Clustering Analysis of Screen Time and Wellness

Luis Dale Gascon
Computer Science
Towson University
lgascon1@students.towson.edu

***Abstract*—We examine how device usage affects personal wellness using a dataset that tracks hours spent on devices, technology habits, and related wellness metrics. Drawing from customer segmentation techniques in marketing, our approach groups users with similar digital behaviors and wellness patterns. Using a trained classifier model, we then recommend personalized strategies to help individuals enhance their well-being.**

***Index Terms*—wellness, clustering, technology usage, mental health, unsupervised learning**

I. Introduction

The rising use of technology has been linked to negative impacts on mental health and overall well-being. This proposal examines how people use technology across various devices such as phones, laptops, tablets, and televisions, and for different activities including social media, work, entertainment, and gaming. We plan to explore how these usage patterns relate to wellness metrics such as sleep quality, mood, stress levels, mental health assessments, and other lifestyle factors. Using this data, we want to identify user segments and train a classification model to recommend personalized strategies that promote healthier technology habits, taking inspiration from user segmentation on the field of marketing.

II. Dataset

The dataset that we'll be working with contains 5,000 rows and is publicly available from Kaggle. It describes device usage hours, usage types, and a range of wellness metrics (sleep quality, mood, stress levels, mental health score, healthy eating, caffeine intake, weekly anxiety, weekly depression, and mindfulness).

III. Related Works

(John, Shobayo and Ogunleye, 2023) compared several leading clustering algorithms, including K-means, Gaussian Mixture Models (GMM), DBSCAN, agglomerative clustering, and BIRCH, for customer segmentation in the UK retail market. Their findings indicated that GMM, when used alongside PCA for dimensionality reduction, outperformed other methods by achieving a Silhouette Score of 0.80, whereas K-means, BIRCH, and agglomerative algorithms all scored 0.64. A score closer to 1 indicates that the clusters are more defined.

Another study by (Ling and Weiling, 2025) evaluated clustering performance using both the Silhouette Score and the Davies-Bouldin Index. A higher Davies-Bouldin Index implies that clusters are less compact and not well separated. This research observed that Gaussian Mixture Models encounter difficulties when handling high-dimensional or large-scale datasets, while K-means++ produced more reliable results even in the presence of high dimensionality and noise.

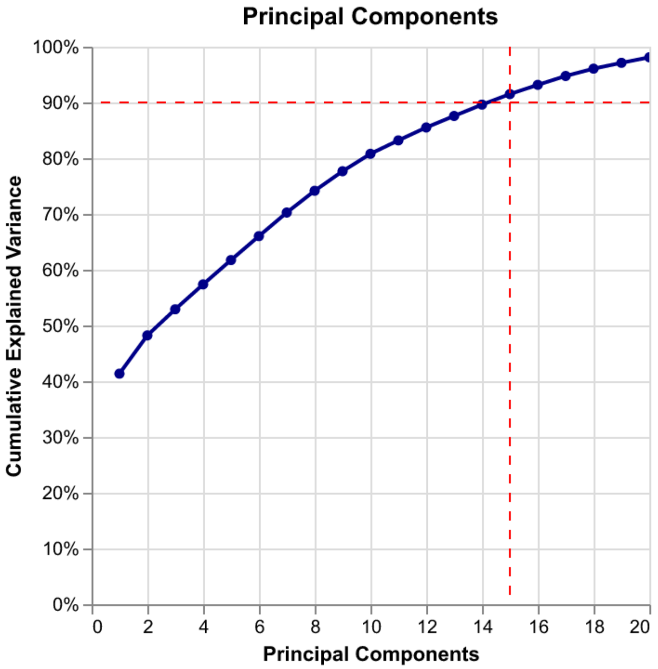
In the context of mental health, research such as this study by (Lee and Žarnic, 2024) explored the relationship between technology use and psychological well-being. The study found that active digital engagement is positively associated with anxiety symptoms, and that access to the internet correlates with higher levels of depression and anxiety, especially among younger individuals. Nevertheless, the paper emphasized that there is no clear causal link established between technology use and mental health outcomes.

IV. Methods

We propose an approach that combines clustering and classification for user segmentation. To capture overlapping user groups, we will be comparing 2 different clustering algorithms: Gaussian Mixture Models (GMMs), which offer soft clustering as opposed to K-Means, which provides hard clusters and fuzzy c-means. This is important because individuals often display multiple, overlapping behaviors in their technology use.

Based on the methodology from (John, Shobayo and Ogunleye, 2023) and the insights from (Ling and Weiling, 2025) regarding GMMs and high dimensionality, we will apply Principal Component Analysis (PCA) to reduce the dimensionality of our dataset before clustering.

To find the optimal number of components to keep during PCA, we will plot the value of cumulative explained variance per components from a range of 1 to 20, with a goal of choosing the number of components that keeps more than 90% of variance. Our results show that 15 components is the lowest value that keeps more than 90% of variance at 91.5%.



Before running the clustering algorithm, we will determine the optimal number of Gaussian components by evaluating a range of components and examining changes in the Bayesian Information Criterion (BIC) as (Lavorini, 2018) showcased in the author's Medium blog. Once the clusters are formed, we will assign labels to each group based on shared characteristics.

After we obtain the probabilities from clustering, we will assign a threshold that assigns that datapoint to 1 or more labels.

Once the preprocessed dataset has each row labeled, we will split the dataset for test train split via holdout and train a multi-label classifier model to predict

A. *Toolset*

Our primary programming language will be Python, chosen for its straightforward syntax and extensive ecosystem of packages available via PyPI. For data manipulation, we will use Pandas; for machine learning algorithms, scikit-learn; and for interactive visualizations, Altair. Marimo will serve as our interactive prototyping notebook, while Streamlit will be used to develop a prototype frontend once the model is ready. To ensure reproducibility, we will manage dependencies with uv (a Rust-based alternative to pip) and devenv.

V. Analysis

A. *GMM*

After obtaining the probabilities generated by We wanted to set a threshold to set This tells us that the majority of users are not overlapping. We'll set the probability threshold at 30%

References

[1] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *Analytics*, vol. 2, no. 4, pp. 809–823, 2023, doi: 10.3390/analytics2040042.

[2] L. S. Ling and C. T. Weiling, "Enhancing Segmentation: A Comparative Study of Clustering Methods," *IEEE Access*, vol. 13, no. , pp. 47418–47439, 2025, doi: 10.1109/ACCESS.2025.3550339.

[3] J. Lee and Ž. Žarnic, "The impact of digital technologies on well-being: Main insights from the literature," *OECD Papers on Well-being and Inequalities*, no. 29, 2024, doi: 10.1787/cb173652-en.

[4] V. Lavorini, "Gaussian mixture model clusterization: How to select the number of components (clusters)." [Online]. Available: https://medium.com/data-science/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4