Universidade de Coimbra

Machine Learning - Challenge Report

Sleep Health and Lifestyle

Miguel Meireles Teixeira

Ângelo da Rocha Rodrigues

Class: PL1
Head Teacher: Catarina Helena Branco Simões da Silva

2023/2024

Contents

1 Introduction: Exploring Sleep Patterns and Lifestyle Concerning Health

As part of the Machine Learning Course, we selected a Machine Learning case study and verified the ability of some simple models to solve it.

The chosen case study will explore sleep patterns and their overall contribution to health.

Sleep is an essential element for maintaining the balance of our human system. During sleep, the brain can rest. Therefore, understanding sleep patterns and lifestyles is a crucial step in preventing diseases that may appear.

The team members chose to study this topic because they also experienced sleep problems and wanted to verify what caused them and how to solve them.

Problem: Sleep is a crucial aspect of health. However, most of the population does not rest sufficiently, which leads to discomfort and health problems.

In fact, according to the National Sleep Foundation of the United States, 2/3 of the world's population suffers from some degree of insomnia, and according to the Portuguese Sleep Association, in Portugal, half of the population has experienced this type of sleep disorder.

Solution: Using Machine Learning models, we will identify the main factors that lead to sleep disorders, classifying sleep diseases as Insomnia or Sleep Apnea before they appear.

Objective: The aim is to predict sleep health problems, such as Insomnia or Sleep Apnea and relate them to lifestyle.

To fulfil the objective, we used the following dataset Sleep Health and Lifestyle Dataset, as it can help in the prevention of sleep disorders.

Throughout this report, we will explore the data provided by the Sleep Health and Lifestyle Dataset, understanding the importance of sleep and how a healthy lifestyle affects the human system.

2 Dataset Overview: Understanding Sleep and Lifestyle for Health Dataset

The Dataset has 374 rows and 14 columns, providing information about the correlation between sleep and lifestyle.

Dataset columns:

- **Person ID:** An identifier for each individual.
- **Gender:** The gender of the person (Male/Female).
- **Age:** The age of the person in years.
- **Occupation:** The occupation or profession of the person.
- **Sleep Duration:** The number of hours the person sleeps per day.
- **Quality of Sleep:** A subjective rating of sleep quality, ranging from 1 to 10.
- **Physical Activity Level:** The number of minutes the person engages in physical activity daily.
- **Stress Level:** 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:** The blood pressure measurement of the person, is indicated as systolic pressure over diastolic pressure.
- **Heart Rate:** The person's resting heart rate in beats per minute.
- **Daily Steps:** The number of daily steps the person takes.
- **Sleep Disorder:** The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).

Dataset columns variable types:

Table 1: Dataset Information

Column	Null Values	Dtype
Person ID	0	int64
Gender	0	object
Age	0	int64
Occupation	0	object
Sleep Duration	0	float64
Quality of Sleep	0	int64
Physical Activity Level	0	int64
Stress Level	0	int64
BMI Category	0	object
Blood Pressure	0	int64
Heart Rate	0	int64
Daily Steps	0	object
Sleep Disorder	0	int64

According to Table 1, there are no null values in the Dataset, so each column has 374 values.

Through Table 1, it is possible to notice three variable types: int64, float64 and object. However, as shown in section A we will perform a conversion from Categorical to Integer variables.

Dataset Statics Information:

	Person ID	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate	Daily Steps
count	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000
mean	187.500000	42.184492	7.132086	7.312834	59.171123	5.385027	70.165775	6816.844920
std	108.108742	8.673133	0.795657	1.196956	20.830804	1.774526	4.135676	1617.915679
min	1.000000	27.000000	5.800000	4.000000	30.000000	3.000000	65.000000	3000.000000
25%	94.250000	35.250000	6.400000	6.000000	45.000000	4.000000	68.000000	5600.000000
50%	187.500000	43.000000	7.200000	7.000000	60.000000	5.000000	70.000000	7000.000000
75%	280.750000	50.000000	7.800000	8.000000	75.000000	7.000000	72.000000	8000.000000
max	374.000000	59.000000	8.500000	9.000000	90.000000	8.000000	86.000000	10000.000000

Figure 1: Dataset descriptive statics

Checking the Dataset Balance:
The Dataset is unbalanced, presenting:

Table 2: Sleep Disorder Value Counts

Sleep Disorder	**Value Count**
None	219
Sleep Apnea	78
Insomnia	77

According to Figure 2, we are dealing with an Unbalanced Dataset with more "None" data. However, as shown further in the report, we will present techniques for balancing the dataset: Oversampling and Undersampling.

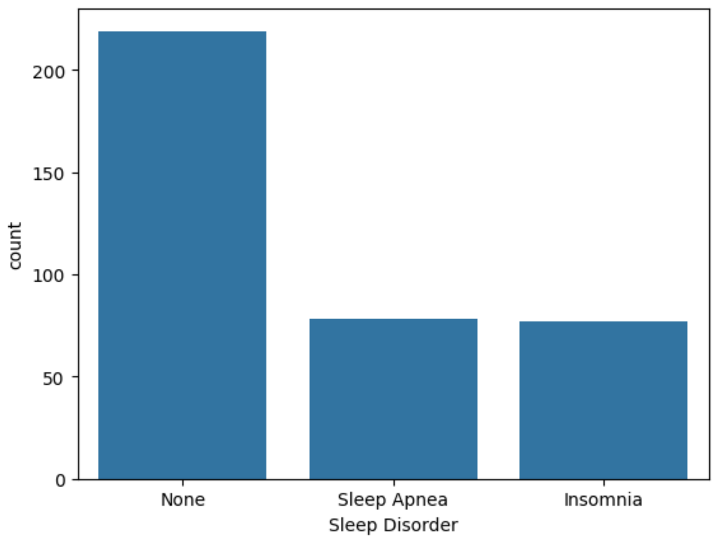

Figure 2: Unbalance Dataset

3 [EDA] - Exploration Data Analysis

3.1 Clean, Prepare and Manipulate Data

In this section, we will identify missing data, process duplicate rows, inconsistent or incorrect data, and standardize data formats.

3.1.1 Data Cleaning

Missing data: As shown in Table 1, "Null Values" are zero. Therefore, the Dataset does not have missing data.

Duplicate rows: There are no duplicate rows.

Creating new columns: We decided to split the *Blood Pressure* column into two new columns: *Systolic Blood Pressure* and *Diastolic Blood Pressure* as it would provide us with more data to correlate with the target class *Sleep Disorder*.

Categorical to Integer transformation: To be able to manipulate our Dataset and extract all the information possible, we decided to create a new Dataframe using (class sklearn.preprocessing.LabelEncoder) since we can encode and transform categorical values into numerical values, as shown in Figure 3.

	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Daily Steps	Sleep Disorder	Systolic Blood Pressure	Diastolic Blood Pressure
0	1	27	9	6.1	6	42	6	3	77	4200	1	126	83
1	1	28	1	6.2	6	60	8	0	75	10000	1	125	80
2	1	28	1	6.2	6	60	8	0	75	10000	1	125	80
3	1	28	6	5.9	4	30	8	2	85	3000	2	140	90
4	1	28	6	5.9	4	30	8	2	85	3000	2	140	90

Figure 3: New Dataframe without categorical data

3.1.2 Univariate Visualization

Person ID: *Person ID* is an identifier for each individual. Therefore, his contribution to *Sleep Disorder* is almost none. The *Person ID* column is not worth studying.

Gender: As shown in Figure 4, the Dataset presents 374 study cases, where 189 are males and 185 are females. Therefore, there is an almost equal division.

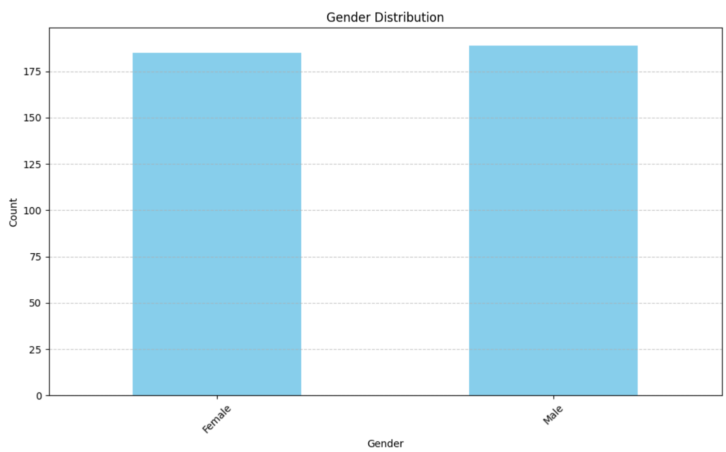

Figure 4: Gender Distribution

As shown in Table 3, of the 189 males, 52 have sleep disorders. However, of the 185 females, 103 suffer from sleep disorders.

Sleep Disorder	**Gender**	**Count**
Insomnia	Male	41
	Female	36
None	Male	137
	Female	82
Sleep Apnea	Female	67
	Male	11

Table 3: Group *Gender* by *Sleep Disorder*

Age: As shown in Figure 5, the Dataset only presents information on individuals between 27 and 59 years old, so we do not have any information on young or older people.

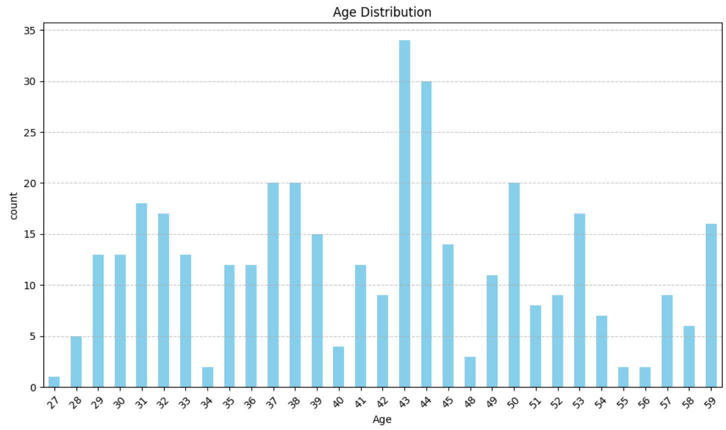

Figure 5: Age Distribution

Sleep Disorder	Age mean
Insomnia	43.52
None	39.04
Sleep Apnea	49.71

Table 4: Group *Age* mean by *Sleep Disorder*

Occupation: As shown in Figure 6, *Occupation* has 11 professions.
The model effectiveness will be later evaluated if we incorporate *Software Engineer* in *Enginner* and *Sales Representative* in *Salesperson*.

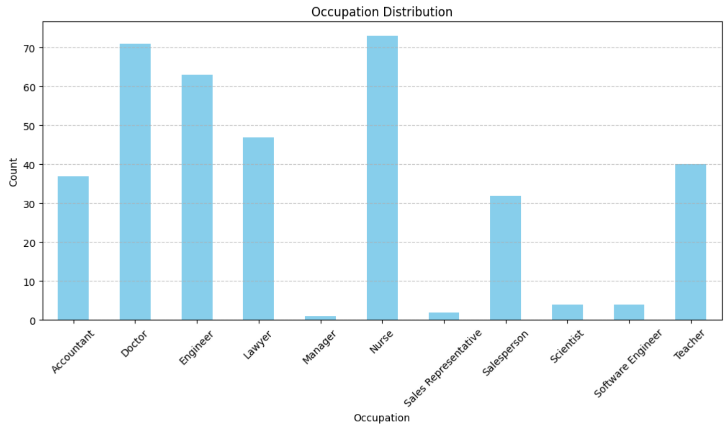

Figure 6: Heart Rate Distribution

As shown in Table 5, *Nurse* is an occupation that suffers a lot from *Sleep Apnea*.

Sleep Disorder	**Occupation**	**Count**
Insomnia	Salesperson	29
	Teacher	27
	Accountant	7
	Engineer	5
	Doctor	3
	Nurse	3
	Lawyer	2
	Software Engineer	1
None	Doctor	64
	Engineer	57
	Lawyer	42
	Accountant	30
	Nurse	9
	Teacher	9
	Software Engineer	3
	Salesperson	2
	Scientist	2
	Manager	1
Sleep Apnea	Nurse	61
	Doctor	4
	Teacher	4
	Lawyer	3
	Sales Representative	2
	Scientist	2
	Engineer	1
	Salesperson	1

Table 5: Group Occupation by Sleep Disorder

Sleep Duration: As shown in Figure 7, people sleep between 5.8 and 8.5 hours. Therefore, we can apply a normalization and test if the model performance increases.

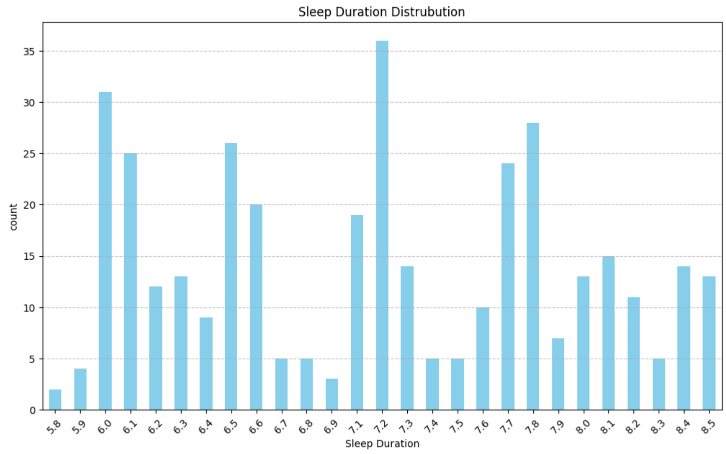

Figure 7: Sleep Duration Distribution

As shown in Table 6, the mean *Sleep Duration* is around 7 hours. According to the Sleep National Foundation, a healthy adult should sleep between 7 and 9 hours.

Sleep Disorder	**Mean Sleep Duration (hours)**
Insomnia	6.59
None	7.36
Sleep Apnea	7.03

Table 6: Group *Sleep Duration* mean by *Sleep Disorder*

Quality of Sleep: As shown in Figure 8, most people rate their sleep as positive.

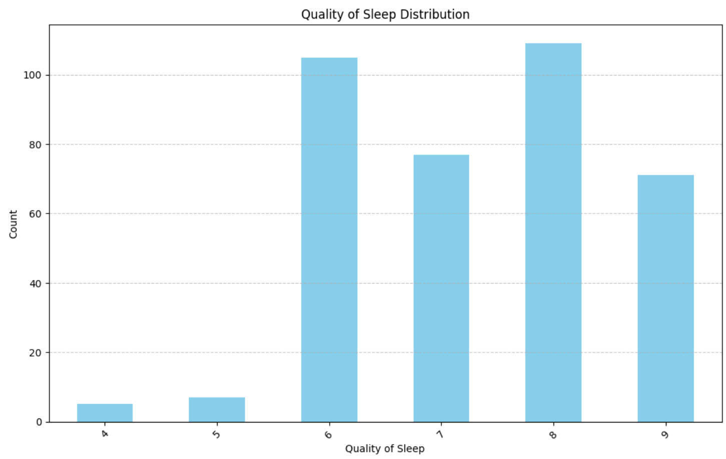

Figure 8: Quality of Sleep Distribution

As shown in Table 7, people with a better *Quality of Sleep* tend to have no *Sleep Disorder*.

Sleep Disorder	Mean Quality of Sleep
Insomnia	6.53
None	7.63
Sleep Apnea	7.21

Table 7: Group Quality of Sleep mean by Sleep Disorder

Physical Activity Level: As shown in Figure 9, all people practice physical activity, with a minimum of 30 minutes per day and a maximum of 90 minutes. In fact, according to the Portuguese National Health Service, a healthy adult should practice 60 minutes of physical activity per day.

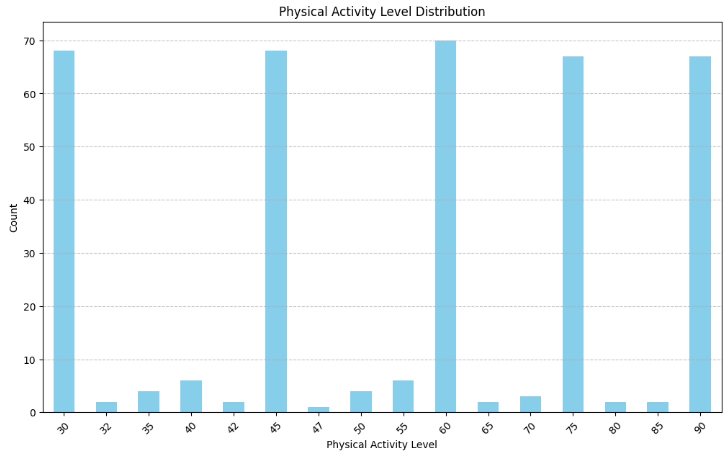

Figure 9: Physical Activity Level Distribution

As shown in Table 8, lack of physical activity is associated with *Insomnia*.

Sleep Disorder	Mean Physical Activity Level (minutes)
Insomnia	46.82
None	57.95
Sleep Apnea	74.79

Table 8: Group *Physical Activity Level* mean by *Sleep Disorder*

Stress Level: As shown in Figure 10, most people rate their stress level as high.

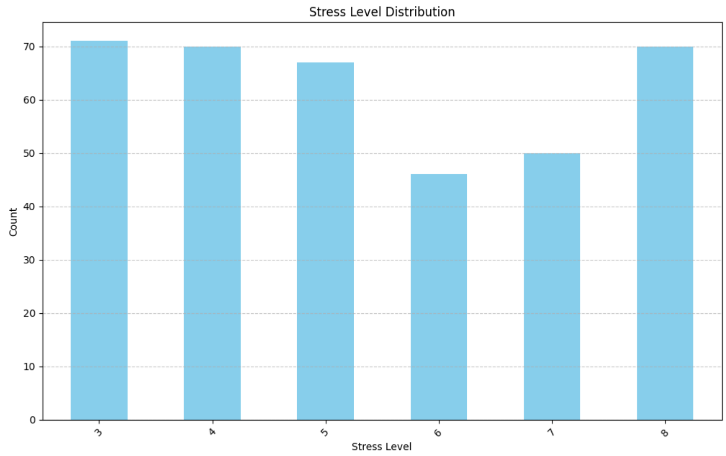

Figure 10: Stress Level Distribution

As shown in Table 9, regardless of the *Sleep Disorder*, people have a high *Stress Level*.

Sleep Disorder	Mean Stress Level
Insomnia	5.87
None	5.11
Sleep Apnea	5.67

Table 9: Group Stress Level mean by Sleep Disorder

BMI Category: As shown in Figure 11, *BMI Category* has 4 categories.

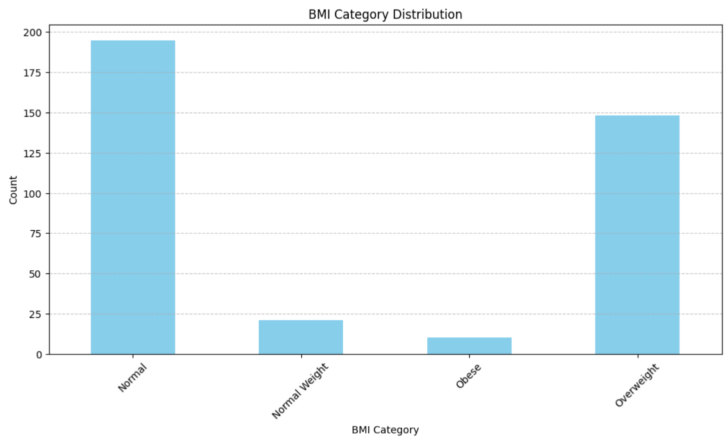

Figure 11: BMI Category Distribution

As shown in Table 10, *Overweight* people develop more sleep disorders.

Sleep Disorder	BMI Category	Count
Insomnia	Overweight	64
	Normal	7
	Obese	4
	Normal Weight	2
None	Normal	183
	Overweight	19
	Normal Weight	17
Sleep Apnea	Overweight	65
	Obese	6
	Normal	5
	Normal Weight	2

Table 10: Group BMI Category by Sleep Disorder

Blood Pressure: As shown in Table 11, low *Diastolic Blood Pressure* is associated with no sleep disorders.

Sleep Disorder	Diastolic Blood Pressure Mean
Insomnia	86.86
None	81.00
Sleep Apnea	92.72

Table 11: Group Diastolic Blood Pressure mean by Sleep Disorder

Similarly, Table 12 presents that low *Systolic Blood Pressure* is associated with no sleep disorders.

Sleep Disorder	Systolic Blood Pressure Mean
Insomnia	132.04
None	124.05
Sleep Apnea	137.77

Table 12: Group Systolic Blood Pressure mean by Sleep Disorder

Heart Rate: As shown in Figure 12, all individuals have a heart rate between 60 and 100 bpm, which is the average for the age group studied in this dataset.

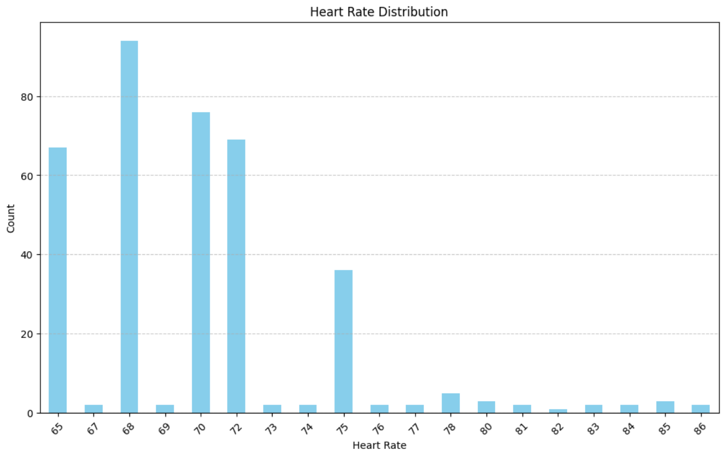

Figure 12: Heart Rate Distribution

As shown in Table 13, the mean heart rate for all values in *Sleep Disorder* is around 70 beats per minute.

Sleep Disorder	**Heart Rate mean**
Insomnia	70.47
None	69.02
Sleep Apnea	73.09

Table 13: Group *Heart Rate* mean by *Sleep Disorder*

Daily Steps: As shown in Figure 13, most of the people walk more than 5000 steps per day.

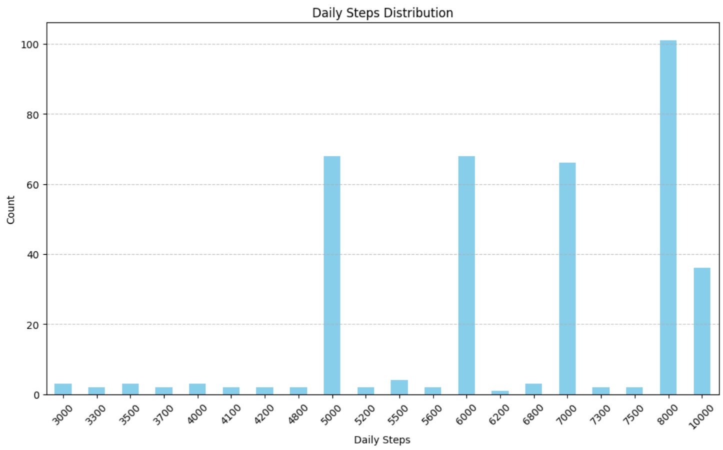

Figure 13: Daily Steps Distribution

As shown in Table 14, for all values in *Sleep Disorder*, the *Daily Steps* mean also exceeds 5000 steps per day.

Sleep Disorder	Daily Steps mean
Insomnia	5901.30
None	6852.97
Sleep Apnea	7619.23

Table 14: Group *Daily Steps* mean by *Sleep Disorder*

Sleep Disorder: As shown in Figure 14, the Dataset only presents information on individuals between 27 and 59 years old, so we do not have any information on young or older people.

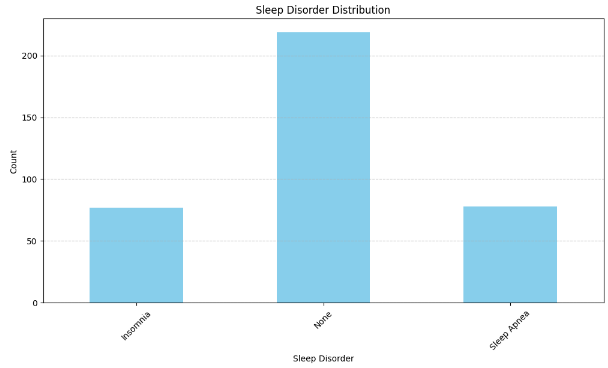

Figure 14: Sleep Disorder Distribution

3.1.3 Multivariate Visualization

Gender vs Occupation: As shown in Figure 15, most Females work as Nurses and have a higher risk of *Sleep Apnea*, while most Males work as Doctors and have a lower risk of *Sleep Disorder*.

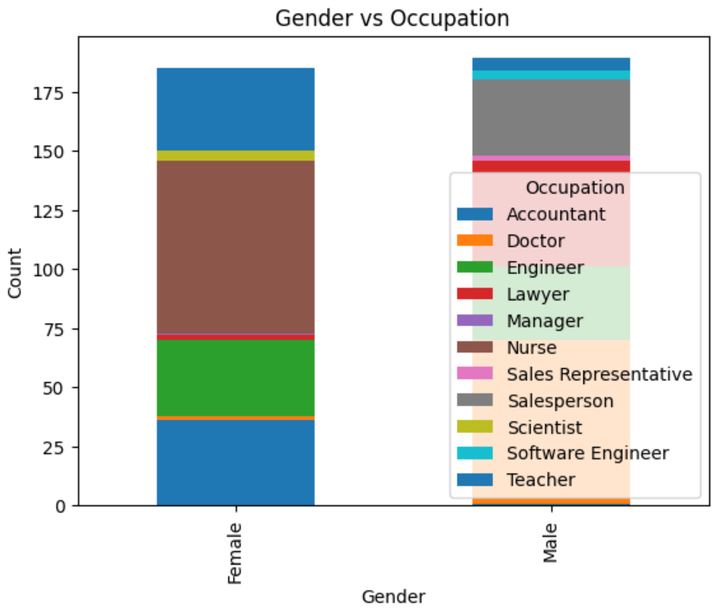

Figure 15: Gender Vs Occupation

Correlation Matrix:
To identify the most correlated classes with our Target Class, we made a Correlation Matrix as

shown in Figure 16. The correlation margin with *Sleep Disorder* must be at least 18%.

As a result, the classes *Occupation*, *Stress Level* and *BMI Category* will be removed because they are the ones that have a lower correlation with the *Sleep Disorder*.

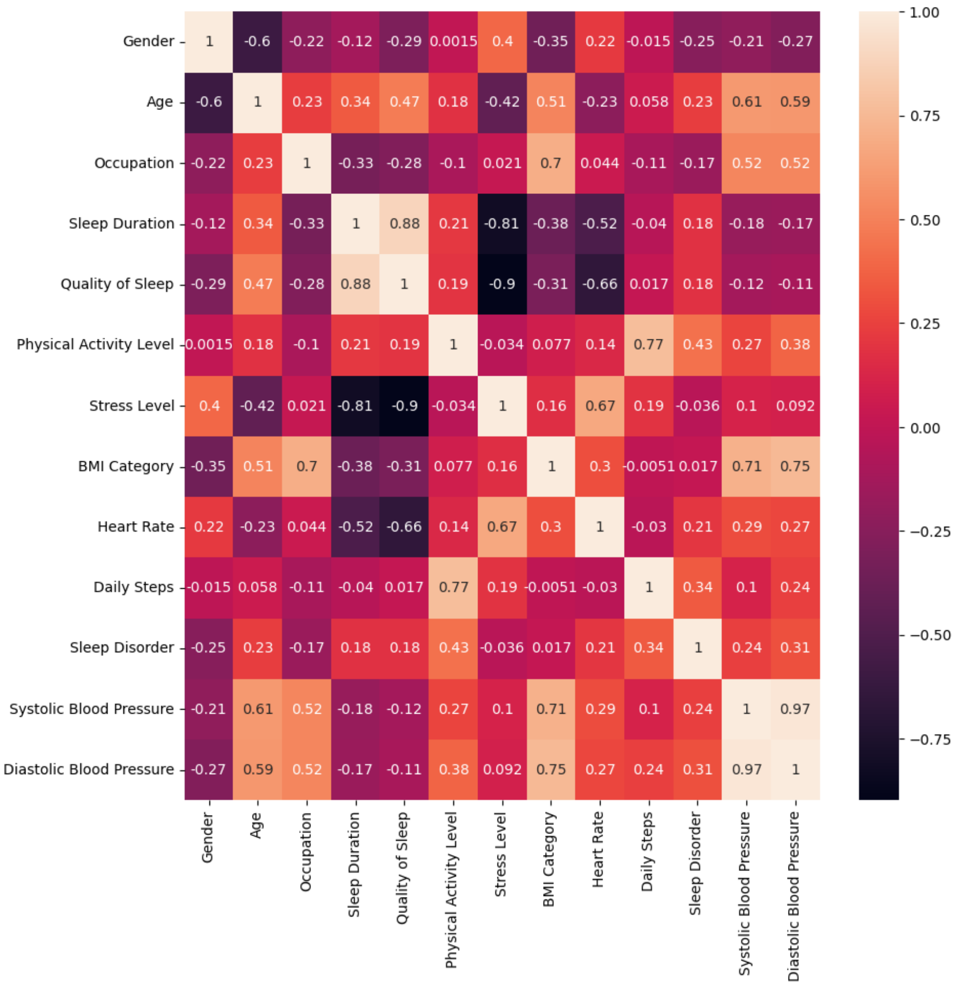

Figure 16: Correlation Matrix

Pairplot plot:
As shown in Figure 17, to identify the most correlated classes we also used a Pairplot plot, which shows the relationship of the different classes with the Sleep Disorder and performs a univariate analysis.

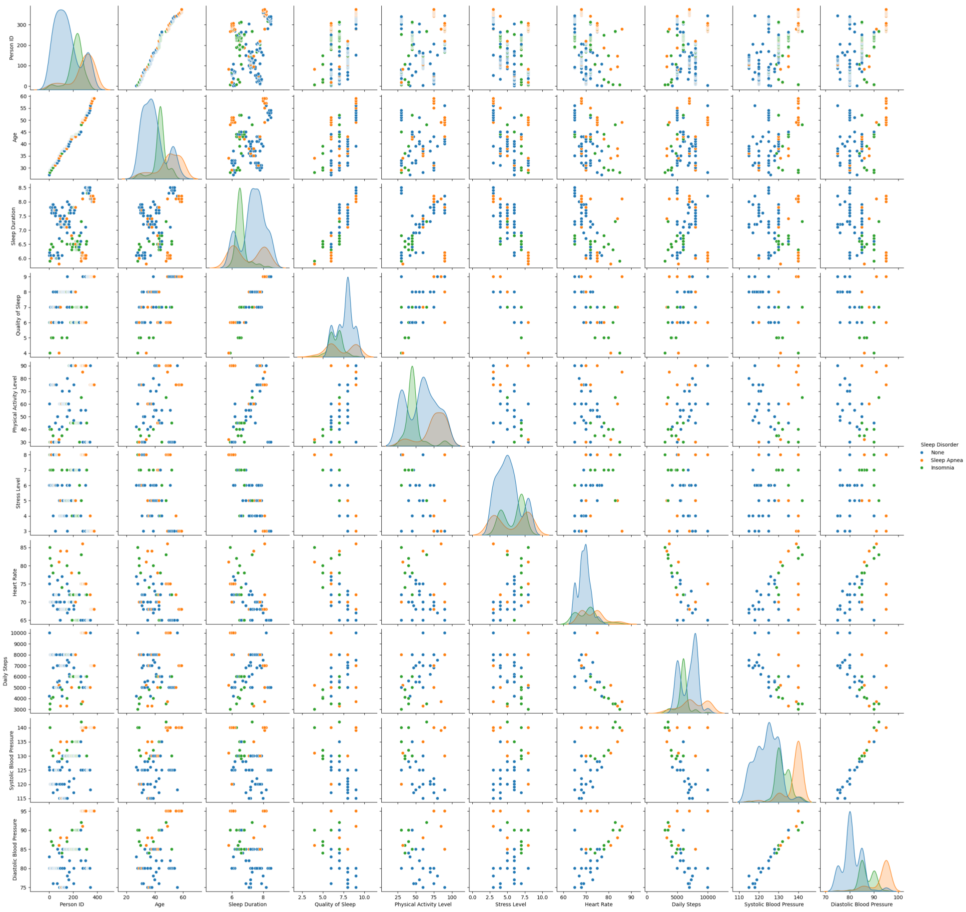

Figure 17: Pairplot plot

3.1.4 Data Manipulation

In this section, we will create and manipulate our data to use in training and testing our Machine Learning Models.

Remove columns with lower correlation: Once the columns with lower correlation have been identified, we proceed to create:

- **X_main:** Input array containing all columns except *Occupation*, *Stress Level*, *BMI Category*, and *Sleep Disorder*.

	Gender	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Heart Rate	Daily Steps	Systolic Blood Pressure	Diastolic Blood Pressure
0	1.0	27	0.111111	0.4	0.20	0.571429	0.171429	0.407407	0.40
1	1.0	28	0.148148	0.4	0.50	0.476190	1.000000	0.370370	0.25
2	1.0	28	0.148148	0.4	0.50	0.476190	1.000000	0.370370	0.25
3	1.0	28	0.037037	0.0	0.00	0.952381	0.000000	0.925926	0.75
4	1.0	28	0.037037	0.0	0.00	0.952381	0.000000	0.925926	0.75
...	...	...	...	...	...	...	...	...	...
369	0.0	59	0.851852	1.0	0.75	0.142857	0.571429	0.925926	1.00
370	0.0	59	0.814815	1.0	0.75	0.142857	0.571429	0.925926	1.00
371	0.0	59	0.851852	1.0	0.75	0.142857	0.571429	0.925926	1.00
372	0.0	59	0.851852	1.0	0.75	0.142857	0.571429	0.925926	1.00
373	0.0	59	0.851852	1.0	0.75	0.142857	0.571429	0.925926	1.00

Figure 18: X_main Array normalized

- **y_main:** Output array containing the *Sleep Disorder* column.

Normalize data: We normalize the *X_main* columns to improve our model.

Split the data: We split X_main and y_main into training and testing sets.
We used 75% of the data for training and 25% of the data for testing.

Oversampling: As we have much more *None* data than *Sleep Apnea* or *Insomnia* in the target class *Sleep Disorder*, we decided to use Oversampling in the tests and check how our model behaved.

However, to avoid the model interpreting the probabilities of the 3 sets *None*, *Sleep Apnea*, or *Insomnia* as equal when using Oversampling, we decided to use three types of oversampling:

- RandomOverSampler
- SMOTEENN
- ADASYN

However, the Baseline will be performed without Oversampling.

Undersampling: As we have much more *None* data than *Sleep Apnea* or *Insomnia* in the target class *Sleep Disorder*, we decided to use Undersampling in the tests and check how our model behaved.

However, to make sure we do not remove important information from our model, we decided to use:

- RandomUnderSampler
- NearMiss
- Edited Nearest Neighbors (ENN):

The last undersampling method removes badly classified examples by k-nearest neighbours (KNN), preventing us from eliminating pertinent information.

After Oversampling or Undersampling we concatenate the array X_res with Y_res and store it in a new Dataframe, where we will evaluate the balance of the Dataset and the new Correlation Matrix.

4 Train Model, Test Data and Improve Model

4.1 [DTC] - Decision Tree Classifier

4.1.1 Model Selection

We start by using the GridSearch CV to find the best parameters for us to put in our DecisionTreeClassifier().

In fact, the values given were as follows:

$$'criterion' :' entropy',' max\_depth' : 100,' splitter' :' best'$$

We also noticed that if we wanted to save computational power, then we could decrease the *depth* of the *Decision Tree*. However, we found cases of Overfitting.

We start with a max_deth=25. However, for the model it is more important to have a correct answer and then set the maximum depth to 100 even if it requires extra computational power.

Best Expriment: The best Experience was Exp 3.

In fact, we are looking for a high Recall and although we have to sacrifice a bit of precision it is important to have fewer false negatives.

It is crucial that the model correctly identifies the majority of positive cases.

4.1.2 Models Evaluation

Experiment	Observations	Accuracy Train | Test	F1	Recall	Precision
Exp 1	No Correlation Criteria GridSearchCV best parameters Normal Sampling	0.94 | 0.84	**0:** 0.84 **1:** 0.88 **2:** 0.74	0.76 0.84 0.95	0.94 0.94 0.61
Exp 2	Correlation Criteria of 18%	0.94 | 0.85	**0:** 0.84 **1:** 0.91 **2:** 0.73	0.76 0.87 0.89	0.94 0.94 0.62
Exp 3	Oversampling - SMOTEENN	1.0 | 0.91	**0:** 0.87 **1:** 0.95 **2:** 0.84	0.81 0.96 0.89	0.94 0.95 0.80
Exp 4	Oversampling - ADASYN	0.96 | 0.83	**0:** 0.80 **1:** 0.87 **2:** 0.76	0.76 0.82 0.94	0.84 0.94 0.63
Exp 5	Oversampling - RandomOverSampler	0.93 | 0.82	**0:** 0.78 **1:** 0.87 **2:** 0.74	0.76 0.80 0.94	0.80 0.96 0.61
Exp 6	Undersampling - EditedNearestNeighbours	0.98 | 0.89	**0:** 0.83 **1:** 0.93 **2:** 0.86	0.81 0.93 0.89	0.85 0.93 0.84
Exp 7	Undersampling - RandomUnderSampler	0.96 | 0.81	**0:** 0.74 **1:** 0.86 **2:** 0.63	0.76 0.78 0.94	0.73 0.96 0.76
Exp 8	Undersampling - NearMiss	0.91 | 0.59	**0:** 0.71 **1:** 0.57 **2:** 0.50	0.74 0.42 0.94	0.68 0.92 0.34

Table 15: Summary of Experiment Results DT

Experiment	Precision Avg | Std	Recall Avg | Std
Exp 1	**0:** 0.92 | 0.048	0.90 | 0.056
	1: 0.95 | 0.022	0.96 | 0.024
	2: 0.90 | 0.055	0.91 | 0.046
Exp 2	**0:** 0.90 | 0.057	0.87 | 0.063
	0: 0.94 | 0.027	0.96 | 0.024
	0: 0.90 | 0.071	0.89 | 0.06
Exp 3	**0:** 0.92 | 0.047	0.88 | 0.059
	1: 0.95 | 0.025	0.94 | 0.016
	2: 0.92 | 0.049	0.90 | 0.047
Exp 4	**0:** 0.93 | 0.055	0.89 | 0.058
	1: 0.93 | 0.056	0.97 | 0.016
	2: 0.92 | 0.049	0.87 | 0.061
Exp 5	**0:** 0.91 | 0.052	0.88 | 0.063
	1: 0.95 | 0.035	0.96 | 0.020
	2: 0.90 | 0.068	0.89 | 0.059
Exp 6	**0:** 0.92 | 0.049	0.90 | 0.060
	1: 0.96 | 0.021	0.97 | 0.015
	2: 0.92 | 0.054	0.90 | 0.063
Exp 7	**0:** 0.92 | 0.048	0.88 | 0.049
	1: 0.95 | 0.034	0.96 | 0.021
	2: 0.90 | 0.057	0.90 | 0.061
Exp 8	**0:** 0.92 | 0.070	0.91 | 0.047
	1: 0.96 | 0.020	0.96 | 0.047
	2: 0.90 | 0.063	0.91 | 0.056

Table 16: Overall Table

4.2 [KNN] - K-Nearest-Neighbours

4.2.1 Model Selection

We start by using the GridSearch CV to find the best parameters for us to put in our KNN.

However, the results were not positive, so we initially started with n_neighbors=5, but if we decrease this same decision number to 4, the accuracy increases and we save computational power avoiding overfitting.

We had also started with weights=distance, but we realized that we would get better results using:

$$'n\_neighbors' : 4,' weights' : uniform,' metric' :' manhattan'$$

Best Expriment: The best Experience was Exp 6.

In fact, we are looking for a high Recall and although we have to sacrifice precision it is important to have fewer false negatives.

It is crucial that the model correctly identifies the majority of positive cases.

Therefore using Undersampling - EditedNearestNeighbours was the best solution to this model.

4.2.2 Models Evaluation

Experiment	Observations	Accuracy Train | Test	F1	Recall	Precision
Exp 1	No Correlation Criteria		**0:** 0.79	0.85	0.73
		0.91 | 0.89	**1:** 0.91	0.90	0.93
	Normal Sampling		**2:** 0.90	0.88	0.93
Exp 2	Correlation Criteria of 18%		**0:** 0.86	0.76	1.0
		0.91 | 0.89	**1:** 0.93	0.95	0.91
			2: 0.82	0.89	0.76
Exp 3	Oversampling - SMOTEENN		**0:** 0.86	0.76	1.0
		1.0 | 0.86	**1:** 0.91	0.89	0.92
			2: 0.74	0.89	0.64
Exp 4	Oversampling - ADASYN		**0:** 0.81	0.81	0.81
		0.91 | 0.79	**1:** 0.81	0.73	0.81
			2: 0.72	0.94	0.59
Exp 5	Oversampling - RandomOverSampler		**0:** 0.81	0.81	0.81
		0.91 | 0.80	**1:** 0.83	0.75	0.93
			2: 0.72	0.94	0.59
Exp 6	Undersampling - EditedNearestNeighbours		**0:** 0.88	0.86	0.90
		0.98 | 0.89	**1:** 0.92	0.91	0.93
			2: 0.84	0.89	0.80
Exp 7	Undersampling - RandomUnderSampler		**0:** 0.86	0.76	1.00
		0.87 | 0.88	**1:** 0.93	0.93	0.93
			2: 0.78	0.89	0.70
Exp 8	Undersampling - NearMiss		**0:** 0.79	0.81	0.77
		0.89 | 0.60	**1:** 0.57	0.42	0.88
			2: 0.50	0.89	0.35

Table 17: Summary of Experiment Results KNN