**Capstone Project Submission**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. Puneet Suthar ([7sutharpuneet@gmail.com](mailto:7sutharpuneet@gmail.com))  * Understanding Data   Understanding different column  Having overview of data   * Data Cleaning * Null value treatement * Dropping unused column. * Feature engineering * Adding new features (no. of directors, no. of casts, no. of countries, no. of genres ) * Adding new features based on date (year, month, day) * Conversion of columns data type to the required ones. * Exploratory Data Analysis * Different plots for all columns. * Clustering * Feature engineering for clustering.  1. Encoding of categorial columns 2. Combining the rating of movies and tv shows. 3. Creating of doc2vec model for decryption column.    1. Text preprocessing.    2. Vectorization.    3. Model training.    4. Creation of new data frame of description column after vectorization.  * Defining some functions used in clustering.   1. find\_optimal\_clusters()   2. plot\_pca\_tsne\_umap() * Clustering the data using different groups of data * Movie recommendation Using Cosine Simillarity and defined cluster for filtering * Conclusions |
| **Please paste the GitHub Repo link.** |
| Github Link:-  https://github.com/Puneet-Suthar/NETFLIX\_MOVIES\_AND\_TV\_SHOWS\_CLUSTERING  Drive Link:-  https://drive.google.com/drive/folders/1-VNbK\_zjCCsFgv8NsZmqYc9Ppz6w0XOJ?usp=sharing |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| The project is based on a Netflix dataset that contains information about movies and TV shows available on Netflix as of 2021. The dataset has 7787 rows and 12 columns. The objective of the project is to analyze the data and draw insights from it.  The first step is to handle the null values in the data. We replace null values in the 'director' column with 'unknown', null values in the 'cast' column with 'no\_cast', null values in the 'country' column with 'mode', and null values in the 'date\_added' column with 'drop'.  Next, we convert some of the columns containing list of values into lists, add new features such as the number of directors, casts, genres, and countries. We also analyze columns such as 'type', 'director', 'cast', 'country', 'date\_added', 'release\_year', 'rating', 'duration', 'genre', 'title', and 'description'.  The data shows that most of the content gets uploaded in the beginning and the middle of the month, and there were notable spikes in the number of releases in the months of October, November, December, and January. The content was mainly released between 2010 and 2021. The most frequent listed duration for TV shows is season 1 and for movies it's between 90 to 120 minutes. In the movies, the international movies genre is the most popular on Netflix, followed by drama and comedy. In the TV shows, the ranking is the same.  We also applied clustering algorithms to find similar shows based on their crew and category information. We divided the data into three groups and applied the k-means clustering algorithm to group the shows based on their genre, country, and release year.  In conclusion, this project provides insights into the Netflix dataset and shows how we can analyze such datasets to extract valuable information. |