Data Mining Fundementals: Final Laboratory Assignment
Due 12/10/17
John M. Warlop

Table of contents #1 => #2 => #3 => #4

1 Choose the area of your preference and create a dataset. For example: actresses/actors, food, movies, sports, music bands, or anything you would like. Create a data file in .arff format (please attach dataset with your submission) containing at least 50 instances, each described by at least 4 attributes, the last attribute containing your preference (or class attribute), e.g.
Appendix

songs.arff

```
@relation food
@attribute calories numeric
@attribute taste {sweet, sour, bitter, salty}
@attribute course {appetizer, main, dessert, drink}
@attribute vegetarian {yes, no}
@attribute like_it {yes, no}
@data
100, sweet, dessert, yes, yes%icecream
80, bitter, drink, yes, yes%beer
2, sweet, dessert,yes, no%cake
```

```
@relation songs
@attribute group
@attribute song_name
@attribute length numeric
@attribute genre {rock, new_age, country, alt_rock, jazz}
@attribute decade {1940,1950,1960,1970,1980,1990,2000,2010}
@attribute lead {male, female}
@attribute horns {no,yes}
@attribute keyboard {no,yes}
@attribute flute_harmonica {no,yes}
@attribute stars {1,2,3,4,5}
@data
```

In your own words please describe the dataset. Use data mining to explore and create models to explain the dataset.

This dataset classifiers songs that I like. The class attribute is “stars” 5 stars is most liked and 1 star is least liked. This dataset has two string type attributes: group and song.

Create and compare at least 3 algorithms on your data set (ex. decision trees, a classification or an association rule learner, naive Bayes, etc.) For each algorithm evaluate the model and discuss your

findings. What was the performance, is the model relevant, which algorithm can explain your personal liking the best, and observe the generated rules and if they tell you anything interesting? If the model is not good, discuss why and some techniques on how you might improve.

I ran J48, One-R and PART and compared and contrasted them.

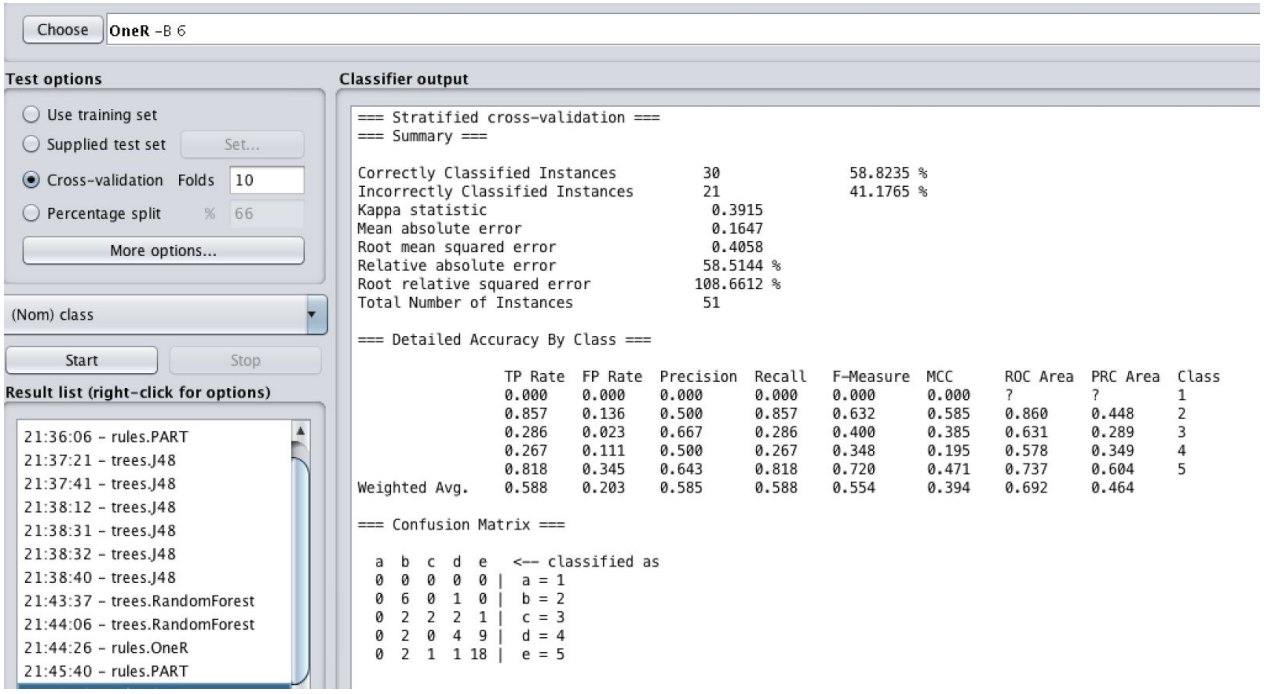
```
=== Summary ===

Correctly Classified Instances          34               66.6667 %
Incorrectly Classified Instances        17               33.3333 %
Kappa statistic                          0.5137
Mean absolute error                      0.17
Root mean squared error                  0.2915
Relative absolute error                 60.7333 %
Root relative squared error             78.4358 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         1
                 0.714    0.000    1.000      0.714    0.833      0.827    0.971     0.859     2
                 0.857    0.136    0.500      0.857    0.632      0.585    0.912     0.683     3
                 0.333    0.056    0.714      0.333    0.455      0.368    0.736     0.514     4
                 0.818    0.310    0.667      0.818    0.735      0.504    0.795     0.655     5
Weighted Avg.    0.667    0.169    0.704      0.667    0.652      0.519    0.818     0.645

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0  0  0  0 |  a = 1
  0  5  1  1  0 |  b = 2
  0  0  6  0  1 |  c = 3
  0  0  2  5  8 |  d = 4
  0  0  3  1 18 |  e = 5
```

#1-J48
Appendix
J-48 gave me the best results. My overall goal was to model which songs I liked and which songs I did not like. I believe for this to work more effectively I would need a much bigger dataset(not just 51) and more attributes.

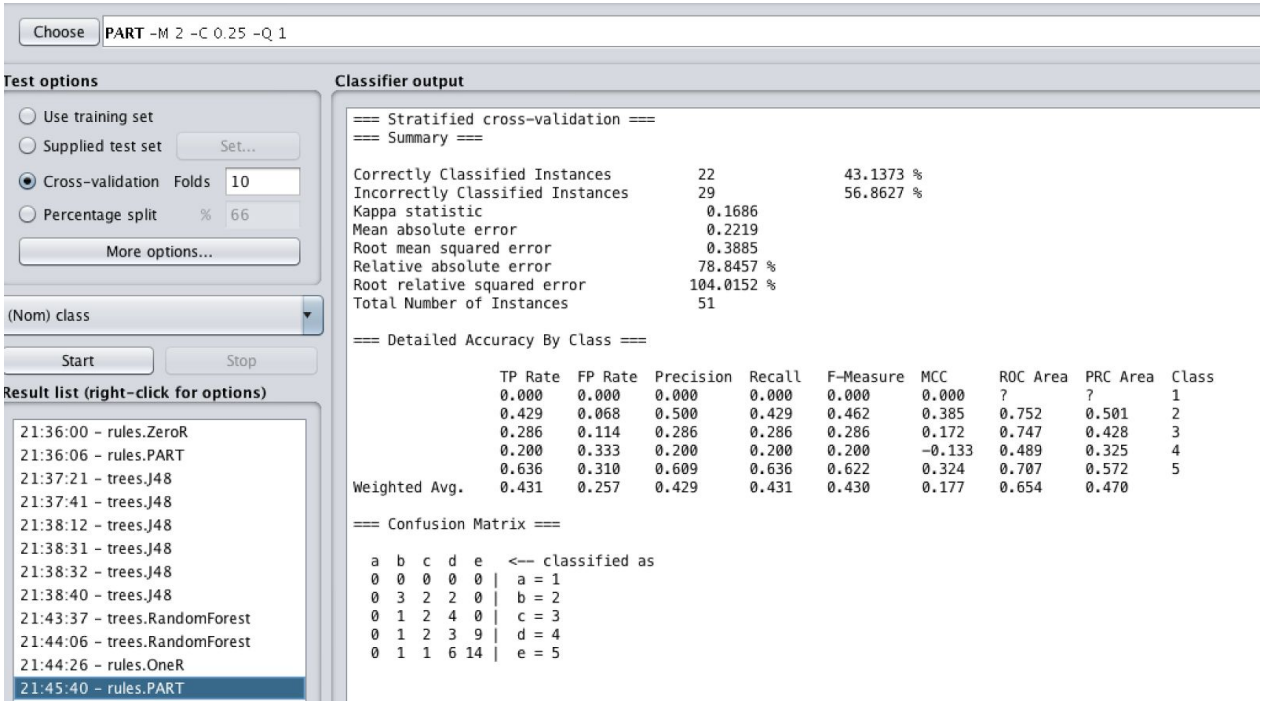
J48 did have a rule that said I like rock, which is true.

One-R also predicted I like rock, which is true.

The PART classifier had a rule that said if it is an 80's song, I'd like it(5 stars) which is true.



#1-One-R
[Appendix]
One-R has only 58.8% classification accuracy



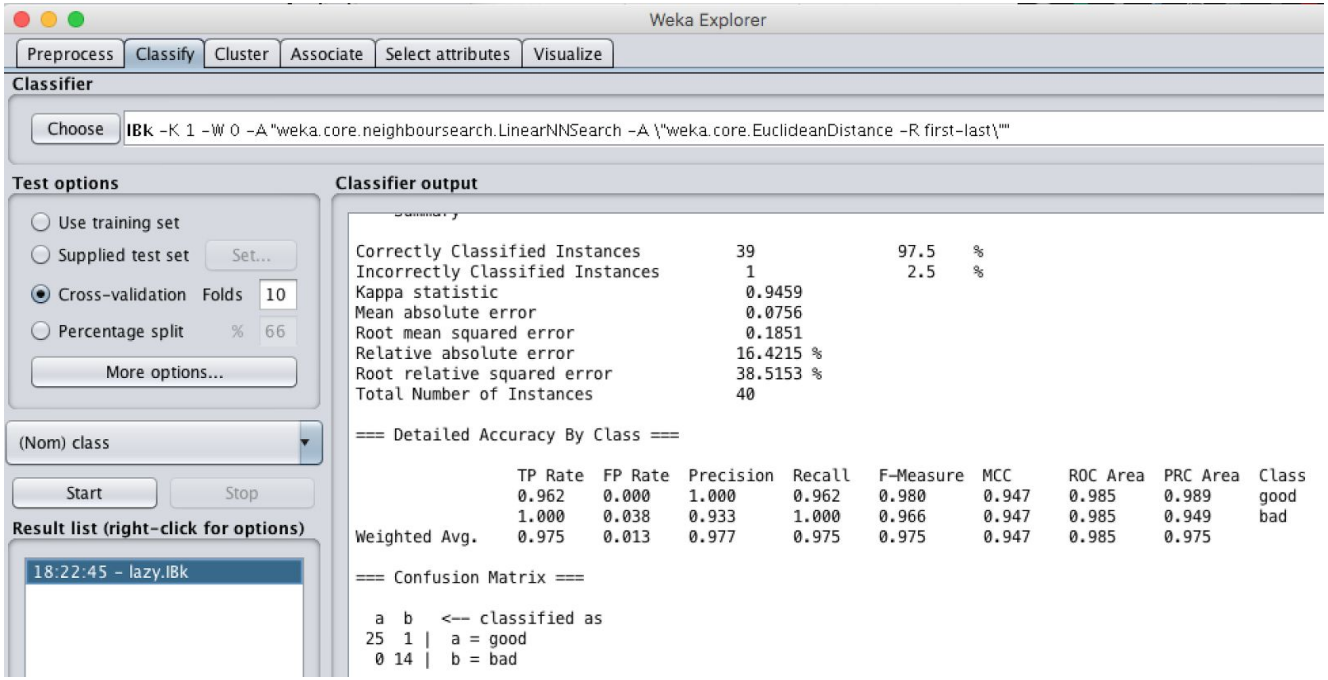
#1-PART

[Appendix]

2 Use the following learning schemes to compare the training set vs. 10-fold stratified cross-validation scores of the labor data in labor_neg_nominal.arff:

- k-nearest neighbors (IBk) with cross-validation and with training set, same with J48(4 runs)
- run J48 with cross-validation and with training set with M=3(2 runs)

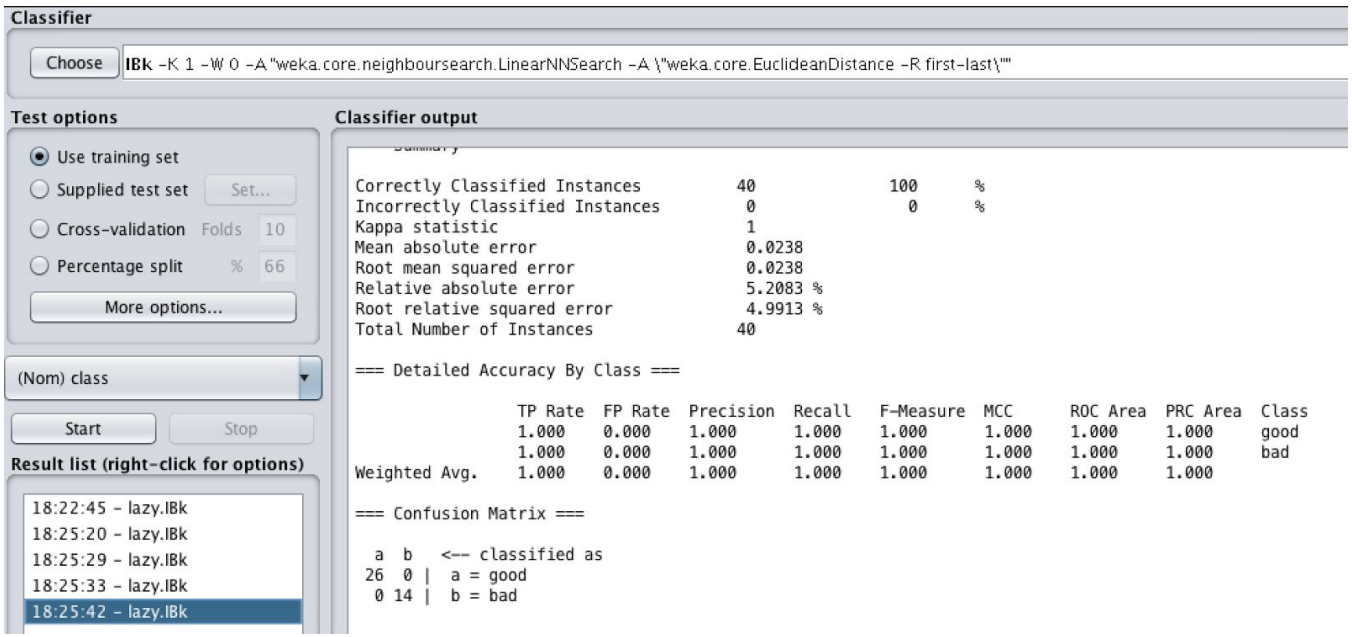
TOC
Appendix



IBk w/10-fold cross validation

IBk 10-fold

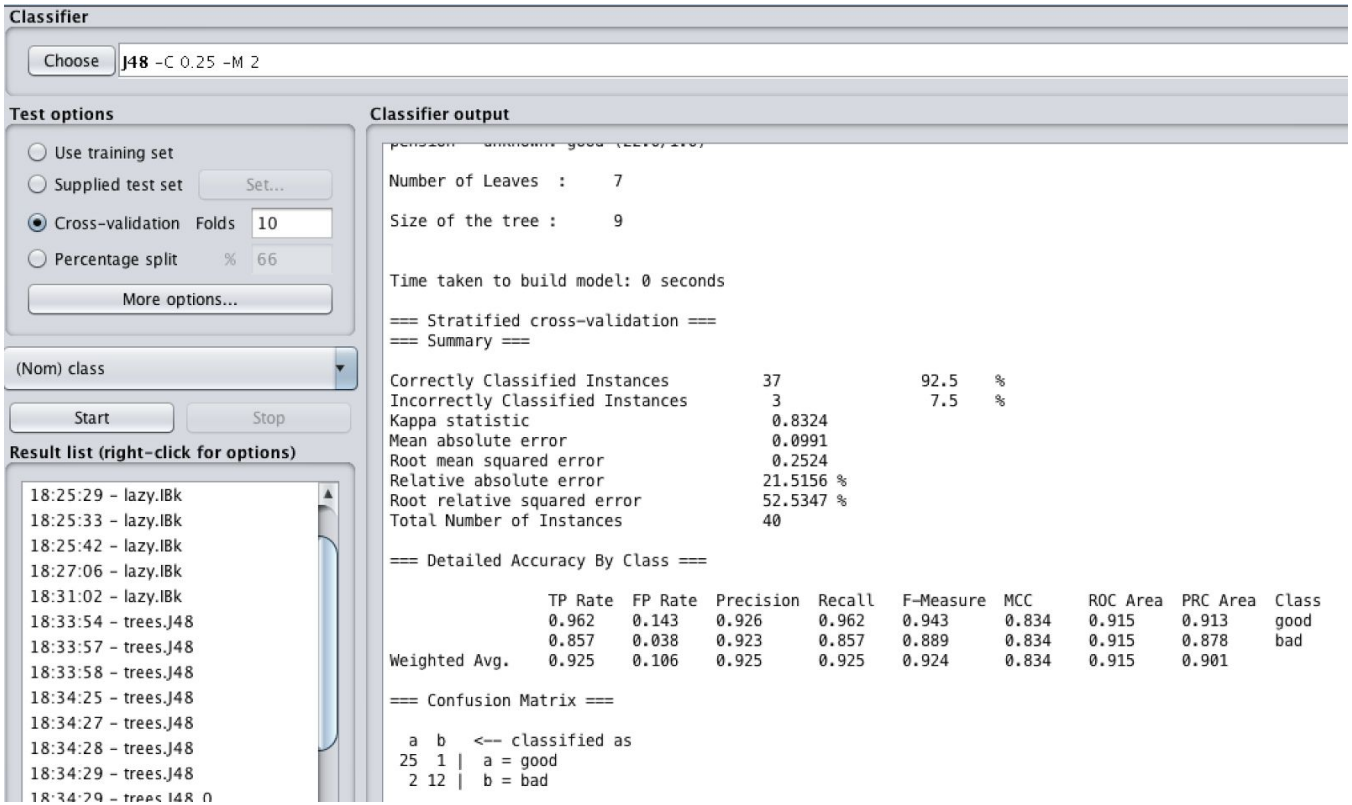
Appendix



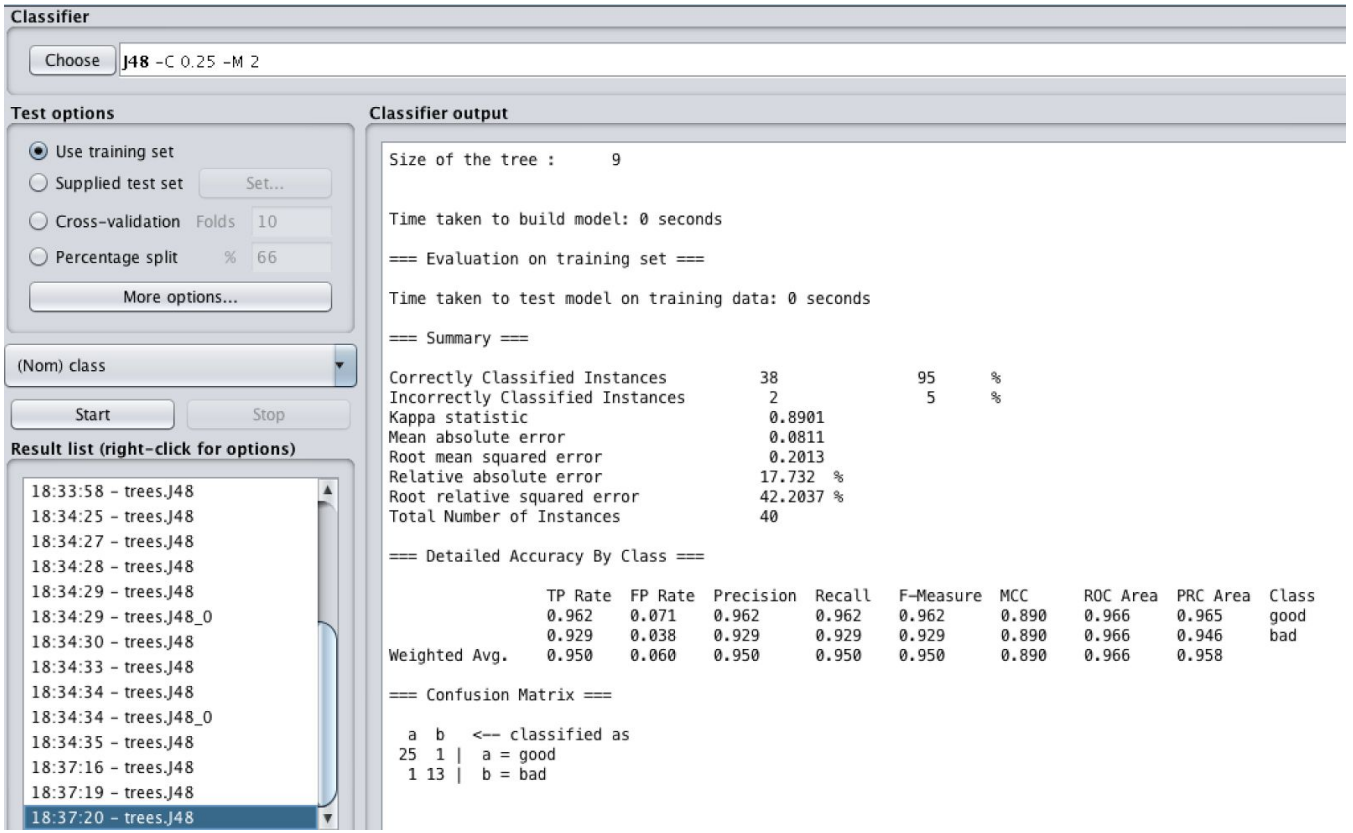
IBk training set [#2](#)

IBk training

[Appendix](#)



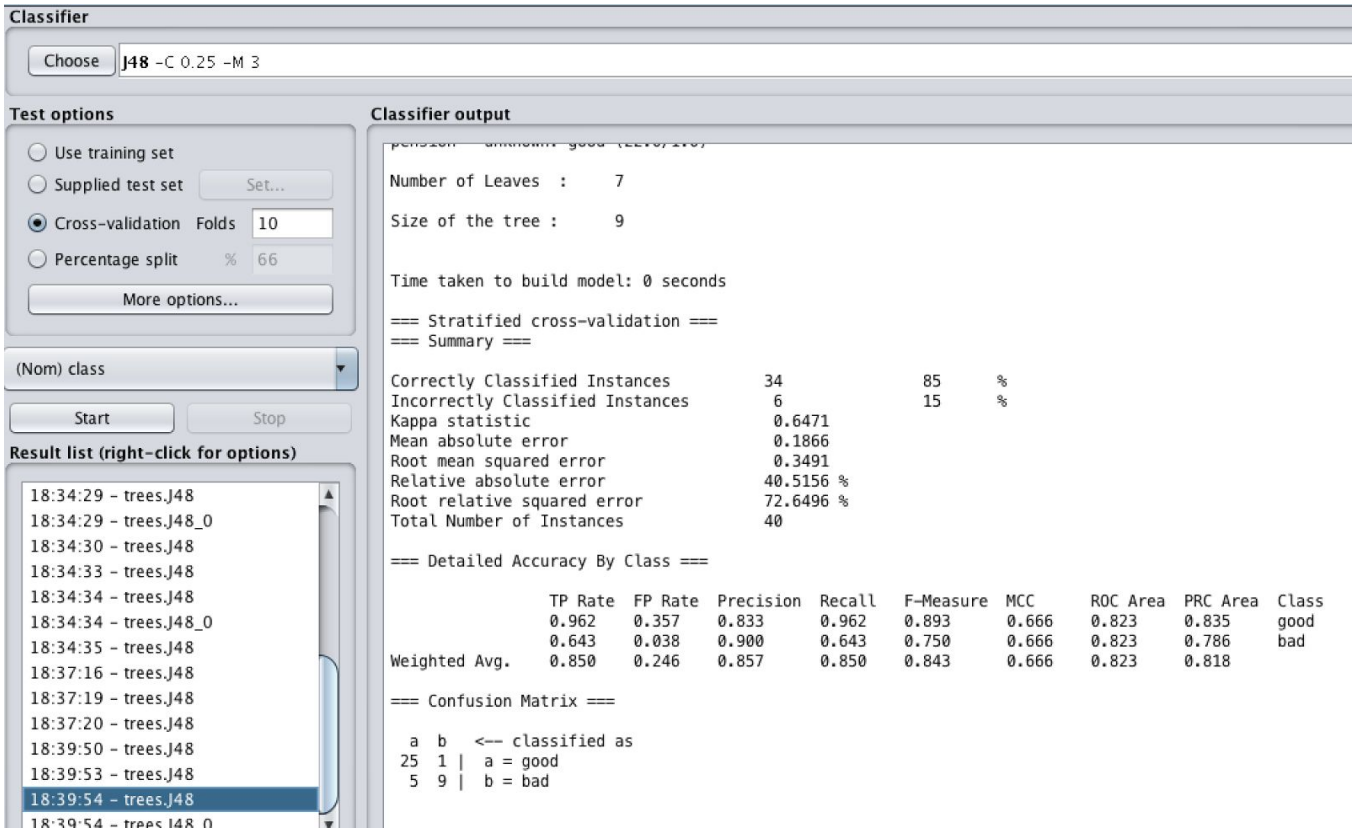
J48 w/10 fold c.v. #2
J48 10-fold
Appendix



J48 w/training set #2

J48 training

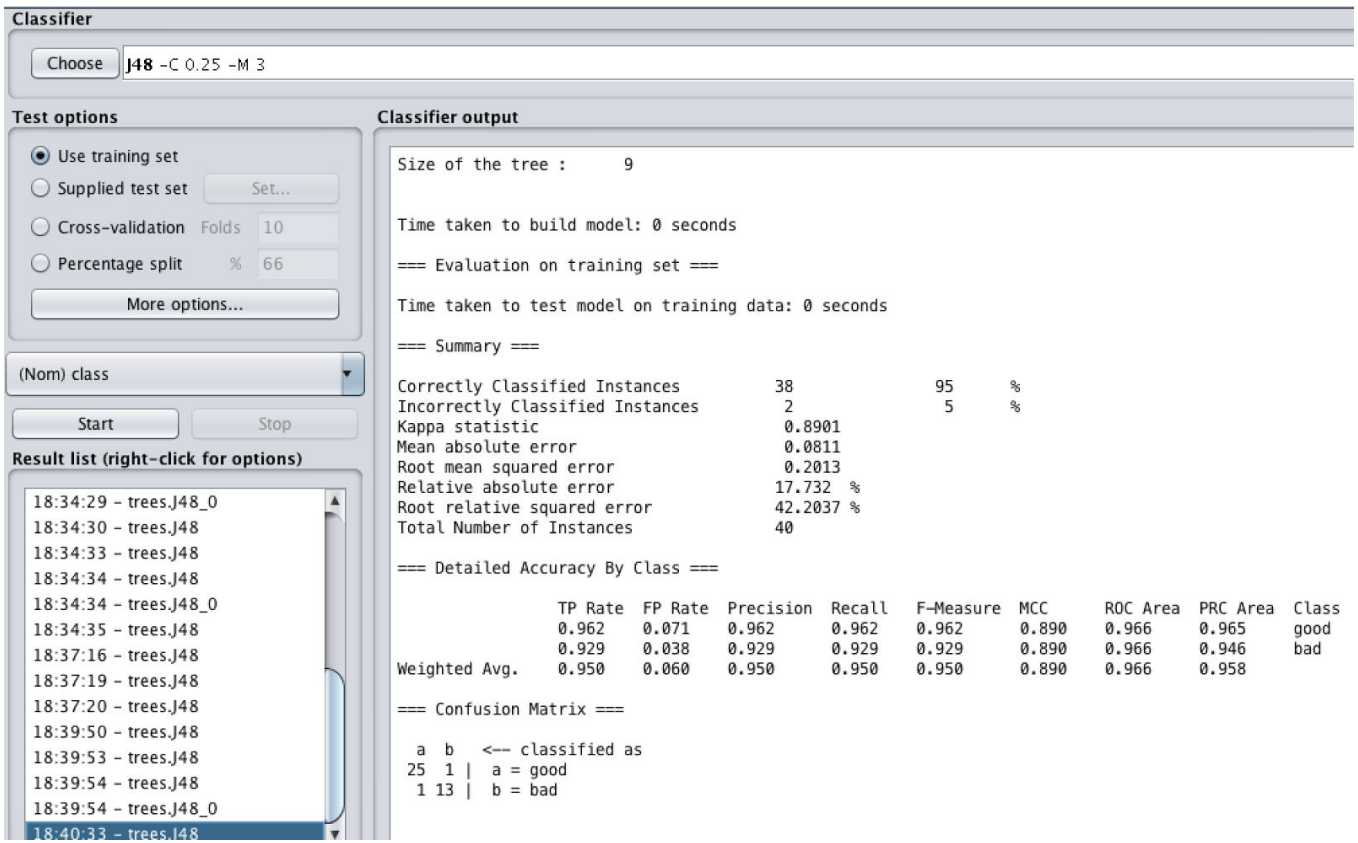
Appendix



J48 w/10-fold M=3 #2

J48 10-fold M=3

Appendix



J48 w/training M=3 #2

J48 training M=3

Appendix

A. What does the training set evaluation score tell you?

It tells you how well the model works when you use the same data to train and to test. This option use training set is often overly optimistic.

B. What does the cross-validation score evaluate?

For each data point in the data set is used for testing and 9 times for training, this gives 10 fold-cross validation.

C. What did you learn from the models about the data?

The pension attribute is the most important predictor.

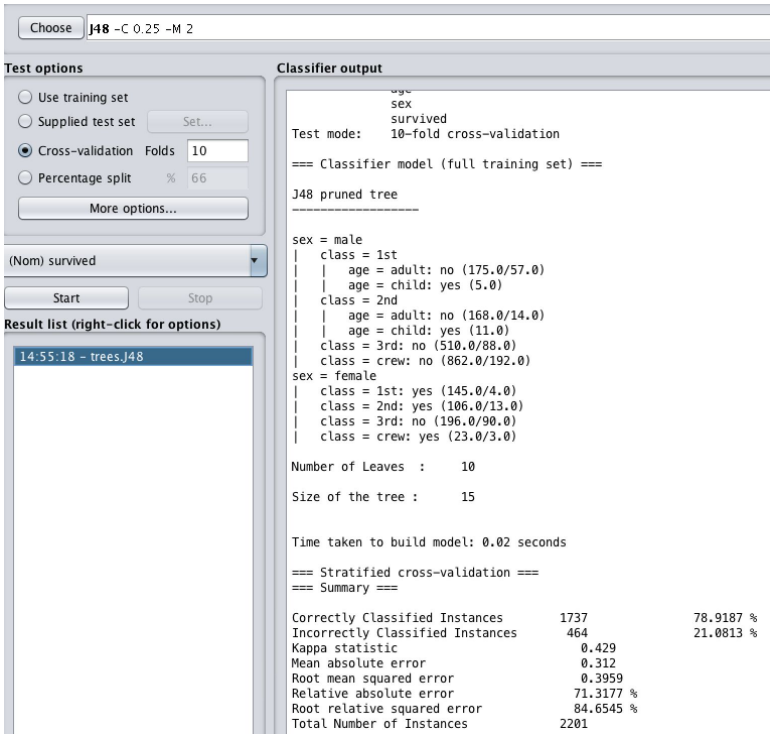
D. Which one of these models would you say is the best? Why?

I'd choose one of the models that used the cross-validation(since we are trying to predict in this case). The best model, would be the J48 M=3 cross-validation

3 Use the following learning schemes to analyze the Titanic data (in titanic.arff).

C4.5 - classifiers.j48.J48 & Assoc. Rules - association.apriori & Dec. List classifiers.PART

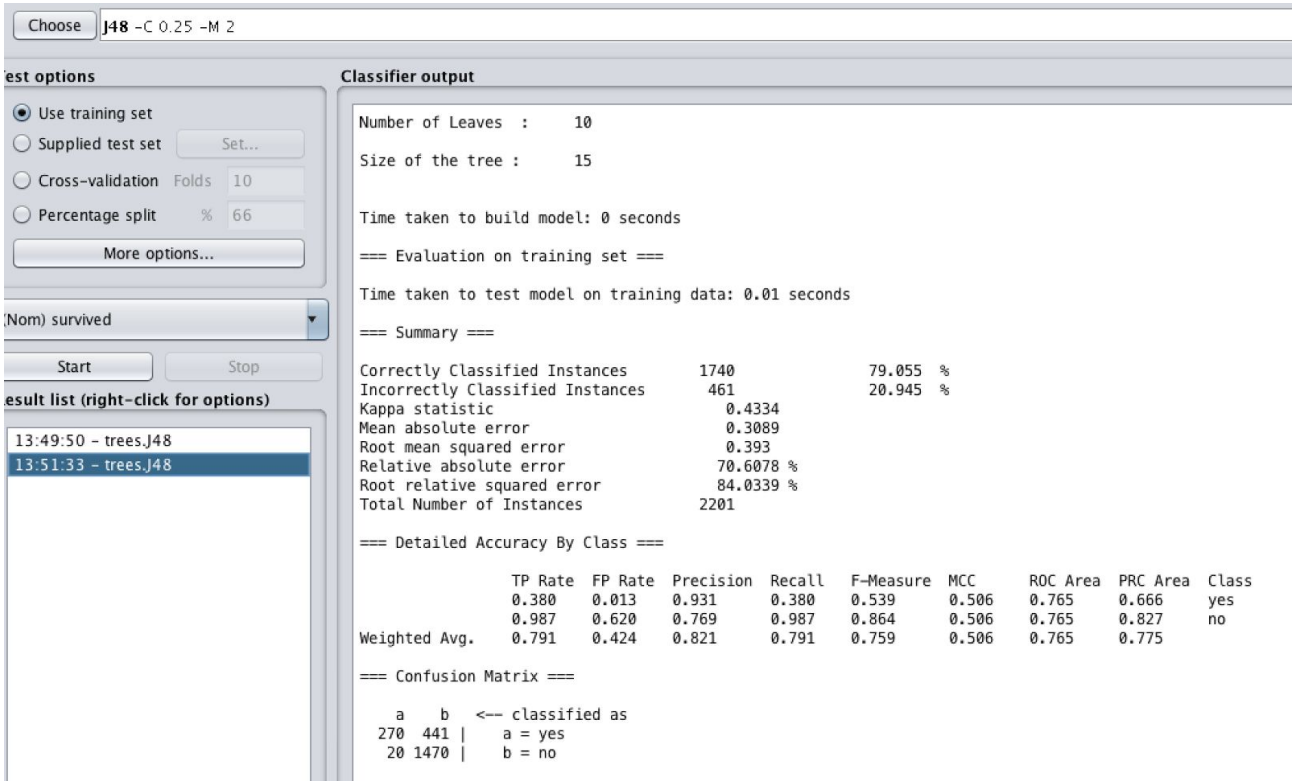
Home => Appendix



J48 >> Appendix

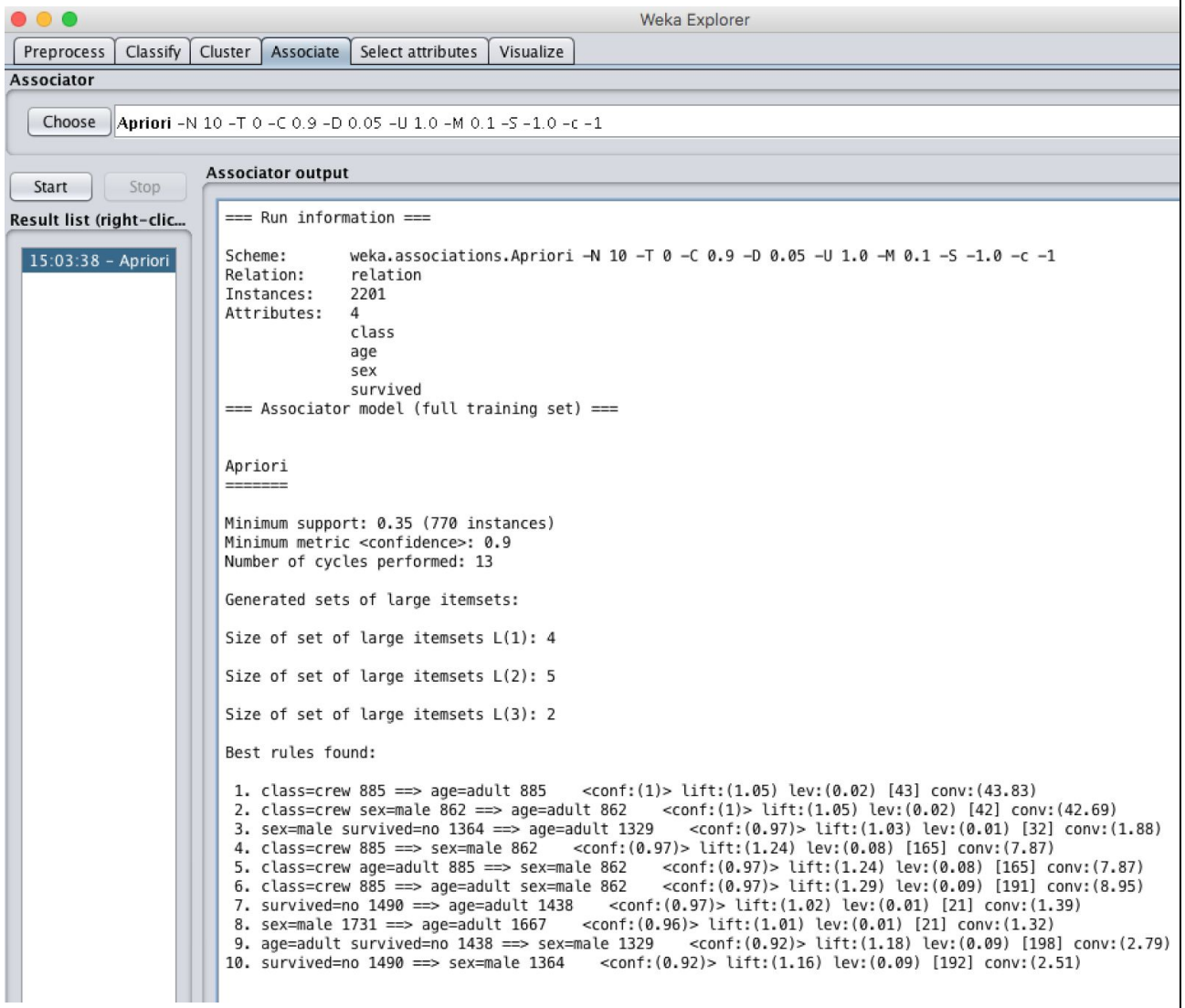
J48 returned a correctly classified value of 78.9%. J48 built a pruned tree with 10 leaves. From the root, the tree built branched to male/famale. From the male/female nodes, the branch was next to class(accomodations). The female branch had no node(s) for age.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.376    0.013    0.930      0.376    0.535      0.503    0.746     0.680     yes
               0.987    0.624    0.768      0.987    0.864      0.503    0.746     0.822     no
Weighted Avg.  0.789    0.427    0.820      0.789    0.758      0.503    0.746     0.777

=== Confusion Matrix ===

   a    b   <-- classified as
 267  444 |    a = yes
  20 1470 |    b = no
```



J48 w/Training Set

Appendix



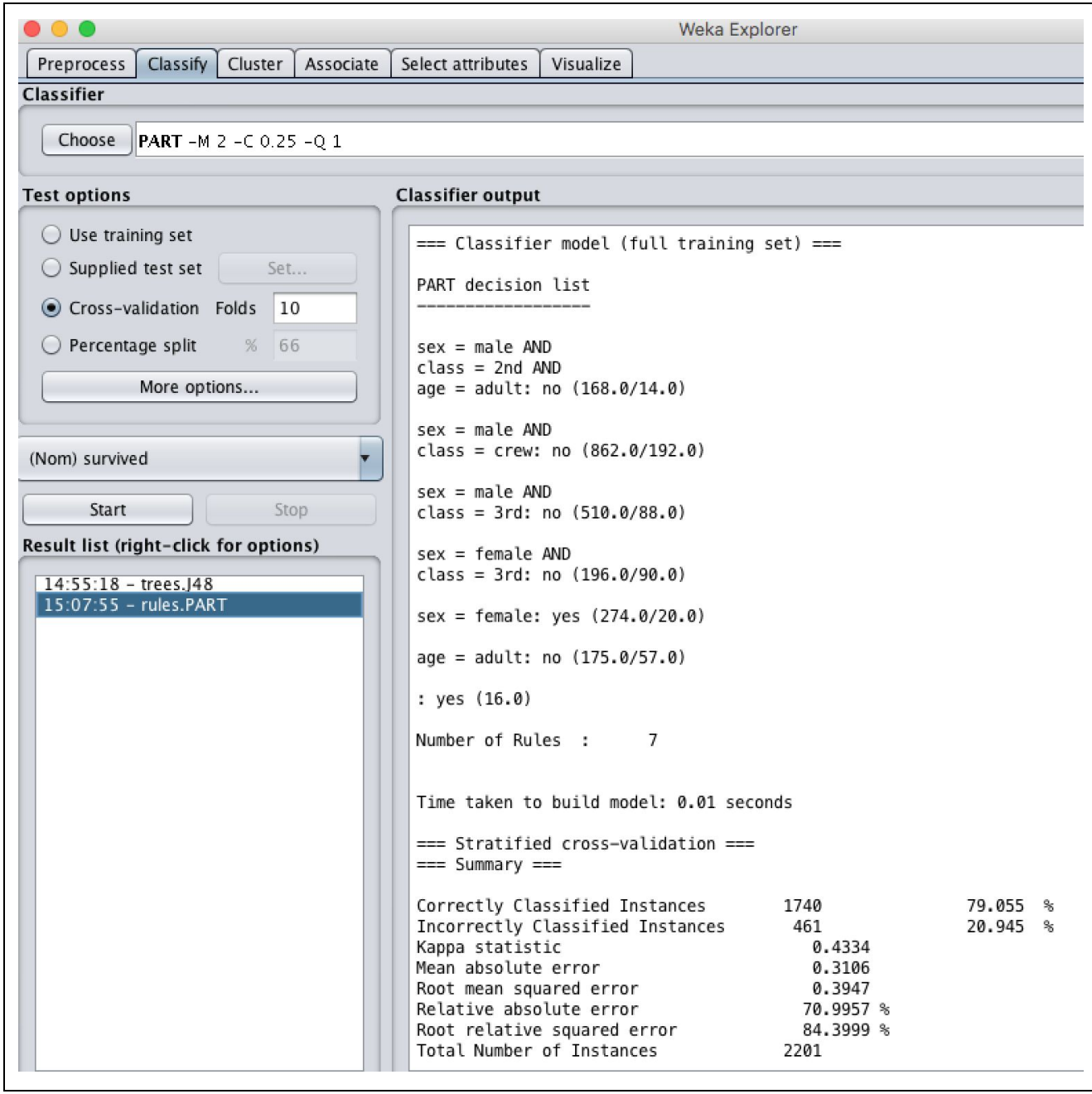
Apriori >> Appendix #3_Top

The Apriori classifier is a different kind of classifier. Even WEKA acknowledges this by giving it its own tab"Association". This type of classifier is used to find association rules in a data set.

When this algorithm comes up with the "best" rules, not all rules are very helpful. For example, the first "best rule" with confidence of 100% is:
If crew => adult
This says that a crew member is an adult, this rule adds no informtion. The Titanic did not have crew complement consisting of children.

give you a greater understanding of your dataset.

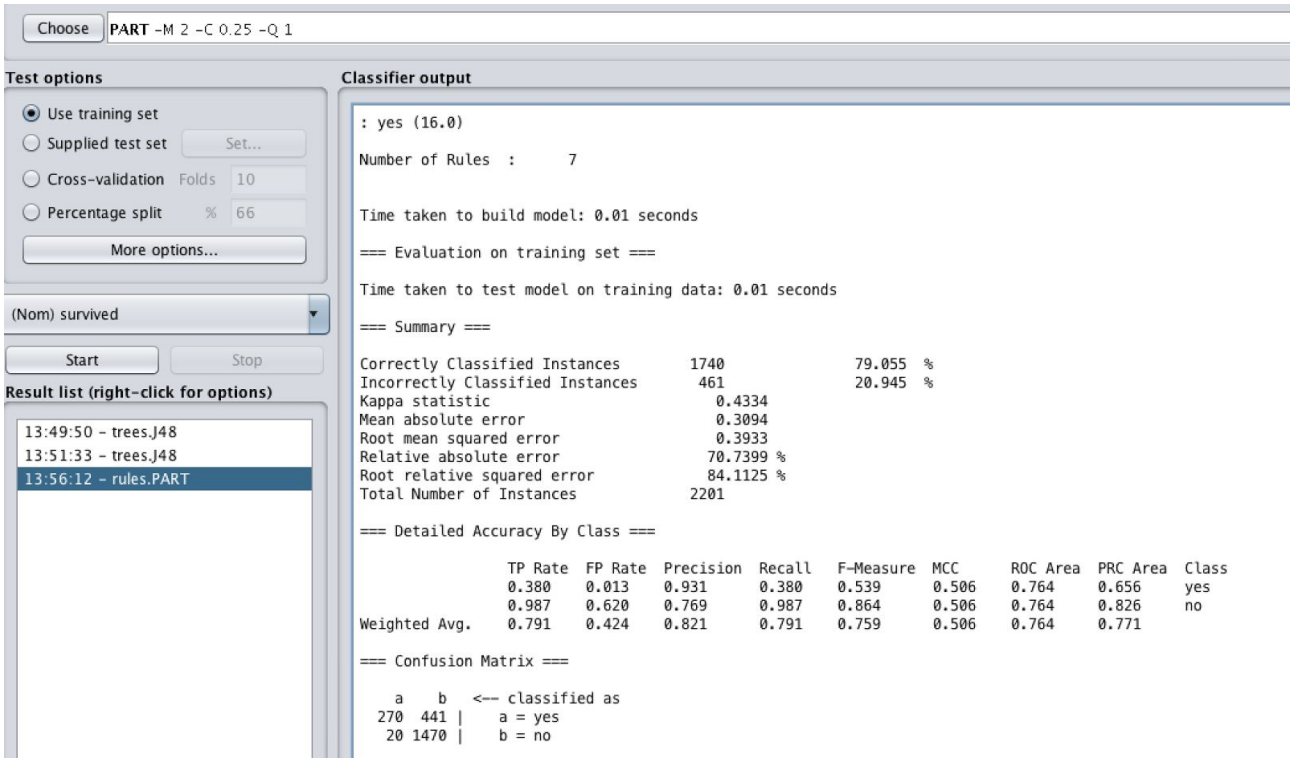
When using Apriori, I believe one should concentrate on rules that give you better insight into your data. A useful rule, I believe is: If male and didn’t survive ==> male (97%) Other rules, while interesting, don’t add too much: If male ==> adult If crew and male => adult	



PART >> Appendix #3_Top

Home3_PART

...



PART w/Training Set
Appendix
...

A. What is the most important descriptor (attribute) in titanic.arff, and how can you tell?
J48 first split criteria is “sex”, therefore I’d say “sex” is the most important descriptor.

B. How well were these methods able to learn the patterns in the dataset? Quantify your answer?
Apriori just finds associations in dataset, some patterns(even with high confidence) are not very useful.

C. Compare the training set and 10-fold cross-validations scores of the methods.

	Cross-Validation	Training Set
Part	79.05%	79.05%
J48	78.9%	79.05

D. Would you trust these models? Did they really learn what was important to survive the Titanic disaster?

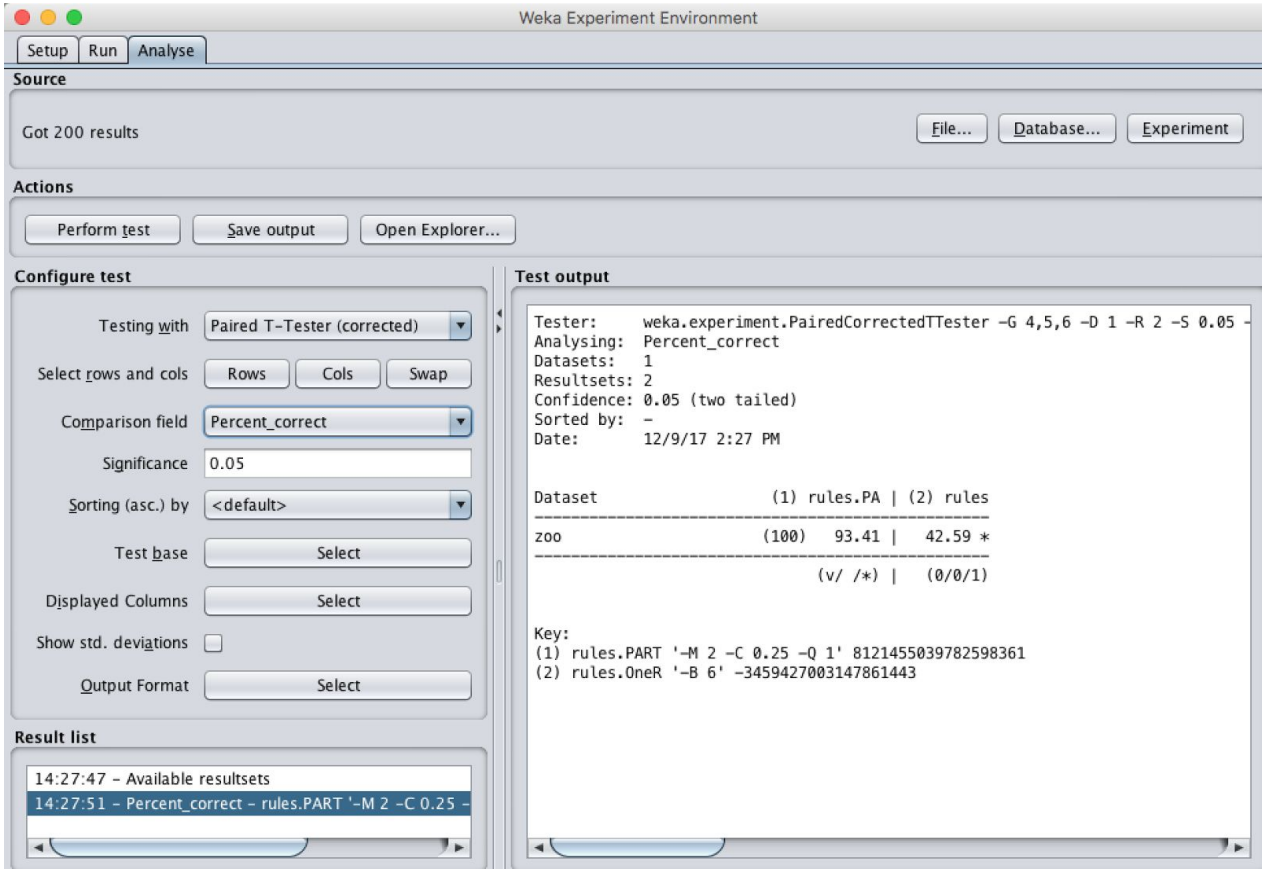
Yes they basically say that if you are male and crew member you are least likely to survive. If you were a male, you'd want to be in first class.

E. Which one would you trust more, even if just very slightly? Why?

I'd trust the J48 w/cross-validation. Even though it is not the highest classification, I understand the rules better. They are more straight forward. Besides, the difference between 78.9 and 79 is very slight.

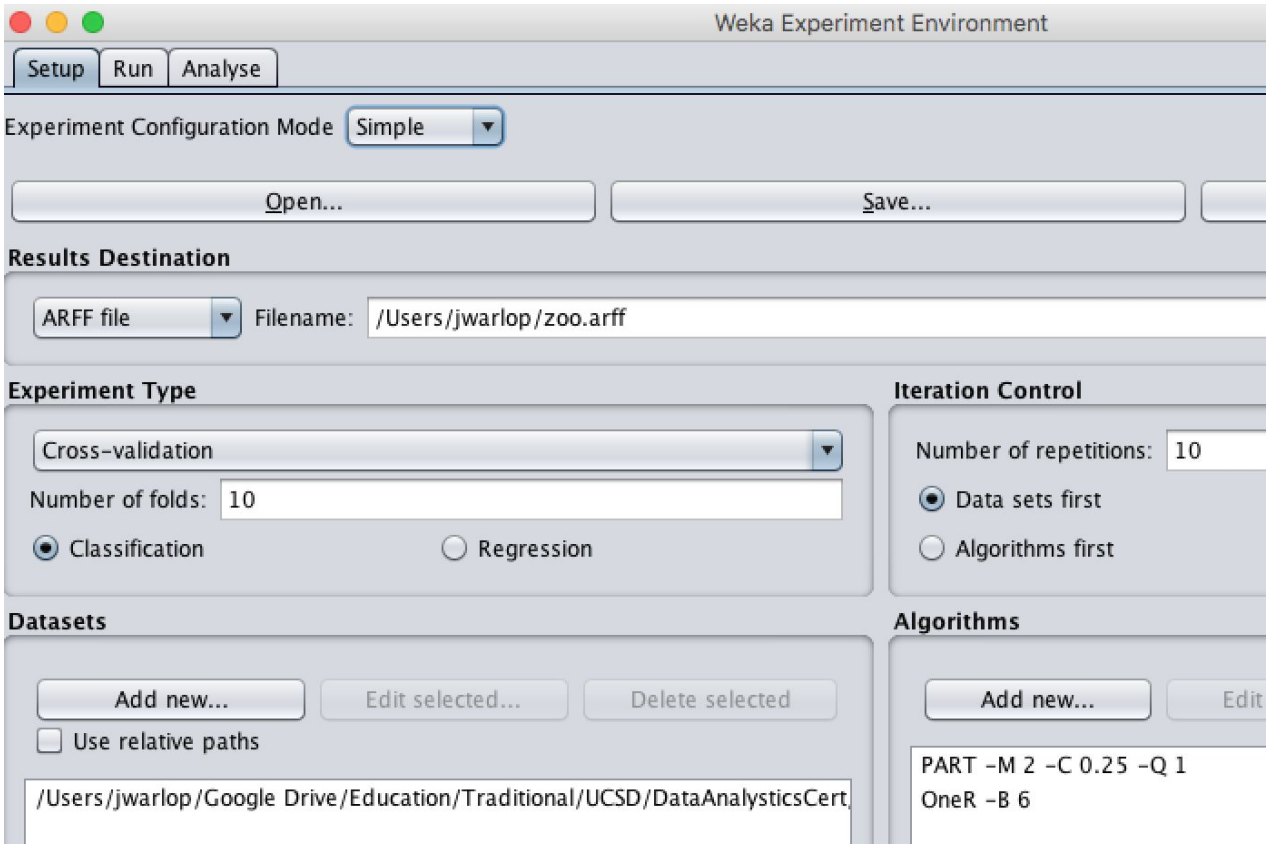
4 Choose one of the following three files: soybean.arff, autoprice.arff, hungarian.arff, zoo.arff or zoo2_x.arff and use any two schemas of your choice to build and compare the models. Evaluate and discuss the models. What was learned by the models (be specific to the dataset)? Which one of the models would you keep? Why?

Home => Appendix



PART >> Appendix
Home
I compared using the experimenter: One-R and PART

I used the experimenter for this test. Based on my experimentation, I'd keep the PART model. The PART model is more accurate and gives more insight into data. One-R is very basic and does not lead to a deeper understnading of the data.



One-R >> [Appendix #4_Top](#)

...

Appendix

1

Home

Question #1 Supporting Material

songs.arff

```
%
%
%
%
@RELATION songs

@ATTRIBUTE group string
```

```
@ATTRIBUTE song_name string
@ATTRIBUTE length numeric
@ATTRIBUTE genre {rock, new_age, country, alt_rock, jazz, pop}
@ATTRIBUTE decade {1940,1950,1960,1970,1980,1990,2000,2010}
@ATTRIBUTE lead {male, female, none}
@ATTRIBUTE horns {yes, no}
@ATTRIBUTE keyboard {yes, no}
@ATTRIBUTE flute_harmonica {yes, no}
@ATTRIBUTE strings {yes, no}
@ATTRIBUTE class {1,2,3,4,5}

@DATA
'neil young'           ,'down by the river'              ,557,rock    ,1960,male  ,no ,no ,no ,no ,5
'neil young'           ,'Ohio'                           ,179,rock    ,1960,male  ,no ,no ,no ,no ,5
'The B-52s'            ,'Love Shack'                     ,321,rock    ,1980,female,yes,yes,no ,no ,4
'Eric Clapton'         ,'Layla'                          ,426,rock    ,1970,male  ,no ,yes,no ,no ,5
'Crosby Stills Nash'   ,'Judy Blue Eyes'                 ,444,rock    ,1960,male  ,no ,no ,no ,no ,4
'Led Zepplin'          ,'Kashmir'                        ,508,rock    ,1970,male  ,yes,yes,no ,yes,5
'Led Zepplin'          ,'When the Levee Breaks'          ,430,rock    ,1970,male  ,no ,no ,yes,no ,5
'Peter Gabriel'        ,'Sledgehammer'                   ,296,rock    ,1980,male  ,yes,yes,yes,no ,5 %panned flute
'Bobby Darin'          ,'Beyond The Sea'                 ,168,jazz    ,1950,male  ,yes,yes,yes,no ,4
'Blind Melon'          ,'No Rain'                        ,217,rock    ,1990,male  ,no ,no ,no ,no ,5
'The Cure'             ,'Lullaby'                        ,253,alt_rock,1980,male  ,no,yes ,no ,yes,5
'Peter Gabriel'        ,'In Your Eyes'                   ,330,rock    ,1980,male  ,yes,yes,yes,yes,5
'Greta Van Fleet'      ,'Black Smoke Rising'             ,181,rock    ,2010,male  ,no ,no ,no ,no ,4
'The Beatles'          ,'Abbey Road'                     ,243,rock    ,1960,male  ,no ,yes,no ,yes,3
'Terry Clark'          ,'Youre easy on the eyes'         ,213,country ,2000,female,no ,no ,no ,yes,3
'Patsy Cline'          ,'Youre Stronger Than Me'         ,174,country ,1960,female,no ,no ,no ,yes,3
'Bob James'            ,'3 AM'                           ,330 ,jazz   ,1980,none  ,no ,yes,no ,yes,2
'Dave Weckl'           ,'7th Ave. South'                 ,360,jazz    ,1990,none  ,yes,yes,no ,yes,2
'D J Shadow'           ,'Fixed Income'                   ,290,jazz    ,2000,none  ,no ,yes,no ,yes,2
'Genesis'              ,'Follow You Follow Me'           ,240,rock    ,1970,male  ,no ,yes,no ,yes,4
'Joni Mitchell'        ,'Free Man in Paris'              ,244,pop     ,1970,female,no ,no ,no ,yes,3
'Diana Kroll'          ,'A Case of You'                  ,420,jazz    ,2000,female,no ,yes,no ,no ,2
'Michael Paulo'        ,'My Heart and Soul'              ,300,jazz    ,1990,none  ,yes,yes,no ,yes,3
'Sarah Vaughn'         ,'Star Dust'                      ,400,jazz    ,1950,female,yes,yes,no ,yes,3
'Loreena KcKennitt'    ,'Dantes Prayer'                  ,431,new_age ,1990,female,no ,yes,no ,yes,2
'The Fixx'             ,'Red Skies'                      ,276,rock    ,1980,male  ,no ,yes,no ,yes,4
```

```
'Frankie Valli'        ,'Sherry'                          ,153,rock    ,1960,male  ,no ,no ,no ,yes,4
'Gerry Rafferty'       ,'Baker Street'                    ,370,rock    ,1980,male  ,yes,yes,no ,yes,5
'Michael Jackson'      ,'Baby Be Mine'                    ,280,pop     ,1980,male  ,no ,yes,no ,no ,2
'Missing Persons'      ,'Destination Unknown'             ,212,pop     ,1980,female,no ,yes,no ,yes,3
'Big Country'          ,'In A Big Country'                ,285,rock    ,1980,male  ,no ,yes,no ,yes,5
'Bob Dylan'            ,'Hurricane'                       ,510,rock    ,1970,male  ,no ,yes,yes,yes,5
'Bryan Adams'          ,'Summer Of 69'                    ,214,rock    ,1980,male  ,no ,no ,no ,yes,5
'Smashing Pumpkins'    ,'1979'                            ,283,rock    ,1990,male  ,no ,no ,no ,yes,5
'Steely Dan'           ,'Reelin In The Years'             ,277,rock    ,1970,male  ,no ,no ,no ,yes,5
'Third Eye Blind'      ,'Semi-Charmed Life'               ,288,rock    ,1990,male  ,no ,no ,no ,yes,5
'Pixies'               ,'Where is my Mind'                ,240,rock    ,1980,male  ,no ,no ,no ,yes,5
'Gotye'                ,'Somebody That I Used To Know'    ,245,alt_rock,2010,male  ,no ,no ,no ,yes,4
'B52s'                 ,'Roam'                            ,240,rock    ,1980,female,no ,no ,no ,yes,5
'Sade'                 ,'Smooth Operator'                 ,260,jazz    ,1980,female,yes,no ,no ,yes,5
'The Motels'           ,'Only the Lonely'                 ,254,rock    ,1980,female,no ,no ,no ,yes,4
'Tom Petty'            ,'Dont Come Around Here No More'   ,280,rock    ,1980,male  ,no ,no ,no ,yes,5
'Garth Brooks'         ,'I Know One'                      ,175,country ,1980,male  ,no ,no ,no ,yes,5
'Miles Davis'          ,'Somethin Else'                   ,492,jazz    ,1950,none  ,yes,yes,no ,yes,5
'Miles Davis'          ,'Tomaas'                          ,336,jazz    ,1980,none  ,yes,yes,no ,yes,4
'Mark Isham'           ,'Raffles In Rio'                  ,280,new_age ,1980,none  ,no ,no ,no ,no ,2
'Canned Heat'          ,'On The Road Again'               ,253,rock    ,1960,male  ,no ,no ,no ,yes,4
'Cat Stevens'          ,'Wild World'                      ,221,pop     ,1970,male  ,no ,no ,no ,yes,4
'Duran Duran'          ,'Rio'                             ,339,pop     ,1980,male  ,yes,no ,no ,yes,4
'Elton John'           ,'Phildephia Freedom'              ,320,pop     ,1970,male  ,yes,yes,no ,yes,4
'The Monkees'          ,'I am a Believer'                 ,225,pop     ,1960,male  ,no ,yes,no ,yes,4
```

J48

Home#1-J48

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
```

```
Relation:     songs-weka.filters.unsupervised.attribute.Remove-R1-2
Instances:    51
Attributes:   9
              length
              genre
              decade
              lead
              horns
              keyboard
              flute_harmonica
              strings
              class
Test mode:    evaluate on training data

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

J48 pruned tree
------------------

genre = rock: 5 (27.0/9.0)
genre = new_age: 2 (2.0)
genre = country: 3 (3.0/1.0)
genre = alt_rock: 4 (2.0/1.0)
genre = jazz
|   horns = yes: 3 (7.0/5.0)
|   horns = no: 2 (3.0)
genre = pop
|   lead = male: 4 (5.0/1.0)
|   lead = female: 3 (2.0)
|   lead = none: 4 (0.0)

Number of Leaves  :   9
```

```
Size of the tree :    12


Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          34               66.6667 %
Incorrectly Classified Instances        17               33.3333 %
Kappa statistic                          0.5137
Mean absolute error                      0.17
Root mean squared error                  0.2915
Relative absolute error                 60.7333 %
Root relative squared error             78.4358 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         1
                 0.714    0.000    1.000      0.714    0.833      0.827    0.971     0.859     2
                 0.857    0.136    0.500      0.857    0.632      0.585    0.912     0.683     3
                 0.333    0.056    0.714      0.333    0.455      0.368    0.736     0.514     4
                 0.818    0.310    0.667      0.818    0.735      0.504    0.795     0.655     5
Weighted Avg.    0.667    0.169    0.704      0.667    0.652      0.519    0.818     0.645
```

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0  0  0  0 |  a = 1
  0  5  1  1  0 |  b = 2
  0  0  6  0  1 |  c = 3
  0  0  2  5  8 |  d = 4
  0  0  3  1 18 |  e = 5
```



```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     songs-weka.filters.unsupervised.attribute.Remove-R1-2
Instances:    51
Attributes:   9
              length
              genre
              decade
              lead
              horns
              keyboard
              flute_harmonica
              strings
              class
Test mode:    10-fold cross-validation
```

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

genre:
      rock  -> 5
      new_age    -> 2
      country    -> 3
      alt_rock   -> 4
      jazz  -> 2
      pop   -> 4
(31/51 instances correct)


Time taken to build model: 0 seconds

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

Correctly Classified Instances          30               58.8235 %
Incorrectly Classified Instances        21               41.1765 %
Kappa statistic                          0.3915
Mean absolute error                      0.1647
Root mean squared error                  0.4058
Relative absolute error                 58.5144 %
Root relative squared error            108.6612 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         1
                 0.857    0.136    0.500      0.857    0.632      0.585    0.860     0.448     2
```

```
                 0.286    0.023    0.667      0.286    0.400      0.385    0.631     0.289     3
                 0.267    0.111    0.500      0.267    0.348      0.195    0.578     0.349     4
                 0.818    0.345    0.643      0.818    0.720      0.471    0.737     0.604     5
Weighted Avg.    0.588    0.203    0.585      0.588    0.554      0.394    0.692     0.464

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0  0  0  0 |  a = 1
  0  6  0  1  0 |  b = 2
  0  2  2  2  1 |  c = 3
  0  2  0  4  9 |  d = 4
  0  2  1  1 18 |  e = 5
```

PART

Home

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation:     songs-weka.filters.unsupervised.attribute.Remove-R1-2
Instances:    51
Attributes:   9
              length
              genre
              decade
              lead
              horns
              keyboard
              flute_harmonica
```

```
              strings
              class
Test mode:    10-fold cross-validation

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

PART decision list
------------------

genre = rock AND
flute_harmonica = no AND
decade = 1980 AND
lead = male: 5 (6.0/1.0)

genre = rock AND
flute_harmonica = yes: 5 (4.0)

genre = rock AND
decade = 1970: 5 (4.0/1.0)

genre = rock AND
keyboard = no AND
decade = 1960 AND
strings = no: 5 (3.0/1.0)

genre = rock AND
keyboard = no AND
decade = 1990: 5 (3.0)

decade = 1960: 4 (5.0/2.0)

decade = 1950: 3 (3.0/2.0)
```

```
decade = 1970: 4 (3.0/1.0)

length > 260 AND
horns = no: 2 (6.0)

decade = 1980 AND
length > 253: 4 (5.0/1.0)

decade = 1980: 5 (4.0/1.0)

decade = 1990: 2 (2.0/1.0)

: 4 (3.0/1.0)

Number of Rules  :    13


Time taken to build model: 0 seconds

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

Correctly Classified Instances          22               43.1373 %
Incorrectly Classified Instances        29               56.8627 %
Kappa statistic                          0.1686
Mean absolute error                      0.2219
Root mean squared error                  0.3885
Relative absolute error                 78.8457 %
Root relative squared error            104.0152 %
Total Number of Instances               51
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?         1
                 0.429    0.068    0.500      0.429    0.462      0.385    0.752     0.501     2
                 0.286    0.114    0.286      0.286    0.286      0.172    0.747     0.428     3
                 0.200    0.333    0.200      0.200    0.200      -0.133   0.489     0.325     4
                 0.636    0.310    0.609      0.636    0.622      0.324    0.707     0.572     5
Weighted Avg.    0.431    0.257    0.429      0.431    0.430      0.177    0.654     0.470

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0  0  0  0 |  a = 1
  0  3  2  2  0 |  b = 2
  0  1  2  4  0 |  c = 3
  0  1  2  3  9 |  d = 4
  0  1  1  6 14 |  e = 5
```

2

Home

Question #2 Supporting Material

IBk w/10-fold cross validation

Home

```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     labor-neg-nominal
Instances:    40
Attributes:   17
              duration
              wage increase first year
              wage increase second year
              wage increase third year
              cost of living adjustment
              working hours
              pension
              standby pay
              shift differential
              education allowance
              statutory holidays
              vacation
              longterm disability assistance
              contribution to dental plan
              bereavement assistance
              contribution to health plan
              class
Test mode:    10-fold cross-validation

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
```

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances          39               97.5    %
Incorrectly Classified Instances         1                2.5    %
Kappa statistic                          0.9459
Mean absolute error                      0.0756
Root mean squared error                  0.1851
Relative absolute error                 16.4215 %
Root relative squared error             38.5153 %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.962    0.000    1.000      0.962    0.980      0.947    0.985     0.989     good
                 1.000    0.038    0.933      1.000    0.966      0.947    0.985     0.949     bad
Weighted Avg.    0.975    0.013    0.977      0.975    0.975      0.947    0.985     0.975

=== Confusion Matrix ===

  a  b   <-- classified as
 25  1 |  a = good
  0 14 |  b = bad
```

IBk w/training set

Home

```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     labor-neg-nominal
Instances:    40
Attributes:   17
              duration
              wage increase first year
              wage increase second year
              wage increase third year
              cost of living adjustment
              working hours
              pension
              standby pay
              shift differential
              education allowance
              statutory holidays
              vacation
              longterm disability assistance
              contribution to dental plan
              bereavement assistance
              contribution to health plan
              class
Test mode:    evaluate on training data

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
```

```
Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          40               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0238
Root mean squared error                  0.0238
Relative absolute error                  5.2083 %
Root relative squared error              4.9913 %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     good
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     bad
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 26  0 |  a = good
  0 14 |  b = bad
```

J48 w/10 fold cross validaiton

Home

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     labor-neg-nominal
Instances:    40
Attributes:   17
              duration
              wage increase first year
              wage increase second year
              wage increase third year
              cost of living adjustment
              working hours
              pension
              standby pay
              shift differential
              education allowance
              statutory holidays
              vacation
              longterm disability assistance
              contribution to dental plan
              bereavement assistance
              contribution to health plan
              class
Test mode:    10-fold cross-validation

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

J48 pruned tree
------------------

pension = none: bad (8.0)
pension = ret_allw: bad (3.0/1.0)
```

```
pension = empl_contr
|   wage increase first year = low: bad (3.0)
|   wage increase first year = medium: good (3.0)
|   wage increase first year = high: good (0.0)
|   wage increase first year = unknown: good (1.0)
pension = unknown: good (22.0/1.0)

Number of Leaves  :   7

Size of the tree :    9


Time taken to build model: 0 seconds

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

Correctly Classified Instances          37               92.5    %
Incorrectly Classified Instances         3                7.5    %
Kappa statistic                          0.8324
Mean absolute error                      0.0991
Root mean squared error                  0.2524
Relative absolute error                 21.5156 %
Root relative squared error             52.5347 %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.962    0.143    0.926      0.962    0.943      0.834    0.915     0.913     good
                 0.857    0.038    0.923      0.857    0.889      0.834    0.915     0.878     bad
```

```
Weighted Avg.    0.925    0.106    0.925      0.925    0.924      0.834    0.915     0.901

=== Confusion Matrix ===

  a  b   <-- classified as
 25  1 |  a = good
  2 12 |  b = bad
```

J48 w/training set

Home

```
=== Run information ===
```

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     labor-neg-nominal
Instances:    40
Attributes:   17
              duration
              wage increase first year
              wage increase second year
              wage increase third year
              cost of living adjustment
              working hours
              pension
              standby pay
              shift differential
              education allowance
              statutory holidays
              vacation
              longterm disability assistance
              contribution to dental plan
              bereavement assistance
              contribution to health plan
              class
Test mode:    evaluate on training data

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

J48 pruned tree
------------------

pension = none: bad (8.0)
pension = ret_allw: bad (3.0/1.0)
pension = empl_contr
```

```
|   wage increase first year = low: bad (3.0)
|   wage increase first year = medium: good (3.0)
|   wage increase first year = high: good (0.0)
|   wage increase first year = unknown: good (1.0)
pension = unknown: good (22.0/1.0)

Number of Leaves  :   7

Size of the tree :    9


Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          38               95      %
Incorrectly Classified Instances         2                5      %
Kappa statistic                          0.8901
Mean absolute error                      0.0811
Root mean squared error                  0.2013
Relative absolute error                 17.732  %
Root relative squared error             42.2037 %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
```

```
                 0.962    0.071    0.962      0.962    0.962      0.890    0.966     0.965     good
                 0.929    0.038    0.929      0.929    0.929      0.890    0.966     0.946     bad
Weighted Avg.    0.950    0.060    0.950      0.950    0.950      0.890    0.966     0.958

=== Confusion Matrix ===

  a  b   <-- classified as
 25  1 |  a = good
  1 13 |  b = bad
```

J48 w/10-fold cross validation and M=3

Home

```
=== Run information ===
```

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 3
Relation:     labor-neg-nominal
Instances:    40
Attributes:   17
              duration
              wage increase first year
              wage increase second year
              wage increase third year
              cost of living adjustment
              working hours
              pension
              standby pay
              shift differential
              education allowance
              statutory holidays
              vacation
              longterm disability assistance
              contribution to dental plan
              bereavement assistance
              contribution to health plan
              class
Test mode:    10-fold cross-validation

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

J48 pruned tree
------------------

pension = none: bad (8.0)
pension = ret_allw: bad (3.0/1.0)
pension = empl_contr
|   wage increase first year = low: bad (3.0)
```

```
|   wage increase first year = medium: good (3.0)
|   wage increase first year = high: good (0.0)
|   wage increase first year = unknown: good (1.0)
pension = unknown: good (22.0/1.0)

Number of Leaves  :   7

Size of the tree :    9


Time taken to build model: 0 seconds

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

Correctly Classified Instances          34               85      %
Incorrectly Classified Instances         6               15      %
Kappa statistic                          0.6471
Mean absolute error                      0.1866
Root mean squared error                  0.3491
Relative absolute error                 40.5156 %
Root relative squared error             72.6496 %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.962    0.357    0.833      0.962    0.893      0.666    0.823     0.835     good
                 0.643    0.038    0.900      0.643    0.750      0.666    0.823     0.786     bad
Weighted Avg.    0.850    0.246    0.857      0.850    0.843      0.666    0.823     0.818
```

```
=== Confusion Matrix ===

  a  b   <-- classified as
 25  1 |  a = good
  5  9 |  b = bad
```

J48 w/training set and M=3

Home

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 3
```

```
Relation:     labor-neg-nominal
Instances:    40
Attributes:   17
              duration
              wage increase first year
              wage increase second year
              wage increase third year
              cost of living adjustment
              working hours
              pension
              standby pay
              shift differential
              education allowance
              statutory holidays
              vacation
              longterm disability assistance
              contribution to dental plan
              bereavement assistance
              contribution to health plan
              class
Test mode:    evaluate on training data

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

J48 pruned tree
------------------

pension = none: bad (8.0)
pension = ret_allw: bad (3.0/1.0)
pension = empl_contr
|   wage increase first year = low: bad (3.0)
|   wage increase first year = medium: good (3.0)
```

```
|   wage increase first year = high: good (0.0)
|   wage increase first year = unknown: good (1.0)
pension = unknown: good (22.0/1.0)

Number of Leaves  :   7

Size of the tree :    9


Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          38               95      %
Incorrectly Classified Instances         2                5      %
Kappa statistic                          0.8901
Mean absolute error                      0.0811
Root mean squared error                  0.2013
Relative absolute error                 17.732  %
Root relative squared error             42.2037 %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.962    0.071    0.962      0.962    0.962      0.890    0.966     0.965     good
                 0.929    0.038    0.929      0.929    0.929      0.890    0.966     0.946     bad
```

```
Weighted Avg.    0.950    0.060    0.950      0.950    0.950      0.890    0.966     0.958

=== Confusion Matrix ===

  a  b   <-- classified as
 25  1 |  a = good
  1 13 |  b = bad
```

LABOR-NEG-NORMAL-V2.ARFF

```
%Date: Tue, 15 Nov 88 15:44:08 EST
%From: stan <stan@csi2.UofO.EDU>
%Message-Id: <8811152044.AA23067@csih.UofO.EDU>
%To: aha@ICS.UCI.EDU
%
%1. Title: Final settlements in labor negotitions in Canadian industry
%
%2. Source Information
%   -- Creators: Collective Barganing Review, montly publication,
%      Labour Canada, Industrial Relations Information Service,
%     Ottawa, Ontario, K1A 0J2, Canada, (819) 997-3117
%     The data includes all collective agreements reached
%        in the business and personal services sector for locals
%        with at least 500 members (teachers, nurses, university
%        staff, police, etc) in Canada in 87 and first quarter of 88.
%   -- Donor: Stan Matwin, Computer Science Dept, University of Ottawa,
%               34 Somerset East, K1N 9B4, (stan@uotcsi2.bitnet)
%   -- Date: November 1988
%
%3. Past Usage:
%   -- testing concept learning software, in particular
%      an experimental method to learn two-tiered concept descriptions.
```

```
%      The data was used to learn the description of an acceptable
%      and unacceptable contract.
%      The unacceptable contracts were either obtained by interviewing
%      experts, or by inventing near misses.
%      Examples of use are described in:
%     Bergadano, F., Matwin, S., Michalski, R.,
%     Zhang, J., Measuring Quality of Concept Descriptions,
%     Procs. of the 3rd European Working Sessions on Learning,
%     Glasgow, October 1988.
%     Bergadano, F., Matwin, S., Michalski, R., Zhang, J.,
%       Representing and Acquiring Imprecise and Context-dependent
%     Concepts in Knowledge-based Systems, Procs. of ISMIS'88,
%     North Holland, 1988.
%4. Relevant Information:
%   -- data was used to test 2tier approach with learning
%from positive and negative examples
%
%5. Number of Instances: 57
%
%6. Number of Attributes: 16
%
%7. Attribute Information:
%   1.  dur: duration of agreement
%       [1..7]
%   2   wage1.wage : wage increase in first year of contract
%       [2.0 .. 7.0]
%   3   wage2.wage : wage increase in second year of contract
%       [2.0 .. 7.0]
%   4   wage3.wage : wage increase in third year of contract
%       [2.0 .. 7.0]
%   5   cola : cost of living allowance
%       [none, tcf, tc]
%   6   hours.hrs : number of working hours during week
%       [35 .. 40]
%   7   pension : employer contributions to pension plan
%       [none, ret_allw, empl_contr]
%   8   stby_pay : standby pay
%       [2 .. 25]
```

```
%   9   shift_diff : shift differencial : supplement for work on II and III shift
%       [1 .. 25]
%  10   educ_allw.boolean : education allowance
%       [yes no]
%  11   holidays : number of statutory holidays
%       [9 .. 15]
%  12   vacation : number of paid vacation days
%       [ba, avg, gnr]
%  13   lngtrm_disabil.boolean :
%       employer's help during employee longterm disabil
%       ity [yes , no]
%  14   dntl_ins : employers contribution towards the dental plan
%       [none, half, full]
%  15   bereavement.boolean : employer's financial contribution towards the
%       covering the costs of bereavement
%       [yes , no]
%  16   empl_hplan : employer's contribution towards the health plan
%       [none, half, full]
%
%8. Missing Attribute Values: None
%
%9. Class Distribution:
%
%10. Exceptions from format instructions: no commas between attribute values.
%
%------------------------------------------------------------------------------


@relation labor-neg-nominal

%  Classes
%  -------

% good, bad.

%  Attributes
%  ----------
```

```
@attribute duration { 1, 2, 3 ,unknown }
@attribute "wage increase first year" { low, medium, high ,unknown }
@attribute "wage increase second year" { low, medium, high ,unknown }
@attribute "wage increase third year" { low, medium, high ,unknown }
@attribute "cost of living adjustment"{ none, tcf, tc ,unknown }
@attribute "working hours" { sub35, sub40, equal40 ,unknown }
@attribute pension { none, ret_allw, empl_contr ,unknown }
@attribute "standby pay" { 2, 4, 8, 12, 13 ,unknown }
@attribute "shift differential"  { 0, 1, 2, 3, 4, 5,6, 10, 11, 25 ,unknown }
@attribute "education allowance" {yes, no ,unknown }
@attribute "statutory holidays"  { 9, 10, 11, 12,13,15  ,unknown }
@attribute vacation { "below average", average, generous ,unknown }
@attribute "longterm disability assistance" { yes, no ,unknown }
@attribute "contribution to dental plan" { none, half, full ,unknown }
@attribute "bereavement assistance" { yes, no ,unknown }
@attribute "contribution to health plan" { none, half, full ,unknown }
@attribute class { good, bad }

@data

1,medium,unknown,unknown,unknown,equal40,unknown,unknown,2,unknown,11,average,unknown,unknown,yes,unknown,good
2,medium,medium,unknown,unknown,sub40,ret_allw,unknown,unknown,yes,11,"below average",unknown,full,unknown,full,good
...........
..............
..................
```

3	Home Question #3 Supporting Material TITANIC.ARFF @RELATION relation @ATTRIBUTE class {"1st","2nd","3rd","crew"} @ATTRIBUTE age {"adult","child"} @ATTRIBUTE sex {"male","female"} @ATTRIBUTE survived {"yes","no"} @DATA 1st,adult,male,yes 1st,adult,male,yes 1st,adult,male,yes ..... crew,adult,female,yes crew,adult,female,yes crew,adult,female,yes crew,adult,female,yes crew,adult,female,yes crew,adult,female,yes

```
crew,adult,female,yes
crew,adult,female,yes
crew,adult,female,no
.....
........
..........
```

J48

NAME
weka.classifiers.trees.J48

SYNOPSIS
Class for generating a pruned or unpruned C4.5 decision tree. For more information, see

Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

OPTIONS
seed -- The seed used for randomizing the data when reduced-error pruning is used.

unpruned -- Whether pruning is performed.

confidenceFactor -- The confidence factor used for pruning (smaller values incur more pruning).

numFolds -- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.

numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model.

batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size.

reducedErrorPruning -- Whether reduced-error pruning is used instead of C.4.5 pruning.

useLaplace -- Whether counts at leaves are smoothed based on Laplace.

doNotMakeSplitPointActualValue -- If true, the split point is not relocated to an actual data

value. This can yield substantial speed-ups for large datasets with numeric attributes.

debug -- If set to true, classifier may output additional info to the console.

subtreeRaising -- Whether to consider the subtree raising operation when pruning.

saveInstanceData -- Whether to save the training data for visualization.

binarySplits -- Whether to use binary splits on nominal attributes when building the trees.

doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime).

minNumObj -- The minimum number of instances per leaf.

useMDLcorrection -- Whether MDL correction is used when finding splits on numeric attributes.

collapseTree -- Whether parts are removed that do not reduce training error.

J48

Home

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     relation
Instances:    2201
Attributes:   4
              class
              age
              sex
              survived
Test mode:    10-fold cross-validation

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

J48 pruned tree
------------------

sex = male
|   class = 1st
|   |   age = adult: no (175.0/57.0)
|   |   age = child: yes (5.0)
|   class = 2nd
|   |   age = adult: no (168.0/14.0)
|   |   age = child: yes (11.0)
|   class = 3rd: no (510.0/88.0)
|   class = crew: no (862.0/192.0)
sex = female
|   class = 1st: yes (145.0/4.0)
|   class = 2nd: yes (106.0/13.0)
```

```
|   class = 3rd: no (196.0/90.0)
|   class = crew: yes (23.0/3.0)

Number of Leaves  :   10

Size of the tree :    15


Time taken to build model: 0.02 seconds

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

Correctly Classified Instances        1737               78.9187 %
Incorrectly Classified Instances       464               21.0813 %
Kappa statistic                          0.429
Mean absolute error                      0.312
Root mean squared error                  0.3959
Relative absolute error                 71.3177 %
Root relative squared error             84.6545 %
Total Number of Instances             2201

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.376    0.013    0.930      0.376    0.535      0.503    0.746     0.680     yes
                 0.987    0.624    0.768      0.987    0.864      0.503    0.746     0.822     no
Weighted Avg.    0.789    0.427    0.820      0.789    0.758      0.503    0.746     0.777

=== Confusion Matrix ===
```

```
    a    b   <-- classified as
  267  444 |    a = yes
   20 1470 |    b = no
```

J48 w/Training Set

Home

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     relation
Instances:    2201
Attributes:   4
              class
              age
              sex
              survived
Test mode:    evaluate on training data

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

J48 pruned tree
------------------

sex = male
|   class = 1st
|   |   age = adult: no (175.0/57.0)
|   |   age = child: yes (5.0)
```

```
|   class = 2nd
|   |   age = adult: no (168.0/14.0)
|   |   age = child: yes (11.0)
|   class = 3rd: no (510.0/88.0)
|   class = crew: no (862.0/192.0)
sex = female
|   class = 1st: yes (145.0/4.0)
|   class = 2nd: yes (106.0/13.0)
|   class = 3rd: no (196.0/90.0)
|   class = crew: yes (23.0/3.0)

Number of Leaves  :   10

Size of the tree :    15


Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances        1740               79.055  %
Incorrectly Classified Instances       461               20.945  %
Kappa statistic                          0.4334
Mean absolute error                      0.3089
Root mean squared error                  0.393
Relative absolute error                 70.6078 %
Root relative squared error             84.0339 %
Total Number of Instances             2201
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.380    0.013    0.931      0.380    0.539      0.506    0.765     0.666     yes
                 0.987    0.620    0.769      0.987    0.864      0.506    0.765     0.827     no
Weighted Avg.    0.791    0.424    0.821      0.791    0.759      0.506    0.765     0.775

=== Confusion Matrix ===

    a    b   <-- classified as
  270  441 |    a = yes
   20 1470 |    b = no
```

Apriori

Home

```
=== Run information ===
```

```
Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     relation
Instances:    2201
Attributes:   4
              class
              age
              sex
              survived
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.35 (770 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 2

Best rules found:

 1. class=crew 885 ==> age=adult 885    <conf:(1)> lift:(1.05) lev:(0.02) [43] conv:(43.83)
 2. class=crew sex=male 862 ==> age=adult 862    <conf:(1)> lift:(1.05) lev:(0.02) [42] conv:(42.69)
 3. sex=male survived=no 1364 ==> age=adult 1329    <conf:(0.97)> lift:(1.03) lev:(0.01) [32] conv:(1.88)
 4. class=crew 885 ==> sex=male 862    <conf:(0.97)> lift:(1.24) lev:(0.08) [165] conv:(7.87)
 5. class=crew age=adult 885 ==> sex=male 862    <conf:(0.97)> lift:(1.24) lev:(0.08) [165] conv:(7.87)
 6. class=crew 885 ==> age=adult sex=male 862    <conf:(0.97)> lift:(1.29) lev:(0.09) [191] conv:(8.95)
 7. survived=no 1490 ==> age=adult 1438    <conf:(0.97)> lift:(1.02) lev:(0.01) [21] conv:(1.39)
 8. sex=male 1731 ==> age=adult 1667    <conf:(0.96)> lift:(1.01) lev:(0.01) [21] conv:(1.32)
 9. age=adult survived=no 1438 ==> sex=male 1329    <conf:(0.92)> lift:(1.18) lev:(0.09) [198] conv:(2.79)
10. survived=no 1490 ==> sex=male 1364    <conf:(0.92)> lift:(1.16) lev:(0.09) [192] conv:(2.51)
```

Classifier/Association	J48	PART	Apriori
Correctly Classified or Highest Confidence	78.9	79.1	1.0

J48 is a tree based classifier and PART is a rules based classifier. Apriori is a special case classifier.

weka.classifiers.rules.PART

NAME
weka.classifiers.rules.PART

SYNOPSIS
Class for generating a PART decision list. Uses separate-and-conquer. Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule.

For more information, see:

Eibe Frank, Ian H. Witten: Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, 144-151, 1998.

OPTIONS
seed -- The seed used for randomizing the data when reduced-error pruning is used.

unpruned -- Whether pruning is performed.

confidenceFactor -- The confidence factor used for pruning (smaller values incur more pruning).

numFolds -- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the rules.

numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model.

batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size.

reducedErrorPruning -- Whether reduced-error pruning is used instead of C.4.5 pruning.

doNotMakeSplitPointActualValue -- If true, the split point is not relocated to an actual data value. This can yield substantial speed-ups for large datasets with numeric attributes.

debug -- If set to true, classifier may output additional info to the console.

binarySplits -- Whether to use binary splits on nominal attributes when building the partial trees.

doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime).

minNumObj -- The minimum number of instances per rule.

useMDLcorrection -- Whether MDL correction is used when finding splits on numeric attributes.

PART

Home

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation:     relation
Instances:    2201
Attributes:   4
              class
              age
              sex
              survived
Test mode:    10-fold cross-validation
```

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

PART decision list
------------------

sex = male AND
class = 2nd AND
age = adult: no (168.0/14.0)

sex = male AND
class = crew: no (862.0/192.0)

sex = male AND
class = 3rd: no (510.0/88.0)

sex = female AND
class = 3rd: no (196.0/90.0)

sex = female: yes (274.0/20.0)

age = adult: no (175.0/57.0)

: yes (16.0)

Number of Rules  :    7


Time taken to build model: 0.01 seconds

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

```
Correctly Classified Instances        1740               79.055  %
Incorrectly Classified Instances       461               20.945  %
Kappa statistic                          0.4334
Mean absolute error                      0.3106
Root mean squared error                  0.3947
Relative absolute error                 70.9957 %
Root relative squared error             84.3999 %
Total Number of Instances             2201

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.380    0.013    0.931      0.380    0.539      0.506    0.749     0.670     yes
                 0.987    0.620    0.769      0.987    0.864      0.506    0.749     0.826     no
Weighted Avg.    0.791    0.424    0.821      0.791    0.759      0.506    0.749     0.775

=== Confusion Matrix ===

    a    b   <-- classified as
  270  441 |    a = yes
   20 1470 |    b = no
```

PART w/Training Set

Home

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation:     relation
Instances:    2201
Attributes:   4
              class
              age
              sex
              survived
Test mode:    evaluate on training data

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

PART decision list
------------------

sex = male AND
class = 2nd AND
age = adult: no (168.0/14.0)

sex = male AND
class = crew: no (862.0/192.0)

sex = male AND
class = 3rd: no (510.0/88.0)

sex = female AND
class = 3rd: no (196.0/90.0)
```

```
sex = female: yes (274.0/20.0)

age = adult: no (175.0/57.0)

: yes (16.0)

Number of Rules  :    7


Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances        1740               79.055  %
Incorrectly Classified Instances       461               20.945  %
Kappa statistic                          0.4334
Mean absolute error                      0.3094
Root mean squared error                  0.3933
Relative absolute error                 70.7399 %
Root relative squared error             84.1125 %
Total Number of Instances             2201

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.380    0.013    0.931      0.380    0.539      0.506    0.764     0.656     yes
```

```
                 0.987    0.620    0.769      0.987    0.864      0.506    0.764     0.826     no
Weighted Avg.    0.791    0.424    0.821      0.791    0.759      0.506    0.764     0.771

=== Confusion Matrix ===

    a    b   <-- classified as
  270  441 |    a = yes
   20 1470 |    b = no
```

4

[Home](#)

Question #4 Supporting Material

ZOO.ARFF

```
% Changes to WEKA Format: SRG - November 1994
%       1. Boolean attributes changed from 1 and 0 to Enumerated attribute with
%          values {true and false}
%       2. Class Number (Attribute 18) changed to an Enumerated type with
%          values {1,2,3,4,5,6,7}
%
% 1. Title: Zoo database
%
% 2. Source Information
%    -- Creator: Richard Forsyth
%    -- Donor: Richard S. Forsyth
%              8 Grosvenor Avenue
%              Mapperley Park
%              Nottingham NG3 5DX
%              0602-621676
```

```
%    -- Date: 5/15/1990
%
% 3. Past Usage:
%    -- None known other than what is shown in Forsyth's PC/BEAGLE User's Guide.
%
% 4. Relevant Information:
%    -- A simple database containing 17 Boolean-valued attributes.  The "type"
%       attribute appears to be the class attribute.  Here is a breakdown of
%       which animals are in which type: (I find it unusual that there are
%       2 instances of "frog" and one of "girl"!)
%
%       Class# Set of animals:
%       ====== ===========================================================
%            1 (41) aardvark, antelope, bear, boar, buffalo, calf,
%                   cavy, cheetah, deer, dolphin, elephant,
%                   fruitbat, giraffe, girl, goat, gorilla, hamster,
%                   hare, leopard, lion, lynx, mink, mole, mongoose,
%                   opossum, oryx, platypus, polecat, pony,
%                   porpoise, puma, pussycat, raccoon, reindeer,
%                   seal, sealion, squirrel, vampire, vole, wallaby,wolf
%            2 (20) chicken, crow, dove, duck, flamingo, gull, hawk,
%                   kiwi, lark, ostrich, parakeet, penguin, pheasant,
%                   rhea, skimmer, skua, sparrow, swan, vulture, wren
%            3 (5)  pitviper, seasnake, slowworm, tortoise, tuatara
%            4 (13) bass, carp, catfish, chub, dogfish, haddock,
%                   herring, pike, piranha, seahorse, sole, stingray, tuna
%            5 (4)  frog, frog, newt, toad
%            6 (8)  flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp
%            7 (10) clam, crab, crayfish, lobster, octopus,
%                   scorpion, seawasp, slug, starfish, worm
%
% 5. Number of Instances: 101
% 6. Number of Attributes: 18 (animal name, 15 Boolean attributes, 2 numerics)
% 7. Attribute Information: (name of attribute and type of value domain)
%    1. animal name:    Unique for each instance
%    2. hair            Boolean
%    3. feathers Boolean
%    4. eggs            Boolean
%    5. milk            Boolean
%    6. airborne Boolean
%    7. aquatic         Boolean
%    8. predator Boolean
%    9. toothed         Boolean
%   10. backbone Boolean
%   11. breathes Boolean
%   12. venomous Boolean
%   13. fins            Boolean
```

```
%   14. legs            Numeric (set of values: {0,2,4,5,6,8})
%   15. tail            Boolean
%   16. domestic Boolean
%   17. catsize         Boolean
%   18. type            Numeric (integer values in range [1,7])
%
% 8. Missing Attribute Values: None
% 9. Class Distribution: Given above

@RELATION zoo

@ATTRIBUTE animal
{aardvark,antelope,bass,bear,boar,buffalo,calf,carp,catfish,cavy,cheetah,chicken,chub,clam,crab,crayfish,crow,deer,dogfish,dolphin,dove,duck,elephant
,flamingo,flea,frog,fruitbat,giraffe,girl,gnat,goat,gorilla,gull,haddock,hamster,hare,hawk,herring,honeybee,housefly,kiwi,ladybird,lark,leopard,lion,
lobster,lynx,mink,mole,mongoose,moth,newt,octopus,opossum,oryx,ostrich,parakeet,penguin,pheasant,pike,piranha,pitviper,platypus,polecat,pony,porpoise
,puma,pussycat,raccoon,reindeer,rhea,scorpion,seahorse,seal,sealion,seasnake,seawasp,skimmer,skua,slowworm,slug,sole,sparrow,squirrel,starfish,stingr
ay,swan,termite,toad,tortoise,tuatara,tuna,vampire,vole,vulture,wallaby,wasp,wolf,worm,wren}
@ATTRIBUTE hair {false, true}
@ATTRIBUTE feathers {false, true}
@ATTRIBUTE eggs {false, true}
@ATTRIBUTE milk {false, true}
@ATTRIBUTE airborne {false, true}
@ATTRIBUTE aquatic {false, true}
@ATTRIBUTE predator {false, true}
@ATTRIBUTE toothed {false, true}
@ATTRIBUTE backbone {false, true}
@ATTRIBUTE breathes {false, true}
@ATTRIBUTE venomous {false, true}
@ATTRIBUTE fins {false, true}
% hakank: changed this since it's simpler for e.g. Apriori
% @ATTRIBUTE legs INTEGER [0,9]
@ATTRIBUTE legs {0,1,2,3,4,5,6,7,8,9}
@ATTRIBUTE tail {false, true}
@ATTRIBUTE domestic {false, true}
@ATTRIBUTE catsize {false, true}
@ATTRIBUTE type { 1,2,3,4,5,6,7 }

@DATA
%
% Instances (101):
%
aardvark,true,false,false,true,false,false,true,true,true,true,false,false,4,false,false,true,1
antelope,true,false,false,true,false,false,false,true,true,true,false,false,4,true,false,true,1
bass,false,false,true,false,false,true,true,true,true,false,false,true,0,true,false,false,4
.......
..........
..............
```

PART

Home

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation:     zoo
Instances:    101
Attributes:   18
              animal
              hair
              feathers
              eggs
              milk
              airborne
```

```
              aquatic
              predator
              toothed
              backbone
              breathes
              venomous
              fins
              legs
              tail
              domestic
              catsize
              type
Test mode:    10-fold cross-validation

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

PART decision list
------------------

feathers = false AND
milk = true: 1 (41.0)

feathers = true: 2 (20.0)

backbone = false AND
airborne = false AND
predator = true: 7 (8.0)

backbone = false AND
legs = 6: 6 (8.0)

fins = true: 4 (13.0)
```

```
backbone = true AND
tail = true: 3 (6.0/1.0)

aquatic = true: 5 (3.0)

: 7 (2.0)

Number of Rules  :    8


Time taken to build model: 0 seconds

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

Correctly Classified Instances          93               92.0792 %
Incorrectly Classified Instances         8                7.9208 %
Kappa statistic                          0.8955
Mean absolute error                      0.0231
Root mean squared error                  0.1435
Relative absolute error                 10.5346 %
Root relative squared error             43.4854 %
Total Number of Instances              101

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     1
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     2
                 0.600    0.010    0.750      0.600    0.667      0.656    0.793     0.420     3
```

```
                 1.000    0.011    0.929      1.000    0.963      0.958    0.994     0.929     4
                 0.750    0.000    1.000      0.750    0.857      0.862    0.872     0.760     5
                 0.625    0.022    0.714      0.625    0.667      0.642    0.927     0.794     6
                 0.800    0.044    0.667      0.800    0.727      0.698    0.978     0.724     7
Weighted Avg.    0.921    0.008    0.923      0.921    0.920      0.914    0.976     0.909

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = 1
  0 20  0  0  0  0  0 |  b = 2
  0  0  3  1  0  0  1 |  c = 3
  0  0  0 13  0  0  0 |  d = 4
  0  0  1  0  3  0  0 |  e = 5
  0  0  0  0  0  5  3 |  f = 6
  0  0  0  0  0  2  8 |  g = 7
```

ONE-R

Home

```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     zoo
Instances:    101
Attributes:   18
              animal
              hair
```

```
              feathers
              eggs
              milk
              airborne
              aquatic
              predator
              toothed
              backbone
              breathes
              venomous
              fins
              legs
              tail
              domestic
              catsize
              type
Test mode:    10-fold cross-validation

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

animal:
      aardvark   -> 1
      antelope   -> 1
      bass  -> 4
      bear  -> 1
      boar  -> 1
      buffalo    -> 1
      calf  -> 1
      carp  -> 4
      catfish    -> 4
      cavy  -> 1
      cheetah    -> 1
```

```
chicken    -> 2
chub  -> 4
clam  -> 7
crab  -> 7
crayfish   -> 7
crow  -> 2
deer  -> 1
dogfish    -> 4
dolphin    -> 1
dove  -> 2
duck  -> 2
elephant   -> 1
flamingo   -> 2
flea  -> 6
frog  -> 5
fruitbat   -> 1
giraffe    -> 1
girl  -> 1
gnat  -> 6
goat  -> 1
gorilla    -> 1
gull  -> 2
haddock    -> 4
hamster    -> 1
hare  -> 1
hawk  -> 2
herring    -> 4
honeybee   -> 6
housefly   -> 6
kiwi  -> 2
ladybird   -> 6
lark  -> 2
```

```
leopard    -> 1
lion  -> 1
lobster    -> 7
lynx  -> 1
mink  -> 1
mole  -> 1
mongoose   -> 1
moth  -> 6
newt  -> 5
octopus    -> 7
opossum    -> 1
oryx  -> 1
ostrich    -> 2
parakeet   -> 2
penguin    -> 2
pheasant   -> 2
pike  -> 4
piranha    -> 4
pitviper   -> 3
platypus   -> 1
polecat    -> 1
pony  -> 1
porpoise   -> 1
puma  -> 1
pussycat   -> 1
raccoon    -> 1
reindeer   -> 1
rhea  -> 2
scorpion   -> 7
seahorse   -> 4
seal  -> 1
sealion    -> 1
```

```
      seasnake   -> 3
      seawasp    -> 7
      skimmer    -> 2
      skua  -> 2
      slowworm   -> 3
      slug  -> 7
      sole  -> 4
      sparrow    -> 2
      squirrel   -> 1
      starfish   -> 7
      stingray   -> 4
      swan  -> 2
      termite    -> 6
      toad  -> 5
      tortoise   -> 3
      tuatara    -> 3
      tuna  -> 4
      vampire    -> 1
      vole  -> 1
      vulture    -> 2
      wallaby    -> 1
      wasp  -> 6
      wolf  -> 1
      worm  -> 7
      wren  -> 2
(101/101 instances correct)


Time taken to build model: 0 seconds

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

```
Correctly Classified Instances          43               42.5743 %
Incorrectly Classified Instances        58               57.4257 %
Kappa statistic                          0.045
Mean absolute error                      0.1641
Root mean squared error                  0.4051
Relative absolute error                 74.8424 %
Root relative squared error            122.7774 %
Total Number of Instances              101

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 1.000    0.967    0.414      1.000    0.586      0.117    0.517     0.414     1
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.198     2
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.050     3
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.129     4
                 0.500    0.000    1.000      0.500    0.667      0.700    0.750     0.520     5
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.079     6
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.099     7
Weighted Avg.    0.426    0.392    0.208      0.426    0.264      0.075    0.517     0.263

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = 1
 20  0  0  0  0  0  0 |  b = 2
  5  0  0  0  0  0  0 |  c = 3
 13  0  0  0  0  0  0 |  d = 4
  2  0  0  0  2  0  0 |  e = 5
  8  0  0  0  0  0  0 |  f = 6
```

```
10  0  0  0  0  0  0 |  g = 7
```