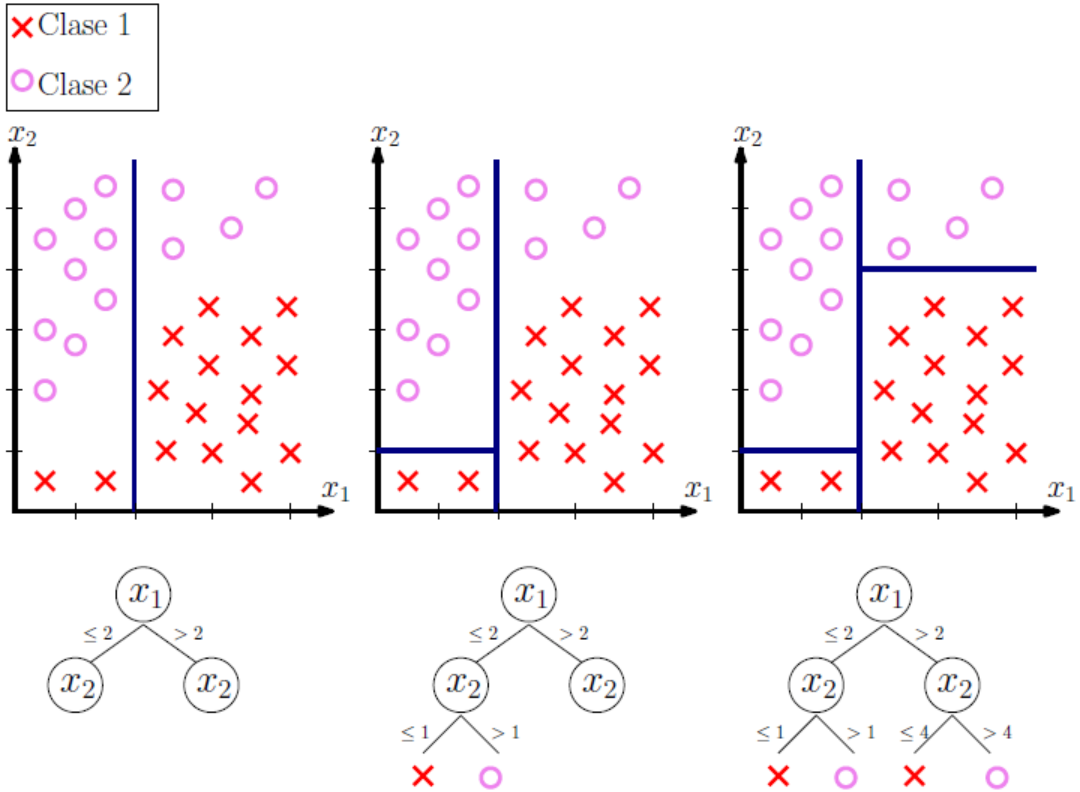
Módulo 3: Aprendizaje Supervisado

3.4. Árboles de decisión

Rafael Zambrano
rafazamb@gmail.com

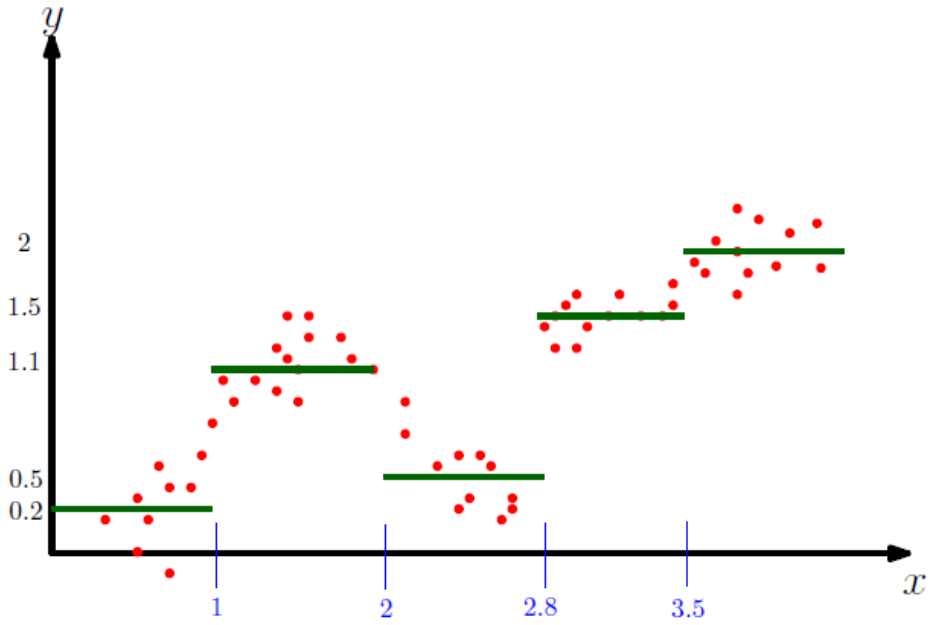
Árboles de decisión

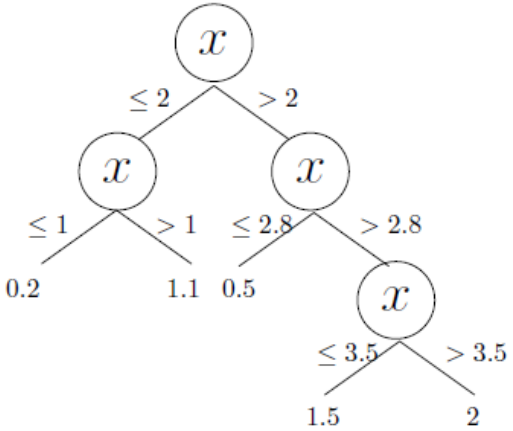
Clasificación



Árboles de decisión

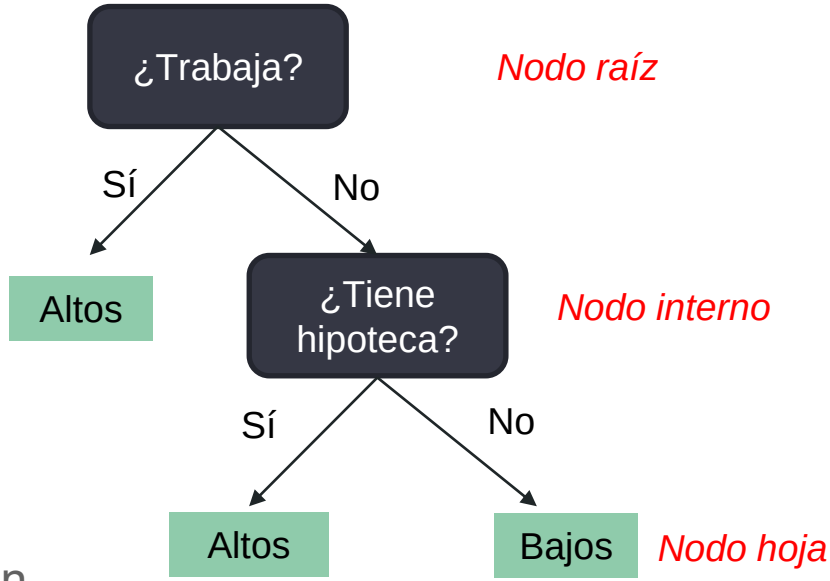
Regresión





Árboles de decisión

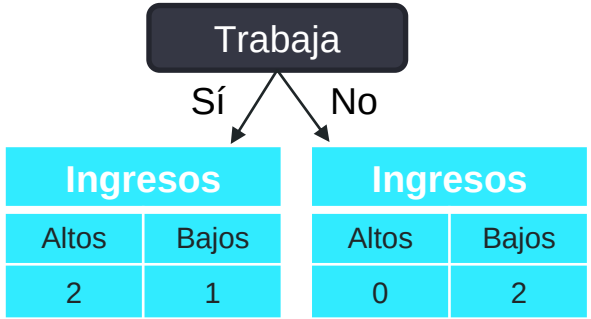
Cliente	Edad	Trabaja	Hipoteca	Ingresos
A	32	SÍ	SÍ	Altos
B	25	SÍ	SÍ	Altos
C	48	NO	NO	Bajos
D	67	NO	SÍ	Bajos
E	18	SÍ	NO	Bajos

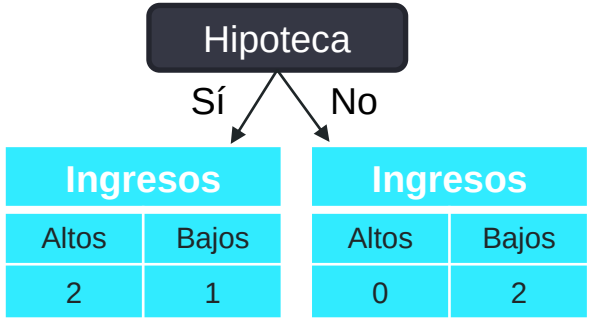


- Pueden utilizarse en regresión y clasificación
- ¿Qué variable utilizar para segmentar en cada nodo?

Árboles de decisión

Cliente	Edad	Trabaja	Hipoteca	Ingresos
A	32	SÍ	SÍ	Altos
B	25	SÍ	SÍ	Altos
C	48	NO	NO	Bajos
D	67	NO	SÍ	Bajos
E	18	SÍ	NO	Bajos

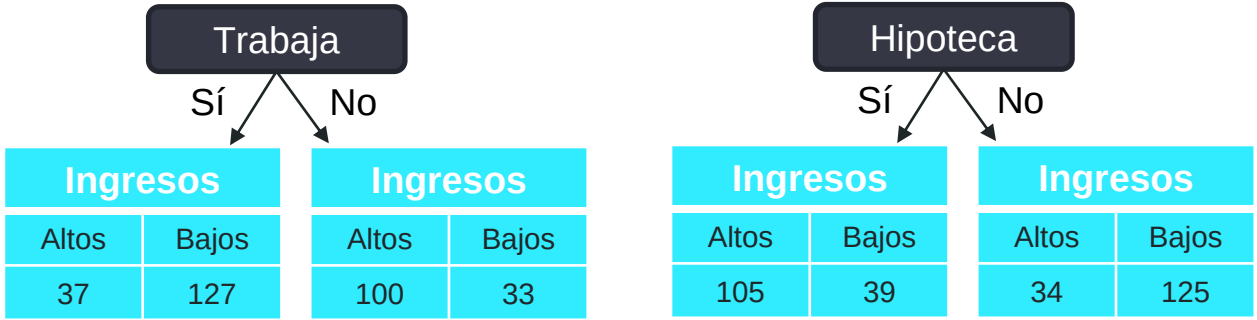




- Hay que medir cómo de bien separan las variables candidatas a la variable objetivo

Árboles de decisión

- Imaginemos que para un dataset completo obtenemos los siguientes resultados

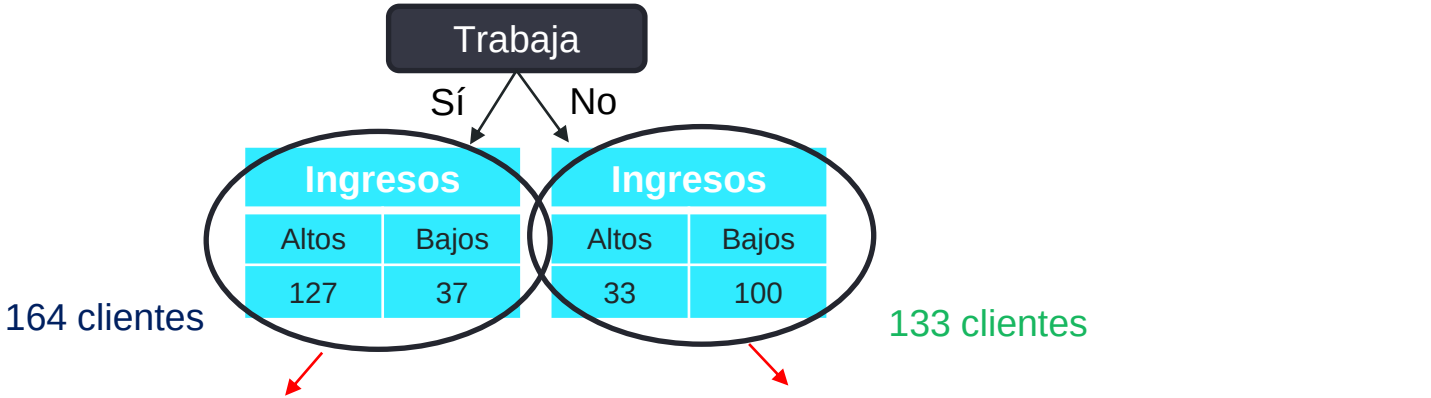


- Normalmente, ninguna de las variables consigue separar perfectamente a la variable objetivo (existe impureza)

- La métrica más común para medir impurezas se conoce como “**Gini**”

Árboles de decisión

- Impureza de Gini, para cada nodo hoja:
  1- (probabilidad de la clase 1)^2 – (probabilidad de la clase 2)^2



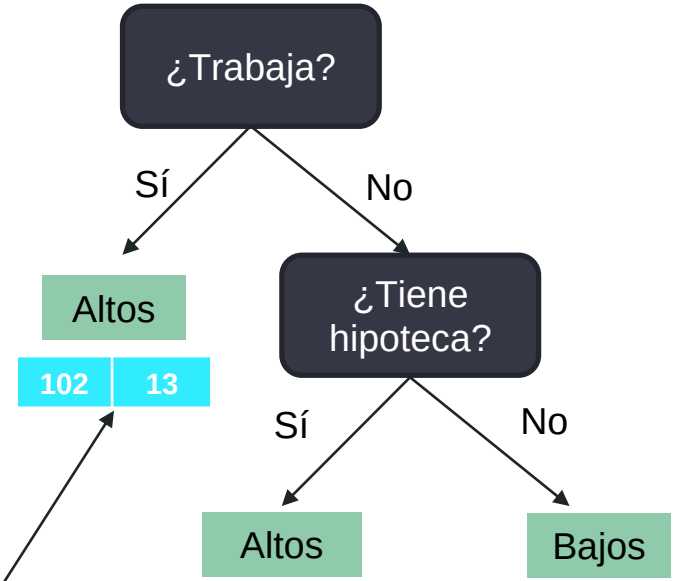
$$1 - \left(\frac{127}{127+37}\right)^2 - \left(\frac{37}{37+127}\right)^2 = 0.35 \qquad 1 - \left(\frac{33}{33+100}\right)^2 - \left(\frac{100}{33+100}\right)^2 = 0.37$$

- Impureza de Gini total de la variable “Trabaja”:

$$0.35 \cdot \left(\frac{164}{164+133}\right) + 0.37 \cdot \left(\frac{133}{164+133}\right) = \mathbf{0.36}$$

Árboles de decisión

- Impureza de Gini total de la variable “Trabaja”: 0.360
- Impureza de Gini total de la variable “Hipoteca”: 0.364
- La variable “trabaja” tiene menos impureza, por lo que funciona mejor a la hora de separar la variable objetivo, utilizándose como nodo raíz
- Este proceso se repite en los nodos intermedios con las variables distintas a la del nodo raíz
- Un nodo se convierte en hoja cuando ninguna variable separa mejor el resultado de ese nodo



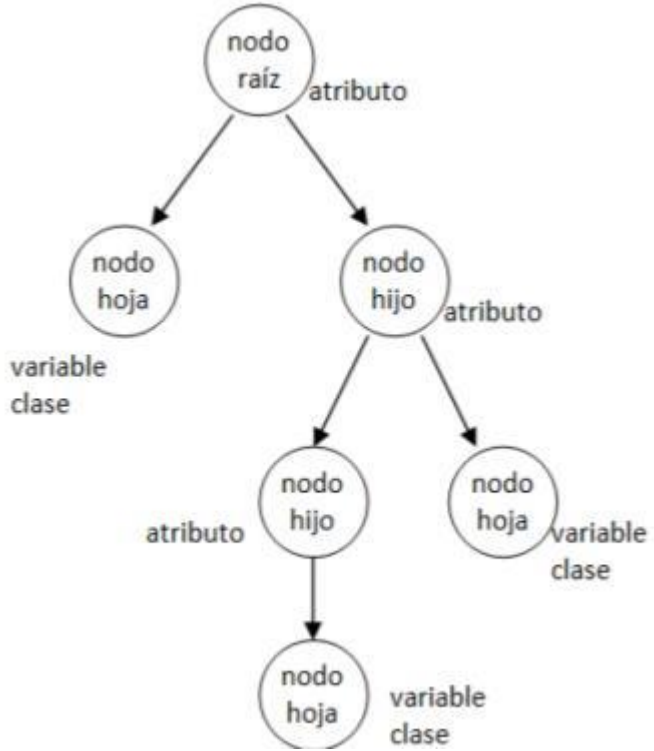
Gini (nodo): $1 - \left(\frac{102}{102+13}\right)^2 - \left(\frac{13}{102+13}\right)^2 = 0.2$

Gini (“Hipoteca”) >0.2

Árboles de decisión

Pasos para construir un árbol de decisión:

1. Calcular el índice de Gini para cada variable
2. Si el nodo en sí tiene el menor Gini, se convierte en hoja
3. Si utilizar una variable para separar mejora el resultado, se utilizará la variable con el menor Gini



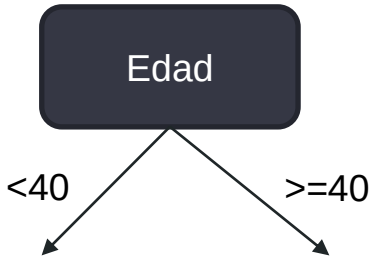
Árboles y variables numéricas

Edad	Ingresos
32	Altos
25	Altos
48	Bajos
67	Bajos
18	Bajos

¿Cómo determinar cuál es el mejor corte para dividir el target?

1) Ordenar de menor a mayor
2) Calcular la media para pares adyacentes
3) Calcular el índice Gini para cada media
4) Escoger el corte que tenga el menor Gini

Edad	Ingresos
18	Bajos
21.5	
25	Altos
28.5	
32	Altos
40	
48	Bajos
57.5	
67	Bajos

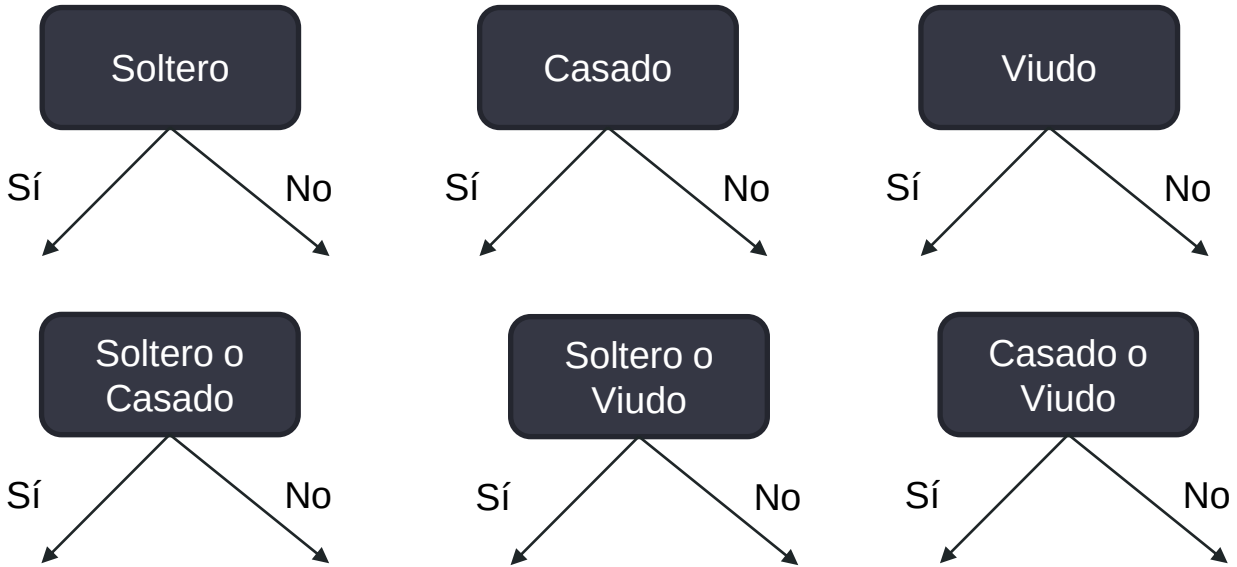


Gini (21.5) = 0.3
Gini (28.5) = 0.47
Gini (40) = 0.27
Gini (57.5) = 0.4

Árboles y variables categóricas

Estado civil	Ingresos
Soltero	Altos
Casado	Altos
Viudo	Bajos
Casado	Bajos
Soltero	Bajos

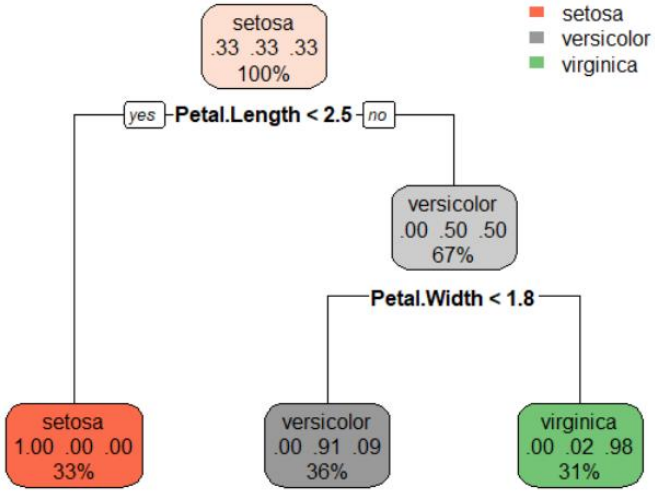
Se calcula el índice Gini para todas las combinaciones y se escoge el menor



Árboles de decisión en R

- En R, podemos crear árboles de decisión con la librería `rpart`

```
library("rpart")
library("rpart.plot")
tree <- rpart(Species ~ ., data = iris, method = "class")
rpart.plot(tree)
```



¡Gracias!

Contacto: Rafael Zambrano

rafazamb@gmail.com