Modelos de Clasificación II: Árboles de Decisión

Objetivo: Comprender y aplicar árboles de decisión para resolver problemas de clasificación de manera interpretable y práctica.

¿Por qué Árboles de Decisión?

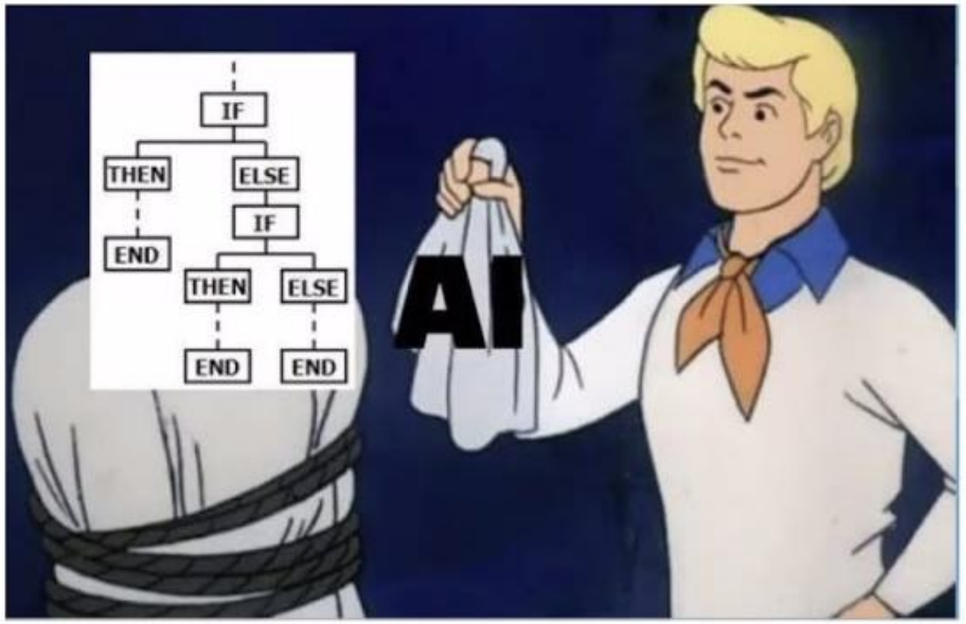
El poder de las decisiones paso a paso

Los árboles de decisión reflejan nuestra forma natural de pensar: tomamos decisiones siguiendo una secuencia lógica de preguntas que nos llevan a una conclusión.

Esta aproximación intuitiva los convierte en una herramienta poderosa tanto para principiantes como para expertos en machine learning.

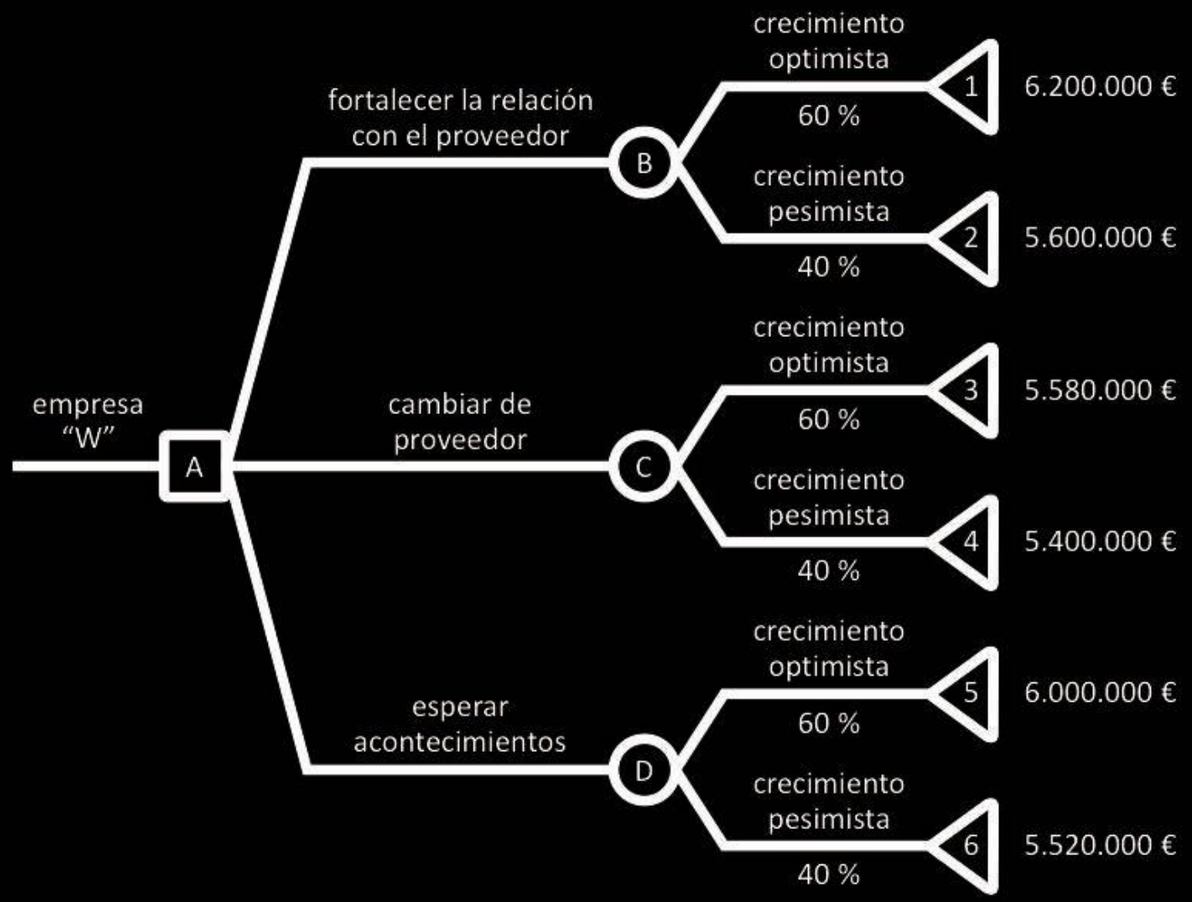
Ejemplo Clásico: Supervivencia en el Titanic


AI


IF
THEN
ELSE
IF
END
THEN
ELSE
END
END
AI

árbol de decisión

ejemplo práctico



Conceptos Fundamentales

Estructura de Árbol

Clasificador organizado como un **árbol binario** donde cada decisión divide los datos en dos grupos más homogéneos.

Nodos de Decisión

Cada nodo interno contiene una **condición** sobre una variable que divide el dataset según un criterio específico.

Hojas Predictivas

Los nodos terminales (hojas) contienen la **predicción final** de la clase para los ejemplos que llegan hasta allí.

Interpretabilidad

Principal ventaja: son **completamente interpretables** - podemos seguir el camino de decisión paso a paso.

Construcción del Árbol

Algoritmo de Construcción

El algoritmo busca de forma automática las **divisiones óptimas** en los datos, evaluando todas las posibles condiciones para encontrar la que mejor separe las clases.

Métricas de Impureza

Índice de Gini

Mide la **pureza** de un nodo. Un valor de 0 indica pureza perfecta (todas las muestras pertenecen a la misma clase).

Entropía

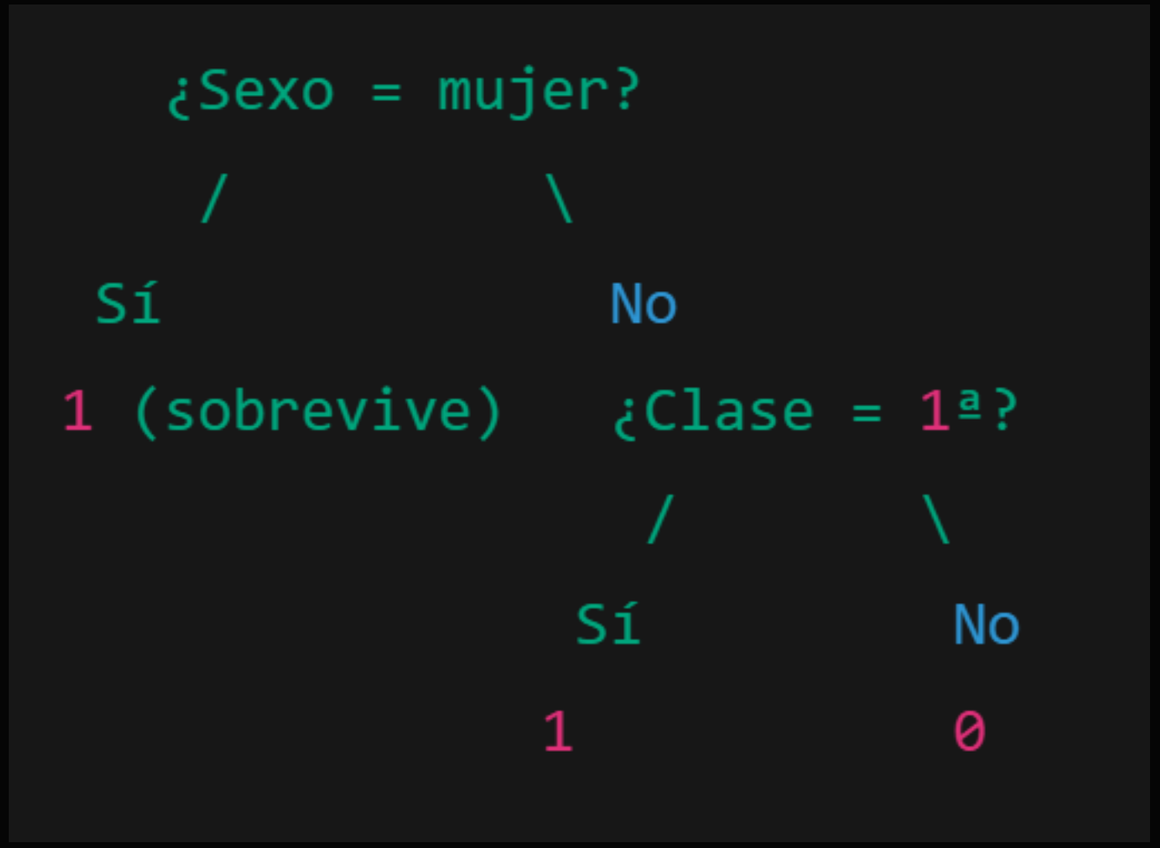
Cuantifica la **incertidumbre** en un nodo. Valores bajos indican mayor certeza en la predicción.

Criterios de Parada

El árbol detiene su crecimiento cuando:

- Todas las hojas son **puras** (una sola clase)
- Se alcanza la **profundidad máxima** definida
- No hay suficientes muestras para dividir
- La mejora es menor al umbral establecido

Ejemplo Visual: Supervivencia en el Titanic



Hiperparámetros Clave

max_depth

Profundidad máxima del árbol. Controla qué tan complejo puede volverse el modelo. Valores típicos: 3-10 para evitar sobreajuste.

min_samples_split

Mínimo de ejemplos requeridos para dividir un nodo interno. Valores más altos previenen divisiones en grupos pequeños.

min_samples_leaf

Mínimo de ejemplos que debe contener una hoja. Garantiza que las predicciones se basen en suficientes datos.

criterion

Medida de impureza utilizada para evaluar las divisiones. Opciones: 'gini' (por defecto) o 'entropy' para mayor precisión.

Ventajas y Desventajas

✅ Fortalezas

Interpretabilidad

Fáciles de visualizar y explicar. Cada decisión puede ser trazada y justificada paso a paso.

Flexibilidad

Capturan relaciones no lineales complejas sin necesidad de transformaciones matemáticas.

Simplicidad

No requieren escalado de variables ni preprocesamiento extensivo de los datos.

⚠️ Limitaciones

Sobreajuste

Tienden a sobreajustar si no se controla su crecimiento con hiperparámetros apropiados.

Inestabilidad

Sensibles a cambios pequeños en los datos de entrenamiento, lo que puede cambiar completamente la estructura.

Métricas de Evaluación

Métricas Estándar

Los árboles de decisión se evalúan con las mismas métricas que otros clasificadores, proporcionando una base común para la comparación.

- **Accuracy**

  Proporción de predicciones correctas sobre el total

- **Precision & Recall**

  Medidas de precisión y exhaustividad para cada clase

- **F1-Score**

  Media armónica entre precisión y recall

- **AUC-ROC**

  Área bajo la curva ROC para evaluar discriminación

Herramientas Específicas

Matriz de Confusión

Visualiza errores de clasificación por clase, especialmente útil para identificar confusiones sistemáticas.

Importancia de Características

¿Qué variables pesan más? Los árboles calculan automáticamente qué características son más importantes para las decisiones.

Ejemplo Práctico: Dataset Titanic

1

Definir Target

Variable objetivo: 'Survived' (0 = no sobrevivió, 1 = sobrevivió)

2

Seleccionar Features

Variables predictoras: edad, sexo, clase, tarifa, tamaño de familia, puerto de embarque

3

Entrenar Modelo

DecisionTreeClassifier con hiperparámetros optimizados para evitar sobreajuste

4

Evaluar y Visualizar

Análisis completo: métricas de rendimiento y visualización del árbol resultante

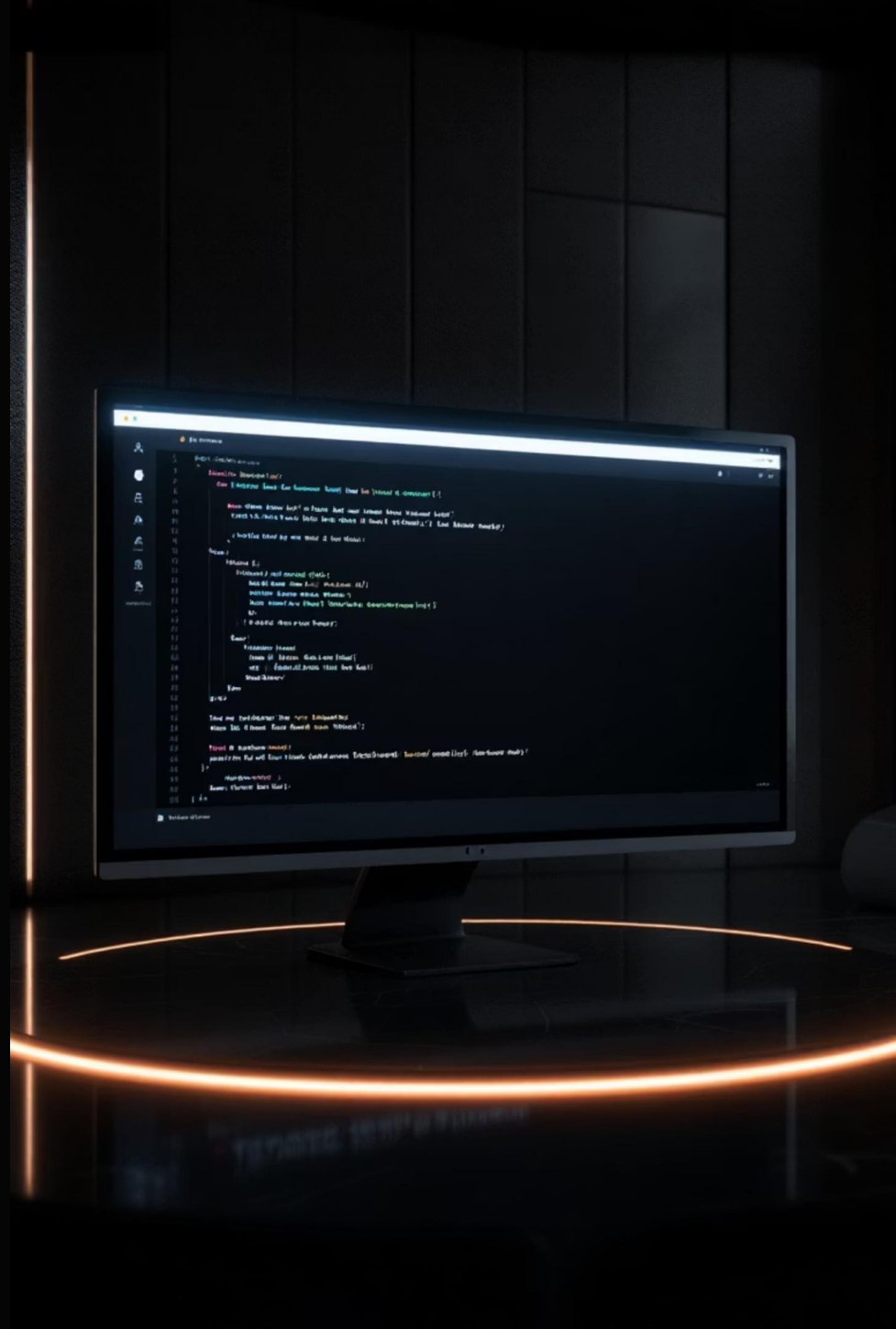
Configuración Recomendada

```
DecisionTreeClassifier(    max_depth=5,
min_samples_split=20,    min_samples_leaf=10,
criterion='gini',    random_state=42)
```

Resultados Esperados

Con esta configuración, típicamente obtenemos:

- **Accuracy:** ~82-85%
- **Interpretabilidad:** Excelente
- **Variables importantes:** Sexo, Clase, Edad



Conclusiones y Próximos Pasos

Fundamento de Algoritmos Avanzados

Los árboles son la **base de métodos ensemble** como Random Forest y Gradient Boosting, que combinan múltiples árboles para mayor precisión.

Herramienta de Explicabilidad

Muy útiles para explicar decisiones a stakeholders no técnicos, especialmente en dominios regulados como medicina o finanzas.

Importancia de la Regularización

Deben regularizarse cuidadosamente para evitar sobreajuste, especialmente con datasets pequeños o ruidosos.

Estrategia Práctica

En la práctica: comenzar con árboles simples para entender los datos, luego combinar con ensambladores para maximizar el rendimiento.

Takeaway clave: Los árboles de decisión son tu puerta de entrada al mundo de machine learning interpretable y la base para técnicas más avanzadas. ¡Domínalos y tendrás una herramienta poderosa para toda la vida!