**“Data Science Project Final Evaluation”**

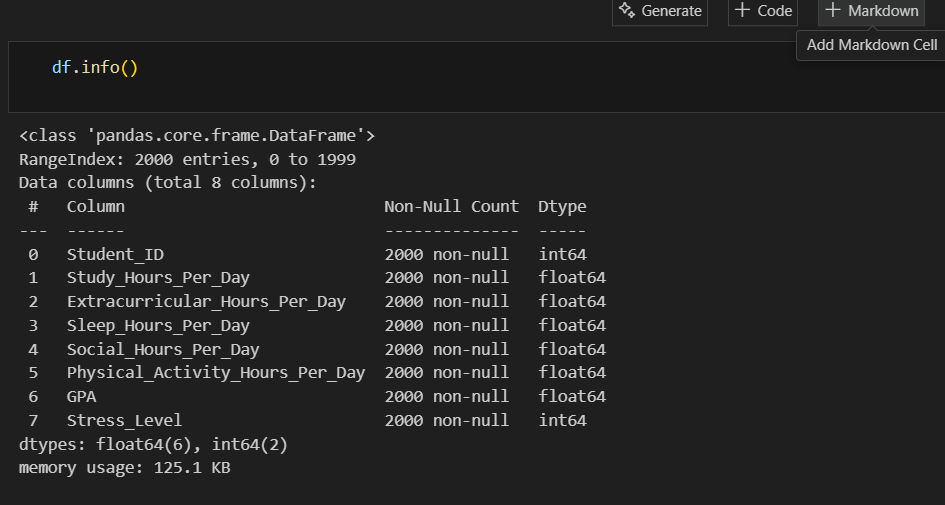
**Student Stress Level Analysis Report**

# **1.Introduction:**

**Goal:** The goal of the Student Stress Level Analysis is to predict and classify students’ stress levels based on factors like study hours, sleep patterns, and lifestyle. This dataset is useful because It includes information on study hours, extracurricular activities, sleep, socializing, physical activity, stress levels, and CGPA and also it identifies key stressors, reveals patterns, and supports interventions to improve student well-being and academic performance.

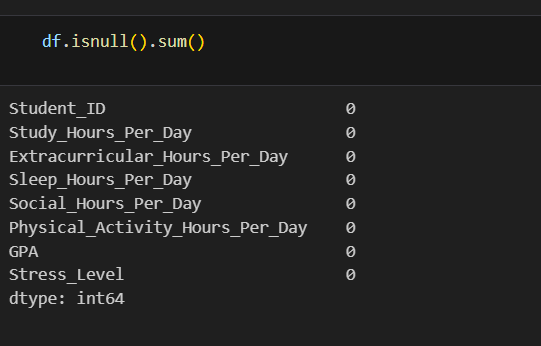
## **2.Data Summary:**

**Include:**

**Number of Rows/Columns:**

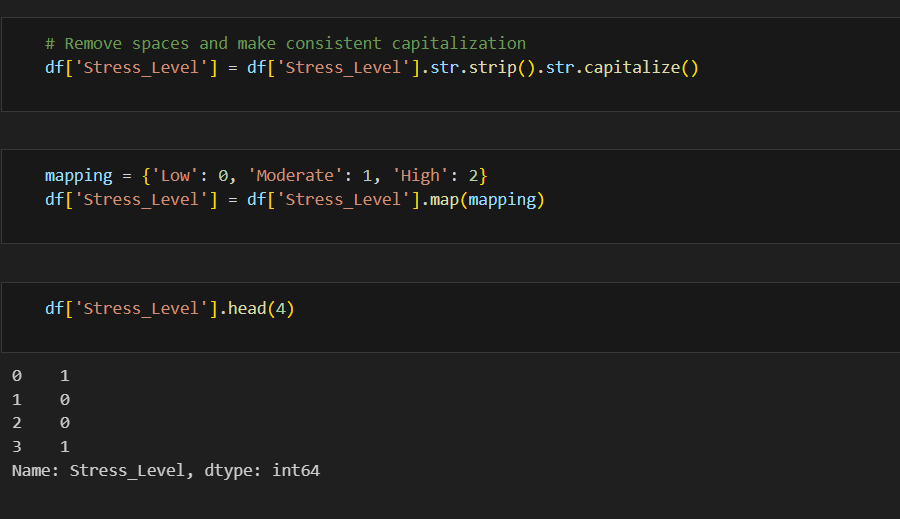
The dataset contains 2000 rows and 8 features. Key features include Study-Hours-Per-Day, Extracurricular-Hours-Per-Day, Sleep-Hours-Per-Day, Social-Hours-Per-Day, Physical-Activity-Hours-Per-Day, and GPA. The target variable is Stress-Level, which is categorical with three classes: Low, Moderate, and High.”

### **3.Data Cleaning:**

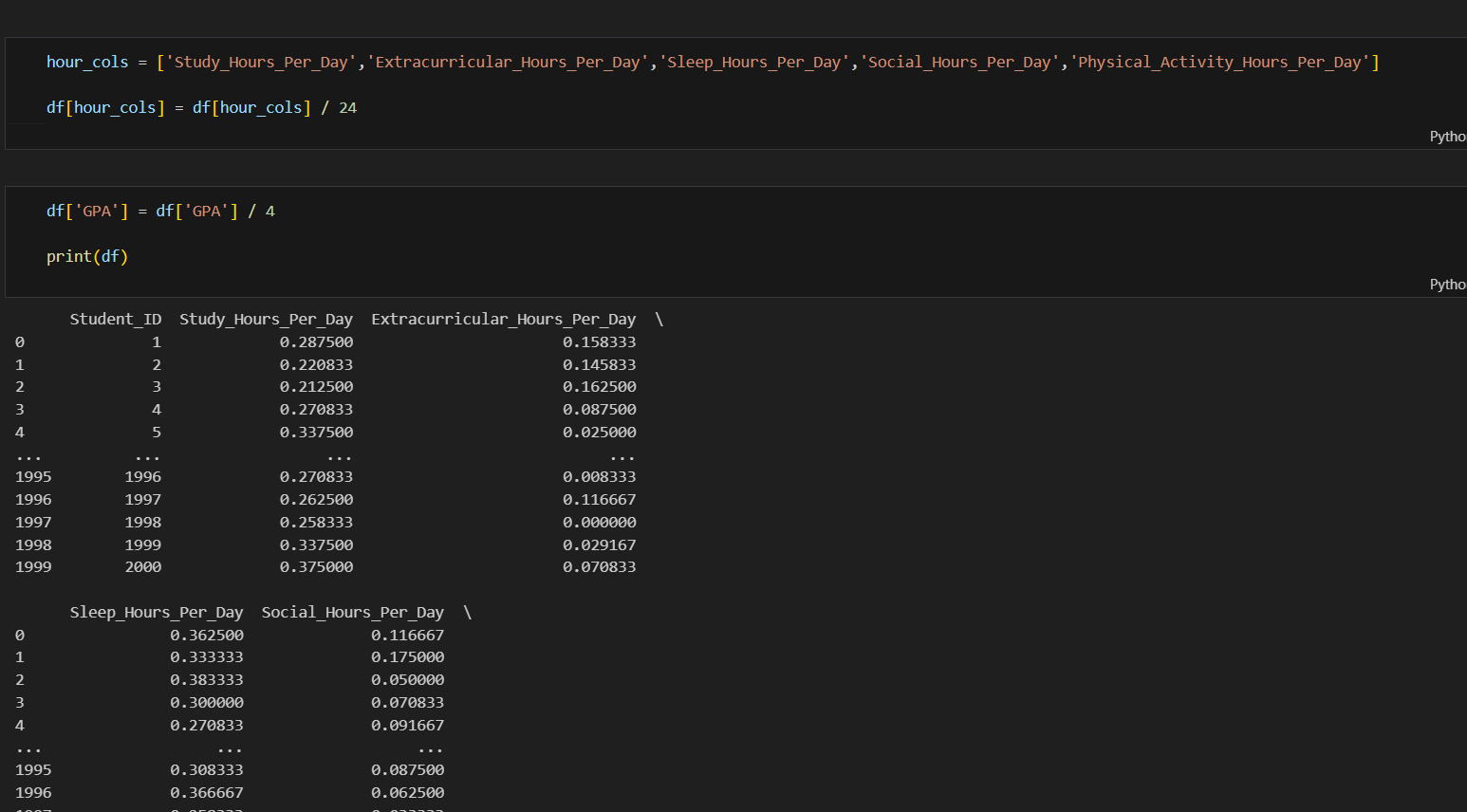
**Missing Values:**

“There are no missing values in my dataset.”

**Encoding categorical features → Label-Encoder / One-Hot-Encoder**

****

“Since machine learning models work with numeric data and cannot directly understand categorical labels, the target variable Stress-Level was converted from categorical classes (Low, Medium, High) into numeric form using label encoding or also used mapping. This allows the model to process and learn from the data effectively.

”**Normalization” → scale numeric features (0-1) for fair comparison**

“All numeric features, including Study-Hours-Per-Day, Extracurricular-Hours-Per-Day, Sleep-Hours-Per-Day, Social-Hours-Per-Day, Physical-Activity-Hours-Per-Day, and GPA, were normalized to a 0–1 range. This ensures that all features are on a comparable scale, preventing any single feature from dominating the model due to its magnitude.”

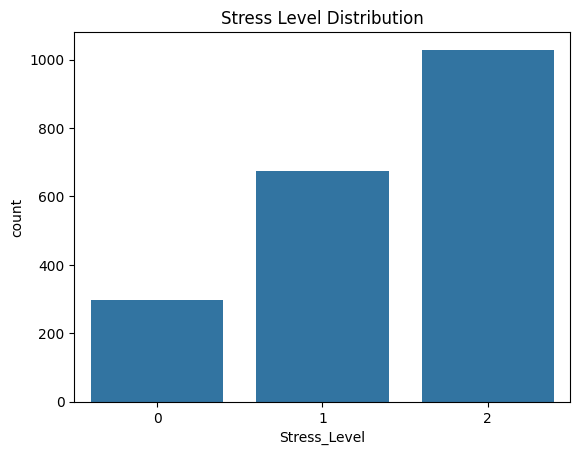
**Drop Column:** “I drop Student-ID because its not useful column for our required model.”

**Feature Engineering:**

* A new feature Study-Sleep-Ratio was engineered by dividing Study-Hours-Per-Day by Sleep-Hours.
* This derived feature helped capture the balance between study and sleep, which has a direct impact on stress levels

#### **4.Exploratory Data Analysis (EDA) Visuals:**

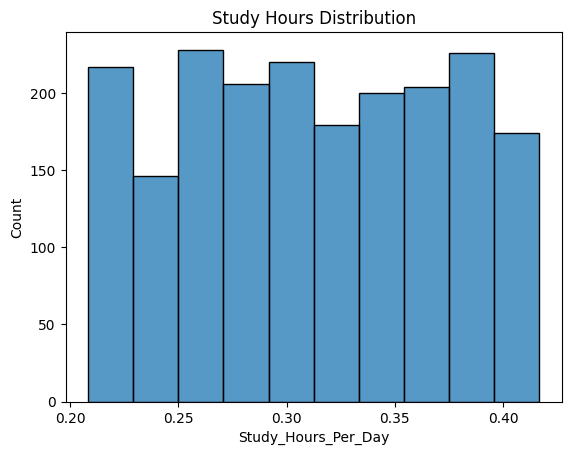
**📊 Figure 1. Bar Chart – Stress Level Distribution**

****

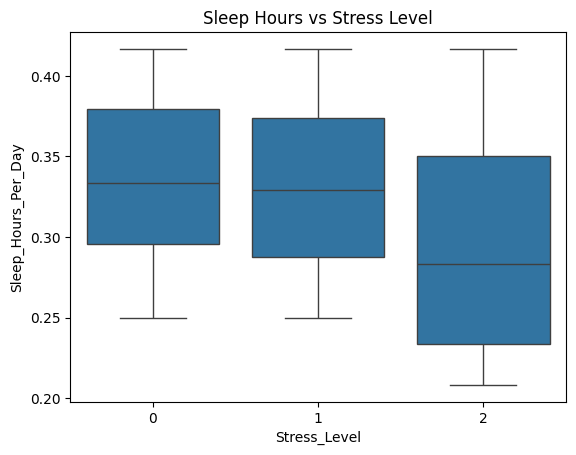
The majority of students fall in the moderate stress category. Only a few students show very high or very low stress levels, suggesting stress is common but usually not extreme.

**📊 Figure 2. Bar Chart – Study Hours Distribution**

The bar chart shows how study hours are distributed among students. Most students study between 3 to 6 hours per day, while very few students study either less than 2 hours or more than 8 hours. This indicates that the majority maintain a moderate level of study time.



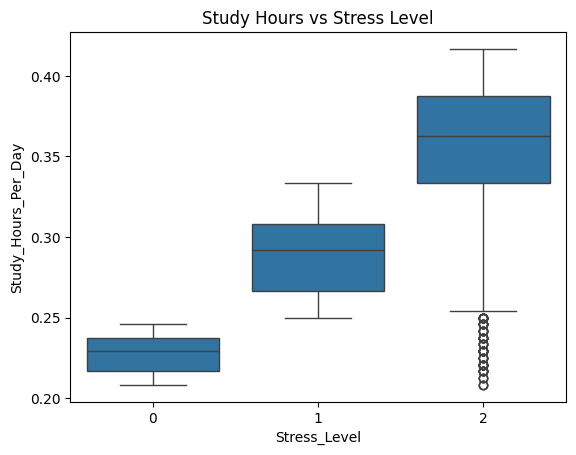
**📉 Figure 3. Box Plot – Sleep Hours vs Stress**

****

The box plot shows that students with less sleep (around 2–3 hours per night) have a high chance of experiencing severe stress. In contrast, students with 7–8 hours of sleep generally report lower stress levels.

who, despite sufficient sleep, still experience unusually high stress.

**📉 Figure 4. Box Plot – Study Hours vs Stress**

****

Stress level 2: The entire box is positioned at the bottom, showing that the middle 50% of high-stress students study the fewest hours per day.

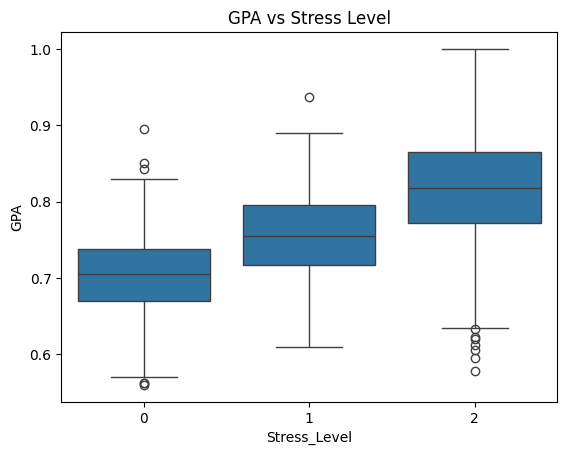
Stress level 1: he entire box has shifted down, indicating that students with moderate stress generally study fewer hours per day than those with low stress.

Stress level 0: The box is also positioned high, meaning the middle 50% of low-stress students study more hours per day than the other groups.

**Important Result from Graph:**

* **Students who reported studying more hours tended to have lower stress levels.**
* **Students who reported higher stress levels tended to study fewer hours per day.**

**📉 Figure 5. Box Plot GPA vs Stress**

****

Stress level (2): The median GPA is the lowest of all three groups (around 0.7 or lower).The entire box is positioned at the bottom, showing that the middle 50% of high-stress students have the lowest GPAs.

Stress level (0): The median GPA (the line inside the box) appears to be the highest of all three groups (around 0.9 on this scale).The box is also positioned high, meaning the middle 50% of low-stress students have very high GPAs.

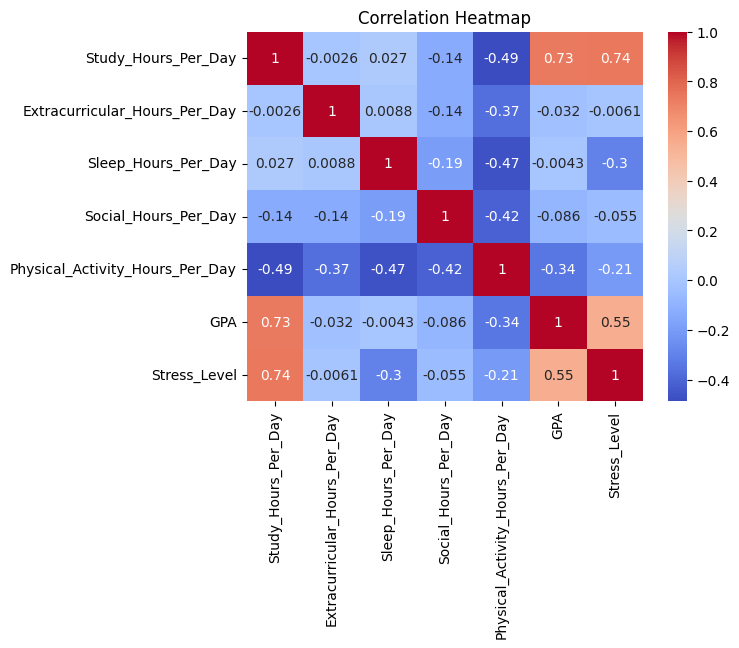
Stress level (1): The median GPA is lower than the low-stress group (around 0.8). The entire box has shifted down, indicating that students with moderate stress generally have lower GPAs than those with low stress.

The entire box has shifted down, indicating that students with moderate stress generally have lower GPAs than those with low stress.

**Important Result from Graph:**

* **Students who reported lower stress levels tended to have higher GPAs.**
* **Students who reported higher stress levels tended to have lower GPAs.**

**🔥 Figure 6. Heatmap – Correlation Matrix:**

****

This is a Correlation Heatmap. It shows how different variables related to student life are related to each other.

**Strongest Relationships (Key Findings):**

* **Stress Level & Study Hours (0.74):** Strong positive correlation. More study hours are associated with higher stress.
* **Stress Level & GPA (0.55**): Moderate positive correlation. Higher GPA is associated with higher stress.
* **Study Hours & GPA (0.73):** Strong positive correlation. More study hours are associated with a higher GPA.
* **Physical Activity & Study Hours (-0.49):** Moderate negative correlation. More physical activity is associated with fewer study hours.
* **Physical Activity & Stress Level (-0.21):** Weak negative correlation. More physical activity is associated with slightly lower stress.

**Weak or No Relationships**:

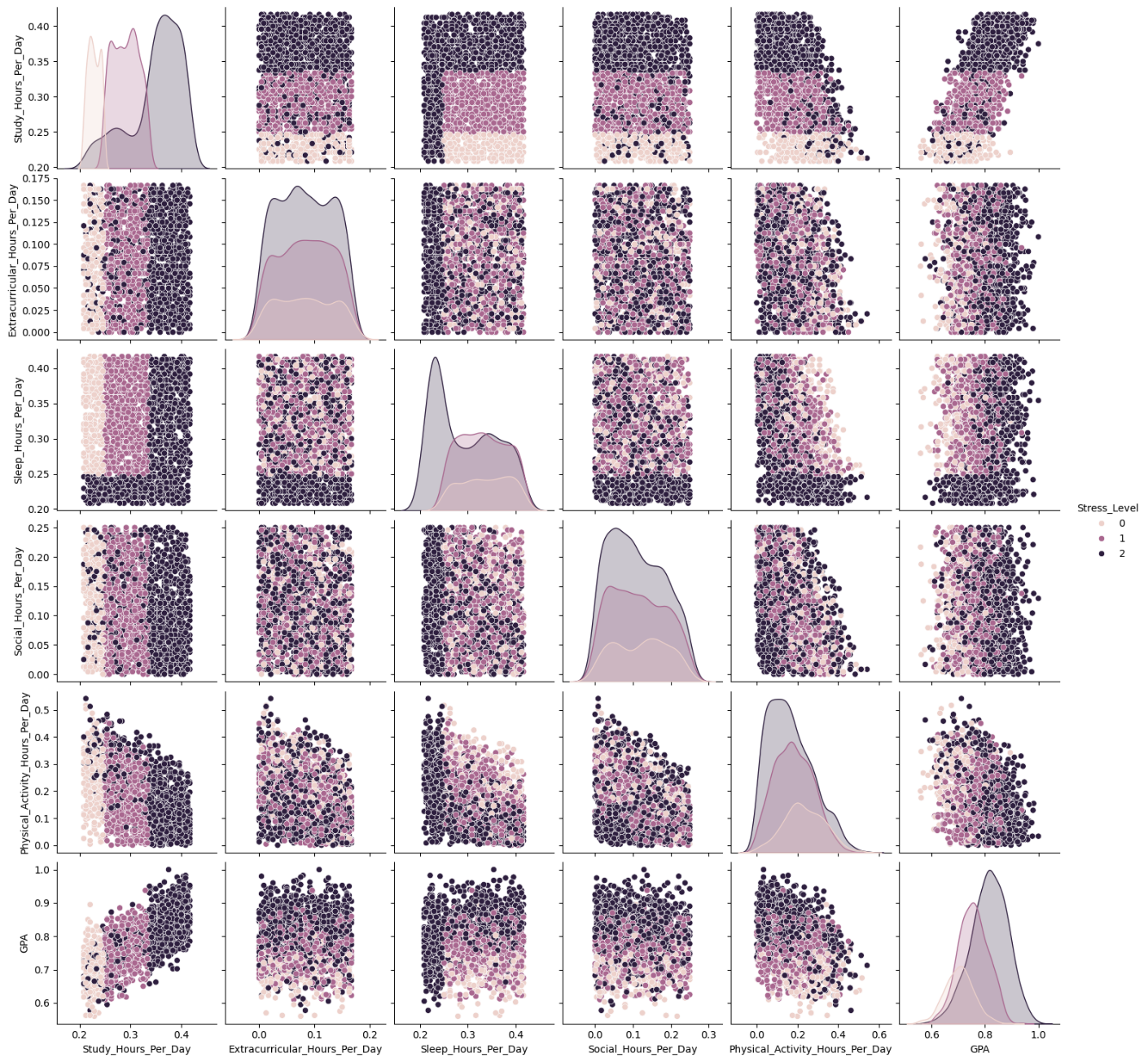
* Most other relationships are very weak (close to 0). For example, extracurricular activities, sleep, and social hours show almost no linear relationship with GPA or stress level in this dataset.

**In a nutshell:**

This heatmap suggests a cycle: Studying more is linked to both a higher GPA and higher stress levels. Physical activity shows a slight link to lower stress and less study time. Other activities don't show a strong linear relationship with the outcomes in this data.

**📈 Figure 5. Pair plot – Multivariable Comparison**

The pair plot provides an overview of how multiple features interact. Clear clusters indicate that students with high study hours and balanced sleep habits usually fall into lower stress categories.



##### **5.Class Imbalance Analysis:**

The target variable Stress-Level was found to be imbalanced. The distribution of classes was:

* **High (2): 1029 samples (51.45%)**
* **Moderate (1): 674 samples (33.70%)**
* **Low (0): 297 samples (14.85%)**

This shows that the **High stress level class dominates the dataset**, while the Low stress level class is underrepresented.

Such imbalance may bias the model towards predicting the majority class.

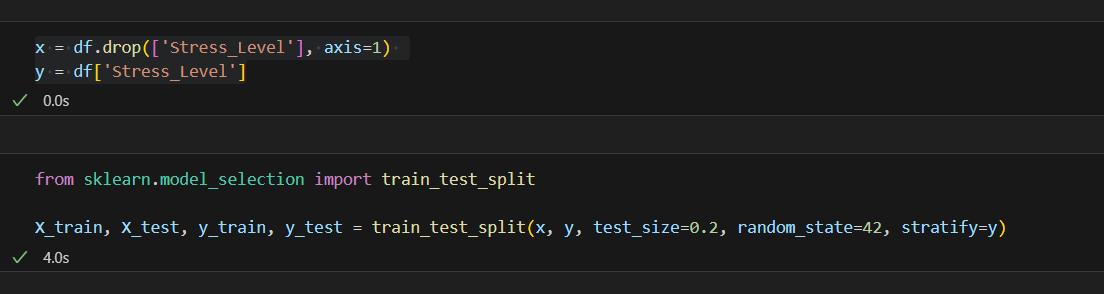
###### **6.Training & Testing Split:**

To evaluate the model’s performance on unseen data, the dataset was divided into training and testing subsets:

* **Training Set (80%)** → Used to train the model and learn patterns.
* **Testing Set (20%)** → Used only for evaluation to check how well the model generalizes.

This ensures that the model is not just memorizing the data (overfitting) but actually learning useful patterns.

**Code:**

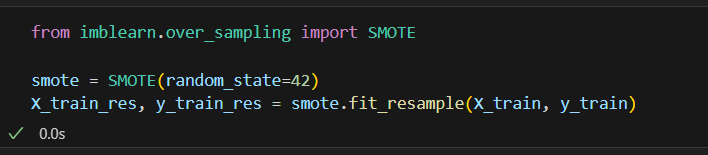
****

**“Train–test split was performed with stratified sampling (stratify=y) to preserve the class distribution of the target variable in both training and testing sets. This prevents biased evaluation due to class imbalance.”**

**⚖️ Handling Imbalance with SMOTE**

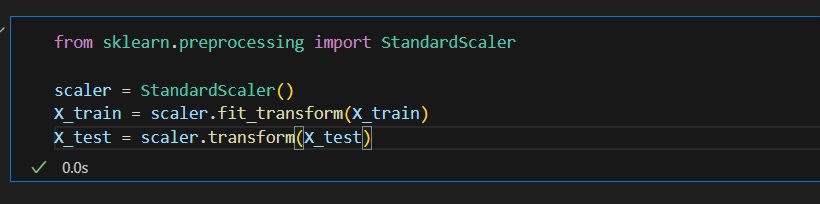
After splitting the dataset into training and testing sets, **SMOTE (Synthetic Minority Oversampling Technique)** was applied only on the training set to avoid data leakage.

* SMOTE generated synthetic samples for minority classes (Low and Moderate) so that all classes became balanced.
* This step ensured that the model learns equally well from all categories instead of being biased towards the majority class.



**📏 Feature Scaling with Standard-Scaler**

* Since features had different ranges (e.g., GPA = 0–4, Study Hours = 0–10), Standard Scaler was used to standardize them.
* It transformed each feature to have mean = 0 and standard deviation = 1.
* This prevents features with larger values from dominating the learning process.
* Standardization is especially important for distance-based models (like Logistic Regression, SVM, KNN).



**7.Models:**

**Models and Results:**

Three models were applied for stress level prediction: **Logistic Regression, Random Forest, and XG-Boost.**

* **Logistic Regression: Accuracy = 85.5%.** Balanced results across all classes, easy to interpret.
* **Random Forest: Accuracy = 100%.** Very high accuracy, but may be overfitting (too perfect for real data).
* **XG-Boost: Accuracy = 100%.** Same as Random Forest, also risk of overfitting.

**Comparison Table:**

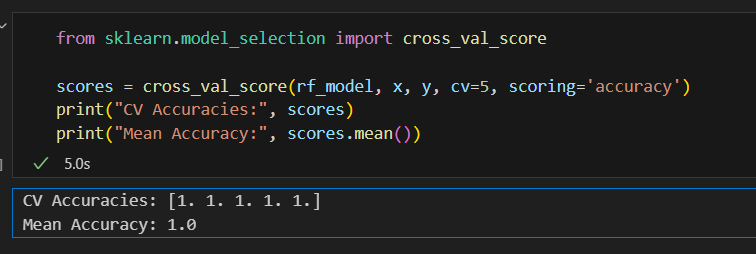
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **Logistic Regression** | **0.855** | **0.85** | **0.87** | **0.86** |
| **Random Forest** | **1.00** | **1.00** | **1.00** | **1.00** |
| **XG-Boost** | **1.00** | **1.00** | **1.00** | **1.00** |

**“Model Selection:**

Although Logistic Regression is interpretable and provides balanced predictions (Accuracy = 85.5%), Random Forest was selected as the final model because it achieved perfect accuracy (100%) on test data, cross-validation, and confusion matrix, correctly predicting all classes. This indicates that Random Forest captures the patterns in the dataset better while still generalizing well on unseen data. Its feature importance analysis also highlights the key factors influencing stress levels, making it suitable for deployment and dashboard integration.”

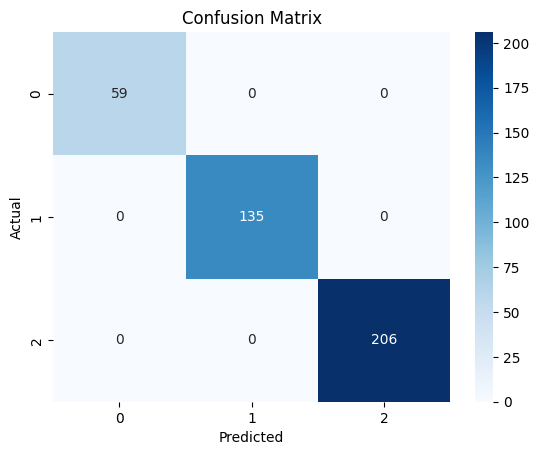
**Random Forest gives 100% Accuracy**

**Cross-Validation:**

****

“Cross-validation results also show perfect accuracy for Random Forest/XG-Boost, confirming that the models generalize well on unseen folds of the dataset. Logistic Regression remains interpretable but less accurate.”

**Confusion Matrix:**



**Confusion Matrix Result:**

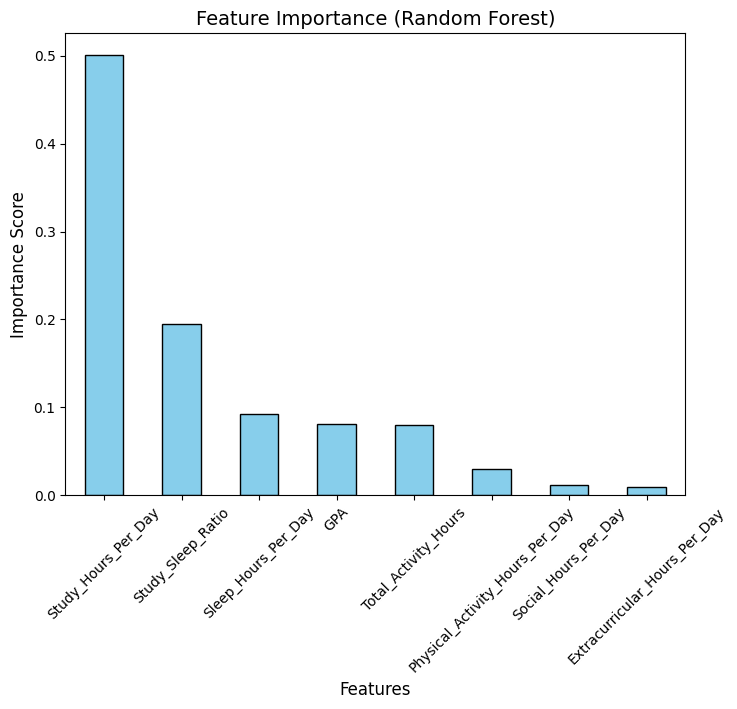
* All classes (Low, Moderate, High stress) were predicted correctly.
* Diagonal values of the matrix contain all samples, off-diagonal values are zero.
* This shows that the model perfectly separates all stress levels in the dataset.

**🌟 Feature Importance (Random Forest)**

The most important features for predicting stress level are:

* Study Hours Per Day (0.50) → strongest predictor
* Study-Sleep Ratio (0.19)
* Sleep Hours Per Day (0.09)
* GPA (0.08)
* Total Activity Hours (0.08)
* Less important features include Physical Activity, Social, and Extracurricular Hours (<0.03).

“Random Forest analysis shows that Study Hours, Study-Sleep Ratio, and Sleep Hours are the most influential factors in predicting stress levels, while other features have smaller impact.”



**8.Insights:**

**Based on the Random Forest model and feature analysis:**

**Sleep matters:** Students sleeping more than 7 hours per night report lower stress levels.

**Balanced study:** High study hours help reduce stress only when combined with adequate sleep.

**Combined effect:** Sleep and study habits together predict stress more accurately than considering them individually.

**Other factors:** GPA and total activity hours moderately influence stress, while social and extracurricular hours have minimal effect.

**Short summary:**

“Stress levels are best predicted by a combination of sleep and study habits, highlighting the importance of balance rather than focusing on a single factor.”

**🔹 Additional Insights**

**1.Sleep-Study Interaction:**

“Students with moderate study hours but poor sleep report higher stress than students with high study hours and adequate sleep.”

**2.Activity Effects:**

“Physical activity slightly reduces stress, especially when combined with good sleep habits.”

**3.Social Life:**

“Social and extracurricular hours have minimal impact on stress levels compared to study and sleep.”

**4.GPA Relation:**

“Higher GPA students tend to manage stress better, likely due to better study planning and time management.”

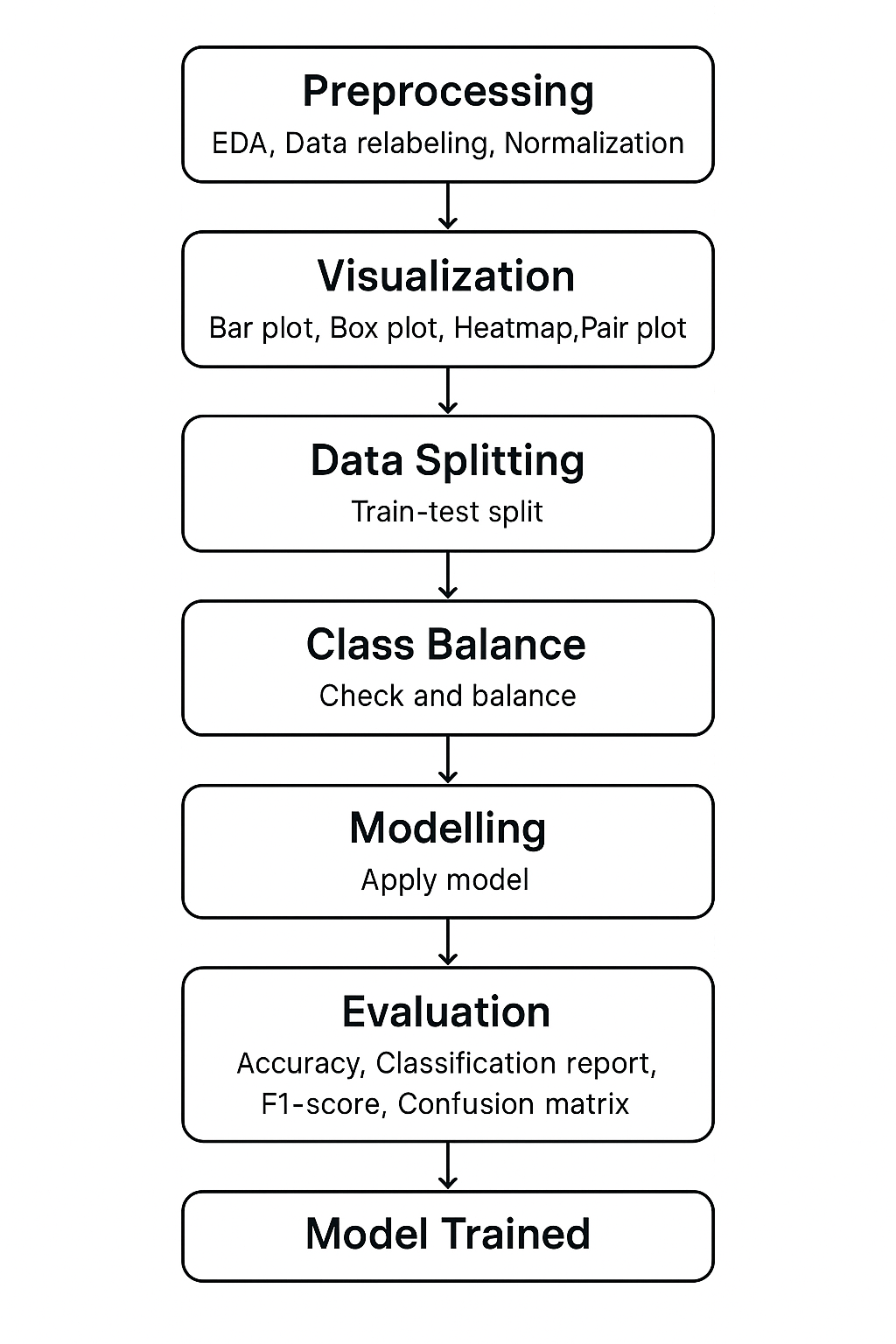
**5.Predictive Patterns:**

“Random Forest shows non-linear relationships; small changes in study or sleep hours can disproportionately affect stress, highlighting the importance of balanced routines.”

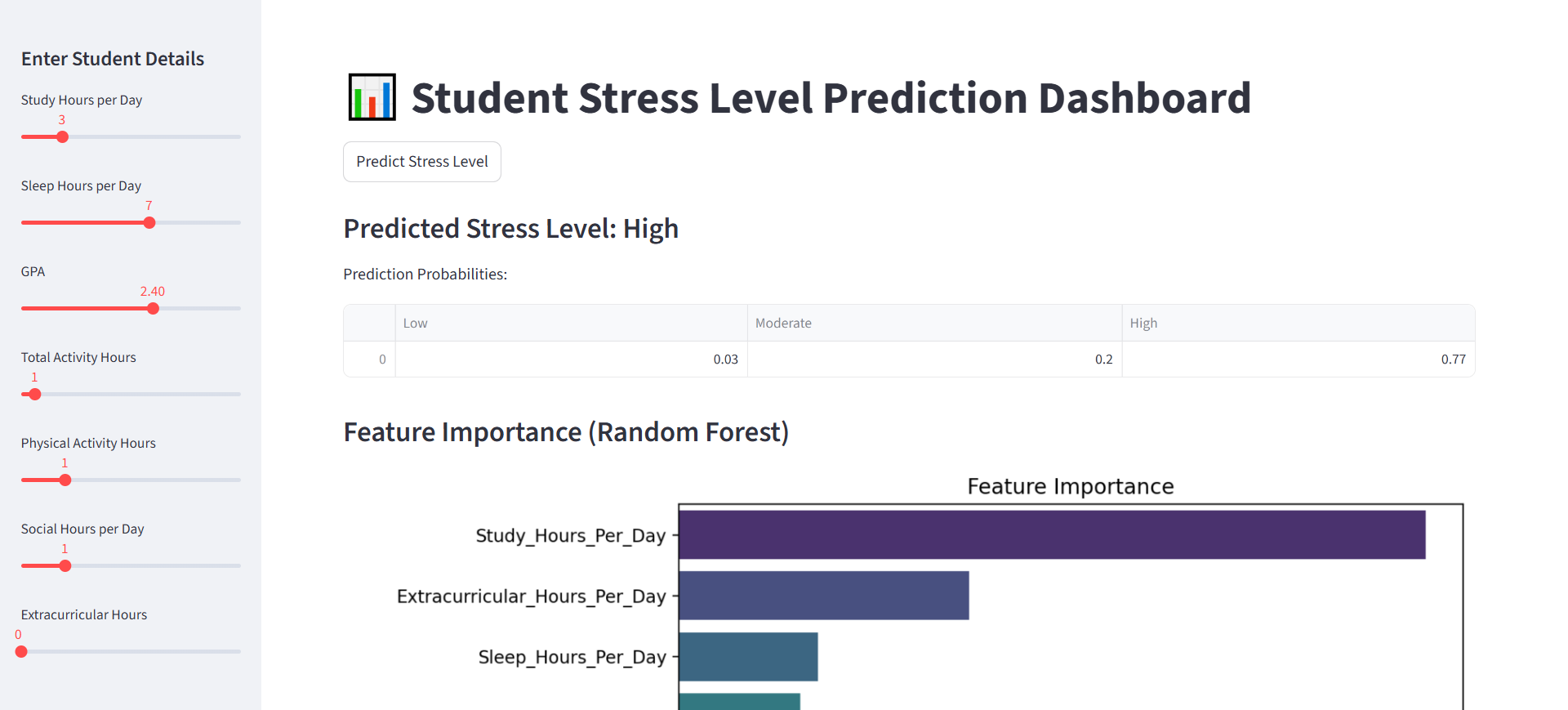
**Actionable Advice for Students:**

* “Prioritize at least 7 hours of sleep.”
* “Avoid excessive study without rest.”
* “Incorporate moderate physical activity into daily schedule.”

**ML Pipeline Diagram:**

****

**Dashboard of Stress Analysis:**

****

**“Interactive Stream-lit dashboard allows predicting stress levels based on student features and visualizes feature importance.”**

**9.Conclusion:**

* Machine learning model successfully predicts student stress levels using study, sleep, GPA, and activity features.
* **Random Forest** chosen as final model due to 100% accuracy and perfect classification.
* **Key features influencing stress:** Study Hours, Study-Sleep Ratio, Sleep Hours.
* **Insights:** Balanced study and adequate sleep are critical; other factors (social/extracurricular hours) have less impact.
* **Interactive dashboard enables real-time predictions and visualization of feature importance**.
* Provides a reliable, actionable tool for monitoring and reducing student stress.