Estimating Students' Academic Performance: Predicting Final Exam Scores

Priscilla Dixon, William McCormack, Olivia Nguyen

Contents

Abstract

This project examines the relationship between student habits, available resources, and demographic influences on academic performance, measured as a binary outcome (above or below the median final exam score). Using exploratory data analysis, logistic regression modeling, and K-Fold Cross-Validation, we investigated key predictors, including Attendance, Hours Studied, Tutoring Sessions, Access to Resources, Parental Involvement, and more.

The logistic regression model with a comprehensive set of predictors achieved high accuracy (90%) and strong discriminative power (AUC = 0.97), with well-balanced sensitivity (91%) and specificity (88%).

Our findings indicate that the odds of above-median performance increased with attendance, hours studied, and tutoring sessions, while they decreased with low access to resources and low parental involvement. External factors such as demographic/background influences and students' resource accessibility were the most significant contributors. Challenges in the modeling process included addressing the apparent high predictive power of nearly all variables due to the large sample size (n = 6607), as well as ensuring model interpretability. These insights provide actionable guidance for identifying critical factors influencing student success and tailoring targeted educational interventions.

1 Statement of the Problem

The problem this project seeks to address is: "What factors most significantly influence academic performance, and how do these factors interact to predict success?"

By analyzing a dataset of student performance metrics, we aim to answer two key research questions:

1. Which individual factors from the dataset serve as the most significant predictors of academic performance, and how do they influence student outcomes?
2. To what extent do the three predictor groups (student habits, available resources, and demographic influences) contribute to the predictive power of our models in predicting academic success?

Through logistic regression modeling and exploratory data analysis, we strive to identify actionable insights that educators and policymakers can use to improve student outcomes, focusing on both individual factors and the broader categories they represent.

2 Variables in the Study

The data analyzed in this study came from a dataset with 20 total variables containing academic and demographic information about 6,607 students. The dataset includes variables that span three major categories: **Student Habits**, **Available Resources**, and **Demographic/Background Influences**. These categories encompass factors such as attendance, hours studied, tutoring sessions, access to resources, and parental involvement, as well as demographic information like family income and parental education levels.

Key variables include a mix of:

- **Categorical variables**: Examples include `Access_to_Resources`, `Peer_Influence`, and `Parental_Education_Level`.
- **Numerical variables**: Examples include `Attendance`, `Hours_Studied`, and `Previous_Scores`.
- **Binary target variable**: `Score_Class`, a categorical version of `Exam_Score`, which classifies students as either "Above the Median" or "Below the Median" based on their final exam scores.

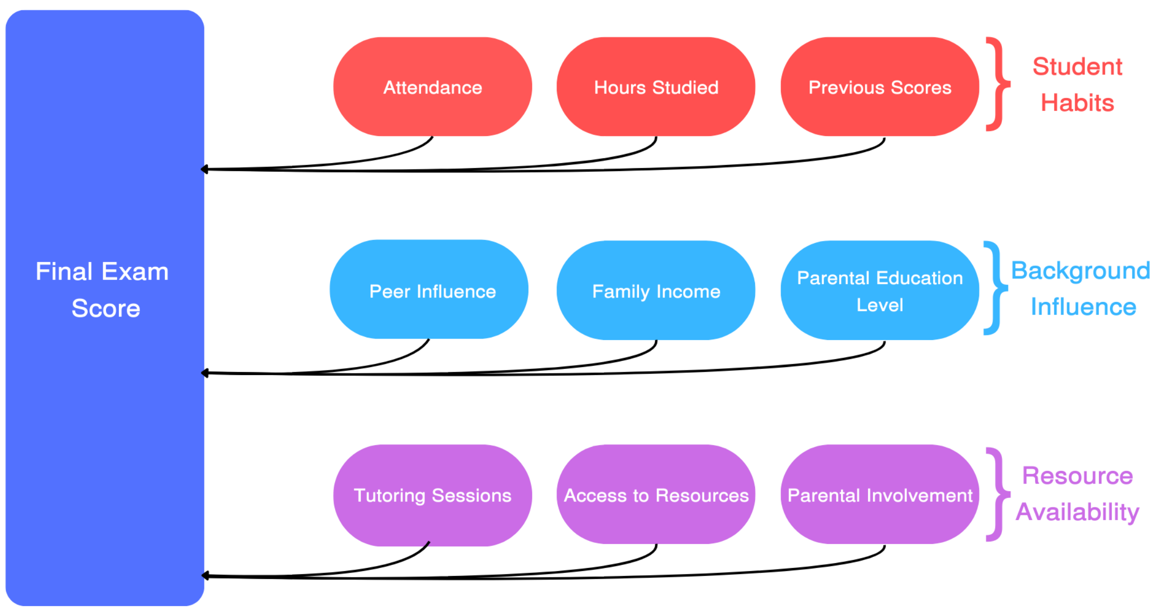
These variables were utilized to develop logistic regression models aimed at predicting academic performance and identifying the most significant factors influencing student success.

Table 1: Overview of All Variables

Variable	Variable Types	Measurement	Description
Score_Class – Response Variable	**Numerical**	**Above the Median, Below the Median**	**Modified Target Variable**
Exam_Score – Original Response/Target Variable	Numerical	50-100	Final exam score
Attendance	Numerical	60-100	Percentage of classes attended
Hours_Studied	Numerical	1-44	Number of hours spent studying per week
Previous_Scores	Numerical	50-100	Scores from previous exams
Tutoring_Sessions	Numerical	0-8	Number of tutoring sessions attended per month
Sleep_Hours	Numerical	5-8	Average number of hours of sleep per night
Physical_Activity	Numerical	5-7	Average number of hours of physical activity per week
Access_to_Resources	Categorical	Low, Medium, High	Availability of educational resources
Parental_Involvement	Categorical	Low, Medium, High	Level of parental involvement in the student's education
Peer_Influence	Categorical	Positive, Neutral, Negative	Influence of peers on academic performance
Family_Income	Categorical	Low, Medium, High	Family income level
Parental_Education Level	Categorical	High School, College, Postgraduate	Highest education level of parents
Motivation_Level	Categorical	Low, Medium, High	Student's level of motivation
Teacher_Quality	Categorical	Low, Medium, High	Quality of the teachers
Distance_from_Home	Categorical	Near, Moderate, Far	Distance from home to school
School_Type	Binary	Public, Private	Type of school attended
Gender	Binary	Male, Female	Gender of the student
Learning_Disabilities	Binary	Yes, No	Presence of learning disabilities
Internet_Access	Binary	Yes, No	Availability of internet access
Extracurricular_Activities	Binary	Yes, No	Participation in extracurricular activities

Schematic

Schematic representation of factors influencing final exam scores, categorized by student habits, background influences, and resource availability.



3 Exploratory Data Analysis

The dataset analyzed in this project contains academic and demographic information for 6,607 students. The data includes a mix of numerical, categorical, and binary variables grouped into three categories: **Student Habits**, **Available Resources**, and **Demographic Influences**. This section outlines the steps taken to preprocess and explore the data, followed by insights derived from visualizations and descriptive statistics.

3.1 Data Cleaning and Preprocessing

Initial data cleaning involved handling missing values by removing a handful of incomplete rows. Categorical variables were one-hot encoded, and a binary target variable, `Score_Category`, was created to classify students as "Above the Median" or "Below the Median" based on their final exam scores.

These transformations ensured compatibility with logistic regression modeling and facilitated meaningful EDA.

3.2 Descriptive Statistics

Descriptive statistics were calculated to summarize the dataset. Key variables, including but not limited to `Attendance` and `Hours_Studied`, showed minimal variability and had clear, positive association with `Exam_Scores`, with students in the Above the Median category consistently scoring higher across these metrics. Categorical variables like `Access_to_Resources` and `Peer_Influence` highlighted notable differences between high and low performers.

3.3 Visualizations

To uncover patterns in the data, we generated several visualizations that highlight the classification nature of the project. These visualizations examine relationships between predictors and the binary outcome variable (`Score_Category`). Below are the selected visualizations with meaningful captions to guide interpretation:

- **Figure 1**: A pie chart visualizing the proportion of students classified as "Above the Median" and "Below the Median."
- **Figure 2**: A scatter plot illustrating the relationship between `Hours_Studied` and `Exam_Score`, with points colored by `Score_Category`.
- **Figure 3**: A histogram displaying the distribution of `Attendance` for each `Score_Category`.
- **Figure 4**: A histogram showing the distribution of `Previous_Scores` for each `Score_Category`.
- **Figure 5**: A boxplot illustrating the distribution of `Hours_Studied` by `Score_Category`.
- **Figure 6**: A stacked bar chart showing the proportion of `Score_Category` across levels of `Parental_Education_Level`.
- **Figure 7**: A bar chart showing the distribution of `Score_Category` by `Access_to_Resources`.
- **Figure 8**: A correlation matrix showing the relationships between numerical predictors and `Exam_Score`.

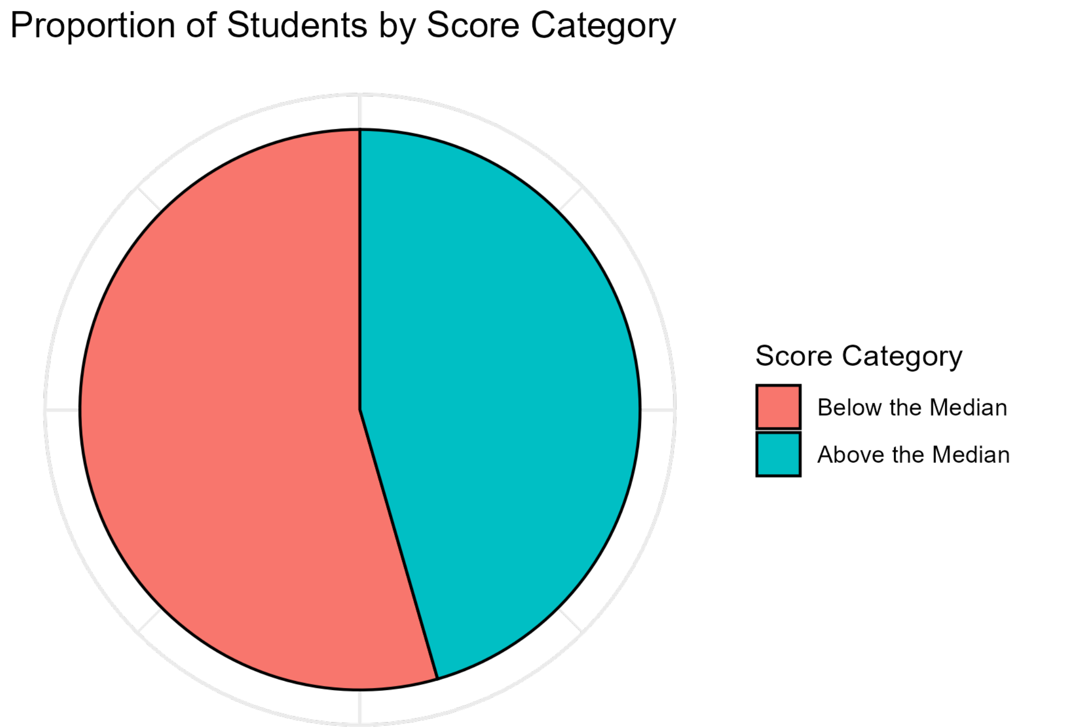


Figure 1: This pie chart shows that the dataset is nearly balanced, with slightly less students classified as "Above the Median" than "Below the Median."

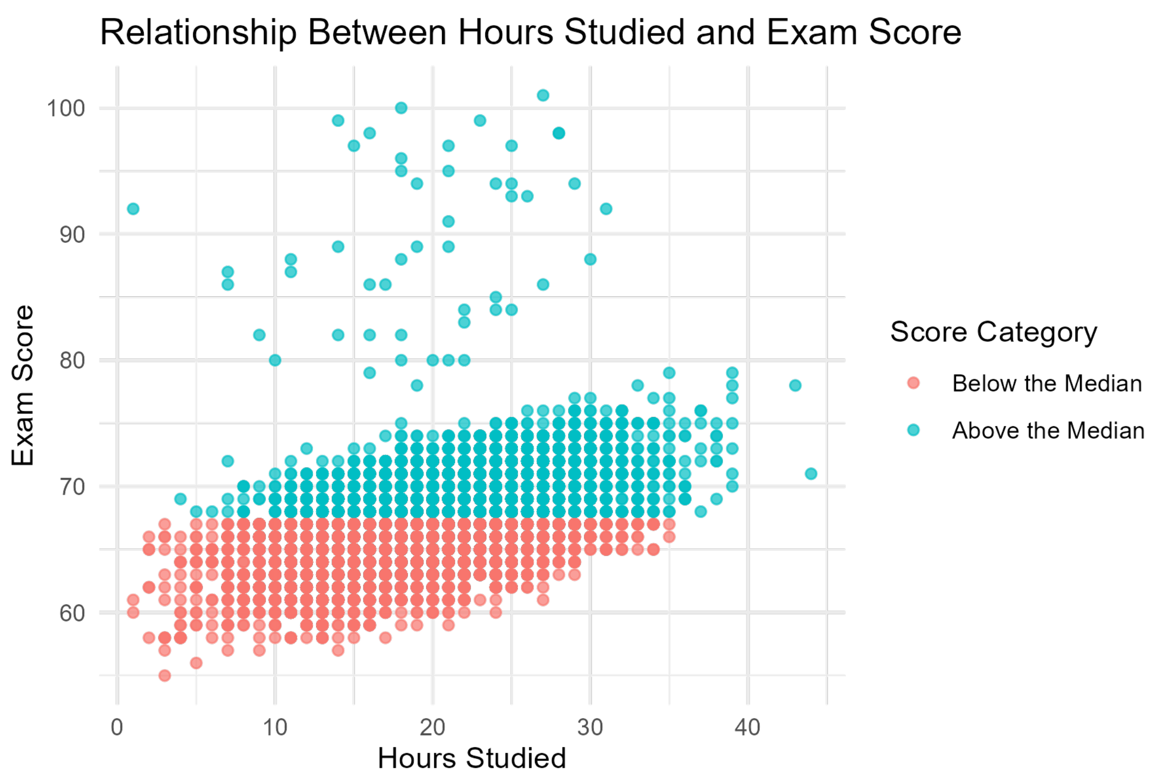


Figure 2: This scatter plot illustrates that students with higher `Hours_Studied` generally achieve higher `Exam_Scores`, with clear separation between "Above the Median" and "Below the Median" categories.

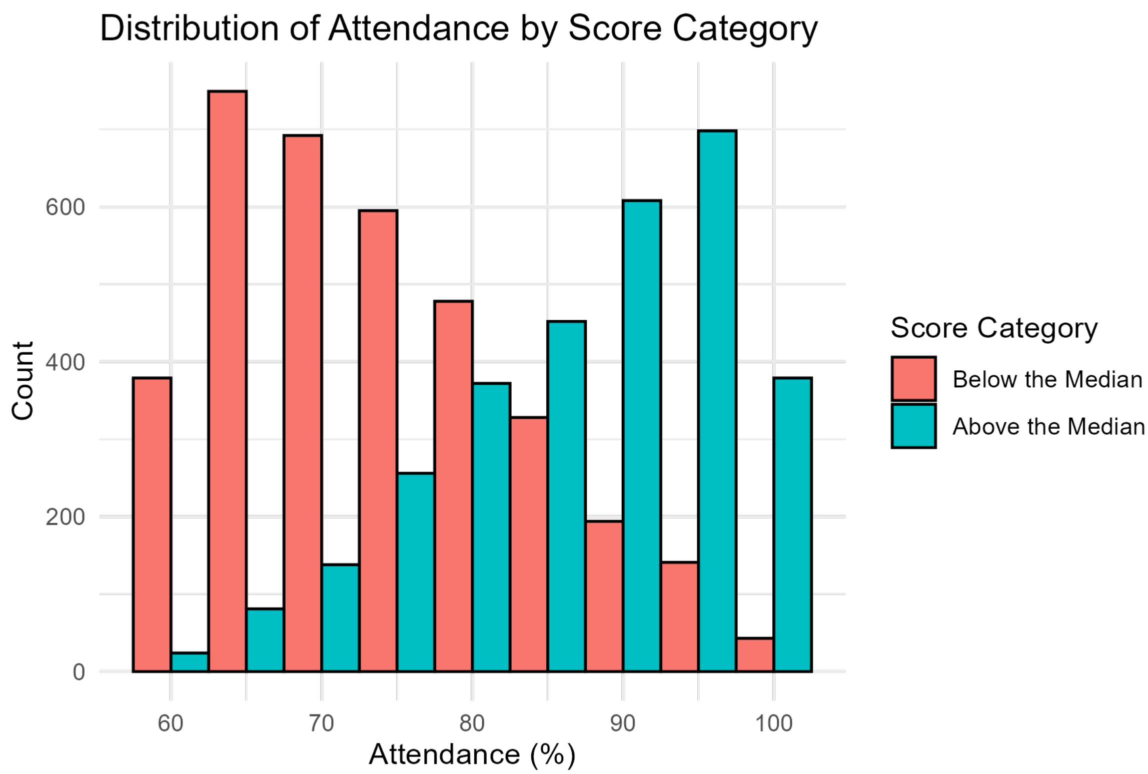


Figure 3: This histogram shows that higher `Attendance` is associated with a greater proportion of students in the "Above the Median" category, highlighting its importance for academic performance.

Figure 4: This histogram illustrates that students with higher `Previous_Scores` tend to fall into the "Above the Median" `Score_Category`, demonstrating the predictive value of prior performance.

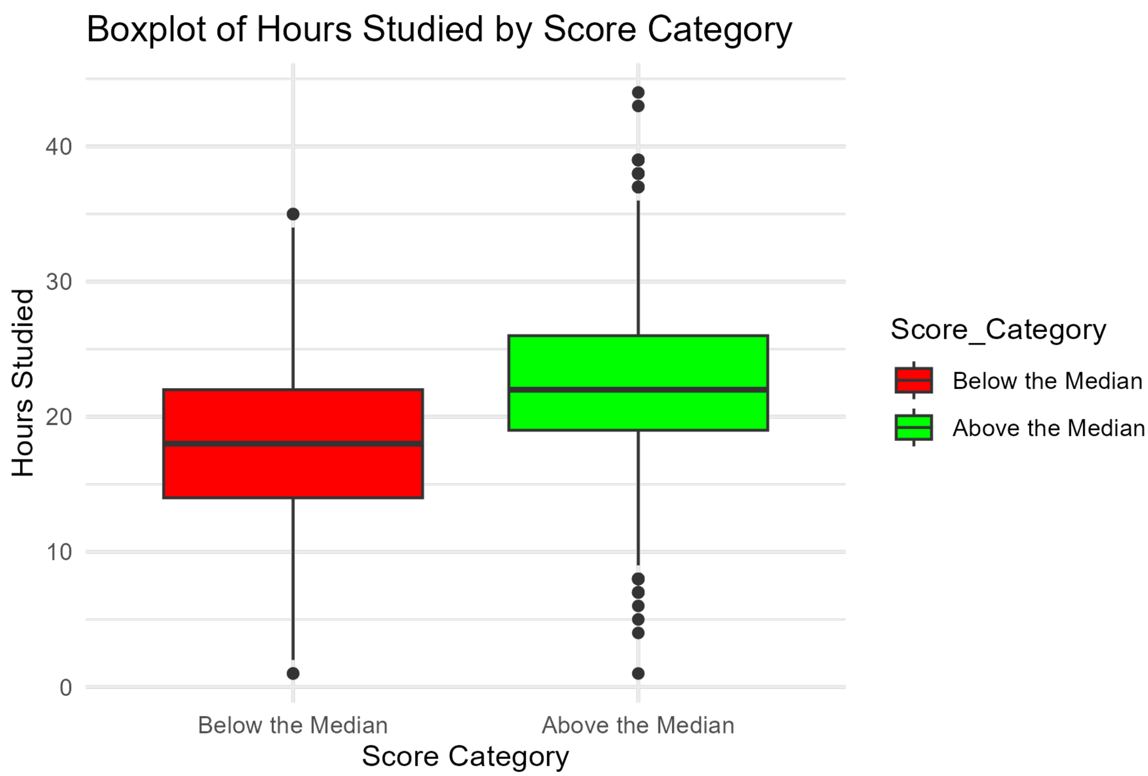


Figure 5: This boxplot highlights that students in the "Above the Median" `Score_Category` have a wider range and higher median for `Hours_Studied` compared to those in the "Below the Median" category.

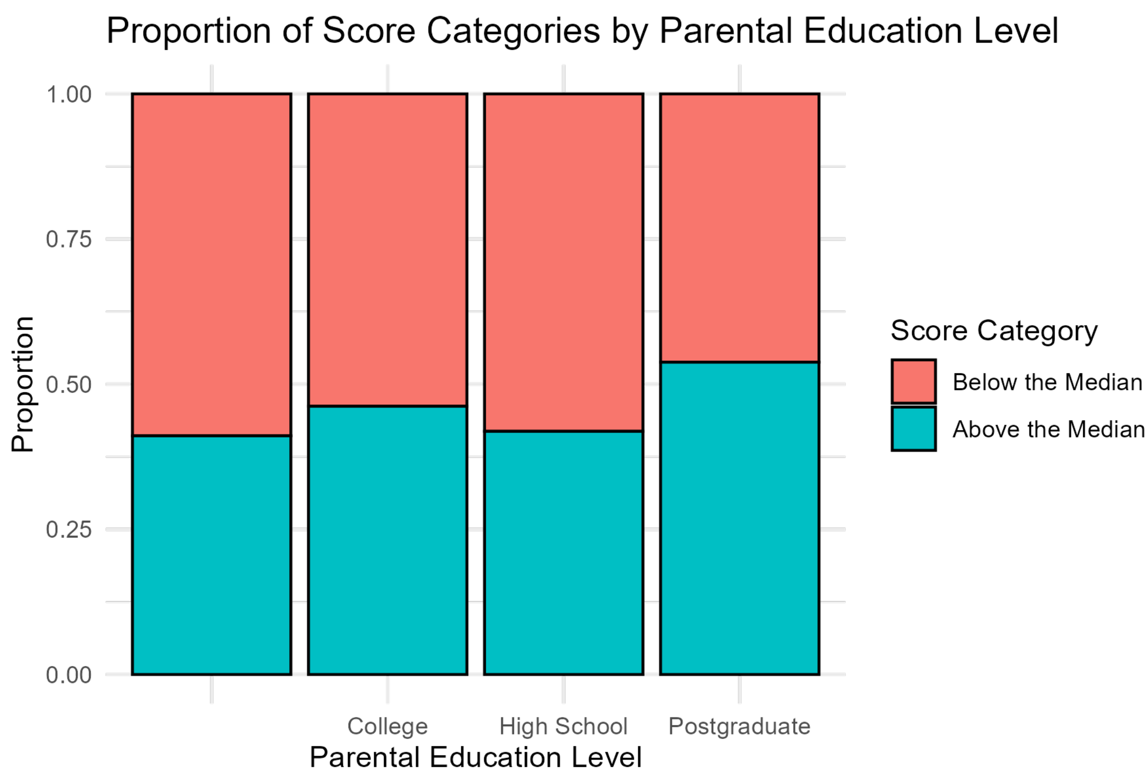


Figure 6: This stacked bar chart shows that students with parents who have a postgraduate education are more likely to be classified as "Above the Median."

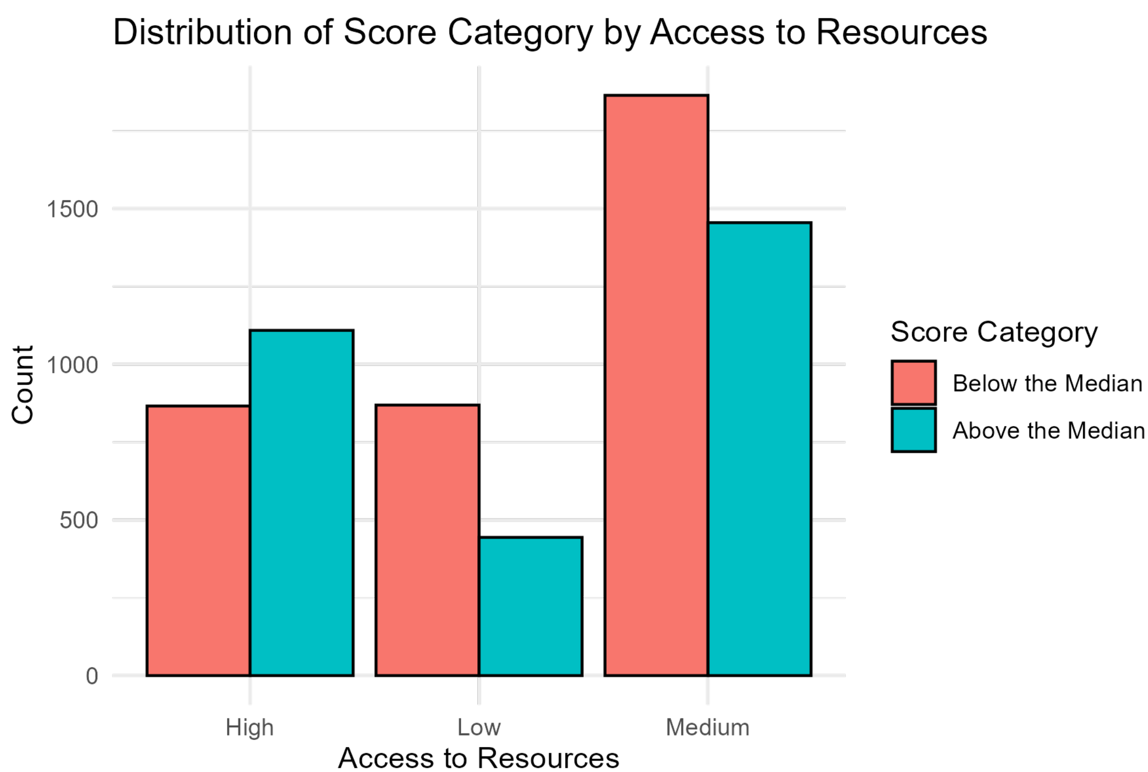


Figure 7: This bar chart indicates that students with high access to resources are more likely to be classified as "Above the Median," while those with low and medium access tend to perform below the median.

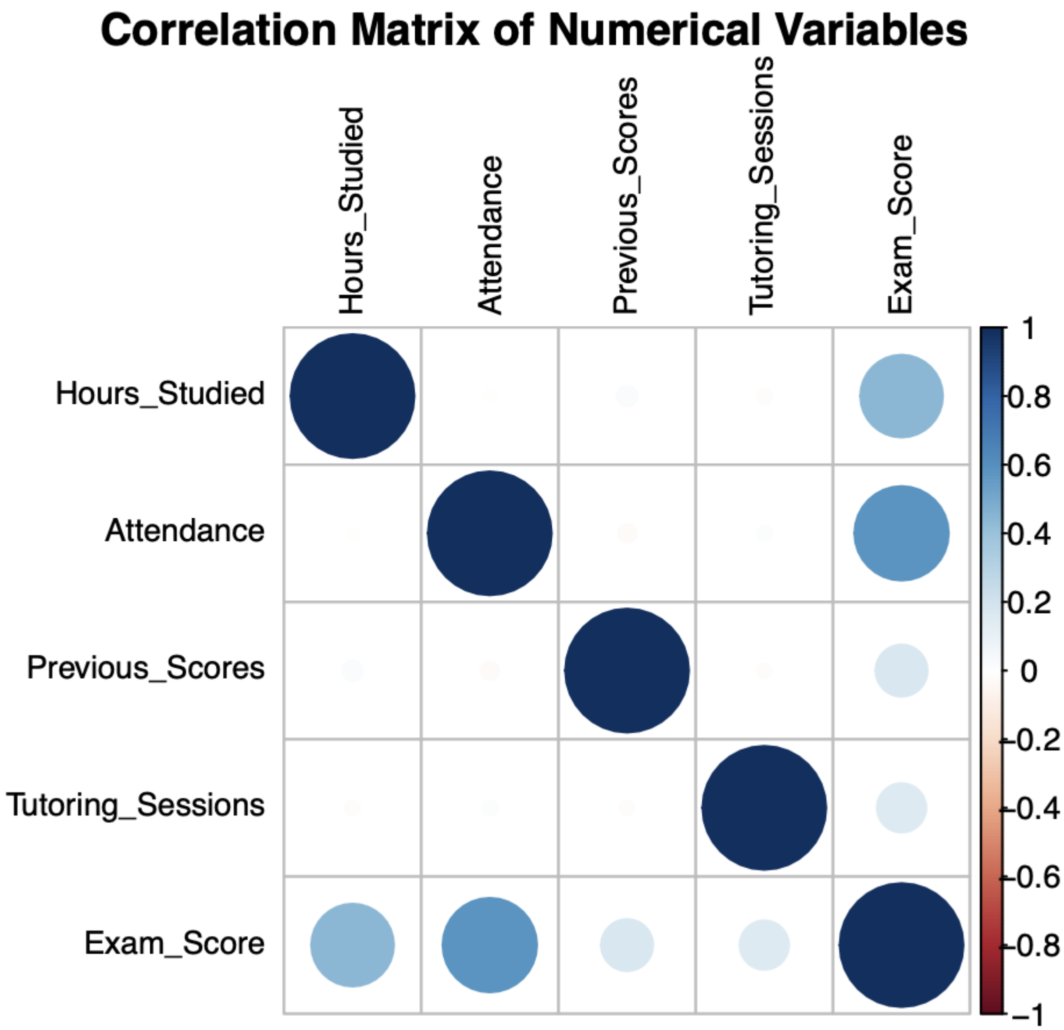


Figure 8: Of the numerical predictors, Attendance and Number of Hours show the strongest positive relationships with Exam Scores. There is very little multicollinearity due to the large sample size.

3.4 Observations

From our exploratory data analysis, several key patterns emerged:

- Students with higher `Attendance` and more `Hours_Studied` were strongly associated with achieving above-median exam scores, as seen in both histograms and scatter plots.
- `Access_to_Resources` played a critical role, with students having high resource availability being more likely to perform above the median, whereas those with low access consistently underperformed.
- `Parental_Education_Level` had a notable influence, with students whose parents achieved postgraduate education showing a higher likelihood of being in the "Above the Median" category.
- Previous academic performance (`Previous_Scores`) was a strong predictor, as students with higher prior scores were more likely to achieve above-median results.
- The interaction between `Hours_Studied` and `Attendance` highlighted that consistent attendance amplified the benefits of study time, emphasizing the combined importance of these behaviors.
- The dataset showed a relatively balanced distribution between students classified as "Above the Median" and "Below the Median," providing a robust basis for classification analysis.

These observations underscore the importance of behavioral factors, resource availability, and demographic influences in predicting academic success, setting the stage for the logistic regression modeling that follows.

3.5 Variable Selection

The dataset used in this study included a diverse set of predictors categorized into three groups: **Student Habits**, **Resource Availability**, and **Demographic/Background Influences**. These categories encompassed variables such as `Attendance`, `Hours Studied`, `Tutoring Sessions`, `Access to Resources`, and `Parental Involvement`.

An initial multiple linear regression (MLR) model was fitted using all 19 predictors in the dataset. The results showed that most variables had statistically significant coefficients, likely due to the large sample size (`n = 6607`). However, this model was not particularly informative in identifying the most impactful predictors due to issues of multicollinearity and the broad scope of significant variables.

To address these limitations, we utilized a random forest model to assess variable importance based on the %IncMSE metric. This approach provided a robust and interpretable ranking of predictors. From the random forest results, we identified the top nine predictors, which were subsequently included in the logistic regression models to ensure both predictive accuracy and interpretability.

The final set of nine predictors included:

Table 2: Variable Groups

Student Habits	**Resource Availability**	**Demographic/Background Influences**
Attendance	Tutoring_Sessions	Peer_Influence
Hours_Studied	Access_to_Resources	Family_Income
Previous_Scores	Parental_Involvement	Parental_Education_Level

By narrowing the scope of predictors to this optimized set, we ensured that our modeling process focused on the most influential factors while maintaining model interpretability. Figure 9 presents a visualization of the variable importance rankings derived from the random forest model.

4 Model Selection

4.1 Model Comparison

To find the best model to predict our response variable, we evaluated the performance of three distinct logistic regression models by employing different combinations of predictors. Through this approach,

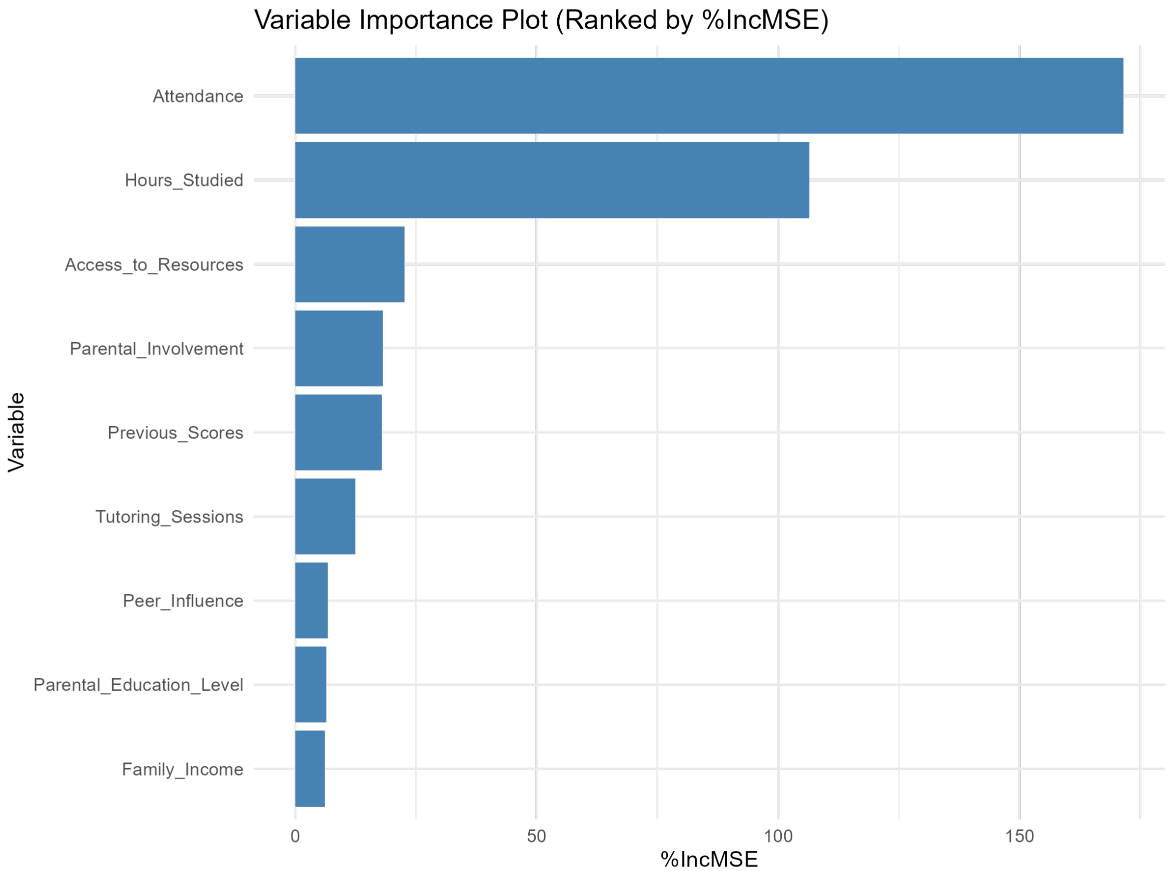


Figure 9: Variable importance rankings for the top nine predictors, based on the %IncMSE metric.

we aim to identify the most accurate and robust logitics regression model for predicting students' academic performance, specifically their final exam scores.

- For Model 1, we included any predictors that met our established threshold: %IncMSE greater or equal to 6. This resulted in the inclusion of 9 predictors. It had the best balance of accuracy (0.9) and AUC (0.97) with a strong ability to distinguish between performance levels. It is reasonably complex but still retains significant predictive power.
- Model 2 introduces less variables and a more standardized and established benchmark of %IncMSE greater or equal to 10, leaving the model with 6 predictors. It has a slightly lower accuracy and AUC score than Model 1 of 0.88 and 0.95, respectively. However, it is simpler with only the most influential predictors, making it more interpretable while maintaining strong results.
- Finally, Model 3 is the simplest and most interpretable as it incorporates only the top 2 predictors, `Attendance` and `Hours_Studied`. It has an accuracy of 0.84 and an AUC of 0.92. While Model 3 is less predictive, it is ideal for when ease of implementation and communication is prioritized.

4.2 Best Model: Model 1

Based on the accuracy and AUC score, our clear winner is Model 1. With an initial accuracy score of 0.9, the model correctly classifies 90% of students as above or below the median performance level. A near perfect AUC score of 0.97 indicates that the model has excellent discriminative ability to distinguish between high and low performers. Overall, a high accuracy and AUC indicates that we have selected a strong predictive model.

Lastly, the inclusion of multiple behavioral, environmental, and resource-related predictors ensures a robust understanding of factors influencing academic success.

5 K-Fold Cross-Validation

5.1 Introduction to K-Fold CV

To further validate our decision to use Model 1, we decided to perform K-fold Cross-Validation to evaluate how well our model predicts the final exam scores based on our predictors. The basic idea of this approach is to split the data into K equally sized divisions, or folds, and then train and test the model K times. Finally, we average the results across all folds. This method ensures the model generalizes well and helps prevent overfitting.

5.2 Results of K-Fold CV

The results of the 10-fold Cross-Validation (CV) for our logistic regression model, which predicts final exam performance, are summarized below. The model was evaluated on three key metrics: Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), Sensitivity (Sens), and Specificity (Spec). These metrics provide insight into the model's ability to generalize across unseen data.

Table 3: Performance Metrics from 10-Fold Cross-Validation

Metric	**Average Value**	**Interpretation**
ROC AUC	0.966	Excellent discriminative ability, indicating the model reliably distinguishes between "AboveMedian" and "BelowMedian" students.
Sensitivity (Sens)	0.911	High true positive rate, showing the model accurately identifies "AboveMedian" students.
Specificity (Spec)	0.884	Strong true negative rate, demonstrating good performance in recognizing "BelowMedian" students.

From the results in Table 3, the model achieved an average ROC AUC of 0.966, which indicates excellent discriminative performance. This metric reflects the model's ability to distinguish between the two classes ("AboveMedian" and "BelowMedian") across all folds. The sensitivity of 0.911 demonstrates the model's effectiveness in correctly identifying students predicted to perform above the median, while a specificity of 0.884 shows its ability to accurately identify those predicted to perform below the median.

The high and balanced sensitivity and specificity values suggest that the model generalizes well and performs robustly across different data splits. These metrics confirm that the predictors included in the model—such as `Attendance`, `Hours_Studied`, and `Access_to_Resources`—are reliable indicators of student performance.

The implications of these results are significant:

- The model demonstrates excellent generalizability, meaning it is likely to perform well on new, unseen data.
- The balance between sensitivity and specificity suggests it is effective for both identifying high-performing students and recognizing those who may need additional support.
- The high ROC AUC value confirms that the selected predictors provide meaningful insights into the factors driving academic success.

Overall, the results of the K-fold Cross-Validation validate the quality and robustness of our logistic regression model, establishing it as a reliable tool for predicting student performance and informing educational interventions.

6 Checking for Multicollinearity

To check for whether there is multicollinearity between the predictors, we calculated the Variance Inflation Factor (VIF) values. The standard threshold to prove that there isn't significant correlation between our predictors is to have an VIF value of less than 5. The closer the VIF value is to 1, the less correlated the predictors are to each other. We do not want to have any VIF values greater than 5 as that means that there is a significant and severe level of correlation and can cause issues when it comes to interpreting an independent variable's predictive power. As seen in **Figure 10**, all of our 9 predictors have VIF values of less than 5 and are very close to 1. Thus, we do not have to worry about multicollinearity issues and that our predictors are reliable.

GVIF_Results

Predictor	GVIF	Df	GVIF..1..2.Df..
Hours_Studied	1.00199500965896	1	1.00099700781719
Attendance	1.00293950450634	1	1.00146867375188
Parental_Involvement	1.00449773771952	2	1.00112254286125
Access_to_Resources	1.00450594643684	2	1.00112458813882
Tutoring_Sessions	1.0010679957019	1	1.00053385535018
Family_Income	1.00279344551581	2	1.00069763100583
Peer_Influence	1.00446207331172	2	1.00111365659982
Parental_Education_Level	1.00490497729476	3	1.00081583045635
Previous_Scores	1.00376112447518	1	1.00187879729795

Figure 10: Visualization of Generalized Variance Inflation Factor (GVIF) results, showcasing the minimal impact of collinearity among predictors.

7 Research Conclusions

7.1 Answers to Research Question 1

Research Question 1: Which individual factors from the dataset serve as the most significant predictors of academic performance, and how do they influence student outcomes?

The analysis of various individual factors reveals several key predictors that significantly influence academic performance. These factors, as identified in the dataset, include:

- `Attendance`: Consistent attendance plays a crucial role in academic success. Students who regularly attend classes are more exposed to course material and engage more effectively with the content, leading to improved academic outcomes. Our model suggests that for every percentage increase of classes attended, the student is 44% more likely to score above the median on their final exam.

- `Hours_Studied`: The amount of time dedicated to studying is a strong predictor of academic performance. For every extra hour spent studying, the student is 72% more likely to score above median on their final exam. This finding highlights the importance of time management and effective study habits in achieving academic success.

OddsRatio and 95% Confidence Interval

Term	OddsRatio	LowerCI_2.5%	UpperCI_97.5%	PValue
(Intercept)	0.00	0.00	0.00	<0.01
Attendance	1.44	1.41	1.48	<0.01
Hours_Studied	1.72	1.65	1.79	<0.01
Previous_Scores	1.10	1.09	1.11	<0.01
Tutoring_Sessions	2.63	2.34	2.96	<0.01
Peer_Influence_Neutral	2.12	1.54	2.94	<0.01
Peer_Influence_Positive	5.48	3.92	7.72	<0.01
Parental_Involvement_Low	0.02	0.02	0.04	<0.01
Parental_Involvement_Medium	0.17	0.12	0.22	<0.01
Parental_Education_Level_High_School	0.66	0.24	1.86	0.43
Parental_Education_Level_College	1.63	0.58	4.64	0.36
Parental_Education_Level_Postgraduate	3.53	1.24	10.19	0.02
Access_to_Resources_Low	0.02	0.01	0.03	<0.01
Access_to_Resources_Medium	0.14	0.10	0.19	<0.01
Family_Income_Low	0.17	0.12	0.23	<0.01
Family_Income_Medium	0.47	0.34	0.65	<0.01

Figure 11: This chart shows the Odds Ratio and its 95% Confidence Level for each predictor. We can see that whether a parent graduated from High School or College does not significantly impacts the academic performance of their child.

- `Access_to_Resources`: The level of access to resources, such as learning materials, internet access, and extracurricular support, is another significant predictor. Students with medium access to these resources are 86% less likely to achieve above-median performance whereas students with low access are 98% less likely. This suggests that limited access to essential academic resources can create barriers to success.
- `Tutoring_Sessions`: Engagement in tutoring sessions has quite a significantly positive impact on student performance. Students who receive personalized guidance through tutoring are 2.63 times more likely to perform above the median, indicating that additional academic support plays a vital role in improving student outcomes.
- `Parental_Involvement`: Higher levels of parental engagement, whether through participation in school events or regular communication with teachers, contribute to improved student outcomes. Our model suggests that students whose parents are less involved in their academics are 98% and 83% less likely to perform above median, for categories Low and Medium. This underscores the importance of a supportive home environment in fostering academic success.
- `Parental_Education_Level`: The educational background of parents is also a significant factor in student performance. Students whose parents have attained postgraduate education are 3.53 times more likely to perform well academically, while students whose parents have only completed high school or college tend to score lower. However, whether the parents completed high school or college show no significant impact on the student's ability to score above median. This finding reflects the influence of parental educational attainment on children's academic expectations and support.

- `Family_Income`: Economic factors, particularly family income, are strongly correlated with academic performance. Students with low and medium family income are 83% and 52% less likely to perform above median compared to those from higher income families. This suggests that financial stability provides access to more educational opportunities and resources, which in turn enhances academic success.

- `Previous_Scores`: How well a student previously performed on their exams is, without a surprise, a powerful indicator of how well they will perform in the future. Our model suggests that for every point increase in previous exam score, the student is 10% more likely to perform above the median in their final exam. This suggests that consistent effort and performance is the key to success.

- `Peer_Influence`: Finally, a student's external environment and who they choose to surround themselves with have the most impact on their academic performance. Students with peers who have a neutral impact on their academic performance are 2.12 times more likely to perform above median whereas students who have peers with positive influences on their academic performance are 5.48 times more likely to perform above median compared to those with negative-impacting peers. This finding confirms the statement that "you are the average of the five people you surround yourself with."

In conclusion, a combination of academic engagement, resource availability, and socioeconomic factors significantly influences student outcomes. The findings underscore the importance of addressing these factors to support academic success and close achievement gaps among students with varying levels of access to resources and familial support.

7.2 Answers to Research Question 2

Research Question 2: To what extent do the three predictor groups (student habits, available resources, and demographic influences, respectively) contribute to the predictive power of our models in predicting academic success?

- `Demographic Influences (Peer_Influence, Family_Income, Parental_Education_Level)`: While , initially, this group seemed to be less influential than the other two, our model actually suggests that environmental influences heavily contributes to the prediction of academic success. Family income is positively correlated with performance, as students from higher-income families often have more access to educational opportunities and resources. Family income also impacts the parents' education level, and we are 95% confident that a student is 1.24 to 10.19 times more likely to perform above the median on their final exam if at least one parent holds a postgraduate degree. Peer influence is another important factor, reflecting the social environment's impact on a student's motivation and behavior. We are 95% confident that students who surround themselves with positive-impacting peers are 3.92 to 7.72 times more likely to academically perform above median than those with negative-impacting peers.

- `Available Resources (Tutoring_Sessions, Access_to_Resources, Parental_Involvement)`: This group emphasizes the impact of external support systems on academic success. Access to resources, such as learning materials and technological tools, provides students with the necessary tools to excel. We are 95% confident that students with low access to these resources are 97% to 99% less likely to perform as well as those with high accessibility. Tutoring sessions also have a direct impact by offering personalized guidance, helping students who may need extra help. Parental involvement adds another layer of support, as we are 95% confident that students whose parents show low levels of involvement are 96% to 98% less likely to perform above median compared to students with very involved parents. The negative effects from the lack of these resources show that external factors, beyond the student's own habits, play a very crucial role in determining academic outcomes.

- `Study Habits (Attendance, Hours_Studied, Previous_Scores)`: Unlike the previous groups, the range of impact from these factors aren't as drastic. However, a student's consistency and persistence is a significant factor in their overall academic performance. Regular attendance and more hours studied are strongly correlated with higher academic performance, underscoring the importance of consistent engagement with the course material. Students who commit more time to studying are more likely to perform better, highlighting the value of effective study habits. Additionally, previous academic performance is a key predictor, suggesting that students who have done well in the past are likely to continue performing well. These habits are central to a student's ability to succeed academically.

In conclusion, external factors such as demographic/background and available resources have the strongest influence on academic success, with students' habits contributing important context. These findings underscore the importance of dismantling a system that prioritizes standardized tests system such as the SAT and ACT as it can amplify educational inequalities rather than provide a fair measure of student ability and potential.

8 Limitations and Considerations for Future Research

Throughout this process, our group noticed various limitations we believe are important to point out:

- **Limited Background Information:** Because our dataset was sourced from Kaggle.com, there lacks detailed information regarding where these observations came from and, thus, we are unable to validate the collection methods. This may introduce bias or limit the generalizability of findings that we may not be aware of. Additionally, there are no clear threshold as to what differentiates categorical classifications, such as "High Family Income" versus "Medium Family Income." For future research, establishing clear thresholds could significantly enhance our findings and help us quantify and understand the magnitude of the impact of categorical variables, such as parents' involvement and resource accessibility, on students' academic performance.
- **Interaction Effects:** Demographic influences, ranked least important, may have indirect effects on resource availability and educational quality, potentially obscuring their true impact on academic performance. Conducting an analysis and including the impact of how certain variables interact with one another in our model should be included in future research.
- **Simplified Modeling Assumptions:** Logistic regression assumes linear relationships between predictors and the outcome, potentially oversimplifying complex interactions; future studies could explore nonlinear methods like random forests or neural networks.

Despite these limitations, our study provides valuable insights into factors influencing academic performance and highlights the importance of predictors such as `Peer_Influence`, `Tutoring_Sessions`, and `Access_to_Resources`. To build upon these findings, future research could address the limitations mentioned by sourcing datasets with well-documented collection methods, examining the interaction effects among predictors in greater detail, and applying more advanced modeling techniques. Incorporating these approaches would strengthen the understanding of how various factors impact student success and potentially lead to more robust predictive models that could inform targeted interventions and educational policies.

References

Data Source (Insights into Student Performance and Contributing Factors; Kaggle):
https://www.kaggle.com/datasets/lainguyn123/student-performance-factors.