R/K/TY

Reg. No: [redacted]

VIT
Vellore Institute of Technology

Final Assessment Test – November 2025

Course: BCSE209L - Machine Learning
Class NBR(s): 2085/2087/2098/2173 Slot: A2+TA2
Time: Three Hours 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	Understand, visualize, analyze and preprocess the data from a real-time source.
CO2	Apply appropriate algorithm to the data.
CO3	Analyze the results of algorithm and convert to appropriate information required for the real – time application.
CO4	Evaluate the performance of various algorithms that could be applied to the data and to suggest most relevant algorithm according to the environment.

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. You are given five real-world application scenarios as follows: **CO1 BL2**

   a) Predicting energy consumption in smart homes based on past usage and weather data
   b) Detecting unusual patterns in server logs to flag potential cybersecurity threats, without labeled data
   c) Recommending products to users on an e-commerce platform based on their browsing and purchase history
   d) Enabling a self-driving car to learn optimal driving strategies through interaction with the environment
   e) Classifying types of crop diseases from smartphone-captured leaf images submitted by farmers

   For **each scenario**, identify the most appropriate machine learning algorithm and its type (***supervised, unsupervised, reinforcement***) from among the following list: *Linear Regression, Multiple Linear Regression, Logistic Regression, Naive Bayes Classifier, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), K-Means, K-Modes, Hierarchical Clustering, Self-Organizing Map (SOM), Random Forest, AdaBoost, XGBoost, Q-Learning*. Provide a proper justification in **2–3 sentences** explaining your choice.

2. You work for a smart home security company, and your job as a machine learning expert is to design a **decision tree** that predicts whether the **alarm should be triggered today** (TriggerAlarm? = 1) or not (TriggerAlarm? = 0). You are given a dataset D with the following attributes: MotionDetected $\in \{0, 1\}$, LightsOn $\in \{0, 1\}$, DoorOpenedYesterday $\in \{0, 1\}$ and TempComfort $\in \{0, 1\}$. **CO2 BL5**

MotionDetected	LightsOn	DoorOpenedYesterday	TempComfort	TriggerAlarm?
1	1	1	0	1
0	1	0	0	0
1	0	0	0	0
1	0	0	1	0
1	0	1	1	0
1	1	0	1	1
0	1	0	0	0
1	0	1	1	0

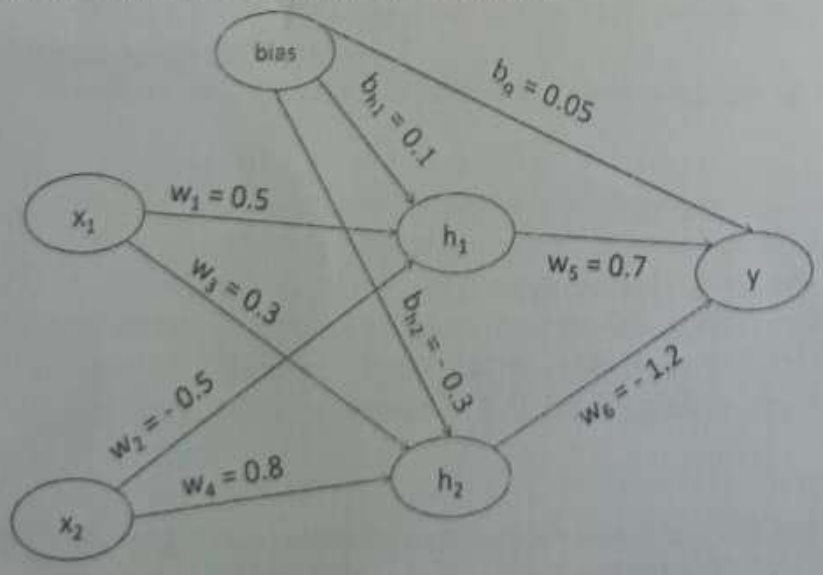
Calculate the entropy of the label 'TriggerAlarm?'. What proportion of the examples should be labeled TriggerAlarm? = 1 in order to maximize the entropy of the label? Compute the Gain of each of these attributes and identify the root node.

3. A company wants to predict the **monthly sales (Y, in units)** of a new product based on two factors: **advertising expenditure (X_1, in INR 000s)** and **number of sales calls made (X_2)**. The following data were collected from **5 regions**: CO2 BL3

Region	Advertising (X_1)	Sales Calls (X_2)	Sales (Y)
1	2	3	28
2	4	2	35
3	3	5	40
4	5	4	50
5	6	6	62

Fit a **multiple linear regression model** of the form $Y = b_0 + b_1X_1 + b_2X_2$ using the **normal equations** (show all key steps). Interpret the coefficients b_1 and b_2. Predict the expected sales for a region where the company spends **₹5,000 on advertising** and makes **7 sales calls**.

4. Consider the Multi-Layer Perceptron (MLP) shown below (weights and biases given in the diagram) with 2 input neurons, 1 hidden layer with 2 neurons (sigmoid activation), and 1 output neuron (sigmoid activation). CO2 BL3



For input $(x_1,x_2) = (1,0)$ with target output t=1, perform a forward pass and calculate the output of the network. Using gradient descent with learning rate $\eta = 0.1$, perform **one iteration of weight update using backpropagation** for both the output and hidden layers. Finally, tabulate the updated weights and biases in the following format;

Weight	*Updated value*
w_1	
w_2	
...	

5. We have the following training dataset with two features (x_1, x_2) and a binary class label (Class = A or B): **CO2 BL5**

ID	x_1	x_2	*Class*
1	2	3	A
2	2	2	A
3	4	3	A
4	6	5	B
5	7	7	B
6	8	6	B

We want to classify a new point:

$(x_1, x_2) = (4, 4)$

Compute the **Euclidean distance** between the test point (4, 4) and each training point. Assign the class of the test point for k=1, k=3 and k=5. Discuss briefly: *How does the choice of k affect the classification result in this dataset?*

6. Suppose that the points (q1, q2, q3, q4, q5, q6) provided in a 2D plane. **Table 1** is the distance matrix which shows the Euclidean distances between pair of data points. **CO3 BL3**

	q1	*q2*	*q3*	*q4*	*q5*	*q6*
q1	0					
q2	0.20	0				
q3	0.55	0.35	0			
q4	0.75	0.40	0.22	0		
q5	0.60	0.18	0.50	0.33	0	
q6	0.42	0.68	0.90	0.56	0.71	0

Table 1

Phase →	*1*	*2*	*3*	*4*	*5*
New cluster	(q2, q5)				
Distance to choose the cluster	0.18				

Table 2

Using **Single and Complete Linkage Hierarchical Agglomerative Clustering** technique, for each phase, find the clusters, show the calculations, and update the distance matrix. Also, fill **Table 2** with the final distances (copy the table for both methods) and draw the final dendrogram. **Note**: Phase 1 result is shown for you.

7. Random Forests are often preferred over single decision trees due to their lower variance. Analyze how the use of bagging and random feature selection in Random Forests contributes to this variance reduction. Additionally, discuss whether Random Forests can still overfit. If so, under what circumstances this might happen. **CO3 BL2**

8. You are given a medical dataset of 1,000 patient records used to predict whether a patient has a rare disease. Out of 1,000 patients, only 50 are diagnosed with the disease (positive class), and the remaining 950 are healthy (negative class). A machine learning model trained on this dataset achieves high overall accuracy but performs very poorly in detecting patients with the disease. **CO4 BL5**

   Based on this scenario, explain how SMOTE can be applied to improve the model's performance. Describe how SMOTE creates synthetic samples, and how this differs from simply duplicating existing minority class records. Also, discuss one possible risk of applying SMOTE without caution in a medical context and suggest a strategy to reduce such risks.

9.a) Given the following 2D data points representing students' performance in two subjects (Math and Science): **CO3 BL3**

(60,58), (65,63), (70,68), (75,85), (80,90), (85,95)

Center the data, compute the covariance matrix of the centered data, calculate the eigenvalues and eigenvectors of the covariance matrix, and identify the **first principal component**. Also, project each centered data point onto the first principal component and fill in a table showing each projection against actual data points.

OR

b) You have six training points (A, B, C, D, E, F) and five classifiers (h1, h2, h3, h4, h5) which make the following misclassifications; CO3 BL3

Classifier	*Misclassified points*
h1	A, D, F
h2	D
h3	B, C
h4	A, B, F
h5	B, C, D

Perform two rounds of boosting using AdaBoost with these classifiers and training data. Start with equal weights. In each round, pick the classifier with the lowest weighted error (ε) [*Note: If there is a tie during the error rate calculation, choose the classifier that comes first in this list: h1, h2, h3, h4, h5*], calculate the classifier's weight (α), update the weights and normalize the weights. At the end, fill the following table;

Round	Classifier	Weighted error (ε)	Weight of weak classifier (α)
1			
2			

10.a) Explain how Q-learning works as a reinforcement learning algorithm, including the roles of the Q-table, learning rate, discount factor, and exploration-exploitation trade-off. Illustrate your explanation with an example of an agent learning to navigate a grid-world environment to reach a goal while avoiding obstacles. Finally, discuss two real-world applications where Q-learning can be effectively used and why it is suitable for those scenarios. CO4 BL4

OR

10.b) Reinforcement Learning (RL) agents learn by interacting with the environment. Explain how the choice of reward function, exploration strategy, and discount factor (γ) can affect the learning and performance of an RL agent. Illustrate your answer with a practical example, such as training an agent to play a simple game or navigate a grid world. CO4 BL4

⇔⇔⇔ R/K/TY ⇔⇔⇔