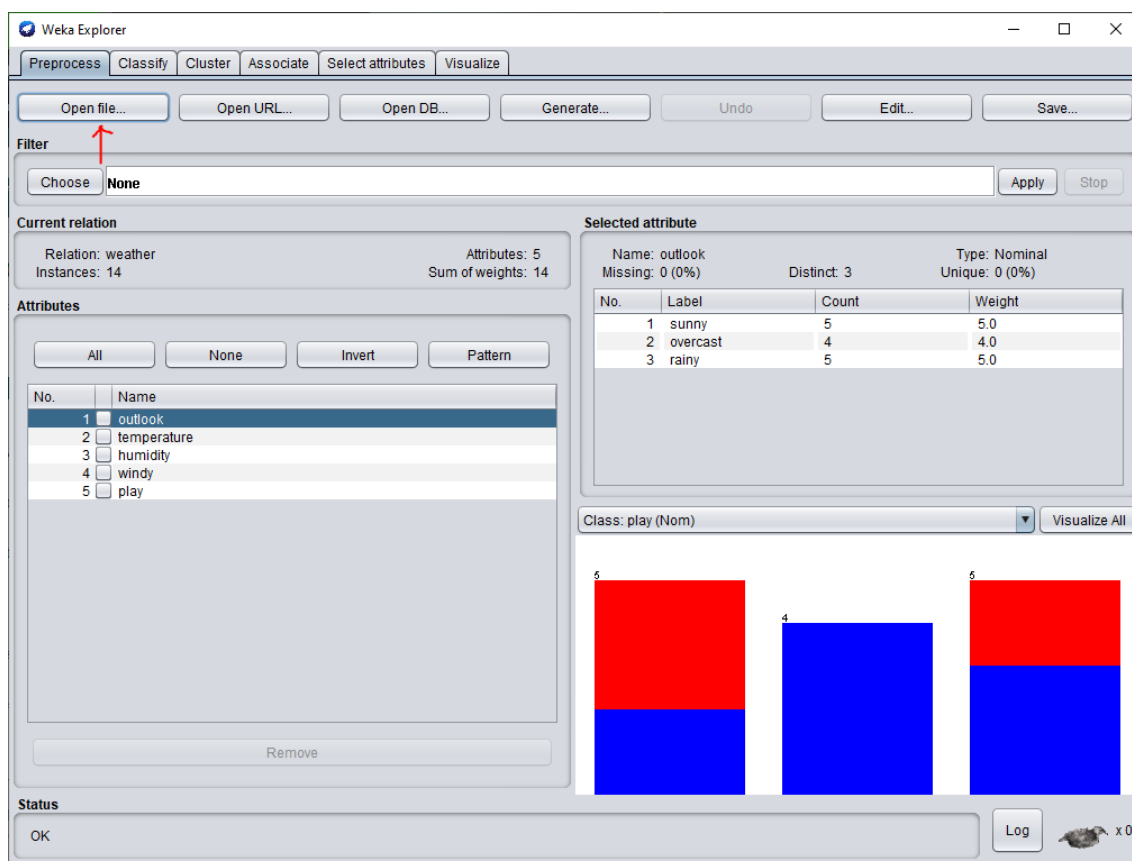


Tarea 28 de abril. Árboles de decisión con Weka.

En esta ocasión vamos a aprender a realizar un árbol de decisiones mediante la herramienta Weka. Vamos a trabajar con el archivo weather.arff

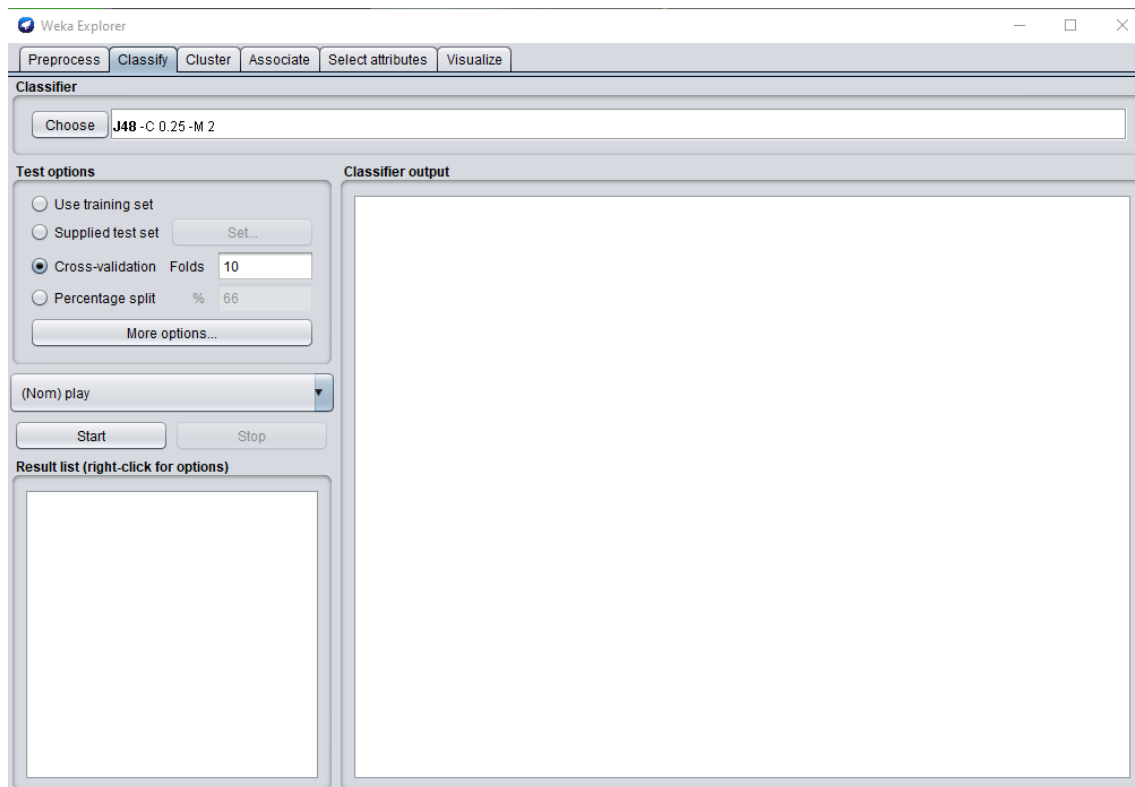
En primer lugar vamos a abrir Weka. En la pantalla principal seleccionaremos la opción “explorer”, y en la nueva ventana que nos aparece abriremos el archivo con el que vamos a trabajar.



En esta ventana observamos los atributos con los que estamos trabajando así como sus valores.

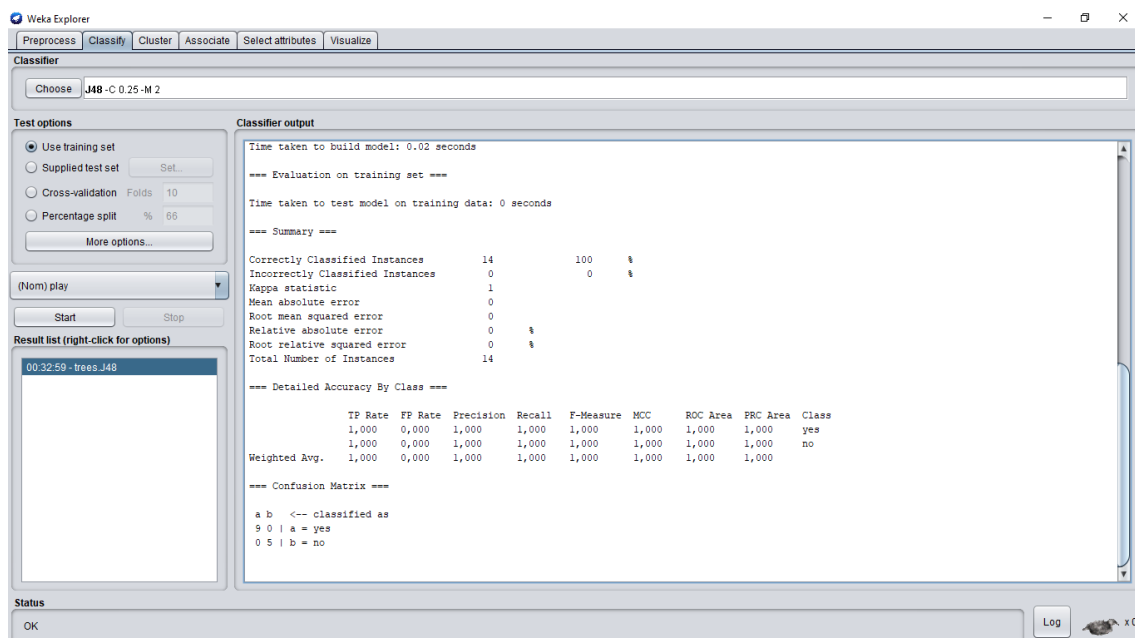
En esta ocasión vamos a trabajar para intentar saber o predecir si un partido se jugará o no teniendo información sobre la temperatura, humedad, viento...

Para ello, iremos a la pestaña “Classify” y buscaremos la subpestaña que dice “trees”. Aquí tendremos una lista de los algoritmos que podemos usar para crear un árbol de decisión. En nuestro caso elegiremos “J48”.

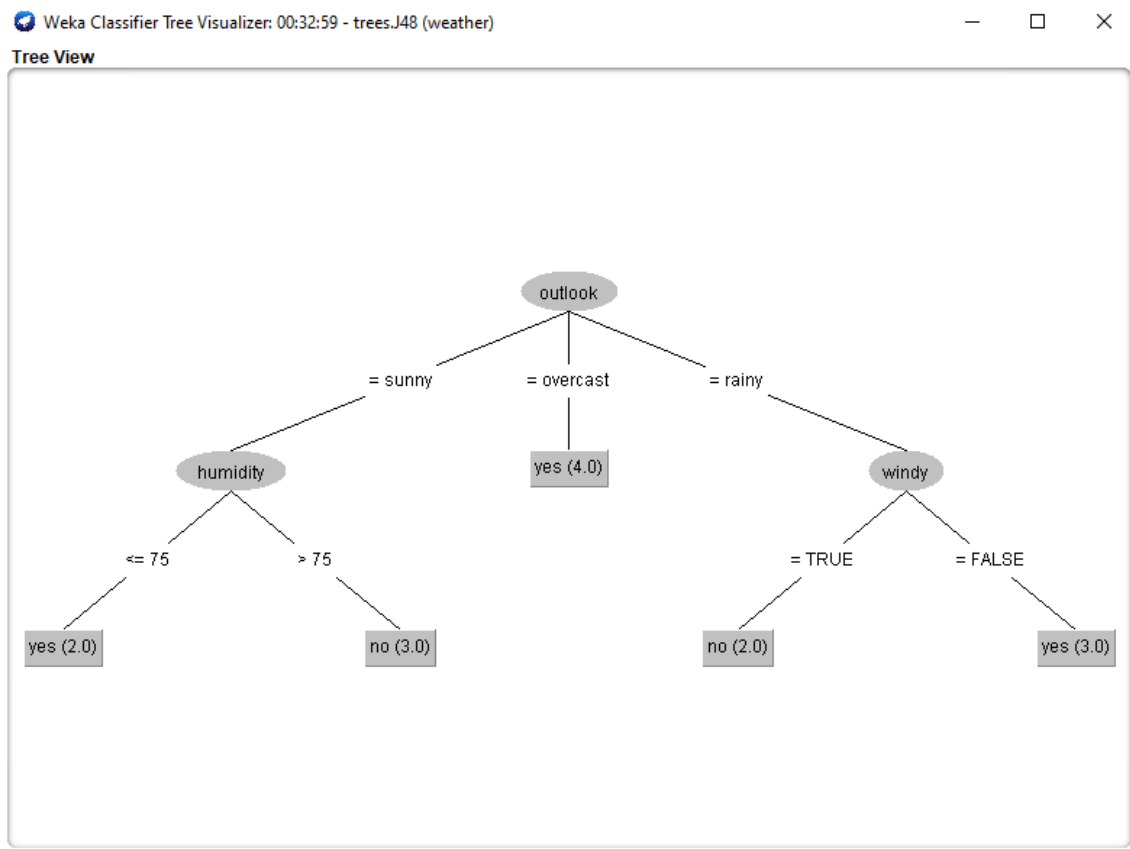


Ahora tenemos que señalar la opción “Use training set”, ya que nosotros usaremos una serie de datos para probar el modelo. También debemos asegurarnos que dónde dice “(Nom)” está señalada la variable que queremos explicar, en este caso “play”.

Tras realizar eso, podemos pulsar “Start”.



En esta pantalla ya tenemos todo lo que buscamos, sin embargo, podemos hacer click con el botón derecho en “tres.J48”, en “Result list” y señalar la opción que dice: “Visualize tree”. De esta manera llegamos a una pantalla con el árbol de decisiones explicado de forma gráfica:

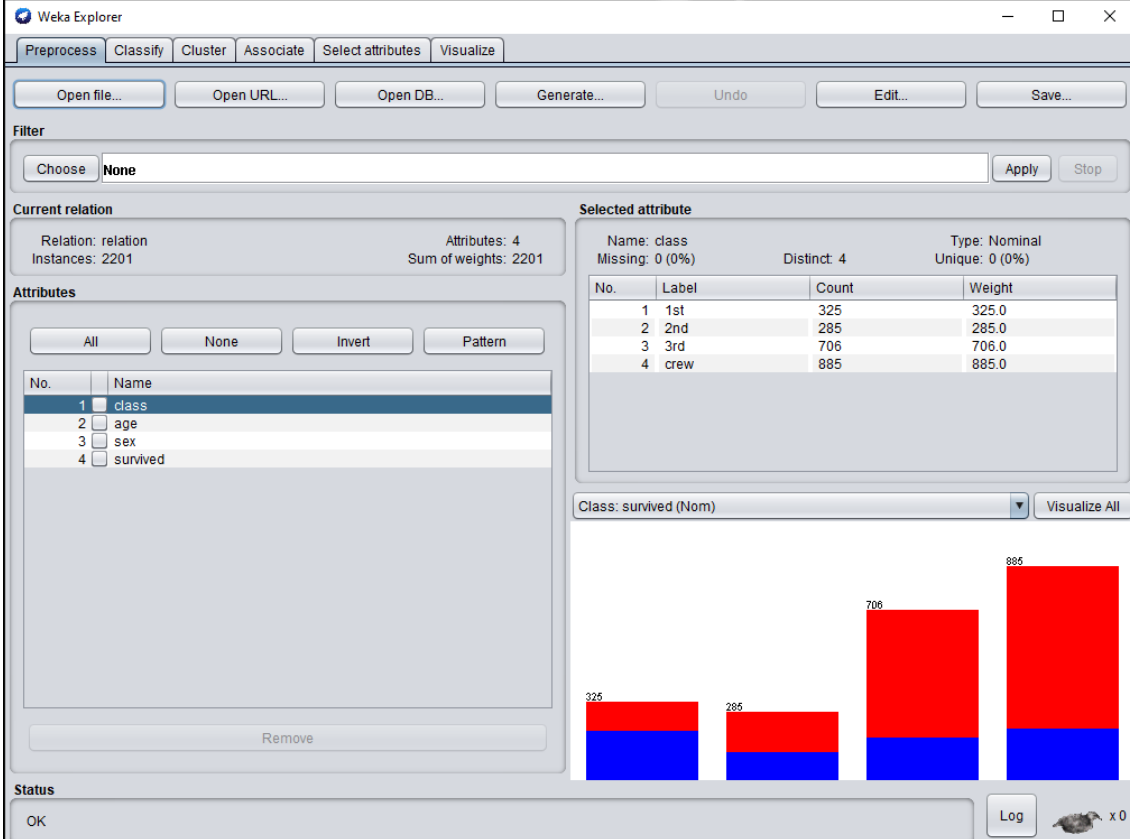


Como podemos observar, el árbol consta de 5 hojas, y la variable que más discrimina al a hora de si un partido se juega o no, es “Outlook”. Es decir, la variable más discriminativa es como se ve el tiempo, si lluvioso, soleado o nublado.

Vemos que una rama por ejemplo es que, si el tiempo es soleado y la humedad es mayor o igual del 75%, entonces se jugará el partido.

Los números que aparecen entre paréntesis indican el número de veces que salió si y el número de veces que salió no.

A continuación crearemos un árbol de decisiones, usando la herramienta Weka, basado en los datos del archivo titanic.arff. Para ello abriremos Weka y pulsaremos “Explorer”, buscaremos el archivo y lo abriremos tal y como hemos hecho con el archivo weather.arff anteriormente.



The screenshot shows the Weka Explorer window with the 'Classify' tab selected. The 'Current relation' section indicates the dataset is 'relation' with 2201 instances and 4 attributes. The 'Attributes' list on the left includes 'class', 'age', 'sex', and 'survived'. The 'Selected attribute' section shows 'Name: class' and 'Type: Nominal'. A bar chart at the bottom displays the distribution of the 'survived' attribute across the four classes (1st, 2nd, 3rd, crew). The chart shows that the 'survived' attribute is more prevalent in the '1st' and '3rd' classes compared to the '2nd' and 'crew' classes.

No.	Label	Count	Weight
1	1st	325	325.0
2	2nd	285	285.0
3	3rd	706	706.0
4	crew	885	885.0

Vamos a la pestaña “Classify” y buscamos el mismo árbol que usamos anteriormente, el J48. Marcamos “Use training set”, nos aseguramos de que el atributo que vamos a usar como clase es “Survived” y pulsamos “Start”.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 10
☐ Percentage split % 66
More options...

(Nom) survived

Start Stop

Result list (right-click for options)

18:40:01 - trees.J48

Classifier output

Time taken to build model: 0.14 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.15 seconds

=== Summary ===

Correctly Classified Instances	1740	79.055 %
Incorrectly Classified Instances	461	20.945 %
Kappa statistic	0.4334	
Mean absolute error	0.3089	
Root mean squared error	0.393	
Relative absolute error	70.6078 %	
Root relative squared error	84.0339 %	
Total Number of Instances	2201	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,380	0,013	0,931	0,380	0,539	0,506	0,765	0,666	yes
	0,987	0,620	0,769	0,987	0,864	0,506	0,765	0,827	no
Weighted Avg.	0,791	0,424	0,821	0,791	0,759	0,506	0,765	0,775	

=== Confusion Matrix ===

```

a    b    <-- classified as
270  441 |    a = yes
 20 1470 |    b = no

```

En esta pantalla podemos observar la matriz de confusión, en la que vemos que un total de 1740 (270 + 1470) están bien clasificados frente a los 461 (441 + 20) que están mal clasificados.

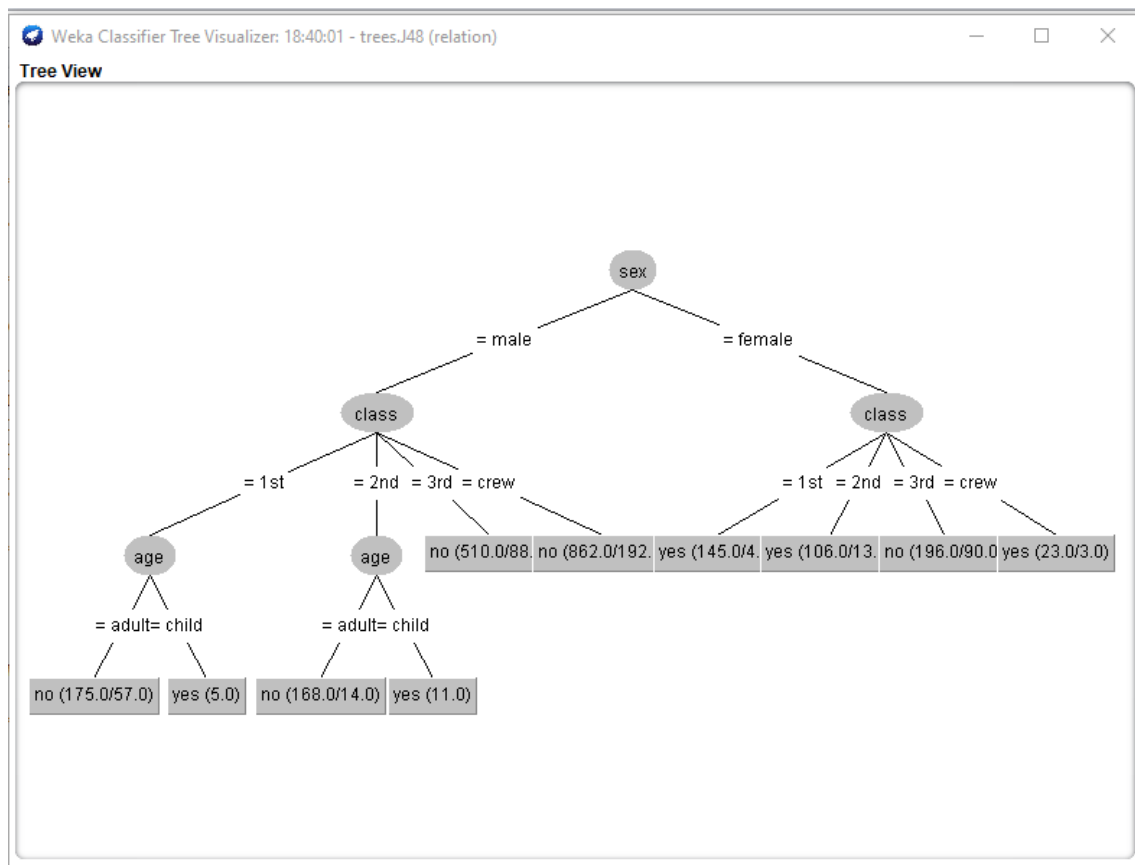
=== Confusion Matrix ===

```

a    b    <-- classified as
270  441 |    a = yes
 20 1470 |    b = no

```

Para ver el árbol de confusión gráficamente hacemos click con el botón derecho en “trees.J48” y seleccionamos la opción “Visualize tree”:



Nuestro árbol tiene 10 hojas, y podemos decir observándolo que el atributo que más discrimina es el sexo del tripulante en cuestión a si sobrevivió o no. El segundo atributo que más discrimina es la clase en la que viajaba, y por último, si el tripulante era un hombre, el atributo que menos discrimina es la edad del susodicho.