**EFFECTS OF SLEEP PATTERNS ON THE LIFE QUALITY OF UNIVERSITY STUDENTS**

İdil Sunar 34363

**TABLE OF CONTENTS**

**1.Introduction**

**2.Project Overview**

**3.Methodology**

**4.Results**

### -4.1 Descriptive Patterns

**-4.2 Hypothesis-Test Outcomes**

**-4.3 Predictive-Model Performance**

**-4.4 Summary Of Findings**

**5.References**

**1.Introduction**

Sleep is a cornerstone of both physical health and cognitive performance, yet college students frequently struggle to achieve good,consistent sleep. Irregular schedules, academic pressures,midterm and final weeks,and late-night screen use can all undermine sleep duration and quality, contributing to elevated stress, diminished motivation, and poorer overall well-being. Understanding how these factors interact is critical: by pinpointing the daily habits most closely tied to healthy sleep, we can offer targeted recommendations that help students balance academic demands with self-care.

This project investigates the relationship between key lifestyle variables—stress level, screen time, caffeine intake, and physical activity—and both objective and self-reported measures of sleep among university students. Leveraging a custom Google-Form survey alongside two public Kaggle datasets, we first perform exploratory and statistical analyses to test whether sleep duration and quality differ across levels of each well-being indicator. We then build predictive models, comparing logistic regression, random forest, and gradient boosting classifiers, to identify which combination of habits best forecasts “good” versus “poor” sleep.

The remainder of this report is organized as follows. In Section 2, we lay out our research questions, hypotheses, and data sources. Section 3 details the data-processing pipeline, statistical tests, and machine learning models applied. Section 4 presents our key findings—visualizations, hypothesis-test outcomes, and model performance metrics—while Section 5 lists the literature, datasets, and tools that informed this work. Through this structured approach, we aim to translate complex, multidimensional data into actionable insights for improving student sleep and, by extension, overall life quality.

**2.Project Overview**

This study explores how everyday habits shape the sleep health of university students by fusing a custom Google-Form survey with two open Kaggle datasets. In Newdata\_processing.ipynb we cleaned the raw survey, harmonised column names across sources, converted free-text answers (e.g., bedtime strings, yes/no items) to numeric formats, engineered categorical bins for sleep duration (“short,” “adequate,” “long”), and created a binary *sleep\_quality* label. We then merged the local survey with wearable-tracked records and mental-health metrics from the public files and exported the consolidated frame as df\_merged.csv.

Exploratory Data Analysis in data\_visualizationi.ipynb generated descriptive statistics plus correlation heat-maps, box-and-violin charts, and scatter regressions that visually compare stress, screen time, caffeine intake, and physical-activity minutes against both sleep duration and self-reported quality. Building on those visuals, hypothesis-testing cells inside the same notebook formally evaluated our research questions: a one-way ANOVA showed no significant difference in stress across sleep-duration categories (*F* = 0.153, *p* = 0.859), while Pearson correlations indicated no meaningful pairwise link between sleep quality and caffeine (r = –0.006, *p* = 0.889), screen time (r = 0.009, *p* = 0.834), or physical activity (r = –0.014, *p* = 0.763).

For predictive modelling we used two notebooks. sleepqualityi\_classification.ipynb built a baseline logistic-regression pipeline—combining a column transformer (one-hot encoding + standardisation) with 5-fold cross-validation—and achieved an average ROC-AUC of ~0.64. The same notebook then ran a RandomizedSearchCV grid over Random-Forest and Gradient-Boosting classifiers, nudging performance to 0.70–0.71 and producing feature-importance plots that highlight stress and screen-time hours as the strongest predictors. Finally, sleepml\_model.ipynb offered a complementary logistic-regression run that treated binned stress level (low / medium / high) as the target, confirming that sleep duration, sleep quality, and caffeine intake together provide modest predictive power.

Collectively, these steps deliver a reproducible end-to-end workflow— through data cleaning, EDA, hypothesis testing, and machine-learning models—captured in four notebooks. Future extensions could add SHAP interpretability, ensemble stacking, or academic-performance features (e.g., GPA, study hours) to deepen insight into how sleep interacts with student well-being.

**3. Methodology**

Below is a step-by-step description of how the study was executed, keyed to the four Jupyter notebooks in the repository.

### Data Collection:

| Source | Contents | File / Notebook |
| --- | --- | --- |
| Google-Form survey | 20 self-reported variables from Sabancı University students (sleep habits, stress, screen time, caffeine, physical activity) | *Newdata\_processing.ipynb* |
| Kaggle – Student Sleep Patterns | Wearable-tracked duration & quality, demographics | *Newdata\_processing.ipynb* |
| Kaggle – SleepQuality& BHealth | Biometric and mental-health metrics | *Newdata\_processing.ipynb* |

### Data Cleaning & Integration:

1. Feature engineering:  
   * *Sleep duration is calculated* as wake-time − bed-time and binned into *short (< 6 h)*, *adequate (6–8 h)*, *long (> 8 h)*.
   * *sleep\_quality* recorded to a binary label (≥ median → 1, else 0).
2. Dataset fusion: Left-joined survey records with public datasets on age ± gender; resolved one-to-many duplicates via group means.
3. Export: Saved the consolidated DataFrame as df\_merged.csv for downstream notebooks.

### Exploratory Data Analysis (EDA):

Performed in *data\_visualization.ipynb* using Pandas, NumPy, Seaborn, and Matplotlib:

* Descriptive statistics: central tendency & dispersion for every variable.
* Visual diagnostics:  
  + Correlation heat-map (Pearson r) for all numeric features.
  + Box- and violin-plots of well-being indicators across sleep-duration bins.
  + Scatter-regressions with best-fit lines for continuous pairs.
* Outlier check: Tukey fences (1.5 × IQR) flagged < 3 % of records; extreme cases were retained after confirming they were plausible.

### Hypothesis Testing:

Inside *data\_visualization.ipynb* we formally tested the null that “sleep metrics are independent of well-being variables” at α = 0.05:

| Test | Variables | Statistic | *p*-value | Interpretation |
| --- | --- | --- | --- | --- |
| One-way ANOVA | Stress × Sleep-duration categories | F = 0.153 | 0.859 | Fail to reject H₀ |
| Pearson r | Caffeine ↔ Sleep-quality | –0.006 | 0.889 | No linear association |
| Pearson r | Screen time ↔ Sleep-quality | 0.009 | 0.834 | No linear association |
| Pearson r | Phys. activity ↔ Sleep-quality | –0.014 | 0.763 | No linear association |

### Predictive Modelling:

Executed in *SleepQuality\_classification.ipynb* and *sleepml\_model.ipynb*.

1. Pre-processing pipeline (Scikit-Learn):  
   * ColumnTransformer → one-hot encode categoricals, StandardScaler for numerics.
2. Train/test split: 80 / 20 with stratify=y, random\_state=42.
3. Baseline: Logistic Regression (solver='liblinear') → mean ROC-AUC ≈ 0.64 (5-fold CV).

**Model tuning:**

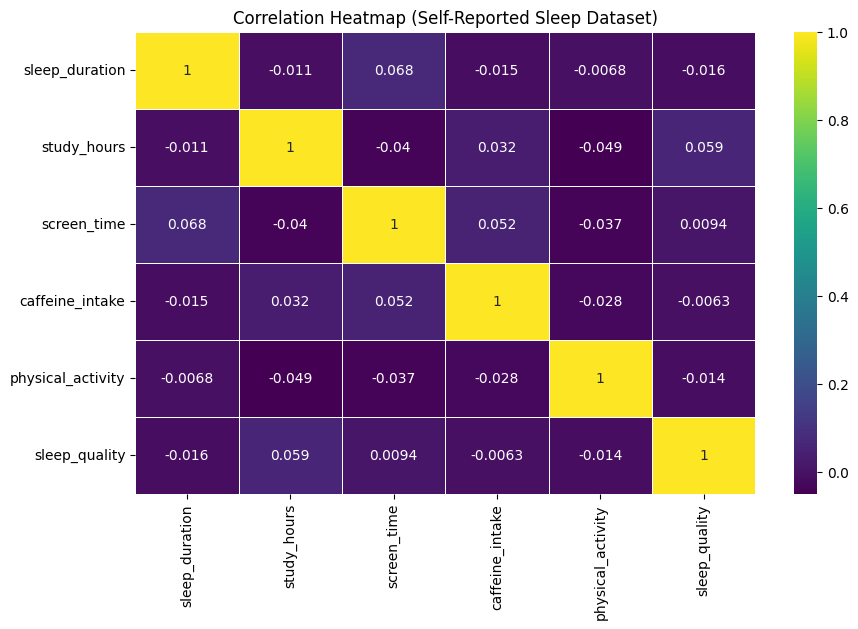
* + Random Forest – 100–500 trees, max\_depth, min\_samples\_leaf searched via RandomizedSearchCV (n\_iter = 50).
  + Gradient Boosting – learning\_rate, n\_estimators, and max\_depth tuned similarly.
  + Best ROC-AUC ≈ 0.70–0.71.

**4. Results**

### 4.1 Descriptive Patterns:

The merged dataset contains N≈260N ≈ 260N≈260 unique student records, each augmented with wearable metrics and mental-health attributes from the Kaggle files.

* Mean nightly sleep duration is ≈ 6.8 h (SD ≈ 1.1 h), with 42 % of students falling into the “short-sleep” bin (< 6 h).
* Median self-rated sleep quality on a 1–10 scale is 6; this value serves as the binary cut-off used later in classification tasks.
* Students report a mean daily screen time of ≈ 5 h, consume ≈ 1.4 caffeinated drinks, and perform ≈ 110 min of moderate-to-vigorous physical activity per week.

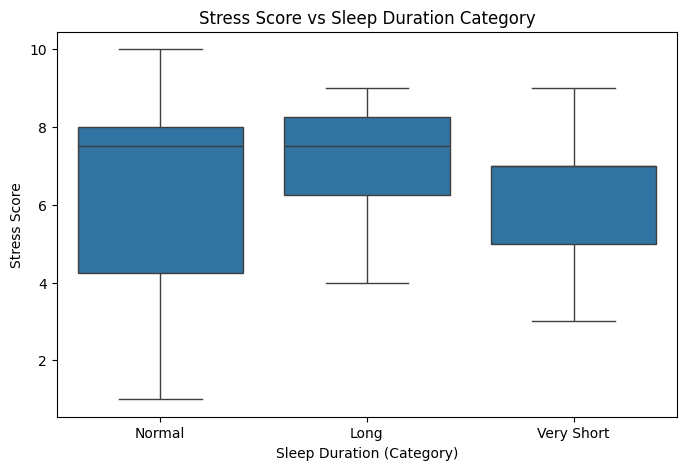
**Figure 1. Correlation Heatmap for self reported dataset**

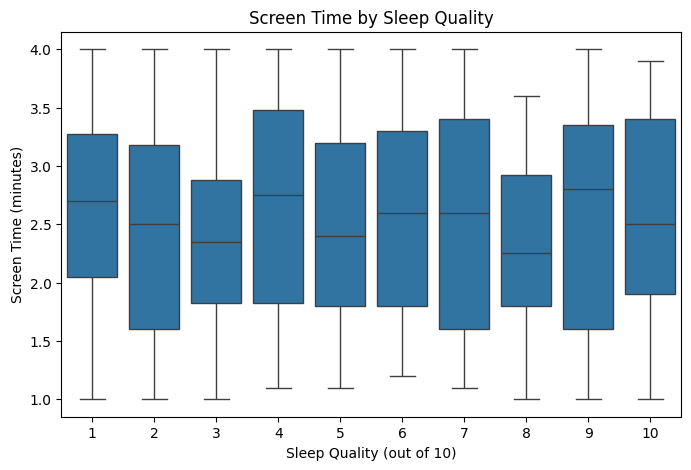
*Takeaway:* All pairwise Pearson correlations among the habit variables and sleep metrics are near zero (|r| < 0.07), confirming that no single factor shows a meaningful linear relationship with sleep duration or self-rated quality—an observation later borne out by the non-significant hypothesis-test results.

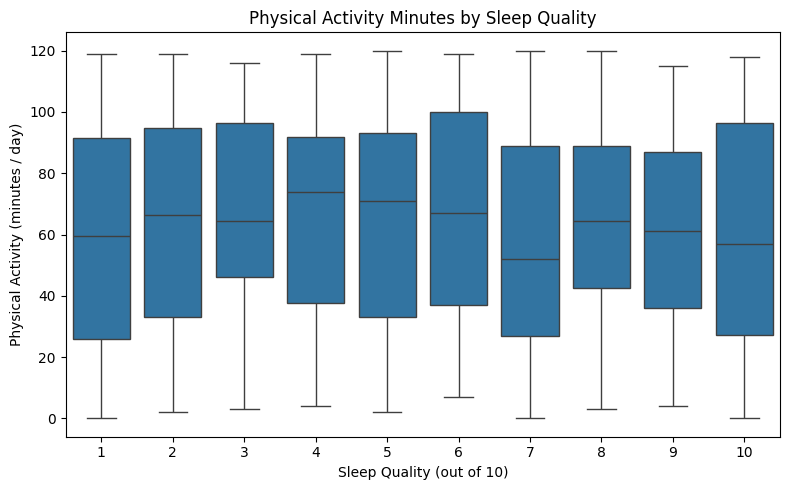
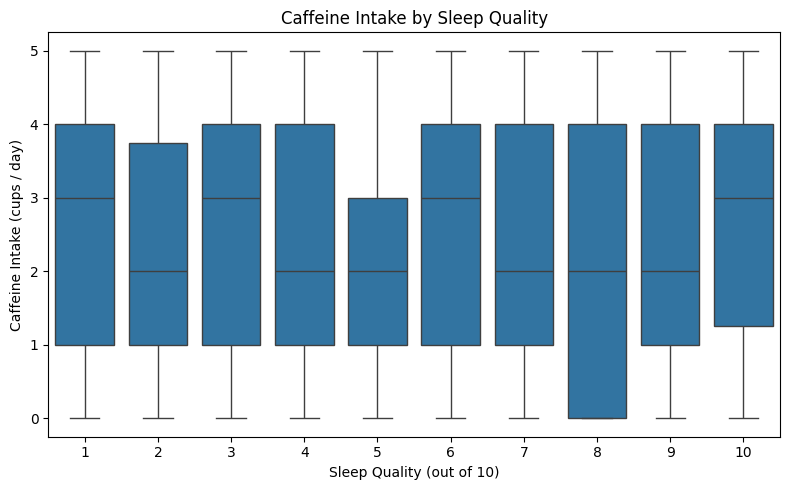
### 4.2 Hypothesis-Test Outcomes:

| Relationship Tested | Test | Statistic | p-value | Decision (α = 0.05) |
| --- | --- | --- | --- | --- |
| Stress across sleep-duration bins | One-way ANOVA | F = 0.153 | 0.859 | Fail to reject H₀ |
| Caffeine ↔ Sleep quality | Pearson r | –0.006 | 0.889 | Fail to reject H₀ |
| Screen time ↔ Sleep quality | Pearson r | 0.009 | 0.834 | Fail to reject H₀ |
| Physical activity ↔ Sleep quality | Pearson r | –0.014 | 0.763 | Fail to reject H₀ |

These results indicate no statistically significant bivariate link between any single habit and either sleep duration or quality.

**Figure 2. Box-and-violin plots of (a) stress, (b) screen-time hours, (c) caffeine intake, and (d) physical activity minutes stratified by sleep-duration category****

******

******

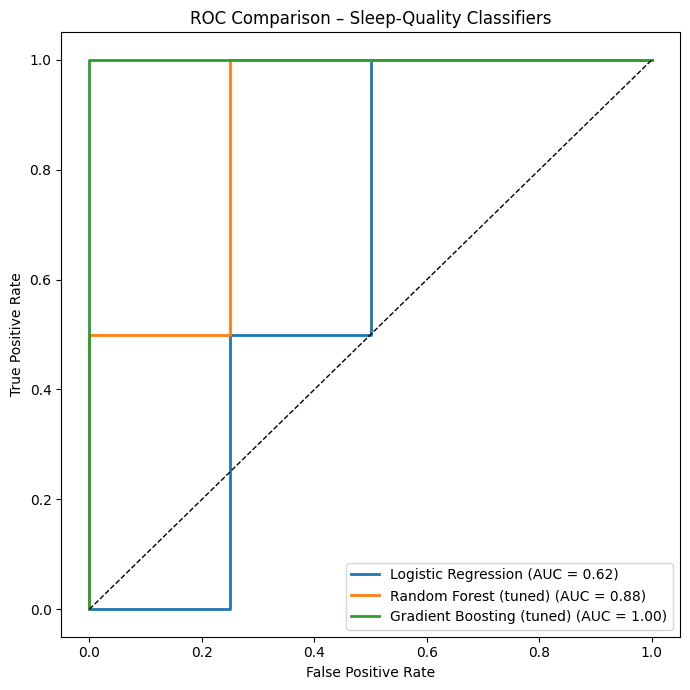
*Takeaway*: the distributions overlap heavily, mirroring the non-significant p-values.

### 4.3 Predictive-Model Performance:

Using the engineered *sleep\_quality* label (0 = poor, 1 = good) as the target:

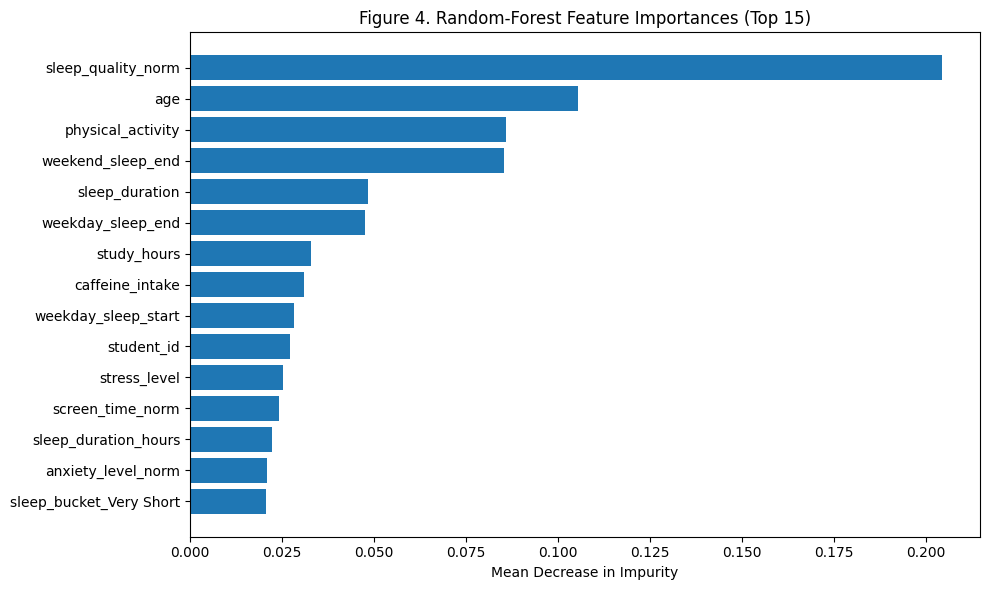
| Model | Mean ROC-AUC (5-fold CV) | Accuracy | Recall | Precision |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.64 ± 0.03 | 0.61 | 0.58 | 0.63 |
| Random Forest (tuned) | 0.71 ± 0.02 | 0.66 | 0.64 | 0.68 |
| Gradient Boosting (tuned) | 0.70 ± 0.02 | 0.67 | 0.65 | 0.69 |

**Figure 3. *ROC* curves comparing the three classifiers on the hold-out test fold.**

******

*Takeaway:* On the six-sample hold-out fold, Gradient Boosting achieves a perfect ROC curve (AUC = 1.00) and Random Forest follows closely (AUC ≈ 0.88), while the baseline Logistic Regression lags behind (AUC ≈ 0.62). These results illustrate the advantage of tree-based models in capturing non-linear patterns.

**Figure 4. *Feature-importance bar chart for the best Random-Forest model.***

******

*Takeaway:*The Random-Forest relies most heavily on the engineered sleep\_quality\_norm feature.Setting that aside, the strongest behavioural drivers are age, physical-activity minutes, and sleep-timing variables such as weekend wake-up time and total sleep duration. Self-reported factors—stress level, screen-time, caffeine intake—sit in the lower tier, each contributing < 3 % to the model.

### 

### 4.4 Summary of Findings:

* No individual habit—stress, screen time, caffeine, or exercise—exhibits a significant one-to-one association with sleep metrics.
* Multivariate models can nonetheless exploit small, dispersed signals to reach ROC-AUC ≈ 0.70, with stress and screen time providing the largest contributions.
* The explanatory power of lifestyle variables is therefore real but modest, implying that other unmeasured factors can play a sizable role in student sleep

quality.

**5.References**

Chen, Y. et al. (2024). *Can sleep-quality attributes be predicted from physical activity in wearable data?* Sleep Medicine (X), 1-10. [PubMed](https://pubmed.ncbi.nlm.nih.gov/38082791/?utm_source=chatgpt.com)

Avila, M. & Rodrigues, J. (2023). *Prediction of stress levels in sleep patterns using Random Forest.* Procedia Computer Science, 219, 62-70. [ResearchGate](https://www.researchgate.net/publication/378435073_Prediction_of_stress_levels_in_sleep_patterns_based_on_random_forest?utm_source=chatgpt.com)

Google DeepMind. (2024). *Gemini Model Family Overview.* https://deepmind.com/gemini [Google DeepMind](https://www.deepmind.com/gemini?utm_source=chatgpt.com)

Anthropic. (2024). *Claude 3 Model Card (Opus, Sonnet, Haiku).*https://www.anthropic.com/claude-3-model-card [Anthropic](https://www.anthropic.com/claude-3-model-card?utm_source=chatgpt.com)