Predicting Movie Success

Team Names :

1. Habiba Khaled Mohammed Mahmoud (3A)
2. Amina Ahmed Mohammed Abounawara (2A)
3. Dina Abdallah Said Ahmed (3A)
4. Ahmed Mostafa Shehata Nobi (1A)
5. Hadeer Mahdy Zein Elabdeen (3B)

Team ID: 11

Machine Learning Project (MileStone1 & 2)

Dr: Dina Khattab

Regression

Preprocessing

Feature Extraction:

- **Techniques:**
  - Drop column [Type] from movies success dataset.
  - Replace Null value with **Expect Mean** for cols in case integer and **first frequency** of cols in case object (list of string).
  - cols: columns ['Directors', 'Genres', 'Country', 'Language', 'Runtime', 'Year', 'Age', 'Rotten Tomatoes','IMDb’, ‘rate”].
  - Apply OneHotEncoding on ['Language', 'Year', 'Directors', 'Genres', 'Country','Age’] using LabelEncoder.
  - Normalize numerical columns that aren’t binary valued ['Directors', 'Genres', 'Country', 'Language', 'Runtime','Year','IMDb’].

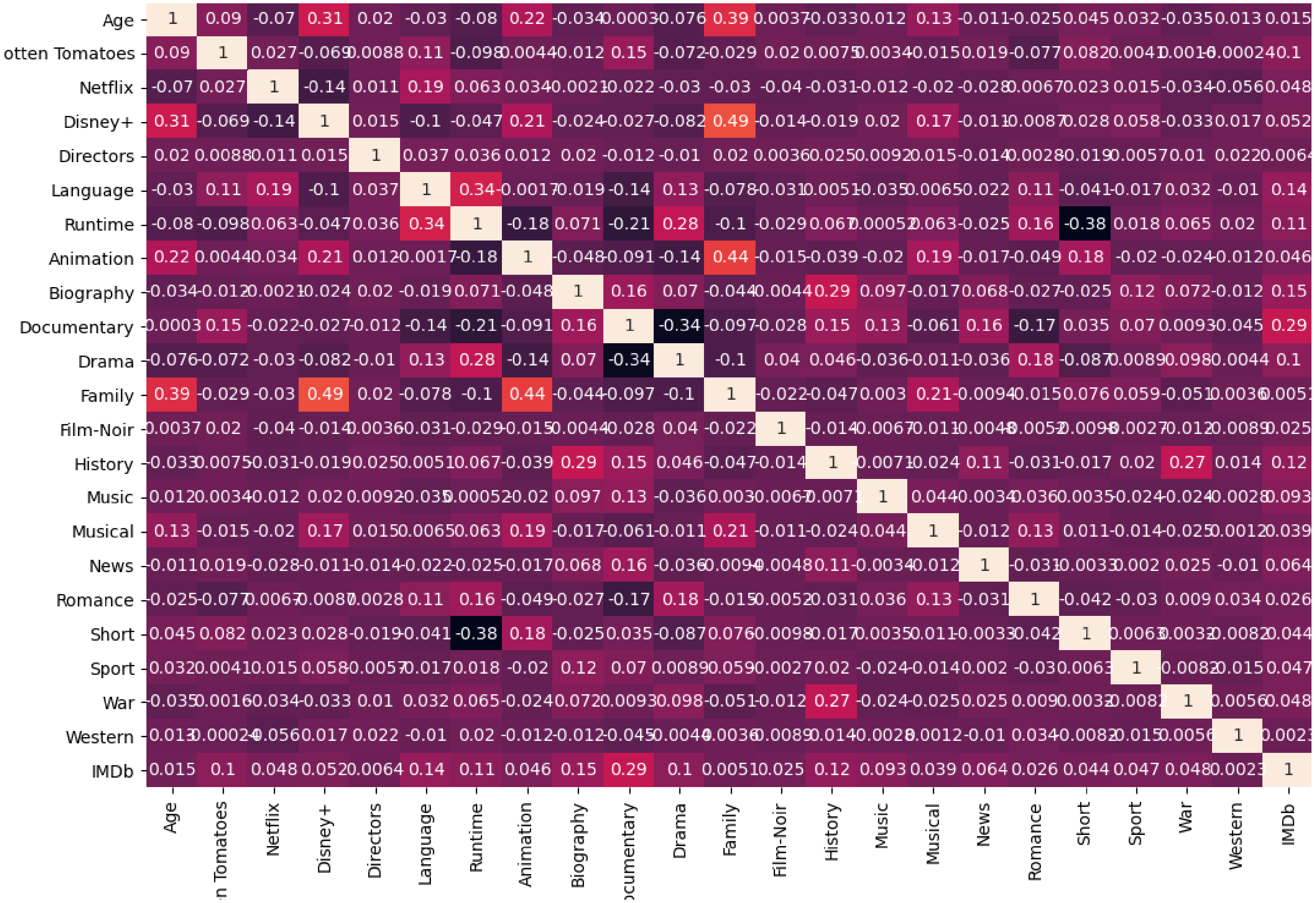
- **Implementation**:
  - Normalization for columns ['Directors', 'Genres', 'Country', 'Language', 'Runtime','Year' , ‘IMDb’] : Using X_normalize=(X - X_min) / (X_Max - X_Min)
  - Replace Null value in data
  - Case type of column object
    - **-** Freq_values = str(col.value_counts()
      [col.value_counts()== col.value_counts().max()])
    - **-** Col= col.replace(np.nan,Freq_values .split(' ')[0])

Else

- Get mean for specific cols moviesFile[idx].mean()
- Col = col.replace(np.nan, Mean_col_value)

- OneHotEncoding function for columns ['Country', 'Language','Year', 'Directors', 'Age']
  - Using Label Encoding to get index of each class
    Label_en = LabelEncoder()
    Id_values = Label_en .fit_transform(col)
  - Using One Hot Encoding to binary class for each rows in cols
    Encode = OneHotEncoder()
    Binary_classes= Encode.fit_transform(Id_values)
- MultilableBinarizer function for column ['Genres']
- Drop col using moviesFile.drop(['Type'], axis=1, inplace=True)

Dataset Analysis: due to correlation results as shown in the below figure.



	Age	n Tomatoes	Netflix	Disney+	Directors	Language	Runtime	Animation	Biography	ocumentary	Drama	Family	Film-Noir	History	Music	Musical	News	Romance	Short	Sport	War	Western	IMDb
Age	1	0.09	-0.07	0.31	0.02	-0.03	-0.08	0.22	-0.034	0.0003	-0.076	0.39	0.0037	-0.033	0.012	0.13	-0.011	-0.025	0.045	0.032	-0.035	0.013	0.015
otten Tomatoes	0.09	1	0.027	-0.069	0.0088	0.11	-0.098	0.0044	-0.012	0.15	-0.072	-0.029	0.02	0.0075	0.0034	-0.015	0.019	-0.077	0.082	0.0041	0.0016	-0.00024	0.1
Netflix	-0.07	0.027	1	-0.14	0.011	0.19	0.063	0.034	0.0021	-0.022	-0.03	-0.03	-0.04	-0.031	-0.012	-0.02	-0.028	0.0067	0.023	0.015	-0.034	-0.056	0.048
Disney+	0.31	-0.069	-0.14	1	0.015	-0.1	-0.047	0.21	-0.024	-0.027	-0.082	0.49	-0.014	-0.019	0.02	0.17	-0.011	-0.0087	0.028	0.058	-0.033	0.017	0.052
Directors	0.02	0.0088	0.011	0.015	1	0.037	0.036	0.012	0.02	-0.012	-0.01	0.02	0.0036	0.025	0.0092	0.015	-0.014	0.0028	-0.019	-0.0057	0.01	0.022	0.0064
Language	-0.03	0.11	0.19	-0.1	0.037	1	0.34	-0.0017	-0.019	-0.14	0.13	-0.078	-0.031	0.0051	-0.035	0.0065	-0.022	0.11	-0.041	-0.017	0.032	-0.01	0.14
Runtime	-0.08	-0.098	0.063	-0.047	0.036	0.34	1	-0.18	0.071	-0.21	0.28	-0.1	-0.029	0.067	0.00052	0.063	-0.025	0.16	-0.38	0.018	0.065	0.02	0.11
Animation	0.22	0.0044	0.034	0.21	0.012	-0.0017	-0.18	1	-0.048	-0.091	-0.14	0.44	-0.015	-0.039	-0.02	0.19	-0.017	-0.049	0.18	-0.02	-0.024	-0.012	0.046
Biography	-0.034	-0.012	0.0021	-0.024	0.02	-0.019	0.071	-0.048	1	0.16	0.07	-0.044	0.0044	0.29	0.097	-0.017	0.068	-0.027	-0.025	0.12	0.072	-0.012	0.15
Documentary	0.0003	0.15	-0.022	-0.027	-0.012	-0.14	-0.21	-0.091	0.16	1	-0.34	-0.097	-0.028	0.15	0.13	-0.061	0.16	-0.17	0.035	0.07	0.0093	-0.045	0.29
Drama	-0.076	-0.072	-0.03	-0.082	-0.01	0.13	0.28	-0.14	0.07	-0.34	1	-0.1	0.04	0.046	-0.036	-0.011	-0.036	0.18	-0.087	0.0089	0.098	0.0044	0.1
Family	0.39	-0.029	-0.03	0.49	0.02	-0.078	-0.1	0.44	-0.044	-0.097	-0.1	1	-0.022	-0.047	0.003	0.21	-0.0094	-0.015	0.076	0.059	-0.051	0.0036	0.0051
Film-Noir	0.0037	0.02	-0.04	-0.014	0.0036	-0.031	-0.029	-0.015	0.0044	-0.028	0.04	-0.022	1	-0.014	-0.0067	-0.011	-0.0048	-0.0052	-0.0098	-0.0027	-0.012	-0.0089	0.025
History	-0.033	0.0075	-0.031	-0.019	0.025	0.0051	0.067	-0.039	0.29	0.15	0.046	-0.047	-0.014	1	-0.0071	-0.024	0.11	-0.031	-0.017	0.02	0.27	0.014	0.12
Music	0.012	0.0034	-0.012	0.02	0.0092	-0.035	0.00052	-0.02	0.097	0.13	-0.036	0.003	-0.0067	-0.0071	1	0.044	-0.0034	0.036	0.0035	-0.024	-0.024	0.0028	0.093
Musical	0.13	-0.015	-0.02	0.17	0.015	0.0065	0.063	0.19	-0.017	-0.061	-0.011	0.21	-0.011	-0.024	0.044	1	-0.012	0.13	0.011	-0.014	-0.025	0.0012	0.039
News	-0.011	0.019	-0.028	-0.011	-0.014	-0.022	-0.025	-0.017	0.068	0.16	-0.036	-0.0094	-0.0048	0.11	-0.0034	-0.012	1	-0.031	0.0033	0.002	0.025	-0.01	0.064
Romance	-0.025	-0.077	0.0067	-0.0087	0.0028	0.11	0.16	-0.049	-0.027	-0.17	0.18	-0.015	-0.0052	-0.031	0.036	0.13	-0.031	1	-0.042	-0.03	0.009	0.034	0.026
Short	0.045	0.082	0.023	0.028	-0.019	-0.041	-0.38	0.18	-0.025	0.035	-0.087	0.076	-0.0098	-0.017	0.0035	0.011	-0.0033	-0.042	1	0.0063	0.0032	0.0082	0.044
Sport	0.032	0.0041	0.015	0.058	-0.0057	-0.017	0.018	-0.02	0.12	0.07	0.0089	0.059	0.0027	0.02	-0.024	-0.014	0.002	-0.03	0.0063	1	-0.0082	0.015	0.047
War	-0.035	0.0016	-0.034	-0.033	0.01	0.032	0.065	-0.024	0.072	0.0093	0.098	-0.051	-0.012	0.27	-0.024	-0.025	0.025	0.009	0.0032	-0.0082	1	0.0056	0.048
Western	0.013	0.00024	-0.056	0.017	0.022	-0.01	0.02	-0.012	-0.012	-0.045	0.0044	0.0036	0.0089	0.014	0.0028	0.0012	-0.01	0.034	0.0082	0.015	0.0056	1	0.0023
IMDb	0.015	0.1	0.048	0.052	0.0064	0.14	0.11	0.046	0.15	0.29	0.1	0.0051	0.025	0.12	0.093	0.039	0.064	0.026	0.044	0.047	0.048	0.0023	1

- ➢ If we look at the figure we would find that the highest correlation to IMDb are Documentary, biography and Language but still a weak correlation. so, the taking of the rest features would enhance the model in our case.

- ❖ **Features we used for our models:** the whole dataset in addition to one-hot encoding the Genres column.

Modelling:

- ➢ **Regression Techniques we used:**
  - ■ Linear Regression:
  - To model the relationship between variables((Language, Runtime, Netflix, Disney+, Genres, Directors) by fitting a linear equation to observed data
  - ■ Polynomial Regression(Degree = 2):

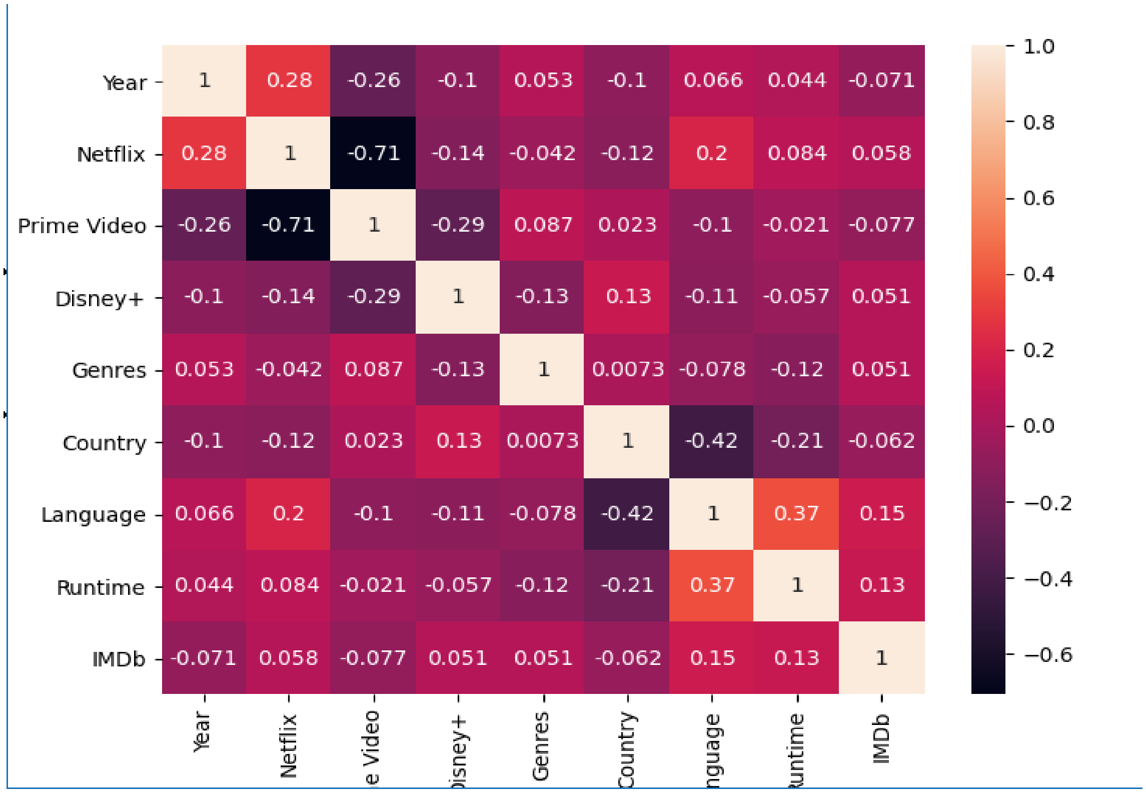
- provides the best approximation of the relationship between the dependent and independent variables in 3 degrees.

- Random forest Regression:
  - technique capable of performing both regression and classification tasks with the use of multiple decision trees.
- Decision tree Regression:
- builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- Ridge Regression:
- To mitigate the problem of multicollinearity in linear regression.
- RBF SVR Regression:
- To find a line that best fits the data using a gaussian kernel.

Regression technique	**MSE**	**Training time**	**Test time**
Random Forest Regression	0.01209	2.1232	0.0618
Decision tree regressor	0.01394	0.01	0.0
Polynomial regression	0.01319	0.1047	0.002
Multi-variable Linear regression	0.01327	0.004	0.001
Ridge Regression	0.01329	0.003	0.0
RBF SVR Regression	0.01200	1.9897	0.2414

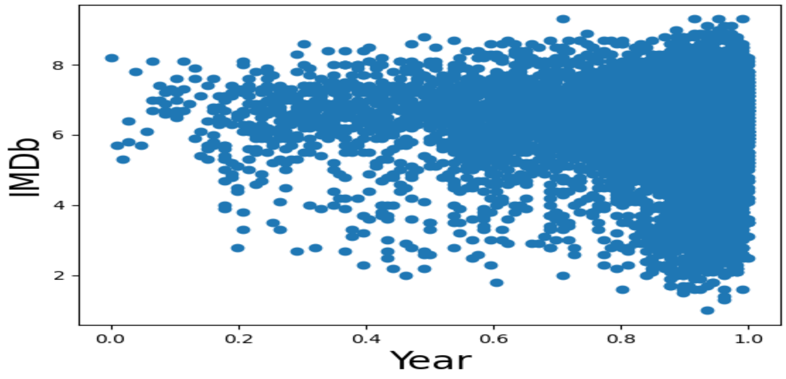
- **Sizes of training, testing sets**: our Training set size is 70% and the testing set is 30%.
- **Number of cross-validation k folds**: 10
- **Techniques used to improve the results**: we used cross-validation on our training set for:
  - Feature selection: features with correlation > 0.05 gave the best average MSE in cross-validation.
  - Parameter tuning: polynomial regression with degree 3 or 4 gave the best average MSE in cross-validation, decision tree regressor with max depth = 3 or higher gave better MSE results.
  - Model Selection: cross-validation scores showed that **Random forest** and **polynomial regression** are the best regression techniques for this dataset.

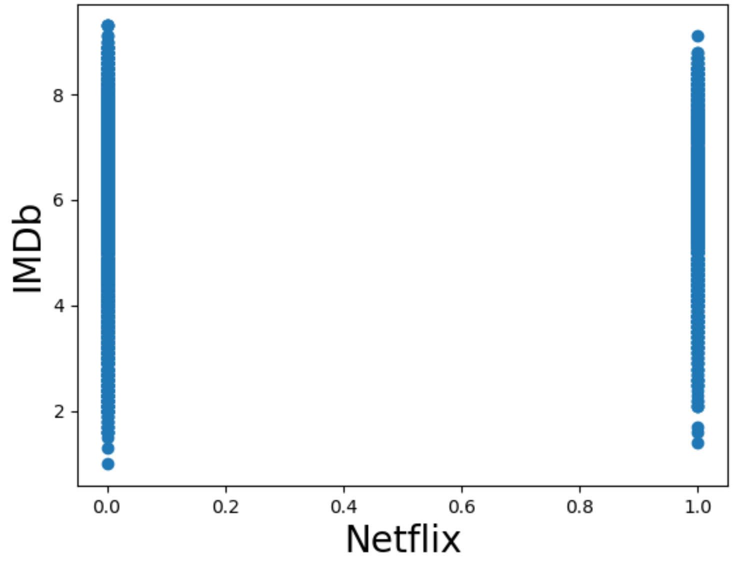
Data visualization:

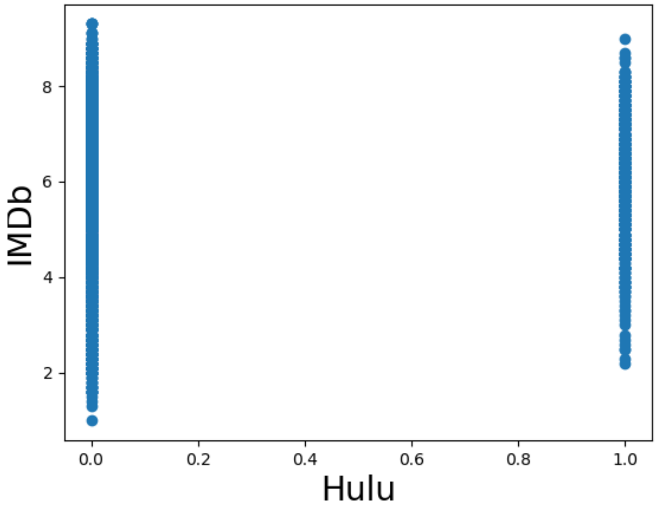


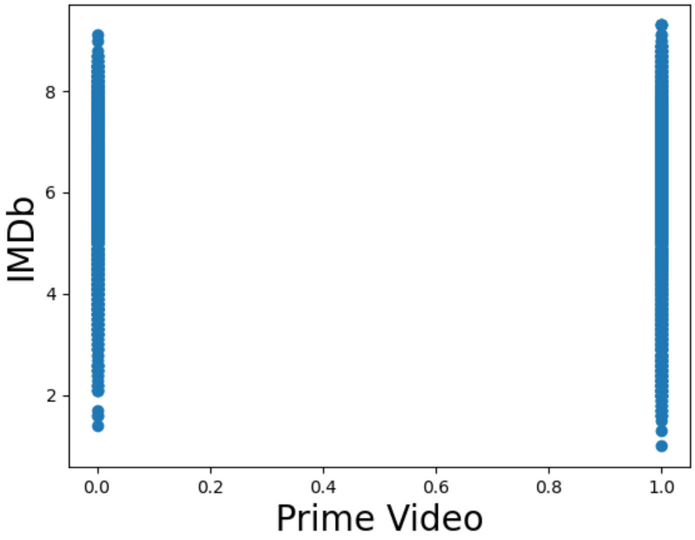
	Year	Netflix	e Video	isney+	Genres	Country	nguage	untime	IMDb
Year	1	0.28	-0.26	-0.1	0.053	-0.1	0.066	0.044	-0.071
Netflix	0.28	1	-0.71	-0.14	-0.042	-0.12	0.2	0.084	0.058
Prime Video	-0.26	-0.71	1	-0.29	0.087	0.023	-0.1	-0.021	-0.077
Disney+	-0.1	-0.14	-0.29	1	-0.13	0.13	-0.11	-0.057	0.051
Genres	0.053	-0.042	0.087	-0.13	1	0.0073	-0.078	-0.12	0.051
Country	-0.1	-0.12	0.023	0.13	0.0073	1	-0.42	-0.21	-0.062
Language	0.066	0.2	-0.1	-0.11	-0.078	-0.42	1	0.37	0.15
Runtime	0.044	0.084	-0.021	-0.057	-0.12	-0.21	0.37	1	0.13
IMDb	-0.071	0.058	-0.077	0.051	0.051	-0.062	0.15	0.13	1

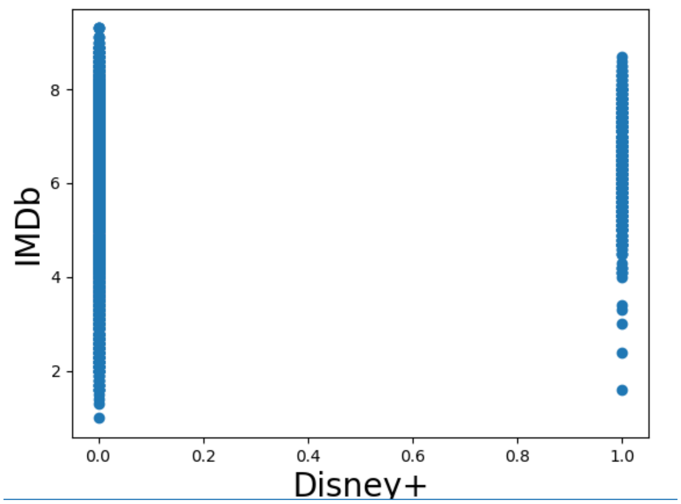
Top features correlation

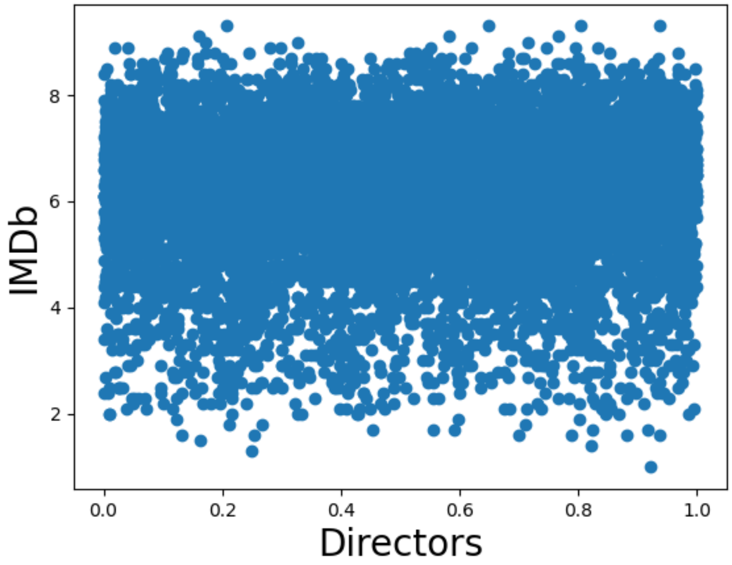


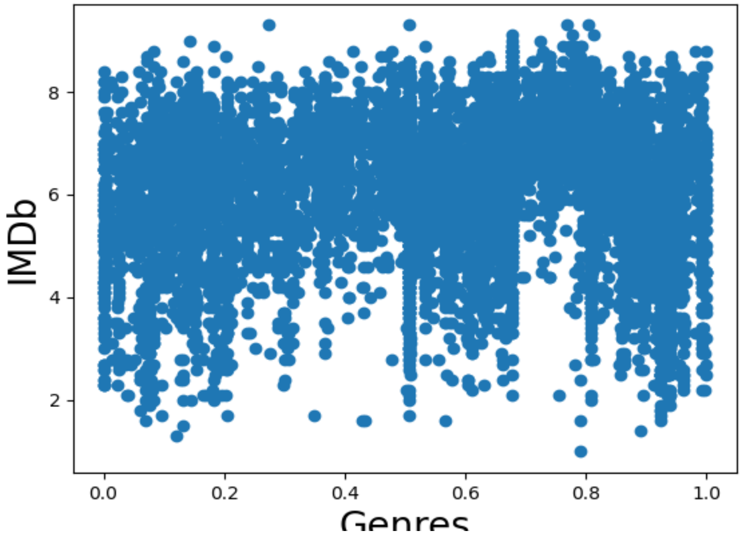

IMDb
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Netflix

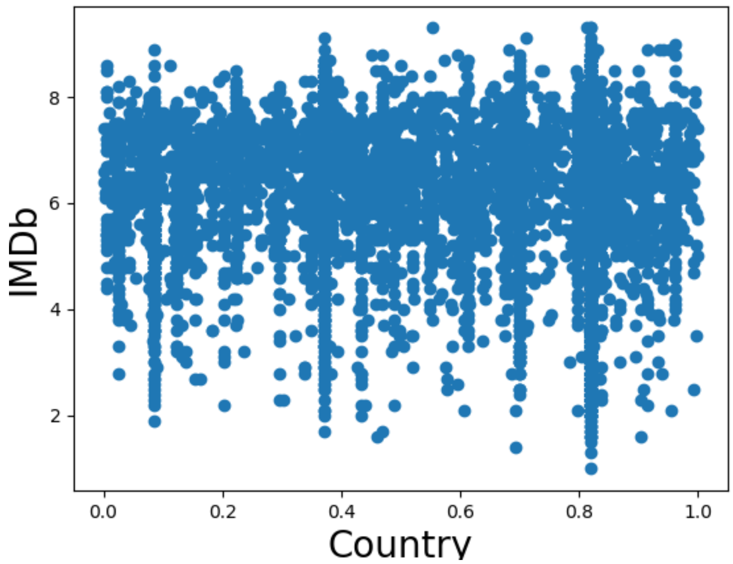

IMDb
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Hulu

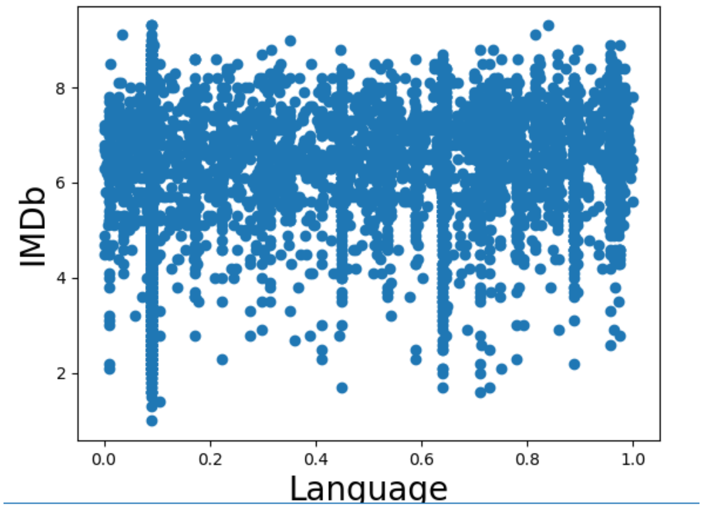

IMDb
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Prime Video

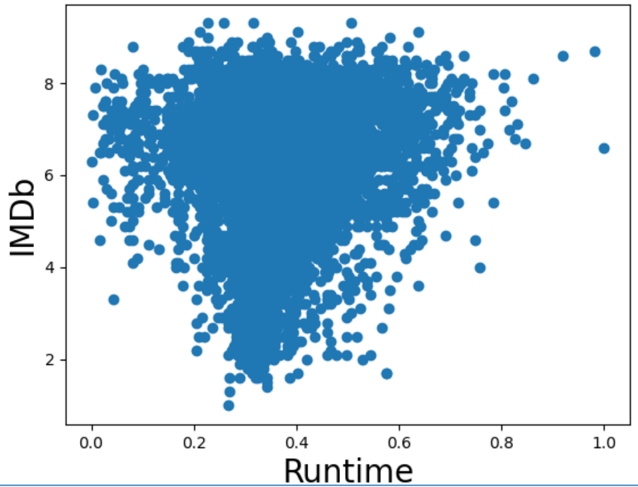

IMDb
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Disney+


IMDb
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Directors


IMDb
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Genres


IMDb
8
6
4
2
0.0
0.2
0.4
0.6
0.8
1.0
Country


IMDb
8
6
4
2
0.0
0.2
0.4
0.6
0.8
1.0
Language


IMDb
8
6
4
2
0.0
0.2
0.4
0.6
0.8
1.0
Runtime

Conclusion:

- Using k-fold cross validation helped decide the best hyperparameters, features and models for this dataset.
- After processing data and implementing 6 Regression techniques, The best techniques that achieved minimization error are random forest regression and polynomial regression.

Classification

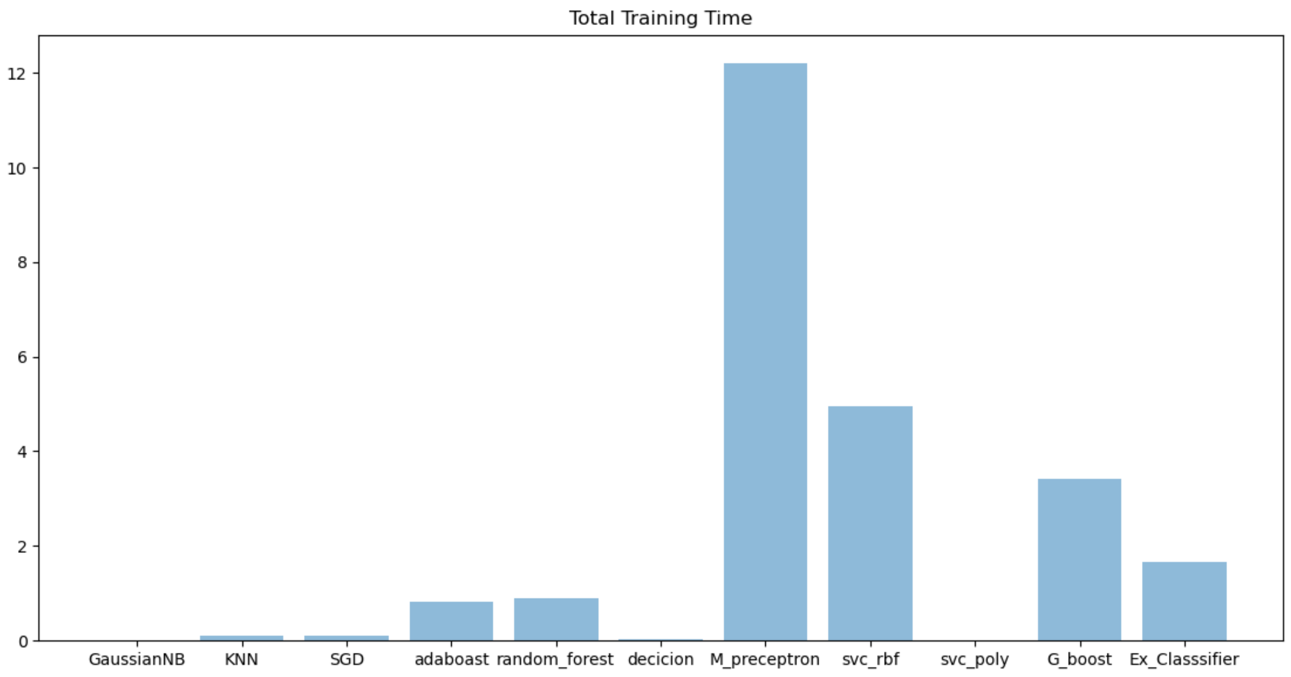
Classification Models

Models:

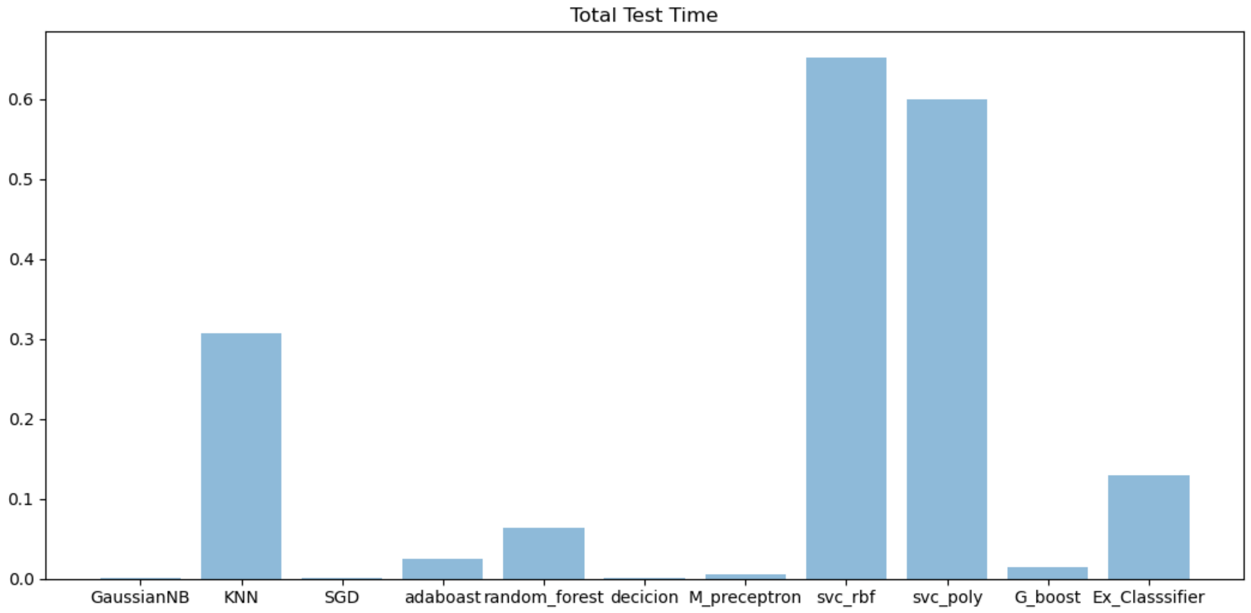
- The support vector Classification in scikit-learn.SVM with kernels {‘Ploy’, ‘linear’, ‘Rbf’ }.
- DecisionTreeClassifier.
- Linear classifiers with stochastic gradient descent (SGD) learning.
- Classifier implemented using the k-nearest neighbours.
- RandomForestClassifier.
- An AdaBoost-SAMME classifier fit on classification (DecisionTreeClassifier)
- ExtraTreesClassifier.
- Gradient Boosting for classification.
- Gaussian Naive Bayes (GaussianNB).
- Multi-layer Perceptron classifier.
- Logistic regression.

Summary:

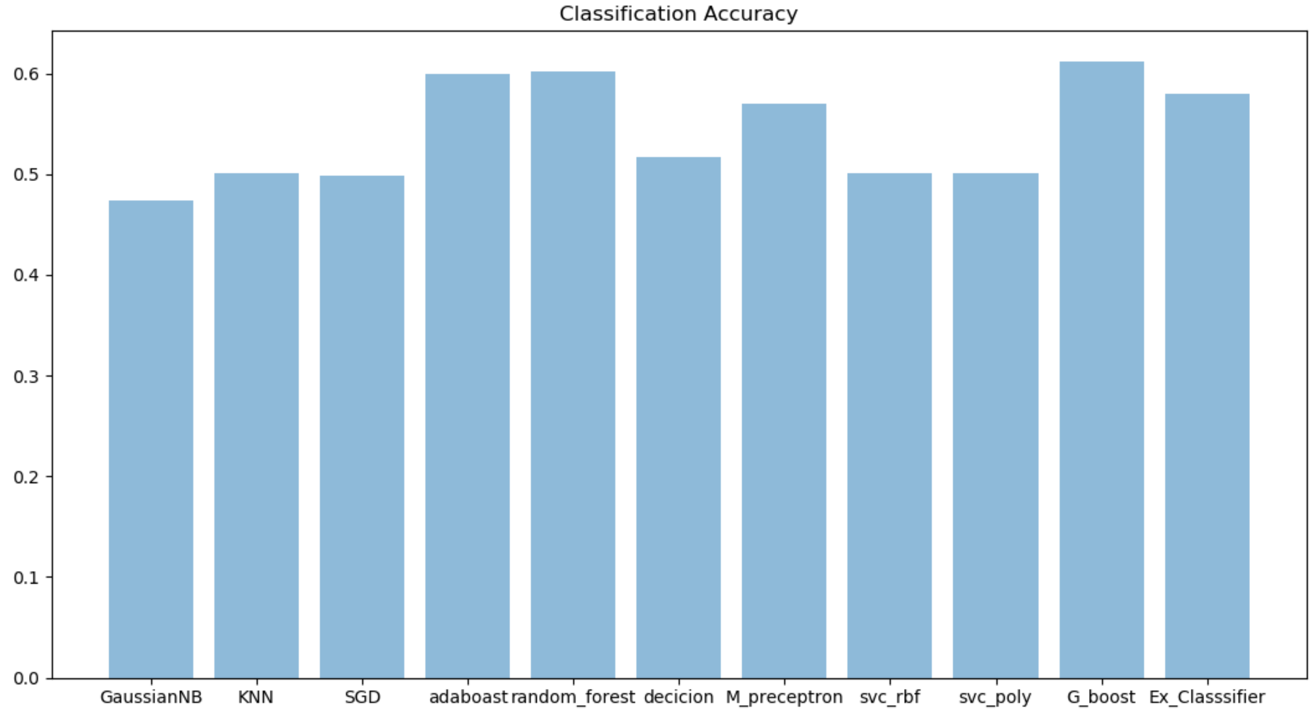
We trained the same preprocessed dataset of regression (Using Multilable Binarizer & Label Encoder) and gives us the best accuracy.
After training all models on the movies success dataset, first-time Gradient Boosting achieved high accuracy with 63% then randomForest achieved an accuracy of 61.1%.



BarGraph for training time

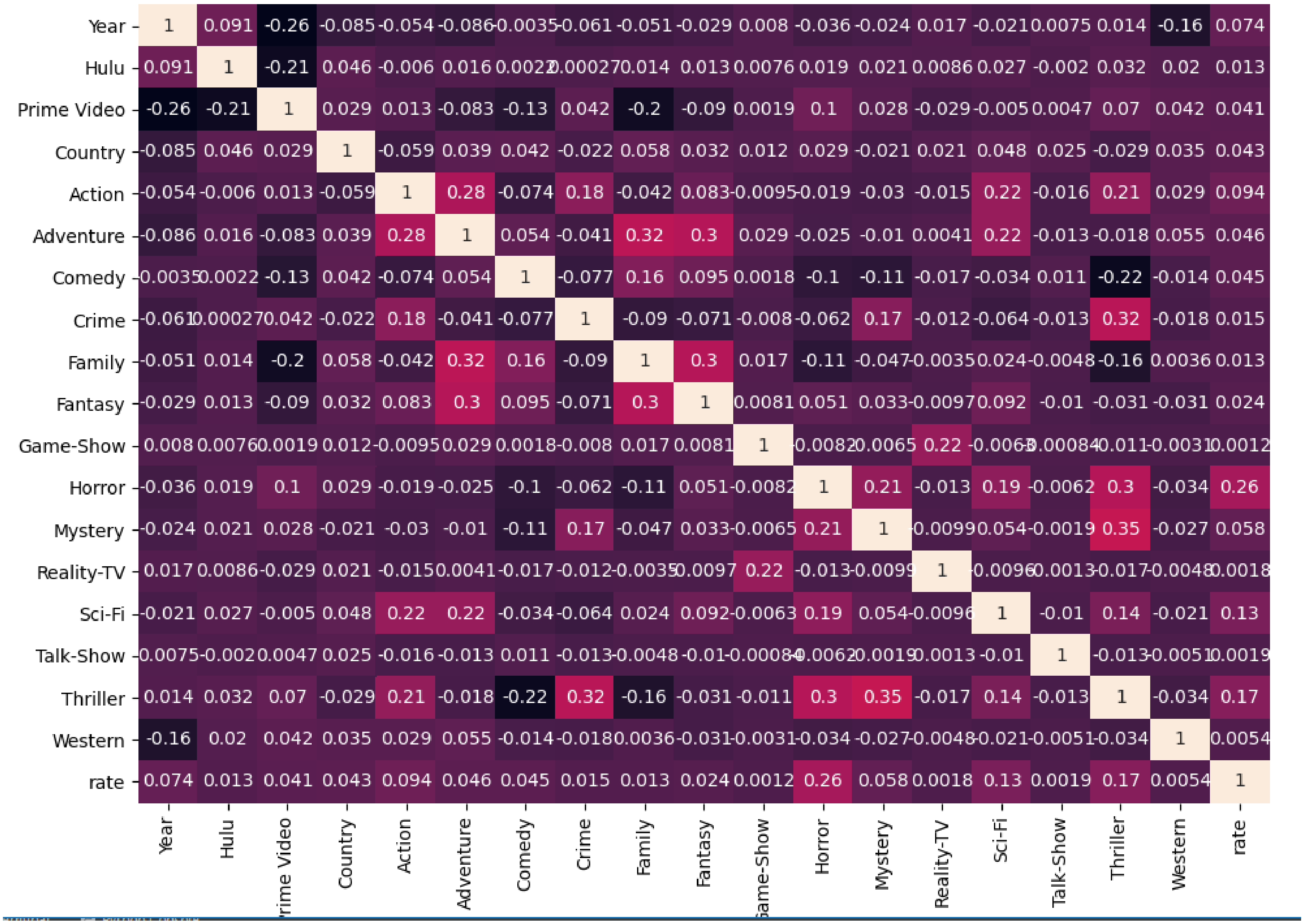


BarGraph for test time



Bar graph for accuracy

Data Analysis:



	Year	Hulu	rime Video	Country	Action	Adventure	Comedy	Crime	Family	Fantasy	ame-Show	Horror	Mystery	Reality-TV	Sci-Fi	Talk-Show	Thriller	Western	rate
Year	1	0.091	-0.26	-0.085	-0.054	-0.086	-0.0035	-0.061	-0.051	-0.029	0.008	-0.036	-0.024	0.017	-0.021	0.0075	0.014	-0.16	0.074
Hulu	0.091	1	-0.21	0.046	-0.006	0.016	0.0022	0.00027	0.014	0.013	0.0076	0.019	0.021	0.0086	0.027	-0.002	0.032	0.02	0.013
Prime Video	-0.26	-0.21	1	0.029	0.013	-0.083	-0.13	0.042	-0.2	-0.09	0.0019	0.1	0.028	-0.029	-0.005	0.0047	0.07	0.042	0.041
Country	-0.085	0.046	0.029	1	-0.059	0.039	0.042	-0.022	0.058	0.032	0.012	0.029	-0.021	0.021	0.048	0.025	-0.029	0.035	0.043
Action	-0.054	-0.006	0.013	-0.059	1	0.28	-0.074	0.18	-0.042	0.083	-0.0095	-0.019	-0.03	-0.015	0.22	-0.016	0.21	0.029	0.094
Adventure	-0.086	0.016	-0.083	0.039	0.28	1	0.054	-0.041	0.32	0.3	0.029	-0.025	-0.01	0.0041	0.22	-0.013	-0.018	0.055	0.046
Comedy	-0.0035	0.0022	-0.13	0.042	-0.074	0.054	1	-0.077	0.16	0.095	0.0018	-0.1	-0.11	-0.017	-0.034	0.011	-0.22	-0.014	0.045
Crime	-0.061	0.00027	0.042	-0.022	0.18	-0.041	-0.077	1	-0.09	-0.071	-0.008	-0.062	0.17	-0.012	-0.064	-0.013	0.32	-0.018	0.015
Family	-0.051	0.014	-0.2	0.058	-0.042	0.32	0.16	-0.09	1	0.3	0.017	-0.11	-0.047	-0.0035	0.024	-0.0048	-0.16	0.0036	0.013
Fantasy	-0.029	0.013	-0.09	0.032	0.083	0.3	0.095	-0.071	0.3	1	0.0081	0.051	0.033	-0.0097	0.092	-0.01	-0.031	-0.031	0.024
Game-Show	0.008	0.0076	0.0019	0.012	-0.0095	0.029	0.0018	-0.008	0.017	0.0081	1	-0.0082	-0.0065	0.22	-0.0063	-0.00084	-0.011	-0.0031	0.0012
Horror	-0.036	0.019	0.1	0.029	-0.019	-0.025	-0.1	-0.062	-0.11	0.051	-0.0082	1	0.21	-0.013	0.19	-0.0062	0.3	-0.034	0.26
Mystery	-0.024	0.021	0.028	-0.021	-0.03	-0.01	-0.11	0.17	-0.047	0.033	-0.0065	0.21	1	-0.0099	0.054	-0.0019	0.35	-0.027	0.058
Reality-TV	0.017	0.0086	-0.029	0.021	-0.015	0.0041	-0.017	-0.012	-0.0035	-0.0097	0.22	-0.013	-0.0099	1	-0.0096	0.0013	-0.017	-0.0048	0.0018
Sci-Fi	-0.021	0.027	-0.005	0.048	0.22	0.22	-0.034	-0.064	0.024	0.092	-0.0063	0.19	0.054	-0.0096	1	-0.01	0.14	-0.021	0.13
Talk-Show	0.0075	-0.002	0.0047	0.025	-0.016	-0.013	0.011	-0.013	-0.0048	-0.01	-0.00084	-0.0062	-0.0019	0.0013	-0.01	1	-0.013	-0.0051	0.0019
Thriller	0.014	0.032	0.07	-0.029	0.21	-0.018	-0.22	0.32	-0.16	-0.031	-0.011	0.3	0.35	-0.017	0.14	-0.013	1	-0.034	0.17
Western	-0.16	0.02	0.042	0.035	0.029	0.055	-0.014	-0.018	0.0036	-0.031	-0.0031	-0.034	-0.027	-0.0048	-0.021	-0.0051	-0.034	1	0.0054
rate	0.074	0.013	0.041	0.043	0.094	0.046	0.045	0.015	0.013	0.024	0.0012	0.26	0.058	0.0018	0.13	0.0019	0.17	0.0054	1

The same feature selection methods used in regression are used in classification.

using correlation type statistical measures between variables as the basis for filter feature selection.

The choice of statistical measures is highly dependent upon the variable data types.
The relation between variables approved after replacing null values with expected means of the column.

HyperParameter tuning:

Hyperparameter tuning is performed to choose a set of optimal hyperparameters for a learning algorithm.
Hyper parameters in Random Forest Classifier using cross validation:

- max_depth = [3, 10, 20, 40]
- n_estimators = [10, 50, 100, 200]

In RandomForestClassifier when n_estimators =10, max_depth=3 accuracy=48%.
In RandomForestClassifier when n_estimators =100, max_depth=3 accuracy=51%.
In RandomForestClassifier when n_estimators =10, max_depth=40 accuracy=56%.
In RandomForestClassifier when n_estimators =100, max_depth=40 accuracy=58%.

When we increase n_estimators and max_depth the classifier achieves higher accuracy till a specific value then it degrades.

Conclusion:

Regression is the task of predicting a continuous quantity while Classification is the task of predicting a discrete class label. Classification predictions can be evaluated using accuracy, whereas regression predictions can be evaluated using mean squared error.
Feature engineering gave a better version of the dataset by extracting the important parts and discarding redundant ones.
Using correlation highlighted the highest correlated features which gave higher accuracy and decreased training time.
For classification, the **Gradient boost** classifier gave the highest accuracy while the **SGD** classifier gave the least accuracy.