Tai Le, Haoming Chen

CMPE 255

10/29/2023

Fall 2023

Project report

Abstract

In this data mining project, we apply data mining techniques to a large public health dataset to find links and patterns between a person's sleep quality and their personal attributes and health habits. To present our findings in this report, we apply various processing techniques, as well as graphing libraries, to the dataset to identify and highlight the most important patterns. As the ultimate goal of this report, we want to inform the public of our findings so people can have a better understanding of their sleep and make a conscious decision to improve their sleep quality.

1. Introduction

Sleeping is an essential activity for maintaining good health and well-being throughout human life. Every day, people spend a portion of their time sleeping to restore their energy and make them feel better for the next day. Because of that, sleeping can take up to one-third of a person's entire life. There are many theories behind why we sleep. Yet, the most relevant theory explains that sleep plays a critical role in helping restore our body through muscle growth, tissue repair, protein synthesis, and hormone growth, and it rejuvenates our brain by eliminating unwanted chemicals and restructuring the neural connections, or synapses.

Because of the necessary functions our body performs during sleep, a lack of quality sleep can have a significant impact on one’s physical and mental functioning. After all, most people have experienced fatigue and mood changes after a night of bad sleep. According to studies, a bad night's sleep can have an immediate negative impact on your emotions and your ability to learn, focus, react, make decisions, and solve problems. Bad-quality sleep in the long term can cause health consequences, including chronic medical conditions like diabetes, obesity, and heart disease. Because of the importance of sleep to our health and well-being, ensuring quality sleep should be a top priority in our lives. This has become the main motivation for this sleep analyst project. As part of our contribution to the community, this project will present our findings on patterns between sleep and various attributes and habits in the hope that readers can be informed with this information and make a conscious decision to improve their sleep quality.

2. Related work

Sleep is a crucial part of our life and it can justify if we have a healthy life. In the research field, several studies address the issue of insomnia and use machine learning to predict the sleep health of people depending on their data. One of the most well-known groups is National Sleep Research Resource. There are a lot of datasets that can be reused in many different research. Every day, there will always be some new techniques that can be developed to improve the algorithms to lead to a more precise result.

In [1], the authors used the Gradient Boosting Machine to detect and predict the sleep patterns of the users. The performance of the Gradient Boosting Machine is also really impressive, yielding 94.9% in accuracy, and 91.9% in precision.

In [2], the authors addressed that, Deep learning will perform better than a traditional logistic regression, under the receiver operating characteristic (ROC) curve (AUC) is 0.9449. It is about 46% improved in performance compared to the traditional logistic regression.

On the other hand, in the study of [3], Random Forests and Hidden Markov Modeling [HMM] also perform exceptionally well on sleep or nap time prediction. Random Forest is great but still lacking the temporal prediction, HMM helped that and yielded a slight improvement in the model.

These are some great examples of successful models that can predict sleep quality or sleep time and can be used as a reference for this project.

3. Analysis

Dataset

The dataset we have used in this project is the 2022 Behavioral Risk Factor Surveillance System (BRFSS) dataset that is publicly available on the CDC. The BRFSS is a system of health-related telephone surveys that collect data from U.S. residents living in every state and non-state territory regarding their personal attributes, health-related risk behaviors, chronic health conditions, and use of preventive services. It is the largest continuously conducted health survey system in the world. The BRFSS collects its data mainly in the form of remote interviews conducted with its participants via landline telephones or cell phones. The questionnaire consisted of various questions on personal attributes, chronic health conditions, and health-related behavioral questions. The response to the questions is transformed and mapped into a total of 326 columns of various binary, categorical, and numeral data. The 2022 BRFSS dataset has a total of 445132 valid data entries.

For the purpose of our sleep analysis, we built a subset of data from the BRFSS dataset that is best suited for our analysis. First, we selected the column “sleep time”, to be the metric to evaluate sleep quality in this analysis. We then examined all 326 columns and hand selected a total of 28 columns of potential features that might correlate to sleep quality. The feature selection process along with additional steps on feature transformation will be further explained in the data preprocess section.

Preprocessing

The first step in data preprocessing is to examine all 326 columns and extract relevant columns that can be used for pattern exploration. We first selected our target column, sleep time, as our metric for analysis, followed by 28 columns that we believed to have some relevance to sleep quality based on known facts and studies. The features are divided into four categories:

- personal traits: states, sex, age, education, income
- Health-related attributes: IBM, exercise, physical health, mental health, stress, life satisfaction
- Health-related habits: cigarettes, tobacco, e-cigarettes, marijuana, drinking
- Chronic disease: heart attack, coronary heart disease, stroke, asthma, kidney disease, arthritis, diabetes, depressive disorder

Once the feature has been selected, we then clean up the data by removing outliers and changing the unknown response to null. During this process, we replaced abnormal, unknown, and placeholder values in each column with NAN values for every feature. Next, we also assigned a numerical order to all categorical features so that the order of the original categorical data can be accounted for in the correlation calculation. An example of this step is shown:

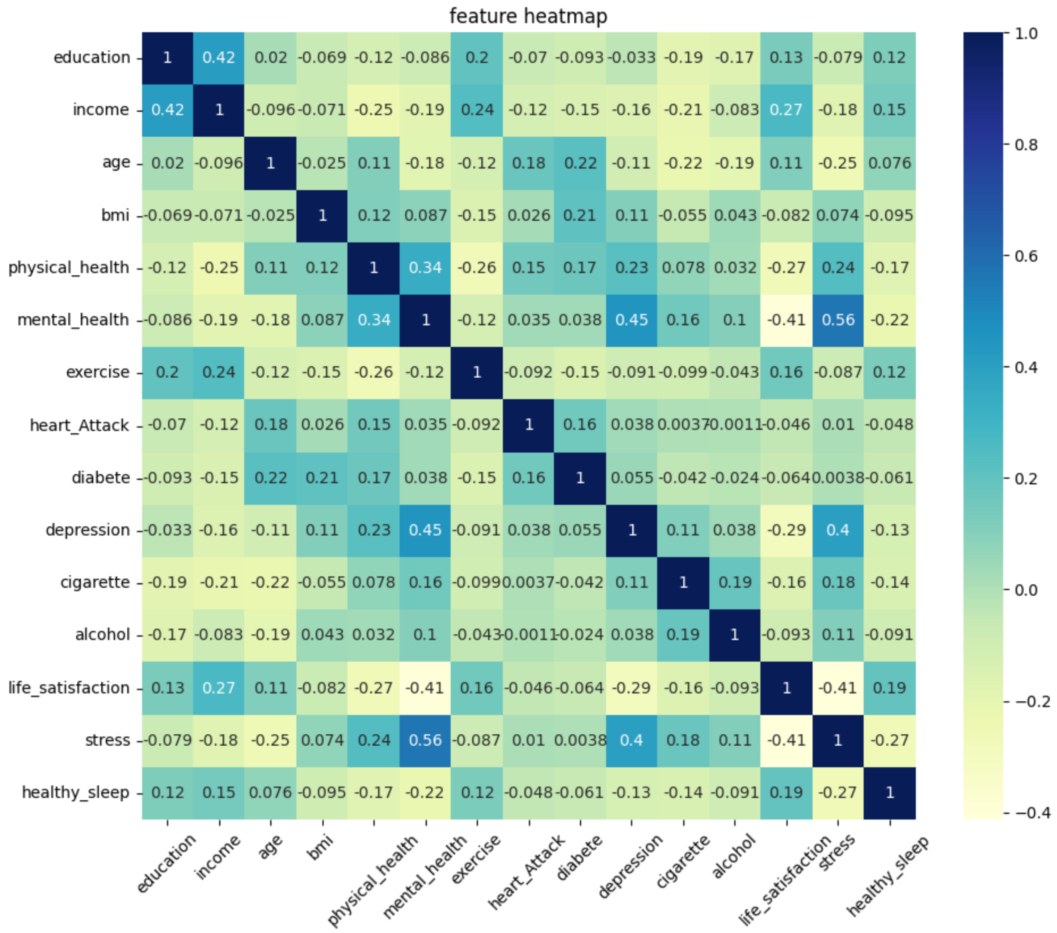
Smoking:

Categorical	numerical
Everyday	2
Some days	1
Not at all	0

In order to measure the sleep quality in this dataset, we need to transform our target column, sleep time, into a new column named healthy_sleep. The reason why sleep time is not used for measuring sleep quality is that sleep time is not linearly related to quality. Although an insufficient amount of sleep can lead to a series of negative effects, studies have also shown that oversleeping on a daily basis can lead to the same negative effect. Sleep experts suggest that the healthy amount of sleep for an adult is between 7 to 9 hours. Based on this fact, we applied a transformation equation below to convert sleep time into a healthy_sleep binary label using the equation below.

$$healthySleep = \begin{cases} 1 & \text{if } 6 < sleeptime < 10 \\ 0 & \text{if } sleeptime < 7, sleeptime > 9 \end{cases}$$

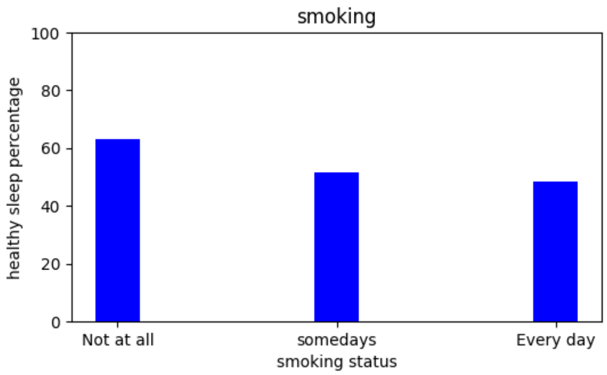
Analysis 1:

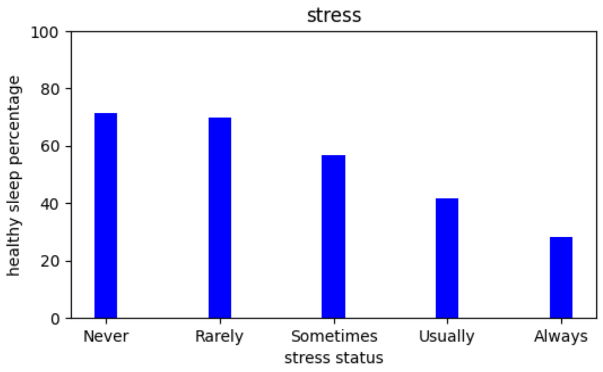


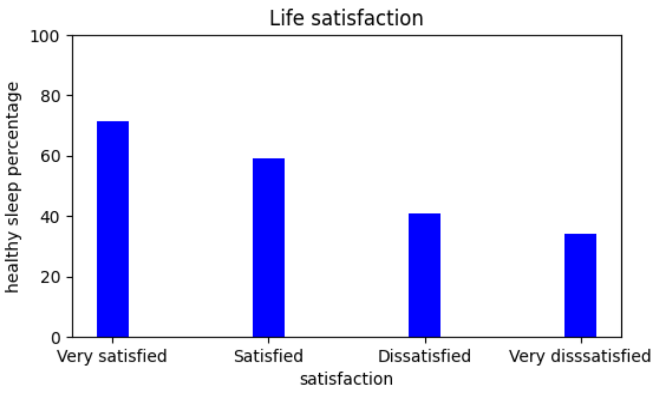
After the EDA process, we created a correlation heatmap based on the significant post-process features along with our metric value of healthy sleep. Here are the most important highlights from this heatmap:

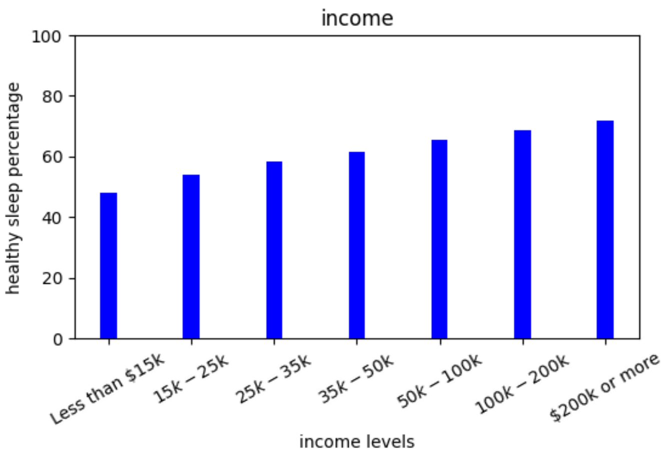
Stress and mental health have the highest relative absolute correlation (> 0.2) with healthy sleep, which indicates that these are the primary determinant factors in healthy sleep. Income, education, physical health, exercise, depression, cigarettes, and life satisfaction all have a small absolute correlation (1.0 > and < 2.0) with healthy sleep, which indicates that these features have some smaller effect on sleep. Aside from these, age, BMI, heart attack, diabetes, and alcohol show minimal correlation (<0.1) with healthy sleep. For all features with a correlation greater than 0.1, stress, mental health, depression, and cigarettes are negatively correlated to healthy sleep, while education, income, and exercise are positively correlated to healthy sleep. Aside from looking at our target feature, there are other interesting correlations we found among all the features: Income has a high positive correlation with education and a negative correlation with mental and physical health. Stress has a positive correlation with mental health, physical health, depression, cigarettes, and alcohol, and a negative correlation with income and age. Lastly, cigarettes have a negative correlation with stress, income, and education.

To better demonstrate the correlation and trend, we tried plotted the percentage of population with healthy sleep of each category within one feature for some of the features from the heatmap:





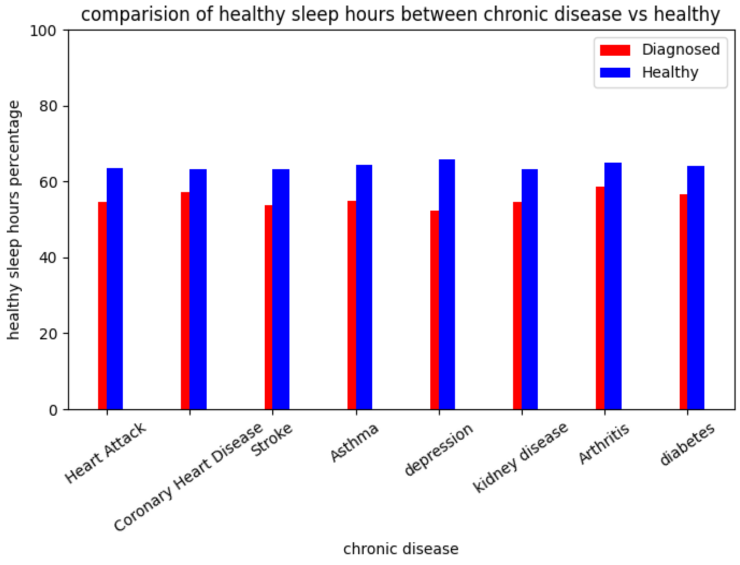




Looking at the bar chart, the graphs for stress, satisfaction, and smoking show a clear downtrend of percentage of healthy sleep individuals as we move across each category. As an example, In the smoking graph, the percentage of healthy sleep individuals who do not smoke at all has a higher percentage at ~61% than those who smoke someday at ~51% , and those who smoke every day at ~48%. Similarly, people who have a high rating with life satisfaction or less stress also have higher healthy sleep hours compared to those who have low satisfaction or higher stress. In the graph for income, there is a clear uptrend in the percentage of healthy sleep individuals as income level goes up.

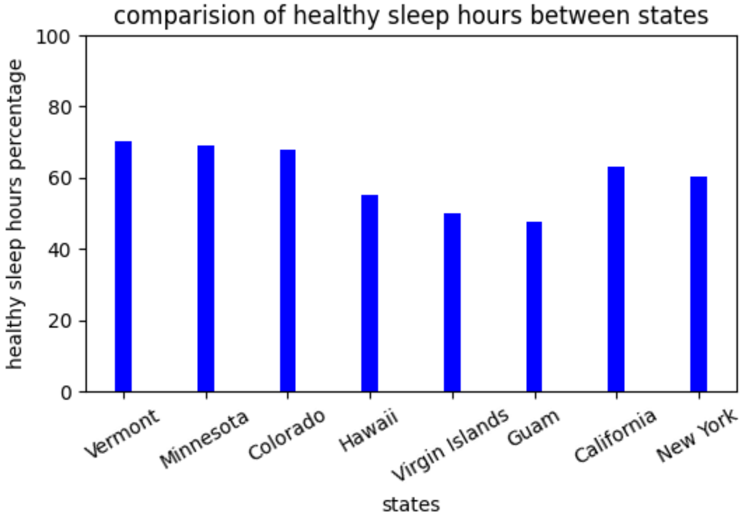
Analysis 2:

There are some abnormalities as well as unexpected findings during the analysis. One unexpected finding we have is that the calculated correlation between chronic diseases like diabetes and heart attack and healthy sleep is minimal. This is opposed to many studies which stated that bad quality sleep can increase the risk of diagnosing chronic diseases. We then tried to prove this fact by plotting bar graph:



Based on the graph, there is a pattern in which people diagnosed with chronic diseases have an overall lower percentage of healthy sleep individuals than normal people. We believed that this is unexpected since the trend is undermined by their correlations.

There is another unexpected finding when we plot the percentage of healthy sleep individuals bar graph by state.



In this bar chart, Vermont, Minnesota, Colorado have the highest percentage out of 50 US states and territories. On the other hand, Hawaii, Virgin islands and Guam are the lowest percentage. Based on this graph, we hypothesize that people in island states and territories might have a less healthy sleep population than inland states.

4. Community Contribution:

Sleep is one of the most essential activities for everyone. But there are so many people who don't realize the consequence of over and undersleep. Oversleeping or undersleeping can lead to depression, and a higher risk of obesity, diabetes, heart disease, stroke, … and so on. We realize that we decided to take action toward improving the sleep health of people or simply raising awareness about oversleeping or undersleeping. Sleep analysis and prediction are crucial and can have a significant impact on improving the overall well-being and health of other people. By collecting data and analyzing it, we gain valuable insight into the factors that can affect sleep health. Using this information, this project will develop a model that can also help people to be aware of their chance of being over or undersleep, enabling them to take proactive measures to improve their sleep quality. Moreover, by sharing the analysis and experiment in the community, we can reach out to many more people and collect more data to refine and enhance the model, making it more accurate and easily accessible to all people. Ultimately, this project has the potential to impact the lives of many people by raising their awareness about sleep health and empowering them to have a healthier and more restful life with better sleep.

5. Conclusion:

In conclusion, this data mining project delved into a big data set about public health, specifically about sleep health, and gained valuable insight into the factors that affect sleep health. Beyond the analytical findings, the project makes a substantial contribution to the community by increasing awareness about the consequence of suboptimal sleep and proposing the development of a model to help others assess their sleep quality, supporting individuals in pursuit of a healthier and more restful life.

References:

[1] M. Kleinsasser *et al.*, “Detecting and predicting sleep activity using biometric sensor data,” *2022 14th International Conference on COMmunication Systems &amp; NETworkS (COMSNETS)*, 2022. doi:10.1109/comsnets53615.2022.9668347 https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9668347

[2] M. Kleinsasser *et al.*, “Detecting and predicting sleep activity using biometric sensor data,” *2022 14th International Conference on COMmunication Systems &amp; NETworkS (COMSNETS)*, 2022. doi:10.1109/comsnets53615.2022.9668347 https://europepmc.org/article/MED/27815231

[3] N. Kuzik, J. C. Spence, and V. Carson, “Machine learning sleep duration classification in preschoolers using waist-worn actigraphs,” *Sleep Medicine*, vol. 78, pp. 141–148, 2021. doi:10.1016/j.sleep.2020.12.019 https://www-sciencedirect-com.libaccess.sjlibrary.org/science/article/pii/S1389945720305694?via%3Dihub

[4] “Why sleep matters: Consequences of sleep deficiency,” Sleep Medicine, https://sleep.hms.harvard.edu/education-training/public-education/sleep-and-health-education-program/sleep-health-education-45 (accessed Oct. 28, 2023). https://sleep.hms.harvard.edu/education-training/public-education/sleep-and-health-education-program/sleep-health-education-45

[5] R. Osmun, “Oversleeping: The effects & health risks of sleeping too much,” Amerisleep, https://amerisleep.com/blog/oversleeping-the-health-effects/#:~:text=Other%20research%20indicates%20that%20getting,increased%20risk%20of%20developing%20dementia.&text=Oversleeping%20is%20considered%20a%20potential%20symptom%20of%20depression (accessed Oct. 28, 2023). https://amerisleep.com/blog/oversleeping-the-health-effects/#:~:text=Other%20research%20indicates%20that%20getting,increased%20risk%20of%20developing%20dementia.&text=Oversleeping%20is%20considered%20a%20potential%20symptom%20of%20depression.

[6] Centers for Disease Control and Prevention. (2014, May 16). *CDC - about BRFSS*. Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/about/index.htm