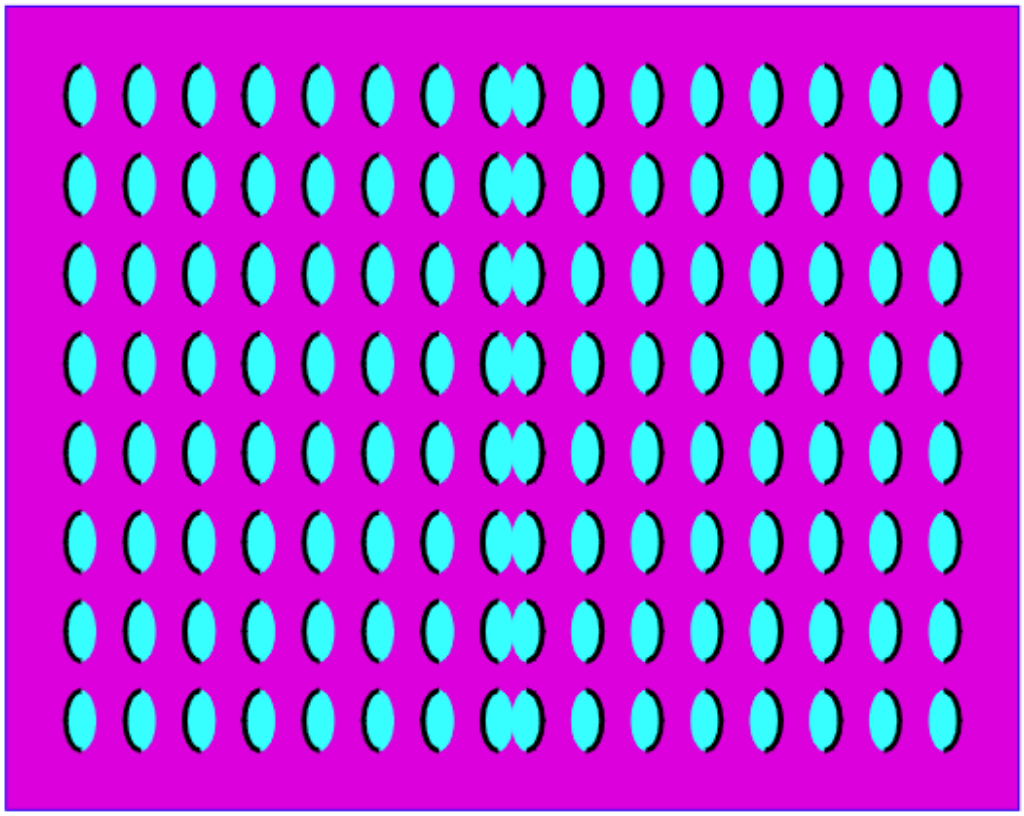
PROGRAMA DE DOCTORADO INTERUNIVERSITARIO

APRENDIZAJE AUTOMÁTICO Y DATA MINING

Práctica 3:
Almacenamiento de modelos con WEKA.

Objetivos:

- Almacenar modelos creados con WEKA.
- Utilizar un modelo previo para clasificar ejemplos no etiquetados.



1. ENTRENAMIENTO Y CLASIFICACIÓN

La utilización real del aprendizaje automático comprende dos pasos:

- **Entrenamiento**: se parte de unos datos conocidos o ejemplos de entrenamiento y se utilizan para construir un modelo (por ejemplo, un árbol de decisión).
- **Clasificación**: una vez creado el modelo con los datos conocidos, se utiliza el modelo para clasificar nuevos datos, para los que no se conoce de antemano la clase.

En las prácticas anteriores únicamente se ha generado el modelo, pero no se ha utilizado para clasificar datos desconocidos. En esta práctica se aprenderá como guardar los modelos y cómo utilizarlos para clasificar nuevos ejemplos.

2. EJEMPLO: RECOMENDACIÓN DE LENTES DE CONTACTO

Supongamos que se conocen datos sobre el tipo de lentes de contacto recomendadas en función de ciertos datos: edad (3 valores), tipo de problema visual (2 valores), existencia de astigmatismo (2 valores) y nivel de producción de lágrimas (2 valores). Parte de estos ejemplos almacenados se muestran a continuación:

Edad	Problema	Astigmatismo	Lágrimas	RECOMENDACIÓN
joven	miope	no	reducido	ninguna
joven	miope	no	normal	blandas
intermedio	hipermétrope	si	reducido	ninguna
intermedio	miope	si	normal	duras
vista cansada	miope	no	normal	ninguna
vista cansada	hipermétrope	no	normal	blandas
…	…	…	…	…

Se trata de un ejemplo disponible en el entorno weka:

```
c:\iarp\weka_java\data\contact-lenses.arff
```

El objetivo es generar un modelo que se ajuste a los ejemplos de entrenamiento y posteriormente usar el modelo para clasificar un ejemplo nuevo.

Trabajaremos desde Matlab, para que sea sencillo. Una vez situados en el directorio de weka, teclearemos la instrucción:

```
!java weka.classifiers.trees.J48 -t data/contact-lenses.arff -d mod.out
```

Se trata de la misma instrucción utilizada en la práctica anterior para generar un árbol de decisión con la diferencia de que el árbol se guarda como un modelo que se puede utilizar

posteriormente para clasificar nuevos ejemplos. La parte de la instrucción que permite hacer esto es:

```
-d mod.out
```

…e indica que el modelo se guarda en el fichero mod.out (podríamos haber dado cualquier otro nombre al modelo).

La forma de utilizar el modelo creado para clasificar un nuevo ejemplo es la siguiente:

1. En primer lugar, se debe crear un fichero .arff con el ejemplo (o los ejemplos) a clasificar. Dado que la clase de estos nuevos ejemplos es desconocida, se indica con una interrogación. Crearemos el siguiente fichero (la cabecera se puede copiar de **`data\contact-lenses.arff`** para ahorrar trabajo):

```
@relation nuevo

@attribute age                {young, pre-presbyopic, presbyopic}
@attribute spectacle-prescrip {myope, hypermetrope}
@attribute astigmatism        {no, yes}
@attribute tear-prod-rate     {reduced, normal}
@attribute contact-lenses     {soft, hard, none}

@data
young, hypermetrope, no, normal, ?
presbyopic, myope, yes, normal, ?
```

…y lo guardaremos como: **`c:\iarp\weka_java\nuevo.arff`**

2. En Segundo lugar, se debe lanzar WEKA para clasificar estos dos nuevos ejemplos de acuerdo con el modelo guardado antes. Para ello se tecleará la siguiente instrucción:

```
!java weka.classifiers.trees.J48 -T nuevo.arff -l mod.out -p 0
```

Donde aparecen nuevos parámetros:

`-T nuevo.arff`	indica el fichero de test (datos no clasificados a utilizar).
`-p 0`	para mostrar sólo los resultados de la clasificación.

El resultado en pantalla será el siguiente:

```
0 soft 0.8333333333333334 ?
1 hard 1.0 ?
```

…e indica que el en el primer caso se deben recomendar lentillas blandas y en el segundo caso lentillas duras. Los valores que aparecen a la derecha se refieren a la confianza de la clasificación, pero no se usarán en esta práctica.

A ENTREGAR: EJERCICIO NÚMERO 1

Se trabajará sobre un problema de clasificación de un tipo de plantas en función de medidas de su sépalo y su pétalo. Los datos del problema son los siguientes:

Longitud del sépalo	Anchura del sépalo	Longitud del pétalo	Anchura del pétalo	PLANTA
Nº real	Nº real	Nº real	Nº real	3 posibilidades: • Iris Setosa • Iris Versicolor • Iris Virginica

El fichero con los datos de entrenamiento está disponible en el directorio de WEKA:

```
c:\iarp\weka_java\data\iris.arff
```

A) Generar un árbol de decisión sobre ese fichero de entrenamiento, con el valor por defecto del el nivel de confianza para la poda; y guardar el modelo.

B) Obtener la clasificación de los siguientes ejemplos:

Longitud del sépalo	Anchura del sépalo	Longitud del pétalo	Anchura del pétalo	PLANTA
4.8	3.0	1.4	0.1	?
6.6	2.9	4.6	1.3	?
5.8	4.0	1.2	0.2	?
5.7	4.4	1.5	0.4	?
6.3	3.3	4.7	1.6	?
4.9	2.4	3.3	1.0	?
4.3	3.0	1.1	0.1	?
5.2	2.7	3.9	1.4	?
7.7	2.6	6.9	2.3	?
6.0	2.2	5.0	1.5	?
6.3	2.7	4.9	1.8	?
5.0	3.3	1.4	0.2	?
7.0	3.2	4.7	1.4	?
6.9	3.2	5.7	2.3	?

Para ello será necesario:

- **Crear un fichero .arff con los valores anteriores.**
- **Llamar a WEKA y clasificar los ejemplos de acuerdo con el modelo creado.**
- **Leer los resultados:**
  - **bien desde pantalla (se recomienda),**
  - **o bien automáticamente: redireccionando el resultado a un fichero y creando un programa Matlab que sea capaz de leerlos.**

C) **Mostrar los resultados sobre dos gráficos como los que se indican a continuación (los datos están inventados), donde Xxxxx Yyyyyy Zzzzzz se corresponden con el nombre y apellidos del alumno. Cada uno de los gráficos muestra la clasificación obtenida en función de dos de los atributos del problema:**