CMR ENGINEERING COLLEGE

(UGC AUTONOMOUS)

(Accredited by NBA,Approved by AICTE NEW DELHI, Affiliated to JNTU, Hyderabad)

Kandlakoya, Medchal Road, Hyderabad-501 401.

Department of Computer Science & Engineering

A Mini Project

on

BEST STREAMING SHOWS SEGMENTATION ANALYSIS

Internal Guide:-

Mr. MD. Azhar,
Assistant Professor,
Dept. of CSE.

Presented By

B. POOJITHA : (218R1A05K4)

D. PRASHANTH : (218R1A05L2)

M. ISHWARYA : (218R1A05N0)

SAI KIRAN : (218R1A05P0)

A.Y 2024-2025

Table Of Contents

Abstract

This abstract delves into the necessity of data analysis on OTT TV shows, highlighting the pivotal role it plays in unraveling viewer patterns and shaping the future of content delivery. The diverse preferences of viewers, influenced by factors such as age, genre affinity, and platform accessibility, necessitate a nuanced understanding for platform providers, content creators, and advertisers alike. In this project, we conducted an in-depth data analysis on an OTT TV shows dataset, leveraging the K-means clustering algorithm to unveil underlying patterns in viewer preferences. The dataset encompassed key features such as age suitability, IMDb ratings, Rotten Tomatoes scores, show titles, and the availability on popular streaming platforms including Netflix, Hulu, Prime Video, Disney Plus, and Hot star.

Introduction

- This project embarks on a comprehensive exploration of viewer behavior within the OTT domain, employing data analysis techniques, specifically the K-means clustering algorithm, to unravel intricate patterns within a rich dataset.
- The diverse preferences of viewers, influenced by factors such as age, genre affinity, and platform accessibility. The dataset encompassed key features such as age suitability, IMDb ratings, Rotten Tomatoes scores.

Literature Survey

Author Name	Title	Journal/ conference	Year	Objective	Methodology / Technology	Key Findings
Dr. K Vengatesan	An Efficient Machine Learning System using Sentiment Analysis for Movie Recommendations	IEEE Transaction on Affective Computing	2022	To improve movie recommendation accuracy by incorporating sentiment analysis from user reviews	Uses natural language processing (NLP) techniques for sentiment analysis and collaborative filtering algorithms for recommendations	Incorporating sentiment analysis significantly boosts recommendation relevance and user engagement.
Md. Balfaqih	.An Intelligent Movies Recommendation System Based Facial Attributes Using Machine Learning	IEEE Transaction on Multimedia	2023	To enhance movie recommendations by analyzing facial attributes and expressions of users	Utilizes facial recognition technology and machine learning models (e.g., convolutional neural networks) to analyze user expressions and preferences	Facial recognition provides an additional layer of personalization, enhancing recommendation precision.

Author Name	Title	Journal/ conference	Year	Objective	Methodology / Technology	Key Findings
Zulfiqaur Ali	Enhancing Performance of Movie Recommendations using LSTM with Meta Path Analysis	IEEE Transactions on Neural Networks and Learning Systems	2023	To enhance the performance of movie recommendations using Long Short-Term Memory (LSTM) networks and meta path analysis	Combines LSTM networks with meta path analysis using graph-based techniques to capture complex user-item interactions over time	LSTM with meta path analysis captures temporal dynamics effectively, improving recommendation accuracy.
Yogesh Kumar	Movie Popularity and Target Audience Prediction using Content-Based Recommendation System	IEEE Transactions on Knowledge and Data Engineering	2022	To predict movie popularity and target audience using content-based recommendation techniques	Uses machine learning models, such as decision trees and support vector machines, to analyze movie content features and predict popularity and target audience	Content-based features are critical in predicting both movie popularity and suitable target audiences

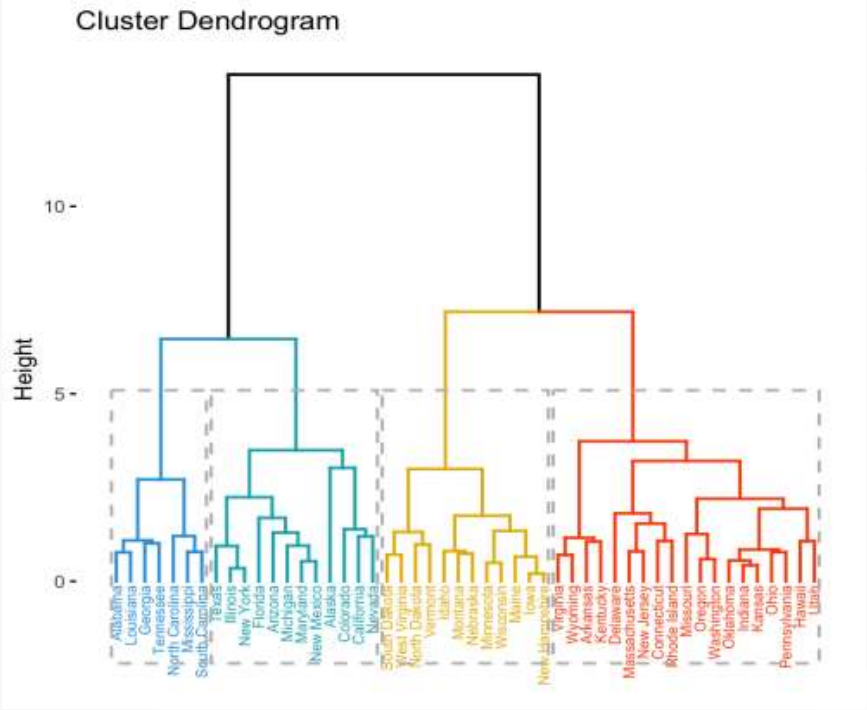
Existing System

Agglomerative Hierarchical Clustering Analysis

- Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

- Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

- In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.



Proposed Methodology

Project Scope

- The scope of the data analysis project on OTT TV shows, utilizing the K-means clustering algorithm, is to gain profound insights into viewer behavior and preferences within the ever-expanding landscape of digital content consumption.
- By categorizing TV shows into distinct clusters based on shared characteristics, the analysis aims to reveal nuanced insights into viewer preferences.

Objectives

- To categorize TV shows into distinct segments based on shared characteristics, enabling a nuanced understanding of viewer preferences.
- To contribute to a more personalized and satisfying viewer experience.
- To inform strategic decisions related to content acquisition, production, and platform differentiation.
- To gain insights that contribute to the competitive positioning of OTT platforms in the market.
- The project aims to extract meaningful patterns from a comprehensive dataset, encompassing key features such as age suitability, IMDb ratings, Rotten Tomatoes scores, show titles, and the presence on major streaming platforms including Netflix, Hulu, Prime Video, Disney Plus Hotstar.

Modules

Data Preprocessing

- Data Cleaning: Handle missing values, outliers, and inconsistencies.
- Feature Engineering: Create new features from raw data (e.g., genre tags, release year,cast).
- Data Transformation: Normalize or standardize data as necessary.

Segmentation Analysis

- Clustering: Apply clustering algorithms (e.g., K-means, DBSCAN) to segment shows based on various attributes (e.g., genre, viewership, ratings).
- Profile Segments: Analyze and describe each segment in terms of demographics,preferences, and behaviors.
- Visualization: Use visualizations to illustrate the segments and their characteristics.

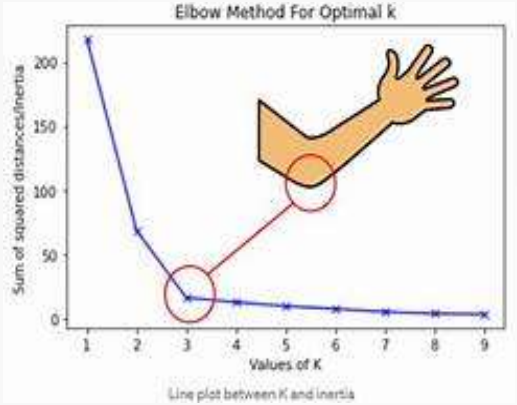
Algorithms

K-Means Clustering Algorithm

- K-Means Clustering is an unsupervised machine learning algorithm, which groups the unlabeled dataset into different clusters.
- It is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties .
- It is a centroid-based algorithm, where each cluster is associated with a centroid.
- The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
- The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms.

Elbow Method

- The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:
- WCSS= $\sum_{\text{Pi in Cluster1}}$ distance(Pi C1)^2 +$\sum_{\text{Pi in Cluster2}}$distance(Pi C2)^2+$\sum_{\text{Pi in CLuster3}}$ distance(Pi C3)^2



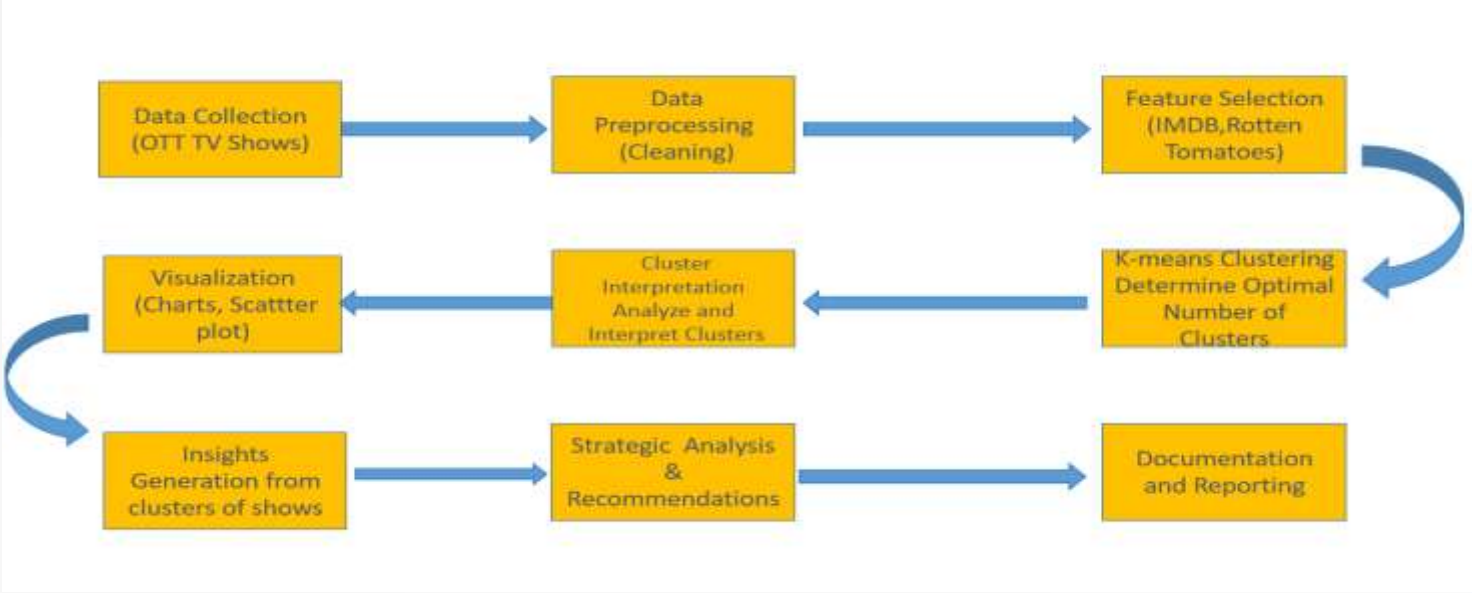
Software Requirements

Operating System	**Windows 7/8/10/11**
Development Software	Python 3.10
Programming Language	Python
Domain	Machine Learning
Integrated Development Environment (IDE)	Visual Studio Code/Jupyter Notebook
Libraries	Pandas , NumPy , Seaborne , Matplotlib

Hardware Requirements

	MY SYSTEM
System	A PC with Windows/Linux OS
Hard Disk	512 GB
Ram	Minimum of 8gb RAM.
Processor	Processor with 2.40GHz 2.50 GHz speed

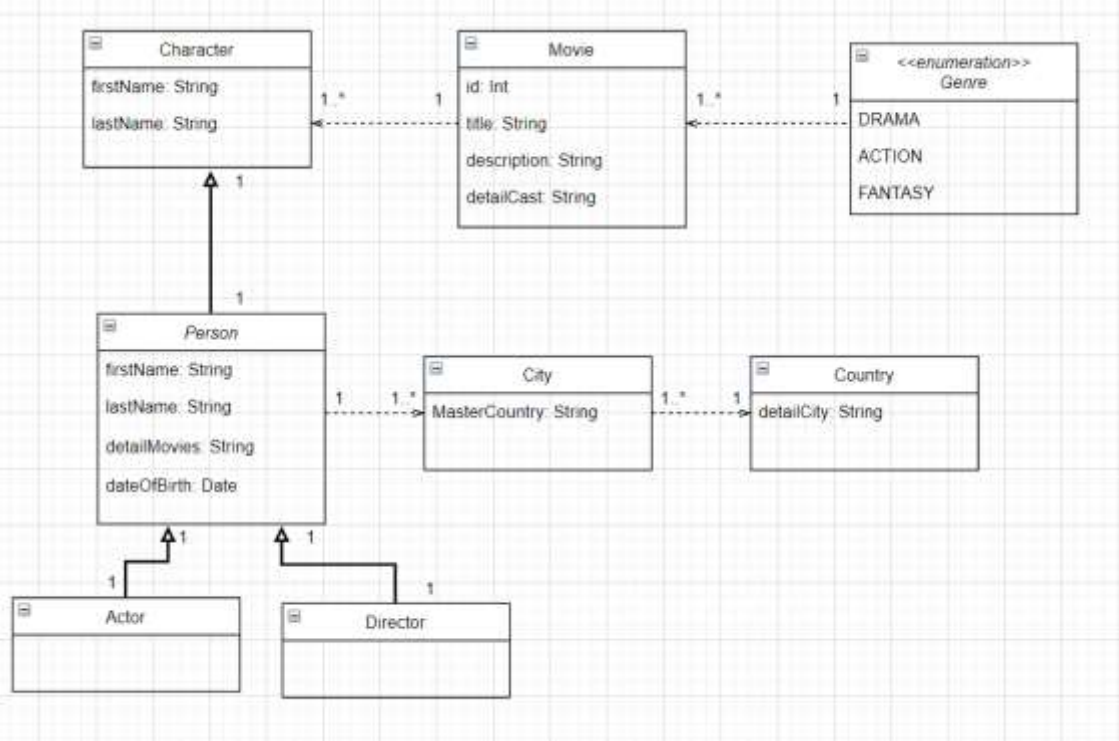
System Architecture



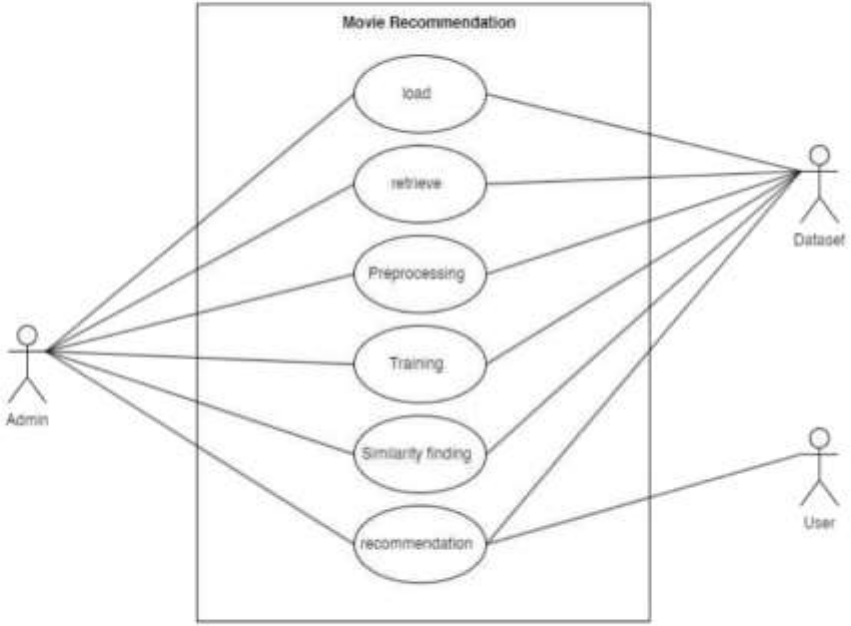
Architecture of Best Streaming Shows Segmentation

System Design

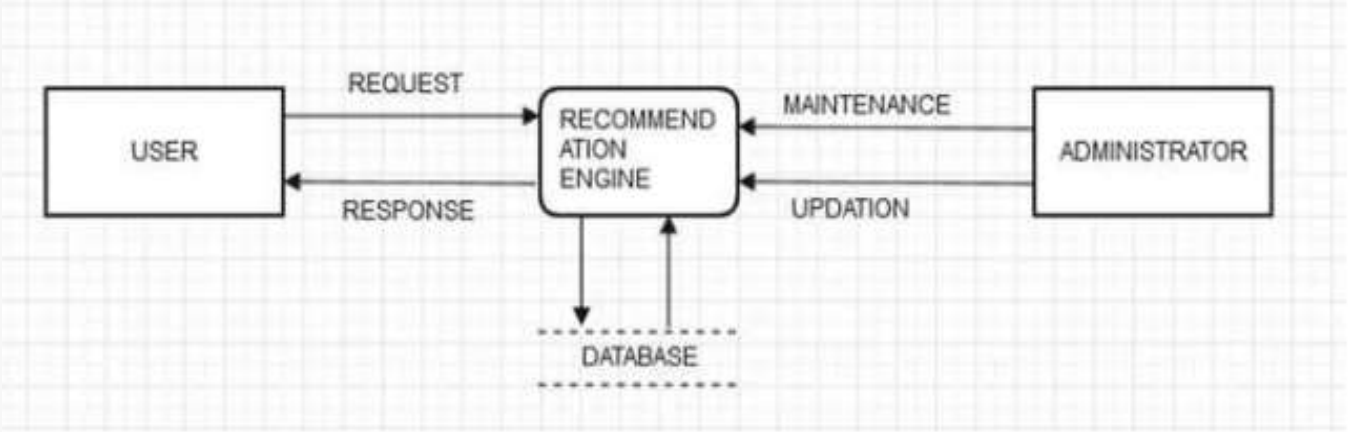
Class Diagram



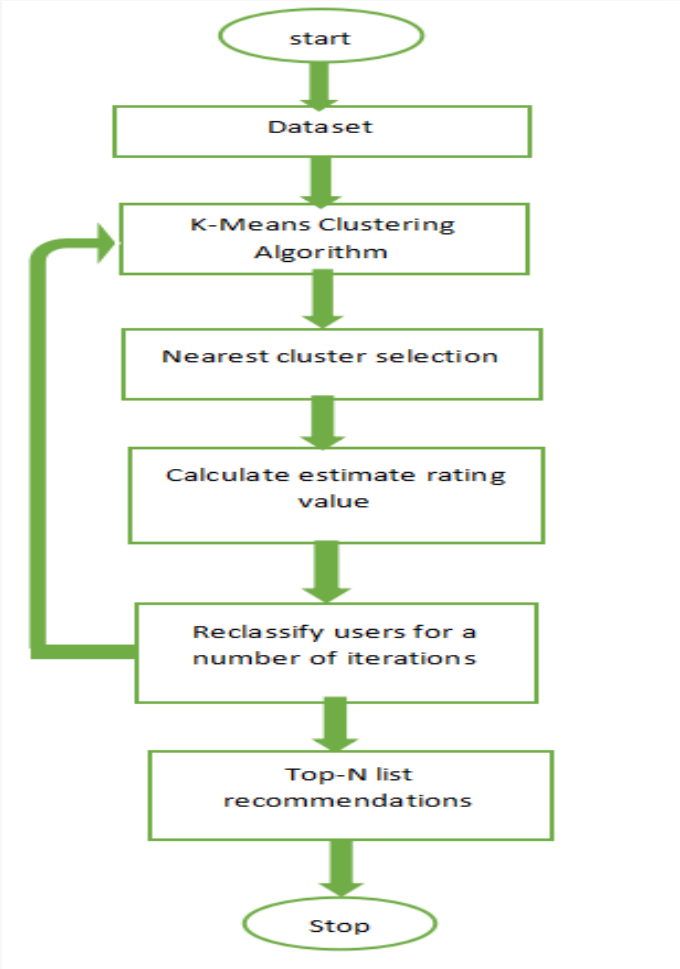
Use Case Diagram



Dataflow Diagram



Activity Diagram



Output:

1.Import and observe dataset

Out[1]:

	rank	title	genre	wiki_plot	imdb_plot
0	0	The Godfather	[u' Crime', u' Drama']	On the day of his only daughter's wedding, Vit...	In late summer 1945, guests are gathered for t...
1	1	The Shawshank Redemption	[u' Crime', u' Drama']	In 1947, banker Andy Dufresne is convicted of ...	In 1947, Andy Dufresne (Tim Robbins), a banker...
2	2	Schindler's List	[u' Biography', u' Drama', u' History']	In 1939, the Germans move Polish Jews into the...	The relocation of Polish Jews from surrounding...
3	3	Raging Bull	[u' Biography', u' Drama', u' Sport']	In a brief scene in 1964, an aging, overweight...	The film opens in 1964, where an older and fat...
4	4	Casablanca	[u' Drama', u' Romance', u' War']	It is early December 1941. American expatriate...	In the early years of World War II, December 1...
...	...	...	...	...	...
95	95	Rebel Without a Cause	[u' Drama']	\n\n\n\n\nJim Stark is in police custody.\n\n \...	Shortly after moving to Los Angeles with his p...
96	96	Rear Window	[u' Mystery', u' Thriller']	\n\n\n\nJames Stewart as L.B. Jefferies\n\n \...	L.B. "Jeff" Jeffries (James Stewart) recuperat...
97	97	The Third Man	[u' Film-Noir', u' Mystery', u' Thriller']	\n\n\n\nSocial network mapping all major chara...	Sights of Vienna, Austria, flash across the sc...
98	98	North by Northwest	[u' Mystery', u' Thriller']	Advertising executive Roger O. Thornhill is mi...	At the end of an ordinary work day, advertisin...
99	99	Yankee Doodle Dandy	[u' Biography', u' Drama', u' Musical']	\n In the early days of World War II, Cohan ...	NaN

100 rows × 5 columns

2.Combine Wikipedia and IMDb plot summaries

Out[2]:

	rank	title	genre	wiki_plot	imdb_plot	plot
0	0	The Godfather	[u' Crime', u' Drama']	On the day of his only daughter's wedding, Vit...	In late summer 1945, guests are gathered for t...	On the day of his only daughter's wedding, Vit...
1	1	The Shawshank Redemption	[u' Crime', u' Drama']	In 1947, banker Andy Dufresne is convicted of ...	In 1947, Andy Dufresne (Tim Robbins), a banker...	In 1947, banker Andy Dufresne is convicted of ...
2	2	Schindler's List	[u' Biography', u' Drama', u' History']	In 1939, the Germans move Polish Jews into the...	The relocation of Polish Jews from surrounding...	In 1939, the Germans move Polish Jews into the...
3	3	Raging Bull	[u' Biography', u' Drama', u' Sport']	In a brief scene in 1964, an aging, overweight...	The film opens in 1964, where an older and fat...	In a brief scene in 1964, an aging, overweight...
4	4	Casablanca	[u' Drama', u' Romance', u' War']	It is early December 1941. American expatriate...	In the early years of World War II, December 1...	It is early December 1941. American expatriate...

3.Tokenization

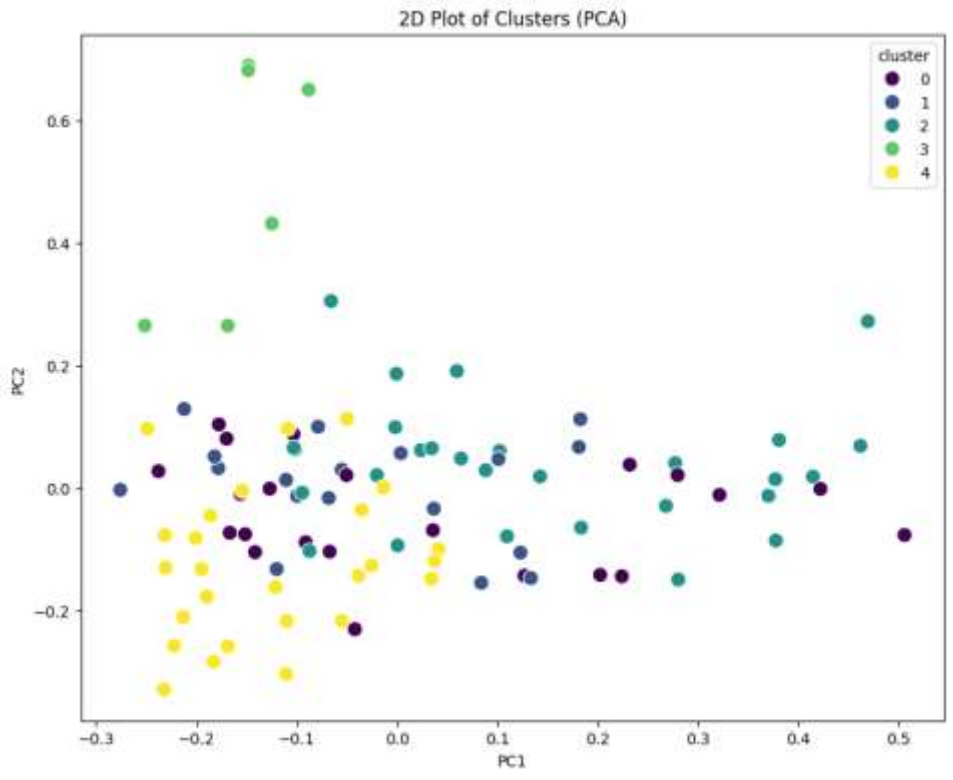
```
Out[3]:  ['Today', 'May', 'is', 'his', 'only', 'daughter', "'s", 'wedding']
```

In [4]:

```
# Import the SnowballStemmer to perform stemming
from nltk.stem.snowball import SnowballStemmer

# Create an English language SnowballStemmer object
stemmer = SnowballStemmer("english")

# Print filtered to observe words without stemming
print("Without stemming: ", filtered)

# Stem the words from filtered and store in stemmed_words
stemmed_words = [stemmer.stem(t) for t in filtered]

# Print the stemmed_words to observe words after stemming
print("After stemming:   ", stemmed_words)
```

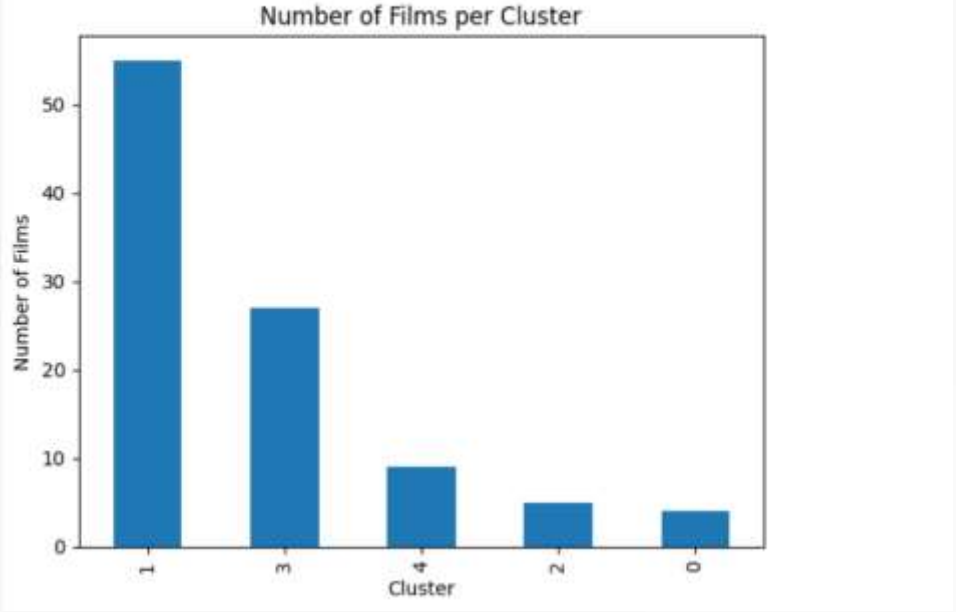
```
Without stemming:  ['Today', 'May', 'is', 'his', 'only', 'daughter', "'s", 'wedding']
After stemming:    ['today', 'may', 'is', 'his', 'onli', 'daughter', "'s", 'wed']
```

4. 2D Plot of clusters



5. Films per Cluster analysis

```
cluster
Collapse Output
4     9
2     5
0     4
Name: count, dtype: int64
```



6. Final Output

```
Apocalypse Now
The Lord of the Rings: The Return of the King
Gladiator
From Here to Eternity
Saving Private Ryan
Raiders of the Lost Ark
Patton
Jaws
Platoon
Dances with Wolves
The Pianist
The Deer Hunter
All Quiet on the Western Front
Mutiny on the Bounty
Enter a movie title: ↑↓ for history, Search history with c-↑/c-↓
```

```
Enter a movie title:  Braveheart

Movies similar to 'Braveheart':
One Flew Over the Cuckoo's Nest
Gone with the Wind
The Wizard of Oz
Titanic
Forrest Gump
E.T. the Extra-Terrestrial
2001: A Space Odyssey
Chinatown
12 Angry Men
To Kill a Mockingbird
My Fair Lady
Ben-Hur
Doctor Zhivago
```

Conclusion

In summary, the project's purpose is to leverage K-means clustering to extract actionable insights from the OTT TV shows dataset, contributing to enhanced content delivery, improved user experiences, and strategic decision-making within the dynamic landscape of over-the-top streaming platforms. The project's findings are expected to contribute to an optimized content ecosystem, enhanced user satisfaction, and strategic decision-making within the dynamic and competitive OTT industry.

References/Bibliography

- R. Ahuja, A. Solanki and A. Nayyar, "Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019, pp. 263-268, doi: 10.1109/CONFLUENCE.2019.8776969.
- Gupta, Meenu et al. “Movie Recommender System Using Collaborative Filtering.” 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (2020): 415-420.
- Ashrita Kashyap, Sunita. B, Sneh Srivastava, Aishwarya PH, Anup Jung Shah (2020), “A Movie Recommender System: MOVREC using Machine Learning”, IJESC Research Article Volume 10 IssueNo.06.
- N Pavitha, Vithika Pungliya, Ankur Raut, Roshita Bhonsle (2022), “Movie Recommendation and Sentiment Analysis Using Machine Learning”, Global Transitions Proceedings.
- Hirdesh Shivhare, Anshul Gupta and Shalki Sharma (2015), “Recommender system using fuzzy c-means clustering and genetic algorithm-based weighted similarity measure”, IEEE International Conference on Computer, Communication and Control.


THANK YOU