**Capstone Project-4 Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Name, Email and Contribution:** |
| **Rushikesh Arjun Mane**   * Introduction * Data cleaning * Null Value Treatment * Exploratory Data Analysis * Univariate * Data cleaning & pre-processing for clustering * Encoding the categorical data * K means clustering * Hierarchical Clustering * Silhouette analysis * Dash App recommendation System * Conclusion |
| **Please paste the GitHub Repo link.** |
| <https://github.com/RushikeshMane9094/Unsupervice-ML-Capstone-Project-Rushikesh-Mane-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  This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from flixable 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. Integrating, this dataset with other external datasets  such as IMDB ratings, rotten tomatoes can also provide many interesting findings.  APPROACH  Initially, in the 1st step imported the data set to carry out the analysis over the data set to comprehend the details of available data and Checked for Null values and treated them. Here, we found more than 30% null values in the director's column. Then, we take appropriate action for null values according to the circumstances.  Performed the Exploratory data analysis and tried to get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step with the help of visualization graph by getting insights from analysis.   * Data preprocessing – in this we remove the punctuation and stops words also used stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language. * We used the k-means clustering algorithm and then checked the model performance using Silhouette’s coefficient and elbow method to find the number of clusters.   Analyzing all the variables of the data set and identifying the solution for given tasks.  Performed hypothesis testing to get the insights on duration of movies and content with respect to different variables.  After doing feature engineering and finding the number of clusters, we used the k-means algorithm and then checked the model performance using Silhouette’s coefficient, to identify the best fit Model.  The number of movies on Netflix is growing significantly faster than the number of TV shows. Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019.   * The project's main goal is to create a model that can perform Clustering on comparable material by   matching text-based attributes.   * As the problem statement says, understanding what type of content is available in different countries   and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering  on similar content by matching text-based features. For that we used Affinity Propagation, Agglomerative Clustering, and K-means Clustering.  CONCLUSION   * ***It was interesting to find that majority of the content available on Netflix is Movies.*** * ***But in the recent years it has been focusing more on Tv-Shows.*** * ***Most of these contents are released either in the year ending or the beginning.*** * ***United States and India are among the top 5 countries that produce all of the available content on the platform.*** * ***Also 6 of the actors among the top ten actors with maximum content are from India.*** * ***TV-MA tops the charts, indicating that mature content is more popular on Netflix.*** * ***k=10 was found to be an optimal value for clusters using which we grouped our data into 10 distinct clusters.*** * ***Using the given data a simple recommender system was created using cosine\_similarity and recommendations for Movies and Tv Shows were obtained.***  **Future Scope:**  * **Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.** * **More time could be given into building a better recommender system, which later can be deployed on web for usage.**  **References:**  1. GeekforGeeks 2. Towards data science 3. Analytics Vidhya 4. Kaggle 5. W3 school 6. Stackoverflow |