Child Mind Institute- Problematic Internet Use

Bettsy Garces, Luis Rojas, Isabel Sanchez

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Introduction

In today's digital era, excessive internet use among adolescents represents a growing public health challenge. Problematic internet use (PIU) is associated with depression, anxiety, academic impairment, and sleep problems. Although early detection is fundamental for effective interventions, identification remains difficult due to the subtle and variable nature of symptoms. Moreover, current solutions depend on costly professional evaluations, limiting access due to cultural and economic factors. This project responds to this need by developing an intelligent prediction system that integrates clinical and biometric data to detect PIU cases early in adolescents, enabling more accessible and effective preventive interventions.

Proposed Solution

The proposed solution is a web application that integrates artificial intelligence with clinical analysis to estimate problematic internet use severity. The system ingests two principal data sources: clinical and questionnaire CSV files, and Parquet files containing thirty days of wrist accelerometer motion data. A backend module performs data cleaning, normalization, and feature extraction on both input types before passing features to a CatBoost-based multiclass ordinal model. The model classifies severity into None, Mild, Moderate, and Severe. The platform returns real-time predictions, allows downloading CSV reports, and persists the trained model for fast reuse in subsequent sessions.

Architecture and Technology

The architecture follows a modular, scalable design that separates presentation, logic, and data processing layers. The frontend, implemented with HTML5, CSS3, and JavaScript, delivers an intuitive responsive interface for uploads and result review. The backend, built with Flask and Python, exposes RESTful APIs and orchestrates data flow between modules. Data processing relies on Pandas and NumPy for cleaning, normalization, and feature engineering. The CatBoost model performs ordinal multiclass prediction, and the trained model is persisted locally in pickle format for rapid reuse. This modular structure facilitates future extensions and maintenance.

Data Analyzed and Technology

The system analyzes multiple complementary dimensions of adolescent behavior and physiology to generate holistic assessments. Clinical inputs include demographics, biometric measures such as BMI and blood pressure, fitness test results, sleep habit summaries, and validated questionnaires such as the PCIAT for problematic internet use and the SDS for daytime somnolence. Sensor-derived data provide thirty days of continuous wrist accelerometer readings that reveal sleep timing, circadian rhythm patterns, and physical activity levels. Integrating these sources enables identification of temporal patterns and individual variability, supporting more personalized and context-aware assessments over extended observation windows.

Model Performance

The CatBoost model was trained on a curated dataset of 2,736 validated samples and showed robust predictive performance across standard metrics. Overall accuracy reached 85%, recall was 79%, and F1-score was 81%, indicating balanced sensitivity and precision. Class discrimination was strong with an AUC of 0.91. Feature importance analysis highlighted the PCIAT problematic internet scale, daytime somnolence measured by SDS, physical activity indices, BMI, and heart rate as primary contributors. These findings support the suitability of a multiclass ordinal approach for classifying problematic internet use severity. The model shows promise for integration into early screening workflows.

Systemic Analysis And Complex Management

The system was designed to manage the inherent complexity and variability of human behavior to ensure dependable operation across diverse real-world settings. Interactions between factors—such as physical activity, screen time, and sleep quality—are explicitly considered. Reliability measures include robust preprocessing pipelines, consistent scale normalization, anomaly detection routines, and automated handling of missing values. The platform continuously monitors data quality and surfaces descriptive error messages for problematic inputs. This systemic approach helps maintain model stability and trustworthiness even when deployed with data that are more heterogeneous than the original training set.

Challenges

Development faced significant technical challenges that strengthened system robustness. Data contained 30% missing values requiring automatic cleaning. Library version incompatibilities were resolved through dependency validation. Architecture was aligned with real prototype capabilities. We implemented automatic error handling, data normalization, and integrity validation. These solutions improved reliability and prepared the system for scalability. Each challenge was systematically documented to guide future improvements and expansions in production environments. The iterative problem-solving process reinforced software engineering best practices and data science principles.

Implemented Features

The application implements practical features tailored for clinical workflows and research needs. It supports drag-and-drop upload of CSV and Parquet files with automatic schema validation and immediate feedback. Real-time prediction yields downloadable CSV reports while interactive visualizations present model metrics, feature importance, and class distributions. Results tables include patient identifiers and summary statistics for quick review. The interface validates data integrity, handles errors with descriptive messages, and is responsive across modern browsers and devices. These capabilities facilitate efficient clinical screening, reproducible research, and straightforward export for downstream analysis or record keeping.

Conclusion

This project demonstrates the feasibility of using Machine Learning to detect problematic internet use early in adolescents. By integrating clinical and biometric data, the CatBoost model achieved 85% accuracy, validating the multiclass ordinal approach effectiveness. The modular architecture enables scalability toward more robust solutions. The system establishes foundations for tools that improve mental health assessment access, reducing clinical disparities. Future expansions include automatic retraining, clinical integration, and multilingual adaptations. This work contributes to demonstrating that well-designed technology can democratize diagnosis, improving global adolescent mental health. Early intervention enabled by this system could prevent more severe conditions.