CAPSTONE PROJECT

Project Title: **Prediction of Sleep Disorder**

Abstract:

The goal of this project is to analyze the given data and predict the occurence of Sleep Disorders such as Insomnia and Sleep Apnea. The data was divided into two parts ie. the training and the test dataset. Models were developed based on training dataset and applied to test dataset to find the accuracy of each model based on the predicted values generated. Based on these values we can determine how good the model suits for prediction of Sleep Disorder.

Submitted by

Kavya Girish

Table of Contents

Introduction 3

Summary of Data 4

Transforming Data 5

Treating Missing Data 5

Label Encode categorical variables 5

Data Visualization 6

Data Standardization 10

Steps performed 11

Results 12

Conclusion 13

References 14

Capstone Project - Prediction of Sleep Disorder

Introduction

The 'Sleep Health and Lifestyle' Dataset comprises 374 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.

Number of observations in the given dataset: 374

Dataset

Categorical Variables:

1. **Gender** (Male/ Female)
2. **Occupation** (Accountant/ Doctor/ Engineer/ Lawyer/ Manager/ Nurse/ Sales Representative/ Salesperson/ Scientist/ Software Engineer/ Teacher)
3. **BMI Category** (Obese/ Normal/ Overweight)
4. **Blood Pressure** (Systole/ Diastole)
5. **Sleep Disorder** (None, Insomnia, Sleep Apnea)

Numerical Variables:

1. **Person ID** (unique identifier)
2. **Age**
3. **Sleep Duration** (Duration in hours)
4. **Quality of Sleep** (Range: 1 to 10)
5. **Physical Activity Level** (Duration in minutes)
6. **Stress Level** (Range: 1 to 10)
7. **Heart Rate** (beats per minute)
8. **Daily Steps**

Overview of 'Sleep Disorder'

- **None**: Person does not exhibit any specific sleep disorder
- **Insomnia**: Person experiences difficulty falling asleep or staying asleep, leading to inadequate or poor-quality sleep
- **Sleep Apnea**: Person suffers from pauses in breathing during sleep, resulting in disrupted sleep patterns and potential health risks

...

Summary of Data

Out[24]:

	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

Out[25]:

	Gender	Occupation	BMI Category	Blood Pressure	Sleep Disorder
count	374	374	374	374	374
unique	2	11	4	25	3
top	Male	Nurse	Normal	130/85	None
freq	189	73	195	99	219

- We can see that there are a total of 374 people in the dataset in which 189 are Male and the rest 185 are Females.
- The Age of people in the dataset ranges from minimum of 27 years to a maximum of 59 years old.
- Majority of the people's profession is Nurse. There are around 73 Nurses.
- There are 219 people who do not suffer Sleep Disorders.
- There are 78 people who suffer from a Sleep Disorder called 'Insomnia'.
- There are 77 people who suffer from a Sleep Disorder called 'Sleep Apnea'.
- The lowest duration taken by a particular person to sleep is 5.8 hours, whereas the highest duration taken by a particular person to sleep is 8.5 hours.
- The lowest number of steps taken by a person per day is 3000, whereas the highest number of steps taken by person per day is 10000.
- Stress level of people in the dataset varies from 3 to 8.
- Blood Pressure in the dataset is provided in the measure of Systolic pressure/ Diastolic pressure.
- The BMI Category of people in the dataset includes Normal weight, Overweight and Obese.

Transforming Data

Treating Missing data:

The missing data in any dataset should be treated to ensure accurate analysis.

```
Out[26]: 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           0
         dtype: int64
```

There is no missing data found in the dataset.

Label Encoding for Categorical Variables:

Label Encoding is a technique used to convert Categorical variables into Numerical so that they can be fitted by machine learning models which only take numerical data.

Data before Label Encoding:

Out[45]:

	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
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
5	6	Male	28	Software Engineer	5.9	4	30	8	Obese	140/90	85	3000	Insomnia
6	7	Male	29	Teacher	6.3	6	40	7	Obese	140/90	82	3500	Insomnia
7	8	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
8	9	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
9	10	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None

...

Data after Label Encoding:

Out[47]:

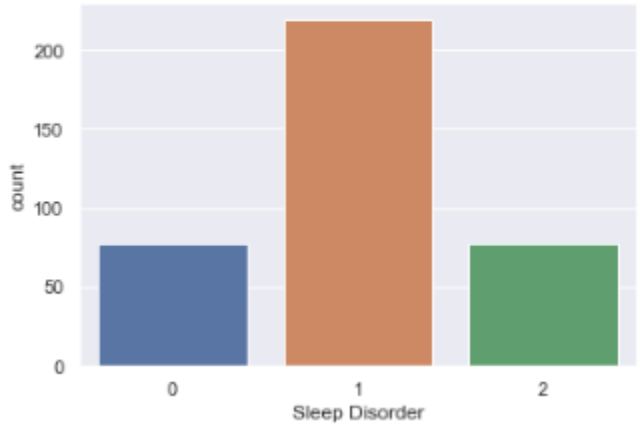
	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	1	27	9	6.1	6	42	6	3	11	77	4200	1
1	2	1	28	1	6.2	6	60	8	0	9	75	10000	1
2	3	1	28	1	6.2	6	60	8	0	9	75	10000	1
3	4	1	28	6	5.9	4	30	8	2	22	85	3000	2
4	5	1	28	6	5.9	4	30	8	2	22	85	3000	2
5	6	1	28	9	5.9	4	30	8	2	22	85	3000	0
6	7	1	29	10	6.3	6	40	7	2	22	82	3500	0
7	8	1	29	1	7.8	7	75	6	0	6	70	8000	1
8	9	1	29	1	7.8	7	75	6	0	6	70	8000	1
9	10	1	29	1	7.8	7	75	6	0	6	70	8000	1

Data Visualization

Class Distribution of Target variable

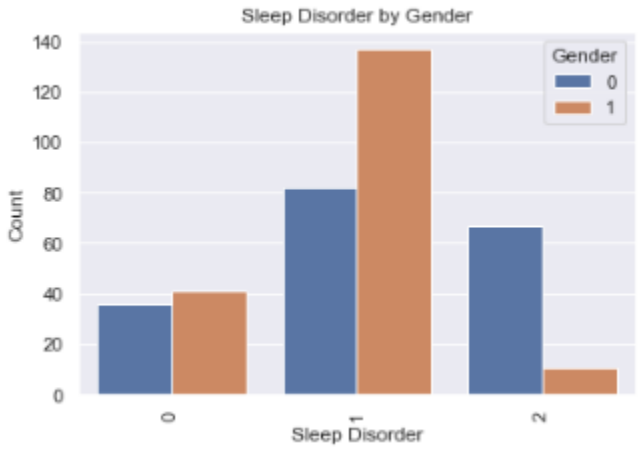
In the given dataset, the target variable is Sleep Disorder.
We will plot a countplot to visualize the distribution of Sleep Disorders suffered by people in the dataset.

```
Count of people not suffering Sleep Disorder:  219
Count of people suffering Insomnia:  78
Count of people suffering Sleep Apnea:  77
```



Multivariate Analysis

Plot to analyze Sleep Disorders based on Gender



Gender:

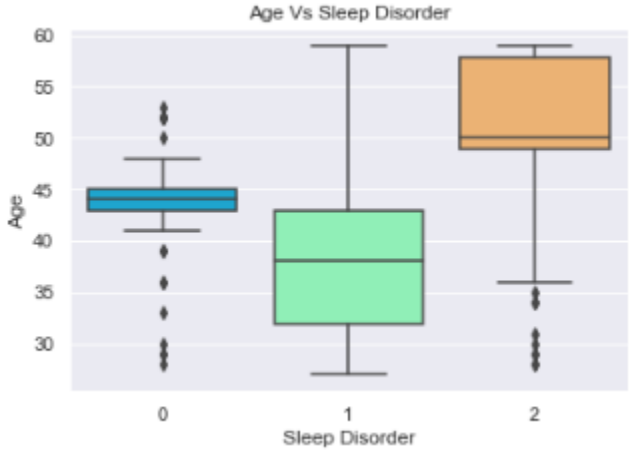
- 0 - Female
- 1 - Male

Sleep Disorder:

- 0 - Insomnia
- 1 - No Sleep Disorder
- 2 - Sleep Apnea

- Most of the people who suffer Insomnia are Male
- Most of the people who suffer Sleep Apnea are Female

Plot to analyze Sleep Disorders based on Gender



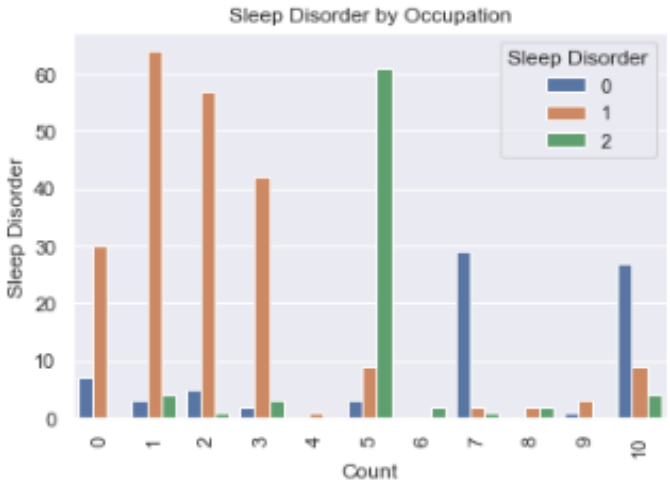
Sleep Disorder:

- 0 - Insomnia
- 1 - No Sleep Disorder
- 2 - Sleep Apnea

- Most people who suffer Sleep Apnea are older than 50

...

Plot to analyze Sleep Disorders based on Occupation



Sleep Disorder:

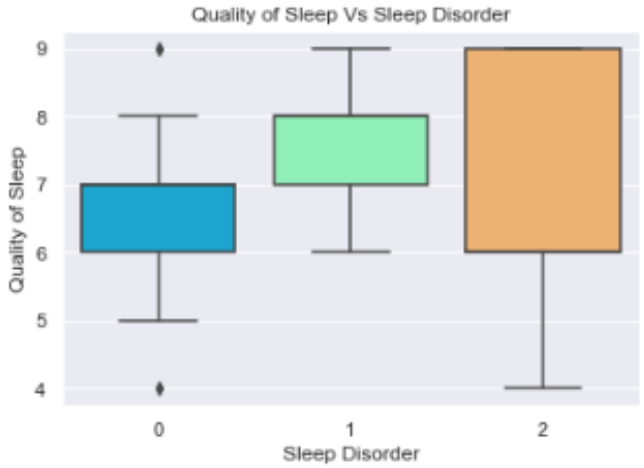
- 0 - Insomnia
- 1 - No Sleep Disorder
- 2 - Sleep Apnea

Occupation:

- 0 - Accountant
- 1 - Doctor
- 2 - Engineer
- 3 - Lawyer
- 4 - Manager
- 5 - Nurse
- 6 - Sales Representative
- 7 - Salesperson
- 8 - Scientist
- 9 - Software Engineer
- 10 - Teacher

- Doctors suffer the highest rate of Insomnia
- More Lawyers, Engineers, Accountants suffer Insomnia
- Nurses suffer highest rate of Sleep Apnea

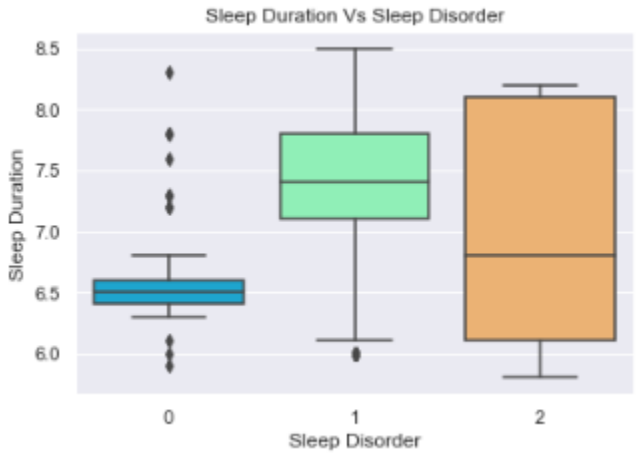
Plot to analyze Sleep Disorders based on Quality of Sleep



Sleep Disorder:

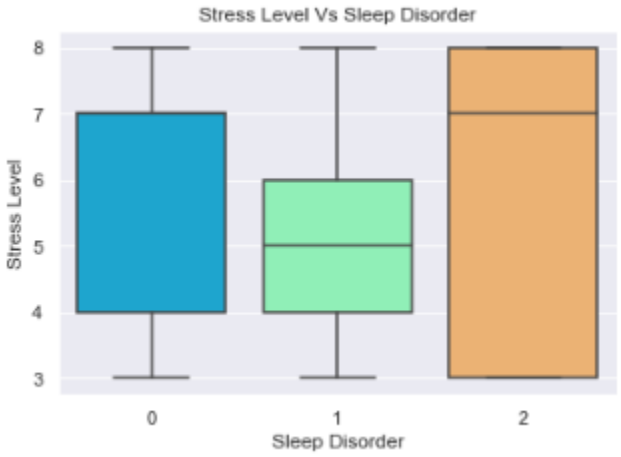
- 0 - Insomnia
- 1 - No Sleep Disorder
- 2 - Sleep Apnea

Plot to analyze Sleep Disorders based on Sleep Duration



- People suffering Insomnia sleep on an average of 6.5 hrs a day
- People suffering Sleep Apnea sleep on an average of 6.9 hrs a day
- People who doesn't suffer Sleep Disorder sleeps on an average of 7.4 hrs a day

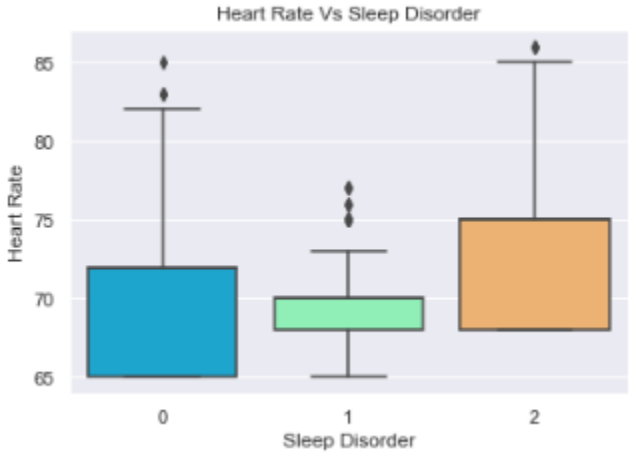
Plot to analyze Sleep Disorders based on Stress Level



- People suffering Sleep Apnea have highest Stress level
- People suffering Insomnia have Stress level higher than Normal

...

Plot to analyze Sleep Disorders based on Heart Rate



- People suffering Insomnia have Heart Rate lower than Normal as well as higher than Normal
- People suffering Sleep Apnea have Heart rate higher than Normal

Data Standardization

Data standardization converts data into a standard format that computers can read and understand. This is important because it allows different systems to share and efficiently use data. It is also essential for preserving data quality. When data is standardized, it is much easier to detect errors and ensure that it is accurate. This is essential for making sure that decision-makers have access to accurate and reliable information.

Steps performed:

1. Installing the necessary Python packages.
2. Fetching the required packages from the Python library.
3. Reading the data from your working directory
   The data 'Sleep_health_and_lifestyle_dataset.csv' is a comma separated value (csv) file that contains 13 variables and 374 observations.
4. Cleaning the data
   a. Treating missing data
   b. Label encode the categorical variables
5. Developing various plots for Exploratory data analysis.
6. Create a cleaned dataset by removing variables based on Correlation matrix
7. Perform Data standardization on cleaned dataset
8. Create Training and Test dataset with 70% being the training data and 30% being the test data.
9. Building models
   a. Model 1 – using BaggingClassifier function. It implements the Bagging Meta- Estimator algorithm for classification.
   b. Model 2 – using AdaBoostClassifier function. It implements the AdaBoost algorithm for classification.
   c. Model 3 – using XGBClassifier function. It implements the XGBoost algorithm for classification.
   d. Model 4 – using NaiveBayes function. It computes the conditional a-posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule.
   e. Model 5 – using RandomForestClassifier function. It implements the Random Forest algorithm for classification.
10. Creating the actual Confusion Matrix based on predicted values.
11. Interpreting results

...

Results

Model-1: Bagging Meta-estimator

	precision	recall	f1-score	support
0	1.00	0.83	0.91	12
1	0.91	0.91	0.91	46
2	0.79	0.88	0.83	17
accuracy			0.89	75
macro avg	0.90	0.88	0.89	75
weighted avg	0.90	0.89	0.89	75

Model-2: AdaBoost

	precision	recall	f1-score	support
0	0.64	0.58	0.61	12
1	0.86	0.91	0.88	46
2	0.93	0.82	0.87	17
accuracy			0.84	75
macro avg	0.81	0.77	0.79	75
weighted avg	0.84	0.84	0.84	75

Model-3: XGBoost

	precision	recall	f1-score	support
0	1.00	0.83	0.91	12
1	0.92	0.98	0.95	46
2	0.94	0.88	0.91	17
accuracy			0.93	75
macro avg	0.95	0.90	0.92	75
weighted avg	0.94	0.93	0.93	75

Model-4: Naive Bayes

	precision	recall	f1-score	support
0	0.71	0.83	0.77	12
1	0.91	0.91	0.91	46
2	0.93	0.82	0.87	17
accuracy			0.88	75
macro avg	0.85	0.86	0.85	75
weighted avg	0.89	0.88	0.88	75

Model-5: Random Forest

	precision	recall	f1-score	support
0	1.00	0.83	0.91	12
1	0.91	0.91	0.91	46
2	0.79	0.88	0.83	17
accuracy			0.89	75
macro avg	0.90	0.88	0.89	75
weighted avg	0.90	0.89	0.89	75

Model Comparison

Out[164]:

	Model	Precision Score	Recall Score	Accuracy Score	f1-score
0	Bagging Meta-estimator	0.893333	0.893333	0.893333	0.893333
1	AdaBoost	0.84	0.84	0.84	0.84
2	XGBM	0.933333	0.933333	0.933333	0.933333
3	Naive Bayes	0.853554	0.856635	0.88	0.852425
4	Random Forest	0.900839	0.876243	0.893333	0.885156

On comparing the five models,

- Accuracy score as well as the other performance measures are highest for Model-3 (XGBoost).
- Hence the best model will be XGBoost, followed by Bagging Meta-estimator, Random Forest and Naive Bayes.
- And the least impressive model will be AdaBoost.

Conclusion

Based on the analysis carried out we can say that Sleep Disorder can be predicted up to 93% accuracy based on the models developed.. However other n models can be developed but in our case we have chosen five models and we have put them to the test. To conclude I would like to say that Python as a programming language for analytics is very powerful and gives immense flexibility to the coder. It helped me to build models far easier than I would have in other languages.

References

- https://www.kaggle.com/datasets (Dataset)
- https://www.scalablepath.com/data-science (Information on Python packages)
- https://www.w3schools.com/python (Information on Python functions)
- https://scikit-learn.org (Information on Python packages)
- https://www.geeksforgeeks.org (Theoritical information)