Unveiling Trends: Data Clustering Analysis of Netflix Tv Shows and Movies Through EDA

Jashandeep Singh
M.tech (Computer Science & Engineering), Baba Farid College of Engineering and Technology, Bathinda, India-151001

Abstract- **This paper explores unsupervised clustering analysis of Netflix's extensive collection of movies and TV shows using advanced techniques such as K-means, Agglomerative Clustering, and Affinity Propagation. Leveraging technologies like Word2Vec for word embedding, the study focuses on optimizing clustering models through meticulous data preprocessing, text cleaning, and hyper-parameter tuning. Key criteria such as Silhouette Score, Elbow Method, and Dendrogram are employed to determine the optimal number of clusters. Insights from exploratory data analysis reveal Netflix's strategic shift towards emphasizing TV content over movies globally. The findings contribute to understanding content preferences across different regions and showcase the platform's effective use of machine learning and AI for personalized recommendations.**

1. INTRODUCTION

Netflix has effectively harnessed AI, data science, and machine learning to personalize user experiences and optimize operations. Personalized recommendations are a core application, where sophisticated algorithms analyze viewing habits to suggest content tailored to individual preferences. This includes customizing thumbnails and ranking content to align with user interests. Netflix also uses AI to ensure high-quality streaming through adaptive streaming technology, dynamically adjusting video quality based on real-time bandwidth to prevent buffering. Additionally, AI plays a significant role in content creation and marketing. Predictive analytics guide decisions on producing or acquiring new content based on viewer trends and preferences. AI-driven targeted marketing and A/B testing enhance user engagement by personalizing promotional strategies and user interface designs. Beyond user-facing applications, AI helps manage content delivery networks and server loads, ensuring operational efficiency and security. This comprehensive integration of AI, data science, and machine learning positions Netflix as a leader in delivering a seamless, personalized streaming experience. The sole purpose of this to ensure whether this machine learning model works accurately or not and what will the accurate percentile of the outcome based on the viewer history.

Data set

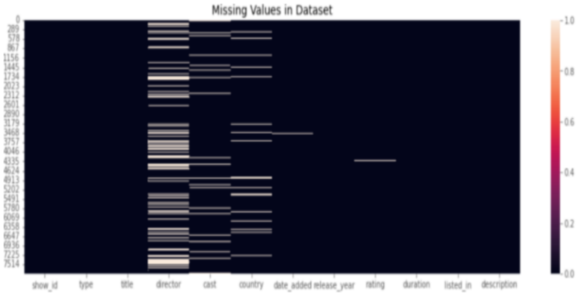


Fig 01. Missing Values in dataset

Our dataset consists of 7,787 entities and 12 features. Missing data is distributed as follows: 30.67% in the 'director' attribute, 9.22% in 'cast', 6.51% in 'country', and 0.0898% in 'rating'. These percentages highlight the extent of missing information across different attributes.

Data Cleaning

Removal of Missing Values in Dataset
0
289
578
867
1156
1445
1734
2023
2312
2601
2890
3179
3468
3757
4046
4335
4624
4913
5202
5491
5780
6069
6358
6647
6936
7225
7514
show_id
type
title
cast
country
date_added
release_year
rating
duration
listed_in
description
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100

Fig 02. Figures shows the missing values

To handle missing values in our dataset, we'll replace 'country' and 'rating' missing values with 'United

States' and 'TV-MA', respectively. Missing 'cast' values will be filled with 'unknown'. Given that 30.68% of the 'director' values are missing, we'll drop this column entirely. These steps will ensure a cleaner and more reliable dataset for analysis.

Exploratory data analysis

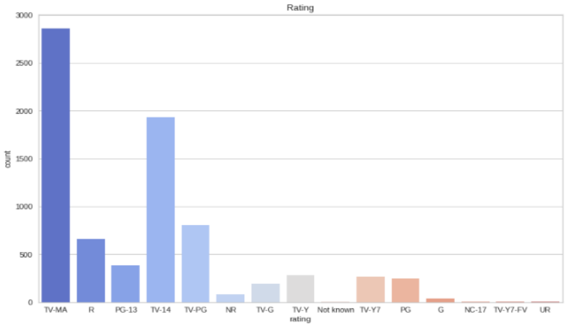


Fig03. Figure shows the exploratory data analysis of different movie and tv shows based on user ratings

2. LITERATURE SURVEY

The paper examines various clustering algorithms like k-Means, DBSCAN, and others, finding that k-Means performs well across datasets from Amazon Prime Video, Netflix, and Disney+. It also integrates Netflix data with additional geographical and review datasets for Exploratory and Sentiment Analysis using Python tools. The study introduces an efficient collaborative filtering method for the Netflix Prize problem, which compresses data into co-clusters and achieves high accuracy in real-time predictions. Additionally, it explores using auto-encoders and K-means clustering to enhance recommendation systems, significantly improving prediction accuracy for popular videos and optimizing video delivery in constrained environment. applying differential privacy to Netflix Prize algorithms, allowing for secure recommendations while protecting user data. It also discusses using Exploratory Data Analysis (EDA) and Sentiment Analysis on Netflix data to gain insights and the role of A/B testing in optimizing recommendation systems. Additionally, it investigates consumer perceptions of recommendation systems, revealing that while users feel in control, they are still influenced. Lastly, it reviews current techniques in recommender systems, reflecting on lessons from the Netflix Prize and suggesting future research directions.

3. HYBRID-CLUSTERING APPROACH

To enhance the clustering and recommendation system, we propose a hybrid approach that combines content-based and collaborative filtering to leverage both show attributes and user interactions. Incorporating deep learning models like neural collaborative filtering (NCF) and auto encoders will capture complex patterns, while context-aware recommendations will use contextual information such as time of day and device type.

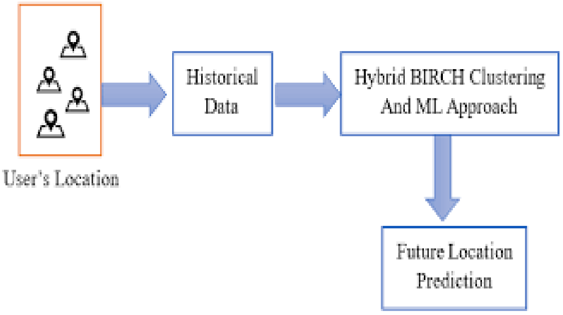
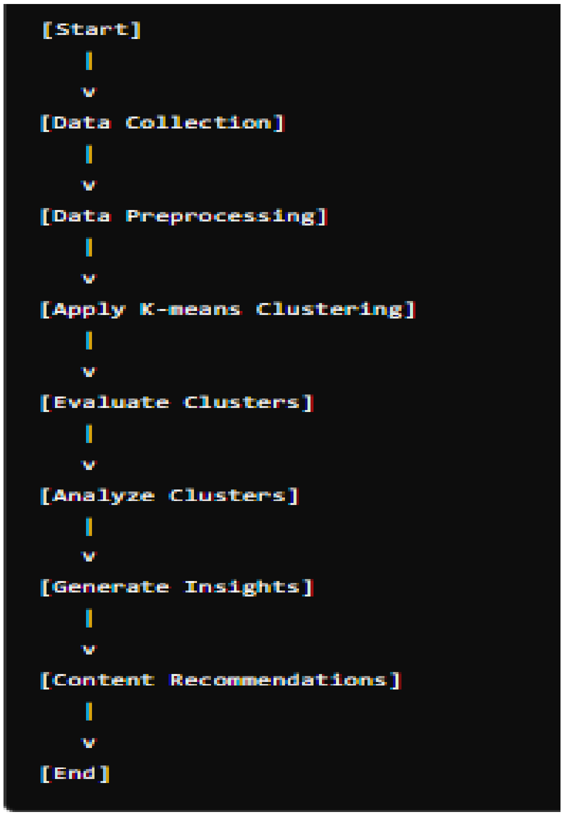

User's Location
Historical Data
Hybrid BIRCH Clustering And ML Approach
Future Location Prediction

Fig 04. Working Model of Hybrid Clustering

We will implement real-time updates and incremental learning to ensure the system adapts quickly to new data, and develop explainable AI models for transparency. A user feedback loop will further refine the recommendations, enhancing overall accuracy and user satisfaction.

4. K-MEANS CLUSTERING

The K-means algorithm is a clustering method that identifies a specified number (k) of centroids and allocates each data point to the nearest cluster, aiming to keep the clusters as compact as possible. Initially, it starts with randomly selected centroids and iteratively adjusts their positions to minimize the distance between data points and centroids. This process continues until the centroids stabilize, meaning their positions no longer change significantly, indicating successful clustering, or until a predefined number of iterations is reached. The goal is to optimize the clustering by ensuring data points are closely grouped around the centroids.



```
[Start]
   |
   v
[Data Collection]
   |
   v
[Data Preprocessing]
   |
   v
[Apply K-means Clustering]
   |
   v
[Evaluate Clusters]
   |
   v
[Analyze Clusters]
   |
   v
[Generate Insights]
   |
   v
[Content Recommendations]
   |
   v
[End]
```

Fig 05. Flow Chart Woking of K-mean Clustering

Here's a flow chart representation of how the K-means algorithm works in a Netflix exploratory data analysis:

1. Start
   - Begin the exploratory data analysis (EDA) process.
2. Data Collection
   - Collect Netflix show data, including attributes such as genre, cast, director, description, etc.
3. Data Preprocessing
   - Handle missing values.
   - Tokenize and vectorize text attributes (e.g., genre, description).
4. Apply K-means Clustering
   - Initialize random centroids.
   - Assign each data point to the nearest centroid.
   - Calculate new centroids based on the mean of the data points in each cluster.
   - Repeat the assignment and centroid calculation steps until convergence or a predefined number of iterations is reached.
5. Evaluate Clusters
   - Assess the quality of clusters using metrics like Silhouette score or Davies-Bouldin index.
6. Analyze Clusters
   - Identify patterns and trends within clusters.
   - Determine common themes or popular attribute combinations.
7. Generate Insights
   - Understand user preferences.
   - Highlight niche categories.
   - Tailor marketing strategies.
8. Content Recommendations
   - Use cluster analysis to improve content recommendations for users.
9. End
   - Conclude the EDA process with actionable insights and improved recommendation systems.

The K-means algorithm can significantly enhance Netflix exploratory data analysis (EDA) by categorizing shows into distinct clusters based on similarities in attributes such as genre, cast, director, and description. This clustering helps identify patterns and trends within the dataset, such as common themes or popular combinations of attributes. For instance, it can reveal groups of shows that appeal to similar audiences or highlight niche categories. By analyzing these clusters, Netflix can better understand user preferences, improve content recommendations, and tailor marketing strategies to specific viewer segments, ultimately enhancing user satisfaction and engagement.

```
# finding optimal number of clusters for K Means

# Instantiate the clustering model and visualizer
model = KMeans(tol=1e-4,random_state = 42)
visualizer = KElbowVisualizer(model, k=(2,22), metric='silhouette', timings=False)

visualizer.fit(transformed_data)
visualizer.show()
```

Fig 06. How K-means code works in analysis.

5. GUSSIAN CLUSTERING

Gaussian Mixture Models (GMMs) can enhance clustering for Netflix by fitting multiple Gaussian distributions to the data, capturing the underlying structure more flexibly than k-means. Unlike k-means, which assigns each data point to the nearest centroid, GMM estimates the parameters (mean, variance, and weight) for each Gaussian distribution and calculates the probability of each data point belonging to each cluster.

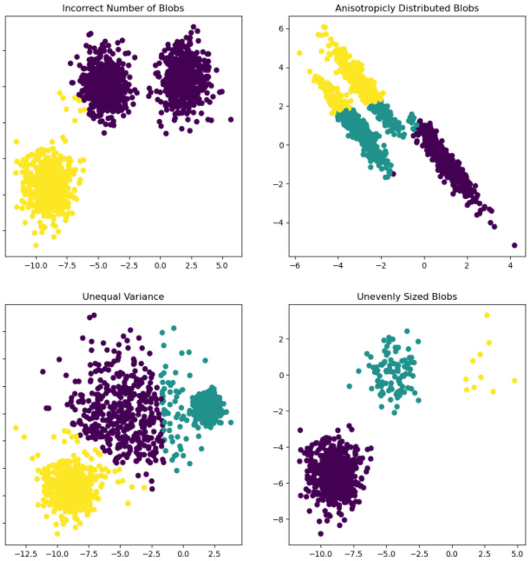


Fig 07. Working Model of Gaussian Clustering

This probabilistic approach allows GMM to model clusters with varying shapes and sizes and handle unequal sample sizes across clusters. By iteratively refining these parameters through the Expectation-Maximization (EM) algorithm, GMM provides a nuanced understanding of user preferences and content characteristics, improving content recommendations and user segmentation on Netflix.

6. RESULTS

The project aims to classify Netflix shows into distinct clusters based on attributes such as director, cast, country, genre, and description. By processing the data with TFIDF vectorization and reducing dimensionality through PCA, the k-means algorithm identified six optimal clusters, while Agglomerative clustering revealed twelve. These clusters reveal distinct groupings of shows, which can inform better content recommendations. Additionally, a content-based recommender system was developed using a cosine similarity matrix to provide personalized show suggestions based on user viewing history. Overall, this approach enhances Netflix's ability to categorize shows accurately and improve recommendation accuracy for users.

7. CONCLUSION

In this project, we aimed to categorize Netflix shows into distinct clusters based on their similarities. Using a dataset of 7,787 records and 11 attributes, we began by addressing missing values and performing exploratory data analysis (EDA), which revealed that Netflix primarily features more movies than TV shows, with a strong focus on U.S. productions for adults and young adults. We tokenized and vectorized attributes such as director, cast, country, genre, and description with TFIDF, generating 20,000 features. To manage this high dimensionality, we applied Principal Component Analysis (PCA), reducing the data to 4,000 components while retaining over 80% of the variance. We determined an optimal number of six clusters using k-means and twelve clusters with Agglomerative clustering. Finally, we created a content-based recommender system using a cosine similarity matrix to recommend ten shows to users based on their viewing history.

8. REFRENCES

[1] C. I. Johnpaul, "Distributed in-memory cluster computing approach in scala for solving graph data applications", Int. Conf. on Advances in Electronics Computers and Communications, pp. 1-6, 2014.

[2] Shoban and N. Samba, "Predicting The Mis-usability Of Data From Malicious Insiders", Int. Journal of Computer Engg & Appl, vol. V, no. II, 2014.

[3] P. Indira Priya and D. K. Ghosh, "A Survey on Different Clustering Algorithms in Data Mining Technique", Int. Journal of Modern Engineering Research, vol. 3, no. 1, pp. 267-274, 2013.

[4] Lin, Y.-T., Yen, C.-C., & Wang, J.-S. (2020). Video Popularity Prediction: An Autoencoder Approach With Clustering. IEEE Access, 8, 129285–129299. https://doi.org/10.1109/access.2020.3009253

[5] Daru, S. (2009) (PDF) pervasive parallelism in data mining: Dataflow solution to co-clustering large and sparse Netflix data. Available at:

https://www.researchgate.net/publication/221653512_Pervasive_parallelism_in_data_mining_Dataflow_solution_to_co-clustering_large_and _sparse _netflix_data.

[6] N. Ramakrishnan and S. Rani S, "Hypergraph based clustering for document similarity using FP growth algorithm", Int. Conf. on Intelligent Computing and Control Systems, pp. 332-336, 2019.

[7] Vadloori, K. B., & Sanghishetty, S. M. (2021). Exploratory and Sentiment Analysis Netflix Data. International Journal of Engineering Research & Technology (IJERT), Vol. 10(09), 213–216.

[8] B. K. Reddy, "Intrusion Detection System Using Computationally Efficient SVM based on K-Prototype Clustering", Third Int. Conf. on Intelligent Computing Instrumentation and Control Technologies, pp. 1153-1158, 2022.