MSE - MLDB - Lab 4 - TREFLE

Authors:

- Romain Claret
- Edward Ransome

Introduction

In this lab, we will use three datasets (Cancer, BCWD, GOLUB) and apply various fuzzy logic models to them using the Trefl library.

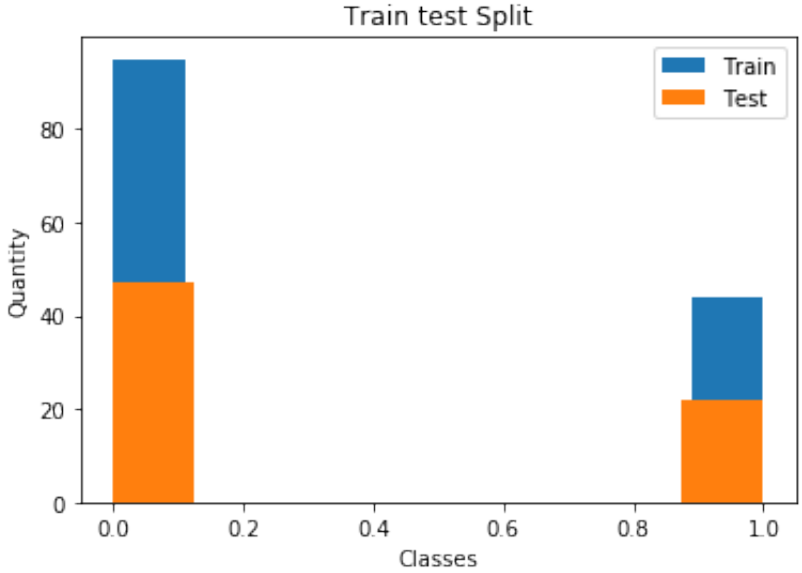
Dataset cancer

Question 1

- The original dataset shape is (208, 17).
  - X shape: (208, 15)
  - y shape: (208,)

The train/test separation is done using a 66/33 split for all datasets.

We can see that there are more data points of class 0 than of class 1 (twice as many), but they are well distributed in the train/test split (the training set is always twice as large as the test set for both classes).



Question 2

The number of rules indicates how many rules will be generated in the model. This will affect calculation time but will increase accuracy on the training set. To choose the optimal number of rules, one must refer to literature. However, one can choose too many rules and overfit on the training data since some rules will represent noise in the data.

Variables per rule indicated how many variables can be used when defining a rule, for example, the rule “IF v11 is low AND v4 is low THEN [0]” contains two variables. This means a rule can be more or less complex based on the number of variables: if many features interact to give a specific class, i.e., the problem is complex, more variables must be used per rule.

Question 3

If we assume each rule uses different features, we can have at maximum 5*6 = 30 features used. However, the same variable can appear in many rules in practice.

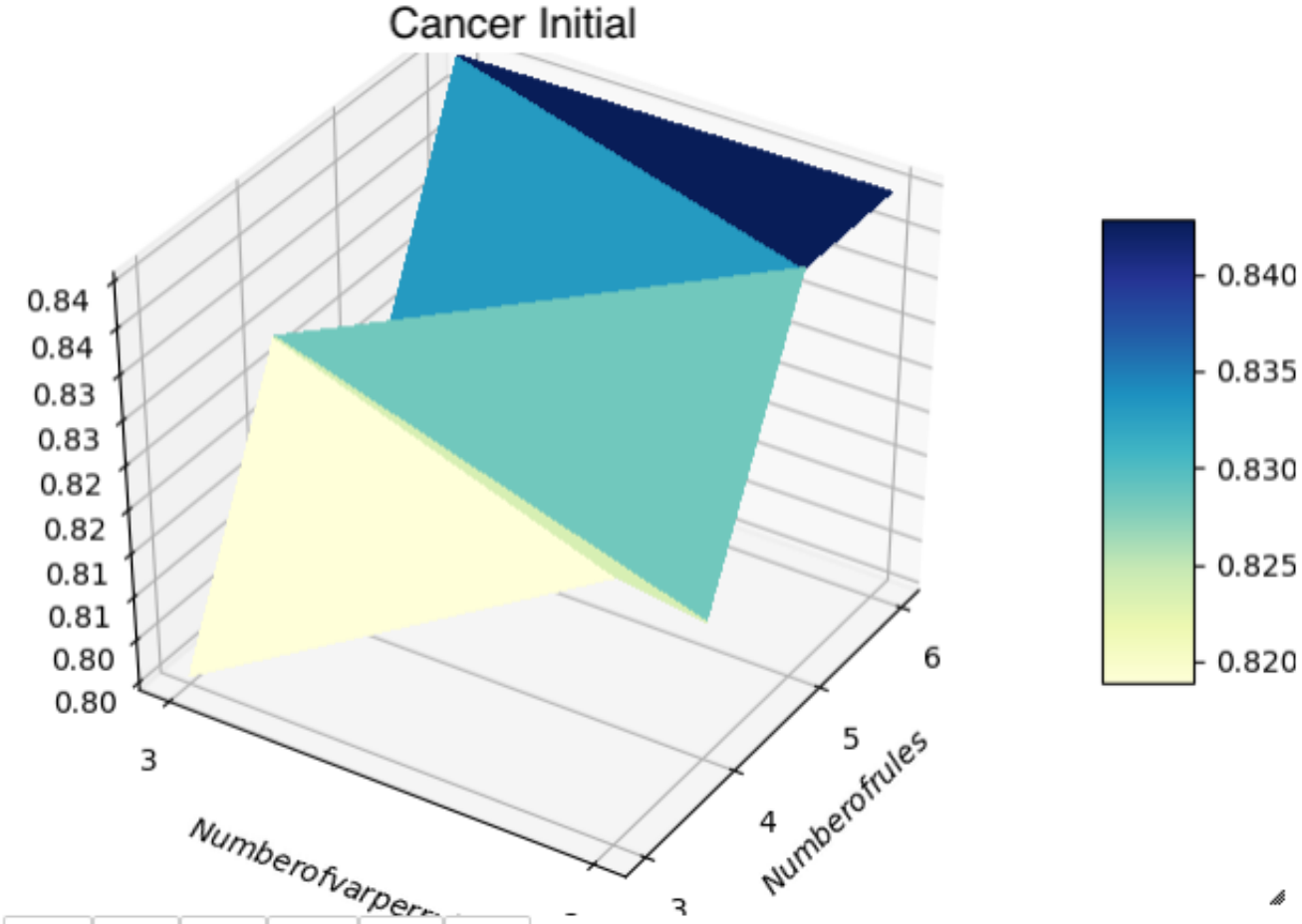
Question 4

Since the dataset has the lowest number of features (15), we decided to test lower amounts of combinations of rules/variables per rule for this dataset. Due to time constraints on the calculations, we decided for this dataset to only test two amounts of variables per rule (2 or 3). This reduced the combinations and allowed us to test from 3–6 rules. In hindsight, however, with more time, we would have preferred to add 4 variables per rule to see the effects, even though with only 15 features we suspect it would lead to rules being based on noise rather than the underlying data.

Question 5

We can see from the 3D visualization that the number of rules increases accuracy until 5 rules are reached, at this point the values flatten out. The number of variables per rule did not highly affect the results between 2 and 3 variables. We, therefore, feel the ideal combination for this dataset is 6

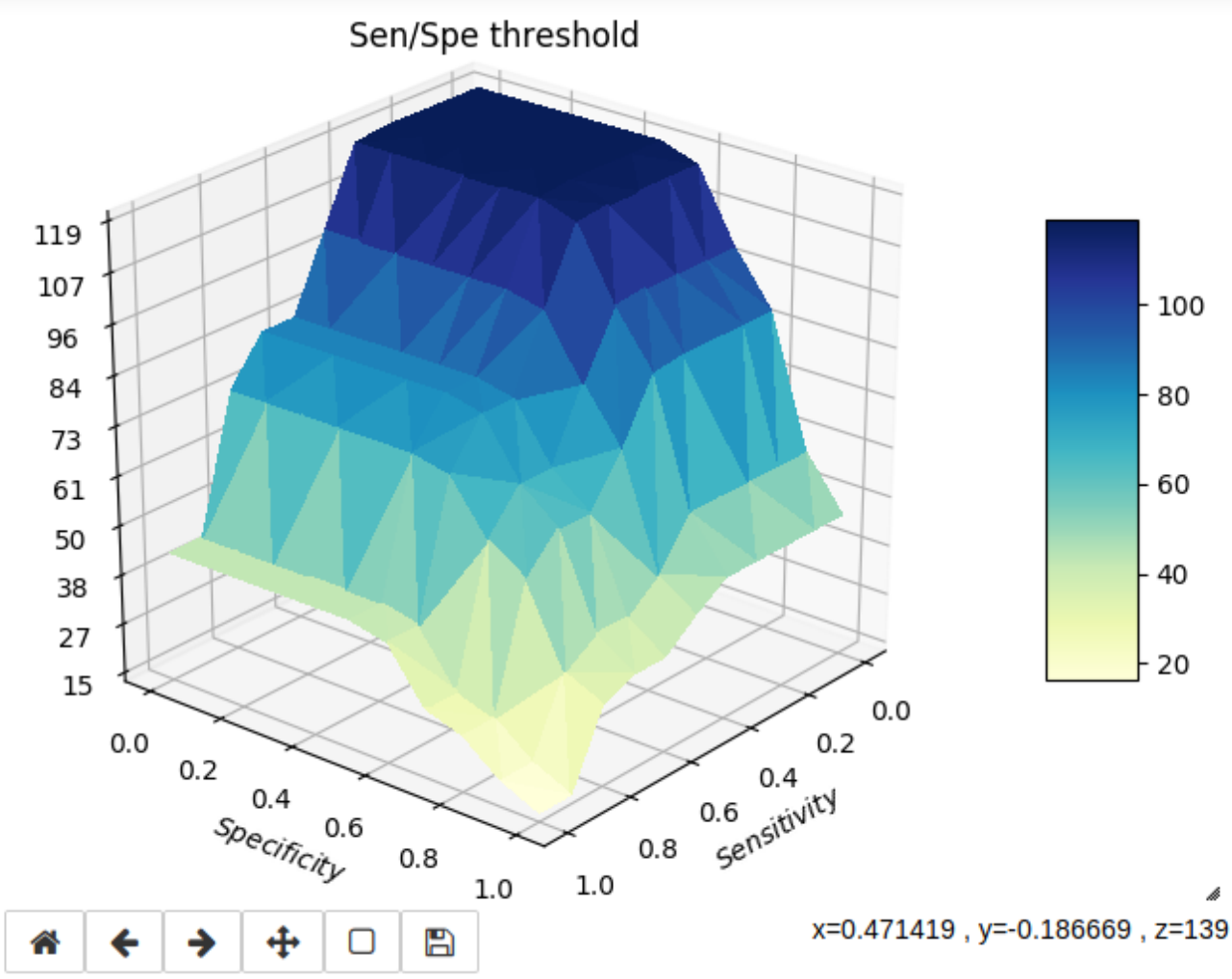
rules with 2 variables per rule.

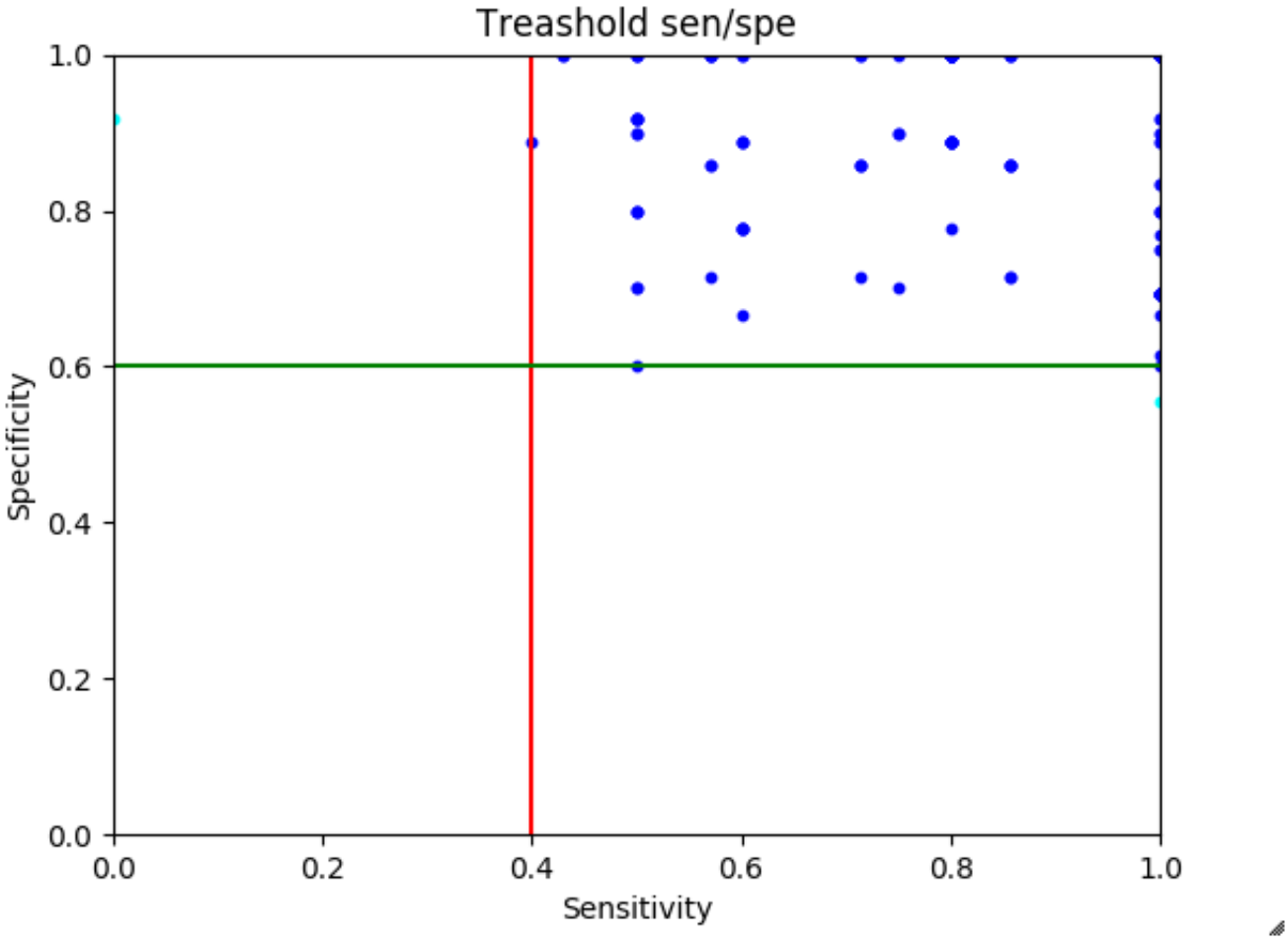


Question 6

Ideally, our first search would have been quite coarse and then refined during this step; however our initial search was quite specific since the dataset has few features. Therefore for this dataset, our “refined” search consists of 2 to 3 variables per rule and 5 to 6 rules. These are settings that already appear in the initial search, however. For the other datasets, the search will be more refined.

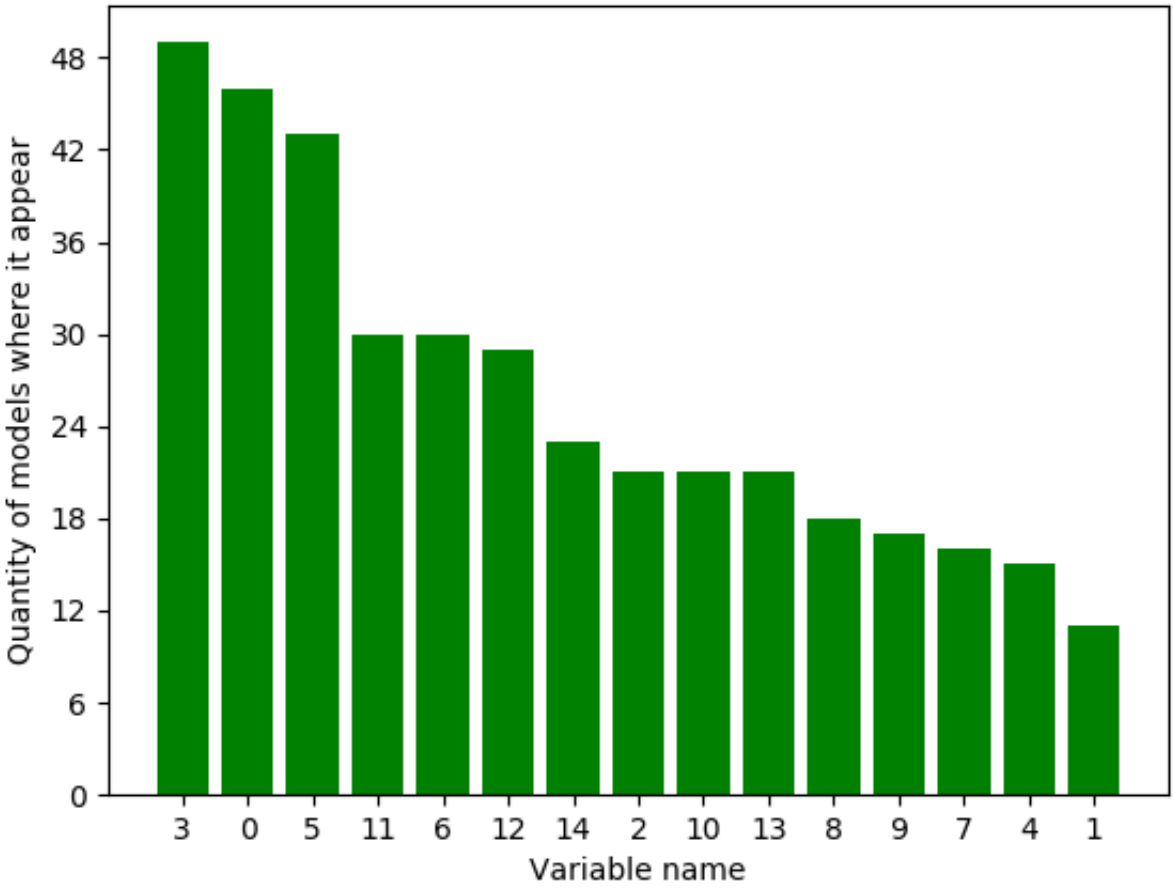
Question 7

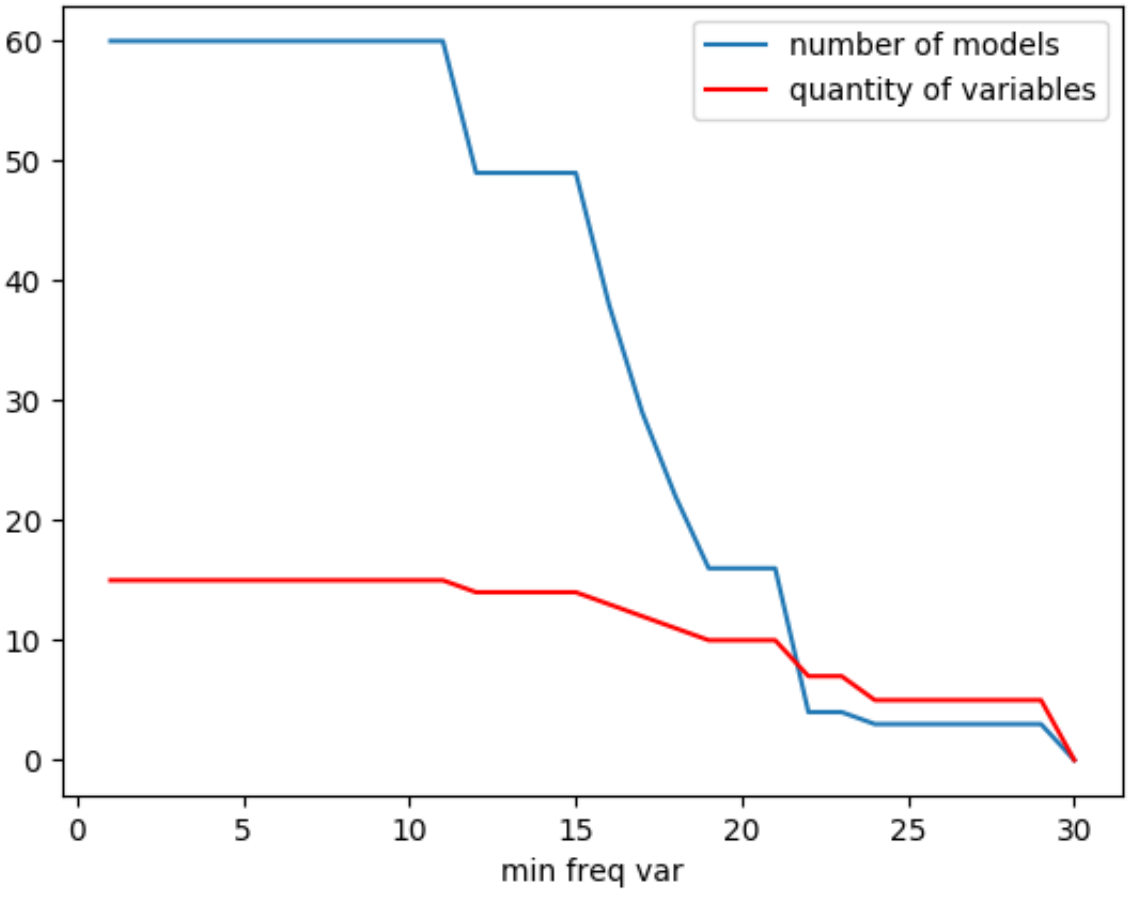

Sen/Spe threshold
119
107
96
84
73
61
50
38
27
15
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
Sensitivity
100
80
60
40
20
x=0.471419 , y=-0.186669 , z=139


Treashold sen/spe
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Sensitivity

Question 8

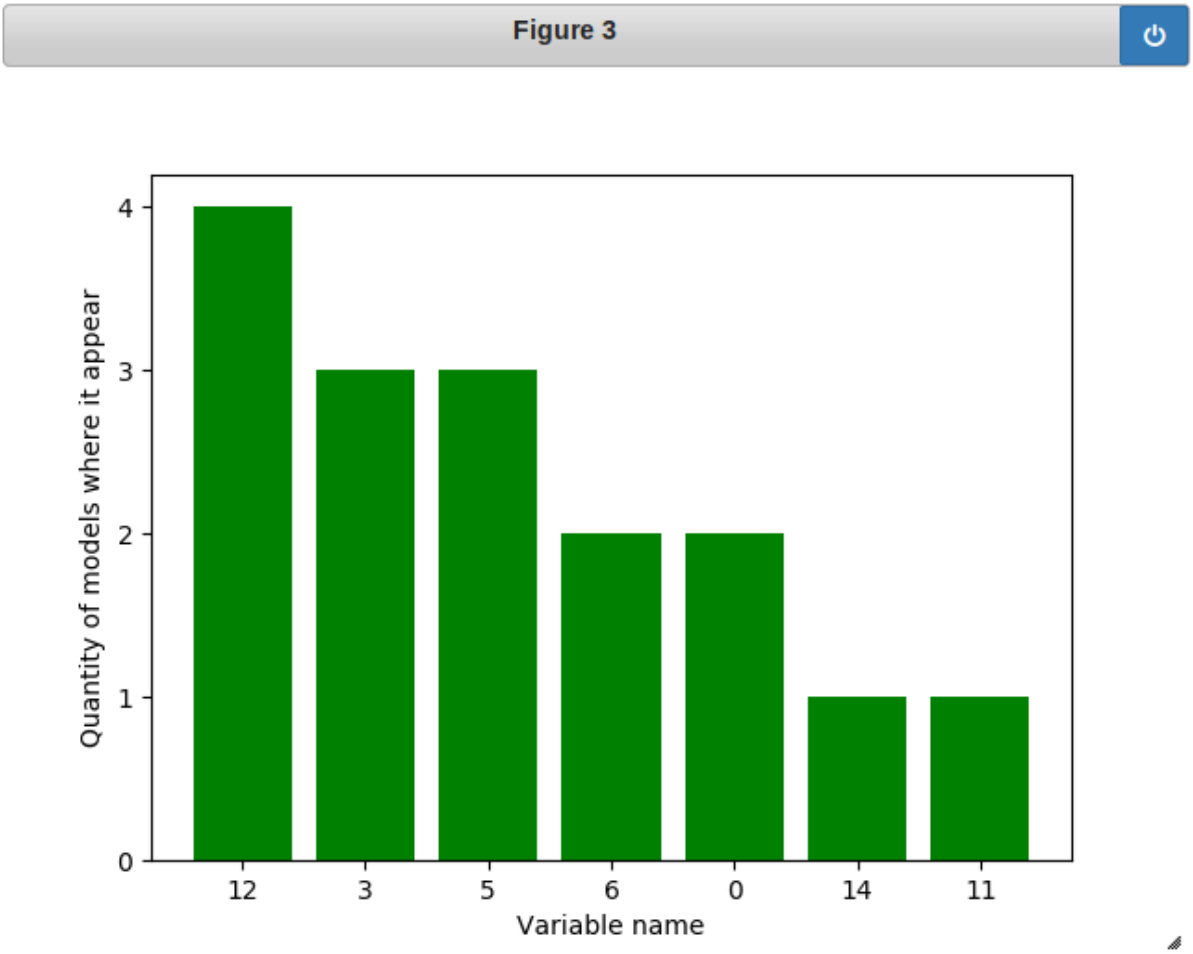
We can see from the frequency chart (first chart) that all variables appear in at least 10 models. The number of appearances seems to follow a relatively linear order; however, only the variables up to 14 (on the x-axis) are included in more than a third of all models. The threshold chart (second chart) and variable table allow us to see that with a variable frequency minimum of 21, we can reduce our 60 models to only 4 models with 7 variables (the number of variables with a higher frequency than 14 on the first chart). This gives us a small number of models that we can accurately test and use definitively.


Quantity of models where it appear
48
42
36
30
24
18
12
6
0
3
0
5
11
6
12
14
2
10
13
8
9
7
4
1
Variable name


number of models
quantity of variables
60
50
40
30
20
10
0
0
5
10
15
20
25
30
min freq var

The remaining variables are the following:

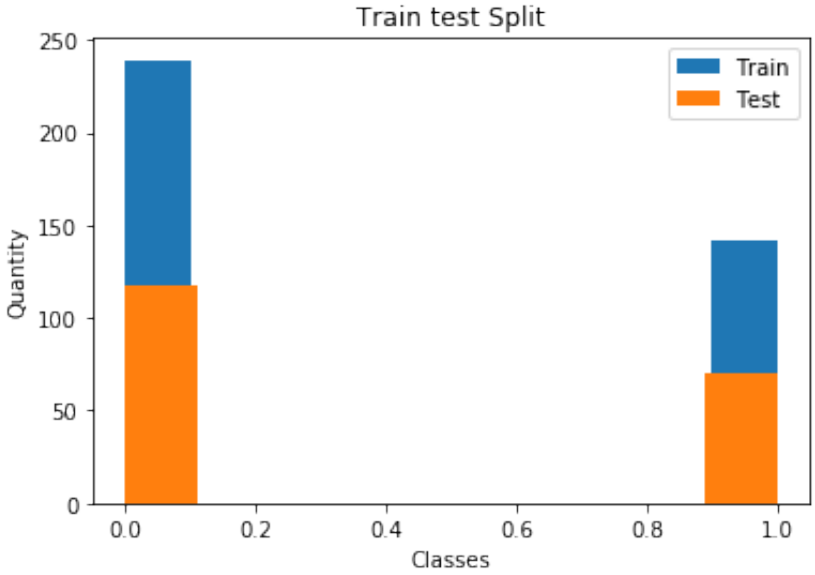
```
You have now 4 models and 7 variables
```



Dataset BCWD

Question 1

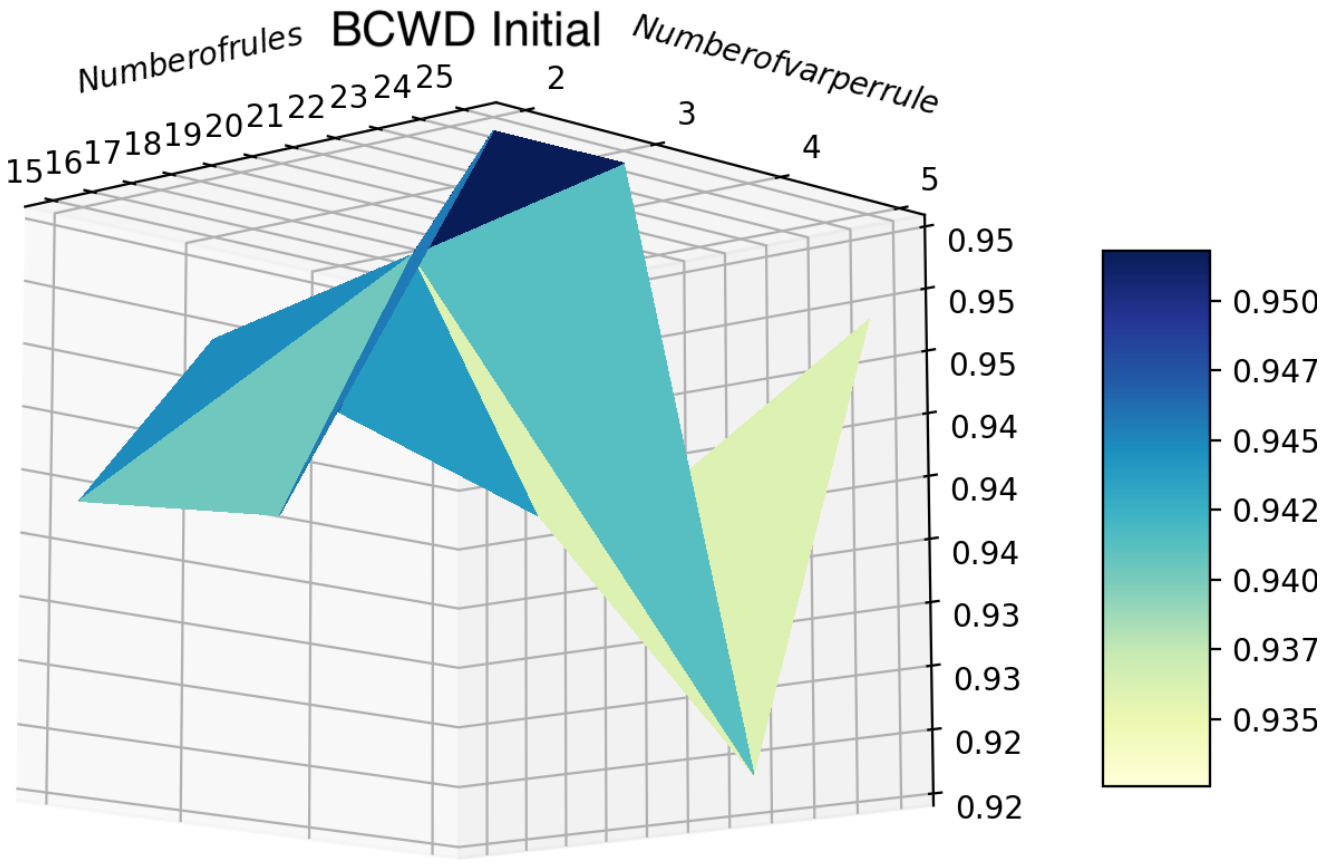
- The original dataset shape is (569, 32).
  - X shape: (569, 30)
  - y shape: (569,)
- We are using a train/test split of 66/33 for whole dataset. (381, 30) (381,) / (188, 30) (188,)
- We see that there are about two times more data points for class 0 than for class 1.



Question 4

Since this dataset has more features than the previous dataset on cancer; we decided to use a more coarse search at this step: we use values farther from each other so that for the next step we can use a more specific range based on the 3D visualization.
- rules_number_vec = [25,20,15]
- var_per_rule_vec = [5,4,3,2]

Question 5

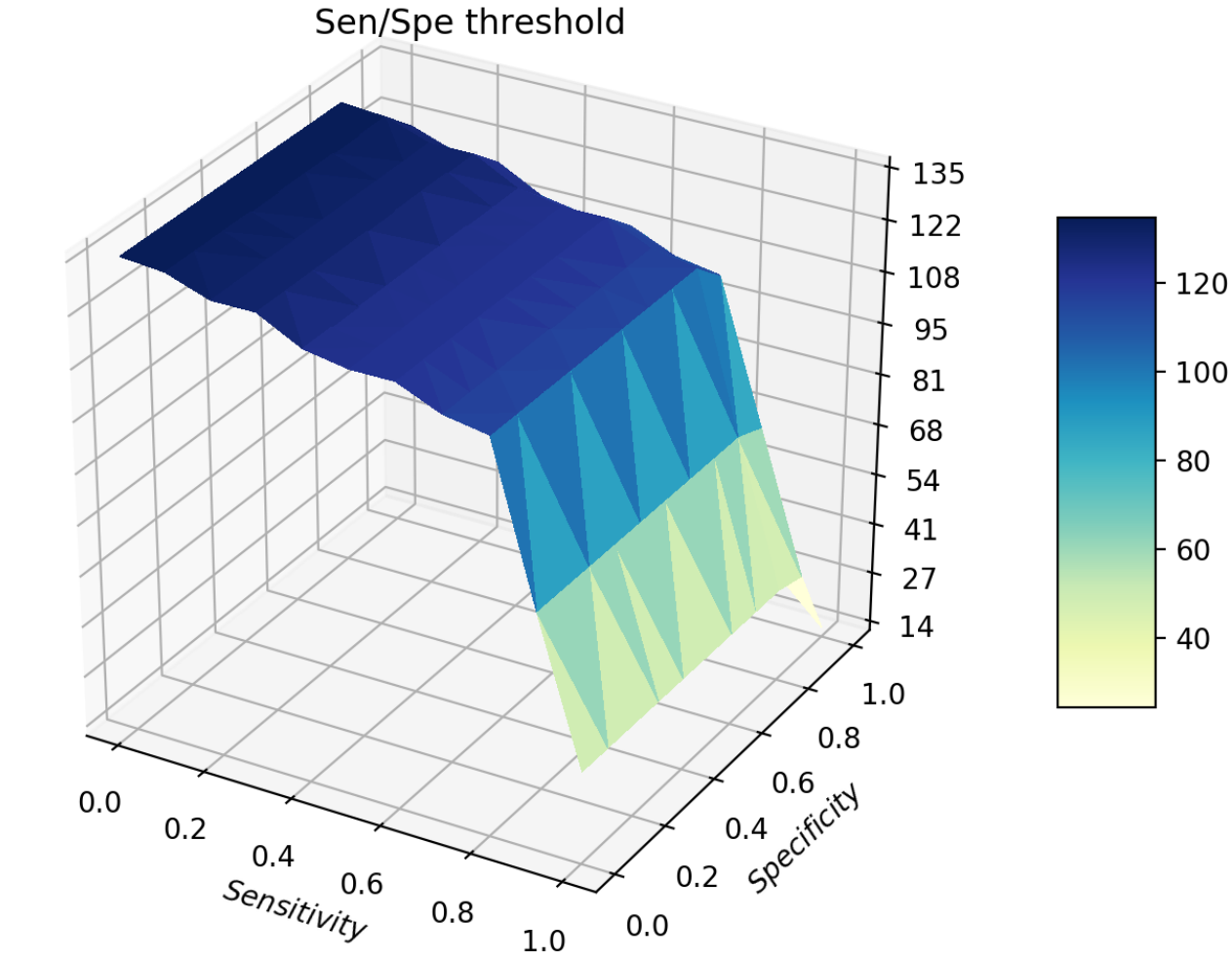


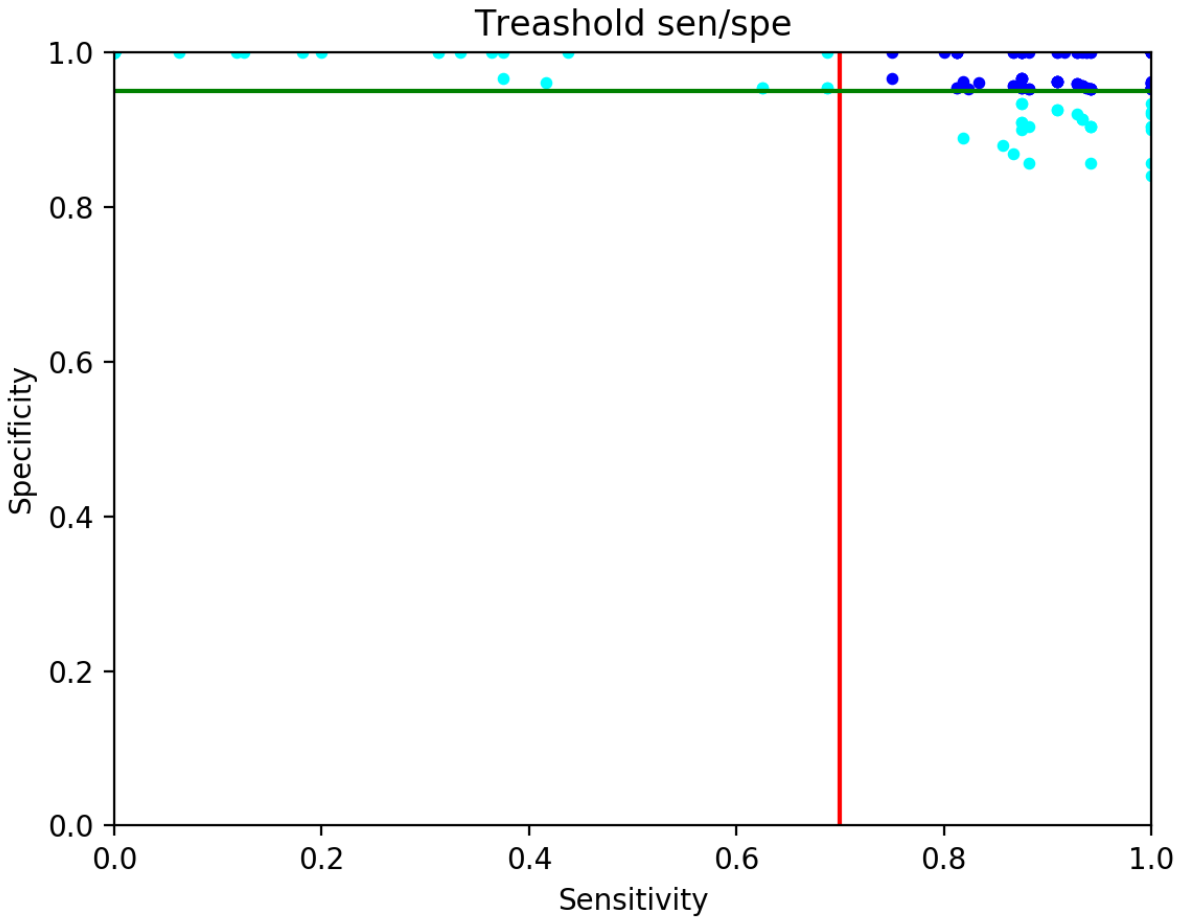
Question 6

Based on the previous outcome, we are expecting a possible increase going further the point of 25 rules in the area of 2 and 3 variables
- rules_number_vec = [2,3]
- var_per_rule_vec = [27,30]

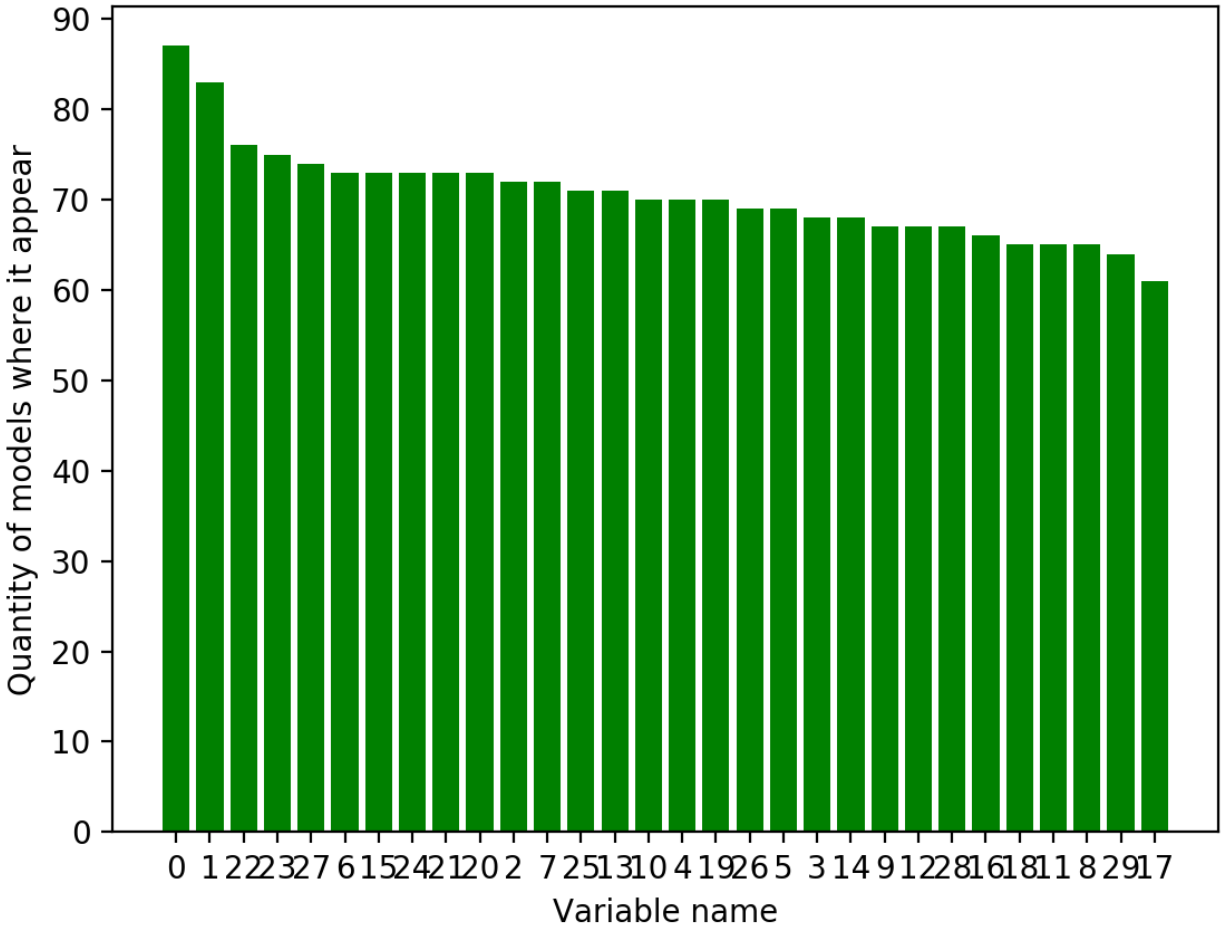
Question 7

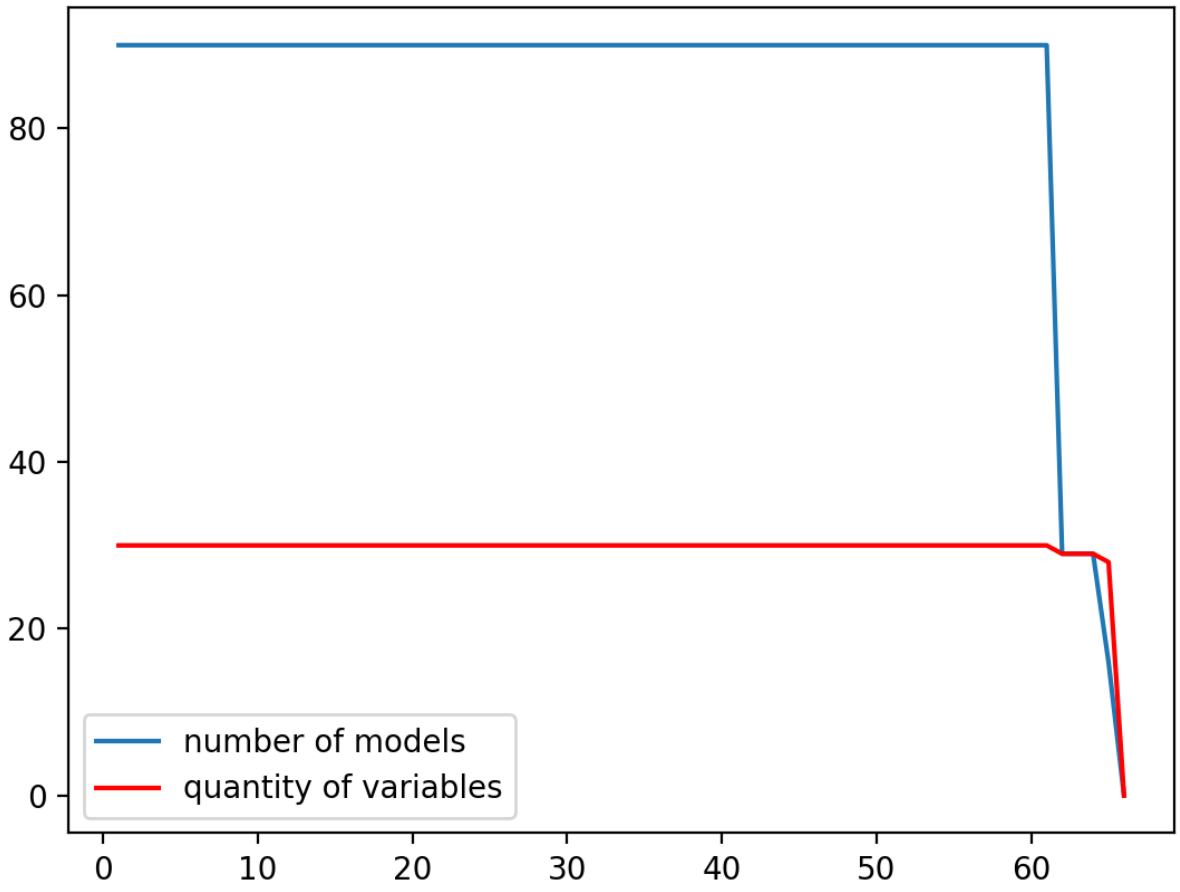
- We have in total 160 models.
- We reduced the amount to 90 models via the following selection, which was helped by the 3D plot.
  - value_sensitivity = 0.70
  - value_specificity = 0.95


Sen/Spe threshold
135
122
108
95
81
68
54
41
27
14
120
100
80
60
40
0.0
0.2
0.4
0.6
0.8
1.0
Sensitivity
Specificity


Treashold sen/spe
Specificity
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0

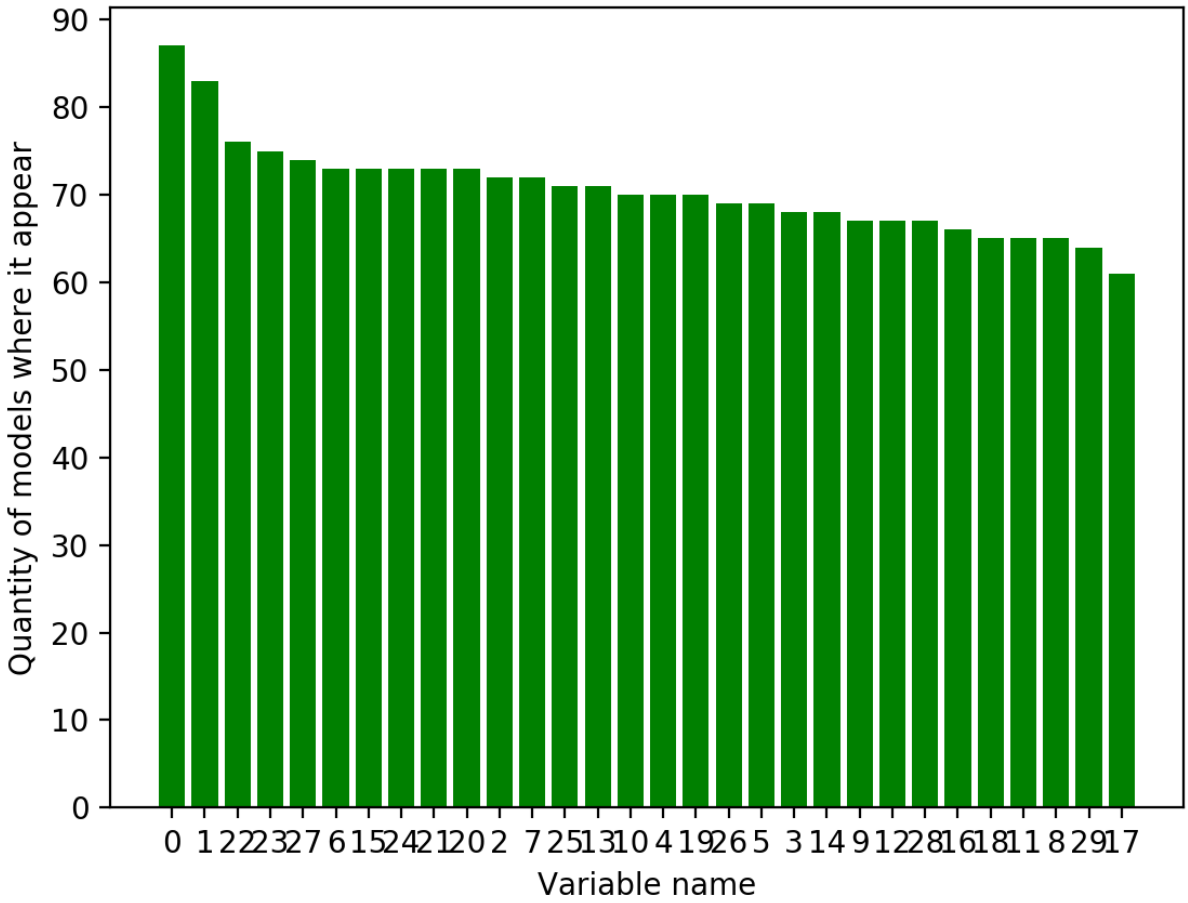
Question 8


Quantity of models where it appear
90
80
70
60
50
40
30
20
10
0
0 1 22 23 27 6 15 24 21 20 2 7 25 13 10 4 19 26 5 3 14 9 12 28 16 18 11 8 29 17
Variable name


80
60
40
20
0
number of models
quantity of variables
0
10
20
30
40
50
60

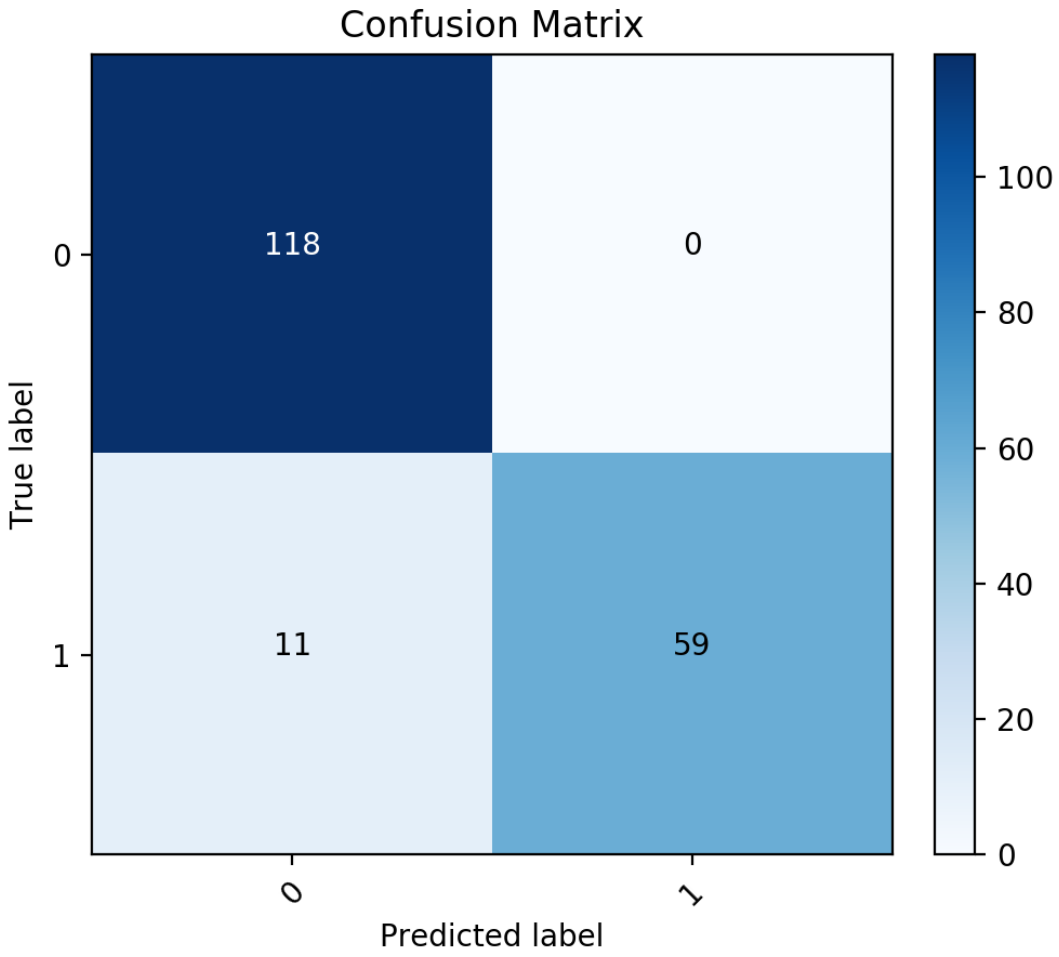
min freq var

- We reduced the frequency to 60, because it’s at this point that the number of models drop below the number of features.
- The remaining variables are the following:

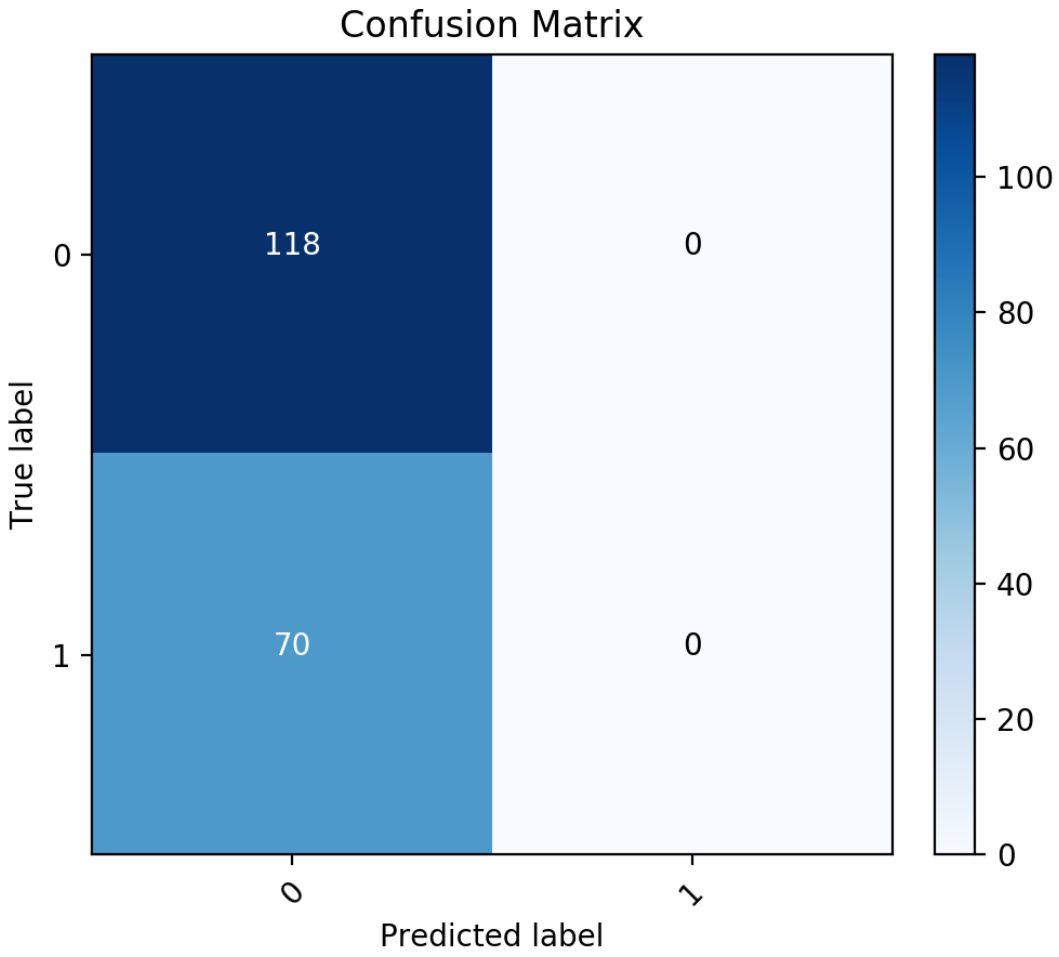


Question 9

- Smallest
  exps_lab0_p1_conf_A_CV_9_rule_15_var_per_rule_2



- Best performence
  exps_lab0_p1_conf_A_CV_9_rule_15_var_per_rule_2



- Average

TODO:

we under estimated the time to build the models and even using dedicated machine we did not manage correctla the time at disposal to finish the following parts.

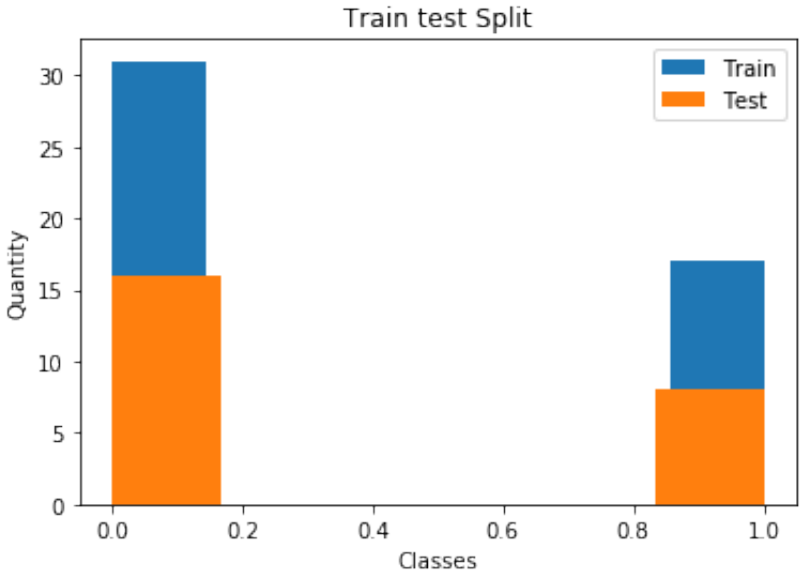
Question 10

Dataset GOLUB

Question 1

- The original dataset shapes:
  - train set is (38, 7129) (38,)
  - test set is (34, 7129) (34,)
  - The default dataset train/test split is 53/47.

- We managed to force a train/test split of 66/33 for whole dataset. (48, 7129) (48,) (24, 7129) (24,)
- We see that there are about two times more data points for class 0 than for class 1.

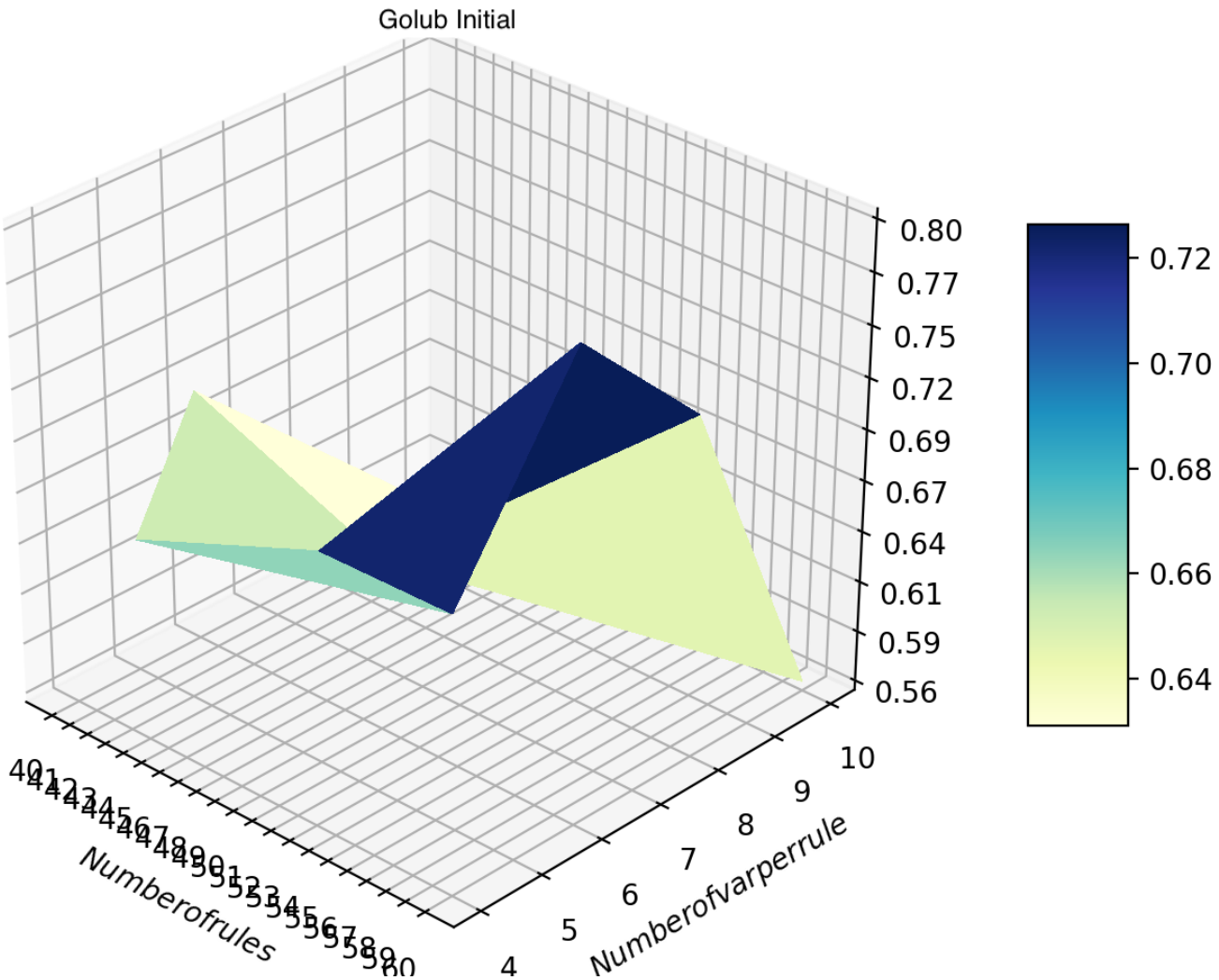


Question 4

- Problems we were faced to with this dataset.
  - Our errors were while building the models with **ValueError: not enough values to unpack (expected 4, got 1)**
- Choice
  - For the same reason that for BCWD, we are using well spread values for rules.
  - Based on our previous experience, since the Golub dataset has a huge amount of features, we started with a higher number of rules.
  - rules_number_vec = [60,50,40,30,20]
  - var_per_rule_vec = [4,6,8,10]

Question 5

- Problems we were faced to with this dataset.
  - Errors came from specific combinations of rules and vars (60, 8) and (60, 6), which had a fn+tp = 0.
  - Our solution was to filter and combine the outcomes manually from multiple runs.
  - Since we had to run this experiment multiple times, the time spent on building models is insane…
- Outcome
  - Based on our previous experience, since the Golub dataset has a massive amount of features, we started with a higher number of rules.
  - As expected we found out that highest number of rules, 60 in this case, is providing the best results. Combined with 6 as the number of variables for this rule.



Question 6

- Choice
  - Based on the previous outcome, we are expecting an increase going further the point of 60 rules in the area of 5,6,7 variables.
  - rules_number_vec = [65,70]
  - var_per_rule_vec = [5,6,7]

TODO:

we under estimated the time to build the models and even using dedicated machine we did not manage correctla the time at disposal to finish the following parts.

Question 7

Question 8

Question 9

Question 10