CASE STUDY

Exploring User Preferences and Opinion Analysis in OTT Platform Selection: A Machine Learning Approach

Abstract:

In this study, we analyze user preferences and opinions regarding Over-The-Top (OTT) platforms using machine learning techniques. We collected data through a survey encompassing factors such as gender, preferred genres, and opinions on popular OTT platforms. Our analysis aims to predict user preferences based on opinion feedback, employing algorithms including KNN, SVM, Decision Trees, Logistic Regression, Random Forest, and Naive Bayes. Through this investigation, we aim to identify the most accurate model for predicting OTT platform preferences, shedding light on user behavior and aiding platform optimization strategies.

Introduction:

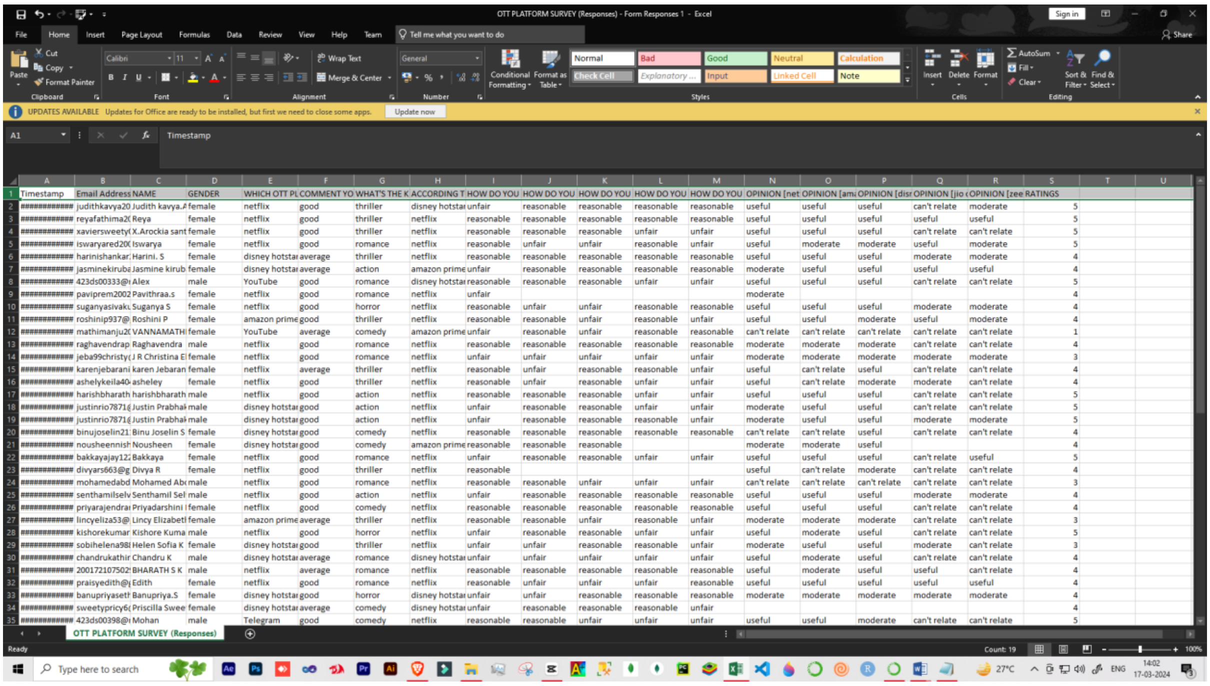
This study delves into user data collected through surveys to analyze preferences and opinions on popular OTT platforms. Leveraging machine learning algorithms, we aim to predict user preferences based on opinion feedback, providing valuable insights for platform optimization.

Problem Definition:

In the rapidly evolving landscape of digital streaming, Over-The-Top (OTT) platforms have become prominent sources of entertainment. However, understanding user preferences and opinions is essential for optimizing platform offerings and maximizing user satisfaction. This study aims to predict user preferences for OTT platforms based on demographic information, genre preferences, and user opinions.

Data Collection:

In the rapidly evolving landscape of digital streaming, Over-The-Top (OTT) platforms have become prominent sources of entertainment. However, understanding user preferences and opinions is essential for optimizing platform offerings and maximizing user satisfaction. This study aims to predict user preferences for OTT platforms based on demographic information, genre preferences, and user opinions.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Timestamp	Email Address	NAME	GENDER	WHICH OTT PL	COMMENT YO	WHAT'S THE K	ACCORDING T	HOW DO YOU	HOW DO YOU	HOW DO YOU	HOW DO YOU	HOW DO YOU	OPINION [net	OPINION [am	OPINION [dis	OPINION [jio	OPINION [zee	RATINGS
2	############	judithkavya20	Judith kavya.A	female	netflix	good	thriller	disney hotstar	unfair	reasonable	reasonable	reasonable	reasonable	useful	useful	useful	can't relate	moderate	5
3	############	reyafathima20	Reya	female	netflix	good	thriller	netflix	reasonable	reasonable	reasonable	reasonable	reasonable	useful	useful	useful	useful	useful	5
4	############	xaviersweety	X.Arockia sant	female	netflix	good	thriller	netflix	reasonable	reasonable	reasonable	unfair	unfair	useful	useful	useful	can't relate	can't relate	5
5	############	iswaryared20	Iswarya	female	netflix	good	romance	netflix	reasonable	unfair	unfair	reasonable	unfair	useful	moderate	moderate	useful	moderate	5
6	############	harinishankar	Harini. S	female	disney hotstar	average	thriller	netflix	reasonable	reasonable	reasonable	reasonable	reasonable	useful	useful	useful	moderate	moderate	4
7	############	jasminekiruba	Jasmine kirub	female	disney hotstar	average	action	amazon prime	unfair	reasonable	reasonable	reasonable	reasonable	moderate	useful	useful	useful	useful	4
8	############	423ds00333@	Alex	male	YouTube	good	romance	disney hotstar	reasonable	reasonable	reasonable	unfair	unfair	useful	useful	useful	can't relate	can't relate	5
9	############	paviprem2002	Pavithraa.s	female	netflix	good	romance	netflix	unfair					moderate					4
10	############	suganyasivaku	Suganya S	female	netflix	good	horror	netflix	reasonable	unfair	unfair	reasonable	reasonable	useful	useful	useful	moderate	moderate	4
11	############	roshinip937@	Roshini P	female	amazon prime	good	thriller	netflix	reasonable	reasonable	unfair	reasonable	unfair	useful	useful	moderate	useful	moderate	4
12	############	mathimanju20	VANNAMATHI	female	YouTube	average	comedy	amazon prime	unfair	reasonable	unfair	reasonable	reasonable	can't relate	can't relate	can't relate	can't relate	can't relate	1
13	############	raghavendrap	Raghavendra	male	netflix	good	romance	netflix	reasonable	reasonable	reasonable	reasonable	reasonable	moderate	moderate	moderate	can't relate	can't relate	4
14	############	jeba99christy	J R Christina E	female	netflix	good	romance	netflix	unfair	unfair	unfair	unfair	unfair	moderate	moderate	moderate	moderate	moderate	3
15	############	karenjebarani	karen Jebaran	female	netflix	average	thriller	netflix	reasonable	unfair	reasonable	reasonable	unfair	useful	can't relate	useful	can't relate	can't relate	4
16	############	ashelykeila40	asheley	female	netflix	good	thriller	netflix	reasonable	unfair	reasonable	unfair	unfair	useful	can't relate	moderate	moderate	can't relate	4
17	############	harishbharath	harishbharath	male	netflix	good	action	netflix	reasonable	reasonable	reasonable	unfair	unfair	useful	useful	useful	can't relate	can't relate	5
18	############	justinrio7871	Justin Prabhak	male	disney hotstar	good	action	netflix	unfair	reasonable	reasonable	unfair	unfair	moderate	useful	useful	can't relate	can't relate	5
19	############	justinrio7871	Justin Prabhak	male	disney hotstar	good	action	netflix	unfair	reasonable	reasonable	reasonable	unfair	moderate	useful	useful	moderate	moderate	5
20	############	binujoselin21	Binu Joselin S	female	disney hotstar	good	comedy	netflix	reasonable	reasonable	reasonable	reasonable	reasonable	can't relate	can't relate	useful	can't relate	can't relate	4
21	############	nousheennish	Nousheen	female	disney hotstar	good	comedy	amazon prime	reasonable	reasonable	reasonable			moderate	moderate	useful			4
22	############	bakkayajay12	Bakkaya	female	netflix	good	romance	netflix	unfair	reasonable	reasonable	unfair	unfair	useful	useful	useful	can't relate	useful	5
23	############	divyars663@g	Divya R	female	netflix	good	thriller	netflix	reasonable					useful	can't relate	moderate	can't relate	can't relate	4
24	############	mohamedabd	Mohamed Ab	male	netflix	good	romance	netflix	reasonable	reasonable	unfair	unfair	unfair	can't relate	can't relate	can't relate	can't relate	can't relate	3
25	############	senthamilselv	Senthamil Sel	male	netflix	good	action	netflix	unfair	reasonable	reasonable	reasonable	reasonable	useful	useful	useful	moderate	moderate	4
26	############	priyarajendra	Priyadarshini	female	netflix	good	comedy	netflix	reasonable	reasonable	reasonable	reasonable	reasonable	useful	useful	useful	can't relate	can't relate	4
27	############	lincyeliza53@	Lincy Elizabet	female	amazon prime	average	thriller	netflix	reasonable	reasonable	unfair	reasonable	unfair	moderate	moderate	moderate	can't relate	can't relate	3
28	############	kishorekumar	Kishore Kuma	male	netflix	good	horror	netflix	unfair	reasonable	reasonable	unfair	unfair	useful	moderate	useful	can't relate	can't relate	5
29	############	sobihelena98	Helen Sofia K	female	disney hotstar	good	thriller	netflix	unfair	unfair	reasonable	reasonable	unfair	moderate	useful	useful	moderate	can't relate	3
30	############	chandrukathir	Chandru K	male	disney hotstar	average	romance	disney hotstar	unfair	unfair	unfair	unfair	unfair	useful	moderate	useful	can't relate	can't relate	4
31	############	200172107502	BHARATH S K	male	netflix	average	romance	netflix	reasonable	reasonable	reasonable	reasonable	reasonable	useful	moderate	useful	useful	can't relate	4
32	############	praisyedith@	Edith	female	netflix	good	romance	netflix	reasonable	unfair	unfair	unfair	reasonable	useful	useful	useful	useful	useful	4
33	############	banupriyaseth	Banupriya.S	female	disney hotstar	good	horror	disney hotstar	unfair	unfair	reasonable	reasonable	reasonable	moderate	moderate	moderate	moderate	moderate	4
34	############	sweetypricy6	Priscilla Swee	female	disney hotstar	average	comedy	disney hotstar	unfair	reasonable	reasonable	reasonable	unfair						4
35	############	423ds00398@	Mohan	male	Telegram	good	comedy	netflix	reasonable	reasonable	unfair	unfair	unfair	useful	useful	moderate	can't relate	can't relate	5

OTT PLATFORM SURVEY (Responses)

Data Preprocessing:

In the data preprocessing stage, null values were addressed using the dropna() function to ensure data integrity and accuracy in subsequent analysis. Any rows containing missing values were removed from the dataset, as they could potentially skew results or introduce bias.

```
In [2]:
import pandas as pd
alex1=pd.read_csv(r"C:\Users\Admin\Desktop\gifta\OTT PLATFORM SURVEY (Responses) - Form Responses 1.csv")
alex1
```

	Timestamp	Email Address	NAME	GENDER	WHICH OTT PLATFORM YOU GENERALLY PREFER WHEN YOUR BORED?	COMMENT YOUR VIEWS ABOUT OTT	WHAT'S THE KIND OF GENRE YOU GENERALLY PREFER ON OTT ?	ACCORDING TO YOU WHICH OTT PLATFORM IS MORE POPULAR AMONG YOUNGER GENERATION	HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [netflix]	HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [amazon prime]	HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [disney hotstar]	HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [jio cinema]	HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [zee 5]
0	01-10-2024 10:23	judithkavya2002@gmail.com	Judith kavya.A	female	netflix	good	thriller	disney hotstar	unfair	reasonable	reasonable	reasonable	reasonable
1	01-10-2024 10:24	reyafathima2003@gmail.com	Reya	female	netflix	good	thriller	netflix	reasonable	reasonable	reasonable	reasonable	reasonable
2	01-10-2024 10:40	xaviersweety07@gmail.com	X.Arockia santhana sweety	female	netflix	good	thriller	netflix	reasonable	reasonable	reasonable	unfair	unfair
3	01-10-2024 10:51	iswaryared2002@gmail.com	Iswarya	female	netflix	good	romance	netflix	reasonable	unfair	unfair	reasonable	unfair
4	01-10-2024	harinishankar2210@gmail.com	Harini. S	female	disney hotstar	average	thriller	netflix	reasonable	reasonable	reasonable	reasonable	reasonable

```
alex2=alex1.dropna()
alex2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 56 entries, 0 to 62
Data columns (total 19 columns):
 #   Column                                                                 Non-Null Count  Dtype
---  ------                                                                 --------------  -----
 0   Timestamp                                                              56 non-null     object
 1   Email Address                                                          56 non-null     object
 2   NAME                                                                   56 non-null     object
 3   GENDER                                                                 56 non-null     object
 4   WHICH OTT PLATFORM YOU GENERALLY PREFER WHEN YOUR BORED?               56 non-null     object
 5   COMMENT YOUR VIEWS ABOUT OTT                                           56 non-null     object
 6   WHAT'S THE KIND OF GENRE YOU GENERALLY PREFER ON OTT ?                 56 non-null     object
 7   ACCORDING TO YOU WHICH OTT PLATFORM IS MORE POPULAR AMONG YOUNGER GENERATION  56 non-null     object
 8   HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [netflix]                56 non-null     object
 9   HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [amazon prime]           56 non-null     object
 10  HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [disney hotstar]         56 non-null     object
 11  HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [jio cinema]             56 non-null     object
 12  HOW DO YOU FEEL ABOUT THE PRICE RATE OF OTT ? [zee 5]                  56 non-null     object
 13  OPINION [netflix]                                                      56 non-null     object
 14  OPINION [amazon prime]                                                 56 non-null     object
 15  OPINION [disney hotstar]                                               56 non-null     object
 16  OPINION [jio cinema]                                                   56 non-null     object
 17  OPINION [zee 5]                                                        56 non-null     object
 18  RATINGS                                                                56 non-null     int64
dtypes: int64(1), object(18)
memory usage: 8.8+ KB
```

64 rows of data reduced to 56 rows by eliminating null values

Converting User Opinions from String to Numerical Values:

In my dataset, user opinions regarding various OTT platforms were collected as string values, including categories such as "useful," "moderate," and "can't relate." To facilitate machine learning analysis, these qualitative opinions needed to be converted into numerical representations using python.

```
list1=[]
for x1 in alex2["OPINION [netflix]"]:
    if x1=="useful":
        list1.append(1)
    elif x1=="moderate":
        list1.append(2)
    else:
        list1.append(3)
print("1-> Useful\n2-> Moderate\n3-> Can't relate")
dict1={"netflix":list1}
print(dict1)
```

```
1-> Useful
2-> Moderate
3-> Can't relate
{'netflix': [1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 2, 2, 1, 1, 1, 2, 2, 3, 1, 3, 1, 1, 2, 1, 2, 1, 1, 1, 2, 1, 1, 2, 2, 3, 1, 2, 2, 1, 1, 2,
3, 1, 1, 1, 1, 2, 1, 1, 1, 3, 2, 1, 2, 1, 1, 1]}
```

```
list2=[]
for x2 in alex2["OPINION [amazon prime]"]:
    if x2=="useful":
        list2.append(1)
    elif x2=="moderate":
        list2.append(2)
    else:
        list2.append(3)
print("1-> Useful\n2-> Moderate\n3-> Can't relate")
dict2={"amazon prime":list2}
print(dict2)
```

```
1-> Useful
2-> Moderate
3-> Can't relate
{'amazon prime': [1, 1, 1, 2, 1, 1, 1, 1, 1, 3, 2, 2, 3, 3, 1, 1, 1, 3, 1, 3, 1, 1, 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 3, 1, 2, 1, 1,
1, 1, 2, 1, 1, 1, 1, 2, 2, 1, 2, 3, 2, 1, 1, 1, 2, 2]}
```

```python
list3=[]
for x3 in alex2["OPINION [disney hotstar]"]:
    if x3=="useful":
        list3.append(1)
    elif x3=="moderate":
        list3.append(2)
    else:
        list3.append(3)
print("1-> Useful\n2-> Moderate\n3-> Can't relate")
dict3={"disney hotstar":list3}
print(dict3)
```

```
1-> Useful
2-> Moderate
3-> Can't relate
{'disney hotstar': [1, 1, 1, 2, 1, 1, 1, 1, 2, 3, 2, 2, 1, 2, 1, 1, 1, 1, 1, 3, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 1, 2, 2, 2, 1,
1, 1, 3, 2, 1, 1, 1, 1, 2, 2, 2, 3, 2, 1, 2, 1, 1, 2]}
```

```python
list4=[]
for x4 in alex2["OPINION [jio cinema]"]:
    if x4=="useful":
        list4.append(1)
    elif x4=="moderate":
        list4.append(2)
    else:
        list4.append(3)
print("1-> Useful\n2-> Moderate\n3-> Can't relate")
dict4={"jio cinema":list4}
print(dict4)
```

```
1-> Useful
2-> Moderate
3-> Can't relate
{'jio cinema': [3, 1, 3, 1, 2, 1, 3, 2, 1, 3, 3, 2, 3, 2, 3, 3, 2, 3, 3, 3, 2, 3, 3, 3, 2, 3, 1, 1, 2, 3, 3, 2, 2, 1, 2, 2, 1, 3, 2,
3, 3, 2, 2, 1, 1, 2, 2, 2, 3, 3, 2, 2, 2, 1, 2, 2]}
```

```python
list5=[]
for x5 in alex2["OPINION [zee 5]"]:
    if x5=="useful":
        list5.append(1)
    elif x5=="moderate":
        list5.append(2)
    else:
        list5.append(3)
print("1-> Useful\n2-> Moderate\n3-> Can't relate")
dict5={"zee 5":list5}
print(dict5)
```

```
1-> Useful
2-> Moderate
3-> Can't relate
{'zee 5': [2, 1, 3, 2, 2, 1, 3, 2, 2, 3, 3, 2, 3, 3, 3, 3, 2, 3, 1, 3, 2, 3, 3, 3, 3, 3, 3, 1, 2, 3, 3, 2, 2, 3, 2, 2, 3, 2, 2, 1, 3,
2, 2, 1, 1, 2, 2, 1, 3, 3, 2, 2, 2, 1, 3, 2]}
```

Feature Selection:

To accomplish this task, we first identified the columns representing user opinions for each OTT platform, including "OPINION [Netflix]," "OPINION [Amazon Prime]," "OPINION [Disney Hotstar]," "OPINION [Jio Cinema]," and "OPINION [Zee 5]." These columns contain qualitative feedback from users regarding their experience with each platform, ranging from "useful" to "can't relate."

Next, we selected these opinion columns as features for our predictive model, recognizing their significance in understanding user preferences. By incorporating user opinions as features, we aim to leverage the insights provided by users to accurately predict the OTT platform that best aligns with their preferences.

Data1

```
dummy1=pd.DataFrame(dict1)
dummy2=pd.DataFrame(dict2)
dummy3=pd.DataFrame(dict3)
dummy4=pd.DataFrame(dict4)
dummy5=pd.DataFrame(dict5)
dummy6=pd.concat([dummy1,dummy2,dummy3,dummy4,dummy5],axis=1)
dummy6
```

	netflix	amazon prime	disney hotstar	jio cinema	zee 5
0	1	1	1	3	2
1	1	1	1	1	1
3	1	2	2	1	2
4	1	1	1	2	2
5	2	1	1	1	1
6	1	1	1	3	3
7	1	1	1	2	2
8	1	1	2	1	2
9	3	3	3	3	3
10	2	2	2	3	3
11	2	2	2	2	2
12	1	3	1	3	3
13	1	3	2	2	3
14	1	1	1	3	3
15	2	1	1	3	3
16	2	1	1	2	2
17	3	3	1	3	3
18	1	1	1	3	1
19	3	3	3	3	3
20	1	1	1	2	2
21	1	1	1	3	3
22	2	2	2	3	3
23	1	2	1	3	3
24	2	1	1	2	3
25	1	2	1	3	3
26	1	2	1	1	3
27	1	1	1	1	1
28	2	2	2	2	2
29	1	1	2	3	3
30	1	1	2	3	3
31	2	2	2	2	2
32	2	2	2	2	2
33	3	3	1	1	3
34	1	1	2	2	2
35	2	2	2	2	2
36	2	1	2	1	3
37	1	1	1	3	2
38	1	1	1	2	2
39	2	1	1	3	1
40	3	2	3	3	3
41	1	1	2	2	2
42	1	1	1	2	2
43	1	1	1	1	1
44	1	1	1	1	1
45	2	2	1	2	2
46	1	2	2	2	2
47	1	1	2	2	1
48	1	2	2	3	3
49	3	3	3	3	3
50	2	2	2	2	2
51	1	1	1	2	2
52	2	1	2	2	2
53	1	1	1	1	1
54	1	2	1	2	3
55	1	2	2	2	2

```
In [48]:
dummy7=alex2["WHICH OTT PLATFORM YOU GENERALLY PREFER WHEN YOUR BORED?"]
dummy7
```

```
0           netflix
1           netflix
2           netflix
3           netflix
4    disney hotstar
5    disney hotstar
6           YouTube
8           netflix
9      amazon prime
10          YouTube
11          netflix
12          netflix
13          netflix
14          netflix
15          netflix
16   disney hotstar
17   disney hotstar
18   disney hotstar
20          netflix
22          netflix
23          netflix
```

Data Splitting for Training and Testing:

we employ the train_test_split function from the metrics library to partition our dataset. we allocate 70% of the data for training and reserve 30% for testing, ensuring a balanced distribution. Setting the random state parameter to 1 ensures reproducibility and consistency in our analysis.

This approach enables us to train our models on a substantial subset while retaining a sizable portion for evaluation. by splitting the data into training and testing sets, we establish a robust framework for model development and evaluation.

```
from sklearn.metrics import accuracy_score as dum
from sklearn.model_selection import train_test_split
a,b,c,d=train_test_split(dummy6,dummy7,test_size=0.3,random_state=1)
```

Model Training:

In this section, we delve into the training process of our predictive models, employing six distinct machine learning algorithms to predict user preferences for OTT platforms. The selected algorithms include

1. Decision Tree Classifier
2. Gaussian Naive Bayes
3. Logistic Regression
4. K Nearest Neighbor (KNN)
5. Support Vector Machine (SVM)
6. Random Forest Classifier

Decision Tree Classifier:

The Decision Tree Classifier utilizes a hierarchical structure of decision rules to classify instances.

```
from sklearn.tree import DecisionTreeClassifier
model1=DecisionTreeClassifier()
model1.fit(a,c)
pred1=model1.predict(b)
print(dum(d,pred1))
```

```
0.35294117647058826
```

Accuracy: 35%

Gaussian Naive Bayes:

Gaussian Naive Bayes relies on probabilistic principles assuming independence among features.

```
from sklearn.naive_bayes import GaussianNB
model2=GaussianNB()
model2.fit(a,c)
pred2=model2.predict(b)
print(dum(d,pred2))
```

```
0.17647058823529413
```

Accuracy: 17%

Logistic Regression:

Logistic Regression estimates the probability of a binary outcome.

```
from sklearn.linear_model import LogisticRegression
model3=LogisticRegression()
model3.fit(a,c)
pred3=model3.predict(b)
print(dum(d,pred3))
```

```
0.35294117647058826
```

Accuracy: 35%

K Nearest Neighbor:

K Nearest Neighbor makes predictions based on the majority class of its nearest neighbors.

```
from sklearn.neighbors import KNeighborsClassifier
model4=KNeighborsClassifier()
model4.fit(a,c)
pred4=model4.predict(b)
print(dum(d,pred4))
```

```
0.47058823529411764
```

Accuracy: 47%

Support Vector Machine (SVM):

SVM constructs hyperplanes in a high-dimensional space to separate classes.

```
from sklearn import svm
model5=svm.SVC()
model5.fit(a,c)
pred5=model5.predict(b)
print(dum(d,pred5))
```

```
0.5882352941176471
```

Accuracy: 58%

Random Forest Classifier:

Random Forest Classifier aggregates the predictions of multiple decision trees to improve accuracy and robustness.

```
from sklearn.ensemble import RandomForestClassifier
model6=RandomForestClassifier()
model6.fit(a,c)
pred6=model6.predict(b)
print(dum(d,pred6))
```

```
0.47058823529411764
```

Accuracy: 47%

Model Evaluation:

SVM Algorithm Outperforms Others with 58% Accuracy

Among the six algorithms utilized, the Support Vector Machine (SVM) algorithm emerged as the top performer, achieving an accuracy of 58%. This result indicates that SVM exhibited the highest predictive power compared to other algorithms, effectively capturing underlying patterns and relationships in the data.

Conclusion:

In conclusion, our comprehensive analysis of machine learning algorithms for predicting user preferences on Over-The-Top (OTT) platforms has provided valuable insights into their performance and effectiveness. Through rigorous model evaluation, we observed that the Support Vector Machine (SVM) algorithm exhibited the highest accuracy among the six algorithms tested, achieving a notable accuracy of 57%