

# Árboles de Decisión

Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e  
Ingenierías - Universidad de Caldas

# Árboles de decisión

- Los arboles de decisión son representaciones gráficas de posibles soluciones a una decisión basadas en ciertas condiciones
- Es uno de los algoritmos de aprendizaje supervisado más utilizados en machine learning y pueden realizar tareas de clasificación o regresión (acrónimo del inglés CART)

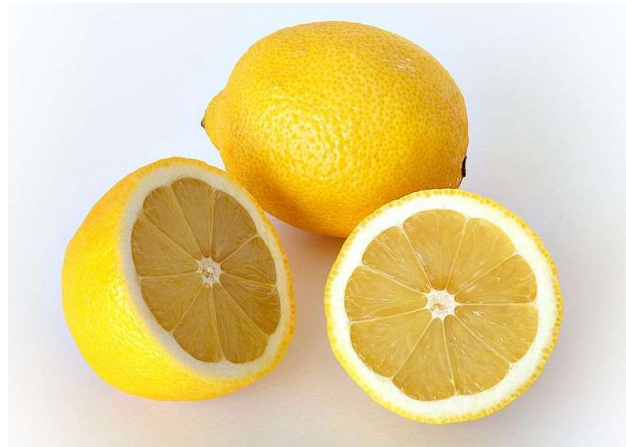
# Árboles de decisión

- Los árboles de decisión tienen un primer nodo llamado raíz (root) y luego se descomponen el resto de atributos de entrada en dos ramas (podrían ser más) planteando una condición que puede ser cierta o falsa.
- Se bifurca cada nodo en 2 y vuelven a subdividirse hasta llegar a las hojas que son los nodos finales y que equivalen a respuestas a la solución: Si/No, Comprar/Vender, o lo que sea que estemos clasificando.

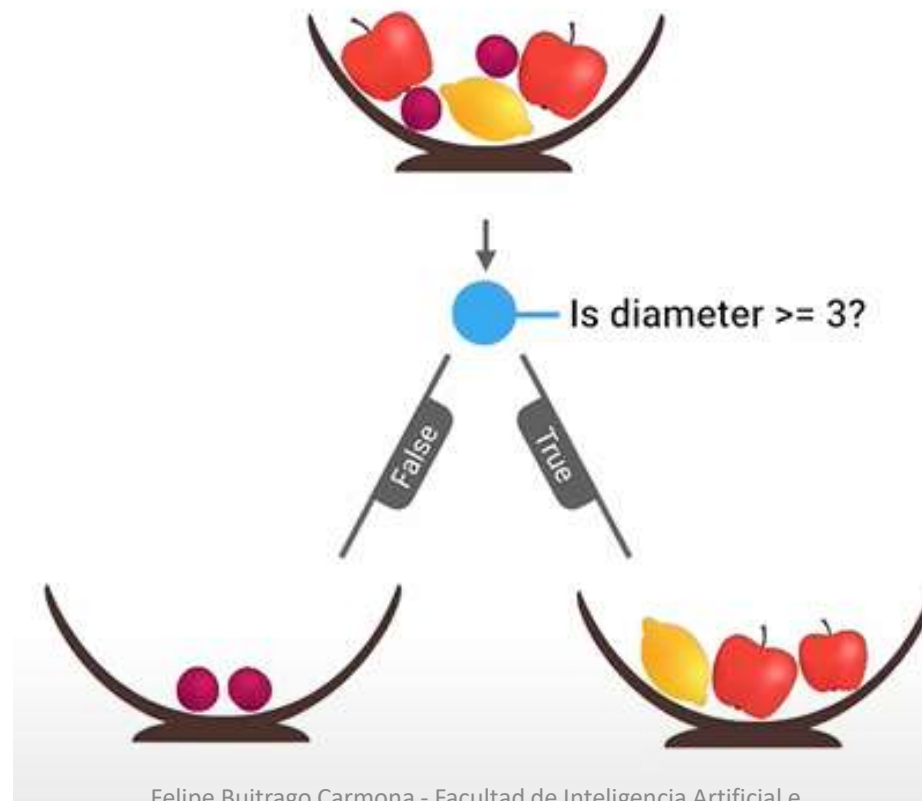
# Ejemplo 1

Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e  
Ingenierías - Universidad de Caldas

# Árboles de decisión

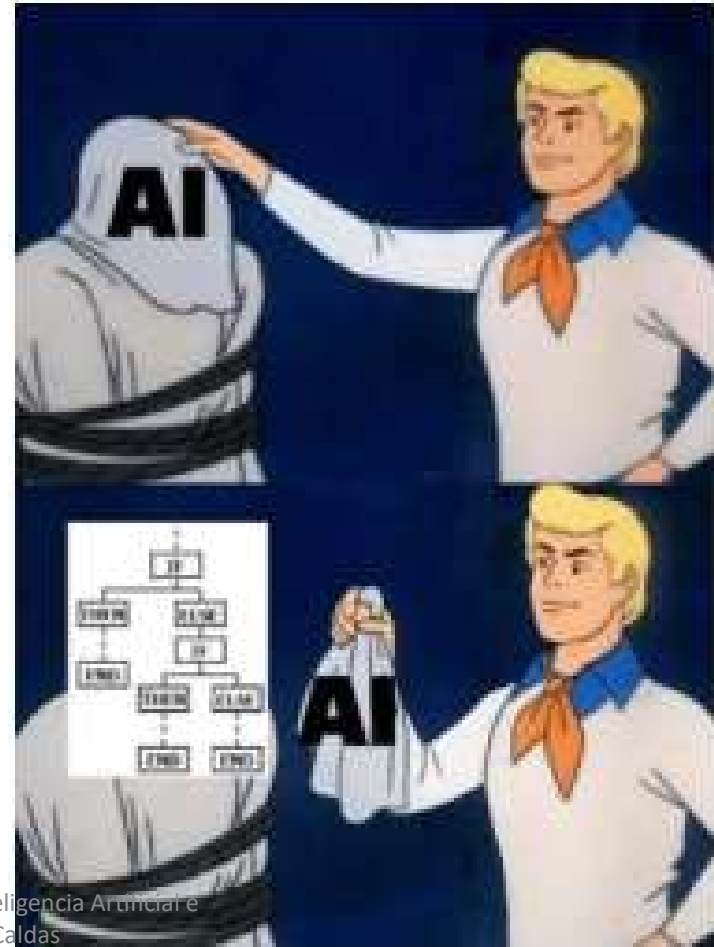
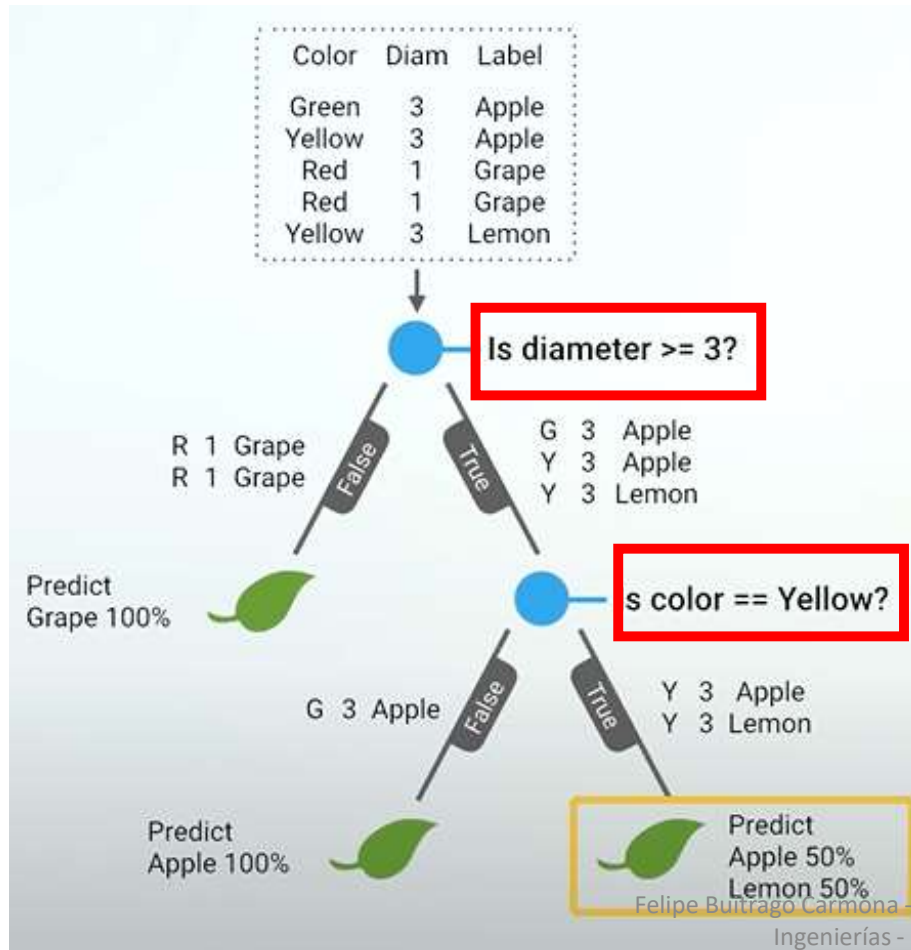


# Árboles de decisión



Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e Ingenierías - Universidad de Caldas

# Árboles de decisión

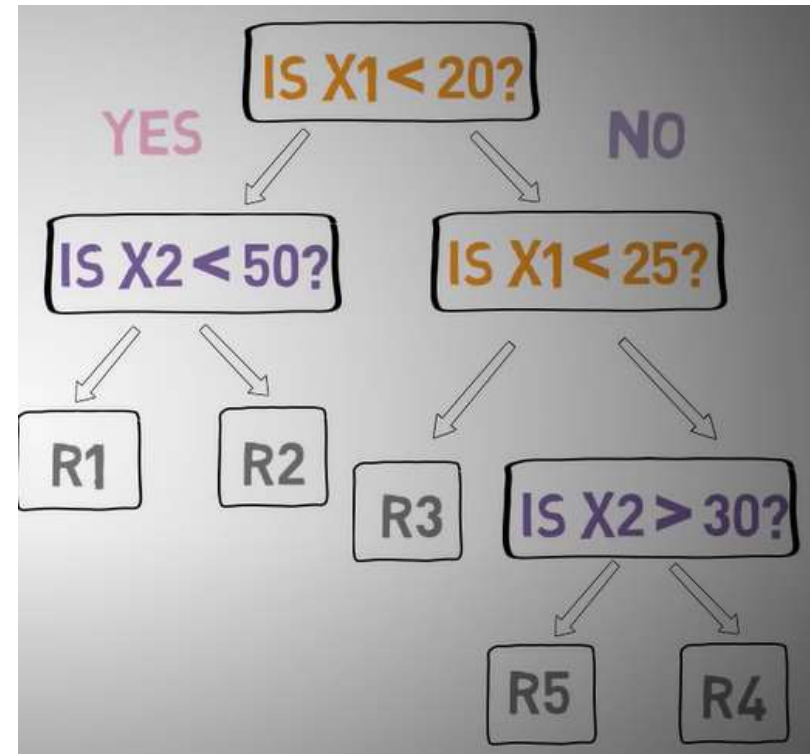
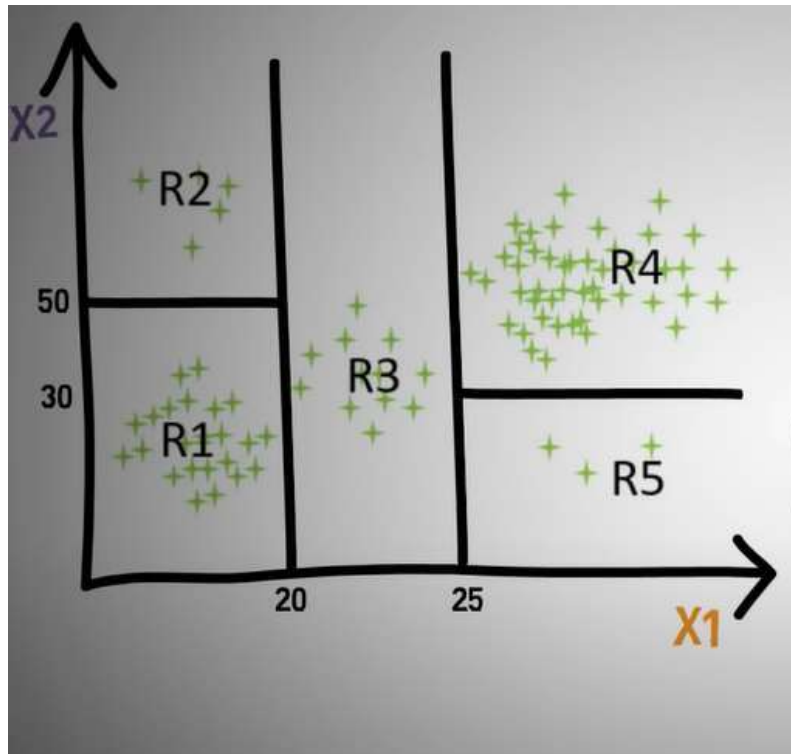


# Ejemplo 2

Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e  
Ingenierías - Universidad de Caldas



# Ejemplo

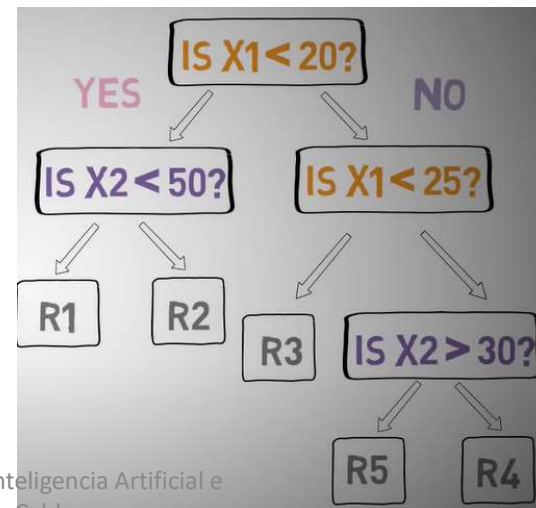
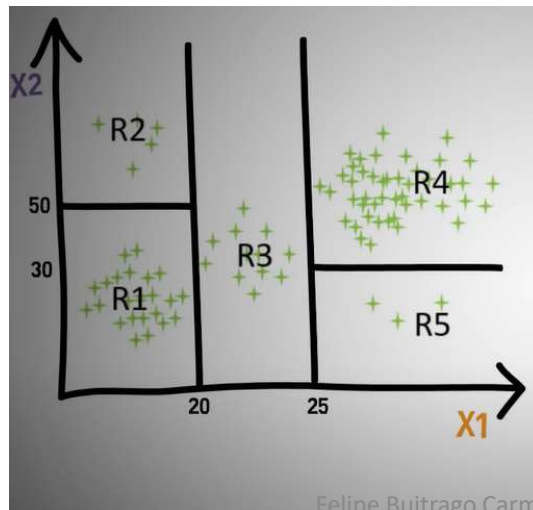


# ¿Cómo elegir las preguntas correctas?

Análisis de los atributos

# Árboles de decisión

- Para obtener el árbol óptimo y valorar cada subdivisión entre todos los árboles posibles y conseguir el nodo raíz y los subsiguientes, el algoritmo deberá medir de alguna manera las predicciones logradas y valorarlas para comparar de entre todas y obtener la mejor.



# Árboles de decisión

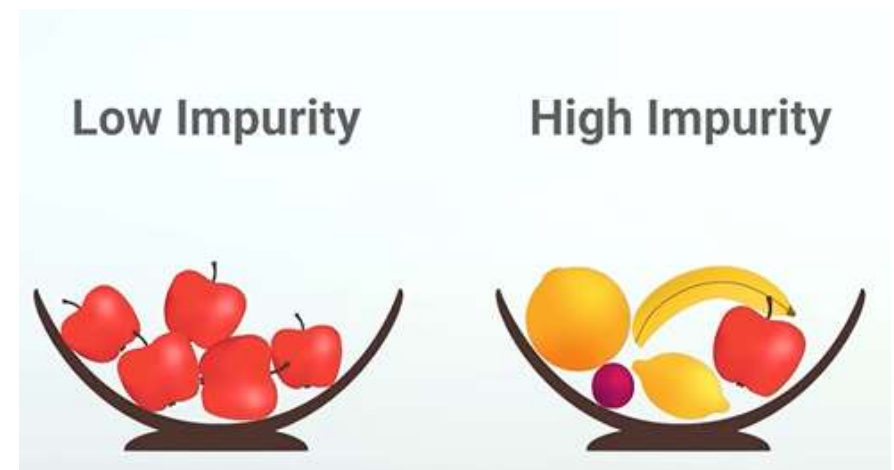
- Para medir y valorar, utiliza diversas funciones, siendo las más conocidas y usadas los
  - Índice gini
  - Ganancia de información,(utiliza la denominada “entropía”).
- La división de nodos continuará hasta que lleguemos a la profundidad máxima posible del árbol ó se limiten los nodos a una cantidad mínima de muestras en cada hoja

# Índice gini

Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e  
Ingenierías - Universidad de Caldas

# Indice Gini

- Se utiliza para atributos con valores continuos (precio de una casa).
- Esta función de coste mide el “grado de impureza” de los nodos, es decir, cuán desordenados o mezclados quedan los nodos una vez divididos.
- Deberemos minimizar el índice GINI



$$I_{Gini} = 1 - \sum_{i=1}^j p_i^2$$

# Ganancia de información

Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e  
Ingenierías - Universidad de Caldas

# Ganancia de información

- Se utiliza para atributos categóricos (cómo en hombre/mujer).
- Este criterio intenta estimar la información que aporta cada atributo basado en la “teoría de la información”.



# Ganancia de información

- Para medir la aleatoriedad de incertidumbre de un valor aleatorio de una variable “X” se define la Entropía.

$$Entropy = \frac{-p}{p+n} \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

- Calcular el promedio de la información

$$I(Attribute) = \sum \frac{p_i + n_i}{p+n} Entropy(A)$$

# Ganancia de información

- Al obtener la medida de entropía de cada atributo, podemos calcular la ganancia de información del árbol. Deberemos maximizar esa ganancia.

$$Gain = Entropy(S) - I(Attribute)$$

# Tipos de Árboles de Decisión

Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e  
Ingenierías - Universidad de Caldas

# Tipos

- ID3 → Manejo de discretos
- C4.5 → Manejo de ambos atributos continuos y discretos
- C5.0 → Recursos Velocidad es significativamente más rápido que el C4.5
- **CART → Puede ser implementado para problemas de regresión y clasificación**

# CART

Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e  
Ingenierías - Universidad de Caldas

# CART

- CART usa “Gini Impurity” en el proceso de división del conjunto de datos en un árbol de decisiones.
- Matemáticamente, podemos escribir la impureza de Gini de la siguiente manera

$$I_{Gini} = 1 - \sum_{i=1}^j p_i^2$$

$$I_{Gini} = 1 - (\text{the probability of target “No”})^2 - (\text{the probability of target “Yes”})^2$$

# Ejemplo

- Esta simulación utiliza un conjunto de datos de enfermedades cardíacas con 297 filas y 13 atributos. Este posee 138 pacientes que SI están enfermos y 165 que NO están enfermos (sanos).

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	Yes
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	Yes
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	Yes
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	Yes
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	Yes

# Pasos

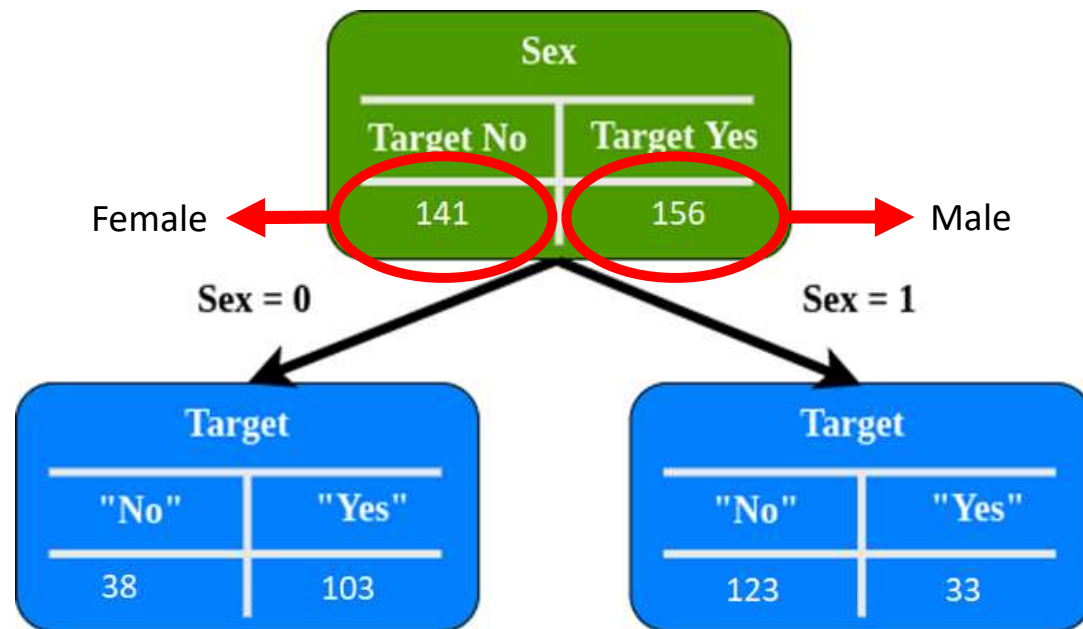
## Para la raíz:

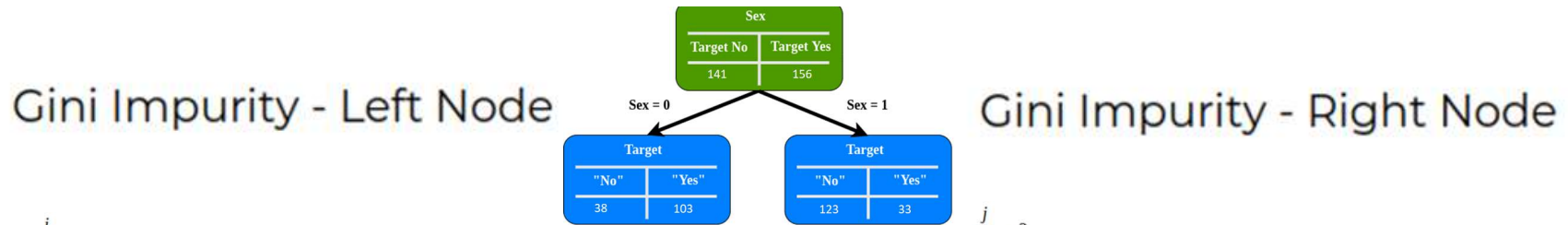
1. Calcule toda la puntuación de impurezas de Gini, para cada una de las características
2. Compare la puntuación de impurezas de Gini, y elija aquella con valor mas pequeño.



# Ejemplo

- Medir la impureza de Gini en el sexo





$$I_{Gini} = 1 - \sum_{i=1}^j p_i^2$$

$$I_{Gini} = 1 - (\text{the probability of target "Yes"})^2 - (\text{the probability of target "No"})^2$$

$$I_{Left} = 1 - \left(\frac{103}{103 + 38}\right)^2 - \left(\frac{38}{103 + 38}\right)^2 = 0,3937$$

$$I_{Gini} = 1 - \sum_{i=1}^j p_i^2$$

$$I_{Gini} = 1 - (\text{the probability of target "Yes"})^2 - (\text{the probability of target "No"})^2$$

$$I_{Right} = 1 - \left(\frac{33}{33 + 123}\right)^2 - \left(\frac{123}{33 + 123}\right)^2 = 0,3335$$

Total Gini Impurity - Leaf Node

## Gini Impurity - Left Node

$$I_{Gini} = 1 - \sum_{i=1}^j p_i^2$$

$$I_{Gini} = 1 - (\text{the probability of target "Yes"})^2 - (\text{the probability of target "No"})^2$$

$$I_{Left} = 1 - \left(\frac{103}{103 + 38}\right)^2 - \left(\frac{38}{103 + 38}\right)^2 = 0,3937$$

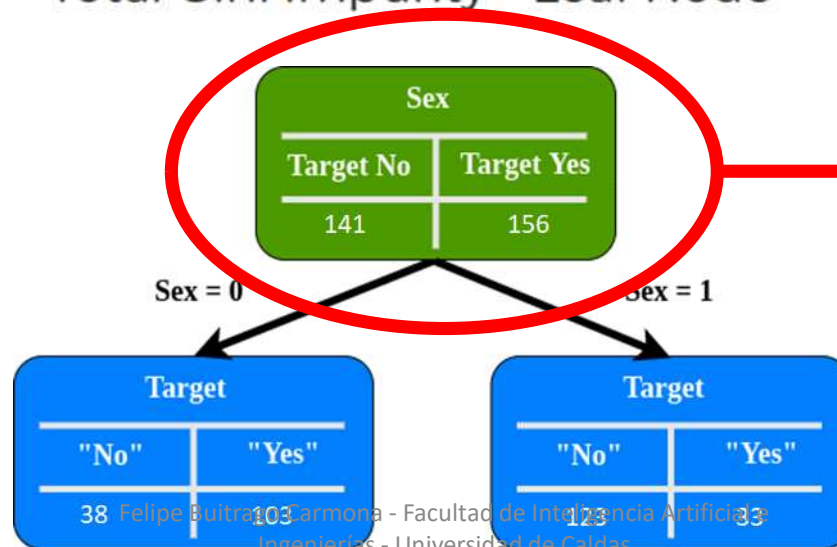
## Gini Impurity - Right Node

$$I_{Gini} = 1 - \sum_{i=1}^j p_i^2$$

$$I_{Gini} = 1 - (\text{the probability of target "Yes"})^2 - (\text{the probability of target "No"})^2$$

$$I_{Right} = 1 - \left(\frac{33}{33 + 123}\right)^2 - \left(\frac{123}{33 + 123}\right)^2 = 0,3335$$

## Total Gini Impurity - Leaf Node



Total impurce?

## Gini Impurity - Left Node

$$I_{Gini} = 1 - \sum_{i=1}^j p_i^2$$

$$I_{Gini} = 1 - (\text{the probability of target "Yes"})^2 - (\text{the probability of target "No"})^2$$

$$I_{Left} = 1 - \left(\frac{103}{103 + 38}\right)^2 - \left(\frac{38}{103 + 38}\right)^2 = 0,3937$$

## Gini Impurity - Right Node

$$I_{Gini} = 1 - \sum_{i=1}^j p_i^2$$

$$I_{Gini} = 1 - (\text{the probability of target "Yes"})^2 - (\text{the probability of target "No"})^2$$

$$I_{Right} = 1 - \left(\frac{33}{33 + 123}\right)^2 - \left(\frac{123}{33 + 123}\right)^2 = 0,3335$$

## Total Gini Impurity - Leaf Node

$I_{Sex}$  = weight average of the leaf node impurities

$$I_{Sex} = \left(\frac{\text{Total Sex 0}}{\text{Total Sex 0} + \text{Total Sex 1}}\right) I_{Left} + \left(\frac{\text{Total Sex 1}}{\text{Total Sex 0} + \text{Total Sex 1}}\right) I_{Right}$$

$$I_{Sex} = \left(\frac{141}{141 + 156}\right) I_{Left} + \left(\frac{156}{141 + 156}\right) I_{Right}$$

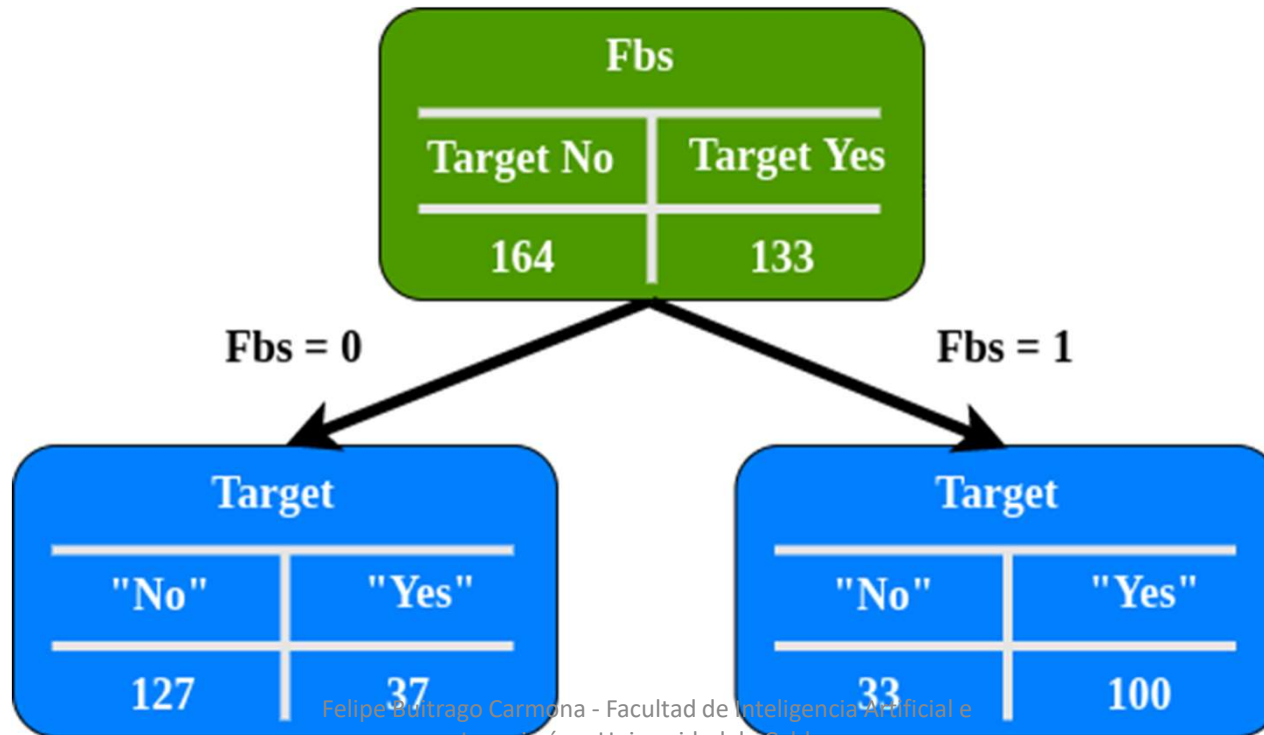
$$I_{Sex} = \frac{141}{141 + 156} \times 0,3937 + \frac{156}{141 + 156} \times 0,3335$$

$$I_{Sex} = 0,36208$$

Sex	
Target No	Target Yes
141	156

# Ejemplo

- Mida la impureza de Gini en Fbs (azúcar en sangre en ayunas)



# Ejemplo

Gini Impurity - Left Node

$$I_{Left} = 1 - \left(\frac{127}{127+37}\right)^2 - \left(\frac{37}{127+37}\right)^2 = 0.349$$

Gini Impurity - Right Node

$$I_{Right} = 1 - \left(\frac{100}{100+33}\right)^2 - \left(\frac{33}{100+33}\right)^2 = 0.373$$

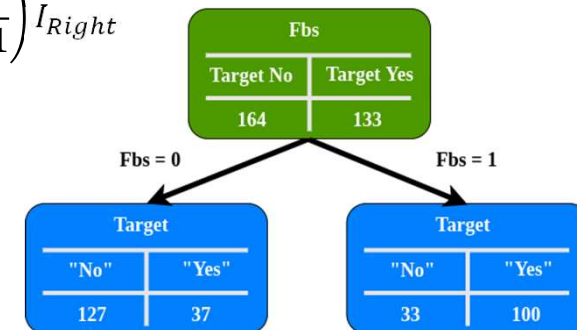
Total Gini Impurity - Leaf Node

$$I = \left(\frac{Total\ Fbs\ 0}{Total\ Fbs\ 0 + Total\ Fbs\ 1}\right) I_{Left} + \left(\frac{Total\ Fbs\ 1}{Total\ Fbs\ 0 + Total\ Fbs\ 1}\right) I_{Right}$$

$$I_{Fbs} = \left(\frac{164}{164+133}\right) I_{Left} + \left(\frac{133}{164+133}\right) I_{Right}$$

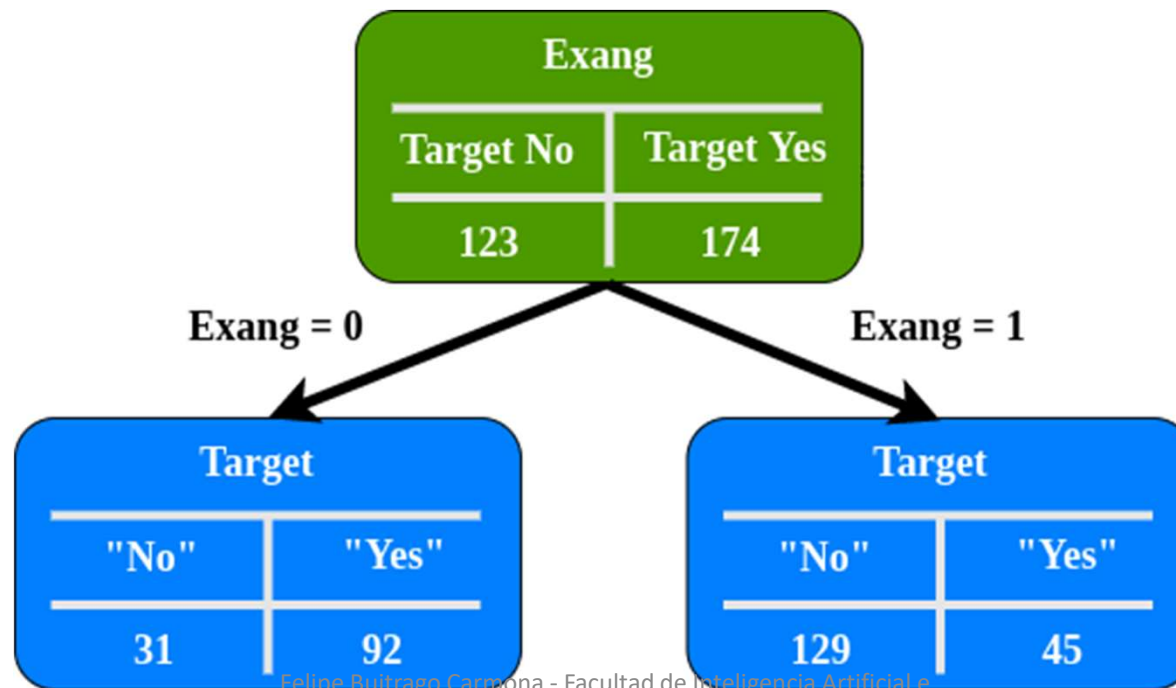
$$I_{Fbs} = \left(\frac{164}{164+133}\right) 0.349 + \left(\frac{133}{164+133}\right) 0.373$$

$$I_{Fbs} = 0.360$$



# Ejemplo

- Medir la impureza de Gini en Exang (angina inducida por el ejercicio)



# Ejemplo

Gini Impurity - Left Node

$$I_{Left} = 1 - \left(\frac{31}{31+92}\right)^2 - \left(\frac{92}{31+92}\right)^2 = 0.377$$

Gini Impurity - Right Node

$$I_{Right} = 1 - \left(\frac{129}{129+45}\right)^2 - \left(\frac{45}{129+45}\right)^2 = 0.383$$

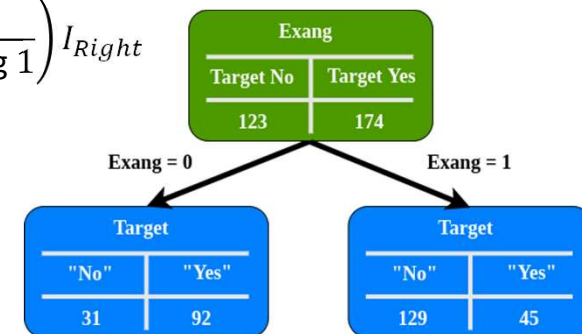
Total Gini Impurity - Leaf Node

$$I = \left(\frac{Total\ Exang\ 0}{Total\ Exang\ 0 + Total\ Exang\ 1}\right) I_{Left} + \left(\frac{Total\ Exang\ 1}{Total\ Exang\ 0 + Total\ Exang\ 1}\right) I_{Right}$$

$$I_{Exang} = \left(\frac{123}{123+174}\right) I_{Left} + \left(\frac{174}{123+174}\right) I_{Right}$$

$$I_{Exang} = \left(\frac{123}{123+174}\right) 0.377 + \left(\frac{174}{123+174}\right) 0.383$$

$$I_{Exang} = 0.381$$





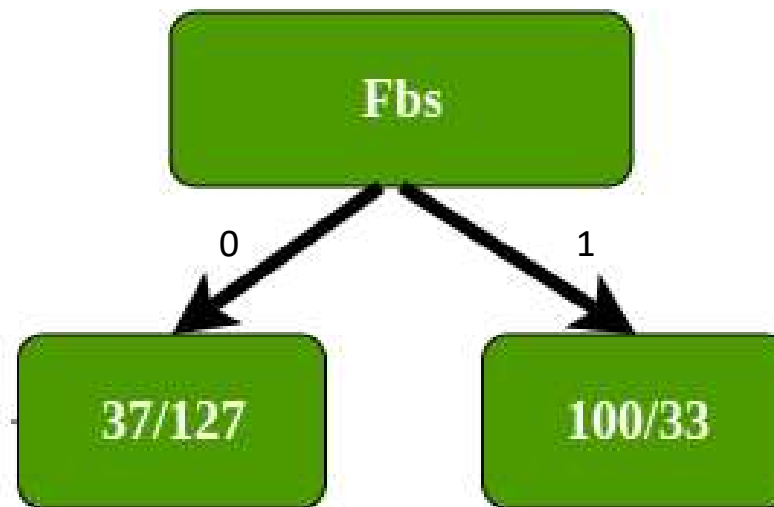
# Ejemplo

- Fbs (azúcar en sangre en ayunas) tiene la impureza de Gini más baja, así que utilícelo en el nodo raíz

Gini Impurity	
Sexo	0.362
Fbs (azúcar en sangre en ayunas)	0.360
Exang (angina inducida por el ejercicio)	0.381

# Ejemplo

- Por tanto elegimos a Fbs como nodo raíz
- Repitamos el proceso anterior por cada una de las ramas teniendo en cuenta al nodo raíz



# Pasos

## Para nodos que no son raíz

1. Calcule la impureza del nodo **sin utilizar ninguna** característica.
2. Calcule toda la puntuación de impurezas de Gini para cada una de las características

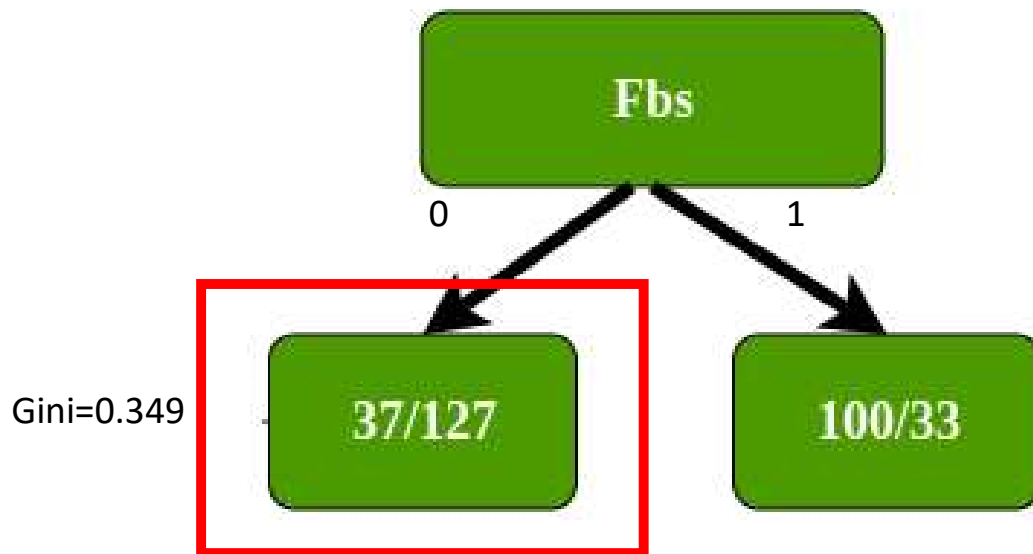
## Nota:

1. Si el nodo en sí tiene la puntuación más baja, entonces no tiene sentido separar los datos
2. Si la separación de los datos da como resultado una mejora, elija la separación con la puntuación de impureza más baja

# ¿Cómo calcular la impureza de Gini en datos categóricos?

# Ejemplo

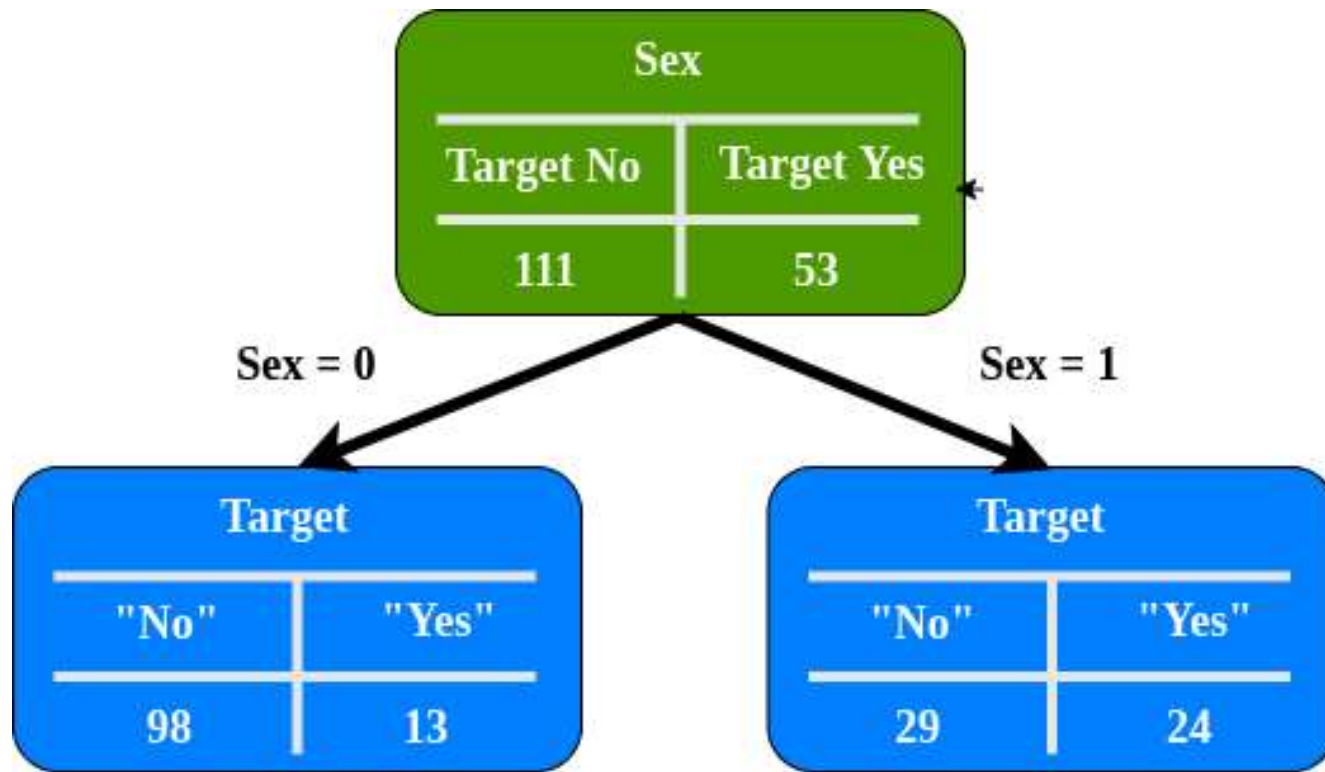
- Calcule la impureza del nodo **sin utilizar ninguna** característica



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

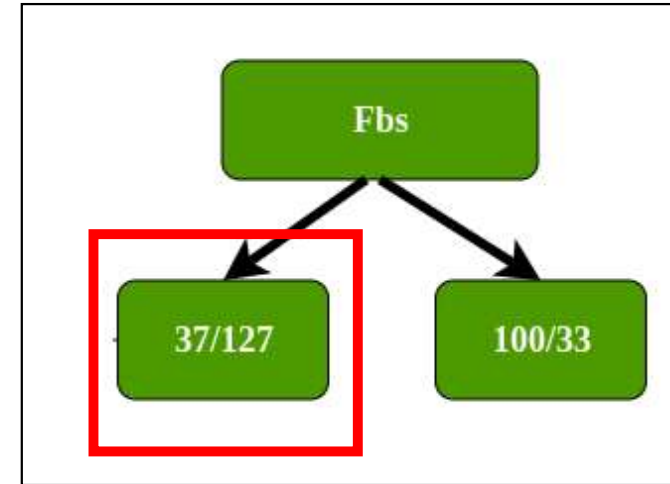
$$I = 0.349$$

# Ejemplo

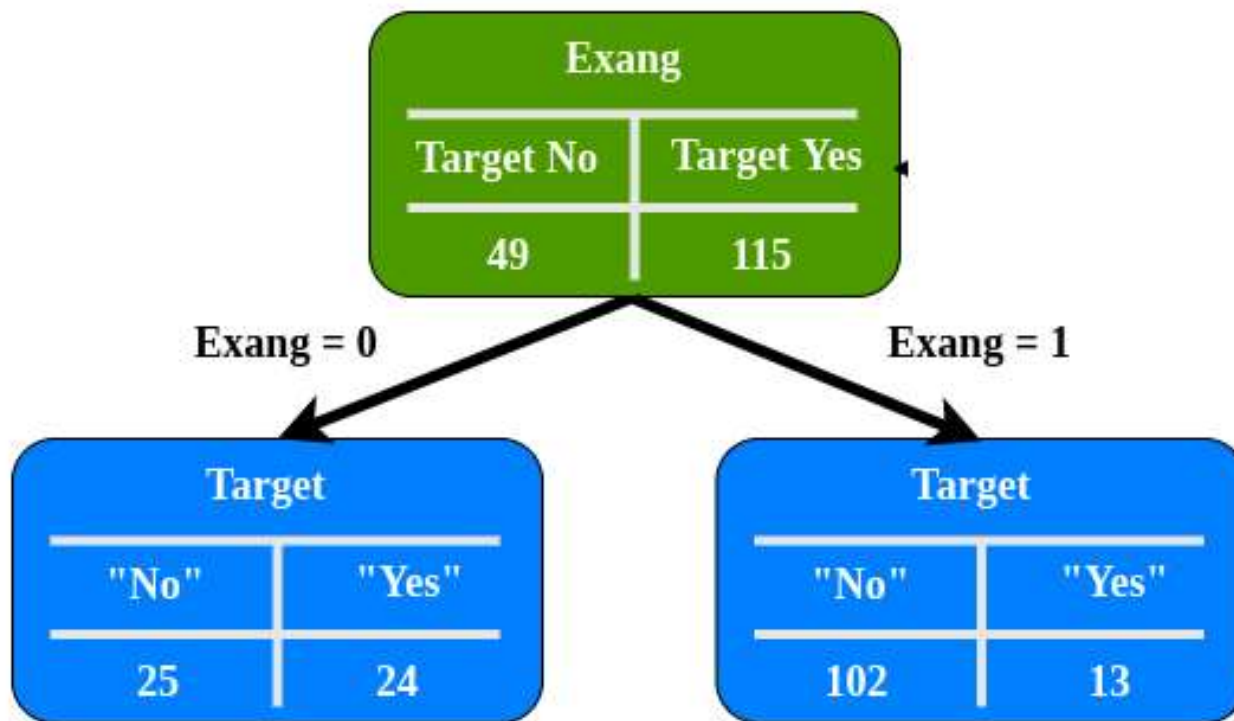


**Gini Impurity - Sex = 0.300**

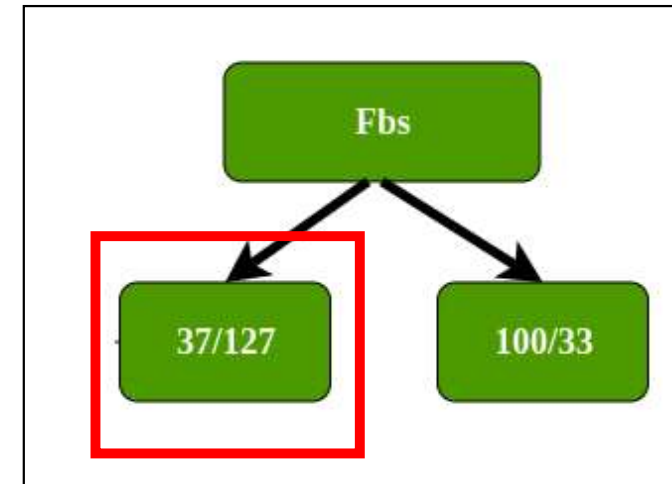
Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e Ingenierías - Universidad de Caldas



# Ejemplo



Gini Impurity - Exang = 0.290



# Ejemplo

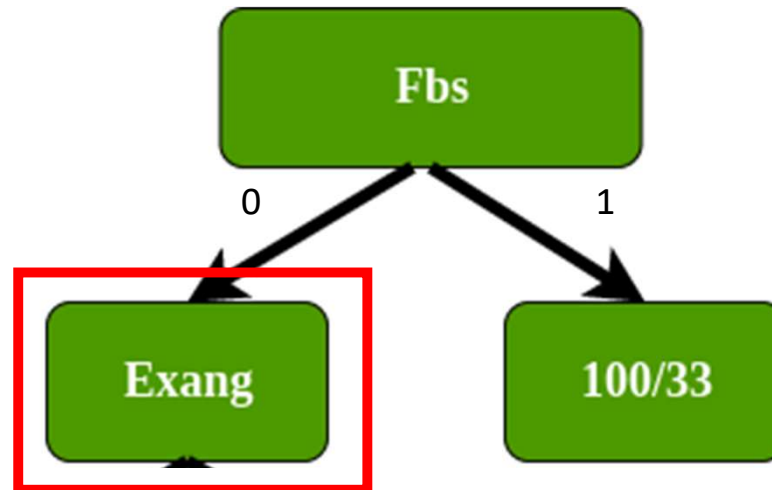
- Exang (angina inducida por ejercicio) tiene la impureza de Gini más baja.

Gini Impurity	
Fbs → Sexo	0.300
Fbs → Exang (angina inducida por el ejercicio)	0.290
Fbs → Hoja	0.349



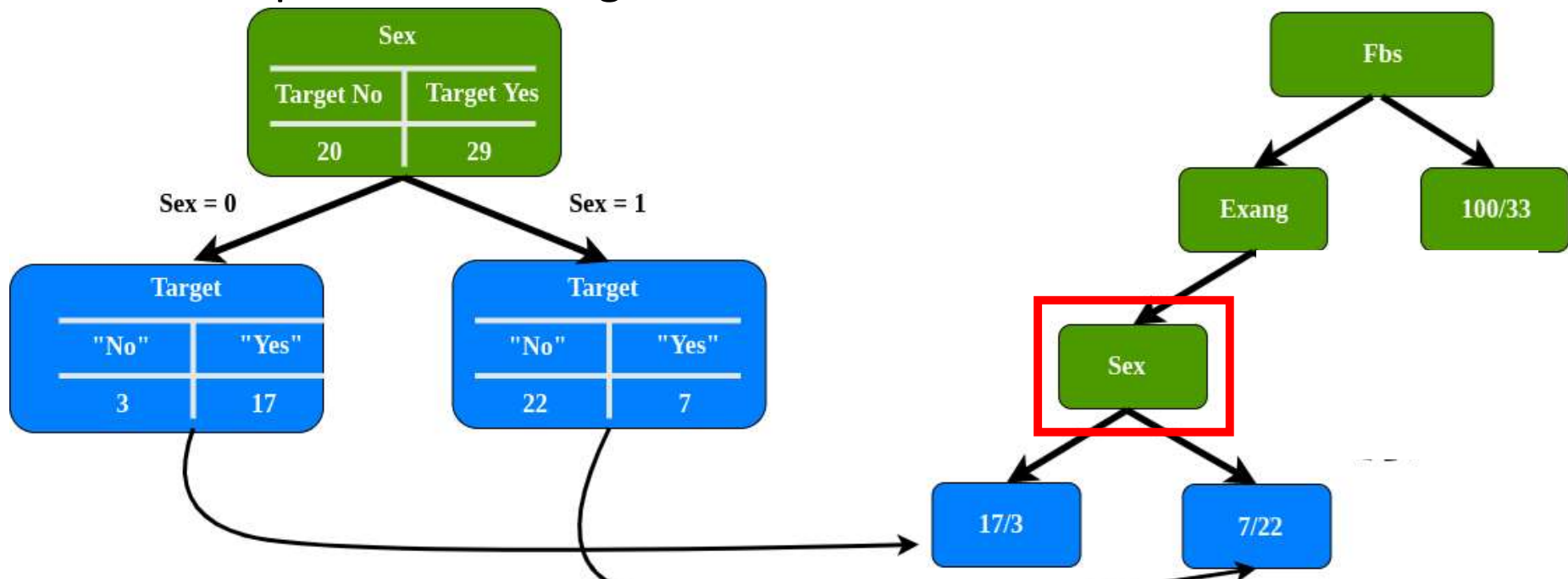
# Ejemplo

- Por tanto elegimos a Exang como nodo de decisión



# Ejemplo

- Como solo queda el atributo sexo, colocamos el atributo sexo en la rama izquierda de Exang



Gini Impurity – Sex= 0.253

Felipe Buitrago Carmona - Facultad de Ingeniería Artificial e Ingenierías - Universidad de Caldas

These are the final leaf nodes on this branch of the tree

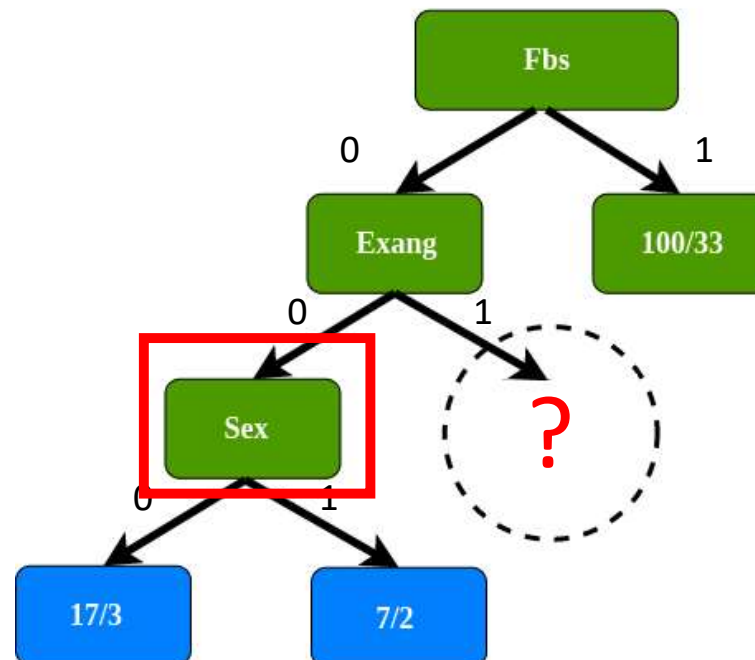
# Ejemplo

- Sexo tiene la impureza de Gini más baja.

Gini Impurity	
Fbs → Exang → Hoja	0.499
Fbs → Exang → Sexo → Hoja	0.253

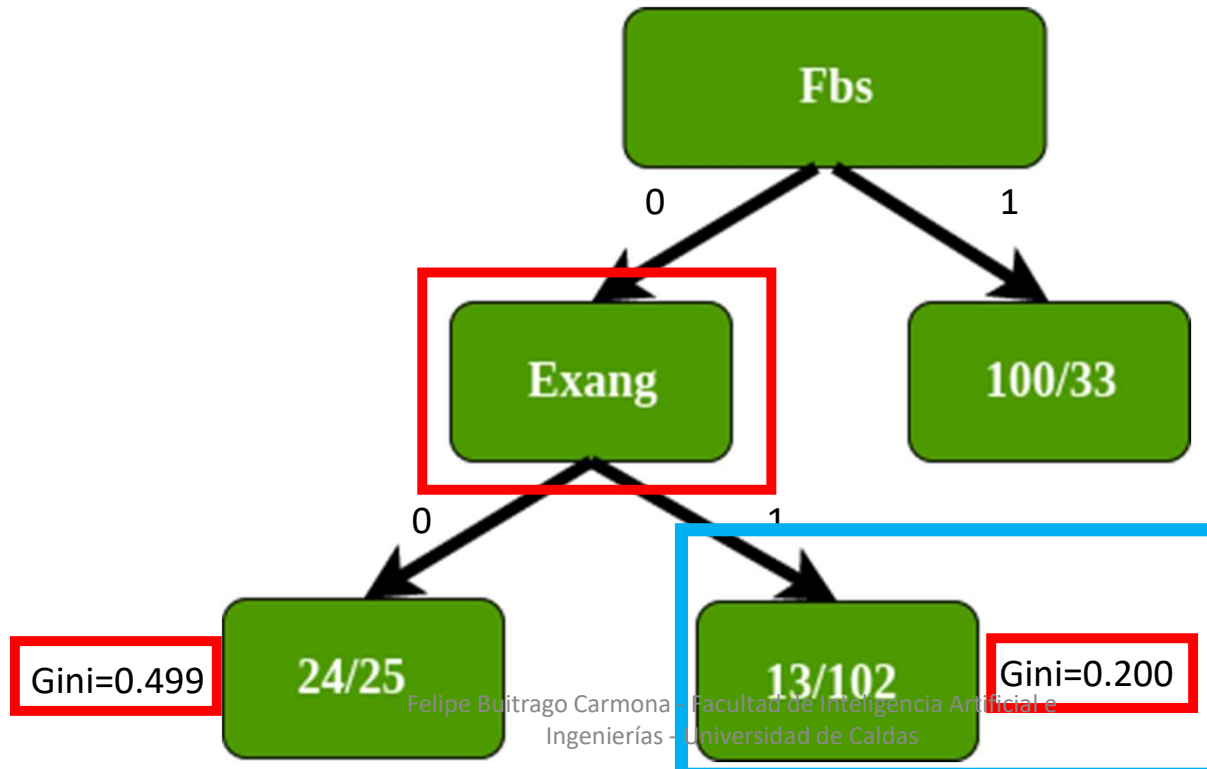
# Ejemplo

- Comparemos las impurezas de Gini para separar a los pacientes. El primer caso es usar el atributo sexo. El segundo es no utilizarlo



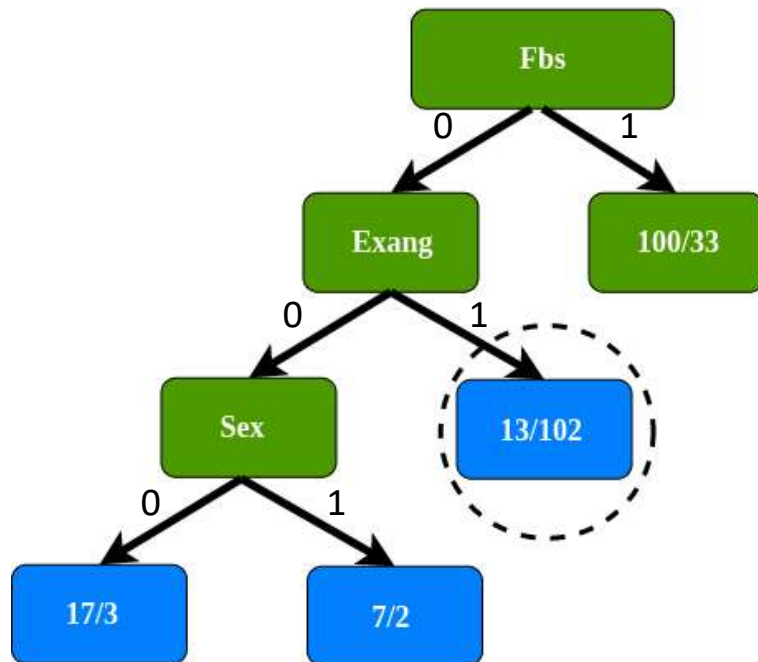
# Ejemplo

- Por tanto elegimos a Exang como nodo de decisión



# Ejemplo

- Comparemos las impurezas de Gini para separar a los pacientes. El primer caso es no usar el atributo sexo.

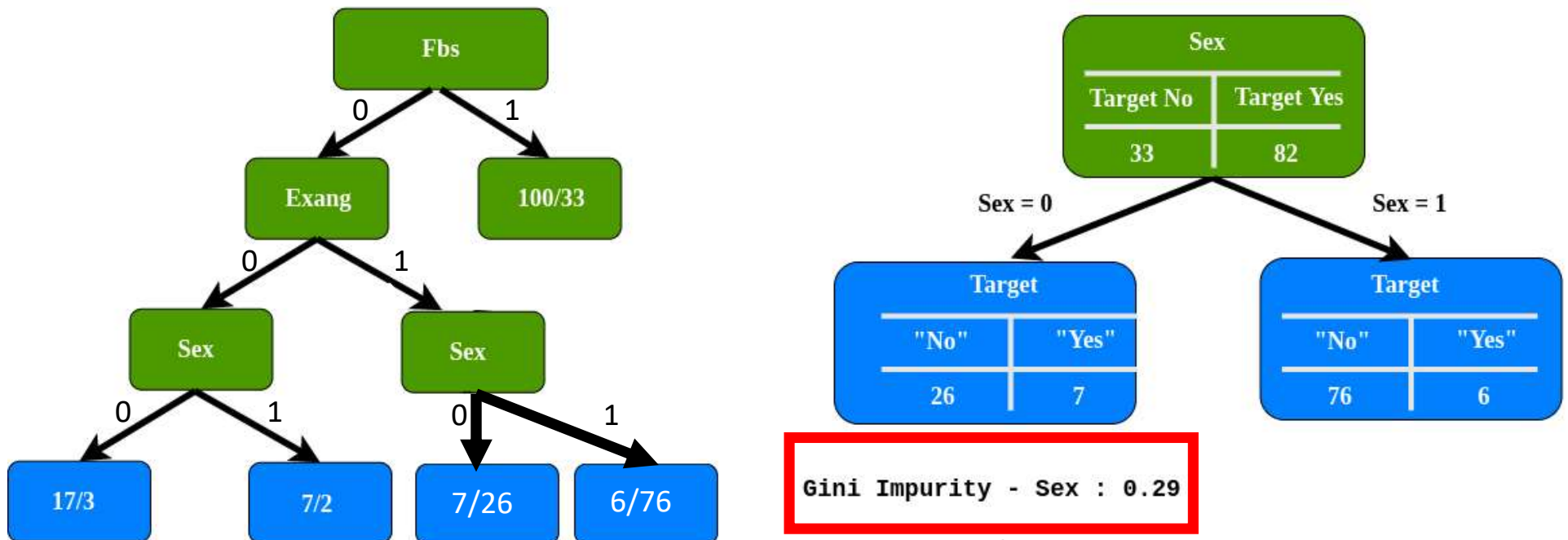


$$I = 1 - \left(\frac{13}{13 + 102}\right)^2 - \left(\frac{102}{13 + 102}\right)^2 = 0.20$$

**Gini Impurity - Sex : 0.20**

# Ejemplo

- Comparemos las impurezas de Gini para separar a los pacientes. **El segundo es utilizar el atributo de sexo**



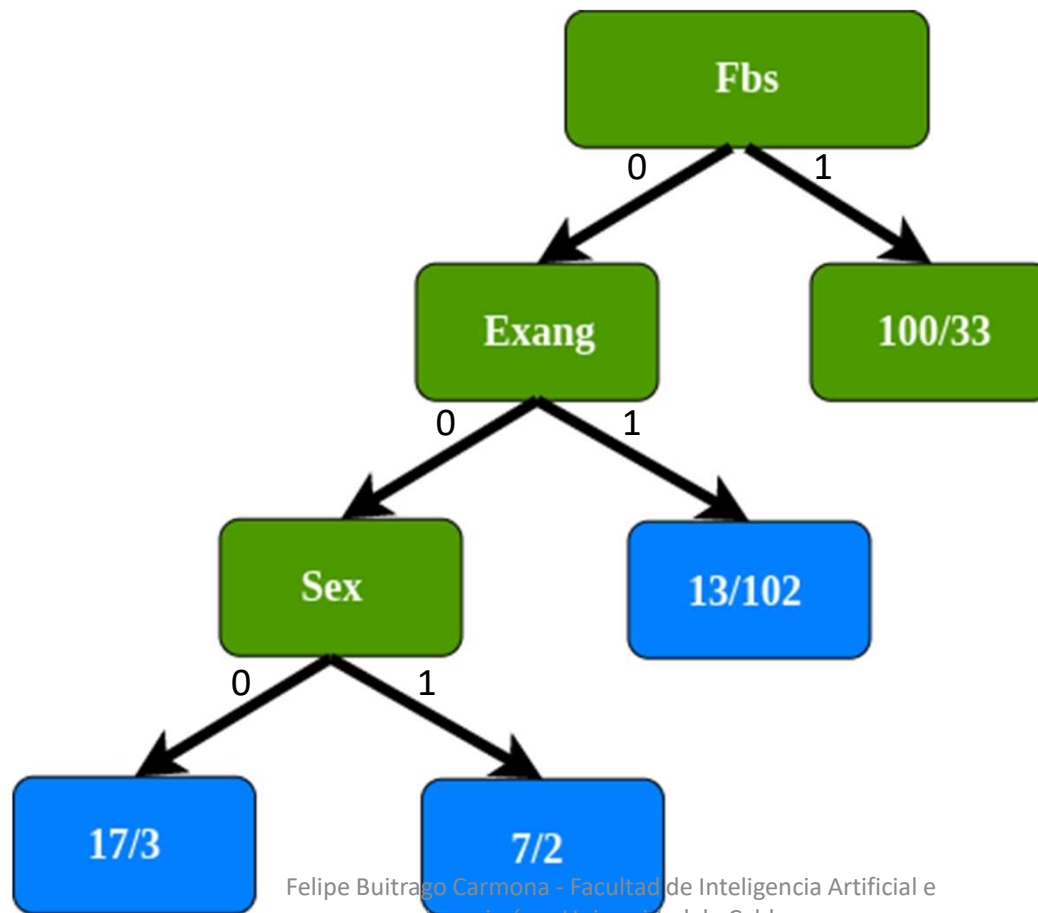
# Ejemplo

- La impureza de Gini **sin usar** el sexo para separar a los pacientes es la más baja, por lo que no la incluiremos.

Gini Impurity	
Fbs→ Exang→Sexo→Hoja	0.29
Fbs→ Exang→ Hoja	0.20

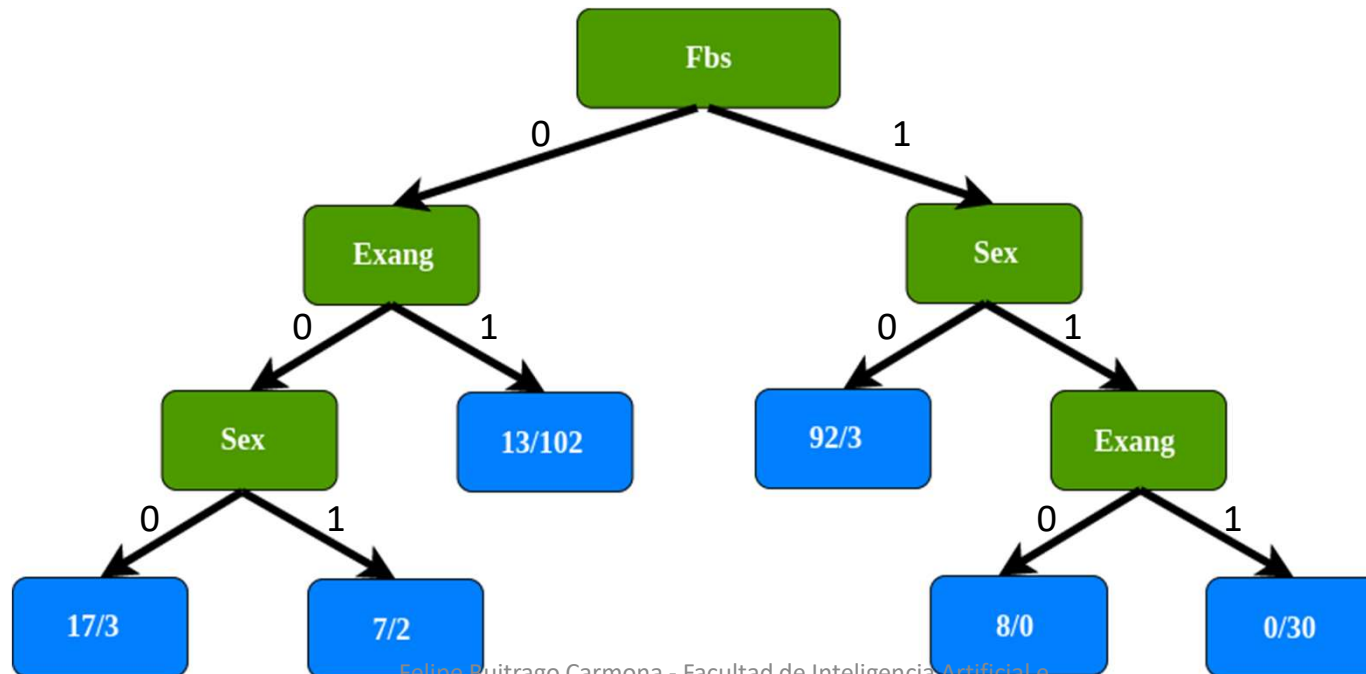


# Ejemplo



# Ejemplo

- Haga lo mismo en la rama derecha, por lo que el resultado final del árbol en este caso es



# ¿Cómo calcular la impureza de Gini en datos continuos?


# Gini en datos continuos

- El peso, que es uno de los atributos para determinar la enfermedad cardíaca, por ejemplo, tenemos el atributo de peso

Weight	Heart Disease
220	Yes
180	Yes
225	Yes
190	No
155	No

# Paso 1: Ordene los datos ascendiendo

**Lowest**



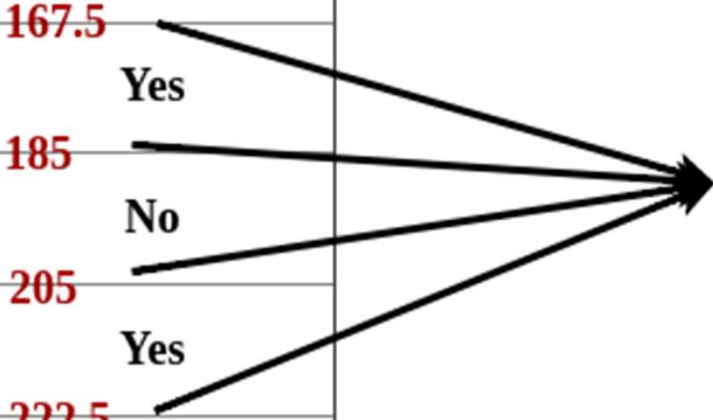
**Highest**

<b>Weight</b>	<b>Heart Disease</b>
<b>155</b>	<b>No</b>
<b>180</b>	<b>Yes</b>
<b>190</b>	<b>No</b>
<b>220</b>	<b>Yes</b>
<b>225</b>	<b>Yes</b>

Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e Ingenierías - Universidad de Caldas

## Paso 2: calcula el peso promedio

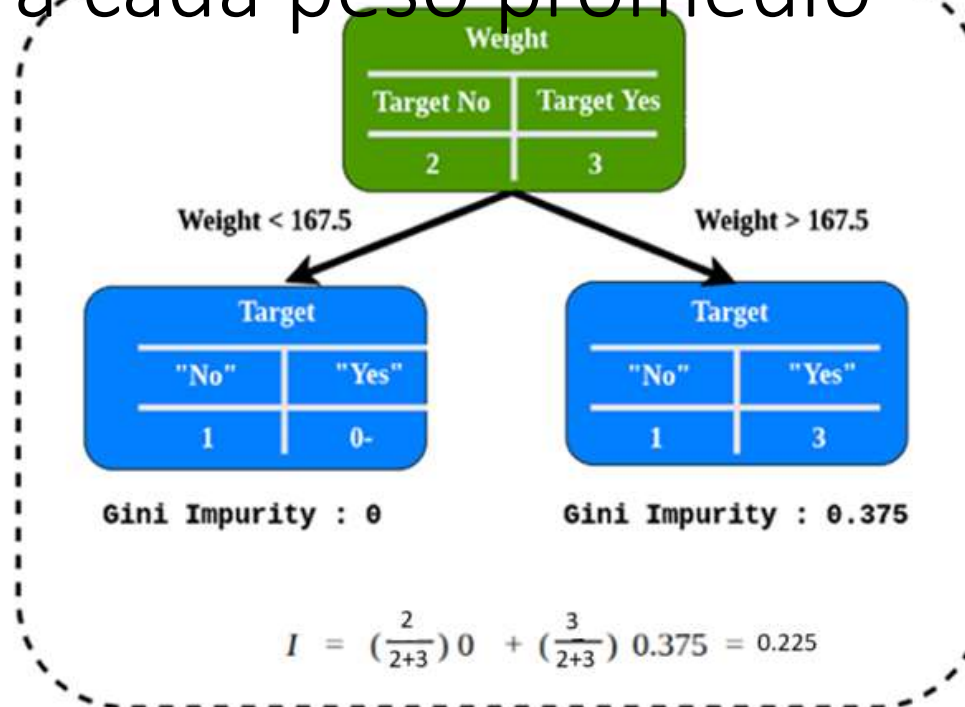
Weight	Heart Disease
155	No
180	Yes
190	No
220	Yes
225	Yes



Calculate the average weight

## Paso 3: Calcule los valores de impureza de Gini para cada peso promedio

Weight	Heart Disease
155	No
180	Yes
190	No
220	Yes
225	Yes



Weight	Heart Disease	Gini Impurity
155	No	
180	Yes	0.225
190	No	0.47
220	Yes	0.27
225	Yes	0.4

La impureza de Gini más baja es **Peso < 225**, este es el valor de corte e impureza si se usa cuando lo comparamos con otro atributo

# Enlaces de interés

- [https://www.youtube.com/watch?v=lacGvKfR28g&ab\\_channel=AMPTech](https://www.youtube.com/watch?v=lacGvKfR28g&ab_channel=AMPTech)
- [https://www.youtube.com/watch?v=coOTEc-0OGw&ab\\_channel=MaheshHuddar](https://www.youtube.com/watch?v=coOTEc-0OGw&ab_channel=MaheshHuddar) (VER EJERCICIO ID3)
- [https://www.youtube.com/watch?v=UdTKxGQvYdc&ab\\_channel=CodeWrestling](https://www.youtube.com/watch?v=UdTKxGQvYdc&ab_channel=CodeWrestling) (complemento anterior)
- [https://www.youtube.com/watch?v=0X3IIPYKRz4&ab\\_channel=VieraClass](https://www.youtube.com/watch?v=0X3IIPYKRz4&ab_channel=VieraClass)
- [https://www.youtube.com/watch?v=269QJ5joMCc&ab\\_channel=AMPTech](https://www.youtube.com/watch?v=269QJ5joMCc&ab_channel=AMPTech)
- <https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-example/> (Ejemplo paso a paso)
- [https://www.youtube.com/watch?v=7VeUPuFGJHk&t=911s&ab\\_channel=StatQuestwithJoshStarmer](https://www.youtube.com/watch?v=7VeUPuFGJHk&t=911s&ab_channel=StatQuestwithJoshStarmer)
- [https://www.youtube.com/watch?v=q90UDEgYqel&ab\\_channel=StatQuestwithJoshStarmer](https://www.youtube.com/watch?v=q90UDEgYqel&ab_channel=StatQuestwithJoshStarmer)



# FIN

Felipe Buitrago Carmona - Facultad de Inteligencia Artificial e  
Ingenierías - Universidad de Caldas