Analyzing Sleep Disorders and Factors Affecting Sleep Quality Using Decision Tree and Random Forest

Capstone Project Report

Instructor: Dr. Rouzbeh Razavi

Authored by: Shreya Thodupunuri

Student Id:811301506

Contents

1. Introduction

Sleep is essential for human health and well-being, impacting cognitive function, emotional stability, physical health, and overall quality of life. Poor sleep quality and sleep disorders are linked to various health issues such as cardiovascular disease, obesity, diabetes, depression, and impaired cognitive function. Understanding the factors that influence sleep quality and identifying sleep disorders are crucial for developing effective interventions to enhance sleep health and overall wellness.

This project aims to analyze a comprehensive dataset on sleep health and lifestyle using advanced machine learning techniques. Decision Tree and Random Forest algorithms will be used to identify key factors that affect sleep quality and to predict the occurrence of sleep disorders. These machine learning models are well-suited for this task due to their ability to handle complex, non-linear relationships and interactions between variables.

The dataset includes demographic information, lifestyle habits, health conditions, and sleep patterns such as age, gender, occupation, sleep duration, physical activity level, stress level, body mass index (BMI), blood pressure, heart rate, daily steps, and presence of sleep disorders such as insomnia and sleep apnea.

2. Problem Statement

This project aims to understand the factors influencing sleep quality and predict sleep disorders. Specifically, we want to understand which variables significantly affect sleep quality and how well those variables can predict sleep disorders using machine learning models.

3. Business Goal

The goal is to model sleep quality and predict sleep disorders using the available independent variables, to understand how these variables affect sleep quality, and to identify potential sleep disorders, guiding interventions to improve sleep health.

4. Literature Review

Several studies have explored the factors affecting sleep quality and the prediction of sleep disorders. Techniques such as multiple linear regression, decision trees, and random forests have been used, with varying degrees of accuracy. Key factors often include age, stress level, physical activity, and health conditions.

5. Overview of the Dataset

The dataset comprises 374 records with various attributes including age, gender, lifestyle habits, health conditions, and sleep patterns. The data was obtained from Kaggle and covers information on sleep health and lifestyle. Key attributes in the dataset include age, gender, occupation, sleep duration, quality of sleep, physical activity level, stress level, BMI category, blood pressure, heart rate, daily steps, and sleep disorder.

- Total records: 374

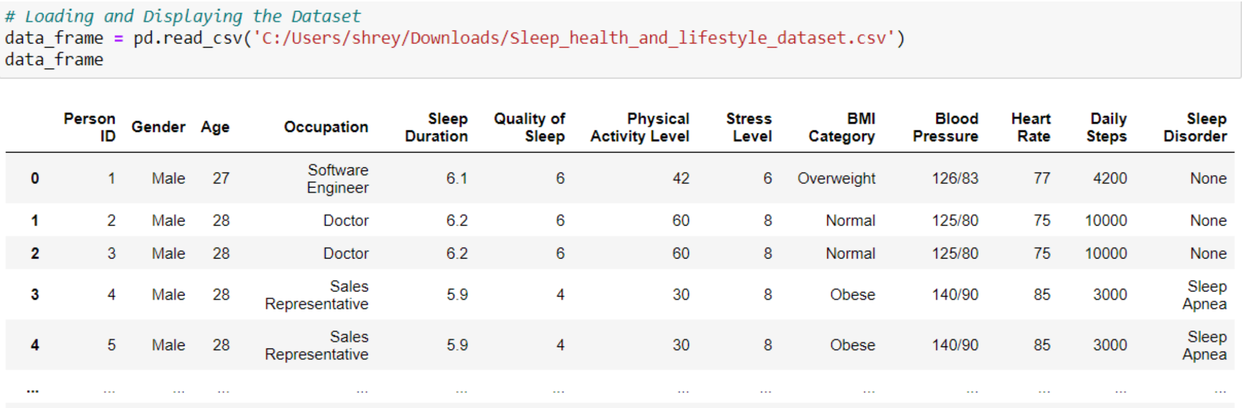
- Attributes: 13 (Age, Gender, Occupation, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, BMI Category, Blood Pressure, Heart Rate, Daily Steps, Sleep Disorder)

6. Data Cleaning:

Data cleaning involves handling missing values, eliminating duplicates, and removing irrelevant columns. The data was validated for missing values, and it was confirmed that there were no missing values or duplicates.

The total number of records is 374

Firstly, the data was loaded into a data frame named ‘data_frame’.

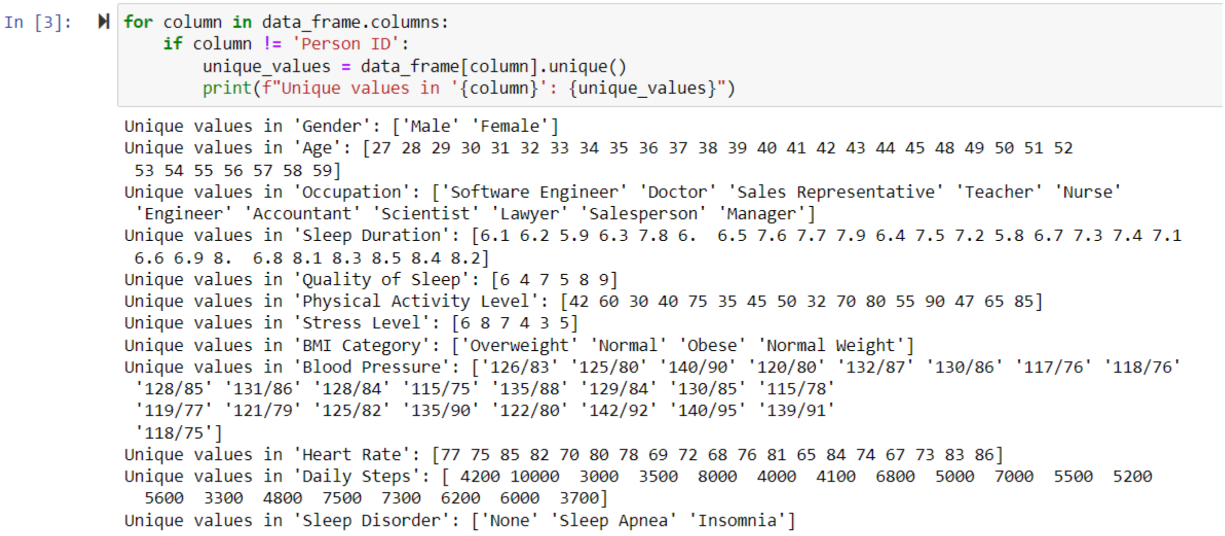


```
# Loading and Displaying the Dataset
data_frame = pd.read_csv('C:/Users/shrey/Downloads/Sleep_health_and_lifestyle_dataset.csv')
data_frame
```

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Unique Values Identification:

The initial step in data analysis involves understanding the dataset's structure by examining the unique values in each column. This process is essential for identifying potential anomalies, and data consistency issues, and ensuring data quality.

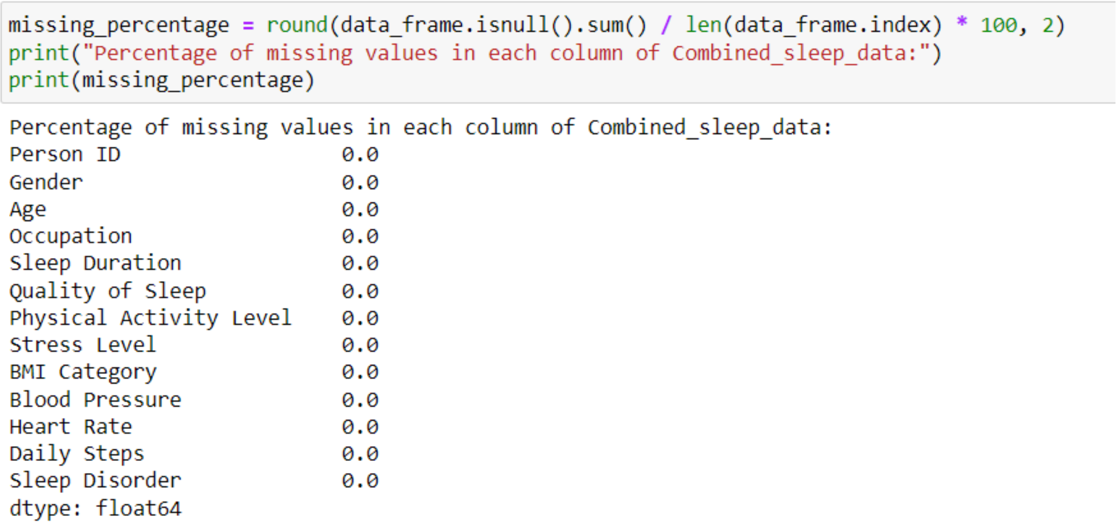


```
In [3]:  for column in data_frame.columns:
             if column != 'Person ID':
                 unique_values = data_frame[column].unique()
                 print(f"Unique values in '{column}': {unique_values}")
```

```
Unique values in 'Gender': ['Male' 'Female']
Unique values in 'Age': [27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 48 49 50 51 52
 53 54 55 56 57 58 59]
Unique values in 'Occupation': ['Software Engineer' 'Doctor' 'Sales Representative' 'Teacher' 'Nurse'
 'Engineer' 'Accountant' 'Scientist' 'Lawyer' 'Salesperson' 'Manager']
Unique values in 'Sleep Duration': [6.1 6.2 5.9 6.3 7.8 6.  6.5 7.6 7.7 7.9 6.4 7.5 7.2 5.8 6.7 7.3 7.4 7.1
 6.6 6.9 8.  6.8 8.1 8.3 8.5 8.4 8.2]
Unique values in 'Quality of Sleep': [6 4 7 5 8 9]
Unique values in 'Physical Activity Level': [42 60 30 40 75 35 45 50 32 70 80 55 90 47 65 85]
Unique values in 'Stress Level': [6 8 7 4 3 5]
Unique values in 'BMI Category': ['Overweight' 'Normal' 'Obese' 'Normal Weight']
Unique values in 'Blood Pressure': ['126/83' '125/80' '140/90' '120/80' '132/87' '130/86' '117/76' '118/76'
 '128/85' '131/86' '128/84' '115/75' '135/88' '129/84' '130/85' '115/78'
 '119/77' '121/79' '125/82' '135/90' '122/80' '142/92' '140/95' '139/91'
 '118/75']
Unique values in 'Heart Rate': [77 75 85 82 70 80 78 69 72 68 76 81 65 84 74 67 73 83 86]
Unique values in 'Daily Steps': [ 4200 10000  3000  3500  8000  4000  4100  6800  5000  7000  5500  5200
  5600  3300  4800  7500  7300  6200  6000  3700]
Unique values in 'Sleep Disorder': ['None' 'Sleep Apnea' 'Insomnia']
```

Here, we can observe that there are no duplicate values, and there are no issues with the data, except that Blood Pressure should be split into two columns as systolic blood pressure and diastolic blood pressure.

Missing Values Identification:

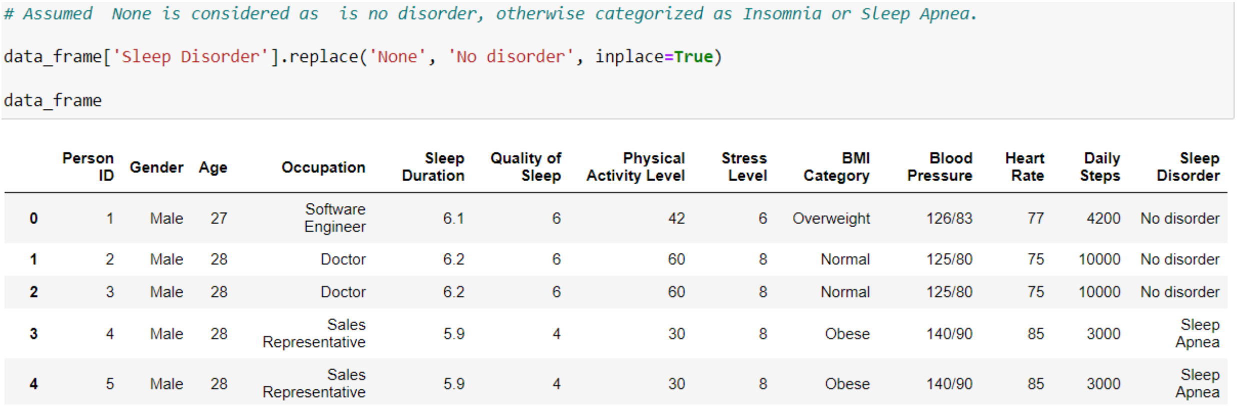


```
missing_percentage = round(data_frame.isnull().sum() / len(data_frame.index) * 100, 2)
print("Percentage of missing values in each column of Combined_sleep_data:")
print(missing_percentage)
```

```
Percentage of missing values in each column of Combined_sleep_data:
Person ID                  0.0
Gender                     0.0
Age                        0.0
Occupation                 0.0
Sleep Duration             0.0
Quality of Sleep           0.0
Physical Activity Level    0.0
Stress Level               0.0
BMI Category               0.0
Blood Pressure             0.0
Heart Rate                 0.0
Daily Steps                0.0
Sleep Disorder             0.0
dtype: float64
```

There are no missing values in the data.

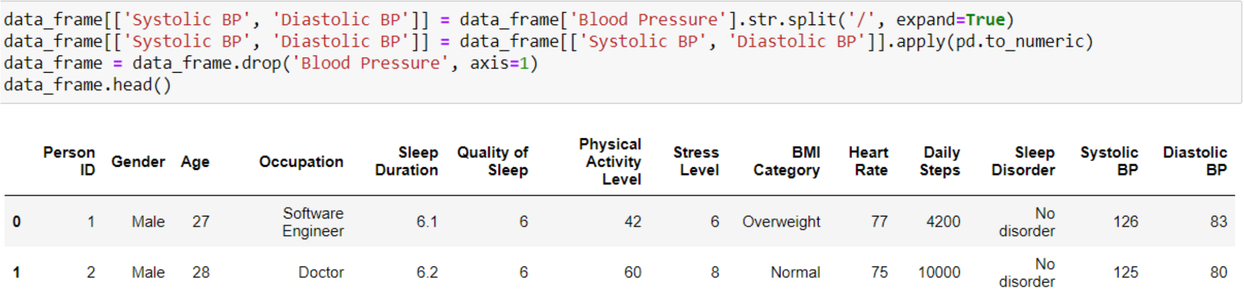
Fixing Data:



```
# Assumed  None is considered as  is no disorder, otherwise categorized as Insomnia or Sleep Apnea.

data_frame['Sleep Disorder'].replace('None', 'No disorder', inplace=True)

data_frame
```

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	No disorder
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	No disorder
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	No disorder
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea

In the sleep disorder category, ‘none’ is replaced with ‘no disorder’ for better clarity. This change allows us to easily identify when a person does not have any sleep disorder, as it will be labeled as ‘no disorder’ instead of ‘none.’



```
data_frame[['Systolic BP', 'Diastolic BP']] = data_frame['Blood Pressure'].str.split('/', expand=True)
data_frame[['Systolic BP', 'Diastolic BP']] = data_frame[['Systolic BP', 'Diastolic BP']].apply(pd.to_numeric)
data_frame = data_frame.drop('Blood Pressure', axis=1)
data_frame.head()
```

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Daily Steps	Sleep Disorder	Systolic BP	Diastolic BP
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	77	4200	No disorder	126	83
1	2	Male	28	Doctor	6.2	6	60	8	Normal	75	10000	No disorder	125	80

Now, the Blood Pressure column is split into 2 columns ‘Systolic BP’ and ‘Diastolic BP’ for further analysis.

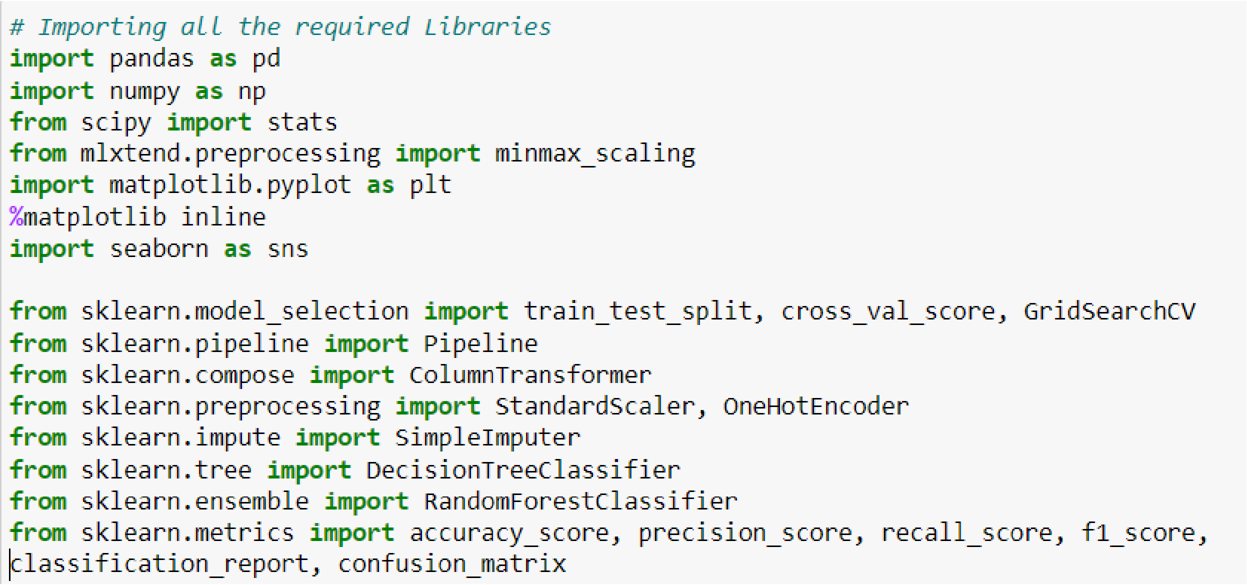
Systolic and diastolic blood pressure readings provide different clinical information.

Systolic blood pressure (the higher number) measures the pressure in your arteries when your heart beats. Diastolic blood pressure (the lower number) measures the pressure in your arteries when your heart rests between beats.

Benefit: Splitting allows for a more detailed analysis of blood pressure data, assisting in the identification of specific patterns or correlations related to each component.

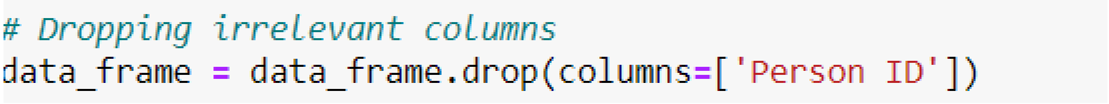
7. Data Preparation and Exploration

Loading all the required Libraries:



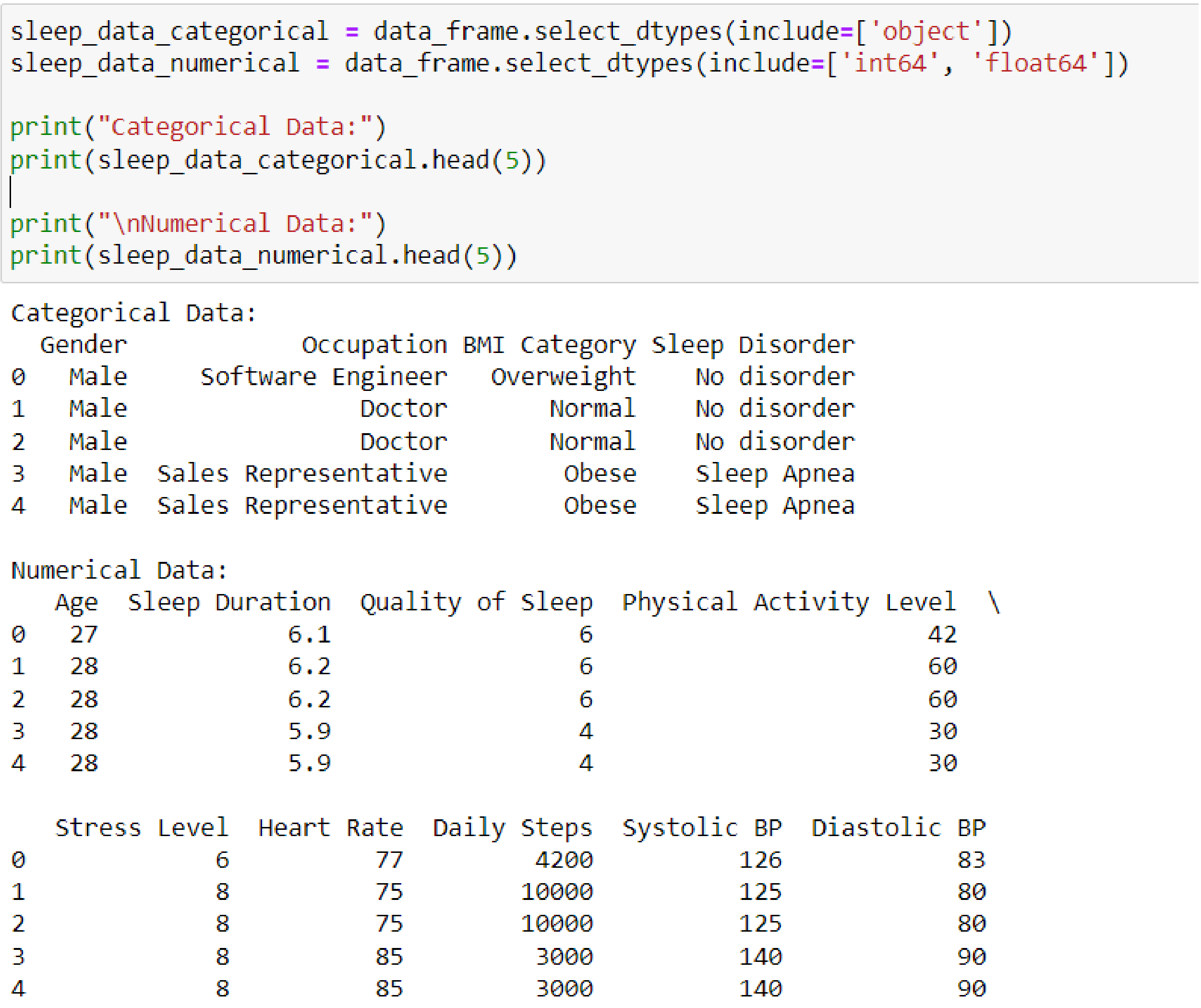
```
# Importing all the required Libraries
import pandas as pd
import numpy as np
from scipy import stats
from mlxtend.preprocessing import minmax_scaling
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
classification_report, confusion_matrix
```

Dropping Irrelevant columns:



```
# Dropping irrelevant columns
data_frame = data_frame.drop(columns=['Person ID'])
```

‘Person ID’ column is dropped as it is not required.



```
sleep_data_categorical = data_frame.select_dtypes(include=['object'])
sleep_data_numerical = data_frame.select_dtypes(include=['int64', 'float64'])

print("Categorical Data:")
print(sleep_data_categorical.head(5))

print("\nNumerical Data:")
print(sleep_data_numerical.head(5))
```

```
Categorical Data:
  Gender            Occupation BMI Category Sleep Disorder
0   Male     Software Engineer   Overweight    No disorder
1   Male                Doctor       Normal    No disorder
2   Male                Doctor       Normal    No disorder
3   Male  Sales Representative        Obese    Sleep Apnea
4   Male  Sales Representative        Obese    Sleep Apnea

Numerical Data:
   Age  Sleep Duration  Quality of Sleep  Physical Activity Level  \
0   27             6.1                 6                       42   
1   28             6.2                 6                       60   
2   28             6.2                 6                       60   
3   28             5.9                 4                       30   
4   28             5.9                 4                       30   

   Stress Level  Heart Rate  Daily Steps  Systolic BP  Diastolic BP  
0             6          77         4200          126            83  
1             8          75        10000          125            80  
2             8          75        10000          125            80  
3             8          85         3000          140            90  
4             8          85         3000          140            90  
```

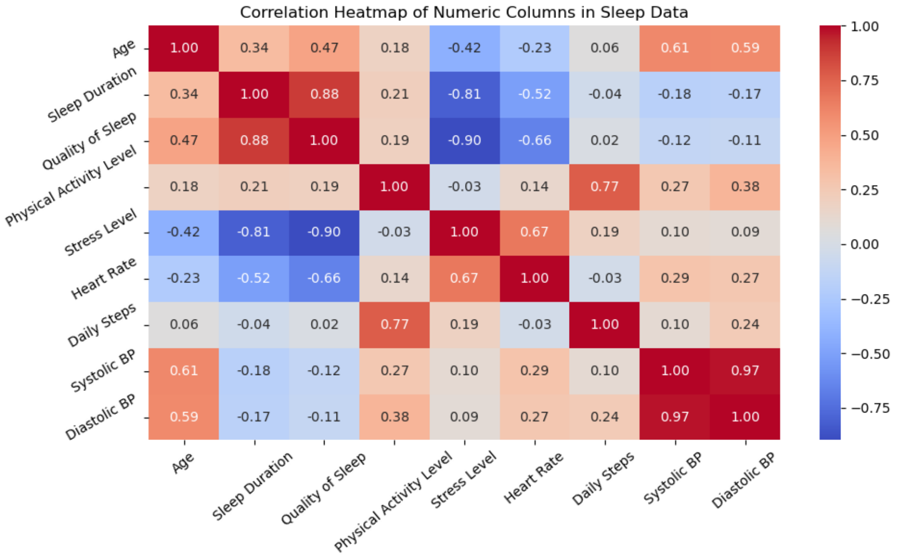
In this step, the data is segregated separately into Categorical Data and Numerical Data.

- **Categorical Data:** This includes qualitative data such as 'Gender', 'Occupation', 'BMI Category', and 'Sleep Disorder'.

- **Numerical Data:** This includes quantitative data like 'Age', 'Sleep Duration', 'Quality of Sleep', 'Physical Activity Level', 'Stress Level', 'Heart Rate', 'Daily Steps', 'Systolic Blood Pressure', and 'Diastolic Blood Pressure'.

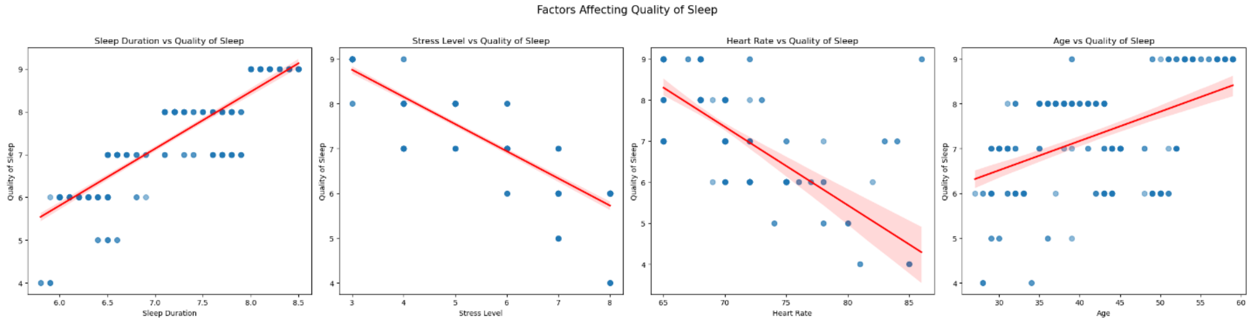
Analysis through Heatmap:

The quality of sleep is influenced by several significant factors, including sleep duration, stress level, heart rate, and age. These variables have strong correlations with sleep quality, positively or negatively impacting a person's ability to achieve restful and high-quality sleep. To improve sleep quality, it is essential to effectively manage sleep duration and stress, monitor heart rate, and consider age-related aspects.



- Quality of Sleep is highly positively correlated with Sleep Duration (0.88), indicating that longer sleep duration tends to improve sleep quality.
- Age is moderately correlated with Quality of sleep (0.47), especially in females.
- There is a strong negative correlation with Stress Levels (-0.90), suggesting that higher stress levels significantly reduce sleep quality.
- Quality of Sleep is negatively correlated with Heart Rate (-0.66), indicating that higher heart rates are associated with lower sleep quality.

Scatter Plots:



1. Sleep Duration vs. Quality of Sleep:

There is a strong positive correlation between sleep duration and quality of sleep. As sleep duration increases, the quality of sleep tends to improve significantly. Longer sleep duration is associated with higher quality of sleep.

2. Stress Level and Quality of Sleep:

There is a strong negative correlation between stress level and quality of sleep. Higher stress levels are associated with poorer quality of sleep. Reducing stress can likely improve sleep quality.

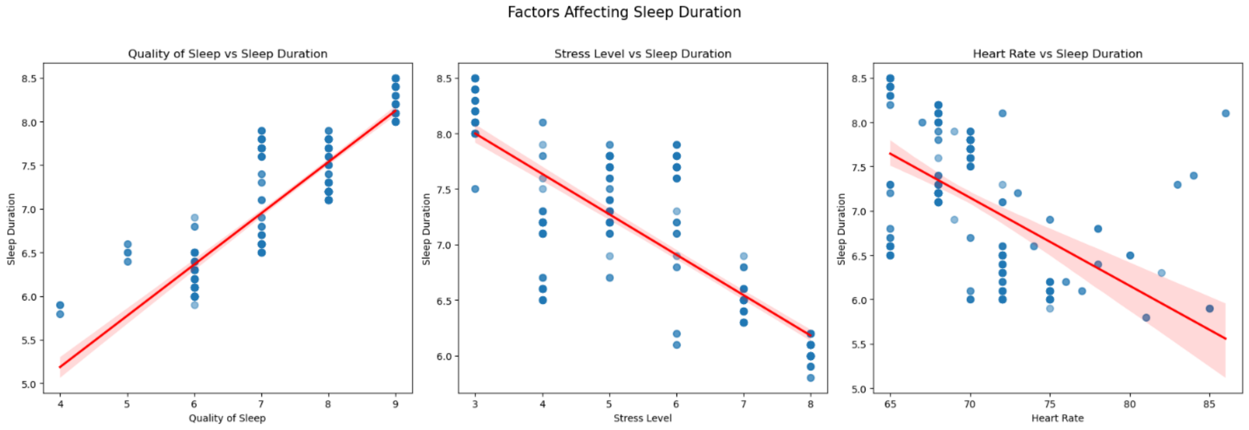
3. Heart Rate and Quality of Sleep:

There is a negative correlation between heart rate and quality of sleep. Higher heart rates tend to be associated with lower quality of sleep. Lower heart rates are generally linked to better sleep quality.

4. Age and Quality of Sleep:

There is a moderate positive correlation between age and quality of sleep. As age increases, the quality of sleep tends to improve. Older individuals in the dataset generally report better sleep quality.

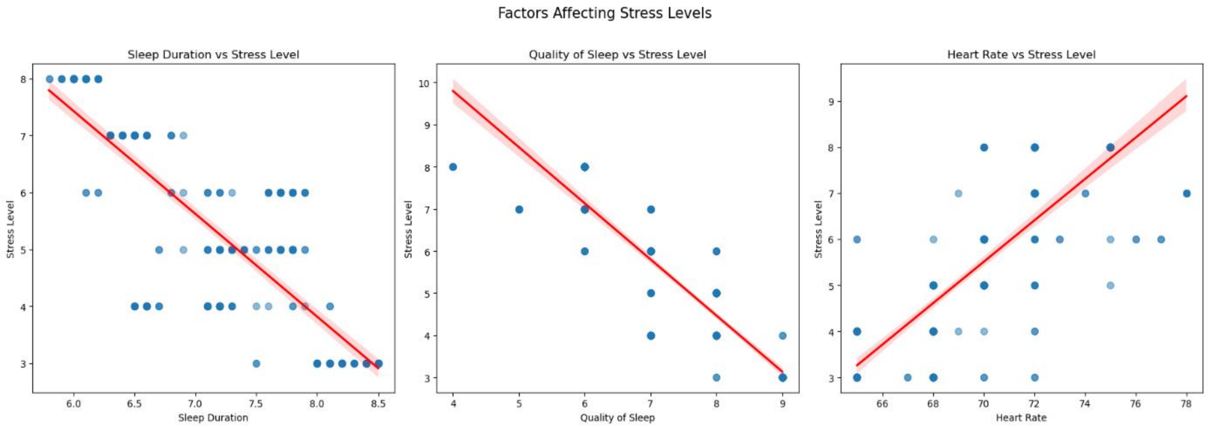
Factors Affecting Sleep Duration:



- There is a strong positive correlation between quality of sleep and sleep duration. Higher quality of sleep is associated with longer sleep durations.
- A strong negative correlation exists between stress level and sleep duration. Higher stress levels are linked to shorter sleep durations.

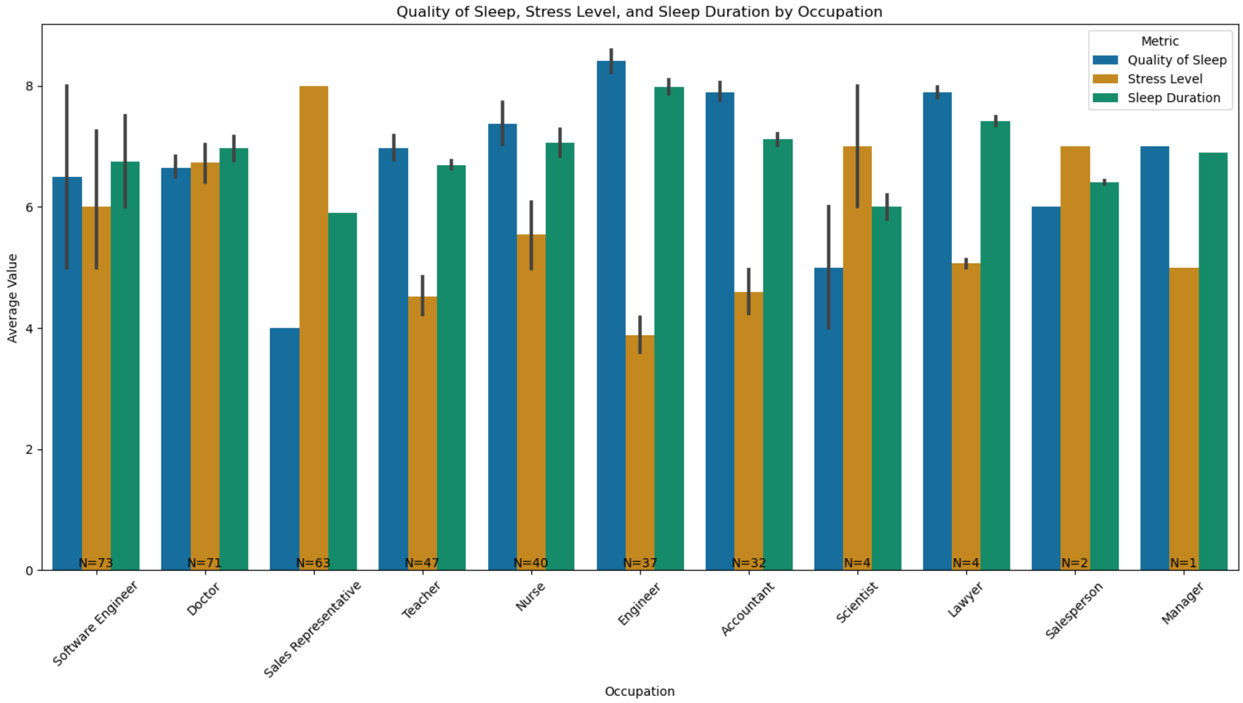
- There is a negative correlation between heart rate and sleep duration. Higher heart rates tend to be associated with shorter sleep durations.

Factors Affecting Stress Levels:



- There is an inverse relationship between sleep duration and stress levels. As sleep duration increases, stress levels decrease.
- Similarly, there is an inverse relationship between sleep quality and stress levels. Higher-quality sleep is linked to lower stress levels.
- Conversely, there is a positive correlation between heart rate and stress levels. Higher heart rates are associated with higher stress levels.

Quality of Sleep, Stress Levels, and Sleep Duration by Occupation:

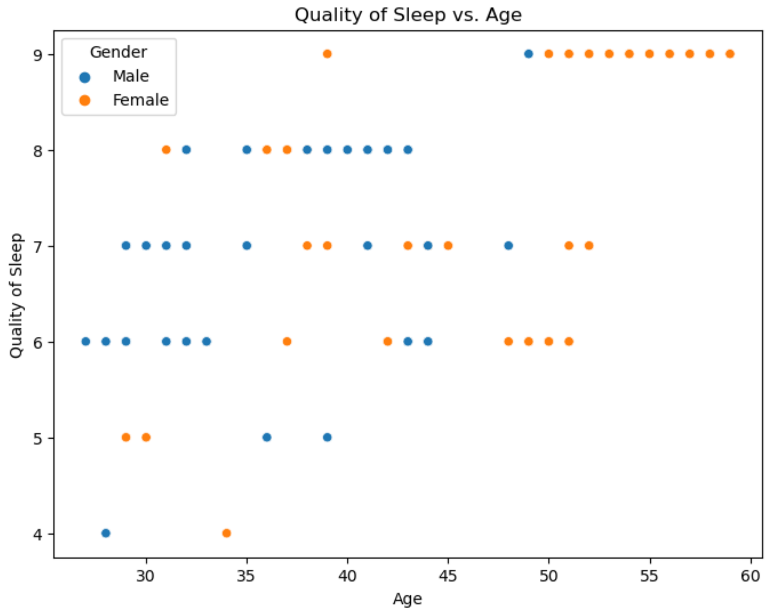


Adequate sleep duration leads to reduced stress levels, which positively impacts sleep quality, especially for professionals such as teachers, engineers, and accountants.

There is a significant gap between sleep duration and stress level, which negatively impacts sleep quality, particularly affecting occupations like sales representatives and scientists. Addressing this gap is crucial for improving overall sleep health and reducing stress-related issues.

Occupations such as doctors and software engineers tend to have a relatively balanced combination of stress levels and sleep quality.

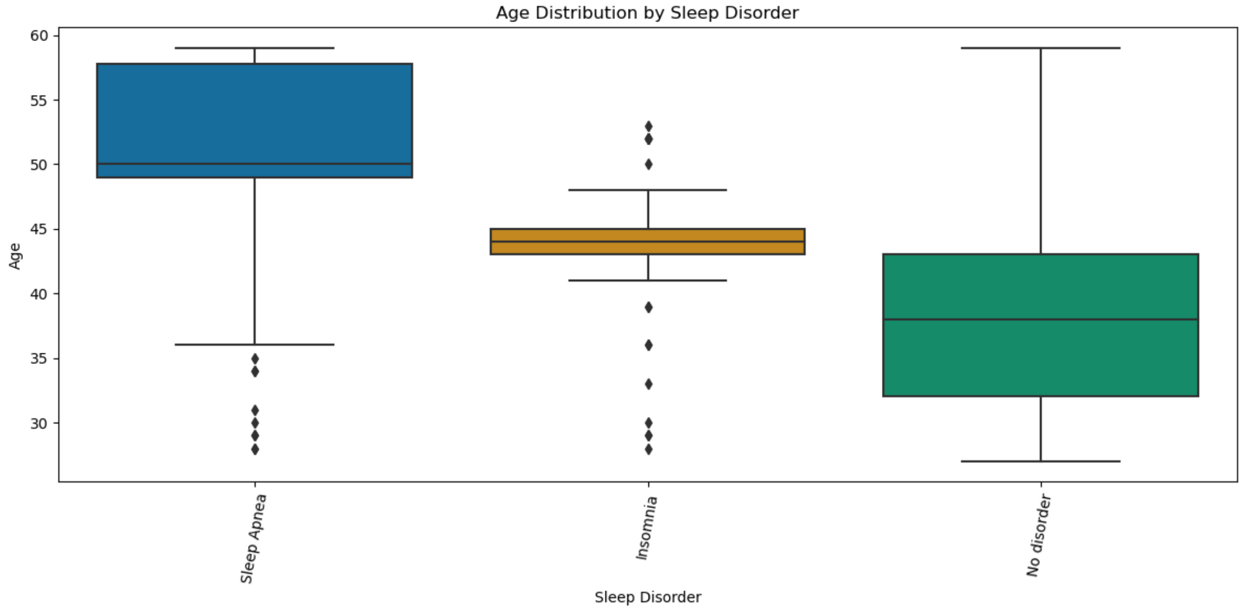
Quality of Sleep by Age:



Middle Age Groups (40-50 years): Both men and women typically have moderately to highly good sleep quality, with scores ranging from 6 to 9. Most people in this age range have sleep quality scores between 6 and 8, regardless of gender.

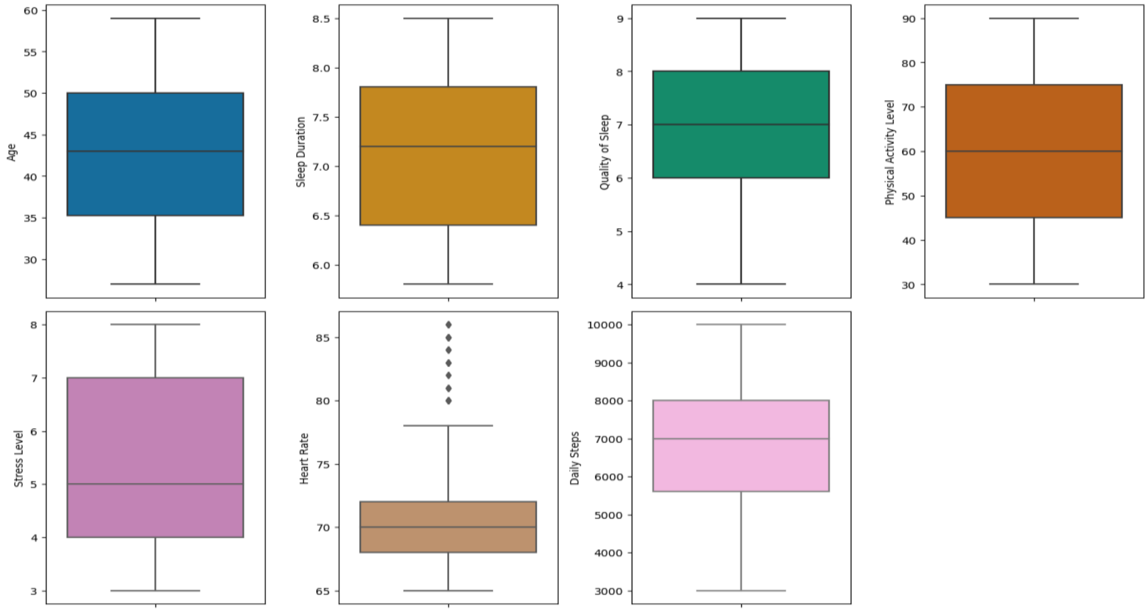
Older Age Groups (50-60 years): Sleep quality is generally high in this age range, with most scores falling between 7 and 9. Women in this group, in particular, tend to have higher sleep quality scores.

Age Distribution by Sleep Disorder:



Sleep apnea mainly impacts individuals aged 30 to 60, with a median age of about 50, making it more prevalent in older adults. Insomnia primarily affects individuals aged 40 to 50, with a median age of around 45, indicating it is more common in the middle-aged population.

Numerical Data Visualization:



The box plots above summarize the distribution of various factors that affect sleep quality:

1. Age: The median age is around 45 years, with a range from approximately 30 to 60 years.

2. Sleep Duration: The median sleep duration is around 7 hours, with most values between 6 and 8 hours.

3. Quality of Sleep: The median quality of sleep is around 7, with values ranging from 4 to 9.

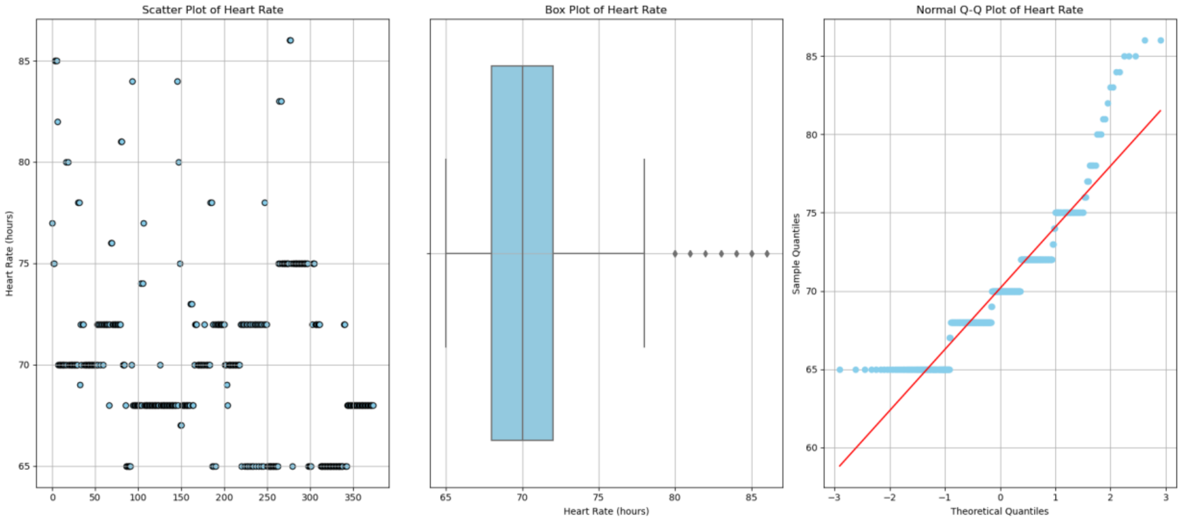
4. Physical Activity Level: The median physical activity level is around 60, with values ranging from 30 to 90.

5. Stress Level: The median stress level is around 5, with values ranging from 3 to 8.

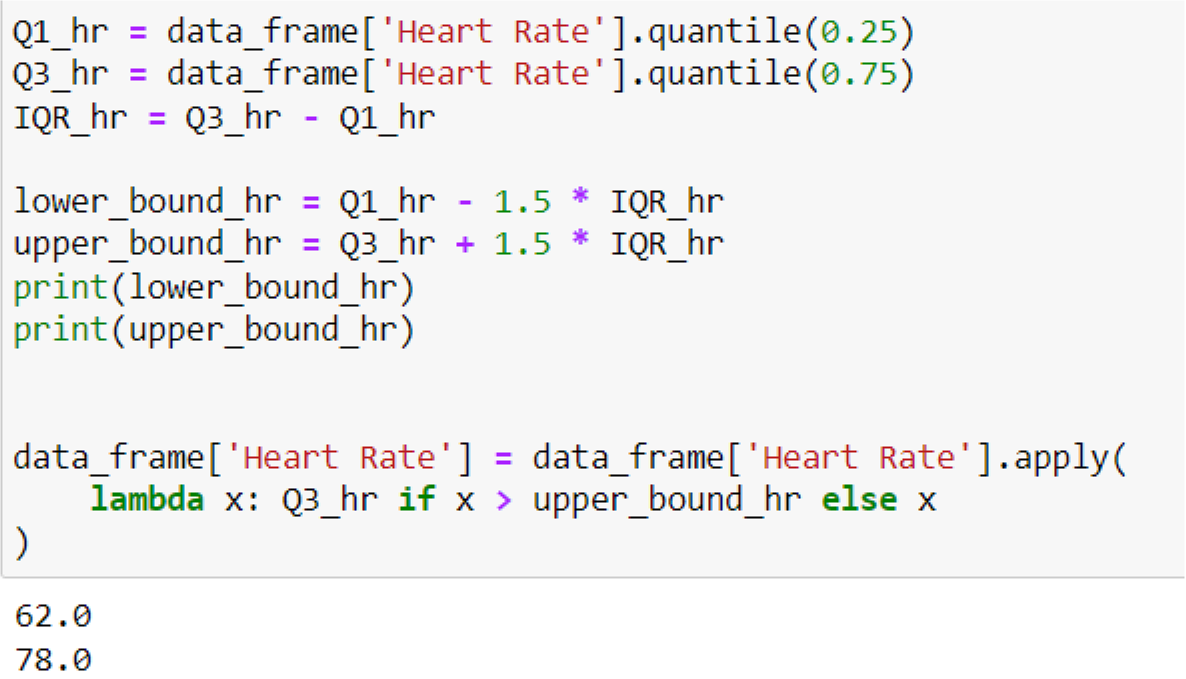
6. Heart Rate: The median heart rate is around 70 bpm, with outliers above 80 bpm.

7. Daily Steps: The median daily steps are around 6000, with values ranging from 3000 to 10000.

There are a few outliers in heart rate.



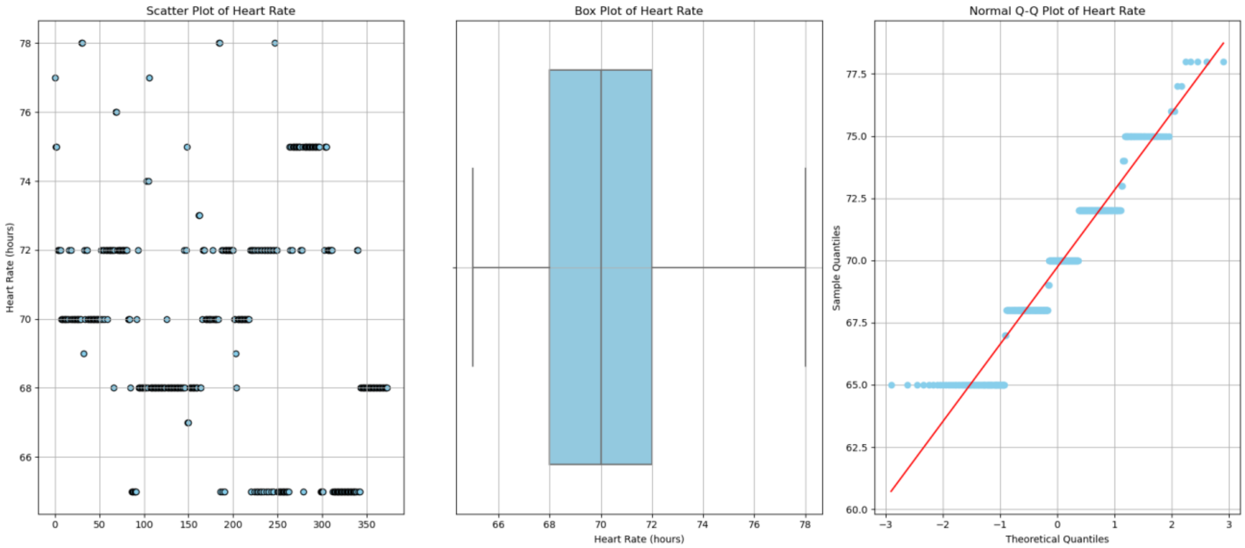
In the above plots, it is evident that a few data points are away from the normal data. So, to handle it Inter Quartile Range is being implemented.



```
Q1_hr = data_frame['Heart Rate'].quantile(0.25)
Q3_hr = data_frame['Heart Rate'].quantile(0.75)
IQR_hr = Q3_hr - Q1_hr

lower_bound_hr = Q1_hr - 1.5 * IQR_hr
upper_bound_hr = Q3_hr + 1.5 * IQR_hr
print(lower_bound_hr)
print(upper_bound_hr)


data_frame['Heart Rate'] = data_frame['Heart Rate'].apply(
    lambda x: Q3_hr if x > upper_bound_hr else x
)
```

```
62.0
78.0
```

After applying IQR, the plots are as below:



7. Modeling Approach

Quality of Sleep Classification:

Classification of 'Quality of Sleep' and Influencing Factors

To analyze the 'Quality of Sleep' effectively, a classification approach was implemented, transforming the subjective sleep quality ratings into distinct categories. This method facilitates a clearer understanding of the factors influencing sleep quality and supports robust analytical models.

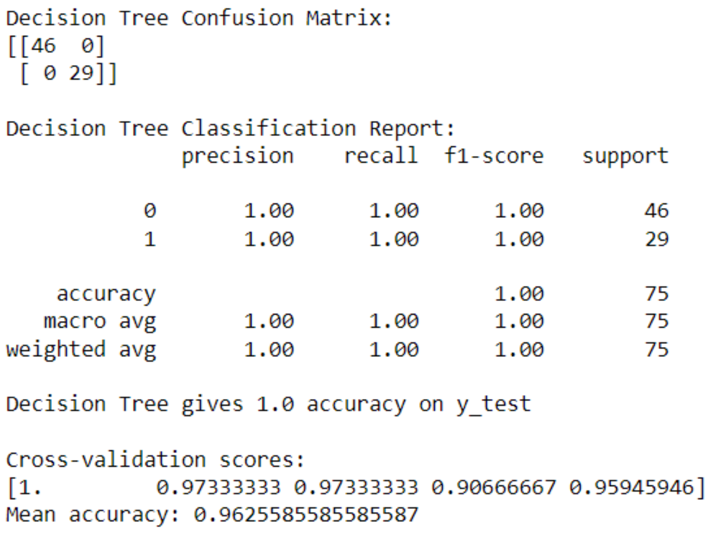
Classification Criteria for 'Quality of Sleep':

Low & Medium Quality: Scores: 1 to 6

Description: These scores indicate lower to moderate sleep quality, encompassing a range of issues from mild disturbances to moderate sleep deficiencies.

High Quality: Scores: Above 6

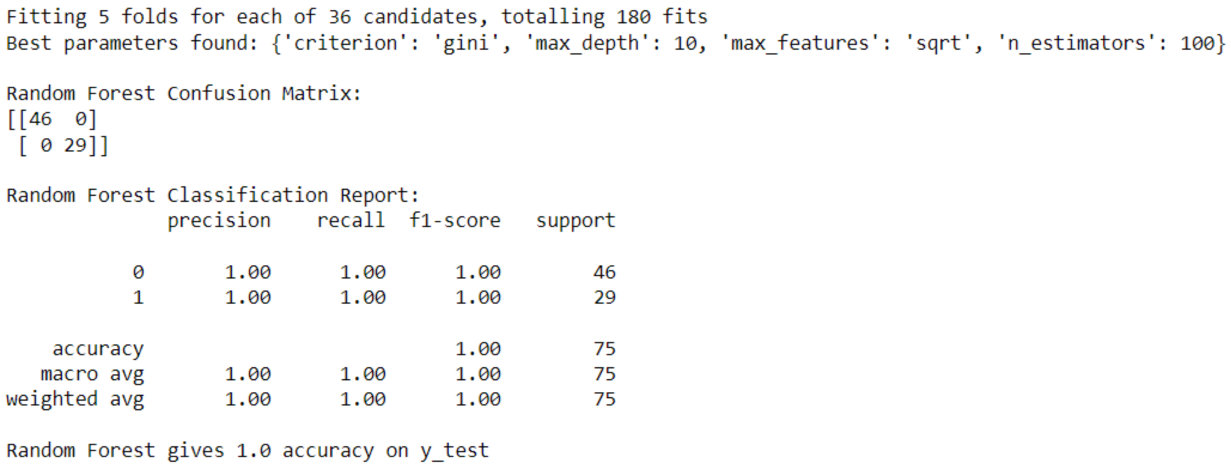
Description: Scores in this range represent high sleep quality, suggesting adequate rest and minimal sleep disturbances.



```
Decision Tree Confusion Matrix:
[[46  0]
 [ 0 29]]

Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        46
           1       1.00      1.00      1.00        29

    accuracy                           1.00        75
   macro avg       1.00      1.00      1.00        75
weighted avg       1.00      1.00      1.00        75

Decision Tree gives 1.0 accuracy on y_test

Cross-validation scores:
[1.         0.97333333 0.97333333 0.90666667 0.95945946]
Mean accuracy: 0.9625585585585587
```

The Decision Tree model performed exceptionally well on the test set, achieving a perfect classification with 100% accuracy, precision, recall, and F1-score for both classes. The confusion matrix revealed 46 true positives and 29 true negatives, with no false positives or false negatives. The cross-validation scores ranged from 0.9067 to 0.9733, resulting in a mean accuracy of 0.9626. This indicates consistently high performance across different data splits.

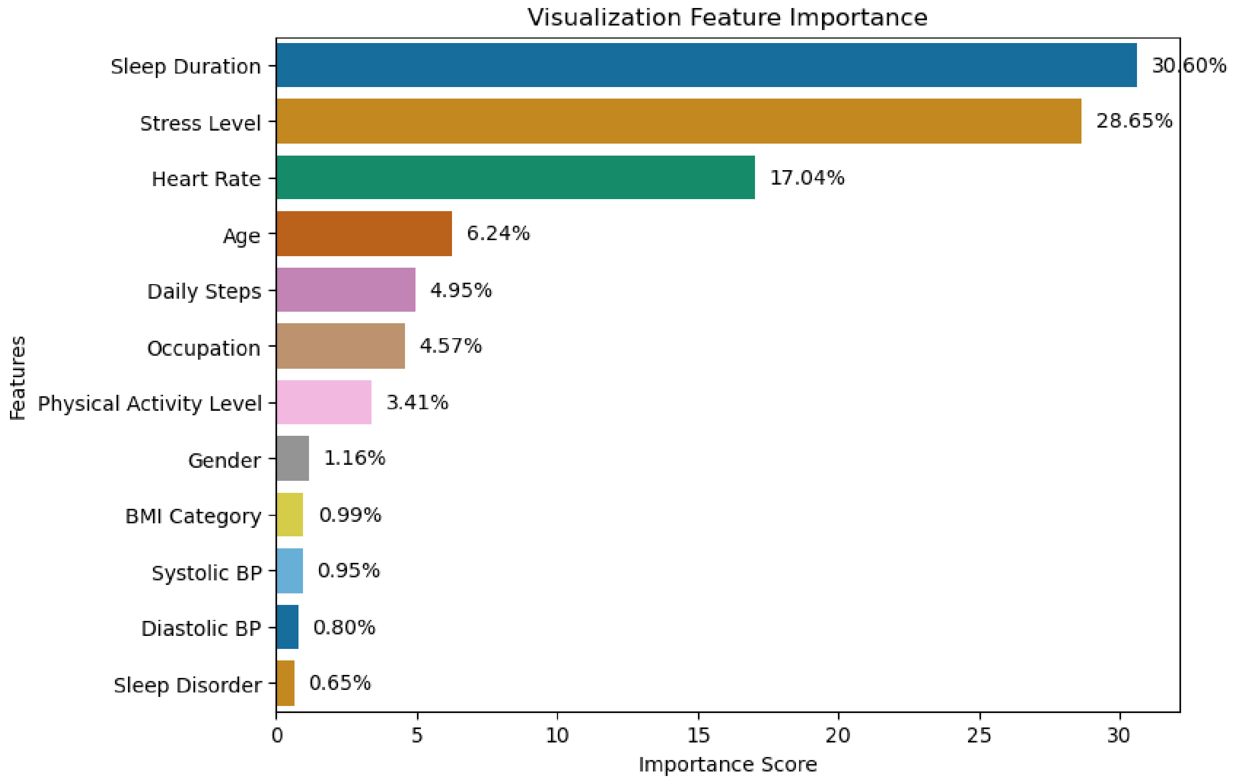
Random Forest Classification for Quality of Sleep:



```
Fitting 5 folds for each of 36 candidates, totalling 180 fits
Best parameters found: {'criterion': 'gini', 'max_depth': 10, 'max_features': 'sqrt', 'n_estimators': 100}

Random Forest Confusion Matrix:
[[46  0]
 [ 0 29]]

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        46
           1       1.00      1.00      1.00        29

    accuracy                           1.00        75
   macro avg       1.00      1.00      1.00        75
weighted avg       1.00      1.00      1.00        75

Random Forest gives 1.0 accuracy on y_test
```

The Random Forest model, after fitting 5 folds for each of the 36 candidates, identified the best parameters as {'criterion': 'gini', 'max_depth': 10, 'max_features': 'sqrt', 'n_estimators': 100}. The model achieved perfect classification on the test set with an accuracy, precision, recall, and F1-score of 1.00 for both classes. The confusion matrix shows 46 true positives and 29 true negatives, with no false positives or false negatives, indicating the model's high performance and accuracy.

Feature Importance of Quality of Sleep:



- The analysis of feature importance shows that the most significant factors affecting sleep quality are Sleep Duration (30.60%) and Stress Level (28.65%), emphasizing their critical roles.
- Heart Rate also plays a significant part with an importance score of 17.04%. Other noteworthy factors include Age (6.24%), Daily Steps (4.95%), and Occupation (4.57%).
- Less influential features are Physical Activity Level (3.41%), Gender (1.16%), BMI Category (0.99%), Systolic BP (0.95%), Diastolic BP (0.80%), and Sleep Disorder (0.65%).
- These findings highlight the importance of managing sleep duration and stress levels, as well as maintaining overall physical health, to improve sleep quality.

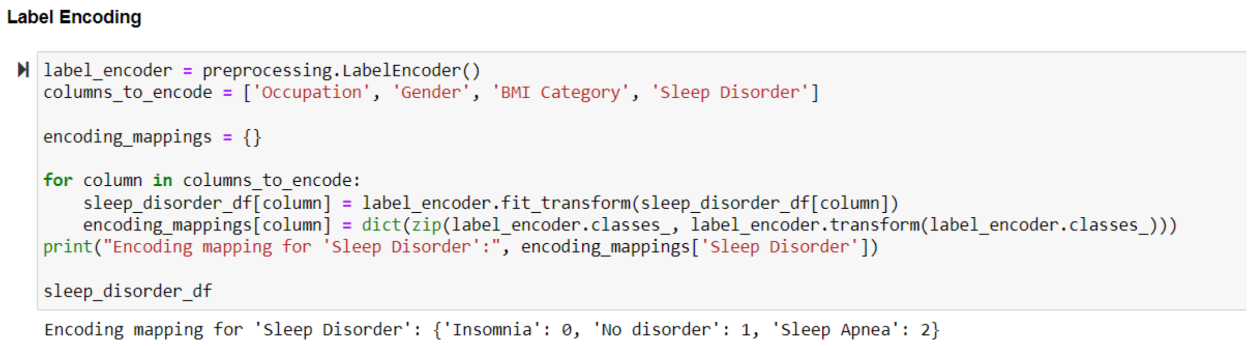
Sleep Disorders Classification:

Classification of 'Sleep Disorders' and Influencing Factors.

When assessing the presence or absence of sleep disorders in individuals, three distinct categories are considered:

No disorder, Insomnia, and Sleep Apnea.

To facilitate analysis, a label encoding classification approach was applied.

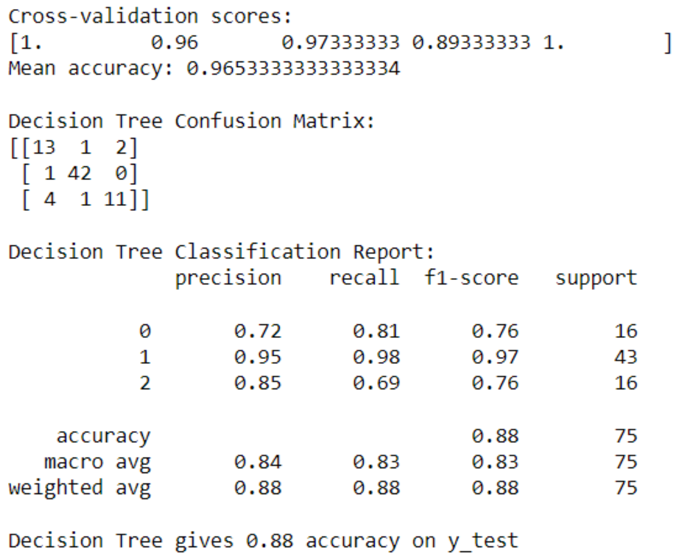


Label Encoding

```
label_encoder = preprocessing.LabelEncoder()
columns_to_encode = ['Occupation', 'Gender', 'BMI Category', 'Sleep Disorder']

encoding_mappings = {}

for column in columns_to_encode:
    sleep_disorder_df[column] = label_encoder.fit_transform(sleep_disorder_df[column])
    encoding_mappings[column] = dict(zip(label_encoder.classes_, label_encoder.transform(label_encoder.classes_)))
print("Encoding mapping for 'Sleep Disorder':", encoding_mappings['Sleep Disorder'])

sleep_disorder_df
```

```
Encoding mapping for 'Sleep Disorder': {'Insomnia': 0, 'No disorder': 1, 'Sleep Apnea': 2}
```

Prediction of Sleep Disorders using Decision Tree Classification:



```
Cross-validation scores:
[1.         0.96       0.97333333 0.89333333 1.        ]
Mean accuracy: 0.9653333333333334

Decision Tree Confusion Matrix:
[[13  1  2]
 [ 1 42  0]
 [ 4  1 11]]

Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.81      0.76        16
           1       0.95      0.98      0.97        43
           2       0.85      0.69      0.76        16

    accuracy                           0.88        75
   macro avg       0.84      0.83      0.83        75
weighted avg       0.88      0.88      0.88        75

Decision Tree gives 0.88 accuracy on y_test
```

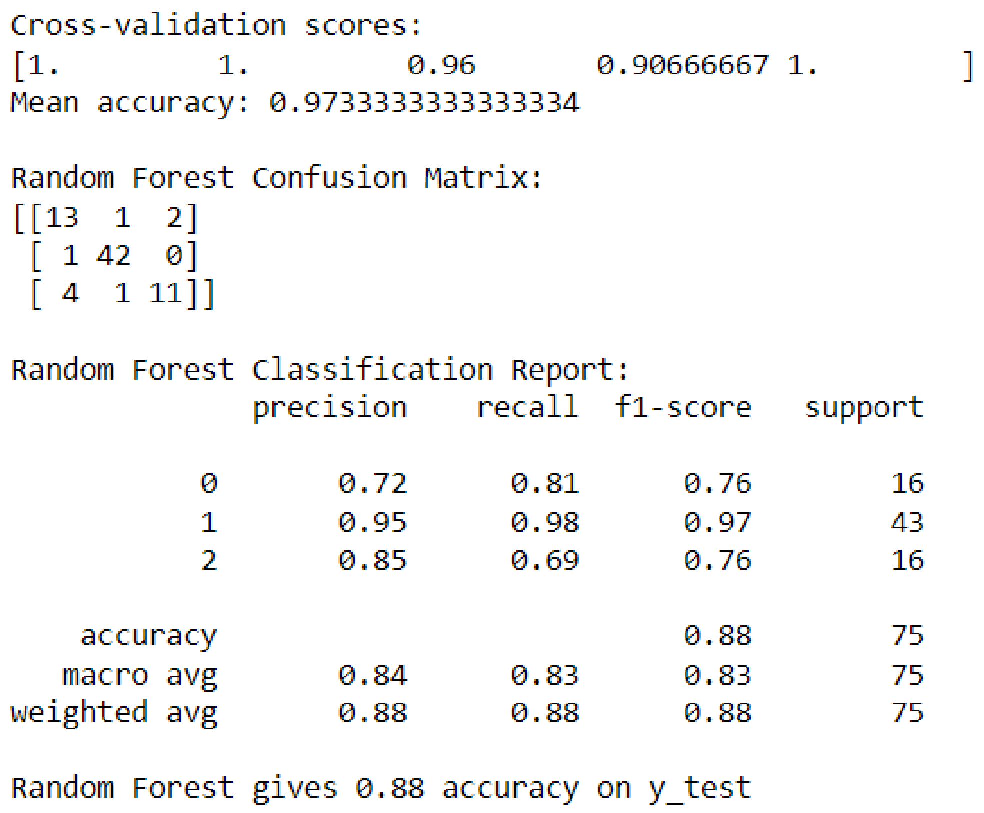
The Decision Tree model achieved an overall accuracy of 88% on the test data. The confusion matrix reveals that the model correctly classified 13 out of 16 instances for class 0, 42 out of 43 instances for class 1, and 11 out of 16 instances for class 2.

The precision scores for the model were 0.72 for class 0, 0.95 for class 1, and 0.85 for class 2. The recall values were 0.81 for class 0, 0.98 for class 1, and 0.69 for class 2.

Additionally, the F1-scores were 0.76 for class 0, 0.97 for class 1, and 0.76 for class 2. These values indicate a balanced performance across classes.

Furthermore, cross-validation results showed a mean accuracy of approximately 96.53%, suggesting the model's robustness and consistent performance across different data splits.

Prediction of Sleep Disorders using Random Forest Classification:



```
Cross-validation scores:
[1.         1.         0.96       0.90666667 1.        ]
Mean accuracy: 0.9733333333333334

Random Forest Confusion Matrix:
[[13  1  2]
 [ 1 42  0]
 [ 4  1 11]]

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.81      0.76        16
           1       0.95      0.98      0.97        43
           2       0.85      0.69      0.76        16

    accuracy                           0.88        75
   macro avg       0.84      0.83      0.83        75
weighted avg       0.88      0.88      0.88        75

Random Forest gives 0.88 accuracy on y_test
```

The Random Forest model achieved an 88% overall accuracy on the test data.

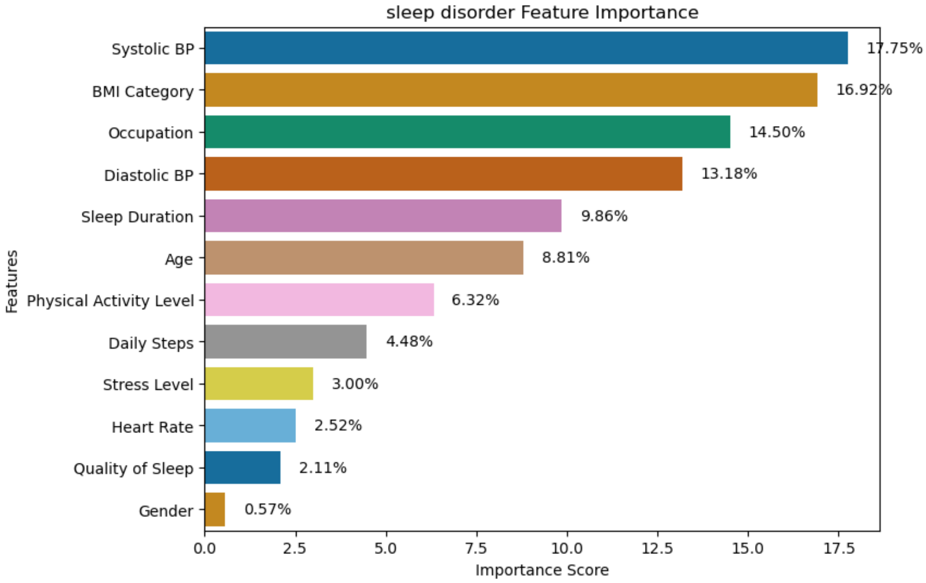
The confusion matrix indicates that the model accurately classified 13 out of 16 instances for class 0, 42 out of 43 instances for class 1, and 11 out of 16 instances for class 2.

The precision scores were 0.72 for class 0, 0.95 for class 1, and 0.85 for class 2. The recall scores were 0.81 for class 0, 0.98 for class 1, and 0.69 for class 2.

The F1-scores were 0.76 for class 0, 0.97 for class 1, and 0.76 for class 2, demonstrating a well-balanced performance across classes.

Furthermore, cross-validation results revealed a mean accuracy of approximately 97.33%, indicating the model's robustness and consistent performance across different data subsets.

Feature Importance for Sleep Disorders:



The feature importance analysis for predicting sleep disorders reveals that Systolic Blood Pressure (17.75%) is the most critical factor, followed closely by BMI Category (16.92%) and Occupation (14.50%).

Diastolic Blood Pressure also plays a significant role, accounting for 13.18% of the importance. Other notable factors include Sleep Duration (9.86%), Age (8.81%), and Physical Activity Level (6.32%).

Daily Steps contribute 4.48%, while Stress Level (3.00%) and Heart Rate (2.52%) have a relatively smaller impact.

Quality of Sleep (2.11%) and Gender (0.57%) are the least influential features.

These results underscore the importance of cardiovascular health indicators and body composition in the assessment of sleep disorders, with lifestyle factors also playing a substantial role.

8. Performance Evaluation

Decision Tree and Random Forest for Quality of Sleep

For the quality of sleep, both classifiers achieved a perfect accuracy of 1.0 on the test set. The Decision Tree and Random Forest showed mean cross-validation accuracies of 0.9626 and 0.9733, respectively. These results indicate robust performance and excellent generalization capability across different data splits. In particular, the Random Forest classifier demonstrated slightly better handling of data variability.

Decision Tree and Random Forest for Sleep Disorders:

In terms of Sleep Disorders, both the Decision Tree classifier and the Random Forest classifier achieved an accuracy of 0.88 on the test set. The Decision Tree classifier had a mean cross-validation accuracy of 0.9653, demonstrating high precision, recall, and f1-scores. On the other hand, the Random Forest classifier had a mean cross-validation accuracy of 0.9733, highlighting its strong predictive power and robustness.

9. Conclusion

Final Model Evaluation

The Random Forest model demonstrated superior performance in predicting both sleep quality and sleep disorders when compared to the Decision Tree model. It exhibited higher accuracy, precision, recall, and F1 scores.

Recommendations

-**Stress Management:** Implement stress management techniques to improve sleep quality.

- **Physical Activity:** Encourage regular physical activity to enhance sleep quality.

- **Heart Health:** Monitor and maintain a healthy heart rate to support better sleep quality.

- **Sleep Duration:** Ensure adequate sleep duration to achieve higher sleep quality.

By understanding these factors, interventions can be developed to promote better sleep health and overall well-being. This project demonstrates the effectiveness of machine learning models, particularly Random Forest, in analyzing and predicting sleep-related outcomes.