**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| * **Mounika Dontula (**[**mounikadontula2795@gmail.com**](mailto:mounikadontula2795@gmail.com)**)**   **• Knowing the Data**   * + - * Importing libraries.       * Loading the data set and knowing the attributes.       * Data set information.   **• Understanding the variables.**   * + - * Statistical summary.       * Variable description and checking the unique values.       * Typecasting ‘Duration’ from string to integer and ‘date\_added’ from string to Date-Time.   **• Data Cleaning.**   * + - * Checking for duplicate values.       * Checking for missing /Null values and Handling them.       * Checking for Outliers and Handling them.   **• Exploratory Data Analysis(EDA)**   * + - * Univariate analysis for columns like ‘type’,’title’,’director’,country’,’release\_year’,’genres’,’description’ and ‘cast’.       * Bivariate analysis for columns like ‘movies and tv shows’,’shows released’,’length of movie’,genre of movie/tv show’,’top directors’,’country and rating’.   **• Textual Data Preprocessing**   * + - * Selecting attributes       * Removing non-ASCII characters       * Removing stop words and lower casing       * Removing punctuations and stemming       * Text vectorization and dimensionality reduction.   .**• Model Implementation**   * K-Means Clustering * Hierarchical Clustering   **• Recommendation system** |
|  |
| **Please paste the GitHub Repo link..0** |
| Github Link:-  https://github.com/MOUNIKADONTULA/Netflix-Movies-and-Tv-Shows-Clustering.git |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| First I decided to take up this project solely due to mutual interest in Netflix Movies and Tv Shows Clustering. Then I decided to gain insights doing some EDA etc. I started to form the questions. After doing some random EDAs. I gained some confidence.  **Short summary:**   * In this project, we worked on a text clustering problem wherein we had to classify/group the Netflix shows into certain clusters such that the shows within a cluster are similar to each other and the shows in different clusters are dissimilar to each other. * The dataset contained about 7787 records, and 12 attributes. * We began by dealing with the dataset's missing values and doing exploratory data analysis (EDA). * It was found that Netflix hosts more movies than TV shows on its platform, and the total number of shows added on Netflix is growing exponentially. Also, the majority of the shows were produced in the United States, and the majority of the shows on Netflix were created for adults and young adults. * It was decided to cluster the data based on the attributes: director, cast, country, genre, and description. The values in these attributes were tokenized, preprocessed, and then vectorized using TFIDF vectorizer.Through TFIDF Vectorization, we created a total of 10000 attributes. * Through TFIDF Vectorization, we created a total of 10000 attributes. * We used Principal Component Analysis (PCA) to handle the curse of dimensionality. 3000 components were able to capture more than 80% of variance, and hence, the number of components were restricted to 3000. * We first built clusters using the k-means clustering algorithm, and the optimal number of clusters came out to be 4. This was obtained through the elbow method and Silhouette score analysis. * Then clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters came out to be 9. This was obtained after visualizing the dendrogram. * A content based recommender system was built using the similarity matrix obtained after using cosine similarity. This recommender system will make 10 recommendations to the user based on the type of show they watched. |