Institución Universitaria
Reacreditada en Alta Calidad
80 Años

Árboles de Decisión

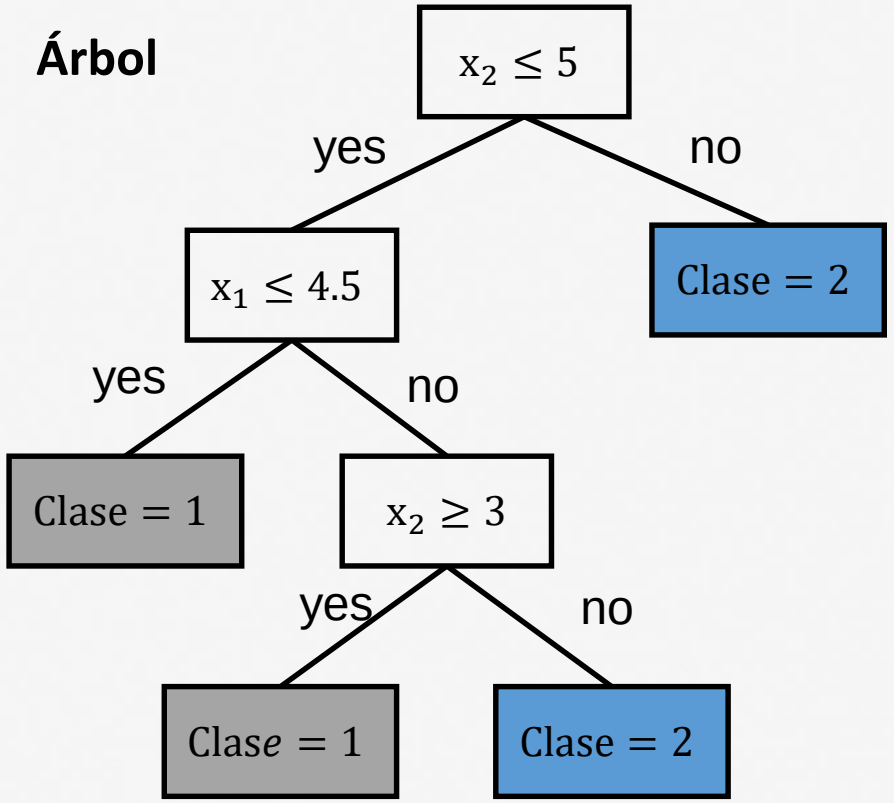
Juego de las 20 preguntas

Notebook

Qué es un árbol de decisión

- Un árbol de decisión es una serie de preguntas si/no que se hacen de forma secuencial en los datos

Árbol



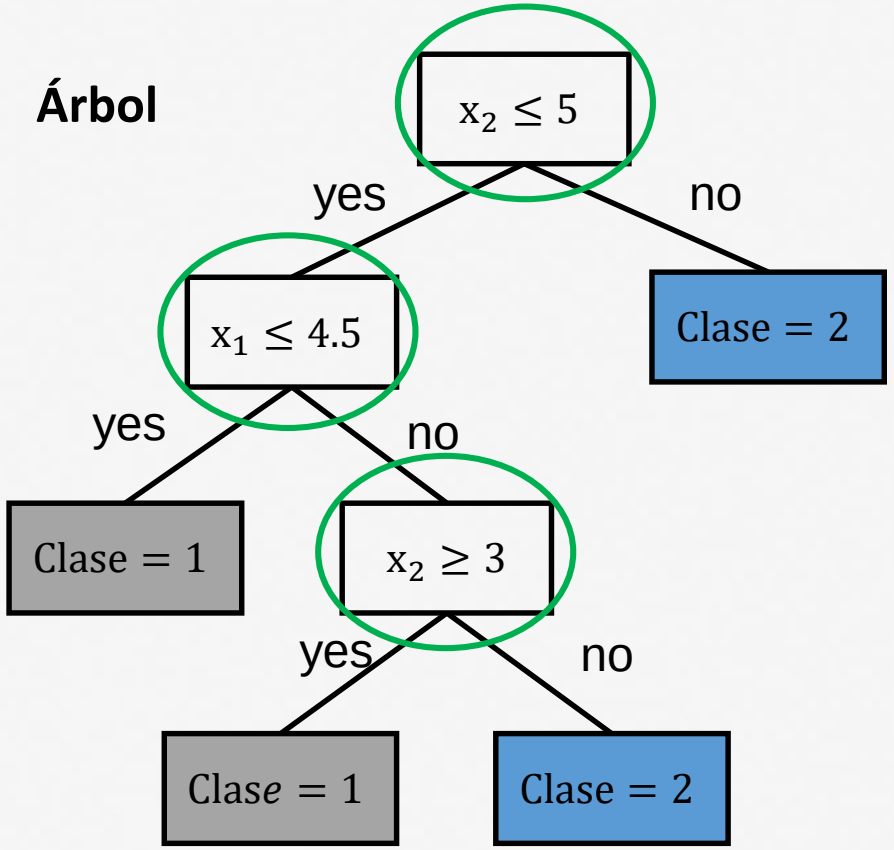
Conjunto de Datos

x_1	x_2	y
3.5	2	1
5	2.5	2
1	3	1
2	4	1
4	2	1
6	6	2
2	9	2
4	9	2
5	4	1
3	8	2

Qué es un árbol de decisión

- Un árbol de decisión es una serie de preguntas si/no que se hacen de forma secuencial en los datos

Árbol



Nodo o rama

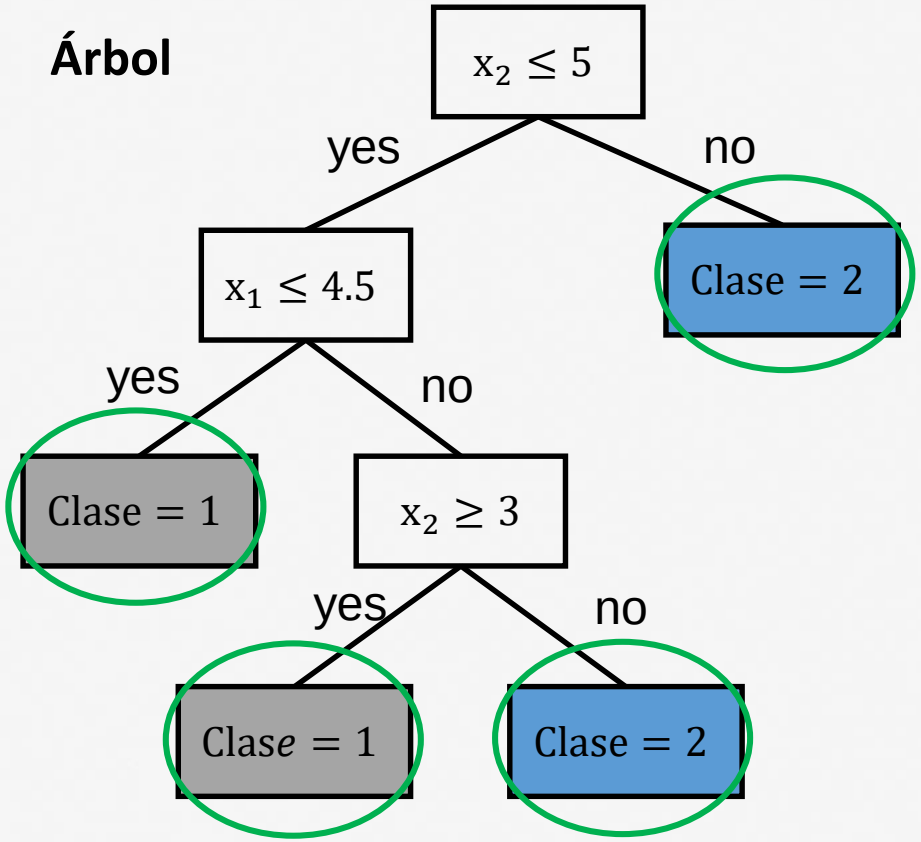
Conjunto de Datos

x_1	x_2	y
3.5	2	1
5	2.5	2
1	3	1
2	4	1
4	2	1
6	6	2
2	9	2
4	9	2
5	4	1
3	8	2

Qué es un árbol de decisión

- Un árbol de decisión es una serie de preguntas si/no que se hacen de forma secuencial en los datos

Árbol



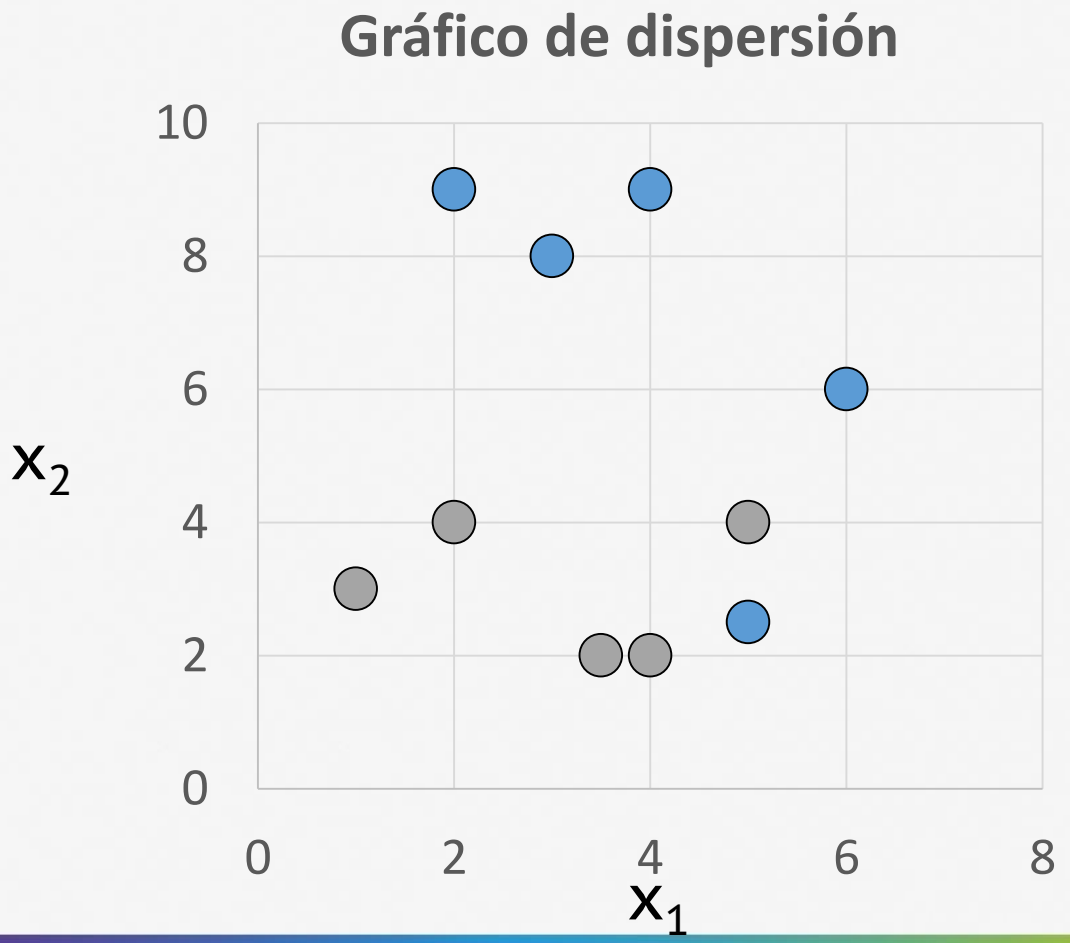
Hojas

Conjunto de Datos

x_1	x_2	y
3.5	2	1
5	2.5	2
1	3	1
2	4	1
4	2	1
6	6	2
2	9	2
4	9	2
5	4	1
3	8	2

Qué es un árbol de decisión

- Un árbol de decisión es una serie de preguntas si/no que se hacen de forma secuencial en los datos

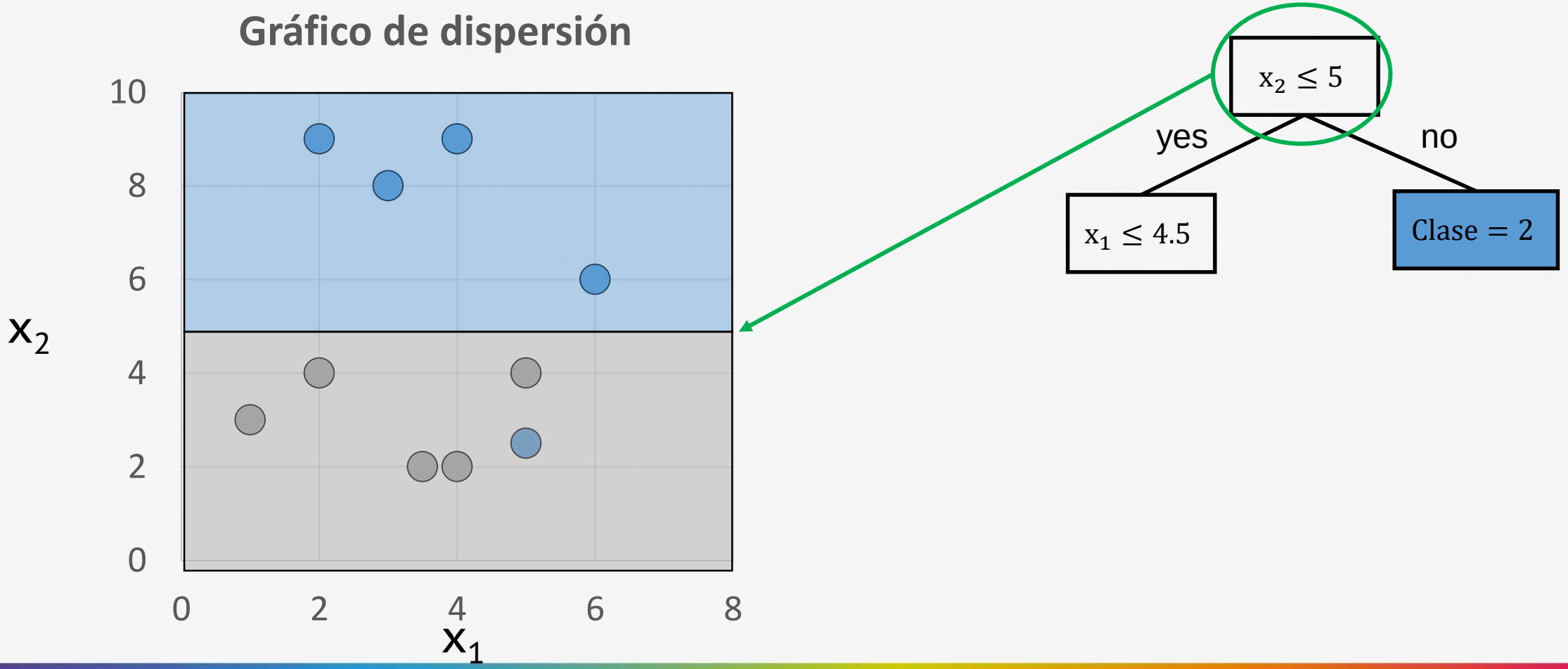


Conjunto de Datos

x_1	x_2	y
3.5	2	1
5	2.5	2
1	3	1
2	4	1
4	2	1
6	6	2
2	9	2
4	9	2
5	4	1
3	8	2

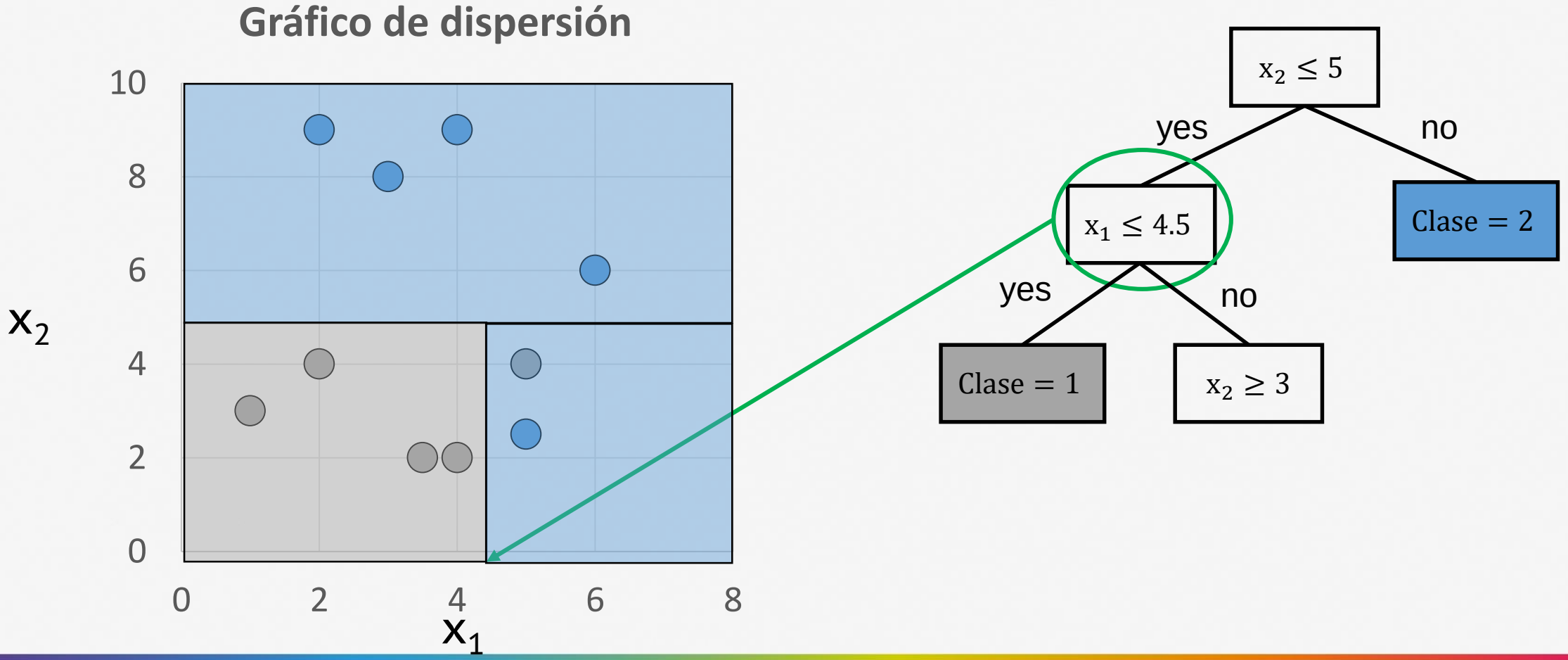
Qué es un árbol de decisión

- Un árbol de decisión es una serie de preguntas si/no que se hacen de forma secuencial en los datos



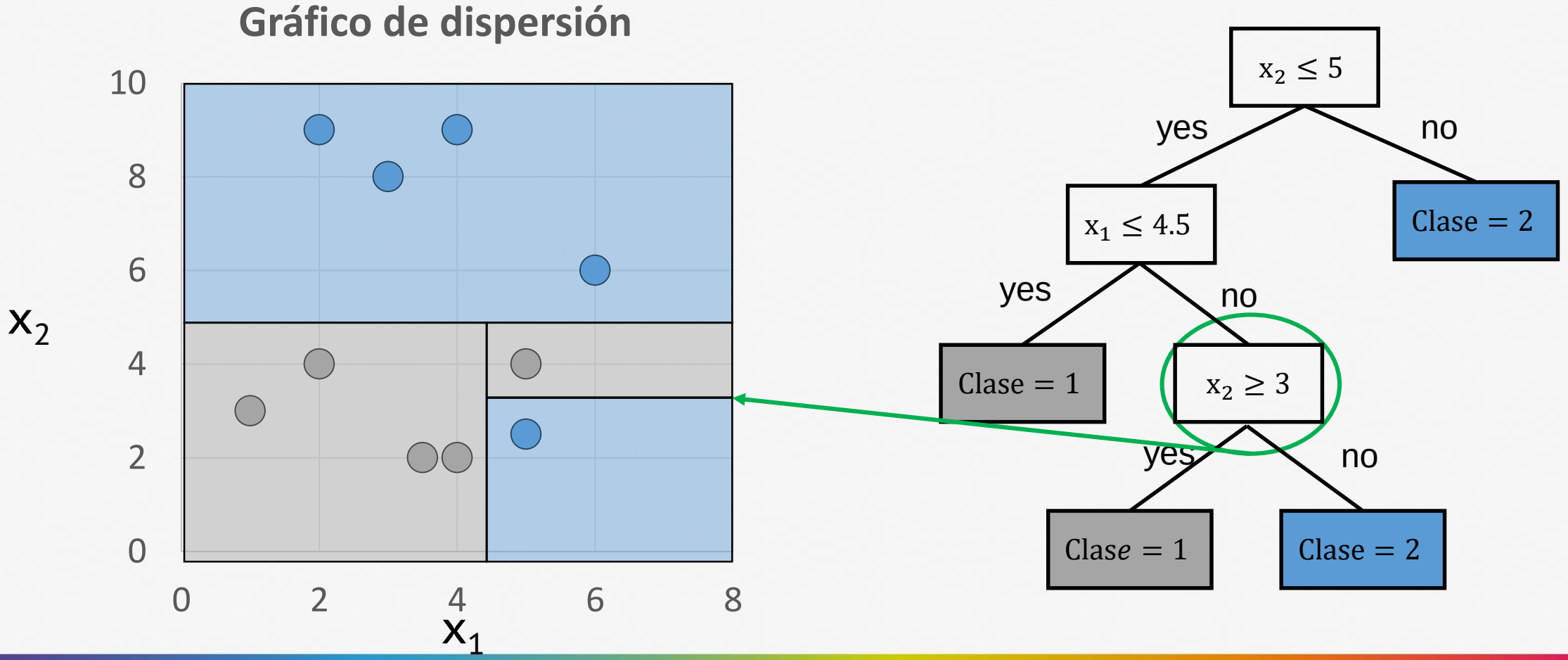
Qué es un árbol de decisión

- Un árbol de decisión es una serie de preguntas si/no que se hacen de forma secuencial en los datos



Qué es un árbol de decisión

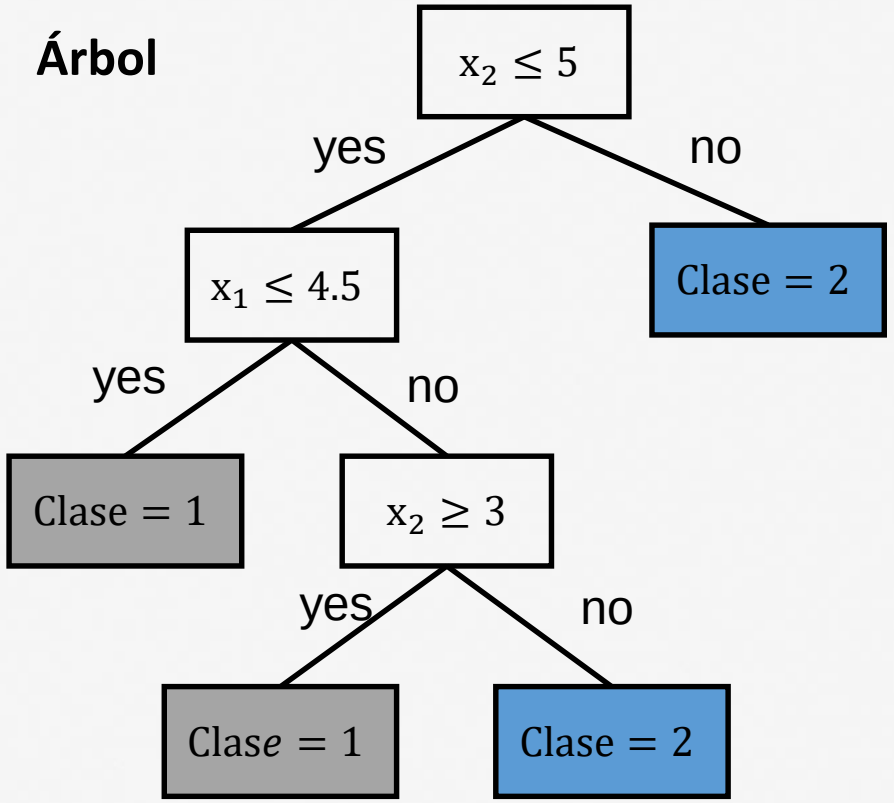
- Un árbol de decisión es una serie de preguntas si/no que se hacen de forma secuencial en los datos



Qué es un árbol de decisión

- Predicción con algún dato

Árbol



Conjunto de Datos

x_1	x_2	y
3.5	2	1
5	2.5	2
1	3	1
2	4	1
4	2	1
6	6	2
2	9	2
4	9	2
5	4	1
3	8	2

Qué es un árbol de decisión

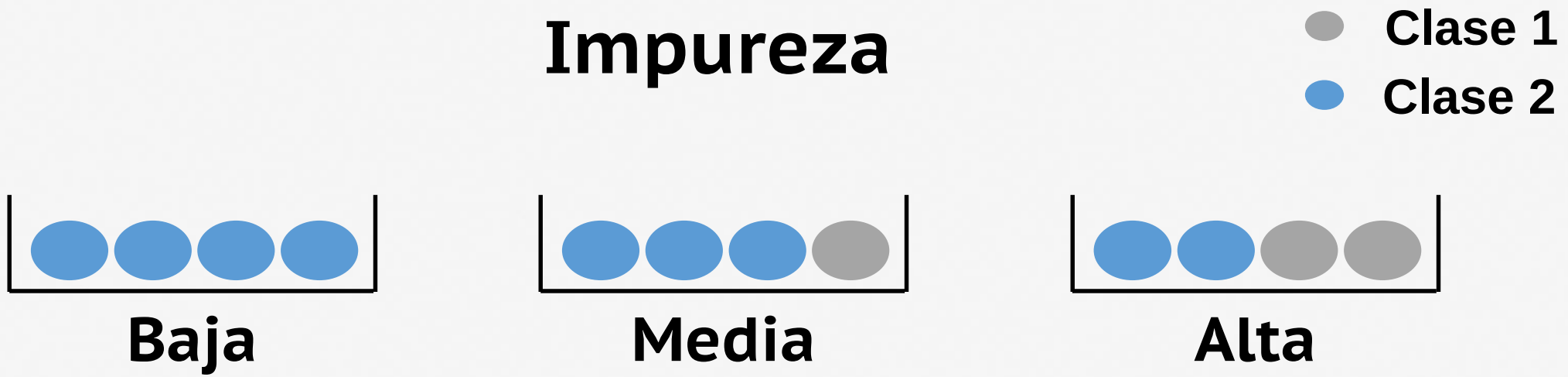
- Los árboles aprenden al dividir recursivamente cada nodo de manera que maximiza el incremento de “pureza”

- El proceso finaliza con un criterio de parada:
  - Cuando se alcanza una profundidad máxima
  - Cuando se alcanza un número máximo de hojas
  - Cuando hay muy pocos datos en una hoja particular
  - Cuando las hojas han alcanzado un nivel de pureza deseado

Impureza

- Para que el aprendizaje es importante manejar alguna noción de impureza. Esto sirve para decidir qué pregunta hacer en cada rama, al considerar la impureza en sus hijos (ramas si/no).

- Se busca que la impureza sea lo más baja posible, o lo que es igual, que la pureza sea alta.



Criterios

- Existen principalmente dos opciones:
  - Entropía: $i(p_1, \ldots, p_k) = -\sum_{j=1}^{k} p_j \log_2(p_j)$
  - Gini: $i(p_1, \ldots, p_k) = \sum_{j=1}^{k} p_j(1 - p_j)$
- Son extremadamente similares, no presentan diferencia en el rendimiento del modelo final. Generalmente se utiliza Gini index.

Ejemplo numérico

Ejemplo de cálculo de Gini:

- ¿Qué valor de Gini nos da si realizamos la pregunta en rama “Llueve?”
- ¿Qué valor de Gini si preguntamos “Tengo dinero”?

$$i(p_1, \ldots, p_k) = \sum_{j=1}^{k} p_j(1 - p_j)$$

Llueve	Tengo dinero	°C	**Salir**
Sí	Sí	7	NO
Sí	No	12	NO
No	Sí	18	SÍ
No	Sí	23	SÍ
Sí	Sí	26	SÍ
Sí	No	32	NO
No	No	34	NO

Institución Universitaria
Reacreditada en Alta Calidad
80 Años

¡MUCHAS GRACIAS!

VIGILADA MINEDUCACIÓN

www.itm.edu.co

Alcaldía de Medellín
Distrito de
Ciencia, Tecnología e Innovación