Predicting Sleep Disorders

Logistic Regression & Random Forest

Description:

The dataset provides insights into various health and lifestyle attributes of individuals, potentially linked with sleep disorders. The goal of this work is to compare two machine learning algorithms; Logistic Regression and Random Forest, and to determine the most effective model that could possibly predict if a person has a sleeping disorder.

Data preparation for the training:

The dataset consists of 13 different columns:

- Person ID: A unique identifier
- Gender
- Age
- Occupation: The job of the individual.
- Sleep duration: Duration of sleep the individual gets per night.
- Quality of sleep: A measure of the individual's self-reported sleep quality.
- Physical activity level: An indicator of the individual's level of physical activity.
- Stress level: A measure of the individual's self-reported stress level.
- BMI category
- Blood pressure
- Heart rate: The resting heart rate of the individual.
- Daily steps
- Sleep disorder

To prepare the dataset for machine learning, the columns with 'String' values must be encoded to be numerical values. The columns: Gender, Occupation, BMI Category, Blood Pressure and Sleep Disorder must be encoded.

Encoding:

Gender:

Since the dataset only has 2 genders (Male and Female), Label encoding is good here. Where Male is 1 and female is 0.

Occupation:

The dataset has many different occupations, so one-hot encoding would be perfect, but I chose to go with label encoding, because when viewing the dataframe of e.g. x_train the occupations would take so much vertical space that it's hard to read.

BMI Category:

For the sake of learning, the BMI column was encoded using ordinal encoding. So, in ordinal encoding we essentially create a map where each BMI label is associated with a number. E.g., Normal: 1 and Obese: 4.

Blood Pressure:

In this dataset, the blood pressure was in a combined format of systolic/diastolic, So to separate the two, Create two new columns called Systolic Pressure and Diastolic Pressure, we can use the Split() method to put them correctly in their own category. Then drop the original combined column.

Sleep Disorder:

In the dataset, sleep disorders were categorized as 'insomnia', 'Sleep Apnea', or 'None'. To simplify the data processing, we used one-hot encoding to convert the sleeping disorders to separate columns. After encoding we combined these into a single sleep disorder column to specify what sleeping disorder the individual has or its absence.

Training data:

X_train dataframe:

Index	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Daily Steps	Systolic Pressure	Diastolic Pressure
192	1	43	7	6.5	6	45	7	3	72	6000	130	85
75	1	33	1	6	6	30	8	1	72	5000	125	80
84	1	35	9	7.5	8	60	5	2	70	8000	120	80
362	0	59	5	8.2	9	75	3	3	68	7000	140	95
16	0	29	5	6.5	5	40	7	2	80	4000	132	87
66	1	32	0	7.2	8	50	6	2	68	7000	118	76
283	0	50	5	6	6	90	8	3	75	10000	140	95
7	1	29	1	7.8	7	75	6	1	70	8000	120	80
113	1	37	3	7.4	8	60	5	1	68	8000	130	85
116	0	37	0	7.2	8	60	4	1	68	7000	115	75

Notes: The Occupation's here are integers, and the associating occupations that are connected with the integer can be viewed from the output of the code.

Y_train dataframe:

y_train - Series

Index	Sleep Disorder
192	Insomnia
75	None
84	None
362	Sleep Apnea
16	Sleep Apnea
66	None
283	Sleep Apnea
7	None
113	None
116	None

No explicit scaling method was applied to the dataset. Dummy variables were created for the 'Sleep Disorder' column, which resulted in columns: 'Sleep_Disorder_Insomnia', 'Sleep_Disorder_Sleep Apnea', and 'Sleep_Disorder_None".

Relevant Metrics for The Cases:

```
Performance of Logistic Regression:
0.9066666666666666
              precision    recall  f1-score   support

    Insomnia       0.82      0.88      0.85        16
        None       0.93      0.98      0.95        43
 Sleep Apnea       0.92      0.75      0.83        16

    accuracy                           0.91        75
   macro avg       0.89      0.87      0.88        75
weighted avg       0.91      0.91      0.90        75

Confusion Matrix for Logistic Regression:
[[14  1  1]
 [ 1 42  0]
 [ 2  2 12]]
Performance of Random Forest:
0.88
              precision    recall  f1-score   support

    Insomnia       0.72      0.81      0.76        16
        None       0.95      0.98      0.97        43
 Sleep Apnea       0.85      0.69      0.76        16

    accuracy                           0.88        75
   macro avg       0.84      0.83      0.83        75
weighted avg       0.88      0.88      0.88        75

Confusion Matrix for Random Forest:
[[13  1  2]
 [ 1 42  0]
 [ 4  1 11]]
```

- Accuracy
  - The proportion of total predictions that were correct.
- Precision
  - The proportion of positive identifications that were correct.
- Recall
  - The proportion of actual positives that were correctly classified.
- F1-score
  - The average (mean) of precision and recall.
- Support
  - The number of actual occurrences of the class in the dataset.
- Macro Avg and Weighted Avg
  - Macro Avg
    - It averages the metric value for each class without considering the class distribution.
  - Weighted Avg
    - It calculates metrics for each label and finds their average weighted by the number of true instances for each label.

- Confusion Matrix:
    - Logistic regression predicted:
        - ‘Insomnia’ correctly 14 times, misclassified it as ‘None’ once and as ‘Sleep Apnea’ once.
        - ‘None’ correctly 42 times and misclassified it as ‘Insomnia’ once.
        - ‘Sleep Apnea’ correctly 12 times, misclassified it as ‘None’ 2 times, and as ‘Insomnia’ once.
    - Random Forest predicted:
        - ‘Insomnia’ correctly 13 times, misclassified it as ‘None’ once, and as ‘Sleep Apnea’ 2 times.
        - ‘None’ correctly 42 times, misclassified it as ‘Insomnia’ once.
        - ‘Sleep Apnea’ correctly 11 times, misclassified it as 'None’ once, and as ‘Insomnia’ 4 times.

Conclusions:

The dataset was split into two sets:

1. Training Data (80%)
2. Test Data (20%)

Looking at the results:

- Logistic Regression model:
    - Accuracy: 90.67%
- Random Forest model:
    - Accuracy: 88 %

Both models performed well, with the Logistic Regression model slightly outperforming the Random Forest model In terms of accuracy.

I think the model is good enough for the planned case, I plugged in my own data to see if the model would predict if I had a sleeping disorder and according to the model, I don’t have a sleeping disorder, so it must work!

The only thing I’m concerned about is this data has 374 rows of data, so 374 unique persons with only data from 1 night, I wonder would the model perform better, if there was data from example a full week of sleep. Because we all have bad nights. I would imagine the data being more valuable if there was from each person like the last 5 nights of data.

To improve the model, I would use cross-validation. Or combine predictions from multiple models.

Dataset:

https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset