WEKA ASSIGNMENT

BY: AARON PULIKKOTTIL

1. DATASET DESCRIPTION:

Dataset Name: Labor

Source of the dataset:
https://github.com/tertiarycourses/Weka/blob/master/Weka%20datasets/labor.arff

Dataset description: The data set contains 1 classes of 57 instances each where each class refers to a type of iris plant.

No of attributes: 17

No of instances: 57

2. INPUT & OUTPUT:

LABOUR	1.Duration 2.Wage increase 1st year 3.Working hours 4.Pension 5.Education allowance 6.Statutory Holidays 7.Vacation 8.Standby Pay 9.Bereavement Assistance 10.Class	Duration: class Wage increase 1st year: class Working hours: class Pension: class Education allowance: class Statutory Holidays: class Vacation: class Standby Pay: class Bereavement Assistance: class

3. DATA PREPARATION:

Removal of null data types, empty instances and false attribute.

4. SELECTING AN ALGORITHM:

Algorithm	Result from Weka	
	Correctly Classified Instances	Mean Absolute Error
Weka Classifiers		
NaiveBayes	51	0.1042
NaiveBayesMultinominalText	37	0.4574
NaiveBayesUpdateable	51	0.1042
Weka Classifiers Functions		
Logistic	53	0.0641
MultilayerPerceptron	49	0.1524
Weka Classifiers Meta		
AdaBoostM1	50	0.1519
Bagging	48	0.2979
Weka Classifier Trees		
J48	42	0.3192
RandomTree	45	0.2438
Weka Classifier Rules		
DecisionTable	43	0.2996
JRip	44	0.2256

5.DATA PREPROCESSING:

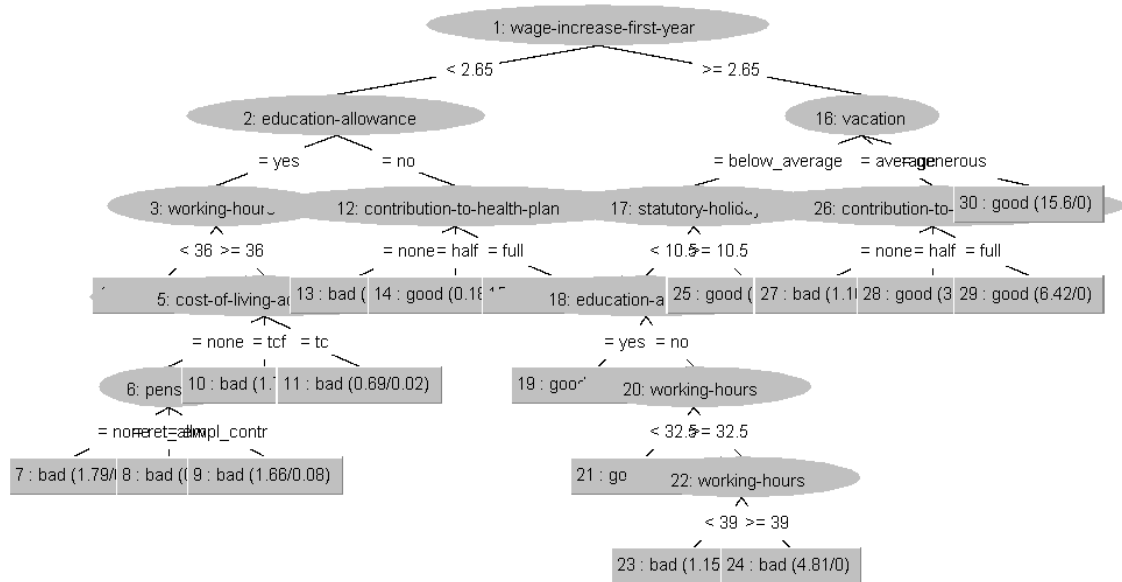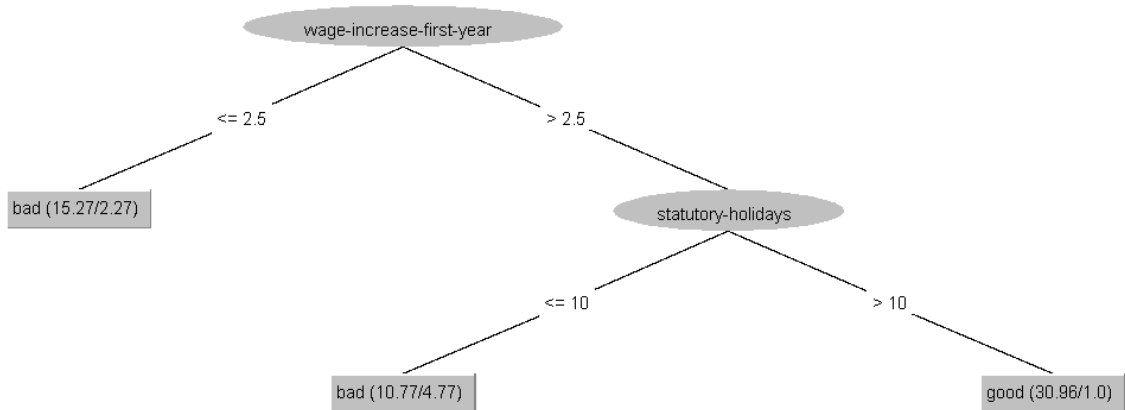|	
Attribute Evaluator	Classifier Subset Eval (Classifier: meta.Bagging with training parameter: classifiers = J48 and the others in their default)
Search Method	BestFirst (Direction: Forward)
Evaluation Mode	10-fold cross-validation (Folds = 10, Seed =1)

6.TREE PRODUCED:

Tree View

1: wage-increase-first-year

< 2.65   >= 2.65

2: education-allowance

16: vacation

= yes   = no

= below_average   = average = generous

3: working-hour   12: contribution-to-health-plan

17: statutory-holiday

26: contribution-to-   30 : good (15.6/0)

< 36 >= 36

= none = half = full

< 10.5 = 10.5

= none = half = full

5: cost-of-living-a   13 : bad (14 : good (0.18   18: education-a   25 : good (27 : bad (1.1   28 : good (3   29 : good (6.42/0)

= none = tcf = tc

= yes = no

6: pens   10 : bad (1.   11 : bad (0.69/0.02)

19 : goo   20: working-hours

= none = ret=allmpl_contr

< 32.5 = 32.5

7 : bad (1.79/   8 : bad (9 : bad (1.66/0.08)

21 : go   22: working-hours

< 39 >= 39

23 : bad (1.15   24 : bad (4.81/0)

Tree View

wage-increase-first-year

<= 2.5   > 2.5

bad (15.27/2.27)

statutory-holidays

<= 10   > 10

bad (10.77/4.77)

good (30.96/1.0)

7. SCREENSHOTS:

Screenshot 1

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose | JRip -F 3 -N 2.0 -O 2 -S 1

Test options

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

More options...

(Nom) class

Start | Stop

Result list (right-click for options)

- 13:51:49 - bayes.NaiveBayes
- 13:52:20 - bayes.NaiveBayesMultinomialText
- 13:54:25 - bayes.NaiveBayesUpdateable
- 16:26:23 - functions.Logistic
- 16:26:45 - functions.MultilayerPerceptron
- 16:27:46 - meta.AdaBoostM1
- 16:28:08 - meta.Bagging
- 16:30:08 - trees.RandomTree
- 16:30:44 - trees.J48
- 16:32:28 - rules.DecisionTable
- 16:32:49 - rules.JRip

Classifier output

```
[total]                     14.0    20.0

contribution-to-health-plan
  none                       9.0     1.0
  half                       3.0     8.0
  full                       7.0    15.0
  [total]                   19.0    24.0


Time taken to build model: 0 seconds

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

Correctly Classified Instances        51            89.4737 %
Incorrectly Classified Instances       6            10.5263 %
Kappa statistic                        0.7741
Mean absolute error                    0.1042
Root mean squared error                0.2637
Relative absolute error               22.7763 %
Root relative squared error           55.2266 %
Total Number of Instances             57

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
               0.900    0.108    0.818      0.900   0.857      0.776  0.965     0.926     bad
               0.892    0.100    0.943      0.892   0.917      0.776  0.965     0.983     good
Weighted Avg.  0.895    0.103    0.899      0.895   0.896      0.776  0.965     0.963

=== Confusion Matrix ===

  a   b   <-- classified as
 18   2 |  a = bad
  4  33 |  b = good
```

Status: OK

Screenshot 2

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose | JRip -F 3 -N 2.0 -O 2 -S 1

Test options

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

More options...

(Nom) class

Start | Stop

Result list (right-click for options)

- 13:51:49 - bayes.NaiveBayes
- 13:52:20 - bayes.NaiveBayesMultinomialText
- 13:54:25 - bayes.NaiveBayesUpdateable
- 16:26:23 - functions.Logistic
- 16:26:45 - functions.MultilayerPerceptron
- 16:27:46 - meta.AdaBoostM1
- 16:28:08 - meta.Bagging
- 16:30:08 - trees.RandomTree
- 16:30:44 - trees.J48
- 16:32:28 - rules.DecisionTable
- 16:32:49 - rules.JRip

Classifier output

```
longterm-disability-assistance=no        2.1493222090504170216
contribution-to-dental-plan=none         6.56351273045078E14
contribution-to-dental-plan=half                         1.2349
contribution-to-dental-plan=full                         0
bereavement-assistance=no                4.3065760813857376E16
contribution-to-health-plan=none         2.14532228874995942E18
contribution-to-health-plan=half                         0
contribution-to-health-plan=full                         0

Time taken to build model: 0.03 seconds

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

Correctly Classified Instances        53            92.9825 %
Incorrectly Classified Instances       4             7.0175 %
Kappa statistic                        0.8494
Mean absolute error                    0.0641
Root mean squared error                0.2438
Relative absolute error               14.0069 %
Root relative squared error           51.0652 %
Total Number of Instances             57

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
               0.950    0.081    0.864      0.950   0.905      0.852  0.970     0.927     bad
               0.919    0.050    0.971      0.919   0.944      0.852  0.981     0.989     good
Weighted Avg.  0.930    0.061    0.934      0.930   0.931      0.852  0.977     0.967

=== Confusion Matrix ===

  a   b   <-- classified as
 19   1 |  a = bad
  3  34 |  b = good
```

Status: OK

Screenshot 1 — Weka Explorer (rules.JRip / MultilayerPerceptron result)

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose | JRip -F 3 -N 2.0 -O 2 -S 1

Test options
- ○ Use training set
- ○ Supplied test set — Set...
- ● Cross-validation   Folds | 10
- ○ Percentage split   % | 66
- More options...

(Nom) class

Start | Stop

Result list (right-click for options)
- 13:51:49 - bayes.NaiveBayes
- 13:52:20 - bayes.NaiveBayesMultinomialText
- 13:54:25 - bayes.NaiveBayesUpdateable
- 16:26:23 - functions.Logistic
- 16:26:45 - functions.MultilayerPerceptron
- 16:27:46 - meta.AdaBoostM1
- 16:28:08 - meta.Bagging
- 16:30:08 - trees.RandomTree
- 16:30:44 - trees.J48
- 16:32:28 - rules.DecisionTable
- 16:32:49 - rules.JRip

Classifier output

```
    Attrib contribution-to-health-plan=half      0.292040505191255
    Attrib contribution-to-health-plan=full     -0.1921735010512306
Class bad
    Input
    Node 0
Class good
    Input
    Node 1

Time taken to build model: 0.23 seconds

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

Correctly Classified Instances          49               85.9649 %
Incorrectly Classified Instances         8               14.0351 %
Kappa statistic                          0.6915
Mean absolute error                      0.1524
Root mean squared error                  0.337
Relative absolute error                 33.3246 %
Root relative squared error             70.5761 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 0.800    0.108    0.800      0.800   0.800      0.692  0.923     0.917     bad
                 0.892    0.200    0.892      0.892   0.892      0.692  0.923     0.951     good
Weighted Avg.    0.860    0.168    0.860      0.860   0.860      0.692  0.923     0.939

=== Confusion Matrix ===

  a  b   <-- classified as
 16  4 |  a = bad
  4 33 |  b = good
```

Status — OK

Screenshot 2 — Weka Explorer (meta.AdaBoostM1 result)

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose | JRip -F 3 -N 2.0 -O 2 -S 1

Test options
- ○ Use training set
- ○ Supplied test set — Set...
- ● Cross-validation   Folds | 10
- ○ Percentage split   % | 66
- More options...

(Nom) class

Start | Stop

Result list (right-click for options)
- 13:51:49 - bayes.NaiveBayes
- 13:52:20 - bayes.NaiveBayesMultinomialText
- 13:54:25 - bayes.NaiveBayesUpdateable
- 16:26:23 - functions.Logistic
- 16:26:45 - functions.MultilayerPerceptron
- 16:27:46 - meta.AdaBoostM1
- 16:28:08 - meta.Bagging
- 16:30:08 - trees.RandomTree
- 16:30:44 - trees.J48
- 16:32:28 - rules.DecisionTable
- 16:32:49 - rules.JRip

Classifier output

```
wage-increase-first-year is missing
bad      good
0.0      1.0

Weight: 1.12

Number of performed Iterations: 10

Time taken to build model: 0.01 seconds

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

Correctly Classified Instances          50               87.7193 %
Incorrectly Classified Instances         7               12.2807 %
Kappa statistic                          0.7273
Mean absolute error                      0.1519
Root mean squared error                  0.3414
Relative absolute error                 33.2028 %
Root relative squared error             71.5013 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 0.800    0.081    0.842      0.800   0.821      0.728  0.870     0.863     bad
                 0.919    0.200    0.895      0.919   0.907      0.728  0.870     0.876     good
Weighted Avg.    0.877    0.158    0.876      0.877   0.876      0.728  0.870     0.871

=== Confusion Matrix ===

  a  b   <-- classified as
 16  4 |  a = bad
  3 34 |  b = good
```

Status — OK

Screenshot 1 (16:49)

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose `JRip -F 3 -N 2.0 -O 2 -S 1`

Test options
- Use training set
- Supplied test set | Set...
- ● Cross-validation | Folds | 10
- Percentage split | % | 66
- More options...

(Nom) class

Start | Stop

Result list (right-click for options)
- 13:51:49 - bayes.NaiveBayes
- 13:52:20 - bayes.NaiveBayesMultinomialText
- 13:54:25 - bayes.NaiveBayesUpdateable
- 16:26:23 - functions.Logistic
- 16:26:45 - functions.MultilayerPerceptron
- 16:27:46 - meta.AdaBoostM1
- 16:28:08 - meta.Bagging
- 16:30:08 - trees.RandomTree
- 16:30:44 - trees.J48
- 16:32:28 - rules.DecisionTable
- 16:32:49 - rules.JRip

Classifier output

```
                contribution-to-health-plan
             class
Test mode:   10-fold cross-validation

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

Bagging with 10 iterations and base learner

weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0

Time taken to build model: 0.03 seconds

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

Correctly Classified Instances          48               84.2105 %
Incorrectly Classified Instances         9               15.7895 %
Kappa statistic                          0.641
Mean absolute error                      0.2979
Root mean squared error                  0.3757
Relative absolute error                 65.1224 %
Root relative squared error             78.6822 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 0.700    0.081    0.824      0.700   0.757      0.646  0.856     0.774     bad
                 0.919    0.300    0.850      0.919   0.883      0.646  0.856     0.886     good
Weighted Avg.    0.842    0.223    0.841      0.842   0.839      0.646  0.856     0.847

=== Confusion Matrix ===

  a  b   <-- classified as
 14  6 |  a = bad
  3 34 |  b = good
```

Status: OK

Screenshot 2 (16:50)

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose `JRip -F 3 -N 2.0 -O 2 -S 1`

Test options
- Use training set
- Supplied test set | Set...
- ● Cross-validation | Folds | 10
- Percentage split | % | 66
- More options...

(Nom) class

Start | Stop

Result list (right-click for options)
- 13:51:49 - bayes.NaiveBayes
- 13:52:20 - bayes.NaiveBayesMultinomialText
- 13:54:25 - bayes.NaiveBayesUpdateable
- 16:26:23 - functions.Logistic
- 16:26:45 - functions.MultilayerPerceptron
- 16:27:46 - meta.AdaBoostM1
- 16:28:08 - meta.Bagging
- 16:30:08 - trees.RandomTree
- 16:30:44 - trees.J48
- 16:32:28 - rules.DecisionTable
- 16:32:49 - rules.JRip

Classifier output

```
|  |  |  |    working hours >= 35 : bad (4.01/0)
|  |  statutory-holidays >= 10.5 : good (7.23/0)
|  vacation = average
|  |  contribution-to-dental-plan = none : bad (1.16/0.16)
|  |  contribution-to-dental-plan = half : good (3.79/0)
|  |  contribution-to-dental-plan = full : good (6.42/0)
|  vacation = generous : good (15.6/0)

Size of the tree : 30

Time taken to build model: 0.01 seconds

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

Correctly Classified Instances          45               78.9474 %
Incorrectly Classified Instances        12               21.0526 %
Kappa statistic                          0.5156
Mean absolute error                      0.2438
Root mean squared error                  0.4353
Relative absolute error                 53.2875 %
Root relative squared error             91.1643 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 0.600    0.108    0.750      0.600   0.667      0.522  0.754     0.654     bad
                 0.892    0.400    0.805      0.892   0.846      0.522  0.754     0.795     good
Weighted Avg.    0.789    0.298    0.786      0.789   0.783      0.522  0.754     0.746

=== Confusion Matrix ===

  a  b   <-- classified as
 12  8 |  a = bad
  4 33 |  b = good
```

Status: OK

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose | JRip -F 3 -N 2.0 -O 2 -S 1

Test options
- Use training set
- Supplied test set   Set...
- Cross-validation Folds 10
- Percentage split % 66

More options...

(Nom) class

Start   Stop

Result list (right-click for options)
13:51:49 - bayes.NaiveBayes
13:52:20 - bayes.NaiveBayesMultinomialText
13:54:25 - bayes.NaiveBayesUpdateable
16:26:23 - functions.Logistic
16:26:45 - functions.MultilayerPerceptron
16:27:46 - meta.AdaBoostM1
16:28:08 - meta.Bagging
16:30:08 - trees.RandomTree
16:30:44 - trees.J48
16:32:28 - rules.DecisionTable
16:32:49 - rules.JRip

Classifier output

```
wage-increase-first-year <= 2.5: bad (19.27/2.27)
wage-increase-first-year > 2.5
|   statutory-holidays <= 10: bad (10.77/4.77)
|   statutory-holidays > 10: good (30.96/1.0)

Number of Leaves  :    3

Size of the tree :    5

Time taken to build model: 0.02 seconds

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

Correctly Classified Instances          42               73.6842 %
Incorrectly Classified Instances        15               26.3158 %
Kappa statistic                          0.4415
Mean absolute error                      0.3192
Root mean squared error                  0.4669
Relative absolute error                 69.7715 %
Root relative squared error             97.7888 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 0.700    0.243    0.609      0.700   0.651      0.444  0.695     0.559     bad
                 0.757    0.300    0.824      0.757   0.789      0.444  0.695     0.738     good
Weighted Avg.    0.737    0.280    0.748      0.737   0.740      0.444  0.695     0.675

=== Confusion Matrix ===

  a  b   <-- classified as
 14  6 |   a = bad
  9 28 |   b = good
```

Status
OK

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose | JRip -F 3 -N 2.0 -O 2 -S 1

Test options
- Use training set
- Supplied test set   Set...
- Cross-validation Folds 10
- Percentage split % 66

More options...

(Nom) class

Start   Stop

Result list (right-click for options)
13:51:49 - bayes.NaiveBayes
13:52:20 - bayes.NaiveBayesMultinomialText
13:54:25 - bayes.NaiveBayesUpdateable
16:26:23 - functions.Logistic
16:26:45 - functions.MultilayerPerceptron
16:27:46 - meta.AdaBoostM1
16:28:08 - meta.Bagging
16:30:08 - trees.RandomTree
16:30:44 - trees.J48
16:32:28 - rules.DecisionTable
16:32:49 - rules.JRip

Classifier output

```
non matches covered by majority class.
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 84
        Merit of best subset found:   92.982
Evaluation (for feature selection): CV (leave one out)
Feature set: 2,13,17

Time taken to build model: 0.05 seconds

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

Correctly Classified Instances          43               75.4386 %
Incorrectly Classified Instances        14               24.5614 %
Kappa statistic                          0.4729
Mean absolute error                      0.2996
Root mean squared error                  0.4014
Relative absolute error                 65.5028 %
Root relative squared error             84.0671 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 0.700    0.216    0.636      0.700   0.667      0.474  0.818     0.722     bad
                 0.784    0.300    0.829      0.784   0.806      0.474  0.818     0.884     good
Weighted Avg.    0.754    0.271    0.761      0.754   0.757      0.474  0.818     0.827

=== Confusion Matrix ===

  a  b   <-- classified as
 14  6 |   a = bad
  8 29 |   b = good
```

Status
OK

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose | JRip -F 3 -N 2.0 -O 2 -S 1

Test options

- Use training set
- Supplied test set | Set...
- Cross-validation | Folds | 10
- Percentage split | % | 66

More options...

(Nom) class

Start | Stop

Result list (right-click for options)

13:51:49 - bayes.NaiveBayes
13:52:20 - bayes.NaiveBayesMultinomialText
13:54:25 - bayes.NaiveBayesUpdateable
16:26:23 - functions.Logistic
16:26:45 - functions.MultilayerPerceptron
16:27:46 - meta.AdaBoostM1
16:28:08 - meta.Bagging
16:30:08 - trees.RandomTree
16:30:44 - trees.J48
16:32:28 - rules.DecisionTable
16:32:49 - rules.JRip

Classifier output

```
(wage-increase-first-year <= 2.5) => class=bad (15.0/2.0)
(statutory-holidays <= 10) and (wage-increase-first-year <= 4) => class=bad (5.0/0.0)
(longterm-disability-assistance = no) => class=bad (2.0/0.0)
 => class=good (35.0/0.0)

Number of Rules : 4

Time taken to build model: 0.03 seconds

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

Correctly Classified Instances          44              77.193 %
Incorrectly Classified Instances        13              22.807 %
Kappa statistic                          0.4535
Mean absolute error                      0.2256
Root mean squared error                  0.456
Relative absolute error                 49.3094 %
Root relative squared error             95.4949 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 0.650    0.162    0.684      0.650   0.667      0.494  0.779     0.718     bad
                 0.838    0.350    0.816      0.838   0.827      0.494  0.779     0.810     good
Weighted Avg.    0.772    0.284    0.770      0.772   0.771      0.494  0.779     0.778

=== Confusion Matrix ===

  a   b   <-- classified as
 13   7 |  a = bad
  6  31 |  b = good
```

Status

OK

Log