Mini Project: IMDb Rating Prediction Using Machine Learning

Problem Statement:

This project aims to predict the IMDb ratings of TV shows based on various features such as movie name, genre, year of release, and the number of votes. The goal is to build a Machine Learning model that can predict the IMDb rating of a given show or movie, using historical data from the IMDb Top 250 dataset.

Dataset Link:
IMDb Top 250 Shows Dataset

Project Requirements:

1. **Python Libraries:**
    - `pandas` for data manipulation.
    - `matplotlib` and `seaborn` for data visualization.
    - `scikit-learn` for machine learning models and preprocessing.
    - `numpy` for numerical operations.
2. **Modeling Approach:**
    - Use Machine Learning models to predict IMDb ratings.
    - Try different models Various Machine Learning Models.
    - Perform model evaluation using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

Project Steps:

1. Data Loading, Exploration and Preprocessing:

Objective:
Explore and prepare the data for modeling.

Task:

- Load the dataset and examine its structure (check for missing values, data types, and basic statistics).
- Understand the features: Title, Genre, Year, Votes, and Rating.
- Convert categorical features (like `Genre`) into numerical features using techniques such as one-hot encoding.
- Handle missing values (if any) and outliers.
- Check for any correlations between the features and the target variable (`Rating`).

2. Data Visualization:

Objective:
Visualize the relationships between different variables to gain insights.

Task:

- Create a scatter plot to show the relationship between the number of votes and IMDb ratings.
- Plot the distribution of the target variable (`Rating`) to understand its spread.
- Visualize the correlation heatmap between all the numeric variables.
- Create bar charts to analyze the distribution of movies across different genres, and how each genre correlates with the ratings.

3. Feature Engineering:

Objective:
Create useful features that could improve the model's performance.

Task:

- Extract the year from the `Year` column and categorize it into different time ranges (e.g., 1990-2000, 2000-2010).
- Process the `Genre` column and one-hot encode it (since the genre is categorical and can have multiple values).
- Consider dropping or transforming features that might not be useful (like `Title`, which may not provide predictive value).

4. Splitting the Dataset:

Objective:
Prepare the data for training and testing.

Task:

- Split the data into training and testing datasets (e.g., 80% training and 20% testing).
- Normalize or scale numeric features (e.g., `Votes`) to ensure the model performs optimally.

5. Model Building:

Objective:
Train different Machine Learning models and compare their performance.

6. Model Evaluation:

Objective:
Evaluate model performance using appropriate metrics.

Task:

- Evaluate the models using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.
- Compare the results of different models and identify the one that performs the best.
- Use residual plots to check the model's prediction errors.

7. Hyperparameter Tuning:

Objective:
Optimize the performance of the models.

Task:

- Perform hyperparameter tuning using techniques like Grid Search or Random Search to find the best set of parameters for the models.
- Tune parameters like `max_depth`, `n_estimators`, and `learning_rate` etc.…

8. Predictions and Final Model:

Objective:
Use the best-performing model to make predictions on unseen data.

Task:

- Apply the final model to predict IMDb ratings for TV shows in the test dataset.
- Analyze the predicted ratings and compare them with the actual ratings.
- Visualize predictions vs actual ratings using a scatter plot.

9. Model Interpretation and Conclusion:

Objective:
Interpret the results and conclude the project.

Task:

- Discuss the importance of different features in predicting IMDb ratings.
- Provide a final comparison of model performances and justify the choice of the best model.
- Highlight the potential improvements or additional features that could be explored to enhance prediction accuracy.

Expected Outcomes:

1. **Data Exploration and Preprocessing:**
    - Cleaned data ready for modeling with categorical features transformed and numerical features scaled.
2. **Model Development:**
    - Different Machine Learning Models built and evaluated.
3. **Model Evaluation:**
    - Identification of the best model based on evaluation metrics and performance.
4. **Prediction Insights:**
    - Predictions for IMDb ratings with detailed error analysis.