KUMARAGURU
college of technology
character is life

Register Number:……

MCA DEGREE EXAMINATIONS: NOV/DEC 2024

(Regulation 2020)

Third Semester

MASTER OF COMPUTER APPLICATIONS

P20CAT2003**:** Data Intensive Computing

COURSE OUTCOMES

After successful completion of this course, the students should be able to

CO1: Understand the fundamentals of Data Mining and Pre-processing

CO2: Apply the regression and classification techniques

CO3: Evaluate the models using performance metrics

CO4: Cluster the high dimensional data and apply the association rules for mining the data

CO5: Apply various methods to detect outliers

CO6: Implement the text analysis

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions:-

PART A (10 x 2 = 20 Marks)

1.	What are two major challenges in data mining?	CO1	[K_2]
2.	What is the difference between classification and clustering?	CO2	[K_4]
3.	Why is data pre-processing important in data mining?	CO1	[K_4]
4.	What is the purpose of data reduction in data mining?	CO1	[K_2]
5.	Define the term "training set" and “testing set” in classification.	CO2	[K_2]
6.	What is the main objective of linear regression?	CO2	[K_2]
7.	What is association rule mining, and how is it used in data analysis?	CO4	[K_2]
8.	What is cluster analysis, and what is its primary goal in data mining?	CO4	[K_2]
9.	Why is outlier detection important in data mining and machine learning?	CO5	[K_4]
10.	What is the term "tokenization," and why is it a critical step in text processing?	CO6	[K_2]

PART B (6 x 5 = 30 Marks)

11.	You have a dataset with several missing values and noisy data. Describe how you would apply data cleaning techniques to prepare the dataset for analysis.	CO1	[K_3]
12.	What are the main steps in data integration? How does it help in combining data from multiple sources?	CO1	[K_3]
13.	Given a dataset with categorical features, explain how you would construct a	CO2	[K_3]

decision tree. What splitting criteria (e.g., Gini index, Information Gain) would you use, and why?

14.	Describe the Apriori algorithm in detail, including its steps for generating frequent itemsets with an example.	CO4	[K_2]
15.	Explain the different types of outliers: global outliers, contextual outliers, and collective outliers.	CO5	[K_2]
16.	Define text mining and discuss its significance in extracting valuable information from unstructured text data.	CO6	[K_2]

Answer any FIVE Questions
PART C (5 x 10 = 50 Marks)

17.	Discuss the role of various technologies in data mining, including machine learning, databases, and statistics. How do these technologies integrate to support the data mining process?	CO1	[K_2]
18.	You are given a large dataset with missing, noisy, and inconsistent data. Explain the different data cleaning techniques you would apply to handle these issues. Provide examples of how each technique would improve the quality of the dataset.	CO1	[K_3]
19.	Describe the Naïve Bayes algorithm in detail, including its underlying assumptions and how it handles classification tasks. Explain the process of calculating prior and conditional probabilities using a given dataset.	CO2	[K_2]
20.	Discuss the various methods for evaluating classification models, focusing on confusion matrices, precision, recall, F1 score, and ROC curves. Explain the significance of each metric in assessing model performance and provide a comparative analysis of how these metrics can be used to make informed decisions about model selection and tuning.	CO3	[K_2]
21.	Explain the K-Means clustering algorithm in detail, including the steps involved in its execution. Discuss how the choice of the number of clusters (k) impacts the results and describe methods to determine the optimal value of k.	CO4	[K_2]
22.	Describe various common representation methods used in data visualization, including bar charts, line graphs, pie charts, heatmaps, and scatter plots. Discuss the strengths and weaknesses of each method.	CO6	[K_4]

****