Capstone Project Submission

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

Team Member's Name, Email and Contribution:

Name: Kabeer Pande

Email: kabeerpande7075@gmail.com

Contribution:
Colab notebook
Project summary
Technical documentation
Project presentation
Presentation video

Please paste the GitHub Repo link.

Github Link:- https://github.com/kabeerrrh/Netflix-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)

Netflix is the most widely used media and video streaming platform. It has around 200 million users World Wide. It offers a large library of movies and TV series that can be accessed at any time via internet services. The data set is given to us in the CSV format in which our main aim is to Cluster the similar content by matching text-based features.

This dataset has 7787 rows and 12 columns. After exploring data we found that there were A lot of null values present in our dataset. The features that contained null values are cast, crew, country and date added . We dropped the rows having null values from the date added column but we didn't drop the null values which were corresponding to cast, crew and country columns because it can impact the data as the null values were huge comparatively to our original dataset. Therefore we swapped those Nan values to unavailable.
After cleaning the data we perform EDA and we got many interesting insights of the data like the content distribution in Netflix ,major countries who produces most number of content and top directors and actors who have worked in most number of Movies and TV-Shows etc.

Then we performed feature Engineering and combined all the important text based features in one column and performed stemming and vectorization in the column. Then we applied PCA to reduce the dimensions.
Finally we implemented K-means Clustering and Agglomerative Clustering , the optimal number of clusters were find out to be 3 through Elbow curve , Silhouette score and Dendrogram.

Then we created a recommender system that will help the customers by similar content that he should watch based on their previous watch list.