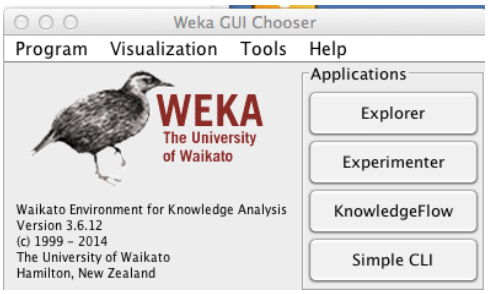
# Introducción al aprendizaje computacional con WEKA

TUTORIAL

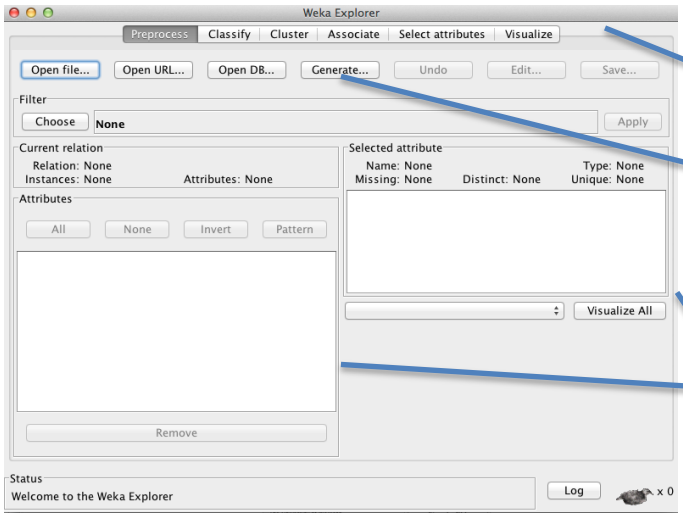
En esta práctica vamos a utilizar el software WEKA para resolver algunos ejercicios de aprendizaje supervisado y no supervisado.

## 1. Introducción y visualización de datos en WEKA

Tras descargar el programa, al ejecutarlo se abre el siguiente interfaz:




Pulsamos el botón *Explorer* y nos encontraremos con el siguiente interfaz:

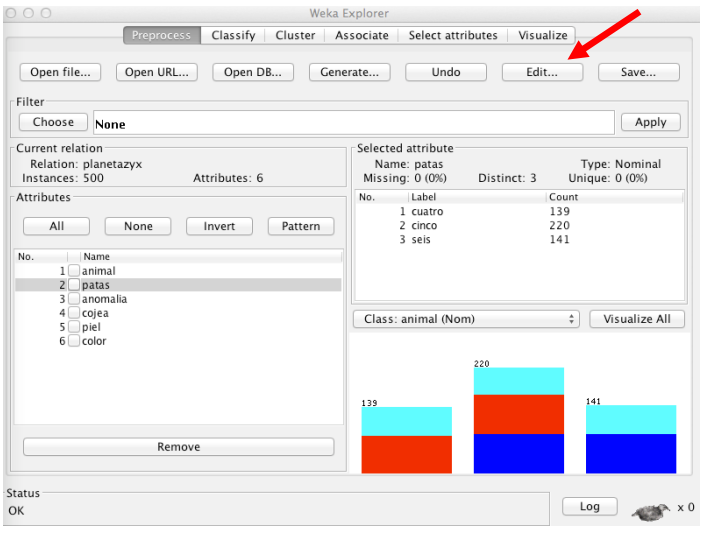


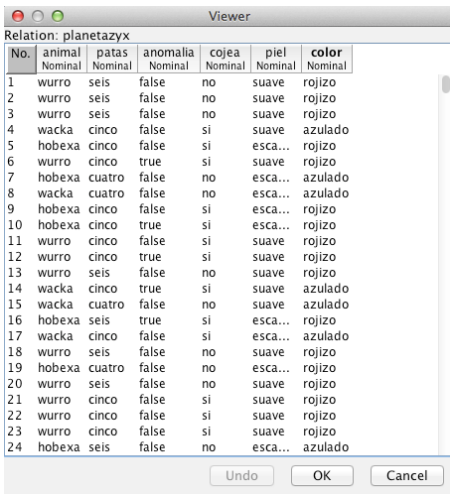

Pestañas: Preprocesamiento, Clasificación, Agrupamientos, Asociación, Selección de atributos, Visualización

Abrir ficheros de datos (desde el disco, desde una url, desde una base de datos, generar un fichero de datos sintético). Una vez abierto un fichero: opciones para deshacer un paso y para editar y guardar los ficheros de datos.

Pantallas que mostrarán los atributos e información sobre ellos (cuando abramos algún fichero de datos)

En WEKA se suele trabajar con ficheros con la extensión arff, pero como vemos también admite otro tipo de formatos. Vamos a abrir nuestro fichero de ejemplo (planetazyx_500.arff), un fichero generado con GeNIe que contiene casos para el problema de clasificación de los animales del planeta ZYX. Para ver (y editar si es necesario) los datos que acabamos de abrir podemos darle al botón Edit y se nos abre el visor de datos "Wiewer", donde también podemos si queremos cambiar los datos y guardar los cambios con el botón Save.



En el interfaz podemos ver que hay 500 registros, con seis atributos de tipo nominal. Seleccionando un atributo conoceremos más información del atributo en cuestión: por ejemplo, en este caso está mostrando datos del segundo atributo (patas), que es de tipo nominal, no tiene valores perdidos y tiene tres valores distintos. Para variables numéricas, se mostrará también su valor mínimo, máximo, la media y la desviación típica. También se muestran los tres valores de atributo (cuatro, cinco, seis) y la distribución de dichos valores en el conjunto de datos (139,220,141).

Por defecto, WEKA considera que *la clase* es el último atributo (color), pero se puede cambiar en la pestaña clase (aparece justo encima del histograma) y especificar que la clase de interés es el atributo animal.

Finalmente, el histograma muestra cuántos animales que tienen cierto número de patas (cuatro, cinco o seis) son wurros (azul), wackas (rojo) u hobexas (turquesa). El botón "Visualize all" muestra los histogramas de todas las variables.

## 2. Aprendizaje supervisado

El aprendizaje supervisado se realiza cuando todos los datos del ejemplo están etiquetados, es decir, para los datos del ejemplo se conoce la clase. Los ejemplos del tema de aprendizaje bayesiano pertenecen a esa categoría, y para ellos se utilizan modelos basados en redes bayesianas o en el clasificador Naive Bayes. Pero hay también otro tipo de modelos, que veremos a continuación.

### 2.1. Árboles de decisión

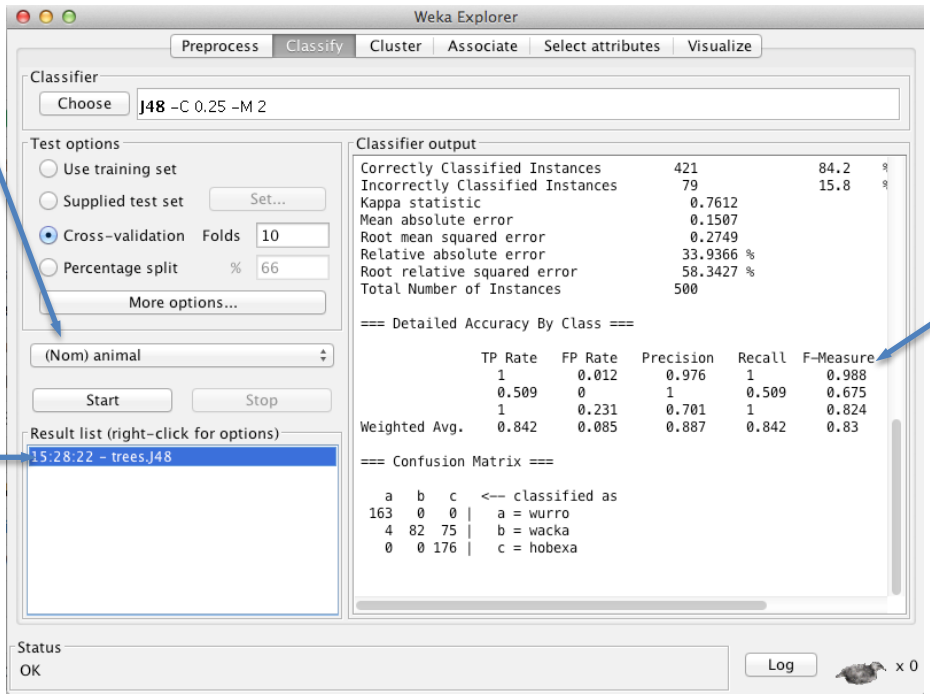
Vamos ahora a aprender modelos. Comenzaremos por aprender un árbol de decisión. Para ello seleccionamos ahora la pestaña "Classify" y vemos que el interfaz cambia. Para elegir el clasificador, pulsamos el botón "Choose" y dentro de la carpeta "trees" seleccionamos el algoritmo J48 (que es una implementación en Java del algoritmo C4.5). Vemos que ahora aparecen diferentes opciones:

- Use training set. El modelo se aprende y evalúa con los mismos datos.
- Supplied test set. Podemos cargar un conjunto de datos para realizar la evaluación del modelo aprendido (normalmente diferente al conjunto de datos utilizado para aprender los modelos).
- Cross-validation. Se realiza la evaluación mediante la técnica de validación cruzada, con el número de iteraciones (folds) que queramos considerar (por defecto es 10).
- Percentage split. Se define un porcentaje con el que se aprende el modelo y el resto del conjunto de datos se utiliza para evaluar su rendimiento.

Todo clasificador se carga por defecto con los valores recomendados para los parámetros, que podemos modificar caso de que lo consideremos necesario (pinchando en la casilla junto al botón Choose). Una vez configurado el clasificador, presionamos el botón "Start" y tendremos el siguiente interfaz:

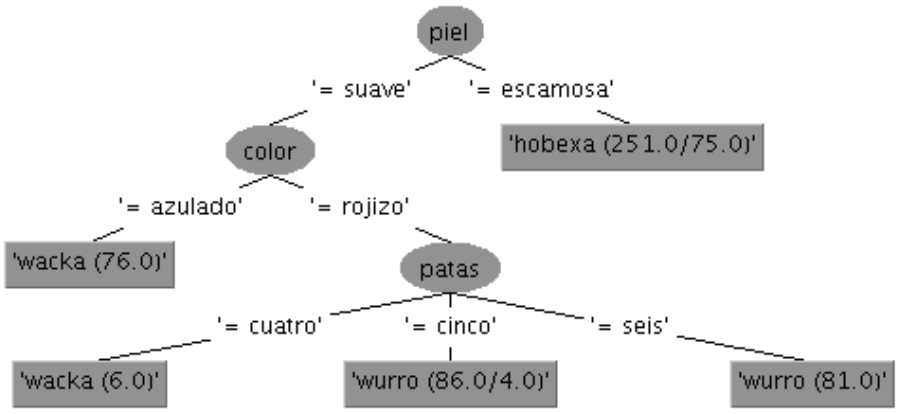
En este menú desplegable podemos cambiar la clase

En la lista de resultados va conservando los resultados de las diferentes ejecuciones. Pinchando en el menú contextual, podemos acceder a diferentes opciones (entre ellas visualizar el árbol aprendido seleccionando la opción "Visualize tree")




El deslizador nos permite desplazarnos por la pantalla, mostrando los resultados del clasificador que hayamos ejecutado: en la parte superior de la pantalla aparece el modelo aprendido (árbol, pero en formato texto), el número de hojas, el tamaño del árbol. En la parte inferior vemos lo que se muestra en esta figura: instancias clasificadas correctamente, incorrectamente y matriz de confusión.

Seleccionando la opción Visualize tree (aparece menú contextual en la ejecución correspondiente de la lista resultados) vemos el modelo aprendido de un modo más atractivo que en la ventana Classifier output:




Este es el modelo que utilizaríamos para clasificar nuevos ejemplos. Vemos que de la base de datos utilizada para el aprendizaje en el proceso *10-fold cross-validation*, clasificó correctamente 421 instancias (un 84%). Mirando la matriz de confusión, podemos ver que todos los wurros y las hobexas fueron correctamente clasificados, mientras que 4 wackas fueron incorrectamente clasificadas como wurros y 75 como hobexas.

Para la entrega, debes responder ahora a las preguntas 1 y 2 (véase el apartado Tarea y entrega)

**2.2. Reglas de clasificación**

Vamos ahora a utilizar un clasificador basado en reglas. Sobre el mismo conjunto de datos, elegimos ahora un nuevo clasificador pulsando el botón "Choose". En este caso, de la carpeta "Rules" vamos a elegir el clasificador JRip, que es una implementación del algoritmo RIPPER. Presionamos "Start" y en la ventana de resultados del clasificador podemos ver los nuevos resultados: el nuevo clasificador ha clasificado correctamente 410 instancias (un 82%). Clasifica correctamente a todos los wurros, sin embargo, se equivoca en 5 wackas que clasifica como wurros, 53 wackas que clasifica como hobexas y 32 hobexas que clasifica como wackas. En la parte superior podemos ver las reglas que ha aprendido el algoritmo.

```
JRIP rules:
===========

(color = azulado) and (piel = suave) => animal=wacka (76.0/0.0)
(color = azulado) and (patas = cinco) => animal=wacka (64.0/31.0)
(patas = cuatro) and (piel = suave) => animal=wacka (6.0/0.0)
(patas = cuatro) and (color = azulado) and (anomalia = false) => animal=wacka (53.0/26.0)
(piel = suave) => animal=wurro (167.0/4.0)
 => animal=hobexa (134.0/15.0)

Number of Rules : 6
```

Utiliza ahora este conjunto de reglas de clasificación para resolver la pregunta 3 de la entrega.

**3. Aprendizaje no supervisado**

En el aprendizaje no supervisado no disponemos de la clase a la que pertenecen los ejemplos (o se dispone de ella, pero no se quiere utilizar). Pese a no tener la clase, aún es posible extraer patrones comunes o relaciones entre las variables que nos puedan ser de utilidad. Si queremos encontrar relaciones entre las variables de la base de datos, usaremos reglas de asociación. También podemos agrupar los datos en grupos con características comunes, e incluso esta agrupación podría ayudarnos a hacer predicciones futuras. Veamos unos ejemplos.

**3.1. Reglas de asociación**

Vamos a estudiar ahora los datos del hundimiento del Titanic[1]. Abre el fichero "titanic.arff" y encontrarás una base de datos con las características de los 2.201 pasajeros a bordo en el momento del hundimiento. Los atributos son:

- Clase (0 = tripulación, 1 = primera, 2 = segunda, 3 = tercera)
- Edad (0 = menor, 1 = adulto)
- Sexo (0 = femenino, 1 = masculino)
- Sobrevivió (1 = sí, 0 = no)

[1] Ejemplo tomado del Curso de Doctorado "Extracción automática del Conocimiento" de la Universidad de Valencia (profesores José Hernández y César Cerri). Datos reales, obtenidos de: *"Report on the Loss of the 'Titanic' (S.S.)" (1990), British Board of Trade Inquiry Report_ (reprint), Gloucester, UK: Allan Sutton Publishing.*

En este ejemplo podríamos también aplicar técnicas de aprendizaje supervisado (si consideramos que la clase es "sobrevivió?"). Pero en lugar de eso vamos a utilizarlos para ver qué reglas de asociación (entre todos los atributos) se pueden extraer para estos atributos.
Para ejecutar los métodos de reglas de asociación en Weka, seleccionamos la ventana "Associate". Veamos las reglas que extrae este algoritmo tomando los valores por defecto:

```
 1. Clase=0 885 ==> Edad=1 885    <conf:(1)> lift:(1.05) lev:(0.02) [43] conv:(43.83)
 2. Clase=0 Sexo=1 862 ==> Edad=1 862    <conf:(1)> lift:(1.05) lev:(0.02) [42] conv:(42.69)
 3. Sexo=1 Sobrevivio?=0 1364 ==> Edad=1 1329    <conf:(0.97)> lift:(1.03) lev:(0.01) [32] conv:(1.88)
 4. Clase=0 885 ==> Sexo=1 862    <conf:(0.97)> lift:(1.24) lev:(0.08) [165] conv:(7.87)
 5. Clase=0 Edad=1 885 ==> Sexo=1 862    <conf:(0.97)> lift:(1.24) lev:(0.08) [165] conv:(7.87)
 6. Clase=0 885 ==> Edad=1 Sexo=1 862    <conf:(0.97)> lift:(1.29) lev:(0.09) [191] conv:(8.95)
 7. Sobrevivio?=0 1490 ==> Edad=1 1438    <conf:(0.97)> lift:(1.02) lev:(0.01) [21] conv:(1.39)
 8. Sexo=1 1731 ==> Edad=1 1667    <conf:(0.96)> lift:(1.01) lev:(0.01) [21] conv:(1.32)
 9. Edad=1 Sobrevivio?=0 1438 ==> Sexo=1 1329    <conf:(0.92)> lift:(1.18) lev:(0.09) [198] conv:(2.79)
10. Sobrevivio?=0 1490 ==> Sexo=1 1364    <conf:(0.92)> lift:(1.16) lev:(0.09) [192] conv:(2.51)
```

En cada regla, tenemos la cobertura de la parte izquierda y de la regla, así como la confianza de la regla (conf) y otras medidas de confianza. Podemos conocer alguna regla interesante, aunque otras los son menos. Por ejemplo, la regla 1 indica que, como era de esperar toda la tripulación es adulta. La regla 2 nos indica lo mismo, pero teniendo en cuenta a los varones. Parecidas conclusiones podemos sacar de las reglas 4, 5 y 6. La regla 3 nos indica que los varones que murieron fueron en su mayoría adultos (97%). La regla 7 destaca que la mayoría de los que murieron fueron adultos (97%). Y finalmente la regla 10 informa que la mayoría de los muertos fueron varones (92%).

Cabe destacar que la calidad de las reglas de asociación que aprendamos muchas veces viene lastrada por la presencia de atributos que estén fuertemente descompensados. Por ejemplo, en este caso la escasa presencia de niños provoca que no aparezcan en las reglas de asociación (son filtradas por tener una baja cobertura). Responde ahora a la pregunta 4 de la entrega.

### 3.2. Agrupamiento

Por último, vamos a ilustrar con un ejemplo el aprendizaje de los agrupamientos. Para ello vamos a utilizar datos reales de un curso Moodle[2] de 20 estudiantes (fichero moodle.arff). Tras abrir el fichero, nos vamos a la pestaña Cluster. Vamos a seleccionar el algoritmo SimpleKMeans. En la configuración del algoritmo elegimos Num_Clusters=2 y presionamos "Start". En este caso y, como se muestra en la siguiente figura, en la ventana de resultados muestra la suma de errores cuadráticos medios de cada valor con respecto al centroide de su cluster y también el valor del centroide de cada cluster. La clasificación de nuevas instancias se haría calculando la distancia de la instancia al centroide y eligiendo el cluster con la distancia mínima.

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 9.03426043493763
Missing values globally replaced with mean/mode

Cluster centroids:
                                    Cluster#
Attribute               Full Data         0         1
                             (20)      (10)      (10)
=====================================================
n_assignment                 7.4       4.1      10.7
n_posts                     0.45       0.2       0.7
n_quiz                       3.5       1.4       5.6
n_quiz_a                    2.75       0.5         5
n_quiz_s                    0.75       0.9       0.6
total_time_assignment     503.25     294.6     711.9
total_time_quiz          1498.85     702.3    2295.4
total_time_forum          1317.6     919.5    1715.7
```

Responde ahora a la pregunta 5 de la entrega.

[2] Facilitados por Cristóbal Romero, de la Universidad de Córdoba

# TAREA Y ENTREGA

**Tarea:** Contesta a las preguntas que figuran a continuación
**Entrega:** Documento pdf con la solución (capturas de pantalla y textos descriptivos)

---

**Pregunta 1.** ¿Cómo clasificaríamos a un bicho rojizo que cojea según este modelo? ¿Y a un bicho de piel escamosa? ¿Y a un bicho de piel suave, rojizo y con cuatro patas?

**Pregunta 2.** ¿Encuentras alguna explicación razonable a que los errores de clasificación se cometan con las wackas?

**Pregunta 3**. ¿Cómo clasificaríamos a un bicho rojizo que cojea según este modelo? ¿Y a un bicho de piel escamosa? ¿Y a un bicho de piel suave, rojizo y con cuatro patas?

**Pregunta 4: ¿**Cuántos varones viajaban en el Titanic? ¿Cuántas mujeres? ¿Cuántos menores de edad? ¿Cuántos viajeros en primera clase? Modifica los parámetros del algoritmo para que aprenda 26 reglas de asociación con una confianza de 0.85, e interpreta el significado de las cinco últimas.

**Pregunta 5:** A la vista de los datos relativos a cada cluster, ¿qué grupo crees que representa mejor a los estudiantes que van a aprobar la asignatura? ¿Y a los que van a suspenderla?