STUDENT PERFORMANCE FACTORS ANALYSIS

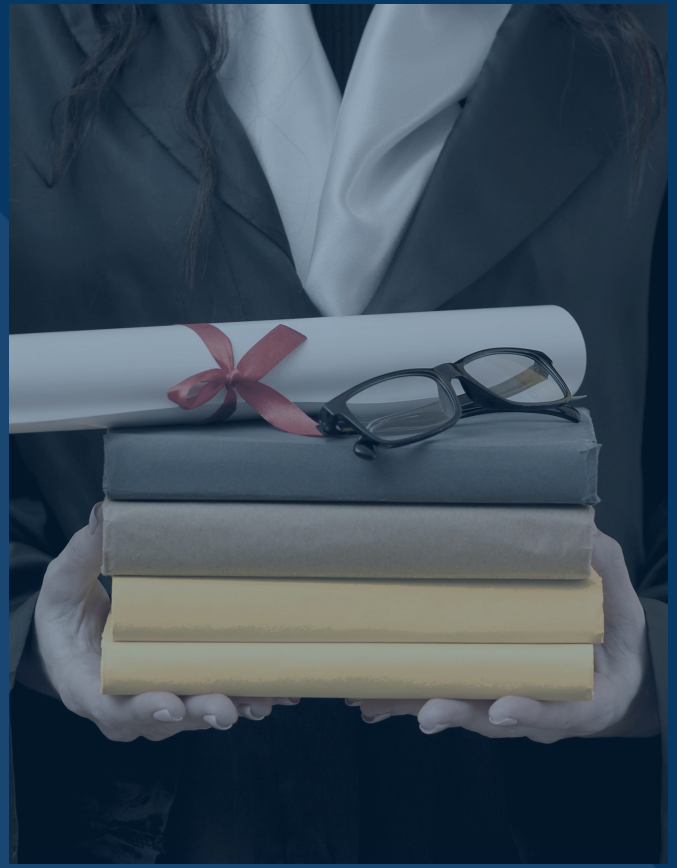
Dana Ortiz

TABLE OF CONTENTS

01 BACKGROUND

Background - Company

Pupil Progress Analytics is a forward-thinking educational technology company focused on improving student outcomes through data-driven insights and innovative educational solutions. The company partners with schools, educational institutions, and tutoring organizations to integrate cutting-edge data analytics into their operations, helping them understand and address various academic challenges.



Background - Stakeholder

As the Director of Data Analytics & Research, the Stakeholder (You!) leads a team that focuses on gathering and analyzing data to provide actionable insights that can improve student learning experiences. This role involves using data to help institutions better understand their students' needs, track academic performance trends, and suggest personalized interventions that can help improve student success rates and thus the school's competitiveness.

02

THE DATASET

THE DATASET

Target Variable:

Final Exam Score (See all variables on next slide)

Our "Student Performance Factors" dataset provides comprehensive insights into various elements that affect student academic performance. It includes data on study habits, parental involvement, attendance, sleep patterns, access to resources, and more, which all contribute to the final exam score of students.

Number of Records:
6,607

Number of Features:
20

Variables

- **Hours_Studied:** Number of hours spent studying per week.
- **Attendance:** Percentage of classes attended.
- **Parental_Involvement:** Level of parental involvement.
- **Access_to_Resources:** Availability of resources.
- **Extracurricular_Activities:** Participation in activities.
- **Sleep_Hours:** Average number of hours of sleep per night.
- **Previous_Scores:** Scores from previous exams.
- **Motivation_Level:** Student's level of motivation.
- **Internet_Access:** Availability of internet access.
- **Tutoring_Sessions:** Number of sessions attended monthly.
- **Family_Income:** Family income level .
- **Teacher_Quality:** Quality of the teachers .
- **School_Type:** Type of school attended (Public, Private).
- **Peer_Influence:** Influence of peers on performance.
- **Physical_Activity:** Average number of hours of physical activity per week.
- **Learning_Disabilities:** Presence of learning disabilities.
- **Parental_Education_Level:** Highest education level of parents (High School, College, Postgraduate).
- **Distance_from_Home:** Distance from home to school.
- **Gender:** Gender of the student (Male, Female).
- **Exam_Score:** Final exam score.

03

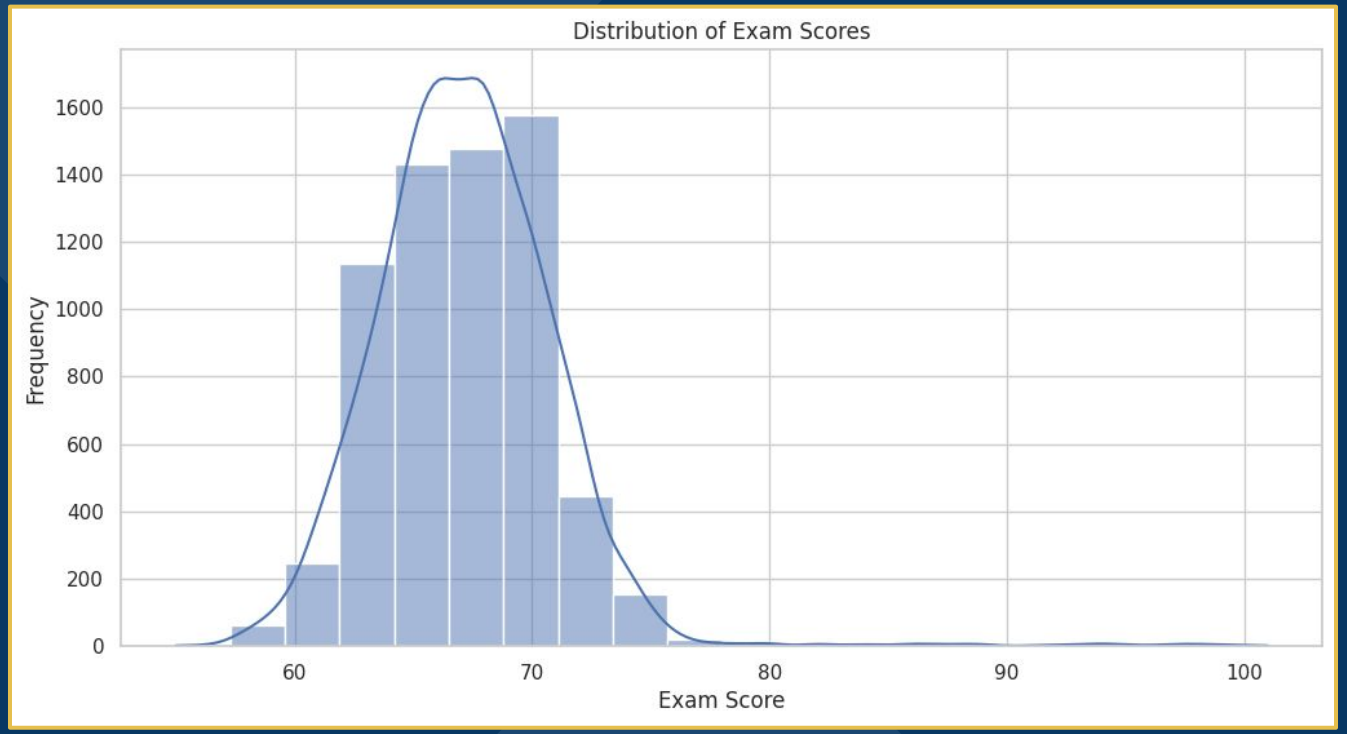
INSIGHTS AND EXPLANATIONS

How did we prep the data?

To prepare the data for analysis, all categorical variables were cleaned and converted into numerical form.

Ordinal variables such as "Parental Involvement" were mapped to numerical values from 1 to 3 based on levels of “Low,” “Medium,” and “High.”

Binary variables like “Internet Access” were encoded as 1 for “Yes” and 0 for “No.”



This histogram shows the overall distribution of student exam scores. The shape of the curve helps us understand how scores are spread out across the population. We observe that the majority of students fall within a middle range, with fewer students at the very low or very high ends. This normal-like distribution suggests a relatively balanced dataset with a few outliers, and it gives us a baseline for interpreting how various factors might shift a student's score.

Key Insights

Strong Positive Influences

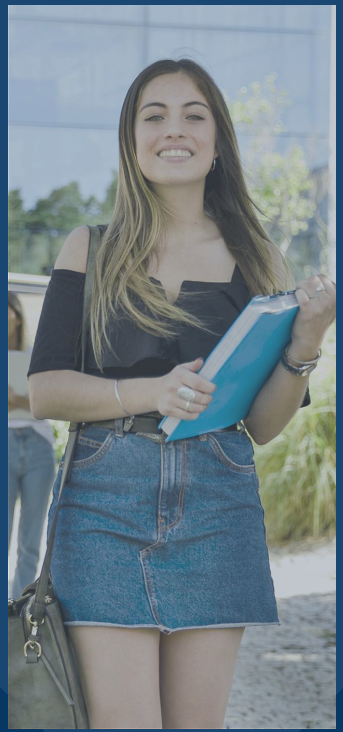
Access_to_Resources (1.02): Students with better access to educational resources tend to score higher.

Parental_Involvement (1.00): Active parental engagement is associated with improved student performance.

Internet_Access (0.90): Reliable internet access facilitates learning and positively impacts scores.

Motivation_Level (0.55): Highly motivated students achieve better exam results.

Teacher_Quality (0.55): Quality teaching contributes significantly to student success.



Negative Influences

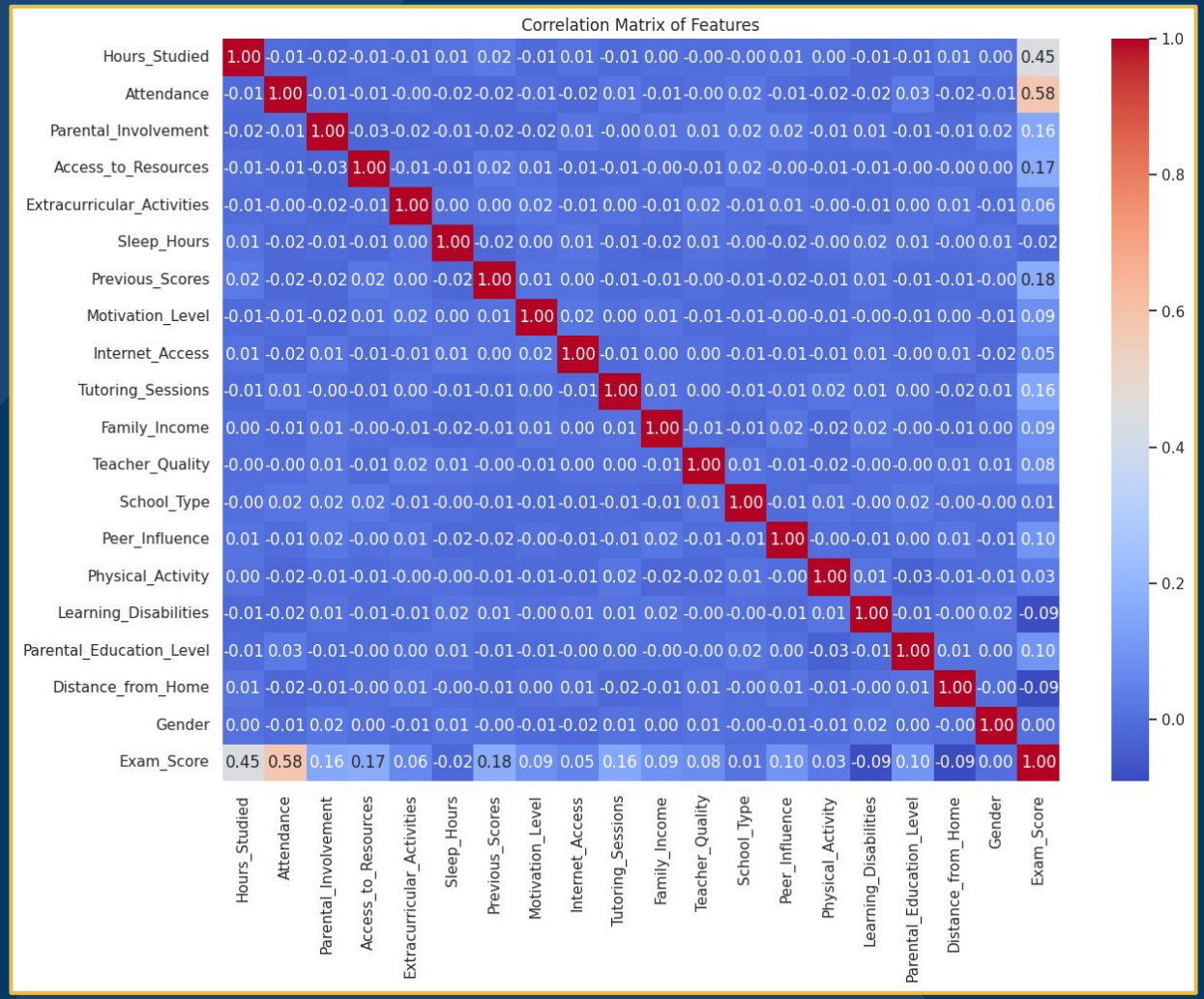
Learning_Disabilities (-0.86): Students with learning disabilities may require additional support to achieve comparable scores.

Distance_from_Home (-0.46): Longer commutes can negatively affect student performance, possibly due to fatigue or reduced study time.

School_Type (-0.05): A slight negative coefficient suggests minimal impact, but further investigation may be warranted.

INSIGHTS AND EXPLANATIONS

One of the most valuable visual tools used in this analysis was the correlation matrix. This heatmap allowed us to see which numerical variables were most strongly associated with final exam scores. Notably, we observed strong positive correlations between Exam Score and variables such as Hours Studied and attendance. There were also many lower positive correlations with Parental Involvement and tutoring. Learning Disabilities and Distance from Home showed slightly negative relationships with exam performance.



Linear Regression

Linear Regression was applied to predict student exam scores based on a combination of behavioral, environmental, and demographic factors. This model showed that certain factors had stronger positive coefficients, confirming their importance in academic outcomes. Although the model assumes linear relationships and doesn't capture complex interactions as well as a decision tree, it provides clear and actionable weightings for each variable, helping prioritize interventions and explain more variance for this specific dataset.

Linear Regression Equation:

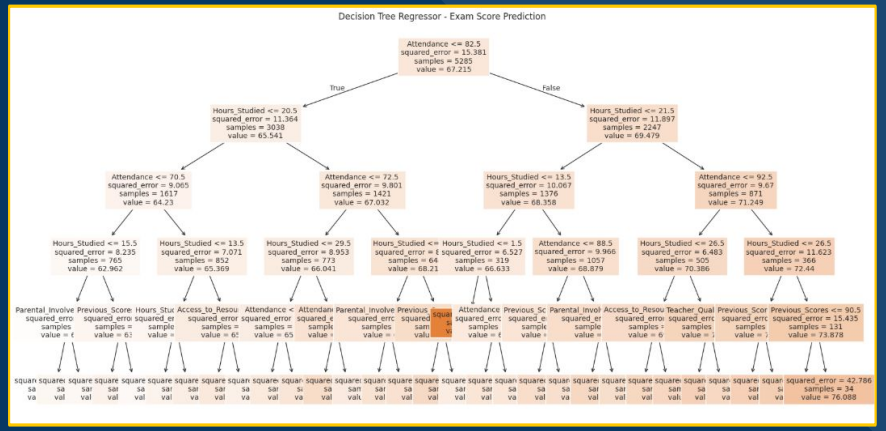
Exam_Score = (0.29 * Hours_Studied) + (0.20 * Attendance) + **(1.00 * Parental_Involvement)** + **(1.02 * Access_to_Resources)** + (0.56 * Extracurricular_Activities) + (-0.01 * Sleep_Hours) + (0.05 * Previous_Scores) + (0.55 * Motivation_Level) + **(0.90 * Internet_Access)** + (0.48 * Tutoring_Sessions) + (0.56 * Family_Income) + (0.55 * Teacher_Quality) + (-0.05 * School_Type) + (0.50 * Peer_Influence) + (0.20 * Physical_Activity) + **(-0.86 * Learning_Disabilities)** + (0.49 * Parental_Education_Level) + (-0.46 * Distance_from_Home) + (0.01 * Gender) + (0.34 * Passed) + 30.49

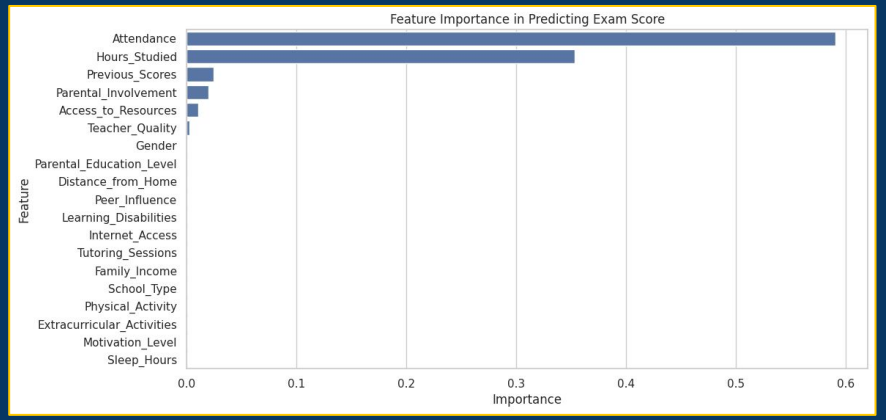
Decision Tree

We also tried a Decision Tree Regressor to model the relationship between student factors and their final exam scores. This model was almost chosen because it captures complex, non-linear interactions between variables and provides a highly interpretable structure, perfect for identifying actionable drivers of student success.

However, our decision tree has a Mean Squared Error of 6.44 and a R^2 Score of 0.54 and our linear regression had a MSE of 4 .15 and R^2 of 0.73. This means that the linear regression is a better fit for our dataset and our goals.

This model reveals which factors most strongly affect performance and can guide interventions. This insight can help the company prioritize resources toward improving these specific areas to boost student success.





04

SUGGESTIONS

Suggestions at a Glance

Based on our model's findings, the following recommendations can be made:

- **Enhance Resource Availability:** Invest in educational materials and ensure students have access to necessary learning tools.
- **Promote Parental Engagement:** Implement programs to encourage parents to participate in their children's education.
- **Improve Internet Infrastructure:** Provide reliable internet access to support digital learning initiatives.
- **Support Motivational Programs:** Develop strategies to boost student motivation, such as goal-setting workshops or mentorship programs.
- **Invest in Teacher Development:** Offer professional development opportunities to enhance teaching quality.
- **Address Learning Disabilities:** Provide specialized support and resources for students with learning challenges.
- **Consider Transportation Solutions:** Explore options to reduce commute times, such as school transportation services or remote learning alternatives.

Providing Access to Open Educational Resources

Implementation Steps:

- **Curate Quality OER Materials:** Identify and select high-quality, curriculum-aligned open resources.
- **Train Educators:** Offer professional development to help teachers effectively integrate OER into their instruction.
- **Develop an OER Repository:** Create a centralized digital library where students and teachers can easily access materials.

Studies have shown that implementing OER can save students significant amounts on textbook costs. For example, students at schools participating in OER initiatives paid at least $65 less per course on average.

Anticipated Benefits:

- **Cost Savings for Students:** Reduces the financial burden of purchasing textbooks and materials.
- **Increased Accessibility:** Ensures all students have equal access to required learning materials.
- **Enhanced Learning Outcomes:** OER has been shown to increase student learning while breaking down barriers of affordability and accessibility.

Influence on Other Factors:

- **Motivation Level:** Easier access to materials can increase student motivation and preparedness.
- **Parental Involvement:** Parents can utilize OER to better support their children's learning at home.

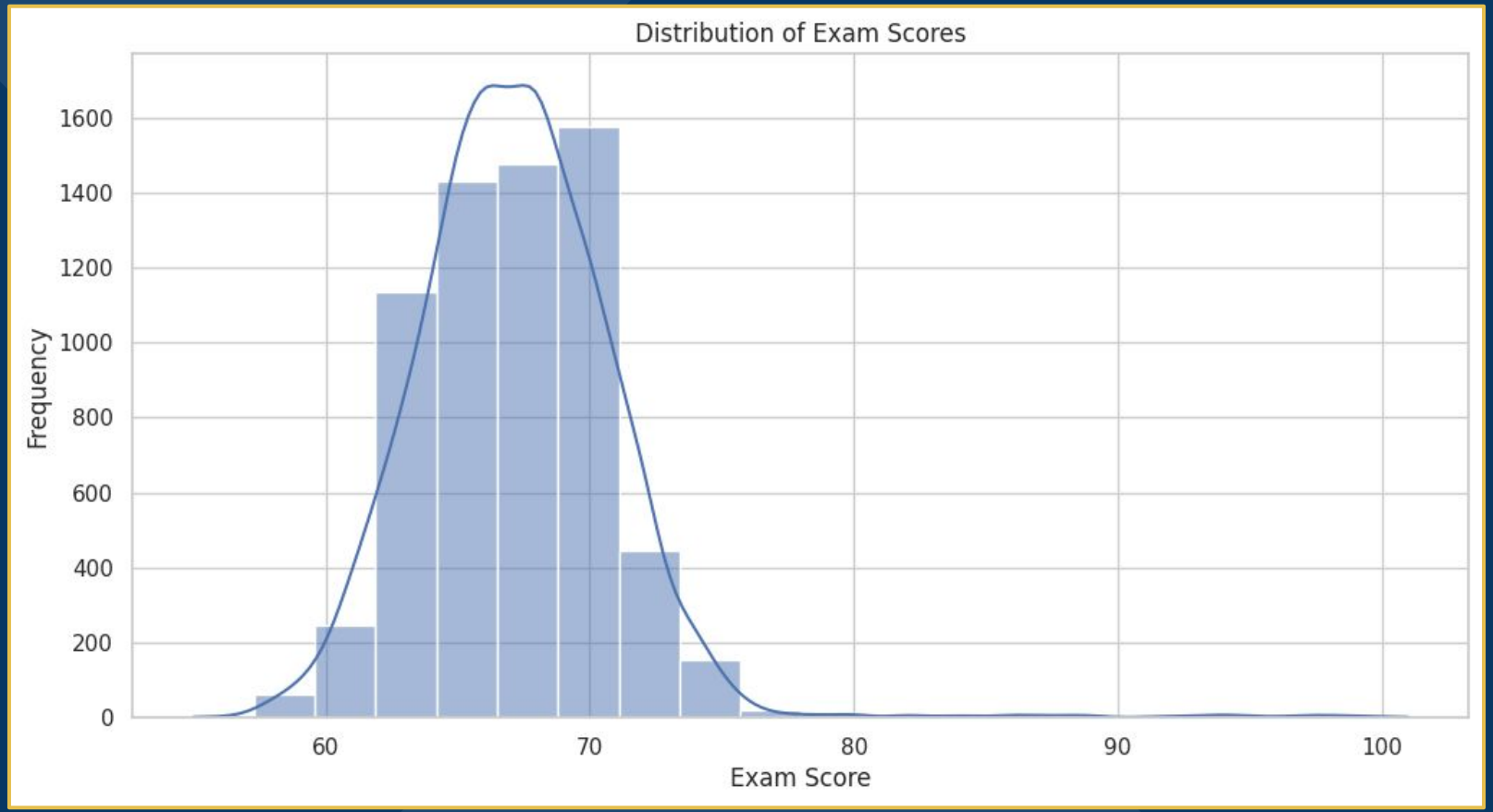
All of our code was written and ran in Google Collab. Compiled here is an ipynb for the code at the date of this presentation and the full dataset csv.

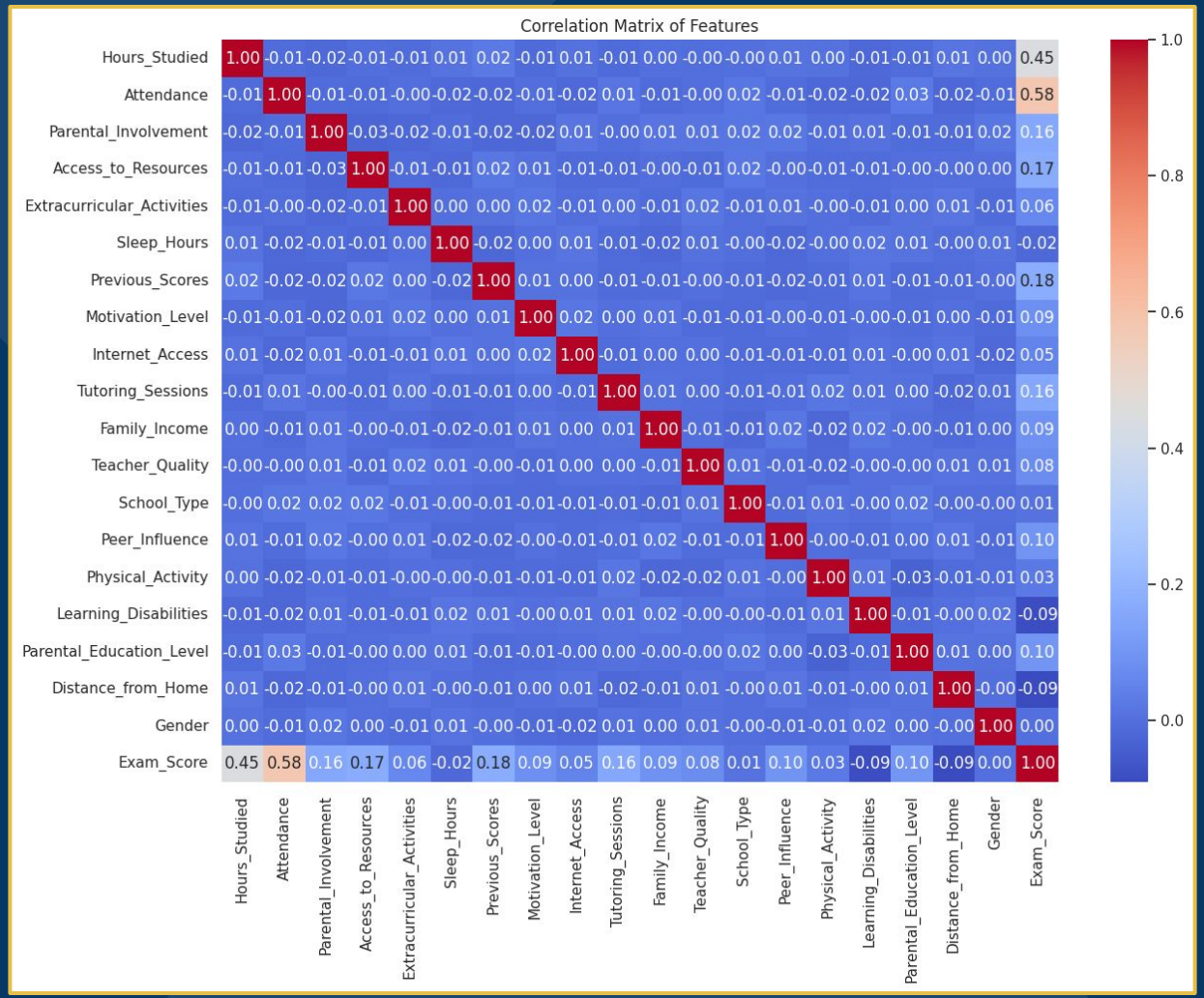
Download ipynb

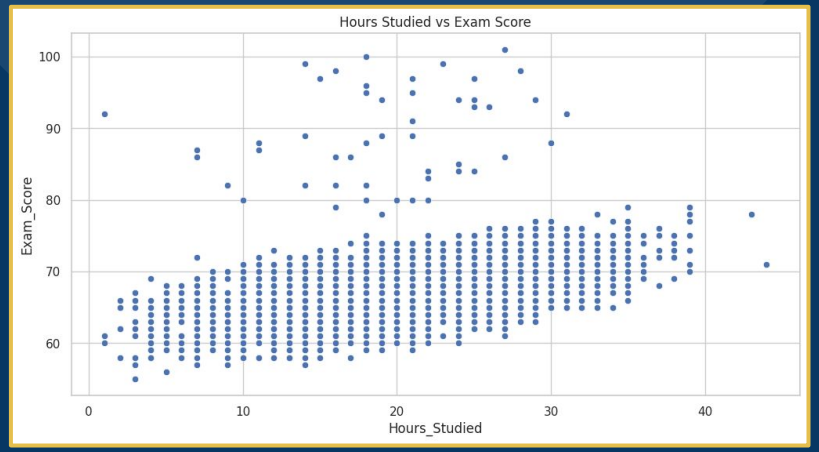
Download csv

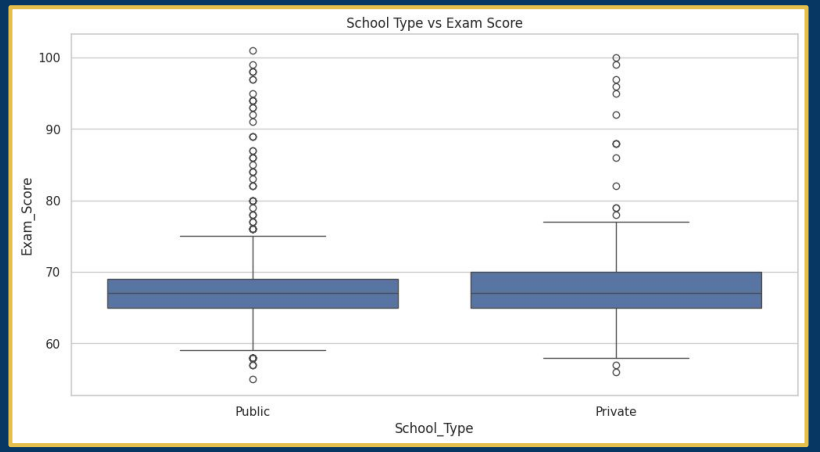

Thank you for your time!

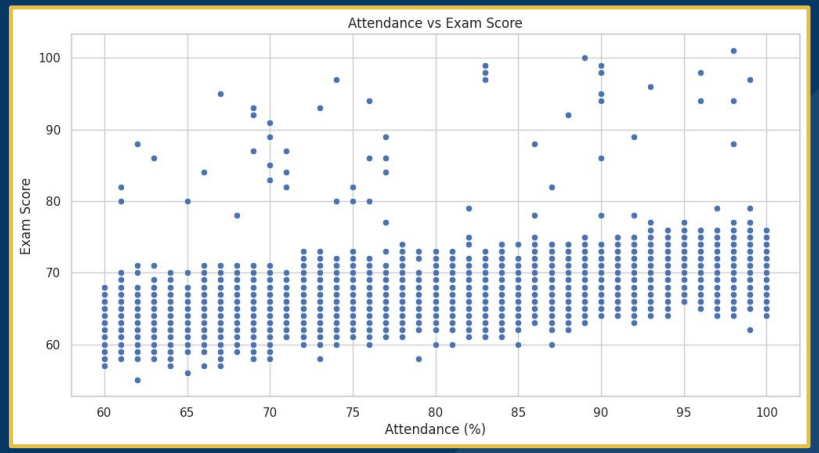
A APPENDIX


Distribution of Exam Scores
Frequency
0
200
400
600
800
1000
1200
1400
1600
60
70
80
90
100
Exam Score


Correlation Matrix of Features
Hours_Studied 1.00 -0.01 -0.02 -0.01 -0.01 0.01 0.02 -0.01 0.01 -0.01 0.00 -0.00 -0.00 0.01 0.00 -0.01 -0.01 0.01 0.00 0.45
Attendance -0.01 1.00 -0.01 -0.01 -0.00 -0.02 -0.02 -0.01 -0.02 0.01 -0.01 -0.00 0.02 -0.01 -0.02 -0.02 0.03 -0.02 -0.01 0.58
Parental_Involvement -0.02 -0.01 1.00 -0.03 -0.02 -0.01 -0.02 -0.02 0.01 -0.00 0.01 0.01 0.02 0.02 -0.01 0.01 -0.01 -0.01 0.02 0.16
Access_to_Resources -0.01 -0.01 -0.03 1.00 -0.01 -0.01 0.02 0.01 -0.01 -0.01 -0.00 -0.01 0.02 -0.00 -0.01 -0.01 -0.00 -0.00 0.00 0.17
Extracurricular_Activities -0.01 -0.00 -0.02 -0.01 1.00 0.00 0.00 0.02 -0.01 0.00 -0.01 0.02 -0.01 0.01 -0.00 -0.01 0.00 0.01 -0.01 0.06
Sleep_Hours 0.01 -0.02 -0.01 -0.01 0.00 1.00 -0.02 0.00 0.01 -0.01 -0.02 0.01 -0.00 -0.02 -0.00 0.02 0.01 -0.00 0.01 -0.02
Previous_Scores 0.02 -0.02 -0.02 0.02 0.00 -0.02 1.00 0.01 0.00 -0.01 -0.01 -0.00 -0.01 -0.02 -0.01 0.01 -0.01 -0.01 -0.00 0.18
Motivation_Level -0.01 -0.01 -0.02 0.01 0.02 0.00 0.01 1.00 0.02 0.00 0.01 -0.01 -0.01 -0.00 -0.01 -0.00 -0.01 0.00 -0.01 0.09
Internet_Access 0.01 -0.02 0.01 -0.01 -0.01 0.01 0.00 0.02 1.00 -0.01 0.00 0.00 -0.01 -0.01 -0.01 0.01 -0.00 0.01 -0.02 0.05
Tutoring_Sessions -0.01 0.01 -0.00 -0.01 0.00 -0.01 -0.01 0.00 -0.01 1.00 0.01 0.00 -0.01 -0.01 0.02 0.01 0.00 -0.02 0.01 0.16
Family_Income 0.00 -0.01 0.01 -0.00 -0.01 -0.02 -0.01 0.01 0.00 0.01 1.00 -0.01 -0.01 0.02 -0.02 0.02 -0.00 -0.01 0.00 0.09
Teacher_Quality -0.00 -0.00 0.01 -0.01 0.02 0.01 -0.00 -0.01 0.00 0.00 -0.01 1.00 0.01 -0.01 -0.02 -0.00 -0.00 0.01 0.01 0.08
School_Type -0.00 0.02 0.02 0.02 -0.01 -0.00 -0.01 -0.01 -0.01 -0.01 -0.01 0.01 1.00 -0.01 0.01 -0.00 0.02 -0.00 -0.00 0.01
Peer_Influence 0.01 -0.01 0.02 -0.00 0.01 -0.02 -0.02 -0.00 -0.01 -0.01 0.02 -0.01 -0.01 1.00 -0.00 -0.01 0.00 0.01 -0.01 0.10
Physical_Activity 0.00 -0.02 -0.01 -0.01 -0.00 -0.00 -0.01 -0.01 -0.01 0.02 -0.02 -0.02 0.01 -0.00 1.00 0.01 -0.03 -0.01 -0.01 0.03
Learning_Disabilities -0.01 -0.02 0.01 -0.01 -0.01 0.02 0.01 -0.00 0.01 0.01 0.02 -0.00 -0.00 -0.01 0.01 1.00 -0.01 -0.00 0.02 -0.09
Parental_Education_Level -0.01 0.03 -0.01 -0.00 0.00 0.01 -0.01 -0.01 -0.00 0.00 -0.00 -0.00 0.02 0.00 -0.03 -0.01 1.00 0.01 0.00 0.10
Distance_from_Home 0.01 -0.02 -0.01 -0.00 0.01 -0.00 -0.01 0.00 0.01 -0.02 -0.01 0.01 -0.00 0.01 -0.01 -0.00 0.01 1.00 -0.00 -0.09
Gender 0.00 -0.01 0.02 0.00 -0.01 0.01 -0.00 -0.01 -0.02 0.01 0.00 0.01 -0.00 -0.01 -0.01 0.02 0.00 -0.00 1.00 0.00
Exam_Score 0.45 0.58 0.16 0.17 0.06 -0.02 0.18 0.09 0.05 0.16 0.09 0.08 0.01 0.10 0.03 -0.09 0.10 -0.09 0.00 1.00
Hours_Studied
Attendance
Parental_Involvement
Access_to_Resources
Extracurricular_Activities
Sleep_Hours
Previous_Scores
Motivation_Level
Internet_Access
Tutoring_Sessions
Family_Income
Teacher_Quality
School_Type
Peer_Influence
Physical_Activity
Learning_Disabilities
Parental_Education_Level
Distance_from_Home
Gender
Exam_Score
1.0
0.8
0.6
0.4
0.2
0.0


Hours Studied vs Exam Score
Exam_Score
Hours_Studied
0
10
20
30
40
60
70
80
90
100

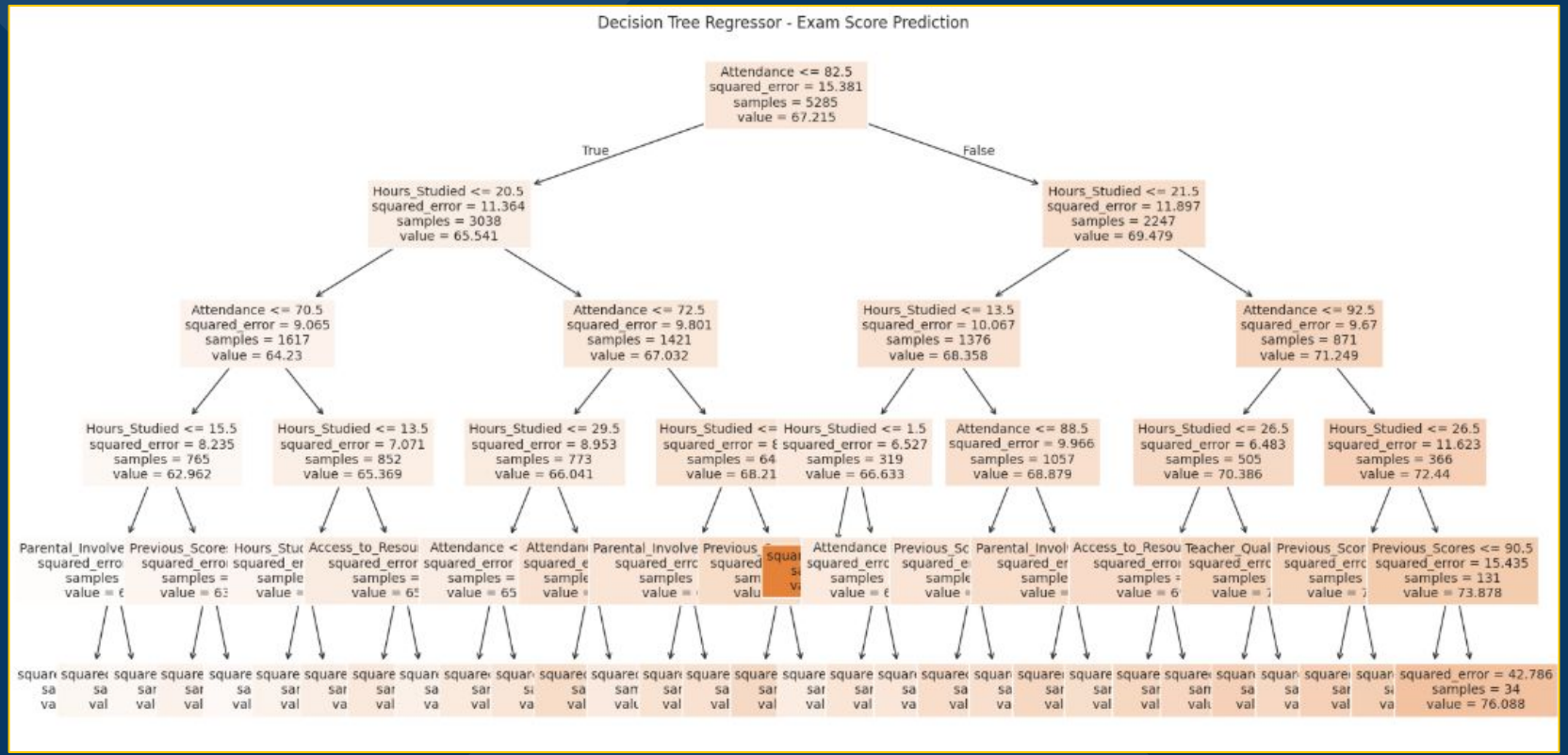

School Type vs Exam Score
Exam_Score
School_Type
Public
Private
60
70
80
90
100

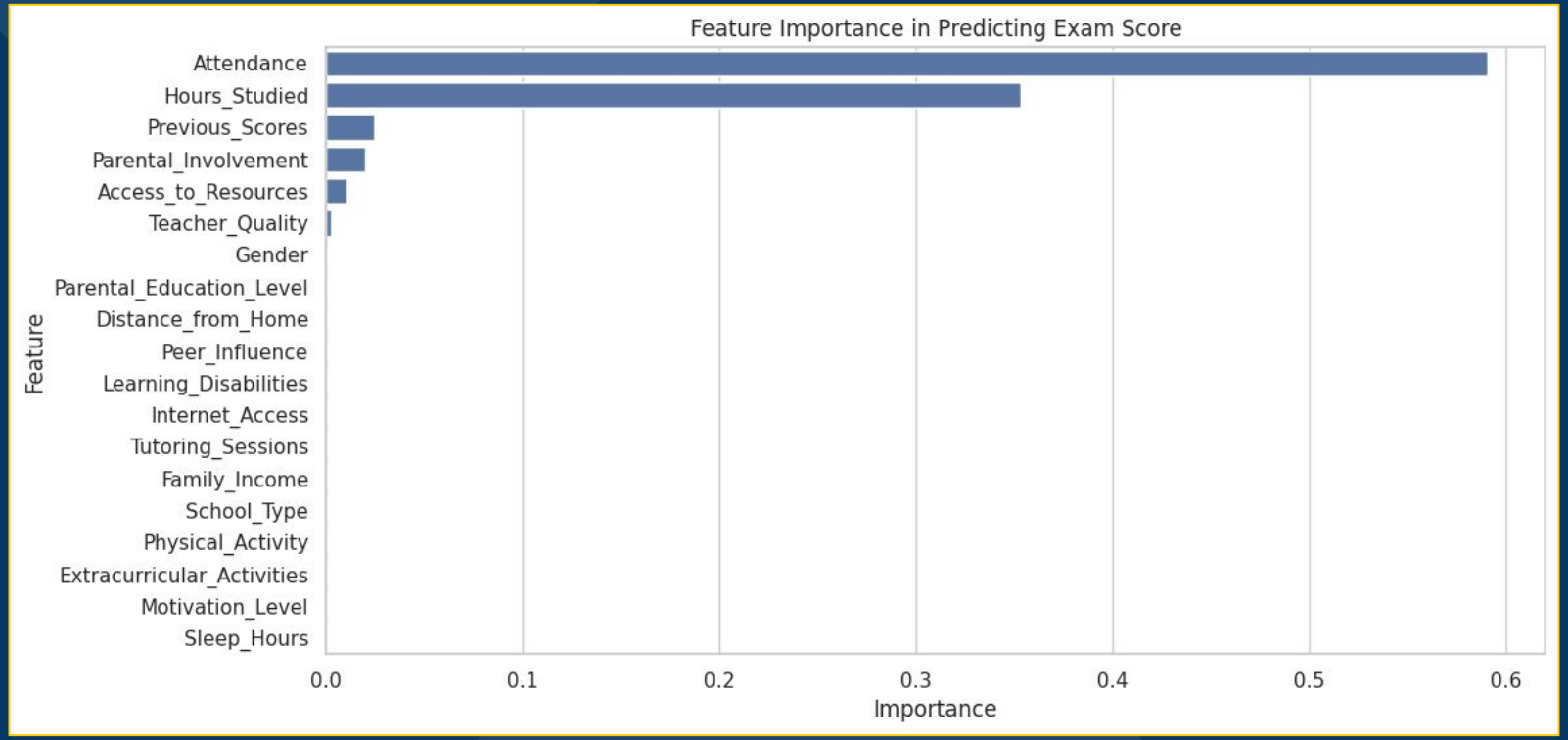

Attendance vs Exam Score
Exam Score
Attendance (%)
60
65
70
75
80
85
90
95
100


Exam Score Distribution by Gender
Exam_Score
Gender
Male
Female
60
70
80
90
100

Linear Regression Equation:

Exam_Score = (0.29 * Hours_Studied) + (0.20 * Attendance) + (1.00 * Parental_Involvement) + (1.02 * Access_to_Resources) + (0.56 * Extracurricular_Activities) + (-0.01 * Sleep_Hours) + (0.05 * Previous_Scores) + (0.55 * Motivation_Level) + (0.90 * Internet_Access) + (0.48 * Tutoring_Sessions) + (0.56 * Family_Income) + (0.55 * Teacher_Quality) + (-0.05 * School_Type) + (0.50 * Peer_Influence) + (0.20 * Physical_Activity) + (-0.86 * Learning_Disabilities) + (0.49 * Parental_Education_Level) + (-0.46 * Distance_from_Home) + (0.01 * Gender) + (0.34 * Passed) + 30.49


Decision Tree Regressor - Exam Score Prediction
Attendance <= 82.5
squared_error = 15.381
samples = 5285
value = 67.215
True
False
Hours_Studied <= 20.5
squared_error = 11.364
samples = 3038
value = 65.541
Hours_Studied <= 21.5
squared_error = 11.897
samples = 2247
value = 69.479
Attendance <= 70.5
squared_error = 9.065
samples = 1617
value = 64.23
Attendance <= 72.5
squared_error = 9.801
samples = 1421
value = 67.032
Hours_Studied <= 13.5
squared_error = 10.067
samples = 1376
value = 68.358
Attendance <= 92.5
squared_error = 9.67
samples = 871
value = 71.249
Hours_Studied <= 15.5
squared_error = 8.235
samples = 765
value = 62.962
Hours_Studied <= 13.5
squared_error = 7.071
samples = 852
value = 65.369
Hours_Studied <= 29.5
squared_error = 8.953
samples = 773
value = 66.041
Hours_Studied <= 1.5
squared_error = 6.527
samples = 319
value = 66.633
Attendance <= 88.5
squared_error = 9.966
samples = 1057
value = 68.879
Hours_Studied <= 26.5
squared_error = 6.483
samples = 505
value = 70.386
Hours_Studied <= 26.5
squared_error = 11.623
samples = 366
value = 72.44
Previous_Scores <= 90.5
squared_error = 15.435
samples = 131
value = 73.878
squared_error = 42.786
samples = 34
value = 76.088


Feature Importance in Predicting Exam Score
Attendance
Hours_Studied
Previous_Scores
Parental_Involvement
Access_to_Resources
Teacher_Quality
Gender
Parental_Education_Level
Distance_from_Home
Peer_Influence
Learning_Disabilities
Internet_Access
Tutoring_Sessions
Family_Income
School_Type
Physical_Activity
Extracurricular_Activities
Motivation_Level
Sleep_Hours
Feature
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance