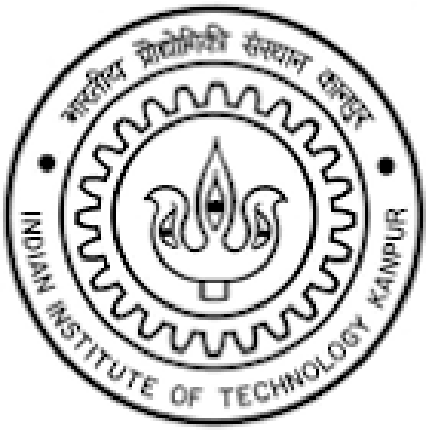

Indian Institute of Technology Kanpur

Social Media Addiction: A Visual Analytics Approach

CS661: Big Data Visual Analytics Project Report

2024–2025 Semester III

Team Members:

- Raj Aryan (230837)
- Anmol Gupta (230156)
- Divya Mhetre (230649)
- Pankhuri Sachan (230734)
- C. Venkata Pranaya (230324)
- Priyanshu Mishra (230806)
- Pranav Bharti (240726)
- Arkajyoti Santra (230194)
- Raj Shekhar (242110609)

Contents

1 Introduction **2**

1.1 Motivation . . . 2

1.2 Objectives . . . 2

2 Dataset Description **2**

2.1 Source . . . 2

2.2 Structure and Fields . . . 2

3 Data Processing **3**

3.1 Missing Data Handling . . . 3

3.2 Feature Engineering . . . 3

3.3 Categorical Encoding . . . 3

3.4 Normalization . . . 3

3.5 Final Dataset . . . 4

4 Tasks and Visual Analytics Goals **4**

5 Results and Visualizations **5**

5.1 Addiction Score Distribution by Demographics . . . 5

5.2 Addiction Score vs Academic Performance . . . 5

5.3 Addiction Score by Gender, Age, and Academic Level . . . 6

5.4 Average Sleep Hours by Addiction Level and Gender . . . 7

5.5 Cluster Grouping of Students . . . 7

5.6 Addiction Score by Most Used Platform . . . 8

6 Dashboard Implementation **9**

6.1 Tech Stack . . . 9

6.2 Features . . . 9

7 Design Justification **9**

8 Conclusion **10**

1. Introduction

1.1 Motivation

Social media addiction among students has emerged as a significant behavioral and academic issue. It not only affects academic performance but also disrupts sleeping habits and increases stress levels. Understanding these patterns through a visual interface is essential for awareness and corrective actions.

1.2 Objectives

The primary objectives of this project are:

- To build a dynamic and insightful visual interface that highlights trends in student social media behavior.
- To identify and segment addiction severity across various demographics such as gender, academic level, and age.
- To explore correlations between high addiction scores and academic decline, sleep deprivation, and mental health deterioration.
- To use clustering techniques to group students into meaningful behavioral profiles.
- To help educational institutions, counselors, and students themselves interpret these patterns and foster healthier digital habits.

2. Dataset Description

2.1 Source

The dataset used for this project is titled **"Students Social Media Addiction"** and is publicly available on Kaggle. It was created through a structured survey to analyze behavioral and psychological patterns among students caused by social media overuse. It can be accessed here: https://www.kaggle.com/datasets/pratyushpuri/students-social-media-addiction

2.2 Structure and Fields

The dataset consists of approximately 1000 entries, with each row representing an individual student's response. The structure is composed of the following key attributes:

- **Demographics:** Gender, Age, Academic Level, Country
- **Platform Preferences:** Most Used Platform, Device Used

- **Usage Metrics:** Average Daily Usage (hours), Sleep Hours Per Night
- **Addiction and Mental Health:** Addicted Score (numeric), Mental Health Score, Addiction Level (Low, Medium, High)
- **Academic Indicators:** GPA, Affects Academic Performance (Yes/No)

These attributes allow for multidimensional analysis, including behavioral clustering, addiction profiling, and correlation studies across personal, social, and academic dimensions.

3. Data Processing

Before performing visualization and analysis, the dataset required several preprocessing steps to clean, transform, and prepare the data for modeling and exploration. These transformations were necessary to ensure that all numerical and categorical variables were in a usable form.

3.1 Missing Data Handling

The dataset was checked for null or missing values. Records with essential fields such as addiction scores or academic performance responses missing were dropped to maintain analytical consistency. Non-critical missing entries (e.g., optional demographic fields) were either filled with mode or left as-is if not required.

3.2 Feature Engineering

To enhance interpretability:

- An `Addiction Level` field was created using binning on the continuous `Addicted Score`, dividing students into "Low," "Medium," and "High" addiction categories.
- The age attribute was transformed into categorical `Age Groups` (16–18, 19–21, 22–24) to aid in age-based visual grouping.

3.3 Categorical Encoding

Categorical variables like Gender, Academic Level, and Platform Used were label encoded for use in clustering algorithms and numerical comparison. These encoded values preserved the categorical structure while making the data machine-readable.

3.4 Normalization

To ensure scale consistency across features like Sleep Hours, Mental Health Score, and Daily Usage, normalization (min-max scaling) was applied prior to clustering and dimensionality reduction steps (e.g., t-SNE).

3.5 Final Dataset

After cleaning and feature transformation, the final dataset had:

- No null or invalid entries
- Properly labeled age, academic, and addiction levels
- Standardized features suitable for visualization and clustering

4. Tasks and Visual Analytics Goals

The following visual and analytical tasks were designed to address the objectives of the project using meaningful data representations:

- **Examine demographic variation in addiction scores:** To understand how addiction severity differs across gender, academic level, and age group.
- **Investigate the impact of social media on academic performance:** To explore the correlation between addiction scores and reported academic decline, using both scatter plots and trend lines.
- **Compare sleep patterns across addiction levels:** To identify whether students with high addiction scores also report lower average sleep durations.
- **Analyze platform preference in relation to addiction severity:** To determine which social media platforms are most associated with high addiction levels.
- **Identify student behavior clusters:** To segment students into behavioral groups based on addiction, sleep, and device usage using t-SNE and clustering algorithms.
- **Enable interactive filtering of visualizations:** To support personalized data exploration through filters based on academic level, age, and gender.

Each of these tasks was carefully mapped to a corresponding visualization technique, ensuring that insights were not just visible but also interpretable and actionable.

5. Results and Visualizations

5.1 Addiction Score Distribution by Demographics

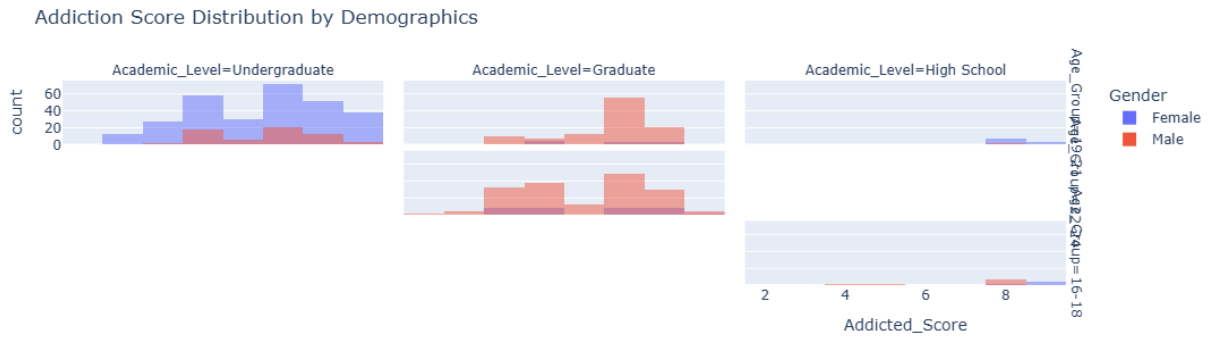

Figure 1: Histogram of Addiction Scores segmented by Gender and Academic Level

- This histogram provides a detailed view of how social media addiction scores are distributed among students across different academic levels and genders.
- It enables comparison of behavioral patterns between males and females, and how those patterns shift across undergraduate, graduate, and high school categories.
- The results show that undergraduate males tend to have both higher frequency and a wider range of addiction scores, suggesting they may be more vulnerable to addictive usage patterns.
- Female students generally demonstrate more centralized distributions, with moderate levels of addiction.
- This visualization supports the identification of key segments for targeted behavioral interventions and digital wellness initiatives.

5.2 Addiction Score vs Academic Performance

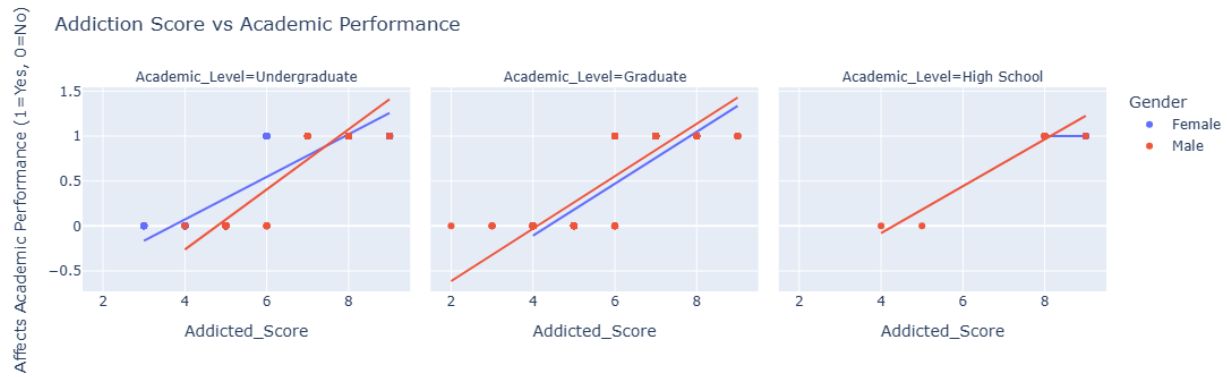

Figure 2: Scatter plot showing correlation between Addiction Score and Academic Impact

- This scatter plot uses regression lines to visualize the relationship between addiction scores and academic disruption, as reported by students.
- The graph is divided by academic level and gender, offering a multi-layered view of how digital overuse affects academic performance across different groups.
- A positive correlation is observed — higher addiction scores are frequently associated with academic struggles, especially in graduate students.
- Regression trends validate the hypothesis that social media addiction negatively impacts academic outcomes.
- This visualization helps stakeholders quantify the trade-off between digital engagement and academic productivity.

5.3 Addiction Score by Gender, Age, and Academic Level

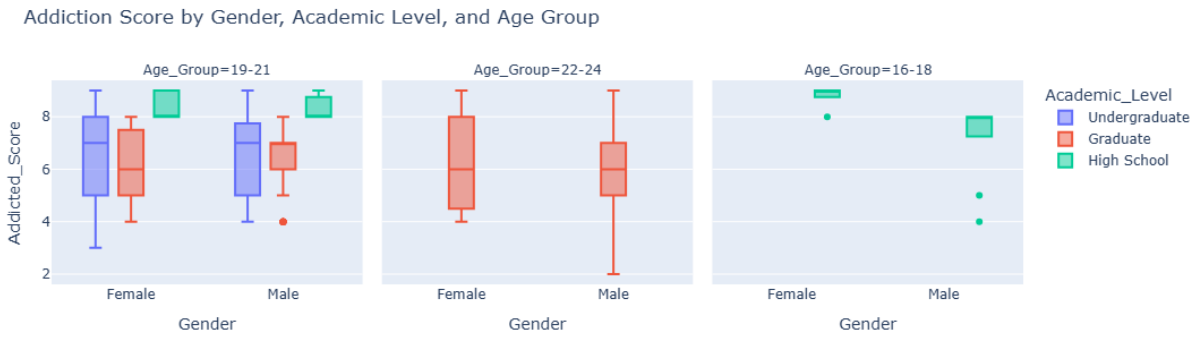

Figure 3: Box plot of Addiction Scores by Gender, Age Group, and Academic Level

- This box plot compares addiction scores based on gender, age group, and academic level, providing a three-dimensional demographic analysis.
- It highlights that students aged 16–18 tend to exhibit higher median addiction scores than older students.
- Male students show wider score variability, indicating a more diverse range of behavioral patterns.
- High school and early undergrad students are generally more prone to extreme usage patterns.
- The visualization makes it easier to identify outliers, medians, and quartiles for different demographic slices, guiding more personalized awareness efforts.

5.4 Average Sleep Hours by Addiction Level and Gender

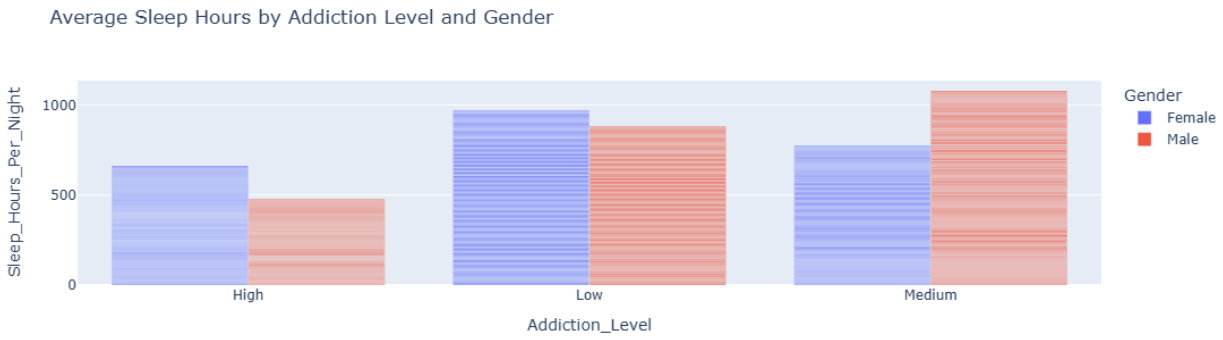

Figure 4: Grouped Bar Chart of Sleep Hours segmented by Addiction Level and Gender

- This grouped bar chart investigates how average sleep hours vary across addiction levels and between genders.
- Students with high addiction scores consistently report lower sleep durations, highlighting the toll of excessive screen time.
- The difference is especially pronounced among male students, who appear more susceptible to sleep disruption from addictive platform usage.
- The visualization effectively bridges behavioral (addiction) and health (sleep) data points to create a holistic view.
- Such analysis helps educators and health professionals link digital habits with wellness risks and recommend lifestyle improvements.

5.5 Cluster Grouping of Students

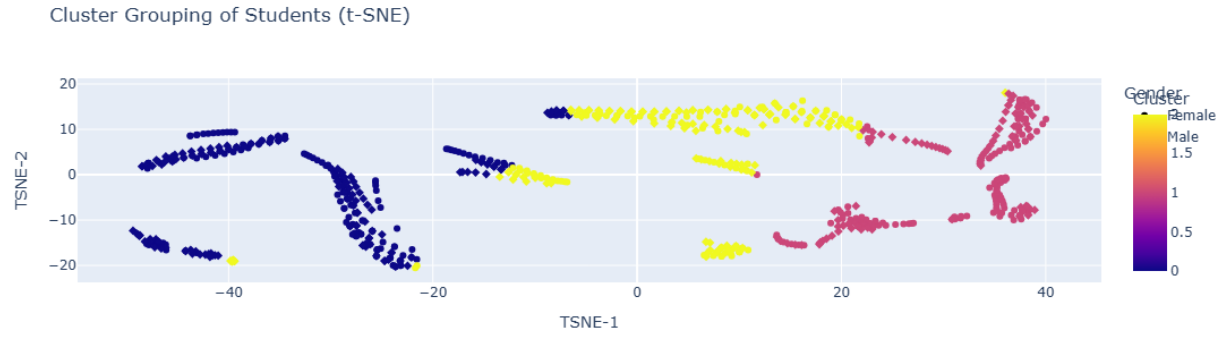

Figure 5: t-SNE Projection of Clusters based on Addiction, Sleep, and Usage

- This scatter plot uses t-SNE (t-distributed stochastic neighbor embedding) to reduce high-dimensional behavioral data into a 2D visualization.
- KMeans clustering is applied to segment students into meaningful behavior groups.
- One cluster includes heavy users with low sleep and high addiction; another includes moderate users with balanced sleep and usage; and a third represents low-risk students.
- This unsupervised learning approach uncovers hidden patterns not easily visible in raw statistics.
- The clusters provide actionable insights for intervention — for example, counseling high-risk groups while promoting behavior seen in the low-risk cluster.

5.6 Addiction Score by Most Used Platform

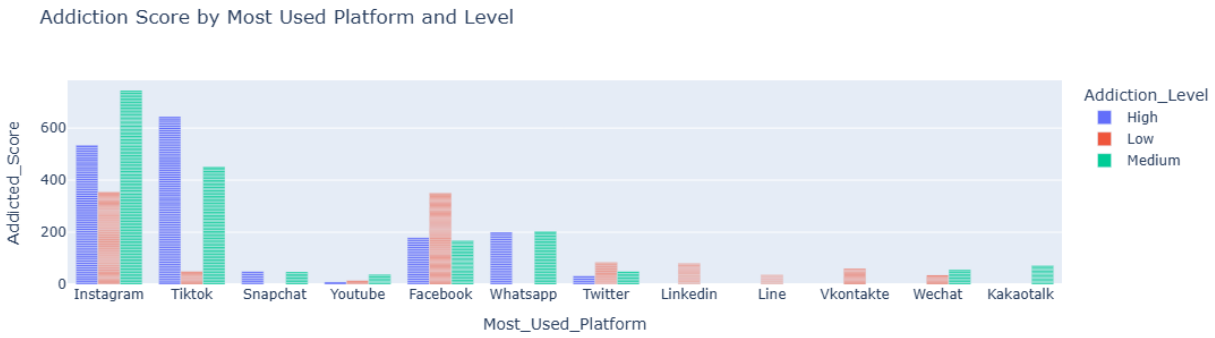

Figure 6: Addiction Score across Platforms and Addiction Levels

- This bar chart shows how students' most-used platforms correlate with their addiction score levels (Low, Medium, High).
- Platforms like Instagram and TikTok have the highest concentration of students in the High Addiction category.
- Platforms with more passive or educational use (like LinkedIn or Twitter) show much lower addiction scores.
- These insights align with the design intent of these platforms — some foster dopamine-driven scrolling, while others encourage targeted engagement.
- The visualization supports recommendations for platform usage limits or digital detox strategies for specific high-risk apps.

6. Dashboard Implementation

6.1 Tech Stack

- Python Dash for frontend interaction
- Pandas and NumPy for data handling
- Scikit-learn for clustering and preprocessing
- Plotly for all visualizations

6.2 Features

- Dynamic plots based on filters (gender, age, academic level)
- Interactive scatter, histogram, and bar charts
- Clustering-based insights through 2D projection
- Responsive layout with hover tooltips
- Modular design with tabs for organized insights

The interface was built to cater to both technical and non-technical users, ensuring that visual insights are intuitive and easy to explore.

7. Design Justification

We deliberately chose specific plot types for different insights, based on:

- **Histogram:** Helps in visualizing the distribution of addiction scores. Useful for detecting skewness, peaks, and gaps across subgroups.
- **Box Plot:** Offers a compact summary of medians, quartiles, and outliers — ideal for comparing addiction scores by gender, academic level, and age simultaneously.
- **Scatter Plot (with Regression):** Allows us to identify correlation between addiction and academic disruption while showing data spread.
- **Grouped Bar Chart:** Best suited for comparative visualizations like average sleep hours across addiction levels and genders.
- **t-SNE Cluster Plot:** Helps project multi-dimensional behavior (e.g., addiction + sleep + usage hours) into 2D for easy visual segmentation.

Each visualization was chosen to support a clear analytical purpose, avoiding clutter or decorative plots with no insight.

8. Conclusion

The project successfully achieves its aim of uncovering and visualizing social media addiction patterns among students. The data reveals compelling insights:

- Undergraduate students, especially males, show higher and more variable addiction scores.
- Instagram and TikTok are most correlated with high addiction levels.
- There is a clear link between high addiction and lower average sleep hours.
- Academic performance is more likely to be affected in students with higher addiction scores.
- Clustering analysis groups students into interpretable behavior segments, showing real value in unsupervised learning.

The dashboard interface makes it easy for users to explore these findings on their own, making this project not just an academic exercise but a practical tool. With rising concerns over digital wellness in academia, this tool provides the foundation for meaningful discussions, research, and interventions.

References

- Students Social Media Addiction Dataset: https://www.kaggle.com/datasets/pratyushpuri/students-social-media-addiction
- Plotly Python Graphing Library: https://plotly.com/python/
- Dash by Plotly: https://dash.plotly.com/
- t-SNE Methodology: van der Maaten and Hinton, 2008