**Capstone Project-4 Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Contributor’s Role:**  **Praful Mahendra Gedam**  **Email: - prafulgedam12@gmail.com**   * **Introduction** * **Data cleaning** * **Exploratory Data Analysis** * **Data cleaning & pre-processing for clustering’s** * **Encoding the categorical data** * **K means clustering** * **Hierarchical Clustering** * **Silhouette analysis** * **Conclusion** |
| **Please paste the GitHub Repo link.** |
| GitHub Link :- <https://github.com/PrafulGedam/Netflix-movies-shows-and-tv-shows-clustering>  Drive Link: -  <https://drive.google.com/drive/folders/1KRYp5tRA02ZvrTchEgZ4AvfCcnhZM5y6?usp=share_link> |
| **Please write a short 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 flixable 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, rotten tomatoes can also provide many interesting findings.  **APPROACH: -**  Initially, in the 1st step imported the data set to carry out the analysis over the data set to comprehend the details of available data and Checked for Null values and treated them. Here, we found more than 30% null values in the director's column. Then, we take appropriate action for null values according to the circumstances.  Performed the Exploratory data analysis and tried to get the 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. has been analyzed in this step with the help of visualization graph by getting insights from analysis.   * Data Exploration – in this we remove the punctuation and stops words 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. * The project's main goal is to create a model that can perform Clustering on comparable material by   matching text-based attributes.   * As the problem statement says, understanding what type of content is available in different countries   and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering  on similar content by matching text-based features. For that we used Affinity Propagation, Agglomerative Clustering, and K-means Clustering.  **For n\_clusters = 2, silhouette score is 0.3394022069127177**  **For n\_clusters = 3, silhouette score is 0.31090374177589386**  **For n\_clusters = 4, silhouette score is 0.29525083370730026**  **For n\_clusters = 5, silhouette score is 0.3004793900229635**  **For n\_clusters = 6, silhouette score is 0.3053700959079982**  **For n\_clusters = 7, silhouette score is 0.316369622686931**  **For n\_clusters = 8, silhouette score is 0.32233548702488246**   * The silhouette score for k means was in the range of 0.35 while in case of HAC we are getting close   **CONCLUSION: -**   * We had lot of operations over the dataset to find out some very useful information from it. we have to conclude dataset in few lines, then we can say that * Netflix has more Movies * Most of the number of Movies and TV shows Produced by united Sates fallowed by India * Most of the content of the Netflix is foe nature Audience * Majority of the movies range in length from 90 to 100 minutes and the majority of TV shows have only one seasons * 2018 is the year in which Netflix released a lot more content as compared to other years * Dramas international movies are the best popular genres on Netflix. * The silhouette score for k means was in the range of 0.35 while in case of HAC we are getting close |