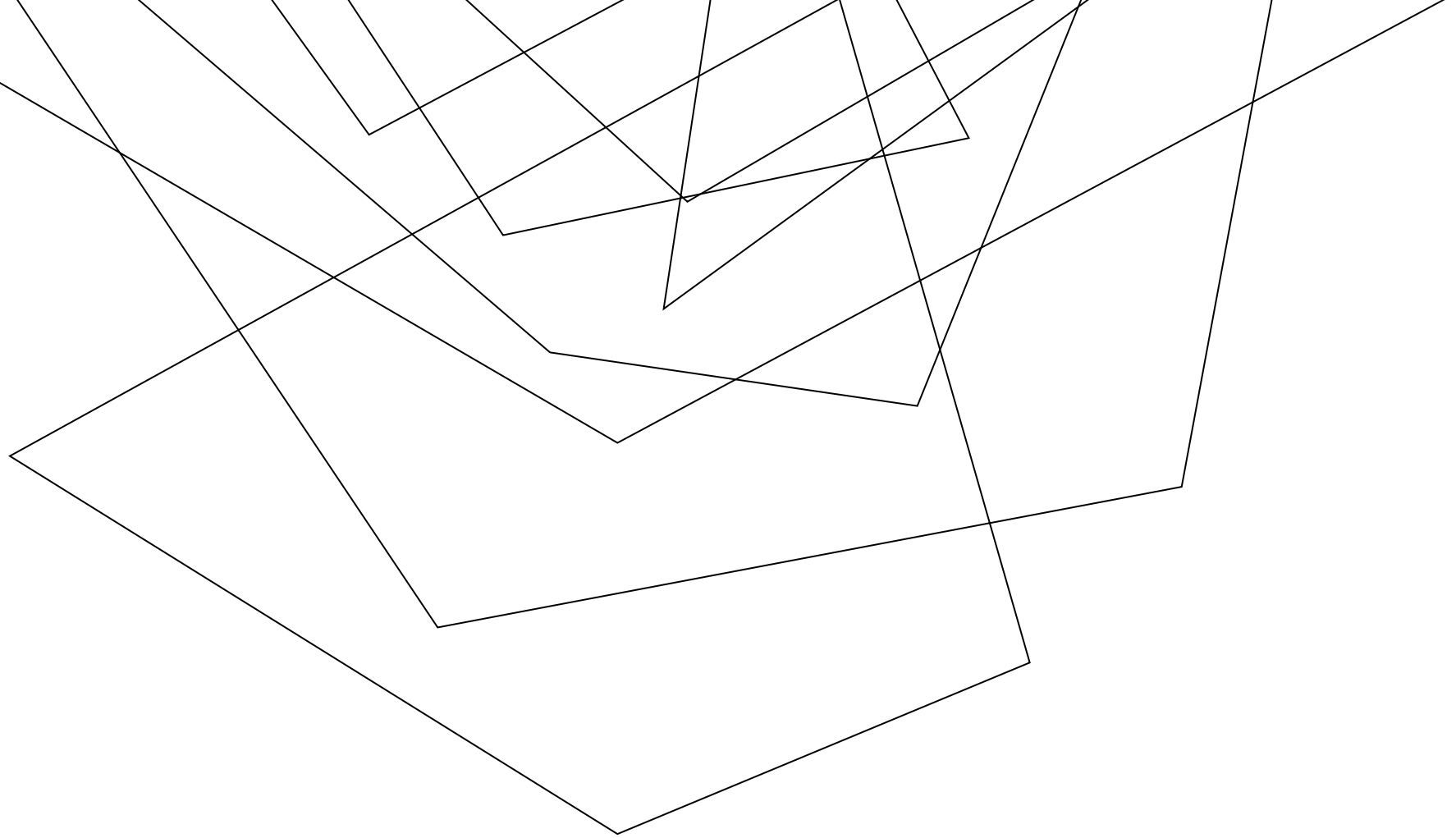
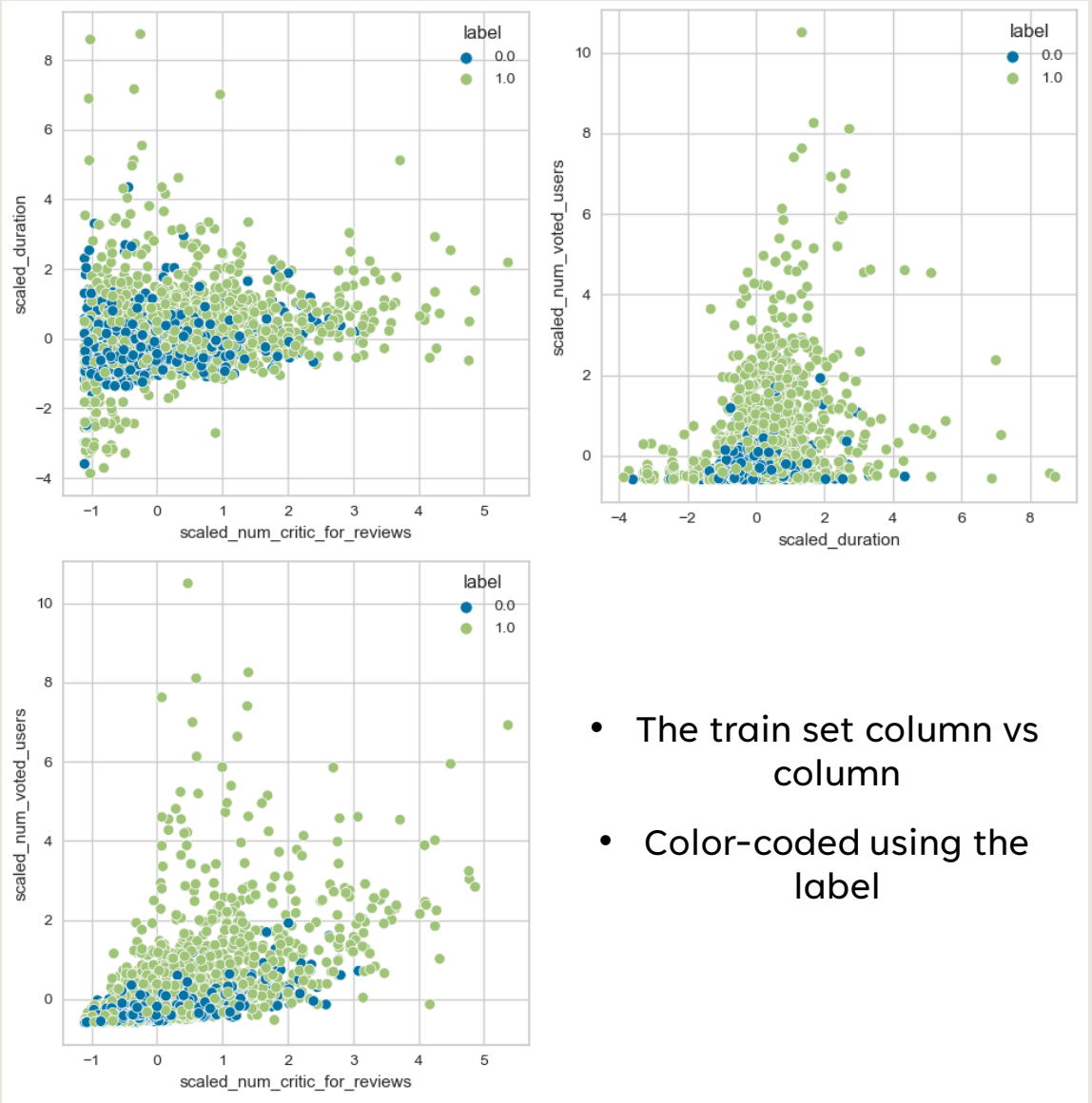


WEEK 6: SVM

6438169421 Pattaradanai Lakkananithiphan

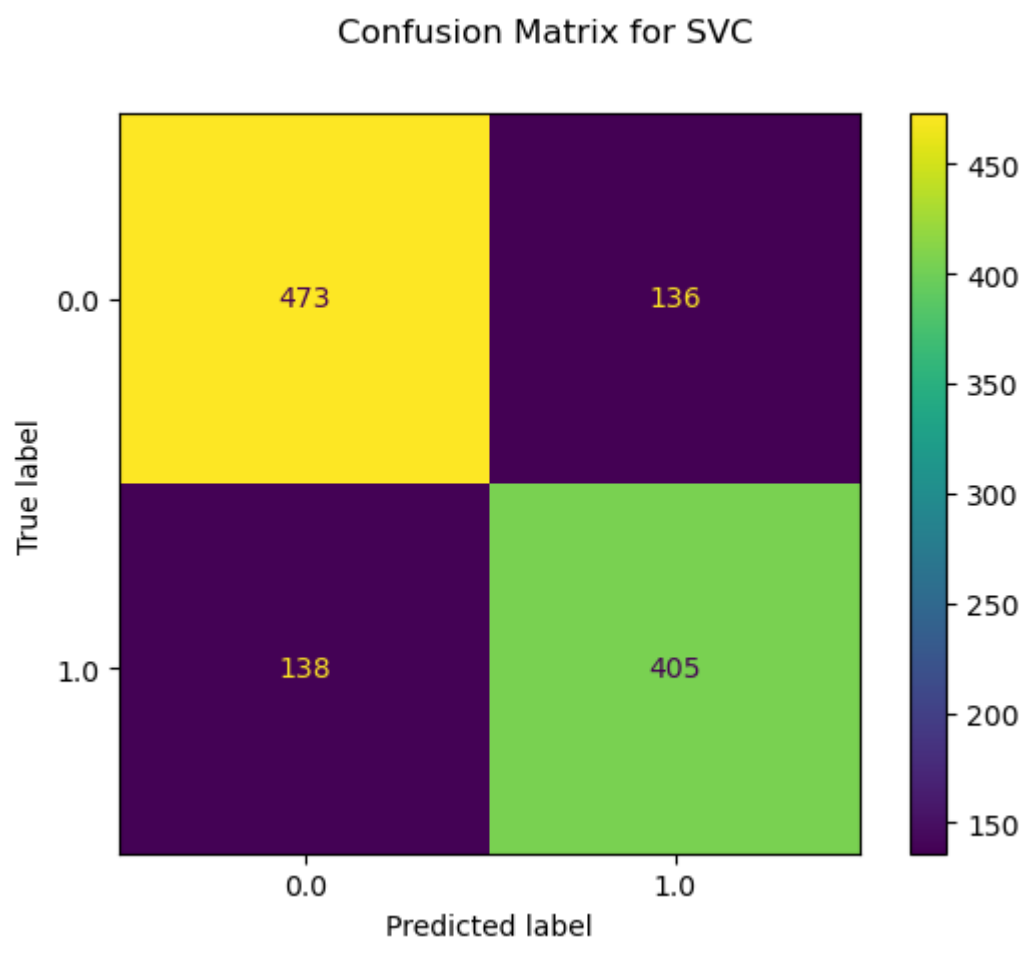
PREPROCESSING

- Drop the rows with “imdb_score” column values in the 40-60 percentile (20% of the data is dropped)
- Add a column of 0/1 according to the “imdb_score” with the condition for 1 being that it exceeds the median of the column -> name it “label”
- Drop the categorical columns as instructed
- Drop all NaN rows (Total of 54 rows = 1.39 % were dropped)
- Drop the columns with low correlation and high multicollinearity
- Scale the features
- Remaining Features: "scaled_num_critic_for_reviews", "scaled_duration", "scaled_num_voted_users“
- Split the train and test set



- The train set column vs column
- Color-coded using the label

MODEL 1: SVC

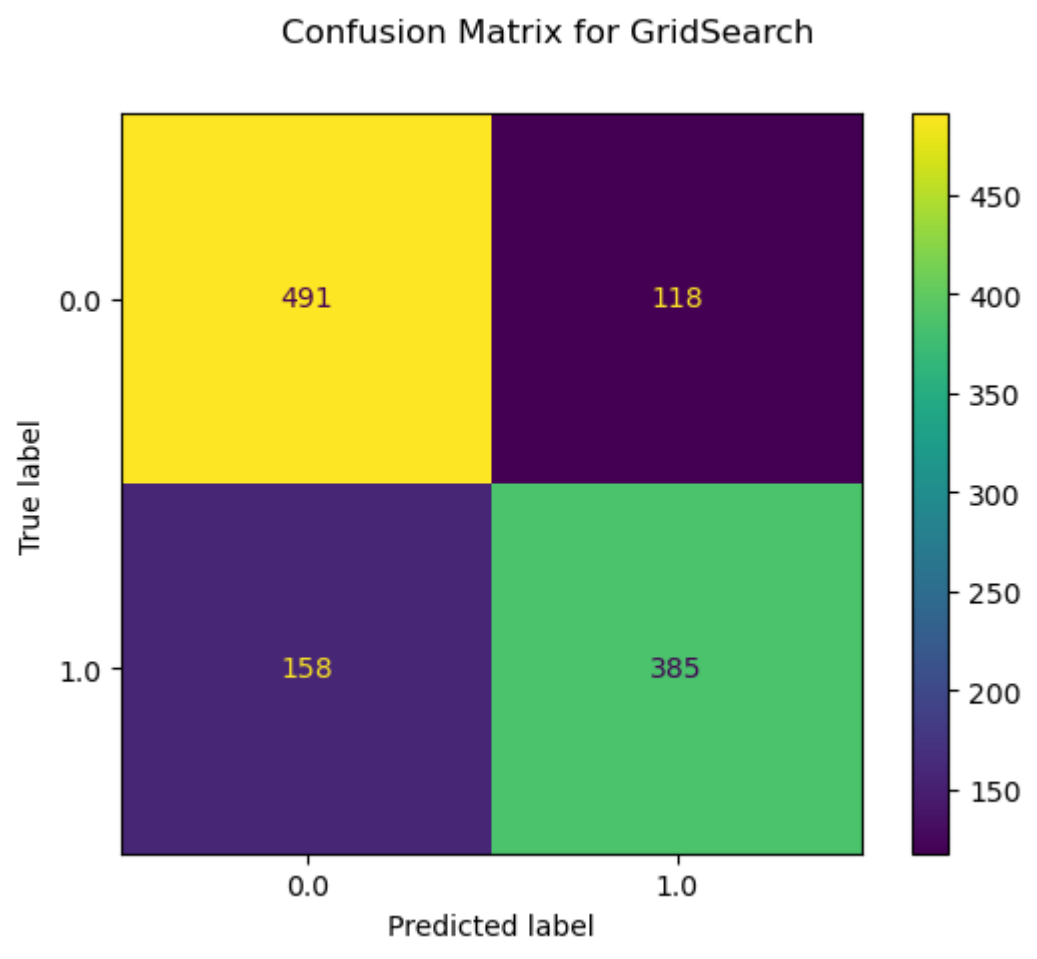



```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.77      0.78      0.78       609
         1.0       0.75      0.75      0.75       543

    accuracy                           0.76      1152
   macro avg       0.76      0.76      0.76      1152
weighted avg       0.76      0.76      0.76      1152
```

```
#Num Support Vectors = 1632 (total vectors = 2685)
```

```
Feature (Permutation) Importances:
scaled_num_critic_for_reviews: 0.04149305555555557
scaled_duration: 0.13237847222222224
scaled_num_voted_users: 0.0921875
```

MODEL 2: SVC + GRIDSEARCHCV



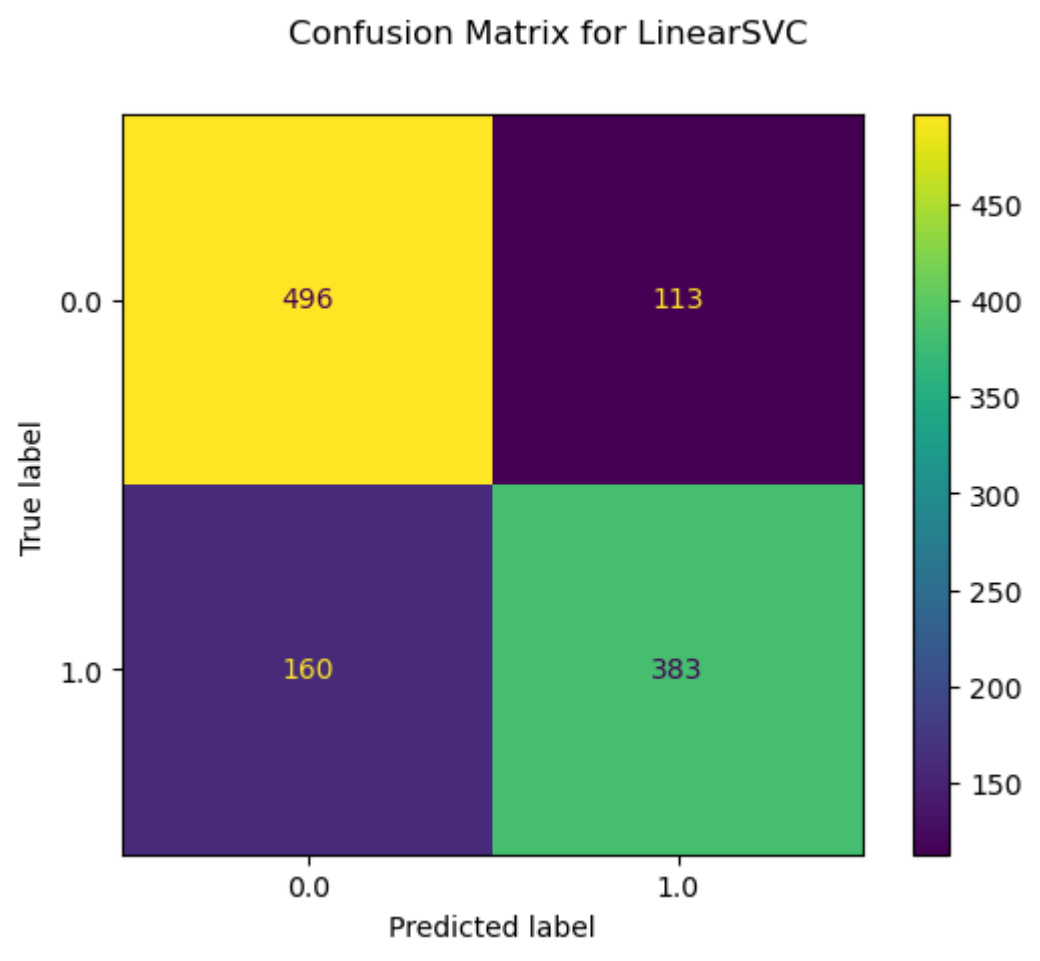
Classification Report:

	precision	recall	f1-score	support
0.0	0.76	0.81	0.78	609
1.0	0.77	0.71	0.74	543
accuracy			0.76	1152
macro avg	0.76	0.76	0.76	1152
weighted avg	0.76	0.76	0.76	1152

```
Best Score:  0.7281191806331471
Best Params:  {'C': 100, 'gamma': 0.1, 'kernel': 'rbf'}
Best Index:  39
Best Estimator:  SVC(C=100, gamma=0.1)
```

```
Feature (Permutation) Importances:
scaled_num_critic_for_reviews: 0.04366319444444441
scaled_duration: 0.12760416666666666
scaled_num_voted_users: 0.0944444444444444
```

MODEL 3: LINEAR SVC + NYSTROEM (25 COMPONENTS)




```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.76      0.81      0.78       609
         1.0       0.77      0.71      0.74       543

    accuracy                           0.76      1152
   macro avg       0.76      0.76      0.76      1152
weighted avg       0.76      0.76      0.76      1152
```

```
Feature Importances:
Feature 1:      0.30088778072932243
Feature 2:      0.6674308536006593
Feature 3:      2.613754430525812
Feature 4:      2.463055947532329
Feature 5:      0.28094293638951395
Feature 6:      13.776151710710904
Feature 7:      1.1371565440374332
Feature 8:      4.022994511654394
Feature 9:      6.6971840852273745
Feature 10:     0.4407185993475042
Feature 11:     0.5858416216739584
Feature 12:     1.1710036534907606
Feature 13:     2.098862908349913
Feature 14:     0.6735231448312492
Feature 15:     0.353278342565955
Feature 16:     2.0532022918775437
Feature 17:     2.9089014350653093
Feature 18:     0.392813516907515
Feature 19:     4.308680461710431
Feature 20:     0.9810660153828189
Feature 21:     2.142410568560043
Feature 22:     1.2585034498801957
Feature 23:     0.48641320927574344
Feature 24:     0.23105179260501005
Feature 25:     8.21760081093035
```

EVALUATION

ALL 3 MODELS PERFORMS SIMILARLY

CHOOSE **LINEAR SVC** FOR ITS SPEED
AND SLIGHTLY HIGHER SCORES