Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
V Rupesh Kumar Patro: (varnasipatro@gmail.com) Upload dataset to Google Colab · Data Cleanup. · Data cleaning. . EDA · Preprocessing · Model Implementation · Project Summary   **Shashank Maindola: (shashank.ddun@gmail.com)** · Upload dataset to Google Colab · EDA · Correction of data types · Data Visualizations . Feature Engineering · Technical Write up · PowerPoint presentation · Technical Documentation
Please paste the GitHub Repo link.

Github Link:- https://github.com/rupeshpatro2001/NETFLIX_MOVIE_AND_TV_SHOWS_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)

Netflix is a popular American subscription-based streaming service that provides a vast library of movies, television series, documentaries, and other forms of visual entertainment. It was founded in 1997 by Reed Hastings and Marc Randolph and has since grown to become one of the world's leading streaming platforms.

Netflix offers a wide range of content, including original productions and licensed content from various studios and distributors. Subscribers can access the service through the Netflix website or mobile applications on various devices, such as smartphones, tablets, smart TVs, gaming consoles, and streaming media players.

The platform has revolutionized the way people consume media by offering a convenient and on-demand way to watch TV shows and movies. Users can stream content instantly without having to wait for traditional television schedules or physical media like DVDs. Netflix's extensive library, personalized recommendations, and the ability to binge-watch entire series have contributed to its widespread popularity across the globe.

We have a given dataset which consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable 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.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings. There are following attributes in the dataset:-

'**show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added', 'release_year', 'rating', 'duration', 'listed_in', 'description'**

Firstly, we imported the library which were required to process our data, then we mounted the data from the drive link folder. We used following steps for appliance energy prediction using machine learning.

- Exploratory Data Analysis
- Pre Processing
- Model Implementation & Explanation
- Conclusion

Firstly, we conducted Exploratory Data analysis under which we started with preliminary data analysis. Then we did missing value treatment. We Broadly summarised our observations in following steps:-

1. Content Type
2. Growth in the content release over years
3. Title
4. Top ten directors on Netflix
5. Top ten actors on Netflix
6. Content by Country
7. Date Added
8. Release Year
9. Rating
10. Duration of Content
11. Top Genres

As part of preprocessing, we first expand contraction, lower casing, removed punctuation, removed URL's & digits and removed stopwords and white spaces.
Then we used other NLP algorithms and performed tokenization, text normalization, text vectorization and dimensionality reduction.
For vectorization used TF-IDF method as classic approach of converting input data from its raw format (i.e. text) For Dimensionality reduction we used PCA with 0.95 so that 95% of data can be present before applying models.

Then we trained our model using various clustering models and checked their performance.
Based on the model performance we selected nearest neighbours as the best model and we summarised our observations.