### **📘 Netflix Movies and TV Shows Clustering**

**Project Report**

#### **1. Project Overview**

This project explores clustering techniques to group similar movies, and TV shows available on Netflix. The goal is to uncover patterns in content types, genres, and other attributes using unsupervised machine learning.

#### **2. Objectives**

* To segment Netflix titles based on their metadata such as genre, type, cast, and release year.
* To identify meaningful content clusters using algorithms like K-Means, DBSCAN, and Hierarchical Clustering.
* To visualize content similarity and trends using dimensionality reduction techniques.

#### **3. Dataset Description**

* **Source**: Netflix content dataset
* **Total Records**: *e.g., 8807 titles*
* **Key Features**:
  + title
  + type (Movie/TV Show)
  + release\_year
  + listed\_in (Genres)
  + cast, country, description

#### **4. Data Preprocessing**

* Handled missing values in cast, country, etc.
* Extracted main genres from listed\_in
* Encoded categorical features using:
  + Label Encoding (type)
  + One-Hot Encoding (genres)
* Text data transformed using **TF-IDF** (e.g., on description or listed\_in)

#### **5. Feature Engineering**

* Constructed genre flags from listed\_in (e.g., genre\_Drama, genre\_Comedy)
* Vectorized text features using TfidfVectorizer
* Scaled numerical data using StandardScaler

#### **6. Clustering Techniques**

* **K-Means Clustering**:
  + Elbow method used to determine optimal k
  + Evaluated using Inertia and Silhouette Score
* **Hierarchical Clustering**:
  + Agglomerative clustering with Ward linkage
  + Dendrogram for visual inspection
* **DBSCAN**:
  + Density-based clustering to capture non-spherical groups

#### **7. Dimensionality Reduction**

* Applied **PCA** for 2D/3D visualization
* Also used **t-SNE** for improved cluster separation

#### **8. Evaluation Metrics**

* **Silhouette Score**
* **Davies-Bouldin Index**
* **Cluster Compactness (Inertia for K-Means)**

#### **9. Results & Insights**

* Identified *N* clusters (e.g., action-heavy shows, drama-centric TV series, etc.)
* PCA proved most effective for visualizing genre-based clusters
* DBSCAN detected noise/outliers not captured by K-Means

#### **10. Conclusion**

Clustering Netflix content by K-Means helps uncover structure in a large content library and can support recommendation systems or catalog reorganization. Future enhancements can include:

* Incorporating user ratings or viewership data
* Advanced NLP on description field

#### **11. Tools Used**

* Python
* Libraries: scikit-learn, pandas, matplotlib, seaborn
* Jupyter Notebook