UIUC CS 412 Project Proposal

Evan Matthews[1], Vikram Ramavarapu[1], and Krishnaveni Unnikrishnan[1]

[1]Siebel School of Computing and Data Science

October 7th, 2024

1 Introduction

The internet has become an integral part of our daily lives, with people of all ages spending a significant amount of time online. This gives rise to concerns about the potential negative impacts of excessive internet use, particularly on children and teens. Problematic Internet Use (PIU) is a condition characterized by excessive or poorly controlled preoccupations, urges, or behaviors regarding computer use and internet access that lead to impairment or distress [3]. PIU has been associated with a range of mental health issues, including depression, anxiety, and impulsivity [2]. Identifying early signs of PIU in children and teens is crucial for prevention and intervention. In this project, we aim to predict early signs of PIU in children and teens using machine learning techniques, leveraging data from the Child Mind Institute's Healthy Brain Network. The project plan consists of three phases: data preprocessing, initial model evaluation, and fine-feature reevaluation. This project is a submission to the Child Mind Institute's Kaggle competition on PIU prediction, and we aim to further publish our results if they outperform competition expectations.

2 Motivation

Despite having multiple studies showing the negative effects of excessive internet use, exact details relating to early signs of PIU and the most at-risk

age groups are still unknown. First, most studies have been conducted on early development groups like children, teens, and young adults [1–4], but a more specific age group could help in preventing and treating PIU early. Second, studies of this nature focus primarily on written or binary feedback from students or parents. These studies can be useful, but they also introduce biases and often fail to show the factors which correlate a participant's estimated internet impact [1, 4].

A major drawback of assessing PIU is in its subjective nature. Problematic internet use is characterized by many different variables that are hard to measure. As such, using quantitative measures such as the Severity Impairment Index (SII) allow for the proper measurement and diagnosis of PIU.

The use of quantitative measures such as the SII allow for the application of data mining methods to aid in the classification of PIU severity. Moreover, other measurable attributes such as sleep quality and duration, physical activity level, and duration of internet usage can all be used to understand correlations with PIU. This project intends to rectify these issues by using a machine-learning approach to predict early signs of PIU using a wider range of variables on children and teens.

3 Data

Our project background and necessary training data originates from the Child Mind Institute (CMI) and the ground-breaking Health Brain Network (HBN) study. CMI has provided these items as part of an ongoing Kaggle competition for PIU prediction given features from children's physical activity. This dataset consists of 81 unique features, including categorical and scoring integers, scoring floating points, and categorical string labels [5]. Data is in timeseries format, per participant, and includes information on physical activity, sleep quality, and duration, and internet usage duration. Given the competition and CS 412 project deadlines, we are committed to a full submission of our project to the Kaggle competition and further publication should intriguing results arise.

4 Plan

Our project development timeline consists of three main phases: data pre-processing, initial model evaluation, and fine-feature reevaluation. In the data preprocessing phase, we will clean up the dataset, cast all features to consistent, comparable types, and determine how much of the dataset can be used in supervised and semi-supervised tasks. Certain values of fields such as 1 for the `no-wear` field or `low` for the battery level field for the heartrate monitor may be excluded from the dataset to avoid skewing the results. Ensemble and decision tree models (XGBoost, Random Forest, Decision Trees, etc.), as well as dimensionality reduction techniques (PCA, LDA, etc.), will be used to determine the most important features for prediction. In the initial model evaluation phase, we will train on all available clean features and submit to Kaggle for initial review. We will also use this phase to perform multi-dimensional comparisons of all features to determine which features are most important for an optimal prediction. Finally, in the fine-feature reevaluation phase, we will retrain our model on a finer set of features as an optimal submission to the Kaggle competition. As previously mentioned, further publication will be considered if results outperform competition expectations.

References

[1] Elias Aboujaoude. Problematic internet use: an overview. *World Psychiatry*, 9(2):85–90, June 2010.

[2] Hilarie Cash, Cosette D Rae, Ann H Steel, and Alexander Winkler. Internet addiction: A brief summary of research and practice. *Curr. Psychiatry Rev.*, 8(4):292–298, November 2012.

[3] Mauro Pettorruso, Stephanie Valle, Elizabeth Cavic, Giovanni Martinotti, Massimo di Giannantonio, and Jon E Grant. Problematic internet use (PIU), personality profiles and emotion dysregulation in a cohort of young adults: trajectories from risky behaviors to addiction. *Psychiatry Res.*, 289(113036):113036, July 2020.

[4] Anita Restrepo, Tohar Scheininger, Jon Clucas, Lindsay Alexander, Giovanni A Salum, Kathy Georgiades, Diana Paksarian, Kathleen R

Merikangas, and Michael P Milham. Problematic internet use in children and adolescents: associations with psychiatric disorders and impairment. *BMC Psychiatry*, 20(1):252, May 2020.

[5] Adam Santorelli, Arianna Zuanazzi, Michael Leyden, Logan Lawler, Maggie Devkin, Yuki Kotani, and Gregory Kiar. Child mind institute — problematic internet use, 2024.