Analyzing Student Performance Using Recognition Models

Matthew Maloney  
Department of Computer Science, Wentworth Institute of Technology

# Abstract

This project explores how various student characteristics, habits, and environments affect academic performance. By analyzing a dataset of student performance from Kaggle, this study investigates the relationship between final grades and factors such as study time, internet access, and parental education. Regression models are used to determine which variables most strongly predict academic success. Results offer insights for educators and policymakers aiming to improve student outcomes.

# Keywords

student performance, regression, education analytics, machine learning, academic success

# I. INTRODUCTION

Understanding the factors that contribute to academic success is vital in improving student outcomes and creating effective education policies. Students’ performance is influenced by a combination of personal habits, socioeconomic background, and learning environments. Prior research has explored these elements individually, but this project aims to evaluate them holistically using data science techniques. Questions such as whether study time matters more than parental education or whether internet access improves learning outcomes are central to this analysis. By applying regression models to a publicly available dataset, this report identifies which attributes have the most significant impact on final student grades.

# II. DATASETS

## A. Source of Dataset

The dataset used in this project was sourced from Kaggle, a reputable platform for machine learning datasets. The dataset, titled \*Students Performance in Exams\*, was uploaded by user "spscientist" and is available at: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams. The data is publicly accessible and has been used in various educational data mining studies.

## B. Character of the Datasets

The dataset is provided in CSV format and includes 1000 entries. Each row represents a student, and the columns capture demographic information, parental education, test preparation status, study habits, and scores in math, reading, and writing. The target variable is the average of the three test scores. A new column was created to represent this average, which serves as the dependent variable for regression analysis. Categorical features such as gender, parental level of education, and internet access were encoded appropriately to be used in model training. Data cleaning included handling missing values, verifying consistency in labels, and creating dummy variables for categorical data.

# III. METHODOLOGY

## A. Linear Regression

Linear Regression was used as a baseline model to understand the direct relationships between features and final grades. The assumption here is that the dependent variable (average grade) has a linear relationship with the independent variables. It is simple to interpret but sensitive to outliers and multicollinearity.

## B. Random Forest Regression

To improve predictive performance and handle non-linear relationships, a Random Forest Regressor was also applied. This ensemble method uses multiple decision trees and averages their outputs, reducing overfitting and handling feature interactions better. Important parameters such as the number of trees (n\_estimators) and maximum tree depth (max\_depth) were tuned using grid search and cross-validation.

## C. Feature Engineering

New features such as total score and average score were added. Categorical variables were one-hot encoded. Data was split into training (80%) and testing (20%) sets, and standard scaling was applied where needed.

# IV. RESULTS

## A. Linear Regression Results

The linear model achieved an R² score of 0.72 on the test data, indicating that 72% of the variance in student performance can be explained by the model. Key predictors included reading score, test preparation status, and parental education.

## B. Random Forest Regression Results

Random Forest yielded a higher R² of 0.85, showing superior performance. Feature importance ranking showed that:  
1. Reading score  
2. Writing score  
3. Math score  
4. Test preparation  
5. Parental education  
6. Internet access  
7. Study time

## C. Visualization

Bar charts and feature importance plots were generated using matplotlib and seaborn. These visuals confirmed that academic habits (like preparation and study time) and support systems (parental education, internet access) contribute significantly to student outcomes.

# V. DISCUSSION

While the regression models performed well, the analysis has limitations. Study time and internet access were self-reported or derived from limited proxies, which could affect accuracy. In future work, larger and more diverse datasets could help validate these findings. Including additional predictors such as classroom environment or teacher quality could further improve model robustness.

# VI. CONCLUSION

This project demonstrated that student performance can be reliably predicted using regression models. Factors such as reading and writing scores, parental education, and test preparation course completion emerged as strong predictors. These results suggest that interventions focused on improving these areas may have the most significant impact on student outcomes.