EARIN Lab4
Task number: 3
Name: Sotaro Suzuki, Chihiro Tomatsu
Number: 317340, K-6929

Introduction:
The objective of this project is to predict the movie critic ratings based on the provided dataset (variant3.csv) using different regression and classification methods. The dataset includes information about the movie's marketing expense, production expense, multiplex coverage, budget, lead actor and actress ratings, director and producer ratings, trailer views, 3D availability, time taken, Twitter hashtags, genre, average age of actors, number of multiplexes, collection, and Start_Tech_Oscar. The Critic_rating variable is the target variable, which can be obtained as either continuous or discrete values.

Methods:
We began by loading and preprocessing the dataset. We split the dataset into training and testing sets, and standardized the features using the StandardScaler function from scikit-learn. Then, we applied different regression and classification methods to predict the Critic_rating variable.

For regression, we used Linear Regression, Support Vector Regression, Random Forest Regression, and XGBoost Regression. For classification, we used Logistic Regression, Support Vector Machine Classifier, Random Forest Classifier, and XGBoost Classifier. We evaluated the performance of each method by calculating the Mean Squared Error and R-squared for regression, and Accuracy and Classification Report for classification. We also compared the performance of different methods to determine the best method for predicting movie critic ratings.

Results:

```
Linear Regression - Mean Squared Error: 0.3601427755840237, R-squared:
0.12081887035337957
Support Vector Regression - Mean Squared Error: 0.39068661096629365, R-squared:
0.04625520972856878
Random Forest Regression - Mean Squared Error: 0.34413479607843117, R-squared:
0.1598975759646858
XGBoost Regression - Mean Squared Error: 0.3851982980522363, R-squared:
0.05965328814291504
The best regression method is: Random Forest Regression
Logistic Regression - Accuracy: 0.5196078431372549
```

```
The best regression method is: Random Forest Regression
Logistic Regression - Accuracy: 0.5196078431372549
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.26      0.39        27
           1       0.53      0.69      0.60        45
           2       0.44      0.50      0.47        30

    accuracy                           0.52       102
   macro avg       0.58      0.48      0.48       102
weighted avg       0.57      0.52      0.50       102

Support Vector Machine Classifier - Accuracy: 0.46078431372549017
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.04      0.07        27
           1       0.46      0.69      0.55        45
           2       0.45      0.50      0.48        30

    accuracy                           0.46       102
   macro avg       0.64      0.41      0.37       102
weighted avg       0.60      0.46      0.40       102
```

```
Random Forest Classifier - Accuracy: 0.5
Classification Report:
              precision    recall  f1-score   support

           0       0.50      0.19      0.27        27
           1       0.48      0.69      0.57        45
           2       0.54      0.50      0.52        30

    accuracy                           0.50       102
   macro avg       0.51      0.46      0.45       102
weighted avg       0.50      0.50      0.47       102

XGBoost Classifier - Accuracy: 0.5196078431372549
Classification Report:
              precision    recall  f1-score   support

           0       0.46      0.22      0.30        27
           1       0.49      0.67      0.57        45
           2       0.61      0.57      0.59        30

    accuracy                           0.52       102
   macro avg       0.52      0.49      0.48       102
weighted avg       0.52      0.52      0.50       102

The best classification method is: Logistic Regression
```

Analysis:

Among the regression methods, Random Forest Regression performed the best with a Mean Squared Error of 0.34 and R-squared of 0.16. This indicates that Random Forest Regression is able to explain 16% of the variance in the Critic_rating variable. Among the classification methods, Logistic Regression performed the best with an accuracy of 0.52. The classification report for Logistic Regression showed that it has higher precision and recall for class 1 (Medium) than the other classes, indicating that it is better at predicting movies with medium critic ratings.