

**VIT**Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

REG.NO.:

SLOT: F1

Programme Name & Branch : BTECH & Common to all branches
 Course Code and Course Name : BCSE208L and Data Mining
 Faculty Name(s) : Sanjiban Sekhar Roy, Lavanya K and Santhi K
 Class Number(s) : VL2025260502546
 Date of Examination : 22/3/2026, TIME: 2PM -3:30PM
 SLOT : F1
 Exam Duration : 90 minutes

Maximum Marks: 50

General instruction(s): Answer all questions (5X10=50 MARKS)

Q. No	Question	Module	Marks	CO	BL												
1.	a) Consider the transaction database given below: <table><thead><tr><th>TID</th><th>Items Purchased</th></tr></thead><tbody><tr><td>T1</td><td>A, B, C</td></tr><tr><td>T2</td><td>A, C, D</td></tr><tr><td>T3</td><td>B, C, E</td></tr><tr><td>T4</td><td>A, B, C, E</td></tr><tr><td>T5</td><td>B, E</td></tr></tbody></table> Assume that the minimum support threshold is 40% and the minimum confidence threshold is 60%. Using the Apriori algorithm, generate all frequent item-sets that satisfy the minimum support threshold, clearly showing the candidate generation and pruning steps at each iteration. Based on the frequent item-sets obtained, derive all strong association rules that satisfy the minimum confidence threshold, and compute the support and confidence for each rule. b) Why does Apriori generate a large number of candidate item sets in dense datasets? Explain how pruning reduces the search space.	TID	Items Purchased	T1	A, B, C	T2	A, C, D	T3	B, C, E	T4	A, B, C, E	T5	B, E	4	(7+3)	3	4
TID	Items Purchased																
T1	A, B, C																
T2	A, C, D																
T3	B, C, E																
T4	A, B, C, E																
T5	B, E																
	A retail supermarket chain aims to analyse customer purchase behaviour using transaction data in order to identify frequently co-occurring product combinations for promotional bundling. The following table represents a transaction database, where each row corresponds to a customer transaction and the items purchased in that transaction.	4	10	3	4												

**VIT**Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

REG.NO.:

SLOT: F1

Transaction ID	Items Purchased
T1	Milk, Bread, Butter
T2	Bread, Diaper, Beer, Eggs
T3	Milk, Diaper, Beer, Cola
T4	Milk, Bread, Diaper, Beer
T5	Bread, Milk, Diaper, Cola

Assume the minimum support count = 2. Construct the FP-Tree for the above transaction database after determining the frequent 1-itemsets. Clearly show the header table, tree construction steps, and conditional pattern bases where required. Using the FP-Growth algorithm, generate all frequent item sets that satisfy the minimum support threshold.

| 3. | a) A financial institution wants to automate its loan approval process using historical applicant data. The bank classifies applicants into two categories: - Approved (Yes) - Rejected (No) Based on past records, the following dataset was collected: | Applicant ID | Income | Credit Score | Employment Status | Loan Status | |--------------|--------|--------------|-------------------|-------------| | 1 | High | Good | Employed | Yes | | 2 | High | Poor | Employed | No | | 3 | Medium | Good | Employed | Yes | | 4 | Low | Good | Unemployed | No | | 5 | Medium | Poor | Employed | No | | 6 | High | Good | Unemployed | Yes | | 7 | Low | Poor | Employed | No | | 8 | Medium | Good | Employed | Yes | | 5 | (8+2) |

**VIT**Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

A new applicant has the profile: Income = Medium, Credit Score = Good, and Employment Status = Employed. Using the given dataset, compute the prior probabilities for each class (Approved and Rejected), determine the likelihood probabilities of each attribute given the class, and apply the Naïve Bayes classifier to calculate the posterior probabilities. Based on the computed posterior values, classify whether the loan should be Approved or Rejected. Finally, state one key assumption of the Naïve Bayes classifier and critically comment on whether that assumption is realistic in this loan approval scenario.

b) Critically analyse how conditional independence assumptions in a Bayesian Belief Network reduce the complexity of joint probability computation. Discuss under what structural conditions (in terms of graph topology) this factorization becomes invalid or misleading for real-world causal modelling.

4. A dataset contains information about whether a student will Pass the Exam based on three attributes: Study Hours, Attendance, and Assignment Submission. The dataset is given below.

Student	Study Hours	Attendance	Assignment	Pas s
S1	High	Good	Yes	Yes
S2	High	Good	No	Yes
S3	Medium	Good	Yes	Yes
S4	Low	Poor	Yes	No
S5	Low	Poor	No	No
S6	Medium	Good	No	Yes
S7	Low	Good	Yes	Yes
S8	Medium	Poor	No	No

Using the ID3 Decision Tree algorithm, compute the entropy of the dataset and calculate the information gain for each attribute (Study Hours, Attendance, Assignment), and then determine which attribute should be selected as the root node of the decision tree, showing all the necessary calculations step-by-step and clearly identifying the attribute with the highest information gain as the best splitting attribute.

5 10 4 4

**VIT**Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

REG.NO.:

SLOT: F1

5. a) A retail analytics company is studying customer movement patterns in a small shopping zone. The goal is to identify representative customer groups and detect dense activity regions for targeted marketing and infrastructure planning. The (X, Y) coordinates represent customer locations in a 2D shopping area map (in km). The analytics team collected the following customer location coordinates:

Customer	X	Y
A	1	1
B	2	1
C	4	3
D	5	4
E	3	2

The company wants to divide the customers into 2 clusters ($k = 2$) and identify representative customers (medoids) using Manhattan distance. You are assigned as a data analyst to perform the clustering. The following tasks have to be accomplished.

1. Perform one full iteration of the K-Medoid algorithm (assume any two initial medoids clearly).
2. Compute the total clustering cost using Manhattan distance.
3. Identify the new medoids after cost minimization.

b) In a large dataset, explain when K-Medoids is preferred over K-Means, considering outliers, choice of distance measure, and interpretability of clusters.

6 (8+2) 4 5



VIT

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

REG.NO.: 238 DS0204

SLOT: F1

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
RE-CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

Programme Name & Branch : BTECH & Common to all branches
Course Code and Course Name : BCSE208L and Data Mining
Faculty Name(s) : Sanjiban Sekhar Roy, Lavanya K and Santhi K
Class Number(s) : VL2025260502546
SLOT : F1
Exam Duration : 90 minutes Maximum Marks: 50

General instruction(s): Answer all questions (5X10=50 MARKS)

Q. No	Question	Module	Marks	CO	BL														
1.	A supermarket maintains the following customer transaction database: <table><thead><tr><th>TID</th><th>Items Purchased</th></tr></thead><tbody><tr><td>T1</td><td>Bread, Milk, Eggs</td></tr><tr><td>T2</td><td>Bread, Butter, Eggs</td></tr><tr><td>T3</td><td>Milk, Eggs, Jam</td></tr><tr><td>T4</td><td>Bread, Milk, Eggs, Jam</td></tr><tr><td>T5</td><td>Bread, Butter</td></tr><tr><td>T6</td><td>Milk, Jam</td></tr></tbody></table> Minimum Support = 50% (count ≥ 3 out of 6 transactions) Minimum Confidence = 70%. Using the Apriori algorithm: (i) Generate all frequent itemsets showing candidate generation and pruning at each level. (ii) Derive all strong association rules satisfying both thresholds, showing support and confidence for each.	TID	Items Purchased	T1	Bread, Milk, Eggs	T2	Bread, Butter, Eggs	T3	Milk, Eggs, Jam	T4	Bread, Milk, Eggs, Jam	T5	Bread, Butter	T6	Milk, Jam	4	10	3	4
TID	Items Purchased																		
T1	Bread, Milk, Eggs																		
T2	Bread, Butter, Eggs																		
T3	Milk, Eggs, Jam																		
T4	Bread, Milk, Eggs, Jam																		
T5	Bread, Butter																		
T6	Milk, Jam																		
2.	Explain the FP-Growth algorithm for mining frequent item-sets. Your answer should address why Apriori fails on large datasets, how an FP-Tree is constructed using a header table and frequency-sorted transactions, how conditional pattern bases and conditional FP-Trees are used to extract all frequent item-sets without candidate generation, and under what conditions FP-Growth may degrade in performance.	4	10	3	4														



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
RE-CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

3/	The table given below represents the Height, Age values and the corresponding Weight of 10 individuals. Find the Weight of the 11th record using K-Nearest Neighbor regression algorithm with 2 different K values (K=3 and K=5). <table><tr><th>ID</th><th>Height</th><th>Age</th><th>Weight</th></tr><tr><td>1</td><td>5</td><td>45</td><td>77</td></tr><tr><td>2</td><td>5.11</td><td>26</td><td>47</td></tr><tr><td>3</td><td>5.6</td><td>30</td><td>55</td></tr><tr><td>4</td><td>5.9</td><td>34</td><td>59</td></tr><tr><td>5</td><td>4.8</td><td>40</td><td>72</td></tr><tr><td>6</td><td>5.8</td><td>36</td><td>60</td></tr><tr><td>7</td><td>5.3</td><td>19</td><td>40</td></tr><tr><td>8</td><td>5.8</td><td>28</td><td>60</td></tr><tr><td>9</td><td>5.5</td><td>23</td><td>45</td></tr><tr><td>10</td><td>5.6</td><td>32</td><td>58</td></tr><tr><td>11</td><td>5.5</td><td>38</td><td>?</td></tr></table>	ID	Height	Age	Weight	1	5	45	77	2	5.11	26	47	3	5.6	30	55	4	5.9	34	59	5	4.8	40	72	6	5.8	36	60	7	5.3	19	40	8	5.8	28	60	9	5.5	23	45	10	5.6	32	58	11	5.5	38	?	5	10	4	5
ID	Height	Age	Weight																																																		
1	5	45	77																																																		
2	5.11	26	47																																																		
3	5.6	30	55																																																		
4	5.9	34	59																																																		
5	4.8	40	72																																																		
6	5.8	36	60																																																		
7	5.3	19	40																																																		
8	5.8	28	60																																																		
9	5.5	23	45																																																		
10	5.6	32	58																																																		
11	5.5	38	?																																																		
4.	A machine learning model is trained to predict tumour in patients. 1349 people who have tumours are predicted positively by the model. 147 people who don't have tumours are predicted negatively by the model. 54 people are predicted as positive of having a tumour, although they don't have a tumour. 1 person who has tumours is predicted as negative. Find F1-Score of the model.	5	10	4	4																																																
5.	A smart city lab is analysing visitor movement across a tech expo using IoT sensors. The (X, Y) coordinates of five visitors on a 2D venue map are given below: <table><tr><th>Visitor</th><th>X</th><th>Y</th></tr><tr><td>P</td><td>2</td><td>4</td></tr><tr><td>Q</td><td>3</td><td>5</td></tr><tr><td>R</td><td>7</td><td>8</td></tr><tr><td>S</td><td>8</td><td>7</td></tr><tr><td>T</td><td>5</td><td>6</td></tr></table> Using $k = 2$ and Manhattan distance, (i) perform one complete iteration of the K-Medoids algorithm with clearly stated initial medoids, showing all distance calculations and cluster assignments, (ii) compute the total clustering cost and identify the optimal medoids after cost minimization, and (iii) briefly justify when K-Medoids is preferred over K-Means in real-world datasets.	Visitor	X	Y	P	2	4	Q	3	5	R	7	8	S	8	7	T	5	6	6	10	4	5																														
Visitor	X	Y																																																			
P	2	4																																																			
Q	3	5																																																			
R	7	8																																																			
S	8	7																																																			
T	5	6																																																			