Machine Learning Project Report

Title

OTT Audience Map for SNU Film Fest: Clustering Students into Viewing Preference Groups

Author

Prasenjit Sasmal
Year 3, Section 09
B.Tech CSE (AI-ML)

Abstract

The Cultural Committee at Student National University (SNU) is organizing the annual Film Fest. A key challenge lies in understanding the diverse viewing preferences of students to plan film screenings and secure relevant OTT partnerships. This project applies **unsupervised machine learning** to cluster students based on their favorite movie genres, series genres, OTT platforms, and content languages. Using **K-Means (evaluated with inertia and silhouette score)** four distinct audience segments were identified. While silhouette scores indicate moderate overlap between clusters, the results provide actionable insights into **screening curation, OTT tie-ups, and targeted student engagement strategies** for the Film Fest.

1. Introduction

The success of cultural events like film festivals depends heavily on aligning programs with student interests. At SNU, the audience base is highly diverse, with preferences ranging from Bollywood to global cinema, Hindi OTT platforms to international streaming giants.

Problem Statement:

- Current planning lacks **data-driven insights** into student preferences.
- Students' favorite genres, languages, and OTT platforms are not systematically analyzed.
- Without segmentation, **screenings and tie-ups may miss key audience groups**.

Research Question:

Can clustering students based on their OTT preferences reveal **meaningful audience groups** to guide event planning?

2. Dataset and Preprocessing

Dataset Features

- **movie_genre_top1** → Student's top movie genre 🎬
- **series_genre_top1** → Student's top web-series genre 📺
- **ott_top1** → Student's preferred OTT platform 📲
- **content_lang_top1** → Student's favorite content language 🌏

Data Cleaning

- Standardized column names (removed extra spaces, stripped symbols).
- Normalized text entries (lowercased, trimmed whitespace).
- Dropped rows with missing values.

Feature Engineering

- Encoded categorical features (Label Encoding).
- Constructed feature matrix from the four columns.
- Scaled values for clustering.

---

3. Methodology

Clustering Approach

- **Algorithms applied:**
  - K-Means (partitioning-based).
  - Agglomerative Clustering (hierarchical).
- **Cluster range tested:** k = 2 to 7.
- **Evaluations:**
  - **Inertia (Elbow Method)** → measured compactness of clusters.
  - **Silhouette Score** → measured separation and cohesion.

Evaluation Metrics

- **Inertia:** Decreased steadily with higher k; the “elbow” appeared at k = 4.
- **Silhouette Score:** Moderate (~0.32) with best performance at k = 4.
- **Visualization:** 3D/4D scatter plots for genre–OTT–language–series preferences.

---

4. Results

Inertia and Silhouette Analysis

- Inertia curve (Elbow Method) indicated **k = 4** as the optimal cluster size.
- Silhouette analysis confirmed moderate cluster quality at **k = 4**.

Figure 1: Inertia vs. Number of Clusters (Elbow at k = 4)

Figure 2: Silhouette Score vs. Number of Clusters

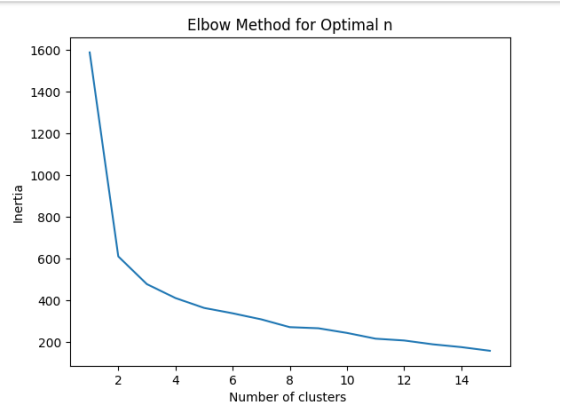

Figure:1

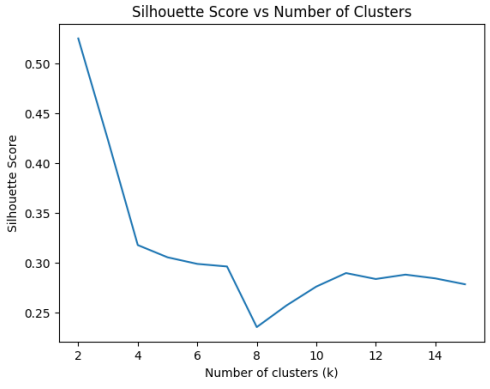

Figure:2

Optimal k = 4

Cluster Profiles

Cluster	Dominant Traits	Interpretation
0	Bollywood movies + Hindi OTT	Bollywood Buffs – Prefer Hindi films and mainstream OTT like Hotstar.
1	English-language content + Netflix/Prime	Global Bingers – Strong preference for international content and platforms.
2	Regional languages + regional OTT	Regional Story Seekers – Rooted in regional culture, favor local OTTs.
3	Diverse genres + web-series focus	Series Enthusiasts – Interested in long-form series across platforms.

Figure 3: Audience Map (2D visualization of 4 clusters)
Figure 4: 3D Cluster Visualization of Student OTT Preferences

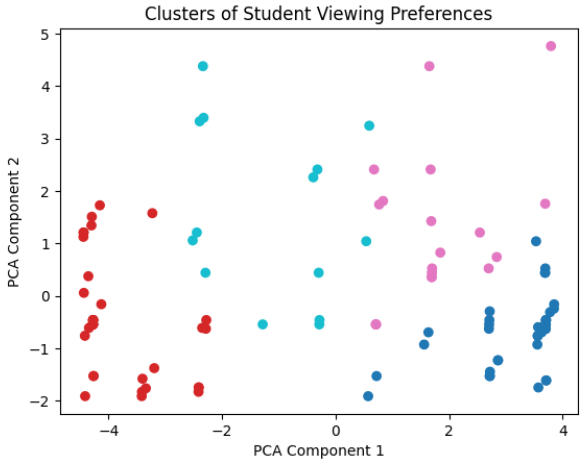

Figure:3

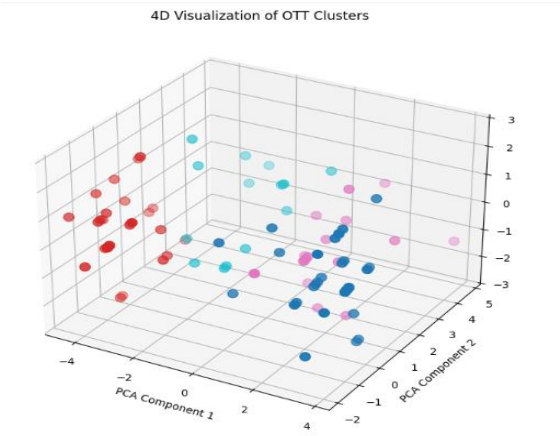

Figure:4

---

5. Discussion

- **Inertia** provided a strong signal for k = 4 (clear elbow).
- **Silhouette scores** were moderate, reflecting natural overlaps in student preferences.
- **Actionable Insights:**
  - Screen a mix of **Bollywood, Regional, and International films**.

- Partner with both **global (Netflix, Prime)** and **regional OTT platforms**.
- Curate **theme-based festival nights**: Bollywood Night, Global Night, Regional Night, Series Marathon.

---

6. Conclusion and Future Work

This project demonstrated the usefulness of clustering in audience analysis for cultural events:

- Identified **4 distinct audience groups** at SNU.
- Used **both inertia and silhouette analysis** to validate cluster quality.
- Showed how clustering supports **screening and OTT tie-up strategies**.

Future Directions

- Collect more features (binge hours, favorite actors, genres ranked).
- Experiment with **DBSCAN, Spectral Clustering** for better separation.
- Apply **association rule mining** to discover hidden content co-preferences.