**PROJECT REPORT**

**Machine Learning**

**CSL313**



Faculty name: Ankita Gupta Student name: Shivam , Gaurav, Garvit

Roll No.: 21csu325 , 21csu316, 21csu322

Section: DS-VI-B

Submission Date: 20-11-2023

**Department of Computer Science and Engineering**

**NorthCap University, Gurugram- 122001, India**

**Session 2023-24**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CONTENTS** | | | | |
| S.N. | **Details** | Page No. | | |
| 1 | **Introduction** | 3 | | |
| 2 | **About The Project** | 4 | | |
| 3 | **Model Experiment** | 6 | |
| 4 | **Libraries Used In Project** | 12 | |
| 5 | **Dataset Overview** | | 13 | |
| 6 | **Screenshots** | | 14 | |
| 7 | **References** | 21 | |

Project Title:

**Enhancing Social Networking Profiles for Students: A Clustering Approach**

This project aims to enhance social networking profiles for students by applying clustering techniques to facilitate the discovery of like-minded individuals and foster deeper connections based on shared interests.connections based on shared interests.

About The Project:

**Problem Statement:**

Social networking platforms often struggle to effectively connect students with peers who share similar interests, hindering the formation of meaningful connections and limiting the overall user experience.

In today's digital age, social networking platforms play a pivotal role in connecting individuals, fostering communities, and facilitating meaningful interactions. Particularly within educational settings, where students seek to engage with peers who share similar interests and aspirations, optimizing social networking profiles can significantly enhance the overall user experience. This project aims to leverage clustering techniques to enhance social networking profiles for students, facilitating the discovery of like-minded individuals and fostering deeper connections based on shared interests.

In this report, we present an in-depth exploration of the project, including its objectives, methodology, results, and potential real-world applications. By applying clustering analysis to a dataset containing various attributes of high school students, such as age, gender, interests, and activities, we demonstrate how the project can provide valuable insights into student demographics, behaviors, and preferences. These insights can then be utilized to improve social networking platforms by offering personalized recommendations, interest-based matching, and community-building features.

Through this project, we aim to address the following objectives:

1. Segmentation and Personalization: Utilize clustering techniques to segment students based on their interests, activities, and demographics, enabling personalized experiences on social networking platforms.
2. Enhanced User Experience: Enhance the user experience of social networking profiles for students by providing tailored content, recommendations, and networking opportunities based on shared interests and preferences.
3. Community Building: Facilitate the formation of interest-based groups and communities within social networking platforms, fostering collaboration, discussion, and engagement among students with similar interests.
4. Insight Generation: Generate actionable insights into student demographics, behaviors, and preferences, empowering educational institutions, social researchers, and marketers to make data-driven decisions and interventions.

In the following sections, we delve into the methodology employed, the dataset used for analysis, the clustering techniques applied, and the implications of the project for social networking platforms in educational contexts. Additionally, we discuss potential real-world use cases, future directions, and the broader significance of leveraging clustering approaches to enhance social networking experiences for students.

Approach:

Our approach involves leveraging clustering techniques to analyze a dataset containing various attributes of high school students. By segmenting students based on their interests, activities, and demographics, we aim to enhance social networking profiles and facilitate the discovery of like-minded individuals. The steps followed in our approach are outlined below:

1. Data Collection and Preparation:
   1. Gathered a dataset containing attributes such as graduation year, gender, age, and interests of high school students.
   2. Preprocessed the data by handling missing values, converting categorical variables into numerical representations, and scaling numerical features.
2. Exploratory Data Analysis (EDA):
   1. Conducted exploratory data analysis to gain insights into the distribution and relationships among different variables in the dataset.
   2. Visualized the data using various techniques such as histograms, box plots, and correlation matrices to understand the underlying patterns.
3. Feature Selection:
   1. Selected relevant features for clustering, focusing on attributes related to students' interests, activities, and demographics.
   2. Identified the subset of features that are most likely to contribute to meaningful clustering results.
4. Clustering Analysis:
   1. Applied KMeans clustering algorithm to group students into distinct clusters based on their selected features.
   2. Experimented with different numbers of clusters and evaluated clustering performance using metrics such as silhouette score and elbow method.
5. Cluster Interpretation:
   1. Interpreted the characteristics of each cluster by analyzing the centroid values and distribution of features within each cluster.
   2. Identified common traits, interests, and demographics associated with each cluster to understand the student segments.
6. Application in Social Networking Profiles:
   1. Translated clustering results into actionable insights for enhancing social networking profiles for students.
   2. Developed strategies for personalized recommendations, interest-based matching, and community-building features based on the identified clusters.
7. Validation and Evaluation:
   1. Validated the clustering results by assessing their consistency and coherence with domain knowledge and real-world observations.
   2. Evaluated the effectiveness of the proposed approach in improving user experience and engagement on social networking platforms.

Machine learning model

**k-Nearest Neighbors (KNN):**

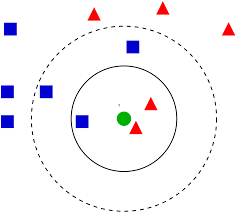
**Type:** Supervised Learning (Classification or Regression)**Use Case**:KNN is a versatile algorithm used for both classification and regression tasks. It is particularly effective when the decision boundary is complex or nonlinear.

**Working Principle:** Given a new data point, KNN classifies or predicts its label based on the labels of its k-nearest neighbors in the feature space.For classification, the majority class among the neighbors is assigned to the new data point. For regression, the average of the neighbors' values is used.

**Hyperparameter**:The crucial hyperparameter is 'k,' the number of neighbors to consider. The choice of 'k' affects the model's sensitivity to noise and its ability to capture patterns.

**Distance Metric**:The choice of distance metric (Euclidean, Manhattan, etc.) influences how the algorithm measures the proximity between data points.

**Decision Boundary:**The decision boundary is non-linear and can adapt to the shape of the data.



**WCSS:**

WCSS (Within-Cluster Sum of Squares):

WCSS is a metric used to evaluate the performance of a clustering algorithm, such as KMeans. It measures the compactness of the clusters formed by the algorithm. Specifically, WCSS calculates the sum of the squared distances between each data point and its nearest cluster centroid.

The goal of KMeans clustering is to minimize WCSS, as smaller WCSS values indicate that data points are closer to their respective cluster centroids, indicating tighter and more cohesive clusters.

**Elbow Method:**

The Elbow Method is a heuristic technique used to determine the optimal number of clusters (K) in a dataset for KMeans clustering. It involves plotting the WCSS values for different values of K and identifying the point where the decrease in WCSS begins to level off, resembling an "elbow" shape in the plot.

The rationale behind the Elbow Method is to select the value of K where adding more clusters does not significantly decrease the WCSS. This point represents the optimal balance between maximizing the cohesion of clusters while avoiding excessive fragmentation.

In practice, the choice of K is subjective, and the Elbow Method provides a visual aid to help select a reasonable value for K based on the inflection point in the WCSS plot. However, it is important to note that the Elbow Method is not always definitive, and other considerations such as domain knowledge and specific use cases should also be taken into account when determining the number of clusters.

A graph with a line and a number of clusters

Description automatically generated

**Libraries Used In Project:**

1. Pandas: A powerful data manipulation library for handling tabular data efficiently.

2. NumPy: Fundamental library for numerical computing, providing support for multi-dimensional arrays and mathematical functions.

3. Matplotlib: Widely-used plotting library for creating static and interactive visualizations in Python.

4. Seaborn: Statistical data visualization library built on top of Matplotlib, offering high-level functions for creating informative plots.

5. Scikit-learn: Comprehensive machine learning library providing tools for various tasks such as classification, regression, and clustering.

6. Warnings: Python module for issuing warning messages to alert users about potential issues or deprecated features.

**Dataset Overview:**

* **gradyear**: The graduation year of the high school student.
* **gender**: The gender of the student (e.g., male or female).
* **age**: The age of the student at the time of the survey.
* **NumberOffriends**: The number of contacts or friends the student had on the social network.
* **basketball**: The frequency or count of mentions of basketball in the student's profile.
* **football**: The frequency or count of mentions of football in the student's profile.
* **soccer**: The frequency or count of mentions of soccer in the student's profile.
* **softball**: The frequency or count of mentions of softball in the student's profile.
* **volleyball**: The frequency or count of mentions of volleyball in the student's profile.
* **swimming**: The frequency or count of mentions of swimming in the student's profile.
* **cheerleading**: The frequency or count of mentions of cheerleading in the student's profile.
* **baseball**: The frequency or count of mentions of baseball in the student's profile.
* **tennis**: The frequency or count of mentions of tennis in the student's profile.
* **sports**: The overall frequency or count of mentions of sports in the student's profile.
* **cute**: The frequency or count of mentions of cute in the student's profile.
* **sex**: The frequency or count of mentions of sex in the student's profile.
* **sexy**: The frequency or count of mentions of sexy in the student's profile.
* **hot**: The frequency or count of mentions of hot in the student's profile.
* **kissed**: The frequency or count of mentions of kissed in the student's profile.
* **dance**: The frequency or count of mentions of dance in the student's profile.
* **band**: The frequency or count of mentions of band in the student's profile.
* **marching**: The frequency or count of mentions of marching in the student's profile.
* **music**: The frequency or count of mentions of music in the student's profile.
* **rock**: The frequency or count of mentions of rock in the student's profile.
* **god**: The frequency or count of mentions of god in the student's profile.
* **church**: The frequency or count of mentions of church in the student's profile.
* **jesus**: The frequency or count of mentions of Jesus in the student's profile.
* **bible**: The frequency or count of mentions of the Bible in the student's profile.
* **hair**: The frequency or count of mentions of hair in the student's profile.
* **dress**: The frequency or count of mentions of dress in the student's profile.
* **blonde**: The frequency or count of mentions of blonde in the student's profile.
* **mall**: The frequency or count of mentions of mall in the student's profile.
* **shopping**: The frequency or count of mentions of shopping in the student's profile.
* **clothes**: The frequency or count of mentions of clothes in the student's profile.
* **hollister**: The frequency or count of mentions of Hollister (a brand) in the student's profile.
* **abercrombie**: The frequency or count of mentions of Abercrombie (a brand) in the student's profile.
* **die**: The frequency or count of mentions of die in the student's profile.
* **death**: The frequency or count of mentions of death in the student's profile.
* **drunk**: The frequency or count of mentions of drunk in the student's profile.
* **drugs**: The frequency or count of mentions of drugs in the student's profile.

Screenshots:

A screenshot of a computer

Description automatically generated

A screenshot of a sports statistics

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generatedA screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated



A group of colorful bars

Description automatically generated with medium confidence

**Observation and Conclusion:**

**Potential Market Use Case:**

The insights derived from this project present a compelling market use case for various industries, particularly those targeting teenage and young adult demographics. Social media platforms, in particular, stand to benefit significantly from implementing the findings to enhance user engagement and attract advertisers.

**Targeting Female-Centric Products:**

* + Social media platforms can leverage clusters with a higher percentage of females (Clusters 1, 2, and 3) to offer targeted advertising opportunities for products and services catering to teenage girls.
  + Brands focusing on fashion, beauty, socializing, or lifestyle can partner with these platforms to reach their desired audience effectively.
  + By understanding the interests and preferences of female-centric clusters, social media platforms can provide tailored content and advertisements that resonate with these demographics, thus increasing engagement and driving revenue from advertisers seeking to target this demographic.

**Understanding Male-Centric Interests:**

* + Platforms can also tap into clusters with a higher percentage of males (Clusters 0 and 4) to offer targeted advertising solutions for products and services tailored to teenage boys.
  + Brands in industries such as sports, technology, gaming, or adventure can collaborate with these platforms to reach male-centric clusters effectively.
  + By analyzing the interests and behaviors of male-centric clusters, social media platforms can deliver content and advertisements that align with the preferences of teenage boys, enhancing engagement and driving advertising revenue from brands targeting this demographic.

**Fashion and Retail Promotions:**

* + Clusters exhibiting a higher interest in shopping (Clusters 2 and 3) present an opportunity for social media platforms to partner with fashion and retail brands.
  + Platforms can offer targeted advertising campaigns and promotions related to fashion, retail, and shopping to these clusters, thereby increasing user engagement and driving revenue from advertisers in the fashion and retail sectors.

**Sports and Lifestyle Brand Partnerships:**

* + Platforms can collaborate with sports brands or organize sports-related events and activities to appeal to clusters showing a strong interest in sports (e.g., Clusters 3 and 4).
  + By aligning with the interests and preferences of these clusters, social media platforms can offer targeted advertising opportunities and branded content partnerships with sports and lifestyle brands, driving engagement and revenue in the process.

In summary, the insights derived from this project offer valuable opportunities for social media platforms to enhance user engagement and attract advertisers by offering targeted advertising solutions and partnerships tailored to specific demographic segments. By leveraging the findings to deliver personalized content and advertisements, platforms can create a more compelling user experience while driving revenue growth through targeted advertising opportunities and brand partnerships.

**In conclusion, this project demonstrates the relevance of leveraging clustering techniques to enhance social networking profiles for students and provide valuable insights for marketing strategies. By segmenting students based on their interests, demographics, and behaviors, we can tailor content and advertisements to specific clusters, thus improving user engagement and satisfaction on social networking platforms. The findings from this analysis offer actionable recommendations for marketing departments, including targeting female-centric products, understanding male-centric interests, promoting fashion and retail offerings, and forging partnerships with sports and lifestyle brands. Overall, this project highlights the potential of data-driven approaches to optimize user experiences and drive business outcomes in the digital landscape.**

**References:**

**Kaggle Dataset** : Student clustering

**Google** : <https://www.google.com/>

**Towards Data Science** : <https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset>

**GeeksforGeeks (To learn about ML Algorithm Models) :** <https://www.geeksforgeeks.org>