**Executive Summary: Problematic Internet Usage**

Aaron Weinberg, Emilie Wiesner, Dan Visscher  
GitHub: https://github.com/aarondweinberg/CMI\_problematic\_internet\_use/

**Introduction**

Internet use has been identified by researchers as having the potential to rise to the level of addiction, with increased rates of anxiety and depression associated with problematic internet use. Identifying cases of problematic internet usage currently requires evaluation by an expert, however, which is a significant impediment to screening children and adolescents across society. One potential solution is to rely on an assortment of data that is more easily and uniformly collected: the kind collected by a family physician or by a smartwatch. The research question this project sets out to answer is: “Can we predict the level of problematic internet usage exhibited by children and adolescents, based on their physical activity?”

**Dataset**

The data comes from a research study conducted by the Child-Mind Institute, with data for 3960 participants. The target variable is a severity impairment index (SII) that measures problematic internet use on a scale from 0 (no impairment) to 3 (severe impairment). There are 33 predictor variables, including demographics (e.g., age, sex), physical measurements often taken by a family physician (e.g., height, weight, blood pressure), results of a fitness test (e.g., sit & reach, endurance time), survey responses and scales (e.g., internet usage in hours per day, sleep disturbance, children’s global assessment), and measures from a bio-electric impedance analysis (e.g., bone mineral content, fat mass index). Additionally, about one month of accelerometer data was provided for almost one thousand of the participants in 5-second intervals. Significantly, participants often completed some parts of the study but not others, so any given participant has entire groups of variables missing.

**Preprocessing and Exploratory Analysis**

Participants were dropped if they did not have an SII score (about a third of the participants). The distribution of remaining participants across SII scores is very skewed: over half of participants had an SII score of 0, while only 30 (about 1% of) participants who were measured with an SII score of 3.

Among the 3000 cases with SII scores, every case was missing data for at least one variable, creating a need to impute predictor variable data. We consider different imputation methods as part of our models.

**Model Selection and Results**

Cohen’s quadratic weighted kappa function measures the accuracy of prediction for ordinal variables; random guessing (or uniform guessing) produces a score of 0, perfect prediction produces a score of 1, and . We couldn’t directly incorporate this into our model training as the loss function, but we did use this to compare models.

Because there was so much missing data, we needed to use imputation, rather than just remove cases with missing values.

Recall that the PCIAT questionnaire is the basis for the SII score, our ultimate target variable. We eliminated all cases that had no PCIAT data, and then imputed any remaining missing data for individual PCIAT questions. All questions were scored on the same scale, so we used KNN imputation here.

For the predictor variables, we tested both KNN and a MICE imputation. Some of the variables, various fitness zones, were derived from the fitness test quantitative measurements. Rather than imputing the zone values, we computed these separately after running the imputation.