Clustering Analysis of Screen Time and Wellness (Proposal)

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 use Gaussian Mixture Models (GMMs), which offer soft clustering as opposed to K-Means, which provides hard clusters. 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. We will start by using the elbow method to identify the optimal number of principal components, with the help of the scree plot for visualization.

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 clustering, we will train a supervised classification model using the cluster labels, with a 70/30 split for training and testing the data.

A. Toolset

Our main tool of choice will be Python for its simple syntax and massive collection of popular packages from PyPi. Polars will serve as our dataframe library for data manipulation, a Rust alternative to Pandas, sklearn for the algorithms mentioned, and Altair for interactive visualization. Marimo will be used as our interactive prototyping notebook, and Streamlit for developing a prototype frontend once we are satisfied with the model. To ensure reproducibility, dependencies will be managed with uv, a Rust alternative to pip, and devenv.

V. Expected Outcomes

We expect GMMs to reveal meaningful user clusters to show distinct profiles of technology use and wellness. Coming up with appropriate labels for these clusters will require thorough analysis and interpretation. Picking the proper strategies to enable individuals to use technology more mindfully, leading to improved well-being will be quite the challenge as usage patterns may be much more complex than expected. The methodology and findings may further inform the development of personalized digital wellness tools and advice.

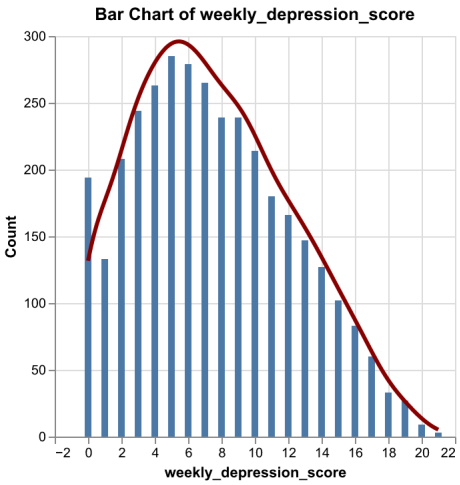
VI. Midterm Progress

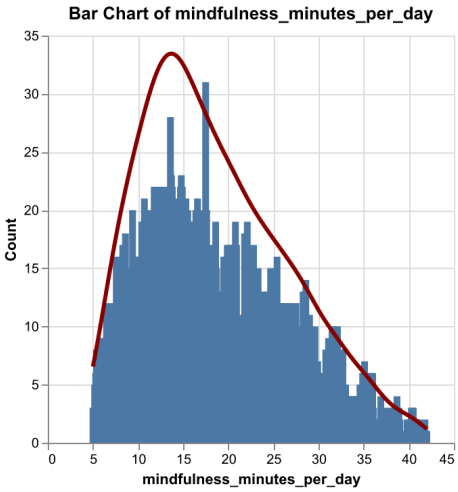
This section reports on any progress made between the submission of this proposal and the due date of this progress report. We will also discuss our plans for the next 45 days and expectations.

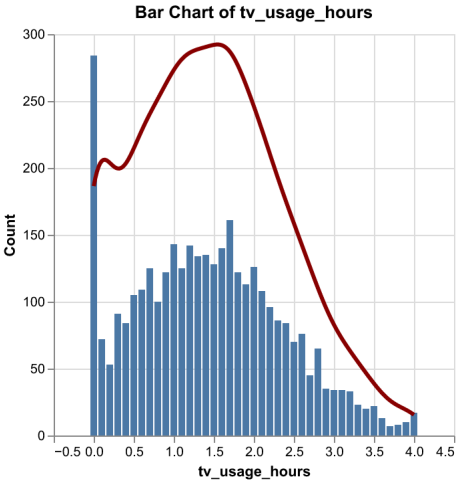
A. Exploratory Data Analysis

To gather insights about the dataset, we generated various visualizations of the dataset. We've generated a histogram of each continuous feature with their respective kernel density estimate (KDE) to determine the skewness of each feature.

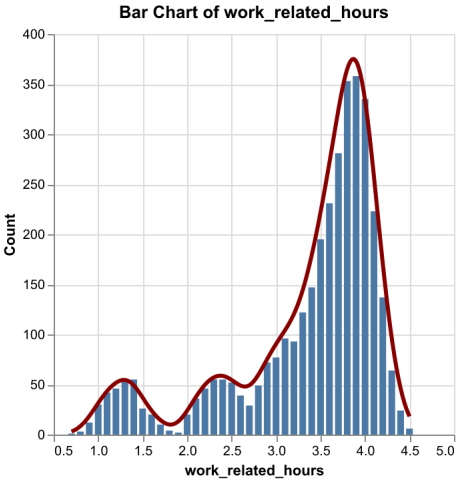
From the histograms, the following features appear positively skewed:







One feature appears negatively skewed:



We've also generated a correlation matrix to analyze any relationships that each pair of features might have:

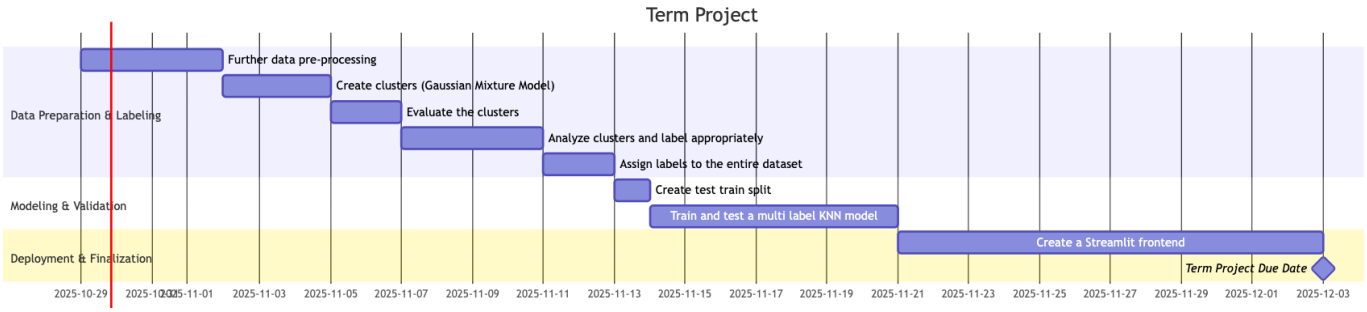


Fig. 2. Gantt chart for the next 45 days

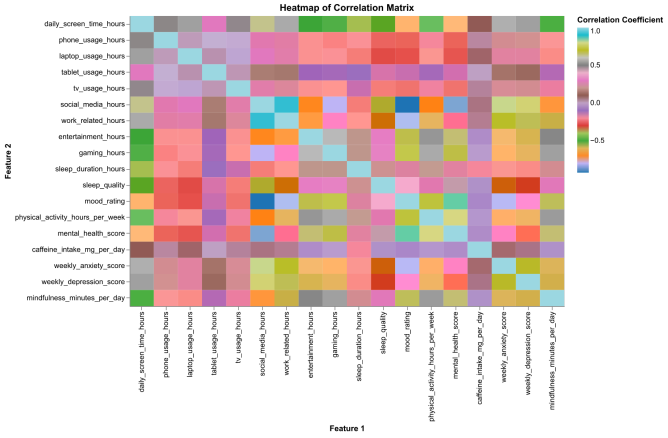


Fig. 1. Correlation Matrix of the dataset

B. Data Preprocessing

a) *Encoding:*

Before training the model, we first have to encode any categorical features within our dataset.

We will use one-hot encoding for categorical features such as location type and gender. Boolean features will be mapped to 0 (false) or 1 (true)

b) *Standardization:*

PCA is highly sensitive to the scale of the features.

For continuous features, we will standardize the data so that all features are on the same scale and equally important.

c) *Normal Distribution:*

Since GMM assumes that our dataset is a mixture of Gaussian distributions, we'll apply appropriate transformations to reduce skewness and approximate a normal distribution for each feature.

C. Changes in the toolset and methods

After working with sklearn, we've decided to use Pandas instead of its alternative, Polars, because sklearn functions do not natively support Polars DataFrames. All other tools in our workflow will remain unchanged.

After generating clusters with GMM, we'll use the resulting probabilities as additional features. We'll assign a target label of 1 to any cluster with a probability greater than 50%. For example:

User ID	P(High Stress)	P(Low Screen)	P(Balanced)	T(High Stress)	T(Low Screen)	T(Balanced)
101	0.85	0.10	0.05	1	0	0
102	0.15	0.35	0.50	0	0	1
103	0.25	0.10	0.65	0	0	1
104	0.60	0.40	0.00	1	0	0

Target Label Generation:
The binary target vectors (0 or 1) are derived from the GMM probabilities by applying a threshold. Here, if the probability for a cluster exceeds 0.50, its target label is set to 1.

In our proposal, we did not specify which supervised classification model we would implement. We now plan to use a multi-label classifier and will experiment with models such as multi-label KNN and ensemble methods like Random Forest, comparing their accuracies.

To evaluate and compare the performance of each multi-label classifiers, we plan to use a few of the evaluation metrics that Rokach, Schclar and Itach (2013) used, namely Hamming Loss, and label-based micro-averaged F-measure.

D. Current Expectations

We expect to change our threshold for the GMM-generated probabilities around as 50% may be a little high. As for the completion of the project, we expect that we'll be able to create a polished UI via streamlit as we don't expect to encounter any significant blockers with our proposed plan. We are concerned with the time that it will take to perform intensive research on remedies for each user segments.

E. Roles

Given that this is a solo project, the sole author will assume all responsibilities and tasks to deliver this project.

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

[5] L. Rokach, A. Schclar, and E. Itach, “Ensemble Methods for Multi-label Classification.” [Online]. Available: https://arxiv.org/abs/1307.1769