Classification Exercises

Exercise 1

1. Load the ionosphere dataset
2. Generate a decision tree (using j48) without pruning and extract the accuracy on the test set (70% splitting for the training set)
3. Find the best parameter for the pruning in order to improve, if possible, the accuracy on the test set (check unpruned, c=0.3, c=0.2, c=0.1).

```
Test mode:    split 70.0% train, remainder test

=== Classifier model (full training set) ===

J48 unpruned tree
------------------

a05 <= 0.0409: b (67.0)
a05 > 0.0409
|   a01 <= 0: b (19.0)
|   a01 > 0
|   |   a08 <= -0.67273
|   |   |   a28 <= -0.21793
|   |   |   |   a06 <= -1: b (2.0)
|   |   |   |   a06 > -1: g (4.0)
|   |   |   a28 > -0.21793: b (11.0)
|   |   a08 > -0.67273
|   |   |   a03 <= 0.26667
|   |   |   |   a03 <= 0.10135: b (9.0)
|   |   |   |   a03 > 0.10135: g (4.0)
|   |   |   a03 > 0.26667
|   |   |   |   a16 <= 0.86284
|   |   |   |   |   a21 <= 0.67213
|   |   |   |   |   |   a19 <= 0.79113
|   |   |   |   |   |   |   a06 <= 0.21908
|   |   |   |   |   |   |   |   a17 <= 0.19672
|   |   |   |   |   |   |   |   |   a07 <= 0.21572: g (4.0)
|   |   |   |   |   |   |   |   |   a07 > 0.21572: b (5.0)
|   |   |   |   |   |   |   |   a17 > 0.19672
|   |   |   |   |   |   |   |   |   a21 <= 0.57399: g (36.0)
|   |   |   |   |   |   |   |   |   a21 > 0.57399
|   |   |   |   |   |   |   |   |   |   a10 <= 0.09237: g (10.0/1.0)
|   |   |   |   |   |   |   |   |   |   a10 > 0.09237: b (2.0)
|   |   |   |   |   |   |   a06 > 0.21908: g (57.0)
|   |   |   |   |   |   a19 > 0.79113
|   |   |   |   |   |   |   a04 <= 0.04528: b (4.0)
|   |   |   |   |   |   |   a04 > 0.04528: g (2.0)
|   |   |   |   |   a21 > 0.67213: g (103.0)
|   |   |   |   a16 > 0.86284
|   |   |   |   |   a27 <= 0.36547: g (6.0)
|   |   |   |   |   a27 > 0.36547: b (6.0)

Number of Leaves  :     18

Size of the tree :      35
```

```
=== Summary ===

Correctly Classified Instances          84               80      %
Incorrectly Classified Instances        21               20      %
Kappa statistic                          0.5914
Mean absolute error                      0.1914
Root mean squared error                  0.4039
Relative absolute error                 39.7584 %
Root relative squared error             78.7771 %
Total Number of Instances              105

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,723    0,138    0,810      0,723    0,764      0,594    0,880     0,829     b
                 0,862    0,277    0,794      0,862    0,826      0,594    0,880     0,867     g
Weighted Avg.    0,800    0,215    0,801      0,800    0,799      0,594    0,880     0,850

=== Confusion Matrix ===

  a  b   <-- classified as
 34 13 |  a = b
  8 50 |  b = g
```

Confidence Factor	Accuracy	Number of Leaves	Size of the tree
0.3	80 %	18	35
0.2	84.7619 %	16	31
0.1	84.7619 %	10	19

Exercise 2

- Perform the classification by using the following classifiers (default parameters) and the iris dataset (66% spit):
    - Jrip (rules)
    - KNN(lazy)
    - Naive Bayes (Bayes)

Which is the most accurate classifier on the test set?

Classification Algorithm	Accuracy	Confusion Matrix
Jrip	92.1569 %	a b c <-- classified as 15 0 0 | a = Iris-setosa 2 17 0 | b = Iris-versicolor 0 2 15 | c = Iris-virginica
KNN	96.0784 %	a b c <-- classified as 15 0 0 | a = Iris-setosa 0 19 0 | b = Iris-versicolor 0 2 15 | c = Iris-virginica
Naive Bayes	94.1176 %	a b c <-- classified as 15 0 0 | a = Iris-setosa 0 18 1 | b = Iris-versicolor 0 2 15 | c = Iris-virginica

Exercise 3

- Perform the classification by using the following classifiers (default parameters) and the Pima Diabetes and and Hepatitis dataset with a 10- fold cross validation:
  - Jrip (rules)
  - J48
  - KNN(lazy)
  - Naive Bayes (Bayes)
  - Random Forests
- Prepare a table resuming the results. Which is the best classifier?

Pima Diabetes					
Algorithm	**Accuracy**	**Precision Negative Class**	**Recall Negative Class**	**Precision Positive Class**	**Recall Positive Class**
Jrip	76.0417 %	0,793	0,856	0,684	0,582
J48	73.8281 %	0,790	0,814	0,632	0,597
KNN	70.1823 %	0,759	0,794	0,580	0,530
Naive Bayes	76.3021 %	0,802	0,844	0,678	0,612
Random Forests	75.7813 %	0,801	0,836	0,667	0,612

Hepatitis					
Algorithm	**Accuracy**	**Precision Negative Class**	**Recall Negative Class**	**Precision Positive Class**	**Recall Positive Class**
Jrip	81.2903 %	0,862	0,911	0,560	0,438
J48	76.129 %	0,831	0,878	0,400	0,313
KNN	80.6452 %	0,891	0,862	0,528	0,594
Naive Bayes	82.5806 %	0,907	0,870	0,568	0,656
Random Forests	83.2258 %	0,876	0,919	0,615	0,500

Algorithm	Pima Diabetes Accuracy	Hepatitis Accuracy
Jrip	76.0417 %	81.2903 %
J48	73.8281 %	76.129 %
KNN	70.1823 %	80.6452 %
Naive Bayes	76.3021 %	82.5806 %
Random Forests	75.7813 %	83.2258 %

Algorithm	Accuracy	Precision Class1	Recall Class1	Precision Class2	Recall Class2	Tree Dimension	Frequent Attributes
C45	71.224 %	0,784	0,770	0,585	0,604	39	
CFSsubestEval + BestFirst	73.6979 %	0,785	0,820	0,634	0,582	29	Plas,mass,pedi,age,class
WrappedC45 + BestFirst	73.8281 %	0,784	0,826	0,639	0,575	768	Plas,pres,mass,age,class
InfoGain + Ranking	73.8281 %	0,771	0,850	0,654	0,530	768	Plas,mass,age,insu,class

Exercise 4

Using the knowledge flow environment, evaluate the ***classification accuracy*** of a C45 classifier on the ionosphere dataset with a 2x10-fold cross validation.

Then, evaluate both accuracy and tree complexity when performing the classification after performing the following attribute selection schemes:

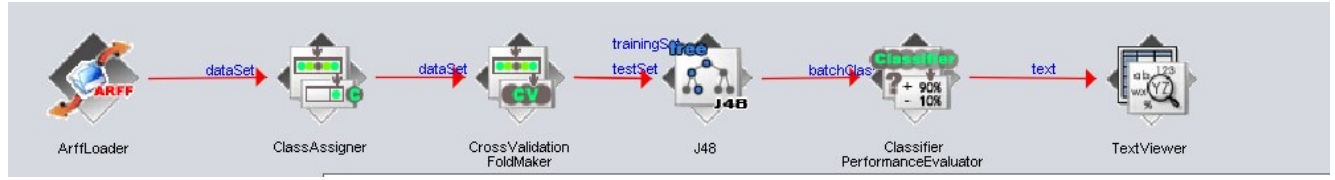
- CFSsubsetEval+BestFirst
- Wrapped Naïve Bayes+BestFirst
- InfoGain+Ranking with a threshold of 0.3

Show the average accuracy and the number of selected features in a table.

Moreover, for each attribute selection method, identify the set of features that are more frequently selected in each training stage of the cross-validation.

Perform a statistical analysis for evaluating the most performing classification scheme (is it actually needed an attribute selection process?). Use The C4.5 without feature selection as control algorithm.

If not, are there any advantages in using a reduced set of features?




```
=== Evaluation result ===

Scheme: J48
Options: -C 0.25 -M 2
Relation: ionosphere

=== Summary ===

Correctly Classified Instances         321               91.453  %
Incorrectly Classified Instances        30                8.547  %
Kappa statistic                          0.8096
Mean absolute error                      0.0938
Root mean squared error                  0.2901
Relative absolute error                 20.36   %
Root relative squared error             60.4599 %
Total Number of Instances              351

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,825     0,036     0,929       0,825   0,874      0,813    0,892     0,855     b
               0,964     0,175     0,908       0,964   0,935      0,813    0,892     0,894     g
Weighted Avg.  0,915     0,125     0,915       0,915   0,913      0,813    0,892     0,880

=== Confusion Matrix ===

   a   b   <-- classified as
 104  22 |   a = b
   8 217 |   b = g
```

```
=== Evaluation result ===

Scheme: J48
Options: -C 0.25 -M 2
Relation: ionosphere

=== Summary ===

Correctly Classified Instances         307               87.4644 %
Incorrectly Classified Instances        44               12.5356 %
Kappa statistic                          0.7238
Mean absolute error                      0.1318
Root mean squared error                  0.3384
Relative absolute error                 28.6307 %
Root relative squared error             70.5428 %
Total Number of Instances              351

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,794     0,080     0,847       0,794   0,820      0,725    0,890     0,835     b
               0,920     0,206     0,888       0,920   0,904      0,725    0,890     0,896     g
Weighted Avg.  0,875     0,161     0,874       0,875   0,874      0,725    0,890     0,874

=== Confusion Matrix ===

   a   b   <-- classified as
 100  26 |   a = b
  18 207 |   b = g
```

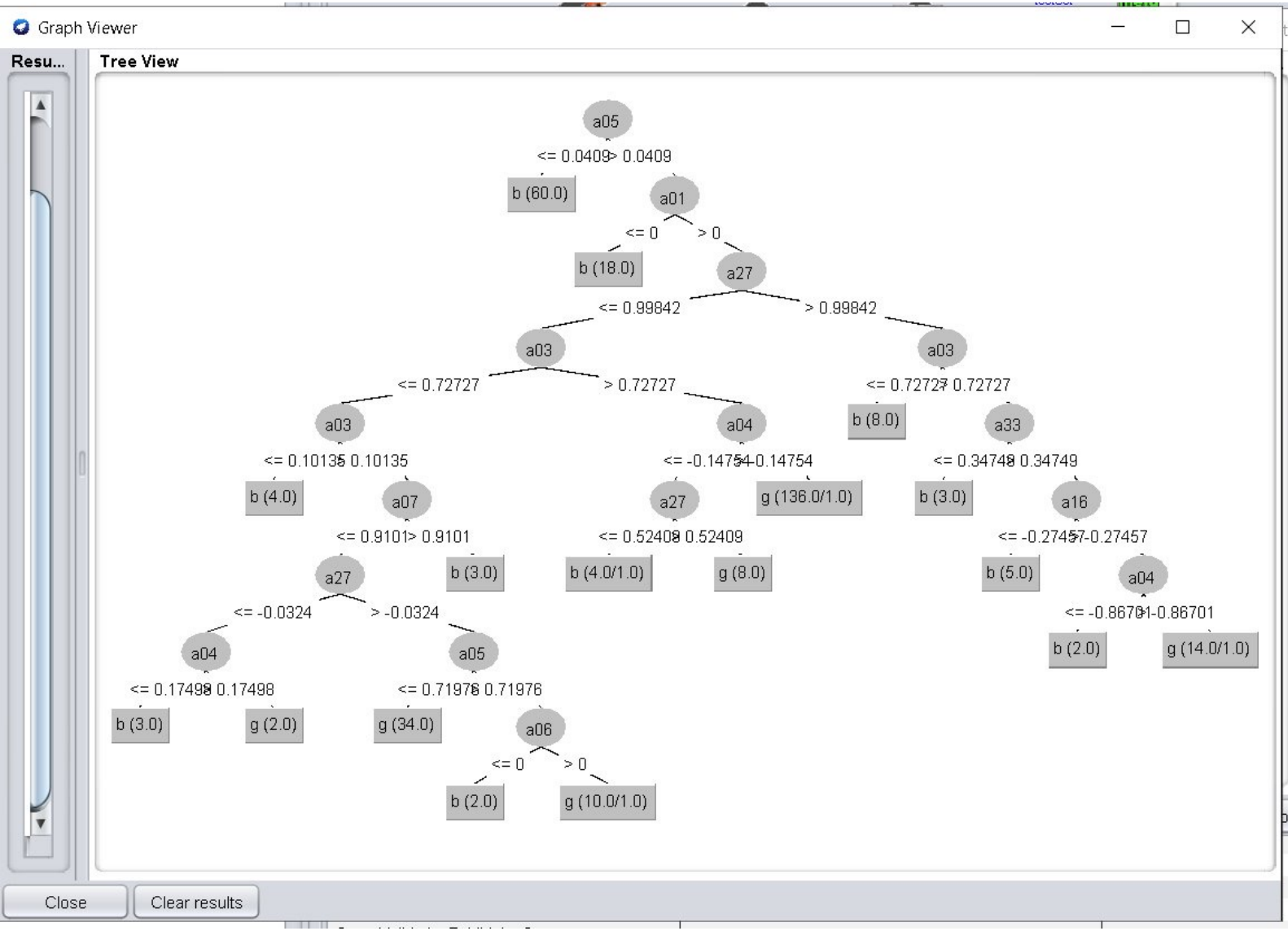
To obtain the 2x10 fold cross-falidation you must do the avg of the 2 results. Accuracy = (91.453 + 87.4644) / 2 = 89,4587

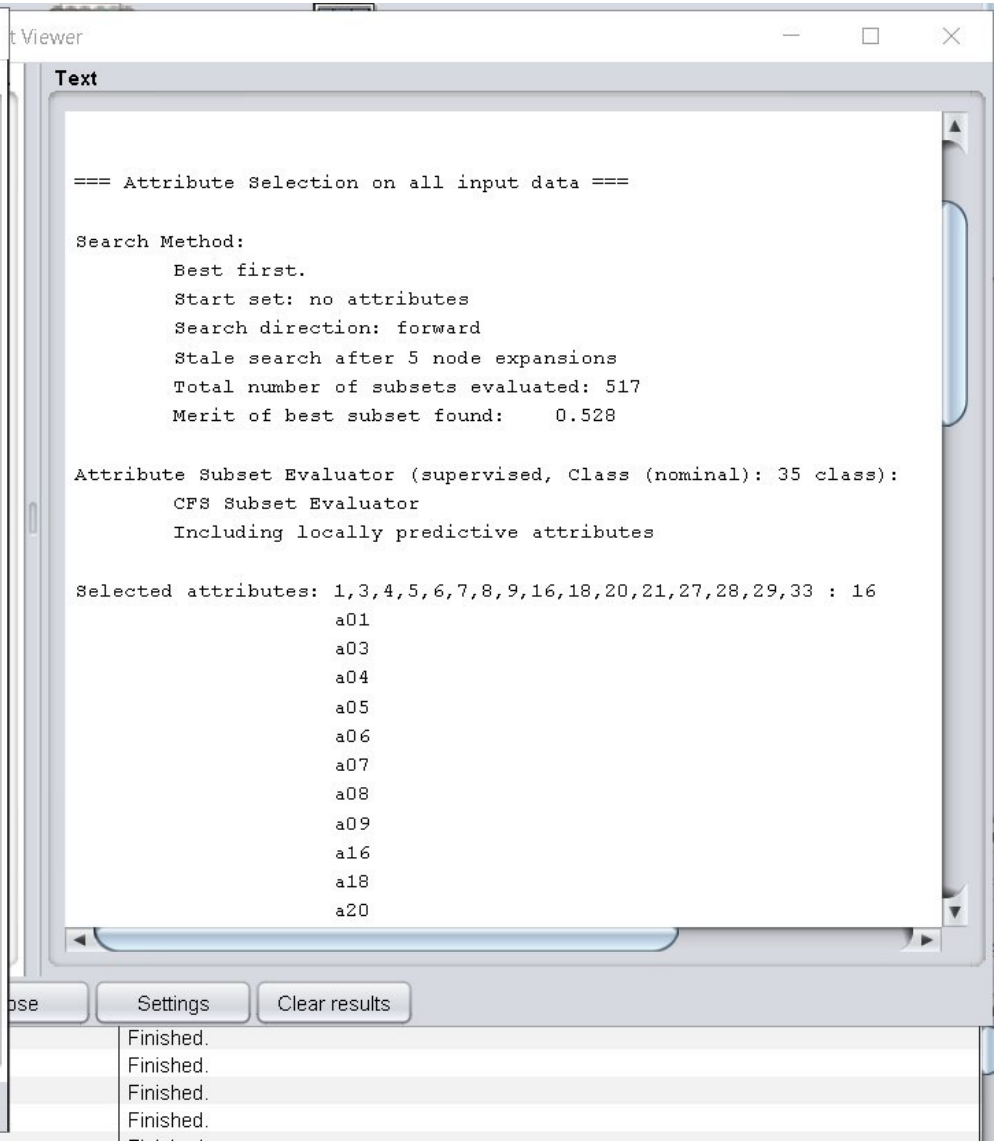
Test output

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-pr
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       08/11/21, 19:18


Dataset                   (4) trees.J4 | (1) meta. (2) meta. (3) meta.
------------------------------------------------------------------------
ionosphere                (20)   89.46 |   89.46     91.33     91.75
------------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) meta.AttributeSelectedClassifier '-E \"CfsSubsetEval -P 1 -E 1\" -S \"BestFirst -D 1 -N 5\" -W trees.J48 -- -C 0.25 -M 2' -1151805453487947577
(2) meta.AttributeSelectedClassifier '-E \"WrapperSubsetEval -B bayes.NaiveBayes -F 5 -T 0.01 -R 1 -E DEFAULT --\" -S \"BestFirst -D 1 -N 5\" -W trees.J48
(3) meta.AttributeSelectedClassifier '-E \"InfoGainAttributeEval \" -S \"Ranker -T 0.3 -N -1\" -W trees.J48 -- -C 0.25 -M 2' -1151805453487947577
(4) trees.J48 '-C 0.25 -M 2' -217733168393644444
```

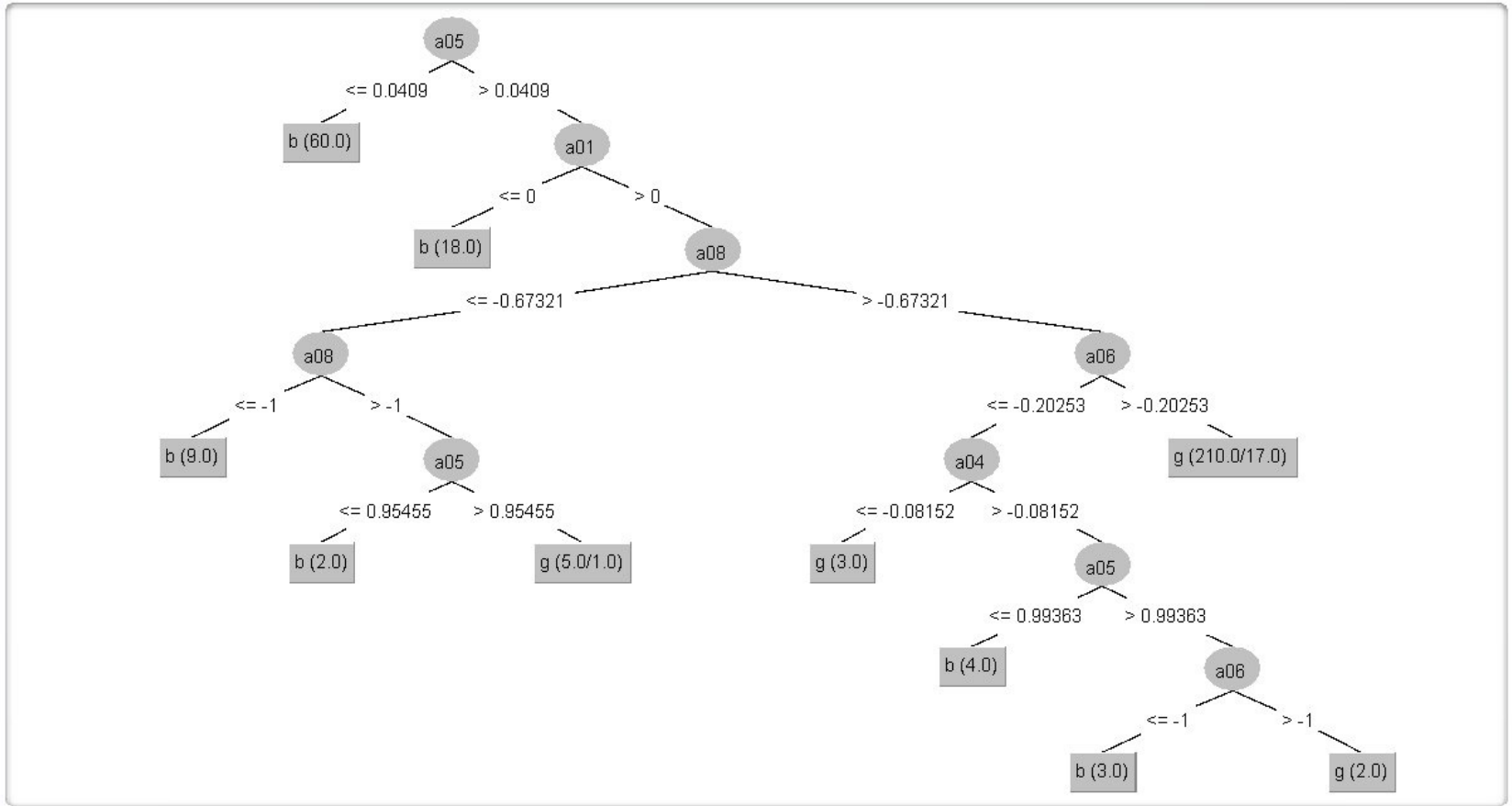

Graph Viewer
Resu...
Tree View
a05
<= 0.0409
> 0.0409
b (60.0)
a01
<= 0
> 0
b (18.0)
a27
<= 0.99842
> 0.99842
a03
a03
<= 0.72727
> 0.72727
<= 0.72727
> 0.72727
a03
a04
b (8.0)
a33
<= 0.10135
> 0.10135
<= -0.14754
> -0.14754
<= 0.34748
> 0.34749
b (4.0)
a07
a27
g (136.0/1.0)
b (3.0)
a16
<= 0.9101
> 0.9101
<= 0.52408
> 0.52409
<= -0.27457
> -0.27457
a27
b (3.0)
b (4.0/1.0)
g (8.0)
b (5.0)
a04
<= -0.0324
> -0.0324
<= -0.86701
> -0.86701
a04
a05
b (2.0)
g (14.0/1.0)
<= 0.17498
> 0.17498
<= 0.71976
> 0.71976
b (3.0)
g (2.0)
g (34.0)
a06
<= 0
> 0
b (2.0)
g (10.0/1.0)
Close
Clear results


t Viewer
Text
=== Attribute Selection on all input data ===
Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 517
Merit of best subset found: 0.528
Attribute Subset Evaluator (supervised, Class (nominal): 35 class):
CFS Subset Evaluator
Including locally predictive attributes
Selected attributes: 1,3,4,5,6,7,8,9,16,18,20,21,27,28,29,33 : 16
a01
a03
a04
a05
a06
a07
a08
a09
a16
a18
a20
Settings
Clear results
Finished.
Finished.
Finished.
Finished.

Graph Viewer

Resu...

Tree View



Close | Clear results

Text

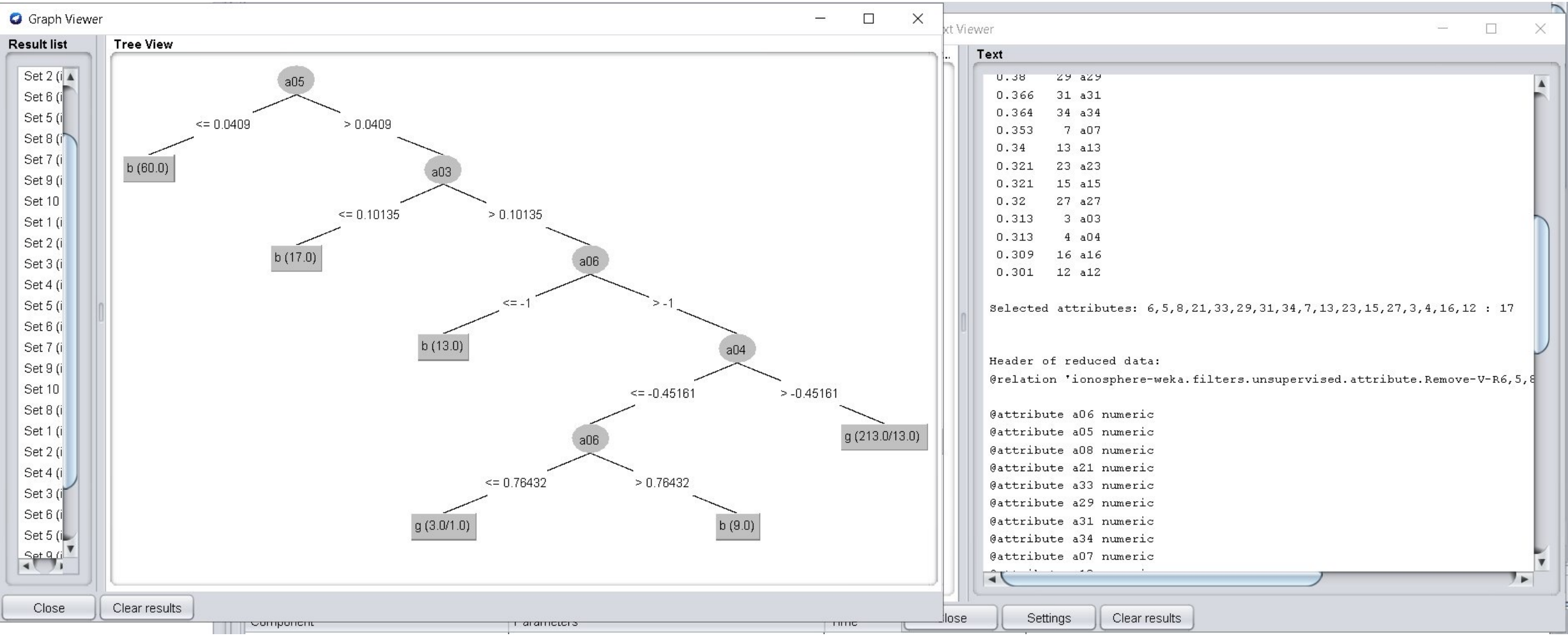
```
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 448
        Merit of best subset found:    0.938

Attribute Subset Evaluator (supervised, Class (nominal):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.bayes.NaiveBay
        Scheme options:
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 1,3,4,5,6,8,12,14,16,26,27,32 : 12
                     a01
                     a03
                     a04
                     a05
                     a06
                     a08
                     a12
                     a14
                     a16
                     a26
                     a27
                     a32

Header of reduced data:
```

Clear results


Graph Viewer
Result list
Tree View
Set 2 (i
Set 6 (i
Set 5 (i
Set 8 (i
Set 7 (i
Set 9 (i
Set 10
Set 1 (i
Set 2 (i
Set 3 (i
Set 4 (i
Set 5 (i
Set 6 (i
Set 7 (i
Set 9 (i
Set 10
Set 8 (i
Set 1 (i
Set 2 (i
Set 4 (i
Set 3 (i
Set 6 (i
Set 5 (i
a05
<= 0.0409
> 0.0409
b (60.0)
a03
<= 0.10135
> 0.10135
b (17.0)
a06
<= -1
> -1
b (13.0)
a04
<= -0.45161
> -0.45161
g (213.0/13.0)
a06
<= 0.76432
> 0.76432
g (3.0/1.0)
b (9.0)
Close
Clear results
Text
0.366 31 a31
0.364 34 a34
0.353 7 a07
0.34 13 a13
0.321 23 a23
0.321 15 a15
0.32 27 a27
0.313 3 a03
0.313 4 a04
0.309 16 a16
0.301 12 a12
Selected attributes: 6,5,8,21,33,29,31,34,7,13,23,15,27,3,4,16,12 : 17
Header of reduced data:
@relation 'ionosphere-weka.filters.unsupervised.attribute.Remove-V-R6,5,8
@attribute a06 numeric
@attribute a05 numeric
@attribute a08 numeric
@attribute a21 numeric
@attribute a33 numeric
@attribute a29 numeric
@attribute a31 numeric
@attribute a34 numeric
@attribute a07 numeric
Settings
Clear results