**Sleep Disorder Prediction**

The aim of the project is to analyze the person's lifestyles and medical variables such as age, BMI, physical activity, sleep duration, blood pressure and many more, to predict the sleep disorder and its type.

# About the Dataset

The Sleep Health and Lifestyle Dataset comprises 400 rows and 13 columns, covering a wide range of variables related to sleep and daily habits. It includes details such as gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders.

**Key Features of the Dataset:**

Comprehensive Sleep Metrics: Explore sleep duration, quality, and factors influencing sleep patterns.

Lifestyle Factors: Analyze physical activity levels, stress levels, and BMI categories.

Cardiovascular Health: Examine blood pressure and heart rate measurements. Sleep Disorder Analysis: Identify the occurrence of sleep disorders such as Insomnia and Sleep Apnea.

# Data Dictionary

**Column Name Description**

Person\_ID Unique ID assigned to each person

|  |  |
| --- | --- |
| Gender | The gender of the person (Male/Female) |
| Age | Age of the person in years |
| Occupation | The occupation of the person |
| Sleep\_duration | The duration of sleep of the person in hours |
| Quality\_of\_sleep | A subjective rating of the quality of sleep, ranging from 1 to 10 |
| Physical\_activity | The level of physical activity of the person (Low/Medium/High) |
| Stress Level | A subjective rating of the stress level, ranging from 1 to 10 |
| BMI\_category | The BMI category of the person  (Underweight/Normal/Overweight/Obesity) |
| Blood\_pressure | The blood pressure of the person in mmHg |
| Heart\_rate | The heart rate of the person in beats per minute |
| Daily Steps | The number of steps taken by the person per day |

**Column Name Description**

The presence or absence of a sleep disorder in the person (None,

Sleep\_disorder

Insomnia, Sleep Apnea)

**Details about Sleep Disorder Column:**

None: The individual does not exhibit any specific sleep disorder.

Insomnia: The individual experiences difficulty falling asleep or staying asleep, leading to inadequate or poor-quality sleep.

Sleep Apnea: The individual suffers from pauses in breathing during sleep, resulting in disrupted sleep patterns and potential health risks.

In [ ]: *#importing the libraries* **import** numpy **as** np **import** pandas **as** pd **import** matplotlib.pyplot **as** plt **import** seaborn **as** sns

|  |  |  |
| --- | --- | --- |
| *#loading the dataset*  df **=** pd**.**read\_csv('Sleep\_health\_and\_lifestyle\_dataset.csv') df**.**head() |  |  |
| **Quality Physical**  **Person Sleep**  **Gender Age Occupation of Activity**  **ID Duration**  **Sleep Level** | **Stress Level** | **BMI Category** |

In [ ]:

Out[ ]:

Software

**0** 1 Male 27 6.1 6 42 6 Overweight Engineer

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 2 | Male | 28 | Doctor |  | 6.2 | 6 | 60 | 8 | Normal |
| **2** | 3 | Male | 28 | Doctor |  | 6.2 | 6 | 60 | 8 | Normal |
| **3** | 4 | Male | 28 | Sales  Representative |  | 5.9 | 4 | 30 | 8 | Obese |
| **4** | 5 | Male | 28 | Sales  Representative |  | 5.9 | 4 | 30 | 8 | Obese |

# Data Preprocessing Part 1

|  |
| --- |
| *#checking for missing values* df**.**isnull()**.**sum() |

In [ ]:

Person ID 0

Gender 0

Age 0

Occupation 0

Sleep Duration 0

Quality of Sleep 0

Physical Activity Level 0

Stress Level 0

BMI Category 0

Blood Pressure 0

Heart Rate 0

Daily Steps 0 Sleep Disorder 219 dtype: int64

In [ ]: *#replacing the null values with 'None' in the column 'Sleep Disorder'* df['Sleep Disorder']**.**fillna('None', inplace**=True**)

The nan/None value in sleep disorder stands for no sleep disorder, so it is not a missing value.

In [ ]: *#drop column Person ID* df**.**drop('Person ID', axis**=**1, inplace**=True**)

In [ ]: *#checking the number of unique values in each column* print("Unique values in each column are:") **for** col **in** df**.**columns: print(col,df[col]**.**nunique())

Unique values in each column are:

Gender 2

Age 31

Occupation 11

Sleep Duration 27

Quality of Sleep 6

Physical Activity Level 16

Stress Level 6

BMI Category 4

Blood Pressure 25

Heart Rate 19

Daily Steps 20

Sleep Disorder 3

## Splitting the blood pressure into systolic and diastolic

|  |
| --- |
| *#spliting the blood pressure into two columns*  df['systolic\_bp'] **=** df['Blood Pressure']**.**apply(**lambda** x: x**.**split('/')[0]) df['diastolic\_bp'] **=** df['Blood Pressure']**.**apply(**lambda** x: x**.**split('/')[1])  *#droping the blood pressure column*  df**.**drop('Blood Pressure', axis**=**1, inplace**=True**) |

In [ ]:

|  |
| --- |
| *#replacing normal weight with normal in BMI column*  df['BMI Category'] **=** df['BMI Category']**.**replace('Normal Weight', 'Normal') |

In [ ]:

|  |
| --- |
| df**.**head() |

In [ ]:

**Quality Physical**

**Sleep Stress BMI Heart**

**Gender Age Occupation of Activity**

**Duration Level Category Rate**

**Sleep Level**

Software

**0** Male 27 6.1 6 42 6 Overweight 77 Engineer

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Male | 28 | Doctor | 6.2 | 6 | 60 | 8 | Normal |  | 75 |
| **2** | Male | 28 | Doctor | 6.2 | 6 | 60 | 8 | Normal |  | 75 |
| **3** | Male | 28 | Sales  Representative | 5.9 | 4 | 30 | 8 | Obese |  | 85 |
| **4** | Male | 28 | Sales  Representative | 5.9 | 4 | 30 | 8 | Obese |  | 85 |

## Checking the unique values from each categorical column

|  |
| --- |
| *#unique values from categorical columns* print(df**.**Occupation**.**unique()) print('\n') print(df['BMI Category']**.**unique()) print('\n') print(df['Sleep Disorder']**.**unique()) |

In [ ]:

['Software Engineer' 'Doctor' 'Sales Representative' 'Teacher' 'Nurse' 'Engineer' 'Accountant' 'Scientist' 'Lawyer' 'Salesperson' 'Manager']

['Overweight' 'Normal' 'Obese']

['None' 'Sleep Apnea' 'Insomnia']

# Explorative Data Analysis

The EDA is divided into two phases:

Phase 1: Understanding the data by plotting its variables

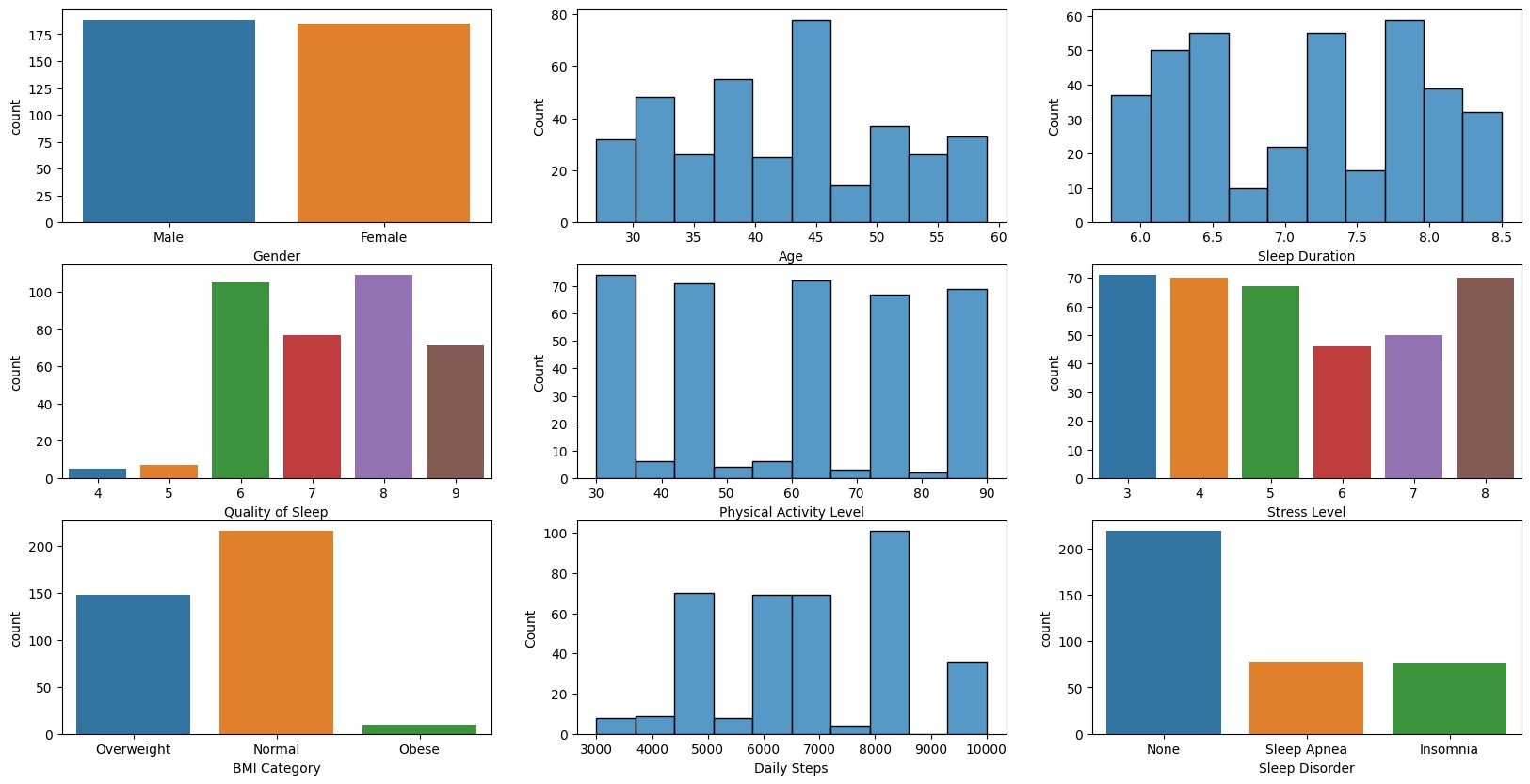
Phase 2: Understanding the correlation between the variables

## Phase 1

|  |
| --- |
| fig,ax **=** plt**.**subplots(3,3,figsize**=**(20,10)) sns**.**countplot(x **=** 'Gender', data **=** df, ax **=** ax[0,0]) sns**.**histplot(x **=** 'Age', data **=** df, ax **=** ax[0,1], bins **=** 10) sns**.**histplot(x **=** 'Sleep Duration', data **=** df, ax **=** ax[0,2], bins **=** 10) sns**.**countplot(x **=** 'Quality of Sleep', data **=** df, ax **=** ax[1,0]) sns**.**histplot(x **=** 'Physical Activity Level', data **=** df, ax **=** ax[1,1], bins **=** 10) sns**.**countplot(x **=** 'Stress Level', data **=** df, ax **=** ax[1,2]) sns**.**countplot(x **=** 'BMI Category', data **=** df, ax **=** ax[2,0]) sns**.**histplot(x **=** 'Daily Steps', data **=** df, ax **=** ax[2,1], bins **=** 10) sns**.**countplot(x **=** 'Sleep Disorder', data **=** df, ax **=** ax[2,2]) |

In [ ]:

<Axes: xlabel='Sleep Disorder', ylabel='count'>



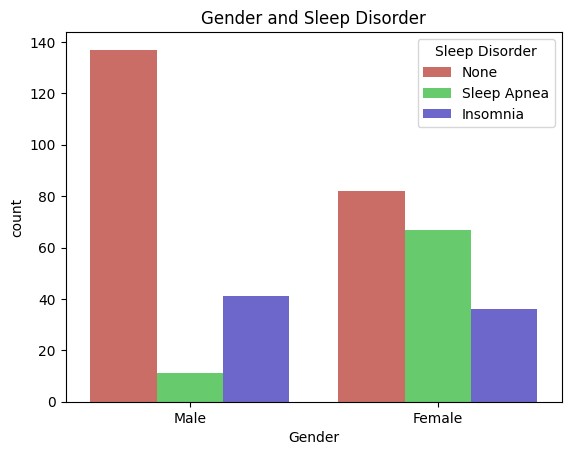
The number of males and females is almost equal, out of which majority of the people have age between 30-45 years. Most of the people have sleep quality greater than 5 which means there are getting sufficient sleep. Moreover, most of the people have normal BMI whci directly relates with the distribution of sleep disorder which shows equal number of people with and without sleep disorder.

**Phase 2**

## Gender and Sleep Disorder

In [ ]: *#Gender count plot* sns**.**countplot(x **=** 'Gender', data **=** df, palette **=** 'hls', hue **=** 'Sleep Disorder')**.**

Out[ ]: Text(0.5, 1.0, 'Gender and Sleep Disorder')



Most of the males and females are not suffering from any sleep disorder. However females tend to have more sleep disorder as compared to males. The number of females suffering from Sleep Apnea is quite high as compared to males. But in contrast to that, greater number of males are suffering from Insomia as compared to females.

## Effect of Occupation on Sleep Disorder

In [ ]: ax **=** sns**.**countplot(x **=** 'Occupation', data **=** df, hue **=** 'Sleep Disorder') ax**.**set\_xticklabels(ax**.**get\_xticklabels(), rotation **=** 90)

Out[ ]: [Text(0, 0, 'Software Engineer'),

Text(1, 0, 'Doctor'),

Text(2, 0, 'Sales Representative'),

Text(3, 0, 'Teacher'),

Text(4, 0, 'Nurse'),

Text(5, 0, 'Engineer'),

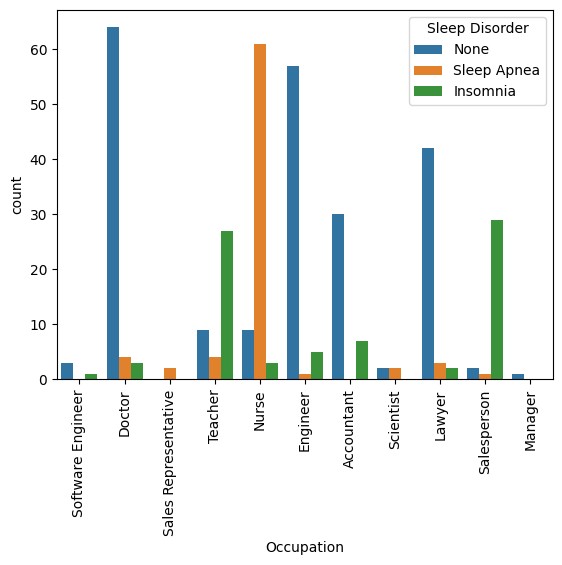
Text(6, 0, 'Accountant'),

Text(7, 0, 'Scientist'),

Text(8, 0, 'Lawyer'),

Text(9, 0, 'Salesperson'),

Text(10, 0, 'Manager')]

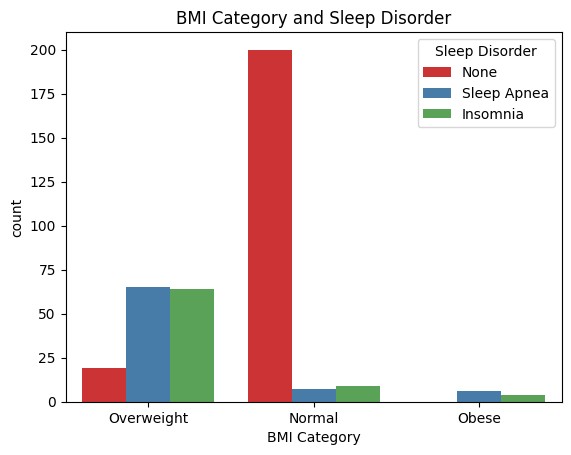


From the graph it is clear that the occupation has huge impact on the sleep disorder. Nurses are more subjected to have Sleep Apenea as compared to other occupations and very few of them have no sleep disorder. After nurses, the next most affected occupation is the Salesperson, which counts for the most suffering from Insomia followed by teachers. However there are some occupations where most of the people have very few instance of Sleep Apenea and Insomia such as Engineers, Doctors, Accountants, Lawyers. The Software ENgineers and Managers are so less in number so I cannot say much about that, But the occupation Sales Representative has shown only Sleep Apenea and no Insomia or No sleep disorder.

## BMI and Sleep Disorder

In [ ]: sns**.**countplot(x **=** 'BMI Category', hue **=** 'Sleep Disorder', data **=** df, palette **=** '

Out[ ]: Text(0.5, 1.0, 'BMI Category and Sleep Disorder')



People with normal BMI are less likely to suffer from any sleep disorder. However, this is opposite in case of Overweight and Obese people. Overweight are more likely to suffer more from sleep disordera than Obese people.

# Data Preprocessing Part 2

## Label Encoding for categorical variables

|  |
| --- |
| **from** sklearn **import** preprocessing  label\_encoder **=** preprocessing**.**LabelEncoder() |

In [ ]:

|  |
| --- |
| vars **=** ['Gender', 'Occupation','BMI Category','Sleep Disorder'] **for** i **in** vars:  label\_encoder**.**fit(df[i]**.**unique()) df[i] **=** label\_encoder**.**transform(df[i]) print(i,':' ,df[i]**.**unique()) |

In [ ]:

Gender : [1 0] Occupation : [ 9 1 6 10 5 2 0 8 3 7 4]

BMI Category : [2 0 1]

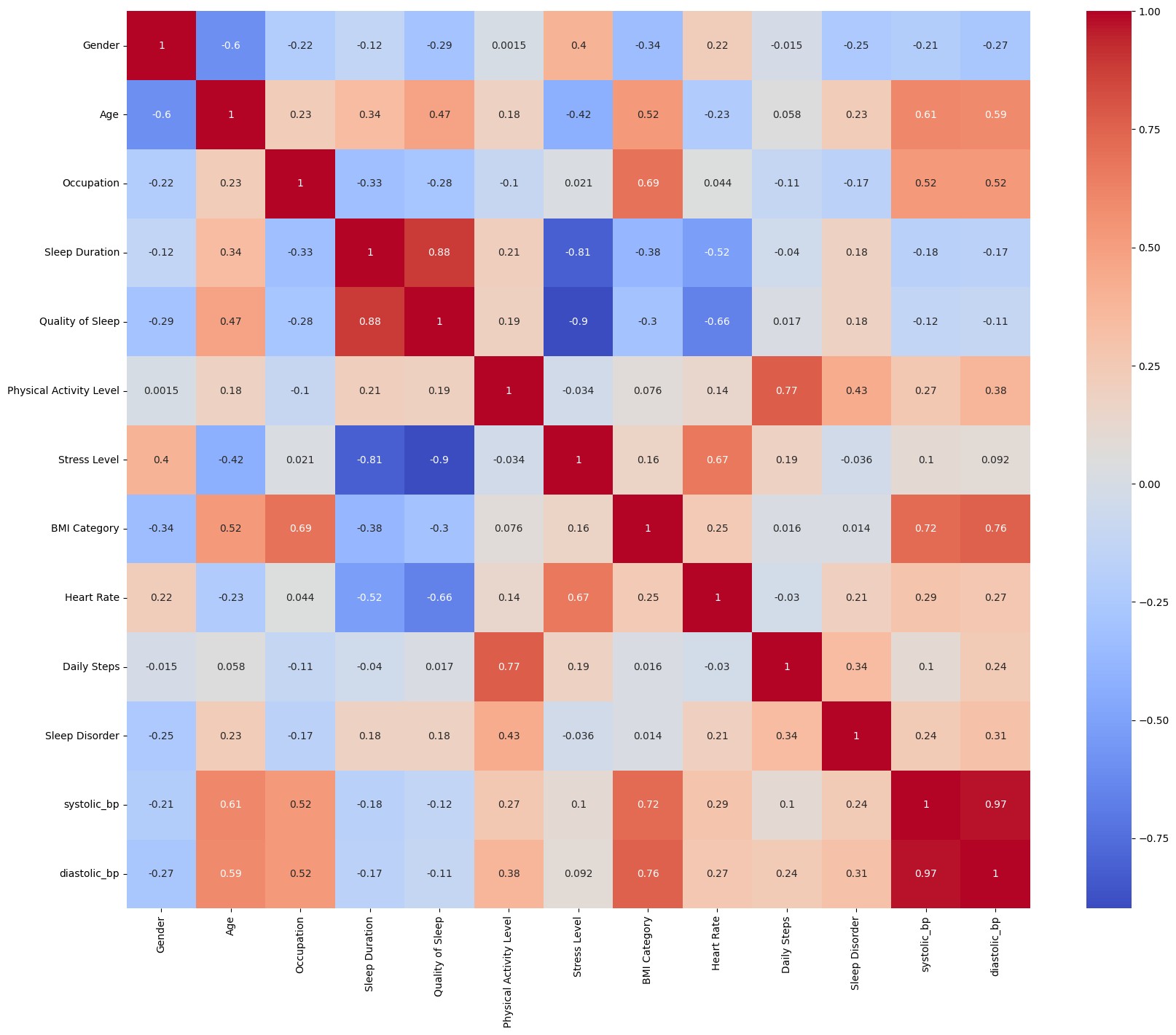
Sleep Disorder : [1 2 0]

# Correlation Matrix Heatmap

|  |
| --- |
| *#Correlation Matrix Heatmap* plt**.**figure(figsize**=**(20, 16)) sns**.**heatmap(df**.**corr(), annot **=** **True**, cmap **=** 'coolwarm') |

In [ ]:

Out[ ]: <Axes: >



# Train Test Split

|  |
| --- |
| **from** sklearn.model\_selection **import** train\_test\_split  X\_train, X\_test, y\_train, y\_test **=** train\_test\_split(df**.**drop('Sleep Disorder',axi |

In [ ]:

# Model Building

For predictiong the sleep disorder thriugh classification algorithms I will use the following algorithms:

1. Decision Tree Classifier
2. Random Forest Classifier

## Decision Tree Classifier

|  |
| --- |
| **from** sklearn.tree **import** DecisionTreeClassifier dtree **=** DecisionTreeClassifier() dtree |

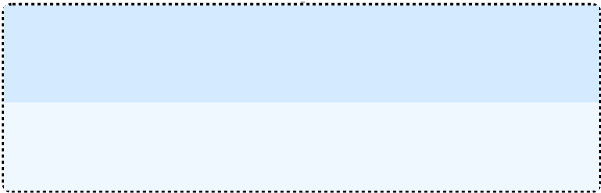
In [ ]:

Out[ ]: ▾DecisionTreeClassifier

DecisionTreeClassifier()

Training the model with train dataset

In [ ]: dtree**.**fit(X\_train, y\_train)

Out[ ]: ▾DecisionTreeClassifier

DecisionTreeClassifier()

In [ ]: *#training accuracy* print("Training Accuracy:",dtree**.**score(X\_train,y\_train))

Training Accuracy: 0.9348659003831418

## Decision Tree Model Evalution

In [ ]: d\_pred **=** dtree**.**predict(X\_test) d\_pred

Out[ ]: array([1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 2, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 2, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 0, 2, 0, 0,

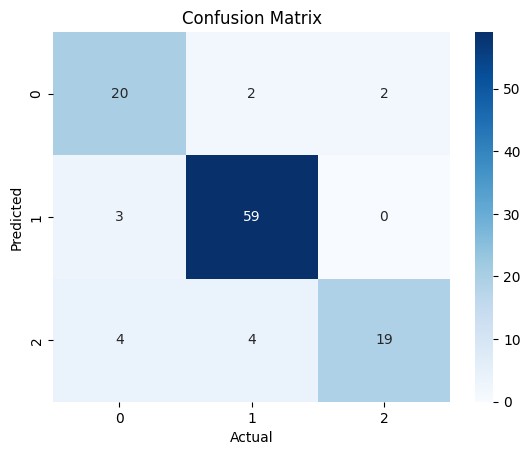
1, 1, 1, 1, 2, 1, 2, 2, 2, 1, 0, 2, 0, 2, 2, 1, 1, 0, 1, 1, 0, 1,

0, 1, 1, 1, 1, 0, 1, 2, 2, 0, 1, 1, 2, 0, 1, 2, 1, 1, 1, 2, 0, 2,

1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 2, 0, 1, 1, 0, 2, 1, 1, 2, 1, 0])

Using Confusion matrix heatmap to visualize the model accuracy

In [ ]: **from** sklearn.metrics **import** confusion\_matrix sns**.**heatmap(confusion\_matrix(y\_test, d\_pred), annot**=True**, cmap**=**'Blues', fmt**=**'g') plt**.**title('Confusion Matrix') plt**.**xlabel('Actual') plt**.**ylabel('Predicted') plt**.**show()

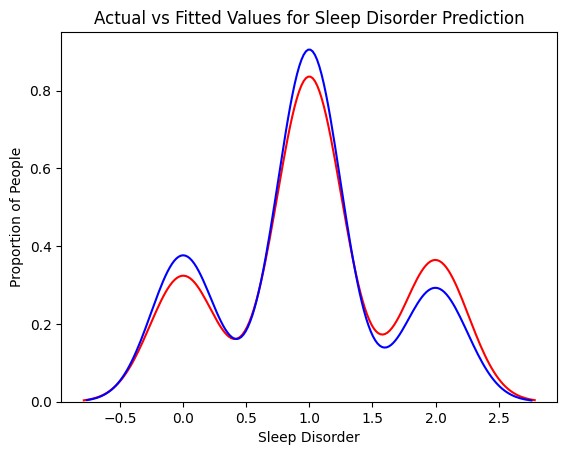


The diagonal boxes show the count of true positive results, i.e correct predictions made by the model. The off-diagonal boxes show the count of false positive results, i.e incorrect predictions made by the model.

## Dsitribution plot for predicted and actual values

|  |
| --- |
| ax **=** sns**.**distplot(y\_test, hist**=False**, color**=**"r", label**=**"Actual Value") sns**.**distplot(d\_pred, hist**=False**, color**=**"b", label**=**"Fitted Values" , ax**=**ax) plt**.**title('Actual vs Fitted Values for Sleep Disorder Prediction') plt**.**xlabel('Sleep Disorder') plt**.**ylabel('Proportion of People') plt**.**show() |

In [ ]:



The actual values are represented with red and the predicted ones with blue. As shown in the graph, the model's prediction are able to follow the curve of actual values but the predicted values are still different from actual ones. Therefore the model is not able to predict the values accurately.

**Classification Report**

|  |
| --- |
| **from** sklearn.metrics **import** classification\_report print(classification\_report(y\_test, d\_pred)) |

In [ ]:

precision recall f1-score support

1. 0.74 0.83 0.78 24
2. 0.91 0.95 0.93 62 2 0.90 0.70 0.79 27

accuracy 0.87 113 macro avg 0.85 0.83 0.84 113 weighted avg 0.87 0.87 0.87 113

The model gives pretty decent results with an accuracy of 87% and an average F1 score of 0.83. The model is able to predict the sleep disorder with a good accuracy.

## Random Forest Classifier

|  |
| --- |
| **from** sklearn.ensemble **import** RandomForestClassifier  rfc **=** RandomForestClassifier(n\_estimators**=**100, random\_state**=**42) |

In [ ]:

Training the model with train dataset

|  |
| --- |
| rfc**.**fit(X\_train, y\_train) |

In [ ]:

Out[ ]: ▾ RandomForestClassifier

RandomForestClassifier(random\_state=42)

|  |
| --- |
| *#Training accuracy*  print("Training accuracy: ",rfc**.**score(X\_train,y\_train)) |

In [ ]:

Training accuracy: 0.9348659003831418

## Random Forest Classifier Evaluation

In [ ]: rfc\_pred **=** rfc**.**predict(X\_test) rfc\_pred

Out[ ]: array([1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 2, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 2, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 0, 2, 0, 0,

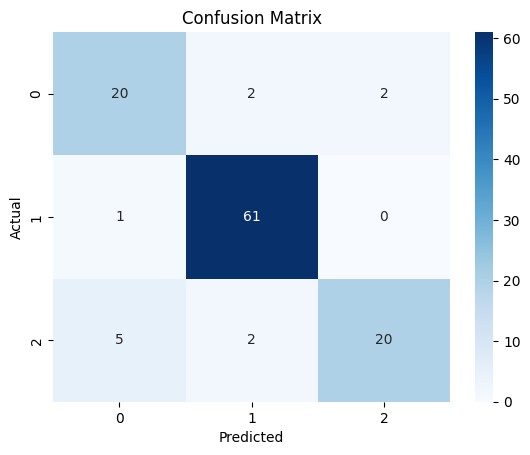
1, 1, 1, 1, 2, 1, 2, 0, 2, 1, 0, 2, 0, 2, 2, 1, 1, 0, 1, 1, 0, 1,

0, 1, 1, 1, 1, 0, 1, 2, 2, 0, 1, 1, 2, 0, 1, 2, 1, 1, 1, 2, 1, 2,

1, 1, 2, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 2, 0, 1, 2, 0, 2, 1, 1, 2, 1, 0])

Using confusion matrix heatmap to visualize the model accuracy

In [ ]: *#confusion matrix heatmap* sns**.**heatmap(confusion\_matrix(y\_test, rfc\_pred), annot**=True**, cmap**=**'Blues') plt**.**title('Confusion Matrix') plt**.**xlabel('Predicted') plt**.**ylabel('Actual') plt**.**show()

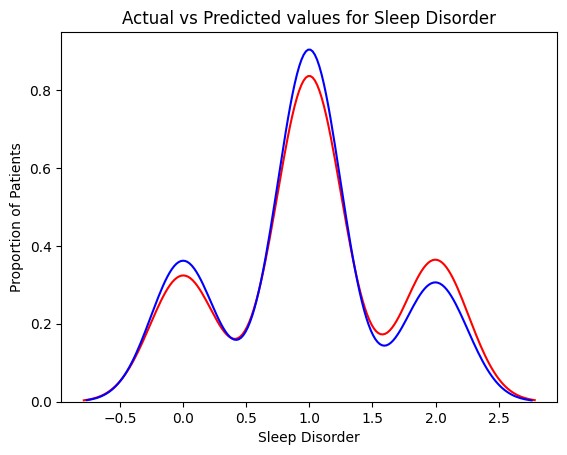


The Random Forest Classifier model has greater accuracy than the Decision Tree Classifier model. The diagonal boxes count for the True Positives i.e correct predictions, whereas the off-diagonal boxes show the count of false positive results, i.e incorrect predictions made by the model. Since the number of false positve value is less, it shows that the model is good at predicting the correct results.

## Distribution plot for predicted and acutal values

|  |
| --- |
| ax **=** sns**.**distplot(y\_test, hist**=False**, color**=**"r", label**=**"Actual Value") sns**.**distplot(rfc\_pred, hist**=False**, color**=**"b", label**=**"Predicted Values" , ax**=**ax) plt**.**title('Actual vs Predicted values for Sleep Disorder') plt**.**xlabel('Sleep Disorder') plt**.**ylabel('Proportion of Patients') plt**.**show() |

In [ ]:



The Random forest classifier has improved accuracy as compared to the Decision Tree which is shown with the gap between the actual and predcited values which was wider incase of Descision Tree Classifier.

### Classification Report

|  |
| --- |
| print(classification\_report(y\_test, rfc\_pred)) |

In [ ]:

precision recall f1-score support

1. 0.77 0.83 0.80 24
2. 0.94 0.98 0.96 62 2 0.91 0.74 0.82 27

accuracy 0.89 113 macro avg 0.87 0.85 0.86 113 weighted avg 0.90 0.89 0.89 113

The Random Forest Classifier model has an accuracy of 89% and an avergae F1 score of 0.86. From the metrics it is quite clear that the model is able to predict the sleep disorder quite effectively, with increased accuracy than Decision Tree Classifer.

# Conclusion

From the exploratory data analysis, I have concluded that the sleep orders depends upon three main factors that are gender, occupation and BMI of the patient. The males have more instance of Insomia whereas femlaes have more instances of Sleep Apnea. In addition the that people with occupation such as nurses are more prone to sleep disorders. The BMI of the patient also plays a vital role in the prediction of sleep disorders. The patients who are either Obese or overweight are more prone to sleep disorders.

Coming to the classfication models, both the models performed pretty good, however the Random Forest Classifier have excellent results with 89% accuracy.