Advanced Machine Learning Project Proposal

To: Prof. Ghosh

From: Jagruta Advani(ja53837), Dameli Aziken(da35254), Samuel Chen (swc872), Shaunak Divine (jsd2672), and Kennedy Zapalac (kmz459)

Subject: Adv. Machine Learning Final Project Proposal

Date: 10/23/2024

Problem Statement (Background)

In recent years, problematic internet usage has become a significant problem for youth development worldwide. Problematic internet usage is often studied in relation to depression and anxiety, but it has also been related to a lack of physical exercise. Current methods to study problematic internet usage involve time intensive surveys that may place an undue burden on research participants. Although it may require more data storage and processing power, there is an incentive to use passive ambulatory assessment to detect problematic internet usage. Ambulatory data is continuously or regularly collected through mobile devices or wearables, and it is passive when users don't have to do anything for the data to be collected. This form of data collection reduces the burden on research participants, and it can be used to assess internet usage in a naturalist context rather than in the laboratory. Particularly given that problematic internet usage is related to reduced physical activity, it is of interest to relate survey data on problematic internet usage to physical activity data collected through wearables. Once the relationship between everyday physical activity and internet usage is better understood, then data collected through wearables could be used to improve quality of life, particularly for impressionable children, by providing timely suggestions to engage in physical activity and disconnect from the internet. Given this background, the problem we are attempting to solve is to predict, with relative accuracy, the rate of problematic internet usage based on user physical activity.

Dataset Chosen

We obtained our dataset from the Kaggle Data Competition titled "Child Mind Institute–Problematic Internet Use," which focuses on understanding the data behind problematic internet usage patterns among children. We have data on approximately 5,000 children aged 5-22 from a clinical study from the Healthy Brain Network (HBN). Our data is subdivided into two types: tabular data and actigraphy data. We have training data for 3,960 participants in the tabular section and 996 participants in the actigraphy section. The tabular data contains 82 features from participant responses to surveys about their:

- Demographics: Basic information like age and sex

- Mental Health and Internet Usage: Internet usage patterns and mental health scores (Children's Global Assessment Scale)
- Physical Activity and Fitness: Metrics such as blood pressure, heart rate, height/weight, and evaluation of physical fitness relating to aerobic capacity, strength, and flexibility (FitnessGram Test)
- Bio-Electric Impedance Analysis: Metrics such as BMI and fat percentage
- Sleep Patterns: Measured by Sleep Disturbance Scale and Physical Activity Questionnaire

The actigraphy data is stored in parquet files for each participant. There is up to 30 days of actigraphy data for each participant, containing, but not limited to:

- Movement Data: Acceleration across X,Y, and Z planes
- Motion Intensity: Metric of how intense the user's motion is measured with Euclidean Norm Minus One (ENMO)
- Light: A measure of light exposure in lux
- Wearable Statistics: Measures to track the physical device including power status and whether the participant was wearing the watch
- Time Data: Time-related metadata to contextualize the other information from the wearable

Our Approach

First, we'll conduct data exploration, cleaning, and feature engineering to understand the data better and prepare it for analysis. Raw actigraphy data would be more useful if it was transformed into aggregate sleep and activity measures for each participant. There are sleep scoring algorithms and physical activity thresholds, such as through the "actigraphr" package in R, that we will try for this purpose. We will also carefully consider how we are going to deal with missing values through imputation. Additionally, some of the labels, Severity of Internet Usage (SII), are missing, so we will undertake semi-supervised methods to complete the dataset. One such semi-supervised method we could use is label propagation, which involves clustering participants and replacing missing labels with labels from similar participants.

Moving onto our analysis, we are initially interested in understanding indicators of problematic internet usage. We will model the relationship between the tabular data and our target variable, SII, using interpretable models such as logistic regression or decision trees. We will include a form of variable selection such as regularization to prevent overfitting since we have many predictors. Trees are particularly useful for the missing values, categorical variables, and potentially non-linear relationships in this dataset.

Once we understand the indicators of problematic internet usage, we would like to relate the actigraphy data to the SII. The indicators we previously found will help us to transform the

actigraphy data in ways that map onto our key indicators. Here we will focus more on accuracy than interpretability since that is the goal of the competition. We will try models such as XGBoost and Neural Networks because of their lack of assumptions and flexibility in modeling relationships, and we'll choose a final model based on test error. To improve accuracy and overcome the class imbalance problem, we may use resampling techniques too.

Key Challenges in Addressing the Problem

This project encounters several critical challenges. The first is the issue of **missing data**, often due to technical failures, skipped survey questions, or incomplete participation in assessments. Properly imputing these missing values is crucial to avoid introducing bias or compromising the validity of the model. Ensuring the accuracy of imputation is especially important given the clinical nature of the dataset.

The next challenge is the **heterogeneity in responses** across a diverse age range (5-22 years). Participants vary significantly in their physical activity, internet usage, and survey responses. This variability complicates **feature selection** and makes it difficult to generalize the model across different age groups without risking overfitting or underfitting the data.

In addition, the dataset includes over 80 variables derived from various assessments, questionnaires, and actigraphy data, making **feature selection** a complex task. Identifying which features contribute meaningfully to the prediction while filtering out noise is essential. The mix of continuous and categorical data further complicates the engineering and interpretation of the features.

Finally, **subjectivity in self-reported data** poses a challenge. Participants may misreport their internet usage or fail to accurately recall their online habits, leading to inconsistencies in the dataset. These inaccuracies introduce noise, which can reduce the predictive accuracy of the model, especially when relying heavily on subjective input.

References

https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use/data

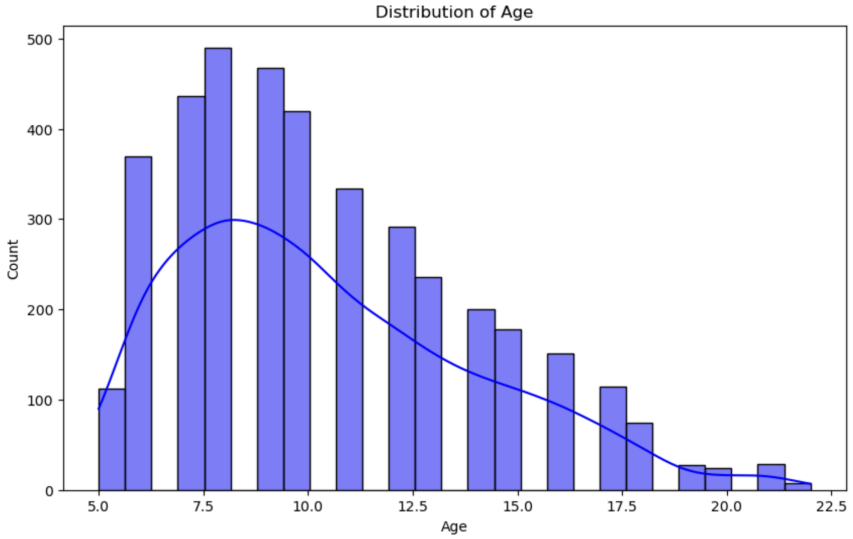
https://github.com/TheTS/actigraphr

Emma Louise Anderson, Eloisa Steen, and Vasileios Stavropoulos. "Internet Use and Problematic Internet Use: A Systematic Review of Longitudinal Research Trends in Adolescence and Emergent Adulthood." *International journal of adolescence and youth* 22.4 (2017): 430–454. Web.

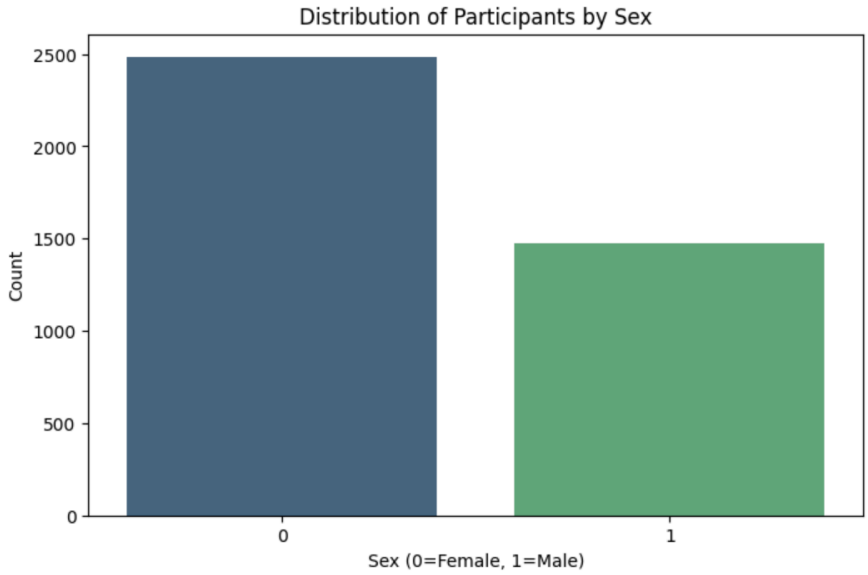
Trull, Timothy J, and Ulrich Ebner-Priemer. “Ambulatory Assessment.” *Annual review of clinical psychology* 9.1 (2013): 151–176. Web.

Appendix

Participant Demographics

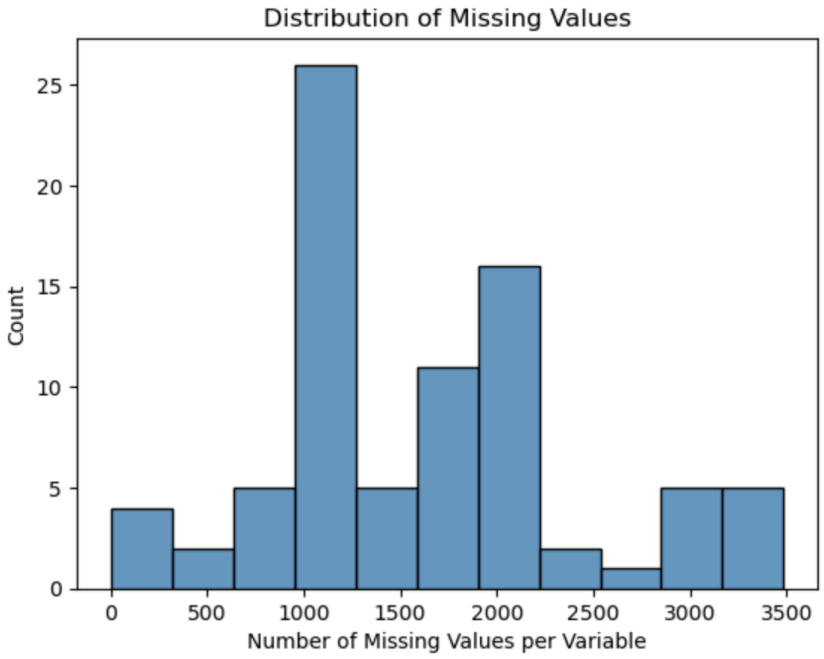


We have a young demographic with a right skew.



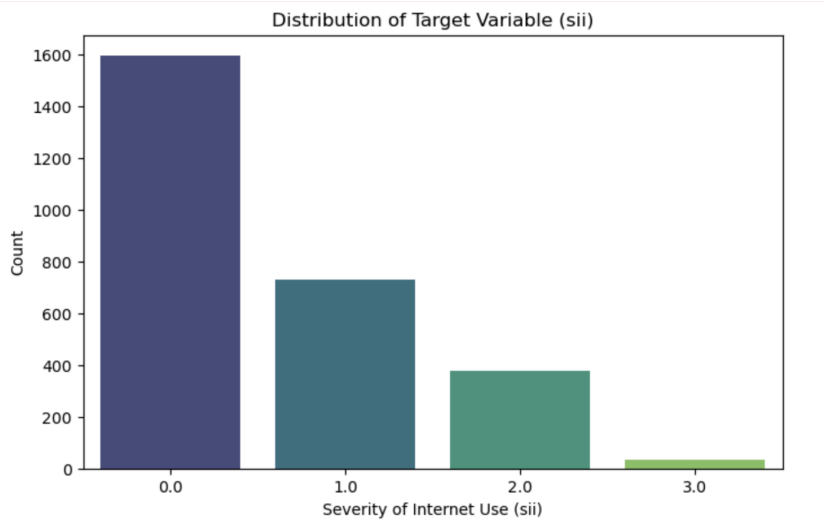
We have more female than male participants.

Missing Data



Some variables are almost completely missing since they have 3000/3960 observations missing.

Class Imbalance



Few participants had severe internet use, which complicates our prediction of SII.