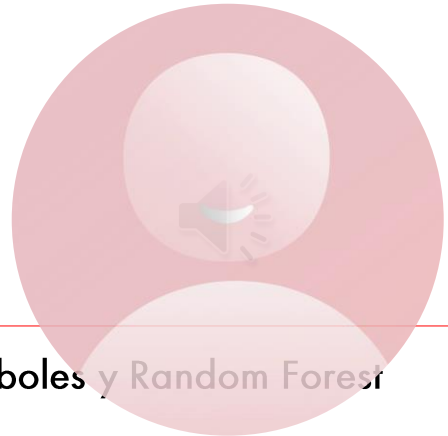
THE BRIDGE

Árboles de Decisión

Construcción (I): Criterios

Árboles de decisión: Construcción

- Queremos construir un árbol que nos ayude a clasificar los ingresos de los clientes de una entidad financiera

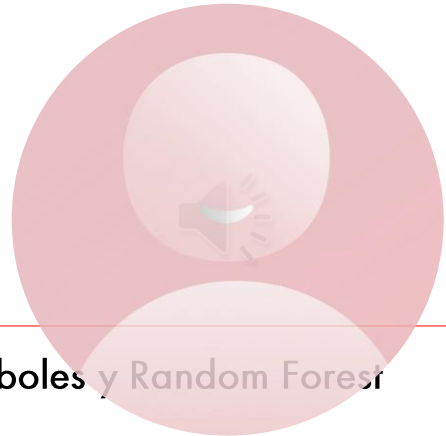


Árboles de decisión: Construcción

- Queremos construir un árbol que nos ayude a clasificar los ingresos de los clientes de una entidad financiera

- A partir de un dataset parecido al siguiente:

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

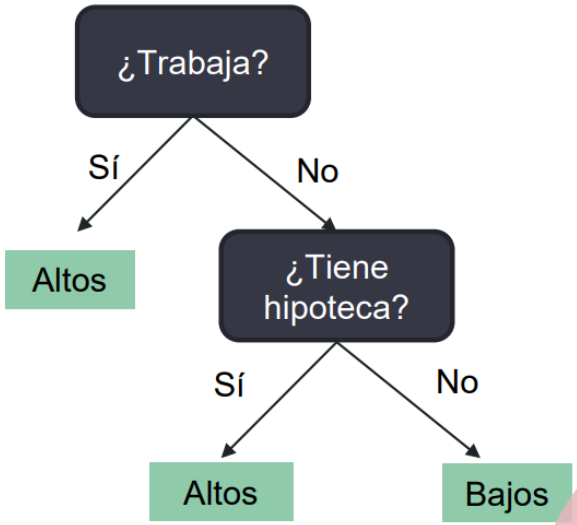


Árboles de decisión: Construcción

- Queremos construir un árbol que nos ayude a clasificar los ingresos de los clientes de una entidad financiera

- A partir de un dataset parecido al siguiente:

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

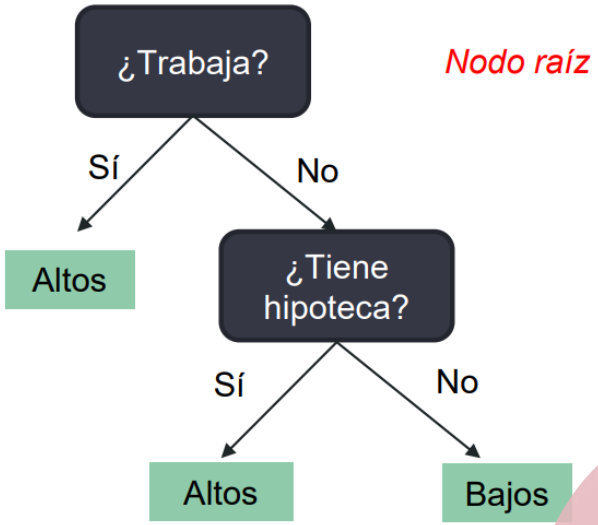


Árboles de decisión: Construcción

- Queremos construir un árbol que nos ayude a clasificar los ingresos de los clientes de una entidad financiera

- A partir de un dataset parecido al siguiente:

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

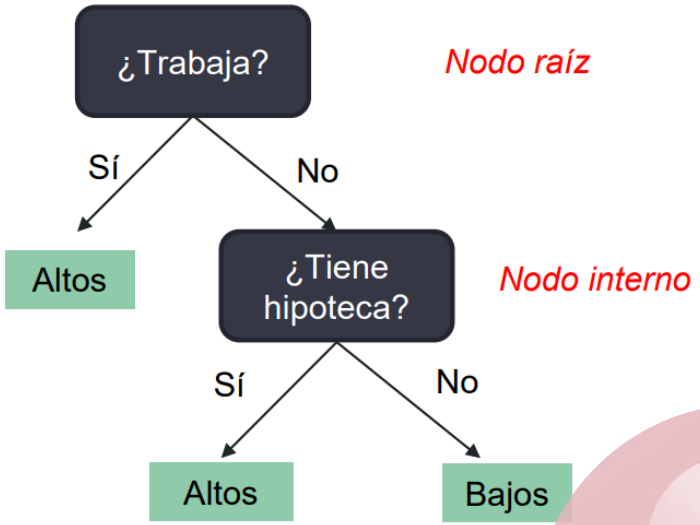


Árboles de decisión: Construcción

- Queremos construir un árbol que nos ayude a clasificar los ingresos de los clientes de una entidad financiera

- A partir de un dataset parecido al siguiente:

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

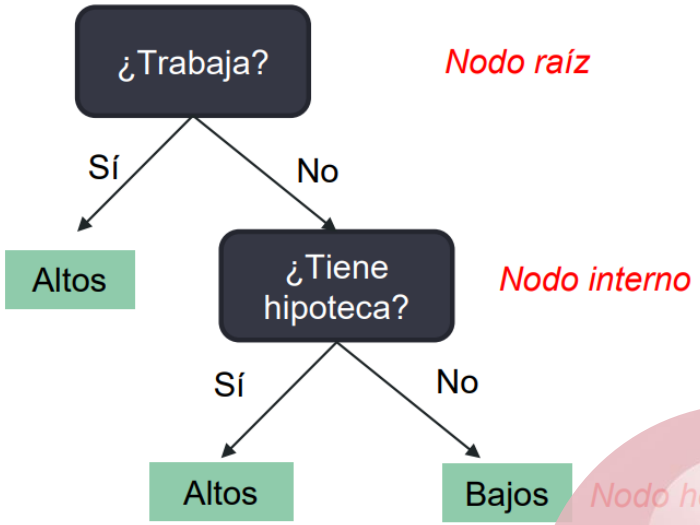


Árboles de decisión: Construcción

- Queremos construir un árbol que nos ayude a clasificar los ingresos de los clientes de una entidad financiera

- A partir de un dataset parecido al siguiente:

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

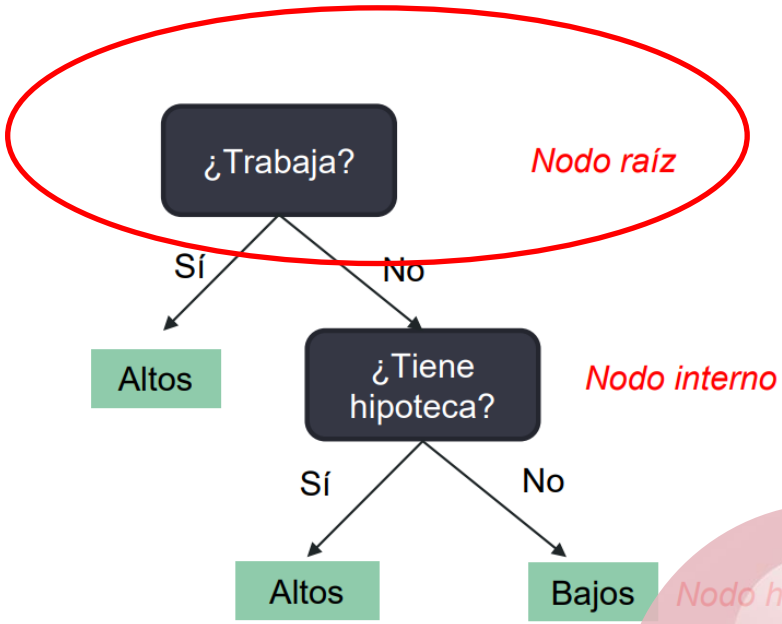


Árboles de decisión: Construcción

- Queremos construir un árbol que nos ayude a clasificar los ingresos de los clientes de una entidad financiera

- A partir de un dataset parecido al siguiente:

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

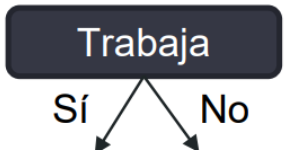


¿Cómo se decide que variable se usa en cada nodo?
Primero para el nodo raíz

Árboles de decisión: Selección Variables

- Medimos cómo de bien separa cada variable o feature a la variable candidata (Ingresos)

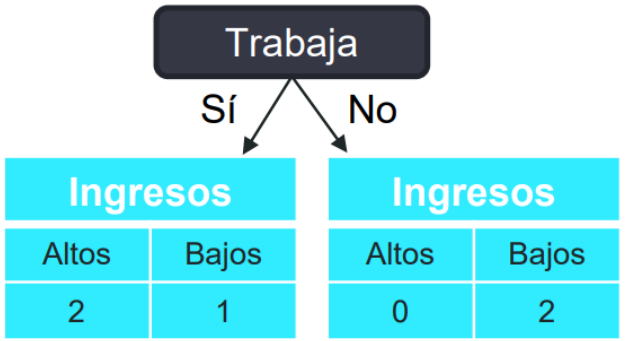
Cliente	Edad	Trabaja	Hipoteca	Ingresos
A	32	SÍ	SÍ	Altos
B	25	SÍ	SÍ	Altos
C	48	NO	NO	Bajos
D	67	NO	SÍ	Bajos
E	18	SÍ	NO	Bajos



Árboles de decisión: Selección Variables

- Medimos cómo de bien separa cada variable o feature a la variable candidata (Ingresos)

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



Árboles de decisión: Selección Variables

- Medimos cómo de bien separa cada variable o feature a la variable candidata (Ingresos)

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

Árboles de decisión: Selección Variables

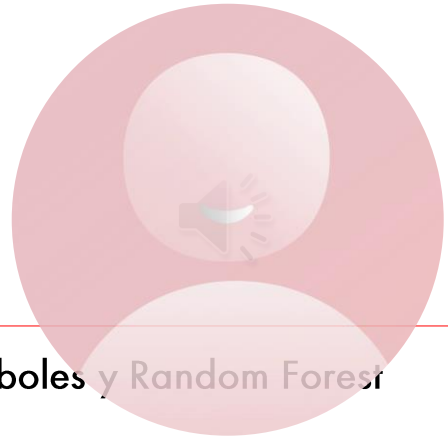
- Medimos cómo de bien separa cada variable o feature a la variable candidata (Ingresos)

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

En este caso, podríamos empezar por cualquier de ellas porque separan igual de bien o mal

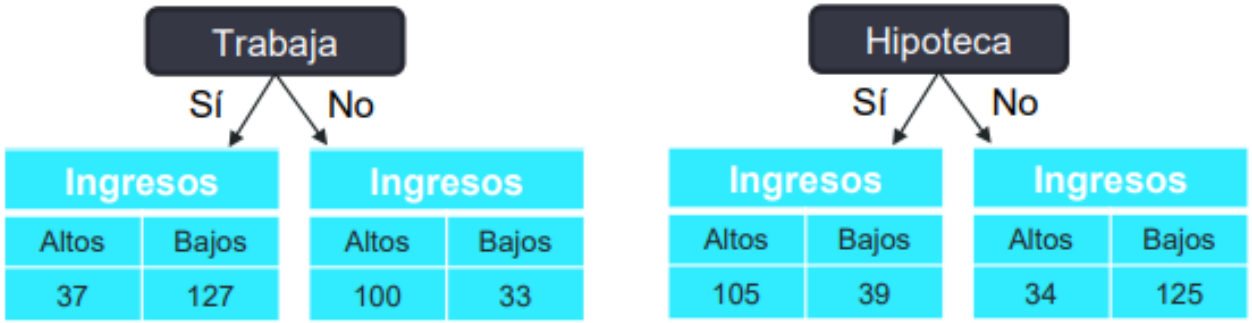
Árboles de decisión: Selección Variables

- Imaginemos que nuestro dataset es un poco más completo y tenemos una situación como la siguiente



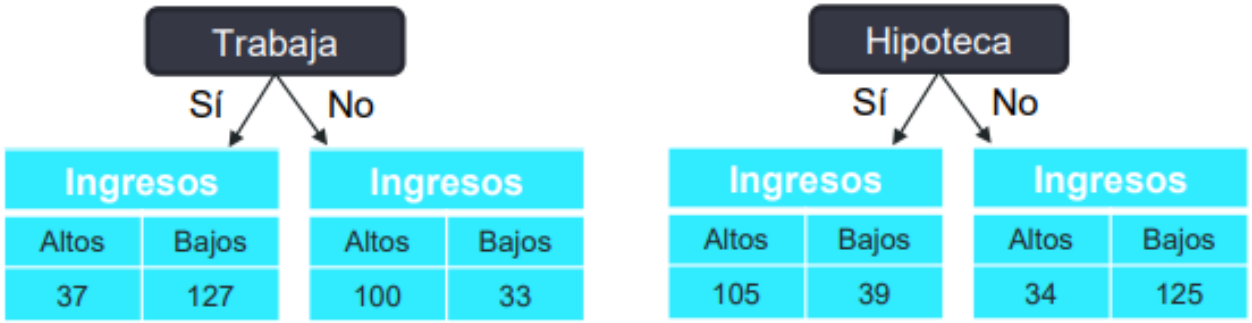
Árboles de decisión: Selección Variables

- Imaginemos que nuestro dataset es un poco más completo y con una situación como la siguiente



Árboles de decisión: Selección Variables

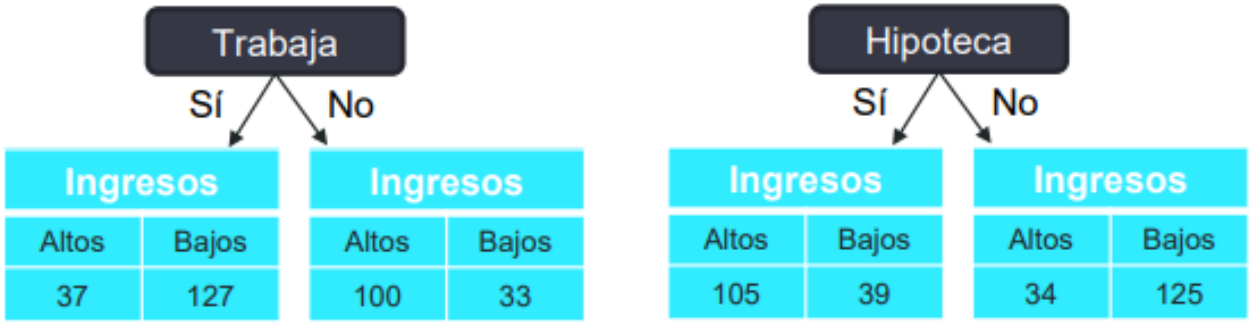
- Imaginemos que nuestro dataset es un poco más completo y con una situación como la siguiente



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

Árboles de decisión: Selección Variables

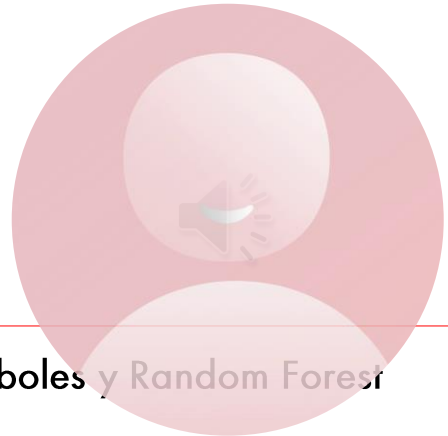
- Imaginemos que nuestro dataset es un poco más completo y con una situación como la siguiente



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

- La métrica más común para medir impurezas se conoce como índice "Gini"

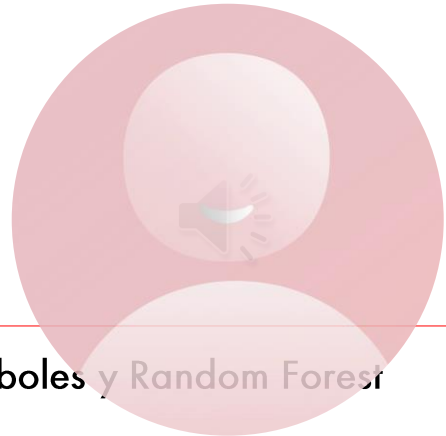
Árboles de decisión: Índice GINI



Árboles de decisión: Índice GINI

- Impureza de Gini, para cada nodo hoja:

$$1 - (\text{probabilidad de la clase 1})^2 - (\text{probabilidad de la clase 2})^2$$

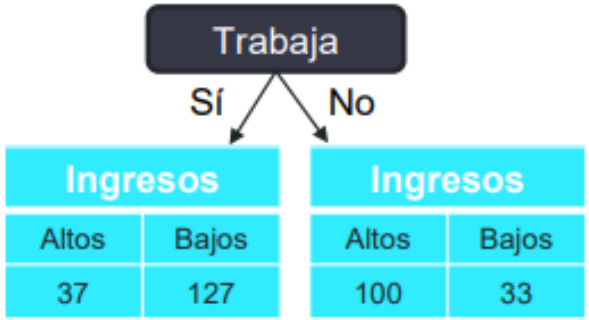


Árboles de decisión: Índice GINI

- Impureza de Gini, para cada nodo hoja:

$$1 - (\text{probabilidad de la clase 1})^2 - (\text{probabilidad de la clase 2})^2$$

- Cálculo sobre la variable "Trabaja":

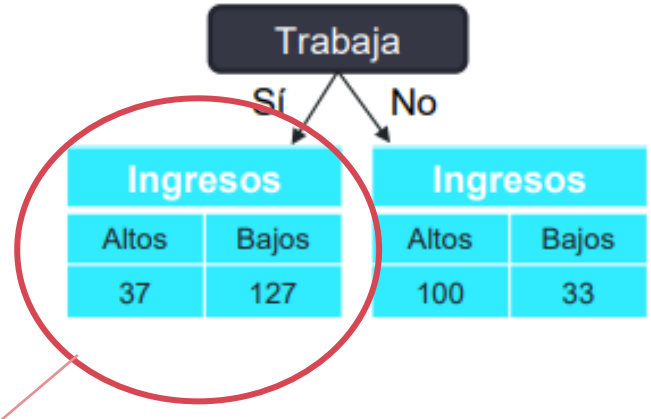


Árboles de decisión: Índice GINI

- Impureza de Gini, para cada nodo hoja:

$$1 - (\text{probabilidad de la clase 1})^2 - (\text{probabilidad de la clase 2})^2$$

- Cálculo sobre la variable "Trabaja":



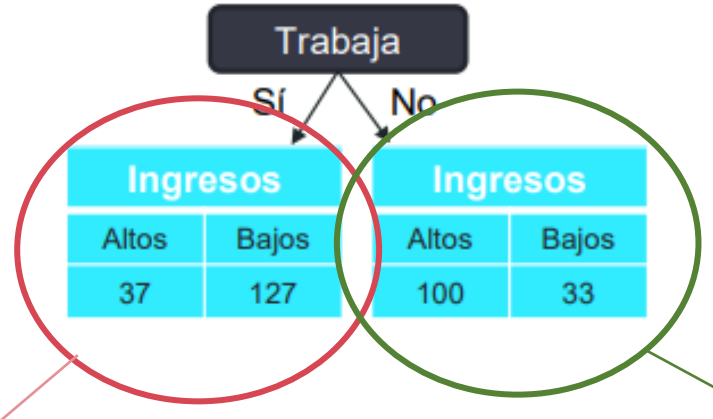
$$1 - \left(\frac{127}{127+37}\right)^2 - \left(\frac{37}{37+127}\right)^2 = 0.35$$

Árboles de decisión: Índice GINI

- Impureza de Gini, para cada nodo hoja:

$$1 - (\text{probabilidad de la clase 1})^2 - (\text{probabilidad de la clase 2})^2$$

- Cálculo sobre la variable "Trabaja":



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

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

Árboles de decisión: Índice GINI

- Impureza de Gini, para cada nodo hoja:

$$1 - (\text{probabilidad de la clase 1})^2 - (\text{probabilidad de la clase 2})^2$$

- Cálculo sobre la variable "Trabaja":

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

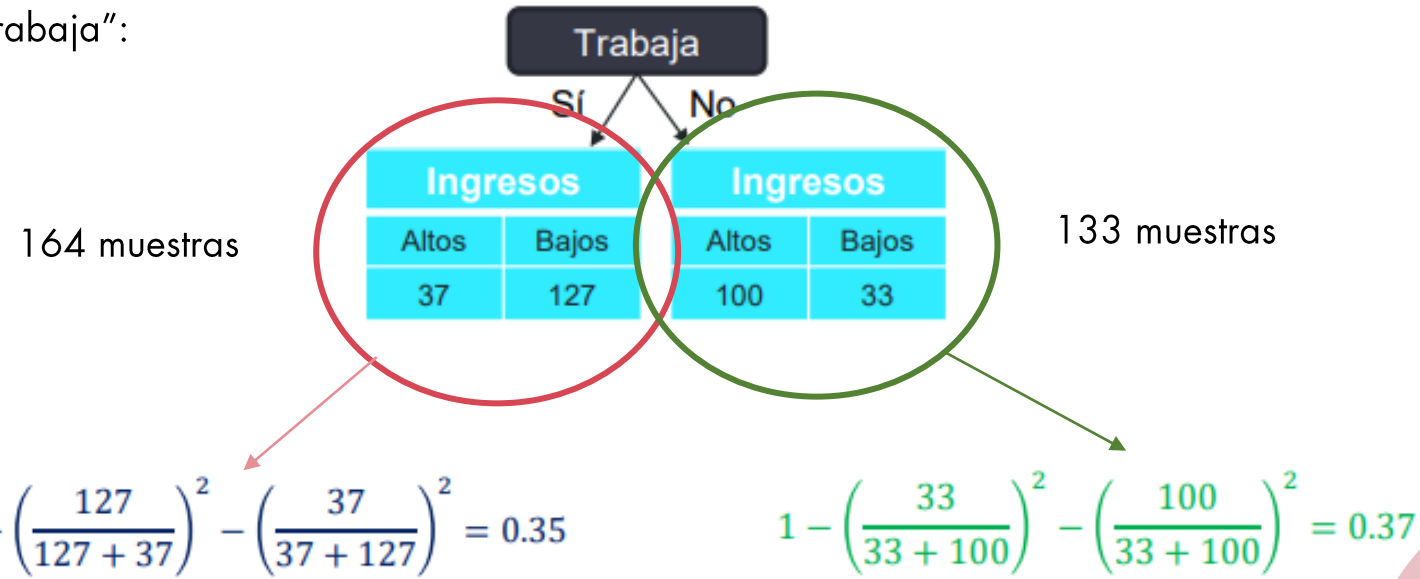
$$1 - \left(\frac{33}{33+100}\right)^2 - \left(\frac{100}{33+100}\right)^2 = 0.37$$

Árboles de decisión: Índice GINI

- Impureza de Gini, para cada nodo hoja:

$$1 - (\text{probabilidad de la clase 1})^2 - (\text{probabilidad de la clase 2})^2$$

- Cálculo sobre la variable "Trabaja":



- El índice Gini de la Variable será la media ponderada:

Árboles de decisión: Índice GINI

- Impureza de Gini, para cada nodo hoja:

$$1 - (\text{probabilidad de la clase 1})^2 - (\text{probabilidad de la clase 2})^2$$

- Cálculo sobre la variable "Trabaja":

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

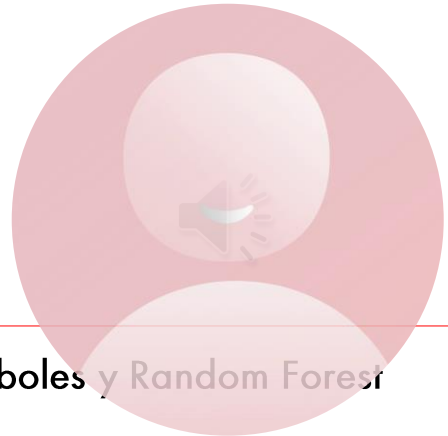
$$1 - \left(\frac{33}{33+100}\right)^2 - \left(\frac{100}{33+100}\right)^2 = 0.37$$

- El índice Gini de la Variable será la media ponderada:

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

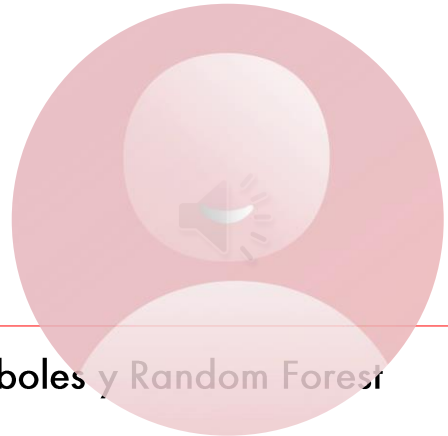
Árboles de decisión: Índice GINI

- Repitiendo el mismo procedimiento obtenemos el índice para la variable "Hipoteca": 0.364



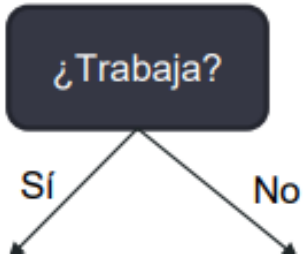
Árboles de decisión: Índice GINI

- Repitiendo el mismo procedimiento obtenemos el índice para la variable "Hipoteca": 0.364
- La impureza generada por utilizar la variable "Trabaja", 0.36, es menor, aunque sea ligeramente, que la de "Hipoteca"



Árboles de decisión: Índice GINI

- Repitiendo el mismo procedimiento obtenemos el índice para la variable "Hipoteca": 0.364
- La impureza generada por utilizar la variable "Trabaja", 0.36, es menor, aunque sea ligeramente, que la de "Hipoteca"
- Escogemos "Trabaja" como la variable del nodo raíz



- Este proceso se repite en los nodos intermedios con las variables distintas a la del nodo raíz

Árboles de decisión: Índice GINI

- Repitiendo el mismo procedimiento obtenemos el índice para la variable "Hipoteca": 0.364
- La impureza generada por utilizar la variable "Trabaja", 0.36, es menor, aunque sea ligeramente, que la de "Hipoteca"
- Escogemos "Trabaja" como la variable del 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

Árboles de decisión: Índice GINI

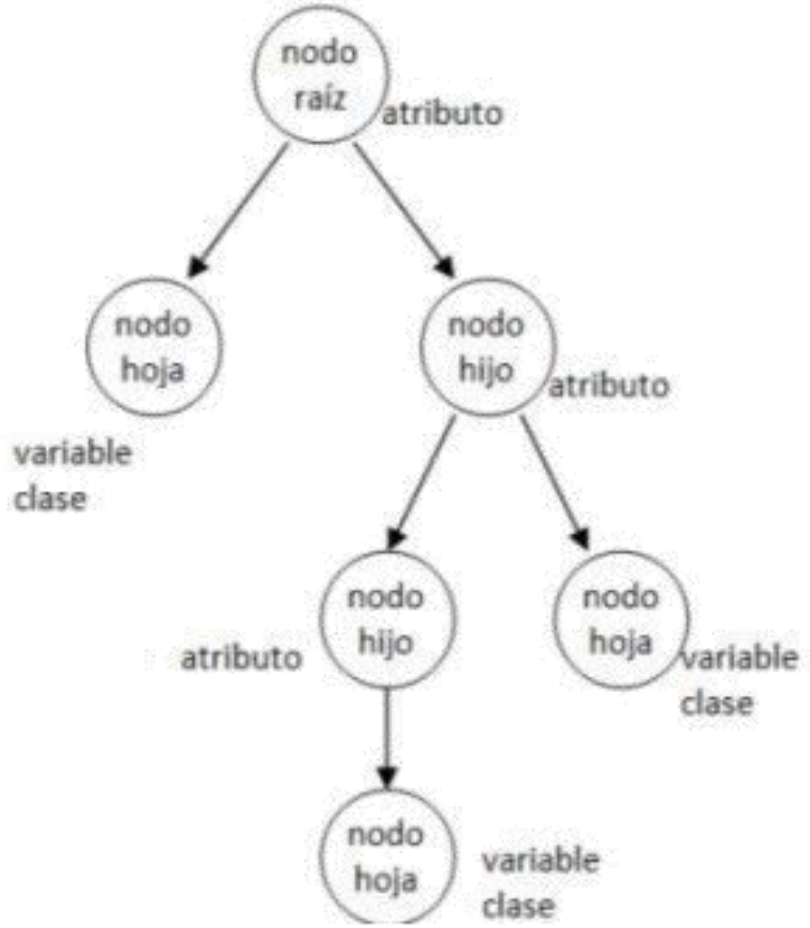
- Repitiendo el mismo procedimiento obtenemos el índice para la variable "Hipoteca": 0.364
- La impureza generada por utilizar la variable "Trabaja", 0.36, es menor, aunque sea ligeramente, que la de "Hipoteca"
- Escogemos "Trabaja" como la variable del nodo raíz
- Este proceso se repite en los nodos 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

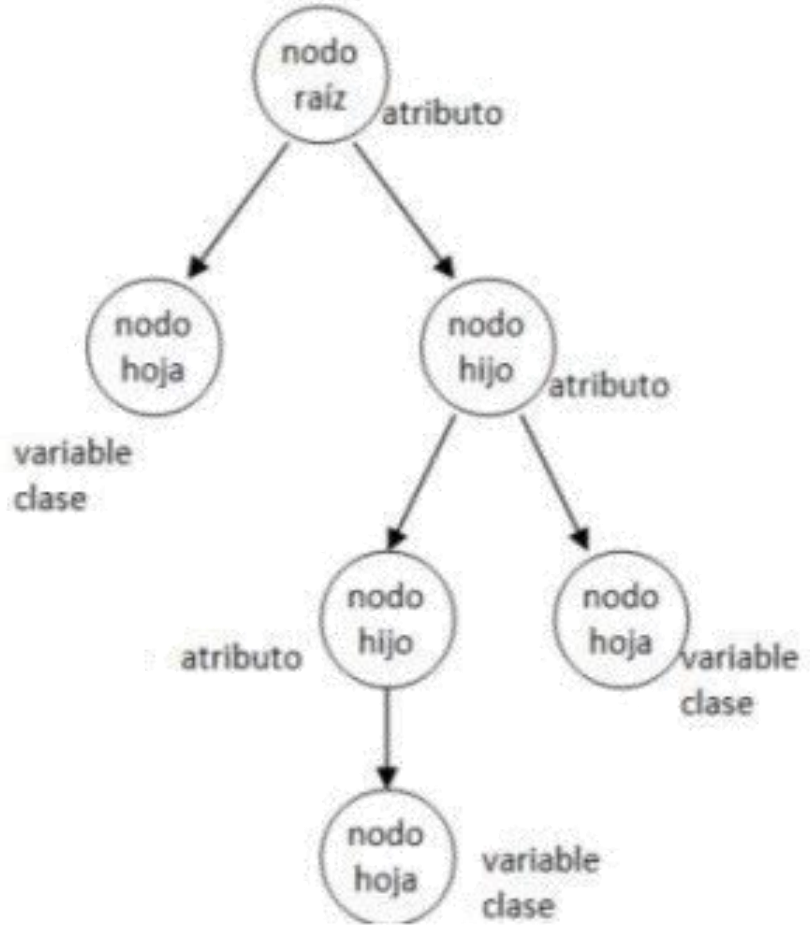
1. Calcular el índice de Gini para cada variable



Árboles de decisión: Pasos

1. Calcular el índice de Gini para cada variable

2. Si el nodo en sí tiene el menor Gini, se convierte en hoja

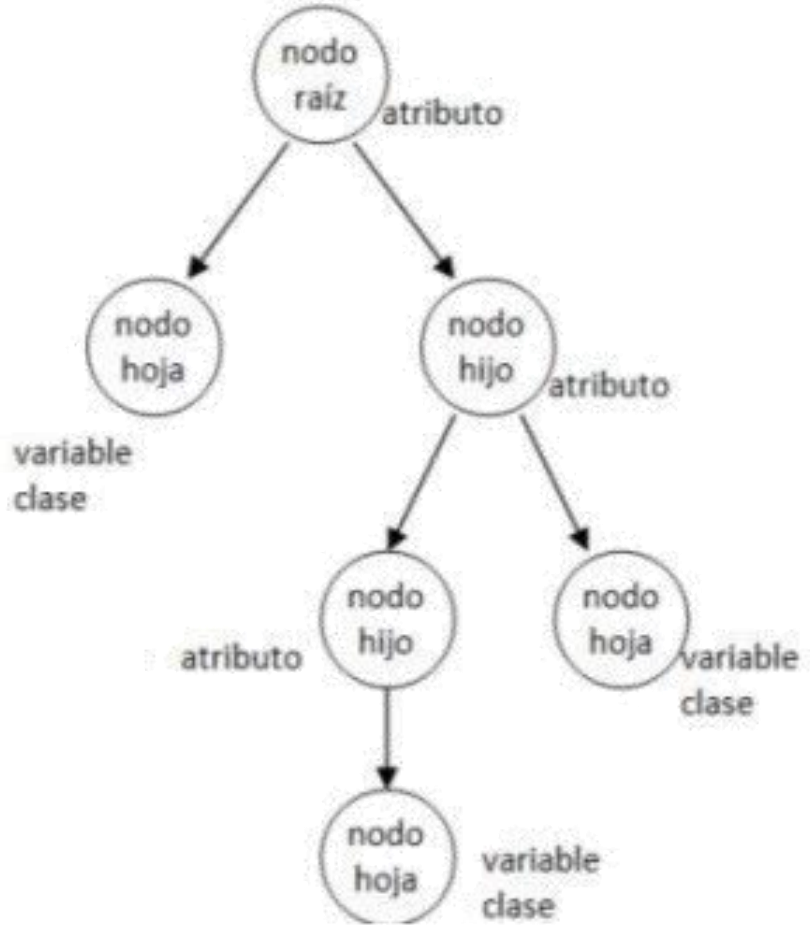


Árboles de decisión: Pasos

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 Gin

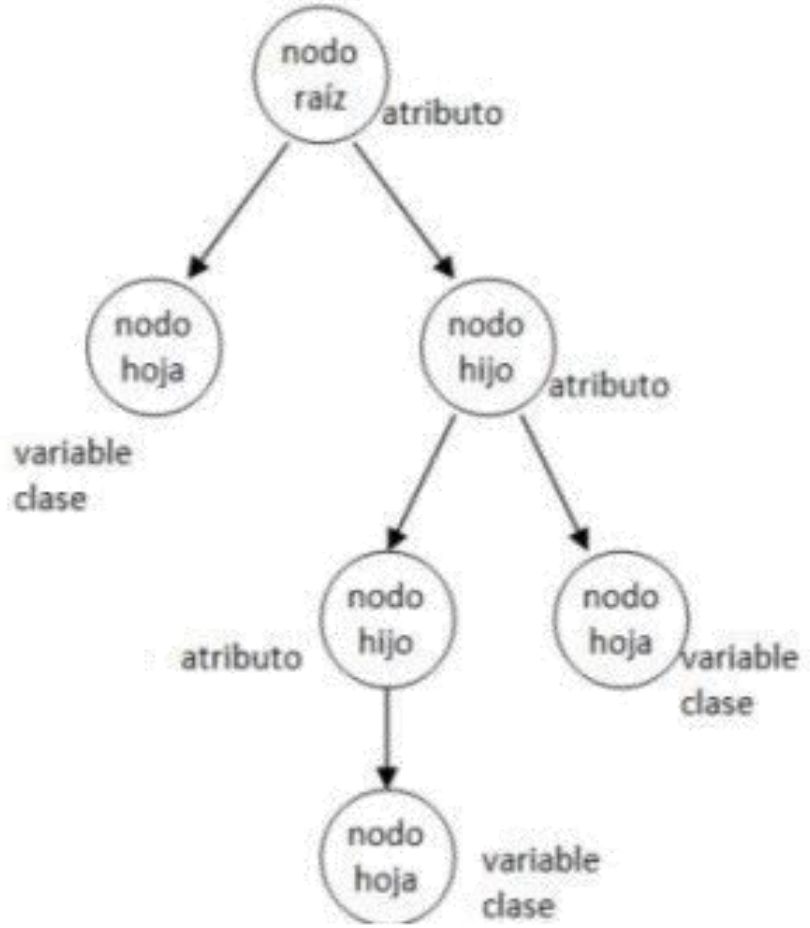


Árboles de decisión: Pasos

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 Gin




THE
BRIDGE