Individual Final Report
Amna Gul

For the final project, I worked mainly on Data Cleaning and GUI. Major details of which are given below:

Data Cleaning:

The raw data that we obtained from Kaggle was in a very bad shape. Although there were 45,000+ rows in it and 24 columns but without any pre-processing it could not be used for exploratory data analysis or modelling so the following steps were performed:

1. Dropped irrelevant columns e.g. "home_page" and "poster_path"
2. Removed all rows with corrupt or invalid data e.g. removed alpha/character values from budget column that should only contain numerical values
3. To keep the level of outliers to a minimum, all unrealistic values/rows were removed e.g. really low values in budget/vote_count column
4. Based on ratio of values contained in two main columns, created a new target/label column for models
5. Our data set was 2 years old so to get the most recent values merged (based on imdb_id) vote_average and vote_count columns from publicly available data on IMDbs website (updated daily).
6. Also since IMDb is the most reliable and authentic source of movies so I replaced/merged Director and startYear column from IMDb into Kaggle's data set.
7. All duplicate values rows were removed.

The above-mentioned steps left us with only 2,222 rows and 17 columns.

Graphical User Interface:

This was the first time ever that I worked with GUI code. I believe I could have done a better job with it had I started working on it but within a limited time frame I was able to:

1. Created the main window with the following drop down menus
    a. File
        i. Contained "Exit Application" button (works when clicked)
    b. EDA Analysis
        i. Contained "Initial Assessment" & "Correlation plot" buttons (both worked upon clicking)
            1. "Initial Assessment" prints bar plot for target variable (output shown at end of this document)

2. "Correlation plot" button prints correlation matrix between numerical features in our data set. (output shown at end of this document). It also provided the option to check/uncheck any features and on clicking "update" would produce the appropriate plot accordingly.

c. ML Models

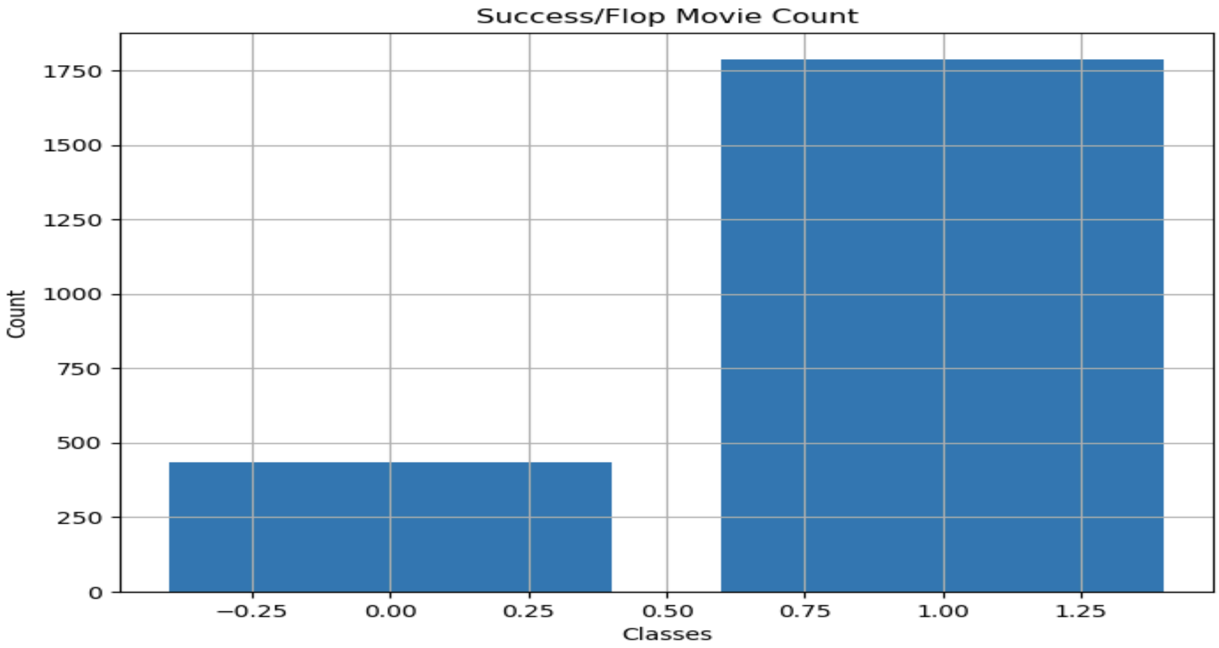
i. Contained "Decision Tree Entropy" & "Random Forest Classifier" buttons (did NOT work upon clicking)

1. "Decision Tree Entropy" GUI button produced super-optimistic results (accuracy of 80%) whereas original accuracy of our DT model was around 69% (output shown at end of this document)

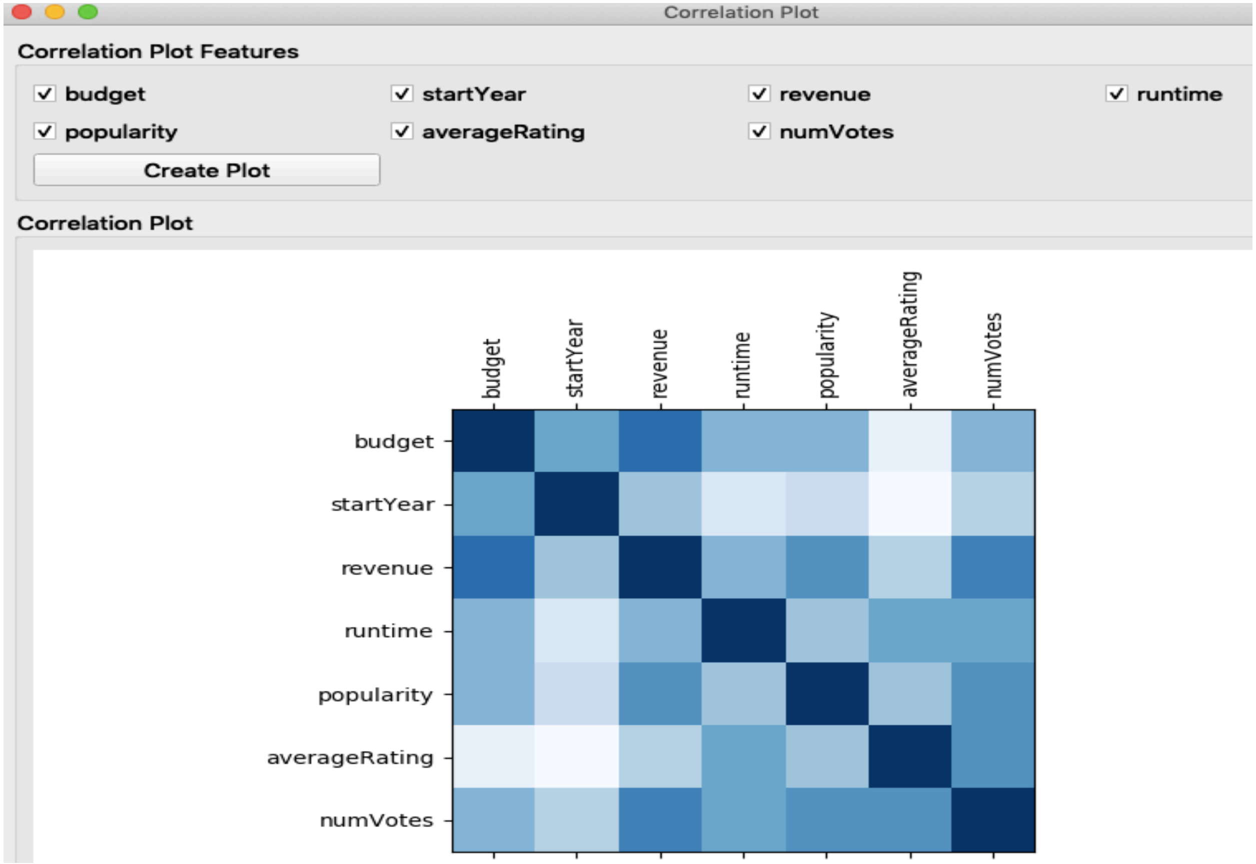
Code Percentage Copied:

Major portion of GUI code was copied from https://github.com/amir-jafari/Data-Mining/tree/master/Demo/PyQt5/Demo and then modified.

"Initial Assessment" button output:



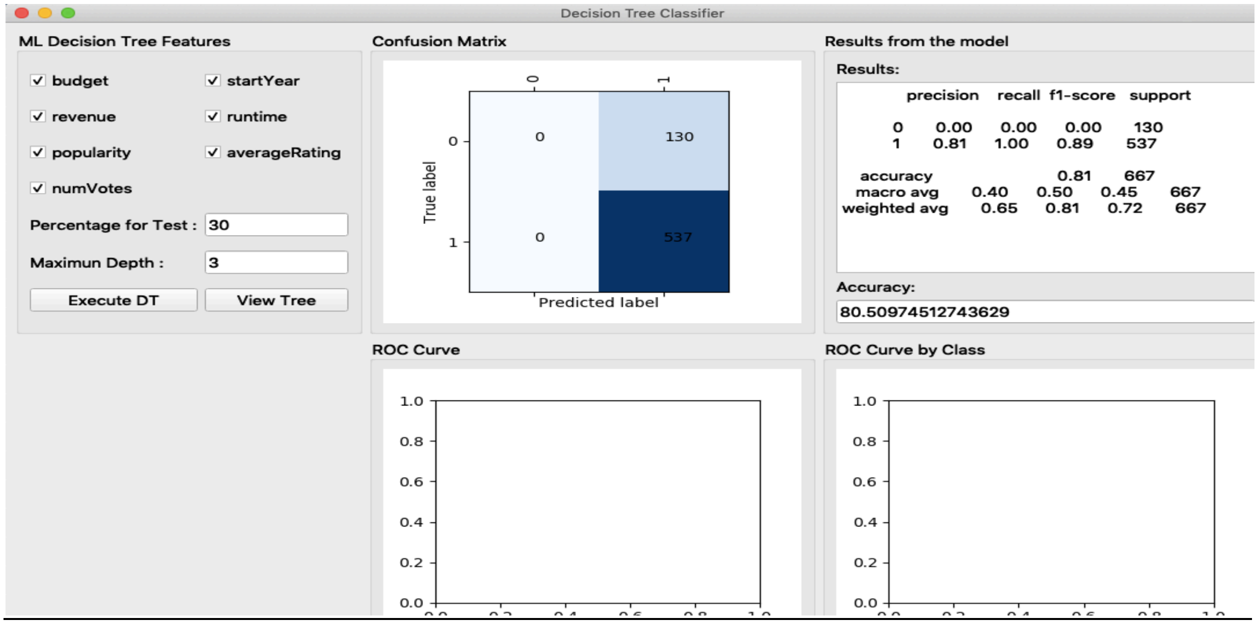
“Correlation plot” button



☑ budget ☑ startYear ☑ revenue ☑ runtime

☑ popularity ☑ averageRating ☑ numVotes

“Decision Tree Entropy” button output:



ML Decision Tree Features

☑ budget ☑ startYear

☑ revenue ☑ runtime

☑ popularity ☑ averageRating

☑ numVotes

Percentage for Test : 30

Maximun Depth : 3

Results from the model

Results:

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       130
           1       0.81      1.00      0.89       537

    accuracy                           0.81       667
   macro avg       0.40      0.50      0.45       667
weighted avg       0.65      0.81      0.72       667
```

Accuracy:

80.50974512743629