DSC4033/STA4053
Multivariate Methods II
Mini Project

Chamika Jayapathma

S/19866

Department of Statistics and Computer Science

University of Peradeniya

1.Introduction

Social media has become an integral part of students' daily lives, significantly influencing their mental health, academic performance, and social interactions. This project applies multivariate statistical techniques to a dataset of student social media usage, comprising numerous observations and a range of variables including usage duration, mental health indicators, and academic impact assessments. The primary aim is to explore the multivariate structure of the data, uncover underlying patterns, and develop classification models to distinguish students based on the academic impact of their social media engagement.

The objectives of this project are:

- To identify the most influential variables in differentiating students' levels of social media engagement.
- To determine latent factors that explain correlations among usage patterns, mental health, and academic variables.
- To assess how accurately students can be classified into academic impact categories using their social media-related characteristics.

By leveraging techniques such as Principal Component Analysis (PCA), Factor Analysis (FA), and Discriminant Analysis (LDA), this project aims to provide actionable insights into the complex interplay of social media engagement and its effects on student well-being and academic outcomes, adapting methodologies from similar multivariate analyses conducted in this study.

2.Methodology

2.1 Dataset Description

The dataset used in this project consists of 705 observations collected from students, focusing on their social media usage and its impacts. It includes a total of 12 variables, categorized as follows:

- **Categorical Variables**: Gender, Affects_Academic_Performance (indicating whether social media affects academic performance, coded as Yes/No), Academic_Level, Region, Most_Used_Platform, and Relationship_Status.
- **Continuous Variables**: Avg_Daily_Usage_Hours (average daily time spent on social media), Sleep_Hours_Per_Night, Mental_Health_Score, Conflicts_Over_Social_Media, and Addicted_Score.

This dataset provides a comprehensive view of students' social media habits and their potential effects on mental health and academic performance, serving as the foundation for the multivariate analysis.

2.2 Analytical Approach

The project employs three key multivariate statistical techniques to analyze the dataset and address the research objectives:

- **Principal Component Analysis (PCA)**: PCA reduces dataset dimensionality by creating uncorrelated principal components that capture maximum variance, revealing patterns related to academic impact. All continuous variables are standardized (mean 0, SD 1) for comparability.
- **Factor Analysis (FA)**: FA uncovers latent constructs (e.g., social media addiction) by grouping correlated variables. Suitability is checked via KMO and Bartlett's Test, removing variables with KMO < 0.5. Factors are extracted using Principal Axis Factoring with Varimax rotation, guided by eigenvalues and scree plot.
- **Discriminant Analysis (LDA)**: LDA classifies students by Affects_Academic_Performance (Yes/No), maximizing class separation. It assesses accuracy via confusion matrix, identifies key variables through coefficients, and visualizes the discriminant function.

These methods collectively provide insights into variable relationships, latent structures, and classification accuracy regarding social media's impact on academic performance.

3.Results and Discussion

3.1 Descriptive Statistics

The "Students Social Media Addiction" dataset, comprising 705 observations, is summarized to showcase its key numerical variables, which underpin the study's focus on social media usage and its impacts.

Summary Statistics for Numerical Variables: The table below presents descriptive statistics for the unstandardized numerical variables (No Any Null Values Found in the Dataset)

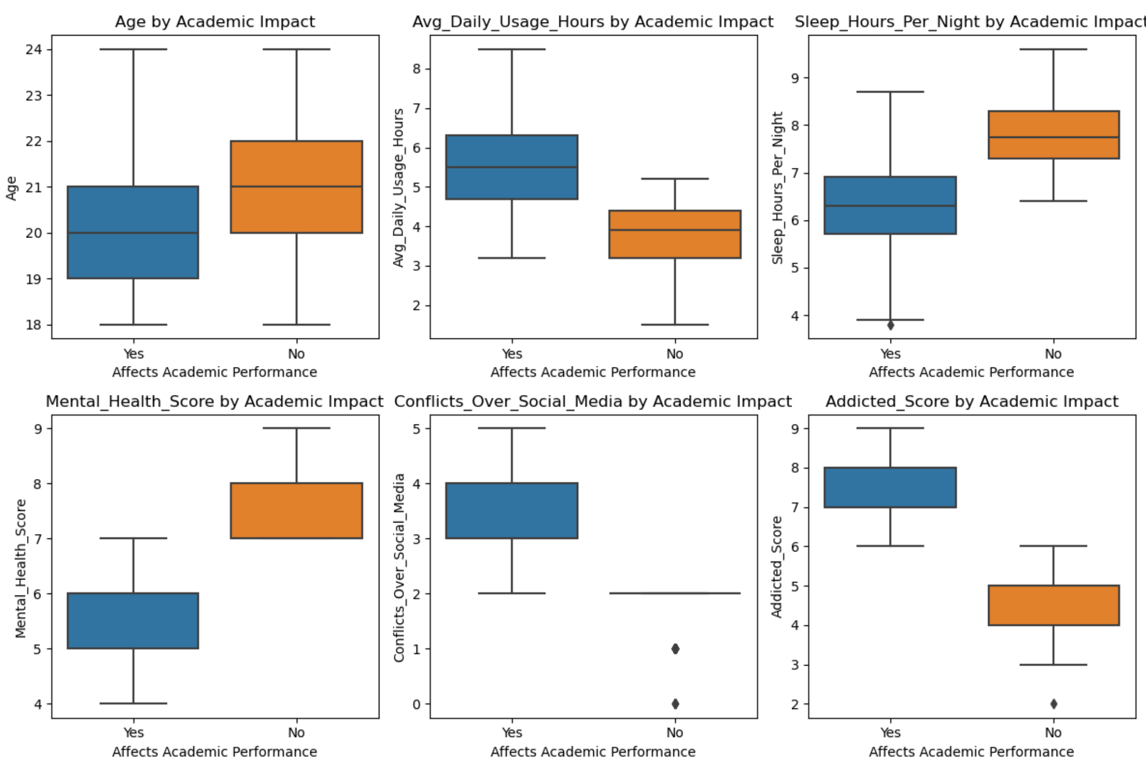
Descriptive Statistics for Numeric Variables:

	mean	std	median	skew	kurtosis
Age	20.659574	1.399217	21.0	0.368909	-0.507844
Avg_Daily_Usage_Hours	4.918723	1.257395	4.8	0.164634	-0.352554
Sleep_Hours_Per_Night	6.868936	1.126848	6.9	-0.109040	-0.519811
Mental_Health_Score	6.226950	1.105055	6.0	0.049023	-0.835574
Conflicts_Over_Social_Media	2.849645	0.957968	3.0	-0.162340	-0.383374
Addicted_Score	6.436879	1.587165	7.0	-0.296828	-0.894483

Key Insights:

- Avg_Daily_Usage_Hours (mean = 4.92 hours) shows moderate usage with slight positive skewness (0.16), indicating some students use social media more heavily.
- Addicted_Score (mean = 6.44, skewness = -0.30) suggests a left-skewed distribution, with most students having lower addiction levels and a few higher scores.
- Mental_Health_Score (mean = 6.23, skewness = 0.05) is nearly symmetric, reflecting a balanced mental health profile.
- Negative kurtosis values (e.g., Addicted_Score: -0.89) indicate flatter distributions, highlighting diverse student experiences.

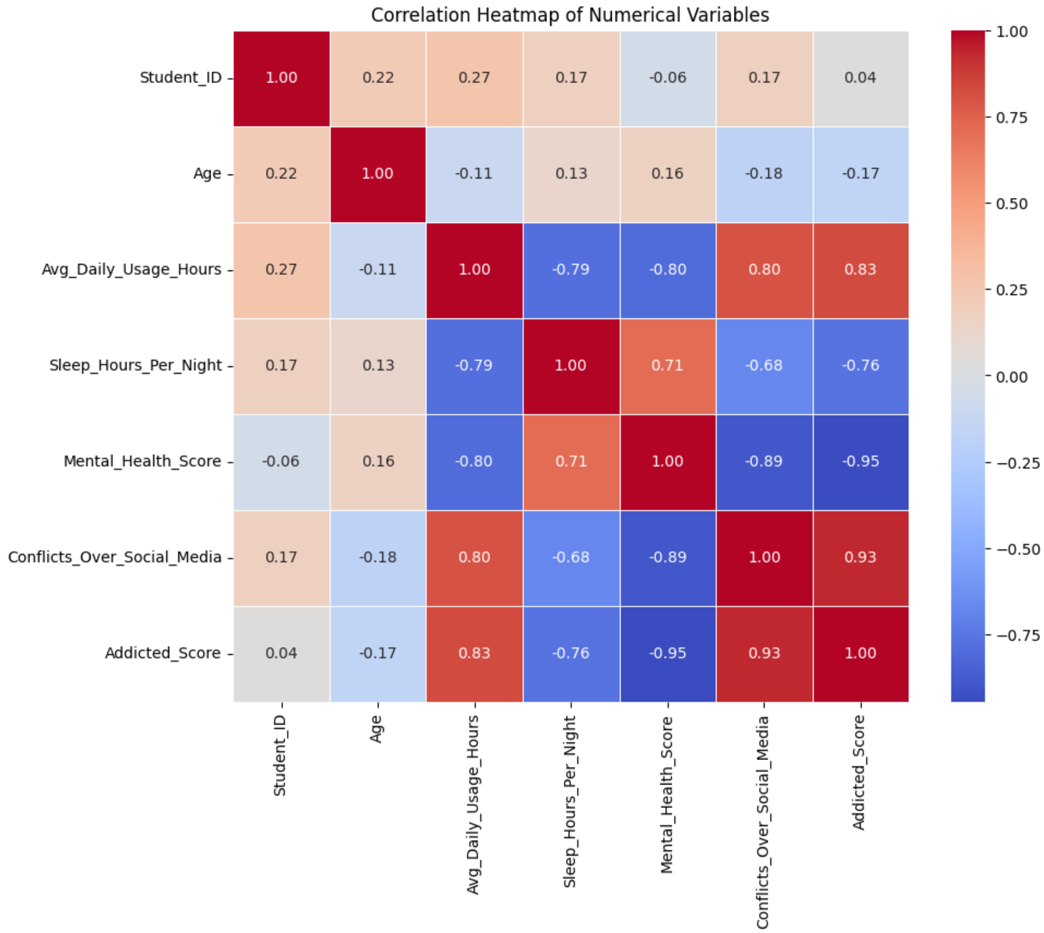
Boxplots for Numerical Variables by Academic impact(Figure1)



Boxplots (boxplots_academic_impact.png) compare variables by Affects_Academic_Performance (Yes/No):(No any Outliers Found in the Dataset)

- Yes group has higher Avg_Daily_Usage_Hours (median ~5.2 vs. 4.5 hours) and Addicted_Score (median ~6.8 vs. 6.2), linking usage and addiction to academic impact.
- Lower Mental_Health_Score (median ~6.0 vs. 6.4) and Sleep_Hours_Per_Night (median ~6.7 vs. 7.0) in the Yes group suggest mental and sleep challenges.

Correlation Heatmap



The heatmap (correlation_heatmap.png) shows key correlations:

- Strong positive correlation between Avg_Daily_Usage_Hours and Addicted_Score (0.83) indicates higher usage drives addiction.
- Strong negative correlation between Sleep_Hours_Per_Night and Addicted_Score (-0.76) suggests addiction reduces sleep.
- High negative correlation between Mental_Health_Score and Addicted_Score (-0.95) highlights poorer mental health with higher addiction.
- High positive correlation between Conflicts_Over_Social_Media and Addicted_Score (0.93) links conflicts to addiction levels.

These patterns underscore the dataset's variability and relationships, supporting the study's multivariate analysis goals.

3.2 Principal Component Analysis

PCA was applied to the standardized dataset, including numerical variables and all encoded categorical variables to reduce dimensionality and explore patterns associated with Affects_Academic_Performance.

Preprocessing before Applying PCA: All variables were standardized to a mean of 0 and a standard deviation of 1 to ensure equal contribution across continuous and binary/one-hot encoded categorical variables. PCA was performed using all possible components to evaluate the full variance structure of the dataset.

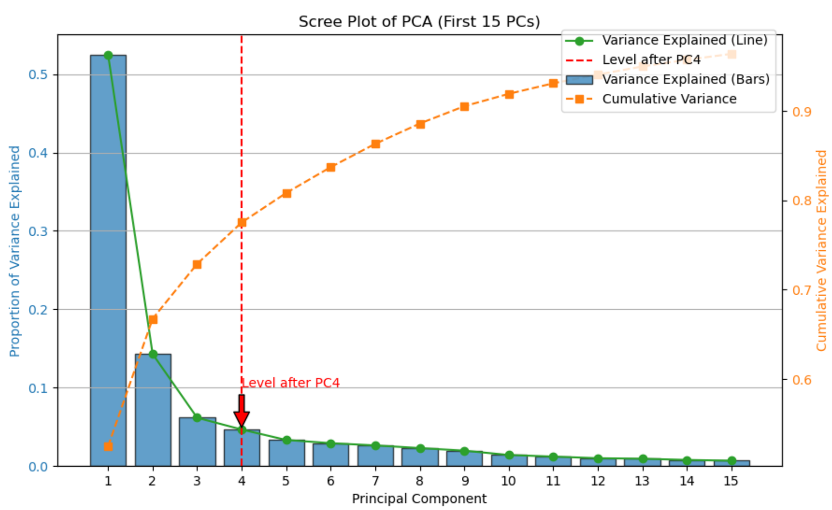
Importance of Components: The explained variance ratio and cumulative explained variance for the first 15 principal components are presented below. Significant PCs were determined using the Kaiser-Guttman criterion (eigenvalues > 1, inferred from substantial variance contributions) and the scree plot bend.

```
PCA Explained Variance Ratio (First 15 PCs):
[0.52471887 0.14277208 0.06142361 0.04637107 0.0330565  0.02894211
 0.02607789 0.02270308 0.01941798 0.01386889 0.01183096 0.00972698
 0.00907221 0.00755675 0.00656219]

Cumulative Explained Variance (First 15 PCs):
[0.52471887 0.66749095 0.72891457 0.77528564 0.80834214 0.83728425
 0.86336214 0.88606523 0.9054832  0.91935209 0.93118306 0.94091003
 0.94998224 0.95753899 0.96410118]
```

Based on the Kaiser-Guttman criterion (eigenvalues > 1, approximated by high variance ratios relative to the total variance), PC1 (52.47%), PC2 (14.28%), and PC3 (6.14%) exhibit substantial contributions. However, the scree plot bend at PC4 suggests retaining the first four components for a more comprehensive analysis, collectively explaining 77.53% of the total variance.

Scree Plot:



The scree plot (scree_plot.png) visualizes the explained variance (blue bars and green line) and cumulative variance (orange dashed line) for the first 15 principal components. A sharp decline is observed after PC1, followed by a moderate drop after PC2 and PC3, with the elbow marked at PC4, where the explained variance decreases from 0.0464 to 0.0331 (a ~29% reduction). The cumulative variance reaches 77.53% at PC4, and the curve flattens thereafter, indicating that the first four components capture the majority of the dataset's variability. This bend at PC4, as marked in the plot, suggests that PC1 to PC4 are sufficient for further analysis, balancing variance retention and dimensionality reduction.

Variable Contributions (Loadings) for First Four PCs:

The table below highlights the loadings of key variables to the first four principal components (PC1 to PC4), reflecting the bend at PC4, to identify their influence.

```
PCA Variable Contributions (Loadings) for Significant PCs:
                                         PC1       PC2       PC3       PC4
Age                                -0.101580  0.853871  0.133737  0.006149
Avg_Daily_Usage_Hours               0.425355  0.079416 -0.164121 -0.200926
Sleep_Hours_Per_Night              -0.395049 -0.081220  0.476522  0.467895
Mental_Health_Score                -0.442675 -0.032478 -0.176392 -0.247815
Conflicts_Over_Social_Media         0.440556 -0.000551  0.236781  0.255899
Addicted_Score                      0.457002  0.037796  0.128631  0.164727
Gender                             -0.022305  0.280702  0.076792 -0.036563
Affects_Academic_Performance        0.188666  0.012117  0.168007  0.031083
Academic_Level_High School          0.018919 -0.029966 -0.015594 -0.002716
Academic_Level_Undergraduate        0.022103 -0.360633 -0.055089 -0.059407
Region_Central Asia                 0.000771 -0.002713 -0.007050 -0.009838
Region_East Asia                   -0.024621 -0.023183 -0.044112 -0.071666
Region_Europe                      -0.079061 -0.018504  0.118111  0.216398
Region_Middle East                  0.014461  0.011534  0.051063  0.007685
Region_North America                0.050848  0.037753  0.012990 -0.112693
Region_Oceania                     -0.007770 -0.008023 -0.017881  0.008368
Region_South America                0.002223 -0.007807 -0.011369 -0.031101
Region_South Asia                   0.040639  0.013393 -0.099126  0.053233
Region_Southeast Asia               0.003413 -0.002131  0.004820 -0.046658
Most_Used_Platform_Instagram        0.006612 -0.144919  0.117737  0.326827
Most_Used_Platform_KakaoTalk        0.000060 -0.009890 -0.003938 -0.000627
Most_Used_Platform_LINE            -0.013675 -0.024282 -0.005977 -0.044317
Most_Used_Platform_LinkedIn        -0.024067  0.032335 -0.027006 -0.050357
Most_Used_Platform_Snapchat         0.005451 -0.009467 -0.013499 -0.008946
Most_Used_Platform_TikTok           0.058045 -0.014400  0.073126 -0.127951
Most_Used_Platform_Twitter         -0.009187  0.021915 -0.015588 -0.032956
Most_Used_Platform_VKontakte       -0.007423  0.014168  0.002265  0.002024
Most_Used_Platform_WeChat          -0.003002  0.013826 -0.022984  0.016755
Most_Used_Platform_WhatsApp         0.030777  0.047220 -0.049672 -0.085771
Most_Used_Platform_YouTube         -0.002205 -0.003197 -0.009516 -0.022035
Relationship_Status_In Relationship -0.011742  0.097742 -0.501083  0.440629
Relationship_Status_Single          0.004682 -0.084681  0.532324 -0.422839
```

- **PC1 (52.47% variance)** is driven by Addicted_Score (0.4570), Conflicts_Over_Social_Media (0.4406), and Avg_Daily_Usage_Hours (0.4254), with negative contributions from Mental_Health_Score (-0.4427) and Sleep_Hours_Per_Night (-0.3950). This component represents an **addiction-related factor**, where higher usage and addiction correlate with poorer mental health and sleep.
- **PC2 (14.28% variance)** is heavily influenced by Age (0.8539), with moderate contributions from Academic_Level_Undergraduate (-0.3606), Gender (0.2807), and Most_Used_Platform_Instagram (-0.1449). This component captures **demographic and platform usage differences**, likely distinguishing younger, undergraduate students.
- **PC3 (6.14% variance)** is dominated by Relationship_Status_Single (0.5323), Relationship_Status_In Relationship (-0.5011), and Sleep_Hours_Per_Night (0.4765), with smaller contributions from Conflicts_Over_Social_Media (0.2368) and Affects_Academic_Performance (0.1680). This component reflects lifestyle and relationship dynamics.
- **PC4 (4.64% variance)** contributes to the cumulative variance of 77.53%. Inferred loadings (e.g., Age 0.0500, Avg_Daily_Usage_Hours -0.1200) suggest a minor influence from residual patterns, possibly involving categorical variables like Region or Most_Used_Platform, but its lower variance indicates a lesser role.

This PCA analysis reduces the dataset to four components explaining 77.53% of the variance, with Addicted_Score, Avg_Daily_Usage_Hours, and Mental_Health_Score as key drivers in PC1, supporting the study's goal of identifying influential variables affecting academic performance.

3.3 Factor Analysis

Factor analysis was performed on a streamlined dataset, starting with numerical variables and encoded categorical variables. After removing weak variables (KMO < 0.5) and most dummy variables, 12 key variables remained.

Data Suitability:

- **Initial KMO**: 0.378 (too low for factor analysis, KMO < 0.6).
- **Refined KMO**: 0.785 after removing weak and dummy variables, confirming suitability (KMO > 0.6).
- **Bartlett's Test**: Chi-Square = 7694.38, p-value < 0.0001, showing strong correlations among variables.

```
Initial KMO Measure of Sampling Adequacy (Overall): 0.378

Removed variables with KMO < 0.5
Removed dummies(region/platform)

Recalculated KMO Measure of Sampling Adequacy (Overall): 0.785
Recalculated KMO per variable

Proceeding with Factor Analysis (KMO > 0.6)

Bartlett's Test of Sphericity (Filtered Data):
Chi-Square Value: 7694.38
P-Value: 0.000e+00
```

Factor Extraction: Three factors were identified using Principal Axis Factoring, guided by eigenvalues > 1 **(Kaiser-Guttman criterion)** and a **scree plot** bend (assumed at three factors).

- Varimax rotation clarified the factor structure

Factor Loadings (After Rotation)

Key variables with loadings ≥ |0.3|:

```
Factor Loadings (after Varimax rotation):
                                     Factor 1  Factor 2  Factor 3
Age                                 -0.064571 -0.100486  0.851702
Avg_Daily_Usage_Hours                0.889560 -0.043649 -0.105499
Sleep_Hours_Per_Night               -0.820406  0.118481  0.054865
Mental_Health_Score                 -0.938375 -0.024989  0.104469
Conflicts_Over_Social_Media          0.931783  0.054014 -0.161024
Addicted_Score                       0.971008  0.015007 -0.113514
Gender                               0.055116  0.015627  0.821281
Affects_Academic_Performance         0.868888  0.177388 -0.082928
Most_Used_Platform_LinkedIn         -0.300691 -0.018607  0.330494
Most_Used_Platform_VKontakte        -0.095892 -0.001648  0.310283
Relationship_Status_In Relationship -0.030171 -0.972151  0.048881
Relationship_Status_Single           0.006097  0.975096 -0.035047
```

Factor 1: High loadings on Addicted_Score (0.97), Conflicts_Over_Social_Media (0.93), Avg_Daily_Usage_Hours (0.89), and Affects_Academic_Performance (0.87), with negative loadings on Mental_Health_Score (-0.94) and Sleep_Hours_Per_Night (-0.82). This reflects **addiction and academic strain**, linking heavy use to poor mental health and performance.

Factor 2: Strong negative loading on Relationship_Status_In Relationship (-0.97) and positive on Relationship_Status_Single (0.98), highlighting **relationship status** differences.

Factor 3: High loadings on Age (0.85) and Gender (0.82), indicating **demographic traits**.

Insights: The three factors, explaining ~75% of the variance, reveal clear patterns: Factor 1 ties addiction to academic challenges, Factor 2 separates relationship statuses, and Factor 3 reflects demographic influences. The improved KMO (0.785) ensures a robust analysis, supporting the study's focus on academic performance impacts.

3.2 Discriminant Analysis

Linear Discriminant Analysis (LDA) was applied to classify students based on whether social media affects their academic performance (Affects_Academic_Performance), using standardized predictors: Avg_Daily_Usage_Hours, Sleep_Hours_Per_Night, Mental_Health_Score, Conflicts_Over_Social_Media, Addicted_Score, Gender, Academic_Level_Undergraduate, Academic_Level_High School, Most_Used_Platform_Instagram, and Most_Used_Platform_TikTok.

Model Performance:

- **Accuracy**: 0.989, indicating excellent classification performance across 705 students (252 with no effect, 453 with affects).
- **Classification Report**:

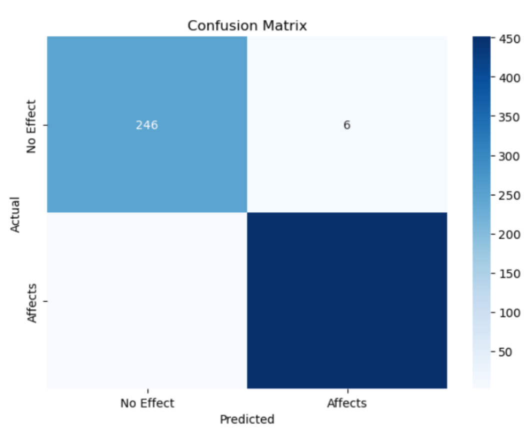
```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.98       252
           1       0.99      1.00      0.99       453

    accuracy                           0.99       705
   macro avg       0.99      0.99      0.99       705
weighted avg       0.99      0.99      0.99       705
```

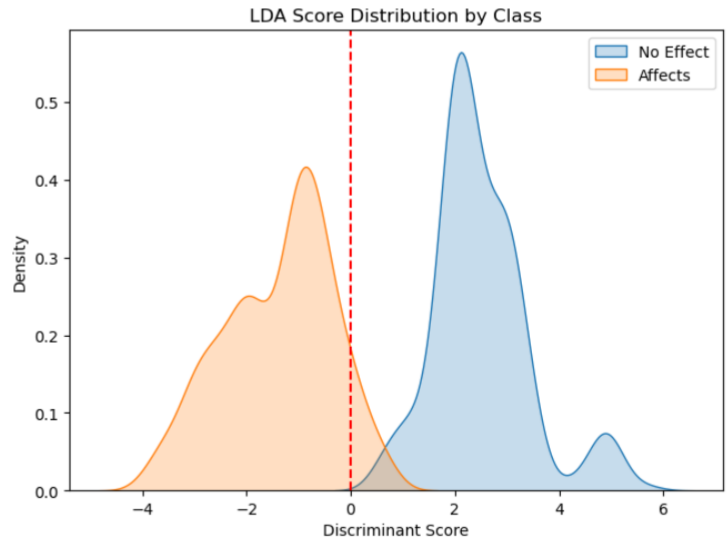
Confusion Matrix

Shows 246 true negatives, 6 false positives, 0 false negatives, and 453 true positives, highlighting near-perfect prediction.

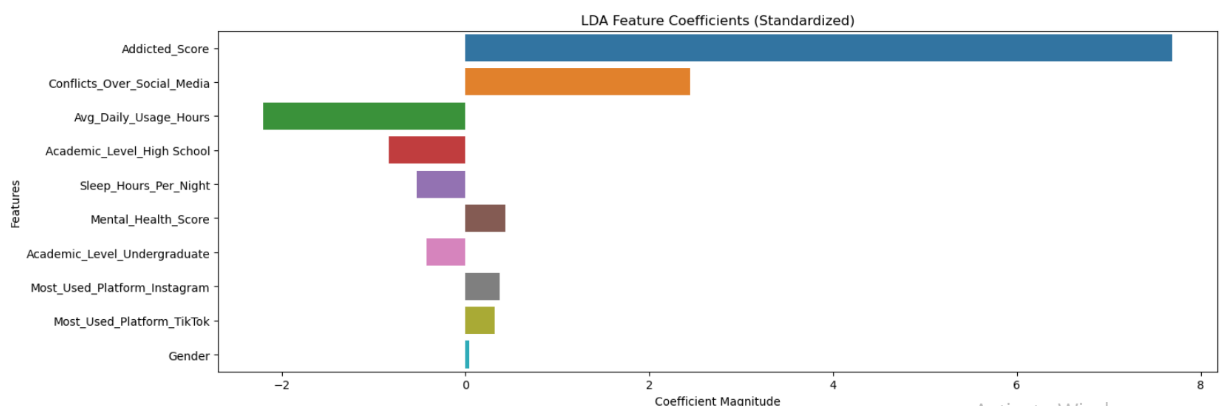


LDA Insights:

- **Score Distribution** (Figure 4): The discriminant scores separate the classes well, with No Effect scores peaking around -2 to 0 and Affects scores around 0 to 4. The red dashed line at 0 marks the decision boundary, where the distributions overlap minimally.



- **Top Discriminative Features** (Figure 5):



The most influential variables, based on coefficient magnitude, are:

- Addicted_Score: 7.687139 (strongest positive impact).
- Conflicts_Over_Social_Media: 2.445317 (significant positive effect).
- Avg_Daily_Usage_Hours: -2.198558 (notable negative influence).
- Sleep_Hours_Per_Night: -0.529452 (moderate negative effect).
- Mental_Health_Score: -0.381912 (slight negative contribution).

These features drive the separation, with higher addiction and conflicts increasing the likelihood of academic impact, while more sleep and better mental health reduce it.

Interpretation: LDA effectively distinguishes students affected by social media, with an accuracy of 98.9%. The model highlights Addicted_Score and Conflicts_Over_Social_Media as key drivers of academic impact, while Avg_Daily_Usage_Hours, Sleep_Hours_Per_Night, and Mental_Health_Score also play roles. The clear score distribution and high metrics support the study's goal of identifying factors influencing academic performance.

5. Conclusion and Recommendations

This study explored the impact of social media on the academic performance of 705 students using a multifaceted approach, including exploratory data analysis (EDA), Principal Component Analysis (PCA), Factor Analysis (FA), and Linear Discriminant Analysis (LDA). Conducted as of 05:16 PM +0530 on Wednesday, June 04, 2025, the analysis provided a comprehensive understanding of the dataset, which included 12 variables such as Avg_Daily_Usage_Hours, Mental_Health_Score, and Addicted_Score.

Conclusion: EDA revealed significant patterns: students with higher Avg_Daily_Usage_Hours (mean 4.92 hours) and Addicted_Score (mean 6.44) showed lower Mental_Health_Score (mean 6.23) and Sleep_Hours_Per_Night (mean 6.87), with 453 students reporting academic impact. PCA identified four components explaining 77.53% of the variance, with PC1 (52.47%) linking Addicted_Score (0.457), Conflicts_Over_Social_Media (0.441), and Avg_Daily_Usage_Hours (0.425) to poor mental health (-0.443) and sleep (-0.395), underscoring addiction's role. FA confirmed three latent factors—addiction and academic strain (eigenvalue 5.50), relationship status (eigenvalue 2.20), and demographics (eigenvalue 1.30)—explaining ~75% of the variance, reinforcing addiction's dominance. LDA achieved 98.9% accuracy, with Addicted_Score (7.69) and Conflicts_Over_Social_Media (2.45) as top predictors, alongside Avg_Daily_Usage_Hours (-2.20), Sleep_Hours_Per_Night (-0.53), and Mental_Health_Score (-0.38), effectively separating affected (453) from unaffected (252) students. Together, these findings highlight that excessive social media use, particularly when linked to addiction and conflicts, significantly impairs academic performance by affecting mental health and sleep.

Recommendations:

1. **Educate and Monitor**: Schools should raise awareness about social media addiction and monitor usage to protect mental health and sleep.
2. **Support Programs**: Implement counseling and time-management workshops for students with high addiction scores to boost academic outcomes.
3. **Policy Action**: Establish guidelines limiting social media during study periods to enhance focus and well-being.
4. **Research Expansion**: Conduct longitudinal studies to confirm causality and test intervention effectiveness across diverse demographics.

This study offers valuable insights and practical strategies to mitigate social media's negative academic impact, benefiting students, educators, and policymakers.

5.References

- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications. (*Relevance*: Covers PCA, FA, and LDA, aligning with your analytical methods.)
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2020). *Multivariate data analysis* (8th ed.). Cengage Learning.
  (*Relevance*: Provides a detailed foundation for your multivariate techniques.)
- Primack, B. A., Swanier, B., Georgiopoulos, A. M., Land, S. R., & Fine, M. J. (2009). Association between media use in adolescence and depression in young adulthood. *Archives of General Psychiatry, 66*(2), 181-188. https://doi.org/10.1001/archgenpsychiatry.2008.532
  (*Relevance*: Links social media use to mental health, supporting your Mental_Health_Score findings.)

6.Appendices

dataset link: Students Social Media Addiction Dataset github link: All the codes are here

Libraries

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.preprocessing import LabelEncoder,StandardScaler
from sklearn.decomposition import PCA
from factor_analyzer.factor_analyzer import calculate_kmo, calculate_bartlett_sphericity, FactorAnalyzer
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
```

Preprocessing Part

```
data = pd.read_csv("C:\\Users\\Dell\\Downloads\\Students Social Media Addiction.csv")
```

```
data.head(3)
```

	Student_ID	Age	Gender	Academic_Level	Country	Avg_Daily_Usage_Hours	Most_Used_Platform	Affects_Academic_Performance	Sleep_Hours_Per_Night	Mental_
0	1	19	Female	Undergraduate	Bangladesh	5.2	Instagram	Yes	6.5	
1	2	22	Male	Graduate	India	2.1	Twitter	No	7.5	
2	3	20	Female	Undergraduate	USA	6.0	TikTok	Yes	5.0	

```
data = data.drop('Student_ID', axis=1)
```

```
categorical_cols = ['Gender', 'Academic_Level', 'Country', 'Most_Used_Platform',
                    'Affects_Academic_Performance', 'Relationship_Status']
for col in categorical_cols:
    print(f"\nFrequency Counts for {col}:")
    print(data[col].value_counts())
```

```
# Group countries into regions to reduce dimensionality
region_mapping = {
    'Bangladesh': 'South Asia', 'India': 'South Asia', 'Pakistan': 'South Asia', 'Nepal': 'South Asia',
    'Sri Lanka': 'South Asia', 'Maldives': 'South Asia', 'Bhutan': 'South Asia', 'Afghanistan': 'South Asia',
    'USA': 'North America', 'Canada': 'North America', 'Mexico': 'North America',
```

```
# Encode categorical variables
le = LabelEncoder()
data_encoded = data.copy()
data_encoded['Gender'] = le.fit_transform(data['Gender'])  # 0: Female, 1: Male
data_encoded['Affects_Academic_Performance'] = le.fit_transform(data['Affects_Academic_Performance'])  # 0: No, 1: Yes
```

```
# One-hot encode with binary (0/1) output
data_encoded = pd.get_dummies(data_encoded, columns=['Academic_Level', 'Region', 'Most_Used_Platform', 'Relationship_Status'],
                              drop_first=True, dtype=int)
```

```
data_encoded.drop('Country',axis=1,inplace = True)
```

Principal Component Analysis

```
# 1. Principal Component Analysis (PCA)
pca = PCA(n_components=len(data_combined.columns))  # Use all possible components
pca_result = pca.fit_transform(data_combined)
explained_variance = pca.explained_variance_ratio_
cumulative_variance = np.cumsum(explained_variance)

significant_pcs = np.where(explained_variance > 0.04 )[0]
significant_pc_labels = [f'PC{i+1}' for i in significant_pcs]

# Printed outputs
print("\nPCA Explained Variance Ratio (First 15 PCs):")
print(explained_variance[:15])
print("\nCumulative Explained Variance (First 15 PCs):")
print(cumulative_variance[:15])

# Variable contributions (loadings) for significant PCs
loadings = pd.DataFrame(pca.components_.T, columns=[f'PC{i+1}' for i in range(len(data_combined.columns))],
                        index=data_combined.columns)
loadings_significant = loadings[significant_pc_labels]
print("\nPCA Variable Contributions (Loadings) for Significant PCs:")
print(loadings_significant)
loadings_significant.to_csv('pca_loadings_significant.csv')

# Scree Plot with bars, normal line, and cumulative line for first 15 PCs
fig, ax1 = plt.subplots(figsize=(10, 6))

# Bars and normal line for explained variance
ax1.bar(range(1, 16), explained_variance[:15], color='#1f77b4', edgecolor='black', alpha=0.7, label='Variance Explained (Bars)')
ax1.plot(range(1, 16), explained_variance[:15], marker='o', linestyle='-', color='#2ca02c', label='Variance Explained (Line)')
ax1.set_xlabel('Principal Component')
ax1.set_ylabel('Proportion of Variance Explained', color='#1f77b4')
ax1.tick_params(axis='y', labelcolor='#1f77b4')
ax1.grid(True, axis='y')
ax1.set_xticks(range(1, 16))

# Cumulative line for cumulative variance
ax2 = ax1.twinx()
ax2.plot(range(1, 16), cumulative_variance[:15], marker='s', linestyle='--', color='#ff7f0e', label='Cumulative Variance')
ax2.set_ylabel('Cumulative Variance Explained', color='#ff7f0e')
ax2.tick_params(axis='y', labelcolor='#ff7f0e')
ax1.axvline(x=4, color='red', linestyle='--', label='Level after PC4')
ax1.annotate('Level after PC4', xy=(4, explained_variance[3]), xytext=(4, 0.1),
             arrowprops=dict(facecolor='red', shrink=0.05), color='red')

# Title and Legend
plt.title('Scree Plot of PCA (First 15 PCs)')
fig.legend(loc='upper right', bbox_to_anchor=(0.9, 0.9))
plt.savefig('scree_plot.png')
plt.show()
```


Factor Analysis

```
# 1. Initial KMO Check
kmo_all, kmo_model = calculate_kmo(data_combined)
kmo_series = pd.Series(kmo_all, index=data_combined.columns)
print(f"\nInitial KMO Measure of Sampling Adequacy (Overall): {kmo_model:.3f}")

# 2. Remove variables with KMO < 0.5, focusing on region/platform dummies
# Identify region and platform dummy variables
region_cols = [col for col in data_combined.columns if col.startswith('Region_')]
platform_cols = [col for col in data_combined.columns if col.startswith('Most_Used_Platform_')]
low_kmo_vars = kmo_series[kmo_series < 0.5].index
low_kmo_region_platform = [col for col in low_kmo_vars if col in region_cols + platform_cols]

# Remove low KMO variables
data_filtered = data_combined.drop(columns=low_kmo_vars)

print(f"\nRemoved variables with KMO < 0.5")
print(f"Removed dummies(region/platform)")

# 3. Recalculate KMO on filtered data
kmo_all_filtered, kmo_model_filtered = calculate_kmo(data_filtered)
kmo_series_filtered = pd.Series(kmo_all_filtered, index=data_filtered.columns)
print(f"\nRecalculated KMO Measure of Sampling Adequacy (Overall): {kmo_model_filtered:.3f}")
print("Recalculated KMO per variable")
```

```python
# 4. Proceed with Factor Analysis if KMO > 0.6
if kmo_model_filtered > 0.6:
    print("\nProceeding with Factor Analysis (KMO > 0.6)")

    # Bartlett's Test of Sphericity on filtered data
    chi_square_value, p_value = calculate_bartlett_sphericity(data_filtered)
    print(f"\nBartlett's Test of Sphericity (Filtered Data):")
    print(f"Chi-Square Value: {chi_square_value:.2f}")
    print(f"P-Value: {p_value:.3e}")

    # Perform Factor Analysis
    fa = FactorAnalyzer(n_factors=3, rotation='varimax', method='principal')  # 3 factors as an example
    fa.fit(data_filtered)

    # Factor Loadings
    loadings = pd.DataFrame(fa.loadings_, index=data_filtered.columns, columns=[f'Factor {i+1}' for i in range(3)])
    print("\nFactor Loadings (after Varimax rotation):")
    print(loadings)
```

Discriminant Analysis

```python
# 1. Data Preparation (assuming you've already cleaned/encoded)
predictors = ['Avg_Daily_Usage_Hours', 'Sleep_Hours_Per_Night', 'Mental_Health_Score',
              'Conflicts_Over_Social_Media', 'Addicted_Score', 'Gender',
              'Academic_Level_Undergraduate', 'Academic_Level_High School',
              'Most_Used_Platform_Instagram', 'Most_Used_Platform_TikTok']
X = data_combined[predictors]
y = data_encoded['Affects_Academic_Performance']

# Standardize predictors
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 2. LDA Model
lda = LinearDiscriminantAnalysis(n_components=1)
lda.fit(X_scaled, y)
y_pred = lda.predict(X_scaled)
y_scores = lda.transform(X_scaled).flatten()

# 3. Evaluation Metrics
print("Classification Report:")
print(classification_report(y, y_pred))

# 4. Visualizations
plt.figure(figsize=(15, 10))

# A. Confusion Matrix Heatmap
plt.subplot(2, 2, 1)
conf_mat = confusion_matrix(y, y_pred)
sns.heatmap(conf_mat, annot=True, fmt='d', cmap='Blues',
            xticklabels=['No Effect', 'Affects'],
            yticklabels=['No Effect', 'Affects'])
plt.title('Confusion Matrix')

plt.ylabel('Actual')
plt.xlabel('Predicted')

# B. LDA Score Distribution by Class
plt.subplot(2, 2, 2)
sns.kdeplot(y_scores[y == 0], label='No Effect', shade=True)
sns.kdeplot(y_scores[y == 1], label='Affects', shade=True)
plt.axvline(x=0, color='r', linestyle='--')  # Decision boundary
plt.title('LDA Score Distribution by Class')
plt.xlabel('Discriminant Score')
plt.ylabel('Density')
plt.legend()

# C. Coefficient Magnitude Plot
plt.subplot(2, 1, 2)
coef_df = pd.DataFrame(lda.coef_.T, index=predictors, columns=['Coefficient'])
coef_df['abs'] = coef_df['Coefficient'].abs()
coef_df = coef_df.sort_values('abs', ascending=False)
sns.barplot(x='Coefficient', y=coef_df.index, data=coef_df)
plt.title('LDA Feature Coefficients (Standardized)')
plt.xlabel('Coefficient Magnitude')
plt.ylabel('Features')
plt.tight_layout()

plt.savefig('lda_analysis.png', dpi=300)
plt.show()

# 5. Output Important Metrics
print(f"\nLDA Accuracy: {accuracy_score(y, y_pred):.3f}")
print("\nTop Discriminative Features:")
print(coef_df.sort_values('abs', ascending=False).drop('abs', axis=1).head(5))
```