**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| Name : Md Ismail Quraishi  Email : [mdismailquraishicse@gmail.com](mailto:mdismailquraishicse@gmail.com)  Contribution : Individual |
| **Please paste the GitHub Repo link.** |
| Github Link:-  <https://github.com/mdismailquraishicse/capstone_project_4_netflix_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 is an Unsupervised Machine learning project. In this project I will have to build a model that can be capable of clustering different different types of data. The dataset is about netflix shows which has 7787 rows and 12 columns like show\_id which represents ID of the show, type represents type of the show, title represents show title, cast represents name of the casting stars, country represents the country of the show, date added represents the date when the show is added to netflix, release\_year represents the year the show was released,rating represents the rating of the show, duration represents the length of the show, listed\_in tells what type and where the show belongs from, description gives short descriptions about the show.  **Approaches** :  My task is to read and understand the data. After that I will have to show some meaningful charts and explain everything about the chart. Then according to the visualization chart I will have to make some hypothesis assumptions about the project then testing the assumptions. Then I will handle missing and null values and outliers after completing these tasks I will look for imbalanced data. If there is any imbalanced data then I will have to deal with that. Then I will select some important features and further I will split the data for test and training purposes.  **Conclusions** :  1. The dataset has 7787 rows and 12 columns.  2. There are 30.68% null values in director, 9.22% in cast column, 6.51% in country, 0.13 in date\_added, and 0.09 on rating columns.  3. The dataset containing only movies has shape (5377,14) and for TV Shows has (2410,14).  4. Total number of features after vectorization is 40255.  5. Total features after dimensionality reduction is 364.  6. Kmeans clustering is giving good clusters therefore I would choose KMeans with n\_clusters = 6 because silhouette in this case is very good .046  7. Agglomerative clustering is also good but for a high number of clusters it is not giving better clusters.  8. DBSCAN is giving a large size of cluster because my data is very noisy therefore in this case dbscan is not good. |