KAGGLE- PROBLEMATIC INTERNET USE

Teammates: **Uku Konsap, Hjalmar Vaiküll, Riika Seeba**

Group: **B1**

Our github repository:

https://github.com/riikaseeba/Problematic-Internet-Use

Competition link: https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use

Business understanding

Identifying our business goals

Background:

In today's digital age, problematic internet use among children and adolescents is a growing concern. Better understanding of this issue is crucial for addressing mental health problems such as depression and anxiety.

Physical & fitness measures are extremely accessible and widely available with minimal intervention or clinical expertise. Changes in physical habits, such as poorer posture, irregular diet, and reduced physical activity, are common in excessive technology users. We propose using these easily obtainable physical fitness indicators as proxies for identifying problematic internet use, especially in contexts lacking clinical expertise or suitable assessment tools. Identifying your business goals

One **business goal** is contributing to a healthier, happier future where children are better equipped to navigate the digital landscape responsibly.

The **business success criteria** would be that when a predictive model is developed that analyses children's physical activity data to detect early indicators of problematic internet and technology use.

Assessing our situation

Inventory of resources:

- We use Jupyter Notebook for the model implementation, including explanations.
- The project itself is Kaggle-competition.
  - https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use

    - From there we will find the necessary background information, guidelines for achieving the ultimate goal of the project, and a communication portal for additional questions.

Requirements, assumptions, and constraints:

Two elements of The Healthy Brain Network (HBN) dataset are being used for this competition: physical activity data (wrist-worn accelerometer data, fitness assessments, and questionnaires) and internet usage behaviour data.

- During participation in the HBN study, some participants were given an accelerometer to wear for up to 30 days continually while at home and going about their regular daily lives.

The competition data is compiled into two sources, parquet files containing the accelerometer (actigraphy) series and csv files containing the remaining tabular data. The majority of measures are missing for most participants. In particular, the target sii is missing for a portion of the participants in the training set.

- Submissions to this competition must be made through Notebooks.
- Deadline – in the kaggle-competition the deadline is on 19th December but we will accomplish it by 11th December.

Risks and contingencies:

We are allowed to submit a maximum of five submissions per day. Each team member's submission counts. Since the model testing may not pass the first few submissions, it is necessary to start early and, if possible, submit at least a few days before the deadline to understand and resolve any issues.

Terminology:

- **Business terms:** (instruments where the tabular data in train.csv and test.csv measurements come from)
    - Demographics - Information about the age and sex of participants.
    - Internet Use - Number of hours of using computer/internet per day.
    - Children's Global Assessment Scale - A numeric scale used by mental health clinicians to rate the general functioning of youths under the age of 18.
    - Physical Measures - Collection of blood pressure, heart rate, height, weight and waist, and hip measurements.
    - FitnessGram Vitals and Treadmill - Measurements of cardiovascular fitness assessed using the NHANES treadmill protocol.
    - FitnessGram Child - Health-related physical fitness assessment measuring five different parameters including aerobic capacity, muscular strength, muscular endurance, flexibility, and body composition.
    - Bio-electric Impedance Analysis - Measure of key body composition elements, including BMI, fat, muscle, and water content.

  - Physical Activity Questionnaire - Information about children's participation in vigorous activities over the last 7 days.
  - Sleep Disturbance Scale - Scale to categorize sleep disorders in children.
  - Actigraphy - Objective measure of ecological physical activity through a research-grade biotracker.
  - Parent-Child Internet Addiction Test - 20-item scale that measures characteristics and behaviors associated with compulsive use of the Internet including compulsivity, escapism, and dependency.
- **Data-mining terms:**
  - Random forest - A machine learning algorithm that creates a "forest" of decision trees during training and combines their outputs to improve prediction accuracy and reduce overfitting.
  - XGBoost - An advanced gradient boosting algorithm known for its speed and performance in predictive modelling tasks, particularly with structured/tabular data.
  - CatBoost - A gradient-boosting algorithm that handles categorical data efficiently without requiring extensive pre-processing.
  - Cross validation - A model evaluation technique that splits data into multiple subsets to train and test models iteratively, ensuring the model generalizes well to unseen data.
  - Root-mean-square error (RMSE) - A metric used to measure the average difference between predicted and observed values, emphasizing larger errors.
  - Quadratic Weighted Kappa (QWK) - A metric used for evaluating the agreement between predicted and actual ordinal ratings, with higher scores indicating better model performance.

Benefits:

There is an opportunity to win the competition:

- 1st Place - $ 15,000
- 2nd Place - $ 10,000
- 3rd Place - $ 8,000
- 4th Place - $ 7,000
- 5th Place - $ 5,000
- 6th Place - $ 5,000
- 7th Place - $ 5,000
- 8th Place - $ 5,000

Costs:

- 90 hours of work
- Poster printing – about 1 or 2 euros

Defining our data-mining goals

Data-mining goal:

The data-mining goal of this project is to predict a participant's Severity Impairment Index (SII) from the given data, providing a standard measure of problematic internet use.

Steps to achieve the goal:

1) Process the dataset so it is suitable for building and validating models.
2) Develop predictive models capable of accurately estimating SII.
3) Deliver a report outlining model performance and its alignment with the competition goal.

Data-mining success criteria:

The model's Quadratic Weighted Kappa score must be greater than 0, demonstrating performance better than random guessing. The model's performance should be as close as possible to the top leaderboard score (0.49) or higher.

Data understanding

Gathering data

Outline data requirements

The project aims to predict the Severity Impairment Index (SII) based on participant data. To achieve this, these following types of data may be necessary:

- Demographic data: age, sex;
- Clinical data: Children's Global Assessment Scale (CGAS) score and season.
- Physical Activity and Measurements: BMI, height, weight, wrist-worn accelerometer data.
- Target Variable: Severity Impairment Index (SII)
- Time Range: Data should cover the participation duration

Verify data availability

The dataset is available in the kaggle-competition environment as follows:

- Tabular Data: train.csv and test.csv include features like demographics, CGAS scores, and physical measures.
- Data Dictionary: The data_dictionary.csv provides detailed descriptions, value ranges, and labels for all features.
- Accelerometer Data: Parquet files contain raw activity data for participants.
- Target Variable: The sample_submission file specifies the structure of the predicted sii scores.

Note: Many fields are incomplete; e.g., the target variable (SII) is missing for some participants in the training set. There is missing demographic or physical measures for some participants.

Define selection criteria

- train.csv: Includes labelled data for training (demographics, CGAS, physical measures, and some target values).
- test.csv: Unlabelled data to be used for predictions and submissions.
- data_dictionary.csv: Serves as a reference for feature definitions and valid ranges.
- parquet files: Provide raw time-series accelerometer data.

Fields used in train.csv and test.csv: Basic_Demos-Enroll_Season, Basic_Demos-Age, Basic_Demos-Sex, CGAS-Season, CGAS-CGAS_Score, Physical-BMI, Physical-Height, Physical-Weight.

Case selection: include participants with valid SII scores for training model.

Describing data

Source: The dataset originates from the Kaggle competition *Child Mind Institute - Problematic Internet Use*.

- Files available: train.csv, test.csv, sample_submission.csv, data_dictionary.csv, parquet files (accelerometer data)

Number of cases: train.csv contains 3960 rows and 82 columns and test.csv contains 20 rows and 59 columns.

Description of fields:

- Basic_Demos-Age: Age of participants (integer).
- Basic_Demos-Sex: Gender of participants (binary: 0 for male, 1 for female).
- Basic_Demos-Enroll_Season: Season when participants enrolled (categorical: Spring, Summer, Fall, Winter).
- CGAS-CGAS_Score: Children's Global Assessment Scale (integer; higher scores indicate better functioning).
- Physical-BMI: Body Mass Index (float).
- Physical-Height: Height in centimetres (float).
- Physical-Weight: Weight in kilograms (float).
- PCIAT fields: Categorical answers to 20 questions
- sii (Severity Impairment Index): Ordinal variable (0, 1, 2, 3)

Suitability: Missing values are present in some features. Distribution and range of certain features need examination to confirm usability.

Exploring data

The data has severe quality issues. About 40% of the data is missing the target variable and some of the existing data is questionable. Thus the data needs to be thoroughly analysed and cleaned. The sii values are occasionally calculated inaccurately because the missing data has a value of 0 from the range of 0 to 5. This results in the sii value to be recalculated in some other way. For example, setting the missing questionnaire data to 5 showcasing the maximal sii value instead of minimal would reduce the risk of under evaluating the subjects. Additionally, physical values like height or weight need to be analysed and cleaned so that they are within acceptable range. 2/3 of the participants are over- or underweight according to the BMI and some blood pressure readings are suspicious and do not fall in normal ranges. The accelerometer data is also at times questionable, because about half the participants had sleep_idle_mode enabled and this results in missing data for some time periods. The battery voltage also seems to affect the readings so we might possibly set a threshold for that.

Verifying data quality

Although the quality of data is low, we can still achieve our goals to some extent. We have enough information to predict with close to 50% accuracy which in the case of this project is acceptable. The main issue is that it is not specified how the data was collected and thus we can not comfortably state the correctness of the results. Furthermore it is not stated in what units the data is collected in, leaving us to predict it during the analysis. The data requires some domain knowledge (BMI, blood pressure etc.) for it to be cleaned efficiently. The biggest challenge is the missing values and figuring out what to keep and what we can discard.

Planning your project

There is a clarifying table for planning our project. Numbers under the names specify how many hours each team member will contribute to each task.

Tasks	Uku Konsap	Hjalmar Vaiküll	Riika seeba
1) Understanding and analysing the data	10	10	10
2) Data cleaning	8	8	6
3) Train different models	6	6	6
4) Teamwork (communication, reports, submissions, etc.)	6	6	8
5) Poster making	3	3	3

The methods and tools that we plan to use in modelling:

Techniques

- **Random forest** - it may be used to predict SII by handling tabular data from CSV files, leveraging its ability to work well with categorical and continuous variables.
- **XGBoost** - it is a strong candidate for predicting SII, as it can handle missing data and optimize complex relationships in features extracted from actigraphy or demographic datasets.
- **CatBoost** – it can be used for models where categorical variables like demographics (age, gender) are prominent, reducing the need for one-hot encoding.

Validation

- **Cross validation** - ensures that the model predicting SII is robust and not overfitted to the training dataset, especially given the missing data challenges.
- **RMSE** – it can be used during model training to assess the accuracy of regression models predicting SII before calculating the final Quadratic Weighted Kappa score.
- **QWK** - is the official competition validation metric for comparing the model's predictions of SII with the actual labels, guiding model optimization efforts.