Diplomado en Minería de Datos

PEUVI, Facultad de Ciencias, UNAM

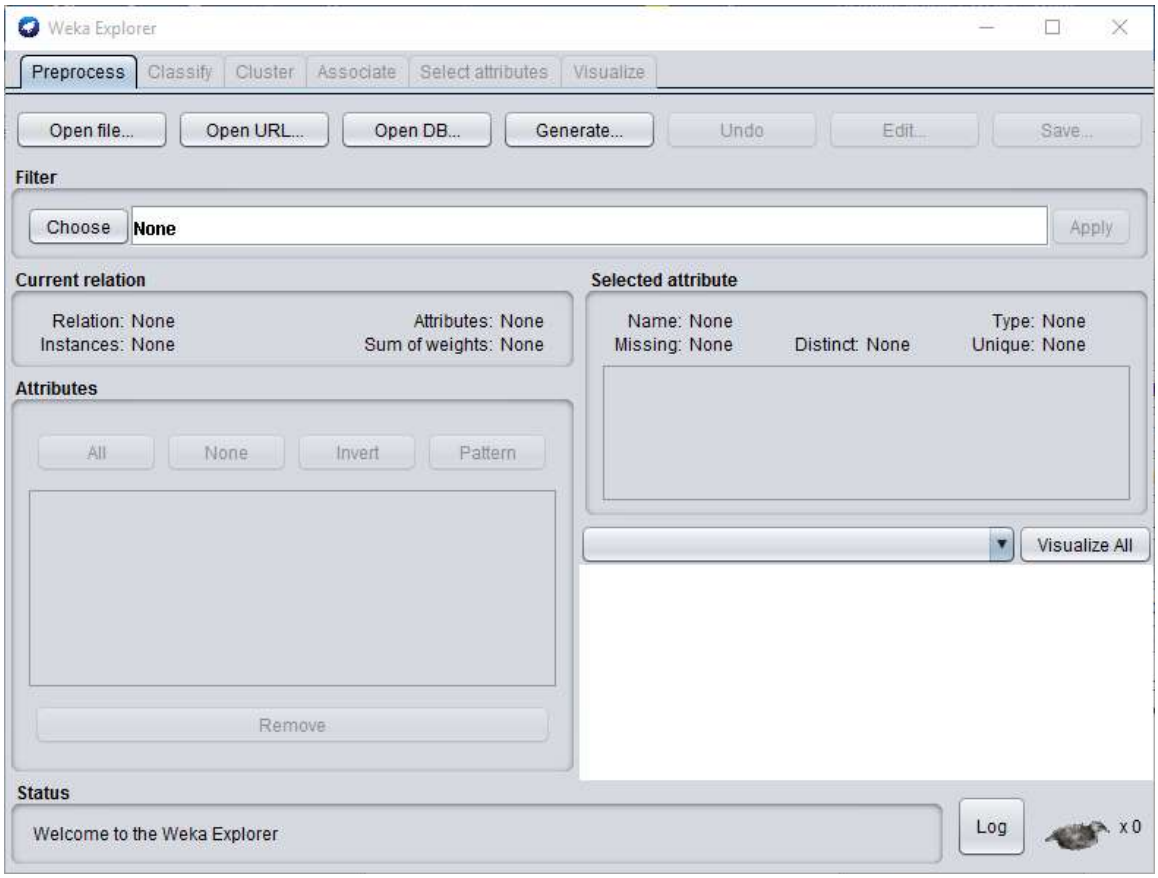
M.I. Gerardo Avilés Rosas

gar@ciencias.unam.mx

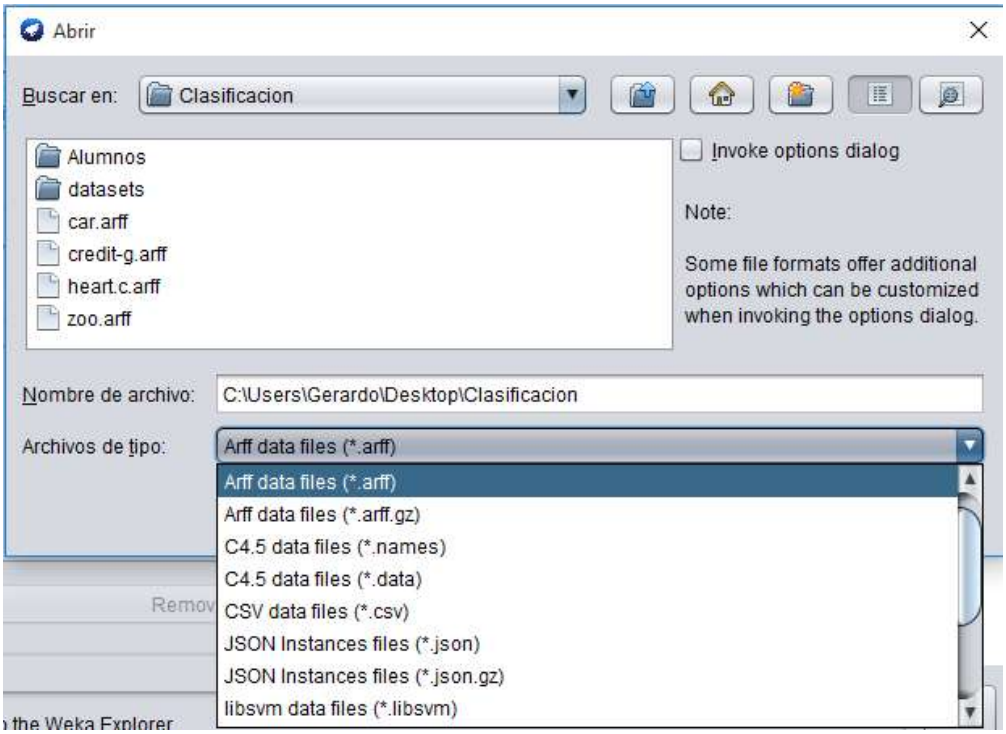
Módulo 5 **Minería de Datos**
Ejemplo de árboles de decisión con Weka

1. Iniciamos **Weka**. Una vez que arrancó, vamos a ejecutar la aplicación **Explorer**:

2. Se muestra la siguiente ventana. Como se puede observar, la mayoría de las opciones se encuentran desactivadas ya que no se ha cargado ningún conjunto de datos:



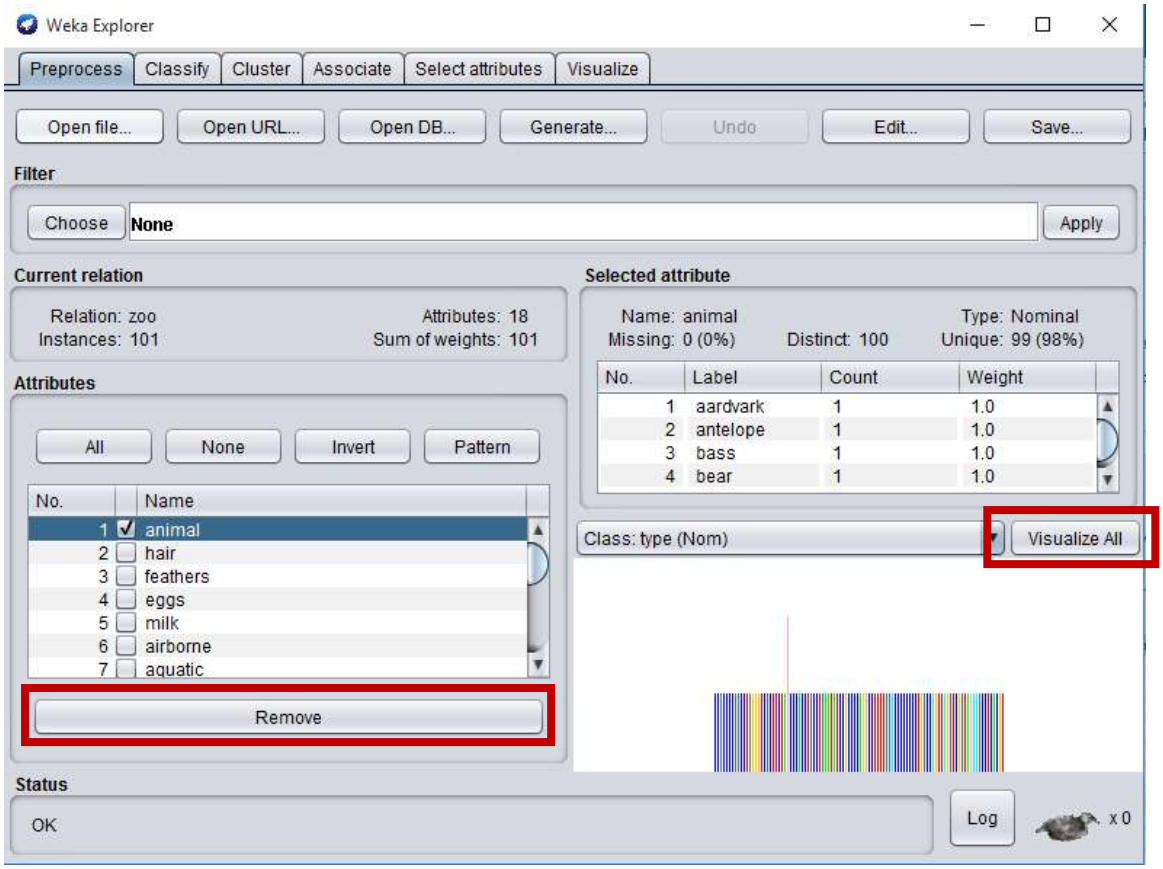
3. Vamos a dar clic en la opción Open file y buscamos el dataset **zoo.arff**, que es un formato nativo de **Weka**. Se trata de un archivo conocido como un **archivo en formato atributo-relación**, es un archivo de texto ASCII que describe una lista de instancias que comparten un conjunto de atributos. Fue desarrollado por el Departamento de Ciencias de la Computación de la Universidad de Waikato para su uso con el software **Weka**:



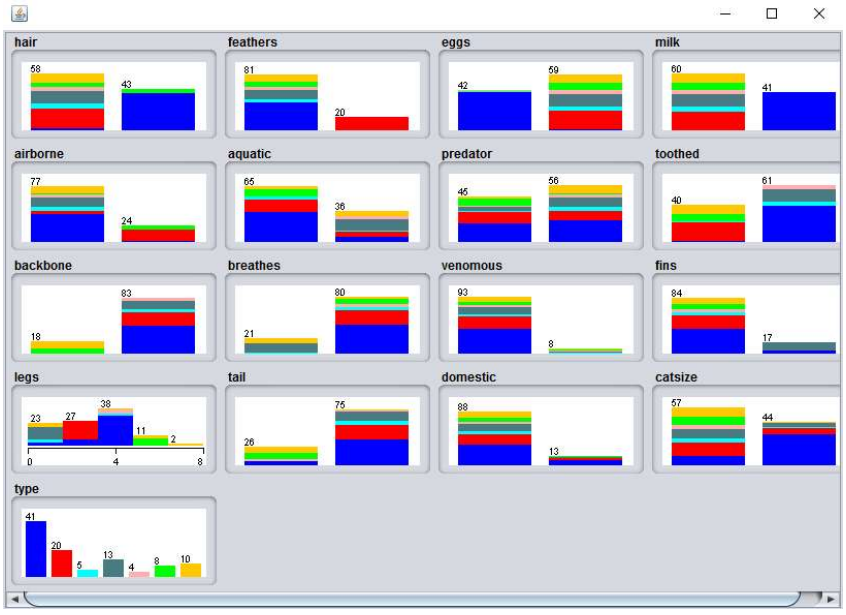
Un ejemplo de este tipo de archivo lo encontrarán a continuación (para el **dataset IRIS**):

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

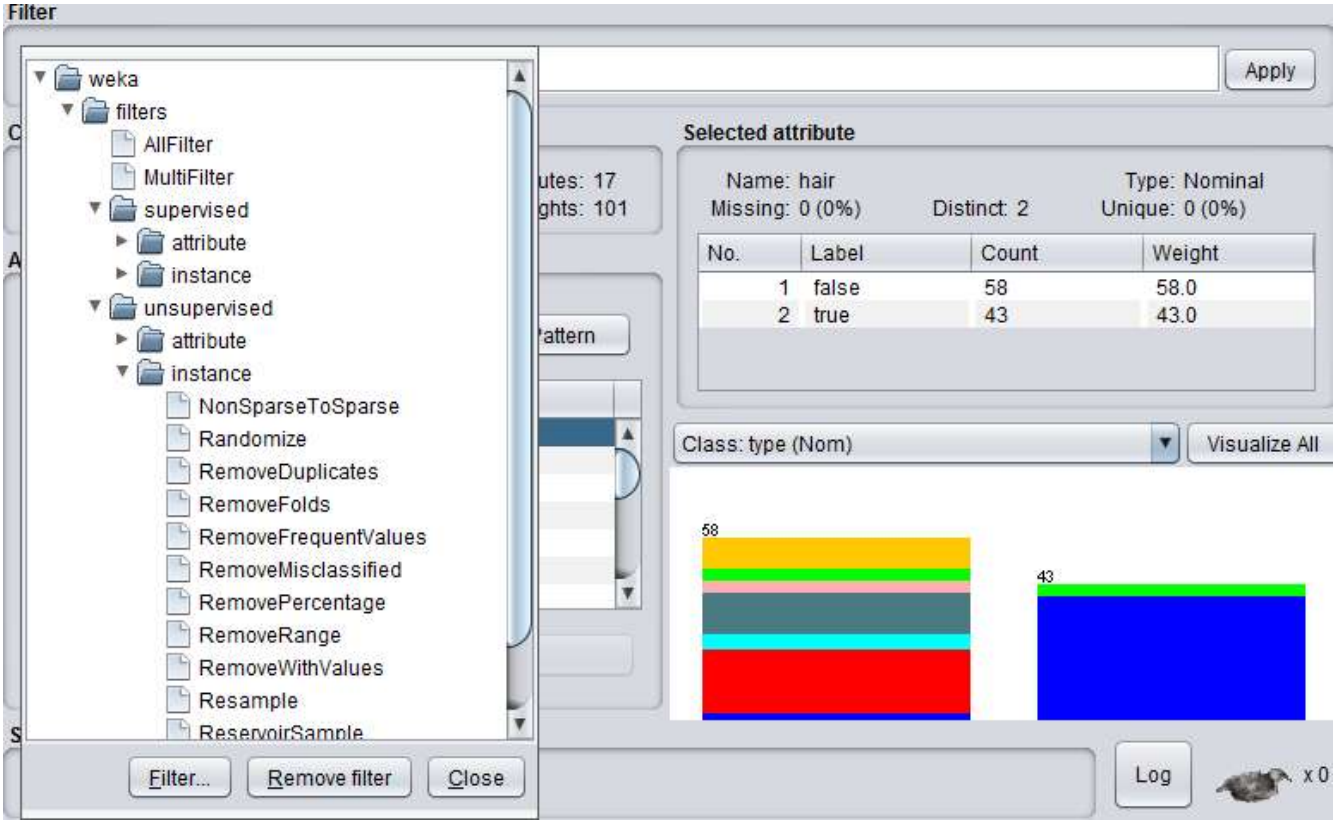
4. Una vez cargado el dataset, vamos a seleccionar el **atributo Animal** y damos clic en la opción **Remove** (se encuentra debajo de la lista de atributos). Esto es para evitar que el árbol de sobreajuste por la cantidad de valores único que tiene (100). No se recomienda agregar al modelo de clasificación atributos que tenga valores del tipo llave primaria, para no crear este efecto:



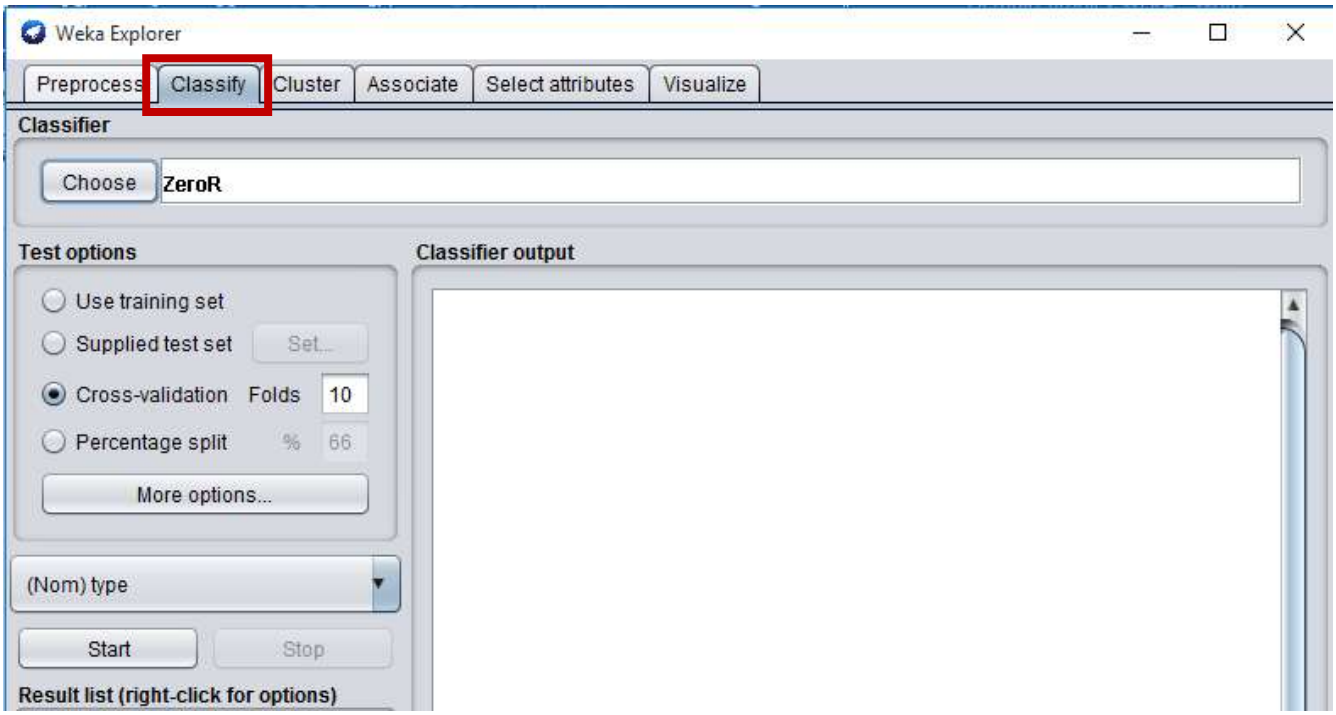
Desde esta ventana, si damos clic en la opción **Visualize all**, podremos ver el comportamiento que tienen las variables que se tienen en el dataset (sin la última que acabamos de quitar):



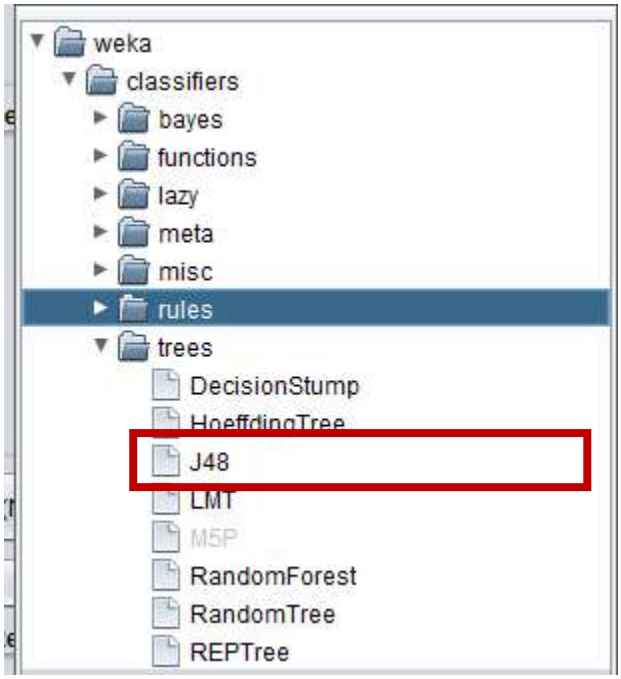
5. Vamos a dejar las variables sin modificar, sin embargo, desde esta perspectiva, tenemos acceso a la sección de **Filtros (Filter)**, que se utilizan para aplicar preprocesamiento de los datos que se tengan actualmente cargados en **Weka**. Los vamos a encontrar en sus versiones supervisados y no supervisados:



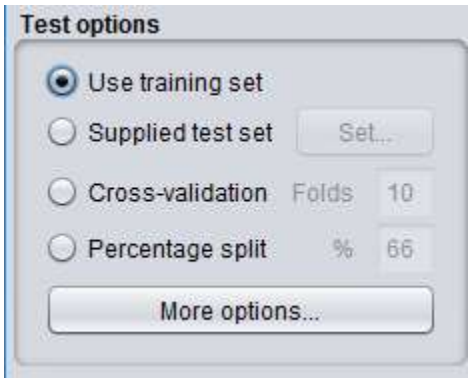
6. Nos vamos a cambiar a la pestaña **Clasificación**:



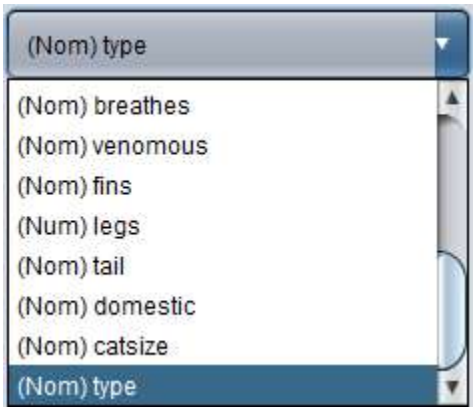
7. Vamos a dar clic en **Choose**, para tener acceso a todos los algoritmos de clasificación. En este caso, iremos a la categoría **trees** y seleccionamos **J48**, que es una implementación del algoritmo **C4.5** que revisamos en clase:



8. Una vez seleccionado, en la sección de **Test options**, vamos a utilizar la primera opción (**Use training set**), que corresponde con una **evaluación del modelo**, utilizando la técnica de **resustitución**. En esta técnica, se hace una evaluación del modelo de clasificación seleccionado, utilizando el mismo conjunto de tuplas con el que se entrenó dicho modelo. No se recomienda esta forma de evaluación, ya que arroja resultados **muy optimistas**:



9. Con esto, estamos listos para poder generar el modelo, para esto, vamos a dar clic en el botón **Start**. Debemos asegurarnos que se tenga selecciona la variable objetivo adecuada, es decir, aquella que tiene las etiquetas de clase. Para el ejemplo que vamos a utilizar, tenemos un dataset que tiene un conjunto de características sobre **100 animales** distintos: ***si tienen pelo, plumas, si ponen huevos, si dan leche, si son acuáticos, si son depredadores, si tienen columna vertebral,*** entre otros. Se desea saber, con base en estas características, **¿qué tipo de animal es?**, las opciones son: ***pez, ave, mamífero, réptil, anfibio, insecto o invertebrado***:



10. Una vez que termina el entrenamiento, podemos ver los resultados arrojados por **Weka**. En primer lugar, encontramos información sobre el modelo utilizado, las variables involucradas, sí se aplicó o no un preprocesamiento y la forma de evaluar el modelo:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     zoo-weka.filters.unsupervised.attribute.Remove-R1
Instances:    101
Attributes:   17
              hair
              feathers
              eggs
              milk
              airborne
              aquatic
              predator
              toothed
              backbone
              breathes
              venomous
              fins
              legs
              tail
              domestic
              catsize
              type
Test mode:    evaluate on training data
```

En segundo lugar, características del árbol: indica sí el árbol esta podado, el árbol resultante, número de nodos hoja y el tamaño del árbol:

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

J48 pruned tree
------------------

feathers = false
|   milk = false
|   |   backbone = false
|   |   |   airborne = false
|   |   |   |   predator = false
|   |   |   |   |   legs <= 2: invertebrate (2.0)
|   |   |   |   |   legs > 2: insect (2.0)
|   |   |   |   predator = true: invertebrate (8.0)
|   |   |   airborne = true: insect (6.0)
|   |   backbone = true
|   |   |   fins = false
|   |   |   |   tail = false: amphibian (3.0)
|   |   |   |   tail = true: reptile (6.0/1.0)
|   |   |   fins = true: fish (13.0)
|   milk = true: mammal (41.0)
feathers = true: bird (20.0)

Number of Leaves  :     9

Size of the tree :      17
```

Enseguida encontramos información de la tasa de la precisión, el error de clasificación y estadísticas detalladas sobre la precisión en cada una de las clases:

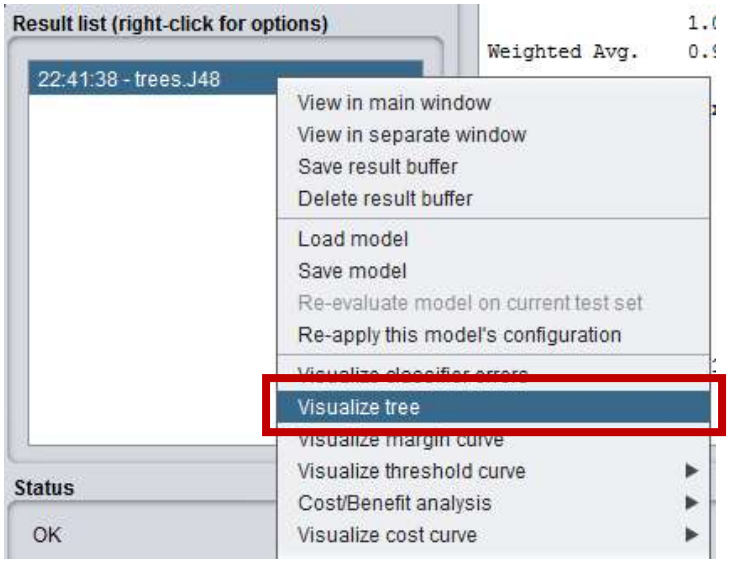
```
=== Summary ===

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

Finalmente encontramos la matriz de confusión, que nos da información sobre las tuplas correctamente clasificadas y en el caso de los errores, podemos saber dónde se confundió el modelo:

```
=== Confusion Matrix ===

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

Más adelante estudiaremos lo aspectos en cuanto evaluación.

11. Finalmente, podemos ver el árbol generado, para esto, vamos a la ventana de resultados y damos clic derecho sobre el modelo ejecutado. Se mostrará el siguiente menú y daremos clic en la opción **Visualize tree**:



12. El árbol generado por Weka es: