# Students Performance and Demographic/Socioeconomic Factors

May 14, 2024

## Introduction

Examining the relationship between student demographics, socioeconomic factors, and academic performance is essential for promoting educational equality and overall student success. It's important that we identify disparities in educational opportunities and outcomes among different groups of students so that educators and policymakers can develop targeted interventions, support programs, and policies to address the specific needs of learners from diverse backgrounds. Understanding the interplay between student backgrounds and academic achievement can inform the creation of a more equitable education system, where every student has access to the resources and support they need to reach their full potential. Moreover, investigating these correlations can provide valuable insights into the factors that contribute to student success, enabling educators to tailor instructional strategies and learning environments to better support all students and address barriers that may hinder academic achievement for students from disadvantaged backgrounds.

### Dataset

Our research is based on a Kaggle dataset whose purpose is to predict college students' end-of-term performances using machine learning techniques. The data was collected from the Faculty of Engineering and Faculty of Educational Sciences students in 2019 from the University of California, Irvine. The data is multivariate and used for classification tasks in the social science subject area.

This dataset contains a total of 31 features (columns), each representing a demographic indicator, socioeconomic factor, or student behavior, including age, sex, parents' education, class attendance, class participation, GPA, etc. More specifically, columns 1 through 10 are personal questions such as the student's partner status, columns 11 through 15 are family questions such as parent occupation, and the rest of the columns include educational habits like reading frequency. Each feature is observed with an integer value - the details of each value will be described within each model we create. There are a total of 145 instances (rows), each representing a student, and there are no missing values.

After analyzing the dataset, we explored the following questions:

- Do demographic factors such as age, gender, ethnicity, socioeconomic status, and parental education) affect a student's performance and their final grade?
- Do academic habits such as listening in class, reading frequency, and preparation for midterms affect students' performance?

## Visualization and Analysis

### Linear Regression, Model 1

In our analysis, we employed a linear regression model (see **Figure 1**) using a formulaic approach because it is particularly advantageous when handling multiple categorical variables. This approach facilitates the direct inclusion of categorical variables and their interaction terms, thereby enhancing the model's interpretability and ease of setup. We selected demographic variables, including student age, sex, school type, total salary, parental education, and parental status, to explore their relationship with students' academic performance. Our aim is to understand how these demographic factors may influence or correlate with students' academic achievements (GPA).

The regression formula used in our model is as follows:

$$GPA = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...$$

$X_1$: 'Student Age' – (1: 18-21, 2: 22-25, 3: above 26)
$X_2$: 'Sex' – (1: female, 2: male)
$X_3$: 'School Type' – (1: private, 2: state, 3: other)
$X_4$: 'Total Salary' – (1: USD 135-200, 2: USD 201-270, 3: USD 271-340, 4: USD 341-410, 5: above 410)

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | GPA | R-squared: | 0.242 |
| Model: | OLS | Adj. R-squared: | 0.113 |
| Method: | Least Squares | F-statistic: | 1.875 |
| Date: | Tue, 14 May 2024 | Prob (F-statistic): | 0.0183 |
| Time: | 14:42:06 | Log-Likelihood: | −223.28 |
| No. Observations: | 145 | AIC: | 490.6 |
| Df Residuals: | 123 | BIC: | 556.0 |
| Df Model: | 21 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 3.0481 | 0.409 | 7.444 | 0.000 | 2.238 | 3.859 |
| C(student_age)[T.2] | 0.2425 | 0.237 | 1.021 | 0.309 | −0.228 | 0.712 |
| C(student_age)[T.3] | 1.2993 | 0.469 | 2.772 | 0.006 | 0.371 | 2.227 |
| C(sex)[T.2] | 0.6398 | 0.224 | 2.859 | 0.005 | 0.197 | 1.083 |
| C(school_type)[T.2] | −0.2016 | 0.334 | −0.604 | 0.547 | −0.863 | 0.459 |
| C(school_type)[T.3] | −0.3581 | 0.448 | −0.799 | 0.426 | −1.246 | 0.530 |
| C(total_salary)[T.2] | −0.6737 | 0.291 | −2.313 | 0.022 | −1.250 | −0.097 |
| C(total_salary)[T.3] | −0.5390 | 0.375 | −1.437 | 0.153 | −1.281 | 0.203 |
| C(total_salary)[T.4] | −0.3704 | 0.713 | −0.520 | 0.604 | −1.781 | 1.041 |
| C(total_salary)[T.5] | 0.1214 | 0.622 | 0.195 | 0.845 | −1.110 | 1.352 |
| C(mother_education)[T.2] | 0.0981 | 0.321 | 0.306 | 0.760 | −0.537 | 0.733 |
| C(mother_education)[T.3] | 0.0573 | 0.279 | 0.205 | 0.838 | −0.496 | 0.610 |
| C(mother_education)[T.4] | 0.3577 | 0.364 | 0.983 | 0.328 | −0.363 | 1.078 |
| C(mother_education)[T.5] | 0.3791 | 1.037 | 0.365 | 0.715 | −1.674 | 2.432 |
| C(mother_education)[T.6] | −1.7429 | 1.011 | −1.724 | 0.087 | −3.744 | 0.259 |
| C(father_education)[T.2] | −0.3114 | 0.347 | −0.898 | 0.371 | −0.998 | 0.375 |
| C(father_education)[T.3] | 0.1421 | 0.319 | 0.446 | 0.657 | −0.489 | 0.773 |
| C(father_education)[T.4] | −0.9537 | 0.380 | −2.508 | 0.013 | −1.707 | −0.201 |
| C(father_education)[T.5] | 0.2613 | 0.682 | 0.383 | 0.702 | −1.088 | 1.611 |
| C(father_education)[T.6] | −0.8669 | 1.361 | −0.637 | 0.525 | −3.560 | 1.827 |
| C(parental_status)[T.2] | −0.0119 | 0.409 | −0.029 | 0.977 | −0.821 | 0.797 |
| C(parental_status)[T.3] | 0.1921 | 0.507 | 0.379 | 0.705 | −0.811 | 1.196 |

| | | | |
|---|---|---|---|
| Omnibus: | 7.698 | Durbin-Watson: | 1.620 |
| Prob(Omnibus): | 0.021 | Jarque-Bera (JB): | 3.767 |
| Skew: | 0.131 | Prob(JB): | 0.152 |
| Kurtosis: | 2.255 | Cond. No. | 22.0 |

**Figure 1:** *OLS regression results from statsmodel.*

$X_5$: 'Mother's Education' – (1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)
$X_6$: 'Father's Education' – (1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)
$X_7$: 'Parental Status' – (1: married, 2: divorced, 3: died - one of them or both)

$X_2$**: Sex** Male (T.2): The coefficient of 0.6398 suggests an increase in GPA for male students compared to females, the reference group. The p-value of 0.005 confirms that this effect is statistically significant, indicating a gender disparity in GPA that favors males.

$X_3$**: School Type Analysis** State School (T.2): The coefficient of -0.2016 suggests a slight decrease in GPA for students from state schools compared to those from private schools. However, with a p value of 0.547, it is not statistically significant, indicating a minimal influence of school type on GPA. Other School Types (T.3): The coefficient of -0.3581 indicates a decrease in GPA for students from other types of schools compared to private schools, with a p-value of 0.426, also showing no significant impact.

$X_4$**: Total Salary** USD 201-270 (T.2): The coefficient of -0.6737 indicates a decrease in GPA for families earning within this salary range, with a p-value of 0.022, indicating that total salary may substantially influence GPA negatively. USD 271-340 (T.3): The coefficient of -0.5390 indicates a decrease in GPA for families earning within this salary range, with a p-value of 0.153, indicating that the impact may not be substantially significant. USD 341-410 (T.4): The coefficient of -0.3704 indicates a decrease in GPA for families earning within this salary range, with a p-value of 0.604, indicating that the impact may not be substantially significant. Above USD 410 (T.5): The coefficient of 0.1214 indicates an increase in GPA for families earning above this salary range, with a p-value of 0.845, indicating that the impact may not be substantially significant.

$X_5$ : **Mother's Education** Secondary School (T.2): The coefficient of 0.0981 indicates a slight increase in GPA for students with mothers educated up to secondary school, with a p-value of 0.760, indicating that the impact may not be substantially significant. High School (T.3): The coefficient of 0.0573 indicates a slight increase in GPA for students with mothers educated up to high school, with a p-value of 0.838, indicating that the impact may not be substantially significant. University (T.4): The coefficient of 0.3577 indicates an increase in GPA for students with mothers educated up to university level, with a p-value of 0.328, indicating that the impact may not be substantially significant. MSc. (T.5): The coefficient of 0.3791 indicates an increase in GPA for students with mothers holding a Master's degree, with a p-value of 0.715, indicating that the impact may not be substantially significant. Ph.D. (T.6): The coefficient of -1.7429 indicates a decrease in GPA for students with mothers holding a Ph.D., with a p-value of 0.087, indicating that the impact may not be substantially significant.

$X_6$ **Father's Education** Secondary School (T.2): The coefficient of -0.3114 indicates a decrease in GPA for students whose fathers attained secondary school education, with a p-value of 0.371, indicating that this variable may not substantially influence GPA. High School (T.3): The coefficient of 0.1421 indicates a slight increase in GPA for students whose fathers attained high school education, with a p-value of 0.657, suggesting that the influence of this educational level may not be substantial. University (T.4): The coefficient of -0.9537 indicates a decrease in GPA for students whose fathers have university education, with a p-value of 0.013, suggesting that this variable may substantially influence GPA negatively. MSc. (T.5): The coefficient of 0.2613 indicates an increase in GPA for students whose fathers have a Master's degree, with a p-value of 0.702, indicating that this variable may not substantially influence GPA. Ph.D. (T.6): The coefficient of -0.8669 indicates a decrease in GPA for students whose

fathers have a Ph.D., with a p-value of 0.525, suggesting that the influence of this educational level may not substantially influence GPA.

$X_7$ **Parental Status** Divorced (T.2): The coefficient of -0.0119 indicates a negligible decrease in GPA for students with divorced parents, with a p-value of 0.977, indicating that this variable may not substantially influence GPA. Died - one of them or both (T.3): The coefficient of 0.1921 indicates a slight increase in GPA for students who have experienced the death of one or both parents, with a p-value of 0.705, suggesting that this variable may not substantially influence GPA.

**Results** Among the 19 variables analyzed, only three yielded p-values lower than the conventional significance threshold of 0.05. These three variables show statistically significant impacts on GPA, indicating that they are likely to be genuine influences on academic outcomes within this dataset. However, the limited number of statistically significant results out of the large pool of variables (3 out of 19) suggests that many of the demographic factors explored may not have strong or consistent effects on GPA as measured in this study.

This scarcity of significant predictors can contribute to the relatively low R-squared and adjusted R-squared values observed in the model that we'll go over in the next section.

After running the linear regression model, we can take a look at the adjusted R-squared value (see **Figure 2**) to gauge how effectively the independent variables explain the variability in the dependent variable while penalizing the inclusion of unnecessary variables. Here, the adjusted R-squared value of 0.113 indicates that approximately 11.3% of the variance in GPA can be attributed to the predictors included in the model. This metric adjusts for the number of predictors, ensuring that the model's explanatory power is not inflated by including irrelevant variables. Consequently, an adjusted R-squared of 0.113 suggests that even after considering the number of predictors, the model still explains only a modest portion of the variability in GPA. The fact that the majority of the variability remains unexplained suggests that the parameters chosen in the model might not be sufficient to account for a significant portion of the data. Additionally, it implies that other factors, not included in the model, could be playing substantial roles in influencing the outcome. This highlights the complexity of the phenomenon being studied and the need for further investigation to uncover additional factors that contribute to the variability observed in the data.

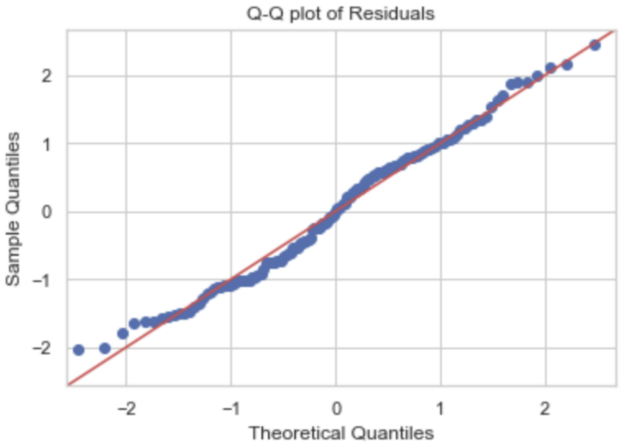



**Figure 2:** *Q-Q plot of adjusted R-squared values.*

With the adjusted R-squared indicating limited explanatory power of the model, it's important to delve into the analysis of residuals. We decided to use a Q-Q plot to plot the quantiles of our model's residuals against the theoretical quantiles of a normal distribution. In an ideal scenario where residuals are perfectly normal, the points should lie exactly on the red line that represents equality between the sample and theoretical quantiles. In our plot, the majority of the data points closely follow the red line, which is indicative of normal distribution of residuals. However, there are slight deviations at both ends of the plot. At the lower end, the points deviate below the line, suggesting that the residuals have slightly heavier tails than a normal distribution. At the upper end, the points deviate above the line, indicating light tails compared to a normal distribution.

Despite the approximate normality in residuals, our model's adjusted R-squared remains relatively low, suggesting limited explanatory power. To improve our model's performance, we will explore interaction terms that could unveil complex relationships between variables that were not previously modeled. By integrating these additional predictors, we aim to capture more nuances in the data, thereby increasing the explanatory power of our model and enhancing the precision of our predictions. In subsequent analyses, we will focus on including these interaction terms and examining their impact on the model's adjusted R-squared and overall fit.

In this improved model, we included several interaction terms. First being $C(\text{sex}) * C(\text{school\_type})$. The relationship between the two terms could stem from societal or cultural norms regarding education. For example, it's plausible that more girls attend private

schools than boys due to cultural expectations regarding the protection or nurturing of female students in environments perceived as safer or more conducive to academic success. Second being $C(\text{total\_salary}) * C(\text{father\_education})$. The relationship between the two terms could stem from socioeconomic status and educational attainment. For example, higher levels of education often translate to increased earning potential, as individuals with advanced degrees or specialized skills typically access better-paying job opportunities and career advancements. Similarly, the interaction term $C(\text{total\_salary}) * C(\text{mother\_education})$ mirrors this relationship with maternal education. Mothers with higher education levels tend to have better-paying jobs, contributing to higher household incomes. Fourth being $C(\text{mother\_education}) * C(\text{father\_education})$. This interaction term likely reflects shared values or educational backgrounds within couples. For example, partners with similar education levels may prefer to marry individuals with comparable or higher educational levels. Fifth being $C(\text{father\_education}) * C(\text{parental\_status})$. This relationship could stem from societal preferences or the perceived desirability of educated partners, suggesting that fathers with higher education levels may be more likely to be married. Similarly, the interaction term $C(\text{mother\_education}) * C(\text{parental\_status})$ follows a similar rationale, wherein mothers with higher education levels may also be more likely to be married, reflecting societal norms or preferences regarding marriage and educational compatibility.

From this improved model, we see that our R squared doubled from 0.242 to 0.509 and our adjusted R squared increased from 0.113 to 0.148. Despite the improvements in our model's explanatory power, it's essential to recognize that these values are still relatively low. This suggests that while demographic factors such as age, gender, ethnicity, socioeconomic status, and parental education do influence students' performance and final grades to some extent, there are likely additional influential factors not captured by our model.

The limited improvement in our model's performance despite the addition of interaction terms can be attributed to several factors. One reason for this limitation is the absence of numerous other potential predictors that may influence academic performance. While demographic factors are important, they represent only a subset of the myriad factors influencing students' grades.

Another reason is the relationships between the included variables and academic performance are likely complex and multifaceted, extending beyond simple linear associations. The model may fail to capture nonlinear relationships, interaction effects beyond those included, or potential moderating factors that could influence the strength or direction of relationships among variables.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | GPA | R-squared: | 0.145 |
| Model: | OLS | Adj. R-squared: | 0.061 |
| Method: | Least Squares | F-statistic: | 1.714 |
| Date: | Tue, 14 May 2024 | Prob (F-statistic): | 0.0649 |
| Time: | 17:41:36 | Log-Likelihood: | -232.02 |
| No. Observations: | 145 | AIC: | 492.0 |
| Df Residuals: | 131 | BIC: | 533.7 |
| Df Model: | 13 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 2.5605 | 0.674 | 3.801 | 0.000 | 1.228 | 3.893 |
| C(weekly_study_hours)[T.2] | -0.1076 | 0.296 | -0.364 | 0.717 | -0.693 | 0.478 |
| C(weekly_study_hours)[T.3] | 0.1278 | 0.366 | 0.349 | 0.728 | -0.597 | 0.852 |
| C(weekly_study_hours)[T.4] | -0.6307 | 0.602 | -1.048 | 0.297 | -1.822 | 0.560 |
| C(weekly_study_hours)[T.5] | -0.8736 | 0.699 | -1.250 | 0.214 | -2.257 | 0.509 |
| C(class_attendance)[T.2] | -0.4927 | 0.262 | -1.880 | 0.062 | -1.011 | 0.026 |
| C(midterm_prep_1)[T.2] | -0.1894 | 0.278 | -0.682 | 0.497 | -0.739 | 0.360 |
| C(midterm_prep_1)[T.3] | -1.1199 | 0.461 | -2.427 | 0.017 | -2.033 | -0.207 |
| C(midterm_prep_2)[T.2] | 0.5052 | 0.383 | 1.318 | 0.190 | -0.253 | 1.264 |
| C(midterm_prep_2)[T.3] | 3.0287 | 1.058 | 2.864 | 0.005 | 0.937 | 5.121 |
| C(notes_taking)[T.2] | 0.5926 | 0.652 | 0.909 | 0.365 | -0.697 | 1.883 |
| C(notes_taking)[T.3] | 0.8130 | 0.636 | 1.279 | 0.203 | -0.444 | 2.070 |
| C(class_listening)[T.2] | 0.0865 | 0.284 | 0.305 | 0.761 | -0.475 | 0.648 |
| C(class_listening)[T.3] | 0.1242 | 0.332 | 0.374 | 0.709 | -0.533 | 0.781 |

| | | | |
|---|---|---|---|
| Omnibus: | 13.541 | Durbin-Watson: | 1.619 |
| Prob(Omnibus): | 0.001 | Jarque-Bera (JB): | 5.371 |
| Skew: | -0.185 | Prob(JB): | 0.0682 |
| Kurtosis: | 2.132 | Cond. No. | 17.8 |

**Figure 3:** *OLS regression results from statsmodel.*

## Linear Regression, Model 2

From our initial analysis, it appears that demographic variables do not significantly impact student grades. Therefore, we decided to explore other potentially influential factors such as study habits and academic behaviors. To conduct this deeper investigation, we employed a linear regression model (see **Figure 3**) using a formulaic approach similar to our previous method but incorporating different variables.

The regression formula used in our model is as follows:

$$GPA = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...$$

$X_1$: 'Weekly study hours' – (1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours)
$X_2$: 'Attendance to classes' – (1: always, 2: sometimes, 3: never)
$X_3$: 'Preparation to midterm exams 1 ' – (1: alone, 2: with friends, 3: not applicable)
$X_4$: 'Preparation to midterm exams 2' – (1: closest date to the exam, 2: regularly during the semester, 3: never)
$X_5$: 'Taking notes in classes' – (1: never, 2: sometimes, 3: always)
$X_6$: Listening in classes' – (1: never, 2: sometimes, 3: always)

$X_1$**: Weekly Study Hours** <5 hours (T.2): The coefficient of -0.1076 suggests a slight decrease in GPA for students studying less than 5 hours compared to those who do not study at all. However, with a p-value of 0.717, this effect is not statistically significant, indicating that light study habits may not substantially influence GPA. 6-10 hours (T.3): The coefficient of 0.1278 shows a slight increase in GPA. However, with a p-value of 0.728, the effect is not statistically significant,suggesting that moderate study hours do not substantially influence GPA either. 11-20 hours (T.4): The coefficient of -0.6307 indicates a decrease in GPA. However, with a p-value of 0.297, it is not statistically significant, indicating that longer study hours do not substantially influence either. >20 hours (T.5): The coefficient of -0.8736 indicates a decrease in GPA. However, with a p-value of 0.214, it is not statistically significant, indicating that much longer study hours do not substantially influence either.

This observation is fairly odd as one would expect study hours to have a significant influence on students' final grades. This is because studying is typically seen as a direct contributor to learning and academic success. This surprising result might suggest issues with the data or underlying factors not captured by the study. While the data may not fully reflect the quality or effectiveness of study time, a basic relationship between hours studied and academic performance is generally anticipated. The lack of such a relationship could indicate diminishing returns, where additional study time does not necessarily correlate with improved grades, potentially due to fatigue or inefficiency. However, the complete absence of any correlation is peculiar and suggests that either key variables are missing or the data collection methods may need reassessment.

$X_2$**: Class Attendance** Sometimes (T.2): The coefficient of -0.4927 suggests a decrease in GPA for attending classes sometimes. However, with a p-value of 0.062, this effect is not statistically significant, indicating that class attendance may not substantially influence GPA.

$X_3$**: Preparation for Midterm Exams 1** With friends (T.2): The coefficient of -0.1894 suggests a decrease in GPA for preparing with friends. However, with a p-value of 0.497, this effect is not statistically significant, indicating that preparing with friends may not substantially influence GPA. Not applicable (T.3): The coefficient of -1.1199 suggests a decrease in GPA. With a p-value of 0.017, this effect is statistically significant, indicating that the other ways of preparing may substantially influence GPA.

$X_4$**: Preparation for Midterm Exams 2** Regularly during the semester (T.2): The coefficient of 0.5052 suggests an increase in GPA for regularly prepping for exams throughout the semester. However, with a p-value of 0.190, this effect is not statistically significant, indicating that preparing regularly may not substantially influence GPA. Never (T.3): The coefficient of 3.0287 suggests an increase in GPA for never prepping for exams throughout the semester. With a p-value of 0.005, this effect is statistically significant, indicating that never prepping may substantially influence GPA. This is incredibly surprising as one would expect that not preparing for exams would likely lead to poorer understanding of the material and, consequently, lower academic performance. This surprisingly positive impact on GPA for students who do not prepare at all may suggest potential anomalies in the data or unaccounted contextual factors within the study environment. For instance, it could be that these students are exceptionally adept at learning through methods not measured by this study, such as having a strong intrinsic understanding of the material or benefiting from alternative learning approaches like collaborative learning outside of traditional study settings.

$X_5$**: Taking Notes in Classes** Sometimes (T.2): The coefficient of 0.5926 suggests an increase in GPA for sometimes taking notes in classes. However, with a p-value of 0.365, this effect is not statistically significant, indicating that sometimes taking notes may not substantially influence GPA. Always (T.3): The coefficient of 0.8130 suggests an increase in GPA for always taking notes in classes. However, with a p-value of 0.203, this effect is not statistically significant, indicating that always taking notes may not substantially influence GPA.

$X_6$**: Listening in Classes** Sometimes (T.2): The coefficient of 0.0865 suggests an increase in GPA for always taking notes in classes. However, with a p-value of 0.761, this effect is not statistically significant, indicating that sometimes listening may not substantially influence GPA. Always (T.3): The coefficient of 0.1242 suggests an increase in GPA for always taking notes in classes. However, with a p-value of 0.709, this effect is not statistically significant, indicating that always listening may not substantially influence GPA.

**Results** Among the eleven variables analyzed, only two yielded p-values lower than the conventional significance threshold of 0.05. These two variables—related

to students reportedly never preparing for midterm exams and the second unexpected variable—stand out not only for their statistical significance but also for their counterintuitive nature. The variable indicating that students who never prepare for exams have a higher GPA is particularly puzzling and goes against typical educational expectations, where studying and exam preparation are directly correlated with better academic performance. This anomalous finding suggests there may be underlying factors or data inconsistencies not captured by the model, such as misreporting, alternative learning methods that are effective yet unaccounted for, or other academic supports in place. The second variable, similarly surprising in its significance, underscores potential gaps in our understanding or measurement of what influences academic success. Like the first model, the scarcity of significant predictors can contribute to the relatively low R-squared and adjusted R-squared values observed in the model which we'll discuss in the next section.

After running the linear regression model, we can take a look at the adjusted R-squared value to gauge how effectively the independent variables explain the variability in the dependent variable while penalizing the inclusion of unnecessary variables. Here, the adjusted R-squared value of 0.061 indicates that approximately 6.1% of the variance in GPA can be attributed to the predictors included in the model. This metric adjusts for the number of predictors, ensuring that the model's explanatory power is not inflated by including irrelevant variables. Consequently, an adjusted R-squared of 0.061 suggests that even after considering the number of predictors, the model still explains only a modest portion of the variability in GPA. The fact that the majority of the variability remains unexplained suggests that the parameters chosen in the model might not be sufficient to account for a significant portion of the data. Additionally, it implies that other factors, not included in the model, could be playing substantial roles in influencing the outcome. This highlights the complexity of the phenomenon being studied and the need for further investigation to uncover additional factors that contribute to the variability observed in the data.

With the adjusted R-squared indicating very low explanatory power of the model, it's important to delve into the analysis of residuals. We decided to use a Q-Q plot (see **Figure 4**) to plot the quantiles of our model's residuals against the theoretical quantiles of a normal distribution. In an ideal scenario where residuals are perfectly normal, the points should lie exactly on the red line that represents equality between the sample

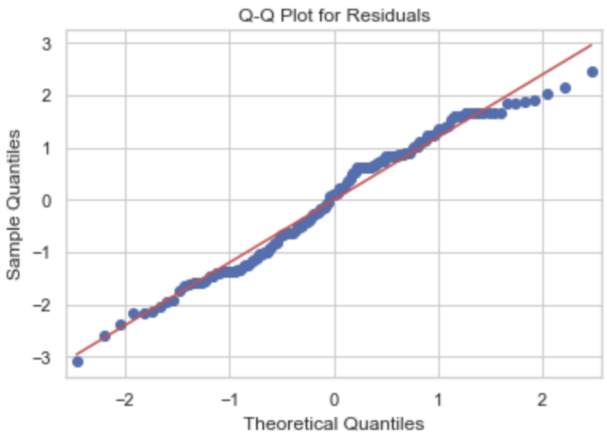



**Figure 4:** *Q-Q plot of adjusted R-squared values.*

and theoretical quantiles. In our plot, the majority of the data points closely follow the red line, which is indicative of normal distribution of residuals. However, there are slight deviations at both ends of the plot. At the lower end, the points deviate below the line, suggesting that the residuals have slightly heavier tails than a normal distribution. At the upper end, the points deviate above the line, indicating light tails compared to a normal distribution. These observations are similar to our prior model with the demographic variables but with a slightly larger deviation from the red line.

Despite the approximate normality in residuals, our model's adjusted R-squared remains incredibly low, suggesting limited explanatory power. To improve our model's performance, we will explore interaction terms that could unveil complex relationships between variables that were not previously modeled. By integrating these additional predictors, we aim to capture more nuances in the data, thereby increasing the explanatory power of our model and enhancing the precision of our predictions. In subsequent analyses, we will focus on including these interaction terms and examining their impact on the model's adjusted R-squared and overall fit.

In this improved model (see **Figure 5**), we included two interaction terms. First of them being $C(\text{midterm\_prep\_1}) * C(\text{midterm\_prep\_2})$. This interaction term compares the relationship between preparation timing with the social context of studying. For example, students who prepare closer to the exam date might be more likely to study in groups, due to the immediacy and shared focus of impending exams. Conversely, those who start preparing earlier may be more inclined to study alone, as most

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | GPA | R-squared: | 0.182 |
| Model: | OLS | Adj. R-squared: | 0.080 |
| Method: | Least Squares | F-statistic: | 1.779 |
| Date: | Tue, 14 May 2024 | Prob (F-statistic): | 0.0407 |
| Time: | 17:41:59 | Log-Likelihood: | −228.85 |
| No. Observations: | 145 | AIC: | 491.7 |
| Df Residuals: | 128 | BIC: | 542.3 |
| Df Model: | 16 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 2.4955 | 0.669 | 3.727 | 0.000 | 1.171 | 3.820 |
| C(weekly_study_hours)[T.2] | −0.1840 | 0.297 | −0.619 | 0.537 | −0.772 | 0.404 |
| C(weekly_study_hours)[T.3] | −0.0156 | 0.376 | −0.042 | 0.967 | −0.761 | 0.729 |
| C(weekly_study_hours)[T.4] | −0.5302 | 0.601 | −0.882 | 0.379 | −1.720 | 0.659 |
| C(weekly_study_hours)[T.5] | −1.0727 | 0.701 | −1.531 | 0.128 | −2.459 | 0.314 |
| C(class_attendance)[T.2] | −0.2934 | 0.174 | −1.689 | 0.094 | −0.637 | 0.050 |
| C(midterm_prep_1)[T.2] | −0.0275 | 0.297 | −0.092 | 0.927 | −0.615 | 0.560 |
| C(midterm_prep_1)[T.3] | −1.4292 | 0.555 | −2.576 | 0.011 | −2.527 | −0.332 |
| C(midterm_prep_2)[T.2] | 0.6837 | 0.437 | 1.564 | 0.120 | −0.181 | 1.549 |
| C(midterm_prep_2)[T.3] | 1.6557 | 0.541 | 3.058 | 0.003 | 0.584 | 2.727 |
| C(notes_taking)[T.2] | 0.7069 | 0.652 | 1.084 | 0.280 | −0.583 | 1.997 |
| C(notes_taking)[T.3] | 0.7709 | 0.632 | 1.219 | 0.225 | −0.480 | 2.022 |
| C(class_listening)[T.2] | 0.2140 | 0.287 | 0.746 | 0.457 | −0.354 | 0.782 |
| C(class_listening)[T.3] | 0.2369 | 0.333 | 0.711 | 0.478 | −0.422 | 0.896 |
| C(midterm_prep_1)[T.2]:C(midterm_prep_2)[T.2] | −1.1194 | 0.792 | −1.413 | 0.160 | −2.687 | 0.448 |
| C(midterm_prep_1)[T.3]:C(midterm_prep_2)[T.2] | 0.8556 | 1.006 | 0.850 | 0.397 | −1.135 | 2.847 |
| C(midterm_prep_1)[T.2]:C(midterm_prep_2)[T.3] | 3.611e−17 | 2.21e−17 | 1.636 | 0.104 | −7.57e−18 | 7.98e−17 |
| C(midterm_prep_1)[T.3]:C(midterm_prep_2)[T.3] | 1.6557 | 0.541 | 3.058 | 0.003 | 0.584 | 2.727 |
| C(class_attendance)[T.2]:C(notes_taking)[T.2] | −0.5400 | 0.270 | −1.999 | 0.048 | −1.074 | −0.006 |
| C(class_attendance)[T.2]:C(notes_taking)[T.3] | 0.2466 | 0.283 | 0.871 | 0.385 | −0.313 | 0.807 |

| | | | |
|---|---|---|---|
| Omnibus: | 6.904 | Durbin-Watson: | 1.717 |
| Prob(Omnibus): | 0.032 | Jarque-Bera (JB): | 3.908 |
| Skew: | −0.197 | Prob(JB): | 0.142 |
| Kurtosis: | 2.299 | Cond. No. | 2.02e+18 |

**Figure 5:** *OLS regression results from statsmodel for the improved model.*

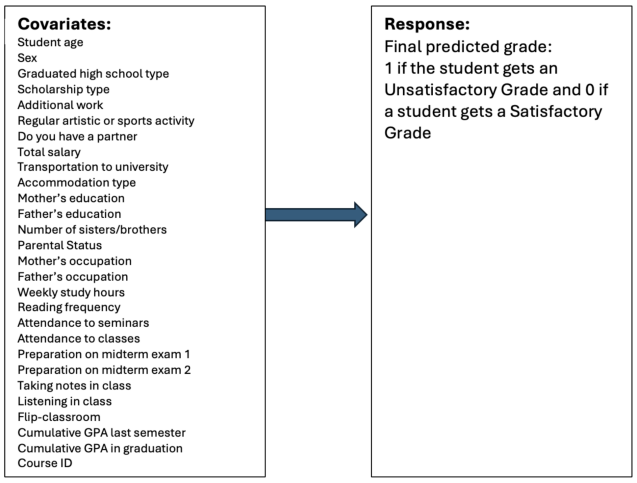



**Figure 6:** *Summary of the covariates and response variables used in our logistic regression analysis.*

students haven't started studying yet. Second being $C(\text{class\_attendance}) * C(\text{notes\_taking})$. This compares the relationship between attending classes with taking notes. It is intuitive that students who are present in class are more likely to engage actively by taking notes.

From this improved model, we see that our R squared increased from 0.145 to 0.182 and our adjusted R squared increased from 0.061 to 0.08. Despite the improvements in our model's explanatory power, it's essential to recognize that these values are still incredibly low. This suggests that while academic habits factors such as attendance, study hours, midterm preparation, note taking, and listening in class do influence students' performance and final grades to some extent, there are likely additional influential factors not captured by our model.

The limited improvement in our model's performance despite the addition of interaction terms can be attributed to several factors. One reason for this limitation is the absence of numerous other potential predictors that may influence academic performance. While the selected academic factors are influential to an extent, they represent only a subset of the myriad factors influencing students' grades.

Despite the inclusion of interaction terms and various predictors related to academic habits, the relatively low R-squared and adjusted R-squared values indicate that a significant portion of the variance in GPA remains unexplained. This suggests that there are other influential factors not captured by the model, which could include unmeasured variables such as student motivation, mental health, quality of teaching, extracurricular activities, or even socio-economic factors that are not given to us in this data

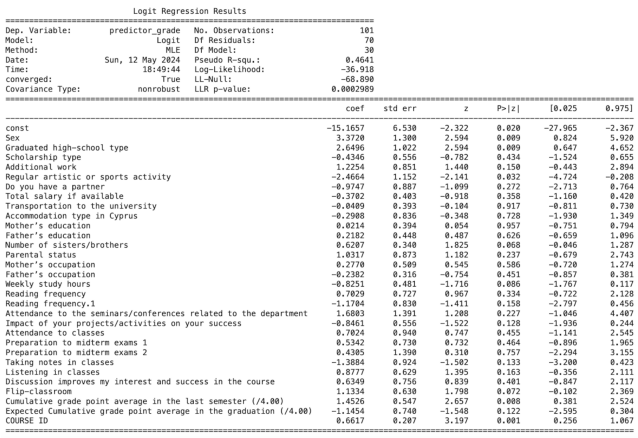
## Logistic Regression

We conducted a Logistic regression analysis to assess the potential of our dataset in predicting whether a student will pass or fail by the end of the semester. In this context, we categorized the outcomes as either Unsatisfactory or Satisfactory. Many universities, including Cornell's College of Engineering, require students to achieve at least a C grade to complete major-approved courses and earn credit. Our dataset includes the predictor OUTPUT Grade, with codes (0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA). For this analysis, grades (0: Fail, 1: DD, 2: DC, 3: CC) are considered Unsatisfactory, while grades (4: CB, 5: BB, 6: BA, 7: AA) are considered Satisfactory.

Given these criteria, we are interested in exploring whether a student would receive a satisfactory or unsatisfactory grade based on various variables in a college environment. **Figure 6** is a is a summary of the covariates and response variables used in our analysis.

**Model Summary** There are many student performance statistics covariates. To train a more accurate model we wanted to choose the covariates that are important in making our prediction. We first split up the data into training and testing. We held out 30% of our data for the test set. We did this using Sklearn: train_test_split. We fit the training data to the logistic regression model using statsmodel's Logit function. We initially trained a logistic regression model that takes in all the covariates stated above.

We then choose the columns/covariates that have a significant p-value (less than 0.05). The significant coefficients were Sex, Graduated high-school type, regular



Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | predictor_grade | No. Observations: | 101 |
| Model: | Logit | Df Residuals: | 70 |
| Method: | MLE | Df Model: | 30 |
| Date: | Sun, 12 May 2024 | Pseudo R-squ.: | 0.4641 |
| Time: | 18:49:44 | Log-Likelihood: | -36.918 |
| converged: | True | LL-Null: | -68.890 |
| Covariance Type: | nonrobust | LLR p-value: | 0.0002989 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -15.1657 | 6.530 | -2.322 | 0.020 | -27.965 | -2.367 |
| Sex | 3.3720 | 1.300 | 2.594 | 0.009 | 0.824 | 5.920 |
| Graduated high-school type | 2.6496 | 1.022 | 2.594 | 0.009 | 0.647 | 4.652 |
| Scholarship type | -0.4346 | 0.556 | -0.782 | 0.434 | -1.524 | 0.655 |
| Additional work | 1.2254 | 0.851 | 1.440 | 0.150 | -0.443 | 2.894 |
| Regular artistic or sports activity | -2.4664 | 1.152 | -2.141 | 0.032 | -4.724 | -0.208 |
| Do you have a partner | -0.9747 | 0.887 | -1.099 | 0.272 | -2.713 | 0.764 |
| Total salary if available | -0.3702 | 0.403 | -0.918 | 0.358 | -1.160 | 0.420 |
| Transportation to the university | -0.0409 | 0.393 | -0.104 | 0.917 | -0.811 | 0.730 |
| Accommodation type in Cyprus | -0.2908 | 0.836 | -0.348 | 0.728 | -1.930 | 1.349 |
| Mother's education | 0.0214 | 0.394 | 0.054 | 0.957 | -0.751 | 0.794 |
| Father's education | 0.2182 | 0.448 | 0.487 | 0.626 | -0.659 | 1.096 |
| Number of sisters/brothers | 0.6207 | 0.340 | 1.825 | 0.068 | -0.046 | 1.287 |
| Parental status | 1.0317 | 0.873 | 1.182 | 0.237 | -0.679 | 2.743 |
| Mother's occupation | 0.2770 | 0.509 | 0.545 | 0.586 | -0.720 | 1.274 |
| Father's occupation | -0.2382 | 0.316 | -0.754 | 0.451 | -0.857 | 0.381 |
| Weekly study hours | -0.8251 | 0.481 | -1.716 | 0.086 | -1.767 | 0.117 |
| Reading frequency | 0.7029 | 0.727 | 0.967 | 0.334 | -0.722 | 2.128 |
| Reading frequency.1 | -1.1704 | 0.830 | -1.411 | 0.158 | -2.797 | 0.456 |
| Attendance to the seminars/conferences related to the department | 1.6003 | 1.391 | 1.208 | 0.227 | -1.046 | 4.407 |
| Impact of your projects/activities on your success | -0.8461 | 0.556 | -1.522 | 0.128 | -1.936 | 0.244 |
| Attendance to classes | 0.7024 | 0.940 | 0.747 | 0.455 | -1.141 | 2.545 |
| Preparation to midterm exams 1 | 0.5342 | 0.730 | 0.732 | 0.464 | -0.896 | 1.965 |
| Preparation to midterm exams 2 | 0.4305 | 1.390 | 0.310 | 0.757 | -2.294 | 3.155 |
| Taking notes in classes | -1.3884 | 0.924 | -1.502 | 0.133 | -3.200 | 0.423 |
| Listening in classes | 0.8777 | 0.629 | 1.395 | 0.163 | -0.356 | 2.111 |
| Discussion improves my interest and success in the course | 0.6349 | 0.756 | 0.839 | 0.401 | -0.847 | 2.117 |
| Flip-classroom | 1.1334 | 0.630 | 1.798 | 0.072 | -0.102 | 2.369 |
| Cumulative grade point average in the last semester (/4.00) | 1.4526 | 0.547 | 2.657 | 0.008 | 0.381 | 2.524 |
| Expected Cumulative grade point average in the graduation (/4.00) | -1.1454 | 0.740 | -1.548 | 0.122 | -2.595 | 0.304 |
| COURSE ID | 0.6617 | 0.207 | 3.197 | 0.001 | 0.256 | 1.067 |

**Figure 7:** *OLS regression results from statsmodel.*

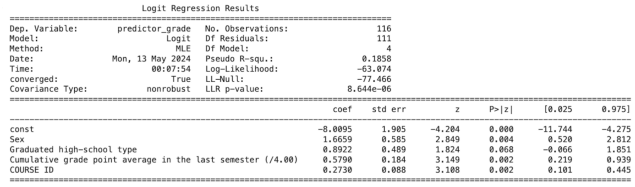
artistic or sports activity, Cumulative GPA in the last semester, and Course ID.

To further verify these features we fit the testing set on a new logistic model with the X-values as the columns with significant p-values (stated above). Sex, Graduated high-school type, Cumulative GPA in the last semester, and Course ID stayed significant. The p-value of regular artistic or sports activity proved to be statistically insignificant (0.07> 0.05). In our visualizations from milestone 1, we made a heat map that shows several different factors and their correlation coefficients with grades received. The coefficient values for Sex, Graduated high-school type, Cumulative GPA in the last semester, and Course ID were 0.34, 0.10, 0.32, and 0.14 respectively showing a correlation.

Through this model selection, we can see that our logistic regression answers our question: **Do demographic factors such as age, gender, ethnicity, socioeconomic status, and parental education) affect a student's performance and their final grade?** As we have shown demographic factors such as Sex and high school type affect whether the student will receive a Satisfactory or Unsatisfactory grade.

**Improved Model** Now that we have been able to determine the most relevant features to use in our logistic regression, we will fit a new logistic regression that takes in Sex, Graduated high-school type, regular artistic or sports activity, Cumulative GPA in the last semester, and Course ID. We trained a new train test split with 20% test data and 80% training data. **Figure 8** is the summary of the logistic regression.

Interpreting the coefficients: A one-unit change from female (coded as 1) to male (coded as 2) is associated with a 1.6659 unit increase in the log odds of the predicted outcome being satisfactory. Graduated high



Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | predictor_grade | No. Observations: | 116 |
| Model: | Logit | Df Residuals: | 111 |
| Method: | MLE | Df Model: | 4 |
| Date: | Mon, 13 May 2024 | Pseudo R-squ.: | 0.1858 |
| Time: | 00:07:54 | Log-Likelihood: | -63.074 |
| converged: | True | LL-Null: | -77.466 |
| Covariance Type: | nonrobust | LLR p-value: | 8.644e-06 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -8.0095 | 1.905 | -4.204 | 0.000 | -11.744 | -4.275 |
| Sex | 1.6659 | 0.585 | 2.849 | 0.004 | 0.520 | 2.812 |
| Graduated high-school type | 0.8922 | 0.489 | 1.824 | 0.068 | -0.066 | 1.851 |
| Cumulative grade point average in the last semester (/4.00) | 0.5790 | 0.184 | 3.149 | 0.002 | 0.219 | 0.939 |
| COURSE ID | 0.2730 | 0.088 | 3.108 | 0.002 | 0.101 | 0.445 |

**Figure 8:** *OLS regression results from statsmodel for the improved model.*

school type has a coefficient of 0.8992, indicating a one-unit change associated with a change from state school to private school increases the log odds of the predicted outcome being satisfactory by 0.8992. Cumulative GPA has a coefficient of 0.5703. Although it does not have a high coefficient it still indicates that a one-unit increase in GPA is associated with a 0.5703 unit increase in the log odds of the predicted outcome being satisfactory The model has an accuracy of 65.5% and a precision of 64.9%. The low R-squared value of 0.186 suggests 18.6% variability in the outcome by an independent variable in the model. The R-squared values are not very high. We have shown that there is a strong variable significance (p-value significance) in our model and we have good accuracy and precision of 65.5% and 64.9% respectively. Although these are not very strong results we think they are good enough to make a pretty good prediction.

**ROC Curve** Another way to evaluate the classification of our logistic regression model is through the ROC curve (see **Figure 9**). ROC curves give us a comprehensive view of the tradeoff between true positive and false positive rates. The green line shows the ROC curve and the dotted black line gives the ROC from a completely random model. Initially, the green line has a steep rise towards the top-left, indicating the model can achieve a high true positive rate while keeping the false positive rate low. At one point, the green line touches the dashed diagonal line, indicating that for those threshold values, the classifier performs no better than random guessing. After touching the diagonal, the green line rises steeply again, indicating the classifier becomes a good model with high true positive rates and low false positive rates in that operating region.

## Random Forest Classification

To further explore possibilities in whether or not we could predict grades based on the demographic and academic characteristics that we have shown to have an effect, we performed a Random Forest classification analysis. For reference, a Random Forest is an ensem-

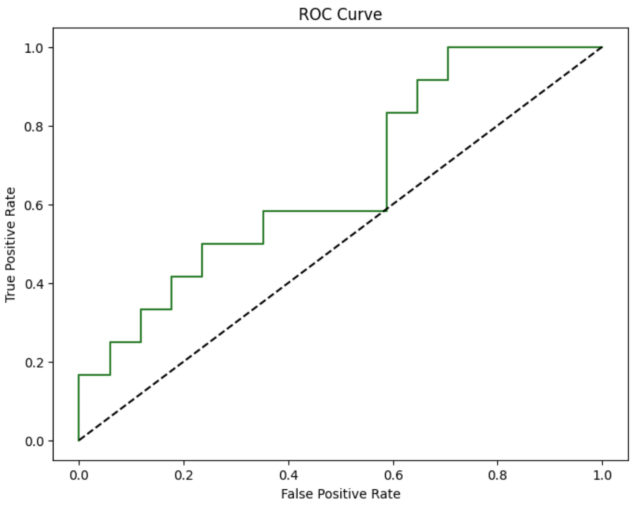



**Figure 9:** *ROC curve of positive rates.*

ble learning method commonly used for classification and regression tasks. It builds upon the concept of decision trees by creating a 'forest' of trees, where each tree is slightly different from the others. The idea is rooted in the principle that a group of weak learners (in this case, decision trees), when combined, can form a strong learner. Each tree in the forest is trained on a random subset of the data with replacement, and at each node, a random subset of features is considered for splitting. This randomness helps in making the model more robust than a single decision tree, significantly reducing the risk of overfitting. The final prediction of the Random Forest is made by averaging the predictions of all the trees for regression tasks or by a majority vote for classification tasks.

The Random Forest model was tested on 15% of the data and trained on the other 85%. It produced an accuracy of approximately 31.82%. This accuracy level, while better than random guessing in a 7-class prediction task (14.3%), suggests significant room for improvement.

**Classification Report Overview** The detailed classification report (see **Figure 10**) provides a breakdown of performance by class:

Class 0: Perfect precision and recall, although it only represents one instance, suggesting limited inference can be drawn.

Class 1-7: Shows varying degrees of precision and recall. Notably, classes with more support (instances) like classes 1 and 3 have recall rates of 40% and 67% respectively, but low precision. This implies that while the model is relatively reliable in identifying these classes, it also makes a significant number of false

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 1 |
| 1 | 0.29 | 0.40 | 0.33 | 5 |
| 2 | 0.50 | 0.25 | 0.33 | 4 |
| 3 | 0.25 | 0.67 | 0.36 | 3 |
| 4 | 0.00 | 0.00 | 0.00 | 1 |
| 5 | 1.00 | 0.33 | 0.50 | 3 |
| 6 | 0.00 | 0.00 | 0.00 | 2 |
| 7 | 0.00 | 0.00 | 0.00 | 3 |
| accuracy | | | 0.32 | 22 |
| macro avg | 0.38 | 0.33 | 0.32 | 22 |
| weighted avg | 0.37 | 0.32 | 0.30 | 22 |

**Figure 10:** *Random forest classification report overview.*

positives.

Classes 4, 6, 7: These classes have zero precision and recall, indicating the model failed to predict these categories correctly at all. This could be due to several reasons, such as insufficient training examples for these classes or features that do not distinguish these categories effectively from others.

Precision: Measures the accuracy of the model when it predicts a class. A low precision across most classes suggests that there are many false positives. Recall: Measures the model's ability to detect all relevant instances of a class. The varied recall rates suggest some classes are better captured than others.

**Strategies for improvement: F1-Score** The harmonic mean of precision and recall. Low F1-scores across the board confirm the challenges the model faces in balancing precision and recall, which is crucial for a good classification system.

**Strategies for improvement: Feature Engineering** There may be a need to revisit the feature selection process to ensure that more predictive features are being utilized. Including interaction terms or polynomial features might capture more complex patterns in the data. Resampling: Given the poor performance on certain classes, particularly those with zero recall and precision, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or random oversampling could be used to address class imbalance.

**Improved Random Forest** To try and improve the accuracy and precision of our Random Forest model, we tried applying SMOTE as well as incorporating new polynomial features of (feature)2 for the demographic and academic predictors that had p values less than .05 in the model selection that was performed in the logistic regression, which were Sex, Graduated high-school

Accuracy: 0.2727272727272727
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 1 |
| 1 | 0.25 | 0.40 | 0.31 | 5 |
| 2 | 0.25 | 0.25 | 0.25 | 4 |
| 3 | 0.40 | 0.67 | 0.50 | 3 |
| 4 | 0.00 | 0.00 | 0.00 | 1 |
| 5 | 0.00 | 0.00 | 0.00 | 3 |
| 6 | 0.00 | 0.00 | 0.00 | 2 |
| 7 | 0.00 | 0.00 | 0.00 | 3 |
| accuracy | | | 0.27 | 22 |
| macro avg | 0.24 | 0.29 | 0.26 | 22 |
| weighted avg | 0.20 | 0.27 | 0.23 | 22 |

**Figure 11:** *Random forest classification report overview for the improved model.*

type, regular artistic or sports activity, Cumulative GPA in the last semester, and Course ID. See **Figure 11**.

The results indicate the model performed worse overall with SMOTE and additional polynomial features. Only class three saw a very small improvement in precision although it is likely insignificant.

Despite these potential improvements to the models, some of the reasons why they may have lead to worse results are potentially that the data set itself is too small, with not enough points to sample on for both testing and training given the complex feature set that is generated via SMOTE and the additional polynomial features. Furthermore, while SMOTE helps to balance the class distribution, it could also lead to overfitting. This occurs because SMOTE can create examples that overly generalize the minority class characteristics, causing the model to learn these artificial nuances too well, which do not generalize to the unseen test data. Similarly, adding additional features could also contribute to overfitting, especially if the added features don't contribute any additional predictive power or have high collinearity.

**Conclusion** The current implementation of the Random Forest model demonstrates that while demographic and academic features can provide some insights into predicting student performance, they alone may not be sufficient to predict how well students will perform accurately. The relatively low accuracy and poor precision and recall for several classes underscore the need for potential further model optimization and exploration of additional features or the potential inability for a Random Forest model to predict the grades solely based on these factors. This study emphasizes the complexity of predicting academic performance and the need for a multifaceted approach that includes more comprehensive data collection, innovative feature engineering, and possibly the integration of more sophisticated modeling techniques.

## References

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

Students performance. (2023, October 17). Kaggle. https://www.kaggle.com/datasets/joebeachcapital/students-performance/data