MINOR PROJECT

IMDb ANALYSIS

Project Overview:

This project aims to analyze an IMDb dataset to predict movie ratings using both regression and classification techniques. The analysis will involve data cleaning, feature engineering, model training, and evaluation. The primary goal is to understand the factors influencing movie ratings and to develop predictive models that can categorize movies based on their ratings.

Standards:

- **Data Quality: Ensure that the dataset is clean, with no missing values or outliers that could skew results.**
- **Model Evaluation: Use appropriate metrics (MSE for regression, accuracy, precision, recall, F1-score for classification) to evaluate model performance.**
- **Reproducibility: Document all steps and code to ensure that the analysis can be replicated.**
- **Visualization: Use clear and informative visualizations to present findings effectively.**

Objectives:

1. **Data Preparation: Clean and preprocess the dataset to make it suitable for analysis.**
2. **Feature Engineering: Identify and create relevant features that can improve model performance.**

3. **Model Development: Train a linear regression model for predicting movie ratings and a random forest classifier for categorizing movies based on ratings.**

4. **Model Evaluation: Assess the performance of both models using appropriate metrics.**

5. **Insights and Visualization: Provide insights into the factors affecting movie ratings and visualize the results for better understanding.**

6. **Data Preparation: Clean and preprocess the dataset to make it suitable for analysis.**

7. **Feature Engineering: Identify and create relevant features that can improve model performance.**

8. **Model Development: Train a linear regression model for predicting movie ratings and a random forest classifier for categorizing movies based on ratings.**

9. **Model Evaluation: Assess the performance of both models using appropriate metrics.**

10. **Insights and Visualization: Provide insights into the factors affecting movie ratings and visualize the results for better understanding.**

Requirements/Task(s):

- **Software: Python, with libraries such as Pandas, Scikit-learn, Matplotlib, and NumPy.**

- **Dataset: An IMDb dataset containing features like movie ratings, votes, year, and genre.**

- **Hardware: A computer capable of running Python and handling data processing tasks.**

Record notes/research here:

- **Data Cleaning: Document any decisions made regarding missing values and outlier handling.**

- **Feature Selection: Keep track of which features were selected and why.**

- **Model Performance: Record the performance metrics for both models, including any hyperparameter tuning results.**

- **Visualizations: Save all plots and visualizations for reference and reporting.**

Outline :

1. **Load the Dataset**: Import the necessary libraries and load the IMDb dataset.
2. **Data Overview**: Display the first few rows and basic statistics of the dataset.
3. **Data Cleaning**: Drop rows with missing values and reset the index.
4. **Encoding Categorical Data**: Convert categorical variables (e.g., Genre) into dummy variables.
5. **Feature Selection**: Identify relevant features for the regression and classification tasks.
6. **Train-Test Split**: Split the dataset into training and testing sets.
7. **Linear Regression Model**:
    - Train the model on the training set.
    - Make predictions and evaluate using MSE.
    - Visualize actual vs. predicted ratings.

8. **Classification Task**:
    - Create a binary target variable for high ratings.
    - Train a Random Forest Classifier.
    - Make predictions and evaluate using accuracy and other metrics.
9. **Feature Importance**: Analyze and visualize the importance of features in the classification model.
10. **Documentation**: Compile findings, visualizations, and insights into a report.

Summary:

This project provides a structured approach to analyzing movie ratings using machine learning techniques. By employing both regression and classification models, the analysis aims to uncover the key factors that influence movie ratings and to develop predictive models that can assist stakeholders in the film industry. The project emphasizes data quality, model evaluation, and effective communication of results through visualizations, ultimately contributing to a deeper understanding of the dynamics of movie ratings.

The code:

```
import kagglehub


# Download latest version

path = kagglehub.dataset_download("yusufdelikkaya/imdb-movie-dataset")


print("Path to dataset files:",path)

import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import mean_squared_error, accuracy_score

import matplotlib.pyplot as plt


# Load the dataset
```

```
file_path =
"C:/Users/sarus/.cache/kagglehub/datasets/yusufdelikkaya/imdb-movie-dataset/versions/1/
imdb_movie_dataset.csv"

df = pd.read_csv(file_path)


# Display basic information

print("Dataset Overview:")

print(df.head())


# Data Cleaning: Drop rows with missing values and reset index

df = df.dropna().reset_index(drop=True)


# Encoding categorical data (e.g., Genre)

df = pd.get_dummies(df, columns=["Genre"], drop_first=True)


# Feature Selection

features = ["Votes", "Year"] + [col for col in df.columns if col.startswith("Genre_")]

target = "Rating"


# Splitting data into training and testing sets

X = df[features]

y = df[target]

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


# Linear Regression for Rating Prediction

lr_model = LinearRegression()

lr_model.fit(X_train, y_train)
```

```
# Predictions and Evaluation
y_pred = lr_model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print(f"\nLinear Regression - Mean Squared Error: {mse:.2f}")

# Visualizing Predictions
plt.scatter(y_test, y_pred, color="blue", alpha=0.5)
plt.title("Actual vs Predicted Ratings")
plt.xlabel("Actual Ratings")
plt.ylabel("Predicted Ratings")
plt.show()

# Classification: Categorizing Movies Based on Rating
# Create a new column for classification (e.g., High Rating: 1, Low Rating: 0)
df["High_Rating"] = (df["Rating"] >= 7.0).astype(int)
X = df[features]
y = df["High_Rating"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Random Forest Classifier
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predictions and Accuracy
y_pred = rf_model.predict(X_test)
```

```
accuracy = accuracy_score(y_test, y_pred)

print(f"\nRandom Forest Classifier - Accuracy: {accuracy:.2f}")


# Feature Importance

feature_importance = pd.Series(rf_model.feature_importances_,
index=features).sort_values(ascending=False)

print("\nFeature Importance:")

print(feature_importance)


# Plot Feature Importance

feature_importance.head(10).plot(kind="bar", color="green", title="Top Features for High
Ratings")

plt.xlabel("Features")

plt.ylabel("Importance")

plt.show()
```

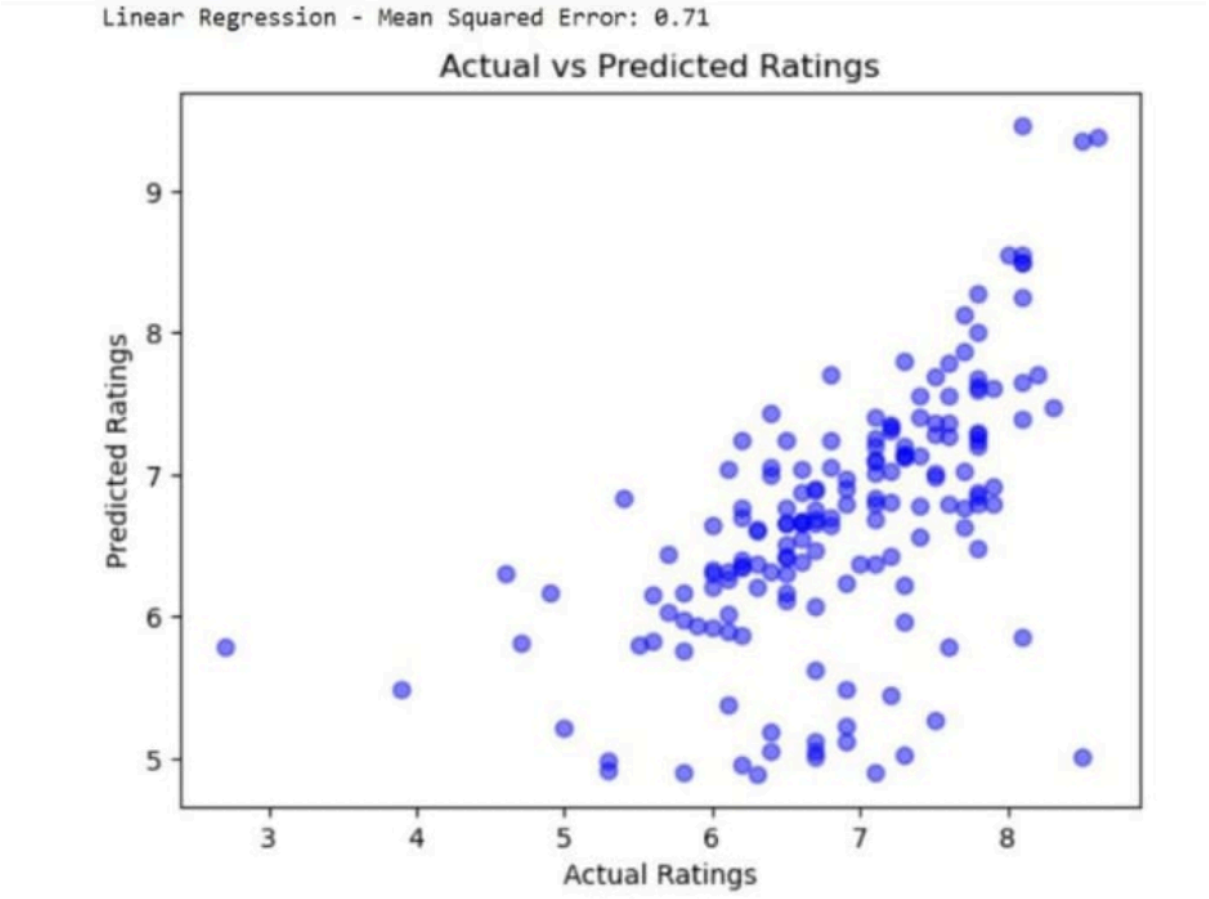
Output :

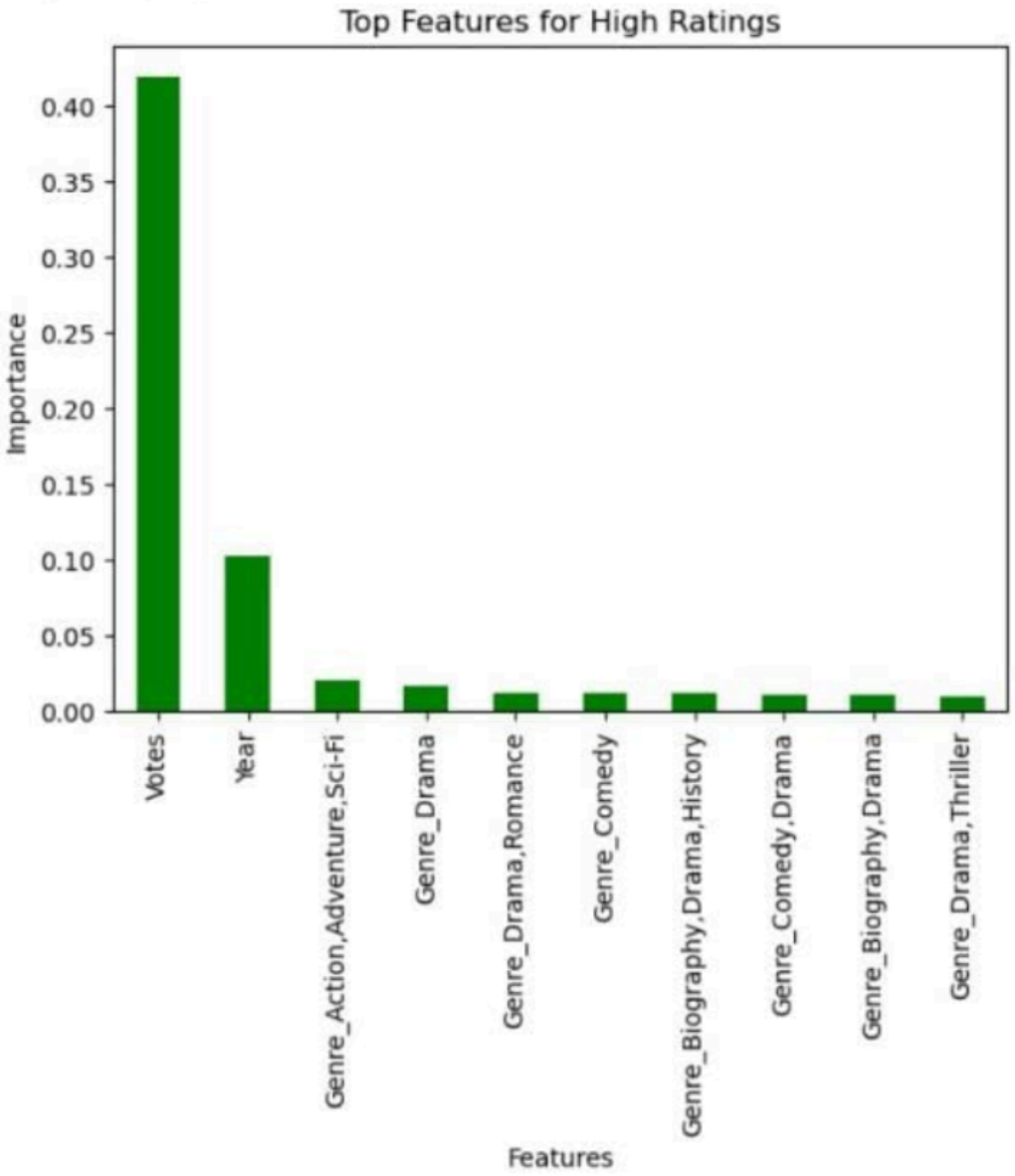
```
Path to dataset files: C:\Users\sarus\.cache\kagglehub\datasets\yusufdelikkaya\imdb-movie-dataset\versions\1
Dataset Overview:
   Rank                    Title                     Genre  \
0     1  Guardians of the Galaxy   Action,Adventure,Sci-Fi
1     2               Prometheus  Adventure,Mystery,Sci-Fi
2     3                    Split           Horror,Thriller
3     4                     Sing   Animation,Comedy,Family
4     5            Suicide Squad  Action,Adventure,Fantasy

                                     Description             Director  \
0  A group of intergalactic criminals are forced ...           James Gunn
1  Following clues to the origin of mankind, a te...          Ridley Scott
2  Three girls are kidnapped by a man with a diag...    M. Night Shyamalan
3  In a city of humanoid animals, a hustling thea...  Christophe Lourdelet
4  A secret government agency recruits some of th...            David Ayer
```

	Actors	Year	Runtime (Minutes)	\
0	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	
1	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	
2	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	
3	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	
4	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	

	Rating	Votes	Revenue (Millions)	Metascore
0	8.1	757074	333.13	76.0
1	7.0	485820	126.46	65.0
2	7.3	157606	138.12	62.0
3	7.2	60545	270.32	59.0
4	6.2	393727	325.02	40.0

Linear Regression - Mean Squared Error: 0.71




Top Features for High Ratings
Importance
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Votes
Year
Genre_Action,Adventure,Sci-Fi
Genre_Drama
Genre_Drama,Romance
Genre_Comedy
Genre_Biography,Drama,History
Genre_Comedy,Drama
Genre_Biography,Drama
Genre_Drama,Thriller
Features