CAPSTONE PROJECT SUBMISSION

INSTRUCTIONS

I. Please fill in all the required information.
II. Avoid grammatical errors.

TEAM MEMBER'S NAME, EMAIL AND CONTRIBUTION:

TEAM - DATA DEFENDERS

S No	Member Name	Email	Contribution
1.	Lokesh Tokas	lokesh.you@gmail.com	Summary
2.	Saraswat Mukherjee	mae21saraswat@gmail.com	Technical documentation
3.	Charan C S	Ccharancs543@gmail.com	Colab Notebook
4.	Shubham Sartape	shubhamns19.pumba@gmail.com	Presentation

GITHUB REPO LINK.

GitHub: https://github.com/shubhamsartape/Netflix-Movies-and-TV-Shows-Clustering

PROJECT SUMMARY

The given dataset is made up of 7787 entries in 12 columns consisting 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 was interesting to explore what all other insights we obtained from the same dataset.

For EDA we imported all the necessary libraries like numpy, pandas, sklearn, seaborn, matplotlib etc.

Next step we fetched the dataset and did a basic analysis of it by checking its data type, features and some basic statistics about them. Then we checked the null values for each feature of the dataset. With this we came to know that there 5 features in the dataset which contain null values which needed to be rectified. We observed that the data added and rating column has only single null, so we will dropped the null values. The null values in the 'directors' column was replaced by unknown, 'country' column by its mode value and the null values were replaced in 'cast' column as 'no cast'. After examining the dataset, 'show id' column seemed quite unimportant so dropping was a logical decision. That's how we made data null values free which subsequently made the dataset much cleaner and crisper.

The data we had also needed to be cleaned which means it should be free of the punctuation marks like comma, semi colon etc. While working with the data those punctuation marks were a source of concern. Similarly we also removed stopped words followed by stemming, lemmatizing and tokenizing the data. This made the dataset and the data in it, more prominent and informative. We created functions of all these activities respectively and called them one by one to get our required work done. This provided us the appropriate data as per our need and made our rest of the work comparatively easy and simple. We also got the frequency of the words which helped us later in this project. Then we applied Principal Component Analysis on our data to reduce its dimensionality. Out of our selected 2000 components, after dimensionality reduction 1500 components covered 90% of the data.

We ran K-Means clustering Elbow method and found the number of optimal clusters *viz.* 10 clusters.

To measure the quality of clusters and how well the clusters are separated, silhouette score was considered.

In K-Means clustering the silhouette score was highest for 3 clusters with a score of 0.510.

We also ran hierarchical clustering and through dendrogram plot we observed optimal clusters and the number of clusters came out to be 6.

The silhouette score for hierarchical cluster is highest for cluster 3 and the silhouette score is 0.505.

By applying K-Means and hierarchical we got the best clusters and optimal cluster equal to 3.