T2 - TDIDT - APRENDIZAJE COMPUTACIONAL

Andrés Matesanz

Octubre 2019

1. Introducción

En este ejemplo vamos a tratar de reconocer bajo qué circunstancias es más probable que se produzca una cesárea atendiendo a las características de la madre antes del parto. Para ello tenemos el siguiente dataset. Este documento se complementa con un Jupyter Notebook donde se puden visualizar más fácilmente los pasos realizados: https://github.com/Matesanz/Decisition-Tree.

Edad	Primeriza	Número de Hijos	Semana de gestación	Parto Múltiple	Cesárea
39	FALSO	1	42	FALSO	FALSO
25	FALSO	2	42	VERDADERO	FALSO
34	VERDADERO	1	41	FALSO	FALSO
42	FALSO	1	42	FALSO	FALSO
17	VERDADERO	1	42	FALSO	FALSO
22	FALSO	3	38	VERDADERO	VERDADERO
25	FALSO	1	42	FALSO	FALSO
35	VERDADERO	4	41	VERDADERO	VERDADERO
28	FALSO	2	34	VERDADERO	VERDADERO
36	FALSO	1	42	FALSO	FALSO
15	FALSO	1	42	FALSO	FALSO
31	FALSO	3	42	VERDADERO	FALSO
25	VERDADERO	2	42	VERDADERO	FALSO
40	FALSO	1	34	FALSO	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:

cesárea sí o no) del resto de variables o *x_dataset*. A continuación echamos un vistazo a la colinealidad y correlación de los datos:

	Edad	Primeriza	Número de Hijos	Semana de gestación	Parto Múltiple	Cesárea
Edad	1.000000	-0.140779	-0.061159	-0.156030	0.201589	0.129737
Primeriza	-0.140779	1.000000	0.188562	0.240192	0.091287	-0.50000
Número de Hijos	-0.061159	0.188562	1.000000	-0.086808	0.860663	0.518545
Semana de gestación	-0.156030	0.240192	-0.086808	1.000000	-0.182720	-0.824660
Parto Múltiple	-0.201589	0.091287	0.860663	-0.182720	1.000000	0.410792
Cesárea	129.737	-50.000	518.545	-824.660	410.792	1.000000

Tabla 2: Correlación entre variables.

Confirmamos que (obviamente) hay una **alta correlación** directa entre el número de recién nacidos y si el parto es múltiple o no. En este ejercicio teórico planteado es más que evidente que 2+ recién nacidos equivale a un parto múltiple, sin embargo, cuando se trabaja con grandes datasets o variables de las que desconocemos su significado, es una práctica altamente recomendada durante el preprocesamiento de datos para árboles de decisión 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 desechamos la variable *"parto múltiple"* ya que nos aporta menos información que el número de recién nacidos.

3. Generación del árbol de decisión

El siguiente paso una vez tenemos listo nuestro conjunto 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:

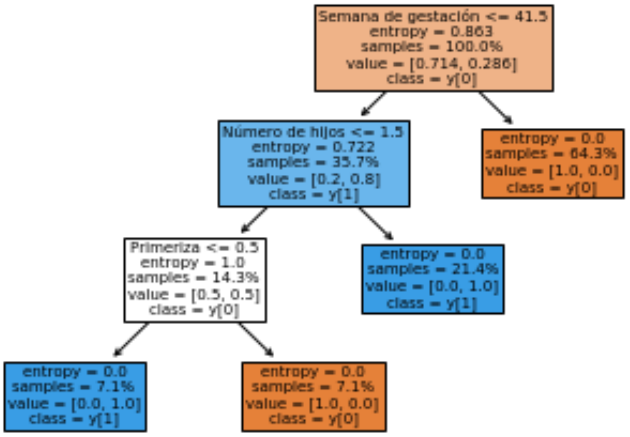

Imagen 1: Árbol de Decisión

De este diagrama podemos concluir que **el mayor condicionante para la realización de una cesárea es la semana de gestación**: si se ha completado el período natural de embarazo, 42 semanas, no será necesaria la intervención (*y[0]*) en el 64.3% de los casos (⅔ aproximadamente). Del 35.7% restante concluimos que el número de neonatos es de 2 o más habrá cesárea mientras que en el resto se verá influido por si la madre es primeriza o no: si ya ha sufrido un parto previamente se realizará intervención (*y[1]*).

Es decir, aunque parezca contraintuitivo, el número de recién nacidos en **menos determinante** que la semana de gestación. Algo que a priori mirando el dataset podría parecer todo lo contrario (parto múltiple = más probabilidades de cesárea). En segundo lugar vemos que generar una **binarización** del número de semanas es el método más efectivo para alcanzar la configuración más efectiva dentro del árbol de decisiones.

En tercer lugar 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 cuarto lugar tiene mucho sentido que el primer nodo sea el relativo a las semanas de gestación ya que es el que mayor correlación tiene con la realización o no de cesárea.

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.

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

Siguiendo la definición de entropía vamos a calcularla para el atributo **semana de gestación** (binarizado en: gestación completa, 42 semanas, o no):

$P_{\text{No Cesárea}} = 10/14 = 0.714$ $P_{\text{Cesárea}} = 4/14 = 0.286$

$P_{\text{Gestación Completada}} = 9/14 = 0.643$ $P_{\text{Gestación No Completada}} = 5/14 = 0.357$

$H_{\text{Primer Nodo}} = - (0.714 \cdot \log_2 0.714) - (0.286 \cdot \log_2 0.286) = 0.347 + 0.514 = 0.861$

$H_{\text{Gestación Completada}} = -(1 \cdot \log_2 1) = 0$

$H_{\text{Gestación No Completada}} = -(0.8 \log_2 0.8) -(0.2 \log_2 0.2) = 0.2576 + 0.464 = 0.7216$

$\mathbf{H_{\text{Semanas de gestación}}} = 0 \cdot 0.643 + 0.7216 \cdot 0.357 = \mathbf{0.257}$

$\Delta_{\text{Semanas de Gestación}} = H_{\text{Primer Nodo}} - H_{\text{Semanas de gestación}} = 0.861 - 0.257 = 0.603$

Para el **número de hijos** binarizando entre parto simple o múltiple

$P_{\text{Simple}} = 8/14 = 0.571$ $P_{\text{Múltiple}} = 6/14 = 0.429$

$H_{\text{Simple}} = -(0.875 \cdot \log_2 0.875) - (0.125 \cdot \log_2 0.125) = 0.168 + 0.375 = 0.543$

$H_{\text{Múltiple}} = -(0.5 \cdot \log_2 0.5) - (0.5 \cdot \log_2 0.5) = 2$

$\mathbf{H_{\text{Número de Hijos}}} = 0.571 \cdot 0.543 + 2 \cdot 0.429 = 0.31 + 0.858 = \mathbf{1.168}$

Para si la madre es **primeriza o no:**

$P_{\text{No Primeriza}} = 10/14 = 0.714$ $P_{\text{Sí Primeriza}} = 4/14 = 0.286$

$H_{\text{No Primeriza}} = -(0.7 \cdot \log_2 0.7) -(0.3 \cdot \log_2 0.3) = 0.36 + 0.5211 = 0.881$

$H_{\text{Sí Primeriza}} = -(0.75 \cdot \log_2 0.75) -(0.25 \cdot \log_2 0.25) = 0.31 + 0.5 = 0.81$

$\mathbf{H_{\text{Primeriza}}} = 0.81 \cdot 0.286 + 0.881 \cdot 0.714 = 0.232 + 0.629 = \mathbf{0.861}$

En cuanto al atributo **edad** podemos ver que en la tabla de correlación entre variables la colinealidad es de 0.12, un valor muy bajo, es decir, podemos concluir que la entropía será alta independientemente de cómo dividamos la variable (salvo que asignemos un nuevo nodo a cada edad: 17, 18, 19…, lo cual no tendría sentido ya que en un árbol de decisiones lo que se busca es la generalización).

Por tanto, comprobamos manualmente que el mejor primer paso es el de dividir el dataset entre si el tiempo natural de gestación se ha completado o no.

Para el resto de nodos se realiza el mismo procedimiento que acabamos de completar. Para visualizar las entropías y los nodos seleccionados a cada paso es recomendable analizar la *imagen 1.*

5. 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 conllevar mucho tiempo, por ello existen a nuestra disposición librerías como scikit-learn que nos ayudan con este objetivo.