

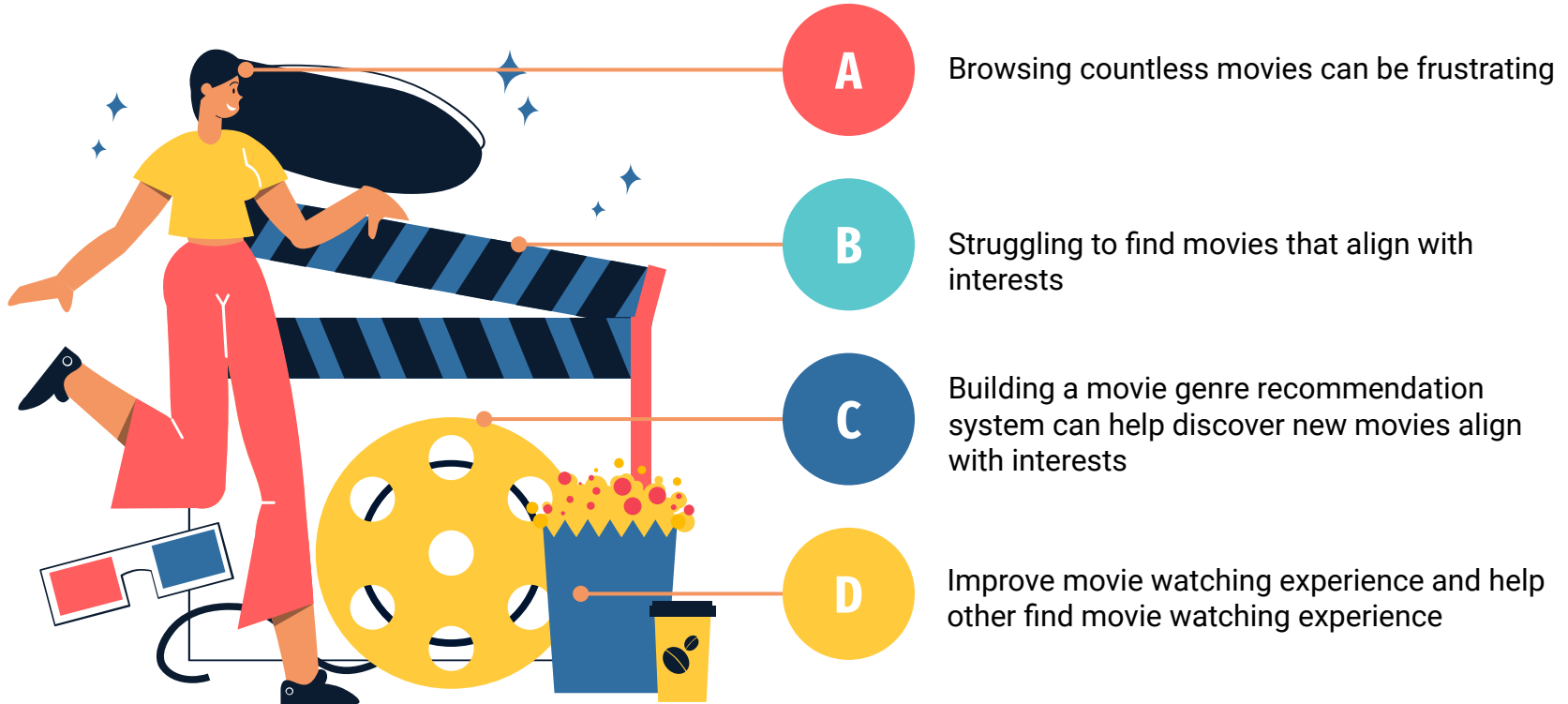
# Movie Genre Recommendation System using Machine Learning

Forward School Applied Data Science  
Capstone Project -11th March 2023

Subashanan Nair



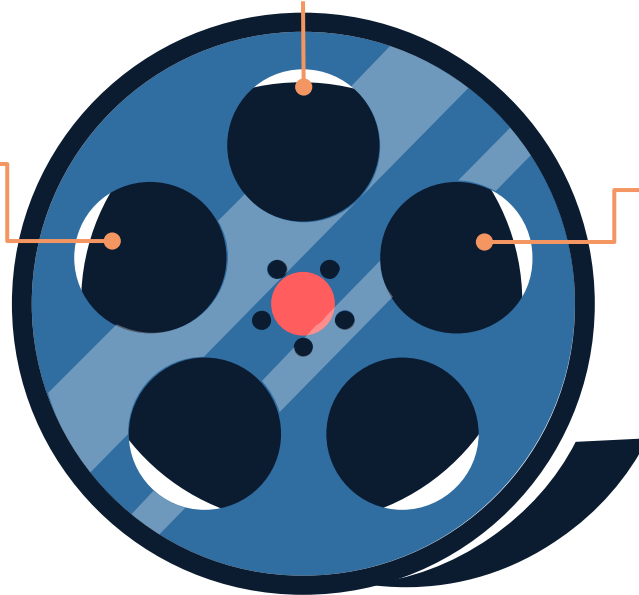
# Why did I choose this project ?



# Problem Statement

**Develop a personalized  
movie genre prediction**

**To enhance user's overall  
viewing experience by  
suggesting movies aligned  
with their interest**



**To improve the  
accuracy of genre  
predictions and  
provide users with  
personalized movie  
recommendation**

# Roadmap

1

Data Collection

2

Data Cleaning



Feature  
Engineering

3

Modelling

4

Deployment

5

## 1

## Data Collection



| Octoparse

Feature Film (Sorted by Popularity Ascending) - IMDb

Show Browser

12 Data Extracted

Running

[Click URLs in the list] Loading webpage https://www.imdb.com/title/tt1016150/?ref\_=adv\_li\_tt

Duplicates: 0 line(s) Time Spent: 1m 48s Avg. Speed: 7 lines/min

Pause

Task Overview Data List Event Log Recent Runs

#	title	title_url	image_url	listitemimage_...	listitemindex	year	certificate	runtime	genre
2	Cocaine Bear	https://www.imdb...	https://m.media-...	https://www.imdb...	2.	(2023)	15	95 min	Comedy, ...
3	The Whale	https://www.imdb...	https://m.media-...	https://www.imdb...	3.	(2022)	15	117 min	Drama
4	Babylon	https://www.imdb...	https://m.media-...	https://www.imdb...	4.	(I) (2022)	18	189 min	Comedy, I
5	Knock at the Cabin	https://www.imdb...	https://m.media-...	https://www.imdb...	5.	(2023)	15	100 min	Horror, M
6	Sharper	https://www.imdb...	https://m.media-...	https://www.imdb...	6.	(2023)	15	116 min	Crime, Dr
7	The Banshees of ...	https://www.imdb...	https://m.media-...	https://www.imdb...	7.	(2022)	15	114 min	Comedy, I
8	Winnie the Pooh: ...	https://www.imdb...	https://m.media-...	https://www.imdb...	8.	(2023)	18	84 min	Horror
9	Avatar: The Way ...	https://www.imdb...	https://m.media-...	https://www.imdb...	9.	(2022)	12A	192 min	Action, Ac
10	Black Panther: W...	https://www.imdb...	https://m.media-...	https://www.imdb...	10.	(2022)	12A	161 min	Action, Ac
11	Everything Every...	https://www.imdb...	https://m.media-...	https://www.imdb...	11.	(2022)	15	135 min	Comedy, ...
12	All Quiet on the ...	https://www.imdb...	https://m.media-...	https://www.imdb...	12.	(2022)	15	148 min	Drama

< 1 > Go to Page



## Avatar (2009)

PG-13

12/18/2009 (US)

Action, Adventure, Fantasy, Science Fiction

2h 42m



User  
Score



Play Trailer

*Enter the world of Pandora.*

### Overview

In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but becomes torn between following orders and protecting an alien civilization.

**James Cameron**

Director, Writer

Multi label classification: Each observation can be classified into multiple classes



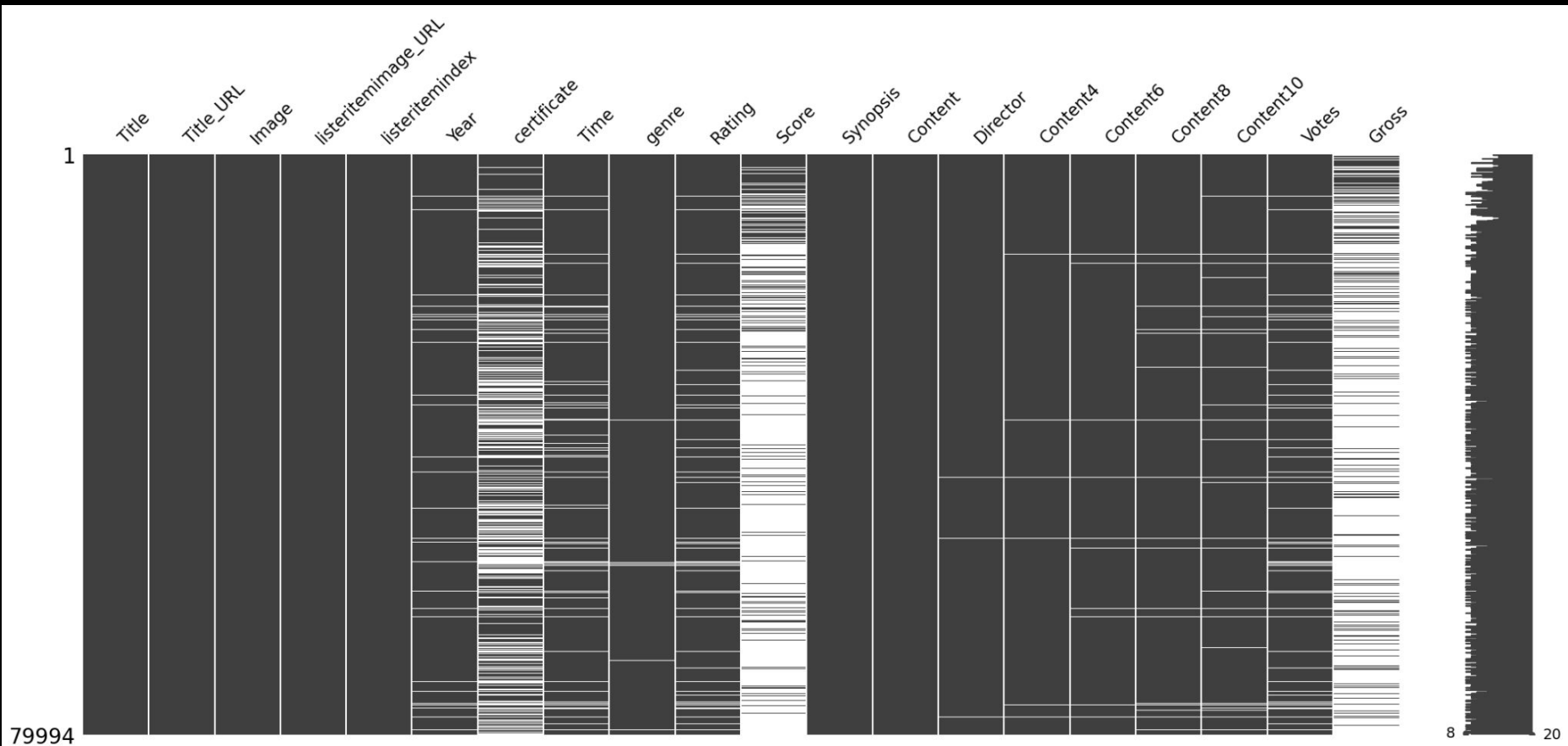
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 71247 entries, 0 to 71246
Data columns (total 20 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Title                 71247 non-null  object
1   Title_URL             71247 non-null  object
2   Image                 71247 non-null  object
3   listeritemimage_URL   71247 non-null  object
4   listeritemindex       71247 non-null  object
5   Year                  68233 non-null  object
6   certificate            38815 non-null  object
7   Time                  65182 non-null  object
8   genre                 70751 non-null  object
9   Rating                65391 non-null  float64
10  Score                 14756 non-null  object
11  Synopsis              71247 non-null  object
12  Content               71246 non-null  object
13  Director              70949 non-null  object
14  Content4              70175 non-null  object
15  Content6              69594 non-null  object
16  Content8              69042 non-null  object
17  Content10             68321 non-null  object
18  Votes                 65392 non-null  object
19  Gross                 13906 non-null  object
dtypes: float64(1), object(19)
memory usage: 10.9+ MB
```



	Title	Title_URL	Image	listitemimage_URL	listitemindex	Year	certificate	Time	genre	Rating	Score	Synopsis	Content	Director	Content4	Content6	Content8	Content10	Votes	
0	Avatar: The Way of Water	<a href="https://www.imdb.com/title/tt1630029/?ref_=adv...">https://www.imdb.com/title/tt1630029/?ref_=adv...</a>	<a href="https://m.media-amazon.com/images/M/MV5BYjhiNj...">https://m.media-amazon.com/images/M/MV5BYjhiNj...</a>	<a href="https://www.imdb.com/title/tt1630029/?ref_=adv...">https://www.imdb.com/title/tt1630029/?ref_=adv...</a>	1.	(2022)	12A	192 min	Action	7.8	\n67\n Metascore\n	\nJake Sully lives with his newfound family fo...	Director:\nJames Cameron\n	James Cameron	Sam Worthington	Zoe Saldana	Sigourney Weaver	Stephen Lang	280,707	
1	The Menu	<a href="https://www.imdb.com/title/tt9764362/?ref_=adv...">https://www.imdb.com/title/tt9764362/?ref_=adv...</a>	<a href="https://m.media-amazon.com/images/M/MV5BMzajNj...">https://m.media-amazon.com/images/M/MV5BMzajNj...</a>	<a href="https://www.imdb.com/title/tt9764362/?ref_=adv...">https://www.imdb.com/title/tt9764362/?ref_=adv...</a>	2.	(2022)	15	107 min	Thriller	7.2	\n71\n Metascore\n	\nA young couple travels to a remote island to...	\n Director:\nMark Mylod\n	Mark Mylod	Ralph Fiennes	Anya Taylor-Joy	Nicholas Hoult	Hong Chau	211,749	
2	Babylon	<a href="https://www.imdb.com/title/tt10640346/?ref_=adv...">https://www.imdb.com/title/tt10640346/?ref_=adv...</a>	<a href="https://m.media-amazon.com/images/M/MV5BNjkYj...">https://m.media-amazon.com/images/M/MV5BNjkYj...</a>	<a href="https://www.imdb.com/title/tt10640346/?ref_=adv...">https://www.imdb.com/title/tt10640346/?ref_=adv...</a>	3.	(I) (2022)	18	189 min	Comedy	7.5	\n60\n Metascore\n	\nA tale of outsized ambition and outrageous e...	Director:\nDamien Chazelle\n	Damien Chazelle	Brad Pitt	Margot Robbie	Jean Smart	Olivia Wilde	47,754	
3	Everything Everywhere All at Once	<a href="https://www.imdb.com/title/tt6710474/?ref_=adv...">https://www.imdb.com/title/tt6710474/?ref_=adv...</a>	<a href="https://m.media-amazon.com/images/S/sash/4Fyxw...">https://m.media-amazon.com/images/S/sash/4Fyxw...</a>	<a href="https://www.imdb.com/title/tt6710474/?ref_=adv...">https://www.imdb.com/title/tt6710474/?ref_=adv...</a>	4.	(2022)	15	139 min	Action	8.0	\n81\n Metascore\n	\nA middle-aged Chinese immigrant is swept up ...	\n Directors:\nDan Kwan,\nDaniel Scheinert...	Dan Kwan	Daniel Scheinert	Michelle Yeoh	Stephanie Hsu	Jamie Lee Curtis	315,971	
4	M3gan	<a href="https://www.imdb.com/title/tt8760708/?ref_=adv...">https://www.imdb.com/title/tt8760708/?ref_=adv...</a>	<a href="https://m.media-amazon.com/images/S/sash/4Fyxw...">https://m.media-amazon.com/images/S/sash/4Fyxw...</a>	<a href="https://www.imdb.com/title/tt8760708/?ref_=adv...">https://www.imdb.com/title/tt8760708/?ref_=adv...</a>	5.	(2022)	15	102 min	Horror	6.4	\n72\n Metascore\n	\nA robotics engineer at a toy company builds ...	Director:\nGerard Johnstone\n	Gerard Johnstone	Allison Williams	Violet McGraw	Ronny Chieng	Amie Donald	52,436	



## Before cleaning



## 2

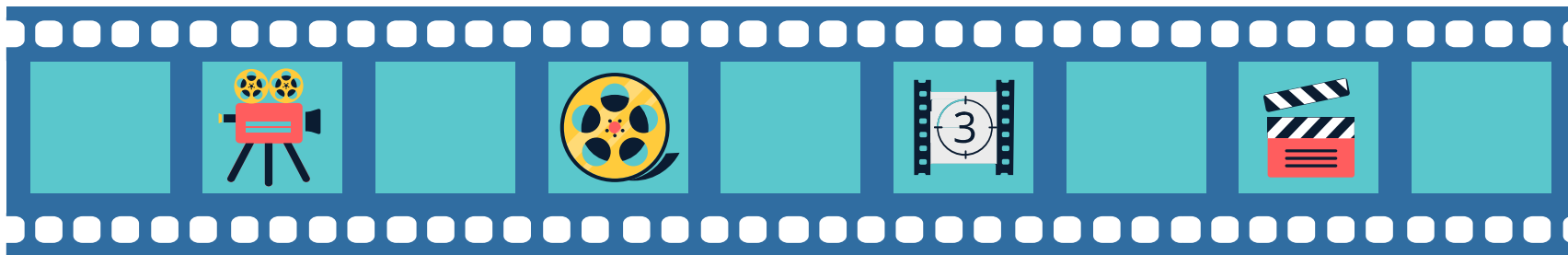
## Data Cleaning

### Step 1

Dropped  
unnecessary  
columns

### Step 3

Clean synopsis using NLTK  
Word\_tokenize  
Lemmatize & PorterStemmer



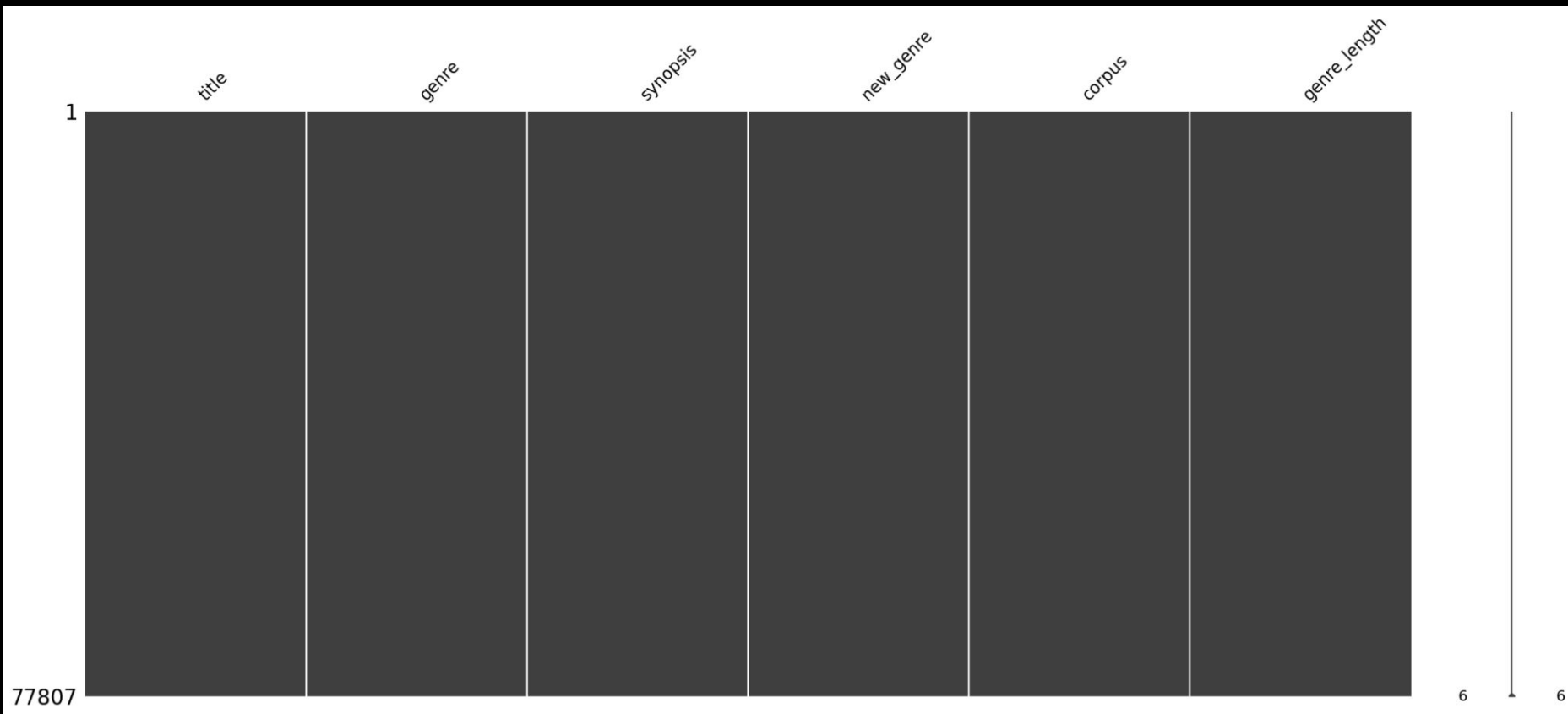
### Step 2

Cleaned genre  
section using  
regex

### Step 4

Check for empty  
string in corpus  
and genre

## After cleaning



3

## Feature Engineering

Unique insights



**Number of movies by  
Genre Length**



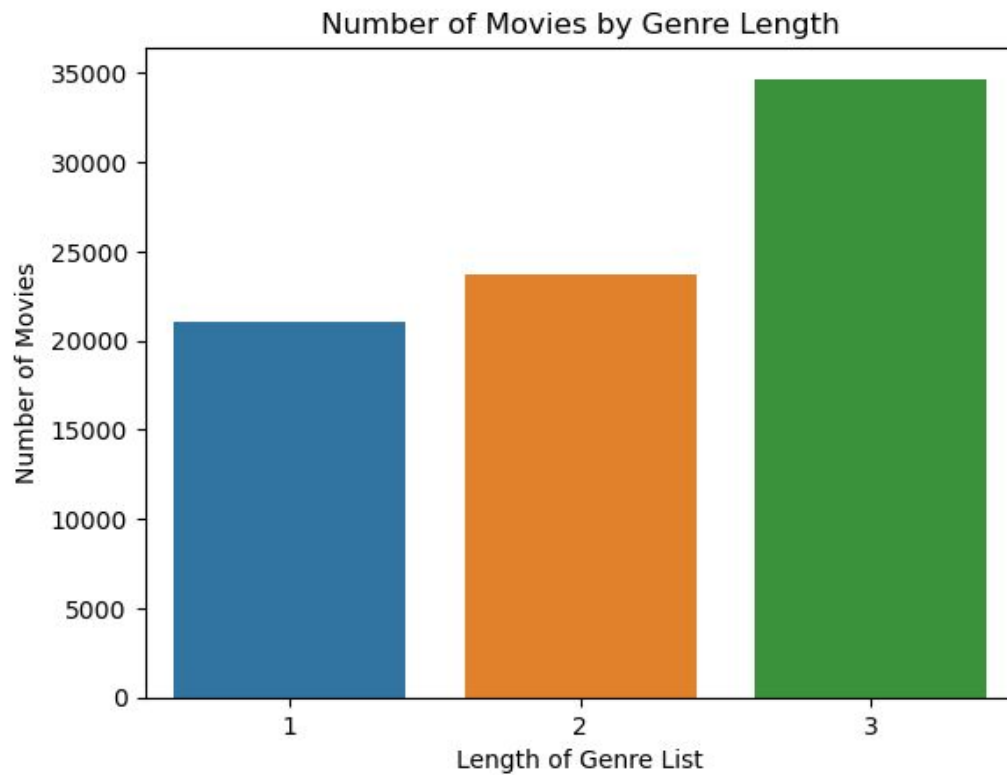
**Top Genres**



**Correlation**

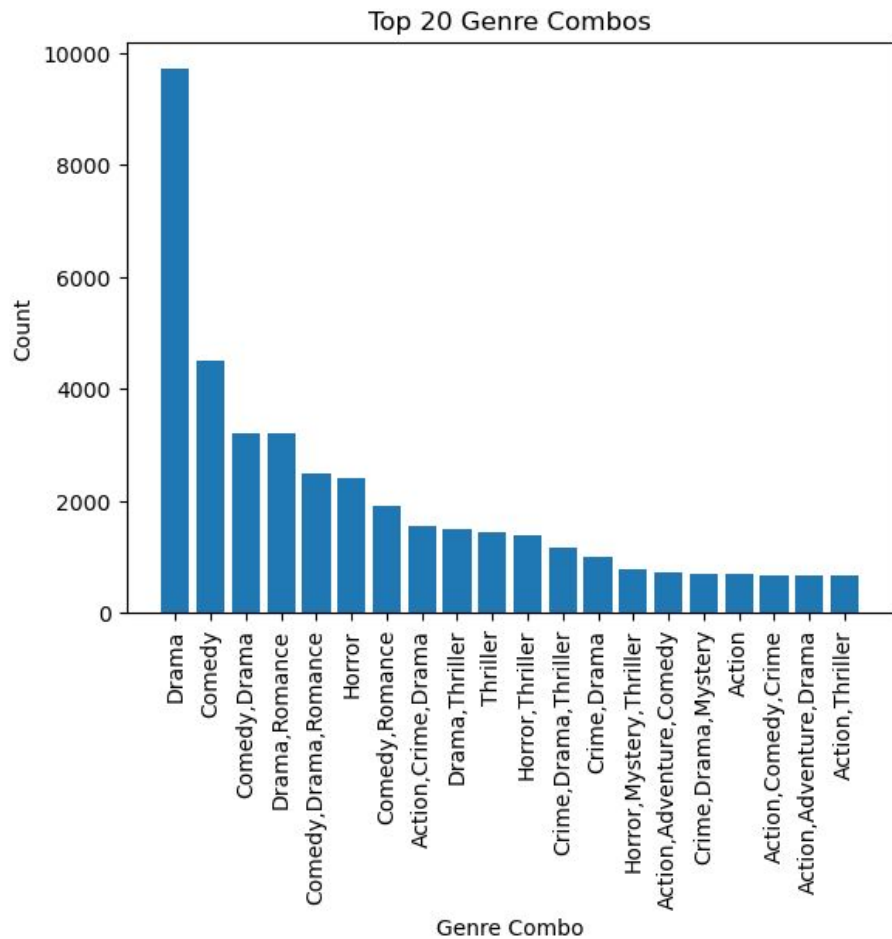


## Number of movies by Genre Length



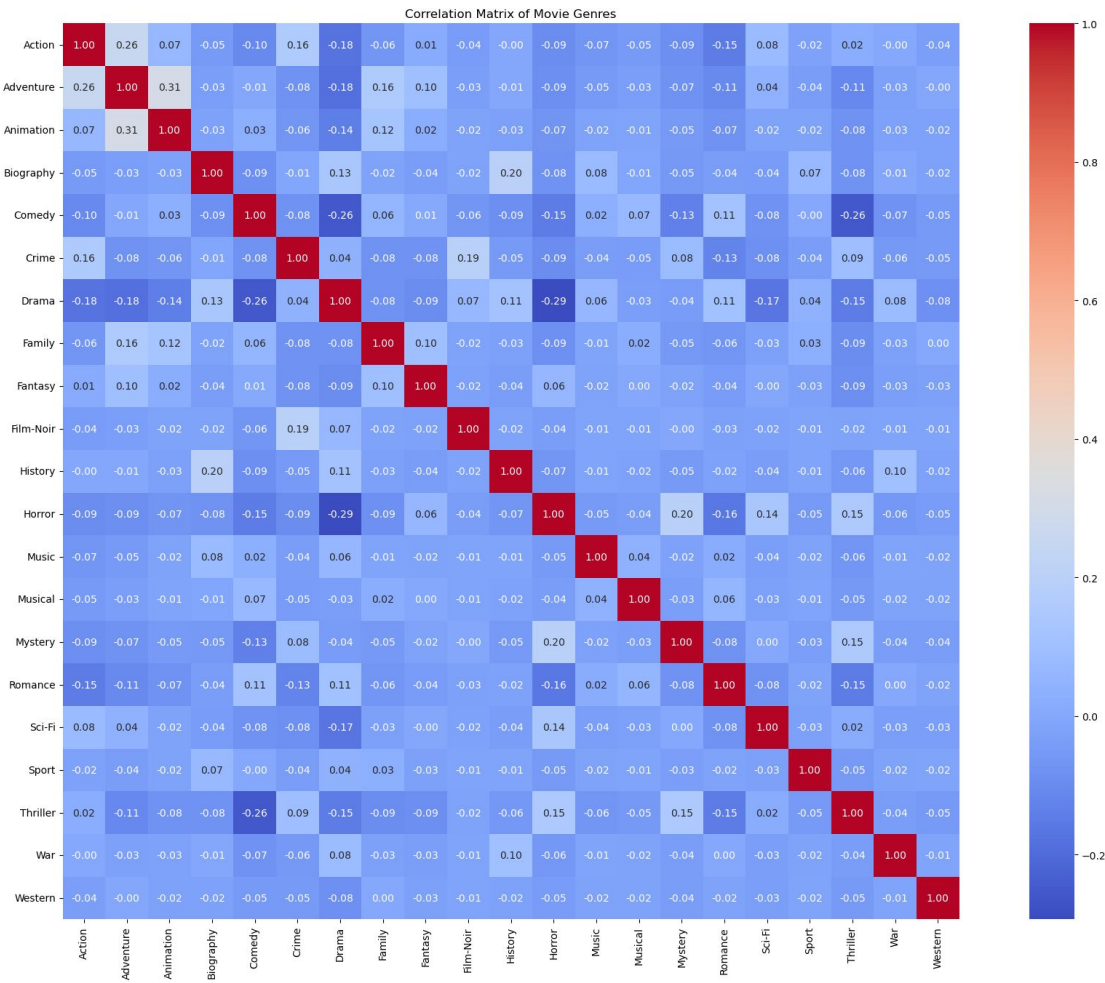


## Top Genres





# Correlation



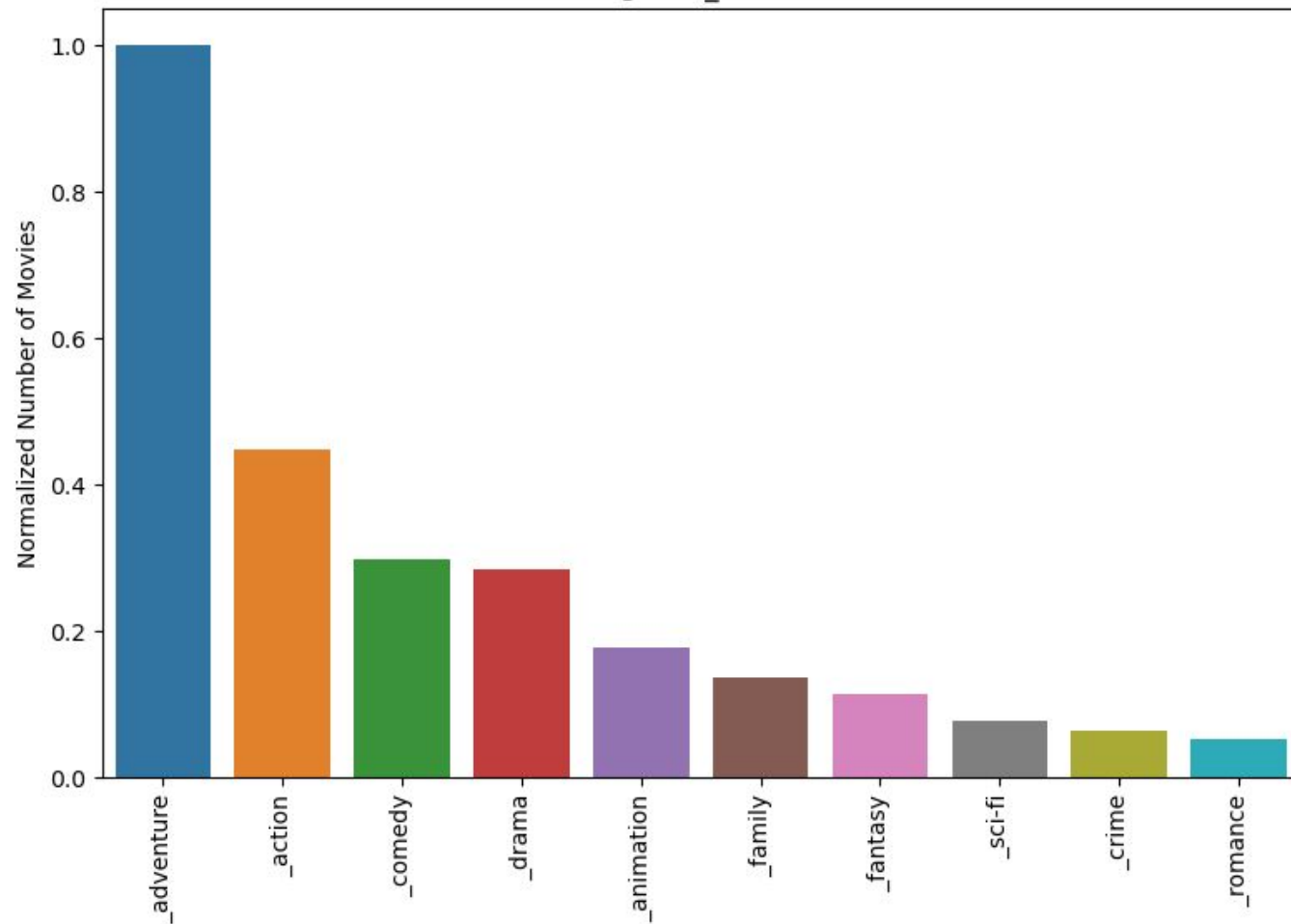
Drama - negative  
corr with many  
genres

Comedy - negative  
with many genre  
except romance and  
family

Horror - positive  
with mystery and  
thriller

Animation - positive  
with family

Distribution given \_adventure Genre





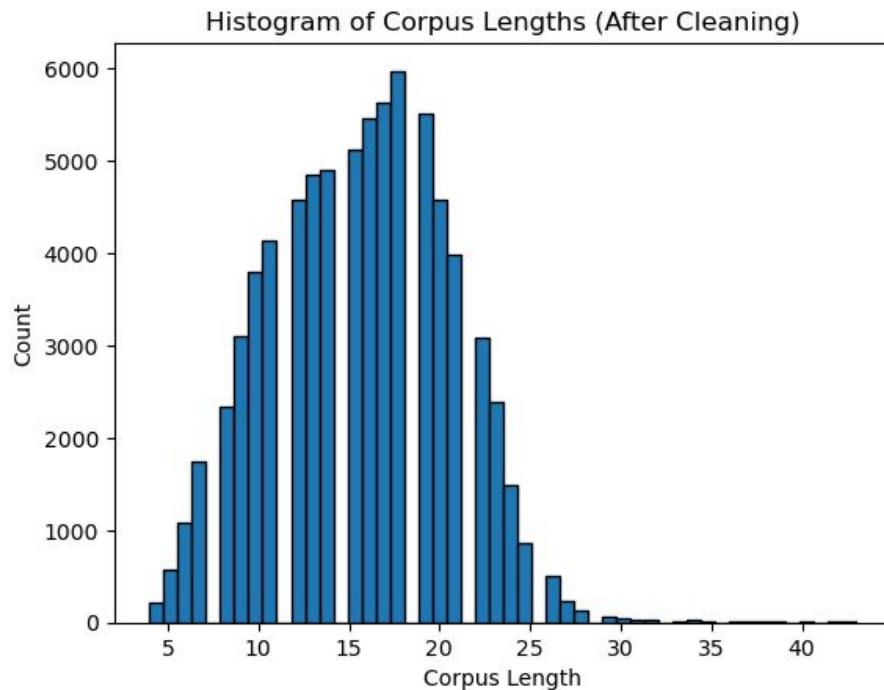
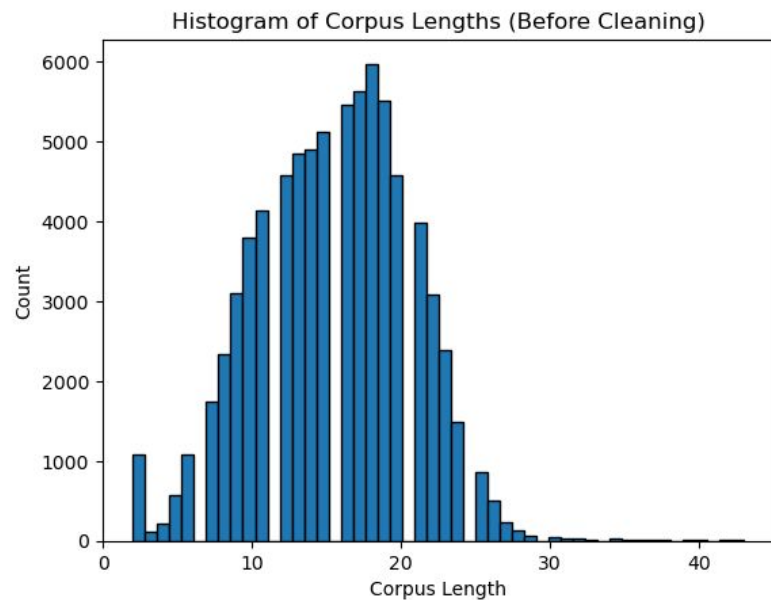
4

## Feature Engineering



## Imputation

1



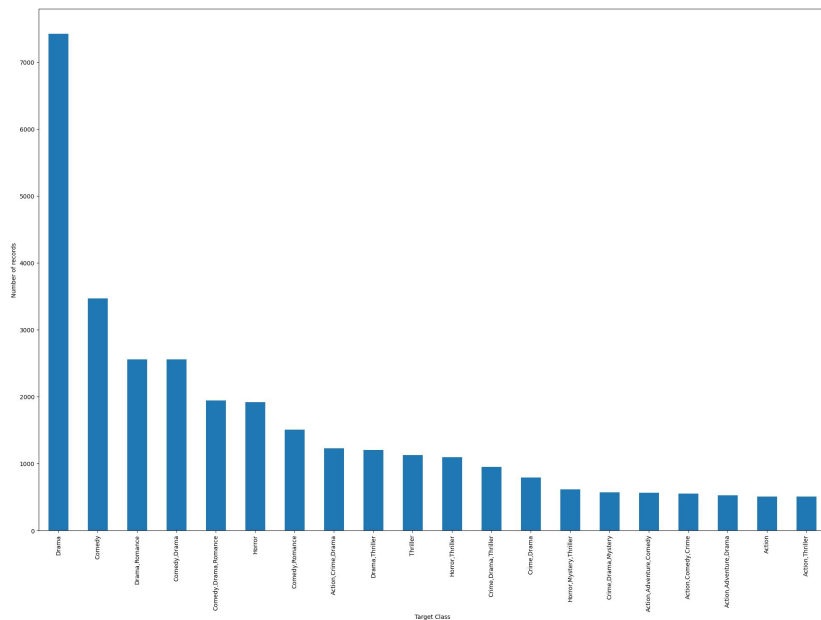
## MultiLabelbinarizer

	action	adventure	animation	biography	comedy	crime	drama	family	fantasy	film-noir	...	horror	music	musical	mystery	romance	sci-fi	sport	thriller	war	western
<b>0</b>	1	1	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
<b>1</b>	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	1	0	0
<b>2</b>	0	0	0	0	1	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
<b>3</b>	1	1	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
<b>4</b>	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	1	0	1	0	0
<b>...</b>	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
<b>76602</b>	0	0	0	1	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
<b>76603</b>	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
<b>76604</b>	1	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
<b>76605</b>	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
<b>76606</b>	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0

76607 rows x 21 columns

## 4

# Train Classification Model



	Classifier	Accuracy	Precision	Recall	F1-Score
0	Decision Tree	0.731724	0.740852	0.731724	0.729321
1	Random Forest	0.749496	0.814880	0.749496	0.755702
2	Logistic Regression	0.358016	0.639328	0.358016	0.286237
3	Support Vector Machine	0.750000	0.765313	0.750000	0.747193

# After resampling (sample\_strategy = 'all')

Model: Single Vector Machine

Accuracy: 0.782

Hamming Loss: 0.218

F1-Score: 0.760

Precision: 0.880

Recall: 0.708

Model: Decision Tree

Accuracy: 0.765

Hamming Loss: 0.235

F1-Score: 0.719

Precision: 0.798

Recall: 0.702

Model: Random Forest

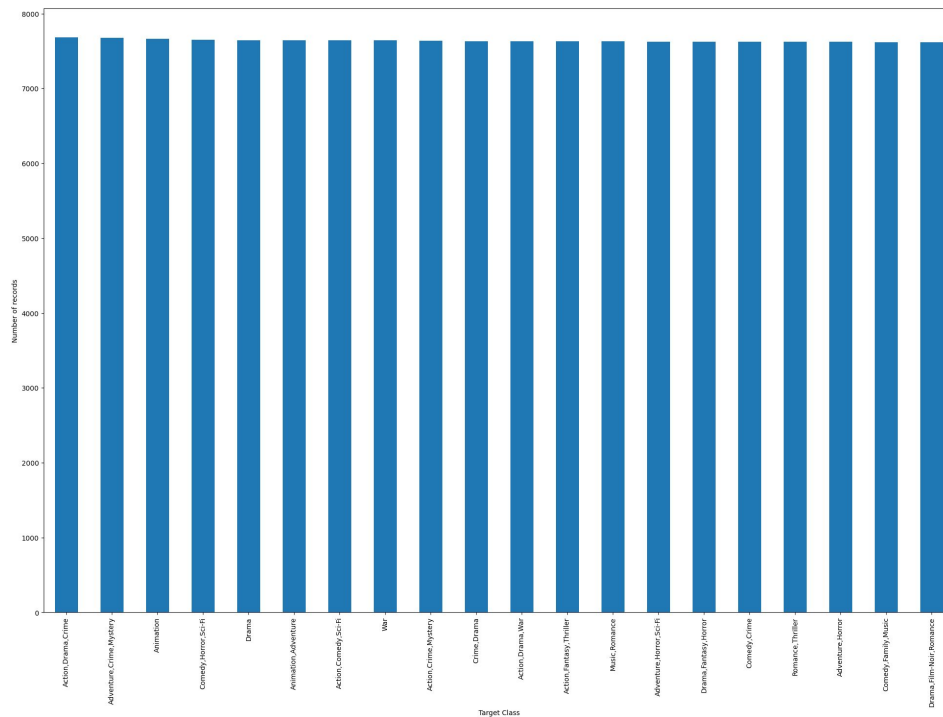
Accuracy: 0.782

Hamming Loss: 0.218

F1-Score: 0.782

Precision: 0.949

Recall: 0.703



# Demo



78%

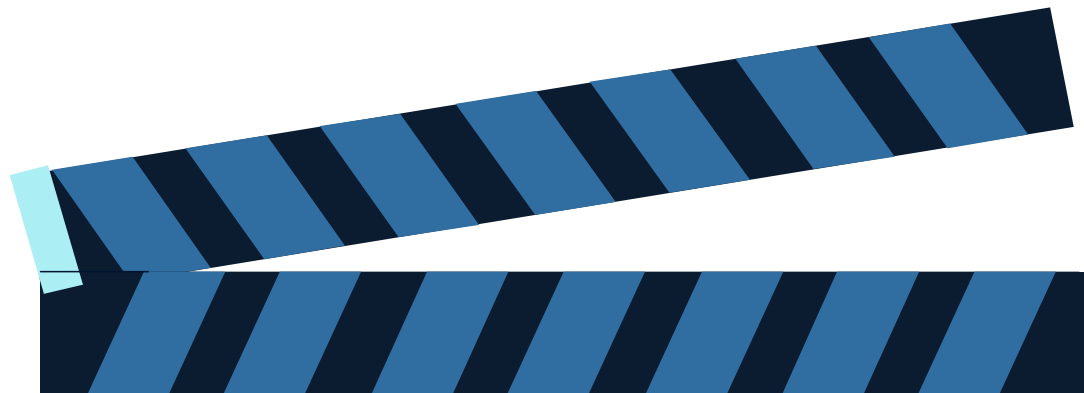
Accuracy



95%

Precision

# QR CODE & QnA



To create a public link, set 'share=True' in 'launch()'.

## GenreOracle (v1.0-beta1)

Enter the synopsis of a movie and get suggestions for similar movies

Enter the synopsis of the movie

Number of suggested movie titles

5 8



Clear

Submit

Predicted genre

Suggested movie titles with the predicted genre

Flag

Use via API  · Built with Gradio 



Directed by  
ROBERT B. WEIDE

*The  
End*