# Project\_4

Assignment - For this project we were tasked with using our machine learning skills to analyze, solve, or visualize a problem of our choice. Our team came across a dataset that was created by a research team from the University of Oxfordshire. Those researchers conducted a study to investigate the impact of lifestyle factors such as caffeine, alcohol, and exercise on sleep patterns and sleep quality. We wanted to see if we could use the dataset to see what impact those lifestyle factors had on sleep quality, and if we could use machine learning to be able to predict our own sleep quality based on our lifestyle factors.

Data - The research team recruited participants from the local community and collected data over a period of several months. The data was collected using a combination of self-reported surveys, actigraphy, and polysomnography which is a sleep monitoring technique. This means that the participants’ sleep was physically monitored by sensors and data was recorded. There were over 400 unique test subjects who participated in the research.

The dataset contains information about a group of test subjects and their sleep patterns. Each test subject is identified by a unique "Subject ID" and their age and gender are also recorded. The "Bedtime" and "Wakeup time" features indicate when each subject goes to bed and wakes up each day, and the "Sleep duration" feature records the total amount of time each subject slept in hours. The "Sleep efficiency" feature is a measure of the proportion of time spent in bed that is spent asleep. The "REM sleep percentage", "Deep sleep percentage", and "Light sleep percentage" features indicate the amount of time each subject spent in each stage of sleep. The "Awakenings" feature records the number of times each subject wakes up during the night.

Additionally, the dataset includes information about each subject's caffeine and alcohol consumption in the 24 hours prior to bedtime, their smoking status, and their exercise frequency. These are the lifestyle factors or as I would like to call them "input" factors for our machine learning algorithm.

Initial Cleaning of the Data - First we dropped the rows with null values, that was approximately 40 rows, bringing our dataset from 450 to roughly 400 participants. These null values were only in the self-reporting part of the dataset which were the Alcohol Consumption, Exercise Frequency, and Caffeine consumption columns. It is important to note that this was done within the excel csv itself because the dataset was small enough to make this the most time effective way to clean the data.

Encoding the Data - As we quickly learned, the machine learning algorithms did not respond well to the Bedtime Hour and Wakeup time columns as they were in the date / time format. So, we decided that it would be best to encode them into categories like "Normal", "Late", "Really Late" etcetera. The dataset made this easy for us as both variables were denoted in 30-minute indices. Additionally, both variables also had narrow timeframes, so we did not have to encode a lot of different categories for each variable.

Bedtime Hours

Normal – 9pm to 10:30pm

Late – 11 pm to 12:30am

Really Late – 1am – 2:30

Wakeup Hours

Very Early – 3am – 5am

Early – 5:30am – 7:30am

Normal – 8am-10am

Additionally, our target variable Sleep Efficiency was also formatted in a way that the machine learning algorithms did not like, they were formatted as numerical values from 0.5 to .99. This proved to be problematic because the machine learning algorithms that we had learned to use preferred their target variables to be in a binary format, Yes or No. So, we decided to encode the Sleep efficiency into a binary, anything less than .74 would be inefficient and anything more than .75, including .75 itself, would be efficient.

Sleep Efficiency

Efficient = .75 - .99

Inefficient = .50 - .74

All encoding was also done within the excel csv itself as this was the determined to be the quickest and most effective way to get this task done.

Additional Dataset - For this project we were tasked with having to use 2 different datasets. This was problematic for our team as our first dataset was hyper unique. This dataset was created by these researchers who conducted this study on 400 unique individuals, meaning the dataset is filled with completely unique data. After an exhaustive search we were unable to find any secondary dataset that would not only be able to merge on comparable variable, but also add to our research of studying the impact of lifestyle factors to sleep quality. So, we conducted our own sleep research mimicking the original dataset. We setup a google sheets document and each of us recorded data from 1 night’s sleep. We each input our own lifestyle data and our interpreted sleep efficiency as we unfortunately did not have machines to monitor our sleep quality. We then were able to concatenate this secondary dataset with the original therefore fulfilling the two-dataset requirement.

Get Dummies - Now that we had our dataset, we needed to do additional encoding. Five variables were classification variables which needed to be converted into numbers so that the machine learning algorithm could properly interpret them. Gender, Bedtime hour, Wakeup Time, Sleep efficiency (Only needed for Random Forest Machine Learning), and Smoking Status.

Additional Data Cleaning - Now that we had the new columns for the get dummies variables, we had to drop the redundant columns. The original 5 columns from the get dummies variables were dropped from the dataset. Additionally, it was decided that "output" variables in the dataset like REM sleep percentage, Deep Sleep percentage, Light Sleep percentage, Awakenings should be dropped so that the machine learning algorithm could focus on the 1 output factor that we wanted to study which was Sleep efficiency.

Variable Correlation to Sleep Efficiency - Once we had a fully cleaned and encoded dataset, we used python to calculate the amount that each variable was correlated with our target variable Sleep efficiency. The results showed that not a single variable was strongly correlated with sleep efficiency more than the minimum statistical threshold of 0.3. All except for 1, Alcohol consumption was negatively correlated at a statistically medium significance of -0.43. This means that the more drinks a person has, the worse their sleep efficiency will be.

|  |  |
| --- | --- |
| Sleep Efficiency | 1 |
| Exercise frequency | 0.222205 |
| Non-Smoker | 0.17351 |
| Normal Wakeup | 0.097512 |
| Really Late Bedtime | 0.085458 |
| Caffeine consumption | 0.050481 |
| Very Early Wakeup | 0.020278 |
| Late Wakeup | 0.011038 |
| Female | 0.008966 |
| Male | -0.008966 |
| Normal Bedtime | -0.018793 |
| Sleep duration | -0.02538 |
| Late Bedtime | -0.063758 |
| Early Wakeup | -0.11724 |
| Smoker | -0.17351 |
| Alcohol consumption | -0.434161 |

However, it's also important to note that none of these variables that were denoted in the study are isolated. Being a smoker and exercising is not mutually exclusive. Neither is drinking and not exercising. When you look at the variables like that, it becomes obvious that the variables can be added up to be statistically significant. Therefore, the combination of lifestyle factors can determine one’s sleep efficiency. And when you look at the chart of correlations a clear pattern emerges, those that live a "healthy" lifestyle are more likely to have better sleep efficiency. Our machine learning algorithms were able to prove this.

Caveat - It is important to note that the quality of one’s sleep is dependent on a multitude of factors, some of which were not recorded in this study. For example, stress, or noise, or the current condition of your health.

Machine Learning:

KNN - We started with the KNN machine learning model. Our first run after training and scaling the data yielded a very low accuracy score, somewhere in the 60% neighborhood. This was mostly due to our random\_state variable being adjusted. We experimented with a lot of different numbers and ultimately settled on a random\_state variable of 1 which stabilized our accuracy to above 70%.

Additionally, we experimented with dropping some of the more neutrally correlated variables like Gender and Age. When we dropped gender our accuracy score dropped significantly. So, we added that back in. However, when we dropped the Age variable our accuracy jumped up.

Unfortunately, with the KNN method we were only able to achieve an accuracy score of 74%.

Random Forest – We then decided to use the Random Forest method of machine learning. This required us to use get dummies to encode our target variable “Sleep efficiency” from words (Efficient / Inefficient) to an integer (1/0). With that done we were able to get accuracy scores in the 75+ range. Again, it was the changing of the random\_state variable that got us to 83%.

User Input - Now that we had a model with fairly high accuracy rate. We decided to code a way for a user to input their own data. We created a variable to hold our answers, then used a while loop to ask a series of user input questions. These questions were designed with the machine learning dataset in mind, so all the questions require a numerical value to be entered. Also, because of the get dummies variables, the same question is asked multiple times in different ways. For example, “Are you a male? Enter 1 for Yes and 0 for No.” This is followed by essentially the same question “Are you a female? Enter 1 for Yes and 0 for No.”

Once all the questions are answered, the array is reshaped so that the machine learning algorithm can understand it, it then takes those inputs and predicts the outcome based on our machine learning algorithm. Of course, it is able to do this with 83% accuracy.