AI project ADA

Session 5

Team

Yuliia Nikolaenko
Eduardo Bonnefemne
Gabriel Pérez García
Elena-Mihaela Grigore

Performance Prediction Challenge

Competition context

‘The project is dedicated to stimulate research and reveal the state-of-the art in "model selection" by organizing a competition followed by a workshop.

The competition will help identifying accurate methods of model assessment, which may include variants of the well-known cross-validation methods and novel techniques based on learning theoretic performance bounds.’

‘The aim of the challenge in performance prediction is to find methods to predict how accuratly a given predictive model will perform on test data, on ALL five benchmark datasets. To facilitate entering results for all five datasets, all tasks are two-class classification problems.’

Part I: Exploring our data

Data

Name: ADA

Domain: Marketing

Size: 0.6 MB

Type: Dense

Features: 48

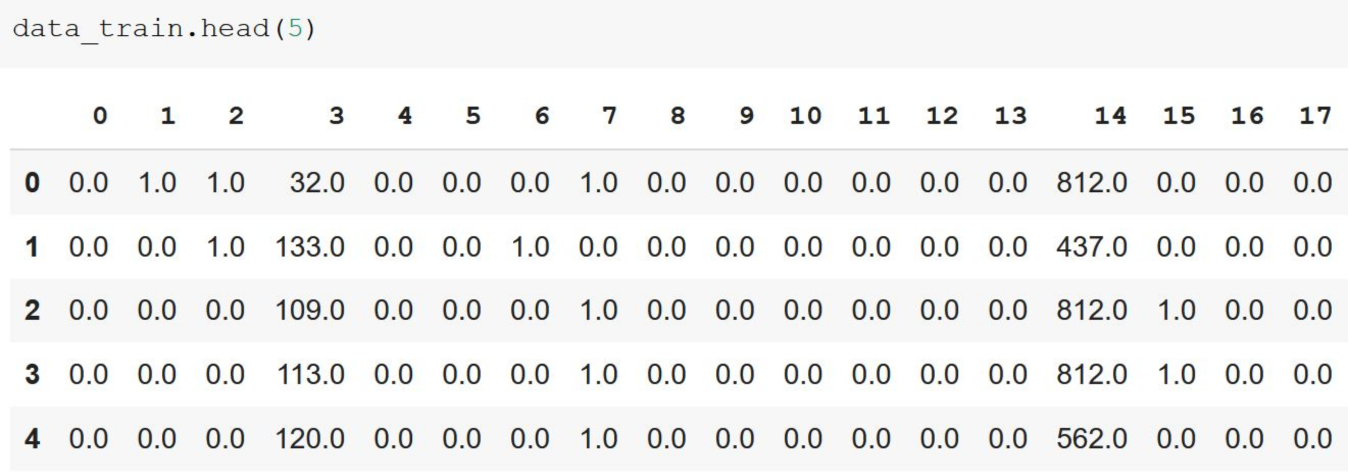
Training Examples: 4147

Validation Examples: 415

Test Examples: 4147

EDA

- Data has no columns description
- Most of our features have values between 0 and 1
- Only 6 of them have values up to 1000 [3, 9, 1, 24, 31, 39] -> explore further
- Two columns have only 0s [13, 20] -> drop

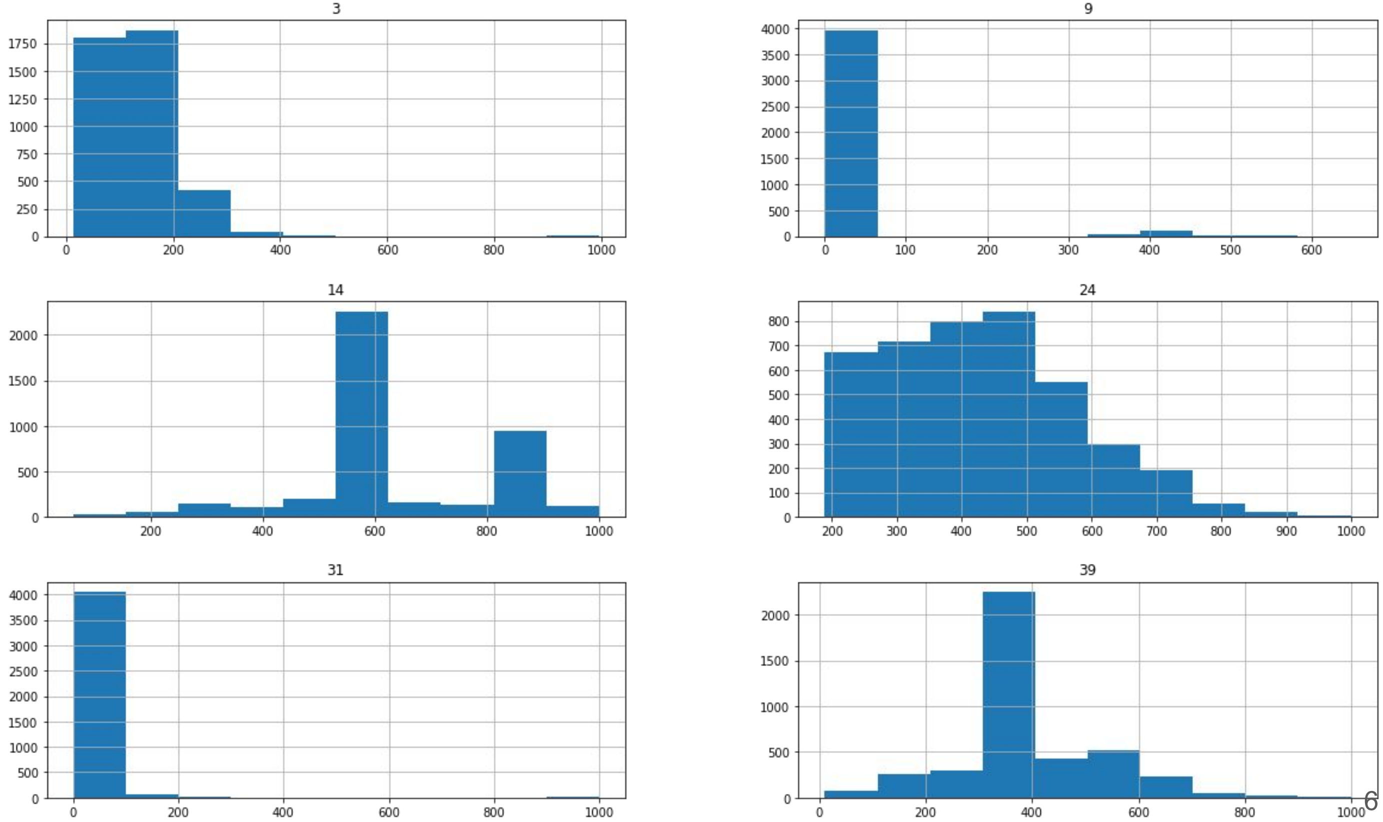


```
data_train.head(5)
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	0.0	1.0	1.0	32.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	812.0	0.0	0.0	0.0
1	0.0	0.0	1.0	133.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	437.0	0.0	0.0	0.0
2	0.0	0.0	0.0	109.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	812.0	1.0	0.0	0.0
3	0.0	0.0	0.0	113.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	812.0	1.0	0.0	0.0
4	0.0	0.0	0.0	120.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	562.0	0.0	0.0	0.0

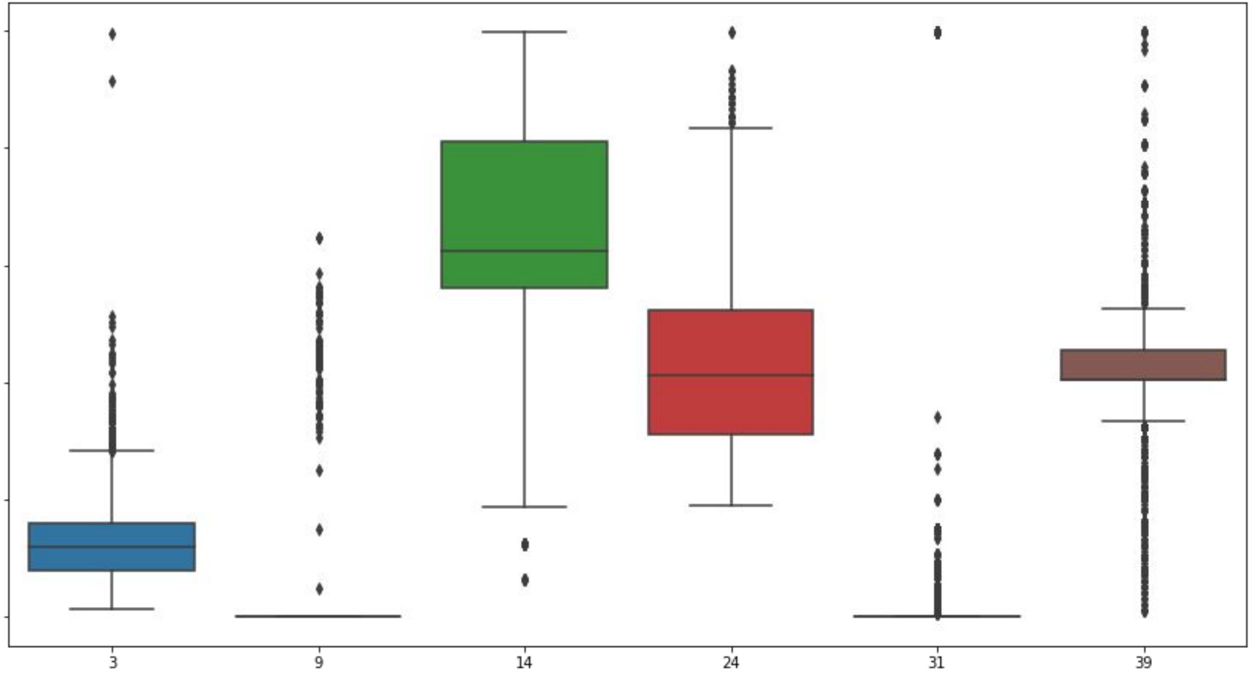
Descriptive statistics: values with outliers

Possible outliers visible in the histograms -> to explore further



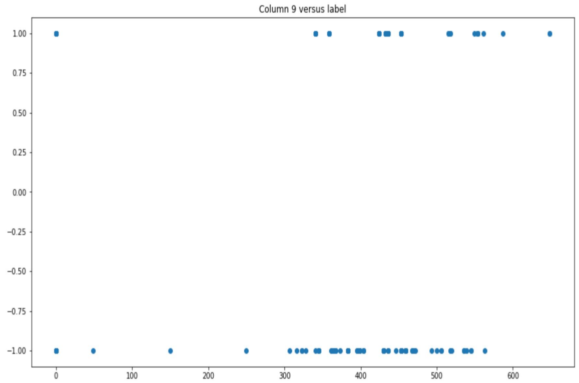
Descriptive statistics: values with outliers

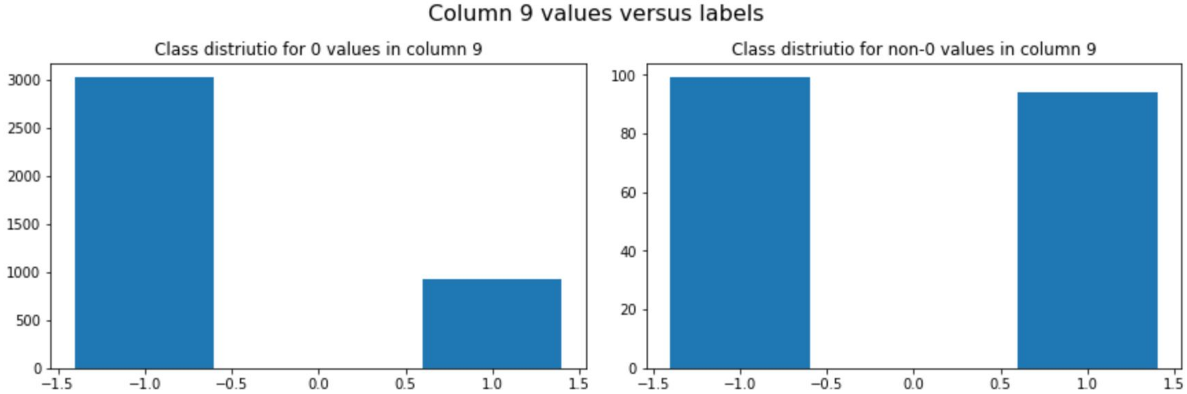
- Too many outliers in all but column 14
- Columns [9, 31] have mostly 0s -> explore the relationship with class labels



Column 9 x class labels

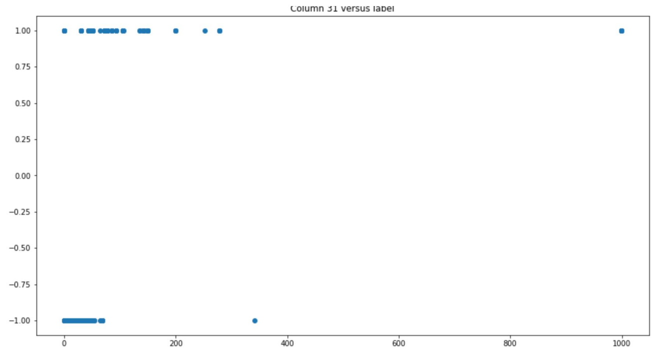
- Exploration of the relationship between non-zero values in column 9 and class label

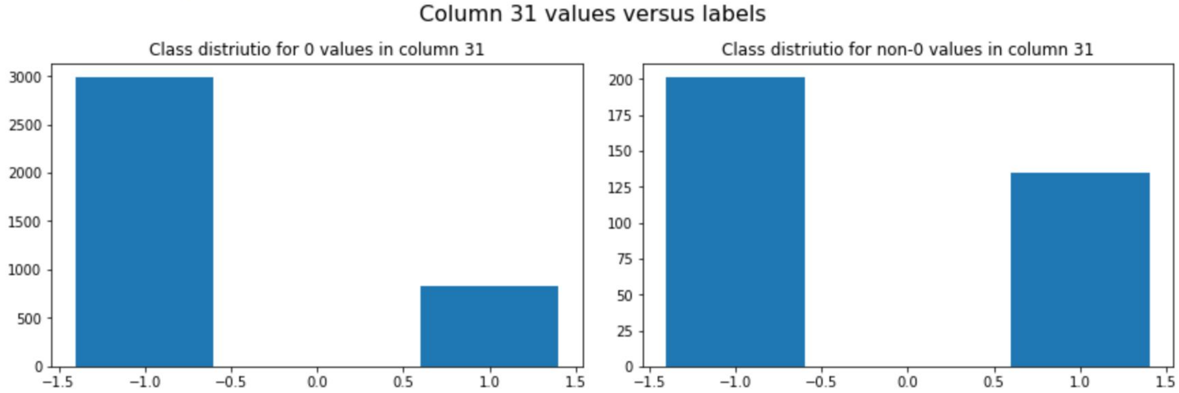



Column 9 values versus labels

Column 31 x class labels

- Exploration of the relationship between non-zero values in column 31 and class label





Data engineering

Dealing with outliers:

1. in columns [3 14 24 39] we have too many values that qualify as outlier. So we will let them be;
2. in columns 9 and 31 we have 90% of values equal to 0 and the rest go up to 1000. We don't see a correlation with class value. So we drop these columns completely.

Normalizing the 4 remaining features [3 14 24 39] that have values between 0 and 1000.

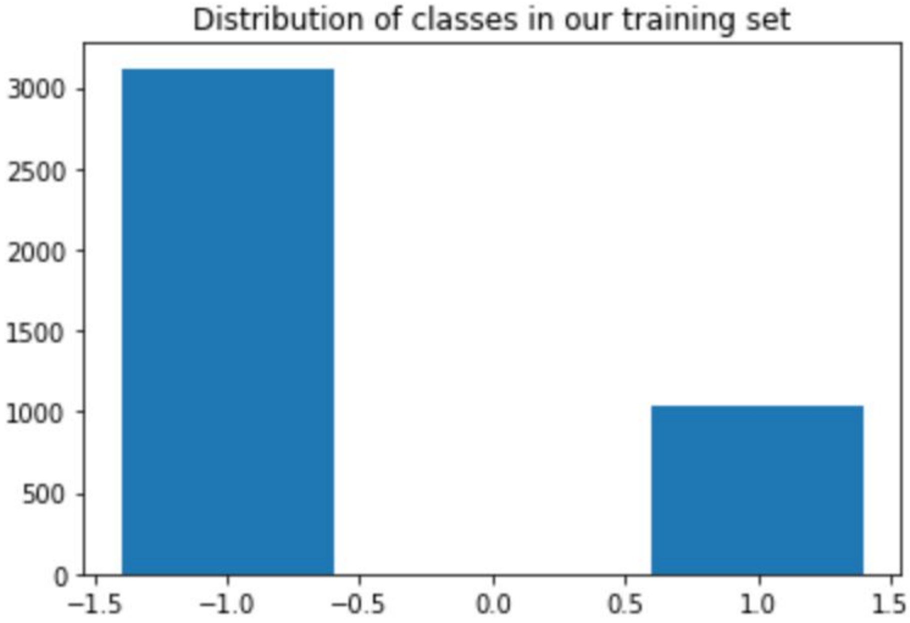
Dropping the features that contain only 0 values [13 20]

Exploring class balance in our training set

3000 x class -1

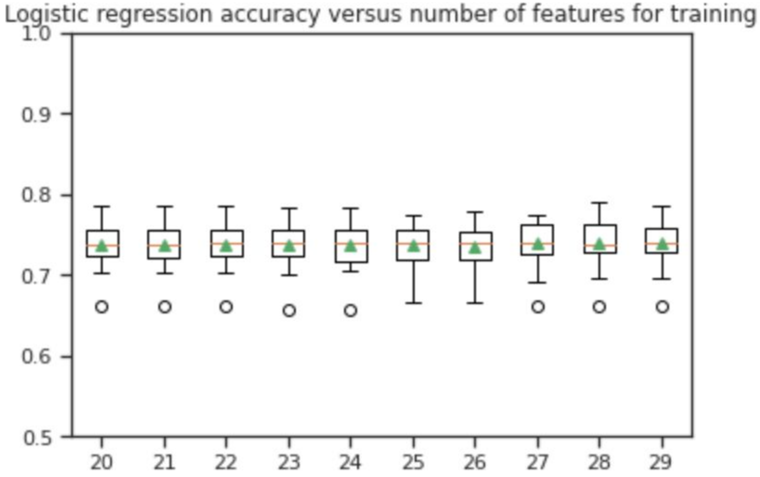
1000 x class 1

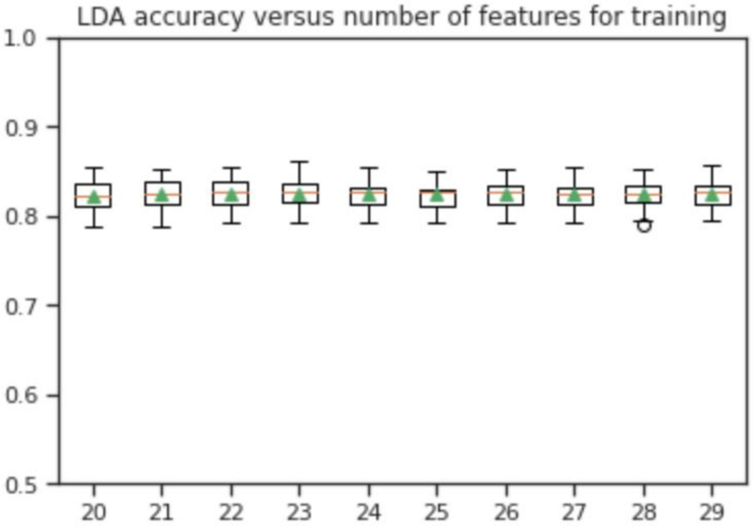
=> careful with scoring metrics, since our classes are not balanced

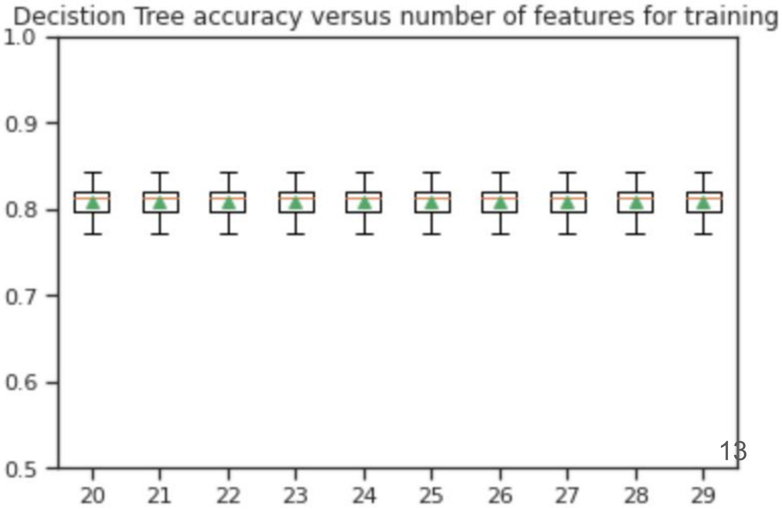


Part II: Model Building

RFE for feature selection



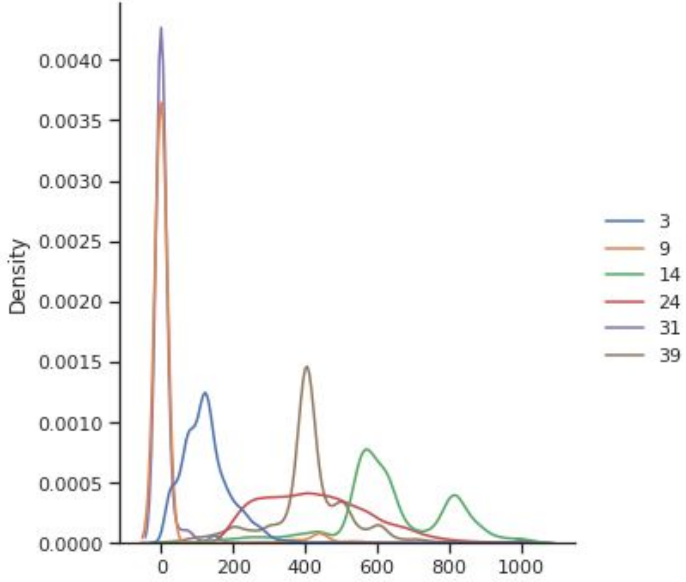


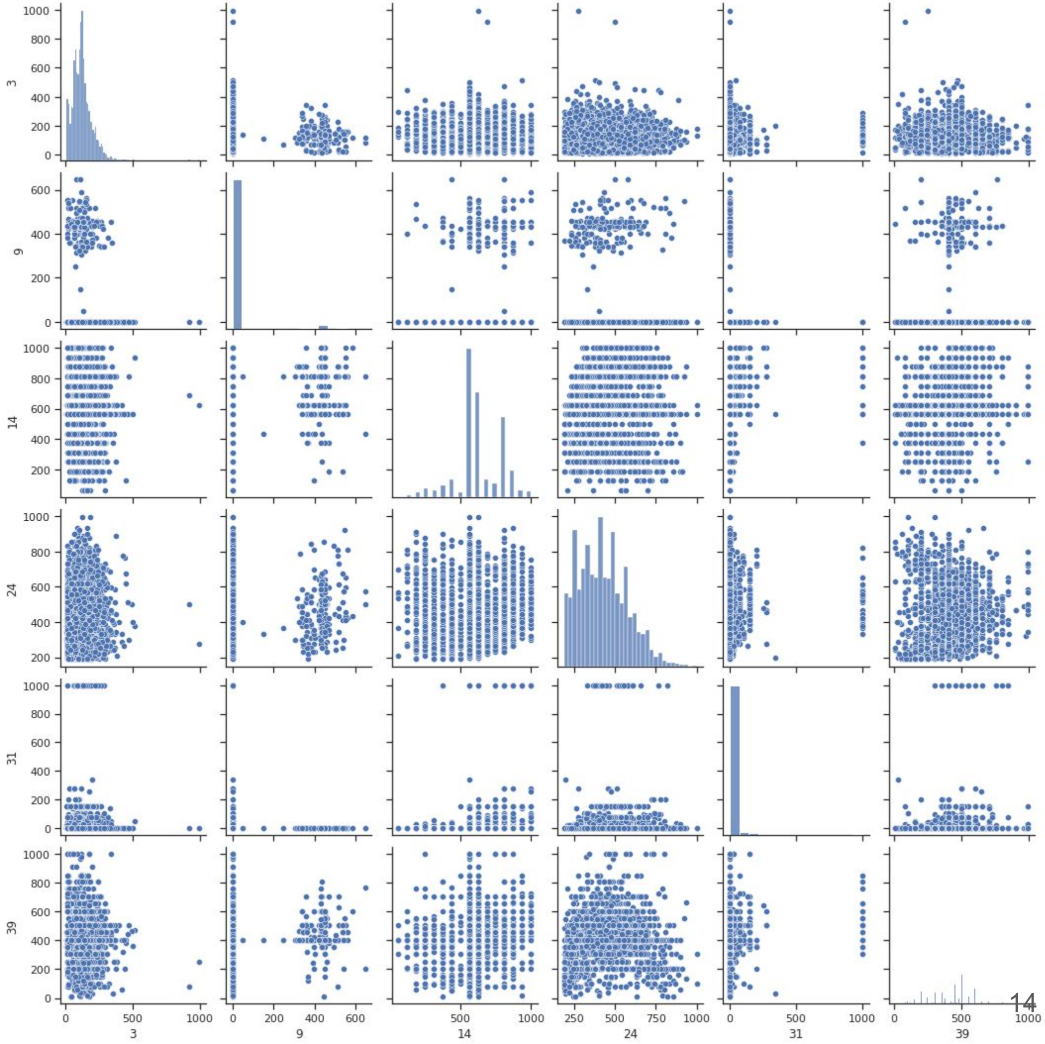


Small differences (on the 3rd decimal)

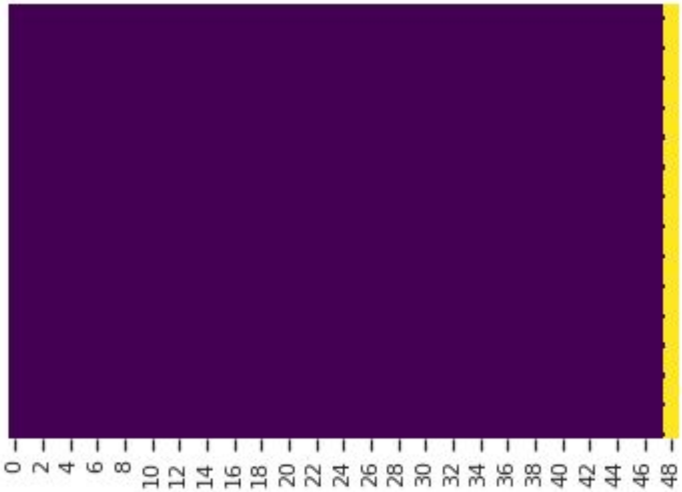
We will use RFE with default

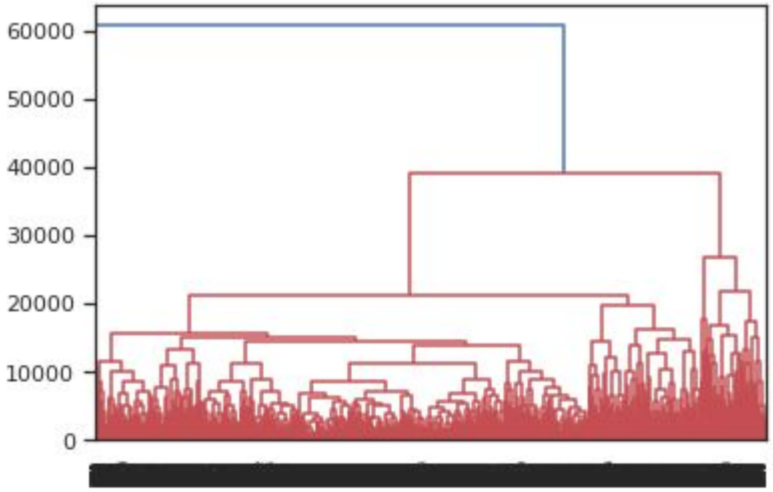
Naive bayes


Density
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
200
400
600
800
1000
3
9
14
24
31
39

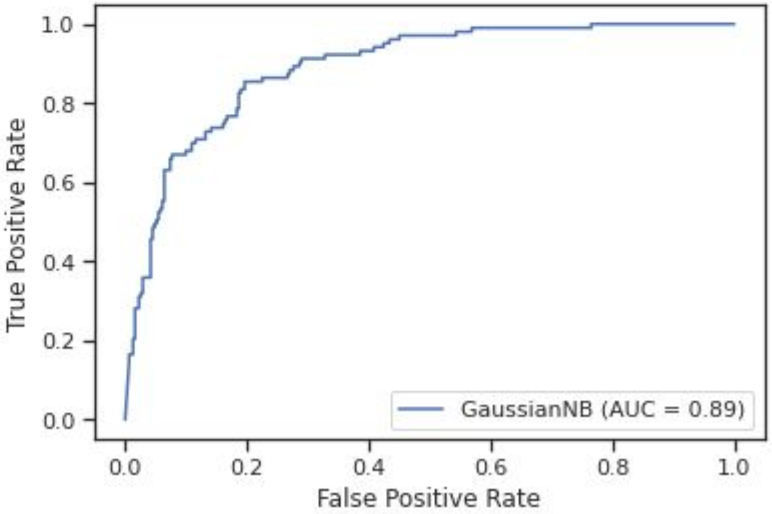

3
9
14
24
31
39
0
200
400
600
800
1000
0
500
1000
0
200
400
600
500
1000
250
500
750
1000
0
500
1000
0
500
1000

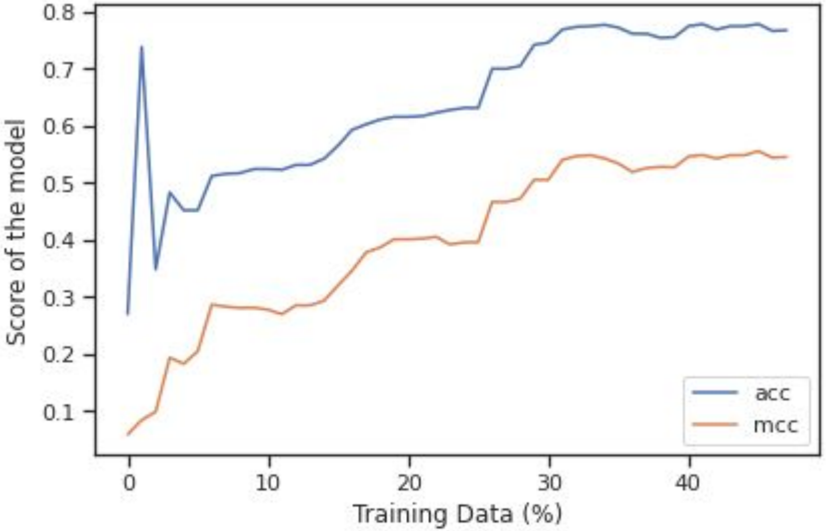
Naive bayes


0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48


60000
50000
40000
30000
20000
10000
0

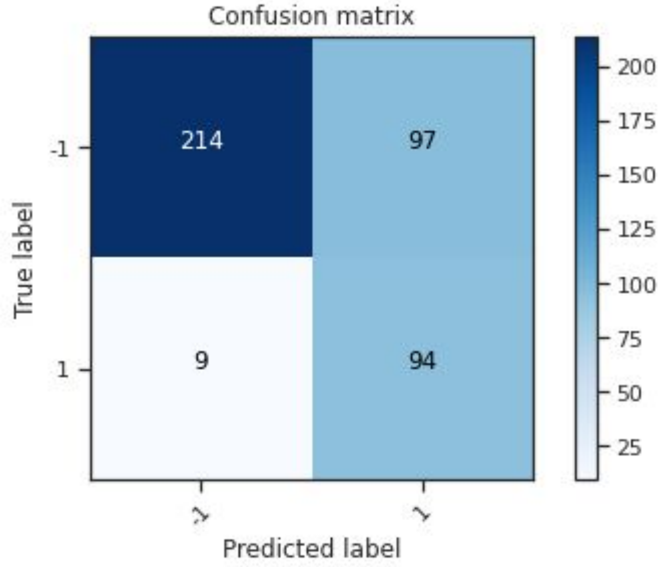
Naive bayes

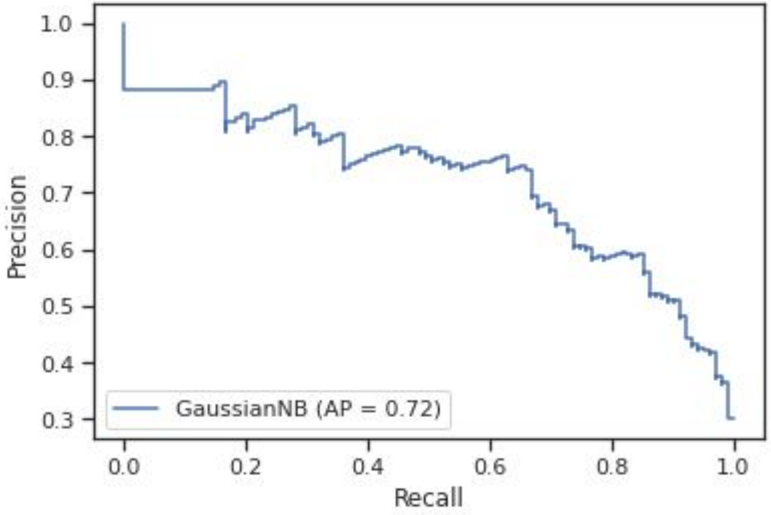




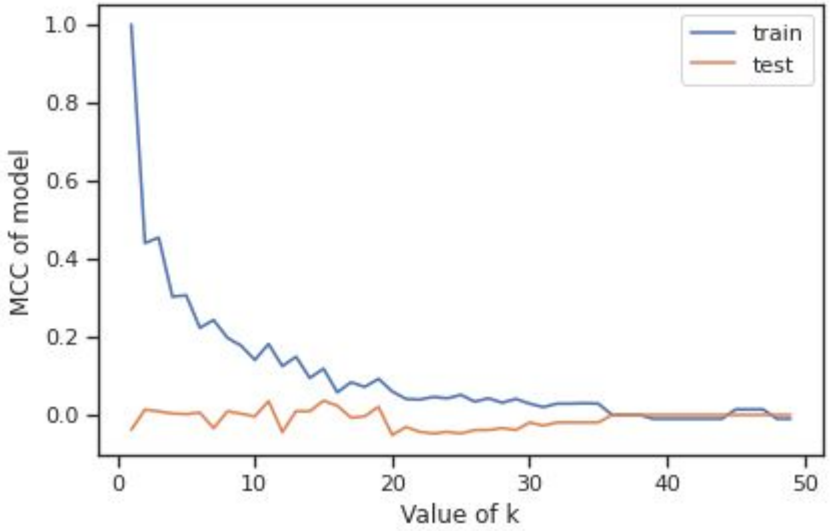
```
The best MCC was 0.55.
```

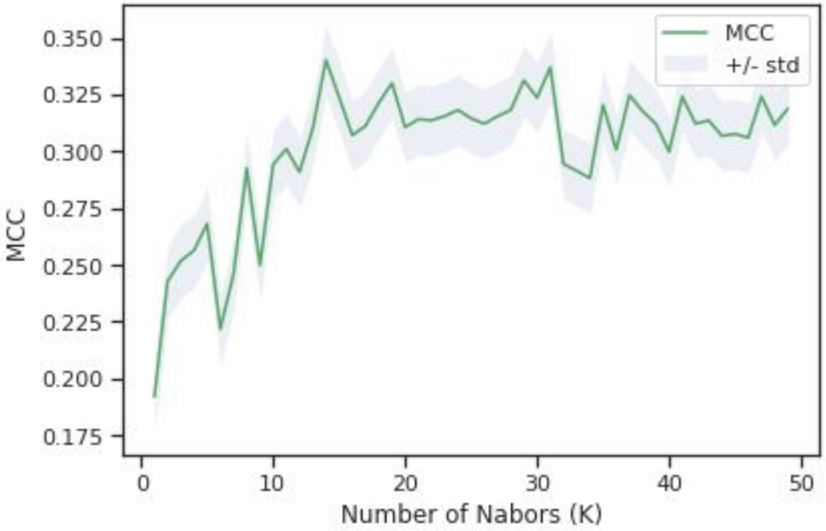
Naive bayes


Confusion matrix
214
97
9
94
-1
1
True label
-1
1
Predicted label
200
175
150
125
100
75
50
25


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall
GaussianNB (AP = 0.72)

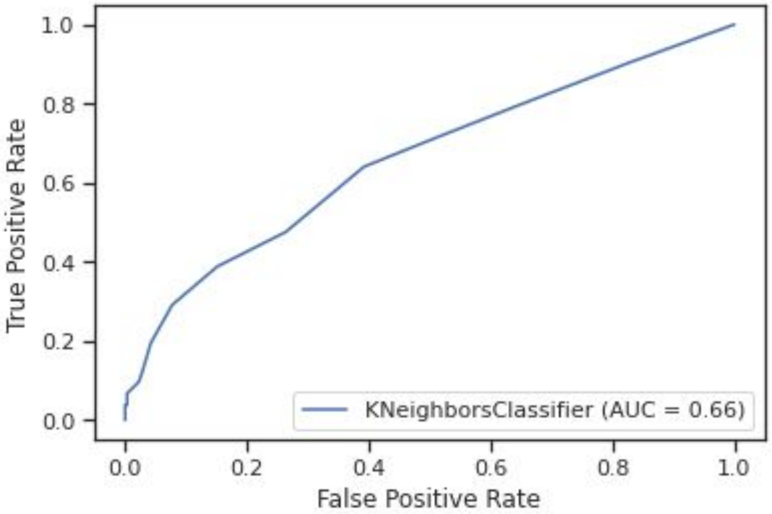
KNN

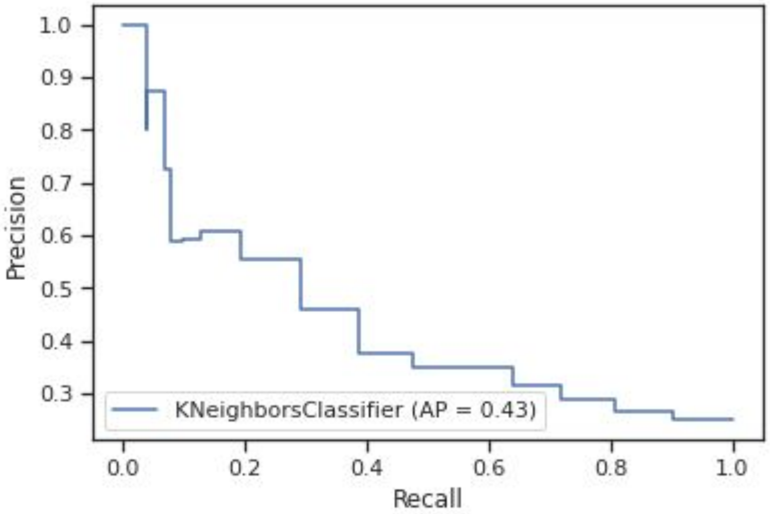




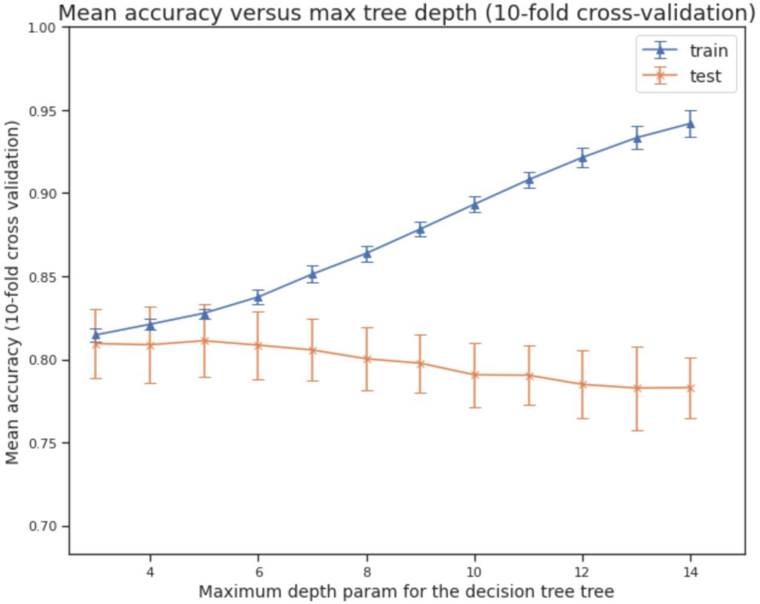
```
The best MCC was 0.3402 with k= 14
```

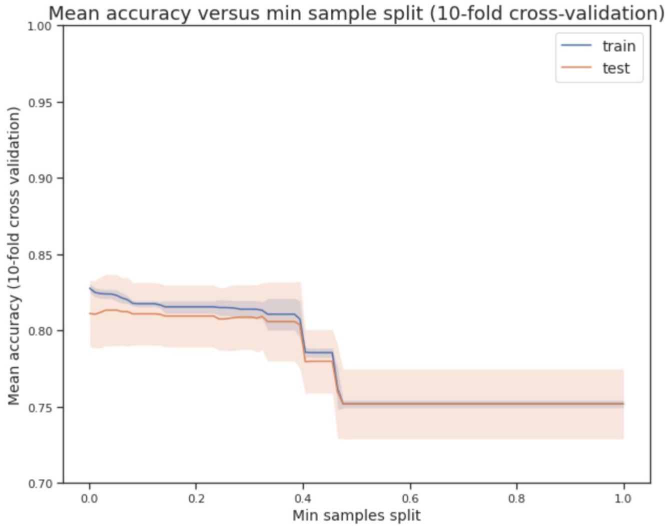
KNN

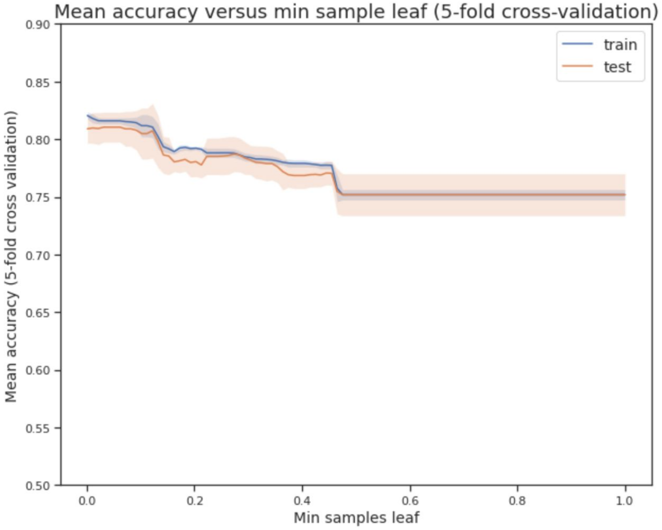

1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
KNeighborsClassifier (AUC = 0.66)


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall
KNeighborsClassifier (AP = 0.43)

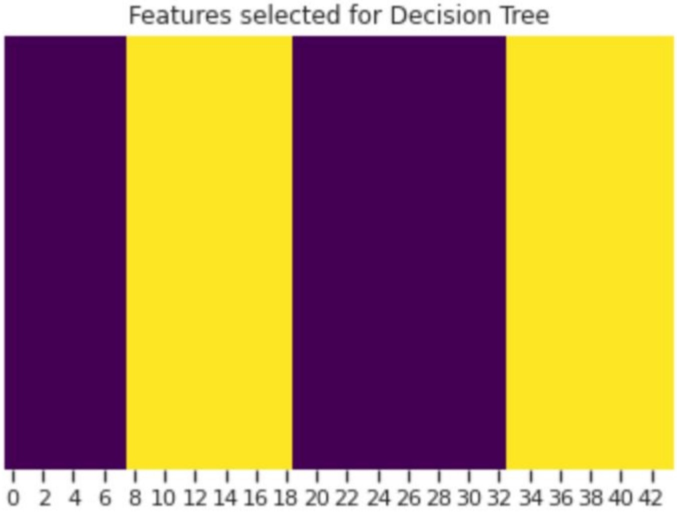
Decision Tree: pre-pruning

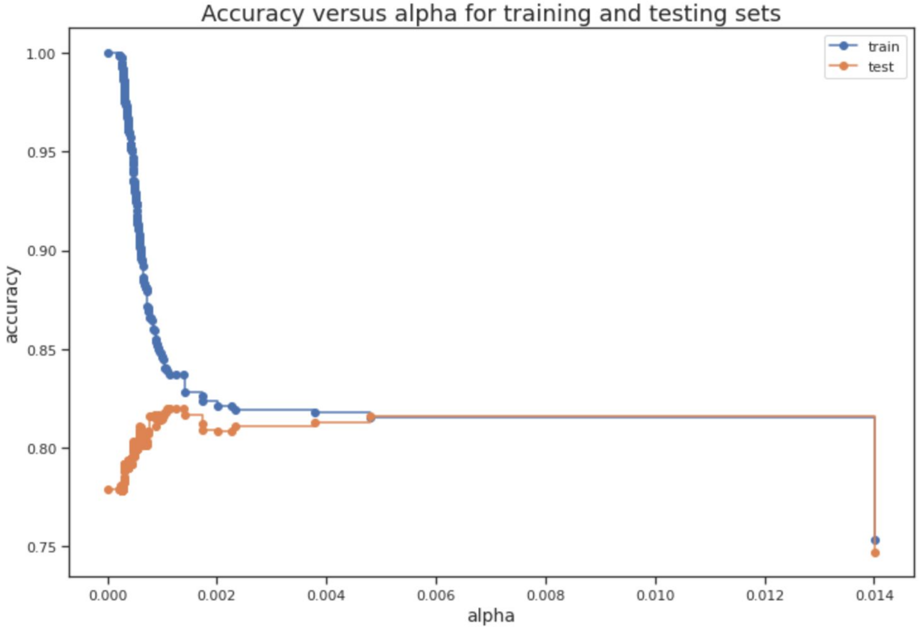


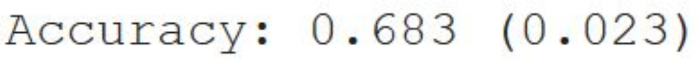




Decision Tree: post-pruning and RFE







```
Accuracy: 0.683 (0.023)
```

Part III: Model Comparison

Model Comparison