MOVIE GENRE CLASSIFICATION

This project involves building a machine learning model to predict the genre of a movie based on its plot summary. The text data is cleaned and preprocessed before being transformed using TF-IDF vectorization. A Multinomial Naive Bayes classifier is trained on the processed data, and its performance is evaluated using a validation set. Finally, the model is used to predict genres for a test dataset, and the results are saved for further analysis.

Steps :

1. Importing Necessary Libraries
First, essential libraries for data manipulation, visualization, text processing, and machine learning are imported. These include libraries like NumPy, pandas, Matplotlib, Seaborn, NLTK, and scikit-learn.

2. Loading the Training Data
The training data is loaded from a text file. The first few rows and the dataset summary are displayed to understand its structure and contents.

3. Loading the Test Data
The test data is similarly loaded from a text file. The first few rows and the dataset summary are displayed.

4. Loading the Test Solution Data
The test solution data is loaded to compare the model's predictions later. Unnecessary columns are removed, and the remaining column is renamed for clarity.

5. Data Visualization
The occurrences of each genre in the training data are counted and visualized using count plots and bar plots. This helps understand the distribution of genres in the dataset.

6. Data Preprocessing
Information about the dataset is displayed. NLTK resources for text processing, such as stop words and tokenizers, are downloaded. The Lancaster Stemmer and stop words are initialized.

7. Defining the Text Cleaning Function
A function is defined to clean the text data by removing unwanted characters, punctuation, and stop words. The text is also converted to lowercase.

8. Applying the Text Cleaning Function
The text cleaning function is applied to the 'Description' column in both the training and test data.

9. Dropping Duplicates
Duplicate rows are removed from the training data to ensure data quality.

10. Calculating Length of Cleaned Text
The length of the cleaned text for each row is calculated and stored in a new column.

11. Visualizing the Distribution of Text Lengths

The distribution of text lengths is visualized using histograms, comparing the lengths before and after cleaning.

12. Text Vectorization using TF-IDF
The TF-IDF vectorizer is initialized. The training data is fitted and transformed, and the test data is transformed using this vectorizer.

13. Splitting Data and Training a Model (Naive Bayes)
The data is split into training and validation sets. A Multinomial Naive Bayes classifier is initialized and trained on the training set. Predictions are made on the validation set, and the model's performance is evaluated.

14. Making Predictions on Test Data and Saving Results
The trained model is used to make predictions on the test data. The test data with predicted genres is saved to a CSV file.

15. Displaying the Test Data with Predicted Genres
The test data with the predicted genres is displayed, showing the results of the model's predictions.