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

Part 2 : Predicting Class Values

1. Observe the output of the algorithm with the training set. Explore different error estimates and record the percentages of misclassifications and classifications.

Misclassification % (Training set) = 98.7654%
Classification % (Training set) = 01.2346%

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

=== Summary ===

Correctly Classified Instances          80               98.7654 %
Incorrectly Classified Instances         1                1.2346 %
Kappa statistic                          0.9831
Mean absolute error                      0.0062
Root mean squared error                  0.0556
Relative absolute error                  2.9101 %
Root relative squared error             17.1647 %
Total Number of Instances               81

=== 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     bird
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     reptile
                 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     amphibian
                 0.833    0.000    1.000      0.833    0.909      0.907    0.992     0.910     insect
                 1.000    0.014    0.875      1.000    0.933      0.929    0.993     0.875     invertebrate
Weighted Avg.    0.988    0.001    0.989      0.988    0.988      0.987    0.999     0.983
```

2. Observe the output of the algorithm with the test data. Explore different error estimates and record the percentages of misclassifications and classifications.

Misclassification % (Supplied test set) = 85.0000%
Classification % (Supplied test set) = 15.0000%

```
Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances          17               85      %
Incorrectly Classified Instances         3               15      %
Kappa statistic                          0.8187
K&B Relative Info Score                 77.3631 %
K&B Information Score                   43.1675 bits      2.1584 bits/instance
Class complexity | order 0              55.7985 bits      2.7899 bits/instance
Class complexity | scheme             2151.5779 bits    107.5789 bits/instance
Complexity improvement     (Sf)      -2095.7794 bits   -104.789  bits/instance
Mean absolute error                      0.0464
Root mean squared error                  0.1965
Relative absolute error                 20.0843 %
Root relative squared error             55.849  %
Total Number of Instances               20

=== 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     bird
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.100     reptile
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     fish
                 1.000    0.053    0.500      1.000    0.667      0.688    0.974     0.500     amphibian
                 0.500    0.000    1.000      0.500    0.667      0.688    0.944     0.667     insect
                 1.000    0.118    0.600      1.000    0.750      0.728    0.941     0.600     invertebrate
Weighted Avg.    0.850    0.020    ?          0.850    ?          ?        0.934     0.792
```

3. Comment on the two results you observed.

- It can be said that, considering the classification percentages , the training set model is more accurate than the test set and it is also worth noticing that the test dataset classification percentage is calculated on unseen data.Therefore test dataset results are closer to the future predictions of the algorithm since future predictions are also unseen for the algorithm.

4. Do the predictions using zoo_test_classmissing.arff as the test set. This file has the same data set with missing class values. Comment on your results with respect to the results you obtained in the step 2 above.

```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Total Number of Instances                0
Ignored Class Unknown Instances                 20

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 ?        ?        ?          ?        ?          ?        ?         ?         mammal
                 ?        ?        ?          ?        ?          ?        ?         ?         bird
                 ?        ?        ?          ?        ?          ?        ?         ?         reptile
                 ?        ?        ?          ?        ?          ?        ?         ?         fish
                 ?        ?        ?          ?        ?          ?        ?         ?         amphibian
                 ?        ?        ?          ?        ?          ?        ?         ?         insect
                 ?        ?        ?          ?        ?          ?        ?         ?         invertebrate
Weighted Avg.    ?        ?        ?          ?        ?          ?        ?         ?
```

All the targets are missing in the dataset. None of the 20 are available. Due to this, the algorithm can't make any predictions.