Aprendizaje automático

Tutorial 2

Grupo 84

Sabrina Riesgo Reyes	100363834	100363834@alumnos.uc3m.es
Laura Yunta García	100363785	100363785@alumnos.uc3m.es

ÍNDICE

1. INTRODUCCIÓN.

El desarrollo de la práctica busca adquirir los conocimientos para utilizar la herramienta Weka y que servirá para hacer un estudio de las distintas clasificaciones según los distintos algoritmos para el análisis y el modelado predictivo para datos de entrenamiento.

Finalmente, el documento y con ello la práctica se dividirá en cuatro partes diferenciadas:

- **Introducción:** En este apartado se comenta brevemente el objetivo a alcanzar con el desarrollo del tutorial y se explica la estructura y contenido de cada uno de los apartados del documento.
- **Respuestas a los ejercicios:** En este apartado se responderán a cada una de las preguntas especificadas en el documento del tutorial.
- **Conclusiones y problemas encontrados:** En este apartado se comentarán las conclusiones sacadas con la realización del ejercicio, además los diversos problemas que se han conseguido solventar durante el trascurso de la práctica.

2. RESPUESTAS A LOS EJERCICIOS.

En esta sección se procederá a responder todas las preguntas y a realizar todos los ejercicios propuestos en el tutorial de la práctica.

EJERCICIO 1: Los ficheros de datos.

a) **¿Cuántos atributos de entrada tiene el fichero de datos? ¿De qué tipo son?**

El fichero de datos de 2 atributos de entrada. A continuación, se muestran en una tabla dichos atributos y el tipo al que pertenecen.

ATRIBUTO	TIPO
name	NOMINAL
class	NOMINAL

b) **¿Podría un algoritmo de aprendizaje automático identificar una función capaz de predecir dicha clase con los datos que hay en el fichero? ¿Por qué?**

No debido a que los atributos no otorgan ninguna información.

EJERCICIO 2: Clasificar con ID3.

a) **En la pestaña Classify, seleccionar el clasificador trees/ID3. En las Test options seleccionar Use training set y pulsar el botón de Start para que se genere el modelo. ¿Cómo de buenos son los resultados?**

Los resultados son perfectos ya que se trata de un acierto al 100%. Esto se debe a que los valores de FN (falsos negativos) y FP (falsos positivos) son 0.

EJERCICIO 3: Generando nuevos atributos.

a) Imagina al menos 6 atributos que te parezca que podrían ser relevantes para este problema. Estos atributos inventados se deberían poder extraer tratando adecuadamente el único atributo de entrada del ejercicio anterior (name). Anótalos y describe en qué consiste cada uno. ¿Por qué has elegido esos atributos?

Los atributos que nos parecen relevantes para este problema son los siguientes:

Atributo	**Descripción**
longitud	Indica el número de caracteres que contiene la palabra. Se trata de un atributo de tipo numeric.
num_vocales	Indica el número de vocales que contiene la palabra. Se trata de un atributo de tipo numeric.
num_consonantes	Indica el número de consonantes que contiene la palabra. Se trata de un atributo de tipo numeric.
num_blancos	Indica el número de espacios en blanco que contiene la palabra. Se trata de un atributo de tipo numeric.
letra_inicial	Indica el valor de la letra inicial del nombre. Se trata de un atributo de tipo nominal que puede tomar un valor dentro del rango {A-Z}.
compuesto	Indica si el nombre es compuesto o no. Se trata de un atributo de tipo nominal que puede tomar los valores 0 (no es compuesto) o 1 (es compuesto).

Hemos seleccionado estos atributos ya que consideramos que son los más representativos en función de la información del fichero y que son los más apropiados para hacer una correcta clasificación de las instancias.

b) Abre el fichero de datos badges1.arff con Weka. ¿Cuántos atributos de entrada tiene el fichero de datos? ¿De qué tipo son?

El fichero de datos tiene 9 atributos de entrada. A continuación, mostraremos en una tabla dichos atributos y el tipo al que pertenecen.

ATRIBUTO	**TIPO**
name	NOMINAL
length	NUMERIC
even_odd	NOMINAL
first_char_vowel	NOMINAL

consonants	NUMERIC
spaces	NUMERIC
dots	NUMERIC
words	NUMERIC
class	NOMINAL

c) ¿Qué otro tipo de información estadística se muestra sobre los atributos? Pulsa el botón "Visualize All". ¿Qué se muestra?

Dependiendo del tipo de atributo se muestra información distinta. En este caso se tienen atributos de tipo numeric y atributos de tipo nominal.

- En cuanto a los de tipo numeric se muestra: el valor máximo y mínimo, la media y la desviación estándar.
- En cuanto a los de tipo nominal se muestra: un identificador, una etiqueta, el número de repeticiones y el peso.

Al pulsar el botón "Visualize All" obtenemos nueve gráficas estadísticas (una por cada atributo) que sirven para comprobar cómo de efectivos son cada uno de los atributos, considerados por separado. Esto te permite elegir el mejor atributo clasificador entre todos ellos (aquel que se encuentra más cerca de la raíz, o como raíz del árbol de decisión). En el caso del atributo name no se muestra ningún gráfico debido a que contiene demasiados valores para que se visualice correctamente.

d) Tratar de generar un clasificador con tree/ID3. ¿Qué es lo que ocurre? ¿Qué se podría hacer para evitar este problema con ID3?

No se puede generar el clasificador debido a que la salida del modelo de clasificación ID3 se representa mediante un árbol de decisión en el que cada rama toma un posible valor. En este caso se tendrían que seleccionar los atributos nominales ya que en el caso de los numéricos el rango de valores sería muy amplio y no se podría representar mediante un árbol.

EJERCICIO 4: Clasificar con ID3: Resolviendo problemas.

a) Seleccionar el filtro Filter/unsupervised/attribute/Discretize, fijar el número de bins a 5 y aplicar al conjunto de datos. ¿Qué efecto este filtro?

Al utilizar el filtro cambian los valores de las etiquetas de valores discretos a rangos, lo que hace que cambien las estadísticas y por consiguiente las gráficas, de forma que no habrá tantos valores representados.

b) ¿Cuántas instancias del conjunto de entrenamiento clasifica bien? ¿Qué porcentaje clasifica bien?

Se clasifican bien 236 instancias del conjunto de entrenamiento y por lo tanto un 80,2721%.

c) ¿Qué crees que indica la matriz de confusión?

La matriz de confusión indica la proporción de errores que se cometen al realizar una clasificación.

d) ¿Cuántas instancias de cada tipo se han clasificado mal?

Se clasifican mal 58 instancias del conjunto de entrenamiento y por lo tanto un 19,7279%.

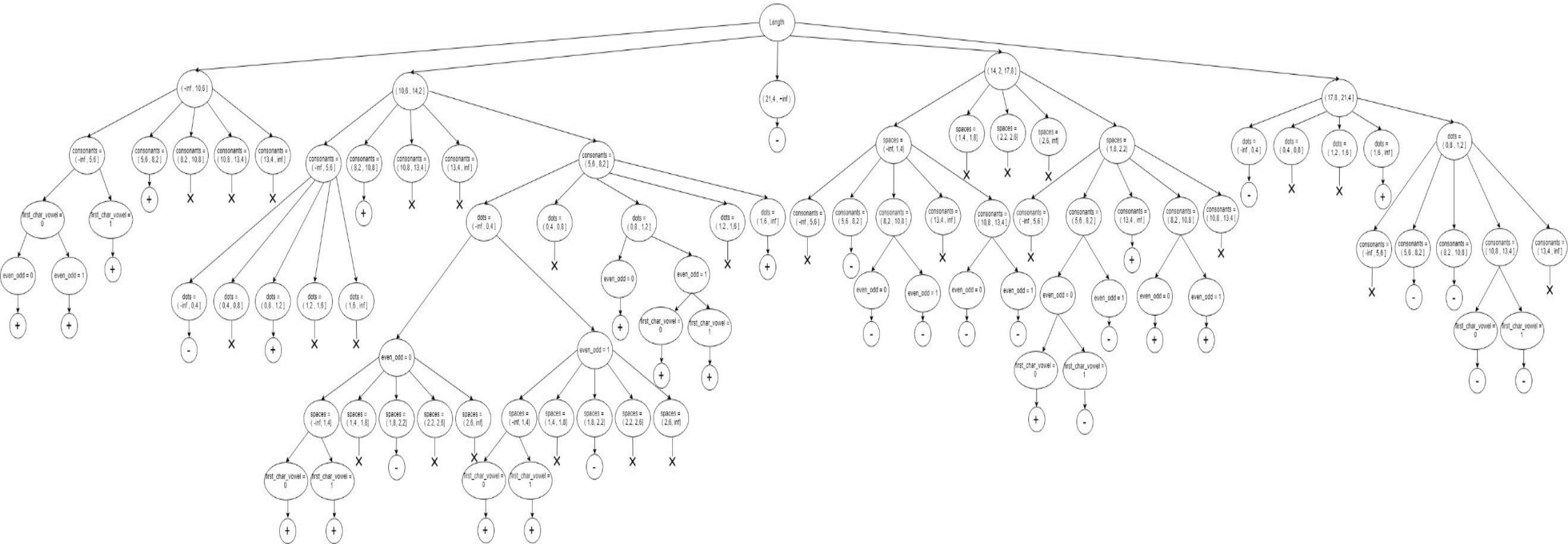
e) Pulsar el botón de More Options y seleccionar la opción Output Predictions (Plain Text). Volver a clasificar y revisar los resultados. ¿Cuál es la primera instancia del conjunto de entrenamiento que se clasifica mal? ¿Por qué?

La primera instancia que se clasifica mal es la instancia 7 debido a que su clase correspondiente es la clase 1: - y sin embargo, la clase predicha es 2: + con un porcentaje de predicción de 0.571.

f) ¿Cómo se clasificaría la instancia "Donald Trump"? ¿Cuáles son los atributos de este nombre? ¿Qué ocurre con los valores de esta instancia si utilizas el filtro usado anteriormente?

Tras haber generado el árbol de decisión con el clasificador ID3 que se muestra a continuación, podemos clasificar la instancia "Donald Trump" en la clase: "+". Debido a que la longitud de la instancia es de 12 caracteres, el número de consonantes es 8, el número de puntos es 0, la instancia tiene una longitud impar, el número de espacios es 1 y el primer caracter no es una vocal.

El árbol de decisión generado es el siguiente:


Length
(-inf , 10.6]
(10.6 , 14.2]
(21.4 , +inf)
(14.2, 17.8]
(17.8 , 21.4]
consonants = (-inf , 5.6]
consonants = (5.6 , 8.2]
consonants = (8.2 , 10.8]
consonants = (10.8 , 13.4]
consonants = (13.4 , inf)
first_char_vowel = 0
first_char_vowel = 1
even_odd = 0
even_odd = 1
dots = (-inf , 0.4]
dots = (0.4 , 0.8]
dots = (0.8 , 1.2]
dots = (1.2 , 1.6]
dots = (1.6 , inf)
spaces = (-inf, 1.4]
spaces = (1.4 , 1.8]
spaces = (1.8, 2.2]
spaces = (2.2, 2.6]
spaces = (2.6, inf]

EJERCICIO 5: Clasificar con ZeroR.

a) ¿Qué modelo genera el clasificador ZeroR?

Genera un modelo con una matriz de confusión que genera los siguientes valores de instancias por clase:

- Positivos verdaderos (TP): 0
- Falsos positivos (FP): 144
- Falsos negativos (FN): 0
- Negativos verdaderos (TN): 150

Por lo tanto, tiene 150 instancias bien clasificadas dando un 51,0204 % de acierto y 144 instancias mal clasificadas dando un 48,9796% de error. Es decir, que es un mal modelo.

b) ¿Cuál es el porcentaje de éxito de este modelo?

El porcentaje de éxito es de un 51,0204 %.

c) ¿Cómo se clasificaría la instancia "Donald Trump"?

Según el modelo generado en el cual se clasifica a todas las instancias en la clase 2: +, podemos decir que según este clasificador la instancia "Donald Trump" estará predicha en la clase 2: +.

EJERCICIO 6: Clasificar con J48.

a) ¿Cuántas hojas tiene el árbol generado con J48?

En total el árbol generado tiene 20 hojas.

b) ¿Cuántas instancias del conjunto de entrenamiento clasifica bien?

El clasificador ha clasificado correctamente 287 instancias.

c) ¿Qué porcentaje de instancias clasifica bien?

Al haber clasificado correctamente 287 instancias, el porcentaje de instancias bien clasificadas es de 97,619%.

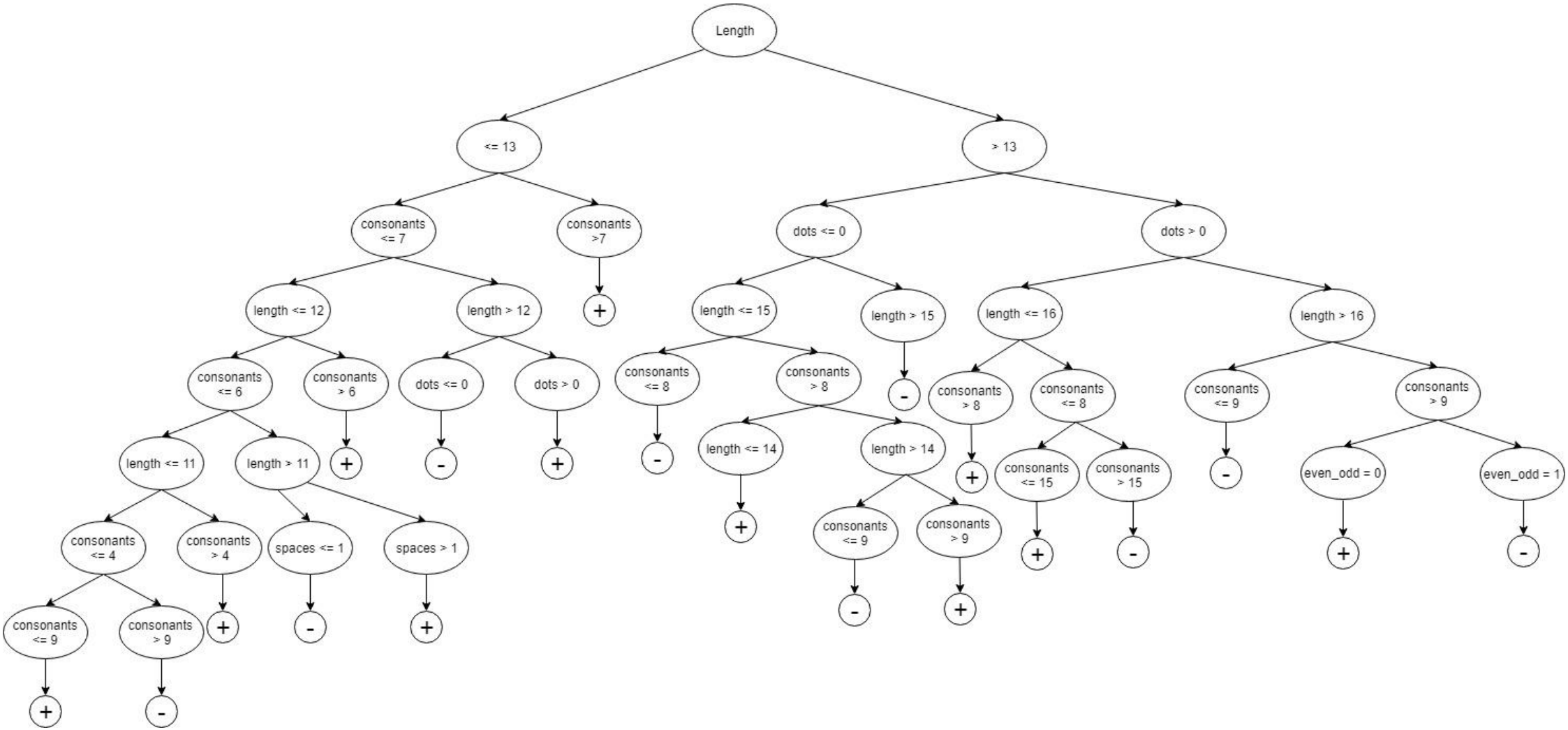
d) ¿Cuántas instancias de cada tipo se han clasificado mal?

Únicamente se han clasificado mal 7 instancias. En este caso son: 3 de la clase a = - y 4 de la clase b = +.

e) ¿Cómo se clasificaría la instancia "Donald Trump"?

Tras haber generado el árbol de decisión con el clasificador ID3 que se muestra a continuación, podemos clasificar la instancia "Donald Trump" en la clase: "+". Debido a que la longitud de la instancia es de 12 caracteres, el número de consonantes es 8.

El árbol de decisión generado es el siguiente:


Length
<= 13
> 13
consonants <= 7
consonants >7
dots <= 0
dots > 0
length <= 12
length > 12
+
length <= 15
length > 15
length <= 16
length > 16
consonants <= 6
consonants > 6
dots <= 0
dots > 0
consonants <= 8
consonants > 8
-
consonants > 8
consonants <= 8
consonants <= 9
consonants > 9
length <= 11
length > 11
+
-
+
length <= 14
length > 14
+
consonants <= 15
consonants > 15
-
even_odd = 0
even_odd = 1
consonants <= 4
consonants > 4
spaces <= 1
spaces > 1
consonants <= 9
consonants > 9
+
-
+
consonants <= 9
consonants > 9
+
+
-
+
-
+
+
-
+
-

f) ¿Elegirías este modelo o el generado por ID3? ¿Por qué?

Es preferible escoger este modelo generado por J48 debido a que genera un porcentaje de acierto mucho mayor al que genera ID3 y por tanto el modelo será mejor.

g) ¿Hemos encontrado la función exacta para generar las etiquetas? ¿Por qué lo sabes?

No porque todavía hay instancias mal clasificadas y por tanto el clasificador no será el perfecto.

EJERCICIO 7: Utilizando más atributos con J48 (C4.5).

a) Volver la pestaña de preproceso y seleccionar el filtro FIlter/unsupervised/attribute/AddExpression para generar un nuevo atributo que calcule el número de vocales.

b) Grabar el conjunto de datos como badges-Ej7.arff.

c) ¿Podrías decir cuál es el rango de vocales más común en el fichero proporcionado?

El rango de vocales más común del fichero es [4.909 - 5.636].

d) Volver a construir un clasificador con J48 con el conjunto de datos del punto anterior. En este caso tienes que seleccionar la clase "class" en el desplegable de la pestaña Classify.

e) Anota el porcentaje de instancias bien clasificadas y la matriz de confusión.

El porcentaje de instancias bien clasificadas es 99.6599%. La matriz de confusión obtenida contiene los siguientes valores:

- Positivos verdaderos (TP): 144
- Falsos positivos (FP): 0
- Falsos negativos (FN): 1
- Negativos verdaderos (TN): 149

f) Haz click con el botón derecho del ratón en el modelo generado que aparece en Result list. Visualiza el árbol generado con Visualize Tree. ¿Qué indican los números que aparecen en las hojas?

Los números que aparecen en las hojas indican el número de instancias que se encuentran en dicha clasificación y el número de instancias mal clasificadas.

g) Ir a la pestaña Visualize.

h) Pinchar en la gráfica que relaciona el atributo creado con la clase.

i) Aumentar el valor de Jitter. ¿Qué efecto tiene?

El aumento del valor de Jitter añade ruido a la posición de cada punto.

j) Tras todos estos resultados, ¿qué características o cualidades crees que deben tener los atributos para maximizar el éxito de los algoritmos de aprendizaje automático?
Consideramos que no resulta sencillo la elección de características o cualidades con el objetivo de maximizar el éxito de los algoritmos de aprendizaje automático ya que tendrían que ser atributos representativos que permitan clasificar adecuadamente cada una de las instancias.

EJERCICIO 8: Balanceado de datos, selección de características y otros filtros.

a) Cargar en Weka el fichero de datos adult-data.arff.

b) ¿Cuántos atributos de entrada tiene este fichero? ¿Cuántas instancias de entrenamiento?
Este fichero tiene 15 atributos de entrada y 32561 instancias de entrenamiento.

c) Ejecuta el clasificador J48. Selecciona en Test Options la opción "Cross-validation" ¿Qué resultados aparecen? Explica el resultado.

Tras ejecutar la cross-validation se obtiene que se han clasificado correctamente 28071 instancias y erróneamente 4490 instancias. Además, se genera una matriz de confusión con los siguientes valores:

- Positivos verdaderos (TP): 4951
- Falsos positivos (FP): 2890
- Falsos negativos (FN): 1600
- Negativos verdaderos (TN): 23120

Teniendo en cuenta los datos anteriores, se puede concluir que el porcentaje de aciertos es del 86,21%, por lo que no se trata de un modelo muy confiable.

d) Ahora vamos a evaluar el clasificador solamente con las instancias que figuren en el fichero adult-test.arff. Para ello selecciona en Test Options la opción "Supplied test set". ¿Qué resultados aparecen? ¿Estos resultados son comparables a los anteriores? ¿Por qué?

Tras ejecutar el supplied test set con las instancias que figuran en "adult-test.arff" se obtiene que se han clasificado correctamente 13977 instancias y erróneamente 2304 instancias. Además, se genera una matriz de confusión con los siguientes valores:

- Positivos verdaderos (TP): 2421
- Falsos positivos (FP): 1425
- Falsos negativos (FN): 879
- Negativos verdaderos (TN): 11556

Se obtiene un porcentaje de error del 85,85%.

Con estos resultados se puede visualizar que los resultados no son comparables entre sí ya que el número de instancias no es el mismo y por tanto no se toma la misma muestra para poder compararlos.

e) **Vuelve a la pestaña Preprocess y haz click en el atributo de salida (la clase). ¿Qué proporción de datos hay de cada clase? ¿Crees que este porcentaje es apropiado para que un algoritmo de aprendizaje automático aprenda bien?**

Para los casos en los que salary > 50k hay 7841 datos, mientras que para los casos en los que salary <= 50k hay 24720 datos, siendo la diferencia entre dichos valores muy significativa, por lo que los valores se encuentran claramente desbalanceados. Por dichas razones dichos valores no serían apropiados para que un algoritmo de aprendizaje automático aprenda bien.

f) **Vamos a modificar las instancias de entrenamiento para que tengan un porcentaje similar entre las dos clases. Para ello selecciona supervised/instance/Resample cambiando el parámetro biasToUniformClass a 1,0. ¿Qué ocurre con el atributo de salida? ¿Ha descendido el número de ejemplos de entrenamiento?**

El número de ejemplos de entrenamiento disminuye en una unidad. Además, la proporción de datos quedaría:

- 16280 datos para salary > 50k.
- 16280 datos para salary <= 50k.

g) **Tras aplicar este filtro, evalúa de nuevo con cross-validation y supplied test set el algoritmo J48. ¿Qué resultado ofrece ahora el algoritmo? ¿Ha mejorado o empeorado?**

Una vez aplicado el filtro se obtiene, en el caso de utilizar cross-validation, un porcentaje de acierto de 87.1898% y una matriz de confusión con los siguientes valores:

- Positivos verdaderos (TP): 14715.
- Falsos positivos (FP): 1565.
- Falsos negativos (FN): 2606.
- Negativos verdaderos (TN): 13674.

En el caso de utilizar supplied test set, se obtiene un porcentaje de acierto del 80.5663% y una matriz de confusión con los siguientes valores:

- Positivos verdaderos (TP): 2951.
- Falsos positivos (FP): 895.
- Falsos negativos (FN): 2269.
- Negativos verdaderos (TN): 10166.

h) **Por último, aplica el filtro de normalización unsupervised/instance/Normalize para los atributos numéricos. ¿Qué resultados se obtienen?**

Una vez aplicado el filtro se obtiene, en el caso de utilizar cross-validation, un porcentaje de acierto de 87.1867% y una matriz de confusión con los siguientes valores:

- Positivos verdaderos (TP): 14715.
- Falsos positivos (FP): 1565.
- Falsos negativos (FN): 2607.
- Negativos verdaderos (TN): 13673.

En el caso de utilizar supplied test set, se obtiene un porcentaje de acierto del 33.1675% y una matriz de confusión con los siguientes valores:

- Positivos verdaderos (TP): 2964.
- Falsos positivos (FP): 882.
- Falsos negativos (FN): 9999.
- Negativos verdaderos (TN): 2436.

i) Después del procesamiento de datos que has realizado en este apartado, ¿crees que esto ayuda al proceso de aprendizaje? ¿Por qué?

Teniendo en cuenta los resultados anteriores se puede concluir que la clasificación que más ayudan al proceso de aprendizaje es la realizada en el apartado g debido a que en esta se obtienen el porcentaje de acierto más elevado.

j) ¿Cuál es el mejor resultado obtenido? Justifícalo.

Comparando los resultados obtenidos en los apartados g y h en los cuales se han aplicado distintos filtros, se puede concluir que el mejor resultado obtenido sería el del apartado g ya que se obtiene el mayor porcentaje de acierto.

3. CONCLUSIONES Y DIFICULTADES.

Una vez realizada la práctica podemos concluir que es de gran importancia la buena elección de los atributos para clasificar correctamente las instancias, ya que las técnicas de aprendizaje automático dependerán directamente de los resultados de esta clasificación.

Cabe destacar que los problemas encontrados en la realización de la práctica finalmente no han tenido gran trascendencia debido a que se trataba de una dificultad inicial a la hora de interpretar los resultados obtenidos con la ejecución de Weka, que finalmente se han solventado a medida que la íbamos realizando y adquiriendo conocimiento sobre el funcionamiento de la misma.