Coffee Consumption, Sleep Patterns, and Energy Levels

DSA210 – Data Science Term Project

1. Motivation

Daily coffee consumption is a common habit among university students, especially during academically demanding periods. While coffee is often used to increase alertness and energy, its interaction with sleep duration and daily routines may produce varying effects on perceived energy levels.

The motivation of this project is to **analyze how coffee consumption and sleep duration jointly influence daily energy levels**, using personal longitudinal data enriched with external weather information. By studying this relationship, the project aims to provide insights into behavioral and lifestyle patterns that affect daily performance.

2. Data Sources and Data Collection

2.1 Personal Dataset

The primary dataset was **self-collected** on a daily basis over multiple weeks. The following variables were recorded:

- `Date`: Calendar date of observation
- `CoffeeAmount_ml`: Total coffee consumption per day (in milliliters)
- `Bedtime`: Time of going to sleep
- `WakeupTime`: Time of waking up
- `EnergyLevel`: Self-reported daily energy level (ordinal scale)

Sleep duration was computed from bedtime and wake-up time, properly handling overnight sleep cases.

2.2 External Data Enrichment

To enrich the personal dataset, **daily precipitation data** was collected using the **Visual Crossing Weather API** for Istanbul, Turkey.
From this data, a binary feature was created:

- `Weather_Precipitation`:
  - 1 → Rainy day
  - 0 → Non-rainy day

Additionally, a behavioral feature was derived:

- `Weekday`:
  - 1 → Weekday
  - 0 → Weekend

3. Data Preparation and Feature Engineering

The following preprocessing steps were applied:

- Conversion of date columns to datetime format
- Parsing of time values and computation of sleep duration in hours
- Handling overnight sleep intervals (e.g., 23:00–07:00)
- Handling missing values using median imputation (for ML models)
- Feature scaling where required (StandardScaler)

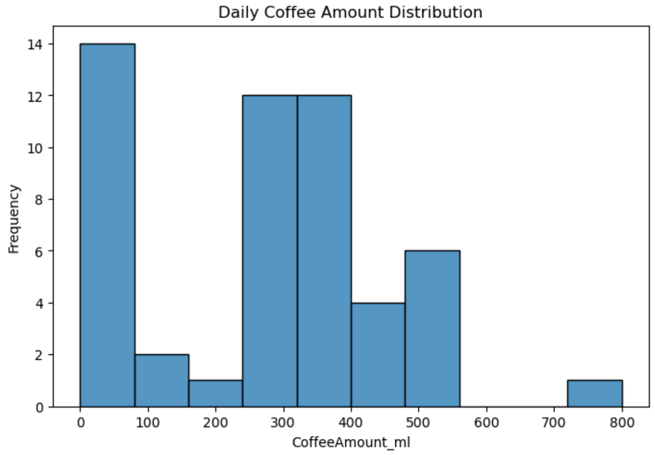
The final dataset included both **behavioral** and **contextual** features.

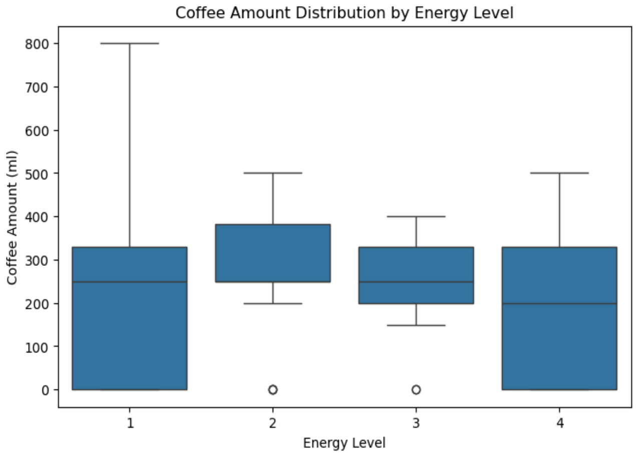
4. Exploratory Data Analysis (EDA)

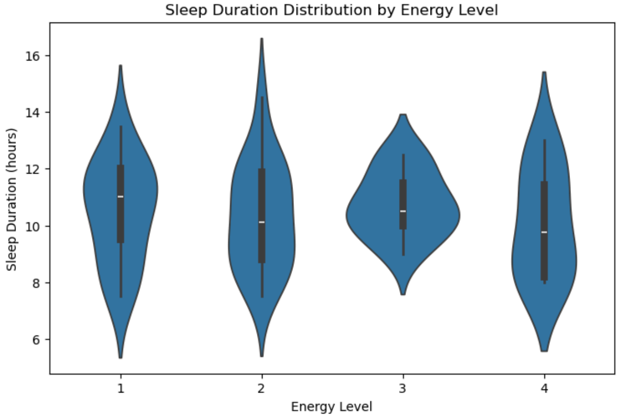
Several visualization techniques were used to explore the data:

4.1 Distribution Analysis

- Histograms for coffee consumption, sleep duration, and energy levels
- Boxplots and violin plots to compare distributions across energy levels

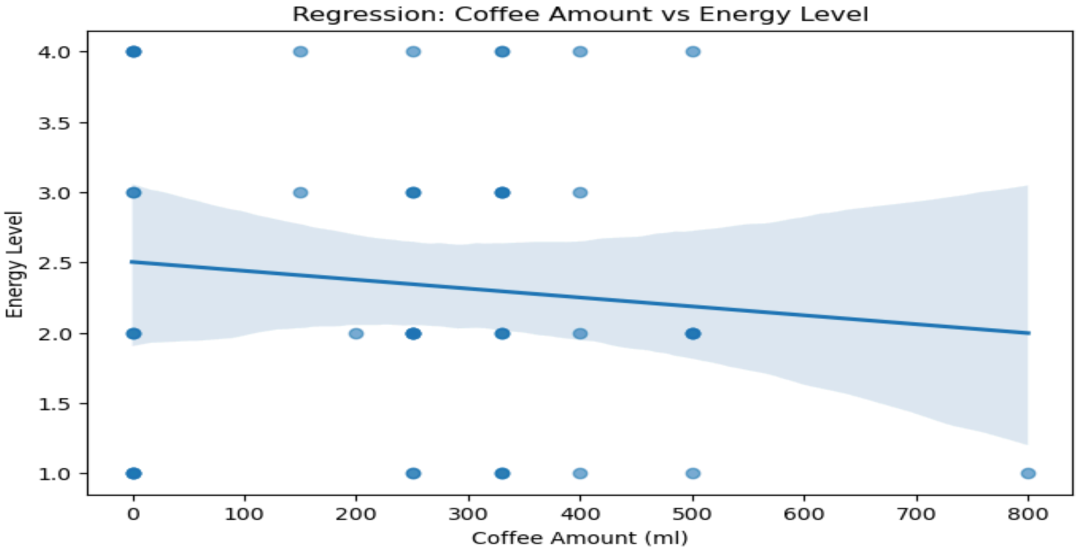




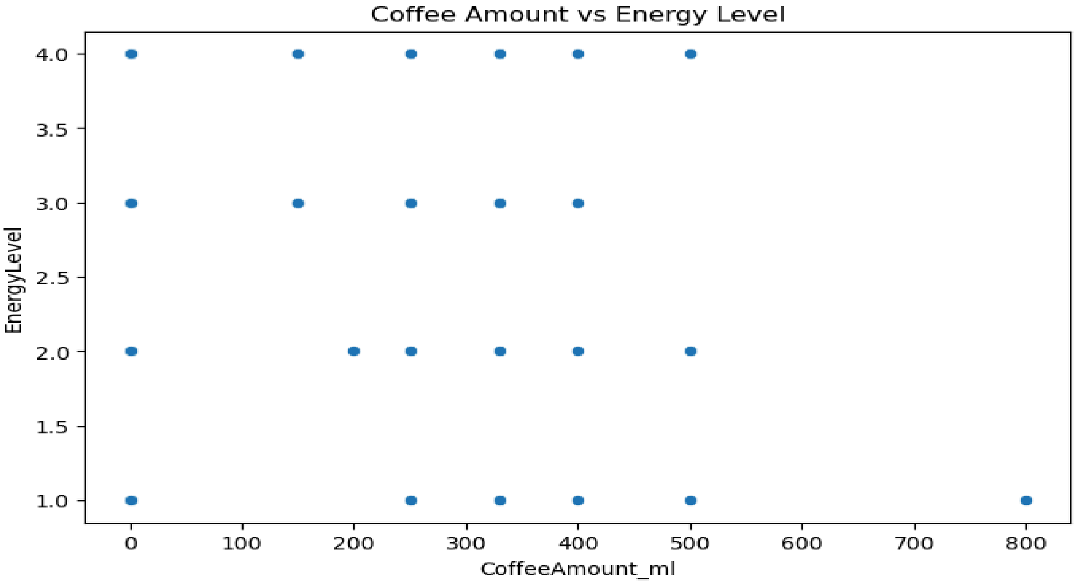


4.2 Relationship Analysis

- Scatter plots between:
  - Coffee amount and energy level
  - Sleep duration and energy level
- Regression plots to visualize overall trends

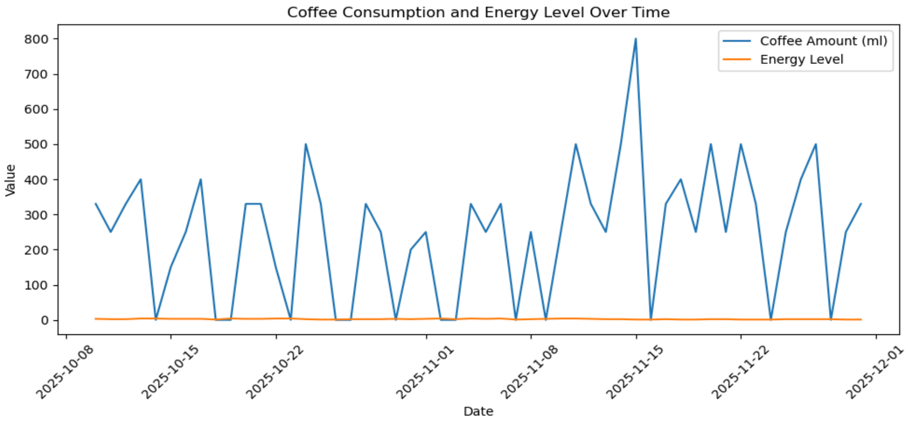

Regression: Coffee Amount vs Energy Level
Energy Level
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
100
200
300
400
500
600
700
800
Coffee Amount (ml)

○


Coffee Amount vs Energy Level
EnergyLevel
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
100
200
300
400
500
600
700
800
CoffeeAmount_ml

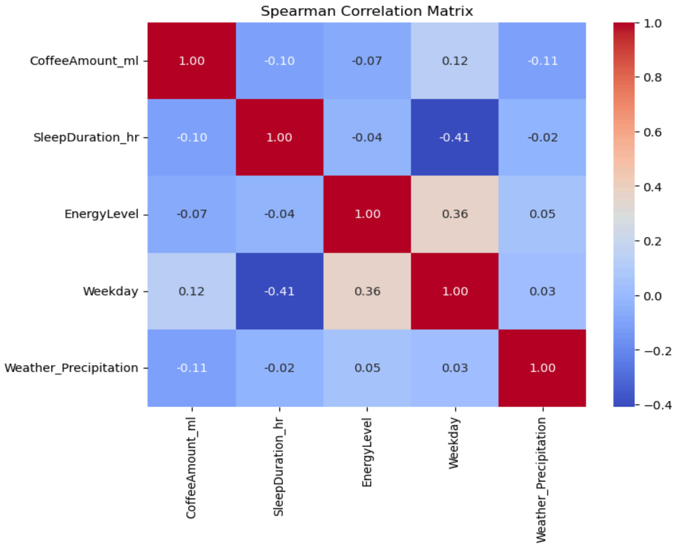
4.3 Temporal Analysis

- Line plots showing coffee consumption and energy level over time



4.4 Correlation Analysis

- Spearman correlation matrix including:
    - Coffee amount
    - Sleep duration
    - Energy level
    - Weekday
    - Weather precipitation



These analyses provided insight into both monotonic relationships and temporal patterns.

5. Statistical Analysis and Hypothesis Testing

To statistically examine the relationship between coffee consumption and energy level, the following tests were applied:

- **Pearson Correlation Test**
- **Spearman Rank Correlation Test**

Null Hypothesis (H_0)

Coffee consumption has no statistically significant relationship with daily energy level.

Alternative Hypothesis (H_1)

Coffee consumption has a statistically significant relationship with daily energy level.

The test results were evaluated using a significance level of $\alpha = 0.05$.
Both correlation coefficients and p-values were reported, and the null hypothesis was either rejected or not rejected accordingly.

6. Machine Learning Methods

To satisfy the machine learning requirement of the project, **supervised classification models** were applied.

6.1 Problem Definition

- **Target variable**: energy level
- **Task**: Multi-class classification

6.2 Feature Set

- CoffeeAmount_ml
- SleepDuration_hr
- Weekday
- Weather_Precipitation

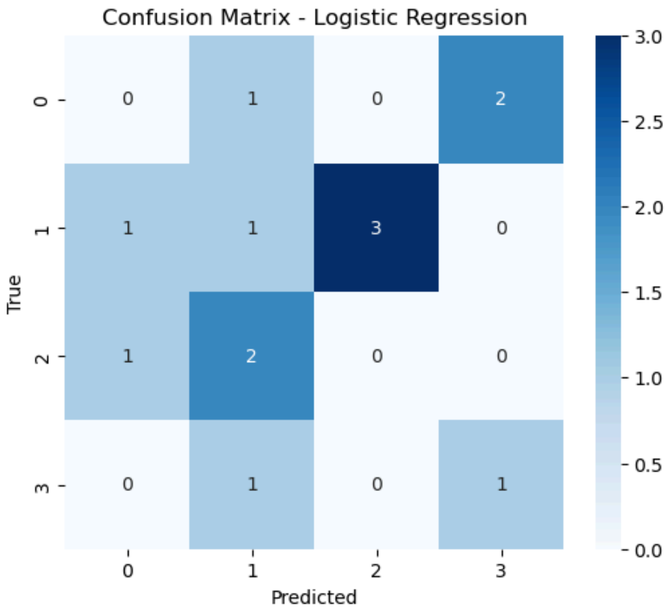
6.3 Models Used

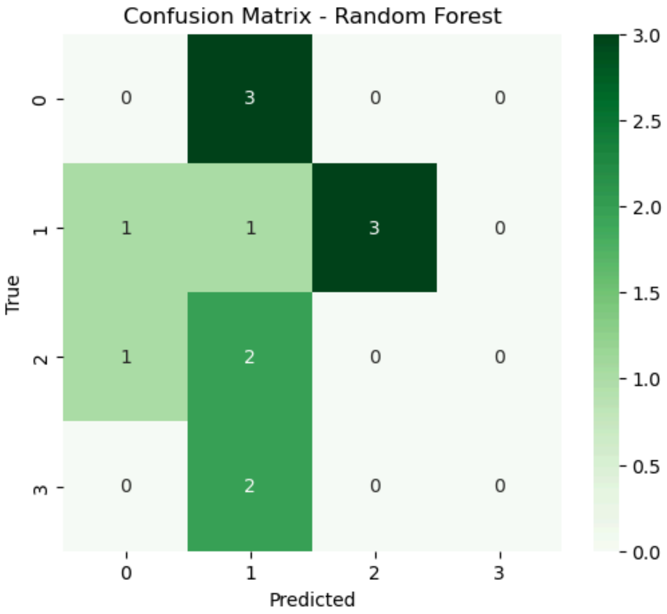
1. **Logistic Regression**
    - Used as a baseline model
    - Feature scaling applied
2. **Random Forest Classifier**
    - Used to capture non-linear relationships
    - Feature importance extracted

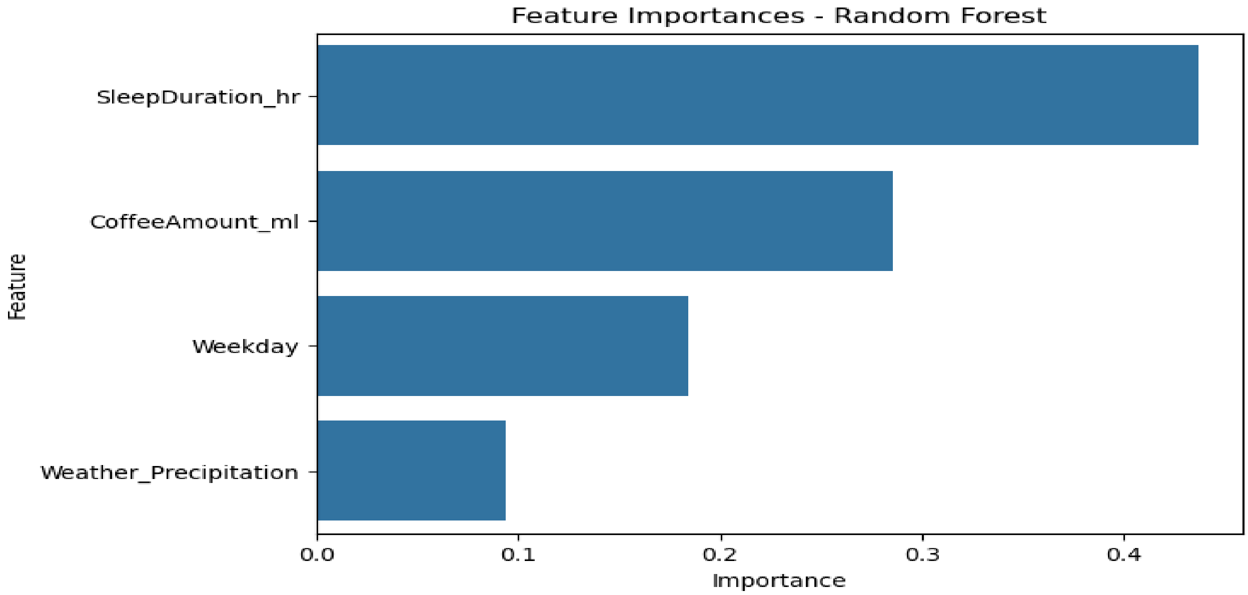
6.4 Evaluation Metrics

- Accuracy
- Precision, Recall, and F1-score
- Confusion matrices
- Feature importance visualization (Random Forest)

Random Forest generally achieved stronger performance and provided interpretability through feature importance scores.


Confusion Matrix - Logistic Regression
True
Predicted
0
1
2
3
0 1 0 2
1 1 3 0
1 2 0 0
0 1 0 1
3.0
2.5
2.0
1.5
1.0
0.5
0.0


Confusion Matrix - Random Forest
True
Predicted
0
1
2
3
0 3 0 0
1 1 3 0
1 2 0 0
0 2 0 0
3.0
2.5
2.0
1.5
1.0
0.5
0.0


Feature Importances - Random Forest
Feature
SleepDuration_hr
CoffeeAmount_ml
Weekday
Weather_Precipitation
Importance
0.0
0.1
0.2
0.3
0.4

7. Findings

The main findings of the project can be summarized as follows:

- Coffee consumption alone does not fully explain daily energy levels.
- Sleep duration plays a crucial role in perceived energy.
- Behavioral factors such as weekdays versus weekends influence energy levels.
- Weather conditions show minor but observable effects.
- Machine learning models demonstrate that **energy level is best explained through a combination of multiple features**, rather than a single variable.

8. Limitations and Future Work

Limitations

- The dataset is based on personal self-reported data, which may introduce subjectivity.
- The sample size is relatively small.
- Energy level is an ordinal variable, which may limit regression-based interpretations.

Future Work

- Collect data over a longer time span
- Include caffeine sources other than coffee (e.g., tea, energy drinks)
- Incorporate academic workload or stress indicators
- Experiment with additional ML models (e.g., Gradient Boosting, XGBoost)

9. Conclusion

This project demonstrates the complete data science pipeline, from data collection and enrichment to exploratory analysis, statistical testing, and machine learning. The results highlight the multifactorial nature of daily energy levels and emphasize the importance of combining behavioral, lifestyle, and contextual data in personal analytics.