PREDICTING IMDB SCORES - PHASE 3

PROBLEM:

The problem is to develop a machine learning model that predicts IMDB scores of movies available on films based on features like genre, premiere data, runtime and language. The objective is to create a model that accurately estimates the popularity of movies, helping user discover highly rated movies that matches their preferences. This project involves data preprocessing, feature engineering, model selection, training and evaluation.

DATA PREPROCESSING:

- Firstly, we import the necessary libraries (StandardScaler from sklearn).
- Then, we join the two dataframes(df2 and y) and we name the combined dataframe as data.
- We call the StandardScaler() function. Now fit the dataframe and transform it.
- We take the null values in the dataframe and add it(isnull().sum()).
- Now we drop the null values by using dropna.
- We sort the values by IMDBScore.
- We take the head value of the sorted dataframe.
- We check whether there are null values in the sorted dataset using isna().
- Lastly, we import MinMaxScaler.

UNIVARIATE ANALYSIS:

- We display the Runtime column from the dataset as column1.
- We display the head of column1.
- We drop the null values in column1 and fill the empty places with "Nan"
- We apply info() and describe() on column1.
- We take the first 10 head values of column1 and store it as column11.
- Now we plot a horizontal bar chart using column11.
- Univariate analysis is completed.

BIVARIATE ANALYSIS:

- We group the dataset (df) by Genre and find the mean of IMDBScore using aggregate function.
- We group the dataset (df) by Runtime and find the mean of IMDBScore using aggregate function.
- We group the dataset (df) by Premiere and find the mean of IMDBScore using aggregate function.
- We group the dataset (df) by Title and find the mean of IMDBScore using aggregate function.
- We group the dataset (df) by Language and find the mean of IMDBScore using aggregate function.