Capstone Project Submission

Instructions:

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

Name:- Tarun
Email:- jangratarun1020@gmail.com
Contribution:- Individual

Github Link:- **https://github.com/tarun422/Netflix-DataSet-Clustring**

The Netflix Movies and TV Show Clustring dataset consist of 7787 observation with 12 features and it is a unsupervised machine learning problem because it does not contain any target column and the main goal is to clustering the dataset into different clusters.

After loading the dataset, first performed data preprocessing and checking data types, missing values, duplicate values and data description. In this dataset there are some missing values . Director colum contain more then 30% of missing values which is not beneficial for the clustering so we drop the director column. After that there are some columns contain missing values which are handled by normal missing values handling technique.

After that feature engineering comes in to add and drop some column which is helpful for this dataset

After that Exploratory Data Analysis is performed to obtain the insights of our Netflix dataset. Various graphs are constructed to comparing columns with other columns. It contain Univariat Analysis, Bivariate Analysis and Multivariate Analysis . Dataset is divided into two parts, first one is Movies and the other one is TV Shows

After that Data preprocessing comes in to pre-process our data. it contains removing special character, removing stopwords and stemming

After that modeling part begins and applying the silhouette score method for n range clusters on dataset and getting the best score which is 0.356 for cluster = 3, it means content explained well on their own clusters. Speaking about other different cluster methods,K-Mean,hierarchical,agglomerative clustering on data, we got the best cluster arrangements and 3 is the best cluster for this dataset.