CS 4375 Final Project Report

Ivan Yau

Introduction

This study explores whether lifestyle habits significantly influence students' academic performance. By analyzing factors such as age, gender, study hours, social media usage, Netflix hours, part-time employment, attendance percentage, sleep duration, diet quality, exercise frequency, parental education level, internet quality, mental health rating, and extracurricular participation, the study examines their impact on exam scores.

As a student myself, I am particularly intrigued by the question of what factors shape academic success. Gaining insights into how different habits affect performance is both academically interesting and personally valuable for my educational journey.

Understanding the connection between lifestyle habits and academic performance can guide the development of more effective student support strategies. A related study found that 71.42% of students with GPAs of 18 or higher had good lifestyles, whereas over 55% of students with GPAs below 14 exhibited poor lifestyles [1]. This highlights the strong link between healthy habits and academic achievement. Just as predictive models in medicine help identify at-risk individuals, educational institutions can use data-driven insights to proactively support students exhibiting poor lifestyle habits.

I will use the following data set from Kaggle:
Student Habits vs. Academic Performance Dataset

This project focuses on predicting students' academic performance using lifestyle-related features. Several machine learning models, including Linear Regression, Random Forest, XGBoost, Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP), were implemented and evaluated. The performance of each model was compared using appropriate metrics, and the evaluation focused on identifying why certain models outperformed others. Multiple feature sets, including original features, interaction terms, and refined subsets based on feature importance, were tested to optimize model performance. Cross-validation was used to fine-tune hyperparameters and avoid overfitting. Evaluation metrics like precision and recall were prioritized, as they provide a more targeted assessment of model effectiveness in high-stakes academic prediction.

Exploratory Data Analysis

Data Formatting and Missing Values

Format:

```
Dataset shape: (1000, 22)

First 5 rows:
  student_id  age  study_hours_per_day  social_media_hours  netflix_hours  \
0      S1000   23                  0.0                 1.2            1.1
1      S1001   20                  6.9                 2.8            2.3
2      S1002   21                  1.4                 3.1            1.3
3      S1003   23                  1.0                 3.9            1.0
4      S1004   19                  5.0                 4.4            0.5

   attendance_percentage  sleep_hours  exercise_frequency  \
0                   85.0          8.0                   6
1                   97.3          4.6                   6
2                   94.8          8.0                   1
3                   71.0          9.2                   4
4                   90.9          4.9                   3

   mental_health_rating  exam_score  ...  gender_Male  gender_Other  \
0                     8        56.2  ...        False         False
1                     8       100.0  ...        False         False
2                     1        34.3  ...         True         False
3                     1        26.8  ...        False         False
4                     1        66.4  ...        False         False

   part_time_job_Yes  diet_quality_Good  diet_quality_Poor  \
0              False              False              False
1              False               True              False
2              False              False               True
3              False              False               True
4              False              False              False

   parental_education_level_High School  parental_education_level_Master  \
0                                 False                             True
1                                  True                            False
2                                  True                            False
3                                 False                             True
4                                 False                             True

   internet_quality_Good  internet_quality_Poor  \
0                  False                  False
1                  False                  False
2                  False                   True
3                   True                  False
4                   True                  False

   extracurricular_participation_Yes
0                               True
1                              False
2                              False
3                               True
4                              False

[5 rows x 22 columns]
```

Examined for missing values or anomalies:

```
Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 22 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   student_id                              1000 non-null   object
 1   age                                     1000 non-null   int64
 2   study_hours_per_day                     1000 non-null   float64
 3   social_media_hours                      1000 non-null   float64
 4   netflix_hours                           1000 non-null   float64
 5   attendance_percentage                   1000 non-null   float64
 6   sleep_hours                             1000 non-null   float64
 7   exercise_frequency                      1000 non-null   int64
 8   mental_health_rating                    1000 non-null   int64
 9   exam_score                              1000 non-null   float64
 10  total_leisure_hours                     1000 non-null   float64
 11  productivity_ratio                      1000 non-null   float64
 12  gender_Male                             1000 non-null   bool
 13  gender_Other                            1000 non-null   bool
 14  part_time_job_Yes                       1000 non-null   bool
 15  diet_quality_Good                       1000 non-null   bool
 16  diet_quality_Poor                       1000 non-null   bool
 17  parental_education_level_High School    1000 non-null   bool
 18  parental_education_level_Master         1000 non-null   bool
 19  internet_quality_Good                   1000 non-null   bool
 20  internet_quality_Poor                   1000 non-null   bool
 21  extracurricular_participation_Yes       1000 non-null   bool
dtypes: bool(10), float64(8), int64(3), object(1)
memory usage: 103.6+ KB
None
```

There were no missing values

Descriptive Statistics:

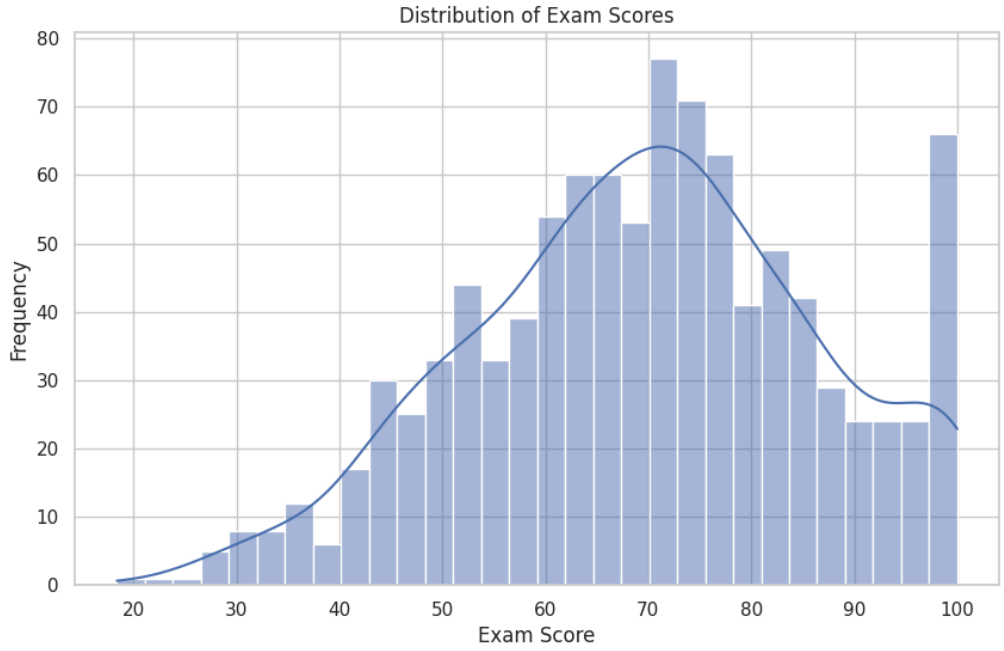
```
Descriptive statistics:
             age  study_hours_per_day  social_media_hours  netflix_hours  \
count  1000.0000           1000.00000         1000.000000    1000.000000
mean     20.4980              3.55010            2.505500       1.819700
std       2.3081              1.46889            1.172422       1.075118
min      17.0000              0.00000            0.000000       0.000000
25%      18.7500              2.60000            1.700000       1.000000
50%      20.0000              3.50000            2.500000       1.800000
75%      23.0000              4.50000            3.300000       2.525000
max      24.0000              8.30000            7.200000       5.400000

       attendance_percentage  sleep_hours  exercise_frequency  \
count            1000.000000  1000.000000         1000.000000
mean               84.131700     6.470100            3.042000
std                 9.399246     1.226377            2.025423
min                56.000000     3.200000            0.000000
25%                78.000000     5.600000            1.000000
50%                84.400000     6.500000            3.000000
75%                91.025000     7.300000            5.000000
max               100.000000    10.000000            6.000000

       mental_health_rating   exam_score  total_leisure_hours  \
count           1000.000000  1000.000000          1000.000000
mean               5.438000    69.601500             4.325200
std                2.847501    16.888564             1.599808
min                1.000000    18.400000             0.200000
25%                3.000000    58.475000             3.300000
50%                5.000000    70.500000             4.400000
75%                8.000000    81.325000             5.400000
max               10.000000   100.000000            10.100000

       productivity_ratio
count         1000.000000
mean             1.031483
std              1.149309
min              0.000000
25%              0.562328
50%              0.812124
75%              1.196581
max             25.714286
```

Descriptive statistics were computed for each feature to better understand the data distribution. Key findings include:

- The average exam score was approximately 69.6, with scores ranging from 18.4 to 100. Most students scored between the 25th percentile (58.7) and the median (70.5), with a standard deviation of 16.9, indicating moderate variability.
- Students reported an average of 6.47 hours of study per day, with a standard deviation of 1.22 hours.
- Leisure time (including social media and Netflix) averaged 4.3 hours per day, with a range of 0.2 to 10.1 hours.

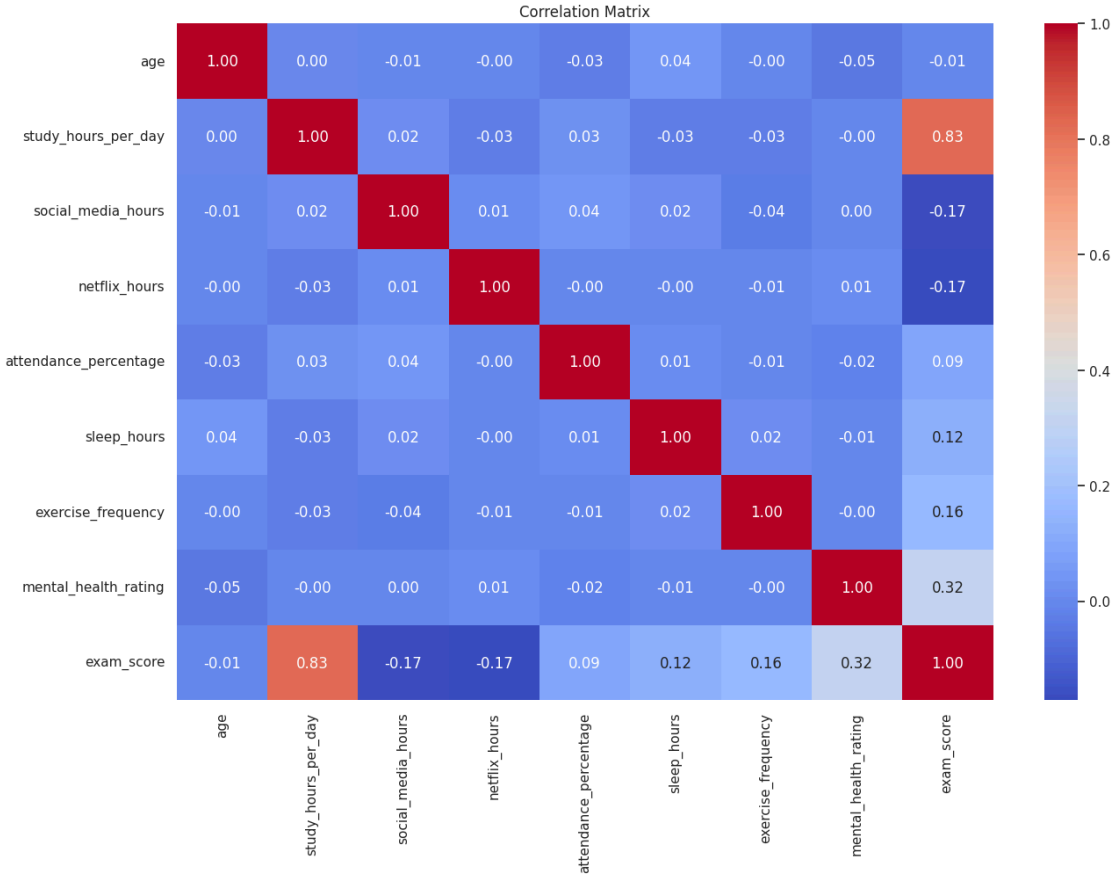
These statistics show notable variation in both academic performance and lifestyle habits, which may significantly affect students' success.

Plotted the distribution of Exam Scores:



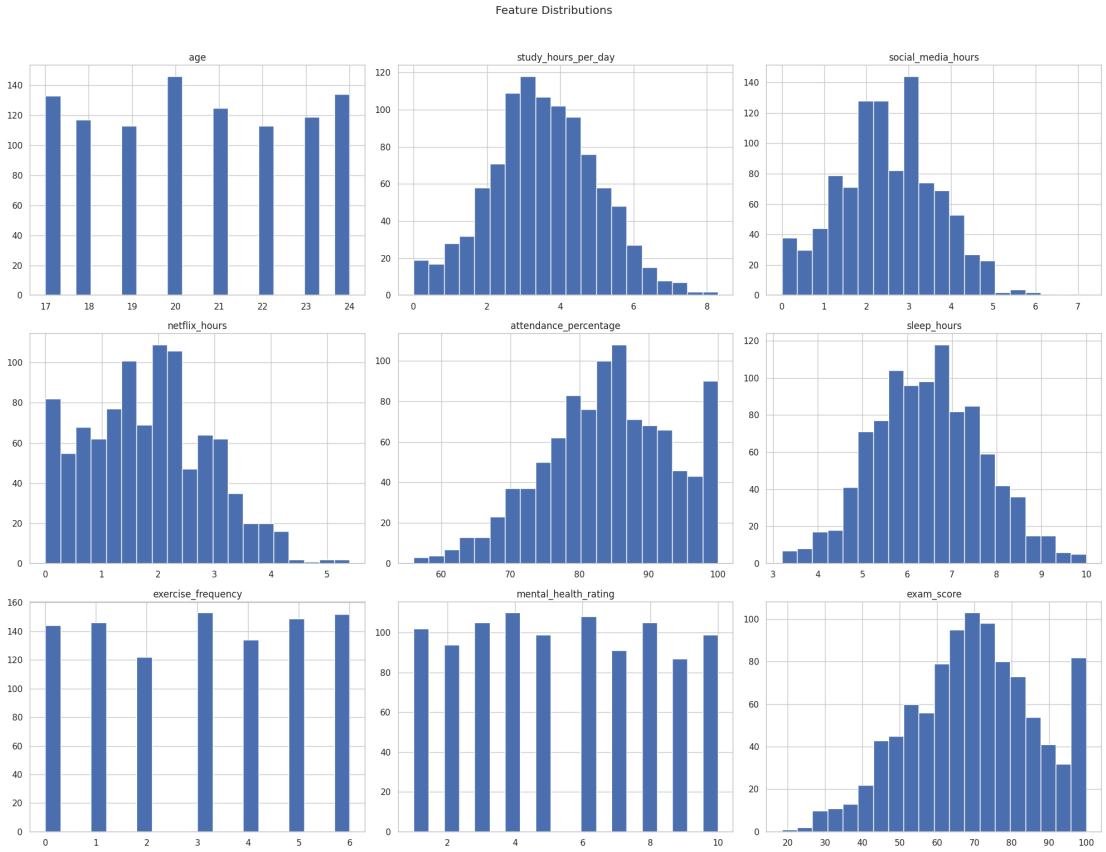
Correlation Analysis:

```
Top correlations with exam_score:
exam_score                1.000000
study_hours_per_day       0.825419
mental_health_rating      0.321523
exercise_frequency        0.160107
sleep_hours               0.121683
attendance_percentage     0.089836
age                      -0.008907
social_media_hours       -0.166733
netflix_hours            -0.171779
```



The correlation matrix revealed that study hours per day had a strong positive correlation (0.83) with exam scores, suggesting that more study time leads to better performance. Factors like mental health rating (0.32), exercise frequency (0.16), and sleep hours (0.12) showed weaker positive correlations. Interestingly, social media hours (-0.17) and Netflix hours (-0.17) displayed weak negative correlations, suggesting that more time spent on these activities slightly decreases exam scores. Age showed no correlation (-0.01) with performance.

Feature Distribution:



The student population is primarily aged between 17 and 24, with high attendance rates but varied lifestyle habits. Many students report studying 2–4 hours per day and sleeping less than the recommended amount, averaging 5–7 hours per night. Exercise is infrequent, and while most students report good mental health, some outliers suggest the need for further attention.

Data Preparation & Feature Engineering

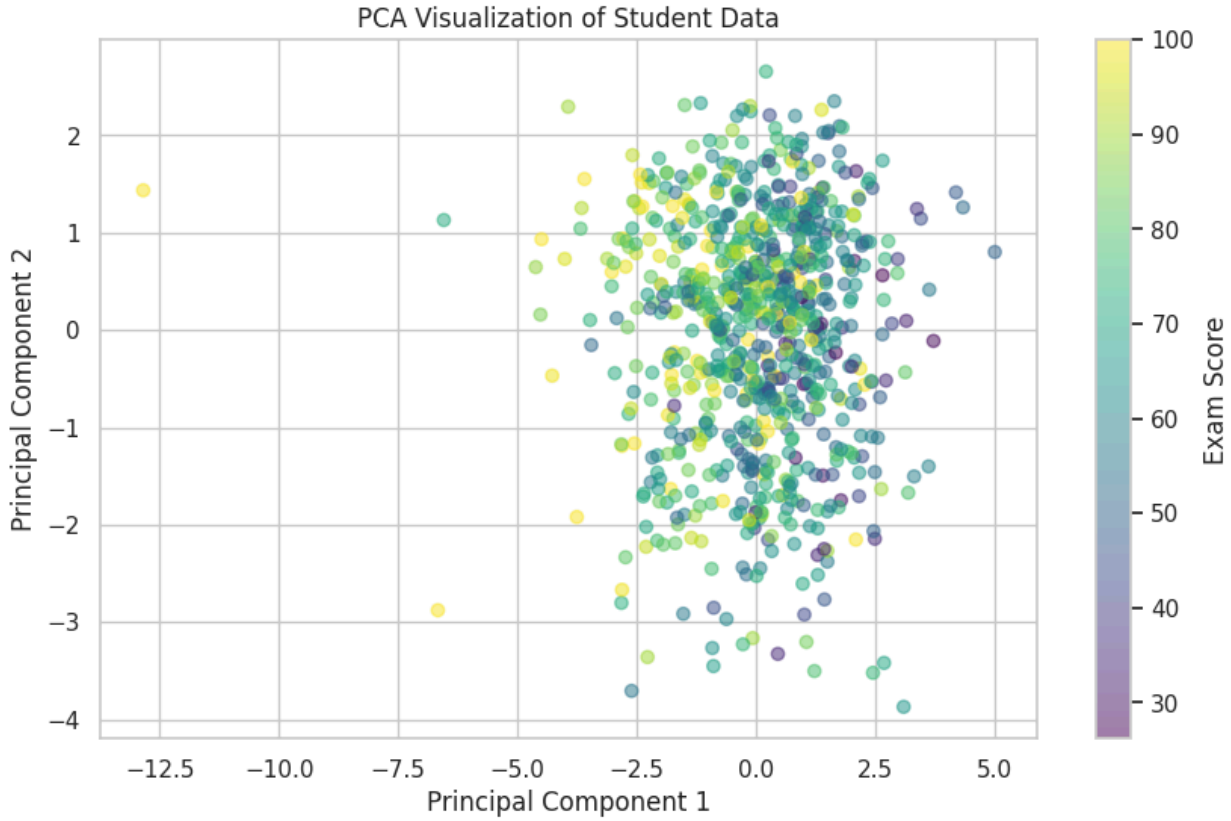
To enhance model performance, two new features were created:

- total_leisure_hours: Combines social media and Netflix usage
- productivity_ratio: Compares study hours to leisure hours.

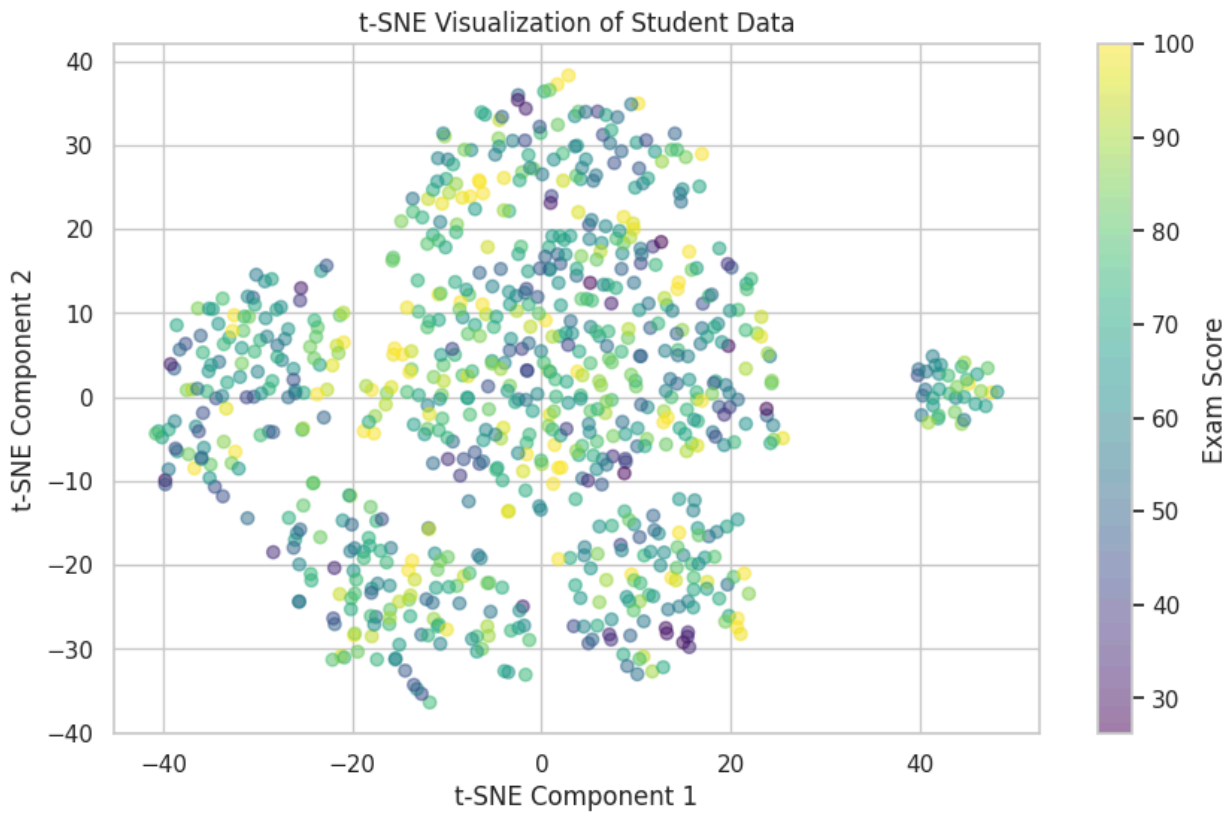
Categorical variables were converted to a numerical format using one-hot encoding.

To standardize the data, all features were normalized using StandardScaler, ensuring that models relying on distance metrics or gradient descent would perform optimally.

Dimensionality reduction techniques, such as PCA and t-SNE, were used to explore the structure of the data:



PCA revealed that most data points were tightly clustered with no clear separation based on exam scores, suggesting that performance differences are not easily captured through linear combinations of features.



t-SNE uncovered several distinct clusters, indicating that student behaviors and characteristics may segment into meaningful groups, which could potentially influence academic performance.

Model Analysis:

The dataset was split into training and test sets (80/20 ratio) to assess model generalization. Feature selection was conducted using SelectKBest with an f_regression score function to identify the top 10 features most predictive of student performance. These included features related to study habits, screen time, attendance, sleep, and health.

```
Top 10 selected features:
Index(['study_hours_per_day', 'social_media_hours', 'netflix_hours',
       'attendance_percentage', 'sleep_hours', 'exercise_frequency',
       'mental_health_rating', 'total_leisure_hours', 'productivity_ratio',
       'diet_quality_Poor'],
      dtype='object')
Linear Regression: Mean R2 = 0.894
Random Forest: Mean R2 = 0.856
XGBoost: Mean R2 = 0.848
SVR: Mean R2 = 0.619
MLP: Mean R2 = 0.847
```

Among the models evaluated using 5-fold cross-validation, Linear Regression achieved the highest mean R^2 score of 0.894, indicating a strong linear relationship between the features and exam scores. Random Forest (R^2 = 0.856) and XGBoost (R^2 = 0.848) also performed well but were slightly less effective than the simpler linear model. More complex models like SVR and MLP showed lower performance.

To optimize model performance, GridSearchCV was used to tune the hyperparameters of the top-performing models:

- Random Forest: Number of trees, tree depth, and split criteria
- XGBoost: Tree depth, learning rate, and number of estimators.
  These optimizations used 5-fold cross-validation to prevent overfitting.

Results:

```
Random Forest (tuned) Test Performance
R2 Score: 0.854
RMSE: 6.121

XGBoost (tuned) Test Performance:
R2 Score: 0.866
RMSE: 5.859
```

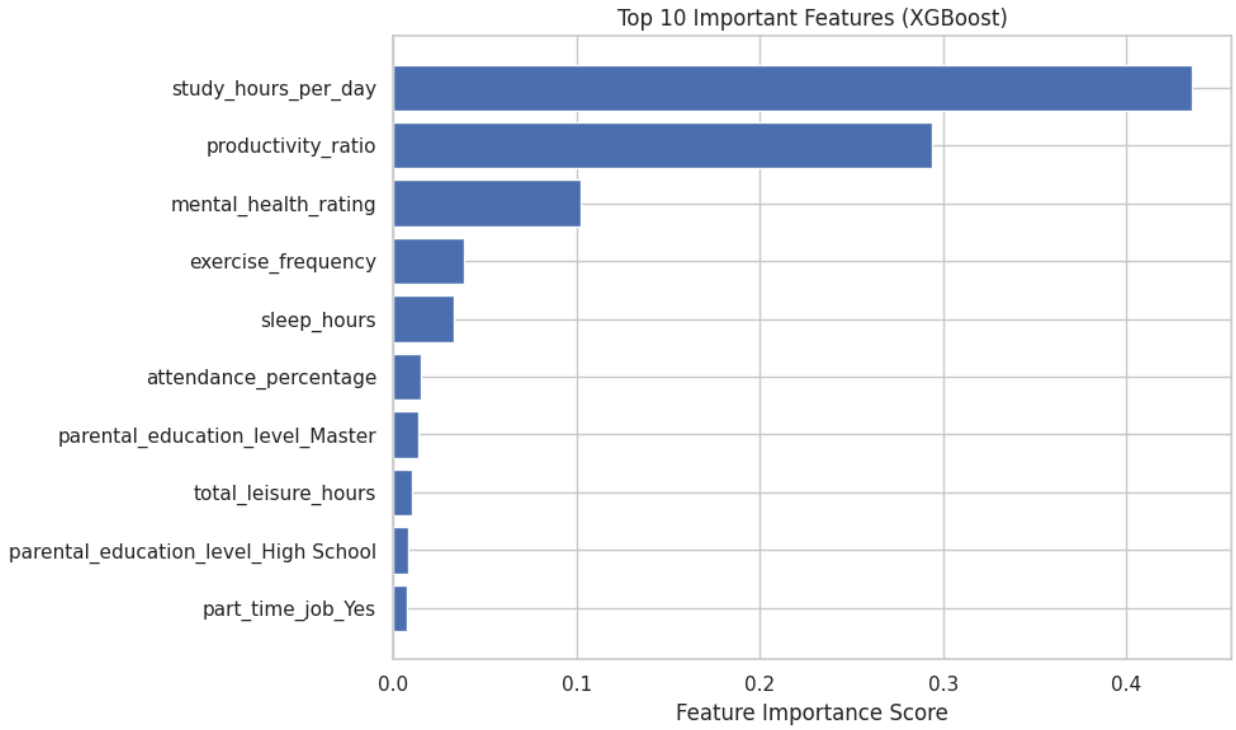
Test Set Evaluation

The tuned Random Forest and XGBoost models were evaluated on an unseen test set. The performance metrics were:

- Random Forest: $R^2 = 0.854$, RMSE = 6.12
- XGBoost: $R^2 = 0.866$, RMSE = 5.859

Both models demonstrated strong generalization ability, with high R^2 scores and low RMSE, indicating that they performed well in predicting student scores.

Feature Importance:



The feature importance visualization revealed the following:

- Study hours per day had the highest importance score (0.45), indicating a strong relationship with academic performance

- Productivity ratio was the second most important feature (0.285), emphasizing the importance of balancing study time and leisure.
- Mental health rating had a moderate importance score (> 0.1), suggesting that well-being impacts academic success.
- Other features, like social media hours and Netflix hours, had lower importance, but still contributed to the model's predictions.

Actual vs Predicted Plot

The scatter plot of Actual vs. Predicted exam scores showed that the XGBoost model's predictions were fairly accurate, with most points clustered close to the diagonal line, indicating minimal prediction errors and strong model performance.

Limitations and Future Directions

While this project provided valuable insights into the relationship between lifestyle habits and academic performance, there are several limitations worth noting:

- Data Source Constraints: The dataset used was limited in size and scope, focusing on a specific population. It may not generalize well to different educational systems, cultural contexts, or age groups.
- Self-Reported Data: Many of the features, such as study hours, mental health rating, and sleep duration, were likely self-reported. This introduces potential biases or inaccuracies due to overestimation, underestimation, or misreporting.
- Feature Coverage: Although the dataset included many important lifestyle factors, other potentially significant variables, such as teaching quality, peer influence, stress levels, or financial background, were not included and may have further improved model accuracy.
- Temporal Limitation: The data represents a snapshot in time rather than a longitudinal view. Academic performance is often influenced by long-term habits and changes over semesters or years, which were not captured here.

Future Directions:

- Longitudinal Studies: Future work could involve tracking students over time to better understand how changing habits affect academic performance.
- Expanded Feature Set: Including more nuanced variables such as stress levels, motivation, course difficulty, and quality of instruction could deepen the model's understanding.
- Personalized Recommendations: Once robust models are developed, they could be used to provide personalized lifestyle recommendations to students to improve their academic outcomes.
- Integration with Educational Platforms: Embedding such predictive models into school or university learning management systems could enable proactive interventions for at-risk students.

Conclusion

This study demonstrates that lifestyle habits, particularly study time, productivity balance, and mental health, play a significant role in predicting students' academic performance. By leveraging machine learning models such as XGBoost and Random Forest, we were able to identify and quantify the impact of these factors on exam outcomes. The analysis suggests that students who manage their study time effectively, limit excessive leisure activities, and maintain good mental health are more likely to succeed academically.

Through this project, I gained hands-on experience with the full machine learning pipeline, including data preprocessing, feature engineering, model training, evaluation, and optimization. Notably, simple models like Linear Regression performed surprisingly well, which reinforced the principle that model complexity should be justified by data behavior, not assumptions. Feature engineering, such as the creation of the productivity ratio, proved critical in improving model accuracy, and exploratory techniques like PCA and t-SNE offered valuable insights into student behavior clustering.

In addition to technical growth, this project underscored the ethical responsibility of using educational data carefully and transparently. Predictive models must be designed to support, not penalize, students, offering insights and interventions without bias.

Overall, the key takeaway is that academic success is multi-faceted, and while habits like consistent studying are critical, broader lifestyle patterns and personal well-being also matter greatly. With further refinement and larger, more diverse datasets, such models could one day assist educators in proactively identifying and supporting students who may be at risk of underperforming.

Reference

[1] Heidari, Mohammad et al. “Relationship of Lifestyle with Academic Achievement in Nursing Students.” *Journal of clinical and diagnostic research : JCDR* vol. 11,3 (2017): JC01-JC03. doi:10.7860/JCDR/2017/24536.9501