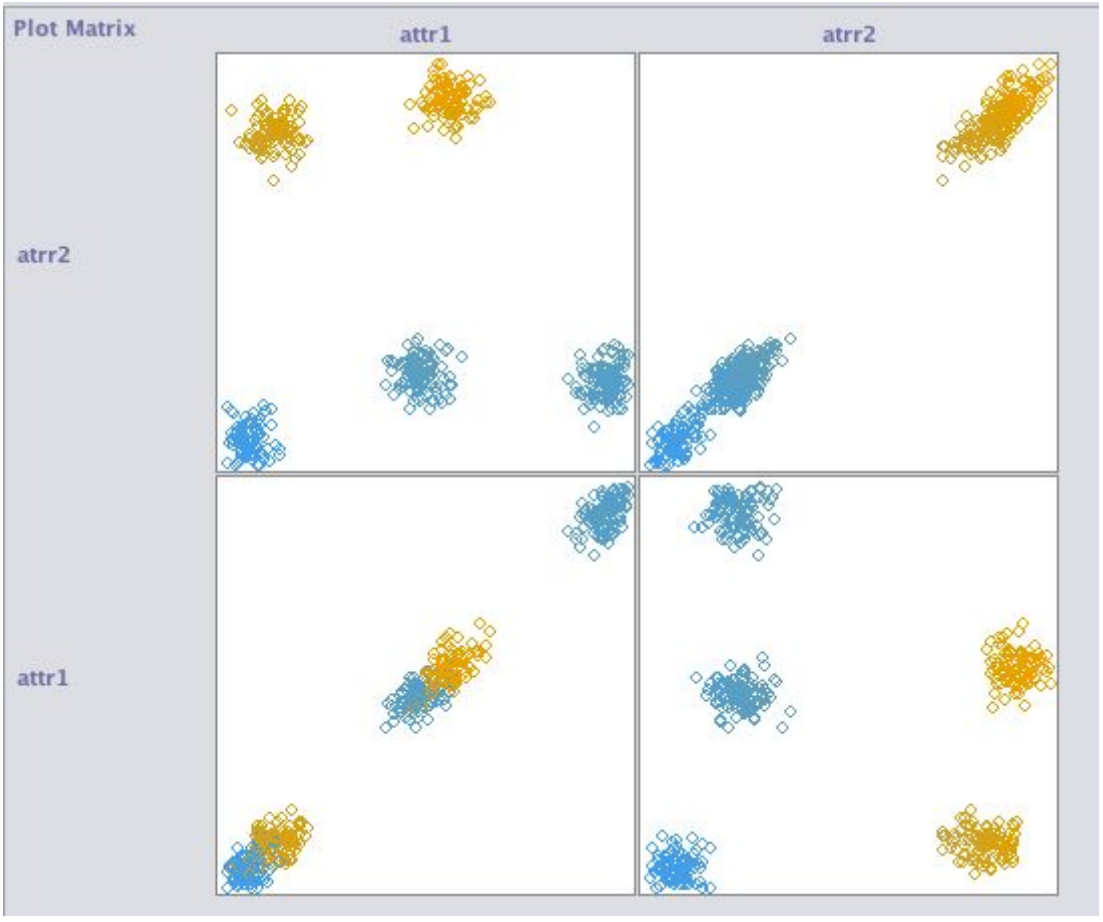
Assignment 03

Team 06

Qinyang Wu, st174540@stud.Uni-Stuttgart.de, 3519174
Huicheng Qian, st169665@stud.uni-stuttgart.de, 3443114
Kuang-Yu Li, st169971@stud.uni-stuttgart.de, 3440829

Task 1

S1.1

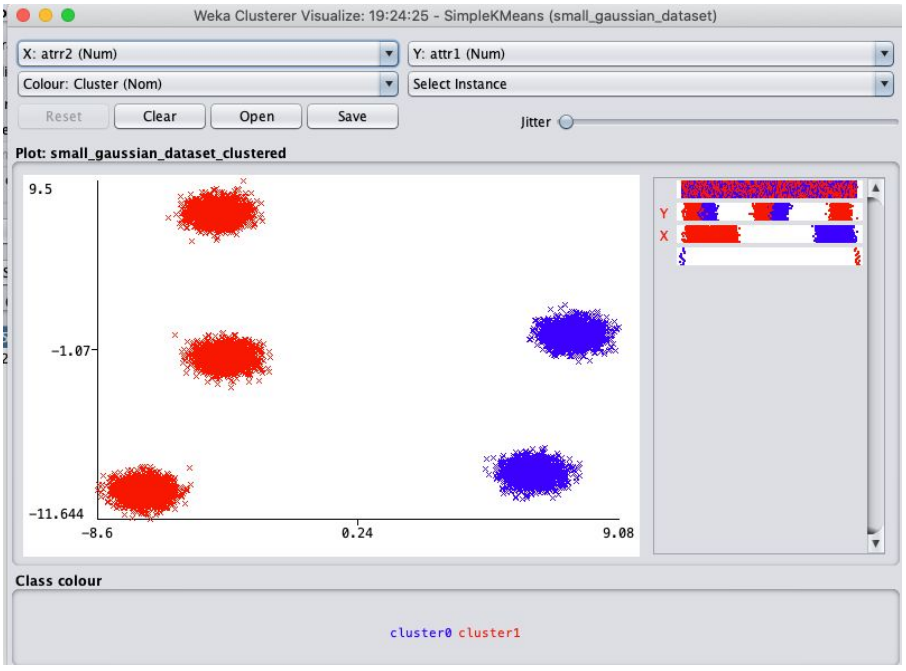


Q1.1

From the attr1 vs attr2, there are 5 clusters in the dataset

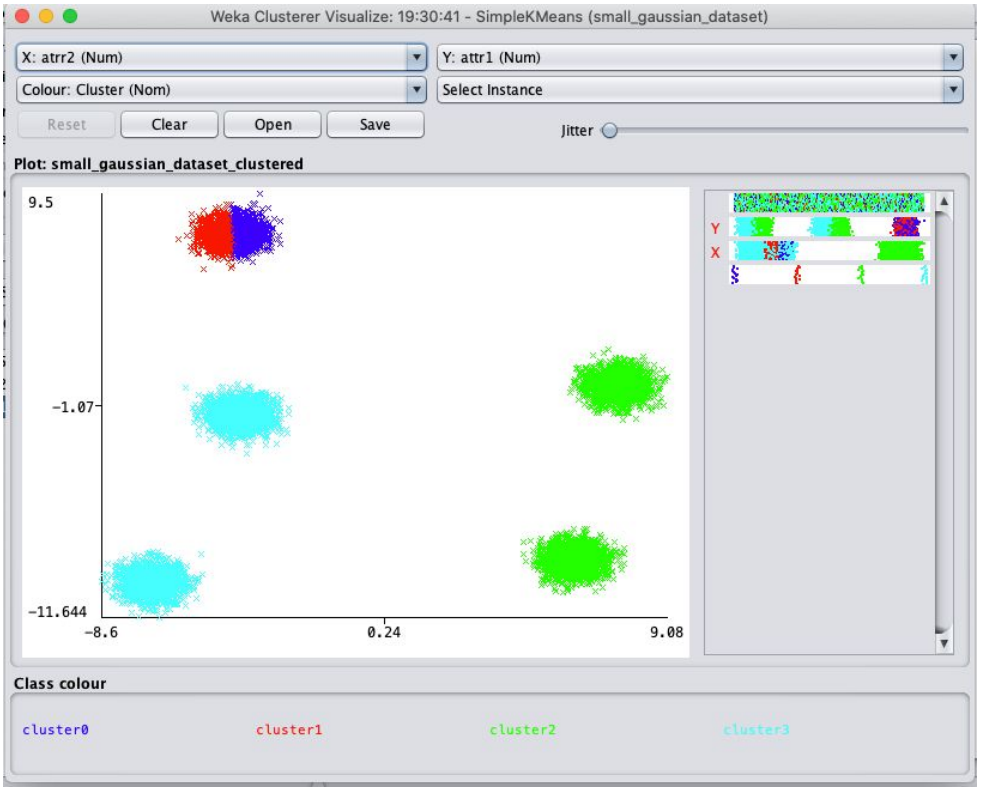
S1.2

K = 2

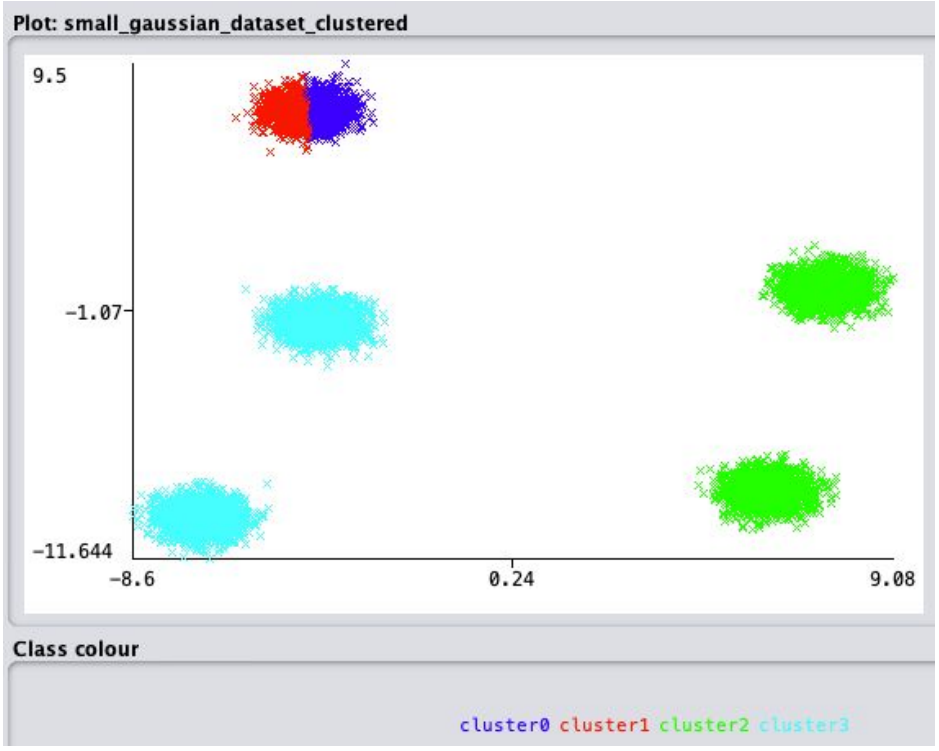


S1.3

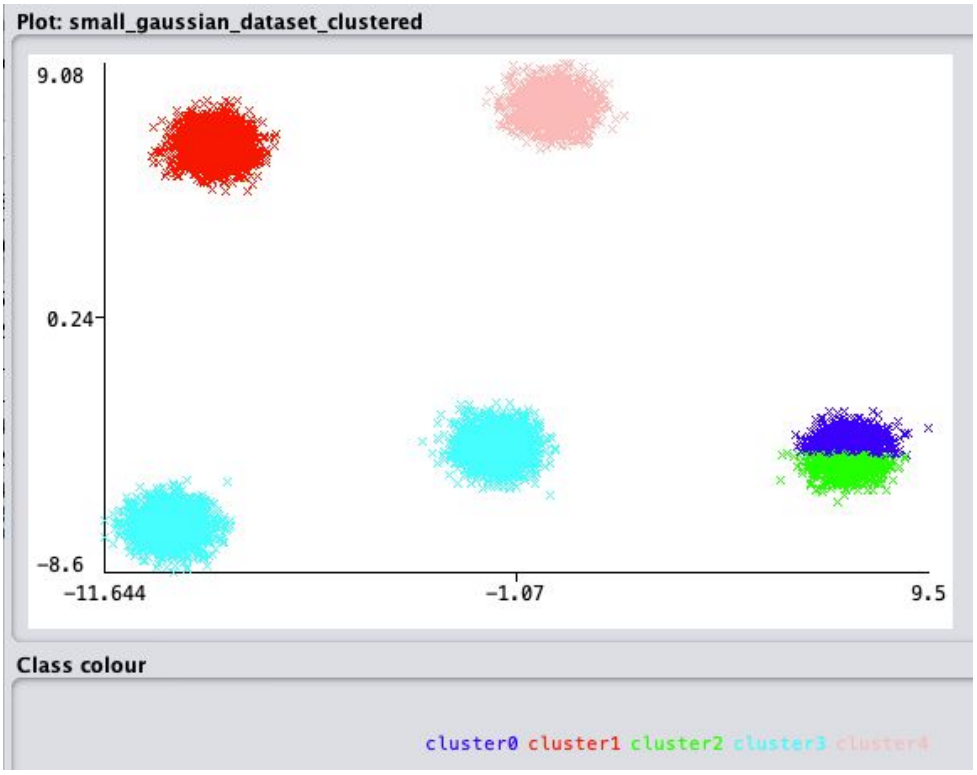
K = 3



K = 4



K = 5

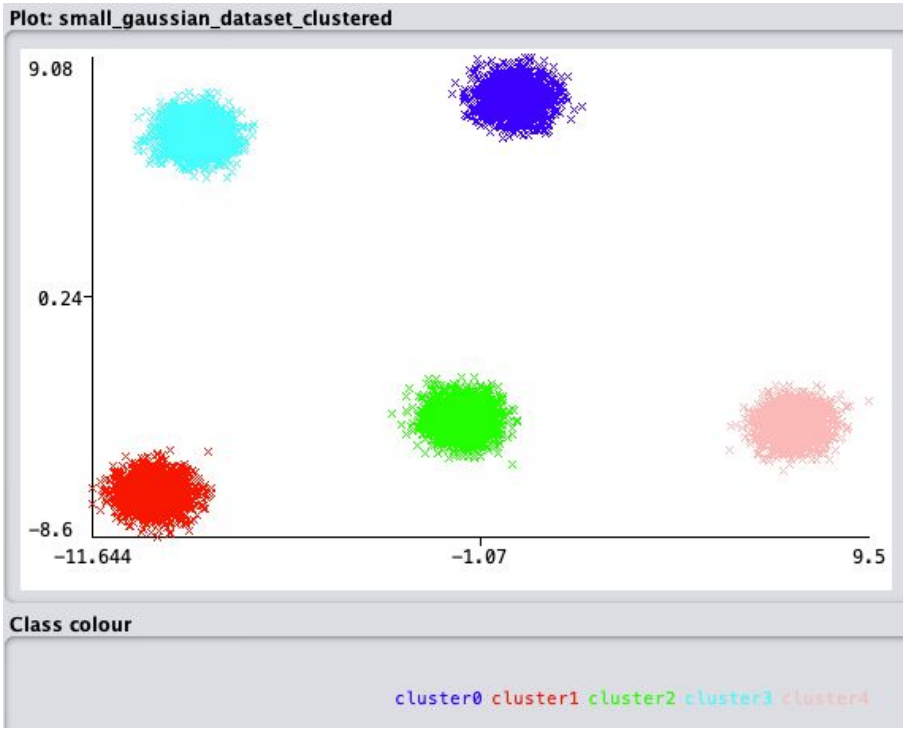


Q1.2

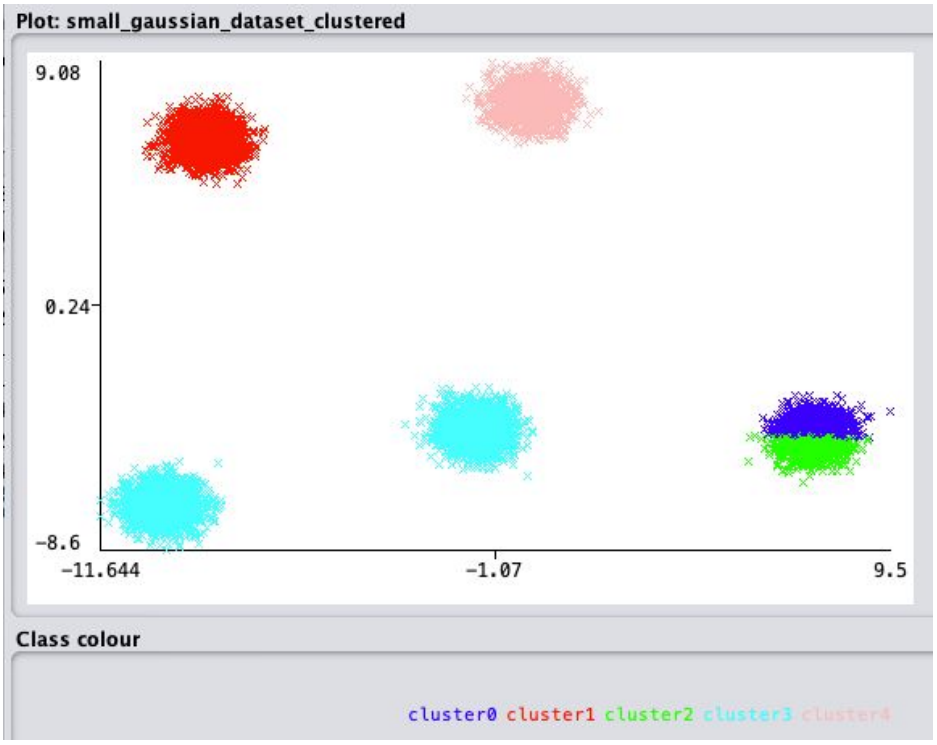
with k increasing, the group of clusters gets increased. KMean didn't detect the actual clusters of the dataset

S1.4

seed = 1



seed = 10

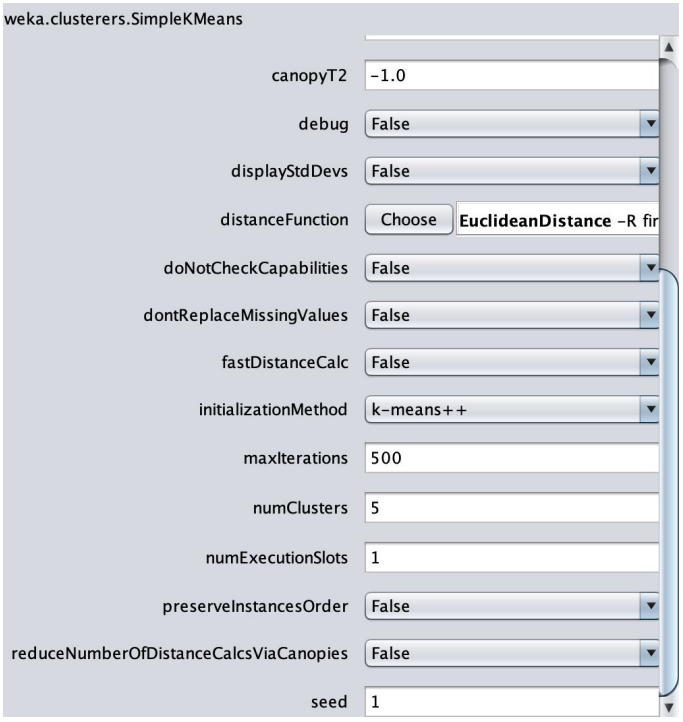


Q1.3

Yes, only seed = 1 detects the correct clusters. The reason is that the K means algorithm strongly depends on the selection of initial points. In this case, random seed selection seed = 1 happens to pick points from a separate group.

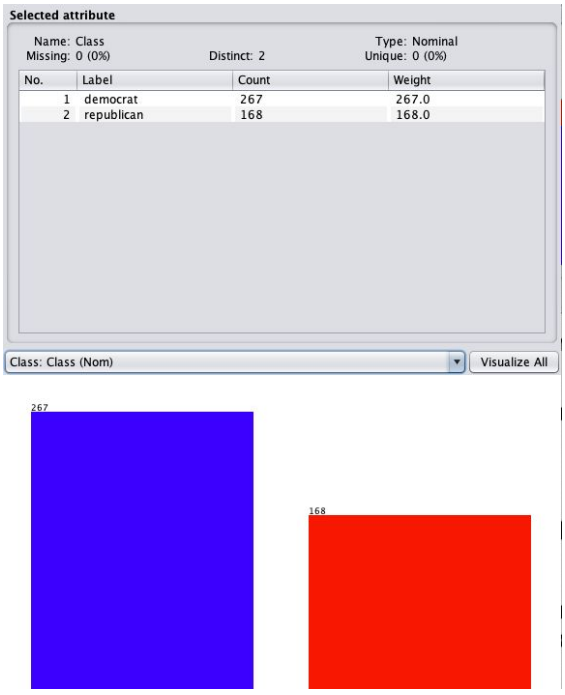
Q1.4

K = 5, and seed = 1



Task 2

S2.1



Q2.1

Values are Democrat, Republican.
By looking at the dataset and attribute, we guess it is the preference of choices from voters belonging to two parties: Democrat and Republican.

Q2.2

Yes, we can use attributes to predict the voters' party.
We cannot use the dataset for association with the class attribute removed

S2.2

```
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.45 (196 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1

Best rules found:

 1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219    <conf:(1)> lift:(1.63) lev:(0.19) [84] conv:(84.58)
 2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198    <conf:(1)> lift:(1.63) lev:(0.18) [76] conv:(76.47)
 3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210    <conf:(1)> lift:(1.62) lev:(0.19) [80] conv:(40.74)
 4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201    <conf:(1)> lift:(1.62) lev:(0.18) [77] conv:(39.01)
 5. physician-fee-freeze=n 247 ==> Class=democrat 245    <conf:(0.99)> lift:(1.62) lev:(0.21) [93] conv:(31.8)
 6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197    <conf:(0.98)> lift:(1.77) lev:(0.2) [85] conv:(22.18)
 7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204    <conf:(0.98)> lift:(1.76) lev:(0.2) [88] conv:(18.46)
 8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198    <conf:(0.98)> lift:(1.72) lev:(0.19) [82] conv:(14.62)
 9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197    <conf:(0.97)> lift:(1.57) lev:(0.17) [71] conv:(9.85)
10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210    <conf:(0.96)> lift:(1.7) lev:(0.2) [86] conv:(10.47)
```

Q2.3

physician-fee-free=“n”	247
physician-fee-free=“n” and class= “democrat”	245

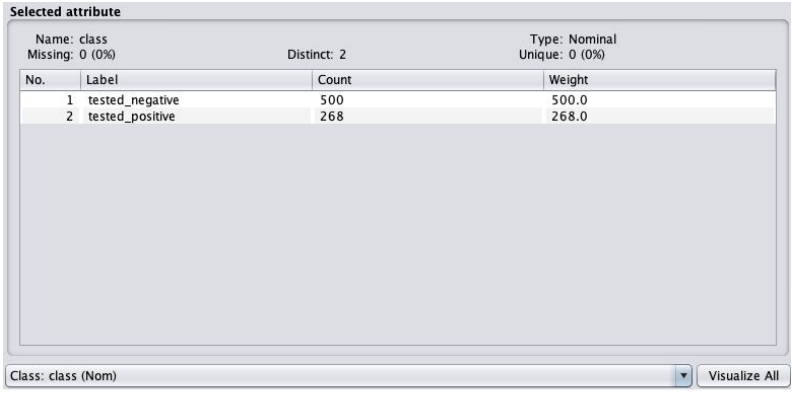
Q2.4

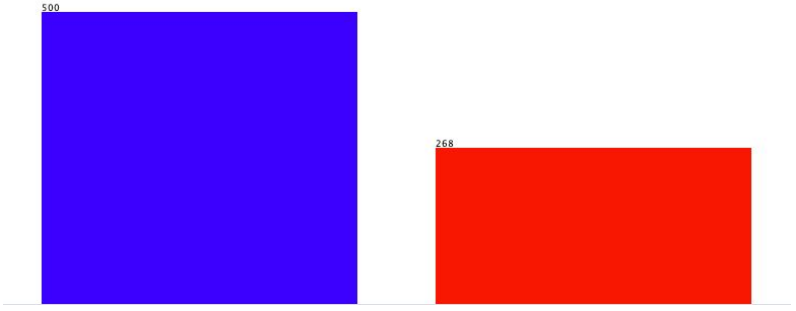
1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 <conf:(1)> lift:(1.63) lev:(0.19) [84] conv:(84.58)
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 <conf:(1)> lift:(1.63) lev:(0.18) [76] conv:(76.47)
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 <conf:(1)> lift:(1.62) lev:(0.19) [80] conv:(40.74)
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 <conf:(1)> lift:(1.62) lev:(0.18) [77] conv:(39.01)
5. physician-fee-freeze=n 247 ==> Class=democrat 245 <conf:(0.99)> lift:(1.62) lev:(0.21) [93] conv:(31.8)

Confidence c(L,R) = s(L,R) / s(L). Confidence indicates the probability of correct classification with the given attributes. Confidence over 0.99 means that correct classification the given rules are over 99%

Task 3

S3.1

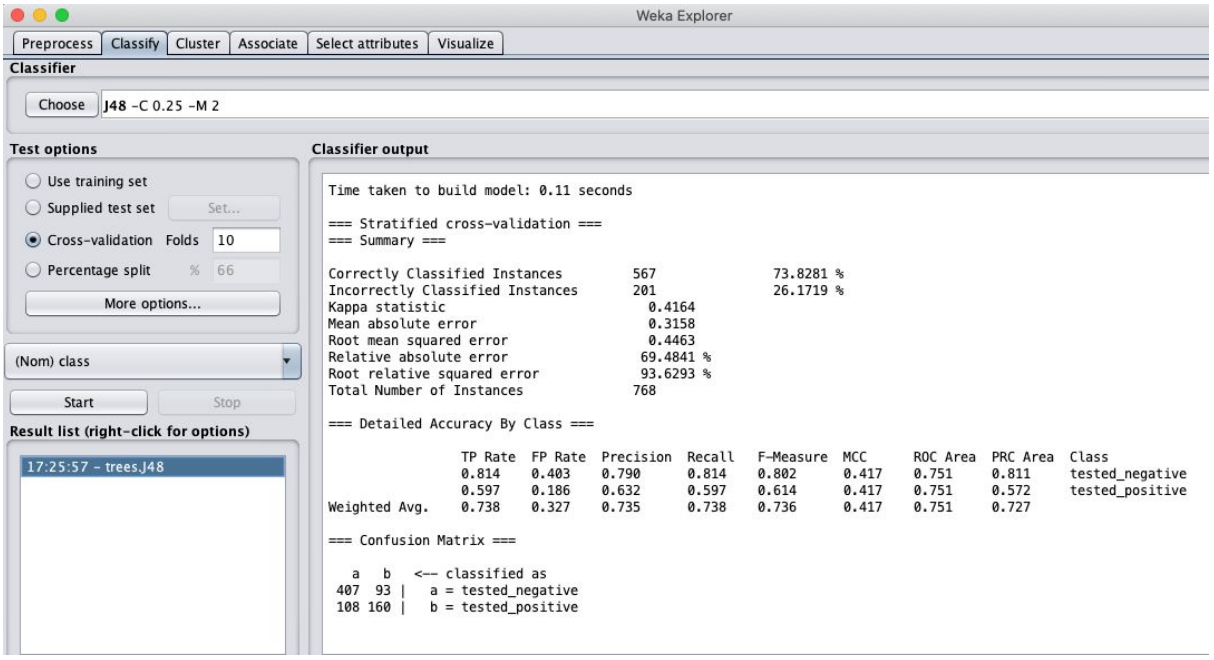




Q3.1

There are two types of value for class attributes: tested positive and tested negative.
tested positive: 268/768 = 34.9%
tested negative: 500/768= 65.1%
Based on other attributes provided, we think that the data comes from the patients who got tested for diabetes. The attributes are the physical situation of each patient, such as skin, age, and so on.

S3.2



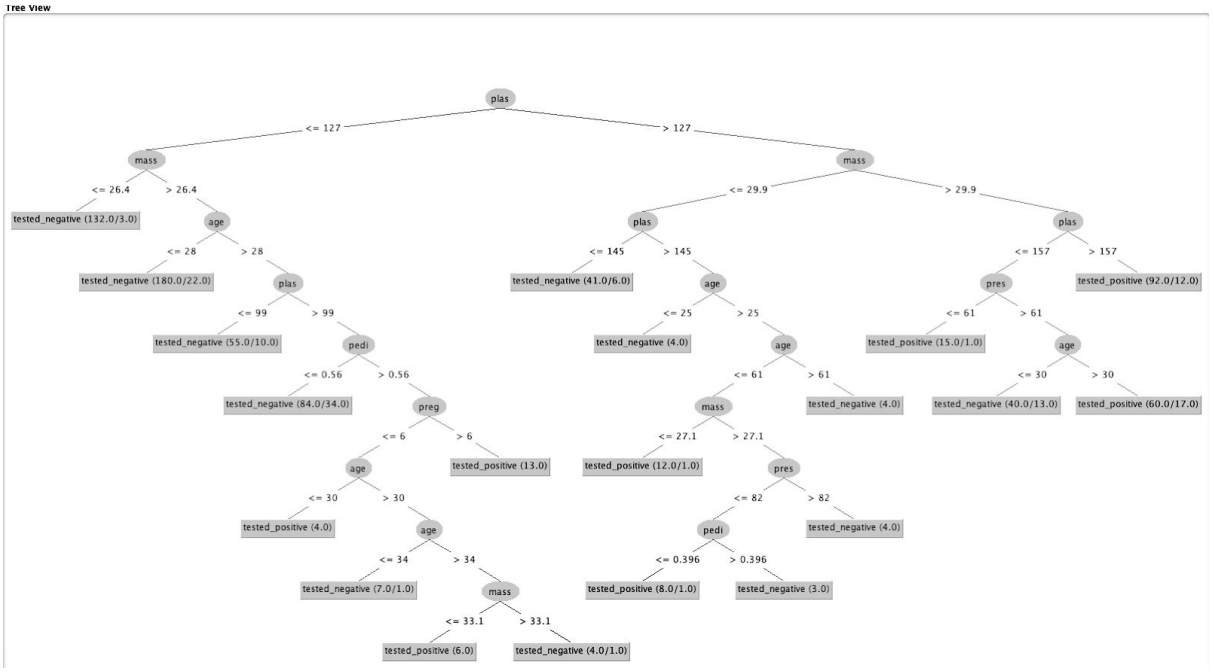
Q3.2

TP: true positive, FP: false positive, FN: false negative

Precision	0.735	TP/(TP+FP)
Recall	0.738	TP/(TP + FN)
F-Measure	0.736	the harmonic mean of Precision and Recall 2 * (Precision * Recall) / (Precision + Recall)

Q3.3

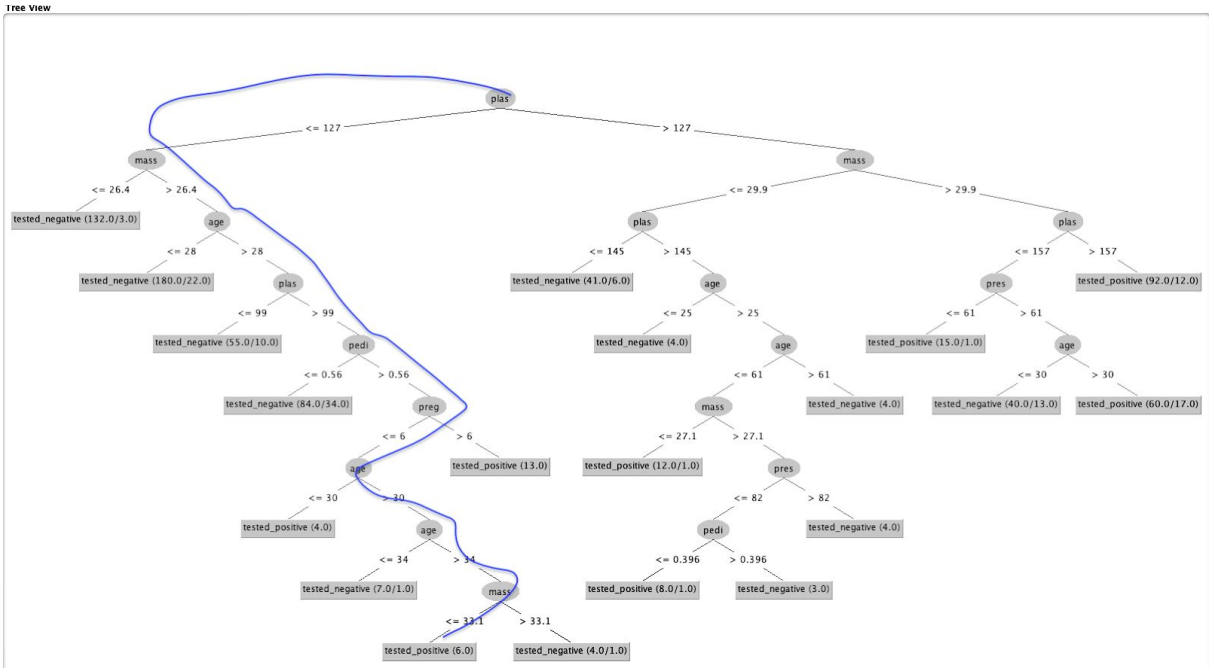
The classes are in the leaf of the tree
The depth of the tree is 10: including leaf



Q3.4

Positive rule:
IF age > 30 and pres > 61 and plas <= 157 and mass > 29.9 THEN tested_positive
Negative rule:
IF plas <= 127 and mass >= 26.4 THEN tested_negative

Q3.5



Traveling along the tree, in the end, it reached tested_positive. Which is not the correct class