Project: Midterm Report

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

[1]CS 412 Group G6

November 6th, 2024

1 Abstract

summarizing the project [1–4].

2 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 trend has given rise to concerns about the potential 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]. As such, 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. We will submit our work to the Child Mind Institute's (CMI) Kaggle competition on PIU prediction, and we also aim to publish our results as a paper should they outperform competition expectations.

3 Motivation

TODO: A few sentences on why the project is of interest from a data mining and/or real world application perspective.

With the rise of machine learning and pattern prediction models, the ability to analyze and predict upon more complex data and parameters becomes much more approachable. Likewise, child development is a multi-facted situation in which parenting and environmental factors can lead to an incredibly high number of outcomes. This field has had great strides in classical research, but a more modern approach could lead to significant development in the success of future generations. Additionally, predictions against an extensive number of possible outcomes like this represents a current roadblock in machine learning- that is, how modern predictive models can adapt to an ever-increasing set of parameters and decreasing set of training data. Finally, child psychology is interested in recognizing patterns in early behavior in order to reduce the impact of harmful effects from a child's environment.

4 Related work

in the literature, and in kaggle/related forum. Having just 1-2 references or no references to papers/books will lead to low scores. TODO: Add references to related work

5 scope

TODO: Any change in scope from original proposal, please see guidelines above.

Given that the original scope of the project was accepted, we are pressing forward with this plan with no significant changes. The most crucial critique provided- that the validation plan and evaluation metric were not clear- are likewise addressed in the methodology section.

6 Methodology

Data for this project has two components: cross-sectional, and time-series. The cross-sectional data is per participant and contains fields described in the following table (etcetc). Each time-series dataset is per participant and each entry of the dataset represents the status of the participant's heartrate monitor at a given point in time.

The project will be divided into three phases: data preprocessing, initial model evaluation, and fine-feature reevaluation. The data preprocessing phase entails dropping fields where survey responses are recorded and are then used to compute the SII, as our model's intention is to compute SII directly from the other metrics. Missing values in the data are filled using iterative imputation, and the missing SII values are filled in using K-Nearest Neighbors (k=5).

Multiple models will be evaluated on the cross-sectional data: Random Forest, XGBoost, SVM, and a feed forward neural network. After this, a sequential model, evaluated amongst transformers or auto-encoders, will be trained on the time-series data. The sequential model will allow us to compute an embedding of the time-series data, which will be used as an additional feature in the cross-sectional model. The final model will be an ensemble of the cross-sectional and sequential models, with the sequential model's embedding as an additional feature in the cross-sectional model. The classifier model will be retrained on the concatenated dataset, to predict the SII.

7 (Current / Preliminary) Results

TODO: what you have so far in terms of initial results and analysis of initial results. Please see comment on figures/tables above, especially the fact that good captions go a long way to making things readable.

Add results here

8 Plan of Work

TODO: what are the next steps before the final report. Please be as precise as possible. Note that you will have about a month to finish the project, so make suitably calibrated plans, e.g., do not over/under promise.

Add plan of work here

9 Conclusions, discussions

add conclusions here

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.