Sleep Disorder Classification

Lecturers : Lay Puthineath

Team number: 5

Teammate's name : Sar Vuthyrak
Thoeun Nimol
Yoeun Youvanneath
Liv Seavhong
Veng Sreynich

Table Contents

I. Overview of Project

In response to the prevalent issue of sleep disorder among our team members, we decided to initiate a project called "**Sleep Disorder Classification**". The primary objective of this project is to know the root causes that highly affect sleep disorder.

In order to achieve the project, we need to explore various dataset which answers to our objectives. After thorough exploration of various datasets related to sleep disorder classification, our team has chosen the "**Sleep Health and Lifestyle Dataset**". This dataset comprises 13 variables such as ***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***.

With a dataset size of **374** records, our team aims to leverage this information to acquire factors influencing sleep disorders and " ***Which variables do they affect sleep disorder the most?***" and by those variables "***Do people tend to have sleep disorders or not?***". At the end of this project, we hopefully can suggest some recommendations to the people for avoiding the sleep disorder as well.

II. EDA (Exploratory Data Analysis)

In order to explore our dataset, we tend to apply three steps of data exploration which includes Data Preprocessing, Data Cleansing, and Data Visualization.

1. Data Preprocessing

We started to preprocess the data by a few steps in order to gather the basic understanding of our dataset :

- **Loading Data** : we used the " **pd.read_csv(file path)**" function to load the data from the dataset by only displaying the first 10 lines to understand the data characteristic.
- **Shaping Data**: by using the "**myData.shape**" function, we can see that our dataset contains **374** rows and **13** columns.
- **Describing Data**: to get the dataset summary, we need to use one function called "**myData.describe**" to get the descriptive statistic of the dataset. In return, we got the mean, standard deviation, maximum, minimum of each 13 columns. We can see the average age of people in the dataset was **42** years old while the oldest was **59** years and the youngest was **27** years old. Moreover, the average sleep duration of them was **7 hours**.

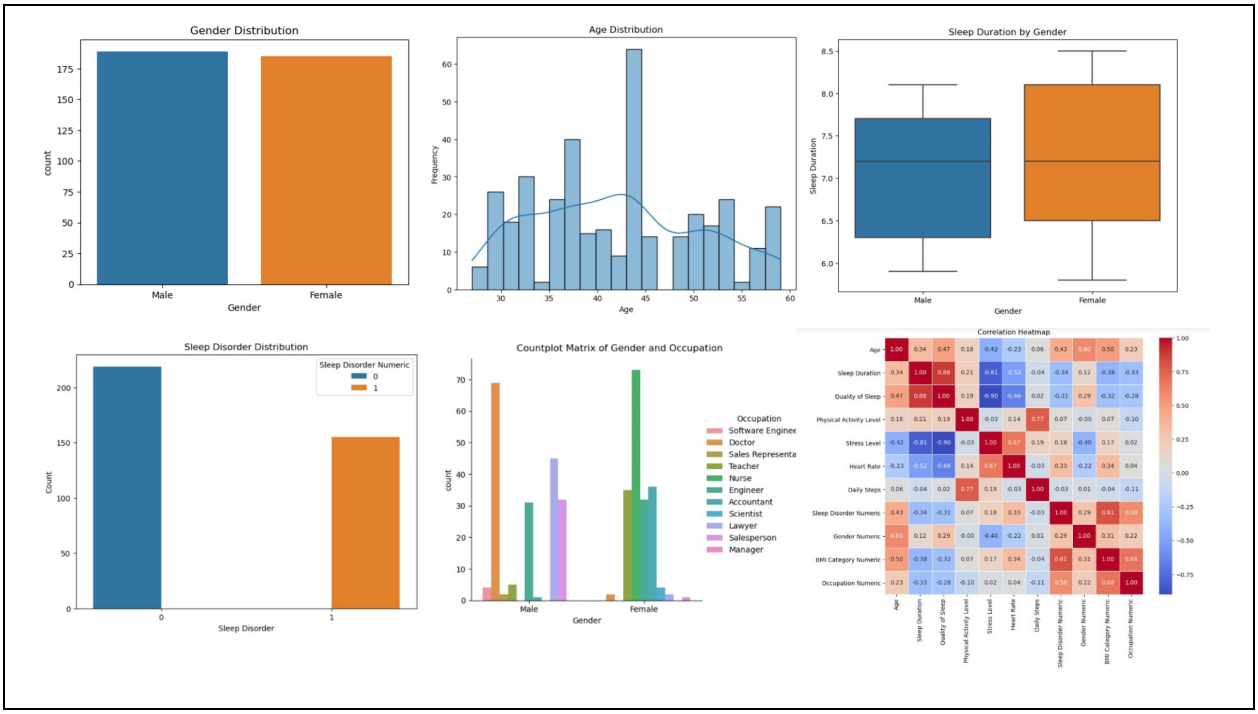
2. Data Cleansing

By ensuring our dataset was cleaned, we need to check it by using a function called "**myData.isna().sum()**" in order to see if our dataset had a null value or not. After applying the function, we see that our dataset had ***no null value*** among 13 columns.

3. Data Visualization

We know that "**Visualization**" is a part of providing an insight of a dataset for easy understanding. In our case, before we visualize our dataset, we need to encode some **non-numeric** variables into **numeric** variables. Among 13 columns, we encode 4 columns such as **Sleep Disorder**, **Gender, BMI Category**, and **Occupation**.

After encoding those variables, we had visualized some variables by using different plots and charts. By doing that, we need to import 2 libraries **Seaborn** and **Matplotlib**. And those are countplot, histoplot, boxplot, catplot, pairplot, and heatmap correlation matrix.



III. Modeling and Evaluation

1. Model Implementation

- **Data Cleanup**: Removing columns prior to feature engineering, we removed 1 column such as 'Person ID'. Person ID is the unique id for every individual.

- **Data Splitting**: We have divided data into independent and dependent variables. And also splitting 80% of data into a training set and 20% into a testing set.
- **Feature Scaling**: We have put training and test data into feature scaling to help in creating a more stable and well-behaved model.
- **Model Training**: We created four models such as Logistic Regression, Support Vector Machine, Decision Tree and Random Forest Classification.

2. Model Evaluation

Accuracy of the models were measured by the confusion matrix on the test data set. We got two models with the same highest accuracy of **0.9333%** which are Decision Tree and Random Forest. And we decided to use **Random Forest**.

IV. Conclusion

In the process of doing this project, it requires us to follow some steps(**Objective Understanding**, **Data Collection**, **Data Exploratory**, **Model Training** , **Evaluation**) in order to provide a clear instruction for our teammate. Especially, data exploratory was an important point that we need to understand its characteristics for building a model that matches with our objectives. Moreover, in the training process we shouldn't try only one model but instead, give the dataset we are trained by more than one model because we can see the accuracy of each model for choosing.

In conclusion, our analysis of sleep disorder classification using four trained models reveals that **BMI** emerges as a significant factor influencing sleep disorders. The models consistently identify **BMI** as a key variable, highlighting its importance in understanding and predicting sleep-related issues. Further exploration and validation of these findings could contribute to enhanced diagnostics and personalized interventions for individuals with sleep disorders, with BMI serving as a valuable feature in predictive models. Therefore, the model that fits with this type of training is **Random Forest** which gives us the most accuracy compared to one another. So, if we want to avoid sleep disorders, we should balance our BMI into standardization .