DS 3010: Applied Data Modeling and Predictive Analysis (Fall 2025) Project Proposal

1 Project Title

Can we predict quality of sleep and presence of sleep disorders based on health and lifestyle factors?

2 Team Members

- Nina De Grandis
- Lucas Sorge

3 Project Details

3.1 Project Objective

- The project objective is to see if machine learning can help predict sleep disorders and the quality of sleep based on lifestyle and health factors.
- The problem we will be trying to solve is whether the above, and if we can use this data to improve people's sleep habits and quality.
- This problem is important because many people have trouble sleeping, which can affect their day to day life. At the end of this project, we will be able to suggest some ways to improve sleep quality and duration.
- Machine learning can help this problem because it can take the data we have so far, and see if we can predict whether this person will have a sleep disorder or not based on their lifestyle and health factors. We will also try to predict their projected quality of sleep.

3.2 Datasets

Please describe your dataset in this section.

- The data set is called "Sleep Health and Lifestyle Dataset" and we found it on Kaggle.
- The data was collected by a Kaggle user from a survey, it is a few years old.
- We will be using all of the features listed which are:

  Person ID: An identifier for each individual.

  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 (e.g., Underweight, Normal, Overweight).

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).

- We will split the dataset into 80% training and 20% testing.

3.3 Machine Learning Algorithm

We will apply two baseline algorithms aligned with our project objectives:

1. Logistic Regression for predicting the presence of a sleep disorder (classification). Logistic regression is appropriate because the target variable is categorical (None, Insomnia, Sleep Apnea). It is interpretable, handles both continuous and categorical features, and provides coefficients that show which lifestyle or health factors most strongly influence the likelihood of a disorder.

2. Linear Regression for predicting the quality of sleep score (1–10 scale, regression). Linear regression is suitable for continuous outcomes and will allow us to estimate the quantitative effect of predictors such as activity level, stress, and blood pressure on sleep quality.

If dataset size or scope limitations make it more practical to focus on a single prediction task, we will prioritize logistic regression for sleep disorder classification, since classification is often more actionable for identifying at-risk individuals.

3.4 Expected Outcomes

Our expected outcome for this project is that increased levels of physical activity and daily steps and decreased levels of stress, blood pressure, and heart rate will lead to a better quality of sleep and a less likely presence of a sleep disorder. We do not think that demographic factors such as gender, age, and occupation will have an effect on the outcome.