Untitled

August 2, 2022

[1]:
```
#Importing important libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
%matplotlib inline
```

[3]:
```
#machine learning

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC, LinearSVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import Perceptron
from sklearn.linear_model import SGDClassifier
from sklearn.tree import DecisionTreeClassifier
```

[4]:
```
#Importing Datasets
```

[5]:
```
#Import Movie dataset
Movies=pd.read_csv("movies.dat",sep="::
 ↪",names=["MovieID","Tittle","Genres"],engine='python')
Movies.head()
```

[5]:

	MovieID	Tittle	Genres
0	1	Toy Story (1995)	Animation|Children's|Comedy
1	2	Jumanji (1995)	Adventure|Children's|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama
4	5	Father of the Bride Part II (1995)	Comedy

[6]:
```
#Importing Ratings dataset
ratings=pd.read_csv("ratings.dat",sep="::
 ↪",names=["UserID","MovieID","Rating","Timestamp"],engine='python')
```

```
ratings.head()
```

[6]:

	UserID	MovieID	Rating	Timestamp
0	1	1193	5	978300760
1	1	661	3	978302109
2	1	914	3	978301968
3	1	3408	4	978300275
4	1	2355	5	978824291

[7]:
```
#Importing Users Dataset
users=pd.read_csv("users.dat",sep="::
 ↪",names=["UserID","Gender","Age","Occupation","Zip-code"],engine='python')
users.head()
```

[7]:

	UserID	Gender	Age	Occupation	Zip-code
0	1	F	1	10	48067
1	2	M	56	16	70072
2	3	M	25	15	55117
3	4	M	45	7	02460
4	5	M	25	20	55455

[8]:
```
#Shapes of Datasets
print("Movies dataset Shape:",Movies.shape)
print("Users dataset Shape:",users.shape)
print("Ratings dataset Shape:",ratings.shape)
```

```
Movies dataset Shape: (3883, 3)
Users dataset Shape: (6040, 5)
Ratings dataset Shape: (1000209, 4)
```

[9]:
```
#CREATING A NEW DATA SET[MASTER-DATA]
```

[10]:
```
#Merging movies and ratings datasets on=Key MovieID
movies_ratings=pd.merge(Movies,ratings,how='inner',on='MovieID')
```

[11]:
```
movies_ratings.head()
```

[11]:

	MovieID	Tittle	Genres	UserID	Rating	\
0	1	Toy Story (1995)	Animation|Children's|Comedy	1	5	
1	1	Toy Story (1995)	Animation|Children's|Comedy	6	4	
2	1	Toy Story (1995)	Animation|Children's|Comedy	8	4	
3	1	Toy Story (1995)	Animation|Children's|Comedy	9	5	
4	1	Toy Story (1995)	Animation|Children's|Comedy	10	5	

	Timestamp
0	978824268
1	978237008

```
2  978233496
3  978225952
4  978226474
```

[12]:
```
#Merging users dataset on the key UserID
df_final = pd.merge(users,movies_ratings,how='inner',on='UserID')
```

[13]:
```
df_final.head()
```

[13]:
```
   UserID Gender  Age  Occupation Zip-code  MovieID  \
0       1      F    1          10    48067        1
1       1      F    1          10    48067       48
2       1      F    1          10    48067      150
3       1      F    1          10    48067      260
4       1      F    1          10    48067      527

                                    Tittle  \
0                         Toy Story (1995)
1                       Pocahontas (1995)
2                        Apollo 13 (1995)
3  Star Wars: Episode IV - A New Hope (1977)
4                   Schindler's List (1993)

                                Genres  Rating  Timestamp
0          Animation|Children's|Comedy       5  978824268
1  Animation|Children's|Musical|Romance       5  978824351
2                               Drama       5  978301777
3      Action|Adventure|Fantasy|Sci-Fi       4  978300760
4                           Drama|War       5  978824195
```

[14]:
```
del df_final['Genres']
```

[15]:
```
del df_final['Timestamp']
```

[16]:
```
del df_final['Zip-code']
```

[17]:
```
df_final.head()
```

[17]:
```
   UserID Gender  Age  Occupation  MovieID  \
0       1      F    1          10        1
1       1      F    1          10       48
2       1      F    1          10      150
3       1      F    1          10      260
4       1      F    1          10      527

                                    Tittle  Rating
0                         Toy Story (1995)       5
```

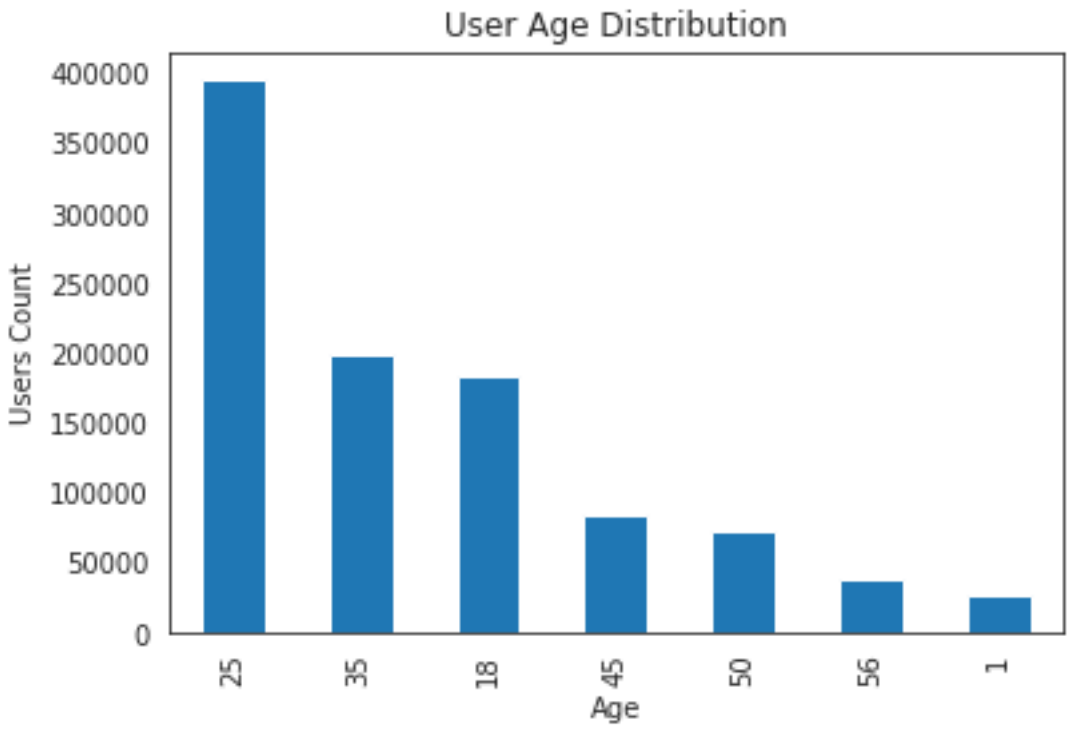
```
1                          Pocahontas (1995)       5
2                           Apollo 13 (1995)       5
3  Star Wars: Episode IV - A New Hope (1977)       4
4                    Schindler's List (1993)       5
```

[18]:
```
#DATA EXPLORATION
```

[21]:
```
#Age Distribution
age_dist = df_final['Age'].value_counts().to_frame()
age_dist.sort_index(inplace=True)
age_dist
```

[21]:
```
       Age
1    27211
18  183536
25  395556
35  199003
45   83633
50   72490
56   38780
```

[28]:
```
#Age distribution visualisation
df_final['Age'].value_counts().plot(kind='bar')
plt.xlabel("Age")
plt.title("User Age Distribution")
plt.ylabel('Users Count')
plt.show()
```



[29]:
```
#Comments
#The data indicates that most of the users are 23-35 in age
#and the least of the users 56+ of age.
```

[31]:
```
#User rating on "Toy Story" movie
Toy_stort_df=df_final[df_final['Tittle'] == "Toy Story (1995)"]
TS_rating = Toy_stort_df['Rating'].value_counts().to_frame()
TS_rating
```

[31]:

	Rating
4	835
5	820
3	345
2	61
1	16

[ ]:
```

```