By Kanaya Deas Aditya

Data-driven Insights :

Netflix Movies and TV Shows Classification

Agenda

1. Introduction
2. Netflix Dataset
3. Data Overview
4. Variables
5. Pre-Processing Data
6. EDA
7. Splitting Data
8. Classification Model

Hi there! I'm Naya!

Introduction ♡

I am a Bachelor of Statistics graduate from Sebelas Maret University with a strong passion for data analysis, statistical modeling, and problem-solving. My academic background has equipped me with expertise in data processing, visualization, and machine learning to generate meaningful insights.
I am eager to contribute my skills in a dynamic environment and continuously grow in the field of data analytics and decision-making. Let's connect and collaborate! 🚀

NETFLIX DATASET

Movies and TV Shows

Netflix is a subscription-based video streaming service that offers a variety of TV shows and movies. Netflix has content that comes from its own productions and other parties. Netflix can be accessed on various devices, such as: Smart TVs, Gaming consoles, Streaming media players, Cable set-top boxes, Smartphones, Tablets, Computers.

The Netflix Movies and TV Shows Dataset available on Kaggle is a dataset that contains information about movies and TV shows available on Netflix, including features such as title, genre, release year, duration, director, actors, country of origin, rating, and short description. This dataset is often used for data exploration, visualisation and trend analysis in the streaming industry.

ibimbing

DATA OVERVIEW

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Columns and Description

Variables	Type	Definition
show_id	string	Unique identifier for each show (s1, s2)
type	string	Content type: Movie or TV Show
title	string	The name of the Netflix title
director	string	The director of the title
cast	string	The main actors involved in the title
country	string	The country where the title was produced
date_added	string	The date when the title was added to Netflix
release_year	integer	The year the title was originally released
rating	string	The content rating ("PG-13", "TV-MA")
duration	string	Duration of the movie (in minutes) or the number of seasons for TV shows
listed_in	string	Genre or content category
description	string	The summary description

Pre-Processing Data

MISSING VALUE

	0
show_id	0
type	0
title	0
director	2634
cast	825
country	831
date_added	10
release_year	0
rating	4
duration	3
listed_in	0
description	0

Handling Missing Value data_added, rating, duration

```
[9]  #Drop Baris Missing Value pada Data Added, Rating, Duration
     data.dropna(subset=['date_added','rating','duration'], inplace = True)
```

Handling Missing Value director

```
[13] data['director'].fillna('Unknown',inplace = True)
```

Handling Missing Value cast

```
[14] data['cast'].fillna('Unknown',inplace = True)
```

Handling Missing Value country

```
[15] data['country'].fillna('Unknown',inplace = True)
```

RESULT

	0
show_id	0
type	0
title	0
director	0
cast	0
country	0
date_added	0
release_year	0
rating	0
duration	0
listed_in	0
description	0

```
[17] data.shape
```

```
(8790, 12)
```

CONVERT TYPE COLUMN

Mapping is used to convert data from one format to another, such as converting from categorical to numerical data.
Mapping helps transform data to make it more ready for use in machine learning models.

```
[28] #Mapping
     data['type'] = data['type'].map({'Movie': 1, 'TV Show': 0})
     data.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	1	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	0	Blood & Water	Unknown	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	0	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	0	Jailbirds New Orleans	Unknown	Unknown	Unknown	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	0	Kota Factory	Unknown	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

SPLITTING THE DURATION COLUMN

```
[29] #Movie
     data['duration_minutes'] = data['duration'].apply(lambda x: int(x.split()[0]) if isinstance(x, str) and 'min' in x else np.nan)

[30] #TV Shows
     data['season_count'] = data['duration'].apply(lambda x: int(x.split()[0]) if isinstance(x, str) and 'Season' in x else np.nan)

[31] #Drop duration column
     data.drop(columns=['duration'], inplace = True)
```

Splitting the duration column is used to separate duration minutes from seasons and convert them to integers. Numerical column 'Duration minutes' for movies and numercial 'Number of Seasons' for TV Shows

Missing Values

duration_minutes	2664
season_count	6126

FILLING MISSING VALUES

```
[34] #Filling missing values
     data['duration_minutes'].fillna(data['duration_minutes'].mean(), inplace  = True)
     data['season_count'].fillna(data['season_count'].mean(), inplace = True)
```

EXPLORATORY DATA ANALYSIS

Conclusion from the descriptive statistics table

Descriptive Statistics

Pie chart between movies and TV shows

Movies and TV Shows Distribution

Bar chart of most popular directors

Top 10 The Director of The Title

Bar chart of actors starring in netflix content

Top 5 The Main Actors Involved in The Title

```
[18] #Statistik deskriptif untuk data kategorikal
     data.describe(include='object')
```

	show_id	type	title	director	cast	country	date_added	rating	duration	listed_in	description
count	8790	8790	8790	8790	8790	8790	8790	8790	8790	8790	8790
unique	8790	2	8790	4527	7679	749	1765	14	220	513	8758
top	s1	Movie	Dick Johnson Is Dead	Unknown	Unknown	United States	January 1, 2020	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	6126	1	2621	825	2809	109	3205	1791	362	4

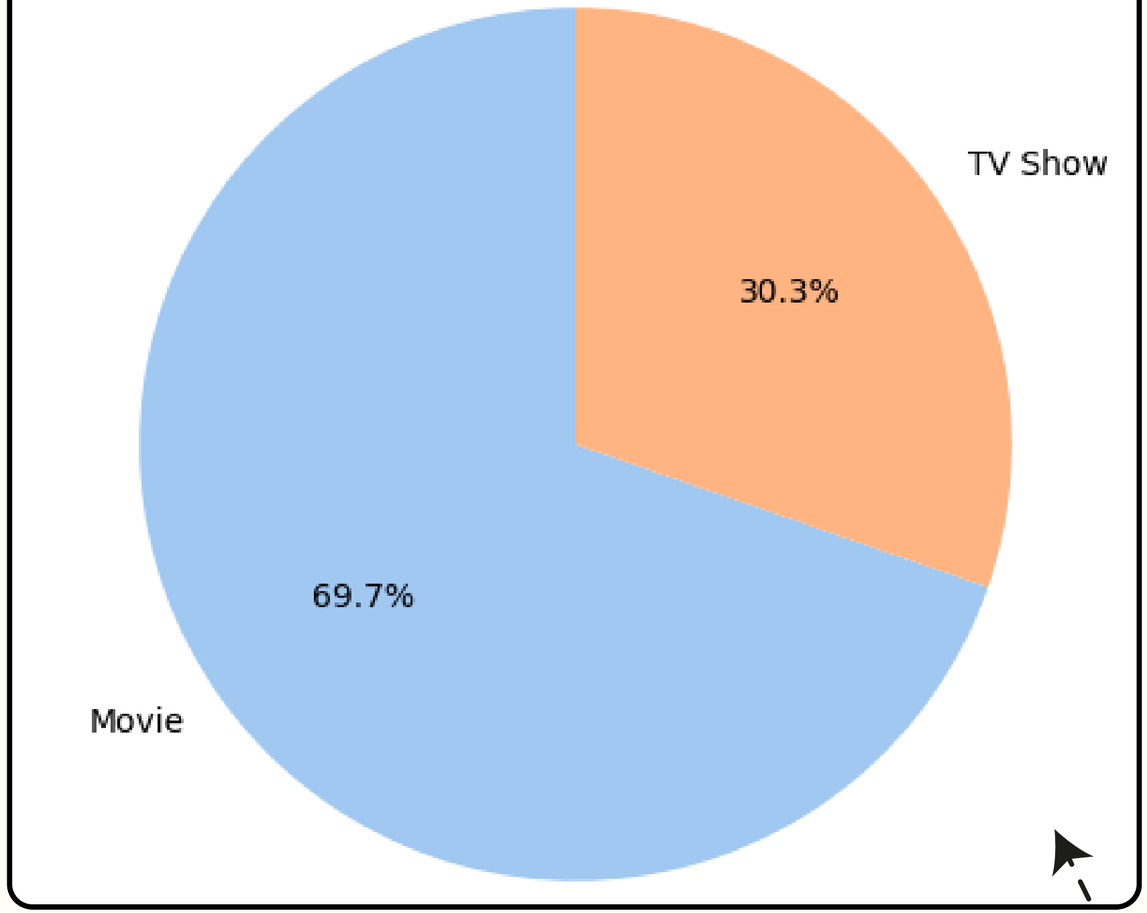
Descriptive Statistics

Descriptive statistics for categorical data aims to analyse the distribution, proportion and frequency of categories in a dataset. Some common techniques used include frequency, mode, proportion, and visualisation.

1. Dataset contains Movie (6126) and TV Show (2664)
2. USA is the country with the highest amount of content at 2809
3. The most common rating is TV-MA with 3205
4. The most popular genre is Dramas, International Movies with 362 titles

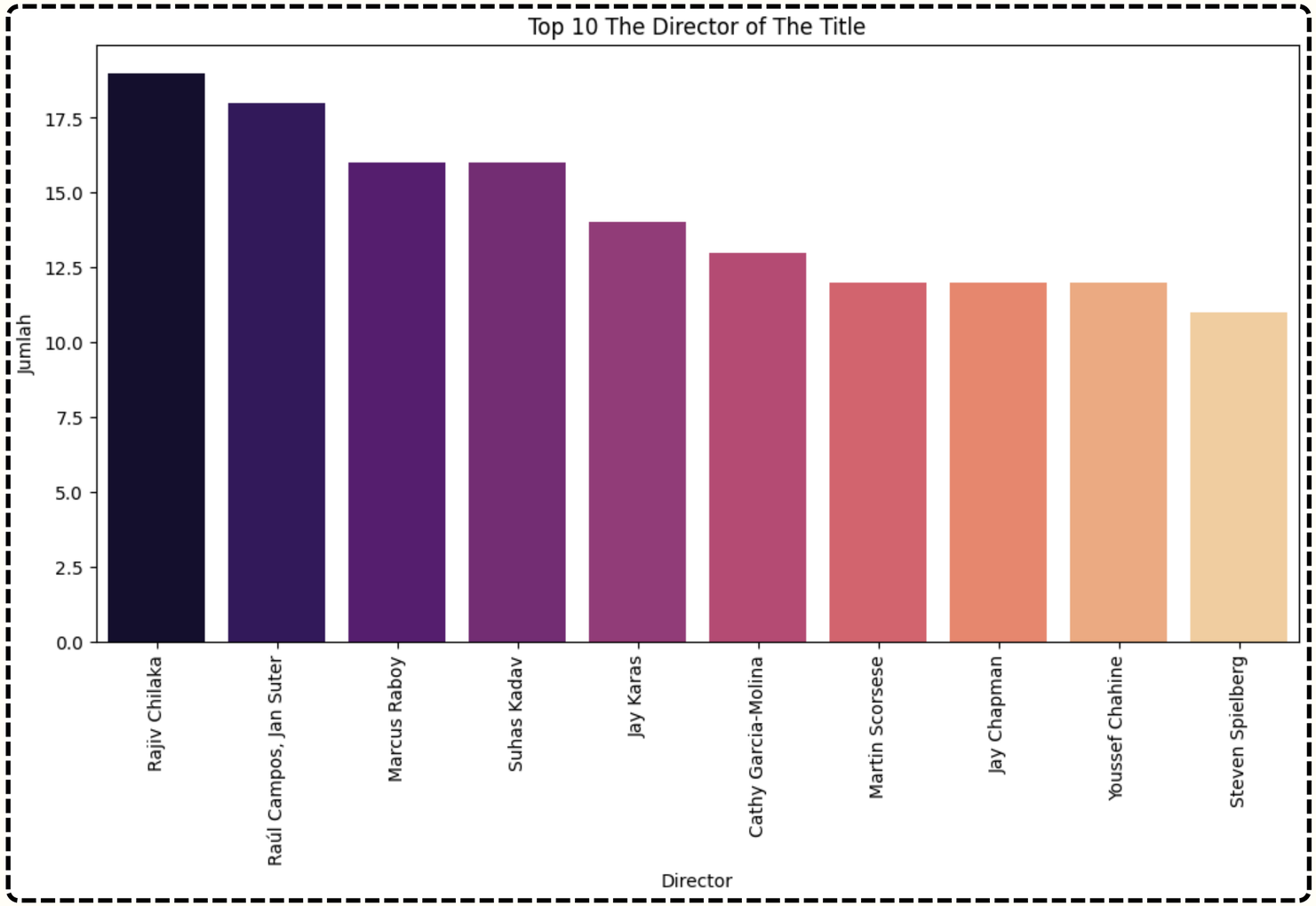
Movies and TV Shows Distribution

The diagram shows the percentage of Movies and TV Shows in the Netflix dataset. It can be seen that the result of the Movie percentage is 69.7% while for TV Shows it is 30.3%. This shows that the film content in the Netflix data is more dominating.

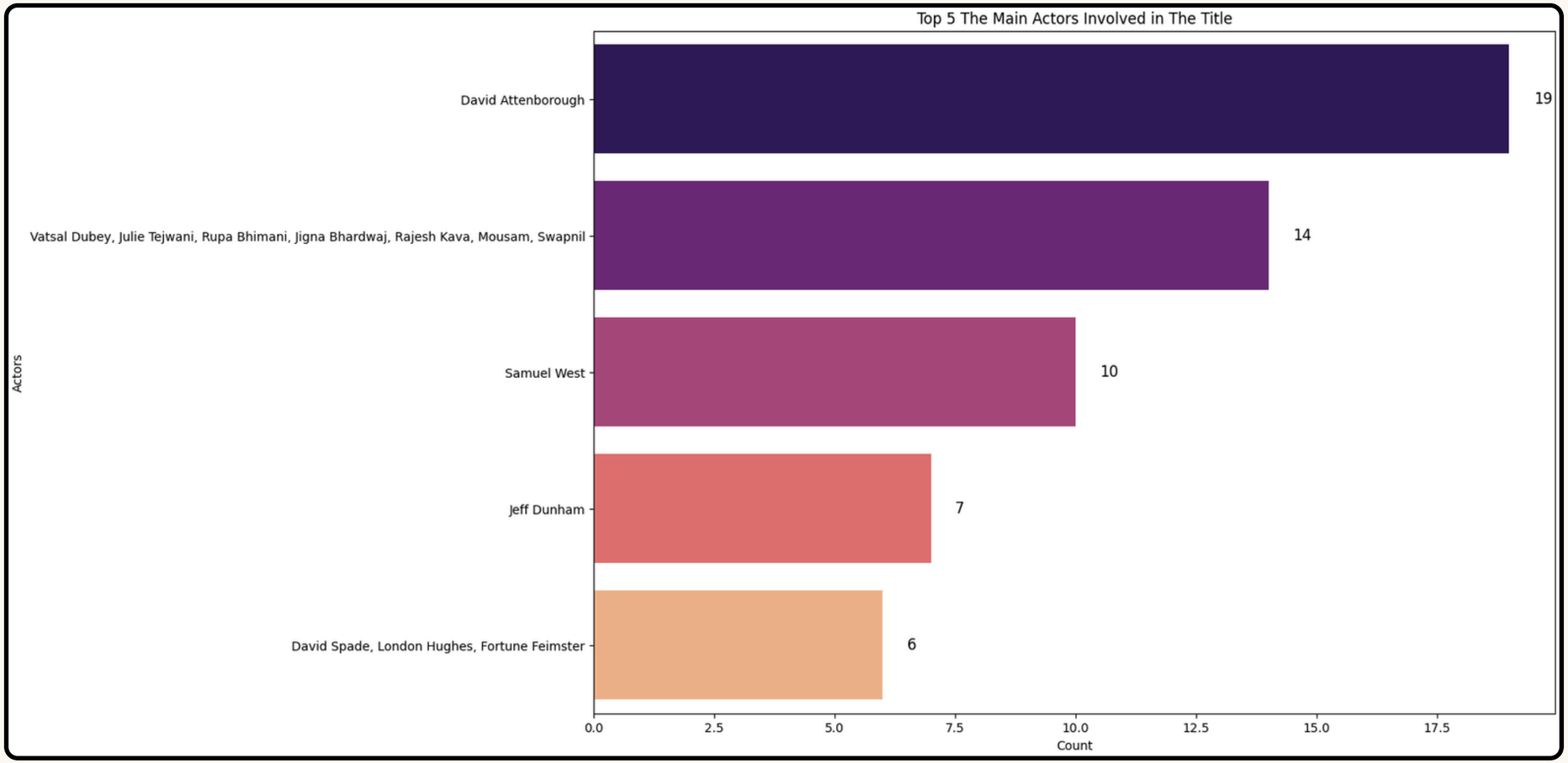


Top 10 Director!

Out of 8790 titles, there are 2621 Netflix contents where the director is unknown. Followed by Rajiv Chilaka with 19 contents, Raúl Campos and Jan Suter with 18 contents, and so on.
In the descriptive statistics, we can see that there are 4527 director names whose content is featured on Netflix.



Top 5 The Main Actors



Actor David Attenborough has the most projects on Netflix with 19 titles that he has starred in. Followed by others with similar numbers. This indicates that Netflix has a relatively even distribution of actors.

Split Data and Classification

SPLIT DATA

Training and Testing

In machine learning, the dataset needs to be divided into training data and testing data so that the model can be learnt and tested properly.

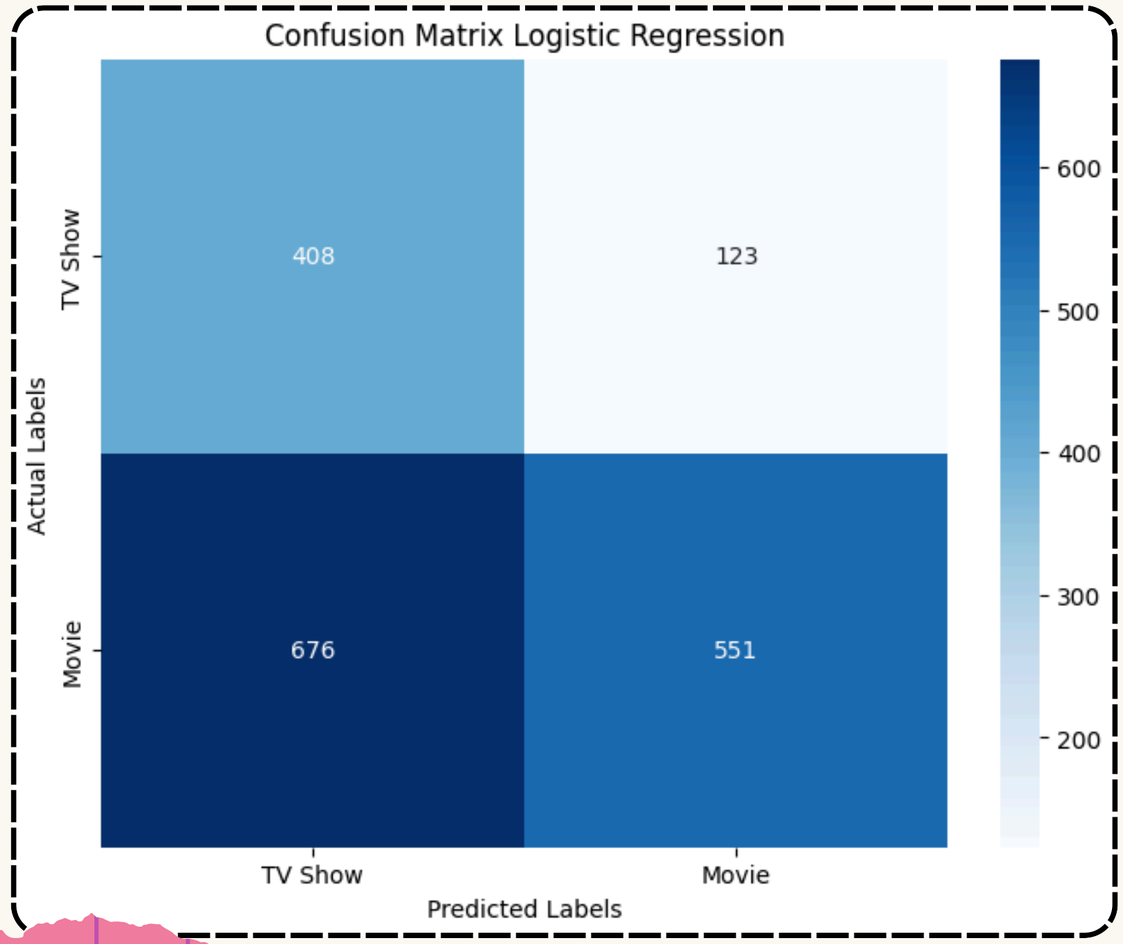
✅ Training Set → Used to train the model to recognise patterns in the data.

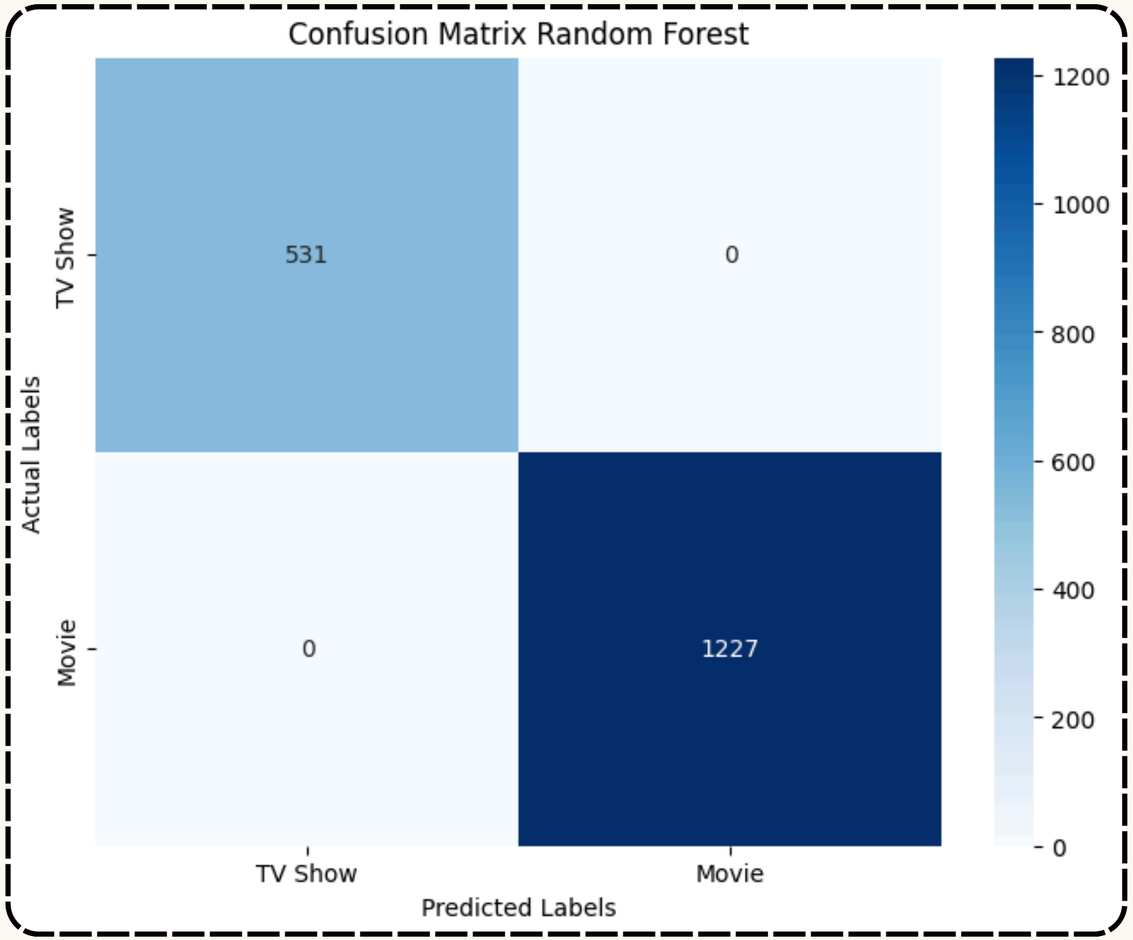
✅ Testing Set → Used to measure the model's performance on data that has never been seen before.

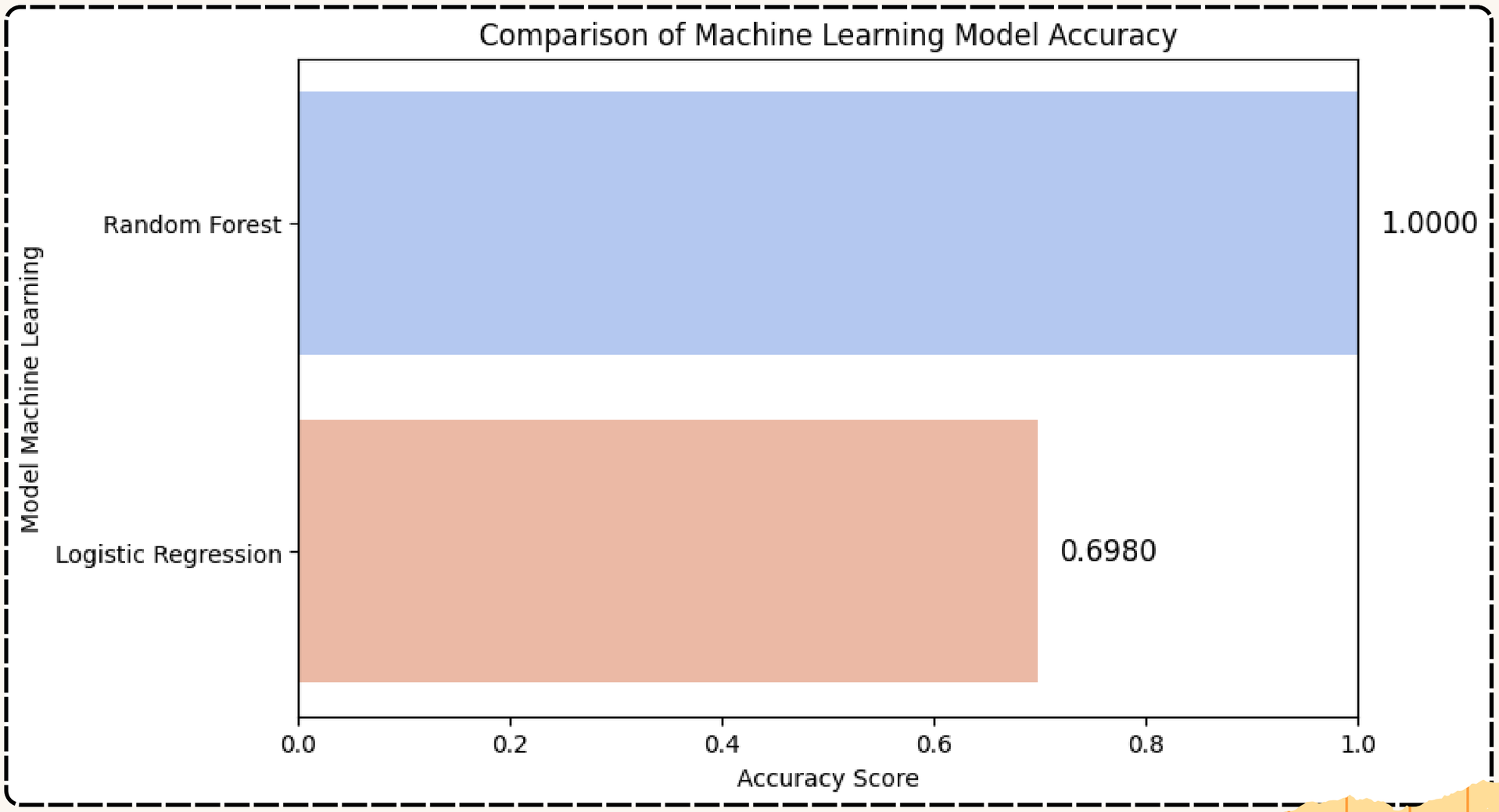
If not separated, the model can experience overfitting, which is over memorising the training data and failing to predict new data.

```
[38] #The target variable
     y = data['type']

[39] #Variabel yang relevan bertipe integer
     X = data[['release_year','duration_minutes','season_count']]

[40] from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

[41] print(f'X_train : {X_train.shape}')
     print(f'X_test : {X_test.shape}')
     print(f'y_train : {y_train.shape}')
     print(f'y_test : {y_test.shape}')

     X_train : (7032, 3)
     X_test : (1758, 3)
     y_train : (7032,)
     y_test : (1758,)
```

Confusion Matrix Comparison






Comparison of Machine Learning Model Accuracy
Model Machine Learning
Random Forest
Logistic Regression
1.0000
0.6980
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy Score

CONCLUSIONS AND SUGGESTIONS

CONCLUSION 1

Netflix produces more movies than TV shows, and most of them come from the United States.

CONCLUSION 2

Random Forest tends to have higher accuracy than Logistic Regression, as it utilises ensemble learning techniques.

CONCLUSION 3

Further development could include more complex models, such as Natural Language Processing (NLP) to analyse movie descriptions and Clustering (K-Means) to group content by genre or duration.

THANK YOU!

kanaya7a@gmail.com
Kanaya Deas Aditya

link google colab: https://bit.ly/ColabNetflixDataset