How Digital and Lifestyle Behaviors Shape Wellness: A Supervised and Unsupervised Learning Approach

Kexin Fang, Lihong Gao, Shirley Shen, Shaibah Raiyan

All group members contributed to the presentation and slides; specifically, **Kexin**, **Lihong**, and **Shirley** divided the work on research questions, coding, and the report, while **Shaibah** was responsible for the conclusion.

1. Introduction

Digital behaviors have become deeply embedded in daily life, and concerns about their impact on mental wellness continue to grow. Public discussions often generalize "screen time" as harmful or helpful, yet mental wellness outcomes are shaped by a broader set of everyday behaviors, including lifestyle habits such as sleep and physical activity. A data-driven approach is therefore needed to better understand these patterns.

This study applies analytic techniques learned in class to investigate the relationship between digital behaviors, lifestyle factors, and wellness outcomes. We structure our analysis around three guiding questions that progress from association, to prediction, to identifying broader patterns of user behavior.

1. 1. Research Questions

1) How are digital usage behaviors and lifestyle associated with mental health scores and stress levels?
2) Can we predict stress level using digital behaviors?
3) Are there distinct user personas that naturally emerge from digital use behaviors?

1. 2. Expected Findings (Before Analysis)

1) Individuals with healthier lifestyles (better sleep and more frequent physical activity) are expected to show higher mental health scores, while heavier digital engagement (greater screen time or gaming) likely corresponds to increased stress levels.
2) More intensive social media use is expected to correspond to higher perceived stress.
3) Meaningful usage-based personas may emerge, such as heavy social media users, moderate users, and low engagement users.

2. Data Description

The dataset was obtained from a publicly available Kaggle repository "Tech Use and Stress Wellness". It contains 5,000 anonymized survey responses capturing individuals' technology-use patterns, lifestyle habits, and self-reported wellness outcomes.

To support our analysis, we organize the variables into three categories:

1) Digital Behavior Variables:

daily_screen_time_hours, social_media_hours, gaming_hours, work_related_hours, entertainment_hours, phone_usage_hours, laptop_usage_hours, tablet_usage_hours, tv_usage_hours (all of them are Numerics)

2) Lifestyle and Demographic Variables:

age (Integer); *gender, location_type* (Characters);

eats_healthy, uses_wellness_apps (Booleans); *physical_activity_hours_per_week, sleep_quality, sleep_duration_hours, caffeine_intake_mg_per_day, mindfulness_minutes_per_day* (Numerics)

3) Wellness Outcomes:

mental_health_score (Numerics), *stress_level* (Integer)

Overall, the dataset provides a broad and varied sample suitable for behavioral analysis. Digital usage variables show wide ranges in daily hours, reflecting diverse engagement levels across respondents. Lifestyle indicators such as *sleep quality, physical activity*, and *caffeine intake* also exhibit meaningful variation, enabling comparisons across different wellness profiles. The two wellness outcomes (*mental_health_score* and *stress_level*) span their full rating scales, offering sufficient contrast for both association and prediction tasks.

3. Methodology

3. 1. Data Preparation

Before applying statistical models, the dataset underwent rigorous preprocessing to ensure data quality and model suitability. The raw data was first cleaned by removing observations with missing values to maintain dataset integrity. To enable classification tasks, we engineered a binary target variable named *High_Stress*. This variable was constructed using a median split on the continuous *stress_level* variable.

- 0 (Low Stress): stress_level < Median
- 1 (High Stress): stress_level > Median

To explore the relationships between variables and determine the most relevant predictors for our Linear Regression model, we generated a correlation matrix.

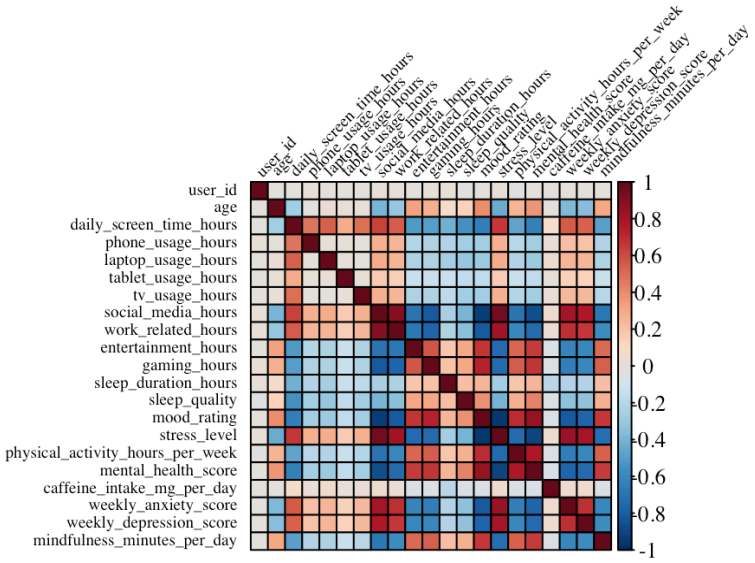

Figure 1 Variables Correlationships Matrix

Based on these results, we identified these highly correlated parameters as the primary candidates for the Linear Regression model, ensuring the analysis focuses on the strongest drivers of people's stress while monitoring for potential multicollinearity.

3. 2. RQ1 – Linear Regression

To comprehensively analyze the factors influencing mental well-being, we constructed two separate Multiple Linear Regression models using Ordinary Least Squares (OLS) estimation. The primary model investigates how lifestyle and digital behaviors affect the overall *mental_health_score*. A secondary model was developed to isolate the specific drivers of *stress_level*, particularly focusing on screen time types and clinical indicators.

3. 3. RQ2 – Classification Models (Predicting High Stress)

To predict high stress using digital behaviors, we employed classification methods to predict whether a person falls into the "High Stress" or "Low Stress" based strictly on their digital usage patterns. The dataset was randomly split into a training set (80%, n=4000) and a test set (20%, n=1000) applying set.seed(123). Two distinct algorithms were trained and evaluated: KNN and Decision Tree.

1) **KNN**: We utilized nine numeric predictors representing digital behaviors: *daily_screen_time_hours*, *phone_usage_hours*, *laptop_usage_hours*, *tablet_usage_hours*,

tv_usage_hours, *social_media_hours*, *work_related_hours*, *entertainment_hours*, and *gaming_hours*. Because KNN relies on Euclidean distance, all predictor variables were standardized. We specified *k=5* for initialization to balance bias and variance.

2) **Decision Tree**: A classification tree was grown to predict High Stress using the same set of digital predictors. The algorithm used recursive binary splitting to minimize node impurity.

3. 4. RQ3 – Clustering Analysis (Identifying User Personas)

To address RQ3 and uncover latent behavioral patterns, we employed **K-Means Clustering**, an unsupervised learning technique suited for partitioning observations into distinct, non-overlapping subgroups. We selected three features: *daily_screen_time_hours, social_media_hours, and gaming_hours*, because they capture distinct modalities of digital behaviors: overall usage, socially driven interaction, and entertainment-oriented activity. To prevent variables with larger magnitudes from dominating the Euclidean distance calculations, we applied Z-score standardization (scale()) to normalize the data prior to clustering.

To determine the optimal number of clusters (k), we employed the Elbow Method by calculating the Total Within-Cluster Sum of Squares (WSS) for k ranging from 1 to 10. This approach allows us to identify the point where adding more clusters provides diminishing returns in variance reduction. As illustrated in *Figure 2*, the scree plot reveals a distinct inflection point ("elbow") at with k = 3, suggesting this value offers the most efficient balance between cluster compactness and interpretability.

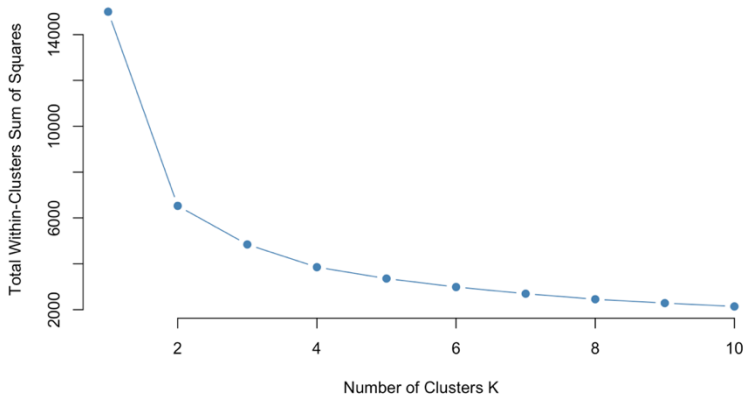

Figure 2 Elbow Method for Optimal k

Consequently, the algorithm was executed with three clusters. To ensure the stability and reproducibility of the results, we fixed the random seed (set.seed(123)) and specified nstart = 20,

forcing the algorithm to initialize with multiple random configurations to avoid local optima. Finally, we characterized the resulting user personas by mapping the cluster assignments back to the original dataset and computing the centroids (means) of the unscaled variables. This allowed for a practical interpretation of behavioral patterns distinct from the standardized values used for computation.

4. Results

4. 1. Lifestyle Predicts Mental Health, While Social Media Dominates Stress

This section applies two regression models to clarify how lifestyle and digital behaviors influence mental wellness.

The first model explains 82 percent of overall mental well-being and identifies physical activity and sleep quality as the strongest positive predictors. Specifically, sleep quality far outweighs sleep duration, which makes logical sense. It also shows that entertainment viewing and gaming are unexpectedly associated with higher mental health scores, suggesting that moderate recreational screen time may function as a constructive coping mechanism.

```
> summary(lm_model1)

Call:
lm(formula = mental_health_score ~ age + entertainment_hours +
    gaming_hours + sleep_duration_hours + sleep_quality + mood_rating +
    physical_activity_hours_per_week + mindfulness_minutes_per_day,
    data = clean_data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.7925  -3.7467   0.0336   3.8034  22.5391

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    31.563160   1.142846  27.618  < 2e-16 ***
age                             0.026667   0.004973   5.363 8.56e-08 ***
entertainment_hours             1.506144   0.164858   9.136  < 2e-16 ***
gaming_hours                    0.936562   0.168969   5.543 3.13e-08 ***
sleep_duration_hours            0.381535   0.150177   2.541   0.0111 *
sleep_quality                   2.228866   0.128923  17.288  < 2e-16 ***
mood_rating                     1.952589   0.057625  33.884  < 2e-16 ***
physical_activity_hours_per_week  1.957856   0.048746  40.164  < 2e-16 ***
mindfulness_minutes_per_day     0.129696   0.013373   9.699  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.521 on 4991 degrees of freedom
Multiple R-squared:  0.8225,    Adjusted R-squared:  0.8222
F-statistic:  2892 on 8 and 4991 DF,  p-value: < 2.2e-16
```

Figure 3 Linear Regression Model 1 Summary

The second model explains 93 percent of stress variation and reveals sharp contrasts across digital behaviors. Social media is the strongest stressor, indicating that passive comparison and

continuous social exposure impose unique psychological pressure. In contrast, work-related digital use slightly reduces stress, likely reflecting better productivity or academic structure. Together, the models show that screen time effects are highly context-dependent, with different digital behaviors producing either beneficial or harmful outcomes.

```
> summary(lm_model2)

Call:
lm(formula = stress_level ~ gaming_hours + daily_screen_time_hours +
    social_media_hours + work_related_hours, data = clean_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.48866 -0.58508 -0.06799  0.53235  2.23906

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.398248   0.106185 -13.168  < 2e-16 ***
gaming_hours             0.094403   0.027479   3.436 0.000596 ***
daily_screen_time_hours  0.148974   0.007939  18.766  < 2e-16 ***
social_media_hours       2.696408   0.025958 103.876  < 2e-16 ***
work_related_hours      -0.778843   0.032905 -23.670  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7949 on 4995 degrees of freedom
Multiple R-squared:  0.9257,    Adjusted R-squared:  0.9256
F-statistic: 1.556e+04 on 4 and 4995 DF,  p-value: < 2.2e-16
```

Figure 4 Linear Regression Model 2 Summary

4. 2. Decision Tree Reveals Social Media Use as Dominant Stress Predictor

This section details the performance of the classification models used to predict High Stress levels based on digital behaviors.

The KNN model (k=5) achieves solid predictive performance with 92.5% accuracy on 1,000 test cases, correctly identifying 541 Low Stress and 384 High Stress individuals. While effective, the “black box” nature of KNN offers limited insight into *which* specific digital behaviors drive these predictions.

The Decision Tree performs better, reaching 96.6% accuracy with only 34 total misclassifications. More importantly, it provides clear interpretive insight. Although multiple device-use predictors were available, the model relied solely on *social_media_hours*, indicating that this single variable maximizes predictive separation.

Table 1 KNN VS. Decision Tree

Model	Test Accuracy	Misclassification Rate	Strongest Predictive Identified
KNN	92.5%	7.5%	N/A (Distance-based)
Decision Tree	**96.6%**	**3.4%**	*Social_media_hours*

The tree centers on a key threshold at 3.85 hours of daily social media use. Individuals below this level are consistently predicted to have Low Stress, with secondary splits reinforcing the stability of this low-risk range. Those above 3.85 hours are overwhelmingly predicted to experience High Stress, regardless of minor usage variation. This structure indicates a pronounced threshold effect: stress risk sharply increases once social media use surpasses a critical level rather than rising incrementally.

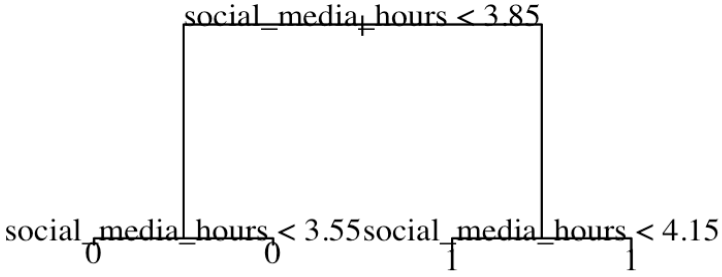

Figure 5 Decision Tree Results

Overall, the Decision Tree is the superior model due to higher accuracy and the ability to isolate a specific, actionable predictor. The analysis highlights social media use as the dominant digital factor influencing stress, with approximately 3.85 hours per day emerging as the tipping point between low and high stress outcomes.

4. 3. Three User Personas from Clustering

The K-means algorithm (k=3) revealed three distinct archetypes of technology engagement. As illustrated in *Figure 6*, the segmentation is driven primarily by the variance in social media usage, which serves as the dominant dividing factor among the groups. To provide a tangible interpretation of these personas, *Table 2* summarizes the average daily hours for each activity across the three clusters, calculated using the unscaled raw data.

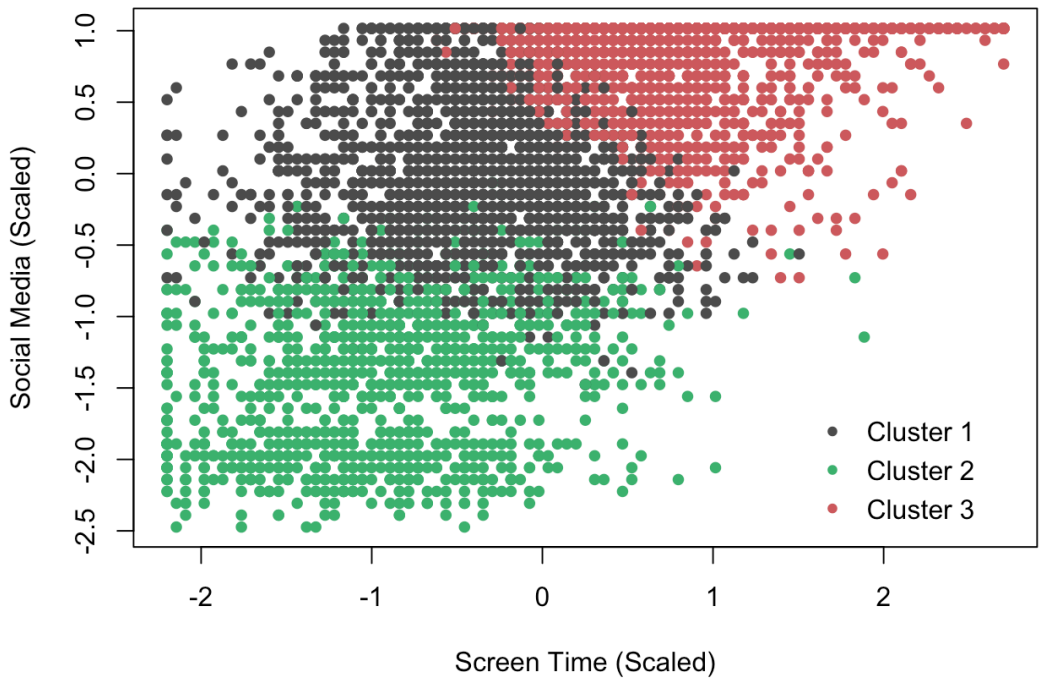

Figure 6 K-means clustering results (k = 3)

Table 2 Mean usage hours and behavioral labels for each cluster (Unscaled Data)

Cluster	Daily Screen Time (h)	Social Media (h)	Gaming (h)	User Persona Label
Cluster 1	4.3	3.3	1.5	Moderate Users
Cluster 2	3.4	1.6	**2.5**	Low-Engagement / Gamer Niche
Cluster 3	**6.6**	**4.3**	1.1	Heavy Social Media Users

- **Cluster 1 (Black): Moderate Users**

This segment reflects balanced usage levels. Their behavior represents an "average" user profile, with daily screen time averaging 4.3 hours, accompanied by moderate social media usage (3.3h) and gaming activity (1.5h). They do not lean strongly toward any single extreme modality.

- **Cluster 2 (Green): Low-Engagement Users/Gamer Niche**

These users consistently show the lowest engagement in digital communication, with daily screen time (3.4 h) and social media use (1.6 h) significantly below the other groups. Interestingly, despite their low overall connectivity, they exhibit the highest gaming activity (2.5h), suggesting a specific preference for offline or solitary digital entertainment.

- **Cluster 3 (Red): Heavy Social Media Users**

Users in this group exhibit high overall screen time (6.6h), driven primarily by intense social media interaction (4.3h). This validates our earlier supervised models linking high social media use to

stress. Notably, their gaming time (1.1h) is the lowest, indicating their digital life is dominated by social platforms rather than gaming.

Collectively, these personas demonstrate that digital engagement is not merely a matter of intensity (high vs. low) but of modality (social vs. gaming). While Cluster 3 validates the link between heavy social media use and potential stress identified in our supervised models, the emergence of the "Gamer Niche" (Cluster 2) highlights that low social connectivity does not imply total digital disengagement. This nuance is critical, as it confirms that the "high-risk" profile is specifically tied to social platforms, rather than screen time in general.

5. Conclusion

5. 1. Key Insights

Our analysis reveals that mental wellness depends on specific digital behaviors rather than total screen time. Social media emerged as the primary stressor, with our Decision Tree identifying a critical 3.85-hour daily threshold for high stress. This finding is reinforced by the "Heavy Social Media" cluster, whose average usage of 4.3 hours places this demographic firmly in the high-risk category. Conversely, recreational gaming functions as a constructive coping mechanism associated with better mental health. This is evident in the distinct "Gamer Niche" cluster (Cluster 2), where high gaming engagement paired with low social media use aligns with positive outcomes. Finally, sleep quality proves to be a significantly stronger predictor of well-being than sleep duration.

5. 2. Practical Implications

Interventions should abandon generic "screen time" limits in favor of targeting high-risk digital behaviors. Policies should specifically address social media usage exceeding the 3.85-hour threshold identified by our Decision Tree model, rather than restricting constructive recreational use. For individuals, "digital hygiene" must distinguish between toxic social comparison and harmless entertainment, while prioritizing sleep quality as a fundamental buffer against academic pressure.

5. 3. Limitations

Three limitations constrain these findings. First, the cross-sectional design establishes correlation but not causation, leaving the directionality between stress and social media use unclear. Second,

reliance on self-reported data introduces potential recall bias regarding specific digital behavior variables. Finally, unobserved external confounders, such as academic workload or financial status, were not controlled for in this public dataset.

5. 4. Future Directions

Future research should adopt longitudinal designs to track stress spikes over time. Replacing self-reported surveys with objective device logging (e.g., Screen Time APIs) would eliminate bias, while integrating contextual data—such as exam schedules—would allow for more precise prediction of student vulnerability.