

ASSIGNMENT 2 – Exercise 1.8

Group 2: Ivana Nworah Bortot, Irene Avezzù

Positive/negative rate

Number of pos instances: 1644

Number of neg instances: 1586

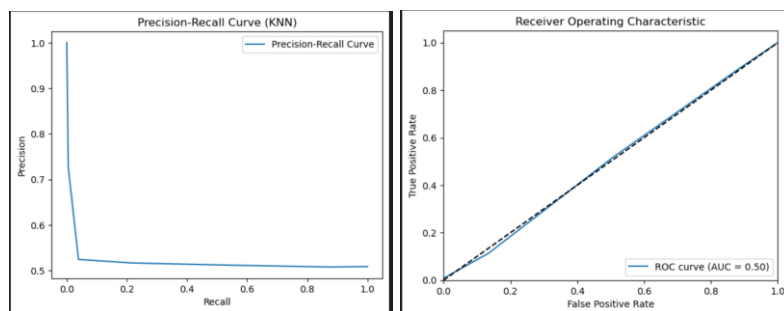
Positive ratio: 0.5089783281733746

Negative ratio: 0.49102167182662537

KNN

Accuracy with KNN classifier: 49.81424148606811

F1 with KNN classifier: 0.30399313009875484



confusion matrix

```
[[1255 331]
```

```
[1290 354]]
```

True negative 1255 --> predicted as negative, and really negative :)

False positive 331 --> predicted as positive, but negative :(

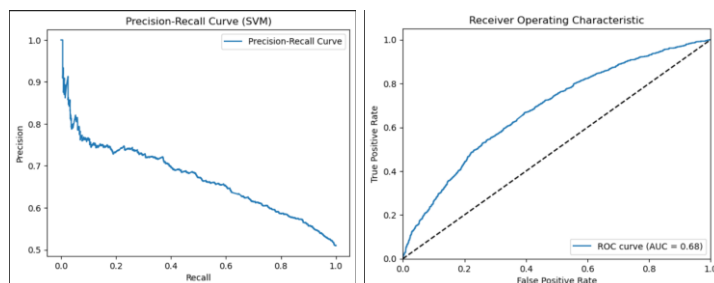
False negative 1290 --> predicted as negative, but positive :(

True positive 354 --> predicted as positive, and really positive :)

SVM

Accuracy with SVM classifier: 62.63157894736842

F1 with SVM classifier: 0.6289578850292038



confusion matrix

```
[[1000 586]
```

```
[ 621 1023]]
```

True negative 1000 --> predicted as negative, and really negative :)

False positive 586 --> predicted as positive, but negative :(

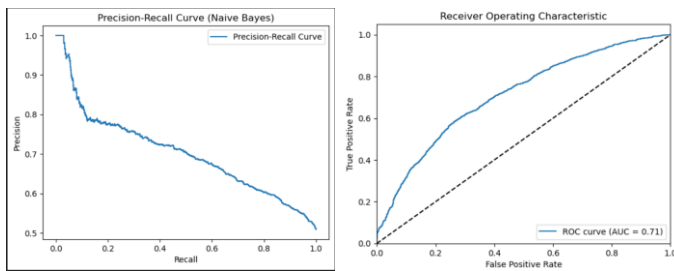
False negative 621 --> predicted as negative, but positive :(

True positive 1023 --> predicted as positive, and really positive :)

Naïve Bayes

Accuracy with Naive Bayes classifier: 64.52012383900929

F1 with Naive Bayes classifier: 0.659131469363474



confusion matrix

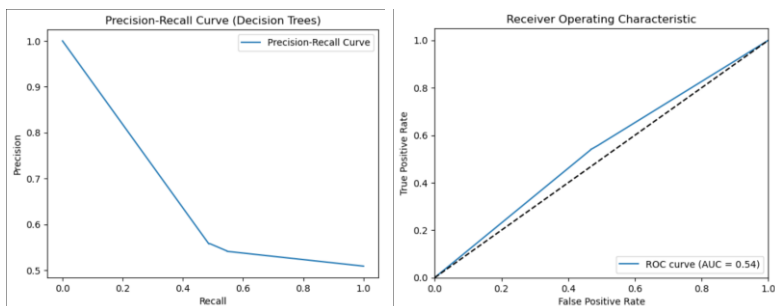
```
[[ 976 610]
 [ 536 1108]]
```

True negative	976 --> predicted as negative, and really negative :)
False positive	610 --> predicted as positive, but negative :(
False negative	536 --> predicted as negative, but positive :(
True positive	1108 --> predicted as positive, and really positive :)

Decision Tree

Accuracy with Decision Trees classifier: 54.27244582043343

F1 with Decision Trees classifier: 0.5209211806681804

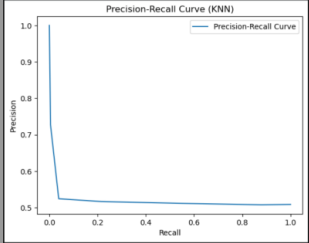
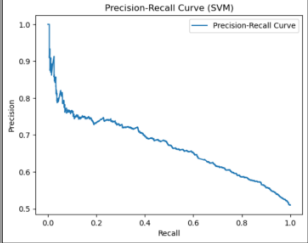
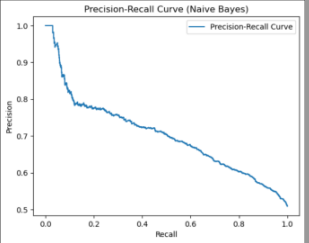
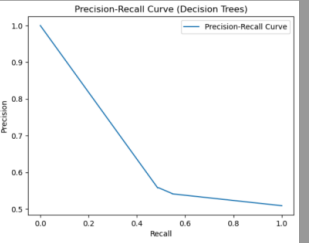
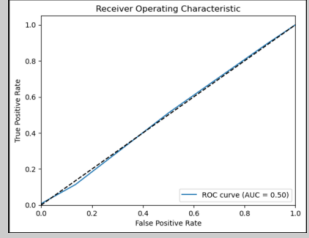
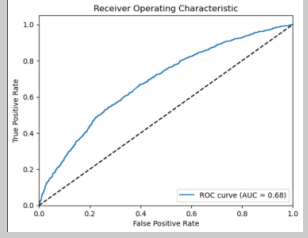
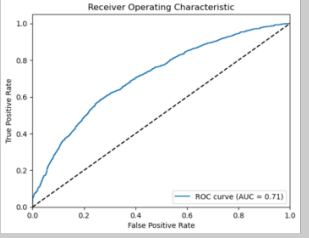
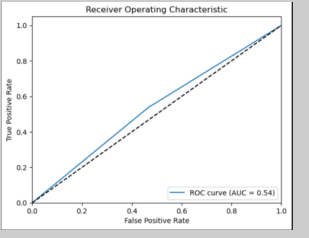


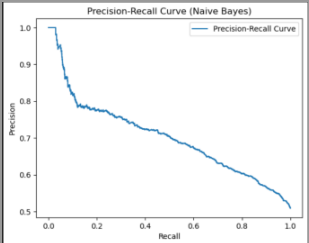
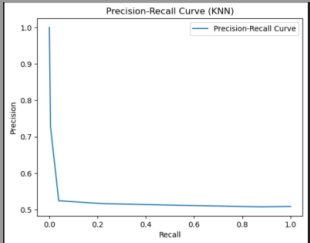
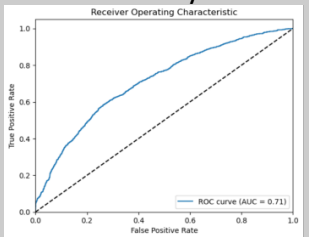
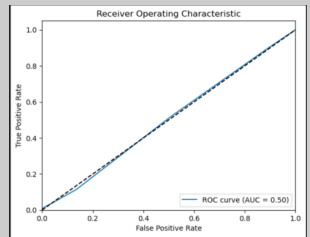
confusion matrix

```
[[950 636]
 [841 803]]
```

True negative	950 --> predicted as negative, and really negative :)
False positive	636 --> predicted as positive, but negative :(
False negative	841 --> predicted as negative, but positive :(
True positive	803 --> predicted as positive, and really positive :)

Summary

	KNN	SVM	Naïve Bayes	Decision Tree
Accuracy	49.81	62.63	64.52	54.27
F1	0.30	0.63	0.66	0.52
TN	1255	1000	976	950
FP	331	586	610	636
FN	1290	621	536	841
TP	354	1023	1108	803
Prec-rec curve				
ROC curve				

	Best	Worst
Accuracy	Naïve Bayes: 64.52	KNN: 49.81
F1	Naïve Bayes: 0.66	KNN: 0.30
TP	Naïve Bayes: 1108	KNN: 354
TN	KNN: 1255	Decision Tree: 950
FP	KNN: 331	Decision Tree: 636
FN	Naïve Bayes: 536	KNN: 1290
Prec-rec curve	Naïve Bayes: 	KNN: 
ROC curve	Naïve Bayes: 	KNN: 

Comment

When observing the negative and positive rate we notice that the two classes are pretty much balanced.

By applying four different classifier we can observe which one is the most accurate.

Overall, the multinomial Naïve Bayes is the best classifier one with the most accurate values for accuracy, F1, TPR, Prec-rec and ROC curve. This classifier gets the best performances on discrete features; in our case we did not have discrete features but using the tdf-tdf vectorizer we are able to represent with numerical features non-discrete values.

Also, by observing the confusion matrix we observe that the multinomial Naïve Baye achieves great rate for TP and TN values with only minor misclassification errors.

On the other hand, the worst classifier is the KNN which, as we can see from both the Prec-rec and ROC curve behaves as a random classifier. KNN is a type of classifier which gets its optimal results when working with image/video recognition but is less useful with labeled data.