KNN Classification on Sleep Data Set

Eric Wu

2023-09-04

Background

The dataset "Sleep health and lifestyle" is from Kaggle.com, an online community with enriched datasets from various contributors for data analysis and machine learning projects. The primary owner of the "housing" data set is Laksika Tharmalingam.

Figure 1: (Cited from https://opendatascience.com/10-tips-to-get-started-with-kaggle/)

Skill Set Objectives

- KNN Classification on multi-class variable
- Cross Validation
- Data Visualization

Dataset

The "Sleep health and lifestyle" dataset consists of data from 374 individuals with 13 variables. The goal of the dataset is to predict the type of sleep disorders based on all other factors, with an emphasis on sleep and

exercise routines. Some of the variables included are sleep duration, quality of sleep, physical activity level, stress level, BMI categories, blood pressure, heart rate, etc. In addition, the dataset also has 0 null values.

Table 1: Basic statistics of quantitative variables

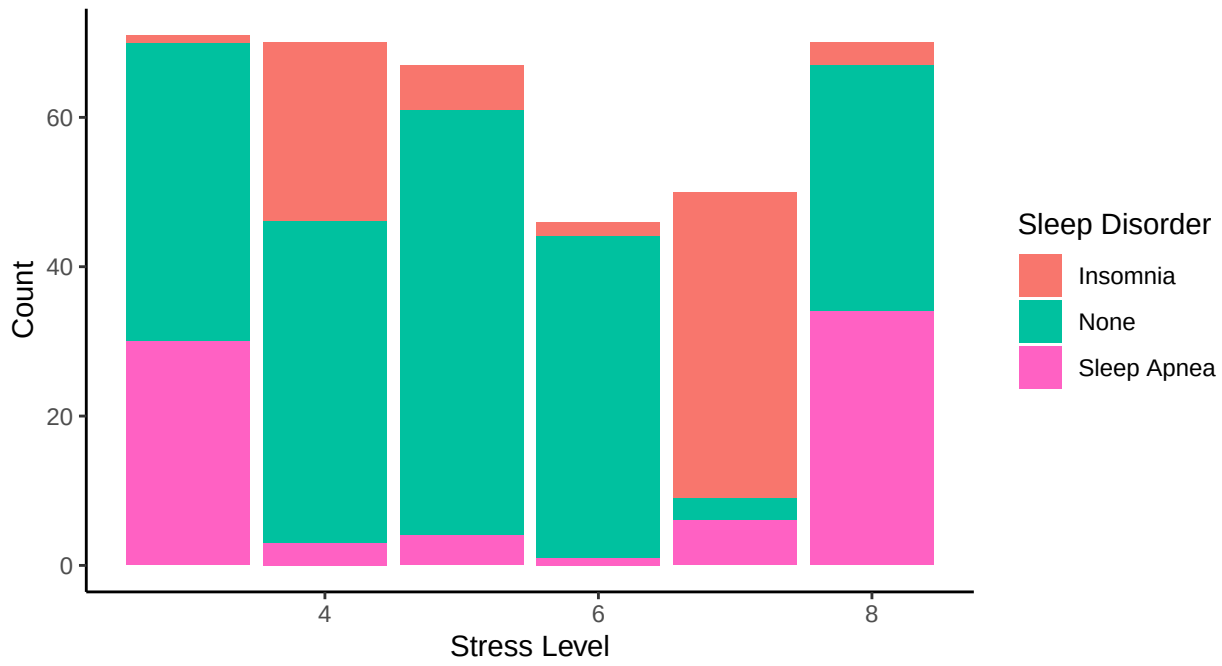
	vars	n	mean	sd	min	max	range	se
Age	1	374	42.18	8.67	27.0	59.0	32.0	0.45
Sleep.Duration	2	374	7.13	0.80	5.8	8.5	2.7	0.04
Quality.of.Sleep	3	374	7.31	1.20	4.0	9.0	5.0	0.06
Physical.Activity.Level	4	374	59.17	20.83	30.0	90.0	60.0	1.08
Stress.Level	5	374	5.39	1.77	3.0	8.0	5.0	0.09
Heart.Rate	6	374	70.17	4.14	65.0	86.0	21.0	0.21
Daily.Steps	7	374	6816.84	1617.92	3000.0	10000.0	7000.0	83.66
Systolic	8	374	128.55	7.75	115.0	142.0	27.0	0.40
Diastolic	9	374	84.65	6.16	75.0	95.0	20.0	0.32

Table 1 demonstrates some sample statistics for the "sleep, health, and lifestyle" dataset, and some of the findings are very interesting. For example, the average value of the quality of sleep is 7.31 (out of 10), which means a majority of people believe they are having quality sleep although only 58.56% don't have a sleep disorder, this means that 41.44% of people who have a sleep disorder doesn't seem to be bothered by the sleep disorder. which suggest the importance of another measurement regarding the severeness of the sleep disorder, although it can be estimated with other variables, or suggest that people filling out the survey are being biased about their quality of sleep.

Exploratory Data Analysis

The Exploratory Data Analysis section will focus on finding the relationships between Sleep Disorder and other important factors.

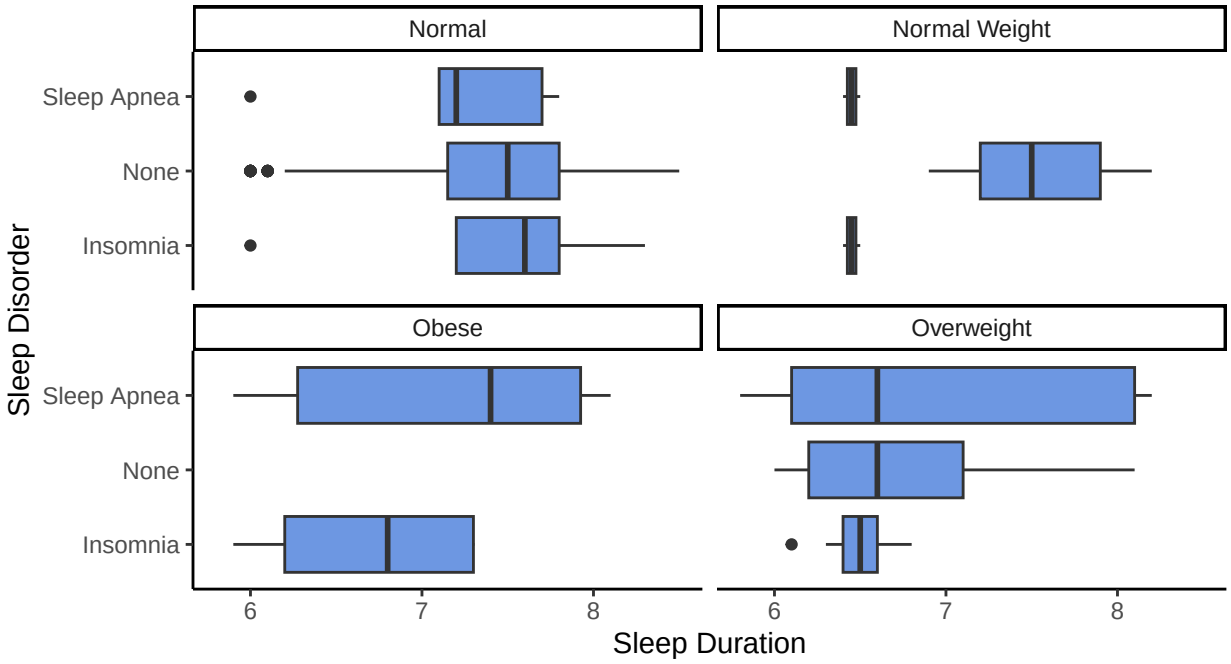
Figure 2:How Sleep Disorder Affect Stress Level



To begin with, the first graph demonstrates how sleep disorders affect people's stress levels. For example, people who are extremely stressed tend to have Sleep Apnea than Insomnia. However, although the total count also decreased, for just one level decrease in stress, people with Insomnia outnumbered those with Sleep Apnea. On the other hand, we can generalize and conclude that both seven and eight are considered extremely stressed, and the reason that people are not selecting stress levels nine and ten is response bias.

Another intriguing finding is that people who have sleep Apnea tend to be in two extremes and people with Insomnia are more evenly spread. This can be because people with Insomnia are growing stressed from time to time, and Sleep Apnea either significantly increases stress or does not affect the stress level at all. One thing to not forget is that only 30% of the extremely stressed population does not have sleep disorders, which is evidence proving that stress levels can lead to sleep disorders.

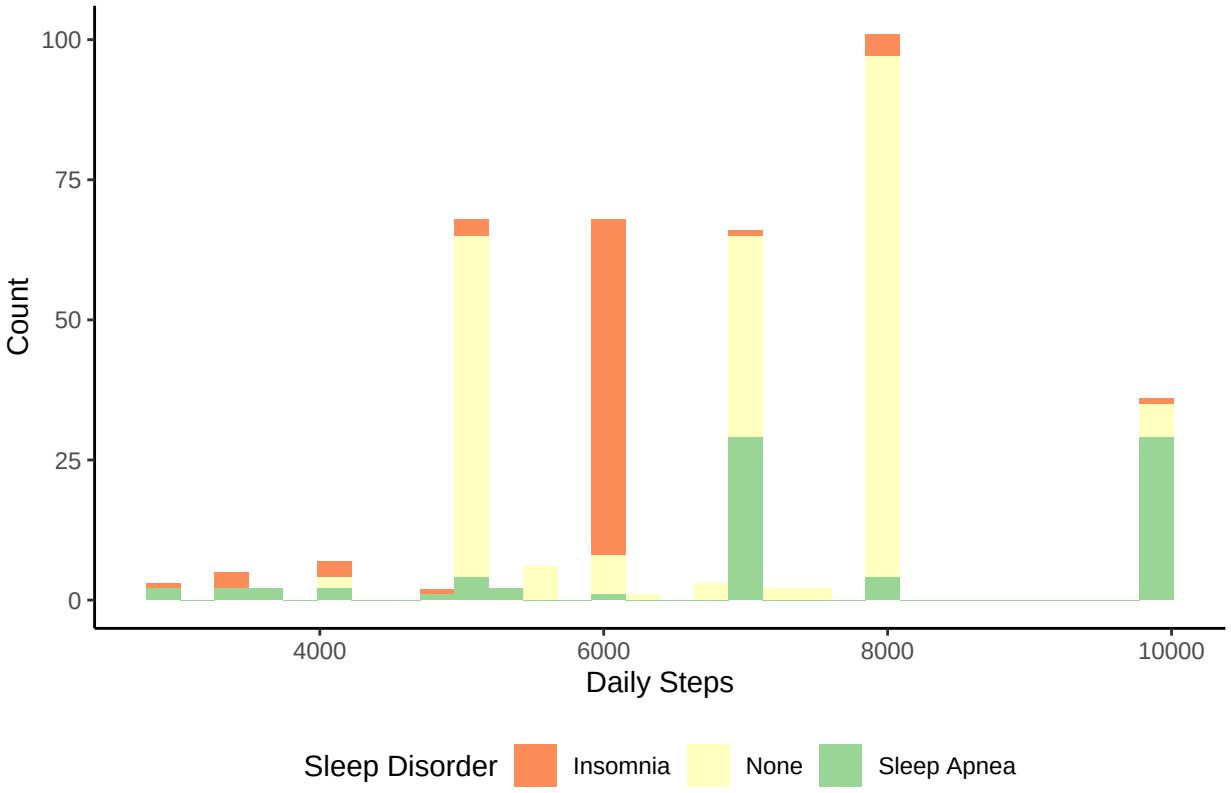
Figure 3: Sleep Duration Among Group With Different BMI Category and Sleep Disorder



The second graph presented a direct comparison of sleep duration between different groups based on sleep disorders and BMI category. From the graph, it's very easy to tell that the BMI category has four different classes with the "Normal" class being the largest (195), "Overweight" second (148), "Normal Weight" third (21), and "Obese" the fewest (10). Oddly, the author addresses the BMI category in four classes, but every piece of data is important in this analysis. On the other hand, It's clear that people who classify as obese either have sleep apnea or insomnia, which suggests that there can be a positive correlation between weight and sleep disorder; however, the reason can be because of the limited data.

In addition, normal people with insomnia had a higher median sleep duration than any other group, this suggests that people classified as normal weight can be less affected by insomnia. However, because the distribution of the data is not normal, and that means median is no longer a great indicator of the average. Moreover, people who classified as overweight and insomnia seem to have a relatively lower average of sleep duration compare to other people who also classified as insomnia.

Figure 4: Histogram of Daily Steps in Relation to Sleep Disorder



The third graph is a histogram of daily steps with sleep disorders as fills. One interesting finding about this graph everyone who walks less than 4,000 steps per day has some sort of sleep disorder, and people who walk the most seem to be dominated by sleep apnea with a small proportion of insomnia. This suggests that either lack of exercise has some kind of association and people with sleep apnea are not that severe that prevent daily activities.

KNN Classification

Next is to implement the actual KNN classification algorithm, the key to KNN classification is that it only support qualitative variables and require scaling to make sure every variable is compared on the same scale. This report will also run a 10-fold cross validation loop to examine the best k (Greatest accuracy).

Figure 5: Graph of the effect of the choice of k