**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. Lavanya Shinde   Email :- [lavanyarshinde@gmail.com](mailto:lavanyarshinde@gmail.com)  Contribution on :- 1) Data Importing  1.1) Uploading on Drive  1.2) 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) Visualize Total Release Movies/Tv shows in Last 10 years  7) Visualize Types of Videos on Netflix  8) Visualize 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. Visualize the top Genres for Movies/TV-Shows on Netflix 2. Visualize the Top Directors on Netflix 3. Visualize 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 focused on TV rather than movies in recent   years?   1. Text Processing (Machine Learning) 2. Machine Learning Clustering 3. K-means Clustering 4. Hierarchical Clustering 5. Conclusion |
| **Please paste the GitHub Repo link.** |
| Github Link:- https://github.com/Lavanya0707/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, My name is Lavanya Shinde. This Netflix Movies/Shows Clustering Machine Learning Capstone Project was successfully completed by me. During the creation of this project, I mostly used the Python programming language and a few Python modules, including Pandas and Numpy for data wrangling and Matplotlib and Seaborn for data visualization. For this research, we also make use of machine learning ideas like DBSCAN, K-means clustering, and hierarchical clustering.  I use a dataset of Netflix users for my project. Beginning in 2006, Netflix began using user data analysis to forecast how much a viewer will enjoy a film based on historical user preferences. The Netflix suggestions algorithm aims to make it as easy as possible for us to find a show or movie to enjoy each time we use the service. All these actions were carried out using user data as inputs for our algorithms. Let’s talk About the Dataset:- The shape of dataset is (7787 x 12) I.e. Total Number of Rows are : 7787, Total Number of Columns are : 12,  The dataset was gathered through the third-party Netflix search engine Flixable.  The amount of TV series available on Netflix has almost tripled since 2010, according to an interesting analysis that was published in 2018. Since 2010, the number of movies available on the streaming service has dropped by more than 2,000, although the number of TV series has nearly tripled. Investigating what other insights may be drawn from the same datasets will be interesting.  While using this Netflix Data we Try to Understand:-  i) Understanding what type of content is available in different countries.  ii) Is Netflix has increasingly focused on TV rather than movies in recent years.  iii) Clustering similar content by matching text-based features  When I started my project, the first stage was to gather information, after which we checked to see if any data was missing. Following exploratory data analysis, we move on to the machine learning portion, where we perform text processing, DBSCAN, K-means clustering, and hierarchical clustering. After putting all of these strategies to use, we can see that TV program signings in 2016 were higher than movie signings. While the number of movies signed has been higher, the number of TV shows signed year is quickly catching up to the number of movies signed annually.  The data were grouped into 10 clusters by DBSCAN, and the silhouette score was 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 |