Pontificia Universidad
JAVERIANA
Bogotá

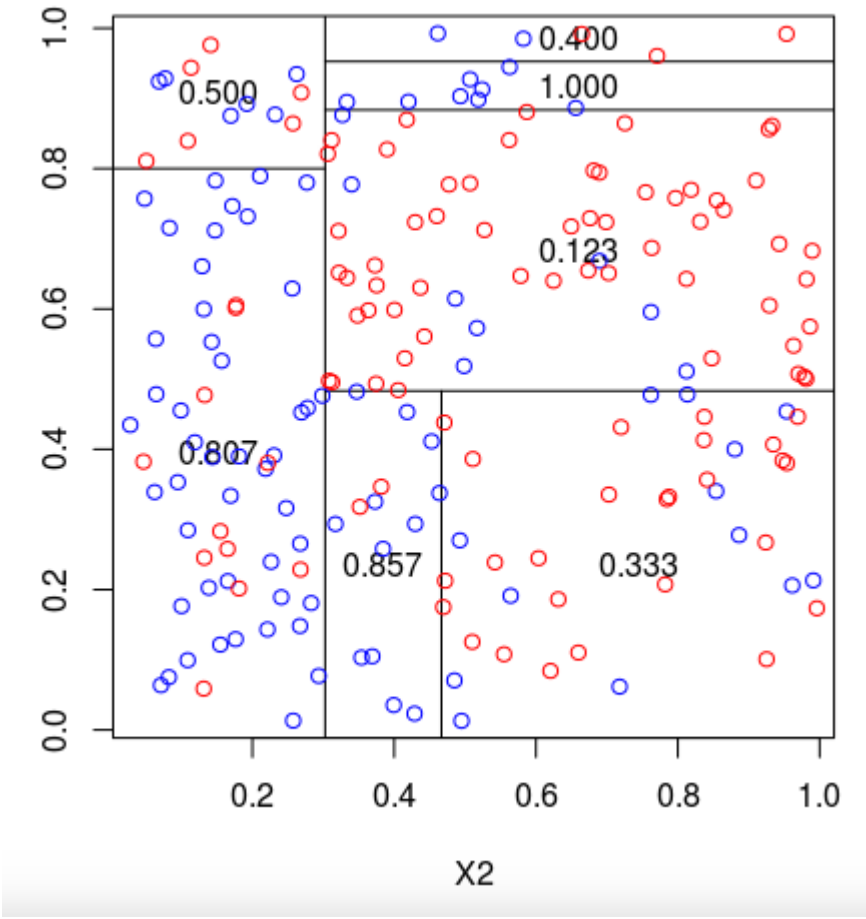
VIGILADA MINEDUCACIÓN

Árboles de Clasificación y Regresión

Pontificia Universidad
JAVERIANA
Colombia

Ideas Fundamentales

- **Objetivo y mecanismo de la tarea:** Tarea de clasificación/estimación que realiza particiones recursivas de las variables predictoras en un conjunto de rectángulos y crea un modelo simple (como una constante) en cada uno que maximizan (o minimizan) una medida de desempeño dada



Fuente: https://www.datacamp.com/community/tutorials/decision-trees-R

Ventajas y desventajas

Ventajas

- Interpretabilidad
- Flexibilidad para el manejo de predictores categóricos y valores perdidos
- Manejo de interacciones
- Particiones binarias (disminuyen la fragmentación de datos)
- Matrices de costos (reemplazo/complemento de balanceo)

Desventajas

- Inestabilidad de árboles (alta varianza: Necesidad de ensambles)
- Variables con muchas categorías (posible sesgo)
- Riesgo de overfitting dependiendo del tamaño del árbol

Árboles de Decisión: Técnica

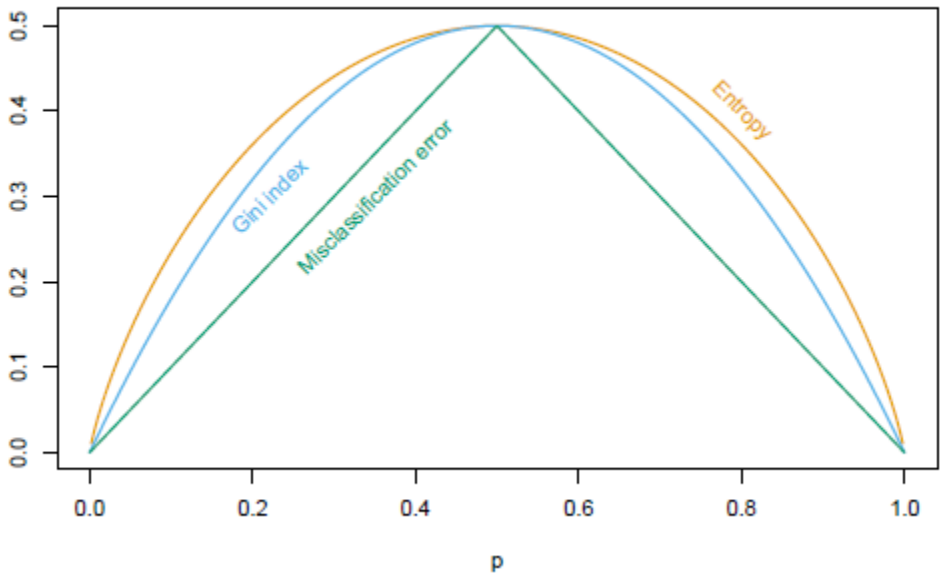
- **CHAID** está basada en tablas de contingencia (Chi-cuadrado). Detiene el crecimiento del árbol basado en p values. Debe ajustar por Bonferroni

- **CART** está basada en algoritmos. Crea el árbol completo y luego lo reduce (poda) hasta hallar el árbol que minimiza el error de clasificación (métrica de impureza) dado un tamaño de árbol basado en costo de la complejidad (número de ramas). Se debe definir el criterio de poda de error en el riesgo. Usa un hiperparámetro de costo de complejidad α.

Métricas de Impurezas

$$\text{entropy}(A) = -\sum_{k=1}^{m} p_k \log_2(p_k)$$

- Se utilizan tres posibles métricas de impureza para crear/podar el árbol:
  - **Information gain** (basado en teoría de la información - entropía)
  - **Índice Gini** $I(A) = 1 - \sum_{k=1}^{m} p_k^2,$
  - **Error de clasificación** (para podar solamente - no diferenciable)



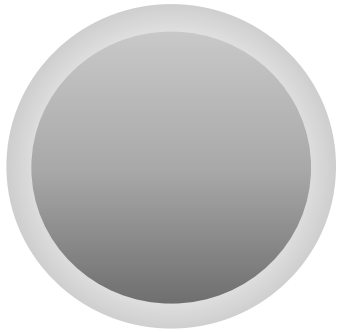
Fuente: Figura 9.3 . Hastie et al. (2017)

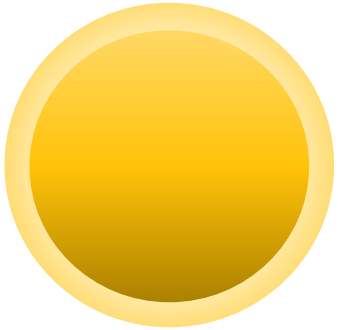
Comparación de Métricas de Impureza

Métrica	Formulación	Máximo	Observaciones
Gini	$1 - \sum_{k=1}^{m} p_k^2$	$\frac{m-1}{m}$	Menor implica más capacidad predictiva
Entropía	$-\sum_{k=1}^{m} p_k * log_2 p_k$	$\log_2(m)$	Menor implica más capacidad predictiva
Chi-Cuadrado	$\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{(n_{ij} - np_ip_j)^2}{np_ip_j}$	No acotado	Mayor implica más capacidad predictiva

Detención del Árbol

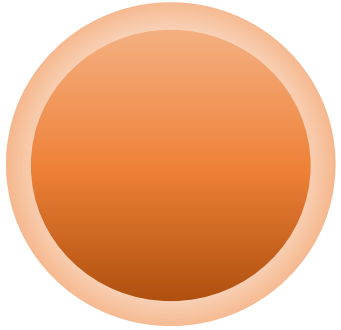
Detener por mínimo número de casos en rama para partición

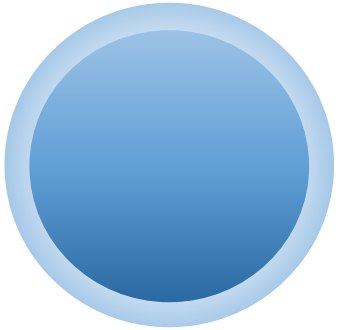




Detener por Cp

Detener por mínimo número en la hoja





Detener por máxima profundidad del árbol