Egypt University of Informatics
Computer and Information Systems
Data Analysis Course

The Analysis of

Submitted by:

Hamdy Elsaeed,
Omar Aboelnaga
Patrick Boules 23-101185
Youssef Ashour 23-101046

25/5/2025

Introduction

Student academic performance is influenced by a wide range of lifestyle and environmental variables. Understanding which factors contribute most significantly to performance can help educators and institutions tailor interventions that improve learning outcomes.This report explores the academic performance of students and its relationship to various lifestyle factors. The goal is to draw meaningful conclusions about the behaviors that are most strongly associated with academic success.

Research Question

What are the most influential lifestyle factors associated with students' exam performance?

Hypothesis

There are 3 main factors that affect academic success performance which are Study Hours, Mental Health and Attendance and there are sub factors that affect them.

Population of Interest

The population of interest is 1000 university students aged 17 to 24.

Sampling Method

The dataset appears to have been compiled through convenience sampling, possibly via a voluntary survey administered to students. While this method allows for fast data collection, it may introduce self-selection bias.

Bias Identification

Potential biases include:

- **Self-reporting bias**: Students may overestimate healthy habits (e.g., study hours).

- **Non-response bias**: Those disinterested in academics may be underrepresented.

To mitigate this, questions were neutrally phrased and the sample includes a broad range of behaviors.

Dataset Description

The dataset includes the following variables:

- student_id: Unique identifier for each student (qualitative)

- age: Student age (quantitative)
  - Mean: 20.49
  - Min: 17
  - Max: 24

- gender: Male/Female (qualitative)
  - Male: 477
  - Female: 481

- study_hours_per_day: Time spent studying (quantitative)
  - Mean: 3.55
  - Min: 0
  - Max: 8.30

- wasted_time: Daily social media use and time spent on Netflix in hours (quantitative)
  - Mean: 4.33
  - Min: 0.20
  - Max: 10.10

- part_time_job: Whether the student works part-time (qualitative)
  - Yes: 215
  - No: 785

- attendance_percentage: Class attendance rate (quantitative)
  - Mean: 84.13
  - Min: 56
  - Max: 100

- sleep_hours: Daily sleep duration (quantitative)
  - Mean: 6.47
  - Min: 3.20
  - Max: 10

- diet_quality: Rated as Poor, Fair, Good (qualitative)
  - Good: 378
  - Fair: 437
  - Poor: 185

- exercise_frequency: Times exercise per week (quantitative)
  - Mean: rounded to 3
  - Min: 0
  - Max: 6

- parental_education_level: High School, Bachelor, Master, etc. (qualitative)
  - High School: 392
  - Bachelor: 350
  - Master: 167

- internet_quality: Rated as Poor, Average, Good (qualitative)
  - Good: 447
  - Average: 391
  - Poor: 162

- **mental_health_rating**: Scale from 1 (low) to 10 (high) (quantitative)
  - Mean: 5.43
  - Min: 1
  - Max: 10

- **extracurricular_participation**: Yes/No (qualitative)
  - Yes: 318
  - No: 682

- **exam_score**: Final exam score out of 100 (quantitative)
  - Mean: 3.55
  - Min: 0
  - Max: 8.30

Analysis

We categorized students according to their exam scores into grade categories and created a PowerBI dashboard and used slicers to present social habits that differ students from each other.
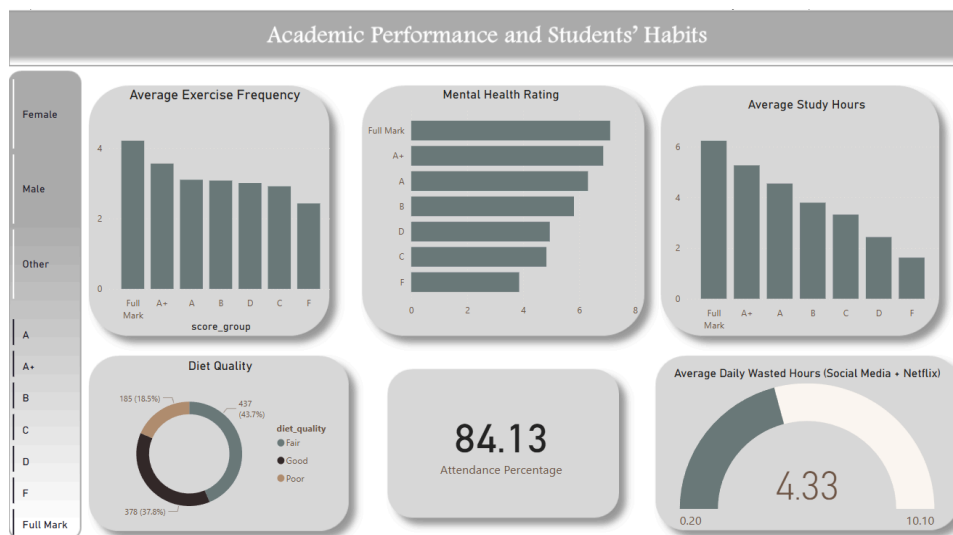
1. **Study Hours Vs Students' Grade**
    ○ Bar chart showing the average study hours for each category for students' grade

2. **Mental Health rating Vs Students' Grade**
    ○ Horizontal bar chart that shows the average mental health rating for every category for students' grade

3. **Average Exercise Frequency vs Exam Score**
    ○ Bar chart showing the average exercise frequency for each category for students' grade
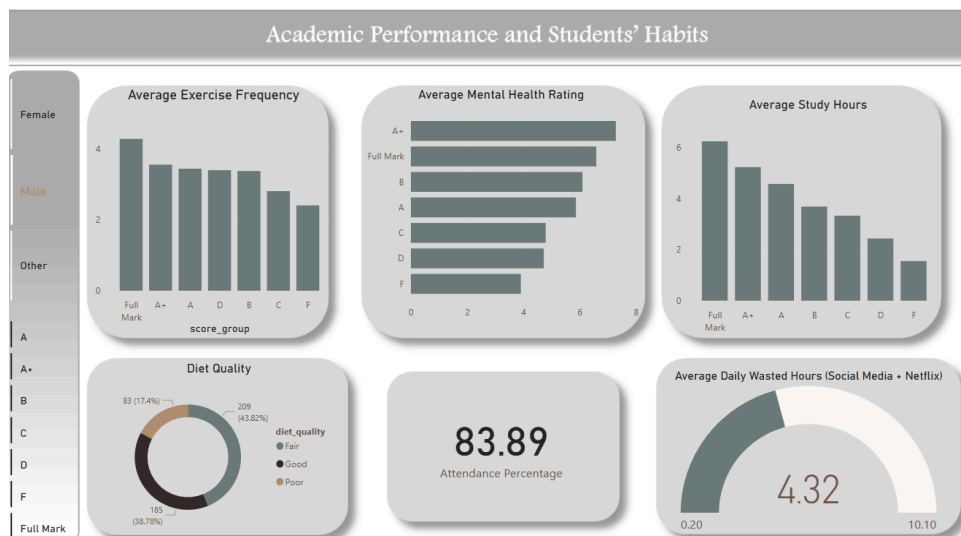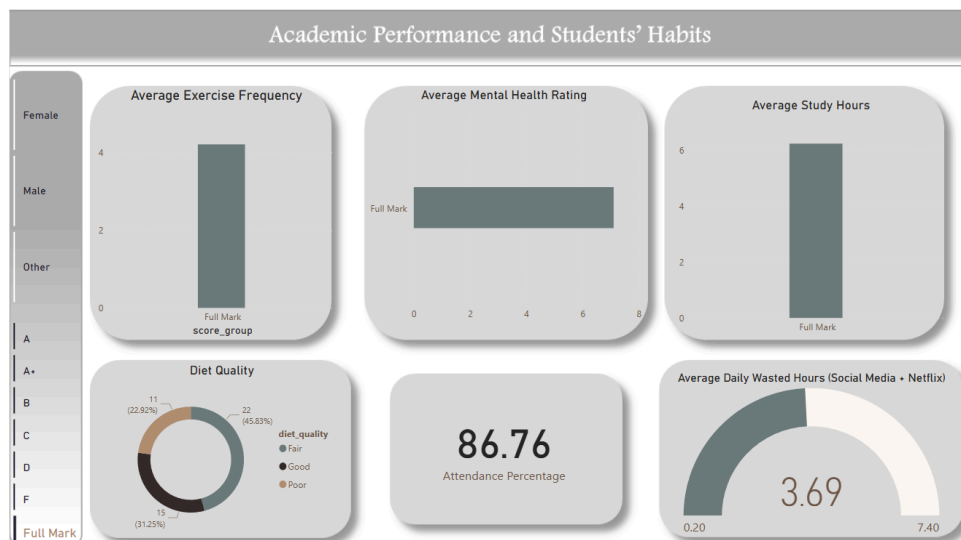
4. **Diet Quality Distribution**
   - A pie chart that shows the distribution of the diet quality and it varies through genders and exam score categories

5. **Average Attendance Percentage**
   - A card that shows the average attendance % of the students of the selected category

6. **Average Daily Wasted Hours**
   - A gauge that shows that min, max, average daily wasted hours (social media + Netflix)


Key Drivers of Exam Score

Based on the correlation analysis, the following variables were associated with exam performance:

- Study Hours per Day: Very strong positive correlation
- Mental Health: Moderate positive correlation
- Wasted Time: Moderate negative correlation
- Exercise Frequency: Weak positive correlation
- Daily Sleep Hours: Weak positive correlation
- Attendance: Very weak positive correlation
- Gender: There is a difference in mean exam score between the two groups

---

Hypothesis Tests
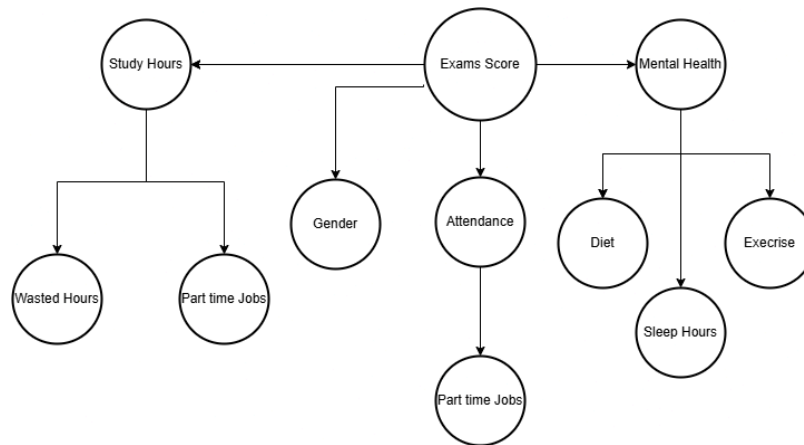
1. <u>Study Hours vs. Exam Grades</u>
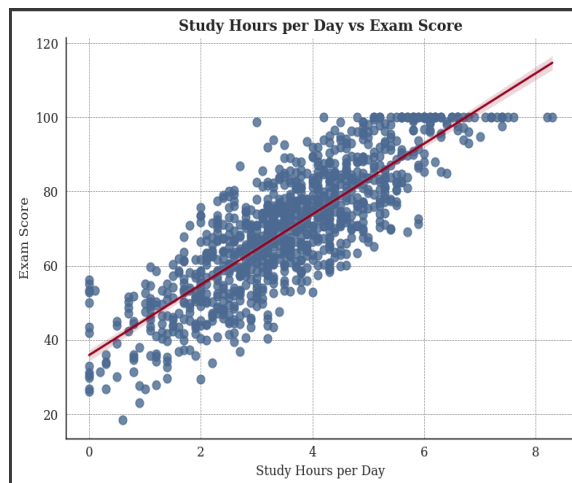
    Test Used: Pearson Correlation Test

    H_0: There is no correlation between study hours and exam grades.

    H_1: There is a positive correlation between study hours and exam grades.

    Result: Correlation Coefficient = 0.825, P-value < 0.001

    Decision: Reject the null hypothesis — strong positive correlation exists.
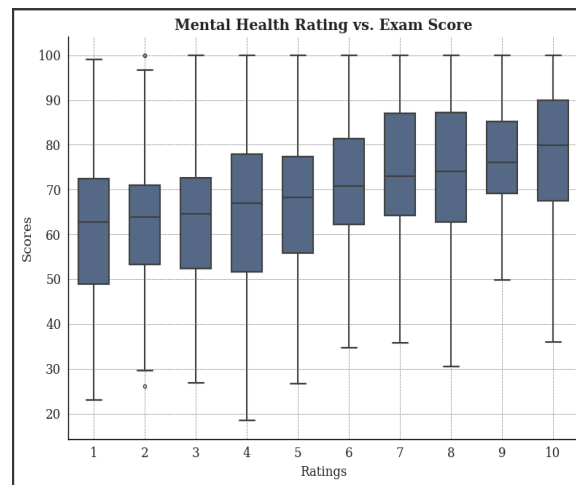
2. Mental Health vs. Exam Grades

   Test Used: Pearson Correlation Test

   H_0: There is no correlation between mental health rating and exam score.

   H_1: Mental health rating is positively correlated with exam score.

   Result: Correlation Coefficient = 0.322, P-value < 0.001

   Decision: Reject the null hypothesis — moderate positive correlation exists.


3. Attendance vs. Exam Grades

   Test Used: Pearson Correlation Test

   H_0: There is no correlation between attendance percentage and exam grades.

   H_1: There is a positive correlation between attendance percentage and exam grades.

   Result: Correlation Coefficient = 0.0898, P-value = 0.0045

   Decision: Reject the null hypothesis — weak but statistically significant positive correlation.
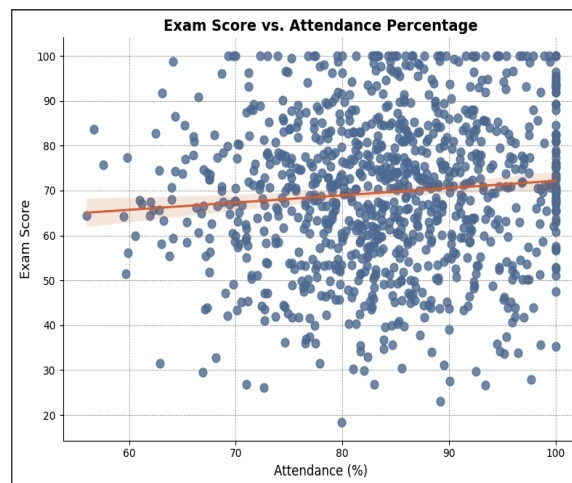
4. Wasted Time vs. Study Time

Test Used: Pearson Correlation Test

H_0: There is no linear correlation between wasted time and study time.

H_1: There is a negative correlation between wasted time and study time.

Result: Pearson Correlation Coefficient = -0.0061, P-value = 0.848

Decision: Fail to reject the null hypothesis. There is no statistically significant correlation between wasted time and study time.
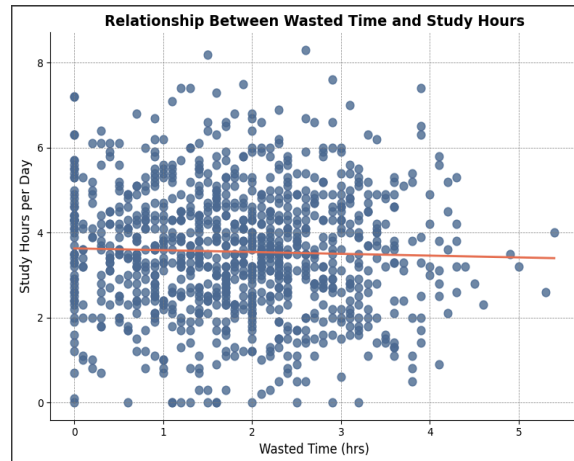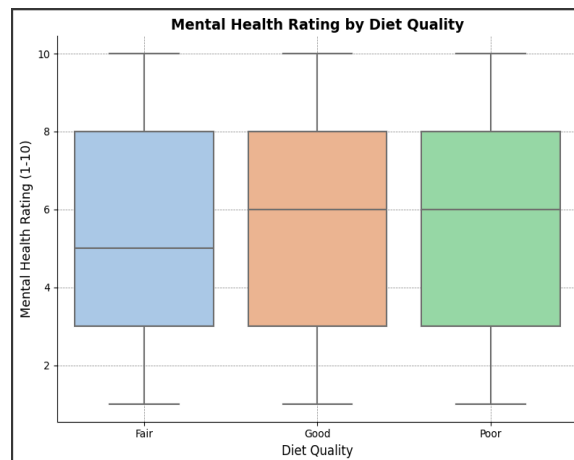

5. Mental Health vs. Diet Quality

Test Used: ANOVA Test

H_0: Mean mental health ratings are equal across levels of diet quality.

H_1: Mean mental health ratings differ across diet quality categories.

Result: F-statistic = 2.595, P-value = 0.075

Decision: Fail to reject the null hypothesis. No significant difference in mental health across diet groups.
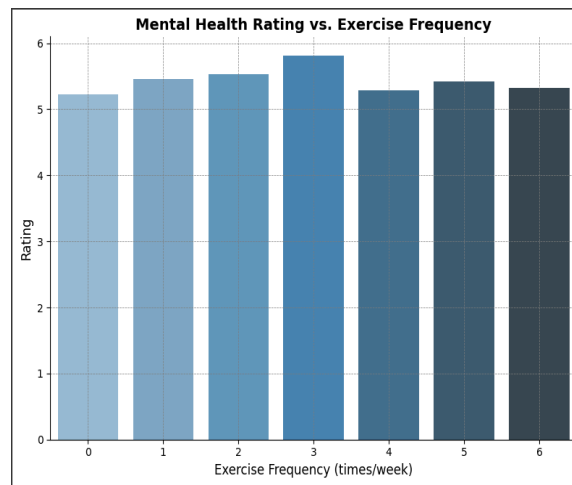
6. <u>Mental Health vs. Exercise Frequency</u>

Test Used: Pearson Correlation Test

H_0: No correlation exists between exercise frequency and mental health.

H_1: A positive correlation exists between exercise frequency and mental health.

Result: Pearson Correlation Coefficient = -0.0002, P-value = 0.994

Decision: Fail to reject the null hypothesis. No significant correlation was found.

7. <u>Mental Health vs. Sleep Hours</u>

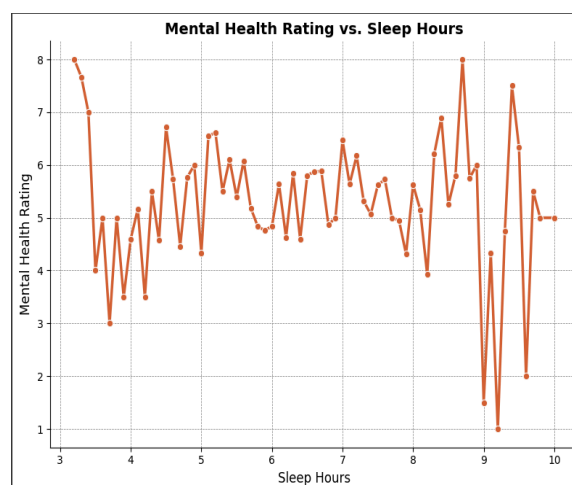Test Used: Pearson Correlation Test

H_0: No correlation exists between sleep duration and mental health.

H_1: A positive correlation exists between sleep and mental health.

Result: Pearson Correlation Coefficient = -0.0065, P-value = 0.837

Decision: Fail to reject the null hypothesis. No significant correlation between sleep and mental health was observed.
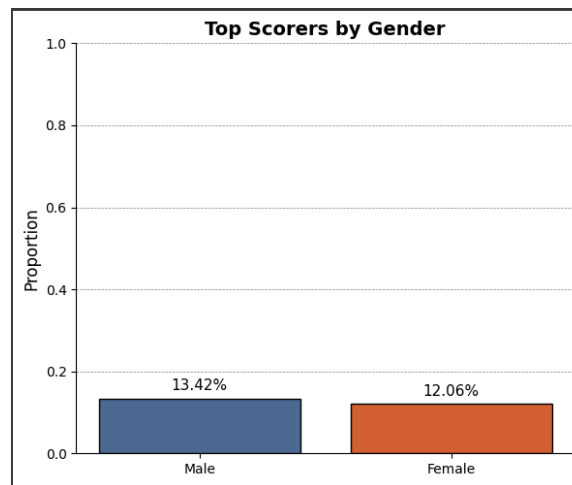
8. Gender vs. High Achievers

Test Used: Z-Test for Proportions

H_0: The proportion of high scorers (A+ or Full Mark) is the same for males and females.

H_1: The proportion of high scorers is different between males and females.

Result: Z-statistic = 0.631, P-value = 0.528

Decision: Fail to reject the null hypothesis — no significant difference found.

Analysis For Hypothesis tests:

After Analyzing the 3 main factors that affect exam scores which are study hours, mental health and attendance and found their significance. Here are the results reached:

- There is a huge positive correlation between number of study hours and exam scores
- There is small positive correlation between mental health rating and exam scores
- There is small positive correlation between attendance percentage and exam scores

The next step, we analyzed what are logical factors that could affect each one of the main factors and to see if they are dependent or independent factors.

1. Study Hours:
   a. Study hours VS Wasted hours: although there is no correlation between study hours yet wasted time has a negative correlation with exam scores
   b. Study hours VS Part-time Jobs: There is no correlation between study hours and part time jobs, similarly there is no correlation between part-time and exam scores.

2. Mental Health:
   a. Mental health VS Diet Food: There is no significance difference and similarly there is no significance difference between Diet and Exam scores.
   b. Mental health VS Exercise Frequency: There is no correlation between mental health and exercise frequency, yet there is small positive correlation between exercise frequency and exam scores.
   c. Mental health VS Sleep Hours: There is no correlation between mental health and sleep hours yet there is a positive correlation between sleep hours and exam scores.
3. Attendance:
   a. Attendance VS Part-Time Job: there is no significant difference, similarly there is no correlation between part-time and exam scores.

P.S: All of the significant relations are tested in the notebook in the "Before ML Hypothesis tests".

Conclusion

The analysis reveals that consistent study habits, adequate sleep, higher class attendance, and better mental health are positively linked with improved academic outcomes. Surprisingly, factors often assumed to be influential—such as having a part-time job, diet quality, internet quality, and extracurricular participation—did not show a strong effect on exam scores. There are other factors has a positive effect on academic outcomes but not a direct effect on the 3 main factors While causality cannot be confirmed from this data alone, the insights gained are valuable for guiding student support strategies and academic policy development.


Study Limitations

- The sample is limited to 1000 students, possibly from a single institution.

- No control for confounding variables like socioeconomic status.

- Convenience sampling reduces generalizability

Potential Issues

- Some variables are self-reported, which can introduce bias.

- Categories like "diet quality" and "mental health rating" are subjective.

- No timestamp on data collection (seasonal effects may apply).

Machine learning model

After performing hypothesis testing between each variable and the target column (exam score), we identified variables that were independent of the target and removed them from the dataset to avoid degrading the model's performance. To prevent overfitting or underfitting, we applied hyperparameter tuning for each model using GridSearch. After comparing the accuracy scores of all models, we found that the Support Vector Regressor achieved the highest accuracy, with a score of **89.4%**.

Model Accuracy Comparison (100 - MAPE)
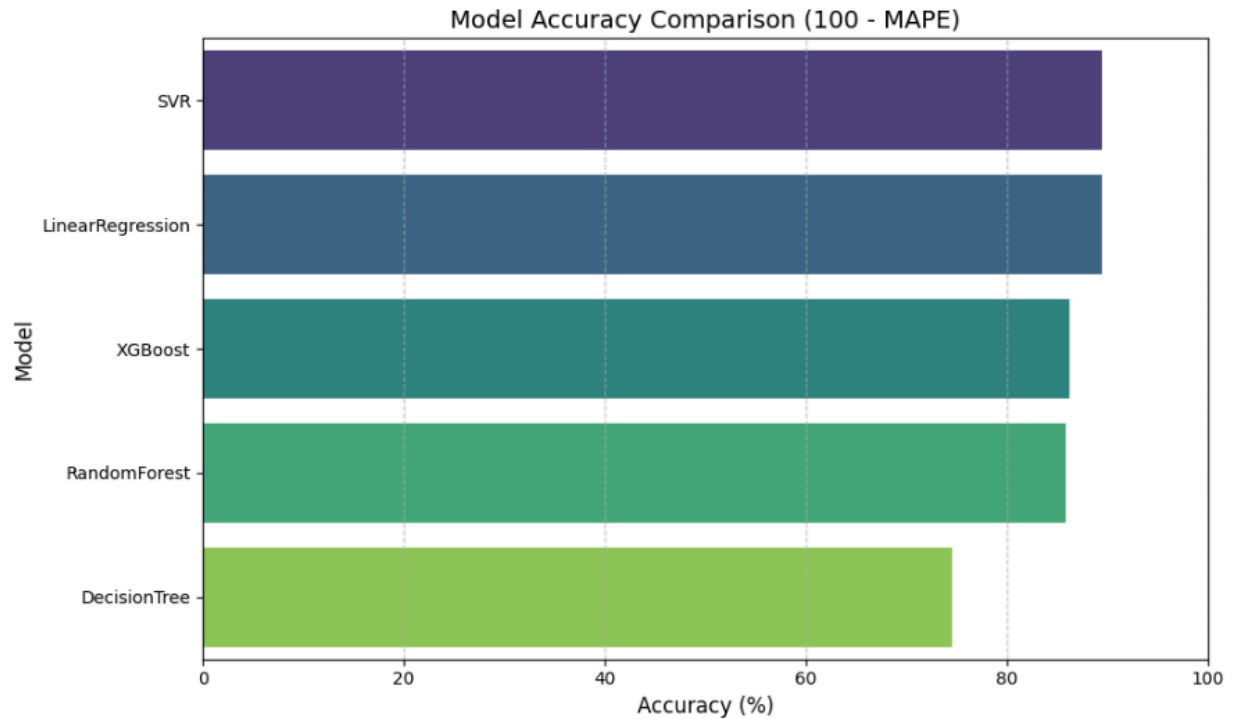
Recommendations

- Enhance Mental Health Support: Universities should expand access to mental health resources and counseling services, as student well-being is closely tied to academic performance.
- Promote Healthy Routines: Encourage students to adopt balanced sleep and study habits through awareness campaigns and structured academic planning.
- Time Management Training: Offer workshops and resources focused on time management to help students minimize distractions, such as excessive screen time, and improve academic efficiency.