# An Exploratory Data Analysis and Prediction Model for Student Performance

Shahaba Alam

shahaba.alam@student.aiu.edu.my

Data Mining

Albukhary International University

# TABLE OF CONTENTS

# Abstract

This project explores the application of data mining techniques to predict student academic performance, focusing primarily on the data mining aspects, including extensive exploratory data analysis (EDA) and data preprocessing. Using a dataset of over 30,000 students sourced from Kaggle, we examined demographic, behavioral, and parental background factors to identify key predictors of success in Math, Reading, and Writing. Through rigorous preprocessing—handling outliers, encoding categorical variables, and normalizing data—we ensured high data quality for effective analysis. The EDA revealed important patterns and correlations, which informed model selection and helped us understand the impact of various features. Our model-building experiments demonstrated that individual academic scores are the strongest predictors, achieving an R-squared of 1.0, while additional demographic factors provided limited improvement. This emphasis on EDA and preprocessing underscores the value of comprehensive data mining for performance prediction, guiding targeted educational interventions based on academic indicators.

# 1.0 Introduction

This project applies data mining techniques to predict student performance by analyzing their demographics and behavior. Our goal is to uncover key factors influencing academic success in Math, Reading, and Writing, using a dataset sourced from Kaggle (Des, 2023). This dataset includes detailed records of 30,641 students, encompassing demographic details, behavioral traits, and academic performance.

## 1.1 Objective

Our objective is to apply data mining techniques to predict student performance based on their demographics and behavior.

## 1.2 Dataset Overview

The dataset used in this project is sourced from "Kaggle", (des, 2023), one of the largest public platforms for machine learning and data science competitions. The dataset contains a detailed record of student demographics, behaviors, and academic performance.

**Dataset Description:**
The dataset contains '30,641 records' of students, with multiple attributes that are relevant for analyzing student performance. These attributes are divided into several categories:

1. **Demographic Information:**
   - Gender: Categorical (Male, Female).
   - Ethnic Group: Categorical (Various categories like group A, group B, etc.).
   - Parent Education Level: Categorical (Levels such as High School, Bachelor's, Master's).
   - Parent Marital Status: Categorical (e.g., married, single).
2. **Behavioral Features:**
   - Lunch Type: Categorical (Standard, Free/Reduced) indicating socioeconomic status.
   - Test Preparation: Categorical (Completed, None), indicating if the student took test preparation courses.
   - Practice Sports: Categorical (Regularly, Sometimes, Never), indicating extracurricular activity.
   - Is First Child: Categorical (Yes, No), whether the student is the firstborn.
   - Number of Siblings: Numerical, indicating the number of siblings.
   - Transport Means: Categorical, showing how students commute to school.
   - Weekly Study Hours: Categorical, representing how much time students spend studying.
3. **Academic Performance:**
   - Math Score: Numerical, representing the student's math score.
   - Reading Score: Numerical, representing the student's reading score.
   - Writing Score: Numerical, representing the student's writing score.

The data is 'structured', with both **categorical** and **numerical** attributes. Most of the categorical features, such as **gender, ethnic group, and parental education, require encoding for use in machine learning models,** while the numerical columns like **'MathScore', 'ReadingScore', and 'WritingScore'** can be directly utilized after preprocessing steps such as handling missing values and outliers.

## 1.3 Problem Statement

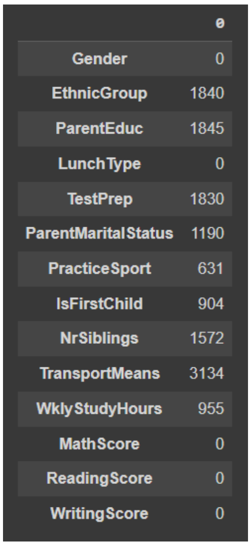
In this project we want to predict 'student academic performance' based on their demographic, behavioral, and parental background information. By analyzing this dataset, we aim to uncover key factors that influence academic success in three subject areas: Math, Reading, and Writing.

The problem is framed as a 'regression problem', where the target variable is a composite score representing the student's overall performance in these three subjects. This will allow us to predict student outcomes and identify the most significant predictors of academic success.

# 2.0 Data Preprocessing

## 2.1 Data Cleaning

In this section, we focus on cleaning the dataset, handling outliers and missing values, transforming features for analysis, and conducting exploratory data analysis (EDA). Below is the detailed breakdown of the steps taken for data preprocessing.



| | 0 |
|---|---|
| Gender | 0 |
| EthnicGroup | 1840 |
| ParentEduc | 1845 |
| LunchType | 0 |
| TestPrep | 1830 |
| ParentMaritalStatus | 1190 |
| PracticeSport | 631 |
| IsFirstChild | 904 |
| NrSiblings | 1572 |
| TransportMeans | 3134 |
| WklyStudyHours | 955 |
| MathScore | 0 |
| ReadingScore | 0 |
| WritingScore | 0 |

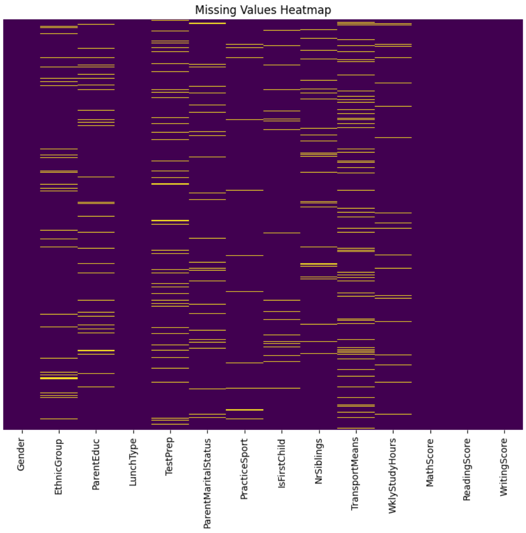
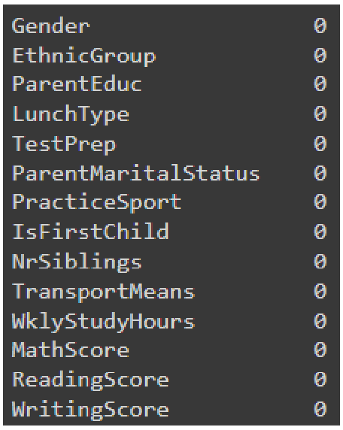


Figure 1.0: Missing Values in our dataset

From figure 1.0, we can see that columns with missing values are **EthnicGroup, ParentEduc, TestPrep, ParentMaritalStatus, PracticeSport, IsFirstChild, TransportMeans, WklyStudyHours and NrSiblings.**

Now for handling the missing values within the dataset, For categorical columns such as **EthnicGroup, ParentEduc, TestPrep, ParentMaritalStatus, PracticeSport, IsFirstChild, TransportMeans, and WklyStudyHours**, missing values were filled with the most frequent value (mode) in each respective column. This method ensures that the imputed values align with the existing distribution of the data.

For the numerical column **NrSiblings**, missing values were replaced with the **median**, as this is a robust statistic that minimizes the influence of outliers. This combination of techniques preserves the integrity of the dataset and allows for accurate analysis moving forward. We can see the result from figure 2.0, now we have no missing values in our dataset.



```
Gender                 0
EthnicGroup            0
ParentEduc             0
LunchType              0
TestPrep               0
ParentMaritalStatus    0
PracticeSport          0
IsFirstChild           0
NrSiblings             0
TransportMeans         0
WklyStudyHours         0
MathScore              0
ReadingScore           0
WritingScore           0
```

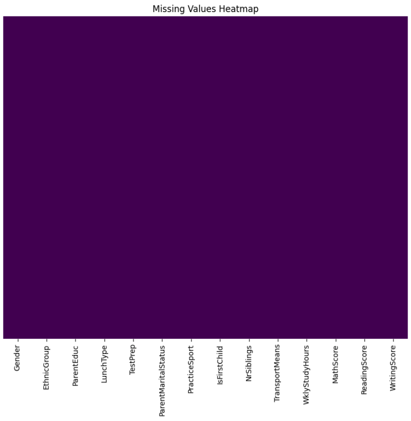


Figure 2.0: After handling Missing Values

### 2.1.1 Handling Outliers

After handling the missing values now we need to manage the outliers in our numerical columns which are NrSibilings, MathScore, ReadingScore and WritingScore.

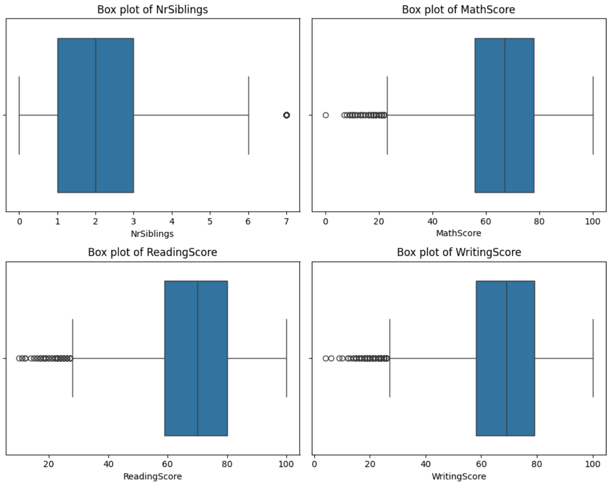


Figure 2.1: BoxPlot before handling outliers

From figure 2.1, we can see that in out numerical columns we have so many outliers, which can distort the results of data analysis and modeling, so we used the 'Interquartile Range (IQR)' method to identify and remove outliers in the numerical columns NrSiblings, MathScore, ReadingScore, and WritingScore. From figure 2.2 we can clearly observe that our numerical columns have no outliers present.

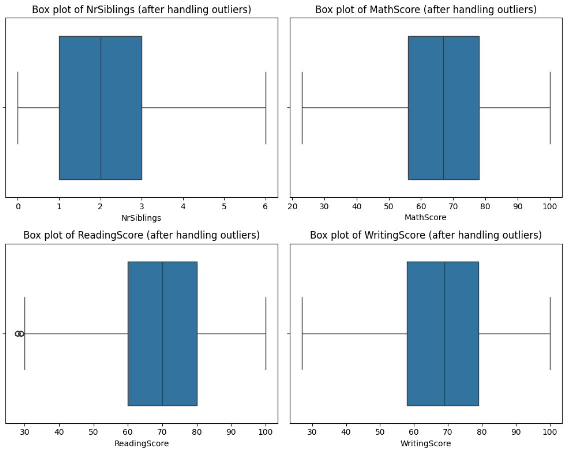
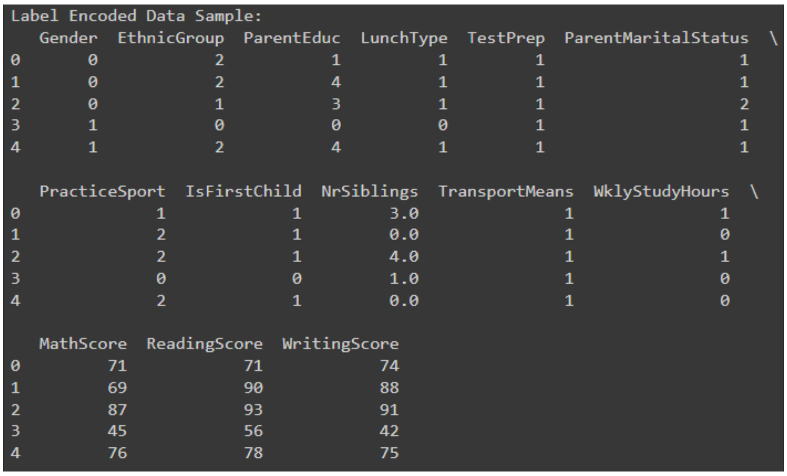


Figure 2.2: Boxplot after handling outliers

## 2.2 Data Transformation

### 2.2.1 Lebel Encoding

To prepare the dataset for machine learning algorithms, label encoding was applied to all categorical columns. A copy of the original dataset was created to preserve its integrity, and the categorical columns (Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport, IsFirstChild, TransportMeans, and WklyStudyHours) were encoded using LabelEncoder from scikit-learn can be seen in figure 2.3. Label encoding converts the categories into numerical values, allowing the model to interpret categorical data effectively. This transformation ensures the data is in a suitable format for modeling while maintaining the original categorical relationships.



```
Label Encoded Data Sample:
   Gender  EthnicGroup  ParentEduc  LunchType  TestPrep  ParentMaritalStatus  \
0       0            2           1          1         1                    1
1       0            2           4          1         1                    1
2       0            1           3          1         1                    2
3       1            0           0          0         1                    1
4       1            2           4          1         1                    1

   PracticeSport  IsFirstChild  NrSiblings  TransportMeans  WklyStudyHours  \
0              1             1         3.0               1               1
1              2             1         0.0               1               0
2              2             1         4.0               1               1
3              0             0         1.0               1               0
4              2             1         0.0               1               0

   MathScore  ReadingScore  WritingScore
0         71            71            74
1         69            90            88
2         87            93            91
3         45            56            42
4         76            78            75
```

Figure 2.3: Data After Label Encoding

Label encoding was chosen for its efficiency in converting categorical variables into numerical values required by machine learning algorithms. It works well for features with few unique values, like Gender and LunchType, and maintains the distinct categories without adding unnecessary complexity. This method is particularly suitable for tree-based models, which can handle the encoded data without being sensitive to the numerical ordering of the categories.

## 2.2.3 Data Normalization

Three normalization techniques Min-Max Scaling, Z-Score Scaling, and Decimal Scaling were applied to NrSiblings, MathScore, ReadingScore, and WritingScore. Min-Max Scaling scales values to a range of [0, 1], Z-Score Scaling standardizes the data to have a mean of 0 and a standard deviation of 1, and Decimal Scaling adjusts values based on their maximum magnitude. Histograms were used to compare the effects of these techniques on the data distributions.

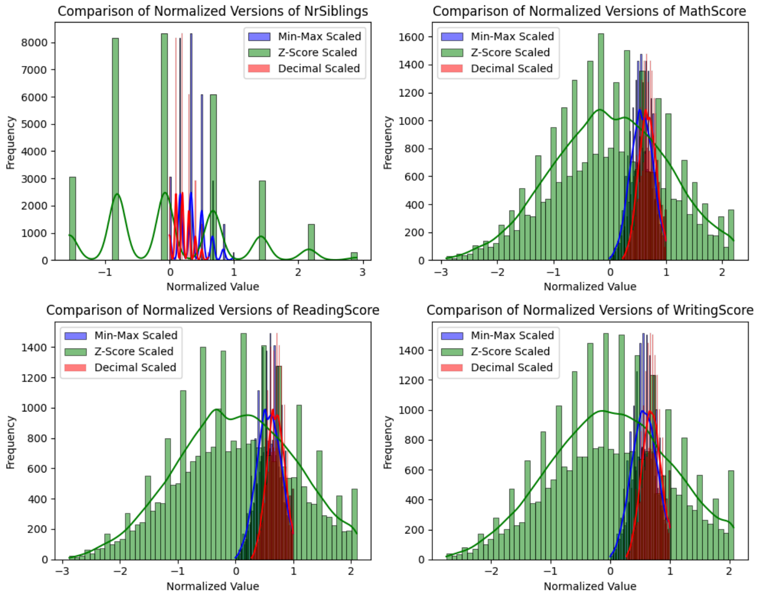


Figure 2.4: Data Normalization Techniques

From figure 2.4 it can be concluded that  For NrSiblings, Min-Max scaling could be more appropriate due to its discrete nature and skewness and for MathScore, ReadingScore, and WritingScore, Z-Score Scaling remains the best choice, given their near-normal distributions.

## 2.3 Exploratory Data Analysis (EDA)

The goal of this Exploratory Data Analysis (EDA) is to understand the dataset and uncover insights related to student academic performance. The dataset includes various socio-demographic and academic preparation factors that influence student performance in Math, Reading, and Writing scores. The focus is on analyzing the distribution of categorical variables, understanding the impact of these factors on academic performance, and exploring relationships between different features.

### 2.3.1 Univariate Analysis: Categorical Data Frequency

Univariate analysis of categorical data frequency is performed to understand the distribution and prevalence of different categories within a single variable. By analyzing the frequency of each category, we can identify patterns, trends, and potential imbalances in the data, which helps in gaining insights into the composition of the dataset (Smith, 2022). This analysis is crucial for informing subsequent data preprocessing steps, such as encoding categorical variables for machine learning models, and for ensuring that any biases or skewed distributions are addressed to improve model accuracy and fairness (Johnson & Lee, 2021). To begin the EDA, we performed a frequency analysis on categorical features. This involves examining the distribution of each categorical variable to understand its unique values and their counts shown in figure 2.5.

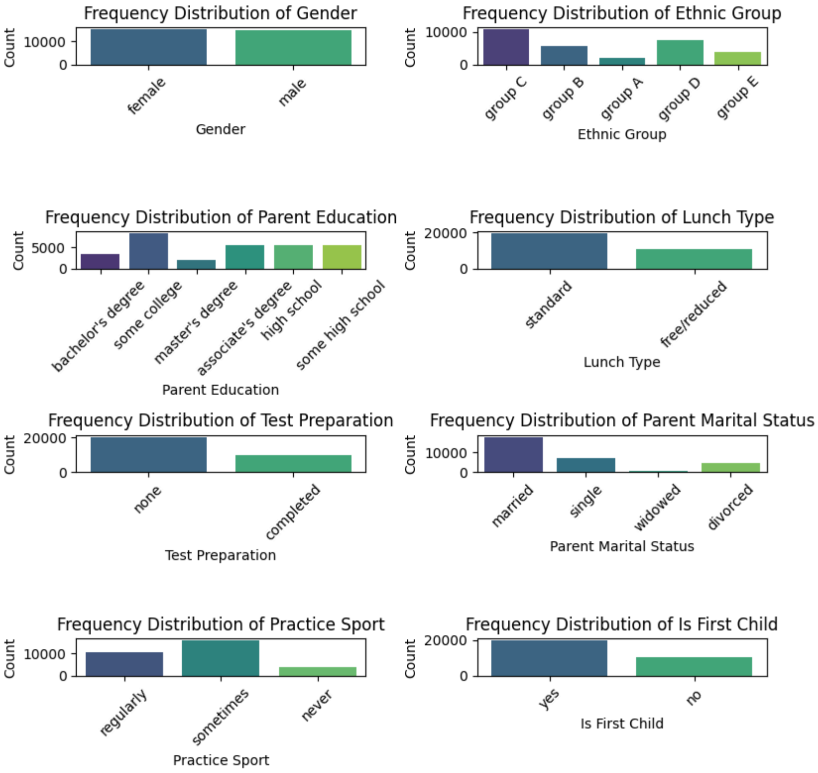


Figure 2.5: Categorical Data Frequency

### 2.3.2 Correlation Matrix

Figure 2.6 illustrates the correlation matrix for the numerical attributes in the dataset, highlighting relationships between the number of siblings and academic performance (Math, Reading, and Writing scores).

The number of siblings (NrSiblings) shows no significant correlation with any academic scores, indicating it does not influence a student's performance in these areas. In contrast, the academic scores are strongly correlated with each other, particularly between Reading and Writing scores (0.95), suggesting that students who perform well in one are likely to excel in the other. Additionally, Math scores are positively correlated with both Reading (0.81) and Writing (0.80) scores, indicating a commonality in the underlying skills required for success across these subjects.

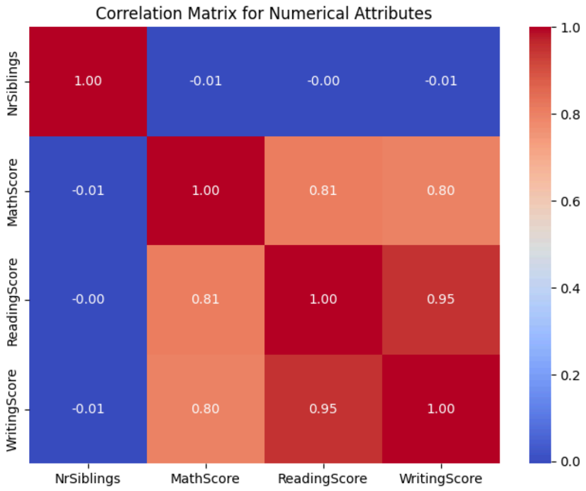


Figure 2.6: Correlation Matrix for Numerical Attributes

Figure 2.7 presents a correlation matrix showing the relationships between various student attributes, such as academic scores, demographic factors, and other background variables. Strong positive correlations are observed between MathScore, ReadingScore, and WritingScore, with coefficients ranging from 0.80 to 0.95, indicating that students who perform well in one subject tend to excel in others. In contrast, attributes like NrSiblings, Gender, EthnicGroup, ParentEduc, and TransportMeans display weak or negligible correlations with scores, typically falling between -0.10 and 0.10. LunchType shows a moderate positive correlation with academic performance, especially in math and reading, while TestPrep reveals a weak negative correlation with writing scores. Other attributes, such as ParentMaritalStatus, PracticeSport, IsFirstChild, and WeeklyStudyHours, exhibit minimal influence on academic performance, with weak correlations across the board. The color gradient, ranging from dark red (strong positive correlation) to dark blue (strong negative correlation), illustrates the varying strengths of these relationships.

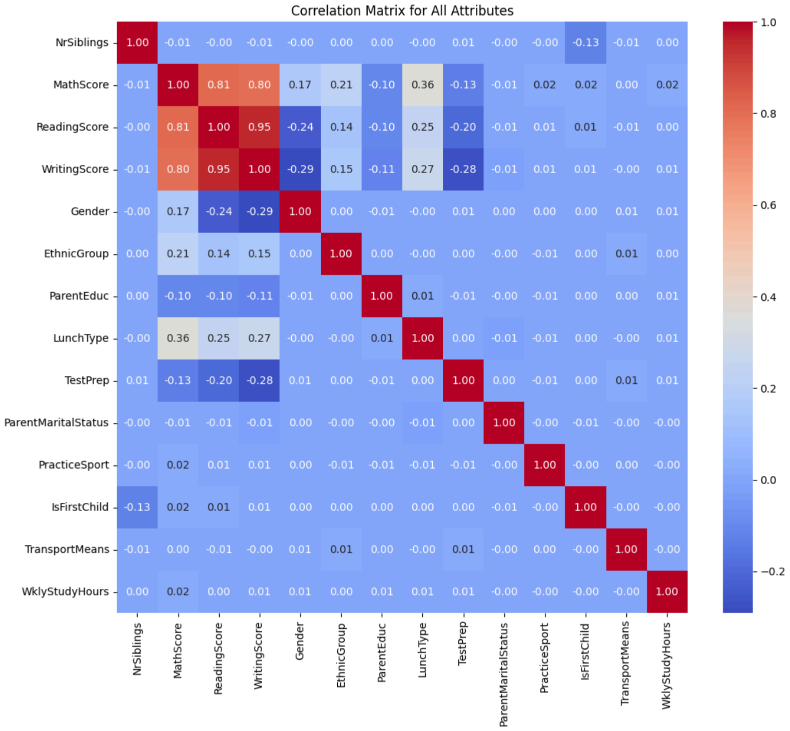


Figure 2.7: Correlation Matrix for All Attributes

### 2.3.3 Scatter Plot

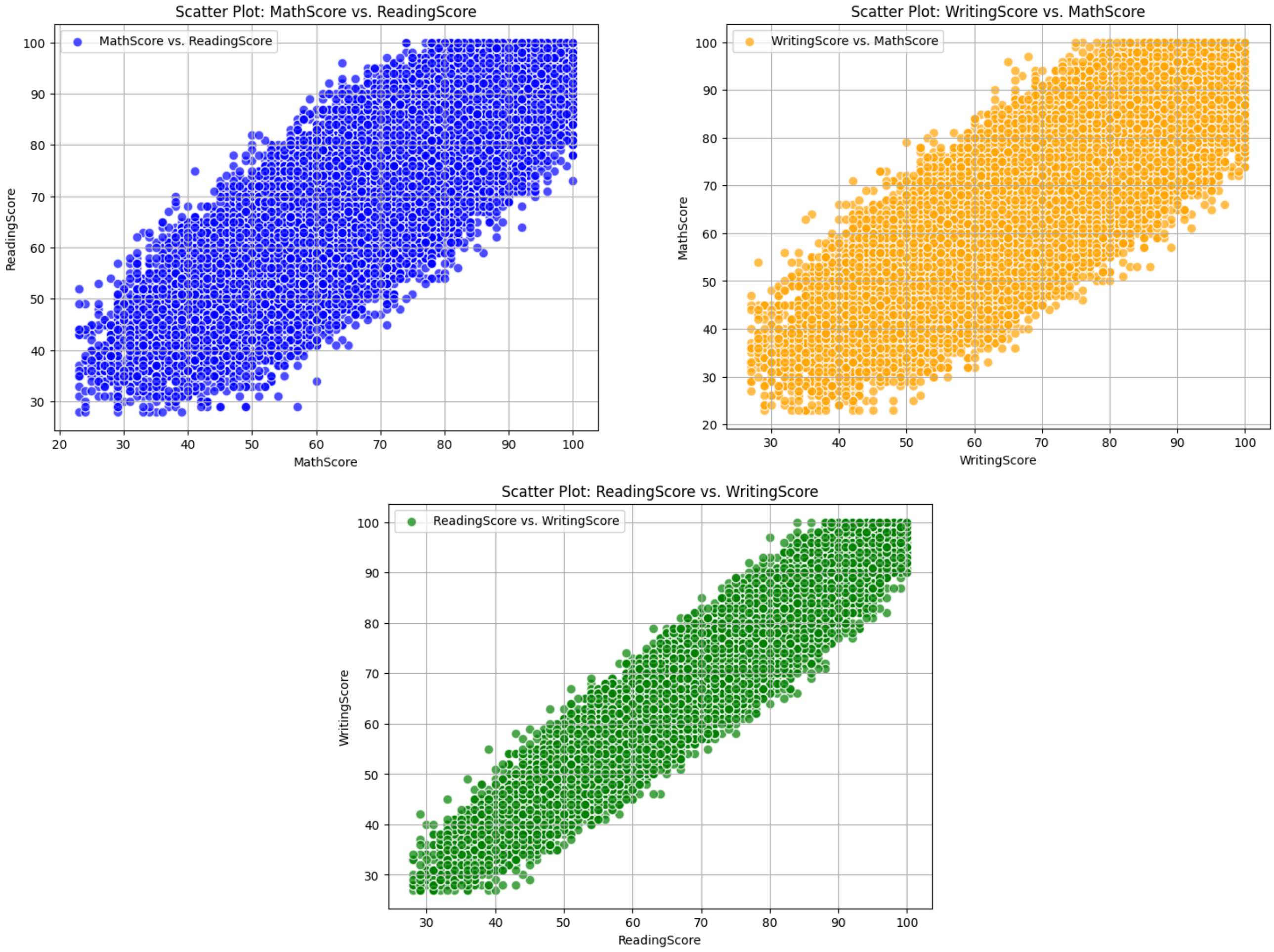


Figure 2.8: Scatter Plot for attribute with higher correlation

Correlation and scatter plots help to identify relationships between variables, guiding feature selection and model design. The correlation matrix, as seen in Figure 2.7, quantifies the linear relationships between variables, indicating how strongly one variable is related to another. This helps in identifying important predictive features and detecting multicollinearity, which can affect model performance. For instance, strong correlations between MathScore, ReadingScore, and WritingScore suggest these variables are closely linked, whereas weak correlations, like those with ParentEduc or Gender, indicate limited predictive power. Scatter plots, also represented by Figure 2.8, visually highlight the nature of relationships between variables, revealing patterns, trends, and outliers. These plots allow modelers to see whether relationships are linear or nonlinear, influencing the choice of model type. Together, correlation and scatter plots provide crucial insights into variable relationships, helping to refine feature selection and inform model-building strategies.

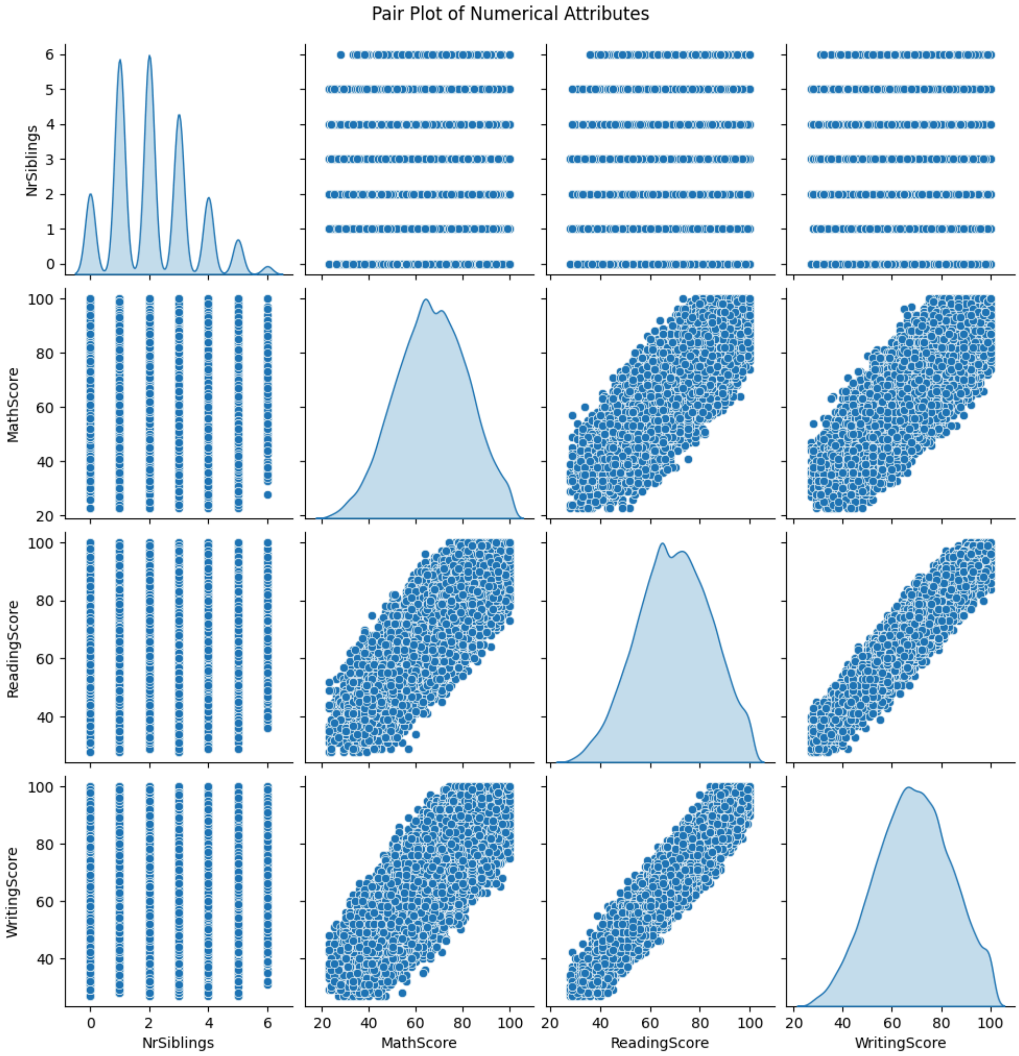


Figure 2.9: Pair Plot of Numerical Attributes

Figure 2.9 presents a pair plot, which displays the relationships between numerical attributes, specifically NrSiblings, MathScore, ReadingScore, and WritingScore. The diagonal elements show the distribution of each variable through histograms, while the scatter plots in the off-diagonal elements illustrate the pairwise relationships between these attributes. The pair plot is a crucial visualization tool for understanding the data, as it allows the identification of correlations and patterns between variables. For example, the strong linear relationships between MathScore, ReadingScore, and WritingScore are clearly visible, reinforcing the strong positive correlations observed in earlier analyses like the correlation matrix in Figure 2.7. Additionally, the distribution of NrSiblings shows a distinct pattern, but it has a weak relationship with the scores, as evident from the scattered points. This type of visualization aids in feature selection and helps identify potential multicollinearity, outliers, or nonlinear relationships, all of which are key considerations in building predictive models.

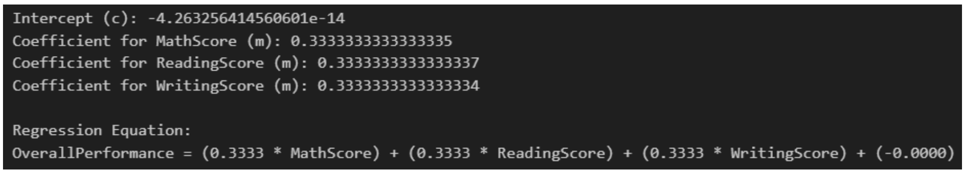
# 3.0 Model Building

## 3.1 Model Selection

For predicting student academic performance, we chose regression models due to their ability to estimate continuous outcomes, such as a composite score representing overall performance in Math, Reading, and Writing. Regression analysis is well-suited to uncover the relationships between academic success and various predictors, including demographic, behavioral, and parental background factors. Linear Regression, in particular, is appropriate for this problem as it allows us to evaluate the impact of multiple predictors on the continuous target variable. Additionally, the model's interpretability helps identify which factors significantly influence student performance. By examining these relationships, we can gain insights into key drivers of academic success and tailor educational strategies accordingly.

## 3.2 Experiment Design

In Experiment 1, the objective is to evaluate the relationship between individual subject scores and overall student performance. To achieve this, a composite score representing overall performance is created by averaging the MathScore, ReadingScore, and WritingScore. This new OverallPerformance score serves as the dependent variable. The independent variables, or predictors, are the individual scores in MathScore, ReadingScore, and WritingScore. The analysis will focus on understanding how well these individual scores can predict the overall performance by examining the correlation and predictive power of these variables.



```
Intercept (c): -4.263256414560601e-14
Coefficient for MathScore (m): 0.3333333333333335
Coefficient for ReadingScore (m): 0.3333333333333337
Coefficient for WritingScore (m): 0.3333333333333334

Regression Equation:
OverallPerformance = (0.3333 * MathScore) + (0.3333 * ReadingScore) + (0.3333 * WritingScore) + (-0.0000)
```

Figure 3.0: Regression Model Equation

The regression equation used to predict OverallPerformance shows that each subject score (Math, Reading, and Writing) equally contributes to the overall academic performance, with each score having a coefficient of 0.3333. This indicates that for every one-unit increase in any of the subject scores, the OverallPerformance increases by approximately 0.3333 units. The intercept is negligible, implying that the baseline OverallPerformance is virtually unaffected when all scores are zero. This model assumes a balanced contribution of each subject score to the overall performance, reflecting their equal importance in the prediction.

Experiment 2 aims to investigate how various demographic, behavioral, and parental background factors influence the overall student performance. A composite score is computed similarly by averaging the MathScore, ReadingScore, and WritingScore, resulting in the OverallPerformance score. This score is used as the dependent variable, while a range of independent variables is considered, including demographic (e.g., Gender, EthnicGroup), behavioral (e.g., PracticeSport, WklyStudyHours), and parental background factors (e.g., ParentEduc, ParentMaritalStatus). The analysis will explore how these factors impact the composite score by employing statistical or machine learning techniques to model and interpret the relationships between the predictors and overall performance.

### 3.2.1 Train-Test Split

The dataset was divided into training and testing subsets using an 80-20 ratio, where 80% of the data was allocated for training the model, and 20% was reserved for evaluation. This approach ensures that the model is trained on a large portion of the data, which helps in capturing the underlying patterns and relationships. The remaining 20% serves as a separate, unseen subset that is crucial for testing the model's ability to generalize to new, unseen data. This separation is essential for assessing the model's performance and robustness, providing an unbiased evaluation of its predictive accuracy.

### 3.2.2 Parameter Tuning

In the case of Linear Regression, there are no hyperparameters to tune in its basic form. However, if model performance needs enhancement or if there are concerns about overfitting, advanced techniques such as Ridge or Lasso regularization can be employed. These methods introduce penalties to the regression coefficients, helping to manage multicollinearity and prevent overfitting. Although parameter tuning is not a primary focus for basic Linear Regression, considering regularization techniques could be beneficial for improving model performance and ensuring more reliable predictions.

## 3.3 Model Evaluation

### 3.3.1 Experiment 1

Table 1.0 Evaluation Metrics

| | |
|---|---|
| **Mean Absolute Error** | 0.002204185301750163 |
| **Mean Squared Error** | 7.347284339166695e-06 |
| **Mean Squared Error** | 0.0027105874527796913 |
| **R-squared** | 1.0 |

The model exhibits excellent performance with an R-squared value of 1.0, indicating perfect prediction accuracy, while the very low Mean Absolute Error and Mean Squared Error values suggest minimal error and high precision in predicting student performance.

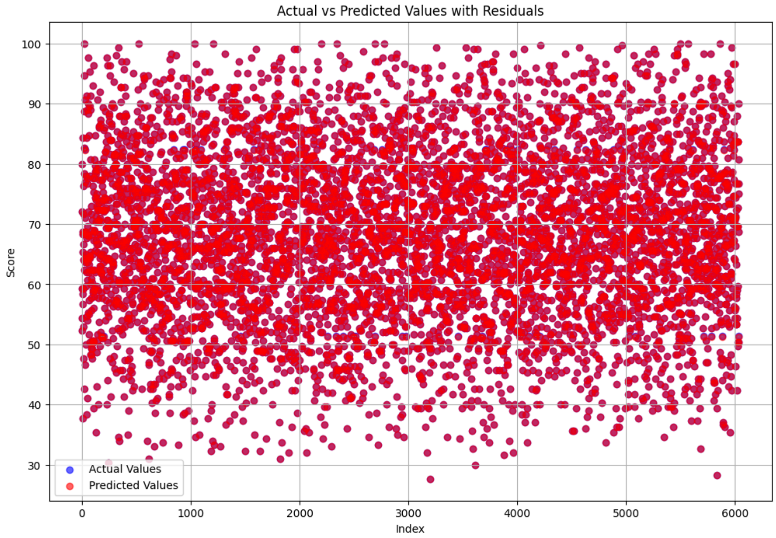


Figure 3.0: Actual vs Predicted Values with Residuals

Figure 3.0 compares actual and predicted values, showing a scatter of residuals that indicates variability in the model's accuracy, with no apparent bias but room for improvement in prediction precision.

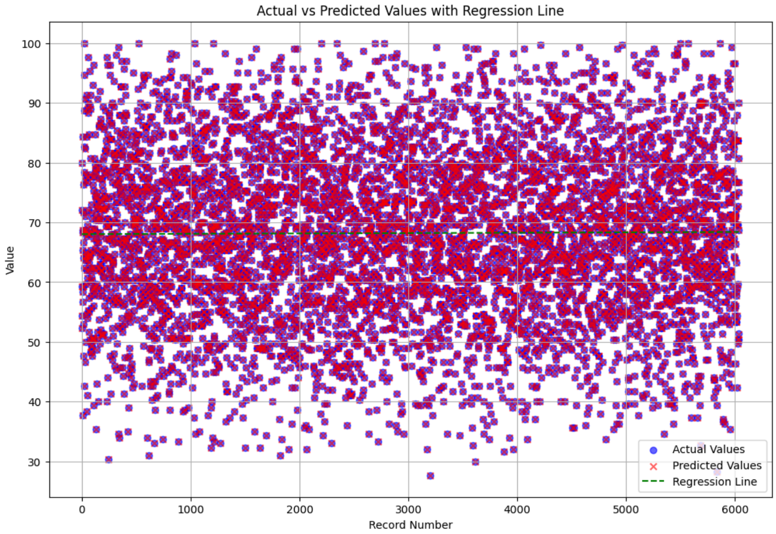


Figure 3.1 Actual vs Predicted value with regression line

Figure 3.1 illustrates the comparison between actual and predicted values using a regression line, where the data points show substantial scatter, indicating high variance, with the predicted values (red "X") closely following the regression line (green dashed line). This plot is used to visually assess the performance of a regression model by comparing how well the predicted values match the actual values. It helps identify patterns, errors, and the model's goodness of fit. The regression line provides a reference for the overall trend of the predictions.

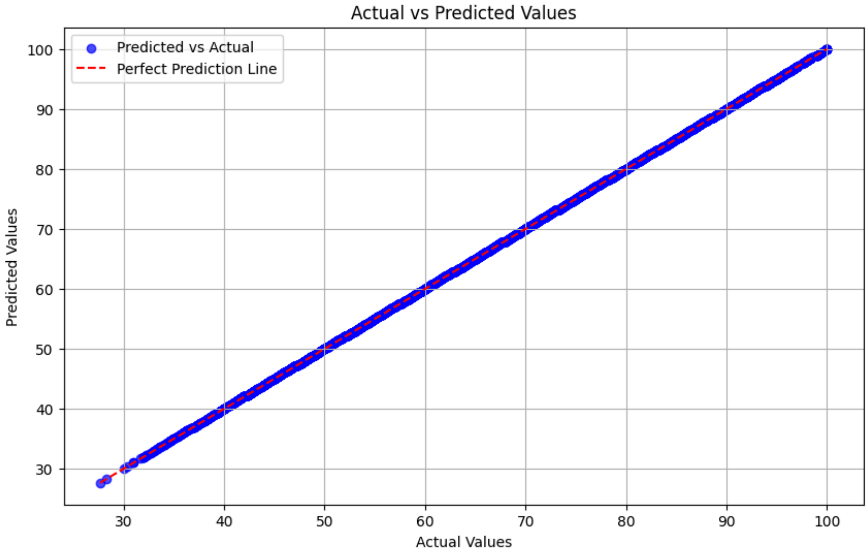


Figure 3.2 Actual vs Predicted Values

Figure 3.3 shows a scatter plot of predicted values versus actual values with a red dashed line representing the perfect prediction line, where predicted and actual values would be identical. The predicted values (blue circles) closely follow this line, indicating a well-fitting model. This plot is used to assess how accurately the regression model predicts actual values. A close alignment of points along the perfect prediction line suggests that the model has minimal errors and performs well, providing a quick visual indicator of model accuracy.

After looking at the result the first question should come why this model performs so well. So to answer that, **why?** Is quite simple by looking at the correlation matrix we can clearly see that those three attributes 'MathScore', 'ReadingScore', 'WritingScore' share a strong correlation with one another. So when using those as independent variables our model performs so well.

### 3.3.2 Experiment 2

As experiment performed so well, let's change few things now we will predict student performance based on other attributes including 'Gender', 'EthnicGroup', 'ParentEduc', 'LunchType', 'TestPrep', 'ParentMaritalStatus', 'PracticeSport', 'IsFirstChild', 'TransportMeans', 'WklyStudyHours', 'NrSiblings', mean our independent variable are those and dependent is OverallPeroformace calculated as the mean of three test score.

Table 2.0 Evaluation Metrics

| | |
|---|---|
| **Mean Absolute Error** | 10.251092465085351 |
| **Mean Squared Error** | 157.2167279203846 |
| **Mean Squared Error** | 12.538609489109414 |
| **R-squared** | 0.19809300600432345 |

As the correlation between those attributes are really low our model performance is relatively poor, with a low R-squared value of 0.20 indicating that only about 20% of the variance in student performance is explained by the model, while the high Mean Absolute Error (10.25), Mean Squared Error (157.22), and Root Mean Squared Error (12.54) suggest substantial prediction errors.

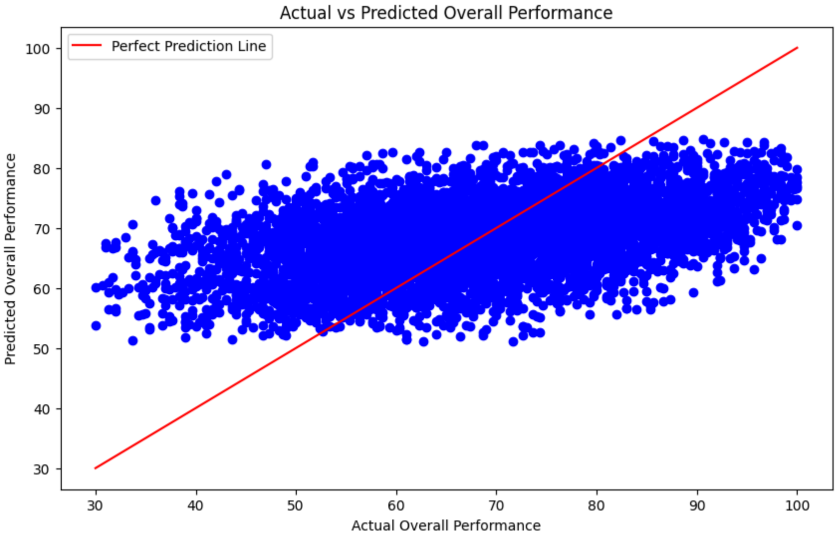


Figure 3.3 Actual vs Predicted overall Performance

Figure 3.3 shows a noticeable spread between predicted and actual performance values, indicating the model's predictions deviate from perfect accuracy with significant variability.

As expected due to weak correlation between those attributes using those as independent variables significantly impacted the performance of our model.

## 3.4 Insights from Model

The models provided valuable insights into predicting student performance based on different sets of predictors. In Experiment 1, the high R-squared value of 1.0 and minimal error metrics indicate that using only the individual subject scores (MathScore, ReadingScore, and WritingScore) for prediction results in near-perfect accuracy. This outcome is expected given the strong correlation among these scores, demonstrating that they are highly predictive of overall performance.

In contrast, Experiment 2, which incorporated a range of demographic, behavioral, and parental background factors, yielded a much lower R-squared value of 0.20 and higher error metrics. This suggests that these additional attributes have a limited impact on predicting overall student performance, likely due to weak correlations with the performance scores. These results highlight the challenge of incorporating diverse predictors and underscore the importance of focusing on the most relevant features for accurate performance prediction.

**Real-World Application:** In practical terms, the findings from Experiment 1 suggest that focusing on individual academic scores is highly effective for performance prediction, which could inform educational strategies and interventions. However, the results from Experiment 2 emphasize the need for further refinement in integrating various student attributes, potentially leading to more nuanced models that combine academic scores with relevant demographic and behavioral factors for a more holistic understanding of student performance.

# 4.0 Recommendations

Based on the results of the data mining process, the following actionable recommendations can be made:

**Focus on Core Academic Support:**

- Based on the analysis from Experiments 1 and 2, the recommendation is to prioritize core academic scores Math, Reading, and Writing when developing performance models, as these scores are highly predictive of overall student performance. While secondary attributes like parental education and student motivation were less impactful in Experiment 2, they should still be considered to provide additional context. Combining these key academic indicators with relevant secondary factors can enhance the model's accuracy and offer a more comprehensive understanding of student performance. Regular evaluation and refinement of the model with new data will help maintain its effectiveness.

**Strengths of the Analysis**

- **High Predictive Power:** The R-squared value of 1.0 in Experiment 1 indicates that the model effectively explains the variance in student performance based solely on Math, Reading, and Writing scores. This suggests that these scores are excellent predictors of overall performance.
- **Effective Preprocessing:** The thorough data cleaning process, including handling missing values and outliers, ensured that the dataset was well-prepared for analysis. This preprocessing minimized noise and improved the accuracy and reliability of the model's predictions.
- **Visual Analysis:** The use of visual tools such as scatter plots, box plots, and correlation matrices provided clear insights into the relationships between variables and the performance of the model. These visualizations helped in understanding the model's effectiveness and identifying key patterns in the data.

**Limitations of the Analysis**

- **Limited Features:** While Math, Reading, and Writing scores were excellent predictors, the inclusion of additional features in Experiment 2, such as parental education level and student motivation, did not improve the model's performance. The low R-squared value of 0.20 suggests that these additional features have limited predictive power for overall student performance, possibly due to weak correlations with the performance scores.
- **Linear Assumptions:** The linear regression model assumes a linear relationship between the input features and the target variable. However, educational factors and student performance may have more complex, non-linear relationships. Exploring other models, such as decision trees or random forests, could capture these non-linear relationships more effectively.
- **Dataset Scope:** The dataset includes a limited range of demographic and performance-related variables. Expanding the dataset to include additional factors, or collecting data over multiple years, could provide a more comprehensive view of student performance and improve the accuracy of predictive models.

# 5.0 Conclusion

This project demonstrated the effectiveness of using individual academic scores to predict overall student performance, achieving high accuracy with minimal errors. Experiment 1 confirmed that MathScore, ReadingScore, and WritingScore are strong predictors of academic success due to their high correlation. Conversely, Experiment 2 revealed that incorporating various demographic, behavioral, and parental background factors resulted in a less effective model, highlighting the challenges in predicting performance with diverse attributes. These findings suggest a need to refine the inclusion of additional predictors and consider advanced modeling techniques to improve performance prediction in educational settings.

# 6.0 References

1. des. (2023). Students Exam Scores: Extended Dataset. Kaggle.com. https://www.kaggle.com/datasets/desalegngeb/students-exam-scores

2. Johnson, M., & Lee, A. (2021). Data preprocessing and analysis for machine learning. Springer.

3. Smith, J. (2022). Introduction to statistical analysis. Oxford University Press.

4. freeCodeCamp. (2020, June 29). How to Build and Train Linear and Logistic Regression ML Models in Python. FreeCodeCamp.org; freeCodeCamp.org. https://www.freecodecamp.org/news/how-to-build-and-train-linear-and-logistic-regression-ml-models-in-python/