2022

MINERÍA DE DATOS

Ejemplos de aplicación de algoritmos con WEKA

Maria N. Moreno García

Contenido

1. Herramienta Weka

Weka es una herramienta de minería de datos de código abierto desarrollada en la Universidad de Waikato. Contiene la implementación de numerosos algoritmos de clasificación, regresión, asociación y agrupamiento. También permite aplicar a los datos diferentes filtros y algoritmos de preprocesamiento y cuenta con herramientas de visualización.

La información sobre la herramienta, así como el software para su instalación se encuentran en:

http://www.cs.waikato.ac.nz/ml/weka/

Al ejecutar la herramienta se obtiene la ventana siguiente:

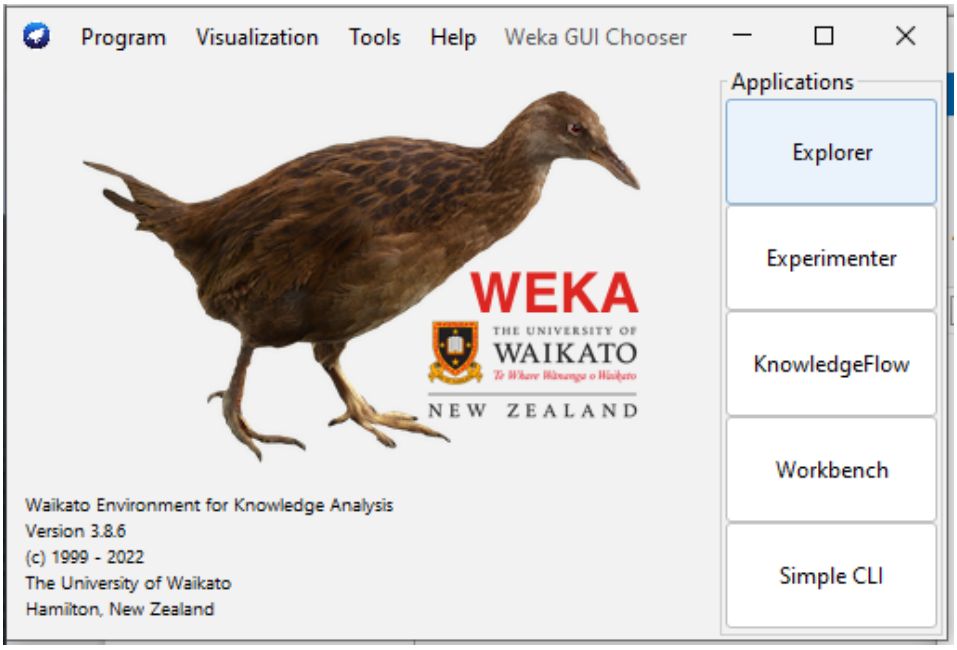

Figura 1. Ventana inicial de Weka

La opción “Explorer” permite aplicar los algoritmos al conjunto de datos que queramos estudiar y visualizar dichos datos de diferentes formas, así como los resultados de algunos algoritmos.

La opción “knowledgeFlow” es un entorno gráfico que se utiliza para aplicar diferentes algoritmos sucesivamente, de manera que la salida de uno sirva de entrada al siguiente. De esta forma se puede diseñar todo el proceso desde el preprocesamiento hasta la visualización de resultados y ejecutar todos los algoritmos que lo forman de una vez.

La propia herramienta nos ofrece varios conjuntos de datos de prueba, uno de los cuales es el que vamos a utilizar como ejemplo en este documento.

2. El conjunto de datos “Iris”

Se trata de un conjunto de datos sobre flores de tres tipos diferentes: Iris setosa, iris virgínica, e iris versicolor. El fichero contiene 150 registros con los atributos longitud del pétalo, anchura del pétalo, longitud del sépalo, anchura del sépalo y el tipo de flor, éste último es el atributo etiqueta (clase).

Al pulsar la opción “Explorer” de Weka obtenemos la ventana que se muestra en la figura 2. Pulsando alguno de los botones “Open file”, “Open URL” u “Open DB” podemos abrir un fichero de datos que se encuentre en cualquier ubicación. Con el botón “Open file” se puede abrir alguno de los ficheros de muestra de la herramienta (figura 3).

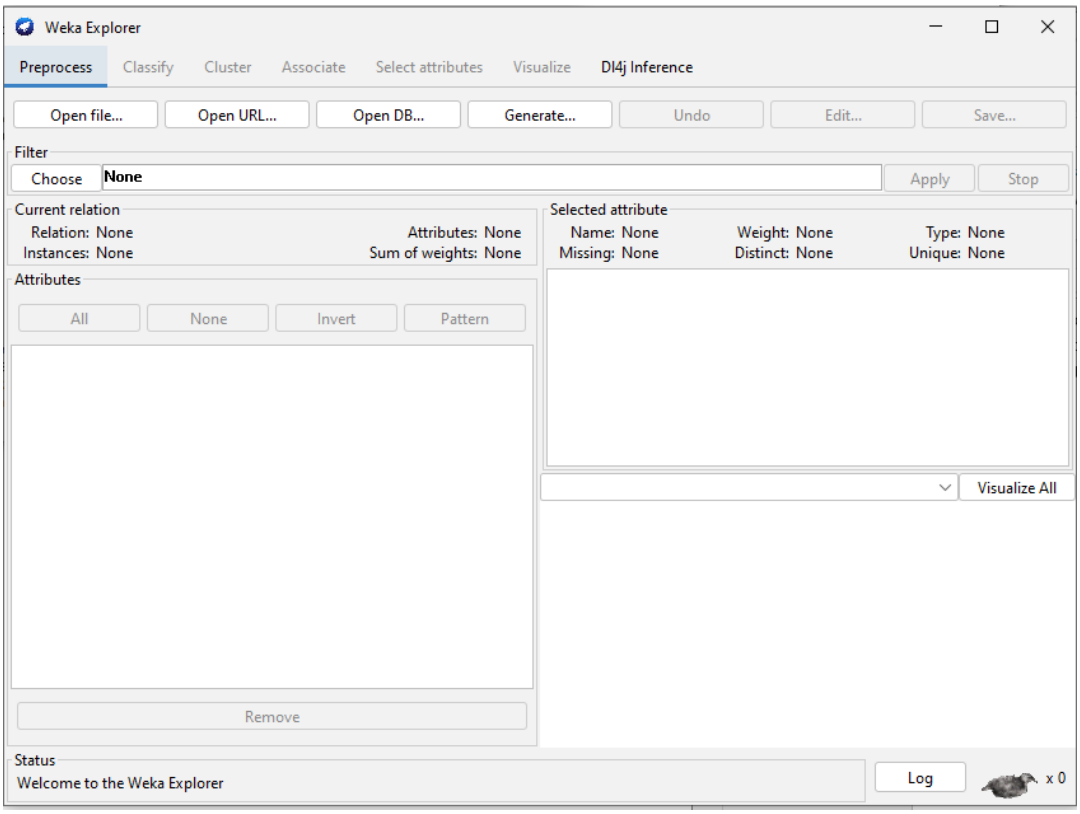

Figura 2. Ventana inicial de la opción “Explorer”

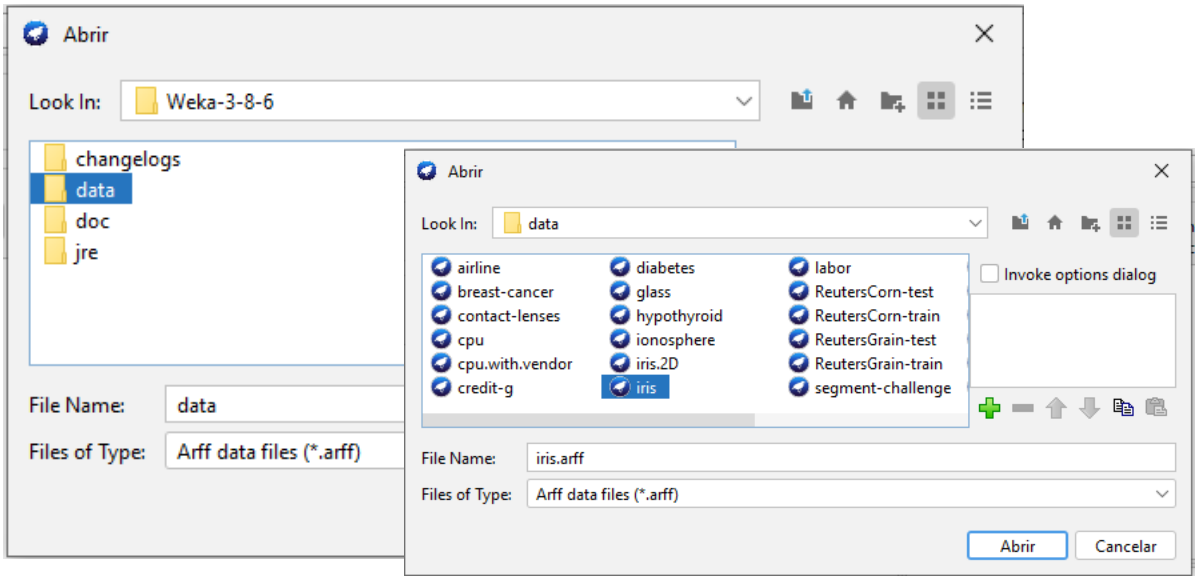

Figura 3. Acceso a los datos de muestra de Weka

Al seleccionar y abrir el fichero “iris.arff” obtenemos la ventana de la figura 4, en la que se muestran las características de los datos de dicho fichero. En la parte izquierda aparecen los 5 atributos: sepallength, sepalwidth, petalength, petalwidth (atributos descriptivos) y el atributo de clase. Al seleccionar uno de ellos aparece en la parte derecha su distribución de valores. En la figura está seleccionado el atributo de clase, por lo que podemos observar sus tres valores (tres clases), su representación en tres colores distintos y su distribución (50 registros de cada valor). Las tres clases se corresponden con tres tipos de flores iris (setosa, virgínica y versicolor).

Es posible editar el archivo para ver y modificar, si es necesario, los valores de los atributos pulsando el botón “Edit” (Figura 5).

También se pueden visualizar las distribuciones de valores de todos los atributos pulsando el botón “visualize All” situado en la parte inferior derecha de la ventana. En la figura 6 podemos ver esa distribución para intervalos de valores de esos atributos y la proporción de registros de cada clase representados en diferentes colores para cada intervalo.

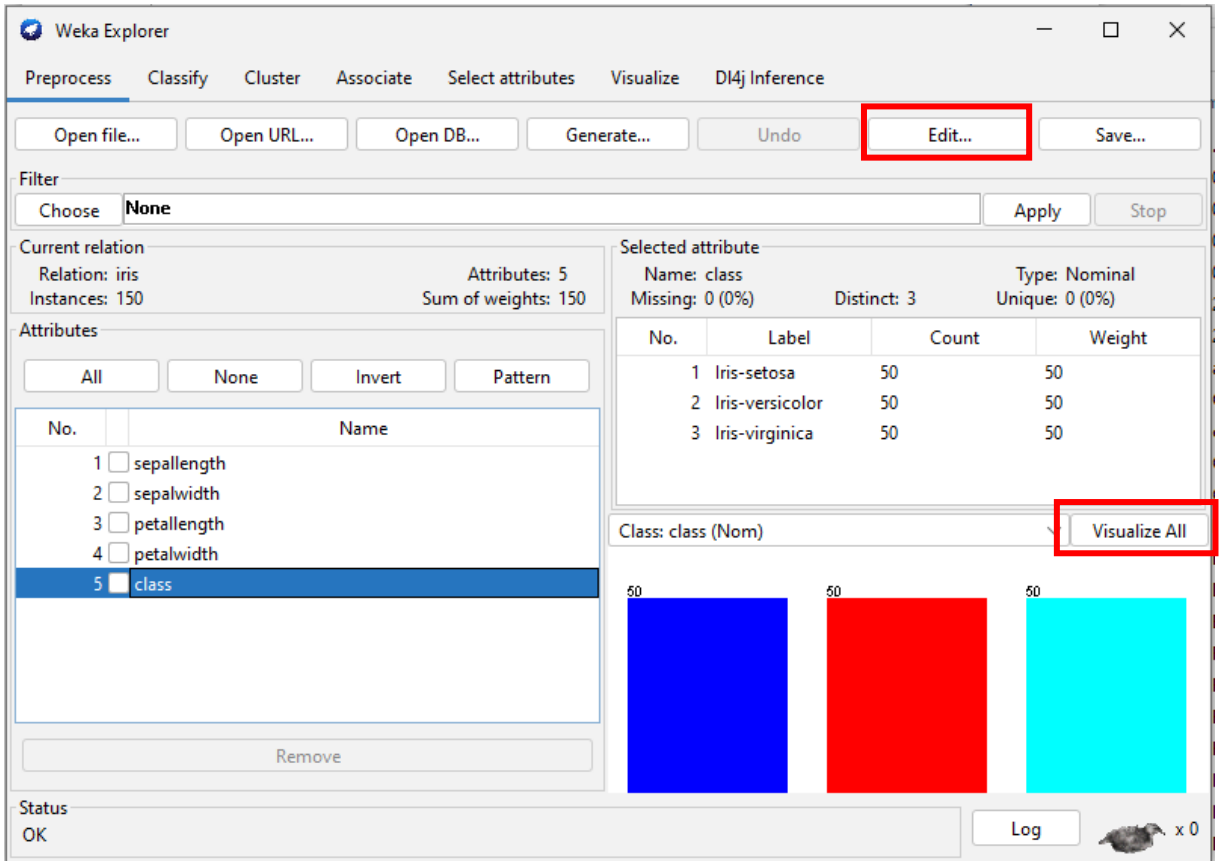

Figura 4. Atributos del archivo “iris”

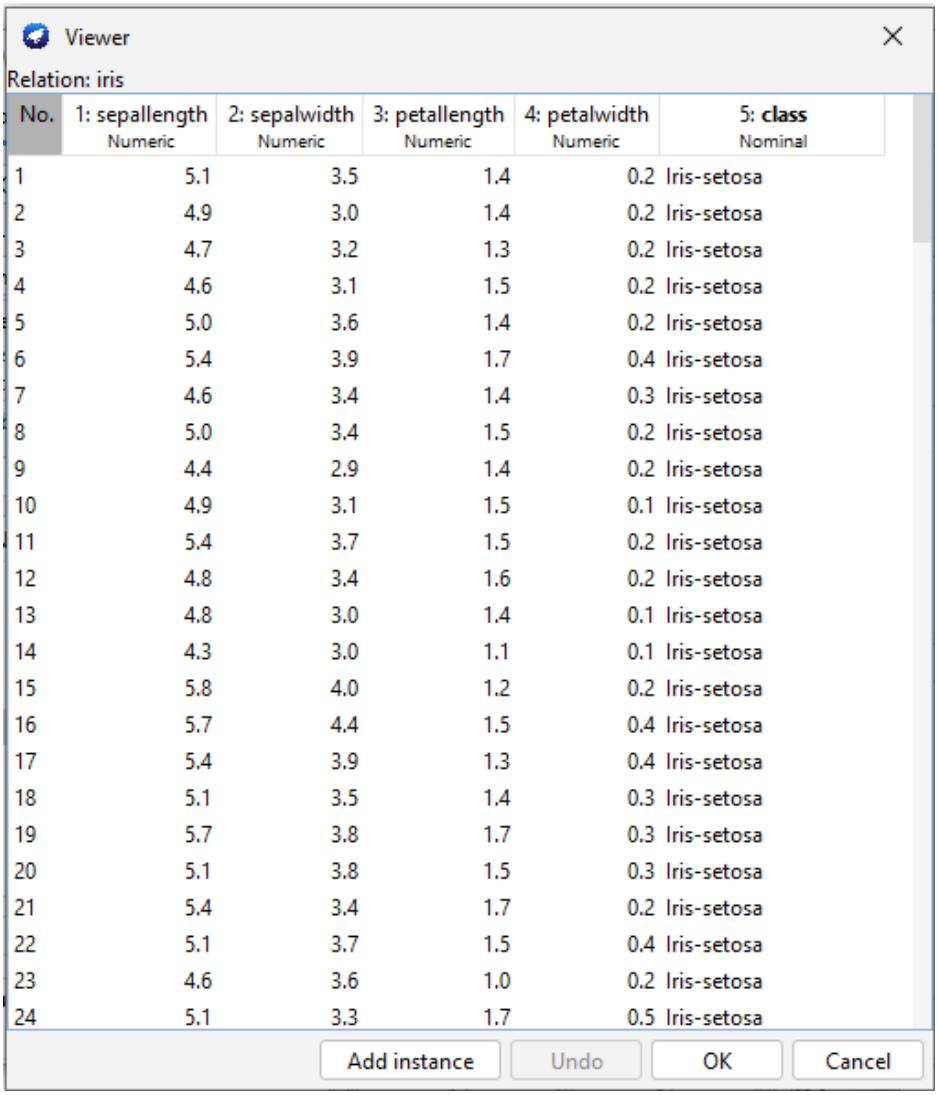

Figura 5. Edición del archivo "iris"

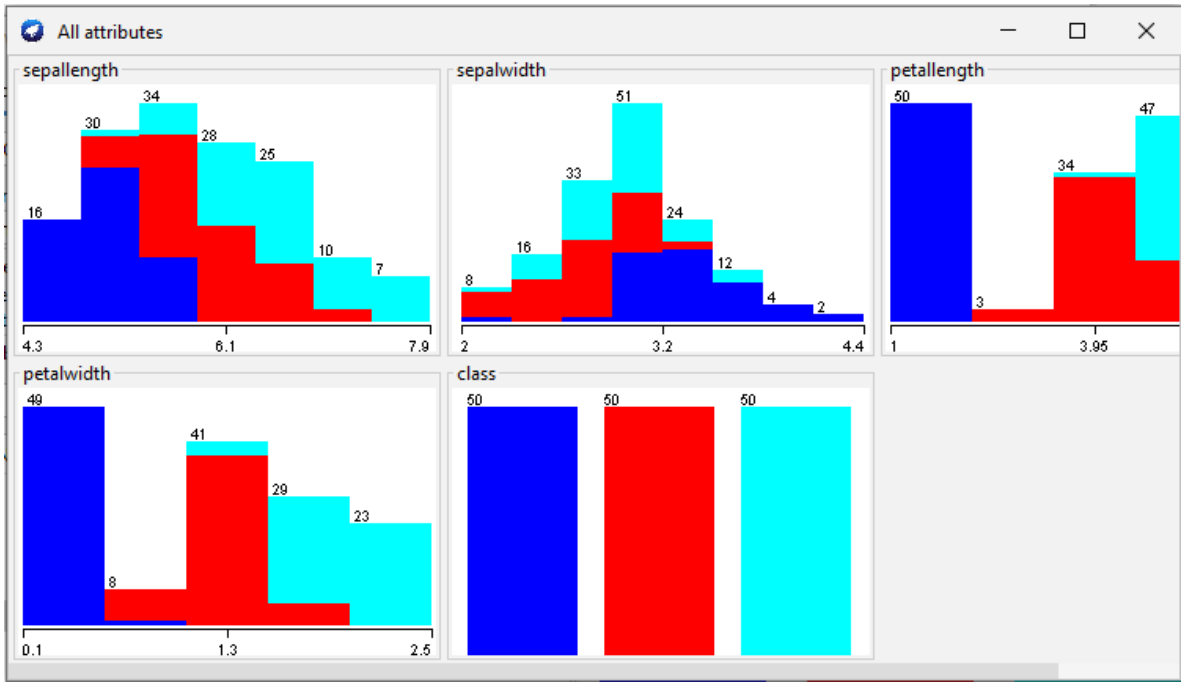

Figura 6. Distribución de valores de todos los atributos

3. Aplicación de algoritmos predictivos (supervisados)

3.1. Clasificación

Weka dispone de numerosos algoritmos de clasificación que pueden aplicarse a los datos del archivo que se ha abierto. Para acceder a ellos hay que seleccionar la pestaña "Classify". En la ventana que aparece se selecciona el algoritmo con el botón "Choose" (figura 7). Debajo se muestran las opciones de prueba del algoritmo seleccionado, siendo la opción de validación cruzada con k=10 la que se da por defecto. Hay que seleccionar también el atributo de clase, por defecto es el que está colocado el último.

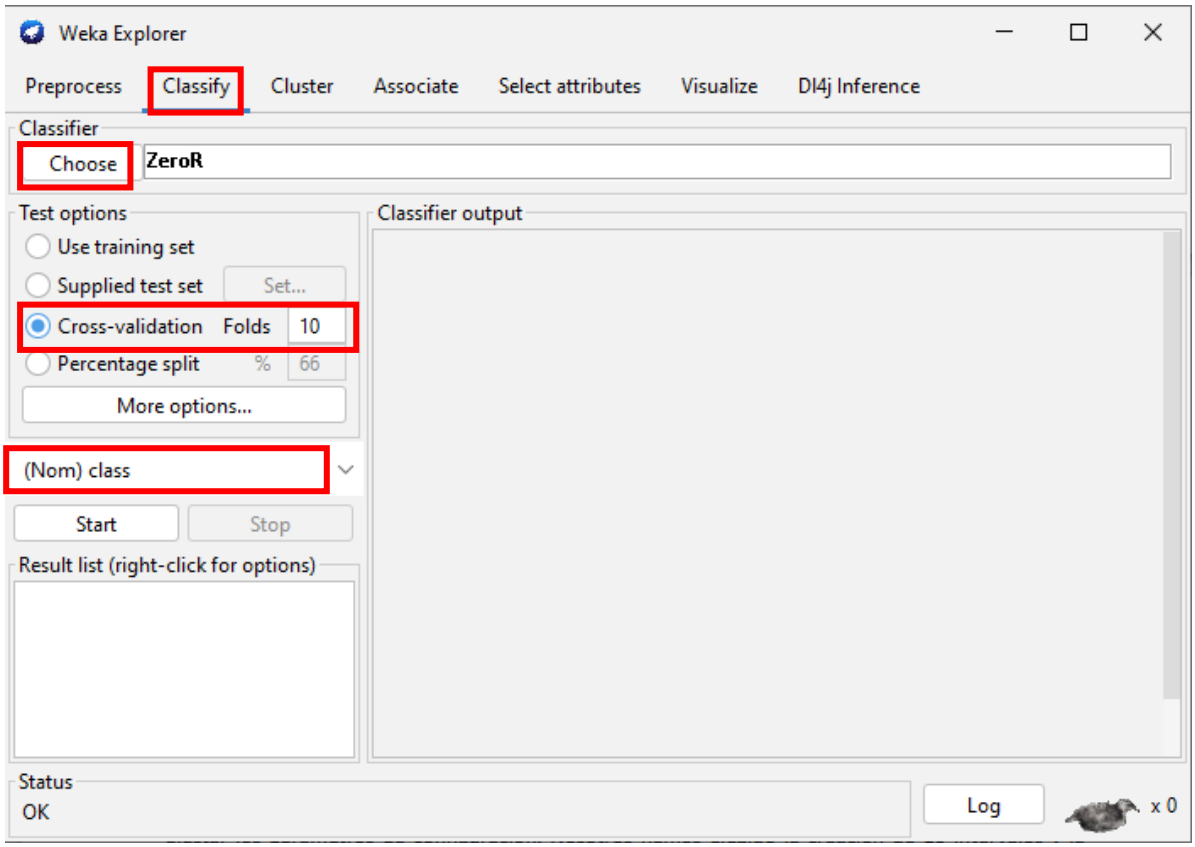

Figura 7. Ventana de los algoritmos de clasificación

Con el botón "Choose" seleccionamos el algoritmo que queremos aplicar (figura 8). Seleccionamos el algoritmo J48 de árboles de decisión. Pulsando el botón izquierdo del ratón encima del nombre del algoritmo aparece un cuadro de diálogo con todas las opciones de configuración del algoritmo, si no lo configuramos se ejecutará con las opciones por defecto.

Al ejecutarlo pulsando el botón "Start" obtenemos los resultados (diferentes tipos de métricas de calidad de la clasificación) en el cuadro de texto de la derecha (figura 9). Podemos ver los resultados de exactitud (porcentaje de instancias correctamente clasificadas), precisión, recall, medida F, AUC, etc. Los resultados de estas métricas se dan para cada una de las clases y también se da el valor medio. Al final se muestra la matriz de confusión con los errores producidos en cada clase.

Pulsando con el botón derecho del ratón sobre la correspondiente entrada de la lista de resultados (figura 10) podemos visualizar el árbol que se ha construido con los datos del archivo "iris" (figura 11).

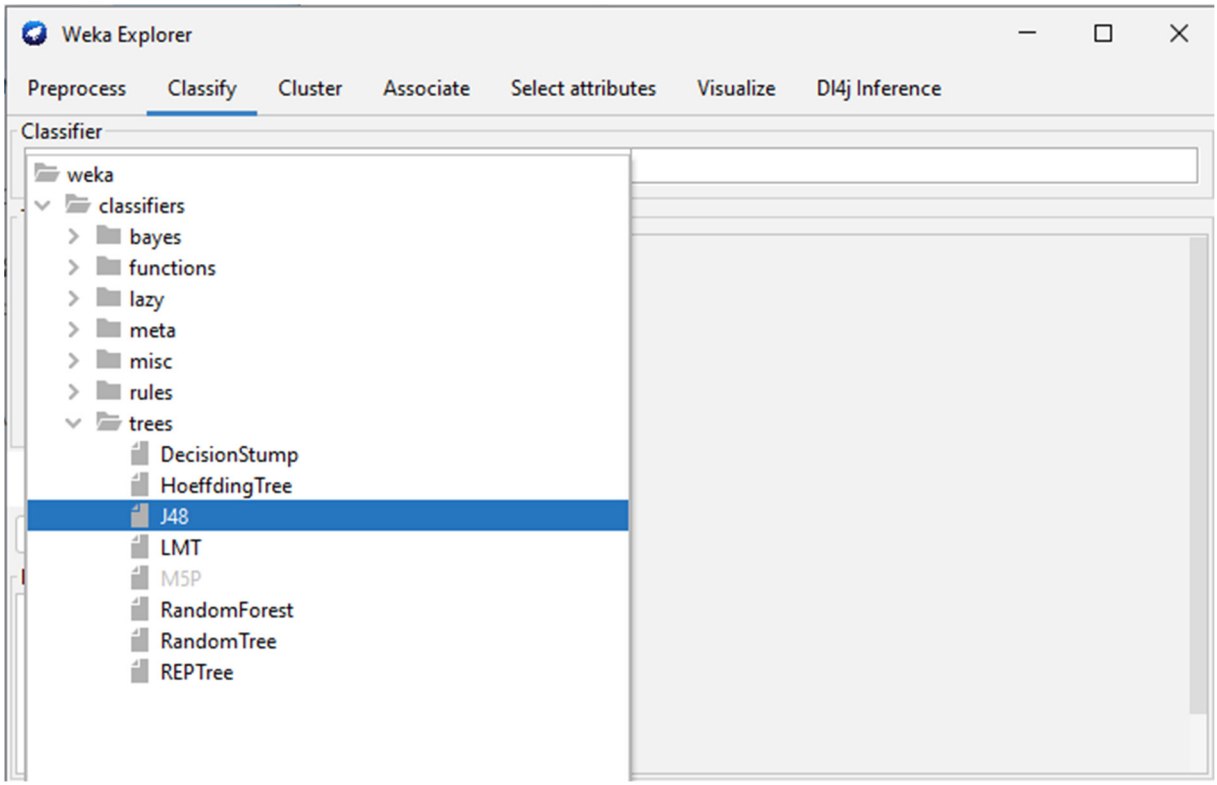

Figura 8. Selección de algoritmos de clasificación

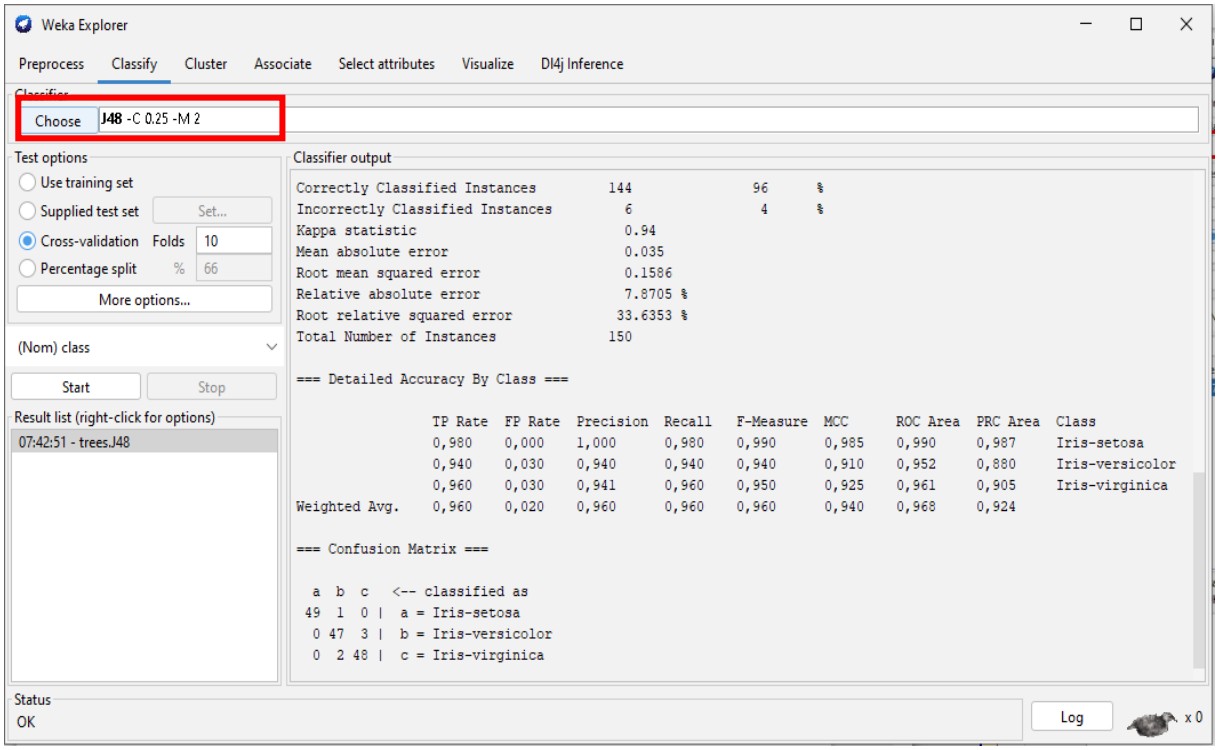

Figura 9. Resultados del algoritmo J48

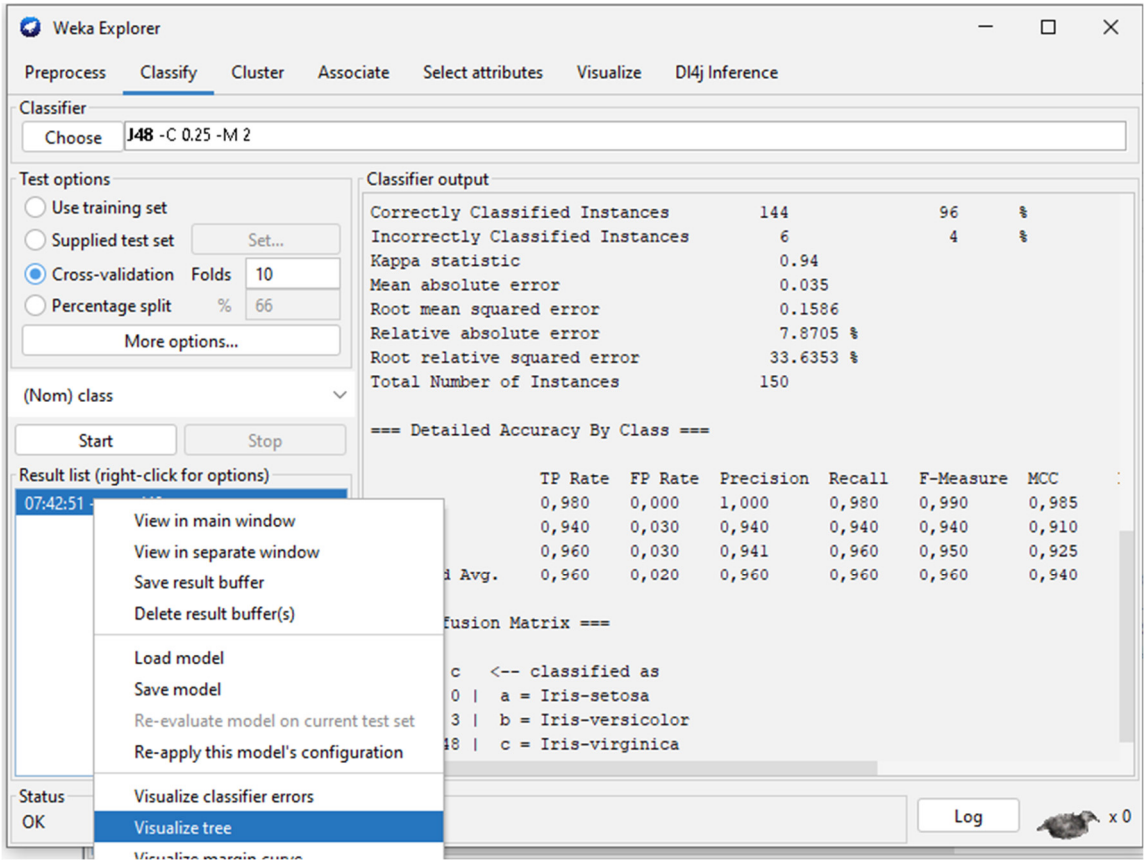

Figura 10. Resultados del algoritmo J48

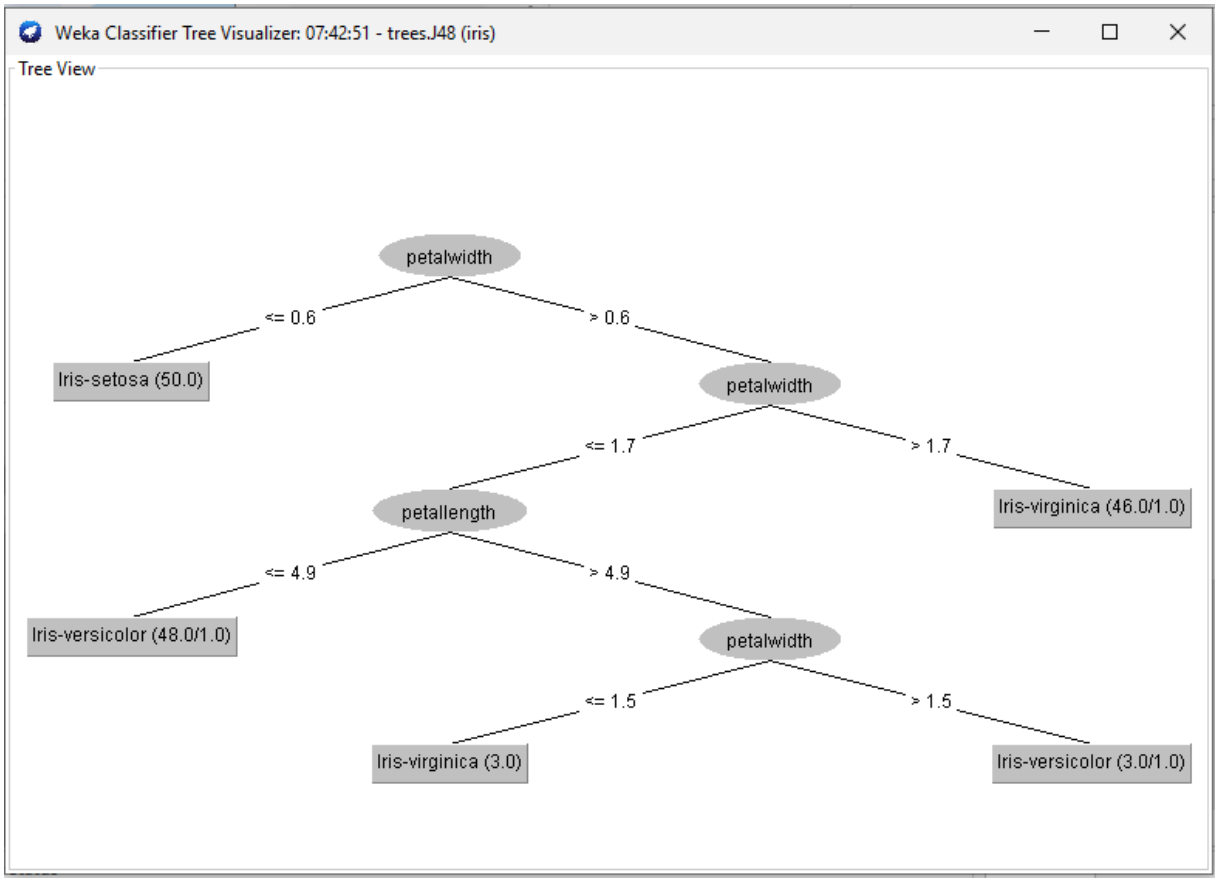

Figura 11. Visualización del árbol de decisión

El árbol de la figura 11 es el modelo de predicción inducido por el algoritmo J48. Consiste en un conjunto de reglas obtenidas a partir de los datos del archivo "iris" de los que se conoce su clase. Las hojas del árbol representan los valores del atributo etiqueta (clase) y el resto de los nodos representan evaluaciones de los atributos descriptivos. Por ejemplo, una de las reglas del árbol se puede expresar de la siguiente forma:

Si **petalwidth > 0.6** y **petalwidth <= 1.7** y **petallength <=4.9** entonces **la clase es iris-versicolor**

Este modelo puede aplicarse posteriormente a datos sin clasificar para predecir la clase.

Podemos seguir el mismo procedimiento para aplicar otros algoritmos de clasificación como redes bayesianas, SVM, K-NN o multiclasificadores (figura 12). Por ejemplo, SMO es un algoritmo de máquinas de vectores de soporte (SVM), IBK implementa un método K-NN y los multiclasificadores se encuentran en el grupo de clasificadores "meta".

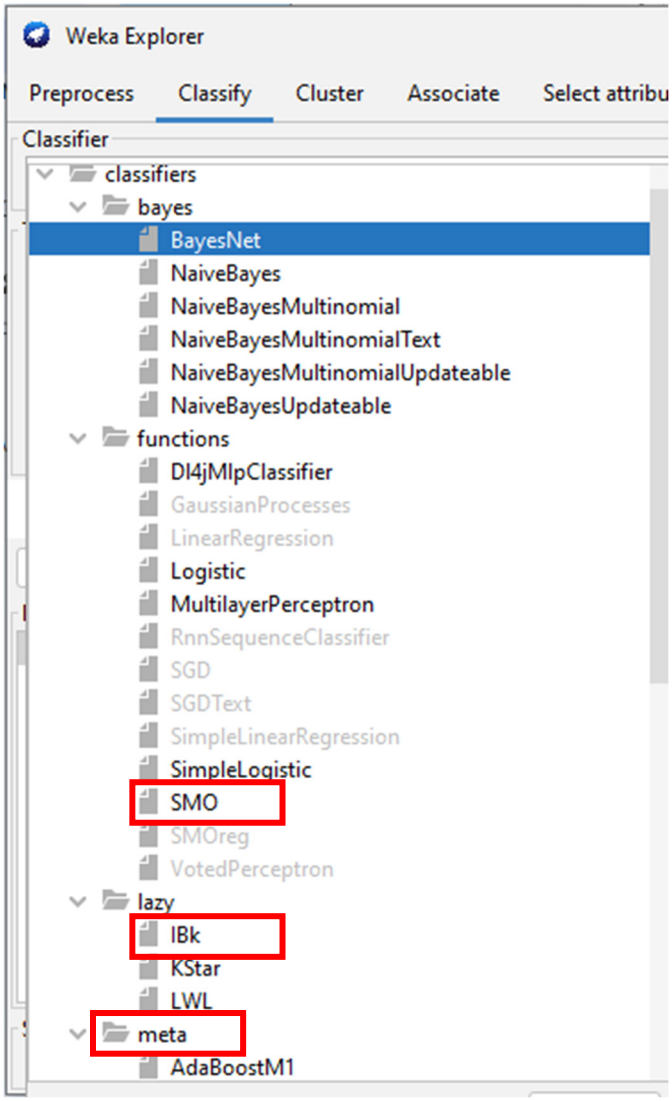

Figura 12. Algoritmos de clasificación de Weka

Aplicamos una **red bayesiana** que configuramos como se muestra en la figura 13. Como algoritmo de búsqueda usamos K2 que es el que viene por defecto. La figura 14 muestra los resultados obtenidos, los cuales son ligeramente peores que los del árbol de decisión ya que hay más instancias clasificadas incorrectamente. La estructura de la red puede verse en la figura 15.

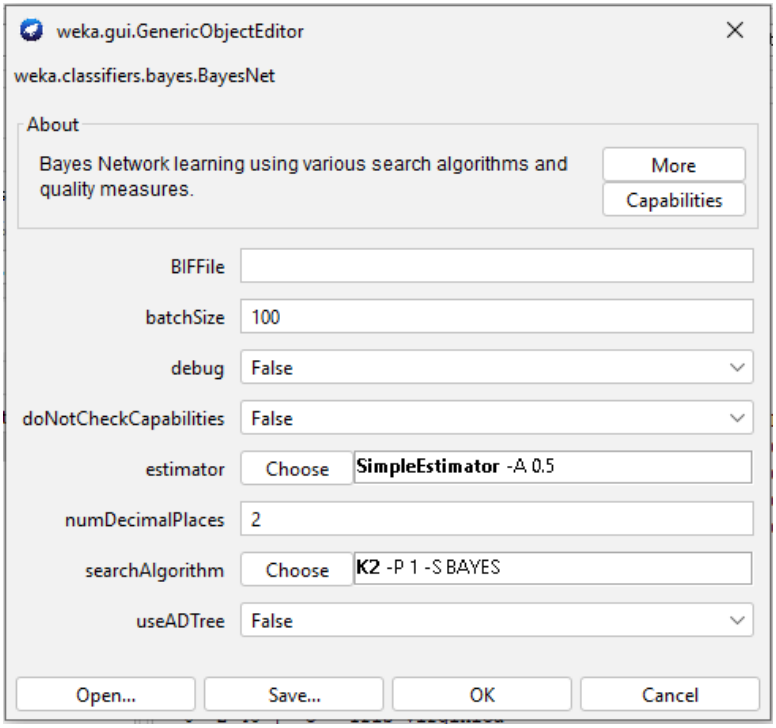

Figura 13. Configuración de la red bayesiana

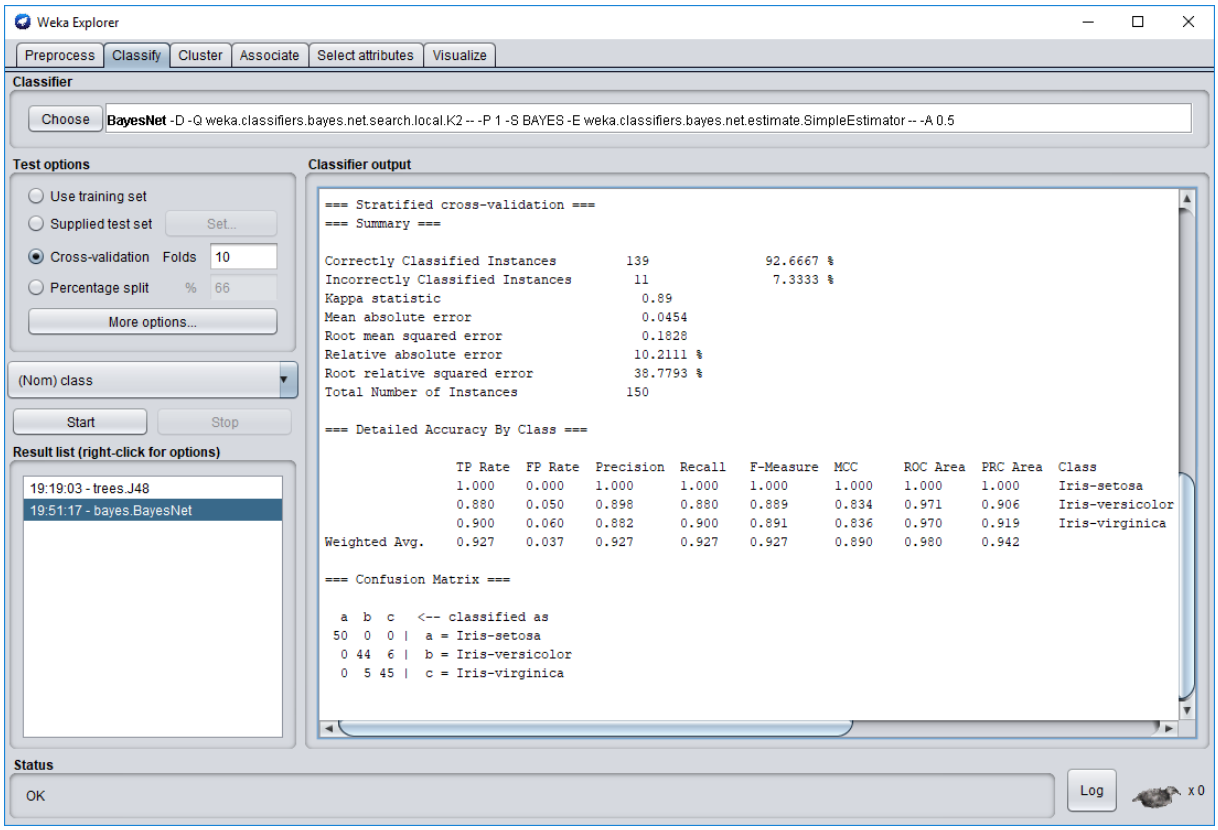

Figura 14. Resultados de la red bayesiana

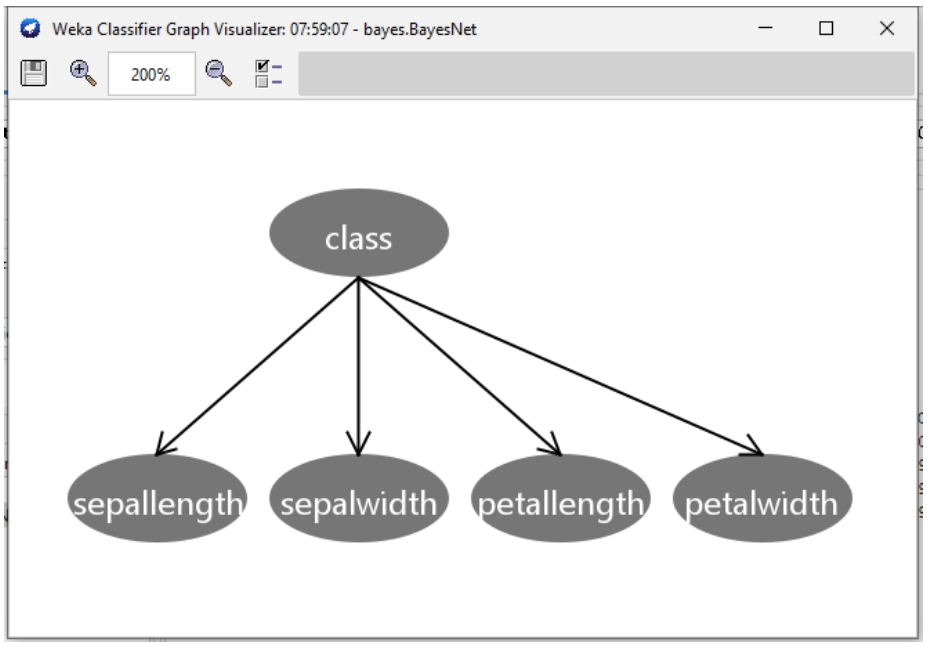

Figura 15. Visualización de la red bayesiana

3.2. Regresión

Weka también tiene implementados algunos algoritmos de regresión para la predicción de valores. Estos algoritmos solo pueden aplicarse cuando el atributo etiqueta tiene valores numéricos continuos, por lo que tenemos que seleccionar otros datos para poder aplicarlos.

Elegimos el conjunto "cpu.arff" que es otro conjunto de datos de muestra de Weka. El archivo contiene información sobre diferentes características de CPUs: Tiempo de ciclo (MYCT), memoria principal mínima (MMIN), memoria principal máxima (MMAX), memoria caché (CACH), canales mínimos (CHMIN), canales máximos (CHMAX). El objetivo será predecir el rendimiento a partir de esos atributos. Por lo tanto, el atributo con el nombre "class" contiene el valor del rendimiento.

En la figura 16 podemos ver que al seleccionar el atributo "class", se proporciona información sobre valores mínimo, máximo, media y desviación estándar de dicho atributo en lugar de aparecer el número de instancias de cada clase como ocurre en problemas de clasificación cuando el atributo etiqueta tiene valores discretos.

Para seleccionar el método de regresión que queremos aplicar pulsamos el botón "Choose" y seleccionamos uno de los métodos de regresión que se encuentran en la carpeta "functions". Para este ejemplo vamos a elegir el método de regresión lineal "LinearRegression" (figura 17). A continuación, configuramos los parámetros de prueba de la misma forma que para los métodos de clasificación y pulsamos "Start" para ejecutar el algoritmo (figura 18).

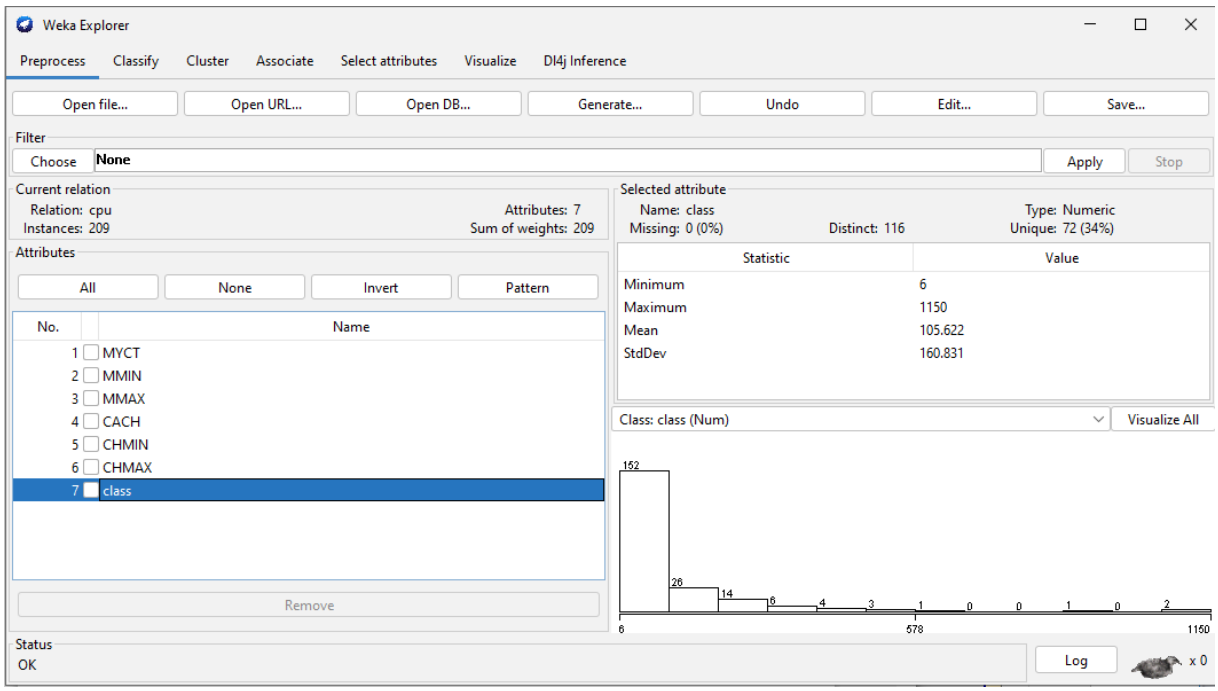

Figura 16. Conjunto de datos cpu.arff

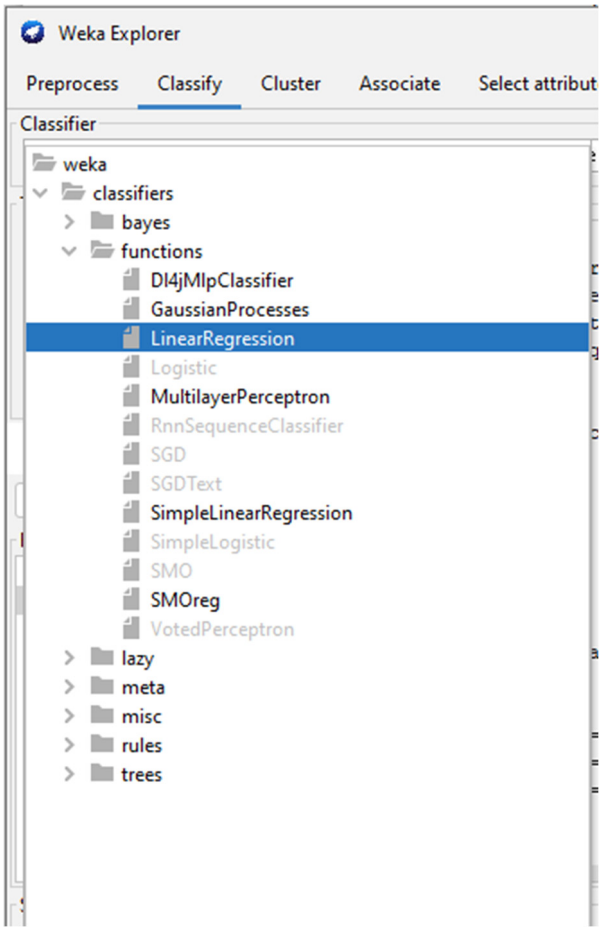

Figura 17. Selección de un método de regresión

Al ejecutar el algoritmo obtenemos unos resultados diferentes a los obtenidos con los métodos de clasificación. En este caso no se pueden utilizar las métricas de exactitud, precisión etc. porque lo que se predice es un valor, mientras que en el caso de la clasificación lo que se predice es una clase y se puede saber si la clasificación es correcta o no. En los problemas de regresión lo que se comprueba es lo cercano que está el valor predicho al valor real. Por eso las métricas que se utilizan están basadas en el error: error absoluto medio, error cuadrático medio, etc.

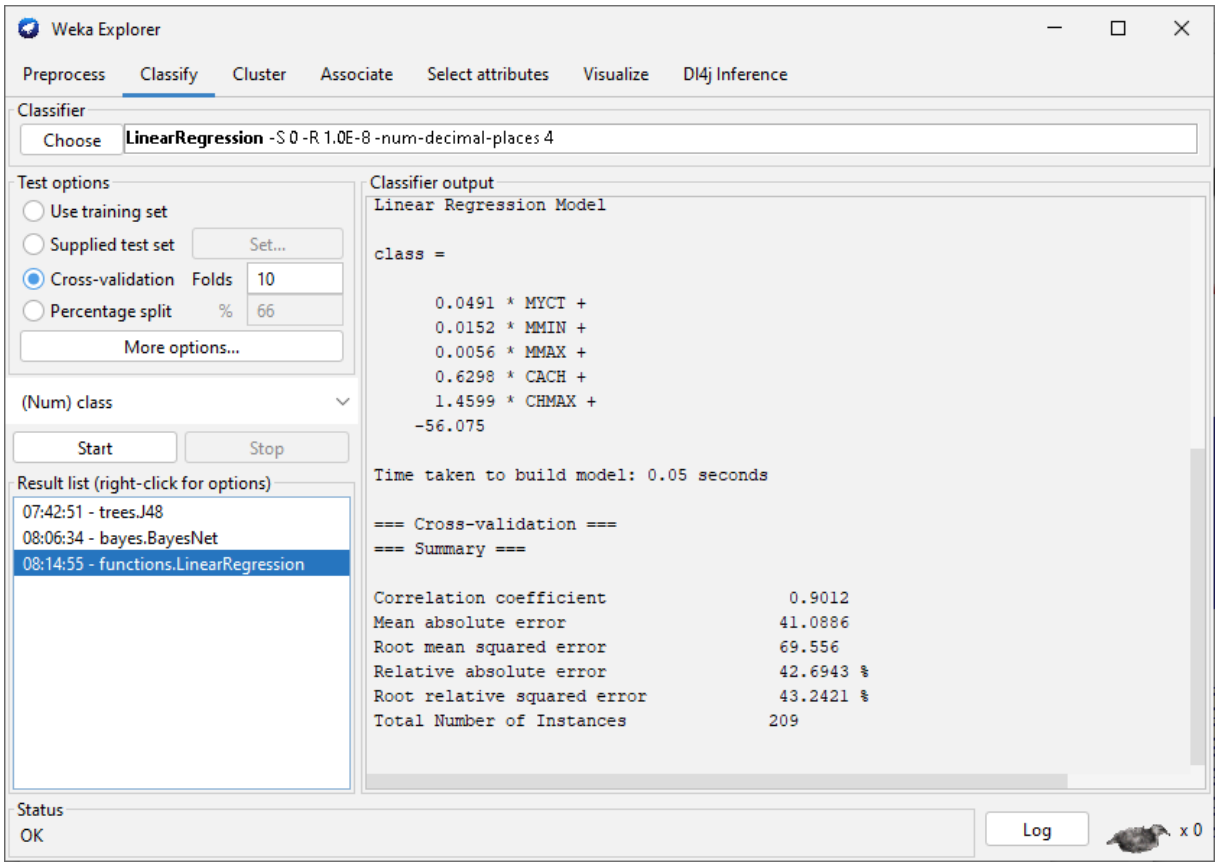

Figura 18. Resultados del algoritmo de regresión

Además de los métodos de regresión clásicos, Weka también implementa algoritmos de **árboles de regresión**. Estos algoritmos se encuentran en el grupo "trees". Seleccionamos REPTree que es uno de los árboles de regresión más conocidos y utilizados (figura 19). Este tipo de árboles también se pueden utilizar en problemas de clasificación.

En la figura 20 vemos los resultados obtenidos tras ejecutar el algoritmo, dados también en función de los diferentes tipos de error. En la pantalla de resultados, por encima de los errores, también se muestra el árbol obtenido en forma de reglas (figuras 20 y 21). Así mismo, se puede visualizar gráficamente el árbol seleccionando la opción "Visualize tree" del menú que se despliega pulsando el botón derecho del ratón con el cursor sobre el algoritmo en la lista de resultados. La representación gráfica del árbol se puede ver en la figura 22.

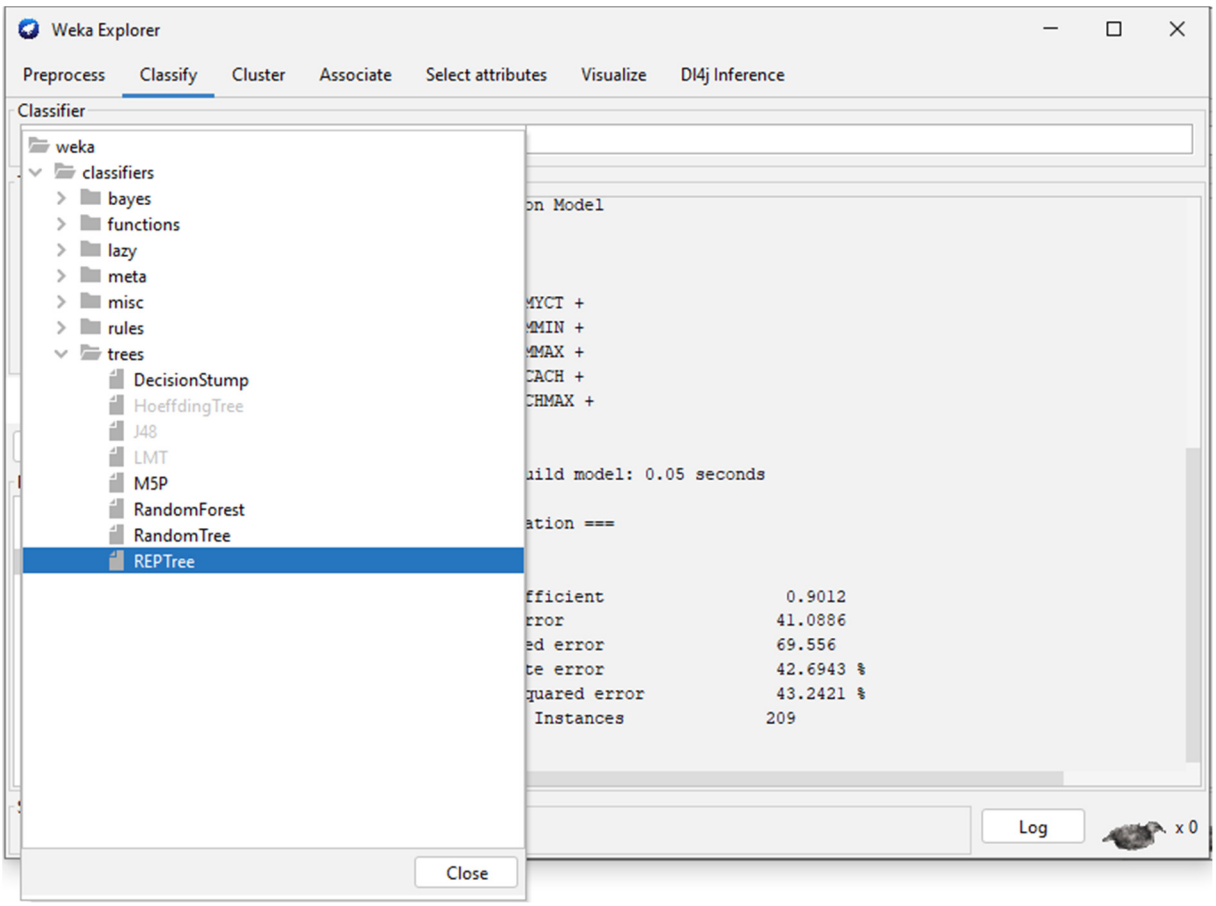

Figura 19. Selección de un algoritmo de árboles de regresión

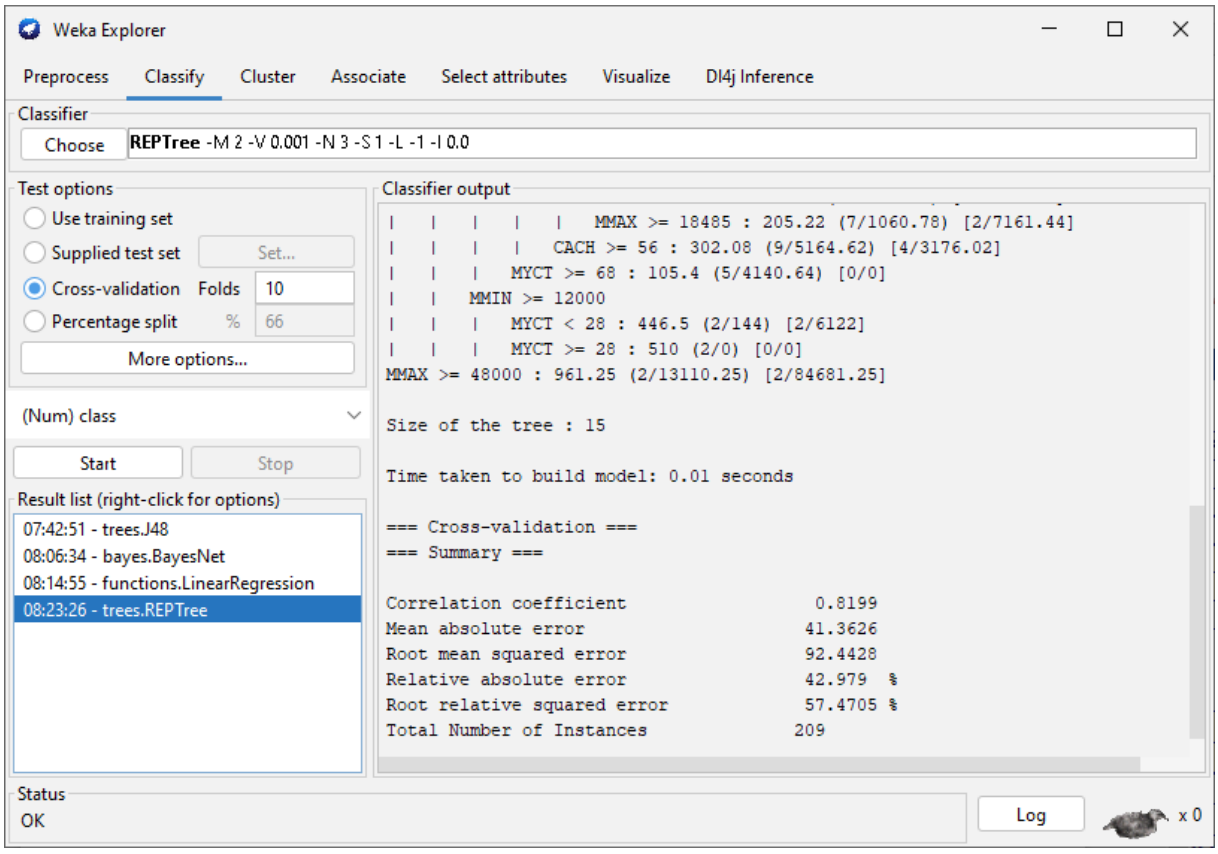

Figura 20. Resultados del algoritmo de árboles de regresión REPTree

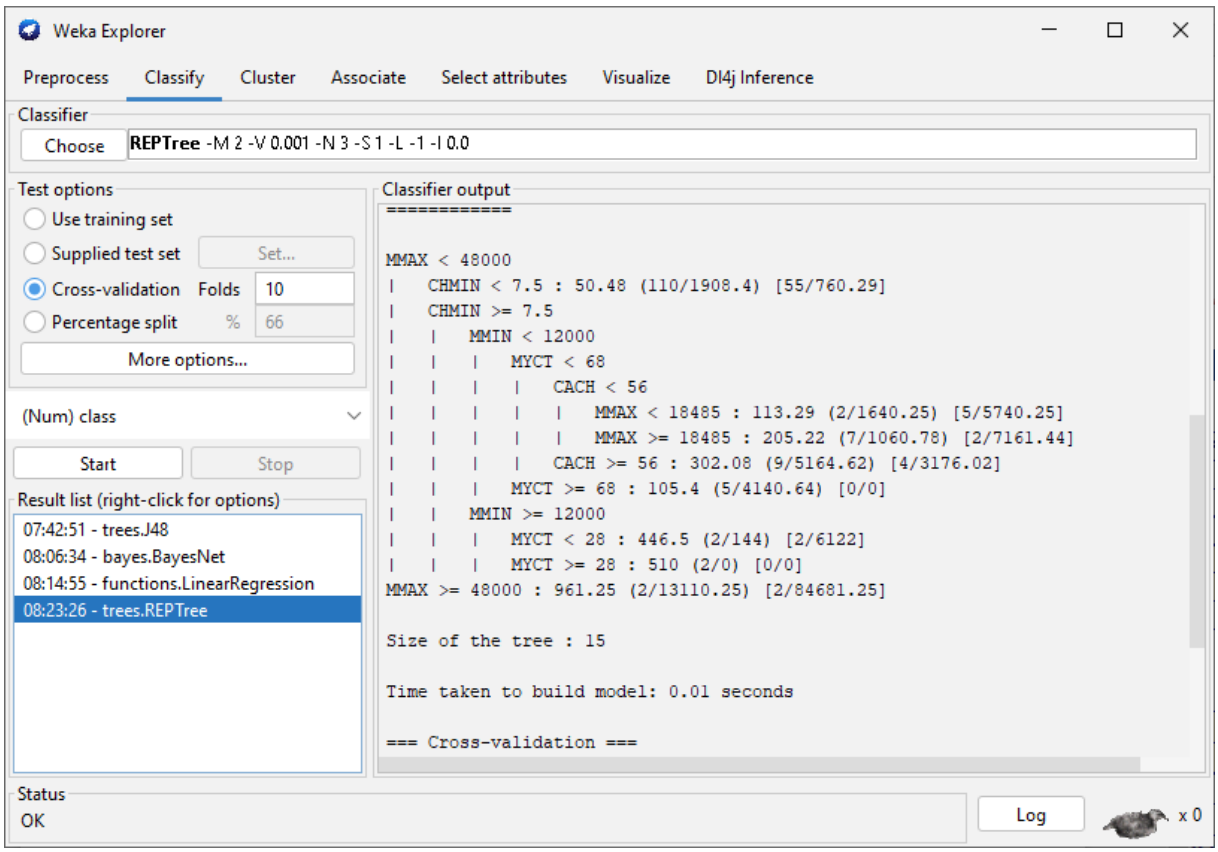

Figura 21. Selección de la opción de visualización del árbol de regresión REPTree

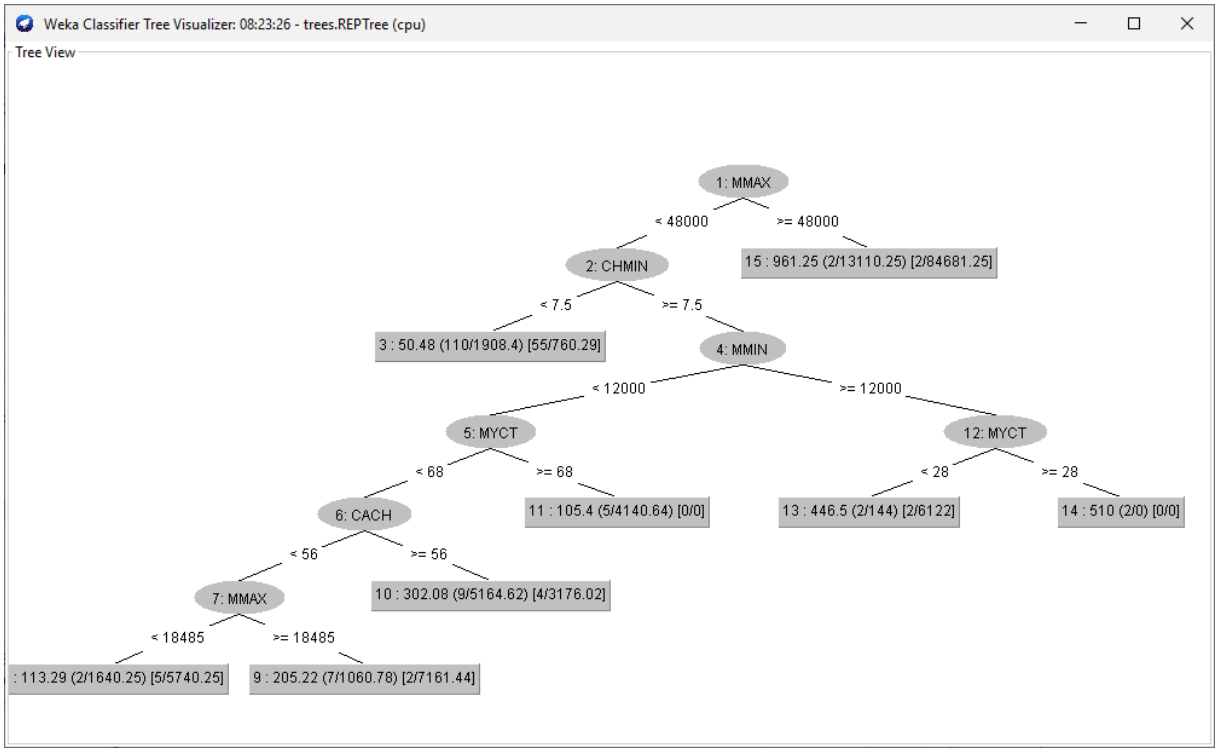

Figura 22. Representación gráfica del árbol de regresión REPTree

4. Aplicación de algoritmos descriptivos (no supervisados)

4.1. Reglas de asociación

Para poder utilizar los algoritmos de asociación de Weka hay que seleccionar la pestaña “Associate”. Sin embargo, para generar reglas de asociación es necesario realizar un preprocesamiento previo de los datos, ya que no se pueden inducir reglas a partir de atributos que tengan valores numéricos continuos, sino que hay que crear intervalos de valores y expresar las reglas en términos de dichos intervalos. Para ello, en la ventana de preprocesamiento, que es la de inicio, en la que se han estado visualizando los atributos, pulsamos el botón “Choose” (a la izquierda de la ventana) para seleccionar un algoritmo de preprocesamiento (figura 23). De todos los algoritmos posibles, elegimos uno no supervisado de discretización (figura 24).

Al pulsar el botón izquierdo del ratón sobre el nombre “Discretice” que aparece en la ventana de preprocesamiento después de elegir el algoritmo (figura 25), nos aparece un cuadro de diálogo para ajustar los parámetros de configuración. Nosotros hemos elegido la creación de 10 intervalos y la discretización de todos los atributos a la vez (“first-last”) (figura 25). Después de cerrar el cuadro de configuración, pulsamos “Apply” para obtener los intervalos de valores. Como puede observarse en la figura 26, al seleccionar uno de los atributos de la parte izquierda de la ventana aparece en la parte derecha su distribución de valores y su representación mediante 10 barras, cada una correspondiente a un intervalo. El color representa la proporción de registros de cada clase en cada uno de los intervalos.

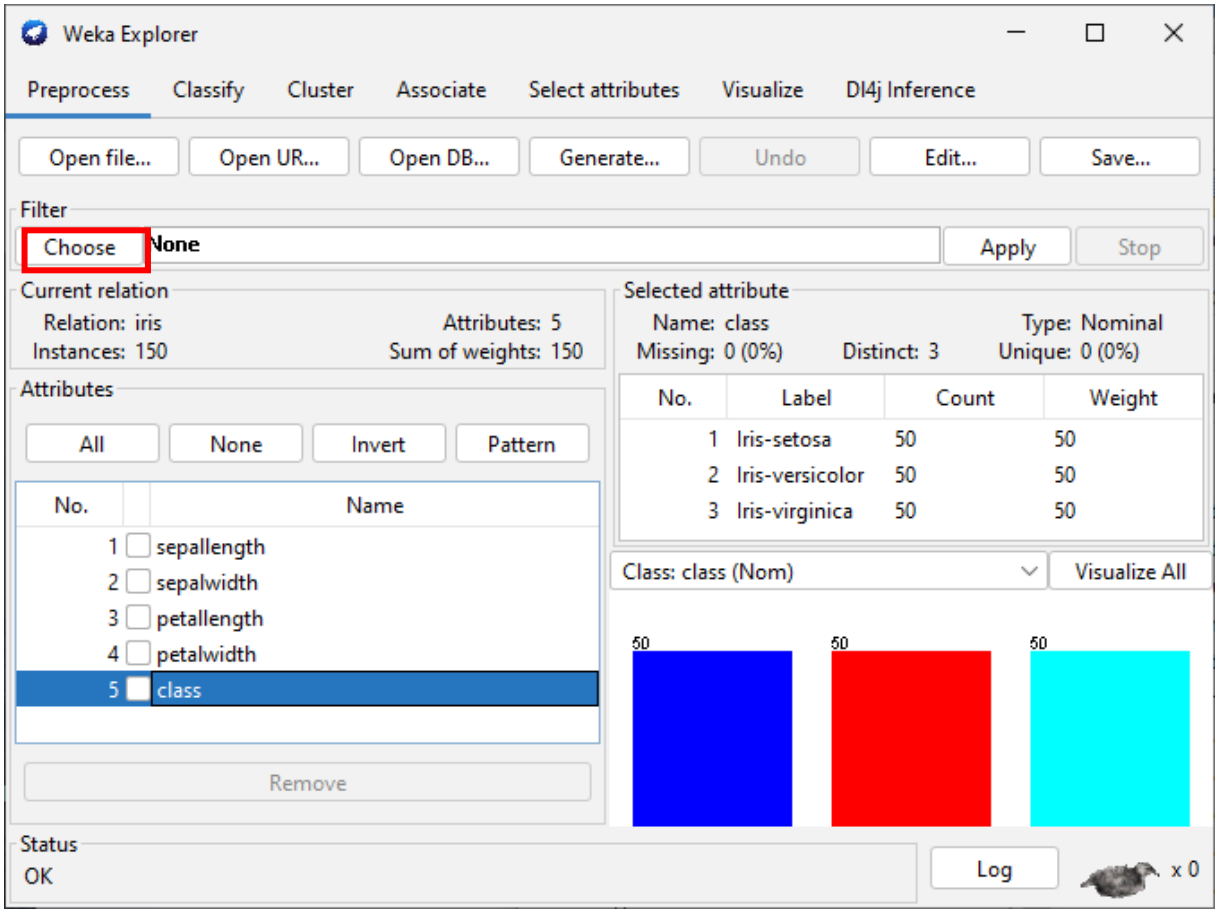

Figura 23. Ventana de preprocesamiento

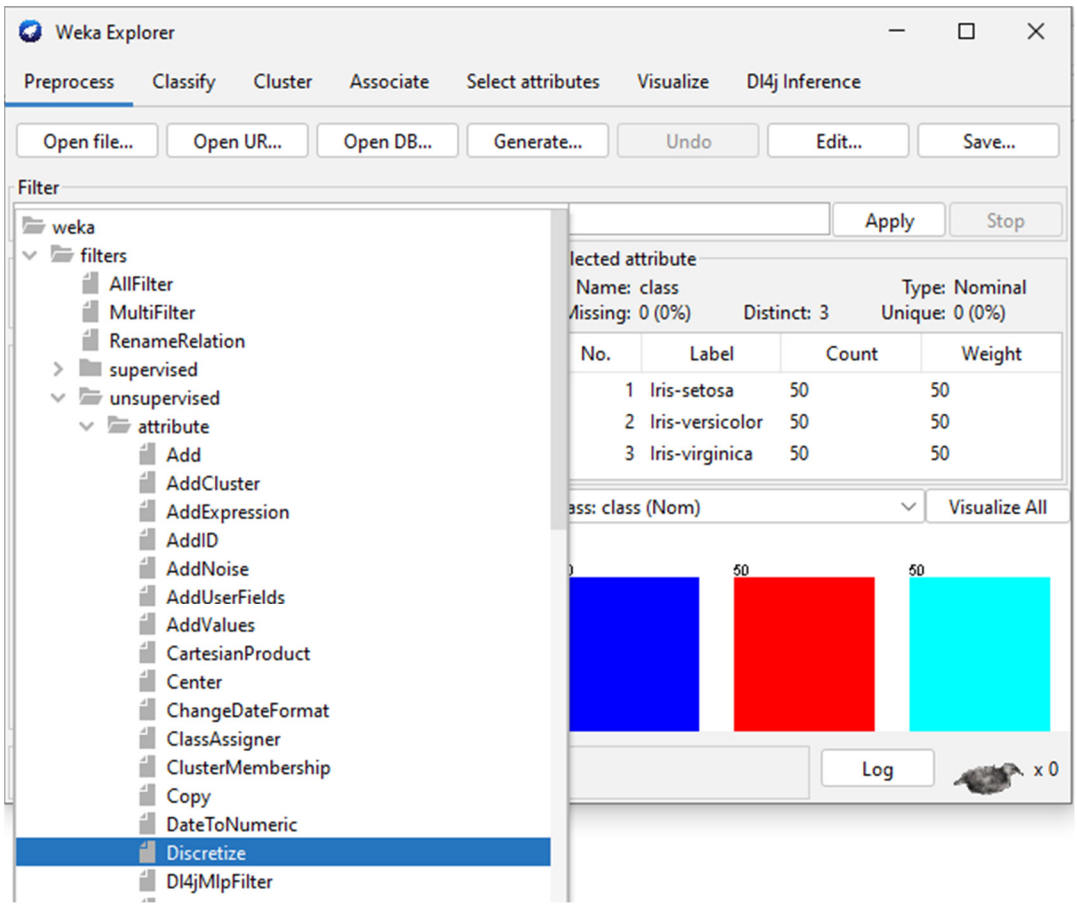

Figura 24. Elección de un algoritmo de preprocesamiento

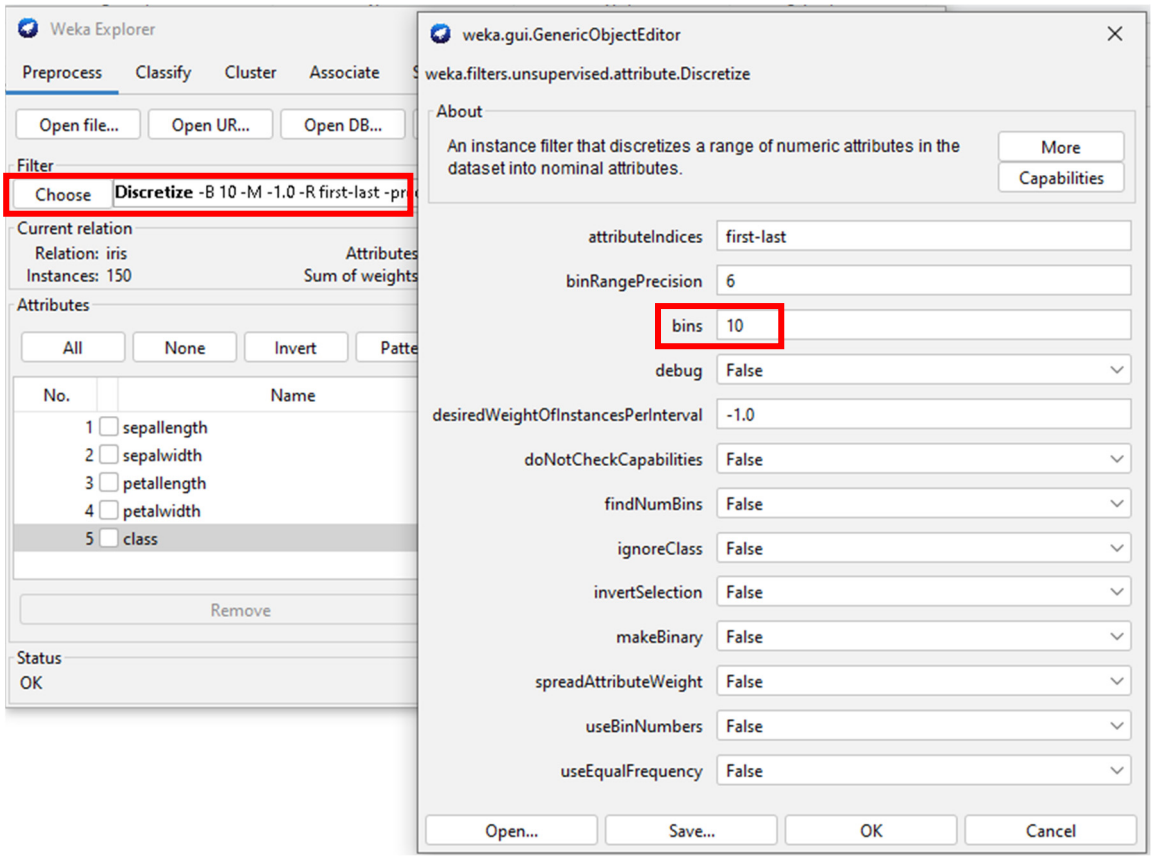

Figura 25. Configuración del algoritmo de discretización

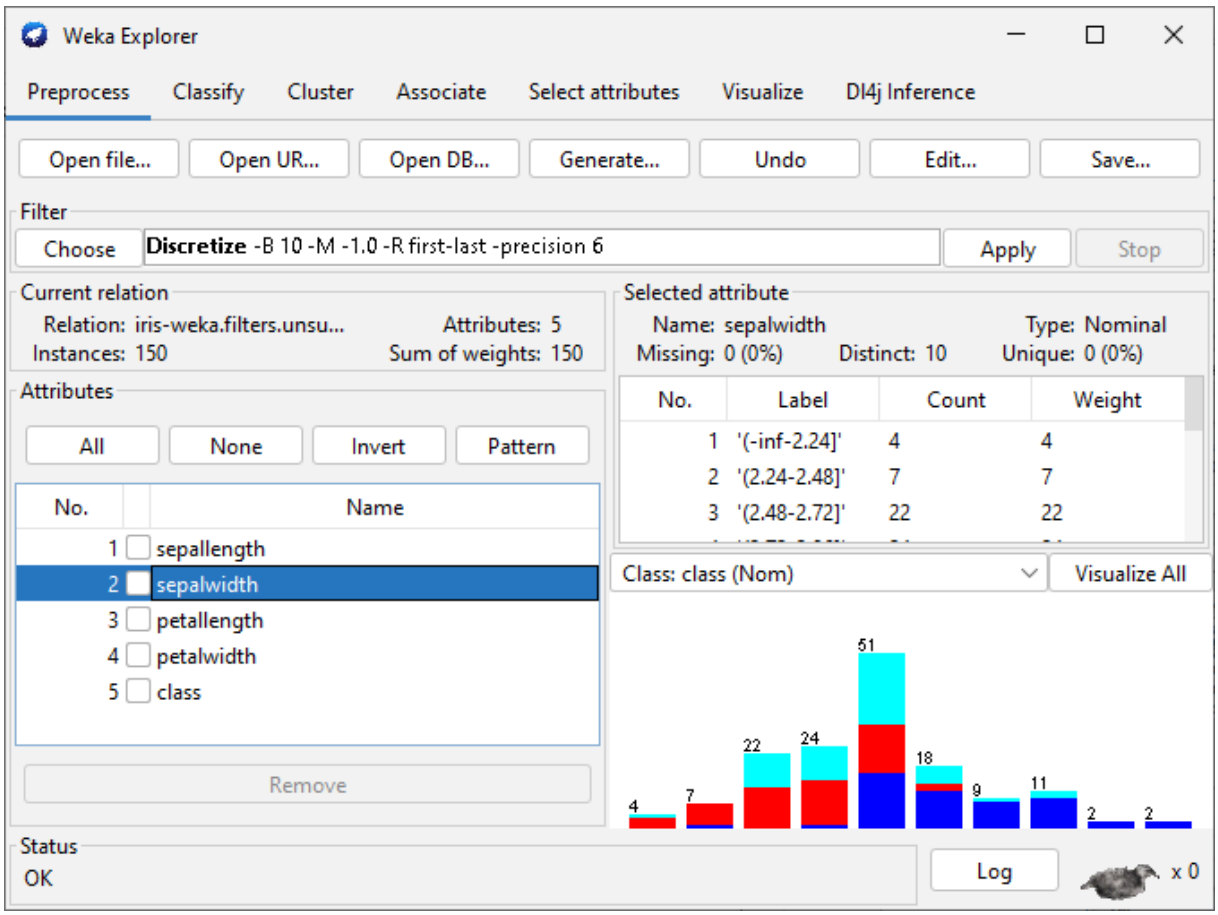

Figura 26. Atributo sepalwidth discretizado en 10 intervalos

Una vez discretizados los atributos podemos aplicarles un algoritmo de reglas de asociación. Para ello seleccionamos la pestaña “Associate” y elegimos el algoritmo pulsando el botón “Choose” en la nueva ventana. En este caso elegimos el algoritmo “Apriori”, como en casos anteriores, pulsando con el botón izquierdo del ratón sobre el nombre accedemos al cuadro de diálogo que nos permite modificar los valores de los parámetros de configuración como pueden ser el número de reglas o el soporte y confianza mínimos (figura 27). Hemos dejado los valores por defecto: 1% de umbral de soporte, 90% de umbral de confianza y 10 reglas (mostrará sólo las 10 mejores reglas). Al pulsar el botón “Start” se ejecuta el algoritmo proporcionando los resultados en el cuadro de texto de la derecha (figura 28). Podemos observar que 8 de las 10 reglas tienen una confianza del 100% y dos del 96%, lo cual es un muy buen resultado. Además, se muestran los valores de otras métricas de calidad vistas en el tema de teoría, como *lift* o convicción.

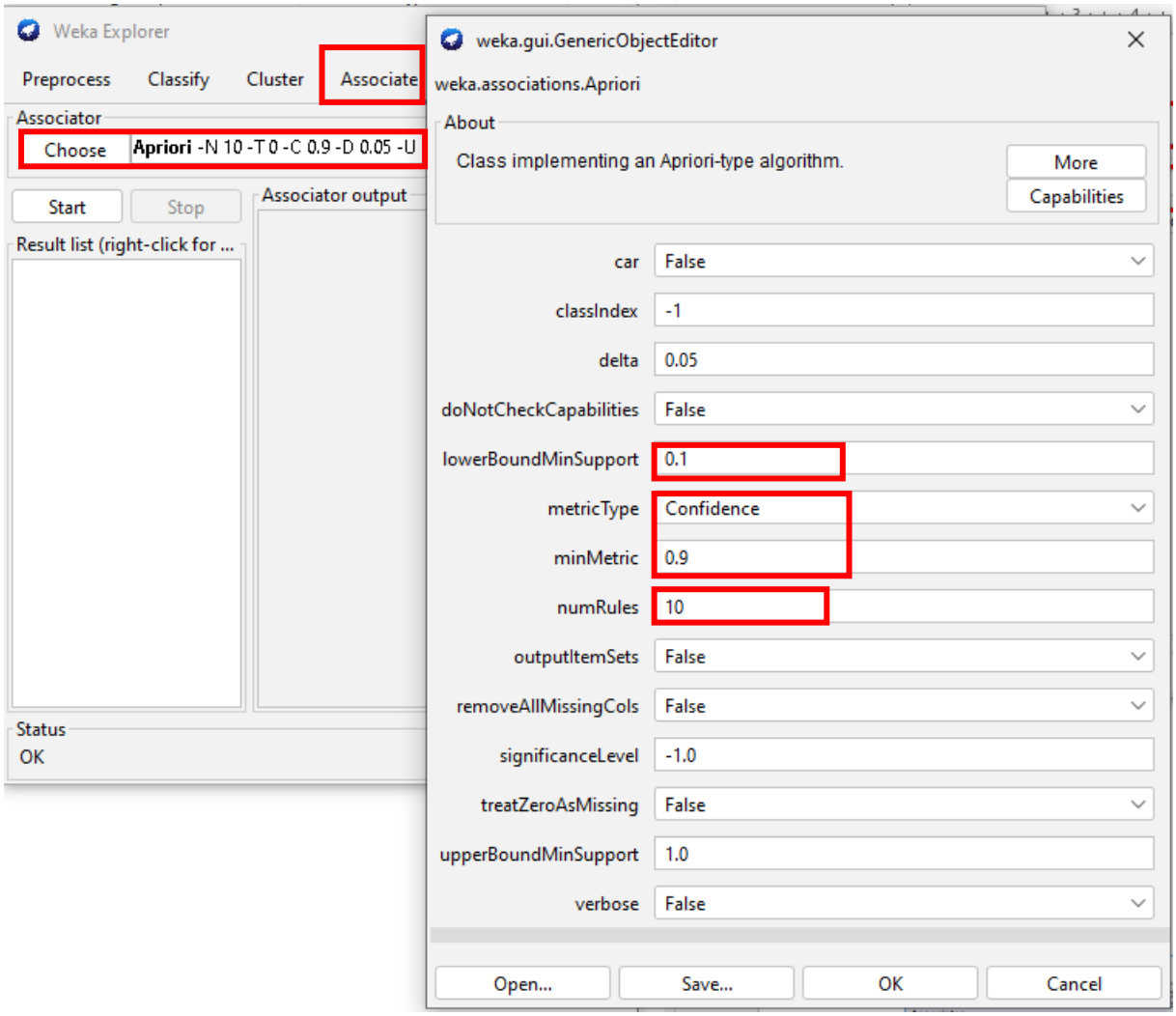

Figura 27. Configuración del algoritmo A priori

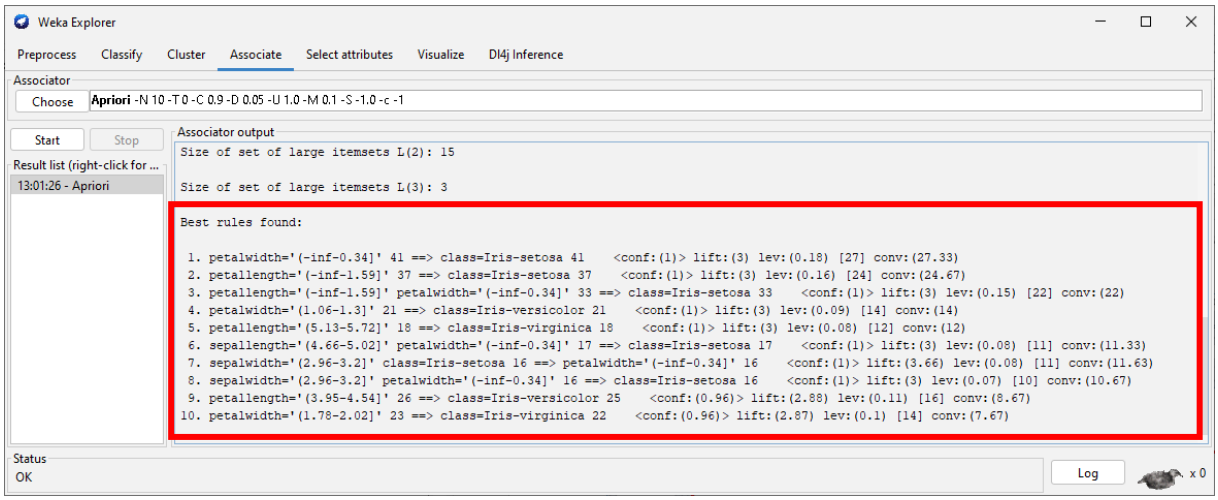

Figura 28. Reglas de asociación generadas con el algoritmo Apriori

4.2. *Clustering*

Seguidamente vamos a aplicar un algoritmo de agrupamiento a los datos "iris". Previamente volvemos a la ventana de preprocesamiento para volver a abrir el archivo original, ya que para aplicar las técnicas de *clustering* es mejor no tener los atributos discretizados.

Una vez que tenemos los datos originales seleccionamos la pestaña "Cluster". La selección y configuración del algoritmo se hace de la misma forma que en los casos anteriores. Hemos elegido el algoritmo "Simple K-means" que es uno de los algoritmos de agrupamiento más sencillo y mejor conocido (figura 29). Además, lo hemos configurado para que muestre la desviación estándar y para obtener tres agrupaciones y de esta forma intentar clasificar los registros en función de las tres clases del ejemplo. Hemos dejado la distancia euclídea, que es la que tiene por defecto. En la figura 30 se muestran los resultados de ejecución del algoritmo con los valores de media y desviación estándar de los centroides de cada uno de los tres *clusters*. Se puede observar también que cada *cluster* contiene 50 registros y en cada uno de ellos hay únicamente registros de una clase.

Podemos elegir la opción de visualización (figura 31) pulsando el botón derecho del ratón y obtenemos la representación de la figura 32 en la que se puede observar claramente los tres *clusters*. Podemos variar la representación cambiando las variables que se ven en el eje *x* e *y*, así como la asignación del color (por defecto son los *clusters*). Si para el eje *y* seleccionamos el atributo de clase podemos ver que hay una correspondencia total entre los *clusters* y las clases (figura 33).

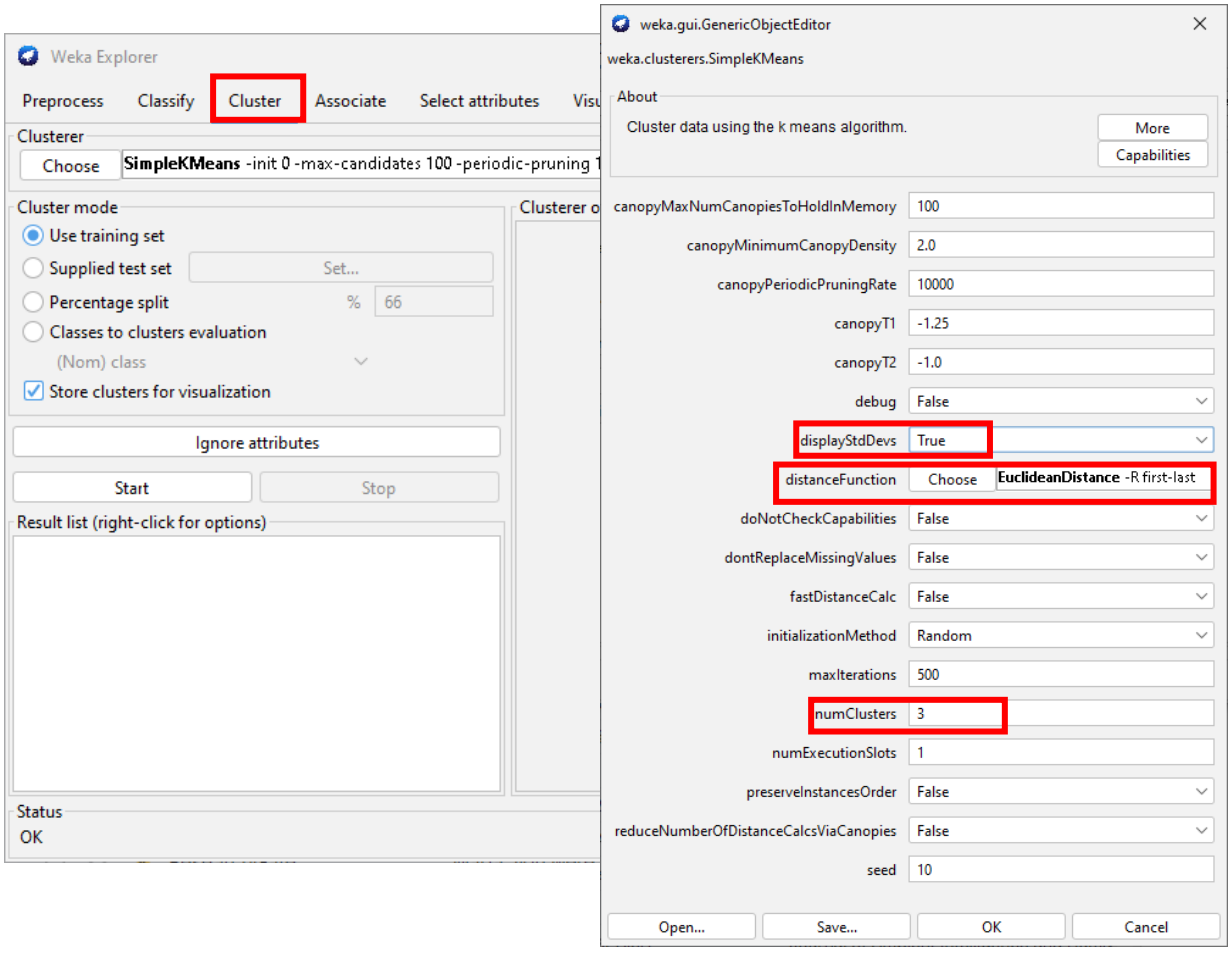

Figura 29. Configuración del algoritmo de *clustering* Simple k-Means

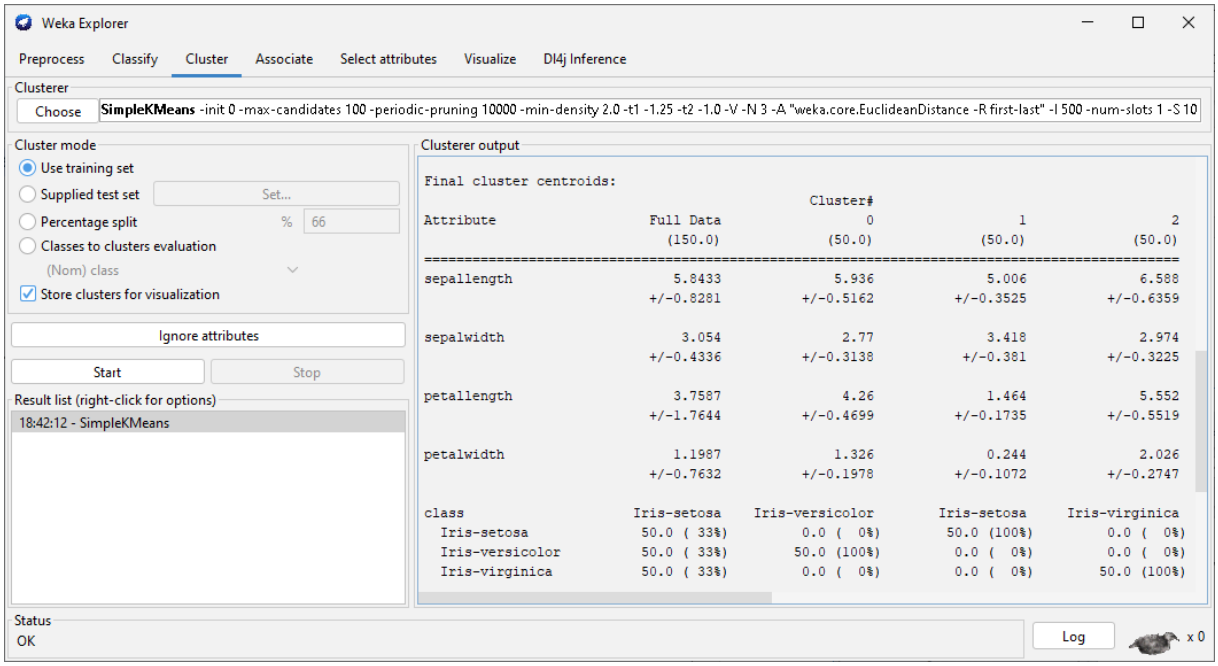

Figura 30. Resultados del algoritmo de *clustering*

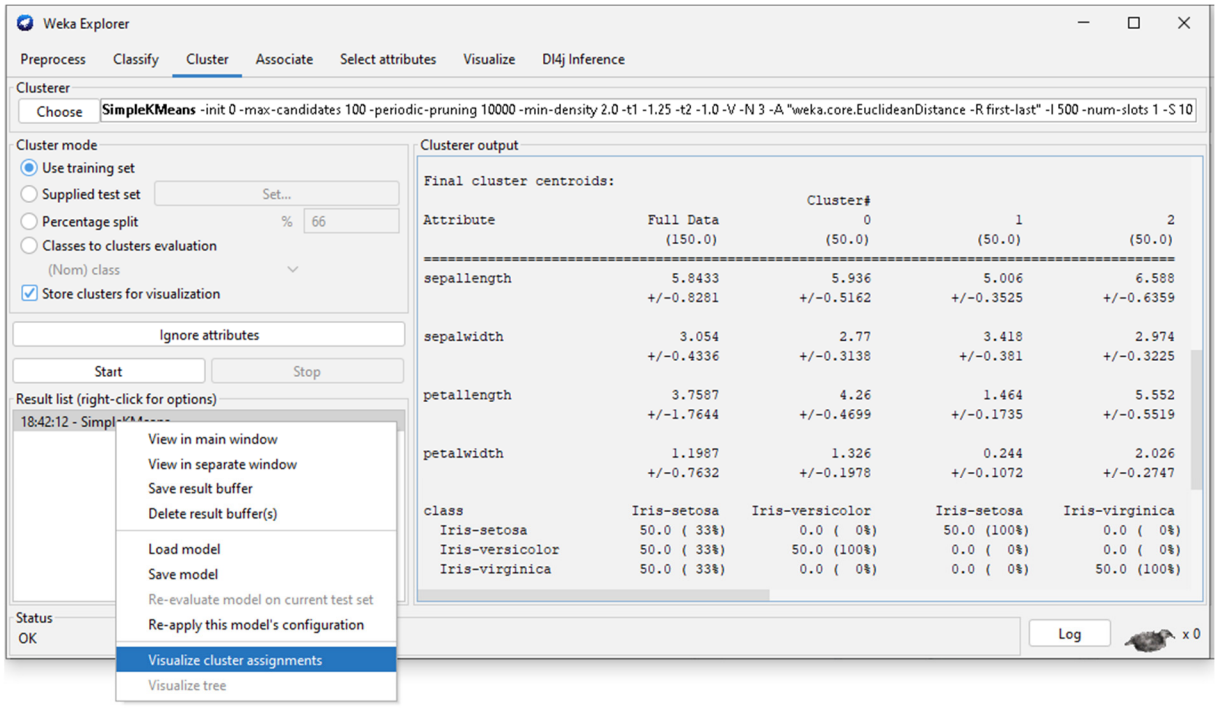

Figura 31. Selección de la opción de visualización de resultados

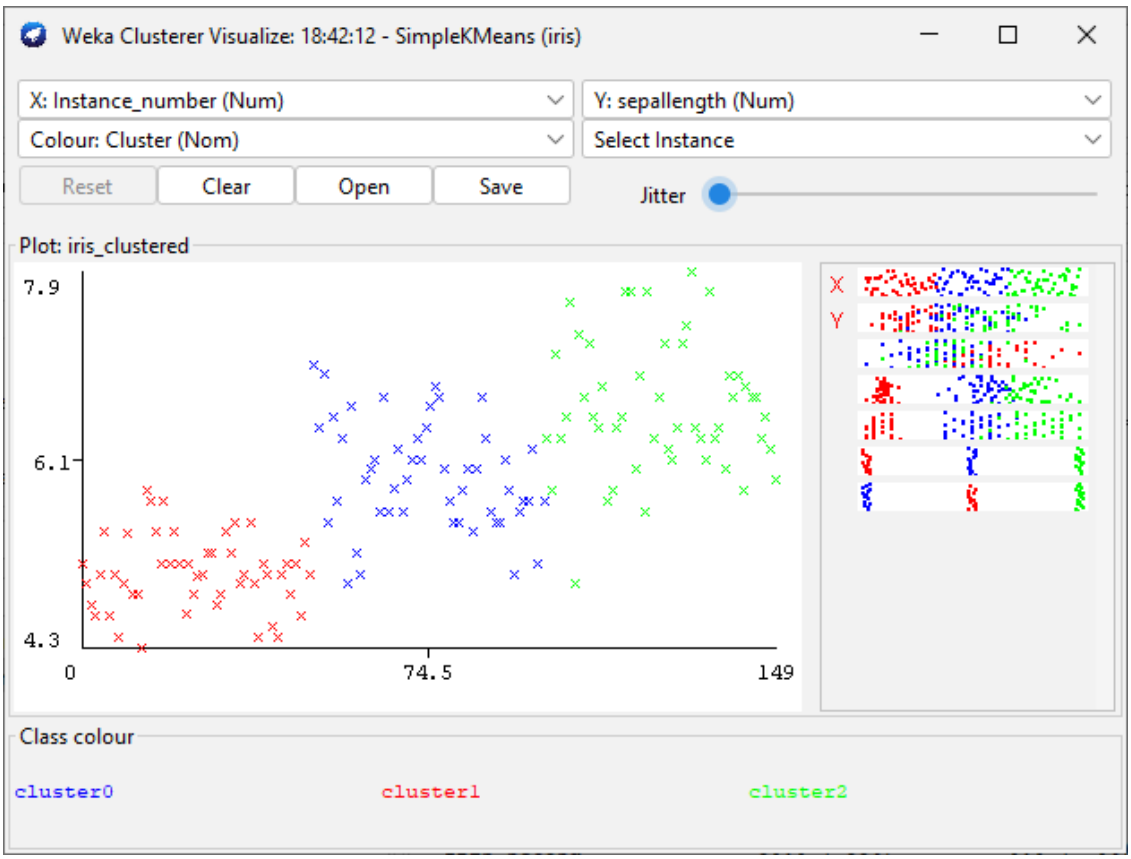

Figura 32. Visualización de los tres *clusters* obtenidos

Figura 33. Relación de los *clusters* con las clases