**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| 1. **Adil Khan (adil.bm009@gmail.com)**    1. Data Processing    2. Data Wrangling       1. Importing some important modules       2. Summarising the data   1.3 Pre-processing the data  1.3.1. Finding the missing values  1.3.2 Searching for the duplicate values  1.3.3 Changing the Column names  1.3.4 Checking the categorical columns and assigning appropriate values  1.3..5 Checking the numerical columns  1.3.6 Changing the Data types  1.3.2 Exploratory Data Analysis  1.3.2.1 Univariate Analysis  1.3.2.2 Analysis of Dependent Variable  1.3.2.3 Analysation of Categorical Variables  1.3.2.4 Data visualisation  1.4 Checking the Correlation  1.5 Unsupervised ML  1.5.1 Silhouette  1.5.2 Elbow curve  1.5.3 DBSCAN  1.6 Observations   1. **Sunil Kumar Panigrahi(**[**sunilpanigrahi104@gmail.com**](mailto:sunilpanigrahi104@gmail.com)**)**   2.1 Data Wrangling  2.2 Exploratory Data Analysis  2.2.1 Analysis of Numerical variables displots  2.2.2 Bivariate analysis  2.2.3 Analysis of important factors to understand the data  2.2.4 Data Visualisation  2.3 Normalised the column data  2.4 Checking of Correlation between Variables  2.5 MultiCollinearity Heatmap  2.5.1 Correlation Heatmap  2.6 Created the Dummy Variables  2.6.1 One hot encoding  2.7 UnSupervised ML  2.7.1 DBSCAN  2.7.2 kMeans Clustering  2.7.3 Agglomerative Clustering  2.8 Observations  2.9 Presentation PPT   1. **Vivek Singh (sviveksingh31@gmail.com)**   3.1 Data Wrangling  3.2 Exploratory Data Analysis  3.2.1 Univariate Analysis  3.2.2 Analysis of important factors to understand the data  3.2.3 Data Visualisation  3.3 Model Building  3.4 UnSupervised ML  3.4.1 Agglomerative  3.4.2 Dendogram  3.4.3 DBSCAN  3.5 Observations  4.1 Conclusions  4.2 Presentation PPT  4.3 Technical Documentation     1. **Shubham Kumar (Samsonknight25@gmail.com)**   4.1 Data Wrangling  4.2 Data Cleaning  4.3 Data pre-processing  4.4 Summarising the data  4.5 Exploratory Data Analysis  4.5.1 Univariate Analysis  4.5.2 Bivariate Analysis  4.5.3 Analysis of important factors to understand the data  4.5.4 Data Visualisation  4.6 UnSupervised ML  4.6.1 UnSupervised ML  4.6.2 DBSCAN  4.6.3 kMEANS  4.6.4 Dendrogram  4.7 Observations  4.8 Conclusions  4.9 Presentation PPT  4.10 Technical Documentation   1. **Sharaffin B (sharaffinb@gmail.com)**   5.1 Data Processing  5.2 Data Wrangling  5.3 Importing some important modules  5.4 Summarising the data  5.5 Pre-processing  5.6 Exploratory Data Analysis  5.6.1 Univariate Analysis  5.6.2 Bivariate Analysis  5.6.3 Analysis of all the important factors to understand the data  5.6.4 Data Visualisation  5.7 UnSupervised ML  5.7.1 Silhouette score  5.7.2 kMEANS  5.7.3 Elbow method  5.8 Technical Documentation  5.9 Presentation PPT  5.10 Observations  5.11 Conclusions |
| **Please paste the GitHub Repo link.** |
| Github Link: <https://github.com/Sharaffin/Netflix-Movies-and-TV-Shows-> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| The most popular media and video streaming platform is none other than Netflix.  It includes over 8000+ movies with tv shows worldwide. Currently Netflix have  over 200M Subscribers Globally. The most popular entertainment service used  by people around the globe is also NETFLIX. It provides a huge collection of  movies and TV shows which are streamed anytime by means of online services.  Netflix observed that in 2018, the number of TV shows has nearly tripled  whereas the number of movies has decreased over 2,000 titles since 2010 and it  will be very interesting to dig into what all other insights can be derived from the  same dataset. The aim of the project is to create a model that can perform  Clustering on comparable material by matching text-based attributes.  The tabular dataset consists of listings of all the movies and tv shows available  on Netflix, along with details such as - cast, directors, ratings, release year,  duration, etc.The data set contains 7787 rows and 12 columns.  The main goal of our project is to create a model that can perform Clustering on  comparable material by matching text-based attributes.  According to the problem statement, the question arises that, 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 have used different clustering algorithms  We are going to perform the data wrangling on the raw data to get the useful data  without NAN values (null values) and observe the summary statistics of the  dataset. We prepared a dataset with feature engineering and feature scaling and  also, dropped out the inessential columns. We executed the exploratory data analysis. In text analysis (NLP) we used stop-words, removed punctuations, stemming & TF-IDF vectorizer and other functions of NLP.  Applied different clustering models like K-means, hierarchical, Agglomerative clustering, DBSCAN on data we got the best cluster arrangements. By applying the silhouette score method for n range clusters on the dataset we got the best score which is 0.3746 for 3 clusters it means content explained well on their own clusters, by using elbow method after k = 3 is the best cluster. |