Studytime Report
(COMP3125 Individual Project)

Cohen Rollins

ABSTRACT— THIS PROJECT USES A REAL-WORLD SECONDARY SCHOOL DATASET TO INVESTIGATE HOW STUDY HABITS AND RELATED BEHAVIORS INFLUENCE STUDENT ACADEMIC PERFORMANCE. USING THE UCI STUDENT PERFORMANCE DATASET, I CREATED A BINARY PASS OR FAIL LABEL FROM THE FINAL GRADE AND TRAINED TWO CLASSIFICATION MODELS, LOGISTIC REGRESSION AND RANDOM FOREST, TO PREDICT WHETHER A STUDENT WOULD PASS A CLASS. AFTER PREPROCESSING THE DATA THROUGH ONE-HOT ENCODING OF CATEGORICAL VARIABLES AND A TRAIN TEST SPLIT, BOTH MODELS ACHIEVED RELATIVELY HIGH PREDICTIVE ACCURACY. LOGISTIC REGRESSION REACHED AN ACCURACY OF ABOUT 0.85, WHILE RANDOM FOREST PERFORMED SLIGHTLY BETTER WITH AN ACCURACY OF ABOUT 0.87 AND AN F1 SCORE OF 0.90. FEATURE IMPORTANCE ANALYSIS SHOWED THAT PRIOR GRADES, NUMBER OF PAST FAILURES, AND ABSENCES WERE THE MOST INFLUENTIAL PREDICTORS OF FINAL PERFORMANCE. THE RESULTS SUGGEST THAT A COMBINATION OF CONTINUOUS ACADEMIC PERFORMANCE AND BEHAVIORAL FACTORS CAN BE USED TO IDENTIFY STUDENTS AT RISK AND COULD SUPPORT DATA-INFORMED INTERVENTIONS IN EDUCATIONAL SETTINGS.

KEYWORDS— STUDENT PERFORMANCE, CLASSIFICATION, LOGISTIC REGRESSION, RANDOM FOREST, EDUCATIONAL DATA

I. INTRODUCTION

Understanding the factors that influence student academic performance is a widely studied topic in education research, as performance outcomes can significantly impact future academic and career opportunities. College students frequently adopt different learning strategies and behaviors, such as attending class regularly, dedicating time to studying, or participating in class, that may contribute to their success. Identifying which of these behavioral elements are most influential can help instructors, advisors, and academic institutions provide more targeted support to students.

Existing studies have examined how variables such as parental involvement, social background, study time, and school environment relate to academic achievement. Many findings suggest that consistent attendance and regular study habits are strongly correlated with higher grades, while poor engagement often predicts academic difficulty. With the increasing availability of public educational datasets, data-driven approaches can now be used to systematically evaluate how these variables interact.

This project explores these relationships using a real-world student academic dataset. The analysis focuses on identifying key behavioral predictors of final grades, determining whether student success can be accurately predicted using machine learning models, and comparing study habits between high-performing and low-performing students. By combining descriptive statistics, visualizations, and classification techniques, this project aims to present a deeper understanding of the academic factors that influence performance and how these insights may support educational decision-making.

II. METHODOLOGY

This section describes the data preprocessing steps and the two classification methods used in this project: Logistic Regression and Random Forest. The overall workflow included data cleaning, feature engineering, train test splitting, model training, and performance evaluation.

A. Data preprocessing and feature engineering

The original CSV file from the UCI repository uses semicolons as separators, so the dataset was loaded with sep=";". I focused on the mathematics course dataset (student-mat.csv), which contains 395 rows and 33 input attributes.

To transform the final grade into a more interpretable target variable, I created a binary feature called **pass_fail** based on the third period grade G3. Students with G3 greater than or equal to 10 were labeled as pass (1), while students with G3 below 10 were labeled as fail (0). This pass or fail label became the response variable for all classification models.

Most of the original variables are categorical, such as school, family support, parental job, and reason for choosing the school. These columns were converted into numerical form using one hot encoding with the pandas.get_dummies function, dropping the first level of each category to avoid perfect multicollinearity. After encoding, I removed the original G1 and G2 columns only from the feature matrix to avoid direct leakage of the target but kept them in the encoded dataset for correlation analysis and feature importance inspection. The final feature matrix **X** consisted of all encoded attributes except G3 and pass_fail, and the target vector **y** was the pass_fail column.

The dataset was split into training and testing subsets using an 80/20 split with stratification on the pass_fail label to preserve the original class balance. This split was implemented using train_test_split from scikit learn. For Logistic Regression, I standardized the numerical features using StandardScaler so that all variables would be on a similar scale, which helps the optimization procedure converge more reliably.

B. Logistic Regression

Logistic Regression is a linear classification method that models the log odds of the probability that an observation belongs to the positive class as a linear combination of the input features. In this project, it estimates the probability that a student will pass the course given their study habits, background, and past performance.

The main assumptions are that there is a linear relationship between the predictors and the log odds of the outcome, that the observations are independent, and that there is no extreme multicollinearity among features. The advantages of Logistic Regression include interpretability and relatively low computational cost, while its main limitation is that it may struggle when the decision boundary is highly nonlinear.

I implemented Logistic Regression using sklearn.linear_model.LogisticRegression with a maximum of 200 iterations and default regularization settings. The model was trained on the scaled training data and evaluated on the scaled test data. Performance was measured using accuracy, precision, recall, and F1 score, along with a confusion matrix that highlights the counts of correctly and incorrectly classified students.

C. Random Forest Classifier

Random Forest is an ensemble learning method that constructs many decision trees on different bootstrap samples of the data and averages their predictions for classification. Each tree is built using a random subset of features at each split, which decorrelates the trees and usually improves generalization.

Random Forest does not rely on linear decision boundaries and can naturally capture complex, nonlinear relationships between features and the target. It is also relatively robust to outliers and noisy features. However, individual trees within the forest are less interpretable than Logistic Regression coefficients, and the model can be more computationally expensive.

In this project, I used sklearn.ensemble.RandomForestClassifier with 300 trees and a fixed random seed for reproducibility. The model was trained on the unscaled training data, and predictions were generated on the test set. The same evaluation metrics as in the Logistic Regression experiment were recorded. In addition, I extracted feature importance scores from the trained Random Forest to identify which variables contributed most to the pass or fail predictions.

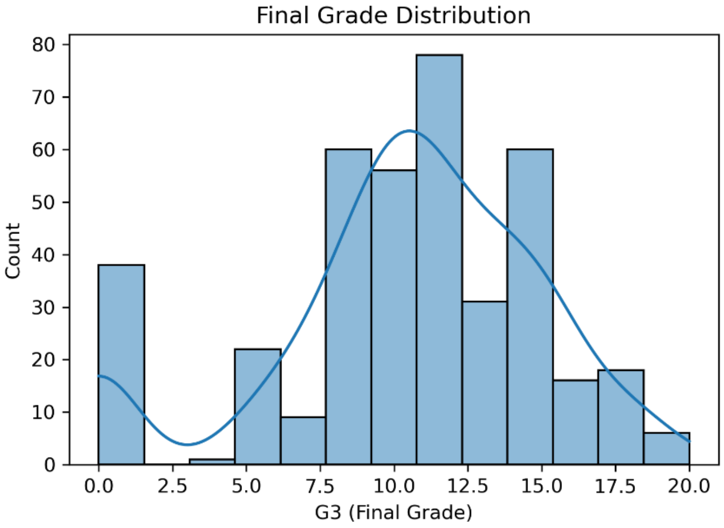
III. Results

This section presents descriptive statistics of the final grades, the performance of the two classification models, and the most influential features associated with student success.

A. Descriptive analysis of final grades

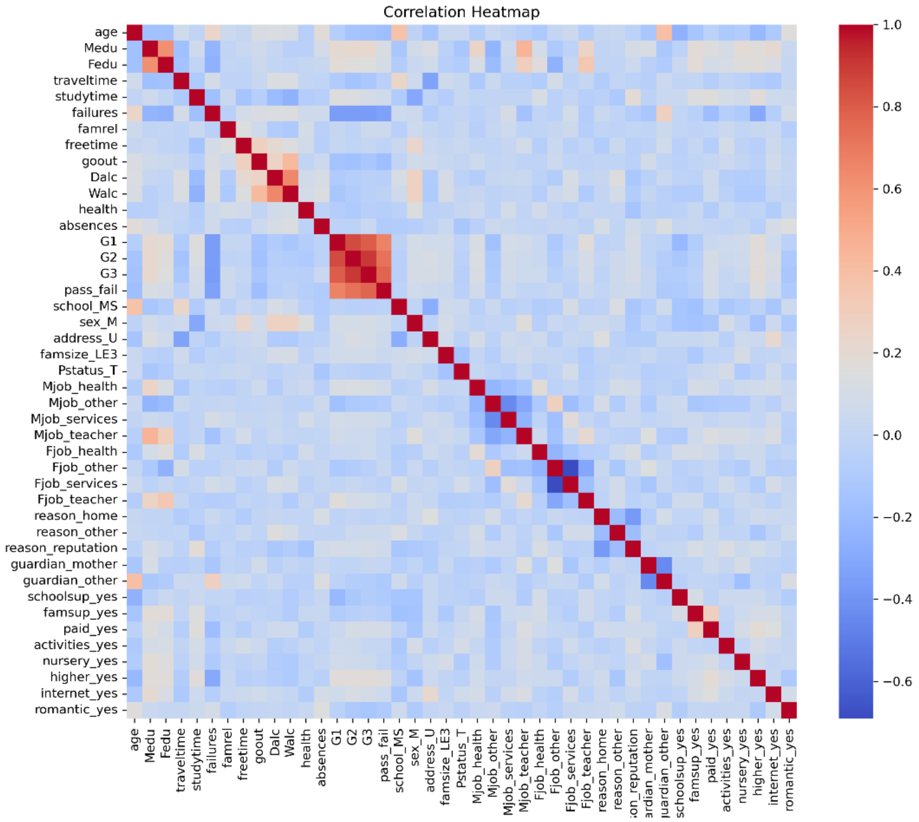
The distribution of the final grade G3 was approximately unimodal with a slight skew toward higher values. Most students scored between 8 and 15 on the 0 to 20 grading scale, and relatively few students achieved the extreme scores near 0 or 20. After converting G3 into the binary pass_fail label, the majority of students in the dataset were labeled as passing, which is reflected in the class counts used during model training and testing.

Fig. 1. Distribution of final grades (G3).



The correlation heatmap of the encoded dataset showed a strong positive correlation among G1, G2, and G3, which is expected because they represent first, second, and final period grades. There were also moderate correlations between G3 and other variables such as number of past failures and absences, whereas social and lifestyle variables such as going out or alcohol consumption were more weakly related to the final grade.

Fig. 2. Correlation heatmap of features.



B. Model performance

1) Table I summarizes the classification performance on the held out test set for both Logistic Regression and Random Forest.

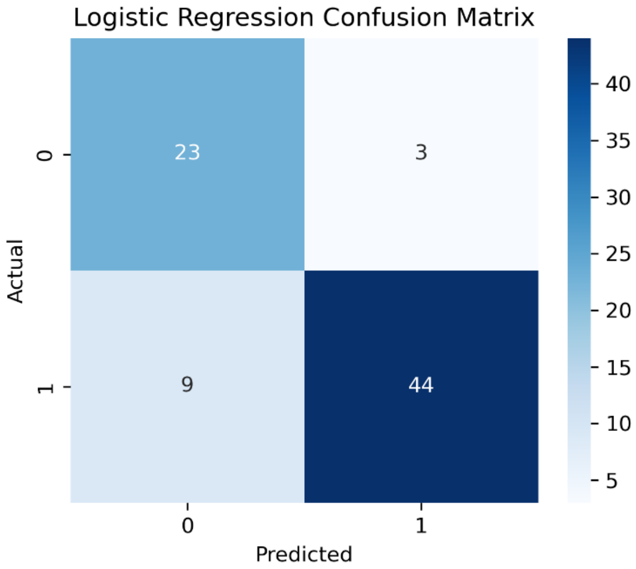
Table I

Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.8481	0.9362	0.8302	0.88
Random Forest	0.8734	0.9574	0.8491	0.90

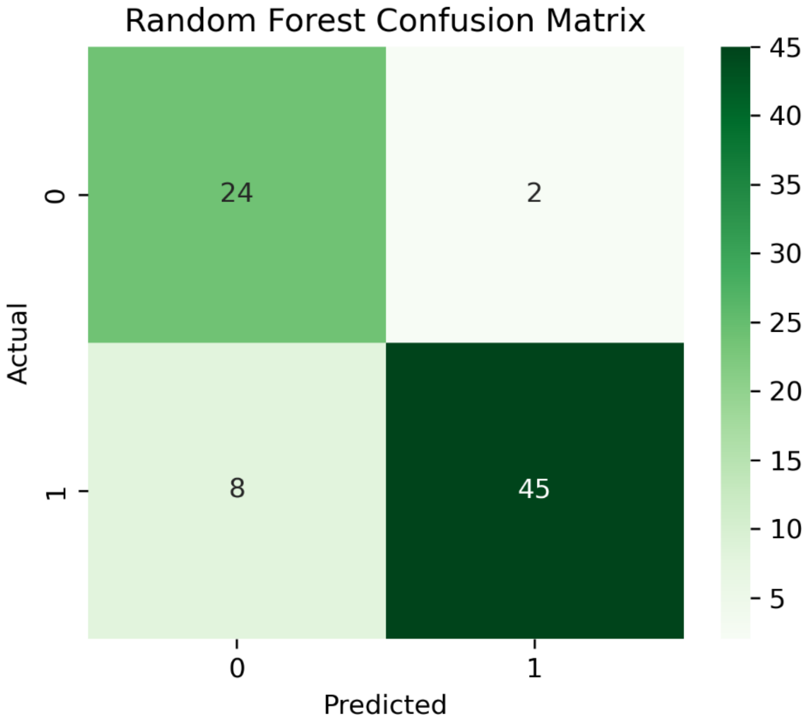
- **Logistic Regression** achieved an accuracy of **0.8481**, precision of **0.9362**, recall of **0.8302**, and an F1 score of **0.88**. The confusion matrix \[\[23,3, 9,44]] shows that the model correctly classified 44 passing students and 23 failing or borderline students, while it misclassified 3 students who actually failed and 9 students who actually passed.

Fig. 3. Confusion matrix for Logistic Regression.



- **Random Forest** achieved slightly higher performance, with an accuracy of **0.8734**, precision of **0.9574**, recall of **0.8491**, and an F1 score of **0.90**. Its confusion matrix \[\[24,2, 8,45]] indicates that the Random Forest correctly identified 45 passing students and 24 failing students, while misclassifying only 2 failing students as pass and 8 passing students as fail.

Fig. 4. Confusion matrix for Random Forest.



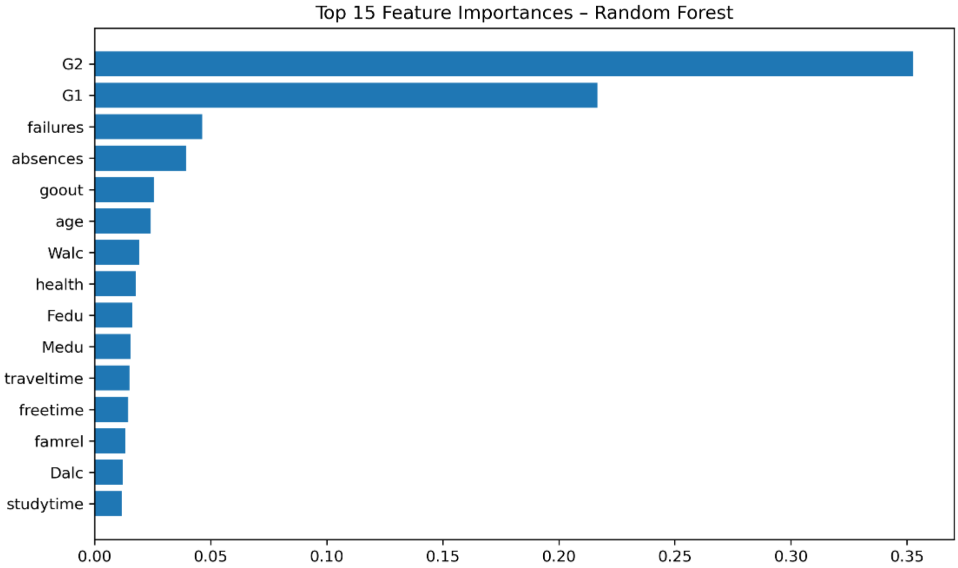
Both models performed well, but the Random Forest classifier consistently produced higher accuracy and F1 scores, suggesting that it captures more of the underlying structure in the data than the linear Logistic Regression model.

C. Feature importance

Feature importance scores from the Random Forest highlight which variables contributed most to the pass or fail predictions. The most important features were the second period grade **G2** and the first period grade **G1**, with importance values of approximately 0.35 and 0.22, respectively. These were followed by the number of past **failures**, **absences**, how often students **go out**, **age**, and both weekend and weekday alcohol consumption (Walc and Dalc). Health status, parental education levels (Medu and Fedu), travel time to school, free time, family relationship quality, and studytime also had nontrivial importance scores.

These results indicate that ongoing academic performance and academic history (G1, G2, failures) are the strongest predictors of final success, but behavioral factors such as attendance, going out, and alcohol use also play noticeable roles.

Fig. 5. Feature importance scores from Random Forest.



IV. Discussion

The results suggest that student performance in this dataset can be predicted with reasonably high accuracy using a combination of academic and behavioral features. Both Logistic Regression and Random Forest achieved accuracies above 0.84, which means that most students were correctly classified as passing or failing based on their attributes.

Random Forest outperformed Logistic Regression across all major metrics, especially in F1 score and accuracy. This difference is likely due to the ability of Random Forest to capture nonlinear interactions between features. For example, the influence of absences may depend on previous grades or the number of past failures, relationships that cannot be fully represented by a single linear decision boundary. The ensemble of decision trees in the Random Forest can model such complex patterns more flexibly.

The feature importance analysis confirms that prior grades G1 and G2 dominate the predictions, which is intuitive: students who do well earlier in the term are much more likely to achieve a passing final grade. The number of past failures and absences also contribute strongly, indicating that repeated struggle and poor attendance are warning signs for lower final outcomes. Lifestyle variables such as going out and alcohol consumption showed smaller but still meaningful influence, suggesting that social behavior may have an indirect connection to academic performance.

Despite these strengths, the models have limitations. The dataset comes from a specific context, Portuguese secondary schools, and may not generalize to other educational systems or college level courses without retraining. Many important factors, such as motivation, mental health, quality of instruction, or access to tutoring, are not captured in the dataset. In addition, using G1 and G2 as predictors could limit the usefulness of the model early in the course when these grades are not yet available.

Future work could explore models that rely more heavily on early semester behavioral data, such as attendance patterns or studytime, and less on later grades, so that at risk students can be flagged sooner. It would also be valuable to compare additional algorithms such as Gradient Boosting or XGBoost, or to apply cross validation and hyperparameter tuning to further improve performance.

V. Conclusion

This project used the UCI Student Performance dataset to examine how study habits, attendance, and related factors influence academic success, and to evaluate whether student pass or fail outcomes can be predicted using supervised learning models. After preprocessing the data and engineering a binary pass_fail target from the final grade, two models were trained and evaluated: Logistic Regression and Random Forest.

Both models achieved strong predictive performance, with Logistic Regression reaching an accuracy of about 0.85 and Random Forest improving this to about 0.87 and an F1 score of 0.90. The analysis showed that prior period grades, number of past failures, and absences are the most influential predictors, while behavioral variables such as going out and alcohol use also contribute to the predictions.

These findings indicate that a combination of academic history and behavioral data can be used to build practical early warning tools for educators. Schools could use similar models to monitor students, identify those at risk of failing, and provide targeted support such as advising, tutoring, or attendance interventions. At the same time, the limitations of the dataset highlight the importance of collecting richer and more diverse information to fully understand the complex factors driving student performance.

References

[1] [1] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in Proc. 5th Future Business Technology Conf., 2008.

[2]

[3] [2] UCI Machine Learning Repository, "Student Performance Data Set," Accessed: Nov. 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/320/student+performance

[4]

[5] [3] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[6]

[4] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.