Netflix Content Clustering Project Report

1. Project Overview

This project involves clustering Netflix Movies and TV Shows using unsupervised machine learning techniques.

The aim is to discover hidden patterns in content types, genres, duration, countries, and other metadata to assist in personalized recommendations, marketing segmentation, and content curation.

2. Data Collection & Cleaning

Dataset: 'NETFLIX MOVIES AND TV SHOWS CLUSTERING.csv'

- Loaded using pandas and inspected for structure.

- Removed duplicates and handled missing values:

- 'director', 'cast', 'country' filled with 'Unknown'.

- 'duration' standardized to numeric (minutes/seasons).

- 'rating' imputed using mode.

3. Feature Engineering

Features created:

- 'duration_numeric': Converted movie/season duration to numbers.

- 'type_encoded': Label encoded Movie/TV Show type.

- One-hot encoded 'rating' and top 10 'country' values.

- 'content_age': Current year - release year.

- 'genre_count': Count of genres per content.

- TF-IDF applied to 'listed_in', 'description', 'cast', 'director' for text features.

- All numerical features standardized using StandardScaler.

Netflix Content Clustering Project Report

4. Clustering Algorithms

Three clustering algorithms were applied:

- K-Means (optimal K = 5 using Elbow and Silhouette)

- Agglomerative Clustering (Hierarchical, K = 5)

- DBSCAN (Density-Based, eps=0.5, min_samples=5)

Dimensionality reduction (PCA, t-SNE) used for 2D visualization.

5. Evaluation Metrics

Model performance metrics:

K-Means:

- Silhouette Score: 0.395

- Davies-Bouldin Index: 1.289

- Inertia: 1824.55

Agglomerative:

- Silhouette Score: 0.378

- Davies-Bouldin Index: 1.348

DBSCAN:

- Silhouette Score: 0.297 (excluding noise)

- Davies-Bouldin Index: 1.515 (excluding noise)

Netflix Content Clustering Project Report

6. Results & Insights

K-Means performed the best with clean cluster boundaries and meaningful groupings.

Cluster Highlights:

- Cluster 0: Older foreign TV shows

- Cluster 1: Recent US-based adult content (TV-MA)

- Cluster 2: Kids content (TV-Y)

- Cluster 3: Short movies with mixed genres

- Cluster 4: Documentaries and unknown categories

Correlation heatmaps showed 'rating', 'duration', and 'country' significantly influence clustering.

7. Conclusion

K-Means is the most effective method for categorizing Netflix content in this case.

Use cases:

- Personalized recommendations

- Regional marketing campaigns

- Inventory analysis for production teams

Future improvements:

- Include user viewing history

- Use deep learning for better text embeddings