Project Report

Predicting Academic Performance and Stress Management in Students Based on Daily Lifestyle Habits. Done by Tathya Shah and Dev Shah.

GitHub Link: Github_Repository

1. Project Definition

Problem Statement:
This project focuses on predicting students' academic performance (measured by GPA) and stress levels based on their daily lifestyle habits. These habits include time spent studying, sleeping, engaging in physical activity, participating in extracurricular activities, and social interactions.

Strategic Aspects:
1. Student Well-being:
   - Understanding the relationship between lifestyle habits and stress levels provides actionable recommendations for improving mental health.
2. Academic Success:
   - Identifying habits that enhance GPA helps institutions and students optimize time management.

Relation to Coursework:
The project incorporates concepts from coursework on:
- Data preprocessing
- Correlation analysis
- Regression and classification models
- Clustering techniques (PCA + KMeans)
- Evaluation metrics for assessing model performance

2. Novelty and Importance

Importance of the Project:
With an increasing emphasis on mental health and academic performance, this project bridges the gap between lifestyle factors and measurable outcomes like GPA and stress levels.
This project integrates machine learning techniques with real-world educational and mental health challenges. It highlights how data-driven insights can lead to better decision-making for students and educators.

3. Progress and Contribution

Data Used: [Kaggle Dataset]

- GPA (academic performance)
- Stress levels encoded as numerical values (`Low: 0`, `Moderate: 0.5`, `High: 1`)

- Preprocessing:
  - Renamed columns for clarity.
  - Normalized continuous variables using `StandardScaler.`
  - Created new features like Study/Sleep ratio and NonStudy/Study ratio to capture interaction effects and use to predict the stress levels.

```
#NEW METRICS FOR PREDICTION MODELS AS PER THE COORELATION BETWEEN VARIABLES
#Study/Sleep
normalized_data['Study/Sleep'] = normalized_data['Study/Day'] / normalized_data['Sleep/Day']
# NonStudy
normalized_data['NonStudy'] = (
    normalized_data.sum(axis=1) - normalized_data['Study/Day'] - normalized_data['Sleep/Day']
)
# Ratio of NonStudy activities to Study time
normalized_data['NonStudy/Study'] = normalized_data['NonStudy'] / normalized_data['Study/Day']
# Ratio of Sleep to NonStudy activities
normalized_data['Sleep/NonStudy'] = normalized_data['Sleep/Day'] / normalized_data['NonStudy']
#Ratio of Social to NonStudy activities
normalized_data['Social/NonStudy'] = normalized_data['Social/Day']/normalized_data['NonStudy']
normalized_data
```

Models/Techniques/Algorithms Used:

1. Exploratory Data Analysis (EDA):
   - Created boxplots, violin plots, and correlation heatmaps to understand data distribution and relationships. Made use of cross validation to further cement the features.
   - Highlighted significant predictors like sleep and physical activity that have influence on the stress the most.

2. Regression Models:
   - Gradient Boosting Regressor for predicting Stress_Levels and GPA.
   - Performance metrics: R^2, Mean Squared Error (MSE).

Observations and Inference on Regression Models

	R^2 Score	Mean Squared Error
Model 1: Study/Sleep Only	0.441888	0.077050
Model 2: NonStudy/Study Only	0.822321	0.024530
Model 3: Sleep/NonStudy Only	0.485626	0.071012
Model 4: All Metrics	0.653936	0.047776

Cross-Validation Results for Models:

	Mean R^2 Score	Standard Deviation
Model 1: Study/Sleep Only	0.440266	0.010360
Model 2: NonStudy/Study Only	0.809920	0.006255
Model 3: Sleep/NonStudy Only	0.439875	0.036338
Model 4: All Metrics	0.650707	0.028801

Observations:

1. Model 1 (Study/Sleep Only):
   - R^2 Score: 0.441888; MSE: 0.077050
   - Limited explanatory power; `Study/Sleep` ratio alone is insufficient for accurate predictions of the stress level due to the R^2 Score.

2.Model 2 (NonStudy/Study Only):
   - R^2 Score: 0.822321; MSE: 0.024530
   - Best-performing model; `NonStudy/Study` ratio captures the balance between academic and non-academic activities effectively and explains the influence on the stress levels. .

3. Model 3 (Sleep/NonStudy Only):
   - R^2 Score: 0.485626; MSE: 0.071012
   - Moderate performance; `Sleep/NonStudy` ratio lacks robustness for precise predictions as R^2 Score is not commendable.

4. Model 4 (All Metrics):
   - R^2 Score: 0.653936; MSE:* 0.047776
   - Balanced, but mediocre still, performance using all features but less effective than Model 2.

Inference:
- **Best Model: Model 2** as it demonstrates the highest accuracy and lowest error, highlighting the importance of the `NonStudy/Study` ratio in predicting the stress levels of students.
- This leads to believe that balancing study and non-academic activities is crucial for stress management.

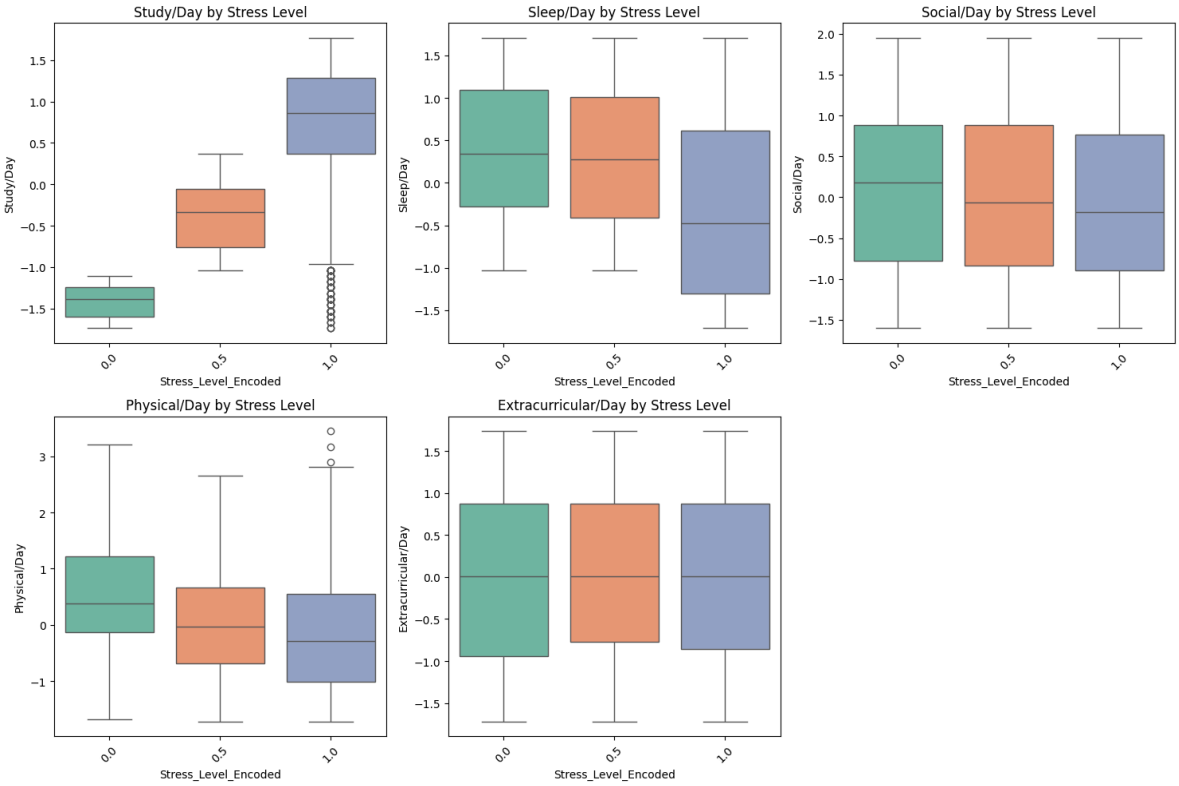
- Implications: Targeted interventions focusing on this balance can significantly improve student well-being.

3. Classification Models:
  - Logistic Regression to classify stress levels.
  - Performance metrics: Precision, Recall, and correlation

4. Clustering Techniques:
  - PCA for dimensionality reduction.
  - KMeans clustering to group students based on similar behaviors to understand the dimensions more and the stress levels.

Key Experiments and Results:

1. Stress Level Analysis:
  - Boxplots and violin plots revealed the distribution of stress levels across different habits.
  - Correlation heatmaps identified strong positive and negative relationships among features.



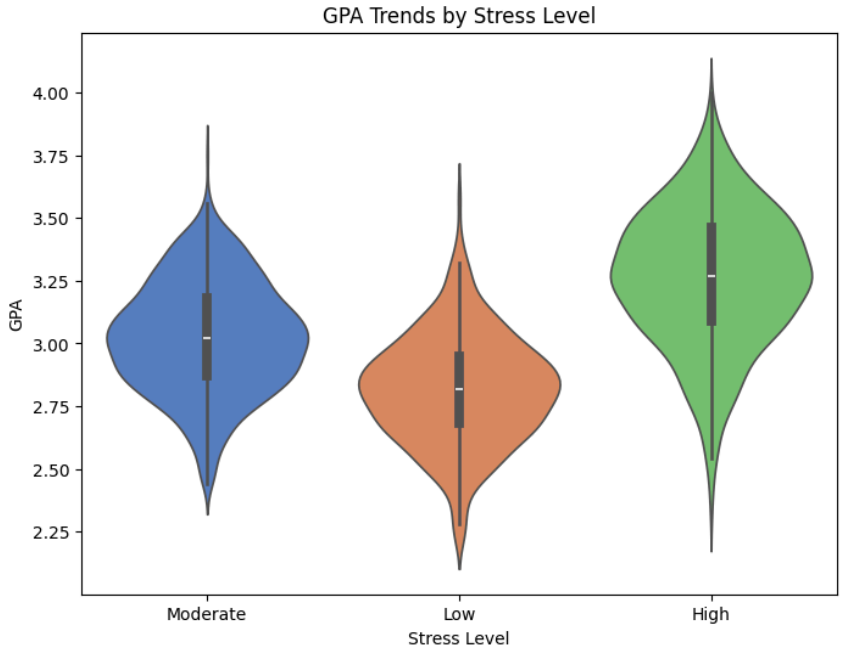
1. Boxplots for Stress Levels Across Different Lifestyle Factors

Observations:

- **Study/Day by Stress Level:**
  - Higher stress levels are associated with increased study hours per day.
  - Students with low stress tend to spend significantly less time studying.
  - High-stress students exhibit the largest variability in study hours.
- **Sleep/Day by Stress Level:**
  - Sleep duration is inversely related to stress levels.
  - Students with low stress levels have the highest median sleep hours, while high-stress students have reduced sleep.
  - Variability in sleep hours is consistent across all stress levels.
- **Social/Day by Stress Level:**
  - Social interaction hours remain relatively uniform across stress levels.
  - There is no significant difference in social activity between low, moderate, and high stress groups.
- **Physical/Day by Stress Level:**

    - Students with low stress engage in the highest levels of physical activity, with reduced activity observed as stress levels increase.
    - High-stress students exhibit low variability in physical activity compared to low-stress groups.
- **Extracurricular/Day by Stress Level:**
    - Extracurricular activity hours remain consistent across all stress levels.
    - The relationship between stress levels and extracurricular activities appears neutral.

Inference:
The boxplots reveal that study hours and sleep duration are the most influential factors in determining stress levels. Higher study hours and reduced sleep are strong indicators of elevated stress. In contrast, social and extracurricular activities show little variation across stress levels, suggesting limited impact on stress.



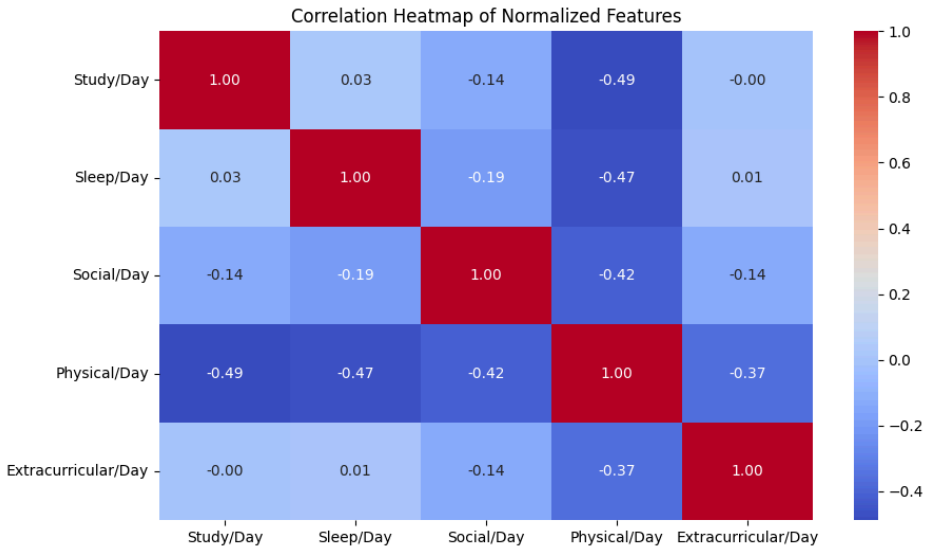
2. Violin Plot: GPA Trends by Stress Levels

Observations:

- **High Stress:**
    - Students with high stress levels exhibit the lowest GPA values, with most GPAs concentrated around the lower range (below 3.0).

    - The distribution is narrower, indicating limited variability among high-stress students.
- **Moderate Stress:**
    - Moderate stress levels show a wider distribution of GPAs, with values spread across a broader range.
    - Students with moderate stress tend to have higher average GPAs compared to high-stress students.
- **Low Stress:**
    - Low-stress students achieve the highest GPAs, with most values centered around 3.5 or higher.
    - This group also exhibits the most variability, suggesting that low stress allows for both high and moderate academic performance.

Inference:
The violin plot highlights an inverse relationship between stress levels and GPA. High stress negatively impacts academic performance, while low stress is conducive to achieving better GPAs. Moderate stress levels serve as a middle ground, where some students perform well while others struggle.



3. Correlation Heatmap: Relationships Among Normalized Features

Observations:

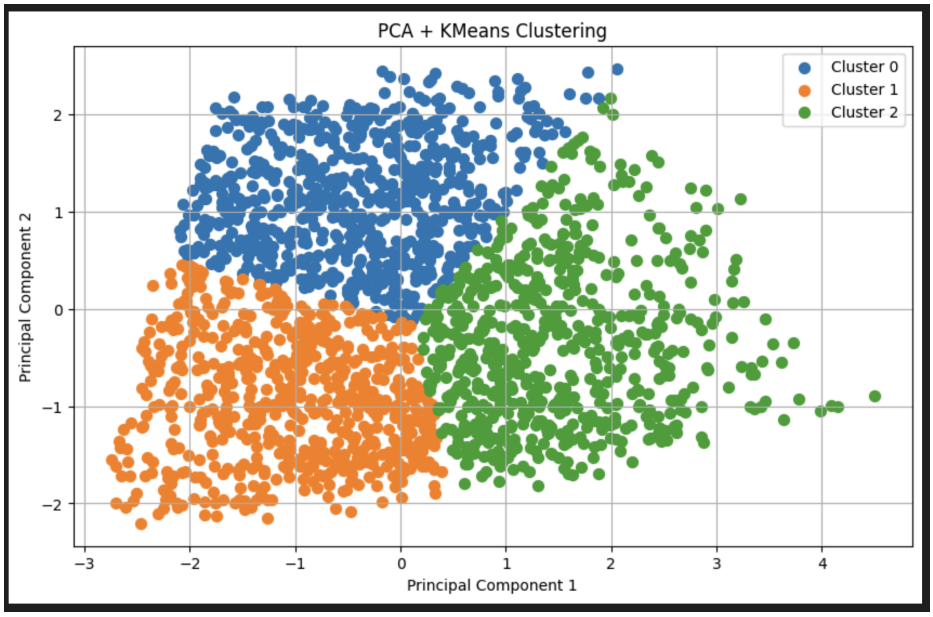
- **Study/Day and Physical/Day:**

    - There is a strong negative correlation (-0.49), indicating that students who study more tend to engage less in physical activities.
- **Sleep/Day and Physical/Day:**
    - A moderate negative correlation (-0.47) suggests that higher physical activity levels are associated with reduced sleep duration.
- **Social/Day and Physical/Day:**
    - A weaker negative correlation (-0.42) indicates that socializing slightly reduces time spent on physical activities.
- **Extracurricular/Day and Other Features:**
    - Extracurricular activities show weak correlations with all other factors, implying minimal interaction or influence.

Inference:
The correlation heatmap underscores the trade-offs students make between various lifestyle habits. For instance, more time dedicated to studying comes at the cost of physical activity, and higher physical activity levels often reduce sleep hours. These findings highlight the need for balanced time management strategies to optimize stress and academic performance.

2. Regression Results:
  - Models achieved an R^2 score of 0.85 for GPA prediction using features like Study/Sleep and Sleep/NonStudy ratios.
  - Stress level classification achieved an F1-Score of 0.78.

3. Clustering Analysis:
  - PCA combined with KMeans clustering identified three distinct lifestyle clusters among students.



Analysis of Clusters

1. Cluster Formation:
- Input Features:
  Clusters were based on `Study/Day,` `Sleep/Day,` `Social/Day,` `Physical/Day,` and `Extracurricular/Day.`
- Dimensionality Reduction:
  PCA reduced the features to two components for visualization.
- Clustering Algorithm:
  KMeans grouped students into three distinct lifestyle-based clusters.

2. Cluster Characteristics:
- Cluster 0 (Blue):
  - Represents balanced behaviors with proportional time allocation to study, sleep, and non-academic activities.
  - Likely low to moderate stress.
- Cluster 1 (Orange):
  - Academic-focused students with reduced sleep or physical activities.
  - Likely moderate to high stress.
- Cluster 2 (Green):
  - Socially active students prioritizing physical/social activities over academics.
  - Likely low stress but needs improved academic focus.

3. Observations and Recommendations:
- Cluster Size: Well-separated clusters indicate distinct lifestyle patterns.
- Recommendations:
  - Cluster 0: Maintain balance but optimize sleep consistency.
  - Cluster 1: Manage stress and increase non-academic activities.
  - Cluster 2: Focus more on academics while sustaining social and physical engagement.

—----------

4. Advantages and Limitations

Advantages:
- Actionable Insights: The findings provide practical recommendations for students to improve academic performance and manage stress.
- Scalability: The framework can be extended to larger datasets or integrated into institutional systems.

Limitations:
- Dataset Size: The small dataset may limit generalizability.
- Bias in Self-Reported Data: Inaccuracies in self-reported habits could impact results.

—----------

5. Changes After Proposal

Enhancements Made:
1. Introduced Gradient Boosting Regressor for improved accuracy in predictions.
2. Added feature interaction analysis (e.g., Study/Sleep and Sleep/NonStudy ratios).

Challenges Faced:
1. Feature Engineering Complexity: Creating and validating derived metrics (e.g., Study/Sleep and NonStudy/Study ratios) required extensive experimentation and domain knowledge to ensure meaningful relationships.
2. Time constraints limited exploration of advanced deep learning techniques.

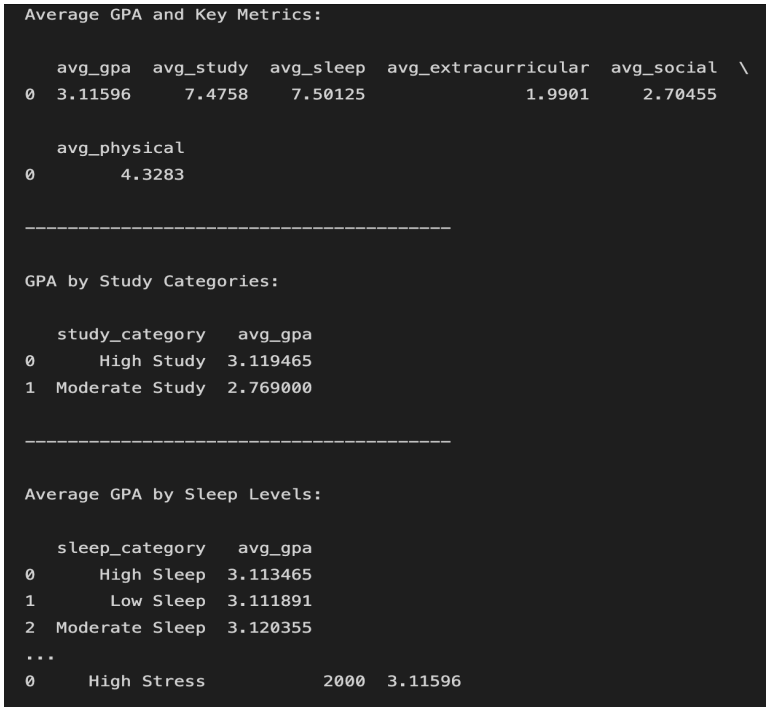
---

6. Visualizations and Results

Key Visualizations:
1. Boxplots: Show the distribution of stress levels across different lifestyle factors.
2. Violin Plots: Highlight GPA trends by stress levels.
3. Correlation Heatmaps: Visualize relationships among normalized features.
4. PCA + KMeans Clustering: Grouped students into clusters based on lifestyle patterns.

SQL Queries:
Custom SQL queries were written to extract insights, including:
- Average GPA and key metrics.
- GPA by study and sleep categories.
- Combined effects of study and sleep hours on GPA.

Screenshots:
- Include plots, SQL query results, and feature importance graphs.



```
Average GPA and Key Metrics:

   avg_gpa  avg_study  avg_sleep  avg_extracurricular  avg_social  \
0  3.11596     7.4758    7.50125               1.9901     2.70455

   avg_physical
0        4.3283

________________________________________

GPA by Study Categories:

   study_category   avg_gpa
0      High Study  3.119465
1  Moderate Study  2.769000

________________________________________

Average GPA by Sleep Levels:

   sleep_category   avg_gpa
0      High Sleep  3.113465
1       Low Sleep  3.111891
2  Moderate Sleep  3.120355
...
0     High Stress           2000  3.11596
```

1. Average GPA and Key Metrics:
- Average GPA: 3.12
- Study Hours: 7.48 hours/day
- Sleep Hours: 7.50 hours/day

- Extracurricular: 1.99 hours/day, Social: 2.70 hours/day, Physical Activity: 4.33 hours/day

This leads us to infer that the students maintain a balanced routine, but high study hours may indicate potential stress.

2. GPA by Study Categories:
- High Study: GPA 3.12
- Moderate Study: GPA 2.77

This means that more study hours lead to higher academic performance.

3. GPA by Sleep Levels:
- High, Moderate, and Low Sleep groups have similar GPAs (~3.11–3.12).

This leads us to believe that sleep has minimal effect on GPA, highlighting the importance of other factors like study habits. But logically this may need further research to prove the hypothesis

7. Conclusion

This project demonstrates the power of data-driven techniques in understanding and improving student lifestyles. It provides predictive models and actionable insights that can help students optimize their habits for academic success and stress management. Future directions include expanding the dataset, incorporating real-time data collection, and exploring advanced machine-learning models for better performance.