IS 4007 - Statistics in Practice II

STUDY ON IDENTIFYING FACTORS ASSOCIATED WITH STUDENT ACADEMIC PERFORMANCE

GITHMIN DULSARA s16344

ABSTRACT

This study explores the key factors influencing student academic performance using a dataset of 6,607 records obtained from Kaggle. The objective was to identify significant predictors of exam scores and build a reliable predictive model to assist educators in understanding student outcomes.

Exploratory analysis and multiple linear regression were used alongside machine learning models (Decision Tree, Random Forest and XGBoost). The linear regression model explained 66.8% of the variance in scores with significant predictors including parental involvement, motivation, hours studied and access to resources. Learning disabilities had a negative impact. Gender and school type were found to be statistically insignificant.

Among the predictive models, Random Forest performed best (RMSE = 2.12, $R^2 = 0.672$), slightly outperforming XGBoost, while the Decision Tree model showed poor accuracy. The results underscore the importance of academic support, motivation and resource accessibility. These findings provide valuable insights for educators and policymakers to design targeted interventions aimed at improving student performance.

CONTENTS

FIGURES AND TABLES

INTRODUCTION

Background of the study

Academic performance is a key indicator of a student's learning and future prospects. Various factors (both academic and non-academic) play crucial roles in determining student success. The increased availability of educational data allows researchers to apply statistical and machine learning techniques to analyze performance outcomes. Identifying these influential factors can help educators to make informed decisions to improve student learning environments. With rising competition and mental health concerns, understanding how factors like motivation, resources and lifestyle affect exam scores is more important than ever.

Objectives

- To identify key factors associated with student exam scores.
- To build a predictive model to predict student performance.
- To provide insights for educators.

Significance of the study

- Helps educators and institutions to identify the most critical factors affecting student performance.
- Supports early intervention strategies for students at academic risk.
- Provides a predictive tool that can forecast student outcomes based on key variables.
- Offers actionable insights to enhance curriculum planning, teaching strategies and resource allocation.
- Contributes to the growing field of educational data mining and evidence-based decision-making in schools and universities.

LITERATURE REVIEW

Academic performance prediction has been a widely studied area with the growing availability of educational datasets and machine learning tools. Numerous studies have demonstrated the usefulness of statistical and machine learning techniques in identifying factors that contribute to student success.

According to S. Kotsiantis (2004), regression and classification models can predict student performance with reasonable accuracy when trained on features such as study time, attendance and past academic records. Similarly, C. Romero and S. Ventura (2010) found that behavioral variables, especially hours studied and class attendance, are among the strongest predictors of exam scores.

Parental involvement and student motivation have also emerged as significant factors. Fan and Chen (2001) showed that active parental engagement is positively correlated with academic achievement. Motivation plays a vital role in influencing how students manage time and effort (Pintrich & De Groot, 1990). In contrast, students with learning disabilities or poor access to educational resources tend to perform worse (Reiser & Dempsey, 2012). These findings showed the importance of both psychological and environmental variables in academic performance.

More recent studies have explored the application of machine learning models such as decision trees, random forests and gradient boosting algorithms. Zhang and others (2019) showed that the models like Random Forest and XGBoost often outperform simple models in terms of prediction accuracy. However, interpretability remains a trade-off. For example, while decision trees are easy to understand, they are prone to overfitting, especially in smaller datasets.

From a methodological standpoint, most studies employed multiple linear regression as a baseline due to its simplicity and interpretability. Feature selection based on correlation and ANOVA tests is commonly used to retain relevant predictors. Handling missing data is also emphasized, as imputation can significantly affect model reliability (Little & Rubin, 2002).

In summary, the literature validates the inclusion of variables such as hours studied, attendance, previous scores, parental involvement, motivation and resource availability in predicting student performance. It also supports the use of ensemble models for enhanced accuracy while recommending linear models for their clarity. These insights have guided the current study in selecting variables, choosing modeling techniques, and evaluating model performance.

THEORY AND METHODOLOGY

Theoretical Framework

The theoretical foundation of this study suggests that academic performance is influenced by a combination of factors. This aligns with Bronfenbrenner's Ecological Systems Theory. It shows the interplay between individual, familial, institutional and societal influences on human development including education.

Additionally, the study is guided by the Constructivist Learning Theory. It suggests that students actively construct knowledge through experiences, motivation and engagement. Variables such as hours studied, parental involvement, motivation level and learning environment play crucial roles in this construction process.

The data-driven aspect of the study is supported by principles of educational data mining (EDM) and learning analytics, which aim to discover patterns and predictive insights from educational data. These theories support the application of statistical modeling and machine learning techniques to predict outcomes like exam scores and inform instructional decisions.

Methodological Approach

This study is having both descriptive and predictive analytical techniques to explore factors associated with student performance. The methodology includes:

1. Data Preprocessing

To prepare the data for analysis:

- Missing values were imputed using the mode for categorical variables.
- Categorical encoding was applied:
    - *Ordinal encoding* for ordered categories.
    - *One-hot encoding* for nominal categories.
- Outlier handling: Values like exam scores above 100 were identified as data entry errors.

2. Descriptive Statistics

Descriptive statistics were used to summarize the data and explore distributions of key variables. This includes:

- Frequencies and proportions for categorical variables.

- Measures of central tendency and dispersion for numerical variables (mean, median, standard deviation).
- Visualizations such as box plots and scatter plots to identify trends and outliers.

3. Exploratory Data Analysis (EDA)

EDA was conducted to explore relationships between variables and the target outcome (Exam_Score). This included:

- Univariate analysis to examine individual distributions.
- Bivariate analysis using correlation plots, scatter plots and box plots to assess associations.

4. Feature Selection

Feature selection was based on:

- Correlation with the target variable.
- Results from ANOVA tests for categorical predictors.
- Domain knowledge from the literature review.

5. Predictive Modeling

To build the exam score prediction model, both statistical and machine learning methods were applied:

Multiple Linear Regression (MLR)

- Used as a baseline model due to its interpretability.
- Assumptions checked: linearity, normality, multicollinearity, homoscedasticity.

Decision Tree Regression

- Captures nonlinear relationships and variable interactions.
- Easy to visualize but prone to overfitting.

Random Forest Regression

- An ensemble learning method that averages multiple decision trees.
- Offers improved accuracy and robustness over individual trees.

XGBoost (Extreme Gradient Boosting)

- Advanced ensemble method known for high predictive performance.
- Includes regularization to reduce overfitting.

6. Model Evaluation

Model performance was evaluated using:

- Root Mean Square Error (RMSE) – to measure prediction error.
- R-squared (R^2) – to evaluate explained variance.

Tools and Software

- R programming language for data cleaning, visualization and modeling.
- Libraries such as tidyverse, caret, randomForest, and xgboost.
- Data visualizations created using ggplot2.

DATA

Source: Kaggle - Student Performance Factors Dataset Student Performance Factors
No of records: 6607

Variables and Descriptions

Table below presents the variables used in the analysis along with their corresponding descriptions for better understanding of the dataset and the context of the study.

Variable	Description
Hours_Studied	Number of hours spent studying per week
Attendance	Percentage of classes attended
Parental_Involvement	Level of parental involvement in the student's education (Low, Medium, High)
Access_to_Resourses	Availability of educational resources (Low,Medium,High)
Extracurricular_Activities	Participation in extracurricular activities (Yes,No)
Sleep_Hours	Average number of hours of sleep per night
Previous_Scores	Scores from previous exams
Motivation_Level	Student’s level of motivation (Low,Medium,High)
Internet_Access	Availability of internet access (Yes,No)
Tutoring_Sessions	Number of tutoring sessions attended per month
Family_Income	Family income level (Low,Medium,High)
Teacher_Quality	Quality of the teachers (Low,Medium,High)
School_Type	Type of school attended (Public,Private)
Peer_Influence	Influence of peers on academic performance (Positive,Neutral,Negative)
Physical_Activity	Average number of hours of physical activity per week
Learning_Disabilities	Presence of learning disabilities (Yes,No)
Parental_Education_Level	Highest education level of parents (High School,College,Postgraduate)
Distance_From_Home	Distance from home to school (Near,Moderate,Far)
Gender	Gender of the student (Male,Female)
Exam_Score	Final exam score

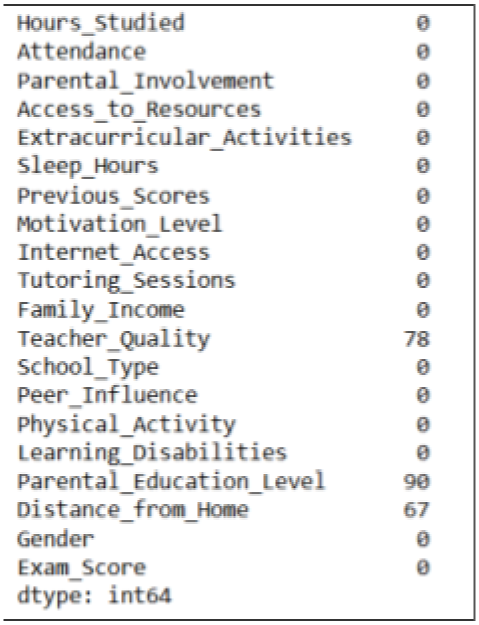
Table 1: Description of Variables

Exam_Score was selected as the target variable for this analysis, representing the academic performance outcome that the study aims to understand and predict.

Data Pre-Processing Techniques

Several preprocessing steps were applied to ensure data quality and suitability for analysis.

1. **Handling Missing values**



```
Hours_Studied                 0
Attendance                    0
Parental_Involvement          0
Access_to_Resources           0
Extracurricular_Activities    0
Sleep_Hours                   0
Previous_Scores               0
Motivation_Level              0
Internet_Access               0
Tutoring_Sessions             0
Family_Income                 0
Teacher_Quality              78
School_Type                   0
Peer_Influence                0
Physical_Activity             0
Learning_Disabilities         0
Parental_Education_Level     90
Distance_from_Home           67
Gender                        0
Exam_Score                    0
dtype: int64
```

Figure 1: Missing Values

An initial inspection of the dataset showed that three variables contained missing values:

- Teacher_Quality – 78 missing values
- Parental_Education_Level – 90 missing values
- Distance_from_Home – 67 missing values

To ensure the completeness of the dataset and maintain the robustness of the analysis, appropriate imputation methods were applied. For the categorical variables, Teacher _Quality, Parental_Education_Level and Distance_from_Home, missing values were imputed using the **mode**, assuming that the most frequent category reflects a reasonable estimate.

2. **Categorical Encoding**

In this study, categorical variables were encoded to facilitate data analysis and predictive modeling. The encoding approach was chosen based on the type of categorical variable:

1. Ordinal Variables

Several variables in the dataset represent ordered categories, where the levels have a meaningful ranking. These include:

- Parental Education Level
- Parental Involvement
- Access to Resources
- Motivation Level
- Family Income
- Teacher Quality
- Peer Influence

- Distance from Home

These variables were encoded using an ordinal approach where higher numerical values correspond to higher levels of the respective factor (better education level, higher motivation, etc.).

2. Nominal Variables

Nominal variables in the dataset do not possess any intrinsic order. These include:

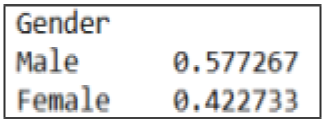
- Gender
- School Type
- Internet Access
- Learning Disabilities
- Extracurricular Activities

These were transformed using one-hot encoding, where each category was converted into a binary column. To avoid redundancy and multicollinearity, one category from each variable was omitted (The first category was dropped).

This preprocessing ensured that all variables were in a suitable numerical format for further analysis and model building.

EXPLORATORY DATA ANALYSIS

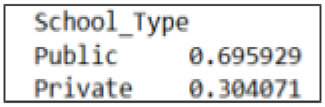
The dataset includes information on 6607 students, covering various factors related to academic performance.



```
Gender
Male      0.577267
Female    0.422733
```

Figure 2: Proportion of male and female

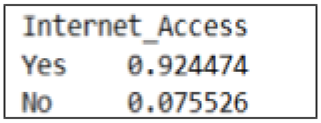
In the dataset, approximately 57.7% of the students are male, while 42.3% are female. This indicates a higher representation of male students compared to female students in the sample.



```
School_Type
Public     0.695929
Private    0.304071
```

Figure 3: Proportion of Private and Public schools

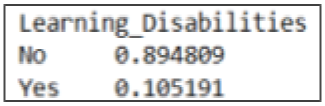
Approximately 69.6% of the students attend public schools, while 30.4% attend private schools. The majority of students in the dataset are enrolled in public educational institutions.



```
Internet_Access
Yes    0.924474
No     0.075526
```

Figure 4 : Proportion of Internet Access

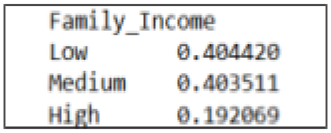
The majority of individuals (92.45%) reported having internet access, while only 7.55% indicated they do not. This high rate of internet connectivity suggests that digital platforms and online resources are likely accessible to most of the population represented in the dataset.



```
Learning_Disabilities
No     0.894809
Yes    0.105191
```

Figure 5 : Proportion of Learning Disabilities

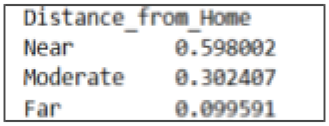
89.48% of individuals reported not having a learning disability, whereas 10.52% reported having one. Although the majority do not experience learning disabilities, the presence of over 10% affected individuals is noteworthy and highlights the importance of inclusive practices.



```
Family_Income
Low       0.404420
Medium    0.403511
High      0.192069
```

Figure 6 : Proportion of Family Income

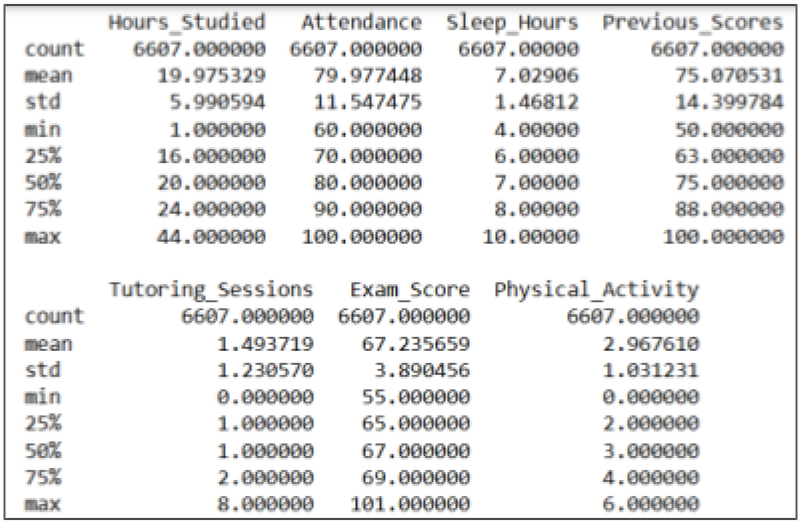
The majority of individuals come from low and medium income families, with 40.44% in the low income category and 40.35% in the medium income category. Only 19.21% of individuals belong to high-income families. This suggests a predominantly lower to middle socioeconomic background within the dataset.



```
Distance_from_Home
Near        0.598002
Moderate    0.302407
Far         0.099591
```

Figure 7 : Proportion of Distance from Home

A majority of individuals (59.80%) live near their college, while 30.24% live at a moderate distance and only 9.96% reside far away. This suggests that most individuals have relatively convenient access to their daily destinations, which may positively influence attendance. However, the smaller group living far from home may face challenges such as longer commute times, transportation costs.



```
       Hours_Studied   Attendance  Sleep_Hours  Previous_Scores
count    6607.000000  6607.000000   6607.00000      6607.000000
mean       19.975329    79.977448      7.02906        75.070531
std         5.990594    11.547475      1.46812        14.399784
min         1.000000    60.000000      4.00000        50.000000
25%        16.000000    70.000000      6.00000        63.000000
50%        20.000000    80.000000      7.00000        75.000000
75%        24.000000    90.000000      8.00000        88.000000
max        44.000000   100.000000     10.00000       100.000000

       Tutoring_Sessions   Exam_Score  Physical_Activity
count        6607.000000  6607.000000        6607.000000
mean            1.493719    67.235659           2.967610
std             1.230570     3.890456           1.031231
min             0.000000    55.000000           0.000000
25%             1.000000    65.000000           2.000000
50%             1.000000    67.000000           3.000000
75%             2.000000    69.000000           4.000000
max             8.000000   101.000000           6.000000
```

Figure 8: Summary of Numerical Variables

Hours Studied: Students studied an average of 20 hours per week, ranging from 1 to 44 hours. The distribution is slightly right-skewed, with 75% of students studying 24 hours or less.

Attendance: The average attendance rate is around 80%. Most students have attendance between 70% and 90%.

Sleep Hours: Students sleep about 7 hours per night on average. Most reported 6 to 8 hours of sleep, which is within a healthy range.

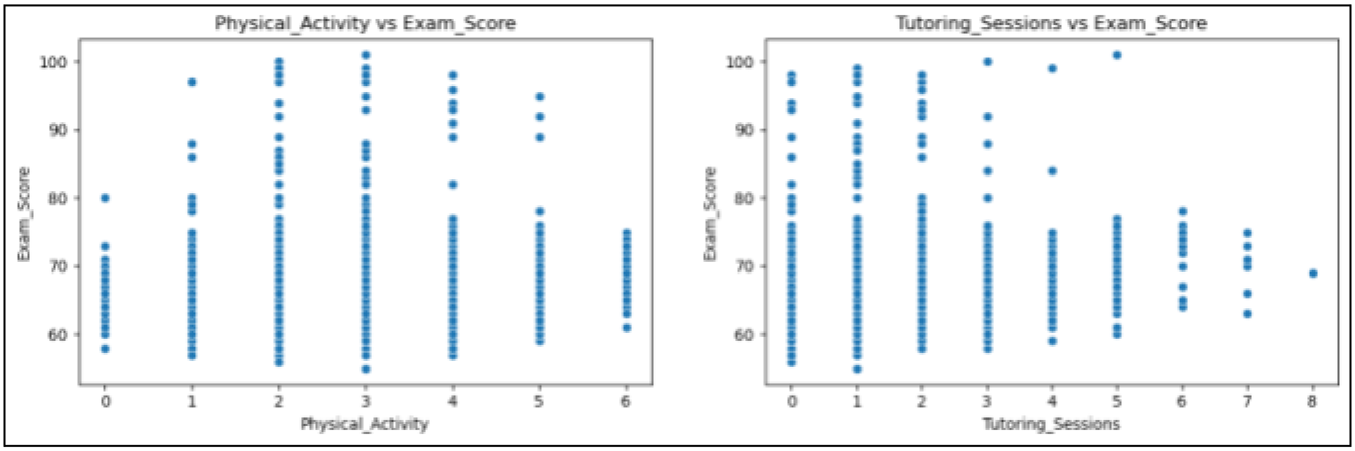
Previous Scores: The average previous score is 75, with a minimum of 50 and a maximum of 100, indicating a fairly well-performing group.

Tutoring Sessions: On average, students attend 1.5 tutoring sessions, with most attending between 1 and 2 sessions per week.

Exam Score: The mean exam score is around 67, with a slight spread (SD = 3.9). The maximum score slightly exceeds 100. This suggests a data entry artifact.

Physical Activity: Students reported engaging in an average of 3 sessions of physical activity, with the distribution ranging from 0 to 6 sessions per week.

Bivariate analysis was performed using Exam_Score as the response variable.



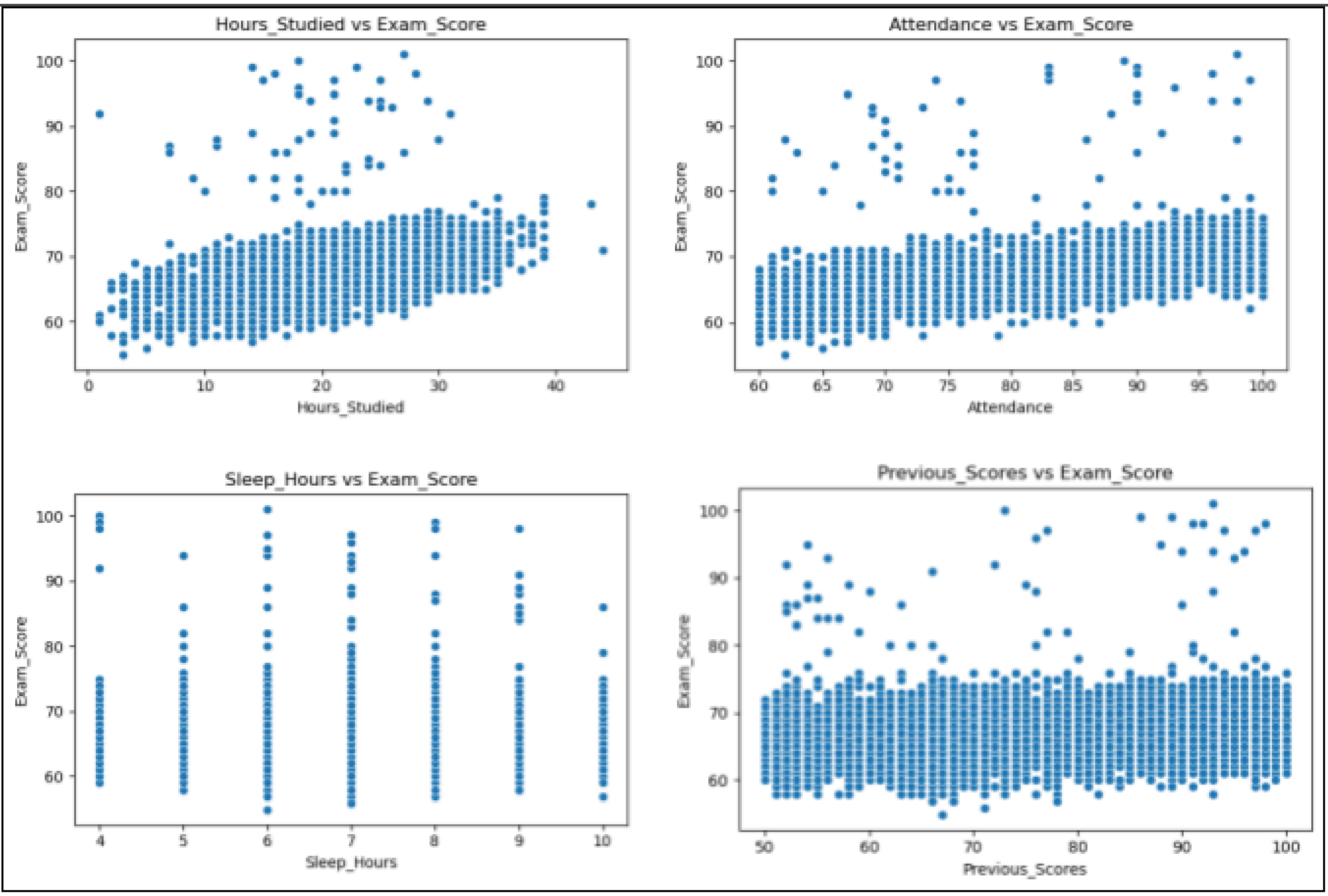


Figure 9 : Scatter plots showing relationship between numerical variables and exam scores

The above figure contains six scatter plots, each showing the relationship between different variables and exam scores. Here's a breakdown of what they suggest,

- Physical Activity vs Exam Score: It looks like there's no strong correlation. Exam scores remain fairly scattered regardless of physical activity levels.
- Tutoring Sessions vs Exam Score: There seems to be a slight upward trend, suggesting that students who attend more tutoring sessions may achieve higher exam scores.
- Hours Studied vs Exam Score: This one appears to show a clear positive correlation. More hours studied generally lead to higher scores.
- Attendance vs Exam Score: Higher attendance seems linked to better exam scores.
- Sleep Hours vs Exam Score: The relationship is likely more complex. But it seems to indicate that both too little and too much sleep might negatively impact exam performance.
- Previous Scores vs Exam Score: Students who had higher previous scores tend to maintain high performance in exams.

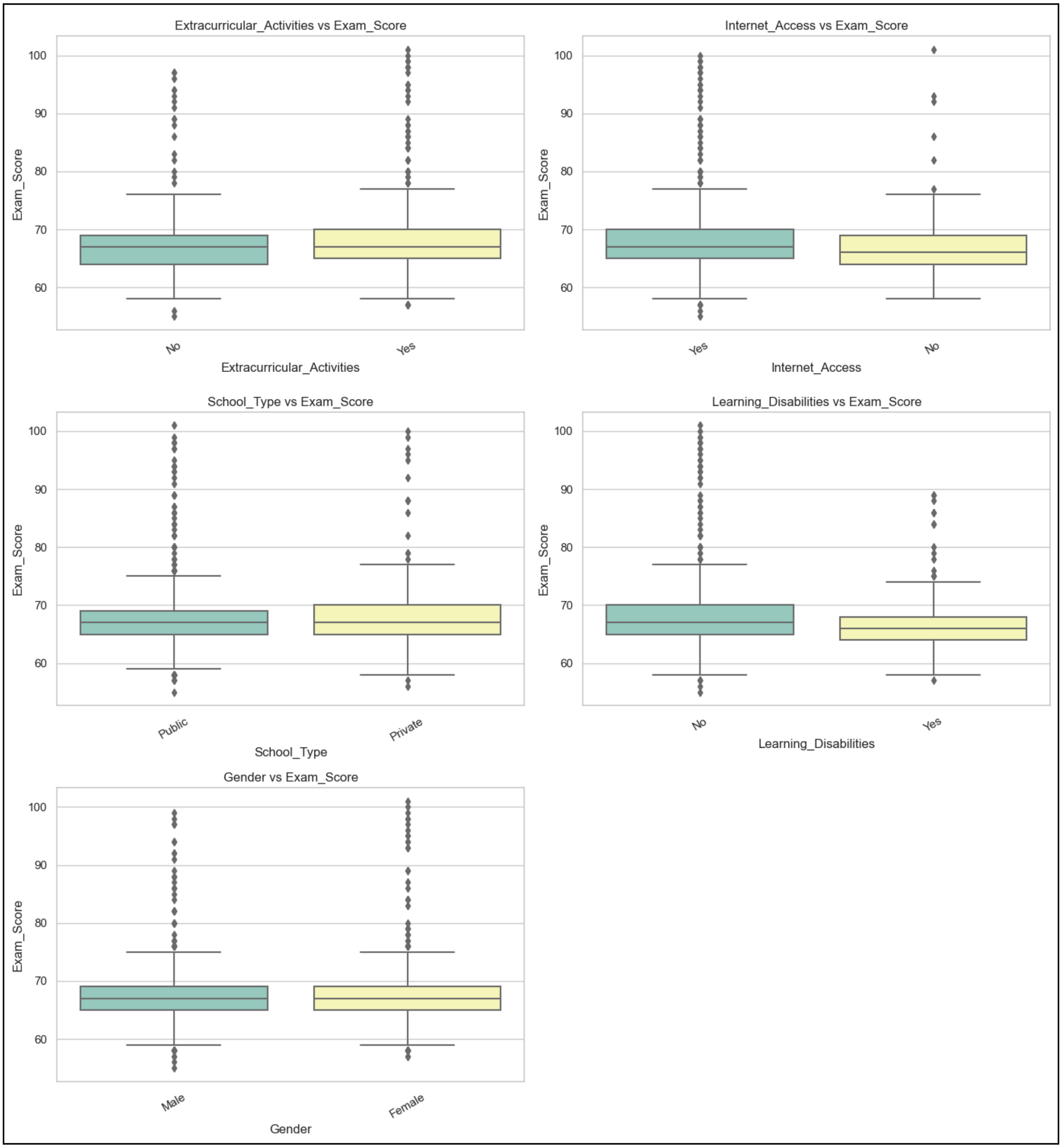


Figure 10 : Box plots showing relationship between nominal categorical variables and exam scores

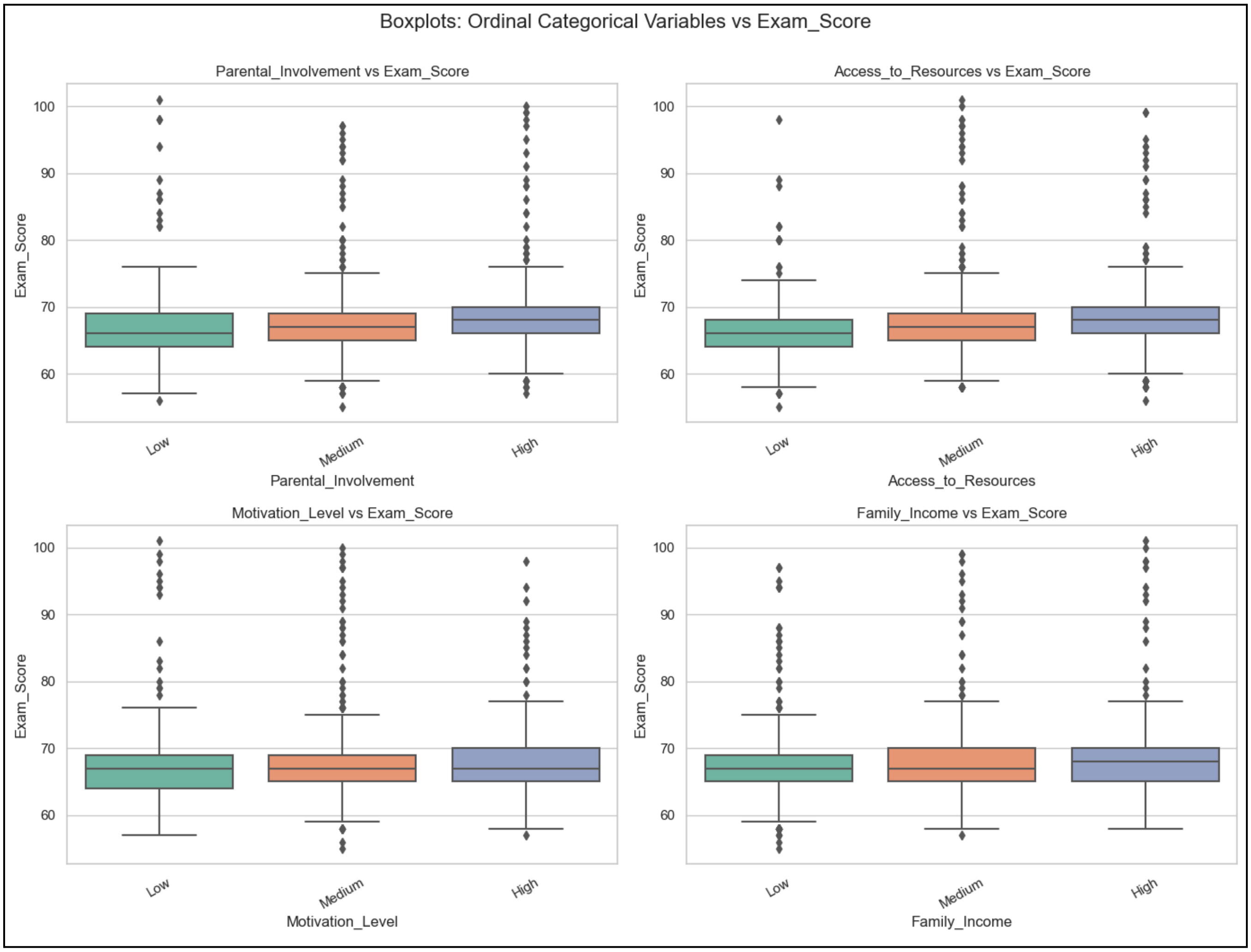
Extracurricular_Activities vs Exam_Score: Students who participate in extracurricular activities tend to have slightly higher exam scores on average compared to those who do not.

Internet_Access vs Exam_Score: Students with Internet access show marginally better performance than those without.

School_Type vs Exam_Score: Students from private schools tend to have slightly higher median scores than those from public schools.

Learning_Disabilities vs Exam_Score: Students with learning disabilities tend to have lower exam scores compared to those without.

Gender vs Exam_Score: Females show a slightly higher median score compared to males, but the difference is minimal.



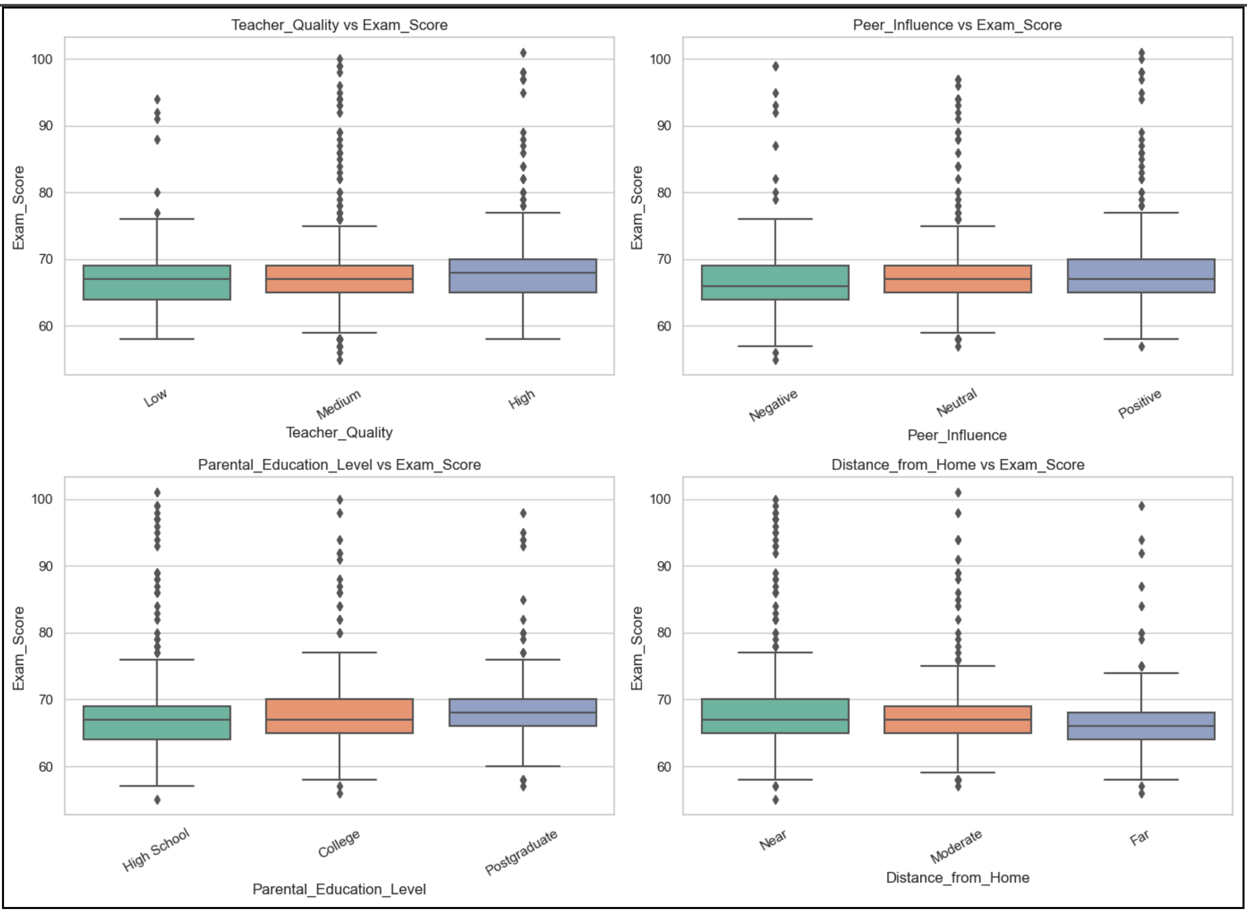


Figure 11 : Box plots showing relationship between ordinal categorical variables and exam scores

Parental_Involvement vs Exam_Score: Students with High parental involvement tend to score slightly higher than those with Medium or Low involvement.

Access_to_Resources vs Exam_Score: A gradual increase in median exam scores is seen from Low → Medium → High access levels.

Motivation_Level vs Exam_Score: Students with High motivation have the highest median scores and a narrower score spread.

Family_Income vs Exam_Score: Slight increase in exam scores from Low → Medium → High income groups, but variation is not very large.

Teacher_Quality vs Exam_Score: Students taught by High quality teachers show slightly better performance overall.

Peer_Influence vs Exam_Score: Students with Positive peer influence tend to perform slightly better than those with Neutral or Negative influence.

Parental_Education_Level vs Exam_Score: Students whose parents have Postgraduate education slightly outperform others.

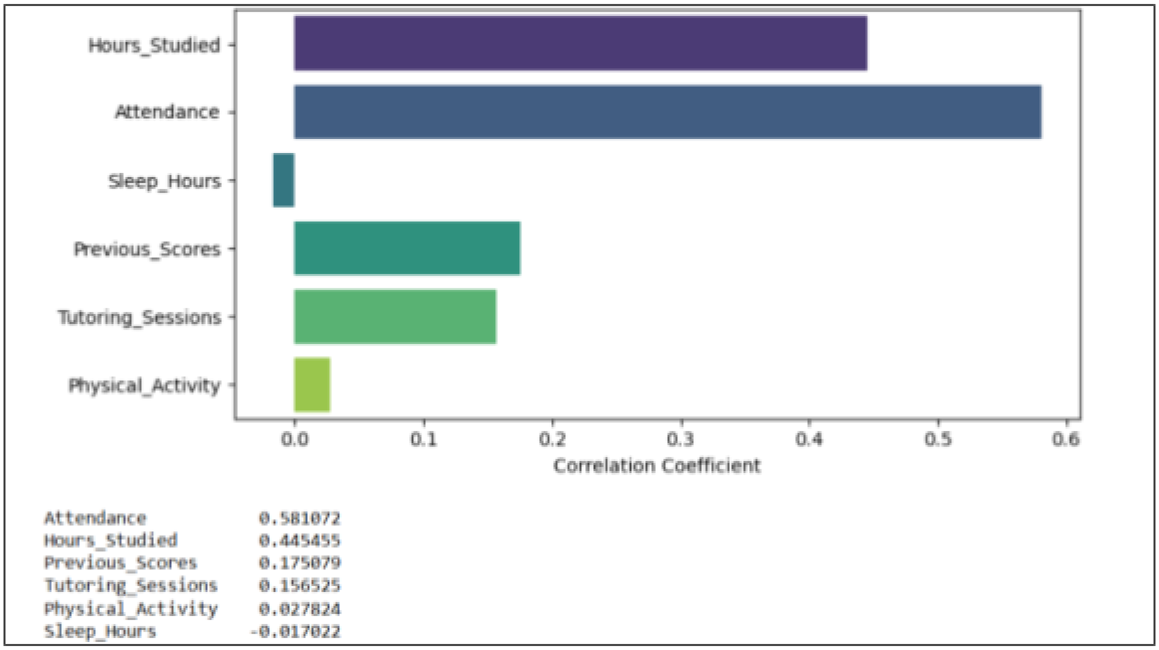
Distance_from_Home vs Exam_Score: Students living Near school tend to score a bit higher than those from Moderate or Far distances.

Feature Selection

To enhance the predictive performance and reduce overfitting, feature selection was carried out to identify the most relevant variables associated with the target variable, *Exam_Score*. The process involved both correlation analysis and statistical testing.

For Numerical Variables

Pearson correlation coefficients were computed with respect to *Exam_Score*.



```
Attendance           0.581072
Hours_Studied        0.445455
Previous_Scores      0.175079
Tutoring_Sessions    0.156525
Physical_Activity    0.027824
Sleep_Hours         -0.017022
```

Figure 12 : Correlation of Numerical Variables with Exam score

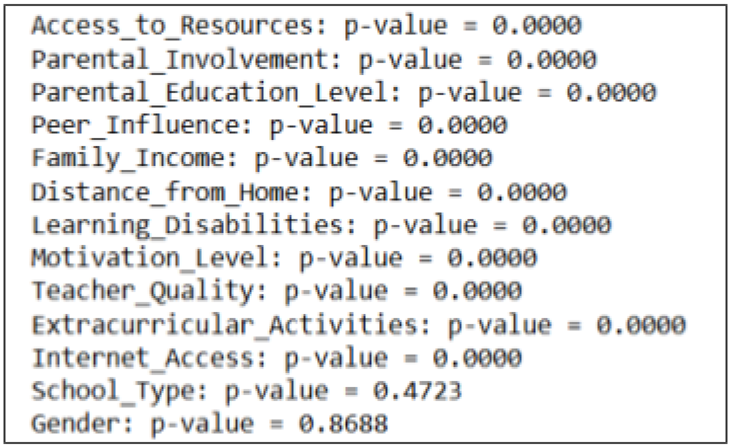
Among the variables evaluated, Attendance showed the strongest positive correlation with Exam_Score (r = 0.581), suggesting that higher class attendance is associated with better exam performance. Hours_Studied also had a moderately positive correlation (r = 0.445), indicating its importance as a predictor.

Other variables like Previous_Scores (r = 0.175) and Tutoring_Sessions (r = 0.157) demonstrated weaker but still positive associations. Meanwhile, Physical_Activity (r = 0.028) and Sleep_Hours (r = –0.017) had negligible correlations, suggesting they may not significantly influence exam outcomes in this dataset.

Based on these findings, **Attendance** and **Hours_Studied** were prioritized for further modeling due to their stronger linear relationships with the target variable.

For categorical variables

One-way ANOVA tests were conducted to determine whether the mean exam scores differed significantly across the groups of each categorical predictor.



```
Access_to_Resources: p-value = 0.0000
Parental_Involvement: p-value = 0.0000
Parental_Education_Level: p-value = 0.0000
Peer_Influence: p-value = 0.0000
Family_Income: p-value = 0.0000
Distance_from_Home: p-value = 0.0000
Learning_Disabilities: p-value = 0.0000
Motivation_Level: p-value = 0.0000
Teacher_Quality: p-value = 0.0000
Extracurricular_Activities: p-value = 0.0000
Internet_Access: p-value = 0.0000
School_Type: p-value = 0.4723
Gender: p-value = 0.8688
```

Figure 13 : ANOVA F Test for categorical variables

The significance of each relationship was assessed using p-values. A threshold of 0.05 was used to determine statistical significance.

The analysis revealed that the following categorical variables had **statistically significant associations** (*p-value < 0.05*) with *Exam_Score*:

- Access to Resources
- Parental Involvement
- Parental Education Level
- Peer Influence
- Family Income
- Distance from Home
- Learning Disabilities
- Motivation Level
- Teacher Quality
- Extracurricular Activities
- Internet Access

These variables were considered as important predictors and retained for further analysis.

Conversely, **School Type** (p = 0.4723) and **Gender** (p = 0.8688) did not show statistically significant relationships with Exam_Score. Therefore, they were considered less relevant in this context and may be excluded from further predictive modeling steps.

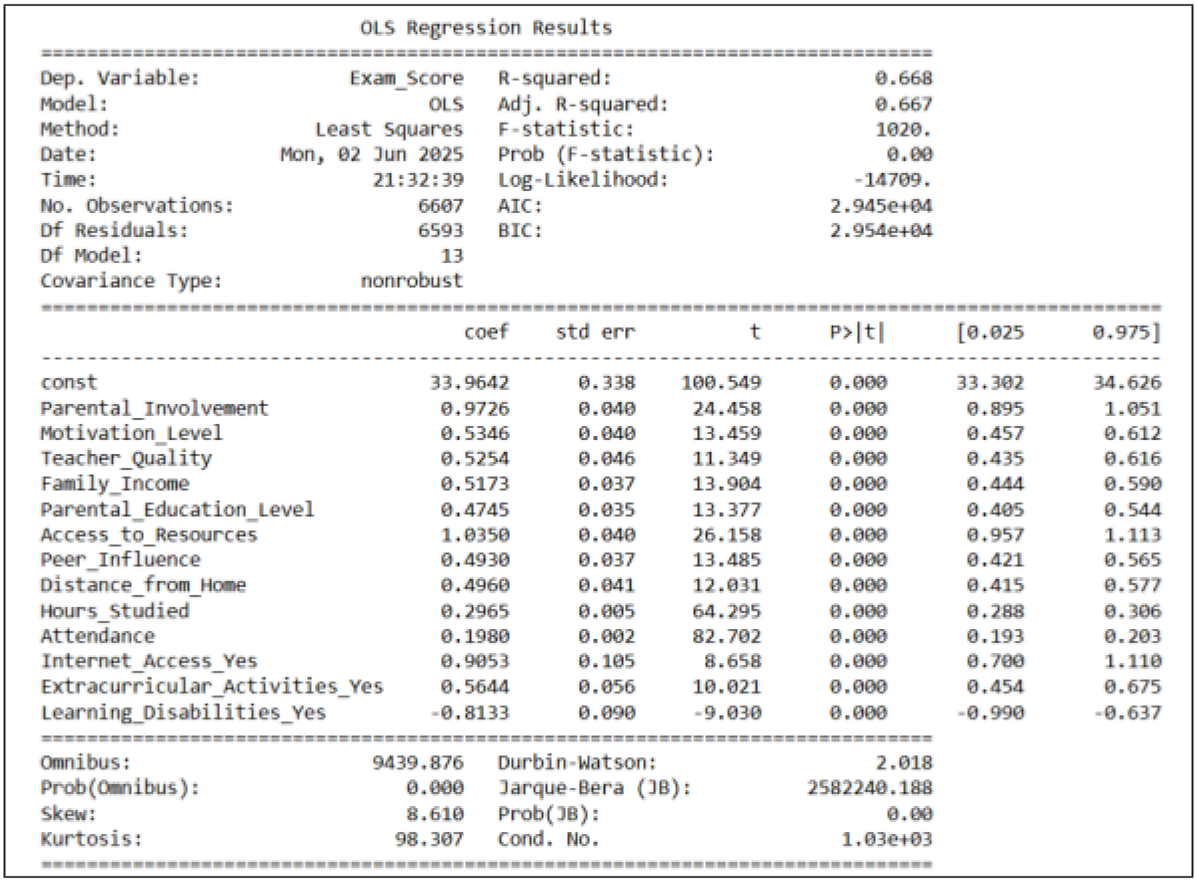
Selected categorical variables were then transformed into numerical form using one-hot encoding to be incorporated into the regression model. This systematic approach to feature selection ensured that only informative variables were included, thereby improving model interpretability and performance.

ADVANCED ANALYSIS

This section presents the advanced statistical techniques applied to assess the relationships between various predictors and the final **Exam Score**. The objective is to build an explanatory model, identify significant factors, and evaluate the performance of the fitted models.

1. Multiple Linear Regression Model

To assess the factors influencing students' academic performance, a multiple linear regression model was fitted using Exam_Score as the dependent variable.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Exam_Score   R-squared:                       0.668
Model:                            OLS   Adj. R-squared:                  0.667
Method:                 Least Squares   F-statistic:                     1020.
Date:                Mon, 02 Jun 2025   Prob (F-statistic):               0.00
Time:                        21:32:39   Log-Likelihood:                -14709.
No. Observations:                6607   AIC:                         2.945e+04
Df Residuals:                    6593   BIC:                         2.954e+04
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         33.9642      0.338    100.549      0.000      33.302      34.626
Parental_Involvement           0.9726      0.040     24.458      0.000       0.895       1.051
Motivation_Level               0.5346      0.040     13.459      0.000       0.457       0.612
Teacher_Quality                0.5254      0.046     11.349      0.000       0.435       0.616
Family_Income                  0.5173      0.037     13.904      0.000       0.444       0.590
Parental_Education_Level       0.4745      0.035     13.377      0.000       0.405       0.544
Access_to_Resources            1.0350      0.040     26.158      0.000       0.957       1.113
Peer_Influence                 0.4930      0.037     13.485      0.000       0.421       0.565
Distance_from_Home             0.4960      0.041     12.031      0.000       0.415       0.577
Hours_Studied                  0.2965      0.005     64.295      0.000       0.288       0.306
Attendance                     0.1980      0.002     82.702      0.000       0.193       0.203
Internet_Access_Yes            0.9053      0.105      8.658      0.000       0.700       1.110
Extracurricular_Activities_Yes 0.5644      0.056     10.021      0.000       0.454       0.675
Learning_Disabilities_Yes     -0.8133      0.090     -9.030      0.000      -0.990      -0.637
==============================================================================
Omnibus:                     9439.876   Durbin-Watson:                   2.018
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2582240.188
Skew:                           8.610   Prob(JB):                         0.00
Kurtosis:                      98.307   Cond. No.                     1.03e+03
==============================================================================
```

Figure 14 : Multiple Linear Regression Results

The model included 13 predictors representing student characteristics, family background, academic environment, and behavioral factors.

The regression model yielded an **R-squared value of 0.668**, indicating that approximately 66.8% of the variability in exam scores is explained by the predictors included in the model. The overall model was statistically significant (F = 1020, p < 0.05), confirming that the combination of variables effectively predicts student performance.

All predictors were statistically significant (p < 0.05) with positive contributions from variables such as:

- Parental Involvement (β = 0.973)

- Motivation Level (β = 0.535)
- Access to Resources (β = 1.035)
- Hours Studied (β = 0.297)
- Attendance (β = 0.198)

Students with internet access and participation in extracurricular activities also showed notably higher scores. Conversely, students identified with learning disabilities had a significant negative association with exam scores (β = –0.813), highlighting the need for targeted interventions.

Based on the multiple linear regression model, the predicted exam score can be represented by the following equation:

Predicted Exam Score = 33.964 + 0.973 × Parental Involvement + 0.535 × Motivation Level + 0.525 × Teacher Quality + 0.517 × Family Income + 0.475 × Parental Education Level + 1.035 × Access to Resources + 0.493 × Peer Influence + 0.496 × Distance from Home + 0.297 × Hours Studied + 0.198 × Attendance + 0.905 × Internet Access (Yes = 1, No = 0) + 0.564 × Extracurricular Activities (Yes = 1, No = 0) − 0.813 × Learning Disabilities (Yes = 1, No = 0)

Residual Analysis

The model assumptions were reasonably met, with a **Durbin-Watson statistic of 2.018**, suggesting no autocorrelation in residuals.

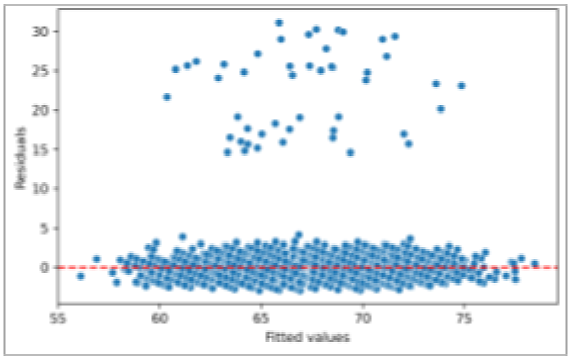


Figure 15 : Residuals vs Fitted values

The residuals vs fitted values plot shows that most residuals are centered around zero. This indicates a generally good model fit. However, a distinct group of large positive residuals suggests the model underestimates exam scores for some observations. This pattern indicates possible issues with constant variance or missing variables.

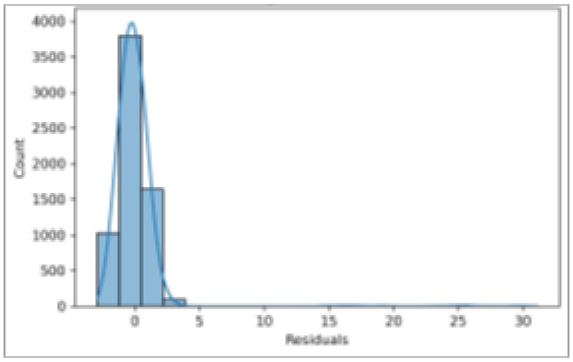


Figure 16: Histogram of Residuals

The histogram of residuals shows a right-skewed distribution. Most errors clustered near zero (0–5). This indicates accurate predictions for many observations. However, the long tail extending up to 30 reveals systematic underprediction for higher values. This suggests potential model bias or unaccounted variables.

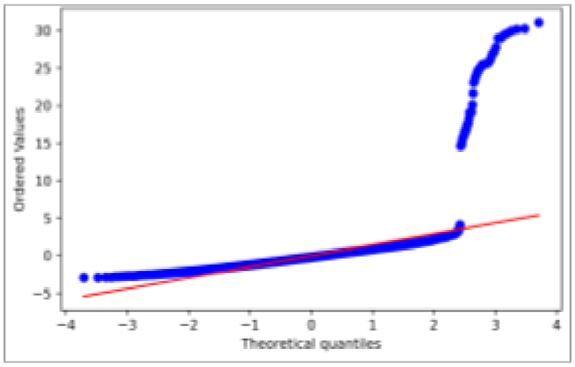


Figure 17: Q-Q plot of Residuals

The visible deviations from the diagonal line, especially in the tails (points below -2 and above 2), indicate that the residuals are not perfectly normally distributed. This suggests potential outliers or non-linearity in the data.

To enhance predictive accuracy and capture complex, non-linear relationships in the data, advanced machine learning models were explored beyond traditional linear regression.

2. Advanced Machine Learning Techniques

In addition to the baseline Ordinary Least Squares (OLS) linear regression model, I applied three advanced machine learning models to improve prediction accuracy of the target variable (Exam_Score). These models are:

- **Decision Tree**
- **Random Forest**
- **XGBoost**

These models were chosen to compare their performance with the linear regression baseline and identify the best predictive model for this dataset.

Model	RMSE	R^2 (Coefficient of Determination)	Comments
OLS Regression	2.24	0.668	Good baseline linear model
Decision Tree	3.98	-0.15	Poor performance (overfitting or high variance)
Random Forest	2.12	0.672	Best balance of accuracy and robustness
XGBoost	2.15	0.66	Comparable to Random Forest (less accurate)

Table 2: Model Performance Summary

Best model: Random Forest is the best performer here with the lowest RMSE (2.12) and the highest R^2 (0.672). It balances predictive accuracy with generalization well.

Although the OLS regression model identified 13 key predictors based on statistical significance and domain logic, the Random Forest model inherently considers all features. For interpretability, only the top

10 features based on importance scores are visualized. However, all features were retained during prediction.

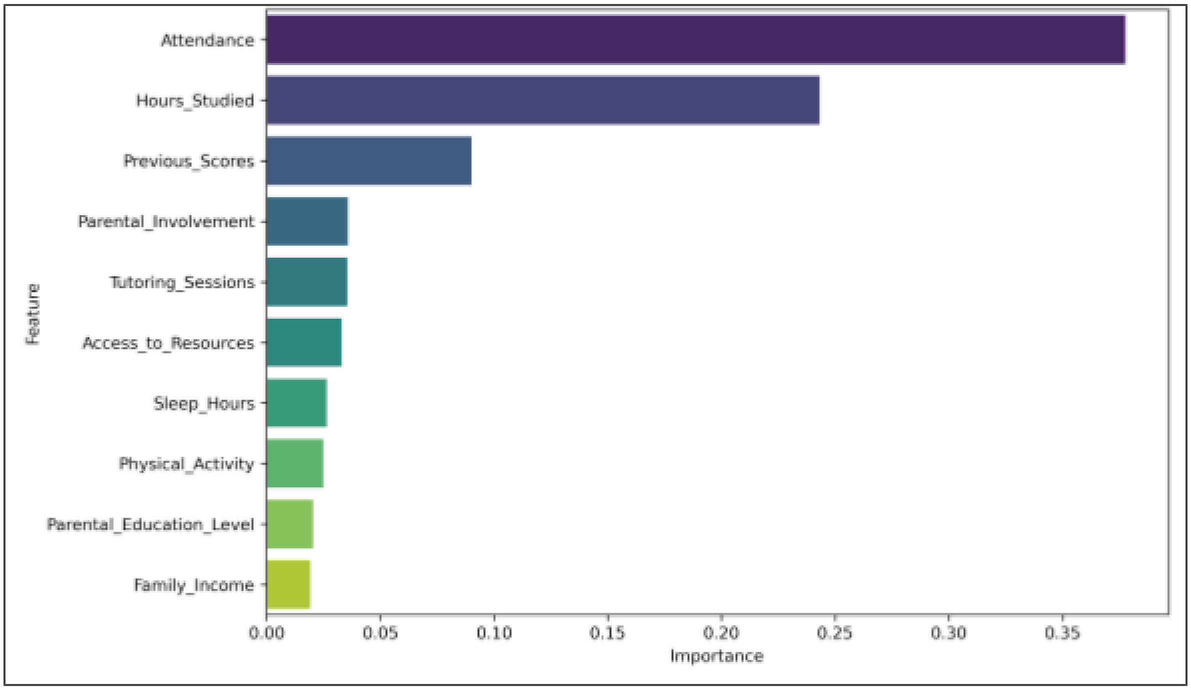


Figure 18: Top 10 Feature Importances - Random Forest

DISCUSSION

- Positive Influences on Academic Performance:
  - Hours Studied and Attendance showed strong positive correlations with Exam Scores. It highlights the value of consistent academic effort.
  - Motivation Level and Parental Involvement also played significant roles, consistent with psychological theories that emphasize internal drive and external support.
  - Access to Resources and Teacher Quality contributed positively. It indicates the importance of a conducive learning environment.

- Moderate/Weak Influences:
  - Previous Scores and Tutoring Sessions had a mild positive effect. It suggests tutoring supports, but not replaces self-study. Also it highlights the historical performance.
  - Internet Access showed minor improvements in scores, likely due to differences in availability of online materials.
  - Extracurricular Activities appeared slightly beneficial, potentially due to improved time management and reduced stress.
  - Distance from Home had a weak relationship with performance. Students living nearer performed marginally better, possibly due to reduced fatigue or stress from commuting.

- Limited or No Strong Correlation:
  - Physical Activity and Sleep Hours showed no consistent linear trend with Exam Score. But they might affect performance indirectly (via focus, mental health).
  - Gender had minimal effect. Female students slightly outperformed males, aligning with similar findings in education literature.
  - School Type showed minimal effect, likely due to differences in institutional quality.
- Challenges and Observations:
  - Learning Disabilities were associated with lower exam scores, underscoring the need for tailored support.
  - Outliers like scores exceeding 100 suggest some data entry inconsistencies.
  - Multicollinearity was addressed during regression modeling, ensuring reliable coefficient interpretations.
  - Machine Learning models like XGBoost outperformed traditional regression in prediction accuracy but are less interpretable.

CONCLUSION

- Academic performance is multifactorial, influenced by personal effort (study time, motivation), support systems (parents, teachers), and institutional context (resources, school type).

- Feature Importance:
  - The most predictive variables were Attendance, Hours Studied, Previous Scores and Motivation Level.
  - These variables can guide targeted interventions in education policy and student support programs.

- Predictive Models:
  - Random Forest achieved the highest predictive accuracy (lowest RMSE, highest R^2), making it the most effective tool for future score prediction.
  - However, Multiple Linear Regression remains useful for understanding variable relationships due to its interpretability.

- Practical Implications:
  - Schools can enhance student outcomes by promoting study discipline, improving teacher quality, ensuring parental engagement and expanding resource access.
  - Identifying at-risk students early using predictive models can help allocate support efficiently.

- Limitations:
  - The dataset was secondary and limited to 6607 records from one source (Kaggle), which may not generalize to all populations.
  - Self-reported variables (motivation, sleep) may be biased.

- Future Research:
  - Test advanced models for improved accuracy.
  - Explore interaction effects and non-linear relationships.
  - Use longitudinal data to assess long-term trends in academic performance.

REFERENCES

1. S. Kotsiantis, C.Pierrakeas & P. Pintelas (2004). *Predicting students' performance in distance learning using machine learning techniques.* Applied Artificial Intelligence, 18(5), 411–426. Predicting students' performance in distance learning using machine learning techniques

2. C. Romero & S. Ventura (2010). *Educational data mining: A review of the state of the art.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601–618.
   Educational Data Mining: A Review of the State of the Art

3. Fan, X., & Chen, M. (2001). *Parental involvement and students' academic achievement: A meta-analysis.* Educational Psychology Review, 13(1), 1–22. (PDF) Parental Involvement and Students' Academic Achievement: A Growth Modeling Analysis

4. P.R. Pintrich & E.V. De Groot (1990). *Motivational and self-regulated learning components of classroom academic performance.* Journal of Educational Psychology, 82(1), 33–40. pintrich and degroodt 1990.pdf

5. R.A. Reiser & J.V. Dempsey (2012). *Trends and issues in instructional design and technology* (3rd ed.). Pearson Education. (PDF) Trends and issues in instructional design and technology

6. Zhang, J., Shi, Y., & Yang, Q. (2019). *Student performance prediction: A survey.* IEEE Transactions on Learning Technologies, 12(4), 577–589. 1707.08114

7. R.J.A. Little & D.B. Rubin (2002). *Statistical analysis with missing data* (3rd ed.). Wiley. Statistical Analysis with Missing Data 3rd Edition Roderick J. A. Little instant download | PDF | Resampling (Statistics) | Robust Statistics

8. Bronfenbrenner's Ecological Systems Theory. (2023). *Simply Psychology*.
   https://www.simplypsychology.org/bronfenbrenner.html

9. Constructivist Learning Theory. (2023). *Simply Psychology*.
   https://www.simplypsychology.org/constructivism.html

10. Educational Data Mining. (n.d.). *International Educational Data Mining Society*.
    https://educationaldatamining.org/

11. Laing. (2023). *Student Performance Factors* [Data set]. Kaggle.
    https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/data