

SRM VALLIAMMAI ENGINEERING COLLEGE

SRM Nagar, Kattankulathur – 603 203



DEPARTMENT OF INFORMATION TECHNOLOGY

MTECH-DATA SCIENCE

QUESTION BANK

I SEMESTER

DS3163- DATA MINING

Regulation – 2023

Academic Year 2026– 2027(Odd Semester)

Prepared by,

Mrs.S.Jonisha, Assistant Professor (O.G) / IT



SRM VALLIAMMAI ENGINEERING COLLEGE

(An Autonomous Institution)

SRM Nagar, Kattankulathur – 603 203.



DEPARTMENT OF INFORMATION TECHNOLOGY

MTECH - DATA SCIENCE

QUESTION BANK

SUBJECT : DS3163- DATA MINING

SEM / YEAR : I / First year

UNIT -1 INTRODUCTION TO DATA MINING AND PREPROCESSING

Data mining - Related technologies - Machine Learning, DBMS, OLAP, Statistics – Data Mining Goals - Stages of the Data Mining Process - Data Mining Techniques – Knowledge Representation Methods – Applications. Data pre-processing- Data cleaning- Data transformation - Data reduction - Discretization and generating concept hierarchies.

PART-A

Q.No.	Questions	Level	Competence
1	Define Data mining. List out the steps in data mining?	BTL1	Remembering
2	List the ways in which interesting patterns should be mined.	BTL1	Remembering
3	Identify the steps involved in the process of KDD. How does it relate to data mining?	BTL2	Understanding
4	How do data mining results assist in decision-making?	BTL2	Understanding
5	List the stages of the data mining process.	BTL1	Remembering
6	What are the main OLAP operations, and how do they enhance multidimensional data analysis?	BTL1	Remembering
7	Distinguish between data cleaning and noisy data.	BTL2	Understanding
8	Differentiate between descriptive, predictive and prescriptive data mining.	BTL2	Understanding
9	What is Support Vector Machine (SVM)?	BTL1	Remembering
10	Why is the goal of anomaly detection crucial in data mining?	BTL2	Understanding
11	State the role of statistics in data mining?	BTL1	Remembering
12	Define clustering and its role in data mining.	BTL2	Understanding
13	What is machine learning in the context of data mining?	BTL1	Remembering
14	Why is understanding customer behavior an important goal of data mining?	BTL2	Understanding
15	Distinguish between classification and regression techniques in data mining.	BTL2	Understanding
16	List out the goals of data mining.	BTL1	Remembering
17	List the common knowledge representation methods used in data mining?	BTL1	Remembering
18	Define concept hierarchy generation.	BTL1	Remembering
19	State the role of a DBMS in data mining.	BTL1	Remembering
20	What is OLAP and how is it related to data mining?	BTL2	Understanding
21	What is association rule mining?	BTL1	Remembering
22	What is the purpose of pattern evaluation?	BTL1	Remembering
23	How does generating concept hierarchies help in summarizing data?	BTL2	Understanding
24	How does knowledge representation improve decision-making in data mining?	BTL2	Understanding

PART-B

Q.No	Questions	Marks	Level	Competence
1	Determine the role of Machine Learning, DBMS, OLAP, and Statistics in Data Mining.	16	BTL4	Analyzing
2	Explain the various data mining steps in knowledge discovery with a detailed description of each stage.	16	BTL3	Applying
3	(i) How Does Data Mining Work? (ii) Apply the concepts of predictive, descriptive, and prescriptive data mining to a business case scenario.	10 6	BTL3	Applying
4	(i) Analyze the various applications of Data Mining across different industries, such as healthcare, finance, retail, and telecommunications. (ii) Explain the various types of OLAP in detail.	10 6	BTL4	Analyzing
5	(i) Analyze the key challenges in the implementation of Data Mining in real-world applications. (ii) Describe the role of statistics in data mining for discovering patterns, correlations, and insights from large datasets.	6 10	BTL4	Analyzing
6	Explain the different types of Machine Learning and examine how each type can be used to address specific problems and the impact they have on various domains, such as healthcare, finance, and robotics.	16	BTL5	Evaluating
7	(i) Discuss the role of OLAP (Online Analytical Processing) in data analysis and decision-making and analyse the advantages of combining OLAP and Data Mining in data analysis and decision-making. (ii) Explain how DBMS supports the Data Mining process, focusing on data storage, retrieval, and management.	8 8	BTL4	Analyzing
8	(i) Discuss the primary goals of data mining and their significance in solving business problems. (ii) Write in brief about Data cleaning?	10 6	BTL3	Applying
9	Describe the key stages of the data mining process and their importance in extracting meaningful insights from data.	16	BTL5	Evaluating
10	Explain the various data mining techniques to a large dataset to extract useful information.	16	BTL3	Applying
11	Discuss how different knowledge representation methods are used to interpret and visualize the results of data mining processes, and evaluate their effectiveness in various applications.	16	BTL3	Applying
12	Discuss the key steps in data pre-processing for data mining and their impact on the overall data analysis process. Justify the selection of specific pre-processing techniques based on the characteristics of the dataset and the objectives of the data mining task.	16	BTL5	Evaluating
13	(i) Explain the concept of data discretization and concept hierarchy generation. (ii) Explain the need for data pre-processing.	8 8	BTL3	Applying
14	(i) Evaluate how data reduction techniques affect the performance of data mining algorithms and the interpretability of results. (ii) Explain the impact of data transformations on the quality of data mining results.	8 8	BTL 5	Evaluating
15	(i) In the context of the data mining process, where does smoothing	8	BTL4	Analyzing

	fit, and why is it important? (ii) Analyze Using Equi-depth binning method, partition the data given below into 4 bins and perform smoothing according to the following methods. 1. Smoothing by bin means 2. Smoothing by bin median 3. Smoothing by bin boundaries 24,25,26,27,28,56,67,70,70,75,78,89,89,90,91,94,95,96,100,102,103,107,109,112.	8		
16	(i) What is Data reduction? Discuss in detail? (ii) Evaluate the effectiveness of different data mining techniques in achieving the goals of a data mining project.	6 10	BTL5	Evaluating
17	Consider the following data for the attribute AGE: 4,8,21,5,21,24,34,28,25,24,30,26. Perform smoothing by bin means, bin boundaries, bin frequencies, equi-depth and equi-frequency using a bin depth of 3.	16	BTL4	Analyzing

UNIT-2 CLASSIFICATION ALGORITHMS I

Association rules: Basic idea: item sets - Generating item sets and rules efficiently - Correlation analysis. Classification: Basic learning/mining tasks - Inferring rudimentary rules: 1R algorithm - Decision trees - Bayes Classification Methods - Rule-Based Classification - Model Evaluation and Selection - Techniques to Improve Classification Accuracy.

PART-A

Q.No	Questions	Level	Competence
1	What are the three key metrics used to evaluate association rules in data mining? Briefly define each.	BTL1	Remembering
2	Write the structure of an association rule in data mining and mention its two parts.	BTL2	Understanding
3	How would you explain the principle of Apriori algorithm? List down the stages involved in the apriori algorithm.	BTL2	Understanding
4	Define Data pruning. State the need for pruning phase in decision tree construction.	BTL2	Understanding
5	Compare the advantages of FP growth algorithm over apriori algorithm.	BTL1	Remembering
6	Define rudimentary rules with example.	BTL1	Remembering
7	What is naive Bayesian classification? How is it differing from Bayesian classification?	BTL2	Understanding
8	Define an itemset in association rule mining and differentiate between a single-item itemset and a multi-item itemset.	BTL1	Remembering
9	Write the Bayes theorem.	BTL1	Remembering
10	Draw a decision tree model for predicting whether a customer will purchase a product based on two features: Age and Income.	BTL2	Understanding
11	What is model evaluation? List out the common evaluation metrics.	BTL2	Understanding
12	What is a Decision Tree?	BTL2	Understanding
13	Define Correlation Analysis. What are the types of correlation?	BTL1	Remembering
14	Illustrate the benefits of correlation.	BTL1	Remembering
15	How would you show your understanding about rule based classification?	BTL2	Understanding
16	Suppose 15 men out of 300 men and 25 women out of 1000 are good orators. An orator is chosen at random. Find the probability that a male person is selected.	BTL2	Understanding
17	Differentiate between Bayes theorem and conditional probability?	BTL1	Remembering
18	What is Correlation?	BTL2	Understanding
19	List out the different classification algorithms.	BTL1	Remembering

20	What is the 1R Algorithm? How the 1R Algorithm Works?	BTL1	Remembering
21	Illustrate the structure of a Decision Tree.	BTL2	Understanding
22	Define classification. What are the two main types of classification?	BTL1	Remembering
23	Prioritize the steps involved in preparing the data for classification.	BTL2	Understanding
24	How to conduct Correlation Analysis?	BTL1	Remembering

PART-B

Q.No.	Questions	Marks	Level	Competence																
1	(i) Explain the concept of association rules in data mining. (ii) Discuss the process of generating frequent itemsets and association rules efficiently using the Apriori algorithm.	6 10	BTL3	Applying																
2	Explain decision trees in detail with an example.	16	BTL3	Applying																
3	Explain Bayesian classification in detail with example.	16	BTL3	Applying																
4	Explain the FP-Growth algorithm in data mining. How does it differ from the Apriori algorithm in terms of generating frequent itemsets? Discuss the steps involved in constructing the FP-tree and mining frequent patterns from it.	16	BTL3	Applying																
5	(i)What is correlation? How to Conduct Correlation Analysis? (ii)A health researcher is studying the relationship between physical exercise (in hours per week) and weight loss (in kg) over a month. The data from 5 individuals is given below: <table><tr><th>Hours of Exercise (X)</th><th>Weight Loss (Y)</th></tr><tr><td>1</td><td>1</td></tr><tr><td>2</td><td>2</td></tr><tr><td>3</td><td>4</td></tr><tr><td>4</td><td>5</td></tr><tr><td>5</td><td>7</td></tr></table> Calculate the Pearson correlation coefficient (r) to determine if there is a correlation between the number of hours of exercise per week and the amount of weight loss.	Hours of Exercise (X)	Weight Loss (Y)	1	1	2	2	3	4	4	5	5	7	6 10	BTL5	Evaluating				
Hours of Exercise (X)	Weight Loss (Y)																			
1	1																			
2	2																			
3	4																			
4	5																			
5	7																			
6	(i) Explain the metrics used to evaluate association rule. (ii) Calculate the support, confidence, and lift for the association rule: Milk⇒Bread <table><tr><th>Transaction ID</th><th>Items Bought</th></tr><tr><td>T1</td><td>{Milk, Bread, Eggs}</td></tr><tr><td>T2</td><td>{Milk, Bread}</td></tr><tr><td>T3</td><td>{Bread, Butter}</td></tr><tr><td>T4</td><td>{Milk, Butter}</td></tr><tr><td>T5</td><td>{Milk, Bread, Butter}</td></tr></table>	Transaction ID	Items Bought	T1	{Milk, Bread, Eggs}	T2	{Milk, Bread}	T3	{Bread, Butter}	T4	{Milk, Butter}	T5	{Milk, Bread, Butter}	6 10	BTL5	Evaluating				
Transaction ID	Items Bought																			
T1	{Milk, Bread, Eggs}																			
T2	{Milk, Bread}																			
T3	{Bread, Butter}																			
T4	{Milk, Butter}																			
T5	{Milk, Bread, Butter}																			
7	Explain and Apply the Apriori algorithm for discovering frequent item sets of the table. Use 0.3 for the minimum support value. <table><tr><th>Trans ID</th><th>Items Purchased</th></tr><tr><td>101</td><td>Milk,bread,eggs</td></tr><tr><td>102</td><td>Milk,juice</td></tr><tr><td>103</td><td>Juice, butter</td></tr><tr><td>104</td><td>Milk,bread,eggs</td></tr><tr><td>105</td><td>Coffee, eggs</td></tr><tr><td>106</td><td>Coffee</td></tr><tr><td>107</td><td>Coffee, Juice</td></tr></table>	Trans ID	Items Purchased	101	Milk,bread,eggs	102	Milk,juice	103	Juice, butter	104	Milk,bread,eggs	105	Coffee, eggs	106	Coffee	107	Coffee, Juice	16	BTL5	Evaluating
Trans ID	Items Purchased																			
101	Milk,bread,eggs																			
102	Milk,juice																			
103	Juice, butter																			
104	Milk,bread,eggs																			
105	Coffee, eggs																			
106	Coffee																			
107	Coffee, Juice																			

	<table><tr><td>108</td><td>Milk,bread,cookies,eggs</td></tr><tr><td>109</td><td>Cookies, butter</td></tr><tr><td>110</td><td>Milk, bread</td></tr></table>	108	Milk,bread,cookies,eggs	109	Cookies, butter	110	Milk, bread																																	
108	Milk,bread,cookies,eggs																																							
109	Cookies, butter																																							
110	Milk, bread																																							
8	Using the Apriori algorithm with a minimum support threshold of 40% and a minimum confidence threshold of 70%, answer the following: 1. Apply the Apriori algorithm to this dataset to identify all frequent 1-itemsets and 2-itemsets. Show your calculations and explain how the algorithm prunes non-frequent itemsets. 2. Based on the frequent itemsets, generate at least two strong association rules with confidence $\geq 70\%$, and calculate the confidence for each rule. <table><tr><td>Transaction ID</td><td>Items Bought</td></tr><tr><td>T1</td><td>{Milk, Bread, Eggs}</td></tr><tr><td>T2</td><td>{Milk, Bread}</td></tr><tr><td>T3</td><td>{Bread, Butter}</td></tr><tr><td>T4</td><td>{Milk, Butter}</td></tr><tr><td>T5</td><td>{Milk, Bread, Butter}</td></tr></table>	Transaction ID	Items Bought	T1	{Milk, Bread, Eggs}	T2	{Milk, Bread}	T3	{Bread, Butter}	T4	{Milk, Butter}	T5	{Milk, Bread, Butter}	16	BTL5	Evaluating																								
Transaction ID	Items Bought																																							
T1	{Milk, Bread, Eggs}																																							
T2	{Milk, Bread}																																							
T3	{Bread, Butter}																																							
T4	{Milk, Butter}																																							
T5	{Milk, Bread, Butter}																																							
9	(i) Classify whether a person buys a car based on their age, income, and student status using Naive Bayes Classification. <table><tr><td>Age</td><td>Income</td><td>Student</td><td>Buys Car?</td></tr><tr><td>Young</td><td>High</td><td>No</td><td>No</td></tr><tr><td>Young</td><td>High</td><td>Yes</td><td>No</td></tr><tr><td>Middle</td><td>High</td><td>No</td><td>Yes</td></tr><tr><td>Old</td><td>Medium</td><td>No</td><td>Yes</td></tr><tr><td>Old</td><td>Low</td><td>Yes</td><td>Yes</td></tr><tr><td>Old</td><td>Low</td><td>No</td><td>No</td></tr><tr><td>Middle</td><td>Low</td><td>No</td><td>Yes</td></tr><tr><td>Young</td><td>Medium</td><td>No</td><td>No</td></tr></table> (ii) Bag1 contains 4 white and 8 black balls and Bag2 contains 5 white and 3 black balls. From one of the bag one ball is drawn at random and the ball which is drawn comes out as black. Find the probability that the ball is drawn from Bag1.	Age	Income	Student	Buys Car?	Young	High	No	No	Young	High	Yes	No	Middle	High	No	Yes	Old	Medium	No	Yes	Old	Low	Yes	Yes	Old	Low	No	No	Middle	Low	No	Yes	Young	Medium	No	No	10 6	BTL4	Analyzing
Age	Income	Student	Buys Car?																																					
Young	High	No	No																																					
Young	High	Yes	No																																					
Middle	High	No	Yes																																					
Old	Medium	No	Yes																																					
Old	Low	Yes	Yes																																					
Old	Low	No	No																																					
Middle	Low	No	Yes																																					
Young	Medium	No	No																																					
10	(i) Describe in detail about correlation analysis. (ii) A researcher collects data on the number of hours studied per week (X) and the corresponding test scores (Y) for a group of students. The data is as follows: <table><tr><td>Hours Studied (X)</td><td>Test Score (Y)</td></tr><tr><td>2</td><td>50</td></tr><tr><td>4</td><td>60</td></tr><tr><td>6</td><td>70</td></tr><tr><td>8</td><td>80</td></tr><tr><td>10</td><td>85</td></tr></table> Calculate the Pearson correlation coefficient (r) to determine the relationship between hours studied and test scores.	Hours Studied (X)	Test Score (Y)	2	50	4	60	6	70	8	80	10	85	8 8	BTL4	Analyzing																								
Hours Studied (X)	Test Score (Y)																																							
2	50																																							
4	60																																							
6	70																																							
8	80																																							
10	85																																							

11	Discuss the different types of basic learning and data mining task. Provide examples for each type.	16	BTL3	Applying																																																																											
12	Explain rule based classification with suitable example.	16	BTL4	Analyzing																																																																											
13	Explain the Techniques to Improve Classification Accuracy.	16	BTL3	Applying																																																																											
14	Write the Bayes theorem. Explain the working of a Bayesian classifier for the following dataset. For a given test tuple, predict whether a person buys computer or not. <table><tr><th>Age</th><th>Income</th><th>Student</th><th>Credit rating</th><th>Buys Computer</th></tr><tr><td><=30</td><td>High</td><td>No</td><td>Fair</td><td>No</td></tr><tr><td><=30</td><td>High</td><td>No</td><td>excellent</td><td>No</td></tr><tr><td>31..40</td><td>High</td><td>No</td><td>Fair</td><td>Yes</td></tr><tr><td>>40</td><td>Medium</td><td>No</td><td>Fair</td><td>Yes</td></tr><tr><td>>40</td><td>Low</td><td>Yes</td><td>Fair</td><td>Yes</td></tr><tr><td>>40</td><td>Low</td><td>Yes</td><td>excellent</td><td>No</td></tr><tr><td>31..40</td><td>Low</td><td>Yes</td><td>excellent</td><td>Yes</td></tr><tr><td><=30</td><td>Medium</td><td>No</td><td>Fair</td><td>No</td></tr><tr><td><=30</td><td>Low</td><td>Yes</td><td>Fair</td><td>Yes</td></tr><tr><td>>40</td><td>Medium</td><td>Yes</td><td>Fair</td><td>Yes</td></tr><tr><td><=30</td><td>Medium</td><td>Yes</td><td>excellent</td><td>Yes</td></tr><tr><td>31..40</td><td>Medium</td><td>No</td><td>excellent</td><td>Yes</td></tr><tr><td>31..40</td><td>High</td><td>Yes</td><td>Fair</td><td>Yes</td></tr><tr><td>>40</td><td>Medium</td><td>No</td><td>excellent</td><td>No</td></tr></table>	Age	Income	Student	Credit rating	Buys Computer	<=30	High	No	Fair	No	<=30	High	No	excellent	No	31..40	High	No	Fair	Yes	>40	Medium	No	Fair	Yes	>40	Low	Yes	Fair	Yes	>40	Low	Yes	excellent	No	31..40	Low	Yes	excellent	Yes	<=30	Medium	No	Fair	No	<=30	Low	Yes	Fair	Yes	>40	Medium	Yes	Fair	Yes	<=30	Medium	Yes	excellent	Yes	31..40	Medium	No	excellent	Yes	31..40	High	Yes	Fair	Yes	>40	Medium	No	excellent	No	16	BTL6	Creating
Age	Income	Student	Credit rating	Buys Computer																																																																											
<=30	High	No	Fair	No																																																																											
<=30	High	No	excellent	No																																																																											
31..40	High	No	Fair	Yes																																																																											
>40	Medium	No	Fair	Yes																																																																											
>40	Low	Yes	Fair	Yes																																																																											
>40	Low	Yes	excellent	No																																																																											
31..40	Low	Yes	excellent	Yes																																																																											
<=30	Medium	No	Fair	No																																																																											
<=30	Low	Yes	Fair	Yes																																																																											
>40	Medium	Yes	Fair	Yes																																																																											
<=30	Medium	Yes	excellent	Yes																																																																											
31..40	Medium	No	excellent	Yes																																																																											
31..40	High	Yes	Fair	Yes																																																																											
>40	Medium	No	excellent	No																																																																											
15	Find all frequent item sets for the given training set using Apriori and FP growth respectively. Compare the efficiency of the two mining processes. Take minimum support as 3 and minimum confidence threshold as 80%. <table><tr><th>TID</th><th>ITEMS BOUGHT</th></tr><tr><td>T100</td><td>{M, O, N, K, E, Y }</td></tr><tr><td>T200</td><td>{D, O, N, K, E, Y }</td></tr><tr><td>T300</td><td>{M, A K, E }</td></tr><tr><td>T400</td><td>{M,U, C, K,Y }</td></tr><tr><td>T500</td><td>{C, O, O,K, I, E }</td></tr></table>	TID	ITEMS BOUGHT	T100	{M, O, N, K, E, Y }	T200	{D, O, N, K, E, Y }	T300	{M, A K, E }	T400	{M,U, C, K,Y }	T500	{C, O, O,K, I, E }	16	BTL4	Analyzing																																																															
TID	ITEMS BOUGHT																																																																														
T100	{M, O, N, K, E, Y }																																																																														
T200	{D, O, N, K, E, Y }																																																																														
T300	{M, A K, E }																																																																														
T400	{M,U, C, K,Y }																																																																														
T500	{C, O, O,K, I, E }																																																																														
16	You are given a dataset of people's characteristics, and your task is to classify whether a person will buy a computer (Yes/No) using 1R algorithm. <table><tr><th>Age</th><th>Income</th><th>Student</th><th>Credit Rating</th><th>Buy Computer</th></tr><tr><td>Youth</td><td>High</td><td>No</td><td>Fair</td><td>No</td></tr><tr><td>Youth</td><td>High</td><td>No</td><td>Excellent</td><td>No</td></tr><tr><td>Middle</td><td>High</td><td>No</td><td>Fair</td><td>Yes</td></tr><tr><td>Senior</td><td>Medium</td><td>No</td><td>Fair</td><td>Yes</td></tr><tr><td>Senior</td><td>Low</td><td>Yes</td><td>Fair</td><td>Yes</td></tr><tr><td>Senior</td><td>Low</td><td>Yes</td><td>Excellent</td><td>No</td></tr><tr><td>Middle</td><td>Low</td><td>Yes</td><td>Excellent</td><td>Yes</td></tr><tr><td>Youth</td><td>Medium</td><td>No</td><td>Fair</td><td>No</td></tr><tr><td>Youth</td><td>Low</td><td>Yes</td><td>Fair</td><td>Yes</td></tr><tr><td>Senior</td><td>Medium</td><td>Yes</td><td>Fair</td><td>Yes</td></tr><tr><td>Youth</td><td>Medium</td><td>Yes</td><td>Excellent</td><td>Yes</td></tr><tr><td>Middle</td><td>Medium</td><td>No</td><td>Excellent</td><td>Yes</td></tr><tr><td>Middle</td><td>High</td><td>Yes</td><td>Fair</td><td>Yes</td></tr><tr><td>Senior</td><td>Medium</td><td>No</td><td>Excellent</td><td>No</td></tr></table>	Age	Income	Student	Credit Rating	Buy Computer	Youth	High	No	Fair	No	Youth	High	No	Excellent	No	Middle	High	No	Fair	Yes	Senior	Medium	No	Fair	Yes	Senior	Low	Yes	Fair	Yes	Senior	Low	Yes	Excellent	No	Middle	Low	Yes	Excellent	Yes	Youth	Medium	No	Fair	No	Youth	Low	Yes	Fair	Yes	Senior	Medium	Yes	Fair	Yes	Youth	Medium	Yes	Excellent	Yes	Middle	Medium	No	Excellent	Yes	Middle	High	Yes	Fair	Yes	Senior	Medium	No	Excellent	No	16	BTL5	Evaluating
Age	Income	Student	Credit Rating	Buy Computer																																																																											
Youth	High	No	Fair	No																																																																											
Youth	High	No	Excellent	No																																																																											
Middle	High	No	Fair	Yes																																																																											
Senior	Medium	No	Fair	Yes																																																																											
Senior	Low	Yes	Fair	Yes																																																																											
Senior	Low	Yes	Excellent	No																																																																											
Middle	Low	Yes	Excellent	Yes																																																																											
Youth	Medium	No	Fair	No																																																																											
Youth	Low	Yes	Fair	Yes																																																																											
Senior	Medium	Yes	Fair	Yes																																																																											
Youth	Medium	Yes	Excellent	Yes																																																																											
Middle	Medium	No	Excellent	Yes																																																																											
Middle	High	Yes	Fair	Yes																																																																											
Senior	Medium	No	Excellent	No																																																																											

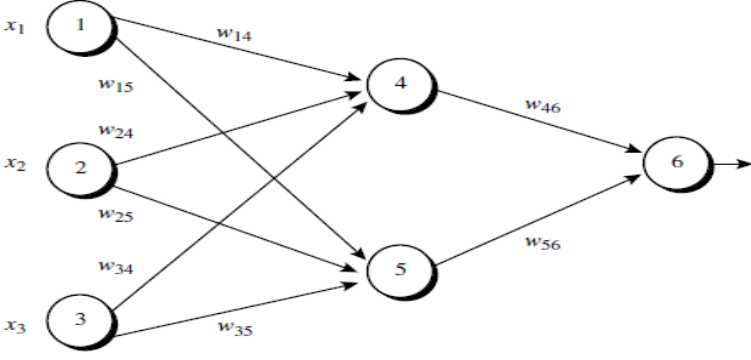
17	Construct the decision tree for the following training dataset using decision tree algorithm. Predict whether a person buys a car based on age, income, and student status.	16	BTL4	Analyzing																																				
<table><tr><th>Age</th><th>Income</th><th>Student</th><th>Buys Car?</th></tr><tr><td>Young</td><td>High</td><td>No</td><td>No</td></tr><tr><td>Young</td><td>High</td><td>Yes</td><td>No</td></tr><tr><td>Middle-aged</td><td>High</td><td>No</td><td>Yes</td></tr><tr><td>Old</td><td>Medium</td><td>No</td><td>Yes</td></tr><tr><td>Old</td><td>Low</td><td>Yes</td><td>No</td></tr><tr><td>Old</td><td>Low</td><td>No</td><td>No</td></tr><tr><td>Middle-aged</td><td>Low</td><td>No</td><td>Yes</td></tr><tr><td>Young</td><td>Medium</td><td>No</td><td>No</td></tr></table>					Age	Income	Student	Buys Car?	Young	High	No	No	Young	High	Yes	No	Middle-aged	High	No	Yes	Old	Medium	No	Yes	Old	Low	Yes	No	Old	Low	No	No	Middle-aged	Low	No	Yes	Young	Medium	No	No
Age	Income	Student	Buys Car?																																					
Young	High	No	No																																					
Young	High	Yes	No																																					
Middle-aged	High	No	Yes																																					
Old	Medium	No	Yes																																					
Old	Low	Yes	No																																					
Old	Low	No	No																																					
Middle-aged	Low	No	Yes																																					
Young	Medium	No	No																																					

UNIT 3- CLASSIFICATION ALGORITHMS II

SYLLABUS: Bayesian Belief Networks - Classification by Back propagation - Support Vector Machines - Classification Using Frequent Patterns - k-Nearest-Neighbour Classifiers - Case-Based Reasoning- Multiclass Classification - Semi-Supervised Classification- Mining Time series Data, Periodicity Analysis for time related sequence data.

PART-A

Q.No.	Questions	Level	Competence
1	List the applications of SVM.	BTL1	Remembering
2	What is Maximum Marginal Hyperplane(MMH) in SVM?	BTL1	Remembering
3	What are called support vectors in Support Vector Machine?	BTL1	Remembering
4	Write the formula for Euclidean distance between the 2 points.	BTL1	Remembering
5	How will you determine the good value of K in K-nearest neighbour?	BTL2	Understanding
6	How you will generate association rules from frequent itemsets.	BTL2	Understanding
7	How would you evaluate accuracy of a classifier?	BTL1	Remembering
8	Give a brief about Bayesian Belief Networks.	BTL2	Understanding
9	Given unit j in a hidden or output layer, write the formula for calculating net input I _j to unit j?	BTL2	Understanding
10	Given a net input I _j , write the formula for output of unit j?	BTL2	Understanding
11	For unit j in the output layer, How the Error is calculated?	BTL2	Understanding
12	For unit j in the hidden layer, How the Error is calculated?	BTL2	Understanding
13	Define Back propagation.	BTL1	Remembering
14	Define support vector machine with example.	BTL1	Remembering
15	How would you show your understanding about Multi class classification?	BTL2	Understanding
16	What is meant by Case based reasoning? Give examples.	BTL1	Remembering
17	What is meant by learning rate?	BTL1	Remembering
18	Apply the concept of back propagation to analyze, why there is a need to update the weights of connections from output layer through hidden layer?	BTL2	Understanding
19	Define Lazy learners with an example.	BTL1	Remembering
20	What are eager learners?	BTL1	Remembering
21	Write down the difference between Lazy learners and Eager learners.	BTL2	Understanding
22	What is meant by periodicity analysis?	BTL2	Understanding

23	Analyse about feed forward neural networks.	BTL2	Understanding																													
24	Why pruning is needed in decision tree?	BTL1	Remembering																													
	PART B	Marks	Level	Competence																												
1.	Calculate the Error at the nodes. Accordingly update the weights and bias of the nodes in the following network using Back propagation algorithm.  An example of a multilayer feed-forward neural network. Initial input, weight, and bias values. <table><tr><th>x_1</th><th>x_2</th><th>x_3</th><th>w_{14}</th><th>w_{15}</th><th>w_{24}</th><th>w_{25}</th><th>w_{34}</th><th>w_{35}</th><th>w_{46}</th><th>w_{56}</th><th>θ_4</th><th>θ_5</th><th>θ_6</th></tr><tr><td>1</td><td>0</td><td>1</td><td>0.2</td><td>-0.3</td><td>0.4</td><td>0.1</td><td>-0.5</td><td>0.2</td><td>-0.3</td><td>-0.2</td><td>-0.4</td><td>0.2</td><td>0.1</td></tr></table>	x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6	1	0	1	0.2	-0.3	0.4	0.1	-0.5	0.2	-0.3	-0.2	-0.4	0.2	0.1	16	BTL5	Evaluating
x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6																			
1	0	1	0.2	-0.3	0.4	0.1	-0.5	0.2	-0.3	-0.2	-0.4	0.2	0.1																			
2.	Explain in detail about classification using frequent patterns.	16	BTL3	Applying																												
3.	Describe in detail about Mining of time series data.	16	BTL3	Applying																												
4.	Explain in detail about semi supervised classification.	16	BTL3	Applying																												
5.	Explain in detail about the cases when data is linearly separable and data is linearly inseparable using SVM classifier.	16	BTL4	Analyzing																												
6.	Explain in detail about Multi-class classification with suitable examples.	16	BTL3	Applying																												
7.	(i) Examine in detail about Lazy learners with examples. (ii) Describe about the process of multi-layer feed-forward neural network classification using back propagation learning.	6 10	BTL4	Analyzing																												
8.	Explain in detail about the K-Nearest Neighbor classifier in detail.	16	BTL4	Analyzing																												
9.	Find all frequent item sets for the given training set using FP growth. Take minimum support as 3 and minimum confidence threshold as 80% . TID ITEMS BOUGHT T1 I1, I2, I3 T2 I2,I3,I4 T3 I4,I5 T4 I1,I2,I4 T5 I1,I2,I3,I5 T6 I1,I2,I3,I4	16	BTL5	Evaluating																												
10.	Explain in detail about the Case Based Reasoning in detail.	16	BTL3	Applying																												
11.	Explain in detail about the Periodicity Analysis for time related sequence data.	16	BTL4	Analyzing																												
12.	(i) With an example explain how support vector machines can be used for classification. (ii) What are the prediction techniques supported by data mining systems?	8 8	BTL6	Creating																												
13.	Analyze how the Bayesian Belief Networks are trained to perform classification.	16	BTL4	Analyzing																												
14.	Describe in detail about the following Classification methods. (i) Support Vector Machine.	8	BTL3	Applying																												

	(ii) Multi-layer feed forward network.	8		
15.	Apply the concept of lazy learners and its various types in detail.	16	BTL3	Applying
16.	Analyze in detail about multiclass classification.	16	BTL4	Analyzing
17.	Explain the back propagation technique with example.	16	BTL4	Analyzing

UNIT - IV CLUSTERING

SYLLABUS: Basic issues in clustering - First conceptual clustering system: Cluster/2 – Partitioning methods: k-means, expectation maximization (EM) - Hierarchical methods: distance-based agglomerative and divisible clustering - Conceptual clustering: Cobweb.

PART-A

Q.No.	Questions	Level	Competence
1	What is the impact of different distance metrics (e.g., Euclidean, Manhattan) on the final clusters in k-means? Provide an example scenario.	BTL2	Understanding
2	What is clustering?	BTL1	Remembering
3	What are the different types of data used for clustering?	BTL1	Remembering
4	What is a conceptual clustering system?	BTL1	Remembering
5	What is the difference between agglomerative and divisible hierarchical clustering?	BTL2	Understanding
6	Differentiate between single-linkage and complete-linkage clustering?	BTL1	Remembering
7	What is the COBWEB algorithm?	BTL2	Understanding
8	What are the main operations used by COBWEB?	BTL2	Understanding
9	What is the main goal of partitioning methods in clustering?	BTL1	Remembering
10	What is the time complexity of the k-means algorithm, and what factors affect it?	BTL2	Understanding
11	What is the key difference between hard clustering in k-means and soft clustering in EM?	BTL2	Understanding
12	What stopping criteria are typically used in the k-means algorithm?	BTL1	Remembering
13	How does single-linkage clustering compute the distance between two clusters?	BTL2	Understanding
14	How does agglomerative clustering build clusters from individual data points?	BTL2	Understanding
15	What is the main principle behind distance-based agglomerative clustering?	BTL1	Remembering
16	In COBWEB, how is the decision made to merge, split, or create a new cluster?	BTL1	Remembering
17	How would you handle the issue of empty clusters in k-means clustering?	BTL1	Remembering
18	How does the order of data input affect the final clustering structure in COBWEB? Can this be corrected?	BTL2	Understanding
19	Under what conditions might the EM algorithm get stuck in a local maximum? Propose a solution.	BTL2	Understanding
20	If the number of clusters is not known beforehand, how would you modify k-means to automatically determine the optimal number of clusters?	BTL1	Remembering
21	In a high-dimensional dataset, what are the challenges k-means faces, and how would you address them?	BTL1	Remembering
22	What is Cluster/2?	BTL2	Understanding
23	Why Cluster/2 is considered a significant step in conceptual clustering?	BTL2	Understanding
24	What is the main advantage of using Cluster/2 for symbolic data?	BTL2	Understanding

Q.No.	PART B	Marks	Level	Competence																																				
1	Explain the K-Means clustering algorithm in detail. Discuss the steps involved in the algorithm, how the centroid is updated, and the criteria for convergence. Additionally, highlight some of the advantages and limitations of the K-Means algorithm.	16	BTL3	Applying																																				
2	(i) Discuss the basic issues in clustering. (ii) Explain the challenges related to the selection of the number of clusters, the handling of different data types, the choice of distance metrics, and the scalability of clustering algorithms.	8 8	BTL4	Analyzing																																				
3	Use k means clustering to cluster the following data into two groups. Data points: {2,4,10,12,3,20,30,11,25}	16	BTL5	Evaluating																																				
4	(i) Explain the concept of the Cluster/2 algorithm, the first conceptual clustering system. (ii) Discuss how it works, its key features, and the steps involved in the clustering process. Additionally, describe its advantages and limitations. Provide examples where Cluster/2 could be applied effectively and discuss potential challenges in using this system for clustering complex datasets.	8 8	BTL3	Applying																																				
5	Explain the K-Means clustering algorithm as a partitioning method. Describe the detailed steps involved, how the clusters are formed, and the process of updating centroids. Additionally, discuss the factors that influence the choice of K, the algorithm's advantages, and its limitations.	16	BTL4	Analyzing																																				
6	Explain the concept of clustering in data analysis. Discuss the different types of clustering methods (e.g., hierarchical, partitioning, density-based), their key characteristics, and when each method is most appropriate to use.	16	BTL4	Analyzing																																				
7	Given the following dataset of points in a 2D plane, apply the k-means algorithm to cluster them into k=3 clusters. Show all steps of the algorithm, including the initialization, distance calculations, and centroid updates until convergence. Dataset: (2,10),(2,5),(8,4),(5,8),(7,5),(6,4),(1,2),(4,9)	16	BTL5	Evaluating																																				
8	Explain the concepts of distance-based agglomerative clustering and divisible clustering. Describe the key steps involved in each method, highlighting how clusters are formed and merged (or split).	16	BTL3	Applying																																				
9	Given the dataset {a, b, c, d, e} and the following matrix. <table><tr><td></td><td>a</td><td>b</td><td>c</td><td>d</td><td>e</td></tr><tr><td>a</td><td>0</td><td>9</td><td>3</td><td>6</td><td>11</td></tr><tr><td>b</td><td>9</td><td>0</td><td>7</td><td>5</td><td>10</td></tr><tr><td>c</td><td>3</td><td>7</td><td>0</td><td>9</td><td>2</td></tr><tr><td>d</td><td>6</td><td>5</td><td>9</td><td>0</td><td>8</td></tr><tr><td>e</td><td>11</td><td>10</td><td>2</td><td>8</td><td>0</td></tr></table> Construct a dendrogram by complete linkage hierarchical clustering using agglomerative method.		a	b	c	d	e	a	0	9	3	6	11	b	9	0	7	5	10	c	3	7	0	9	2	d	6	5	9	0	8	e	11	10	2	8	0	16	BTL6	Creating
	a	b	c	d	e																																			
a	0	9	3	6	11																																			
b	9	0	7	5	10																																			
c	3	7	0	9	2																																			
d	6	5	9	0	8																																			
e	11	10	2	8	0																																			
10	Given a one-dimensional data set {1, 5, 8, 10, 2}, use the agglomerative clustering algorithms with the complete linkage with	16	BTL5	Evaluating																																				

	Euclidean distance to establish a hierarchal grouping relationship. By using the cutting threshold of 5, how many clusters are there?																																													
11	Given the dataset {a, b, c, d, e} and the corresponding distance matrix, apply the Divisive Analysis (DIANA) clustering algorithm to divide the dataset into clusters. Explain each step of the process and show how clusters are formed at each stage. <table><tr><td></td><td>a</td><td>b</td><td>c</td><td>d</td><td>e</td></tr><tr><td>a</td><td>0</td><td>9</td><td>3</td><td>6</td><td>11</td></tr><tr><td>b</td><td>9</td><td>0</td><td>7</td><td>5</td><td>10</td></tr><tr><td>c</td><td>3</td><td>7</td><td>0</td><td>9</td><td>2</td></tr><tr><td>d</td><td>6</td><td>5</td><td>9</td><td>0</td><td>8</td></tr><tr><td>e</td><td>11</td><td>10</td><td>2</td><td>8</td><td>0</td></tr></table>		a	b	c	d	e	a	0	9	3	6	11	b	9	0	7	5	10	c	3	7	0	9	2	d	6	5	9	0	8	e	11	10	2	8	0	16	BTL4	Analyzing						
	a	b	c	d	e																																									
a	0	9	3	6	11																																									
b	9	0	7	5	10																																									
c	3	7	0	9	2																																									
d	6	5	9	0	8																																									
e	11	10	2	8	0																																									
12	Imagine we have a dataset about animals with attributes such as "has fur," "lays eggs," "can fly," "has feathers," and "is warm-blooded." The goal is to group animals into meaningful clusters based on these attributes using conceptual clustering. Dataset: <table><tr><th>Animal</th><th>Has Fur</th><th>Lays Eggs</th><th>Can Fly</th><th>Has Feathers</th><th>Is Warm-Blooded</th></tr><tr><td>Dog</td><td>Yes</td><td>No</td><td>No</td><td>No</td><td>Yes</td></tr><tr><td>Cat</td><td>Yes</td><td>No</td><td>No</td><td>No</td><td>Yes</td></tr><tr><td>Sparrow</td><td>No</td><td>Yes</td><td>Yes</td><td>Yes</td><td>Yes</td></tr><tr><td>Penguin</td><td>No</td><td>Yes</td><td>No</td><td>Yes</td><td>Yes</td></tr><tr><td>Crocodile</td><td>No</td><td>Yes</td><td>No</td><td>No</td><td>No</td></tr><tr><td>Snake</td><td>No</td><td>Yes</td><td>No</td><td>No</td><td>No</td></tr></table>	Animal	Has Fur	Lays Eggs	Can Fly	Has Feathers	Is Warm-Blooded	Dog	Yes	No	No	No	Yes	Cat	Yes	No	No	No	Yes	Sparrow	No	Yes	Yes	Yes	Yes	Penguin	No	Yes	No	Yes	Yes	Crocodile	No	Yes	No	No	No	Snake	No	Yes	No	No	No	16	BTL6	Creating
Animal	Has Fur	Lays Eggs	Can Fly	Has Feathers	Is Warm-Blooded																																									
Dog	Yes	No	No	No	Yes																																									
Cat	Yes	No	No	No	Yes																																									
Sparrow	No	Yes	Yes	Yes	Yes																																									
Penguin	No	Yes	No	Yes	Yes																																									
Crocodile	No	Yes	No	No	No																																									
Snake	No	Yes	No	No	No																																									
13	Using the COBWEB algorithm, which processes data incrementally, build clusters for the given dataset of animals? The dataset has three attributes—Size (Small or large), Type (Mammal or Bird), and Diet (Herbivore or Carnivore)—for five animals: Rabbit, Wolf, Deer, Sparrow, and Hawk. Calculate the Category Utility (CU) after each step, as each animal is added to the clusters. Explain each step and the CU calculation. <table><tr><th>Animal</th><th>Size</th><th>Type</th><th>Diet</th></tr><tr><td>Rabbit</td><td>Small</td><td>Mammal</td><td>Herbivore</td></tr><tr><td>Wolf</td><td>Large</td><td>Mammal</td><td>Carnivore</td></tr><tr><td>Deer</td><td>Large</td><td>Mammal</td><td>Herbivore</td></tr><tr><td>Sparrow</td><td>Small</td><td>Bird</td><td>Herbivore</td></tr><tr><td>Hawk</td><td>Small</td><td>Bird</td><td>Carnivore</td></tr></table>	Animal	Size	Type	Diet	Rabbit	Small	Mammal	Herbivore	Wolf	Large	Mammal	Carnivore	Deer	Large	Mammal	Herbivore	Sparrow	Small	Bird	Herbivore	Hawk	Small	Bird	Carnivore	16	BTL4	Analyzing																		
Animal	Size	Type	Diet																																											
Rabbit	Small	Mammal	Herbivore																																											
Wolf	Large	Mammal	Carnivore																																											
Deer	Large	Mammal	Herbivore																																											
Sparrow	Small	Bird	Herbivore																																											
Hawk	Small	Bird	Carnivore																																											
14	Explain the COBWEB conceptual clustering algorithm. Discuss how it constructs clusters based on the hierarchical organization of data and how it uses a tree structure to represent clusters.	16	BTL3	Applying																																										
15	Given the dataset {a, b, c, d, e} and the following matrix.	16	BTL4	Analyzing																																										

	<table><tr><td></td><td>a</td><td>b</td><td>c</td><td>d</td><td>e</td></tr><tr><td>a</td><td>0</td><td>9</td><td>3</td><td>6</td><td>11</td></tr><tr><td>b</td><td>9</td><td>0</td><td>7</td><td>5</td><td>10</td></tr><tr><td>c</td><td>3</td><td>7</td><td>0</td><td>9</td><td>2</td></tr><tr><td>d</td><td>6</td><td>5</td><td>9</td><td>0</td><td>8</td></tr><tr><td>e</td><td>11</td><td>10</td><td>2</td><td>8</td><td>0</td></tr></table> Construct a dendrogram by Single-linkage hierarchical using agglomerative method.		a	b	c	d	e	a	0	9	3	6	11	b	9	0	7	5	10	c	3	7	0	9	2	d	6	5	9	0	8	e	11	10	2	8	0			
	a	b	c	d	e																																			
a	0	9	3	6	11																																			
b	9	0	7	5	10																																			
c	3	7	0	9	2																																			
d	6	5	9	0	8																																			
e	11	10	2	8	0																																			
16	Given a dataset of five points on a 2D plane: (1,1),(2,1),(4,3),(5,4),(3,2), apply the k-means clustering algorithm to group these points into k=2 clusters. 1. What are the initial centroids if we choose (1,1) and (5,4) as starting points? 2. After the first iteration, what are the new centroids? 3. Repeat the process until the centroids no longer change. What are the final cluster centroids, and which points belong to each cluster?	16	BTL4	Analyzing																																				
17	Given the dataset with six points, P1 to P6, and their coordinates (X,Y) as shown in the table, apply the agglomerative clustering algorithm to group these points until only one cluster remains. <table><tr><td></td><td>X</td><td>Y</td></tr><tr><td>P1</td><td>0.40</td><td>0.53</td></tr><tr><td>P2</td><td>0.22</td><td>0.38</td></tr><tr><td>P3</td><td>0.35</td><td>0.32</td></tr><tr><td>P4</td><td>0.26</td><td>0.19</td></tr><tr><td>P5</td><td>0.08</td><td>0.41</td></tr><tr><td>P6</td><td>0.45</td><td>0.30</td></tr></table>		X	Y	P1	0.40	0.53	P2	0.22	0.38	P3	0.35	0.32	P4	0.26	0.19	P5	0.08	0.41	P6	0.45	0.30	16	BTL5	Evaluating															
	X	Y																																						
P1	0.40	0.53																																						
P2	0.22	0.38																																						
P3	0.35	0.32																																						
P4	0.26	0.19																																						
P5	0.08	0.41																																						
P6	0.45	0.30																																						

UNIT V OUTLIER DETECTION

SYLLABUS: Outliers and Outlier Analysis, Outlier Detection Methods, Statistical Approaches, Proximity Based Approaches, Clustering-Based Approaches, Classification-Based Approaches, Mining Contextual and Collective Outliers, Outlier Detection in High-Dimensional Data.

PART-A

Q.No.	Questions	Level	Competence
1	How does identification of outliers aid in mining?	BTL2	Understanding
2	What is an outlier? Mention its applications.	BTL1	Remembering
3	Mention the methods of detecting outliers.	BTL1	Remembering
4	Differentiate between global outliers and contextual outliers.	BTL1	Remembering
5	Mention the role of outlier analysis in anomaly detection.	BTL1	Remembering
6	How does a classification-based approach detect outliers in a dataset?	BTL1	Remembering

7	What is the Interquartile Range (IQR)-based method for outlier detection, and why is it considered robust?	BTL2	Understanding
8	What is the Z-Score method for statistical outlier detection, and how does it identify outliers?	BTL1	Remembering
9	How Grubbs' test detects outliers?	BTL2	Understanding
10	How is the Interquartile Range (IQR) used to define outlier thresholds, and what are the criteria for identifying outliers?	BTL2	Understanding
11	Why is the Interquartile Range (IQR) method commonly used for outlier detection?	BTL1	Remembering
12	What is the difference between supervised and unsupervised classification-based outlier detection?	BTL2	Understanding
13	How does the k-nearest neighbours (k-NN) method detect outliers?	BTL2	Understanding
14	What is density-based outlier detection?	BTL2	Understanding
15	How does the choice of proximity measure impact outlier detection?	BTL1	Remembering
16	What is the fundamental concept behind proximity-based outlier detection methods?	BTL2	Understanding
17	What is the difference between centroid-based and density-based clustering methods in outlier detection?	BTL2	Understanding
18	How does a classification model identify outliers in labeled datasets?	BTL1	Remembering
19	What is the DBSCAN algorithm, and how does it identify outliers?	BTL1	Remembering
20	How can the concept of density help in detecting outliers using clustering methods?	BTL2	Understanding
21	How can clustering techniques be used for outlier detection?	BTL2	Understanding
22	In a dataset with 200 dimensions, applying a proximity-based approach is resulting in almost all points appearing as outliers. What concept could explain this phenomenon?	BTL1	Remembering
23	You discover outliers in your dataset due to entry errors where decimal points are misplaced (e.g., 10000 instead of 100.00). What is an effective way to address this issue?	BTL1	Remembering
24	In a social network dataset, you find a small group of users forming connections in an unusual pattern over a short period, unlike the general user behavior. How would you categorize these connections?	BTL2	Understanding

PART-B

Q.No	Questions	Marks	Level	Competence
1	i) What are outliers in a dataset, and why is it important to perform outlier analysis? ii) Explain the different types of outliers that can be identified in a dataset (e.g., point outliers, contextual outliers, and collective outliers).	8 8	BTL4	Analyzing
2	Describe at least three methods of outlier detection. For each method, explain how it works and give an example of a scenario where it might be useful.	16	BTL3	Applying
3	i) Explain about the various statistical methods of outlier detection. ii) Using the Z-score method, determine if there are any outliers in the following dataset of exam scores: Scores = [50, 55, 60, 65, 70, 100]	6 10	BTL4	Analyzing

4	Using the Z-score method, determine if there are any outliers in the following dataset of daily sales revenue for a retail store over the past 30 days. {100,150,120,125,140,130,110,135,130,150,140,100,95,80,120,125,130,100,140,135,130,145,110,120,130,135,140,125,130,120}	16	BTL5	Evaluating																																			
5	Consider the following dataset representing the monthly number of visitors (in thousands) to a website: Visitors = [45, 60, 50, 80, 120, 75, 90, 95, 200, 400] Calculate the first quartile (Q1), third quartile (Q3), and the Interquartile Range (IQR) for the dataset. Are there any outliers in this dataset? If so, which data points are outliers and why?	16	BTL4	Analyzing																																			
6	Given the following 2D dataset of points: [(1,2),(2,3),(3,4),(4,5),(100,100)] Explain how you would identify the anomalous point (outlier) in this dataset using Euclidean distance-based anomaly detection. Provide the step-by-step process and reasoning for your answer.	16	BTL5	Evaluating																																			
7	Given the following 2D dataset of points: [(1,2),(2,2),(2,3),(3,3),(10,10)] Explain how you would identify the anomalous point (outlier) in this dataset using Euclidean distance-based anomaly detection.	16	BTL5	Evaluating																																			
8	Data Points: P1: (3, 7) P2: (4, 6) P3: (5, 5) P4: (6, 4) P5: (7, 3) P6: (6, 2) P7: (7, 2) P8: (8, 4) P9: (3, 3) P10: (2, 6) P11: (3, 5) P12: (2, 4) Apply the DBSCAN algorithm to the given data points and Create the clusters with minPts = 4 and epsilon (ϵ) = 1.9	16	BTL6	Creating																																			
9	(i) Explain the working of the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm. How does it identify core points, border points, and noise (outliers)?	8	BTL4	Analyzing																																			
	(ii) Given the similarity matrix below, apply the DBSCAN algorithm with a similarity threshold of 0.8 and MinPts = 2. Identify the core points, border points, and noise (outliers) in the dataset. Explain your reasoning. <table><tr><td></td><td>P1</td><td>P2</td><td>P3</td><td>P4</td><td>P5</td></tr><tr><td>P1</td><td>1.00</td><td>0.10</td><td>0.41</td><td>0.55</td><td>0.35</td></tr><tr><td>P2</td><td>0.10</td><td>1.00</td><td>0.64</td><td>0.47</td><td>0.98</td></tr><tr><td>P3</td><td>0.41</td><td>0.64</td><td>1.00</td><td>0.44</td><td>0.85</td></tr><tr><td>P4</td><td>0.55</td><td>0.47</td><td>0.44</td><td>1.00</td><td>0.76</td></tr><tr><td>P5</td><td>0.35</td><td>0.98</td><td>0.85</td><td>0.76</td><td>1.00</td></tr></table>		P1	P2	P3	P4	P5	P1	1.00	0.10	0.41	0.55	0.35	P2	0.10	1.00	0.64	0.47	0.98	P3	0.41	0.64	1.00	0.44	0.85	P4	0.55	0.47	0.44	1.00	0.76	P5	0.35	0.98	0.85	0.76	1.00	8	BTL4
	P1	P2	P3	P4	P5																																		
P1	1.00	0.10	0.41	0.55	0.35																																		
P2	0.10	1.00	0.64	0.47	0.98																																		
P3	0.41	0.64	1.00	0.44	0.85																																		
P4	0.55	0.47	0.44	1.00	0.76																																		
P5	0.35	0.98	0.85	0.76	1.00																																		
10	i) Explain the concept of clustering-based approaches for anomaly detection. ii) Consider the following 2D data points: (1,2),(2,3),(3,3),(5,5),(6,5),(8,8),(10,10) Perform outlier detection using k-means clustering with k=2.	6 10	BTL4	Analyzing																																			

11	i) Consider the following dataset representing the values of a sample: Dataset = {50, 55, 60, 65, 70, 100} Discuss how the calculation of Q1, Q3, and IQR helps in identifying potential outliers in the data. ii) Given the following 2D dataset of points: [(1,2),(2,3),(3,4),(4,5),(100,100)] Explain how you would identify the anomalous point (outlier) in this dataset using Euclidean distance-based anomaly detection. Provide the step-by-step process and reasoning for your answer.	8	BTL3	Applying
12	Discuss about the density based methods of outlier detection.	16	BTL4	Analyzing
13	Discuss about the classification based approaches of outlier detection.	13	BTL5	Evaluating
14	(i) Explain what contextual outliers are and how they differ from global outliers. (ii) Provide an example of a dataset where contextual outliers might occur, such as time-series data or seasonal data.	8	BTL4	Analyzing
15	How can clustering techniques like DBSCAN or K-means be used to detect collective outliers in a dataset? Provide an example scenario where collective outliers might be relevant.	16	BTL3	Applying
16	Explain the challenges associated with outlier detection in high-dimensional data. What issues arise from having too many dimensions?	16	BTL3	Applying
17	(i) How can contextual and collective outliers be detected simultaneously in high-dimensional datasets? (ii) Provide an example of a real-world problem where contextual and collective outliers might exist in high-dimensional data (e.g., in customer behavior, medical data, or sensor networks).	8	BTL4	Analyzing

STAFF INCHARGE

VERIFIED BY

HOD