

Inteligencia en Sistemas Electrónicos

GUIÓN PRÁCTICA Introduccion a Weka y sistemas de aprendizaje 1.2

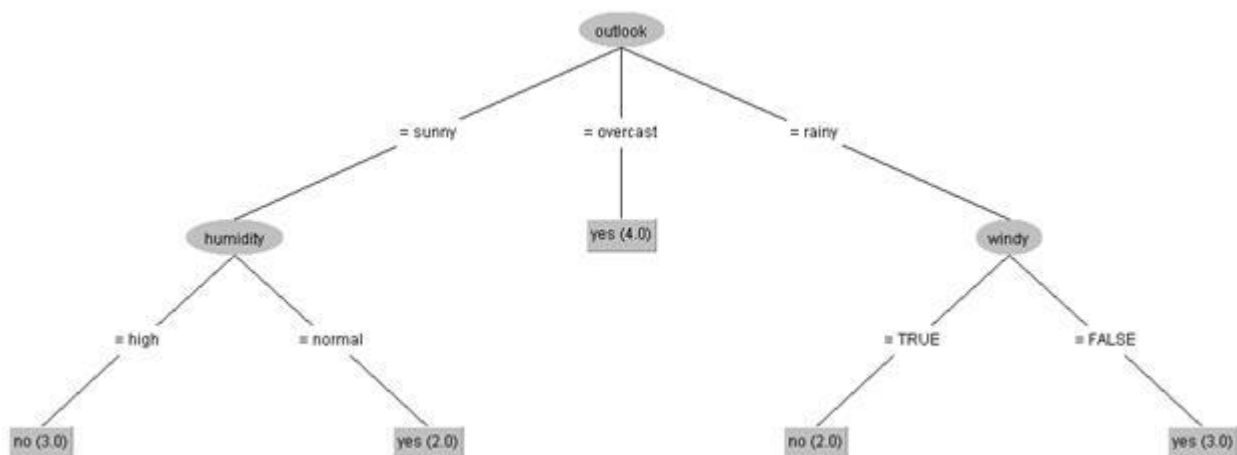
Alumno 1 _Jezael Pérez Herrera_____

Alumno 2 _Elena Cantero Molina_____

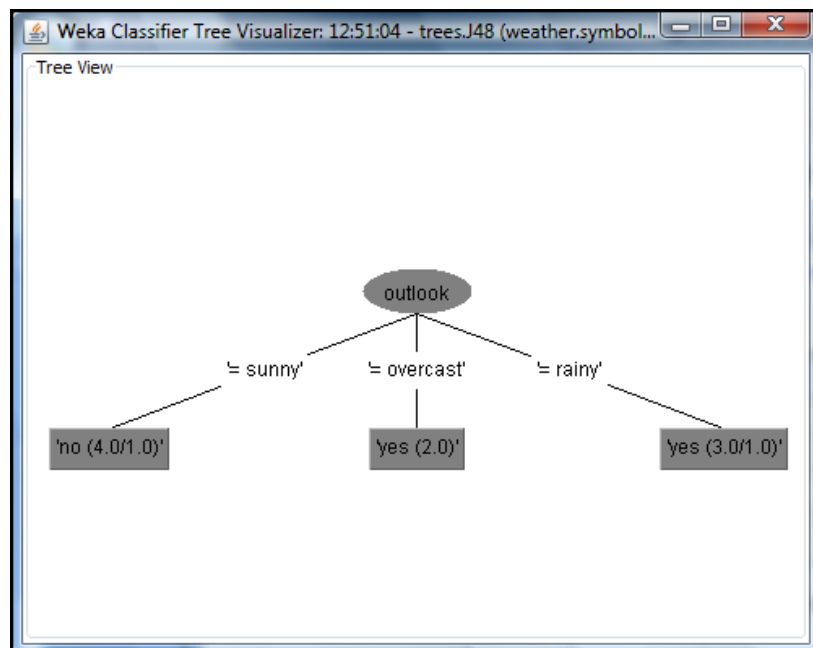
Grupo _INSE15__

Clasificación

1. Aplicaremos ahora un clasificador en árbol a los datos *weather.nominal.arff*. Carguemos esos datos con la pestaña *Openfile*. Después pasaremos a la pestaña *Classify*. Usaremos el clasificador en árbol J48 que es una versión evolucionada del que se ha visto en clase. Seleccione este clasificador en la pestaña *Choose-Trees-J48*. En la ventana aparece *J48 -C0.25 -M 2* que son los parámetros por defecto que utiliza. Éstos parámetros pueden cambiarse si clickamos en la ventana blanca donde aparece *J48 -C0.25 -M 2* pero por el momento lo dejaremos así. En “*Test options*” a la izquierda seleccionamos “*Use training set*” por lo que utilizaremos todo el conjunto para entrenamiento y todo el conjunto para test. Pulse ahora en el botón “*Start*”.
 - a. ¿Cuál es el porcentaje de acierto sobre estos datos? 100%
 - b. Pulse ahora con el ratón derecho en la ventana a la izquierda abajo en *trees.J48*. Aparecerá un nuevo menú. Pulse en *Visualize tree*. Vemos un diagrama del árbol creado y cómo el árbol clasifica cada uno de los ejemplos. Incluya abajo una captura de pantalla.



- c. Seleccione ahora sólo los primeros 9 ejemplos de la base de datos (cargando el fichero weather.nominal.primeros9.arff del Moodle- si no lo puede descargar, copie el fichero en el block de notas y guárdelo con extensión .arff) y aplique de nuevo el clasificador J48 y visualice el árbol que vemos a continuación.



¿Cuántos errores se producen ahora? Enumere el indicador de cada error.

2, el indicador de un error es el 6 y el del otro es el 9.

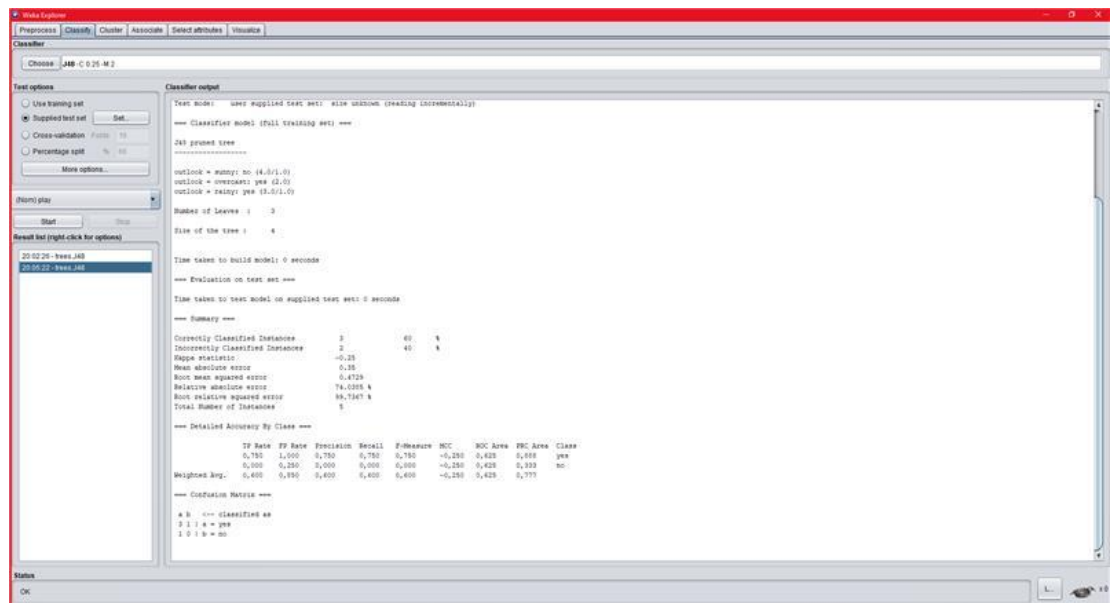
- d. Calcule ahora mirando este árbol los errores que se cometen al usar este árbol para clasificar los últimos 5 datos del conjunto. Para ello vaya al apartado 2 y vea uno por uno los resultados de clasificación para los 5 restantes.

- ¿Cuántos errores se producen? 2
- Enumere el indicador de cada uno según su índice.
El indicador de un error es el 11 y el del otro es el 14.

- e. El número de errores también puede calcularlo si en el recuadro Test options cambia a "Supplied test set". Elija ahora el fichero weather.nominal.ultimos5.arff del Moodle y utilícelo. Pulse Start e incluya una captura de pantalla de los resultados. Mire en la matriz de clasificación y compruebe si los errores son los mismos que encontró manualmente.

¿son los mismos? Diga el indicador de los mismos según su nuevo índice en el grupo de 5 de test

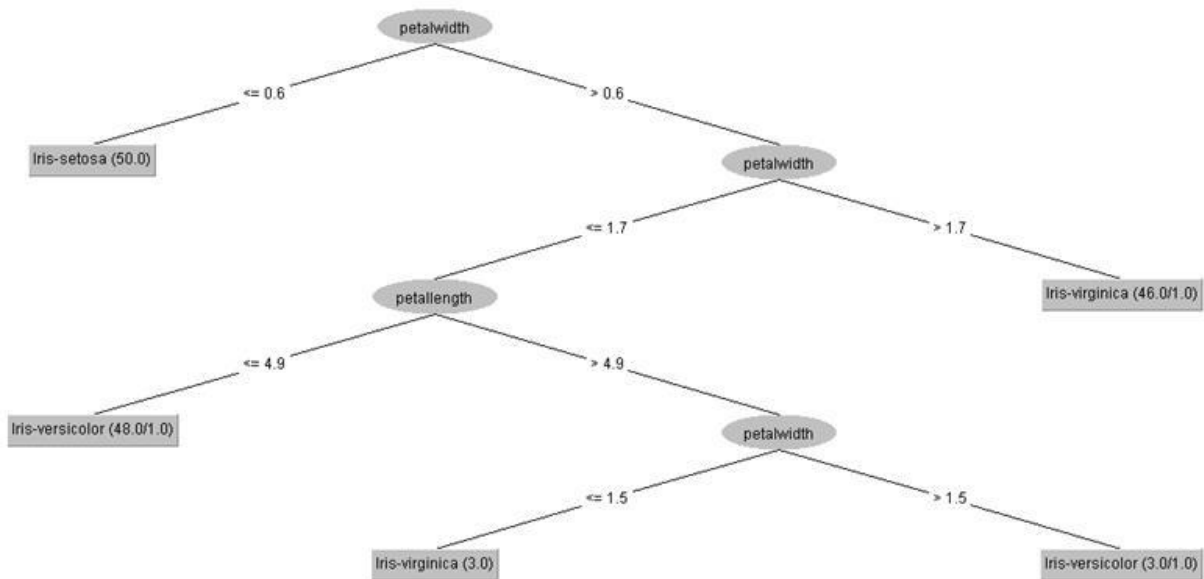
No, no son los mismos. En este caso son el 2 y el 5.



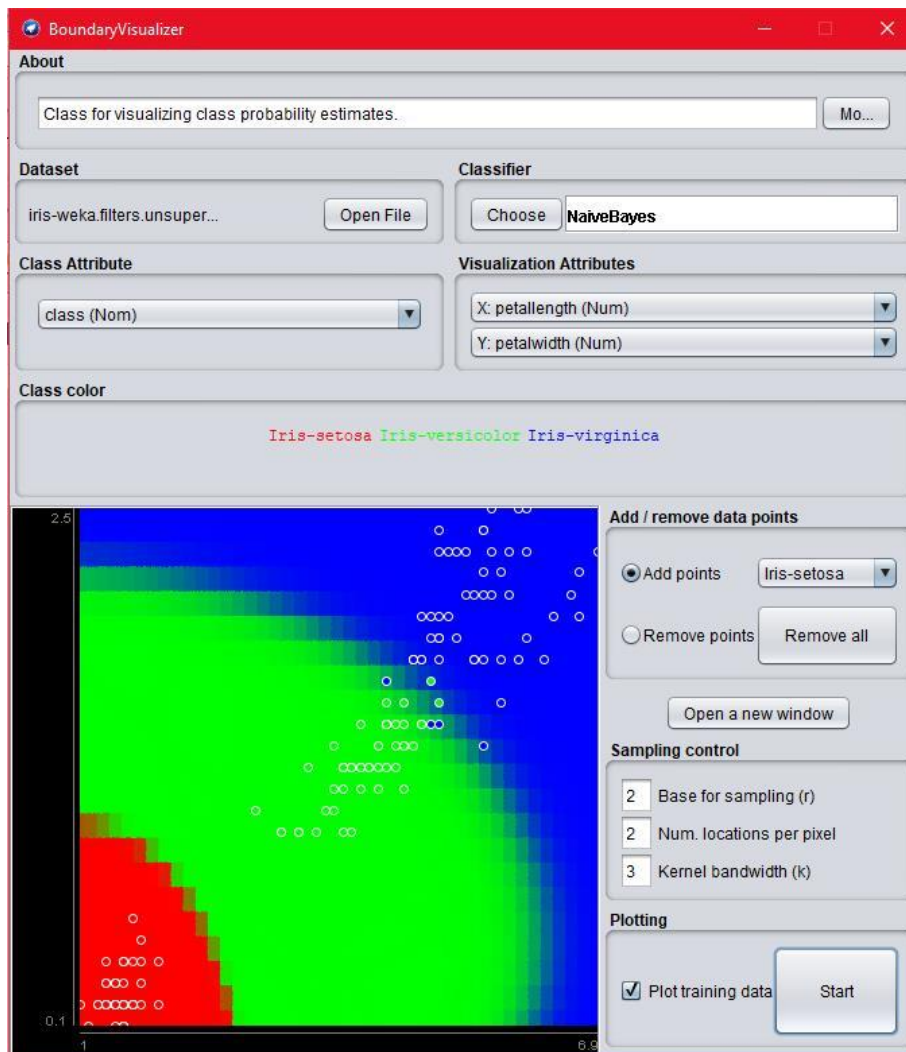
2. Vamos a modificar ahora un parámetro de la creación del árbol que es el número mínimo de objetos en cada rama final. Para ello entre en la ventana de características del árbol a crear (botón derecho en la ventana del clasificador, en “show properties” y ponga en minNumObj=1 y genere ahora el nuevo árbol. Incluya una captura de pantalla



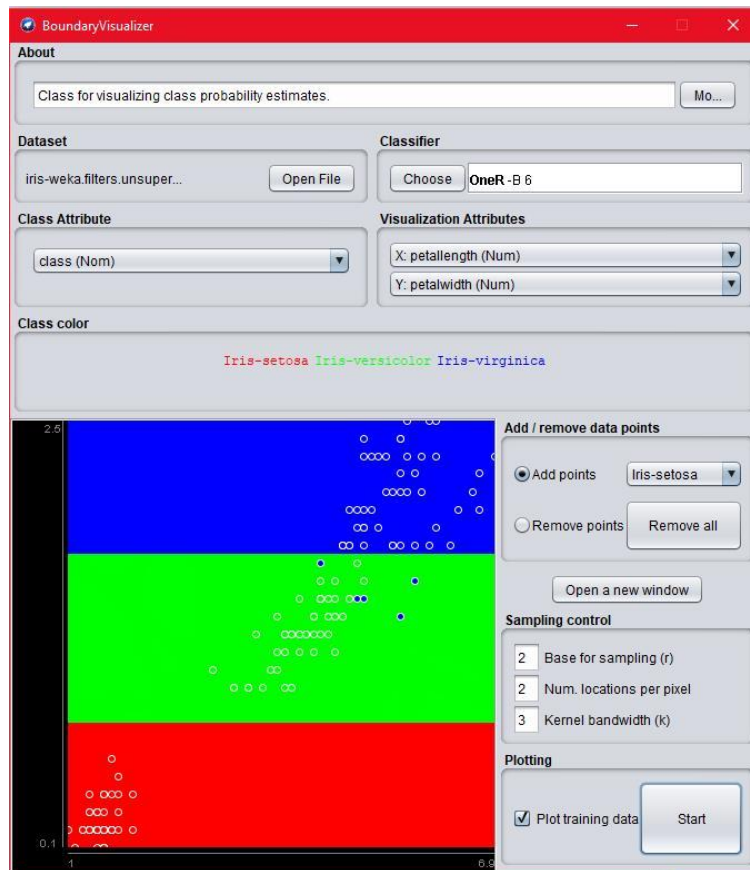
- a. Compare ahora este árbol con el original entrenado con los 14 casos y el entrenado con los 9 casos
 - i. ¿Cambia en algo? Es el mismo árbol con menos casos.
 - ii. Calcule la tasa sobre “Use training set”. 100%
 - iii. Razone si es bueno que en cada rama del árbol creado haya solo un elemento- como ahora- y que todos los ejemplos del training los clasifique correctamente o es mejor que en cada rama haya más de un elemento como en el caso anterior aunque no clasifique los datos de entrenamiento al 100%.
3. Cargue ahora los datos *iris.arff* y clasifíquelos con el árbol J48. Vuelva al valor por defecto del MinNumObj=2
 - a. ¿Cuál es la tasa de acierto para el test *Use training set*. 98%
 - b. Incluya abajo una captura del árbol



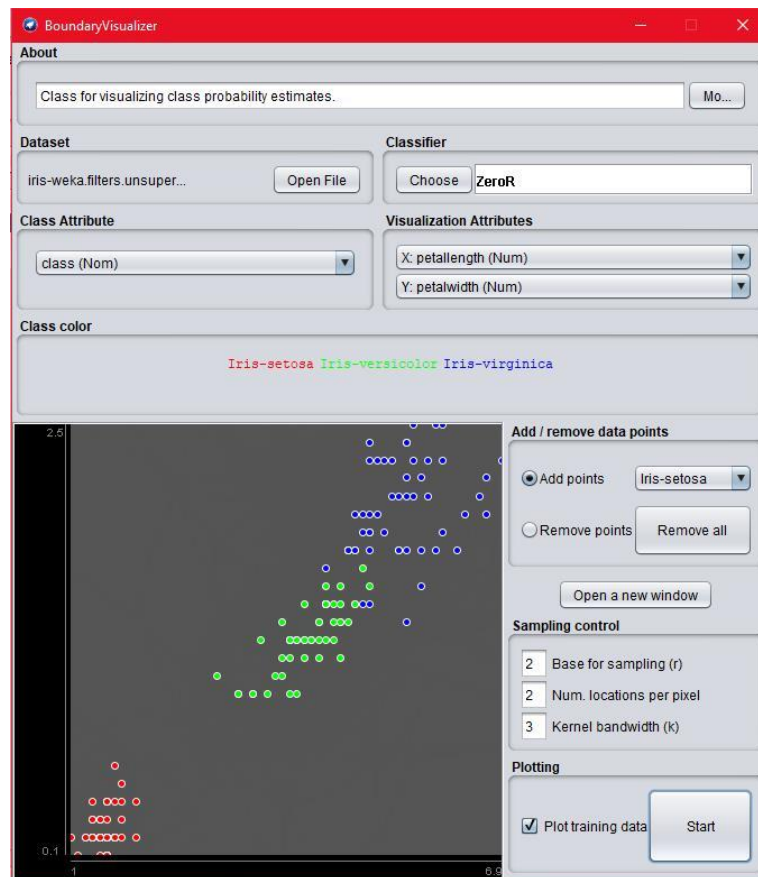
- c. Visualizando el árbol enumere en qué casos se producen los errores
 Un error se produce para petalwidth mayor que 1.7, iris-virginica.
 Otro error se produce para un petalwidth entre 0.6 y 1.7 y un petallenght menor o igual que 4.9, iris-versicolor.
 Y otro error se produce para un petalwidth entre 1.5 y 1.7 y un petallenght mayor que 4.9, iris-versicolor.
 - d. Calcule ahora la tasa de acierto con percentage Split del 66%
 96,0784%
 - e. Cambie ahora Ramdom Seed = 2 y calucle la tasa y explique el resultado
 98.0392% de tasa de acierto. Esto sucede porque se eligen los datos de manera aleatoria y distinta a tener Random Seed = 1
 - f. ¿Cuál sería la tasa que explicaría mejor su sistema, la de "use training set" la de Split 66% con Ramdom Seed =1 o la de Split 66% con Random Seed = 2
 La tasa que explicaría mejor el sistema es de "use training set", ya que tenemos tasa de acierto de un 98%. Aunque no sea la más alta, siendo la diferencia insignificante, se elige debido a que usa todos los datos.
4. Vaya ahora al panel de Visualización de la primera pantalla del programa y teclee *Visualization y Boundary Visualizer*. Open file *IRIS.2D.arff*. Choose classifier *Bayes -Naïve Bayes*, seleccione *Plot Training Data* y después pulse el botón *Start*.
 - a. Incluya una captura de pantalla



- b. ¿Visualiza los límites de las clases? ¿cuántos errores ha identificado?
Sí, hemos identificado 5 errores.
- c. Con los mismos datos, utilice ahora el clasificador *OneR*. Incluya una captura de pantalla con los resultados.

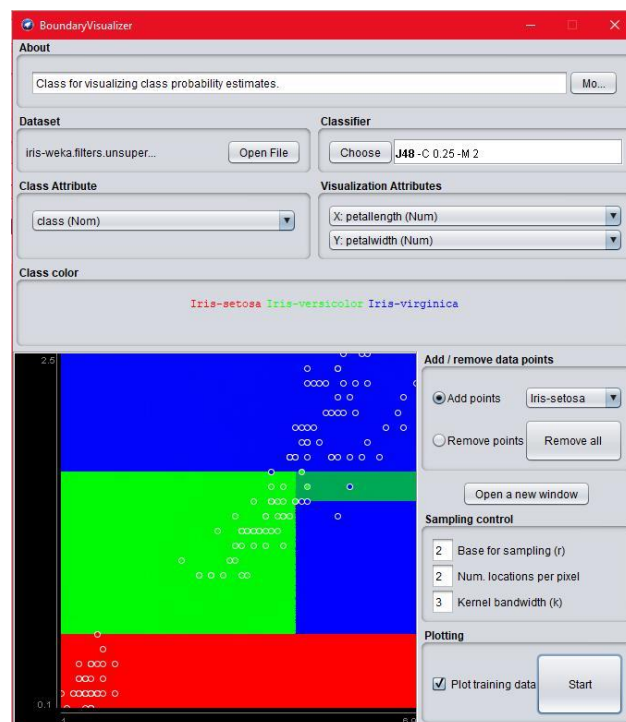


- i. ¿cuántos errores encuentra? Hay 5 errores
 - ii. Razone las diferencias entre los tipos de fronteras entre clases encontrados.
Las fronteras entre clases en esta última captura (1R) son tan marcadas (rectas) ya que sólo se contempla un atributo mientras que en el caso Bayes como se tienen en cuenta todos los atributos, las fronteras entre clases son más progresivas.
5. Repita la cuestión anterior usando ahora el clasificador OR
- a. Incluya una Captura de Pantalla

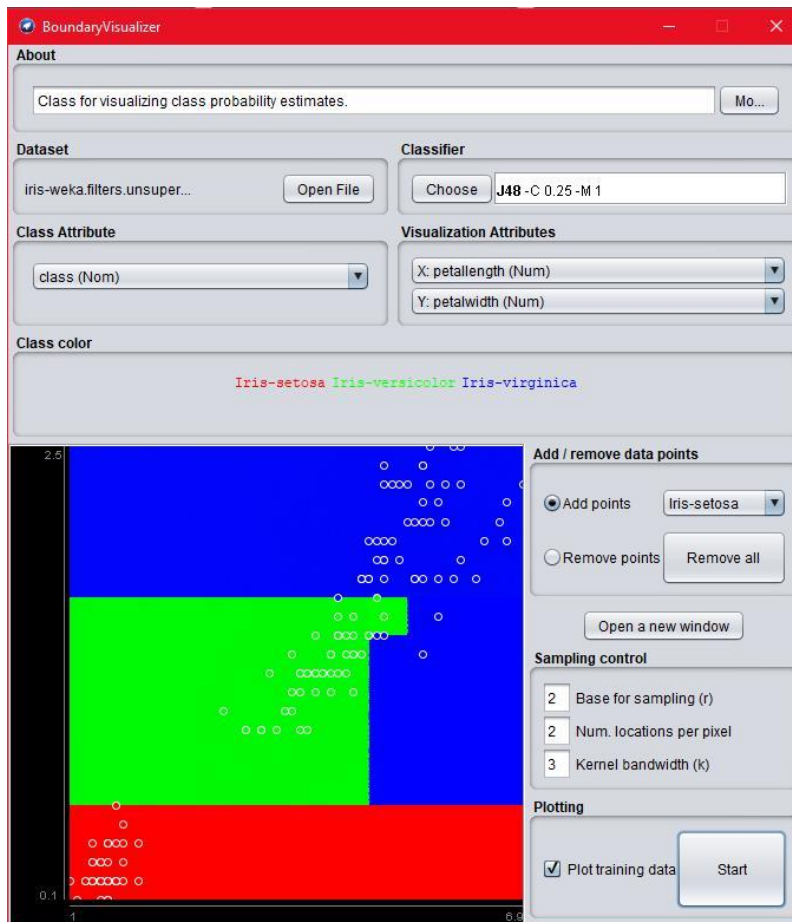


- b. Explique lo que ve en la Pantalla según el resultado esperado del algoritmo 0R.
En este caso no tenemos fronteras puesto que no se considera ningún atributo.

6. Repita la cuestión anterior usando ahora el clasificador árbol J48.



- a. Identifique las diferencias entre los tipos de fronteras entre clases encontrados, especialmente las que el color no es puro.
Las líneas horizontales delimitan el ancho del pétalo y las verticales delimitan el largo del pétalo. La zona donde el color no es puro corresponde con el final de la rama del árbol con tasa de error alta.
- b. Cambie ahora en las propiedades del árbol el MinNumObj a 1 y haga una captura de pantalla



- c. Explique qué sucede y relaciónelo con lo obtenido en apartados anteriores
Como sólo hay un elemento por rama, la zona de error disminuye.
- d. Compare las fronteras encontradas con las encontradas con los métodos 1R y Naive Bayes
Las fronteras son rectas ya que las clasificaciones son discretas. Sin embargo, el 1R usa todos los atributos y eso hace que en vez de franjas sean secciones.

7. Regresión lineal

- a. Abra el fichero de datos cpu.arff y aplique una regresión lineal.
 - i. Diga la fórmula de regresión que se obtiene

$$0.0491 * MYCT + 0.0152 * MMIN + 0.0056 * MMAX + 0.6298 * CACH + 1.4599 * CHMAX + -56.075.$$
 - ii. Ponga el error cuadrático medio que se obtiene en la predicción
58.9899

Haga un pdf y súbalo a la tarea de Moodle.