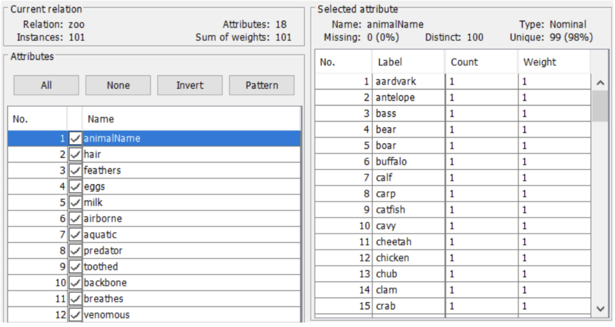
CO544 : Machine Learning and Data Mining
Lab 05: Classification, Predictions, Clustering and Association Learning
Nawarathna K.G.I.S.
E/17/219

Part 1 : Classification using WEKA

1. Load the Zoo dataset, Observe the attributes and their values.



Current relation
Relation: zoo
Instances: 101
Attributes: 18
Sum of weights: 101

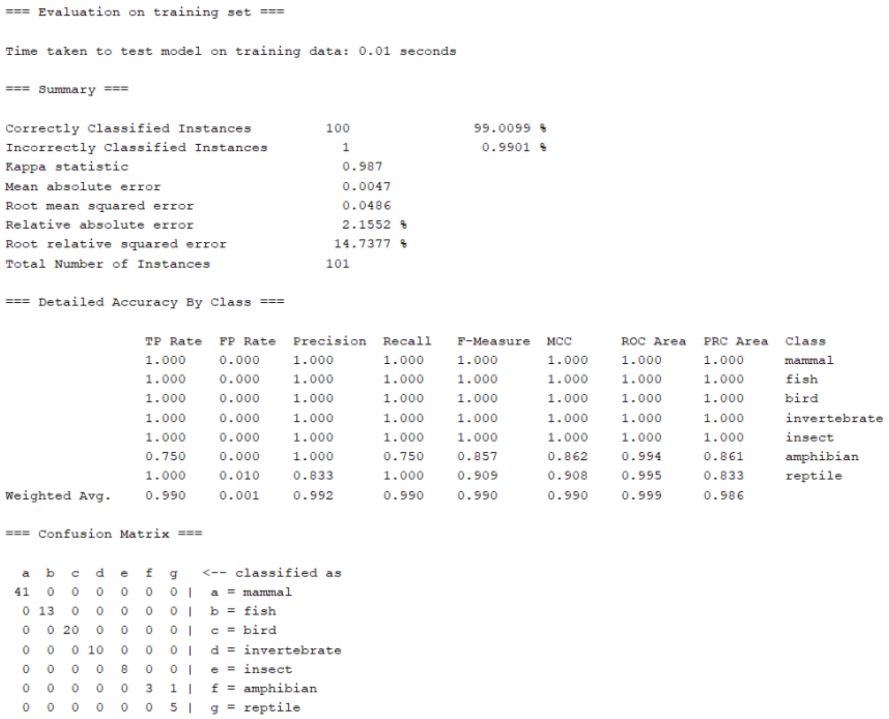
Selected attribute
Name: animalName
Missing: 0 (0%)
Distinct: 100
Type: Nominal
Unique: 99 (98%)

Attributes

No.		Name
1	☑	animalName
2	☑	hair
3	☑	feathers
4	☑	eggs
5	☑	milk
6	☑	airborne
7	☑	aquatic
8	☑	predator
9	☑	toothed
10	☑	backbone
11	☑	breathes
12	☑	venomous

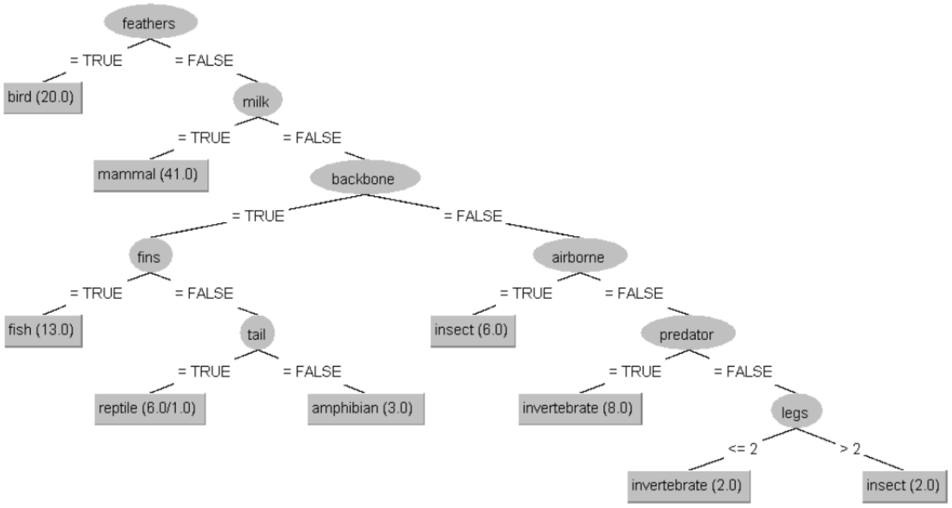
No.	Label	Count	Weight
1	aardvark	1	1
2	antelope	1	1
3	bass	1	1
4	bear	1	1
5	boar	1	1
6	buffalo	1	1
7	calf	1	1
8	carp	1	1
9	catfish	1	1
10	cavy	1	1
11	cheetah	1	1
12	chicken	1	1
13	chub	1	1
14	clam	1	1
15	crab	1	1

2. Build the C4.5 decision tree using default parameters and test options. Obesere the output of the algorithm.



```
=== Evaluation on training set ===

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
```

3. Visualise the output of C4.5 by right-clicking on the experiment in the result list and then choosing the Visualie tree option.Examine the true positive (TP) rates , the false positive(FP) rates and the confusion matrix. Explain misclassification observed in the confusion matrix.



Evaluation parameters(from Q2 answer)

Mean Abs. error : 0.0047

Root mean squared error : 0.0486

Relative absolute error : 2.1552 %

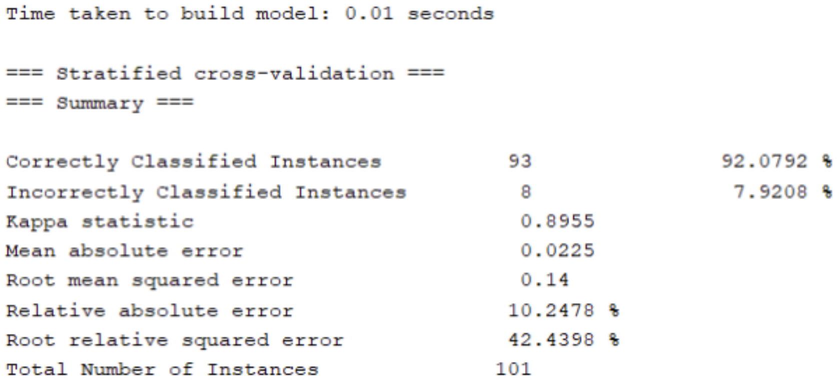
Root relative squared error : 14.7377 %

Classification accuracy : 99.0099 %

True and false positive rates for each column, Confusion matrix can be found using the answer to Q2 and only one miscalculation is found on the confusion matrix.

4. Evaluate C4.5 algorithm using the following testing options.
   a. The training set
   b. 10-fold cross validation

Record the classification accuracies using both the methods. WHich one provides more realistic future performance? Why?



```
Time taken to build model: 0.01 seconds

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

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

Accuracy and evaluation criteria for C4.5 Training set model is obtained previously.

For 10-Fold Cross validation :

Accuracy : 92.0792%

10- Fold cross validation model is more usable for future usages.

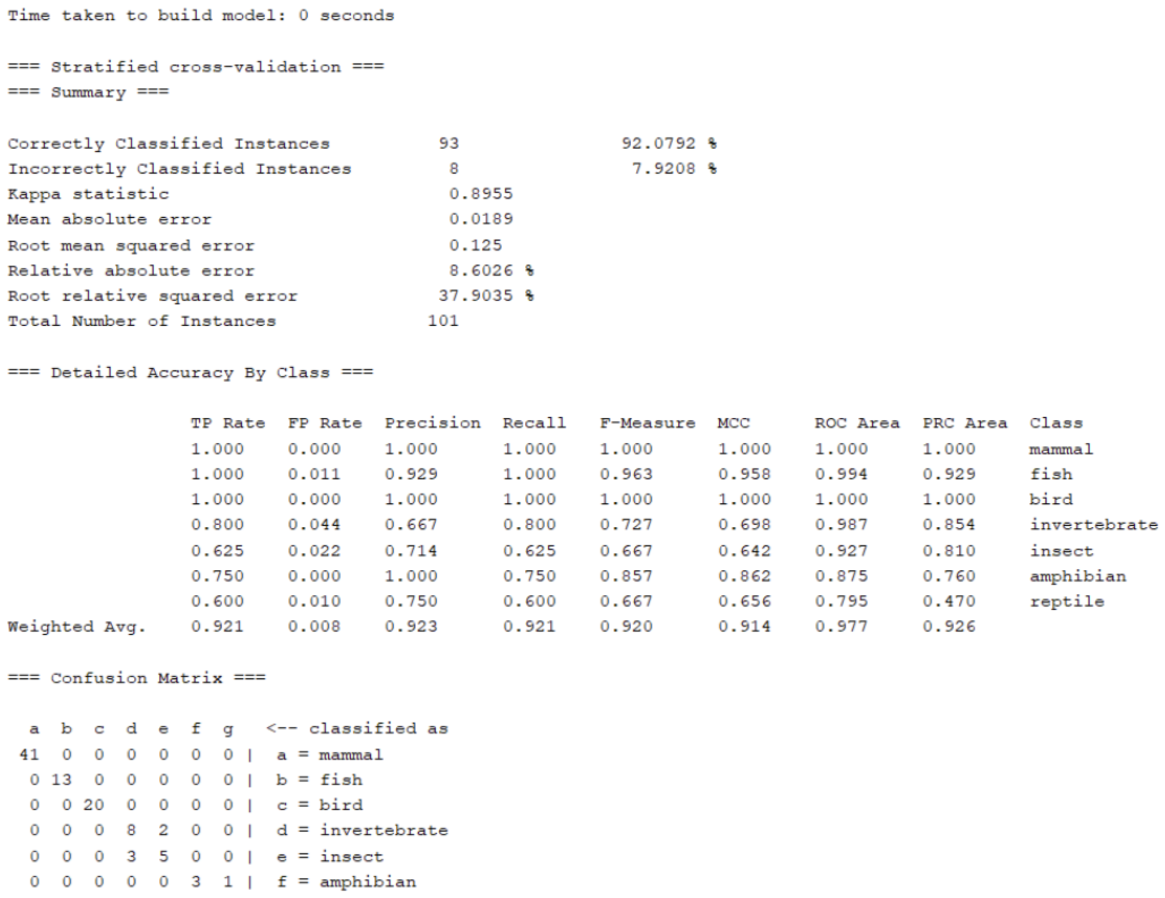
10-fold cross validation also seems better in reliability considering the classification accuracy since this gives 92% accuracy on the train set. It is also worth pointing out that, training set model is tested using the same dataset.

5. Can you apply the ID3 (Iterative Dichotomiser 3) learning algorithm on this dataset? Explain your answer.

It is not possible to use ID3 for this.
C4.5 works with both discrete and continuous values but ID3 is only usable for nominal values.

7. Build the ID3 decision tree. Examine the output. Record the 10-fold Cross Validation accuracy.

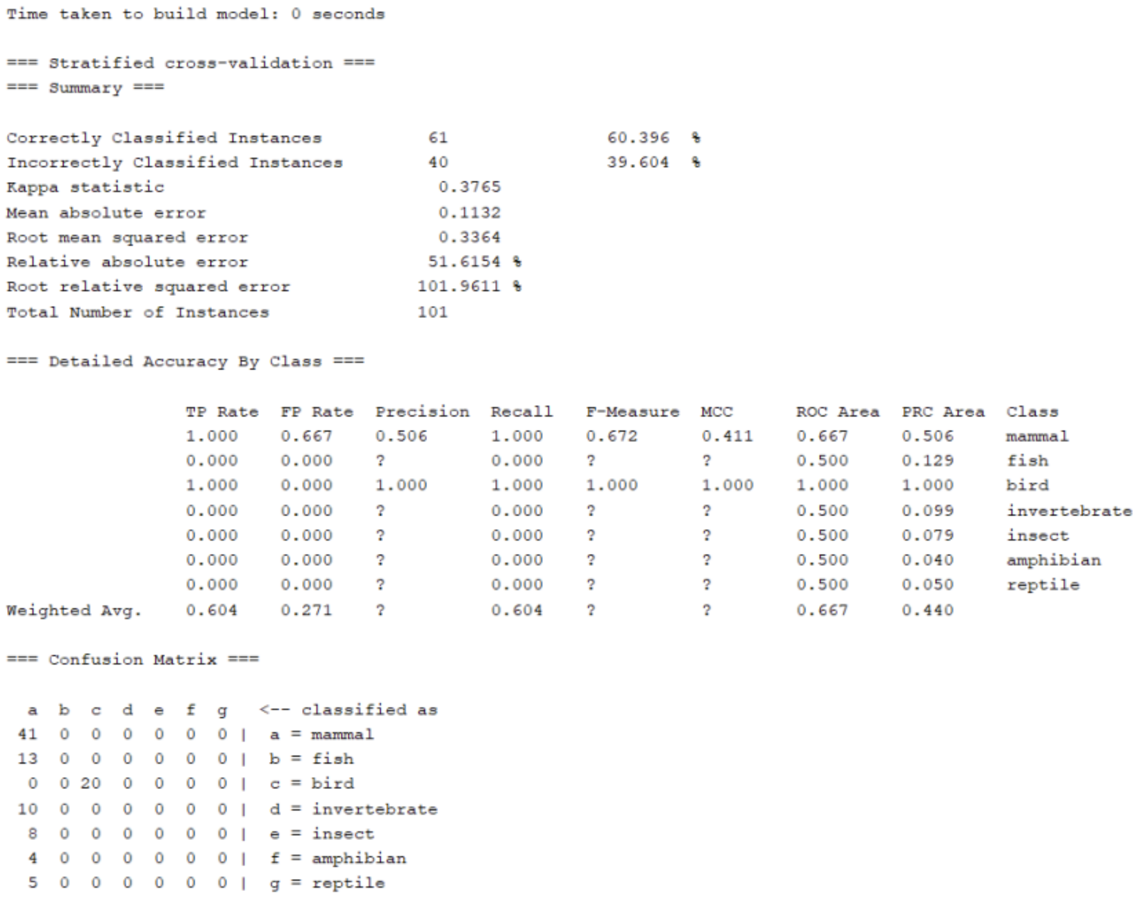
Accuracy : 92.0792% (93/101)



```
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                      0.0189
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     invertebrate
                 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     amphibian
                 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
```

8. Use the OneR algorithm and explain the classifier output. Record the 10-fold Cross Validation accuracy.

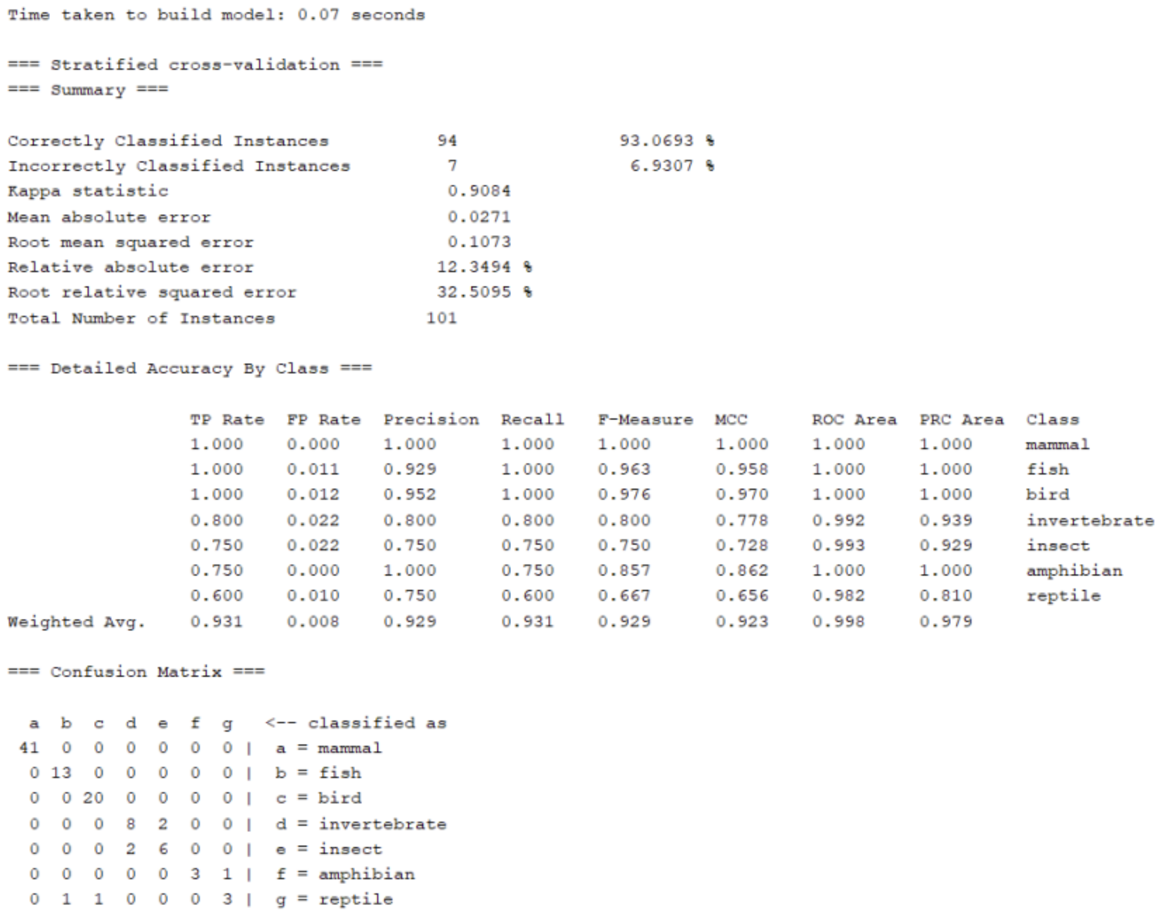
Accuracy : 60.396% (61/101)



```
Time taken to build model: 0 seconds

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

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     invertebrate
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.079     insect
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.040     amphibian
                 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
```

9. Use another classification algorithm of your choice and observe the output of the algorithm. Compare the results of the chosen algorithm with previous outputs.

Accuracy : 93.0693% (94/101)



```
Time taken to build model: 0.07 seconds

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

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     invertebrate
                 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     amphibian
                 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
```