Assignment-1 Report

Course: CS502 – Advanced Pattern Recognition Submitted by: Vinay Sadupati (2201ai44)

Title

Predicting Student Final Grades using Linear Regression

1. Introduction

Academic performance prediction is an important task in the field of educational data mining. Institutions often want to identify students who might be at risk of poor performance so that timely intervention can be provided.

In this assignment, we use Linear Regression to model and predict student final grades based on the Student Performance Dataset (UCI Machine Learning Repository). The dataset includes demographic, behavioral, and academic factors of students. Among these, we focus on academic and behavioral variables such as study time, past failures, absences, and earlier grades (G1 and G2).

The main objective is to build a regression model that predicts the final grade (G3) and to analyze how different factors contribute to academic success.

2. Dataset Description

- Source: UCI Machine Learning Repository – Student Performance Dataset(UCIstudentPerformance)
- File Used: student-mat.csv (data related to Math course performance)
- Target Variable:
  - G3 → Final Grade (scale: 0–20)

- Features Selected for Modeling:
  1. studytime → Weekly study time (categorical: 1–4)
  2. failures → Number of past class failures (0–3)
  3. absences → Number of school absences (continuous)
  4. G1 → First period grade (0–20)
  5. G2 → Second period grade (0–20)

The features were chosen because they directly influence academic outcomes.

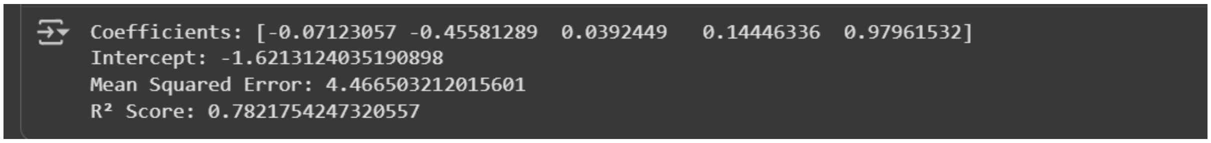
3. Methodology

The steps followed are:

1. Data Loading: Used pandas to load the dataset.
2. Feature Selection: Selected only numerical features relevant to academic performance.
3. Data Splitting: Split dataset into 80% training and 20% testing.
4. Model Training: Applied Linear Regression from scikit-learn.
5. Model Evaluation: Used R^2 score and Mean Squared Error (MSE).
6. Visualization: Compared actual vs predicted G3 using scatter plot.

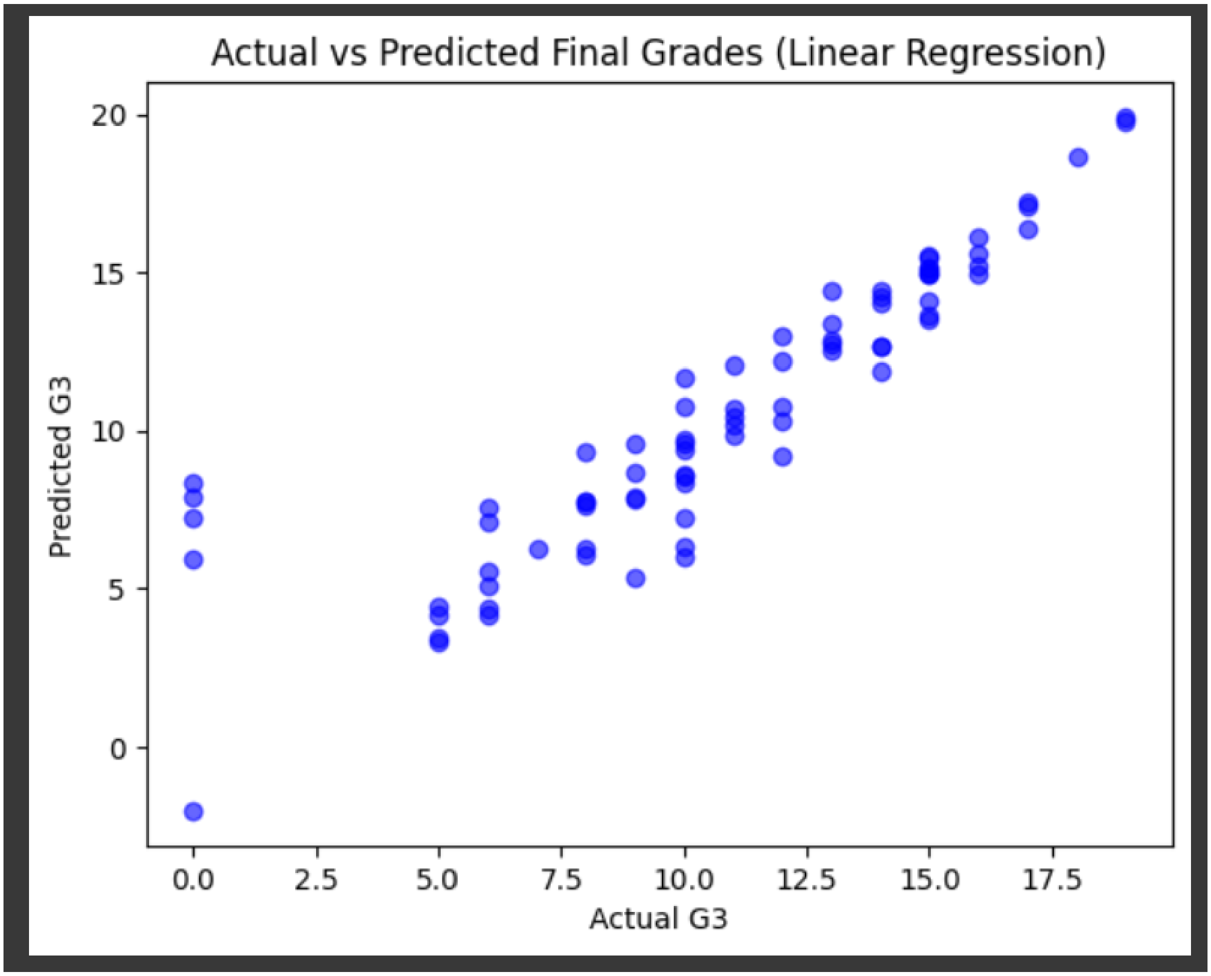
4. Code Implementation

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Load dataset
data = pd.read_csv("student-mat.csv", sep=";")

# Features and target
X = data[["studytime", "failures", "absences", "G1", "G2"]]
y = data["G3"]

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
   X, y, test_size=0.2, random_state=42
)

# Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Predictions
y_pred = model.predict(X_test)

# Evaluation metrics
print("Coefficients:", model.coef_)
print("Intercept:", model.intercept_)
print("Mean Squared Error:", mean_squared_error(y_test, y_pred))
print("R² Score:", r2_score(y_test, y_pred))
```



```
Coefficients: [-0.07123057 -0.45581289  0.0392449   0.14446336  0.97961532]
Intercept: -1.6213124035190898
Mean Squared Error: 4.466503212015601
R² Score: 0.7821754247320557
```

```
# Visualization
plt.scatter(y_test, y_pred, color="blue", alpha=0.6)
plt.xlabel("Actual G3")
plt.ylabel("Predicted G3")
plt.title("Actual vs Predicted Final Grades")
plt.show()
```



5. Results

- Model Parameters:
  - Coefficients:
    - • Study Time → -0.0712
    - • Failures → -0.4558
    - • Absences → +0.0392
    - • G1 → +0.1445
    - • G2 → +0.9796
  - Intercept: -1.6213

- Performance Metrics:
  - Mean Squared Error (MSE): 4.4665
  - R^2 Score: 0.7822

- Visualization: Scatter plot of Actual vs Predicted G3 shows predictions are close to the diagonal line.

6. Interpretation of Results

1. Strong Predictor: G2 (second-period grade) is the strongest predictor of G3.
2. Moderate Predictor: G1 contributes positively but less than G2.
3. Weak Predictors: Study time, failures, and absences have minimal effect when G1 and G2 are included.
4. Model Accuracy: With R^2 = 0.78, the model explains most variance. MSE ~4.47 means average prediction error is around ±2 points.

7. Conclusion

The Linear Regression model predicts student final grades with good accuracy. Past academic performance (G1 and G2) are the most important predictors. Other behavioral factors have weaker influence.