**DSA210 Final Project Report**

**Name:** Ertuğrul Soydal  
**Student ID:** 32175  
**Course:** DSA210  
**Submission Date:** May 30, 2025

📌 Project Title:

**Personal Data Analysis: Study Behavior Under Exam Pressure Based on Sleep, Caffeine, and Meal Patterns**

This project presents an analysis of a self-collected dataset to explore how exam pressure affects daily routines such as sleep duration, caffeine consumption, meal frequency, and study time. The analysis is conducted using Python, including data preprocessing, statistical hypothesis testing, and machine learning models.

**DSA210 Term Project Description**

**Project Aim**

As a university student, after lots of exam period, I have realized some significant changes in my life when exam dates are coming.Therefore; I want to look deeper to these changes This project aims to analyze the correlation between sleep patterns, eating habits,study hour and exam dates. By collecting my personal data, I will try to answer questions below:

* **How sleep duration affects study performance?**
* **When getting closer to the exam date, how sleeping hour and eating habbits vary?**
* **Is there any significant correlation between sleeping hour and studying hour?**
* **Is there any significant correlation between eating habbit and studying hour?**
* **How fast studying hour and coffee intake change while reaching to the exam date?**
* **Which exam date made the most significant change in terms of habbits and studying hour?**

**Data Description**

* **Date:** Date of The Given Day
* **Wake-up Time:** Waking up time in the given day
* **Sleeping Time:** Sleeping time in the given day
* **Sleep Duration:** Longness of the sleep (minute) in the given day
* **Number of Meal:** Eaten meal number in the given day
* **Coffee Intake:** Total drinking coffee in the given day (Each coffee is Starbucks Latte so number multiplied with 75 mg Caffeine) (This 75 mg information is taken from official Starbucks website)
* **Study Time:** How long studied (minute) in the given day
* **Lecture Name and Exam Date:** Lecture name(s) and closest exam date(s)

**The Key Idea**

The idea of the project is tracing the significant changes while being close to exam and representing them in terms of data and statistics.After using these statistics it will be clear to see correlations and significant changes.Then different ML techniques will be applied to the given dataset to observe which one will predict studying hour.

**DATA VISUALIZATON AND HYPOTHESIS TESTING**

**Question 1: Does sleep duration significantly correlate with study time?**

**H₀**: There is no significant correlation between sleep duration and study time.  
**H₁**: There is a significant correlation between sleep duration and study time.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Question 2: Does number of exams in 7 days affect sleep duration?**

**H₀**: There is no significant difference in average sleep duration between students with 0, 1, or 2+ upcoming exams.  
**H₁**: At least one group differs significantly in average sleep duration.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Question 3: Does the number of upcoming exams affect eating habits?**

**H₀**: There is no significant difference in the number of meals among students with 0, 1, or 2+ upcoming exams.  
**H₁**: At least one group differs significantly in the number of meals.

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Question 4: Does caffeine intake significantly change depending on number of upcoming exams?**

**H₀**: There is no significant correlation between number of upcoming exams and caffeine intake.  
**H₁**: There is a significant correlation between number of upcoming exams and caffeine intake.

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Question 5: How much is the change in the studying hour depending on the active exams (within 7 days)?**

**H₀**: There is no significant difference in average study time across groups.  
**H₁**: At least one group differs significantly.

metin, diyagram, ekran görüntüsü, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Question 6: Which exam affects studying hours most (based on average daily study hours)?**

This question is answered by comparing the **average daily study time** associated with each exam (excluding exam days).

metin, ekran görüntüsü, diyagram, paralel içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**ML METHODS**

**Random Forest Generator**

**Supervised Learning:** Regression Model

**Algorithm:**Random Forest Generator

**Aim:**Predict Study Hour

📊 **Random Forest Hyperparameter Tuning Summary**

**n\_estimators = 100, max\_depth = None, max\_features = None, min\_samples\_split = 2**  
→ R² Score: 0.659  
→ RMSE: 83.36 minutes  
✅ This configuration gave the best overall performance. (Best fitting configuration for Random Forest Generator)

**n\_estimators = 100, max\_depth = 10, max\_features = sqrt, min\_samples\_split = 2**  
→ R² Score: 0.635  
→ RMSE: 86.31 minutes  
Slight performance drop compared to the default configuration.

**n\_estimators = 200, max\_depth = 20, max\_features = sqrt, min\_samples\_split = 5**  
→ R² Score: 0.567  
→ RMSE: 94.00 minutes  
Deeper trees and increased estimators did not improve performance.

**n\_estimators = 300, max\_depth = 30, max\_features = log2, min\_samples\_split = 2**  
→ R² Score: 0.605  
→ RMSE: 89.76 minutes  
Slight improvement over some configs, but still worse than the default.

**n\_estimators = 500, max\_depth = 20, max\_features = None, min\_samples\_split = 2**  
→ R² Score: 0.649  
→ RMSE: 84.57 minutes  
Close to the best; higher number of trees performed well.

**n\_estimators = 500, max\_depth = None, max\_features = log2, min\_samples\_split = 10**  
→ R² Score: 0.516  
→ RMSE: 99.31 minutes  
Worst result; likely underfitting due to high split threshold and reduced feature set.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**XGBoost**

**Supervised Learning:** Regression Model

**Algorithm:**XGBoost

**Aim:**Predict Study Hour

**📊 XGBoost Hyperparameter Tuning Summary**

**n\_estimators = 100, max\_depth = 6, learning\_rate = 0.2, subsample = 1.0**  
→ R² Score: 0.712  
→ RMSE: 76.68 minutes  
✅ Best performing configuration. (Best fitting configuration for XGBoost)

**n\_estimators = 250, max\_depth = 8, learning\_rate = 0.02, subsample = 0.85**  
→ R² Score: 0.682  
→ RMSE: 80.57 minutes  
Strong performance with deeper trees and slower learning.

**n\_estimators = 300, max\_depth = 5, learning\_rate = 0.03, subsample = 0.8**  
→ R² Score: 0.632  
→ RMSE: 86.67 minutes  
Moderately good, stable configuration.

**n\_estimators = 200, max\_depth = 4, learning\_rate = 0.05, subsample = 0.9**  
→ R² Score: 0.580  
→ RMSE: 92.52 minutes  
Lower depth and slower learning yield moderate results.

**n\_estimators = 200, max\_depth = 6, learning\_rate = 0.1, subsample = 0.7**  
→ R² Score: 0.535  
→ RMSE: 97.39 minutes  
Weaker results likely due to subsampling too aggressively.

**n\_estimators = 100, max\_depth = 3, learning\_rate = 0.1, subsample = 1.0**  
→ R² Score: 0.441  
→ RMSE: 106.78 minutes  
Shallow trees underfit the data.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**K-Nearest Neighbors (KNN)**

**Supervised Learning:** Regression Model

**Algorithm:** K-Nearest Neighbors (KNN) Regressor

**Aim:** Predict Study Hour

**📊 K-Nearest Neighbors (KNN) Hyperparameter Tuning Summary**

**n\_neighbors = 3**  
→ R² Score: 0.557  
→ RMSE: 95.09 minutes  
✅ Best performing value for K. (Best fitting configuration for KNN)

**n\_neighbors = 5**  
→ R² Score: 0.525  
→ RMSE: 98.39 minutes  
Slightly weaker than k=3.

**n\_neighbors = 15**  
→ R² Score: 0.442  
→ RMSE: 106.69 minutes  
Larger k leads to underfitting.

**n\_neighbors = 7**  
→ R² Score: 0.409  
→ RMSE: 109.80 minutes  
Underfits the data.

**n\_neighbors = 10**  
→ R² Score: 0.403  
→ RMSE: 110.31 minutes  
Same underfitting trend continues.

**n\_neighbors = 1**  
→ R² Score: 0.022  
→ RMSE: 141.24 minutes  
Overfits badly to local noise.

metin, çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**COMPARISON FOR ML METHODS**

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**📊 Comparison of Regression Models (Best Results)**

🔷 XGBoost Regressor

* **R² Score**: 0.712 → Best explained variance
* **RMSE**: 76.68 minutes → Lowest prediction error ✅

🟩 Random Forest Regressor

* **R² Score**: 0.659 → Strong explanation of variance
* **RMSE**: 83.36 minutes → Slightly higher error

🟦 KNN Regressor (k = 3)

* **R² Score**: 0.557 → Less variance explained
* **RMSE**: 95.09 minutes → Highest prediction error ⚠️

📘 **Interpretation Notes**

* **R² Score**: Indicates how much of the variation in study time is explained by the model.  
  → Higher is better.
* **RMSE**: Average difference between predicted and actual study time.  
  → Lower is better.

**📌 Final Summary**

* ✅ **XGBoost** showed the best overall performance in both accuracy and variance explanation.
* ✅ **Random Forest** also performed consistently and is a solid alternative.
* ⚠️ **KNN**, while simple, did not perform as well on this dataset due to its non-parametric nature.

**What Can Be Done In The Future?**

To enhance the accuracy of machine learning estimations, future work will focus on expanding the dataset. Collecting data over a one-year period will allow for a more comprehensive analysis of annual study habits, enabling the identification of long-term behavioral patterns when Ertuğrul Soydal (data collector) is in undergraduate years. With a larger dataset, more robust and precise estimations of study hours can be achieved, thereby improving the reliability and generalizability of the predictive models.

**Final Notes About The Final Report**

* The data is collected by Ertuğrul Soydal.
* The designs and test parameters are decided by Ertuğrul Soydal.
* Using appropriate test methods and interpreting the results are also done by Ertuğrul Soydal.
* Appropriate ML methods are decided and applied by Ertuğrul Soydal.
* In the visualization and hypothesis testing part, ChatGPT assistance was used to make the visualizations clearer and the test results more understandable.
* This explanation is included to avoid any potential academic dishonesty.