CodSoft: Data Science Internship

Predicting Movie Ratings with Python

Description: The objective of this project is to build a predictive model that estimates the rating of a movie based on its key attributes. Ratings are crucial indicators of a movie's success and are influenced by various factors such as:

- **Year of Release**: Older movies might have different rating trends compared to newer ones.
- **Number of Votes**: More votes might indicate a broader audience, possibly leading to higher or more stable ratings.
- **Duration**: The length of the movie could impact its reception; both very short and very long movies may have distinctive rating patterns.
- **Genre**: Different genres appeal to different audiences, which can affect ratings.
- **Director**: The reputation and past work of a director can greatly influence audience expectations and ratings.
- **Actors**: Lead actors can attract their fan base, potentially boosting ratings.

Dataset : [Movie Prediction]

Tools and Libraries:

- Python: Programming language used for the entire pipeline.
- Pandas & NumPy: For data manipulation and preprocessing.

- Matplotlib & Seaborn: For data visualization.
- Scikit-learn: For machine learning algorithms and model evaluation.

Key Features:

- **User Input**: The function starts by collecting input from the user. It asks for details like the year of release, number of votes, duration of the movie, genre, director's name, and the names of the lead actors.
- **Data Preparation**: After gathering the input, the details are stored in a dictionary and then converted into a pandas DataFrame. This is essential for aligning the input data with the training data used in the machine learning model.
- **Data Cleaning**: To handle any potential errors, the function converts the 'Votes' and 'Duration' to numeric values. If any of these fields contain invalid data, they are replaced with the median values from the training dataset.
- **Feature Encoding**: The input DataFrame is then transformed using one-hot encoding to handle categorical variables such as Genre, Director, and Actors. This ensures that the input data is in the same format as the data used to train the model.
- **Model Prediction**: The encoded input data is aligned with the features used during the model's training phase. If any columns are missing, they are filled with zeros. Finally, the pre-trained model is used to predict the movie's rating based on the provided details.
- **Output**: The predicted rating is then returned and printed out, giving the user an estimate of the movie's rating.