Analyzing Student Academic Performance

By: William Chen, Alexander Cheung, Shivtej Lakkakula, Steven Pham, Nicholas Tang

UNIVERSITY OF DELAWARE

Table Of Contents

Problem Statement

Purpose of Our Analysis

- The purpose of analyzing a dataset for student performance is to identify factors that influence academic success.
    - Specifically, the analysis aims to understand how students allocate their time across various activities (study, leisure, work, etc.) and how these allocations correlate with academic outcomes, such as exam scores.
- By gaining insights into these patterns, educational strategies can be improved to support student success, enhance time management practices, and identify areas where interventions may be most effective.

Data Context

Limitations

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

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	76
School_Type	0
Peer_Influence	0
Physical_Activity	0
Learning_Disabilities	0
Parental_Education_Level	90
Distance_from_Home	0
Gender	0
Exam_Score	0

dtype: int64

Data Dictionary

Strings:

[‘Parental_Involvement', 'Access_to_Resources', 'Extracurricular_Activities', 'Motivation_Level', 'Internet_Access', 'Family_Income', 'Teacher_Quality', 'School_Type', 'Peer_Influence', 'Learning_Disabilities', 'Parental_Education_Level', 'Distance_from_Home', 'Gender']

Integers:

['Hours_Studied', 'Attendance', 'Sleep_Hours', 'Previous_Scores', 'Tutoring_Sessions', 'Physical_Activity', 'Exam_Score']

Data Preparation Process

Research Questions

1) **How does the time allocated to different types of activities (study, leisure, work, etc.) influence student success?**
    - This would help uncover the balance between study time and other commitments, and its impact on academic results
2) **How does a student's attendance and motivation affect student success and learning outcomes?**
    - Investigating the role of mental capacity and performance can lead to recommendations on student lifestyle adjustments.
3) **How does the distance from home to school impact student performance, and what measures can mitigate potential negative effects (e.g., fatigue or reduced study time)?**
    - This can lead to insights on how transportation or proximity to school affects academic success.

Data Cleaning & Preprocessing

- Data Cleaning
  - Identified missing or inconsistent data type values

Preprocessing

```
df.loc[df['Motivation_Level'] == 'Low', 'MotivationCode'] = 0
df.loc[df['Motivation_Level'] == 'Medium', 'MotivationCode'] = 1
df.loc[df['Motivation_Level'] == 'High', 'MotivationCode'] = 2
```

```
dfr_1.loc[df1['Extracurricular_Activities'] == 'No', 'Extracurricular'] = 0
dfr_1.loc[df1['Extracurricular_Activities'] == 'Yes', 'Extracurricular'] = 1
```

```
df.loc[df['Exam_Score'] < 70, 'ExamCode'] = 0
df.loc[df['Exam_Score'] >= 70, 'ExamCode'] = 1
```

Exploratory Data Analysis & Tools

Applications

- Python
- Jupyter Notebook
- Tableau

Libraries

- Pandas
- Seaborn
- Sklearn
- Matplot

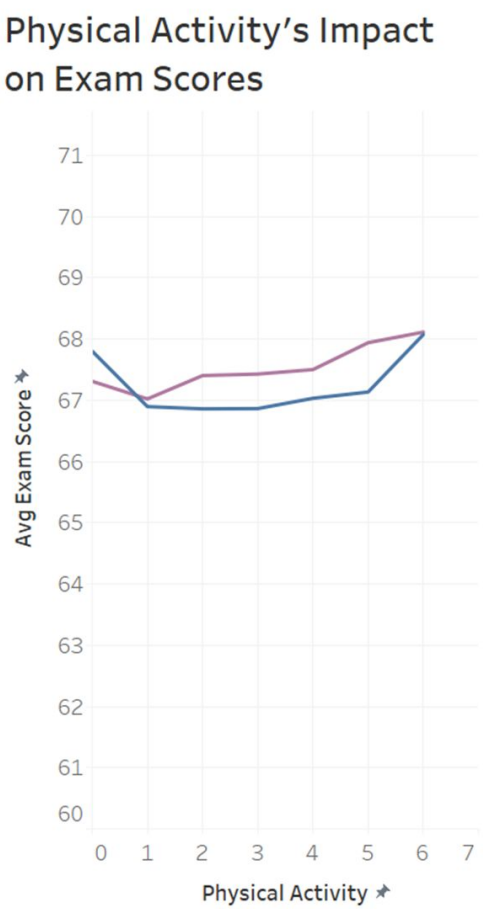
Research Question Answers

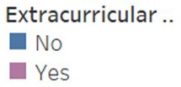
UNIVERSITY OF DELAWARE

Research Question 1

How does the time allocated to different types of activities (study, leisure, work, etc.) influence student success?

UNIVERSITY OF DELAWARE



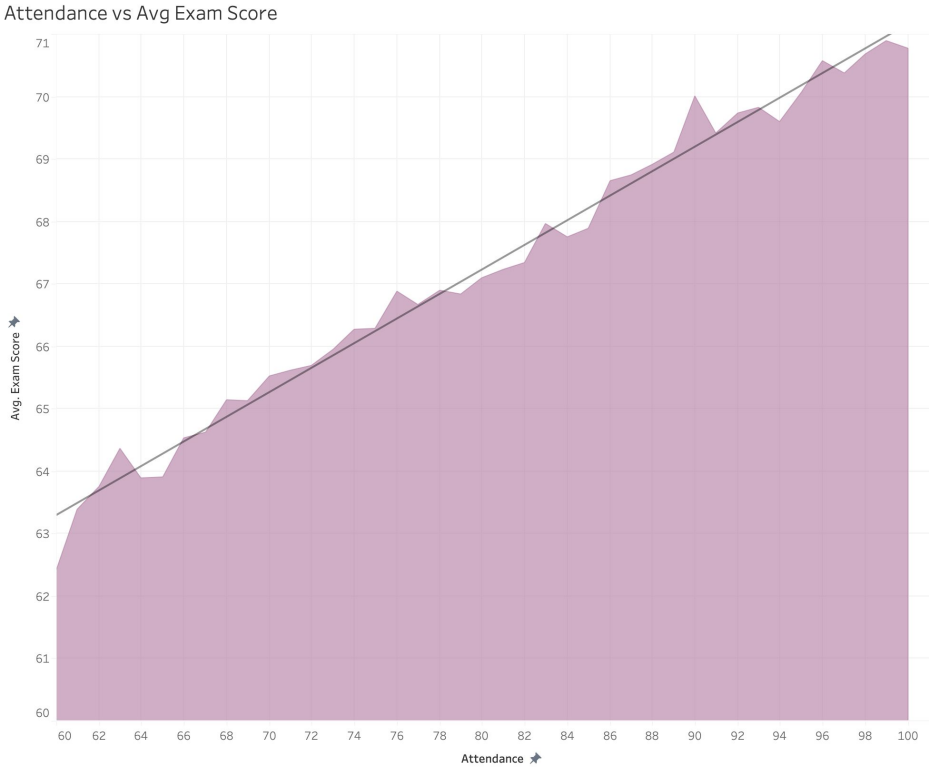


- Data divided into 2 categories
  - EC Participation vs No ECs
- Both physical activity and tutoring sessions will boost exam scores
  - However, too many tutoring sessions may cause burn out
- Overall impact of tutoring sessions is higher than Physical Activity
  - Around a 7.6% boost in scores from 0 sessions to 5 tutoring sessions

Research Question 2

How does a student's attendance and motivation affect student success and learning outcomes?

Attendance VS. AVG Exam Score



The plot of average of Exam Score for Attendance.

AVG Exam score is 67.089%

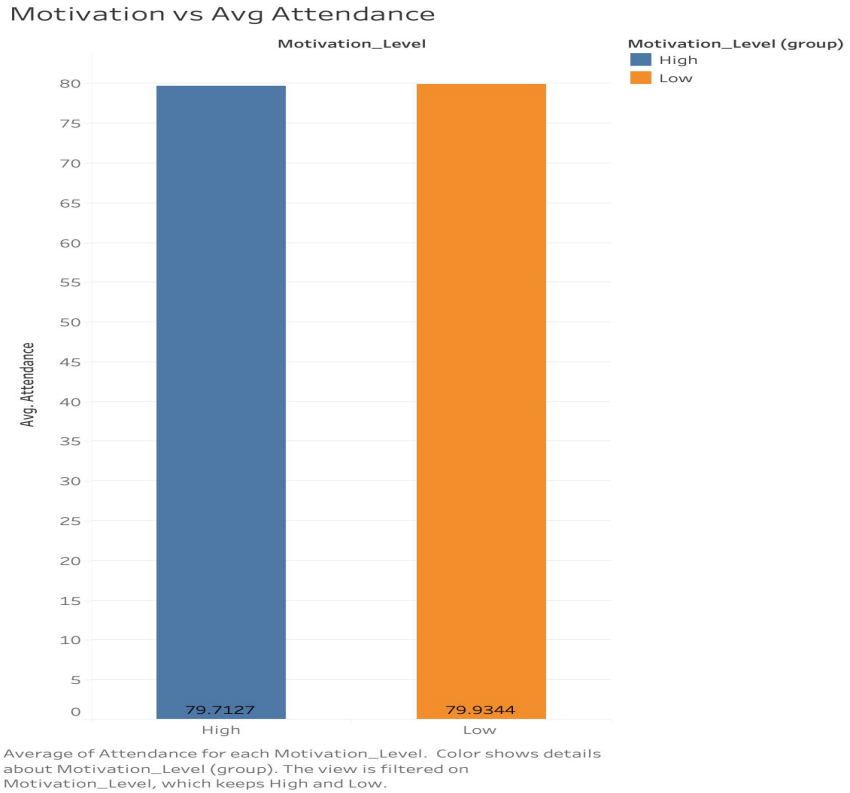
As attendance increases, the average exam score also increases, showing a clear correlation.

Positive Trend Line

At lower attendance levels (around 60–70%), average exam scores are between 60–65.

Higher attendance (80–100%), scores range from 67–71.

Motivation vs. AVG Attendance



Average Attendance Levels:

- The average attendance for the High Motivation Level group is 79.71%.
- The average attendance for the Low Motivation Level group is 79.93%.

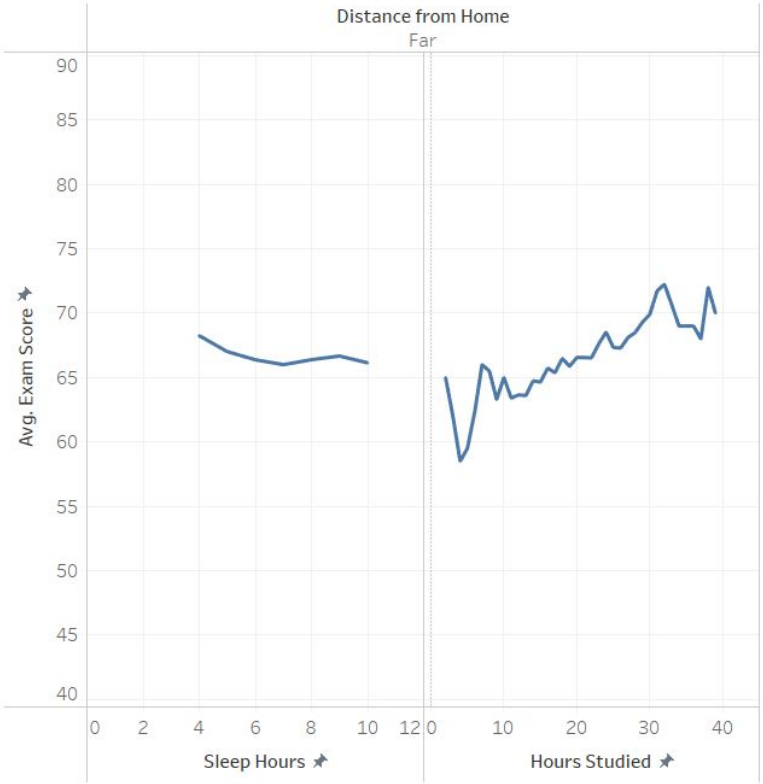
Key Insight:

- The difference in average attendance between the two groups is negligible. Despite the expectation that high motivation would correlate with significantly better attendance, this data suggests both groups maintain nearly identical attendance levels.

Research Question 3

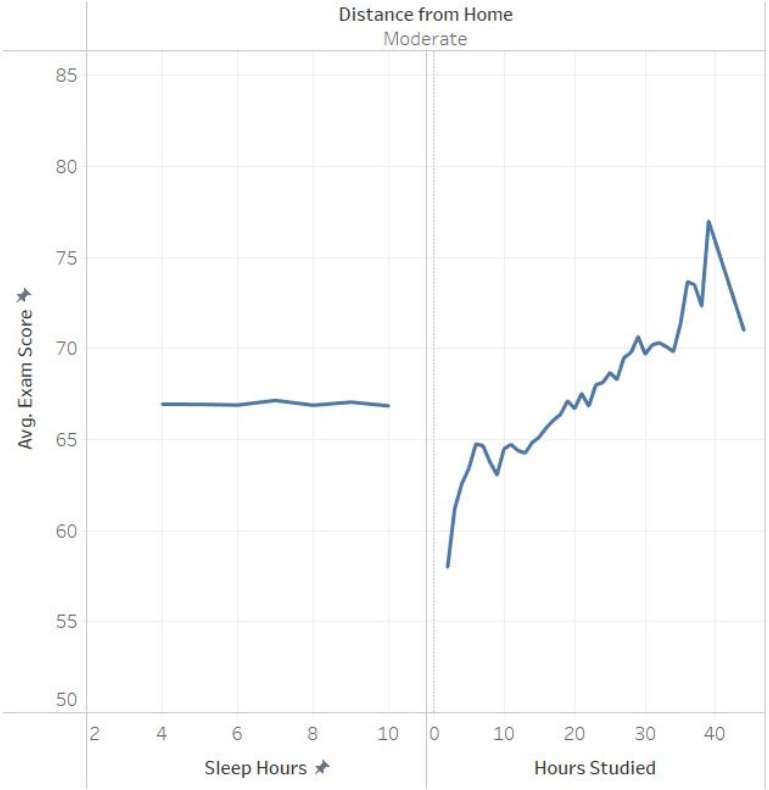
How does the distance from home to school impact student performance, and what measures can mitigate potential negative effects (e.g., fatigue or reduced study time)?

The relationship between the average exam score and hours studied and hours of sleep with distance from home-far



- There is a noticeable increase in average exam scores with increasing study hours, indicating a positive correlation between study time and performance for students living far from home.
- Sleep hours, however, seem to have a less consistent relationship with exam scores, as shown by the relatively flat line, suggesting that sleep does not significantly impact exam scores for these students.

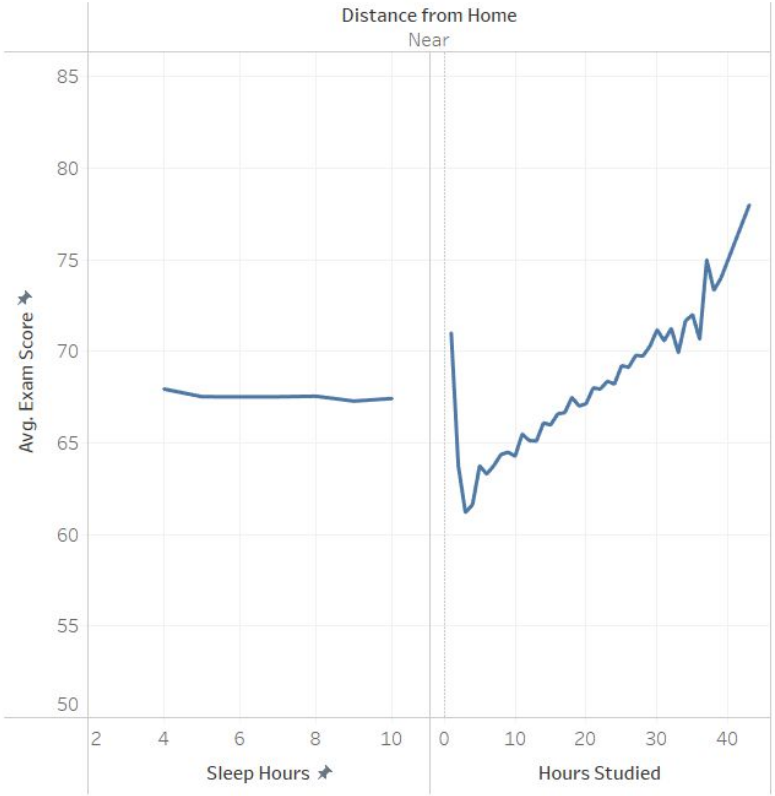
The relationship between the average exam score and hours studied and hours of sleep with distance from home-moderate



- There is a dramatic rise in average exam scores as study hours increase, peaking at around 40 hours before seeing a sharp decline. This could indicate that a moderate amount of studying is optimal, with too much potentially leading to burnout or decreased efficiency.

- Similar to the "far" group, sleep does not display a strong correlation with exam scores, maintaining a steady trend across different sleep hours.

The relationship between the average exam score and hours studied and hours of sleep with distance from home-near



- There is a steep increase in exam scores with more hours studied, suggesting that students who live near their place of education might find more effective or intensive study patterns.
- As with the other groups, the relationship between sleep hours and exam scores remains relatively flat, implying little to no impact of sleep on exam performance for these students.

Models

Random Forests Model (Q1)

Random Forests Tuning

```
With n_estimators=50 and with max_depth=1, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=50 and with max_depth=2, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=50 and with max_depth=3, the Cross validation scores mean is 0.7538206027632615:
With n_estimators=50 and with max_depth=4, the Cross validation scores mean is 0.7518543441429832:
With n_estimators=50 and with max_depth=5, the Cross validation scores mean is 0.7522910253656907:
With n_estimators=100 and with max_depth=1, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=100 and with max_depth=2, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=100 and with max_depth=3, the Cross validation scores mean is 0.7538206027632615:
With n_estimators=100 and with max_depth=4, the Cross validation scores mean is 0.7531655809292004:
With n_estimators=100 and with max_depth=5, the Cross validation scores mean is 0.7514171856730378:
With n_estimators=150 and with max_depth=1, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=150 and with max_depth=2, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=150 and with max_depth=3, the Cross validation scores mean is 0.7538206027632615:
With n_estimators=150 and with max_depth=4, the Cross validation scores mean is 0.7531655809292004:
With n_estimators=150 and with max_depth=5, the Cross validation scores mean is 0.7511988450616842:
With n_estimators=200 and with max_depth=1, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=200 and with max_depth=2, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=200 and with max_depth=3, the Cross validation scores mean is 0.7538206027632615:
With n_estimators=200 and with max_depth=4, the Cross validation scores mean is 0.7531655809292004:
With n_estimators=200 and with max_depth=5, the Cross validation scores mean is 0.7516360035316295:
With n_estimators=250 and with max_depth=1, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=250 and with max_depth=2, the Cross validation scores mean is 0.7542577612332069:
With n_estimators=250 and with max_depth=3, the Cross validation scores mean is 0.7538206027632615:
With n_estimators=250 and with max_depth=4, the Cross validation scores mean is 0.7531655809292004:
With n_estimators=250 and with max_depth=5, the Cross validation scores mean is 0.7518543441429832:

Best parameters: {'n_estimators': 50, 'max_depth': 1}
Best mean score: 0.7542577612332069
```

Training/Testing Data (70/30)

Extracurricular','Tutoring_Sessions','Physical_Activity']]

- Extracurricular: 34%
- Tutoring_Sessions: 40%
- Physical Activity: 26%

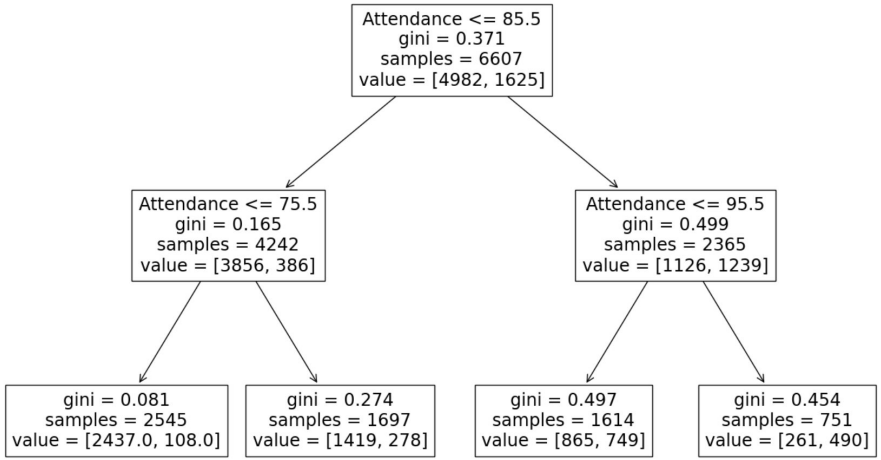
```
accuracy = clf.score(X_train,Y_train) # make it a percentage and round to 2 places
print("The Random forest is {:.2f}% accurate for train dataset".format(accuracy*100))
```

```
The Random forest is 75.43% accurate for train dataset
```

```
accuracy=clf.score(X_test,Y_test)
print("The Random forest is {:.2f}% accurate for test dataset".format(accuracy*100))
```

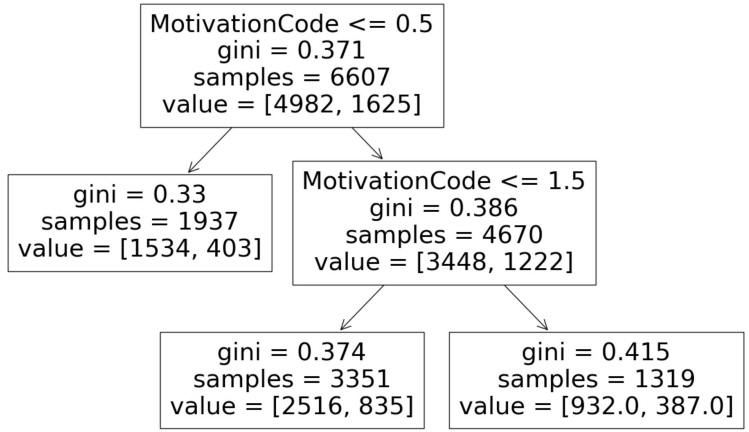
```
The Random forest is 74.79% accurate for test dataset
```

Decision Trees (Q2)




```
accuracy = clf.score(X,Y)
print("The decision tree is {:.2f}% accurate.".format(accuracy*100))
```

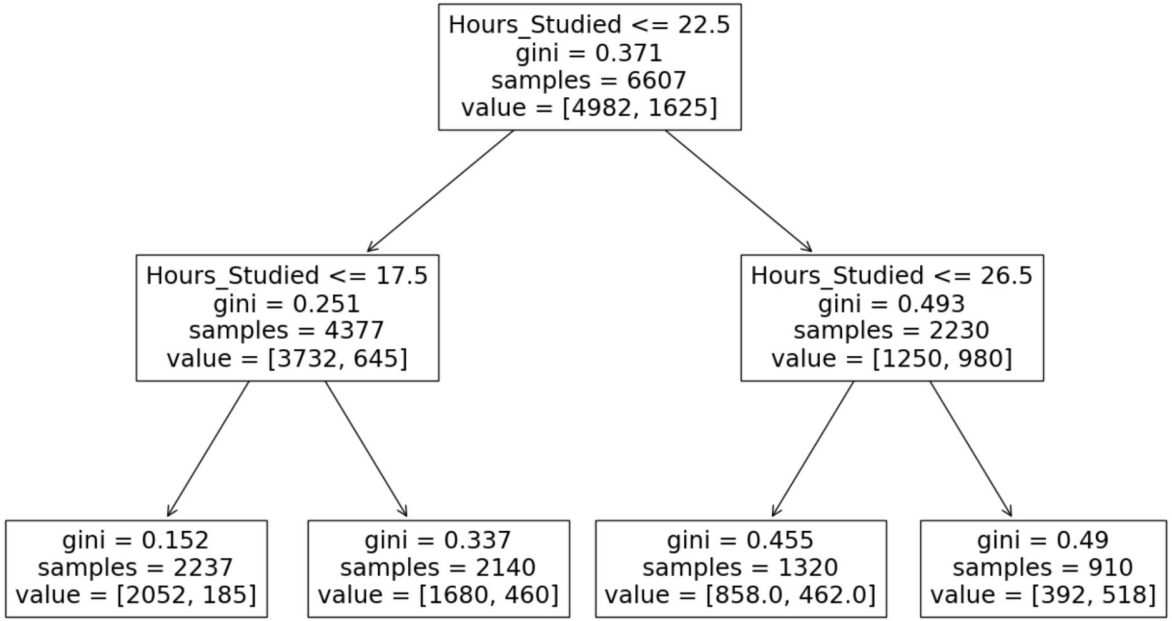
```
The decision tree is 78.87% accurate.
```




```
accuracy = clf.score(X,Y)
print("The decision tree is {:.2f}% accurate.".format(accuracy*100))
```

```
The decision tree is 75.40% accurate.
```

Decision Trees (Q3)




```
accuracy = clf.score(X,Y)
print("The decision tree is {:.2f}% accurate.".format(accuracy*100))
```

```
The decision tree is 77.31% accurate.
```

Results And Recommendations

Results

- Attendance, the number of hours studied, and tutoring sessions were factors that had the biggest influence on student performance
- Distance from home, sleep hours, extracurriculars, and physical activity did not have a significant influence on student performance

Recommendations

- If students want to increase their exam scores, effective ways of doing so would include getting to class, attending tutoring sessions, and focusing on studying
- Additional Recommendations:
  - Pair Attendance With Engagement
  - Join Study Groups
  - Incorporate Healthy Habits
    - Social Connections, Exercise, Time Management

Questions?