ISWE209L - Data mining

Assignment Report

Submitted by

23MIS1133 – Hema Priyadharshni G

23MIS1147 – Premkumar R

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

School of Computer Science and Engineering

Vellore Institute of Technology, Chennai

Oct 2025

Content

Problem Description

Smartphone use has become a necessary component of daily life in today's hyperconnected world, encompassing everything from work and education to communication and entertainment. But too much screen time and continual device use have raised worries about stress, digital addiction, and declining mental health.

The purpose of this project is to examine smartphone usage patterns and how they relate to indicators of mental health. The project divides users into groups according to how they use their devices, finds behavioural patterns among them using clustering and regression techniques, and uses machine learning algorithms to predict mental health scores.

Dataset

The dataset combines smartphone usage statistics and behavioural data, simulating metrics obtained from **Android Digital Wellbeing**, **iOS Screen Time**, and **self-reported lifestyle and mental health surveys**.

It provides insights into users' digital behaviour, lifestyle habits, and their possible correlations with mental well-being indicators such as stress, anxiety, and depression levels.

Attribute	Description
User_ID	Unique identifier for each user.
Country	Country of residence of the user.
Age	Age of the user in years.
Gender	Gender of the user (Male/Female/Other).
Occupation	User's profession or type of employment.
Education_Level	Highest level of education attained.
Income_USD	Monthly or annual income in U.S. dollars.
Daily_Screen_Time_Hours	Total average screen time per day (in hours).
Phone_Unlocks_Per_Day	Number of times the phone is unlocked daily.
Social_Media_Usage_Hours	Average daily time spent on social media apps (in hours).
Gaming_Usage_Hours	Average daily time spent on mobile games (in hours).
Streaming_Usage_Hours	Average daily time spent on streaming services like Netflix or YouTube (in hours).
Messaging_Usage_Hours	Average daily time spent on messaging platforms such as WhatsApp, Telegram, etc.
Work_Related_Usage_Hours	Daily hours spent on work or study-related applications.
Sleep_Hours	Average hours of sleep per day.
Physical_Activity_Hours	Time spent in physical or fitness-related activities daily.
Mental_Health_Score	Composite score representing overall mental well-being (higher score = better mental health).

Depression_Score	Self-reported or derived score indicating depression levels.
Anxiety_Score	Score representing anxiety levels of the user.
Stress_Level	Perceived stress level on a defined scale.
Relationship_Status	Relationship category (Single, Married, etc.).
Has_Children	Indicates whether the user has children (Yes/No).
Urban_or_Rural	Living area type (Urban/Rural).
Time_Spent_With_Family_Hours	Average daily time spent with family.
Online_Shopping_Hours	Average daily time spent on e-commerce platforms.
Internet_Connection_Type	Type of internet connection (Wi-Fi, Mobile Data, etc.).
Primary_Device_Brand	Brand of the main smartphone device used.
Has_Screen_Time_Management_App	Indicates whether the user uses screen-time limiting apps (Yes/No).
Self_Reported_Addiction_Level	User's self-assessed level of smartphone addiction (scale value).
Monthly_Data_Usage_GB	Total monthly mobile data usage in gigabytes.
Has_Night_Mode_On	Whether the user keeps night/dark mode enabled (Yes/No).
Age_First_Phone	Age at which the user got their first smartphone.
Push_Notifications_Per_Day	Average number of push notifications received daily.
Tech_Savviness_Score	Measure of the user's familiarity and comfort with technology.

Objectives

1. To identify different **user behavior clusters** based on smartphone usage patterns.
2. To understand how factors such as **screen time, sleep, and social media use** relate to mental wellbeing.
3. To build a **predictive model** that estimates mental health scores from usage data.
4. To generate **visual insights** through Tableau dashboards for better interpretation of user habits and risk profiles.
5. To support awareness and potential interventions in **digital wellbeing management**.

Pre-processing

1. **Data Cleaning:**
    - Checked for missing or inconsistent values and replaced or removed them appropriately.
    - Standardized categorical variables (e.g., "Gender", "Urban_or_Rural") for uniformity.
2. **Feature Selection:**

- Selected numeric behavioral features such as *Daily_Screen_Time_Hours*, *Phone_Unlocks_Per_Day*, *Sleep_Hours*, etc., for clustering and regression.

3. **Scaling:**
   - Applied **StandardScaler** to normalize numeric values so that high-range features (like unlocks/day) don't dominate smaller-scale features (like hours).
4. **Clustering Preparation:**
   - Extracted relevant usage features for **K-Means clustering** to group users based on their behavior patterns.
5. **Feature Engineering:**
   - Created derived fields such as *Predicted_Mental_Health_Score* from regression models and included *Cluster* labels for Tableau visualization.

Data Mining Algorithms Used

1. **Clustering using K-Means Algorithm:**

   The K-Means clustering algorithm was used to segment users based on their smartphone usage behaviour. Numeric features such as *Daily Screen Time*, *Phone Unlocks per Day*, *Social Media Usage Hours*, *Gaming Usage Hours*, *Sleep Hours*, and *Physical Activity Hours* were used to form the clusters.

   - The algorithm automatically grouped users into **four clusters** (Cluster 0 to Cluster 3), each representing a different digital behaviour pattern.
   - This helped identify patterns such as *digital burnout* and *healthy usage balance* among user groups.
2. **Regression using Random Forest Regressor:**

   The Random Forest regression model was applied to predict the **Mental Health Score** based on multiple smartphone usage features.

   - This ensemble method builds multiple decision trees and averages their results, reducing overfitting and improving prediction stability.
   - The top contributing predictors for mental health were:
     - Streaming_Usage_Hours
     - Daily_Screen_Time_Hours
     - Time_Spent_With_Family_Hours
     - Messaging_Usage_Hours
     - Physical_Activity_Hours

Evaluation Metrics

For the **Random Forest Regression** model, the following metrics were used to evaluate performance:

1. **R^2 Score (Coefficient of Determination):**
   - Measures how well the model explains the variance in the target variable (Mental Health Score).

    - Obtained **R^2 = -0.045**, indicating that the model has low predictive power — mental health may depend on more complex, non-linear, or external factors beyond digital usage.
2. **RMSE (Root Mean Squared Error):**
    - Measures the average prediction error magnitude.
    - Obtained **RMSE = 28.57**, suggesting moderate variation between predicted and actual mental health scores.

For **Clustering**, evaluation was done using **interpretative analysis**:

- The number of users per cluster was examined.
- Mean feature values across clusters were compared to interpret behavioral differences.

Interpretation

1. **High Screen Time Correlates with Poor Mental Health**

   Users in **Cluster 0** and **Cluster 3** show significantly higher *daily screen time* and *phone unlocks per day*, combined with lower *sleep hours* and *mental health scores*.

   → This indicates **digital burnout** and the negative impact of excessive smartphone use on mental wellbeing.

2. **Physical Activity and Family Time Improve Mental Health**

   Clusters with higher *Physical Activity Hours* and *Time Spent with Family Hours* (particularly **Cluster 2**) have **better mental health scores**.

   → Suggests that maintaining offline social connections and regular physical activity promotes better mental balance.

3. **Streaming and Messaging Usage Strongly Influence Mental Health**

   The **feature importance ranking** from the Random Forest model shows that *Streaming_Usage_Hours* and *Messaging_Usage_Hours* are among the **top predictors** of mental health.

   → Indicates that passive entertainment (streaming) and constant communication (messaging) patterns are strong behavioral signals for stress or relaxation balance.

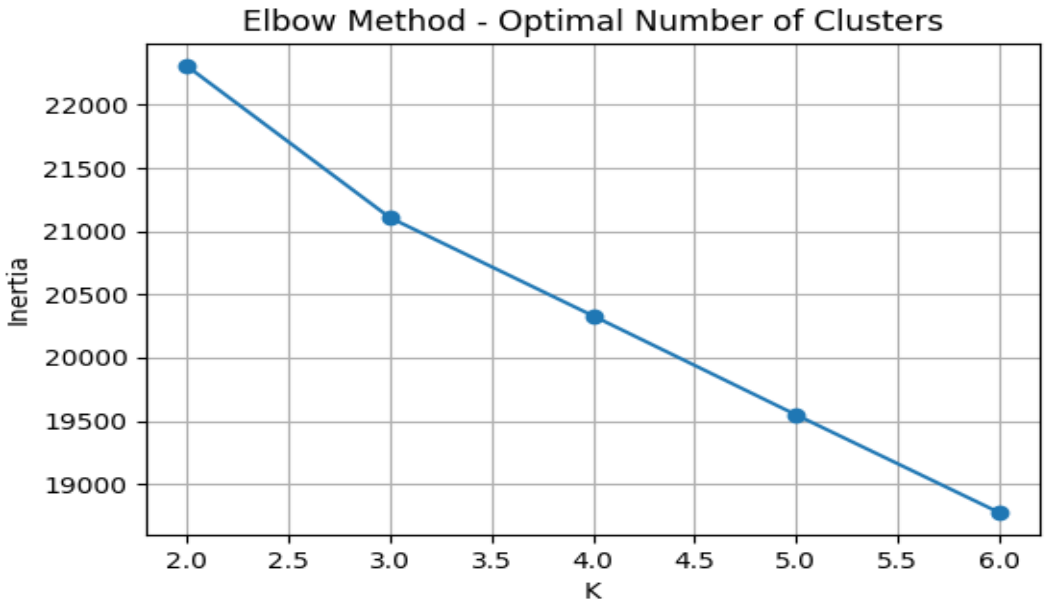
4. **Cluster-Based Behavior Segmentation Helps Identify At-Risk Groups**

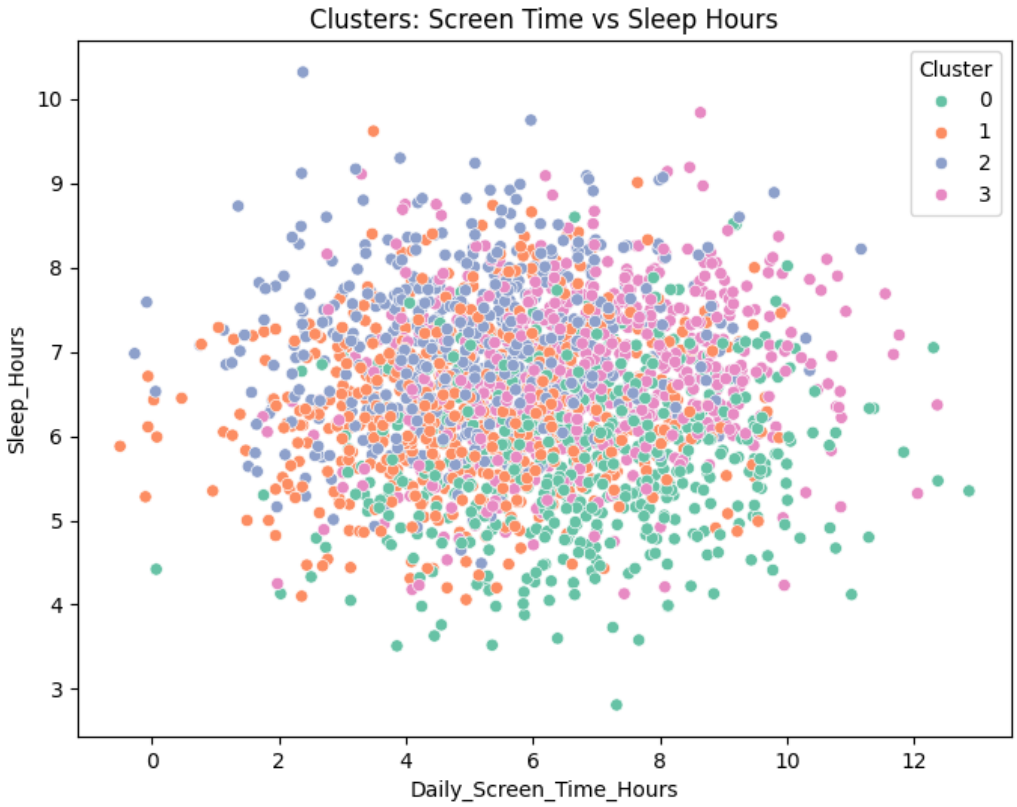
   The K-Means clustering effectively grouped users into **distinct lifestyle segments**, helping pinpoint risk categories like *Digital Burnout Users* (Cluster 0) versus *Healthy Balanced Users* (Cluster 2).

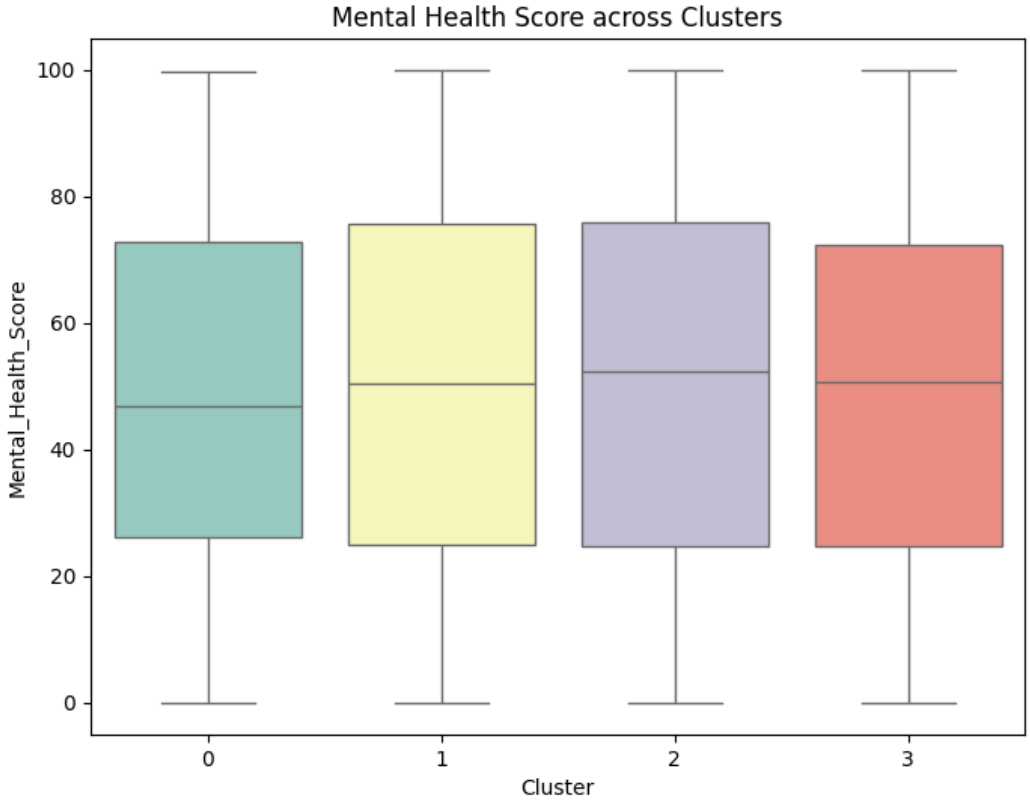
   → This segmentation can help design **targeted digital wellbeing interventions**, such as screen time reminders or wellness app recommendations.

```
hemapriyadharshni@Hemas-MacBook-Air Data Mining DA % python3 digitalwellbeing.py
✅ Dataset Loaded Successfully!
Columns in dataset:
 ['User_ID', 'Country', 'Age', 'Gender', 'Occupation', 'Education_Level', 'Income_USD', 'Daily_Screen_Time_Hours', 'Phone_Unlocks_Per_Day', 'Social_Media_Usage_Hours', 'Gaming_Usage_Hours', 'Streaming_Usage_Hours', 'Messaging_Usage_Hours', 'Work_Related_Usage_Hours', 'Sleep_Hours', 'Physical_Activity_Hours', 'Mental_Health_Score', 'Depression_Score', 'Anxiety_Score', 'Stress_Level', 'Relationship_Status', 'Has_Children', 'Urban_or_Rural', 'Time_Spent_With_Family_Hours', 'Online_Shopping_Hours', 'Internet_Connection_Type', 'Primary_Device_Brand', 'Has_Screen_Time_Management_App', 'Self_Reported_Addiction_Level', 'Monthly_Data_Usage_GB', 'Has_Night_Mode_On', 'Age_First_Phone', 'Push_Notifications_Per_Day', 'Tech_Savviness_Score']

✅ Numeric features selected for clustering:
 ['Daily_Screen_Time_Hours', 'Phone_Unlocks_Per_Day', 'Social_Media_Usage_Hours', 'Gaming_Usage_Hours', 'Streaming_Usage_Hours', 'Messaging_Usage_Hours', 'Work_Related_Usage_Hours', 'Sleep_Hours', 'Physical_Activity_Hours', 'Time_Spent_With_Family_Hours']

📊 Cluster Summary (Mean Values):
         Daily_Screen_Time_Hours  Phone_Unlocks_Per_Day  Social_Media_Usage_Hours  Gaming_Usage_Hours  ...  Sleep_Hours  Physical_Activity_Hours  Time_Spent_With_Family_Hours  Mental_Health_Score
Cluster                                                                            ...
0                          6.80                  73.56                      2.27                1.55  ...         5.79                     0.96                          0.84                48.74
1                          5.11                  78.17                      1.75                0.90  ...         6.35                     1.01                          1.47                50.60
2                          5.12                  76.74                      1.44                1.79  ...         7.05                     1.18                          1.37                50.73
3                          6.87                  89.94                      2.50                1.70  ...         6.75                     0.83                          2.07                48.95

[4 rows x 11 columns]

Number of users per cluster:
 Cluster
0    568
1    625
2    583
3    612
Name: count, dtype: int64
/Users/hemapriyadharshni/Documents/Data Mining DA/digitalwellbeing.py:79: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(x="Cluster", y="Mental_Health_Score", data=data, palette="Set3")

💡 Automatic Insights by Cluster:

Cluster 0 (568 users):
  • Digital Burnout: High screen time + low sleep.
  • Mental Health Concern: Lower mental wellbeing detected.

Cluster 1 (625 users):

Cluster 2 (583 users):

Cluster 3 (612 users):
  • Mental Health Concern: Lower mental wellbeing detected.

📊 Regression Model Evaluation:
R² Score: -0.045
RMSE: 28.57
```

```
Top Feature Importances:
Streaming_Usage_Hours          0.112735
Daily_Screen_Time_Hours        0.103475
Time_Spent_With_Family_Hours   0.100547
Messaging_Usage_Hours          0.099237
Physical_Activity_Hours        0.099174
Social_Media_Usage_Hours       0.099135
Gaming_Usage_Hours             0.097823
Sleep_Hours                    0.097788
Work_Related_Usage_Hours       0.096490
Phone_Unlocks_Per_Day          0.093597
dtype: float64

✅ CSV saved successfully for Tableau at:
/Users/hemapriyadharshni/Documents/Data Mining DA/processed_smartphone_usage.csv
👉 Columns available for Tableau Dashboard:
['User_ID', 'Country', 'Age', 'Gender', 'Occupation', 'Education_Level', 'Income_USD', 'Daily_Screen_Time_Hours', 'Phone_Unlocks_Per_Day', 'Social_Media_Usage_Hours', 'Gaming_Usage_Hours', 'Streaming_Usage_Hours', 'Messaging_Usage_Hours', 'Work_Related_Usage_Hours', 'Sleep_Hours', 'Physical_Activity_Hours', 'Mental_Health_Score', 'Depression_Score', 'Anxiety_Score', 'Stress_Level', 'Relationship_Status', 'Has_Children', 'Urban_or_Rural', 'Time_Spent_With_Family_Hours', 'Online_Shopping_Hours', 'Internet_Connection_Type', 'Primary_Device_Brand', 'Has_Screen_Time_Management_App', 'Self_Reported_Addiction_Level', 'Monthly_Data_Usage_GB', 'Has_Night_Mode_On', 'Age_First_Phone', 'Push_Notifications_Per_Day', 'Tech_Savviness_Score', 'Cluster', 'Predicted_Mental_Health_Score']
```




Clusters: Screen Time vs Sleep Hours
Cluster
0
1
2
3
Sleep_Hours
Daily_Screen_Time_Hours
10
9
8
7
6
5
4
3
0
2
4
6
8
10
12


Mental Health Score across Clusters
Mental_Health_Score
Cluster
100
80
60
40
20
0
0
1
2
3

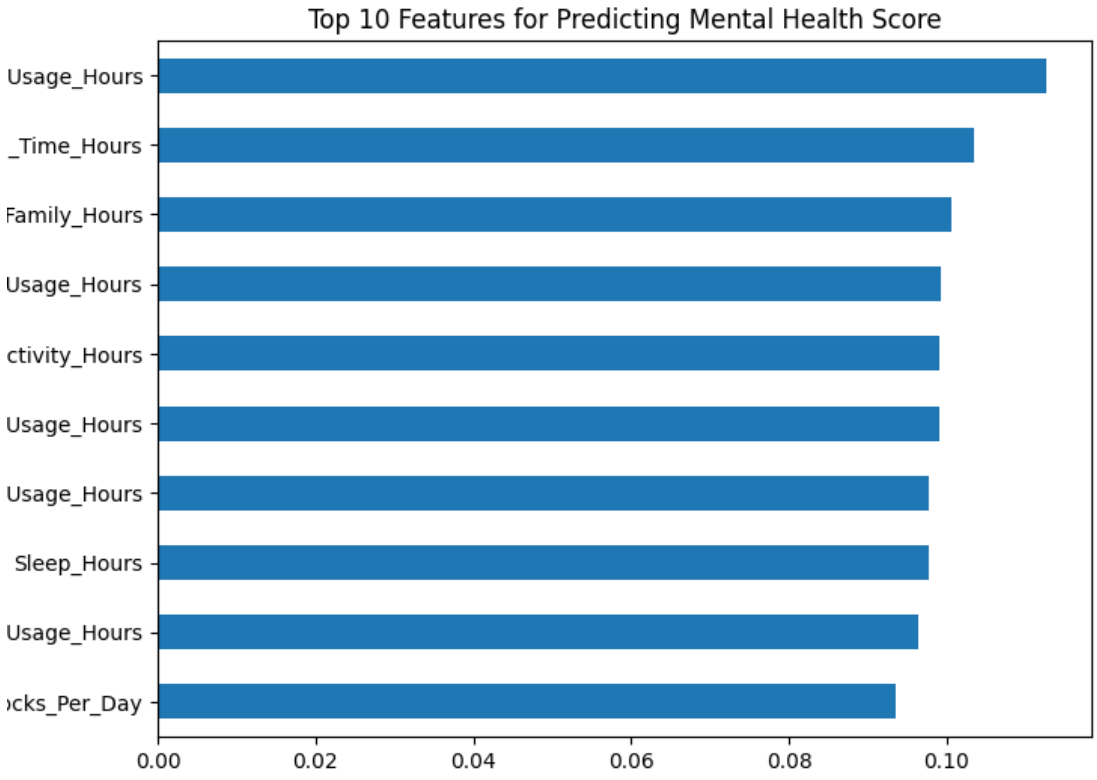
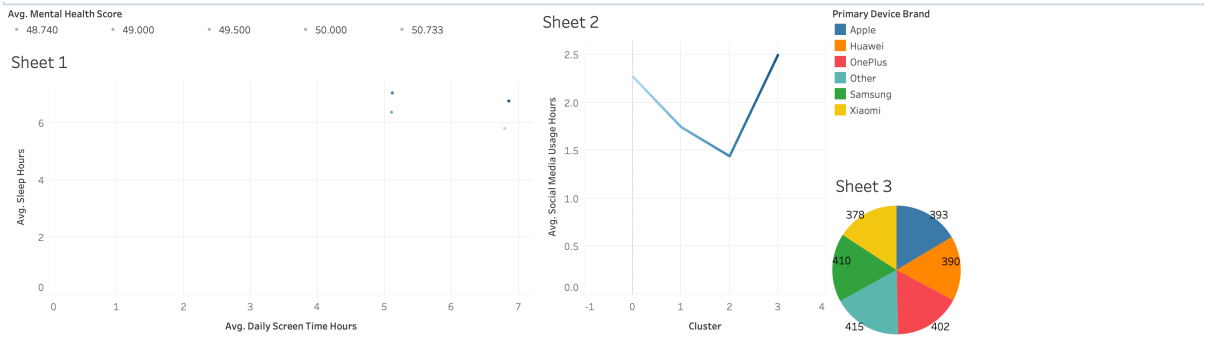


Tableau Dashboard



Summary:

- “Cluster 0 users show signs of digital burnout — high screen time, poor sleep, low mental health.”
- “Cluster 2 demonstrates a balanced lifestyle and strong mental well-being.”