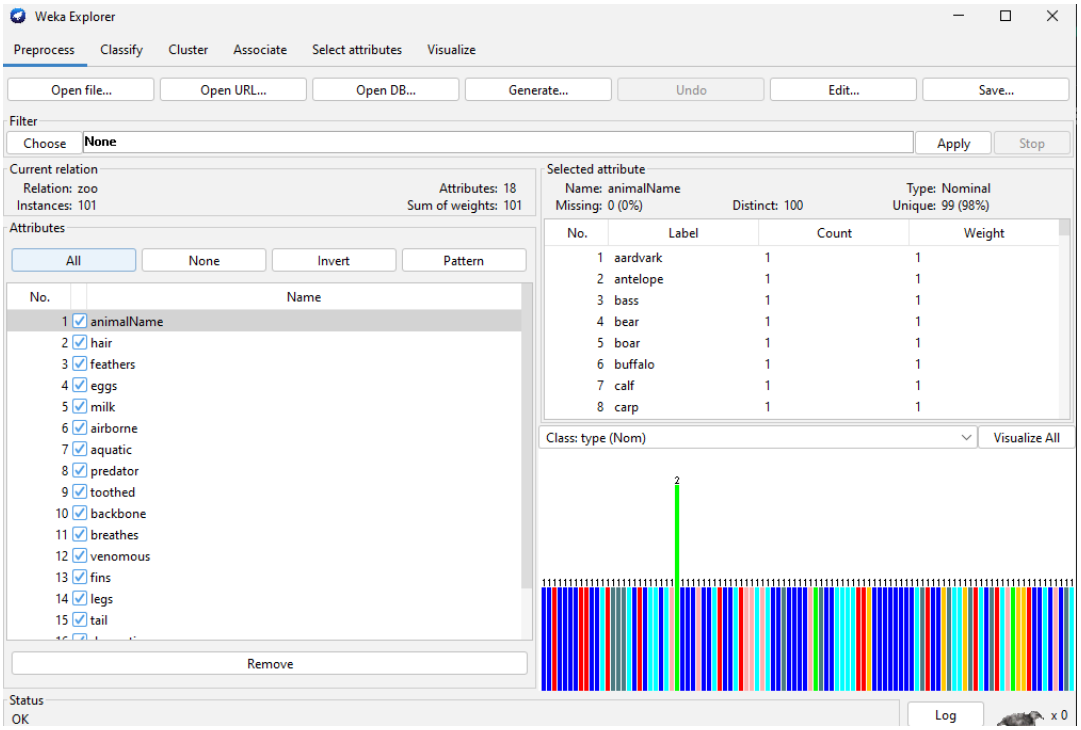
E/17/153 : Part 1 – Classification using WEKA

1.As it shows here there are 101 instances and 18 attributes in the “Zoo” dataset.



2. Output of the C4.5 algorithm:

Correctly classified instances = 100

Incorrectly classified instances = 1

3. Accuracy = 99.0099%

Mean absolute error = 0.0047

Classifier output

```
=== 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
```

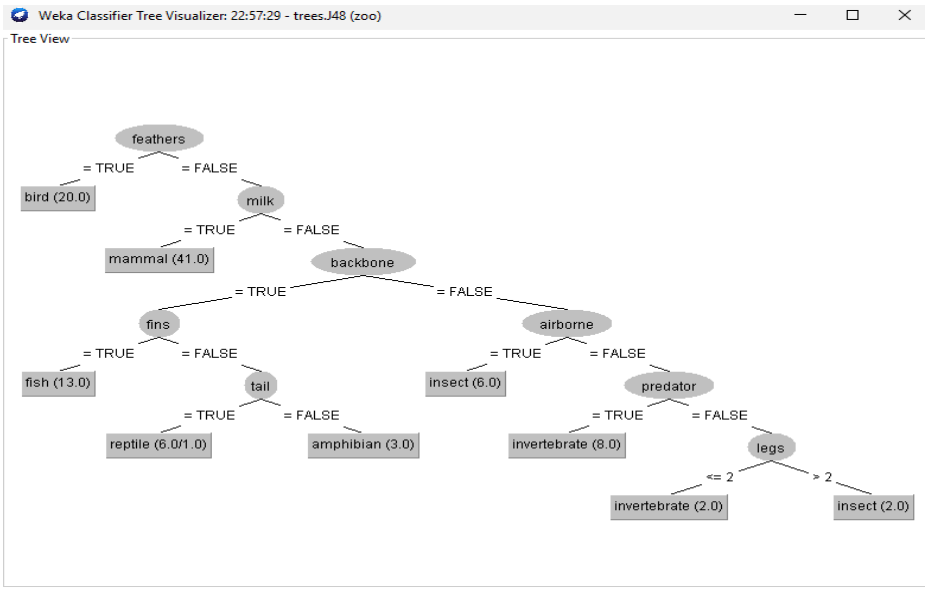
These are the True Positive and False Positive rates for each class. As it is observed, weighted average TP rate is 0.990 and weighted average FP rate is 0.001. The only FP happened involved in the 'Reptile' class.

```
=== 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     invertebra
               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
```

This is the confusion matrix obtained from the classification. If the accuracy is hundred percent only diagonal elements are filled with the numbers in the confusion matrix. But here there is an element in (g,f) position in addition to the diagonal elements. This means 1 'f' classed animal (amphibian) has classified as a 'g' classed (reptile) animal. This detail was partially obtained from the FP rate column in the previous table too.

```
=== 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
```

Decision tree



4.

Training set

```
=== Summary ===

Correctly Classified Instances         100               99.0099 %
Incorrectly Classified Instances         1                0.9901 %
```

10-fold cross validation

```
=== Summary ===

Correctly Classified Instances          93               92.0792 %
Incorrectly Classified Instances         8                7.9208 %
```

When comparing accuracy values, it can be seen that,

Accuracy (Training set) > Accuracy (10-cross validation)

Hence, training test model will give a better future performance with compare to 10-fold cross validation model.

Cross validation is usually used for small datasets. Here, it randomly divides the set of observations into 10 folds and one of them is treated as test set. This is run 10 times. Since this run 10 times shown performance is the average across 10 times. So, the incorrectly classified instances might be higher than in the training set.

5. We can't apply ID3 learning algorithm on this dataset since this algorithm only deals with nominal attributes. Here we got an attribute which has numerical values. So, the original dataset does not support the ID3 algorithm.

7. After removing animal name and legs attributes, ID3 decision tree was built. As the summary shows, 93 instances were classified correctly and 8 were incorrect. So, the accuracy was 92.0792% and this is still a lesser accuracy than the training set model.

```
=== 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     invertebra
               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
```

By studying the confusion matrix, incorrectly classified data could be identified.

```
=== 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. OneR algorithm

Only 61 instances were classified correctly and it decreased the accuracy into 60.396%

```
=== 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     invertebra
               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
```

From the confusion matrix we can see that only mammals and birds have classified into their true category and all the others are mis classifies.

```
=== 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
```

9. Prism algorithm

PRISM is a separate and conquer algorithm based on ID3's cons.

ID3 doesn't consider whether an attribute might be highly relevant to only one classification and irrelevant to the others. In PRISM a branch could be considered as an attribute-value pair. It considers the relevance between an attribute-value pair and the specific classification.

Since there is no big effect of the relevance in this dataset, PRISM algorithm doesn't show much difference in the classification compared to ID3. But it obviously gives a better performance than OneR algorithm in this case.

```
Correctly Classified Instances          92               91.0891 %
Incorrectly Classified Instances         5                4.9505 %
Kappa statistic                          0.9307
Mean absolute error                      0.0147
Root mean squared error                  0.1214
Relative absolute error                  7.06   %
Root relative squared error             37.8906 %
UnClassified Instances                   4                3.9604 %
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
               0.889    0.034    0.727      0.889    0.800      0.782    0.884     0.602     invertebra
               0.625    0.000    1.000      0.625    0.769      0.778    0.813     0.655     insect
               1.000    0.011    0.750      1.000    0.857      0.861    0.870     0.572     amphibian
               0.667    0.011    0.667      0.667    0.667      0.656    0.695     0.296     reptile
Weighted Avg.  0.948    0.004    0.957      0.948    0.948      0.947    0.960     0.900
```

```
=== 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  0  0  1 |  d = invertebrate
  0  0  0  3  5  0  0 |  e = insect
  0  0  0  0  0  3  0 |  f = amphibian
  0  0  0  0  0  1  2 |  g = reptile
```