**Kids Screen Time Project**

Asrith Krishna Vejandla

Student ID: 005033913

MSCS-634 Advanced Big Data and Data Mining

Residency Project

Professor Satish Penmatsa

July 13th, 2025

**Abstract**

The present report summarizes a four-stage data-mining project that investigated daily screen-time behavior and associated health impacts among Indian school-age children. A publicly available Kaggle dataset containing 9668 records and eight key attributes was cleaned, explored, and modeled. Linear, regularized, and ensemble methods were applied to predict screen-time quantity and health outcomes. Unsupervised techniques revealed usage archetypes, and association-rule mining highlighted actionable patterns. Results show that the educational share of screen time consistently mitigates negative health reports, whereas prolonged recreational use, particularly on portable devices, raises the likelihood of eye strain and poor sleep. The report concludes with recommendations for parents, educators, and app developers, together with ethical reflections on privacy and algorithmic fairness.

**Introduction**

Digital media has become a central component of childhood leisure and learning. Excessive or unbalanced screen use has been linked to physical and psychological complaints, yet blanket time‐limit guidelines can be difficult to enforce and may overlook content quality. This project sought to identify quantitative and qualitative factors that best explain children’s screen habits and their self-reported health impacts. The guiding questions were threefold. First, which demographic or behavioral attributes predict average daily screen time. Second, can specific health complaints be anticipated from observable features. Third, do natural clusters and association patterns reveal modifiable risk profiles.

**Data and Preparation**

The Indian Kids ScreenTime 2025 dataset was selected because it contains both usage metrics and outcome variables while exceeding the minimum requirement of five hundred observations and eight attributes. Variables include age, gender, primary device, urban or rural residence, average daily screen hours, educational-to-recreational ratio, a binary flag for exceeding recommended limits, and a multi-label field of health impacts (anxiety, eye strain, obesity risk, and poor sleep). After loading the CSV into Pandas the column types and ranges were verified. Missing values in the health-impact column were replaced with the literal string “None” to preserve intentional non-impact reports. Numeric gaps were imputed using median values and three records with impossible ages were removed. Duplicate rows were dropped. Exploratory analysis revealed a right-skewed distribution of screen hours and a moderate negative correlation between age and educational ratio.

A graph of a distribution

AI-generated content may be incorrect.

Figure

A graph of a screen time

AI-generated content may be incorrect.

Figure

A screenshot of a graph

AI-generated content may be incorrect.

Figure

A graph of age versus screen time

AI-generated content may be incorrect.

Figure

**Regression Modeling**

Deliverable 2 addressed the continuous target of average daily screen time. A simple linear regression using age alone explained less than two percent of the variance, indicating that age is a weak stand-alone predictor. A multiple regression incorporating gender, device type, locality, educational ratio, and an age-by-ratio interaction raised the explained variance only marginally. Ridge regression with cross-validated regularization delivered the lowest root-mean-square error of 1.65 hours, but still under-predicted heavy users.

A graph with blue lines

AI-generated content may be incorrect.

Figure

The educational-to-recreational ratio emerged as the strongest negative coefficient, while tablet use and age contributed positively. Approximately twenty-eight percent of hold-out predictions fell within one hour of the actual value. Figure 5 provides the predicted versus actual scatter plot for the ridge model.

A graph with a line and dotted line

AI-generated content may be incorrect.

Figure

**Classification, Clustering, and Association Patterns**

Deliverable 3 employed multi-label classification to forecast health impacts. One-vs-rest logistic regression and a tuned random-forest ensemble both achieved macro area-under-the-curve scores near 0.55, reflecting moderate recall but low precision.

A graph of a graph showing different types of data

AI-generated content may be incorrect.

Figure

Although predictive power was limited, feature-importance analysis confirmed that higher educational ratios reduce the probability of every health complaint, whereas tablet usage elevates risks of eye strain and poor sleep.

Principal-component analysis followed by three-means clustering revealed three balanced groups. Cluster 0 consisted of light, mainly educational users who rarely exceeded time limits or reported complaints. Cluster 2 represented moderate, mixed users with some over-limit incidence. Cluster 1 comprised heavy, recreational users with the lowest educational ratio and the greatest share of health complaints. The below figure illustrates the cluster separation on the first two principal components that together account for seventy-four percent of the variance.

A graph of different colored dots

AI-generated content may be incorrect.

Figure

Association-rule mining with the Apriori algorithm identified high-lift relationships linking low educational ratio and specific devices to complaints. The strongest rule showed that children who use laptops and maintain less than forty percent educational content experience poor sleep sixty-four percent of the time, a rate thirty-three percent higher than expected by chance. Additional rules linked eye strain with smartphone use under similar content conditions.

**Ethical Considerations**

All analyses relied on an anonymized public dataset devoid of personal identifiers, minimizing privacy concerns. Model fairness was considered by inspecting coefficients and feature importances for unintended demographic effects. No evidence suggested that gender or locality alone drove negative classifications. Nevertheless, the report emphasizes that recommendations target behavior, not identity, to avoid stigmatizing specific demographic groups. Transparency is maintained through open notebooks and explicit documentation of cleaning and modeling steps.

**Recommendations**

Parents and guardians should focus on increasing the educational proportion of children’s screen time rather than imposing blanket hour caps alone. Encouraging high-quality content appears to buffer the negative effects of extended use. Heavy recreational users, especially those relying on tablets and smartphones, benefit from structured breaks, blue-light filtering, and bedtime device restrictions. App developers can incorporate usage-ratio tracking and push educational content when recreational exposure dominates. Policymakers might refine guidelines by combining duration thresholds with content-quality targets.

**Future Work**

Future research should incorporate temporal usage logs, screen brightness, and ergonomic factors to improve health-impact prediction. Non-linear models such as gradient-boosted trees may capture interactions missed by linear and shallow tree ensembles. Experimental interventions that shift the educational ratio could validate causal interpretations suggested by the association rules.

**References**

1. American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). American Psychological Association.
2. Kaggle. (2025). Indian Kids ScreenTime 2025 \[Data set]. Kaggle. <https://www.kaggle.com/datasets/ankushpanday2/indian-kids-screentime-2025>