Forecasting Sleep Efficiency with a Machine Learning-Based Analysis of Lifestyle and Sleep Behavior Data

Team members:
Brandon Ismalej
Jittapatana (Patrick) Prayoonpruk
John Lee

Table Contents

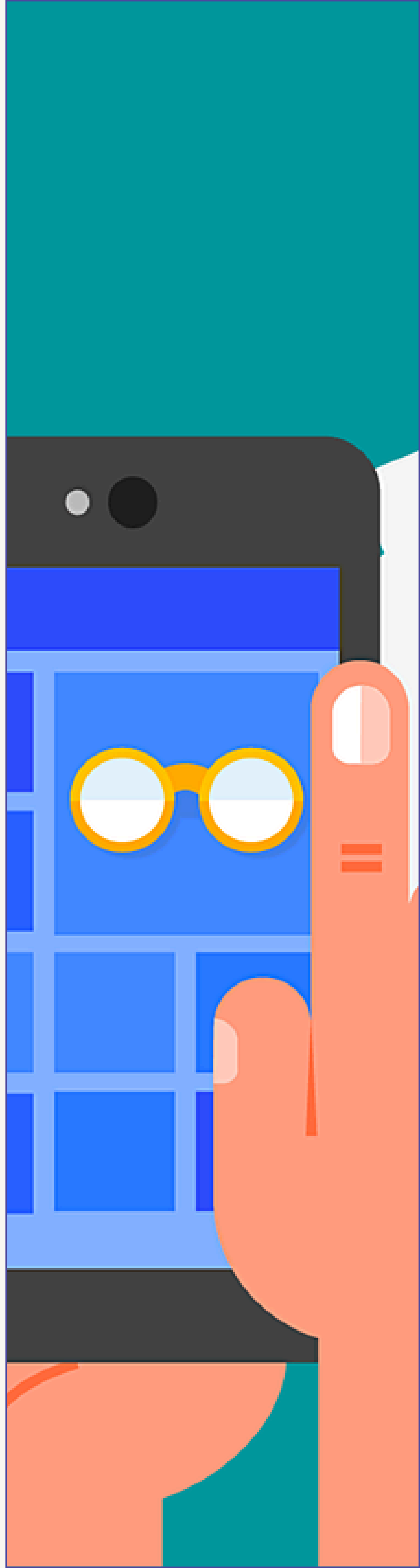
Introduction

- Project Overview
  - Utilize a machine learning approach to forecast sleep efficiency based on behavioral, physiological, and lifestyle factors.

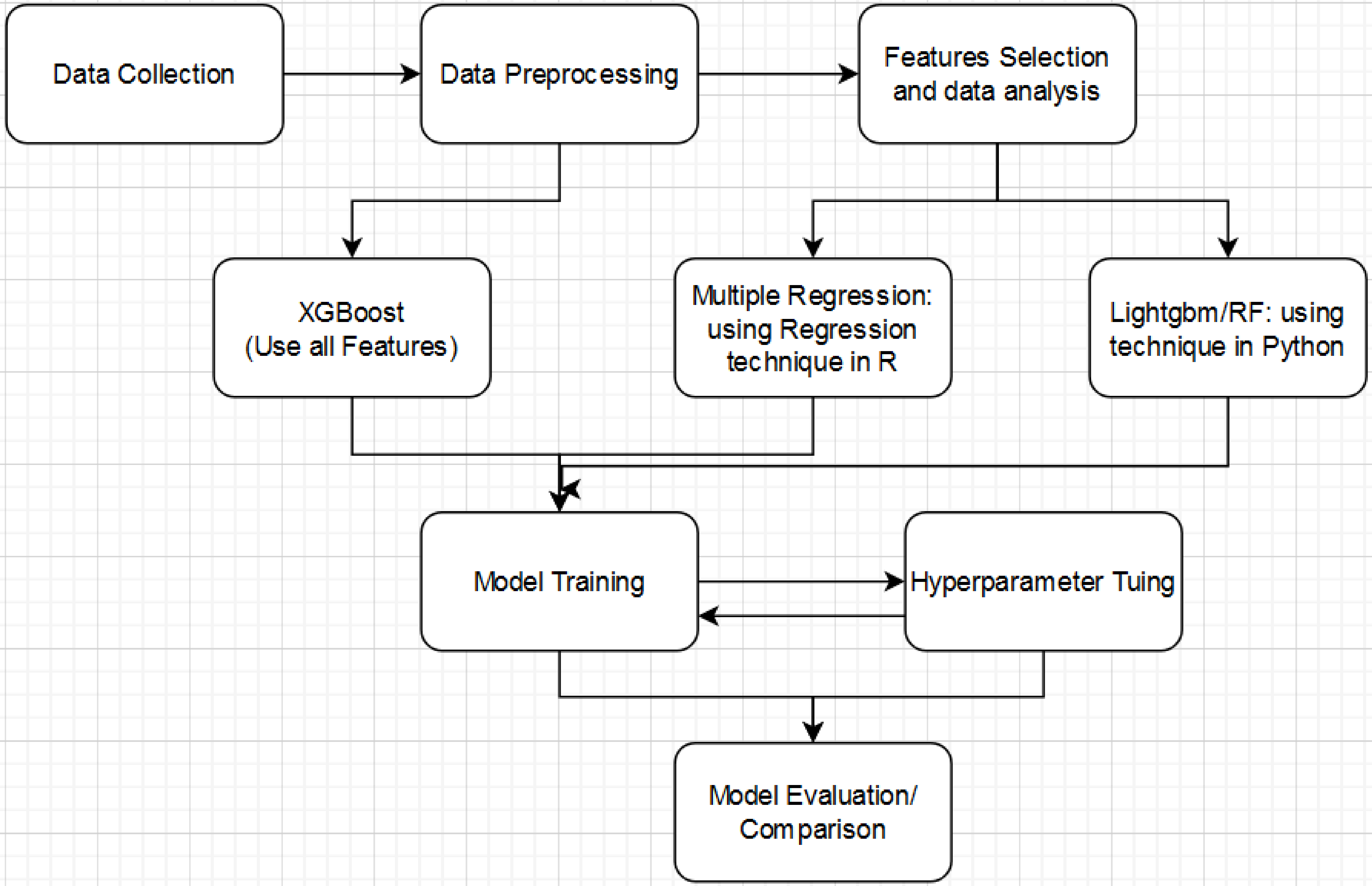
- Why This Project is Useful
  - Offers a data-driven approach to address challenges in understanding sleep quality.
  - Provides valuable insights for individuals to improve their sleep and overall health.
  - Supports healthcare professionals in identifying and managing sleep-related conditions

Introduction

- Possible Applications:
  - Development of personalized sleep improvement tools
  - Enhancement of health-monitoring devices with predictive capabilities.
  - Integration into fitness and wellness platforms for lifestyle optimization.
  - Healthcare applications to identify and mitigate factors impacting sleep disorders.



Methodology



Data Collection

- Publicly available from Kaggle - Sleep Efficiency Dataset
- 14 features including sleep and lifestyle factors (e.g., sleep duration, REM percentage, caffeine consumption, etc.).
- Number of Samples: 452 observations.
- test size: 20% - 91 samples
- train size: 80% - 361 samples
- cross validation: k=5

Data Preprocessing

LABEL ENCODING

categorical variables were converted to binary: gender and smoking status

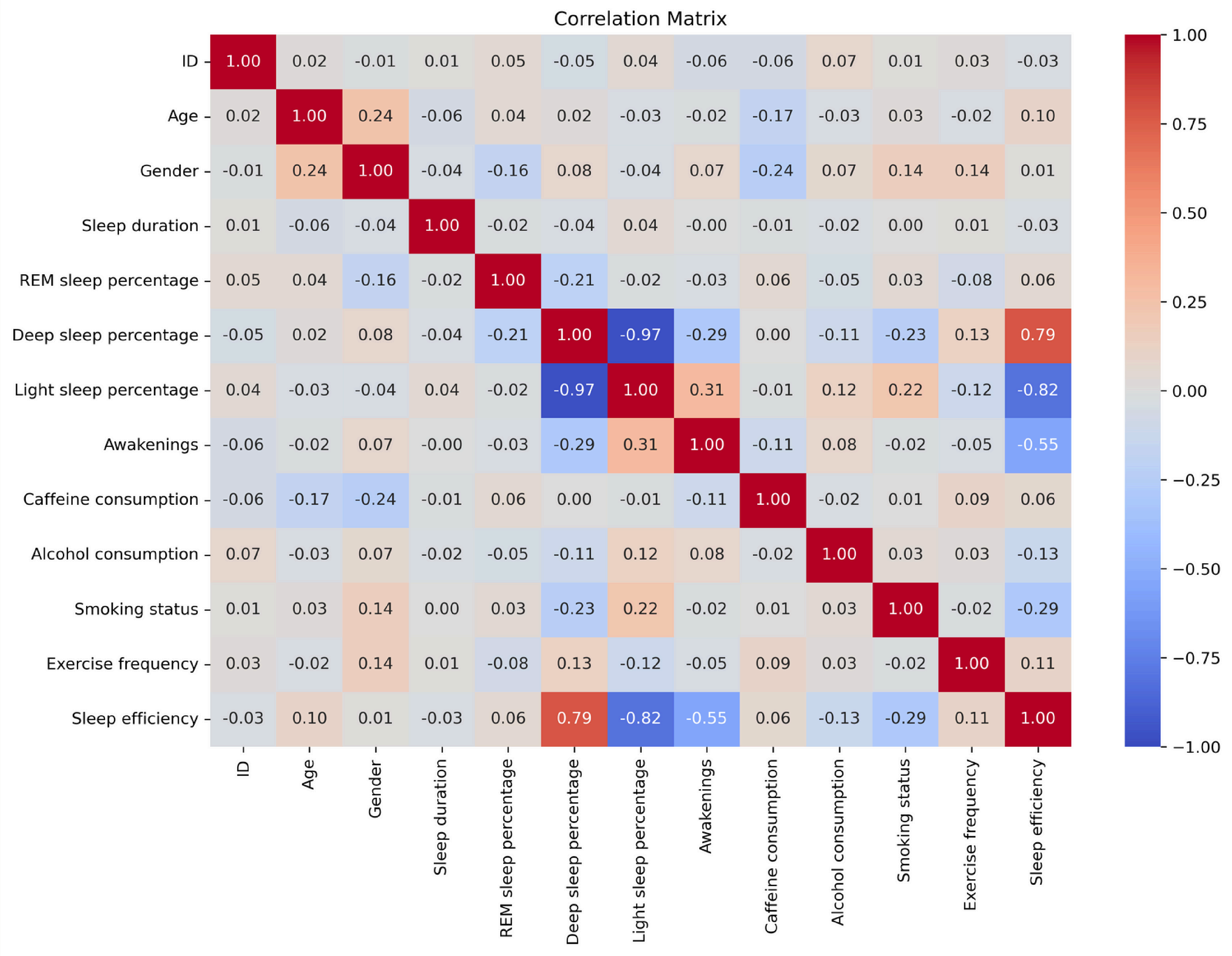
HANDLE MISSING VALUES

Missing values were handled using appropriate techniques, such as imputing with the mean or median based on the relevant group to ensure data consistency and accuracy

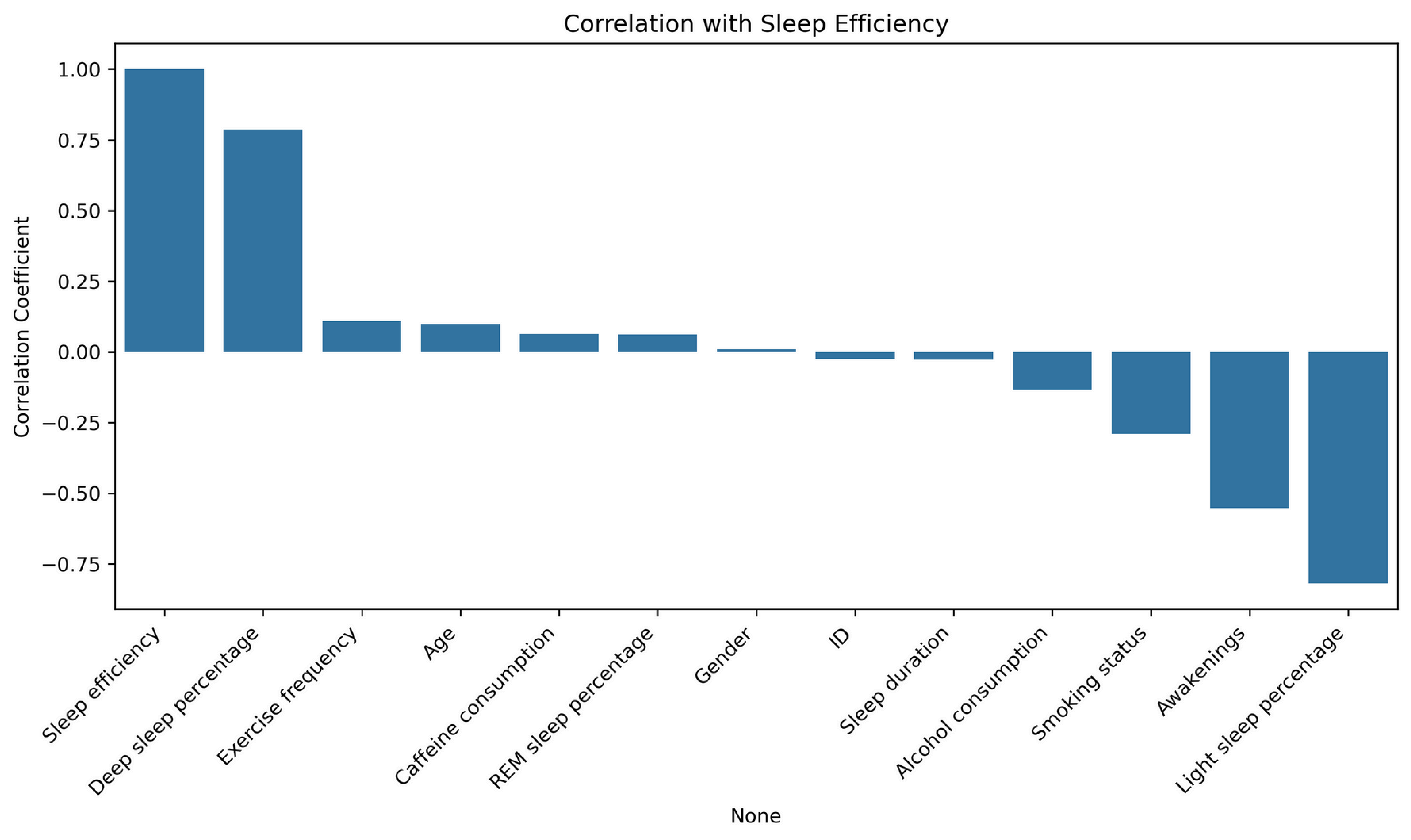
Feature Selection

Method	Idea	Strengths	Weaknesses
Correlation Analysis	Measures the linear relationship between features and the target (or among features).	- Useful for initial filtering. - Simple and quick to compute.	- Captures only linear relationships. - Does not consider multicollinearity among features.
Mutual Information (MI)	Measures dependency between a feature and the target variable, capturing both linear and non-linear relationships.	- Captures non-linear relationships. - Provides a quantitative measure of dependency.	- Computationally intensive, especially for large datasets or complex models.
Recursive Feature Elimination (RFE)	Iteratively removes the least important features based on a model's performance, leaving the most predictive subset.	- Considers feature interactions. - Works well with small to medium-sized datasets.	- Depends on the choice of the underlying model.
Random Forest Feature Importance	Ranks features based on their contribution to reducing impurity (e.g., variance, Gini) in decision tree splits across a forest.	- Handles non-linear relationships and feature interactions effectively.	- Biased toward features with more levels or higher variance.
Correlation Matrix (Redundancy Check)	Identifies highly correlated feature pairs to remove redundancy.	- Helps reduce redundancy in features.	- Only considers linear relationships; might miss complex dependencies.

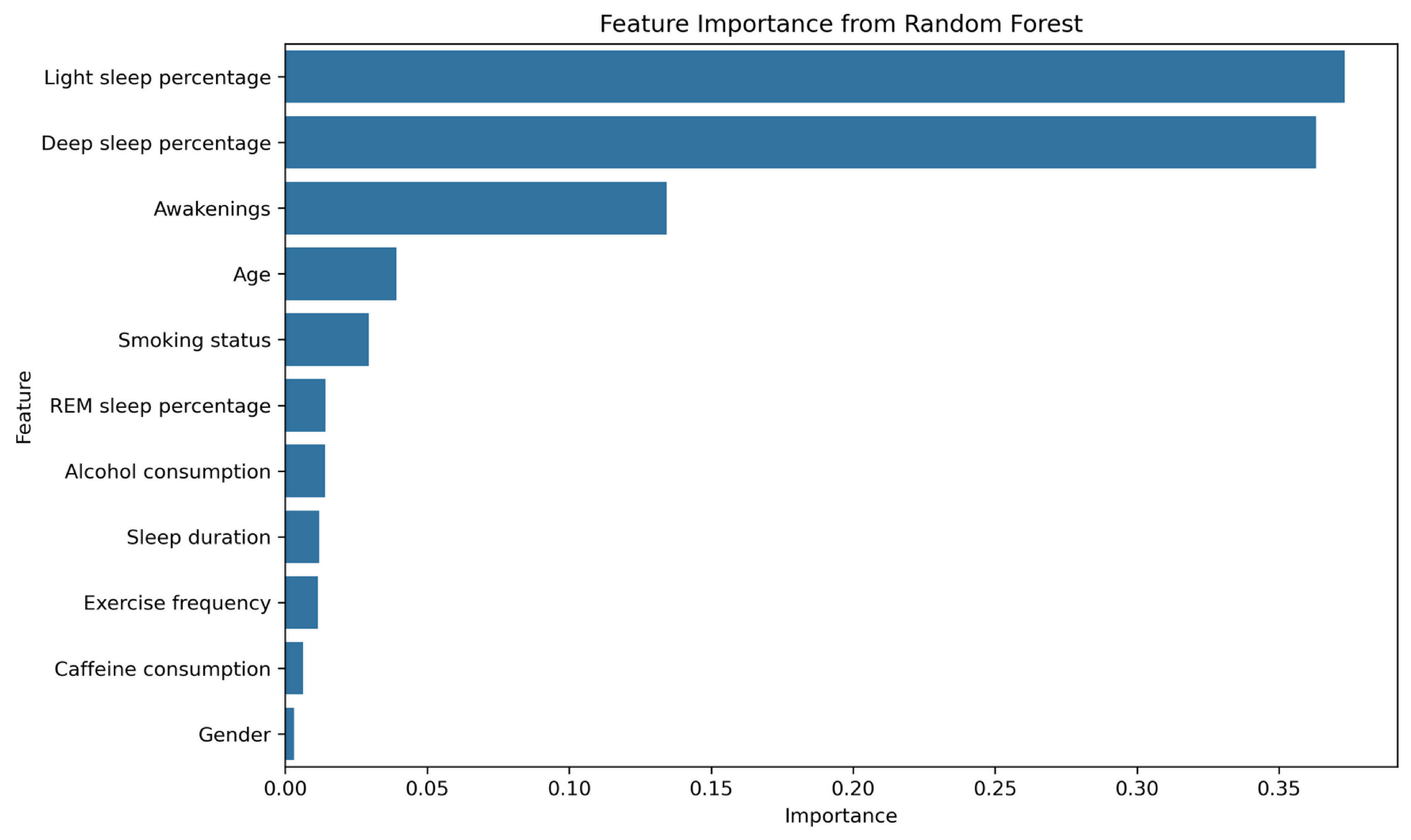
Feature Selection



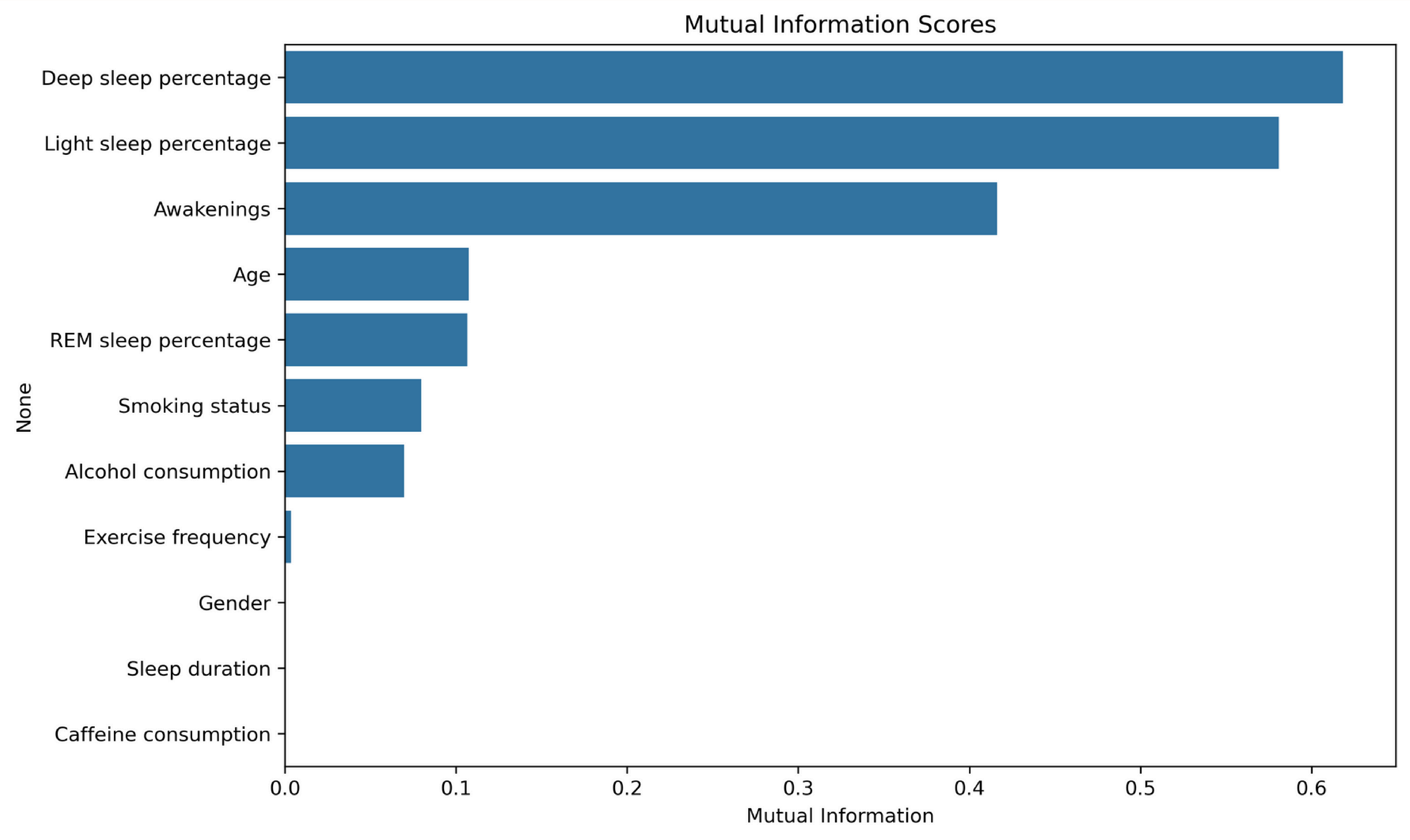
Feature Selection



Feature Selection



Feature Selection



Feature Selection-Final Features

- **Light Sleep Percentage**
  - *Key Insight:* Strong negative correlation (-0.82) with Sleep Efficiency.
  - Significant in Correlation, MI, RFE, and Random Forest.
  - *Impact*: Indicates less restorative sleep.
- **Deep Sleep Percentage**
  - *Key Insight*: Strong positive correlation (+0.79) with Sleep Efficiency.
  - High importance across all methods.
  - *Impact*: Represents the most restorative sleep phase.
- **Awakenings**
  - *Key Insight*: Moderately negative correlation (-0.55) with Sleep Efficiency.
  - Important across all methods.
  - *Impact*: Frequent awakenings disrupt sleep quality.
- **Smoking Status**
  - *Key Insight*: Moderate negative correlation (-0.29) with Sleep Efficiency.
  - Highlighted by RFE and domain knowledge.
  - *Impact:* Behavioral factor linked to poor sleep quality.

Multiple Regression

A statistical method that models the relationship between one dependent variable (e.g., sleep efficiency) and multiple independent variables (predictors).

Why Multiple Regression?

- Simple and Interpretable: Easy to understand and implement.
- Quantifies Relationships: Estimates the impact of each predictor on the outcome.
- Works Well for Linear Relationships: Effective when predictors have a linear association with the target.

Key Features:

- Predicts outcomes by fitting a linear equation to the data:

$y=\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta nxn+\epsilon$

- Assumes independence, linearity, and normality of residuals.

Ideal for Sleep Data:

- Suitable for understanding how specific factors (e.g., hours slept, activity level, stress) contribute to sleep efficiency, especially when relationships are linear.

XGBoost

A powerful machine learning algorithm based on gradient boosting.

Why XGBoost?

- Fast and Efficient: Optimized for speed and performance.
- Prevents Overfitting: Uses L1/L2 regularization for better generalization

Key Features:

- Builds decision tree ensembles to minimize prediction errors.
- Highly customizable with tunable hyperparameters.
- Provides feature importance insights to understand key factors influencing sleep efficiency.

Ideal for Sleep Data:

- Works well with structured, tabular datasets like those often used in sleep analysis.

LightGBM

A gradient boosting framework that builds decision tree models with a focus on speed and efficiency.

Why LightGBM?

- Fast and Scalable: Handles large datasets with low memory usage.
- Efficient with Large Features: Optimized for high-dimensional data.
- Prevents Overfitting: Includes built-in regularization and early stopping.

Key Features:

- Uses leaf-wise tree growth for deeper, more accurate trees.
- Highly customizable with a wide range of hyperparameters.

Ideal for Sleep Data:

- Perfect for structured datasets and scenarios where computational efficiency is critical, such as real-time sleep efficiency predictions.

Random Forest

An ensemble machine learning algorithm that builds multiple decision trees to improve prediction accuracy and robustness.

Why Random Forest?

- Accurate and Robust: Reduces overfitting by averaging multiple decision trees.
- Interpretable: Provides feature importance for understanding key predictors.

Key Features:

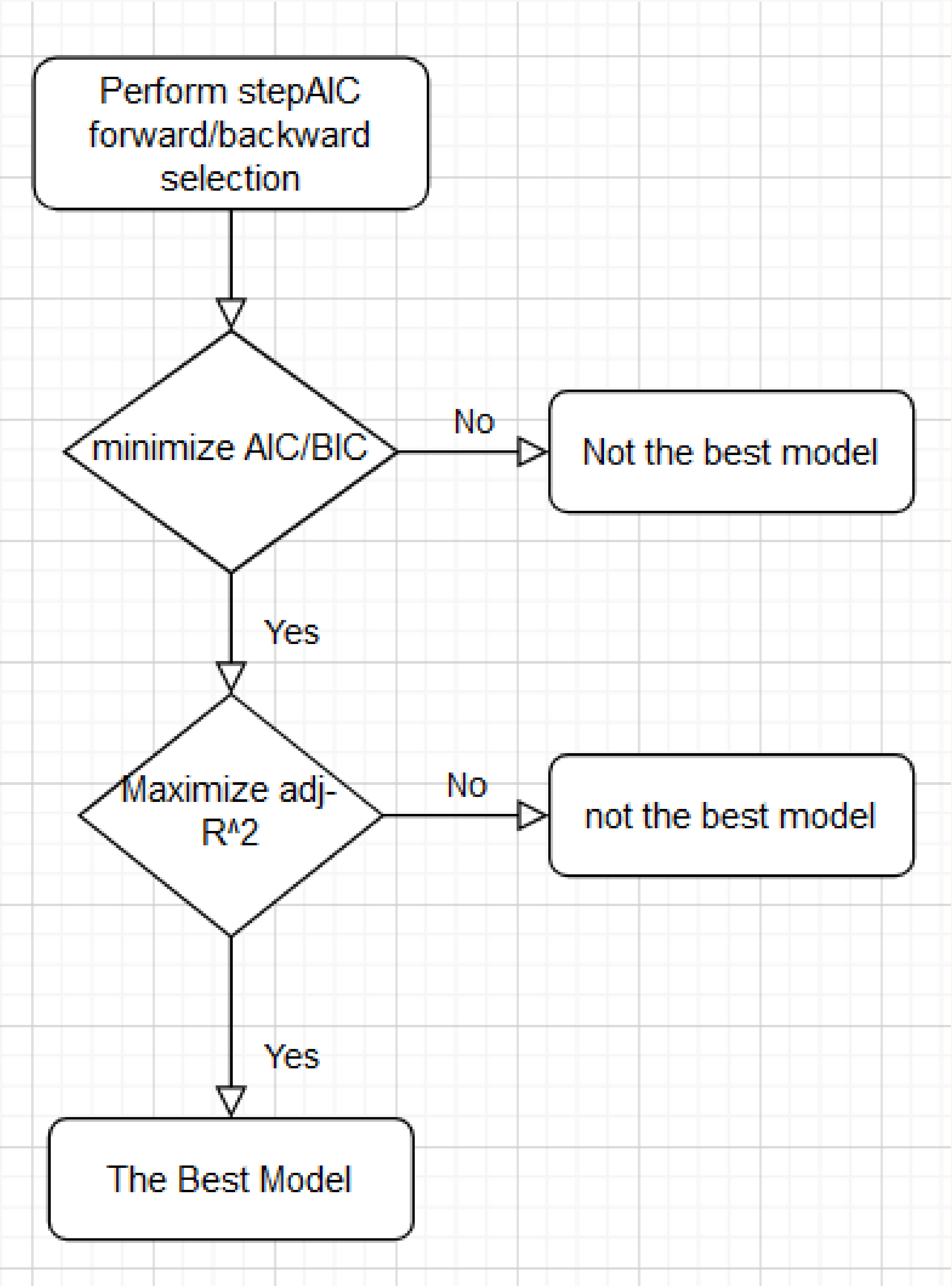
- Constructs multiple decision trees using random subsets of data and features.
- Combines predictions from all trees (majority vote for classification, averaging for regression).
- Resistant to overfitting by leveraging randomness and averaging.

Ideal for Sleep Data:

- Excels with tabular data and datasets with complex feature interactions, making it suitable for analyzing sleep efficiency predictors.

R Analysis in Multiple Regression

AIC	BIC	ADJ-R^2
• A metric that evaluates the quality of a model by balancing goodness-of-fit and complexity. • Lower AIC is better; it indicates a simpler, more accurate model. • Used to avoid overfitting by penalizing models with more predictors.	• Similar to AIC but imposes a stronger penalty for models with more predictors. • Favored when you aim for simplicity, especially with larger datasets. • Lower BIC means a more parsimonious model.	• Measures the proportion of variance in the dependent variable explained by the predictors. • Adjusted for the number of predictors to prevent overestimation of explanatory power. • Higher Adjusted R^2 is better; it indicates better model performance without overfitting

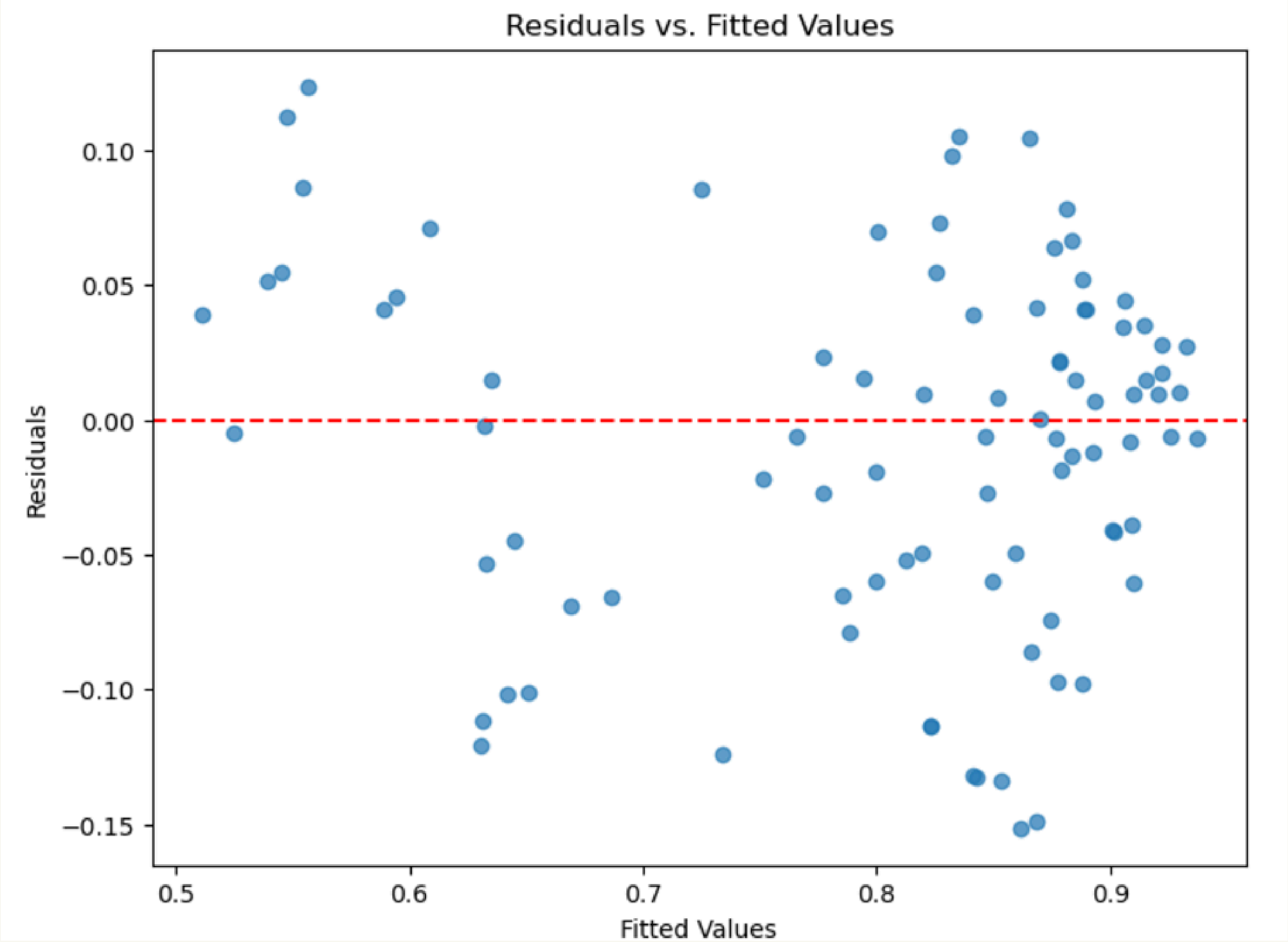


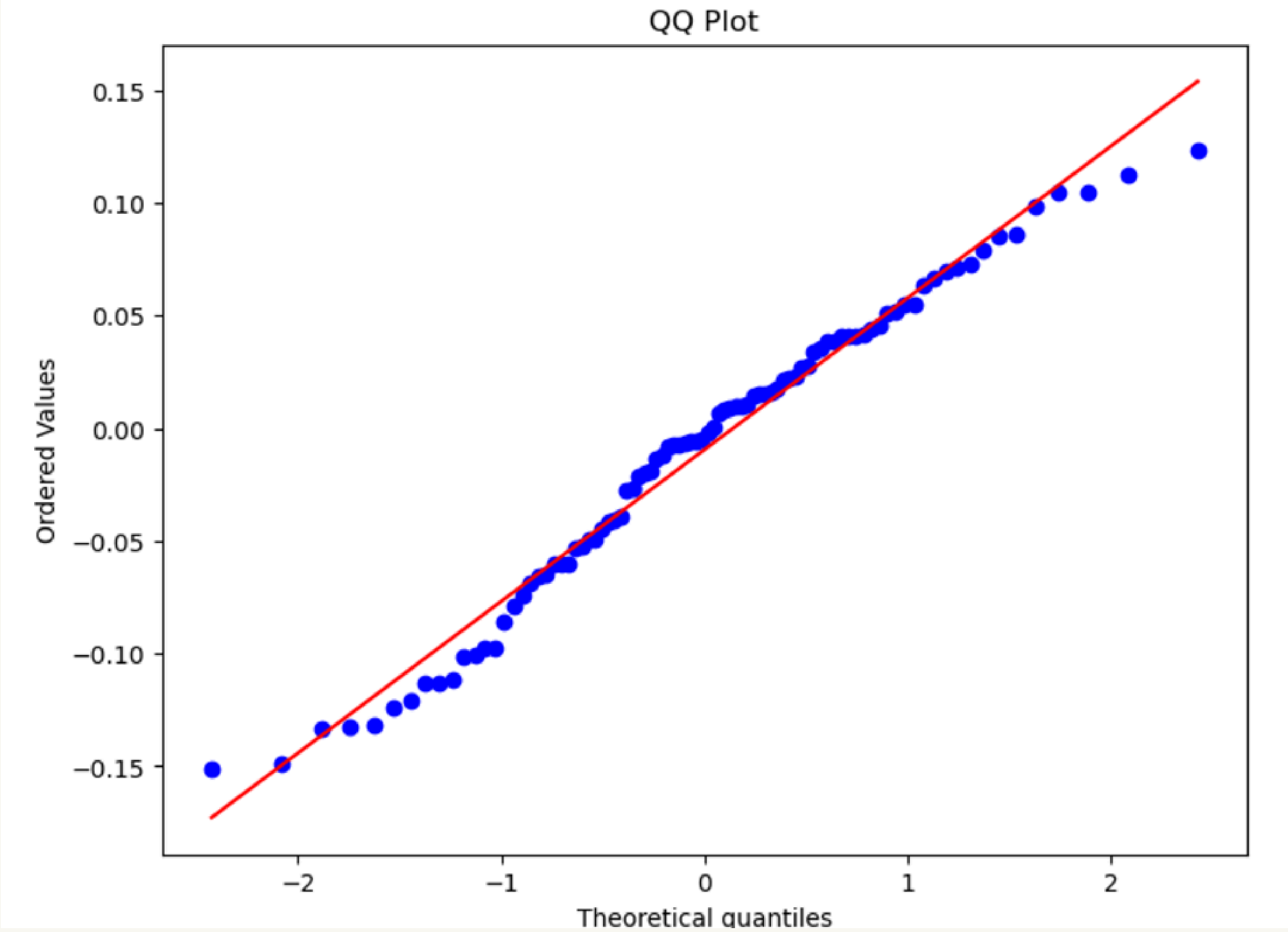
R Analysis in Multiple Regression

Best Final Model

```
Call:
lm(formula = Sleep.efficiency ~ Light.sleep.percentage + Awakenings +
    Smoking.status + Age + REM.sleep.percentage, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16139 -0.03957  0.00794  0.04281  0.13765

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.9386711  0.0211662  44.348  < 2e-16 ***
Light.sleep.percentage -0.0059781  0.0002038 -29.331  < 2e-16 ***
Awakenings            -0.0349651  0.0022809 -15.330  < 2e-16 ***
Smoking.status        -0.0422649  0.0062643  -6.747  4.7e-11 ***
Age                    0.0007580  0.0002191   3.460 0.000592 ***
REM.sleep.percentage   0.0016509  0.0008182   2.018 0.044214 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06115 on 446 degrees of freedom
Multiple R-squared:  0.7978,    Adjusted R-squared:  0.7955
F-statistic: 351.9 on 5 and 446 DF,  p-value: < 2.2e-16
```

Best Final Model



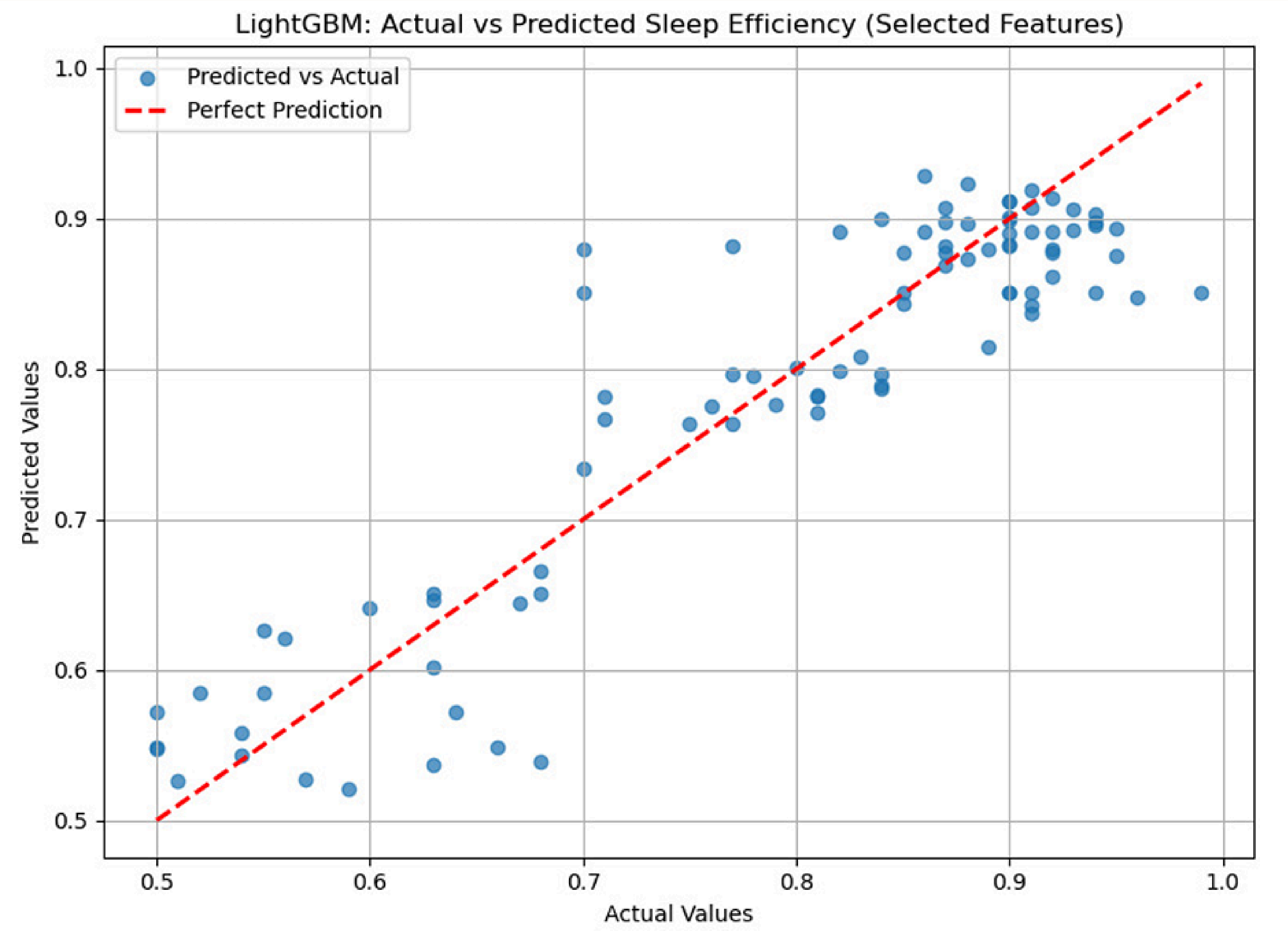


Model Evaluation

RMSE: Root Mean Squared Error

It measures the average magnitude of error between predicted and actual values

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$



Results

Model	Features	RMSE
Multiple Regression	Best Model Based on stepAIC	0.0617
XGBoost (without tuning)	All Features	0.0563
XGBoost (with tuning)	All Features	**0.0494**
LightGBM	All Features	0.0511
LightGBM	Feature Selection	0.0525
Random Forest	All Features	0.0503
Random Forest	Feature Selection	0.0557

Thank you

Q&A Session