**NETFLIX MOVIES AND TV SHOWS CLUSTERING**

**Lavanya Shinde**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract:**

Netflix is an American subscription-based streaming service that allows us to watch TV shows and movies without commercials on an internet-connected device. Additionally, we can download TV series and films to our iOS, Android, or Windows 10 device and view them offline.

This Netflix Movies/Shows Clustering Machine Learning Capstone Project was successfully completed by me. During the creation of this project, I mostly used the Python programming language and a few Python modules, including Pandas and Numpy for data wrangling and Matplotlib and Seaborn for data visualization. Additionally, we leverage machine learning concepts like DBSCAN, K-means clustering, and hierarchical clustering for this project.

***Keywords: exploratory data analysis, unsupervised ml clustering, DBSCAN, Hierarchical Clustering, K-means Clustering***

**1.Problem Statement :**

1) This dataset includes of Netflix-eligible television series and motion pictures as of 2019. The dataset was gathered through the third-party Netflix search engine Flixable.

2) The amount of TV series available on Netflix has almost tripled since 2010, according to an interesting analysis that was

published in 2018. Since 2010, the number of movies available on the streaming service has dropped by more than 2,000, although the number of TV series has nearly tripled. Investigating what further insights may be drawn from the same dataset will be intriguing.

3) While using this Netflix Data we Try to Understand :-

i) Understanding what type content is available in different countries.

ii) Is Netflix has increasingly focusing on TV rather than movies in recent years.

iii) Clustering similar content by matching text-based features.

**2. Introduction :**

1) Netflix began using user data analysis in 2006 to forecast how much a viewer will enjoy a film based on previous user preferences.

2) Whenever we access the Netflix service, the Netflix suggestions system aims to make it as easy as possible for us to find a show or movie to enjoy..

3) All of these actions were carried out using user data as inputs for our algorithms. A procedure or set of guidelines used in a problem-solving operation is known as an algorithm.

4) In this project, clustering is also used. Unsupervised learning is a common method for statistical data analysis that is utilized in many different domains.

I use a dataset of Netflix users for my project. Beginning in 2006, Netflix began using user data analysis to forecast how much a viewer will enjoy a film based on historical user preferences. The Netflix suggestions algorithm aims to make it as easy as possible for us to find a show or movie to enjoy each time we use the service. All of these actions were carried out using user data as inputs for our algorithms. Let's discuss the dataset as follows:- The shape of dataset is (7787 x 12) I.e. Total Number of Rows are : 7787, Total Number of Columns are : 12.

1. **Work of Flow :**

In the workflow process, I first gather and comprehend the data. I discovered there are 7787 rows and 12 numbers of columns while interpreting the data. The dataset is next checked to see whether any data are missing. I attempt to correct any missing data. Next, I begin the exploratory data analysis (EDA). Once the analysis is complete, i go on to the machine learning section, where i perform text processing, DBSCAN, K-means clustering, hierarchical clustering, and finally come to our conclusion.

1. **Data Review :**

1) The shape of dataset is (7787 x 12)

2) Total Number of Rows are : 7787

3) Total Number of Columns are : 12

4) Missing Value in Columnswise :-

i) director - 2389

ii) cast - 718

iii) country - 507

iv) date\_added - 10

v) rating - 7

5) The Total number of duplicate values in the data set is = 0



1. **Points We Cover With this Project :**

1) Visualize Total Release Movies/Tv shows in Last 10 years

2) Visualize Types of Videos on Netflix

3) Visualize the top 10 Countries that produced Highest Number of Movies/Shows on Netflix

4) Visualize Top 5 Rating Distribution for Movies and Shows on Netflix

5) Visualize the top Genres for Movies/TV-Shows on Netflix

6) Visualize the Top Directors on Netflix

7) Visualize the Top cast on Netflix till Year of 2020

8) Top Duration of Movies on Netflix

9) Highest Duration of TV Shows on Netflix

10) What type of content is available in different countries

11) Is Netflix has increasingly focusing on TV rather than movies in recent years ?

1. **Performing EDA :**

So we move on to our Next part of this project which is Exploratory data analysis (EDA).

Data analysis utilizing visual methods is called exploratory data analysis. With the aid of statistical information and graphical representations, it is used to identify trends and patterns as well as to test hypotheses. It entails making an effort to comprehend the provided data much better so that we can extract more meaning from it. Additionally, it is used to generate a value distribution to spot outliers and missing numbers.

A statistical model may or may not be employed in statistics, but in general, exploratory data analysis (EDA) is used to discover what the data may tell us beyond formal modelling or hypothesis testing tasks in Python. Data visualization is then utilized to identify relevant patterns and insights.All Visualisation Images Attach in PPT.

1. **Text Processing (Machine Learning ) :**

What is text Processing ?

:- It is the process of drawing important conclusions from texts. This technique is also used to group text into clusters and find other patterns in data. We lowercase all text, eliminate punctuation, and eliminate superfluous words. Here, comparable words are combined to conserve memory and processing time. Additionally, words and word groups are gathered to extract context-relevant information.

The steps involved in text preprocessing are :

● Tokenization :- Tokenization divides the original text into tokens, which are words and sentences. These tokens aid in context comprehension or model development for NLP. By examining the word order in the text, tokenization aids in comprehending the text's meaning.

● Punctuation Removal :- All the punctuations from the text are removed.

● Stopword Removal One of the most popular preprocessing techniques across various NLP applications is stop word removal. The simple idea is to exclude words that appear frequently throughout all of the corpus's documents..

● Stemming Words :- Stemming is the process of stripping a word down to its root, or lemma, which attaches to suffixes, prefixes, and other word parts..

All Visualisation Images Attach in PPT.

1. **Machine Learning Clustering :**

DBSCAN Clustering Algorithm

:- Density-Based Spatial Clustering of Applications with Noise is known as DBSCAN. According to the concept that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density, density-based clustering refers to unsupervised learning approaches that discover unique groups/clusters in the data. Two parameters are used by the DBSCAN algorithm.:

\* minPts: The minimum number of points (a threshold) clustered together for a region to be considered dense.

\* eps (ε): A distance measure that will be used to locate the points in the neighbourhood of any point.

After Performing DBSCAN clustered the data into 10 clusters with a silhouette score is 0.43875

All Visualisation Images Attach in PPT.

1. **Means Clustering :-**

K-means clustering, as we are aware, is a kind of unsupervised learning that is employed when we have unlabelled data (i.e. data which is either taken from nature or created by human to explore the scientific patterns behind it). This algorithm's primary goal is to identify groups in the data, with the variable K representing the approximate number of groups.

The technique separates the unlabelled dataset into K number of clusters, then iteratively searches for better clusters until it runs out of options.

K-means is performed after the elbow was clustered, and the eight clusters with the best silhouette score (0.474), Davies-Bouldin Index (0.884), and Calinski-Harbaz Score (2932.28) were discovered.

All Visualisation Images Attach in PPT.

**Hierarchical Clustering :-**

A cluster analysis technique called hierarchical clustering aims to create a hierarchy of groups. It is the most well-liked and frequently utilised technique for studying social network data. Nodes are compared to one another using this method based on how similar they are. By combining groups of nodes based on their similarity, larger groupings are created. The proximity matrix for each cluster is regenerated after merging the clusters that are extremely similar or close to one another.

The dendrogram distance was optimal at 20 after performing hierarchical clustering, with eight clusters yielding a silhouette score of 0.4705, Davies-Bouldin Index of 0.8839, and Calinski-Harbaz Score of 2930.84.All Visualisation Images Attach in PPT.

1. **Conclusion :**

1) After Visualize Total Release Movies/Tv shows in Last 10 years, we find out:-

i) 2018 is a year where Maximum number of movies have been released which is Total 1121.

ii) 2017 is a Second Highest year For Maximum number of movies have been released which is Total 1012.

iii) 2019 is a Third Highest year where 996 number of movies have been released.

2) A total of 5377 movies, or 69% of Netflix's content, are movies. There are 2410 TV shows total on Netflix, or 31% of the total content.

3) i) United State is a Country which is produced Highest Number of Movies/Shows on Netflix. Total Number of Movies/Show produced by US is 3296.

ii) India is a Second Highest Country which is produced Highest Number of Movies/Shows on Netflix. Total Number of Movies/Show produced by India is 990.

iii) United Kingdom is a Third Highest Country which is produced Highest Number of Movies/Shows on Netflix. Total Number of Movies/Show produced by UK is 722.

4) i) Tv-MA is a Highest Rating Distribution For Movies and Shows on Netflix, which is Total 2863.

ii) Tv-14 is a Second Highest Rating Distribution For Movies and Shows on Netflix, which is Total 1931.

iii) Tv-PG is a Third Highest Rating Distribution For Movies and Shows on Netflix, which is Total 806.

5) i) Drama is a Top Genres For Movies/TV Shows on Netflix. That is, the content of the movie in the drama

genre has been produced the most which is 2810 Times.

ii) Comedy is a Second Highest Genres For Movies/TV Shows on Netflix which is 2377 Times.

iii) Documentary is a Third Highest Genres For Movies/TV Shows on Netflix which is 1139 Times.

iv) Action Adventure are in 4th Position & Romance are in 5th Position.

6) i) Jan Suter is the top director in the Netflix industry & he has Directed 21 Movies/Shows.

ii) Raul Compose is the Second top director in the Netflix industry & he has Directed 19 Movies/Shows.

iii) Marcus Raboy is the Third top director in the Netflix industry & he has Directed 16 Movies/Shows.

iv) Jay Karas is the 4th position & he has Directed 15 Movies/Shows and Cathy Garcia-Molina is the 5th Position & he has Directed 13 Movies/Shows.

7) i) Anupam Kher is the Top Cast on Netflix As per our Visualization.

ii) Shah Rukh Khan is the second highest cast on Netflix.

iii) Naseeruddin Shah is the third highest cast on Netflix.

iv) Om Puri is the 4th Highest cast on Netflix.

v) Akshay Kumar is the 5th Highest cast on Netflix.

8 ) Most of the movies on Netflix have a duration range from 85 to 115 minutes.

9 ) Most TV shows on Netflix have a length of 1 season only.

10) i) Drama is the most produced genre in a lot of Non-English speaking countries.

ii) Comedy is the most produced genre in English speaking countries like United States of America and United Kingdom and Canada.

iii) Drama and Comedy are the most produced genres in the top countries with exceptions of Japan and South Korea.

iv) Japan is the biggest producer of Anime. Anime is also the most produced in genre in Japan.

v) Most South Korean content are from the Romance genre.

vi) Documentaries are mainly produced in United Kingdom and United States of America.

11) i) The above graph depicts seasons of TV shows signed vs the movies signed.

iiThis distinction is important since TV shows need ongoing funding for each season. As a result, TV viewership has increased in step with the seasons. because they were once viewed as a single entity.

iii) We can see that more TV programmes than movies were signed in 2016. While the number of movies signed has been higher, it is clear that the number of TV shows signed year is quickly catching up to the number of movies signed annually.

12) After Performing DBSCAN clustered the data into 10 clusters with a silhouette score is 0.43875.

13) The elbow and best silhouette score were discovered at 8 clusters after K-means clustering was completed. These clusters had a silhouette score of 0.474, a Davies-Bouldin Index of 0.884, and a Calinski-Harbaz Score of 2932.28.

14) The dendrogram distance was optimal at 20 after performing hierarchical clustering, with eight clusters yielding a silhouette score of 0.4705, Davies-Bouldin Index of 0.8839, and Calinski-Harbaz Score of 2930.84.

1. **References-**

1) Clustering in Machine learning By GeekforGeeks.

2) k-means Clustering in machine learning From TowardsDataScience By Team Education Ecosystem(LEDU).

3) DBSCAN Clustering Algorithm for Machine Learning From AnalyticsVidya By Abhishek Sharma.

4) Hierarchical Clustering in Machine Learning From JavaPoint .

5) Text Processing From Datarobot.