**Netflix Movies and TV Shows Clustering**

**Shubham Narendra Kadu**

**Data science trainees**

**AlmaBetter**

**Abstract:**

The study conducted was on the proposed “Netflix Movies and TV Shows Clustering”. The main purpose of this study was to build a DEC (Deep Embedded Cluster) with the given movie/tv show descriptions to find similar movies/tv shows as a recommender system.

Netflix is one of the leading OTT platforms, not only in India but also internationally Netflix manages a large collection of TV shows and movies, streaming them anytime online.

The success of the OTT platforms depends on two things- the variety of content and appropriate recommendations to the users. This business is profitable because users make a monthly payment to access the platform. Exploratory Data Analysis is done on the dataset to get insights from the information however the principal invalid qualities are taken care of. There are 12 features and around 7787 observations in the dataset and are mostly textual features. Clustering is a useful technique to achieve the best possible recommendations and increase the viewership of the platform

**1. Problem Statement**

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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, and rotten tomatoes can also provide many interesting findings.

**2. Introducti****on**

Netflix is a media distribution company and a prominent OTT platform with a wide variety of content to view from a variety of nations and genres. It started with DVD distribution via mail but has evolved substantially throughout its existence. Today, Netflix is focused on streaming video.

Netflix originally focused on movies, but television shows are probably the more common format today. Its works on a subscription model, where users get unlimited access to content with a paid subscription.

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

This project aims to analyze and perform clustering to determine patterns related to the content available on Netflix. Based on the attributes related to the Tv shows or movies, we will implement different clustering algorithms that come under the unsupervised Machine learning category.

**3. Dataset Description**

1. show\_id: Unique ID for every Movie / Tv Show
2. type: Identifier - A Movie or TV Show
3. title: Title of the Movie / Tv Show
4. director: Director of the Movie
5. cast: Actors involved in the movie/show
6. country: The country where the movie/show was produced
7. date\_added: Date it was added on Netflix
8. release\_year: Actual Releaseyear of the movie/show
9. rating: TV Rating of the movie/show
10. duration: Total Duration - in minutes or number of seasons
11. listed\_in : Genre
12. description: The Summary description

**4. Handling missing values**

There is a total of 3631 null values in the dataset where 2389 null values in the director column, 718 null values in the cast column,507 null values in a country column 10 in date\_added, and 7 in rating. so, we need to handle the null values.

Replacing the null values in the director, cast, and country columns. And 'date\_added' and 'rating' contains an insignificant portion of the data so we will drop them from the dataset.

**5. Exploratory Data Analysis**

* Exploratory Data Analysis (EDA) plays a vital role in the analysis of the data variables which are important from the aspect of feature engineering.
* After mounting our drive and fetching and reading the dataset given, we performed the Exploratory Data Analysis for it.
* To get an understanding of the data and how the content is distributed in the dataset, its type, and details such as which countries are watching more and which type of content is in demand etc. have been analyzed in this step.
* The United States is the most prolific generator of Netflix content, with India and the United Kingdom trailing far behind.

**6. Data Preprocessing:**

• Removing Punctuation:

Punctuations do not carry any

meaning in clustering, so removing

punctuations help to get rid of

unhelpful parts of the data, or noise.

• Removing stop-words:

Stop-words are basically a set of commonly used words in any language, not just in English. If we remove the words that are very commonly used in a given language,

we can focus on the important words instead.

• Stemming:

Stemming is the process of removing

a part of a word, or reducing a word to its stem or root. Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

**7) Clustering**

Clustering (also called cluster analysis) is a

The task of grouping similar instances into clusters. More formally, clustering is the task of grouping the population of unlabeled data points into clusters in a way that data points in the same cluster are more similar to each other than to data points in other clusters. The clustering task is probably the most important in unsupervised learning, since it has many applications, for example:

• data analysis: often a huge dataset contains several large clusters, analyzing which separately, you can come to interesting insights.

• anomaly detection: as we saw before, data points located in the regions of low density can be considered anomalies

• semi-supervised learning:

clustering approaches often help you to automatically label partially labeled data

for classification tasks.

• Indirectly clustering tasks (tasks where

clustering helps to gain good results): recommender systems, search engines,

etc.

• Direct clustering tasks: customer segmentation, image segmentation, etc.

**7.1) K- Means Clustering**

K-Means Clustering is an unsupervised learning algorithm that is used to solve clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

**7.1.1) K-means algorithm :**

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

“It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.”

It allows us to cluster the data into different groups and is a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

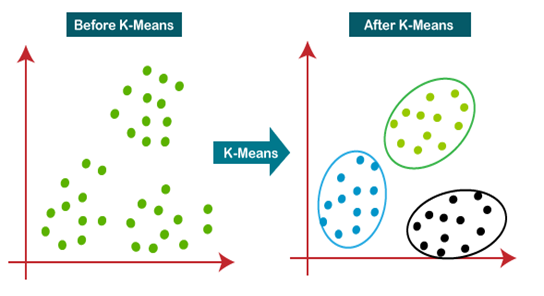
The k-means clustering algorithm mainly performs two tasks:

● Determines the best value for K center points or centroids by an iterative process.

● Assigns each data point to its closest k-center. Those data points which are near the particular k-center, create a cluster.

Hence each cluster has data points with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



**7.1.2) K-Means Algorithm Work:**

The working of the K-Means algorithm is explained in the below steps:

**Step 1:** Select the number K to decide the number of clusters**.**

**Step 2:** Select random K points or centroids. (It can be other than the input dataset).

**Step 3:** Assign each data point to its closest centroid, which will form the predefined K clusters**.**

**Step 4:** Calculate the variance and place a new centroid of each cluster.

**Step 5:** Repeat the third steps, which means reassigning each datapoint to the new closest centroid of each cluster.

**Step 6:** If any reassignment occurs, then go to step 4 else go to FINISH**.**

**Step 7:** The model is ready**.**

**7.1.3) Methods to find k value**

**A) Silhouette score:**

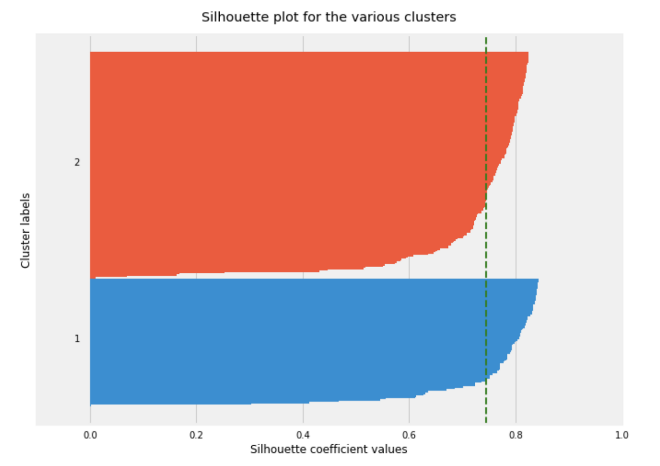
Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K Means in terms of how

Well, samples are clustered with other samples that are similar to each other.

Coefficients for a single sample are then

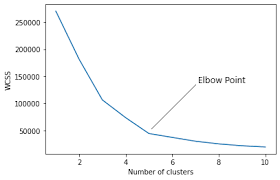
given as:

𝑠 = 𝑏 – 𝑎 /𝑚𝑎𝑥(𝑎, 𝑏)



**B) Elbow Curve:**

The Elbow Curve is one of the most popular methods to determine this optimal value of k.

The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.

**7.2)** **Dendrogram**

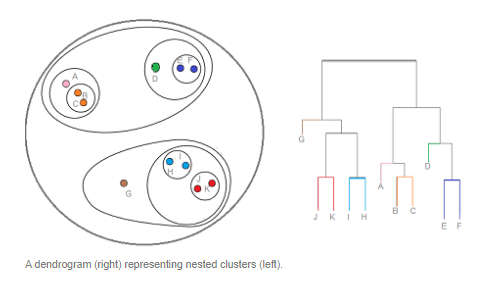
A dendrogram is a branching diagram that represents the relationships of similarity among a group of entities.

**7.2.1) Hierarchical Clustering:**

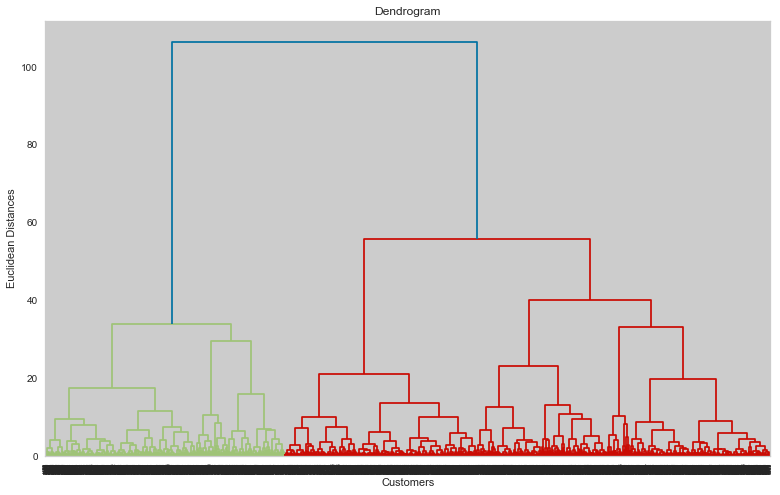
Hierarchical clustering is where you build a cluster tree (a dendrogram) to represent data, where each group (or “node”) links to two or more successor groups. The groups are nested and organized as a tree, which ideally ends up as a meaningful classification scheme.

Each node in the cluster tree contains a group of similar data; Nodes group on the graph next to other, similar nodes. Clusters at one level join with clusters at the next level up, using a degree of similarity; The process carries on until all nodes are in the tree, which gives a visual snapshot of the data contained in the whole set. The total number of clusters is not predetermined before you start the tree creation.

**Dendrogram plot:**



A dendrogram is a type of tree diagram showing hierarchical clustering relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data

****

**8) Conclusion:**

1) Data set contains 7787 rows and 12 columns in that cast and director features contain a large number of missing values so we can fill it and other features like ‘date\_added’ and ‘rating’ contain an insignificant portion of the data so we will drop them from the dataset.

2) We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies), there are more number movies on Netflix than TV shows.

3) TV-MA has the highest number of ratings for tv shows. i.e., adult ratings

4) The most number of movies and TV shows released in 2017 & 2018 or 2020 respectively.

5) United States has the highest number of content on Netflix, followed by India and India has the highest number of movies on Netflix.

6) The number of movies on Netflix is growing significantly faster than the number of TV shows. We saw a huge increase in the number of movies and television episodes after 2015 n our datasets.

7) Kids tv is the top TV show genre on Netflix.

8) Most of the movies have a duration of between 50 to 150 minutes long.

9) The most content is added to Netflix from October to January.

10) Documentaries are the top most genre on Netflix which is followed by standup comedy, Dramas, and international movies.

11) When it comes to movies having a TV-Y rating, they have the shortest runtime on average.

12) By applying the elbow and silhouette score, the optimal of 10 clusters formed, K Means is best for identification than Hierarchical as the evaluation metrics also indicate the same.