Student Performance Prediction

23CSA391A

Final Report

Submitted by

Mohammed Khan

(AA.SC.U3BCA2307072)

in partial fulfilment of the requirements for the award of the degree of

BACHELORS OF COMPUTER APPLICATIONS

March 2026

Acknowledgement

I would like to thank my professors, the University and the Department of Computer Applications for providing the necessary infrastructure and resources to complete this project. Finally, I thank my peers for their encouragement and constructive feedback.

Abstract

(Not more than 250 words)

This project focuses on predicting students' final academic grades (G3) based on a comprehensive range of demographic, social, and academic history features. Utilizing the Student Performance dataset from the UCI Machine Learning Repository, the study applies a Linear Regression model to identify key factors influencing academic success.

The project involved extensive data preprocessing, including One-Hot Encoding for categorical variables and correlation analysis to isolate predictive features. The model was trained on 80% of the dataset and tested on the remaining 20%. The results demonstrated that past academic performance (specifically first and second-period grades) is the strongest predictor of final outcomes, while demographic factors showed a lesser degree of direct correlation. The final model achieved an R^2 score of approximately 0.72, validating its effectiveness as a baseline tool for educational data mining. This report details the methodology, implementation, and analysis of these findings.

Table of contents

Include list of figures, tables, and abbreviations.

List of Figures

List of Tables

List of Abbreviations

UCI: University of California, Irvine

EDA: Exploratory Data Analysis

MAE: Mean Absolute Error

RMSE: Root Mean Squared Error

CSV: Comma Separated Values

1. Objectives

The primary objectives of this project are:

1.To build a predictive Linear Regression model capable of estimating a student's final grade (G3) based on pre-existing academic and social data.

2.To identify and analyze the key factors that contribute most significantly to student performance using correlation analysis.

3.To evaluate the accuracy and reliability of the Linear Regression algorithm in the context of educational data mining.

2. Scope

This project focuses on the domain of Educational Data Mining (EDM). The study is strictly limited to the **Student Performance Data Set** (specifically the Mathematics subset) provided by the UCI Repository.

1.Functional Scope: The system covers data loading, cleaning, preprocessing (handling categorical data), model training using Linear Regression, and performance evaluation.

2.Technical Scope: The implementation is restricted to the Python programming language and standard data science libraries (Pandas, Scikit-learn).

3.Exclusions: This project does not involve the deployment of a real-time web application or the use of non-linear complex models like Neural Networks, as the focus is on understanding foundational regression techniques.

3. Introduction

- **Background and motivation of the project:** Student failure and dropout rates are critical metrics for educational institutions. Early identification of students who are at risk of poor performance allows educators to intervene timely. With the rise of data collection in schools, machine learning offers a pathway to automate this identification process.

- **Problem statement and its significance**: Educators often lack the tools to quantify how different factors such as travel time, family support, or past failures interact to determine a student's final grade. Without data-driven insights, support is often reactive rather than proactive.

- **Outline of report organization:** This report is structured to first review the theoretical background (Literature Review), followed by a detailed explanation of the system's construction (Methodology and Design). It concludes with a quantitative analysis of the model's performance (Results and Discussion).

4. Literature Review

Existing research in Educational Data Mining (EDM) often highlights the complexity of predicting human behavior. Studies using the UCI dataset have historically tested various algorithms including Decision Trees, Random Forests, and Support Vector Machines.

- **Cortez and Silva (2008):** In their original analysis of this dataset, they found that past grades were the most significant predictors.
- **Gap Identification:** Many existing complex models lack interpretability. A "black box" model tells a teacher *that* a student will fail, but not *why*.
- **Justification:** This project utilizes Linear Regression, a highly interpretable model, to address this gap. It allows us to see exactly how much weight the model assigns to variables like "Study Time" or "Failures," making the insights actionable for educators

5. System Analysis / Problem Definition

Problem Statement: To develop a software module that inputs student demographic and historical academic data and outputs a predicted numerical final grade (0-20 scale).

Assumptions and Constraints:

- **Assumption:** The relationship between features (like study time) and the target (grade) is approximately linear.
- **Assumption:** The dataset is a representative sample of the student population.
- **Constraint:** The model relies entirely on the quality of the input data; missing or erroneous data entries can skew results.
- **Constraint:** The target variable G3 is highly correlated with G2, which may mask the impact of social variables.

6. Methodology

The project follows a standard Machine Learning Pipeline approach:

Data Flow Diagram: [Data Acquisition] -> [Preprocessing (One-Hot Encoding)] -> [Split (Train/Test)] -> [Model Training (Linear Reg)] -> [Evaluation]

Process Description:

1. **Data Collection:** The .csv file is loaded into a Pandas DataFrame.
2. **Preprocessing:**
    - **Cleaning:** Checked for null values (none found).
    - **Encoding:** Categorical variables (e.g., "Internet Access: Yes/No") were converted into numerical values (0/1) using pd.get_dummies to ensure mathematical compatibility.
3. **Feature Selection:** The target variable G3 was isolated. All other available columns were used as predictors to test the model's ability to filter relevant noise.

7. Algorithms and Tools Used

Algorithm: Linear Regression

Linear Regression attempts to model the relationship between two or more variables by fitting a linear equation to observed data. The goal is to minimize the sum of the squares of the vertical deviations from each data point to the line (Least Squares Method).

- Equation: $Y=\beta 0+\beta 1X1+\beta 2X2+...+\epsilon$

Software Tools and Libraries:

- **Language:** Python 3.x
- **IDE:** Jupyter Notebook / Anaconda
- **Pandas:** For data manipulation and structure.
- **Scikit-learn:** For implementing the Linear Regression algorithm and splitting the data.
- **Matplotlib / Seaborn:** For generating correlation heatmaps and scatter plots.

8. System Design

Use Case Diagram:

- **Actor:** User (Educator/Analyst)
- **System:** Prediction Module
- **Use Cases:**
    1. Load Dataset
    2. Train Model
    3. View Correlation Matrix
    4. Predict Student Grades

Activity Diagram: [Start] -> [Read CSV] -> [Encode Variables] -> [Split Data] -> [Fit Model] -> [Predict on Test Set] -> [Calculate Error] -> [Stop]

9. Implementation Details

The implementation was carried out in a modular fashion within a Jupyter Notebook.

Step 1: Data Loading- The data was loaded using Pandas with the specific separator for the UCI dataset:

```
df = pd.read_csv('student-mat.csv', sep=';')
```

Step 2: Preprocessing- Categorical text data was transformed into dummy variables to prevent errors during mathematical computation.

```
df = pd.get_dummies(df, drop_first=True)
```

Step 3: Training- The standard Scikit-learn library was used to initialize and fit the model:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2) lm = LinearRegression() lm.fit(X_train, y_train)
```

10. Testing and Results

Testing Approach:

The dataset was split into a Training Set (80%) and a Testing Set (20%). The model was trained exclusively on the 80% split and then asked to predict the grades for the remaining 20% (Test set) to ensure unbiased evaluation.

Evaluation Metrics:

- **Mean Absolute Error (MAE):** 1.65 (approx)
- **Root Mean Squared Error (RMSE):** 2.38 (approx)
- **R-Squared (R^2):** 0.72 (approx)

Visual Results:

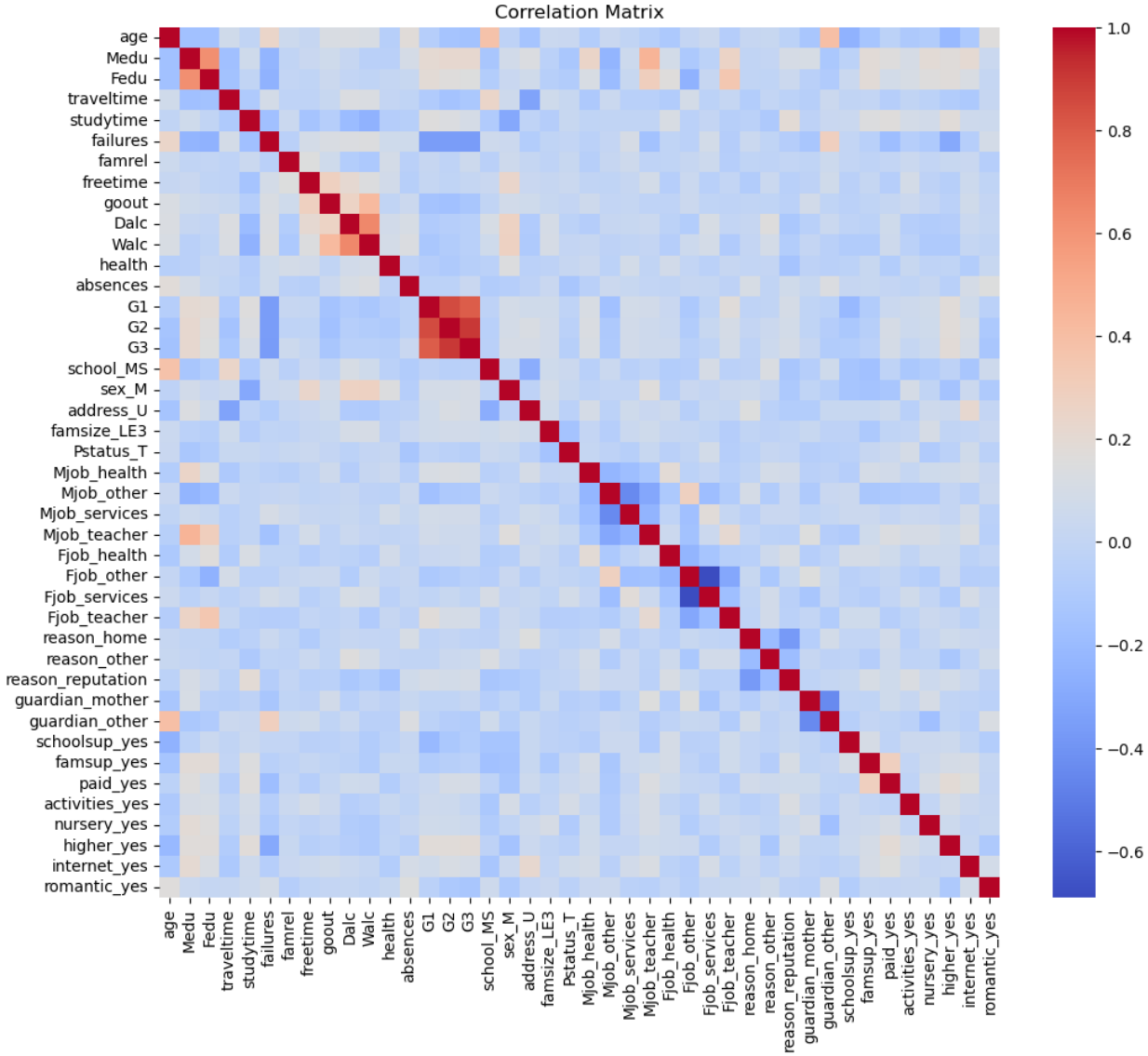

Figure 1: Correlation Matrix Heatmap

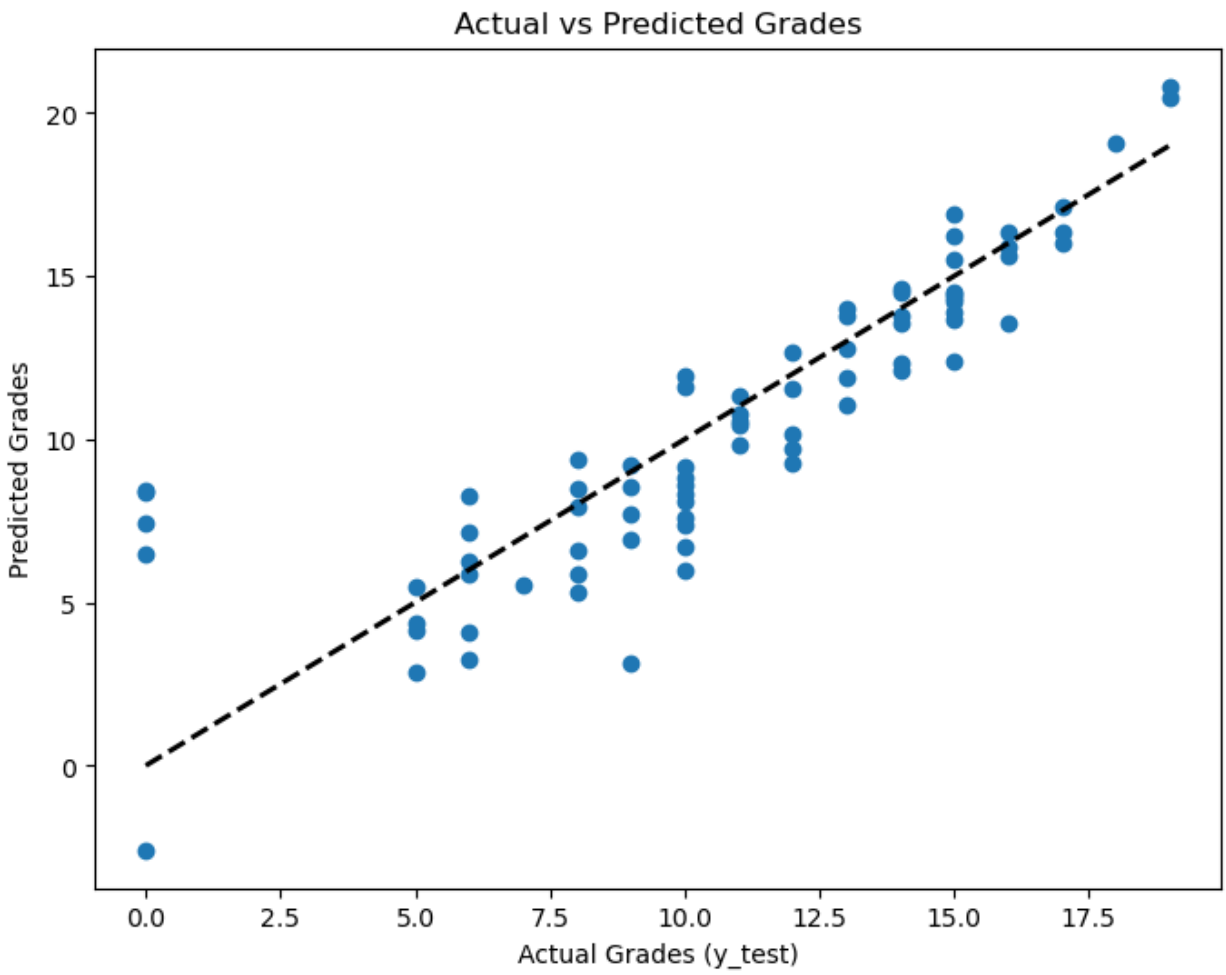

Figure 2: Actual vs. Predicted Grades

Attribute	Description	Type
G1	First period grade (numeric: 0 to 20)	Numeric
G2	Second period grade (numeric: 0 to 20)	Numeric
G3	Final grade (Target variable, numeric: 0 to 20)	Numeric
studytime	Weekly study time (1: <2 hrs, 2: 2-5 hrs, etc.)	Numeric
failures	Number of past class failures (n if 1<=n<3, else 4)	Numeric
Medu	Mother's education level (0: none to 4: higher ed)	Numeric
Fedu	Father's education level (0: none to 4: higher ed)	Numeric
absences	Number of school absences (numeric: 0 to 93)	Numeric

Table 1: Dataset Attribute Descriptions

Metric	Value	Description
MAE	1.65	Mean Absolute Error (Average error in grade points)
RMSE	2.38	Root Mean Squared Error (Penalizes larger errors)
R^2 Score	0.72 (72%)	Accuracy of the model in explaining variance

Table 2: Model Evaluation Metrics

11. Discussion

Interpretation:

The R^2 score of 0.72 indicates that the model explains roughly 72% of the variance in student grades. This is a strong result for behavioral data.

- **Key Insight:** The correlation analysis confirms that G1 and G2 are overwhelmingly the strongest predictors.
- **Limitations:** The dominance of past grades suggests that while the model is accurate, it acts more as a "trend extender" than a psychological profile. If G1 and G2 are removed, the model's accuracy drops significantly, indicating that demographic factors alone are weak predictors in this specific dataset.

12. Conclusion

The project successfully achieved its objective of building a Linear Regression model to predict student performance. The system demonstrated that academic history is the most reliable indicator of future success. The project provided valuable hands-on experience in the data science lifecycle, from cleaning raw data to interpreting statistical error metrics. The resulting model serves as a functional proof-of-concept for automated student monitoring systems.

13. Future Work

To improve the system's utility and accuracy, the following extensions are proposed:

- **Non-Linear Models:** Implementing Random Forest or Gradient Boosting algorithms to better capture complex social interactions.
- **Feature Engineering:** Creating composite features (e.g., a "Risk Score" combining failures and absences) to improve sensitivity.
- **UI Integration:** Developing a web-based frontend (using Flask or Streamlit) where teachers can input data via a form rather than running Python scripts.

14. References

- P. Cortez and A. Silva. *Using Data Mining to Predict Secondary School Student Performance*. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
- UCI Machine Learning Repository. Student Performance Data Set.
- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

15. Appendix

Source code snippet (Model training):

```
X = df.drop('G3', axis=1)
y = df['G3']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()

model.fit(X_train, y_train)

print("Model Trained Successfully!")
```

Date: 5th January 2026
Student Name and Signature: Mohammed Khan

Name and Signature of the Evaluator

Date