Social Media Usage and Academic Factors

by
Audwit Nafi Anam

Analysis Objective

Draw Insights on impact of social media usage on students' academic life and social interactions.

Dataset Summary

- **1019 Samples, 8 Columns**
  - **Age (numerical)**
  - **Gender (categorical: Male, Female, Other)**
  - **Major (categorical: Computer Science, Business, Engineering, etc.)**
  - **Hours Spent on Social Media per Day (numerical, in hours)**
  - **Primary Social Media Platform (categorical: Facebook, Instagram, Twitter, TikTok, etc.)**
  - **Frequency of Posting (numerical: 0 – Never, 1 – Rarely, 2 – Sometimes, 3 – Often, 4 – Daily)**
  - **Number of Friends (numerical)**
  - **Academic Performance: (categorical: Good, Average, Poor).**

Dataset Preprocessing

- Deleting Rows with Null Values (32) and Duplicates (13)
- One-hot Encoding Categorical Features
  - Gender, Major, Primary Platform,
- Label Encoding Ordinal Features
  - posting_frequency, academic performance
  - 20 Columns after Encoding
- Outlier Removal
  - Keeping Numerical Values within 1.5*IQR
  - Keeping Categorical Values having count greater than threshold
  - 969 Samples after removal
- Challenges
  - Duplicate values with slightly different names
  - Trailing/Leading Whitespaces
- Solutions
  - Used lambda exprsssions, regular expressions, etc.

Exploratory Data Analysis (EDA)

- Majority primarily use Facebook, Instagram, X and LinkedIn in order
- Majority (44.6%) rarely post on social media
- Majority studied in CSE, then EEE, BBA and Arts in order
- Most respondents have follower count from 0 to 2000
- Majority have average academic performance followed by good
- Roughly, less time on social media corresponds to better performance. No correlation with follower count
- Poor performers post from rarely to often, rest post rarely to sometimes
- TikTok and Snapchat have the highest hours
- FB, X and LinkedIn have highest follower count

Model Development and Evaluation

- DBSCAN (eps=15, min_pts = 3)
  - n_clusters: 11, silhouette score: 46.98%, davies_bouldin_score: 3.13
- KMeans
  - Elbow method reveals suitable n_clusters to be around 2, 3, 4
  - n_clusters = 2, silhouette score: 60.57%, davies_bouldin_score: 0.5441
  - n_clusters = 3, silhouette score: 56.43%, davies_bouldin_score: 0.5390
  - n_clusters = 4, silhouette score: 55.96%, davies_bouldin_score: 0.4993

Key Insights and Recommendations

Insights

- DBSCAN performs worse due to irregular density and high dimensional data
- Kmeans outperform DBSCAN due to fixing number of clusters and large dataset
- Kmeans perform best with n_clusters=3
- KMeans have scores better than DBSCAN indicating better clustering

Recommendations

- Gather more data
- Balanced data
- Modularize, Containerize for scalable deployment

Thank You