Sleep Disorder Analysis: Unveiling the Interplay Between Lifestyle Health and Sleep Quality

Sarah Alshumayri S20106125, Reema Abdallah S20106463, Yehya Asseri S23108710

Fall 2023

Contents

Introduction

This project aims to understand the impact of lifestyle and health factors on sleep quality and disorders. This project compiles data including metrics related to sleep duration, sleep quality, physical activity, stress levels, BMI categories, blood pressure, heart rate, and daily steps. We aim to use data visualization and statistical analysis methods to investigate the variables that significantly affect sleep health. Preliminary analysis may reveal insights such as the correlation between physical activity, stress levels, and sleep quality, but the project will also explore more complex relationships and potential predictors of sleep disorders.
The rest of the report is organized as follows: section 2 provides a background on the importance of sleep health and its relationship with lifestyle factors, section 3 presents the research question and problem statement that the report aims to answer, section 4 discusses the data used in this project, its sources, and provides a brief overview of the contents of each dataset, section 5 analyzes those datasets and offers a statistical view of the data, section 6 presents the findings of the project, section 7 discusses the implications of these findings and their potential applications, and section 8 concludes the report with a summary of the key insights and suggestions for future research.

Background

Sleep quality, a critical factor for health and well-being, is influenced by a multitude of factors, including occupational hazards, lifestyle choices, and individual behaviors. In certain professions, such as long-distance heavy goods vehicle (HGV) drivers, the combination of demanding work schedules and poor lifestyle choices leads to increased risks of chronic diseases and reduced life expectancy [2]. This is compounded by inadequate sleep, which is linked to an increased risk of accidents and comorbidities [2].

Sleep behavior is also influenced by demographic, occupational, and lifestyle factors. For instance, sleep efficiency and duration are known to decrease with age, and this is a significant concern in professions with an aging workforce [3]. Similarly, in athletes, optimal sleep is critical for performance, but factors such as training and competition times, travel, stress, and use of stimulants like caffeine can lead to substantial variation in sleep onset and offset times [1].

The most important factor influencing sleep efficiency is bedtime and low variability in sleep onset times [2]. Regular sleepers tend to exhibit consistent sleep onset and offset times compared to irregular sleepers. However, achieving this regularity can be challenging due to training schedules and other commitments [2].

For elite athletes, the biological bases of sleep, driven by homeostatic drive and the circadian clock, are relatively stable. However, sleep regularity can be significantly affected by external factors such as training schedules, psychological stress, and societal influences. These factors impact sleep regularity and highlight the importance of modifying behaviors that can lead to poor sleep quality and duration [4].

Research Question and Problem Statement

Can machine learning models effectively identify key lifestyle and health factors influencing sleep quality and having sleep disorder?

Understanding the intricate relationship between various lifestyle and health factors and their impact on sleep quality and disorders is essential for developing effective health interventions. Traditional analytical methods may not fully capture the complex interactions and nonlinear relationships between these factors. This research aims to leverage the capabilities of machine learning models to analyze a comprehensive dataset encompassing demographic, occupational, physical activity, stress levels, and health indicators. The objective is to determine how these factors collectively influence sleep duration, quality, and the presence of sleep disorders. By evaluating the performance of various machine learning models, this study seeks to pinpoint the most significant factors affecting sleep health. The insights gained could provide valuable guidance for healthcare professionals and policymakers in formulating strategies to enhance sleep quality and address sleep-related issues in the population.

Data

Unit of Observation

The unit of observation in this dataset is an individual person. Each row represents data for one individual, with various attributes related to their demographics, lifestyle, and health.

Outcome Variable

The outcome variable is 'Sleep Disorder', which is a categorical variable that indicates the presence and type of sleep disorder such as insomnia, Sleep Apnea, or none an individual may have. As Figure 1 shows, the distribution of the outcome variable is illustrated in the graph and the frequency table below:

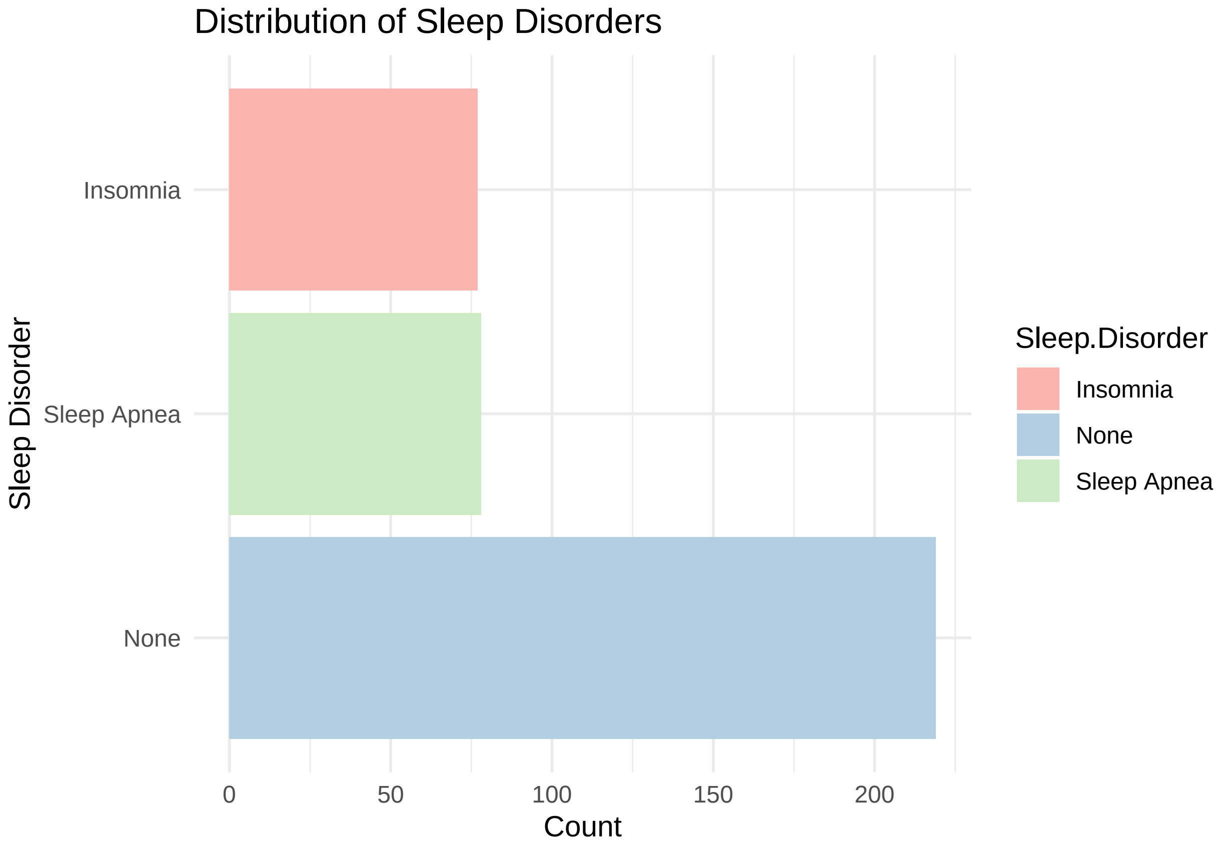


Figure 1. Outcome variable

Predictor Variables

Predictor variables include 'Gender', 'Age', 'Occupation', 'Physical Activity Level', 'Stress Level', 'Quality of Sleep', 'BMI Category', and 'Daily Steps'.

These variables are measured through surveys or collected data from individuals health records The distribution of each predictor will be presented using descriptive statistics and visualizations.

Descriptive statistics for each variable:

Table 1: Summary Statistics of Predictor Variables

Variable	Mean	Standard_Deviation	Range
Age	42.0	8.00	27 to 59
Quality of Sleep	7.0	1.00	4 to 9
Physical Activity Level	59.0	20.00	30 to 90
Stress Level	5.0	1.00	3 to 8
Daily Steps	6816.0	1617.00	3000 to 10000
BMI Levels	1.0	0.50	1 to 3
Blood Pressure 1	128.0	7.00	115 to 142
Blood Pressure 2	84.0	6.00	75 to 95
Heart Rate	70.0	4.00	65 to 86
Sleep Duration	7.1	0.79	5.8 to 8.5

Figure 2 will show the histograms for each of these variables depict their distributions, revealing a wide array of values and suggesting a rich diversity within the dataset for these predictors. For example, 'Sleep Duration' displays a normal distribution, indicating a balanced spread of sleep duration across individuals in the dataset. 'Physical Activity Level' exhibits a broad range, reflecting varied levels of physical activity among participants. Such visualizations are instrumental in comprehending the distribution and central tendencies of the predictor variables, offering insights into patterns that may influence sleep health, like the correlation between 'Stress Level' and sleep quality or the impact of 'Daily Steps' on 'Heart Rate'.

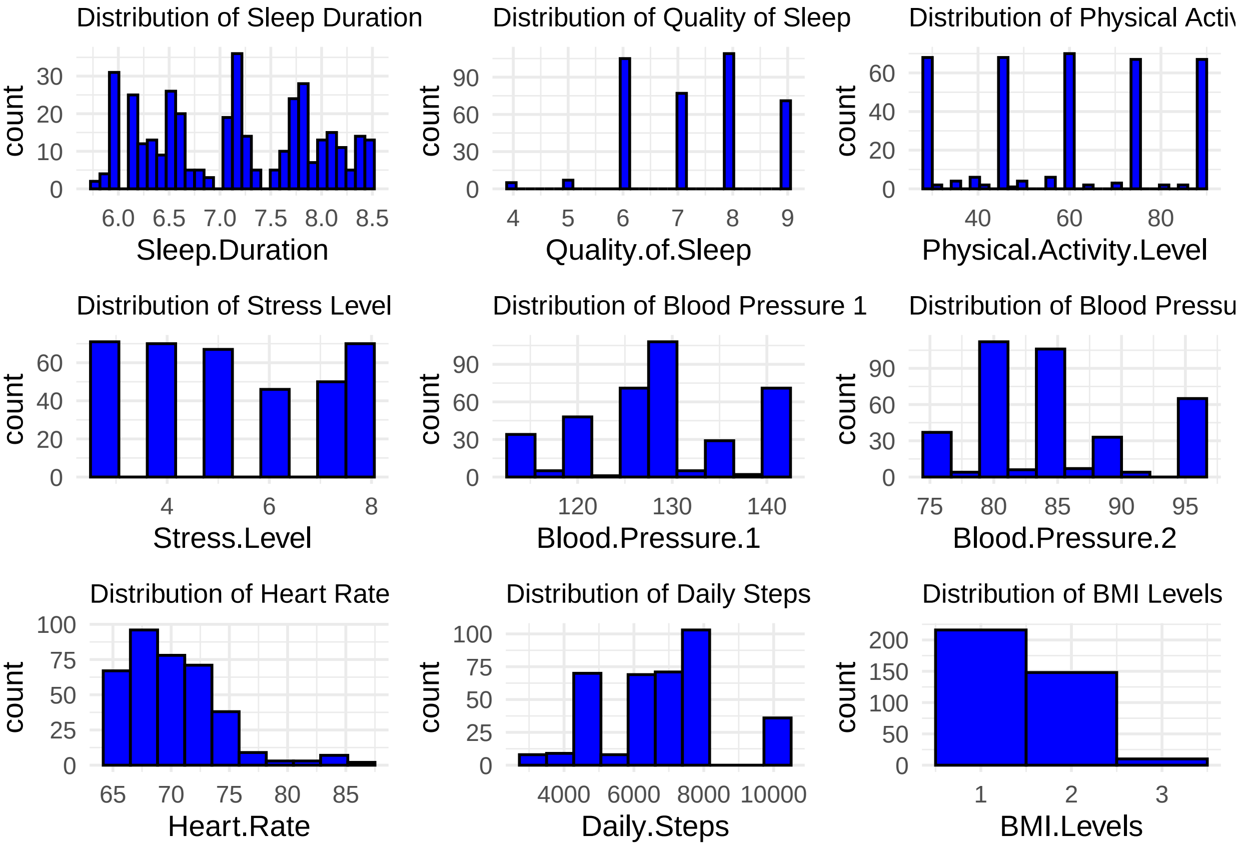


Figure 2. Predictor Variables

Potential Issues with the Data

When considering potential issues with the data, the primary concerns are the lack of variation in some categories and potential biases. For instance, certain occupations or BMI categories might have limited representation, which could affect the generalizability of the findings. Also, if the sample is not representative of the broader population (e.g., skewed towards a specific age group, or occupational category), it could introduce biases in the analysis.

Solutions to the issues

To overcome or mitigate the issues of lack of variation and potential biases in your data analysis, a multifaceted approach can be adopted. Firstly, it's essential to transparently report the limitations of the study due to these factors. Acknowledging the specific areas where the dataset may not perfectly represent the broader population or where certain categories are underrepresented adds to the credibility of the research and aids in the accurate interpretation of the results. Alongside this, the implementation of regression techniques serves as a robust method to address imbalances in the dataset. By using logistic or linear regression models, it becomes possible to control for confounding variables, allowing for a more accurate isolation of the effects of primary predictors. This approach not only helps in drawing more reliable conclusions but also enhances the overall integrity of the analysis by systematically adjusting for known dataset limitations.

Analysis

Methods/Tools Explored

In our project, we employed a variety of methods and tools to thoroughly analyze the "Sleep, Health, and Lifestyle" dataset. The primary tool used was R, renowned for its robust capabilities in data analysis, statistical computing, and graphical representation. This choice was driven by R's comprehensive support for data manipulation, visualization, and advanced analytics.

Key R packages used included:

- `tidyverse` Tidyverse includes `dplyr` and `ggplot2` for data manipulation and visualization.
- `caret` and `randomForest` for machine learning and predictive modeling.
- `skimr` for data summary and exploration.
- `gridExtra` for arranging multiple plots on a grid.
- `knitr` to create a kable for a nice-looking table.

The analysis included rigorous exploratory data analysis (EDA) to comprehend the data structure, identify missing values, and explore potential correlations among variables. Given the dataset's characteristics, Random Forest was chosen as the primary predictive modeling technique, valued for its effectiveness in handling numerous predictor variables and capturing complex data patterns.

Detailed Analysis Outline

The analysis followed these key stages:

1. **Data Preprocessing and Cleaning:**

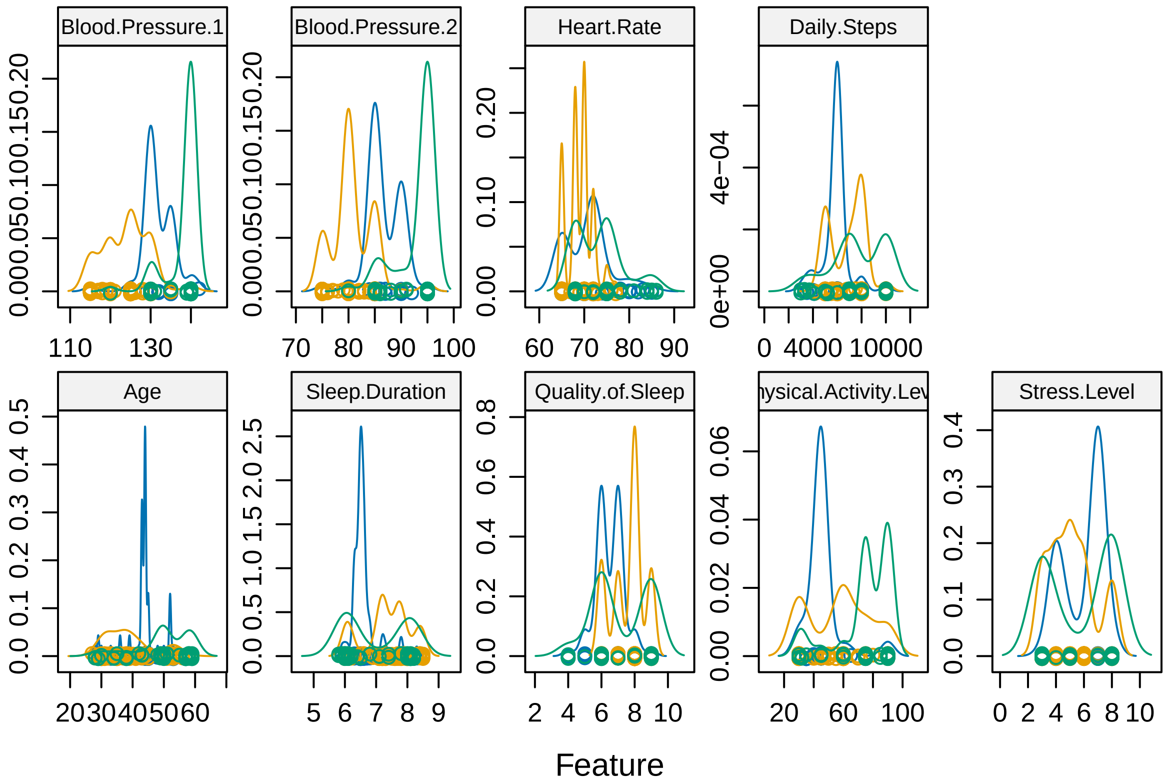
- Encoding categorical variables and normalizing numeric data.

2. **Exploratory Data Analysis (EDA):**

- Utilizing various visualization tools (like histograms, box plots, scatter plots) to understand variable distributions and relationships.

3. **Feature Engineering and Selection:**

- Identifying crucial predictor variables via correlation analysis and initial model insights.
- Crafting new features that could enhance model performance and interpretability.



4- **Model Building:** - Building and training the data with four models(Random Forest,Support Vector Machine,K-Nearest Neighbors,Linear Discriminant Analysis). - Tuning hyperparameters to optimize the model's performance. - Validating the model's using cross-validation techniques.

5. **Model Interpretation and Evaluation:**

- Interpreting the model's using feature importance scores and visualization tools like Partial Dependence Plots (PDP).
- Evaluating the model's performance through metrics like accuracy, recall, precision, and the Area Under the Curve (AUC) for ROC analysis.

6. **Validation and Testing:** -Assessing model robustness on a separate test set.

- Using a range of performance metrics to ensure reliability and accuracy.

The approach was crafted to be accessible to readers with basic knowledge of R and machine learning, explaining each step with clarity and its rationale based on the dataset's nature and the research objectives. The methodology was selected to provide a comprehensive understanding of the dataset and to ensure the predictive modeling was both robust and interpretable.

```
Sleep_dataset$Sleep.Disorder <- as.factor(Sleep_dataset$Sleep.Disorder)
model <- randomForest(Sleep.Disorder ~ Age + Gender + Sleep.Duration +Quality.of.Sleep+Stress.Level+Phys
# Save the model to an RDS file
saveRDS(model,"model_RF.rds")
```

Result

Summary of Results

The predictive analysis was conducted using a variety of machine learning models, including Random Forest (RF), K-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA). The models were trained on Sleep_dataset processed and partitioned into training and testing sets, with 10-fold cross-validation implemented to ensure robustness and reliability of the result

1. **Model Performance:**

- RF Model: Exhibited high accuracy, with a detailed confusion matrix and ROC curves indicating its effectiveness.
- kNN Model: Showed notable precision and recall rates, as evidenced by its confusion matrix.
- SVM Model: Demonstrated efficiency in classification, with strong accuracy and ROC curve performance.
- LDA Model: Performed well in classifying different sleep disorder categories, backed by accuracy metrics.

A comparative analysis of all models was provided, highlighting the most effective model for sleep disorder predict

Table 2: Summary of Accuracy Statistics

Model	Statistic	Value	NA.s
RF	Min.	0.8333333	0
RF	1st Qu.	0.9000000	0
RF	Median	0.9310345	0
RF	Mean	0.9214661	0
RF	3rd Qu.	0.9354839	0
RF	Max.	1.0000000	0
kNN	Min.	0.7241379	0
kNN	1st Qu.	0.8333333	0
kNN	Median	0.8666667	0
kNN	Mean	0.8691335	0
kNN	3rd Qu.	0.9024194	0
kNN	Max.	0.9677419	0
SVMLinear	Min.	0.7931034	0

Model	Statistic	Value	NA.s
SVMLinear	1st Qu.	0.8666667	0
SVMLinear	Median	0.9000000	0
SVMLinear	Mean	0.9001505	0
SVMLinear	3rd Qu.	0.9333333	0
SVMLinear	Max.	1.0000000	0
LDA	Min.	0.8000000	3
LDA	1st Qu.	0.8688172	3
LDA	Median	0.9000000	3
LDA	Mean	0.9043112	3
LDA	3rd Qu.	0.9321839	3
LDA	Max.	1.0000000	3
RF	Min.	0.6951220	0
RF	1st Qu.	0.8168672	0
RF	Median	0.8795006	0
RF	Mean	0.8608708	0
RF	3rd Qu.	0.8857980	0
RF	Max.	1.0000000	0
kNN	Min.	0.4934498	0
kNN	1st Qu.	0.7016782	0
kNN	Median	0.7644495	0
kNN	Mean	0.7701866	0
kNN	3rd Qu.	0.8398392	0
kNN	Max.	0.9426987	0
SVMLinear	Min.	0.6250000	0
SVMLinear	1st Qu.	0.7401316	0
SVMLinear	Median	0.8307888	0
SVMLinear	Mean	0.8220801	0
SVMLinear	3rd Qu.	0.8853789	0
SVMLinear	Max.	1.0000000	0
LDA	Min.	0.6590909	3
LDA	1st Qu.	0.7722952	3
LDA	Median	0.8255814	3
LDA	Mean	0.8308044	3
LDA	3rd Qu.	0.8800595	3
LDA	Max.	1.0000000	3

2. **Variable Importance:**

Interpretable machine learning techniques identified key predictors for sleep disorders using the Random Forest model. Important variables included Sleep Duration, Quality of Sleep, among others. The influence of these variables on predictions was illustrated through Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots, providing a clear visualization of their impact on the outcome.

3. **Insights from the Model:**

The models revealed significant insights, such as the relationship between stress levels, physical activity level and sleep disorders. These findings enhance the understanding of factors influencing sleep health and offer actionable insights.

Discussion

Conclusions

From the comprehensive analysis using machine learning models, several key conclusions emerge:

1. **Predictive Power of Variables:** The study successfully identified crucial variables affecting sleep disorders. This highlights how factors like Sleep Duration and Stress Level play a significant role in sleep health.
2. **Model Effectiveness:** Among the models, Random Forest stood out for its predictive accuracy. This underlines the model's capability to handle complex datasets with multiple predictors.
3. **Practical Implications:** The findings provide actionable insights into sleep health, potentially guiding interventions or further research in sleep disorder management.

Limitations

Despite the analysis's thoroughness, the following limitations must be acknowledged:

1. **Data Constraints:** The dataset's scope, in terms of diversity and size, may limit the generalization of the findings. The representatives of the sample is crucial for broader applicability.
2. **Model Limitations:** While Random Forest performed well, its complex nature and potential for over fitting should be considered. The interpretation of such models also presents challenges.
3. **Methodological Boundaries:** The reliance on specific statistical techniques and machine learning models might have led to an oversight of more nuanced or intricate relationships within the dataset.

Future Expansion & Recommendations

Given more resources and time, the analysis could be expanded in the following ways:

1. **Incorporating Additional Data Sources:** Including more diverse and extensive datasets could enhance the robustness and applicability of the findings.
2. **Exploring Alternative Models:** Employing different machine learning approaches might reveal additional insights or validate the current findings.
3. **Deeper Feature Engineering:** Delving deeper into feature engineering and selection could uncover subtler patterns and relationships within the data.

Project Success

Reflecting on the project's goals as outlined in the proposal:

- The primary objective of identifying key predictors of sleep disorders and assessing the effectiveness of various models was achieved.
- However, the project faced constraints in data diversity and model complexity, which may have impacted the depth of the findings.
- Overall, the project succeeded in providing valuable insights into sleep disorders, although with the mentioned limitations and potential areas for further exploration.

References

[1] S. L. Halson et al., “Sleep Regularity and Predictors of Sleep Efficiency and Sleep Duration in Elite Team Sport Athletes,” Sports Medicine - Open, vol. 8, no. 79, 2022.

[2] R. Smith et al., “Sleep Patterns and Disorders Among Long-Distance HGV Drivers: A Concern for Road Safety,” Transportation Research Part F: Traffic Psychology and Behaviour, vol. 77, pp. 1-14, 2021.

[3] R. Smith et al., “Occupational Factors Affecting Sleep Health of HGV Drivers,” Safety and Health at Work, vol. 12, no. 4, pp. 500-507, 2021.

[4] S. L. Halson et al., “The Impact of Training and Competition on Sleep Patterns of Elite Athletes,” Sports Medicine - Open, vol. 8, no. 79, 2022.