Child Mind Institute — Problematic Internet Use
WorkShop 1

Team Members:

Jan Henrik Sánchez Jerez – 20231020130

Juan David Quiroga - 20222020206

Sebastián Villarreal Castro - 20221020059

System Analysis

Presented to:

Carlos Andrés Sierra Virgüez

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

Faculty of Engineering

Systems Engineering

BOGOTÁ D.C

2025

COMPETITION OVERVIEW

Summary: Develop a machine learning model that predicts the Severity Impairment Index (sii) a measure of problematic internet use in children and adolescents, using their physical activity and fitness data. The goal is to enable early detection and promote healthier digital habits.

Tabular Data (train.csv, test.csv)

Demographics: Age, sex

Fitness & Health: Heart rate, BMI, treadmill test

Behavioral Surveys: PCIAT (internet addiction test), sleep, physical activity

Time-Series Data (series_train.parquet, etc.)

Accelerometer readings: X, Y, Z, ENMO, light, non-wear flag

Temporal context: Timestamps, weekday, quarter, days since PCIAT

Supporting Files

data_dictionary.csv: Description of fields and instruments

sample_submission.csv: Required submission format (id,sii)

Significant Constraints

Missing Data: Many participants have incomplete measurements

Derived Labels: sii is calculated from subjective PCIAT responses

Data Integration: Requires aligning tabular and time-series formats

Evaluation Metric: Quadratic Weighted Kappa—sensitive to label ordering

Population Bias: Dataset is based on a specific clinical population (NYC)

SYSTEMIC ANALYSIS

ELEMENTS

Participants:

The participants are children and adolescents aged 5 to 22. This age range introduces significant variability in physical development, behavior, and digital habits, which increases the modeling complexity.

- **Instruments:**

The instruments that we are going to use manage to relate various states both physical and mental depending on the person as well as their time of use on the internet among other things, these instruments are:

1. **Demographics:** Data to know about the sex of the participant between male and female.
2. **Internet Use:** This data saves the number of hours using internet per day.
3. **Children's Global Assessment Scale**: Represented on a numerical scale by measuring the mental health clinicians to rate the general functioning of youths
4. **Physical Measures:** A basic data collection such as blood pressure, height, weight, pulse and others.
5. **FitnessGram Vitals and Treadmill:** Cardiovascular fitness measurements assessed using the NHANES treadmill protocol.
6. **FitnessGram Child:** It assesses health-related physical fitness which measures five different parameters, namely aerobic capacity, muscular strength, muscular endurance, flexibility and body composition.
7. **Bio-electric Impedance Analysis:** Measurement of key elements of body composition as well as the elements of BMI, fat, muscle and water content are also included.
8. **Physical Activity Questionnaire:** Information on participant's participation in vigorous physical activity during the past 7 days.
9. **Sleep Disturbance Scale: A** scale to categorize sleep disorders in the participants.
10. **Actigraphy:** Objective measurement of physical activity in real-world contexts using a research-grade bio-recorder.
11. **Parent-Child Internet Addiction Test:** A 20-item scale that assesses characteristics and behaviors associated with compulsive internet use, including compulsivity, escapism and dependence.

In relation to the types of relationships, we can see that since it is a physical and emotional state, the state is easily alterable, which shows the type of relationships between them.

Direct Causality:

- Internet use → Parent-Child Internet Addiction Test.
- Internet use → Sleep disturbance scale.
- Children's Global Assessment Scale → Internet Use.
- Physical Activity Questionnaire → FitnessGram Vitals and Treadmill.

Correlation:

- Demographics ↔ Internet Use, Physical Measures.
- Internet use ↔ Sleep disturbance scale.

- Physical measures ↔ FitnessGram Child, Bioelectrical impedance analysis.
- Bioelectrical Impedance Analysis ↔ Sleep Disturbance Scale.

Functional Dependence:

- Children's Global Assessment Scale relies on data from other instruments (e.g., mental health and physical activity).
- Parent-Child Internet Addiction Test relies on Demographics for segmentation.

Hierarchy:

- FitnessGram Child is a detailed subset of Physical Measures.

Feedback:

- FitnessGram Vitals results ↔ adjustments to Physical Activity Questionnaire.

Complementarity:

- Actigraphy ↔ Physical Activity Questionnaire (objective vs. subjective data).
- FitnessGram Child ↔ Bioelectrical Impedance Analysis (fitness and body composition parameters).

Restriction:

- Demographics limits gender/age interpretations in analysis.
- Actigraphy is restricted by the availability of devices.

Synergy:

- Combination of FitnessGram Child and Bio-electric Impedance Analysis generates a comprehensive fitness profile.

Target Variable (sii)

The target to predict is the Severity Impairment Index (SII), a 0–3 scale representing the level of internet addiction. It is derived from the Parent-Child Internet Addiction Test (PCIAT), which measures compulsive digital behaviors.

Feedback Mechanism

Feedback is delivered through the Kaggle leaderboard, where models are scored using the Quadratic Weighted Kappa metric. This measures how closely predictions match the true SII levels, with penalties based on prediction error severity.

Modeling Pipeline

We follow a modeling pipeline that transforms raw physical and behavioral data into accurate severity predictions. The following steps outline how we can process, analyze, and model the data:

1. Data Acquisition: Collect the main datasets: tabular data (train.csv, test.csv), actigraphy time-series (series_train.parquet), and supporting files like the data dictionary and sample submission.
2. Data Preprocessing: Clean and prepare the data by handling missing values, scaling numerical features, encoding categories, and aligning different data sources by id.
3. Feature Engineering: Extract meaningful features from the raw data. For actigraphy, calculate statistics like mean ENMO or activity variability. From tabular data, create derived metrics like BMI or sleep irregularity.
4. Exploratory Data Analysis (EDA): Explore distributions, detect imbalances, examine correlations, and visualize patterns in both physical and behavioral variables to guide modeling choices.
5. Model Selection: Choose suitable machine learning algorithms based on the data type and project goals. Common choices include XGBoost, LightGBM, and neural networks, especially for time-series data.
6. Training & Validation: Train your model using cross-validation (e.g., stratified k-fold) to ensure robust performance. Adjust for class imbalance and test different hyperparameters.
7. Evaluation: Use the Quadratic Weighted Kappa to assess how well the model predicts ord:ered severity levels. Analyze error patterns to refine your approach.
8. Submission: Generate predictions for the test set, format them in id,sii structure, and submit through Kaggle. Use the public leaderboard to assess your model's effectiveness.

RELATIONSHIPS

- Participant data is collected and labeled based on PCIAT scores → transformed into the sii variable.
- Physical behavior and health features are used as proxies for psychological traits.
- ML models transform raw and derived features into predicted sii values.
- The leaderboard ranks solutions, feeding back into model optimization.

COMPLEXITY AND SENSITIVITY

Missing Data

Many features have **incomplete records** across participants.

Especially common in fitness assessments and actigraphy.

Multi-Modal Data Fusion

Requires integrating **tabular** (structured) and **time-series** (actigraphy) data.

Different formats and resolutions increase modeling complexity.

Label Ambiguity

The target sii is **derived indirectly** from PCIAT responses.

Subjective test introduces noise or inconsistency in the labels.

Ordinal Target Sensitivity

The output variable (sii: 0–3) is **ordinal**.

Simple classifiers may treat it as categorical, ignoring label order.

Quadratic Weighted Kappa penalizes misclassifications based on *distance* from the true label.

Age and Developmental Variability

The dataset includes children and adolescents (ages 5–22).

Large physiological and behavioral differences across age groups.

Sensor Noise and Artifacts

Actigraphy may contain:

Periods where the device was not worn (non-wear_flag)

External disturbances (e.g., shaking, low light)

Temporal Irregularities

Length of actigraphy recordings varies per participant.

Not all recordings are synchronized or aligned by calendar or event (e.g., date of PCIAT).

Evaluation Metric Sensitivity

Quadratic Weighted Kappa is **strict** with wrong-order predictions.

Small model adjustments can cause large shifts in score.

Class Imbalance

Fewer examples may exist in high severity (sii = 3).

This affects model learning and may bias predictions toward common classes.

Feature Correlation and Confounding

Many physical features (e.g., BMI, sleep) are correlated.

Confounding variables (e.g., socioeconomic status) are not present in the data but may influence both behavior and SII.

Preprocessing the data

As we've mentioned before, there are quite a few missing values (NaNs) in the dataset, and we need a strategy to deal with them properly. In this case, the best approach isn't to remove rows with missing data—dropping them could result in losing a significant amount of useful information, which isn't ideal. Instead, we should investigate ways of **filling in** those missing values.

One common and effective method is to fill them using the **median** of the column. We could use the **mean**, but there's a problem with that: if the data contains outliers (like one value that's way too high or low), the mean can get pulled in that direction and end up misrepresenting the rest of the data. That would cause the filled-in values to be inconsistent with the general trend. The median, on the other hand, is more robust to outliers and gives a more balanced central value, especially in skewed distributions.

Another option is to use **interpolation**, which works particularly well with time-based or ordered data. Interpolation basically estimates missing values by looking at the values before and after them and filling in something that makes sense in between. This can be helpful when working with data that changes gradually over time—like sensor readings or movement data from wearable devices.

So overall, our plan for dealing with missing values would be to avoid dropping any data if possible and instead use smarter techniques like filling with the median or applying interpolation when appropriate.

Multiclass Classification: The "Normal" Approach

In this problem, we're dealing with a multiclass classification task—more specifically, a "standard" or "vanilla" version of it. This is a type of supervised learning where the model learns to assign one single label to each input, picking from a fixed set of classes. In our case, those classes are the integers 0 through 4, which represent the severity levels of problematic internet use.

Now, while those numbers might look ordered to us (like 0 = low severity, 4 = high severity), the model doesn't know that. In normal multiclass classification, the model treats each class as completely independent. So it's not really aware that predicting a 4 instead of a 3 is a "smaller mistake" than predicting a 0 instead of a 4. It sees them kind of like arbitrary labels—like red, green, blue, etc.—with no sense of order between them.

So, how does this works?

1. During training:
Each sample in the training set has features (X) and a target label (y), where y is one of {0, 1, 2, 3, 4}. The model looks at the patterns in the features and tries to learn how likely each input is to belong to each of the five classes. This is often done using softmax output layers (in neural nets, for example), or probability estimators in other types of classifiers.

2. During prediction:
When the model sees a new input, it gives a **probability score for each class**. Let's say it predicts something like:

- P(0) = 0.1
- P(1) = 0.2
- P(2) = 0.3
- P(3) = 0.3
- P(4) = 0.1

Then it picks the class with the highest probability—in this case, it might pick class 2 or class 3. If there's a tie, some models will just go with the first one that reaches that maximum.

This method works well in general, but it ignores the ordinal nature of the labels. That's why metrics like Quadratic Weighted Kappa (which we discussed earlier) are used to evaluate the predictions more fairly—it takes into account how far off a prediction is, not just whether it's wrong.

So, while the model is doing a standard multiclass classification, our evaluation method (QWK) helps add some structure and fairness by considering the fact that these labels actually do have a meaningful order.

CHAOS AND RANDOMNESS

- **Feedback:**

In this competition, the way we get feedback on how well our model is doing is through a specific evaluation metric called Quadratic Weighted Kappa (QWK). This metric doesn't just check if the prediction is correct or not—it also penalizes the model more when it's *way off* from the actual value. For example, predicting a 2 when the true label is 3 isn't too bad, but predicting a 0 when it should've been a 4 is a bigger problem.

So, how does this method actually work?

To compute the QWK score, we use three key matrices: O, W, and E.

The O matrix (Observed matrix) is like a confusion matrix with size N x N, where *N* is the number of possible labels. Each element O(i, j) represents the number of times the model predicted label *j* when the true label was *i*. Ideally, most of the values should be on the diagonal, since those are the correct predictions. Any value off the diagonal means a prediction was wrong.

Next is the W matrix (Weight matrix), which is used to assign penalties to prediction errors. This matrix is also N x N, and it punishes predictions more the further they are from the correct value. The formula to calculate each element is:

$W(i, j) = ((i - j)^2) / ((N - 1)^2)$

This squared difference increases the penalty for larger errors. The denominator normalizes the result so all the values fall between 0 and 1. We square the difference to make the system more sensitive to big errors, but not *too* sensitive raising it to a higher power would make the score drop too much even with just a few mistakes, which isn't ideal.

The third one is the E matrix (Expected matrix). This matrix simulates what would happen if predictions were made completely at random, while still keeping the same distribution of real and predicted labels. It's built using the outer product between the histogram of the actual labels and the histogram of the predicted labels. Then it's normalized so that the total number of elements matches that of the O matrix. This makes sure both matrices are comparable and operate under the same scale.

Now that we have these three matrices, we can finally calculate the QWK score using this formula:

$K = 1 - (\text{sum of } W(i, j) * O(i, j)) / (\text{sum of } W(i, j) * E(i, j))$

This formula gives us a score that tells how much better (or worse) our model is compared to random guessing, taking into account how serious each prediction error was.

The value of Kappa can go from $-\infty$ to 1:

- If Kappa = 1, the model made perfect predictions.
- If Kappa = 0, it performed no better than random.
- If Kappa < 0, it did worse than random guessing.
- And anything between 0 and 1 means it's doing better than chance, and the closer it is to 1, the better.

So, in short, QWK is a solid metric for this kind of problem because it considers not just if a prediction is correct, but also *how wrong* it was when it's not. That helps build models that are not only accurate, but cautious and thoughtful in their predictions.

CONCLUSION

In conclusion, this project highlights the complexity and interdisciplinary nature of detecting and analyzing problematic internet use in children and adolescents. By integrating physical, behavioral, and psychological data, we aim to develop machine learning models capable of predicting the Severity Impairment Index (SII) with accuracy and sensitivity.

Throughout the analysis, we encountered key challenges such as missing data, multi-modal integration of tabular and time-series sources, and the subjectivity inherent in label generation through psychological scales like PCIAT. Furthermore, the diverse age range of participants and the ordinal nature of the target variable added layers of complexity to both data preprocessing and model evaluation.

Despite these obstacles, the use of robust preprocessing strategies, thoughtful feature engineering, and the implementation of evaluation metrics like Quadratic Weighted Kappa enables us to construct models that go beyond simple classification. These models can meaningfully reflect the severity of internet addiction and its underlying behavioral patterns, fostering opportunities for early intervention.

Ultimately, the insights gained from this workshop not only contribute to our technical understanding of data science in healthcare contexts but also reinforce the importance of addressing digital well-being in today's youth through informed, data-driven approaches.

References

https://www.kaggle.com/code/reighns/understanding-the-quadratic-weighted-kappa
https://www.healthyplace.com/psychological-tests/parent-child-internet-addiction-test