Wearable Health Predictor Project Proposal

Authors: Jason Vasquez, Dylan Skinner, Jeff Hansen, Dallin Stewart

Project Overview

We will explore factors that contribute to human health using a dataset created by wearable health and fitness trackers on Kaggle. We will answer questions such as the following. What is the most impactful change for an individual to make to improve their health? What are the best predictors of sleep quality? What medical condition does an individual have? What type of exercise is an individual performing at a given time?

We will use regression techniques to find correlations between features in the data and the values we are interested in such as health and sleep quality. We will also use random forests and k-means clustering to make predictions and classifications about these values. We will evaluate our answers with confusion matrices enabled by making train-test splits and correlation coefficients. If they are sufficiently large and statistically significant, we can be more confident in the reliability and accuracy of the answers. We will divide the work by assigning ourselves to analyze individual research questions and develop different common helper functions.

Data Summary

The dataset has 10,000 samples with over 40 features including age, gender, weight, height, medical condition, sleep quality, stress, health score, steps, exercise, and caloric intake among others. Subsets of these features will help answer each of our research questions. While we are most likely to use these features, we will omit features like notifications, screen time, and user ID. We will use user ID as a unique key to combine the three datasets into a single dataframe with an inner join. After joining the data, we will clean the data by putting continuous data into bins, one-hot encoding the categorical data we are interested in, and using regression to identify important features and assist in our feature engineering.