Total No. of Pages 02 *Roll No. ……..*

FIFTH SEMESTER **B.Tech. (CSE)**

MID SEMESTER EXAMINATION *Sept, 2025*

CO303 MACHINE LEARNING

Time: 1:30 Hours ***Max. Marks: 20***

Note:	Answer ***ALL*** questions. Assume suitable missing data, if any. CO# is course outcome(s) related to the question.

1. A city transport authority wants to develop a machine learning model to predict whether a passenger will arrive on time (Yes/No) when using public buses. The dataset includes features such as: travel distance (km), number of transfers (integer), day of week (weekday/weekend), bus stop ID (numeric code), average traffic delay (minutes), and weather condition (clear/rain/snow). Answer the following (**3–4 sentences maximum** for eahc): **[1+1+1+2] [CO1,2,4][BTL-2,3]**

[a] Should this problem be formulated as a classification or regression task?

[b] Identify any feature(s) that should be removed during feature selection and explain why.

[c] Explain why feature scaling may be important for some algorithms in this dataset, giving one example of an algorithm that requires scaling and one that does not.

[d] Critically evaluate the use of linear regression for this task and suggest a more appropriate algorithm with justification.

2. A company is using a decision tree to predict whether a customer will buy a product (Yes/No) based on two categorical features: Age Group (Young, Middle, Senior) and Income Level (Low, High). The training dataset is shown below: **[1+1+1+2] [CO1,34] [BTL-3,4]**

Table I

Age Group	Income Level	Purchase
Young	Low	No
Young	High	Yes
Middle	Low	No
Middle	High	Yes
Senior	High	Yes

Later, after reviewing sales records, the company discovers that one customer was misclassified. The last record should actually be Senior–High–No instead of Senior–High–Yes.

[a] For the original dataset, calculate the Gini Index for splits on both 'Age Group' and 'Income Level', and identify which feature the tree will choose at the root.

[b] Repeat the calculations for the corrected dataset.

[c] Compare the two trees and explain how a single change in one record alters the structure.

[d] What general drawback of decision trees does this example highlight, and why could this be problematic in real-world business decision-making (e.g., churn prediction, loan approvals)?

3. A research team is developing an SVM classifier to identify whether a crop plant is diseased (1) or healthy (0) using features such as Leaf Color Index, Moisture Level, and Growth Rate. The data collected is not linearly separable. To improve classification, the team experiments with different values of the regularization hyperparameter *C*. With C = 0.1 (softer margin), the test confusion matrix is: **[1+2+2] [CO3,4] [BTL-3,4]**

Table II: Confusion Matrix with *C* = 0.1

	Predicted Diseased	**Predicted Healthy**
Actual Diseased	40	20
Actual Healthy	5	135

With C = 100 (harder margin), the test confusion matrix is**:**

Table III: Confusion Matrix with *C* = 100

	Predicted Diseased	**Predicted Healthy**
Actual Diseased	55	5
Actual Healthy	25	115

Answer the following:

[a] Calculate the precision and recall for both C=0.1and C=100. Show your working.

[b] Suppose the classifier is used for detecting crop diseases early in a large farm, where missing a diseased plant is very costly because infection can spread quickly. Based on your calculations, which value of C is more suitable? Explain your reasoning. (**3–4 sentences maximum**)

[c] Suppose the classifier is used for certifying crops for export, where wrongly labeling healthy crops as diseased leads to large financial loss. Based on your calculations, which value of C is more suitable? Explain. (**3–4 sentences maximum**)

4. A bank uses a logistic regression model to predict whether a customer will default on a loan (1) or repay (0). The model outputs predicted probabilities of default.

The true outcomes and predicted probabilities for 6 customers are shown below: **[3+2] [CO1,3,4][BTL-3,4]**

Table IV

Customer	**True Label (*y*)**	**Predicted Probability (*p*)**
C1	1	0.90
C2	0	0.80
C3	1	0.70
C4	0	0.60
C5	1	0.40
C6	0	0.30

[a] For thresholds t=0.5and t=0.7: Construct the confusion matrix and compute True Positive Rate (TPR = Recall) and False Positive Rate (FPR).

[b] Interpret how changing the threshold affects the trade-off between catching defaulters (recall) and avoiding false alarms (specificity) in the banking context.

---Best of Luck---