Curso: Inteligencia Artificial

Unidad 2: Aprendizaje automático

Sesión 9: Aprendizaje de árboles de decisión

Docente: Carlos R. P. Tovar

Universidad Tecnológica del Perú

INICIO

¿Tienen dudas o consultas sobre la clase previa?

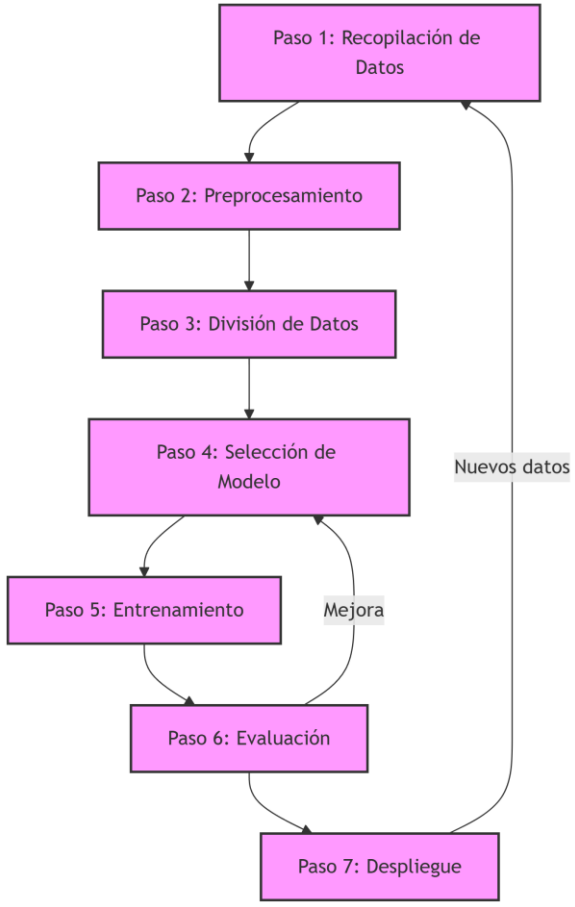
Objetivo de la sesión

Al finalizar la sesión, el alumno será capaz de:

- Comprender la teoría detrás de los árboles de decisión (impurezas, ganancia de información).
- Construir y evaluar árboles de clasificación y regresión.
- Aplicar estrategias para variables continuas/categóricas, missing values y poda.
- Implementar y visualizar árboles con Python y entender limitaciones/prácticas.

Pasos del Aprendizaje Supervisado

1. **Recopilación de Datos**
2. **Preprocesamiento**
3. **División de Datos**
4. **Selección de Modelo**
5. **Entrenamiento**
6. **Evaluación**
7. **Despliegue**



UTILIDAD
¿Por qué usar árboles de decisión?

Ventajas

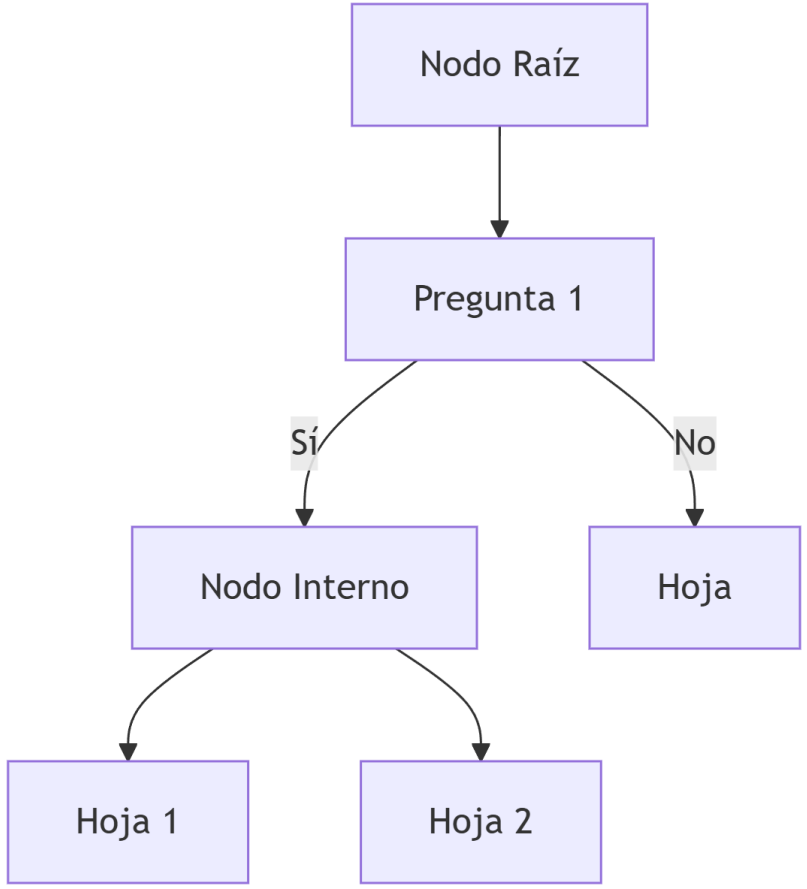
- Interpretabilidad.
- Flexibilidad.
- Sin necesidad de escalado.
- Base para modelos ensemble.

Casos de uso

- Diagnóstico médico.
- Scoring crediticio.
- Segmentación de clientes.
- Auditoría de procesos.

TRANSFORMACIÓN
¿Qué es un árbol de decisión?

- **Modelo predictivo supervisado**
- Estructura jerárquica: Nodo raíz, nodos internos y hojas.
- Recorrido de condiciones hasta la predicción.
- Parámetros clave.
- Analogía: Flujograma de decisiones médicas.



Algoritmos principales

Algoritmo	Año	Característica
ID3	1986	Usa ganancia de información
C4.5	1993	Maneja valores continuos
CART	1984	Usa índice Gini

Impureza: Entropía

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Donde:

- S: Conjunto de datos
- c: Número de clases
- p_i: Proporción de elementos de la clase i en S

Impureza: Entropía

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

- Ejemplo: Conjunto: [9 Sí, 5 No] →

$$H = -\frac{9}{14}\log_2 \frac{9}{14} - \frac{5}{14}\log_2 \frac{5}{14} = 0.940$$

Impureza: Índice Gini

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

Donde:

- S: Conjunto de datos
- c: Número de clases
- p_i: Proporción de elementos de la clase i en S

Matemáticas: Índice Gini

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

- Ejemplos:

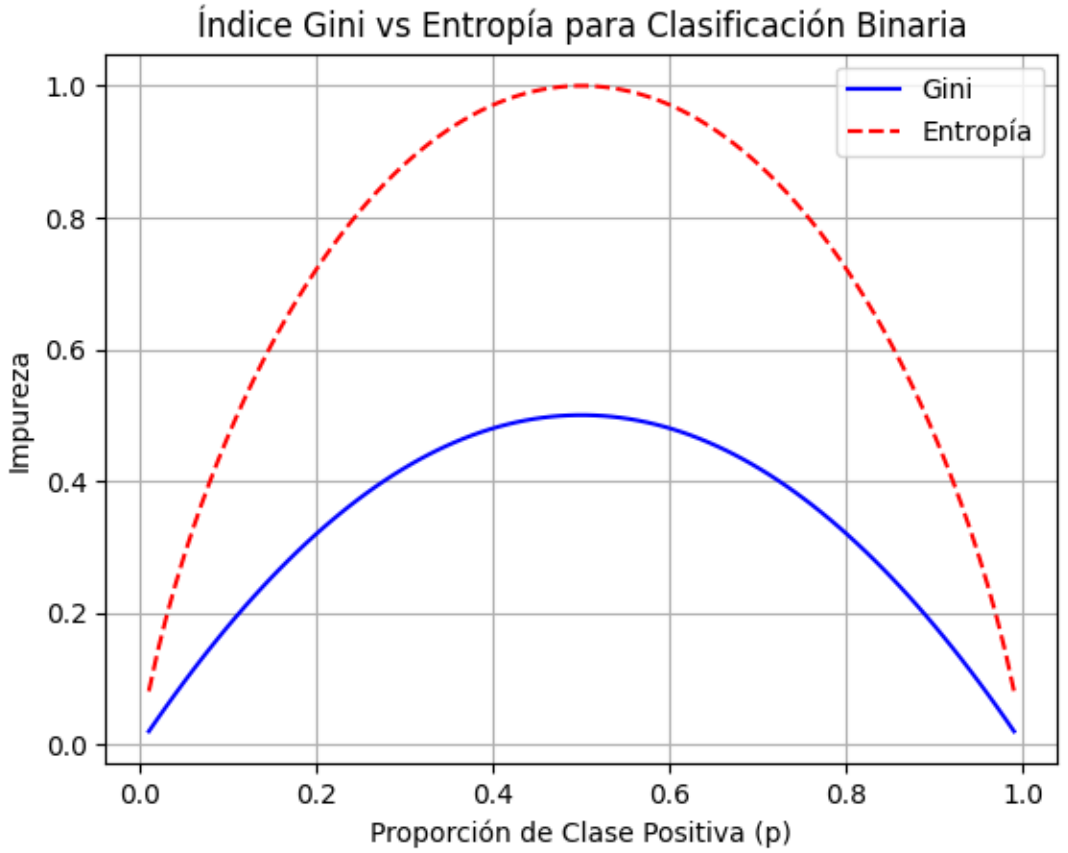
- Conjunto puro: [10 Sí, 0 No] → Gini = 0
- Conjunto mezclado: [5 Sí, 5 No] → Gini = 0.5

Matemáticas: Índice Gini vs Entropía

```
import numpy as np
import matplotlib.pyplot as plt

p = np.linspace(0.01, 0.99, 100)
gini = 1 - (p**2 + (1-p)**2)

plt.plot(p, gini, 'b-', label='Gini')
plt.plot(p, -p*np.log2(p) - (1-p)*np.log2(1-p), 'r--',
label='Entropía')
plt.title('Índice Gini vs Entropía para Clasificación Binaria')
plt.xlabel('Proporción de Clase Positiva (p)')
plt.ylabel('Impureza')
plt.legend()
plt.grid(True)
plt.show()
```



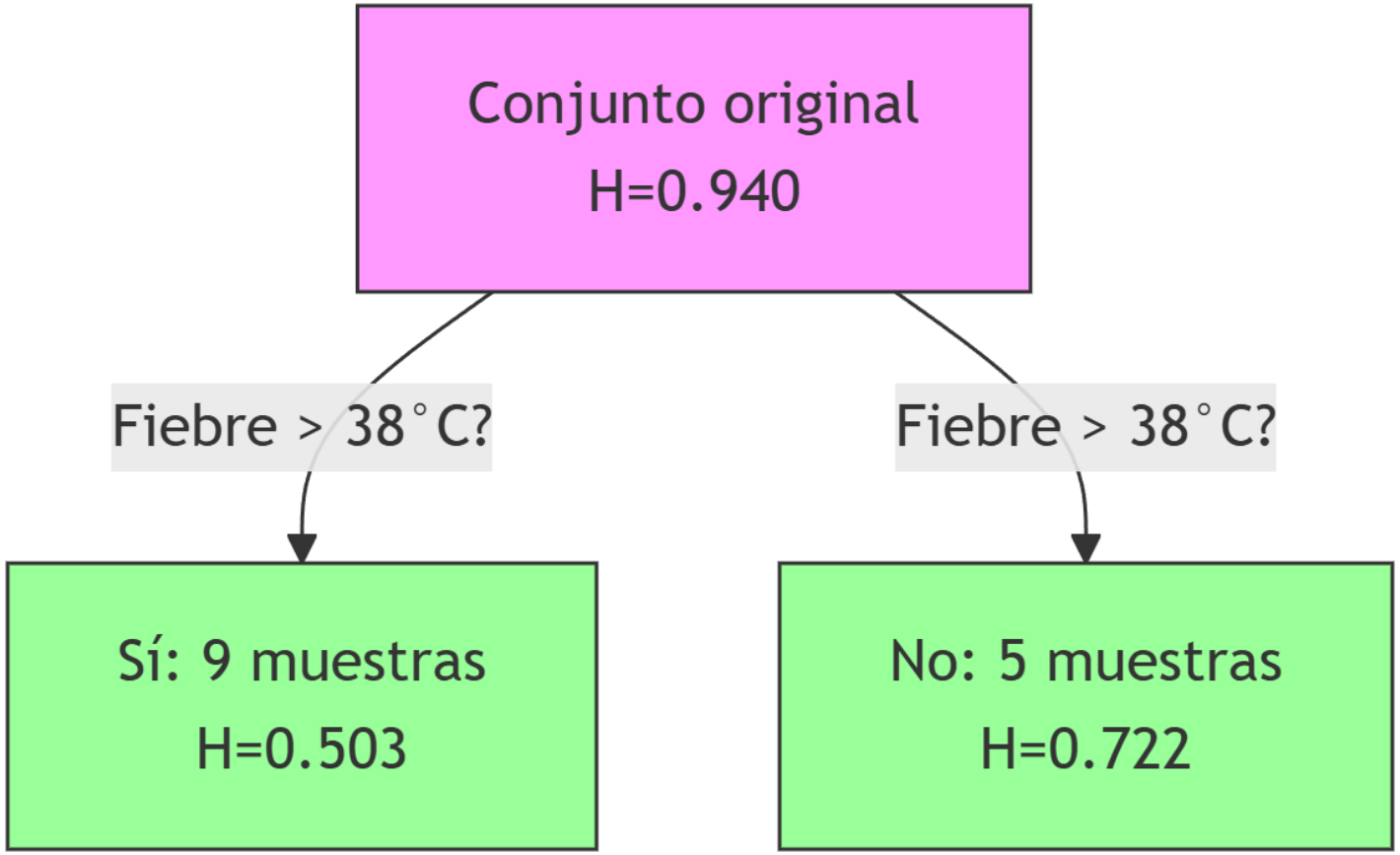
Matemáticas: Ganancia de Información

$$IG(S, A) = H(S) - \sum_{v \in Val(A)} \frac{|S_v|}{|S|} H(S_v)$$

Interpretación:

Reducción de incertidumbre al dividir por atributo A

Matemáticas: Ganancia de Información



Proceso de Aprendizaje

- Seleccionar mejor atributo (max IG o min Gini)
- Dividir dataset
- Repetir recursivamente
- Criterio de parada: pureza o profundidad máxima

Poda de árboles (Pruning)

¿Por qué podar?

- Evitar sobreajuste
- Simplificar modelo

Técnicas:

- **Pre-poda:** Limitar parámetros
- **Post-poda:** Eliminar ramas poco importantes

Ventajas y desventajas

Ventajas	Desventajas
Fácil interpretación	Inestable (peq. cambios datos)
Maneja datos categóricos	Tendencia a overfitting
Requiere poco preprocesamiento	Sesgo con clases dominantes

Cierre
Conclusiones

- Árboles = Modelos interpretables y versátiles
- Fundamentos matemáticos: Gini/Entropía
- Implementación práctica con sklearn

Recursos adicionales

- Scikit-learn Decision Trees
- Dataset UCI Repository
- Libro: "The Elements of Statistical Learning" (Cap. 9)

Universidad
Tecnológica
del Perú