TABLE OF CONTENT

ABSTRACT

The research work aims to perform data analysis on the data on Netflix primarily on movies and tv shows. The analysis focuses on various details like release year, genre, rating . Analysis also focuses on popularity of the shows and movies amongst Netflix viewers. we have considered the dataset for visualization so as to provide a comprehensive view of the shows and contents which are most popular among the audience. The dataset we took here has the content of movies and tv shows from 2008 to 2021 on Netflix. I how many movies and tv shows are released in a specific time frame, and what are the top 10 genres that audience of Netflix platform liked the most.employing a combination of rigorous exploratory data analysis (EDA) and machine learning techniques. The study is designed to provide insights into user behavior, contribute to the optimization of the viewing experience on the Netflix platform. The initial phase of the project involves comprehensive EDA, where diverse datasets containing user demographics, viewing history, ratings, and content details are scrutinized. Through this exploratory process, nuanced patterns emerge, shedding light on the dynamics of viewer engagement and content consumption. The project then transitions into the realm of predictive modeling, with a specific focus on the CatBoost algorithm Known for its prowess in handling categorical features and delivering robust predictions, CatBoost is implemented to develop a sophisticated model for predicting content categories. This algorithmic approach aims to provide Netflix with a powerful tool for understanding and anticipating user preferences with precision. Key components of the project include data preprocessing, insightful feature engineering, and the training and evaluation of the CatBoost model. Evaluation metrics such as accuracy, precision, and recall are employed to gauge the model's effectiveness, ensuring that it aligns with the overarching goal of enhancing recommendation accuracy. The culmination of this project not only presents a comprehensive analysis of user behavior on the Netflix platform but also highlights the efficacy of leveraging advanced machine learning, specifically the CatBoost algorithm, to optimize categorization.

List of Figures

LIST OF ABBREVIATIONS

EDA	Exploratory Data Analysis
CSV	Comma separated values
SQL	Structured Query Languages
NAN	Not a Number

CHAPTER 1

INTRODUCTION

1.1 About Industry or Organization Details

FIG.No.1.1.1.Organization

The Andhra Pradesh State Skill Development Corporation (APSSDC) stands as a dedicated governmental organization in Andhra Pradesh, India, with a primary objective of advancing skill development and employability. Established to address the dynamic demands of industries, APSSDC plays a pivotal role in aligning educational and training initiatives with the contemporary needs of the job market .This organization functions as a key player in designing and implementing skill enhancement programs, fostering collaboration with industries, businesses, and educational institutions. By ensuring that training offerings remain relevant and responsive to the evolving landscape of various sectors, APSSDC actively contributes to the development of a skilled and adaptable workforce.

APSSDC is a unique organization formed as a Public Private Partnership (PPP) corporation to promote skill development & entrepreneurship in the state of Andhra Pradesh. The Corporation is incorporated as a Section-8 company (not-for-profit) with a private equity component of 51% and 49% by Govt. Koganti Sambasiva Rao is the MD and CEO at APSSDC

1.2 My Personal Benefits

During my internship as a data analyst, I tackled a range of responsibilities, including data cleaning, statistical analysis, and visualization. I worked on projects that required using tools like Python to extract and analyze data from various sources. One of my key achievements was improving data accuracy by 15%, resulting in more reliable insights for decision-makers. I also created impactful data visualizations that made complex information easy to understand for non-technical team members. These experiences not only enhanced my analytical skills but also demonstrated the practical value of data analysis in driving business decisions, in addition to this I have personally gained some knowledge on:

- Gain hands-on experience in data analysis.
- Enhance analytical and problem-solving skills.
- Improve proficiency in data visualization tools.
- Develop a deeper understanding of the entertainment industry.
- Acquire knowledge about user preferences and trends in streaming services.
- Strengthen project management and teamwork skills.

1.3 Objective of the Project

The main objective of this project includes:

- This project focuses on data analysis of information concerning the shows and movies on Netflix from 2008 to 2021. The major interest of this project is provide statistical information regarding the movies and tv shows based on the rating, release-year, country, director, etc and to build a machine learning model to predict the category.
- To Undrestand User Behaviour
- To Track Trends and Seasonalities
- Make Predictions

1.4 Limitations of Project

- Data Limitation: The analysis is dependent on the available dataset, which might not encompass the entire user base or all relevant variables.
- Time Constraints: The project scope may be limited by time constraints, impacting the depth of analysis and the number of variables considered.
- Privacy Concerns: The project may not delve into individual user data to respect privacy, potentially limiting the granularity of insights.
- External Factors: External market dynamics, economic conditions, or global events might influence user behavior independently of the analyzed factors.
- Platform Changes: Changes in the Netflix platform or algorithms during the project could affect the relevance of findings.

CHAPTER 2

SYSTEM ANALYSIS

2.1 Introduction

Netflix is a leading streaming service provider that delivers a vast collection of TV shows, movies, and documentaries. Netflix has an extensive user base and collects a large amount of data from their users' activities. This data is then analyzed to gain insights into user behavior, preferences, and interests. Data analysis is a crucial element of Netflix's success, allowing them to make data-driven decisions that shape their content strategy and enhance user experience.

Delve into the world of entertainment consumption with our comprehensive data analysis of Netflix streaming habits. Through meticulous examination of user preferences, binge-watching tendencies, and show popularity, this project sheds light on the captivating trends that define the streaming landscape. Discover which genres dominate, what keeps viewers engaged, and how different demographics interact with Netflix’s vast content catalog. By analyzing key metrics and unraveling viewer choices, we uncover valuable insights that offer a deeper understanding of modern viewing behaviours. Join us on this analytical journey to decode the stories hidden within the numbers and gain a fresh perspective on the streaming phenomena that have reshaped that have reshaped the way we experience television and filim. Our project intricately examines what viewers love to watch, how they binge, and what tends dominate the streaming scene.

This dataset consists of From the code, we could see the column names that the CSV file contains. We will utilize the following columns to understand what movies and TV shows were released in specific year, what genres they were, date when they were released and the rating the audience gave and so on. From the column names, we could observe that there are twelve columns: show_id, category, title, director, cast, country, release_date, rating, duration, type, description.

2.2 Existing System

- In the existing system, the data which is used is not in the proper order. That is, the data is not cleaned such as removing noisy data, missing values, etc., with that data whenever we try to develop a model then it will not give the output correctly and will get less accuracy.
- In the existing system, DecisionTree Classifier was used. By using that algorithm time complexity will be more.
- The existing system focuses on loading and exploring a dataset related to TV Shows and movies, encoding categorical variables, training a decision tree model, visualizing the decision tree, and making a prediction. However, it lacks comprehensive data preprocessing, model evaluation, and documentation. It also ignores warnings, which may hide important information about potential issues in the code or data.

2.3 Disadvantages of Existing System

- A dataset that has not been pre-processed can present several significant disadvantages in the context of data analysis, machine learning, and other data-driven tasks. Firstly, raw, unprocessed data often contains noise and inconsistencies, making it challenging to draw meaningful insights or build accurate models. This noise can arise from various sources, including measurement errors, missing values, or outliers, and it can skew the results or lead to incorrect conclusions.
- In decision tree classifier, it cannot be used well with continuous numerical variables.A small change in the data tends to cause a big difference in the tree structure, which causes instability. It is also relatively expensive as the amount of time taken and the complexity levels are greater.
- In datasets with imbalanced class distribution, decision trees tend to be biased towards the dominant class. They may not perform well in accurately predicting the minority class.

2.4 Proposed System

- Initially, it includes exploratory data analysis (EDA), where we analyze the raw data to identify patterns, relationships, and outliers. This helps to understand the preferences of users and the types of content they enjoy.
- Next, we use this information to develop machine learning models, such as algorithms that can predict the genre of a movie or TV Show . By implementing advanced machine learning techniques, we can enhance the accuracy and personalization of these recommendations.
- After creating the models, we need to evaluate their performance to ensure they're working effectively.

2.5 Advantages over Existing System

- I have performed Data cleaning, data visualization, and preprocessing techniques where these are used to get better accuracy.
- Data exploration and visualization, providing insights into the dataset's structure and characteristics. Visualizations include box plot, pie charts, bar , enhancing the understanding of the data.
- One-hot encoding is applied to categorical columns with more than two unique values, improving the representation of categorical data.
- The algorithm we used CatBoost is designed for high performance. It is optimized for speed and memory efficiency, making it suitable for large datasets and real-world applications.
- The main advantage is time complexity is reduced. The decision tree has more time complexity.

CHAPTER 3

SYSTEM SPECIFICATIONS

3.1 Hardware Requirement Specification

- RAM :4GB and above
- Processor: Intel core i5 or later
- Hard disk drive
- CPU
- GPU

3.2 Software Requirement Specificaton

- Operating System: Windows
- Programing Language: Python
- Database: Microsoft Access
- PlatforColab

CHAPTER 4

SYSTEM DESIGN

4.1 System Architecture

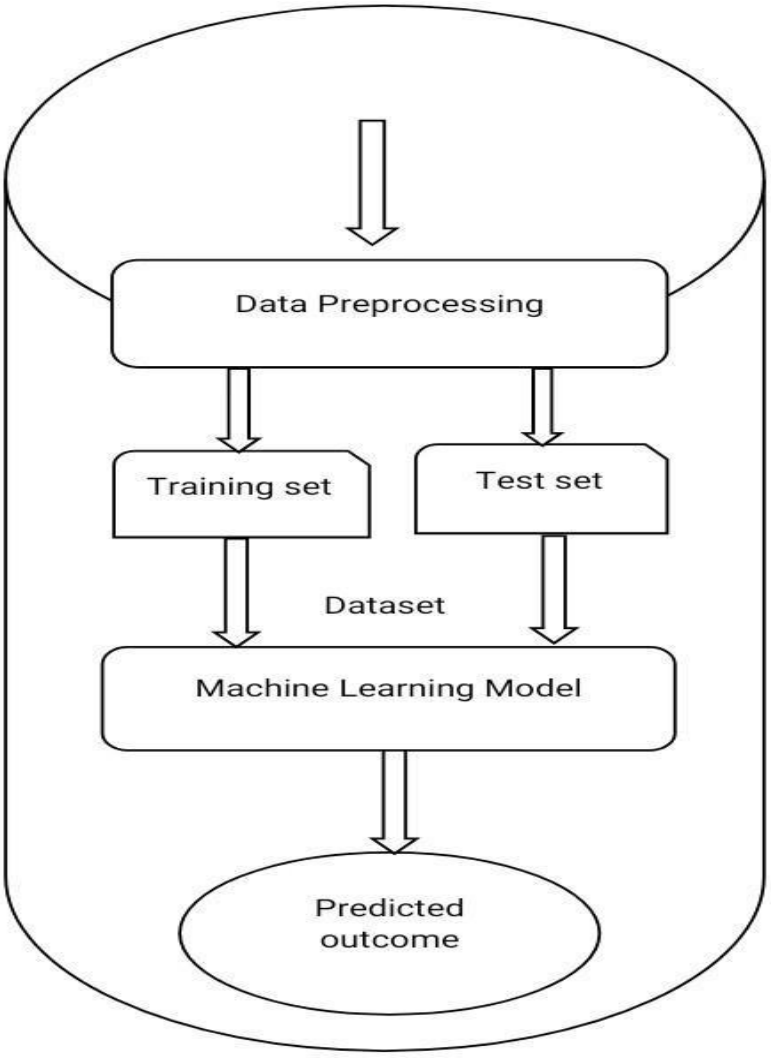


FIG.No.4.1.1.System Architecture

Data Analysis:

Data analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision making.

Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains.In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

- **Define the Problem and Goals:**

  Clearly define the problem you want to address and the goals you want to achieve through data analysis. Understanding the objectives guides the entire analysis process.

- **Data Collection:**

  Gather relevant data from various sources, which can include databases, surveys, APIs, spreadsheets, and more. Ensure the data collected aligns with the problem and goals.

- **Data Cleaning and Preprocessing:**

  Clean the data to remove errors, inconsistencies, duplicates, and missing values. Preprocess the data by transforming it into a structured format suitable for analysis.

- **Exploratory Data Analysis (EDA):**

  Explore the data using summary statistics, visualizations (e.g., histograms, scatter plots, box plots), and charts. EDA helps you understand the distribution, relationships, and patterns in the data.

- **Data Transformation and Feature Engineering**:

  Create new variables or features from existing ones to enhance the predictive power of the data. This step might involve normalization, scaling, encoding categorical variables, and creating derived features.

- **Data Modeling:**

  Apply statistical or machine learning models to the preprocessed data to extract insights. Choose models based on the problem, such as regression for predicting values or classification for categorizing data.

- **Model Training and Validation:**

  Split the data into training and validation sets. Train the model on the training data and validate its performance using the validation data. This step helps you assess the model's accuracy and generalization to new data.

- **Model Evaluation:**

  Evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, F1-score, etc. Compare the model's results to the goals you defined earlier.

- **Interpretation of Results:**

  Analyze the model's output to gain insights into the relationships between variables, the significance of features, and how they contribute to the outcomes.

- **Visualization and Reporting:**

  Create visualizations and reports to communicate your findings effectively. Visualizations help stakeholders understand complex data patterns and trends.

- **Draw Conclusions:**

  Based on the analysis and interpretation, draw conclusions about the problem you initially set out to solve. Address questions like whether the goals were achieved, what patterns were discovered, and what actionable insights can be derived.

- **Decision-Making and Implementation:**

  Use the insights gained to make informed decisions, whether it's optimizing a process, refining a strategy, or taking specific actions based on the analysis.

- **Iterative Process:**

  Data analysis is often an iterative process. You might need to revisit and refine earlier steps based on new insights or changing requirements.

4.2 Modules Flow diagrams

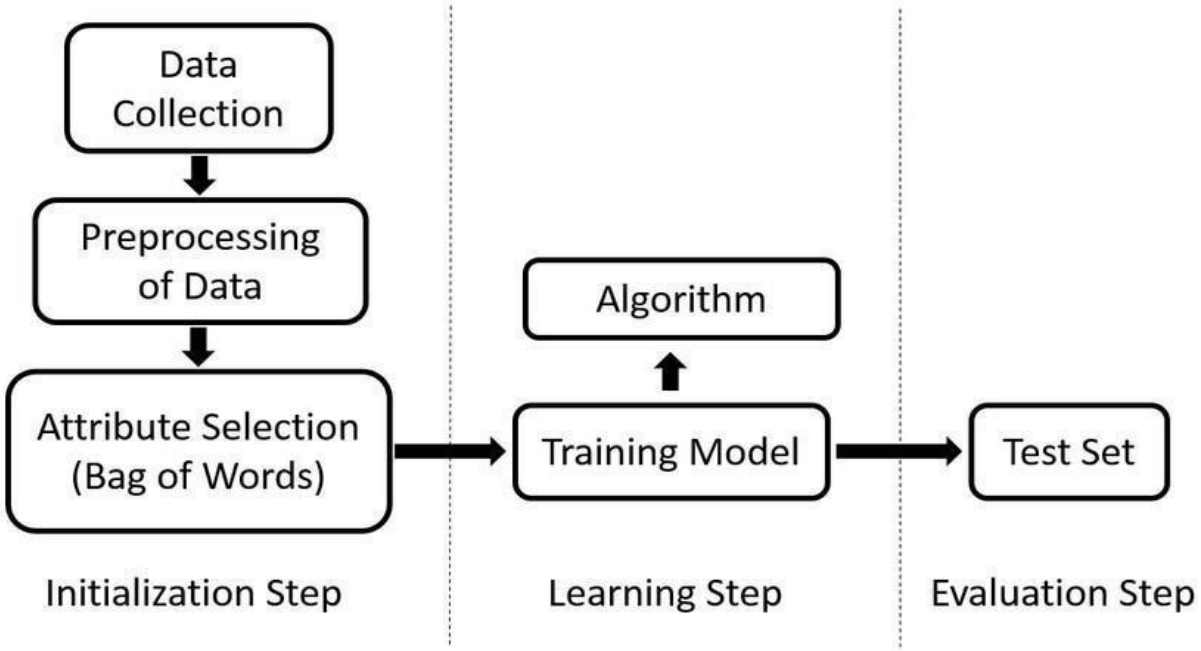


FIG.No.4.2.1.Flow Diagrams

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 Introduction

Dataset plays a crucial role in the Data analysis.So we should be careful while choosing the dataset for the project.

Dataset:

- The dataset utilized in this project is a dataset of the various shows available in the Netflix from the year 2000 to 2021.
- The dataset was downloaded from the following website:
  "https://www.kaggle.com/datasets/shivamb/netflix-shows"
- The dataset Utilized in this data analysis contain significant features to gain the insights of the Netflix The results of the analysis were also informative and insightful.
- The purpose of this dataset is to test my data cleaning and visualization skills. This data set has 7789 rows and 11variables.
- Show _id: unique id of each show.
- Category: show category. Could be either movie or a tv show.
- Title: name of the show.
- Director: name of the director(s) of the show.
- Cast: the group of actors who make up a show.
- Country: countries where the show was added on Netflix.
- Release_Date: date when the show was added on Netflix.
- Rating: show rating on Netflix.
- Duration: time duration of the show.
- Genre: genre of the show.
- Description: About the show.

5.2.1 Implementation of key functions

- Data preprocessing -It is most important process. Mostly data contains missingvale and other impurities that can cause effective- ness of data. To improve qualityand effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset, we need to perform preprocessing in two steps
- Missing Values removal- Remove all the instances that have zero (0) as worth. Having zeros worth is not possible. Therefore, this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this processis called features subset selection, which reduces dimensionality of data and helpto work.
- Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically,aim of normalization is to bring all the attributes under same scale.

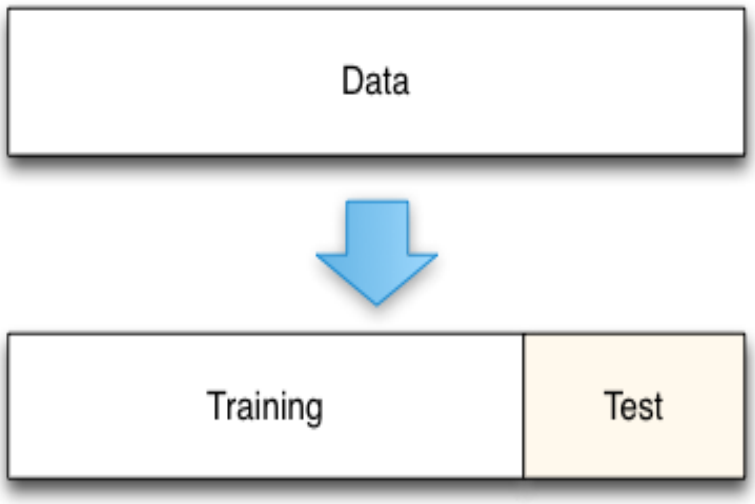


FIG.No.5.2.1. Data Splitting

Splitting of data:

The Processed data is splitted into the training set and testing set based on the size of the dataset

Apply Machine Learning:

- When data has been ready, we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict category. The methods applied Netfix dataset.
- Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.

The commands that used in this project:

- head()-It shows the first N rows in the data (by default, N=5).
- tail 0-It shows the last N rows in the data (by default, N=5).
- shape-It shows the total no. of rows and no. of columns of the dataframe.
- size-To show No. of total values(elements) in the dataset.
- columns-To show each Column Name.
- dtypes-To show the data-type of each column.
- info()-To show indexes, columns, data-types of each column, memory at once.
- value_counts - In a column, it shows all the unique values with their count. It can be applied on a single column only.
- unique()-It shows the all unique values of the series
- nunique() - It shows the total no. of unique values in the series.
- duplicated()-To check row wise and detect the Duplicate rows.
- isnull()-To show where Null value is present.
- dropna()-It drops the rows that contains all missing values.
- isin()-To show all records including particular elements.
- str.contains()-To get all records that contains a given string. str.split()-It splits a column's string into different columns.
- to datetime()-Converts the data-type of Date-Time Column into datetimelna] datatype.
- dt.year.value_counts()-It counts the occurrence of all individual years in Time column.

- groupby()-Groupby is used to split the data into groups based on some criteria.
- sns.countplot(df[Col_name])-To show the count of all unique values of any column in the form of bar graph.
- max(), min() - It shows the maximum/minimum value of the series.
- mean() - It shows the mean value of the series

CATBOOST :

CatBoost is a machine learning algorithm developed by Yandex, specifically designed for categorical feature support and efficient handling of large datasets. It is particularly popular in tasks such as classification, regression, and ranking.

Advantages of CatBoost:

- **Categorical Feature Support:**

CatBoost efficiently handles categorical features without the need for pre-processing like one-hot encoding. This simplifies the data preparation phase and often results in better model performance.

- **Robust to Overfitting:**

CatBoost includes built-in regularization techniques, such as gradient-based and ordered boosting, which help prevent overfitting. This is especially useful when working with complex datasets.

- **Handling of Missing Data:**

CatBoost has a robust method for handling missing data, reducing the need for imputation techniques. The algorithm can naturally handle missing values during the training process.

- **High Performance:**

CatBoost is designed for high performance. It is optimized for speed and memory efficiency,

- **Built-in Cross-Validation:**

  CatBoost includes a built-in cross-validation feature, simplifying the process of model evaluation and hyperparameter tuning. This is crucial for ensuring the robustness of the model.

- **Effective for Time-Series Data:**

  CatBoost performs well on time-series data, making it suitable for applications such as financial forecasting, stock price prediction, and other time-dependent tasks.

- **User-Friendly Parameters:**

  CatBoost provides a relatively small set of hyperparameters, which makes it more user-friendly for practitioners, especially those who are less experienced in machine learning.

- **Wide Range of Applications:**

  CatBoost has been successfully applied in various domains, including finance, healthcare, marketing, and more. Its versatility makes it a valuable tool for a wide range of problem.

USES OF DATA ANALYSIS LIBRARY:

The Libraries plays a vital role in the process of data analysis

- Libraries provide a centralized location for storing and organizing data.This makes it easy to access and analyze data,and it also helps to ensure that data is consistent and accurate.
- Libraries offer a variety of tools and resourses that can be used to analyze data.These tools can inclue statistical software ,data visualization tools,and machine learning algorithms.
- Libraries can provide a community of experts who can help with data analysis .This community of experts who can help with data analysis.

- It can provide the guidance and support ,and it can also help to identify new and innovative ways to analyze data.
- Overall ,Libraries are an essential resource for data analysis .It provide centralized location for storing and organizing data,they offer a variety of tools and resources ,and they provide a community of experts who can help with data analysis.

Need for libraries in data analysis:

- **Efficiency and Speed**

  Libraries are often developed by experts to optimize and streamline common data analysis tasks. They are typically implemented in efficient programming languages (such as Python or R) and are well-tested for performance, which can significantly speed up data processing and analysis.

- **Reusable code**

  Libraries encapsulate complex algorithms and methods into reusable functions and classes. Analysts don't need to reinvent the wheel each time they perform a specific analysis task, as they can leverage existing libraries to quickly achieve their goals .

- **Complex Algorithms and Models**

  Data analysis often involves complex statistical analyses, machine learning algorithms, and modeling techniques. Libraries provide ready-to-use implementations of these methods,saving analysts the effort of implementing them from scratch.

- **Data Manipulation and Cleaning**

  Data rarely comes in a clean and structured format. Libraries like Pandas in Python provide powerful tools for data cleaning, transformation, and manipulation, making it easier to prepare data for analysis.

- **Visualization**

  Data visualization is essential for gaining insights and communicating results effectively. Libraries like Matplotlib, Seaborn, and ggplot2 provide a wide range of plotting options to help analysts create informative visualizations.

- **Domain -Specific Tasks**

  Libraries can cater to specific domains and industries. For example, the Bioconductor library in R focuses on genomics and bioinformatics, while NLTK and SpaCy are libraries for natural language processing tasks.

- **Community Contributions**

  Open-source libraries benefit from contributions by a global community of developers, researchers, and practitioners. This collective effort leads to more robust and well maintained tools.

- **Interoperability**

  Libraries are often designed to work well with each other. This means that you can combine different libraries to take advantage of their specific strengths in various aspects of data analysis .

- **Learning Resources**

  Many libraries have extensive documentation, tutorials, and examples that aid analysts in learning how to use them effectively. This lowers the barrier to entry for newcomers in the field of data analysis.

- **Consistency and Best practices**

  Reputable libraries are developed following best practices and coding standards. This promotes consistent and reliable analysis processes across projects.

- **Scalability**

  Some libraries are optimized to handle large datasets and can take advantage of parallel processing and distributed computing resources, which is crucial when working with big data.

- **Update and Enhancements**

  Libraries are actively maintained and updated by their communities, ensuring that they stay relevant with new technologies, methodologies, and data analysis challenges.

NumPy Library:

NumPy is a library that provides support for mathematical operations on arrays. It is the fundamental package for scientific computing with Python. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is also the basis for many other libraries, such as Pandas, which we will discuss later.
In data analysis, NumPy is used to manipulate, analyze, and manipulate large sets of numerical data. It provides a fastand efficient way to perform mathematical operations on arrays, and it is widely used in scientific computing, machine learning, and data science.
Here are some of the main features of NumPy:

- Fast vectorized operations: NumPy provides a fast and efficient way to perform mathematical operations on arrays. It is designed to work with arrays of any dimension, and it provides a wide range of mathematical functions that can be applied to arrays.
- Linear algebra: NumPy provides a linear algebra library that can be used to perform linear algebra operations on arrays. It provides functions for solving systems of linear equations and eigenvalue problems, among others.

- Random number generation: NumPy provides a random number generator that can be used to generate random numbers. It can be used to simulate real-world events, such as the results of dice rolls, and it can be used to generate statistical distributions.
- Broadcasting: NumPy provides a way to perform operations on arrays of different shapes. It can automatically broadcast arrays of different shapes to make them compatible for mathematical operations. NumPy is a powerful tool for data analysis, and it is essential for anyone working with large sets of numerical data.

MATPLOT LIBRARY:

Matplotlib is a library that provides support for creating 2D plots and graphics. It is the most commonly used library for data visualization in Python. It provides a wide range of plot types, and it can be used to create complex 3D plots and animations.

In data analysis, Matplotlib is used to create visualizations of data. It can be used to create line plots, scatter plots, bar charts, and other types of plots. It can also be used to create histograms, heat maps, and other types of visualizations.

Here are some of the main features of Matplotlib:

- Easy to use interface: Matplotlib provides a simple and easy-to-use interface for creating plots. It provides functions for creating various types of plots, such as line plots, scatter plots, and bar charts.
- Customizable plots: Matplotlib provides a wide range of options for customizing plots. It can be used to set the size, color, and other properties of plots.
- High-quality output: Matplotlib provides high-quality output, both in terms of appearance and in terms of speed. It can be used to create publication-quality graphics.
- Interactive plots: Matplotlib can be used to create interactive plots, such as zooming and panning. It can also be used to create animations.

Matplotlib is a powerful tool for creating visualizations, and it is essential for anyoneworking with data

PANDAS LIBRARY:

Pandas is a library that provides support for data manipulation and analysis in Python. It is built on top of NumPy, which is a library for scientific computing. Pandas provides a number of features that make it well-suited for data analysis, including:

- DataFrames: A DataFrame is a tabular data structure that can be used to store and manipulate data. It is similar to a spreadsheet, but it is more powerful and flexible.
- Time Series: Pandas provides support for time series data. This makes it easy to analyze data that changes over time.
- Data Visualization: Pandas provides a number of tools for visualizing data. This can be helpful for identifying patterns and trends in data.
- Pandas is a powerful tool for data analysis. It is easy to use and it provides a number of features that make it well-suited for this task.

Here are some examples of how Pandas can be used for data analysis:

- Cleaning data: Pandas can be used to clean data. This can involve removing duplicate rows, dealing with missing values, and correcting errors.
- Transforming data: Pandas can be used to transform data. This can involve changing the format of data, or creating new variables from existing data.
- Analyzing data: Pandas can be used to analyze data. This can involve performing statistical tests, or creating visualizations**.**

Pandas is a powerful tool for data analysis. It is easy to use and it provides a number of features that make it well-suited for this task.

SEABORN LIBRARY:

Seaborn is a library for statistical data visualization in Python. It builds on top of Matplotlib, and it provides a number of features that make it easy to create beautiful and informative data visualizations.

Here are some of the main features of Seaborn:

- Built-in themes: Seaborn provides a number of built-in themes that can be used to customize the appearance of plots.
- Easy to use interface: Seaborn provides a simple and easy-to-use interface for creating plots. It provides functions for creating various types of plots, such as scatter plots, line plots, and bar charts.
- Customizable plots: Seaborn provides a wide range of options for customizing plots. It can be used to set the size, color, and other properties of plots.
- High-quality output: Seaborn provides high-quality output, both in terms of appearance and in terms of speed. It can be used to create publication-qualityaphics.

- Interactive plots: Seaborn can be used to create interactive plots, such as zooming and panning. It can also be used to create animations.

Seaborn is a powerful tool for creating statistical data visualizations, and it is essential for anyone working with data.

5.3 Method of Implementation

- IMPORTING LIBRARIES

```
import pandas as pd
data=pd.read_csv(r"/content/Netflix Dataset.csv")
```

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from scipy import stats
import pylab
import warnings # Used to supressed the warnings
warnings.filterwarnings('ignore')
```

- DATASET

1. head()

```
In [3]: data.head()                    # to show top-5 records of the dataset
```

Out[3]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	07:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

- BASIC INFORMATION OF DATA

```
[ ] data.shape

    (7789, 11)

[ ] data.size

    85679

[ ] data.columns

    Index(['Show_Id', 'Category', 'Title', 'Director', 'Cast', 'Country',
           'Release_Date', 'Rating', 'Duration', 'Type', 'Description'],
          dtype='object')
```

```
data.dtypes

Show_Id         object
Category        object
Title           object
Director        object
Cast            object
Country         object
Release_Date    object
Rating          object
Duration        object
Type            object
Description     object
dtype: object
```

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7789 entries, 0 to 7788
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Show_Id       7789 non-null   object
 1   Category      7789 non-null   object
 2   Title         7789 non-null   object
 3   Director      5401 non-null   object
 4   Cast          7071 non-null   object
 5   Country       7282 non-null   object
 6   Release_Date  7779 non-null   object
 7   Rating        7782 non-null   object
 8   Duration      7789 non-null   object
 9   Type          7789 non-null   object
 10  Description   7789 non-null   object
dtypes: object(11)
memory usage: 669.5+ KB
```

- IS THERE ANY DUPLICATE RECORD IN THE DATASET? IF YES, THEN REMOVE THE DUPLICATE RECORDS.

```
data.duplicated()

0       False
1       False
2       False
3       False
4       False
        ...
7784    False
7785    False
7786    False
7787    False
7788    False
Length: 7789, dtype: bool
```

```
data.drop_duplicates(inplace=True)      #To Remove Duplicates the permenently
```

```
[ ] data[data.duplicated()]
```

Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description

- IS THERE ANY NULL VALUE PRESENT IN ANY COLUMN? SHOW WITH HEAT-MAP.

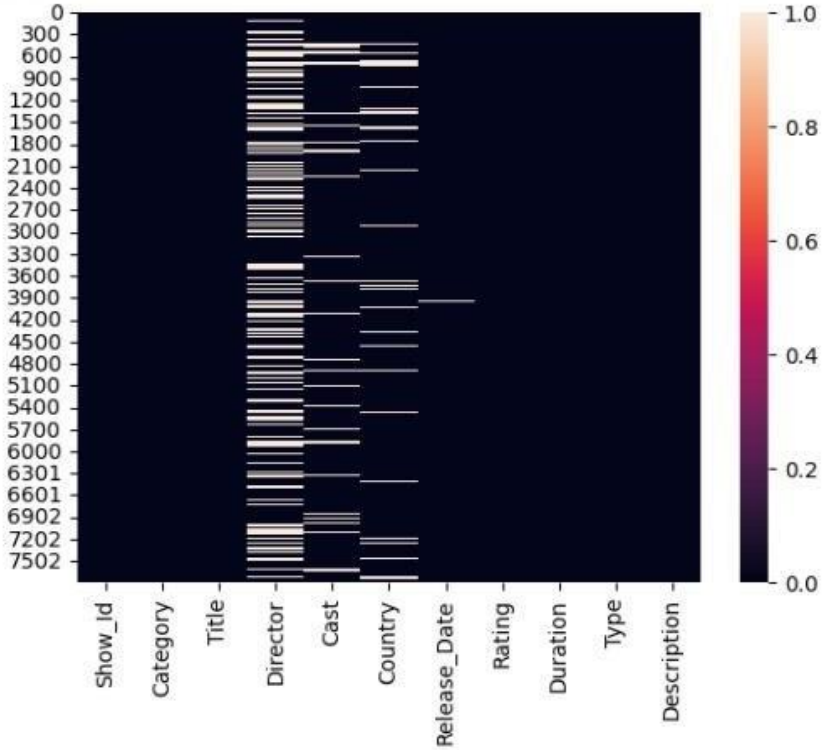
```
[ ]  data.isnull().sum()      #Count of Null values
```

```
Show_Id            0
Category           0
Title              0
Director        2388
Cast             718
Country          507
Release_Date      10
Rating             7
Duration           0
Type               0
Description        0
dtype: int64
```

```
import seaborn as sns     # SEABORN LIBRARY
```

```
[ ]  sns.heatmap(data.isnull())    #Count of Null values using seaborn
```

```
<Axes: >
```



- FOR 'HOUSE OF CARDS,WHAT IS THE SHOW ID AND WHO IS THE DIRECTOR OF THIS SHOW.

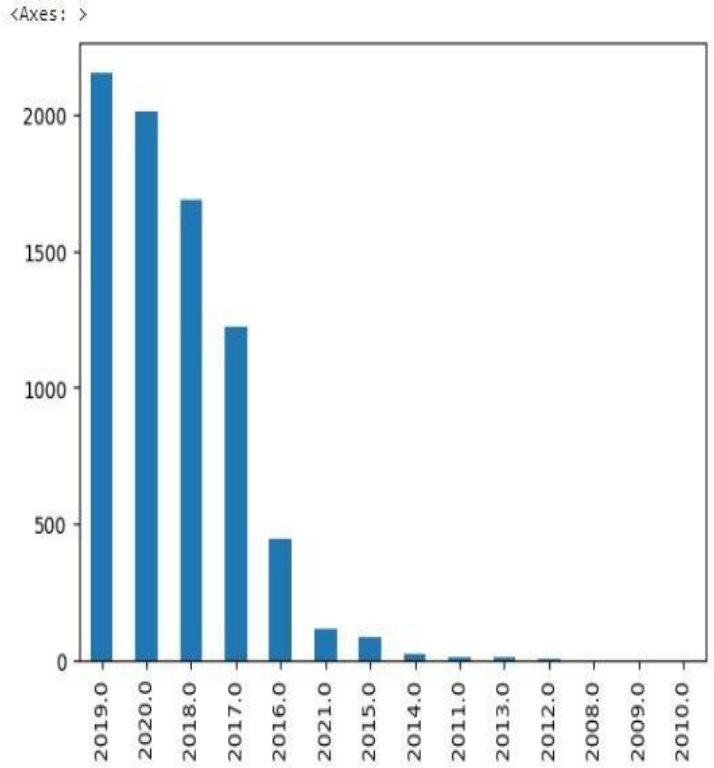
```
In [23]: data[data['Title'].isin(['House of Cards'])]      # To show all records of a particular item in any column
```

Out[23]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description
2832	s2833	TV Show	House of Cards	Robin Wright, David Fincher, Gerald McRaney, J...	Kevin Spacey, Robin Wright, Kate Mara, Corey S...	United States	November 2, 2018	TV-MA	6 Seasons	TV Dramas, TV Thrillers	A ruthless politician will stop at nothing to ...

- IN WHICH YEAR HIGHEST NUMBER OF THE TV SHOWS &MOVIES WERE REALEASED.SHOW USING BAR GRAPH.

```
[ ] data['Data_N'].dt.year.value_counts().plot(kind='bar')
```



- HOW MANY MOVIES & TV SHOWS ARE IN THE DATASET.

```
[ ] data.groupby('Category').Category.count() #group all unique items of a column and show their count
```

```
Category
Movie      5377
TV Show    2410
Name: Category, dtype: int64
```

- SHOW ALL THE MOVIES THAT WERE REALEASEED IN YEAR 2000.

```
[ ] data[(data['Category']=='Movie') & (data['Year']==2000)]
```

Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	Data_N	Year

- SHOW ONLY THE TITLES OF ALL TV SHOWS THAT WERE RELEASED IN INDIA ONLY.

```
[ ] data[(data['Category']=='TV Show') & (data['Country']=='India')]['Title']
```

```
86          21 Sarfarosh: Saragarhi 1897
132                          7 (Seven)
340                       Agent Raghav
364                       Akbar Birbal
533                 Anjaan: Rural Myths
                        ...
6249                The Creative Indians
6400    The Golden Years with Javed Akhtar
6469               The House That Made Me
7294                         Typewriter
7705                    Yeh Meri Family
Name: Title, Length: 71, dtype: object
```

- SHOW TOP 10 DIRECTORS WHO HAVE HIGHEST NUMBER OF TV SHOWS AND MOVIES TO NETFLIX.

```
[ ]  data['Director'].value_counts()

Raúl Campos, Jan Suter     18
Marcus Raboy               16
Jay Karas                  14
Cathy Garcia-Molina        13
Jay Chapman                12
                           ..
Vibhu Virender Puri         1
Lucien Jean-Baptiste        1
Jason Krawczyk              1
Quinn Lasher                1
Sam Dunn                    1
Name: Director, Length: 4050, dtype: int64
```

- SHOW ALL THE RECORDS,WHERE "CATEGORY IS MOVIE AND TYPE IS COMEDIES' OR 'COUNTRY IS UNITED KINGDOM.

```
data[(data['Category']=='Movie') & (data['Type']=='Comedies') | (data['Country']=='United Kingdom')]
```

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	Data_N	Year
19	s20	Movie	'89	NaN	Lee Dixon, Ian Wright, Paul Merson	United Kingdom	May 16, 2018	TV-PG	87 min	Sports Movies	Mixing old footage with interviews, this is th...	2018-05-16	2018.0
33	s34	Movie	#realityhigh	Fernando Lebrija	Nesta Cooper, Kate Walsh, John Michael Higgins...	United States	September 8, 2017	TV-14	99 min	Comedies	When nerdy high schooler Dani finally attracts...	2017-09-08	2017.0
58	s59	TV Show	100% Hotter	NaN	Daniel Palmer, Melissa Sophia, Karen Williams,...	United Kingdom	November 1, 2019	TV-14	1 Season	British TV Shows, International TV Shows, Real...	A stylist, a hair designer and a makeup artist...	2019-11-01	2019.0
72	s73	Movie	17 Again	Burr Steers	Zac Efron, Leslie Mann, Matthew Perry, Thomas...	United States	January 1, 2021	PG-13	102 min	Comedies	Nearing a midlife crisis, thirty-something Mik...	2021-01-01	2021.0
82	s83	Movie	2036 Origin Unknown	Hasraf Dulull	Katee Sackhoff, Ray Fearon, Julie Cox, Steven...	United Kingdom	December 20, 2018	TV-14	95 min	Sci-Fi & Fantasy	Working with an artificial intelligence to inv...	2018-12-20	2018.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7670	s7669	TV Show	World War II in Colour	NaN	Robert Powell	United Kingdom	August 1, 2017	TV-MA	1 Season	British TV Shows, Docuseries, International TV...	Footage of the most dramatic moments from Worl...	2017-08-01	2017.0
7671	s7670	TV Show	World's Busiest Cities	NaN	Anita Rani, Ade Adepitan, Dan Snow	United Kingdom	February 1, 2019	TV-PG	1 Season	British TV Shows, Docuseries	From Moscow to Mexico City, three BBC journali...	2019-02-01	2019.0
7688	s7687	Movie	XV: Beyond the Tryline	Pierre Deschamps	NaN	United Kingdom	March 18, 2020	TV-14	91 min	Documentaries, Sports Movies	Set against the 2015 Rugby World Cup, this doc...	2020-03-18	2020.0
7725	s7724	Movie	You Can Tutu	James Brown	Lily O'Regan, Jeannettsy Enriquez Borges, Joel...	United Kingdom	December 31, 2017	TV-G	87 min	Children & Family Movies	A gifted young ballet dancer struggles to find...	2017-12-31	2017.0
7740	s7739	TV Show	Young Wallander	NaN	Adam Pålsson, Richard Dillane, Leanne Best, Eli...	United Kingdom	September 3, 2020	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Dramas	An incendiary hate crime stirs civil unrest, f...	2020-09-03	2020.0

485 rows × 13 columns

- IN HOW MANY MOVIES/SHOWS,TOM CRULSE WAS CAST.

```
In [55]: # data_new[data_new['Cast'].str.contains('Tom Cruise')]

         data_new[data_new['Cast'].str.contains('Tom Cruise')]
```

Out[55]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	Date_N	Year
3860	s3861	Movie	Magnolia	Paul Thomas Anderson	Jeremy Blackman, Tom Cruise, Melinda Dillon, A...	United States	January 1, 2020	R	189 min	Dramas, Independent Movies	Through chance, human action, past history and...	2020-01-01	2020.0
5071	s5071	Movie	Rain Man	Barry Levinson	Dustin Hoffman, Tom Cruise, Valeria Golino, Ge...	United States	July 1, 2019	R	134 min	Classic Movies, Dramas	A fast-talking yuppie is forced to slow down w...	2019-07-01	2019.0

- WHAT ARE THE DIFFERENT RATINGS DEFINED BY NETFLIX.

nunique()

```
In [59]: # data.Rating.nunique()
         data.head(2)
```

Out[59]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	Date_N	Year
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...	2020-08-14	2020.0
1	s2	Movie	07:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...	2016-12-23	2016.0

```
In [ ]: data['Rating'].nunique()
```

Out[61]: 14

- HOW MANY MOVIES GOT THE 'TV-14 RATING IN CANADA.

```
data[(data['Category']=='Movie') & (data['Rating']=='TV-14') & (data['Country']=='Canada')].shape
```

```
(11, 13)
```

- HOW MANY TV SHOWS GOT THE 'R' RATING AFTER YEAR 2018.

```
data[(data['Category']=='Movie') & (data['Rating']=='R') & (data['Year']>2018)]
```

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	Data_N	Year
7	s8	Movie	187	Kevin Reynolds	Samuel L. Jackson, John Heard, Kelly Rowan, Cl...	United States	November 1, 2019	R	119 min	Dramas	After one of his high school students attacks ...	2019-11-01	2019.0
14	s15	Movie	3022	John Suits	Omar Epps, Kate Walsh, Miranda Cosgrove, Angus...	United States	March 19, 2020	R	91 min	Independent Movies, Sci-Fi & Fantasy, Thrillers	Stranded when the Earth is suddenly destroyed ...	2020-03-19	2020.0
65	s66	Movie	13 Sins	Daniel Stamm	Mark Webber, Rutina Wesley, Devon Graye, Tom B...	United States	January 13, 2019	R	93 min	Horror Movies, Thrillers	A man agrees to appear on a game show with a $...	2019-01-13	2019.0
68	s69	Movie	14 Blades	Daniel Lee	Donnie Yen, Zhao Wei, Wu Chun, Law Kar-Ying, K...	Hong Kong, China, Singapore	April 3, 2019	R	113 min	Action & Adventure, International Movies	In the age of the Ming Dynasty, Quinglong is t...	2019-04-03	2019.0
83	s84	Movie	20th Century Women	Mike Mills	Annette Bening, Elle Fanning, Greta Gerwig, Lu...	United States	June 28, 2019	R	119 min	Dramas, Independent Movies	In 1979, single bohemian mom Dorothea, hoping ...	2019-06-28	2019.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7659	s7658	Movie	Woodshock	Kate Mulleavy, Laura Mulleavy	Kirsten Dunst, Joe Cole, Pilou Asbæk, Jack Kil...	United States	June 21, 2020	R	101 min	Dramas, Independent Movies, Thrillers	Shattered after her mother's death, a woman fl...	2020-06-21	2020.0
7712	s7711	Movie	Yes, God, Yes	Karen Maine	Natalia Dyer, Timothy Simons, Wolfgang Novogra...	United States	October 22, 2020	R	78 min	Comedies, Dramas, Independent Movies	A devoutly religious teen grapples with her ow...	2020-10-22	2020.0
7738	s7737	Movie	Young Adult	Jason Reitman	Charlize Theron, Patton Oswalt, Patrick Wilson...	United States	November 20, 2019	R	94 min	Comedies, Dramas, Independent Movies	When a divorced writer gets a letter from an o...	2019-11-20	2019.0
7776	s7775	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...	2019-11-20	2019.0
7780	s7779	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...	2019-11-01	2019.0

449 rows × 13 columns

- WHAT IS THE MAXIMUM DURATION OF A MOVIE/SHOW ON NETFLIX.

```
data['Duration'].unique()
```

```
array(['4 Seasons', '93 min', '78 min', '80 min', '123 min', '1 Season',
       '95 min', '119 min', '118 min', '143 min', '103 min', '89 min',
       '91 min', '149 min', '144 min', '124 min', '87 min', '110 min',
       '128 min', '117 min', '100 min', '2 Seasons', '84 min', '99 min',
       '90 min', '102 min', '104 min', '105 min', '56 min', '125 min',
       '81 min', '97 min', '106 min', '107 min', '109 min', '44 min',
       '75 min', '101 min', '3 Seasons', '37 min', '113 min', '114 min',
       '130 min', '94 min', '140 min', '135 min', '82 min', '70 min',
       '121 min', '92 min', '164 min', '53 min', '83 min', '116 min',
       '86 min', '120 min', '96 min', '126 min', '129 min', '77 min',
       '137 min', '148 min', '28 min', '122 min', '176 min', '85 min',
       '22 min', '68 min', '111 min', '29 min', '142 min', '168 min',
       '21 min', '59 min', '20 min', '98 min', '108 min', '76 min',
       '26 min', '156 min', '30 min', '57 min', '150 min', '133 min',
       '115 min', '154 min', '127 min', '146 min', '136 min', '88 min',
       '131 min', '24 min', '112 min', '74 min', '63 min', '38 min',
       '25 min', '174 min', '60 min', '153 min', '158 min', '151 min',
       '162 min', '54 min', '51 min', '69 min', '64 min', '147 min',
       '42 min', '79 min', '5 Seasons', '40 min', '45 min', '172 min',
       '10 min', '163 min', '9 Seasons', '55 min', '72 min', '61 min',
       '71 min', '160 min', '171 min', '48 min', '139 min', '157 min',
       '15 min', '65 min', '134 min', '161 min', '62 min', '8 Seasons',
       '186 min', '49 min', '73 min', '58 min', '165 min', '166 min',
       '138 min', '159 min', '141 min', '132 min', '52 min', '67 min',
       '34 min', '66 min', '312 min', '180 min', '47 min', '6 Seasons',
       '155 min', '14 min', '177 min', '11 min', '9 min', '46 min',
       '145 min', '11 Seasons', '7 Seasons', '13 Seasons', '8 min',
       '12 min', '12 Seasons', '10 Seasons', '43 min', '50 min', '23 min',
       '185 min', '200 min', '169 min', '27 min', '170 min', '196 min',
       '33 min', '181 min', '204 min', '32 min', '35 min', '167 min',
       '16 Seasons', '179 min', '193 min', '13 min', '214 min', '17 min',
       '173 min', '192 min', '209 min', '187 min', '41 min', '182 min',
       '224 min', '233 min', '189 min', '152 min', '19 min', '15 Seasons',
       '208 min', '237 min', '31 min', '178 min', '230 min', '194 min',
       '228 min', '195 min', '3 min', '16 min', '5 min', '18 min',
       '205 min', '190 min', '36 min', '201 min', '253 min', '203 min',
       '191 min'], dtype=object)
```

- WHICH INDIVIDUAL COUNTRY HAS THE HIGHEST NO OF TV SHOWS.

```
data_tvshow.Country.value_counts()
```

```
United States                                                705
United Kingdom                                               204
Japan                                                        157
South Korea                                                  147
India                                                         71
                                                            ...
Canada, United States, United Kingdom, France, Luxembourg      1
United States, Italy                                           1
Chile, Italy                                                   1
Canada, United Kingdom                                         1
United States, France, South Korea, Indonesia                  1
Name: Country, Length: 183, dtype: int64
```

- HOW CAN WE SORT THE DATASET BY YEAR.

```
data.sort_values(by='Year',ascending=False).head(2)
```

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	Data_N	Year	Minutes	Unit
5564	s5564	Movie	Sherlock Holmes	Guy Ritchie	Robert Downey Jr., Jude Law, Rachel McAdams, M...	United States, Germany, United Kingdom, Australia	January 1, 2021	PG-13	128 min	Action & Adventure, Comedies	The game is afoot for an eccentric detective w...	2021-01-01	2021.0	128	min
5919	s5919	Movie	Surf's Up	Ash Brannon, Chris Buck	Shia LaBeouf, Jeff Bridges, Zooey Deschanel, J...	United States	January 1, 2021	PG	86 min	Children & Family Movies, Comedies, Sports Movies	This Oscar-nominated animated comedy goes behi...	2021-01-01	2021.0	86	min

- FIND ALL THE INSTANCES WHERE: CATEGORY IS 'MOVIE' AND TYPE IS 'DRAMA OR CATEGORY IS TV SHOW AND TYPE IS 'KID'S TV.

```
data[(data['Category']=='Movie') & (data['Type']=='Dramas') | (data['Category']=='TV Show') & (data['Type']=="Kids' TV")]
```

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	Data_N	Year	Minutes	Unit
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	PG-13	123 min	Dramas	A brilliant group of students become card-coun...	2020-01-01	2020.0	123	min
7	s8	Movie	187	Kevin Reynolds	Samuel L. Jackson, John Heard, Kelly Rowan, Cl...	United States	November 1, 2019	R	119 min	Dramas	After one of his high school students attacks ...	2019-11-01	2019.0	119	min
111	s112	TV Show	44 Cats	NaN	Sarah Natochenny, Suzy Myers, Simona Berman, E...	Italy	October 1, 2020	TV-Y7	2 Seasons	Kids' TV	Paw-esome tales abound when singing furry frie...	2020-10-01	2020.0	2	Seasons
170	s171	Movie	A Family Man	Mark Williams	Gerard Butler, Gretchen Mol, Allison Brie, Will...	Canada, United States	December 15, 2019	R	110 min	Dramas	A ruthless corporate headhunter battles his ri...	2019-12-15	2019.0	110	min
232	s233	Movie	A Stoning in Fulham County	Lary Elkann	Ken Olin, Jill Eikenberry, Maureen Mueller, Gr...	United States	October 1, 2011	TV-14	95 min	Dramas	After reckless teens kill an Amish child, a pr...	2011-10-01	2011.0	95	min
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7668	s7667	TV Show	World of Winx	Iginio Straffi	Haven Paschall, Alysha Deslorieux, Jessica Paq...	Italy, United States	June 16, 2017	TV-Y7	2 Seasons	Kids' TV	The reality show "WOW!" engages the Winx in th...	2017-06-16	2017.0	2	Seasons
7717	s7716	TV Show	Yoko	NaN	Eileen Stevens, Alyson Leigh Rosenfeld, Sarah ...	NaN	June 23, 2018	TV-Y	1 Season	Kids' TV	Friends Mai, Oto and Vik's games at the park b...	2018-06-23	2018.0	1	Season
7719	s7718	TV Show	YOM	NaN	Sairaj Devyani Dagaonkar, Ketan Singh, Mayur ...	NaN	June 7, 2018	TV-Y7	1 Season	Kids' TV	With the mind of a human being, and the body o...	2018-06-07	2018.0	1	Season
7758	s7757	TV Show	Zé	NaN	Apollo Costa, Gabriel Santana, Matheus Lustosa...	Brazil	February 22, 2019	TV-PG	2 Seasons	Kids' TV	Fading music biz veteran Zé realizes he has ju...	2019-02-22	2019.0	2	Seasons
7761	s7760	TV Show	Zak Storm	NaN	Michael Johnston, Jessica Gee-George, Christin...	United States, France, South Korea, Indonesia	September 13, 2018	TV-Y7	3 Seasons	Kids' TV	Teen surfer Zak Storm is mysteriously transpor...	2018-09-13	2018.0	3	Seasons

322 rows × 15 columns

5.4 Output Screens and Result Analysis

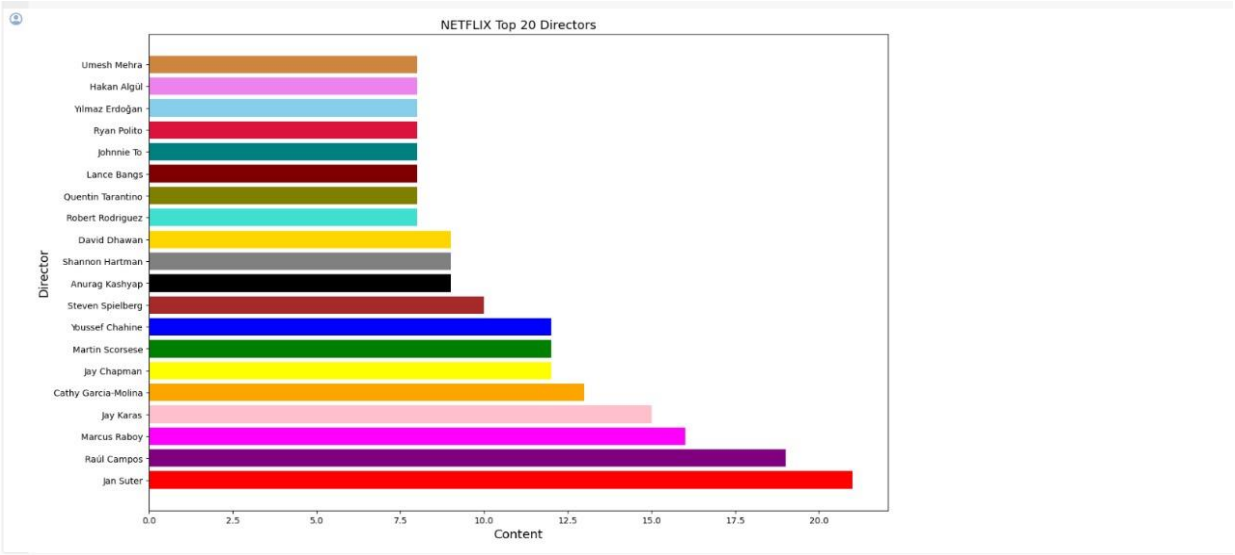


FIG.No.5.4.1.Top 20 Directors

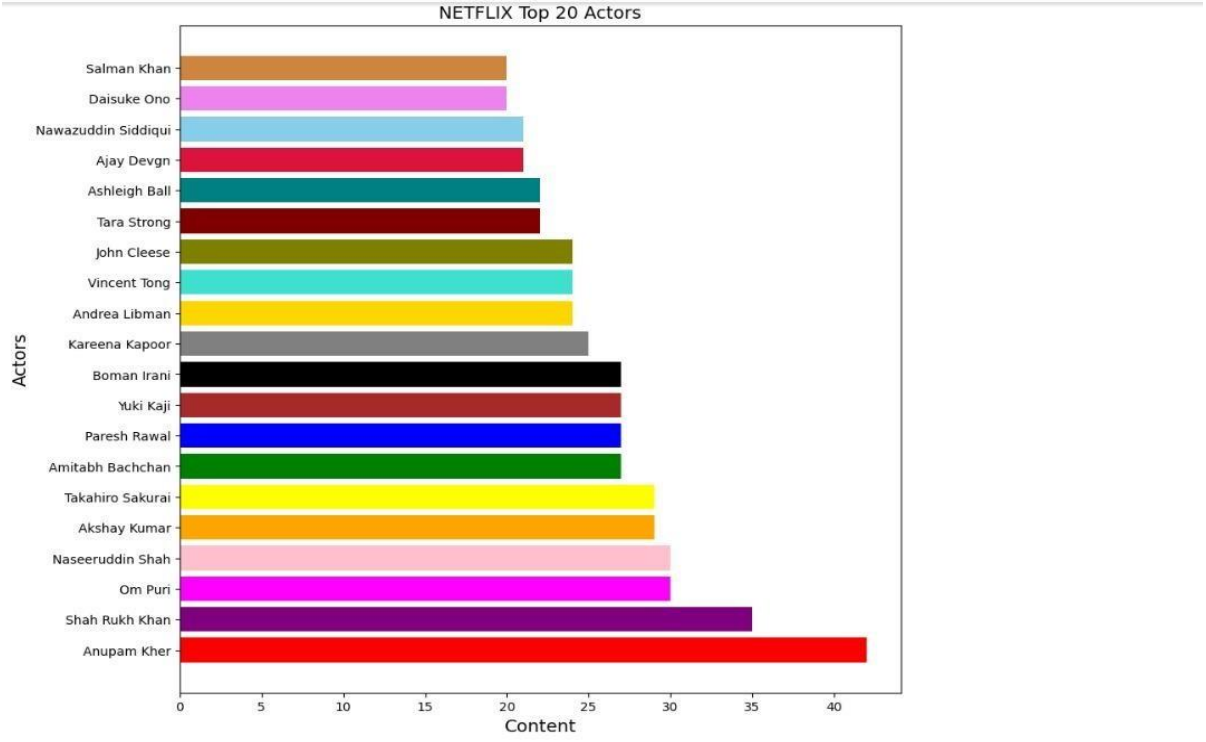


FIG.No.5.4.2.Top 20 Actors

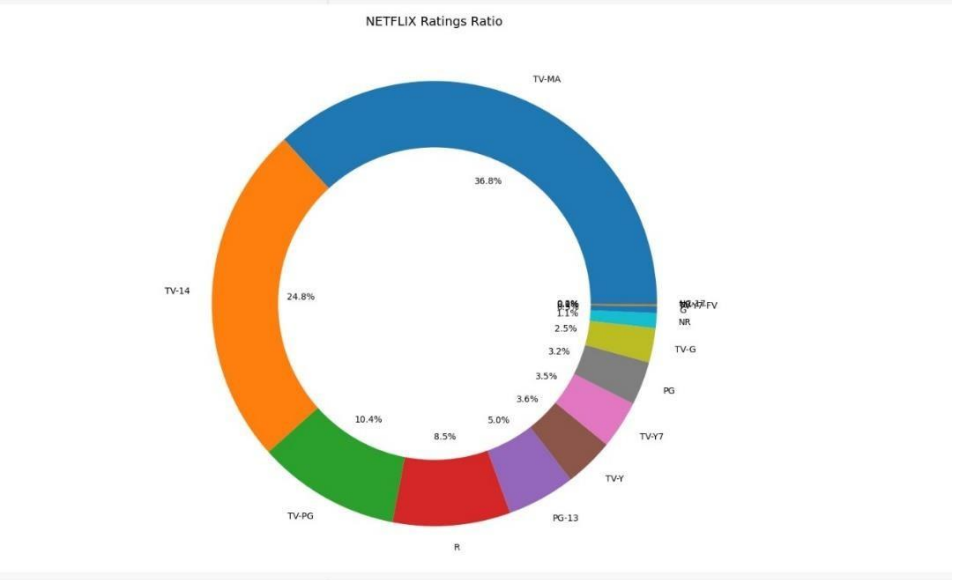


FIG.No.5.4.3.Rating Ratio

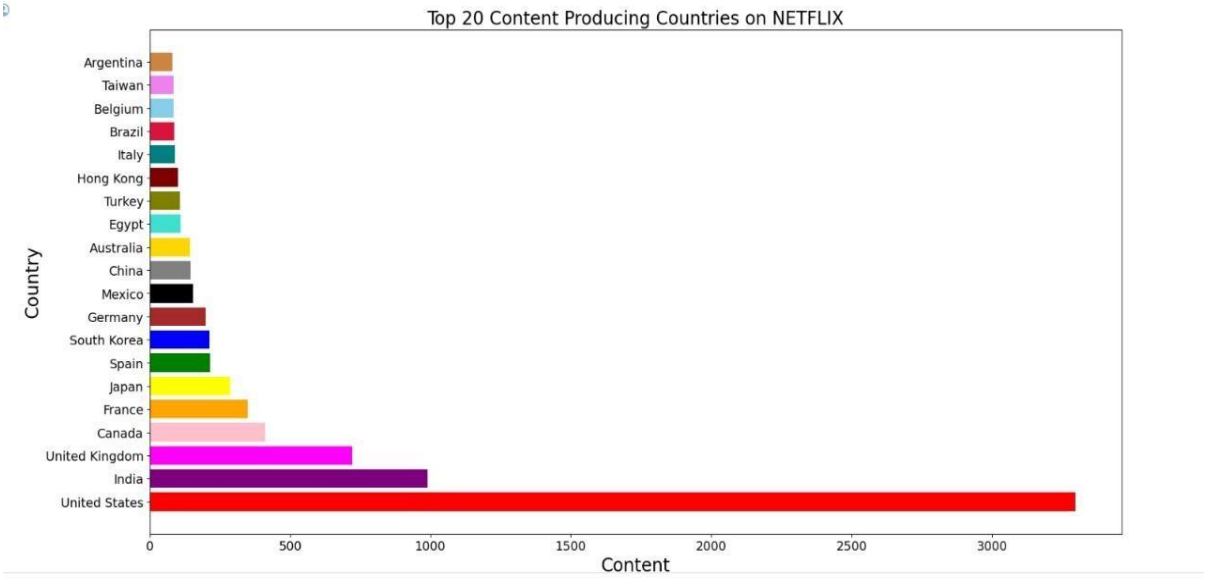


FIG.No.5.4.4.Top 20 content producing countries

5.5 Conclusion

Thus, Exploring datasets of Netflix for Future Release of TV shows and Movies on the Platform. how many movies and TV shows are release in the recent ten years on the platform, and what were the top 10 genres that the audience of the Netflix platform liked the most. From our data analysis we conducted on R markdown, we have discovered that there were a wide variety of genres that movie directors produced worldwide and we have observed many cast members and genres .

CHAPTER 6

TESTING AND VALIDATION

6.1 Introduction

Testing and validation are crucial steps in the development of machine learning models, including those built using the CatBoost algorithm:

- Data Splitting:

- Training Set:

  - Typically, 70-80% of your data is used for training the model.

- Validation Set:

  - 10-15% is allocated for validation to tune hyperparameters and assess model performance during training.

- Test Set:

  - The remaining 10-20% is reserved for final evaluation after model training.

Cross-Validation:

- Implement k-fold cross-validation if you have a limited dataset.

- CatBoost has built-in cross-validation support, allowing you to specify the number of folds (cv) during training.

Hyperparameter Tuning:

- Use the validation set to tune hyperparameters. CatBoost has several hyperparameters, including the depth of the trees, learning rate, and regularization parameters.

- Leverage techniques like grid search or random search to find optimal hyperparameter combinations.

Evaluation Metrics:

- Choose appropriate evaluation metrics based on the nature of your problem

- Classification: Accuracy, Precision, Recall, F1 Score, ROC-AUC.

- Regression: Mean Squared Error (MSE), R-squared.

6.2 Design of Test cases and Scenarios

CODE:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

%matplotlib inline

import seaborn as snstrain=pd.read_csv(r"/content/netflix1.csv")

train=pd.read_csv(r"/content/netflix1.csv")

train.shape

train.columns

pip install pandas scikit-learn matplotlib

import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import LabelEncoder

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy_score, classification_report

import matplotlib.pyplot as plt

import catboost
```

```
import pandas as pd

from catboost import CatBoostClassifier

from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score

pip install catboost

train.dropna(inplace=True)

label_encoders = {}

categorical_columns = ["Category", "Director", "Cast", "Country", "Type", "Rating"]

for col in categorical_columns:

    le = LabelEncoder()

    train[col] = le.fit_transform(train[col])

    label_encoders[col] = le

train["Release_Date"] = pd.to_datetime(train["Release_Date"])

train["Release_Year"] = train["Release_Date"].dt.year

train["Release_Month"] = train["Release_Date"].dt.month

train["Release_Day"] = train["Release_Date"].dt.day

X = train.drop(columns=["Show_Id", "Title", "Duration", "Release_Date"])

y = train["Category"]

def get_categorical_indicies(X):

  cats = []
```

```
  for col in X.columns:
    if is_numeric_dtype(X[col]):
      pass
    else:
      cats.append(col)
    cat_indicies = []
for col in cats:
  cat_indicies.append(X.columns.get_loc(col))
  return cat_indicies
categorical_indicies = get_categorical_indicies(X)
def convert_cats(X):
  cats = []
  for col in X.columns:
    if is_numeric_dtype(X[col]):
      pass
    else:
      cats.append(col)
  cat_indicies = []
  for col in cats:
    X[col] = X[col].astype('category')
```

```
convert_cats(X)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=4,stratify=y)

train_dataset = cb.Pool(X_train,y_train, cat_features=categorical_indicies)

test_dataset = cb.Pool(X_test,y_test, cat_features=categorical_indicies)

model = cb.CatBoostClassifier(loss_function='Logloss', eval_metric='Accuracy')

grid = {'learning_rate': [0.03, 0.1],

'depth': [4, 6, 10],

'l2_leaf_reg': [1, 3, 5,],

'iterations': [50, 100, 150]}

model.grid_search(grid,train_dataset)

model.get_params()
```

6.3 Validation

Accuracy for CatBoost alogorithm

```
pred = model.predict(X_test)
print(classification_report(y_test, pred))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       936
           1       1.00      1.00      1.00        27

    accuracy                           1.00       963
   macro avg       1.00      1.00      1.00       963
weighted avg       1.00      1.00      1.00       963
```

```
[ ] accuracy = accuracy_score(y_test, pred)
    classification_report_str = classification_report(y_test, pred)
```

```
[ ] print(f"Accuracy: {accuracy}")
    print("Classification Report:")
    print(classification_report_str)
```

```
Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       936
           1       1.00      1.00      1.00        27

    accuracy                           1.00       963
   macro avg       1.00      1.00      1.00       963
weighted avg       1.00      1.00      1.00       963
```

6.4 Conclusion

Thus, Exploring new machine learning techniques to make more accurate predictions. Experiment with new data sources and improve data quality. Model building for Netflix Data Analysis using CatBoost algorithm for predicting the category like TV Show, Movie is completed successfully.

CHAPTER 7

CONCLUSION

7.1. Conclusion

Thus, Exploring datasets of Netflix for Future Release of TV shows and Movies on the Platform. In this project, we are going to explore the dataset from Kaggle and, how many movies and TV shows are released in specific time frame, how many movies and TV shows are release in the recent ten years on the platform, and what were the top 10 genres that the audience of the Netflix platform liked the most. From here, we would like to apply a machine learning approach to understand the data fully and provide a great solution where the platform should be headed to. From our data analysis we conducted on R markdown, we have discovered that there were a wide variety of genres that movie directors produced worldwide and we have observed many cast members and genres.

References

Dataset: https://www.kaggle.com/datasets/shivamb/netflix-shows

[1] Netflix.In Wikipedia. Retrieved 09/08/2019, from https://en.wikipedia.org/wiki/Netflix

[2] Linden G, Smith B, York J (2003). Amazon.com Recommendations: Item-to-Item Collaborative Filtering. Published by the IEEE Computer Society, IEEE InternetComput. 7(1):76-80.

[3] Adomavicius G, Tuzhilin A (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEETrans. Knowl. Data Eng. 17(6):734- 749.

[4] Pazzani M.J., Billsus D. (2007) Content-Based Recommendation Systems. In: Brusilovsky P., Kobsa A., Nejdl W. (eds) The Adaptive Web. Lecture Notes in Computer Science, vol 4321. Springer, Berlin, Heidelberg

[5] Balabanovic, M., Shoham Y. (1997). FAB: Content-based, Collaborative Recommendation. Com-munications of the Association for Computing Machinery 40(3), 66-72.