DOSSO Aminata
EL GARHI Myriam
OURIF Bilal

Introduction to Data Science with Python : Final Project Submission

DOSSO Aminata
EL GARHI Myriam
OURIF Bilal

Summary

DOSSO Aminata
EL GARHI Myriam
OURIF Bilal

I. Problem Definition - Sleep Health and Lifestyle Dataset

Introduction

Sleep is a major concern among students, who often struggle to balance adequate rest with academic demands.

This project investigates how various factors (including medical, social, and lifestyle-related aspects) impact sleep quality and contribute to disorders such as insomnia or sleep apnea.

Using the '*Sleep Health and Lifestyle Dataset*' from Kaggle, we aim to identify key correlations between lifestyle choices and overall sleep health.

Through data visualization, we seek to clearly illustrate these relationships, highlighting which behaviors may increase the risk of sleep disorders and which may help prevent them.

In addition, we apply machine learning models to predict the likelihood of a person developing a sleep disorder based on a combination of personal and behavioral characteristics. By combining exploratory analysis with predictive modeling, this project offers a comprehensive approach to understanding and anticipating sleep-related health outcomes.

Dataset Description

The dataset selected for this analysis is the '*Sleep Health and Lifestyle Dataset*', including 374 rows and 13 variables/columns.

These features cover a broad spectrum of information, including sleep habits, demographic details (such as gender, age, and occupation), health status, physical activity, and social influences.

The data combines both numerical and categorical variables, offering a solid foundation to analyse the complex relationships between lifestyle factors and sleep-related health outcomes.

Columns	Data type	Numerical/categorical	Lifestyle aspect	Example of values
Person ID	Integer	Numerical	Identity	1,2,45…
Gender	String	Categorical	Identity	Male,Female
Age	Integer	Numerical	Identity	36,45,50
Occupation	String	Categorical	Social	Doctor,Lawyer…
Quality of Sleep (scale: 1-10)	Integer	Numerical	Health	6,8…
Physical Activity Level (minutes/day)	Integer	Numerical	Health	42,60,30
Stress Level (scale: 1-10	Integer	Numerical	Health	6,7,3…
BMI Category	String	Categorical	Health	Obsese, Normal, Overweight
Blood Pressure (systolic/diastolic)	String	Categorical	Health	140/90 , 125/80
Heart Rate (bpm)	Integer	Numerical	Health	70,82,85…
Daily Steps	Integer	Numerical	Health	3000,3500,8000
Sleep Disorder	String	Categorical	Health	Sleep Apnea, Insomnia

Columns of dataset

II. Preprocessing

Data cleaning

Data cleaning is the process of identifying and handling missing, inaccurate, or irrelevant data to improve its quality and reliability.

To prepare our dataset for modeling and ensure accurate analysis, we performed several cleaning steps.

Firstly we used the '*.fillna()*' method to replace all *'NaN'* values of the *'Sleep Disorder'* column with the label *'No sleep disorder'*. This allowed us to compare lifestyle habits between individuals suffering from sleep disorders and those without.

Feature Engineering

To reduce feature complexity and enhance data clarity, we grouped similar variables together. We also applied label encoding to convert categorical variables into numerical values suitable for machine learning models.

Below are the updated columns created as a result of these modifications :

➢ Physical Activity Level:

To simplify data visualization, we scaled down the original values by dividing them by 10, resulting in integers values ranging from 0 to 10.

➢ BMI Category:

The original dataset included four BMI categories: '*Overweight*', '*Obese*', '*Normal*', and '*Normal Weight*'. We merged them into two groups: '*Normal*' (combining Normal and Normal Weight), and '*Obese*' (combining Overweight and Obese).

➢ Occupation:

Initially, there were 11 distinct occupation types: *'Software Engineer'*, *'Doctor'*, *'Teacher'*, *'Nurse'*, *'Engineer'*, *'Scientist'*, *'Lawyer'*, *'Sales Representative'*, *'Salesperson'*, *'Accountant'*, *'Manager'*.
We grouped these into three broader categories: *'Technical', 'Business' and 'Healthcare/ Education'*.

➢ Blood Pressure Quality:

This new feature categorizes blood pressure readings into two medically recognized categories based on clinical standards: *'Hypertensive'* and *'Non-hypertensive'*.

III. Data visualization

Identity and social characteristics

The dataset includes 189 males and 185 females participants. Among them, 184 are employed in healthcare/ educational fields, 119 are in technical fields and 71 in business-related roles. The data reveals a strong predominance of healthcare/ educational roles among participants.

Gender	count
Male	189
Female	185

Occupation_Group	count
Healthcare/Education	184
Business	119
Technical	71

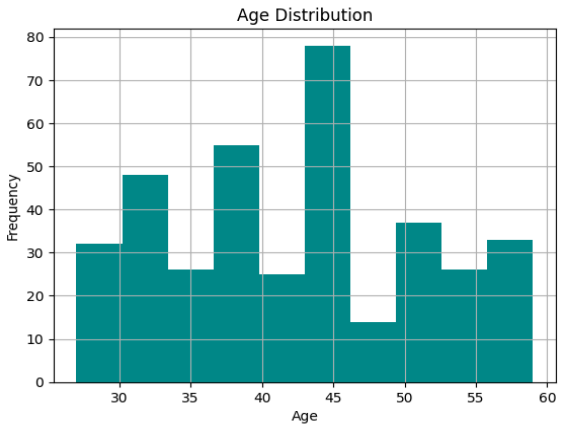
Code :

```
gender = df["Gender"].value_counts()
gender
```

```
occupation = df["Occupation_Group"].value_counts()
occupation
```

We created an initial histogram to visualize the age distribution of the study participants. The most represented age group is around 45 years old, with over 75 individuals, accounting for approximately 20% of the total population.

The ages in the dataset range from around 10 years old to nearly 60, showing a very wide distribution across age groups. The average age of participants is approximately 42 years.



Code:

```
plt.hist(df["Age"], color='darkcyan')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Age Distribution')
plt.grid()
plt.show()
```

Health characteristics

In our analysis of blood pressure quality, we found that 88% of participants are classified as hypertensive according to medical standards.
Additionally, 216 participants out of 374 (approximately 58%) have a normal weight, while 158 participants (about 42%) are overweight.

Blood Pressure quality	count
Hypertensive	332
Non-hypertensive	42

BMI Category	count
Normal Weight	216
Overweight	158

Code:

```
df["Blood Pressure quality"].value_counts()
```

```
BMI_count = df.value_counts("BMI Category")
```

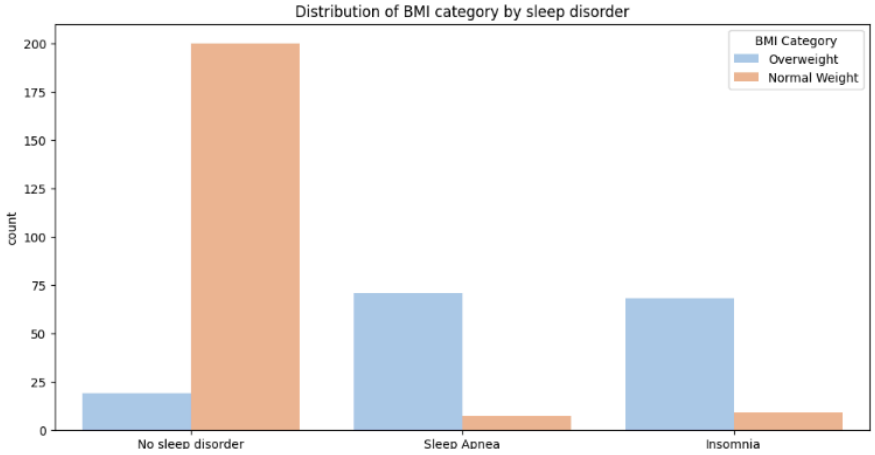
We examined the relationship between blood pressure quality and BMI categories.

BMI Category	Sleep Disorder	count
Normal Weight	No sleep disorder	200
	Insomnia	9
	Sleep Apnea	7
Overweight	Sleep Apnea	71
	Insomnia	68
	No sleep disorder	19

Code :

```
BMI_count_d = df.groupby("BMI Category")["Sleep Disorder"].value_counts()
```

The data suggest that the BMI category affects the prevalence of sleep disorders, as people with a normal weight tend to have fewer sleep issues compared to overweight individuals.



Code:

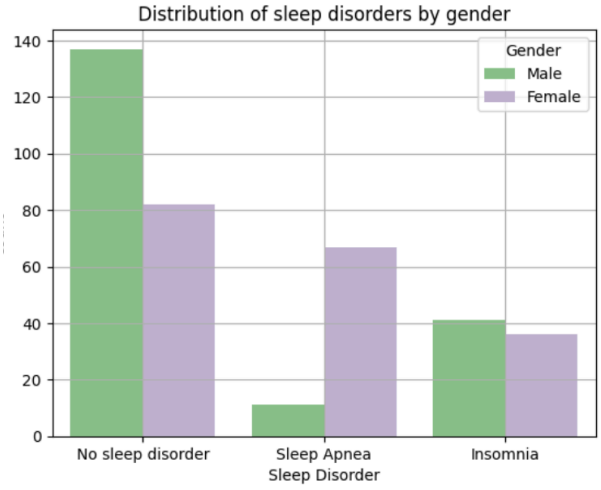
```
plt.figure(figsize=(12, 6))
sns.countplot(data=df, x="Sleep Disorder", hue="BMI Category", palette="pastel")
plt.title("Distribution of BMI category by sleep disorder")
```

Sleep disorders by gender

We created an histogram using Seaborn that shows sleep disorders sorted by gender.

Out of the 374 individuals in the dataset, 219 (58%) do not suffer from any sleep disorder, while 78 (21%) are affected by sleep apnea and 77 (21%) by insomnia.

Sleep Disorder	count
No sleep disorder	219
Sleep Apnea	78
Insomnia	77



Code :

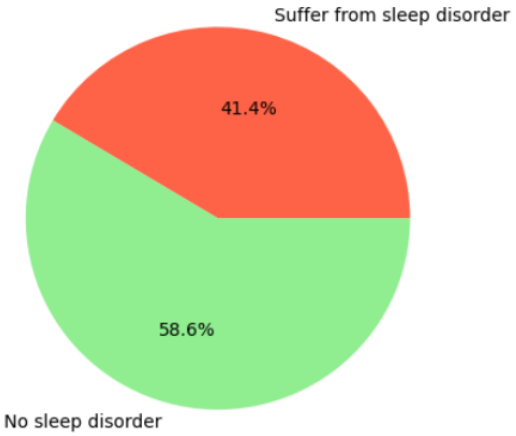
```
sns.countplot(df, x="Sleep Disorder", hue = "Gender", palette = "Accent")
plt.grid()
plt.title("Distribution of sleep disorders by gender")
```

Approximately 140 out of 189 men do not suffer from any sleep disorder, compared to 80 out of 185 women. From this, we can infer that women are more prone to sleep disorders than men.

Overall, approximately 16% of men suffer from either insomnia or sleep apnea, compared to 33% of women.

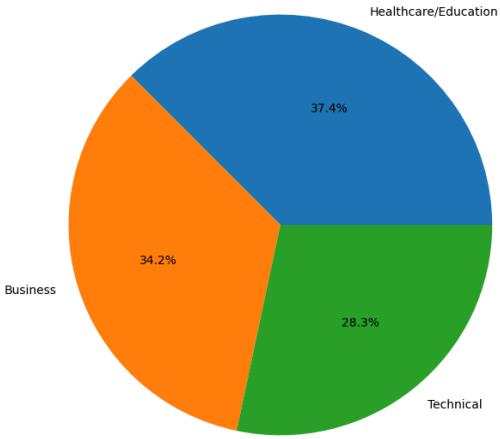
In general, men are more affected by insomnia than by sleep apnea (11% versus 3%).

This means that around 42% of the population is affected by a sleep disorder.



Occupation vs sleep disorders

We observe that occupation influences the prevalence of sleep disorders. For example, employees in the healthcare/educational sector are significantly more prone to sleep apnea (88%) compared to other occupational groups.
However, despite this high rate, around 37% of individuals in this category report no sleep disorders.
It is also important to consider the sample size of each group: the dataset contains a larger number of individuals in healthcare and education roles. This distribution helps explain why some categories may show high rates of certain disorders, yet still include a substantial number of unaffected individuals.



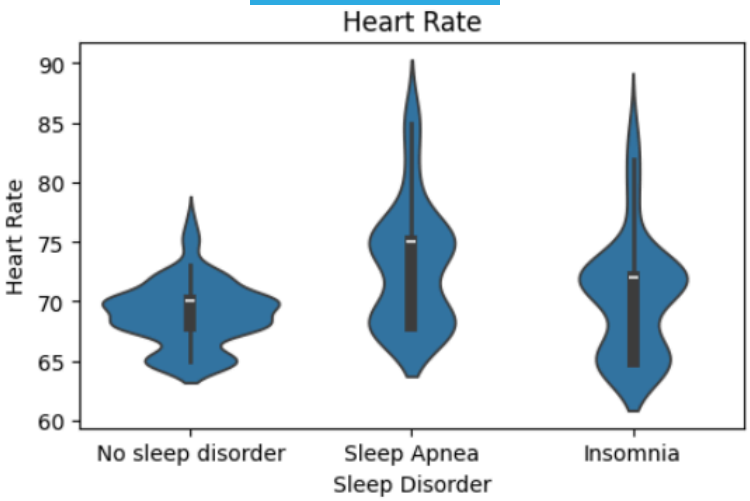
We also identified other correlations with sleep disorders, such as stress levels, highlighting the complex and multifactorial nature of sleep issues as revealed through our data visualization.

Correlations through data visualization

We created a correlation matrix to explore the relationships between various features and sleep disorders.
The matrix highlights several strong correlations. For instance, quality of sleep and sleep duration are highly positively correlated (0.88), indicating that longer sleep is generally associated with better sleep quality. In contrast, stress level shows strong negative correlations with both sleep duration (-0.81) and sleep quality (-0.90), confirming that stress is a major factor affecting sleep.
We also observed that heart rate is negatively correlated with sleep quality (-0.66), while BMI category (overweight) is positively correlated with sleep apnea (0.51). These associations are consistent with known medical insights.
Gender also appears to influence several factors; for example, being female is

negatively correlated with stress level (-0.40) and positively correlated with being overweight (0.31). Overall, the correlation matrix supports the idea that stress, BMI, and heart rate are key factors linked to sleep disorders, while occupation and gender show more nuanced patterns.

For heart rate for instance this graph illustrates that elevated heart rates are more common among individuals with sleep disorders



IV. Machine Learning Model (Training/ Evaluation Metrics)

To predict whether a patient suffers from a sleep disorder, we trained a supervised machine learning model using the Keras API. The input features included *Age*, *Sleep Duration*, *Quality of Sleep*, *Stress Level*, *Heart Rate,and Physical Activity Level,* which were selected for their clinical relevance and the observations we made on the correlation matrix. We didn't include the feature called "BMI Category_Overweight" since it was affecting our model performance. Although it has a strong correlation with sleeping disorders, we believe that its inclusion in our model could create overfitting.

The target variable '*Sleep Disorder*' was label-encoded into three classes: '*no disorder'*, '*insomnia'*, and '*sleep apnea'*. One-hot encoding was applied to the target labels to fit the softmax output of the model.

```
le = LabelEncoder()
df["Sleep Disorder numeric"] = le.fit_transform(df["Sleep Disorder"])
```

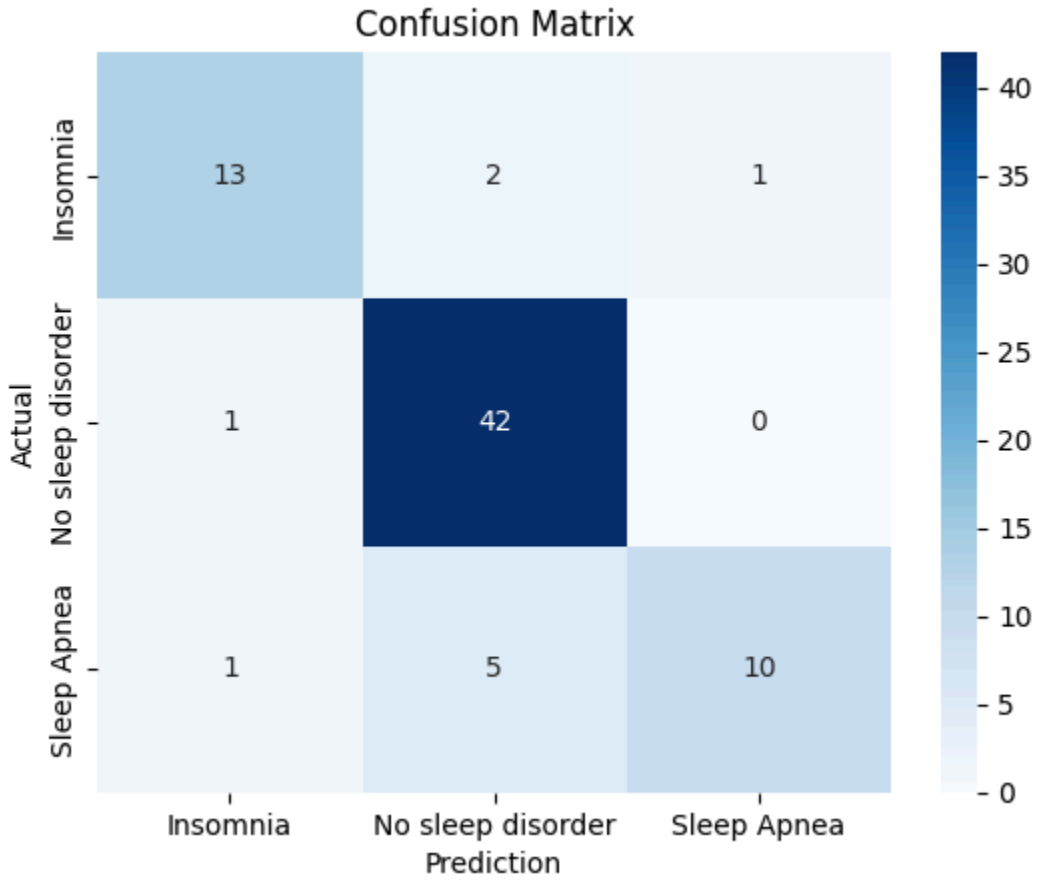
We used a simple neural network architecture consisting of an input layer, one hidden dense layer with 64 units and *ReLU* activation, and an output layer with softmax activation to handle multi-class classification. The model was compiled with the categorical cross-entropy loss function and the Adam optimizer. Early stopping was implemented to prevent overfitting.

The dataset was split into training and testing sets, and the model achieved an accuracy of 0.880, a precision of 0.878, a recall of 0.880, and an F1-score of 0.876 (weighted average), indicating strong and consistent performance. These numbers could change when the code is executed several times.

Our model demonstrates very strong overall performance (above 85% across all metrics), indicating it is reliable for predicting sleep disorders within our dataset.

A confusion matrix was plotted to visualize classification performance across all classes, along with a learning curve to monitor training and validation accuracy over epochs.

```
conf_mat = confusion_matrix(y_test, y_pred)
```



The confusion matrix shows that the model performs well overall, especially in detecting individuals with no sleep disorders, with 42 correct predictions out of 43. For insomnia, the model correctly identified 13 out of 16 cases, and for sleep apnea, it correctly classified 10 out of 16 cases.

V. Responsible AI Practices

Integration of responsible AI practices

Our implementation incorporates several fundamental principles of responsible AI, including fairness, transparency, and reliability.

To promote fairness, we analyzed class distribution within sensitive features such as occupation, and acknowledged imbalances (a higher number of healthcare/education professionals compared to technical and business roles). This helped us remain aware of potential representation bias when interpreting model performance across groups.

To preserve interpretability, we retained the original class labels throughout the classification process.

In evaluating model performance, we did not rely solely on accuracy but also included additional metrics such as precision, recall, and F1-score.

This multi-metric evaluation allowed us to assess different dimensions of prediction quality, including false positives and false negatives.

Furthermore, we implemented an early stopping mechanism with validation monitoring, which helped prevent overfitting and ensured the model maintained robustness when exposed to new, unseen data.

Opportunities for ethical improvement

While our current setup demonstrates an initial commitment to ethical AI practices, several enhancements could further align the model with advanced responsible AI standards.

First, to protect sensitive medical data, implementing differential privacy (such as TensorFlow Privacy's differentially private optimizers) would help ensure strong privacy guarantees for individuals in the dataset.

Additionally, incorporating bias detection metrics across demographic groups (age, gender) could help identify and address any disparities in model performance.

These enhancements would strengthen the system's fairness, privacy, and trustworthiness, especially in clinical or socially sensitive applications.