Netflix Movies and TV Shows Clustering

To apply unsupervised machine learning for clustering Netflix's movies and TV shows based on descriptive features enhancing content discoverability, enabling nuanced genre exploration, and supporting personalized recommendations without user interaction data.

Data Acquisition, Data Cleaning and Preprocessing

Netflix Movies & TV Shows datasets with Title, Description, Genre, Director, Cast, Country, etc. Handling missing values; Standardizing & normalizing text data. Removing unnecessary columns.

Feature Engineering (NLP)

Converting text fields (description, cast, etc.) to features;

Techniques: TF-IDF for vectorization

Dimensionality Reduction

Applying PCA to reduce high-dimensional vectors; Balancing interpretability and information retention

Clustering, Cluster Evaluation

K-Means Clustering; Hierarchical Agglomerative Clustering; Metrics: Silhouette, Calinski-Harabasz, Davies-Bouldin ; Assessing quality and separation of clusters.

Visualization and Content-Based Recommendation

Scatter plots; Word clouds and Bar charts for distribution in clusters. Finding similar titles using cosine similarity. Suggesting items within or across clusters.

Understanding the data

Knowing the data

Understanding the variables

Finding Unique values

Description

Variables Description

- show_id : Unique ID for every Movie/Show
- type : Identifier - Movie/Show
- title : Title of the Movie/Show
- director : Director of the Movie/Show
- cast : Actors involved in the Movie/Show
- country : Country where the Movie/Show was produced
- date_added : Date it was added on Netflix
- release_year : Actual Release year of the Movie/Show
- rating : TV Rating of the Movie/Show
- duration : Total Duration - in minutes or number of seasons
- listed_in : Genre
- description : The Summary description
- The unique values of all the features were analysed.
- By utilizing these features, we plan to create a cluster column and implement both K-means and Hierarchical clustering algorithms.

Exploratory data analysis and data cleaning

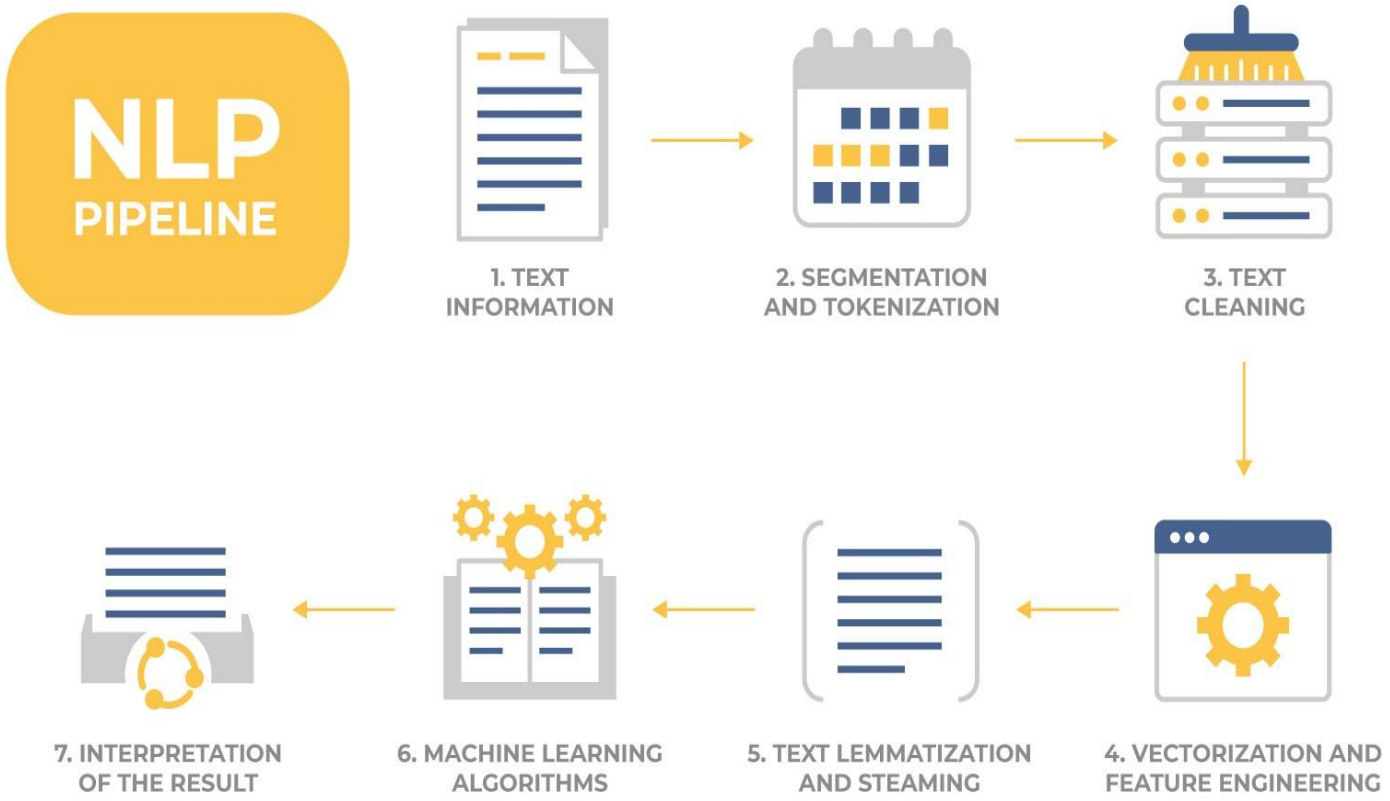
Exploratory data analysis (EDA)

- Analysis of all columns
- Removing the missing values
- Removing the Duplicate values
- Handling Outliers

Description

- The goal of EDA is to gain insights into the data, identify patterns, and discover relationships and trends. It is an iterative process that helps to identify outliers, missing values, and any other issues that may affect the analysis and modeling of the data.
- EDA revealed key content trends—movies dominate the catalog (69%), with genres like dramas, comedies, and international films being most common. Frequent keywords such as “love,” “life,” and “family” highlight emotional and relational themes central to Netflix’s content.
- Data cleaning may include tasks such as removing duplicate records, filling in missing values, correcting errors, and standardizing data formats. Except for the release year, almost all of the data are presented in text format.
- The textual format contains the data we need to build a cluster/building model. Therefore, there is no need to handle outliers.

NLP/Text Transformation

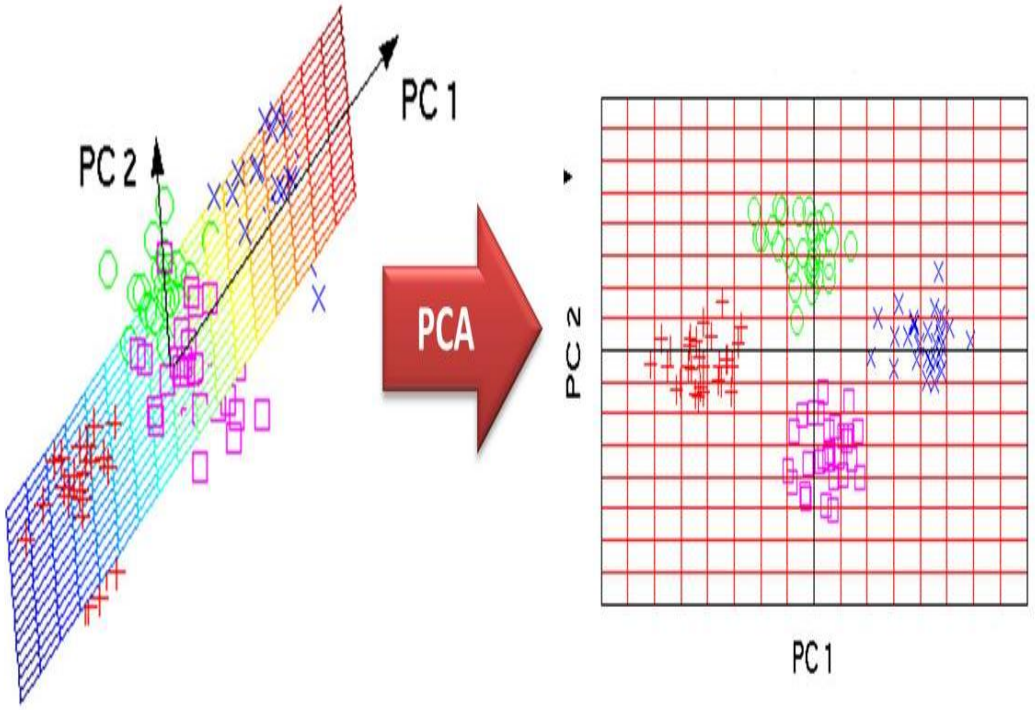


Modeling Approach

Choose the attributes that you want to cluster.

- Text Preprocessing: Change all textual data to lowercase and eliminate all punctuation marks and stopwords. Removing commonly occurring words such as "the", "and", "a", etc. that don't carry much meaning.
- Stemming or Lemmatization: Normalizing the words by reducing them to their base form.
- Tokenization: Breaking the text into smaller units, such as sentences or words.
- Dimensionality reduction.
- Unsupervised training ML algorithms to cluster the movies and various techniques to determine the optimal number of clusters.
- Visualization of optimal number of clusters (using scatter plot, bar plot) and use wordclouds to display the contents of each cluster.

Principal Component Analysis



Dimensionality Reduction:

- Dimensionality reduction simplifies high-dimensional data while preserving important patterns and structures.
- It helps improve model performance
- Reduce overfitting
- Enable better visualization
- Techniques included: PCA (Principal Component Analysis)t-SNE, Autoencoders, etc.
- In our case, we use PCA to project the data into lower dimensions for easier clustering and visualization.
- Also, we have used t-SNE to scatter-plot our clustered shows in 2D space, showing how different genres or types group together after clustering.

Unsupervised Machine Learning

Algorithm	Key Idea	Best For	Core Parameter(s)
K-Means Clustering	Partitions data into K clusters by minimizing intra-cluster variance.	Spherical, equally sized clusters	n_clusters
Hierarchical Clustering	Builds a tree (dendrogram) of nested clusters via merging or splitting.	Hierarchical or nested groupings	linkage, distance_threshold

Selection of optimum value of K

- Elbow method
- Silhouette analysis

Evaluation Metrics

- Calinski-Harabasz Score
- Davies-Bouldin Score

- K-Means is an unsupervised learning algorithm that partitions data into K clusters based on similarity.
- Each cluster is represented by a centroid, and points are grouped by their proximity to it.
- Objective: Minimize the distance between data points and their cluster centroids.
- Evaluation: Elbow Method: Optimal K ≈ 6 (point of diminishing returns).
- Silhouette Score: Highest at K = 6, indicating well-separated clusters
- Result: Built 6 optimal clusters capturing distinct content groupings.

- To assess how well the data has been clustered, we use three key evaluation metrics:

- **Silhouette Score**
  Measures how similar each point is to its own cluster vs. others. Range: –1 to 1 → Higher is better (well-separated clusters).

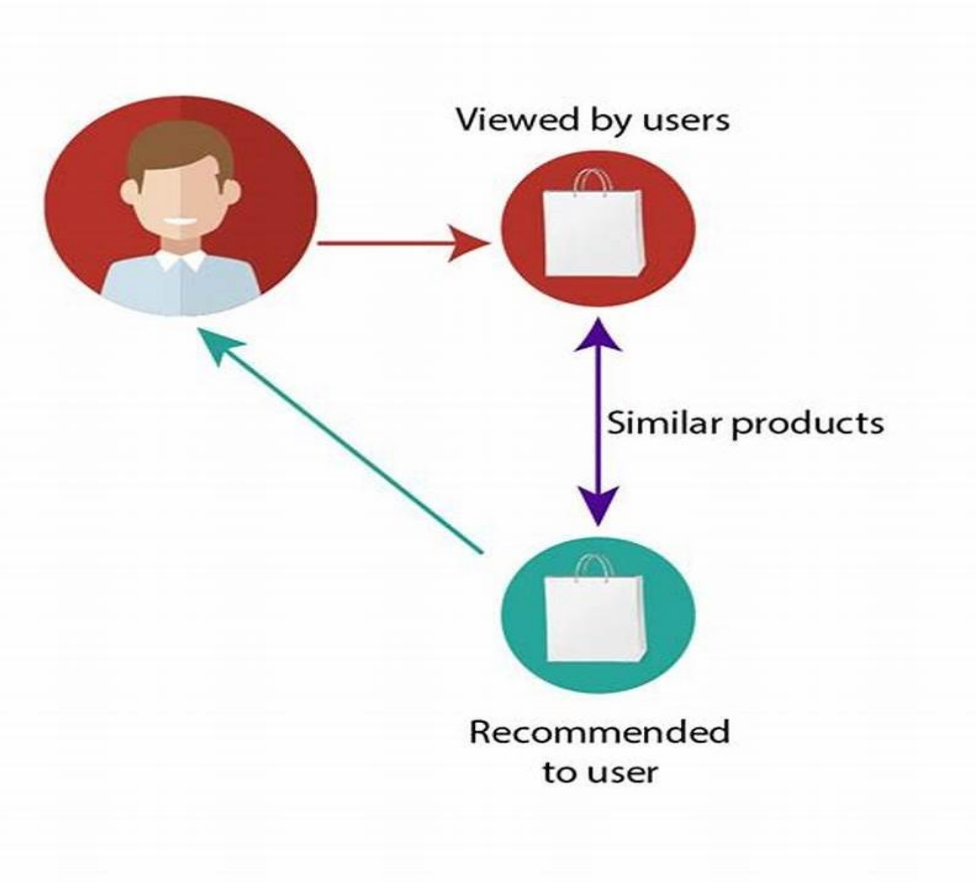
- **Calinski-Harabasz Score**
  Ratio of between-cluster dispersion to within-cluster dispersion. Higher values indicate better-defined clusters.

- **Davies-Bouldin Score**
  Measures average similarity between clusters (lower is better). Lower values imply better separation and compactness.

Evaluation Metrics

Algorithm	Calinski-Harabasz Score	Davies-Bouldin Score	Silhouette score	Distortion (approx)
K-Means Clustering	61.958	10.33	0.013	6274.23
Hierarchical Clustering	22.75	8.541	0.0054	6205.99

User Recommendation System



- A content-based recommender suggests shows similar to those a user already liked, using show attributes.
- We have used cosine similarity to measure how alike two shows are?
- A higher similarity score means closer content match.
- Helps personalize suggestions based on user preferences.

Conclusion

- In this project, we focused on clustering Netflix shows based on textual attributes to group similar content together.
- Starting with a dataset of 7,787 records and 11 features, we handled missing data and performed exploratory data analysis (EDA).
- We found that Netflix primarily hosts movies, with most content produced in the United States, and the overall content volume is rapidly increasing.
- We selected key textual attributes, such as cast, country, genre, director, rating, and description and processed them using TF-IDF vectorization, resulting in around 10,000 features.
- To manage high dimensionality, PCA was applied, reducing it to 3,000 components that captured over 80% of the variance.
- Clustering was performed using K-Means, where the optimal number of clusters was found to be 4, based on the elbow method and Silhouette score.
- We also applied Agglomerative Hierarchical Clustering, with the dendrogram suggesting 13 clusters.

Conclusion

- Finally, we built a content-based recommendation system using cosine similarity, which provides personalized suggestions by recommending 10 similar shows based on user input.

- Regional/Genre-Specific Trends: Example: Cluster 0: Documentaries, International Movies; Cluster 2: International TV Shows, TV Dramas; Impact: Enables targeted content acquisition by understanding audience preferences by region and genre.

- Improved User Experience: Thematic Clusters help viewers explore content based on mood or theme rather than rigid genres. Examples: Users interested in Comedies & Family Movies can explore Cluster 1 easily.Encourages organic discovery and longer engagement on the platform.