High School Student Academic Performance

Group 4: Debanshi Misra, Maeve Horan-Portelance,
Anika Chowdhury, Angelo Desiderio, Ashley Chan
Stats 112: A Window to Understanding Diversity
Discussion 1B

Table of Contents

I. Abstract

In this report, we present our analysis of the Students Performance Dataset, sourced from Kaggle, which provides an overview of how demographic, familial, and behavioral factors influence high school student academic performance. The dataset includes features such as age, gender, ethnicity, parental education and support, weekly study time, absences, tutoring, and extracurricular participation. Our objective was to train and evaluate a multinomial logistic regression model to classify students into three GPA categories (low, medium, and high) based on these predictors.

We trained a multinomial logistic regression model to classify students into GPA categories (low, medium, high) after removing highly correlated predictors and non-significant variables based on p-values and Variance Inflation Factor (VIF) analysis. The results indicated that study time, tutoring, and participation in extracurricular activities were positively related, absences were negatively related, and student demographics were not related to medium/high GPA compared to low GPA. The model achieved an average accuracy of 87.21% based on 10-fold cross-validation, demonstrating strong predictive performance. To enhance reliability, we validated the model using a train-test split and Leave-One-Out Cross-Validation (LOOCV), with similar accuracy results. Below, we detail our data preprocessing steps, experimental design, and analysis.

II. Introduction

Research Question

How are age, gender, ethnicity, parental background/involvement, study habits, and extracurricular involvement associated with student academic performance?

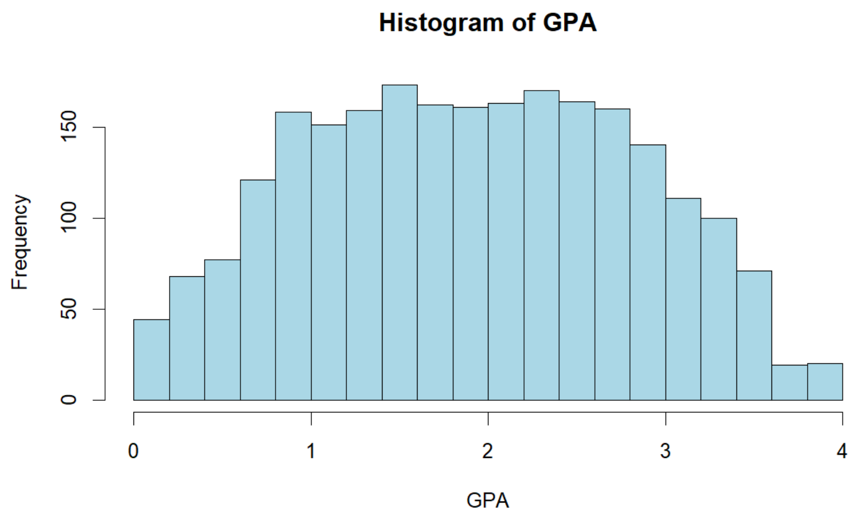
Dataset and Variables

This dataset profiles high school students, presenting their grades/GPA and detailing factors like demographics, study habits, and extracurriculars for educational analysis. Note that the data most likely comes from a district characterized by low academic performance, potentially influenced by socioeconomic conditions that hinder schools from providing adequate resources, fostering effective teaching practices, or supporting comprehensive student engagement programs.

Variable Name	Variable Type	Description
GPA	Outcome, Categorical	RESPONSE VARIABLE, Categories: - Low GPA (< 2.5) - Medium GPA (2.5 - 3.0) - High GPA (> 3.0)
Age	Predictor, Numerical	The age of the students ranges from 15 to 18 years.
Gender	Predictor, Categorical	Gender of the students, where 0 represents Male and 1 represents Female.
Ethnicity	Predictor, Categorical	The ethnicity of the students, coded as follows: - 0: Caucasian - 1: African American - 2: Asian - 3: Other
ParentalEducation	Predictor, Categorical	The education level of the parents, coded as follows: - 0: None - 1: High School

		- 2: Some College - 3: Bachelor's - 4: Higher
StudyTimeWeekly	Predictor, Numerical	Weekly study time in hours, ranging from 0 to 20.
Absences	Predictor, Numerical	Number of absences during the school year, ranging from 0 to 30.
Tutoring	Predictor, Categorical	Tutoring status, where 0 indicates No and 1 indicates Yes.
ParentalSupport	Predictor, Categorical	The level of parental support, coded as follows: - 0: None - 1: Low - 2: Moderate - 3: High - 4: Very High
Extracurricular	Predictor, Categorical	Participation in extracurricular activities, where 0 indicates No and 1 indicates Yes.
Sports	Predictor, Categorical	Participation in sports, where 0 indicates No and 1 indicates Yes.
Music	Predictor, Categorical	Participation in music activities, where 0 indicates No and 1 indicates Yes.
Volunteering	Predictor, Categorical	Participation in volunteering, where 0 indicates No and 1 indicates Yes.
GradeClass	Outcome, Categorical	Classification of students' grades based on GPA: - 0: 'A' (GPA >= 3.5) - 1: 'B' (3.0 <= GPA < 3.5) - 2: 'C' (2.5 <= GPA < 3.0) - 3: 'D' (2.0 <= GPA < 2.5) - 4: 'F' (GPA < 2.0)

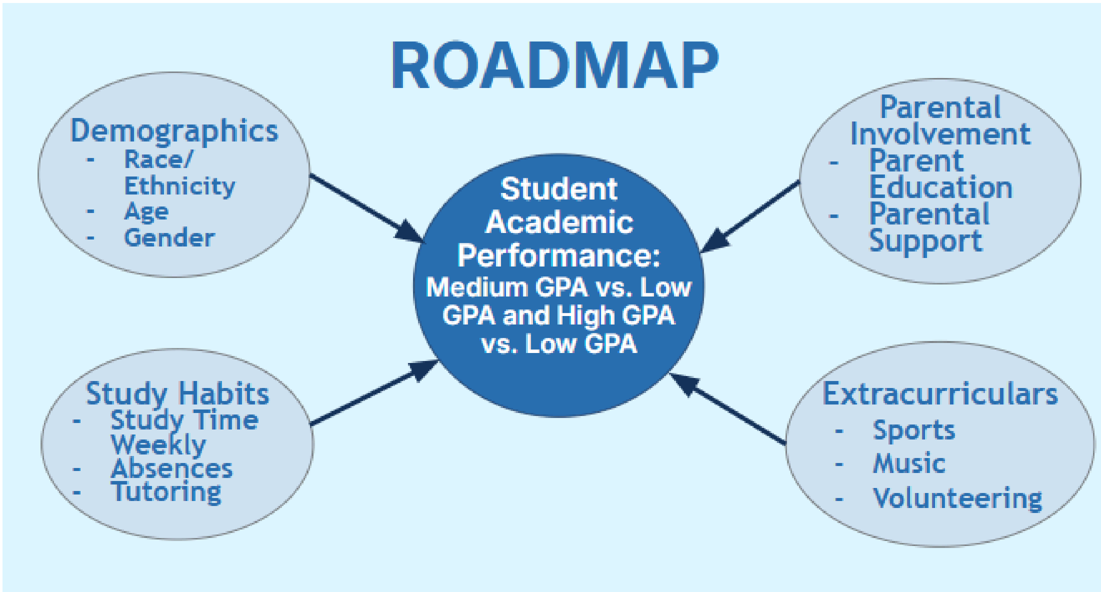
Outcome Variable - GPA



Measure of Academic Performance, influenced by study habits, parental involvement, extracurricular activities, and more. We split it into three categories: Low GPA (< 2.5), Medium GPA (2.5 - 3.0), and High GPA (> 3.0). This split maintains an approximately even spread of data throughout the different categories.

We chose these categorizations to remain true to real-world definitions of low, medium, and high GPAs. Simply using high vs low GPA categorizations is too vague, and three categorizations would provide more real-world context.

Roadmap

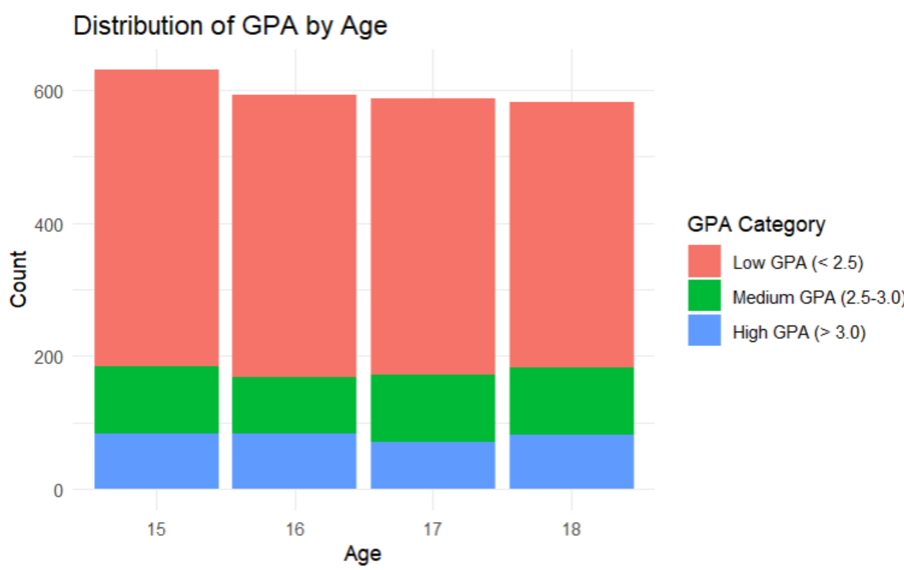


Our roadmap focuses on understanding how the factors of demographics, parental involvement, study habits, and extracurricular participation are associated with student academic performance. This is measured in three GPA categories, low, medium, and high. Demographic factors such as age, gender, and ethnicity are considered for our research as they can influence academic outcomes due to developmental and societal contexts. Parental involvement, including parent education and support, is considered for the influences of family background on a student's academic success. Study habits such as study time, attendance, and tutoring services reflect behaviors and routines connected to academic performance. Participation in extracurricular activities, such as volunteering, athletics, and music, is also taken into account since extracurriculars can improve time management and discipline, which may have an indirect impact on students' academics. This roadmap allows us to see how these factors influence academic performance to address our research question.

III. Methods

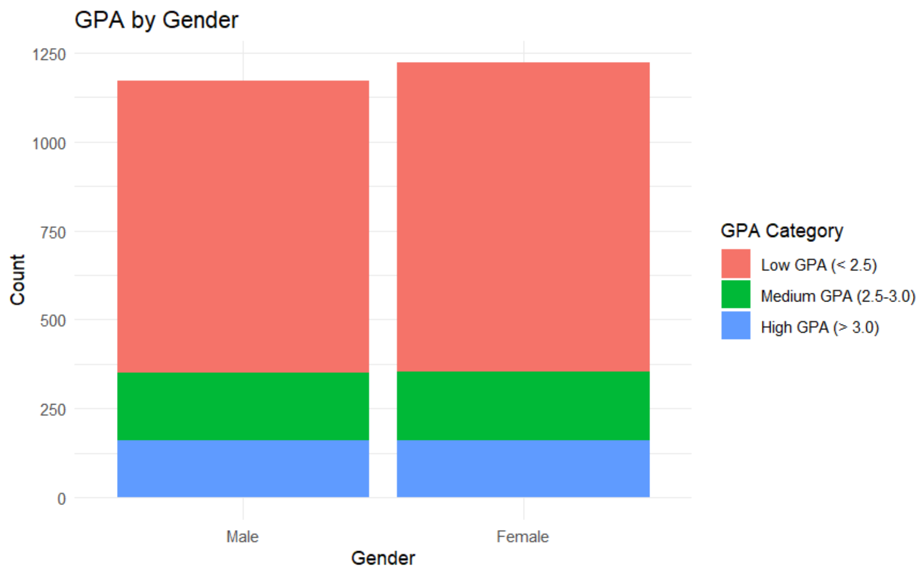
Exploratory Data Analysis for Predictors

Age



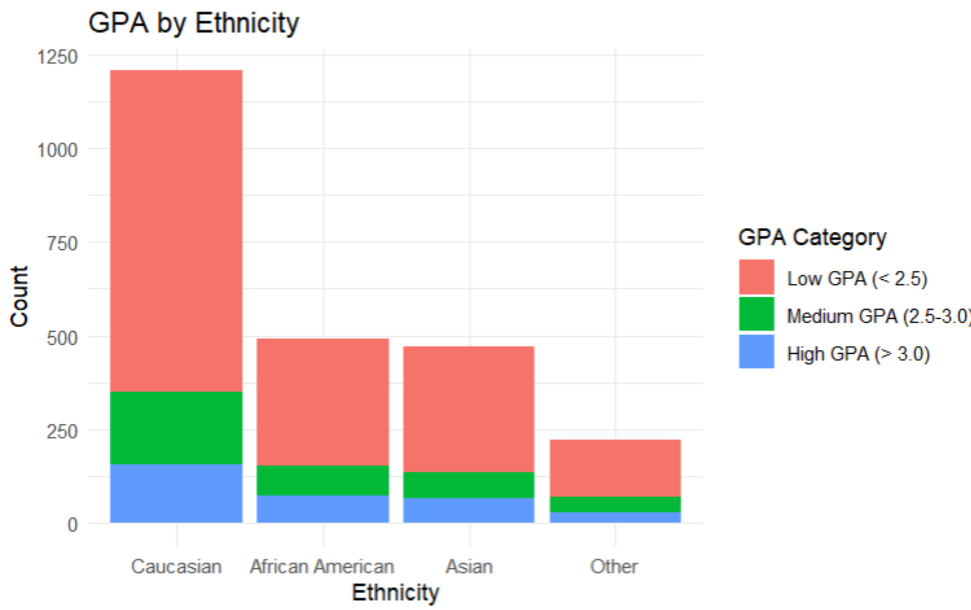
The majority of the students have low GPAs, below 2.5, for each age group. A smaller proportion of students per age have medium GPAs and the smallest proportion have high GPAs, above 3.0. With such little variation in the GPA distribution for the different age groups, it suggests that age may not be a factor that influences student performance, there may be external factors or consistent challenges across all age groups.

Gender



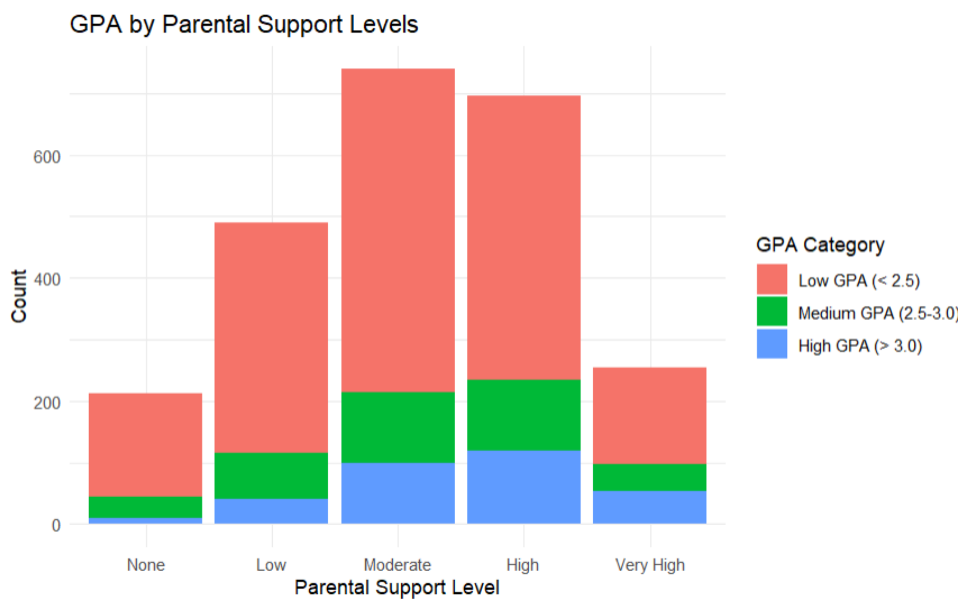
This histogram of the distribution of GPA categories across Male and Female students shows that the majority of students are in the Low GPA category. The Medium GPA and High GPA categories have smaller proportions for both male and female students. However, the Medium GPA category has a slightly higher amount than the High GPA. There is not much variation of different GPAs between male and female students, this indicates that gender alone may not be a strong predictor for a student's academic performance.

Ethnicity



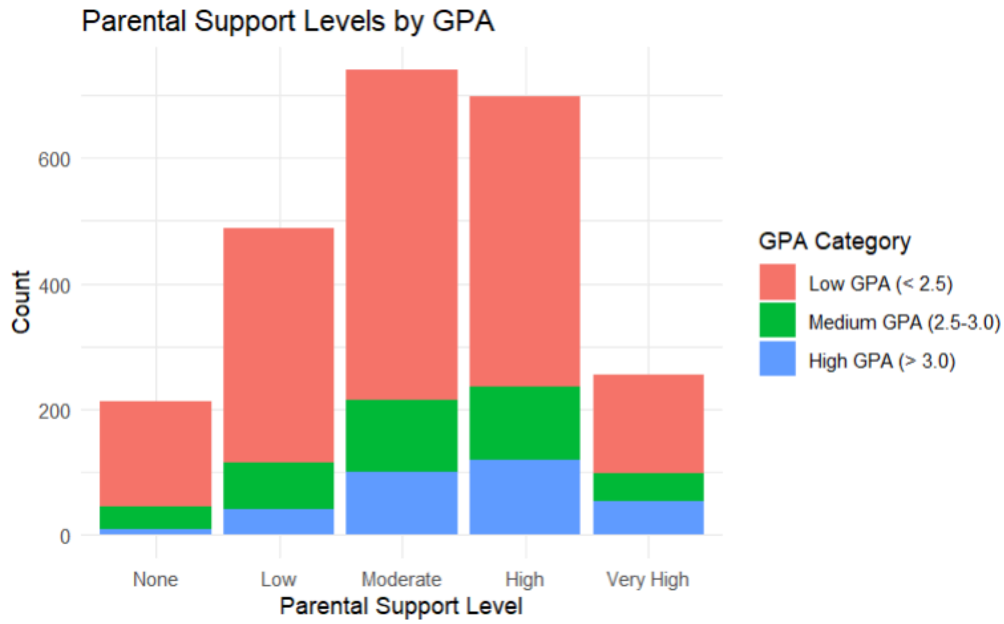
The distribution in the histogram of GPA levels across different ethnicities shows that the Low GPA category is the largest for all ethnic groups. There is a larger proportion of Caucasian students with low GPAs which is due to the sample continuing a larger amount of students in this ethnic group. Caucasian students, African American students, and students that fall in the Other category have smaller proportions in the Medium and High GPA levels, with a slightly larger proportion of students in all groups falling in the Medium GPA level. Asian students have a balanced proportion of students in the Medium and High GPA levels. Considering the distribution, the ethnicity of students does not seem to show significant variation in the GPA levels for this dataset.

Parental Education



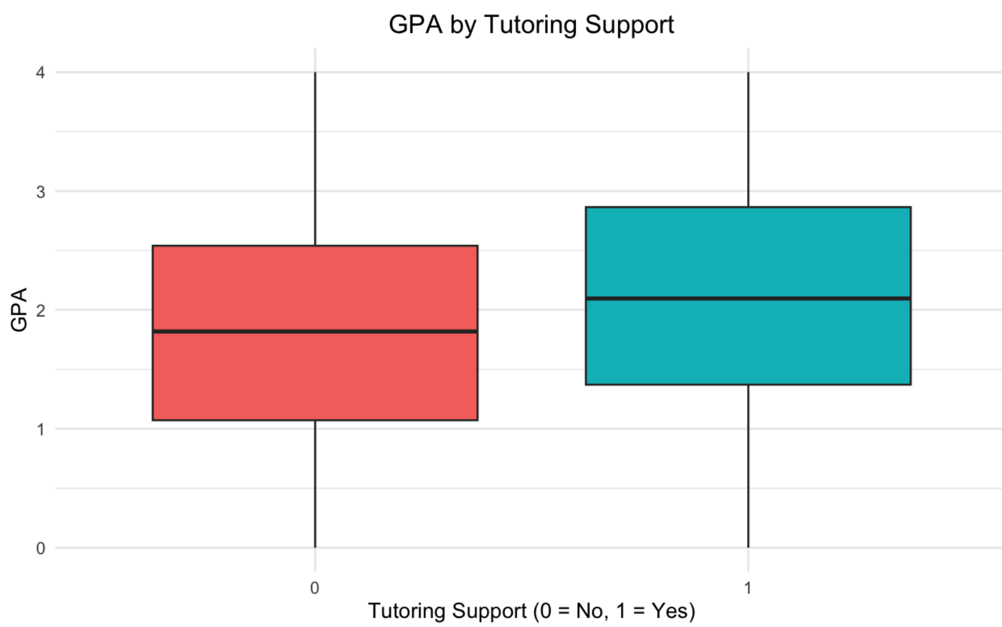
With this histogram of GPA levels by student's parental education levels, we see that a majority of students have parents under the "High School" and "Some College" levels of education and a majority of those students fall under the Low GPA category. The majority of students with parents of the other education levels also make up the Low GPA category. Only a small proportion of students in all parental education levels have Medium or High GPAs, with slightly more students having Medium GPAs. This distribution suggests that parental education levels may not be a significant predictor for this sample of students all on its own.

Parental Support



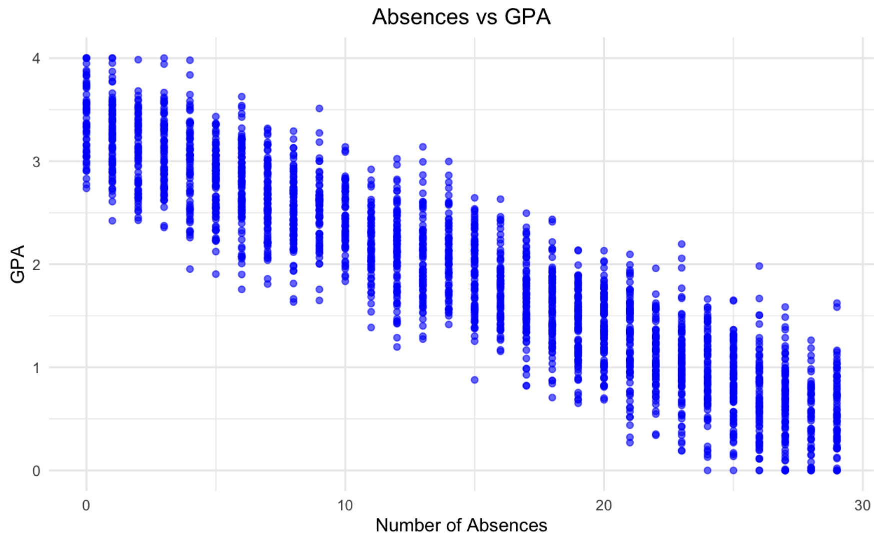
This histogram of GPA levels across different parental support levels shows that as the level of parental support increases, the proportion of high GPA increases up until the parental support level is “Very High”, and the proportion of low GPA decreases. The majority of students across the different parental support levels fall under the Low GPA category. There is a smaller number of students across the parental support levels who have Medium GPAs. From the parental support level from “None” to “High,” there are slight increases in the number of students with High GPAs in proportion to the total number of students under each parental support level. With parental support being “Very High,” there is a balanced amount of students with Medium and High GPA. The distribution of GPA categories varies across the levels of parental support, this indicates that it could be a significant predictor in predicting students’ academic performance.

Study Habits - Tutoring



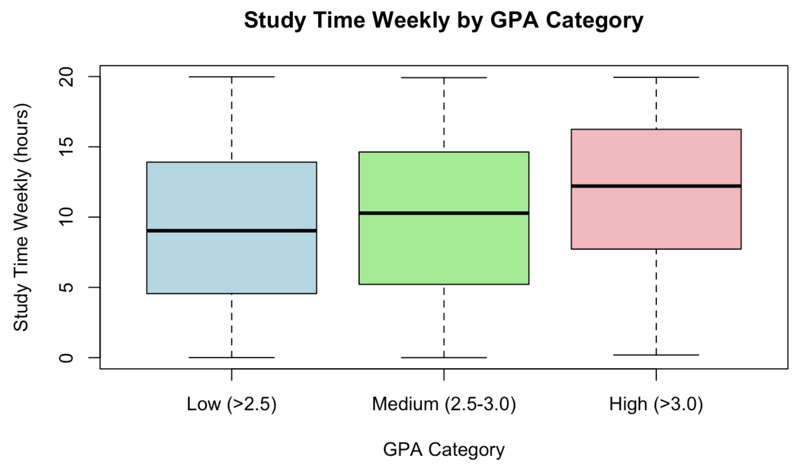
From the observed box plot we compare the two student groups, one group not involved in tutoring and the other group that is. On average tutored students have a higher GPA than the non-tutored group. Although both groups have overall similar GPA ranges to one another suggesting that tutoring does not eliminate student performance variability whatsoever.

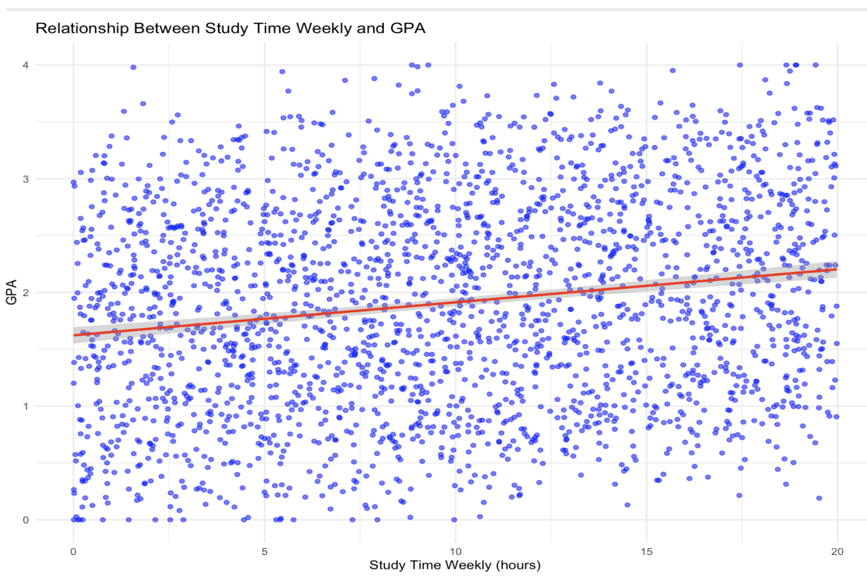
Study Habits - Absences



This plot reveals a clear negative correlation between the number of absences and GPA. Implying that as the number of absences increases the value of GPA will tend to decrease. The highest performers in the dataset or students who had a GPA close to a 4.0 hovered around the 0 to 5 absences range, while students who were classified as low performers (under a 2.0 GPA) were predominantly clustered around the 20 or more absences range. There were also a few outliers – students who had a fair amount of absences, 10 or more, and possessed a decent GPA (3.0 to 4.0). But these outliers are fairly rare in occurrence.

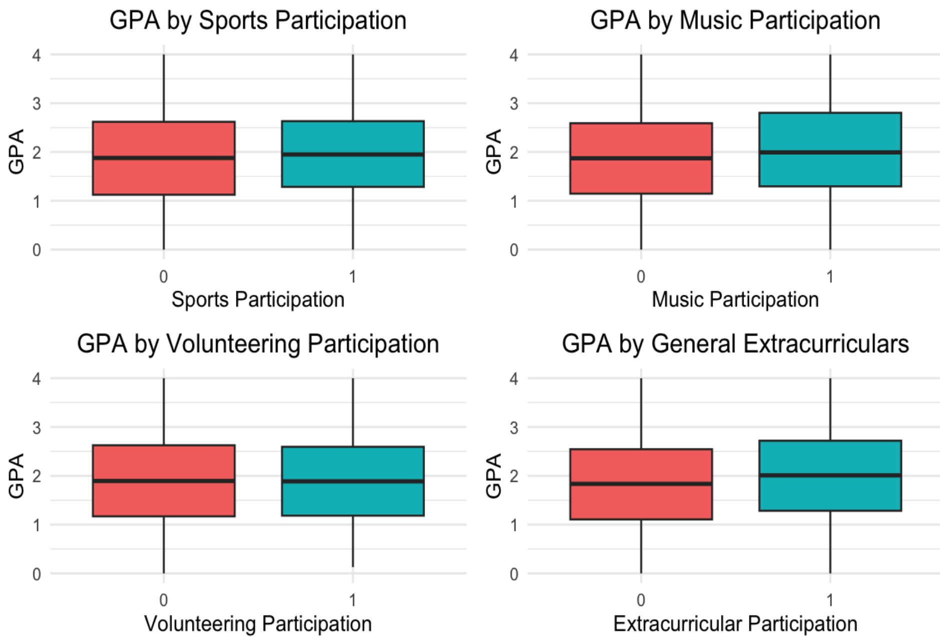
Study Habits - StudyTimeWeekly





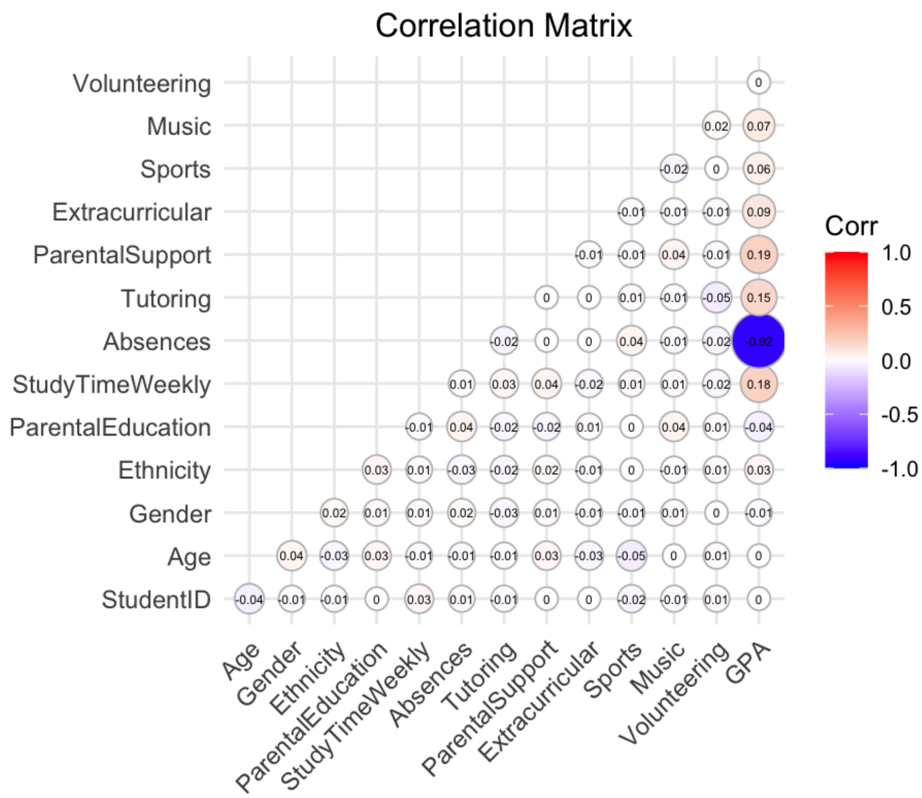
To examine the relationship between the amount of time spent studying per week and GPA, two plots will be used. The boxplot showcases a consistent increase in GPA parallel to an increase in hours spent studying weekly. The scatter plot further reinforces the idea of a positive but there is fairly weak relationship between study time and GPA. A single blue dot represents a specific student in our data and with the plot being fairly scattered it implies the high presence of variability in the data, suggesting the idea that there are other variables affecting student performance than just the amount of time studied per week, but with that said it still plays a part.

Extracurriculars



All extracurricular activities in our dataset whether it be sports, music, volunteering, or "general" extracurriculars are encoded as binary variables, representing whether a student is involved in an activity or not. Although the variables do not tell us the specific level of involvement for a particular student, we observed using the boxplots, a slight increase in GPA for students who were involved in at least one extracurricular activity. Because these plots do not tell us the significance of extracurricular activities this warrants deeper analysis which will be explored further in our report.

Correlation Matrix



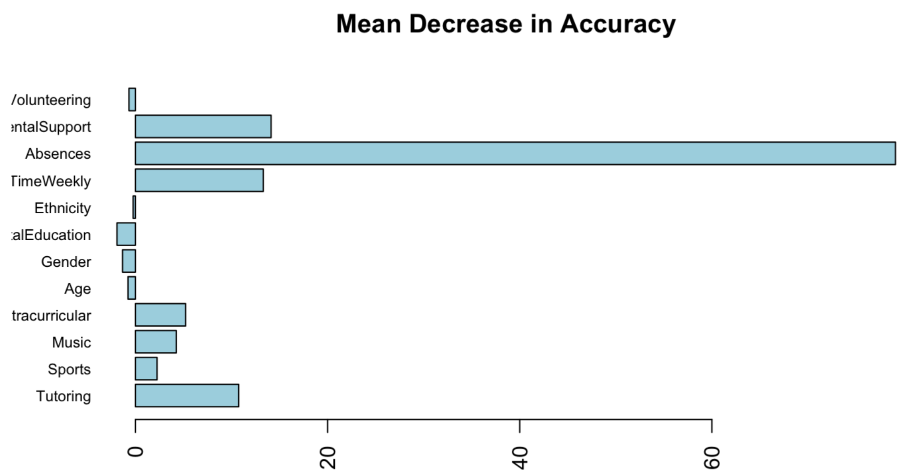
With our correlation matrix, we can explore how the variables interact with one another and more importantly GPA. Variables such as ParentalSupport, Tutoring, and StudyTimeWeekly all exhibited positive but overall weak correlation values with GPA, ranging from (0.2 to 0.15). Secondly, we have the Absences variable which is the only variable that has a negative correlation in the entire dataset and is by far the strongest, with GPA having a value of (-0.92). Other variables such as Ethnicity, Gender, and Age all exhibited negligible correlation values with GPA, suggesting that these may not play an important role in predicting student performance.

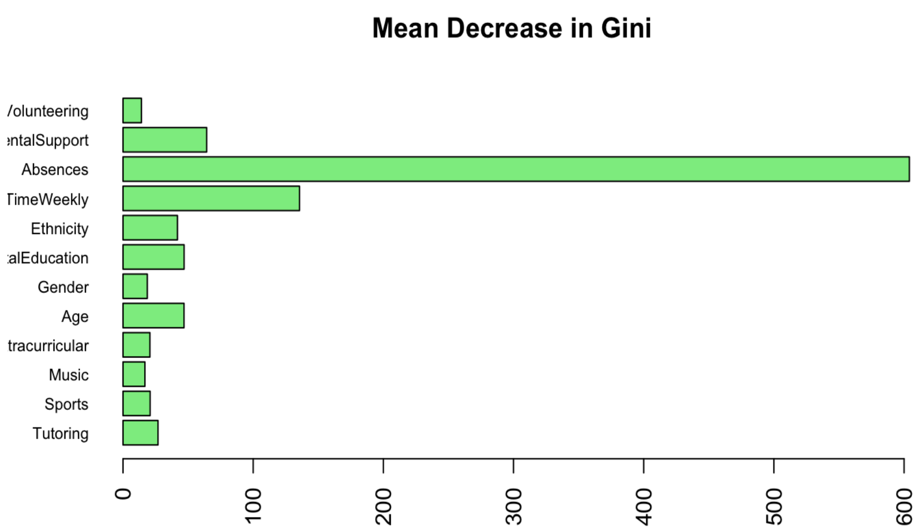
Model Selection

Multinomial regression, an extension of logistic regression, is used to predict a categorical dependent variable with more than two levels. It models the relationship between predictors and the probability of each category of the dependent variable relative to a reference category, estimating log-odds ratios to compare each category against the reference. In our analysis, the dependent variable, GPA, has three categories: Low (the reference category), Medium, and High, while the predictors include both continuous and categorical variables. Multinomial regression enables us to evaluate how these predictors influence the likelihood of being in the Medium or High GPA categories compared to the Low GPA category. To apply this method, we estimated two equations: the log-odds of Medium GPA relative to Low GPA, and the log-odds of High GPA relative to Low GPA. This approach is appropriate because logistic or linear regression would not effectively capture the structure of our data, given the three-category outcome variable.

Significant Predictors

Random Forest Variable Importance Plots





Using Random Forest, we evaluated the importance of our initial predictors by analyzing the mean decrease in both accuracy and Gini importance. This method allowed us to identify the most influential predictors in our model. The most important predictors were Absences, StudyTimeWeekly, Parental Support, Extracurricular activities, Sports participation, and involvement in Music. These predictors are considered most important because they produce the highest mean decrease in both accuracy and Gini importance, meaning they have high predictive power and should not be removed from the model.

	Intercept	Age	Gender	Ethnicity	Parental Education	StudyTime Weekly	Absences	Tutoring	Parental Support	Extracurricular	Sports	Music	Volunteering
Medium GPA	0.27974223	0.1882610	0.45263685	0.8524296	0.7161515	0	0	0	0	4.551914e-14	1.912141e-09	9.095482e-05	0.3026991
High GPA	0.01555246	0.5838568	0.07741703	0.6218271	0.8356795	0	0	0	0	0.000000e+00	0.000000e+00	5.800556e-10	0.7653241

Based on the p-values from the initial model, we should only include predictors that are statistically significant ($p < 0.05$) for at least one of the GPA categories. This reduces the model's complexity while retaining its predictive power. Notice that these align with the Random Forests variable importance plot from above.

We needed to remove ParentalSupport because it was highly correlated with tutoring (HIGH VIF: 187) and after removing it, our model was much better. In real life, this makes sense, as those with parental support, typically, will also have tutoring.

Thus, the predictors used in our final model were: StudyTimeWeekly, Absences, Tutoring, Extracurriculars, Sports, and Music.

Final Model

Because we had three categories, we get two equations for our final multinomial regression model that retains statistically significant predictors. They are:

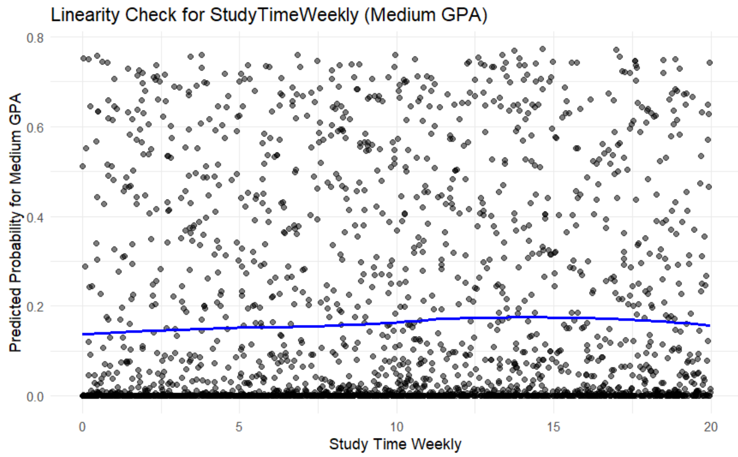
log(P(Medium GPA)/P(Low GPA)) = 2.0076249 + 0.1988159 * StudyTimeWeekly - 0.6383501 * Absences + 1.585080 * Tutoring + 1.407852 * Extracurricular + 0.912152 * Sports + 0.8456578 * Music

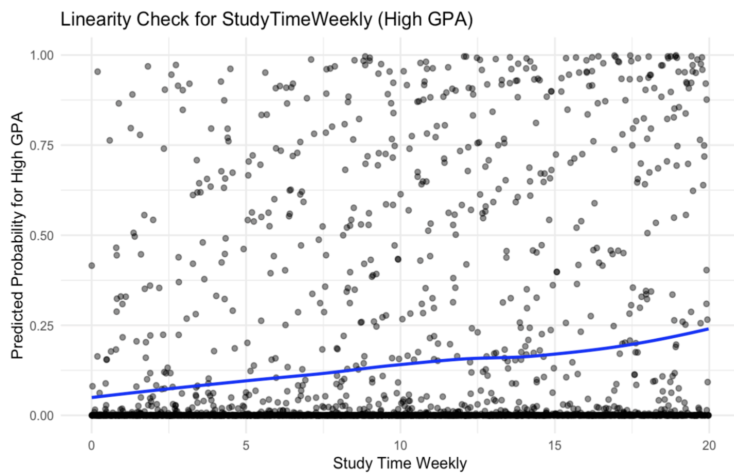
log(P(High GPA)/P(Low GPA)) = 0.6241777 + 0.4007474 * StudyTimeWeekly - 1.3141799 * Absences + 3.560722 * Tutoring + 2.581424 *Extracurricular + 2.814718 * Sports + 2.0922313 * Music

This final model shows that higher study time, tutoring, and involvement in extracurricular activities, sports, and music participation increase the likelihood of achieving a Medium or High GPA compared to a Low GPA, with stronger effects observed for the High GPA. Conversely, more absences reduce the likelihood of being in the Medium or High GPA categories relative to Low GPA.

Validating Model Assumptions

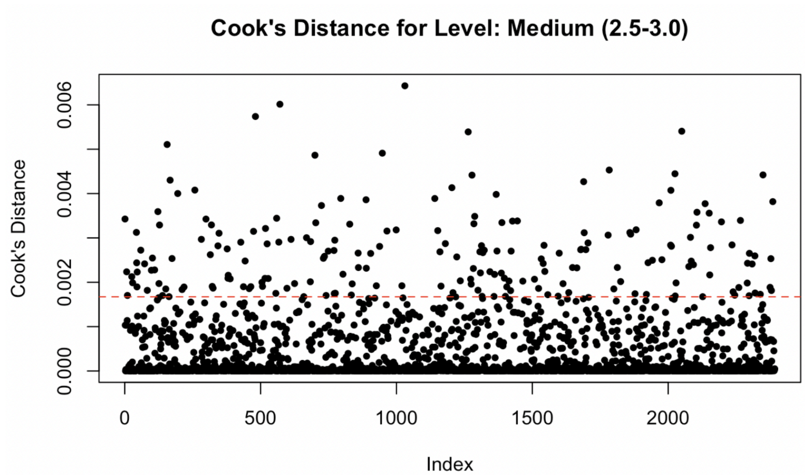
Linearity Assumption

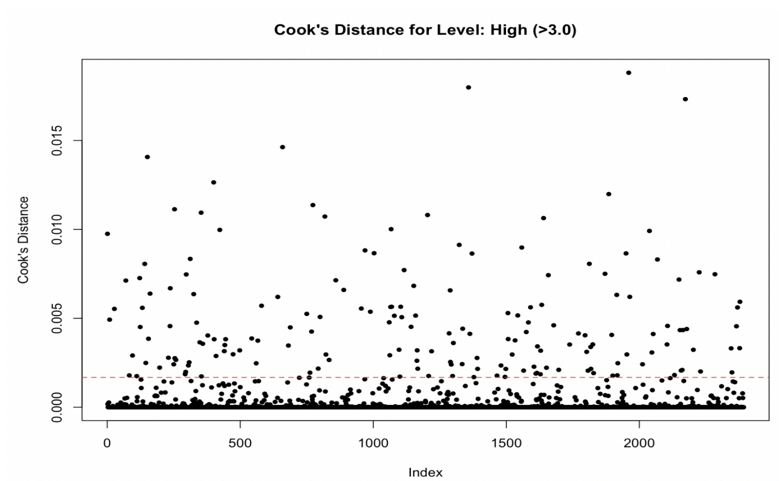




Logistic regression assumes a linear relationship between the log odds of the outcome (GPA) and the predictor variables, so we checked the linearity assumption for our continuous numerical predictor, StudyTimeWeekly. The above graphs show the log odds of the GPA categories (Low vs. Medium and Low vs. High) plotted against the variable StudyTimeWeekly. Because there is no observable trend in the graphs, and the lines are approximately straight and linear, we can assume there is a linear relationship between our continuous predictor and the log-odds of the outcome. This validates our model's interpretability and predictive performance and means a transformation is not needed for this variable.

Cook's Distance





Next, we wanted to identify if influential observations have a significant impact on the regression model's coefficients, and whether or not they should be removed. Upon evaluating Cook's Distance, we discovered that 175 out of 2,392 observations fell past the line on the plots, which is 7.3% of the data. Between the 2 plots, not a single observation fell beyond a distance of 0.006, suggesting that even the points that fall past the line are not disproportionately affecting the model. Because they likely represent valid variability in the dataset, removing them could bias the model by excluding important subgroups, thus we chose to retain these observations in our data and proceeded without removing any outliers.

VIF for Multicollinearity

Variance Inflation Factor

Absences	Tutoring	Extracurricular	Sports	Music	Weekly Study Time
1.002037	1.000442	1.000372	1.002321	1.000818	1.00125

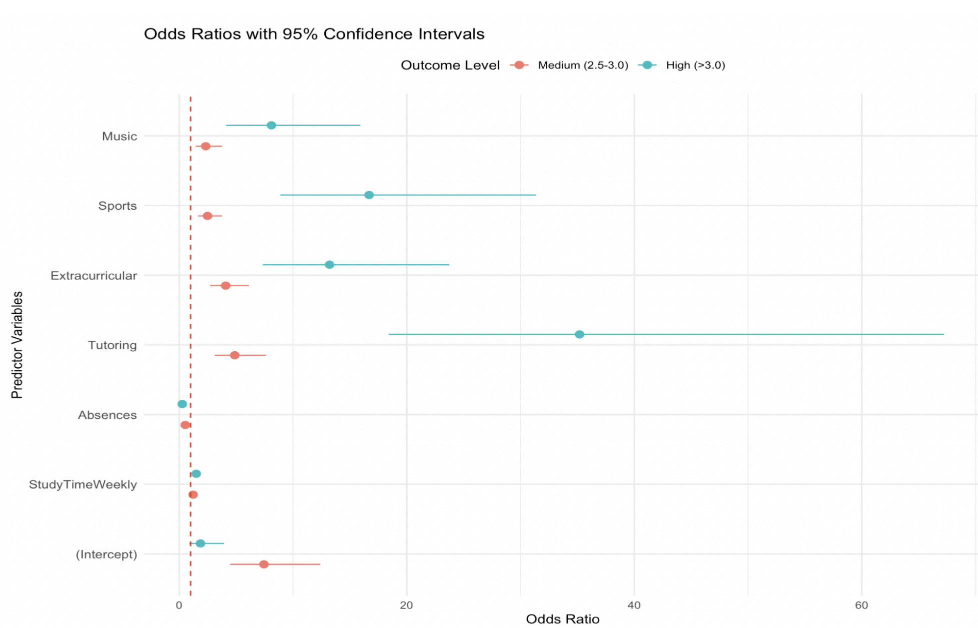
For our next assumption, we calculated the Variance Inflation Factor (VIF) for each predictor in our final model–a statistical measure used to detect multicollinearity. For each predictor, we measured how well it is explained by the other predictors, and because all of the values are close to 1, we can be sure that multicollinearity is not an issue in our model (the threshold for high VIF is usually greater than 5). Thus, no predictors needed to be removed or combined from our final model. Small VIF values make the model more interpretable, minimizes coefficient variances, enhances predictive performance, and creates a more stable model overall.

Odds Ratios

Table with Odds Ratios

term <chr>	y.level <chr>	Odds_Ratio <dbl>	Lower_95CI <dbl>	Upper_95CI <dbl>
(Intercept)	Medium (2.5–3.0)	7.4456124	4.4741947	12.3904185
(Intercept)	High (>3.0)	1.8667103	0.8847431	3.9385530
StudyTimeWeekly	Medium (2.5–3.0)	1.2199574	1.1727398	1.2690761
StudyTimeWeekly	High (>3.0)	1.4929401	1.4070187	1.5841084
Absences	Medium (2.5–3.0)	0.5281631	0.4909429	0.5682051
Absences	High (>3.0)	0.2686946	0.2364575	0.3053268
Tutoring	Medium (2.5–3.0)	4.8796816	3.1181920	7.6362497
Tutoring	High (>3.0)	35.1886048	18.4178921	67.2301641
Extracurricular	Medium (2.5–3.0)	4.0871687	2.7318301	6.1149294
Extracurricular	High (>3.0)	13.2159401	7.3591978	23.7337107
Sports	Medium (2.5–3.0)	2.4896746	1.6448076	3.7685134
Sports	High (>3.0)	16.6884656	8.8803046	31.3620867
Music	Medium (2.5–3.0)	2.3295097	1.4394078	3.7700333
Music	High (>3.0)	8.1029749	4.1253117	15.9159373

Plot of Odds Ratio



Once our final model was confirmed, we examined the odds ratios for each non-base category to understand how each predictor influences the odds of being in the Medium and High categories, relative to the base category, Low GPA. These ratios are calculated by exponentiating the coefficients, or log-odds from the multinomial equation. As shown above, the predictors have different effects on the odds of being in different categories, for example, Absences, Tutoring, and Extracurricular are the most significant predictors for distinguishing Medium GPA from Low GPA, and Tutoring, Sports, Absences, Music and Extracurricular are the most significant in distinguishing High GPA from Low GPA. As Absences increase, GPA tends to decrease, while increasing all other predictors tend to increase GPA as well. It is important to note that Absences and Weekly Study Time are numerical predictors, while the rest are categorical, so the odds ratios must be interpreted within context to distinguish which are most important. Absences are a very significant predictor but do not appear so on the plot compared to some of the categorical variables, as the odds ratios have slightly different interpretations.

We also calculated 95% confidence intervals for each of the odds ratios, to assess reliability and see which predictors are significant. None of the 95% confidence intervals for the predictors include 1, thus we can be 95% confident that, holding all other variables constant, a one-unit increase in the numerical predictors, or a change in level for the categorical predictors, has a statistically significant impact on a student's GPA.

Interpretations Of Odds Ratios Within Context

As previously stated, it is important to understand what the odds ratios mean within the context of the research question, data, and model. Below are the interpretations of each of our odds ratios for both categories:

For Medium GPA (2.5-3.0):

- StudyTimeWeekly: 1.22
    - For each additional hour of study per week, the odds of being in the Medium GPA category increase by 22%.
- Absences: 0.53
    - For each additional absence, the odds decrease by 47%.
- Tutoring: 4.88
    - Students receiving tutoring are 4.88 times more likely to be in the Medium GPA category.
- Extracurricular: 4.08
    - Participation in extracurricular activities increases the odds of Medium GPA by 308%.
- Sports: 2.49
    - Participation in sports increases the odds of Medium GPA by 149%.
- Music: 2.33
    - Participation in music increases the odds of Medium GPA by 133%.
- Intercept: 7.45
    - Baseline odds of being in the Medium GPA category

For High GPA (>3.0):

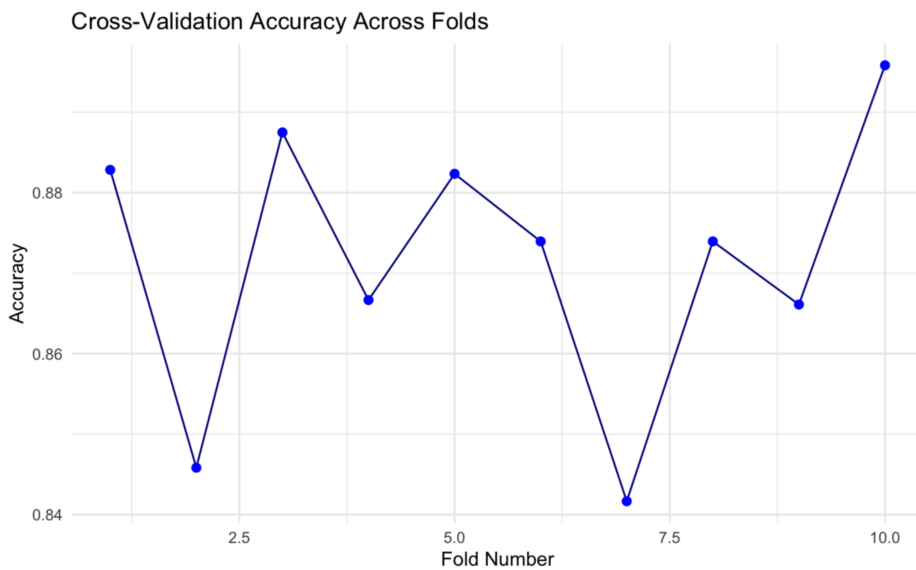
- StudyTimeWeekly: 1.49
    - For each additional hour of study per week, the odds of being in the High GPA category increase by 49%.
- Absences: 0.27
    - For each additional absence, the odds decrease by 73%.
- Tutoring: 13.23
    - Students receiving tutoring are 13.23 times more likely to achieve a High GPA.
- Extracurricular: 35.18
    - Participation in extracurricular activities increases the odds of a High GPA by 3418%.
- Sports: 16.69
    - Participation in sports increases the odds of a High GPA by 1569%.
- Music: 8.10
    - Participation in music increases the odds of a High GPA by 710%.
- Intercept: 1.87
    - Baseline odds of being in the High GPA category.

These interpretations make understanding the impact and predictive power of each of the predictors less complicated.

Strength of the Model

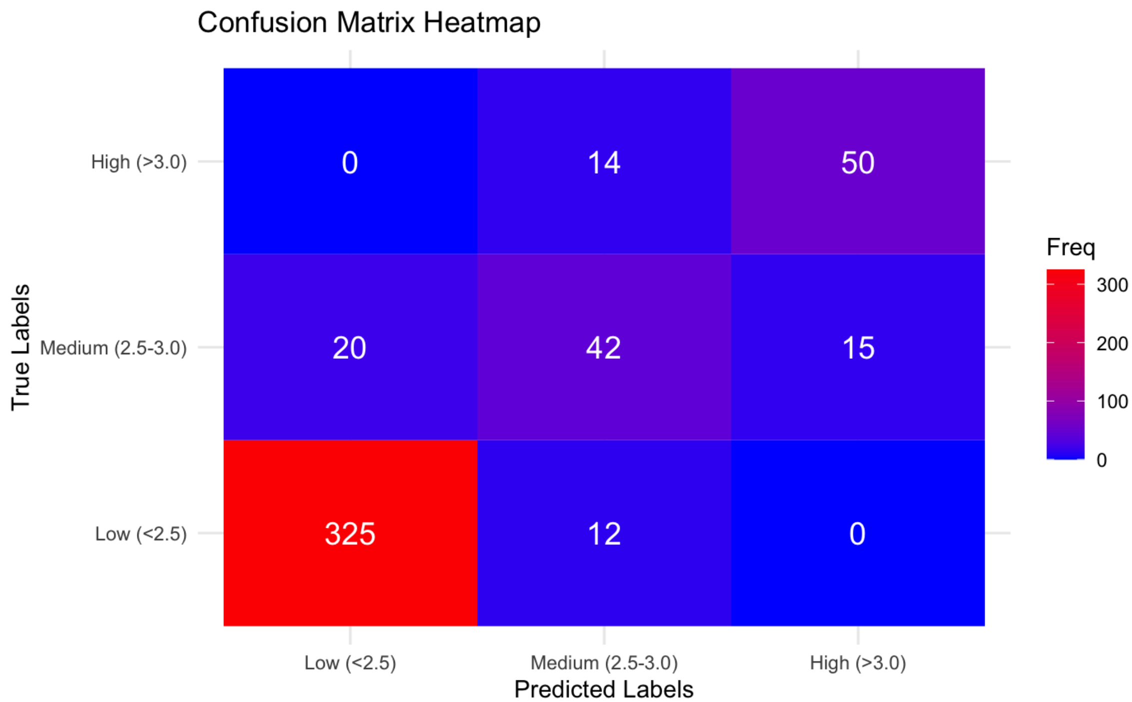
Cross Validation

To assess the strength of our model, we used two cross validation methods: k-fold cross-validation with 10 folds, and leave one out cross-validation (LOOCV). Both evaluate the performance of the model by testing it on multiple subsets of the data to assess how well the model generalizes to unseen data, reducing the risk of overfitting.



K-fold cross validation produced the above visual, with the accuracy of the model pictured at each fold, ranging from 84% to 91%. Although there is a slight variance between folds, the range of the y-axis is small, so it is all within a 7% range. The average accuracy across all folds was 87.17%, and the accuracy from the leave one out cross-validation was 87.25%. Thus, the data was correctly categorized into low, medium, and high GPAs around 87% of the time, indicating high model accuracy.

Confusion Matrix



For further model assessment, we divided the data into an 80-20 train-test split and evaluated its performance, visualizing it with the above confusion matrix heatmap. The overall accuracy achieved was 87.24%, which is similar to that of the cross-validation methods. Because the data is skewed toward the low category, our model is best at distinguishing these data points, as there is more to train it on–the accuracy for classifying Low GPA is 94.02%, for Medium GPA is 61.76%, and High GPA is 76.92%. The class imbalance in the data leads the model to focus more on the dominant class–this could be resolved by adding in more data points from the Medium or High classes, as well as readjusting our GPA cutoff for the categories. However, overall, the accuracy of the model averages at 87.24%, which is relatively high.

IV. Results

Our research highlights that while demographic characteristics like age, gender, and ethnicity have little bearing on GPA results, behavioral factors like study time, absences, tutoring, and extracurricular involvement have a statistically significant impact on high school academic performance. Our multinomial logistic regression model achieved an average accuracy of around 87% in classifying students into GPA categories throughout various validation methods (80/20 test/train split, 10-fold, LOOCV). These results indicate that behavioral factors are strong predictors of academic performance, highlighting the importance of fostering engagement in academic and extracurricular activities to improve academic performance. Despite some limitations, our analysis offers valuable insights for both educators and policymakers, especially when addressing academic challenges within underperforming academic environments.

V. Discussion

Shortcomings

While our research produced valuable insights, it faced several limitations related to data constraints and modeling assumptions. A key challenge was the significant class imbalance in the dataset, with a majority of students falling into the low GPA category, which limited the model's ability to accurately predict medium and high GPA outcomes. Because the data likely comes from a low-income, underperforming school district, we were faced with the choice to either have this class imbalance or redefine the GPA categories so that the cutoff for Medium and High is much lower. However, we aimed to stay true and realistic to common understandings of these categories and we did not want to skew our findings too much because of the naturally skewed nature of the data. Additionally, parental support–a potentially influential variable– had to be excluded due to high multicollinearity with tutoring, which may have oversimplified the analysis. The dataset also included a limited number of numerical predictors (only three–one continuous and two discrete), which may have restricted the ability to uncover more nuanced relationships between variables. Lastly, there was no description of where the data was collected from, but from our observation, its scope is likely specific to schools with predominantly poor academic performance, limiting the generalizability of the findings to more diverse or higher-achieving schools.

In terms of modeling, multinomial regression made the most sense for the structure of our data, but we could further explore relationships via more complex models such as neural networks or other clustering techniques. We also had a smaller number of data points, so we are unable to remove more influential points–these leverage points may have skewed the model's predictions and influenced the interpretation of certain predictors. Increasing the number of observations would help resolve this issue, as well as make the findings more robust. These limitations highlight areas for improvement in future studies, which could refine the robustness and applicability of the results.

Recommendations

To address the factors influencing academic performance, we propose several targeted interventions. Tutoring was strongly associated with improved GPA outcomes, emphasizing the need for schools to develop academic support programs, especially targeted tutoring initiatives for students with low GPAs. Encouraging more structured study routines can also help students reach their weekly study time goals, which correlates with higher academic achievement overall. Additionally, schools should promote active participation in extracurricular activities, such as sports and music, as these are linked to better GPA outcomes and overall personal growth. Addressing absenteeism is vital as it is strongly correlated with academic performance; schools should implement attendance incentives, such as rewards for students who show consistent progress, to motivate struggling students to remain engaged in their education, particularly when there is a lack of parental support or involvement. Finally, in low-income areas, tutoring and after-school programs should be enhanced, as they sometimes lack structure, support, and meaningful time usage, and are there instead to appease parents and superintendents.

To enhance the predictive power of similar studies, future research should incorporate a wider range of predictors, such as socioeconomic factors, to better capture multifaceted influences on academic success. There must be monitoring and evaluation systems implemented in schools to track the progress of new programs. By adopting these strategies, education and policymakers can create a more engaging, supportive learning environment. These recommendations provide a framework for supporting academic growth at both the individual and institutional levels.