PREDICTING IMDB SCORES - ADS_PHASE3

TEAM NUMBER

Problem Statement : Loading and Preprocessing

In this part we will begin building our project by loading and preprocessing the dataset.

We have begin building the IMDb score prediction model by loading and preprocessing the dataset.

```
#importing necessary libraries
import pandas as pd
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.impute import SimpleImputer from
sklearn.model_selection import train_test_split import
warnings
warnings.simplefilter(action='ignore', category=FutureWarning)

#importing the netflix dataset
file_path = r"C:\Users\Saranya\Desktop\IBM\NetflixOriginals.csv"
encoding = "ISO-8859-1"
df = pd.read_csv(file_path, encoding=encoding)
df
```

```
                                          Title
Genre  \
0                                  Enter the Anime
                                   Documentary
1                                  Dark Forces
Thriller
2                                  The App  Science fiction/Drama
3                                  The Open House        Horror
                                   thriller
4                                  Kaali Khuhi
Mystery
..                                           ...                    ..
.
579    Taylor Swift: Reputation Stadium Tour          Concert Film
580    Winter on Fire: Ukraine's Fight for Freedom
Documentary
581    Springsteen on Broadway           One-man show
582    Emicida: AmarElo - It's All For Yesterday
Documentary
```

```
583    David Attenborough: A Life on Our Planet            Documentary

              Premiere  Runtime  IMDB Score                  Language

0        August 5, 2019       58         2.5          English/Japanese

1       August 21, 2020       81         2.6                   Spanish

2     December 26, 2019       79         2.6                   Italian

3      January 19, 2018       94         3.2                   English

4      October 30, 2020       90         3.4                     Hindi

..                  ...      ...         ...                       ...

579  December 31, 2018      125         8.4                   English

580    October 9, 2015       91         8.4  English/Ukranian/Russian

581  December 16, 2018      153         8.5                   English

582   December 8, 2020       89         8.6                Portuguese


583    October 4, 2020       83         9.0                   English


[584 rows x 6 columns]
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 584 entries, 0 to 583
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  -
--  ------      --------------  -----   0
Title       584 non-null    object
1   Genre       584 non-null    object
2   Premiere    584 non-null    object
3   Runtime     584 non-null    int64   4   IMDB Score  584 non-null
    float64  5   Language    584 non-null    object dtypes:
    float64(1), int64(1), object(4) memory usage: 27.5+ KB
```

```
d              Title                    Genre           Premiere  Runtime

   Enter the Anime            Documentary     August 5, 2019        58


\
0

1       Dark Forces                Thriller    August 21, 2020        81

2          The App  Science fiction/Drama  December 26, 2019        79

3    The Open House         Horror thriller   January 19, 2018        94
4       Kaali Khuhi                 Mystery   October 30, 2020        90
   IMDB Score          Language
0         2.5  English/Japanese
1         2.6           Spanish
2         2.6           Italian
3         3.2           English
4         3.4             Hindi
#to display null values
df.isnull()
     Title  Genre  Premiere  Runtime  IMDB Score  Language
0     False  False     False    False       False     False
1     False  False     False    False       False     False
2     False  False     False    False       False     False
3     False  False     False    False       False     False
4     False  False     False    False       False     False..
...     ...        ...       ...          ...        ... 579
False  False     False    False       False     False
580  False  False     False    False       False     False
581  False  False     False    False       False     False
582  False  False     False    False       False     False
583  False  False     False    False       False     False
[584 rows x 6 columns]
#handling null values

df.fillna(df.mean(), inplace=True) df.dropna(inplace=True)

#Display distinct languages
value_lang = df['Language'].value_counts()
print("\nDistinct languages:")
print(value_lang)
```

```
Distinct languages:

English                         401
Hindi                            33
Spanish                          31
French                           20
Italian                          14
Portuguese                       12
Indonesian                        9
Japanese                          6
Korean                            6
German                            5
Turkish                           5
English/Spanish                   5
Polish                            3
Dutch                             3
Marathi                           3
English/Hindi                     2
Thai                              2
English/Mandarin                  2
English/Japanese                  2
Filipino                          2
English/Russian                   1
Bengali                           1
English/Arabic                    1
English/Korean                    1
Spanish/English                   1
Tamil                             1
English/Akan                      1
Khmer/English/French              1
Swedish                           1
Georgian                          1
Thia/English                      1
English/Taiwanese/Mandarin        1
English/Swedish                   1
Spanish/Catalan                   1
Spanish/Basque                    1
Norwegian                         1
Malay                             1
English/Ukranian/Russian          1
Name: Language, dtype: int64
```

```
distinct_lang = df['Language'].unique()
print(distinct_lang)
```

```
['English/Japanese' 'Spanish' 'Italian' 'English' 'Hindi' 'Turkish'
 'Korean' 'Indonesian' 'Malay' 'Dutch' 'French' 'English/Spanish'
'Portuguese' 'Filipino' 'German' 'Polish' 'Norwegian' 'Marathi' 'Thai'
 'Swedish' 'Japanese' 'Spanish/Basque' 'Spanish/Catalan'
```

```
'English/Swedish'
 'English/Taiwanese/Mandarin' 'Thia/English' 'English/Mandarin'
'Georgian'
 'Bengali' 'Khmer/English/French' 'English/Hindi' 'Tamil'
 'Spanish/English' 'English/Korean' 'English/Arabic' 'English/Russian'
 'English/Akan' 'English/Ukranian/Russian']
```

```
#label encoder for language column

label_encoder = LabelEncoder()
df['Language'] = label_encoder.fit_transform(df['Language'])
df
```

```
                                          Title
Genre  \
0                                Enter the Anime
                                 Documentary
1                                Dark Forces
Thriller
2                                The App  Science fiction/Drama
3                                The Open House        Horror
                                 thriller
4                                Kaali Khuhi
Mystery
..                                          ...                       ..
.
579        Taylor Swift: Reputation Stadium Tour           Concert
Film
580        Winter on Fire: Ukraine's Fight for Freedom
Documentary
581        Springsteen on Broadway           One-man show
582        Emicida: AmarElo - It's All For Yesterday
Documentary
583        David Attenborough: A Life on Our Planet
Documentary

              Premiere  Runtime  IMDB Score  Language  0
August 5, 2019       58         2.5         6
1      August 21, 2020       81         2.6        29
2      December 26, 2019       79         2.6        20
3     January 19, 2018       94         3.2                        2
4     October 30, 2020       90         3.4        18              2
..                 ...      ...         ...       ...  579  December
31, 2018      125         8.4                                      2
580    October 9, 2015       91         8.4        13              2
581  December 16, 2018      153         8.5                        2
582   December 8, 2020       89         8.6        28              2
583    October 4, 2020       83         9.0                        2
```

```
[584 rows x 6 columns]
```

```
#scaling

scaler = StandardScaler()
df['Runtime'] = scaler.fit_transform(df['Runtime'].values.reshape(-1,
1))
df
```

```
                                        Title
Genre  \
0                              Enter the Anime
                                        Documentary
1                                       Dark Forces
Thriller
2                                       The App  Science fiction/Drama
3                                       The Open House        Horror
                                        thriller
4                                       Kaali Khuhi
Mystery
..                                                ...                    ..
.
579        Taylor Swift: Reputation Stadium Tour           Concert
Film
580        Winter on Fire: Ukraine's Fight for Freedom
Documentary
581        Springsteen on Broadway           One-man show
582        Emicida: AmarElo - It's All For Yesterday
Documentary
583        David Attenborough: A Life on Our Planet
           Documentary

              Premiere   Runtime  IMDB Score  Language
0       August 5, 2019 -1.282615         2.5         6
1       August 21, 2020 -0.453425         2.6        29
2       December 26, 2019 -0.525528         2.6        20
3       January 19, 2018  0.015248         3.2         2
4       October 30, 2020 -0.128959         3.4        18  ..
        ...       ...         ...       ...  579  December 31, 2018
        1.132852         8.4         2
580    October 9, 2015 -0.092907         8.4        13
581    December 16, 2018  2.142301         8.5         2
582    December 8, 2020 -0.165011         8.6        28
583    October 4, 2020 -0.381321         9.0         2

[584 rows x 6 columns]
```

```
#train_test split
```

```
X = df.drop('IMDB Score', axis=1)
y = df['IMDB Score']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

print("\n X_test info")
print(X_test.info())
```

```
 X_test info
<class 'pandas.core.frame.DataFrame'>
Int64Index: 117 entries, 383 to 362
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype  ---
------    --------------  -----
0   Title     117 non-null    object
1   Genre     117 non-null    object
2   Premiere  117 non-null    object  3   Runtime   117 non-null
    float64  4   Language  117 non-null    int32  dtypes: float64(1),
    int32(1), object(3)
memory usage: 5.0+ KB
None
```