Analysis of Children's Problematic Internet Use

Data Science 2 - Final Project

Chase Ivancic, J. Simon Richard, and Quynh Tran

Introduction

Background Information

The Child Mind Institute's Problematic Internet Use dataset, known officially as the Healthy Brain Network (HBN) dataset, is a sample of five-thousand 5-22 year olds who underwent clinical and research screenings. The goal of this dataset is to use the research findings to appropriately identify those with unhealthy internet habits. The classifier is a categorical variable called the participant's Severity Impairment Index (sii). This variable rates each participants' problematic internet use on a scale from 0 (None) to 3 (Severe).

Dataset Description

The Severity Impairment Index is calculated using one of the assessments within the dataset. The Parent-Child Internet Addiction Test (PCIAT) is a 20-question long test taken by the parent of the participant. The questions contain information relating to internet use and problematic behavioural symptoms. Each question prompts a response of zero (Does Not Apply) through five (Always) and sums the total together for a score out of 100. This score is then divided into the four sii variables.

Another assessment present in the dataset is the Children's Global Assessment Scale. This scale is used by mental health clinics to rate the general functioning of youths under 18. The rating is on a scale from 0 to 100. This feature is described in more detail in our exploratory data analysis and is used to answer the first of our questions listed below.

Questions

1. Does the CGAS depend on the season it was administered?
2. Do activity scores tend to diverge before/after the PCIAT is administered?
3. Does PCIAT correlate with age?
4. Can we predict SII without using a complex ML / NN model?

Exploratory Data Analysis

Children's Global Assessment Scale

The Children's Global Assessment Scale (CGAS) is a "numeric scale used by mental health clinicians to rate the general functioning of youths under the age of 18" [1]. It ranges between 0 and 100, with values from 1 to 10 corresponding with "needs constant supervision" and values from 91 to 100 corresponding with "superior functioning" [2].

When looking at the distribution of scores within the CMI dataset, we found a single data entry error recording a CGAS score of 999. After replacing that with `NaN`, we plotted the feature's distribution (Figure 1). Based on this plot, it seems that the distribution is similar to a normal distribution, but it includes some "psuedo-quantization." That is, round numbers (multiples of 5) are recorded far more frequently than the surrounding numbers. This is unsurprising since this is a score assigned by human clinicians.

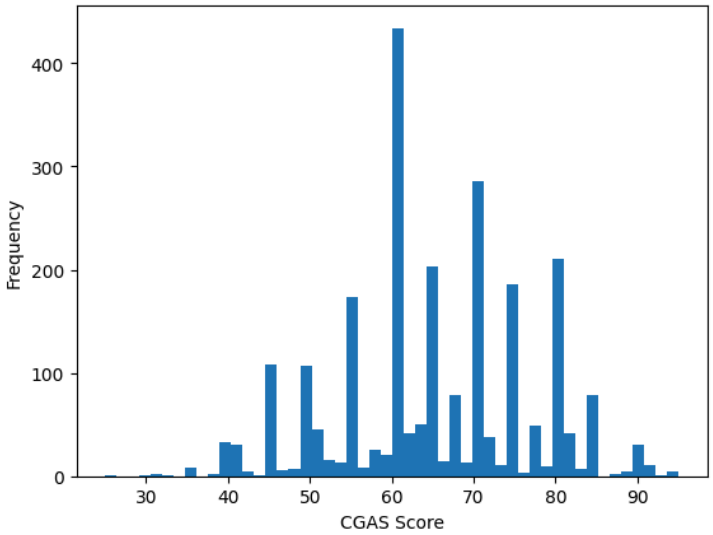

Figure 1: CGAS Distribution

As a part of our analysis of CGAS, we tested whether the score is correlated significantly with the season during which it was recorded (question 1). Using a one-way ANOVA test, we found that there is in fact a significant correlation ($p = 0.0156$). and further pair-wise T-tests revealed that Winter was the "odd-one-out" as illustrated in Figure 2.

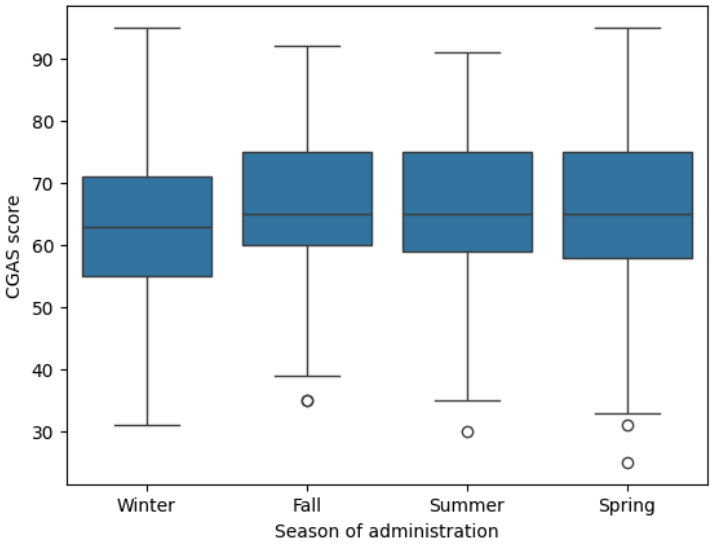

Figure 2: CGAS distributions during different seasons.

Bio-electric Impedance Analysis

The measurements of 16 key body characteristics including BMI and muscle content were measured for some of the patients within the CMI's dataset. A complete list of features is included below.

- Activity Level
- Bone Mineral Content
- Body Mass Index
- Basal Metabolic Rate
- Daily Energy Expenditure
- Extracellular Water
- Fat Free Mass
- Fat Free Mass Index
- Fat Mass Index
- Body Fat Percentage
- Body Frame
- Intracellular Water
- Lean Dry Mass
- Lean Soft Tissue
- Skeletal Muscle Mass
- Total Body Water

Because we had limited time and because we don't have domain experience, we opted to remove all extreme outliers (under the 2nd percentile or over the 98th) instead of researching reasonable ranges for each feature. Following that, we plotted the distribution of each (Figure 3).

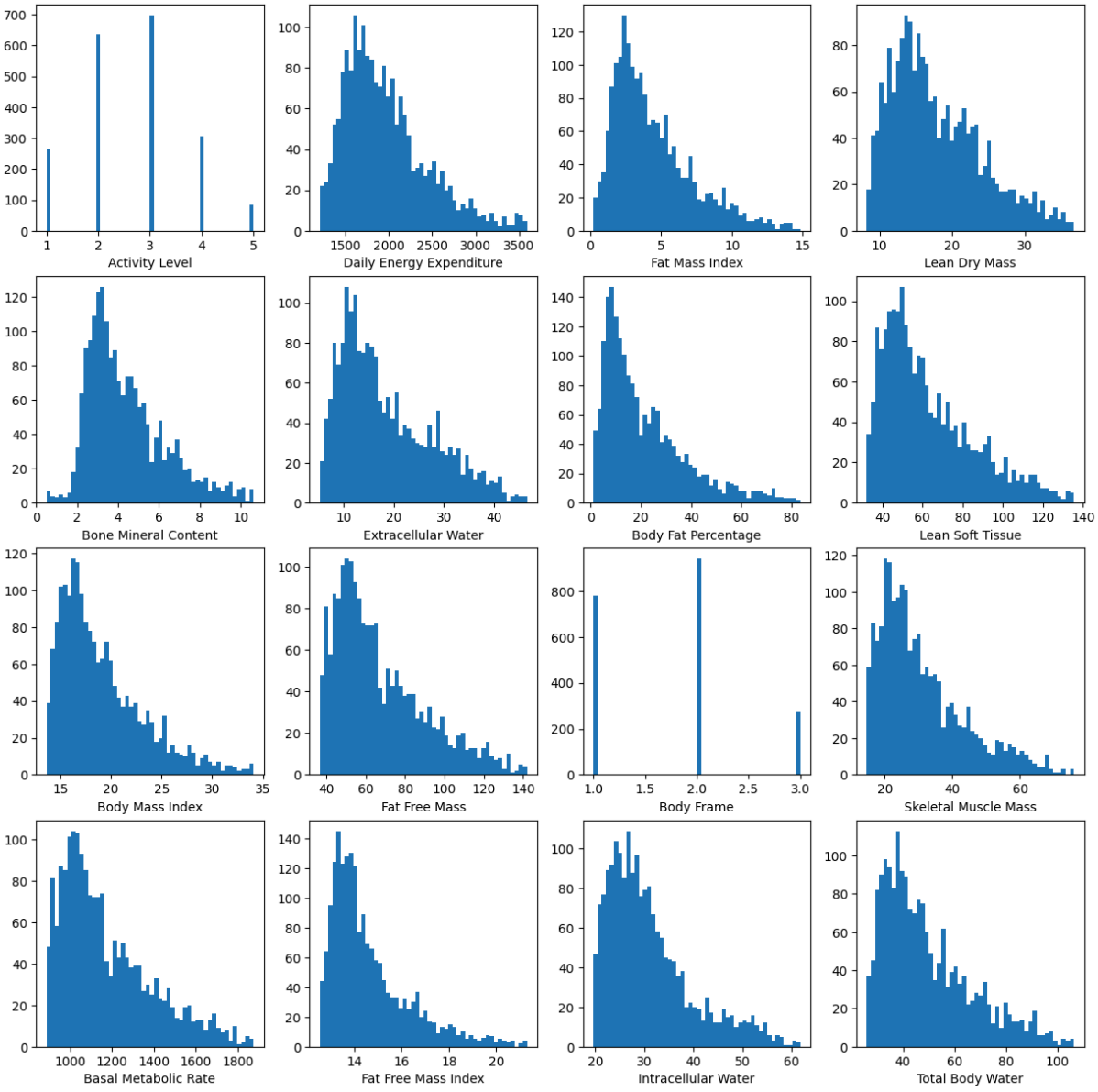

Figure 3: Distributions of the Bio-electric Impedance Features

These plots show that each of the bio-electric impedance features are skewed right. We also tested for variations between the seasons using one-way ANOVA, but found no significant difference.

Actigraphy

Some patients in the CMI's dataset wore fitness watches recording acceleration data and ambient light levels (which could be used as a proxy for whether the patient is inside or outside). The data collected exceeded 6GB, which made analysis quite difficult. For this reason, we did not use this data to train our final model for question 4. However, we still investigated question 2, which requires the use of this data.

The distribution of lengths for which patients wore these devices is visualize by Figure 4.

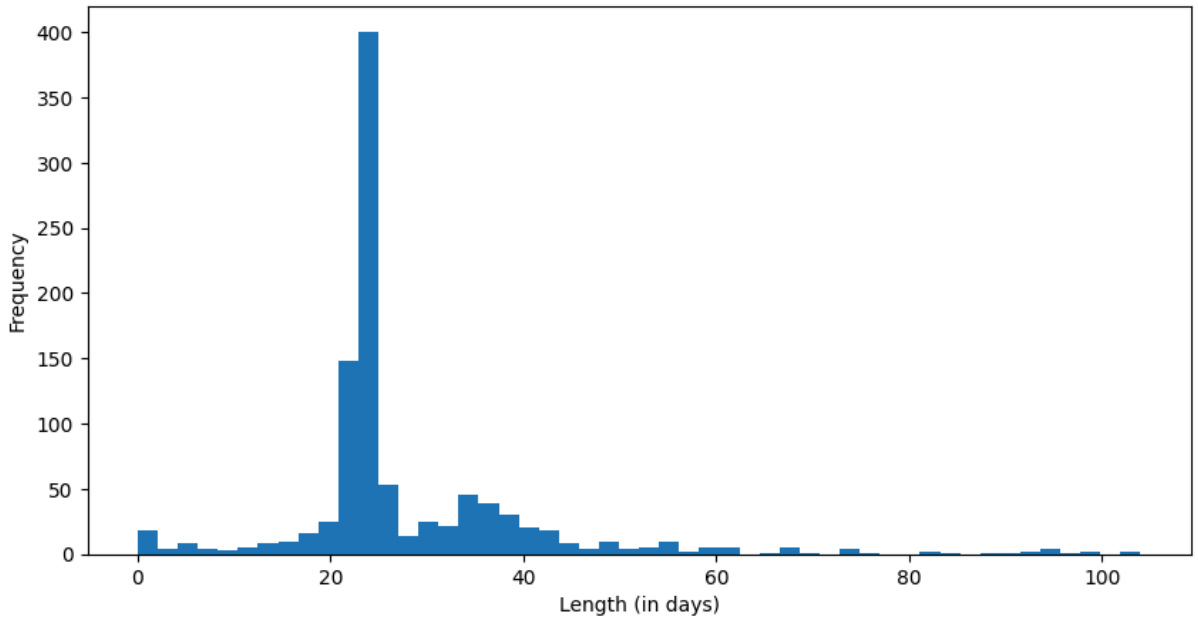

Figure 4: Distribution of lengths for which patients wore their Actigraphy devices

We also visualized the Euclidean Norm Minus One (ENMO) feature, which is derived from the accelerometer's three-axis acceleration data and used to represent general levels of activity. A single patient's ENMO is plotted in Figure 5.

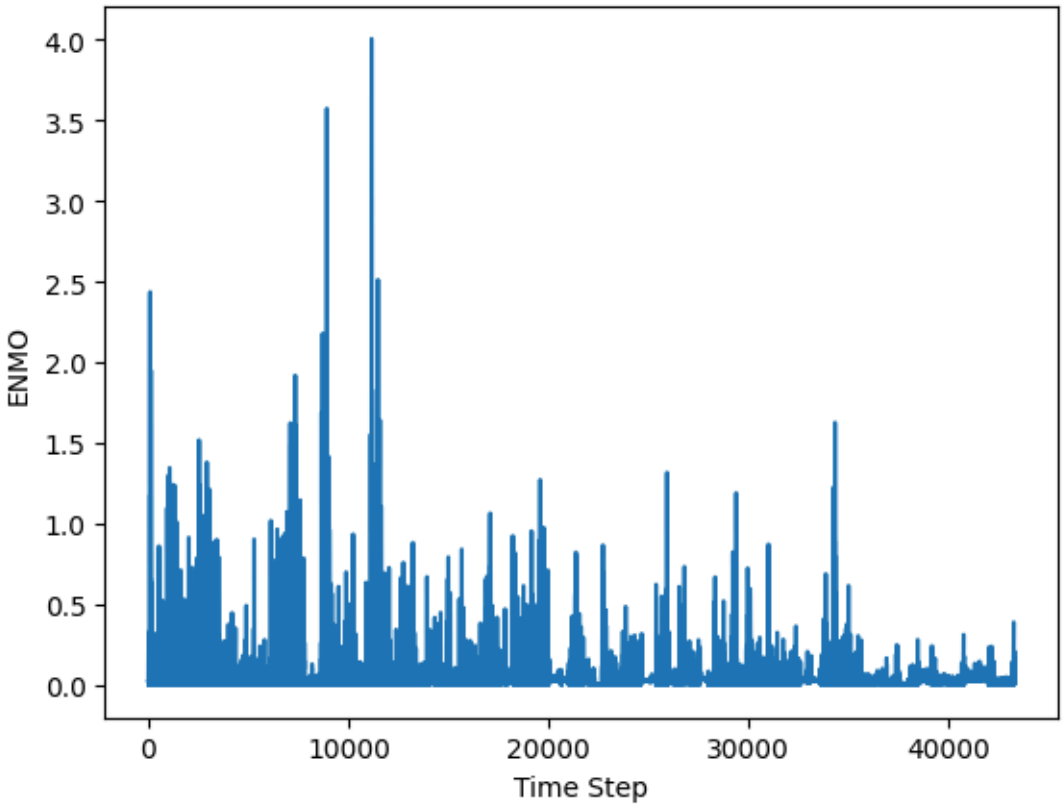

Figure 5: A single patient's ENMO plotted across the entire duration of their

Figure 6 depicts the distribution of ENMO means calculated for each patient, a rough indicator of the average activity level that the patient had while the fitness watch was worn.

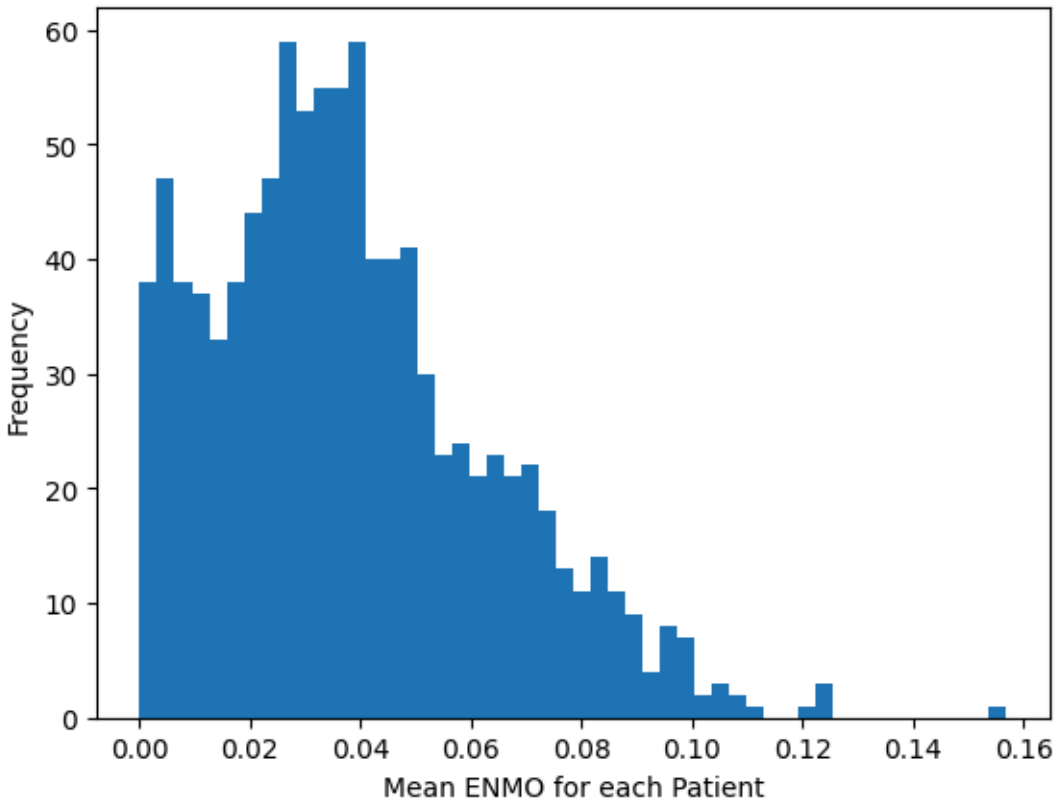

Figure 6: The distribution of ENMO means (calculated for each patient)

Internet Use

Participants of the CMI's examinations were asked how frequently they currently use the internet. The responses consisted of a numerical value from zero to three. zero meant an internet use of less than an hour per day, one meant use around one hour per day, two meant use around two hours per day, and three meant use over three hours per day.

Figure 7 displays the response frequency of each value from all the participants. Roughly 46% of participants claimed to use the internet for less than one hour per day. There is an interesting spike in participants using the internet for two hours per day, over double those of just one hour. This could be from participants confused on the difference between "Less than 1h/day" and "Around 1h/day."

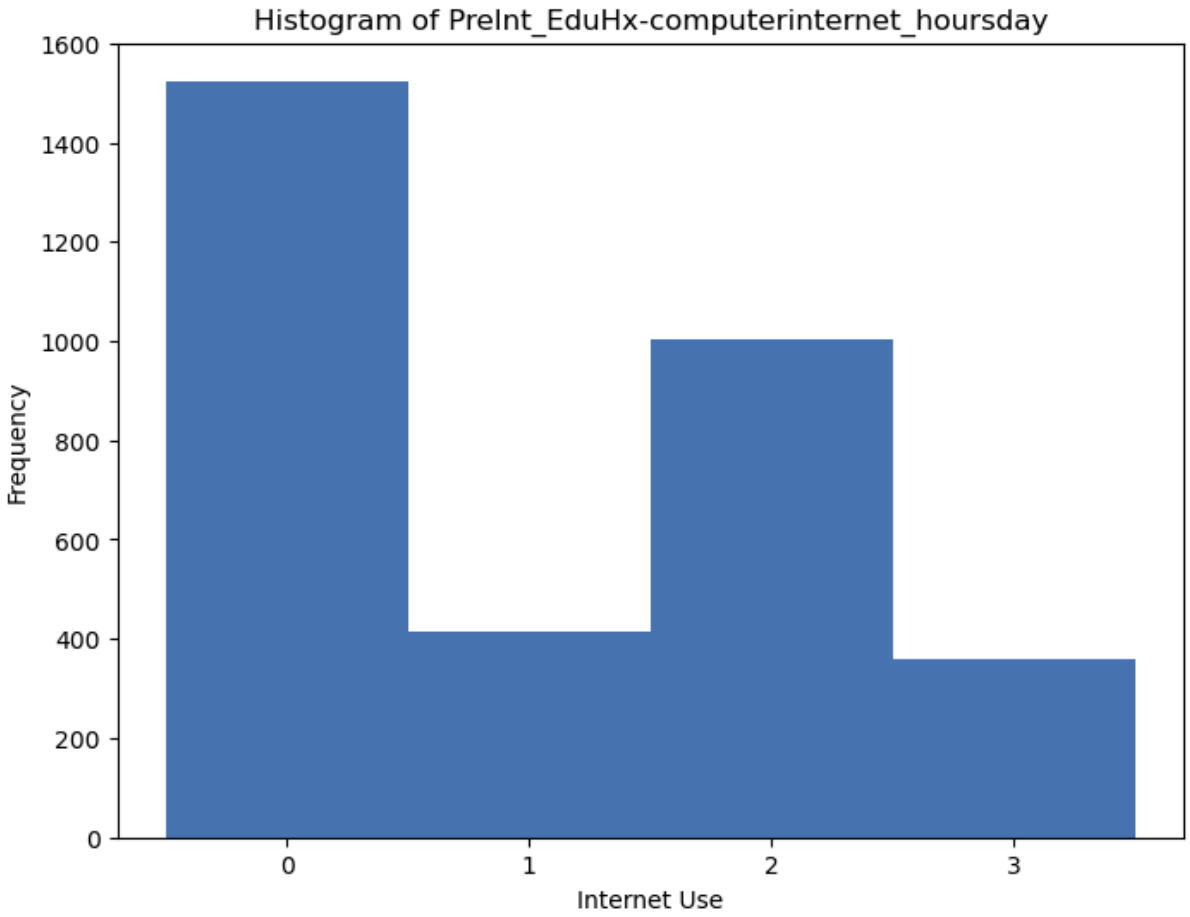

Figure 7: The distribution of Internet Use scores

FitnessGram

Many participants took part in numerous FitnessGram tests. FitnessGram measures a variety of strength, flexibility, and endurance benchmarks to ensure children are in good physical health.

FitnessGram consisted of "FitnessGram Vitals" and "FitnessGram Child" tests. FitnessGram Vitals tested the muscular endurance of participants with a treadmill run. Not many of the Child Mind Institute's participants performed this test so it has been dropped from our analysis. FitnessGram Child tested for strength and flexibility with their tests in the curl up, trunk lift, grip strength, push up, and sit & reach. The distribution of performance on each test is shown in Figure 8

Based on these scores, participants were given either a zero (Needs Improvement) or a one (Healthy Fitness Zone).

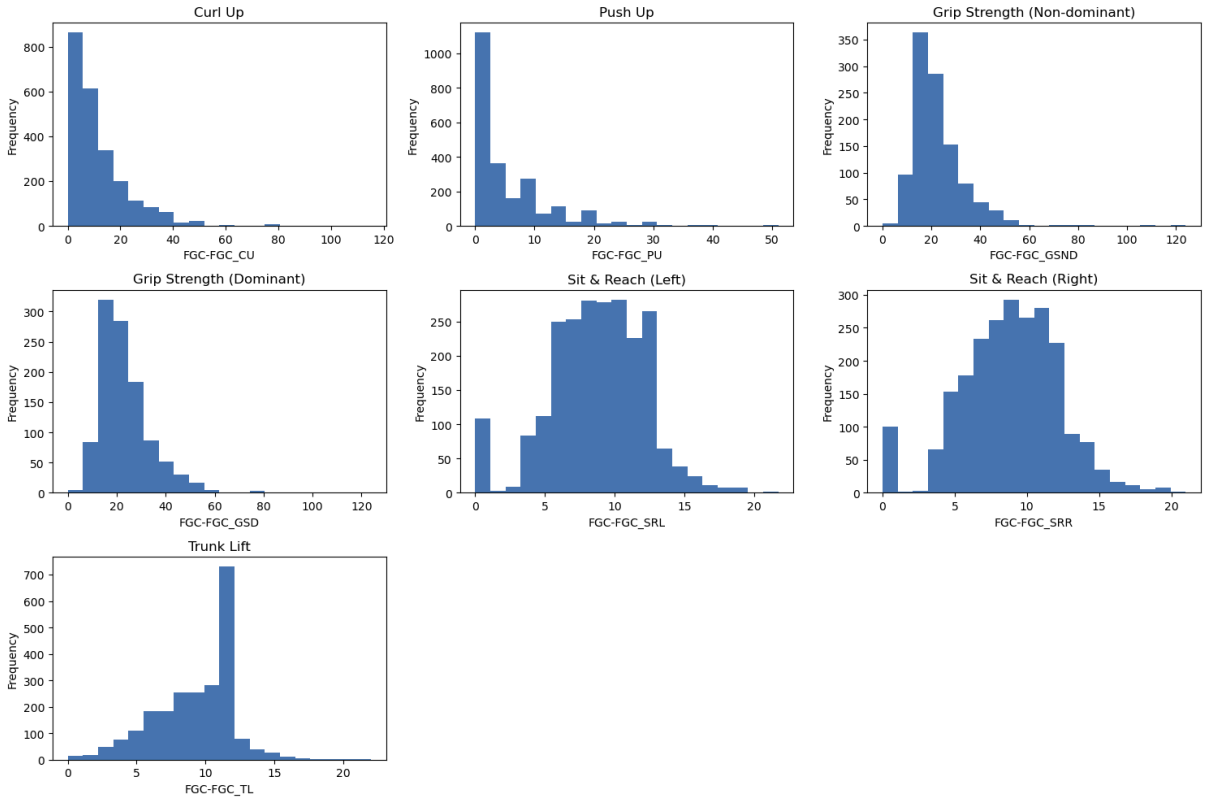

Figure 8: The distribution of different tests provided by FitnessGram

Physical Activity Questionnaire

The Physical Activity Questionnaire (PAQ) was administered to the Child Mind Institute's participants. This questionnaire asked the participants about their recent activity levels over the past seven days and returned a score from zero to five. The PAQ was split into two categories based on age named "Child" and "Adolescent" The distribution of results from this questionnaire are shown in Figure 9. The season of participation in the questionnaire was also recorded for both child and adolescent segments.

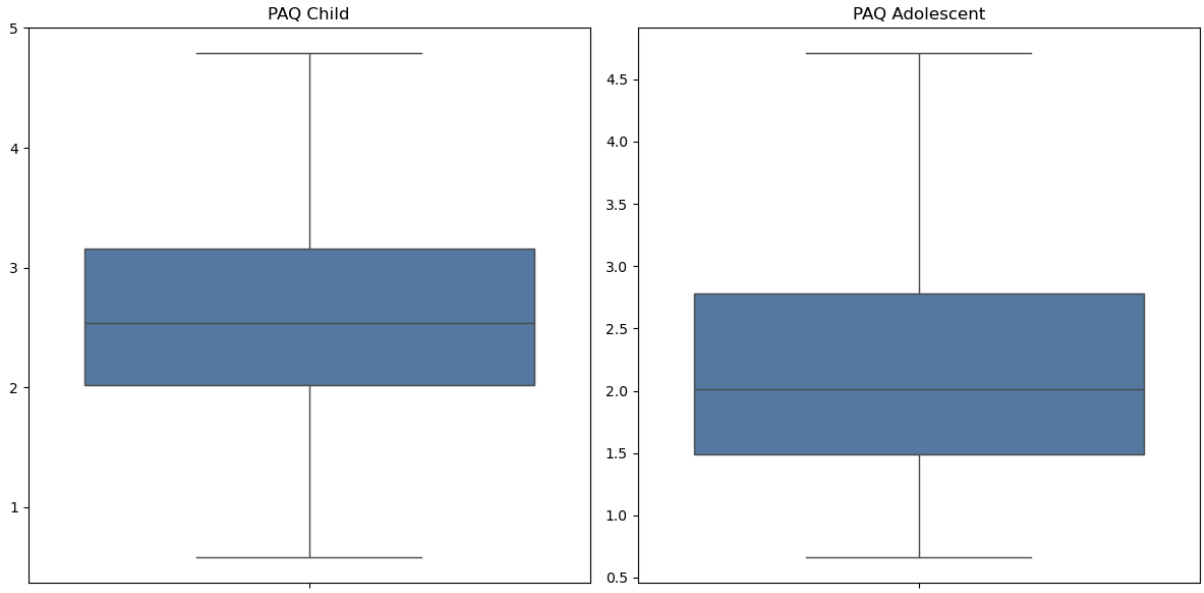

Figure 9: Boxplot distributions for the Physical Activity Questionnaire Child and Adolescent

The adolescent group contained roughly 88% null values and in turn were left out of the analysis.

Demographics

The demographics data for this study included data on participants' age and sex, as well as the season of enrollment in the study. The feature "Basic_Demos-Sex" had categorical int values, 0 for females and 1 for males. These 3 features in the dataset contained no null values, needing no further cleaning and preprocessing. After EDA, it was found that the dataset includes a disportionately high number of younger participants with the mean age being 10 years old. However, participants in this dataset ranged from as young as 5 years old to as old as 22 years old. This may limit the generalizability of findings to older participants. The analysis also revealed that over 60% of partipants were males. For the remaining feature, enrollment season, the analysis showed that enrollment was evenly distributed between all seasons (Spring, Summer, Fall, and Winter). The figures shown below help us visualize these results.

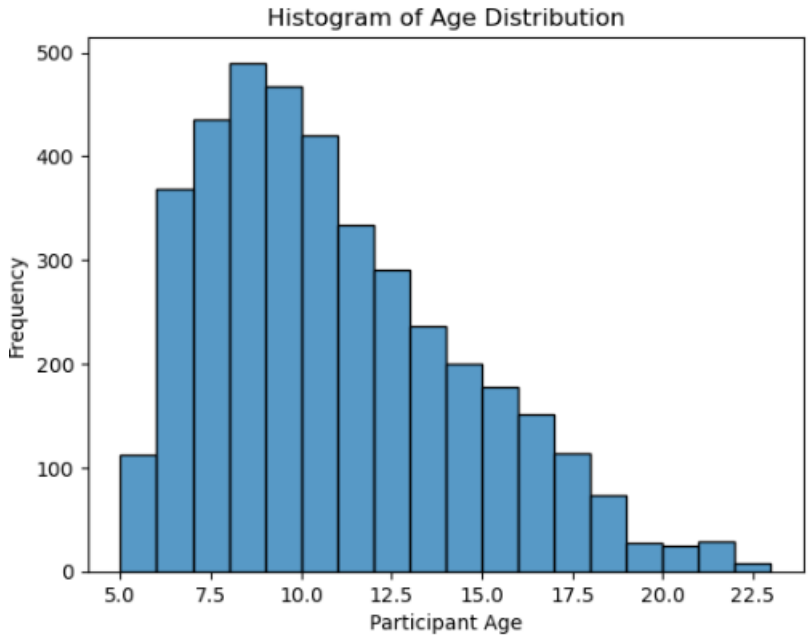

Figure 10: The distribution for Participant Age

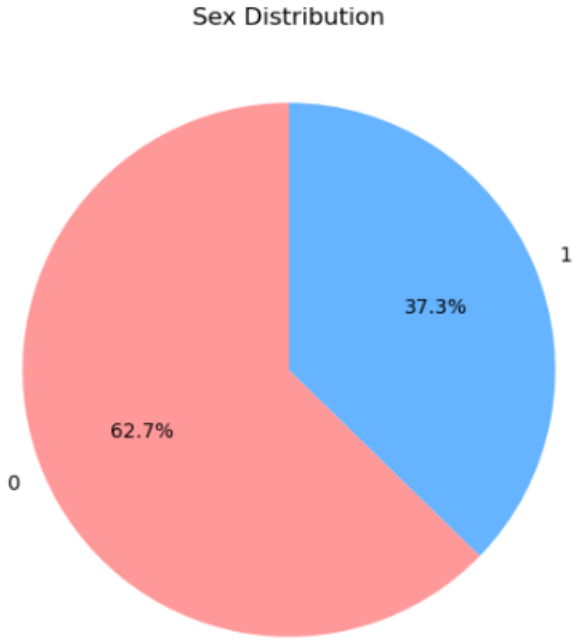

Figure 11: Pie Chart for Participant Sex Distribution

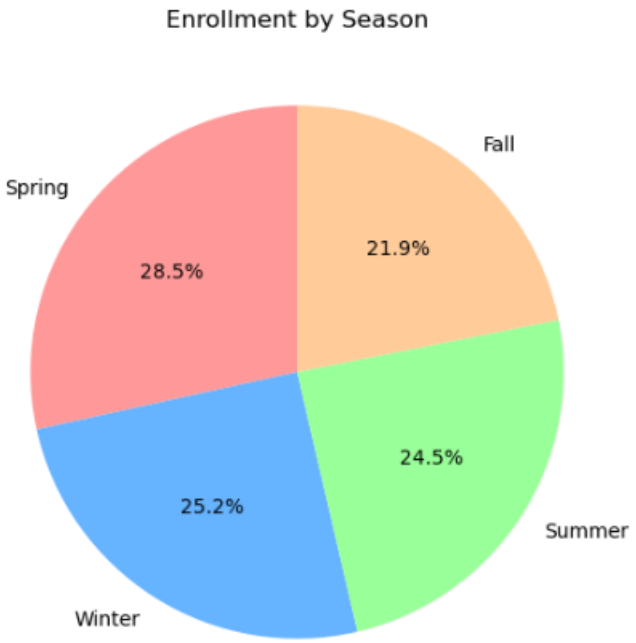

Figure 12: Pie Chart for Enrollment Season Distribution

Physical Measures

Physical measures are features of recorded data on height, weight, BMI (Body Mass Index), Blood Pressure (Diastolic and Systolic), Heart Rate, and Waist Circumference. The season in which these were recoreded is included as well. During analysis, the Waist Circumference feature had to be removed from the dataset as it contained over 77% missing values and would impact further analysis. The rest of the features contained 16% - 25%, which is low enough to keep in the dataset for analysis. After deciding to keep these features, imputation resulted in the replacement of missing values with the median of each feature, expect for season, as they are numeric and not normally distributed data. For season, missing values were imputed with the mode, which was Spring. After cleaning the date, exploratory analysis revealed that seasonality for physical measures were also evenly distributed and histrograms were created to visualize the distributions of all the numeric features.

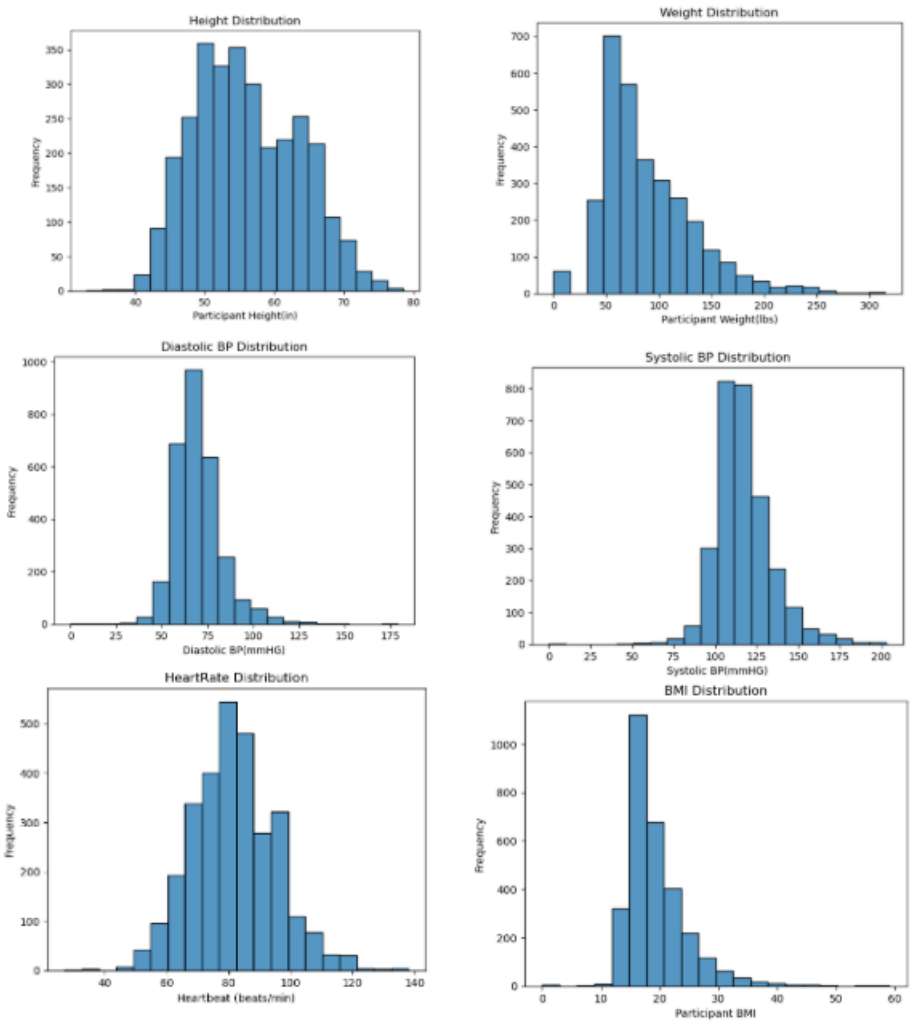

Figure 13: Distribution of Physical Measures

As seen in the figure above, BMI and Weight are right-skewed, further supporting that there is a high number of younger participants in the data. Surprisingly, Height appears to be normal.

Sleep Disturbance Scale

The Sleep Disturbance Scale captures data sleep disorders in children, using a scale to categorize them. There are a total of two features under this scale, the Total Raw Score and Total T SCore. The season of when this data was captured is also included and is also evenly distributed, which is consistent with other seaosonality features in the dataset. Histograms were created to show the Raw score and the Total score.

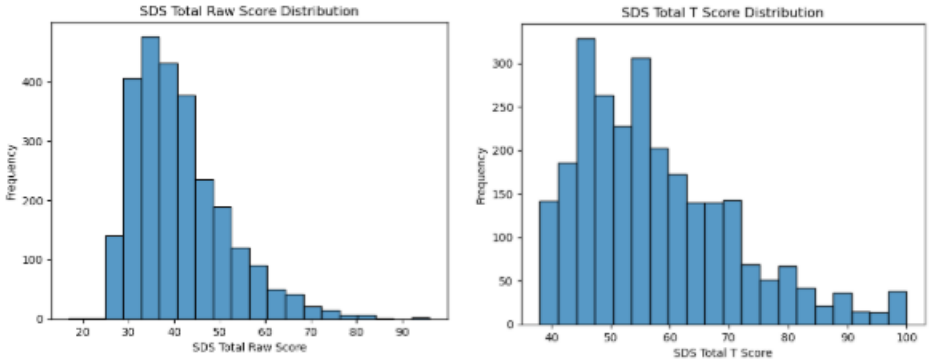

Figure 14: Distribution of Sleep Disturbance Scale Raw and Total Scores

Methods and Results

Since question 1 has already been answered, so we will move on to the second.

Question 2: Activity and PCIAT Score Divergence

Another way to state with question is: do PCIAT scores correlate with the Actigraphy data, and does that correlation deteriorate the further the Actigraphy data collection time is from when the PCIAT was administered?

To answer this, we found the correlation between patients' PCIAT Total Scores and patients' ENMO means for each day the ENMO was recorded. Then, we plotted these correlations against the corresponding distance from the PCIAT administration date. Are results are shown in Figure 15. However, we are not sure how to interpret these results, and we believe that there may be errors in our calculations.

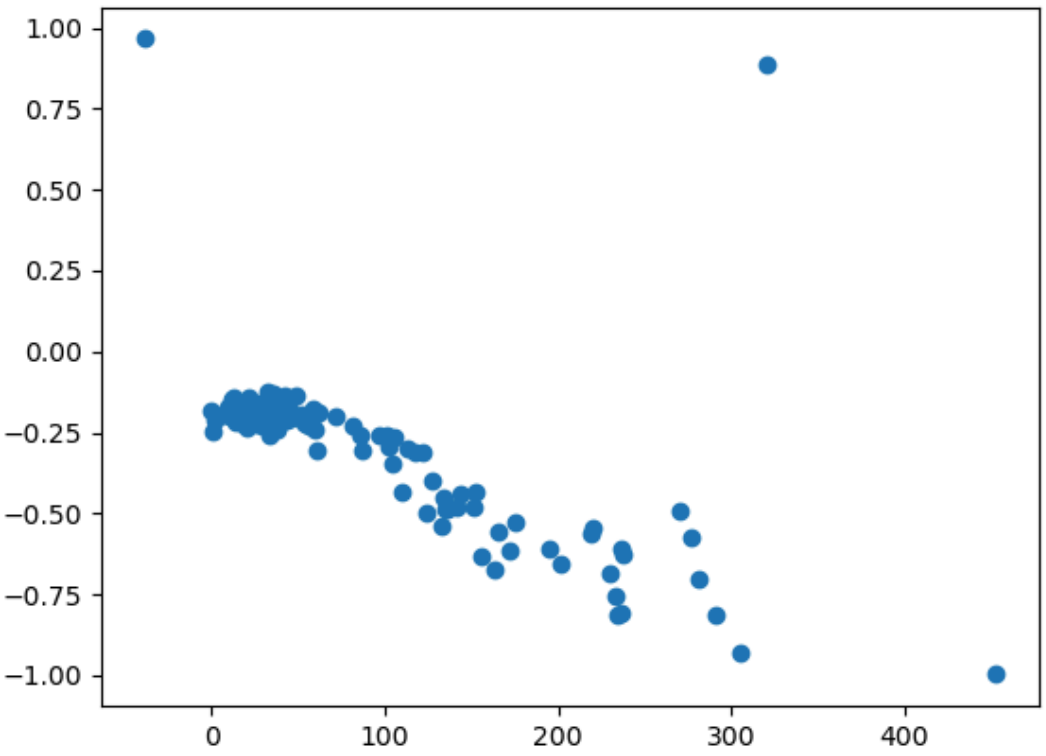

Figure 15: Correlation between PCIAT and ENMO plotted again distance from the PCIAT administration date (in days)

Question 3: PCIAT / Age Correlation

Question 4: Modeling SII

Finally, we investigated whether it possible to successfully model SII scores using simpler machine learning models.

We started by removing the data meant for unsupervised learning since it did not contain SII ground truth values. Following that, we broke the remainder into three datasets:

Training: 64% **Validation:** 16% **Testing:** 20%

Then, we chose a unified preprocessing pipeline for the CMI's dataset:

- One-hot encoding for all categorical features
- The robust MICE Forest imputer [3], [4]
- The Standard Scaler module from Scikit-learn [5], which replaces each value with its z-score.

We chose to train two models: a Random Forest Classifier and Logistic Regression classifier.

To give our models as much feedback as possible, we trained them to predict the output of each PCIAT question individually. The SII would be derived from those predictions. To evaluate these models, we chose two metrics:

- Raw proportional accuracy
- The quadratic weighted version of Cohen's Kappa Score [6], as used by the Kaggle competition conencted with this dataset. The score ranges from -1 to 1; a score of -1 corresponds with complete disagreement, 0 with random guessing, and 1 with complete agreement.

Results

After training the Random Forest and Logistic Regression models on the training set, we evaluated them on the validation set, producing the scores listed in Table 1.

	Raw Accuracy	Cohen's Kappa
Random Forest Classifier	**0.619**	0.186
Logistic Regression	0.598	**0.359**

Table 1: Validation scores for our two models

Because Cohen's Kappa is the score used by the Kaggle competition, we consider it the more valuable metric and therefore chose the Logistic Regression model as our final model.

Finally, we evaluated our Logistic Regression model on the test dataset, producing the scores listed in Table 2.

	Raw Accuracy	Cohen's Kappa
Logistic Regression	0.624	0.397

Table 2: Test scores for our Logistic Regression model

Conclusion

Project Code

All code for this project (including the Typst source code for this report) can be found in https://github.com/jsimonrichard/ds2-problematic-internet-use.

Bibliography

[1] A. Santorelli *et al.*, "Child Mind Institute — Problematic Internet Use." 2024.

[2] "Children's Global Assessment Scale (CGAS)." Accessed: Dec. 07, 2024. [Online]. Available: https://www.corc.uk.net/outcome-experience-measures/childrens-global-assessment-scale-cgas/

[3] D. J. Stekhoven and P. Bühlmann, "MissForest–non-parametric Missing Value Imputation for Mixed-Type Data," *Bioinformatics (Oxford, England)*, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/bioinformatics/btr597.

[4] "Miceforest: Multiple Imputation by Chained Equations with LightGBM." Accessed: Dec. 07, 2024. [Online]. Available: https://github.com/AnotherSamWilson/miceforest

[5] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] "A Coefficient of Agreement for Nominal Scales - Jacob Cohen, 1960." Accessed: Dec. 07, 2024. [Online]. Available: https://journals.sagepub.com/doi/10.1177/001316446002000104