Understanding the Stress Levels Among University Students: A Data-Driven Analysis

A. Abstract

University students often face mental health challenges, with stress being a common problem. This study looks at key factors that affect stress levels, such as sleep habits, device use, and the balance between academics and social life. Data from a survey of 367 students was analyzed using statistical methods like Pearson correlation and ANOVA, as well as machine learning models. The results showed strong links between less sleep ($\rho = -0.63$) and more device use ($\rho = 0.62$) with higher stress levels. However, factors like gender and study level did not have a significant effect. Machine learning models predicted stress levels with moderate accuracy (~55%), showing the need for more detailed data. These findings underline the importance of better sleep and limiting device use to help reduce stress.

B. Introduction

Mental health among university students is a growing concern, with stress being a leading factor affecting academic performance and overall well-being. Despite numerous studies, identifying the behavioral determinants of stress remains critical for targeted interventions. This study aims to analyze survey data to:

- Identify key behavioral factors influencing stress.
- Explore relationships between stress levels, sleep patterns, and device usage.
- Develop machine learning models to predict stress levels based on behavioral data.

Our contributions include leveraging statistical methods and machine learning models to provide actionable insights into student stress management.

C. Methodology

1. Data Collection

A survey was conducted among 367 university students, collecting data on:

- **Demographics**: Gender, age, study level, and UIR student status.
- **Behavioral Factors**: Sleep hours, device time, academic-social balance, activity frequency.
- **Stress Levels**: Frequency, evaluation, and symptoms.

2. Data Preprocessing

- **Cleaning**: Ensured no missing values.
- **Encoding**: Applied label encoding for binary variables and one-hot encoding for nominal variables.
- **Scaling**: Used StandardScaler to standardize numerical features.

3. Statistical Analysis

- **Pearson Correlation**: Assessed relationships between numerical variables.
- **Chi-Square Test**: Examined associations between categorical variables and stress levels.
- **ANOVA**: Tested differences in stress levels across study levels and activity frequencies.

4. Machine Learning

Several models were tested to predict stress levels:

- **Logistic Regression**: A simple model used as a baseline. It achieved an accuracy of 55% when using **sleep hours** and **device time** as features.
- **K-Nearest Neighbors (KNN)**: Provided slightly better performance, with an accuracy of 57% using the same features. Including additional variables like **academic-social balance** improved accuracy to 59%.
- **Support Vector Machines (SVM)**: Achieved 55% accuracy using sleep hours and device time but showed no improvement with additional features.
- **Random Forest Classifier**: The best-performing model, achieving 62% accuracy when using all features (sleep hours, device time, academic-social balance, and stress frequency).

Performance metrics such as precision, recall, and F1-score were used to evaluate the models. Random Forest showed better handling of imbalanced classes, particularly in predicting low and high stress levels.

D.Results

1. Statistical Findings

a. Distribution of Stress Frequency

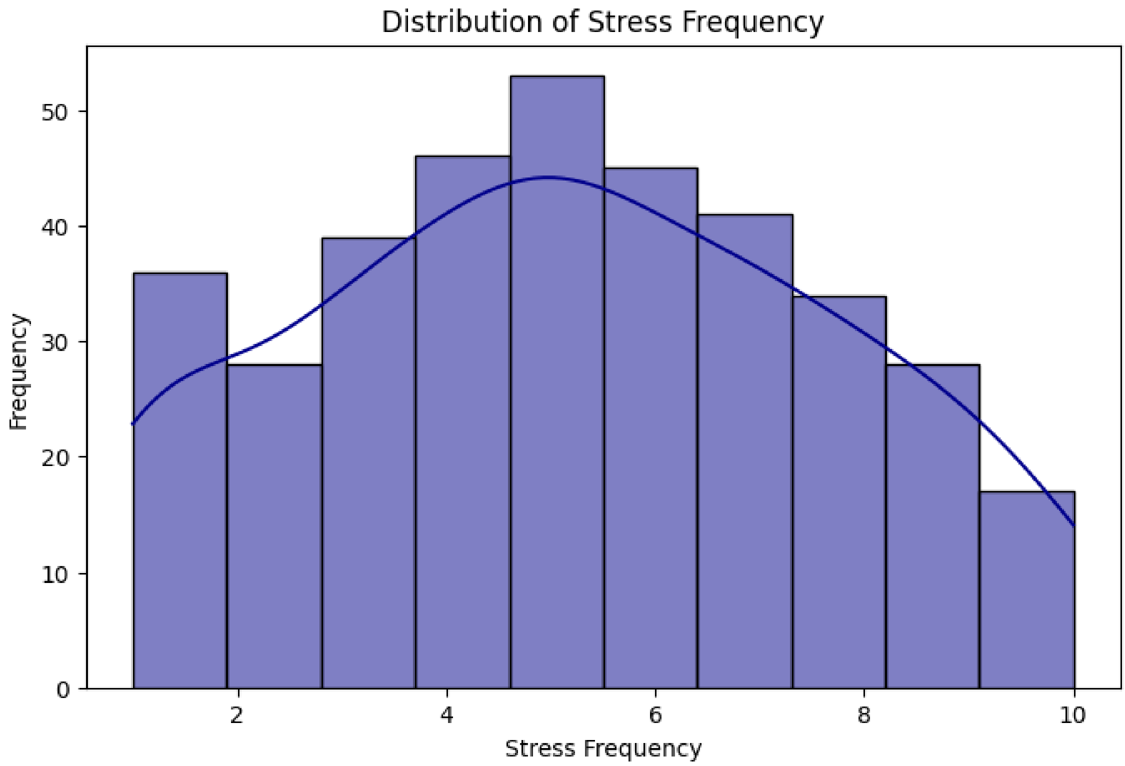

Figure 1: Distribution of stress frequency among university students.

1. The distribution of stress frequency among students is relatively symmetric and approximately bell-shaped, indicating a near-normal distribution.
2. Most students experience moderate stress levels, as indicated by the peak near the middle of the range.

b. Stress Frequency by Gender

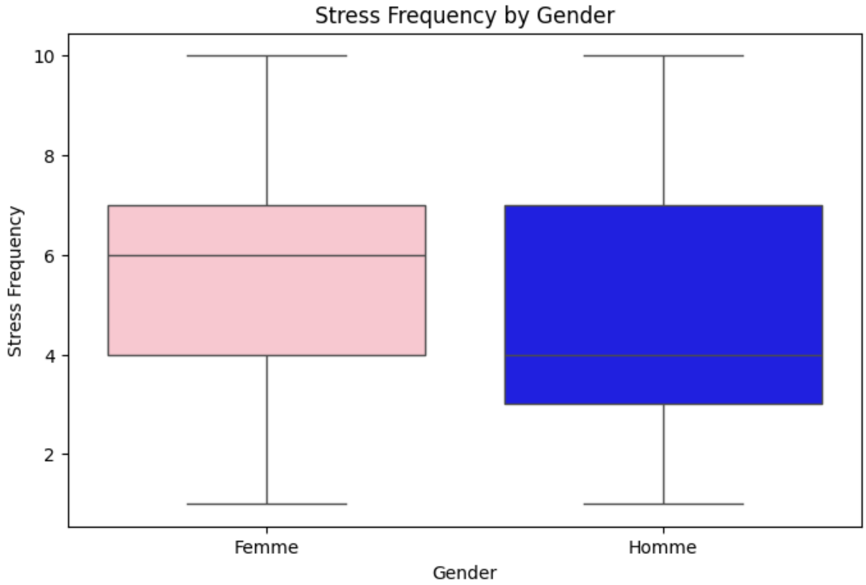

Figure 2: box plot comparing stress frequency across gender groups. This highlights any notable differences between male and female students.

1. females: Tend to report slightly higher stress frequencies (median around 6) with a narrower interquartile range, indicating more consistent stress levels.
2. males: Show lower median stress frequency (around 4) but greater variability in stress levels.
3. Both genders have similar overall ranges, but females show a concentration of higher stress levels compared to males.

c. Stress Levels and Sleep Hours:

A strong negative correlation ($\rho = -0.63$, $p < 0.05$) indicates that less sleep is associated with higher stress levels.

d. Stress Levels and Device Time:

- A strong positive correlation ($\rho = 0.62$, $p < 0.05$) suggests that excessive device usage elevates stress.

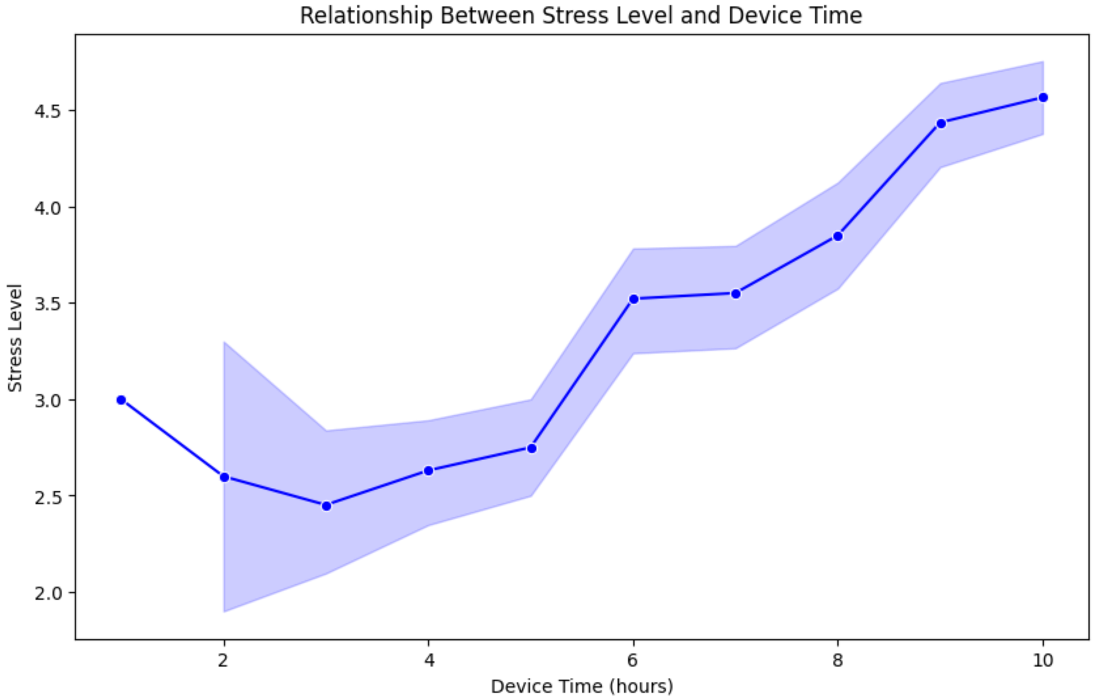

Figure 3: Line plot showing the relationship between device time and stress levels. This graph highlights that longer device usage tends to correspond with higher stress levels.

e. Gender and Stress Levels:

- No significant association was found (Chi-Square $p = 0.74$).

f. Study Levels:

- ANOVA results ($F = 1.13$, $p = 0.34$) showed no significant differences in stress levels across study levels.

g. Correlation Matrix

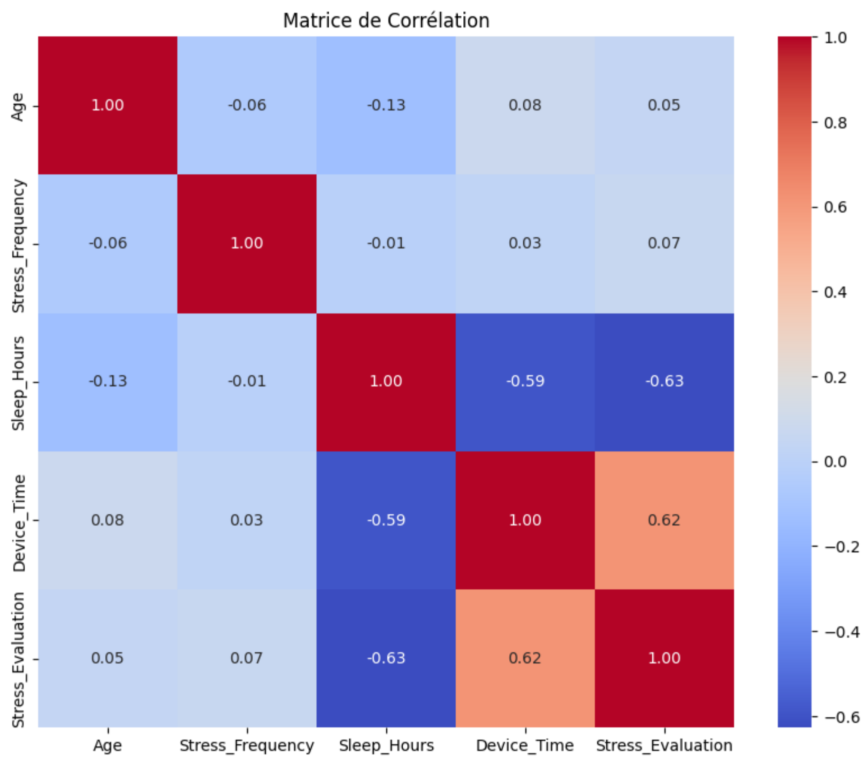

Figure 4: Correlation matrix including age, stress frequency, sleep hours, device time, and stress evaluation. This provides an overview of relationships among key variables.

h. Stress Levels by UIR Student Status

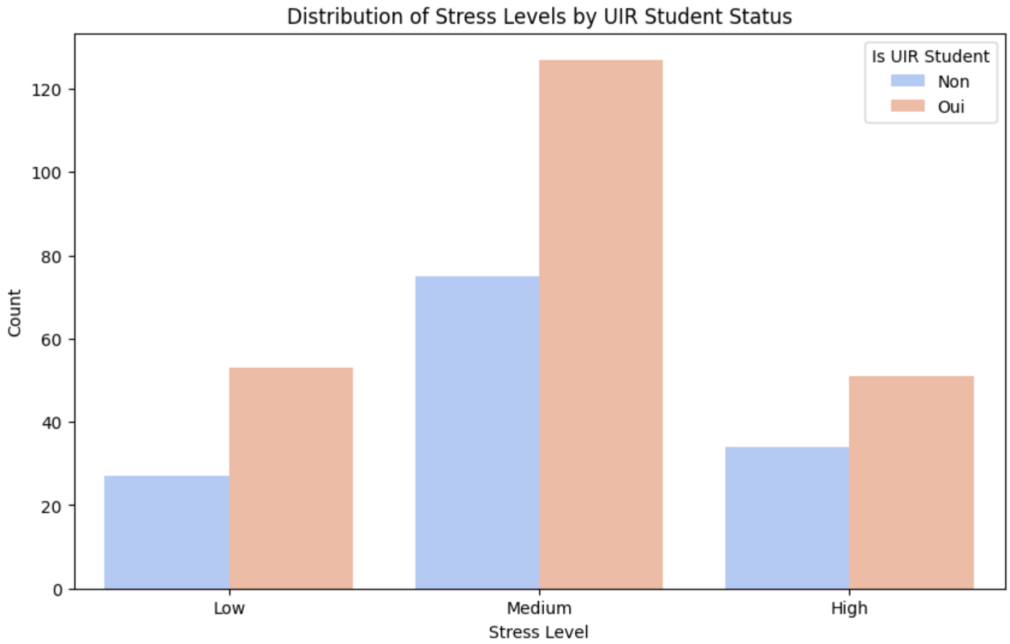

Figure 5: Distribution of stress levels among UIR and non-UIR students, illustrating potential differences between the two groups.

- Both UIR and non-UIR students predominantly fall into the 'Medium' stress level category, with UIR students having a significantly higher count.
- UIR students also show a higher count in the 'Low' stress level category compared to non-UIR students.

2. Machine Learning Performance

Using **Sleep Hours** and **Device Time** as features:

a. Logistic Regression: Accuracy = 55%, F1-score = 49%.
b. KNN: Accuracy = 57%, F1-score = 52%.
c. SVM: Accuracy = 55%, F1-score = 48

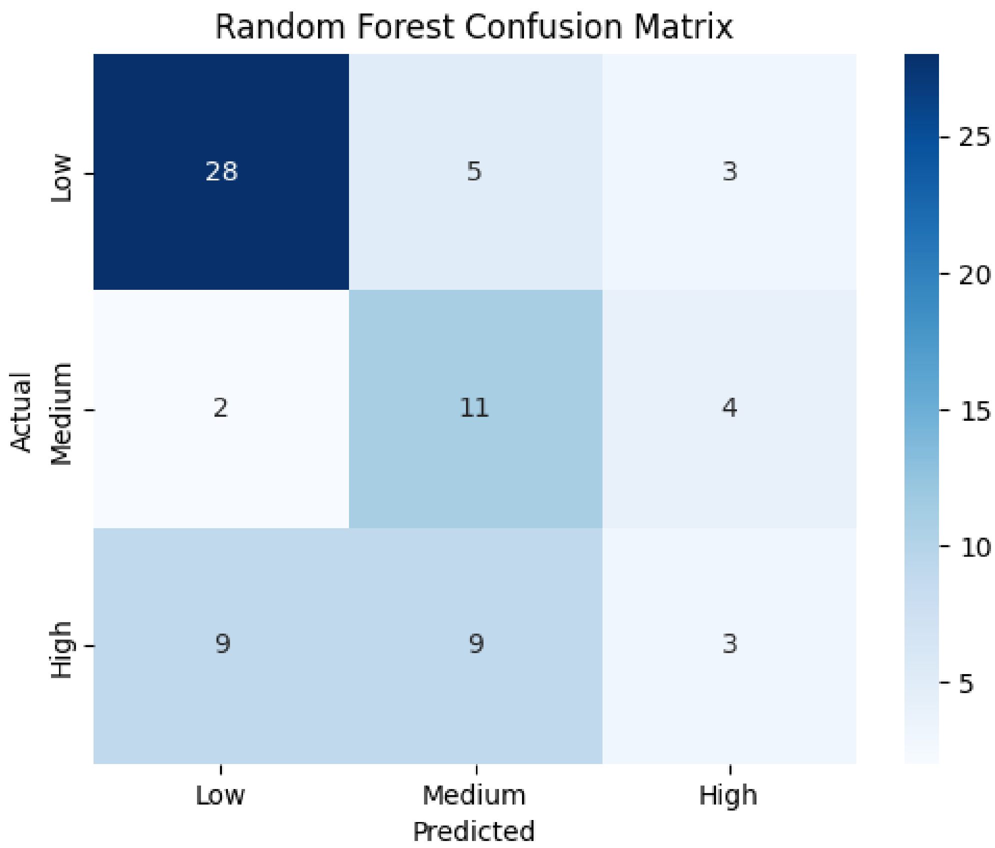

Figure 6: Confusion matrix of the Random Forest model's performance on the test data, illustrating misclassifications across low, medium, and high stress levels.

- Models struggled to predict medium stress levels accurately, highlighting the need for richer data features.

E. Discussion

1. Key Insights

1. Behavioral factors such as sleep, and device usage significantly influence stress levels. Promoting better sleep hygiene and moderating device use could reduce stress.
2. Machine learning models achieved moderate accuracy, suggesting potential for predictive tools but requiring more comprehensive datasets.

2. Implications

- Universities should prioritize student well-being by addressing behavioral stressors.
- Digital detox initiatives and time management workshops may help reduce stress.

3. Limitations

- Self-reported data may introduce bias.
- Sample size is limited to one university.

4. Future Work

- Incorporate additional features such as dietary habits or mental health interventions.
- Explore advanced machine learning models for improved predictions.

F. Conclusion

This study highlights the significant impact of sleep and device time on stress levels among university students. While statistical and machine learning analyses provided valuable insights, future research should expand datasets and features to enhance predictive accuracy. These findings underscore the importance of behavioral interventions in fostering student well-being.

G. References

Scikit-learn Documentation: Machine Learning in Python.
Matplotlib Documentation: Visualization in Python.
Seaborn Documentation: Statistical Data Visualization.