T5 - TDIDT II - APRENDIZAJE COMPUTACIONAL

Andrés Matesanz

1. Introducción

En este ejemplo vamos a tratar de generar un árbol de decisión que nos ayude a seleccionar a aquellas personas a las que **conceder un crédito**. Este documento se complementa con un Jupyter Notebook donde se puden visualizar más fácilmente los pasos realizados: https://github.com/Matesanz/Decision-Tree. Para ello tenemos el siguiente dataset:

Pareja	Propiedades	Trabaja	Crédito	Concedido
FALSO	VERDADERO	VERDADERO	156.000,00	VERDADERO
VERDADERO	FALSO	FALSO	3.000,00	VERDADERO
FALSO	VERDADERO	FALSO	45.000,00	VERDADERO
FALSO	FALSO	VERDADERO	115.000,00	FALSO
VERDADERO	FALSO	FALSO	215.000,00	FALSO
FALSO	VERDADERO	VERDADERO	60.500,00	VERDADERO
FALSO	FALSO	FALSO	120.400,00	FALSO
VERDADERO	VERDADERO	VERDADERO	94.000,00	VERDADERO
VERDADERO	VERDADERO	FALSO	124.000,00	VERDADERO
FALSO	FALSO	FALSO	3.950,00	FALSO
FALSO	VERDADERO	FALSO	2.150,00	VERDADERO
VERDADERO	FALSO	VERDADERO	165.000,00	VERDADERO
FALSO	VERDADERO	VERDADERO	87.400,00	VERDADERO
FALSO	FALSO	FALSO	15.300,00	FALSO
VERDADERO	VERDADERO	VERDADERO	4.600,00	VERDADERO
FALSO	VERDADERO	VERDADERO	7.525,00	VERDADERO
VERDADERO	FALSO	FALSO	3.950,00	VERDADERO
FALSO	VERDADERO	FALSO	223.000,00	FALSO
VERDADERO	VERDADERO	VERDADERO	105.000,00	VERDADERO
VERDADERO	FALSO	VERDADERO	76.500,00	VERDADERO
FALSO	VERDADERO	FALSO	190.600,00	FALSO
VERDADERO	VERDADERO	VERDADERO	184.000,00	VERDADERO

Tabla 1: Dataset

2. Preprocesamiento de los datos

Antes de realizar ningún paso del árbol de decisiones, vamos a analizar nuestro conjunto de datos separando las etiquetas *y_dataset* (en este caso: crédito sí o no) del resto de variables o *x_dataset*. A continuación echamos un vistazo a la colinealidad y correlación de los datos:

	Pareja	**Propiedades**	**Trabaja**	**Crédito**	**Concedido**
Pareja	1.000000	-0.168790	0.182574	0.135363	0.427618
Propiedades	-0.168790	1.000000	0.277350	0.092033	0.424043
Trabaja	0.182574	0.277350	1.000000	0.125659	0.487950
Crédito	0.135363	0.092033	0.125659	1.000000	-0.273451
Concedido	0.427618	0.424043	0.487950	-0.273451	1.000000

Tabla 2: Correlación entre variables.

Podemos ver que hay cierta correlación entre si el interesado o la interesada tienen trabajo, pareja o propiedades y relación indirecta con la cantidad solicitada. Sin embargo **no hay relaciones fuertes entre las distintas variables** (mayor de 0.85), por lo que, en principio, no deberíamos prescindir de ninguna. En este ejercicio teórico planteado no se han dado colinealidades entre variables, sin embargo, durante el preprocesamiento de datos para árboles de decisión cuando se trabaja con grandes datasets o variables de las que desconocemos su significado, es una práctica altamente recomendada eliminar parte de la datos mientras mantenemos la información porque:

1. Evita duplicidades.
2. Trabajar con menos datos al tiempo que mantenemos la misma información es más eficiente
3. Se reduce el ruido.
4. Obtendremos árboles de decisión más sencillos.

En este caso mantenemos el dataset completo ya que, como hemos mencionado, no se dan las circunstancias para reducir datos.

3. Generación del árbol de decisión

El siguiente paso una vez tenemos listo nuestro dataset es calcular la entropía para cada conjunto de experiencias según atributos. Este tipo de tareas son sencillas gracias a herramientas como *scikit-learn* (ver Jupyter notebook adjunto). El árbol de decisión resultante es el siguiente:

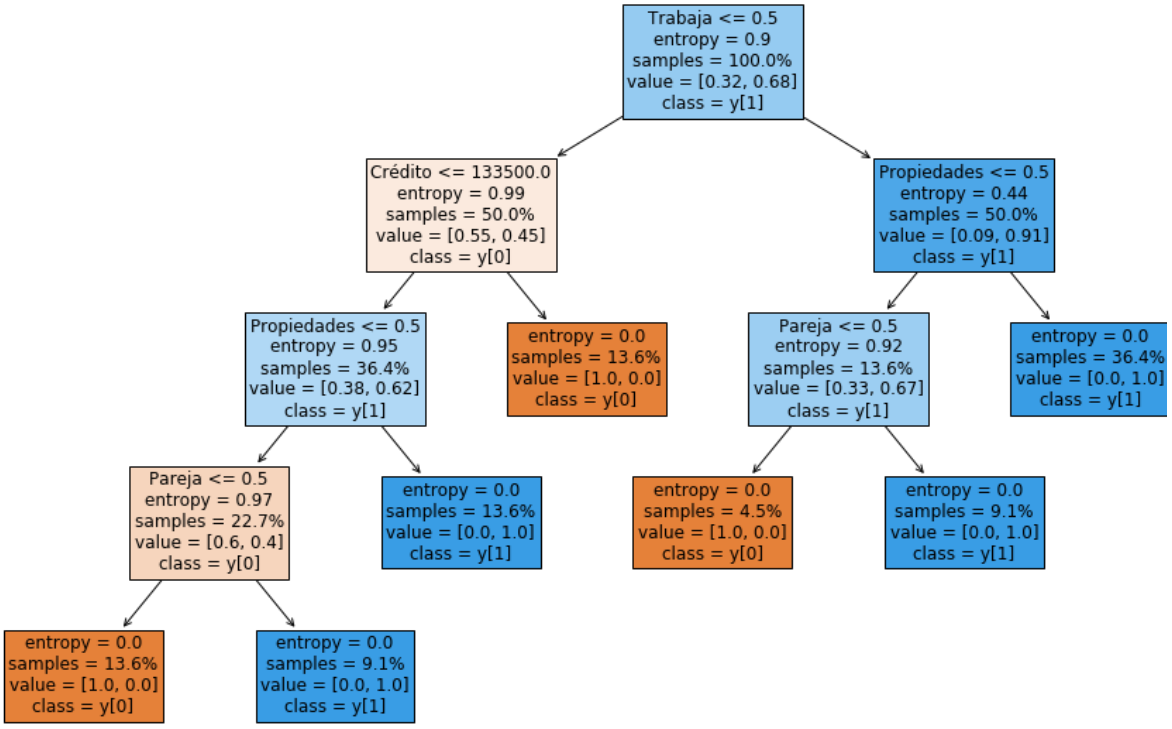

Imagen 1: Árbol de Decisión

En este caso podemos ver que **el mayor condicionante para la concesión de un crédito** depende de si el interesado **trabaja o no**: si a mayores posee propiedades o tiene pareja el crédito le es concedido (y[1]), sin valorar el montante solicitado, en esta situación se encuentra el 44,5% de los clientes. Por otra parte, si no trabaja, entra en otra casuística en la que se valora el importe solicitado al banco. Para ello **binarizamo**s el dataset en función del montante que piden los interesados: alto, si la cifra supera los 133.500€, y bajo, si la cifra es inferior.

Para cifras bajas **se concede** (y[1]) únicamente cuando el interesado **tiene pareja y/o propiedades**, lo que sucede en un 36,4% de las ocasiones. Mientras que para cifras altas el crédito es rechazado automáticamente en ausencia de trabajo.

De este árbol podemos observar que **la entropía llega a cero** en al menos una hoja de cada nodo, lo que indica un buena progresión sobre el árbol de decisiones, es decir, que hay una buena categorización. Recordemos que la

entropía es una medida objetiva del desorden a cada paso del árbol, por lo que, a menor entropía mejor son los resultados.

En segundo lugar tiene mucho sentido que el primer nodo sea el relativo a si el interesado trabaja o no ya que es el que mayor correlación tiene con la concesión o no de un crédito..

4. Cálculo de la entropía

A continuación vamos a analizar cómo y porqué se ha realizado este árbol de decisión calculando manualmente la entropía del primer nodo.

$$H = -\sum_i p_i \log_2 p_i$$

Calculamos la Entropía **del primer nodo** para a continuación obtener la del resto de atributos.

$P_{Concedido} = 7/22 = 0.318$ $P_{No\ Concedido} = 15/22 = 0.682$

$\mathbf{H_{Primer\ Nodo}} = - (0.318 \cdot \log_2 0.318) - (0.682 \cdot \log_2 0.682) = 0.526 + 0.376 = 0.902$

Siguiendo la definición de entropía vamos a calcularla para el atributo de si el interesado **trabaja o no**:

$P_{Sí\ Trabaja} = 9/22 = 0.643$ $P_{No\ Trabaja} = 5/22 = 0.357$

$H_{Sí\ Trabaja} = - (0.91 \cdot \log_2 0.91) - (0.09 \cdot \log_2 0.09) = 0.124 + 0.313 = 0.131$

$H_{No\ Trabaja} = - (0.545 \cdot \log_2 0.545) - (0.454 \cdot \log_2 0.454) = 0.477 + 0.517 = 0.994$

$\mathbf{H_{Trabaja}} = 0.131 \cdot 0.643 + 0.994 \cdot 0.357 = 0.084 + 0.355 = \mathbf{0.439}$

$\mathbf{\Delta_{Trabaja}} = H_{Primer\ Nodo} - H_{Trabaja} = 0.902 - 0.439 = \mathbf{0.463}$

Para si tiene **pareja o no:**

$P_{Sí\ Pareja} = 10/22 = 0.454$ $P_{No\ Pareja} = 12/22 = 0.545$

$H_{Sí\ Pareja} = - (0.9 \cdot \log_2 0.9) - (0.1 \cdot \log_2 0.1) = 0.137 + 0.332$

$H_{No\ Pareja} = - (0.5 \cdot \log_2 0.5) - (0.5 \cdot \log_2 0.5) = 1$

$\mathbf{H_{Pareja}} = 0.141 \cdot 0.454 + 1 \cdot 0.545 = 0.064 + 1.09 = \mathbf{0.61}$

$\mathbf{\Delta_{Pareja}} = H_{Primer\ Nodo} - H_{Trabaja} = 0.902 - 0.61 = \mathbf{0.292}$

Para si **tiene propiedades o no:**

$P_{Sí\ Propiedades}$ = 13/22 = 0.591 $P_{No\ Propiedades}$= 9/22 = 0.41

$H_{Sí\ Propiedades}$ = - (0.846 · $\log_2$ 0.846) - (0.154 · $\log_2$ 0.154) = 0.2 + 0.416 = 0.616

$H_{No\ Propiedades}$= - (0.555 · $\log_2$ 0.555) - (0.445 · $\log_2$ 0.445) = 0.471 + 0.52 = 0.991

$\mathbf{H_{Propiedades}}$ = 0.991 · 0.41 + 0.616 · 0.591 = 0.406 + 0.32 = **0.726**

$\mathbf{\Delta_{Propiedades}}$ = $H_{Primer\ Nodo}$ - $H_{Trabaja}$ = 0.902 - 0.726 = **0.176**

En el caso del atributo **crédito** vamos a **binarizar** entre aquellos que tienen piden más de 133.500€ o menos de 133.500€**:**

$P_{Credito\ Alto}$ = 5/22 = 0.227 $P_{Crédito\ Bajo}$= 17/22 = 0.772

$H_{Credito\ Alto}$ = -(0.6 · $\log_2$ 0.6) - (0.4 · $\log_2$ 0.4) = 0.442 + 0.529 = 0.971

$H_{Credito\ Bajo}$ = -(0.235 · $\log_2$ 0.235) - (0.765 · $\log_2$ 0.765) = 0.49 + 0.296 = 0.786

$\mathbf{H_{Crédito}}$ = 0.786 · 0.772 + 0.971 · 0.227 = 0.607 + 0.220 = **0.827**

$\mathbf{\Delta_{Crédito}}$ = $H_{Primer\ Nodo}$ - $H_{Trabaja}$ = 0.902 - 0.827 = **0.075**

Cada uno de los atributos produce un descenso en la entropía progresivamente menor, siendo la binarización en este caso la menos oportuna para el primer nodo. Finalmente es el atributo **trabaja** el que nos da una reducción mayor de entropía respecto al resto, concretamente pasando de 0.902 a 0.439 tal y como se ve reflejado en el árbol de decisión.

Para el resto de nodos se realiza el mismo procedimiento siguiendo las mismas reglas empleadas. Para visualizar las entropías y los nodos seleccionados a cada paso es recomendable analizar la *imagen 1.*

5. Modificación de la discretización

El siguiente paso consiste en elegir una discretización distinta a la propuesta en el apartado anterior, en este caso vamos a comparar el salto de entropía del atributo **crédito** cuando este se binariza como alto para importes superiores a **50.000€,** y como bajo para aquellos que no alcanzan dicha cantidad. En el apartado anterior como hemos podido comprobar la línea de corte se encontraba en 133.500€.

$P_{Credito\ Alto}$ = 11/22 = 0.5 $P_{Crédito\ Bajo}$= 11/22 = 0.5

$H_{Credito\ Alto}$ = -(0.454 · $\log_2$ 0.454) - (0.545 · $\log_2$ 0.545) = 0.517 + 0.477 = 0.994

$H_{\text{Credito Bajo}} = -(0.182 \cdot \log_2 0.182) - (0.818 \cdot \log_2 0.818) = 0.447 + 0.237 = 0.684$

$\mathbf{H_{Crédito}} = 0.5 \cdot 0.994 + 0.5 \cdot 0.684 = 0.497 + 0.342 = \mathbf{0.84}$

$\mathbf{\Delta_{Crédito}} = H_{\text{Primer Nodo}} - H_{\text{Trabaja}} = 0.902 - 0.84 = \mathbf{0.062}$

En nuestro caso la cifra aleatoria de 50.000€ **no modifica en gran medida** el descenso de entropía para el primer nodo **ni la continuación del árbol**

6. Utilizar otra medida distinta a la Entropía

En este apartado vamos a emplear Gini. El índice de Gini se calcula restando la suma de las probabilidades al cuadrado de cada clase. La ganancia de información multiplica la probabilidad de cada clase por el registro de esa probabilidad de clase.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

La ganancia de información favorece particiones más pequeñas con muchos valores distintos. Veamos cómo se comporta con nuestro dataset:

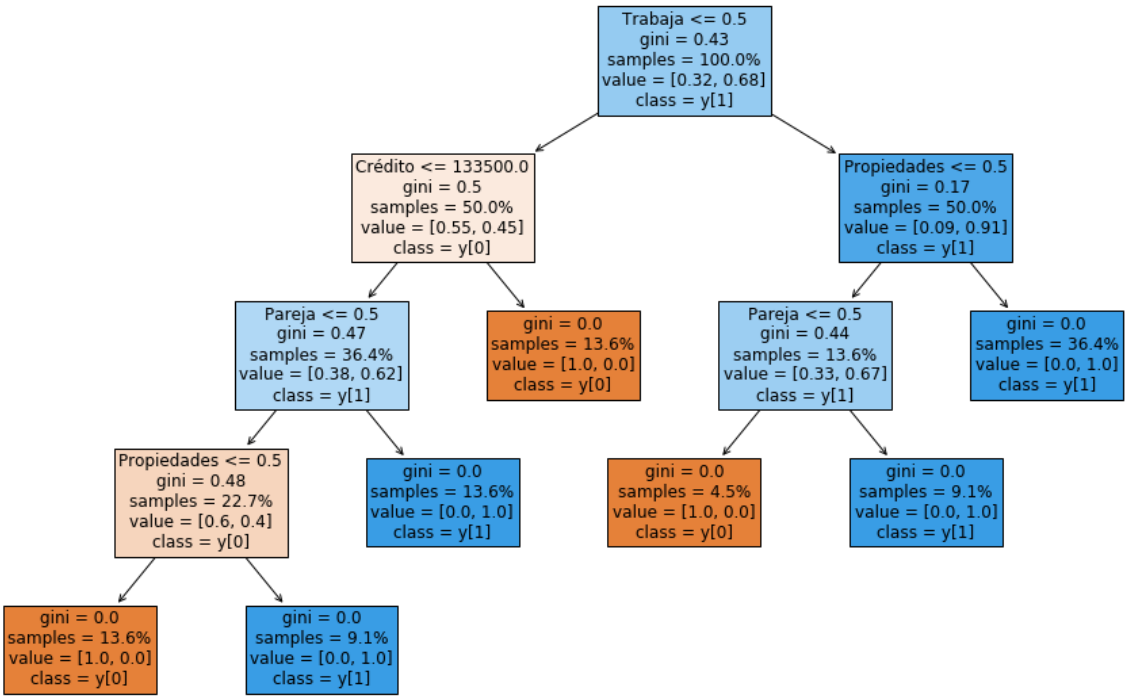

Imagen 2: 2º Árbol de decisión

Vemos que para un dataset tan pequeño, en el que todos los nodos terminan en hoja, el criterio gini **no hace variar el árbol de decisión**.

7. Conclusiones

Un árbol de decisiones es una herramienta muy útil a la hora de categorizar los datos y valorar qué acciones tomar en función de ellos. Sin embargo, su cálculo es un proceso tedioso que realizado manualmente puede llevar mucho tiempo, por ello existen a nuestra disposición librerías como scikit-learn que nos ayudan con este objetivo.