THE BRIDGE

Árboles de Decisión

Ejemplo Resumen

Árboles de decisión: Un "problema"

- Queremos construir un clasificador del nivel de ingresos de los jugadores de fútbol.

Árboles de decisión: Un "problema"

- Queremos construir un clasificador del nivel de ingresos de los jugadores de fútbol.
- "Entrenando" a partir de un dataset como el de la tabla

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Árboles de decisión: Un "problema"

- Queremos construir un clasificador del nivel de ingresos de los jugadores de fútbol.
- "Entrenando" a partir de un dataset como el de la tabla

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Vamos a repasar los pasos para obtener las variables y los nodos de un árbol de decisión que no permita tener ese clasificador

Árboles de decisión: Dosis de recuerdo

- Queremos construir un clasificador del nivel de ingresos de los jugadores de fútbol.
- "Entrenando" a partir de un dataset como el de la tabla

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

- Hay que medir cómo de bien separan las variables candidatos a la variable objetivo
- 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: Selección

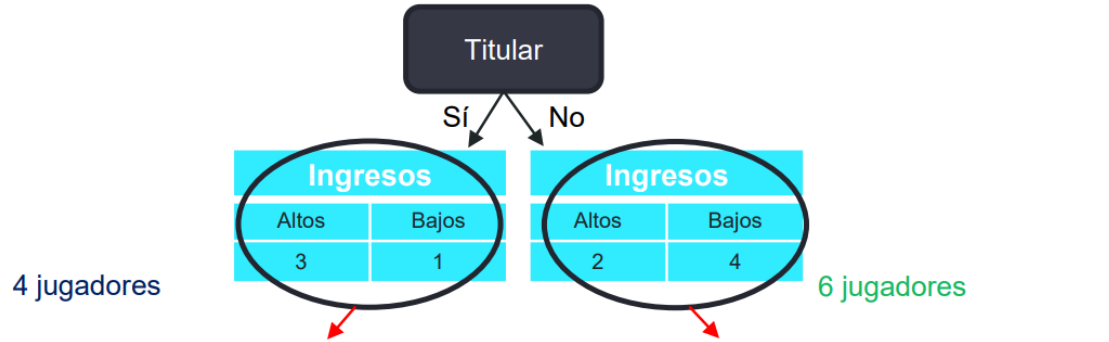
- Calculamos el índice Gini para las tres variables y escogemos la variable para nuestro nodo raíz

Árboles de decisión: Selección

- Calculamos el índice Gini para las tres variables y escogemos la variable para nuestro nodo raíz
- Escogemos una para empezar: “Titular”

Árboles de decisión: Selección

- Calculamos el índice Gini para las tres variables y escogemos la variable para nuestro nodo raíz
- Escogemos una para empezar: "Titular"



$$1-\left(\frac{3}{3+1}\right)^2-\left(\frac{1}{3+1}\right)^2=0.375$$

$$1-\left(\frac{2}{2+4}\right)^2-\left(\frac{4}{2+4}\right)^2=0.44$$

$$0.375\cdot\left(\frac{4}{10}\right)+0.44\cdot\left(\frac{6}{10}\right)=\mathbf{0.414}$$

Su índice Gini es: 0.414

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

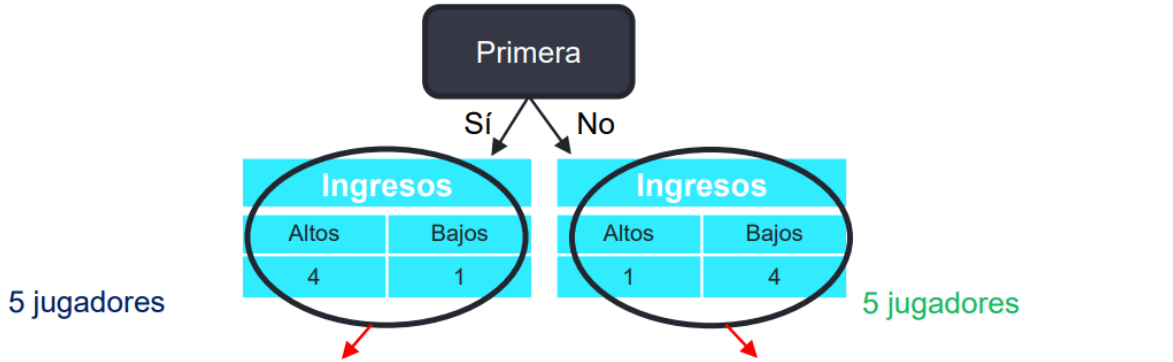
Árboles de decisión: Ginis

- Continuamos con “Primera”

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Árboles de decisión: Ginis

- Continuamos con "Primera"



$$1-\left(\frac{4}{4+1}\right)^2-\left(\frac{1}{4+1}\right)^2=0.32$$

$$1-\left(\frac{1}{1+4}\right)^2-\left(\frac{4}{1+4}\right)^2=0.32$$

$$0.32\cdot\left(\frac{5}{10}\right)+0.32\cdot\left(\frac{5}{10}\right)=\mathbf{0.32}$$

Su índice Gini es: 0.32

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Árboles de decisión: Ginis

- Terminamos con "Edad"
- Al ser una variable numérica no binaria, tendremos que seguir los pasos que ya conocemos

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

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Árboles de decisión: Ginis

- Terminamos con "Edad"
- Al ser una variable numérica no binaria, tendremos que seguir los pasos que ya conocemos

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

Edad	INGRESOS
18	BAJOS
19	ALTOS
19	BAJOS
20	ALTOS
20	BAJOS
24	BAJOS
28	ALTOS
29	ALTOS
30	ALTOS
31	BAJOS

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Árboles de decisión: Ginis

- Terminamos con "Edad"
- Al ser una variable numérica no binaria, tendremos que seguir los pasos que ya conocemos

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

Edad	INGRESOS
18	BAJOS
18.5	
19	ALTOS
19	
19	BAJOS
19.5	
20	ALTOS
20	
20	BAJOS
22	
24	BAJOS
26	
28	ALTOS
28.5	
29	ALTOS
29.5	
30	ALTOS
30.5	
31	BAJOS

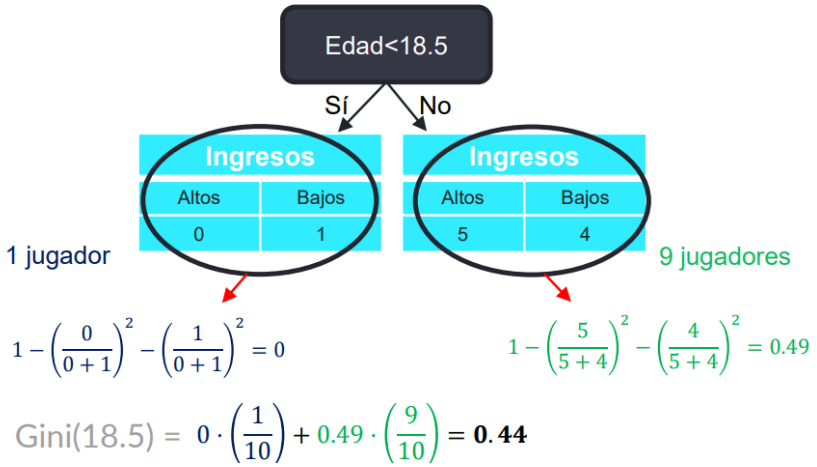
Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Árboles de decisión: Ginis

- Terminamos con "Edad"
- Al ser una variable numérica no binaria, tendremos que seguir los pasos que ya conocemos

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

Edad		INGRESOS
18	18.5	BAJOS
19	19	ALTOS
19	19.5	BAJOS
20	20	ALTOS
20	22	BAJOS
24	26	BAJOS
28	28.5	ALTOS
29	29.5	ALTOS
30	30.5	ALTOS
31		BAJOS



Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

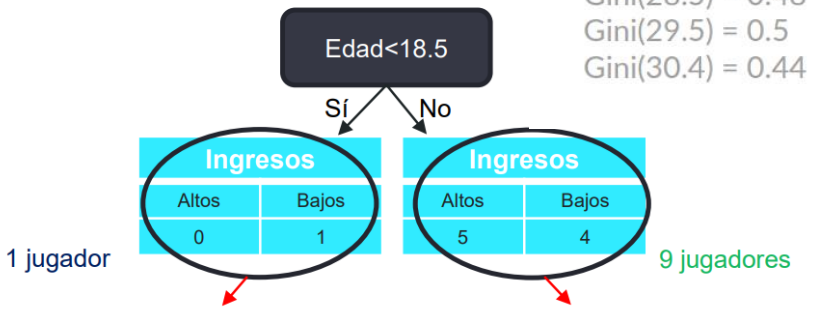
Árboles de decisión: Ginis

- Terminamos con "Edad"
- Al ser una variable numérica no binaria, tendremos que seguir los pasos que ya conocemos

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

Edad		INGRESOS
18	18.5	BAJOS
19	19	ALTOS
19	19.5	BAJOS
20	20	ALTOS
20	22	BAJOS
24	26	BAJOS
28	28.5	ALTOS
29	29.5	ALTOS
30	30.5	ALTOS
31		BAJOS

Gini(18.5) = 0.44
Gini(19) = 0.44
Gini(19.5) = 0.48
Gini(20) = 0.48
Gini(22) = 0.48
Gini(26) = 0.42
Gini(28.5) = 0.48
Gini(29.5) = 0.5
Gini(30.4) = 0.44



$$1-\left(\frac{0}{0+1}\right)^2-\left(\frac{1}{0+1}\right)^2=0$$

$$1-\left(\frac{5}{5+4}\right)^2-\left(\frac{4}{5+4}\right)^2=0.49$$

$$\text{Gini}(18.5) = 0\cdot\left(\frac{1}{10}\right)+0.49\cdot\left(\frac{9}{10}\right)=\mathbf{0.44}$$

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

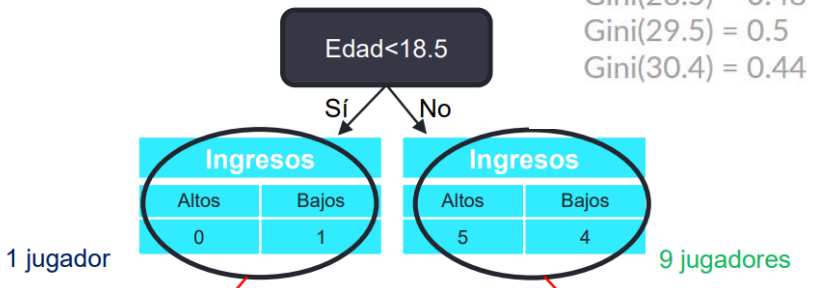
Árboles de decisión: Ginis

- Terminamos con "Edad"
- Al ser una variable numérica no binaria, tendremos que seguir los pasos que ya conocemos

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

Edad		INGRESOS
18	18.5	BAJOS
19	19	ALTOS
19	19.5	BAJOS
20	20	ALTOS
20	22	BAJOS
24	26	BAJOS
28	28.5	ALTOS
29	29.5	ALTOS
30	30.5	ALTOS
31		BAJOS

Gini(18.5) = 0.44
Gini(19) = 0.44
Gini(19.5) = 0.48
Gini(20) = 0.48
Gini(22) = 0.48
Gini(26) = 0.42
Gini(28.5) = 0.48
Gini(29.5) = 0.5
Gini(30.4) = 0.44



$$1-\left(\frac{0}{0+1}\right)^2-\left(\frac{1}{0+1}\right)^2=0$$

$$1-\left(\frac{5}{5+4}\right)^2-\left(\frac{4}{5+4}\right)^2=0.49$$

$$\text{Gini}(18.5) = 0\cdot\left(\frac{1}{10}\right)+0.49\cdot\left(\frac{9}{10}\right)=\mathbf{0.44}$$

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Árboles de decisión: Selección Final

- Impureza Gini de la variable " Edad ": 0.42
- Impureza Gini de la variable " Primera ": 0.32
- Impureza Gini de la variable " Titular ": 0.41

Árboles de decisión: Selección Final

- Impureza Gini de la variable " Edad ": 0.42
- Impureza Gini de la variable " Primera ": 0.32
- Impureza Gini de la variable " Titular ": 0.41

Árboles de decisión: Selección Final

- Impureza Gini de la variable " Edad ": 0.42
- Impureza Gini de la variable " Primera ": 0.32
- Impureza Gini de la variable " Titular ": 0.41

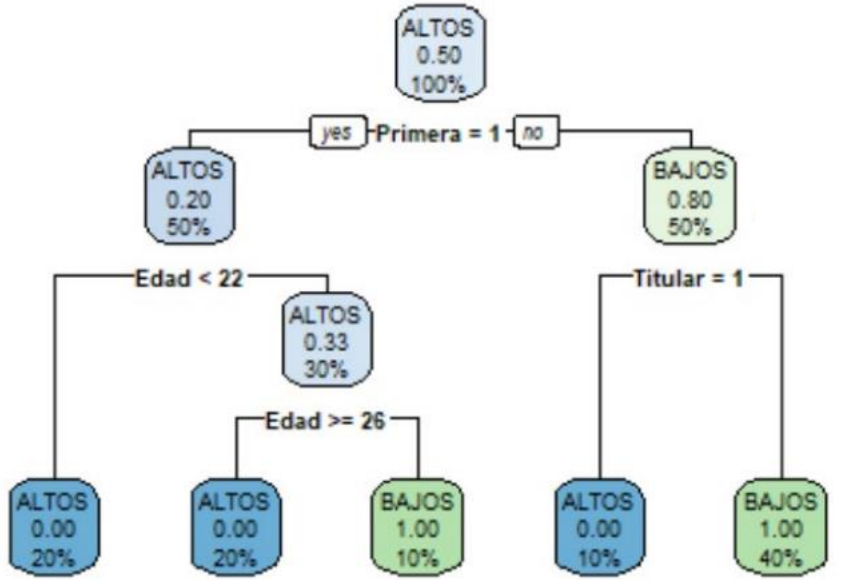
- Repetiremos el proceso con los nodos intermedios, pero ya solo con las muestras que correspondan a cada nodo

Árboles de decisión: Selección final

- Impureza Gini de la variable " Edad ": 0.42
- Impureza Gini de la variable " Primera ": 0.32
- Impureza Gini de la variable " Titular ": 0.41

- Repetiremos el proceso con los nodos intermedios, pero ya solo con las muestras que correspondan a cada nodo

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

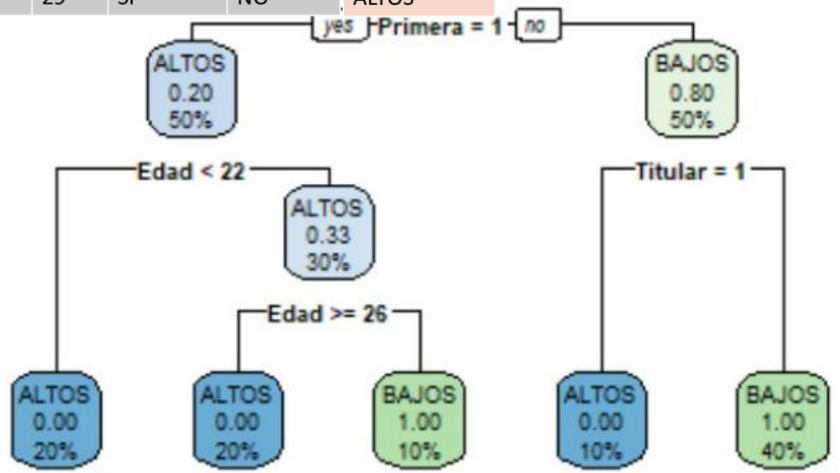


Árboles de decisión: Selección final

- Impureza Gini de la variable " Edad ": 0.42
- Impureza Gini de la variable " Primera ": 0.32
- Impureza Gini de la variable " Titular ": 0.41

- Repetiremos el proceso con los nodos intermedios, pero ya solo con las muestras que correspondan a cada nodo

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
8	29	SI	NO	ALTOS

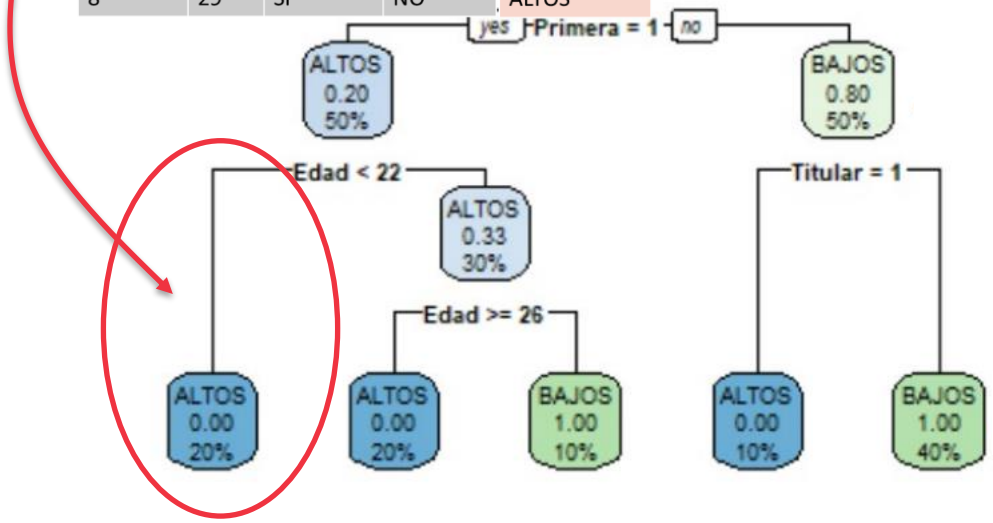


Árboles de decisión: Selección final

- Impureza Gini de la variable " Edad ": 0.42
- Impureza Gini de la variable " Primera ": 0.32
- Impureza Gini de la variable " Titular ": 0.41

- Repetiremos el proceso con los nodos intermedios, pero ya solo con las muestras que correspondan a cada nodo

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
8	29	SI	NO	ALTOS

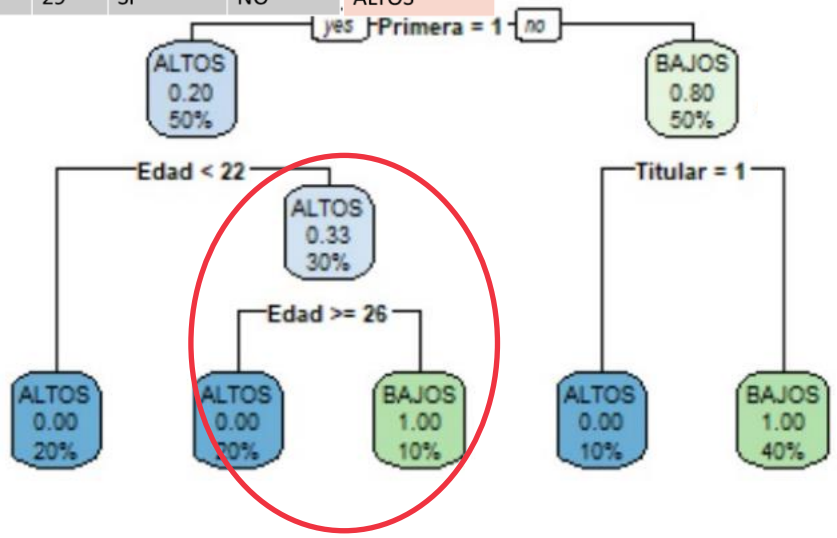


Árboles de decisión: Selección final

- Impureza Gini de la variable " Edad ": 0.42
- Impureza Gini de la variable " Primera ": 0.32
- Impureza Gini de la variable " Titular ": 0.41

- Repetiremos el proceso con los nodos intermedios, pero ya solo con las muestras que correspondan a cada nodo

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
8	29	SI	NO	ALTOS

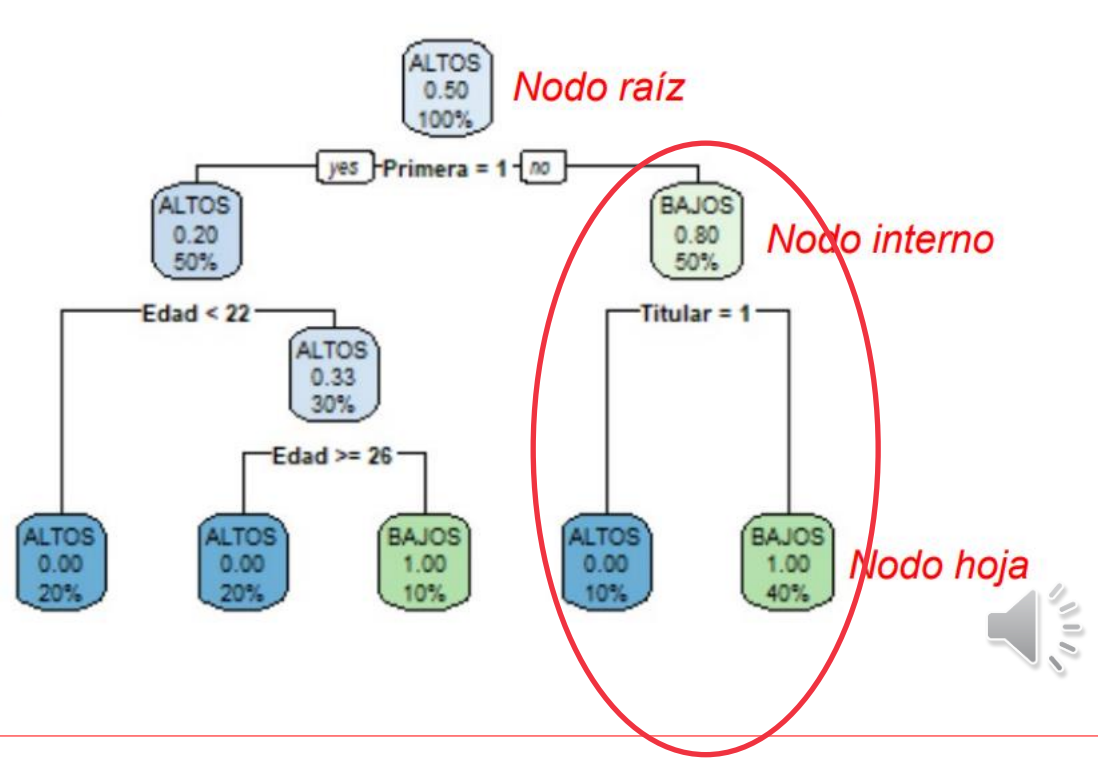


Árboles de decisión: Selección final

- Impureza Gini de la variable " Edad ": 0.42
- Impureza Gini de la variable " Primera ": 0.32
- Impureza Gini de la variable " Titular ": 0.41

- Repetiremos el proceso con los nodos intermedios, pero ya solo con las muestras que correspondan a cada nodo

- Un nodo se convierte en hoja cuando ninguna variable separa mejor el resultado de ese nodo




THE
BRIDGE