Multiclassification: Problematic Internet Use Study

Camilla Ren
Department of Computer Science
Western University
yren345@uwo.ca

Yizhou Chen
Department of Computer Science
Western University
yche2593@uwo.ca

Maxwell Ding
Department of Computer Science
Western University
jding263@uwo.ca

***Abstract*— In today's digital age, problematic internet use among children and adolescents is a growing concern. To understand and address mental health problems such as depression and anxiety, some machine learning techniques can be used to help uncover trends and relationships between internet use and physical activity. The Severity Impairment Index (SII) was introduced and aids in predicting problematic behaviors among youth, thereby allowing for more focused interventions and preventive actions. This project used the Healthy Brain Network dataset, encompassing detailed physical activity data captured via wrist-worn accelerometers, alongside internet usage metrics, demographic information, fitness assessments, and health questionnaires from over 5,000 children and adolescents. This study combined both unsupervised and supervised learning techniques to develop predictive models, enhancing feature engineering and improving classification accuracy. The study successfully demonstrated how sophisticated data analytics can identify early signs of excessive internet use, providing a basis for interventions to promote healthier digital habits. Future work will focus on refining these models to increase their effectiveness and scalability for real-world applications.**

I. Introduction

Growing concern about problematic internet use among children and adolescents underscores the importance of this study. It uses physical activity and health data to predict and address unhealthy digital behaviors. Existing methods fall short when it comes to early detection, which is crucial for preventing the long-term negative effects of excessive digital engagement. This research addresses a significant gap by using physical activity data to understand and predict patterns of digital addiction in the youth population.

Our goal is to develop and validate the most effective predictive models for the early identification of problematic internet use in children based on their physical activity levels and other relevant health data. This involves integrating insights from unsupervised clustering methods to enhance feature engineering, applying advanced supervised machine learning techniques to improve classification accuracy, and effectively handling data imbalances and high-dimensional data. The study's findings demonstrate that our integrated machine learning approach significantly enhances the ability to identify early signs of problematic internet use, providing actionable insights that can be leveraged to design interventions aimed at promoting healthier digital habits. This research not only fills a crucial gap by enhancing early detection capabilities but also sets a new standard for utilizing machine learning in behavioral science to impact both theory and practice positively.

The structure of the report is divided into sections covering Exploratory Data Analysis (EDA), which outlines initial data findings and pre-processing steps; Unsupervised Learning, which details the clustering techniques and feature transformation; and Supervised Learning, which describes model building, evaluation and results. This comprehensive format allows the methods and efficacy to be fully understood, paving the way for future studies to build on this ground-breaking work.

II. Background & Related Work

Problematic Internet Use (PIU) among children and adolescents is a growing concern, but current assessment methods require complex professional evaluations that many can't access due to cultural, language and practical barriers. As young people spend more time online, rates of PIU are on the rise and this has implications for their mental and physical health. Studies show that excessive cyber-use may lead to social isolation, depression and sleep deprivation, affecting young people's wellbeing.

In order to develop predictive models, studies have looked at how PIU relates to behaviors such as impulsivity and compulsivity. For example, Ioannidis et al (2016) showed that the use of measures of impulsivity and compulsivity from specific scales could improve machine learning predictions and achieve good ROC-AUC scores. However, the focus of this research has largely been on adults, to the exclusion of younger groups who are increasingly affected by PIU.

There's an important gap in how different studies assess PIU, as they use different tools and criteria. This makes it difficult to compare the results and to create a standard treatment (King et al., 2013). Despite its success in other medical fields, machine learning is still new to psychiatry, particularly for PIU research. This presents an opportunity for growth, particularly in the use of techniques such as K-means or K-mode clustering, which can find hidden patterns and improve predictions by adding cluster-based features to the models (Jovič et al., 2024).

Relationships in complex datasets can be distorted by traditional methods of dealing with missing data, such as using simple mean, median or mode values. More advanced methods, such as K-Nearest Neighbours (KNN), provide better solutions by preserving the quality of the data and reducing bias, leading to more reliable and interpretable results (Hinić, 2012). By addressing these methodological and application gaps, this research helps to better understand PIU and supports the development of targeted treatments that are both effective and widely accessible.

Our study used data from the Healthy Brain Network (HBN), which has collected information from more than 5,000 young people aged 5 to 22 in clinical settings. The aim of the HBN project is the identification of biological markers that could improve the diagnosis and treatment of mental health and

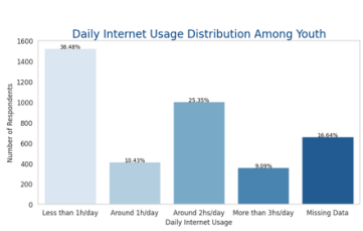
learning disorders. The dataset integrates different types of information: Physical activity data from wrist-worn devices, health assessments and detailed questionnaires about internet use. The physical activity measures include step counts, activity intensity and orientation of the wrist-worn device. The health data include basic demographic details and comprehensive fitness assessments using FitnessGram tests. These tests assess cardiovascular fitness, muscle strength, endurance, flexibility and body composition. In addition, the dataset includes results from the Parent-Child Internet Addiction Test, which assesses compulsive internet use behaviors. This rich set of data allows for the modelling of both physical activity and internet use patterns, making it easier to predict the SII, which indicates problematic internet use among participants.

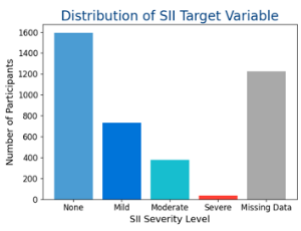
III. Exploratory Data Analysis

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar. The bar charts illustrate the distribution of daily internet usage among youth and the distribution of the SII target variable, categorized by severity levels.

The most common category, reported by 38.48% of respondents, is using the internet for less than 1 hour daily. A smaller percentage, 9.09%, indicated spending more than 3 hours online daily. Notably, 16.64% of responses fall under the "Missing Data." This indicates a significant variation in internet usage behavior, with a considerable proportion of participants falling into lower usage categories. The inclusion of "Missing Data" ensures transparency regarding incomplete records and allows for proper handling during analysis.

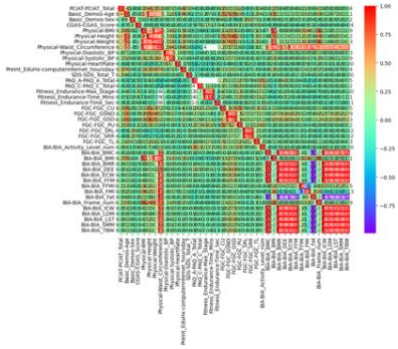
Also, most participants, approximately 1600, fall into the "None" category, indicating no signs of severity. The "Mild" and "Moderate" severity levels account for a smaller yet significant portion of the dataset. In contrast, only a few participants are classified in the "Severe" category. Additionally, a notable number of responses, around 1200, are categorized as "Missing Data." This distribution reveals an imbalance in the target variable, with "None" being the dominant category. Such imbalance must be carefully addressed in downstream analyses, such as clustering and predictive modeling, to prevent biased predictions.



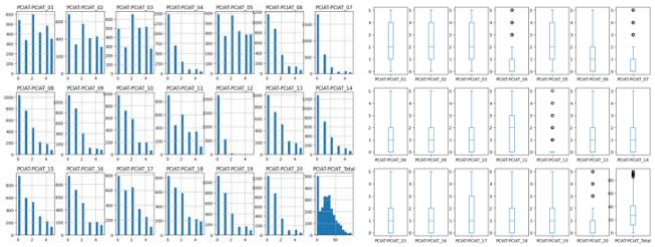


The distribution of daily Internet use highlights different usage patterns among young people, which will be further analyzed to explore potential associations between Internet use behavior and SII target variables. At the same time, the SII distribution highlights the need to address imbalances in the proportion of missing data and target classes. We will mitigate these issues in later analyses to ensure robust and unbiased predictive modeling.

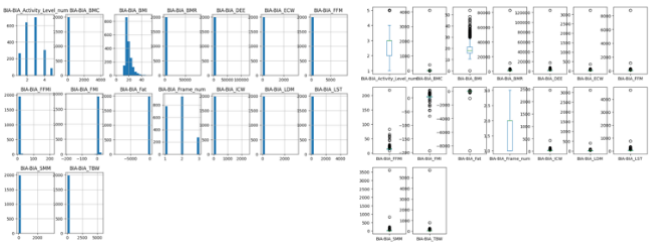
A correlation heatmap highlighted strong relationships within and across feature groups. For instance, BIA variables, including BMI and fat content, demonstrated high correlations, confirming their interdependence. Similarly, FGC variables exhibited moderate correlations, supporting the rationale for using zone-based clustering in further analyses.



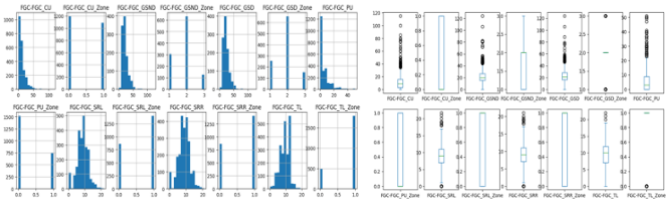
The PCIAT (Parent-Child Internet Addiction Test) reveals that most participants report minimal compulsive internet use, as indicated by skewed distributions in individual items (PCIAT_01 to PCIAT_20) and a broader range in the total score (PCIAT_Total). Box plots highlight outliers in individual items and the total score, suggesting a small subgroup with distinct behavioral patterns. These outliers may be valuable for clustering or supervised learning to uncover unique trends in internet addiction.



BIA (Bioelectrical Impedance Analysis) variables capture key body composition indicators such as BMI, fat content, muscle mass, and water content. The histogram of discrete features, such as BIA_Activity_Level_num, shows a tiered pattern consistent with the categorical data. The box chart highlights a wider range of features, such as BIA_SMM and BIA_TBW, indicating variability in muscle and water composition. We will normalize or transform outliers and wide value ranges, while discrete features will be applied directly to clustering.



The FGC (FitnessGram Child Assessment) variable measures aspects of physical fitness such as strength and endurance. Histograms display region-based variables (for example, FGC_CU_Zone) that exhibit significant classification separation and are therefore well suited for clustering. The box chart confirms that there are no outliers in the region-based variables but shows skewness and outliers in the continuous variables, highlighting the variability of fitness level. We will unsupervised learning with only the regional variables.



IV. Research Objectives

In our research to develop and validate the most effective predictive models for identifying early signs of problematic internet use in children, we utilized a comprehensive approach involving both unsupervised and supervised machine learning methods. Our research aims to address key challenges in data analysis and preprocessing. We focus on methods to deal with missing data and to optimize clustering techniques for different data subsets. At the same time, we aim to address the challenges of modelling complex and high-dimensional datasets using supervised learning models to advance the application and performance of multi-class classification tasks.

1) Objective 1: Prioritize addressing missing data challenges by implementing robust imputation techniques to enhance data completeness and ensure reliable analysis. This is essential for mitigating biases and improving model performance.

2) Objective 2: Emphasize evaluating and selecting optimal imputation strategies, such as mode filling, K-Nearest Neighbors (KNN), and iterative methods, to achieve a balance between computational efficiency and result consistency.

3) Objective 3: Explore clustering techniques like K-Means, K-Medians, K-Modes, DBSCAN, and Gaussian Mixture Models (GMM), aiming to align the chosen method with the data type for effective pattern identification.

4) Objective 4: Tailor preprocessing and clustering approaches to specific data subgroups (e.g., BIA, FGC, PCIAT), ensuring the outputs reflect meaningful insights that can support subsequent supervised learning tasks.

5) Objective 5: Prepare data inputs for the training of sophisticated models to support more accurate and meaningful analytical results through several pre-processing steps, including feature scaling, missing value handling, and the use of SMOTE to address class imbalance.

6) Objective 6: Integrate unsupervised learning cluster labels to enrich the dataset, adding structure and improving the model's ability to capture nuanced patterns in the data.

7) Objective 7: Evaluate and optimize advanced classification algorithms such as Random Forest, Gradient Boosting and Support Vector Machines to improve accuracy and handle multiple classes effectively, with a focus on refining model architectures and tuning parameters to improve generalization capabilities and prevent overfitting.

8) Objective 8: To investigate the integration of ensemble methods to create robust predictive models, with the aim of reducing the variance of predictions and improving accuracy, and to demonstrate the value of ensemble strategies in complex classification scenarios.

V. Research Methodology

A. Unsupervised Machine Learning - Data Preprocessing

To address missing data features with over 50% missing values were removed. This threshold was chosen to minimize noise and avoid unreliable imputations for highly incomplete features. The cleaned dataset retained a core set of features with enough data to lay the foundation for subsequent input and clustering tasks (Van Buuren, S., & Groothuis-Oudshoorn, K. ,2011).

We used the pattern-filling method, the KNN method and the iteration method to deal with the missing data and make a comparison (Zhang, S., Wang, L., & Ford, J. C.,2016). Mode filling replaced missing values with the most frequent value for each feature, providing a simple but limited approach. KNN Imputation estimates missing values based on the values of the nearest neighbors, effectively capturing nonlinear patterns by considering relationships between features. Iterative Imputation uses iterative regression to predict missing values based on other features, offering high accuracy but at the cost of increased computational complexity and a higher risk of overfitting in highly correlated datasets. The following density plot showed that all methods produced similar distributions for the imputed values, indicating comparable performance.

B. Clustering Techniques

We used three different clustering algorithms, their suitability tailored to the dataset's characteristics. First of all, we use K-means and K-modes to do the clustering. K-means was applied to numeric variables, such as PCIAT scores, as it excels with continuous data, while K-modes were used for categorical variables like BIA and FGC, leveraging its suitability for non-numeric data. To ensure fair weighting, numerical variables were standardized using z-score normalization, and categorical variables for K-Modes were one-hot encoded to preserve category information. The optimal number of clusters k was determined using the elbow method, which identifies the point of diminishing returns, and silhouette scores, with higher scores indicating better-defined clusters.

Secondly, we used DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to identify clusters of arbitrary shapes and handle noise effectively. This method was particularly useful for datasets with overlapping clusters or outliers. We standardized the numerical features to avoid bias due to scale differences. We also did parameter tuning, multiple

combinations of eps (neighborhood radius) and min_samples (minimum number of points in a cluster) were tested.

Finally, we used Gaussian Mixture Models (GMM) to assume clusters follow Gaussian distributions, making it suitable for continuous data with overlapping clusters. We also used scaled data with the number of components and silhouette scores to find the best model.

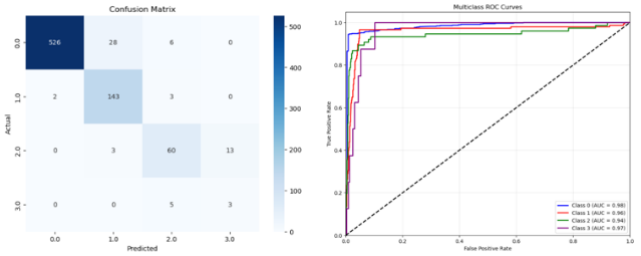
C. *Supervised Machine Learning - Data Preprocessing*

Medians were chosen to make imputed values robust to skewed distributions and outliers, preserving the integrity of underlying data patterns (Van Buuren & Groothuis-Oudshoorn, 2011). To ensure that we had all the necessary data, we first found all columns beginning with 'Physical' and then systematically replaced the null values with the median of each column. We also used the median to fill in any missing values for the target variable 'sii'. In this way, we were able to retain all the important data and maintain the representativeness of the dataset. By using this imputation strategy, our dataset was complete and better to train and test with. The use of medians prevented the data from being biased, making the machine learning models work better. It also provided a solid foundation for the classification tasks that followed, as keeping as many observations as possible maintained statistical power and improved the robustness of the predictive analysis.

Additionally, cluster labels derived from the unsupervised phase (K-means, DBSCAN, and GMM clusters) were integrated as new predictors to provide the model with latent structural information. After this, the data was split into training and testing sets, and SMOTE was applied to counter class imbalance, improving the model's ability to accurately classify multi-categories.

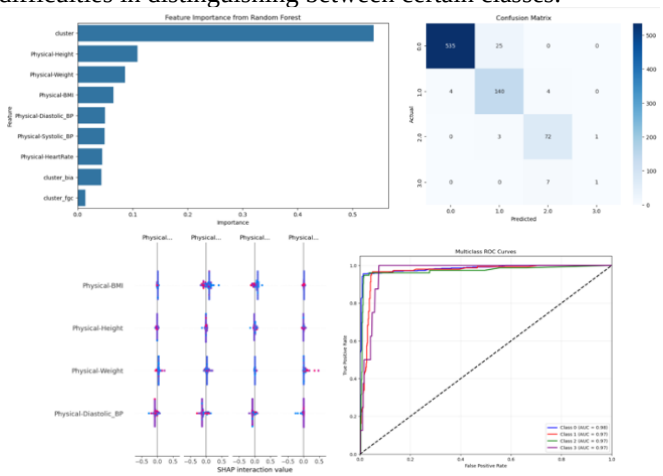
D. *Baseline Model*

The baseline model chosen was multinomial logistic, which is simple, easy to interpret, and has low computational costs. The data were standardized and balanced to provide solid starting points for later model comparisons (Hosmer, Lemeshow, & Sturdivant, 2013). The model's confusion matrix showed good performance in many cases, but we found some errors in classifying between classes that shared similar features. The model achieved good results in terms of ROC and AUC, with AUC values close to 1.0 for each of the classes. However, we consider this model to be just a starting point, and we plan to test more advanced models to see if they can provide any further performance improvements.



E. *Random Forest*

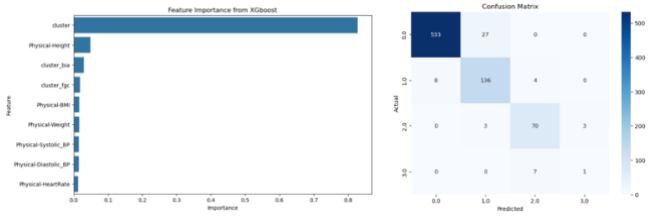
We used Random Forest because it works well with multi-class problems and can handle many features and complex relationships without much data preparation (Liaw & Wiener, 2002). Parameter tuning using GridSearchCV optimized the model for both accuracy and generalization across multiple classes: trees with 20 levels of depth, a minimum split size of 2, and a total of 500 trees. These settings helped to achieve an accuracy of around 97.4%. The model provided useful insights through its feature importance analysis, which showed that 'physical height' and 'physical weight' were the strongest predictors. After training the model, SHAP values were calculated to help us understand how each feature affected the model's predictions, giving us a clearer view of how the model makes decisions. Our analysis showed that physical measurements such as height, weight, and BMI played a key role in determining classifications. When we looked at the confusion matrix and the ROC curves, we found that the Random Forest was performing very well, with high AUC values between 0.97 and 0.98, although it did have some minor difficulties in distinguishing between certain classes.

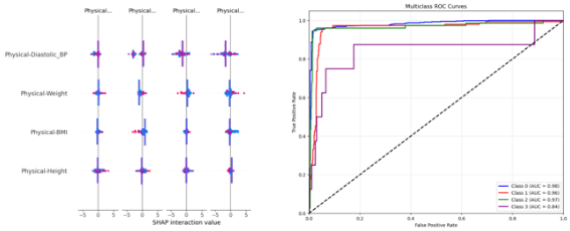


F. *Gradient Boosting Techniques*

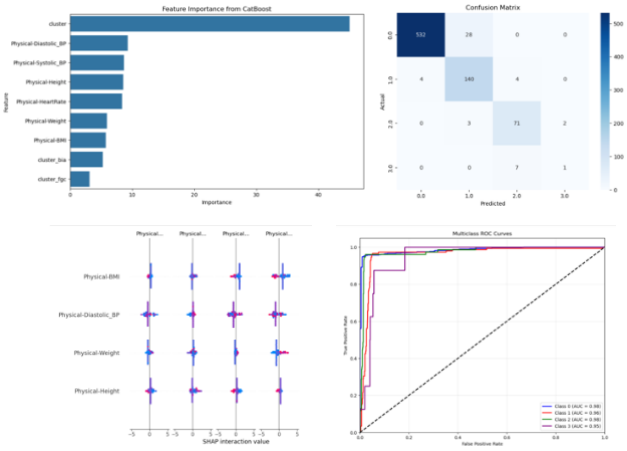
To further our analysis with gradient boosting techniques, we used XGBoost, CatBoost and LightGBM for their high efficiency in multiclass classification tasks. These models excel at handling high-dimensional data with complex feature interactions and help prevent overfitting.

XGBoost has a strong record of efficient and accurate handling of large datasets, and has demonstrated its efficiency on our large dataset. Feature importance plot showed that the "cluster" generated by the K-means clustering had a strong influence on the model predictions. This suggests that the model's decisions were significantly influenced by the K-Means clustering of similar data points, combined with health measures such as 'physical height'. With high AUC scores and accurate predictions across different classes, the model showed strong performance.

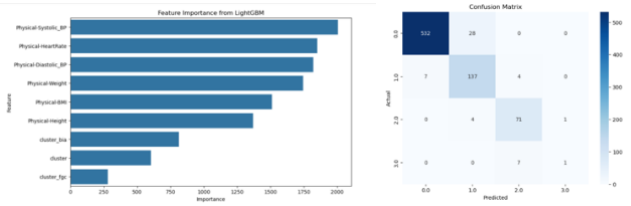


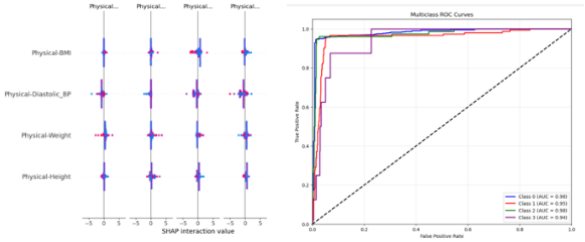


CatBoost effectively handled both categorical and numerical data with minimal preprocessing. To handle our complex multi-class problem and to optimise the processing of both categorical and numerical features, we optimised the model using a tree depth of 6 and a learning rate of 0.1. Similar to the plot from XGboost, feature importance also indicated that 'cluster' was highly influential, affecting the model's predictions significantly. SHAP analysis further identified 'Physical-Diastolic_BP' as a critical predictor and showed how this health measure had a strong influence on predictions across classes. While the model had a high level of accuracy overall, it had some weaknesses in the separation of classes with very similar characteristics.



LightGBM offers fast performance and uses less memory when processing large datasets. In this study, the model was configured with a learning rate of 0.1 and 100 estimators to provide optimal speed and efficiency in processing the numerous decision trees. This allowed the model to efficiently handle multi-class classification without significant resource consumption. Both the feature importance plot and SHAP analysis identified 'Physical-Height' as the most important predictor. This is consistent with their known importance in health assessments. Although the model performed well in general, its AUC values showed that it had some difficulty in distinguishing between certain classes that were similar.

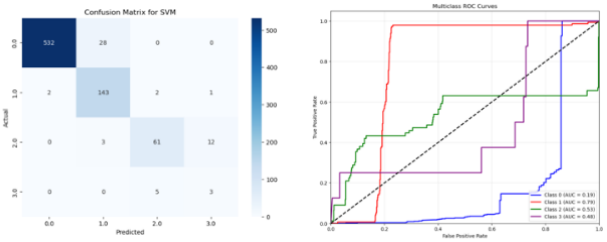




G. *Neural Network and Support Vector Machines*

We built our neural network as a multi-layer sequential model for multi-class classification. The structure included a dense input layer, multiple hidden layers using ReLU activation, and a softmax output layer. ReLU was chosen because, as noted by Goodfellow, Bengio, and Courville (2016), it handles nonlinear relationships well without changing the input scale. By generating probabilities that add up to one, the softmax layer helped us manage multiple classes. We used the Adam optimiser because it effectively adjusts the rate at which it learns, which improves the speed at which it learns (Kingma and Ba, 2014). The model used categorical cross-entropy for training with multiple classes. To prevent overfitting and improve accuracy with new data, the model was refined through repeated training and testing cycles. One challenge was to ensure that the model structure matched the format of our input data. This required adjusting the layer settings.

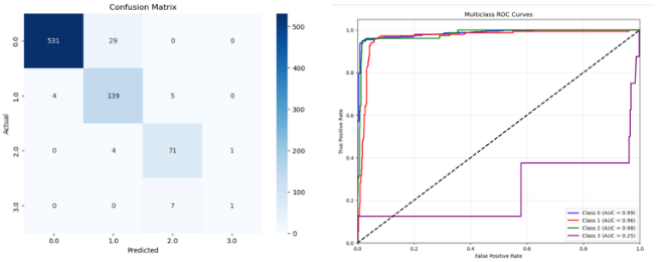
Our support vector machine (SVM) model used a linear kernel to balance simplicity and speed, which works well for many features, as James et al. (2013) suggested. To improve the performance of the SVM, we prepared the data using StandardScaler for normalisation. We used Principal Component Analysis (PCA) to reduce the number of dimensions, keeping the most important features while reducing the amount of overlapping information. The model created one classifier for each class, using a one-vs-rest approach to handle multiple classes. The C parameter was adjusted to balance accuracy and complexity. We evaluated performance by measuring true and false positives using accuracy, precision, recall, and ROC curves. A key challenge was selecting the right number of PCA components to retain important class discriminative features. Specifically, the ROC curves showed different levels of success for each class. The model performed best with classes 0 and 1, as indicated by their high AUC values. However, with Class 2 and especially Class 3, where the AUC scores were lower, the model had more difficulty. The performance of class 3 was particularly weak, with results that were close to chance, as shown by the curve close to the diagonal line. It is possible that the uneven class sizes in our data, or the fact that some classes are more distinct than others, explain these differences in performance.



H. *Ensemble Technique*

The ensemble method was implemented by integrating several boosting algorithms - XGBoost, CatBoost and LightGBM - into a stacked ensemble model, using a gradient boosting classifier as a meta-learner. This approach reduces errors and improves overall accuracy by leveraging the strengths of each model. We chose this method because it helps the model learn from different patterns in the data. At the same time, it prevents overfitting by combining different prediction strategies.

Looking at the plots, the ensemble performed extremely well for Class 0 and Class 2, achieving AUC values close to 1.00, showing that it correctly identified these classes with few errors. However, it struggled significantly with class 3, where it showed a low AUC of 0.25, indicating that it had some difficulty in correctly identifying this class. This weak performance for class 3 may reflect the way the features represent this class or a non-uniform distribution of classes in the data, suggesting a different model fit or balance for the data set.

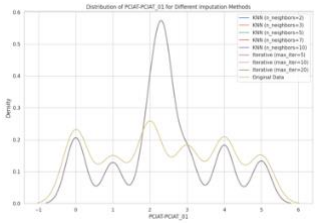


VI. Results

A. *Unsupervised Methods*

Multiple imputation methods were implemented and tested, including KNN with various numbers of neighbors (n = 2, 3, 5, 7, 10) and iterative imputation with different maximum iterations (max_iter = 5, 10, 20). Density plots comparing the distributions of the imputed data across these methods revealed minimal differences, demonstrating that all methods performed similarly in preserving the data's original structure.

Given this consistency, KNN with n = 5 was selected as the optimal method due to its balance of simplicity, computational efficiency, and ability to handle nonlinear relationships without the risk of overfitting, which can occur with iterative methods. This choice was validated by the close alignment between the distributions of the imputed data and the original data. The novelty of this result lies in the use of distributional analysis as a practical validation tool for selecting an imputation method, ensuring both accuracy and efficiency.



We used three different methods to do the clustering of PCIAT variables, which are K-Means, DBSCAN, and Gaussian Mixture Models (GMM) to identify the best approach. Silhouette analysis and visual inspection of the clusters were performed to determine the optimal method and parameters. Although K-Means with k = 2 produced the highest silhouette score, it was deemed unsuitable due to the small number of clusters, which failed to capture the complexity and variation within the PCIAT dataset.

K-means with k = 3 was chosen as the optimal solution, achieving a silhouette score of 0.3927. This configuration balanced cluster compactness and separation while providing more meaningful subgroup distinctions. Compared to DBSCAN and GMM, K-Means offered stable results and computational efficiency, making it the preferred method for PCIAT variables. The novelty of this result lies in prioritizing practical interpretability and subgroup differentiation over the highest silhouette score, demonstrating the importance of balancing statistical metrics with domain relevance.

Clustering the BIA data presented unique challenges due to the dataset's mixed categorical and numerical variables. The requirements included identifying clusters that reflected the inherent structure of the data while avoiding over- or under-clustering. K-Medians, DBSCAN, and GMM were tested, with silhouette scores and cluster compactness serving as evaluation metrics. While K-Medians and GMM with k = 2 achieved high silhouette scores, they failed to capture sufficient variation in the data, making them unsuitable for this application.

Ultimately, DBSCAN with eps = 2 was selected as the optimal method for clustering BIA variables, achieving a silhouette score of 0.6777. DBSCAN's ability to identify arbitrary-shaped clusters and handle noise effectively made it a superior choice, especially for datasets where the number of clusters is not predefined. This result underscores the flexibility of DBSCAN in handling complex data structures, marking a significant advancement in clustering mixed-variable datasets.

For the clustering of FGC variables, we do the same steps. While K-Medians with k = 5 provided reasonable results, it lacked sufficient separation between clusters, making it less suitable. DBSCAN (eps = 0.5) performed well in identifying noise but struggled to consistently form well-separated clusters due to its sensitivity to parameter tuning.

GMM emerged as the optimal method for clustering FGC variables, achieving the highest silhouette score of 0.696. It provided well-defined and interpretable clusters, aligning closely with the domain knowledge of FGC data. Despite being computationally more intensive, GMM's ability to model the data effectively justified its selection. The novelty of this result lies in the successful application of GMM to zone-based data,

showcasing its potential for clustering spatial or mixed-variable datasets.

In summary, we will use KNN imputation (n = 5) to ensure consistent handling of missing data. For clustering, PCIAT variables were grouped using K-Means (k = 3), BIA variables were clustered with DBSCAN, and FGC variables were best modeled using GMM. Then we will use these results to do the supervised learning.

B. Supervised Methods

In our analysis of supervised learning models, we evaluated several approaches for our multiclass classification problem. We compared Baseline, Random Forest, XGBoost, Neural Network, CatBoost, LightGBM, SVM and an Ensemble method to evaluate several supervised learning models for our multiclass classification problem. Our evaluation thoroughly assessed each model using a comprehensive set of metrics including accuracy, precision, recall, F1 score, Cohen's Kappa and ROC-AUC.

Random Forest and Ensemble performed best in accuracy, precision and F1 scores for multiclass classification. Random Forest stood out with the highest ROC-AUC of 97.21%, demonstrating superior ability to discriminate between classes. The Neural Network, which effectively handles complex data patterns, also showed strong performance across all metrics. While LightGBM and CatBoost performed well, they fell slightly short of the Random Forest and Ensemble methods, highlighting the advantages of the more sophisticated ensemble approaches. SVM performed weaker, particularly on ROC-AUC (49.72%), indicating that it struggled to separate classes using its linear approach.

In our project, we introduced a novel ensemble method that combines the predictions of XGBoost, CatBoost and LightGBM by using a gradient boosting classifier as a meta-learner. By leveraging the strengths of each model for better class separation, this method achieved 93.99% accuracy and 93.58% F1 score. However, Class 3 presented a challenge with a significantly lower ROC-AUC than its base models, suggesting possible overfitting or integration issues.

Our models were tested using separate training and test data sets to ensure that the models worked well with new data. Our system used a modular design, training each model separately before stacking to combine predictions, which allowed different model combinations to be easily tested and simplified system maintenance.

Overall, the three best performing models for multiclass classification in our analysis were Random Forest, CatBoost and Neural Network. Random Forest achieved the highest scores in almost all metrics. It had the highest accuracy at 94.44%, the highest recall at 94.44%, and an impressive ROC-AUC of 97.21%. Its high Cohen's Kappa showed a strong agreement between predicted and actual values, exceeding random chance. CatBoost performed almost as well, showing its strength in handling complex data with categorical features, with 93.94% accuracy, 94.12% precision and 96.72% ROC-AUC. The neural network also performed well with 93.81% accuracy and 93.86% F1 score. However, its ROC-AUC was slightly lower at 96.04%. These results are valuable for important applications where prediction accuracy is critical, and demonstrate how advanced algorithms can provide reliable and accurate predictions.

	Model	Accuracy	Precision	Recall	F1 Score
0	Baseline	0.924242	0.937667	0.924242	0.928944
1	Random Forest	0.944444	0.945840	0.944444	0.942920
2	XGBoost	0.934343	0.935165	0.934343	0.933720
3	Neural Network	0.938131	0.942163	0.938131	0.938577
4	CatBoost	0.939394	0.941240	0.939394	0.938528
5	LightGBM	0.935606	0.937511	0.935606	0.934335
6	SVM	0.933081	0.945974	0.933081	0.937482
7	Ensemble	0.936869	0.939937	0.936869	0.935824

	Model	Cohen's Kappa	ROC-AUC
0	Baseline	0.841642	0.962942
1	Random Forest	0.881564	0.972142
2	XGBoost	0.859703	0.937948
3	Neural Network	0.869325	0.960373
4	CatBoost	0.871368	0.967168
5	LightGBM	0.862689	0.963532
6	SVM	0.858750	0.497228
7	Ensemble	0.866164	0.793867

	Model	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa	\
1	Random Forest	0.944444	0.945840	0.944444	0.942920	0.881564	
4	CatBoost	0.939394	0.941240	0.939394	0.938528	0.871368	
3	Neural Network	0.938131	0.942163	0.938131	0.938577	0.869325	

	ROC-AUC	Average Rank
1	0.972142	0.938559
4	0.967168	0.932849
3	0.960373	0.931117

VII. Conclusion & Future work

A. Conclusion

To address the challenge of early detection of problematic internet use in children based on physical activity data and other health metrics, this comprehensive study successfully integrated advanced machine learning techniques. By effectively combining unsupervised and supervised learning, the study significantly advanced predictive modelling applications in behavioural science. Unsupervised methods such as KNN, DBSCAN and GMM improved the quality and structure of the data set, which in turn enhanced the ability of the supervised models to accurately identify nuanced patterns of behaviour. In particular, Random Forest and Ensemble models were particularly effective in this complex, high-dimensional data scenario due to their high accuracy and robust performance in multi-class classification.

The results of this study highlight the critical role that thoughtful data pre-processing and the innovative use of machine learning models can play in addressing the challenges of digital health. Key findings from this study highlight the critical importance of tailoring data pre-processing and clustering methods to the specific characteristics of the dataset, a process that was validated through rigorous metrics such as silhouette scores and detailed visual analysis. The integration of unsupervised learning to refine feature engineering significantly improved the performance of supervised models, as evidenced by high ROC-AUC scores, particularly for random forest and ensemble methods. This demonstrates a successful blend of methods that improves model accuracy and interpretability while managing computational demands. These results promise to have a significant impact on practice. They will provide a scalable, effective tool for early intervention strategies aimed at curbing excessive internet use among young people. The future direction of this research will focus on refining these models and exploring hybrid imputation techniques to further improve their predictive accuracy and efficiency.

B. Future Study

In order to further enhance the robustness, accuracy, and interpretability of the model, we propose several directions for future research. These efforts will address current limitations and introduce advanced methods for data preprocessing, feature engineering, and model optimization.

In this study, we rely on training sets and test sets for model evaluation. However, the lack of a dedicated validation set makes it difficult to assess the risk of overfitting. Future work could integrate validation sets to mitigate overfitting and provide more reliable generalizations. In addition, addressing category imbalances in the "none" category of the Severity Impairment Index (SII) through oversampling techniques, such as category weight adjustment, will improve clustering effectiveness. Outliers for continuous variables such as BMI and fat percentage and skewed distributions of PCIAT and BIA variables were identified as significant challenges. Transformations such as logarithmic or Box-Cox normalization, followed by repartitioning, can improve clustering accuracy and better align features with model requirements.

We removed variables such as seasonal data and columns with more than 50% missing values, including "FitnessGram Vitals and Treadmill." While this simplifies preprocessing, it may exclude potentially useful information. Future research could explore advanced interpolation techniques to retain key features and apply methods such as mutual information and the importance of feature arrangement to systematically evaluate and retain the most influential variables. We can also add feature engineering to address underfitting, such as building interaction terms between BMI and activity levels or aggregating time data to capture underlying patterns.

Although hyperparameter tuning was applied to random forests, future work could extend the tuning work to other models, such as gradient lift trees and neural networks, for more extensive optimization using GridSearchCV or RandomizedSearchCV. In addition, the exercise and physiological indicators that the participants received on a daily basis provided unique opportunities for time series modeling. We can use time analysis using methods such as ARIMA or Long short-term memory (LSTM) networks to enhance predictions of SII variables

By addressing these future directions, researchers can create more accurate, interpretable, and practical models based on this research. These improvements could not only improve the performance of existing models, but also provide a deeper understanding of the relationship between Internet use, physical activity, and mental health outcomes.

REFERENCES

[1] Buuren, S. van, & Groothuis-Oudshoorn, C. G. M. (2017, May 12). *Mice: Multivariate imputation by chained equations in R*. University of Twente Research Information. https://research.utwente.nl/en/publications/mice-multivariate-imputation-by-chained-equations-in-r

[2] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.https://www.researchgate.net/publication/269935079_Adam_A_Method_for_Stochastic_Optimization

[3] Hosmer , D. W., Sturdivant, R. X., & Lemeshow , S. (2013, March 22). *Applied logistic regression | wiley series in probability and statistics*. https://doi.org/10.1002/9781118548387.

[4] *Ioannidis, K., Chamberlain, S. R., Treder, M. S., Kiraly, F., Leppink, E. W., Redden, S. A., Stein, D. J., Lochner, C., & Grant, J. E. (2016). Problematic internet use (PIU): Associations with the impulsive-compulsive spectrum. An application of machine learning in psychiatry. Journal of Psychiatric Research, 83(Complete), 94–102. https://doi.org/10.1016/j.jpsychires.2016.08.010*

[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer Science & Business Media.

[6] Jović, J., Ćorac, A., Stanimirović, A., Nikolić, M., Stojanović, M., Bukumirić, Z., & Ignjatović Ristić, D. (2024). Using machine learning algorithms and techniques for defining the impact of affective temperament types, content search and activities on the internet on the development of problematic internet use in adolescents' population. Frontiers in Public Health, 12. https://doi.org/10.3389/fpubh.2024.1326178

[7] King, D. L., Haagsma, M. C., Delfabbro, P. H., Gradisar, M., & Griffiths, M. D. (2013). Toward a consensus definition of pathological video-gaming: A systematic review of psychometric assessment tools. Clinical Psychology Review, 33(3), 331–342. https://doi.org/10.1016/j.cpr.2013.01.002

[8] Liaw , A., & Wiener, M. (2022, December). *Classification and Regression by randomForest*. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf.

[9] Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. https://www.researchgate.net/publication/51057381_Multiple_Imputation_by_Chained_Equations_What_is_it_and_how_does_it_work

[10] Zhang, S., Wang, L., & Ford, J. C. (2016). KNN imputation for missing values in high-dimensional datasets. Proceedings of IEEE Big Data Conference. Imputation Method of Missing Values for Dissolved Gas Analysis Data Based on Iterative KNN and XGBoost | Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence