Capital University of Science and Technology

Department of Artificial Intelligence

Course Title: Ai

ASSIGNMENT NO.3

Semester: Fall 2025 **Instructor: SIr Adnan**

Assigned Date: 2026-04-1 **Due Date: 2026-04-1**

Name: Abdullah Luqman
Reg. No. BAI243042

1. Introduction

This assignment focuses on applying **Linear Regression** to a **binary classification problem** using a student performance dataset. Although Linear Regression is primarily used for predicting continuous values, it is adapted here for classification by applying a fixed threshold to convert predictions into binary outcomes.

The purpose of this project is to understand model evaluation, performance metrics, and visualization techniques in machine learning.

---

2. Dataset Description

- **Dataset Name:** Student Performance Dataset
- **Source:** Kaggle
- **File Name:** student_mat.csv
- **Domain:** Education

Attributes

The dataset includes:

- Student demographic information
- Study time and attendance
- Academic and personal factors
- Final grade (G3)

---

3. Tools and Libraries Used

- **Pandas:** Data loading and preprocessing
- **NumPy:** Numerical computations

- **Scikit-learn:** Model training and evaluation
- **Matplotlib:** Visualization
- **Seaborn:** Statistical plots

---

4. Data Preprocessing

4.1 Binary Target Creation

A binary target variable named **above_median_performance** is created using the median of the final grade:

- **1:** Above median performance
- **0:** Below or equal to median performance

4.2 Feature Handling

- Original grade column (G3) is removed
- Categorical variables are encoded using one-hot encoding
- Dataset is converted into numeric format

---

5. Train-Test Split

The dataset is split as follows:

- **Training Set:** 80%
- **Testing Set:** 20%
- **Random State:** Fixed for reproducibility

---

6. Model Training

A **Linear Regression** model is trained on the training dataset. The model outputs continuous prediction values.

---

7. Prediction Strategy

7.1 Continuous Predictions

The model generates real-valued predictions.

7.2 Binary Classification

Predictions are converted into binary classes using a threshold:

- Value ≥ 0.5 → Class 1
- Value < 0.5 → Class 0

---

8. Evaluation Metrics

Regression Metrics

- Mean Squared Error (MSE)
- R-squared (R^2) Score

Classification Metrics

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix

---

9. ROC Curve and AUC

A **Receiver Operating Characteristic (ROC)** curve is plotted using continuous predictions. The **Area Under the Curve (AUC)** summarizes the model's classification performance.

---

10. Visualizations

- Confusion Matrix Heatmap
- ROC Curve
- Classification Metrics Bar Chart

---

11. Results Summary

- The model achieves moderate accuracy
- Recall is reasonable for high-performing students
- R^2 score is low, showing limitations of Linear Regression
- ROC AUC is slightly better than random guessing

---

12. Limitations

- Linear Regression is not ideal for classification
- Fixed threshold can cause misclassification
- Relationships may not be linear

---

13. Future Improvements

- Use Logistic Regression

- Apply Random Forest or SVM
- Perform feature scaling and hyperparameter tuning