

Year: B. Tech III (Semester V)

Subject Name: Data Mining and Business Intelligence

Subject Code: BTIT13501

Type of course: Professional Core Course

Prerequisite (if any): Database Management Systems, Design and Analysis of Algorithms

Rationale: Data mining is the study of algorithms for finding patterns in large amounts of data. It is an interdisciplinary topic involving databases, statistics, machine learning and algorithms. Data Mining is useful for gaining business insight by processing large volumes of customer data. Data Mining helps businesses to learn more about their customers' buying preferences, develop more effective marketing strategies, increase sales and decrease costs. This course will cover important aspects of data mining including algorithms for data pre-processing, association rule mining, classification and clustering. It also explores various applications of data mining, with a special focus on business-related applications.

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:

Sr. No.	Contents	Total Hrs
1.	Introduction to Data Mining and Data Warehousing: Motivation for Data Mining, Mining patterns from data – finding what patterns can be mined and what are interesting patterns, Data Mining and Knowledge Discovery Process (KDD), Data mining as a confluence of multiple disciplines, data warehouses and data marts, data warehouse architecture, warehouse schemas, multi-dimensional data model, OLAP operations, OLAP vs OLTP, Major issues in Data Mining.	08
2.	Data Pre-processing: Data cleaning, pre-processing for noisy data, missing values, Data Integration and Transformation, Data Reduction, Data Discretization and Concept Hierarchy generation.	05

3.	Association Rule Mining: Market basket analysis, finding frequent itemsets, Apriori algorithm, Association Rule Mining, Methods to improve efficiency of Apriori algorithm- Hashing, Transaction Reduction, Partitioning, Sampling; FP Growth algorithm, FP Tree; Apriori vs FP Growth for Market Basket Analysis, Implications of market basket analysis to Inventory Control, Online Shopping and Sales.	08
4.	Linear Regression and Prediction: Linear Regression and its applications to business- Using linear regression to predict stock prices and trends in business	04
5.	Clustering: Categorization of clustering methods, Partition-based clustering algorithms- k-means, k-medoids, Hierarchical clustering, Density based methods, Outlier detection and Analysis, Application of clustering to Customer Segregation and Targeted Marketing	06
6.	Privacy Preserving Data Mining: Privacy implications of data mining on consumers, Challenges to privacy preservation, Methods for Privacy Preserving Data Publishing- randomization, k-anonymity, l-diversity; Changing the results of Data Mining applications to preserve privacy - Association Rule hiding, Query auditing.	08
7.	Business Applications and Case Studies of Data Mining: Applications and case studies of data mining in Banking & Finance, Fraud Detection, Retail Sector, Recommendation Systems, Customer Relationship Management and maintaining customer loyalty- predicting churn, Telecommunications Industry etc.;	06

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
15	20	10	5	5	5

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

Reference Books:

Sr No.	Title of book /article	Author(s)	Publisher and details like ISBN
1.	Data Mining: Concepts and Techniques	J. Han, M. Kamber	Morgan Kaufmann
2.	Data Mining for Business Intelligence	Shmueli, Bruce, Yahav, Patel, Lichtendahl Jr.	Wiley
3.	Data Mining for Business Applications	Longbing Cao, Philip S. Yu,	Springer

		Chengqi Zhang, Huaifeng Zhang	
4.	Data Mining: Introductory and Advanced Topics	M. Dunham	Pearson Education

Note: Students should refer to the latest editions of books

Course Outcomes (CO):

Sr. No.	CO statements	Marks % weightage
CO-1	Explain the basic concepts of data mining and business intelligence, and the major issues in data mining	25%
CO-2	Demonstrate the ability to pre-process data needed for data mining algorithms	20%
CO-3	Select and implement appropriate data mining methods like classification, clustering and association rule mining on large data sets.	30%
CO-4	Apply data mining methods and open source tools to solve practical problems.	25%

List of Open learning website:

- <https://nptel.ac.in/courses/106105174>
- WEKA (The Data Platform for AI) <https://www.weka.io/>

List of Experiments:

1. Study the following data mining applications and make a detailed note on them:

- Mining Time-Series Data
- Social Network Mining
- Recommendation Systems in Retail
- Web Mining
- Text Mining

2. Implement routines to normalize the data in the sample data file using:

- Min-Max normalization
 - Map to range [0,1]
 - Map to range [-1,1]
- z-score normalization
- decimal scaled normalization

Sample DataSet: 10, 12, 3, 6, 5, 25, 17, 100, 1000, 98, 11, 27, 78, 33, 9, 18, 23, 44, 690, 200

3. Implement Binning methods for data smoothing in following dataset using bin depth of 3
Demonstrate (a) smoothing by bin means, (b) smoothing by bin medians (c) bin boundaries.
Sample Dataset: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 5, 35, 35,
36, 40, 45, 46, 52, 70
4. Study and implement the k-means clustering algorithm over the given dataset.
5. Build a naive Bayesian Classifier for the Playtennis dataset. (Use WEKA toolkit)
6. Generate a Decision Tree Classifier for the Playtennis dataset. (Use WEKA toolkit)
7. Study and implement the Apriori algorithm over the given dataset. (Use WEKA toolkit)
8. Generate a linear regression based model to predict stock prices.