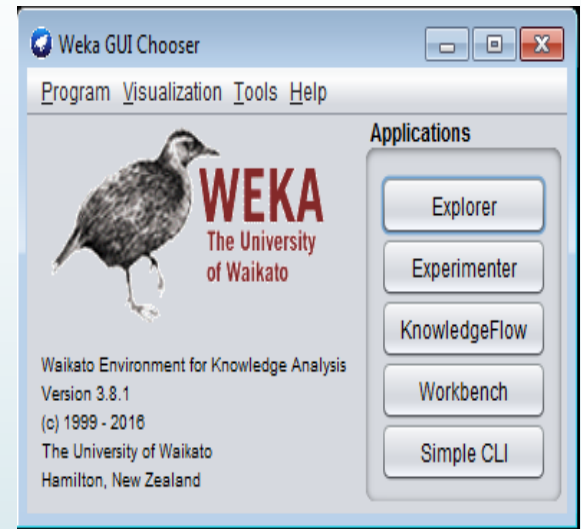


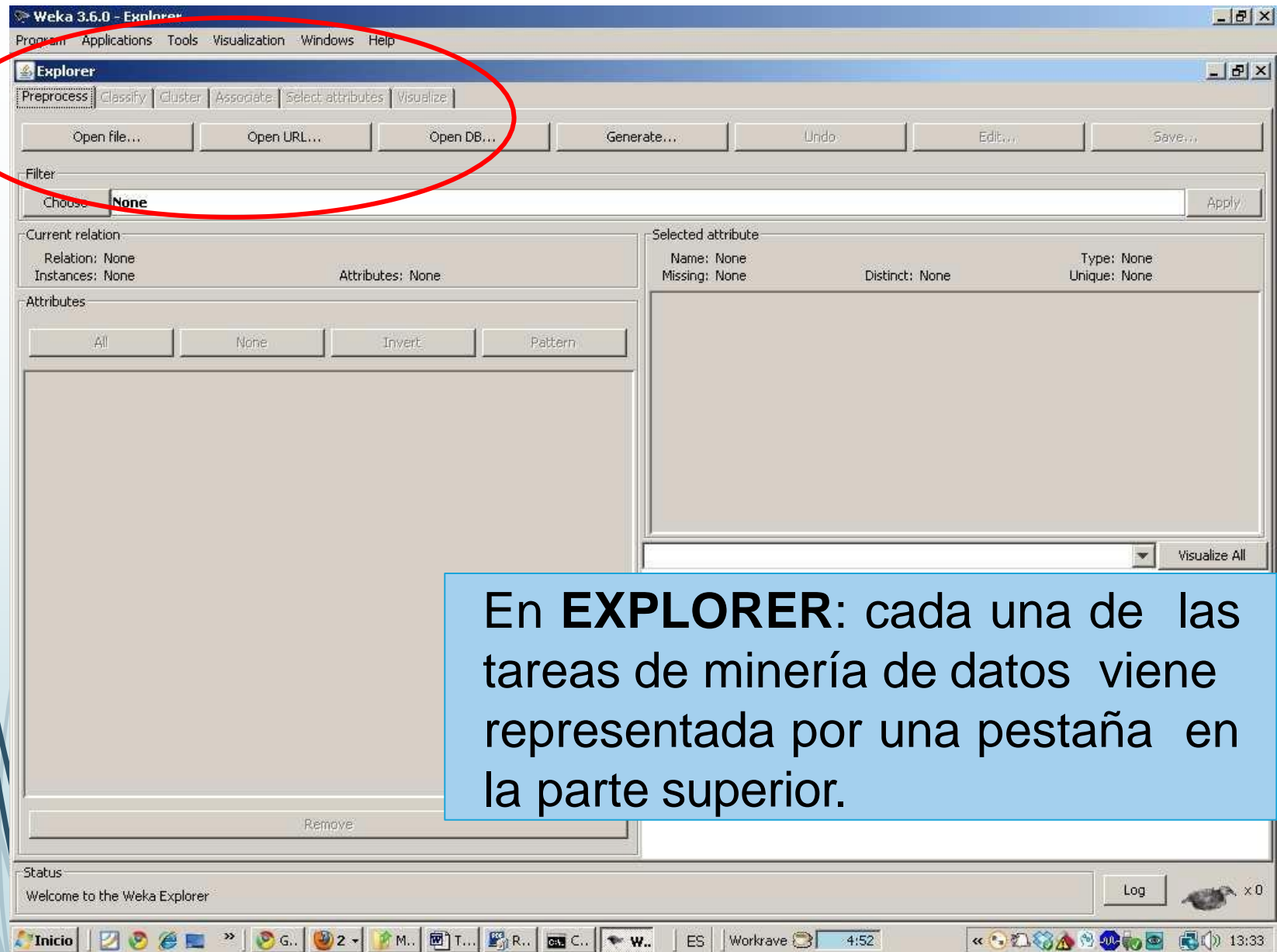
UNSE – 2019



Weka (Waikato Environment for Knowledge Analysis - *Entorno para Análisis del Conocimiento de la Universidad de Waikato*) es una plataforma de software para aprendizaje automático y minería de datos escrito en Java. Weka es un software libre distribuido bajo licencia GNU-GPL.

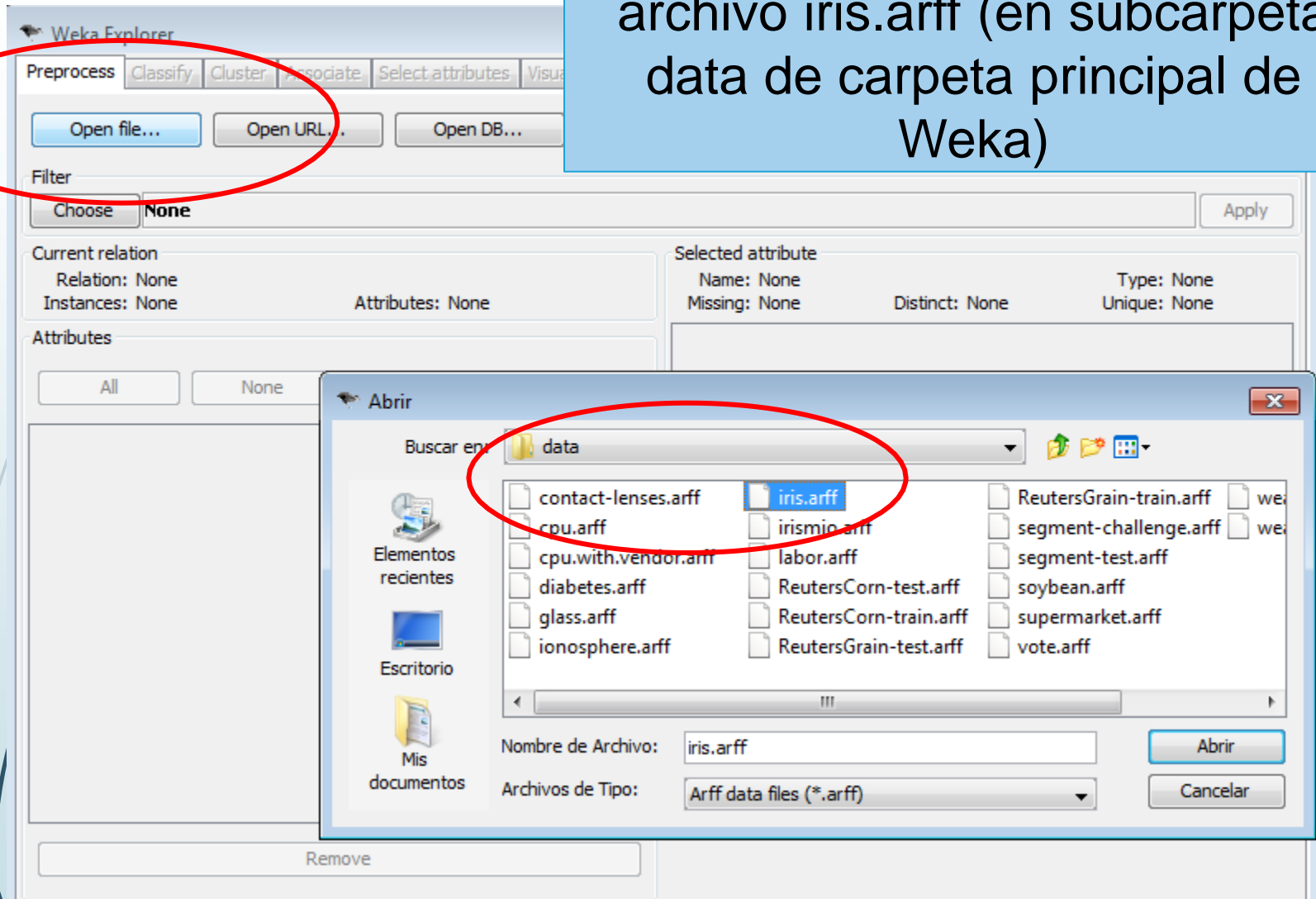
- **Explorer:** es la opción que permite llevar a cabo la ejecución de los algoritmos de análisis implementados sobre los ficheros de entrada, una ejecución independiente por cada prueba.
- **Experimenter:** esta opción permite definir experimentos más complejos, con objeto de ejecutar uno o varios algoritmos sobre uno o varios conjuntos de datos de entrada, y comparar estadísticamente los resultados
- **KnowledgeFlow:** permite llevar a cabo las mismas acciones del "Explorer", con una configuración totalmente gráfica, inspirada en herramientas de tipo "data-flow" para seleccionar componentes y conectarlos en un proyecto de minería de datos, desde que se cargan los datos, se aplican algoritmos de tratamiento y análisis, hasta el tipo de evaluación deseada
- **Workbench:** Una aplicación todo-en-uno que combina todas las demás funcionalidades, dentro "perspectivas" seleccionables por el usuario
- **Simple CLI:** entorno de consola para invocar directamente con java a los paquetes de weka





En **EXPLORER**: cada una de las tareas de minería de datos viene representada por una pestaña en la parte superior.

Seleccionar **Open file** y abrir archivo iris.arff (en subcarpeta data de carpeta principal de Weka)



Objetivo: Construir un **clasificador** que clasifique plantas en tres clases: *setosa*, *versicolor* y *virginica*

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None** Apply

Current relation
Relation: iris
Instances: 150
Attributes: 5

Attributes: All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> sepalength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petallength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Remove

Selected attribute
Name: sepalength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

Bin Range	Count
4.3 - 4.8	16
4.8 - 5.3	30
5.3 - 5.8	34
5.8 - 6.3	28
6.3 - 6.8	25
6.8 - 7.3	10
7.3 - 7.9	7

Status: OK

Log x 0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file...

Filter
Choose None

Current relation
Relation: iris
Instances: 150

Attributes
All

No.	Name
1	<input checked="" type="checkbox"/> sepalwidth
2	<input checked="" type="checkbox"/> sepalwidth
3	<input checked="" type="checkbox"/> petalwidth
4	<input checked="" type="checkbox"/> petalwidth
5	<input checked="" type="checkbox"/> class

Status
See error log

Viewer

Relation: iris

No.	sepalwidth Numeric	sepalwidth Numeric	petalwidth Numeric	petalwidth Numeric	class Nominal
1	5.1	3.5	1.4	0.2	Iris-se...
2	4.9	3.0	1.4	0.2	Iris-se...
3	4.7	3.2	1.3	0.2	Iris-se...
4	4.6	3.1	1.5	0.2	Iris-se...
5	5.0	3.6	1.4	0.2	Iris-se...
6	5.4	3.9	1.7	0.4	Iris-se...
7	4.6	3.4	1.4	0.3	Iris-se...
8	5.0	3.4	1.5	0.2	Iris-se...
9	4.4	2.9	1.4	0.2	Iris-se...
10	4.9	3.1	1.5	0.1	Iris-se...
11	5.4	3.7	1.5	0.2	Iris-se...
12	4.8	3.4	1.6	0.2	Iris-se...
13	4.8	3.0	1.4	0.1	Iris-se...
14	4.3	3.0	1.1	0.1	Iris-se...
15	5.8	4.0	1.2	0.2	Iris-se...
16	5.7	4.4	1.5	0.4	Iris-se...
17	5.4	3.9	1.3	0.4	Iris-se...
18	5.1	3.5	1.4	0.3	Iris-se...
19	5.7	3.8	1.7	0.3	Iris-se...
20	5.1	3.8	1.5	0.3	Iris-se...
21	5.4	3.4	1.7	0.2	Iris-se...
22	5.1	3.7	1.5	0.4	Iris-se...
23	4.6	3.6	1.0	0.2	Iris-se...
24	5.1	3.3	1.7	0.5	Iris-se...

Undo Edit...

Distinct: 35 Type: Numer Unique: 9 (6%)

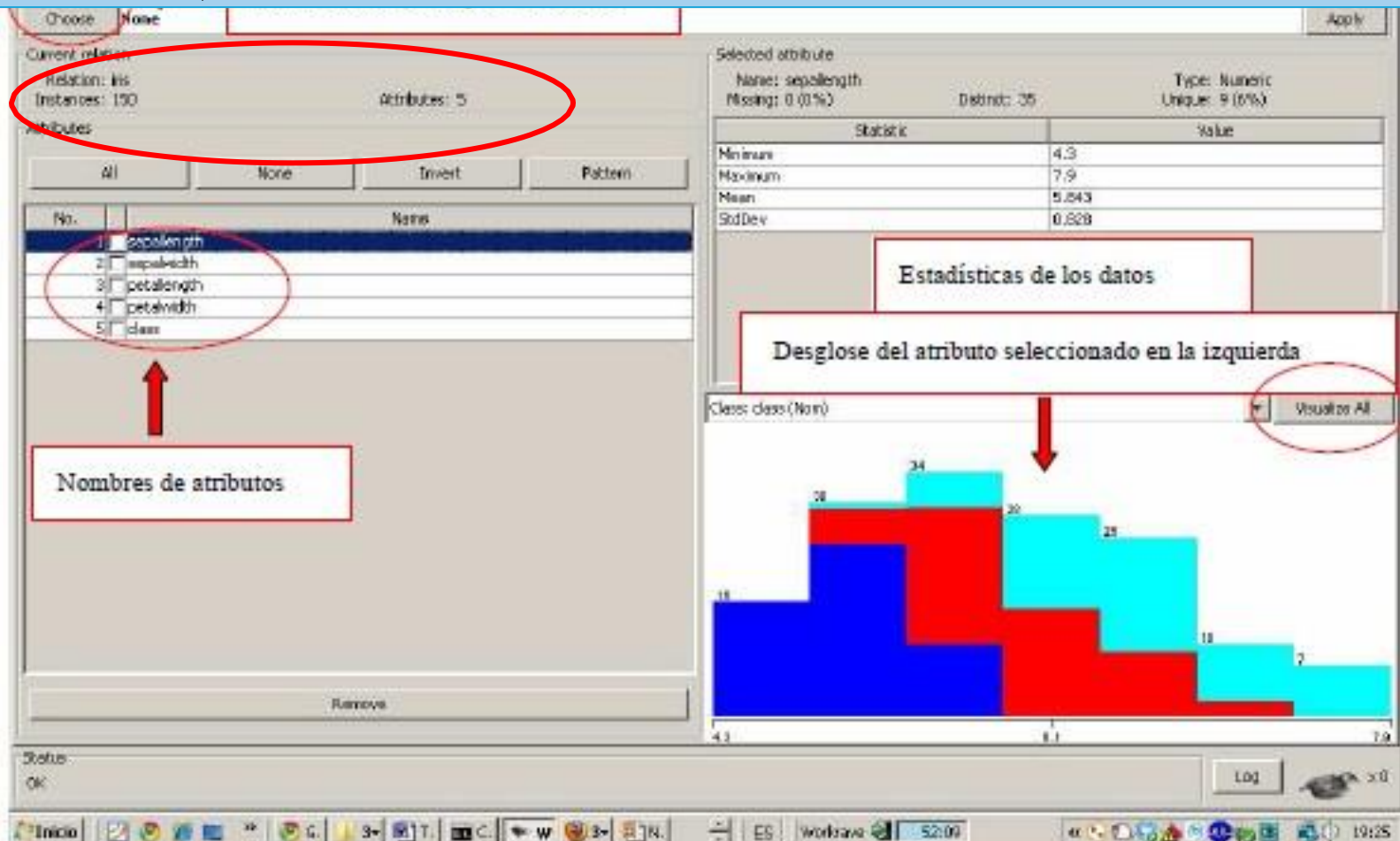
Value
4.3
7.9
5.843
0.828

34 28 25 10

6.1

Presionando sobre **Edit** aparece una representación *tabular* del archivo *iris.arff* abierto. Desplazando la barra vertical es posible ver los valores de atributos de todas las instancias

Parte izquierda: Informa nombre de la relación (iris), cantidad de instancias (150) y cantidad de atributos (5). Además, lista los nombres de los diferentes atributos.



Parte derecha: Brinda información estadística sobre los atributos. Seleccionando del lado izquierdo cada atributo, aparecen sus datos a derecha: nombre, tipo de dato, valores perdidos, valores diferentes que se repiten (distinct) y valores que aparecen una sola vez (unique).

Selección de filtros para los datos

Nombres de atributos

Estadísticas de los datos

Desglose del atributo seleccionado en la izquierda

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Visualize All

Parte derecha: Un poco mas abajo aparece el detalle estadístico: valor máximo, valor mínimo, media y desvío.

The screenshot shows the Weka 3.6.0 Explorer window. The 'Preprocess' tab is active. The 'Filter' button is circled in red, with a red arrow pointing to it from a text box labeled 'Selección de filtros para los datos'. The 'Current relation' section shows 'Relation: iris' and 'Instances: 150'. The 'Attributes' list on the left has a red circle around the first five attributes: 'sepalength', 'sepalwidth', 'petalength', 'petalwidth', and 'class'. A red arrow points from a text box labeled 'Nombres de atributos' to this list. The 'Selected attribute' section on the right shows 'Name: sepalength' with statistics: 'Minimum: 4.3', 'Maximum: 7.9', 'Mean: 5.843', and 'StdDev: 0.828'. This section is circled in red. Below it, a text box labeled 'Estadísticas de los datos' points to the statistics table. The 'Class: class (Nom)' section shows a histogram with a red arrow pointing to it from a text box labeled 'Desglose del atributo seleccionado en la izquierda'. The histogram has a 'Visualize All' button circled in red. The histogram shows four bars with counts: 18 (blue), 34 (red), 20 (red), and 2 (cyan). The x-axis is labeled with values 4.3, 5.1, and 7.9. The status bar at the bottom shows 'OK' and a 'Log' button.

Weka 3.6.0 - Explorer

Program Applications Tools Visualizations Windows Help

Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open File... Open URL... Open DB... Generate... Undo Edit... Save...

Filter Choose None

Selección de filtros para los datos

Current relation
Relation: iris
Instances: 150
Attributes: 5

Attributes

All None Invert Pattern

No.	Name
1	sepalength
2	sepalwidth
3	petalength
4	petalwidth
5	class

Nombres de atributos

Remove

Selected attribute
Name: sepalength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Estadísticas de los datos

Desglose del atributo seleccionado en la izquierda

Class: class (Nom)

Visualize All

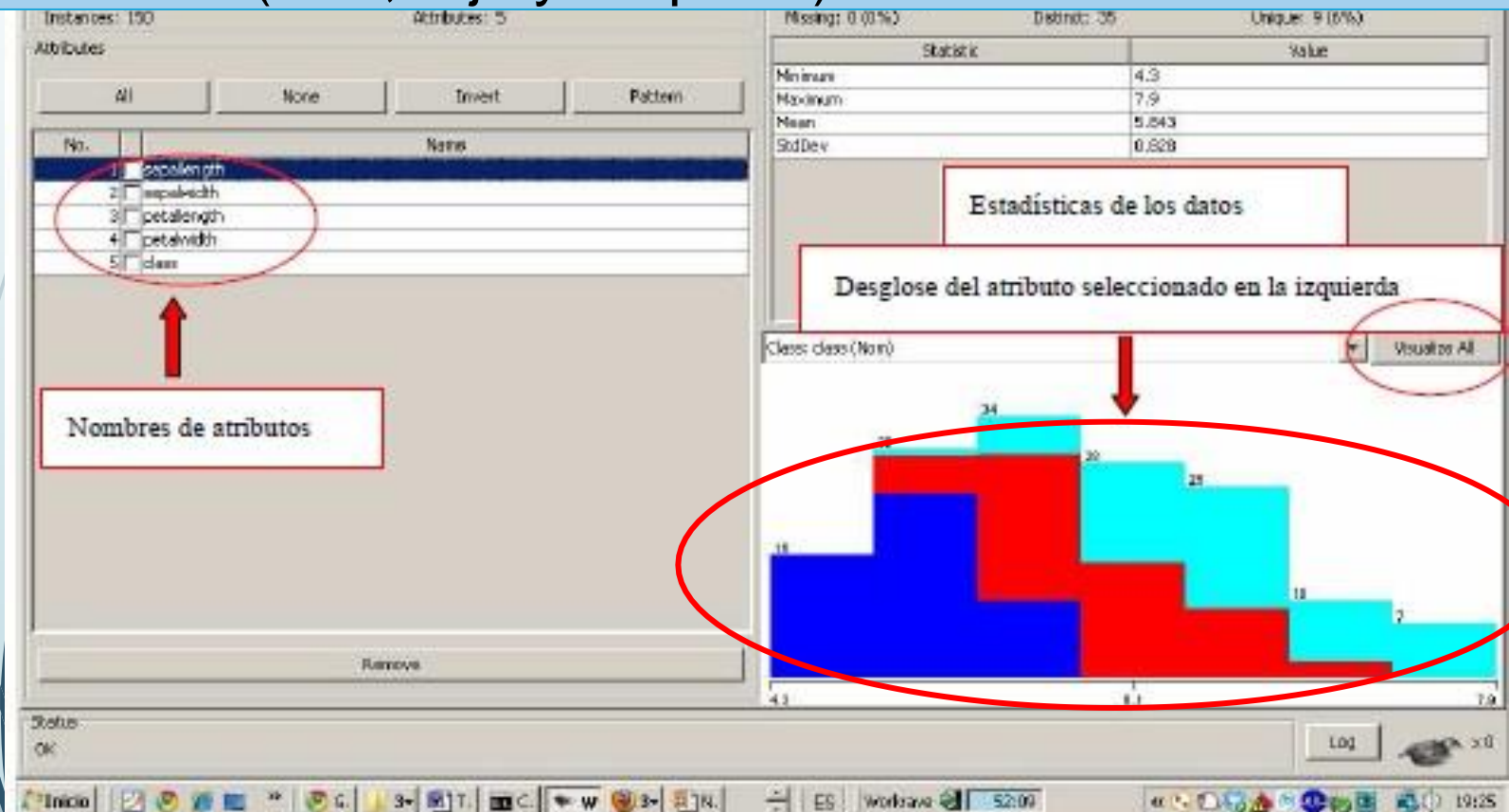
18 34 20 2

4.3 5.1 7.9

Status: OK

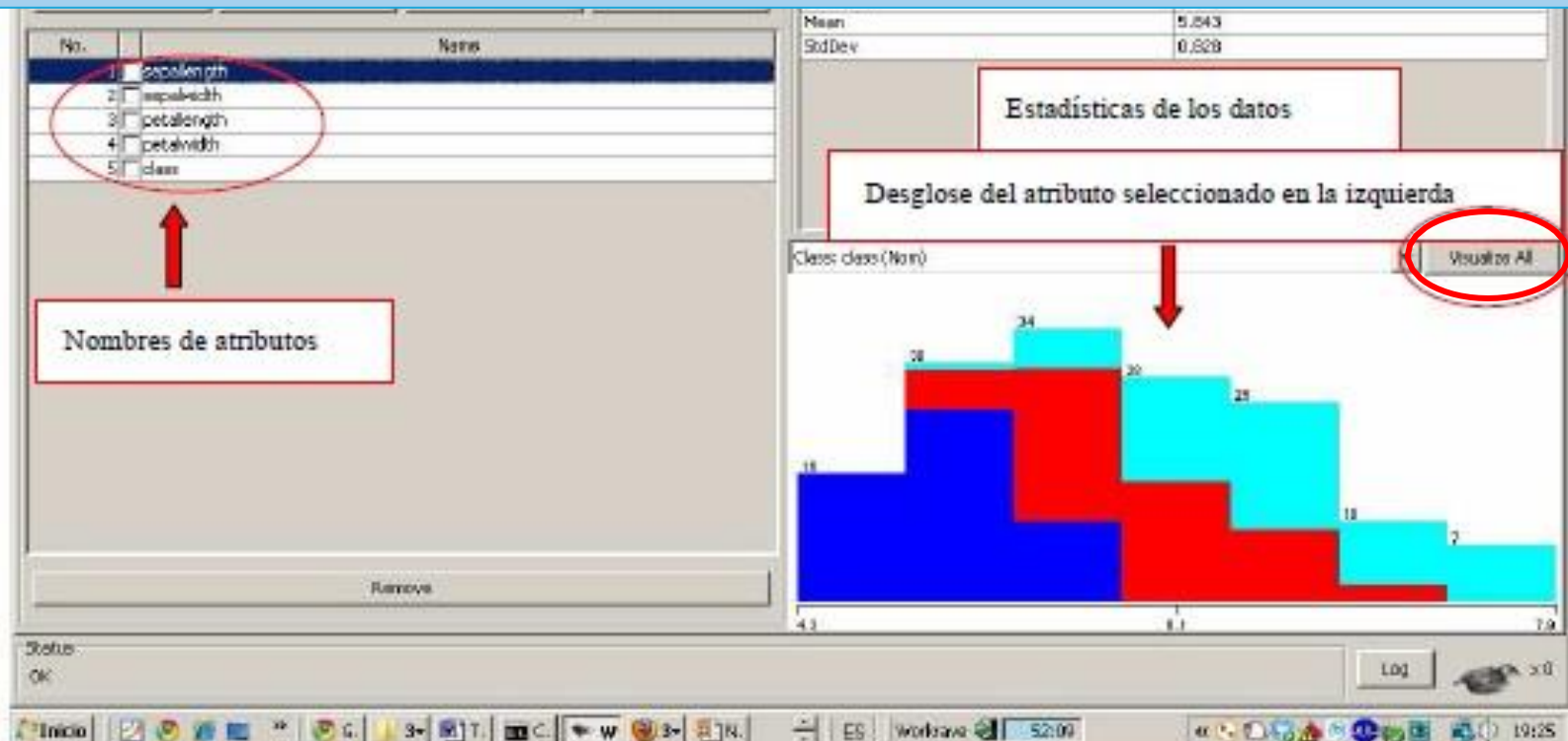
Log

Parte derecha: Finalmente, en el gráfico se ve el desglose de los valores del atributo seleccionado (*sepal.length*) en las tres clases (azul, rojo y turquesa)

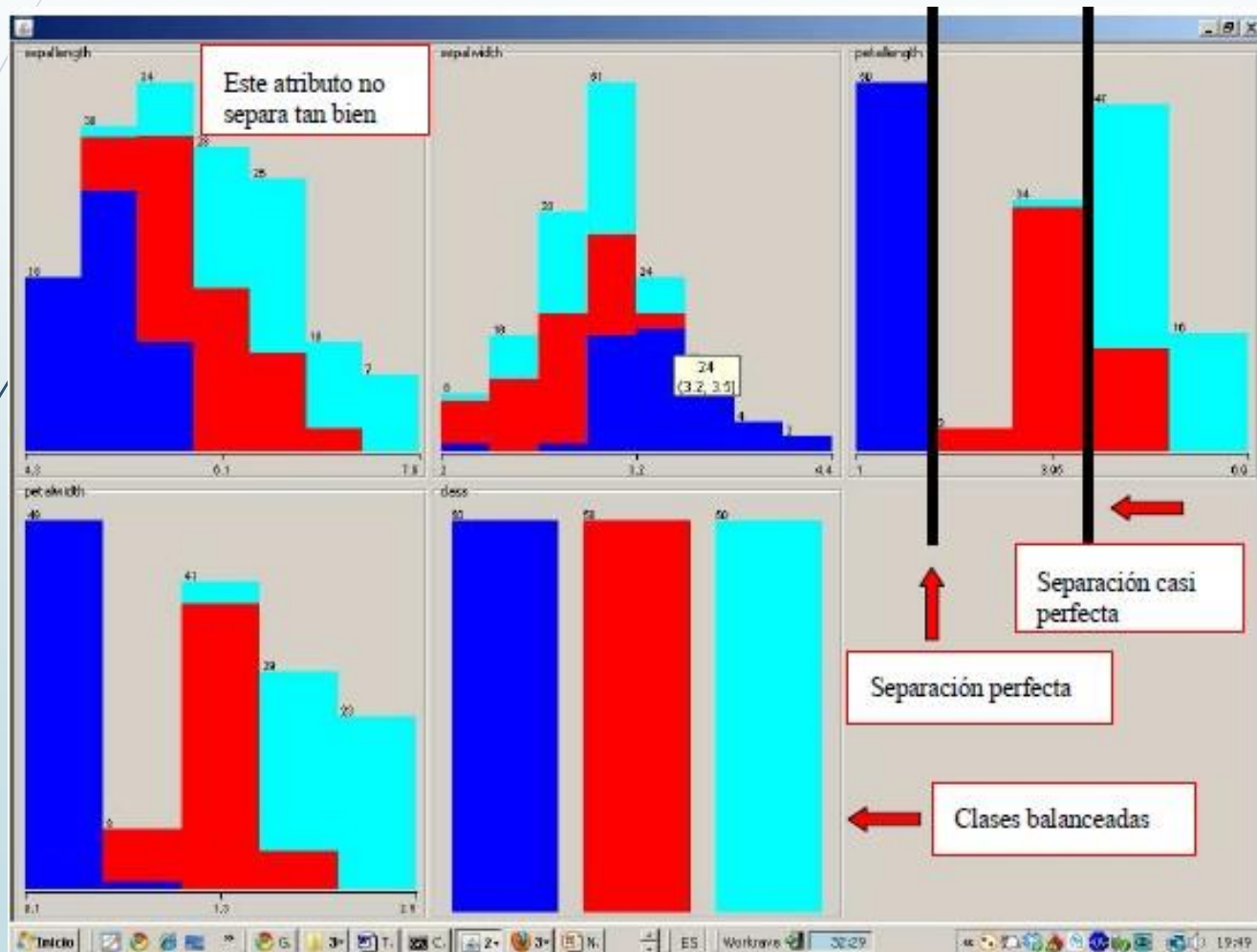


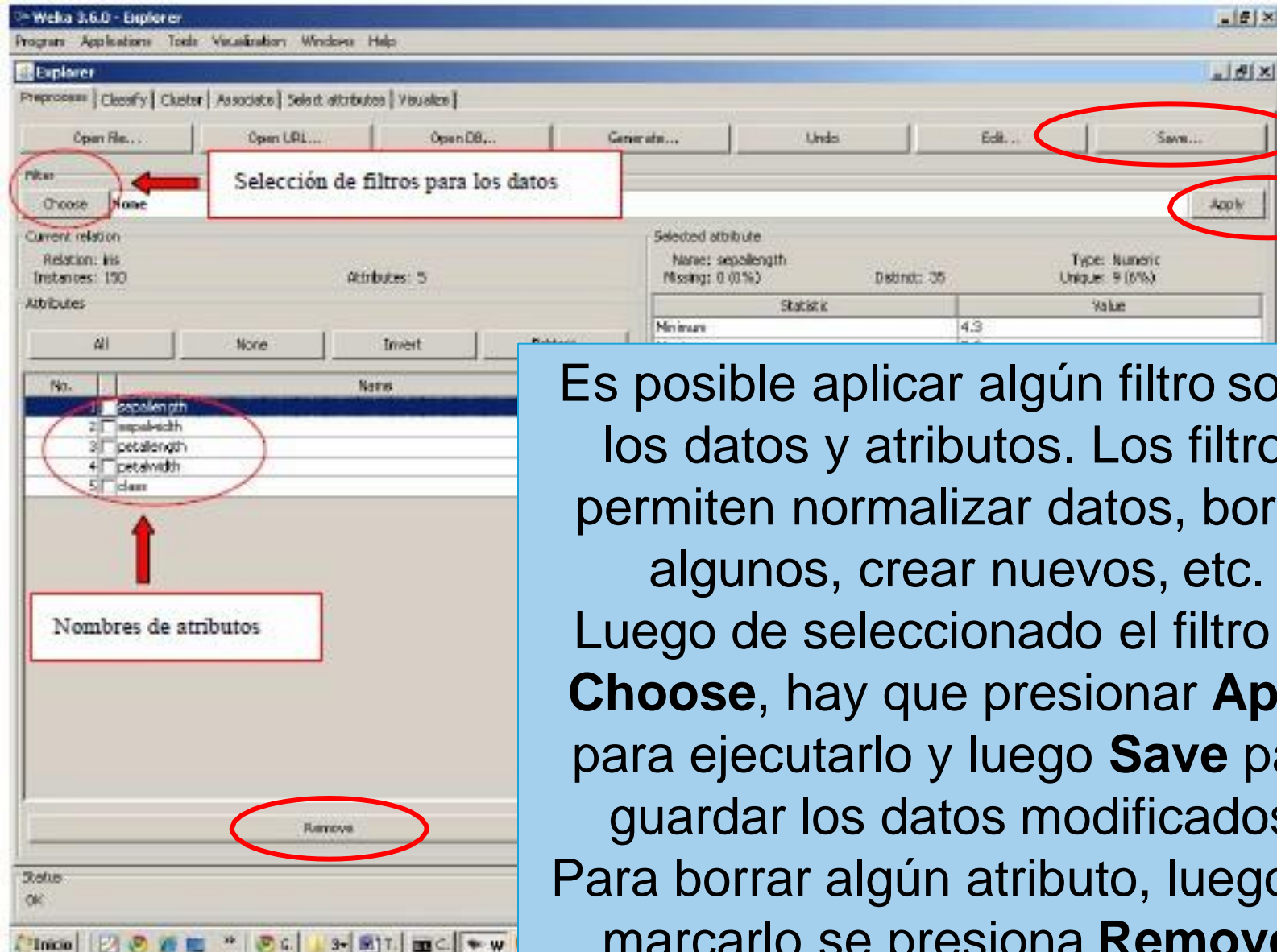


Esto puede hacerse para cada uno de los atributos de la izquierda... pero también para todos juntos. Para esto hay que seleccionar **Visualize All** (en vértice superior derecho de la gráfica)



Se puede ver que el mejor es *petalength* pues separa perfectamente clase azul de clase roja, y casi perfectamente clase roja de turquesa. Obviamente el atributo *clase* tiene separación perfecta.



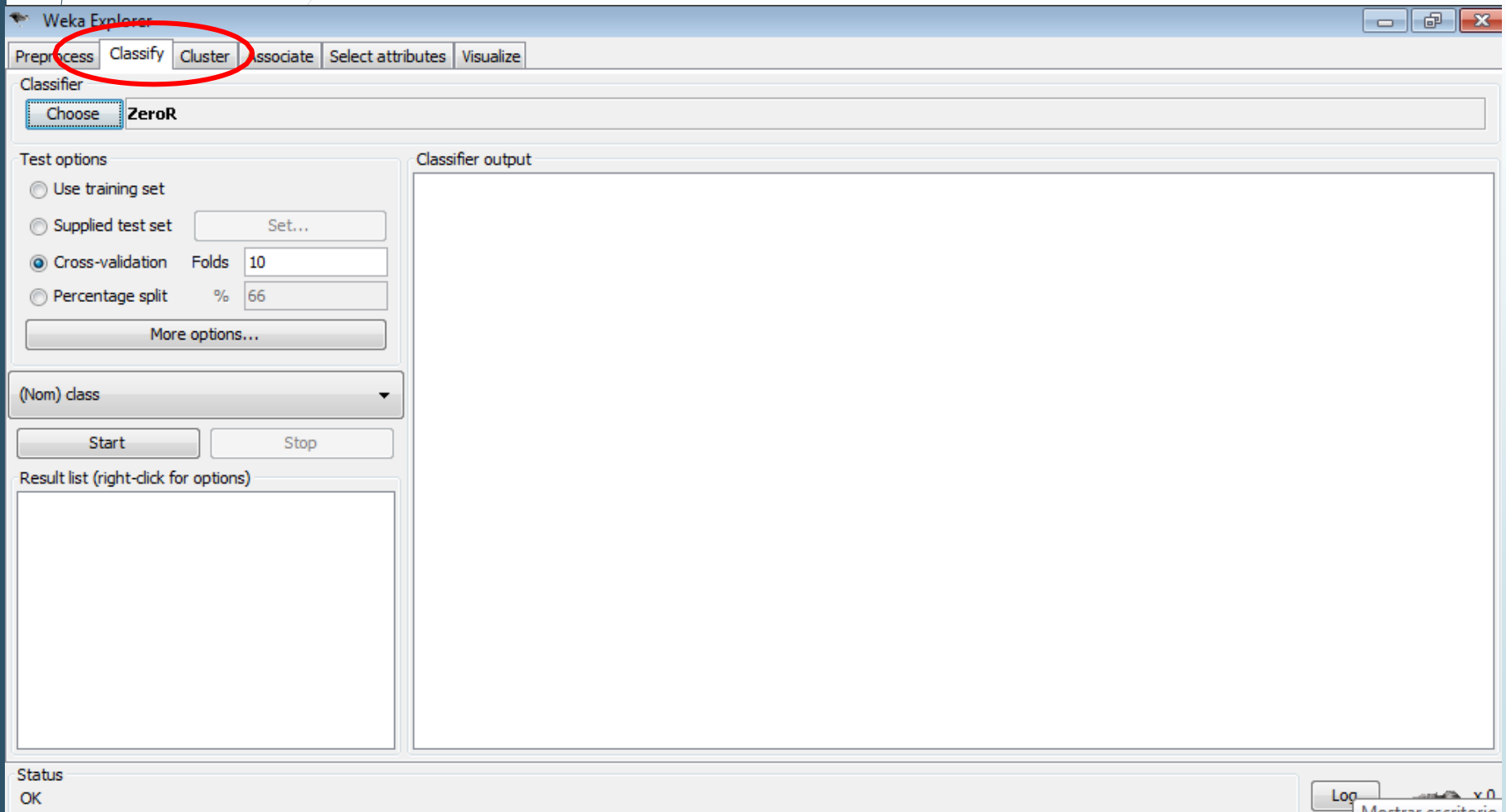


Es posible aplicar algún filtro sobre los datos y atributos. Los filtros permiten normalizar datos, borrar algunos, crear nuevos, etc. Luego de seleccionado el filtro en **Choose**, hay que presionar **Apply** para ejecutarlo y luego **Save** para guardar los datos modificados. Para borrar algún atributo, luego de marcarlo se presiona **Remove**.

Técnicas de Clasificación

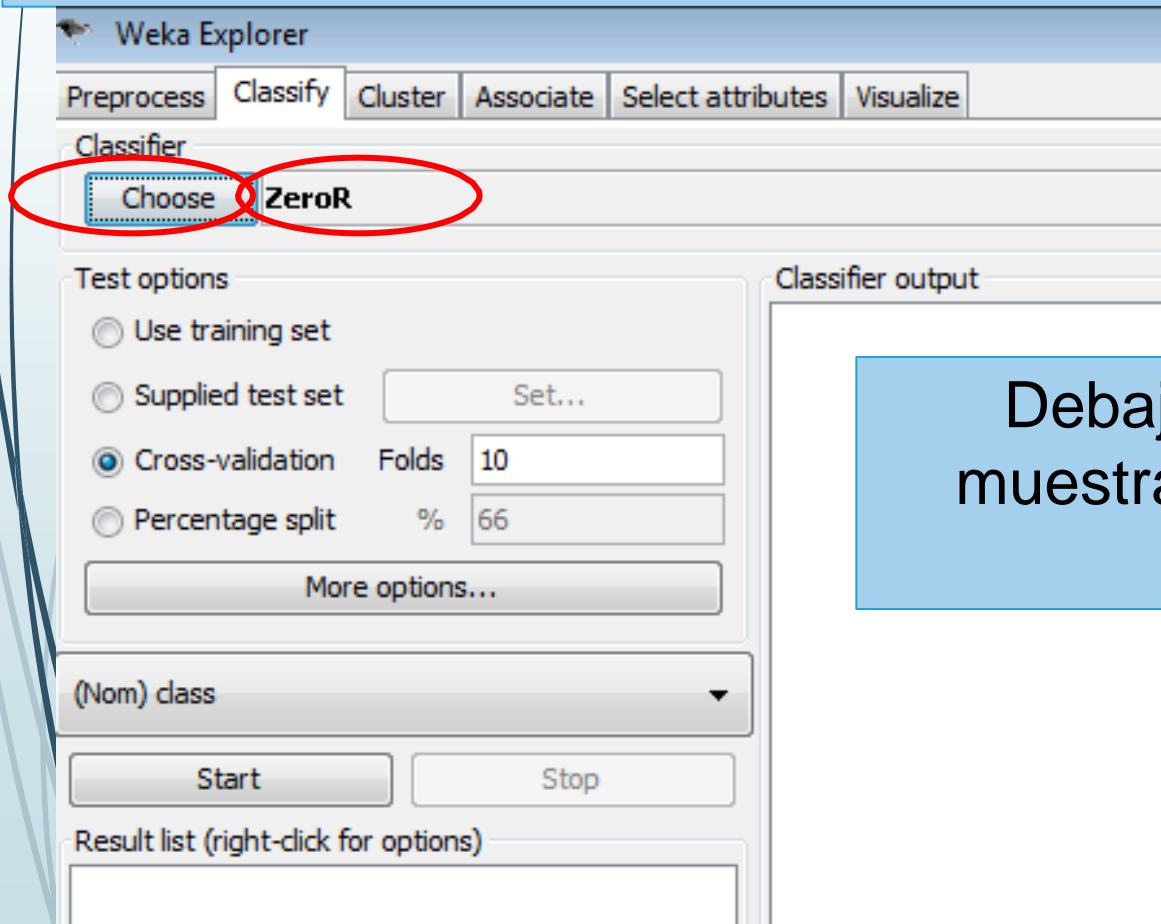


Presionando la pestaña **Classify** en la ventana del **Explorer**...

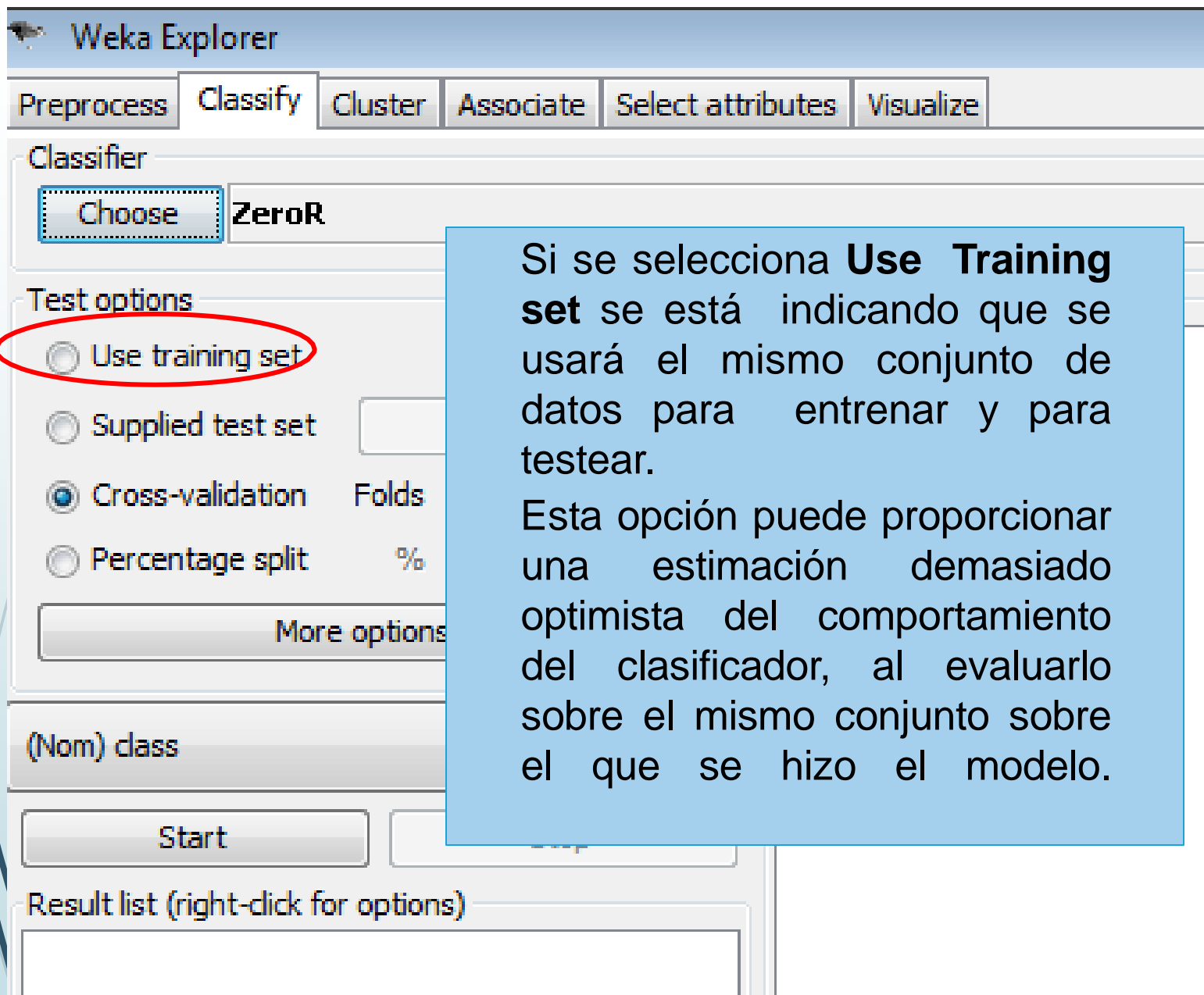


Para construir un clasificador primero hay que seleccionarlo...

Por defecto aparece **ZeroR**, pero presionando en **Choose** pueden elegirse otras opciones.



Debajo de **Choose** se muestran las opciones de testeo.



Si se selecciona **Use Training set** se está indicando que se usará el mismo conjunto de datos para entrenar y para testear.

Esta opción puede proporcionar una estimación demasiado optimista del comportamiento del clasificador, al evaluarlo sobre el mismo conjunto sobre el que se hizo el modelo.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose ZeroR

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds 10

☐ Percentage split % 66

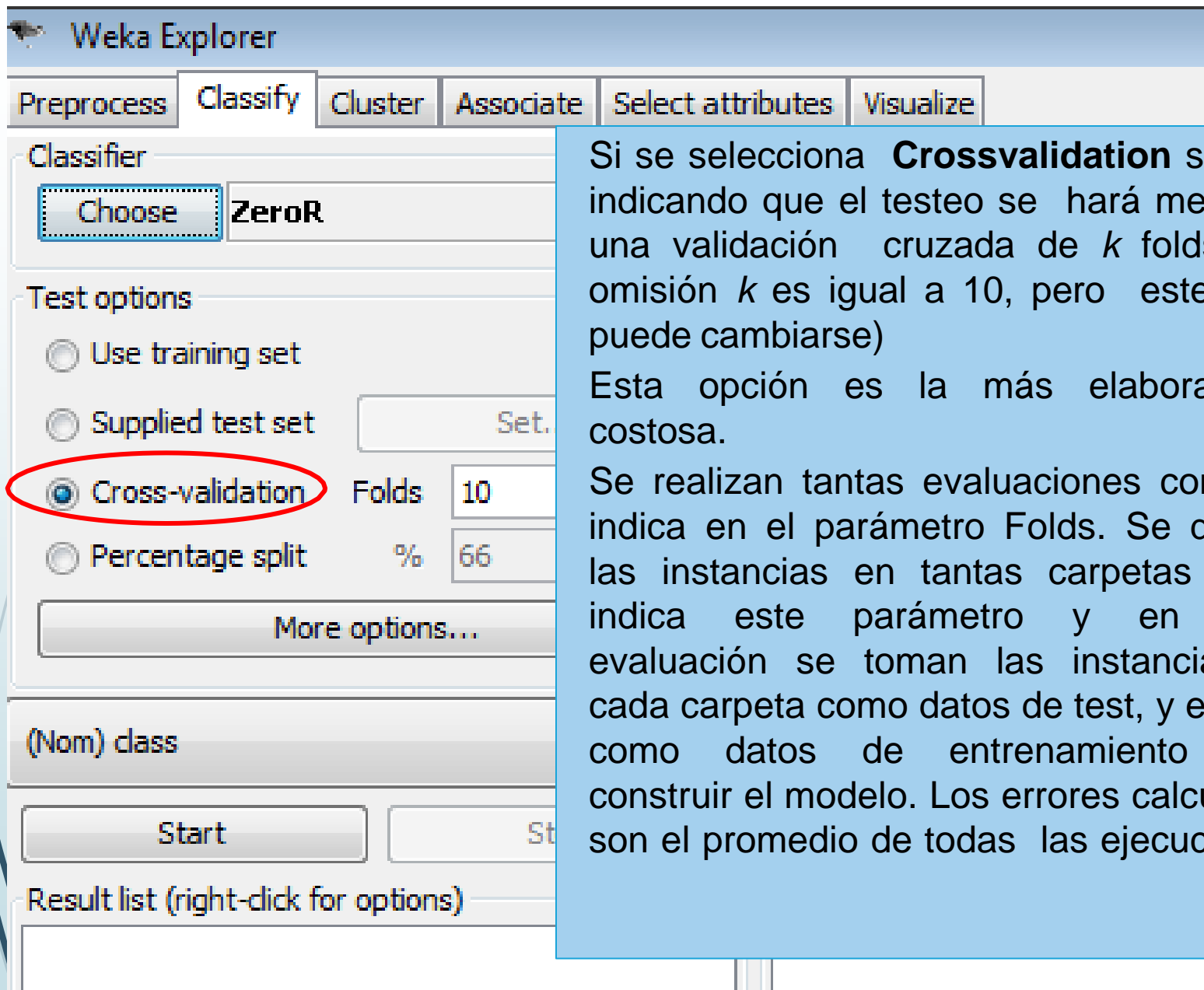
More options...

(Nom) class

Start Stop

Result list (right-click for options)

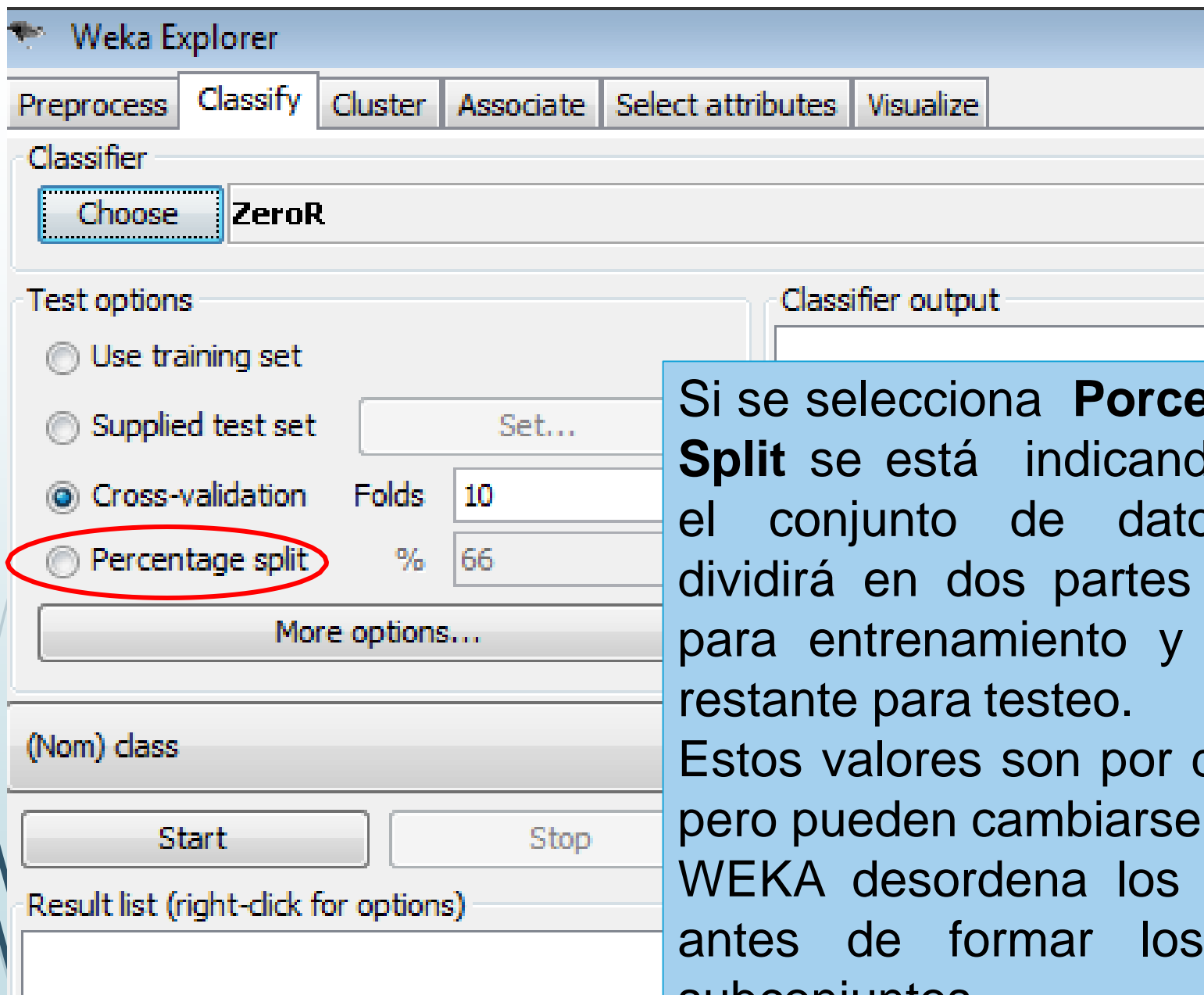
Si se selecciona **Supplied test set** se está indicando que el testeo se hará con otros datos. Debe darse allí el ingreso del archivo que los contiene.



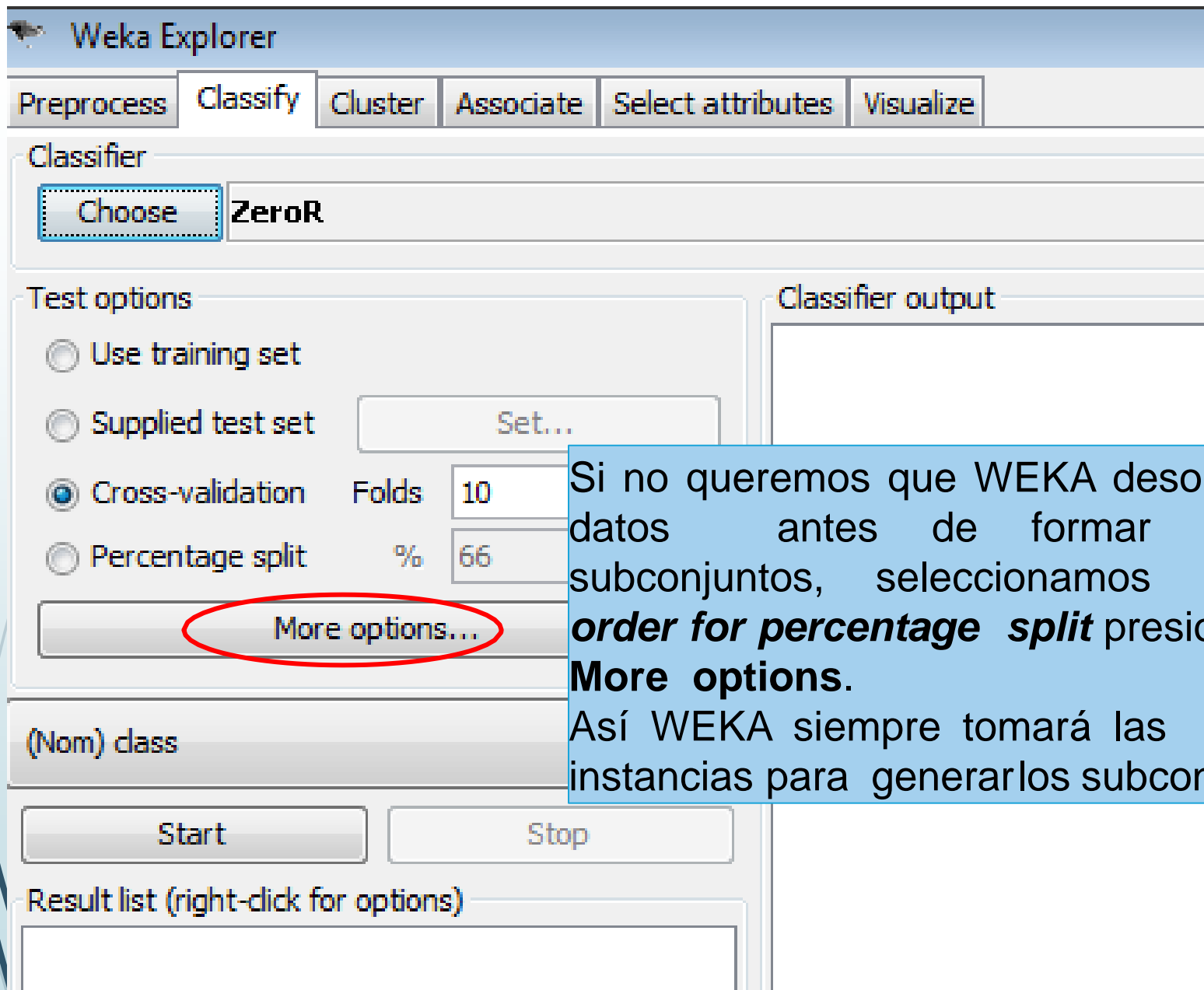
Si se selecciona **Crossvalidation** se está indicando que el testeo se hará mediante una validación cruzada de k folds (por omisión k es igual a 10, pero este valor puede cambiarse)

Esta opción es la más elaborada y costosa.

Se realizan tantas evaluaciones como se indica en el parámetro Folds. Se dividen las instancias en tantas carpetas como indica este parámetro y en cada evaluación se toman las instancias de cada carpeta como datos de test, y el resto como datos de entrenamiento para construir el modelo. Los errores calculados son el promedio de todas las ejecuciones.

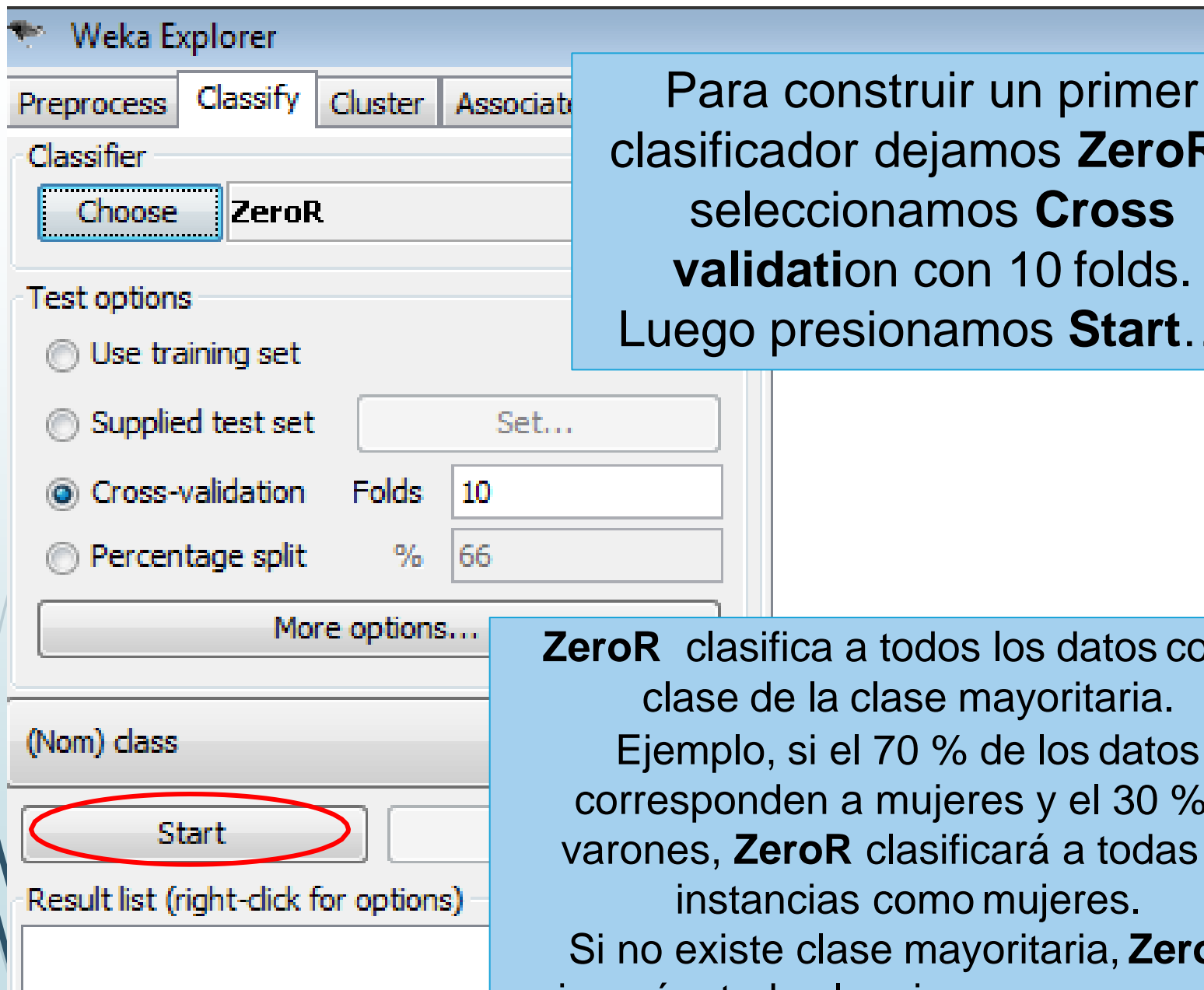


Si se selecciona **Percentage Split** se está indicando que el conjunto de datos se dividirá en dos partes 66% para entrenamiento y 33% restante para testeo. Estos valores son por defecto pero pueden cambiarse. WEKA desordena los datos antes de formar los dos subconjuntos.



Si no queremos que WEKA desordene los datos antes de formar los dos subconjuntos, seleccionamos ***Preserve order for percentage split*** presionando el **More options**.

Así WEKA siempre tomará las mismas instancias para generarlos subconjuntos.



Para construir un primer clasificador dejamos **ZeroR** y seleccionamos **Cross validation** con 10 folds. Luego presionamos **Start....**

ZeroR clasifica a todos los datos con la clase de la clase mayoritaria. Ejemplo, si el 70 % de los datos corresponden a mujeres y el 30 % a varones, **ZeroR** clasificará a todas las instancias como mujeres. Si no existe clase mayoritaria, **ZeroR** asignará a todos la primera que encuentre.

Observemos los resultados...

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **ZeroR**

Test options

- ☐ Use training set
- ☐ Supplied test set Set...
- ☒ Cross-validation Folds 10
- ☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

18:02:28 - rules.ZeroR

Classifier output

Correctly Classified Instances	50	33.3333 %
Incorrectly Classified Instances	100	66.6667 %
Kappa statistic	0	
Mean absolute error	0.4444	
Root mean squared error	0.4714	
Relative absolute error	100	%
Root relative squared error		
Total Number of Instances		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate
1	1	1
0	0	0
0	0	0
Weighted Avg.	0.333	0.333

=== Confusion Matrix ===

a	b	c	<-- classified as
50	0	0	a = Iris-setosa
50	0	0	b = Iris-versicolor
50	0	0	c = Iris-virginica

Status: OK

Se logra un 33,33 % de aciertos (instancias correctamente clasificadas). Esto es lógico dado lo expuesto antes sobre **ZeroR** (como no hay clase mayoritaria, reconoció correctamente a iris-setosa con 50 instancias, a las 100 restantes las clasificó incorrectamente en esta clase también).

Desglosado por clases, vemos una tasa de aciertos (**TP rate** = tasa de verdaderos positivos) igual a 1 para la clase *iris-setosa* y cero para las otras dos clases (*iris-versicolor* e *iris-virginica*)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose ZeroR

Test options

- ☐ Use training set
- ☐ Supplied test set Set...
- ☒ Cross-validation Folds 10
- ☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

18:02:28 - rules.ZeroR

Classifier output

Correctly Classified Instances
Incorrectly Classified Instance
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error 100 %
Root relative squared error 100 %
Total Number of Instances 150

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.333	1	0.5	0.5	Iris-setosa
	0	0	0	0	0	0.5	Iris-versicolor
	0	0	0	0	0	0.5	Iris-virginica
Weighted Avg	0.333	0.333	0.111	0.333	0.167	0.5	

=== Confusion Matrix ===

a	b	c	<-- classified as
50	0	0	a = Iris-setosa
50	0	0	b = Iris-versicolor
50	0	0	c = Iris-virginica

Status OK Log x 0

Más abajo, en la matriz de confusión, se observa que todas las instancias (150) fueron clasificadas como tipo **a** , es decir, *iris-setosa*.

Lo correcto hubiera sido que aparezca una diagonal con valor igual a 50 en cada clase

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose ZeroR

Test options:

- ☐ Use training set
- ☐ Supplied test set Set...
- ☒ Cross-validation Folds 10
- ☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

18:02:28 - rules.ZeroR

Classifier output

Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error 0.4714
Relative absolute error 100 %
Root relative squared error 100 %
Total Number of Instances 150

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.333	1	0.5	0.5	Iris-setosa
	0	0	0	0	0	0.5	Iris-versicolor
	0	0	0	0	0	0.5	Iris-virginica
Weighted Avg.	0.333	0.333	0.111	0.333	0.167	0.5	

=== Confusion Matrix ===

a	b	c	<-- classified as
50	0	0	a = Iris-setosa
50	0	0	b = Iris-versicolor
50	0	0	c = Iris-virginica

Status: OK

Log x 0

Probemos ahora a superar el 33.33 % de aciertos de **ZeroR** seleccionando en **Choose** otro clasificador.

Elegimos ahora **Part**...

The screenshot shows the Weka Explorer interface. The 'Classifier' tab is active, and the 'rules' folder is expanded in the left pane. The 'PART' classifier is circled in red. The right pane displays the performance metrics for the selected classifier (ZeroR).

Classifier

- weka
 - classifiers
 - bayes
 - functions
 - lazy
 - meta
 - mi
 - misc
 - rules
 - ConjunctiveRule
 - DecisionTable
 - DTNB
 - JRip
 - M5Rules
 - NNge
 - OneR
 - PART**
 - Prism
 - Ridor
 - ZeroR
 - trees

Classified Instances 50
Classified Instances 100
Classification Error 0.4444
Classification Squared Error 0.4714
Classification Absolute Error 100 %
Classification Squared Error 100 %
Number of Instances 150

Accuracy By Class ==

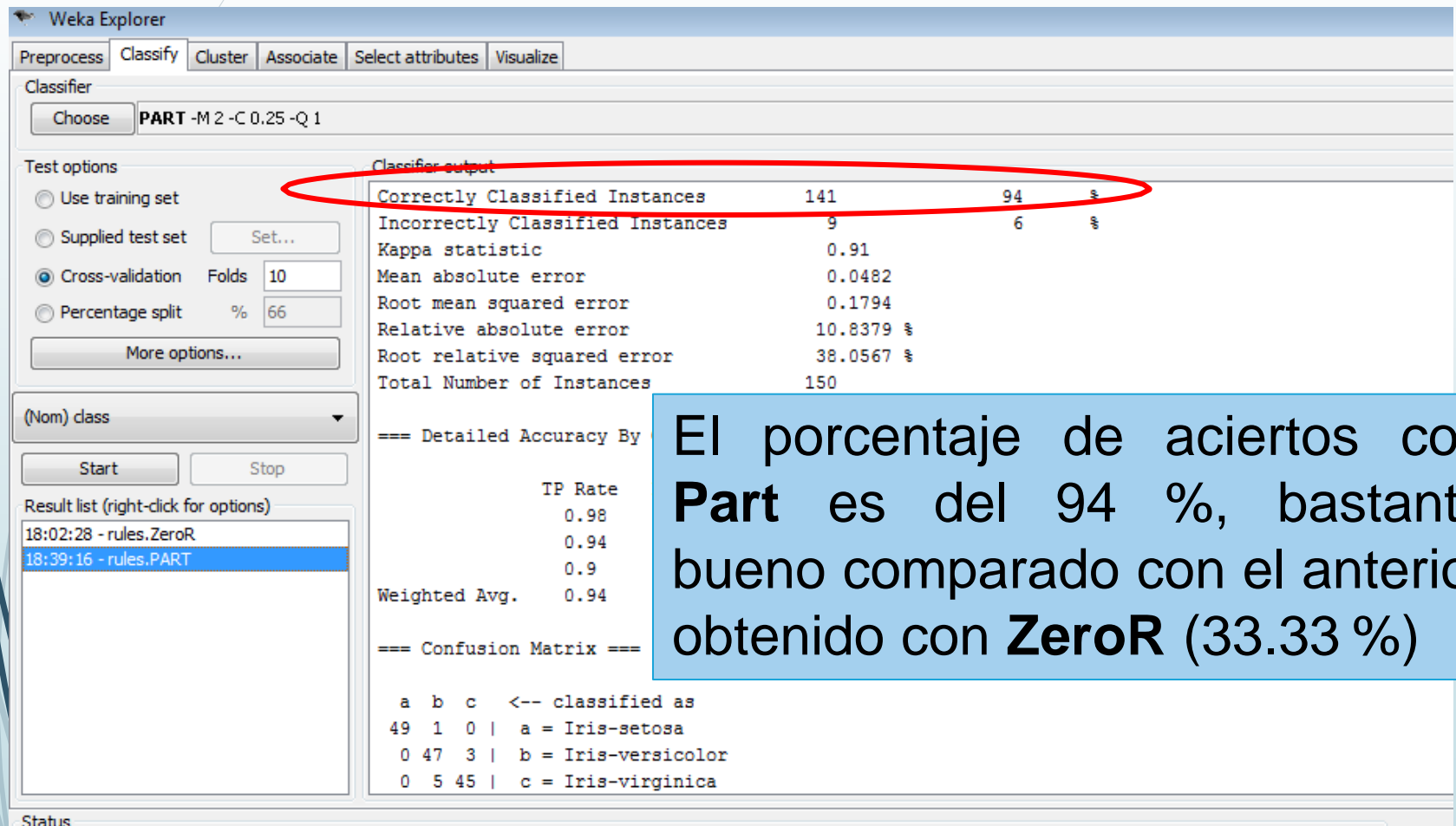
TP Rate	FP Rate	Precision	Recall
1	1	0.333	1
0	0	0	0
0	0	0	0
0.333	0.333	0.111	0.333

Confusion Matrix ==

<-- classified as
a = Iris-setosa
b = Iris-versicolor
c = Iris-virginica

Status: OK

Una vez seleccionado Part presionamos en Start y luego analizamos los resultados....



The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'PART -M 2 -C 0.25 -Q 1'. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' section is circled in red, showing the following results:

Classifier output			
Correctly Classified Instances	141	94	%
Incorrectly Classified Instances	9	6	%
Kappa statistic	0.91		
Mean absolute error	0.0482		
Root mean squared error	0.1794		
Relative absolute error	10.8379	%	
Root relative squared error	38.0567	%	
Total Number of Instances	150		

Below the 'Classifier output' section, the 'Detailed Accuracy By' table shows the TP Rate for each class:

TP Rate
0.98
0.94
0.9
Weighted Avg. 0.94

The 'Confusion Matrix' section shows the following data:

a	b	c	<-- classified as
49	1	0	a = Iris-setosa
0	47	3	b = Iris-versicolor
0	5	45	c = Iris-virginica

The 'Result list' on the left shows two entries: '18:02:28 - rules.ZeroR' and '18:39:16 - rules.PART', with the latter selected.

El porcentaje de aciertos con **Part** es del 94 %, bastante bueno comparado con el anterior obtenido con **ZeroR** (33.33 %)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose PART -M 2 -C 0.25 -Q 1

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

18:02:28 -rules.ZeroR

18:39:16 -rules.PART

Classifier output

Correctly Classified Instances 141 94 %

Incorrectly Classified Instances 9 6 %

Kappa statistic 0.91

Mean absolute error 0.0482

Root mean squared error 0.1794

Relative absolute error 10.8379 %

Root relative squared error 38.0567 %

Total Number of Instances 150

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.98	0	1	0.98	0.99	0.99	Iris-setosa
0.94	0.06	0.887	0.94	0.913	0.954	Iris-versicolor
0.9	0.03	0.938	0.9	0.918	0.959	Iris-virginica
Weighted Avg.	0.94	0.03	0.941	0.94	0.968	

=== Confusion Matrix ===

a b c <-- classified as

49 1 0 | a = Iris-setosa

0 47 3 | b = Iris-versicolor

La tasa de aciertos por clase indica un 98 % para iris-setosa, 94 % para iris-versicolor y 90 % para iris-virgínica.

Estos resultados están mejor que los anteriormente obtenidos. Además observamos que están equilibrados respecto a las tres clases, lo cual se debe a que las instancias de datos por clase también están balanceadas (50 por clase).

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **PART -M 2 -C 0.25 -Q 1**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

18:02:28 - rules.ZeroR

18:39:16 - rules.PART

Classifier output

Correctly Classified Instances	141	94	%
Incorrectly Classified Instances	9	6	%
Kappa statistic	0.91		
Mean absolute error			
Root mean squared error			
Relative absolute error			
Root relative squared error			
Total Number of Instances			

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate
	0.98	0
	0.94	0.06
	0.9	0.03
Weighted Avg.	0.94	0.03

=== Confusion Matrix ===

a	b	c	<-- classified as
49	1	0	a = Iris-setosa
0	47	3	b = Iris-versicolor
0	5	45	c = Iris-virginica

Status

Respecto a la **matriz de confusión**, vemos que se clasificaron correctamente 49 instancias de iris-setosa, 47 de iris-versicolor y 45 de iris-virgínica. 1 instancia de setosa se clasificó incorrectamente como versicolor, 3 instancias de versicolor como virgínicas, y 5 instancias virgínicas como versicolor.

Ahora nos centraremos en el modelo obtenido por el clasificador **Part**....

Se obtienen tres reglas que permiten clasificar a las nuevas instancias en alguna de las tres clases. Cabe resaltar que las reglas solo invocan el valor de los atributos *petalwidth* y *petalength* (considerados los mejores unas diapos atrás)

The screenshot shows the Weka Explorer interface. The 'Classify' tab is active, and the 'PART' classifier is selected with parameters '-M 2 -C 0.25 -Q 1'. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Result list' on the left shows two entries: '18:02:28 - rules.ZeroR' and '18:39:16 - rules.PART', with the latter selected. The 'Classifier output' pane on the right displays the 'PART decision list' which contains three rules for classifying Iris species based on petal width and length. A red oval is drawn around the decision list rules.

```
petalwidth <= 0.6: Iris-setosa (50.0)

petalwidth <= 1.7 AND
petallength <= 4.9: Iris-versicolor (48.0/1.0)

: Iris-virginica (52.0/3.0)

Number of Rules :      3

Time taken to build model: 0.14 seconds

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

Intentemos con un tercer clasificador, seleccionemos en **Choose** opción **trees**, el algoritmo **J48**...
Presionemos **Start** y analicemos los resultados obtenidos...

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

18:02:28 - rules.ZeroR

18:39:16 - rules.PART

19:09:20 - trees.J48

Classifier output

Correctly Classified Instances	144	96	%
Incorrectly Classified Instances	6	4	%
Kappa statistic	0.94		
Mean absolute error	0.035		
Root mean squared error	0.1586		
Relative absolute error	7.8705	%	
Root relative squared error	33.6353	%	
Total Number of Instances	150		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.98	0	1	0.98	0.99	0.99	Iris-setosa
	0.94	0.03	0.94	0.94	0.94	0.952	Iris-versicolor
	0.96	0.03	0.941	0.96	0.95	0.961	Iris-virginica
Weighted Avg.	0.96	0.02	0.96	0.96	0.96	0.968	

=== Confusion Matrix ===

```
a b c <-- classified as
49 1 0 | a = Iris-setosa
0 47 3 | b = Iris-versicolor
0 2 48 | c = Iris-virginica
```

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

18:02:28 - rules.ZeroR

18:39:16 - rules.PART

19:09:20 - trees.J48

Classifier output

Correctly Classified Instances 144 96 %

Incorrectly Classified Instances 6 4 %

Kappa statistic 0.94

Mean absolute error 0.035

Root mean squared error 0.1586

Relative absolute error

Root relative squared error

Total Number of Instances

=== Detailed Accuracy By Class

	TP Rate	FP Rate					
	0.98	0					
	0.94	0.03	0.94	0.94	0.94	0.952	Iris-versicolor
	0.96	0.03	0.941	0.96	0.95	0.961	Iris-virginica
Weighted Avg.	0.96	0.02	0.96	0.96	0.96	0.968	

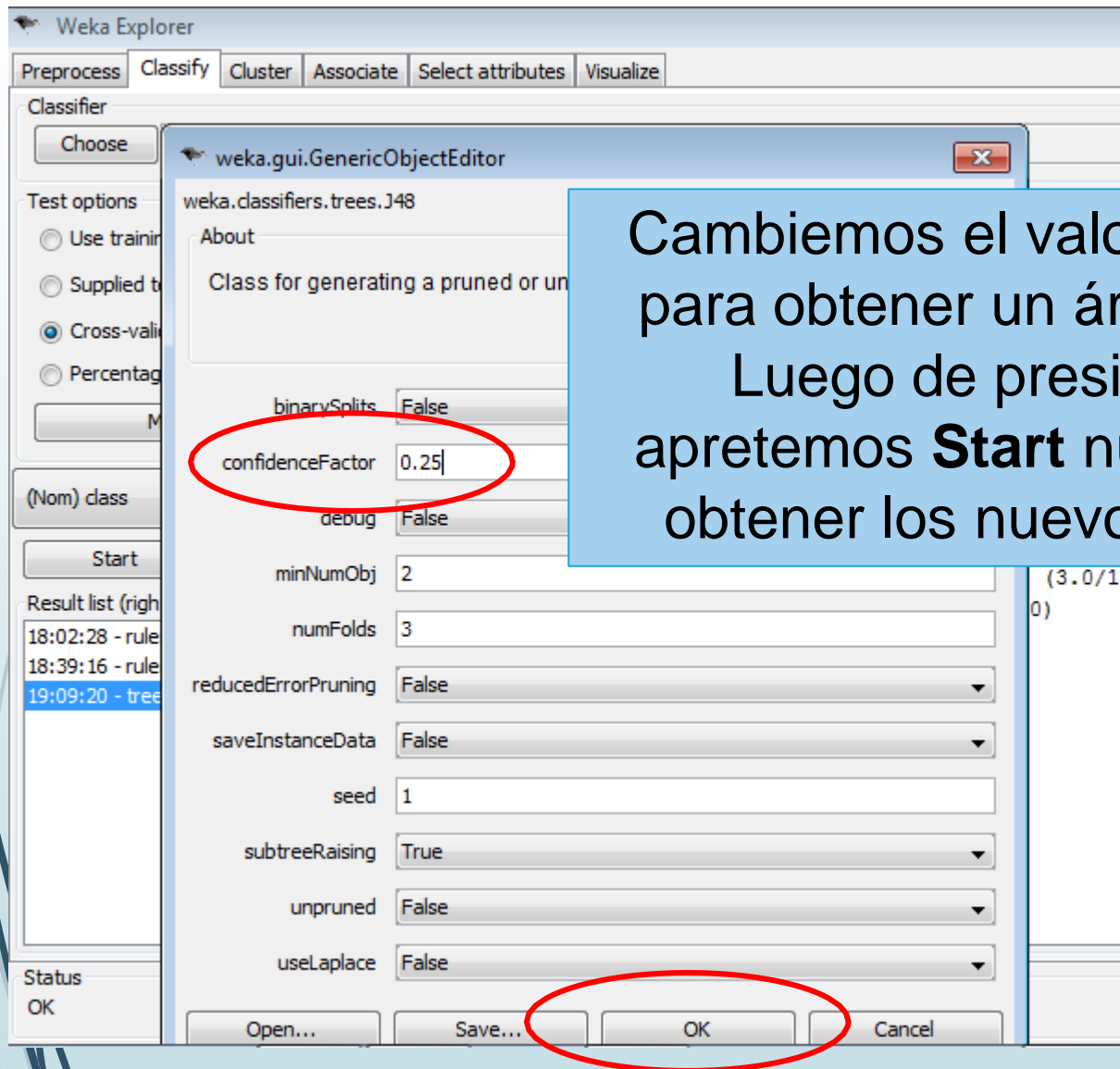
=== Confusion Matrix ===

a	b	c	-- classified as
49	1	0	a = Iris-setosa
0	47	3	b = Iris-versicolor
0	2	48	c = Iris-virginica

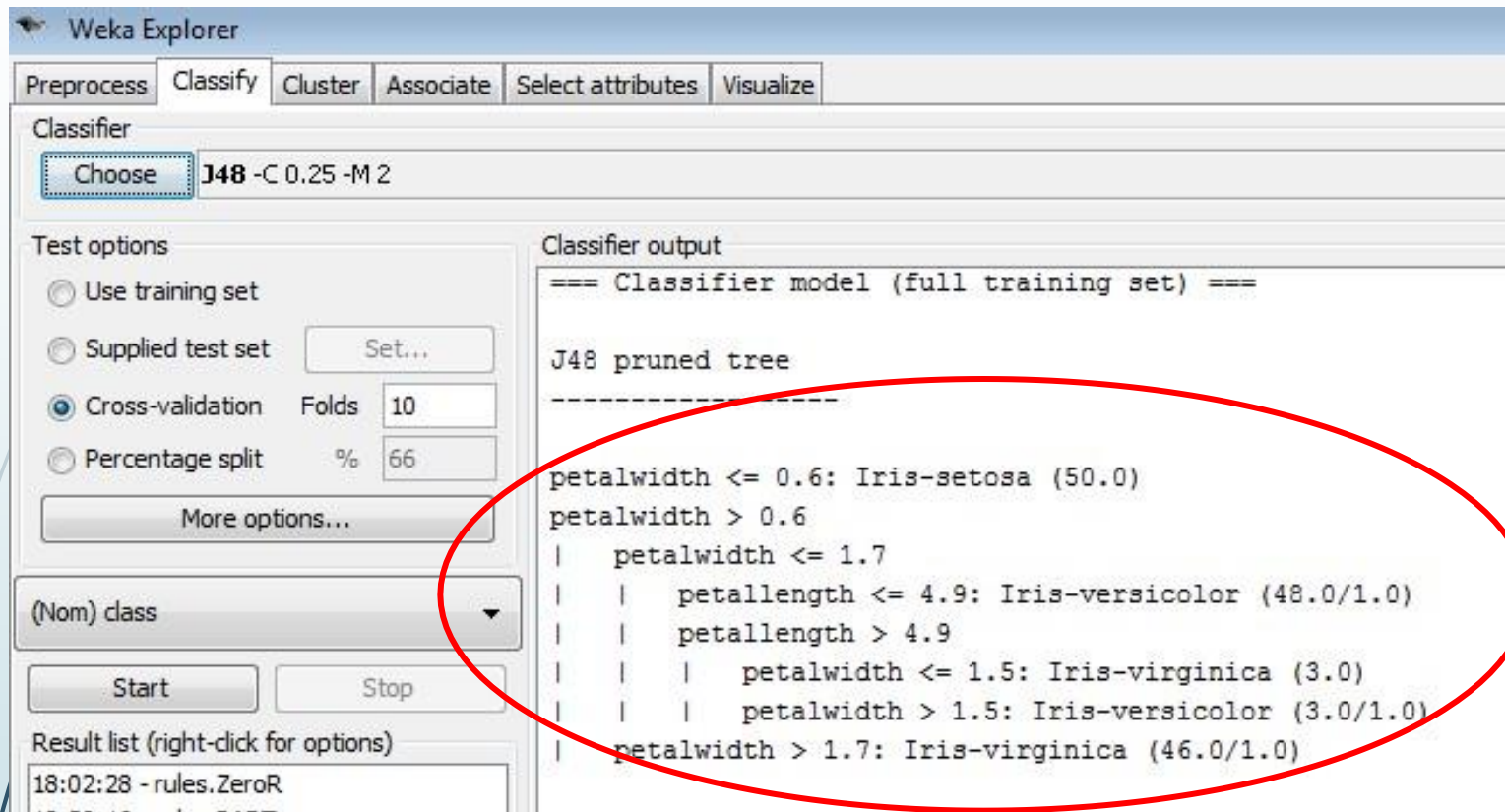
Con J48 el % de aciertos sube a 96

La tasa de acierto se mantiene para las iris-setosa (98%) e iris-versicolor (94%), pero mejora a 96% para iris-virginica.

Los cambios en la matriz de confusión se dan solo en la tercera clase (respaldando el aumento en la tasa de acierto).



El modelo obtenido por el clasificador **J48** es un árbol de 5 niveles y 9 nodos...



No se puede controlar el número de nodos, pero cuanto más pequeño es su factor de confianza más simple resulta el clasificador. Este parámetro varía entre 0 y 1, por defecto se iguala a 0.25 pero puede ajustarse clickeando sobre **J48**...

Con el factor de confianza en 0.001 el árbol resultante tiene 4 niveles y 7 nodos, pero su % de instancias correctamente clasificadas descendió a 94.

Test options

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds
- ☐ Percentage split %
-

(Nom) class

Result list (right-click for options)

- 18:02:28 - rules.ZeroR
- 18:39:16 - rules.PART
- 19:09:20 - trees.J48
- 19:47:57 - trees.J48

Classifier output

J48 pruned tree

```
-----
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petallength <= 4.9: Iris-versicolor (48.0/1.0)
| | petallength > 4.9: Iris-virginica (6.0/2.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)
```

Number of Leaves : 4

Size of the tree : 7

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	141	94	%
Incorrectly Classified Instances	9	6	%
Kappa statistic	0.91		

Generalmente, un árbol más pequeño es más fácil de entender pero es menos preciso en la clasificación.

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **J48 -C 0.25 -M 2**

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds

☐ Percentage split %

(Nom) class

Result list (right-click for options)

- 18:02:28 - rules.ZeroR
- 18:39:16 - rules.PART
- 19:09:20 - trees.J48**

Classifier output

Correctly Classified Instances	144	96	%
Incorrectly Classified Instances	6	4	%
Kappa statistic	0.94		
Mean absolute error	0.035		
Root mean squared error	0.1586		
Relative absolute error	7.8705 %		
Root relative squared error	33.6353 %		
Total Number of Instances	150		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.98	0	1	0.98	0.99	0.99	Iris-setosa
	0.94	0.03	0.94	0.94	0.94	0.952	Iris-versicolor
	0.96	0.03	0.941	0.96	0.95	0.961	Iris-virginica
Weighted Avg.	0.96	0.02	0.96	0.96	0.96	0.968	

=== Co

a b

49 1

0 47

0 2

Un detalle interesante de WEKA...
podemos retornar a los resultados
obtenidos en cualquiera de las
ejecuciones anteriores solo
seleccionándola...

Volvamos a los resultados obtenidos con **Part** para intentar mejorarlos....

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **PART -M 2 -C 0.25 -Q 1**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds

☐ Percentage split %

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 18:02:28 - rules.ZeroR
- 18:39:16 - rules.PART

Classifier output

Correctly Classified Instances 141 94 %

Incorrectly Classified Instances 9 6 %

Kappa statistic 0.91

Mean absolute error 0.0482

Root mean squared error 0.1794

Relative absolute error 10.8379 %

Root relative squared error 38.0567 %

Total Number of Instances 150

=== Detailed Accuracy By Class ===

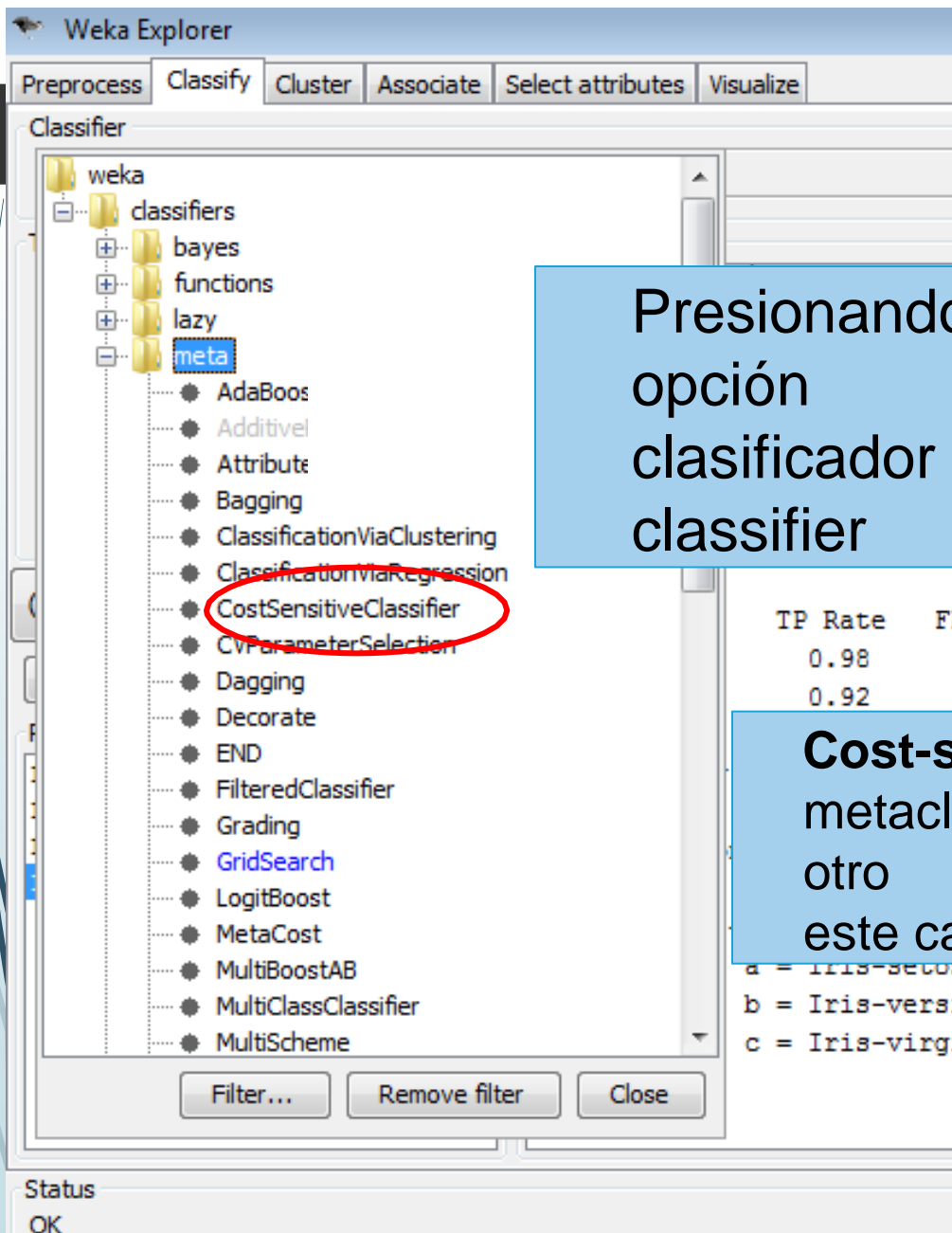
TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.98	0	1	0.98	0.99	0.99	Iris-setosa
0.94	0.06	0.887	0.94	0.913	0.954	Iris-versicolor
0.9	0.03	0.938	0.9	0.918	0.959	Iris-virginica
Weighted Avg.	0.94	0.941	0.94	0.94	0.968	

=== Confusion Matrix ===

a	b	c	<-- classified as
49	1	0	a = Iris-setosa
0	47	3	b = Iris-versicolor
0	5	45	c = Iris-virginica

Status

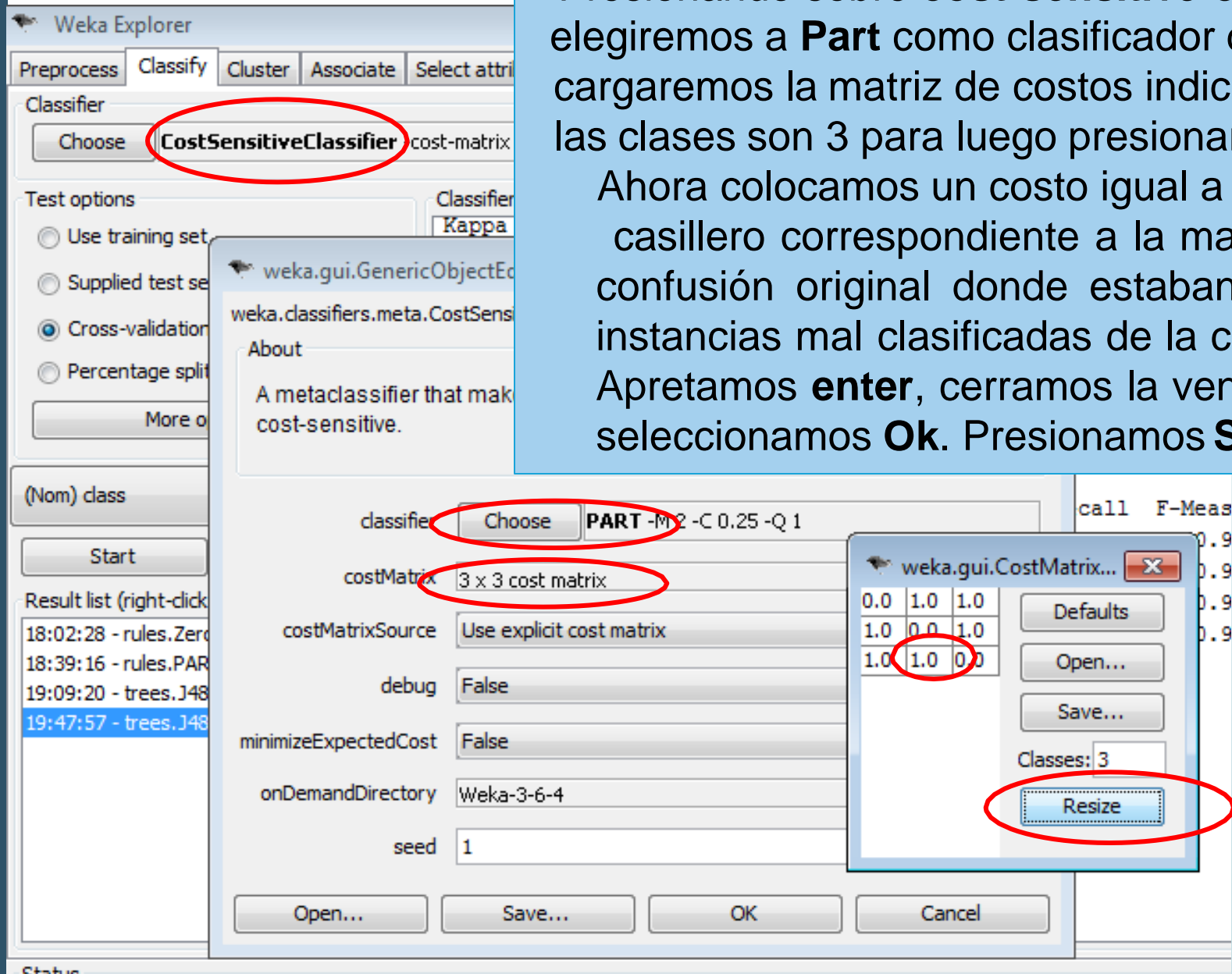
Intentaremos mejorar la clasificación de las instancias de *iris-virgínica*



Presionando sobre Choose, opción meta, elegimos el clasificador cost-sensitive-classifier

Cost-sensitive-classifier es un metaclassificador porque utiliza otro clasificador de base (en este caso será Part).

Antes debemos realizar algunos ajustes....
Presionando sobre **cost-sensitive-classifier** elegiremos a **Part** como clasificador de base y cargaremos la matriz de costos indicando que las clases son 3 para luego presionar **Resize**.
Ahora colocamos un costo igual a 2 en el casillero correspondiente a la matriz de confusión original donde estaban las 5 instancias mal clasificadas de la clase 3. Apretamos **enter**, cerramos la ventana y seleccionamos **Ok**. Presionamos **Start**...



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **CostSensitiveClassifier** -cost-matrix "[0.0 1.0 1.0; 1.0 0.0 1.0]"

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

18:02:28 - rules.ZeroR

18:39:16 - rules.PART

19:09:20 - trees.J48

19:47:57 - trees.J48

20:23:29 - meta.CostSensitiveClassifier

Classifier output

Incorrectly Classified Instances: 10
Kappa statistic: 0.98
Mean absolute error: 0.02
Root mean squared error: 0.141
Relative absolute error: 6.309%
Root relative squared error: 1.054
Total Number of Instances: 150

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure
a = Iris-setosa	0.98	0.00	1.000	0.980	0.989
b = Iris-versicol	0.88	0.03	0.936	0.880	0.908
c = Iris-virginica	0.96	0.06	0.889	0.960	0.924
Weighted Avg.	0.94	0.03	0.942	0.940	0.941

=== Confusion Matrix ===

	a	b	c	<- class
a	49	1	0	a = Iris-setosa
b	0	44	6	b = Iris-versicol
c	0	2	48	c = Iris-virginica

Observamos los % de acierto por clase cambiaron al utilizar este metaclassificador...

Iris-setosa se mantiene en 98% pero *Iris-versicola* bajó de 94% a 88%, e *iris-virgínica* subió de 90% a 96%.

Esto es congruente con los cambios en la matriz de confusión. Mientras los valores para *iris-setosa* se mantienen, las instancias mal clasificadas de *iris-versicola* aumentaron así como disminuyeron las de *iris-virgínica*.

Un archivo *.arff* internamente...

```
1 @relation weather
2
3 @attribute outlook {sunny, overcast, rainy}
4 @attribute temperature real
5 @attribute humidity real
6 @attribute windy {TRUE, FALSE}
7 @attribute play {yes, no}
8
9 @data
10 sunny,85,85,FALSE,no
11 sunny,80,90,TRUE,no
12 overcast,83,86,FALSE,yes
13 rainy,70,96,FALSE,yes
14 rainy,68,80,FALSE,yes
15 rainy,65,70,TRUE,no
16 overcast,64,65,TRUE,yes
17 sunny,72,95,FALSE,no
18 sunny,69,70,FALSE,yes
19 rainy,75,80,FALSE,yes
20 sunny,75,70,TRUE,yes
21 overcast,72,90,TRUE,yes
22 overcast,81,75,FALSE,yes
23 rainy,71,91,TRUE,no
```

ARFF (Attribute-Relation File Format) es un archivo de texto enASCII.

```
1 @relation weather
```

```
2
```

Todo archivo *.arff* debe comenzar con esta declaración en su primera línea (no es válido dejarla en blanco). Se requiere una cadena de caracteres, y si contiene espacios se los debe colocar entre comillas.

```
2  
3 @attribute outlook {sunny, overcast, rainy}  
4 @attribute temperature real  
5 @attribute humidity real  
6 @attribute windy {TRUE, FALSE}  
7 @attribute play {yes, no}  
8
```

Se incluye una línea por cada atributo que se vaya a incluir en el conjunto de datos, indicando su nombre y el tipo de dato.

Con *@attribute* se informa el nombre del atributo (debe comenzar por una letra y si contiene espacios tendrán que estar entrecomillados). Luego se indica el **tipo de dato**.

Cuando no es numérico, se indican entre { } los valores posibles. El formato de la fecha será del tipo "yyyy-MM-dd'T'HH:mm:ss".


```
9 @data
10 sunny,85,85,FALSE,no
11 sunny,80,90,TRUE,no
12 overcast,83,86,FALSE,yes
13 rainy,70,96,FALSE,yes
14 ...
```

Después de **@data** se incluyen los datos propiamente dichos. Cada atributo se separa con coma, todas las filas deben tener el mismo número de atributos (en coincidencia con la declaración **@attribute**).

Cuando no se disponga de algún dato, colocaremos un signo de interrogación (?) en su lugar. El separador de decimales tiene que ser obligatoriamente el punto y las cadenas de tipo string tienen que estar entre comillas simples.

