**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email, and Contribution:** |
| 1. Name: - Shubham Narendra Kadu 2. Email - [shubhamkadu.ds@gmail.com](mailto:shubhamkadu.ds@gmail.com) 3. Contribution **–**  * Contributed to this project that first I self-done. * Data processing * Data Cleaning * Data Analysis * EDA * Data Preprocessing * Machine Learning * Clustering * K-means clustering(Elbow Method, Silhouette Score) * Dendrogram * Agglomerative hierarchical Clustering * Conclusion * PPT |
| **Please paste the GitHub Repo link.** |
| GitHub Link:- https://github.com/shubhsk98/Netflix-Movies-and-TV-Shows-Clustering.git |
| **Please write a summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)** |
| **Problem Statement:** This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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 dataset. Integrating this dataset with other external datasets such as IMDB ratings, and rotten tomatoes can also provide many interesting findings.  **Approach:**  • Initially, in the 1st step import the data set to carry out the analysis of the data set to comprehend and handle the missing values and duplicate values.  • Performed the Exploratory data analysis and tried to get an understanding of the data and how the content is distributed in the dataset, its type, and details such as which countries are watching more and which type of content is in demand, etc. have been analyzed in this step with the help of visualization graph by getting insights from analysis.  • Data preprocessing – in this we remove the punctuation and stop words and also used stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.  • We used the k-means clustering algorithm and then checked the model performance using Silhouette’s coefficient and elbow method to find the number of clusters  **Conclusion:**  1) Data set contains 7787 rows and 12 columns in that cast and director features contain a large number of missing values so we can fill it and other features like ‘date\_added’ and ‘rating’ contain an insignificant portion of the data so we will drop them from the dataset.  2) We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies), there are more number movies on Netflix than TV shows.  3) TV-MA has the highest number of ratings for tv shows, i.e. adult ratings  4) The most number of movies and TV shows released in 2017 & 2018 or 2020 respectively.  5) United States has the highest number of content on Netflix, followed by India and India has the highest number of movies on Netflix.  6) The number of movies on Netflix is growing significantly faster than the number of TV shows. We saw a huge increase in the number of movies and television episodes after 2015 n our datasets.  7) Kids tv is the top TV show genre on Netflix.  8) Most of the movies have a duration of between 50 to 150 minutes long.  9) The most content is added to Netflix from October to January.  10) Documentaries are the top most genre on Netflix which is followed by standup comedy, Dramas, and international movies.  11) By applying the elbow and silhouette score, the optimal of 10 clusters formed, K Means is best for identification than Hierarchical as the evaluation metrics also indicate the same. |