Capstone Project Submission Summary

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1.Harshit Kumar Email: harshitgangwar427@gmail.com Contribution: 1. Data Cleaning 2.Technical Document 3.Exploratory Data Analysis 4.PPT **2. Harshit Kumar** Email:chauhanh8439@gmail.com Contribution: 1.Feature Engineering 2. ML Modelling 3. Clustering 4. Summary
Please paste the GitHub Repo link.
GitHubLink – https://github.com/HarshitKumar-git/Netflix-Movies-And-TV-Shows-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)

Project Title: NETFLIX_MOVIES_AND_TV_SHOWS_CLUSTERING

Netflix, Inc. is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a film and television series library through distribution deals as well as its own productions, known as Netflix Originals.

Netflix can be accessed via internet browser on computers, or via application software installed on smart TVs, set-top boxes connected to televisions, tablet computers, smartphones, digital media players, Blu-ray Disc players, video game consoles and virtual reality headsets on the list of Netflix-compatible devices. It is available in 4K resolution.

Netflix has become dominant company in the on-demand media industry, with 167 million paying subscribers around the world. By creating compelling original programming, analysing its user data to serve subscribers better, and above all by letting people consume content in the ways they prefer, Netflix disrupted the television industry and forced cable companies to change the way they do business.

Our main objectives of this project are to do exploratory analysis and find useful insights from dataset, to understand what type content is available in different countries, also to find out is Netflix has increasingly focused on TV rather than movies in recent years and at last to do clustering of similar content by matching text-based features from dataset.

We have been provided a dataset collected from Flixable 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 data set.

We started this project with the intention to obtain some useful insights related to the type of Netflix content. For this, we performed exploratory data analysis on our data after cleaning and making it easy to analyse. This analysis helped us to understand the trend. We found that most of the content on Netflix are of TV-MA and TV-14 rating. USA and India are two countries producing the maximum number of contents. Documentaries and stand up are top genre in terms of number of contents they have on platform. Further we found number of movies on Netflix outnumbers TV-shows.

Our next job was to make an unsupervised clustering model. For this, we processed our text by removing unuseful characters like - stop words, punctuation and did stemming. After getting the length for each text feature we rescaled them for generalisation and started applying algorithms. We first used K-means clustering. In order to find appropriate cluster number, we used elbow method and finally got the best silhouette score of around 0.35. Next, we applied Hierarchal Agglomerative Clustering for which we made dendrogram. We also obtained silhouette score of around 0.32. With this we achieved our objectives of the project.