Exploring Factors Affecting Student Performance

Selena Buttery, Chaminade University of Honolulu, DS495, Fall 2024

Introduction

Student performance is a critical topic today as academic success increasingly shapes future opportunities. Understanding the most influential factors behind student performance can inform targeted interventions to support students. This study aims to identify the key variables that predict students' grades, using machine learning to develop models that give us a better understanding of academic success.

Background

Previous studies have shown that student performance is influenced by factors such as socioeconomic background,study time, and preparation for courses/exams. This research builds on existing literature by incorporating datasets that include demographics and academic features to predict students performance by averages. In a study conducted in 2005, they focused on socioeconomic status and the relationship to achievement using a meta-analysis approach (Sirrin, 2005).

Research Question/Hypothesis

Research Question: What are the key factors that influence student performance the most?

Hypothesis: Students who engage in test preparation courses and have studied more will perform better academically.

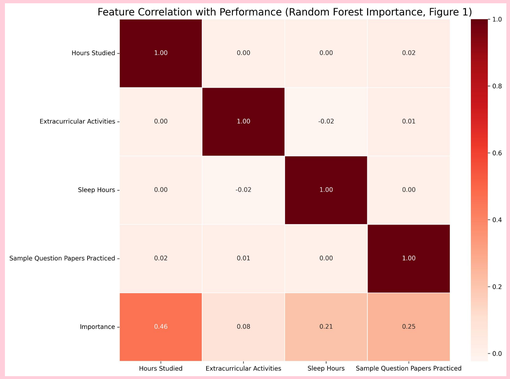
Methodology

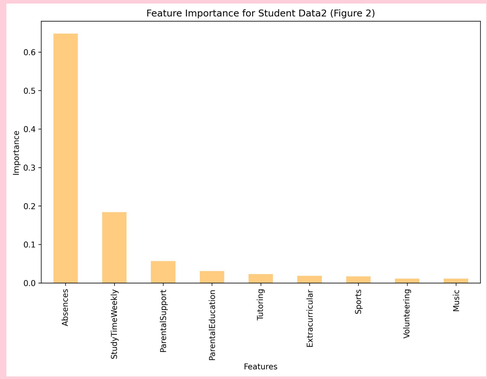
Data and Tools: Data was sourced from Kaggle and analyzed using Python in Google Colab.

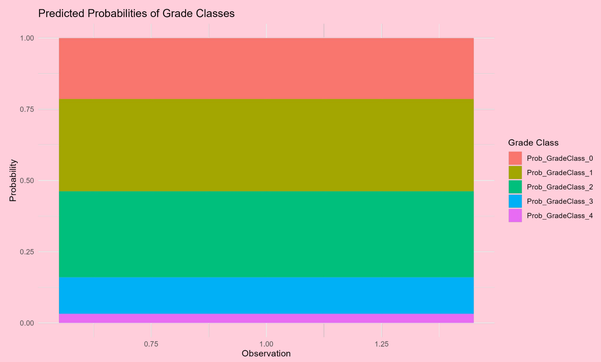
Analyses:

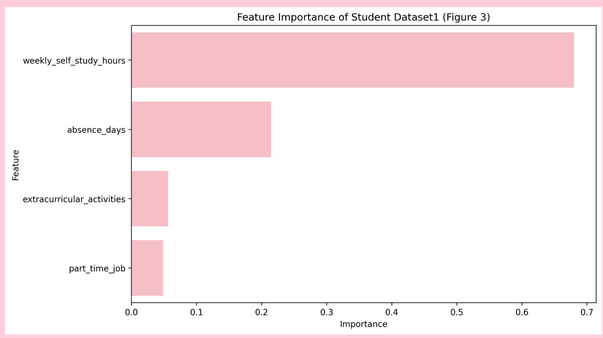
- data cleaning
- encoding categorical variables
- normalizing numeric features
- Using random forest algorithms, I analyzed variables like study hours, parental education, and test preparation to predict student performance.
- Used random forest to create probabilities of being in the 5 different grade classes.
- Machine learning models provided the analytical depth required to uncover patterns in academic success based on complex variables

Figures/Analyses on Student Performance Data









Preliminary Results/ Findings

- **Key Influences** : "Hours studied" showed a moderate positive influence on grades (correlation: 0.46), while practicing sample questions also contributed positively (correlation: 0.25).
- **Attendance Impact** : In two separate datasets, absences were the most important features correlating to lower grades which I did not predict in my hypothesis.
- **Probabilities of Grades** : In the bottom visual there is a snippet of code from the final dataset used. This was found using random forest models and shows the proabbiltiy of getting certain grades (GradeClass 0 = A, 1 = B, 2 = C, 3 = D, 4 = F)

Discussion and Results

My hypothesis suggested that test preparation and more study hours improve academic performance. The findings partially support this: "hours studied" showed a moderate correlation (0.46) with grades, and practicing sample questions also had a positive effect (0.25). However, in two different datasets, it showed that absences were the most important among features as well as study time weekly for both. These results suggest a complex relationship between study habits, attendance, and performance. These results can be seen in the figures, feature importance was run separately for Figures 2 and 3 with separate datasets, as figure 1 is also a different dataset for the correlation matrix.

Limitations:

- data quality (self-reported culd be unreliable)
- Sample size limitations
- Correlation vs causation

Acknowledgement

- Chaminade University Data Science Analytics and Visualization Program
- NSF ALL-SPICE ALLIANCE program

References

Ansodariya, D. (2022, May 26). *Student performance dataset*. Kaggle. https://www.kaggle.com/datasets/devansodariya/student-performance-data

Narayan, N. (2023, June 29). *Student performance (multiple linear regression)*. Kaggle. https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression

Rattanaporn, K. (2023, March 12). *Student performance prediction*. Kaggle. https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics