



VIT

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

REG.NO.:

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
WINTER SEMESTER 2024-2025

SLOT: F2

Programme Name & Branch : B.Tech- CSE, BDS
Course Code and Course Name : BCSE208L - Data Mining
Faculty Name(s) : Dr. Chellatamilan T, Dr. Lavanya K
Class Number(s) : VL2024250501792, VL2024250501801
Date of Examination : 21-March-25
Exam Duration : 90 minutes
Maximum Marks: 50

General instruction(s):

Answer All Questions

M - Max mark; CO - Course Outcome; BL - Blooms Taxonomy Level (1 - Remember, 2 - Understand, 3 - Apply, 4 - Analyse, 5 - Evaluate, 6 - Create)

Course Outcomes (Type the CO statements covered in this question paper. Use the CO number as per the syllabus copy)

CO3. Discover interesting patterns from large amounts of data using Association Rule Mining.

CO4. Extract useful information from the labeled data using various classifiers and Compile unlabeled data into clusters applying various clustering algorithms.

unlabeled data into clusters applying various clustering algorithms.		M	CO	BL																
Q. No	Question																			
1.	A healthcare centre aims to analyse symptom patterns among patients to identify frequent symptom combinations. This analysis can help in early diagnosis, clustering diseases, or planning better treatments. The dataset provided includes records of patients along with the symptoms they reported. Your task is to use the FP-Growth algorithm to find frequent symptom sets through the built FP-Growth tree based on the transactions provided considering the minimum support count threshold as 2. <table><tr><th>Patient ID</th><th>Symptoms</th></tr><tr><td>1</td><td>Headache, Fever, Cold</td></tr><tr><td>2</td><td>Smell Loss, Fever, Diarrhea</td></tr><tr><td>3</td><td>Headache, Smell Loss, Fever, Diarrhea</td></tr><tr><td>4</td><td>Smell Loss, Diarrhea</td></tr><tr><td>5</td><td>Body Pain</td></tr></table>	Patient ID	Symptoms	1	Headache, Fever, Cold	2	Smell Loss, Fever, Diarrhea	3	Headache, Smell Loss, Fever, Diarrhea	4	Smell Loss, Diarrhea	5	Body Pain	10	CO3	3				
Patient ID	Symptoms																			
1	Headache, Fever, Cold																			
2	Smell Loss, Fever, Diarrhea																			
3	Headache, Smell Loss, Fever, Diarrhea																			
4	Smell Loss, Diarrhea																			
5	Body Pain																			
2.	You are a data analyst for a retail company called "ShopSmart," and you want to predict whether a customer is loyal based on two categorical attributes: "Membership Status" (Gold, Silver, Bronze) and "Purchase Frequency" (High, Medium, Low). The label attribute "Loyalty" indicates whether the customer is loyal (Yes) or not (No). Apply the CART algorithm using the Gini index to construct the decision tree <table><tr><th>Customer</th><th>Membership Status</th><th>Purchase Frequency</th><th>Loyalty</th></tr><tr><td>C1</td><td>Gold</td><td>High ↗</td><td>Yes</td></tr><tr><td>C2</td><td>Silver</td><td>Medium ↘</td><td>No</td></tr><tr><td>C3</td><td>Bronze ↘</td><td>Low ↗</td><td>No</td></tr></table>	Customer	Membership Status	Purchase Frequency	Loyalty	C1	Gold	High ↗	Yes	C2	Silver	Medium ↘	No	C3	Bronze ↘	Low ↗	No	10	CO4	4
Customer	Membership Status	Purchase Frequency	Loyalty																	
C1	Gold	High ↗	Yes																	
C2	Silver	Medium ↘	No																	
C3	Bronze ↘	Low ↗	No																	



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
WINTER SEMESTER 2024-2025

SLOT: F2

		<table><tr><td>C4</td><td>Gold ✓</td><td>Medium ✓</td><td>Yes ✓</td></tr><tr><td>C5</td><td>Silver ✓</td><td>High ✓</td><td>Yes</td></tr><tr><td>C6</td><td>Bronze ✓</td><td>Medium ✓</td><td>No</td></tr><tr><td>C7</td><td>Gold ✓</td><td>Low ✓</td><td>No</td></tr><tr><td>C8</td><td>Silver ✓</td><td>Medium ✓</td><td>No</td></tr><tr><td>C9</td><td>Bronze ✓</td><td>High ✓</td><td>Yes</td></tr><tr><td>C10</td><td>Gold ✓</td><td>High ✓</td><td>Yes</td></tr></table> Let's consider a test case to traverse the constructed decision tree and determine the loyalty status of a customer based on the attributes "Membership Status (Silver)" and "Purchase Frequency(High)."	C4	Gold ✓	Medium ✓	Yes ✓	C5	Silver ✓	High ✓	Yes	C6	Bronze ✓	Medium ✓	No	C7	Gold ✓	Low ✓	No	C8	Silver ✓	Medium ✓	No	C9	Bronze ✓	High ✓	Yes	C10	Gold ✓	High ✓	Yes									
C4	Gold ✓	Medium ✓	Yes ✓																																				
C5	Silver ✓	High ✓	Yes																																				
C6	Bronze ✓	Medium ✓	No																																				
C7	Gold ✓	Low ✓	No																																				
C8	Silver ✓	Medium ✓	No																																				
C9	Bronze ✓	High ✓	Yes																																				
C10	Gold ✓	High ✓	Yes																																				
3.	A binary classification result, where the correct labels are [T, T, F, T, F, T, F, T] and the predicted labels are [T, F, T, T, F, F, F, T]. Assume T means "true" (the desired class) and F ("false") is the "default" class. Compute accuracy, recall, precision, and F1 for T. Write down the formula for each and draw the Confusion matrix. Highlight portions of distributions of FP,TP,TN and FN using a chart of coverage of each.			10	CO4	3																																	
4.	A retail company wants to segment its customers based on their shopping behaviour. The company has collected transactional data for each customer, including their standardized total purchase amount(X) and the standardized average number of items bought per visit(Y). There might be clusters of customers with similar shopping patterns, while some customers may be outliers (e.g., big spenders or occasional shoppers).Apply the DBSCAN algorithm to cluster the following dataset using the Euclidean distance and parameters Eps=2 and MinPts =4. How many clusters will be formed and how many noise points will be identified? <table><tr><td>Customer</td><td>C1</td><td>C2</td><td>C3</td><td>C4</td><td>C5</td><td>C6</td><td>C7</td><td>C8</td><td>C9</td><td>C10</td></tr><tr><td>X</td><td>0</td><td>1</td><td>2</td><td>2</td><td>3</td><td>3</td><td>4</td><td>6</td><td>5</td><td>6</td></tr><tr><td>Y</td><td>0</td><td>1</td><td>2</td><td>3</td><td>2</td><td>3</td><td>4</td><td>6</td><td>6</td><td>5</td></tr></table>			Customer	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	X	0	1	2	2	3	3	4	6	5	6	Y	0	1	2	3	2	3	4	6	6	5	10	CO4	5
Customer	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10																													
X	0	1	2	2	3	3	4	6	5	6																													
Y	0	1	2	3	2	3	4	6	6	5																													
5.	To analyse the effectiveness of the clustering, you need to determine the inter-cluster distance (distance between the centroids of the clusters) and the intra-cluster distance (average distance of points within each cluster to the centroid of that cluster). Use the following clusters assignment table and calculate each of the inter and intra cluster distance with plot visualization. <table><tr><td>Points</td><td>P1</td><td>P2</td><td>P3</td><td>P4</td><td>P5</td><td>P6</td></tr><tr><td>X</td><td>1</td><td>2</td><td>2</td><td>8</td><td>8</td><td>9</td></tr><tr><td>Y</td><td>2</td><td>2</td><td>3</td><td>7</td><td>8</td><td>7</td></tr><tr><td>Cluster Number</td><td>C1</td><td>C1</td><td>C1</td><td>C2</td><td>C2</td><td>C2</td></tr></table>			Points	P1	P2	P3	P4	P5	P6	X	1	2	2	8	8	9	Y	2	2	3	7	8	7	Cluster Number	C1	C1	C1	C2	C2	C2	10	CO4	4					
Points	P1	P2	P3	P4	P5	P6																																	
X	1	2	2	8	8	9																																	
Y	2	2	3	7	8	7																																	
Cluster Number	C1	C1	C1	C2	C2	C2																																	
