San Diego State University

STAT 520: Applied Multivariate Analysis

Mental Health and Technology Usage Analysis

Miguel Ángel Bravo Martínez del Valle
LinkedIn Website

Fernanda Carrillo Escarcega
LinkedIn

Professor: Ryan Paul Lafler

Date: December 11, 2024

Abstract

This project explores the relationship between technology usage and mental health using the Kaggle dataset on technology usage and mental health. Through statistical analysis and machine learning models, we investigate correlations between various variables such as screen time, stress level, and access to support systems.

Contents

1 Introduction

The rapid advancement of technology has significantly transformed modern life, bringing both benefits and challenges. As digital devices become increasingly integrated into daily routines, understanding the impact of technology usage on mental health is of growing importance. Excessive screen time, high engagement on social media platforms, and reduced physical activity have all been highlighted as factors that may negatively affect mental well-being, contributing to higher levels of stress, anxiety, and even depression [1].

This project investigates the relationship between technology usage patterns and mental health by leveraging the Mental Health and Technology Usage Dataset from Kaggle [1]. Through statistical analysis and machine learning techniques, this study aims to address several key questions:

1. Is there a significant relationship between screen time and stress levels?
2. How does physical activity affect mental health outcomes among high social media users?
3. What impact do support systems and work environment have on overall mental health status?

To address these questions, the project applies inferential statistics, supervised learning models for predictive analysis, and unsupervised clustering to identify user segments with distinct technology usage and mental health profiles. This research not only highlights the potential risks associated with excessive technology use but also explores factors that may help mitigate negative mental health outcomes. Ultimately, these insights can inform both individuals and organizations in fostering healthier interactions with technology.

2 Methodology

2.1 Dataset Description

The Mental Health and Technology Usage Dataset from Kaggle includes various variables that capture both technology usage behaviors and mental health indicators. This dataset

provides a comprehensive view of how digital habits, such as screen time and social media usage, correlate with mental well-being factors like stress levels, sleep, and access to support systems. Each variable in the dataset is briefly described in Table 1, highlighting its type and observed range based on an initial exploration of the data.

Table 1: Description of Dataset Variables

Variable	**Description**	**Type**	**Range**
User_ID	Unique identifier for each user.	Categorical	"USER-00001" to "USER-10000"
Age	Age of the user.	Numerical	18 - 65
Gender	Gender of the user.	Categorical	"Female", "Male", "Other"
Technology_Usage_Hours	Total hours of technology usage per day.	Numerical	1.0 - 12.0
Social_Media_Usage_Hours	Hours of social media usage per day.	Numerical	0.0 - 8.0
Gaming_Hours	Hours of gaming per day.	Numerical	0.0 - 5.0
Screen_Time_Hours	Total screen time (all devices) per day.	Numerical	1.0 - 15.0
Mental_Health_Status	General mental health status.	Categorical	"Good", "Poor", "Fair", "Excellent"
Stress_Level	User's level of stress.	Categorical	"Low", "High", "Medium"
Sleep_Hours	Average sleep hours per night.	Numerical	4.0 - 9.0
Physical_Activity_Hours	Daily hours of physical activity.	Numerical	0.0 - 10.0
Support_Systems_Access	Access to support systems (family, friends, therapy).	Categorical	"No", "Yes"
Work_Environment$_{I}$*mpact*	Impact of work environment on mental health.	Categorical	"Negative", "Positive", "Neutral"
Online_Support_Usage	Usage of online support systems for mental health (e.g., forums, teletherapy).	Categorical	"Yes", "No"

2.1.1 Covariance Matrix

To explore how the variables vary together, we calculate the covariance matrix in Tables 2 and 3. Each value in this matrix represents the covariance between two variables, showing how changes in one variable are associated with changes in another. The diagonal elements represent the variance of each variable.

Table 2: Covariance Matrix (Part 1)

	Age	Tech Usage	Social Media	Gaming	Screen Time	MH Status	Stress
Age	193.77	0.77	0.29	0.10	0.40	0.005	-0.11
Tech Usage	0.77	10.04	0.17	0.07	0.10	-0.03	-0.02
Social Media	0.29	0.17	5.35	0.02	-0.08	-0.01	0.03
Gaming	0.10	0.07	0.02	2.09	-0.05	0.009	-0.01
Screen Time	0.40	0.10	-0.08	-0.05	16.34	0.02	-0.05
MH Status	0.005	-0.03	-0.01	0.009	0.02	1.25	-0.0006
Stress	-0.11	-0.02	0.03	-0.01	-0.05	-0.0006	0.67
Sleep	-0.03	-0.05	0.01	0.02	-0.07	-0.02	0.01
Physical Activity	-0.18	0.09	0.02	-0.002	0.36	0.01	0.04
Work Impact	0.008	0.010	-0.0002	0.012	0.051	-0.007	0.002
Support Systems (Yes)	-0.01	0.003	-0.02	-0.005	0.02	-0.009	0.005
Online Support (Yes)	-0.02	-0.03	-0.01	-0.004	0.01	0.002	-0.005

Table 3: Covariance Matrix (Part 2)

	Sleep	Physical Activity	Work Impact	Support Systems (Yes)	Online Support (Yes)
Age	-0.03	-0.18	0.008	-0.01	-0.02
Tech Usage	-0.05	0.09	0.010	0.003	-0.03
Social Media	0.01	0.02	-0.0002	-0.02	-0.01
Gaming	0.02	-0.002	0.012	-0.005	-0.004
Screen Time	-0.07	0.36	0.051	0.02	0.01
MH Status	-0.02	0.01	-0.007	-0.009	0.002
Stress	0.01	0.04	0.002	0.005	-0.005
Sleep	2.11	-0.04	-0.03	0.007	-0.01
Physical Activity	-0.04	8.44	0.018	0.03	-0.03
Work Impact	-0.03	0.018	0.67	-0.003	0.001
Support Systems (Yes)	0.007	0.03	-0.003	0.25	-0.004
Online Support (Yes)	-0.01	-0.03	0.001	-0.004	0.25

Comments: The covariance matrix reveals that 'Age' has relatively high variance compared to other variables (193.77 on the diagonal). The covariances, such as between 'Screen Time Hours' and 'Physical Activity Hours' (0.36), indicate mild relationships between certain variables, suggesting some patterns that may be further investigated in the modeling process. High variances (such as for 'Screen Time Hours') indicate the diversity in the dataset for these metrics, which may impact model selection.

2.1.2 Correlation Matrix

The correlation matrix in Tables 4 and 5 measures the strength and direction of linear relationships between variables, with values ranging from -1 (perfect negative correlation)

to 1 (perfect positive correlation). This matrix provides a standardized view of associations that are independent of the scale of the variables.

Table 4: Correlation Matrix (Part 1)

	Age	Tech Usage	Social Media	Gaming	Screen Time	MH Status	Stress
Age	1.000	0.017	0.009	0.005	0.007	0.000	-0.010
Tech Usage	0.017	1.000	0.023	0.014	0.008	-0.007	-0.009
Social Media	0.009	0.023	1.000	0.006	-0.008	-0.005	0.014
Gaming	0.005	0.014	0.006	1.000	-0.008	0.005	-0.006
Screen Time	0.007	0.008	-0.008	-0.008	1.000	0.005	-0.016
MH Status	0.000	-0.007	-0.005	0.005	0.005	1.000	-0.001
Stress	-0.010	-0.009	0.014	-0.006	-0.016	-0.001	1.000
Sleep	-0.002	-0.010	0.004	0.010	-0.011	-0.009	0.011
Physical Activity	-0.005	0.010	0.002	-0.0004	0.031	0.004	0.016
Work Impact	0.001	0.004	-0.0001	0.010	0.015	-0.007	0.003
Support Systems (Yes)	-0.002	0.002	-0.017	-0.007	0.010	-0.016	0.012
Online Support (Yes)	-0.003	-0.016	-0.008	-0.006	0.005	0.004	-0.013

Table 5: Correlation Matrix (Part 2)

	Sleep Hours	Physical Activity	Work Impact	Support Systems (Yes)	Online Support (Yes)
Age	-0.002	-0.005	0.001	-0.002	-0.003
Tech Usage	-0.010	0.010	0.004	0.002	-0.016
Social Media	0.004	0.002	-0.0001	-0.017	-0.008
Gaming	0.010	-0.0004	0.010	-0.007	-0.006
Screen Time	-0.011	0.031	0.015	0.010	0.005
MH Status	-0.009	0.004	-0.007	-0.016	0.004
Stress	0.011	0.016	0.003	0.012	-0.013
Sleep	1.000	-0.010	-0.023	0.010	-0.019
Physical Activity	-0.010	1.000	0.008	0.022	-0.023
Work Impact	-0.023	0.008	1.000	-0.009	0.003
Support Systems (Yes)	0.010	0.022	-0.009	1.000	-0.014
Online Support (Yes)	-0.019	-0.023	0.003	-0.014	1.000

Comments: The correlation matrix indicates very weak correlations between most variables. For instance, 'Screen Time Hours' shows a slight positive correlation with 'Physical Activity Hours' (0.031), and 'Stress Level' has a minimal positive correlation with 'Social Media Usage Hours' (0.014). These weak relationships suggest that while there are some associations, no strong linear relationships are immediately apparent, indicating that nonlinear models may be beneficial in this analysis.

2.2 Data Preparation and Cleaning

The analysis started by exploring the data and looking for missing values or data inconsistencies. After looking at the data it was confirmed that no additional data cleaning was necessary since there were no missing values or data anomalies. This allowed the team to focus on feature selection and encoding techniques.

The Pandas library was imported for data manipulation and analysis in Python and a list of column names was created to only select the relevant fields for this project such as Age, Technology Usage Hours, Social Media Usage Hours, Screen Time Hours, Mental Health Status, and others. Then, a new data frame was created that only contained the desired columns for this project and the research questions.

Once the new data frame was created, the age column was transformed into categorical values, creating to facilitate demographic analysis. These age groups were defined as intervals (0-18, 18-25, 26-35, etc.) and labeled accordingly. This process allows to analyze different patterns and trends among different groups of people. For modeling purposes the new created column Age Category was included into the process mentioned below.

Next, label encoding was applied to convert ordinal values into numerical values that the models could interpret. We mapped each ordinal value to a numerical value based on the order. For instance, Stress Level was mapped so that "Low" became 1, "Medium" became 2, and "High" became 3. Similarly, Mental Health Status and Work Environment Impact were mapped to numerical scales reflecting their order.

Then, the remaining categorical columns were transformed by using One-hot encoding which converts categorical values into binary columns, where each category is represented by either a 0 or a 1. This process resulted in a data frame fully prepared with all categorical and ordinal variables encoded as numerical values, making it suitable for a classification model.

Multiple time related variables such as Technology Usage Hours and Social Media Usage Hours had different ranges which could disproportionately influence the model. To make sure that all variables contribute equally to the model feature scaling was applied to standardize the data. The standard scaler from the Scikit learn library was used to perform this task, which ensured that the algorithm will treat all features equally improving the overall

performance of the model.

Finally, a few interactions were added to capture relationships between features that could potentially influence the target variable. These aim to reveal potential effects that individual features alone might not be able to explain. By adding these interaction the model could better identify non-linear relationships within the data.

The following interaction terms were developed based on logical hypotheses about the relationships between features:

1. **Screen_Stress_Interaction**
    - **Formula**: Screen_Time_Hours * Stress_Level
    - **Purpose**: To explore the potential impact of high screen time on stress levels, hypothesizing that increased screen time might correlate with higher stress.
2. **Social_Physical_Interaction**
    - **Formula**: Social_Media_Usage_Hours * Physical_Activity_Hours
    - **Purpose**: To investigate whether balancing social media usage with physical activity could mitigate negative effects on mental health or stress.
3. **Stress_Sleep_Interaction**
    - **Formula**: Stress_Level * Sleep_Hours
    - **Purpose**: To examine how stress and sleep hours interact, given that higher stress levels might reduce sleep duration or quality.
4. **Work_Stress_Interaction**
    - **Formula**: Work_Environment_Impact * Stress_Level
    - **Purpose**: To capture how workplace environment factors combine with stress to influence mental health outcomes.

5. **Tech_Social_Interaction**

   - **Formula**: Technology_Usage_Hours * Social_Media_Usage_Hours
   - **Purpose**: To assess whether heavy use of technology and social media jointly contribute to stress or mental health challenges.

6. **Gaming_Stress_Interaction**

   - **Formula**: Gaming_Hours * Stress_Level
   - **Purpose**: To analyze the relationship between gaming hours and stress levels, hypothesizing that gaming might either exacerbate or alleviate stress.

7. **Support_Online_Interaction**

   - **Formula**: Support_Systems_Access * Online_Support_Usage
   - **Purpose**: To explore the combined effects of accessing support systems and online support on mental health.

8. **Age_Tech_Interaction**

   - **Formula**: Age_Category * Technology_Usage_Hours
   - **Purpose**: To examine how technology usage varies with age and whether this interaction affects mental health or stress outcomes.

3 Inferential Statistical Analysis

The inferential statistical analysis aims to draw conclusions about the broader population based on the sample data provided in the Mental Health and Technology Usage Dataset. By using hypothesis testing and other inferential techniques, this analysis investigates relationships between variables such as screen time, stress levels, physical activity, and mental health status. These statistical methods help determine whether observed patterns in the sample data are statistically significant or if they could have occurred by chance.

To address the research questions, we will apply three main statistical tests: ANOVA, t-tests, and chi-squared tests. Each method is briefly defined below:

- **ANOVA (Analysis of Variance)**: ANOVA is a statistical method used to determine whether there are significant differences between the means of three or more independent groups. It is suitable for comparing multiple categories within a variable, such as levels of screen time or physical activity, and assessing their impact on mental health outcomes.

- **t-tests**: A t-test is used to assess whether there is a significant difference between the means of two groups. For instance, we use t-tests to compare mental health outcomes between two levels of physical activity (e.g., low vs. high) or between users with and without access to support systems.

- **Chi-Squared Test**: The chi-squared test is a non-parametric test used to evaluate associations between categorical variables. It is helpful in determining whether variables such as support systems and work environment are associated with different mental health outcomes.

In the following subsections, we address each research question by formulating specific hypotheses and applying these statistical tests to evaluate the relationships between key variables.

3.1 Research Question 1: Relationship between Screen Time and Stress Levels

For this question, we aim to determine if there is a significant relationship between screen time and stress levels.

- **Null Hypothesis (H_0):** There is no significant relationship between screen time and stress levels.

- **Alternative Hypothesis (H_1):** There is a significant relationship between screen time and stress levels.

Since *Screen_Time_Hours* is a numerical variable with a wide range (1.0 to 15.0) and *Stress_Level* is categorical (Low, Medium, High), we can categorize *Screen_Time_Hours* into three groups (e.g., Low, Medium, High) to perform an **ANOVA** test. If we decide to analyze two groups (e.g., Low vs. High screen time), an independent sample **t-test** will be used to assess differences in mean stress levels.

3.2 Research Question 2: Impact of Physical Activity on Mental Health among High Social Media Users

This question investigates whether different levels of physical activity have a significant effect on mental health outcomes among high social media users.

- **Null Hypothesis (H_0):** Physical activity has no significant effect on mental health outcomes among high social media users.
- **Alternative Hypothesis (H_1):** Physical activity has a significant effect on mental health outcomes among high social media users.

In this case, we will focus on users who are high social media users, as indicated by *Social_Media_Usage_Hours*. With a range of 0.0 to 10.0, *Physical_Activity_Hours* can be divided into categories (e.g., Low, Medium, High). An **ANOVA** test is appropriate if we analyze three or more levels of physical activity. If there are only two levels (e.g., Low vs. High), we will use an independent sample **t-test** to evaluate differences in mental health outcomes.

3.3 Research Question 3: Influence of Support Systems and Work Environment on Mental Health Status

This question explores whether the presence of support systems and the perceived impact of the work environment are significantly associated with overall mental health status.

- **Null Hypothesis (H_0):** Support systems and work environment impact have no significant association with overall mental health status.

- **Alternative Hypothesis (H_1):** Support systems and work environment impact are significantly associated with overall mental health status.

Here, both *Support_Systems_Access* and *Work_Environment_Impact* are categorical variables, as is *Mental_Health_Status* (categorized as Good, Poor, Fair, Excellent). We will apply a **Chi-Squared Test** to evaluate associations between each of these variables and mental health status. This test will help determine whether having access to support systems or being in a positive, negative, or neutral work environment is associated with different mental health outcomes.

4 Machine Learning Modeling

To gain deeper insights into the relationship between technology usage and mental health, this project leverages machine learning techniques to create predictive and descriptive models. Machine learning models can help identify patterns and associations that are not immediately apparent through traditional statistical analysis. By applying both supervised and unsupervised learning approaches, this study aims to (1) predict mental health outcomes based on technology usage and lifestyle factors, and (2) identify clusters of individuals with similar digital behavior profiles and mental health characteristics.

This section outlines the objectives for each approach, justifies the selected models, and details their relevance to the project goals.

4.1 Objective 1: Predicting Mental Health Outcomes (Supervised Models)

The first objective focuses on using supervised learning to predict mental health outcomes, specifically *Mental Health Status* (Good, Poor, Fair, Excellent) and *Stress Level* (Low, Medium, High). Supervised learning involves training a model on labeled data to make predictions on new, unseen data. Given the nature of these categorical outcomes, the selected models aim to capture both simple and complex relationships between predictors such as *Technology Usage Hours*, *Sleep Hours*, and *Physical Activity Hours*.

Parametric and Non-Parametric Models: We incorporate both parametric and non-parametric models to balance interpretability with flexibility. Parametric models, such as Logistic Regression, make assumptions about the data distribution, which can simplify interpretation but may underperform if these assumptions do not align with the data. Non-parametric models, like Random Forest, adapt better to complex patterns without assuming a fixed data distribution but require careful tuning to avoid overfitting.

The supervised models selected for this study are:

4.1.1 Logistic Regression

Logistic Regression is a parametric classification model that assumes a logistic relationship between features and the probability of belonging to a specific class. This model is chosen for its simplicity and interpretability, making it valuable for understanding the influence of factors like *Screen Time Hours* and *Physical Activity Hours* on mental health categories. Logistic Regression will serve as a baseline model, providing a straightforward interpretation of each feature's contribution to the prediction.

4.1.2 Decision Trees and Ensemble Methods

Decision Trees are non-parametric models that partition the feature space based on decision rules. Here, Decision Trees will help capture non-linear relationships between variables like *Social Media Usage Hours* and *Mental Health Status*. However, single trees can be prone to overfitting, so we implement the following ensemble methods to improve stability and accuracy:

- **Random Forest**: Random Forest extends bagging by introducing feature randomness, selecting a random subset of features at each split to reduce correlation between trees. This method is particularly useful for capturing interactions between variables and will help improve predictive accuracy when dealing with factors like *Sleep Hours* and *Stress Level*.

4.2 Objective 2: Identifying Behavioral Clusters (Unsupervised Models)

The second objective is to identify clusters of individuals with similar technology usage and mental health profiles, which may provide insights into distinct user segments. Unsupervised learning is used here to discover patterns without predefined labels, helping us understand how different behaviors correlate with mental health outcomes.

The unsupervised models applied in this study are:

4.2.1 K-Means Clustering

K-Means Clustering is a partitioning algorithm that divides data into a predefined number of clusters based on feature similarity. This model will group users by attributes such as *Screen Time Hours*, *Social Media Usage Hours*, and *Physical Activity Hours*. By analyzing these clusters, we aim to identify profiles of technology usage that correspond to similar mental health statuses, potentially uncovering at-risk groups.

4.2.2 Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that transforms correlated variables into a set of uncorrelated components, capturing the main variance in the data. Here, PCA will be used to simplify complex relationships among features like *Technology Usage Hours* and *Support Systems Access*. This reduction allows for clearer visualization of clusters and helps identify prominent patterns in user behavior, which can then be analyzed alongside mental health indicators.

4.3 Model Evaluation and Interpretation

For supervised models, we will evaluate performance using metrics such as accuracy, F1-score, and ROC-AUC, with particular attention to the interpretability of each model. This will help determine the effectiveness of each model in predicting mental health outcomes based on technology usage and lifestyle variables.

For unsupervised models, we will examine cluster quality and cohesion to ensure meaningful segmentation. K-Means results will be visualized using PCA, facilitating interpretation of user segments with similar behaviors.

Together, these models will provide a comprehensive analysis, with supervised models predicting mental health outcomes and unsupervised models identifying behavioral clusters. This multi-faceted approach allows us to understand both individual risks and group patterns related to technology usage and mental health.

5 Results and Discussion

5.1 Inferential Statistical Analysis

5.1.1 Relationship between Screen Time and Stress Levels

The aim of this analysis is to determine if there is a significant relationship between *Screen Time Hours* and *Stress Level*. The ANOVA test was used to compare the mean screen time hours across different levels of stress (Low, Medium, and High).

Hypotheses:

- **Null Hypothesis (H_0):** There is no significant relationship between screen time and stress levels.
- **Alternative Hypothesis (H_1):** There is a significant relationship between screen time and stress levels.

Code Implementation:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
import pandas as pd
# Ensure Stress_Level is categorical and Screen_Time_Hours is
   numeric
df['Stress_Level'] = df['Stress_Level'].astype('category')
df['Screen_Time_Hours'] = pd.to_numeric(df['Screen_Time_Hours'],
   errors='coerce')
```

```
# Calculate means for each Stress_Level
mean_values = df.groupby('Stress_Level')['Screen_Time_Hours'].mean()
print(mean_values)
# Perform ANOVA
model = ols('Screen_Time_Hours ~ C(Stress_Level)', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
# Display ANOVA results
print(anova_table)
```

The following table shows the ANOVA results for screen time across the three stress levels.

Table 6: ANOVA results for Screen Time Hours by Stress Level

	sum_sq	**df**	**F**	**PR(>F)**
C(Stress_Level)	52.593039	2.0	1.609267	0.200086
Residual	163357.837508	9997.0	NaN	NaN

Interpretation:

The mean values of *Screen Time Hours* for each stress level category are as follows:

- High Stress: 7.92 hours
- Low Stress: 8.08 hours
- Medium Stress: 7.93 hours

These means represent the average daily screen time reported for each stress level. The mean for low stress (8.08) is slightly higher than that of the other two groups, but the difference between means is minimal. This suggests that, overall, screen time does not vary significantly with stress levels in this dataset. In other words, participants report similar screen time usage regardless of their stress level.

The ANOVA test results show an F-statistic of 1.609 with a p-value of 0.200. Since the p-value is greater than the standard significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant difference in screen time

hours between the different stress levels. Thus, based on this analysis, screen time does not appear to vary meaningfully with changes in stress level.

5.1.2 Impact of Physical Activity on Mental Health among High Social Media Users

This analysis aims to determine if there is a significant relationship between *Physical Activity Hours* and *Mental Health Status* among high social media users (those who report 5 or more hours per day on social media). An ANOVA test was conducted to compare mean physical activity hours across the mental health status categories (Excellent, Fair, Good, and Poor).

Hypotheses:

- **Null Hypothesis (H_0):** There is no significant relationship between physical activity and mental health status among high social media users.
- **Alternative Hypothesis (H_1):** There is a significant relationship between physical activity and mental health status among high social media users.

Code Implementation:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
import pandas as pd
# Filter data to include only high social media users
high_social_df = df[df['Social_Media_Usage_Hours'] > 4]
# Calculate and display mean physical activity hours for each mental
    health status
mean_values = high_social_df.groupby('Mental_Health_Status')['
   Physical_Activity_Hours'].mean()
print("\nMean Physical Activity Hours by Mental Health Status:")
print(mean_values)
# Perform ANOVA
model = ols('Physical_Activity_Hours ~ C(Mental_Health_Status)',
   data=high_social_df).fit()
```

```
anova_table = sm.stats.anova_lm(model, typ=2)
# Display ANOVA results
print(anova_table)
```

The following tables show the mean values of physical activity hours by mental health status and the ANOVA results for this analysis.

Table 7: ANOVA Results for Physical Activity Hours by Mental Health Status

	sum_sq	**df**	**F**	**PR(>F)**
C(Mental_Health_Status)	12.216	3.0	0.479	0.697
Residual	41917.100	4930.0	NaN	NaN

Interpretation:

The mean values of *Physical Activity Hours* for each mental health status category among high social media users are as follows:

- Excellent: 5.04 hours
- Fair: 4.92 hours
- Good: 4.97 hours
- Poor: 5.04 hours

These means reflect the average daily physical activity hours among high social media users within each mental health category. The means are quite similar across the four mental health categories, ranging from about 4.92 to 5.04 hours. This suggests that daily physical activity levels do not vary significantly among individuals with different mental health statuses in this group of high social media users, which aligns with the ANOVA result showing no significant differences.

The ANOVA test results show an F-statistic of 0.479 with a p-value of 0.697. Since the p-value is greater than the standard significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant difference in physical activity hours across the different mental health statuses for high social media users. Thus,

physical activity does not appear to vary meaningfully with changes in mental health status within this group.

5.1.3 Influence of Support Systems and Work Environment on Mental Health Status

This analysis evaluates the association between *Support Systems Access* and *Work Environment Impact* with *Mental Health Status*.

Hypotheses

- **Null Hypothesis (H_0):** There is no significant association between support systems access or work environment impact and mental health status.
- **Alternative Hypothesis (H_1):** There is a significant association between support systems access or work environment impact and mental health status.

Code Implementation:

```
import pandas as pd
import scipy.stats as stats

# Load the dataset
file_path = '/content/mental_health_and_technology_usage_2024.csv'
df = pd.read_csv(file_path)

# Ensure categorical data type for relevant variables
df['Support_Systems_Access'] = df['Support_Systems_Access'].astype('category')
df['Work_Environment_Impact'] = df['Work_Environment_Impact'].astype('category')
df['Mental_Health_Status'] = df['Mental_Health_Status'].astype('category')

# Contingency table for Support Systems Access and Mental Health Status
```

```
contingency_table_support = pd.crosstab(df['Support_Systems_Access'], df['Mental_Health_Status'])
chi2_support, p_support, dof_support, expected_support = stats.chi2_contingency(contingency_table_support)

# Contingency table for Work Environment Impact and Mental Health Status
contingency_table_work = pd.crosstab(df['Work_Environment_Impact'], df['Mental_Health_Status'])
chi2_work, p_work, dof_work, expected_work = stats.chi2_contingency(contingency_table_work)

# Display results
print("Chi-Squared Test for Support Systems Access and Mental Health Status:")
print(f"Chi2 Statistic: {chi2_support}, p-value: {p_support}")
print("\nExpected frequencies table for Support Systems Access and Mental Health Status:")
print(expected_support)

print("\nChi-Squared Test for Work Environment Impact and Mental Health Status:")
print(f"Chi2 Statistic: {chi2_work}, p-value: {p_work}")
print("\nExpected frequencies table for Work Environment Impact and Mental Health Status:")
print(expected_work)
```

Interpretation

- **Support Systems Access and Mental Health Status**:
  - **Chi-Squared Statistic**: 8.33
  - **p-value**: 0.040

Since the p-value (0.040) is below the significance level of 0.05, we reject the null hypothesis, suggesting a significant association between *Support Systems Access* and *Mental Health Status*. Individuals with access to support systems are more likely to experience different mental health outcomes, indicating that support systems may play a role in influencing mental health.

- **Work Environment Impact and Mental Health Status**:
  - **Chi-Squared Statistic**: 7.34
  - **p-value**: 0.291

  The p-value for the association between *Work Environment Impact* and *Mental Health Status* is 0.291, which is greater than the significance level of 0.05. Thus, we fail to reject the null hypothesis, indicating no significant association between the impact of the work environment and mental health status in this sample.

This section outlines the findings for the third research question, showing that support system access appears to have a notable effect on mental health status, while work environment impact does not display a statistically significant relationship.

5.2 Machine Learning Modeling

5.2.1 Supervised Models

Predicting Mental Health Status

Three machine learning models were implemented to predict Mental Health Status using the selected features: Logistic Regression, Decision Trees, and Random Forest . All model had similar performance, with low accuracy scores between 0.23 and 0.25.

Although there were small differences in precision, recall and F1 scores across all classes none of the models significantly outperformed the others, indicating limited predictive power of the dataset for this specific target variable.

Logistic Regression

Accuracy: 0.24 with constant but low precision and recall in all classes. The macro average F-1 score of 0.24 indicates difficulty in differencing mental health status categories.

Decision Trees

Accuracy: 0.23 with better recall for class 1 (0.47) but low performance for the other classes. This suggest that the Decision Tree overfit to certain patterns, while failing to generalize across all classes

Random Forest

Accuracy: 0.25 , the highest score achieved, with precision, recall and F-1 scores consistent around 0.25 for most classes. This ensemble method slightly improved performance compared to single trees.

Common trends

Low accuracy scores for all models: All models performed a little better than random guessing (which would yield 25 percent accuracy with 4 classes).

Balanced F-1 scores: Some models achieved slightly higher recall for some of the classes, and no model achieved a good performance for all the classes.

Class overlap: The low precision and recall for all classes suggest significant overlap in the feature space, which causes the model to struggle distinguishing between different categories of mental health status.

Potential Explanation

Limited predictive features: the available features may not capture the complexity of mental health status. Additional key features are missing.

Feature relationships: Nonlinear relationships among features might not have been fully explored. Several interactions were included in the models; however, this did not help with model performance.

Next steps and recommendations

Feature engineering: explore additional interactions or feature transformation. Moreover, including other relevant features might lead to model performance, for example, socioeconomic status or access to mental health care.

Predicting Stress Level

Three machine learning models were implemented to predict Stress Level: Decision Trees, Random Forest and Logistic Regression. The interaction columns that included Stress Level information were dropped from the dataset before modeling. The accuracy scores fall between 0.33 and 0.35, suggesting that the model had a hard time differentiating between stress levels accurately. Precision, recall and F1 scores were low, no model outperforming the others.

Decision Tree

Accuracy: 0.35 the highest of all models. Recall for class 2 was significantly higher at 0.65 indicating identified patterns specific to this stress level.

However, the precision and recall for all the other classes was way lower with a macro F-1 score of 0.32 indicating poor overall performance.

Random Forest

Accuracy: 0.33 with constant precision, recall, and F1-scores for all three classes at approximately 0.33.

The ensemble nature of Random Forest did not provide significant gains over the Decision Tree, suggesting that the dataset lacks the complexity for Random Forest to leverage.

Logistic Regression

Accuracy: 0.34, with constant F1 scores for all classes. The highest recall was for Class 3 at 0.44, but precision was low. Logistic Regression linear nature might struggle to capture complex relationships for this dataset.

Potential Explanation

Limited predictive power: None of the models outperformed the others. This suggests that its hard to to separate different stress levels in the feature space.

Class overlap: Although the class is not imbalanced, there seems to be significant overlap for different stress level classes, causing the model to misclassify often.

Non-linear relationships: The poor performance by logistic regression indicates that the relationship between the target variable and and features is insufficient to predict stress level

accurately.

Next steps and recommendations

Feature engineering and feature selection: explore additional interactions that might contribute meaningfully to the models. Also, looking to add other key features could lead to models performance given that the current features do not seem to be very good predictors for the target variable.

5.2.2 Unsupervised Models

K-Means Clustering and Principal Component Analysis (PCA)

Our goal is to group individuals based on their behavioral, mental health, and technology usage habits. K-means clustering and Principal Component Analysis (PCA) were both implemented to identify different groups within the data.

The PCA indicated that the variance was equally distributed among components. The first ten components explained these proportions: [10.64%, 10.45%, 10.20%, 10.08%, 8.67%, 8.56%, 6.69%, 6.28%, 5.39%, 5.38%]. This suggests that there is no single dominant feature in our dataset and that several factors contribute to the variability in the data. Clustering was implemented on those principal components using $k = 3$. The plot can be observed below.

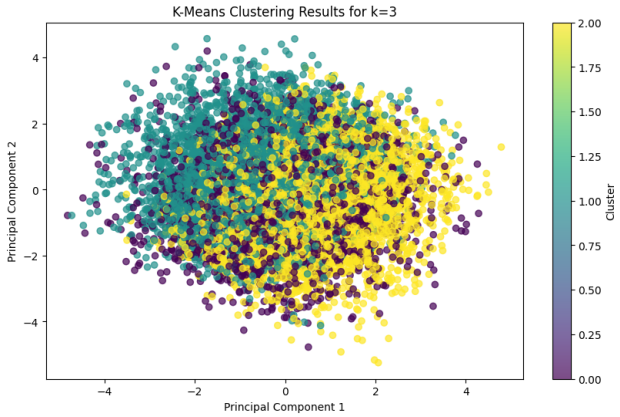

Figure 1: **Clustering of Individuals Based on PCA Projection and Key Features**

The cluster analysis showed the following group-level characteristics based on the feature means:

Cluster 0: Strong Social Support and Online Engagement

- **Support Systems:** This cluster shows stronger access to support systems (`Support_Systems_Access = 1.0`) and frequent usage of online support platforms (`Online_Support_Usage = 1.0`).
- **Stress and Screen Time:** Individuals in this group showed moderate stress levels and the highest screen time.
- **Work and Stress Interaction:** Lower `Work_Stress_Interaction` indicates that work environments in this cluster are less influenced by stress.

Cluster 1: Limited Social Support, Low Screen Time

- **Support Systems:** This group had limited access to support systems (`Support_Systems_Access = 0.334`) and less online presence.
- **Screen Time:** Noticeably lower screen time compared to other groups (`Screen_Time_Hours = -0.867`).
- **Stress Levels:** Moderate stress levels, with positive relationships between work environments and stress.

Cluster 2: High Screen Time, Moderate Social Support

- **Support Systems:** Moderate access to support systems (`Support_Systems_Access = 0.338`), similar to Cluster 1, but with higher online engagement.
- **Screen Time:** Higher screen time compared to other groups (`Screen_Time_Hours = 0.871`).
- **Work and Stress Interaction:** A positive relationship between work environment and stress indicates a more stress-influenced work context.

Implications:

This clustering result allows us to understand behavioral and mental health patterns:

- **Cluster 0:** These individuals, with strong support systems and moderate stress levels, may benefit from keeping their current engagement with online resources.
- **Cluster 1:** This group may need targeted interventions to enhance access to support systems and online resources.
- **Cluster 2:** Maintaining a balance between screen time and work-related stress could have a positive impact on individuals in this group.

By adding multiple interactions to our dataset, some relationships between features were identified, such as how screen time correlates with stress levels in different groups.

6 Conclusions

This project explores the complex relationships between technology usage and mental health, using statistical analysis and machine learning methods to find patterns and insights. While inferential analysis showed limited statistical relationships, like no significant connection between screen time and stress levels, the chi-squared test highlighted the relevance of support systems in influencing mental health outcomes. Machine learning models showed limited predictive accuracy, which suggest that more robust features are needed and a deeper exploration of non-linear relationships.

Cluster analysis provided insights about user behaviors, and helped identify groups with varying access to support systems, screen time habits, and work-related stress dynamics. These findings indicate potential ways pathways for targeted interventions to promote mental well-being in multiple demographic and behavioral segments.

Future research would require incorporating more features, such as socio-economic variables or qualitative data, in order to better capture the complexity of mental health predictors. By improving the dataset and applying advanced modeling techniques gives us the chance to improve predictive power and deepen our understanding of these important topic. Ultimately, this study highlights the potential of data-driven approaches to inform policies and strategies aimed at improving technology interactions and mental health outcomes.

7 The Authors

This section provides a brief overview of the authors' academic backgrounds, professional interests, and current research endeavors. Each author brings a unique perspective and expertise to the study, contributing to a well-rounded analysis of the topic.

Miguel Ángel Bravo Martínez Del Valle is a second-year student at San Diego State University, in the Big Data Analytics Master's program, set to graduate in Spring of 2025. He took a bachelor's in Electronics, Robotics, and Mechatronics Engineering at the University of Málaga, back in Spain. Miguel has an interest in Data Analytics, Machine Learning, and Artificial Intelligence. He is currently in two Research Groups, one in Data Visualization working with an Ireland company and he is also in the AI4Businnes team, starting his research in Spring 2024.

Email: miguelangelbravo2000@gmail.com

Fernanda Carrillo is currently pursuing a graduate degree in Big Data Analytics at San Diego State University, with an anticipated graduation in Spring 2025. She previously earned a bachelor's degree in International Business with a focus on Management Information Systems and an emphasis on English and North America, also from San Diego State. Fernanda brings experience from both the technology and finance sectors, with a primary focus on business intelligence roles. She is actively involved in the Metabolism of Cities Living Lab, where she contributes to synthesis science, data accessibility, and mentoring future scientists in alignment with the UN Sustainable Development Goals. Her interests lie in the intersection of machine learning engineering and using data science to drive healthcare and social impact initiatives.

Email: fernandacarrilloe@gmail.com

References

[1] W. Mushtaq, "Mental health and technology usage dataset," 2023, accessed: 2024-10-31. [Online]. Available: https://www.kaggle.com/datasets/waqi786/mental-health-and-technology-usage-dataset

A Code in GitHub

GitHub Repository