparameters: Need for and importance of hyperparameter tuning, Overfitting and underfitting, Hyperparameters in algorithms, Different methods of Hyperparameter tuning techniques: Explainable AI in Data Mining: Introduction to Explainable AI (XAI), Need for explainability in data mining, Black-box vs interpretable models, Explainable AI techniques.	5
7.	Advance data mining techniques: Web Mining, Spatial Mining, Temporal Mining, Spatial temporal Mining, Stream Mining, Text Mining, And Multimedia Mining. Applications of Distributed and parallel Data Mining.	5

Suggested Specification Table with Marks(Theory):

Distribution of Theory Marks (in%)					
RLevel	**ULevel**	**ALevel**	**NLevel**	**E Level**	**CLevel**
20	**20**	**25**	**25**	**10**	**-**

Where R:Remember; U:Understanding; A:Application; N:Analyz; E:Evaluate; C: Create (as per Revised Bloom's Taxonomy)

References books:

1. J. Han, M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann
2. M. Kantardzic, "Data mining: Concepts, models, methods and algorithms, John Wiley &Sons Inc.
3. M. Dunham, "Data Mining: Introductory and Advanced Topics", Pearson Education.
4. Ning Tan, Vipin Kumar, Michael Steinbanch Pang, "Introduction to Data Mining", Pearson Education

Open-source software and website:

5. WEKA
6. Orange
7. R
8. RapidMiner

Online Courses:

1. Data Mining by Prof. Pabitra Mitra, IIT Kharagpur
2. Business Analytics and Data Mining Modeling Using R by Prof. Gaurav Dixit, IIT Roorkee

Suggested Course Practical List:(List can be change according to Latest Development)

1.Data Cleaning and Data Transformation
Handle missing values, noisy data, and inconsistencies using techniques like binning and smoothing. Apply normalization techniques such as Min-Max, Z-score, and Decimal Scaling on datasets.

2. Data Reduction and Feature Selection
Implement data reduction techniques and perform feature selection to identify important attributes.

3.Preprocessing using Tools
To perform hand on experiments of data preprocessing with sample data with Orange tool.

4. Frequent Pattern Mining (Apriori Algorithm)
Implement Apriori algorithm and generate frequent itemsets and association rules.

5. Association Rule Mining using Tools
Apply association rule mining on datasets using WEKA/Orange and evaluate rule quality.

6. Classification and Prediction Techniques
Implement classification algorithms (Decision Tree / KNN / Naive Bayes) and evaluate accuracy.

7. Classification/ Prediction using Tools
Apply classification /Prediction on datasets using WEKA/Orange and evaluate rule quality.

8. Clustering Techniques
Apply K-Means and Hierarchical clustering algorithms and analyze clustering performance.

9. Advanced Data Mining Application / Case Study
Perform web mining / spatial / temporal / multimedia data using any tool.

10. Real Life Application
To solve a real-world problem using data mining in Python with preprocessing, model building, hyperparameter tuning.

List of suggested activities for Problem Based Learning:

Sr. No	Name of the activity	No. of hours	Evaluation Criteria
1.	Assignment writing	3 assignments of 5h each. Total – 15h	Based on correctness of solutions, conceptual understanding and timely submission of assignments.
2.	Seminar / Presentation	Duration for study and preparation=10h Report writing=3h Presentation=2h Total=15h	Topic can be selected from the technical content beyond syllabus
3.	Poster/chart/power point preparation on technical topics	Duration = 15h	Based on poster/chart preparation and presentation skills
4.	Real world case studies-based learning	Duration of data collection/study = 5h Report preparation = 10h Total = 15h	Based on in-depth study, technical depth, data collected, fact finding, etc.
5.	Self-learning on-line course related to Data Mining((NPTEL / Coursera).	Minimum duration of the course should be 15h.	Examination based assessment at the end of course. Based on the certificate produced.
6.	Project/Research work	Duration = 15 h	Depending on the complexity of the Project/Research
7.	Research paper-based learning related to the subject	Duration of Learning = 15h	Based on understanding of research paper, interpretation, technical depth, presentation
8.	Group discussion on emerging applications of Data Mining.	Duration – 3hrs for each discussion (Total 5 discussions in semester) (Topic preparation – 1 hrs, group discussion – 1 hrs, summary/report preparation – 1 hrs. Total – 3 hrs)	Based on participation in discussion, technical depth of arguments and summarized conclusions.

Note:

- All the suggested activity should be related to the subject.
- One activity must be carried out as per the availability of faculties and students.
- The number of hours is suggestive. Faculty can sub-divide the number of hours based on the activity. However, total number of hours is fixed.
- Rubrics for the evaluation can be prepared by the faculty.
- All records pertaining to the evaluation and assessment of self-learning activities must be properly maintained and preserved at the institute level. These records should be made available to the university upon request.
- Institutes are encouraged to utilize digital platforms, such as Microsoft Teams, for effective record-keeping and to ensure transparency in the evaluation and assessment of self-learning activities.

* * * * * *