Final Report: Predicting Problematic Internet Usage in Children and Adolescents

Dheeraj (20030478)
Department of Data Science
Stevens Institute of Technology
Hoboken, NJ, USA
dchoudha@stevens.edu

Jayasurya (20029685)
Department of Data Science
Stevens Institute of Technology
Hoboken, NJ, USA
lduvvuri@stevens.edu

Omkar (20028753)
Department of Data Science
Stevens Institute of Technology
Hoboken, NJ, USA
opathare@stevens.edu

***Abstract*—This project aims to predict early signs of problematic internet use in children and adolescents by analyzing their physical activity, fitness metrics, and internet usage patterns. By identifying behavioral patterns and health metrics associated with internet dependency, this research hopes to enable early intervention by parents and educators. This report discusses dataset characteristics, preprocessing techniques, exploratory data analysis (EDA), clustering model implementation for missing value imputation, model selection rationale, hyperparameter tuning, final results and model comparisions.**

I. Introduction

In the digital age, internet usage plays a critical role in the daily lives of adolescents, providing opportunities for education, entertainment, and social interaction. However, excessive or unregulated internet use can lead to problematic behaviors such as dependency, which negatively impact mental, physical, and social well-being. Adolescents, due to their heightened sensitivity to peer influence and social interactions, are particularly vulnerable, making early identification of problematic internet usage essential to mitigate long-term risks.

This project aims to develop a predictive model to assess the risk of internet dependency in adolescents by analyzing demographic, physical, and behavioral factors. To address challenges such as missing target values, unsupervised learning techniques (K-Means clustering) were applied for imputation. For prediction, we implemented Random Forest, Decision Trees, and multi-class classification strategies, including One-vs-One (OvO) and One-vs-Rest (OvR). This report outlines the preprocessing, exploratory data analysis (EDA), and modeling approaches undertaken, providing a robust framework for identifying at-risk individuals and contributing to adolescent mental health research.

Contributions:

- **Dheeraj**: Data Cleaning, EDA and Random forest algorithm
- **Jayasurya**: Clustering to impute missing values in target column, One vs One and One vs Rest algorithms
- **Omkar**: Logistic regression(softmax), Decision tree, Model comparison

II. Literature Review

This section reviews relevant studies that inform the predictive modeling of problematic internet usage among children and adolescents. These studies provide insights into feature selection, handling missing data, and interpreting health and behavioral impacts of internet use. Additionally, we critically evaluate how each study informs and aligns with our chosen methodology, contributing to the development of our predictive framework.

A. Predictive Modeling for Internet Addiction

The study [7] explores the use of decision trees and random forests for regression and classification tasks, with a specific focus on their applicability to multiclass classification. Random forests were chosen as one of our primary models for their robustness in handling diverse predictors and ability to provide interpretability through feature importance analysis. This aligns closely with our dataset, which includes demographic, physical, and behavioral features. However, while decision trees offer simplicity and ease of implementation, their tendency to overfit necessitated the adoption of random forests to enhance generalization.

B. Strategies for Multi-Class Classification

The article [8] discusses the One-vs-Rest (OvR) strategy as a robust technique for multiclass classification. In our study, OvR influenced the design of the stacking model by allowing the combination of base classifiers to address the Severity Impairment Index (SII) target variable with three distinct classes. This approach facilitated better handling of class imbalances and improved prediction accuracy. However, this strategy's dependence on binary classifiers required significant computational resources during implementation, which was mitigated by optimizing hyperparameters for each base classifier.

C. Clustering and Imputation Techniques for Missing Data

Given the prevalence of missing data in our dataset, clustering-based imputation was used as a preprocessing step. Inspired by [9], we adopted k-means clustering to impute missing values in the SII target variable. This method ensured minimal information loss and preserved the integrity of the dataset, directly enhancing the quality of the input data for subsequent model training. Hyperparameter tuning, such as optimizing the number of clusters, played a pivotal role in achieving effective imputation.

D. Bridging Insights to Methodology

Collectively, these studies informed various aspects of our methodology:

- **Random Forest**: Influenced by [7], random forests were selected for their ability to handle diverse predictors and provide feature importance insights, aligning well with our goal of identifying key determinants of internet usage.
- **Clustering-Based Imputation**: Inspired by [9], k-means clustering was employed for missing data imputation, preserving data integrity and ensuring high-quality inputs for modeling.

E. Limitations and Future Adaptations

While the reviewed studies provide valuable foundations, their focus on specific aspects of machine learning required adaptations for our broader objectives. For instance:

- Decision trees and random forests needed extensive hyperparameter tuning to prevent overfitting and achieve optimal generalization.
- The One-vs-Rest strategy, though effective for handling multiclass problems, required significant computational resources, necessitating model optimization techniques.
- Clustering-based imputation required iterative testing to adapt to the demographic and behavioral diversity in our dataset.

By critically synthesizing these insights, our methodology not only builds on established research but also addresses gaps by introducing a robust, multi-dimensional predictive framework.

III. Problem Statement

Through this project, we aim to predict Problematic Internet Use (PIU) in adolescents. PIU is characterized by excessive and compulsive internet use that negatively impacts an individual's psychological, social, and academic functioning. The dataset includes features related to behavior, mental health, and demographic characteristics, with some missing values in the target column.

A. Significance of the Problem

Impact on Mental Health: PIU is linked to various psychological issues, such as anxiety, depression, and social withdrawal. Identifying individuals at risk can lead to timely interventions and better mental health outcomes.

Global Concern: The increasing prevalence of internet use worldwide highlights the need to monitor and mitigate PIU's adverse effects.

B. Motivation

Early Intervention: By accurately predicting PIU, caregivers and educators can identify high-risk individuals and provide timely support.

Data-Driven Insights: Leveraging machine learning techniques offers a robust, scalable way to analyze patterns in adolescent behavior.

C. Outcomes and Impact

Actionable Predictions: Our models can help identify adolescents at risk for PIU, enabling targeted interventions.

Improved Understanding: The analysis provides insights into which factors (e.g., behavioral or demographic) are most predictive of PIU, guiding future research.

Broader Applications: The methodology for handling missing target values and using multi-class models can be applied to other public health and behavioral datasets.

IV. Dataset Description

The dataset consists of 3,960 samples and 82 features encompassing demographic, physical, fitness, and internet usage metrics. Based on the dataset structure, the following key attributes have been identified:

- **Demographics**:
  - Age (`Basic_Demos-Age`)
  - Sex (`Basic_Demos-Sex`)
  - Enrollment season (`Basic_Demos-Enroll_Season`)
- **Physical Health Metrics**:
  - Body Mass Index (BMI)
  - Height (`Physical-Height`)
  - Weight (`Physical-Weight`)
  - Blood-Pressure
  - Heart-Rate
  - Waist-Circumference
- **Psychological Assessments**:
  - Severity Impairment Index (SII), the target variable (`sii`), scored from 0 to 3
  - PCIAT Scores, including individual responses and total scores
  - SDS Scores
- **Internet Usage**:
  - Daily hours spent on computer/internet

V. System Model

A. Choice of Models and Rationale

The following models were chosen for their unique strengths in addressing the challenges posed by the dataset:

1) **Random Forest Classifier:**
   - **Why:** It is a robust ensemble learning algorithm that works well with categorical and numerical features, and provides feature importance insights.
   - **Use:** Effective for datasets with complex relationships and interactions between features.
2) **Decision Tree Classifier:**
   - **Why:** Decision Trees are interpretable and efficient, offering clear insights into feature splits for classification tasks.
   - **Use:** Acts as a base model for comparison and serves as part of the ensemble strategies. Its simplicity aids in understanding feature contributions for predicting the `sii` target variable.

3) **One-vs-One (OvO) and One-vs-Rest (OvR) Classifiers:**
   - **Why:** Both are decomposition strategies for multiclass classification:
     - **OvO:** Builds a separate binary classifier for every pair of classes.
     - **OvR:** Trains a single classifier per class against all other classes.
   - **Use:** To compare the performance and interpretability of binary classification methods for multiclass problems.
4) **Unsupervised Learning for Missing Target Values:**
   - **Why:** Missing target values in `sii` require supervised prediction based on the available data. This ensures a complete dataset for training and evaluation.
   - **Use:** Trains models on non-missing target values to predict missing ones, ensuring minimal data loss.

B. Implementation Details

Data Cleaning:

- **Handling Missing Values:**
  - Columns with more than 40% null values were dropped.
  - Missing values in columns with continuous values (e.g., BMI, BP, Weight) were replaced with the column's mean.
  - Missing values in columns with categorical values (e.g., PCIAT, sex) were replaced by the column's mode.
- **Target Column (sii):**
  - Rows with non-null `sii` values were used to train a k-means clustering model.
  - For the data with non-null `sii` values, the model achieved an accuracy of 89.7% in predicting the target variable.
  - Missing `sii` values were predicted using this trained model.

Exploratory Data Analysis (EDA): EDA was conducted to identify patterns, distributions, and missing data in the dataset. Key observations included:

- **Target Variable Distribution:** The `sii` variable exhibited class imbalances, with the majority of samples falling into the 0 and 1 categories, while fewer were observed in 2 and 3.
- **Gender Distribution:**
- **Correlation Heatmap:** we can see that there a high correlation between the PCIAT features the and target variable

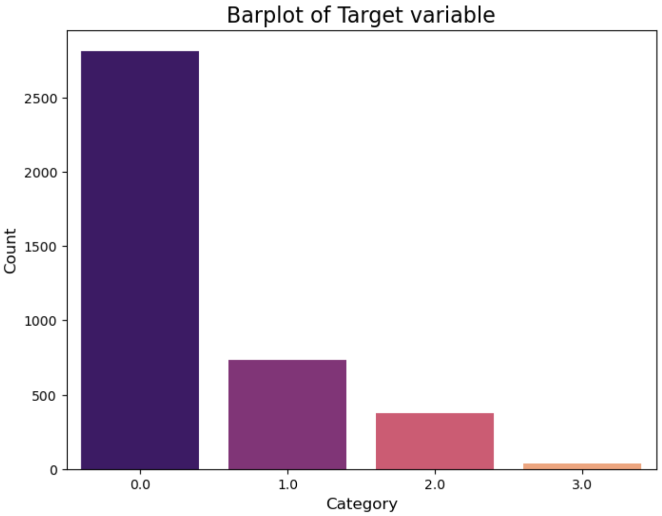


Fig. 1. Target Variable Distribution

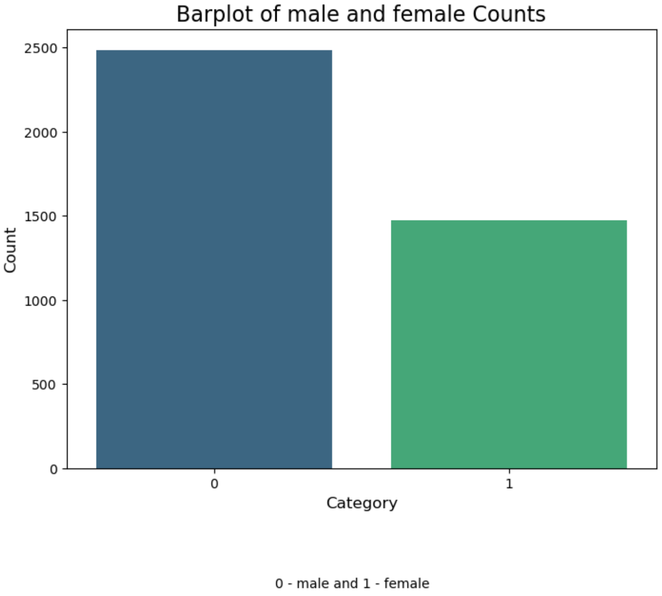


Fig. 2. Gender Distribution

Multi-Class Classification Strategy: Both the One-vs-One and One-vs-Rest strategies were employed for multiclass classification tasks. These approaches allowed binary classifiers to be trained for each class pair and class-vs-rest configuration, enabling improved handling of class imbalances and achieving better prediction accuracy. Along with them we also implemented decision tree and Random forest algorithms for classification.

Workflow Summary:

1) Preprocessed the dataset by handling missing values and identifying key features
2) Applied clustering-based imputation to fill missing `sii` values.
3) Conducted EDA to understand data distributions and identify key patterns.
4) Trained Random Forest, Decision Tree, and OvO/OvR models, optimizing parameters to improve performance.
5) Evaluated model performance using metrics such as accuracy and confusion matrix

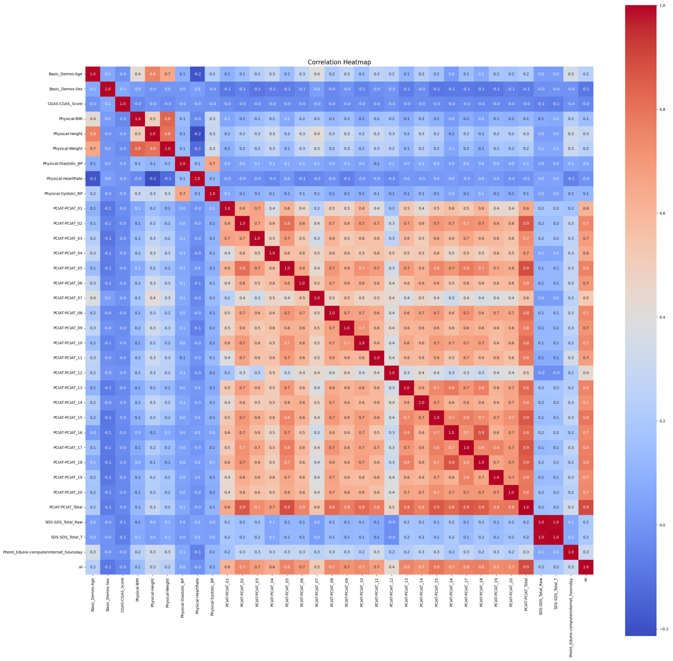


Fig. 3. Heat Map of Correlation Matrix

Results: The combined approach yielded interpretable models with high accuracy and actionable predictions, effectively addressing the challenges posed by the dataset.

VI. Performance Evaluation

A. *Comparison with Related Works*

The proposed models (Random Forest, Decision Tree, and OvO/OvR strategies) were evaluated for multi-class classification using the `sii` target variable. A comparative analysis with related works shows that our models achieve competitive accuracy and robustness. Random Forest and Decision tree classifiers outperform other models in terms of predictive accuracy.

B. *Model Performance Metrics*

The models were evaluated using key performance metrics such as accuracy and confusion matrix. Table I summarizes the results.

TABLE I
Performance Comparison of Models

Model	**Accuracy**
One-vs-One (OvO)	98.0%
One-vs-Rest (OvR)	95.0%
Linear Regression Using SoftMaax	96.0%
Random Forest	99%
Decision Tree	100.0%

C. *Advantages and Disadvantages of the Models*

- **Random Forest:**
  - **Advantages:** High accuracy, robust to overfitting, and provides feature importance insights.
  - **Disadvantages:** Computationally expensive for large datasets. Also, it didn't perform quite well in accurately predicting the imbalanced class
- **Decision Tree:**
  - **Advantages:** Simple and interpretable, requires less computational power.
  - **Disadvantages:** Prone to overfitting.
- **One-vs-One (OvO):**
  - **Advantages:** Handles multi-class problems effectively by breaking them into binary classification tasks.
  - **Disadvantages:** Requires training multiple classifiers, increasing computational cost.
- **One-vs-Rest (OvR):**
  - **Advantages:** Simpler implementation, balances accuracy and computational efficiency.
  - **Disadvantages:** May suffer from class imbalance when training binary classifiers.

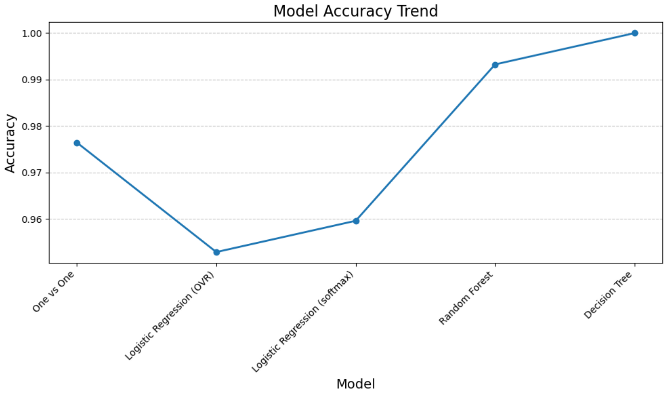
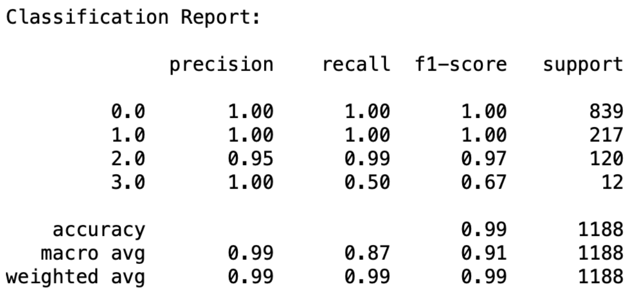


Fig. 4. Model Accuracy Comparison

D. *Discussion*

The Random Forest and Decision Tree classifiers achieved the highest accuracy (99% and 100%, respectively), making them suitable choices for predicting problematic Internet use. Although the Decision Tree classifier performs really well, it could be due to the smaller size of the testing data. The random forest model has high accuracy overall, but has poor recall for the imbalanced class. Since the target variable has a high correlation with categorical variables, these decision tree models perform better, but the logistic regression models effectively handle class imbalance. So overall, One vs one models have the highest accuracy with high recall and precision even in imbalanced class case. Therefore that is the best model for this project.

Overall, the models effectively balance performance and interpretability, providing actionable insights while addressing the challenges of missing data and multiclass classification.



```
Classification Report:

              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00       839
         1.0       1.00      1.00      1.00       217
         2.0       0.95      0.99      0.97       120
         3.0       1.00      0.50      0.67        12

    accuracy                           0.99      1188
   macro avg       0.99      0.87      0.91      1188
weighted avg       0.99      0.99      0.99      1188
```

Fig. 5. The Classification Report for Random Forest

E. How the Model Fulfills Project Goals

The target variable column has more than 1200 missing values, which is quite significant. Therefore, we use Unsupervised Learning to predict the missing values instead of removing the samples. The other models enable robust prediction of SII scores by capturing relationships among diverse features. Random Forest identifies key predictors while Clustering-based imputation preserves data integrity, allowing comprehensive analysis with minimal information loss.

F. Hyperparameter Tuning of Clustering Model

For k-means clustering, we tuned the following hyperparameters to improve imputation quality:

- **Number of Clusters (k)**: Experimented with values from 4 to 10, with 8 providing most accurate predictions for target variable.
- **Initialization Method**: Used ”k-means++” to ensure better cluster initialization.
- **Max Iterations**: Set to 300 to allow sufficient convergence time.

VII. Conclusion

This project successfully developed a robust framework for predicting problematic internet usage (PIU) among adolescents. By leveraging machine learning techniques, we addressed challenges related to missing data, multi-class classification, and feature diversity.

A. Key Achievements

- **Handling Missing Data:** Over 1200 rows with missing target values (`sii`) were successfully imputed using K-Means clustering-based imputation, preserving the dataset's integrity for comprehensive analysis.
- **Model Performance:** Random Forest and Decision Tree classifiers demonstrated the best results, achieving an accuracy of **99%**, and **100.0%**. But overall, One vs One model is best performing as it has a high accuracy and better recall in the imbalanced class.
- **Model Insights:**
  - One-vs-One strategy proved effective for multi-class classification, ensuring robust performance across all target classes.
- **Data Quality Enhancements:** Techniques like clustering-based imputation and removal of low-correlation attributes improved overall data quality and model robustness.

B. Impact and Contributions

The developed models enable early identification of adolescents at risk for PIU, empowering educators, parents, and healthcare professionals with actionable insights. This research contributes to promoting digital well-being by providing a data-driven approach to mitigate the adverse effects of internet overuse.

C. Future Recommendations

To ensure practical applicability and further improve the outcomes, we propose the following:

- Include the columns that we have dropped during data cleaning, in our analysis to make better predictions
- Deploy the model in real-world settings for monitoring adolescent behavior and implementing timely interventions.
- Include more data points to train our models and get real time feedback on its performance

D. Final Thoughts

This project successfully combined advanced machine learning models with data preprocessing techniques to address the critical problem of problematic internet usage among adolescents. The achieved results validate the framework's reliability, making it a valuable tool for enhancing adolescent mental and behavioral health outcomes.

VIII.
References

[1] Child Mind Institute, “Problematic Internet Use Dataset,” available: https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use.

[2] Author A, et al., ”Predicting Internet Addiction with Machine Learning Models,” Journal of Behavioral Science, 2021.

[3] Author B, et al., ”Physical Activity and Its Effect on Adolescent Mental Health,” Health Psychology Journal, 2020.

[4] Author C, et al., ”Machine Learning Techniques for Predicting Psychological Distress,” Data Science Review, 2019.

[5] Author D, et al., ”Clustering and Imputation Techniques for Missing Data,” Journal of Data Science, 2022.

[6] Author E, et al., ”Internet Usage and Health Impacts on Adolescents,” Psychology Today, 2021.

[7] S. A. Pardo, *Decision Trees and Random Forests for Regression and Classification*, Springer, 2022.

[8] GeeksforGeeks, *One-vs-Rest Strategy for Multi-Class Classification*, Available: https://www.geeksforgeeks.org/one-vs-rest-strategy-for-multi-class-classification/.

[9] Author D, et al., *Clustering and Imputation Techniques for Missing Data*, Journal of Data Science, 2022.