**Internship Report**

**External guide Name : Mallikarjun Kumbar**

**Designation : director**

**Mail :** [**info@takeitsmart.in**](mailto:info@takeitsmart.in)

**Phone number : 8792697647**

**ABOUT COMPANY**

**TAKE IT SMART (OPC) PVT.LTD** is an Indian based engineering and Software Company headquartered in Bangalore, Karnataka, India. It is both product and service oriented software company. All offices employ an experienced team of professionals, with an outstanding track record of handling complex web & Apps development projects.

**2.1 HISTORY**

The company was legally registered in the year 2021, but it made its humble beginning in the year 2018 with a team of Two members.

**2.2** **COMPANY STRATERGY**

* **Purpose:** To be a leader in the software Industry by providing enhanced services, relationship and profitability.
* **Mission:** To build long term relationships with our customers and clients and provide exceptional customer services by pursuing business through innovation and advanced technology
* **Vision:** To provide quality services that exceeds the expectations of our esteemed customers.
* **Core values:**
* To incorporate good business practices in order to achieve customer satisfaction and treating the customers with respect and faith.
* To grow through creativity, invention and innovation.
* To integrate honesty, integrity and business ethics into all aspects of the business functioning.

**Goals:**

* To improve, grow and become more efficient in the field electronics engineering and software development and develop a strong base of key clients.
* To understand customer requirements and fulfill them.
* Increase the assets and investments of the organization to support the development of services and expansion of the organization.
* To increase the productivity and improve the customer service satisfaction.
* To do Innovations in Software field and provide quality services to deliver a range of products.

**2.3 COMPANY SERVICES**

**TAKE IT SMART (OPC) PVT.LTD** have its own services such as,

* Embedded Applications development
* Web design and development
* IT Service
* Android app Development
* Web Bases Software Solutions
* Web Based ERP
* Web Based Ads Mobile Based Services: Mobile Web Apps a. Android Apps b. Windows Apps c. IOS Apps d. Cross Plate forms Apps
* Native Apps
* Hybrid apps Get trained for industry requirements while you pursuing degree The Different verticals that we operate in are: ¬ Internship & Software Training

**2.4 DOMAINS**

TAKE IT SMART (OPC) PVT.LTD have working with several domains like-

* IT
* Digital marketing

**2.5 DEPARTMENTS**

* **Marketing:** These are the main section of the market departments:
* **Sales department** is responsible for the sales and distribution of the products to the different regions.
* **Promotion department** decides on the type of promotion method for the products, arranges advertisements and the advertising media used.
* **Distribution department** distributes the products across the industries.
* **Embedded System and Internet of Things (IOT) department.**
* **Machine learning and web development department.**

**Business Address:** **Take It Smart (OPC) Pvt.Ltd**

14,SGN Arcade, 1st Floor, 2nd stage, 1st Main Rd,

RPC Layout, Hoshalli Extension, Stage 1,

Vijayanagar, Bengaluru, Karnataka 560040

**Mobile**: +91-8050104212

**Email**: [info@takeitsmart.in](mailto:info@takeitsmart.in)

**Website**: www.takeitsmart.in

**Programmers and opportunities:**

The Institute combines pioneering research with top class education. An innovative curriculum allows the student flexibility in selecting courses and projects. Students, even at the undergraduate level, get to participate in on-going research and technology development - an opportunity unprecedented in India. As a result, a vibrant undergraduate programmer co- exists with a strong postgraduate programmer.

**NETFLIX MOVIES AND TV SHOWS CLUSTERING**

**Exploring Netflix Content:**

**Introduction**

Netflix, the renowned streaming giant, has transformed the way we consume entertainment. As of 2019, it offered a vast library of TV shows and movies to cater to diverse tastes and preferences. This dataset, sourced from Flixable, a third-party Netflix search engine, provides a snapshot of the content available on Netflix at that time.

The ever-evolving landscape of streaming platforms has witnessed intriguing shifts over the years. In 2018, Netflix reported a substantial increase in the number of TV shows, nearly tripling since 2010, while the number of movies decreased significantly. Beyond these statistics, there lies a treasure trove of insights and patterns waiting to be discovered within the dataset.

In this project, we embark on an exploratory data analysis (EDA) journey to delve into Netflix's content landscape. We aim to uncover valuable information about the types of content available across different countries, Netflix's focus on TV shows versus movies, and the potential for clustering similar content based on text-based features. Additionally, the integration of external datasets, such as IMDB ratings and Rotten Tomatoes, may yield further intriguing findings.

**Data Preparation**

Before we embark on our exploration, it is imperative to prepare the data. This involves tasks such as data cleaning, feature engineering, and handling missing values to ensure that the dataset is in optimal condition for analysis. Given the mix of textual and numeric data, various preprocessing steps will be applied.

**Some key data preparation steps include:**

* Data Cleaning: Identifying and rectifying any inconsistencies or anomalies in the dataset.
* Feature Engineering: Creating new features and extracting relevant information from existing columns.
* Text Data Processing: Cleaning and tokenizing text-based features for analysis and clustering.
* Missing Value Handling: Dealing with missing data points through imputation or removal.

**Objectives**

In this project, we set out to achieve the following objectives:

* Exploratory Data Analysis (EDA): Our primary goal is to perform a thorough EDA of the Netflix dataset. This involves gaining insights into the distribution of content types, analyzing the country-wise availability of content, and visualizing trends in content addition over time.
* Content by Country: We aim to understand the diversity of content available in different countries. This includes identifying popular genres, directors, and actors associated with specific regions.
* Focus on TV Shows vs. Movies: We will investigate whether Netflix has shifted its focus toward TV shows in recent years compared to movies. This analysis will provide insights into Netflix's content strategy.
* Text-Based Content Clustering: Leveraging text-based features like descriptions and genres, we will explore the possibility of clustering similar content. This could help in content recommendation and understanding content patterns.
* Integration of External Datasets: If applicable, we will explore the integration of external datasets, such as IMDB ratings or Rotten Tomatoes scores, to enhance our analysis and gain a more comprehensive understanding of content quality.

**Model Selection**

The next phase of our Netflix data exploration project involves employing machine learning models to gain further insights and perform clustering tasks. Given the nature of the dataset and our objectives, here are some suitable models and techniques for consideration:

**K-Means Clustering:**

K-means is a versatile and widely-used clustering algorithm that can group similar content based on features like genre, description, or cast. We can use it to identify content clusters with similar characteristics.

**Silhouette Score:**

The Silhouette Score is a metric used to measure the quality of clusters formed by a clustering algorithm like K-Means. It quantifies how similar each data point is to its own cluster compared to other clusters. Here's how it's calculated:

For each data point, calculate two values:

a: The average distance from the data point to all other points within the same cluster.

b: The smallest average distance from the data point to all points in a different cluster, minimizing over clusters.

The Silhouette Score for the data point is then calculated as (b - a) / max(a, b).

The overall Silhouette Score for the entire dataset is the average of the Silhouette Scores for individual data points. It ranges from -1 (poor clustering) to +1 (good clustering), with higher values indicating better-defined clusters.

A high Silhouette Score indicates that the data points within a cluster are close to each other, and the clusters are well-separated. In contrast, a low or negative score suggests overlapping or poorly defined clusters.

**Agglomerative Clustering:**

Agglomerative Clustering is another clustering technique used for grouping data points. Unlike K-Means, which starts with individual data points as clusters and merges them, Agglomerative Clustering starts with each data point as a separate cluster and recursively merges them into larger clusters. Here's how it works:

Initialization: Start with each data point as a single-cluster, treating them as individual clusters.

Merge: Identify the two closest clusters based on a distance metric (e.g., Euclidean distance) and merge them into a single cluster.

Repeat: Continue merging the closest clusters until there is only one cluster containing all data points, or until a specified number of clusters is reached.

Agglomerative Clustering is versatile and can produce hierarchical cluster structures. It's often used for tasks where the number of clusters is not predefined, and the algorithm determines the optimal clustering structure based on distance measures.

* Text Vectorization: To extract meaningful information from text-based features like descriptions, we can employ techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) or Word Embeddings (Word2Vec, GloVe) to convert text data into numerical vectors for modeling.
* Hierarchical Clustering: This technique can help us create hierarchical clusters of content, which can be useful for understanding content relationships at different levels.
* Time Series Analysis: To investigate Netflix's content strategy over time, time series models can be employed to identify trends in content additions, both in terms of TV shows and movies.

**Model Training**

Once we've selected the appropriate models, we'll proceed with model training and clustering. This involves the following steps:

* Data Transformation: Ensure that the dataset is prepared, including the extraction and transformation of relevant features for modeling, especially text data.
* Model Initialization: Initialize the selected models with appropriate hyperparameters, such as the number of clusters (K-means), text vectorization settings, or PCA components.
* Training: Train the models using the transformed data. For K-means, this involves finding the optimal cluster centroids, while for text-based models, it includes creating vector representations of the text features.
* Evaluation: Assess the quality of clusters or patterns generated by the models. For unsupervised clustering tasks, metrics like silhouette score, Davies-Bouldin index, or visual inspection of clusters can be used to evaluate the model's performance.

**Evaluation**

Evaluation is a critical step to determine the effectiveness and validity of our model outcomes. Here's how we can approach the evaluation process:

* Cluster Interpretability: For K-means and hierarchical clustering, interpret the meaning and characteristics of clusters. Analyze whether the content within clusters makes sense and provides meaningful insights.
* Visualization: Visualize the clusters using techniques like scatter plots, dendrogram plots (for hierarchical clustering), or t-SNE (t-Distributed Stochastic Neighbor Embedding) for dimensionality reduction visualization.
* Validation Metrics: Calculate clustering metrics, such as silhouette score and Davies-Bouldin index, to measure the quality and separation of clusters. A higher silhouette score indicates better-defined clusters.
* Content Profiling: Examine the content within clusters to identify patterns and commonalities. For example, do certain clusters predominantly contain movies from specific countries or TV shows with particular genres?
* Temporal Analysis: If time series analysis is performed, visualize trends in Netflix's content additions over time. Identify periods when the focus shifted from movies to TV shows or vice versa.
* External Dataset Integration: If external datasets (e.g., IMDB ratings) are integrated, analyze correlations or patterns between content quality (ratings) and clustering results.
* User Feedback: Consider gathering user feedback or domain experts' opinions to validate the relevance and meaningfulness of content clusters.

**Cluster of Same Types of Shows**

****

**Conclusion**

* From elbow and sillhoute score ,optimal of 26 clusters formed , K Means is best for identification than Hierarchical as the evaluation metrics also indicates the same.in kmean cluster 0 has the highest number of datapoints and evnly distributed for other cluster
* Netflix has 5372 movies and 2398 TV shows. There are more movies on Netflix than TV shows.
* TV-MA has the highest number of ratings for tv shows i,e adult ratings
* Highest number of movies released in 2017 and 2018 highest number of movies released in 2020 The number of movies on Netflix is growing significantly faster than the number of TV shows. We saw a huge increase in the number of movies and television episodes after 2015. there is a significant drop in the number of movies and television episodes produced after 2020. It appears that Netflix has focused more attention on increasing Movie content than TV Shows. Movies have increased much more dramatically than TV shows
* The most content is added to Netflix from october to january
* Documentaries are the top most genre in netflix which is fllowed by standup comedy and Drama and international movies
* Kids tv is the top most TV show genre in netflix
* Most of the movies have duration of between 50 to 150
* Highest number of tv\_shows consistig of single season
* Those movies that have a rating of NC-17 have the longest average duration.
* When it comes to movies having a TV-Y rating, they have the shortest runtime on average
* 30% movies released on Netflix. 70% movies added on Netflix were released earlier by different modes.

