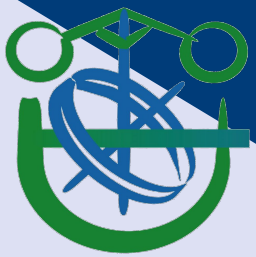


Escuela Politécnica Superior (EPS)

UNIVERSIDAD D CÓRDOBA
Departamento de Informática y Análisis Numérico

Práctica 1 - 1 Introducción al aprendizaje automático y Weka

Juan Carlos Fernández Caballero (jfcaballero@uco.es)

Introducción al Aprendizaje Automático (IAA)
3º de Grado de Ingeniería Informática
Especialidad en Computación
Curso 2019-2020

GRUPO DE INVESTIGACIÓN AYRNA
APRENDIZAJE **Y** **R**EDES **N**EURONALES **A**RTIFICIALES
uco.es/ayrna

Índice de contenidos

Introducción

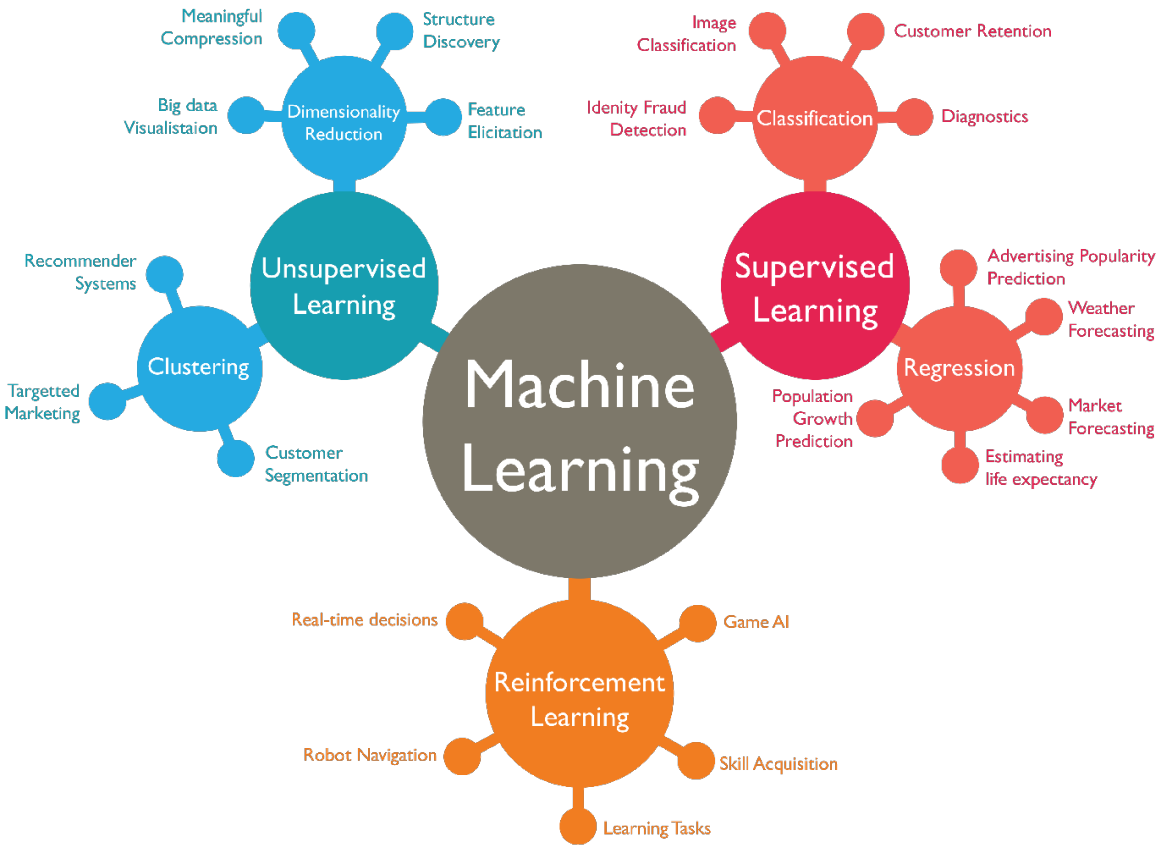
La investigación personal como herramienta principal

- Uno de los objetivos de esta asignatura y de estas prácticas es que sea **protagonista** activo en el **papel de la investigación**.
- La investigación es el acto de llevar a cabo **estrategias para descubrir algo**, y un conjunto de actividades de índole intelectual y experimental, con la intención de incrementar los conocimientos sobre un determinado asunto.
- Sin investigación no hay avance en Inteligencia Artificial ni aplicaciones de la misma.
- En esta asignatura usted debe ser lo más **autónomo** posible, buscando en la Web la información necesaria que le permita comprender las técnicas que use y cómo interpretarlas, y **aumentar sus conocimientos** en Aprendizaje Automático.

¿Qué es el aprendizaje automático?

- El **Aprendizaje Automático** (*Machine Learning*) es un subcampo de las Ciencias de la Computación y una **rama de la Inteligencia Artificial (IA)**.
- **Objetivo:** Desarrollar técnicas que permitan a las computadoras **aprender**, **inferir** información o **encontrar relaciones y conclusiones**.
- **¿Cómo aprenden?:** Mediante la generación de **modelos** matemáticos que generalizan comportamientos o mediante la **extracción de patrones de información y relaciones entre los datos**.
- **¿A partir de qué?:** A partir de una información suministrada en forma de **ejemplos por parte de un experto**, o a partir de **grandes cantidades de datos**.

Tipos de aprendizaje automático



El aprendizaje por refuerzo es un aprendizaje dinámico en base a errores y recompensas, donde puede no haber datos previos de un experto.

Utilidad del aprendizaje automático

Algunos ejemplos

- **Predicción** (Clasificación, Regresión, Agrupamiento).
  - Medicina: transplante hepático.
  - Predicción de polen y del tiempo.
  - Detectar correos spam.
  - Detección de olas extremas.
  - Sistemas de recomendación: Publicidad dirigida.
- **Optimización** (No es Aprendizaje Automático en cuanto a que no hay un modelo general final).
  - Juegos: Optimización del mejor movimiento de ajedrez.
  - Optimización de rutas: Viajante de comercio que debe recorrer N ciudades.
- **Reglas de asociación** (Tampoco hay un modelo general final).
  - Pretende obtener patrones de información y relaciones entre los datos.
  - Podría quedar fuera del Aprendizaje Automático, algunos autores lo encajan en el término **Mineria de Datos** (*Data Mining*) $\longrightarrow$ acoje al termino Aprendizaje Automático.
  - Usado también en *Big Data*: Cantidades enormes de datos estructurados y No estructurados.
  - Empresas de supermercados: Obtener información sobre el comportamiento de compra de sus clientes.

Tipos de aprendizaje automático

En las prácticas de IAA nos centraremos en estos tipos de Aprendizaje Automático:

- Aprendizaje **supervisado**.
- Aprendizaje **no supervisado.**

Aprendizaje supervisado

- **Clasificación** y **Regresión**. ¿Qué es cada cosa?
  - Dados una serie de datos de entrada **hay una salida determinada**, ya sea una **etiqueta** o un **valor numérico**.
- Creación de **modelos a partir de datos de entrenamiento**.
- Capacidad de **generalizar a partir de los datos no vistos previamente** con los datos de entrenamiento.

Tipos de aprendizaje automático

Aprendizaje NO supervisado (Agrupamiento o *Clustering*)

- Dados una serie de datos de entrada, **no hay una salida determinada** (no hay etiqueta ni valor numérico).
- Objetivo: Asignar una etiqueta de clase a cada patrón de entrada.
- **No hay conocimiento *a priori*** por parte de un experto.
- Los modelos **descubren** en los datos de entrada:
  - Características.
  - Regularidades.
  - Correlaciones.
  - Comportamientos.
  - Categorías.
- Aplicaciones:
  - Agrupar genes y proteinas con similar funcionalidad.
  - Agrupar documentos para catalogarlos.
  - Determinar tipos de clientes similares en internet en función de comportamientos.

Tipos de aprendizaje automático

Aprendizaje NO supervisado (Reglas de asociación)

- Dados una serie de datos de entrada, **tampoco hay una salida determinada**.
- Objetivo: Extracción de patrones de información, de conocimientos interesantes y de relaciones a partir de grandes cantidades de datos.
- Estos patrones y relaciones pueden estar implícitos en los datos y ser no triviales, desconocidos y potencialmente útiles.
- Las reglas de asociación se utilizan en la **Minería de Datos** y ***Big Data***.

Tipos de aprendizaje automático

Aprendizaje NO supervisado (Reglas de asociación)

- Un supermercado quiere obtener información sobre el comportamiento de compra de sus clientes.
- Se piensa que de esta manera se puede mejorar el servicio, colocando ciertos productos juntos, etc.
- ¿Qué productos suelen ir juntos en las cestas de la compra? ¿Qué probabilidad hay de que una persona que compre el producto A compre el producto B?

Tipos de aprendizaje automático

Aprendizaje NO supervisado (Reglas de asociación)

Id	Hue-vos	A-cei-te	Pa-ña-les	Vino	Leche	Man-te-qui-lla	Sal-món	Le-chu-gas	...
1	Si	No	No	Si	No	Si	Si	Si	...
2	No	Si	No	No	Si	No	No	Si	...
3	No	No	Si	No	Si	No	No	No	...
4	No	Si	Si	No	Si	No	No	No	...
5	Si	Si	No	No	No	Si	No	Si	...
6	Si	No	No	Si	Si	Si	Si	No	...
7	No	No	No	No	No	No	No	No	...
8	Si	Si	Si	Si	Si	Si	Si	No	...

Tipos de aprendizaje automático

Aprendizaje NO supervisado (Reglas de asociación)

Conocimiento obtenido ⇒ reglas de asociación:

- **SI** pañales, **ENTONCES** leche=si (100 %, 37 %)
- (a,b) = (precisión, cobertura)
  - Precisión ("*confidence*"): porcentaje de veces que la regla es correcta.
  - Cobertura ("*support*"): porcentaje de ocurrencia de la regla en los datos.
- Algunos analistas confirman que las empresas que adopten técnicas de analítica de *Machine Learning* y *Big Data* tendrán **una ventaja competitiva de 20 % en todas las métricas financieras** sobre sus competidores.

Introducción

Datasets

Validación

Weka

Datasets en Weka

a .excell

a .arff

Entregables

Bibliografía

Atributos, patrones y dataset (base de datos)

Dataset (Sinónimos: *base de datos*)

- **Colección de patrones y atributos** a partir de los cuales se aplicarán procedimientos de aprendizaje automático.

Patron	Atributo1	Atributo2	Atributo3	Atributo4	Atributo5	Salida/Clase
1	13.71	5.65	2.45	20.5	95.0	1.68/A
2	12.96	3.45	2.35	18.5	106.0	1.39/A
3	12.77	2.39	2.28	19.5	86.0	1.39/A
4	12.85	3.27	2.58	22.0	106.0	1.65/A
5	12.79	2.67	2.48	22.0	112.0	1.48/A
6	13.48	1.67	2.64	22.5	89.0	2.6/C
7	12.51	1.24	2.25	17.5	85.0	2.0/B
8	13.11	1.9	2.75	25.5	116.0	2.2/B
9	12.84	2.96	2.61	24.0	101.0	2.32/C
10	14.16	2.51	2.48	20.0	91.0	1.68/A
11	12.82	3.37	2.3	19.5	88.0	1.48/A
12	13.45	3.7	2.6	23.0	111.0	1.7/A
13	12.53	5.51	2.64	25.0	96.0	1.79/A
14	13.69	3.26	2.54	20.0	107.0	1.83/A
15	12.81	2.31	2.4	24.0	98.0	1.15/A
16	12.45	3.03	2.64	27.0	97.0	1.9/A

Atributos, patrones y dataset (base de datos)

Atributos (Sinónimos: *características*, *descriptores*, *variables de entrada*)

- **Nominales**: Son un conjunto (lista) de categorías.
  - Color: blanco, rojo, amarillo, verde, azul.
  - Tiempo: soleado, lluvioso, nublado.
- **Binarios**: Existen o no Existen (Subcaso de nominales).
  - Branquias: Ausencia de Branquias, Presencia de branquias
  - Estado de los ojos: Correcto, No correcto.
- **Numéricos**: Son valores reales (incluye enteros).
  - Altura en metros: 1.5, 1.75, 1.90.
  - Edad: 12, 45, 68.

Patrones (Sinónimos: *instancias*, *objetos*)

- **Colección** (posiblemente ordenada y estructurada) **de atributos** que representan un objeto.
- Cada patrón está representado por un conjunto de atributos, un vector de "d" dimensiones, llamado vector de atributos.

Dependiente Vs Independiente

- Una variable independiente es aquella cuyo valor no depende del de otra variable.
  - Todas las variables de entrada son independientes, ya que su valor viene dado en la base de datos y no se deriva de otras variables.
  - La i-ésima variable independiente se suele representar con x_i $\mathbf{x} = \{x_1, \ldots, x_K\}$ es el conjunto de las K variables independientes.
- Una variable dependiente es aquella cuyo valor depende del valor que toman otras variables.
  - En problemas de **clasificación** y **regresión**, la variable objetivo (o **variable de salida**) es la variable dependiente, porque su valor se **modela** en función de las variables independientes.
  - Se suele representar con y.

Categóricas Vs Cuantitativas

- Una variable categórica (o cualitativa) se refiere a características o cualidades que no pueden ser medidas con números.
  - Toman valor en un **conjunto de categorías**.
  - Hay dos tipos:
    - Categóricas nominales: las categorías no presentan un orden (p.ej. el estado civil, con las modalidades: soltero, casado, separado, divorciado y viudo).
    - Categóricas ordinales: las categorías presentan un orden (p.ej. la nota en un examen, suspenso, aprobado, notable, sobresaliente; puesto conseguido en una prueba deportiva: primero, segundo, tercero; ...).

Categóricas Vs Cuantitativas

- Una variable cuantitativa (o numérica) es la que se expresa mediante un número, por lo que se pueden realizar operaciones aritméticas con sus valores.
  - Toman valor en un intervalo.
  - Aunque a menudo se tratan de la misma forma, hay dos tipos de variables cuantitativas:
    - Cuantitativa discreta: es aquella que toma valores aislados, es decir no admite valores intermedios entre dos valores específicos (p.ej. el número de hijos: 0, 1, 2, 3, 4, ...)
    - Cuantitativa continua: es aquella que puede tomar infinitos valores (p.ej. la altura de una persona: 1.60m, 1.65m, ...).

Tratamiento de variables categóricas

- La mayoría de **algoritmos de aprendizaje automático** necesitan que las **variables categóricas se conviertan a cuantitativas**, para poder aplicar operaciones aritméticas sobre las mismas.
- Tipo de animal $\in$ {mamífero, reptil, ave}
- **Posible conversión**: 1 variable categórica discreta con **tres valores**.

Tipo	mamífero	reptil	ave
Valores	1	2	3

- Inconvenientes:
  - Se ha **asumido** que existe un **orden entre las categorías** (reptil>mamífero, ave>reptil, ...).
  - Se ha **asumido** una **distancia** entre cada una de las categorías.

Tratamiento de variables categóricas

Tipo	Nº Extrem.	Peso	...
mamífero	4	20	...
mamífero	4	15	...
reptil	0	5	...
ave	2	0.5	...
reptil	4	2	...
...	...	...	...

Tipo	Nº Extrem.	Peso	...
1	4	20	...
1	4	15	...
2	0	5	...
3	2	0.5	...
2	4	2	...
...	...	...	...

Tratamiento de variables categóricas

Tipo	Nº Extrem.	Peso	...
mífero	4	20	
ro	4	15	
	0		
av	2		...
reptil			...
...	...	...	...

Tipo	N		...
1			
	4	15	
	0	5	
	2	0.5	...
2	4	2	...
...	...	...	...

Tratamiento de variables categóricas

- Binarización de variables categóricas a través de la representación 1-de-k, donde k es el número de categorías.
  - La variable categórica genera k variables binarias.
  - Por elemento de la base de datos (o patrón), la i-ésima variable binaria será igual a 1 si el patrón es de esa categoría y a 0 si no lo es.
  - Es decir, todas las variables binarias serán 0, salvo aquella que corresponda a la categoría del patrón (en la que habrá un 1).
- Binarización de la variable anterior:

Tipo	mamífero	reptil	ave
1-de-k	$\{1,0,0\}$	$\{0,1,0\}$	$\{0,0,1\}$

Tratamiento de variables categóricas

Tipo	Nº Extrem.	Peso	...
mamífero	4	20	...
mamífero	4	15	...
reptil	0	5	...
ave	2	0.5	...
reptil	4	2	...
...	...	...	...

Tipo= mamífero	Tipo= reptil	Tipo= ave	Nº Extrem.	Peso	...
1	0	0	4	20	...
1	0	0	4	15	...
0	1	0	0	5	...
0	0	1	2	0.5	...
0	1	0	4	2	...
...	...	...	...	...	...

Métodos de validación

Patron	Atributo1	Atributo2	Atributo3	Atributo4	Atributo5	Salida/Clase
1	13.71	5.65	2.45	20.5	95.0	1.68/A
2	12.96	3.45	2.35	18.5	106.0	1.39/A
3	12.77	2.39	2.28	19.5	86.0	1.39/A
4	12.85	3.27	2.58	22.0	106.0	1.65/A
5	12.79	2.67	2.48	22.0	112.0	1.48/A
6	13.48	1.67	2.64	22.5	89.0	2.6/C
7	12.51	1.24	2.25	17.5	85.0	2.0/B
8	13.11	1.9	2.75	25.5	116.0	2.2/B
9	12.84	2.96	2.61	24.0	101.0	2.32/C
10	14.16	2.51	2.48	20.0	91.0	1.68/A
11	12.82	3.37	2.3	19.5	88.0	1.48/A
12	13.45	3.7	2.6	23.0	111.0	1.7/A
13	12.53	5.51	2.64	25.0	96.0	1.79/A
14	13.69	3.26	2.54	20.0	107.0	1.83/A
15	12.81	2.31	2.4	24.0	98.0	1.15/A
16	12.45	3.03	2.64	27.0	97.0	1.9/A

- Antes de construir un modelo, se divide el conjunto de datos disponible en:
  - Un conjunto de **entrenamiento** o *training* (para construirlo).
  - Un conjunto de **generalización** o *testing* (para evaluarlo).
- Comparando los patrones etiquetados del conjunto de generalización con el resultado de aplicar el modelo,se obtienen métricas de calidad y redimiento como el **porcentaje de patrones correctamente clasificados**.

Métodos de validación

Sobreaprendizaje

- El conjunto de generalización debe ser siempre independiente del conjunto de entrenamiento.
- En ocasiones los modelos tienden a ajustarse demasiado al conjunto de entrenamiento utilizado en su construcción, provocando lo que se conoce como (**sobreaprendizaje**), lo que los hace menos útiles para clasificar nuevos datos.
- El error de clasificación en el conjunto de entrenamiento NO es un buen estimador del rendimiento del modelo, hay que usarlo en el **conjunto de generalización**.

Métodos de validación

Dos formas de crear conjuntos de entrenamiento y generalización

- Validación cruzada (*crossvalidation*): 1) *K-fold* y 2) *N-Holdout*.
- Wikipedia - Consultar enlace
- Suele hacerse de forma **estratificada** (proporcional) en cuanto al número de patrones de cada clase, es decir, cada fold debería tener patrones de todas las clases en función de su proporción sobre el total de patrones.
- Weka dispone de unos "**filtros**" para crear conjuntos *K-fold* y *N-Holdout*.

Métodos de validación

Validación cruzada mediante *K-fold*:

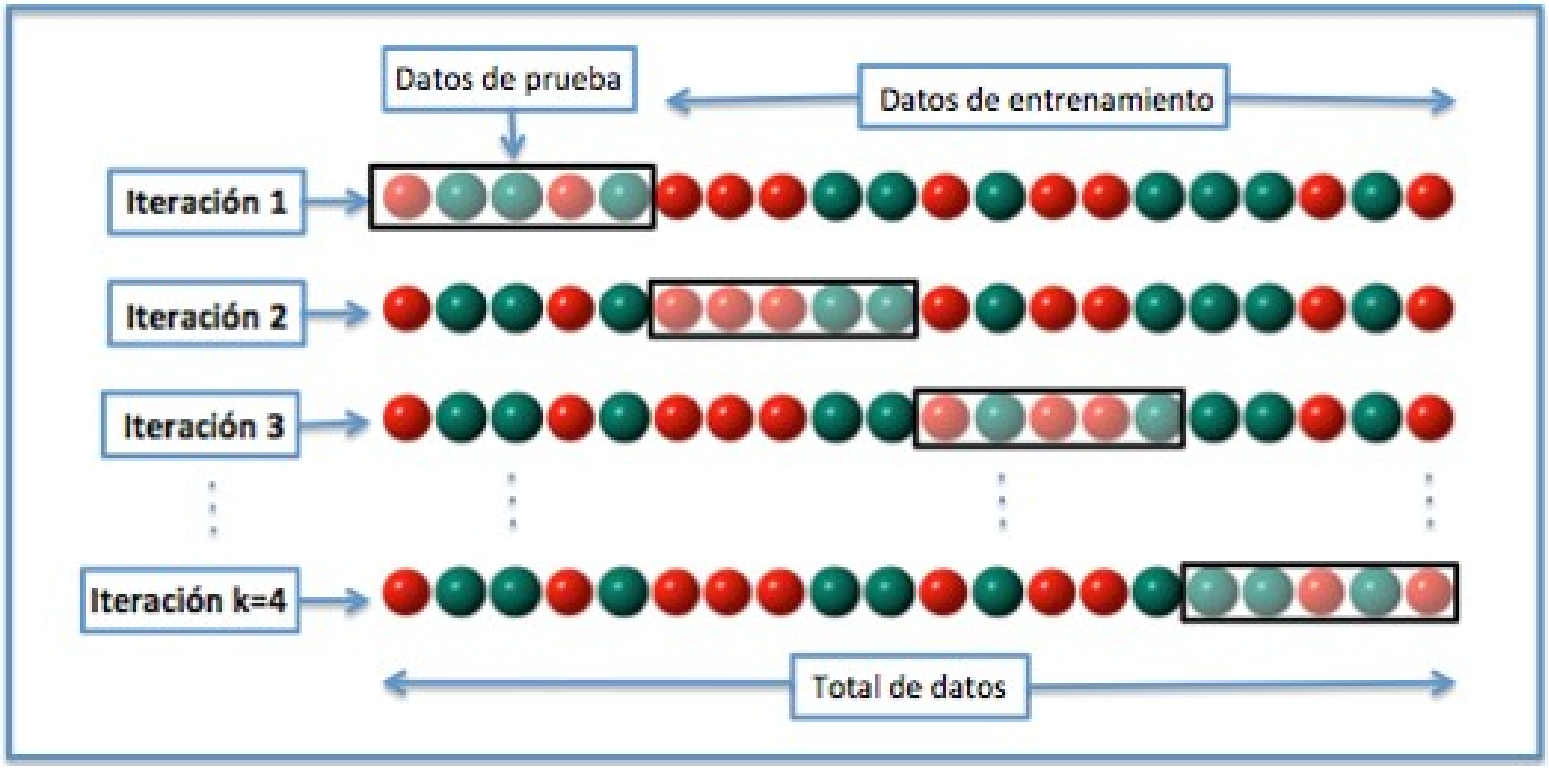


Figura: Validación cruzada mediante *K-fold*

Métodos de validación

Validación cruzada mediante un *hold-out:*

- Suele hacerse aleatoriamente y de forma estratificada, por ejemplo 75 % de patrones para training y 25 % para testing.
- Esa operación se hace N veces, dando lugar a un *N-holdout*.

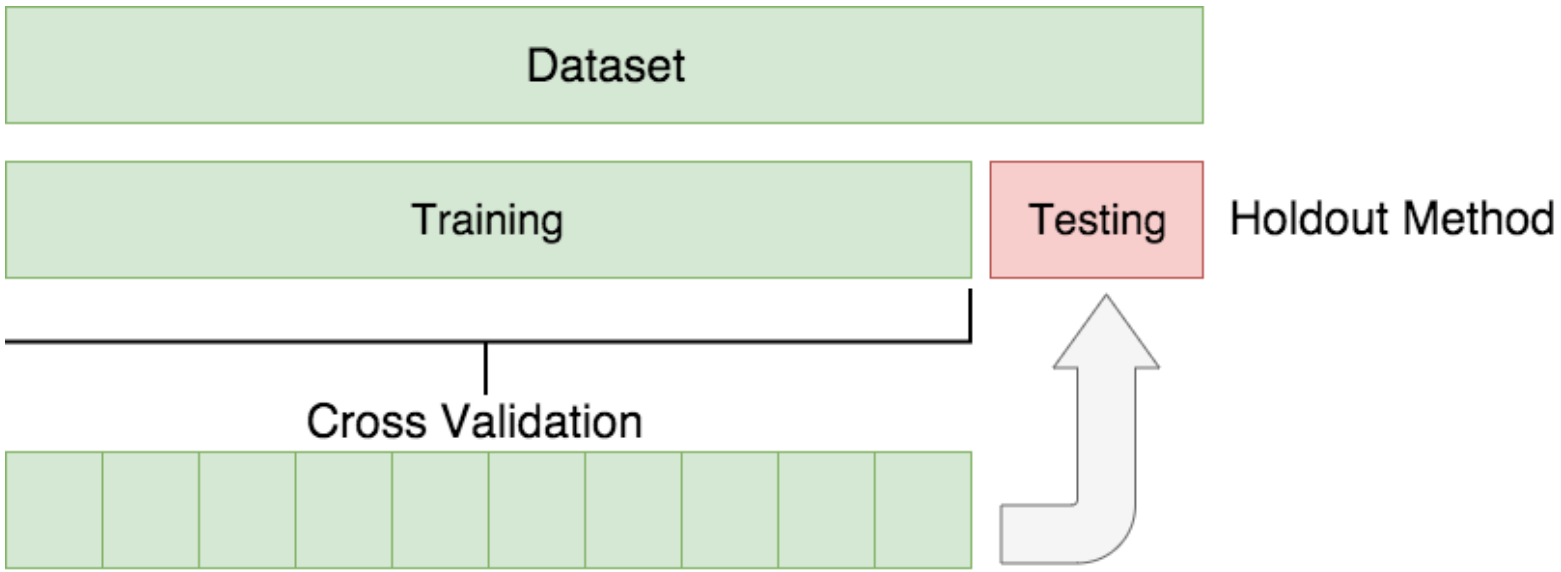


Figura: Validación cruzada mediante un *hold-out*

El software Weka

- Herramienta de *Machine Learning* desarrollada por la Universidad de Waikato, Nueva Zelanda [1].
- Multiplataforma.
- Tres formas de operar: GUI, línea de comandos y API de Java.
- **Waikato Web**

Weka

- Weka incorpora algoritmos que son capaces de aprender a partir de datos.
- Información extraída automáticamente de los datos, mediante métodos estadísticos o computacionales usados en dichos algoritmos.
- Weka tiene algoritmos de aprendizaje **supervisado** y **no supervisado**.
- Tómese su tiempo y estudie los manuales de Weka disponibles en Moodle para aprender a discernir entre los distintos módulos que posee Weka y para qué sirve cada uno de ellos.

Herramientas de las que dispone Weka

- Métodos para el **preprocesamiento** de los datos.
  - Métodos de **selección de atributos o características**.
- Algoritmos de **clasificación/regresión**.
- Algoritmos de **agrupamiento**.
- Algoritmos para encontrar **reglas de asociación.**
- **Visualización** de los datos.

GUI o módulos de Weka

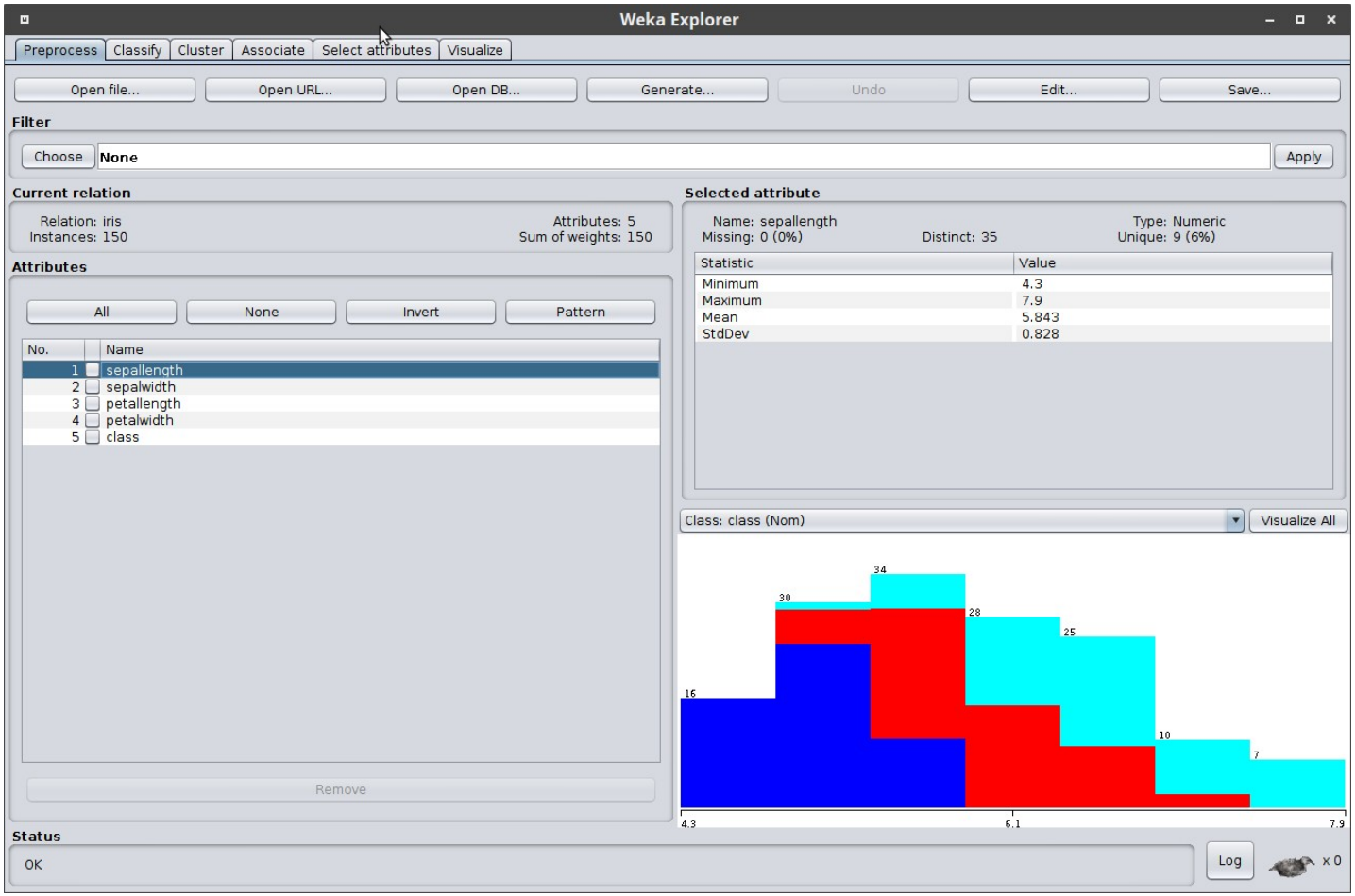
Weka tiene cuatro módulos para el usuario para poder usar las herramientas comentadas anteriormente:

- **Explorer**: Para preprocesamiento de datos y experimentos usando algoritmos de todo tipo.
- **Experimenter**: Realiza experimentos como en el entorno Explorer, pero a gran escala y permitiendo comparaciones en los resultados.
- **Knowledge Flow**: Entorno que permite arrastrar componentes de Weka y conectarlos para hacer experimentos.
- **Simple CLI**: Interfaz simple de comandos.

En las prácticas se usarán los módulos **Explorer** y **Experimenter**.

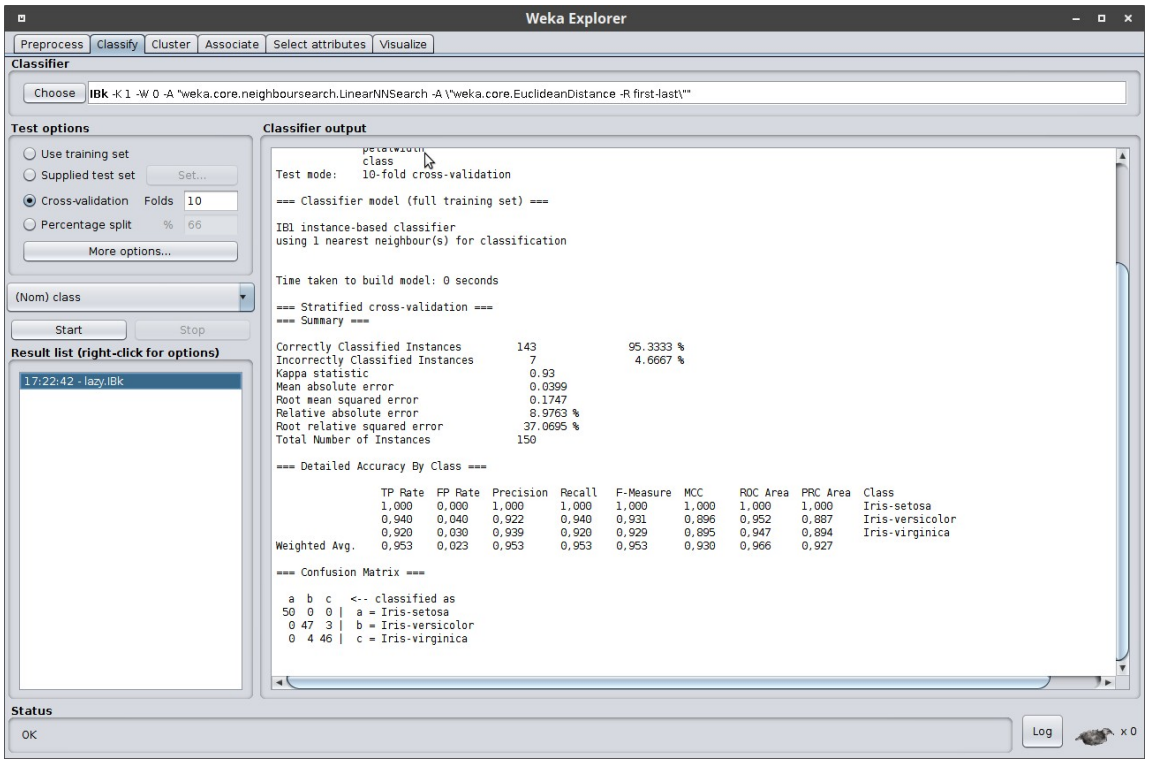
Entorno Explorer $\longrightarrow$ Preprocess

- Entorno que permite cargar una base de datos y tratar sus atributos y clase.
- Permite el uso de múltiples tipos de filtros (**importante**).



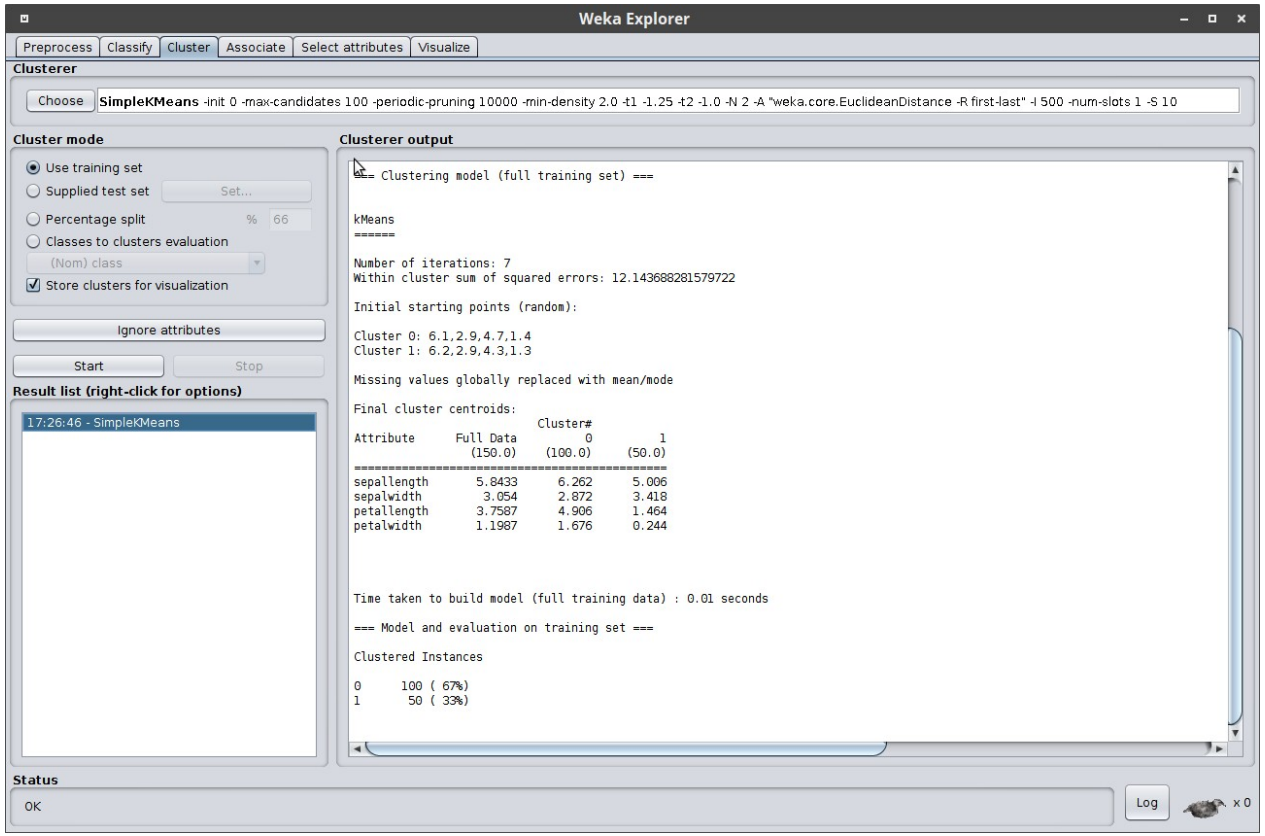
Entorno Explorer $\longrightarrow$ Classify

- Entorno que permite seleccionar algoritmos de clasificación/regresión y aplicarlos sobre una base de datos cargada.
- Permite evaluar los modelos mediante varias opciones: Usando el conjunto de entrenamiento, aportando conjunto de test, haciendo validación cruzada (*k-fold*) o haciendo un *hod-out* (*percentage split*).



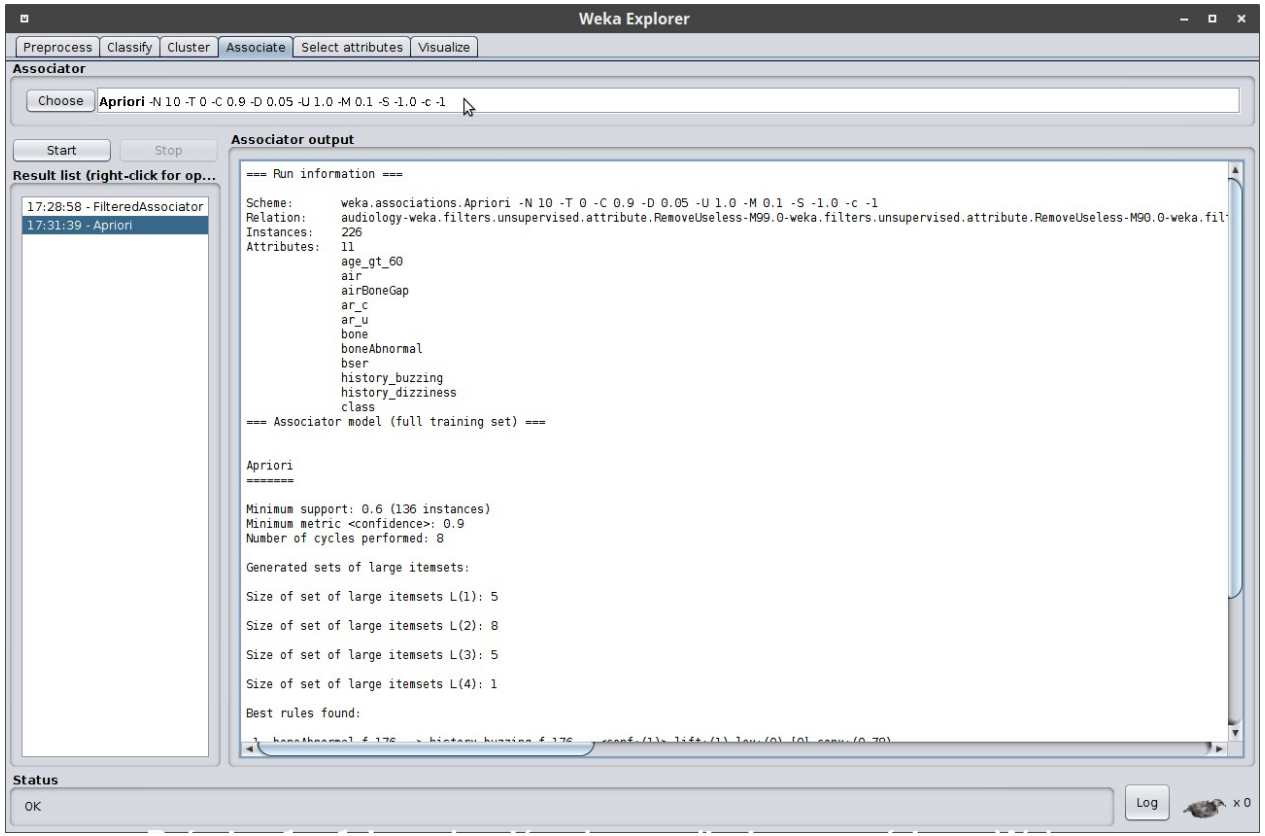
Entorno Explorer $\longrightarrow$ Cluster

- Entorno que permite realizar aprendizaje no supervisado (*Clustering*) mediante la aplicación de diversos algoritmos.
- Permite elegir que atributos formarán parte del *Clustering*.
- Este entorno se usará en otra práctica.



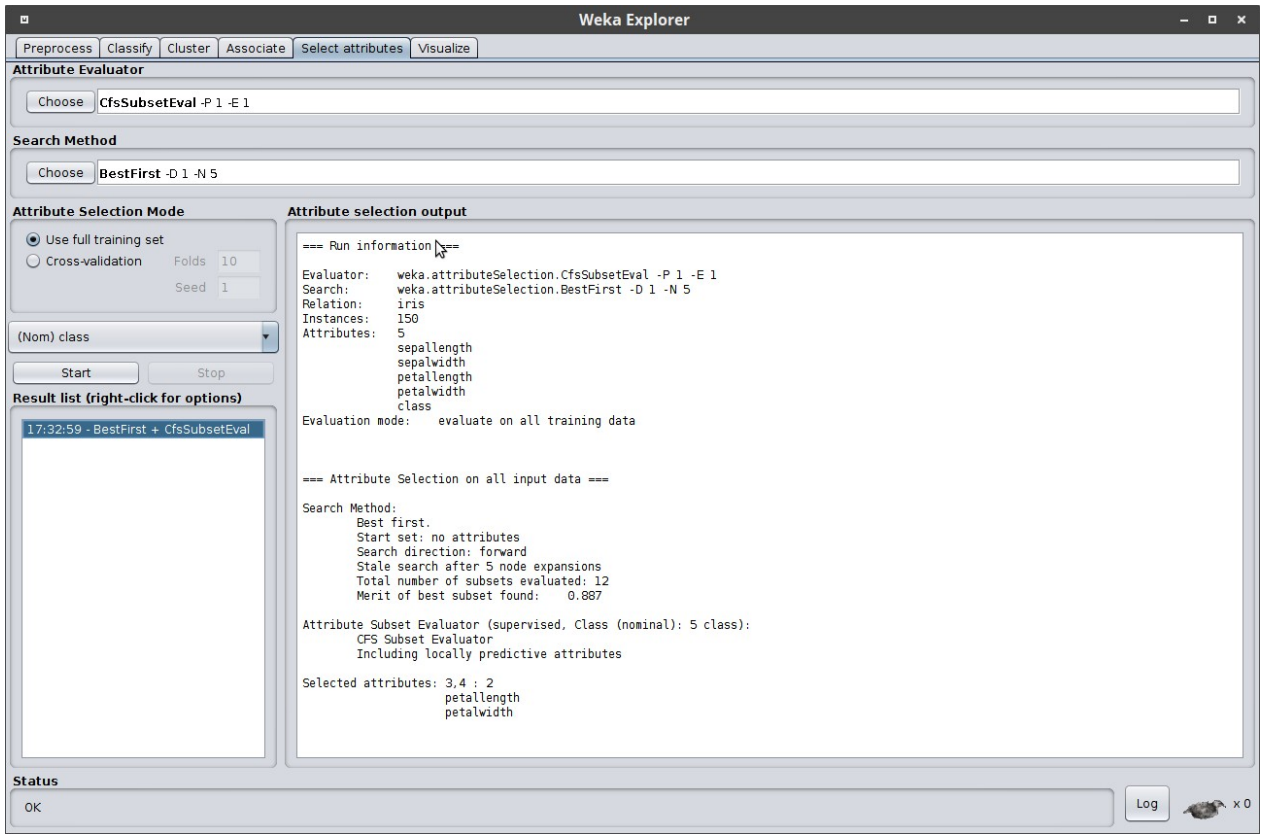
Entorno Explorer $\longrightarrow$ Associate

- Entorno que permite aplicar métodos orientados a buscar asociaciones entre datos.
- *A priori* este entorno no lo usaremos en las prácticas y está más relacionado con el termino **Minería de Datos** y ***Big Data***, aunque encontrará literatura donde al aprendizaje supervisado y no supervisado también lo incluyen en el termino Minería de Datos.



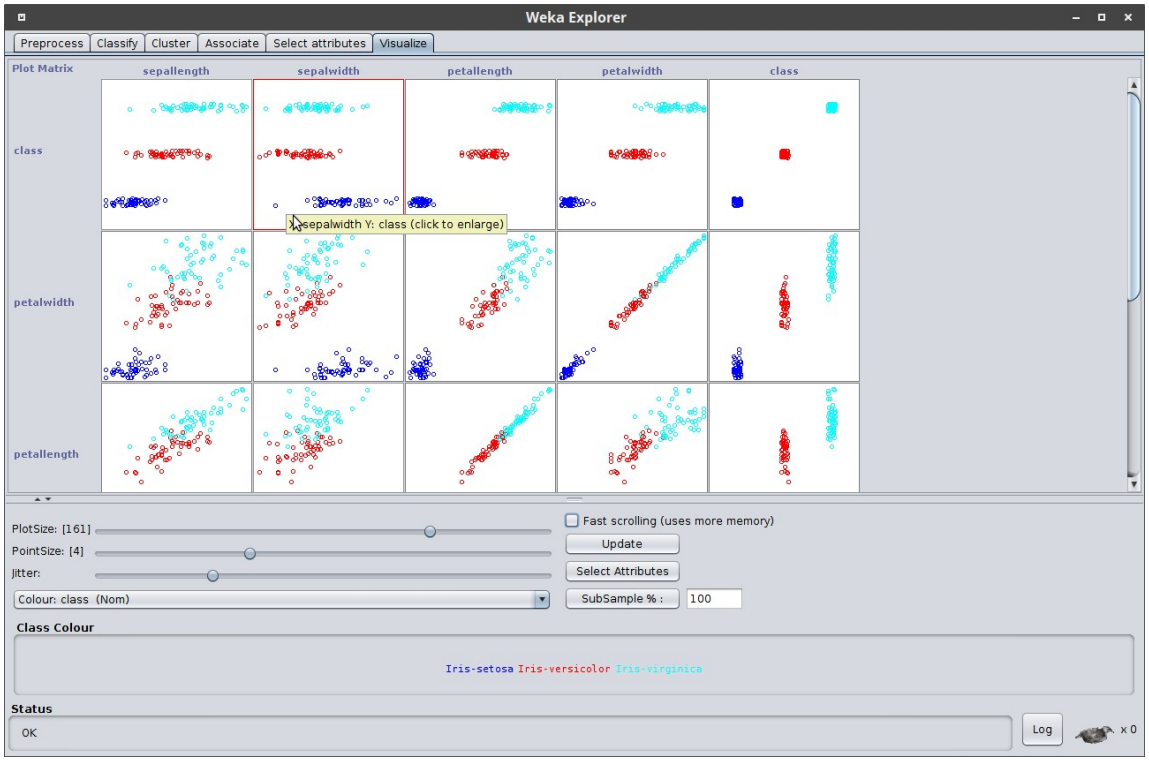
Entorno Explorer⟶Select Attributes

- Entorno que permite identificar qué atributos pueden ser más importantes a la hora de discriminar una clase, y cómo pueden influir unos y otros sobre ello.
- Este entorno también es importante en el preprocesado de datos y selección de características.



Entorno Explorer $\longrightarrow$ Visualize

- Este entorno muestra gráficamente la distribución de todos los atributos, mostrando gráficas en dos dimensiones.
- En los ejes de las gráficas se representan todos los posibles pares de combinaciones de los atributos.
- Permite ver correlaciones y asociaciones entre los atributos de una forma gráfica.



Atributos, patrones y dataset (base de datos) en Weka

Las bases de datos en Weka se almacenan en ficheros `.arff`, que son ficheros de texto con el siguiente aspecto:

```
@relation nombe_relación

@attribute nombre_atributo_1 tipo_atrtibuto_1
@attribute nombre_atributo_2 tipo_atrtibuto_2
@attribute nombre_atributo_3 tipo_atrtibuto_3
@attribute nombre_atributo_4 tipo_atrtibuto_4

@data
Fila_de_datos_1
Fila_de_datos_2
Fila_de_datos_3
Fila_de_datos_4
...
```

Desde la Web de Weka: https://waikato.github.io/weka-wiki/formats_and_processing/arff_stable/

Formato de los ficheros de entrada

Atributos soportados: numéricos, nominales, *string*, *date*.

```
@relation enfermedad_corazon

@attribute edad numeric
@attribute sexo {hombre, mujer}
@attribute tipo_dolor_toracico {esofagico, pleuropulmonar, osteomuscular,
    neuritico}
@attribute colesterol numeric
@attribute clase {si, no}

@data
63,hombre,esofagico,233,no
67,hombre,neuritico,286,si
67,mujer,pleuropulmonarasympt,229,no
38,hombre,pleuropulmonar,?,si
...
```

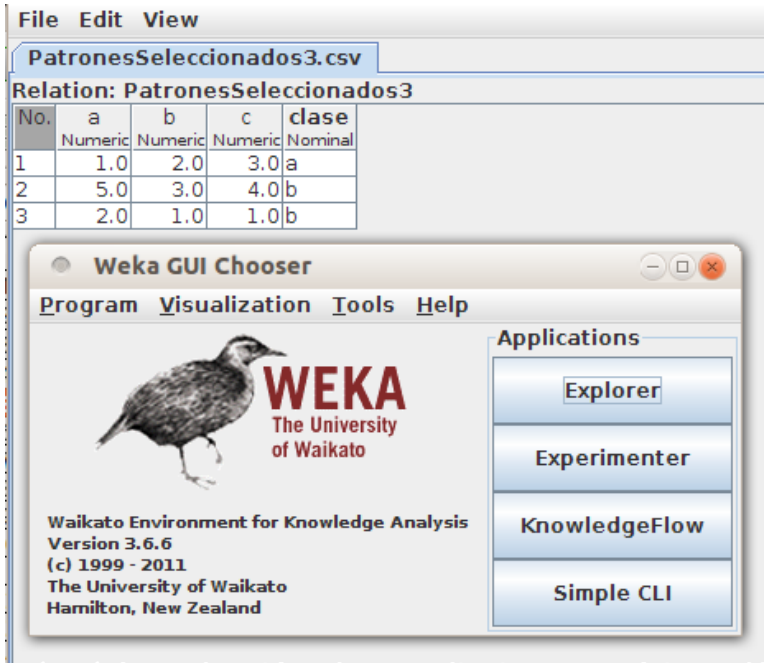
ATRIBUTO NUMÉRICO

ATRIBUTO NOMINAL

VALOR PERDIDO

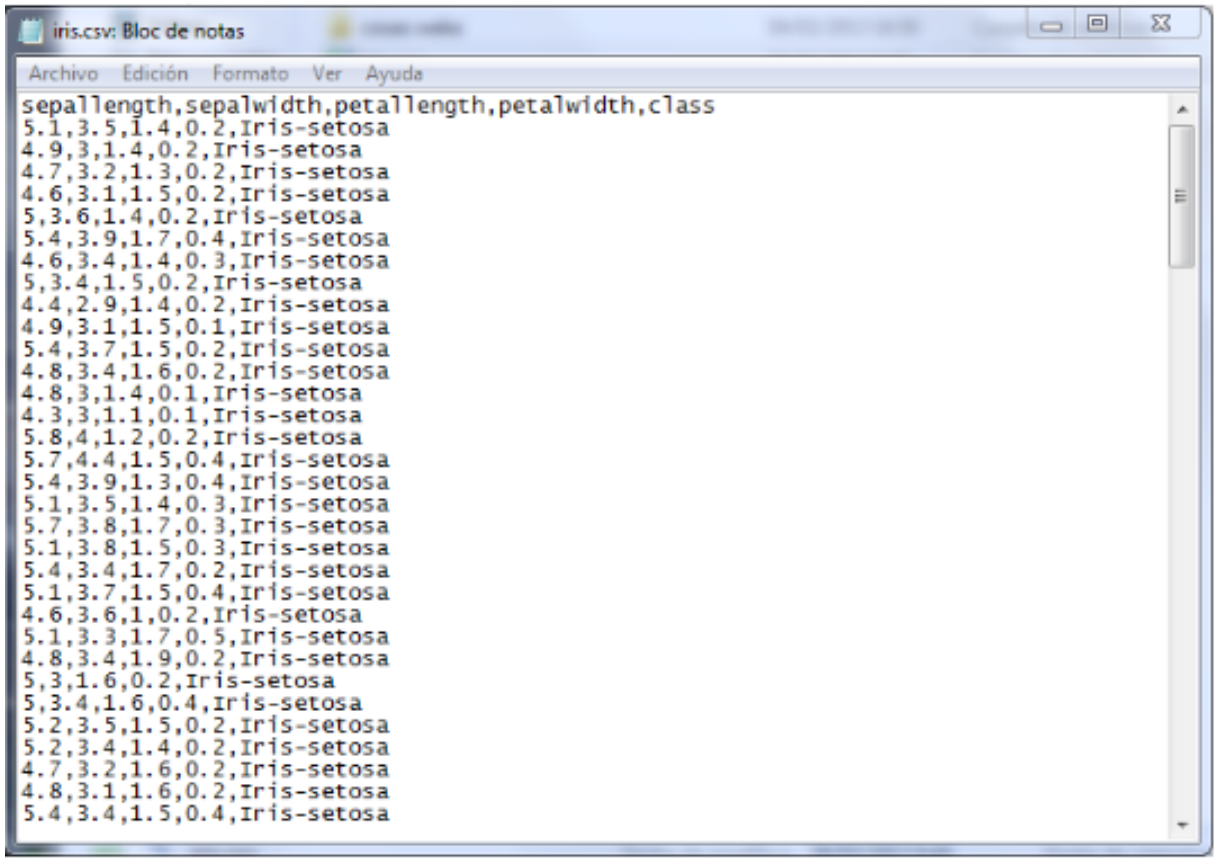
Bases de datos en formato .csv

- Muchas de las bases de datos disponibles en la red se encuentran en formato `.csv`.
- Además de usar ficheros `.arff`, Weka también permite usar ficheros `.csv`, aunque hay que hacerlo con precaución.
- Se pueden visualizar los ficheros con un editor de texto o mediante `Tools`→`ArffViewer`.



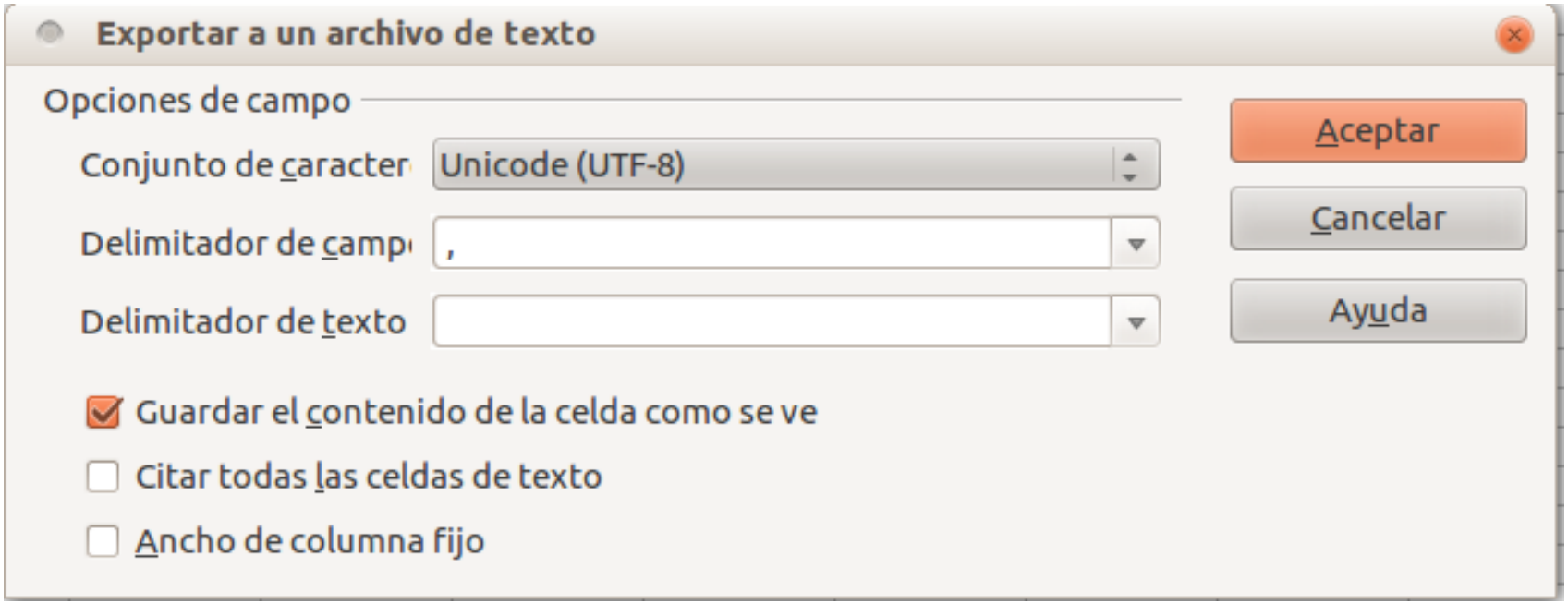
Extensiones comúnmente utilizadas

Formato .csv (*comma separated values*) soportado por Weka.



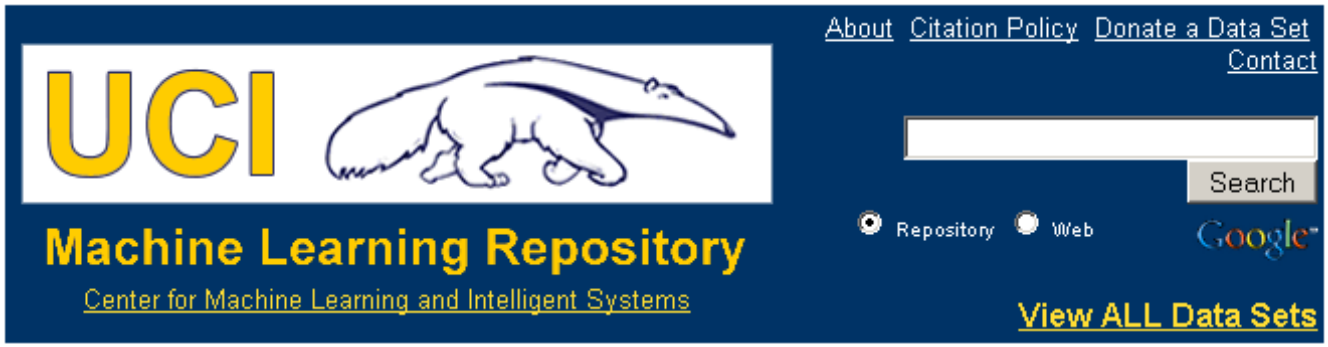
Exportar a CSV

Todas las hojas de cálculo soportan exportar sus datos a `.csv`, formato que si que se puede abrir desde Weka y transformar despues a `.arff`.



Importar a Excel

- Bajamos los datos de la base de datos `annealing` de la UCI.



Annealing Data Set

Download: Data Folder, Data Set Description

Abstract: Steel annealing data

Data Set Characteristics:	Multivariate	**Number of Instances:**	798	**Area:**	Physical
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	38	**Date Donated**	N/A
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	70570

Source:

Importar a Excel

- `anneal.names` contiene información sobre los atributos.
- `anneal.data` contiene los datos propiamente dichos:

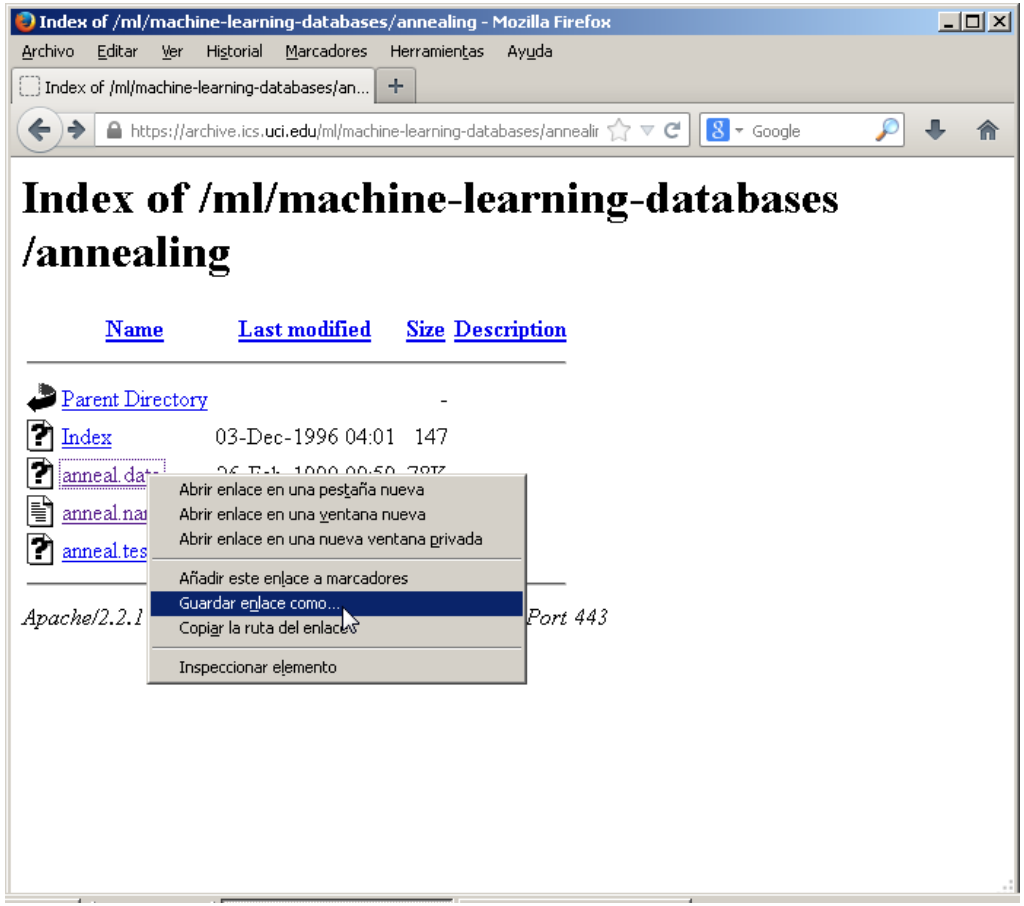
Index of /ml/machine-learning-databases /annealing

Name	Last modified	Size	Description
Parent Directory		-	
Index	03-Dec-1996 04:01	147	
anneal.data	26-Feb-1990 09:59	78K	
anneal.names	15-Mar-1990 09:17	2.7K	
anneal.test	26-Feb-1990 10:00	9.8K	

Apache/2.2.15 (CentOS) Server at archive.ics.uci.edu Port 443

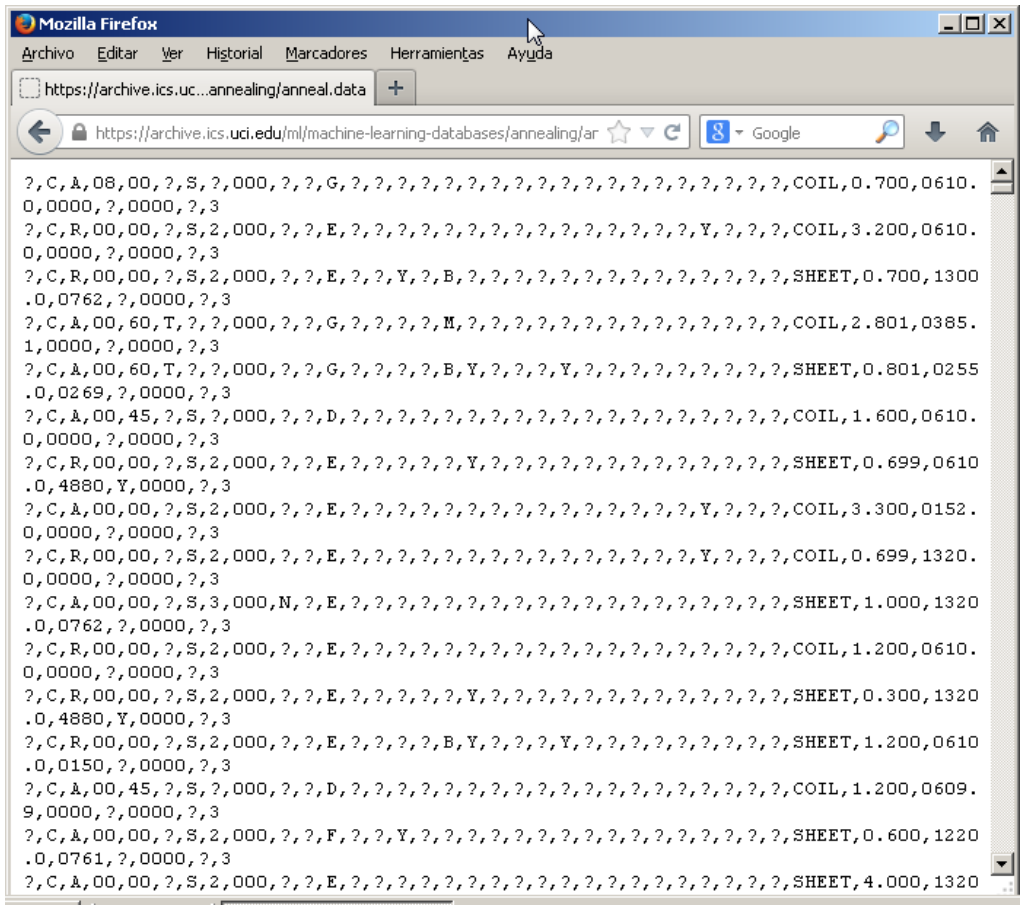
Importar a Excel

- Guardamos los datos en el escritorio:



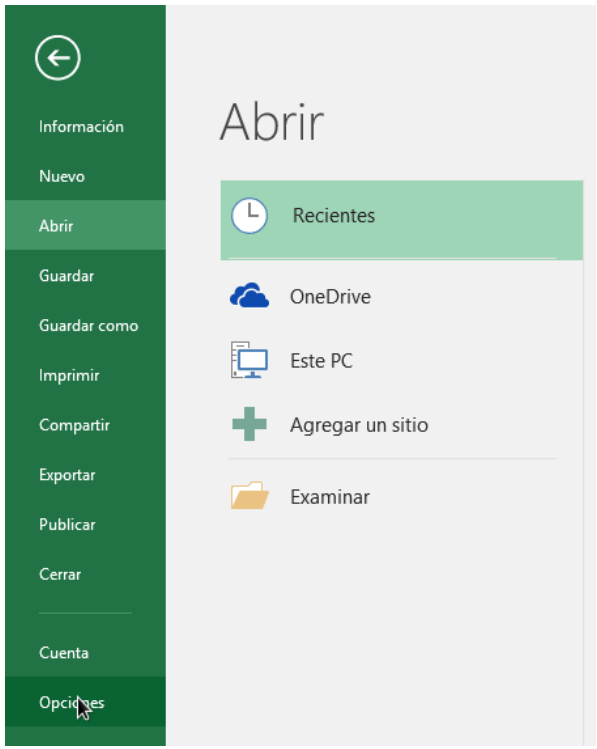
Importar a Excel

- Echamos un vistazo a los datos:

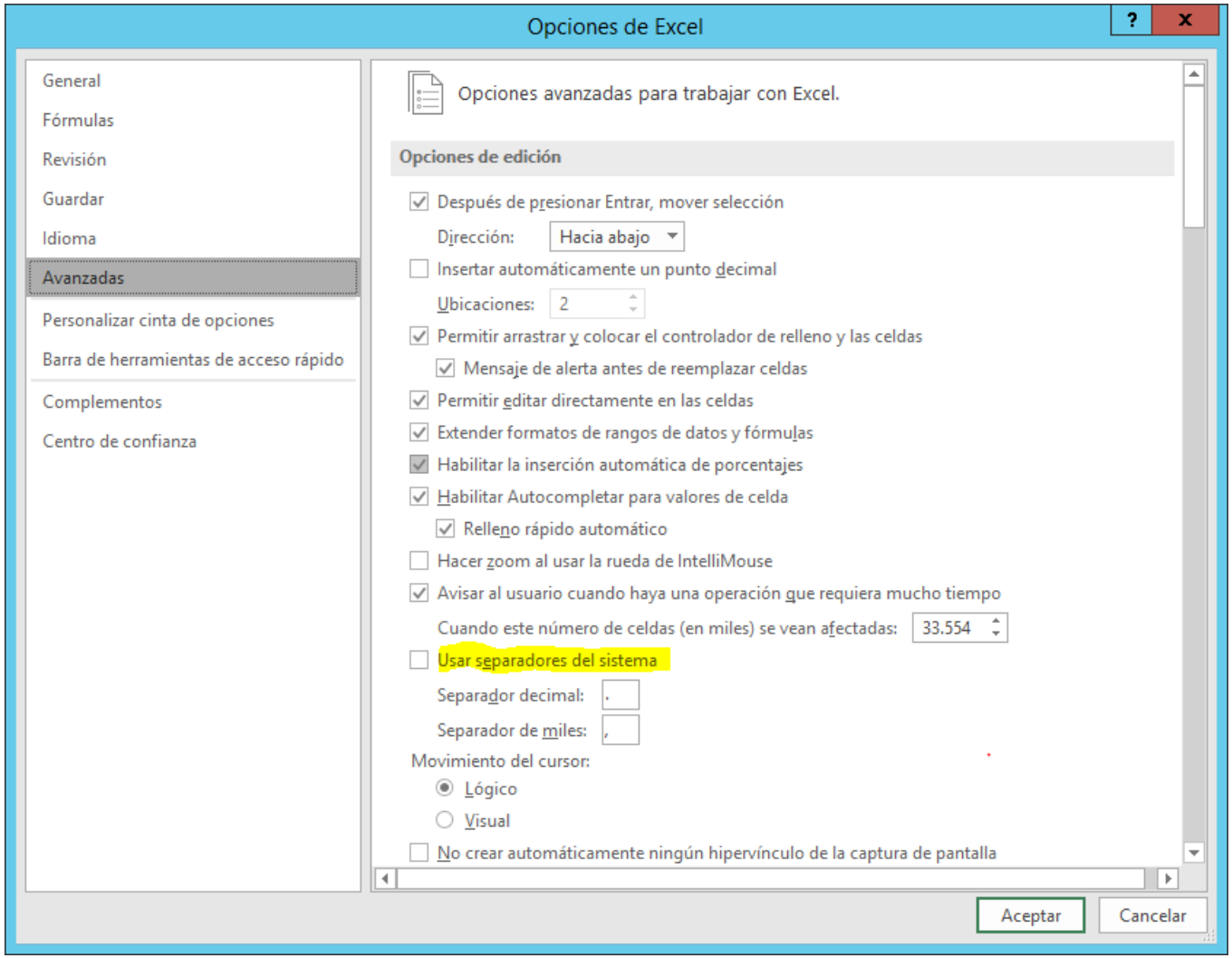


Importar a Excel (separador decimal)

- Ahora vamos a trabajar con Excel.
- Cambiamos la configuración regional para que se utilice el "." como separador decimal y la "," como separador de miles:

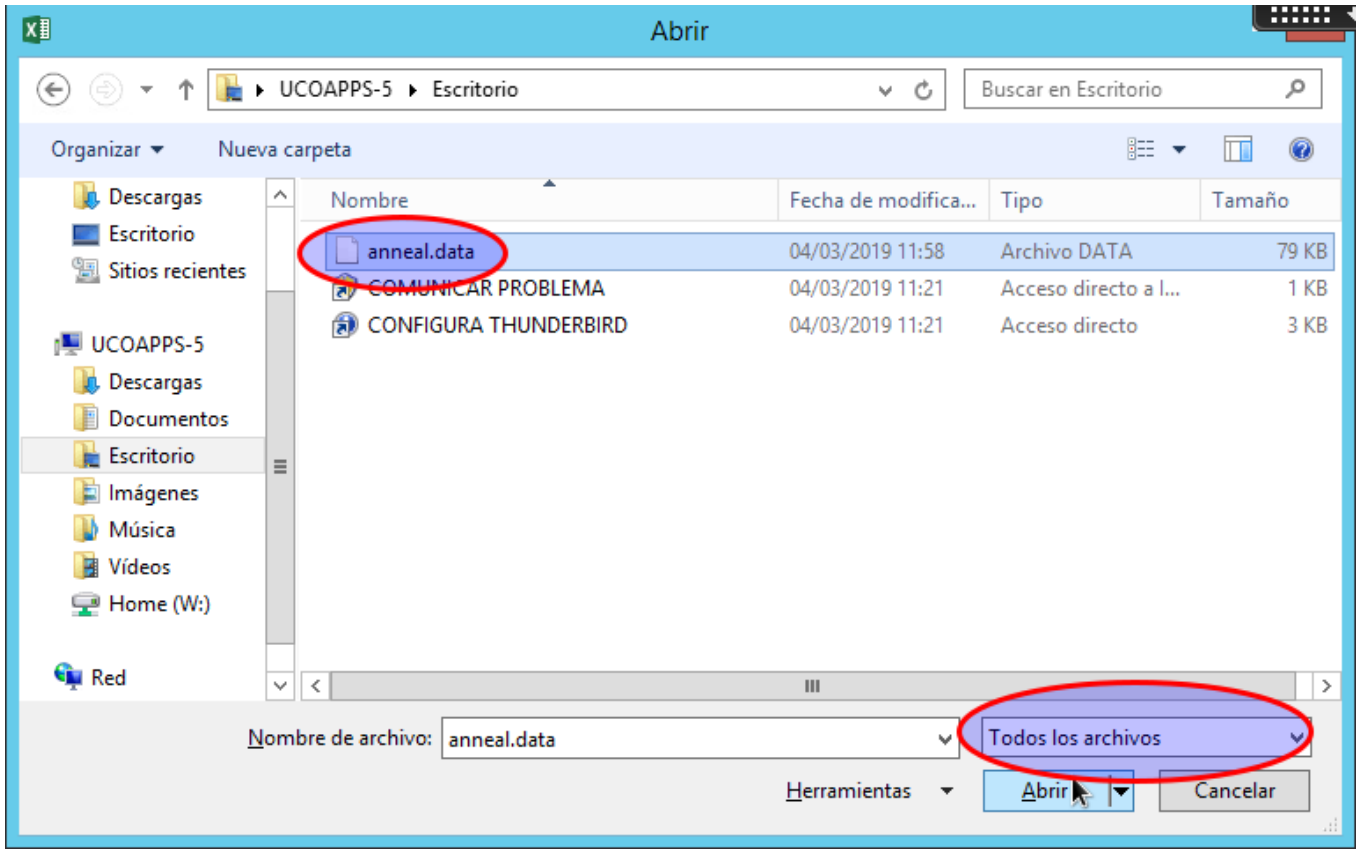


Importar a Excel (separador decimal)



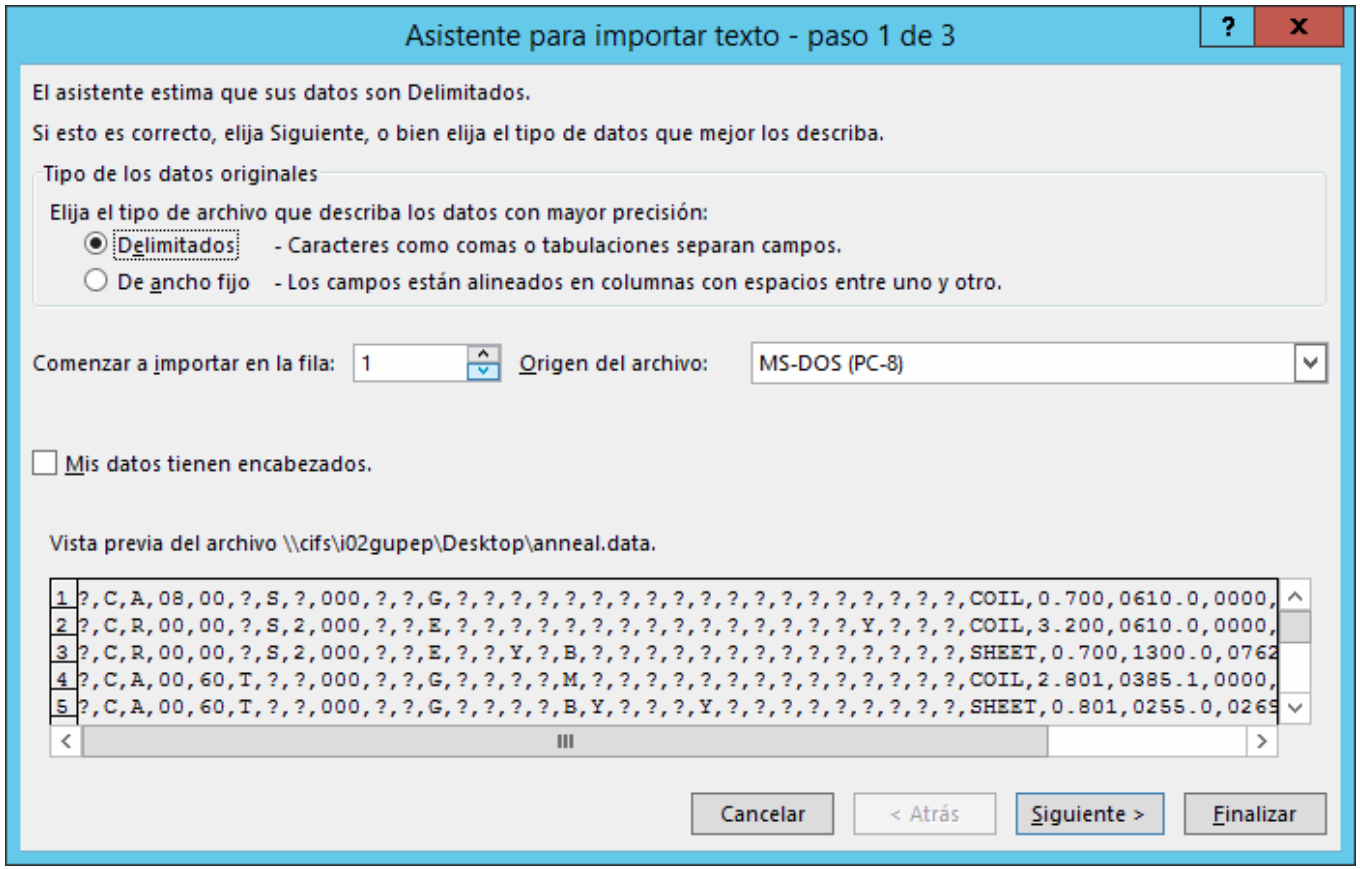
Importar a Excel

- Abrimos el archivo desde Excel (seleccionar "`Todos los archivos`"):



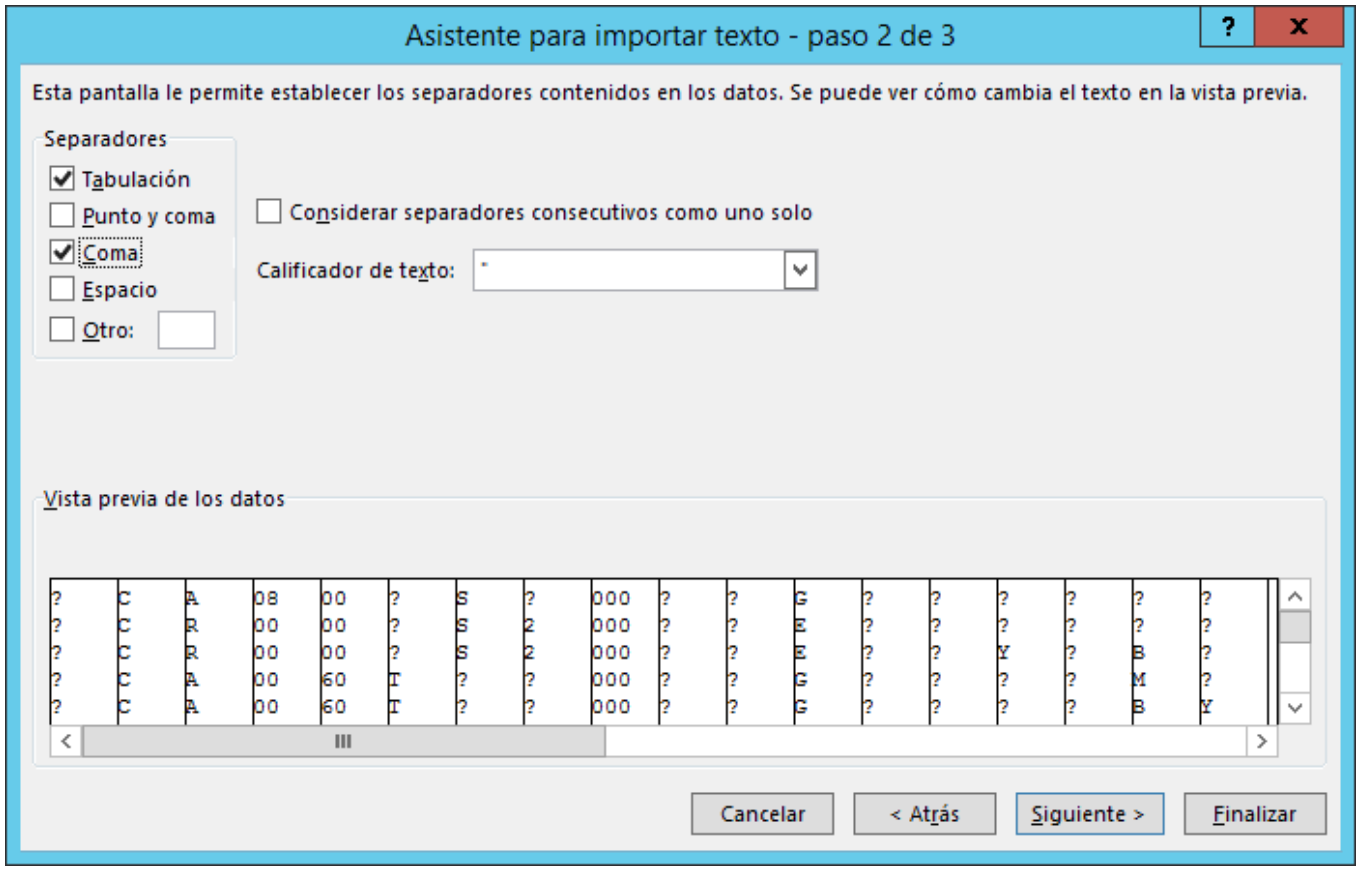
Importar a Excel

- Se invoca el asistente de conversión de texto:



Importar a Excel

- Es importante elegir correctamente el separador de campos (en este caso, es la “,”).

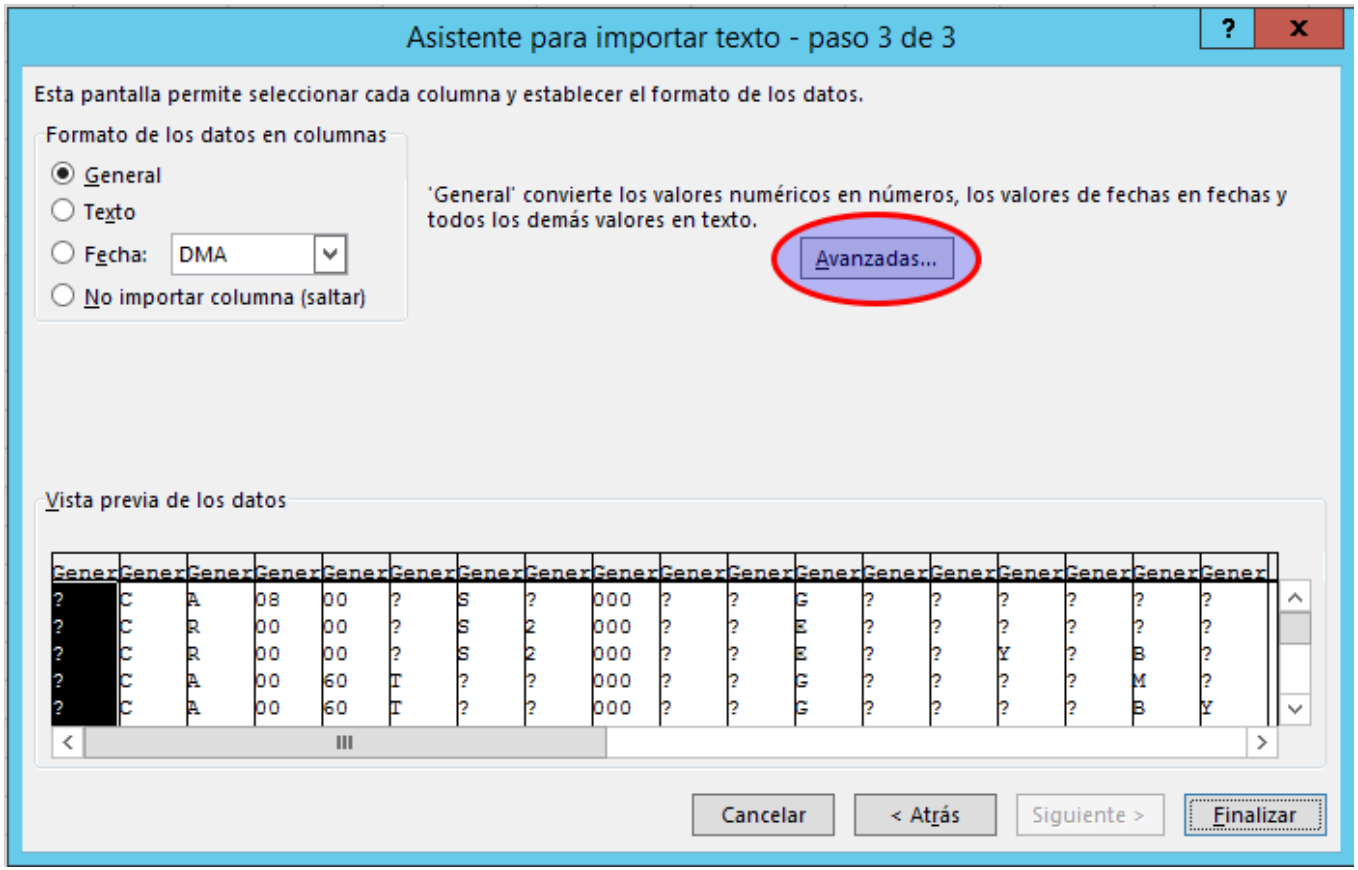


Importar a Excel (separador decimal)

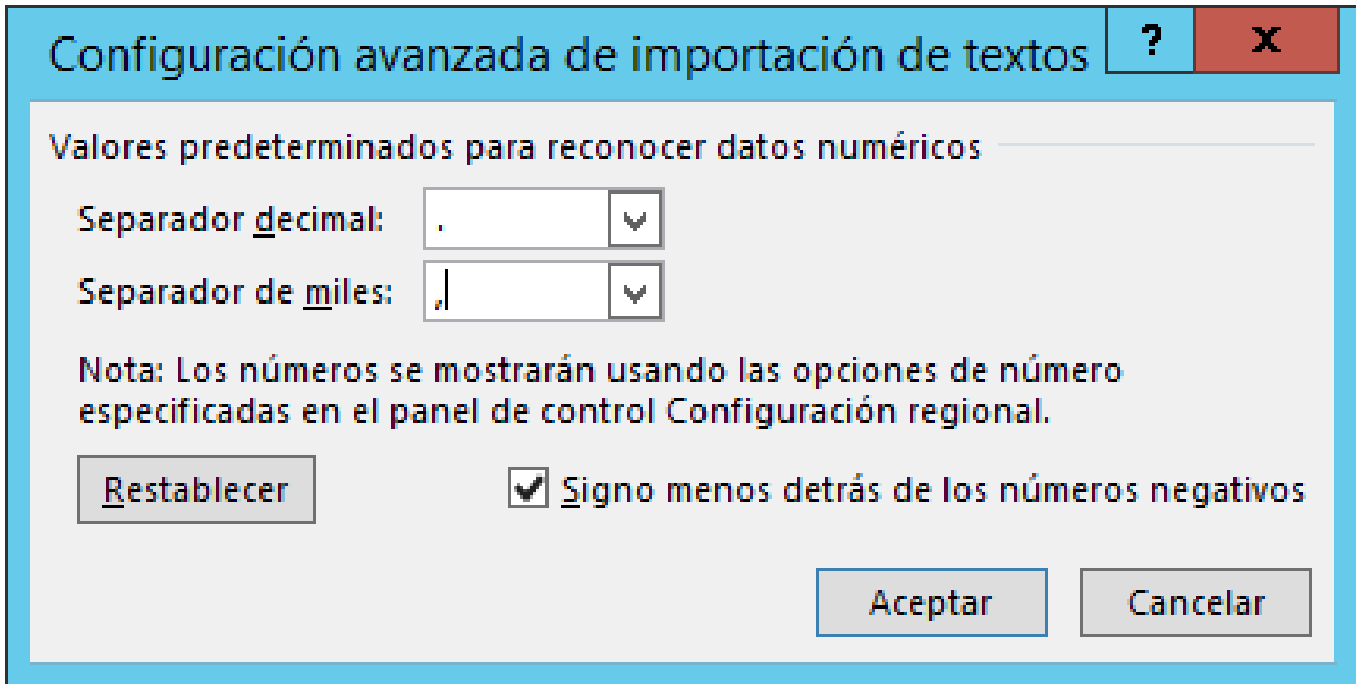
- Cuando nos traemos datos de fuentes externas, es habitual que los decimales se expresen con “.” (punto) en lugar de “,”.
- Por ello, hay que seleccionar el separador decimal.
- Ídem para el separador de miles, que en fuentes externas suele ser la “,”.

Importar a Excel

- Pulsar Avanzadas.

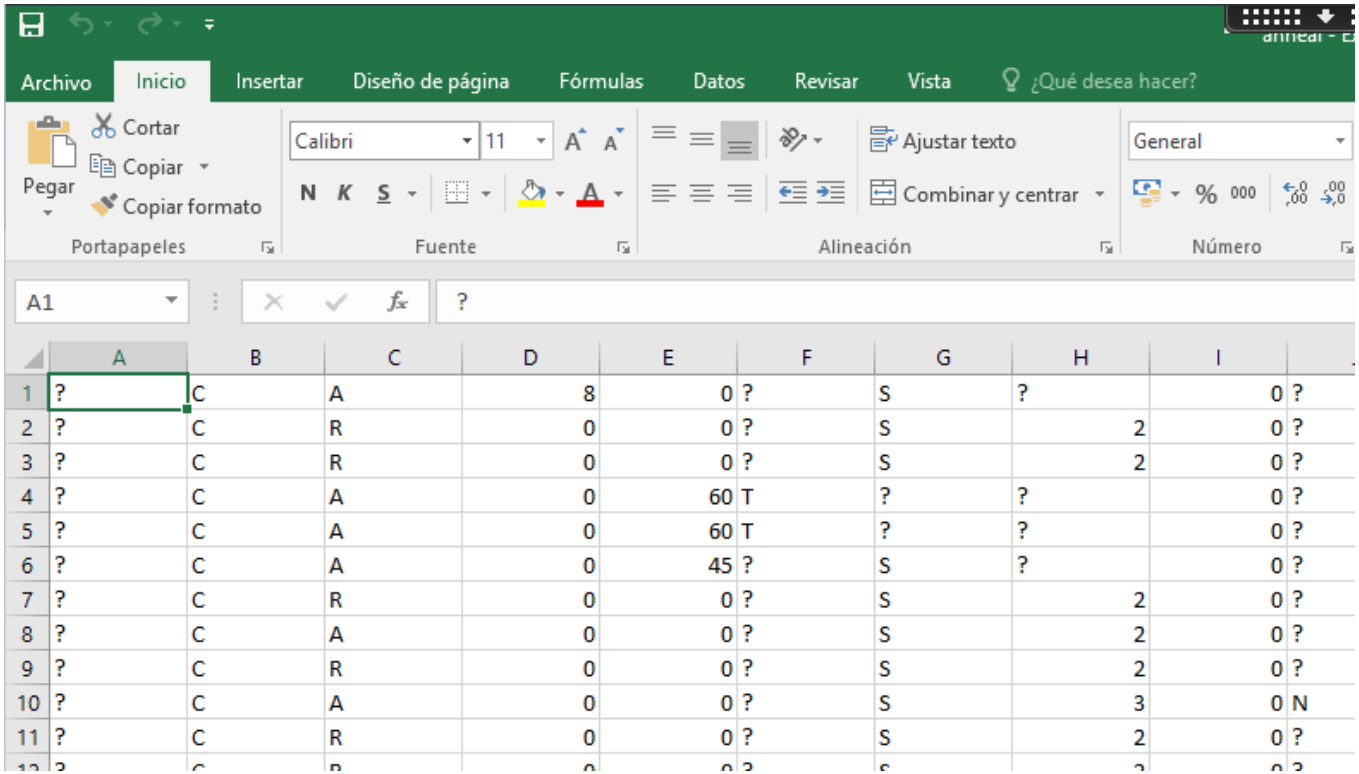


Importar a Excel (separador decimal)



Importar a Excel

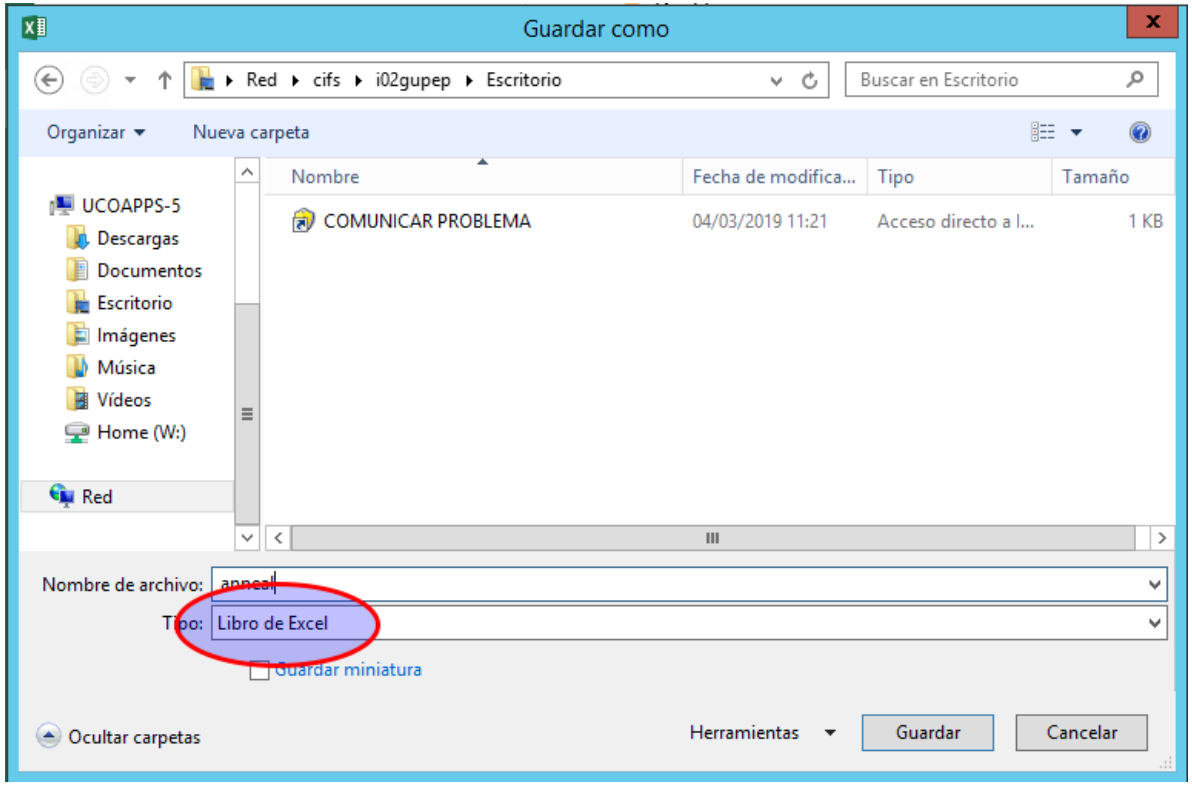
- Ya tenemos los datos importados.
- Todas esas interrogaciones son valores perdidos (ya veremos como tratarlos).



	A	B	C	D	E	F	G	H	I	
1	?	C	A	8	0	?	S	?	0	?
2	?	C	R	0	0	?	S	2	0	?
3	?	C	R	0	0	?	S	2	0	?
4	?	C	A	0	60	T	?	?	0	?
5	?	C	A	0	60	T	?	?	0	?
6	?	C	A	0	45	?	S	?	0	?
7	?	C	R	0	0	?	S	2	0	?
8	?	C	A	0	0	?	S	2	0	?
9	?	C	R	0	0	?	S	2	0	?
10	?	C	A	0	0	?	S	3	0	N
11	?	C	R	0	0	?	S	2	0	?

Importar a Excel

- Antes de seguir, conviene guardar los datos en el formato nativo de Excel (`.xls`) para tenerlos para futuras modificaciones.

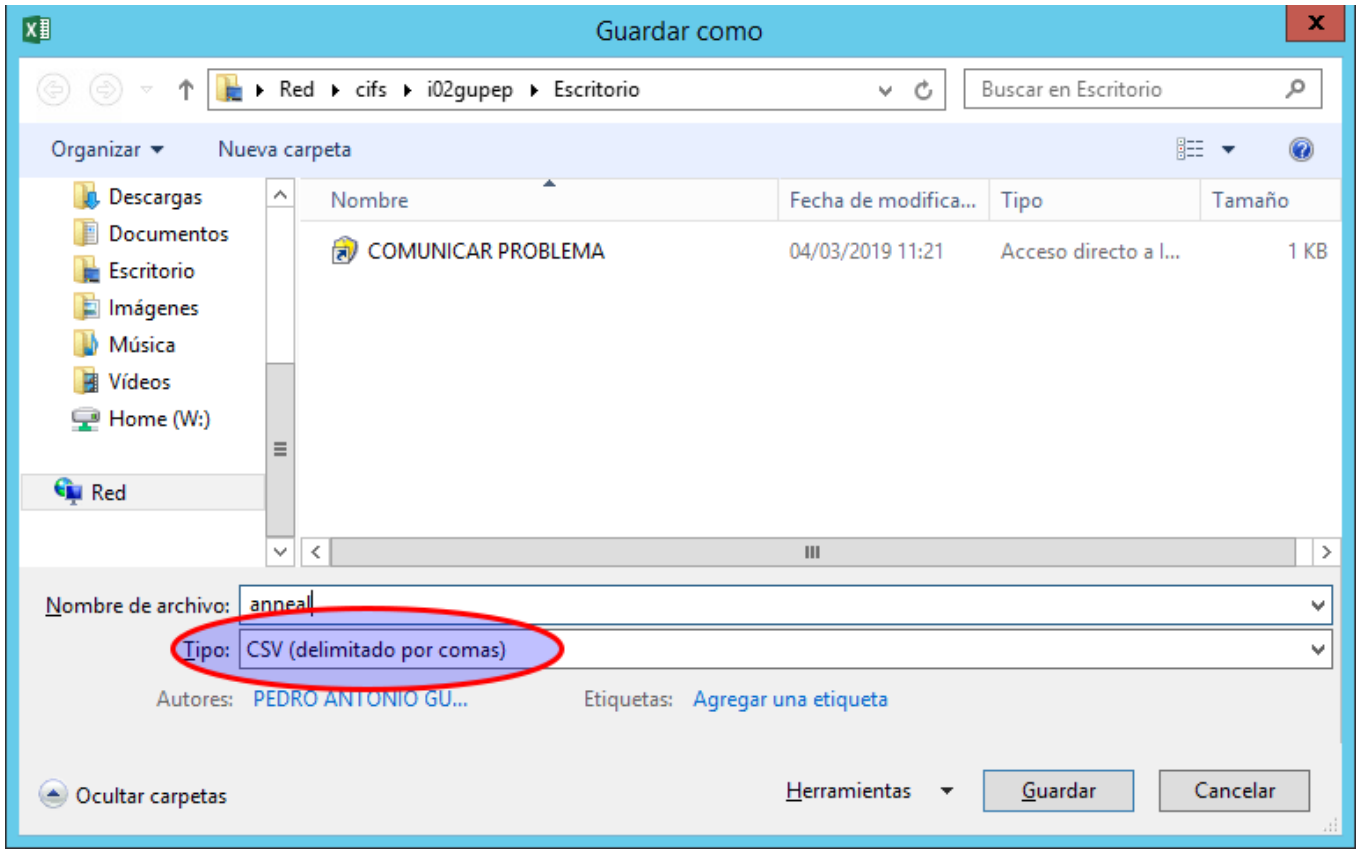


Entender los datos

- En archivo `anneal.names` tenemos información sobre cada atributo.
- Vemos que la última columna es la clase a predecir, con 5 clases distintas (1, 2, 3, 4, 5 o U).
- Los valores perdidos se representan con "?" y los valores no aplicables con "-" (y deben considerarse como categoría).

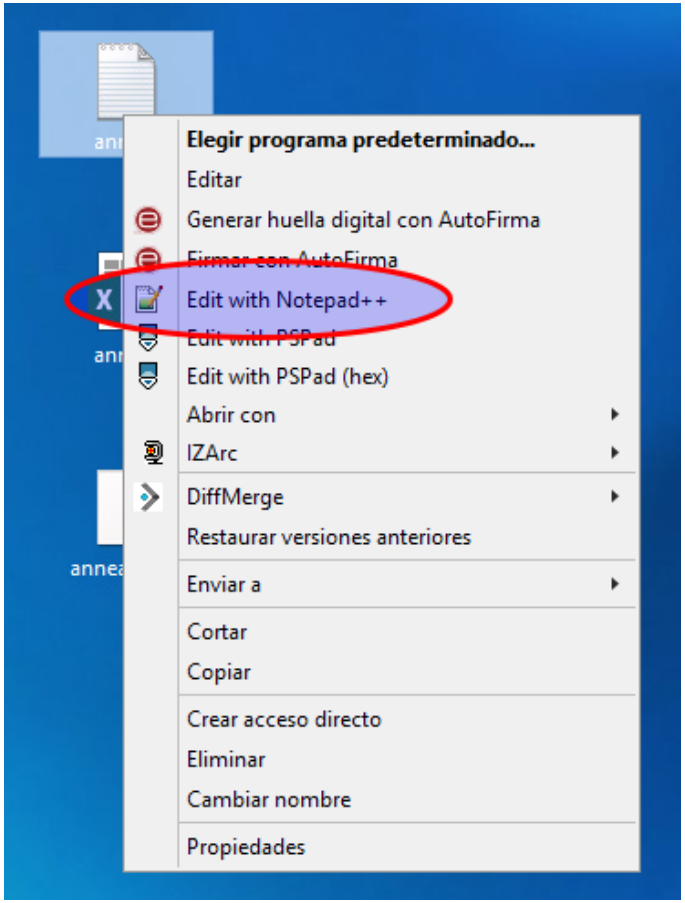
De Excel a .csv

- Ahora convertimos los datos a .csv (que si que lo puede leer Weka) (Archivo, Guardar como...).



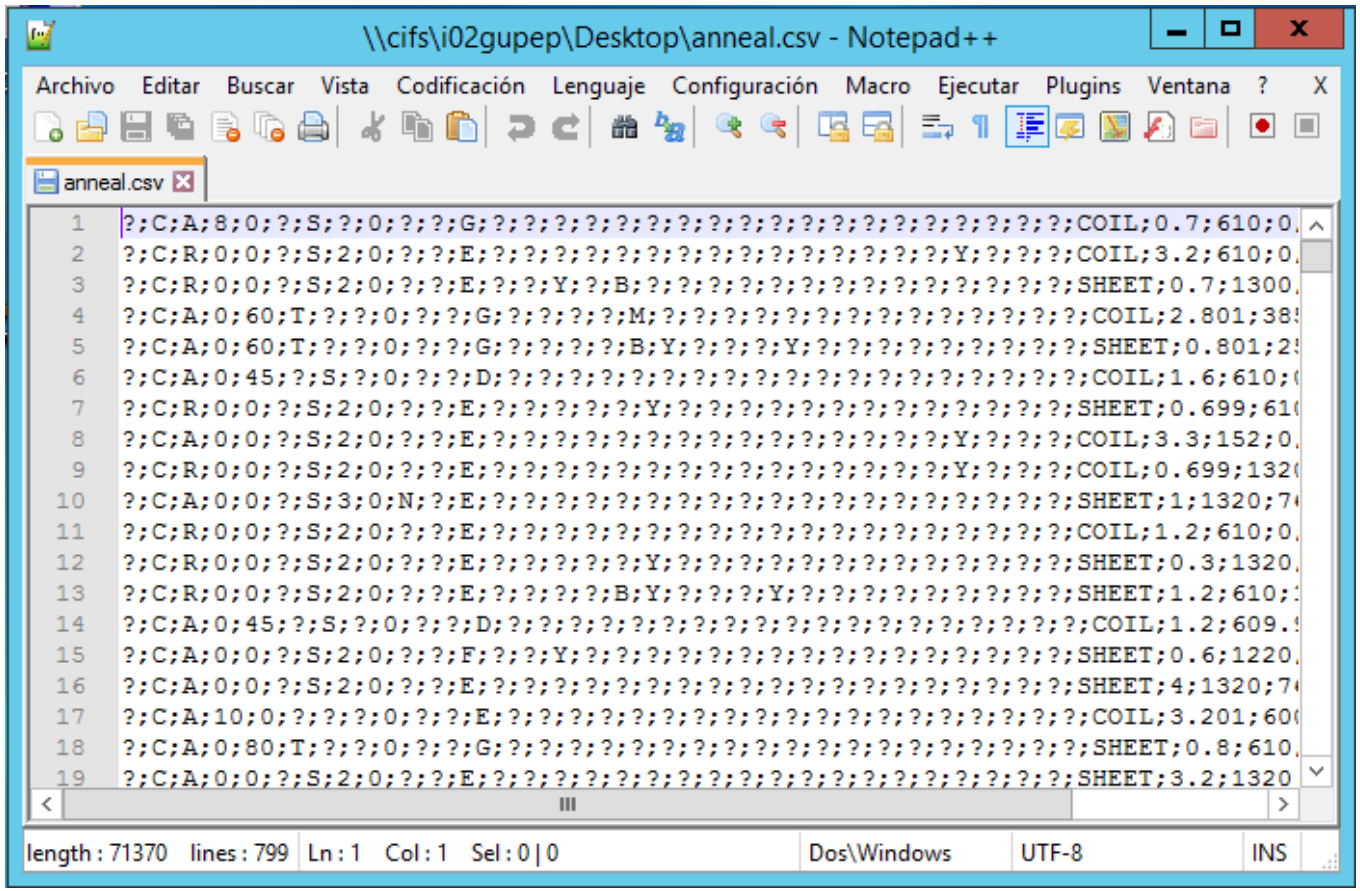
De Excel a .csv

- Abrimos el archivo con el "Bloc de notas", para ver como se ha generado:



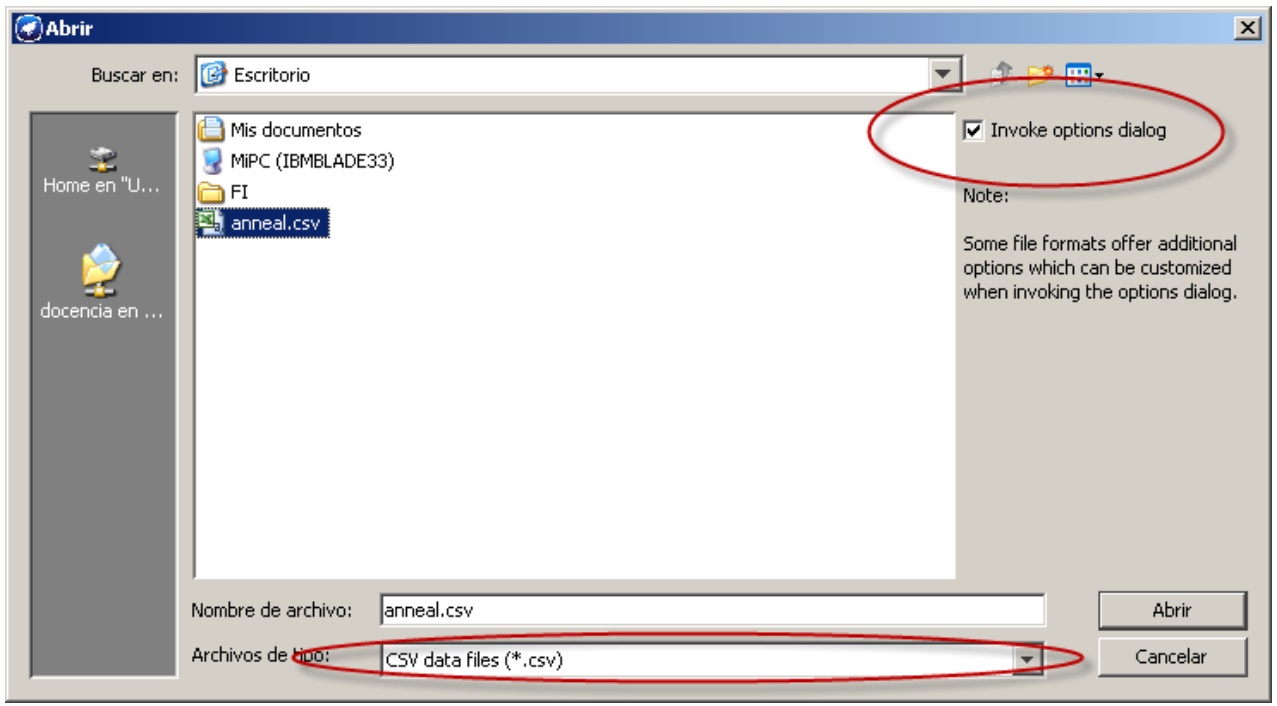
De Excel a .csv

- Vemos que ha aparecido el ";" como separador de campo:



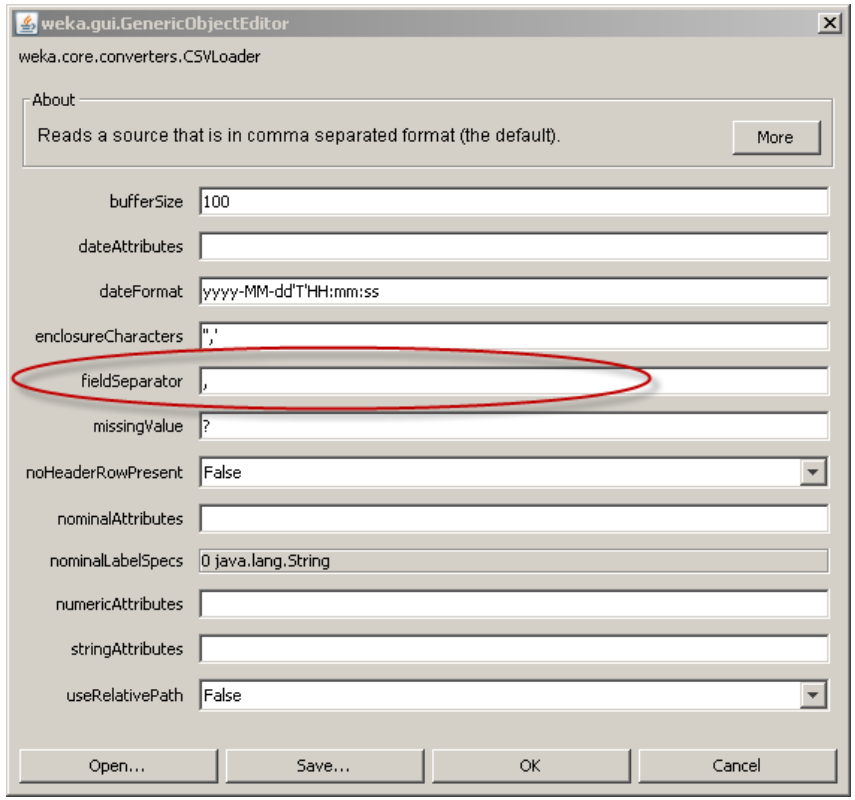
Cargar `.csv` en Weka

- Abrimos Weka y accedemos al `Explorer`.
- Pulsamos `Open file...` y elegimos "CSV" como tipo de archivos.
- Importante: Pulsar "Invoke options dialog".



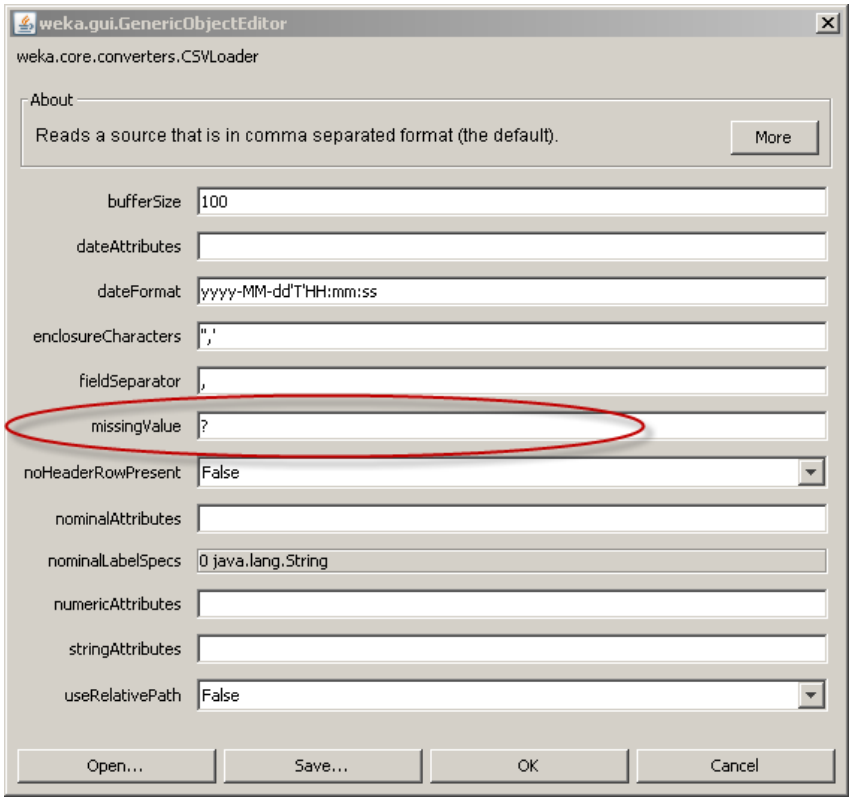
Cargar .csv en Weka

- Elegir el separador de campos (cambiar por “;”).



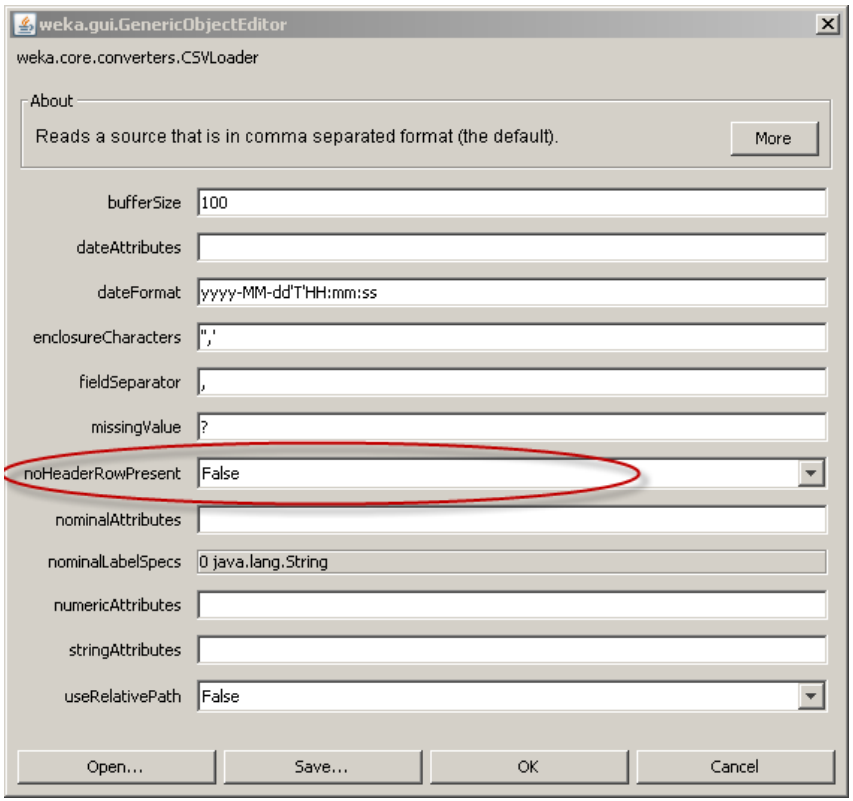
Cargar .csv en Weka

- Elegir el carácter que usamos para valores perdidos (en nuestro caso, "?".



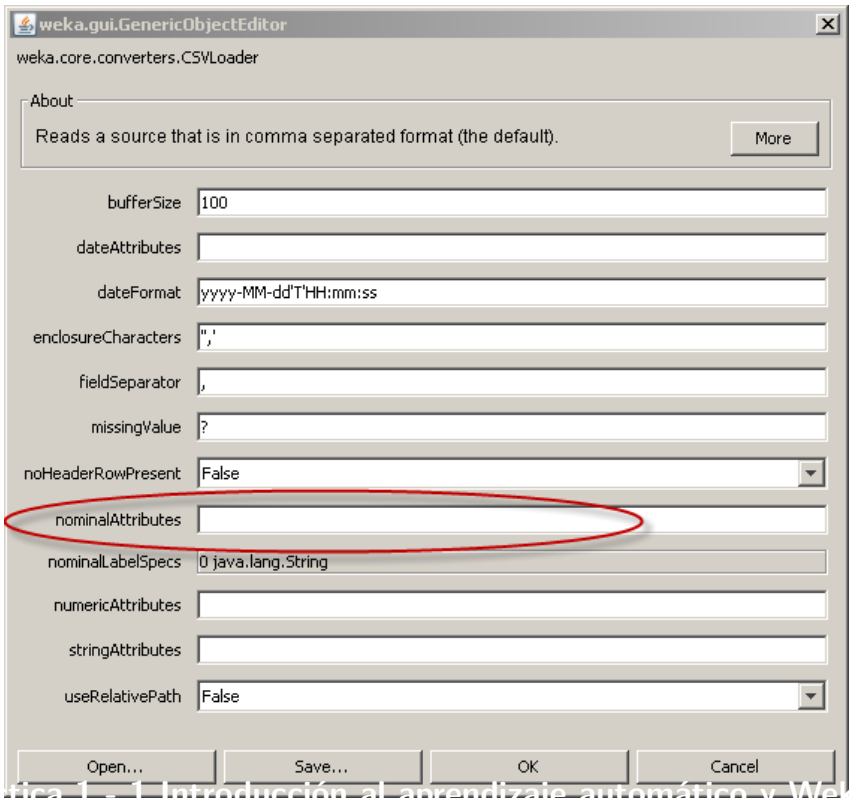
Cargar .csv en Weka

- Decimos si el fichero .csv no trae una fila al principio con los nombres de las variables (en nuestro caso, no la trae, decimos "True").



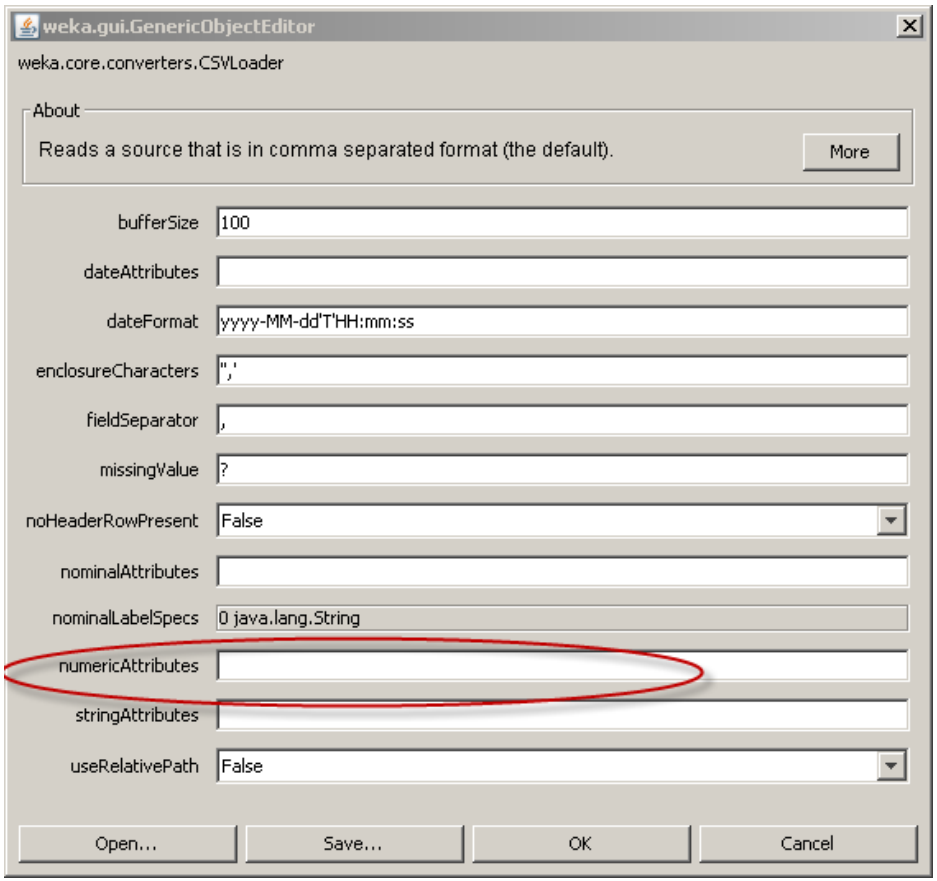
Cargar .csv en Weka

- Lista de variables nominales (hay que poner su índice).
- Se pueden poner rangos (p.ej. 1-3,6-8,10-32,36-39).
- Es recomendable escribir los rangos en otro documento, para copiar y pegar.



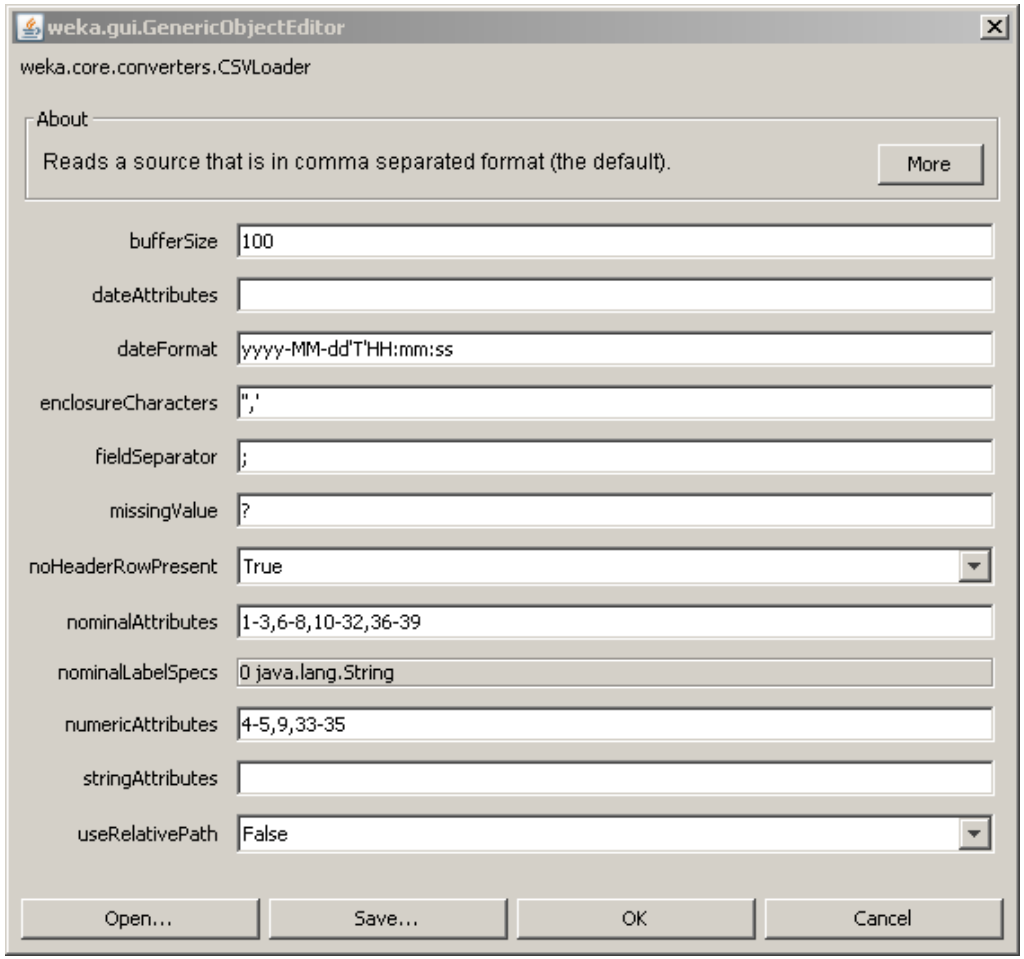
Cargar `.csv` en Weka

- Lista de variables numéricas (hay que poner su índice, `4-5,9,33-35`).



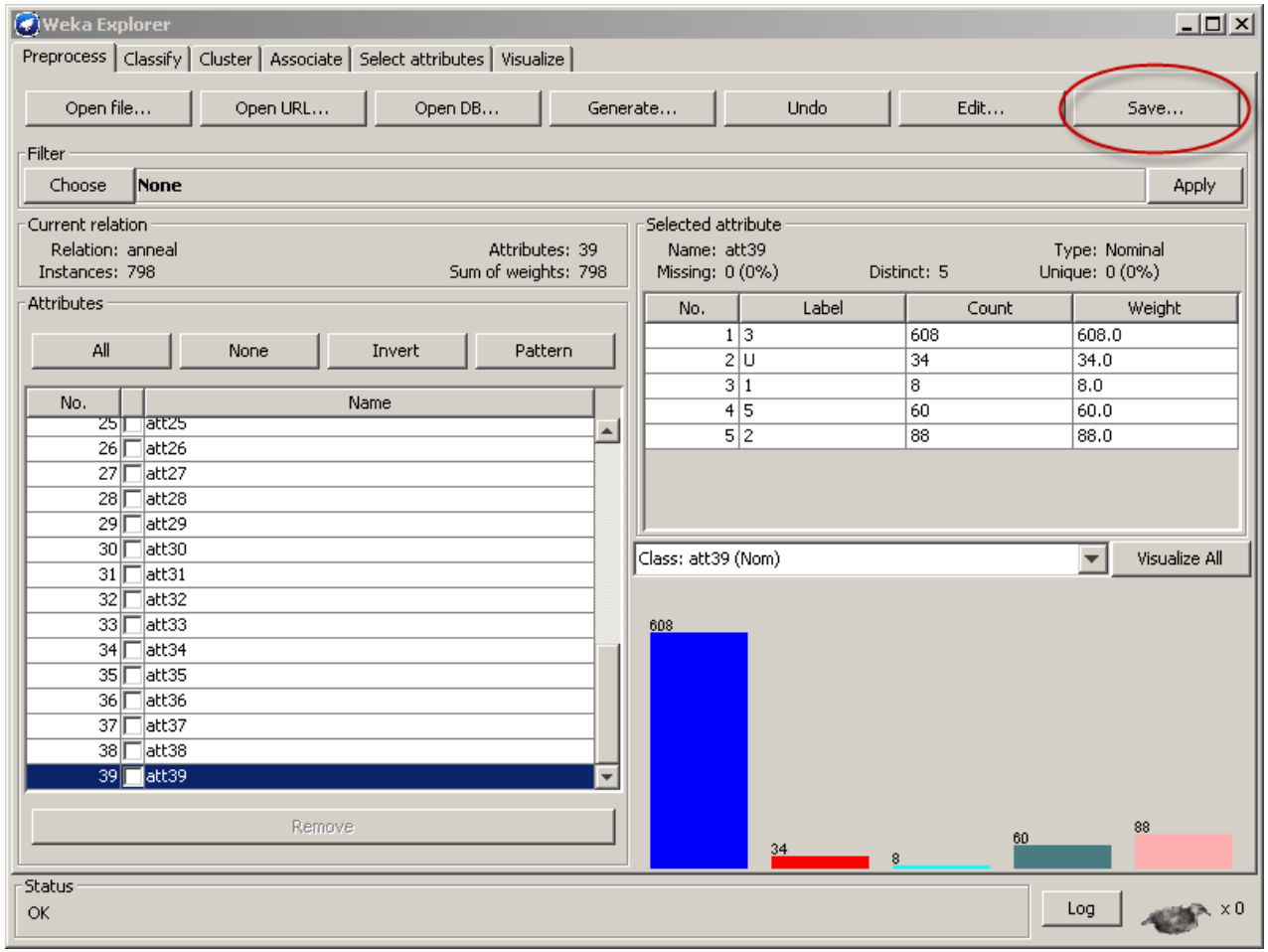
Cargar .csv en Weka

Para el caso de anneal, debemos dejarlo todo así:



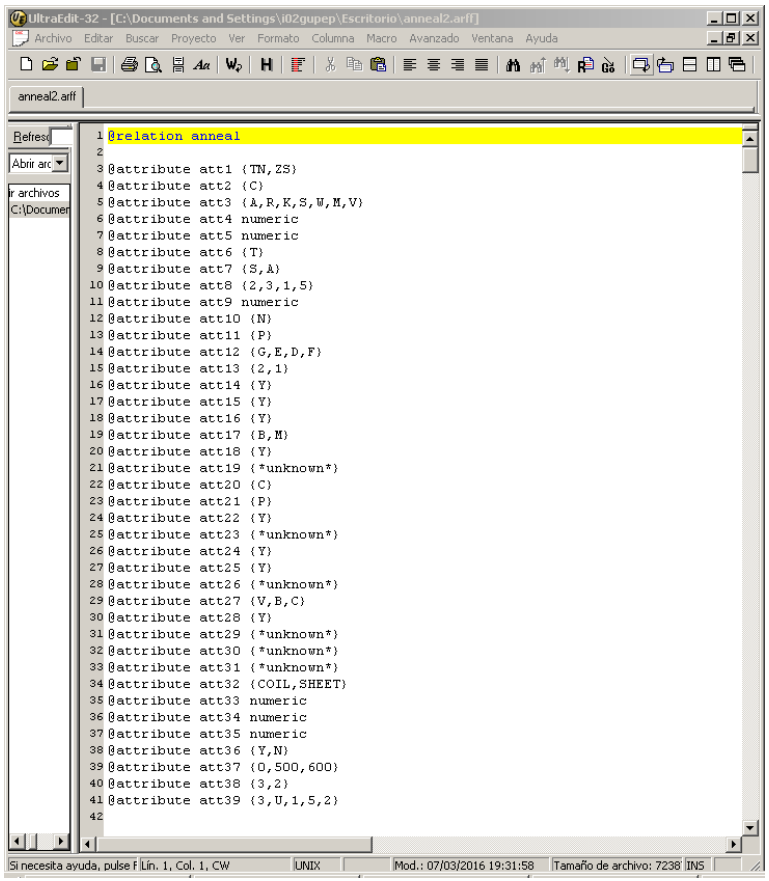
De .csv a .arff

- Salvamos los datos en formato .arff:



De .csv a .arff

- Fichero .arff generado (botón derecho en el archivo, Notepad ++):



Entregables

1. Elija 3 bases de datos de la *UCI Machine Learning Repository* de las que hay en Moodle y transformelas a .arff, indicando en cada una de ellas qué procedimiento ha seguido.

Bibliografía adicional a la de la asignatura y al material de Moodle

 Weka 3: Data Mining Software in Java, 2019.
https://www.cs.waikato.ac.nz/ml/weka.

¿Preguntas?