**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Sneha Raikar  [sneharaikar650@gmail.com](mailto:sneharaikar650@gmail.com)  Individual Project |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/sneraikar/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)** |
| **Problem Statement**  This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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. In this project, you are required to do –  • Exploratory Data Analysis  • Understanding what type of content is available in different countries  • Is Netflix increasingly focused on TV rather than movies in recent years?  • Clustering similar content by matching text-based features  Our goal here is to make an unsupervised clustering model, which will help in garnering insights on Netflix and how its content is being consumed. **Approach :**  * On the given dataset of OTT platform -"**Netflix Movies and Tv shows"** clustering was performed. * First Data cleaning was done. Then Feature Engineering was done.Then some interesting insights were found by **Exploratory Data Analysis**. * United States and India top the countries that produce all the available content on Netflx. * TV-MA tops the graphs, indicating that mature content is more popular on Netflix. Then to perform clustering based on matching text features -Unsupervised Machine learning models were used. * Then text pre-processing was done by removing unusual characters like-stopwords, punctuation and stemming.   **Conclusion :**   * Majority of **content available on Netflix is Movies**. * In recent years though many TV shows have been added, **number Movies outpower the number of TV shows**.   • Firstly **K-Means clustering** unsupervised Machine learning technique was applied. For this elbow method was used to find K value and Silhouette score of 0.35 was obtained.  • Next **Hierarchical clustering** was applied for which dendogram was obtained. Silhouette score of 0.32 was obtained.  • So **K-Means clustering** performs better on the dataset. |