**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |  |
| --- | --- |
| **Team Member’s Name, Email and Contribution:** | |
| 1. **VISHAKHA KUMARI (**[**vishakhakumari0598@gmail.com**](mailto:vishakhakumari0598@gmail.com)**)**   **Contribution:**   * **Outlining project planning** * **Univariate analysis.** * **Data preparation & cleaning.** * **Technical documentation.**  1. **ABHINAV AKOTKAR (**[**akotkarabhinaw@yahoo.com**](mailto:akotkarabhinaw@yahoo.com)**)**   **Contribution:**   * **Decision making on dataset** * **Handling imbalance dataset** * **Fundamental step for analyzing dataset.** * **Project summary**  1. **SHIVAM PATHAK (shivamself1997@gmail.com)**   **Contribution:**   * **Analysis of the data based on visualization.** * **Model selection.** * **Location visualization on map** * **Maintaining colab notebook** | |
| **Please paste the GitHub Repo link @ drive link** | |
| Github Link:- https://github.com/shivam0070/-NETFLIX  Drive Link:- https://drive.google.com/drive/folders/1qeSI0f23G8xb22ULA7ZIHe6UWAK4Ov8- | |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** | |
| Netflix Movies and TV shows Prediction is done on the basis of clustering techniquesalthough it isunsupervised machine learning. In this prediction we understanding what type content is available in different countries,  Is Netflix has increasingly focusing on TV rather than movies in recent years,  Clustering similar content by matching text-based features. | |
| 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.  Database contain 7787 entries and total 12 columns contain -:show\_id, type, title, director, cast ,country ,date\_added,release\_year,rating ,duration, listed\_in ,description  First step is to Importing some important libraries ,sklearn, preparing the problems and  summarized data  Then we Exploring features name, here we can see that most of data in object format.  And there are some features are missing some of the records. **Then we do Data Cleaning like missing values, Duplicated Values, Making a copy of our data frame for further operations, Filing null values with "unknown", Descriptive Statics.** Then we done **Analysis and Visualization of dataset by using matplotlib. like some Feature Engineering, Univariate Analysis, Genres, Content Addition, Text Visualization, Stemming,**  **Multivariate Analysis.**  **After that we working on the content available in dataset like exploring countries, differentiate between Ratings & Countries.** **Then we working on the data that is Netflix has increasingly focusing on TV rather than movies in recent years such points to be noticed in it.**After that we done **Clustering similar content by matching text-based features it includes Feature Engineering, Stemming, Text cleaning****TF-IDF (Term frequency-inverse document frequency) is a text vectorization that transforms the text into a usable vector. It include PCA-Principal Component Analysis to reduce dimensions, Cumulative Explained Variance.**By applying **K Means Clustering algorithm by using Silhouette Score & Elbow method for K-Means Clustering it used for various analysis like Elbow Method to get number of clusters, for Model fitting, to Predicting**. Using all above factors then we show Conclusion, Limitation & Future scope. |