CART

(Classification And Regression Trees)

Analítica de Datos, Universidad de San Andrés

Si encuentran algún error en el documento o hay alguna duda, mandenmé un mail a rodriguezf@udesa.edu.ar y lo revisamos.

1. Introducción a CART

El algoritmo CART (Classification And Regression Trees) es una técnica de aprendizaje automático utilizada para la clasificación y la regresión. Su principal objetivo es dividir un conjunto de datos en grupos homogéneos.

El índice de Gini es una medida de impureza o pureza de un nodo en un árbol de decisión. Se utiliza para evaluar la calidad de una división en el árbol. La fórmula del índice de Gini es:

$$\text{Gini} = 1 - \sum_{k=1}^{K} p_k^2,$$

donde p_k es la proporción de la clase k en el nodo. Un Gini de 0 indica un nodo puro (todas las instancias pertenecen a una sola clase), mientras que un Gini de 0.5 indica máxima impureza (las instancias están distribuidas uniformemente entre las clases).

En el caso de las variables categóricas el índice Gini se calcula más facilmente ya que es necesario nada más ver cuantos valores pertenecen a cada clase. En el caso de las variables continuas, el algoritmo CART busca los puntos de corte que minimizan el índice de Gini. Esto a mano y con pocos datos es fácil, pero en la práctica se hace impráctico. En computadora, el algoritmo de CART se fija para cada punto de corte el Gini. A mano nosotros lo haremos de una manera que no es óptima, pero que se acerca mucho.

A chequear cada punto y determinar si es útil o no sin tener en cuenta todo lo que viene después se le llama búsqueda **greedy** (voraz). Es como si en cada paso tomáramos la mejor decisión localmente, sin pensar en el futuro. Esto puede llevar a soluciones subóptimas, pero en muchos casos es suficiente.

2. Ejercicio 1

2.1. Datos

Tenemos la siguiente tabla donde queremos predecir si una persona es aprobada o no para un crédito. La variable objetivo es "Aprobado" y las variables predictoras son "Edad", "Ingreso", "Historial" y "Deuda Actual".

Edad	Ingreso	Historial	Deuda Actual	Aprobado
25	3.5	Bueno	10	Sí
40	5.0	Malo	20	No
30	4.2	Bueno	30	Sí
22	2.8	Malo	40	No
35	6.1	Bueno	15	Sí

2.2. Partición por Variable Objetivo

Para calcular el índice de Gini para la variable "Aprobado", consideramos las instancias en la tabla:

- **Aprobado = Sí**: 3 instancias.
- **Aprobado = No**: 2 instancias.

Calculamos entonces el índice de Gini, viendo que tenemos 5 instancias en total:

$$\text{Gini}_{\text{Aprobado}} = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0{,}48$$

Esto nos sirve para tener un benchmark de la calidad de la partición. A partir de este valor, intentaremos mejorarlo.

2.3. Partición por Edad

Cuando tenemos una variable continua, debemos ordenar los valores y calcular el gini en cada corte. Para nuestro algoritmo a mano, tomaremos en cuenta entonces solo los puntos de corte que son relevantes. En este caso, tenemos la siguiente tabla, ordenada de menor a mayor por Edad:

Edad	Aprobado
22	No
25	Sí
30	Sí
35	Sí
40	No

Notamos entonces dos puntos de corte, uno entre 22 y 25, y otro entre 35 y 40 (que es donde los valores objetivo cambian). Para calcular el índice de Gini vamos entonces a considerar la mitad entre los dos puntos de corte, es decir entre 22 y 25 tomamos 23.5, y entre 35 y 40 tomamos 37.5.

- **Edad** $\leq$ **23.5**: 1 instancias (0 aprobados, 1 aprobado).
- **Edad** $>$ **23.5**: 4 instancias (3 aprobados, 1 no aprobado).

$$\text{Gini}_{\leq 23{,}5} = 1 - \left[\left(\tfrac{0}{1}\right)^2 + \left(\tfrac{1}{1}\right)^2\right] = 0$$

$$\text{Gini}_{>23{,}5} = 1 - \left[\left(\tfrac{3}{4}\right)^2 + \left(\tfrac{1}{4}\right)^2\right] = 0{,}375$$

$$\Rightarrow \text{Gini}_{\text{Edad, } 23.5} = \tfrac{1}{5} \times 0 + \tfrac{4}{5} \times 0{,}375 = 0{,}3$$

- **Edad** $\leq$ **37.5**: 4 instancias (3 aprobados, 1 no aprobado).
- **Edad** $>$ **37.5**: 1 instancias (0 aprobados, 1 no aprobado).

$$\text{Gini}_{\leq 37{,}5} = 1 - \left[\left(\tfrac{3}{4}\right)^2 + \left(\tfrac{1}{4}\right)^2\right] = 0{,}375$$

$$\text{Gini}_{>37{,}5} = 1 - \left[\left(\tfrac{0}{1}\right)^2 + \left(\tfrac{1}{1}\right)^2\right] = 0$$

$$\Rightarrow \text{Gini}_{\text{Edad, } 37.5} = \tfrac{4}{5} \times 0{,}375 + \tfrac{1}{5} \times 0 = 0{,}3$$

2.4. Partición por Historial

Para calcular el índice de Gini para la variable "Historial", consideramos las instancias en la tabla. Tenemos que tener en cuenta si fue aprobado o no.

- **Historial = Bueno**: 3 instancias (3 aprobados, 0 no aprobados).
- **Historial = Malo**: 2 instancias (0 aprobados, 2 no aprobados).

$$\text{Gini}_{\text{Bueno}} = 1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right] = 0$$

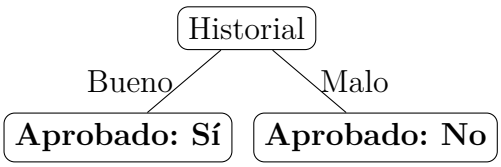
$$\text{Gini}_{\text{Malo}} = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$\Rightarrow \text{Gini}_{\text{Historial}} = \frac{3}{5} \times 0 + \frac{2}{5} \times 0 = 0$$

Dado que tenemos un Gini de 0 en este nodo, no es necesario continuar porque encontramos una forma de dividir perfectamente los datos. Un **Gini = 0** indica un nodo puro (división perfecta), mientras que **Gini = 0.5** indica máxima impureza (no se gana información). Si tenemos alguna categoría con un Gini = 0 entonces podemos generar un árbol de un solo nodo.

2.5. Árbol resultante

Para armar el árbol de decisión en este caso es fácil ya que tenemos una clase cuyo Gini es 0. Por lo tanto, podemos hacer un árbol de un solo nodo y no es necesario seguir dividiendo cada rama. Esto en la práctica jamás sucede, pero es un buen ejemplo para entender cómo funciona el algoritmo.



CART siempre nos garantiza que el árbol que encontremos sea el que menos nodos tenga, siempre y cuando hayamos buscado todos los Gini de cada punto de corte, y no como hicimos nosotros a mano.

3. Ejercicio 2

3.1. Datos

Tenemos la siguiente tabla:

Departamento	Antigüedad	Cumplimiento de Objetivos	Recibe Bono
Ventas	2	80	No
Soporte	5	60	No
Logística	3	75	Sí
Ventas	1	50	No
Soporte	4	85	Sí
Logística	6	70	No

3.2. Partición por Variable Objetivo

$$\text{Gini}_{\text{Recibe Bono}} = 1 - \left[\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right] = 0{,}444$$

3.3. Partición por Departamento

Cada departamento tiene 2 observaciones, pero con diferentes distribuciones de "Recibe Bono":

- **Ventas**: 2 observaciones (0 Sí, 2 No).
- **Soporte**: 2 observaciones (1 Sí, 1 No).
- **Logística**: 2 observaciones (1 Sí, 1 No).

$$\text{Gini}_{\text{Ventas}} = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$\text{Gini}_{\text{Soporte}} = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0{,}5$$

$$\text{Gini}_{\text{Logística}} = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0{,}5$$

$$\Rightarrow \text{Gini}_{\text{Departamento}} = \frac{2}{6} \times 0 + \frac{2}{6} \times 0{,}5 + \frac{2}{6} \times 0{,}5 = 0{,}333$$

3.4. Partición por Cumplimiento de Objetivos

De la misma manera que hicimos anteriormente, ordenamos los valores de "Cumplimiento de Objetivos" y encontramos los puntos de corte. En este caso, tenemos la siguiente tabla:

Cumplimiento de Objetivos	Recibe Bono
50	No
60	No
70	No
75	Sí
80	No
85	Sí

Los puntos de corte son entre 70 y 75, entre 75 y 80, y entre 80 y 85. Tomamos entonces los puntos 72.5, 77.5 y 82.5.

- **Cumplimiento de Objetivos $\leq$ 72.5**: 3 instancias (0 Sí, 3 No).
- **Cumplimiento de Objetivos $>$ 72.5**: 3 instancias (2 Sí, 1 No).

$$\text{Gini}_{\leq 72{,}5} = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = 0$$

$$\text{Gini}_{> 72{,}5} = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0{,}444$$

$$\Rightarrow \text{Gini}_{\text{Cumplimiento de Objetivos, } 72.5} = \frac{3}{6} \times 0 + \frac{3}{6} \times 0{,}444 = 0{,}222$$

- **Cumplimiento de Objetivos $\leq$ 77.5**: 4 instancias (1 Sí, 3 No).
- **Cumplimiento de Objetivos $>$ 77.5**: 2 instancias (0 Sí, 2 No).

$$\text{Gini}_{\leq 77{,}5} = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right] = 0{,}375$$

$$\text{Gini}_{> 77{,}5} = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0{,}5$$

$$\Rightarrow \text{Gini}_{\text{Cumplimiento de Objetivos, } 77.5} = \frac{4}{6} \times 0{,}375 + \frac{2}{6} \times 0{,}5 = 0{,}417$$

- **Cumplimiento de Objetivos $\leq$ 82.5**: 5 instancias (1 Sí, 4 No).

- **Cumplimiento de Objetivos** $>$ **82.5**: 1 instancia (1 Sí, 0 No).

$$\text{Gini}_{\leq 82,5} = 1 - \left[\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right] = 0{,}32$$

$$\text{Gini}_{>82,5} = 1 - \left[\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right] = 0$$

$$\Rightarrow \text{Gini}_{\text{Cumplimiento de Objetivos, 82.5}} = \frac{5}{6} \times 0{,}48 + \frac{1}{6} \times 0 = 0{,}4$$

3.5. Partición por Antigüedad

Ordenamos los valores de "Antigüedad" y encontramos los puntos de corte. En este caso, tenemos la siguiente tabla:

Antigüedad	Recibe Bono
1	No
2	No
3	Sí
4	Sí
5	No
6	No

Los puntos de corte son entre 2 y 3, y entre 4 y 5. Tomamos entonces los puntos 2.5 y 4.5.

- **Antigüedad** $\leq$ **2.5**: 2 instancias (0 Sí, 2 No).
- **Antigüedad** $>$ **2.5**: 4 instancias (2 Sí, 2 No).

$$\text{Gini}_{\leq 2,5} = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$\text{Gini}_{>2,5} = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0{,}5$$

$$\Rightarrow \text{Gini}_{\text{Antigüedad, 2.5}} = \frac{2}{6} \times 0 + \frac{4}{6} \times 0{,}5 = 0{,}333$$

- **Antigüedad** $\leq$ **4.5**: 4 instancias (2 Sí, 2 No).

- **Antigüedad $>$ 4.5**: 2 instancias (0 Sí, 2 No).

$$\text{Gini}_{\leq 4{,}5} = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0{,}5$$

$$\text{Gini}_{>4{,}5} = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$\Rightarrow \text{Gini}_{\text{Antigüedad, 4.5}} = \frac{4}{6} \times 0{,}5 + \frac{2}{6} \times 0 = 0{,}333$$

- *Observación*: **CART** analiza cada punto de corte en búsqueda del mejor índice de Gini. Sin embargo, realizar este análisis a mano es impráctico, por lo que en este ejemplo probamos únicamente con los puntos de corte más relevantes, como se hizo anteriormente.

Si hicieramos para el punto de corte 3.5 quedaría de la siguiente manera:

- **Antigüedad $\leq$ 3.5**: 3 instancias (2 Sí, 1 No).
- **Antigüedad $>$ 3.5**: 3 instancias (0 Sí, 3 No).

$$\text{Gini}_{\leq 3{,}5} = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0{,}444$$

$$\text{Gini}_{>3{,}5} = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = 0$$

$$\Rightarrow \text{Gini}_{\text{Antigüedad, 3.5}} = \frac{3}{6} \times 0{,}444 + \frac{3}{6} \times 0 = 0{,}222$$

Vemos que tiene un Gini tan bajo como el de "Cumplimiento de Objetivos" cuando tomamos 72,5 como punto de corte y por lo tanto es una buena opción para dividir los datos. Este punto específico sabía que era útil porque cuando corrí el algoritmo de CART con Python y estos datos nos arroja eso, pero a mano es imposible saber a menos que tengas ganas y tiempo de chequear cada punto. Lo importante acá es entender cómo funciona el algoritmo, y no tanto la implementación a mano (que jamás lo van a hacer excepto en el exámen).

3.6. Árbol resultante

Para armar el árbol de decisión, comenzamos con la variable que tiene el menor índice de Gini. La siguiente tabla muestra los índices de Gini calculados para cada partición:

Variable	Partición	Gini
Cumplimiento de Objetivos	72.5	0.222
Antigüedad	2.5	0.333
Antigüedad	4.5	0.333
Departamento	-	0.333
Cumplimiento de Objetivos	82.5	0.4
Cumplimiento de Objetivos	77.5	0.417

En este caso, la variable Cumplimiento de Objetivos (72.5) es la que mejor divide los datos. Luego iremos agregando variables a medida que necesitemos hasta que no quede ningún dato sin clasificar.