Title

Screen Time, Smartphone Usage, and Well-Being: Comparing Advanced Classical Models for Tabular Data

---

1. Introduction and Project Statement

Smartphones and laptops are deeply integrated into daily life, especially for students and younger adults. While screen time enables productivity and social connection, excessive use has been associated with worse sleep, increased anxiety, and poorer overall mental health in many studies.(PMC) However, most public conversations about screen time focus on simple thresholds ("more than 4 hours is bad") rather than understanding which *patterns* of use (e.g., late-night usage, social media vs. productivity apps, notifications, etc.) are most predictive of negative outcomes.

The goal of this project is to use **tabular machine learning** to model the relationship between daily screen time patterns and indicators of well-being (e.g., self-reported mental wellness or phone addiction scores). Specifically, we will:

- Build predictive models that estimate mental wellness / phone addiction categories from screen-time and usage features.
- Compare **advanced classical algorithms** (XGBoost, Support Vector Machines) to models studied earlier in the course (Logistic Regression, k-NN, Random Forest).
- Analyze which aspects of screen use (total time, app category usage, notifications, etc.) are most strongly associated with poorer well-being.

This directly addresses the instructor's suggestion to **investigate advanced classical algorithms for tabular data** while grounding the work in a topic that is highly relevant to real life: how our device habits relate to our mental health.

---

2. Data Sources and Technologies Used

Data Sources

I plan to use at least one (possibly two) publicly available screen-time datasets that include both **usage metrics** and **well-being indicators**:

1. **ScreenTime vs Mental Wellness Survey (400 Users)** – Kaggle dataset with self-reported daily screen usage and mental wellness scores/labels.(Kaggle)

2. **Smartphone Usage and Behavioral Dataset** – Tabular dataset of ~1,000 users with features such as daily screen-on time, app usage time, and engagement patterns; may be used for additional modeling or feature engineering.(Kaggle)

These datasets give us both **input features** (screen hours, app categories, notifications, device usage patterns) and **targets** (e.g., mental wellness level, addiction score, or wellness categories), enabling classification or regression tasks.

Technologies

- **Python** and **Jupyter Notebook** for development
- **pandas**, **NumPy** for data cleaning and preprocessing
- **scikit-learn** for baseline models (Logistic Regression, k-NN, Random Forest, SVM)
- **xgboost** (or `XGBClassifier` via `xgboost`/`sklearn`) for gradient-boosted tree models
- **matplotlib** / **seaborn** for visualization (EDA and result plots)

Optionally, I may use:

- **SHAP** or permutation feature importance to interpret XGBoost models
- **GridSearchCV** or **RandomizedSearchCV** for hyperparameter tuning

---

3. Methods Employed

3.1 Data Preprocessing and Exploration

- **Load and inspect data**: Examine shape, variable types, and target distribution (e.g., “low/medium/high mental wellness” or “addicted vs. not-addicted”).

- **Handle missing values**: Impute or drop as appropriate (e.g., median imputation for numeric fields, mode for categorical).

- **Feature engineering** (depending on dataset):

    - Aggregate or normalize screen time (e.g., total daily hours, late-night usage ratio).

    - Derive meaningful ratios (social media time / total time, productivity time / total time).

- **Encoding & scaling**:

    - One-hot encode categorical features (e.g., gender, device type, main usage category).

    - Standardize/scale numeric features for SVM and k-NN.

- **EDA**:

    - Histograms / KDE plots of screen time distribution.

    - Boxplots of screen time vs wellness category.

    - Correlation heatmaps between usage metrics and wellness indicators.

3.2 Model Design and Training

I will treat this as a **supervised classification problem**, where the target might be:

- Binary (e.g., “good vs poor mental wellness”, “low vs high addiction risk”), or

- Multi-class (e.g., low / moderate / high wellness).

Models:

1. **Baseline Models from Unit 2**

    - Logistic Regression

    - k-Nearest Neighbors (k-NN)

    - Random Forest

2. **Advanced Classical Models**
    - **Support Vector Machine (SVM)** with RBF and/or linear kernel
    - **XGBoost (Gradient Boosted Trees)**

Training procedure:

- Stratified **train/validation/test split** to preserve class balance.
- Use **cross-validation** on the training data for hyperparameter tuning:
    - SVM: C, kernel, gamma
    - XGBoost: learning rate, max depth, number of estimators, subsampling, etc.
- Compare models on the **test set** after tuning.

3.3 Evaluation and Interpretation

Metrics:

- Accuracy
- Precision, Recall, F1-score (especially important if classes are imbalanced)
- ROC–AUC (for binary tasks)

Analysis:

- Confusion matrices to see which wellness categories are misclassified.
- Feature importance:
    - XGBoost feature importance / SHAP values to understand which screen-time features matter most.
- Sensitivity analysis: Does prediction performance change significantly if we exclude one group of features (e.g., social media time) versus another?

Key questions:

- Do XGBoost/SVM clearly outperform simpler models like Logistic Regression and k-NN?
- Which patterns of screen behavior are most predictive of poor wellness/addiction risk?
- Is there a “threshold” range of daily screen hours where risk increases sharply, consistent with prior research?(CDC)

---

4. Results (Planned / Expected)

Since this is an initial proposal, there are no empirical results yet. However, I expect to report on:

1. **Model comparison table**
    - Test performance (Accuracy, F1, ROC–AUC) for each algorithm.
    - Discussion of trade-offs: e.g., XGBoost may achieve higher accuracy but require more tuning than Logistic Regression.
2. **Importance of different screen-time factors**
    - Which variables (total hours, late-night use, social media time, notifications, etc.) show the strongest association with low wellness or high addiction risk?
    - Visualization of feature importance (bar plots, SHAP summary plots).
3. **Practical interpretation**
    - Based on the models, are there clear patterns like “>4 hours/day of screen time plus high notification volume is strongly associated with poor mental health,” in line with existing literature?(PMC)
    - Brief discussion of limitations (e.g., self-reported data, cross-sectional nature, possible confounders).

These results will be presented in the final report with supporting charts and tables.

---

5. References *(initial list)*

- Kaggle – **ScreenTime vs Mental Wellness Survey (400 Users)**.(Kaggle)
- Kaggle – **Smartphone Usage and Behavioral Dataset**.(Kaggle)
- Santos, R. M. S. et al. *The associations between screen time and mental health in adolescents: a systematic review.* BMC Psychology, 2023.(SpringerLink)
- Zablotsky, B. et al. *Daily Screen Time Among Teenagers.* CDC NCHS Data Brief, 2025.(CDC)
- XGBoost documentation and scikit-learn documentation for implementation details.