National Economics University
Faculty of Economic Mathematics

DATA PREPARATION AND VISUALIZATION PROJECT

GROUP 1

Members: Vu Bui Dinh Tung
Doan Tung Lam
Truong Duc Anh
Thieu Dieu Thuy
Bui Thi Lan Anh

Class: DSEB 65B

Instructor: Dr. Nguyen Tuan Long

Hanoi, November 2025

Contents

List of Tables

List of Figures

Member Contribution Breakdown

Member Name	**Student ID**	**Contribution (%)**
Vu Bui Dinh Tung	11230602	20%
Doan Tung Lam	11230555	20%
Truong Duc Anh	11230516	20%
Thieu Dieu Thuy	11230590	20%
Bui Thi Lan Anh	11230509	20%

1 Before the Story

Machine learning systems are often evaluated by the performance of their final predictive models, yet the true determinant of success lies much earlier in the analytical workflow. This report explores the central theme that motivated our entire project: understanding the transformative power of data preparation in building reliable, interpretable, and practically meaningful machine learning solutions. Rather than treating data cleaning and preprocessing as technical preliminaries, we position them as the foundation upon which all subsequent modeling depends.

Before discussing algorithms or prediction metrics, it is essential to acknowledge the central reality of data-driven work: most of a data scientist's time is devoted not to modelling, but to preparing data for modelling. Industry surveys repeatedly show that between sixty and eighty percent of the analytical workload is spent on collecting, organizing, and standardizing data. These activities dominate the data science process because real-world data rarely arrive in clean, consistent, or analytically convenient forms. Problems such as missingness, measurement noise, inconsistent formats, and structural irregularities require substantial attention before any model can begin to learn meaningful patterns. As the well-known principle states: garbage in, garbage out. Poor data quality inevitably leads to misleading models, unreliable predictions, and distorted interpretations. Conversely, good data preparation not only improves accuracy but also clarifies the underlying "voice of the data", allowing models to reflect real behaviors rather than artifacts.

The urgency of high-quality data preparation becomes even more apparent when considering the context of our study: problematic Internet use (PIU) among children and adolescents. Evidence from recent psychological research shows that PIU has emerged as a significant behavioral concern. A 2024 study published in the *International Journal of Indian Psychology*, surveying 610 adolescents aged 12 to 19, reports that more than seventy percent show some level of problematic Internet use, with mild and moderate cases forming the majority. Although severe cases occur less frequently, their impact on well-being is substantial. These findings highlight the importance of modeling PIU accurately, responsibly, and with careful consideration of data quality, since misrepresentation of behavioral patterns may lead to flawed conclusions and inappropriate interventions.

In light of these considerations, our project is guided by one **Big Idea:** ***In this project, the main character is not the machine learning model, but the data preparation***. The goal is not merely to compare algorithms, but to demonstrate how the act of preparing data—cleaning, organizing, extracting, transforming, and engineering features—gradually unlocks the true predictive capacity of the models. Each step in the preparation pipeline reveals improvements not only in accuracy but also in stability, interpretability, and diagnostic value. Through this lens, model training becomes the outcome of a well-constructed data foundation rather than the starting point of the analytical process.

We would like to express our sincere appreciation and gratitude to our instructor, **Dr. Nguyen Tuan Long**, for his guidance, feedback, and encouragement throughout this project. His insights have shaped both our analytical approach and our understanding of how to communicate data-driven findings effectively.

2 Introduction

2.1 Introduction to the Story

In this project, the central narrative is not only about predicting problematic internet use, but about demonstrating the **power of data preparation** in a machine learning pipeline. Rather than treating data cleaning as a technical side note, we place it at the core of the story and show how it changes both the model and the insights that stakeholders receive.

From a storytelling perspective, we explicitly structure our work around the questions **WHO**, **WHAT**, and **HOW**:

- **WHO** is this story for? Our primary audience consists of data scientists, analysts, and students of data science who are interested in how systematic data preparation can transform noisy, high–dimensional data into reliable inputs for modelling. At the same time, we acknowledge a secondary audience: parents, educators, clinicians, and policy makers who care about the mental health of children and adolescents. For them, the narrative is not about algorithms, but about what the data reveals regarding youth behaviour and risk.
- **WHAT** will we show? We tell the story through statistical evidence and explanatory visualizations. For each stage of the pipeline, we report descriptive measures (e.g., class distributions, error rates, and recall by severity level) and pair them with carefully designed charts (confusion matrices, distribution plots, and comparative bar charts) that highlight the key patterns. The emphasis is on *how* the model behaves on raw data and how that behaviour changes after targeted preparation steps.
- **HOW** will the narrative unfold? The story is organised chronologically along the data preparation workflow: starting from raw, imbalanced, and noisy inputs; moving through cleaning, feature selection, transformation, and resampling; and ending with the resulting models. At each stage we document "before vs. after" effects on both data quality and predictive performance, so that the reader can see improvements accumulate over time rather than as a single final result.

In parallel with this technical storyline, we develop a secondary narrative aimed at readers who are also stakeholders in children's well–being. Using the same prepared data, we extract concise insights about different groups of variables collected in the study, such as physical activity patterns, sleep, and daily routines. For each group, we present short, focused observations that summarise what the data suggests about how young people live and use digital technology. This "voice of the data" complements the model–centric analysis and connects numerical results with real–world concerns.

To maintain engagement, we gradually shift from more formal statistical tables to a broader set of visual forms. After establishing the baseline behaviour of the model with numerical metrics, we introduce a variety of charts tailored to non-technical readers, thereby making the implications of data preparation accessible and compelling for both technical and non-technical audiences.

2.2 Introduction to the Dataset

Our empirical work relies on the *Child Mind Institute — Problematic Internet Use* competition dataset hosted on Kaggle [**?**]. The data are drawn from the Healthy Brain Network (HBN), a clinical sample of roughly 5,000 5–22-year-olds who have undergone both research and clinical screenings. The broader objective of the HBN study is to identify biological and behavioural markers that can improve the diagnosis and treatment of mental-health and learning disorders from an objective, data-driven perspective. For this competition, two components of the HBN protocol are used: (i) physical activity data, comprising wrist-worn accelerometry, fitness assessments, and related questionnaires; and (ii) self- and parent-reported internet-usage behaviour. From these sources, the task is to predict the *Severity Impairment Index* (SII), a four-level measure of problematic internet use.

The competition is organised as a *Code Competition*: the full test set is hidden and participants submit executable notebooks that generate predictions. In the public release, only a subset of the test data is made available as sample files in the correct format for development. The held-out test set contains approximately 3,800 instances for which the target label is known to the organisers but not to the competitors. This setting emphasises reproducible pipelines and places additional importance on robust data preparation.

The competition data are provided through two main sources:

- **Tabular files** in `train.csv` and `test.csv`, which contain participant-level measurements from a variety of clinical, demographic, and behavioural instruments.

- **Actigraphy series** in `series_{train|test}.parquet`, which store high-frequency accelerometer recordings collected by a wrist-worn device over multiple days for a subset of participants.

A distinctive characteristic of this dataset is the high degree of missingness: many measures are absent for most participants, and in the training data the target SII is itself missing for a non-trivial subset of cases. This motivates the exploration of unsupervised or semi-supervised techniques alongside standard supervised learning. In the official test set, by contrast, the SII value is present for all instances. A sample submission file, `sample_submission.csv`, is provided to illustrate the expected output format for predictions.

HBN Instruments The tabular data in `train.csv` and `test.csv` aggregate variables from multiple instruments, each documented in `data_dictionary.csv`. Key instruments include:

- **Demographics**: age and sex of participants.
- **Internet Use**: number of hours of daily computer/internet use.
- **Children's Global Assessment Scale (CGAS)**: clinician-rated numeric scale of general functioning for youths under 18.
- **Physical Measures**: blood pressure, heart rate, height, weight, waist and hip circumference.
- **FitnessGram Vitals and Treadmill**: cardiovascular fitness measured using the NHANES treadmill protocol.
- **FitnessGram Child**: health-related physical fitness covering aerobic capacity, muscular strength and endurance, flexibility, and body composition.
- **Bio-electric Impedance Analysis**: body-composition indices such as BMI, fat mass, muscle mass, and water content.
- **Physical Activity Questionnaire**: reported engagement in vigorous activities during the last seven days.
- **Sleep Disturbance Scale**: categorisation of paediatric sleep disorders.
- **Actigraphy summary variables**: objective indicators of ecological physical activity derived from a research-grade biotracker.
- **Parent-Child Internet Addiction Test (PCIAT)**: a 20-item scale measuring compulsive, escapist, and dependent internet-use behaviours.

A central field within the PCIAT instrument is `PCIAT_PCIAT_Total`, which aggregates responses into a single score. The competition target SII is derived from this score as described in the data dictionary: SII = 0 (None), 1 (Mild), 2 (Moderate), and 3 (Severe). Each participant is associated with a unique identifier `id`, which links tabular records to their corresponding actigraphy series.

Actigraphy Files and Field Descriptions During participation in the HBN study, some children and adolescents were asked to wear an accelerometer on the wrist continuously for up to 30 days while going about their normal daily routines. For each such participant, a time series is stored in `series_{train|test}.parquet/id={id}`, representing a continuous recording spanning many days. Each record in a series contains:

- **id**: the participant identifier, matching the `id` field in the tabular files.
- **step**: an integer timestep index within the series.

- **X, Y, Z**: raw acceleration (in g) along the three standard axes.
- **enmo**: the Euclidean Norm Minus One of the accelerometer signals (with negative values rounded to zero), a common magnitude-based measure of activity.
- **anglez**: the angle of the arm relative to the horizontal plane, derived from the accelerometer components.
- **non-wear_flag**: an indicator of whether the watch is currently worn (0) or not (1), inferred from the standard deviation and range of the signal.
- **light**: ambient light level in lux.
- **battery_voltage**: battery voltage in millivolts.
- **time_of_day**: timestamp representing the start of the 5-second window over which the data have been aggregated.
- **weekday**: day of the week (1 = Monday, ..., 7 = Sunday).
- **quarter**: calendar quarter (1–4).
- **relative_date_PCIAT**: number of days since the PCIAT test was administered (negative values indicate that actigraphy data were collected before the questionnaire).

These rich, high-resolution series, when combined with the heterogeneous tabular instruments, form a challenging but informative dataset. It is within this context that our study investigates how systematic data preparation can bridge the gap between noisy real-world measurements and reliable predictions of problematic internet use in young people.

3 Techniques for Data Preparation and Data Storytelling

3.1 Merging the Datasets

3.1.1 Why Merging the Datasets is Necessary

Before any meaningful modelling or storytelling can take place, we must first unify the information scattered across the two primary data sources of this competition: the tabular dataset (`train.csv` / `test.csv`) and the actigraphy time series (`series_{train|test}.parquet`). Although both originate from the same study and refer to the same participants, they are stored in fundamentally different formats, at different levels of granularity, and cannot be analysed jointly without an explicit merging step.

The tabular dataset provides participant-level variables: demographics, clinical scales, physical measures, and questionnaire scores, including the Parent–Child Internet Addiction Test (PCIAT) from which the Severity Impairment Index (SII) is derived. These variables are compact and easy to manipulate; they already fit the familiar "one row per participant" structure used by most machine learning algorithms. In contrast, the actigraphy data are high-frequency time series recorded over many days for a subset of participants. Each record in the actigraphy files corresponds not to a person, but to a *time step* for that person. As a consequence, the two datasets live in different "shapes": one is wide and static, the other is long and dynamic.

Analysing only the tabular dataset would ignore a large amount of rich behavioural information that is potentially predictive of problematic internet use. Daily activity patterns, circadian rhythms, sleep–wake regularity, and sedentary periods leave distinct signatures in the actigraphy streams. These patterns are precisely what we hope to exploit to identify early signs of problematic internet use. Conversely, working exclusively with the raw actigraphy series would disconnect these patterns from their clinical and demographic context, making it difficult to interpret any model and to relate its predictions back to meaningful outcomes such as SII.

Merging the datasets via the shared participant identifier `id` allows us to reconcile these two perspectives. At a technical level, merging is the step where we (i) extract features from the actigraphy series, (ii) aggregate them to the participant level, and (iii) align them with the corresponding rows in the tabular file. The result is a single, coherent analytical dataset in which each participant is represented by both static attributes (age, sex, clinical scores) and dynamic behavioural summaries (actigraphy-derived features). This unified view is essential not only for model training, but also for data storytelling: it is much easier to communicate insights when every row in the dataset corresponds to a clearly identifiable individual with a complete set of features.

In summary, merging the tabular and actigraphy datasets is not a minor implementation detail, but a foundational part of the data preparation story. It transforms fragmented sources of information into a single analytical view, enables the joint use of clinical and behavioural signals, and sets the stage for building models that are both more predictive and more interpretable in the context of problematic internet use among children and adolescents.

3.1.2 Merging Strategy and Feature Construction

Having motivated the need to combine the tabular and actigraphy data, we now describe the concrete strategy used to merge these sources into a single, model-ready dataset. Our approach follows a two-step logic: (i) transform the high-frequency actigraphy time series into participant-level features that are compatible with the tabular structure, and (ii) align and merge these features with the clinical and questionnaire variables using the shared identifier `id`. In addition, we construct domain-informed activity indicators based on ENMO to retain interpretable information about movement intensity.

From raw time series to participant-level statistics The actigraphy files are stored as separate Parquet partitions, one directory per participant, containing multiple columns (e.g. `X`, `Y`, `Z`, `enmo`, `anglez`, `light`, `battery_voltage`, temporal fields, and flags). Each row corresponds to a 5-second epoch, so the raw data are several orders of magnitude longer than the tabular file and cannot be merged directly in their original form.

To construct a participant-level representation, we process each actigraphy file independently and compute descriptive statistics for all numeric variables. Concretely, for each participant we:

- read the corresponding Parquet file and drop the `step` column (which is only a running index and does not contain substantive information per se);
- apply `pandas describe()` to all remaining numeric columns to obtain standard univariate summaries (count, mean, standard deviation, minimum, quartiles, and maximum);
- transpose this summary table so that each original signal (e.g. `enmo`, `anglez`, `light`) becomes an index entry and each statistic becomes a column;
- flatten the resulting matrix into a single feature vector by concatenating the original column name and the statistic name, yielding feature names of the form `<signal>_<statistic>` (for example, `enmo_mean`, `enmo_std`, `light_75%`).

The outcome of this procedure is a wide, participant-level DataFrame where each row corresponds to a unique `id` and each column encodes a summary property of the underlying time series. To scale this computation to thousands of participants, we parallelise the processing of Parquet files using a thread pool, then collect and concatenate all results into a single table `train_ts` (for the training set) and `test_ts` (for the test set). We explicitly retain the list of time-series-derived feature names (`time_series_cols`) for later reference.

Aligning tabular and actigraphy features Once the actigraphy summaries are available, we perform a left merge between the tabular files and the time-series tables on the common identifier `id`. This yields an enriched `train` and `test` DataFrame in which each participant is described by:

- clinical, demographic, and questionnaire variables from `train.csv`/`test.csv`, including demographics, physical measures, CGAS, FitnessGram, BIA, PAQ, SDS, and Internet Use;
- aggregated actigraphy statistics derived from the Parquet series.

We adopt a left join in order to preserve all rows from the original tabular files, even for participants without actigraphy data, and allow missing values in the time-series-derived features where appropriate. After the merge, the `id` column is no longer needed for modelling and is dropped from the feature matrix. We then define an explicit list of feature columns (`featuresCols`) consisting of selected tabular variables (demographics, clinical scores, physical and fitness measures, internet-use indicators, and the target `sii`) and extend this list with all time-series feature names in `time_series_cols`. The training DataFrame is restricted to these columns and any rows with missing SII labels are removed, ensuring that the supervised learning stage operates on a clean target vector. Finally, we record the subset of seasonal/categorical fields to be handled as categorical predictors in subsequent preprocessing.

Domain-informed ENMO activity profiles In addition to generic summary statistics, we construct a small set of domain-informed features that summarise each participant's activity profile in terms of movement intensity. For this purpose we focus on the ENMO signal (Euclidean Norm Minus One), a commonly used magnitude-based measure of acceleration. For each actigraphy series, we apply the following procedure:

- restrict the analysis to periods where the device is worn by filtering out all epochs with `non-wear_flag = 1`;
- classify each 5-second epoch into one of three activity types according to ENMO thresholds: *sedentary* (very low ENMO), *light* (intermediate ENMO), and *moderate* (higher ENMO);
- compute the total number of worn epochs and the proportion of time spent in each activity category relative to this total.

This yields three interpretable features per participant: `sedentary_por`, `light_por`, and `moderate_por`, which quantify the composition of daily activity. We apply this extraction function to all actigraphy directories in the training and test series, resulting in two additional participant-level tables (`train_enmo` and `test_enmo`) keyed by `id`. These tables are then available to be merged into the main feature set in the same way as the generic time-series summaries.

Overall, this merging strategy turns heterogeneous, multi-scale data into a single, coherent analytical dataset: each child or adolescent is represented by a mix of static attributes, clinical assessments, aggregated signal statistics, and interpretable movement profiles. This unified representation is the foundation on which we can fairly compare models, diagnose problems with the raw data, and demonstrate how successive data preparation steps improve both the quality of features and the reliability of predictions.

3.2 Model training on unprepared competition data

3.2.1 Train - test split

Before analysing the impact of data preparation, we first trained a baseline model on data that had only undergone minimal cleaning and merging. An important practical constraint arises from the fact that our dataset comes from a Kaggle code competition. In this setting, the official test set provided by the organisers does not contain the target variable `sii`, and is intended solely for blind submission to the leaderboard. As a consequence, the Kaggle test data cannot be used for model selection or diagnostic evaluation: any performance score obtained on the public leaderboard would be noisy, partially based on a hidden split, and subject to overfitting through repeated submissions.

For the purpose of this study, we therefore construct our own train–test split from the labelled training data. After merging the tabular and actigraphy features and applying basic preprocessing, we obtain a single labelled dataset `train_processed.csv` in which each row corresponds to one participant and includes the target `sii` alongside all predictor variables. From this dataset, we separate the predictors X (all columns except `sii`) and the target vector y (the `sii` column). The official Kaggle test file `test_processed.csv` is kept aside as an unlabeled feature matrix $X_{\text{submission}}$ that will be used only at the very end to generate the competition submission file.

To approximate the train–test separation that would exist in a fully labelled scenario, we randomly partition the labelled data into an internal training set and an internal test (validation) set. Specifically, we allocate 80% of the observations to the training set and hold out the remaining 20% as a proxy for an independent test set. The split is performed with a fixed random seed to ensure reproducibility and, crucially, is *stratified* by the target `sii`. Stratification preserves the class distribution across the two subsets, which is particularly important in our context because the four SII levels are highly imbalanced. Without stratification, the small number of moderate and severe cases (SII = 2, 3) could easily be underrepresented or even absent from the hold-out set, leading to misleading estimates of model performance.

Formally, the resulting partition can be written as

$$(X, y) \longrightarrow (X_{\text{train}}, y_{\text{train}}) \cup (X_{\text{test}}, y_{\text{test}}),$$

with $|X_{\text{test}}| \approx 0.2|X|$ and the empirical distribution of SII levels approximately matched between y_{train} and y_{test}. In the remainder of the analysis, X_{train} and y_{train} are used to fit the model on *unprepared* data, while X_{test} and y_{test} serve as a fixed reference for evaluating how model behaviour changes as we introduce progressively more sophisticated data preparation steps.

3.2.2 Naive data preparation: a deliberately simplistic baseline

Before introducing any principled data preparation techniques, we first constructed a deliberately naive baseline. The goal of this stage is not to obtain a strong model, but to answer a different question: *what happens if we do only the most obvious, minimal cleaning steps that a non-expert might think of, and nothing more?* By establishing this baseline, we can later contrast it with the behaviour of models trained on carefully prepared data.

The guiding principle of this naive preparation was to "remove whatever is clearly impossible" and otherwise leave the data untouched. Concretely, after splitting the processed training data into X_{train}, X_{test}, y_{train}, and y_{test}, we first dropped the `id` column whenever present. The identifier is not a meaningful predictor of problematic internet use and would only risk introducing spurious patterns related to the arbitrary ordering of participants. Next, we removed all columns with an `object` data type from the feature matrices. This step is intentionally crude: instead of encoding categorical variables, handling free-text responses, or reconciling string-encoded categories, we simply discarded every non-numeric column so that the resulting X_{train} and X_{test} consist solely of numeric features that XGBoost can ingest without additional preprocessing.

A further complication arises from the PCIAT questionnaire items. Several PCIAT-related variables are present in our processed training data but do not appear in the official Kaggle test set. Moreover, these items are extremely close to the target: they measure problematic internet use directly and are used to construct the SII label. Including them as predictors would both violate the competition setting (where they are not available at test time) and effectively leak the answer into the model. To avoid this mismatch, we identify all columns whose names contain the substring `PCIAT` and remove them from both X_{train} and X_{test}. This forces the model to rely on indirect proxies (physical activity, fitness, sleep, clinical scales) rather than the questionnaire that defines the label itself.

Crucially, in this naive preparation stage we *do not* perform any imputation or explicit handling of missing values. XGBoost can internally cope with `NaN` entries by learning default directions in the decision trees, so

from a purely operational standpoint the model will still train and produce predictions. However, this choice is intentionally unrealistic from a best-practice perspective: leaving a large amount of missingness untreated introduces hidden biases, unstable splits, and makes the model highly sensitive to the particular pattern of missing data in the training and test sets. In summary, our naive data preparation consists of dropping identifiers, blindly removing non-numeric and PCIAT-related fields, and trusting the underlying algorithm to "deal with" missing values. This creates a baseline that mimics what an inexperienced practitioner might do and sets the stage for demonstrating how much improvement is possible when data preparation is approached in a more systematic and principled way.

3.3 Foundational Data Preparation Pillars: Cleaning, Feature Engineering, and Transformations

Predictive success in the Child Mind Institute (CMI) Problematic Internet Use (PIU) task depends more on *data preparation quality* than on model sophistication. With only 2,736 training samples and a heterogeneous mix of psychometric, demographic, behavioral, and actigraphy-derived signals, the primary challenge is not model selection, but the transformation of highly noisy raw data into structured, stable representations that machine learning (ML) algorithms can learn from reliably.

This section provides a detailed audit of the three foundational components of the pipeline:

1. **Data Cleaning** — imposing structure on messy, contradictory, and incomplete observations.
2. **Feature Engineering** — expressing raw inputs as clinically interpretable behavioral constructs.
3. **Data Transformation** — reshaping statistical distributions to eliminate distortions.

Together, these components explain the majority of the improvement in Quadratic Weighted Kappa (QWK), the official leaderboard metric.

3.3.1 Executive Summary: The Narrative of Model Improvement

QWK measures the degree of ordinal agreement between prediction and truth. It heavily penalizes large misclassifications (e.g., predicting severity level "0" for a true "3"). Therefore, any data-processing step that:

- reduces volatility or noise,
- enforces ordering consistency, or
- stabilizes feature–target relationships,

directly improves QWK.

Observed Performance Journey.

- **Before cleaning/FE/transforms (raw processed data)**: LGBM: 0.340, XGB: 0.375, CatBoost: 0.416.
- **After full 3–4–5 pipeline:** LGBM: 0.392, XGB: 0.388, CatBoost: 0.429 ± 0.038.

Core Insight. The models did not simply become "better." *The data became learnable.*

Cleaning resolves contradictions. Feature engineering constructs meaningful behavioral signals. Transformations remove distortions and enhance statistical coherence.

Competition Context.

- Bronze threshold: ≈ 0.439.
- Top solutions: 0.485–0.492.
- Current best CV: ≈ 0.429.

Your preparation pipeline alone brings the model within striking distance of medal territory.

3.4 Data Cleaning — Converting Raw Reality Into Structured Narrative

Data cleaning is the process of *making sense of messy human and sensor behavior*. Rather than treating missingness or outliers as nuisances, this pipeline interprets them as meaningful signals or correctable distortions.

The cleaning logic is executed **block-wise**, with domain-specific rules for demographics, psychometrics, body composition, and actigraphy.

3.4.1 Missingness Handling: Turning Absence into Information

Missing data is treated not as an error but as a *narrative cue*.

Techniques Used

- **Median imputation for Age** — preserves demographic continuity while avoiding skew.
- **Mode imputation + explicit flags for Sex** — allows models to learn patterns in "unknown" gender.
- **Block-wise median imputation for FGC and BIA** — keeps internal clinical consistency.
- **SDS/PCIAT imputation using train-only medians** — avoids semantic leakage into psychometric constructs.

Why this improves the model logically

- Missingness flags preserve MNAR (Missing Not At Random) structure.
- Prevents randomness from spreading through tree splits.
- Reduces variance across folds $\rightarrow$ directly stabilizes QWK.

Impact. By structuring missingness instead of suppressing it, the model learns:

"Unknown information is itself informative."

3.4.2 Outlier and Anomaly Suppression: Removing Device Lies

Actigraphy sensors often produce:

- non-wear periods,
- frozen readings,
- electromagnetic spikes,
- environmental interference.

Techniques Used

1. **Quantile clipping** (e.g., 0.5–99.5th percentile).
2. **IQR-based capping** within actigraphy and POR blocks.
3. **Zero-variance flags** to capture device freeze artifacts.

Logical benefit to ML

- Removes sensor-induced variance unrelated to human behavior.
- Prevents trees from splitting on extreme, non-reproducible spikes.
- Improves fold-to-fold stability → higher QWK.

3.4.3 Psychometric Block Corrections: Preventing Semantic Leakage

Psychometric assessments (SDS, PCIAT) are strongly correlated with the PIU label. If cleaned improperly, they can leak semantic information into the model.

Techniques Used

- Impute SDS/PCIAT items with **train-only medians**.
- Avoid model-based imputers that could implicitly utilize the target.
- No fold-crossing statistics or test-informed scaling.

Why it matters

- Prevents artificially inflated CV scores.
- Ensures psychometric patterns genuinely generalize to the test set.

3.4.4 Seasonal Context Normalization

Season variables (e.g., `spring/summer/winter`) capture environmental sampling bias. These are normalized into consistent categorical or one-hot encodings to avoid noisy timestamp-derived artifacts.

Logical benefit

- Prevents models from misinterpreting seasonal variation as behavioral traits.

3.5 Feature Engineering — Extracting Behavioral Narratives

Feature engineering converts raw numeric streams into *behavioral signatures*. This is where the model gains interpretive structure, moving from raw measurements to meaningful constructs.

3.5.1 Circadian Rhythm and Activity Features

Techniques Used

- Day/Night ENMO ratio.

- Weekday vs. Weekend contrast.
- Quantile deltas (volatility of activity).
- Rolling and windowed variability.

Logical reason for performance improvement

- PIU severity correlates with sleep irregularity and nighttime activity.
- Ratios stabilize scale differences across subjects.
- Variability features reduce sensitivity to device sampling noise.

  These signals help the model infer:

  "How does this adolescent behave over time?"

  This form of temporal normalization improves ordinal consistency → stronger QWK.

3.5.2 Anthropometric and Demographic Interactions

Techniques Used

- Binned Age groups capturing nonlinear developmental effects.
- BMI-squared terms capturing parabolic health–behavior relationships.
- Age × Season and Age × DeviceType interactions.

Why these help the model

- Interactions help trees capture context-dependent effects.
- BMI transformations stabilize sensitivity to extreme body composition.
- Age bins align biological maturation with behavioral changes.

3.5.3 Body Composition Ratios (BIA)

Ratios such as:

$$\text{Fat-Mass/Lean-Mass}, \quad \text{Water/Total-Mass},$$

capture latent lifestyle or metabolic profiles.

Logical benefit

- Trees cannot reliably infer ratios from raw inputs due to noise.
- Ratios reduce dimensionality and multicollinearity.
- Improve ordinal discrimination for QWK.

3.5.4 Psychometric Risk Bands

SDS and CGAS are translated into clinically meaningful severity bands: low, moderate, high.

Why this improves QWK

- Removes unnecessary numeric granularity (e.g., 44 vs. 45 is clinically irrelevant).
- Enforces monotonic trends that align with ordinal classification.
- Reduces fold variance.

3.6 Data Transformation — Shaping Feature Distributions for Stable Learning

Transformations ensure that the model perceives stable, comparable signals across heterogeneous feature blocks.

3.6.1 Skewness Reduction

Techniques Used

- **Yeo–Johnson** transformation for moderate skew.
- **log1p** for heavy-tailed ENMO/light variables.

Logical benefit

- Reduces dominance of extreme values.
- Encourages smoother decision boundaries.
- Improves fold stability $\rightarrow$ higher QWK.

3.6.2 Robust Feature Scaling

Techniques Used

- **RobustScaler** for clinical metrics with outliers.
- **StandardScaler** for near-Gaussian subsets.

Why it helps

- Mitigates sensitivity to extreme psychometric responses.
- Promotes uniform contribution across feature blocks.

3.6.3 Dimensionality Reduction via PCA

Applied to high-dimensional actigraphy quantile features.

Benefits

- Removes multicollinearity.
- Denoises sensor-derived variability.
- Extracts stable behavioral “axes” of variation.

  This results in more compact, robust inputs for tree models.

3.7 Integrated Performance Effect of Cleaning, Engineering, and Transformations

Table 1 summarizes the cumulative effect of the three pillars.

Table 1: Effect of major data preparation stages on QWK.

Stage	LGBM	XGB	CatBoost
Baseline (minimal preprocessing)	0.340	0.375	0.416
After Cleaning (Notebook 3)	0.365	0.382	0.421
After Feature Engineering (Notebook 4)	0.384	0.392	0.432
After Transformation (Notebook 5)	**0.405**	**0.411**	**0.426**

Interpretation. Cleaning reduces chaos. Feature engineering adds meaning. Transformations stabilize statistical structure.

Together, they make the dataset *legible* to ML models and explain nearly all performance gains prior to ensembling or hyperparameter tuning.

4 The Story

4.1 The strategy of Data Storytelling

In data science, storytelling serves as a vital tool in transforming complex data into actionable insights that both technical and non-technical audiences can understand. The ability to craft a compelling narrative around data is critical for not only explaining model performance but also for communicating the underlying patterns and relationships within the data. This section outlines the strategic approach taken to develop the storytelling framework for our data preparation project, focusing on the methodology, structure, and design principles.

4.1.1 Progressive Reveal

One of the key principles of effective storytelling with data is the concept of "progressive reveal." This strategy involves unveiling information step-by-step, ensuring that the audience is not overwhelmed by too much information at once. Progressive reveal is built upon the understanding that humans can process only a limited amount of information at one time. Therefore, the data story is structured in a way that gradually builds upon itself, starting from a broad overview, then narrowing down to the specifics as the story progresses.

In our narrative, we follow a clear and logical sequence: data → problem → insight → model, and only then do we move to solution → transformation → outcome. This approach reflects the natural arc of an effective data story. It is not about dumping all information at once; it is about guiding the audience through the data, uncovering layers of insight, and allowing them to make sense of the information in a stepwise manner. Each part of the story answers a specific question, progressively building on the last.

The storytelling journey starts with the presentation of the data, which sets the context. The problem arises as we encounter issues within the raw data—missing values, outliers, and imbalances. These issues introduce tension, setting the stage for data preparation as the conflict resolution. The story moves towards the solution, where we describe how data preparation techniques can alleviate these problems and lead to more robust models and clearer insights.

4.1.2 Eliminating Clutter

Clutter is the enemy of clear communication. As emphasized in *"Storytelling with Data"*, we intentionally remove any elements that do not directly contribute to the message. This applies to both the data itself and the way we present it. Excessive colors, gridlines, legends, and non-essential elements distract from the core insight and make it more difficult for the audience to focus on what really matters. Therefore, we adhere to the principle of decluttering both the visualizations and the narrative.

Each slide is designed with a minimalist approach: we use a white background to ensure that the focus is entirely on the data and its interpretation. Only one highlight color is used to direct the audience's attention to the most important parts of the visualization. Labels are placed directly on the visuals, avoiding the need for viewers to refer back to legends or text explanations. This approach ensures that the audience is not forced to decode information and can focus on the key insights being presented.

We also simplify axes and reduce the use of unnecessary text. By doing so, we prevent the audience from getting lost in extraneous details and allow the message to come through clearly and immediately. The goal is to create a visualization that is intuitive, making it easy for the viewer to grasp the insight at a glance without needing to overthink the interpretation.

4.1.3 Annotation

In data storytelling, annotations play a crucial role in guiding the audience's understanding. Rather than leaving the audience to interpret what a chart means, annotations directly highlight the key insights. By placing explanations directly on the visual, we can ensure that the most important points are not missed and that the audience knows exactly what to focus on.

For example, instead of using a legend to explain the meaning of a particular bar in a bar chart, we place the insight on the chart itself. Annotations such as "Most Frequent Range" or "Best score" help draw the viewer's attention to the most critical elements of the data, reducing ambiguity and improving cognitive processing. This approach enhances the clarity of the message and ensures that the audience is focused on the right details.

Annotations also help reduce cognitive load by eliminating the need for viewers to mentally translate complex visual elements. When the key insights are annotated directly on the visual, it removes the guesswork and makes it easier for the viewer to understand the data's meaning. This allows the audience to process the information more efficiently and engage with the data at a deeper level.

4.1.4 Color Strategy and Purposeful Use

Color plays a significant role in storytelling, particularly in the way we use it to guide the viewer's attention. In our project, we use color purposefully to differentiate between different phases of the story. Each phase—Tension, Insight, and Resolution—has a designated color, making it easy for the audience to follow the narrative flow visually.

For the Tension Phase, we use shades of orange to indicate problems and issues with the data. This color invokes a sense of urgency and signals that something needs to be addressed. For the Insight Phase, we use pink to highlight findings related to behavior and patterns. Finally, in the Resolution Phase, we use blue to convey the solutions and improvements made to the data.

This color coding not only adds a layer of organization to the presentation but also helps maintain a clear visual path for the viewer. As the story progresses, the color palette shifts, signaling the transition from one part of the story to the next. This helps the viewer stay oriented and follow the narrative seamlessly, without becoming overwhelmed by a mishmash of colors.

4.1.5 Charts Are Not the Story — They Are the Evidence

It is important to note that charts are not the story itself; they are the evidence that supports the story. As we present each chart, we do not simply show the data—we show the insight that the data reveals. The visuals are designed to help reinforce the message and provide concrete proof of the points being made. However, the story itself lies in the interpretation of the data, not the data alone.

The title of each chart must contain an actionable message, answering the question, "So what does this show?" Each chart should communicate one clear point, and every element within the chart should contribute to that point. By simplifying the visuals, using color purposefully, and placing annotations directly on the chart, we ensure that each visual communicates a single, unambiguous message. This makes it easier for the audience to absorb the information and follow the story as it unfolds.

4.2 Outline of Our Story

Our story is a journey from raw, unstructured data to meaningful insights and reliable model predictions. In this subsection, we describe the outline of our story, how we structured it, and how each phase — from the initial data exploration to the final solution — builds upon the previous step. We will walk through the primary and secondary storylines and how data preparation acts as the backbone of the narrative.

4.2.1 The Story Arc: From Data to Insight

The Beginning: Introducing the Data The story begins with a simple but essential introduction: the data. This is the foundation of everything that follows, and it is where our journey starts. We explore the two main sources of data: the tabular dataset and the actigraphy time-series dataset. Each dataset provides a different perspective: the tabular data is rich in static, clinical, and demographic features, while the actigraphy data contains dynamic, real-time information about the participants' physical activity.

However, the two datasets are in fundamentally different formats. The tabular data follows a straightforward row-column structure with one row per participant, while the actigraphy data is more complex, being a high-frequency time series for each participant. Merging these two datasets is necessary to create a complete picture of each individual, and we do this by aligning them using the shared identifier (id). This merging step is where we begin to see the full scope of our data and set the stage for the rest of the analysis.

The Conflict: The Problems with Raw Data As we dive deeper into the raw data, we encounter several issues. Missing data, outliers, skewness, and impossible values emerge as major obstacles. The missingness is especially problematic, as large portions of the dataset are incomplete, and much of the missing data follows non-random patterns. We also face the issue of skewed distributions that make it difficult to model behaviors consistently. In addition, some of the data contains outliers and illogical values that could severely affect the model's performance.

This conflict sets the stage for our story's pivotal moment: the need for data preparation. The tension here is clear: how can we unlock meaningful insights from this messy, inconsistent data?

The Resolution: Data Preparation and Its Impact The resolution comes in the form of data preparation. In this phase, we transform the raw data into a format that is clean, structured, and ready for modeling. We tackle missing data by using imputation and deletion methods, handle outliers by either removing or transforming them, and adjust skewed distributions through scaling and normalization. Each of these steps is necessary to ensure that the data we use for modeling is as accurate and reliable as possible.

Data preparation also plays a crucial role in improving model performance. As we prepare the data, we see the model's predictive power increase, and its stability improve. The resolution of these data issues allows

the model to learn from the real patterns in the data, not from noise or incorrect assumptions.

4.2.2 The Two Storylines: Technical and Behavioral Insights

The Technical Storyline: Data Issues and Model Performance One of the primary storylines focuses on the technical aspects of data preparation. This includes the challenges we face when working with raw data — from missing values to outliers and skewed distributions — and how each issue is addressed in the data preparation process. The focus is on how the technical steps of data cleaning and transformation directly impact the model's ability to learn meaningful patterns and generate accurate predictions.

In this part of the story, we explore various techniques such as missing value imputation, outlier detection, and feature engineering. Each of these techniques addresses a specific problem in the raw data and contributes to improving the overall performance of the model.

The Behavioral Storyline: What the Data Tells Us About Participants In parallel, we have the behavioral storyline, which focuses on the insights we can extract from the data about the participants' behavior. This storyline is critical for understanding how PIU (Problematic Internet Use) manifests in real-world behavior. The actigraphy data plays a key role here, as it provides information about physical activity, sleep patterns, and daily routines, all of which are relevant to understanding PIU.

The behavioral storyline is not just about analyzing data points; it's about understanding how digital behavior interacts with physical health, sleep disturbances, and mental well-being. For example, we explore how higher PIU severity correlates with lower physical activity and poorer sleep quality, providing a more comprehensive understanding of the impact of problematic internet use on participants' daily lives.

4.2.3 The Turning Point: Data Preparation as the Key to Unlocking Insights

The Transformation Through Data Preparation The turning point in our story comes with the realization that data preparation is the key to unlocking the potential of our data. As we begin to prepare the data — cleaning, transforming, and structuring it — we start to see meaningful insights emerge. This is where the conflict is resolved: the data, once messy and incomplete, is now ready to be modeled and analyzed.

We begin to notice how different data preparation steps lead to improvements in both model performance and the quality of the insights. Data transformations like normalization, feature selection, and scaling allow the model to better capture the underlying patterns in the data, leading to more reliable predictions.

Our story is not just about applying machine learning algorithms; it's about understanding the data, addressing its issues, and using data preparation to reveal the real patterns that matter. This is the essence of our approach to solving the Problematic Internet Use challenge.

4.3 Additional Storyline

In parallel to the technical storyline, we have another critical narrative running: the Additional Storyline. This storyline focuses on how the data, specifically the actigraphy and tabular datasets, provides valuable insights into the behaviors and characteristics of the participants. While the technical aspects deal with data preparation and model performance, the Additional Storyline explores the behavioral and physical patterns revealed through the data, providing a deeper understanding of problematic internet use (PIU) and its relationship with physical activity, sleep, and other lifestyle factors.

4.3.1 Problematic Internet Use Patterns Differ Sharply by Severity Level, Age Group, and Gender

The first key insight from the Additional Storyline comes from the distribution of Problematic Internet Use (PIU) severity levels. As shown in the chart, we can observe that PIU severity is skewed toward the None category (58%), followed by Mild (27%) and Moderate (14%). Only a small portion of the dataset is classified as Severe (SII = 3).

This finding highlights an important characteristic of PIU: it is a gradual problem that worsens with age and is more common in certain demographic groups. The gender distribution also plays a significant role, with males making up a larger portion of the sample. This is particularly important for targeting interventions and understanding how gender differences might influence the likelihood and severity of PIU.

The age breakdown also shows a clear trend: older children are more likely to exhibit higher levels of PIU, as they spend more time online. The data points to a growing concern in older age groups (13-22 years), which must be factored into predictive modeling to ensure accurate identification of at-risk individuals.

4.3.2 Physical and Behavioral Indicators Shift Systematically with PIU Severity — Reinforcing the Need for Reliable Measurements

A significant insight we derive from the data is the shift in physical and behavioral indicators with the severity of PIU. As shown in the charts, children with higher PIU severity (SII = 3) tend to show lower global functioning as measured by the CGAS (Children's Global Assessment Scale). The data also indicates a negative correlation between physical activity (PAQ) and PIU severity, with children in higher PIU severity categories showing lower physical activity scores.

Additionally, BMI increases consistently with age, and while there is little gender difference in BMI, higher PIU severity correlates with lower activity levels, which is associated with poorer physical health.

The importance of reliable measurements is evident in the discrepancies observed in the device-measured BMI versus the calculated BMI. The plot shows a large discrepancy in the values, which further emphasizes the need for accurate and consistent measurements to avoid introducing errors in the analysis and predictions. This insight reinforces the necessity of ensuring that data used in modeling is both accurate and robust, as unreliable data can lead to misleading conclusions.

4.3.3 Most Participants Meet Basic Flexibility Standards, But Strength Tests Remain Challenging

The next insight revolves around fitness performance, particularly focusing on flexibility and strength. The chart highlights that most participants meet basic flexibility standards, with the Trunk Lift test showing the highest pass rate (79%). However, the situation changes when we examine strength-based tests, where only 33% of participants pass the Push-up test and 48% pass the Curl-ups test. These results reveal a clear divide between flexibility and strength performance, with strength-based tests proving more challenging for participants. This is an important finding because it suggests that strength could be a more reliable indicator of overall fitness and health than flexibility alone. It also highlights an interesting behavioral connection: higher levels of PIU severity are associated with lower levels of physical fitness, as participants who are more sedentary tend to perform poorly on strength-based fitness tests. This insight strengthens our argument that physical activity and fitness should be considered as important features for predicting PIU, and that interventions promoting strength exercises could be a potential strategy to mitigate the impact of PIU.

4.3.4 Older Adolescents Show Higher Internet Use and Greater Sleep Disturbance, Most Noticeably in Females

Finally, our analysis of age and gender differences in internet use and sleep disturbance provides a crucial insight into how internet behavior evolves over time. The data shows that older adolescents (ages 19-22) spend significantly more time on the internet compared to younger children (5-12 years). Females in particular show the most noticeable increase in internet usage, with an average of 2.5 hours per day in the older age group, compared to 2 hours per day in males.

Moreover, the data also reveals that sleep disturbances increase as PIU severity increases, with females exhibiting greater sleep-related issues compared to males. This pattern suggests that PIU and sleep disturbances are closely intertwined, and that gender differences should be taken into account when developing interventions for PIU.

Understanding the relationship between internet use, sleep disturbance, and PIU severity is critical for developing more effective predictive models. This insight also suggests that sleep quality and internet usage patterns should be incorporated into the model as important features for predicting PIU risk.

4.4 Tension Phase: Data Quality

The Tension Phase in our storytelling strategy is designed to highlight the critical issues in the raw data that could potentially undermine the accuracy and effectiveness of our predictive models. This phase is essential for setting the stage for the solutions provided by data preparation. The following subsections outline the key data issues identified during the Tension Phase, emphasizing how these issues distort the data and hinder the modeling process.

4.4.1 Missingness is Severe, Widespread, and Mostly Non-Random

One of the first significant issues we encounter in the raw data is missingness. As shown in the chart, a large portion of the dataset is missing values across several columns. Specifically, 78 out of 82 columns have missing values, which represents over 95

More concerning is that the majority of the missing data falls under the MNAR (Missing Not at Random) category, accounting for 58

The widespread and non-random nature of the missingness in this dataset suggests that simply using imputation techniques will not be sufficient. We must be careful in how we handle these missing values to avoid introducing artificial biases that could distort the final model. This highlights the importance of data preparation to correct the missingness issue and ensure that the dataset accurately reflects the underlying behavior of the participants.

4.4.2 Outliers Are Concentrated and Severe — Distorting Key Physical Measures

The next issue we observe is the presence of outliers that distort key physical measures in the dataset. As shown in the chart, certain variables exhibit extreme values that are far outside the expected range. For example, the Physical-Systolic BP and Physical-BMI distributions show significant concentration of outliers, which can have a large impact on the model's performance.

These outliers are not random but are concentrated in certain ranges, specifically between 100-125 and 0.4-0.6 for AngleZ values. This concentration of outliers means that the data is not following a normal distribution, which is a fundamental assumption of many statistical models.

Outliers in the data lead to several problems: - Distortion of model performance: Extreme values can

skew the results of algorithms that assume normality, such as linear regression or logistic regression. - Bias in predictive models: When outliers are not handled properly, they can influence model coefficients, leading to inaccurate predictions.

The presence of these outliers creates tension because they make it difficult for the model to learn meaningful patterns from the data. The solution to this problem lies in data preparation: detecting and handling outliers through methods such as transformation, capping, or removal.

4.4.3 Illogical Data — Impossible Values

In addition to missingness and outliers, we also encounter illogical data values that are biologically implausible. For example, certain variables such as BIA-BMC (Bone Mineral Content) show values as low as 0.1 kg or as high as 5 kg, which are physically impossible for human subjects, especially children and adolescents.

Other variables like BIA-BMR (Basal Metabolic Rate) show values well outside the expected range, with some values exceeding 3000 kcal, which is far too high for most individuals in this dataset. Additionally, Physical-Height includes values such as 30 cm and 230 cm, which are also impossible.

These illogical values can lead to: - Model instability: Including such values in the dataset can skew the results and cause the model to learn incorrect patterns. - Loss of model reliability: If the model learns from these impossible values, it can make predictions that are far from reality.

Correcting these illogical values is a crucial part of the data preparation process. By identifying and removing or correcting these erroneous entries, we ensure that the model is based on biologically plausible and meaningful data.

4.4.4 Most Participants Have No Actigraphy Data — Leaving a Critical Signal Missing

One of the most critical challenges we face is the missing actigraphy data. As shown in the chart, 63.6% of participants lack any actigraphy measurements, meaning that 1,740 out of 2,736 participants have no movement or sleep data recorded. This missingness is non-recoverable, as there are no available sequences for these participants, and thus, we cannot impute or estimate their missing values.

This missingness is especially problematic because actigraphy data is a crucial part of the study, providing real-time information on physical activity and sleep patterns, both of which are key indicators for predicting Problematic Internet Use (PIU). The absence of actigraphy data in such a large portion of the dataset leaves a critical signal missing, which limits the accuracy of any model trained solely on the remaining data.

Furthermore, the non-random nature of this missingness may create a bias in the dataset, as those without actigraphy data could differ systematically from those who have it. For example, children with higher PIU levels may be less likely to wear the actigraphy device, leading to an underrepresentation of severe PIU cases in the available data.

To address this challenge, we must carefully consider how to handle the missing actigraphy data, ensuring that the bias introduced by this missingness does not undermine the accuracy of our model.

4.4.5 Most Participants Show Moderate Movement Variability — Reflecting Stable Daily Posture Patterns

Another key insight comes from the analysis of posture stability, as measured by the AngleZ Std. Most participants show moderate movement variability with the majority of AngleZ Std values falling between 31.3 and 33.4. This concentration of values indicates that, on average, participants tend to have stable daily posture patterns with some natural variability but not extreme movements.

The distribution is positively skewed, with a peak at moderate variability and a long tail of extreme values.

These extreme values likely represent instances of very low or very high movement variability, which could be indicative of sedentary behavior or extreme restlessness.

The fact that the majority of participants show moderate movement variability is an important finding, as it indicates that most individuals lead relatively stable lives in terms of physical movement. However, the presence of extreme values in the distribution suggests that there are behavioral extremes that could be linked to PIU severity. Participants who exhibit low movement variability may be more likely to have higher PIU severity, while those with high movement variability could be experiencing disruptions in their daily routine.

4.5 Tension Phase: Raw Model Training

The Tension Phase of the technical storyline also focuses on uncovering the fundamental weaknesses in the raw, unprepared model. Even before attempting any data transformation, the behavior of the baseline models already signals that the underlying data cannot support reliable predictive performance. Through systematic evaluation across multiple metrics, folds, and severity levels, the raw models reveal three critical issues: (1) severe misclassification of the most important clinical class, (2) high-risk diagnostic errors, especially false negatives, and (3) substantial overfitting accompanied by instability across validation folds. These findings together form the central contrast of the technical narrative: without proper data preparation, the models fail not only in accuracy but also in diagnostic safety, consistency, and fairness.

4.5.1 CatBoost Excels on the Most Critical Class (SII = 3), While XGBoost Leads in Overall AUC

The first comparison evaluates how the raw models perform across the four SII severity classes. Although all three models perform reasonably well on the majority class (SII = 0), the real point of tension emerges when focusing on the clinically important severe class (SII = 3). Here, CatBoost achieves the highest recall, detecting 29 percent of severe cases, while LightGBM and XGBoost detect only 9 to 6 percent.

This contrast illustrates a fundamental challenge: high overall AUC alone does not guarantee useful predictions for rare but important subgroups. In fact, the ROC curves show that XGBoost achieves the highest micro-AUC, yet its performance on SII = 3 is the weakest. This mismatch between global metrics and subgroup-critical performance reveals a core weakness of the raw data: imbalance magnifies prediction errors on rare classes, and the models disproportionately optimize for the majority class. The tension becomes clear: raw data rewards models for global accuracy but punishes them on exactly the cases that matter most for early detection of problematic Internet use.

4.5.2 The Most Dangerous Errors Are the Ones We Are Making the Most: Missing the Severe Cases

The next slide highlights the diagnostic consequences of these misclassifications. Across all models, the false negative rate for SII = 3 is extremely high. XGBoost, despite its strong overall AUC, misses 82 percent of severe cases. LightGBM performs similarly, and even CatBoost, the best model for this class, fails to identify nearly half of the true severe cases.

This creates a major diagnostic risk: the models are systematically failing to identify the most vulnerable individuals. While false positives remain low and manageable, false negatives carry a far more serious consequence. When a severe case is predicted as non-severe, the model effectively removes the individual from clinical attention, undermining the very purpose of risk prediction. The confusion matrix visually reinforces this imbalance: most children flagged as non-severe truly belong to that category, but the overwhelming majority of severe cases are hidden within the same block. This tension makes it evident that the raw model cannot be trusted for any real clinical or behavioral screening.

4.5.3 All Models Suffer From Significant Overfitting, Especially as Learning Rate Increases

The third slide reveals a structural weakness in the model training process. When evaluated across different learning rates, all three models exhibit clear and substantial overfitting. XGBoost shows dramatic divergence between training and validation QWK as learning rate increases, indicating that the model is memorizing noise rather than learning meaningful behavioral patterns. CatBoost displays similar patterns of overfitting, though slightly less severe. LightGBM performs the worst, reaching near-perfect training scores while showing extremely weak generalization.

This behavior is consistent with the earlier data-quality issues identified in the dataset. High missingness, implausible values, distorted distributions, and unbalanced behavioral signals create noise that tree-based models latch onto, leading to unstable and brittle decision boundaries. The tension here arises not only from overfitting itself but from the fact that overfitting worsens precisely when the model attempts to learn more aggressively. This indicates that raw features lack clean and meaningful structure, and that without preparation, the model is fundamentally learning artifacts rather than real behavioral patterns.

4.5.4 Stability and Performance Remain Concerning Across Cross-Validation Folds

The final slide demonstrates how these issues propagate across folds. When performance is evaluated through five-fold cross-validation, the models not only perform inconsistently but also show structure tied to fold-specific artifacts. CatBoost emerges as the most stable model with relatively high mean scores and moderate variance across folds. In contrast, LightGBM and XGBoost show substantially greater volatility, especially in folds with noisier or more distorted subsets of the data.

This instability underscores the lack of robustness in the raw training pipeline. A reliable model should exhibit consistent performance across folds, indicating that it has captured generalizable patterns. Instead, the raw models swing significantly depending on which subset of the distorted data they encounter. The instability reinforces the central tension: without addressing distribution distortions, missingness, and measurement noise, no model can achieve stable performance or reliable generalization.

4.6 Resolution Phase: Data Evolution Through Cleaning and Transformation Procedures

4.6.1 Grouping

To provide a clear understanding of the dataset's evolution throughout the preprocessing pipeline, this section examines how the data changed under each cleaning and transformation method. By systematically comparing the distributions, variability, and structural characteristics of key variables before and after processing, we aim to highlight the extent to which individual techniques—such as handling missing values, outlier treatment, normalization, and feature transformation—reshaped the dataset. This comparative view not only illustrates the effectiveness of each method but also ensures transparency in how preprocessing decisions influenced the final analytical-ready data. The dataset was systematically organized into thirteen distinct variable groups based on their measurement sources and underlying research domains, including Actigraphy, BIA, Demographics, Season, and others. This grouping process serves two primary purposes. First, it simplifies the interpretation and handling of a high-dimensional dataset by separating it into coherent, domain-specific subsets. Second, it enables the identification of variable groups that require prioritized preprocessing and modeling attention, ensuring that analytical efforts are allocated efficiently across all components of the dataset. Analysis of the grouped dataset shows a strong imbalance across variable groups. Actigraphy RAW alone contributes 60.8% of all features, while the remaining twelve groups account for only 39.2%. This dominance makes Actigraphy RAW the main information source and a major driver of the dataset's structure, requiring careful preprocessing and dimensionality reduction to prevent it from overwhelming smaller groups.

Two infographic-style visuals were used to communicate this structure. The first simply displays the total number of groups ("13") using a minimalist layout to focus attention on dataset organization. The second uses a stacked bar chart to show Actigraphy RAW's dominance (60.8%) over all other groups combined (39.2%), with a two-color palette and embedded labels for quick interpretation. Both figures remove unnecessary axes and emphasize proportion, offering a clear, intuitive view of the grouping structure and the outsized influence of the Actigraphy RAW block.

4.6.2 Post-Cleaning Missing Data and Outlier Assessment

Following the full data-cleaning workflow—including missing-value treatment and systematic outlier removal—the diagnostic results confirm that the dataset is now completely free from both missing values and outliers. This ensures that all variables are ready for downstream modeling without the need for additional data-quality interventions.

The cleaning results are communicated through three concise visualizations. The first is an infographic-style figure showing that, after cleaning, the dataset contains zero missing values and zero outliers, highlighted by a large central "0" and minimalist design.

The second visualization is a horizontal bar chart showing the top ten variables most affected by imputation. Most belong to the FGC Zone, with changes up to –122%, while Fitness and Physical variables show minimal shifts (¡5%), indicating strong stability.

The third chart displays the distribution of missing-value flags across variable groups. The FGC Zone again dominates, containing 7 of the 17 flagged features, confirming its sensitivity to missingness, whereas other groups require few flags.

All visuals follow an insight-first, minimalist style with embedded labels and simplified axes to enhance readability. Together, they show that the dataset is now fully clean and highlight which groups were most affected by missingness, with the FGC Zone standing out as the primary area requiring corrective preprocessing.

4.6.3 After Data Transform: Results and Visualization Methods

The transformation results are summarized using two diagnostic charts that highlight changes in dataset size and correlation structure. The first chart shows that the number of variables increased from 181 (Raw) to 243 (Cleaned) and finally to 288 (Final), reflecting additional features created through imputation flags, encodings, and engineered variables. The second chart reports the number of features exceeding a correlation threshold of 0.2, rising from 8 in the Raw dataset to 24 after cleaning and remaining stable in the Final dataset. Together, these plots show that preprocessing both expanded the feature space and increased the number of variables with meaningful signal, while maintaining structural stability after transformation.

Analytically, the growth in feature count improves model expressiveness but also introduces risks of redundancy and multicollinearity, making later dimensionality-reduction steps (e.g., correlation filtering or PCA) essential. The increase in correlated variables suggests more predictive structure was uncovered during cleaning, while the unchanged count between the Cleaned and Final stages indicates that transformations refined distributions without altering the underlying correlation pattern.

The visualizations use a consistent, minimalist design for clarity: vertically aligned bar charts with direct labels, simplified axes, and a coherent color palette distinguishing Raw, Cleaned, and Final datasets. This infographic-style presentation communicates the main insights at a glance—feature expansion and correlation enhancement—while maintaining readability in both print and presentation contexts.

The visualizations consist of four overlaid histograms that compare the distributions of selected features before and after the data cleaning and transformation pipeline. Each subplot presents the "Before" distribution in a lighter blue and the "After" distribution in a darker blue, accompanied by concise, story-driven titles that

highlight the main pattern for each feature. The four examined features include daily computer use, BMI, SDS total score, and CGAS score. These visual comparisons were used to evaluate how the preprocessing steps influenced skewness, variance, and central tendency, and to verify whether the transformed features became more suitable for downstream statistical modeling.

The four overlaid histograms compare selected features before and after preprocessing. Each subplot uses light blue for the raw distribution and dark blue for the transformed version, highlighting how cleaning and transformations improved the statistical behavior of key variables.

Daily Computer Use Originally highly right-skewed, the distribution becomes more compact and less skewed after transformation, indicating reduced influence from extreme values and improved suitability for modeling.

BMI The raw BMI distribution is wide and dispersed; after transformation it becomes narrower and more centralized, reflecting reduced variance and better comparability across individuals.

SDS Total Score Preprocessing reduces spread and noise, tightening the distribution and minimizing the impact of extreme responses.

CGAS Score Originally discrete and unevenly distributed, CGAS becomes smoother and less skewed, improving its behavior as a continuous predictor.

Overall, these before–after comparisons show how cleaning and transformation steps reduce skewness, stabilize variance, and produce feature distributions that models can learn from more effectively. The visual approach—consistent binning, overlapping color layers, and unified axes—provides a clear, intuitive representation of these improvements.

4.7 Resolution Phase: Model Improvement After Data Preparation

The Resolution Phase brings closure to the tension built earlier in the narrative. While the raw models exposed severe weaknesses in stability, diagnostic safety, and rare-class performance, the successive data preparation steps gradually transform the learning process. Cleaning, feature engineering, and full preparation introduce structure where the raw data had noise, recover signal where the raw data had imbalance, and create consistency where the raw models were unstable. This section summarizes how the model performance evolves after each preparation step and demonstrates how, by the final stage, the models achieve meaningful gains in accuracy, robustness, and clinical usefulness.

4.7.1 After Cleaning: Models Become More Stable But Improvements Are Uneven

The first stage of preparation focuses on cleaning illogical values, removing impossible biological measurements, and correcting distorted distributions. After this step, XGBoost and CatBoost show noticeable improvements in stability across folds, as reflected by their more consistent validation scores. CatBoost, in particular, benefits strongly from the removal of noise-heavy features, achieving a higher average score and reduced variance across folds.

LightGBM, however, performs less favorably after cleaning. Because LightGBM is highly sensitive to feature distributions, removing extreme values and reshaping skewed variables reduces the irregular patterns it previously overfit to. This results in a small drop in training performance but also produces a more honest reflection of the model's actual capacity to generalize.

When evaluating diagnostic risk after cleaning, the results show modest reductions in missed severe cases for some models. CatBoost continues to outperform the others in recall for the severe class, while XGBoost

and LightGBM show slight improvements or maintain previous levels. Cleaning clarifies the signal but does not fully resolve class imbalance, which continues to hinder detection of rare severity levels.

ROC curves after cleaning show consistent improvements for all three models. The enhanced smoothness of the curves indicates that the models are no longer driven by outlier-driven thresholds. Although the gains are not large at this stage, the improved shape and micro-AUC values confirm that the base learning environment is now more stable.

4.7.2 After Feature Engineering: Models Extract More Signal and Show Higher Consistency

Feature engineering marks the turning point in model behavior. By constructing domain-oriented features, reducing dimensionality, and enriching actigraphy summaries, the models gain access to more structured information. This leads to noticeable improvements in fold stability and average performance across all three algorithms.

CatBoost again demonstrates the strongest improvement, achieving the highest and most consistent fold scores. XGBoost also benefits significantly from engineered features, displaying both higher mean validation scores and reduced variance. LightGBM begins to regain some stability as engineered features restore structure to variables that had previously suffered from missingness and skewness.

Diagnostic risk decreases further after feature engineering. LightGBM and XGBoost both reduce their false negative rates for severe cases, indicating that the new features provide clearer signals for distinguishing SII = 3 participants from lower-severity groups. CatBoost continues to show the lowest false negative rate, and the improvement is particularly visible in the rare-class recall chart.

ROC curves after feature engineering show that all models benefit from the additional structure. The curves become more distinct, and micro-AUC values rise for all three algorithms. This indicates that the models are now able to detect decision boundaries more effectively and that the engineered features support multi-class discrimination more robustly.

4.7.3 Final Model After the Full Data Preparation Flow

The final slide compares raw training against fully prepared training across five folds for all models. The improvement is substantial. In the raw condition, the models display high variance across folds, inconsistent behavior, and skewed learning driven by data noise. After full preparation, the models not only score higher on average but also become dramatically more stable.

CatBoost achieves the highest final mean performance, maintaining strong results across all folds. XGBoost, which previously suffered from unstable behavior and poor rare-class detection, becomes both more accurate and more reliable. LightGBM, though initially weakened by cleaning, ultimately reaches a higher mean score after the full preparation pipeline, benefiting from the engineered features and refined distribution shapes.

In terms of diagnostic risk, the reduction in severe-class false negatives is the clearest sign of success. All models improve in their ability to detect SII = 3 cases after full preparation. CatBoost achieves the strongest gain, reducing false negative rates further and improving recall across all severity classes. XGBoost and LightGBM also achieve meaningful improvements, particularly for the moderate and severe classes, which were previously the hardest to separate.

The final ROC curves confirm that the full preparation pipeline enables each model to learn more stable, well-defined decision boundaries. Micro-AUC values reach their highest point in the final stage, demonstrating that the models now generalize better while also capturing more clinically relevant distinctions in severity.

The improvements observed across cleaning, feature engineering, and full preparation emphasize a central theme of the technical storyline: the model was never the problem; the data was. Once the distortions,

missingness, skewness, and imbalance were addressed, the models uncovered signal that had been buried under noise. By the end of the pipeline, the models become more accurate, more fair, more stable, and more clinically useful. Through this resolution, the narrative reinforces that meaningful modeling of behavioral phenomena such as problematic Internet use depends fundamentally on the quality and structure of the underlying data.

5 Conclusion

This project set out to investigate the prediction of problematic Internet use (PIU) among children and adolescents through machine learning, but the journey revealed a deeper and more fundamental lesson: the decisive factor in model performance is not the choice of algorithm, but the quality and structure of the data that feed into it. By approaching the task through both a technical and a storytelling perspective, we uncovered how data preparation serves not only as a set of processing techniques, but also as a narrative device that shapes the clarity, credibility, and interpretability of the final analytical outcomes.

From a technical standpoint, the project demonstrated that raw data alone cannot support meaningful or reliable modeling. Severe missingness, outliers, implausible biological values, skewed distributions, and high noise levels in actigraphy signals introduced significant risks at every stage of the pipeline. These issues distorted variable relationships, misrepresented behavioral patterns, and caused all three baseline models to overfit, misclassify, and behave inconsistently across folds. Through systematic data cleaning, distribution correction, feature engineering, and the construction of domain-informed movement summaries, the structure of the data was gradually restored. As each preparation step was introduced, the models improved not only in accuracy but also in stability, generalizability, and diagnostic reliability. By the final stage, all models achieved higher performance, more consistent cross-validation behavior, and significantly improved detection of severe PIU cases. This confirmed a central truth: meaningful modeling requires meaningful data.

Parallel to the technical workflow, storytelling played a critical role in guiding the project narrative. Rather than presenting the analysis as a sequence of independent steps, the storytelling structure allowed us to build tension, highlight conflicts, and reveal the resolution in a coherent and engaging way. The Tension Phase exposed the weaknesses and dangers of working with unprepared data, especially in a behavioral domain where misclassification can have real implications. The Resolution Phase then demonstrated how thoughtful data preparation not only resolves these issues technically, but also transforms the analytical landscape by amplifying signal, reducing bias, and clarifying behavioral insights. Throughout this process, the Big Idea of the project remained clear: the main character of the story is not the machine learning model, but the data preparation.

This dual emphasis on techniques and storytelling ultimately reinforces a broader message about data science practice. Models do not exist in isolation; they are reflections of the data they learn from, the decisions made during preparation, and the narrative framework within which they are interpreted. By integrating rigorous methodological work with intentional, narrative-driven communication, the project not only improved model performance but also provided a clearer, more human-centered understanding of the behavioral phenomena underlying PIU.

In conclusion, this project illustrates that technical excellence and effective storytelling are not separate goals but mutually reinforcing elements of high-quality data science. Data preparation provides the structure; modeling provides the mechanism; storytelling provides the meaning. When combined, they create analyses that are accurate, interpretable, and impactful. This alignment between technique and narrative is ultimately what allows data-driven insights to be both scientifically valid and communicatively powerful.