Project Brief: Clustering High School Students Based on Social Media Interests for Edulytics.

Dataset: https://www.kaggle.com/datasets/zabihullah18/students-social-network-profile-clustering

File to Use: The main dataset file.

Background & Business Context

Edulytics is a forward-thinking ed-tech analytics startup aiming to revolutionize how educational institutions understand and support their student populations. One of their primary goals is to help schools and policymakers make informed decisions by identifying underlying behavioral patterns and interest groups among students.

To achieve this, Edulytics has obtained a unique dataset comprising 15,000 high school students' social media profiles collected from 2006–2009. These profiles include demographic details and the frequency of certain keywords related to hobbies, lifestyle, social behaviors, and cultural interests.

Understanding clusters of students with similar interests can help:

- Tailor academic and extracurricular programs
- Develop targeted counseling and support services
- Predict potential behavioral trends
- Foster inclusive and diverse school environments

As a data science intern at Edulytics, your task is to develop an **unsupervised learning model** that can identify distinct **student interest-based clusters**. Your

model will be the first step in building a behavioral analytics tool to help educators and administrators uncover hidden insights from online student behavior.

Objective

Design and evaluate an unsupervised clustering model that groups students into distinct clusters based on:

- Online interest indicators (e.g., sports, music, religion, fashion, substance-related terms, etc.)
- Basic demographics (age, gender, grad year, number of friends)

The ultimate goal is to provide a clustering solution that is:

- **Interpretable**: Understand what defines each group
- **Actionable**: Inform decisions related to student engagement and support
- **Valid**: Grounded in real, distinguishable patterns within the data

---

Dataset Overview

Download:
https://www.kaggle.com/datasets/zabihullah18/students-social-network-profile-clustering

Structure:
The dataset contains 41 columns including:

- **Demographics**: gradyear, gender, age, NumberOffriends

- **Interest Terms** (37 columns): Frequency counts of mentions such as football, shopping, drugs, music, jesus, rock, death, etc.

All features are numerical (either counts or categorical encoded), making the dataset relatively clean and ready for vector-based clustering techniques.

Deliverables

1. Exploratory Data Analysis (EDA)

- Summary statistics and distributions for demographics and interest frequencies
- Identify high-frequency and low-frequency interests
- Analyze correlation or co-occurrence among interest terms
- Visualize potential groupings or anomalies using PCA or t-SNE

2. Data Preprocessing

- Handle outliers or skewed distributions
- Normalize/scale features to ensure fair clustering
- Encode categorical variables (e.g., gender if not already binary)
- Consider dimensionality reduction for better clustering visualization

3. Clustering Model Development

- Apply and compare at least 3 clustering techniques (e.g., KMeans, DBSCAN, Agglomerative Clustering)

- Determine the optimal number of clusters using Elbow Method, Silhouette Score etc.
- Use dimensionality reduction (PCA) to visualize clusters in 3D space

4. Evaluation Metrics

- Internal validation: Silhouette Score, Inertia etc.
- Visualizations of cluster separability
- Profile interpretation of each cluster:
    - What interests define each group?
    - What demographic characteristics are common within each group?

5. Cluster Interpretation

- Provide a detailed narrative for each cluster (e.g., "Music Lovers", "Religious-Oriented", "Risk-Takers")
- Highlight surprising or counterintuitive findings if any
- Suggest how a school or educational app might use these groupings to create targeted content, support systems, or policy recommendations

---

Final Submission Requirements

Each team or individual must submit:

- A **GitHub Gist** containing:
    - Clean, modular Python code (preferably in Jupyter Notebook format)
    - Clear sections for EDA, preprocessing, clustering, and evaluation

    - Code should be well-commented and reproducible using Colab
    - Save a copy of the notebook as a GitHub Gist and submit the generated link
- A **Medium Article** summarizing:
    - Your approach and major decisions
    - Key visualizations and insights
    - Final clustering results and business implications
    - Suggestions for how schools or youth programs can apply these findings

Links to the Gist and Medium article should be submitted via the portal.

Key Skills To Be Evaluated

- Unsupervised learning and clustering
- Dimensionality reduction and data visualization
- Feature engineering and normalization
- Interpretability and business reasoning
- Communication of technical insights to non-technical stakeholders

---

GOOD LUCK. Make your analysis meaningful, ethical, and insightful.