ML Classification Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1. **Abhishek Mishra (Abhishekmishra@gmail.com):**

Data Wrangling

- Missing Values/Null Values
- Null value treatment
- Duplicate Values

Exploratory Data Analysis (EDA)

- Analysis 'type' column
- Analysis 'director' column
- Analysis 'cast' column
- Analysis 'country' column

Feature Engineering & Data Pre-processing

- Feature Engineering
- Text cleaning
- Applying PCA to reduce dimensions

Cluster Model Implementation

- Silhouette Score Elbow for KMeans Clustering
- Dendogram
- Agglomerative Clustering
- KMeans Clustering

Evaluation Metrics

- Silhouette Score

Movie Recommendation System

- Code for Movie Recommendation System
- Movie Recommendation System

2. **Kurva Mallesh (kurvamallesh36@gmail.com):**

Data Wrangling

- Missing Values/Null Values
- Null value treatment
- Duplicate Values

Exploratory Data Analysis (EDA)

- Analysis 'date_added' column
- Analysis 'release_year' column
- Analysis 'rating' column
- Analysis 'duration' column

Feature Engineering & Data Pre-processing

- Using TF-IDF
- Cumulative Explained Variance
- Stemming

Cluster Model Implementation

- Silhouette Score Elbow for KMeans Clustering
- Dendogram
- Agglomerative Clustering
- KMeans Clustering

Evaluation Metrics

- Silhouette Score

Movie Recommendation System

- Code for Movie Recommendation System
- Movie Recommendation System

4. **Arunesh Mishra (Arunesh12mishra@gmail.com):**

Exploratory Data Analysis (EDA)

- Analysis 'listed_in'(Genera) column
- Analysis on 'title' column
- Analysis on 'description' column
- Analysis 'release_year' column

Feature Engineering & Data Pre-processing

- Using TF-IDF
- Cumulative Explained Variance
- Stemming

Cluster Model Implementation

- Silhouette Score Elbow for KMeans Clustering
- Dendogram
- Agglomerative Clustering
- KMeans Clustering

Evaluation Metrics

- Silhouette Score

Movie Recommendation System

- Code for Movie Recommendation System
- Movie Recommendation System

GithubRepository link:

Abhishek Mishra- https://github.com/abhishekmishra-bareilly/ML-unsupervised-Capistone-project

Kurva Mallesh - https://github.com/kurvamallesh/ML-unsupervised-Capistone-project

Arunesh Mishra - https://github.com/kajuun/Unsupervised-ML-capstone-project-#unsupervised-ml-capstone-project-

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Fixable 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. According to the graph we have 5377(69.14%) movies. And 2400(30.86%) as TV Show in this dataset. According to plot we can say Raul Campos and Jan Sulter collectively have the most content on Netflix. Marcus Raboy have the second most content on Netflix. Now we can say in this data Anupam Kher having 38 number of listing.takahiro Sakurai is the second most listed actor on netflix.Shah Rukh Khan is the 3rd most listed actor on Netflix. According to the plot we can understand United States have 2080 Movies and 975 TV Show. INDIA have second most listed country with 852 movies and 71 TV Show on Netflix.
We have so many content release in October (785), November (738), December (833) and January (757) may be it is because of Holiday season. The number of release has significantly increased after 2015 to 2020.
But sudden drop in 2021 may be it is because of covid 19. We have 744 movies and 268 TV Show release in 2017Also 734 movies and 386 TV Show release in 2018.82%(6431) of the content was released between 2010 and 2021.17.28%(1346) of the content was released before 2010.Most number of movies rated TV-MA i.e. Adult Rating. Most number of TV Shows rated TV-MA i.e. Adult Rating. We have most listed duration as season 1 with 1608 listing. We have second most listed duration as season 2 with 378 listing. Mainly the movie duration is in b/w 55 to 150 minutes. Most of the movies list for 90 to 120 minutes. In Movies Documentaries is the most popular genera on Netflix.
Comedy is the second most popular genera on netflix.In TV Shows Drama is the most popular genera. International TV shows is the second most popular genere.Most repeated words in title column are love, Christmas, World, Man, and life. Most repeated words in the description of the TV shows and movies are Family, new, Love, Life, mother, find.
Data pre-processing

For train the model we use description column, listed_in column, rating column, country column, title column, director column, cast column. We performed Text cleaning as our next step. Convert all words in lowercase. We performed Stemming as our next step. We remove all stopwords.Also use stemming function. We performed TF-IDF vectorizer.Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. Applying PCA-Principal Component Analysis to reduce dimensions. We will use 3000 components.

Applying models

WE use Elbow method for finding k values. Also use Silhouette Score for best score. Also use Dendogram for finding the value of clusters. Use Agglomerative Clustering. Use KMeans Clustering. here are few clusters with there word cloud graph.