**CapstoneProjectSubmission**

|  |
| --- |
| **TeamMember’sName,EmailandContribution:** |
| **1.Lucky Jain Email-id:-**[jainluckycool@gmail.com](mailto:jainluckycool@gmail.com)  **Contribution-**   * EDA on type and ratings and to evaluate the highest rating with content preferred by theaudience. * Analysisonthecountriesandcastintheshowusingwordcloudvisualization. * AnalysisdonetoobserveNetflixhasincreasinglyfocusingonTV showsratherthanmoviesinrecentyears   ornot.   * Data preparation using nlp. * Modelimplementationforagglomerativeclustering. * Correlationmatrix. * Elbowmethod&Box-plotforK-meansclustering   **2.Debashish Das Email-id:-devashishdas40@gmail.com**  **Contribution-**   * CheckingtheNAN valuesinthedataset. * EDAonhighestwatchedgenreonNetflix. * Hypothesisfindingsbasedonreleaseyearand datefeature. * Data Preprocessing (USING NLTK) * Standardizedthedataintoscalartransformation. * ModelimplementationforK-meansClustering. * Dataengineeringonthebasisofclusteringtextbasedfeatures * Silhouette score method&Box-plotforK-meansclustering * Dbscan   **3. VivekKatolkar Email-id:-vivekkatolkar7@gmail.com**  **Contribution-**   * ScatterplotwithtypefeatureandwithclustersafterfittingK-meansclusteringmodel. * EDAonhighestwatchedgenreonNetflix. * Convertingthedateintodate-timeformat. * Understandingthetypeofcontentavailableindifferentcountries * dendogram |
| **GitHub link:**  https://github.com/vivekkatolkar/Netflix-Movies-And-Tv-Shows-Clustering.git |

**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 is by far the most widely used media and video streaming service. It includes over 8000+ movies with tv shows worldwide. Currently, Netflix has more than 200 million subscribers worldwide. NETFLIX is also the most widely utilized entertainment platform worldwide. It offers a vast library of films and TV series that may be seen at any time through internet services.

Netflix is without a doubt the most used media and video streaming service. There are more than 8000 films and television programmers included. At the moment, Netflix has over 200 million subscribers worldwide. Additionally, NETFLIX is the most widely used entertainment service worldwide. Through online services, it offers a vast library of movies and TV series that may be accessed at any time.

The tabular dataset includes listings for all Netflix movies and TV shows, together with information about the actors, directors, ratings, release year, duration, and other factors. It has 7787 rows and 12 columns.

Our project's main objective is to build a model that can cluster similar data by matching text-based features.

Accordingtotheproblemstatement,thequestionarisesthat,understandingwhattype of content is available in different countries and Is Netflix increasinglyfocused on TV rather than movies in recent years we have to do clustering onsimilar content by matching text-based features. For that we have used K-meansClustering.

To obtain relevant data free of NAN values (null values), we will manipulate the raw data. We will next look at the dataset's summary statistics. With feature engineering, feature scaling, and the removal of unnecessary columns, we prepared a dataset. To use the model in the study, the data was transformed into a common scalar form.

We executed the exploratory data analysis. Reviewing all the data processingthen the model was trained to form Collections. In which we remark as below-k-means clustering model gave insights of silhouette analysis consisting of2,3,4,5,6 clusters.

For n\_clusters = 2 The average silhouette score is 0.42541313028836003

For n\_clusters = 3 The average silhouette scoreis: 0.3940500353920696

For n\_clusters = 4 The average silhouette scoreis: 0.38498095158183726

For n\_clusters = 5 The average silhouette scoreis: 0.3962372504377786

For n\_clusters = 6 The average silhouette scoreis: 0.392565886828632

Inthe end, weplot boxplotto predictthe hypothesis-

* After clustering, we can state that the number of TV series that have been released over the past few years is not increasing, which is our alternative hypothesis.
* Oursecondalternativehypothesis isthe numberofTVshows addedto Netflixis high.

We calculated that whereas movies make up 97.2 percent of the total, TV series only make up 2.8 percent. When compared to films made in the last 10 years, Netflix has added a lot more movies and TV episodes in the past years, but the numbers are still small. In many nations, people choose to watch movies over TV programming.

|  |
| --- |
| **Drive link**  https://drive.google.com/drive/folders/12lrhEaHK9DVmz1GntpvGWoQrp6fSM5-7?usp=sharing |