CapStone Project

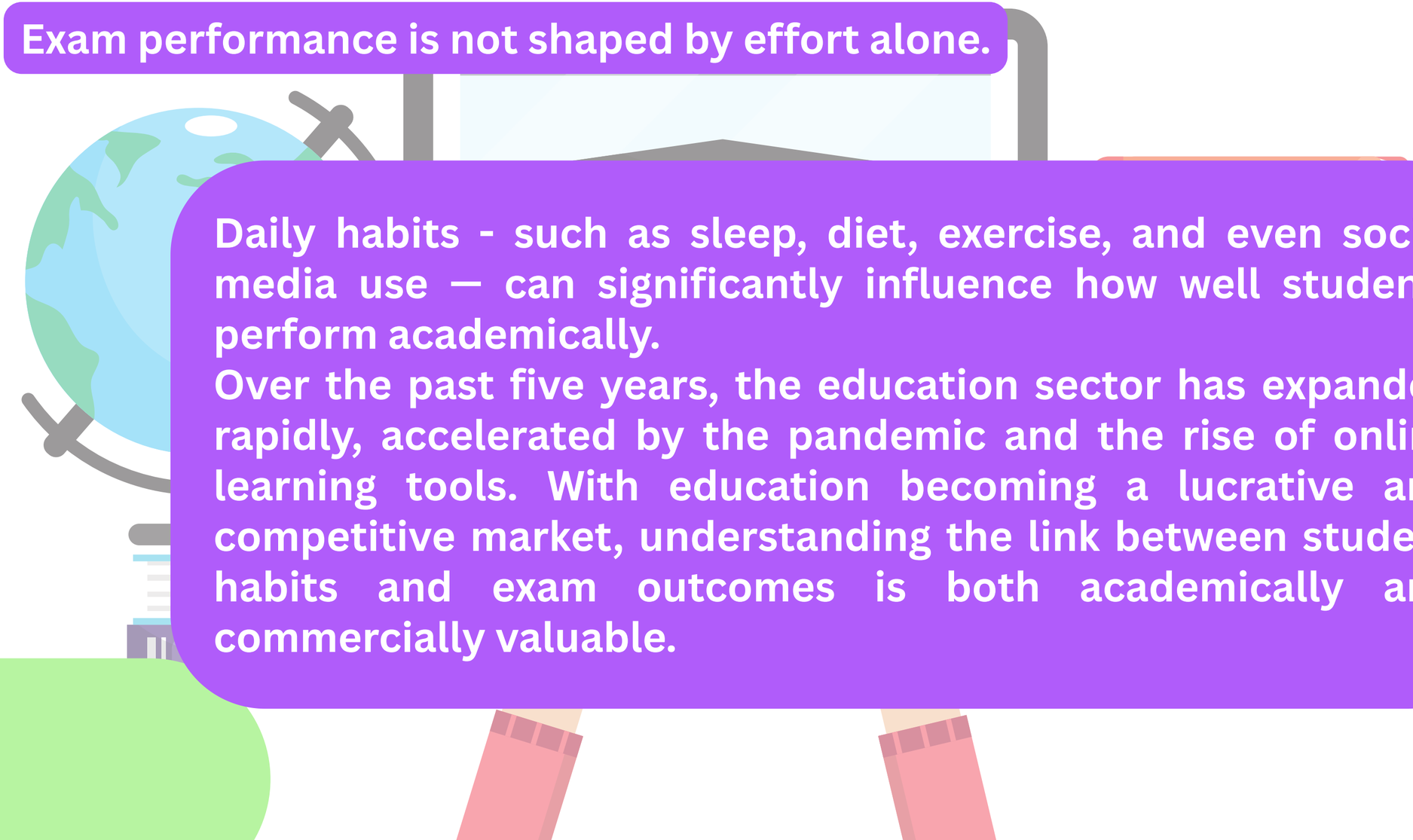
NICHE MARKET RESEARCH: TOP STUDENT HABITS EFFECTING EXAM SCORES

By Lee-Anne van der Merwe

SUMMARY

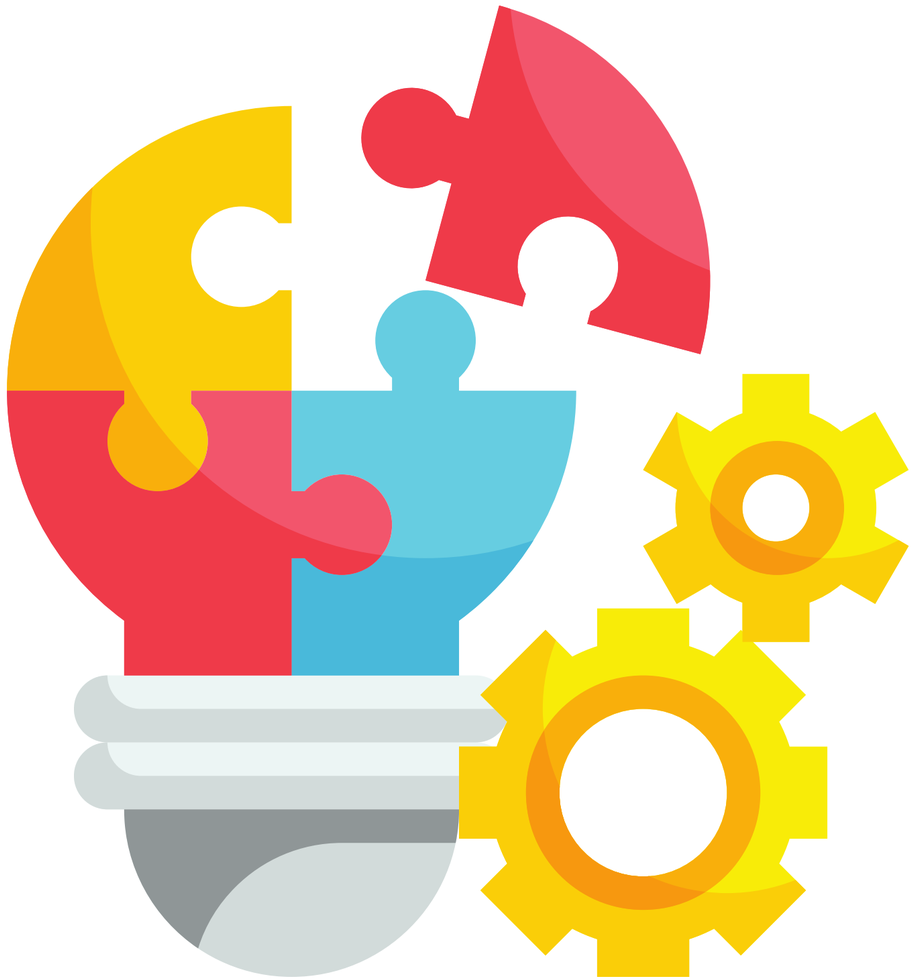
INTRODUCTION

Exam performance is not shaped by effort alone.



Daily habits - such as sleep, diet, exercise, and even social media use — can significantly influence how well students perform academically.

Over the past five years, the education sector has expanded rapidly, accelerated by the pandemic and the rise of online learning tools. With education becoming a lucrative and competitive market, understanding the link between student habits and exam outcomes is both academically and commercially valuable.



THE PROJECT GOAL

This Project Focuses on Uncovering Which Student Habits Affect Exam Scores the Most.

By identifying these drivers, we not only gain insights into student success but also highlight opportunities for innovation in the education space — from wellness apps to study-support platforms.

DATA PREPARATION

1

Data Cleaning

1. Replaced NANs with Mean values
2. Dropped duplicates

2

Data Filtering

1. Selected data where hours of time spent does not exceed 24 hours.
2. Filtered to ages between 16-24.

3

Data Normalisation

1. Replaced spaces with ‘_’
2. Converted everything to lowercase

student_habits_EDA.CSV

4

Data Standardisation / Scaling

1. Encoded categorical data
2. Scaled columns selected for machine learning

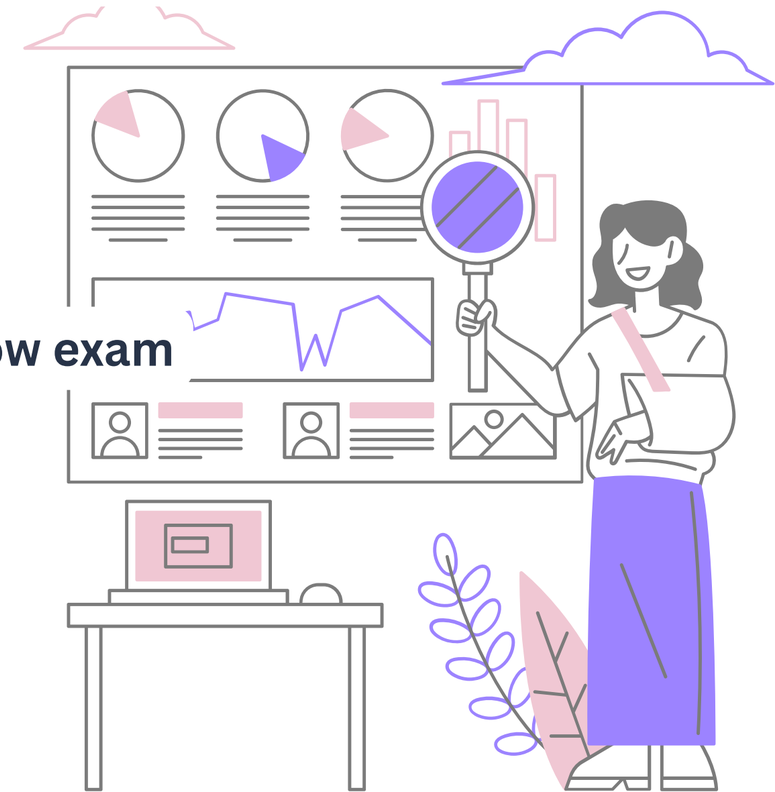
Encoded_student_habits.CSV

EXPLORATORY DATA ANALYSIS

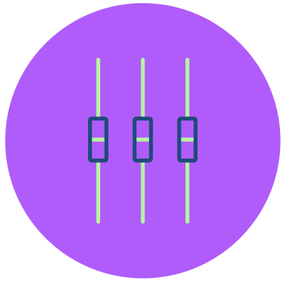
The effect of different habits recorded in the data was explored using various graphical methods, limited to “Seaborn, Matplotlib, and plotly”.

Scatter Plot → “Does more sleep, studying, and regular workouts add up to higher exam scores?”

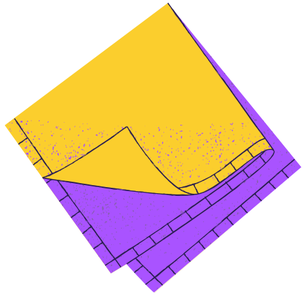
Stacked Bar Plot → “Same internet, different story: how exam scores and Wi-Fi quality stack up across genders.”



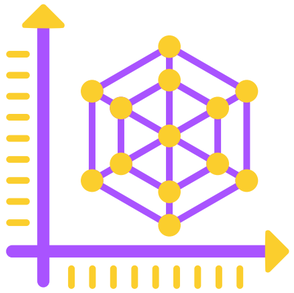
EXPLORATORY DATA ANALYSIS



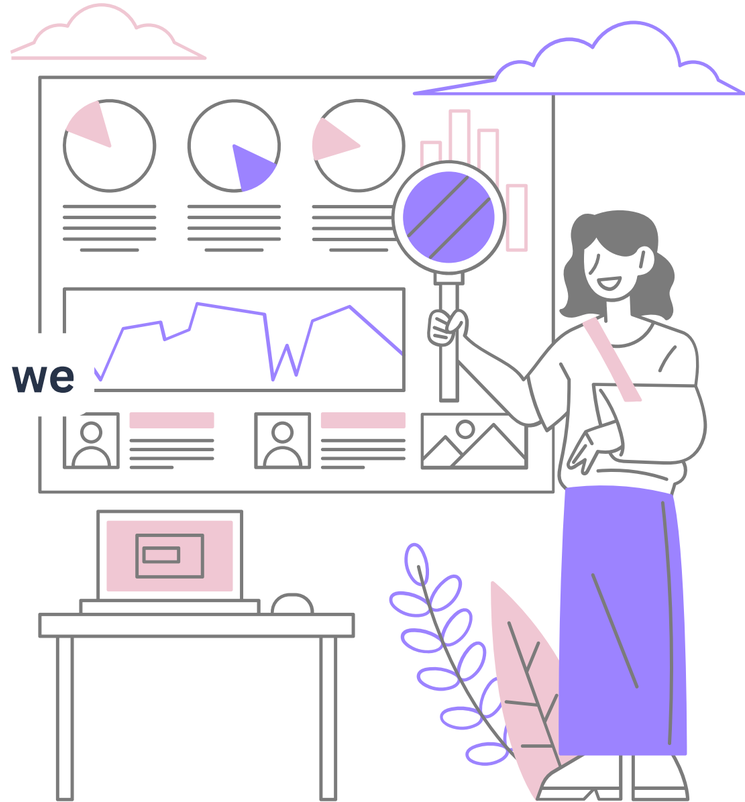
Violin Plot → “Mental health — here’s how it plays out across age groups.”



Heatmap → “A map of hidden patterns in the data.”



Radar Chart → “A day in the life, visualized: how we spread our time across daily habits.”



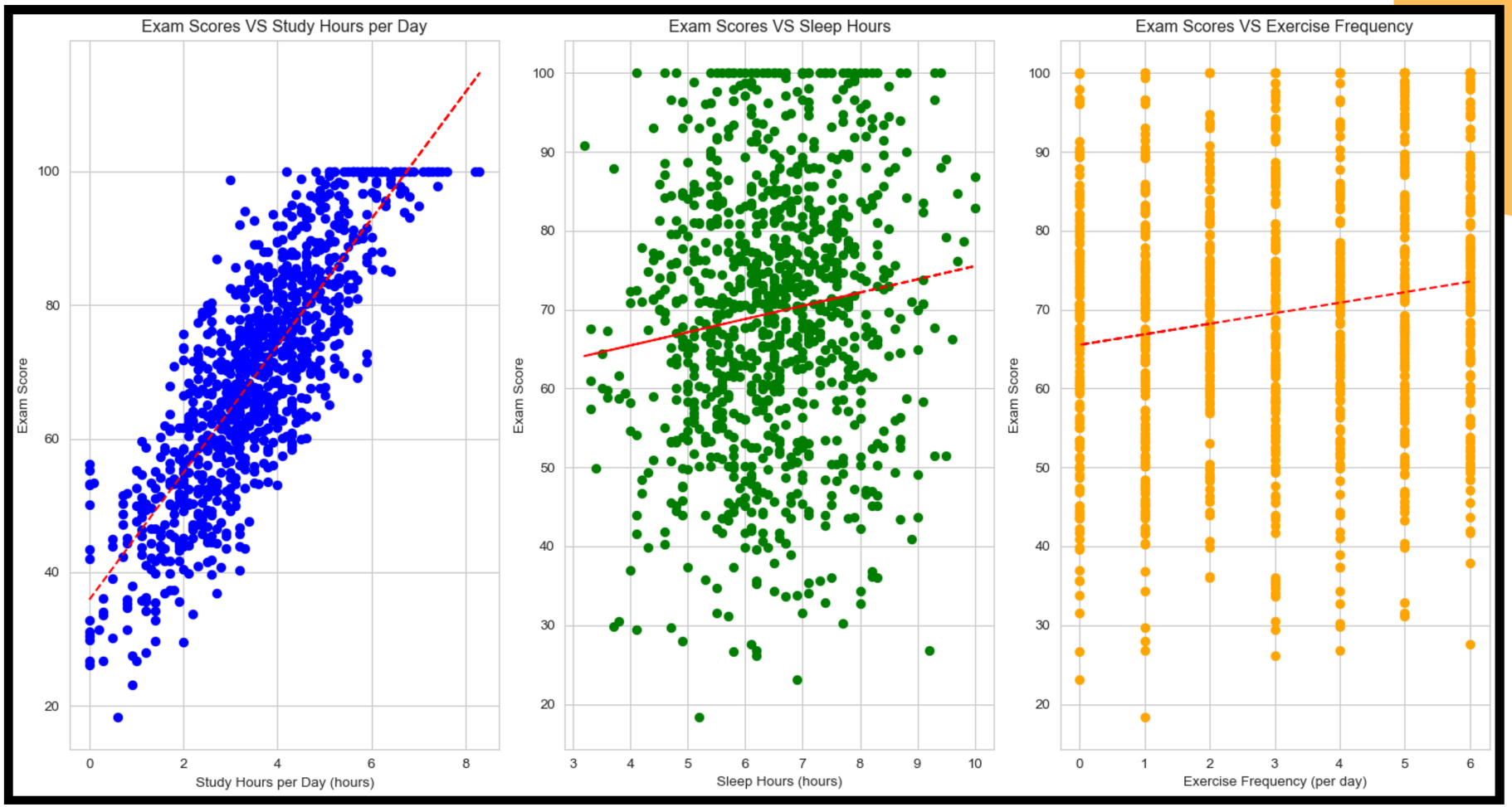
DOES MORE SLEEP, STUDYING, AND REGULAR WORKOUTS ADD UP TO HIGHER EXAM SCORES?

Correlation?

A comparison of the scatterplots and correlation values reveals that the time spent **studying has the strongest correlation with exam scores** at 0.83.

Meanwhile, exercise frequency and sleep have **very little affect**.

When aiming for high exam scores, students may incorporate different study methods, time management devices or innovative study material - potential business.

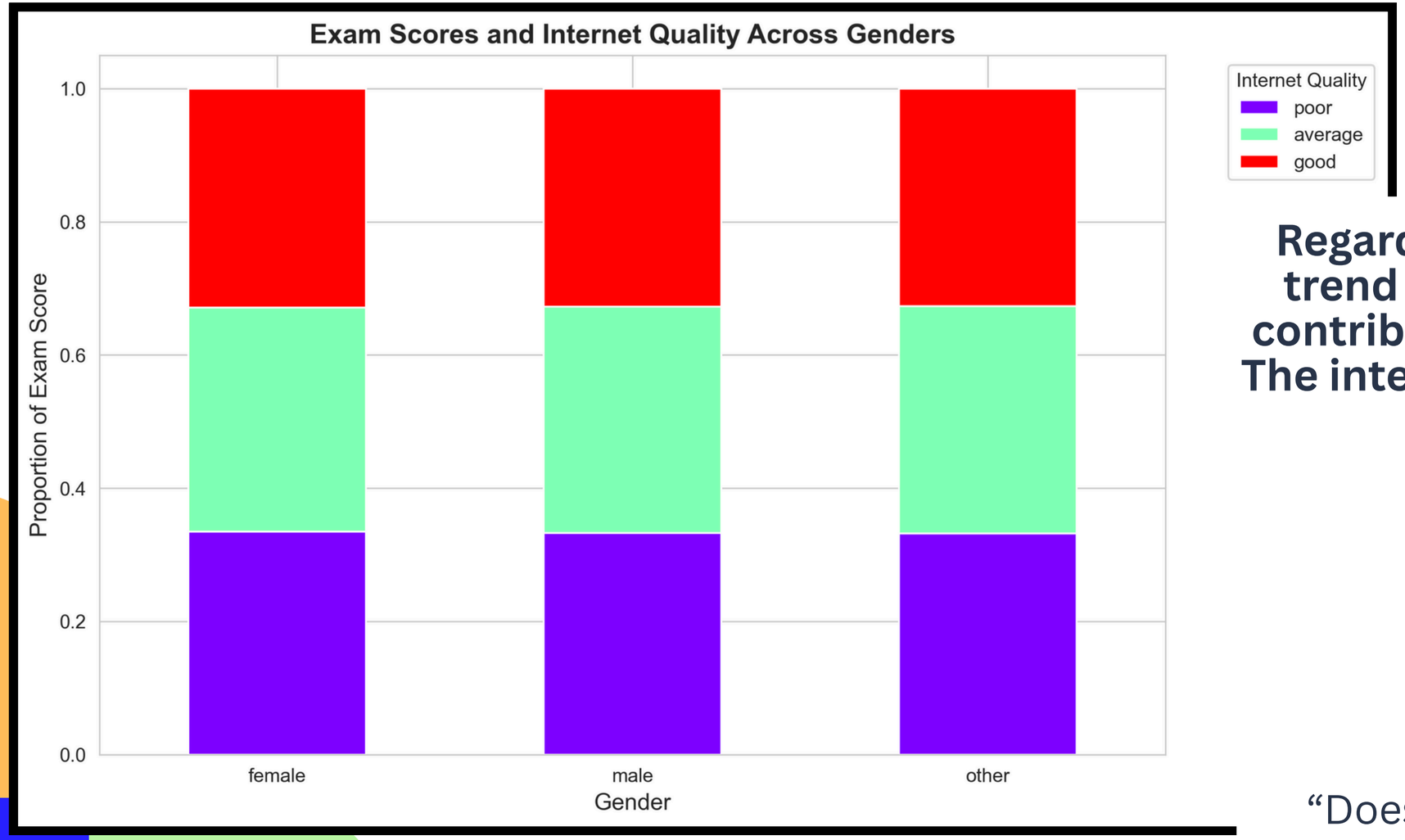


Correlation value: 0.83

Correlation value: 0.12

Correlation value: 0.16

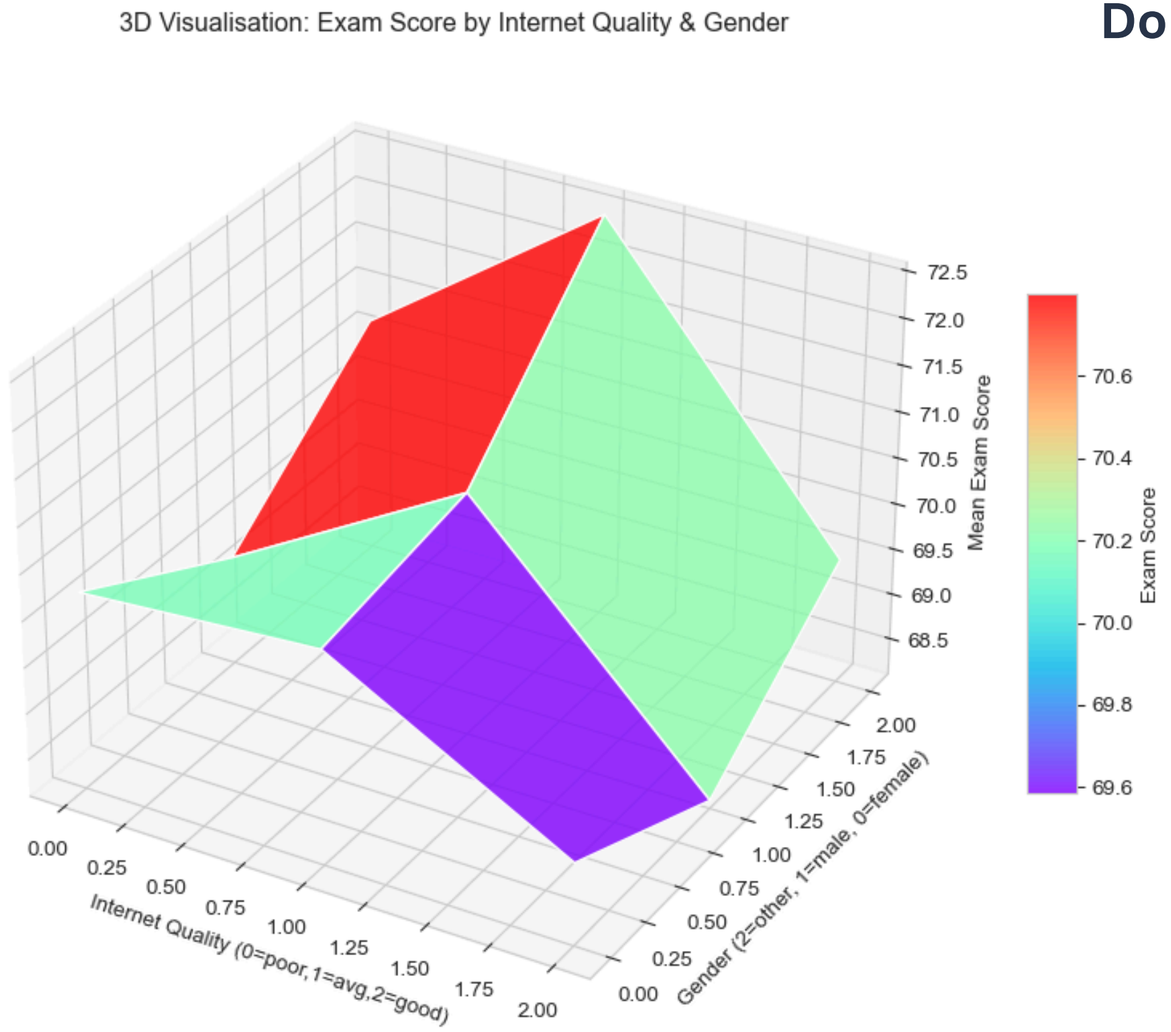
SAME INTERNET, DIFFERENT STORY: HOW EXAM SCORES AND WI-FI QUALITY STACK UP ACROSS GENDERS.



Regardless of gender, the same trend appears: improved Wi-Fi contributes to most exam scores. The internet gap affects all groups equally.

“Does better Wifi lead to better exam scores?”

SAME INTERNET, DIFFERENT STORY: HOW EXAM SCORES AND WI-FI QUALITY STACK UP ACROSS GENDERS.



Do we focus on the narrow 2% margin or the broader 10% scale?

The 3D surface illustrates that exam scores generally rise as internet quality improves, across all genders, before showing a slight decline at the highest levels.

Importantly, the gap between male and female exam scores at different levels of internet quality remains under 2%—a difference so small it is practically negligible.

Overall, this suggests that while internet quality does correlate with higher average exam scores, it does not explain causation.

MENTAL HEALTH — HERE'S HOW IT PLAYS OUT ACROSS AGE GROUPS.

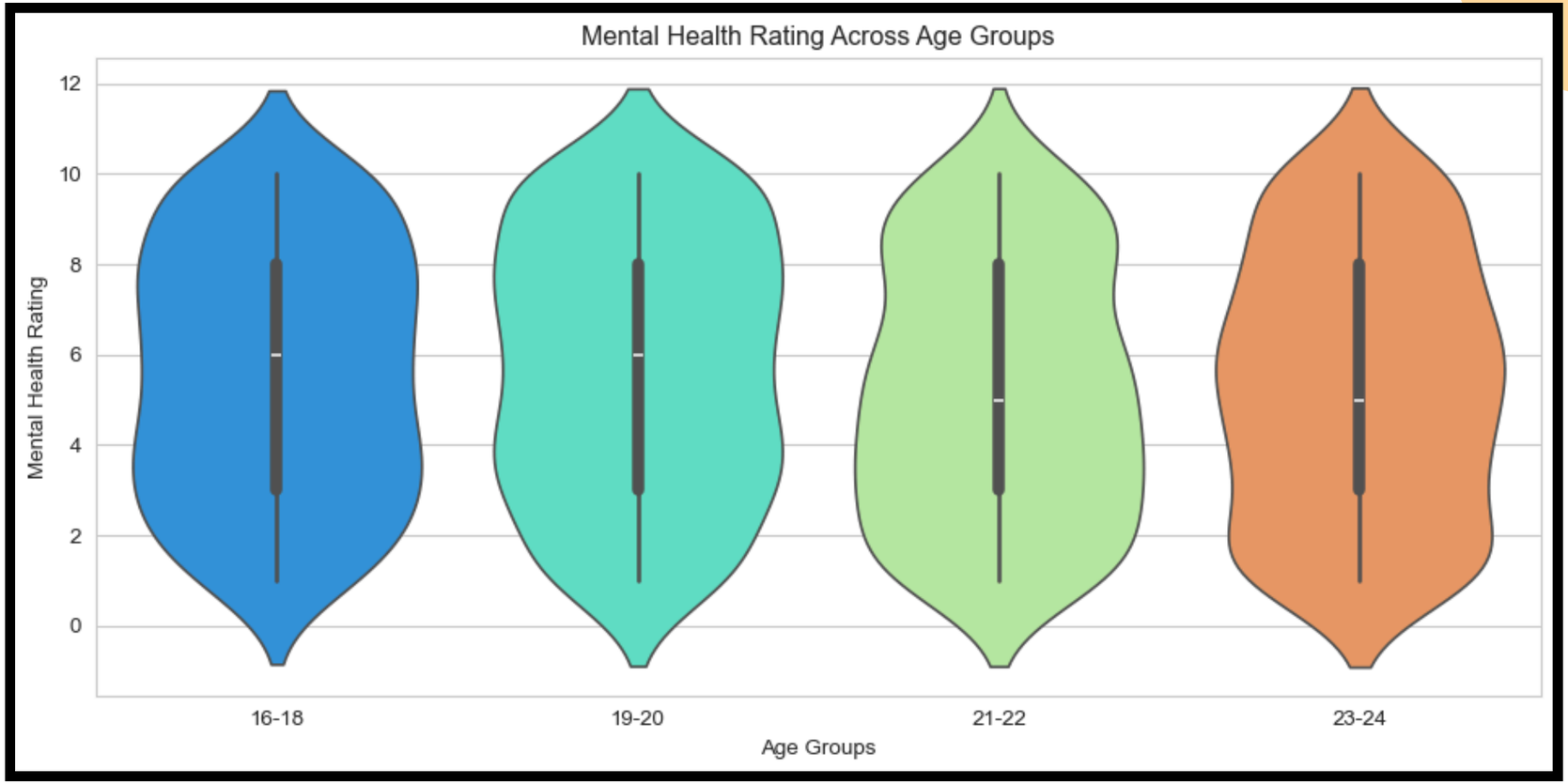
The median mental health rating for both age groups is a 6, with the interquartile ranges between a rating of 3-8. No extreme outliers were present in either groups.

Age 16-18

Very Symmetrical but slightly bimodal data distribution.

Age 19-20

Slightly asymmetrical data distribution, heavier on top and narrower at the bottom.



MENTAL HEALTH — HERE'S HOW IT PLAYS OUT ACROSS AGE GROUPS.

The 19–20 year age group shows a greater concentration of responses at the higher end of the mental health ratings, indicating that more individuals in this group report stronger mental health compared to lower ratings.

In contrast, the 16–18 year age group displays a more balanced distribution, with similar amounts of data across both higher and lower mental health ratings, suggesting less distinction between the two ends of the scale.

MENTAL HEALTH — HERE'S HOW IT PLAYS OUT ACROSS AGE GROUPS.

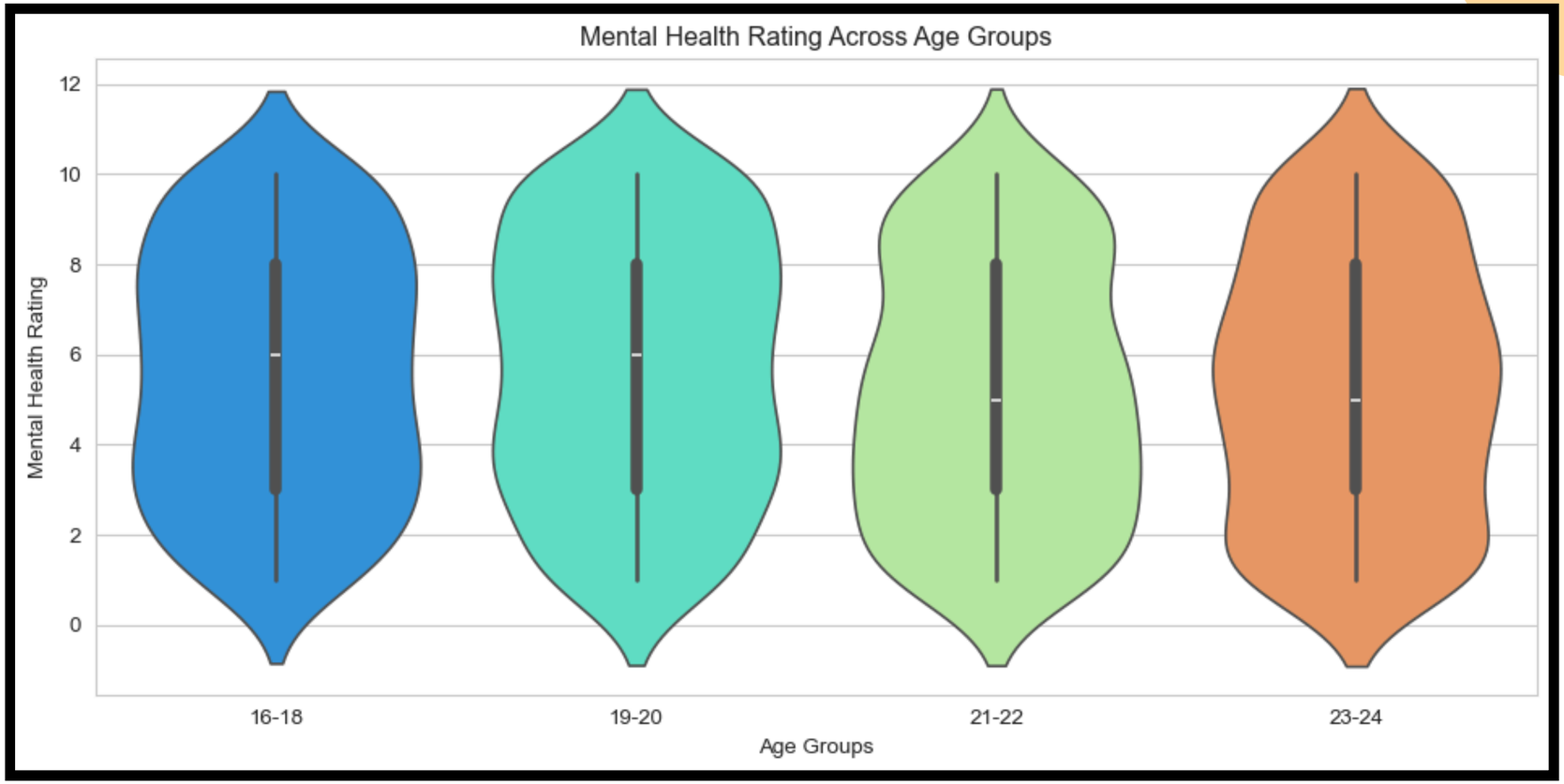
The median for both datasets is lower compared to the other age categories at a mental health rating of 5. Both categories also have slightly narrower distributions.

Age 21-22

The distribution is slightly narrower at the top compared to the bottom, suggesting more low mental health ratings among this age category.

Age 23-24

The distribution is slightly wider at the middle, suggesting more mental health ratings near the median among this age category.

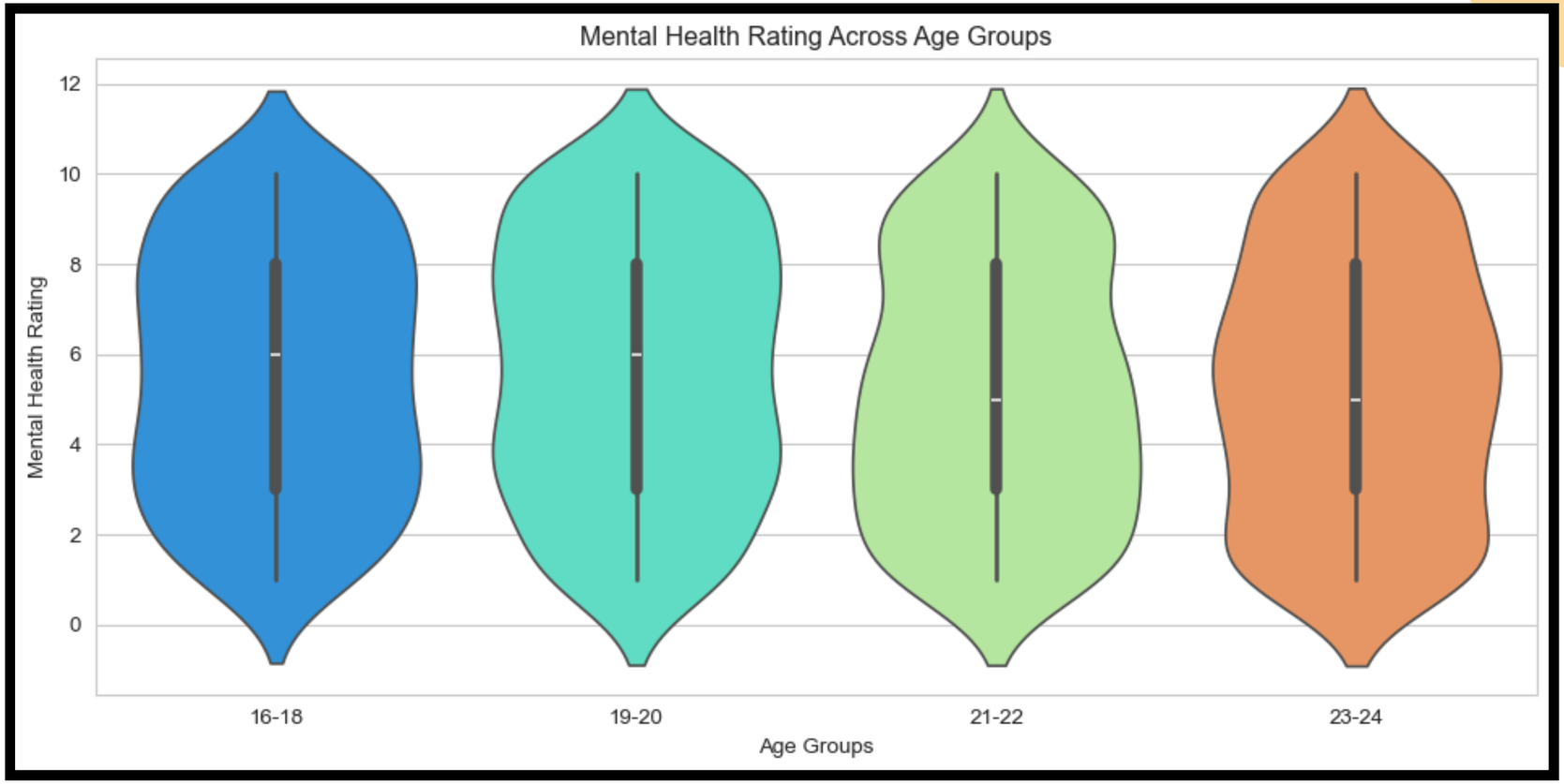


MENTAL HEALTH — HERE'S HOW IT PLAYS OUT ACROSS AGE GROUPS.

Overall, mental health ratings appear higher among the younger age categories, though the distributions show considerable variability. In contrast, the older age categories display more stability, with less variation across the distribution, even though their median ratings decline slightly with age.

Insight

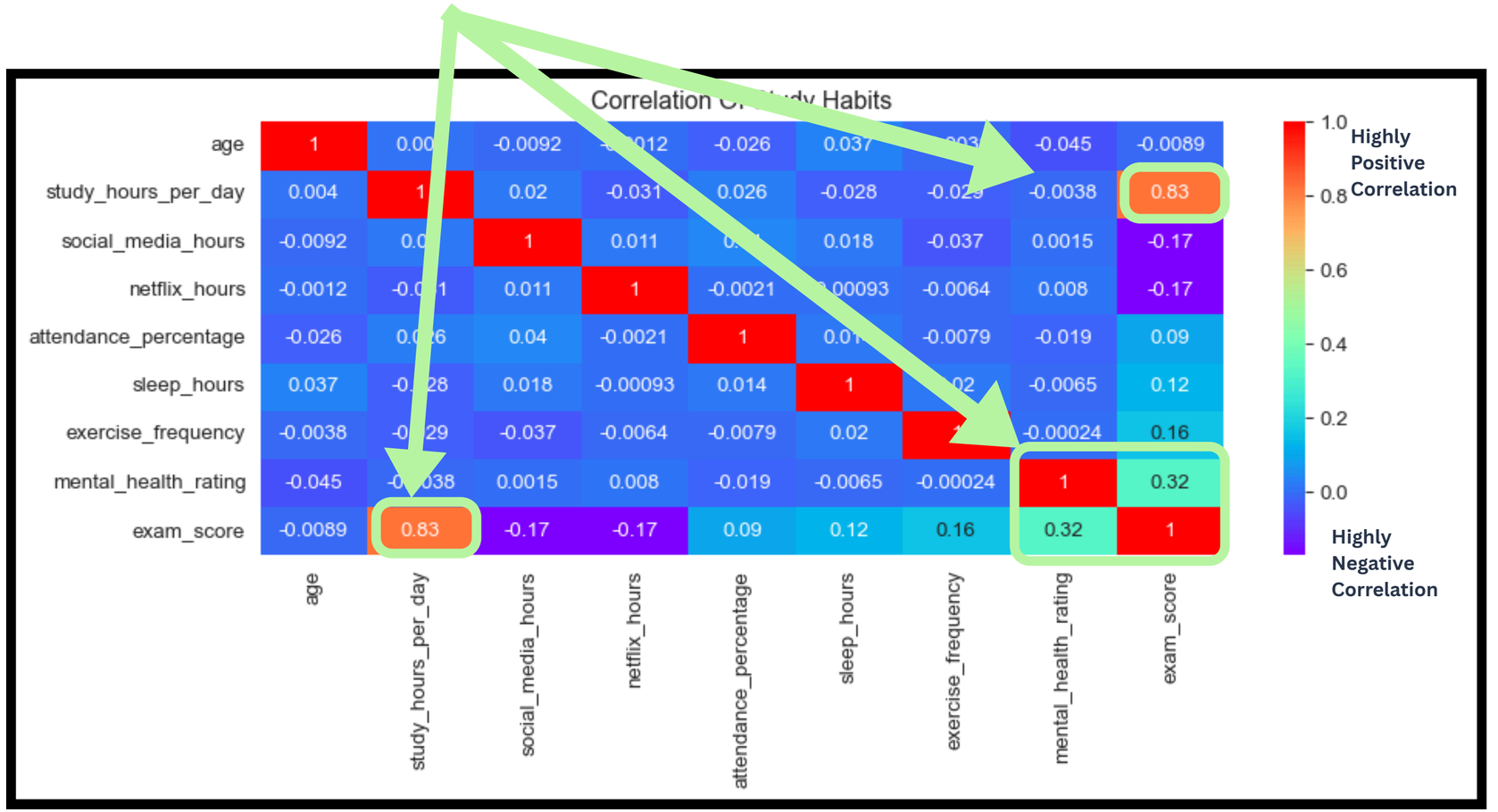
Older categories may experience more stress with added responsibilities like a job search, family responsibilities, or financial pressure (student loan repayment), explaining a lower median.



MULTIVARIATE DATA ANALYSIS: “A MAP OF HIDDEN PATTERNS IN THE DATA.”

THE GOOD

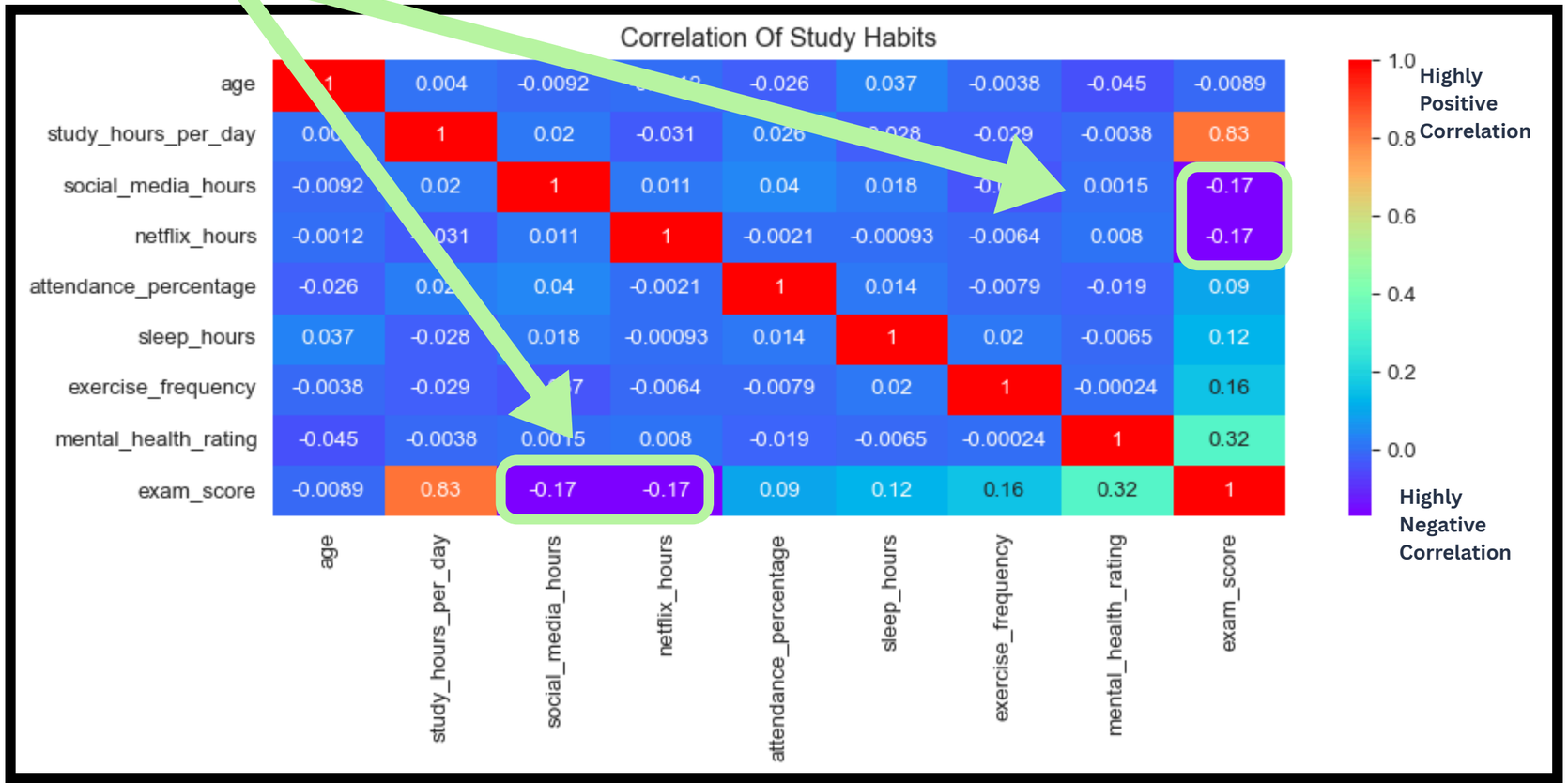
The most positive correlation was found between study hours and exam scores, followed by the positive correlation between mental health rating and exam scores. This suggests that if mental health affects exam scores it is on a slightly lower scale than time spent studying.



MULTIVARIATE DATA ANALYSIS: “A MAP OF HIDDEN PATTERNS IN THE DATA.”

THE BAD

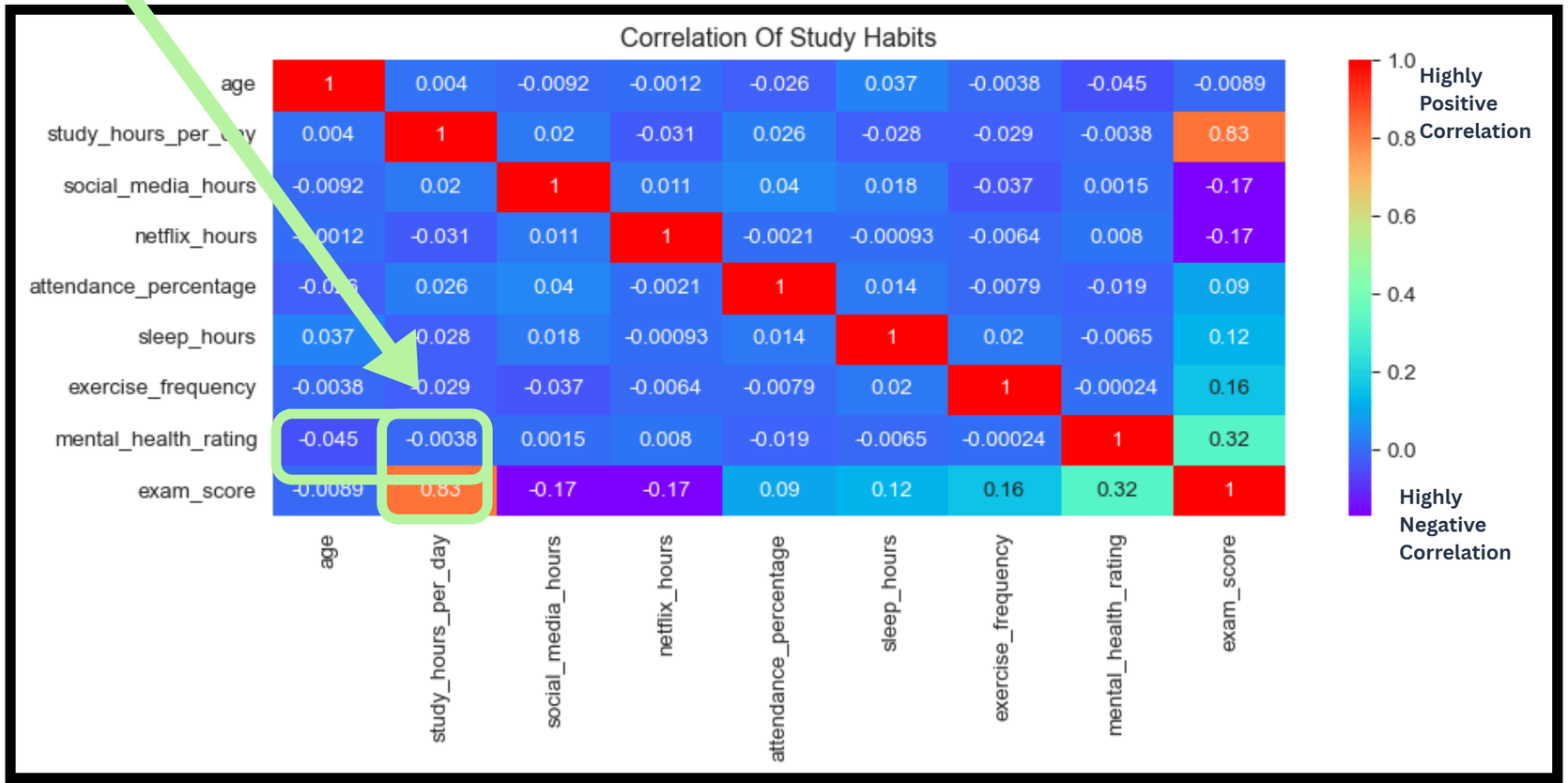
Two distinct clusters of negative correlation coefficients are evident in the data. The most prominent of these is the negative relationship between exam scores and time spent on Netflix and social media, suggesting that increased usage of these platforms is associated with lower academic performance.



MULTIVARIATE DATA ANALYSIS: “A MAP OF HIDDEN PATTERNS IN THE DATA.”

THE INSIGHT

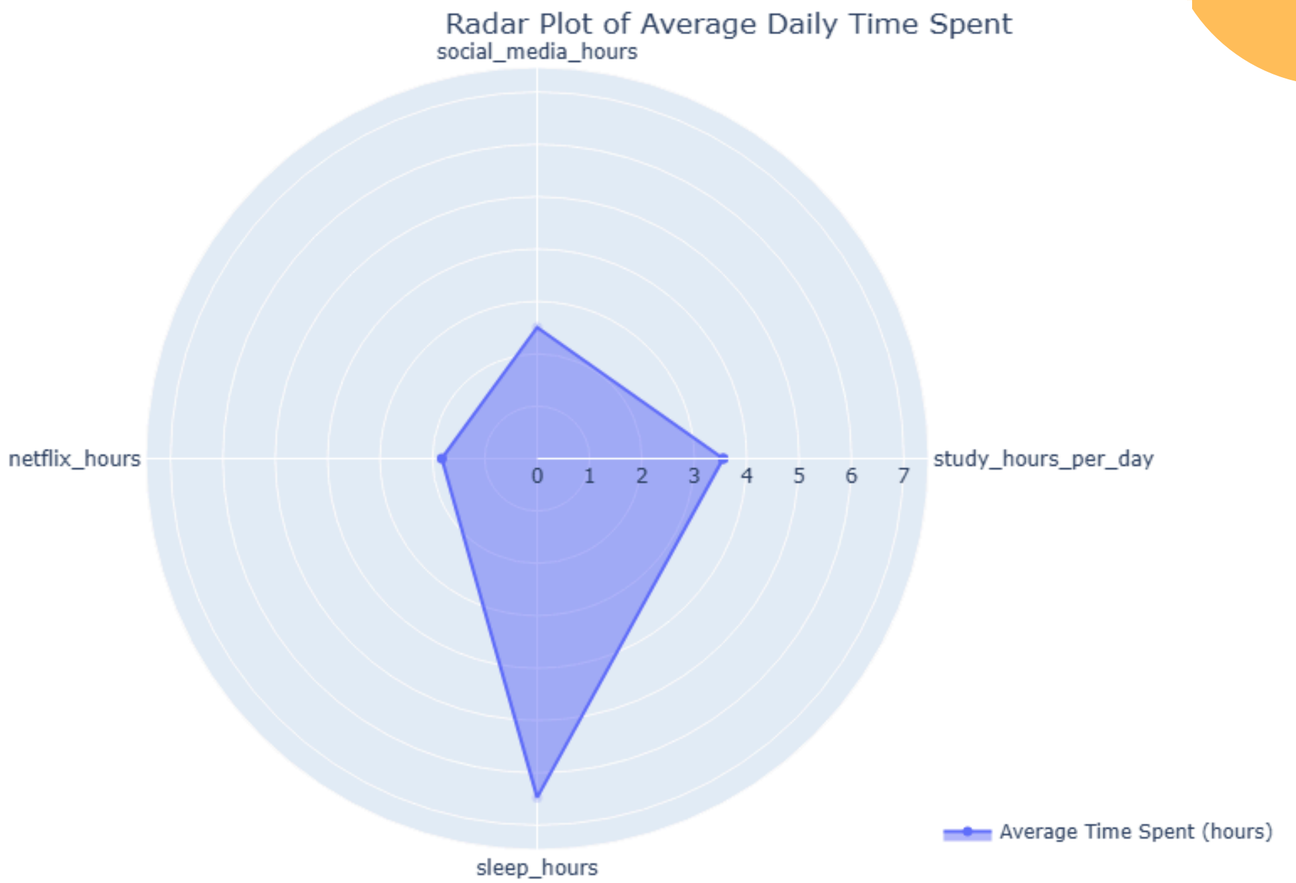
Interestingly, mental health ratings show a slight negative correlation with time spent studying. This suggests that maintaining a balance may be necessary in order to achieve the positive associations observed when each factor is considered individually.



A DAY IN THE LIFE, VISUALIZED: HOW WE SPREAD OUR TIME ACROSS DAILY HABITS

The largest share of time is allocated to sleep, averaging 6.5 hours per day, followed by study activities at 3.5 hours.

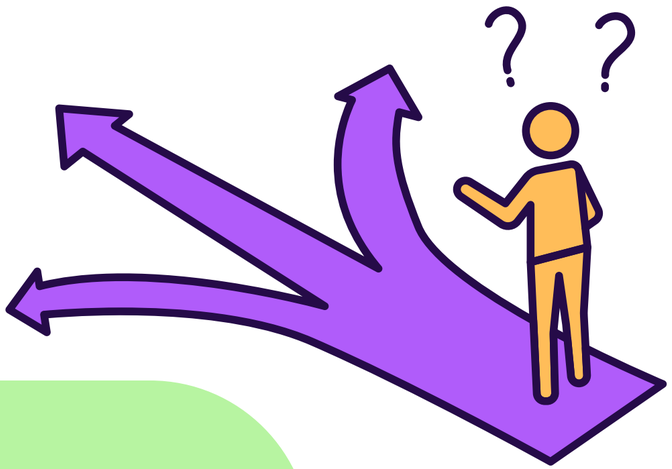
In comparison, social media usage accounts for approximately 2.5 hours daily, while time spent on the streaming platform Netflix remains under 2 hours.



MACHINE LEARNING

A Linear Regression model and Decision Tree Classifier were built, trained and tested to predict the following:

The Linear regression provided a clear, quantitative perspective on the effects of lifestyle habits on exam scores.



The decision tree revealed how combinations of student habits influence performance, clearly showing how these factors separate students into higher- or lower-performing groups. The exam score threshold was set at 75% for the decision tree.

LINEAR REGRESSION

Built via Sci-Kit Learn

Process

The dataset was divided into a 20% test set and an 80% training set. An OLS regression summary and a Variance Inflation Factor (VIF) check were conducted. Following the initial run, the variable **age** was excluded due to its **high** VIF value, as well as its limited commercial relevance and lack of alignment with habitual factors.

Finally, the performance of the trained model was evaluated against the test set by comparing their respective R-squared and Mean Squared Error (MSE) values.

LINEAR REGRESSION

Model Evaluation

OLS Regression Results

Dep. Variable:	exam_score	R-squared:	0.900
Model:	OLS	Adj. R-squared:	0.898
Method:	Least Squares	F-statistic:	641.2
Date:	Sun, 02 Nov 2025	Prob (F-statistic):	0.00
Time:	12:26:45	Log-Likelihood:	-2476.4
No. Observations:	800	AIC:	4977.
Df Residuals:	788	BIC:	5033.
Df Model:	11		
Covariance Type:	nonrobust		

The regression model explains 89.8% of the variance in exam scores, as indicated by the adjusted R^2. This suggests the model captures nearly all of the meaningful patterns in the data.

Another indicator of its significance is the probability F-statistic of 0.00, which shows that our predictors together explain exam scores far better than chance.

Even though attendance percentage and sleep hours had high VIF values, they could not be excluded from the model due to their commercial potential and effect on R-squared value when removed.

--- Variance Inflation Factors ---

	Features	VIF
0	attendance_percentage	33.08
4	sleep_hours	23.08
1	study_hours_per_day	6.77
2	social_media_hours	5.45
6	mental_health_rating	4.53
9	internet_quality_encoded	4.19
3	netflix_hours	3.89
10	diet_quality_encoded	3.60
5	exercise_frequency	3.17
8	extracurricular_encoded	1.46
7	part_time_job_encoded	1.29

LINEAR REGRESSION

Model Evaluation

Predictors with p-values greater than 0.05 were not statistically significant, meaning they did not have a measurable effect on exam scores.

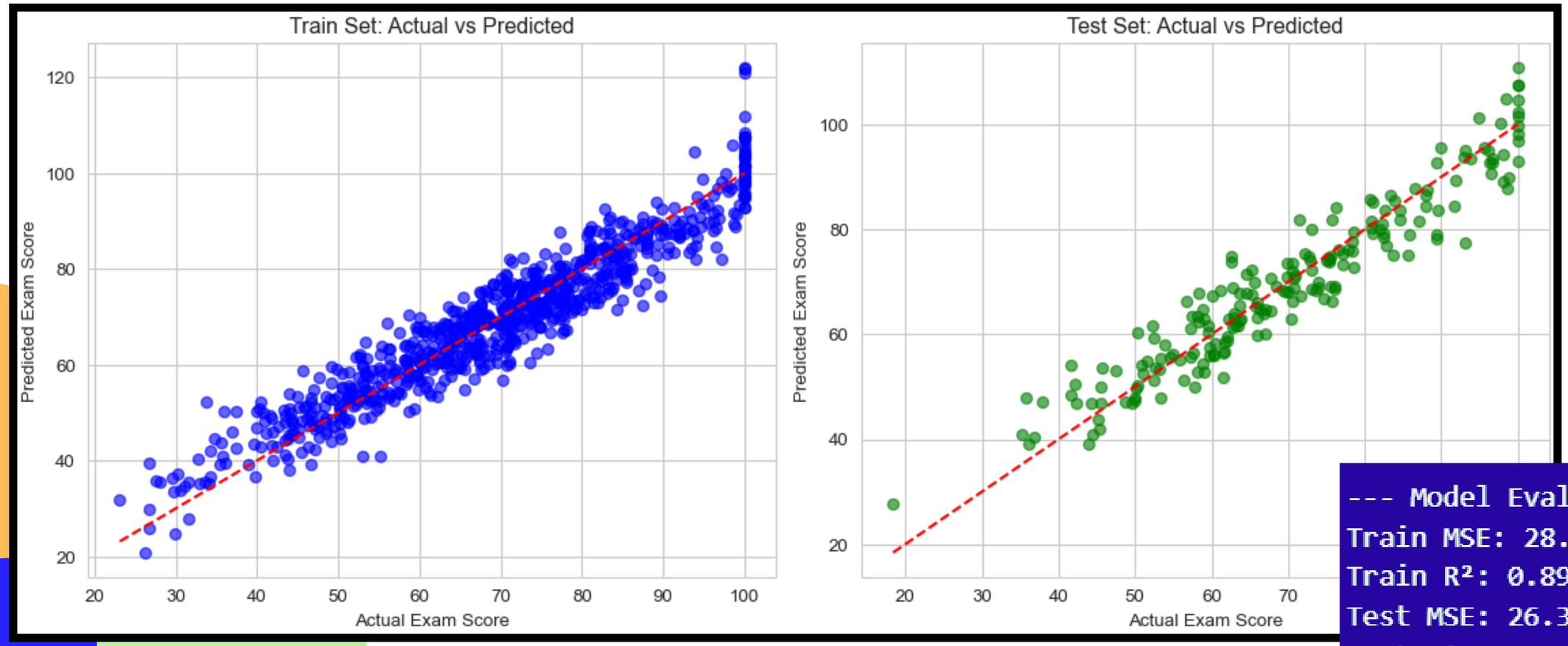
	coef	std err	t	P>|t|	[0.025	0.975]
const	6.3828	2.222	2.873	0.004	2.022	10.744
attendance_percentage	0.1488	0.020	7.281	0.000	0.109	0.189
study_hours_per_day	9.5776	0.132	72.592	0.000	9.319	9.837
social_media_hours	-2.6584	0.164	-16.202	0.000	-2.981	-2.336
netflix_hours	-2.1425	0.178	-12.010	0.000	-2.493	-1.792
sleep_hours	1.9704	0.155	12.727	0.000	1.666	2.274
exercise_frequency	1.4586	0.094	15.477	0.000	1.274	1.644
mental_health_rating	1.9514	0.068	28.869	0.000	1.819	2.084
part_time_job_encoded	0.4823	0.460	1.047	0.295	-0.422	1.386
extracurricular_encoded	0.0374	0.410	0.091	0.927	-0.768	0.843
internet_quality_encoded	-0.1913	0.266	-0.719	0.472	-0.713	0.331
diet_quality_encoded	-0.3600	0.265	-1.359	0.174	-0.880	0.160

Omnibus:	12.211	Durbin-Watson:	1.991
Prob(Omnibus):	0.002	Jarque-Bera (JB):	15.916
Skew:	-0.174	Prob(JB):	0.000350
Kurtosis:	3.597	Cond. No.	992.

LINEAR REGRESSION

R-squared Results

Our model explains nearly 90% of the variance in exam scores on the training data. Importantly, the test data R-squared is slightly higher, which shows the model generalizes well and isn’t overfitting.

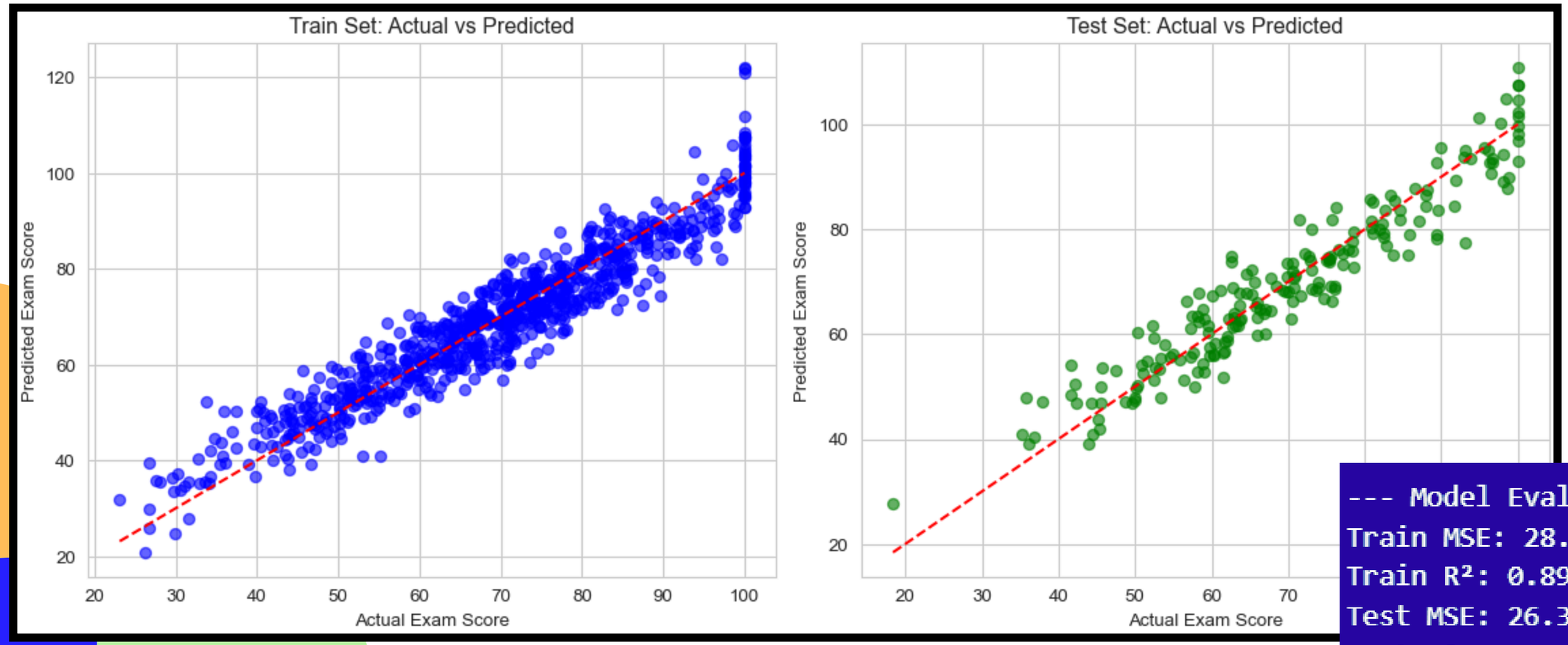



```
--- Model Evaluation ---
Train MSE: 28.589952902206925
Train R²: 0.8995046747076838
Test MSE: 26.368471963300962
Test R²: 0.908012218543634
```

LINEAR REGRESSION

MSE Results

On average, our predictions are off by about 26–28 points squared. Importantly, the test error is slightly lower than the training error, which shows the model is not overfitting and performs reliably on unseen data. However, this might be improved by removing attendance percentage as a predictor.



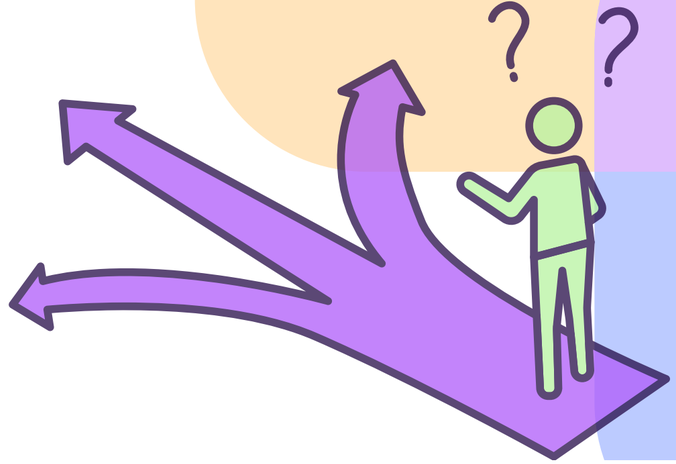

```
--- Model Evaluation ---
Train MSE: 28.589952902206925
Train R²: 0.8995046747076838
Test MSE: 26.368471963300962
Test R²: 0.908012218543634
```

DECISION TREE

Built via Sci-Kit Learn

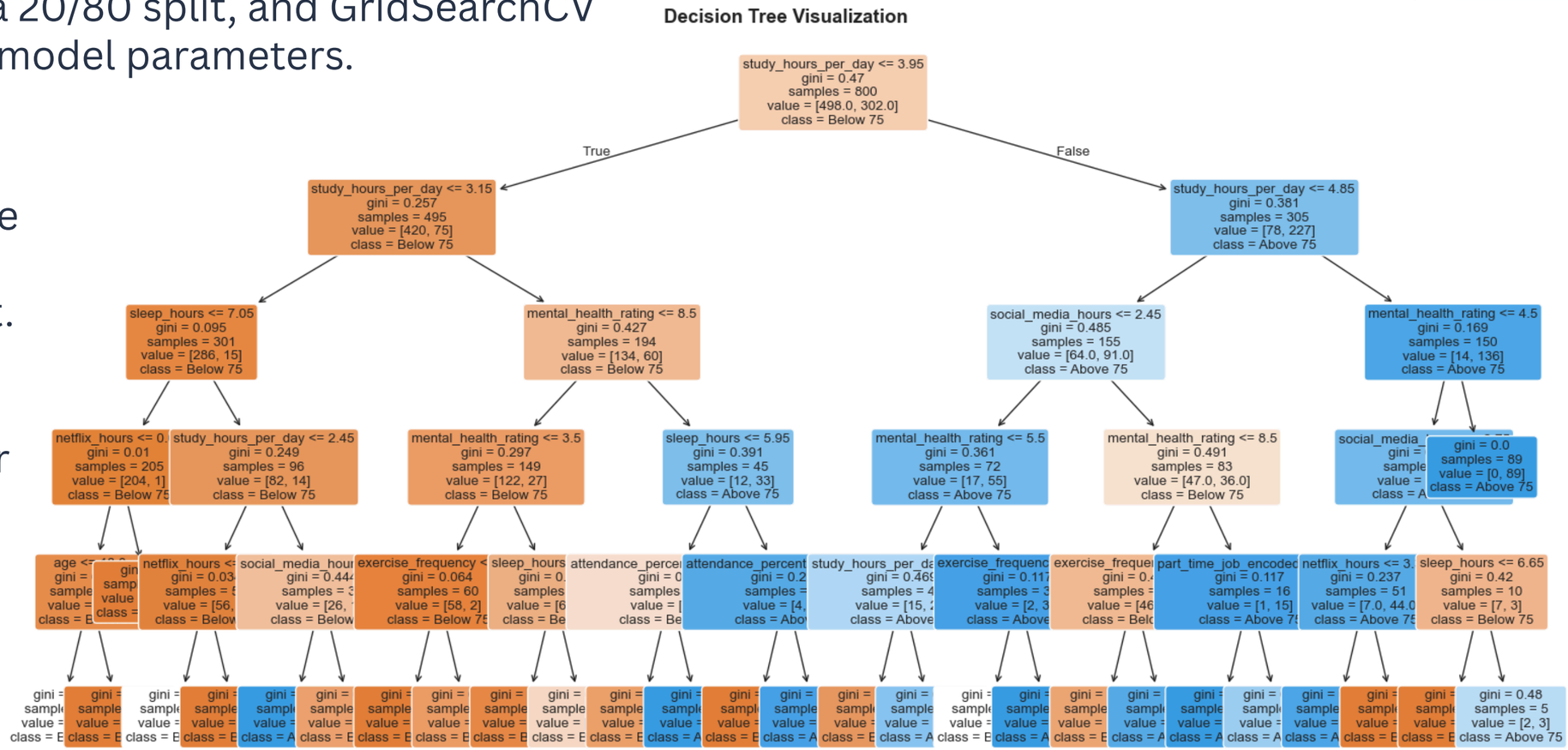
Process

The same feature categories applied in the Linear Regression model were retained, with the addition of a binary variable indicating whether exam scores were above or below 75%. The dataset was partitioned using a 20/80 split, and GridSearchCV was employed to optimize the model parameters.

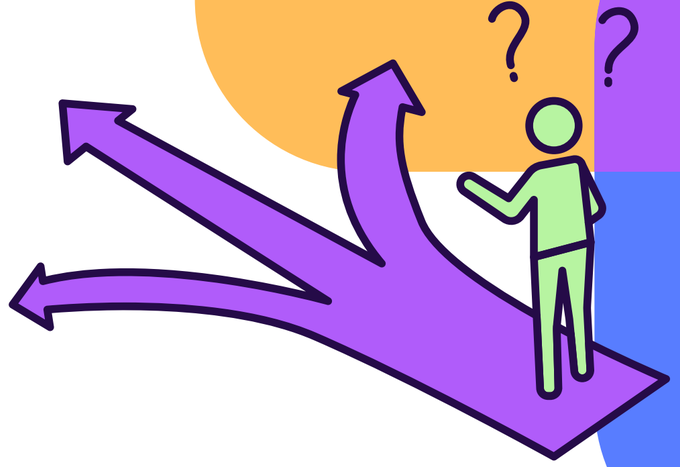


To evaluate the model's performance, we calculated the confusion matrix, accuracy score, and classification report.

Finally, we compared the linear regression model to the decision tree model using a ROC curve and AUC score.



DECISION TREE

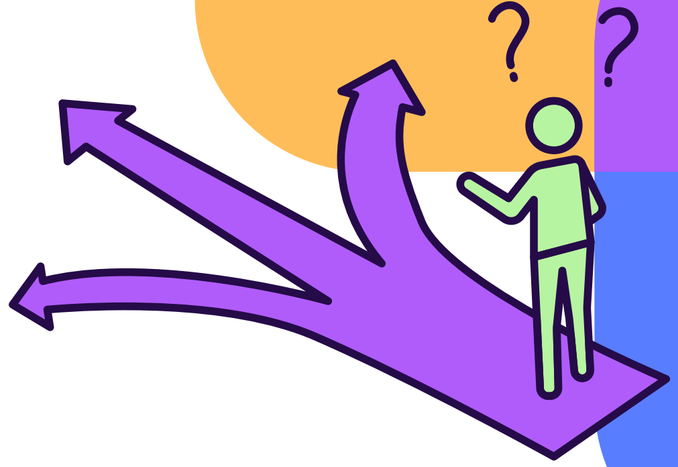


Confusion Matrix and Classification Report

```
Best Parameters: {'max_depth': 5, 'min_samples_leaf': 2, 'min_samples_split': 2}
Cross-validated Best Score: 0.8300000000000001
Test Accuracy: 0.885

Confusion Matrix:
 [[117   7]
 [ 16  60]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.94      0.91       124
           1       0.90      0.79      0.84        76

    accuracy                           0.89       200
   macro avg       0.89      0.87      0.87       200
weighted avg       0.89      0.89      0.88       200
```

The confusion matrix allowed us to see how well the model distinguished between the two classes, the accuracy score provided an overall measure of correct predictions, and the classification report gave deeper insights into precision, recall, and F1-score for each class.

DECISION TREE



Confusion Matrix and Classification Report

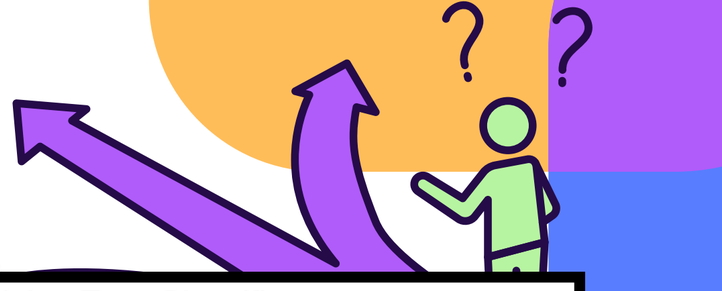
```
Best Parameters: {'max_depth': 5, 'min_samples_leaf': 2, 'min_samples_split': 2}
Cross-validated Best Score: 0.8300000000000001
Test Accuracy: 0.885

Confusion Matrix:
 [[117   7]
 [ 16  60]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.94      0.91       124
           1       0.90      0.79      0.84        76

    accuracy                           0.89       200
   macro avg       0.89      0.87      0.87       200
weighted avg       0.89      0.89      0.88       200
```

Exam score below 75%

Exam score above 75%

The model predicted that 90% of students would score above 75%, while in reality 79% did. Despite this slight overestimation, the classification report shows an F1 score of 0.84 for the ‘above 75%’ class — confirming the model is both accurate and reliable at identifying high performers.

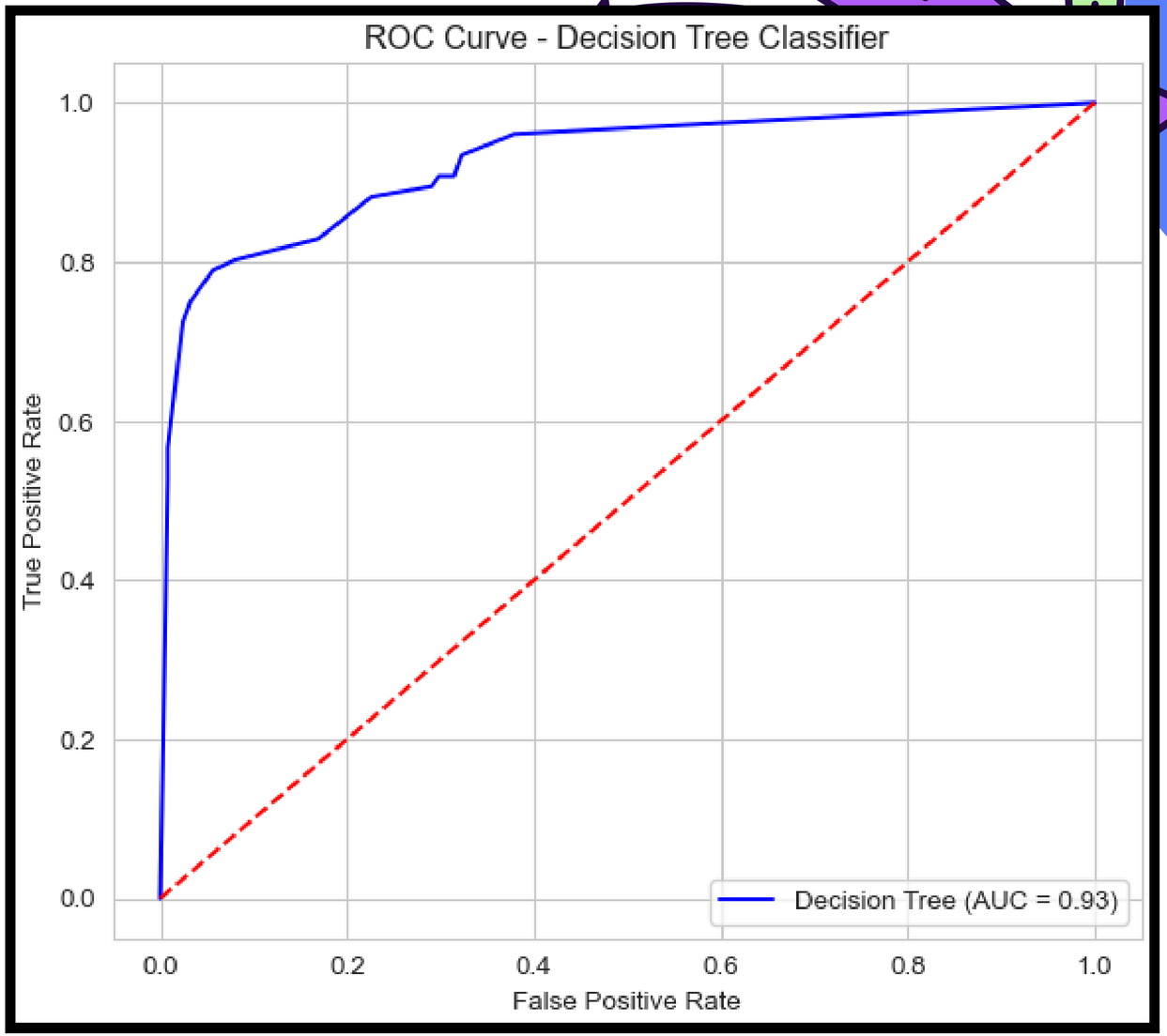
DECISION TREE



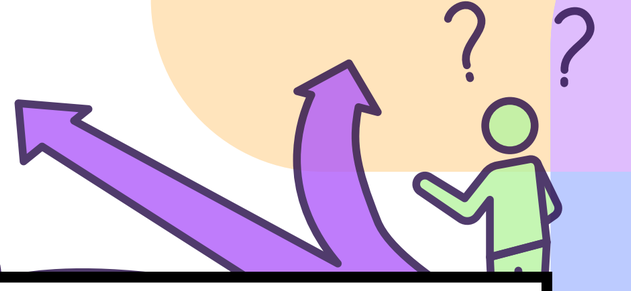
ROC Curve and AUC Result

The ROC curve shows our Decision model is highly effective at distinguishing between classes. With an AUC close to 1, the model achieves near-perfect classification performance.

This means the model is not only accurate overall, but also balances sensitivity and specificity extremely well.



DECISION TREE

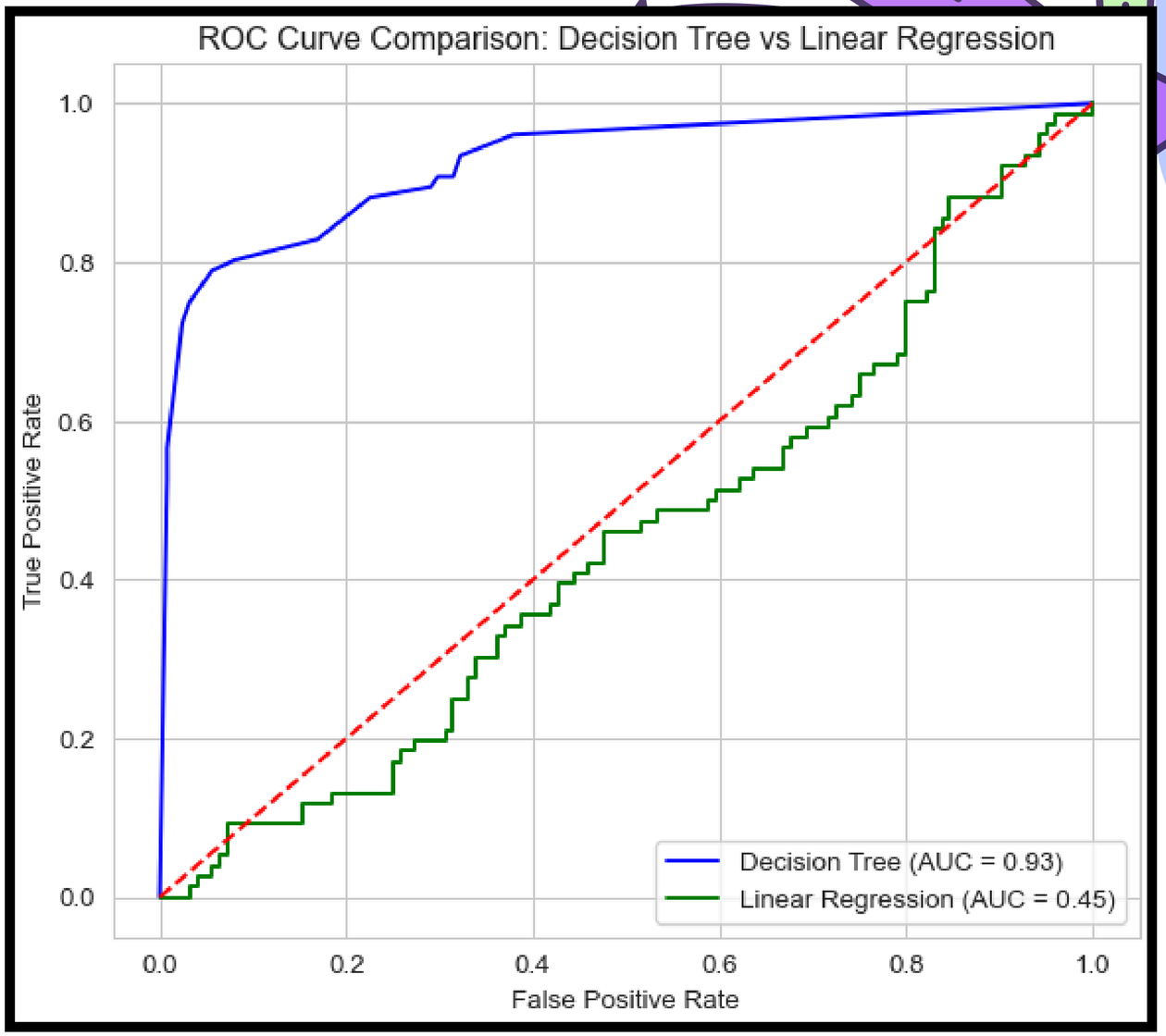


ROC Curve and AUC Result

While the Decision Tree model achieves near-perfect classification performance (AUC ≈ 1), the Linear Regression model performs poorly on this metric (AUC ≈ 0.45).

This is expected, since linear regression is not optimized for classification tasks. Also, continuous exam scores were used for the curve.

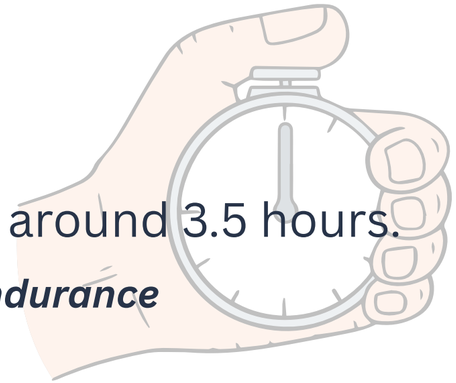
Instead, its strength lies in explaining variance (R^2 ≈ 0.898) and mean squared errors (MSE ≈ 26–28).



INSIGHTS

1. Books open doors, but rest keeps you standing.

Suggested Niche Industries: Health Focused



2. Exam success isn’t random — it scales with study time, peaking around 3.5 hours.

Suggested Niche Industries: Gamifying Studying, Study Timers, Endurance Supplements

3. Connection quality peaks early — after that, it’s all you.

Suggested Niche Industries: Study Cafes, Habit Reinforcement Apps

CONCLUSION

Although time-consuming activities played a role, exam outcomes were shaped most by the balance of sleep, mental health, and dedicated study hours.

Industries that may improve these factors would be profitable.

THANK YOU!