ECS 171 Group 10 Project

Leader: Ryan Yu

Amira Basyouni, Calvin Chen, Alexis Lydon, Tianming Tan

Github Repository: `https://github.com/ryyu444/ECS-171-Group-Project`

November 11, 2023

Introduction and Background

Literature Review

Machine Learning has great capabilities within the realm of pattern recognition and categorization, two fields which we hope to utilize as we explore the connection between sleep and stress. It is important to address some of the pre-existing studies which were conducted with a similar objective. Presented below are examples of such research and methodologies as well as their overall findings.

Jayawickrama and Rupasingha, researchers from Sabaragamuwa University of Sri Lanka, conducted a study to examine the impact of sleep habits on stress levels. (1) They applied various machine learning models, including Naïve Bayes, Random Forest, Decision Trees, Multi-layer Perceptrons (MLPs), Support Vector Machines (SVMs), and Logistic Regression mixed with Cross Validation. It was revealed that five of the six models achieved accuracy rates exceeding 80%, suggesting that there is a strong correlation between sleep and stress. Notably, Naïve Bayes was the best predictor with an accuracy of 91.27%.

A different research article called "sleepWell: Stress Level Prediction Through Sleep Data" was predicting stress levels using data collected from sleeping participants. (2) Some of the models they utilized include Gradient Boosting, Decision Trees, Random Forest, Gaussian Naive Bayes, and Linear Support Vector Machine. Their models achieved an accuracy rate of over 95%, with Naive Bayes and SVM as the top performers. Although the focus of this research focused on data that predicts stress during sleep, our model will focus on predicting stress of participants who are awake.

Dataset Description and Exploratory Data Analysis

The dataset being used in this project is the "Sleep Health and Lifestyle Dataset" by Laksika Tharmalingam on Kaggle [link to reference this dataset]. The dataset comes in the form of a CSV file and contains 400 rows and 13 columns, encompassing various sleep and lifestyle variables. These include gender, age, occupation, sleep duration, sleep quality, physical activity, stress levels, BMI category, blood pressure, heart rate, daily steps, and sleep disorder status. We chose this dataset because, as students, we can relate to the common sleep issues and stress levels many of us face. This dataset encompasses the aspects of sleep and health that can be useful to predict stress. Our project focuses on investigating the relationship between sleep and health, making a dataset on sleep and health the most suitable choice. A limitation of the dataset is its synthetic nature, generated artificially rather than from real-world observations. While it may lack some real-world nuances, high-quality synthetic data can effectively train and test machine learning models in this context, and can be a step in the right direction to begin to build models based off of real data collected.

First, some basic data preprocessing needed to be performed in order to get all the categorical data into numerical values, which are easier to work with in prediction models. For this, we performed label encoding on the categorical attributes from the dataset, including BMI and Gender. We chose label encoding rather than one hot encoding in order to keep the number of columns or attributes the same in our data set and to see a better representation in our pair plot and correlation matrix. After the data preprocessing, we created a pair plot matrix in order to see if there were any linear relationships between attributes in predicting Stress Level. From the pair plot, it was

easy to see that Sleep Quality with Sleep Duration and Sleep Quality with Age show a linear relationship when predicting Stress Level. Sleep Duration and Gender also have somewhat of a linear relationship in predicting stress. Next, we made a correlation matrix in order to see attributes that had high correlations with each other, and more specifically with Stress Level since this is the dependent variable for our models. It was easy to see similar relationships as the pair plot, and Sleep Quality and Stress Level had the strongest correlation at -0.9. Stress Level and Sleep Quality also had a high correlation at -0.81. What was also important to note, is that Sleep Quality and Sleep Duration have a high correlation with each other, at 0.88. From this knowledge, it became clear that Sleep Quality and Sleep Duration should be our main predictors of stress in our models. It also became clear that using these together to predict stress could also be beneficial, as well as possibly using Age and Sleep Quality.

We also performed some outlier detection on these important variables using box plot models. We found no significant outliers in the predictor variables or in the dependent variable Stress Level.

Proposed Methodology

Experimental Results

Conclusion and Discussion

References