The Impact of Lifestyle Behaviors on Sleep Health

Katherine Pastva and Tyler Mui

1. Motivation

The primary motivation for this project is to see what factors are most significant to our sleep health, given different variables. Taking up roughly ⅓ of our lives, sleeping is a quintessential part of human existence. This insight may be valuable to anyone trying to become healthier and change their lifestyle. So let's maximize this time, quantifying which factors impact our sleep the most!

2. Data Sets

2.1. Data Set Description

In this project, we are working with a dataset scraped from Kaggle.com, covering a wide range of variables related to sleep and daily habits. It includes 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.

2.2 Data Preparation

Some initial cleaning has been carried out on the dataset. For instance, after importing the original data set into Python, the 'None' labels in the Sleep Disorder column are imported as 'NaN'. Using the *fillna()* method within pandas, the missing values in the sleep disorder column were labeled as 'No Disorder', meaning that the individual does not exhibit any sleeping disorders. The values 'Normal Weight' were changed to 'Normal', and 'Sales Representative' to 'Salesperson' for consistency. Additionally, the 'Person ID' column was removed and duplicate rows were removed using the *drop_duplicates* function. The cleaned dataset comprises 12 columns, with each row representing an individual. The columns are as follows:

Dataset Columns:

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 (hours): The number of hours the person sleeps per day.

Quality of Sleep (scale: 1-10): A subjective rating of the quality of sleep, ranging from 1 to 10.

Physical Activity Level (minutes/day): The number of minutes the person engages in physical activity daily.

Stress Level (scale: 1-10): A subjective rating of the stress level experienced by the person, ranging from 1 to 10.

BMI Category: The BMI category of the person (BMI < 25.0 -- Normal, BMI < 30.0 -- Overweight, BMI < 35.0 -- Obese)

Blood Pressure (systolic/diastolic): The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.

Heart Rate (bpm): The resting heart rate of the person in beats per minute.

Daily Steps: The number of steps the person takes per day.

Sleep Disorder: The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).

Details about Sleep Disorder Column:

No Disorder: The individual does not exhibit any specific sleep disorder.

Insomnia: The individual experiences difficulty falling asleep or staying asleep, leading to inadequate or poor-quality sleep.

Sleep Apnea: The individual suffers from pauses in breathing during sleep, resulting in disrupted sleep patterns and potential health risks.

3. Research Methods and Related Work

3.1. Research Questions

What factors are most significant to sleep health?

Is there any correlation between factors and different sleeping disorders?

3.2. Related Work

A few works on Kaggle use this data set as well, but our research differs from those in various ways. In our work, we are diving in depth to see what BMI category has the most sleeping disorders. Moreover, we are looking at the distribution between occupations and the stress related to each gender. We also use machine learning models such as Decision Tree and KNN to look at important factors in the data.

4. Research Methods

Decision Tree Classification:

We implemented a decision tree classifier to predict the target variable, Quality of Sleep, based on the features Gender, Sleep Duration, Stress Level, and BMI Category. Since Gender and BMI Category are categorical variables, we converted them into numerical representations using get_dummies. The dataset was split into training and test sets, with the test set comprising 40% of the data. We defined the decision tree classifier with hyperparameters max_leaf_nodes=6, max_depth=6, and random_state=0. To evaluate model performance more robustly, we utilized cross-validation to assess generalization performance across multiple folds of the data.

The classifier was trained (fit) on the training set, and predictions for the test set were made. Finally, we calculated the best model's accuracy to measure its predictive performance.

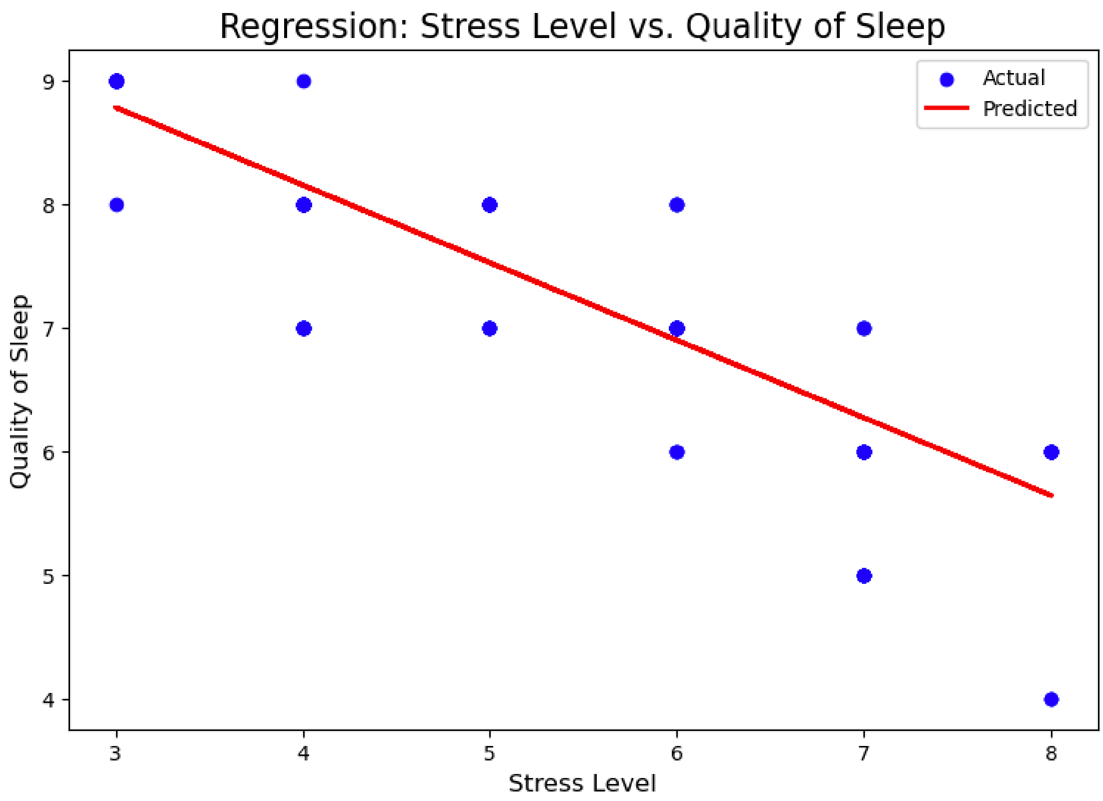
KNN Classification:

We prepared the X and Y inputs in the same way as we did for the decision tree model. The data was split into training and testing sets. The training set was then used to train the KNN model, employing n-fold cross-validation to identify the best-performing configuration. Once the optimal KNN model was determined, its performance was evaluated using the test set, allowing us to compare it with the Decision Tree model.

To optimize the KNN model, we utilized GridSearchCV, which systematically trains the model across a range of specified hyperparameters. This approach evaluates each hyperparameter combination to identify the model for best performance. For our model, we explored different values for the n_neighbors parameter by creating a dictionary where n_neighbors served as the key result. Using NumPy, we generated an array of values ranging from 1 to 24 to determine which value works best for our KNN model. The optimal hyperparameter for performance was a KNN of 1 and a cross-validation value of 2.

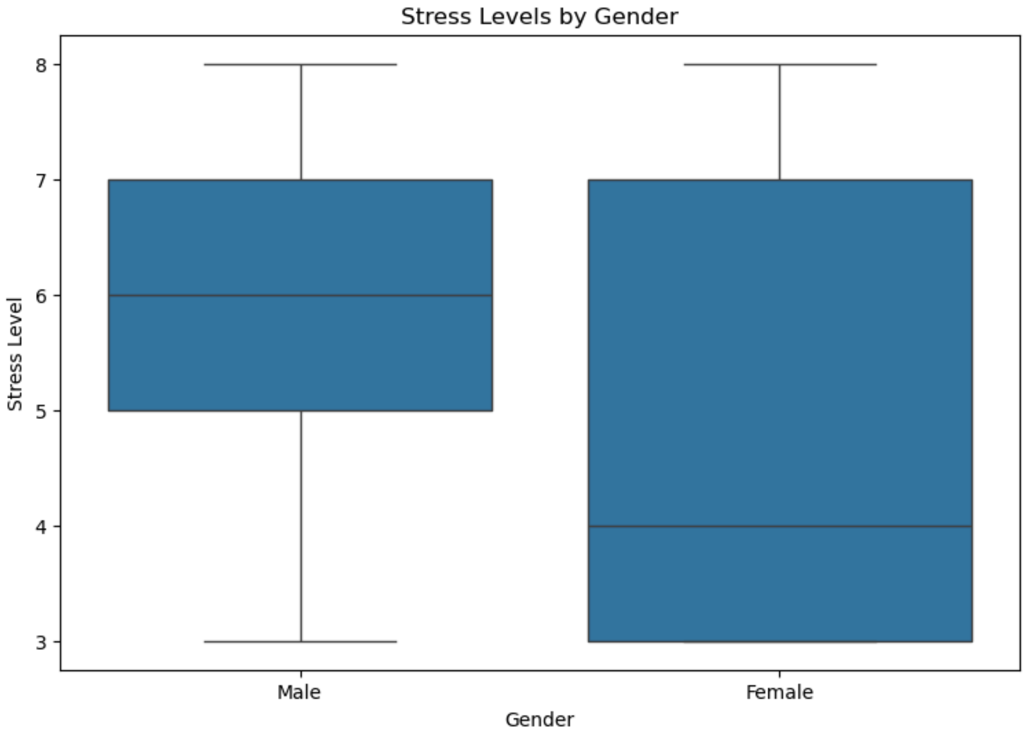
5. Findings/Results

5.1. Visualizations

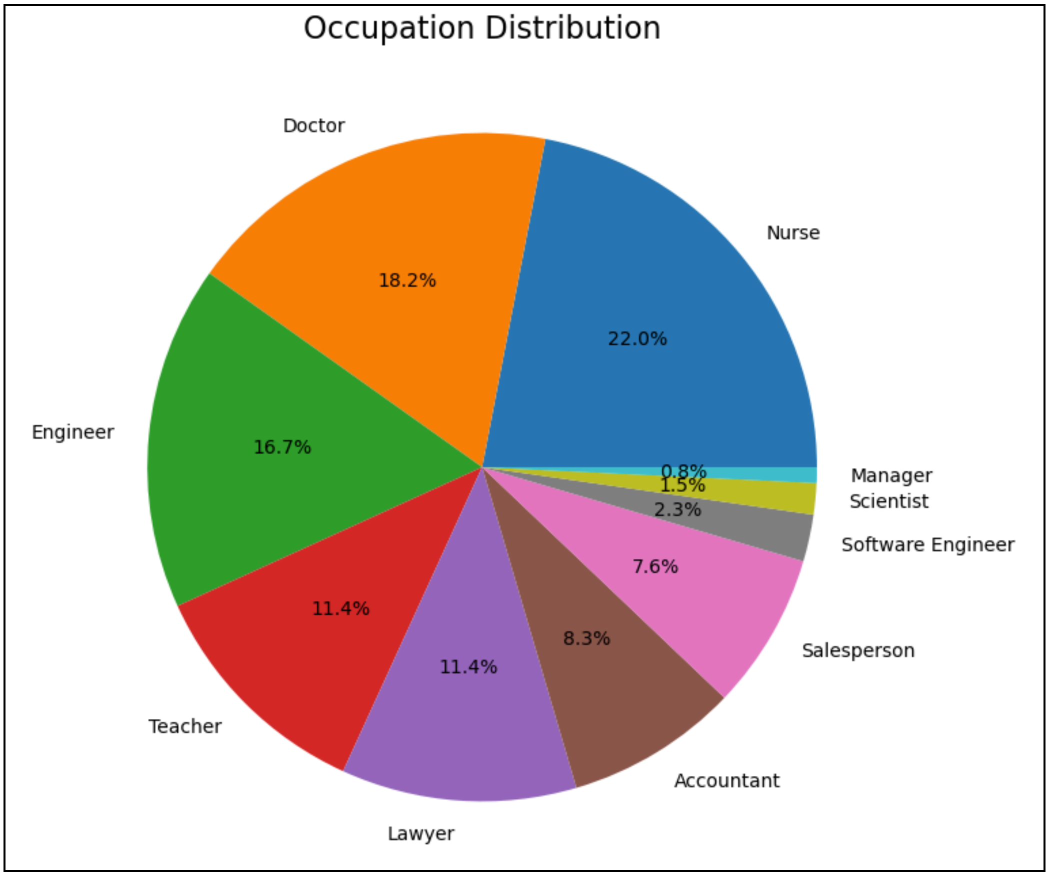


We can see that there is a negative linear relationship between stress levels and Quality of Sleep.

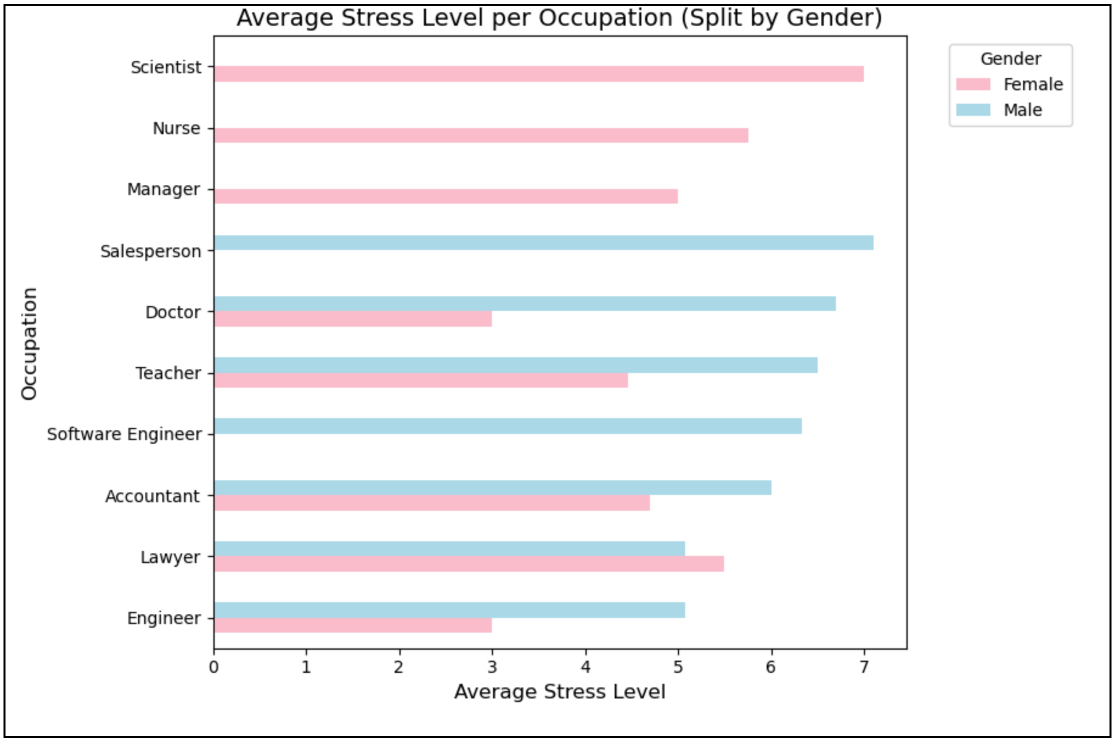
As we are more stressed, we tend to have poorer sleep



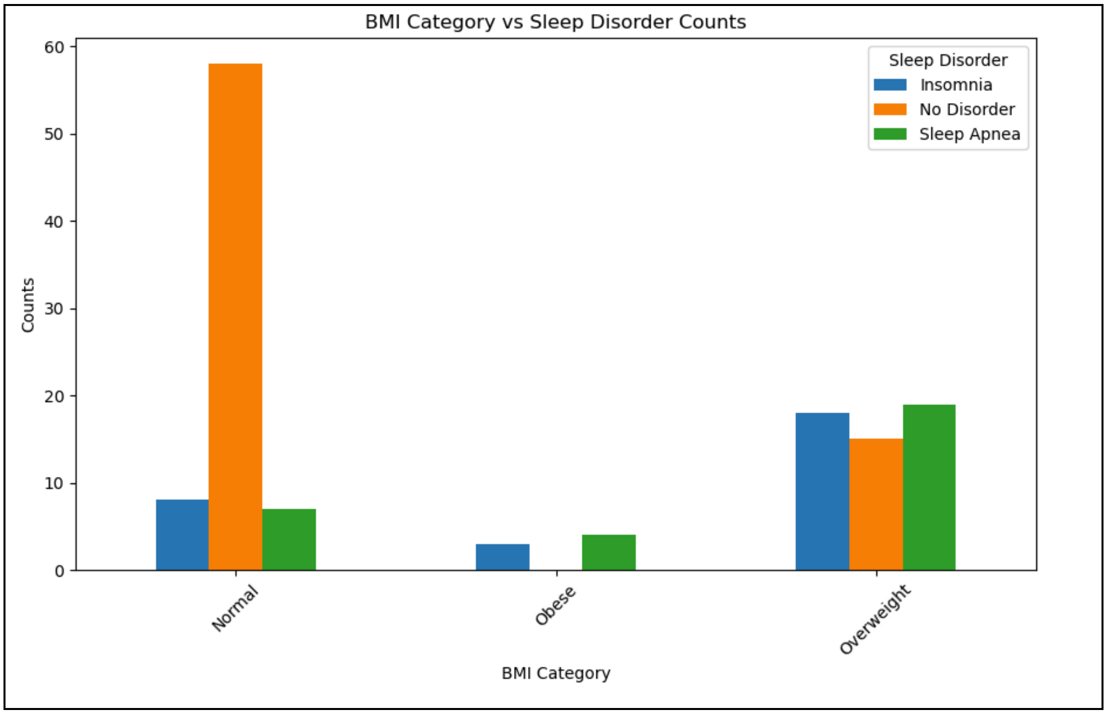
Given this box plot, both males and females have high stress, but the median stress for a male is higher than a female, in this dataset. We would use the median in this case since it is less affected by extreme values and provides a better representation of the "typical" stress level within the data, whereas the mean can be significantly pulled towards the outliers. For example, different occupations may be more stressful than others.



This pie chart shows the occupation distribution for the data set. The majority of the occupations are doctors, nurses, and engineers. This is also important to note the amount of people in each profession as we continue to analyze our data set so that reported stress levels are weighted differently depending on the number of people that reported from that profession. This led us to investigate if certain careers are more stressful than others, impacting sleep.



The previous pie chart led us to believe that occupation can affect stress levels, although given that professions can have significant differences between their makeups of male and female, we wanted to investigate if males and females are distributed differently among the professions, given their stress level. Given this bar chart comparing average stress levels per occupation, split by gender, we can see that the average stress levels for females (represented by the pink bars) are less than the average stress levels for males. In particular, Accountants, Engineers, Teachers, and Doctors all present less stress than their male counterparts. The only outlier is lawyers, presenting higher stress for females than males.



As shown in this bar chart, people of normal weight are less likely to have a sleep disorder compared to Overweight people. People who report being obese also have a sleep disorder.

These visualizations led us to regress our data based on the factors: BMI, Gender, Stress Level, and Sleep Duration.

5.2. Analysis and Results:

Based on our target variable and our research question, we compared the **decision tree** and **k-nearest neighbor** models to determine which model would have the best predictive accuracy.

The performance metric used to compare these models was the test accuracy, which is defined as the proportion of predictions that were correct.

$$\text{Accuracy} = \frac{\#\text{Correctly predicted}}{\#\text{Total}} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Although accuracy is overly simplistic, this was used in combination with three other metrics. The precision of a classifier is the proportion of correct positive predictions, defined by the following function:

$$\text{Precision} = \frac{\#\text{True positive}}{\#\text{Predicted positive}} = \frac{TP}{TP + FP}.$$

Recall is the proportion of positives that were correctly predicted and can be found by the formula:

$$\text{Recall} = \frac{\#\text{True positive}}{\#\text{Actual positive}} = \frac{TP}{TP + FN}.$$

The F1 Score combines our precision and accuracy, making it a more accurate measure if our data is imbalanced. An F1 score closer to 1 indicates a perfect model.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Classification Report:

Metric	**Decision Tree**	**KNN**
Accuracy	0.83	0.91
Weighted Avg Precision	0.84	0.91
Weighted Avg Recall	0.83	0.91
Weighted Avg F1-Score	0.82	0.90

Since the KNN metric exhibits higher values for all accuracy measures than the decision tree model, we determined that the KNN model would predict more accurately based on our data.

5.3. Determining Variable Importance:

To address the research question, it is important to determine variable importance. Selecting the variables sleep duration, stress level, gender, and BMI as the predictors for the models, we can determine their importance from the following table:

Feature	**Decision Tree**	**KNN**
Sleep Duration	0.791246	0.402532
Stress Level	0.085437	0.369620
is_Male	0.067310	0.260759
BMI	0.056007	0.331646

From this table, the variable sleep duration displays the greatest importance to the decision tree model, but all variables share equally strong importance in the KNN model. This means that depending on the model variables exhibit differences in their importance to create accurate predictions.

6. Conclusions

6.1. Summary of Findings

Sleep duration is a strong feature for quality of sleep, followed by stress levels

Stress levels additionally play a factor in your quality of sleep. There seems to be a negative correlation between the two. In this dataset, the range of stress between males and females tends to be higher in males. Moreover, different occupations tend to have different levels of stress, which may lead back to your quality of sleep.

6.2. Limitations

This project was only able to test four different factors and their impact on sleep health, all factors from one data set. Because we selected to use one data set, this limited the scope of data we had access to analyze. In addition, the data set in itself had its own credibility issues due to an uneven distribution of occupations being recorded, therefore skewing reporting results. Additionally, this data was a synthetic dataset generated from Kaggle, and may not represent the true population distribution of our factors.

7. Future Work

It would be more beneficial to find a larger, non-synthetic, dataset about sleep and lifestyle because the limited rows after cleaning are not an accurate representation of the population's sleep health. Adding another data set that includes dementia and Alzheimer's would be an interesting expansion of this project. Additionally, finding a set where there are underweight individuals would allow for a broader investigation into how weight impacts sleep health.

8. Acknowledgment and Reflection on the Use of LLMs

We were assisted by the LLM ChatGPT. Throughout our work, we used it for troubleshooting, the explanation of technical concepts, and the generation of some code, including some of the graphs. AI is one of the most useful tools in today's day and age if used right. If you don't know how to prompt AI or understand how it works, you will fall behind in the workplace. Most companies are implementing AI within their work, including software engineers using Github Copilot, so it is at the point where it is becoming an industry standard.

9. References

Dataset:

https://www.kaggle.com/datasets/informateur234/sleep-health-and-lifestyle-dataset

Coding references:

https://www.kaggle.com/code/ibrahimelgmmal/sleep-health-and-lifestyle/notebook