ourdiu.com | Shared by SUPAN ROY

Daffodil International University
Department of Software Engineering
Faculty of Science & Information Technology
Midterm Examination, Fall 2025
Course Code: SE544; Course Title: Introduction to Machine Learning
Sections & Teachers: 40(B,D,G,F)(JJS), 40(E,H)(FA), 40(A,C)(NS), 40I(DB)

Time: 1 Hour 30 Mins **Marks: 25**

Answer ALL Questions

[The figures in the right margin indicate the full marks and corresponding course outcomes. All portions of each question must be answered sequentially.]

1.	*a)*	A data scientist at TechSecure Inc. trains a model for predicting spam emails for his company. The model gives 98% accuracy on training data but only 60% accuracy on test data. Later, a simpler model gives low accuracy on both training and test data. Based on the scenario, **identify** which model is overfitting and which is underfitting. Briefly explain your answer.	**[Marks-2]**	**CLO-1** *Level-2*
	b)	**Explain** the bias-variance tradeoff observed in this scenario. Suggest two techniques the data scientist can use to achieve a better balance between bias and variance.	**[Marks-3]**	
2	*a)*	A food delivery startup "QuickBite" wants to understand the relationship between delivery distance (in km) and delivery time (in minutes) to optimize their service. The operations manager collected data from 7 recent deliveries: i) **Examine** the data and construct a simple linear regression model by calculating Slope and Intercept. Using the model, predict the delivery time for a new order that needs to be delivered 9 km away. Also, calculate coefficient of determination to evaluate how well distance explains delivery time variance.	**[Marks-3]**	**CLO-2** *Level-3*
		ii) Additionally, **calculate** the performance metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).	**[Marks-2]**	

Delivery	Distance (X Km)	Time (Y Minutes)
1	2	12
2	3	15
3	5	16
4	6	18
5	4	15
6	7	22
7	6	20

ourdiu.com

ourdiu.com | Shared by SUPAN ROY

2 **b)** An e-commerce platform "ShopSmart" wants to predict product ratings based on two features: Price (in $) and Number of Reviews. They have historical data from 8 products:

Product	Price ($)	Reviews (#)	Rating (out of 5.0)
P1	20	60	3.5
P2	35	110	4.2
P3	45	130	4.7
P4	18	72	4.1
P5	50	130	3.5
P6	28	90	4.8
P7	48	143	4.3
P8	36	106	3.9

New Product: Price =30, Reviews = 95

i) **Calculate** the Euclidean distance from the new product to all 06 existing products. Using K=3, identify the 3 nearest neighbors and predict the rating using simple averaging (unweighted method). **[Marks-3]** **CLO-2** *Level-3*

ii) Now predict the rating using weighted averaging where weight = 1/distance) **Compare** both predictions and explain which method gives more reliable results in this scenario. **[Marks-2]**

3 **a)** A fintech company "SecurePay" developed a machine learning model to detect fraudulent transactions in real-time. The model classifies transactions into two categories: Legitimate and Fraud.

The model was tested on 900 transactions with the following actual distribution:

=> Legitimate transactions: 870, => Fraud transactions: 30

Model Performance Results:

=> Out of 870 Legitimate transactions 845 correctly classified as Legitimate, and the rest are misclassified as Fraud

=> Out of 30 Fraud transactions 16 correctly classified as Fraud, and the rest are misclassified as Legitimate

Construct a 2×2 confusion matrix based on the given results. Then calculate the Precision and Recall for the Fraud class only. Show all calculations clearly. **[Marks-3]** **CLO-3** *Level-3*

b) In this fraud detection system, which type of error is more critical: classifying Fraud as Legitimate (False Negative) or classifying Legitimate as Fraud (False Positive)? **Validate** your answer by discussing the real-world consequences for the company and customers. **[Marks-2]**

4. **a)** DataDrive AI has three client projects:

Project A: Group millions of social media users for advertising (no predefined labels)

Project B: Train delivery drones to learn optimal flight paths through trial and feedback

Project C: Diagnose pneumonia from 100,000 labeled X-ray images

Interpret each project with the appropriate ML type (Supervised, Unsupervised, or Reinforcement Learning) and justify each choice. **[Marks-3]** **CLO-4** *Level-2*

b) Which project would **require** the most computational resources during training? Why? **[Marks-2]**