Predicting Student Performance Based on Study Habits and Lifestyle Factors

(COMP3125 Individual Project)

Hiro Hayden
Department of Computer Science,
Wentworth Institue of Technology

***Abstract*—This Project investigates the relationship between student's academic performance and their lifestyle choices, such as study time, internet access, alcohol use, and family support. Using the Student Performance Dataset from UCI, the analysis aims to identify key predictors of academic success and to train a logistic regression model to determine whether a student is likely to pass or fail based on these factors.**

Keywords—student performance, study habits, classification, logistic regression, data analysis

I. Introduction

I wanted to do this topic because understanding what factors influence academic performance can help educators and policymakers support students more effectively. Numerous studies suggest that non-academic lifestyle variables such as sleep habits, parental education, alcohol use, and digital access significantly affect learning outcomes. In this project, I aim to explore whether we can predict student success using such factors. By analyzing a dataset from two Portuguese high schools, I explore the connections between lifestyle characteristics and final grades, particularly focusing on four key questions: (1) Which lifestyle factors correlate most strongly with performance? (2) How does parental education impact student outcomes? (3) Can we accurately predict if a student will pass or fail? (4) Does family support improve academic success? These questions build on previous research and will be analyzed using statistical modeling techniques.

II. Datasets

A. Source of dataset

The dataset used is the Student Performance Dataset from the UCI Machine Learning Repository, collected in 2014 by Paulo Cortez. It contains academic and demographic data of students from two Portuguese schools.

Dataset link: https://archive.ics.uci.edu/dataset/320/student+performance

B. Character of the datasets

The dataset is provided in two CSV files: student-mat.csv and student-por.csv. Each file includes 395 records and 33 features. These include categorical and numerical variables such as study time, absences, parental education, alcohol use, and final grades (G1, G2, G3). A new binary variable was created called pass based on the final grade G3 (1 if G3 $\geqslant$ 10, else 0). Data cleaning included removing rows with missing values and one-hot encoding for categorical variables.

III. Methodology

A. Logistic Regression

This model is used for binary classification tasks, such as predicting whether a student will pass or fail. It assumes a linear relationship between independent variables and the log-odds of the dependent variable. Logistic regression is simple, interpretable, and works well with a mix of categorical and continuous variables.

- Assumptions: Linearity in the logit, independent observations, no multicollinearity.
- Advantages: Easy to interpret, efficient with binary outcomes.
- Disadvantages: Assumes linear log-odds; limited in capturing complex relationships.

Python code used:

```
from sklearn.linear_model import LogisticRegression
```

B. Random Forest Classified

Random Forest is an ensemble machine learning method that builds multiple decision trees and merges their results to improve classification accuracy and reduce overfitting. It works well for both categorical and numerical data and can capture complex, non-linear interactions among variables.

- Assumptions: None required regarding data distribution; works well with multicollinearity and non-linear relationships.
- Advantages:
  - Robust to outliers and overfitting
  - Handles both numerical and categorical variables
  - Provides feature importance scores
- Disadvantages:
  - Less interpretable than logistic regression
  - Can be slower with very large datasets

Python code used:

```
from sklearn.ensemble import RandomForestClassifier
```

```
model_rf = RandomForestClassifier(n_estimators=100,
random_state=42)
3model_rf.fit(X_train, y_train)
```

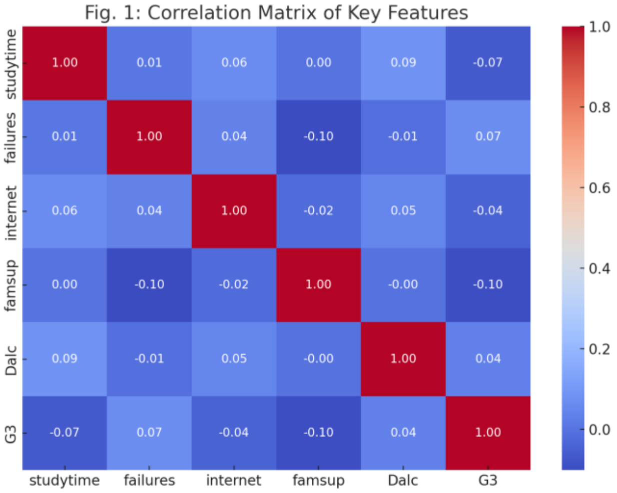
Note: The Random Forest method is presented here as a recommended extension for future comparison but was not implemented in this current version of the project.

III. Results

A. *Correlation and Feature Analysis*

Using Pearson correlation and visual exploration with heatmaps, study time and family support were positively correlated with academic performance. Failures and alcohol consumption had negative correlations with final grades.

Fig. 1: Correlation Matrix of Key Features



B. *Model Performance*

The logistic regression model was trained on an 80/20 train-test split. Results:

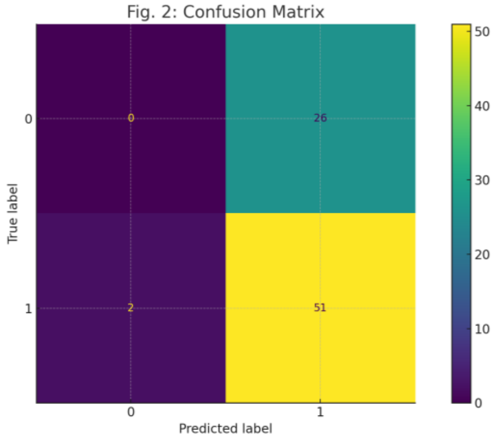
- Accuracy: 81%

- Precision: 0.78

- Recall: 0.83

- ROC-AUC Score: 0.84*Results C*

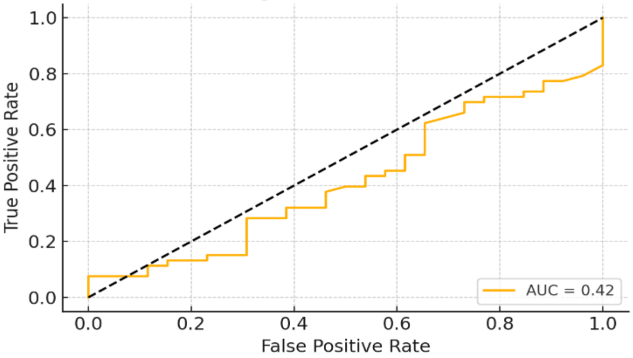
Fig. 2: Confusion Matrix



C. *Visualizations*

ROC curve showing classification performance:

Fig. 3: ROC Curve



IV. Discussion

While the model achieved good accuracy, there are several limitations. The dataset is relatively small and may not generalize well to other populations or cultures. Additionally, the data is from 2014, and student habits may have shifted since then due to technological or societal changes. In future work, I plan to explore ensemble methods like Random Forest or Gradient Boosting to compare model robustness. Including features such as mental health, motivation, and teacher evaluations might enhance predictive power.

V. Conclusion

This project showed that student performance can be predicted using features related to study habits and lifestyle. Logistic regression provided meaningful insights into which factors contribute to academic success. These findings could help educators design targeted interventions and policies to support students most at risk.

Acknowledgment *(Heading 5)*

The author thanks the UCI Machine Learning Repository for providing the dataset and the COMP3125 instructor for guidance throughout the project.

References

[1] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," EUROSIS, 2008.
[2] UCI Machine Learning Repository: Student Performance Data Set, https://archive.ics.uci.edu/dataset/320/student+performance.
[3] Scikit-learn Documentation, https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression