**Introducción a la ciencia de datos**

**PRÁCTICA**

Pregunta 1

Enunciado

1. A lo largo de la clase hemos visto más de una definición de diversos términos relacionados con la ciencia de datos. Quizás la más importante es la inicial, la definición de dato. Datos hay de muchos tipos, pero a nivel de procesamiento los podemos clasificar en simples, estructurados, semiestructurados o no estructurados. Valorar en los siguientes casos de qué tipo de dato se trata y justifica tu respuesta (recuerda que, a menudo, el tipo también puede depender de la interpretación o uso que se haga del mismo).

* La lista de notas de la asignatura, con nombres y calificaciones.

Dato estructurado, pues estarían en una lista o tabla.

* Un fichero en PDF.

Dato no estructurado, pues serían muchos tipos de datos sin ninguna estructura.

* Vuestro DNI.

Dato compuesto, pues sería la combinación de datos simples (números y una letra)

* Un poema.

Dato semiestructurado, no lo tengo muy claro aquí.

* El saldo de vuestra cuenta corriente.

Dato simple, pues es un número

Pregunta 2

Enunciado

Los datos «abiertos» (u *Open Data*) se han popularizado en los últimos años y permiten a cualquier persona acceder a conjuntos de datos públicos para analizarlos. Para saber algo más podéis dirigiros al siguiente enlace, que contiene además varios casos de uso y ejemplos.

<https://www.paradigmadigital.com/techbiz/open-data-los-datos-estan-cambiando-mundo/>

Los casos de uso que habéis leído se centran en la transparencia, la agricultura, el medio ambiente, la política, la ciencia y el periodismo. Escoged uno de los ámbitos y su caso (por ejemplo, si os gusta el Open Data en las ciudades, el ejemplo que cita el artículo es la EMT) y responded a las siguientes cuestiones:

He elegido el portal <https://www.globalforestwatch.org/>

* Describe el portal, los contenidos y servicios que ofrece.

Global Forest Watch es una página web que ofrece datos sobre la evolución de las masas forestales en el tiempo. También ofrece tecnología y herramientas. Permite monitorizar los bosques. Tiene un mapa super interesante dónde se pueden controlar y analizar un montón de variables.

* ¿De qué forma se puede acceder a los datos abiertos ofrecidos por el portal?

Se puede acceder de manera online mediante un mapa interactivo en el que se pueden controlar y analizar un montón de variables. Además tiene un dashboard dónde se analizan datos. También permite filtrar por temas. Y descargar los datos en formato Excel.

* ¿Qué tipo o tipos de datos podemos encontrar en él?

Todo tipo de datos.

Ahora vamos a ir un poco más al detalle. Elegid un único conjunto de datos que exista en el portal elegido (por ejemplo, los datos de uso de BICIMAD) y responded a lo siguiente:

* Describe brevemente el conjunto de datos elegido.

He elegido los datos de “Alertas de deforestación (GLAD)”. Dónde prevalecen en Indonesia, Malasia, Sudáfrica y Sudamérica. Son datos desde el 2015 hasta el 2020.

* ¿A qué tipo de preguntas podríais dar respuesta con el conjunto elegido?

Qué zonas están deforestándose, dónde podríamos plantar, qué terrenos presentan problemas, etc.

* ¿Qué datos se te ocurren que podrías coger de algún otro sitio o repositorio para combinarlos con los tuyos?

Los datos forestales del Ministerio de Agricultura ( <https://www.mapa.gob.es/es/desarrollo-rural/temas/politica-forestal/inventario-cartografia/inventario-forestal-nacional/default.aspx> ). Datos de incendios forestales de epdata (<https://www.epdata.es/datos/incendios-forestales-datos-estadisticas-cifras/267>)

Pregunta 3

Enunciado

Actualmente existen muchos lenguajes de programación estadística que podemos utilizar para realizar tareas analíticas en proyectos de ciencia de datos, entre los que destacan **R** o **Python**. Estos lenguajes o entornos de programación estadística tienen en común la capacidad de implementar algoritmos de aprendizaje automático sobre los datos a estudiar, permitiendo obtener rápidamente resultados que de otro modo sería muy costoso de conseguir.

Sin embargo, es conveniente que antes de ponernos a utilizar estas herramientas, dediquemos tiempo a conocer la materia prima de la que disponemos; los datos. Conocer los datos, en la fase inicial de todo proyecto de ciencia de datos, es imprescindible para posteriormente poder interpretar correctamente las diferentes soluciones obtenidas con el uso de algoritmos, y finalmente realizar ajustes de precisión en nuestros proyectos.

En este ejercicio, vamos a utilizar una herramienta que nos permitirá realizar esta fase inicial de reconocimiento de los datos, y posteriormente también nos permitirá implementar algoritmos de *machine* *learning* realizando así, un experimento completo de ciencia de datos. La herramienta en particular que vamos utilizar es **Weka**.

Para ello deberéis seguir el **tutorial de Weka adjunto**, deberéis responder solamente a los siguientes apartados del tutorial:

**0.Instalación de Weka**: En este apartado bastará con instalar la aplicación.

**1.Visualización de un dataset**: En este apartado deberéis responder a las preguntas 1,2,3,4 y 5 del tutorial.

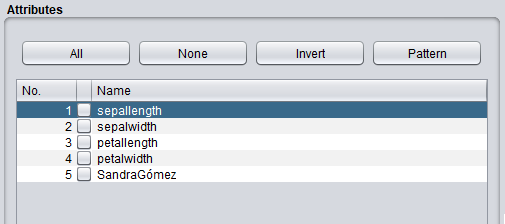
**Pregunta 1 - Investiga el link y explica los datos que contiene y a qué tipo de experimentos puede dar respuestas:**

Iris Data Set. Es un data Set que tiene 3 clases de 50 instancias (150 instancias) y 5 atributos del área “life”. Cada clase se refiere a un tipo de planta Iris..

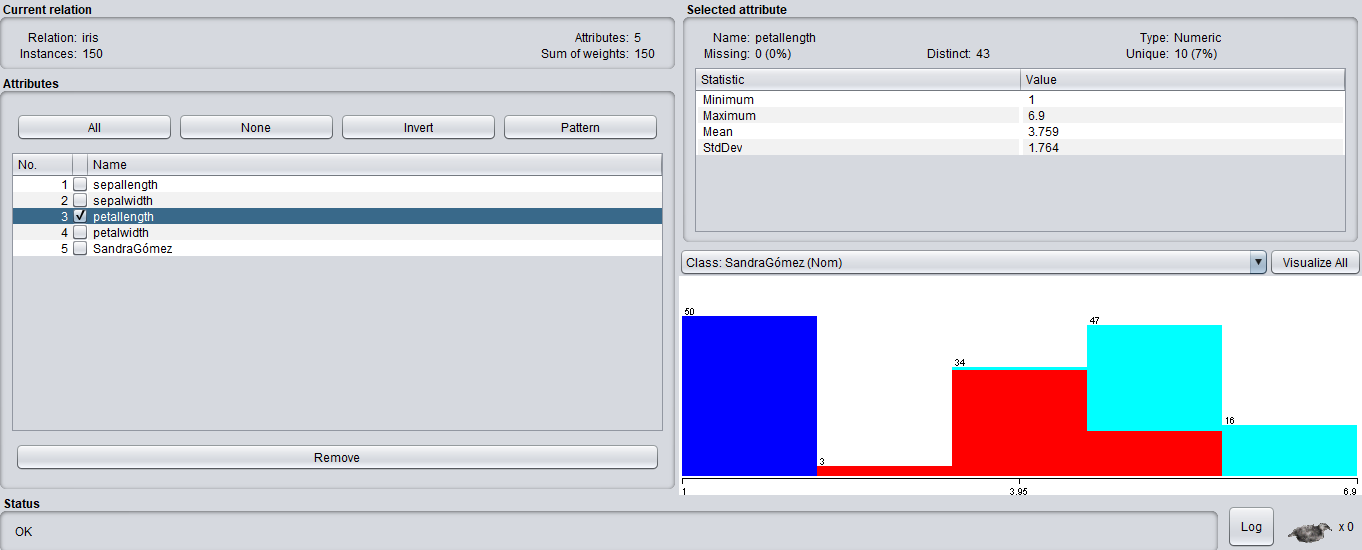
Los datos nos dan la información de la longitud y ancho de los sépalos en cm, la longitud y ancho de los pétalos en cm, y la clase ( Iris Setosa, Iris Versicolour, Iris Virginica).

Con estos datos podemos ver diferencias y similitudes entre clases diferentes de plantas Iris. Con estos datos si los asociamos a insectos podríamos ver cómo es el tipo de planta iris que los insectos prefieren.

**Pregunta 2 - Realiza los pasos anteriores, y renombra el atributo class con tu nombre y primer apellido. Haz una captura de pantalla del explorador una vez modificado el nombre del atributo (como la pantalla anterior).**



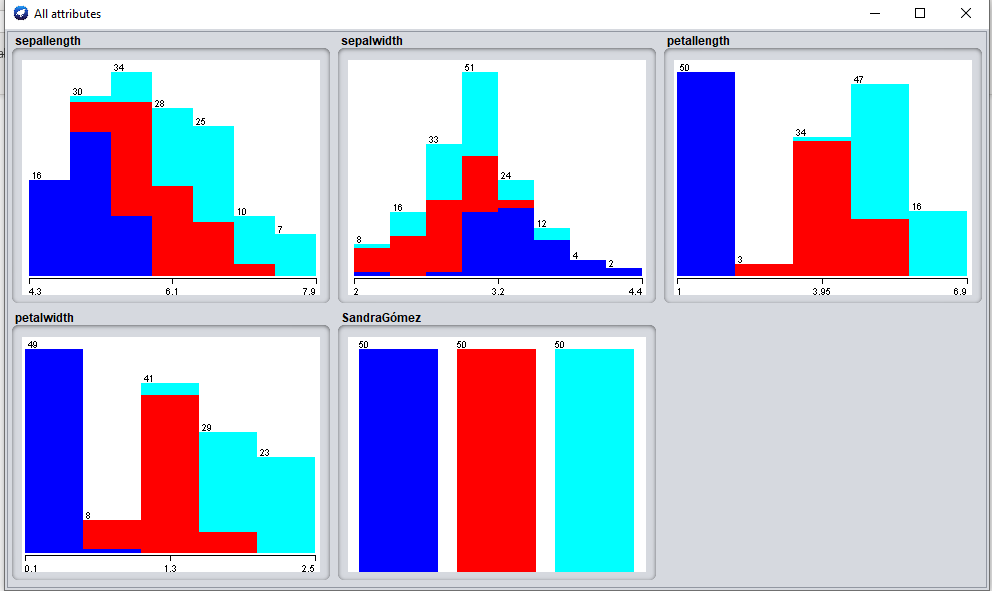
**Pregunta 3 - A continuación, selecciona el atributo petallength haz una captura de pantalla del explorador de WEKA, e igual que se ha realizado anteriormente, haz un análisis del atributo, realizando una descripción de los valores estadísticos y del histograma. ¿Qué información relevante podemos extraer del análisis?**



* El rango de valores está comprendido en el intervalo amplio [1, 6.9], la media es de 3.759, y la desviación típica es de 1.764.
* Todas las instancias están informadas, pues missing está al 0 % (No hay datos con el valor del atributo sin informar). De los 150 valores, 43 son no nulos distintos y solo hay 10 instancias (un 7%) con un único valor del atributo.
* En el histograma observamos:
  + Todos los valores comprendidos entre [1, 2.18] son de la clase iris Setosa y que sus 50 muestras comprenden ese tamaño.
  + El total de las instancias, 3, entre el (2.18, 3.36] son Iris-versicolor.
  + Entre el (3.36, 4.54](4.54, 5.72] encontramos 34 muestras de Iris Versicolor y Iris Virginica. Dónde en el segundo intervalo crece la predominancia de Iris Virginica.
  + Del (5.72,6.9] encontramos 16 muestras de Iris Virginica

Con esto podemos extraer que de las plantas Iris las que tienen una longitud de pétalo más pequeña es la Iris Setosa, y la más grande Iris Virginica, mientras que en el medio se sitúa la Iris Versicolor.

**Pregunta 4 - Tras una visualización del conjunto de histogramas anterior, ¿qué información relevante destacarías? Puedes aportar capturas para argumentar tu razonamiento.**



Azúl: Iris Setosa. Azúl clarito: Iris Virginica. Roja: Iris Versicolor.

* Observamos cómo las plantas Iris Virginica tienen los sépalos y pétalos más largos de los tres, sus pétalos tienen un ancho grande pero sus sépalos tienen un ancho medio.
* De las Iris Setosa observamos cómo son las plantas con una longitud de sépalos y pétalos menores, con una anchura de pétalo menor a las demás, pero con una anchura se sépalo MAYOR a las demás.
* De las Iris Versicolor observamos cómo suelen tener siempre valores intermedios en comparación a las otras dos clases. Tienen una longitud de sépalos y pétalos entre medias de las dos clases. La anchura del sépalo es menor a la Iris Setosa y parecida a la Iris Virginica. Mientras que la anchura del pétalo también se encuentra entre ambas clases.

**Pregunta 5 - Realiza una inspección visual a la matriz de gráficos de 5x5, y de entre todos escoge aquel cuyos atributos creas que clasifican mejor las instancias. Explica el histograma seleccionado y haz una exploración de las zonas donde se acumulan instancias de distintas clases y analiza alguna instancia (como hemos hecho anteriormente). Haz capturas de pantalla de los aspectos relevantes del proceso.**

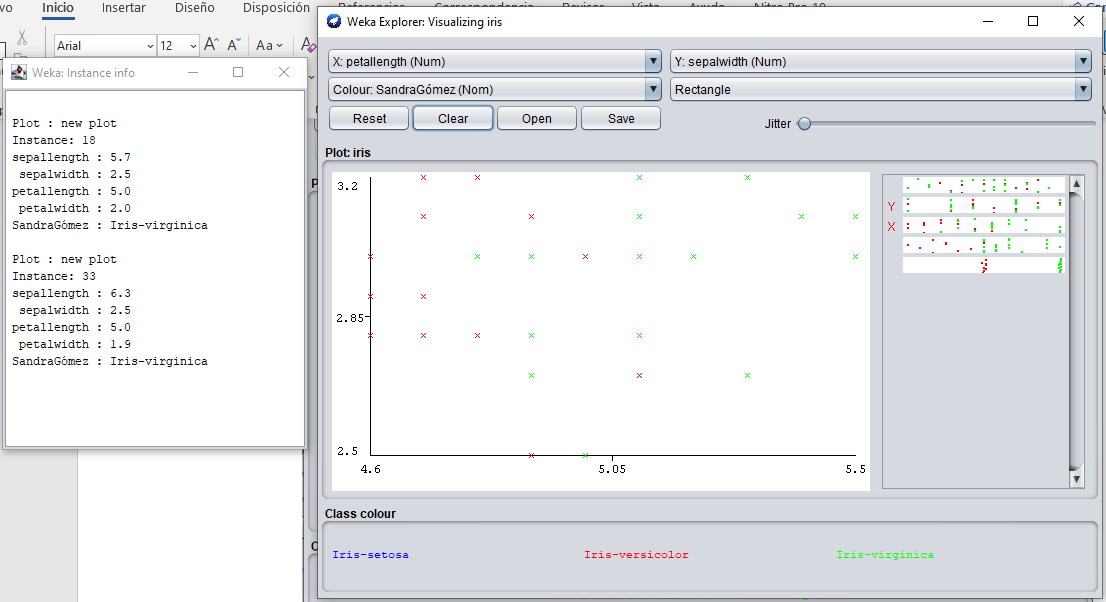
**IMPORTANTE Recuerda renombrar la clase (atributo class) con tu nombre y primer apellido (NombreApellido1) de manera que este aparezca en todas las capturas de pantalla que adjuntes.**

He elegido petallenght (X) con sepalwidth (Y)



Observamos como los valores de Iris Setosa se concentran al principio de petallength pero se encuentran a lo largo de todo el sepalwidth. Los valores de Iris Versicolor se concentran en la mitad del petallength y tienen valores desde el principio hasta la mitad del totalwidth. Iris Versicolor comparte valores medios con Iris Virginica. Iris Virginica la encontramos al final del petallength y hasta la mitad del totalwidth.

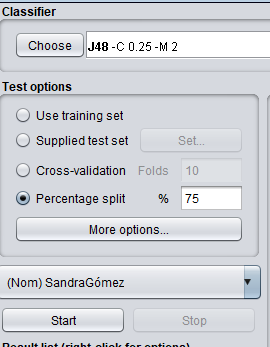
Si hacemos una ampliación rectangular dónde los valores de Iris Versicolor y Iris Virginica están más concentrados, encontramos instancias que comparten mismos valores:



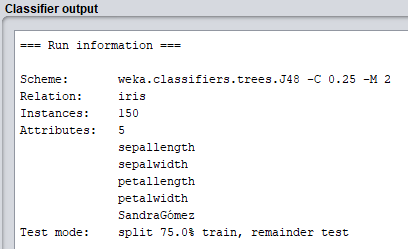
Aquí observamos cómo la instancia 18 y la 33 comparten mismo (x,y). Aunque son Instancias del mismo tipo de planta.

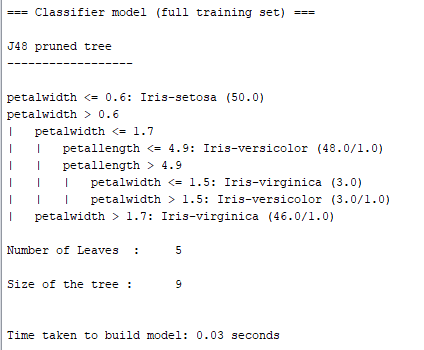
**2.Algoritmo supervisado J48:** En este apartado deberéis responder a la pregunta 6 del tutorial.

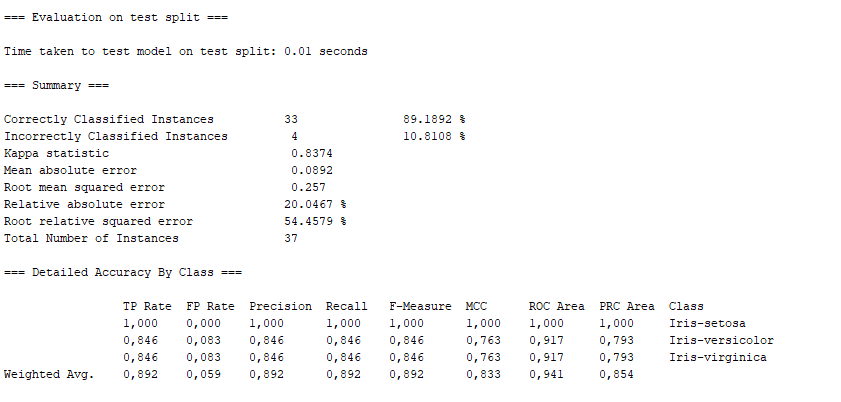
**Pregunta 6 - Realiza el mismo experimento, pero cambiando el porcentaje de instancias del conjunto de entrenamiento (Percentage Split). Esta vez define un 75% para el entrenamiento del modelo, dejando el 25% restante para el test. Repite los pasos anteriores y analiza los resultados obtenidos destacando la información más relevante y haciendo capturas de las pantallas relevantes para tu argumentación. ¿Qué diferencias observas respecto al anterior experimento?**

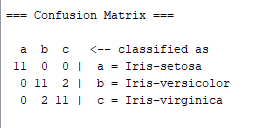


Los datos obtenidos con 75% de entrenamiento y 25% de test son:





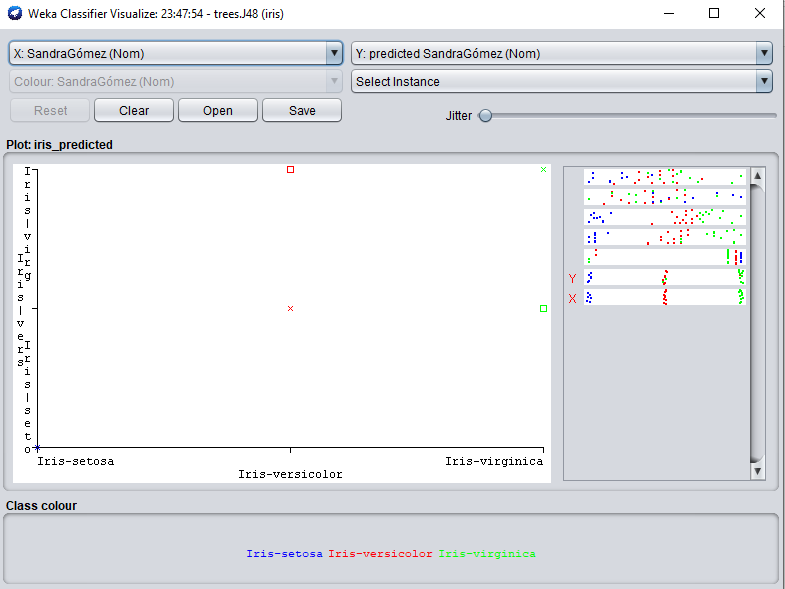




* En Summary podemos ver:
* Se han clasificado correctamente 33 de 37 instancias, lo que supone un acierto del 89.1892%. 4 no se han clasificado correctamente, por tanto un 10.8108% de error.
* El valor de “Kappa Statistic” es 0.8374, como es cercano a 1 se trata de un buen clasificador.
* Errores:: No los analizamos ya que son más utilizados para precisión numérica que para clasificación.
* Detailed accuracy by class / Precisión por clase:
* TP Rate (positivos correctos, instancias clasificadas correctamente): Iris Setosa al 100%, Iris Versicolor e Iris Virginica al 84%
* FP Rate ( falsos positivos, instancias clasificadas incorrectamente): Iris Setosa al 0%, Iris Versicolor e Iris Virginica al 8%
* Precision ( instancias correctas / nº de instancias clasificadas): Iris Setosa al 100%, Iris Versicolor e Iris Virginica al 84%
* Recall( nº instancias clasificadas/ total de instancias), F-Measure(combina Precision y Recall): Igual que Precision
* ROC área (equilibrio del data set. Clasificador óptimo si tiene valores cercanos a 1, clasifica al azar si tiene valores cercanos a 0.5): Iris Setosa al 100%, Iris Versicolor e Iris Virginica al 91%. Lo que indica que el clasificador es óptimo, es bastante bueno.
* Matriz de confusión:

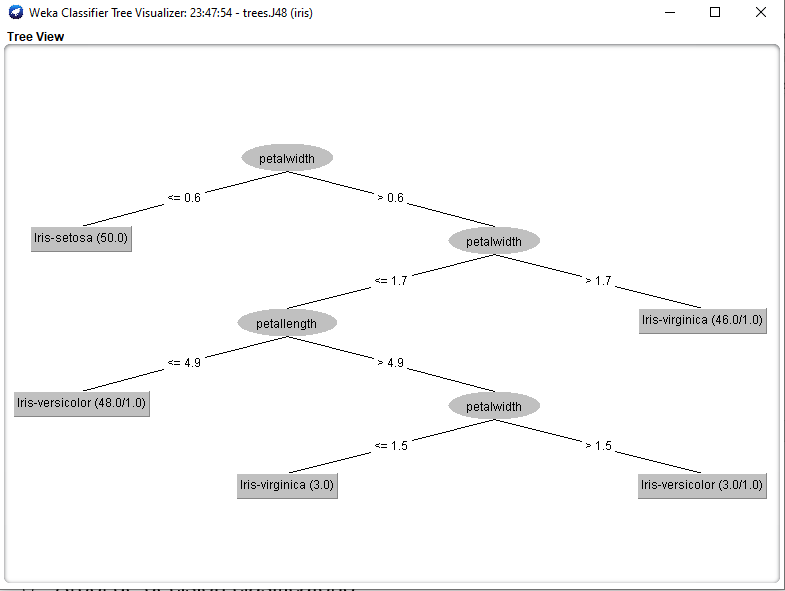
Vemos como 11 instancias de Iris Setosa se han clasificado correctamente, 11 instancias de Iris Versicolor se han clasificado correctamente y 2 instancias de Iris Versicolor se han clasificado incorrectamente como Iris Virginica, 11 instancias de Iris Virginica se han clasificado correctamente y dos instancias de Iris Virginica se han clasificado como Iris Versicolor.

* Errores de clasificación:



(Instancias mal clasificadas marcadas con un cuadro)

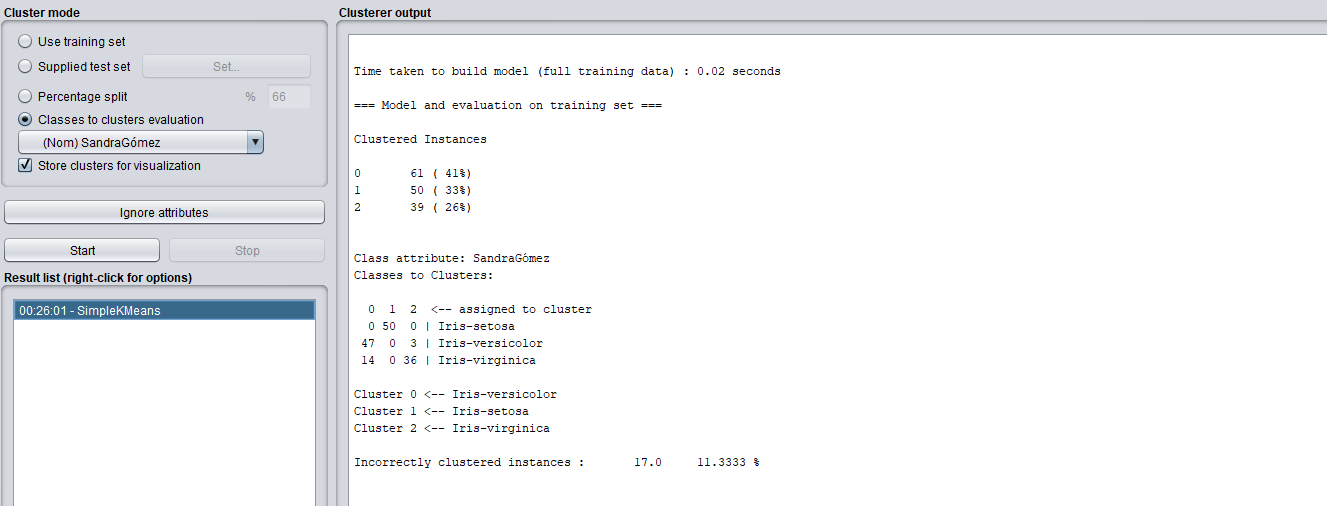
* Árbol de decisión clasificatorio:



* En comparación al otro ejemplo observamos como el acierto es menor, es decir menos muestras se han clasificado correctamente, ambas han utilizado un buen clasificador, en este caso ha habido más muestras mal clasificadas de Iris Versicolor e Iris Virginica.

**3.Algoritmo no supervisado Kmeans:** En este apartado deberéis responder a la pregunta 7 y 8 del tutorial.

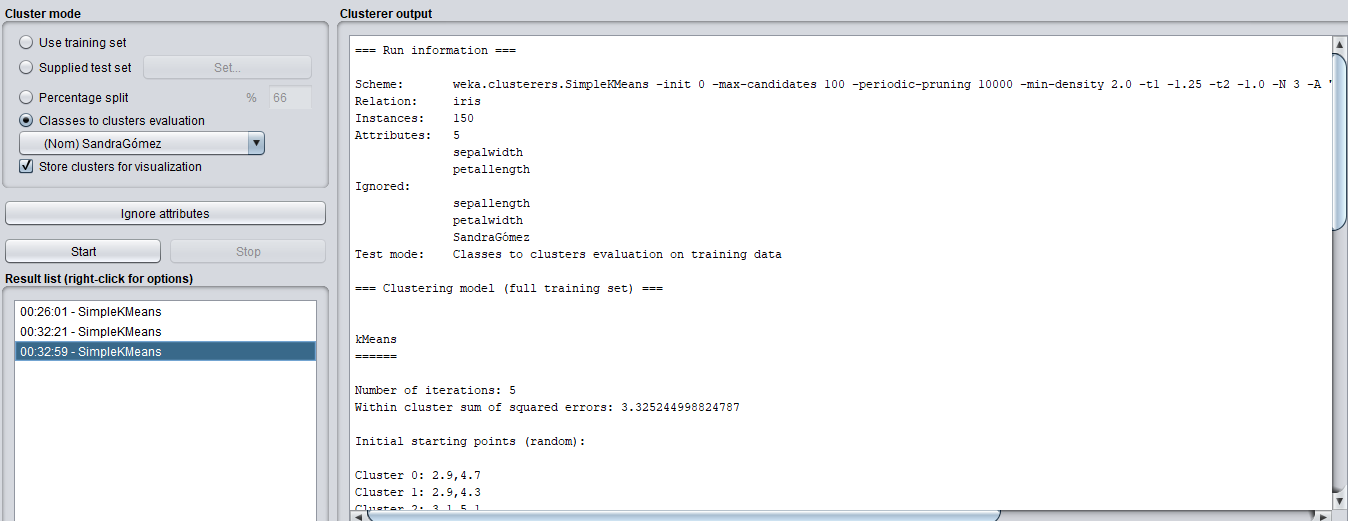
**Pregunta 7 - Realiza el mismo ejercicio anterior y captura la última pantalla del proceso(imagen superior), de manera que se pueda ver tu nombre y apellido en la clase para evaluar los clusters y el número final de instancias incorrectamente agrupadas**

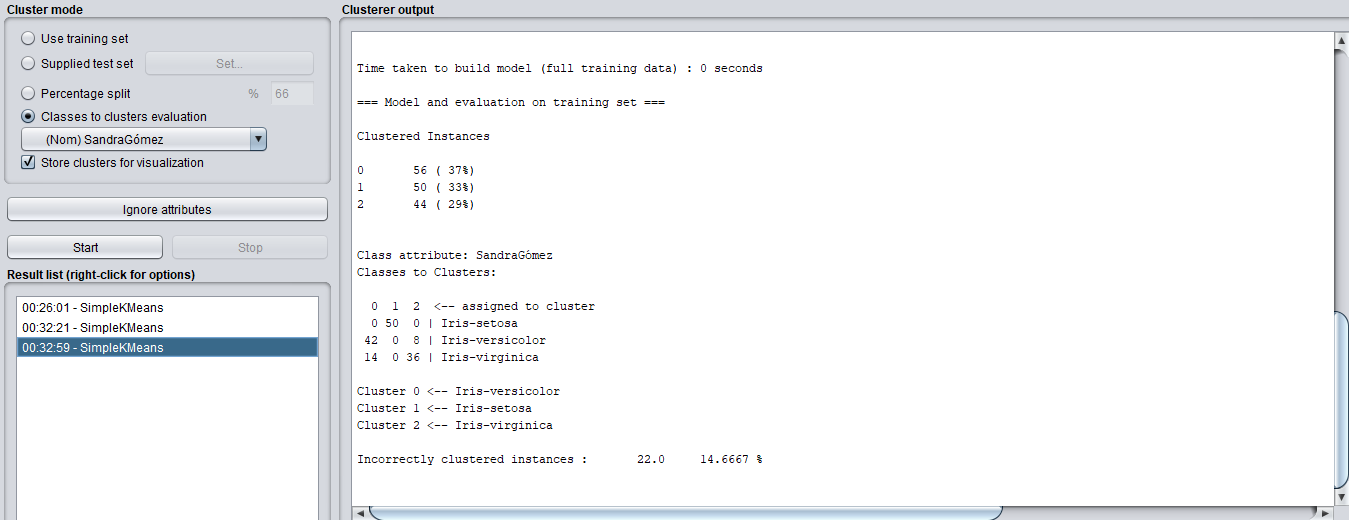


**Pregunta 8 - Captura dos pantallas como las dos de arriba, de manera que se pueda ver los atributos ignorados (en la primera pantalla), el detalle de la agrupación de las instancias y el número de estas mal agrupadas (en una segunda pantalla). Argumenta los resultados obtenidos, en comparación con el resultado de la ejecución anterior.**

En el ejercicio 5 yo elegí los atributos petallength (X) con sepalwidth (Y), por tanto ifnore petalwidth y sepallength.

Con estos ignorados obtenemos:





Lo que hace que aumenten las instancias mal agrupadas. Por ende, no fue una buena elección.