Capstone Project Submission

Instructions:
i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Team members:

Member 1: Saaquib Mustafa
Email: *saaquibmustafa26@gmail.com*
Contributions:
1. Clean and prepare the data for analysis.
2. Added Useful Codes to simplify the analysis.
3. Done Initial analysis and helped in visualization.
4. Prepared Project Summary
5. Prepared Key Notes, conclusion and PPT

Member 1: Sahil Kolambkar
Email: *sahilrajkolambkar@gmail.com*
Contributions:
1. Prepared Project Presentation
2. Helped in Data Cleaning
3. Help in Summary Preparation & Technical Documentation
4. Helped in Key Notes and Conclusion

Please paste the GitHub Repo link.

Github Link:- **https://github.com/Saaquib01/Capstone-Machine-Learning-Unsupervised-Netflix-and-Tv-Show-Clustering-**

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

PROBLEM:

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from fixable, a third-party Netflix search engine. In 2018, they released an interesting report showing that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has dropped by more than 2,000 titles since 2010, while the number of TV shows has nearly tripled. It will be interesting to explore what other insights can be gained from the same data set. By integrating this dataset with other external datasets such as IMDB ratings, Rotten Tomatoes can also provide many interesting findings.

APPROACH:

Initially, in the 1st step imported the data set to carry out the analysis over the data set to comprehend the details of available data and Checked for Null values and treated them. Here, we

found more than 30% null values in the director's column. Then, we take appropriate action for null values according to the circumstances.

Performed the Exploratory data analysis and tried to get the 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. has been analyzed in this step with the help of visualization graph by getting insights from analysis.

- Data preprocessing – in this we remove the punctuation and stops words also used stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.
- We used the k-means clustering algorithm and then checked the model performance using silhouette's coefficient and elbow method to find the number of clusters.

Analyzing all the variables of the data set and identifying the solution for given tasks. Performed hypothesis testing to get the insights on duration of movies and content with respect to different variables. After doing feature engineering and finding the number of clusters, we used the k-means algorithm and then checked the model performance using Silhouette's coefficient, to identify the best fit Model. The number of movies on Netflix is growing significantly faster than the number of TV shows. Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019.

CONCLUSION:

The ratio of Tv show and movie is 31% to 69% in most watched-on Netflix
Here, top 20 Directors with the greatest number of Movies/Shows on Netflix. Highest is Raul Campos and lowest is Justin G. Dyck
The most of movies acted by any actor is Anupam kher followed by Takahiro and Shahrukh khan
Highest number in production, Tv shows and movies is in USA country followed by India.
Most movies/tv show got released in 2018 and least 2006
More content with Mature content is available on Netflix.

Cluster value we got from knee elbow method is **9**

- Cluster 0: Dramas & International TV Shows
- Cluster 1: Sc-Fi, Adventure and Action
- Cluster 2: International Movies
- Cluster 3: Comedy
- Cluster 4: Romantic Movies
- Cluster 5: Stand-Up Comedy
- Cluster 6: Documentary
- Cluster 7: International Movies
- Cluster 8: Children and Family movies