**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Individual Project done By-:**  **NAME-** MONICA PATEL  **EMAIL –** [Monicapatel826@gmail.com](mailto:Monicapatel826@gmail.com)  **CONTRIBUTION -**  DATA CLEANING, EXPLORATORY DATA ANALYSIS, DATA VISUALISATION, REPORT GENERATION,PPT, TECHNICAL DOCUMENT,CLUSTERING , SUMMARY |
| **Please paste the GitHub Repo link.**  Github Link:- https://github.com/1993monica |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| **Name of The Project - NETFLIX\_MOVIES\_AND\_TV\_SHOWS\_CLUSTERING** Netflix, Inc. is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a film and television series library through distribution deals as well as its own productions, known as Netflix Originals. Netflix can be accessed via internet browser on computers, or via application software installed on smartTVs, set-top boxes connected to televisions, tablet computers, smartphones, digital media players, Blu-ray Disc players, video game consoles and virtual reality headsets on the list of Netflix-compatibledevices. It is available in 4K resolution.  Our main objectives of this project are to do exploratory analysis and find useful insights from dataset, to understand what type content is available in different countries, also to find out is Netflix has increasingly focused on TV rather than movies in recent years and at last to do clustering of similar content by matching text-based features from dataset. This analysis helped us to understand the trend. We found that most of the content on Netflix are of TV-MA and TV-14 rating. USA and India are two countries producing the maximum number of contents. Documentaries and stand up are top genre in terms of number of contents they have on platform. Further we found number of movies on Netflix outnumbers TV-shows. Our next job was to make an unsupervised clustering model. For this, we processed our text by removing unuseful characters like - stop words, punctuation and did stemming. After getting the length for each text feature we rescaled them for generalisation and started applying algorithms. We first used K-means clustering. In order to find appropriate cluster number, we used elbow method and finally got the best silhouette score of around 0.35. Next, we applied Hierarchal Agglomerative Clustering for which we made dendrogram. We also obtained silhouette score of around 0.32. With this we achieved our objectives of the project.  The dataset contains 7787 rows and 12 columns, cast and director columns have a lot of missing values so we dropped them and we have 10 features for the further analysis. 2.We have two types of content: movies and TV shows. 3.Netflix has 69% of its content as movies, so we can say that movies are clearly more popular on Netflix than TV shows.  I have used 2 clustering algorithms namely 1) ELBOW METHOD 2) DB SCAN |