Final Project: Classification of Sleep Disorders Based on Lifestyle and Physiological Metrics

Course: STAT 385 – Final Project

Date: November 22, 2025

Tan Ho

1. Scientific Questions and Objectives

Our group aims to address a critical health classification problem using the "Sleep Health and Lifestyle" dataset.

Scientific Question: Can we accurately predict the specific class of sleep disorder (None, Insomnia, or Sleep Apnea) an individual has based on their daily habits (such as physical activity and steps) and physiological health metrics (such as Blood Pressure, BMI, and Stress Level)?

Objective: By identifying the most significant predictors of sleep disorders, we aim to provide a data-driven approach for early diagnosis. We implemented and compared three statistical learning methods—Multinomial Logistic Regression, Decision Trees, and Random Forests—to determine which model offers the best balance of predictive accuracy and interpretability.

2. Preliminary Exploratory Data Analysis (EDA)

Before modeling, we performed rigorous data cleaning and exploratory analysis to guide our method selection.

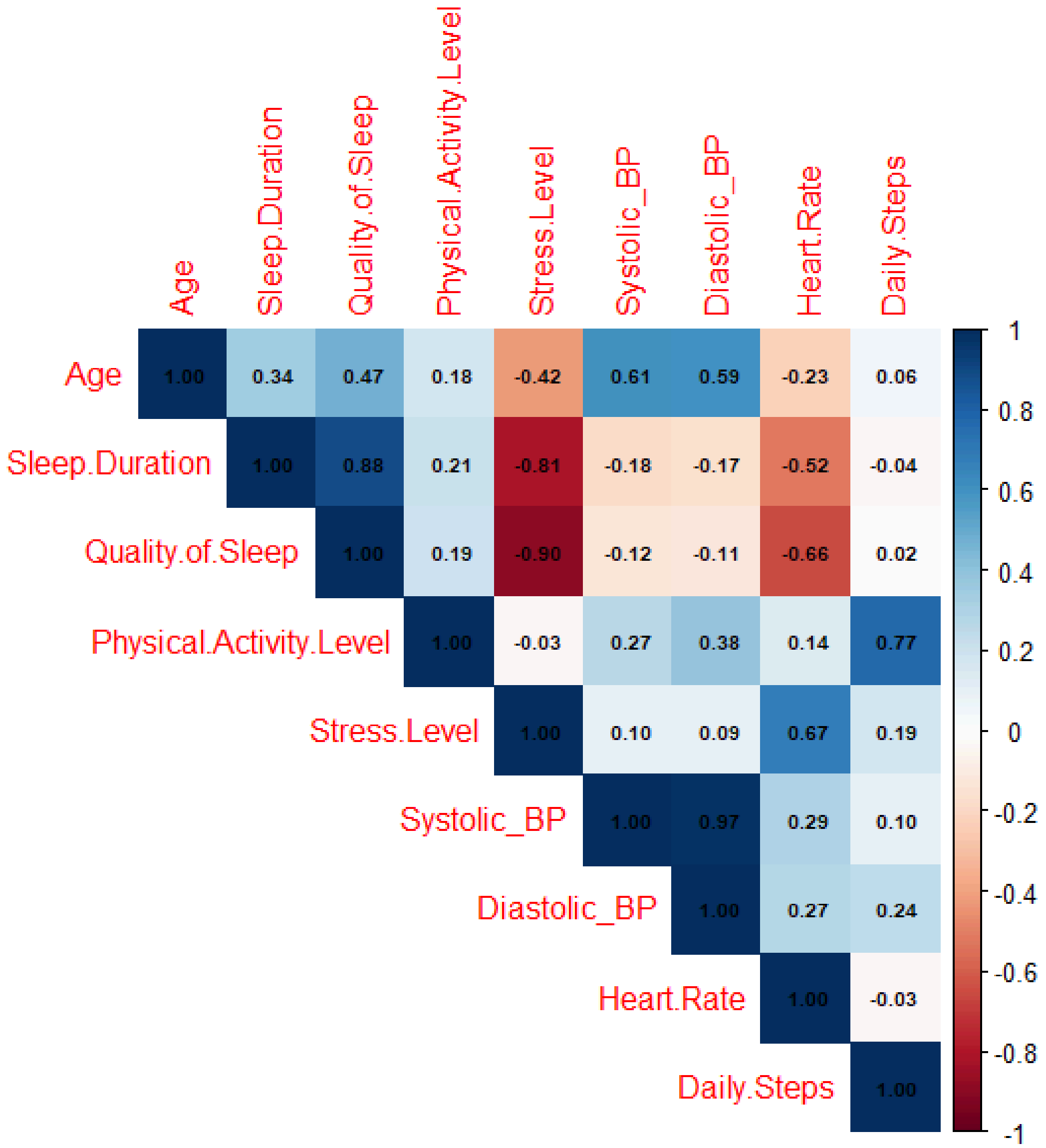
Data Preprocessing & Justification

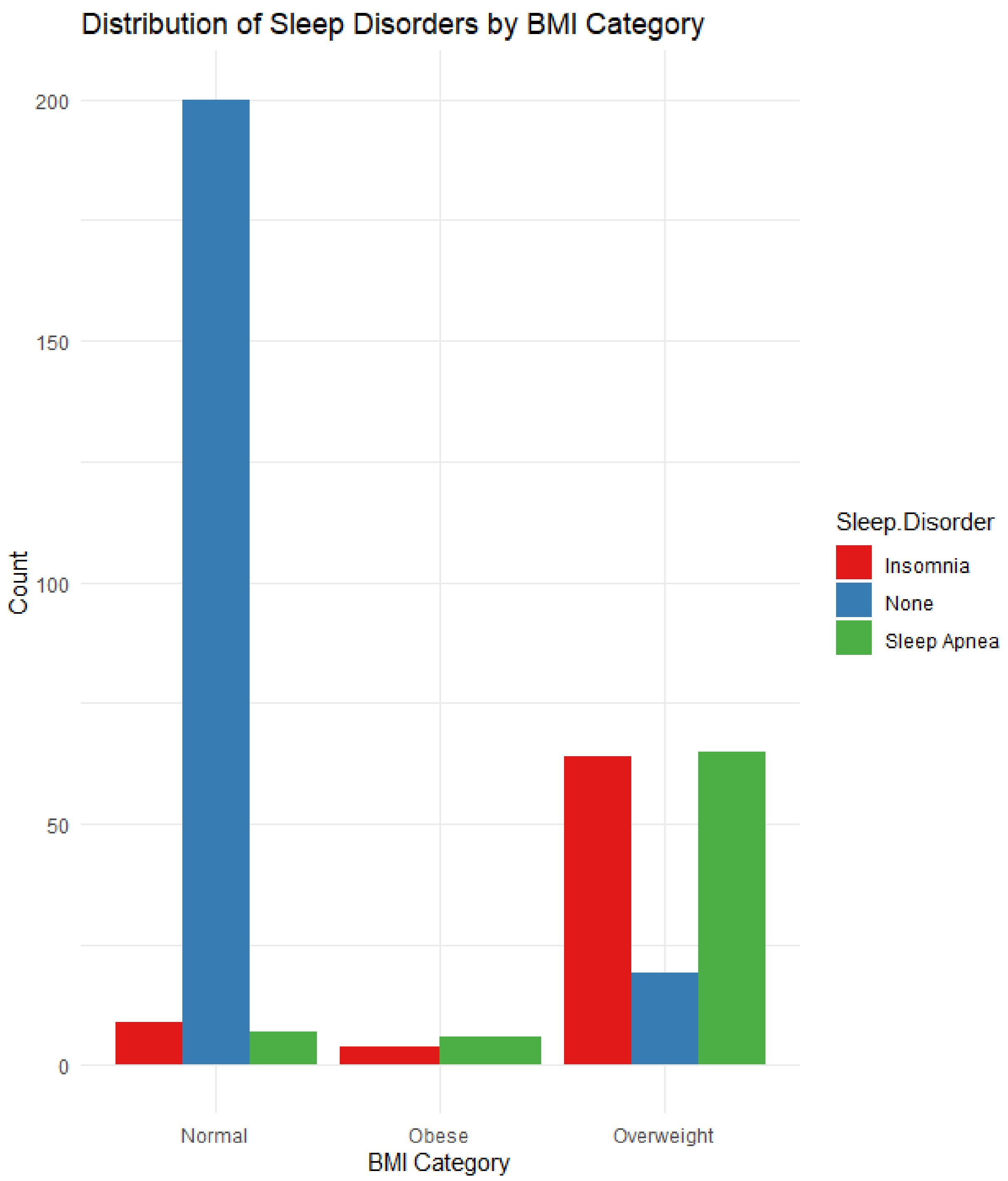
- **Blood Pressure:** The original Blood Pressure column (e.g., "126/83") was split into two numerical features: Systolic_BP and Diastolic_BP to allow for mathematical modeling.
- **BMI Categorization:** We merged inconsistent labels in the BMI Category column (e.g., "Normal" and "Normal Weight") into a single "Normal" category.
    - **Justification:** According to World Health Organization (WHO) standards, both labels correspond to the same healthy BMI range (18.5–24.9). Keeping them separate would arbitrarily split the baseline group and dilute the statistical power of the model.

EDA Findings

- **Correlation Analysis:** We generated a correlation matrix which revealed a strong negative correlation between Stress Level and Sleep Duration. This suggests that high stress is a direct inhibitor of sleep quantity.
- **BMI and Disorders:** A bar chart analysis of Sleep Disorder by BMI Category revealed a stark contrast. Individuals in the "Normal" weight category predominantly showed "None" (no disorder). In contrast, the "Obese" category was dominated by "Sleep Apnea." This guided our decision to include BMI as a primary predictor in all models.

Correlation Heatmap of Health Metrics





3. Methods and Model Implementation

To ensure a fair comparison, we split the data into a **70% training set** and a **30% testing set**. We set a random seed (set.seed(123)) to ensure all results are reproducible.

Method 1: Multinomial Logistic Regression

Since our target variable Sleep Disorder has three nominal levels (None, Insomnia, Sleep Apnea), we fitted a Multinomial Logistic Regression model.

- **Model Equation:** The probability of a specific sleep disorder class k relative to the reference class "None" is modeled as:

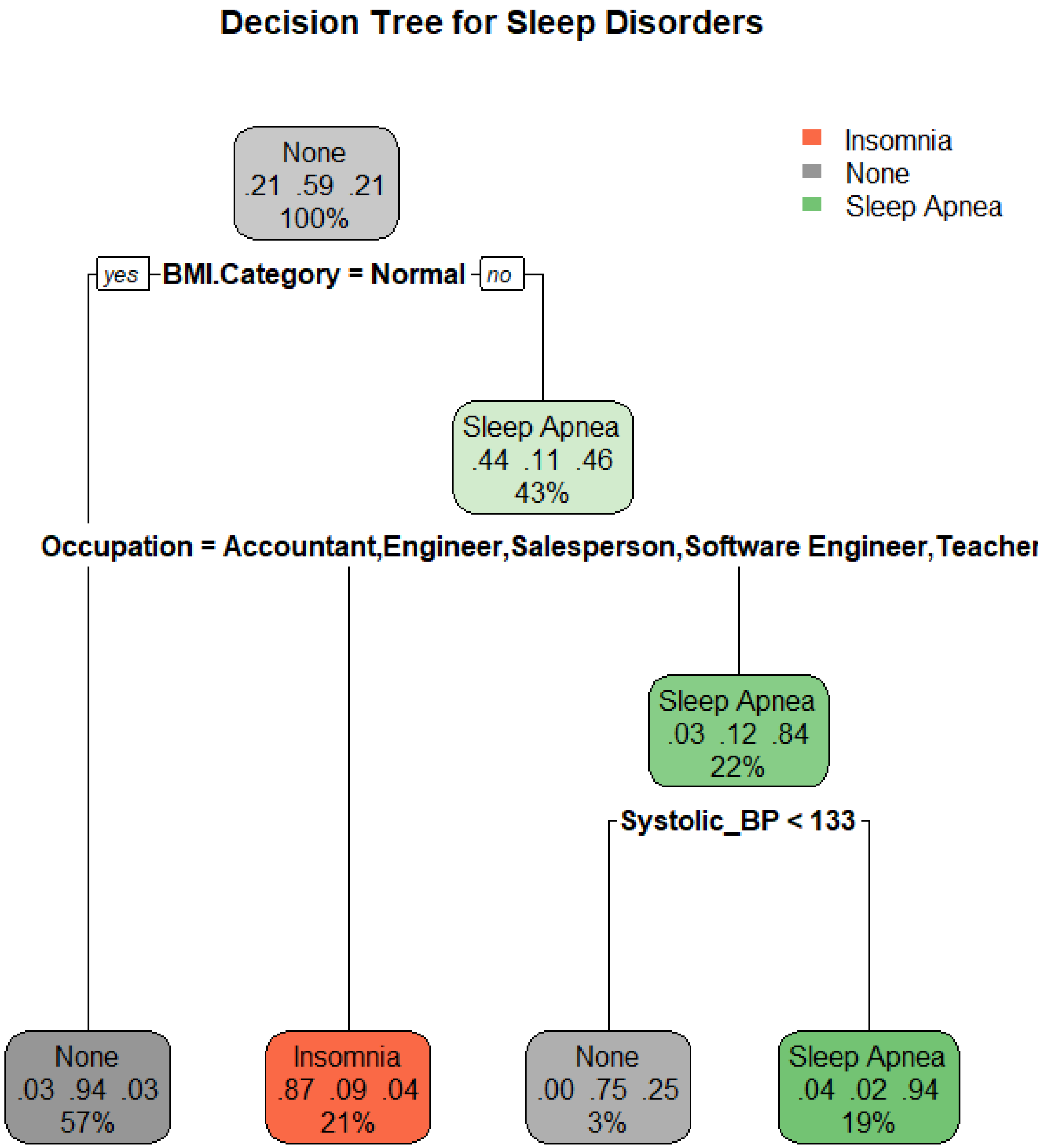
$$\ln(P(Y=k) / P(Y=None)) = \beta_0 + \beta_1(Age) + \beta_2(BMI) + \beta_3(Stress)$$

- **Assumptions & Multicollinearity:** We assessed multicollinearity and observed an extremely high correlation (0.97) between Systolic_BP and Diastolic_BP. While both were included in the initial fit, we acknowledge this inflates standard errors. In clinical practice, we would recommend using only Systolic_BP to resolve this redundancy.
- **Results & Interpretation:** The model achieved an accuracy of **86.49%**. Parameter estimations revealed a positive coefficient for Stress Level relative to the 'None' baseline. This indicates that as an individual's stress level increases, the probability of being diagnosed with Insomnia or Sleep Apnea significantly increases, confirming our EDA findings.

Method 2: Decision Tree

We implemented a Decision Tree to capture non-linear relationships and interaction effects between variables.

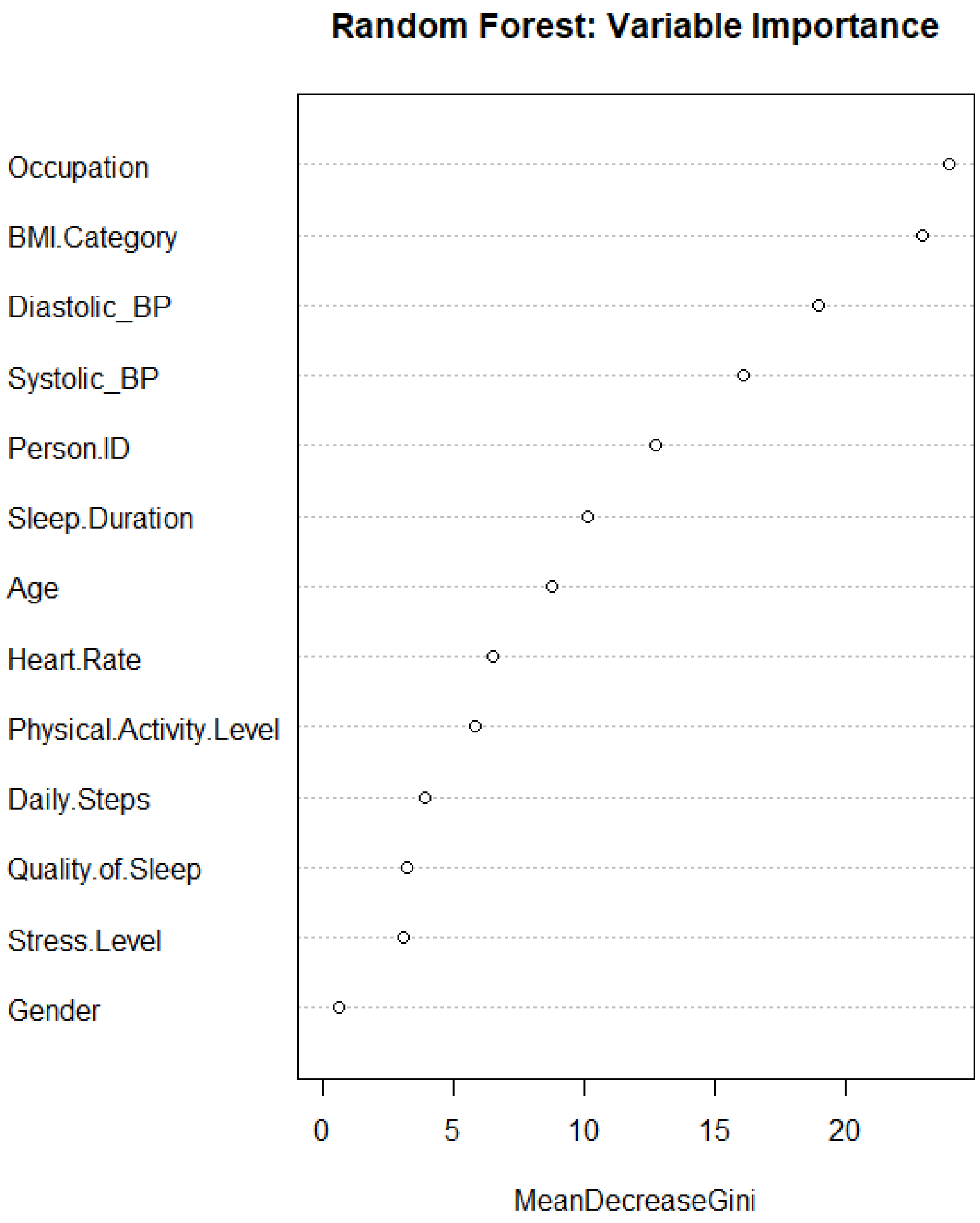
- **Tuning Parameter:** We tuned the complexity parameter (cp) to **0.01** to prune the tree and prevent overfitting.
- **Interpretation:** The tree visualization identified **Systolic Blood Pressure** as the most critical splitting feature.
  - *Rule 1:* If Systolic BP < 133, the individual is classified as "None" (Healthy).
  - *Rule 2:* If Systolic BP ≥ 133, the tree checks BMI. If "Obese," the prediction is "Sleep Apnea."
- **Results:** This method achieved an accuracy of **84.68%**.



Method 3: Random Forest

To address the high variance of single decision trees, we implemented a Random Forest ensemble.

- **Tuning Parameters:** We tuned the model by setting ntree (number of trees) to 100, which is sufficient for convergence on this dataset size, and mtry (variables sampled per split) to 3, approximating the square root of predictors ($\sqrt{p}$).
- **Variable Importance:** Analysis of the Random Forest feature importance (using Mean Decrease in Gini) confirmed that **BMI Category** and **Systolic Blood Pressure** were the most significant predictors, aligning perfectly with our EDA.
- **Results:** The model achieved an accuracy of **85.59%**, performing slightly better than the single decision tree but slightly worse than the logistic regression.



4. Comparison and Conclusion

Model Evaluation

We compared the three methods using the same test dataset.

Method	**Accuracy**	**Kappa**
Logistic Regression	**86.49%**	**0.762**
Random Forest	85.59%	0.749
Decision Tree	84.68%	0.732

Conclusion and Recommendations

Our analysis successfully demonstrated that sleep disorders can be classified with high accuracy using lifestyle data.

- **Best Performing Model:** We recommend the **Multinomial Logistic Regression** model. It achieved the highest accuracy (86.49%) and Kappa score (0.762), indicating it is the most reliable method.
- **Clinical Recommendation:** While Logistic Regression was the most accurate, we also recommend the **Decision Tree** for clinical settings. Its visual nature provides doctors with an immediate, interpretable rule: "Check if Systolic BP ≥ 133."
- **Limitations:** The Decision Tree was the weakest performer (84.7%) likely because its rectangular decision boundaries could not capture the smooth, continuous increase in risk associated with rising stress levels as effectively as the regression model.

Future Improvements

- **Data Balance:** The "Obese" class was relatively small but highly predictive of Apnea. Collecting more data on obese individuals would help confirm if this correlation holds in larger populations.
- **Feature Engineering:** Given the high multicollinearity found between Systolic and Diastolic blood pressure (0.97), future work should combine them into a single "Mean Arterial Pressure" feature to simplify the model without losing information.