Phase 4

APPLIED DATA SCIENCE

Model Training and Evaluation

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
data = pd.read_csv(r"imdb_top_1000 - Copy.csv")
print(data.head())
```

```
                                         Poster_Link  \
0  https://m.media-amazon.com/images/M/MV5BMDFkYT...
1  https://m.media-amazon.com/images/M/MV5BM2MyNj...
2  https://m.media-amazon.com/images/M/MV5BMTMxNT...
3  https://m.media-amazon.com/images/M/MV5BMWMwMG...
4  https://m.media-amazon.com/images/M/MV5BMWU4N2...

              Series_Title  Released_Year  Runtime                Genre  \
0  The Shawshank Redemption           1994  142 min                Drama
1             The Godfather           1972  175 min          Crime, Drama
2           The Dark Knight           2008  152 min  Action, Crime, Drama
3    The Godfather: Part II           1974  202 min          Crime, Drama
4              12 Angry Men           1957   96 min          Crime, Drama

   IMDB_Rating                                           Overview  Meta_score  \
0          9.3  Two imprisoned men bond over a number of years...          80
1          9.2  An organized crime dynasty's aging patriarch t...         100
2          9.0  When the menace known as the Joker wreaks havo...          84
3          9.0  The early life and career of Vito Corleone in ...          90
4          9.0  A jury holdout attempts to prevent a miscarria...          96

              Director           Star1           Star2          Star3  \
0       Frank Darabont     Tim Robbins  Morgan Freeman     Bob Gunton
1  Francis Ford Coppola   Marlon Brando       Al Pacino     James Caan
2    Christopher Nolan  Christian Bale    Heath Ledger  Aaron Eckhart
3  Francis Ford Coppola       Al Pacino  Robert De Niro  Robert Duvall
4         Sidney Lumet     Henry Fonda     Lee J. Cobb  Martin Balsam

            Star4  No_of_Votes        Gross Certificate  Label
0  William Sadler      2343110  2,83,41,469           A      1
1    Diane Keaton      1620367  13,49,66,411          A      1
2   Michael Caine      2303232  53,48,58,444         UA      2
3    Diane Keaton      1129952   5,73,00,000          A      1
4    John Fiedler       689845     43,60,000          U      3
```

```
print(data.columns)

feature=data[['Meta_score','IMDB_Rating' ]]

#independent var
x=np.asarray(feature)

#dependent var
y=np.asarray(data['Label'])
```

```
Index(['Poster_Link', 'Series_Title', 'Released_Year', 'Runtime', 'Genre',
       'IMDB_Rating', 'Overview', 'Meta_score', 'Director', 'Star1', 'Star2',
       'Star3', 'Star4', 'No_of_Votes', 'Gross', 'Certificate', 'Label'],
      dtype='object')
```

```
from sklearn.model_selection import train_test_split
from sklearn import tree
clf = tree.DecisionTreeClassifier()

clf.fit(x_train, y_train)
#decision tree

y_predict2=clf.predict(x_test)
x_train, x_test , y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=1 )
```

```
from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_predict2)
print(cm)
```

```
[[4 0 5]
 [3 1 0]
 [4 0 1]]
```

```
from sklearn.metrics import accuracy_score

accuracy = accuracy_score(y_test, y_predict)
print('Accuracy (Linear Kernel): ', "%.2f" % (accuracy*100))
```

```
Accuracy (Linear Kernel):  50.00
```

```
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
#calculating precision and reall
precision = precision_score(y_test, y_predict2, average='weighted')
recall = recall_score(y_test, y_predict2, average='weighted')
print('Precision: ', "%.2f" % (precision*100))
print('Recall: ', "%.2f" % (recall*100))
```

```
Precision:  45.03
Recall:  33.33
```