Minería de Datos

Árbol de decisión

José Horacio Ahuactzin García

Matrícula: 202224450

27 - agosto - 2025

Introducción

Un árbol de decisión es un modelo predictivo utilizado en minería de datos y aprendizaje automático para clasificar datos. Su estructura se asemeja a un árbol, donde cada nodo interno representa el resultado de esa prueba, y cada hoja representa una clase final o decisión tomada. Es una herramienta intuitiva que permite visualizar cómo se toma una decisión a partir de una serie de preguntas binarias.

En este ejercicio se utilizó el algoritmo ID3 (Iterative Dichotomiser 3), el cual selecciona en cada nivel del árbol el atributo que más reduce la entropía del conjunto de datos. La entropía es una medida matemática que representa el nivel de incertidumbre o desorden de un conjunto. Cuanto mayor sea la entropía, más mezclada están las clases; y cuanto menor sea, más homogéneo es el conjunto.

El objetivo principal de este ejercicio es construir un árbol de decisión a partir de una base de datos con cinco atributos binarios y una variable de clase multiclase. Para ello, se realiza manualmente el cálculo de entropía y ganancia de información para cada atributo, y se construye el árbol seleccionando los atributos que mejor separan las clases, con el fin de lograr una clasificación eficiente y comprensible.

Cálculo de Entropía General

Tabla de clases:

Clase	Frecuencia	Proporción
C1	2	0.167
C2	4	0.333
C3	4	0.333
C4	2	0.167

Fórmula:

$$H(S) = -\sum p_i \cdot \log_2(p_i)$$

$$H(S) = -(0.167 \cdot \log_2(0.167) + 0.333 \cdot \log_2(0.333) + 0.333 \cdot \log_2(0.333) + 0.167 \cdot \log_2(0.167))$$

$$H(S) = 1.918$$

Entropía condicional y ganancia de información por atributo

A2 (mejor atributo)

Valor	Clases	Proporciones	Entropía	Peso	Entropía Ponderada
0	C2 = 4, C4 = 2	0.667, 0.333	0.918	0.5	0.459
1	C3 = 4, C1 = 2	0.667, 0.333	0.918	0.5	0.459

A2 = 0

- *Proporciones:*
  - *C2:* $\frac{4}{6} = 0.667$
  - *C4:* $\frac{2}{6} = 0.333$
- *Fórmula de entropía:*
  - $H = -(0.667 \cdot \log_2(0.667) + 0.333 \cdot \log_2(0.333))$
- *`Resultado:*
  - $H = -(0.667 \cdot -0.585 + 0.333 \cdot -1.585) = 0.390 + 0.528 = 0.918$
- *Peso de grupo:*
  - $\frac{6}{12} = 0.5$
- *Entropía ponderada:*
  - $0.5 \cdot 0.918 = 0.459$

A2 = 1

- *Proporciones:*
  - *C3:* $\frac{4}{6} = 0.667$
  - *C1:* $\frac{2}{6} = 0.333$
- *Fórmula de entropía:*
  - $H = -(0.667 \cdot \log_2(0.667) + 0.333 \cdot \log_2(0.333)) = 0.918$
- *Peso de grupo:*
  - $\frac{6}{12} = 0.5$
- *Entropía ponderada:*
  - $0.5 \cdot 0.918 = 0.459$

Resultado Final

Entropía condicional: $H(S|A2) = 0.459 + 0.459 = 0.918$

Ganancia $= H(S) - H(S|A2) = 1.918 - 0.918 = 1.000$

A1

Valor	Clases	Proporciones	Entropía	Peso	Entropía Ponderada
1	C2 = 4, C1 = 2, C4 = 2	0.5, 0.25, 0.25	1.5	0.667	1.000
0	C3 = 4	1.0	0.0	0.333	0.000

A1 = 1

- Proporciones:
  - C2: $\frac{4}{8} = 0.5$
  - C1: $\frac{2}{8} = 0.25$
  - C4: $\frac{2}{8} = 0.25$

- Fórmula de entropía:
  - $H = -(0.5 \cdot \log_2(0.5) + 0.25 \cdot \log_2(0.25) + 0.25 \cdot \log_2(0.25))$

- Resultado:
  - $H = -(0.5 \cdot -1 + 0.25 \cdot -2 + 0.25 \cdot -2) = 0.5 + 0.5 + 0.5 = 1.5$

- Peso de grupo:
  - $\frac{8}{12} = 0.667$

- Entropía ponderada:
  - $0.667 \cdot 1.5 = 1.000$

A1 = 0

- Proporciones: 1.00
- Fórmula de entropía:
  - $H = -(1.0 \cdot \log_2(1.0)) = 0.0$

- Peso de grupo:
  - $\frac{4}{12} = 0.333$

- Entropía ponderada:
  - $0.333 \cdot 0 = 0.000$

Resultado Final

Entropía condicional: $H(S \mid A1) = 1.000 + 0.000 = 1.000$

Ganancia $= 1.918 - 1.000 = 0.918$

A3

Valor	Clases	Proporciones	Entropía	Peso	Entrotopía
0	C2 = 4, C3 = 4, C1 = 2	0.4, 0.4, 0.2	1.522	0.833	1.268
1	C4 = 2	1.0	0.0	0.167	0.000

A3 = 0

- Proporciones:
  - C2: $\frac{4}{10} = 0.4$
  - C3: $\frac{4}{10} = 0.4$
  - C1: $\frac{2}{10} = 0.2$
- Fórmula de entropía:
  - $H = -(0.4 \cdot \log_2(0.4) + 0.4 \cdot \log_2(0.4) + 0.2 \cdot \log_2(0.2))$
- Resultado:
  - $H = -(0.4 \cdot -1.322 + 0.4 \cdot -1.322 + 0.2 \cdot -2.322)$
  - $= 0.529 + 0.529 + 0.464 = 1.522$
- Peso de grupo:
  - $\frac{10}{12} = 0.833$
- Entropía ponderada:
  - $0.833 \cdot 1.522 = 1.268$

A3 = 1

- *Fórmula de entropía:*
  - $H = -(1.0 \cdot \log_2(1.0)) = 0.0$
- *Peso de grupo:*
  - $\frac{2}{12} = 0.167$
- *Entropía ponderada:*
  - $0.167 \cdot 0 = 0.000$

Resultado Final

Entropía condicional: $H(S \mid A3) = 1.268 + 0.000 = 1.268$

Ganancia $= 1.918 - 1.268 = 0.650$

A4

Valor	Clases (ambos lados iguales)	Proporciones	Entropía	Peso	Entropía Ponderada
1	C1 = 2, C2 = 4, C3 = 4, C4 = 2	mismas	1.918	0.5	0.959
0	C1 = 2, C2 = 4, C3 = 4, C4 = 2	mismas	1.918	0.5	0.959

A4 = 1

- *Proporciones:*
  - *C1:* 0.167
  - *C2:* 0.333
  - *C3:* 0.333
  - *C4:* 0.167

- *Fórmula de entropía;*
  - $H = -\left(0.167 \cdot \log_2(0.167) + 0.333 \cdot \log_2(0.333) + 0.333 \cdot \log_2(0.333) + 0.167 \cdot \log_2(0.167)\right) = 1.918$
- *Peso de grupo:*
  - $\frac{6}{12} = 0.5$
- *Entropía ponderada:*
  - $0.5 \cdot 1.918 = 0.959$

A4 = 0

- *Proporciones:*
  - *C1:* 0.167
  - *C2:* 0.333
  - *C3:* 0.333
  - *C4:* 0.167
- *Fórmula de entropía:*
  - $H = -\left(0.167 \cdot \log_2(0.167) + 0.333 \cdot \log_2(0.333) + 0.333 \cdot \log_2(0.333) + 0.167 \cdot \log_2(0.167)\right) = 1.918$
- *Peso de grupo:*
  - $\frac{6}{12} = 0.5$
- *Entropía ponderada:*
  - $\frac{6}{12} = 0.5$

Resultado Final

Entropía condicional: $H(S \mid A4) = 0.959 + 0.959 = 1.918$

Ganancia $= H(S) - H(S \mid A4) = 1.918 - 1.918 = 0.000$

A5

Valor	Clases	Proporciones	Entropía	Peso	Entropía Ponderada
1	C2, C3, C4, C1	0.29,0.29,0.29, 0.14	1.950	0.582	1.138
0	C2 = 2, C3 = 2, C1 = 1	0.4, 0.4, 0.2	1.522	0.417	0.634

A5 = 1

- Proporciones:
  - C2: 0.286
  - C3: 0.286
  - C4: 0.286
  - C1: 0.143
- Fórmula de entropía:
  - $H = -\left(0.286 \cdot \log_2(0.286) + 0.286 \cdot \log_2(0.286) + 0.286 \cdot \log_2(0.286) + 0.143 \cdot \log_2(0.143)\right) = 1.950$
- Peso de grupo:
  - $\frac{7}{12} = 0.582$
- Entropía ponderada:
  - $0.582 \cdot 1.950 = 1.138$

A5 = 0

- Proporciones:
  - C2: 0.4
  - C3: 0.4
  - C1: 0.2
- Fórmula de entropía:
  - $H = -\left(0.4 \cdot \log_2(0.4) + 0.4 \cdot \log_2(0.4) + 0.2 \cdot \log_2(0.2)\right) = 1.522$

- Peso de grupo:
  - $\frac{5}{12} = 0.417$
- Entropía ponderada:
  - $0.417 \cdot 1.522 = 0.634$

Resultado Final

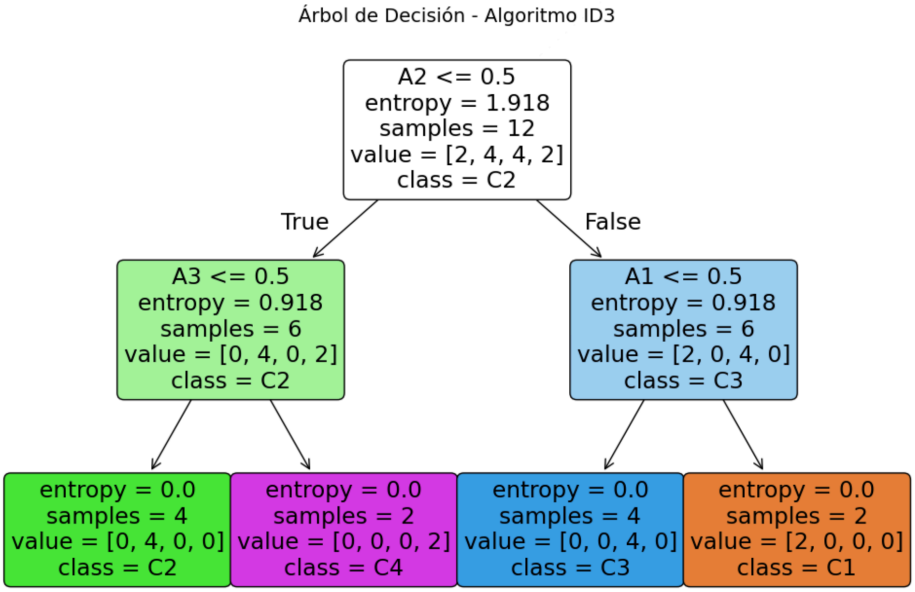
Entropía condicional: $H(S|A5) = 1.138 + 0.634 = 1.772$

Ganancia $= H(S) - H(S|A5) = 1.918 - 1.772 = 0.147$

Construcción del Árbol de Decisión (manual)

1. Se selecciona A2 como atributo raíz, ya que tiene la mayor ganancia de información (1.0).
2. Se divide el conjunto:
   - Si A2 = 0, se selecciona el siguiente mejor atributo: A3
     - Si A3 =0 -> todos los ejemplos son de clase C2 -> nodo hoja con clase C2 (entropía = 0)
     - Si A3 = 1 -> todos los ejemplos son de clase C4 -> nodo hoja con clase C4 (entropía = 00
   - Si A2 = 1 -> se selecciona el siguiente mejor atributo: A1
     - Si A1 = 0 -> todos los ejemplos son de clase C3 -> nodo hoja con clase C3 (entropía = 0)
     - Si A1 = 1 -> todos los ejemplos son de clase C1 -> nodo hoja con clase C1 (entropía = 0)

3. El árbol finaliza porque todas las hojas tienen entropía = 0, es decir, son puras y no se requiere seguir dividiendo.



Conclusión

Este proyecto me ayudó a comprender cómo una fórmula tan sencilla como la entropía puede tener un impacto tan importante en la toma de decisiones dentro de un modelo de clasificación. Aunque al principio parecía algo tedioso, al hacer los cálculos manualmente entendí con claridad por qué el árbol elige ciertos atributos y cómo se va formando paso a paso.

Algo que me sorprendió fue cómo el atributo A2 logró dividir tan bien los datos desde la raíz, haciendo que las demás divisiones fueran más amplias. Me di cuenta de que estos cálculos a mano no sólo sirve para aprender, sino que también me ayudó a interpretar mejor cómo funcionan los algoritmos de machine learning, más allá de simplemente correr código de Python.