**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. Harsh Durugkar   Email :- [durugkarharsh@gmail.com](mailto:durugkarharsh@gmail.com)  Contribution on :- 1) Data Importing   * 1. ) Uploading on Drive   2. ) Give access to google colab   2 ) Find Missing value  3 ) Fill missing value  4 ) Check if every value is filled or not.  5 ) Start EDA :-  6 ) Visualise Total Release Movies/Tv shows in Last 10 years  7 ) Visualise Types of Video on Netflix  8 ) Visualise the top 10 Countries that produced Highest Number of  Movies/Shows on Netflix.  9 ) Visualize Top 5 Rating Distribution for Movies and Shows on Netflix.   1. Visualise the top Genres For Movies/TV-Shows on Netflix 2. Visualise the Top Directores on Netflix 3. Visualise the Top cast on Netflix Till Year of 2020 4. Top Duration of Movies on Netflix 5. Highest Duration of TV Shows on Netflix 6. What type content is available in different countries 7. Is Netflix has increasingly focusing on TV rather than movies in recent   years ?   1. Text Processing (Machine Learning) 2. Machine Learning Clustering 3. K-means Clustering 4. Hierarchical Clutering 5. Conclusion |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/harshdurugkar/NETFLIX-MOVIES-AND-TV-SHOWS-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)** |
| Hi I’m Harsh Durugkar. I completed this Machine Learning Capstone Project on Netflix Movies/Shows Clustering. While making this project I mainly use python programming language and some python modules such as Pandas & Numpy for Data Wrangling and for Data visualisation i use Matplotlib & Seaborn. As well as For this Project we use some machine learning concept such as DBSCAN , K-means Clustering, Hierarchical Clustering.  For this project I use a Netflix user dataset. From 2006 Netflix start the analysing user data to predict how much a viewer would like a movie, based on previous user Preferences. Whenever we access the Netflix service, Netflix recommendations system strives to help us to find a show or movie to enjoy with minimal effort. All of these done by using user data as an inputs that we process in our algorithms. Lets talking About the Datset :- The shape of dataset is (7787 x 12) I.e. Total Number of Rows are : 7787, Total Number of Columns are : 12,  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.  While using this Netflix Data we Try to Understand :-  i) Understanding what type content is available in different countries.  ii) Is Netflix has increasingly focusing on TV rather than movies in recent years.  iii) Clustering similar content by matching text-based features  When we start project first step is to find out collect the data then we check is any data miss or not.Then we start EDA ( Exploratory Data Analysis ) after analysis we move on our Machine Learning Part where we Perform Text Processing, DBSCAN, K-means Clustering, Hierarchical Clustering. After Performing all of this techniques we can Observe that TV shows signed have been higher than movies in 2016. While the the movies signed have been higher, it is blatantly visible that the TV shows signed per year is catching up to the movies signed by the year.  In DBSCAN clustered the data into 10 clusters with a silhouette score is 0.43875  In K-means Clustering the elbow and optimal silhouette score were found at 8 clusters with a  silhouette score of 0.474, Davies-Bouldin Index of 0.884 and Calinski-Harbaz Score of 2932.28  In Hierarchical Clustering the dendrogram distance was optimal at a distance of 20 with eight clusters producing a silhouette score of 0.4705, Davies-Bouldin Index of 0.8839 and Calinski-Harbaz Score of 2930.84 |