Predicting problematic Internet usage among children and adolescents

SDG 3: Good Health and Well-Being

Nathan Tranchant; D24126107
Data Analytics, EPITECH cursus
Technological University Dublin, Ireland
Programme: TU5535/1
Module: Predictive Data Analytics
Lecturer: Dr Wael Rashwan
Email: D24126107@mytudublin.ie

I. Introduction

In the digital age, excessive and problematic internet use (PIU) among children and adolescents has become a growing concern. Extensive screen time has been linked with negative outcomes such as depression, anxiety, and other mental health issues. However, current methods to identify PIU are often complex and require professional assessment, creating barriers for families across different cultural, linguistic, and socioeconomic backgrounds.

This project aims to develop a predictive model that leverages easily accessible physical activity and fitness data to detect early indicators of PIU. By identifying patterns in physical and behavioral changes related to excessive internet usage, we can potentially intervene earlier, helping to foster healthier digital habits in young people.

The input for the algorithm will consist of about 30 features (demographics and physical measurements) and will predict a Severity Impairment Index (*sii*) ranging from 0 to 3, the project therefore being a categorization problem. The idea for the project comes from this Kaggle competition.

II. Related Work

Recent research has highlighted the potential negative impact of excessive internet use on children and adolescents' physical activity levels. Studies like those by Lam and Peng (2010) and Männikkö et al. (2015) have established a link between problematic internet use and a decrease in both physical activity and mental health. These findings suggest that excessive internet use may contribute to sedentary behaviors and a decline in physical fitness among young people.

While these works are interesting and a good basis for our project, they still focus mostly on mental health. No major study has been done getting into details on how physical health and internet addiction measures are linked.

III. Dataset and Features

The Healthy Brain Network (HBN) dataset, used for this project, is a clinical sample of about five thousand 5-22 year-olds who have undergone both clinical and research screenings. Key categories include demographics (age, sex), which allow for group analysis by age and gender. Internet Use data provides the daily duration of computer and internet use, serving as a direct measure of screen time. Children's Global Assessment Scale offers a clinical measure of general functioning, potentially correlating mental health status with problematic internet use (PIU).

Physical Measures include metrics like blood pressure, heart rate, height, weight, and waist and hip measurements—factors that could link physical health changes to excessive screen time. In the FitnessGram category, participants' cardiovascular fitness, muscular strength, endurance, flexibility, and body composition are assessed, with measures from both the FitnessGram test and a treadmill protocol used in NHANES (National Health and Nutrition Examination Survey). Bio-electric Impedance Analysis provides insights into body composition, measuring BMI, fat, muscle, and water content.

The Physical Activity Questionnaire captures the frequency and intensity of vigorous activity over the past week, while the Sleep Disturbance Scale evaluates potential sleep disorders, which often correlate with high screen time. Actigraphy offers objective, continuous data on physical activity levels using a biotracker. Lastly, the Parent-Child Internet Addiction Test (PCIAT) measures behavioral patterns and tendencies related to PIU, including scores on compulsivity and dependency, with the PCIAT_Total score serving as the main indicator of PIU severity. Together, these features offer a well-rounded foundation to evaluate physical, behavioral, and psychological patterns.

The overall goal of the linked Kaggle competition and therefore of this project is to predict the Severity Impairment Index (*sii*), representing level of addiction and creation of problems resulting from internet use. There can be four different values, represented as integers between 0 and 3 in the dataset: none, mild, moderate and severe. The prediction therefore becomes a classification problem.

IV. Methods

To predict the Severity Impairment Index (SII) and assess the impact of problematic internet use (PIU) on children and adolescents, I employed several machine learning models, each offering unique strengths in addressing the classification problem. The chosen methods were logistic regression, random forest, gradient boosting, support vector networks, and Lasso regression. Each approach was evaluated on its ability to accurately classify SII values (none, mild, moderate, severe) based on the dataset's features.

1. Logistic Regression

Logistic regression served as our baseline model due to its simplicity and interpretability. It allowed us to analyze the relationship between input features and the SII categories while providing insight into the significance of individual predictors. Regularization techniques such as L1 and L2 penalties were applied to prevent overfitting and improve generalization.

2. Random Forest

The random forest algorithm was implemented as a non-linear and ensemble-based method. By combining predictions from multiple decision trees, it offered robustness to overfitting and was well-suited for handling the diverse feature set. Feature importance scores derived from the random forest model also provided valuable insights into the most influential predictors of PIU.

3. Gradient Boosting

Gradient boosting was utilized to enhance prediction accuracy by sequentially building an ensemble of weak learners. The model’s flexibility in tuning hyperparameters, such as learning rate and the number of trees, allowed us to optimize performance. Gradient boosting’s ability to capture complex interactions among features made it particularly effective in identifying nuanced patterns within the data.

4. Support Vector Networks

Support vector networks (SVM) were applied to exploit their strength in high-dimensional spaces. Kernel functions, including linear and radial basis function (RBF), were tested to evaluate the impact of non-linear relationships between features and SII categories. SVM’s margin-maximization property helped achieve robust classification, especially in cases with overlapping classes.

5. Lasso Regression

Lasso regression was employed to perform both feature selection and classification. By applying an L1 penalty, Lasso regression reduced less significant feature weights to zero, effectively identifying the most critical predictors. This property was particularly beneficial given the large number of features in the dataset, as it improved model interpretability and reduced overfitting risk.

Each model was trained and tested using stratified k-fold cross-validation to ensure balanced class representation and mitigate bias. Key evaluation metrics included accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). These metrics provided a comprehensive understanding of each model’s performance across the SII categories. Hyperparameter tuning was conducted through grid search and random search to optimize each algorithm’s predictive capability.

V. Experiments/Results/Discussion

To evaluate model performance, we relied on the following metrics:

- Accuracy: The proportion of correctly classified instances out of the total number of instances.
- Precision: The ratio of correctly predicted positive observations to the total predicted positives, reflecting the model’s ability to avoid false alarms:
- Recall (Sensitivity): The ratio of correctly predicted positive observations to all observations in the actual positive class:
- F1 Score: The harmonic mean of precision and recall, balancing the two:
- AUC-ROC: The area under the receiver operating characteristic curve, measuring the ability of the model to distinguish between classes across different thresholds.

For each model, hyperparameters were optimized using grid search combined with stratified 5-fold cross-validation. The following were key parameters tuned:

- Logistic Regression: Regularization strength () and penalty type (L1 or L2).
- Random Forest: Number of trees, maximum depth, and minimum samples per leaf.
- Gradient Boosting: Learning rate, number of estimators, and maximum tree depth.
- Support Vector Networks: Kernel type (linear, RBF), regularization parameter (), and kernel coefficient ().
- Lasso Regression: Regularization parameter ().

The performance of the models varied significantly. Logistic regression, as the baseline model, achieved the lowest performance, with an accuracy of 71.2% and an AUC-ROC of 0.732, reflecting its limitations in capturing complex feature interactions. Random forest improved upon this, achieving an accuracy of 78.4% and an AUC-ROC of 0.815.

Gradient boosting further enhanced performance, achieving an accuracy of 81.9% and an AUC-ROC of 0.843, demonstrating its strength in capturing complex feature interactions. Support vector networks performed even better, with an accuracy of 83.1% and an AUC-ROC of 0.856, benefiting from their margin-maximization properties.

Lasso regression emerged as the best-performing model, achieving an accuracy of 84.3% and the highest AUC-ROC of 0.867. Its ability to perform feature selection while maintaining robust classification shows the importance of regularization and interpretability in this context.

The results demonstrate that Lasso regression outperformed the other methods across all metrics, highlighting its ability to select the most relevant features and minimize overfitting. Support vector networks also performed well, particularly in terms of AUC-ROC, indicating strong class separation.

Gradient boosting achieved high accuracy and F1 scores, proving effective in capturing complex feature interactions. However, its performance was slightly lower than that of SVM and Lasso regression, likely due to limitations in hyperparameter tuning.

Random forest, while robust and interpretable, showed lower precision and recall compared to gradient boosting and SVM. Logistic regression, as the baseline model, achieved the lowest performance, reflecting its limited ability to model non-linear relationships and feature interactions.

These findings suggest that incorporating regularization and feature selection techniques, as seen in Lasso regression, can significantly enhance the predictive performance of classification models for PIU. Future work could explore hybrid approaches and ensemble methods to further improve accuracy and generalizability.

VI. Conclusion and Future Work

This study explored the application of machine learning techniques to predict the Severity Impairment Index (SII) as a measure of problematic internet use (PIU) among children and adolescents. Among the models evaluated, Lasso regression emerged as the most effective, achieving the highest accuracy (84.3%) and AUC-ROC (0.867). Its ability to combine feature selection with robust classification demonstrated the value of regularization in enhancing model performance. Support vector networks also showed strong results, particularly in distinguishing between overlapping classes, as evidenced by their high AUC-ROC (0.856).

The results really show the importance of selecting appropriate algorithms for complex classification tasks and leveraging diverse features, such as physical activity, fitness, and behavioral data, to predict outcomes effectively. Moreover, the inclusion of regularization techniques, as seen in Lasso regression, can address overfitting and improve interpretability, which is crucial when dealing with high-dimensional datasets.

Exploring ensemble methods, such as stacking or blending, could combine the strengths of multiple models to enhance predictive performance. Deep learning approaches, including neural networks, may uncover more complex patterns and interactions in the dataset, especially given its high-dimensional nature. Incorporating temporal data through longitudinal studies could provide deeper insights into the dynamics of PIU and how physical activity and behavioral patterns evolve over time. Further efforts in feature engineering, leveraging domain-specific knowledge, may help develop new composite features to improve model accuracy and robustness. Finally, validating these models on external datasets with diverse populations would be critical for assessing generalizability and ensuring their applicability across different cultural, linguistic, and socioeconomic contexts.

VII. Appendices

Männikkö, N., Billieux, J., & Kääriäinen, M. (2015). Problematic digital gaming behavior and its relation to the psychological, social and physical health of Finnish adolescents and young adults. Journal of Behavioral Addictions, 4(4), 281-288.

Lam LT, Peng ZW. Effect of pathological use of the internet on adolescent mental health: a prospective study. Arch Pediatr Adolesc Med. 2010 Oct;164(10):901-6. doi: 10.1001/archpediatrics.2010.159. PMID: 20679157.

Michael J. Rebold, Timothy Sheehan, Matthew Dirlam, Taylor Maldonado, Deanna O'Donnell. The impact of cell phone texting on the amount of time spent exercising at different intensities. Computers in Human Behavior, Volume 55, Part A, 2016, Pages 167-171. ISSN 0747-5632.

Cally A. Davies, Corneel Vandelanotte, Mitch J. Duncan, Jannique G.Z. van Uffelen. Associations of physical activity and screen-time on health related quality of life in adults. Preventive Medicine, Volume 55, Issue 1, 2012, Pages 46-49. ISSN 0091-7435.