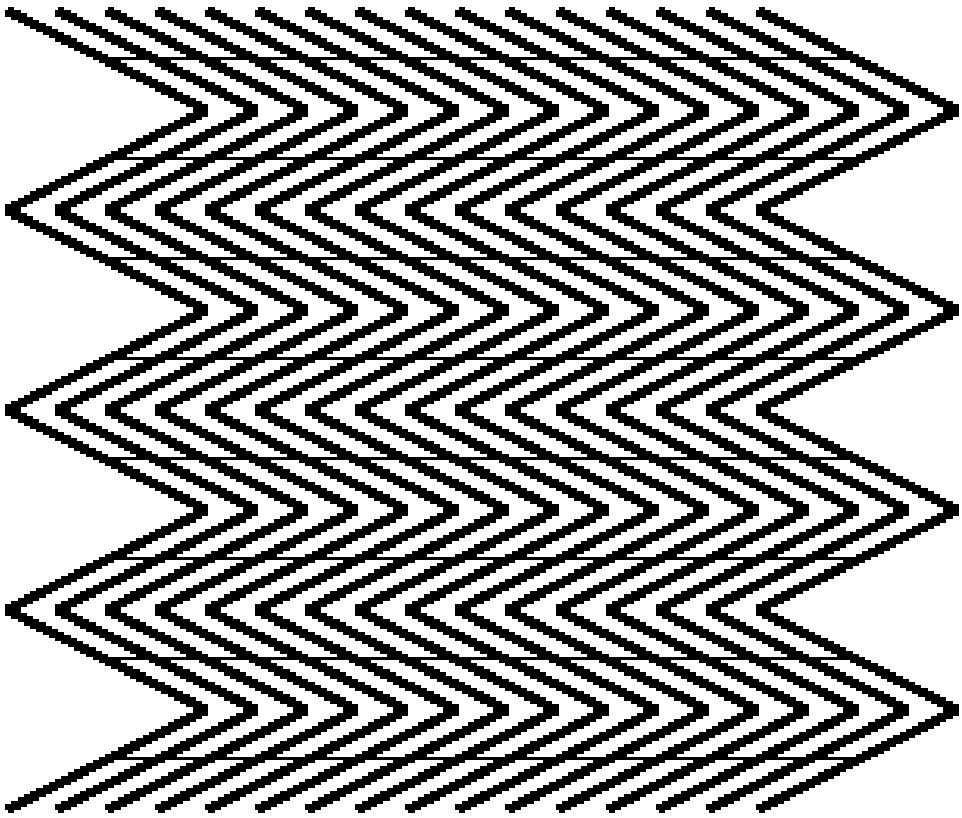
PROGRAMA DE DOCTORADO INTERUNIVERSITARIO

APRENDIZAJE AUTOMÁTICO Y DATA MINING

Práctica 1:
Entorno WEKA de aprendizaje automático y data mining.

Objetivos:

- Utilización de funciones de visualización de datos en WEKA.
- Lanzamiento de clasificadores en WEKA.



1. INTRODUCCIÓN

WEKA es una herramienta de aprendizaje automático y *data mining*, escrita en lenguaje Java, gratuita y desarrollada en la Universidad de Waikato (WEKA = Waikato Environment for Knowledge Analysis).

El programa WEKA se puede descargar desde:

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

2. FORMAS DE UTILIZAR WEKA

WEKA se puede utilizar de 3 formas distintas:

A: Desde la línea de comandos

Cada uno de los algoritmos incluidos en WEKA se pueden invocar desde la línea de comandos de MS-DOS como programas individuales. Los resultados se muestran únicamente en modo texto.

B: Desde uno de los interfaces de usuario

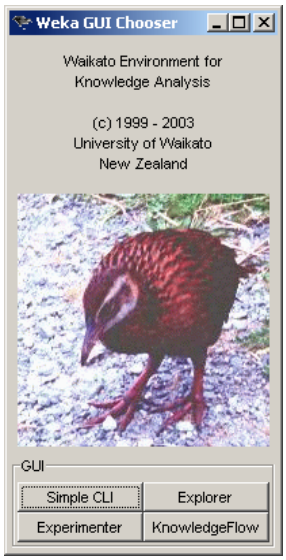
WEKA dispone de 4 interfaces de usuario distintos, que se pueden elegir después de lanzar la aplicación completa. Los interfaces son:

- **Simple CLI** (command line interface)**:** interfaz en modo texto.
- **Explorer:** interfaz gráfico básico.
- **Experimenter:** interfaz gráfico con posibilidad de comparar el funcionamiento de diversos algoritmos de aprendizaje.
- **KnowledgeFlow**: interfaz gráfico que permite interconectar distintos algoritmos de aprendizaje en cascada, creando una red.

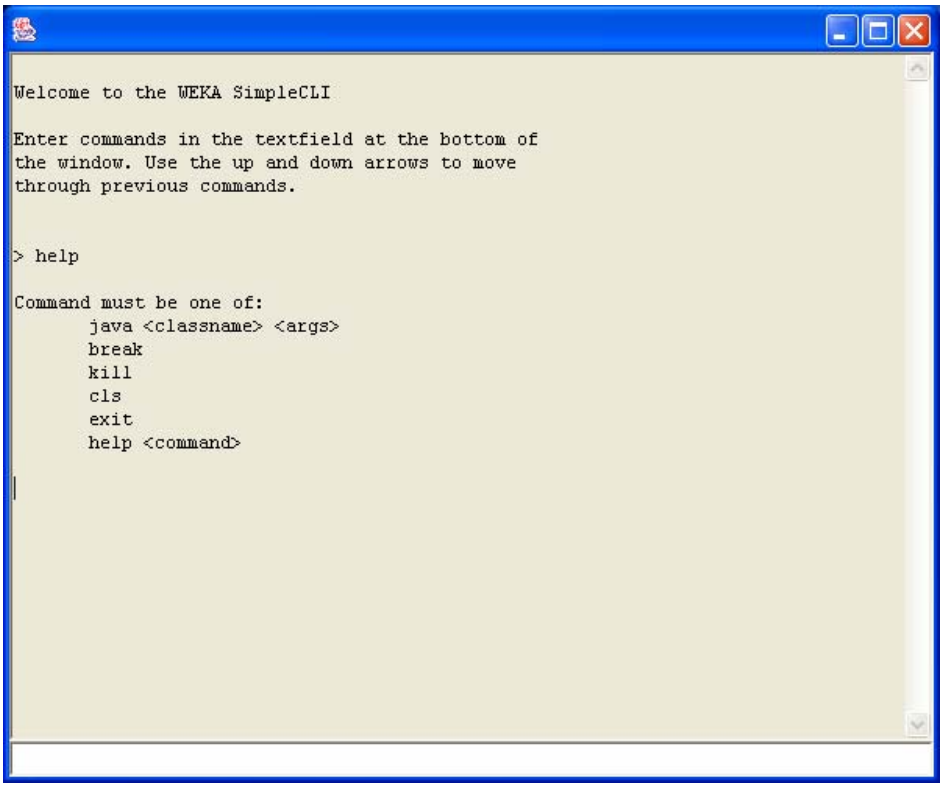
Para comprobar el aspecto de cada interfaz, se lanzará la aplicación. En el directorio donde se encuentre instalado WEKA, se deberá ejecutar el programa java **weka.jar**. En algunos ordenadores bastará con hacer doble clic sobre el fichero weka.jar; en otros ordenadores será necesario teclear desde la ventana de comandos la sentencia:

```
java –jar weka.jar
```

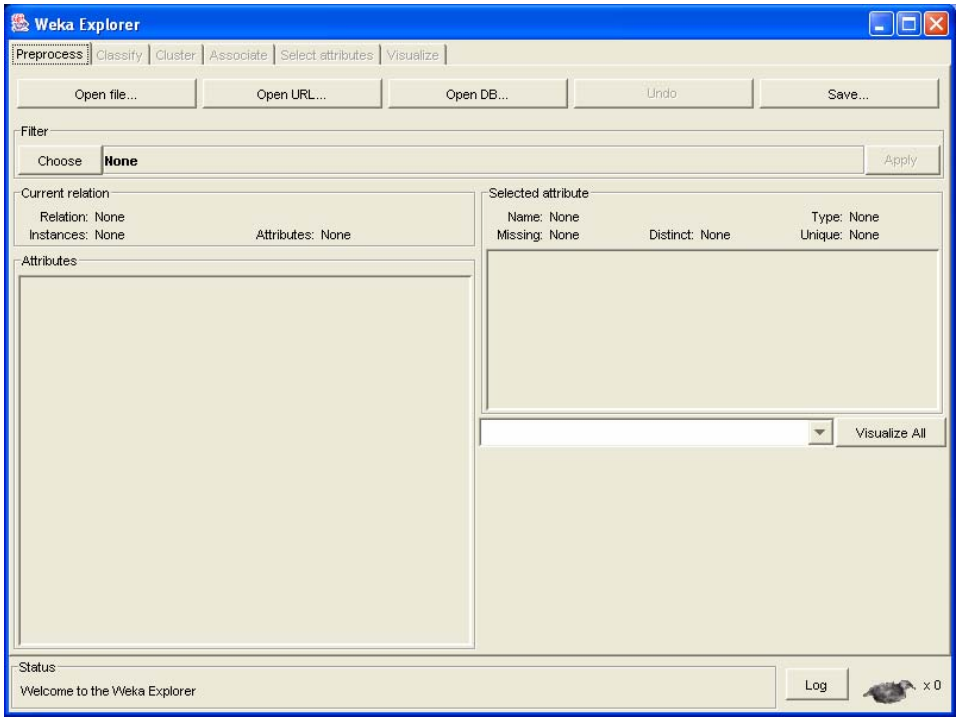
El aspecto de la pantalla inicial debe ser el siguiente:



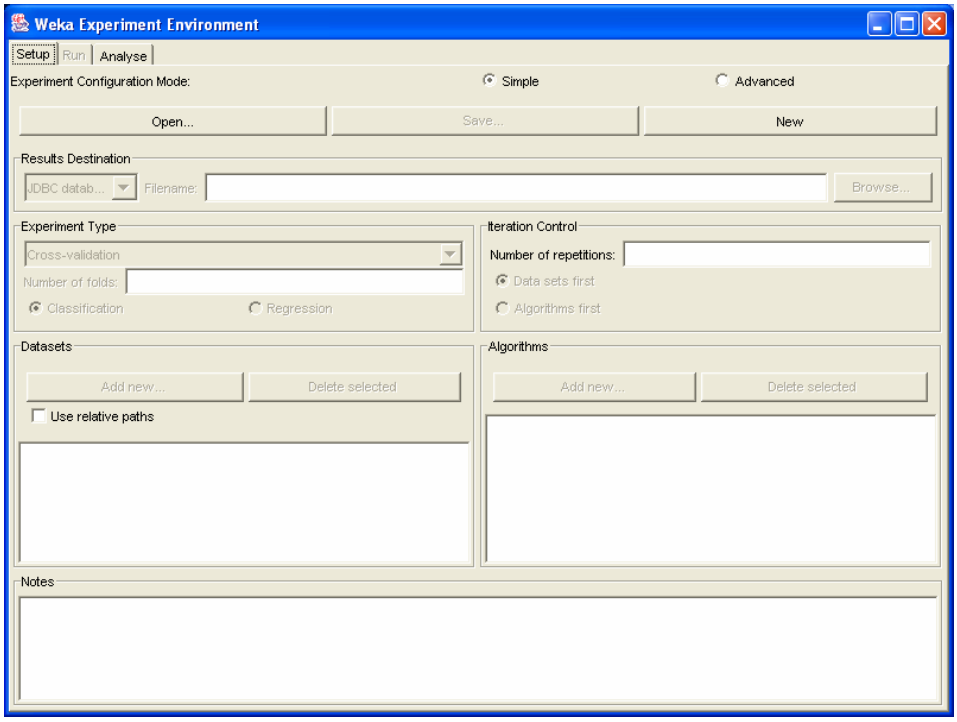
Los botones de la parte inferior permiten elegir uno de los cuatro interfaces. El aspecto de cada uno de ellos se muestra en las figuras siguientes:



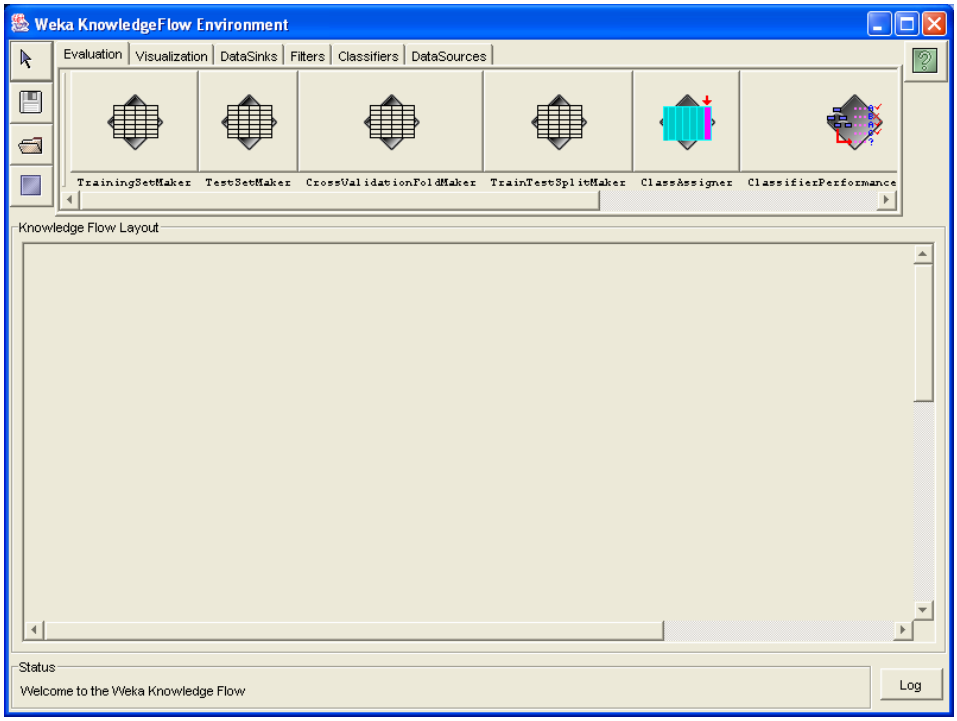
Interfaz Simple CLI



Interfaz Explorer



Interfaz Experimenter



Interfaz KnowledgeFlow

C: Creando un programa Java

La tercera forma en la que se puede utilizar el programa WEKA es mediante la creación de un programa Java que llame a las funciones que se desee. El código fuente de WEKA está disponible, con lo que se puede utilizar para crear un programa propio.

3. DÓNDE ENCONTRAR AYUDA Y DOCUMENTACIÓN SOBRE WEKA

- La información básica en forma de presentación en PowerPoint se puede encontrar en la siguiente página de internet:

  http://prdownloads.sourceforge.net/weka/weka.ppt

- Para información más detallada, se puede acceder a la página principal de WEKA:

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

4. PRIMER EJEMPLO DE UTILIZACIÓN DE WEKA

Como primer ejemplo, se trabajará sobre una base de datos clásica incluida en el propio programa. Se trata de una base de datos en la que se pretende determinar cuáles son los factores que hacen que una cierta persona practique o no el tenis.

Cada instancia de la base de datos se corresponde con un cierto día en el que la persona se plantea si jugar o no al tenis, y recoge los siguientes atributos:

- Aspecto del cielo: {soleado, cubierto, lluvioso}.
- Temperatura: medida en grados.
- Humedad: medida en %.
- Viento: {si, no}.
- Juega al tenis: {si, no}.

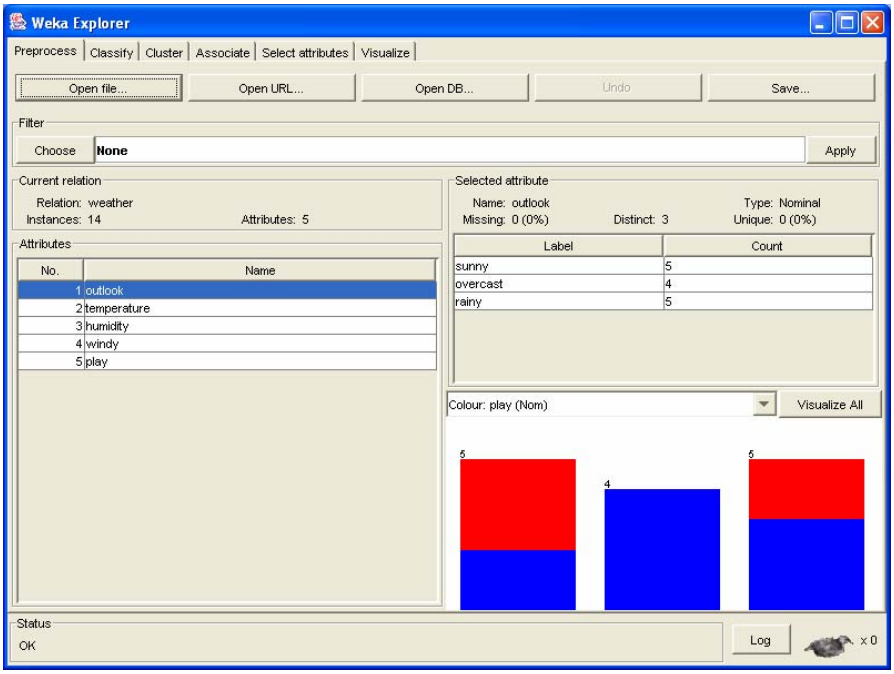
Se dispone de datos recogidos durante 14 días distintos, y el objetivo es determinar cuál es la relación entre las condiciones del tiempo y la decisión de jugar o no al tenis.

PASO 1: Lanzar el interfaz Explorer

En esta primera práctica se utilizará WEKA desde el interfaz Explorer. Se lanzará este interfaz de acuerdo con lo indicado en la introducción.

PASO 2: Cargar la base de datos

Para cargar la base de datos se utilizará el botón **OPEN FILE** del interfaz Explorer (pestaña **Preprocess**), se seleccionará el directorio **data** y dentro de él, el fichero **weather.arff**. El resultado será una pantalla como la que se muestra en la figura:



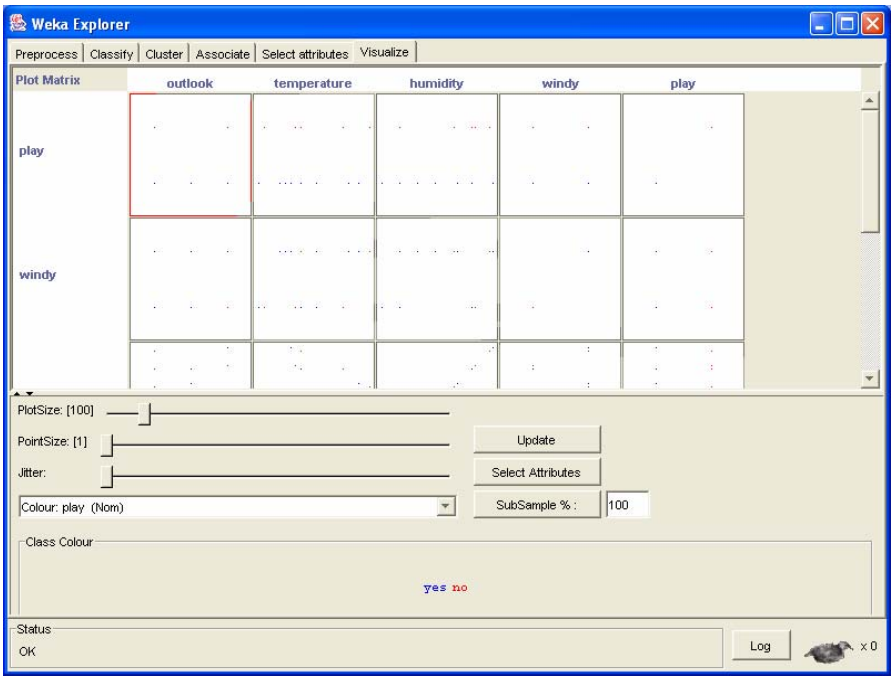
En la parte izquierda de la pantalla aparecen los cinco atributos mencionados:

- Outlook
- Temperature.
- Humidity.
- Windy.
- Play.

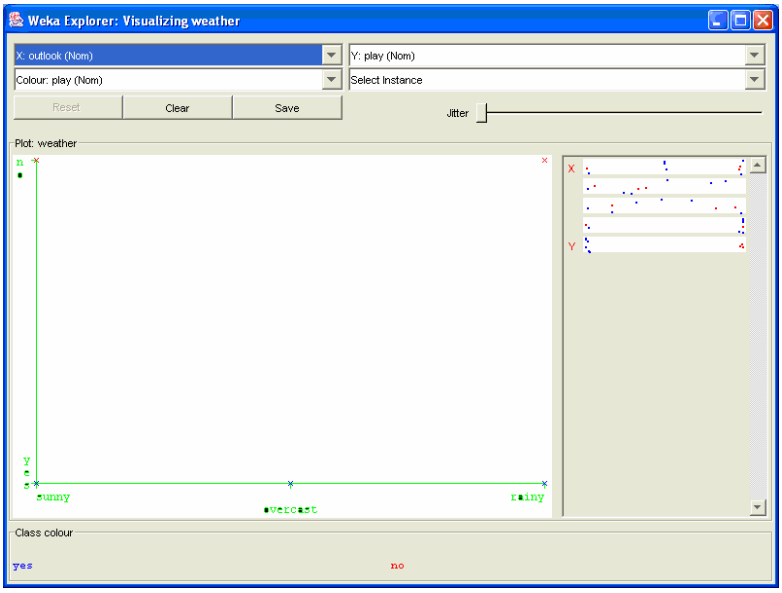
Haciendo clic sobre cada uno de los atributos, se muestra información sobre el mismo en la parte derecha de la ventana. En el caso de atributos discretos se indica el número de instancias que toman cada uno de los valores posibles; y en el caso de atributos reales se muestran los valores máximo, mínimo, medio y la desviación estándar. Asimismo, se muestra un gráfico en el que las distintas clases (juega o no juega) se representan con colores distintos, en función de los valores del atributo elegido.

PASO 3: Generación de graficos

Para generar gráficos con los datos del ejemplo, se seleccionará la pestaña **Visualize**. Por defecto, se muestran gráficos para todas las combinaciones de atributos tomadas dos a dos, de modo que se pueda estudiar la relación entre dos atributos cualesquiera. El aspecto de la pantalla es el mostrado en la figura siguiente:

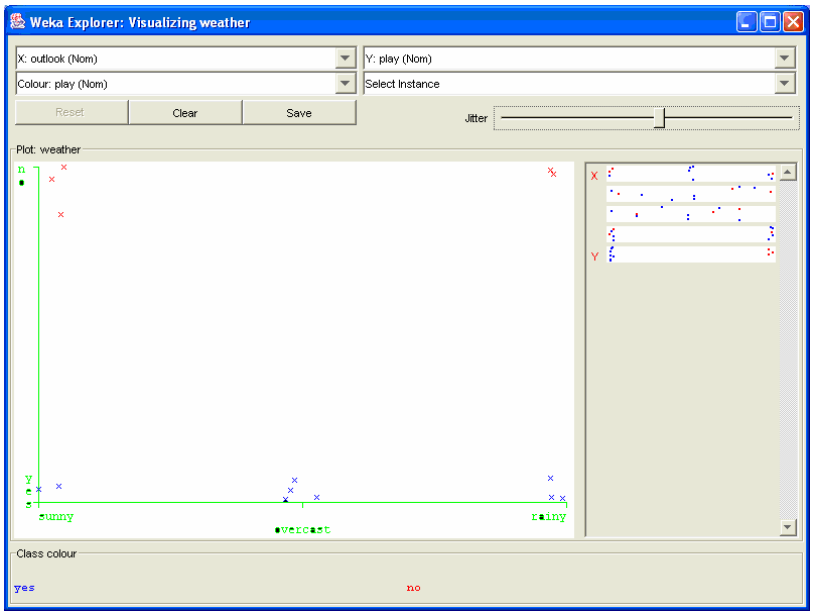


Si se desea mostrar un gráfico concreto, basta con hacer doble clic sobre él. Por ejemplo, haciendo doble clic sobre el gráfico que relaciona el aspecto del cielo con la decisión de jugar o no (play / outlook) se muestra el gráfico de la figura siguiente:



Según el gráfico, cuando el día es soleado puede tomarse la decisión de jugar o no (aparecen ejemplos de valor ‘si’ (azules) y ejemplos de valor ‘no’ (rojos). Cuando el cielo está cubierto, se juega siempre; y cuando el día es lluvioso también se pueden tomar las dos decisiones.

Un problema que presenta el gráfico es que los puntos se superponen, con lo cual es imposible determinar cuántos ejemplos representa cada cruz. Para solucionar este problema, se introduce un ruido en el gráfico (perturbaciones aleatorias de los valores) de modo que los puntos superpuestos se separen. Para introducir el ruido se utiliza la función **jitter**, desplazando el cursor hasta que la visualización sea la deseada. Una posible visualización se muestra en la figura siguiente:



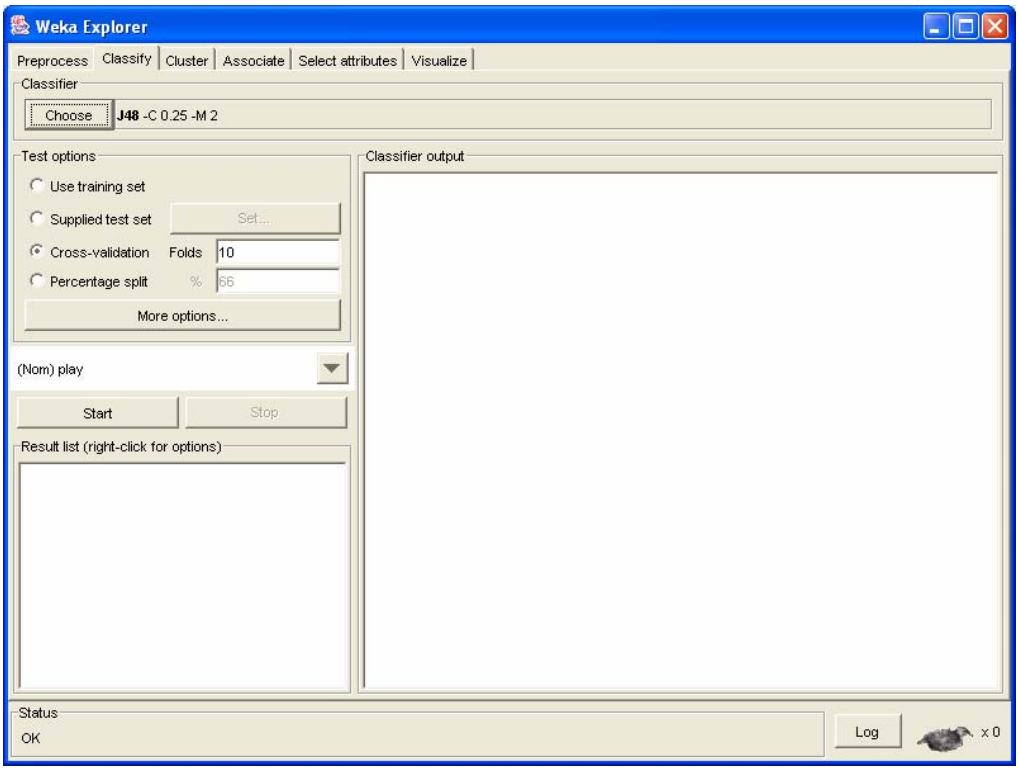
A ENTREGAR: EJERCICIO NÚMERO 1

Representar los gráficos que relacionan cada uno de los atributos con la decisión de jugar o no e indicar si influyen o no influyen en esta decisión. En total debe haber 4 gráficos. Se debe utilizar el ruido aleatorio cuando sea necesario y se deben pegar los gráficos en un documento Word con la tecla *ImprPant* del teclado.

4. GENERACIÓN DE UN ÁRBOL DE DECISIÓN CON WEKA

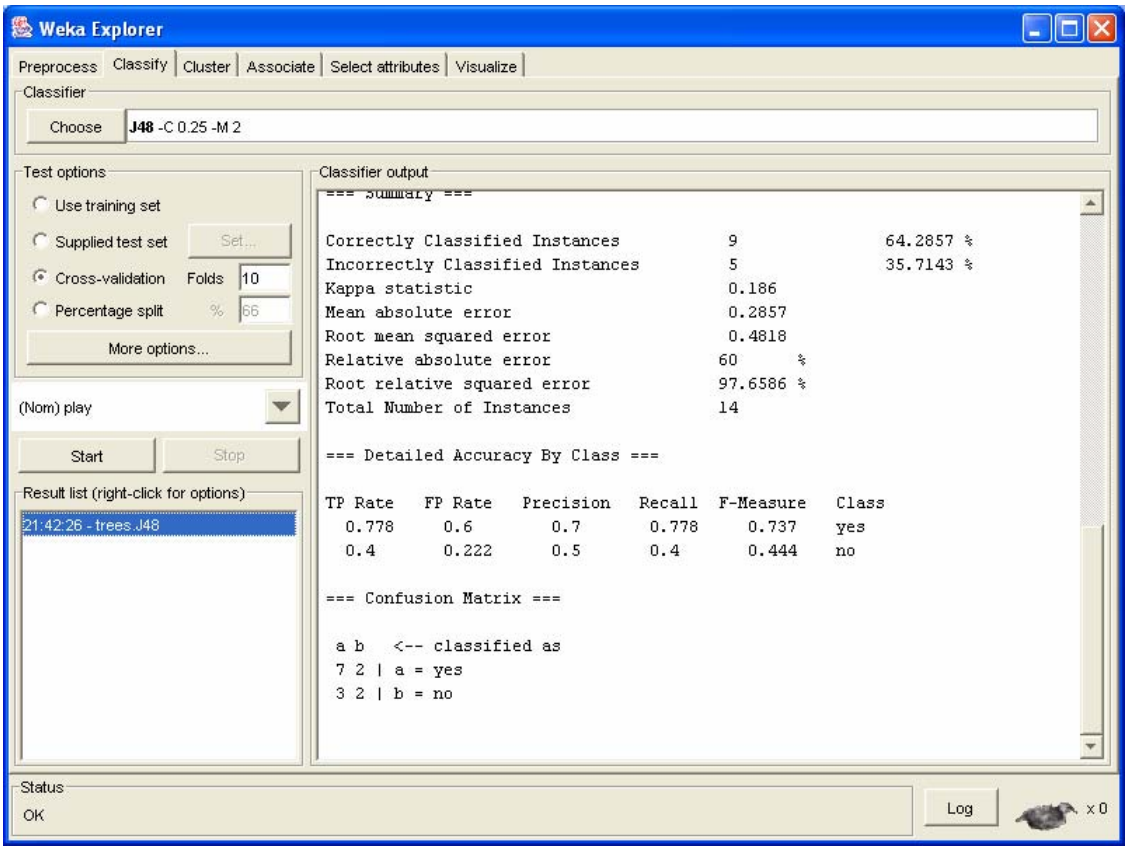
Una vez vistas las posibilidades de representación gráfica, se probará uno de los algoritmos de aprendizaje automático incluidos en WEKA: los árboles de decisión. Se intentará generar un árbol de decisión que se corresponda con los ejemplos de la base de datos anterior.

Para ello se seleccionará la pestaña **Classify** y se elegirá un clasificador pulsando el botón **Choose**. Aparecerá una estructura de directorios en la que se seleccionará el directorio **trees** y dentro de él, el algoritmo **J48**. Se mantendrán las opciones por defecto del clasificador (**J48 –C 0.25 –M 2**), tal y como muestra la pantalla siguiente.



El resto de opciones para el experimento también se mantendrán en los valores por defecto: activa la opción de test ‘**cross validation**’ e inactivas las restantes. Para generar el árbol se pulsará **Start**.

El resultado será el que muestra la pantalla siguiente, donde se muestran en modo texto tanto el árbol generado como la capacidad de clasificación del mismo:



Si se analiza la información que se ofrece en modo texto, se puede destacar lo siguiente:

En primer lugar, se muestra información sobre el tipo de clasificador utilizado (algoritmo **J48**), la base de datos sobre la que se trabaja (**weather**) y el tipo de test (**cross validation**).

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    10-fold cross-validation
```

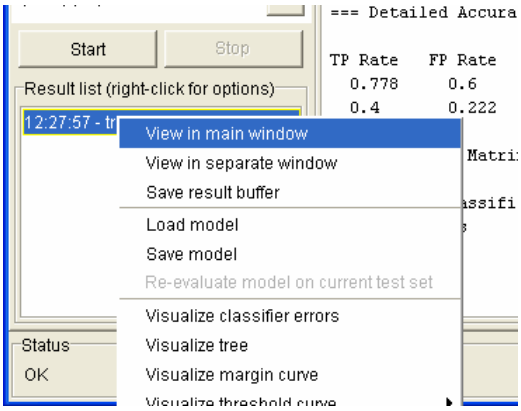
A continuación se muestra el árbol que se ha generado y el número de instancias que clasifica cada nodo:

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

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

outlook = sunny
|   humidity <= 75: yes (2.0)
|   humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :     5

Size of the tree :      8


Time taken to build model: 0 seconds
```

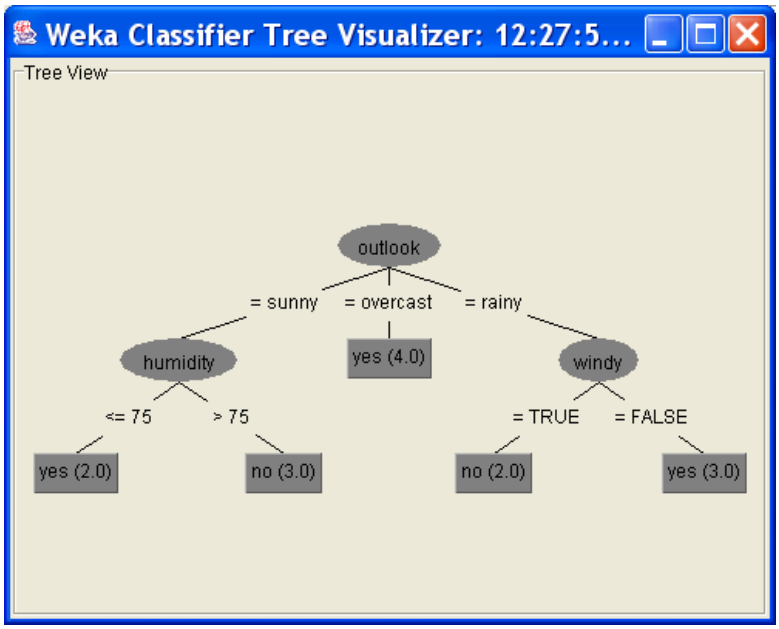
Y por ultimo se muestran los resultados del test (indican la capacidad de clasificación esperable para el árbol y la matriz de confusión):

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

Correctly Classified Instances           9               64.2857 %
Incorrectly Classified Instances         5               35.7143 %
Kappa statistic                          0.186
Mean absolute error                      0.2857
Root mean squared error                  0.4818
Relative absolute error                 60      %
Root relative squared error             97.6586 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.778     0.6        0.7       0.778     0.737    yes
  0.4       0.222      0.5       0.4       0.444    no

=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 3 2 | b = no
```

También es posible visualizar el árbol de decisión de una forma más legible. Para ello se debe hacer clic con el botón derecho en la ventana de resultados, sobre el resultado de la generación del árbol. Aparecerá un menú desplegable:



Y dentro de ese menú se deberá seleccionar la opción **'Visualize tree'**. El resultado se muestra en la figura siguiente:



5. INTRODUCCIÓN DE DATOS PROPIOS EN WEKA

A continuación se hará funcionar WEKA con datos propios, distintos de los datos de ejemplo utilizados hasta ahora. Para ello será fundamental guardar los datos en el tipo de formato utilizado por WEKA (ficheros con extensión **.arff**)

Como ejemplo, se muestra el fichero utilizado en la primera parte de la práctica:

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

El encabezado del fichero contiene el nombre de la base de datos, en este caso **weather**.

```
@relation weather
```

A continuación se indican los atributos que existen y los valores que pueden tomar cada uno de ellos. Pueden ser atributos reales (cualquier valor) o atributos discretos (se deben especificar los posibles valores entre paréntesis).

```
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
```

Por último, la línea **@data** y a continuación todos los datos, uno por cada línea, y en el orden indicado para los atributos.

```
@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
…
```

Crearemos un fichero para WEKA con un ejemplo propio: datos sobre la posibilidad de fallo de una máquina en función de ciertos atributos medidos: vibraciones, temperatura,

tiempo desde la última revisión y horas de funcionamiento. Estos datos se muestran a continuación:

Temperatura	Vibraciones	Horas funcionamiento	Días desde revisión	FALLO
55	si	500	55	si
23	no	30	17	no
45	no	1500	72	no
47	no	650	43	no
32	si	700	58	no
35	si	2500	93	si
50	si	150	21	si
53	si	550	50	si
21	no	35	12	no
47	no	1200	75	no
43	no	750	51	no
35	si	680	63	no
30	si	2300	87	si
52	si	180	23	si

Para evitar teclearlos, los datos están disponibles también en una hoia Excel en la siguiente dirección de internet:

http://isa.umh.es/isa/es/asignaturas/aprendizaje/maquina.xls

A ENTREGAR: EJERCICIO NÚMERO 2

Crear un fichero con los datos anteriores en formato WEKA y guardarlo con la extensión .arff (copiar y pegar desde Excel para evitar teclear, luego adaptar formato). En la cabecera del fichero debe aparecer la línea:

@relation XXXX_YYYY_ZZZZ

… donde XXXX, YYYY y ZZZZ deben ser el nombre y apellidos del alumno. Este fichero se copiará en el informe de la práctica.

Abrir el fichero .arff creado desde WEKA y generar un árbol de decisión sobre esos datos. Copiar en el informe tanto el resultado ofrecido en modo texto como la representación gráfica del árbol de decisión.