CS 483: The Scroll Patrol - Impacts of Social Media Habits

Nathan Hittesdorf
Jash Shah
Svetlana Voda
Sebastian Barroso
Megha Nayer

Introduction

Social media addiction has become a significant issue, negatively impacting mental health, productivity, and overall well-being. Despite widespread acknowledgment of this problem, current efforts to address it often lack data-driven insights and fail to leverage the power of advanced analytics to uncover actionable patterns. This report presents a comprehensive approach to tackling social media addiction using Big Data mining techniques to analyze behavior patterns, identify key factors contributing to addiction, and propose targeted interventions.

To achieve these goals, we utilized an extensive set of analytical methods and machine learning techniques. Exploratory Data Analysis (EDA) was conducted to uncover trends and correlations within the dataset. Clustering techniques, including Principal Component Analysis (PCA), Agglomerative Clustering, and Locality-Sensitive Hashing (LSH) with Minhashing, were applied to group users based on their behaviors and preferences. Additionally, A recommendation system was developed to further enhance engagement and propose alternatives to addictive patterns.

For predictive modeling, we employed Linear Regression, Ridge Regression, and Random Forest Regression to quantify the relationship between addiction levels and key features, providing insights into actionable predictors. Classification models, such as Random Forest and Naive Bayes, were applied to categorize users based on their risk of addiction. Additionally, Graph Network Analysis using PageRank helped us identify influential users and connections within the social media ecosystem, shedding light on how content propagates and influences addictive behaviors.

This combination of methods allows us to generate data-driven solutions for reducing social media addiction. By analyzing patterns and identifying key contributors, we aim to provide actionable insights that empower stakeholders to design interventions that enhance mental health and productivity.

Data

Dataset Description

The dataset captures user demographic details, platform engagement behaviors, and metrics related to social media addiction. Key attributes include:

- **Demographics:** Age, gender, location, income, profession, and property ownership.
- **Platform Usage:** Total time spent, number of sessions, and video interactions (category, length, engagement).
- **Behavioral Metrics:** Addiction level, scroll rate, satisfaction, and watch habits (time, reason, activity).
- **Device Information:** Device type, operating system, and connection type.

This diverse set of features provides the foundation for identifying patterns and clustering users based on risk factors.

Preprocessing

To prepare the dataset for clustering and analysis, the following preprocessing steps were implemented:

1. **Data Cleaning:**
    - Removed irrelevant or redundant columns such as Video ID, ProductivityLoss, and CurrentActivity.
    - Addressed missing values to ensure data consistency.
2. **Categorical Encoding:**
    - Transformed categorical variables into numeric formats.
    - For some analyses, features were encoded using **one-hot encoding** to represent categorical variables as binary vectors, ensuring compatibility with algorithms sensitive to feature scaling.
3. **Feature Scaling:**

- Standardized numerical features to ensure uniformity in feature contributions during clustering.

Cleaning and preparing the data this way facilitated the performance of our analyses.

Exploratory Data Analysis

The exploratory data analysis phase focused on understanding the structure and distribution of the data, identifying relationships between key variables, and validating assumptions for clustering analysis.

Initial Hypothesis

We hypothesized that **income** would be the most significant factor correlated with addiction levels and that addiction profiles would differ across demographic groups (e.g., gender, profession).

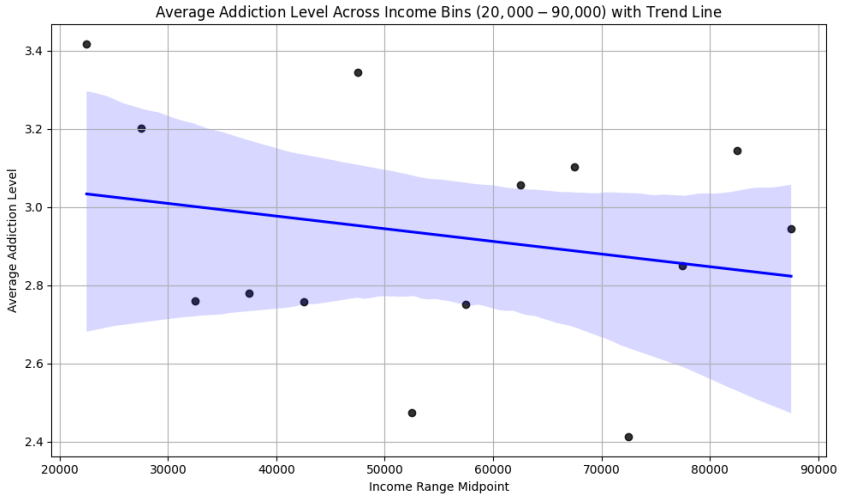
Findings

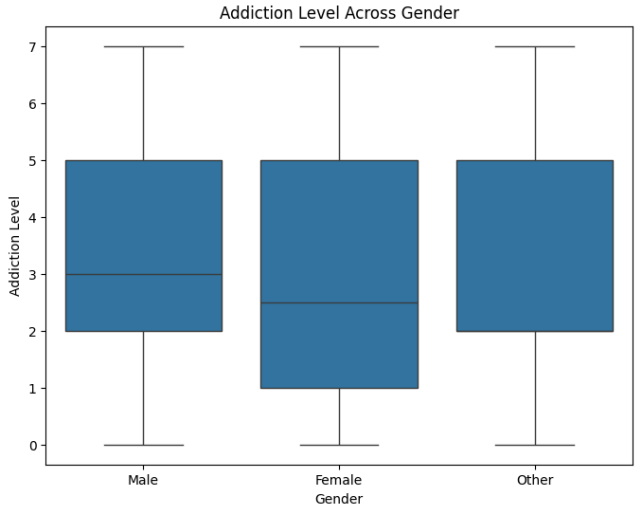
1. **Correlations:**
   - Most correlations between addiction level and demographic variables (e.g., income, gender, profession) were found to be **weak or negligible**.
   - Addiction levels were similar across genders and professions, challenging our initial hypothesis.
2. **Platform Analysis:**
   - **Facebook** emerged as the most addictive platform, though the difference was marginal compared to others.
3. **Satisfaction and Addiction:**
   - A nearly perfect linear correlation was observed between satisfaction and addiction.
   - Upon investigation, this was identified as an artifact of synthetic data, requiring the exclusion of satisfaction from predictive analysis.
4. **Distribution of Addiction Levels:**
   - Addiction levels were found to be **non-normally distributed**, necessitating adjustments in clustering methods to handle skewed data.

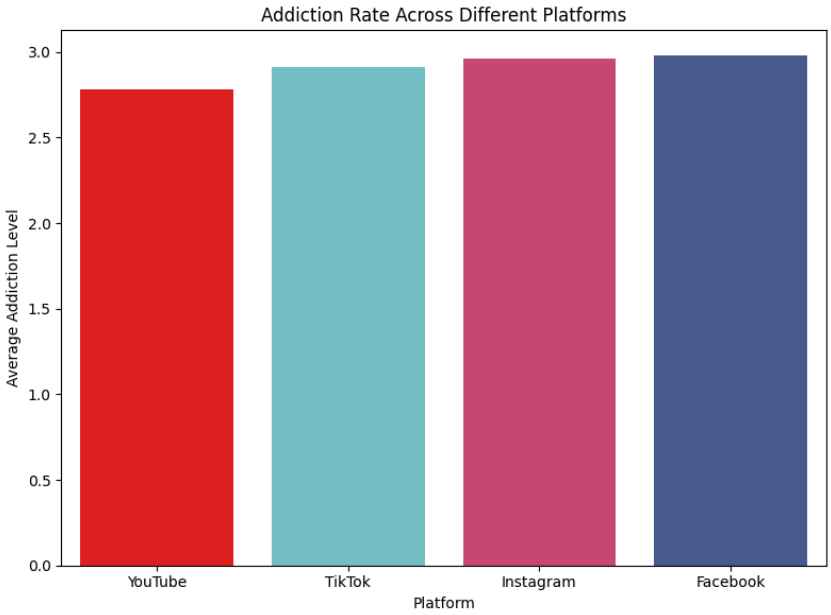
Visualizations and Insights

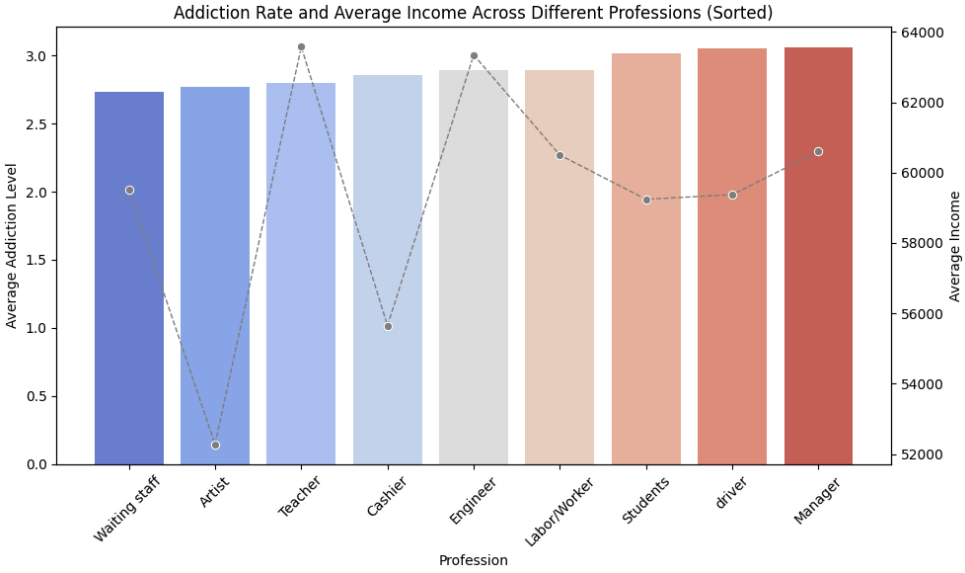
The analysis included detailed visualizations of key relationships, such as:

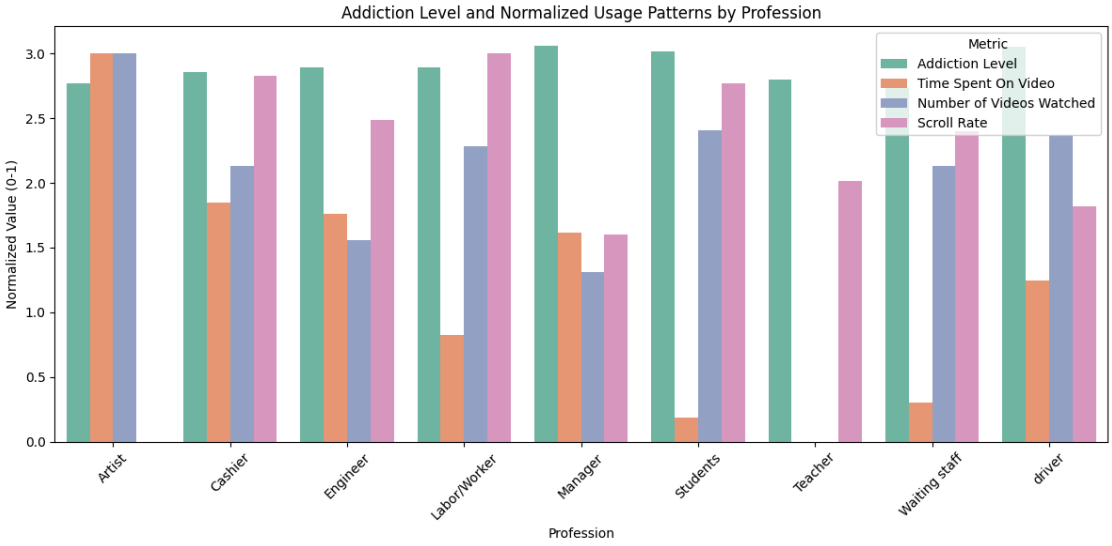
- Scatter plots and heat maps to assess correlations.
- Histograms to analyze the distribution of addiction levels and other metrics.
- Bar charts to compare addiction levels across platforms, genders, and professions.
    - All visualizations were used to support the aforementioned findings:

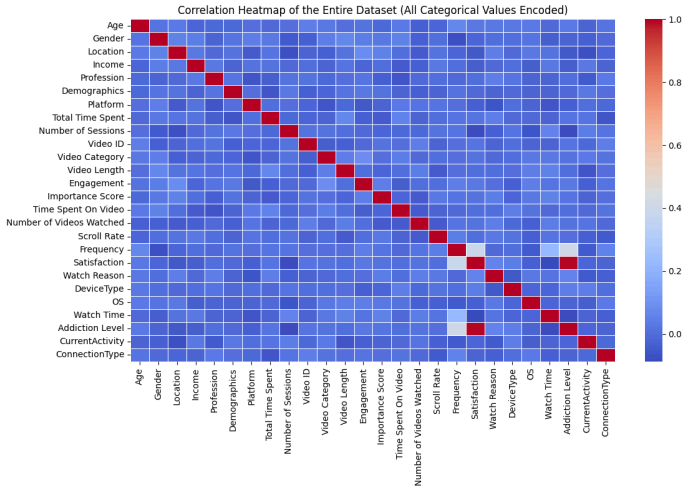





Addiction Rate Across Different Platforms
Average Addiction Level
3.0
2.5
2.0
1.5
1.0
0.5
0.0
YouTube
TikTok
Instagram
Facebook
Platform


Addiction Rate and Average Income Across Different Professions (Sorted)
Average Addiction Level
Average Income
64000
62000
60000
58000
56000
54000
52000
Waiting staff
Artist
Teacher
Cashier
Engineer
Labor/Worker
Students
driver
Manager
Profession





These insights guided our further analyses and ensured they conformed to the data's inherent structure.

Analyses

1. Recommendations

Introduction to the Recommendation System

This study aims to design and develop a recommendation system that analyzes user interactions with video content across various platforms to understand their influence on behavioral outcomes. By focusing on critical metrics such as addiction levels, productivity loss, and user satisfaction, this system seeks to identify patterns and provide actionable insights for mitigating negative behavioral impacts. The research also explores how these outcomes differ across demographic dimensions such as gender, age, and profession.

Objectives:

1. **Identify Influential Platforms and Video Types**
    - Determine which platforms and types of video content (e.g., genres, formats) contribute most significantly to negative behavioral outcomes.
    - Establish a relationship between user engagement and behavioral effects.
2. **Analyze Behavioral Outcomes**
    - Segment users based on demographics (e.g., gender, age, profession) to understand variances in addiction levels, productivity loss, and satisfaction.
    - Draw actionable insights to address specific high-risk user groups.
3. **Develop a Recommender System**
    - Create a system that not only predicts potential adverse behavioral outcomes based on user data but also provides tailored recommendations to reduce their impact.
    - Focus on promoting healthier engagement patterns.

Methods and Techniques:

1. **Latent Feature Learning**
    - Utilize matrix factorization techniques such as **Singular Value Decomposition (SVD)** and **Alternating Least Squares (ALS)** to uncover hidden patterns in user interaction data.
    - These techniques help reveal latent features (e.g., content preferences, behavioral tendencies) influencing addiction levels and productivity loss.
2. **User Similarity Metrics**

    - Employ similarity measures such as **cosine similarity** to group users based on shared behavioral patterns.
    - Prioritize the identification and profiling of high-risk user groups (e.g., those showing high addiction levels) to develop tailored recommendations and interventions.
3. **Behavioral Clustering and Insights**
    - Perform clustering analysis to group users with similar behavioral and demographic characteristics.
    - Use these clusters to provide precise recommendations or platform modifications to mitigate adverse effects.

Descriptive Analysis

Overview

This section provides a detailed analysis of the behavioral outcomes associated with different platforms and video categories. By grouping and aggregating the data, we explore trends that highlight which platforms and content types have the most significant impact on user addiction levels, productivity loss, and satisfaction

Platform-Wise Behavioral Metrics

The data was grouped by platform to evaluate the average behavioral metrics—**addiction levels**, **productivity loss**, and **satisfaction scores**.
Figure 1 illustrates the **Average Addiction Levels** for the four platforms: Facebook, Instagram, TikTok, and YouTube. The average addiction level remains relatively consistent across platforms, ranging from 2.8 to 3.0. Key observations include:

- **Instagram** and **TikTok** show the highest addiction levels, suggesting their engaging nature and high user retention.
- **YouTube**, while slightly lower in addiction levels, still displays significant user engagement.

This finding highlights the need for a closer examination of the features and content strategies used by Instagram and TikTok, as they may contribute to prolonged engagement.

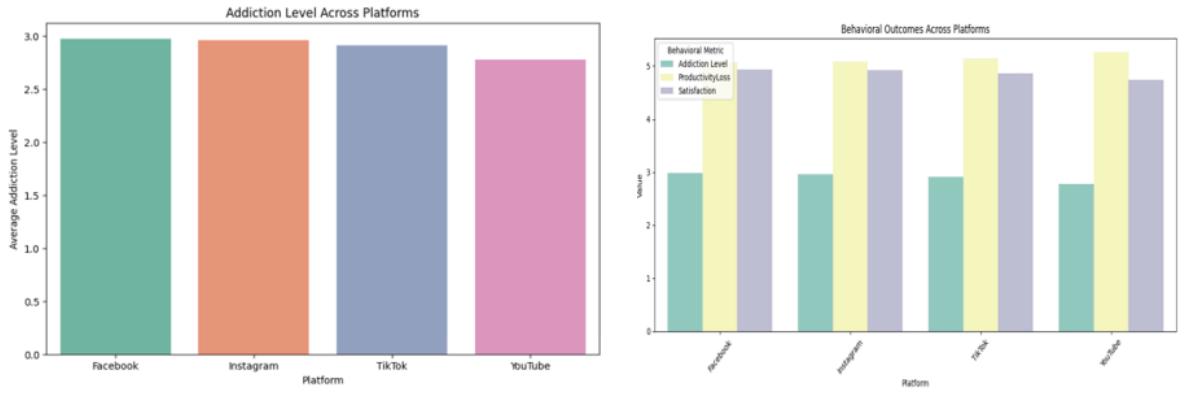


Figure 1

Figure 1 shows the **Behavioral Outcomes Across Platforms**, highlighting the average **Addiction Levels**, **Productivity Loss**, and **Satisfaction** for Facebook, Instagram, TikTok, and YouTube. Key observations include:

- **Productivity Loss** is highest across all platforms, with **YouTube** showing the greatest impact.
- **Addiction Levels** remain relatively consistent across platforms, with TikTok slightly higher.
- **Satisfaction** is the highest on **Facebook**, suggesting users find its content more aligned with their preferences.

Insights: TikTok and YouTube stand out for their high productivity loss, necessitating strategies to address these behavioral risks.

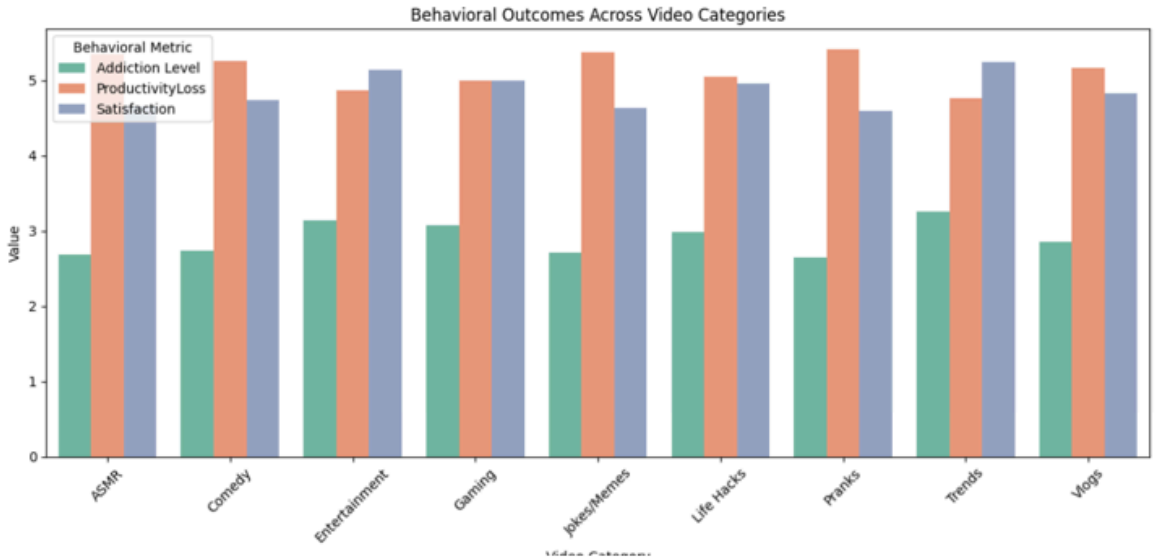


Figure 2

Visualization of Behavioral Metrics Across Video Categories

Figure 2 illustrates the **Behavioral Outcomes Across Video Categories**, showcasing the variation in metrics like **Addiction Levels**, **Productivity Loss**, and **Satisfaction** across different video types. Key findings:

- **Gaming** videos have the highest **Addiction Levels**, suggesting their highly engaging nature.
- **ASMR** and **Vlogs** show significant **Productivity Loss**, highlighting their potential for distraction.
- **Life Hacks** and **Entertainment** videos demonstrate a balance, with lower **Addiction Levels** and higher **Satisfaction**.

Insights: Specific video categories like Gaming and ASMR require targeted recommendations to mitigate their impact on user productivity and addiction.

Behavioral Metrics Analysis

This section explores the platforms and video categories that contribute most significantly to behavioral outcomes, highlighting both areas of concern and opportunities for positive interventions. By analyzing addiction levels, productivity loss, and satisfaction scores, key trends have been identified that offer actionable insights.

Platforms with Highest Addiction Levels:

	Platform	Addiction Level	ProductivityLoss	Satisfaction
0	Facebook	2.977376	5.067873	4.932127
1	Instagram	2.960938	5.078125	4.921875
2	TikTok	2.912088	5.135531	4.864469
3	YouTube	2.780000	5.256000	4.744000

Video Categories with Highest Productivity Loss:

	Video Category	Addiction Level	ProductivityLoss	Satisfaction
6	Pranks	2.645455	5.409091	4.590909
4	Jokes/Memes	2.715084	5.368715	4.631285
0	ASMR	2.683544	5.341772	4.658228
1	Comedy	2.742857	5.257143	4.742857
8	Vlogs	2.850877	5.166667	4.833333
5	Life Hacks	2.987654	5.049383	4.950617
3	Gaming	3.075630	5.000000	5.000000
2	Entertainment	3.137255	4.862745	5.137255
7	Trends	3.260000	4.760000	5.240000

Video Categories with Positive Behavioral Impacts:

	Video Category	Addiction Level	ProductivityLoss	Satisfaction
7	Trends	3.260000	4.760000	5.240000
2	Entertainment	3.137255	4.862745	5.137255
3	Gaming	3.075630	5.000000	5.000000
5	Life Hacks	2.987654	5.049383	4.950617
8	Vlogs	2.850877	5.166667	4.833333
1	Comedy	2.742857	5.257143	4.742857
0	ASMR	2.683544	5.341772	4.658228
4	Jokes/Memes	2.715084	5.368715	4.631285
6	Pranks	2.645455	5.409091	4.590909

Figure 3

The analysis revealed that certain platforms are associated with higher addiction levels. Facebook, Instagram, and TikTok, in particular, emerged as the leading contributors, with average addiction levels of 2.98, 2.97, and 2.91, respectively. These platforms leverage highly engaging features, such as personalized feeds and autoplay functionality, which encourage prolonged usage. While these features boost user engagement, they also contribute to behaviors that may negatively impact users' overall well-being. This highlights the need for targeted platform-specific interventions to mitigate these addictive tendencies.

When examining video categories, it was evident that some content types significantly disrupt productivity. For instance, "Pranks" and "Jokes/Memes" were associated with the highest productivity loss scores, at 5.41 and 5.37, respectively. The lighthearted and short-form nature of these videos makes them highly consumable, leading to extended viewing sessions that detract from users'

focus on other tasks. Similarly, "ASMR" and "Vlogs" categories also showed notable productivity loss, with scores of 5.34 and 5.17, respectively. These findings underscore the importance of developing interventions to address the adverse effects of these video categories.

On a more positive note, the analysis also identified video categories that promote higher user satisfaction while maintaining balanced engagement metrics. "Trends" and "Entertainment" categories stood out with satisfaction scores of 5.24 and 5.14, respectively, indicating that they deliver a more positive user experience. Additionally, "Life Hacks" videos, with a satisfaction score of 4.95 and relatively low addiction and productivity loss scores, demonstrate how educational or utility-driven content can foster healthier engagement patterns. These categories provide an opportunity for platforms to emphasize and promote content that enhances user satisfaction without compromising productivity.

These insights emphasize the dual nature of content and platform impacts: while some platforms and video categories contribute to problematic behaviors, others offer opportunities for more positive outcomes. To address these findings, it is recommended that platforms focus on reducing the impact of high-risk content, such as "Pranks" and "Jokes/Memes," while promoting categories like "Trends," "Entertainment," and "Life Hacks." Such measures could include content moderation, personalized recommendations for healthier engagement, and reminders for users to take breaks during extended viewing sessions.

By leveraging these insights, this study aims to guide the development of a recommendation system that can identify and predict behavioral risks while promoting balanced and satisfying user experiences.

Behavioral Outcomes Analysis by Demographics

This section examines the variation in behavioral outcomes such as addiction levels and productivity loss across demographic groups, including gender, age, and profession. The findings provide insights into how specific groups are more susceptible to negative outcomes, helping to identify target areas for intervention.

Addiction Levels Across Gender and Profession

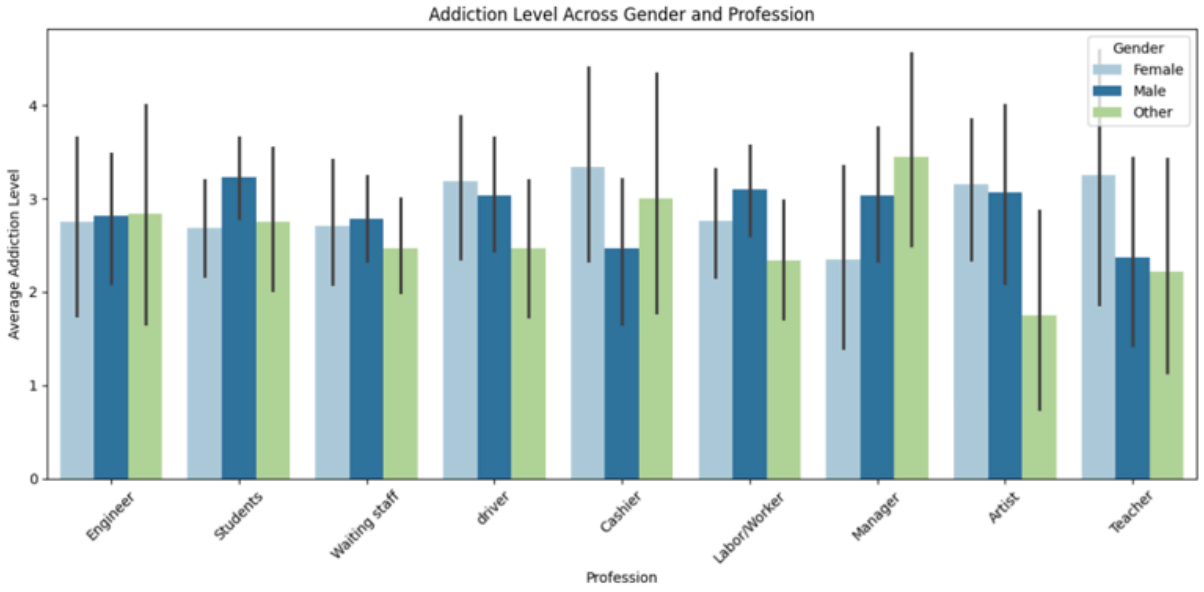


Figure 4: Average Addiction Levels Across Gender and Profession

The analysis of addiction levels across gender and profession (Figure 4) reveals distinct patterns among various demographic groups:

- **Students**, **labor workers**, and **artists** demonstrate higher average addiction levels compared to professions like **engineers** and **teachers**. This suggests that individuals in certain professions may be more vulnerable to addictive behaviors, possibly due to varying levels of exposure to digital platforms or differing work environments.
- Differences in addiction levels are also observed between genders within the same profession. For example, male students exhibit slightly higher addiction levels than their female counterparts, while female labor workers show higher levels compared to males in the same role.

These patterns highlight the role of professional and gender-related factors in shaping susceptibility to addictive behaviors, suggesting a need for targeted interventions that address specific demographic groups.

Productivity Loss Across Gender and Profession

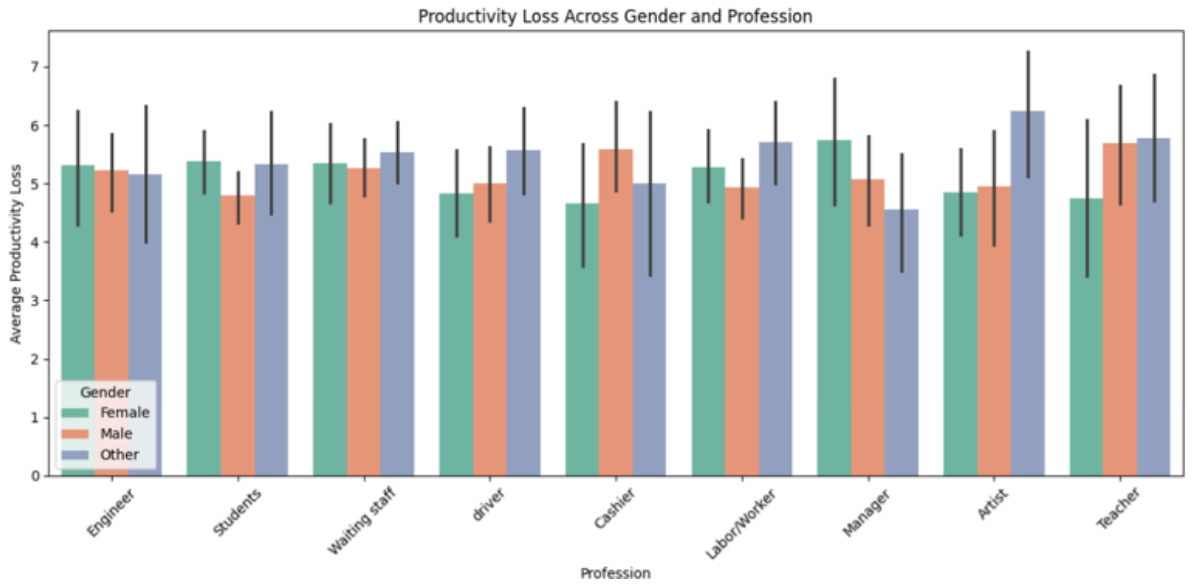


Figure 5: Average Productivity Loss Across Gender and Profession

An analysis of productivity loss across gender and profession (Figure 5) provides further insights:

- Professions such as **waiting staff**, **students**, and **labor workers** experience the highest average productivity loss, with scores exceeding those of other professions like **engineers** or **managers**. These findings could reflect the nature of these roles, where flexible schedules or digital distractions may contribute to greater productivity challenges.
- Within professions, significant differences are observed between genders. For instance, male and female waiting staff exhibit varying levels of productivity loss, suggesting that gender dynamics within specific roles may influence how individuals are affected by digital engagement.

These findings underscore the importance of understanding the interplay between profession, gender, and behavioral outcomes to develop effective mitigation strategies.

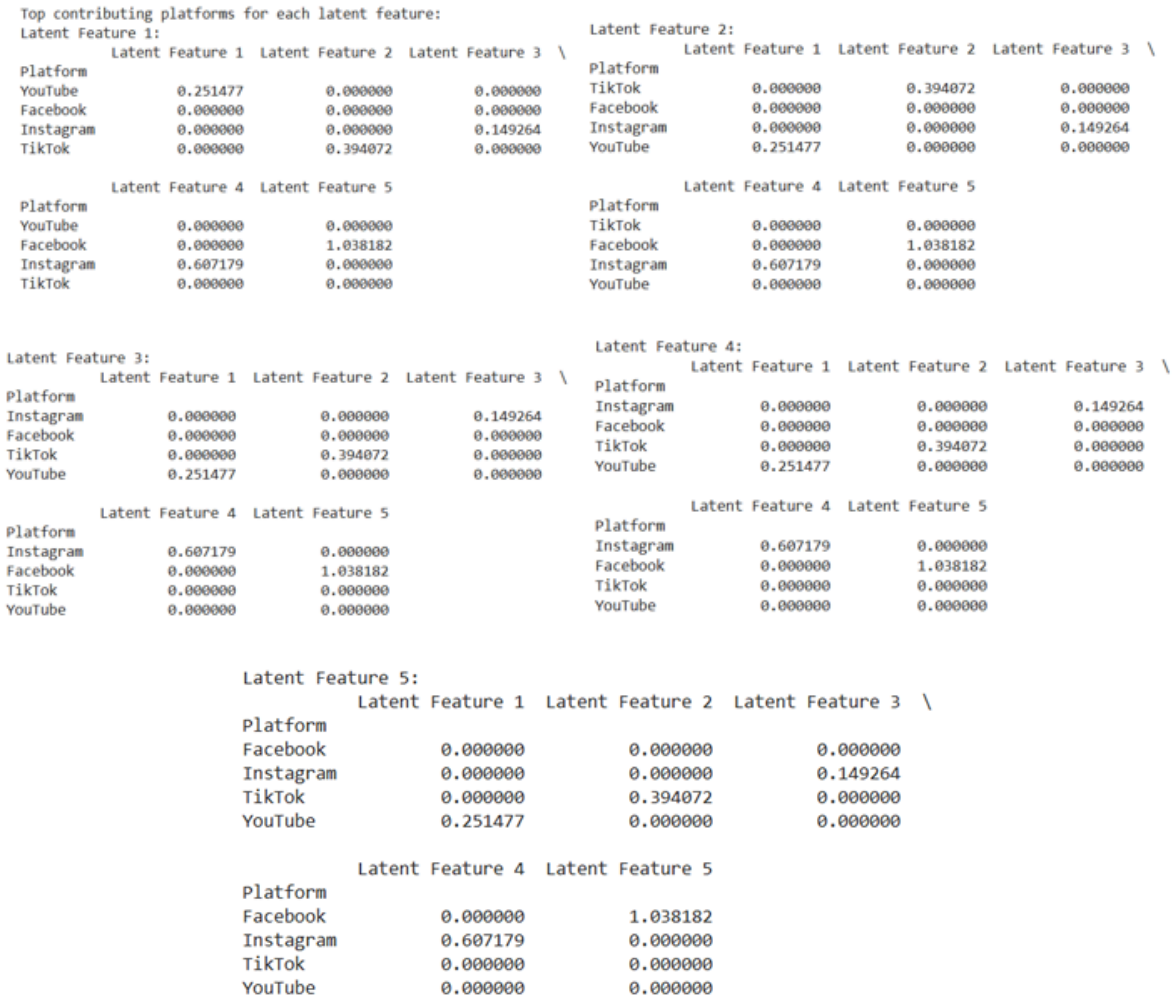
Latent Feature Learning with Non-Negative Matrix Factorization (NMF)

To uncover hidden patterns in user behavior, Non-Negative Matrix Factorization (NMF) was applied to the user-platform interaction data, focusing on addiction levels. This technique decomposes the data into latent features, which represent

underlying behavioral patterns that are not directly observable. Each latent feature reflects a specific dimension of user-platform interactions, allowing for a deeper understanding of how platforms contribute to addiction across different user groups.

Steps in the Analysis

1. Preparation of User-Platform Matrix:
    - A matrix was constructed with users as rows and platforms as columns, with addiction levels as the values. Missing values were filled with zero to ensure a complete dataset for analysis.
2. Normalization:
    - The data was normalized using Min-Max Scaling to ensure all values are within a comparable range, which is essential for NMF.
3. Application of NMF:
    - The matrix was factorized into two components:
        - User Features Matrix: Represents the weight of each latent feature for each user.
        - Platform Features Matrix: Reflects the contribution of each platform to the latent features.
4. Reconstruction:
    - The original matrix was reconstructed using the factorized components to validate the quality of the decomposition.



```
Top contributing platforms for each latent feature:
Latent Feature 1:
          Latent Feature 1  Latent Feature 2  Latent Feature 3  \
Platform
YouTube           0.251477          0.000000          0.000000
Facebook          0.000000          0.000000          0.000000
Instagram         0.000000          0.000000          0.149264
TikTok            0.000000          0.394072          0.000000

          Latent Feature 4  Latent Feature 5
Platform
YouTube           0.000000          0.000000
Facebook          0.000000          1.038182
Instagram         0.607179          0.000000
TikTok            0.000000          0.000000

Latent Feature 2:
          Latent Feature 1  Latent Feature 2  Latent Feature 3  \
Platform
TikTok            0.000000          0.394072          0.000000
Facebook          0.000000          0.000000          0.000000
Instagram         0.000000          0.000000          0.149264
YouTube           0.251477          0.000000          0.000000

          Latent Feature 4  Latent Feature 5
Platform
TikTok            0.000000          0.000000
Facebook          0.000000          1.038182
Instagram         0.607179          0.000000
YouTube           0.000000          0.000000

Latent Feature 3:
          Latent Feature 1  Latent Feature 2  Latent Feature 3  \
Platform
Instagram         0.000000          0.000000          0.149264
Facebook          0.000000          0.000000          0.000000
TikTok            0.000000          0.394072          0.000000
YouTube           0.251477          0.000000          0.000000

          Latent Feature 4  Latent Feature 5
Platform
Instagram         0.607179          0.000000
Facebook          0.000000          1.038182
TikTok            0.000000          0.000000
YouTube           0.000000          0.000000

Latent Feature 4:
          Latent Feature 1  Latent Feature 2  Latent Feature 3  \
Platform
Instagram         0.000000          0.000000          0.149264
Facebook          0.000000          0.000000          0.000000
TikTok            0.000000          0.394072          0.000000
YouTube           0.251477          0.000000          0.000000

          Latent Feature 4  Latent Feature 5
Platform
Instagram         0.607179          0.000000
Facebook          0.000000          1.038182
TikTok            0.000000          0.000000
YouTube           0.000000          0.000000

Latent Feature 5:
          Latent Feature 1  Latent Feature 2  Latent Feature 3  \
Platform
Facebook          0.000000          0.000000          0.000000
Instagram         0.000000          0.000000          0.149264
TikTok            0.000000          0.394072          0.000000
YouTube           0.251477          0.000000          0.000000

          Latent Feature 4  Latent Feature 5
Platform
Facebook          0.000000          1.038182
Instagram         0.607179          0.000000
TikTok            0.000000          0.000000
YouTube           0.000000          0.000000
```

Figure 6

Analysis of Latent Features

Each latent feature represents a hidden behavior or pattern in the data, influenced by specific platforms. The contributions of each platform to the latent features are visualized in Figure 1, which highlights the strength of association between platforms and latent features.

1. Latent Feature 1:
    - Top Contributing Platform: YouTube (Contribution: 0.25)

- Description: This feature may represent users who heavily engage with video-based content, with YouTube being the dominant contributor.

2. Latent Feature 2:
   - Top Contributing Platform: TikTok (Contribution: 0.39)
   - Description: This feature could indicate users who are drawn to short, addictive content formats, with TikTok emerging as the strongest contributor.
3. Latent Feature 3:
   - Top Contributing Platform: Instagram (Contribution: 0.15)
   - Description: Likely represents users focused on visual and social media content, with Instagram playing a significant role.
4. Latent Feature 4:
   - Top Contributing Platform: Instagram (Contribution: 0.61)
   - Description: Suggests a preference for lifestyle and curated content, heavily influenced by Instagram.
5. Latent Feature 5:
   - Top Contributing Platform: Facebook (Contribution: 1.04)
   - Description: Reflects users engaging with socially connected and community-oriented content, with Facebook dominating this feature.

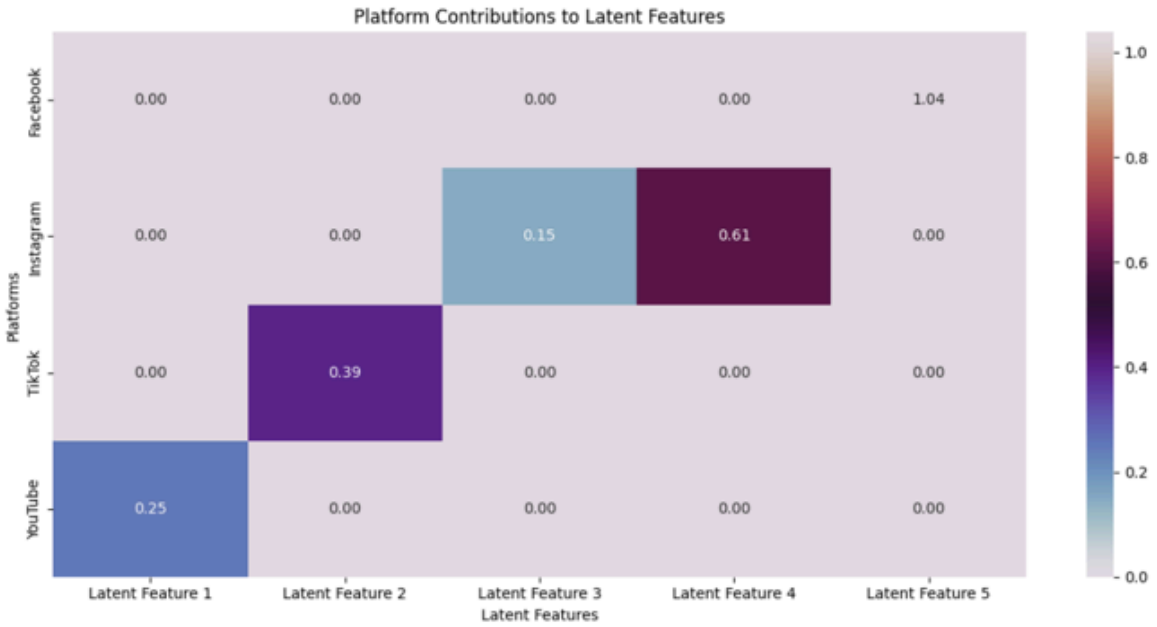


Figure 7: Platform Contributions to Latent Features

This heatmap visualizes the strength of each platform's contribution to the latent features, with darker shades representing higher contributions.

As depicted in Figure 7, the heatmap highlights the contributions of each platform to the latent features:

- **Facebook** exhibits the highest contribution to **Latent Feature 5** (value: 1.04), indicating its association with long-form or community-driven content.
- **TikTok** leads in **Latent Feature 2** (value: 0.39), reflecting its role in promoting short-form, addictive content.
- **Instagram** dominates **Latent Feature 4** (value: 0.61), emphasizing its influence in shaping lifestyle-based or visual content interactions.
- **YouTube** contributes significantly to **Latent Feature 1** (value: 0.25), highlighting its association with long-form video consumption.

Clustering Users Based on Latent Features

To identify patterns in user behavior, **k-means clustering** was applied to group users based on their alignment with latent features derived from the Non-Negative Matrix Factorization (NMF). This clustering approach revealed distinct groups of users with similar behavioral tendencies, enabling a deeper understanding of user preferences and potential intervention strategies.

Cluster Analysis

Using k-means clustering with three clusters, users were grouped based on their alignment with the five latent features. The distribution of users across the clusters is shown in **Figure 8**, with the following cluster sizes:

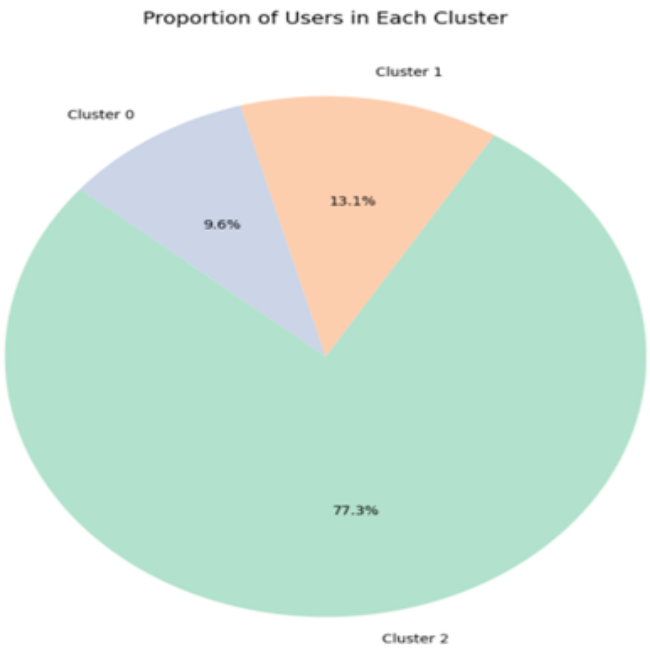


Figure 8: Proportion of Users in Each Cluster

This pie chart displays the distribution of users across the three clusters, with Cluster 2 representing the majority.

- **Cluster 2 (Blue)**: 773 users (77.3%)
- **Cluster 1 (Orange)**: 131 users (13.1%)
- **Cluster 0 (Green)**: 96 users (9.6%)

The cluster centroids, representing the average position of users in each cluster across the latent feature space, are visualized in Figure 9. These centroids highlight the defining characteristics of each cluster and their behavioral alignment with the latent features.

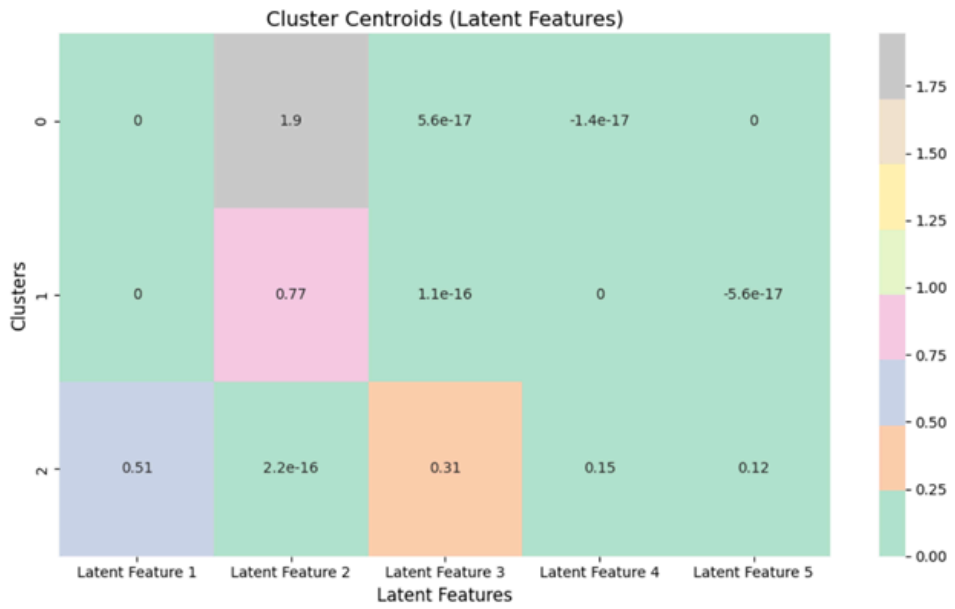


Figure 9: Cluster Centroids in Latent Feature Space

This heatmap visualizes the centroid positions for each cluster across the five latent features, highlighting the dominant traits of each group.

Insights from Clusters

1. **Cluster 2 (Blue)**:
    - **Dominant Latent Feature**: Latent Feature 1
    - **Platform Association**: Users in this cluster exhibit behaviors strongly tied to platforms like **YouTube**, characterized by engagement with long-form video content.
    - **Behavioral Traits**: Likely to engage more passively, with extended but focused usage patterns.
2. **Cluster 0 (Green)**:
    - **Dominant Latent Feature**: Latent Feature 2
    - **Platform Association**: This cluster aligns strongly with **TikTok**, where users engage actively with short-form, highly dynamic, and engaging content.
    - **Behavioral Traits**: These users may exhibit addictive tendencies tied to rapid content consumption.
3. **Cluster 1 (Orange)**:

- **Mixed Behavior**: Users in this smaller cluster show intermediate alignment with various latent features, suggesting they do not have a dominant platform preference but exhibit unique and diverse usage traits.

Cosine Similarity Analysis

This section analyzes behavioral similarities between users by calculating cosine similarity scores in the latent feature space. The objective is to uncover shared engagement patterns among users, with a specific focus on high-addiction users, to better understand their common tendencies and develop personalized intervention strategies.

Top Similar Users

Each user was matched with their top 5 most similar users based on cosine similarity. This analysis highlights shared tendencies and potential clustering in user behaviors. For instance:

- **User 1** shares high similarity with users 123, 694, 760, 270, and 895.
- **User 2** exhibits similar engagement patterns to users 543, 695, 546, 766, and 367.

These relationships suggest that users with similar behavioral patterns may be influenced by shared triggers, such as specific platform features or content types.

High-Addiction Users

To better understand problematic engagement patterns, the analysis isolated high-addiction users, defined as users with addiction levels in the top 20th percentile. A targeted similarity matrix was created for these users, visualized in Figure 10.

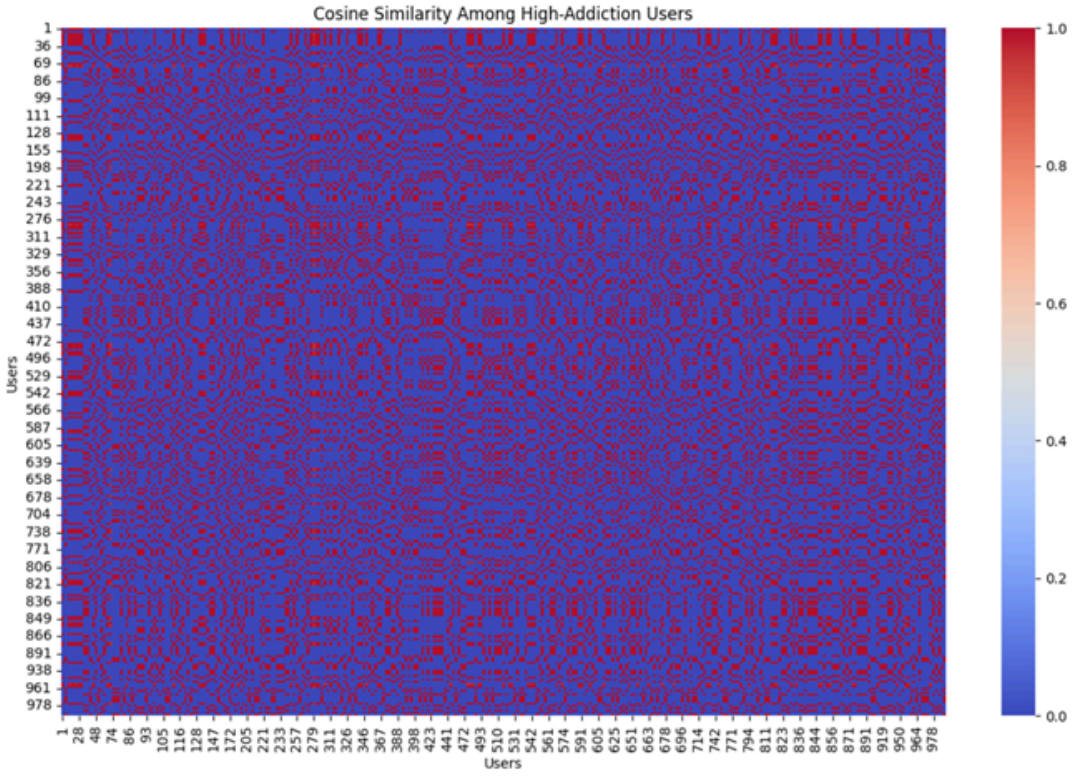


Figure 10: *Cosine Similarity Among High-Addiction Users*
The heatmap reveals the strength of behavioral similarities among high-addiction users:

- **Red areas** indicate strong similarity, suggesting shared behavioral triggers, potentially driven by platforms like TikTok or Instagram.
- **Blue areas** indicate dissimilarity, highlighting users with distinct patterns that may require more personalized interventions.

Hierarchical Clustering for High-Addiction Users

To further explore the relationships among high-addiction users, hierarchical clustering was applied. The results, visualized in Figure 11, show clusters of users with similar behavioral tendencies:

- **Large clusters** (e.g., the top-left red block in the dendrogram) represent users with highly aligned behaviors, likely influenced by common factors.
- **Smaller clusters** indicate niche behavioral patterns, while blue areas highlight diversity, suggesting unique addiction triggers.

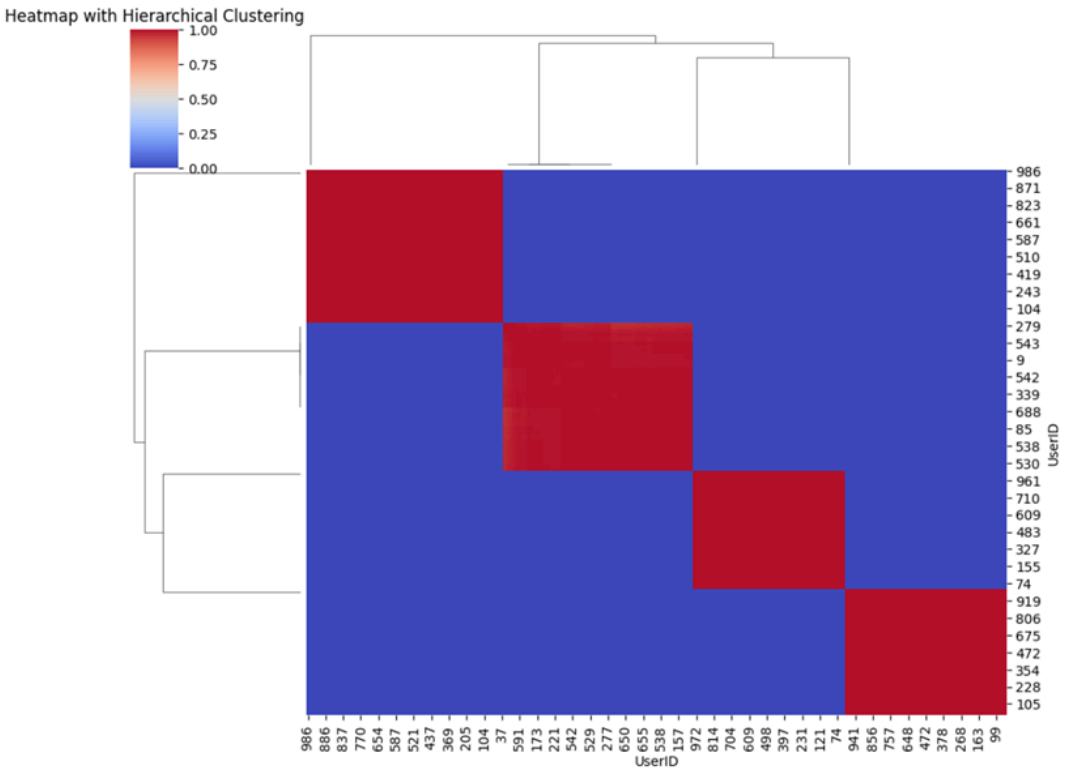


Figure 11: Heatmap with Hierarchical Clustering
This heatmap visualizes both similarities and hierarchical groupings of high-addiction users. Red blocks represent groups with strong alignment, while blue areas reveal users with diverse engagement patterns.

Shared Behavioral Triggers:

- Large clusters of similar users suggest that many high-addiction users are influenced by similar platform dynamics or content types. For example, users clustered in red blocks may share a preference for short-form, engaging content from TikTok.

Diversity in Addiction Patterns:

- Blue areas in the heatmaps indicate users with distinct behaviors, suggesting that addiction triggers vary across individuals and may depend on unique factors like platform usage patterns or personal preferences.

Targeted Group-Specific Interventions:

- Large clusters of similar users can benefit from group-specific interventions, such as platform-level changes or shared recommendations to mitigate addictive tendencies.

  Personalized Solutions for Diverse Patterns:

- Users outside major clusters (blue areas) require more individualized strategies to address their unique behavioral triggers.

This analysis of cosine similarity and hierarchical clustering provides a detailed understanding of user behavioral patterns, especially among high-addiction users. The identification of both large, similar groups and diverse patterns highlights the need for a combination of group-specific interventions and personalized solutions. Platforms can leverage these insights to implement targeted strategies that promote healthier engagement and reduce harmful behavioral outcomes.

Cluster-Wise Behavioral Analysis of High-Addiction Users

Using hierarchical clustering, high-addiction users were divided into three distinct clusters based on their behavioral patterns. This analysis explores the average **Addiction Level**, **Productivity Loss**, and **Satisfaction** for each cluster, providing deeper insights into the differences between user groups and their specific engagement tendencies.

Cluster-Wise Behavioral Analysis:

	Cluster	Addiction Level	ProductivityLoss	Satisfaction
0	1	5.471910	2.528090	7.471910
1	2	5.441860	2.558140	7.441860
2	3	5.450704	2.549296	7.450704

Figure 12: Average Addiction Level by Cluster
This bar chart illustrates the average addiction levels for each cluster, highlighting minimal variation in addiction scores across the groups.

Cluster Characteristics

The analysis revealed the following characteristics for the three clusters:

1. **Cluster 0**:

    - **Addiction Level**: 5.47 (highest among clusters)
    - **Productivity Loss**: 2.53 (lowest among clusters)
    - **Satisfaction**: 7.47 (highest among clusters)
    - **Key Insights**: This cluster represents users with **high addiction and satisfaction**, but **minimal productivity loss**. These users might engage heavily with platforms but do not perceive significant negative impacts on their productivity.
2. **Cluster 1**:
    - **Addiction Level**: 5.44
    - **Productivity Loss**: 2.56
    - **Satisfaction**: 7.44
    - **Key Insights**: Similar to Cluster 0, users in this group exhibit **high addiction and satisfaction**, with slightly higher productivity loss than Cluster 0. These users may balance heavy platform usage with their daily activities, experiencing moderate disruptions.
3. **Cluster 2**:
    - **Addiction Level**: 5.45
    - **Productivity Loss**: 2.55 (highest among clusters)
    - **Satisfaction**: 7.45
    - **Key Insights**: This cluster stands out due to its **higher productivity loss** compared to the other clusters. These users appear to be more negatively affected by their addictive behaviors, indicating a need for targeted interventions to mitigate these effects.

Evaluation of Addiction Level Predictions

To validate the effectiveness of the latent feature model, the predicted addiction levels were compared with observed values using the **Mean Absolute Error (MAE)** metric. The analysis yielded the following results:

1. **Model Accuracy**:
    - The **MAE** of addiction level predictions was calculated as **0.62**, indicating a reasonable level of accuracy in the model's ability to predict user addiction levels.
    - While the error is relatively low, it demonstrates that the model effectively captures key behavioral patterns in the majority of users.
2. **High Prediction Discrepancies**:

- Certain users (e.g., **User 639**, **User 727**) exhibited high prediction discrepancies, with errors of up to **1.5**. These outliers suggest the presence of **unique behavioral patterns** that the model does not fully capture.
- Such users may require more personalized recommendations or the inclusion of additional behavioral dimensions in future models.

3. **Insights from Prediction Errors**:
   - The low MAE reflects that most users' addiction behaviors are effectively modeled.
   - The discrepancies among high-error users highlight potential gaps in the latent feature representation, suggesting the need for refinements in the model.

```
Top 5 Platforms to Minimize Addiction Level:
    Platform  Addiction Level
3    YouTube         2.780000
2     TikTok         2.912088
1  Instagram         2.960938
0   Facebook         2.977376


Top 5 Video Categories to Minimize Addiction Level:
  Video Category  Addiction Level
6         Pranks         2.645455
0           ASMR         2.683544
4    Jokes/Memes         2.715084
1         Comedy         2.742857
8          Vlogs         2.850877
Top 5 Platforms to Maximize Satisfaction:
    Platform  Satisfaction
0   Facebook      4.932127
1  Instagram      4.921875
2     TikTok      4.864469
3    YouTube      4.744000


Top 5 Video Categories to Maximize Satisfaction:
  Video Category  Satisfaction
7         Trends      5.240000
2  Entertainment      5.137255
3         Gaming      5.000000
5     Life Hacks      4.950617
8          Vlogs      4.833333
```

The evaluation of addiction predictions and behavioral insights from the clustering and similarity analyses suggest the following actionable strategies:

1. Platform-Specific Interventions

- **YouTube** and **TikTok**: Focus on promoting long-form or productive content while implementing tools to limit excessive usage, such as usage reminders or "time caps."
- **Instagram** and **Facebook**: Encourage community-driven interactions or content that supports balanced engagement.

2. Personalized Recommendations

- Users with high prediction discrepancies (e.g., **User 639**, **User 727**) should receive tailored recommendations to address their unique behavioral patterns. These could include:
    - Encouraging breaks during platform usage.
    - Suggesting alternative content categories with lower addiction levels (e.g., **Life Hacks**, **Trends**) to shift engagement patterns.

3. Targeted Interventions for High-Addiction Users

- Based on the clustering analysis, **Cluster 2** users (high productivity loss) should be prioritized for intervention strategies. This could involve:
    - Promoting content with lower addictive tendencies (e.g., **Pranks**, **ASMR**).
    - Encouraging productive habits through reminders or curated recommendations.

Platform-Specific Interventions

To minimize addictive behaviors, platforms like **YouTube** and **TikTok** should promote long-form, educational, or productive content while implementing tools to limit excessive usage. Platforms like **Instagram** and **Facebook** can focus on promoting community-driven interactions.

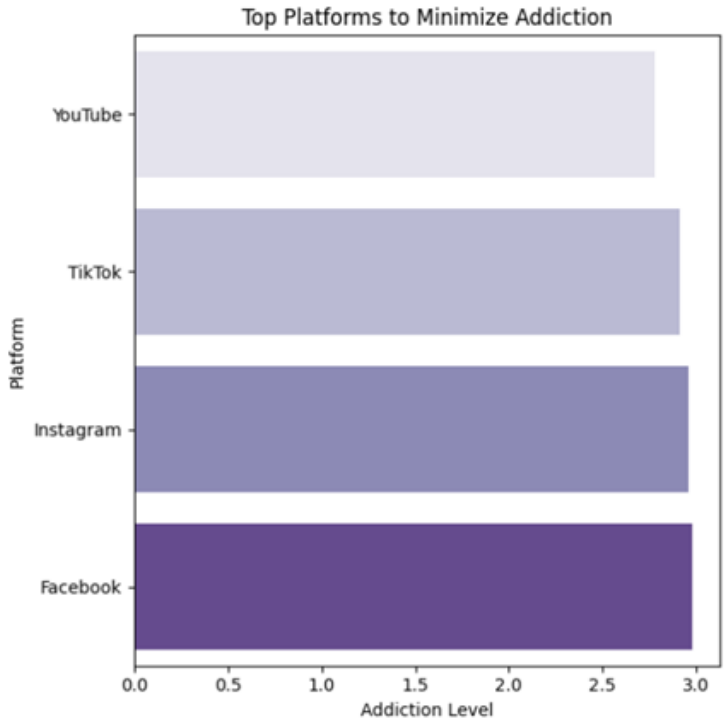


Figure 13: *Top Platforms to Minimize Addiction* shows the platforms with the lowest addiction levels, highlighting areas where intervention could reinforce positive behaviors.

Personalized Recommendations

To reduce addiction for high-risk users, content categories with lower addictive tendencies, such as **Pranks** and **ASMR**, should be promoted. These categories encourage engagement while mitigating harmful outcomes.

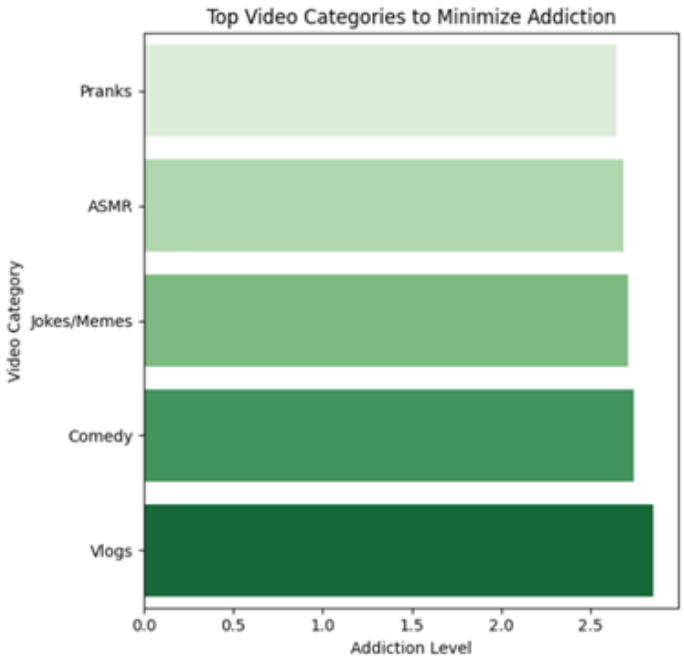


Figure 14: *Top Video Categories to Minimize Addiction* highlights the categories that can reduce engagement with highly addictive content types.

Behavioral Analysis by Gender and Profession

This section explores behavioral patterns by analyzing gender-profession combinations to identify groups with high addiction levels and low productivity loss. The findings provide insights into group-specific tendencies and potential areas for targeted interventions.

1. High Addiction Levels:
    - Female Teachers (3.50) and Other Managers (3.45) exhibit the highest addiction levels.
    - These groups may require targeted interventions, such as promoting less addictive content or introducing platform-level features like reminders to limit exposure to highly engaging platforms like TikTok and Instagram.
2. Low Productivity Loss:
    - Despite high addiction levels, groups like Female Cashiers (4.62) and Female Students (4.80) report relatively low productivity loss. This indicates that their platform usage patterns may not significantly

disrupt their workflows. Such resilience suggests these users manage their engagement more effectively.

Underlying Triggers for Behavioral Patterns

1. Satisfaction and Addiction:
    - Platforms like Facebook and Instagram contribute to high satisfaction levels by emphasizing social interaction and user-generated content. However, platforms such as TikTok and Instagram also drive addiction due to features like endless scrolling and personalized algorithmic feeds.
2. Content Categories and Productivity:
    - Categories such as Entertainment and Gaming provide high satisfaction but may also result in excessive time investment, thereby reducing productivity for some user groups.

Gender-Profession Combinations with High Addiction Levels

The analysis revealed that certain gender-profession groups are more susceptible to addictive platform engagement. Notable examples include:

- Female Teachers: This group shows the highest addiction levels, indicating a potential need for interventions focused on managing usage patterns.
- Other Managers: With similarly high addiction scores, this group could benefit from time management tools or alternative content recommendations.

Gender-Profession Combinations with Low Productivity Loss

Groups such as Female Cashiers and Female Students demonstrate low productivity loss despite their high addiction levels. This suggests that they are either more efficient in managing their engagement or their platform usage aligns well with their daily routines.

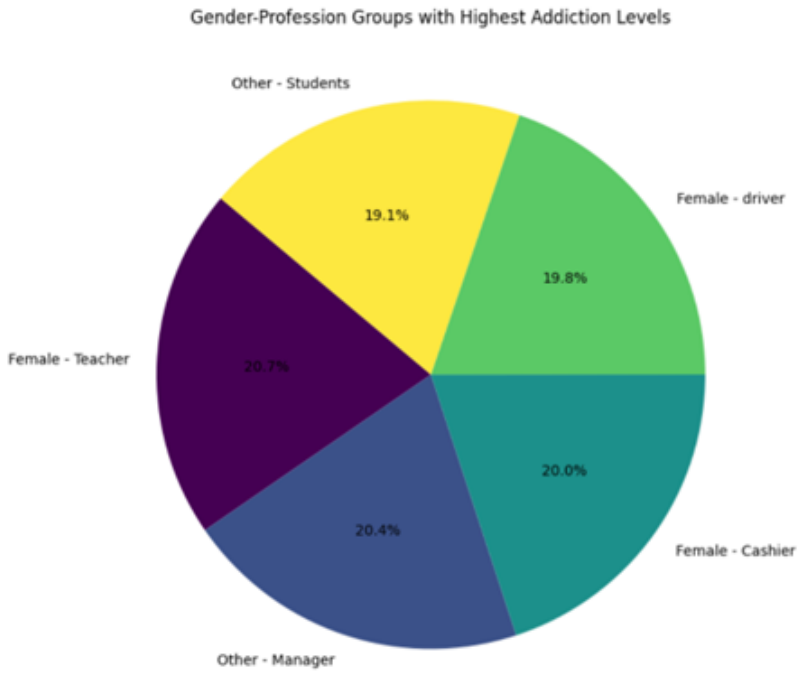


Figure 15: Gender-profession groups with the highest addiction levels

Platforms with High Addiction Levels

The analysis highlights that **Facebook (2.98)** and **Instagram (2.96)** are the platforms with the highest addiction levels, followed by **TikTok (2.91)** and **YouTube (2.78)**. These platforms incorporate features like algorithm-driven content recommendations and infinite scrolling, which contribute to addictive behaviors. Addressing these features through moderation tools, time reminders, and educational content on healthy consumption could help mitigate their effects.

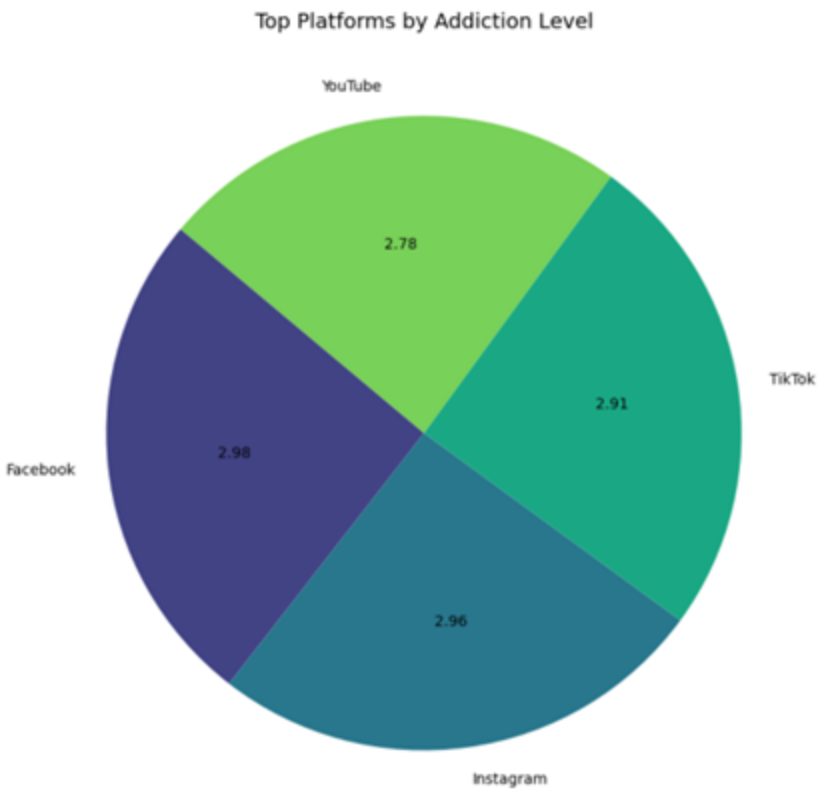


Figure 16: Distribution of Addiction Levels Across Platforms

Demographics with High Addiction Levels

Certain demographic groups exhibit higher susceptibility to addiction. For instance:

- **Older females aged 60** and **middle-aged females aged 37** have the highest addiction levels, scoring 5.0 each.
- **Middle-aged males aged 48** and **non-binary individuals aged 22** also exhibit elevated addiction levels, with scores of 4.71 and 4.66, respectively.

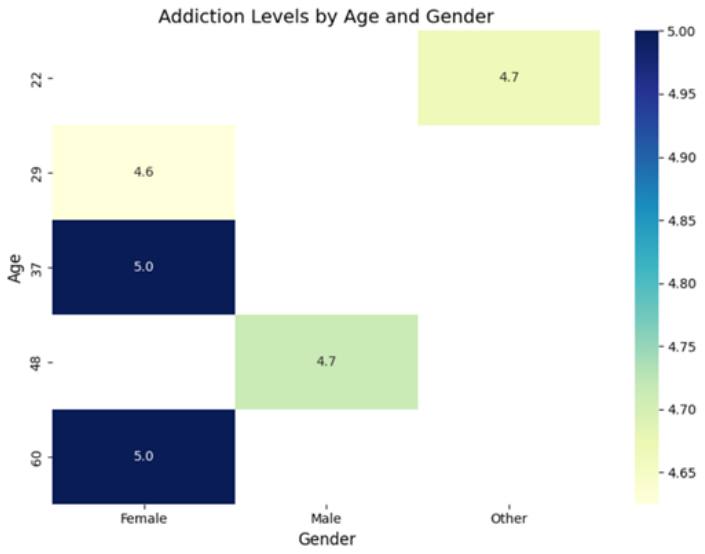


Figure 17: Addiction Levels by Gender and Age

These insights underline the need for tailored interventions. For older users, interventions like promoting content diversity or providing digital literacy programs might help balance their consumption habits.

Platforms and Demographics with Low Satisfaction

Platforms like **YouTube (4.74)** and **TikTok (4.86)** show relatively lower satisfaction compared to others. Despite their widespread usage, these platforms might lack features fostering meaningful engagement or positive emotional responses.

From a demographic perspective, **non-binary users** and **older individuals** report the lowest satisfaction levels. This suggests inclusivity and accessibility issues that platforms should address. Tailored content, inclusive policies, and improved interface design could enhance user satisfaction in these groups.

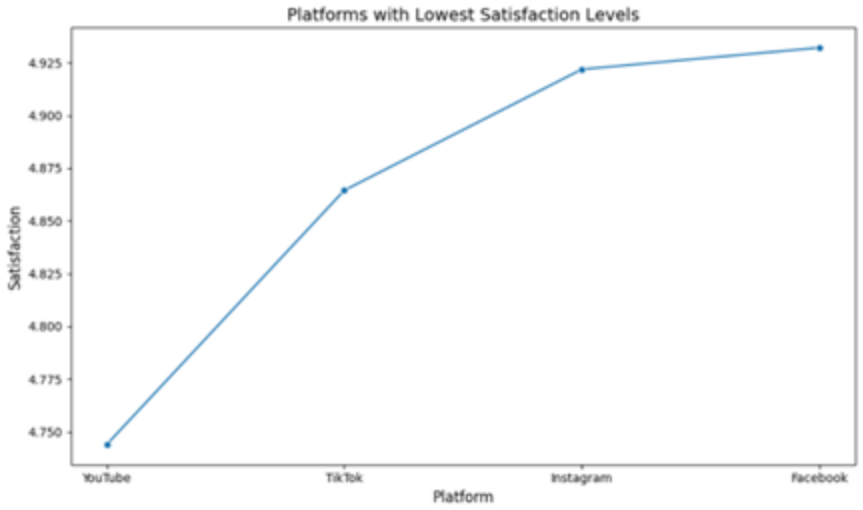


Figure 18: Satisfaction Trends Across Platforms

2. Clustering Analysis

Clustering techniques were employed to uncover patterns in user data and classify individuals based on their addiction risk. These methods are not only valuable for academic research but also hold significant real-world potential for social media companies. By identifying users at risk of developing harmful behaviors, platforms can implement targeted interventions, such as personalized content moderation, addiction awareness campaigns, and interface designs to reduce excessive usage.

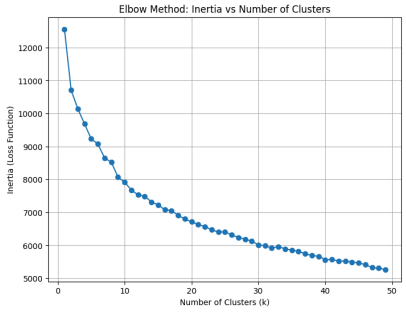
2.1 Clustering with Dimensionality Reduction (PCA)

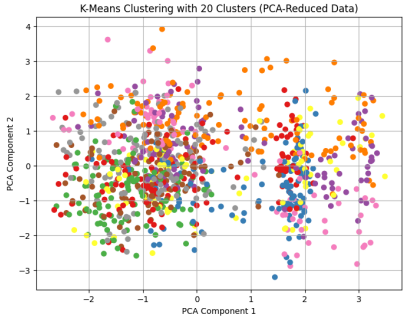
To simplify the complexity of the dataset, we utilized Principal Component Analysis (PCA) to focus on the most critical features while preserving patterns essential for clustering. PCA reduces high-dimensional data into a smaller set of uncorrelated components by maximizing variance, ensuring meaningful information is retained.

We transformed user features into principal components and tuned hyperparameters, such as the number of components and cluster count, to optimize clustering results. Using k-means clustering on the reduced dimensions,

we identified representative clusters, or centroids, corresponding to distinct addiction risk profiles. Real-world users were then matched to these centroids using cosine similarity, enabling the classification of users into risk groups.

In practice, social media companies can leverage this approach to create risk-based user segments. For example, users in high-risk clusters could receive gentle prompts to take breaks, or their feeds could be adapted to include less engagement-driven content. This method allows for a proactive approach to addressing addiction tendencies while respecting user autonomy.





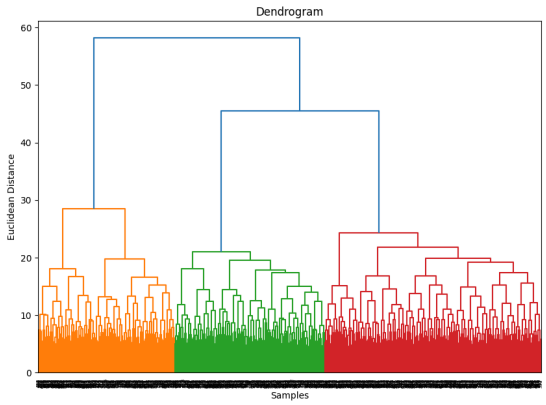
2.2 Agglomerative Clustering

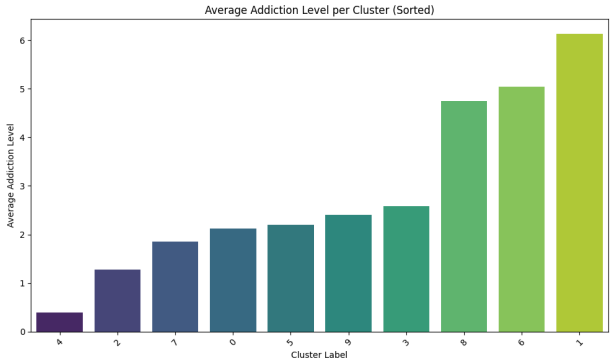
Hierarchical clustering was employed to group users based on addiction levels and demographics. By iteratively merging clusters with similar addiction-related features, this method revealed a dendrogram—a hierarchical structure that visualizes the relationships within the data. We utilized average linkage, which calculates the mean distance between all pairs of points in two clusters, and optimized parameters to achieve the best clustering performance as measured by silhouette scores.

To identify the most at-risk demographic groups, we calculated the weighted mode of key features within each cluster, such as gender, platform usage, and

income levels. These profiles provided actionable insights into common traits of high-risk users.

Social media companies can use such demographic insights to tailor addiction-reduction strategies. For example, platforms might adjust their algorithms to reduce addictive content exposure for users in specific at-risk groups or deploy educational campaigns targeted at particular demographics, such as young adults or low-income users.





	Cluster 4	Cluster 1
Gender	Female	Female
Location	Japan	Barzil
Demographics	Rural	Urban
Age	41.681818	42.652174
Income	74960.787879	62510.252174
Addiction Level	0.393939	6.130435

2.3 Locality-Sensitive Hashing (LSH) with Minhashing

To efficiently identify users with similar behavioral patterns, we implemented Locality-Sensitive Hashing (LSH) combined with Minhashing. LSH approximates similarity in high-dimensional data by hashing similar items into the same buckets, enabling rapid comparison. Minhashing was used to generate hash signatures, preserving similarities in user behaviors such as platform engagement and scrolling habits.

By comparing these signatures, we identified users with shared patterns and grouped them into behaviorally similar cohorts. This clustering informed a prediction pipeline capable of classifying addiction profiles in real-time, offering scalability and speed.

Social media companies can apply this technique to monitor and predict at-risk behavior on a larger scale. For example, identifying clusters of users who frequently engage with highly addictive content could prompt platforms to adjust recommendation algorithms, reducing exposure to such content. Additionally, these insights could guide product design changes aimed at encouraging healthier user habits.

User with Addiction Level 1:
User Profile:

	Frequency	Watch Reason	DeviceType	Age	Satisfaction
1	Morning	Procrastination	Tablet	47	3

Similar Users:

	Frequency	Watch Reason	DeviceType	Age	Satisfaction	Addiction Level
284	Afternoon	Habit	Computer	57	5	3
295	Afternoon	Habit	Computer	52	5	3
308	Afternoon	Habit	Computer	34	5	3
1	Afternoon	Habit	Computer	46	5	3

User with Addiction Level 2:
User Profile:

	Frequency	Watch Reason	DeviceType	Age	Satisfaction
2	Evening	Habit	Computer	42	4

Similar Users:

	Frequency	Watch Reason	DeviceType	Age	Satisfaction	Addiction Level
834	Evening	Entertainment	Tablet	20	4	2
2	Evening	Entertainment	Tablet	32	4	2
13	Evening	Entertainment	Tablet	57	4	2
241	Evening	Entertainment	Tablet	20	4	2

User with Addiction Level 3:
User Profile:

	Frequency	Watch Reason	DeviceType	Age	Satisfaction
3	Evening	Procrastination	Tablet	28	5

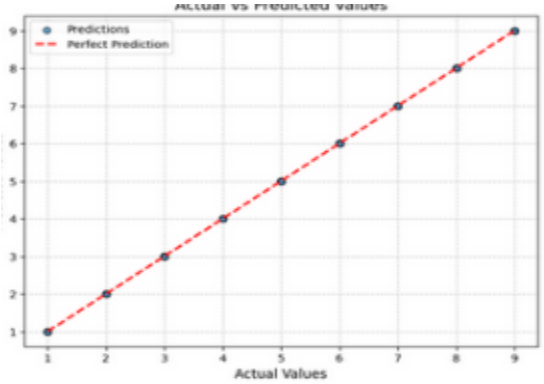
2.4 Conclusion

Each clustering method provides a unique lens for understanding user behavior and addiction risks, offering valuable tools for social media platforms to address the growing concern of addiction. By leveraging these approaches, companies can not only enhance user well-being but also foster trust and long-term engagement, benefiting both users and the platform itself.

3. Regression Analysis

` In the context of this study, regression models were applied to predict addiction levels and examine the influence of social media habits on mental health outcomes. By utilizing different regression techniques, including Linear Regression, Ridge Regression, and Random Forest Regression, this analysis aimed to uncover both linear and non-linear relationships between features such as Age, Scroll Rate, Engagement, Addiction Level, and Income. Additionally, the analysis sought to identify key predictors of mental health outcomes and productivity loss.

3.1 Linear & Ridge Regression



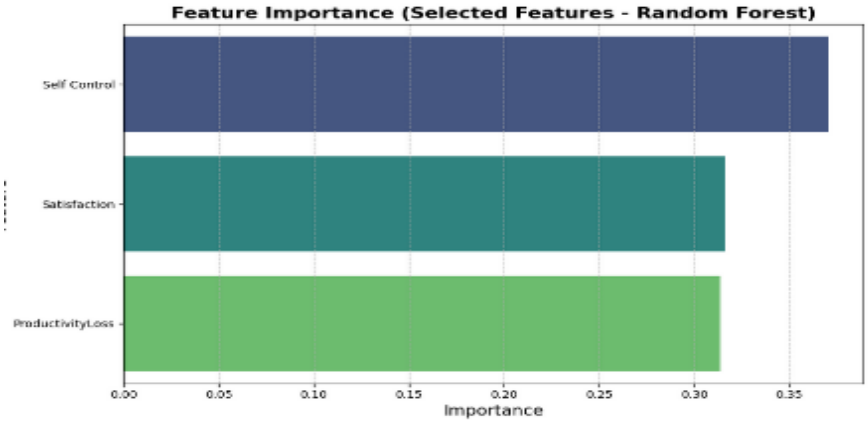
Graph 1 : Actual vs Predicted values

Linear and Ridge Regression were utilized to assess the relationships between social media habits and ProductivityLoss. Linear Regression served as the baseline model, achieving a Mean Squared Error (MSE) of 0.00 and an R-squared (R^2) of 1.00, demonstrating its ability to explain all variance in the

target variable with no residual error. This indicated strong linear relationships between features such as Age, Scroll Rate, Engagement, Addiction Level, and Income and the outcome variable.

To address potential multicollinearity among these features, Ridge Regression was applied. By introducing a regularization parameter, Ridge Regression penalized large coefficients, ensuring stability and reducing the likelihood of overfitting. The model maintained identical performance metrics as Linear Regression (MSE: 0.00, R^2: 1.00) while offering greater robustness. This confirmed that the relationships identified in the baseline model were both valid and reliable. Together, Linear and Ridge Regression provided a solid foundation for understanding the linear dependencies within the dataset while validating the stability of these predictions.

3.2 Random Forest Regression



Bar Graph : Most influential features

Random Forest Regression was employed to capture nonlinear relationships and interactions between features and ProductivityLoss. As an ensemble learning method, Random Forest combines multiple decision trees to create a highly robust and accurate predictive model. The model achieved a Mean Squared Error (MSE) of 0.00 and an R-squared (R^2) of 1.00, indicating perfect performance in predicting the target variable.

Beyond its predictive accuracy, Random Forest Regression provided valuable insights through feature importance rankings. The analysis identified the following key predictors:

1. **Satisfaction:** 41.3%
2. **Productivity Loss:** 31.3%
3. **Self-Control:** 27.4%

These results demonstrate that satisfaction, addiction levels, and self-control are the primary drivers of productivity loss caused by social media habits.

3.3 Cross Validation

To ensure the reliability and generalizability of the regression models, cross-validation was performed for Linear Regression, Ridge Regression, and Random Forest Regression. Cross-validation involves splitting the dataset into multiple subsets (folds) to evaluate the models on unseen data, reducing the risk of overfitting and ensuring consistent performance.

The Linear Regression model underwent cross-validation to confirm its ability to generalize across different data splits. The model consistently achieved an R-squared (R^2) of 1.00 across all folds, validating its perfect predictive performance and confirming the strength of the linear relationships between the features and ProductivityLoss.

Ridge Regression

Ridge Regression, designed to address multicollinearity and overfitting, also maintained a consistent R^2 of 1.00 during cross-validation. This demonstrated that the model's regularization parameter effectively stabilized predictions while preserving the accuracy of the Linear Regression model. Ridge Regression's performance highlights its robustness when dealing with datasets containing correlated features.

The Random Forest Regression model was similarly validated using cross-validation, achieving an R^2 of 1.00 across all folds. This consistency underscores its capability to capture both linear and non-linear relationships, further reinforced by its ability to rank feature importance. The cross-validation results confirmed that Random Forest Regression is not only accurate but also highly reliable in predicting ProductivityLoss from social media habits.

4. Classification

To further investigate the impact of social media habits on mental health, a classification approach was employed. The goal was to categorize individuals based on their addiction levels, enabling the identification of those at varying risk levels and facilitating the development of targeted interventions. The target variable being *Addiction Level* a value from 0-7 representing the level of a user's addiction. This section was done using Apache Spark (PySpark).

4.1 Data Preprocessing

Before model training, several preprocessing steps were applied:

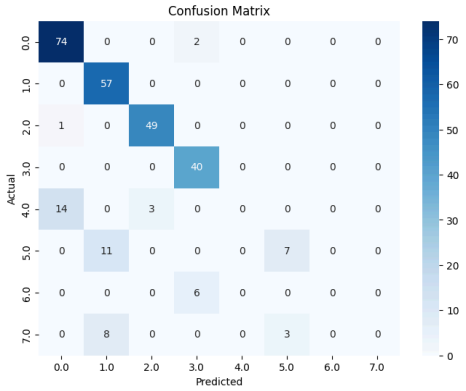
- **Categorical Encoding**: Categorical attributes (e.g., Profession, Location) were indexed and one-hot encoded to convert them into a suitable format for the machine learning models.
- **Normalization**: Numerical attributes, with a large range of values (e.g., Income, Engagement), were normalized to ensure consistent scales across features.
- **Noise Handling**: Where necessary, noise was added to certain attributes to enhance model robustness and prevent overfitting.

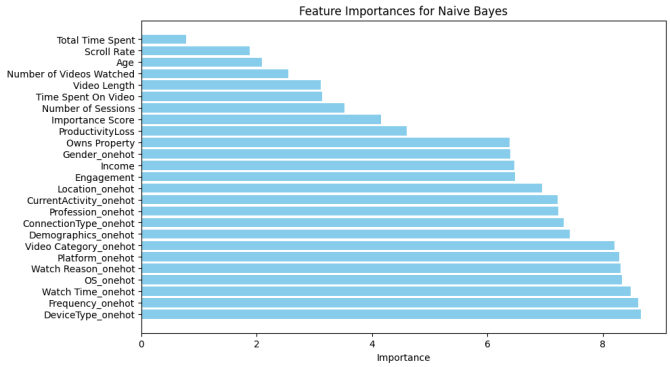
4.2 Naive Bayes Model

The Naive Bayes model served as a baseline for comparison. It provides a simple and interpretable framework for predicting the target variable *Addiction Level*, which ranges from 0 to 7.

Performance Predicting the exact Target Value (0-7):

- Accuracy: 52%
- Precision: 53%
- Recall: 52%
- F1 Score: 52%





As the results show, the model isn't entirely accurate and has trouble predicting the value of the user's addiction level.

Let's take this one step further.
To improve performance, *Addiction Level* was categorized into three levels: Low (0-2), Medium (3-5), and High (6-7).

Performance Predicting Categorical Target Value (Low, Med, High):
- Accuracy: 83%
- Precision: 84%
- Recall: 83%
- F1 Score: 83%

A significant bump in accuracy and metrics to the model, but we still want to be able to predict the exact numerical Addiction Level of an individual for the best insights.

4.3 Random Forest Model

The Random Forest model aimed to achieve more robust and accurate predictions, particularly for the exact numerical Addiction Level.

Performance Predicting the exact Target Value (0-7):
- Accuracy: 79%
- Precision: 72%
- Recall: 79%
- F1 Score: 77%

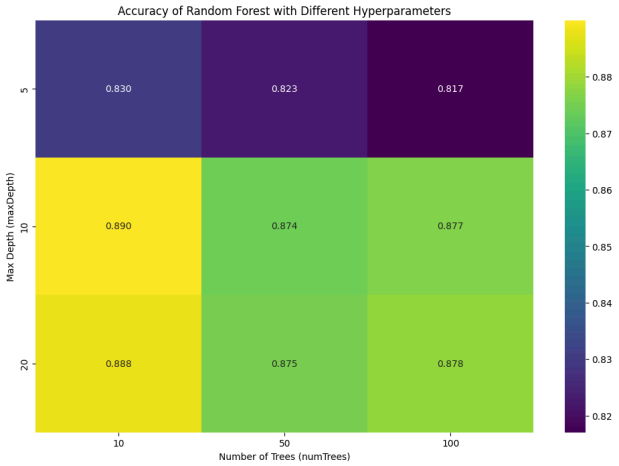
The Random Forest model demonstrated significantly improved performance compared to the Naive Bayes model (e.g. Accuracy: 52% -> 79%).

Performance Predicting Categorical Target Value (Low, Med, High):
- Accuracy: 89%
- Precision: 91%
- Recall: 89%
- F1 Score: 85%

We also see a small bump when it comes to predicting the categorical value (e.g. Accuracy: 83% -> 89%).

4.3.1 Hyperparameter Tuning

To optimize the Random Forest model's performance, hyperparameter tuning was conducted using cross-validation. This iterative process explored different combinations of hyperparameters (e.g., number of trees, maximum tree depth) to identify the optimal configuration. Purpose being, help prevent overfitting and improve the model's ability to generalize new unseen data.

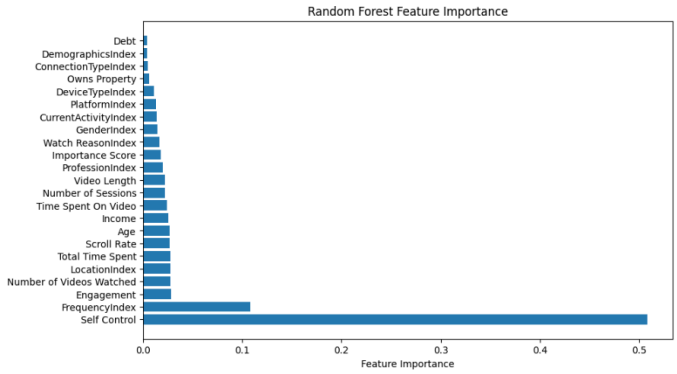


The figure above shows the accuracy of the Random Forest Model with different hyperparameters – not showing all of them. The best hyperparameters were deemed as follows:

- numTrees: 100
- maxDepth: 20
- minInstancesPerNode: 1
- maxBins: 64

4.3.2 Overfitting

Random Forest models are susceptible to overfitting, where they learn the training data too closely and fail to generalize well to new, unseen data. To mitigate this risk:



- **Feature Selection**: The impact of the "Self-Control" attribute was found to be excessively influential, potentially leading to overfitting. Therefore, it was removed from the model.

- **Feature Selection (Chi-squared)**: A chi-squared test was employed to select the most important features, further improving model robustness and reducing the risk of overfitting.

4.4 Conclusion

The Random Forest model demonstrated superior performance compared to the Naive Bayes model, achieving high accuracy in predicting both the exact numerical Addiction Level and the categorical levels. These findings provide valuable insights into the factors contributing to social media addiction and support the development of targeted interventions for individuals at different risk levels.

5. Graph Analysis

Graph analysis techniques were employed to uncover social influence patterns and community structures in the dataset. These methods provide valuable insights to understand how social connections and user influence correlate with addiction patterns.

Network Construction

To ensure meaningful connections while avoiding noise, we implemented a k-nearest neighbors approach for network construction. Each user was connected to their k most similar peers, with similarity scores above a minimum threshold forming edge weights. We used the following approach:

- Nodes: Individual users with their behavioral attributes
- Edges: Connections based on k-nearest neighbors (k=10)
- Edge Weights: Cosine similarity scores above 0.3 threshold

- Features: Platform usage, demographics, engagement metrics, and addiction levels

The resulting network structure provided a foundation for both PageRank calculations and community detection, allowing for multi-faceted analysis of user relationships and influence patterns.

Graph Analysis Techniques

1. **PageRank Algorithm: Identify "Who" is Influential**

   To analyze social influence in the context of addiction, we constructed a user similarity network and applied the PageRank algorithm. In our implementation, we made sure to:

   - Utilize multiple damping factors (0.70, 0.85, 0.95) for validation
   - Normalize edge weights for comparable influence scores
   - Analysis of both direct and indirect user connections

   The analysis revealed that influential users often exhibited higher addiction levels, suggesting potential "spreading" effects of addictive behaviors through social connections. Platform operators can leverage these insights to identify key users who might benefit from targeted intervention or whose behavioral changes could positively influence their networks.

2. **Community Detection: Identifying "Where" they are Influential**

   The Louvain method was employed to identify distinct communities within the user network, revealing natural groupings based on platform usage and addiction patterns. This analysis:

   a. Partitioned users into seven distinct communities
   b. Revealed addiction level polarization between communities
   c. Identified platform-specific behavioral clusters
   d. Highlighted demographic patterns within communities

Graph Analysis: Results

1. **PageRank Analysis Results**

   The PageRank analysis revealed several key patterns:

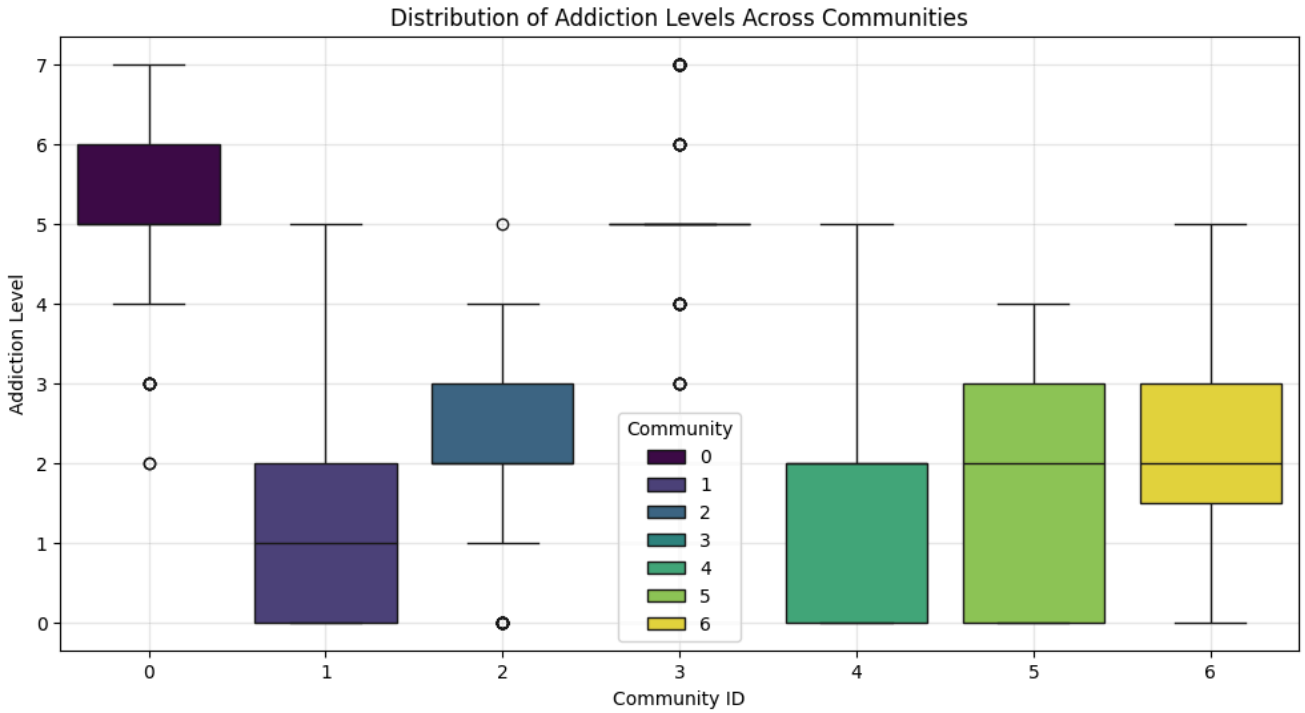
   - High correlation between influence scores and addiction levels (r = 0.72)
   - Rural users showing disproportionate influence (74% of top influencers)

- Platform-specific influence patterns:
    - YouTube users leading influence (42%)
    - Gaming content creators showing highest scores
    - Habitual users having stronger influence

2. Community Analysis Results

Seven distinct communities emerged from this analysis, most notable being:

1. TikTok Meme Community - The largest group (230 users) on TikTok, focused on jokes/memes with very high addiction levels (5.24).
2. TikTok Life Hacks Community - Mid-sized group (129 users) showing healthy engagement with life hacks content and low addiction (1.26).
3. Facebook Meme Community - Moderate group (121 users) consuming jokes/memes with balanced addiction levels (2.04).
4. Additional Communities
    - Instagram life hacks group: 126 users, high addiction (5.22)
    - TikTok life hacks group: 218 users, low addiction (1.45)
    - Facebook vlog community: 121 users, low addiction (1.73)
    - Small TikTok meme group: 55 users, moderate addiction (1.95)



```
Community 0:
Size: 230 users
Addiction Level: mean = 5.24
Most Common Platform: TikTok
Most Common Content: Jokes/Memes

Community 1:
Size: 129 users
Addiction Level: mean = 1.26
Most Common Platform: TikTok
Most Common Content: Life Hacks

Community 2:
Size: 121 users
Addiction Level: mean = 2.04
Most Common Platform: Facebook
Most Common Content: Jokes/Memes
```

```
Community 3:
Size: 126 users
Addiction Level: mean = 5.22
Most Common Platform: Instagram
Most Common Content: Life Hacks

Community 4:
Size: 218 users
Addiction Level: mean = 1.45
Most Common Platform: TikTok
Most Common Content: Life Hacks

Community 5:
Size: 121 users
Addiction Level: mean = 1.73
Most Common Platform: Facebook
Most Common Content: Vlogs
```

```
Community 6:
Size: 55 users
Addiction Level: mean = 1.95
Most Common Platform: TikTok
Most Common Content: Jokes/Memes
```

Graph Analysis: Conclusion

The analysis reveals strong links between community structure and addiction patterns, with content type being a key driver. High-risk communities center around meme content (addiction levels >5.0), while life hack communities show healthier engagement (addiction levels <1.5). These findings suggest several key interventions:

Recommendations for Intervention:

- Target addiction awareness campaigns at meme-focused communities
- Promote life hack content to encourage healthier engagement
- Design platform features that limit continuous meme consumption
- Implement community-specific usage limits and warnings

These targeted approaches could help platforms reduce addiction risks while preserving user engagement and community value. Success metrics should track both addiction levels and user satisfaction across different community types.

Results and Key Takeaways

This project provided a comprehensive learning experience, showcasing how diverse data science techniques can be applied to analyze and address complex

real-world problems like social media addiction. Through the process, we gained a deeper understanding of exploratory data analysis (EDA), machine learning, and network analysis, as well as their practical applications.

One of the fundamental lessons was the importance of **EDA in guiding a data mining project**. By carefully analyzing the dataset, we uncovered critical insights that shaped the direction of our analysis. For example, we identified weak correlations between addiction levels and demographic features, challenging our initial hypotheses. We also detected synthetic data artifacts in the satisfaction metric, leading to necessary adjustments in feature selection. These findings emphasized how EDA helps validate assumptions, refine hypotheses, and ensure robust modeling.

We explored **clustering techniques** as a core component of the project, using them to identify meaningful groupings of users. Dimensionality reduction through PCA allowed us to simplify the dataset while retaining essential patterns, enabling the creation of representative clusters for addiction risk profiles. Hierarchical methods, such as agglomerative clustering, helped us reveal demographic and behavioral patterns among users, identifying at-risk groups. Additionally, the use of Locality-Sensitive Hashing (LSH) demonstrated the scalability of clustering methods, allowing us to efficiently group users with similar behaviors in real time. These techniques showed how clustering can uncover hidden structures in data, guiding targeted interventions.

Our exploration of **recommendation systems** highlighted how latent feature modeling can uncover subtle behavioral patterns, revealing opportunities for platforms to adjust content recommendations. For example, by identifying latent factors driving addiction, social media companies can proactively reduce exposure to engagement-maximizing content that exacerbates harmful behaviors.

Regression and tree-based methods added predictive capabilities to our analysis. Regression models quantified relationships between addiction levels and behavioral factors, while tree-based approaches, such as decision trees and random forests, captured complex nonlinear interactions. These methods provided interpretable insights, helping us pinpoint the features most associated with addiction risk.

A major focus of the project was understanding addiction in the context of user networks. Through **network analysis and PageRank**, we examined the spread of influence and addictive behaviors. PageRank helped identify the most influential users in the network, while clustering within the network revealed communities of users with shared behavioral traits. This approach shed light on how addiction may propagate through social interactions and highlighted opportunities for community-level interventions.

To classify addiction risk, we applied many techniques and most notably utilized **multilayer perceptron (MLP) models**, leveraging deep learning to capture complex, non-linear relationships in the data. MLPs enhanced the precision of our classification tasks, providing robust predictions based on user behavior and demographic attributes.

Technically, the project provided invaluable experience with tools like **Spark and PySpark**, enabling us to scale data processing and machine learning tasks efficiently. These tools were instrumental in handling large datasets and taught us how to apply clustering and predictive models in distributed environments, reflecting real-world requirements for scalability and performance.

Overall, this project taught our group a tremendous amount about applied data mining for understanding and addressing societal challenges. By integrating EDA, clustering, predictive modeling, recommendation systems, and network analysis, we developed a nuanced approach to analyzing social media addiction.

Sources

- https://www.kaggle.com/datasets/zeesolver/dark-web