Daffodil International University

**Faculty of Science & Information Technology**

**Department of Computer Science & Engineering**

**Midterm Examination, Summer 2026**

**Course Code: CSE431, Course Title: Machine Learning**

**Level: 4 Term: 2 Batch: 64**

**Time: 01:30 Hrs** **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 subscription-based online learning platform offers both Free and Premium membership plans. The company aims to develop a machine learning model to predict whether a newly registered user will purchase a Premium subscription within the next 30 days.<br>The dataset contains the following features:<br>• Number of courses viewed<br>• Number of completed lessons<br>• Total learning hours<br>• Average session duration<br>• Number of promotional emails opened<br>• Device type<br>• Age<br>• Previous subscription history<br><br>During exploratory data analysis, the data science team discovers that Number of Courses Viewed, Number of Completed Lessons, and Total Learning Hours are highly correlated, indicating multicollinearity problem. They are also evaluating different predictive models and planning to incorporate additional user-behavior features in future versions of the system. | | **CO1** |
| | *a)* | The team observes strong correlations among some features. Explain how this situation affects model training and the interpretation of model coefficients. | **[2]** | |
| | *b)* | For predicting Yes/No subscription, explain the suitability of Logistic Regression compared to Linear Regression. | **[2]** | |
| | *c)* | To support Premium subscription prediction, an SVM model is trained on 100 user-behavior data samples. After training, only 4 samples have non-zero $\alpha$ values. Analyze the role of these 4 samples in determining the SVM decision boundary and the effect of removing one of them from the training dataset. | **[3]** | |

**2.** A student wants to predict house prices using House Size (sq. ft.) as the only feature. The dataset is given below: **[8]** **CO2**

| House Size (100 sq.ft.) | Number of Rooms | Price ($1000) |
|---|---|---|
| 10 | 3 | 200 |
| 12 | 3 | 220 |
| 15 | 4 | 270 |

The student initially applied a Linear Regression model, but found that the predictions had high error (MSE).

Assist the student in improving the model's accuracy. Apply a suitable error-minimization method using the House Size feature and show two complete parameter update step. Comment on whether the error is improving and how to further improve predictions.
Use initial weight w=0.5 and bias b=10, Learning rate $\alpha$=0.01, Loss function: MSE

**3.** A bank wants to develop a machine learning model to predict whether a loan application should be Approved or Rejected. **CO2**
The historical dataset is given below:

| Employment Type | Marital Status | Annual Income (K USD) | Loan Status |
|---|---|---|---|
| Government | Married | 60 | Approved |
| Private | Single | 35 | Rejected |
| Business | Married | 75 | Approved |
| Private | Married | 55 | Approved |
| Business | Single | 30 | Rejected |
| Government | Single | 80 | Approved |

A new applicant has the following profile:

| | | | |
|---|---|---|---|
| Business | Married | 50 | ? |

***a)*** Apply a suitable machine learning model to the given dataset and explain the rationale for selecting that model. **[5]**

***b)*** Using the model applied in part (a), determine the loan status of the new applicant. **[2]**

***c)*** A Decision Tree model achieves 100% training accuracy but performs poorly on unseen data. Explain how the Ensemble Learning Technique can address this issue and why it generally provides better generalization than a Decision Tree. **[3]**