Capstone Project Submission

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

Team Member's Name, Email and Contribution:

Team Member's Role: -

- Sahil Pardeshi(8623879021.sp@gmail.com)

Contribution:

- o Data understanding
- o Data visualization
- o Feature engineering
- o TF-IDF vectorizer
- o Elbow method

- Pravin Bejjo(praveen.bejo.pb@gmail.com)

Contribution:

- o Data understanding
- o Data visualization
- o Removing punctuation and stop words
- o Dendogram
- o K-means clustering
- o Visualize silhouette score and clusters

- Kirtesh Verma(kirteshverma12345@gmail.com)

Contribution:

- o Data understanding
- o Handling null & missing values
- o Performing EDA
- o Data preprocessing
- o Silhouette score

Please paste the GitHub Repo link.

GitHub Link :https://github.com/Sahilpardeshi1/Netflix-movies-and-Tv-shows

Google Drive Link: https://drive.google.com/drive/u/0/folders/1vt24QSY16917_92tUrIH-38UtPIDwq_C

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 a subscription-based streaming service that allows our members to watch TV shows and movies without commercials on an internet-connected device. You can also download TV shows and movies to your iOS, Android, or Windows 10 device and watch without an internet connection.
The dataset is collected from Flexible which is a third-party Netflix search engine. It has 7787 entries and 12 attributes.

Our objective is to do text-based clustering on the given dataset so that we can cluster similar content on Netflix. For this we began by performing Exploratory Data Analysis(EDA) on the given data thus drawing meaningful insights which helped us to understand the type of content available on Netflix, the targeted audience, top genres etc.

We have used featured engineering columns for better intuition and followed data preprocessing steps to make data read for model building. We converted the data type of certain columns. We have used it by handling null values. Maximum null values were present in the ‘director’ and ‘cast’ column, as this are not important for model building, we dropped them.

For the columns of text based we have removed stop words and punctuation, performing stemming and calculated TF - IDF. We performed steps on ‘descriptions’ and ‘listed_in’ columns. We have used various method like data understanding, feature engineering , TF - IDF vectorizer, Elbow method, Handling null values and missing values , performing EDA, Data preprocessing , Silhouette score, Dendogram , K - mean clustering as this is an unsupervised Machine learning problems, visualize Silhouette score and clustering.

We have used the elbow method to derive the most suitable number of clusters giving us the acceptable error term with less computational cost. The number of clusters for k-means clustering turned out to be 5.To measure the goodness of clusters we used Silhouette score method. The score for 3 clusters is 0.34 which is good.

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. In the United States, the company provides DVD and Blu-ray rentals delivered individually via the United States Postal Service from regional warehouses