Reg. No: 47

BG/D/TZ

Final Assessment Test – April 2026

VIT Vellore Institute of Technology

Course: BITE312E - Data Mining
Class NBR(s): 2163 / 2165
Time: Three Hours

Slot: C1
Max. Marks: 100

- KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

COs	CO Statements
CO1	Analyses various real-time problems and design the data visualizations.
CO2	Apply the data mining concepts to conduct data pre-processing and to improve the quality of data for training the models.
CO3	Use and apply important methods for finding frequent item sets and association rule Mining.
CO4	Investigate the concept of data classification methods and advanced classification techniques.
CO5	Understand the unsupervised learning techniques and the algorithm used for data clustering.

BL – Blooms Taxonomy Level (1 – Remember, 2 – Understand, 3 – Apply, 4 – Analyse, 5 – Evaluate, 6 – Create)

Answer ALL Questions
(10 X 10 = 100 Marks)

1. The Reliance retail chain maintains a sales data warehouse with a multidimensional schema. The fact table Sales stores the measured Sales_Amount and is linked to three dimensions: **CO1 BL2**

   i) Time(Year → Quarter → Month)

   ii) Product(Category → Subcategory → Item)

   iii) Location(Country → Region → City)

   An analyst, initially views a 3D OLAP cube that shows total Sales_Amount by Product Category, Quarter, and Region.

   Answer the following:

   a) Describe one roll-up operation and one drill-down operation the analyst can perform on this cube. State clearly on which dimension(s) each operation is applied and what the resulting view shows.

   b) The analyst now wants to focus only on Electronics products for Q1 2026 in the South region. Explain how slice and dice operations can be used on this cube to obtain such focused views, and clearly distinguish between the two.

5.b) Consider the following data set:

CO4 BL4

ID	Income	Student	Credit_Rating	Approve
1	High	No	Fair	No
2	High	No	Excellent	No
3	Medium	No	Fair	Yes
4	Low	Yes	Fair	Yes
5	Low	Yes	Excellent	No
6	Medium	Yes	Excellent	Yes
7	High	Yes	Fair	Yes
8	Low	No	Fair	No

i) Build a Naïve Bayes Classifier and if any partial probability is zero correct and then process.

ii) Classify the example X=<Low, Yes, Yes, Excellent>.

6. Build a neural-network classifier using backpropagation algorithm to predict whether a patient is at high risk of heart disease. For one training example (single patient): CO5 BL3

x1=0.9 (high cholesterol)

x2=0.7 (high blood pressure)

x3=0.4 (middle age)

Let there is one hidden layer having two neurons (h1, h2) with sigmoid activation and an Output layer one neuron (o) with sigmoid activation, output interpreted as probability of "High-Risk = Yes" and learning rate η=0.5. Consider the following initial weights:

Hidden layer: $w_{h1,1}$=0.1, $w_{h1,2}$=0.2, $w_{h1,3}$=−0.1,b b_{h1}=0.0, $w_{h2,1}$=0.1, $w_{h2,2}$=0.2, $w_{h2,3}$=−0.1, b_{h2}=0.0

Output layer: $w_{o,h1}$=0.1, $w_{o,h2}$=0.2, b_o=0.0

a) Perform the forward pass and compute the outputs of h1, h2, and the final output neuron o for this patient. (Show intermediate net inputs and activations.) [4]

c) The analyst decides to pivot the cube to better compare product performance across quarters. Explain what pivoting means in this context and give one specific example of how the row and column axes could be rearranged.

d) Briefly justify (2–3 lines) why multidimensional OLAP analysis is more suitable than flat relational queries for this kind of sales analysis.

2. Let the following table describes the test data of four patients for 3 tests. Find the similarity matrix with respect to Test 1, Test 2 and Test3. Compute the dissimilarity matrix based on the given table. CO2 BL5

O_ID	Test1 (Nominal)	Test2 (Ordinal)	Test3 (Numeric)
1	Code A	Excellent	45
2	Code B	Fair	22
3	Code C	Good	64
4	Code D	Excellent	28

3. Consider a hospital monitoring patients' blood glucose levels (mg/dL) before surgery. The raw data for 8 patients is: CO3 BL3

Patient IDs: 1, 2, 3, 4, 5, 6, 7, 8

Glucose: 92, 180, 140, 210, 75, 250, 130, 160

The hospital wants to:

a) Normalize the data for Patient ID 2, 4 and 7 to a new scale [5, 10].

b) Group the readings into clinical risk categories into Low, Medium, High risk using Equal-width binning.

4. A garment store records the following 10 transactions for one weekend sale: CO4 BL4

TID	Items bought
T1	Men-Shirt, Men-Jeans, Belt
T2	Men-Shirt, Men-Jeans, Socks
T3	Women-Kurti, Leggings, Scarf
T4	Men-Jeans, T-Shirt, Cap
T5	Women-Kurti, Leggings, Dupatta
T6	Men-Shirt, Men-Jeans, Belt, Socks
T7	T-Shirt, Men-Jeans, Socks

CO4 BL4

b) Compute the error term at the output neuron (δo) and the gradient of the error w.r.t. each output-layer weight $w_{o,h1}$=0.1 and $w_{o,h2}$=0.2. [3]

c) Using backpropagation, compute the error terms for the hidden neurons h1 and h2 ($\delta h1$, $\delta h2$), and the gradients for the hidden-layer weights connected to x1 (ie., $w_{h1,1}$, $w_{h2,1}$). [3]

7. Compare and contrast eager learner and lazy learner classifiers. Support your answer with a comparison table and briefly explain the scenarios where each approach is preferable. CO3 BL3

8. State how Prediction is different than the Classification. Let a real estate agent wants to predict house prices (₹ lakhs) based on size (sq.ft. in hundreds) using Simple Linear Regression. Historical sales data: CO4 BL5

House	Size(X)	Price(Y)
H1	20	45
H2	25	52
H3	30	58
H4	35	65
H5	40	72

Find the regression line and Predict price for new a house X=28 sq.ft. hundreds.

9.a) Construct the clusters by applying Agglomerative algorithm for the data objects described by the following table. CO5 BL6

Objects	X	Y
O1	1	1
O2	1.5	1.5
O3	5	5
O4	3	4
O5	4	4
O6	3	3.5

Use Euclidian distance to derive the distance matrix and single link distance as clustering criterion.

	T8	Women-Kurti, Scarf, Handbag
	T9	Men-Shirt, T-Shirt, Men-Jeans
	T10	Women-Kurti, Leggings, Scarf, Dupatta

Perform the following tasks by applying Apriori algorithm with minimum support = 30% and minimum confidence = 70%.

a) Using the frequent 2-itemsets, generate all possible candidate association rules of the form X → Y where X and Y are single garment items and identify which rules are strong. **[4]**

b) Select any two strong rules that the store can use for cross-selling (for example, tops with bottoms, or main garments with accessories). **[2]**

c) Based on your strong rules, suggest two concrete business actions for the manager, such as: **[4]**

(i) Changes in rack/shelf layout, and

(ii) Design of bundle discounts or festival offers. Explain briefly how each action is supported by the discovered rules.

5.a) A bank wants to predict whether to approve a credit card for new applicants. Build a decision tree with the ID3 algorithm using the following data. **CO4 BL4**

ID	Income	Student	Credit_Rating	Approve
1	High	No	Fair	No
2	High	No	Excellent	No
3	Medium	No	Fair	Yes
4	Low	Yes	Fair	Yes
5	Low	Yes	Excellent	No
6	Medium	Yes	Excellent	Yes
7	High	Yes	Fair	Yes
8	Low	No	Fair	No

OR

OR

9.b) A telecom company wants to place 2 service centres (medoids) to minimize total travel distance for 6 cell towers. Use K-Medoids (PAM algorithm) with K=2 and Manhattan distance. Let Cell Tower Coordinates (km) are as follows: A(1,2), B(2,1), C(8,8), D(9,7), E(3,9), F(4,10) CO5 BL6

Consider the initial medoids as : A(1,2), C(8,8)

Perform the below tasks:

i) Compute total cost for initial medoids (sum of distances)

ii) Perform 1 swap iteration - swap A with closest non-medoid to A, recompute assignments and total cost

iii) State final medoids and clusters.

10. A retail chain wants to segment its customers based on annual spending (₹ in thousands) and average purchase frequency (purchases/month) for targeted marketing. Use K-Means clustering (K=2) to identify customer segments. CO5 BL6

C1: (2,10), C2: (2,5), C3: (8,4), C4: (5,8), C5: (7,5), C6: (6,4), C7: (1,2), C8: (4,9)

Apply K-Means clustering with initial centroids C3 (8,4), C7(1,2).

Show exactly 2 iterations with:

i) Distance calculations,

ii) Point assignments,

iii) New centroid computation,

iv) Final customer clusters.

BG/D/TZ