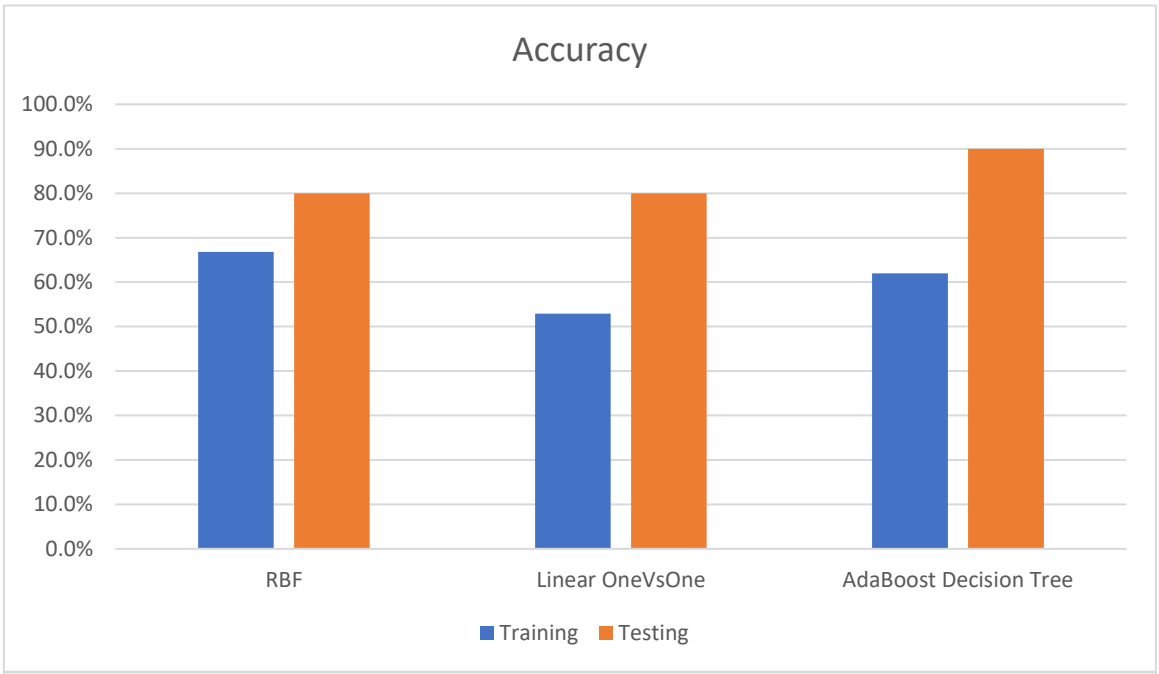
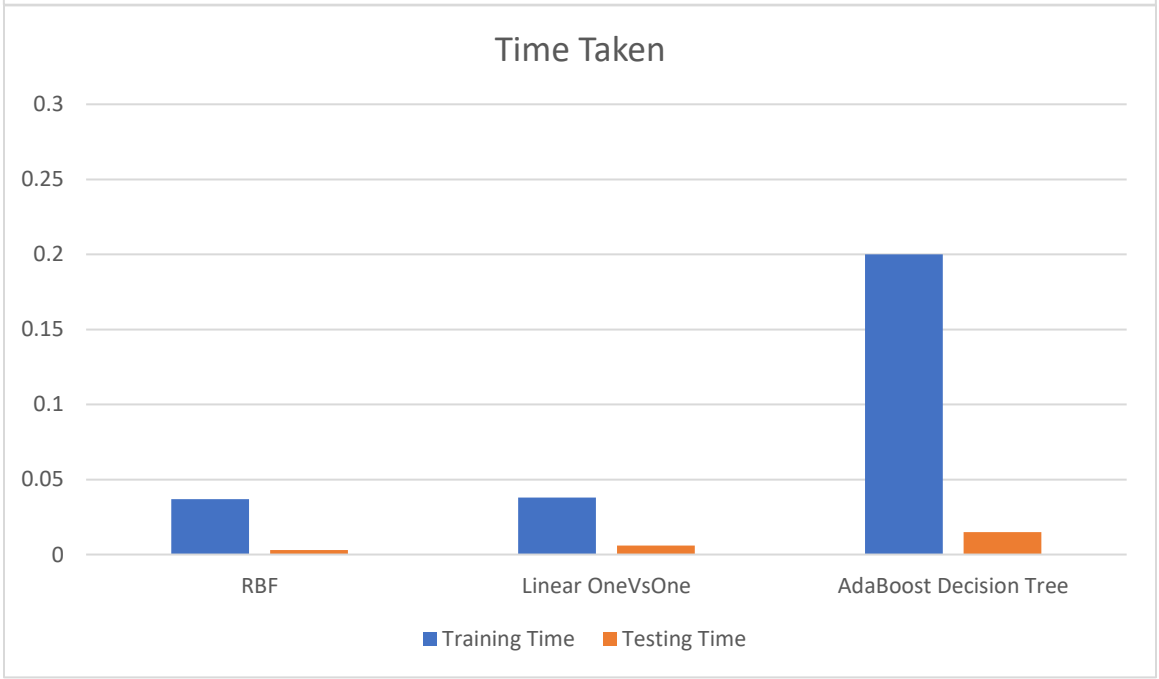
Faculty of Computer & Information Sciences
Ain Shams University
Subject: Machine Learning & Pattern Recognition
Year: (3rd year) undergraduate (CS)
Academic year: 2nd term 2021-2022

Milestone 2 Report
(Classification)

Bar Graphs:



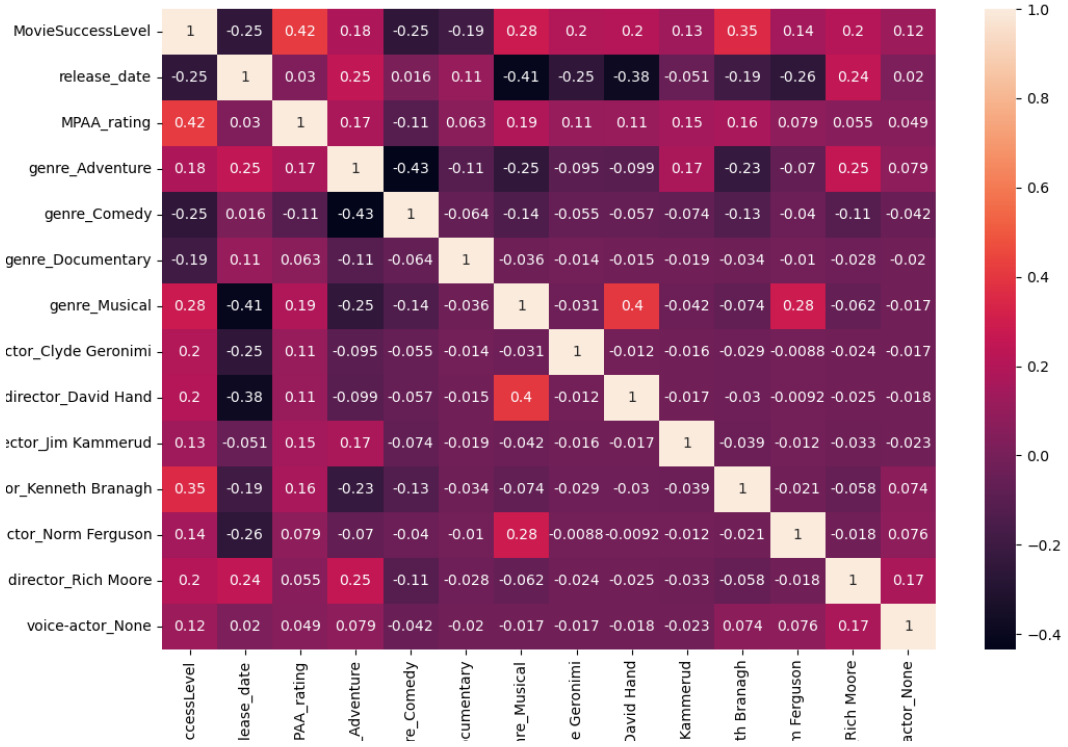


Preprocessing:

- We removed the is_animation feature and merged actor's csv file with our preprocessed csv.
- We used ordinal encoding for both the rating as (S = 4, A = 3, B = 2, C = 1, D = 0) and MPAA rating but MPAA get it dynamically which get the rating available in the current read csv file for training. We used feature scaling on the release date to normalize its data.
- Because we are in classification phase we also didn't need the revenue column anymore, so we dropped it and merged the full preprocessed csv file with the classification one to get the rating using function (classification_preprocessing(df, prev_df, is_drop)).

Feature Selection:

Like Milestone 1 regression, we used correlation to get the best features for classification, but the difference is that we now get the correlation higher than 0.12 instead of 0.17.



Used Algorithms:

RBF, Linear OneVsOne and Decision Tree with AdaBoost.

Any algorithm that is not written here always had lower accuracy than the ones stated above so we didn't include them.

RBF got an accuracy of 66.8%, linear OneVsOne got a 52.9% while decision tree has a range between 57 – 66 %.

Hyperparameters:

The dataset is split at a random state of 21 because it gave the highest accuracy for our models, while the regularization parameter (C) is different for each model.

RBF:

Best result was at C = 1 and a gamma of 1.

Linear OneVsOne:

Best was at C = 0.13

For Decision Trees, a max depth of 4 with 100 estimator had the best result.

Conclusion:

We noticed that RBF have the highest training accuracy at 66.8% which higher than OneVsOne by 13.9% and beats the highest accuracy gotten in the decision trees by a 1% difference.