Netflix prediction

Nouf Faisal Alghamisi

Objective

The purpose of this research was to apply prediction models to predict the trend of watching. I worked with data given by Netflix to get promising results for this multiclass challenge by combining different types of modeling and categorical feature engineering with a random forest, KNN Algorithms and Logistic Regression.

Summary of the Project

- Exploring the Data
- Check the variables that contains missing values
- Clean and Correct the missing values
- Exploratory Data Analysis (EDA) and Visualization
- Data Cleaning and Preprocessing with Pandas for modification
- Feature Selection
- Predictive Machine Learning Modeling

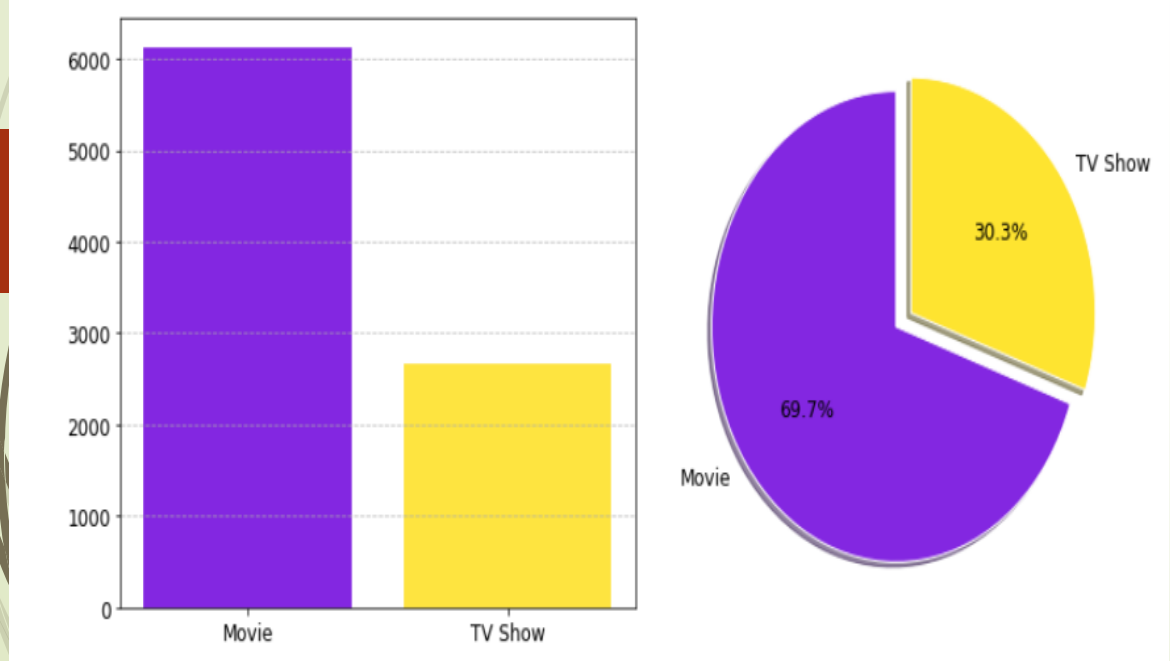
Data Description

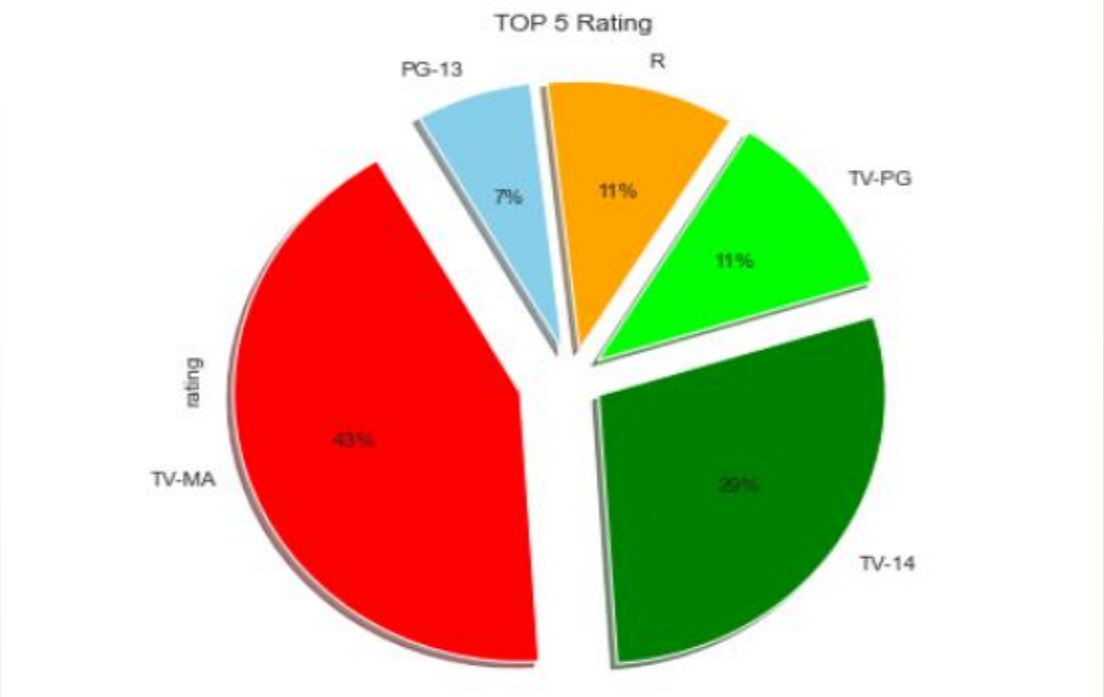
This dataset consists of TV shows and movies available on Netflix. It contains 7787 rows and 12 columns.

	type	title	country	date_added	release_year	rating	duration	listed_in	year_adc
0	Movie	Dick Johnson Is Dead	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	2(
1	TV Show	Blood & Water	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	2(
2	TV Show	Ganglands	United States	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	2(
3	TV Show	Jailbirds New Orleans	United States	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	2(
4	TV Show	Kota Factory	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	2(

Exploratory data analysis

Exploratory data analysis on NETFLIX MOVIES AND TV SHOWS, EDA done using python, **numpy**, **pandas**, **matplotlib**, **seaborn** and **plotly**. You will find many useful visualizations and Tables in the project. I have tried to analyze most of the features of the dataset to derive insights.





Cleaning and Preprocessing

Cleaning and Preprocessing of the data before modulization, using different libraries and different method. as converting the text to numerical variables.

```
In [318]: df.head(5)
```

Out[318]:

	type_enc	title_enc	country_enc	date_added_dates	month_added_enc	year_added_enc	release_year_enc	rating_enc	listed_in_
0	0	1973	603	25	12	13	72	7	
1	1	1089	426	24	12	13	73	11	
2	1	2647	603	24	12	13	73	11	
3	1	3501	603	24	12	13	73	11	
4	1	3855	251	24	12	13	73	11	

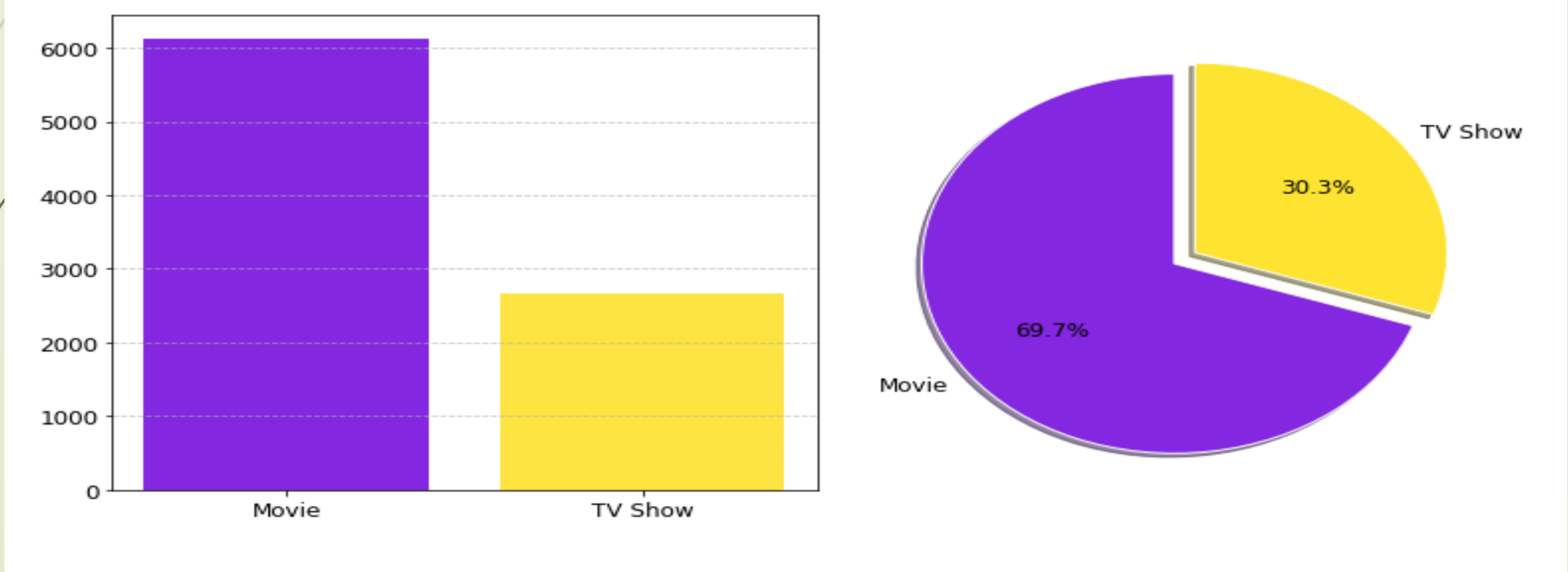
Machine Learning Algorithms

Analysis using different types of Machine Learning Algorithms for classifying and predicting. using three types of Algorithms; KNN, Random Forest and Logistic Regression.

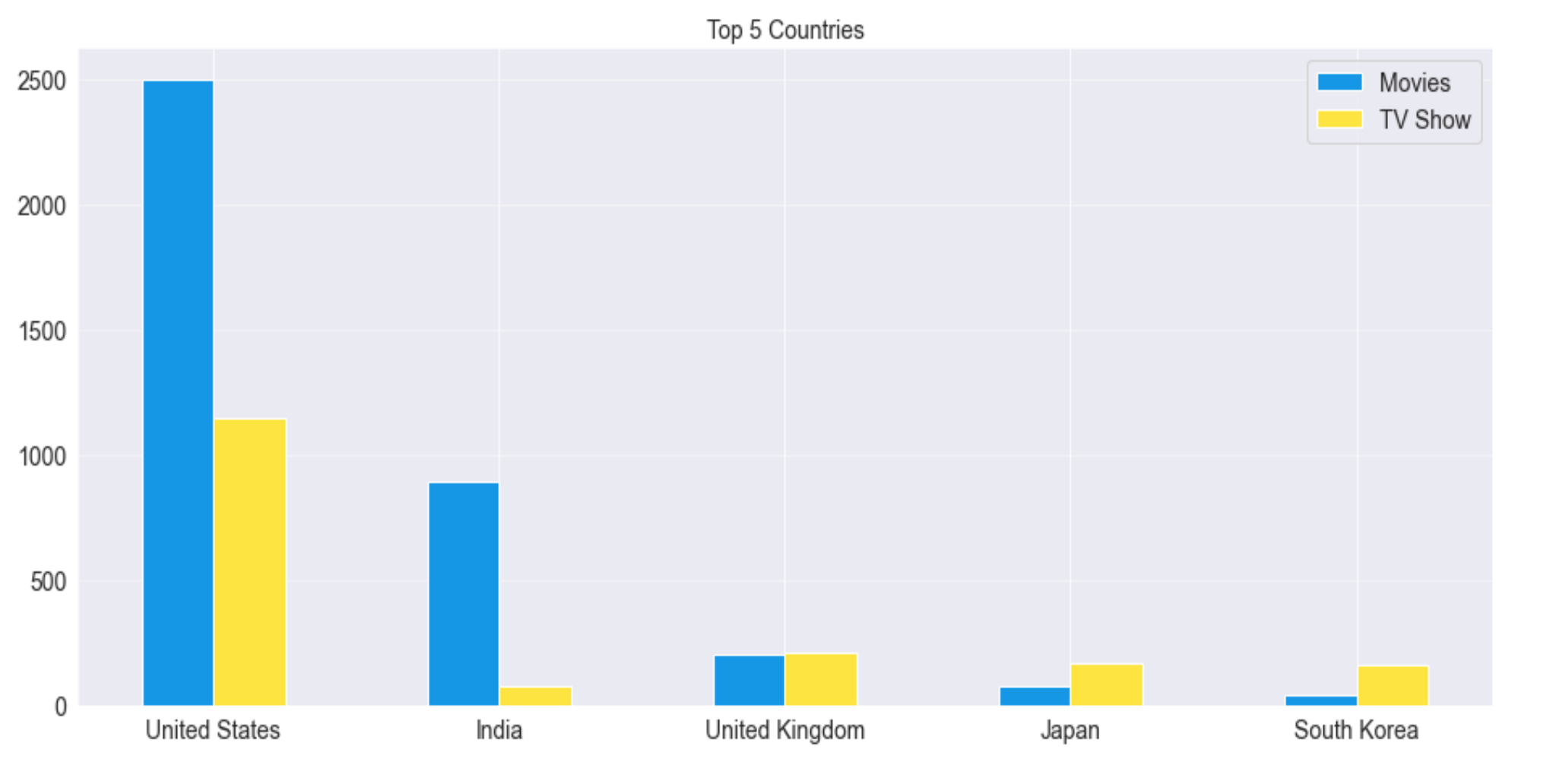
Machine Learning Algorithms

Logistic regression, k-Nearest Neighbors, and Random Forest classifiers were used before settling on random forest as the model with strongest cross-validation performance. Random forest feature importance ranking was used directly to guide the choice and order of variables to be included as the model underwent refinement.

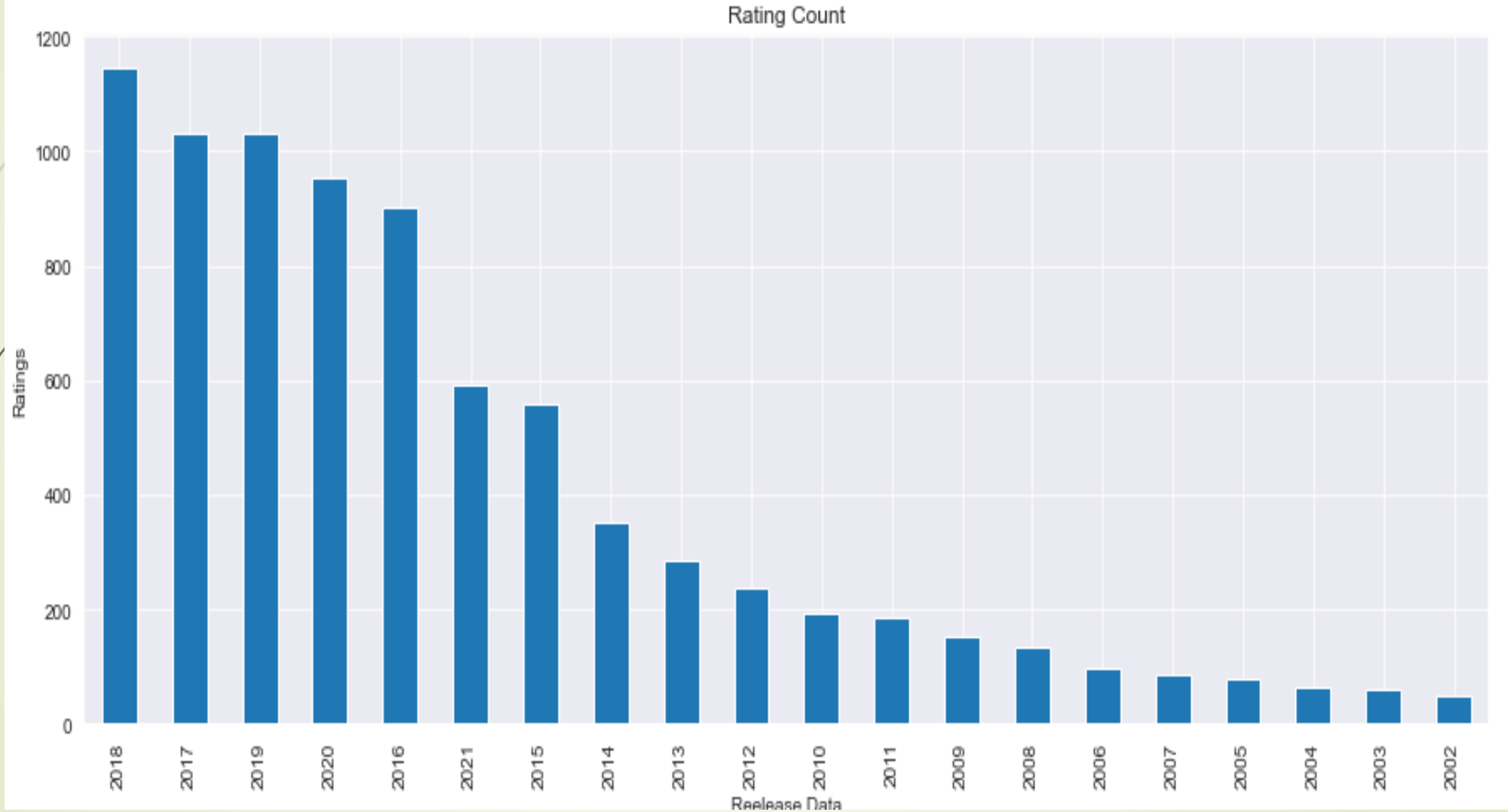
Result eht taht ereh ees nac ew :
rehgih si seivom rof sreweiv fo egatnecrep
69.7% compared to tv show 30.3%



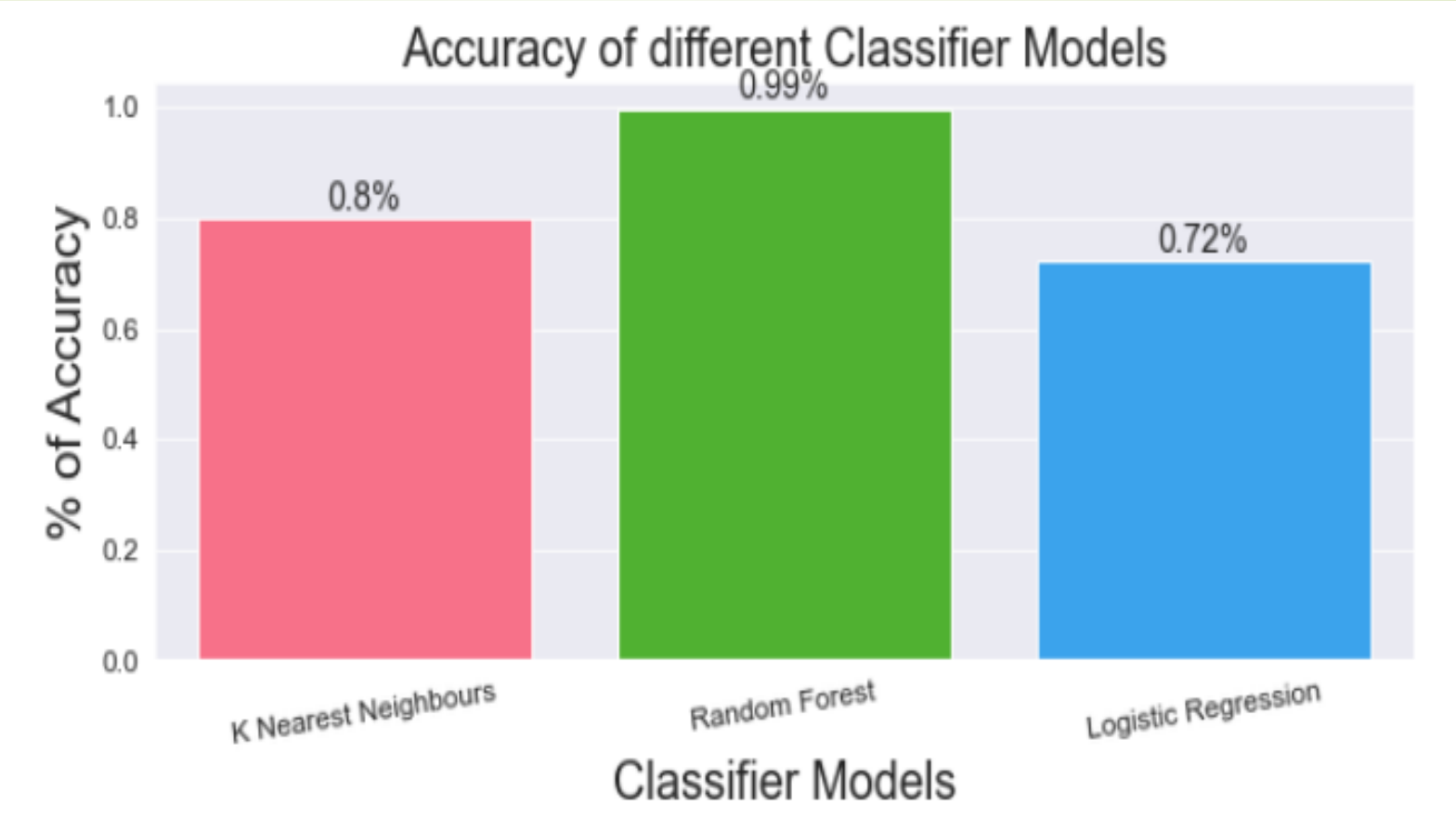
Result: the top 5 product the movies and tv showes



Result:


Rating Count
Ratings
1200
1000
800
600
400
200
0
2018
2017
2019
2020
2016
2021
2015
2014
2013
2012
2010
2011
2009
2008
2006
2007
2005
2004
2003
2002
Reelease Data

Result2:



According the Figure we see that using Random Forest Algorithm gives a high accuracy results compared to KNN Algorithm and the Logistic Regression.