Netflix Movies and Tv Shows Clustering

Vishal Mahadev Potdar
(Potdar15v@gmail.com)

Abstract :

The Netflix Movies and TV Shows Clustering project is a machine learning project that aims to group similar movies and TV shows available on the Netflix platform into clusters based on their shared attributes. The project uses unsupervised learning techniques to analyse the dataset, including k-means, hierarchical clustering, and DBSCAN algorithms.

By clustering the data, the project aims to identify patterns in the content available on Netflix, which can be used to improve the platform's recommendation system and provide better viewing suggestions to its users.

Table of contents :

List of terms :

Feature engineering :- the pre-processing step of machine learning, which extracts features from raw data.

Univariate analysis :- Univariate analysis explores each variable in a data set, separately.

Bivariate analysis :- The bivariate analysis is will measure the correlations between the two variables.

Multivariate analysis :- is a Statistical procedure for analysis of data involving more than two types of measurement or observation.

Acknowledgments :

Thank you AImabetter for your help and support in completing my ML project. Your guidance and advice was invaluable and I am grateful for all the help you provided.

Problem Statement :

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

In this project, you are required to do :

- Exploratory Data Analysis
- Understanding what type content is available in different countries
- Is Netflix has increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text-based features

Attribute Information :

- **show_id :** Unique ID for every Movie / Tv Show
- **type :** Identifier - A Movie or TV Show
- **title :** Title of the Movie / Tv Show
- **director :** Director of the Movie / TV Show

Netflix Movies and Tv Shows Clustering

Vishal Mahadev Potdar
(Potdar15v@gmail.com)

- **cast :** Actors involved
- **country :** Country of production
- **date_added :** Date it was added on Netflix
- **release_year :** Actual Release year of the movie / TV show
- **rating :** TV Rating of the movie / TV show
- **duration :** Total Duration in minutes or number of seasons
- **listed_in :** Geners
- **description :** The Summary description

Introduction :

The Netflix Movies and TV Shows Clustering project aims to cluster similar movies and tv shows available on Netflix into different clusters based on their content. The project begins with data collected from the third-party Netflix search engine, which contains information about more than 7,000 movies and TV shows available on the platform.

Background :

By clustering the data, the project aims to identify patterns in the content available on Netflix, which can be used to improve the platform's recommendation system and provide better viewing suggestions to its users.

The results of this project can also be useful to content creators and producers who are looking to understand what types of content are popular among viewers.

Materials and apparatus:

Materials:

1. Computer: A desktop or laptop computer with an operating system capable of running machine learning algorithms is essential for any machine learning project.
2. Data: Datasets are an essential component of any machine learning project. The type of data required for a specific project will depend on the type of machine learning task it is being used for.
3. Software: Machine learning algorithms need to be implemented in a programming language. Popular choices are Python used.
4. Libraries: Libraries are collections of pre-written functions that can be used in the implementation of machine learning algorithms. Popular libraries include scikit-learn used.

Apparatus:

1. GPU: Graphics processing units (GPUs) are specialized hardware used to speed up the training process of machine learning algorithms. GPUs can significantly reduce the time required to train a machine-learning model.

Procedure :

[1] Import, Loading and Inspection of Data:

After importing the dataset, we look at its columns and shape. The info() method is used to verify variables and associated datatypes for null values. Using the describe() function, we can determine the fundamental characteristics of each *variable, such as the mean, median, count, and so on.*

We can better comprehend the meaning of the variable thanks to the supplied variable description. This helped us understand datasets.

[2] Handling duplicated values :

Fortunately, there aren't any duplicate values in the dataset, but if there are, you can get rid of them with the drop_duplicated() method or by replacing it.

[3] Handling null values :

Because there are numerous null values for features such as director, cast, and country, those null values cannot be dropped; instead, they have been replaced with Director Unavailable, Cast

Netflix Movies and Tv Shows Clustering

Vishal Mahadev Potdar
(Potdar15v@gmail.com)

Unavailable, and Country Unavailable, respectively.

Features such as date_added and rating have a very low number of null values, so we dropped those null values.

[4] Handling outliers :

Boxplot and distplot are used to detect outliers. The interquartile range approach is used to eliminate outliers from data.

[5] Feature engineering and data wrangling:

Using feature engineering, we generate new variables from the original one.

[6] Exploratory data analysis :

We use a count plot, bar plot, line plot, heatmap, box plot, and distribution plot for exploratory data analysis.

Univariate, bivariate, and multivariate analyses were performed using statistical techniques, and the results revealed insightful information.

[7] Data Pre-processing **:**

[a] Textual Data Preprocessing :

We processed text data from the description variable by removing punctuation, stopwords, whitespace, emails, html tags, urls, special characters, and digits. Vectorized after lemmanization using TFIDF vectorizer and got a TFIDF matrix for feeding to the model as input.

[8] Fitting different ML models :

We trained various clustering algorithms, such as KMeans clustering, Hierarchical clustering, and DBSCAN. The several clustering models were trained using the Sklearn library, and predictions were made.

[9] Evaluation of Model :

Metrics including silhouette_score, Calinski-Harabasz score, and davies bouldin score were used to further assess the model. Model KMeans clustering tops all classification evaluation metrics among all different implemented models.

Conclusion :

We tested numerous machine-learning models and assessed them using different classification evaluation metrics. The KMeans clustering model comes the closest to having the highest scores on all classification evaluation metrics.

The K-Means Clustering model is the optimal model and well-trained for clustering TV shows and movies based on the content due to its high Calinski-Harabasz score (9.039247) and silhouette score (0.004634), which are closer to 1 than other builded models.