# Abstract:

The goal of the Netflix TV shows clustering project is to develop a machine learning model that can group TV shows on the platform based on their popularity and audience preferences. The project will involve several steps, including data collection, preprocessing, feature engineering, clustering, and evaluation. The dataset will include various features related to TV shows, such as their genre, rating, cast, production budget, and release year. Several clustering algorithms, including K-means, hierarchical clustering, and DBSCAN, will be applied to the dataset to group TV shows based on their similarities. The effectiveness of the clustering will be evaluated using metrics such as silhouette score and inertia. Ultimately, the project aims to help Netflix users discover new TV shows that align with their viewing preferences by providing them with curated groups of similar shows based on their attributes. By providing personalized recommendations to users, the project has the potential to improve user satisfaction and engagement with the platform.

# INTRODUCTION

The aim of the Netflix movies and TV shows clustering project is to develop a machine learning model that can group movies and TV shows based on their similarities using various features and attributes. These factors include genre, rating, cast, production budget, release year, and other features that affect audience preferences and popularity. The model will use unsupervised learning algorithms such as K-means, hierarchical clustering, or DBSCAN to group movies and TV shows into clusters based on their similarities.

The clusters will be labeled based on their characteristics, such as "action movies," "romantic comedies," or "sci-fi TV shows," allowing users to easily find content that aligns with their preferences. Like the personalized recommendation algorithm, the Netflix movies and TV shows clustering model will be dynamic, adjusting the clusters based on the latest viewer trends and feedback. The output of the model will be a set of clusters, each containing movies and TV shows that share similar characteristics, which could be communicated to the user.

The model will be trained on a dataset containing various movies and TV shows and their corresponding features and attributes. This dataset will be used to train the model using unsupervised learning algorithms to group similar movies and TV shows into clusters. Once the model is trained and validated, it can be deployed as an application or integrated into the Netflix platform, allowing users to easily discover and explore new content that aligns with their viewing preferences.

This project has the potential to increase viewer satisfaction and engagement, leading to higher retention rates for the platform. By providing users with curated groups of similar movies and TV shows based on their attributes, they will be more likely to find content that interests them, ultimately leading to a better viewing experience.

# PROBLEM STATEMENT

The dataset under consideration is a comprehensive collection of TV shows and movies available on the popular streaming service Netflix, as of 2019. This dataset was sourced from Almabetter School. In recent years, Netflix has emerged as a key player in the entertainment industry, and this dataset reflects the company's focus on original programming. According to a report released by the company in 2018, the number of TV shows on the platform had nearly tripled since 2010, while the number of movies had decreased by over 2,000 titles.

The dataset contains a vast array of variables, including information on the title, director, cast, country, rating, genre, and more. By analyzing this data, it is possible to gain insights into trends in the types of content that are popular on the platform, as well as explore the characteristics of highly rated movies and TV shows. This information can be invaluable for content creators, producers, and marketers looking to understand consumer preferences and improve their offerings.

Moreover, the dataset can be used to build predictive models that can help anticipate the success of future content on the platform. With this data, it is possible to gain a better understanding of the type of content that resonates with viewers, and use this knowledge to develop programming that is more likely to attract and retain audiences.

Overall, this dataset provides a unique opportunity to explore and analyze the rapidly-evolving entertainment industry, and can be used to gain valuable insights into the preferences and behaviors of viewers.

### Static vs Dynamic Pricing: Choosing the Right Strategy for Your Industry

In the context of Netflix movies and TV show clustering, companies may use different pricing strategies to set the prices for their streaming services. Static pricing is a fixed pricing scheme that remains the same regardless of changes in demand or supply. This pricing strategy is often used in traditional media settings where the cost of content production and distribution is relatively stable. Dynamic pricing, on the other hand, is a flexible pricing strategy that adjusts prices in real-time based on changes in demand and supply. This strategy is becoming increasingly popular in the streaming industry as companies seek to optimize revenue and meet the needs of customers in real-time. Similar to surge pricing in the taxi industry, dynamic pricing in the streaming industry adjusts prices based on fluctuations in demand, supply, and other external factors such as promotions, competitors' pricing, and consumer behavior. By leveraging data and analytics to track market trends, companies can make strategic pricing decisions and optimize revenue. Overall, understanding the pros and cons of static and dynamic pricing can help companies develop effective pricing strategies and stay competitive in a rapidly evolving market for streaming services.

## Price Drivers

In the context of clustering Netflix content, several factors can impact how movies and TV shows are categorized and recommended to users. The first factor is genre, which can have a significant influence on how content is categorized and recommended. The genre of a movie or TV show, such as action, comedy, drama, or sci-fi, can help to group similar content together and recommend it to users who have shown a preference for that genre.

Another factor that can impact content clustering is user ratings and reviews. High user ratings and positive reviews may result in content being recommended more frequently to users. Viewer demographics, such as age, gender, and viewing history, can also play a role in how content is categorized and recommended. For example, if a user frequently watches romantic comedies, they may be recommended more content in that genre.

Popularity is another factor that can influence content clustering and recommendations. Popular content with high viewership may be recommended more frequently to users, even if it does not align with their viewing history or preferences. Finally, release date can also impact how content is categorized and recommended. New releases may be promoted more heavily to users to generate interest in the latest content.

By understanding these factors, Netflix can develop more effective content clustering and recommendation strategies that improve the user experience and increase user engagement. Effective clustering and recommendations can help users discover new content they may enjoy and ultimately lead to increased viewer satisfaction and retention rates for the platform.

## Dynamic Pricing in Netflix Movie and TV Show Clustering

Dynamic pricing plays a significant role in Netflix movie and TV show clustering by optimizing revenue and meeting customer needs. The company uses personalized pricing strategies to tailor prices to individual users based on their past viewing habits, preferences, and behavior. For instance, a user who frequently watches action movies may be shown a higher price point for an action movie than a user who rarely watches that genre. This personalized pricing strategy helps Netflix optimize revenue and offer tailored recommendations to users.

Netflix also uses dynamic pricing to adjust prices based on external factors such as market demand, competitor pricing, and seasonality. During the holiday season, prices for popular movies and TV shows may increase due to higher demand. Similarly, if a competitor lowers their prices, Netflix may choose to adjust their prices in response to remain competitive. By leveraging data and analytics to track market trends and consumer behavior, Netflix can make strategic pricing decisions and optimize revenue.

The use of dynamic pricing allows Netflix to better serve the needs of their customers by offering personalized pricing and tailored recommendations while still remaining competitive in a rapidly evolving market. With personalized pricing and dynamic adjustments, Netflix can keep their pricing strategies flexible and dynamic, ensuring that their revenue remains optimized and their customers remain satisfied.

# Clustering

## K-Means Clustering:

K-means clustering and hierarchical clustering are two popular unsupervised machine learning algorithms used for grouping similar data points into clusters. K-means clustering partitions the dataset into k number of clusters, where k is predefined by the user. It starts by randomly selecting k initial centroids and then assigns each data point to the nearest centroid. The algorithm then recalculates the centroids by taking the mean of all the data points in each cluster, and this process continues until the centroids no longer change or a specified number of iterations have been reached.

On the other hand, hierarchical clustering does not require the user to define the number of clusters beforehand. It starts with each data point as its own cluster and then recursively merges the two closest clusters until only one cluster remains. This process results in a hierarchical tree-like structure known as a dendrogram, where each leaf node represents an individual data point, and the internal nodes represent clusters. The user can choose the number of clusters by selecting a cut-off point on the dendrogram.

Both algorithms have their strengths and weaknesses. K-means clustering is faster and more scalable than hierarchical clustering, making it suitable for large datasets. However, it requires the user to specify the number of clusters beforehand and is sensitive to the initial centroids' selection. Hierarchical clustering, on the other hand, does not require the user to specify the number of clusters beforehand and produces a dendrogram that can provide insight into the data's underlying structure. However, it can be computationally expensive and less scalable than k-means clustering.

Overall, both algorithms are useful tools for clustering similar data points and can provide valuable insights into the data's underlying structure. Choosing the right algorithm depends on the specific needs and characteristics of the dataset and the desired outcome of the analysis.

# CONCLUSIONS

Welcome to our exciting journey into the world of Netflix shows! Our objective was to cluster the shows into groups based on their similarities and differences and ultimately create a content-based recommender system that suggests 10 shows based on the user's viewing history.

With a dataset comprising over 7787 records and 11 attributes, we began our adventure by exploring the dataset's missing values and performing exploratory data analysis (EDA). Our analysis revealed that Netflix offers more movies than TV shows, with a rapidly growing collection of shows from the United States. To cluster the shows, we focused on six key attributes: director, cast, country, genre, rating, and description. We transformed these attributes into a 10000-feature TFIDF vectorization and used Principal Component Analysis (PCA) to tackle the curse of dimensionality. By reducing the components to 3000, we were able to capture more than 80% of the variance.

Next, we utilized two clustering algorithms, K-Means and Agglomerative clustering, to group the shows. K-Means determined that the optimal number of clusters was 5, as confirmed by the elbow method and Silhouette score analysis. Meanwhile, Agglomerative clustering suggested 7 clusters, which we visualized using a dendrogram.

But our exploration didn't stop there. We created a content-based recommender system using the similarity matrix obtained through cosine similarity. This system offers personalized recommendations based on the type of show the user has watched, providing them with 10 top-notch suggestions to explore.

Join us in discovering the diverse world of Netflix shows, and let our recommender system guide you to your next binge-worthy obsession.