A Predictive Model for Problematic Internet Use Using Physical Fitness Data in Adolescents

[*]This competition is participated in by a Zenith³ team from the University of Engineering and Technology, Vietnam National University

Thi Linh Luong, Anh Tuan Mai, Thu Trang Thi Nguyen
Faculty of Information Technology
VNU University of Engineering and Technology
Hanoi, Vietnam
22028202@vnu.edu.vn, 22028144@vnu.edu.vn, 22028254@vnu.edu.vn

***Abstract*—This report details our participation in a Kaggle competition aimed at predicting the level of problematic internet usage among children and adolescents, using data on physical activity and fitness. Identifying problematic internet use in youth is inherently challenging due to the complex, high-dimensional nature of the data. In this study, we frame the task as a multi-class classification problem, drawing from a diverse set of physical activity and fitness indicators. These features often introduce noise and exhibit non-linear relationships, adding to the difficulty of accurate prediction. To address this, we employ advanced classification models and ensemble learning techniques to enhance robustness and improve predictive accuracy. Our results show that ensemble methods, which combine the strengths of multiple classifiers, consistently outperform individual models by capturing subtle patterns in the data. This approach demonstrates the potential to effectively manage noisy, multi-dimensional datasets, contributing to early intervention efforts aimed at fostering healthier digital habits among children and adolescents.**

***Index Terms*—Classification, Ensemble Learning, Multi-dimensional Data, Problematic Internet Use, Predictive Modeling.**

I. Introduction

A. Problem Statement

In this section, we address the issue of problematic internet use (PIU) among children and adolescents—an escalating public health concern. As digital device usage continues to rise, so do worries about its detrimental effects on youth, manifesting in both physical and mental health issues. Excessive internet engagement has been linked to conditions such as depression, anxiety, and social isolation, underscoring the urgent need for effective detection and early intervention strategies.

Accurately predicting and diagnosing PIU is crucial for timely intervention, yet it remains a challenging task. The primary difficulty lies in handling high-dimensional, noisy data used to assess behavioral patterns—ranging from physical activity levels to mental health indicators and digital usage behaviors. These features are not only interdependent but also exhibit complex, non-linear relationships, making it difficult to discern underlying patterns. Additionally, inconsistencies in participant behavior, such as device errors or incomplete recordings, introduce further noise, complicating the prediction process.

To overcome these challenges, our objective is to develop a predictive model capable of robustly classifying PIU in youth. We approach this as a multi-class classification problem, reflecting the diverse manifestations of PIU. By leveraging accessible and widely available data—particularly physical fitness indicators—we aim to create a scalable, cost-effective solution for early detection.

Our work is grounded in the Child Mind Institute's Problematic Internet Use dataset, which encompasses both time-series data capturing activity and behavior over time, as well as static demographic information. This rich dataset provides a comprehensive foundation for our analysis, allowing us to extract meaningful insights and construct a predictive model that is both accurate and reliable.

B. Background

Since the competition began on November 25, 2024, our team has made significant strides in developing predictive solutions. Over the course of nearly one month, we submitted 60 solutions, with 11 key versions selected for analysis in this paper. Among these, two submissions achieved private scores of 0.407 on Kaggle's leaderboard. Our best-performing model attained a private score of 0.449, demonstrating notable improvement and highlighting the effectiveness of our iterative approach to model refinement.

The remainder of this paper is organized as follows: Section II provides an analysis of the dataset, addressing missing values, feature distributions, and discrepancies between the training and testing datasets. Section III delves into modeling and optimization, covering single-model approaches, parameter tuning, and ensemble methods to enhance predictive accuracy. Section IV and V concludes with a summary of the results and an evaluation of the proposed methods' effectiveness in identifying early signs of problematic internet use through physical activity data.

II. Data Analysis

The data used for this study was provided by the **Healthy Brain Network**, a landmark mental health initiative based in New York City, aimed at supporting children worldwide. This initiative is a collaboration between families, community leaders, and the Child Mind Institute, working together to unlock the secrets of the developing brain. The project is further supported by the Kaggle team and the California Department of Health Care Services (DHCS) through the Children and Youth Behavioral Health Initiative (CYBHI).

A. Data Overview

This study utilizes a comprehensive dataset that includes both tabular and time series data, capturing multidimensional information about children and adolescents. The dataset covers a wide range of characteristics, including demographic details, physical fitness indicators, internet usage habits, and behavioral assessments. Below is a detailed breakdown of the dataset components:

1) Tabular Data: The tabular dataset includes observations for each participant, represented in the files **train.csv** and **test.csv** . These observations are gathered from various measurement tools, which are described in detail below:

- **Demographics**: Contains basic participant information, including age and gender.
- **Internet Use**: Provides data on the average daily time spent on internet or computer usage.
- **Children's Global Assessment Scale (CGAS)**: A numerical scale used by clinicians to evaluate the general functioning of adolescents under 18.
- **Physical Measures**: Includes vital signs and anthropometric data such as blood pressure, heart rate, height, weight, waist circumference, and hip circumference.
- **Fitness Assessments**:
  - *FitnessGram Vitals and Treadmill*: Measures cardiovascular fitness using standardized treadmill protocols.
  - *FitnessGram Child*: Assesses health-related fitness attributes such as endurance, muscular strength, flexibility, and body composition.
  - *Bio-electrical Impedance Analysis (BIA)*: Provides metrics for BMI, body fat percentage, muscle mass, and water content.
- **Physical Activity Questionnaire (PAQ)**: Summarizes the frequency and intensity of physical activity over the past seven days.
- **Sleep Disturbance Scale**: Categorizes potential sleep disorders in children.
- **Parent-Child Internet Addiction Test (PCIAT)**: A 20-item questionnaire assessing compulsive internet use, avoidance behaviors, and dependency. The overall PCIAT score (**PCIAT_Total**) serves as the target variable, categorized into four severity levels:
  - 0: None
  - 1: Mild
  - 2: Moderate
  - 3: Severe

2) Time-Series Data: Participants wore accelerometer devices to collect longitudinal activity data, which is stored in a series of **.parquet** files. These files contain detailed time-series measurements gathered over a 30-day period, with each file corresponding to an individual participant. The data fields include:

- **id**: A unique participant identifier, matching the ID in the tabular dataset.
- **step**: The timestep of the measurement.
- **X, Y, Z**: Acceleration values along the x-, y-, and z-axes (measured in gravitational units, g).
- **enmo**: The Euclidean Norm Minus One (ENMO) of accelerometer signals, with negative values rounded to zero, representing movement intensity.
- **anglez**: The angle of the arm relative to the horizontal plane.
- **non-wear_flag**: Indicates whether the device was worn during the measurement period (0: worn, 1: not worn).
- **light**: Environmental light intensity (measured in lux).
- **battery_voltage**: The device's battery voltage (in millivolts, mV).
- **time_of_day**: The timestamp of the measurement (in the format %H:%M:%S.%9f).
- **weekday**: Day of the week (1: Monday, 7: Sunday).
- **quarter**: Calendar quarter (1 to 4).
- **relative_date_PCIAT**: The number of days relative to the PCIAT assessment date (negative values indicate data collected prior to the assessment).

3) Target Variable: The primary target variable (**sii**) is derived from the **PCIAT_Total** score and represents the severity of problematic internet use. This variable is divided into four ordinal categories (0–3), making the task suitable for multi-class classification.

B. Challenges of the Data

1) Missing Data:: A significant portion of the dataset has missing entries, particularly in time-series measurements. This requires the application of robust data imputation techniques to fill gaps while minimizing the risk of introducing bias.

2) Class Imbalance:: The target variable (SII) is unevenly distributed across its four levels, with some categories likely underrepresented. This class imbalance makes it challenging to build models that generalize well to less common severity levels.

3) Heterogeneous Data Types:: The dataset comprises time-series data, static demographics, and behavioral metrics, each with unique preprocessing and modeling requirements. The diversity in data types adds complexity to feature engineering and integration.

4) High Dimensionality:: The dataset contains numerous features, including detailed activity logs and multiple demographic and health-related variables. Identifying the most relevant predictors while avoiding overfitting poses a challenge.

5) Temporal Patterns:: The time-series data captures dynamic behavior, necessitating the use of models that can effectively leverage temporal dependencies to improve predictive performance.

6) Feature Interactions:: The interaction between physical activity, behavioral metrics, and demographic factors may involve complex, non-linear relationships that are difficult to model directly.

C. Data Analysis

We will present the results of the data analysis, focusing on duplicates, missing values, and feature distributions in both the training and testing datasets.

1) Duplicates in the Dataset: The training dataset had 137 duplicate rows, while the testing dataset had none. We removed the duplicates from the training data to avoid overfitting and improve model performance.

2) Missing Data Analysis: Despite higher missing rates, the test dataset benefits from the absence of duplicate rows. Missing values were analyzed and visualized. In the training dataset, columns like ‘CGAS-Season‘ and ‘PreInt_EduHx-Season‘ had up to 35.48% and 10.61% missing values, respectively. Other columns also had significant missing data, suggesting the need for imputation or removal. Similar trends were observed in the testing dataset.

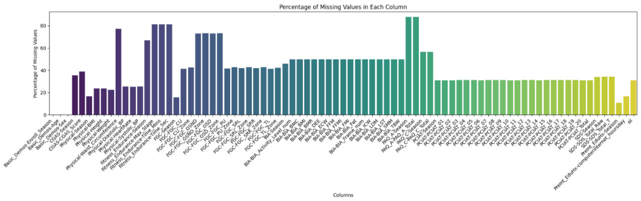

Fig. 1. Percentage of Missing Values in Each Column (Training Dataset)

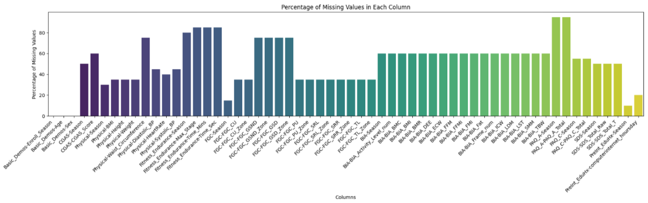

Fig. 2. Percentage of Missing Values in Each Column (Testing Dataset)

When comparing the training and test datasets: The test dataset generally exhibits higher missing rates across most features. For instance, CGAS-Season shows a missing rate of 35.5% in the training dataset and 50% in the test dataset, while BIA-BIA_BMI increases from 49.7% to 60%. Features such as PAQ_A-Season and PAQ_A-Total display extremely high missing rates in both datasets, with the test dataset missing 95% of data compared to 88% in the training dataset. Despite higher missing rates, the test dataset benefits from the absence of duplicate rows.

3) Detailed Exploration of some Features and the Correlation Matrix: We will delve deeper into 4 aspects of the training data: **PCIAT-PCIAT_Total (Total Score)**, **SII scores**, **Age group** distribution, **Basic_Demos-Enroll_Season (Season of Enrollment)**, **Internet Usage**. Additionally, we will examine the correlations between the 58 features and the SII scores, which serve as the target variable for prediction.

First of all, the Figure 3 demonstrates the distribution of SII scores and PCIAT scores. The Severity Impairment Index (SII) shows 30.9% missing data, raising concerns about its impact on model reliability. Among observed data, most individuals report no impairment (40.3%), followed by mild (18.4%), moderate (9.5%), and severe impairments (0.9%). The PCIAT_Total histogram reveals a right-skewed distribution, with most individuals scoring low, indicating minimal problematic internet use. Both distributions are skewed, with most data clustered at lower severity or PCIAT scores. However, the SII is affected by missing data, unlike the PCIAT_Total, which includes only complete responses.

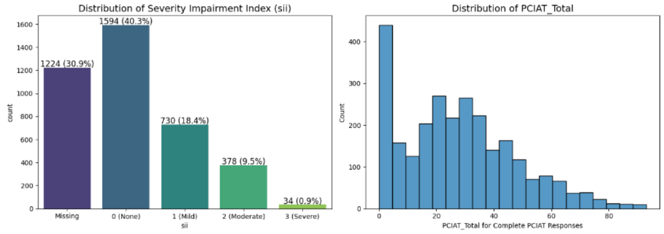

Fig. 3. The Distribution of SII and PCIAT_Total Scores

Secondly, we further analyze the SII distribution across different ages groups.

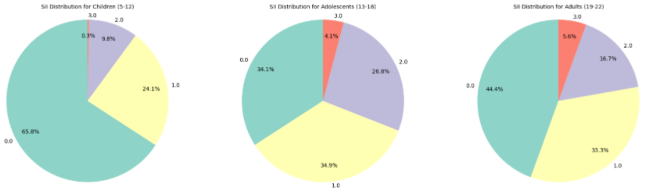

Fig. 4. The Distribution SII for Three Age Groups

The SII distributions across age groups show distinct patterns in severity levels and missing data: Children (ages 5–12): Most have no impairment (65.8%), with 24.1% reporting mild, 9.8% moderate, and 0.3% severe impairments. Adolescents (ages 13–18): Impairments increase, with 34.9% reporting mild, 26.8% moderate, and 4.1% severe, while no impairment decreases to 34.1%. Adults (ages 19–22): Missing data is significant (44.4%), but the distribution shows 33.3% mild, 16.7% moderate, and 5.6% severe impairments, with only 44.4% reporting no impairment. These findings highlight rising impairment severity with age, particularly in adolescents and adults, alongside increased missing data proportions.

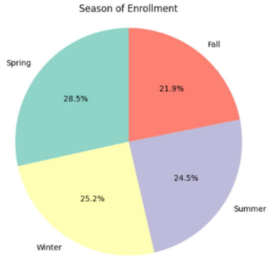

Fig. 5. Season of Enrollment

When examining the distribution of enrollment across each season, the data suggests that enrollment is relatively evenly distributed, with no significant peaks or drops. Spring leads with the highest enrollment at 28.5%, possibly reflecting alignment with academic calendars or student preferences. Surprisingly, Fall shows the lowest enrollment at 21.9%, which contrasts with its traditional role as a primary start period. Summer (24.5%) and Winter (25.2%) display similar percentages, indicating steady student engagement, likely driven by shorter courses or specialized programs.

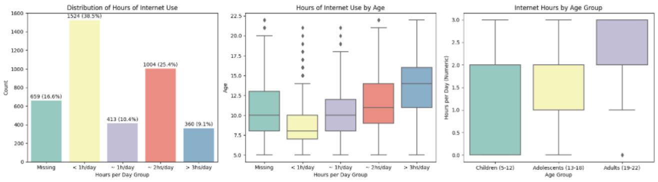

Fig. 6. Hours of Internet Use

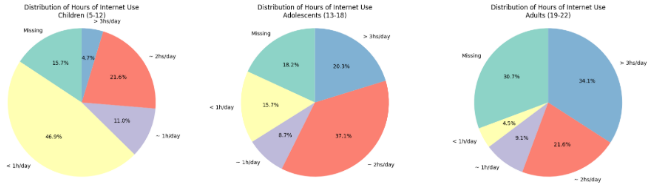

Fig. 7. Distribution of Hours of Internet Use

The distribution of internet use hours per day shows that the largest portion of participants (38.3%) report using the internet for less than one hour per day, totaling 1,524 individuals. A smaller percentage (16.6%) report missing data regarding their internet usage. Usage between one to two hours per day accounts for 10.4% (413 individuals), while 25.4% (1,004 individuals) report spending more than two but less than three hours online daily. The smallest group, 9.1% (360 individuals), spends over three hours online each day.

When examining the use of the Internet by age, the box plot reveals that younger children tend to use the Internet for shorter durations, with higher variability in older age groups. Adolescents and adults show an increasing trend in internet use, with more consistent usage patterns as age increases.

The final plot illustrates internet usage across different age groups. Children (ages 5–12) generally report fewer hours online compared to adolescents (ages 13–18) and adults (ages 19–22), where internet use gradually increases. Adults show the highest median usage, suggesting that internet engagement intensifies with age.

Finally, the analysis explores the correlation of all 58 features with SII scores.

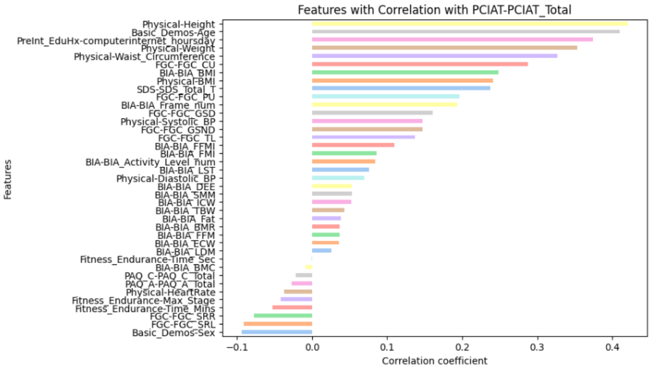

Fig. 8. Correlation matrix

The analysis of correlations with SII reveals several significant trends. Among the features, Basic Demos-Age shows the strongest positive correlation (0.41), indicating that older participants tend to have higher SII scores. Similarly, physical health indicators such as Physical-Height (0.42), Physical-Weight (0.35), Physical-BMI (0.24), and Physical-Waist Circumference (0.33) also exhibit positive correlations, suggesting a relationship between physical characteristics and SII. Additionally, PreInt EduHx-computer/internet hours per day (0.37) demonstrates a notable link between increased screen time and higher SII scores.

On the other hand, activity-related metrics such as PAQ A-PAQ A Total (-0.03), PAQ C-PAQ C Total (-0.02), Fitness Endurance-Max Stage (-0.04), and Fitness Endurance-Time Mins (-0.05) display weak negative correlations, implying minimal influence of physical activity on SII. Lastly, Gender (Basic Demos-Sex) has a slight negative correlation (-0.09), suggesting that females tend to have slightly lower SII scores compared to males. These findings highlight the complex interplay of demographic, physical, and behavioral factors in influencing SII variability.

III. Model Training and Evaluation

A. Base Models

1) Decision Trees: Decision Trees split data into subsets based on feature values to make predictions. They are easy to interpret and work with both numerical and categorical data. However, they can overfit if too deep, which can be mitigated using pruning or ensemble methods like RF.

B. Ensemble Methods

1) Random Forest (RF): Random Forest builds multiple decision trees and averages their predictions to improve accuracy and avoid overfitting. It's effective for both classification and regression and is robust with a large number of trees, though it can be slow for prediction.

2) Gradient Boosting (GB): Gradient Boosting sequentially builds weak learners (usually decision trees), each correcting errors of the previous one. It reduces bias and increases accuracy but requires careful tuning to avoid overfitting. It's flexible and powerful but can be slow in training.

3) XGBoost (Extreme Gradient Boosting): XGBoost is an efficient gradient boosting algorithm optimized for structured data. It offers regularization to reduce overfitting and is effective with missing data. Though sensitive to hyperparameter tuning, its strong performance makes it a key tool in predictive modeling.

4) CatBoost (Categorical Boosting): CatBoost is optimized for datasets with categorical and continuous features. It simplifies preprocessing by handling categorical data natively. While it offers high accuracy, it has longer training times and limited interpretability, making it valuable for complex datasets.

5) LightGBM (Light Gradient Boosting Machine): LightGBM is a fast, memory-efficient gradient boosting algorithm designed for large datasets. It supports categorical features natively and speeds up training with a histogram-based approach. Though sensitive to hyperparameters, its scalability makes it ideal for large-scale classification tasks.

6) Voting Regressor: The Voting Regressor is an ensemble method that combines the strengths of multiple models to improve predictive accuracy and reduce overfitting. The aggregation of model outputs helps balance biases, resulting in a more stable and robust predictive framework. This approach leverages the complementary strengths of various algorithms, enhancing the prediction of PIU.

C. Model Optimization And Evaluation

In the model evaluation phase, we utilized the **Quadratic Weighted Kappa (QWK)** as the primary metric for assessing the model's performance. The QWK is particularly effective for problems involving ordinal data, where the target variable has a natural order but the intervals between classes are not necessarily uniform. This evaluation method provides a more nuanced understanding of model performance, as it takes into account both the agreement between predicted and true values and the severity of prediction errors.

To further refine the evaluation process, we implemented the following techniques:

1. Threshold Rounding: The model's raw predictions were continuous values, which required mapping to discrete ordinal categories. This was achieved using a threshold-based rounding approach. Specifically, a function was defined to classify predictions into discrete classes based on predefined thresholds. For instance:

- Predictions below the first threshold were assigned to the lowest category.
- Predictions between thresholds were mapped to intermediate categories.
- Predictions above the final threshold were classified into the highest category.

This step ensured the model's outputs aligned with the ordinal nature of the target variable.

2. Cross-Validation: To assess the generalization capability of the model, we employed **Stratified K-Fold Cross-Validation**.

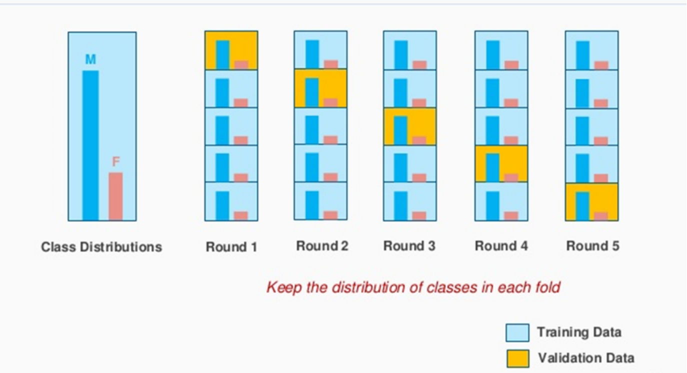

Fig. 9. Stratified K-Fold Cross-Validation (k=5)

This method divides the dataset into k folds while maintaining the class distribution within each fold. The model was trained on $k-1$ folds and validated on the remaining fold, iteratively across all folds. For each iteration, we computed:

- The QWK score on the training set, to evaluate how well the model fit the data.
- The QWK score on the validation set, to assess the model's ability to generalize.

The average QWK scores across all folds provided a robust measure of model performance.

3. Threshold Optimization: To enhance the accuracy of the predictions, we optimized the decision thresholds using a numerical optimization technique. By minimizing the negative QWK score, the optimization process determined the set of thresholds that maximized agreement between the predicted and true labels. The optimized thresholds were then applied to the validation predictions, resulting in an improved QWK score.

4. Performance Metrics: The evaluation process yielded the following key metrics:

- **Mean Training QWK:** An average of QWK scores across all training folds, indicating the model's fit to the training data.
- **Mean Validation QWK:** An average of QWK scores across all validation folds, reflecting the model's generalization ability.
- **Optimized QWK:** The final QWK score after applying the optimized thresholds, showcasing the model's refined performance.

The final model predictions on the test data were computed using the optimized thresholds. These predictions were averaged across all cross-validation folds, and the tuned thresholds were applied to produce discrete class labels. The results were formatted into a submission file for further evaluation.

D. Model training

1) Base Models: In this study, we initially tested several machine learning models, including Decision Tree and Logistic Regression. However, due to convergence issues with the Logistic Regression model, we ultimately decided to focus solely on the Decision Tree model for further evaluation.

Model Parameters:

```
DT_Params = {
    "max_depth": 5,
    "min_samples_split": 10,
    "min_samples_leaf": 4,
    "random_state": 42
}
```

The Decision Tree model was implemented with the above parameters. We used these hyperparameters to control the complexity of the tree, ensuring it could generalize well while avoiding overfitting. The `max_depth` limits the depth of the tree, `min_samples_split` and `min_samples_leaf` regulate the minimum number of samples required for splits and leaf nodes, respectively, and `random_state` ensures the results are reproducible.

Initially, Logistic Regression was considered as an alternative model, but due to convergence issues during training, it was excluded from the final analysis.

Model Parameters for Ensemble Model

```
# Model parameters for LightGBM
Params = {
    ’learning_rate’: 0.046,
    ’max_depth’: 12,
    ’num_leaves’: 478,
    ’min_data_in_leaf’: 13,
    ’n_estimators’: 300,
    ’feature_fraction’: 0.893,
    ’bagging_fraction’: 0.784,
    ’bagging_freq’: 4,
    ’lambda_l1’: 10,
    ’lambda_l2’: 0.01,
    ’random_state’: SEED,
    "verbose": -1,
    ’device’: ’cpu’
}

# XGBoost parameters
XGB_Params = {
    ’learning_rate’: 0.05,
    ’max_depth’: 6,
    ’n_estimators’: 200,
    ’subsample’: 0.8,
    ’colsample_bytree’: 0.8,
    ’reg_alpha’: 1,
    ’reg_lambda’: 5,
    ’random_state’: SEED,
    ’tree_method’: ’gpu_hist’,

}

# CatBoost parameters
CatBoost_Params = {
    ’learning_rate’: 0.05,
    ’depth’: 6,
    ’iterations’: 200,
    ’random_seed’: SEED,
    ’verbose’: 0,
    ’l2_leaf_reg’: 10,
    ’task_type’: ’GPU’
}

# RF parameters
RF_Params = {
    "n_estimators": 493,
    "max_depth": 5,
    "min_samples_split": 6,
    "min_samples_leaf": 1,
    "max_features": None,
    "bootstrap": True,
    "random_state": SEED,
}

# GB parameters
GB_Params = {
    "n_estimators": 295,
    "learning_rate": 0.012,
    "max_depth": 3,
    "min_samples_split": 18,
    "min_samples_leaf": 5,
    "subsample": 0.6,
    "max_features": None,
    "random_state": SEED,
}
```

2) Single Ensemble Model: In this study, machine learning models were implemented and evaluated independently to examine the predictive capabilities of each algorithm. Specifically, five standalone models were utilized, including: LightGBM (LGBM), XGBoost (XGB), CatBoost, Random Forest (RF), Gradient Boosting (GB).

a) Data Processing::

- Missing values were imputed using the median strategy via **SimpleImputer**.
- Categorical features (if any) were encoded appropriately

according to the requirements of each model.

b) Training Individual Models::

- Data was split using **K-Fold Cross Validation** (5 folds) to ensure stability and minimize bias during evaluation.
- Each model was trained and hyperparameters were optimized individually to achieve the best performance.

c) Performance Evaluation::

- The performance of each model was measured using **Mean Absolute Error (MAE)** and **Quadratic Weighted Kappa (QWK)**, providing a comprehensive view of prediction accuracy and agreement levels.

3) Voting Ensemble: To leverage the strengths of multiple algorithms, an ensemble approach was implemented using a **Voting Regressor** framework.

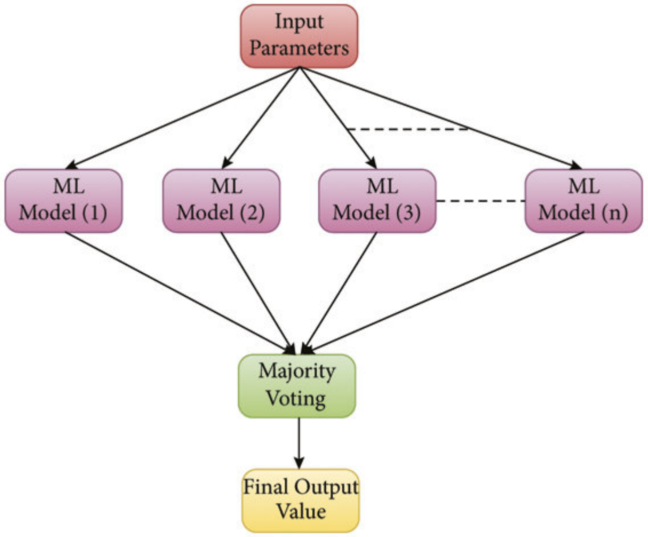

Fig. 10. Working of Majority Voting Ensemble algorithm

Process of Majority Voting:

- **Model Predictions:** For each instance in the dataset, multiple models make a prediction.
- **Comparison of Predictions:** The predictions from all the models are collected. Each model's prediction is treated as a "vote" for a particular class or value.
- **Selection of the Majority Vote:** The prediction that appears most frequently among the models is selected as the final prediction for that instance. If there is a tie (i.e., two predictions have the same frequency), a predefined rule (such as choosing the smallest or largest value) may be applied to break the tie.
- **Final Output:** The final prediction for each instance is determined by the majority vote, which combines the strengths of the different models and reduces the risk of errors from individual models.

This method combined the predictions from a variety of models and gradient boosting techniques. The primary goal was to create robust predictors capable of handling tabular data with high dimensionality and noise. Below is a detailed breakdown of the ensemble models developed for this study.

a) First Model:

- **LightGBM (LGB):** A gradient boosting framework that is highly efficient and effective for tabular data.
- **XGBoost (XGB):** A powerful boosting algorithm known for its performance in competitive machine learning tasks.
- **CatBoost:** A gradient boosting model designed to handle categorical features natively, reducing preprocessing requirements.

Pipeline:

- Missing values were imputed using a **SimpleImputer** with the median strategy.
- Each algorithm was encapsulated in a pipeline to ensure consistent preprocessing.

Ensemble Method:

- Predictions from the three models were combined using a **VotingRegressor**, which averaged their outputs.

Objective:

- To establish a baseline ensemble model utilizing diverse gradient boosting techniques.

b) Second Model:

- **Added Random Forest (RF):** A tree-based ensemble model known for its ability to reduce overfitting and handle non-linear relationships.

Pipeline:

- Maintained the same preprocessing and pipeline structure as the First Model.

Ensemble Method:

- The ensemble now included four algorithms: LightGBM, XGBoost, CatBoost, and Random Forest.

Objective:

- To increase the ensemble's diversity by incorporating a fundamentally different algorithm (RF), potentially improving performance on noisy data.

c) Third Model:

- **Added Gradient Boosting (GB):** A traditional boosting algorithm that, while slower, can provide additional predictive power when combined with modern algorithms.

Pipeline:

- Similar preprocessing and structure as the previous models.

Ensemble Method:

- Combined five algorithms: LightGBM, XGBoost, CatBoost, Random Forest, and Gradient Boosting, using the **VotingRegressor** framework.

Objective:

- To maximize the ensemble's predictive power by leveraging the strengths of five diverse algorithms.

IV. Results

Version	Model	Mean Train QWK	Mean Validation QWK	Optimized QWK Score	Private Score	Public Score
Version 1	Decision Tree	0.4586	0.3013	0.308	0.351	0.342
Version 2	LightGBM	0.7842	0.4081	0.454	0.351	0.342
Version 3	CatBoost	0.5525	0.3666	0.447	0.355	0.375
Version 4	XGBoost	0.8939	0.3897	0.445	0.365	0.379
Version 5	Random Forest	0.8327	0.3489	0.440	0.440	0.433
Version 6	Gradient Boosting	0.4778	0.3515	0.464	0.447	0.434
Version 14	Model 1	0.6281	0.3441	0.413	0.404	0.457
	Model 2	0.7619	0.4008	0.403		
	Model 3	0.9197	0.3850	0.449		
Version 28	Model 1	0.7240	0.4613	0.519	0.423	0.486
	Model 2	0.7595	0.3926	0.459		
	Model 3	0.9175	0.3803	0.450		
Version 31	Model 1	0.6407	0.4510	0.527	0.433	0.477
	Model 2	0.6819	0.3796	0.460		
	Model 3	0.8544	0.3708	0.443		
Version 39	Model 1	0.7240	0.4613	0.519	0.407	0.494
	Model 2	0.7595	0.3926	0.457		
	Model 3	0.6575	0.3708	0.462		
Version 47	Model 1	0.7486	0.3899	0.463	**0.449**	**0.455**
	Model 2	0.6973	0.3754	0.457		
	Model 3	0.6575	0.3708	0.462		

1) Base Model: In this study, the Decision Tree was used as the base model to compare with more complex machine learning models. Although the Decision Tree did not achieve the best performance across indicators like Train QWK and Validation QWK, it serves as a baseline model, providing a foundation for comparison with more sophisticated models.

2) Single Models: The performance of each model was evaluated using Mean Train QWK, Mean Validation QWK, Optimized QWK Score, Private Score, and Public Score. The results are summarized as follows:

- **LightGBM:** Achieved the highest Mean Train QWK (0.7842) but exhibited a significant drop in Mean Validation QWK (0.4081). Its Optimized QWK Score was 0.454, with Private and Public Scores of 0.351 and 0.342, respectively. This indicates overfitting to the training data.
- **CatBoost:** Performed more consistently across validation and test sets with a Mean Validation QWK of 0.3666 and an Optimized QWK Score of 0.447. Its Private and Public Scores were 0.355 and 0.375, respectively, making it slightly more generalizable than LightGBM.
- **XGBoost:** Demonstrated the highest Mean Train QWK (0.8939) but suffered from overfitting, with a Mean Validation QWK of 0.3897. Despite this, its Optimized QWK Score (0.445) and Public Score (0.379) were competitive.
- **Random Forest:** Showed balanced performance with a Mean Train QWK of 0.8327 and Mean Validation QWK of 0.3489. Its Optimized QWK Score was 0.440, with Private and Public Scores of 0.440 and 0.433, respectively, indicating robustness across datasets.
- **Gradient Boosting:** Had the lowest Mean Train QWK (0.4778), reflecting slower training dynamics. However, it achieved the highest Optimized QWK Score (0.464), with Private and Public Scores of 0.447 and 0.434.

The results reveal that each model has distinct strengths and weaknesses. LightGBM and XGBoost achieved high training QWK scores but showed signs of overfitting.

In contrast, Gradient Boosting, while slower in training, demonstrated the best generalization with the highest Optimized QWK Score and strong performance in both private and public test sets. CatBoost and Random Forest exhibited consistent and balanced performance, with Random Forest particularly excelling in robustness. These insights provide valuable guidance for selecting appropriate models based on specific project needs.

3) Single Ensemble Models: From version v14 to v28, the single ensemble model approach exemplified a continuous effort to enhance predictive accuracy through iterative refinement and strategic design. This methodology centers on integrating multiple base models into a cohesive framework, leveraging their collective predictive strengths to improve overall performance. The core objective was to harness the statistical and methodological diversity of the constituent models, fostering greater accuracy, generalization capability, and resilience in handling heterogeneous data.

Version v14 marked the initial implementation of this strategy, utilizing a VotingRegressor to aggregate predictions from several base models. This approach capitalized on the diverse decision-making processes of individual models, significantly boosting predictive accuracy to 0.457. The inclusion of varying model architectures enhanced the system's ability to manage missing values and mixed data types, contributing to more robust performance across different datasets. However, this ensemble technique introduced notable challenges, particularly in terms of computational cost and the necessity for meticulous tuning of model weights to achieve optimal performance. The increased complexity required careful calibration to balance contributions from each model effectively.

In subsequent iterations, the ensemble approach underwent substantial refinement, culminating in notable breakthroughs.

A key milestone emerged with version v28, where the cross-validation process was expanded to 10 folds. This adjustment aimed to enhance the reliability of the model's generalization estimates, boosting the public score to 0.486 and the private score to 0.423. However, this gain introduced potential risks, primarily an increased likelihood of overfitting. Additionally, the more extensive cross-validation process significantly heightened computational demands, raising concerns about scalability—especially when applied to larger datasets or more complex ensemble architectures.

By version v31, the ensemble had matured to incorporate a sophisticated blend of advanced learners, including LightGBM, CatBoost, and XGBoost—models celebrated for their speed and predictive accuracy. This iteration sought to balance performance gains with computational efficiency. The strategic integration of these gradient boosting algorithms allowed the ensemble to capitalize on their complementary strengths, resulting in a robust predictive framework. Despite these improvements, version v31 recorded a slight dip in public score to 0.477, while the private score rose to 0.433. This subtle decline highlighted the intricate trade-offs between model complexity and the iterative pursuit of optimal performance.

While the ensemble model demonstrated notable progress, it also revealed several limitations. The escalating complexity of the ensemble architecture demanded extensive computational resources and prolonged training durations. Overfitting persisted as a critical concern, especially when incorporating a large number of models without sufficient regularization or validation safeguards. Additionally, the challenge of managing missing values remained prominent. Certain imputation techniques, such as the KNN Imputer, inadvertently introduced bias by influencing the target variable, thereby compromising the integrity of predictions.

In conclusion, the single ensemble model approach delivered significant advancements in predictive accuracy through meticulous design and optimization across successive versions. However, it also underscored essential trade-offs between model complexity, computational efficiency, and generalizability. These findings highlight the importance of balancing these factors to develop practical, scalable, and reliable ensemble modeling solutions.

4) Voting Esemble Model: The transition to a voting ensemble model, implemented from version v39 to v47, marked a pivotal shift aimed at boosting predictive accuracy by consolidating outputs from multiple ensemble models. This approach is grounded in the principle that aggregating predictions from diverse models can yield superior performance, as each model offers distinct strengths and varying perspectives that collectively enhance the final result.

In version v39, the ensemble approach transitioned from single models to a voting-based framework, driven by the pursuit of greater robustness and accuracy. LightGBM, CatBoost, and XGBoost were selected as the core components, chosen for their complementary strengths and proven performance on structured datasets. Their predictions were aggregated using a weighted voting mechanism, enabling the system to harness the unique advantages of each model. This shift resulted in a public score increase to 0.494, though the private score declined to 0.407. The inclusion of multiple models enhanced resilience against missing data and fluctuations in feature importance, but it also introduced higher computational demands and resource requirements.

Development continued with further refinements. By version v47, an automated pipeline was implemented to streamline preprocessing and model training, minimizing manual intervention and boosting efficiency. Additionally, Random Forest and Gradient Boosting models were integrated into the ensemble, enriching the diversity of learners. Uniform imputation techniques were applied across all models, mitigating the impact of missing data without introducing bias. These enhancements culminated in the highest recorded private score of 0.449. However, the increasing complexity of the ensemble—driven by the addition of new models and finer tuning—prolonged training times and elevated the risk of overfitting, underscoring the delicate balance between accuracy and scalability.

While the voting ensemble approach delivered clear improvements in accuracy, it also exposed significant challenges. Computational costs and lengthy training durations posed scalability issues, limiting the feasibility of deploying the model on large datasets or in real-time applications. Additionally, balancing the weights assigned to each model within the voting mechanism required careful experimentation. Misaligned weight distributions could undermine the benefits of model diversity, potentially diminishing overall performance gains.

In conclusion, the voting ensemble model approach demonstrated the power of collaborative model predictions in driving accuracy improvements. However, the project underscored the necessity of balancing model complexity with computational efficiency. Future implementations will need to carefully manage these factors to ensure the ensemble framework remains practical, scalable, and adaptable to a wide range of scenarios.

V. Conclusion

This study highlights the potential of utilizing physical fitness and activity data to predict problematic internet use (PIU) among children and adolescents. By approaching the task as a multi-class classification problem and applying advanced machine learning techniques, we developed a robust and effective predictive framework. Ensemble methods—particularly the integration of diverse gradient boosting algorithms and voting-based ensembles—proved instrumental in capturing the complex, non-linear relationships within the dataset.

Our findings emphasize the value of combining model diversity with optimization strategies such as threshold tuning and cross-validation to enhance predictive accuracy. Notably, the refined voting ensemble demonstrated a strong balance between generalizability and performance, positioning it as a viable solution for addressing the nuances of real-world data.

Additionally, the source code is available on our GitHub repository: **CMI-PIU-Zenish**

In summary, we present a comprehensive workflow for tackling the Problematic Internet Use prediction task. Our process spans exploratory data analysis (EDA), data preprocessing, modeling, and optimization. Through EDA, we identified key trends that guided subsequent steps. In the preprocessing stage, we addressed missing values using various techniques to optimize time series data. During modeling, we iteratively refined single models, optimized parameters, and employed rounding strategies. The transition to ensemble models—including both single and voting-based approaches—marked a significant improvement in overall performance.

References

[1] Adam Santorelli, Arianna Zuanazzi, Michael Leyden, Logan Lawler, Maggie Devkin, Yuki Kotani, and Gregory Kiar. Child Mind Institute — Problematic Internet Use. https://kaggle.com/competitions/child-mind-institute-problematic-internet-use, 2024. Kaggle.