**NETFLIX MOVIES AND TV SHOWS CLUSTERING**

**1. Ms.Madhuri Bonela**

**2. Mr.Pratik Ghodke**

DataScience Traine Almabetter, Bangalore



# INTRODUCTION

Netflix is world's largest entertainment platform with many entertaining tv shows ,movies in it that are streaming across the world. it was founded on august 29 ,1997 in los gatos ,california by Mark and Reed .

Netflix had approximately 223 million paid subscribers worldwide as of the third quarter of 2022 and still its count is increasing to 2.5 million susubscribersfor the next quarter .Netflix is available in more than 190 countries with more than 100 million hours of tv shows and movies per day.

Netflix has many series ,documentaries,dramatic comedy shows and many feature films online entertaining every age group people around the world. It had approximately 223 million paid subscribers

Worldwide as of the third quarter of 2022 and still its count is increasing to 2.5 million susubscribersfor the next quarter.

Netflix uses data science for correct reccomendations to people by using clustering and classification techniques they trace out what kind of content is watched by most of their people in certain location and their age group and what kind of content mostly they like and by using reccomendation systems they also trace out and predict the future preference of people .

As Netflix Recommender recommends Netflix movies and TV shows based on a user's favorite movie or TV show. It uses a a K-Means Clustering model to make these recommendations. These models use information

About movies and TV shows such as their plot descriptions and genres to make suggestions.

**PROBLEM STATEMENT**

1.Exploratory Data Analysis

2.Understanding what type of content is availablein different countries

3.Is Netflix has increasingly focusing on Tv rather than movies in recent years

4.Clustering similar content by matching text -based features

Existing System

Whenever you access the Netflix service, the recommendations system strives to help you find a show or movie to enjoy with minimal effort. We estimate the likelihood that you will watch a particular title in our catalog based on a number of factors including:

your interactions with the service (such as your viewing history and how you rated other

titles),other members with similar tastes and preferences on our service, and information about the titles, such as their genre, categories, actors, release year, etc.

In addition to knowing what you have watched on Netflix, to best personalize and the recommendations also look at things like:

1.the time of day you watch,

2.the devices you are watching Netflix on

3.how long you watch.

All of these pieces of data are used as inputs that process in algorithms. (An algorithm is a process or set of rules followed in a problem solving operation.) The recommendations system does not include demographic information (such as age or gender) as part of the decision making process.

# INSPECTING DATASET

1. Importing the dataset and checking the head and tail rows to get an overall idea.
2. After that exploration of the dataset, by checking the info which gives us some intuition about null values and data types of columns present in it.
3. And then we checked the descriptive summary of the data frame which gives some quantitative idea about the dataset i.e., average values of the columns, frequency of the values, variability, and dispersion concerns on how to spread out the values.

# DATA CLEANING

Data Cleansing is the process of detecting and changing raw data by identifying incomplete, wrong, repeated, or irrelevant parts of the data. 

The null values in director is of about 30.68%.

The null values in cast is of about 9.22%.

The null values in country is about 6.51%.

The null values in date added is about 0.13%.

The null values in rating is about 0.09%.

The null values are dropped for the columns of

date added and rating .

# EXPLORATORY DATA ANALYSIS

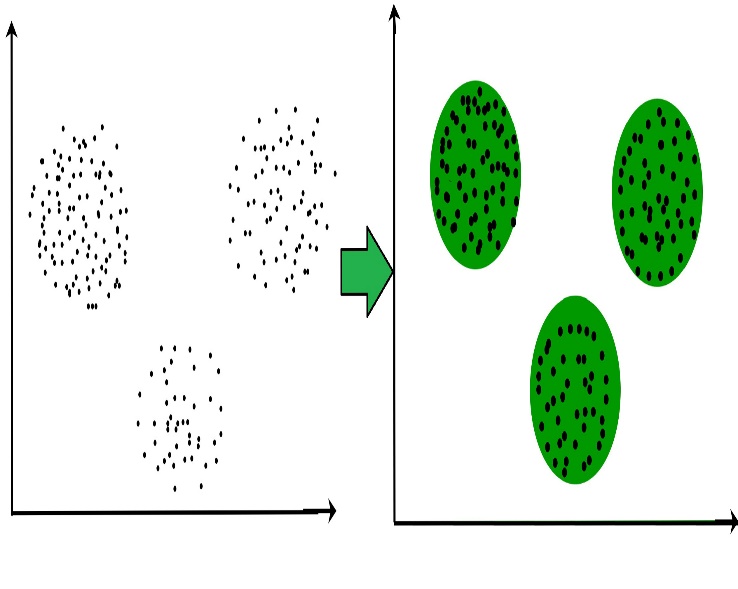
# NORMALISATION OF THE DATA

In machine learning, some feature values at times differ from others multiple times. The features with higher values will always dominate the learning process. Before giving our data to clustering algorithms we need to perform the data normalizationtask (i.e StandardScaler which will give equal importance to each variable so that no single variable drives the model performance.

# CLUSTERING

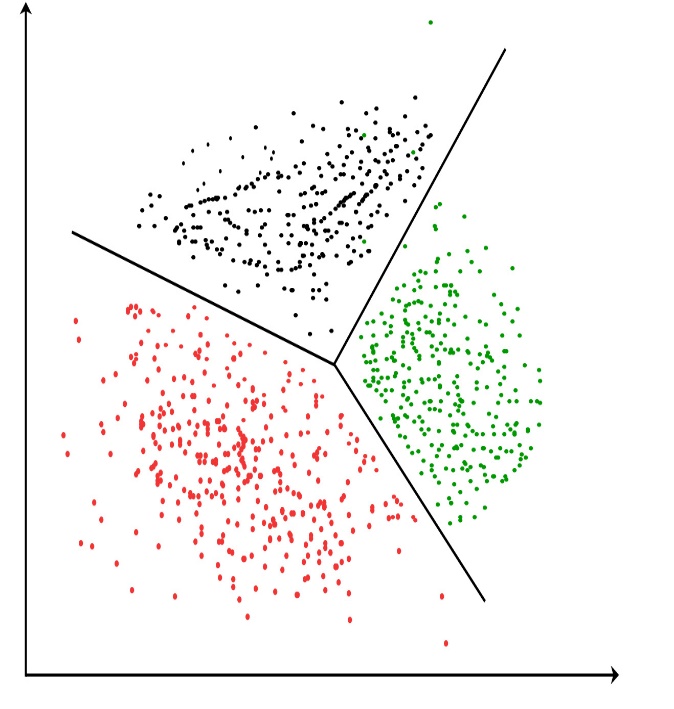
Customer segmentation has been demonstrated to benefit from clustering. Clustering is a sort of unsupervised learning that allows us to locate clusters in unlabelled datasets. Clustering techniques include Binning, Quantile based, K- means, hierarchical clustering, and DBSCAN clustering.

It is basically a type of [unsupervised learning method](https://www.geeksforgeeks.org/supervised-unsupervised-learning/). An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.



**The Elbow method**.

**1.**[**K-means clustering algorithm**](https://www.geeksforgeeks.org/k-means-clustering-introduction/) – It is the simplest unsupervised learning algorithm that solves clustering problem-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



**Applications of Clustering in different fields**

* **Marketing:** It can be used to characterize & discover customer segments for marketing purposes.
* **Biology:** It can be used for classification among different species of plants and animals.
* **Libraries:** It is used in clustering different books on the basis of topics and information.
* **Insurance:** It is used to acknowledge the customers, their policies and identifying the frauds.

# 2. The elbow method

# The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of *k*. As you know, if *k* increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as *k* increases. The value of *k* at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters

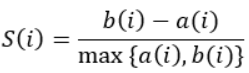
# https://www.oreilly.com/api/v2/epubs/9781788295758/files/assets/995b8b58-06f1-4884-a2a1-f3648428e947.png

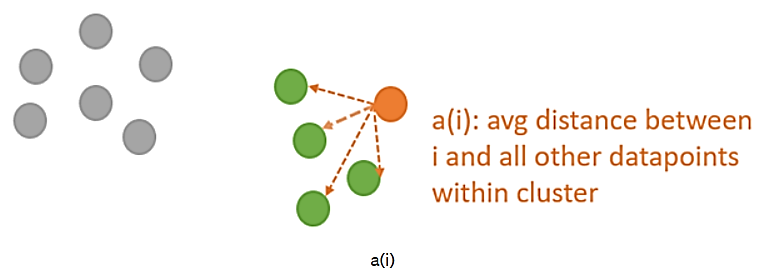
## 3. **Silhouette analysis**

The silhouette coefficient is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation).

* Select a range of values of k (say 1 to 10).
* Plot Silhouette coefﬁcient for each value of K.

The equation for calculating the silhouette coefﬁcient for a particular data point:



* (i) is the silhouette coefficient of the data point i.
* a(i) is the average distance between i and all the other data points in the cluster to which i belongs.
* b(i) is the average distance from i to all clusters to which i does not belong. 

We will then calculate the average\_silhouette for every k.

# Avg silhouette

# ****DBSCAN:** **Density-based Spatial Clustering of Applications with Noise****

# These data points are clustered by using the basic concept that the data point lies within the given constraint from the cluster center. Various distance methods and techniques are used for the calculation of the outliers.

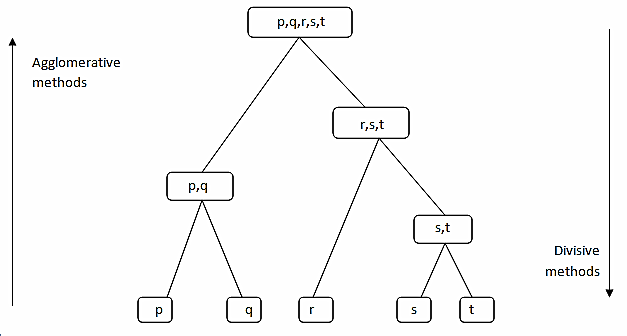
* **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)*, *OPTICS (Ordering Points to Identify Clustering Structure)*, etc.

# 

**Hierarchical Agglomerative Clustering Algorithm :-**

Hierarchical clustering algorithms group similar objects into groups called clusters. There are two types of hierarchical clustering algorithms:

* Agglomerative — Bottom up approach. Start with many small clusters and merge them together to create bigger clusters.
* Divisive — Top down approach. Start with a single cluster than break it up into smaller clusters.

****

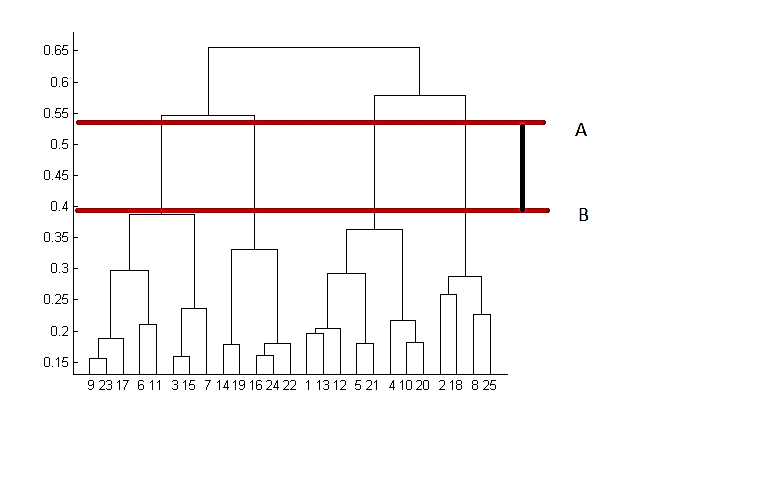
**Some pros and cons of Hierarchical Clustering**

**Pros**

* No assumption of a particular number of clusters (i.e. k-means)
* May correspond to meaningful taxonomies

**Cons**

* Once a decision is made to combine two clusters, it can’t be undone
* Too slow for large data sets, O(𝑛2 log(𝑛))



# CONCLUSION:-

# 1.Netflix has 69% of its content as movies, so we can say that movies are clearly more popular on Netflix than TV shows. Netflix has more movies than TV Shows

# 2. United States provides the most number of movies and shows followed by India and United Kingdom.

# 3. TV-MA rated content is maximum in number in the dataset. This rating indicates that the content is for mature and adult audience above the age of 18

# 4. There is an exponential raise in the number of TV shows and movies distributed by Netflix in the recent years.

# 5. Optimal number of clusters were found out to be 15 with silhouette coefficient value of 0.040

# 6.Recommendation based on cosine similiarity is also done on the same transformed data.

# 7.Indian market is one of the biggest market after USA so Netflix should more focus on Indian market