Mini Project: Segmenting Students Based on Social Network Profiles

1. Project Objective

The objective of this project is to group high-school students into meaningful clusters based on the words and topics mentioned on their social network profiles. These keywords reflect students' interests, personality traits, and social behaviors. The project demonstrates how unsupervised learning can uncover hidden patterns without predefined labels.

2. Dataset Description

The dataset consists of student demographic information and keyword counts extracted from social network profiles.

Key attributes include:
- gradyear: Expected graduation year
- gender: Gender of the student
- age: Age at the time of data collection
- NumberOfFriends: Number of social connections
- Sports-related keywords (basketball, football, soccer, etc.)
- Appearance-related keywords (cute, sexy, clothes, etc.)
- Social activities (music, shopping, dance, mall)
- Religion-related keywords (god, church, jesus, bible)
- Risk-related keywords (drunk, drugs, death)

3. Data Preprocessing

The dataset was preprocessed before clustering. Missing values in age and gender were handled, categorical variables were encoded numerically, and remaining missing values were replaced with zero. Feature scaling was applied using StandardScaler to ensure all variables contributed equally to the clustering process.

4. Methodology

K-Means clustering was used to segment students. Since clustering is sensitive to scale, all features were standardized. The optimal number of clusters was determined using the Silhouette Score method, which measures how well data points fit within their assigned cluster.

5. Model Implementation

The K-Means algorithm was trained using the optimal number of clusters. Each student was assigned a cluster label based on similarities in keyword usage and profile characteristics.

6. Results and Cluster Interpretation

The clustering results revealed distinct behavioral groups:
- Sports-oriented students

- Appearance and fashion-focused students
- Music and social activity enthusiasts
- Religiously inclined students
- Risk-prone behavior group

These clusters highlight how students naturally group based on online expressions.

7. Conclusion

This project demonstrates the effectiveness of unsupervised learning in discovering hidden patterns in social network data. The insights gained from clustering can be used for targeted interventions, behavioral analysis, and understanding student interests at scale.