# Capstone Project Submission

**Instructions:**

- Please fill in all the required information.
- Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Contribution Roles:

i). **Gaurav Bhakte (bhaktegaurav1999@gmail.com) :**

- Data Description
- Model Implementation
  - K- means Clustering

ii). **Pratiksha Kharode (pratikshakharode1312@gmail.com) :-**

- Data Cleaning
- Movie Recommendation
- Topic Modeling
  - LSA(Latent Semantic Analysis)

iii).**Vinit Ladse (ladsevinit7@gmail.com) :-**

- General Analysis
- EDA (Exploratory Data Analysis)
- Data Preprocessing
- Topic Modeling
  - LDA(Latent Dirichlet Allocation)

Please paste the GitHub Repo link.

**Github Link :-** https://github.com/GauravBhakte/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)

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 Netflixsearch 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 explote 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.

- show_id : Unique ID for every Movie / Tv Show
- type : Identifier - A Movie or TV Show
- title : Title of the Movie / Tv Show
- director : Director of the Movie
- cast : Actors involved in the movie / show
- country : Country where the movie / show was produced
- date_added : Date it was added on Netflix
- release_year : Actual Releaseyear of the movie / show
- rating : TV Rating of the movie / show
- duration : Total Duration - in minutes or number of seasons
- listed_in : Genere
- description: The Summary description

Things to be done-

1. Exploratory Data Analysis.
2. Understanding what type content is available in differentcountries.
3. Is Netflix has increasingly focusing on TV rather than movies inrecent years.
4. Clustering similar content by matching text-based features.

Discussion of Netflix Movies And Tv Shows Clustering will involve various steps such as:

- Loading the data into data frame
- Data Description
- Exploratory Data Analysis
- Data Preprocessing
- Topic Modeling
  - LSA(Latent Semantic Analysis)
  - LDA(Latent Dirichlet Allocation)
- K- means Clustering
- Conclusion

That's how we have accomplished our team work in Netflix Movies And Tv Shows Clustering .Throughout the project we learn many new things right from takingproblem statement to understand the technical side of a data to analysis. We deal with Netflix data.

- There are about 70% movies and 30% TV shows on Netflix.
- The United States has the highest number of content on Netflix by a huge margin followed by India.
- LDA and LSA has sorted much more similar titles in a group of genre.
- Recommendation system works perfectly well with description column.
- After applying K - means optimal value of number of clusters is 5.
- Silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are.

Drive Link:-
https://drive.google.com/drive/folders/1wsKwtQ-9CPBufrmX5vxWLVl2HilwgLs5?usp=sharing