A REPORT ON SLEEP HEALTH AND LIFESTYLE

Purpose : The main objective of this project is to analyze, provide insights on sleep health and lifestyle of people, and to predict sleep disorders.

Introduction : Sleep is one important factor that affects the health and lifestyle of people, but is often neglected by a lot of people. In this project, my goal is to preprocess the data, analyze and understand different patterns among the dataset and the relationships between the predictive variables and to predict sleep disorder using multiple classification models.

Dataset :

The dataset contains data of approximately 400 rows and 13 columns, covering a wide range of features related to sleep and daily habits. The columns are

Columns	Description
Person ID	An identifier for each individual.
Gender	Gender of the individual
Age	Age of the individual
Occupation	The gender of the person (Male/Female).
Sleep Duration (hours)	The number of hours the person sleeps per day.
Quality of Sleep	A subjective rating of the quality of sleep, ranging from 1 to 10.
Physical Activity Level (minutes/day)	The number of minutes the person engages in physical activity daily.
Stress Level (scale: 1-10)	A subjective rating of the stress level experienced by the person, ranging from 1 to 10.

BMI Category	The BMI category of the person (e.g., Underweight, Normal, Overweight).
Blood Pressure (systolic/diastolic)	The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.
Heart Rate (bpm)	The resting heart rate of the person in beats per minute.
Daily Steps	The number of steps the person takes per day.
Sleep Disorder	The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).

Data Preprocessing :

I have dropped the Person ID column as it is merely a repetition of the indexing column

The column ‘Sleep Disorder’ has 219 null values that need to be dealt with. I’ve filled these null values with ‘None’ .

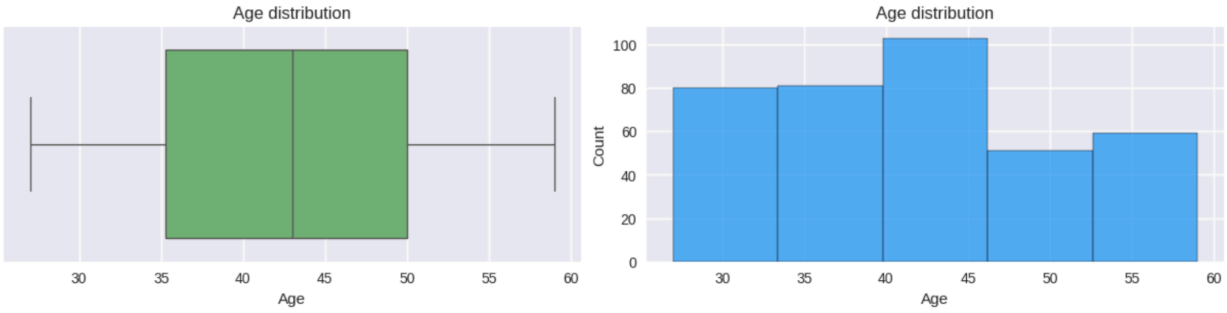
From the sample data I have observed that the ‘BMI Category’ column has values 'Overweight', 'Normal', 'Obese', 'Normal Weight'. ‘Normal’ and ‘Normal Weight’ both refer to the same. Hence, I have replaced ‘Normal Weight with Normal’

The dataset contains Blood Pressure columns which have values like 123/80, 124/79 etc. To easily process Blood Pressure I have created a new column ‘Blood Pressure Category’ and categorized ‘Blood Pressure’ values to Normal, Elevated and Hypertension.

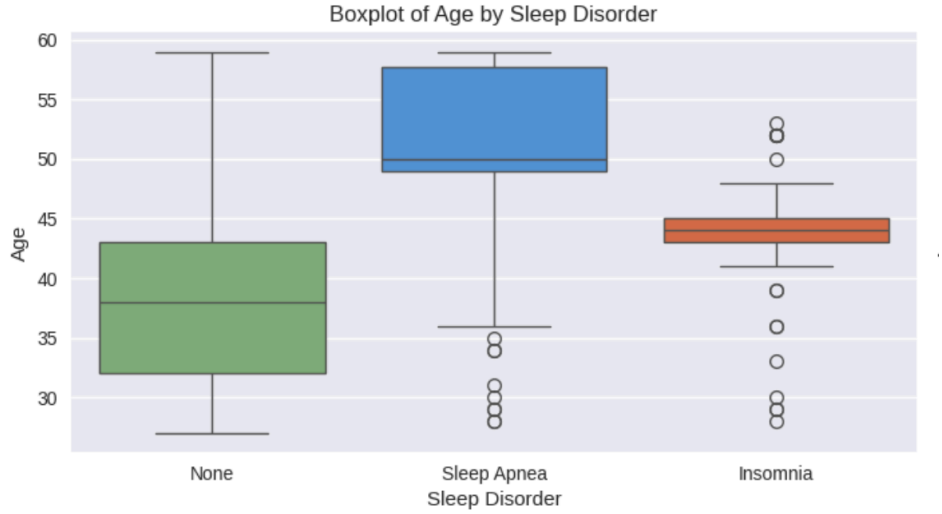
Exploratory Data Analysis :

In this section, we will explore different patterns in the dataset and relationships among features

Age : The distribution of age can be observed as below. It has interquartile range from 35 to 50 with median as 43 approximately. The right side graph shows that more than 100 people are in age group 40 to 45yrs.



From the below boxplot of Age vs Sleep Disorder, the age distribution across different sleep disorders: "None," "Sleep Apnea," and "Insomnia." The median age is highest for Sleep Apnea which is around 50, followed by Insomnia which is 44, and then None which is 38. The data also shows varying interquartile ranges and outliers for each category.

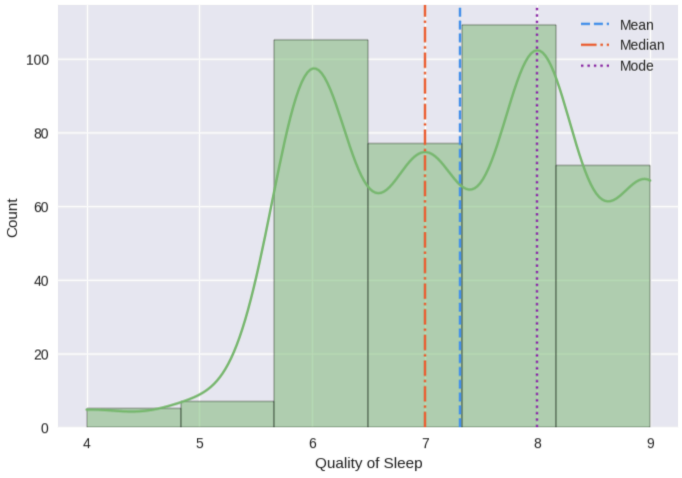


Gender :

After plotting the Sleep Quality vs Gender box plot, we can observe that the median Sleep quality of Female is superior to that of Male and indicates that most men have a Sleep Quality index between 6 to 8, whereas women have it for 7 to 9. Indicating women have greater avg sleep quality compared to Men.

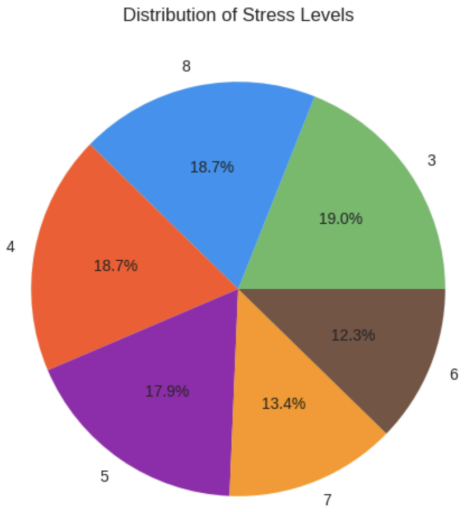
Quality of Sleep :

The Sleep Quality histogram helps us depict the following values. The mean of the sleep quality is approximately 7, the median value is also approximately 7 and mode of sleep quality is 8. Indicating most people have sleep quality rating as 8.



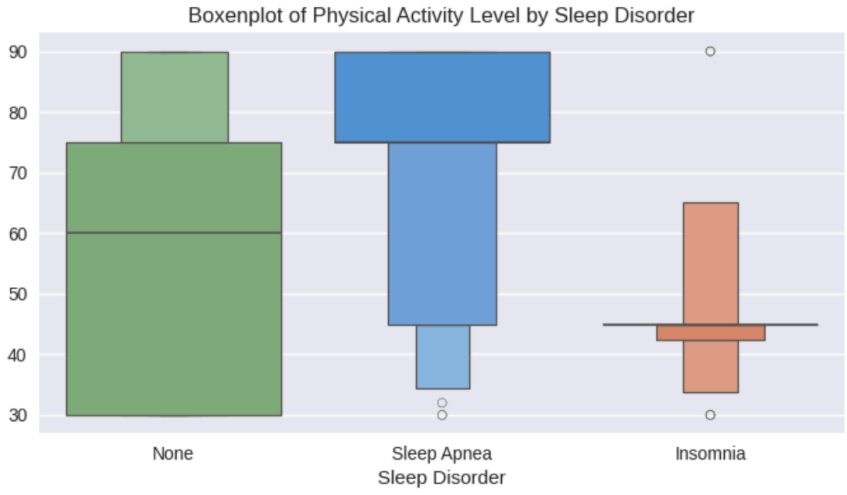
Stress Levels :

The distribution of stress levels from the dataset can be seen as below. The pie chart depicts that 19% of people have a stress level of 3, and 18.7% of people have a stress level of 8 and so on.



Physical Activity Level vs Sleep Disorder :

The boxen plot compares the physical activity levels of individuals with different sleep disorders: "None," "Sleep Apnea," and "Insomnia.". For sleep disorder ("None"), the distribution ranges from 30 to 90, with a median around 65 and an interquartile range (IQR) from 60 to 80. For those with sleep apnea, the distribution skews downward, with a median below 50 and an IQR from 40 to 60. The box is inverted, indicating lower activity levels and some outliers. For individuals with insomnia, the distribution is narrow, with a median around 45 and an IQR ranging from 40 to 50, showing limited variations in activity levels. We can see that less physical activity has a good chance of getting sleep disorder.

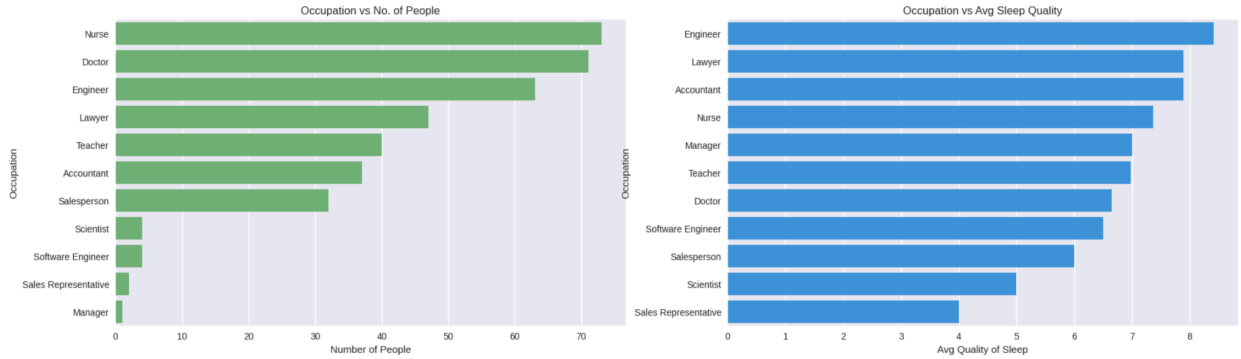


Occupation :

The two plots Occupation vs Number of People and Occupation vs Avg. Sleep Quality helps us to understand the role of occupation and how it affects sleep quality.

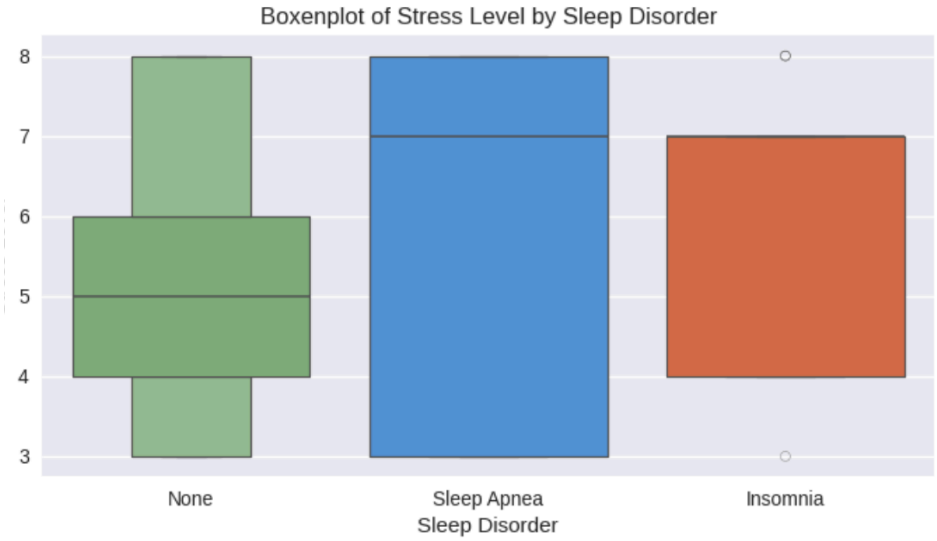
The first graph shows that the most pursued occupation from the graph is Nurse(around 74), followed by Doctor,and so on. The least number (around 2) of people are found for Manager occupation

The second graph plots Occupation vs Avg. Sleep Quality. We can observe that Engineers have a good sleep quality i.e. 8 and Sales Representatives have worst sleep quality i.e. 4 which helps us understand how occupations affect sleep quality



Stress Level by Sleep Disorder :

The boxen plot of Stress Level by Sleep Disorder indicates that people who have No sleep disorder (categorized as None) have a median stress level of 5. Whereas, people having sleep disorder Sleep Apnea have a median Stress Level of 7 . The stress levels for individuals with insomnia are very similar, leading to a compressed or nonexistent boxenplot. Indicating Higher stress levels leads to Sleep Disorders

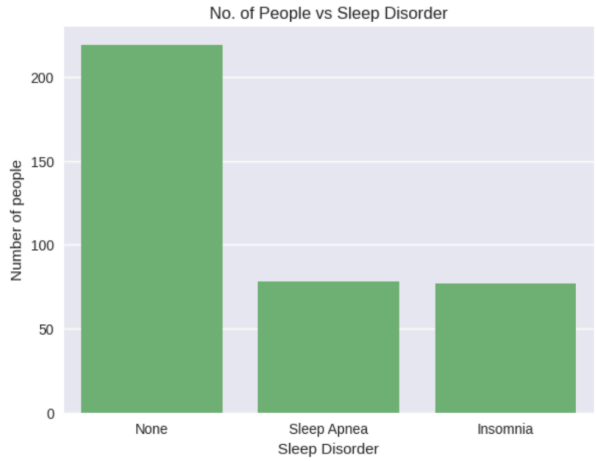


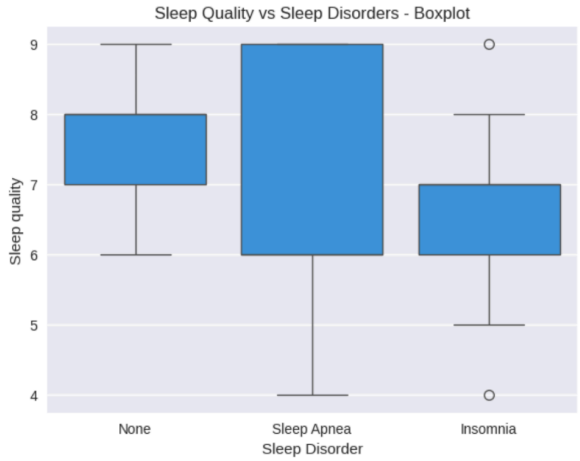
Sleep Disorder :

Below are the two graphs, Number of People vs Sleep Disorder and Sleep Quality vs Sleep Disorder.

From the first graph, we can observe that while over 200 people have a no sleep disorder, around 75 people and 70 people have Sleep Apnea and Insomnia respectively

From the second graph, we can observe that individuals with no sleep disorder have a higher and more consistent sleep quality, while those with sleep apnea or insomnia have lower and more varied sleep quality.



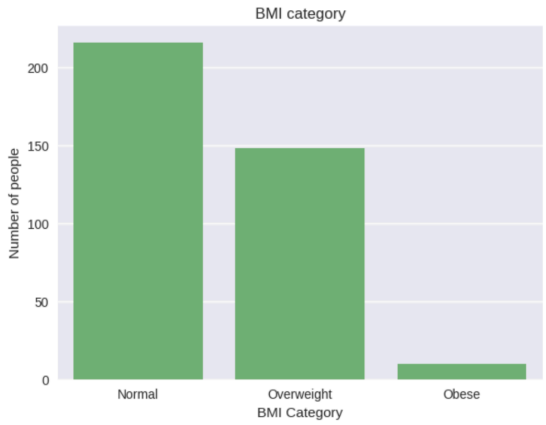


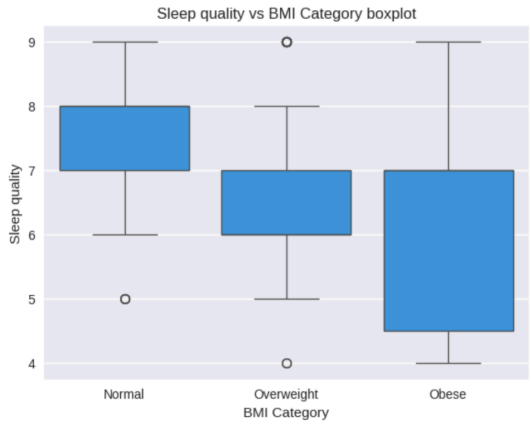
BMI Category :

BMI Category can have an impact on individual sleep. Let us analyze the BMI category with respect to Number of People and Sleep Quality.

The first graph shows that over 200 people are normal and around 150 people are overweight and around 10 people are Obese.

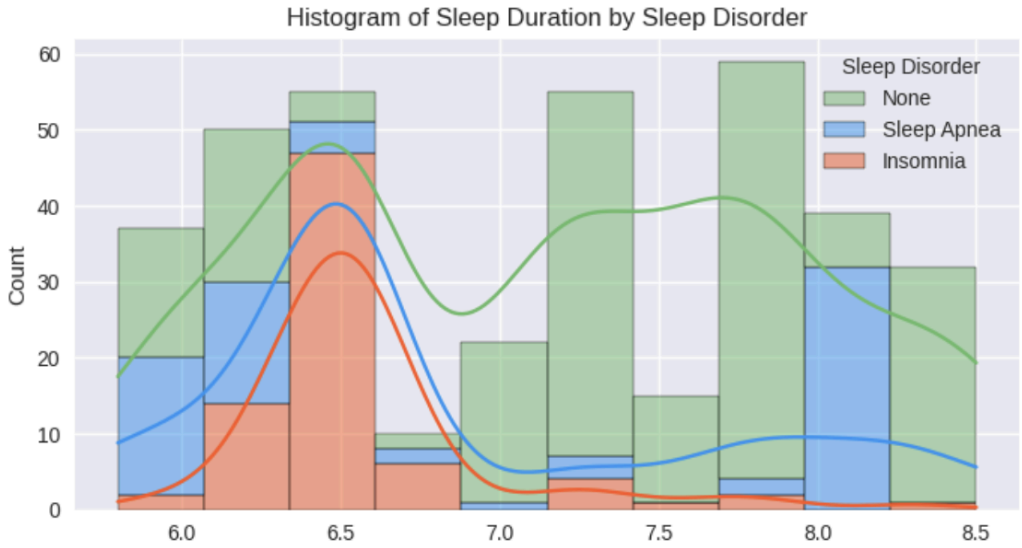
The box plot of Sleep Quality vs BMI Category boxplot shows interquartile range for Normal category is 7-8, Overweight is 6 to 7, and Obese is 4.5 to 7. It indicates that Normal Weight people have a good sleep quality compared to Overweight and Obese Category people have a low sleep quality interquartile.





Sleep Duration :

From the Histogram of Sleep Duration by Sleep Disorder we can observe that when sleep duration is around 7.0 to 8.5 there are very less Sleep Disorders (both Sleep Apnea and Insomnia)



The histogram plots various sleep durations and sleep disorders that occur for individuals who have that average sleep duration.

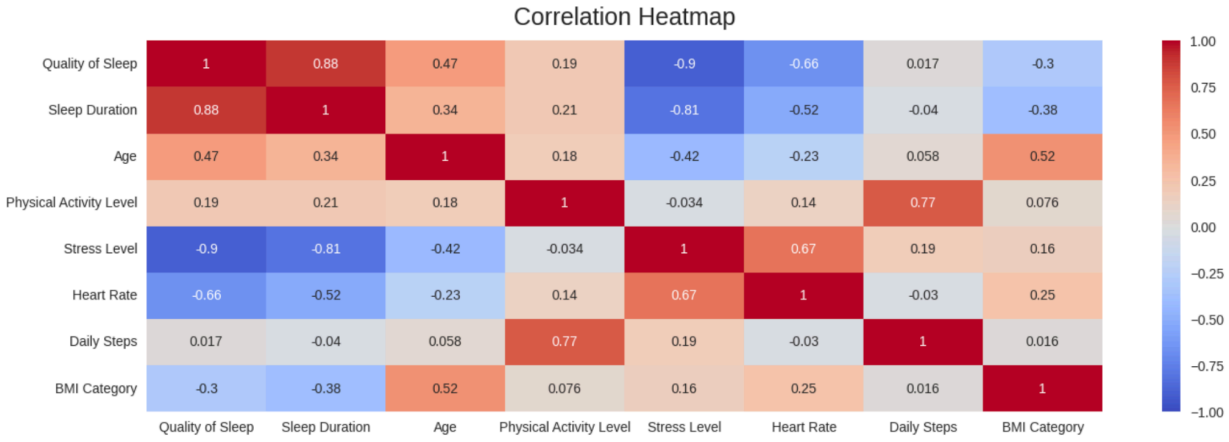
Correlation :

The correlation heatmap shows how strongly features are connected with each other. It ranges from -1 to 1. Dark red color in the heatmap shows stronger correlations (positive correlation values) and Dark blue shows weak correlations (negative correlation values).

Positive correlation indicates that change in one variable increases other variable as well. Conversely, Negative Correlation indicates that with change in one variable the other variable decreases.

We can observe Strong positive correlation between Quality of Sleep and Sleep Duration, Physical Activity Level and Daily Steps, Stress Level and Heart Rate

We can also observer Strong negative correlation between Quality of Sleep and Stress level, Sleep Duration and Stress level, Quality of Sleep and Heart Rate.



Data Preparation :

To test and train the data for predictions, I have used 'train_test_split' method of sklearn and splitted the data into 75% (training) and 25% (training).

Classification Method :

I have used three classification models to predict the Sleep Quality based on other features and compared their results.

1. Logistic Regression : Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables.

2. KNN Classifier : k-Nearest Neighbors (KNN) is a simple and intuitive classification algorithm where a data point is classified based on the majority class of its nearest neighbors.
3. Random Forest :Random Forest is an ensemble learning method that builds multiple decision trees during training and combines their predictions through averaging or voting to improve accuracy and reduce overfitting.

After using these three classifications methods I have obtained following accuracies

Accuracy :

- Logistic Regression : 0.86
- KNN Classifier : 0.88
- Random Forest : 0.89

Classification Report :

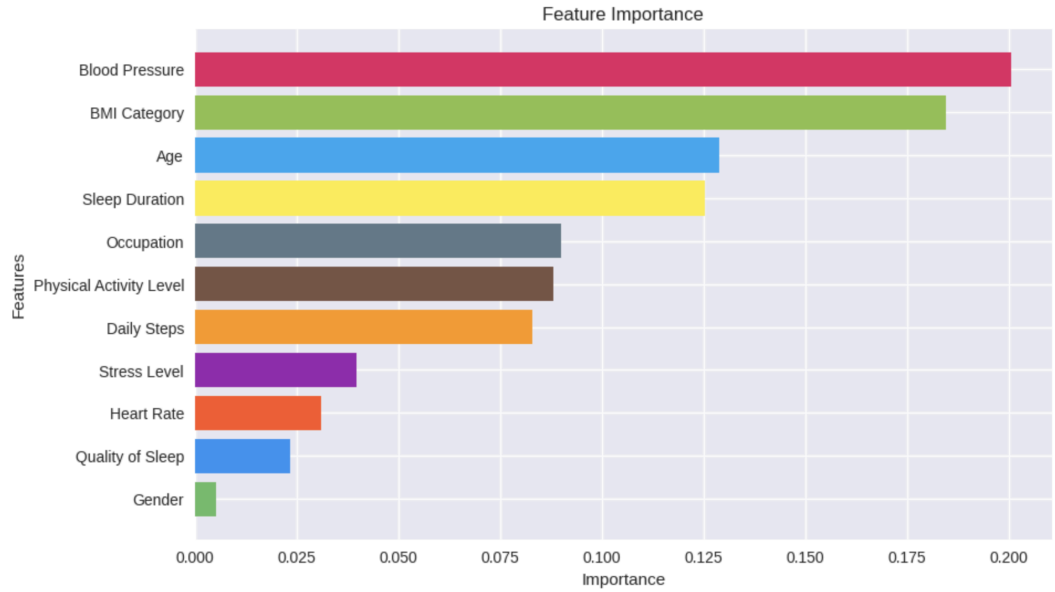
Below table shows average weighted classification report of each model

Model	Accuracies	Recall	Precision	F1 Score
Logistic Regression	0.86	0.86	0.88	0.86
KNN Classifier	0.88	0.88	0.90	0.88
Random Forest	0.89	0.89	0.90	0.89

Analyzing this report, I have selected Random Forest as the best model for this prediction as it has the higher accuracy, recall, precision and f1 score.

Feature Importance :

The trained Random Forest Classifier gives the below plot for indicating feature importance in prediction of sleep disorder. From the plot, we can conclude that Blood Pressure, BMI Category, and Age are the top three features that contribute to predicting the predictor variable - Sleep Disorder.



Conclusion :

After using multiple classification models to predict Sleep Disorder with other target variables, the most optimal model was the Random Forest Classifier with an accuracy of **89%**. Also from feature importance we can understand that **Blood Pressure**, **BMI Category, and Age** are three most important features contributing to the prediction of Sleep Disorder.