CO544 – Machine Learning and Data Mining

Lab 06 – Part 01

E/17/407

WIJESOORIYA H.D

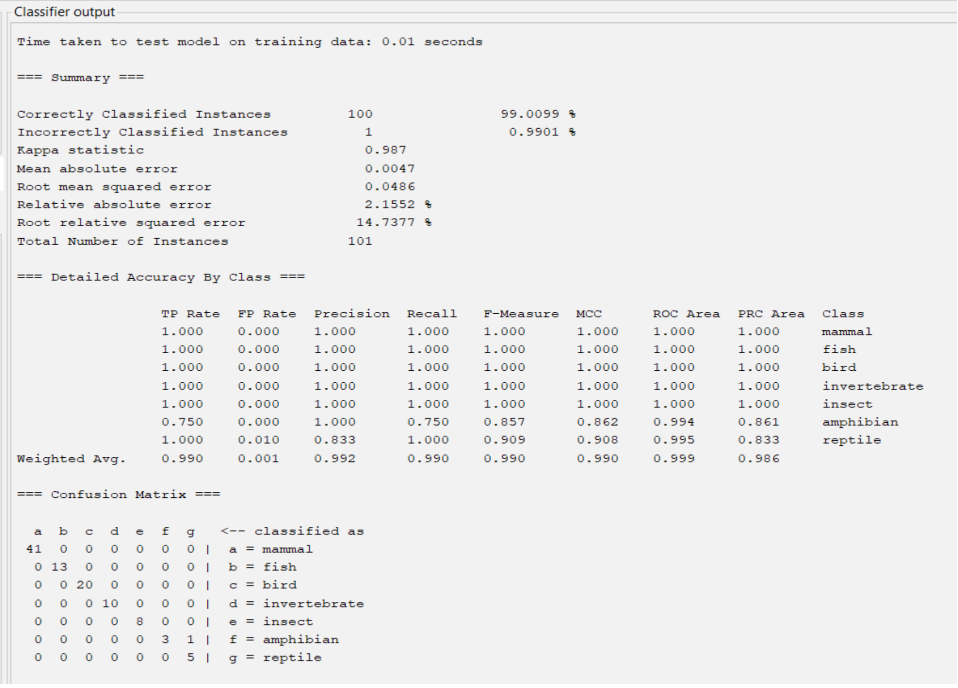
(01) Attributes and their values

Attribute	No of distinct records
animalName	100
hair	2
feathers	2
eggs	2
milk	2
airborne	2
aquatic	2
predator	2
toothed	2

Attribute	No of distinct records
backbone	2
breathes	2
venomous	2
fins	2
legs	6
tail	2
domestic	2
catsize	2
type	7

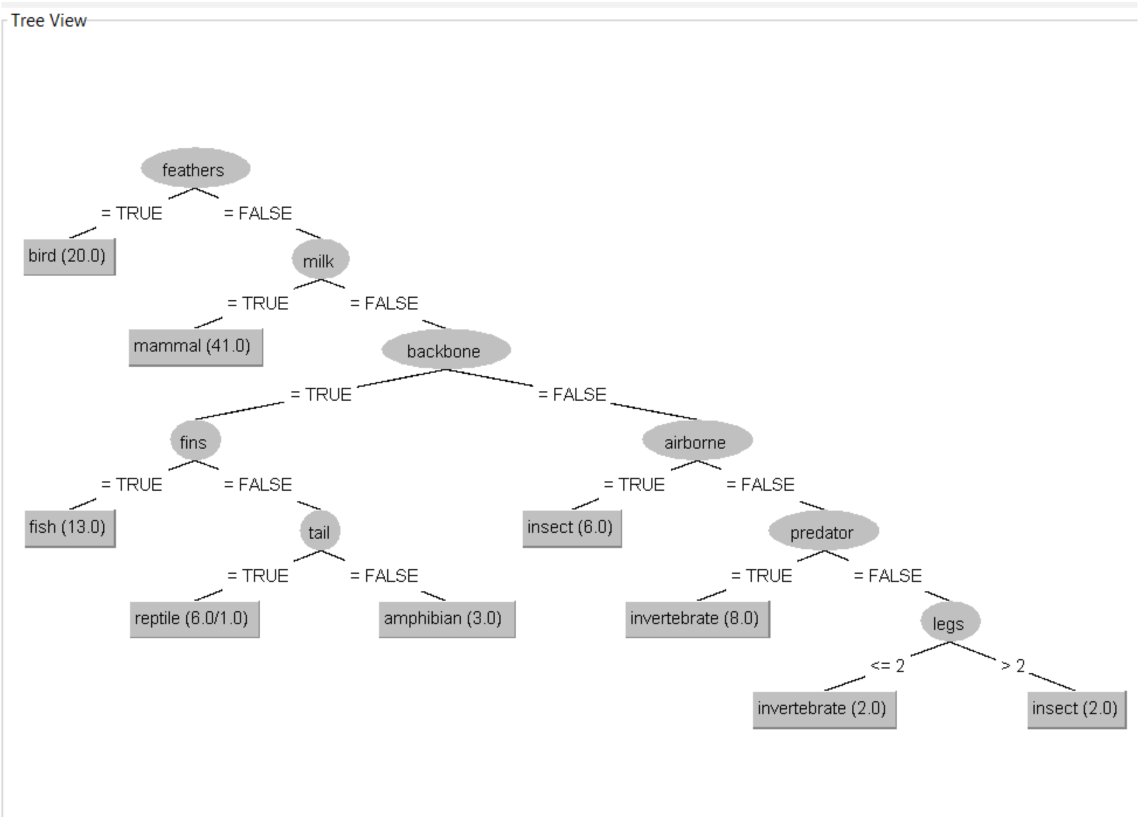
Here the attribute 'type' is the target and the other attributes are the features.

(02) Output of the C4.5 algorithm



```
Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         100               99.0099 %
Incorrectly Classified Instances         1                0.9901 %
Kappa statistic                          0.987
Mean absolute error                      0.0047
Root mean squared error                  0.0486
Relative absolute error                  2.1552 %
Root relative squared error             14.7377 %
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     mammal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     fish
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     bird
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     invertebrate
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     insect
                 0.750    0.000    1.000      0.750    0.857      0.862    0.994     0.861     amphibian
                 1.000    0.010    0.833      1.000    0.909      0.908    0.995     0.833     reptile
Weighted Avg.    0.990    0.001    0.992      0.990    0.990      0.990    0.999     0.986

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 13  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
  0  0  0 10  0  0  0 |  d = invertebrate
  0  0  0  0  8  0  0 |  e = insect
  0  0  0  0  0  3  1 |  f = amphibian
  0  0  0  0  0  0  5 |  g = reptile
```

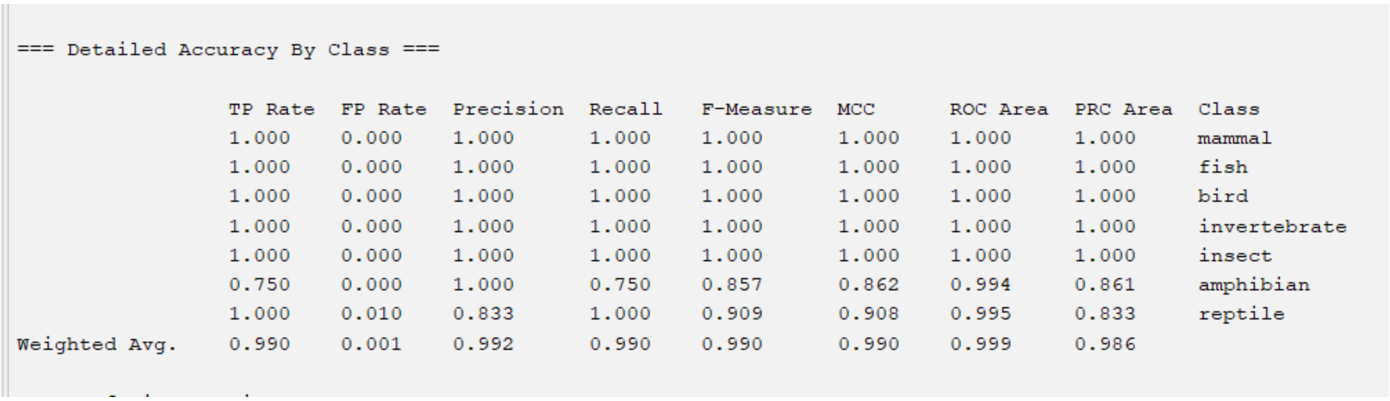
(03) visualize tree



Classification accuracy : 99.0099 %

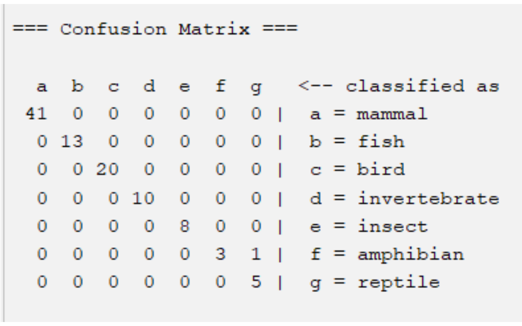
There were 101 instances in the data set, among them 100 instances are correctly classified.

- TP and FP rates



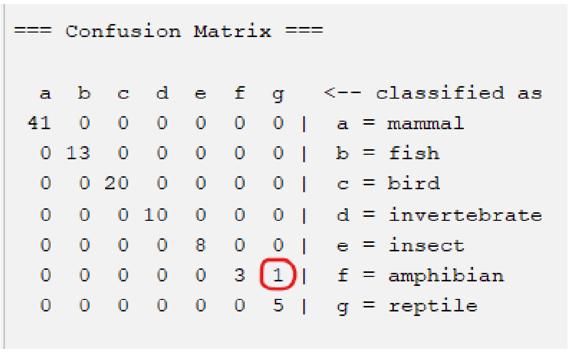
```
=== Detailed Accuracy By Class ===

                 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     mammal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     fish
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     bird
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     invertebrate
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     insect
                 0.750    0.000    1.000      0.750    0.857      0.862    0.994     0.861     amphibian
                 1.000    0.010    0.833      1.000    0.909      0.908    0.995     0.833     reptile
Weighted Avg.    0.990    0.001    0.992      0.990    0.990      0.990    0.999     0.986
```

- Confusion matrix



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 13  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
  0  0  0 10  0  0  0 |  d = invertebrate
  0  0  0  0  8  0  0 |  e = insect
  0  0  0  0  0  3  1 |  f = amphibian
  0  0  0  0  0  0  5 |  g = reptile
```

- Misclassification observed in the confusion matrix

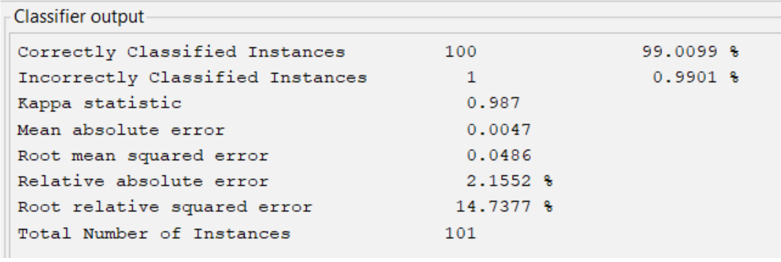


```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 13  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
  0  0  0 10  0  0  0 |  d = invertebrate
  0  0  0  0  8  0  0 |  e = insect
  0  0  0  0  0  3  1 |  f = amphibian
  0  0  0  0  0  0  5 |  g = reptile
```

According to the confusion matrix we can see that an amphibian has been misclassified as a reptile. (the misclassified element is circled in red colour)

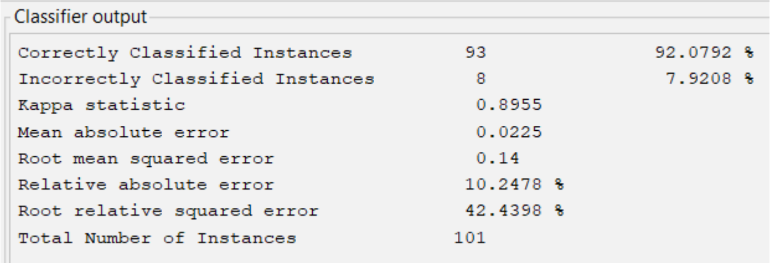
(04) training set option vs 10-fold cross validation option

- training set option – accuracy



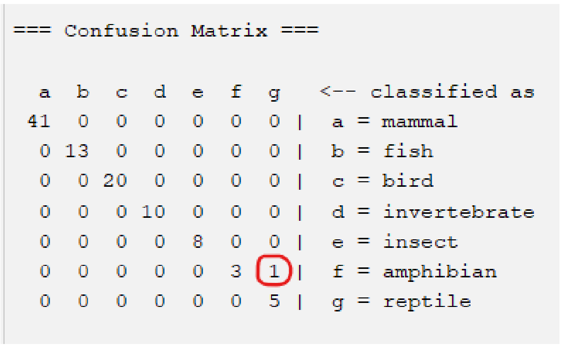
```
Correctly Classified Instances         100               99.0099 %
Incorrectly Classified Instances         1                0.9901 %
Kappa statistic                          0.987
Mean absolute error                      0.0047
Root mean squared error                  0.0486
Relative absolute error                  2.1552 %
Root relative squared error             14.7377 %
Total Number of Instances              101
```

- 10-fold cross validation option

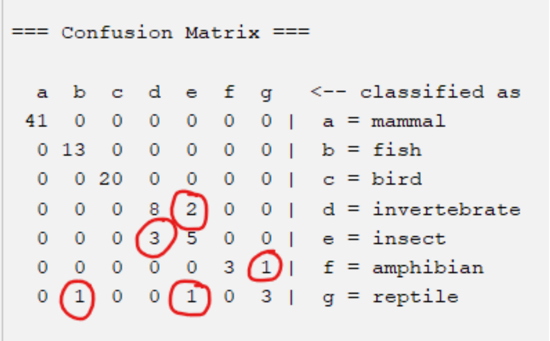


```
Correctly Classified Instances          93               92.0792 %
Incorrectly Classified Instances         8                7.9208 %
Kappa statistic                          0.8955
Mean absolute error                      0.0225
Root mean squared error                  0.14
Relative absolute error                 10.2478 %
Root relative squared error             42.4398 %
Total Number of Instances              101
```

- Misclassification observed in confusion matrices (circled in red colour)



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 13  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
  0  0  0 10  0  0  0 |  d = invertebrate
  0  0  0  0  8  0  0 |  e = insect
  0  0  0  0  0  3  1 |  f = amphibian
  0  0  0  0  0  0  5 |  g = reptile
```



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 13  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
  0  0  0  8  2  0  0 |  d = invertebrate
  0  0  0  3  5  0  0 |  e = insect
  0  0  0  0  0  3  1 |  f = amphibian
  0  1  0  0  1  0  3 |  g = reptile
```

Training set option

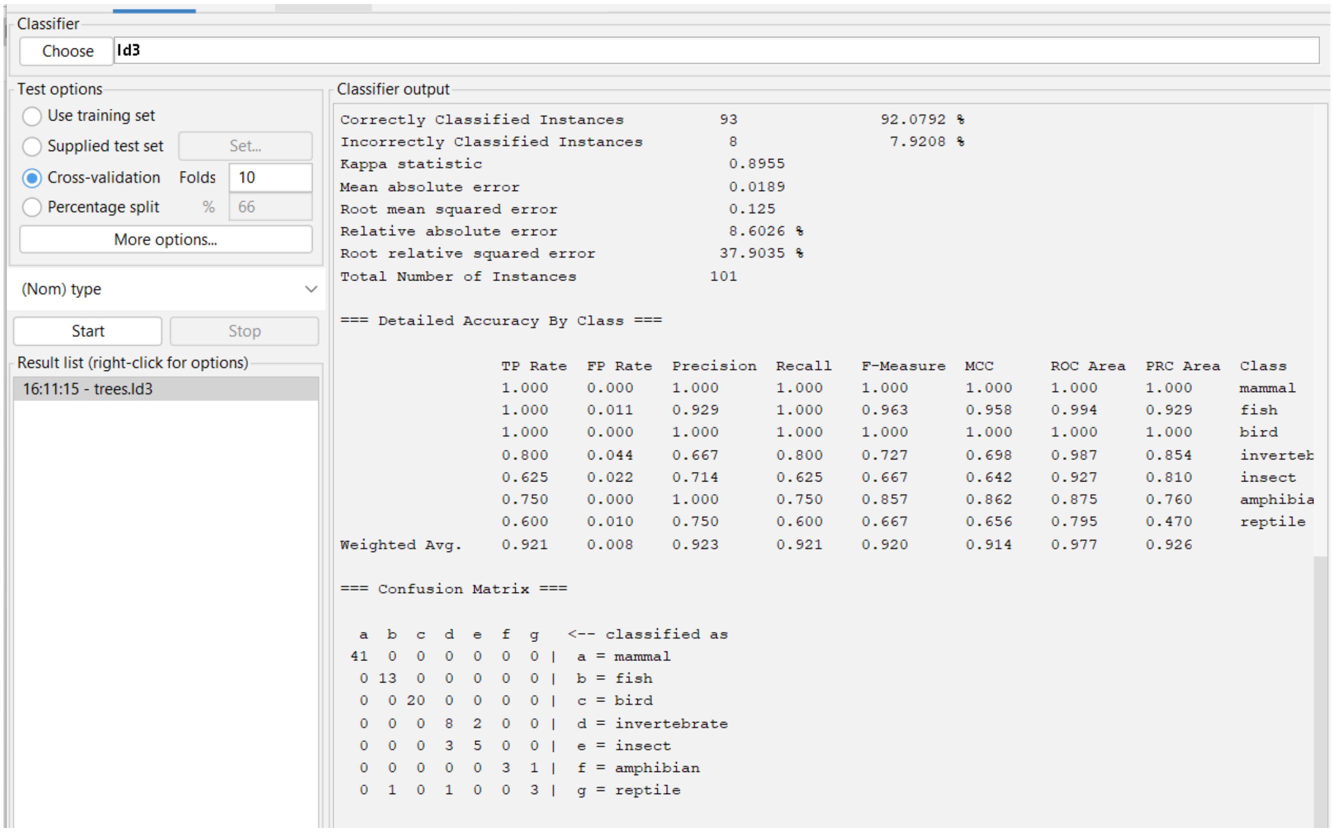
10-fold cross validation option

- According to the above figures we can say that the test option ‘training test’ provides more realistic future performance. Because it has a higher accuracy (99.0099%) compared to the 10-fold cross validation option (92.0729%).

(05) ID3 learning algorithm

We can’t apply ID3 learning algorithm on this data set, because ID3 only works with Discrete or nominal data, and it does not work with continuous data.

(07) ID3 decision tree (10 fold cross validation accuracy)



Classifier

Choose Id3

Test options

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

More options...

(Nom) type

Start Stop

Result list (right-click for options)

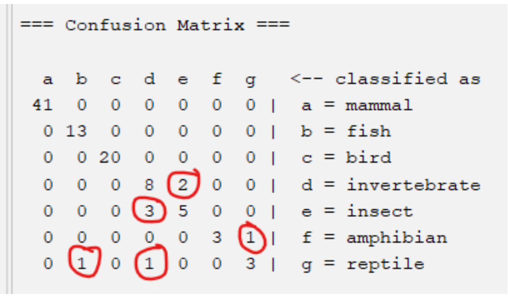
16:11:15 - trees.Id3

Classifier output

```
Correctly Classified Instances          93               92.0792 %
Incorrectly Classified Instances         8                7.9208 %
Kappa statistic                          0.8955
Mean absolute error                      0.0189
Root mean squared error                  0.125
Relative absolute error                  8.6026 %
Root relative squared error             37.9035 %
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     mammal
                 1.000    0.011    0.929      1.000    0.963      0.958    0.994     0.929     fish
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     bird
                 0.800    0.044    0.667      0.800    0.727      0.698    0.987     0.854     invertek
                 0.625    0.022    0.714      0.625    0.667      0.642    0.927     0.810     insect
                 0.750    0.000    1.000      0.750    0.857      0.862    0.875     0.760     amphibia
                 0.600    0.010    0.750      0.600    0.667      0.656    0.795     0.470     reptile
Weighted Avg.    0.921    0.008    0.923      0.921    0.920      0.914    0.977     0.926

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 13  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
  0  0  0  8  2  0  0 |  d = invertebrate
  0  0  0  3  5  0  0 |  e = insect
  0  0  0  0  0  3  1 |  f = amphibian
  0  1  0  1  0  0  3 |  g = reptile
```

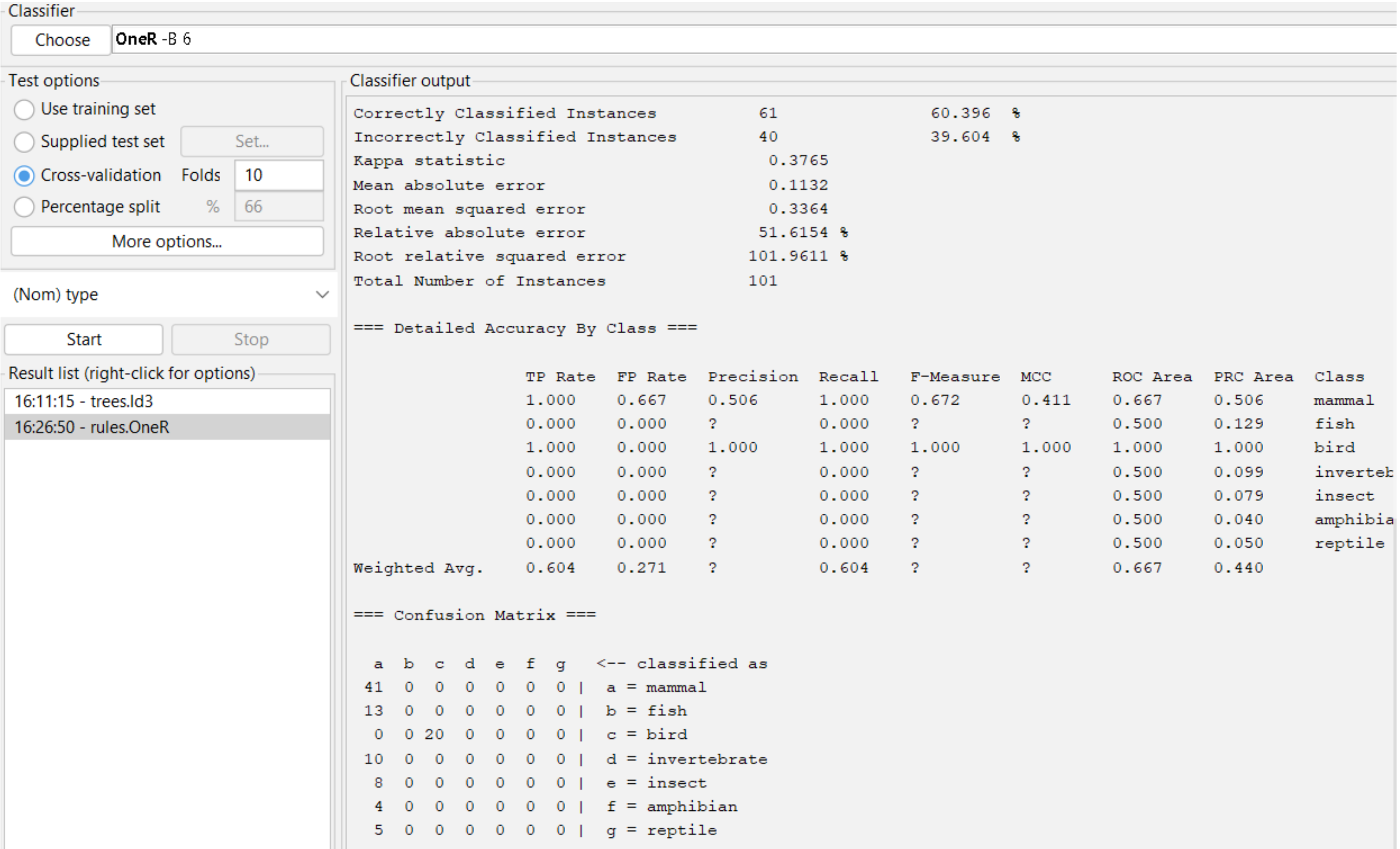
10-fold cross validation accuracy for ID3 decision tree algorithm = 92.0792 %

93 instances are correctly classified out of 101 instances. Misclassifications that can be observed in the confusion matrix are shown below (circled in red colour).



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 13  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
  0  0  0  8  2  0  0 |  d = invertebrate
  0  0  0  3  5  0  0 |  e = insect
  0  0  0  0  0  3  1 |  f = amphibian
  0  1  0  1  0  0  3 |  g = reptile
```

(08) OneR algorithm



OneR -B 6

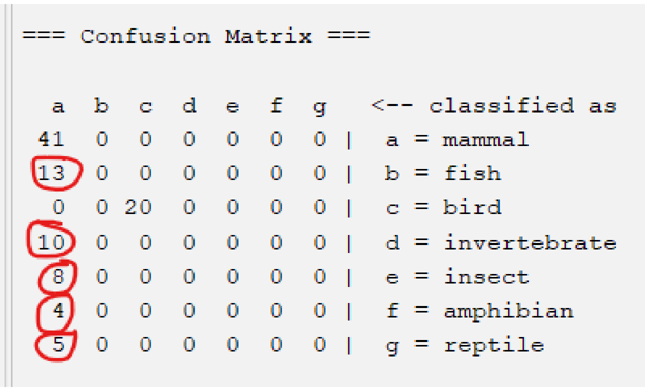
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) type

16:11:15 - trees.Id3
16:26:50 - rules.OneR

```
Correctly Classified Instances          61               60.396  %
Incorrectly Classified Instances        40               39.604  %
Kappa statistic                          0.3765
Mean absolute error                      0.1132
Root mean squared error                  0.3364
Relative absolute error                 51.6154 %
Root relative squared error            101.9611 %
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    0.667    0.506      1.000    0.672      0.411    0.667     0.506     mammal
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.129     fish
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     bird
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.099     inverteb
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.079     insect
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.040     amphibia
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.050     reptile
Weighted Avg.    0.604    0.271    ?          0.604    ?          ?        0.667     0.440

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
 13  0  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
 10  0  0  0  0  0  0 |  d = invertebrate
  8  0  0  0  0  0  0 |  e = insect
  4  0  0  0  0  0  0 |  f = amphibian
  5  0  0  0  0  0  0 |  g = reptile
```

According to the above figure we can say that the 10-fold cross validation accuracy of the OneR algorithm is very low compared to the other algorithms and the accuracy is about 60.396%. 40 instances are incorrectly classified and those instances are show in the following confusion matrix. (misclassified elements are circled in red colour)

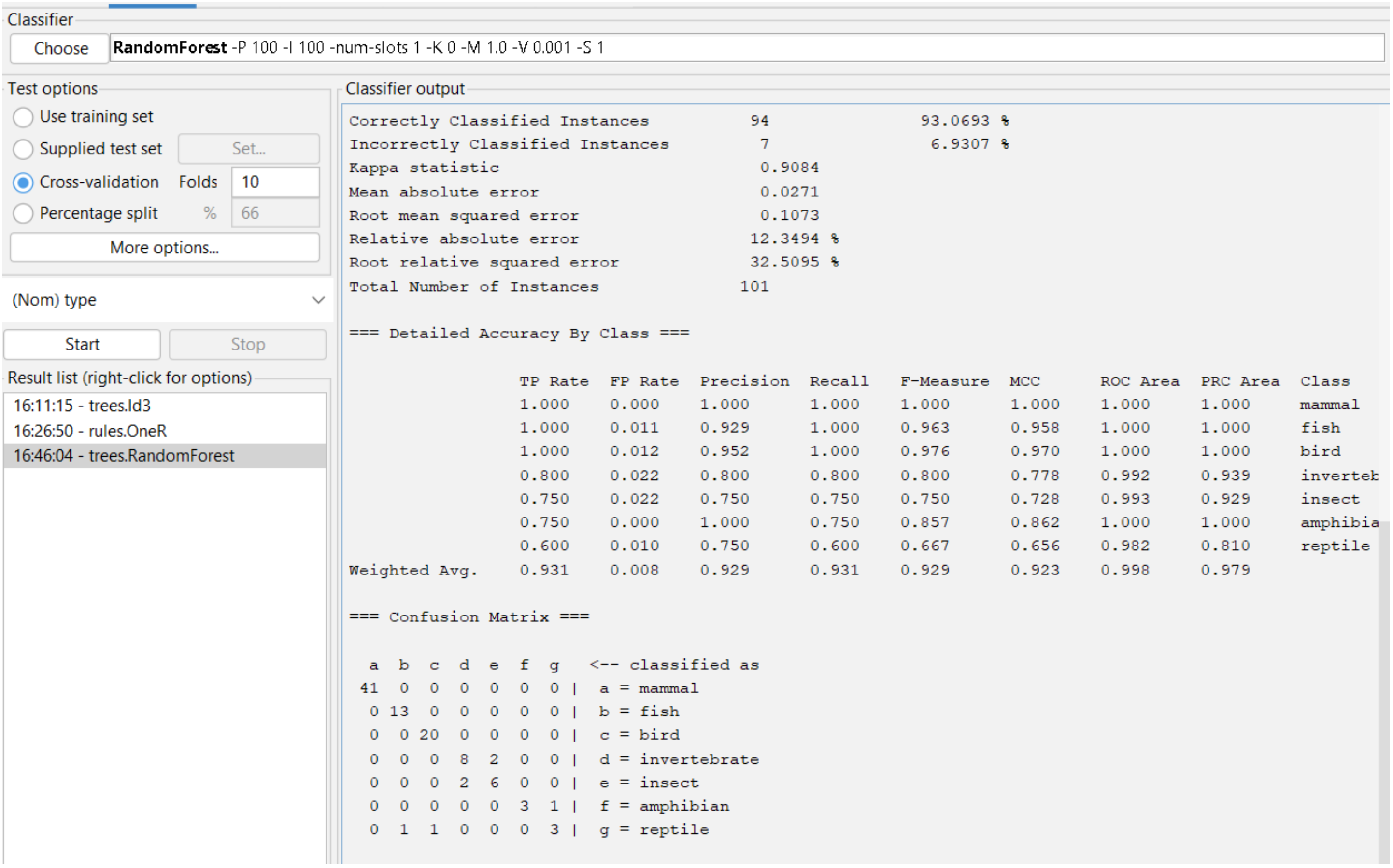


```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
 13  0  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
 10  0  0  0  0  0  0 |  d = invertebrate
  8  0  0  0  0  0  0 |  e = insect
  4  0  0  0  0  0  0 |  f = amphibian
  5  0  0  0  0  0  0 |  g = reptile
```

13 instances which are actually fish, are classified as mammals.

10 invertebrate animals,8 insects,4 amphibians and 5 reptiles are classified as mammals.

(08) Random Forest algorithm



Classifier

Choose RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

More options...

(Nom) type

Start Stop

Result list (right-click for options)

16:11:15 - trees.Id3

16:26:50 - rules.OneR

16:46:04 - trees.RandomForest

Classifier output

```
Correctly Classified Instances          94               93.0693 %
Incorrectly Classified Instances         7                6.9307 %
Kappa statistic                          0.9084
Mean absolute error                      0.0271
Root mean squared error                  0.1073
Relative absolute error                 12.3494 %
Root relative squared error             32.5095 %
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     mammal
                 1.000    0.011    0.929      1.000    0.963      0.958    1.000     1.000     fish
                 1.000    0.012    0.952      1.000    0.976      0.970    1.000     1.000     bird
                 0.800    0.022    0.800      0.800    0.800      0.778    0.992     0.939     invertek
                 0.750    0.022    0.750      0.750    0.750      0.728    0.993     0.929     insect
                 0.750    0.000    1.000      0.750    0.857      0.862    1.000     1.000     amphibia
                 0.600    0.010    0.750      0.600    0.667      0.656    0.982     0.810     reptile
Weighted Avg.    0.931    0.008    0.929      0.931    0.929      0.923    0.998     0.979

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 13  0  0  0  0  0 |  b = fish
  0  0 20  0  0  0  0 |  c = bird
  0  0  0  8  2  0  0 |  d = invertebrate
  0  0  0  2  6  0  0 |  e = insect
  0  0  0  0  0  3  1 |  f = amphibian
  0  1  1  0  0  0  3 |  g = reptile
```

According to the above figure we can see that the accuracy of the random forest algorithm is 93.0693%. 7 instances are misclassified. This has a high accuracy compared to the ID3 and OneR algorithms.