MOVIE RATING PREDICTION WITH PYTHON

INTRODUCTION:

Predicting movie ratings can be an interesting and challenging task in the field of data science and machine learning. With the help of Python and various machine learning techniques, it's possible to create models that predict how well a movie might be rated based on different features such as cast, crew, genre, release date, and more.

To start building a movie rating prediction system using Python, you'll typically follow these steps:

1. **Data Collection**: Obtain a dataset containing information about movies, including features like title, genre, director, actors, budget, box office earnings, release date, ratings, etc. This data can be collected from various sources like IMDb, Kaggle, or other movie databases.
2. **Data Preprocessing**: Clean the dataset by handling missing values, removing duplicates, converting categorical variables to numerical format (using techniques like one-hot encoding or label encoding), scaling numerical features if needed, and performing other necessary transformations.
3. **Feature Selection/Engineering**: Analyze the dataset to identify important features that might significantly influence movie ratings. This step involves selecting relevant features or creating new ones based on domain knowledge and statistical analysis.
4. **Splitting Data**: Divide the dataset into training and testing sets to train the machine learning model on a portion of the data and evaluate its performance on unseen data.
5. **Model Selection and Training**: Choose appropriate machine learning algorithms (such as linear regression, decision trees, random forests, gradient boosting, etc.) for regression (since movie ratings are continuous values) and train the model using the training data.
6. **Model Evaluation**: Evaluate the trained model's performance using various metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared score, etc., on the test dataset to assess how well the model generalizes to new data.
7. **Hyperparameter Tuning (Optional)**: Fine-tune the model by adjusting its hyperparameters to improve its performance.
8. **Prediction**: Use the trained model to make predictions on new or unseen movie data to estimate their ratings.

LOAD THE DATASET:

Import padas

Import sklearn

Import matplotlib

PROGRAM:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.metrics import mean_absolute_error,
mean_squared_error, r2_score

from sklearn.linear_model import SGDRegressor

from sklearn.preprocessing import StandardScaler

from sklearn.pipeline import Pipeline

df =
pd.read_csv('C:/Users/abcd/Downloads/movie_data.csv
')

df.head()

def dataoveriew(df, message):

    print(f'{message}:\n')

    print("Rows:", df.shape[0])

    print("\nNumber of features:", df.shape[1])

    print("\nFeatures:")
```

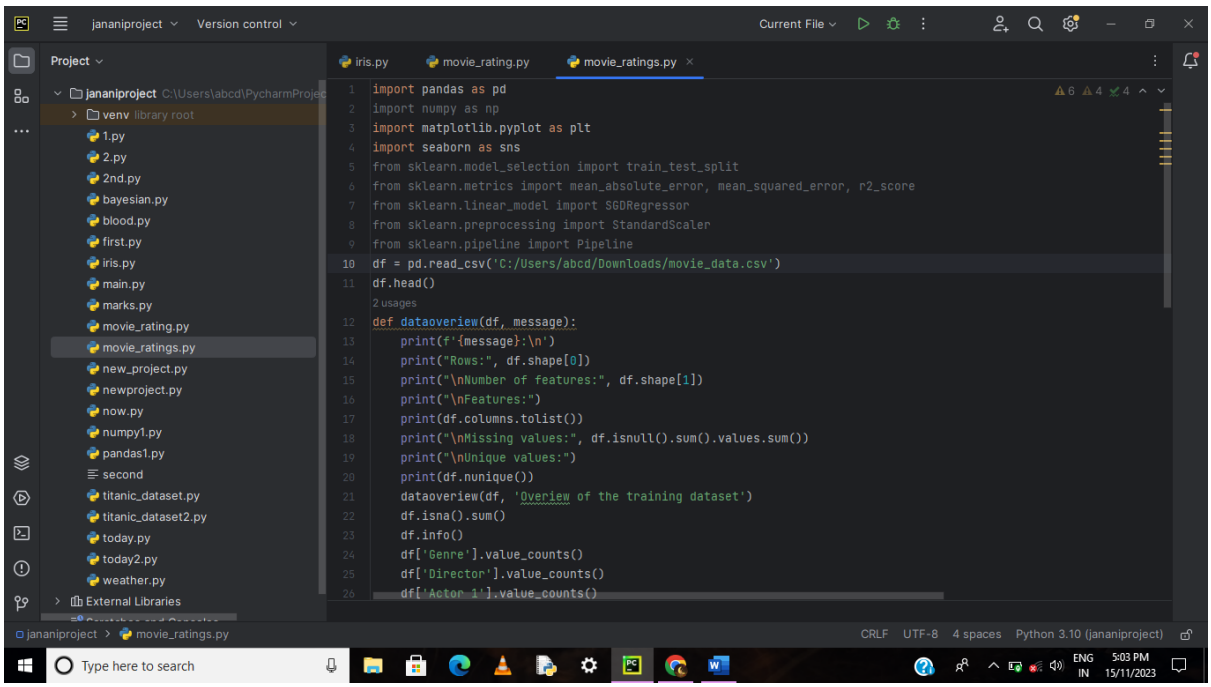
```
print(df.columns.tolist())
print("\nMissing values:", df.isnull().sum().values.sum())
print("\nUnique values:")
print(df.nunique())
dataoveriew(df, 'Overiew of the training dataset')
df.isna().sum()
df.info()
df['Genre'].value_counts()
df['Director'].value_counts()
df['Actor 1'].value_counts()
df.head(10)
# As we are going to predict movie ratings based on features, we need to remove null values from features that can directly influence the results.
df.dropna(subset=['Name', 'Year', 'Duration', 'Votes', 'Rating'], inplace=True)
df.isna().sum()
df.head()
dataoveriew(df, 'Overiew of the training dataset')
# Remove parentheses from 'Year' column and convert to integer
df['Year'] = df['Year'].str.strip('()').astype(int)
```

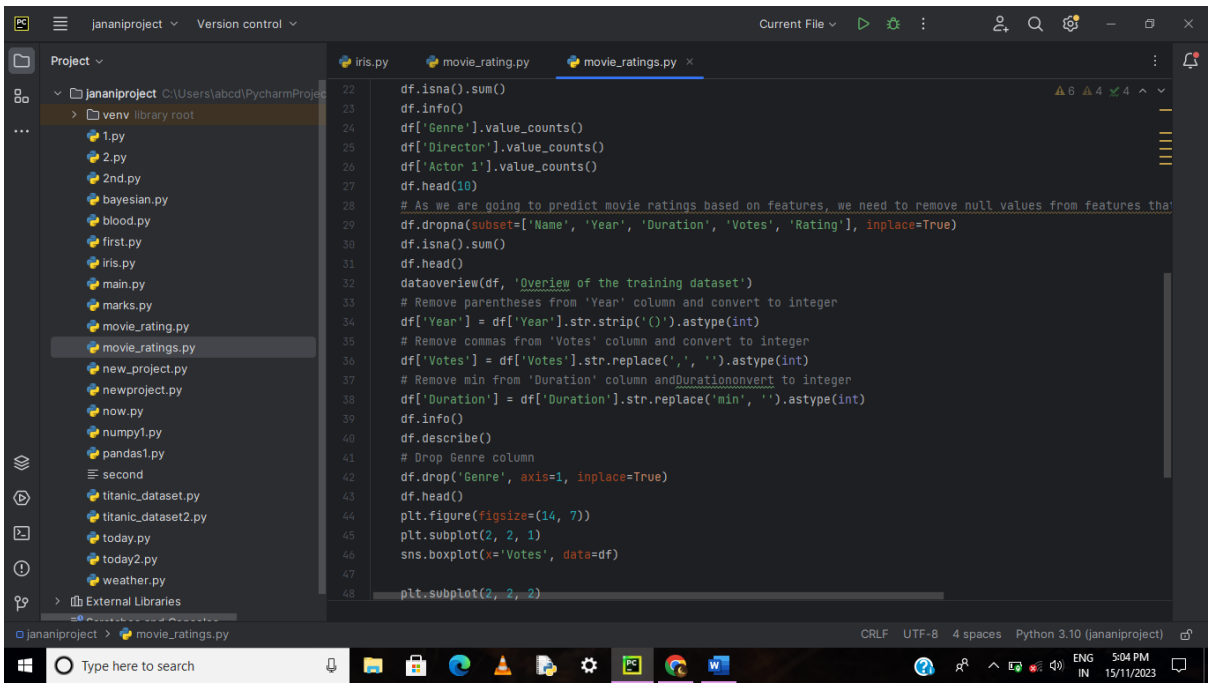
```
# Remove commas from 'Votes' column and convert to integer
df['Votes'] = df['Votes'].str.replace(',', '').astype(int)
# Remove min from 'Duration' column andDurationonvert to integer
df['Duration'] = df['Duration'].str.replace('min', '').astype(int)
df.info()
df.describe()
# Drop Genre column
df.drop('Genre', axis=1, inplace=True)
df.head()
plt.figure(figsize=(14, 7))
plt.subplot(2, 2, 1)
sns.boxplot(x='Votes', data=df)

plt.subplot(2, 2, 2)
sns.distplot(df['Year'], color='g')

plt.subplot(2, 2, 3)
sns.distplot(df['Rating'], color='g')

plt.subplot(2, 2, 4)
```

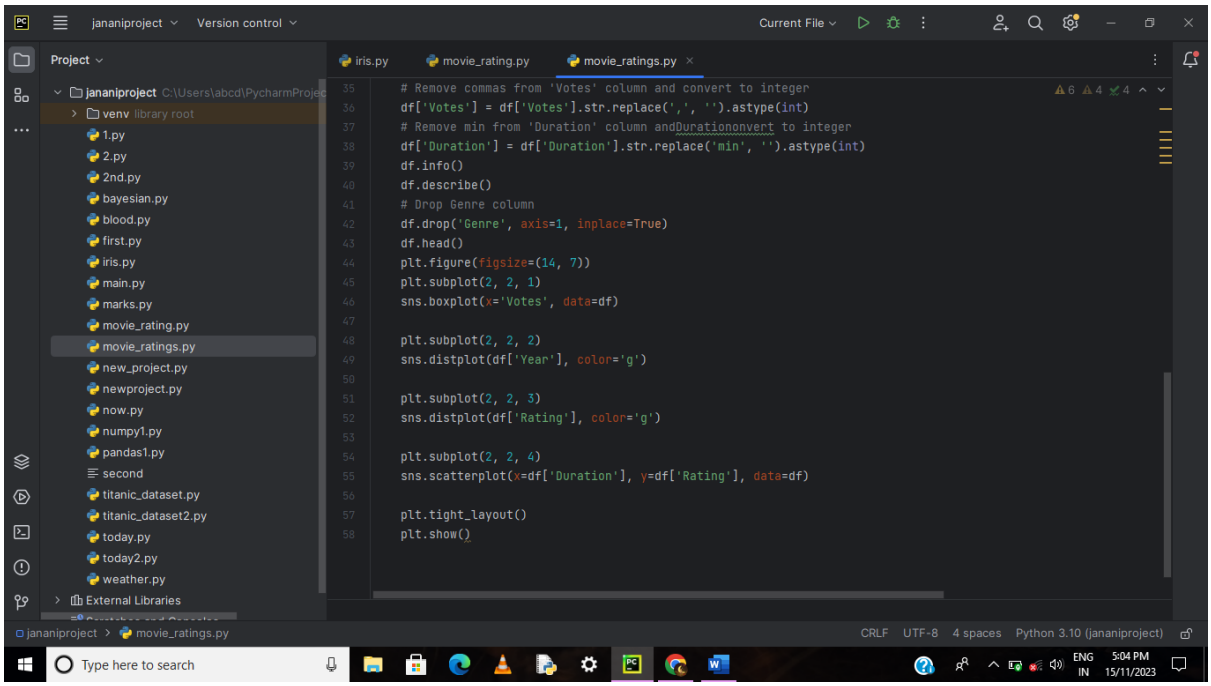
sns.scatterplot(x=df['Duration'], y=df['Rating'], data=df)

plt.tight_layout()

plt.show()







OUTPUT:

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
0		NaN	NaN	Drama	NaN	NaN	J.S. Randhawa	Manmauji	Birbal	Rajendra Bhatia
1	#Gadhvi (He thought he was Gandhi)	(2019)	109 min	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
2	#Homecoming	(2021)	90 min	Drama, Musical	NaN	NaN	Soumyajit Majumdar	Sayani Gupta	Plabita Borthakur	Roy Angana
3	#Yaaram	(2019)	110 min	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor
4	...And Once Again	(2010)	105 min	Drama	NaN	NaN	Amol Palekar	Rajat Kapoor	Rituparna Sengupta	Antara Mali

```
Overiew of the training dataset:

Rows: 15509

Number of features: 10

Features:
['Name', 'Year', 'Duration', 'Genre', 'Rating', 'Votes', 'Director', 'Actor 1', 'Actor 2', 'Actor 3']

Missing values: 33523

Unique values:
Name        13838
Year          102
Duration      182
Genre         485
Rating         84
Votes        2034
Director     5938
```

```
Name           0
Year         528
Duration    8269
Genre       1877
Rating      7590
Votes       7589
Director     525
Actor 1     1617
Actor 2     2384
Actor 3     3144
dtype: int64
```

```
Drama                          2780
Action                         1289
Thriller                        779
Romance                         708
Drama, Romance                  524
                               ...
Action, Musical, War              1
Horror, Crime, Thriller           1
Animation, Comedy                 1
Romance, Action, Crime            1
Adventure, Fantasy, Sci-Fi        1
Name: Genre, Length: 485, dtype: int64
```

```
Jayant Desai        58
Kanti Shah          57
Babubhai Mistry     50
Mahesh Bhatt        48
Master Bhagwan      47
                    ..
Naeem Siddiqui       1
Shadaab Khan         1
Mystelle Brabbee     1
Kunal Shivdasani     1
Kiran Thej           1
Name: Director, Length: 5938, dtype: int64
```

```
Ashok Kumar           158
Dharmendra            140
Jeetendra             140
Mithun Chakraborty    133
Amitabh Bachchan      129
                     ...
Vatsal Sheth            1
Ujala Baboria           1
Dimple Sewak            1
Komal Leels             1
Sangeeta Tiwari         1
Name: Actor 1, Length: 4718, dtype: int64
```

```
Name          0
Year          0
Duration      0
Genre        31
Rating        0
Votes         0
Director      1
Actor 1      75
Actor 2     117
Actor 3     163
dtype: int64
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
1	#Gadhvi (He thought he was Gandhi)	(2019)	109 min	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
3	#Yaaram	(2019)	110 min	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor
5	...Aur Pyaar Ho Gaya	(1997)	147 min	Comedy, Drama, Musical	4.7	827	Rahul Rawail	Bobby Deol	Aishwarya Rai Bachchan	Shammi Kapoor
6	...Yahaan	(2005)	142 min	Drama, Romance, War	7.4	1,086	Shoojit Sircar	Jimmy Sheirgill	Minissha Lamba	Yashpal Sharma
8	?: A Question Mark	(2012)	82 min	Horror, Mystery, Thriller	5.6	326	Allyson Patel	Yash Dave	Muntazir Ahmad	Kiran Bhatia

```
Overiew of the training dataset:

Rows: 5851

Number of features: 10

Features:
['Name', 'Year', 'Duration', 'Genre', 'Rating', 'Votes', 'Director', 'Actor 1', 'Actor 2', 'Actor 3']

Missing values: 387

Unique values:
Name        5570
Year          91
Duration     178
Genre        393
Rating        83
Votes       2030
Director    2549
```

|:

	Year	Duration	Rating	Votes
count	5851.000000	5851.000000	5851.000000	5851.000000
mean	1996.416852	132.294480	5.931875	2611.273116
std	19.914640	26.555826	1.389942	13433.828528
min	1931.000000	21.000000	1.100000	5.000000
25%	1983.000000	117.000000	5.000000	28.000000
50%	2002.000000	134.000000	6.100000	119.000000
75%	2013.000000	150.000000	7.000000	862.500000
max	2021.000000	321.000000	10.000000	591417.000000

[20]:

	Name	Year	Duration	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
1	#Gadhvi (He thought he was Gandhi)	2019	109	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
3	#Yaaram	2019	110	4.4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor
5	...Aur Pyaar Ho Gaya	1997	147	4.7	827	Rahul Rawail	Bobby Deol	Aishwarya Rai Bachchan	Shammi Kapoor
6	...Yahaan	2005	142	7.4	1086	Shoojit Sircar	Jimmy Sheirgill	Minissha Lamba	Yashpal Sharma
8	?: A Question Mark	2012	82	5.6	326	Allyson Patel	Yash Dave	Muntazir Ahmad	Kiran Bhatia

CONCLUSION:

This is the program of movie rating prediction using python. Here the output of these.