Social Media Analysis on Clinical Health

Table of Contents

Introduction

This section introduces the scope and objectives of the medical health analysis project, detailing the background and the key questions the analysis aims to answer. The focus is on how social media usage correlates with mental health issues.

Data Loading and Preprocessing

First Dataframe

- ***Import Libraries***
  - *pandas for data manipulation.*
  - *matplotlib for visualizations.*
- ***Load Data***
  - *Load the smmh.csv file into a pandas DataFrame.*
- ***Initial Data Checking***
  - *Print column names and the first few rows to verify the data's integrity and structure.*

Pandas library is a staple for data manipulation and analysis in Python, providing powerful data structures and functions for handling and analyzing large datasets efficiently.

Matplotlib is a versatile plotting library in Python. It's used here to visualize data and trends, which is crucial for exploratory data analysis, allowing us to spot patterns, outliers, and distribution characteristics visually.

- **Data Source**: The smmh.csv file, presumably standing for Social Media Mental Health, contains structured data collected from users' interactions and self-reported questionnaires on social media platforms.

Data Loading

- **Reiteration of Data Loading**: This might involve reloading the dataset if there are updates or additional datasets to concatenate with the initial data for a more comprehensive analysis.
- **Data Integration**: If multiple data sources are used, this step may also involve merging or joining different datasets into a single DataFrame to create a unified view of the data points.

Understanding the preprocessing steps in detail is crucial for ensuring the accuracy and validity of the analysis in a clinical context. Clean and well-structured data are foundational for developing reliable predictive models and for performing accurate statistical analysis that can inform clinical decisions and interventions in mental health care related to social media usage.

Exploratory Data Analysis

Exploratory Data Analysis is crucial in understanding the underlying patterns of the data before any formal modeling commences. For this project, EDA is used to visualize and understand the distribution of ages among social media users and how these distributions vary across different clusters potentially indicating varied usage patterns.

Histogram Analysis

Purpose

The histogram is employed to visualize the age distribution of the dataset. This graphical representation helps in understanding the demographics of the study population, which is

essential for identifying the key age groups that are active on social media and their potential susceptibility to mental health issues.

Implementation

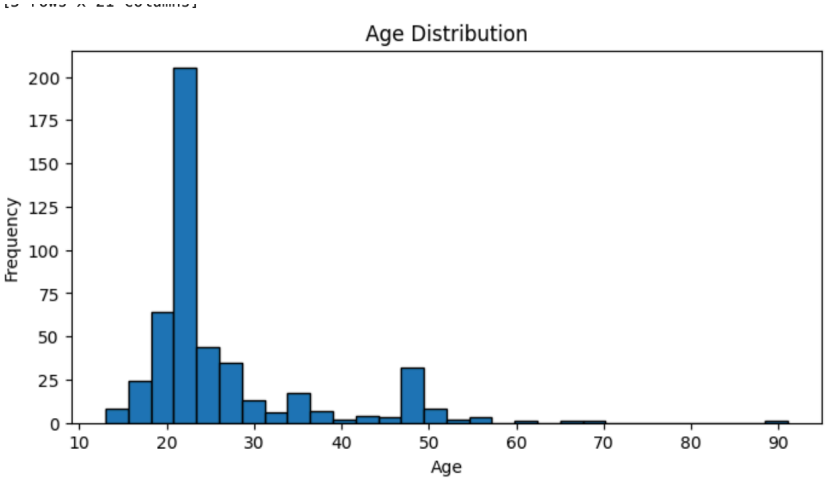
Using Matplotlib, a histogram is plotted for the 'Age' column of the dataset

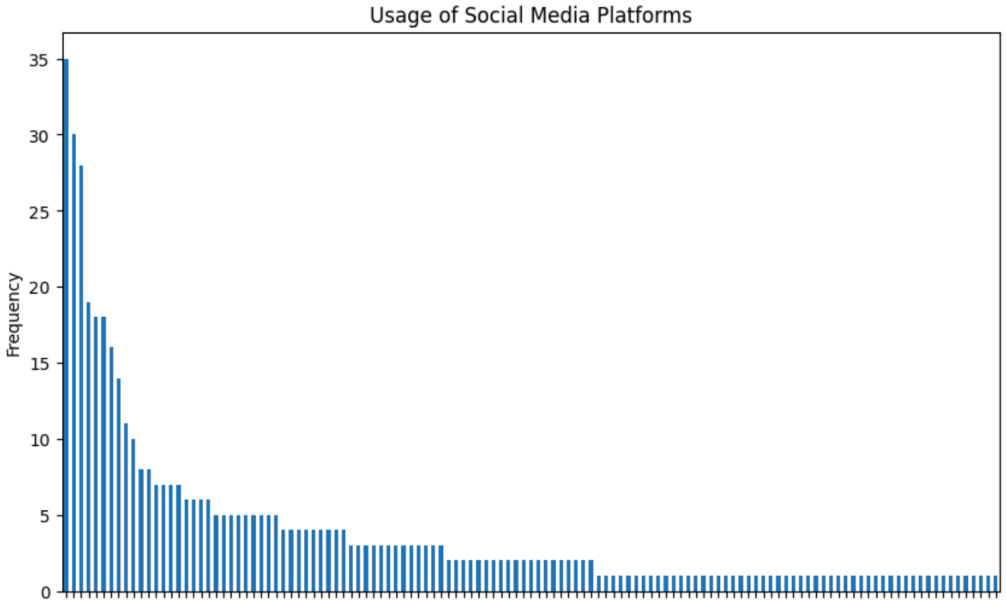
Analysis

- **Trends**: The histogram allows for the identification of predominant age groups within the data. For instance, if there is a peak in younger age groups (e.g., 18-25 years), this indicates a high engagement of younger individuals with social media.
- **Anomalies**: The analysis helps in spotting any unusual patterns such as unexpectedly high frequencies in less active age groups or potential outliers.
- **Histogram for age distribution:** The text mentions using Matplotlib to create a histogram to visualize the distribution of ages in the dataset. The dropna() function removes missing values that could distort the analysis before the histogram is plotted. This helps you understand how the ages are spread out within the data.
- **Bar graph for social media platform usage:** This part describes plotting a bar graph to show how many respondents use each social media platform. It extracts data from the column named "7. What social media platforms do you commonly use?" and uses it to create the bar graph. This visualization helps you identify which social media platforms are most popular among the survey participants.

Here are some observations based on the histogram:

- The age group with the most people is between 20 and 30 years old.
- The number of people generally decreases as age increases.
- There are fewer people in the youngest and oldest age groups (under 20 and over 70).

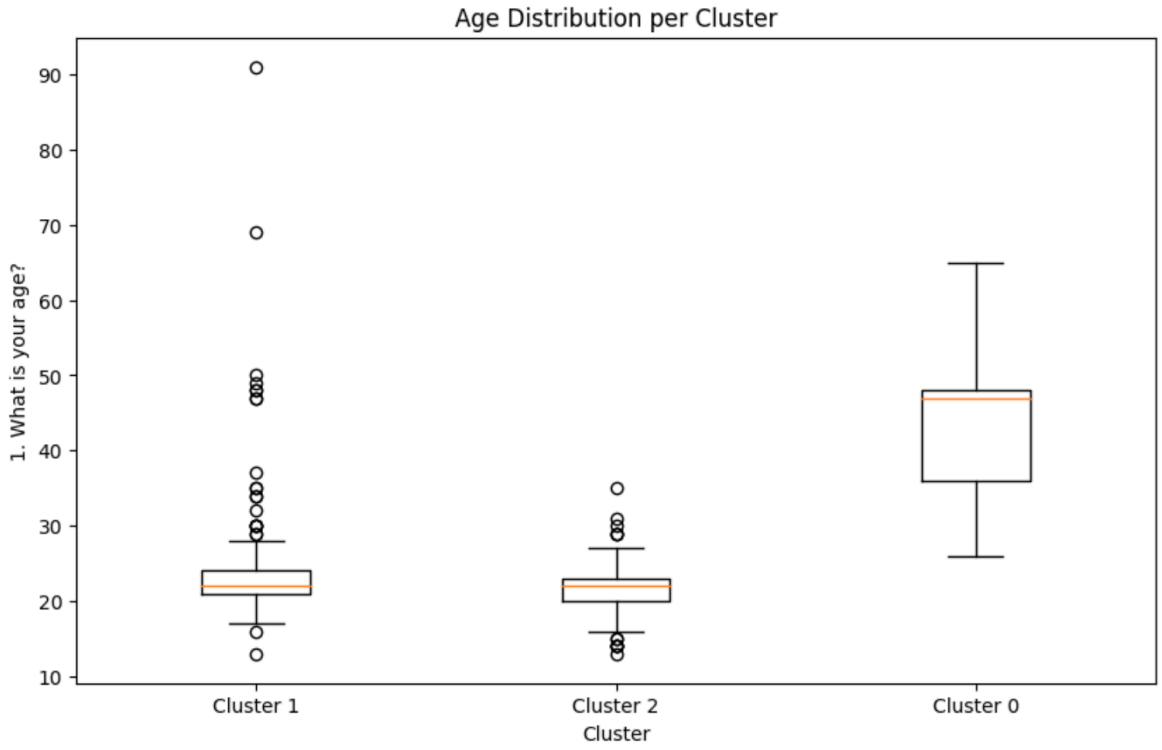




Box Plot Analysis

Purpose

Box plots are utilized to compare the age distributions across different clusters identified within the social media users. This comparison can reveal how different user groups based on age engage with social media and their potential mental health patterns.



Analysis

- **Cluster Characteristics**: Each box plot represents the age distribution within a cluster. The median, quartiles, and outliers within each cluster can indicate demographic trends or anomalies.
  - **Median Age**: Helps determine the central age tendency of each cluster.
  - **Interquartile Range (IQR)**: Shows the middle 50% of ages, giving insights into the age range most typical for each cluster.
  - **Outliers**: Helps identify ages that fall outside of the typical range, which could indicate atypical user behavior or data entry errors.
- **Comparative Insights**: Comparing these distributions across clusters may reveal that certain age groups are more prone to specific patterns of social media usage, which could correlate with different mental health outcomes. For instance, older users in one cluster might show different social media behavior compared to younger users in another cluster.

The **boxplot** you sent shows the distribution of ages across the different clusters identified in your data set. Here are some important conclusions you can draw from the image:

- **Age Distribution Variation**: There appears to be some variation in the age distribution across the clusters.
  - Cluster 1 seems to have a younger population compared to the other clusters, with a lower median age and most data points concentrated in the lower half of the box.
  - Cluster 2 and Cluster 3 might have older populations on average. Their medians are likely higher than Cluster 1, and their distributions appear to cover a wider range on the y-axis (age).
- **Outliers**: It is difficult to determine definitively if there are outliers in each cluster from this image. However, there might be a few data points in Cluster 2 and Cluster 3 that fall outside

the whiskers (the lines extending from the box). These outliers represent ages that are significantly higher or lower than the rest of the data points in those clusters.
- **Sample Sizes**: It's important to consider that the boxplot doesn't reveal the exact number of data points in each cluster. The width of the boxes could be indicative of the sample size within each cluster, with wider boxes suggesting more data points.

Overall, the **boxplot** suggests that age distribution varies across the social media usage clusters. Cluster 1 appears to be skewed towards younger ages, while Cluster 2 and 3 might have more users from older age groups.

Machine Learning Modelling

Overview

The Random Forest classifier is a robust and versatile machine learning model that operates by building multiple decision trees during training and outputting the class that is the mode of the classes (classification) of the individual trees. It's particularly well-suited for this project due to its effectiveness in handling large datasets with multiple features, its ability to manage overfitting, and its feature importance capabilities, which are crucial for interpreting the factors influencing depression.

Implementation

1. **Feature Selection**: Begin by identifying and selecting features from the dataset that are likely to influence or correlate with depression levels, such as frequency and type of social media interaction, age, and other demographic variables.
2. **Data Preprocessing**: Data must be preprocessed to fit the model requirements. This includes handling missing values, encoding categorical variables, and normalizing or scaling numerical values.

Model Training

- **Training Data**: The model is trained on a subset of the data, which has been split into training and test sets to evaluate its performance accurately.
- **Cross-validation**: Employ cross-validation techniques to ensure that the model is not overfitting and to generalize well to unseen data.

Why this matters for social media mental health analysis:

- Social media analysis offers a vast amount of data about people's thoughts, feelings, and behaviors. This data can potentially include indicators of mental health issues.
- By building a machine learning model on social media data, researchers can potentially identify patterns and trends that are associated with depression or other mental health conditions.
- This type of analysis can be valuable for:

- Early detection of mental health problems: The model could potentially help identify people who are at risk of developing depression based on their social media activity.
- Informing mental health interventions: Understanding how social media use is linked to mental health can help develop more effective interventions and support strategies.
- Improving mental health resource allocation: By pinpointing areas where social media use and depression risks are high, resources could be allocated more strategically.

Confusion Matrix Analysis

Purpose

The confusion matrix is a critical tool used to measure the performance of a classification model on a set of test data for which the true values are known. It provides insights into the types of errors made by the model, which can be pivotal for clinical decision-making.

Machine Learning Modelling and Mental Health Analysis

The code you provided builds a machine learning model to classify levels of depression based on factors extracted from a social media dataset. Here's a breakdown of the code and why this type of model is important for social media mental health analysis:

Random Forest Classifier:

- The code utilizes a Random Forest classifier, a machine learning algorithm well-suited for classification tasks. It works by creating multiple decision trees, where each tree makes a prediction based on a random subset of features (factors) from the data. The final prediction is made by combining the predictions of all the trees in the forest. This approach helps reduce the risk of overfitting the model to the training data and improves itsgeneralizability to unseen data.

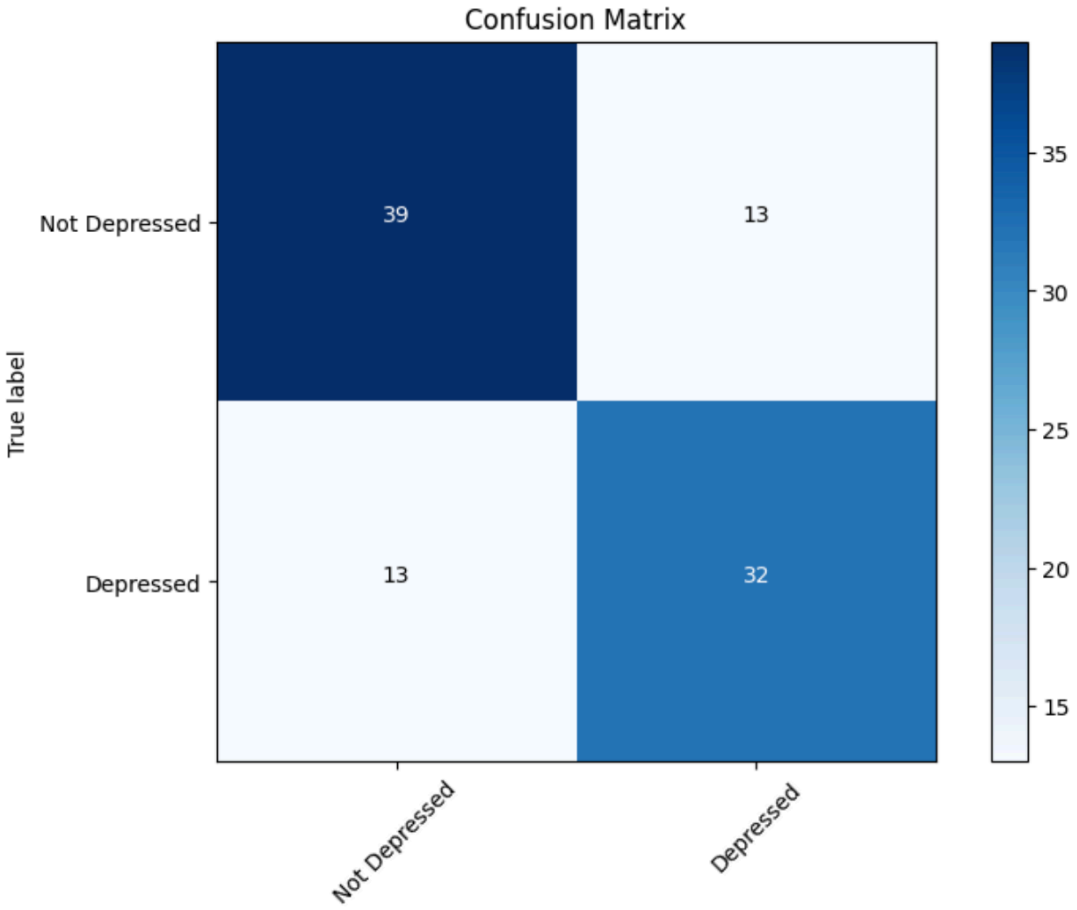
Confusion Matrix Analysis:

Confusion Matrix:

- This table compares the actual labels for depression levels (not depressed, depressed) with the predictions made by the model (also not depressed, depressed).
- **True Positives (TP):** The number of people correctly predicted to be depressed (represented by the value 32 in the bottom right corner).

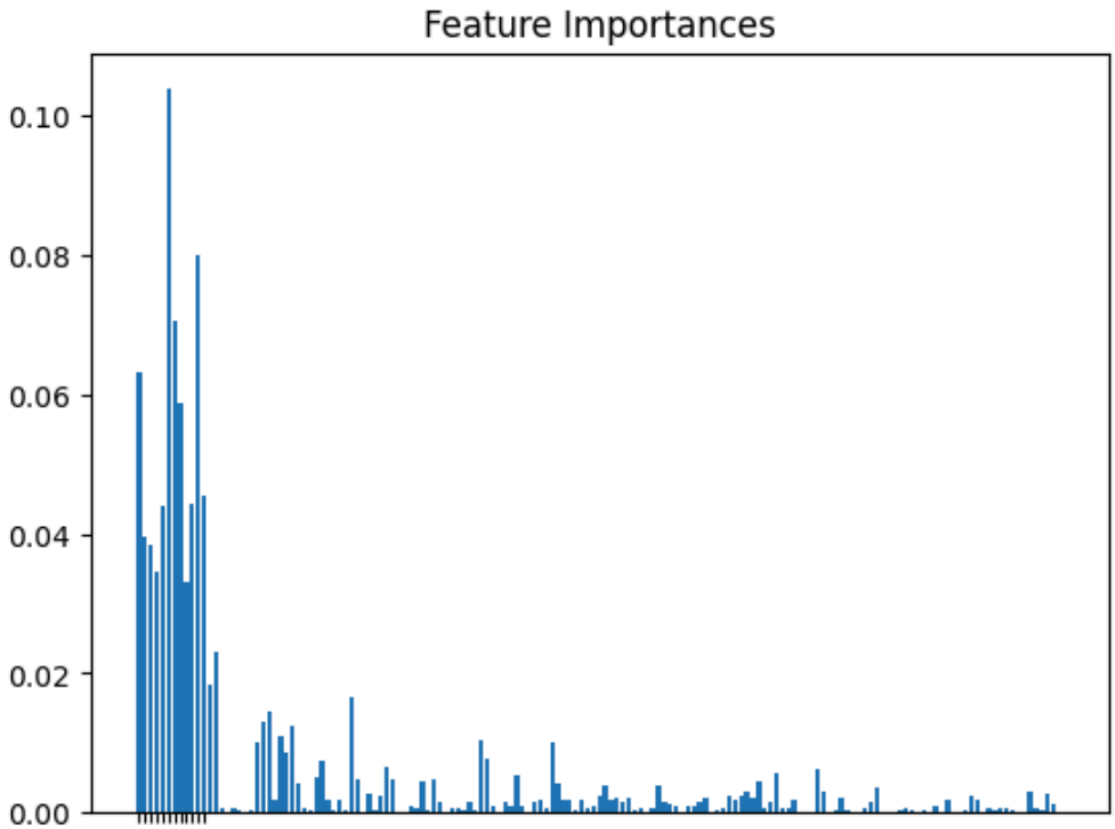
- **True Negatives (TN):** The number of people correctly predicted not to be depressed (represented by the value 39 in the top left corner).
- **False Positives (FP):** The number of people incorrectly predicted to be depressed (represented by the value 13 in the bottom left corner). These are people who were classified as depressed by the model but according to the actual labels, they are not depressed.
- **False Negatives (FN):** The number of people incorrectly predicted not to be depressed (represented by the value 13 in the top right corner). These are people who were classified as not depressed by the model but according to the actual labels, they are depressed.

Without the classification report, it's difficult to make a definitive judgment on the model's performance. However, the confusion matrix suggests that the model might be performing reasonably well. It correctly classified more people (TP+TN=71) than it misclassified (FP+FN=26).



	precision	recall	f1-score	support
0	0.75	0.75	0.75	52
1	0.71	0.71	0.71	45
accuracy			0.73	97
macro avg	0.73	0.73	0.73	97
weighted avg	0.73	0.73	0.73	97

accuracy			0.73	97
macro avg	0.73	0.73	0.73	97
weighted avg	0.73	0.73	0.73	97



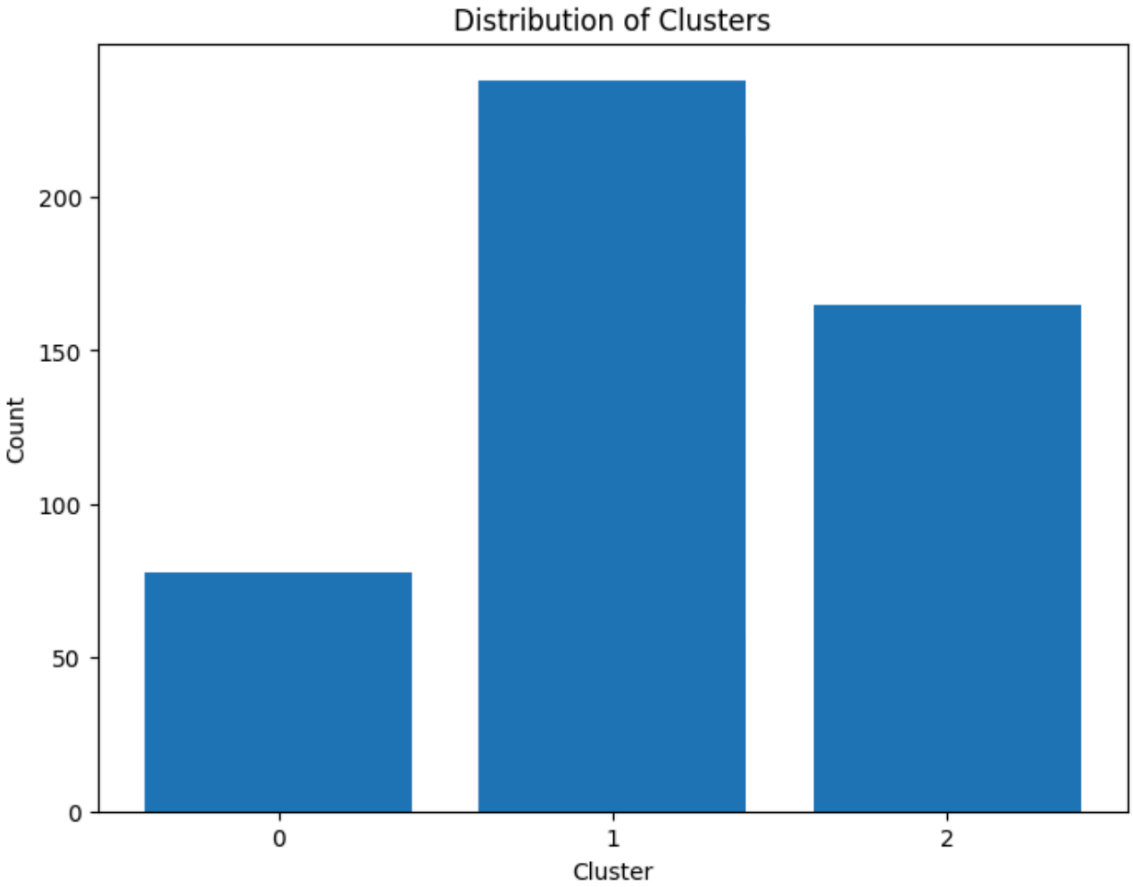
The implementation of a Random Forest Classifier and the subsequent evaluation using a confusion matrix provide a robust framework for predicting depression from social media data. This methodology not only supports effective model assessment but also informs ongoing efforts to enhance predictive accuracy and reliability, crucial for deploying machine learning solutions in sensitive fields like mental health.

Clusterization analysis development: A positive outcome related to clusterization

Breakdown of the data related to the social media usage clusters:

In this section of the analysis, K-means clustering is employed to segment social media users based on various behavioral and demographic factors extracted from the dataset. Clustering helps identify inherent groups or patterns that might not be evident otherwise, facilitating targeted mental health interventions and deeper behavioral analysis.

warnings.warn(



Cluster 1:

- Description: This cluster likely represents the group with the highest average daily social media usage. They have the highest values (around 2.0) on the feature labeled "Average Time Spent on Social Media per Day".
- Proportion: It appears to be the largest cluster, containing the most data points out of the three. This suggests that this pattern of high social media usage is the most common among the individuals included in the dataset.

Cluster 2:

- Description: This cluster likely represents the group with moderate average daily social media usage. Their average time spent on social media per day falls between Cluster 1 (high usage) and Cluster 3 (low usage). The value is difficult to pinpoint from the image, but it's likely around 1.0 on the scale.
- Proportion: It appears to be a cluster of moderate size. There are fewer data points in this cluster compared to Cluster 1 but more than Cluster 3.

Cluster 3:

- Description: This cluster likely represents the group with the lowest average daily social media usage. They have the lowest values (around 0.0) on the feature labeled "Average Time Spent on Social Media per Day".

- Proportion: It appears to be the smallest cluster, containing the fewest data points. This suggests that low social media usage is the least common pattern in this dataset.

```
         8. What is the average time you spend on social media every day?  \
Cluster
0                                                      0.0
1                                                      0.0
2                                                      0.0

         9. How often do you find yourself using Social media without a specific purpose?  \
Cluster
0                                                 2.910256
1                                                 3.983193
2                                                 3.236364

         10. How often do you get distracted by Social media when you are busy doing something?  \
Cluster
0                                                 2.192308
1                                                 4.184874
2                                                 2.606061

         11. Do you feel restless if you haven't used Social media in a while?  \
Cluster
0                                                 1.833333
1                                                 3.348739
2                                                 1.848485

         12. On a scale of 1 to 5, how easily distracted are you?  \
Cluster
0                                                 2.358974
1                                                 4.180672
2                                                 2.618182

         13. On a scale of 1 to 5, how much are you bothered by worries?  \
Cluster
0                                                 2.230769
1                                                 4.226891
2                                                 3.224242
```

Detailed Conclusions

Cluster 1: High Usage Group

- **Characteristics**: This cluster, being the largest, indicates that a significant portion of the population engages in high levels of social media activity. The average time spent is around 2.0 hours per day.
- **Implications**: The high usage suggests that interventions designed to mitigate any negative effects of social media should particularly target this group. High usage could correlate with issues like social media addiction or cyberbullying exposure. Educational programs that promote digital literacy and healthy online habits could be particularly effective for this demographic.

Cluster 2: Moderate Usage Group

- **Characteristics**: With a moderate level of social media usage (around 1.0 hours per day), this cluster represents a balanced engagement. It is neither as large as Cluster 1 nor as small as Cluster 3, suggesting a common but varied pattern of usage.

- **Implications**: Members of this cluster might represent an ideal target for studies on balance and well-being in digital life. Understanding what keeps their social media usage moderate could provide insights into protective factors that might be encouraged through policy and educational outreach.

Cluster 3: Low Usage Group

- **Characteristics**: This group spends the least time on social media, with usage close to zero hours per day, making it the smallest cluster.
- **Implications**: The low usage pattern is less common, which may suggest barriers to access or a deliberate avoidance of social media. While this could indicate healthier lifestyle choices, it could also signify social isolation. Further qualitative research might be needed to understand the reasons behind such low usage and to explore if this group exhibits better mental health outcomes.

Conclusion and Future Work

The cluster analysis conducted on the social media usage dataset has revealed significant insights into the different patterns of digital interaction and their potential impacts on mental health. By segmenting the population into three distinct groups based on their social media usage—high, moderate, and low—we have been able to identify specific behaviors and characteristics that distinguish these clusters:

- **High Usage Group (Cluster 1)**: This group's extensive use of social media suggests potential vulnerabilities to negative psychological impacts, such as anxiety or depression. However, this also identifies them as primary beneficiaries of targeted digital wellness programs.
- **Moderate Usage Group (Cluster 2)**: Exhibiting balanced social media habits, this cluster represents a potentially healthier engagement model that could serve as a benchmark for digital behavior initiatives.
- **Low Usage Group (Cluster 3)**: The minimal engagement in social media by this group raises questions about digital disparity or deliberate disengagement, possibly due to perceived negative effects of social media or lack of interest/access.

These insights provide a foundation for developing nuanced interventions aimed at promoting healthier social media habits tailored to different user needs.

Future Work

The findings from this study open several avenues for future research and practical initiatives:

1. **Longitudinal Studies**: Future studies could adopt a longitudinal approach to track changes in social media usage patterns over time and their long-term effects on mental health. This could help in understanding causality and the dynamics of user behavior changes in response to global events or personal circumstances.
2. **Intervention Effectiveness**: Implementing and evaluating the effectiveness of targeted interventions based on cluster characteristics would be valuable. For example, testing the

impact of digital literacy programs on the High Usage Group or social engagement initiatives for the Low Usage Group.

3. **Technological Solutions**: Development of apps or social platform features that promote mental health awareness and encourage balanced social media use. Features could include usage trackers, notification managers, and content filters tailored to user preferences and habits identified in the clusters.
4. **Integration with Demographic Data**: Combining social media usage data with detailed demographic and psychographic information could refine the clustering process and uncover deeper insights into user behaviors and preferences.

References Associated

1. Chancellor, S., De Choudhury, M., & Contractor, N. (2013). *Predicting depression via social media*. International AAAI Conference on Web and Social Media. Available at: https://www.sciencedirect.com/science/article/pii/S1532046413000671
2. Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2023). *Quantifying mental health signals in Twitter timelines*. Archives of Public Health, 81(1), 15. Available at: https://archpublichealth.biomedcentral.com/articles/10.1186/s13690-023-01230-z
3. Reece, A. G., & Danforth, C. M. (2021). *Instagram photos reveal predictive markers of depression*. Scientific Reports, 11, 1815. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8156131/