Solution to Laboratory Week 6.

1. You will need to have access to the WEKA package.

2. The data files for this lab can be found at /KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data

3. Load the file /arff/UCI/ionosphere.arff

(a) Get the classification accuracy with J48.

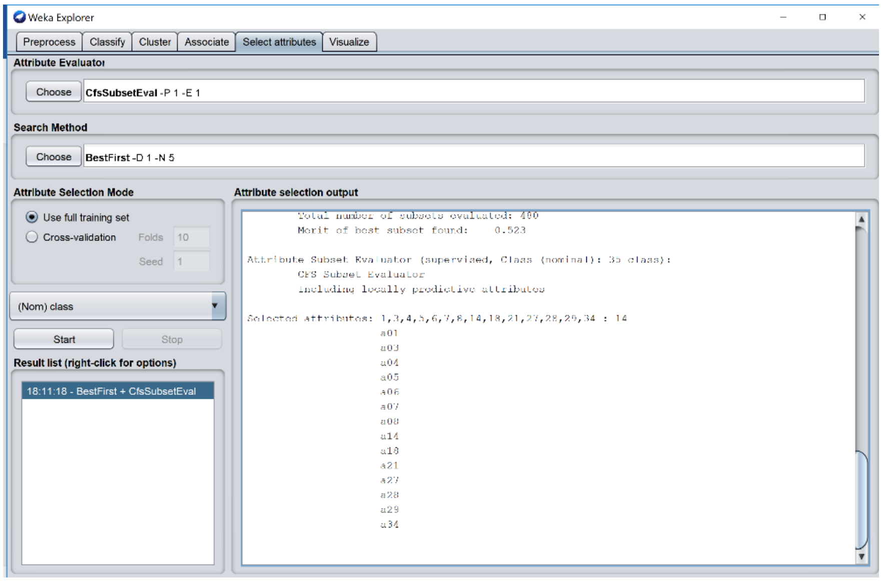
The classification accuracy of J48, using 10-fold cross-validation on ionosphere data set is as follows:

Total number of instances: 351

Correctly Classified Instances 321 91.453 %

Incorrectly Classified Instances 30 8.547 %

(b) Apply attribute selection with the default settings, ie CfsSubsetEval and BestFirst. Go back to Preprocess, remove all but the selected attributes and rerun J48. What is the accuracy?



Attribute Evaluator: CfsSubsetEval -P 1 -E 1

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

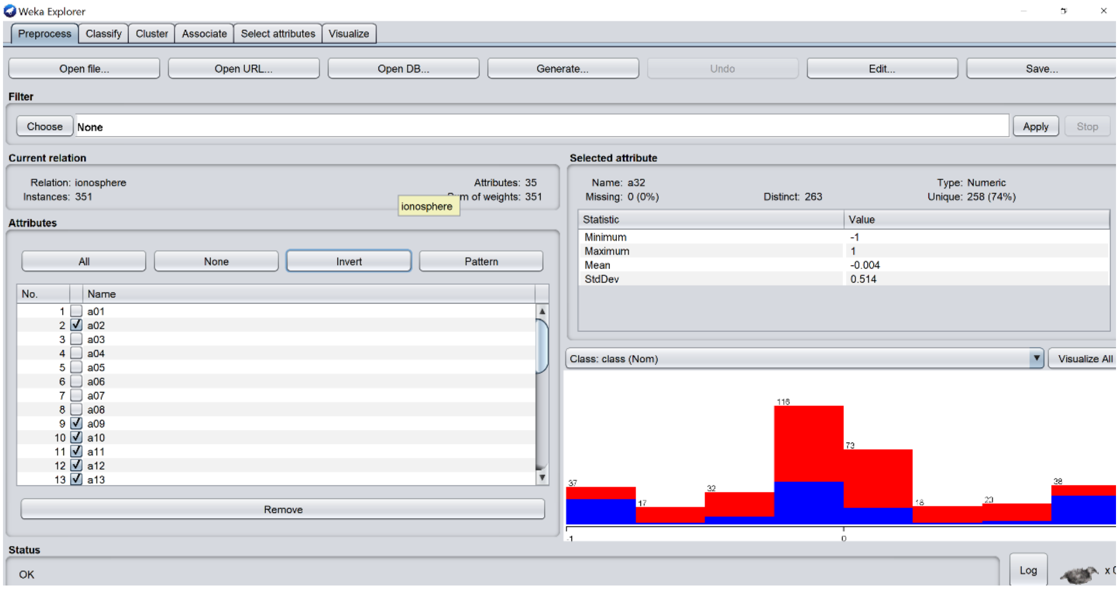
(Nom) class

Result list (right-click for options): 18:11:18 - BestFirst + CfsSubsetEval

Attribute selection output

```
        Total number of subsets evaluated: 400
        Merit of best subset found:    0.523

Attribute Subset Evaluator (supervised, Class (nominal): 35 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,3,4,5,6,7,8,14,18,21,27,28,29,34 : 14
                     a01
                     a03
                     a04
                     a05
                     a06
                     a07
                     a08
                     a14
                     a18
                     a21
                     a27
                     a28
                     a29
                     a34
```

Figure 1. List of selected attributes by running CfsSubsetEval and BestFirst attribute selection algorithms.



Filter: None

Current relation
Relation: ionosphere
Instances: 351
Attributes: 35
Sum of weights: 351

Attributes

No.		Name
1	☐	a01
2	☑	a02
3	☐	a03
4	☐	a04
5	☐	a05
6	☐	a06
7	☐	a07
8	☐	a08
9	☑	a09
10	☑	a10
11	☑	a11
12	☑	a12
13	☑	a13

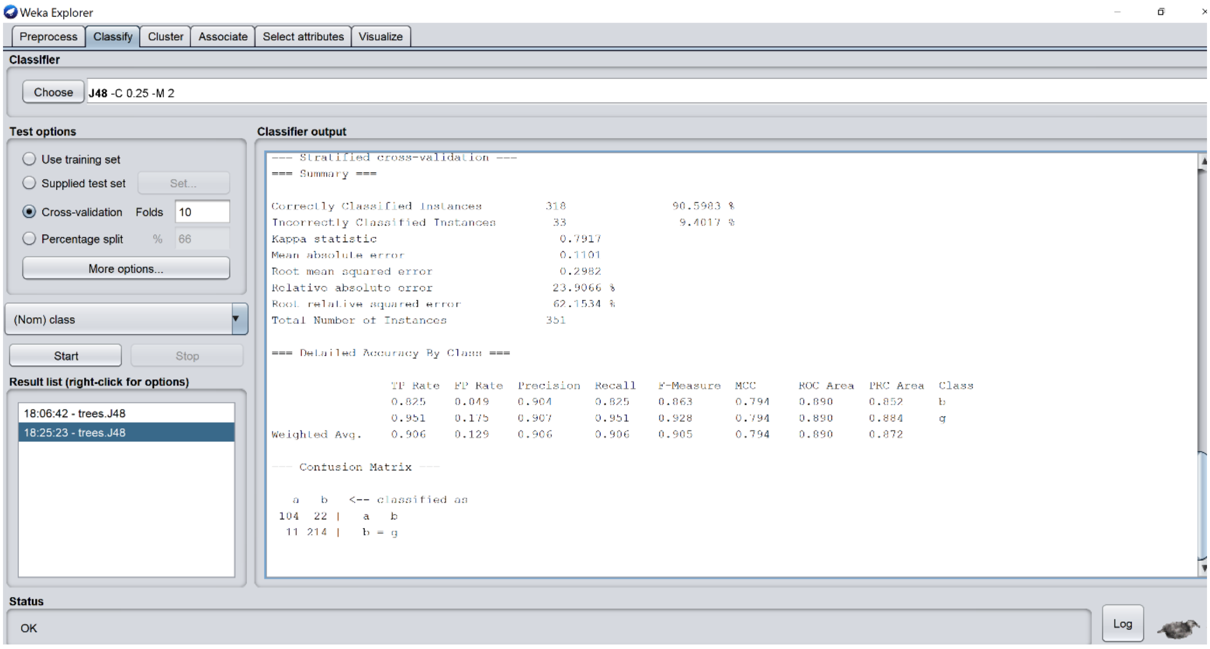
Selected attribute
Name: a32 Type: Numeric
Missing: 0 (0%) Distinct: 263 Unique: 258 (74%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	-0.004
StdDev	0.514

Class: class (Nom)

Status
OK

Figure 2. Remove all but the selected attributes from preprocess tab by clicking on Remove button. Be careful do not remove class label.



Classifier: J48 -C 0.25 -M 2

Cross-validation Folds 10; Percentage split % 66

Result list: 18:06:42 - trees.J48; 18:25:23 - trees.J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         318               90.5983 %
Incorrectly Classified Instances        33                9.4017 %
Kappa statistic                          0.7917
Mean absolute error                      0.1101
Root mean squared error                  0.2982
Relative absolute error                 23.9066 %
Root relative squared error             62.1534 %
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    0.049    0.904      0.825    0.863      0.794    0.890     0.852     b
                 0.951    0.175    0.907      0.951    0.928      0.794    0.890     0.884     g
Weighted Avg.    0.906    0.129    0.906      0.906    0.905      0.794    0.890     0.872

=== Confusion Matrix ===

   a   b   <-- classified as
 104  22 |   a = b
  11 214 |   b = g
```

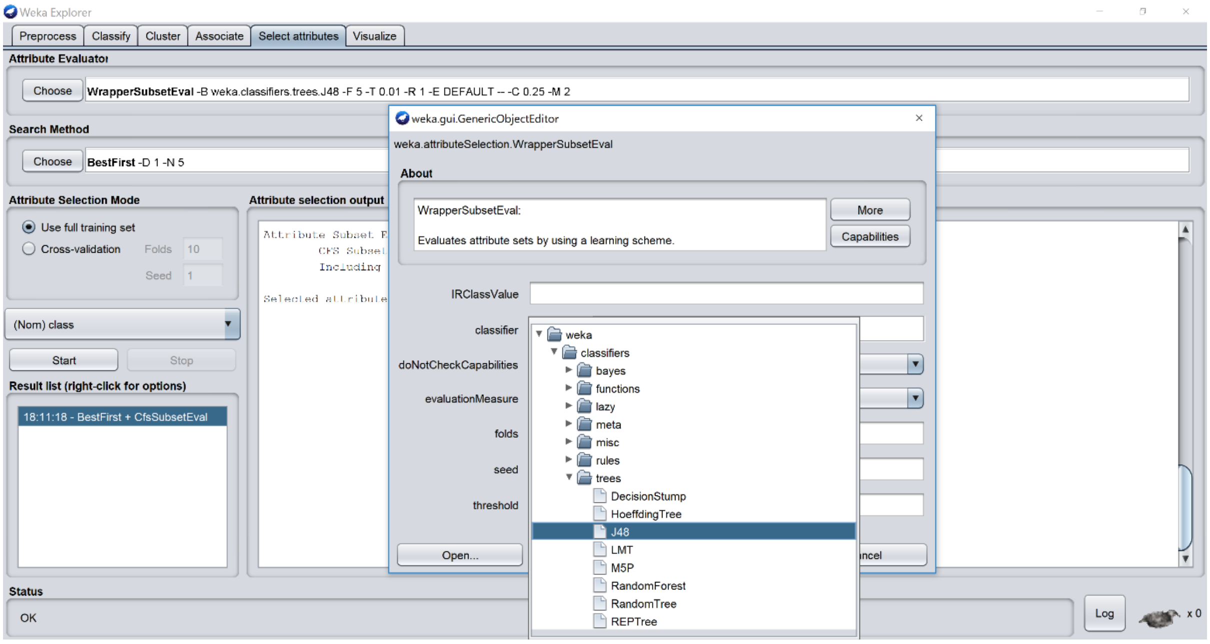
Status: OK

Figure 3. Output of running J48 after attribute selection.

As it can be seen from figure 3, accuracy of J48 is reducing to 90.593% after attribute selection.

[Useful hint] Go to Preprocess -->Filters--> Supervised -Attribute and select the AttributeSelection filter. Using this filter (with Undo) will save the tedious task of manually selecting the results of attribute selection.

(c) Reload ionosphere.arff Apply attribute selection with WrapperSubsetEval, BestFirst and J48 as the classifier in WrapperSubsetEval. Go back to Preprocess, remove all but the selected attributes and rerun J48. What is the accuracy?



Attribute Evaluator: WrapperSubsetEval -B weka.classifiers.trees.J48 -F 5 -T 0.01 -R 1 -E DEFAULT -- -C 0.25 -M 2

Search Method: BestFirst -D 1 -N 5

Cross-validation Folds 10 Seed 1

Result list: 18:11:18 - BestFirst + CfsSubsetEval

```
Attribute Subset E
        CFS Subset
        Including

Selected attribute
```

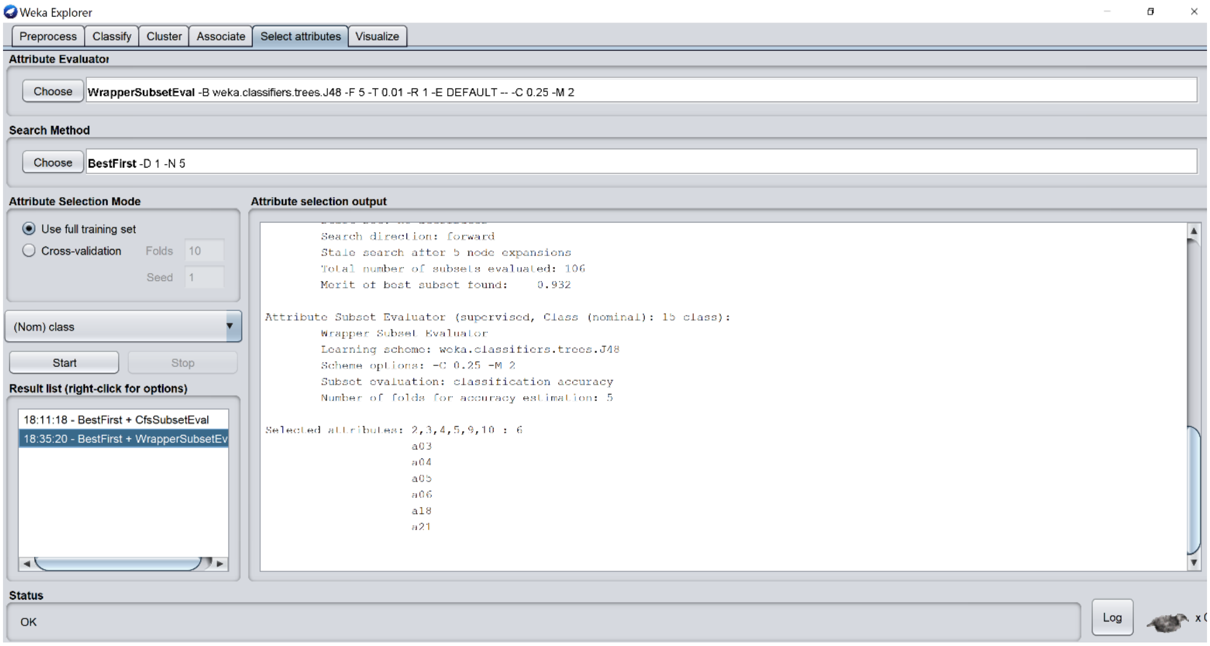
weka.attributeSelection.WrapperSubsetEval

WrapperSubsetEval:

Evaluates attribute sets by using a learning scheme.

Status: OK

Figure 4. Snapshot of setting attribute selection algorithm to WrapperSubsetEval and setting its classifier parameter to J48.



Attribute Evaluator: WrapperSubsetEval -B weka.classifiers.trees.J48 -F 5 -T 0.01 -R 1 -E DEFAULT -- -C 0.25 -M 2

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) class

Result list (right-click for options)

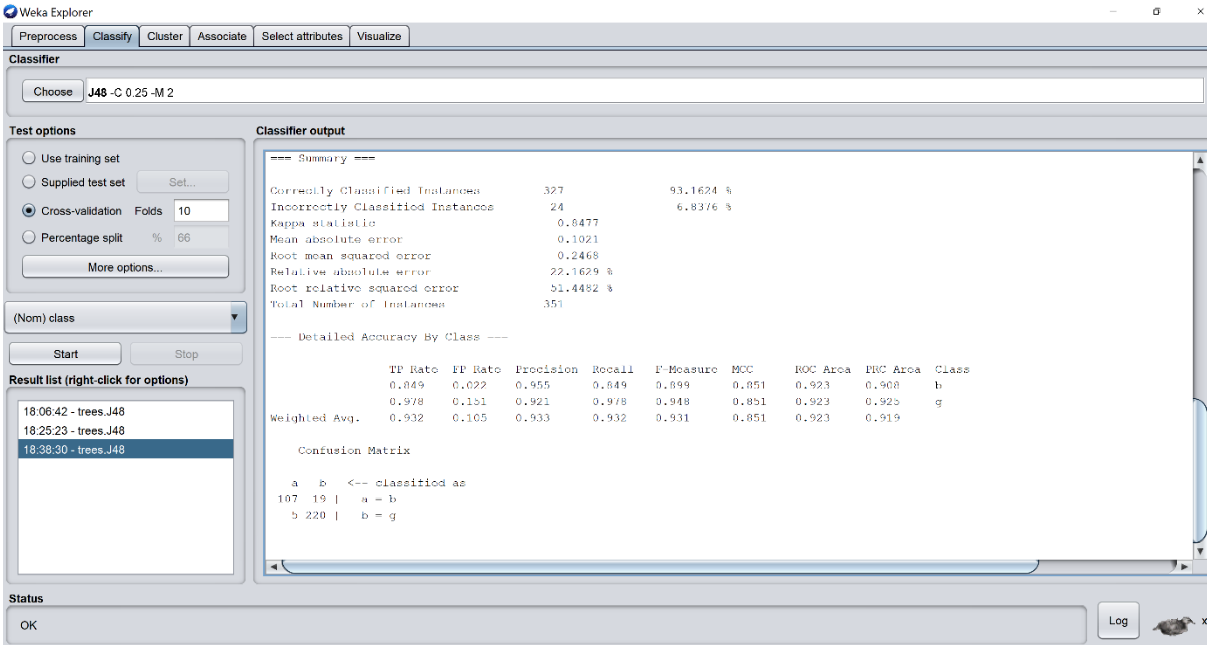
18:11:18 - BestFirst + CfsSubsetEval
18:35:20 - BestFirst + WrapperSubsetEv

Attribute selection output

```
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 106
        Merit of best subset found:    0.932

Attribute Subset Evaluator (supervised, Class (nominal): 15 class):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 2,3,4,5,9,10 : 6
                     a03
                     a04
                     a05
                     a06
                     a18
                     a21
```

Status: OK

Figure 5. Output of attribute selection from WrapperSubsetEval.



Classifier: J48 -C 0.25 -M 2

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

Result list (right-click for options)

18:06:42 - trees.J48
18:25:23 - trees.J48
18:38:30 - trees.J48

Classifier output

```
=== Summary ===

Correctly Classified Instances         327               93.1624 %
Incorrectly Classified Instances        24                6.8376 %
Kappa statistic                          0.8477
Mean absolute error                      0.1021
Root mean squared error                  0.2468
Relative absolute error                 22.1629 %
Root relative squared error             51.4482 %
Total Number of Instances              351

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.849    0.022    0.955      0.849    0.899      0.851    0.923     0.908     b
                 0.978    0.151    0.921      0.978    0.948      0.851    0.923     0.925     g
Weighted Avg.    0.932    0.105    0.933      0.932    0.931      0.851    0.923     0.919

    Confusion Matrix

   a   b   <-- classified as
 107  19 |   a = b
   5 220 |   b = g
```

Status: OK

Figure 6. Output of running J48 after wrapper attribute selection from figure 5.

As it can be seen from figure 6, the accuracy of classifier is increased to 93.1624% by using wrapper attribute selection algorithm.

(d) What do you conclude about the value of attribute selection?

Using attribute selection algorithm could improve accuracy of classifier or make it worse. In this case, using the CfsSubsetEval filter results in fewer attributes, but lower accuracy. Using the wrapper method results in fewer attributes and higher accuracy.

(e) Explore other combinations of evaluator and search method. Can you find anything better?

4. The file /arff/UCI/isolet.arff has 618 attributes. Explore a variety of attribute selection techniques to reduce the number of attributes without reducing accuracy. What is your best result?

Using isolet1+2+3+4.arff

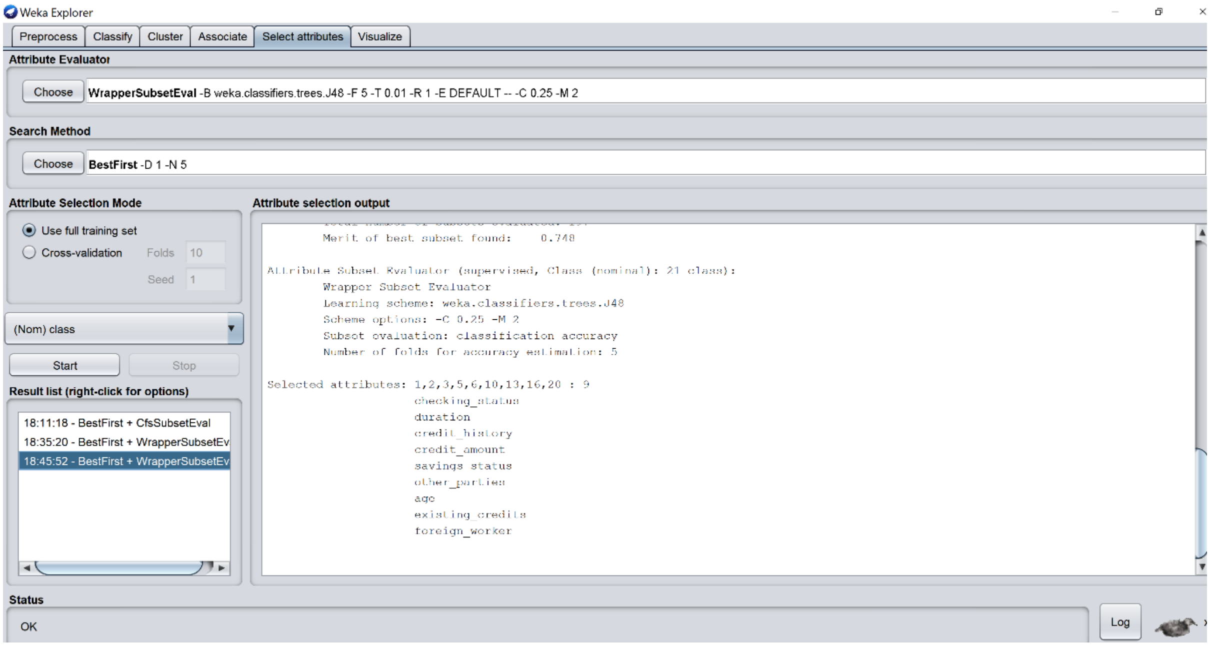
Evaluator	Search	Num Features	Accuracy
Full Feature set		617	82.27
CfssubsetEval	Best First	190	83.39
CfssubsetEval	Best First	191	83.42
CorrelationAttrEval	Ranker	190	80.15
GainRatio	Ranker	200	81.65
InfoGain	Ranker	200	78.83
OneR	Ranker	200	79.25
Wrapper J48, M=10	Best First	???	??.??

Wrapper has run for 24 hours and didn't finish.

Results so far indicate that slightly better accuracy can be achieved with a reduced number of attributes. There is a fair variation in the performance of the attribute selection methods.

5. Load the file /arff/UCI/credit-g.arff

(a) Repeat (3) on this file.



Attribute Evaluator

WrapperSubsetEval -B weka.classifiers.trees.J48 -F 5 -T 0.01 -R 1 -E DEFAULT -- -C 0.25 -M 2

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) class

Result list (right-click for options)

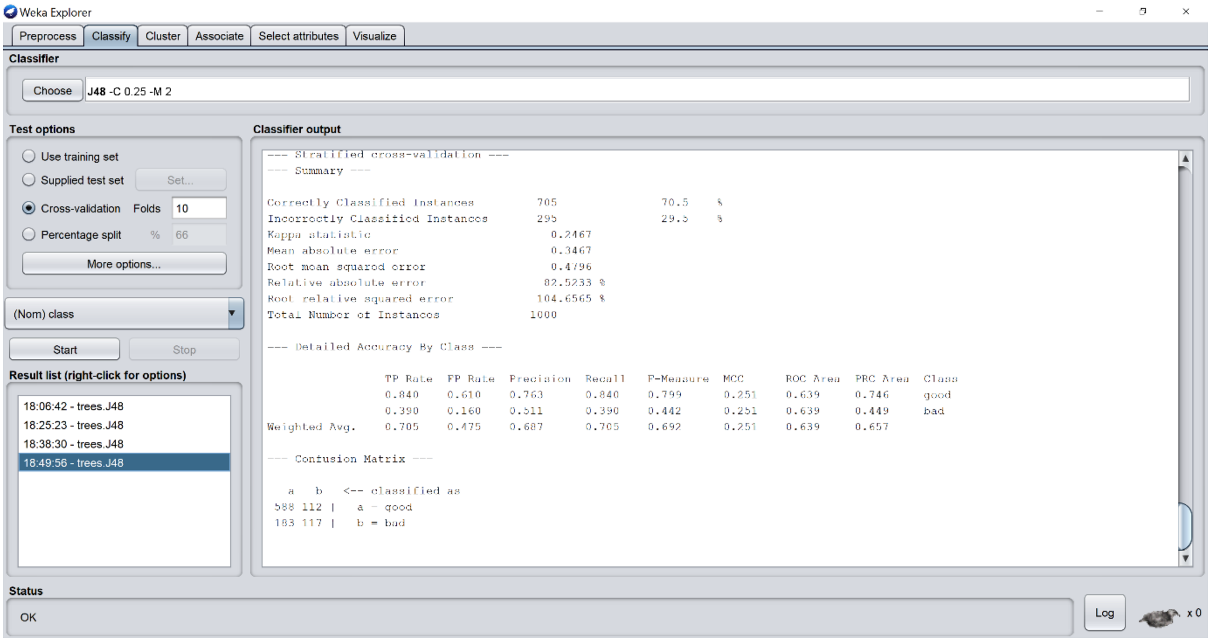
- 18:11:18 - BestFirst + CfsSubsetEval
- 18:35:20 - BestFirst + WrapperSubsetEv
- 18:45:52 - BestFirst + WrapperSubsetEv

Attribute selection output

```
        Merit of best subset found:    0.748

Attribute Subset Evaluator (supervised, Class (nominal): 21 class):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 1,2,3,5,6,10,13,16,20 : 9
                     checking_status
                     duration
                     credit_history
                     credit_amount
                     savings status
                     other_parties
                     age
                     existing_credits
                     foreign_worker
```

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Result list (right-click for options)

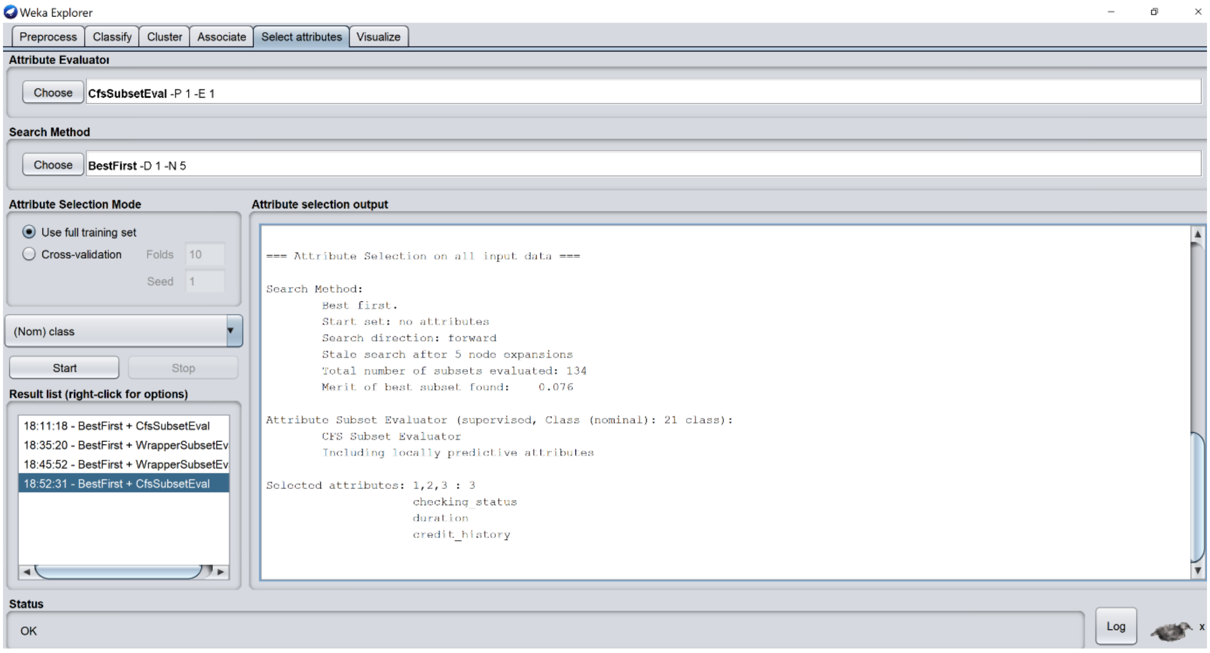
18:06:42 - trees.J48
18:25:23 - trees.J48
18:38:30 - trees.J48
18:49:56 - trees.J48

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         705               70.5    %
Incorrectly Classified Instances       295               29.5    %
Kappa statistic                          0.2467
Mean absolute error                      0.3467
Root mean squared error                  0.4796
Relative absolute error                 82.5233 %
Root relative squared error            104.6565 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.840    0.610    0.763      0.840    0.799      0.251    0.639     0.746     good
                 0.390    0.160    0.511      0.390    0.442      0.251    0.639     0.449     bad
Weighted Avg.    0.705    0.475    0.687      0.705    0.692      0.251    0.639     0.657

=== Confusion Matrix ===

   a   b   <-- classified as
 588 112 |   a = good
 183 117 |   b = bad
```

Status

OK



Attribute Evaluator

CfsSubsetEval -P 1 -E 1

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) class

Result list (right-click for options)

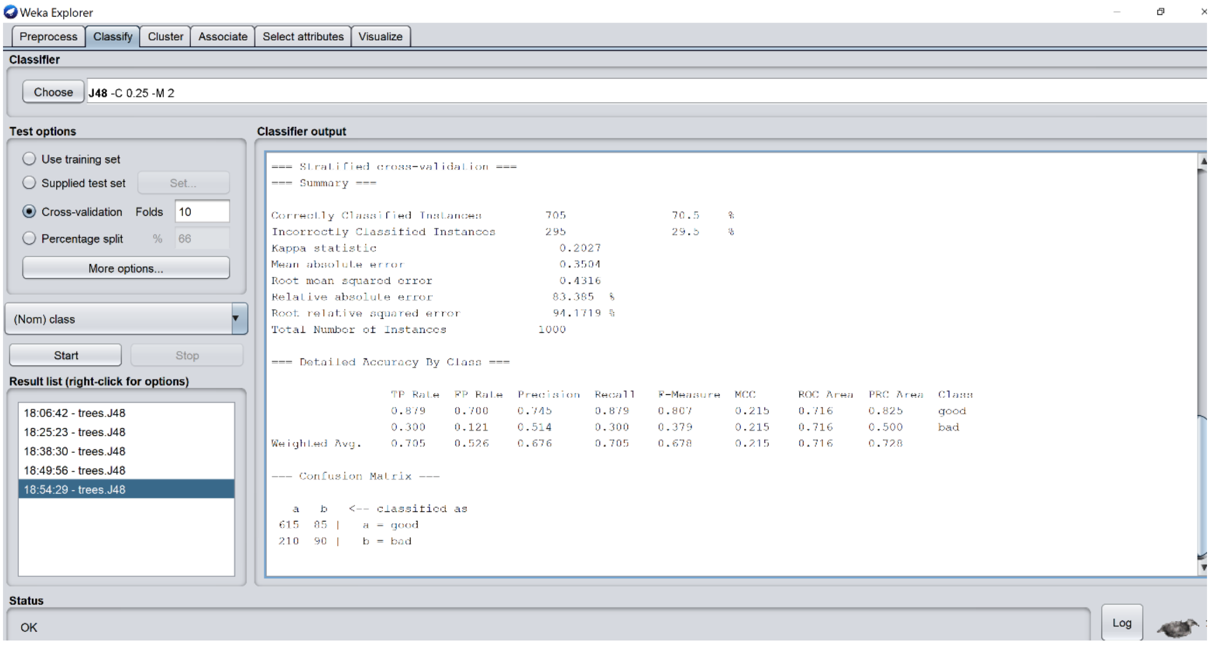
18:11:18 - BestFirst + CfsSubsetEval
18:35:20 - BestFirst + WrapperSubsetEv
18:45:52 - BestFirst + WrapperSubsetEv
18:52:31 - BestFirst + CfsSubsetEval

Attribute selection output

```
=== 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: 134
        Merit of best subset found:    0.076

Attribute Subset Evaluator (supervised, Class (nominal): 21 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,3 : 3
                     checking_status
                     duration
                     credit_history
```

Status

OK



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

18:06:42 - trees.J48
18:25:23 - trees.J48
18:38:30 - trees.J48
18:49:56 - trees.J48
18:54:29 - trees.J48

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         705               70.5    %
Incorrectly Classified Instances       295               29.5    %
Kappa statistic                          0.2027
Mean absolute error                      0.3504
Root mean squared error                  0.4316
Relative absolute error                 83.385  %
Root relative squared error             94.1719 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.879    0.700    0.745      0.879    0.807      0.215    0.716     0.825     good
                 0.300    0.121    0.514      0.300    0.379      0.215    0.716     0.500     bad
Weighted Avg.    0.705    0.526    0.676      0.705    0.678      0.215    0.716     0.728

=== Confusion Matrix ===

   a   b   <-- classified as
 615  85 |   a = good
 210  90 |   b = bad
```

Status

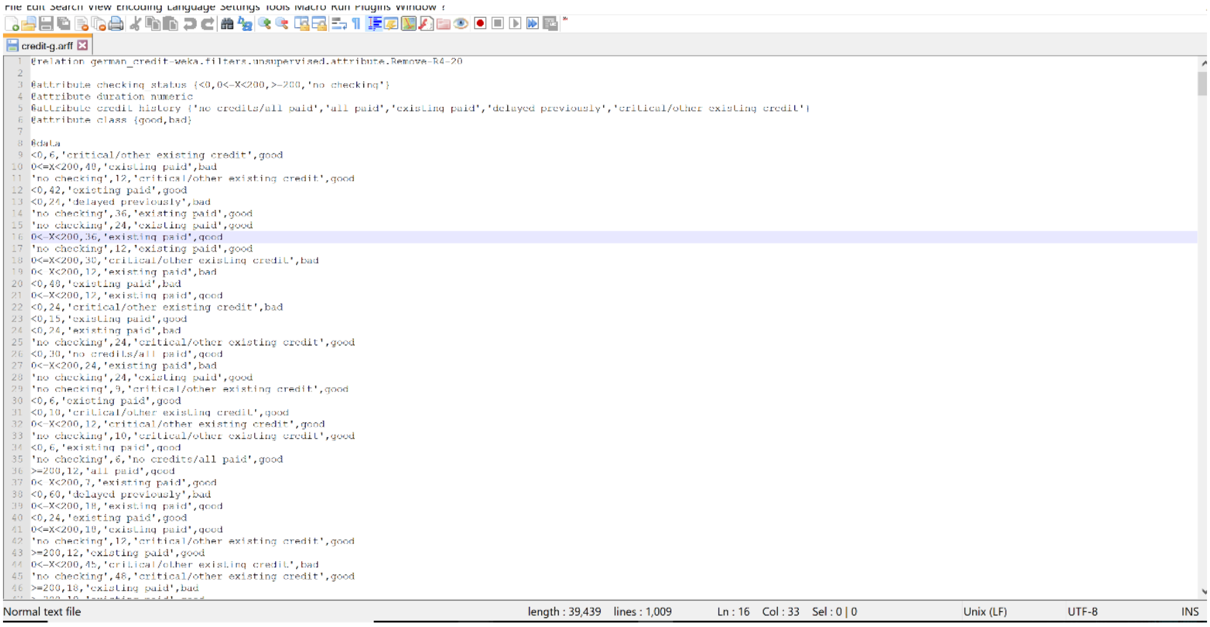
OK

Using credit-g.arff

Evaluator	Search	Num Features	Accuracy
Full Feature set		20	70.5
CfssubsetEval	Best First	3	70.5
Wrapper J48	Best First	9	70.5

The same accuracy is achieved with 3, 9 and 20 features. Looking at the file below, the 9 features look reasonable. Not sure what to make of the 3 features. Maybe an artefact of the data.

(b) Open the file with a text editor and read the descriptions of the attributes. Does the set of selected attributes make sense from what you know about credit ratings?


File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
credit-g.arff

```
@relation german_credit-weka.filters.unsupervised.attribute.Remove-R4-20

@attribute checking_status {<0,0<=X<200,>=200,'no checking'}
@attribute duration numeric
@attribute credit_history {'no credits/all paid','all paid','existing paid','delayed previously','critical/other existing credit'}
@attribute class {good,bad}

@data
<0,6,'critical/other existing credit',good
0<=X<200,48,'existing paid',bad
'no checking',12,'critical/other existing credit',good
<0,42,'existing paid',good
<0,24,'delayed previously',bad
'no checking',36,'existing paid',good
'no checking',24,'existing paid',good
0<=X<200,36,'existing paid',good
'no checking',12,'existing paid',good
0<=X<200,30,'critical/other existing credit',bad
0<=X<200,12,'existing paid',bad
<0,48,'existing paid',bad
0<=X<200,12,'existing paid',good
<0,24,'critical/other existing credit',bad
<0,15,'existing paid',good
<0,24,'existing paid',bad
'no checking',24,'critical/other existing credit',good
<0,30,'no credits/all paid',good
0<=X<200,24,'existing paid',bad
'no checking',24,'existing paid',good
'no checking',9,'critical/other existing credit',good
<0,6,'existing paid',good
<0,10,'critical/other existing credit',good
0<=X<200,12,'critical/other existing credit',good
'no checking',10,'critical/other existing credit',good
<0,6,'existing paid',good
'no checking',6,'no credits/all paid',good
>=200,12,'all paid',good
0<=X<200,7,'existing paid',good
<0,60,'delayed previously',bad
0<=X<200,18,'existing paid',good
<0,24,'existing paid',good
0<=X<200,18,'existing paid',good
'no checking',12,'critical/other existing credit',good
>=200,12,'existing paid',good
0<=X<200,45,'critical/other existing credit',bad
'no checking',48,'critical/other existing credit',good
>=200,18,'existing paid',bad
```

Normal text file   length : 39,439   lines : 1,009   Ln : 16   Col : 33   Sel : 0 | 0   Unix (LF)   UTF-8   INS