How do student traits affect academic performance in Cyprus?

Project Technical Report

John Carrillo

Michal Kuderski

Ethan Meadows

05/08/2025

1. Abstract Summary

a. The Higher Education Students Performance data set was collected from students in the Faculty of Engineering and Faculty of Educational Sciences in 2019. The data sets' primary objective is to predict the success or failure of a student; this is dependent on their responses to a survey regarding their background traits and academic habits. There were 145 students surveyed, and 30 predictor variables were quantified through the survey. These variables ranged from personal, familial, and educational matters. These include but are not limited to, demographic details, study habits, and overall academic engagement. The dataset is specifically designed for classification tasks, which makes it useful for educational research and creating projective models in student performance assessments. Lastly, no missing data values were reported, ensuring the level of data integrity for a thorough analysis.

2. Introduction and Objectives statement

- Student Performance Prediction is a growing field for multiple higher-level institutes across the globe. Through this they hope to understand better what will bolster a student's rates of success and what will hinder them. The Higher Education Students Performance data set offers valuable insights through the surplus sample size and wide range of variables that highlight areas of concern, both academically and personal, that will impact a student's outcome. Ultimately this data set aims to achieve the following objectives:
    a. **Predict student performance** based on various demographic, academic, and behavioral features of students
    b. **Identify correlations** between these features and determine the strength of impact on the outcome
    c. **Support Educational Research** that aims to identify strategies that can be implemented to improve student success rates
    d. **Facilitate Machine Learning Approaches** for better classification and performance modeling of students in higher education environments

3. Data description / data preprocessing / sources, etc.

a. Data Description:

i. **Data Source:**

1. Our dataset contains information on 145 students enrolled in various courses at a higher education institution in Cyprus. The data was collected through student surveys and administrative records. Each observation is recorded as a unique student, with variables considering factors, such as demographic information, study habits, academic performance, and family background.

ii. **Variables:**

   1. The dataset includes the following variables:

iii. **X1-X30:**

   1. Categorical survey responses (coded as numerical values for simplicity)

iv. ***COURSE.ID*:**

   1. Course identifier (values 1 - 9)

v. ***GRADE*:**

   1. Student grade (values 0-7, in ascending performance levels)

      a. The grade scale (for the response variable) is defined as:

         i. 0 = Fail
         ii. 1 = DD
         iii. 2 = DC
         iv. 3 = CC
         v. 4 = CB
         vi. 5 = BB
         vii. 6 = BA
         viii. 7 = AA

vi. **Key Predictor Variables:**

   1. ***Student Age* (X1):**

      a. 1 = 18 - 21 years, 2 = 22 - 25 years, 3 = 26+ years

   2. ***Weekly Study Hours* (X3):**

      a. 1 = None, 2 = < 5 hours, 3 = 6-10 hours, 4 = 11 – 20 hours, 5 = >20 hours

   3. ***Scholarship Type* (X4):**

      a. 1 = None, 2 = 25%, 3 = 50%, 4 = 75%, 5 = Full

   4. ***Class Attendance* (X19):**

      a. 1 = Always, 2 = Sometimes, 3 = Never

   5. **Parents' Education (X11: Mother's Education, X12: Father's Education):**

      a. 1 = Primary school, 2 = Secondary school, 3 = High school, 4 = University, 5 = MSc, 6 = PhD

b. Data Preprocessing:

i. We performed the following preprocessing steps to prepare the data for analysis:

   1. **Data Cleaning:** We verified that there were no duplicate student IDs and checked for missing values. No missing values were found in the dataset.

2. **Variable Recoding**: We properly recoded *COURSE.ID* and *GRADE* as categorical variables, which is crucial given their nominal (unordered) and ordinal (ordered) nature. This addresses the feedback we received in the exploratory analysis about effectively encoding these variables.
3. **Variable Relabeling**: We renamed the numerical variables X1 – X30 to match their corresponding survey questions, making the analysis more interpretable and, therefore, meaningful. This resolves the professor's concern about using meaningful variable names instead of the meaningless and uninterpretable X1 – X30 labels.
4. **Data Validation**: We verified the ranges and distributions of all variables to guarantee they fall within the expected parameters. All categorical variables were checked to confirm they contained only valid category codes.

c. **Sources**

i. Yilmaz, N. & Şekeroğlu, B. (2019). Higher Education Students Performance Evaluation [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C51G82.

ii. *****NOTE:** Our dataset from the site: https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation includes the following citation under "Citations/Acknowledgements:"

1. YÄ±lmaz N., Sekeroglu B. (2020) Student Performance Classification Using Artificial Intelligence Techniques. In: Aliev R., Kacprzyk J., Pedrycz W., Jamshidi M., Babanli M., Sadikoglu F. (eds) 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019. ICSCCW 2019. Advances in Intelligent Systems and Computing, vol 1095. Springer, Cham.

a. (This does NOT include the DOI as other citation, which should be the proper format for our source. I assume this citation was edited by the creators and not the one in APA format, which I found from clicking the yellow "Cite" button found above in our dataset site. ***

4. Exploratory Analysis

a. Summary Statistics:

i. Located in the other PDF file containing all the code and visualizations, the summary statistics table shows the following information pertaining to the key predictor variables:
    1. ***Student Age:***
        a. Mean: 1.68
        b. SD (Standard Deviation): 0.57
        c. Min (Minimum): 1
        d. Max (Maximum): 3
        e. Description: Age category
    2. ***Weekly Study Hours:***
        a. Mean: 2.11
        b. SD: 0.51
        c. Min: 1
        d. Max: 3
        e. Description: Hours studying weekly
    3. ***Scholarship Type:***
        a. Mean: 3.57
        b. SD: 0.82
        c. Min: 1
        d. Max: 5
        e. Description: Scholarship percentage
    4. ***Mother's Education:***
        a. Mean: 2.35
        b. SD: 1.16
        c. Min: 1
        d. Max: 6
        e. Description: Education level
    5. ***Father's Education:***
        a. Mean: 2.71
        b. SD: 1.19
        c. Min: 1
        d. Max: 6
        e. Description: Education level
    6. ***Class Attendance:***
        a. Mean: 2.06
        b. SD: 0.56
        c. Min: 1
        d. Max: 3

e. Description: Attendance frequency

7. ***COURSE.ID:***

a. Mean: 4.19
b. SD: 3.26
c. Min: 1
d. Max: 9
e. Description: Course identifier

8. ***GRADE:***

a. Mean: 3.14
b. SD: 2.20
c. Min: 0
d. Max: 7
e. Description: Grade performance

ii. **Results & Interpretations:**

1. *Scholarship Type* (X4) has the highest mean (3.57) and ranges from 1 – 5. This indicates that many students in Cyprus receive substantial financial aid. The distribution is skewed right, meaning that more students receive higher scholarship percentages.
2. *Seminar Attendance* (X20) displays the lowest variability, with a standard deviation of only 0.41. This suggests consistency (a constant value) in student participation in various departmental seminars and conferences.
3. *Parents' Education (Mother's & Father's Education:* X11 & X12) both have large ranges (1 – 6), signifying diversity in family educational backgrounds.
4. The grade distribution spans the full range (0 – 7), with a mean of 3.14 (close to a "CC" grade; like our "C" or "C-" grades)

b. <u>**Distribution Analysis**</u>

i. **Course Enrollment:**

1. The distribution of students across courses is highly skewed, which can be seen in our file containing the complete code/visualizations. Course 1 has the highest enrollment with over three times more students than the next most enrolled course (Course 9). This course enrollment imbalance was explained in our modeling approach, initially commented about potential bias by the professor.

ii. **Distribution:**

1. The skewed distribution of students across courses presents a limitation, as some courses have limited sample sizes for reliable

inference. We accounted for this downside in our statistical modeling by using appropriate techniques that are stable for uneven course sample sizes (i.e., number of students enrolled per course).

iii. ***Grade* Distribution:**

1. The grade distribution exhibits a multimodal pattern with peaks at grades 1, 5, and 7 (DD, BB, and AA). This implies distinct performance clusters among the student population. The relative lack of failing grades (0) proposes that most students achieve at least marginal passing performance.

iv. **Categorical Variable Distributions:**

1. We analyzed the distributions of key categorical predictor variables to better understand the student population in this dataset.

   a. ***Student Age* Distribution:**

      i. This distribution is mainly skewed toward younger students, with most falling in the 18 – 21 or 22 – 25 age ranges. This is expected, due to typical undergraduate demographics, but indicates a relatively small population of older students.

   b. ***Mother's Education Level* Distribution:**

      i. The bimodal distribution of the mother's education level suggests two different groups within the student population:

         1. Those whose mothers have basic education (primary or secondary school)
         2. Those whose mothers have higher education (university degree or beyond)

      ii. This pattern may showcase a broader socioeconomic division within this dataset of only 145 students (relative to the entire student population of Cyprus).

c. Correlation Analysis

i. We performed correlation analysis to pinpoint relationships between predictor variables and grades:

1. **Correlation Matrix Heatmap:**

   a. We noticed several critical relationships:

      i. Strong Positive Correlations:

1. *Mother's Education* (X11) and *Father's Education* (X12) show a strong positive correlation (r = ~ 0.82). One could claim that this relationship represents educational homogamy in parents, people's tendency to marry those with similar educational backgrounds.
2. *Taking Notes in Class* (X25) and *Listening in Class* (X26) are strongly correlated (r = ~ 0.75), denoting consistent study behaviors.

ii. Correlations with Grade:

1. Cumulative GPA from previous semester shows the strongest positive correlation with current grades
2. *Parents' education levels* (X11, X12) both represent moderate positive correlations with grade.
3. *Scholarship Type* (X4) shows a weak negative correlation with grade (r = ~ 0.12), which was unexpected for us. This discrepancy led to further investigation into the accuracy of our correlation matrix heatmap.

d. Course-Specific Analysis

i. To comprehend course-specific effects, we examined grade distributions and student characteristics across different courses.

1. Grade Distribution by Course:

a. The analysis results in considerable variability in grade distributions across courses:

i. Courses 3, 5, and 7 have higher median grades

ii. Course 8 shows the lowest median grade and highest variability

iii. Course 1, despite having the highest enrollment, has moderate grade variability

2. Student Characteristic Distribution Boxplots:

a. We also looked at student characteristics varying across courses through boxplots comparing X1 – X3 (*Student Age*, *Sex*, *Weekly Study Hours*) across *COURSE.ID.* These boxplots showed:

i. Relatively consistent median student age (X1) across courses (1.5 - 2.0)
ii. Low variability in sex distribution (X2) across courses
iii. Some variability in weekly study hours (X3), with Course 5 having two outliers

5. formal analysis using statistical learning and model/variable selection methods

The methods used by Lasso Regression and Stepwise Selection for variable selection and model conservatism showed considerable variations. Known for its regularization method, Lasso was more conservative, keeping only five important predictors: cumulative GPA from the previous semester, student age, sex, high school type, and extra work. By decreasing less significant coefficients to zero, this selection method implies Lasso gave parsimony top priority, which could lessen the effect of multicollinearity.

Stepwise Selection, on the other hand, considered a wider range of twelve predictors, such as transportation, the father's kind of work, listening in class, and overall income—variables that Lasso completely disregarded. Even though multicollinearity or redundancy may have existed, the Stepwise model's inclusion of these factors suggests that it recognized their significant impact on student performance. This disparity offers a broader perspective of analysis and reflects Stepwise's propensity to be less forceful in eliminating predictors.

Despite these disagreements, Lasso and Stepwise both acknowledged the importance of high school type and the previous semester's cumulative GPA as crucial markers of academic achievement. This agreement emphasizes how reliable these variables are across various model selection strategies and how crucial they are to comprehending student outcomes.

An important finding from the initial analysis was the recognition of students 24, 78, 94, and 113 as prominent outliers. Their responses to the survey and their academic results did not correspond with the overarching trends seen in the larger dataset. These discrepancies indicate unusual behaviors or inconsistencies that may require further exploration to grasp their root causes.

6. conclusion:

The analysis conducted provides strong evidence that student performance is influenced by both academic background and socio-economic factors. Key predictors such as class engagement, prior GPA, and high school type emerged consistently across both LASSO and Stepwise methods, underscoring the importance of historical academic performance and structured learning environments. Gender-based analysis revealed different predictors for males and females, suggesting that tailored educational strategies may be effective in improving outcomes.

Additionally, outlier removal enhanced model stability, validating the importance of clean datasets in predictive analysis. While LASSO focused on a minimal subset of impactful variables, Stepwise introduced a broader perspective, capturing socio-economic factors like parental occupation and transportation.

Overall, the models explained approximately 45–48% of the variance in student grades, indicating that while significant progress was made in understanding academic success drivers, further exploration with non-linear methods or deeper socio-economic variables could provide even more insights.

7. The Coding file is being included in a separate PDF file as the R coding file’s size deems it necessary