

# **ARBOLES DE DECISIÓN**

**Mg Sc. Meza Rodríguez, Aldo Richard**

## TÉCNICAS SUPERVISADAS

### Árboles de decisión



### Supervised

- **Clasificación**

- Naive Bayes
- Logística
- Árboles de clasificación
- Bagging
- Adaboost
- SVM
- Redes neuronales, etc.

- **Regresión**

- Lineal
- Árboles de regresión
- Etc.



Label

### Unsupervised

- **Clustering**

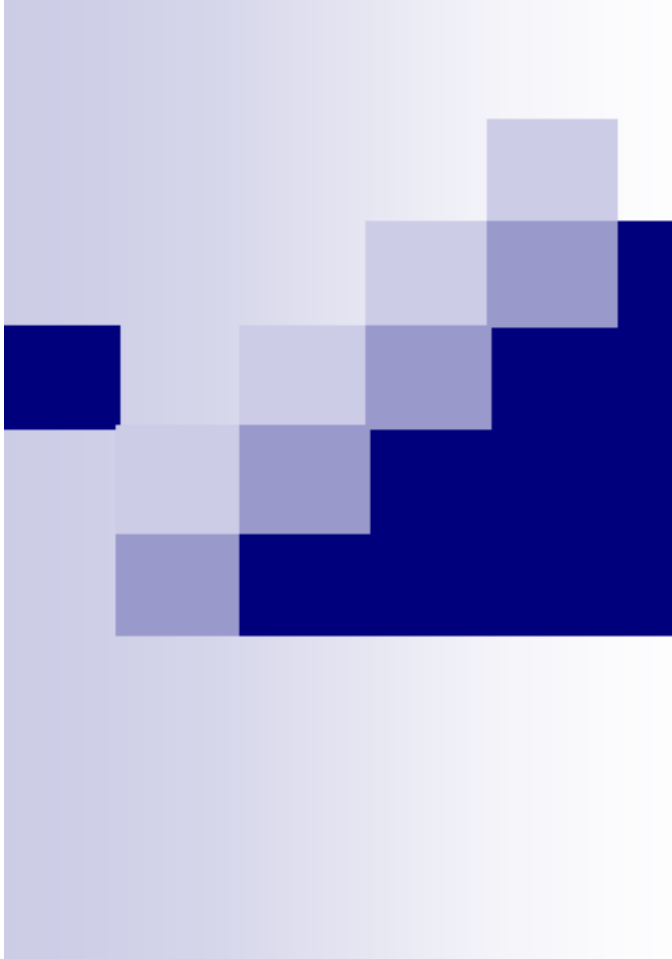
- Kmean
- Clara
- Fanny
- Mona, etc.

- **Reglas de asociación**

- **Reducción de dimensiones**

- Análisis de compones principales
- SVD





## Indicadores para evaluación de modelos de clasificación

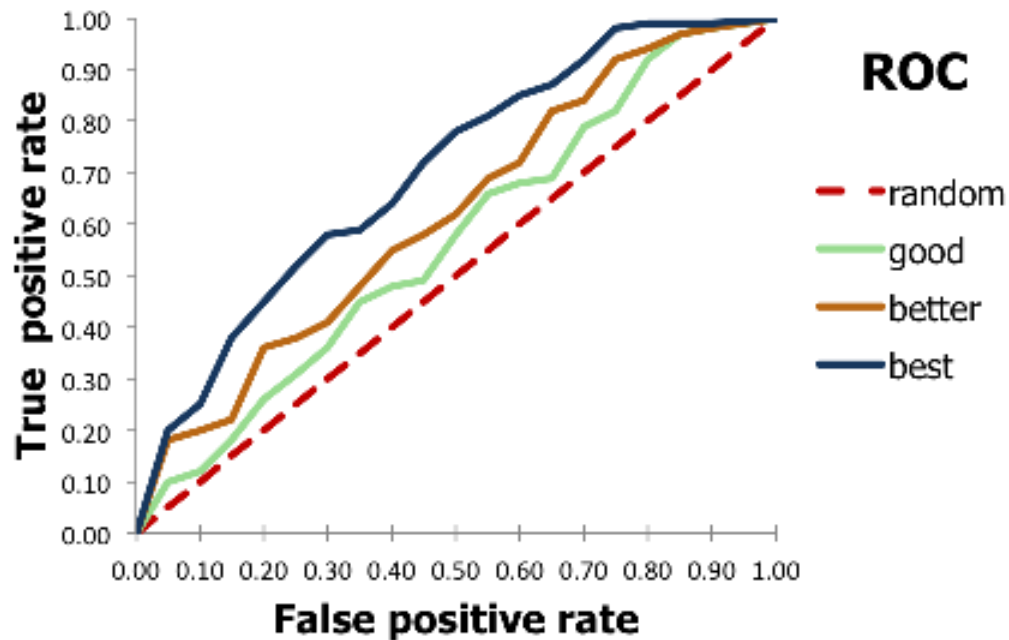
## MATRIZ DE CONFUSIÓN

		CLASE OBSERVADA (ACTUAL)	
		Positivo	Negativo
CLASE PREDICHA	Positivo	Verdadero Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadero Negativo (VN)

- ❑ **Accuracy:** Es la tasa de predicción correcta  $(VP + VN)/\text{Total de casos}$
- ❑ **Tasa de error:**  $(FP + FN)/\text{Total de casos} = 1 - \text{Accuracy}$
- ❑ **Sensibilidad:** Es la probabilidad de clasificar correctamente a un valor positivo.  $VP/(FN+VP)$
- ❑ **Especificidad:** Es la probabilidad de clasificar correctamente a un valor negativo.  $VN/(VN+FP)$
- ❑ **Precisión= (VP+):** Probabilidad de clasificar incorrectamente a un valor negativo =  $VP/(FP+VP)$
- ❑ **Valor Predictivo Negativo (VP-):** Probabilidad de clasificar incorrectamente a un valor positivo =  $VN/(VN+FN)$

## CURVAS ROC

- Se forma con la tasa verdadera positiva (sensitividad) en función de la tasa falsa positiva ( $1$ -especificidad) para diferentes puntos de corte de un modelo.
- El clasificador tiene un mejor desempeño si la curva ROC se aleja más de la diagonal principal.
- El área bajo la curva (AUC) la probabilidad de que el clasificador pronostique correctamente a dos individuos. Uno que pertenece a la categoría  $Y=1$  y otro que pertenece a  $Y=0$ .

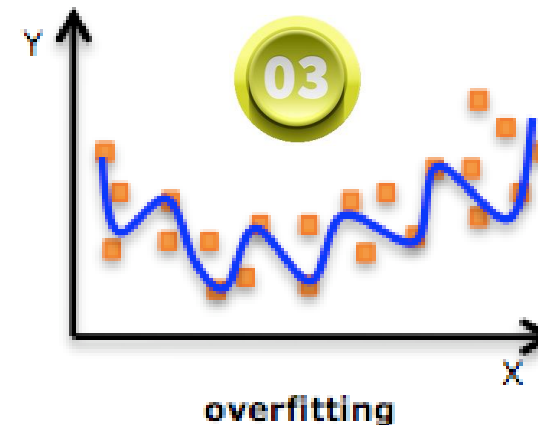
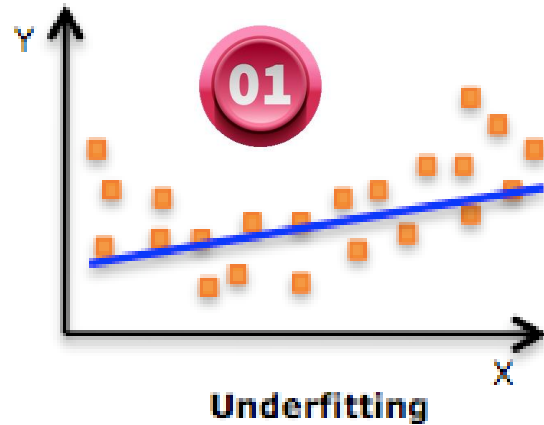


AUC ROC	Conclusión
0.5	No discrimina
0.6 - 0.7	Pobre
0.71 - 0.8	Aceptable
0.81 - 0.9	Excelente (bueno)
> 0.9	Muy bueno



## Métodos de Validación

## Problemas en los modelos

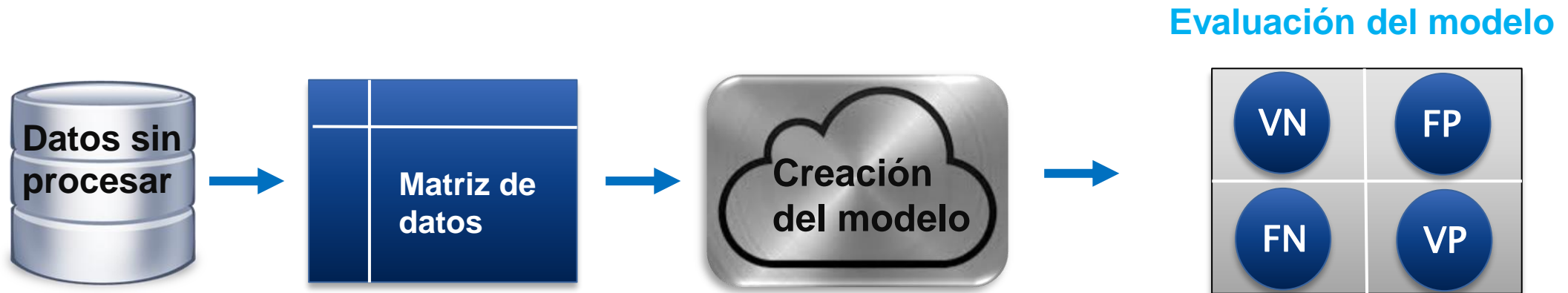


**1. Underfitting:** El modelo no captura la tendencia subyacente de los datos. La ecuación lineal tiene un alto error de los puntos de datos de entrenamiento. Esto no funcionará bien en la clasificación.

**2.** La relación entre las variables es correcta, el error de entrenamiento es bajo, existe buena generalización de los datos.

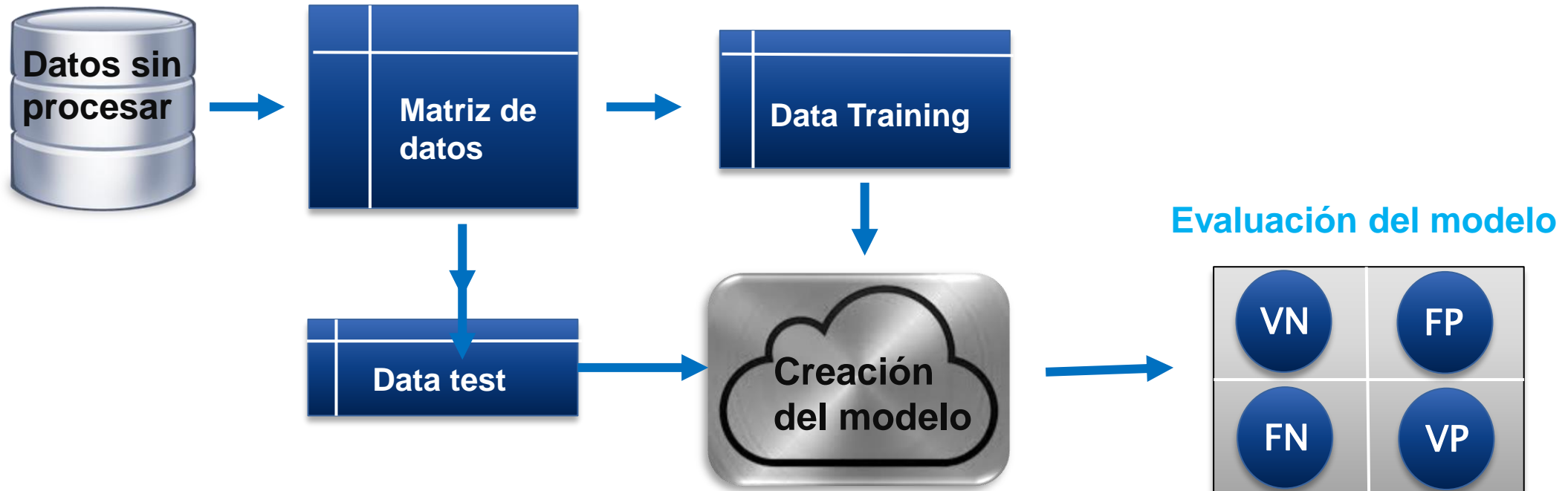
**3. Overfitting:** funciona bien en los datos de entrenamiento pero no tiene un buen rendimiento en datos no vistos. Usual en modelos **no paramétricos y no lineales**, sucede porque el modelo está memorizando la relación entre las variables  $X$  y  $Y$ , no puede generalizar bien los datos.

## VALIDACIÓN POR RESTITUCIÓN

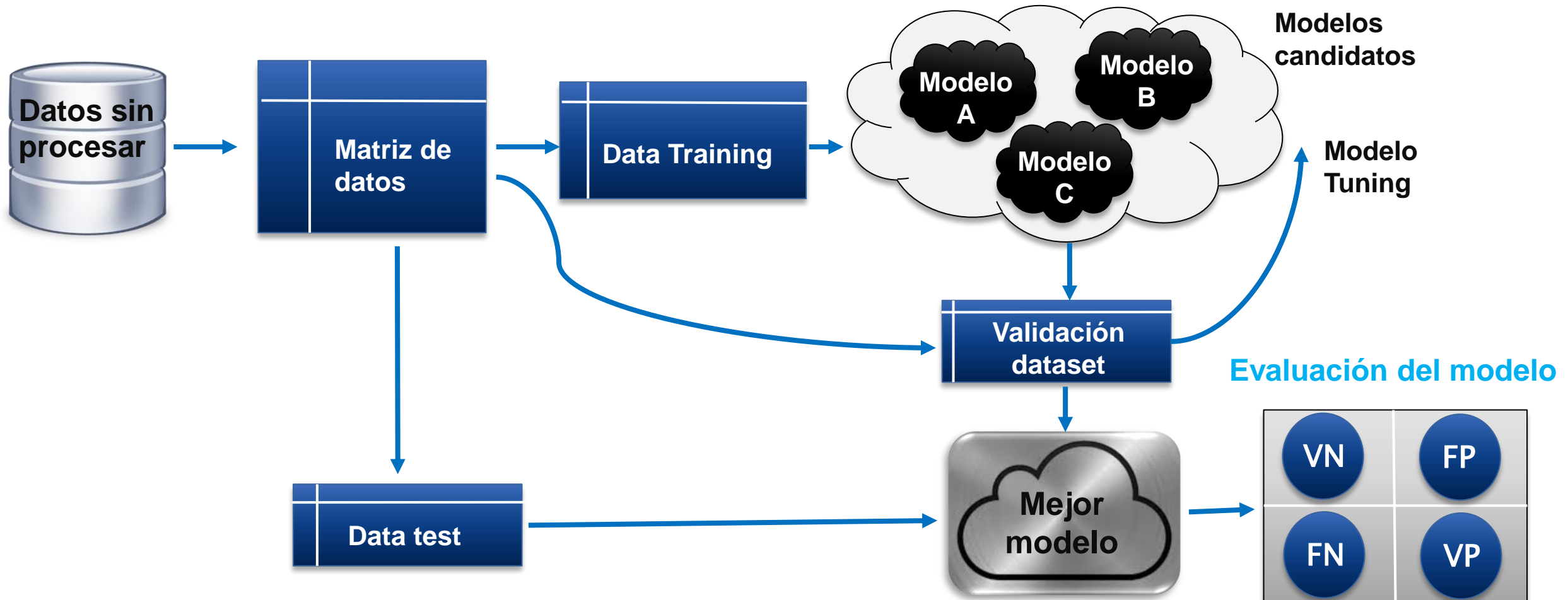




## VALIDACIÓN POR CONJUNTO DE PRUEBA



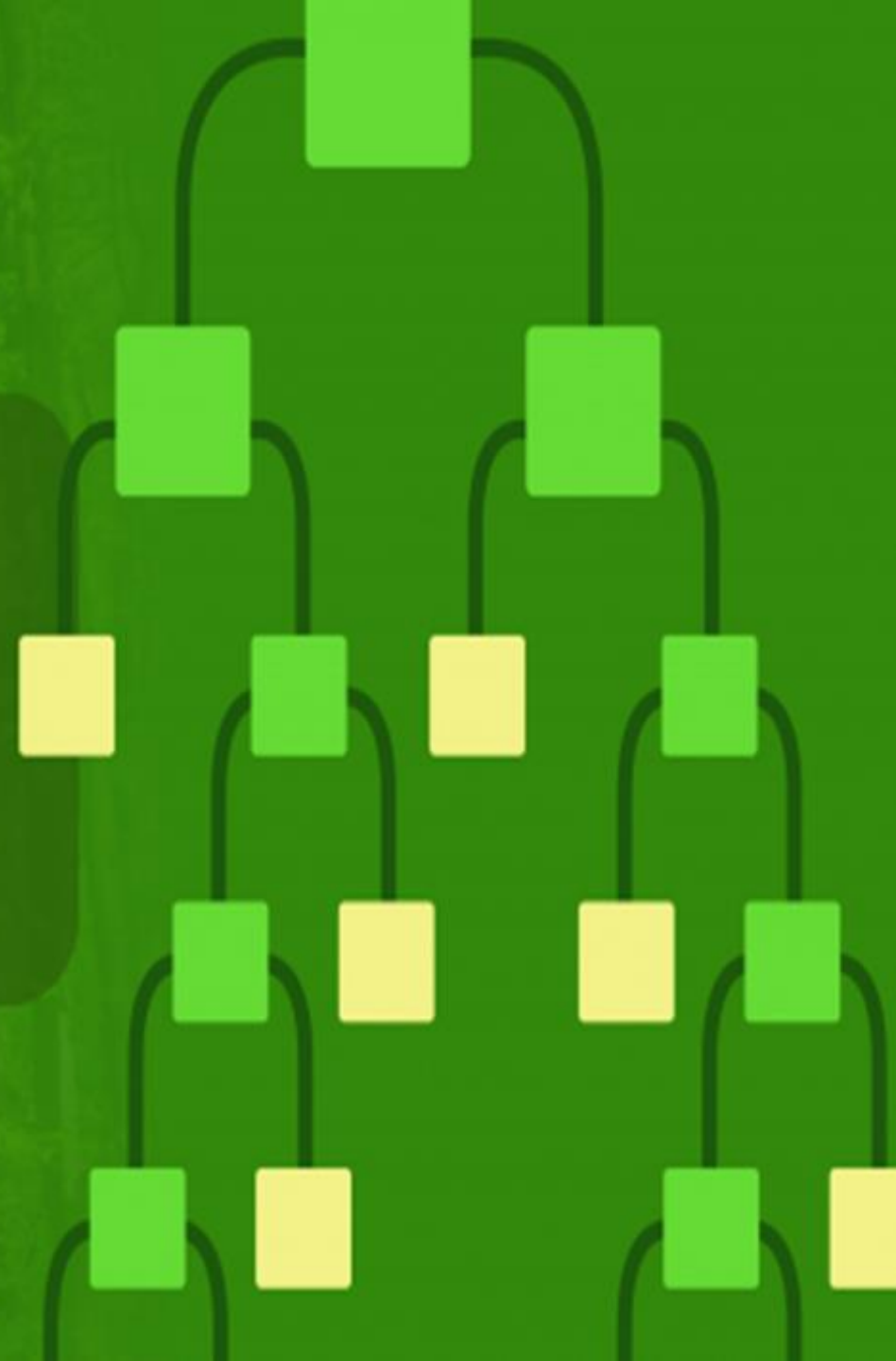
## Validación por Conjunto de Prueba y Modelo Tuning



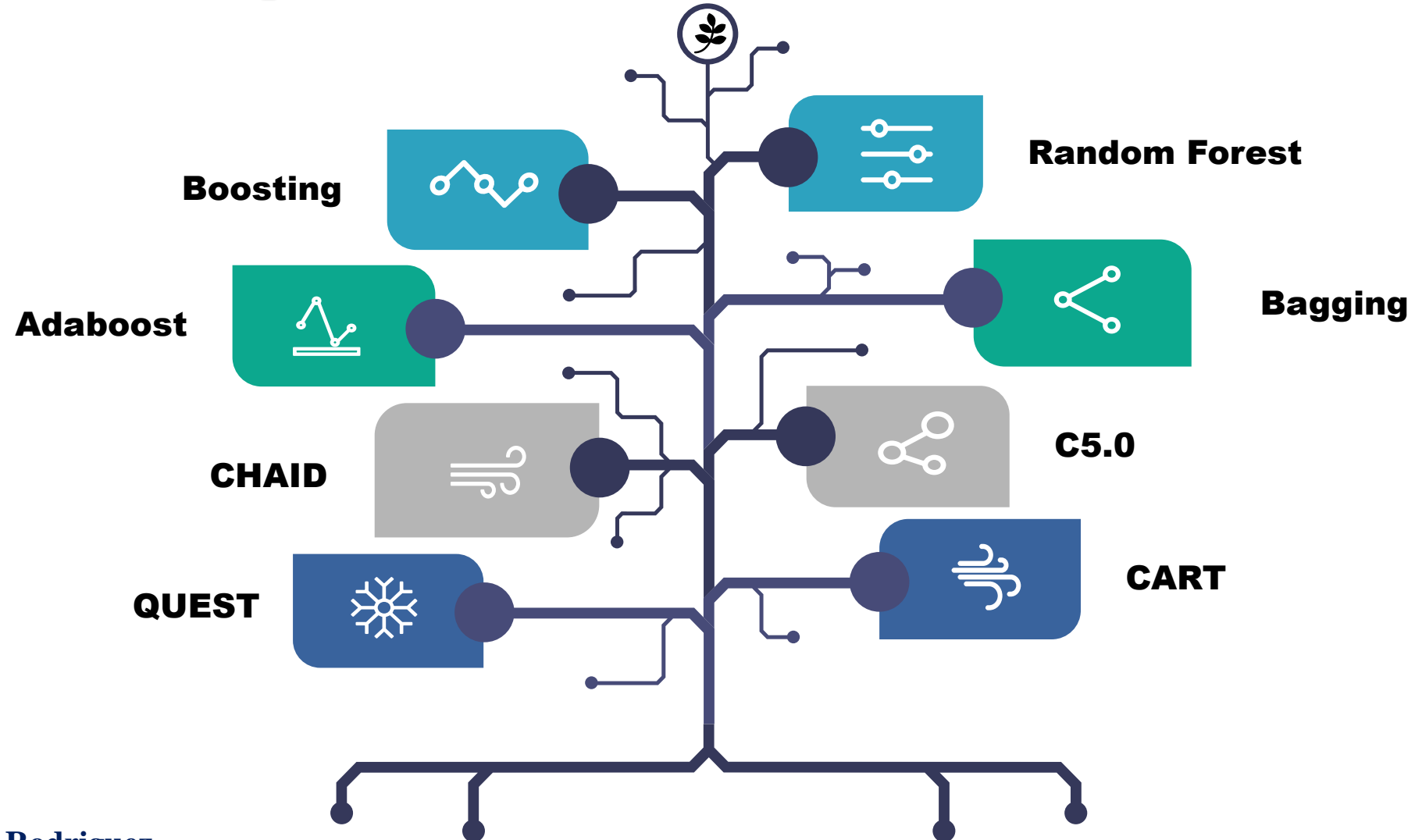
Árboles de decisión

# ÁRBOLES DE DECISIÓN

Mg.Sc Aldo Meza Rodriguez



## Principales árboles de clasificación



## Los árboles de clasificación





## FORMACIÓN DEL ÁRBOL



### Construir el árbol

Al inicio los ejemplos de entrenamiento están en la raíz. Partir los ejemplos en forma recursiva basado en los atributos seleccionados.



### Detener la construcción



### Podar el árbol

Identificar y eliminar ramas que reflejan ruido o valores atípicos.

## Árbol de decisión



## Comparación de diferentes algoritmos

MÉTODOS	CART	C5.0	CHAID	QUEST
<b>Medida utilizada para seleccionar la variable de entrada</b>	Índice de Gini; Dos criterios	Entropía Información-de nuevo	Chi-cuadrado	Chi-cuadrado para variables categóricas; ANOVA J-vias para variables continuas / ordinales
<b>Poda</b>	Pre-Poda utilizando un algoritmo de pase único	Pre-Poda utilizando un algoritmo de pase único	Pre-Poda usando la prueba de Chi-cuadrado de independencia	Post-poda
<b>Variable dependiente</b>	Categórico / Continuo	Categórico / Continuo	Categorica	Categórico
<b>Variables de entrada</b>	Categórico / Continuo	Categórico / Continuo	Categórico / Continuo	Categórico / Continuo
<b>División en cada nodo</b>	Binario; Dividir en combinaciones lineales	Múltiple	Múltiple	Binario; Dividir en combinaciones lineales



## El Algoritmo CART

Clasificación y Regresión Árboles o CART para abreviar es un término introducido por **Leo Breiman** para referirse a los algoritmos del Árbol de Decisión que pueden usarse para problemas de modelado predictivo de clasificación o regresión.



Uno de los métodos de aprendizaje supervisado no paramétrico más utilizado



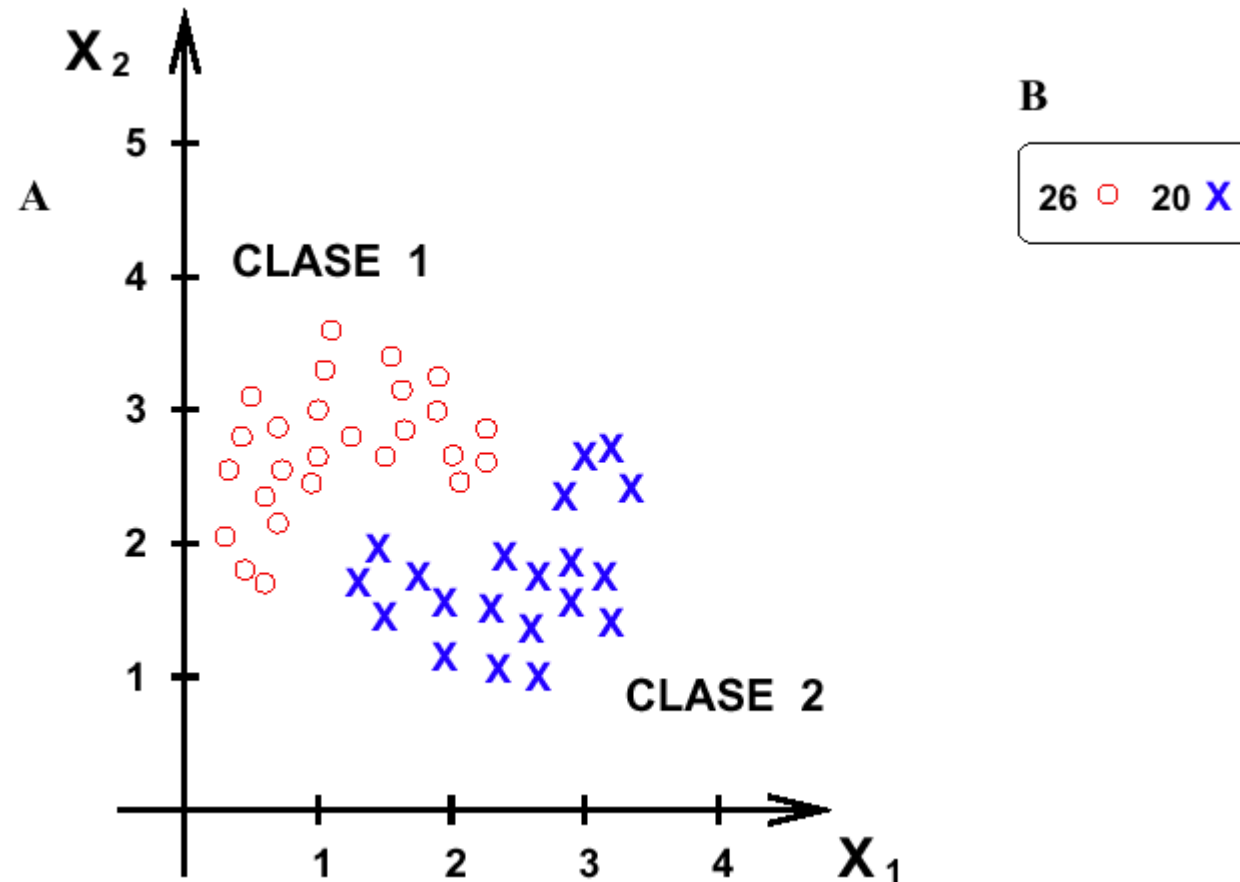
Realizar sucesivas divisiones binarias en un conjunto de datos guiado por un criterio



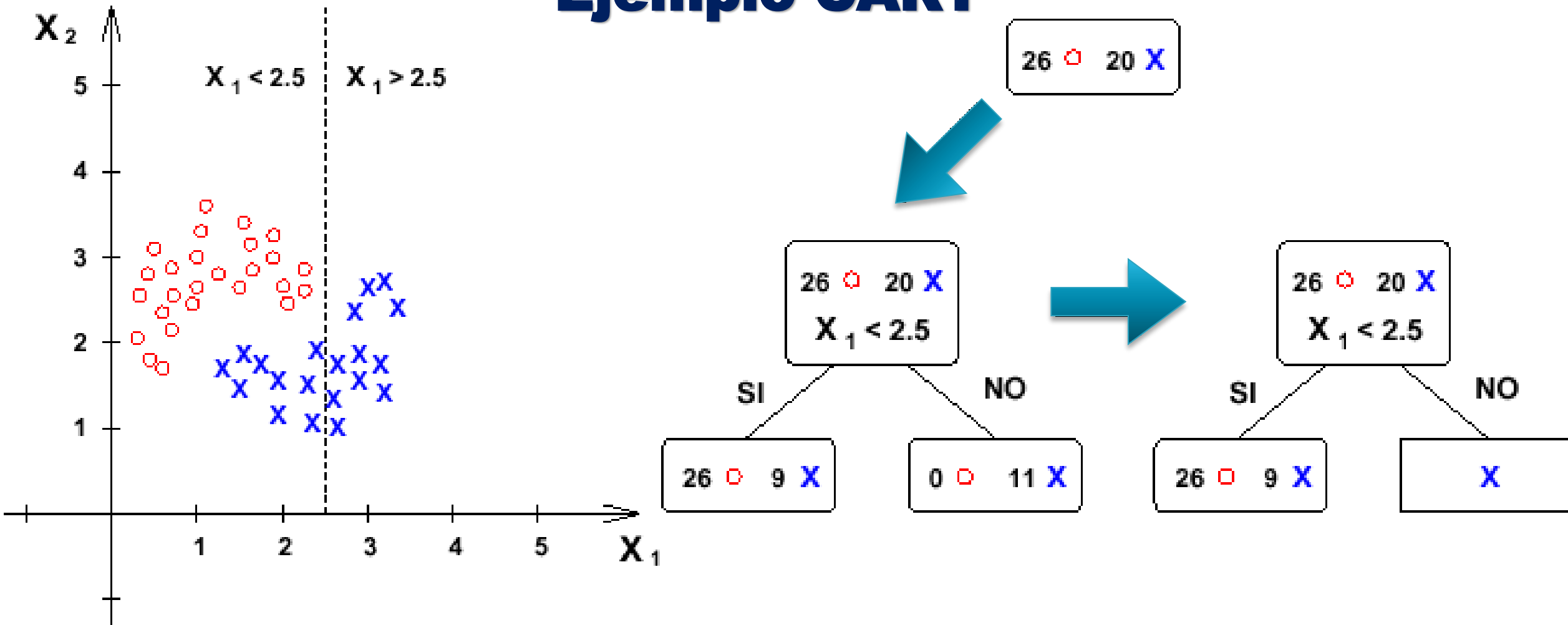
En cada nodo selecciona a la variable indep. que proporciona el mejor desempeño en el criterio para particional a los datos

## Ejemplo CART

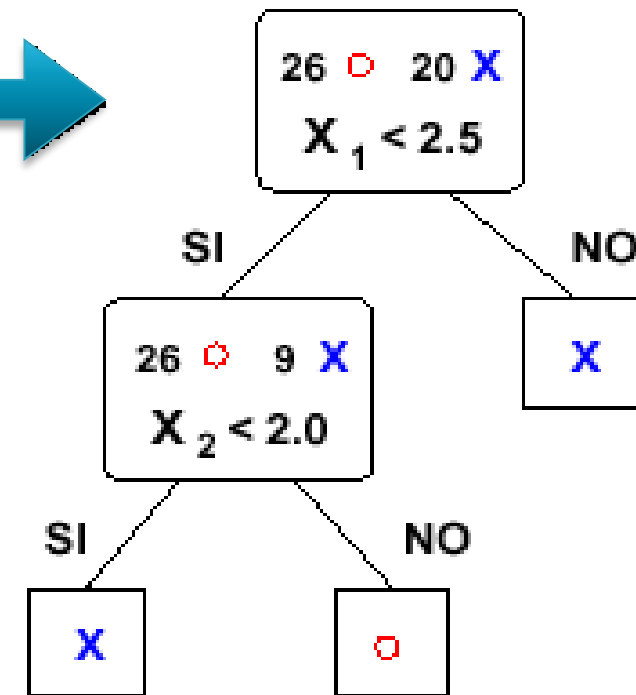
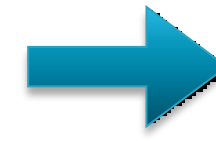
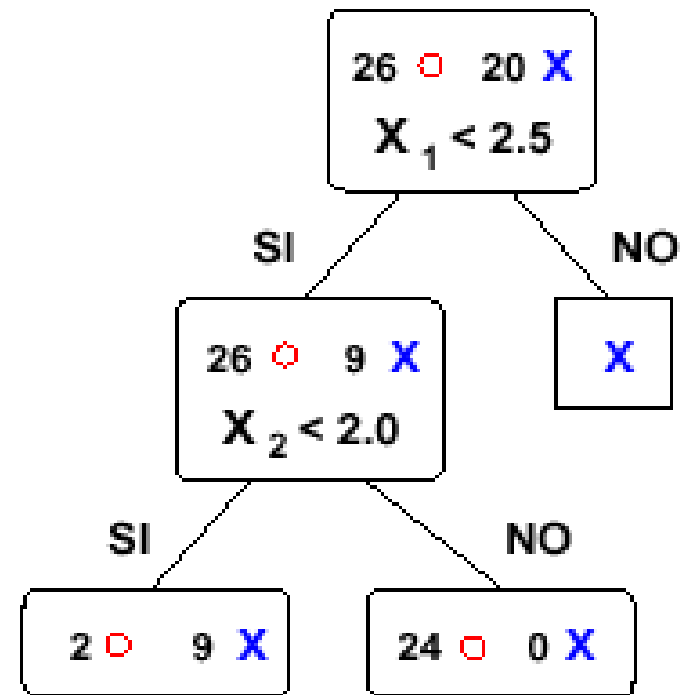
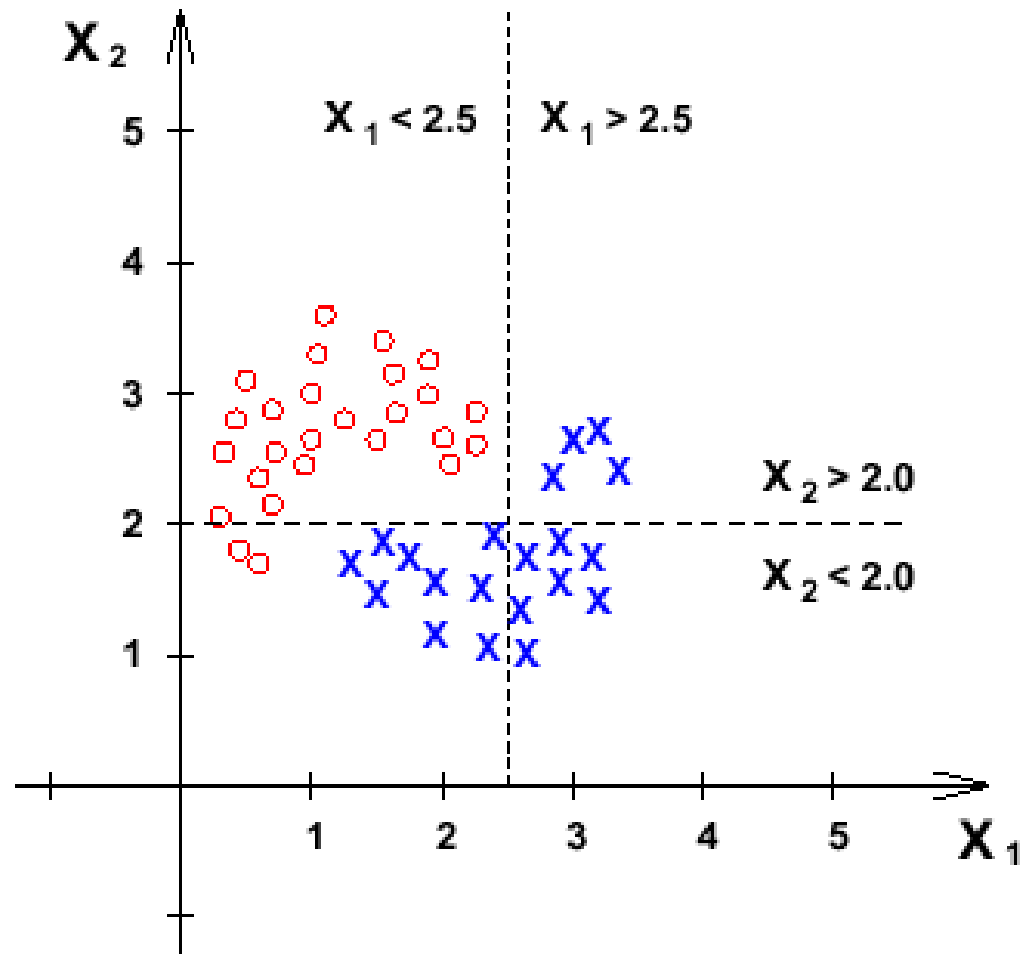
- Se tienen dos variables predictoras  $X_1$  y  $X_2$ .
- Se cuentan con 46 observaciones divididas en dos clases.  $N=46$  ( $N_1=26$  y  $N_2=20$ )



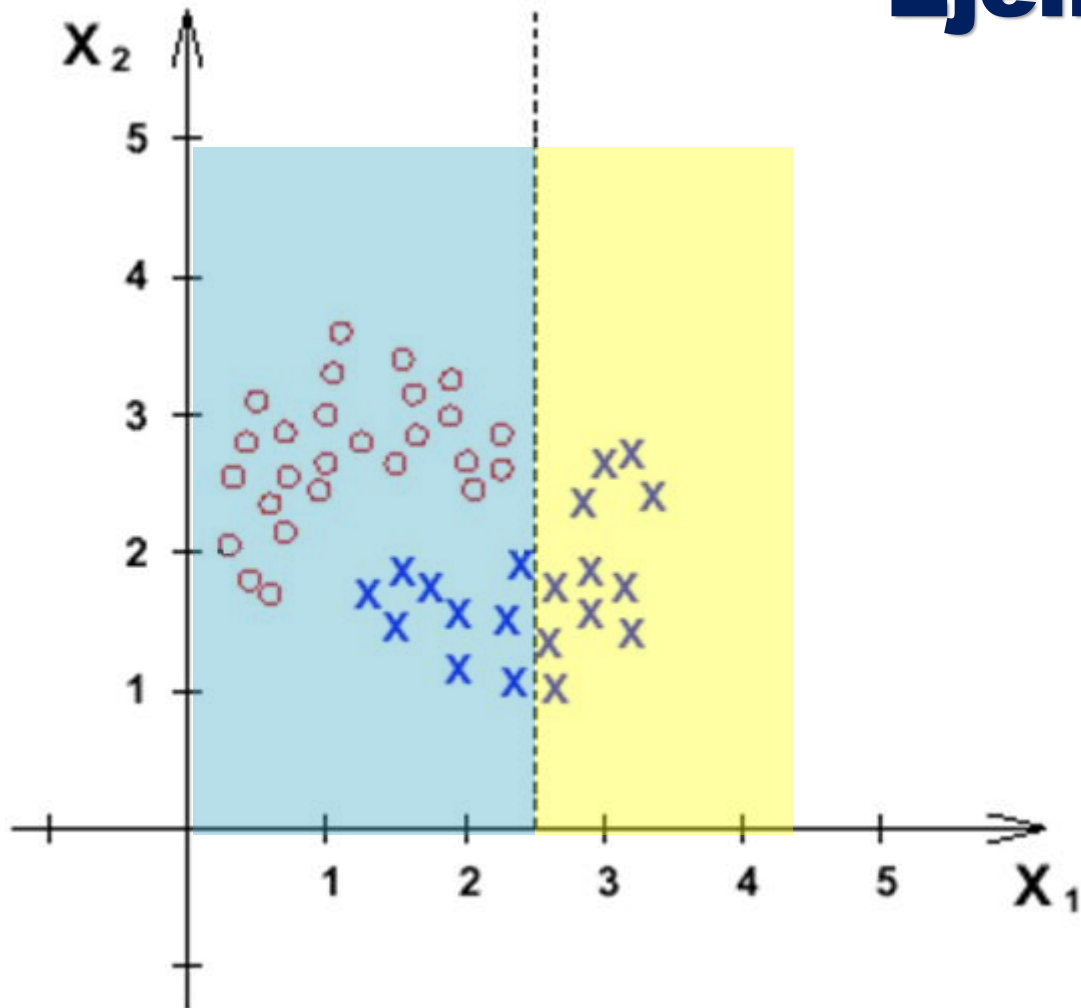
## Ejemplo CART



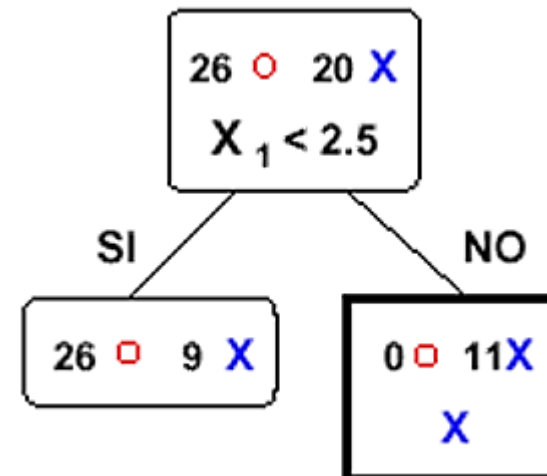
## Ejemplo CART



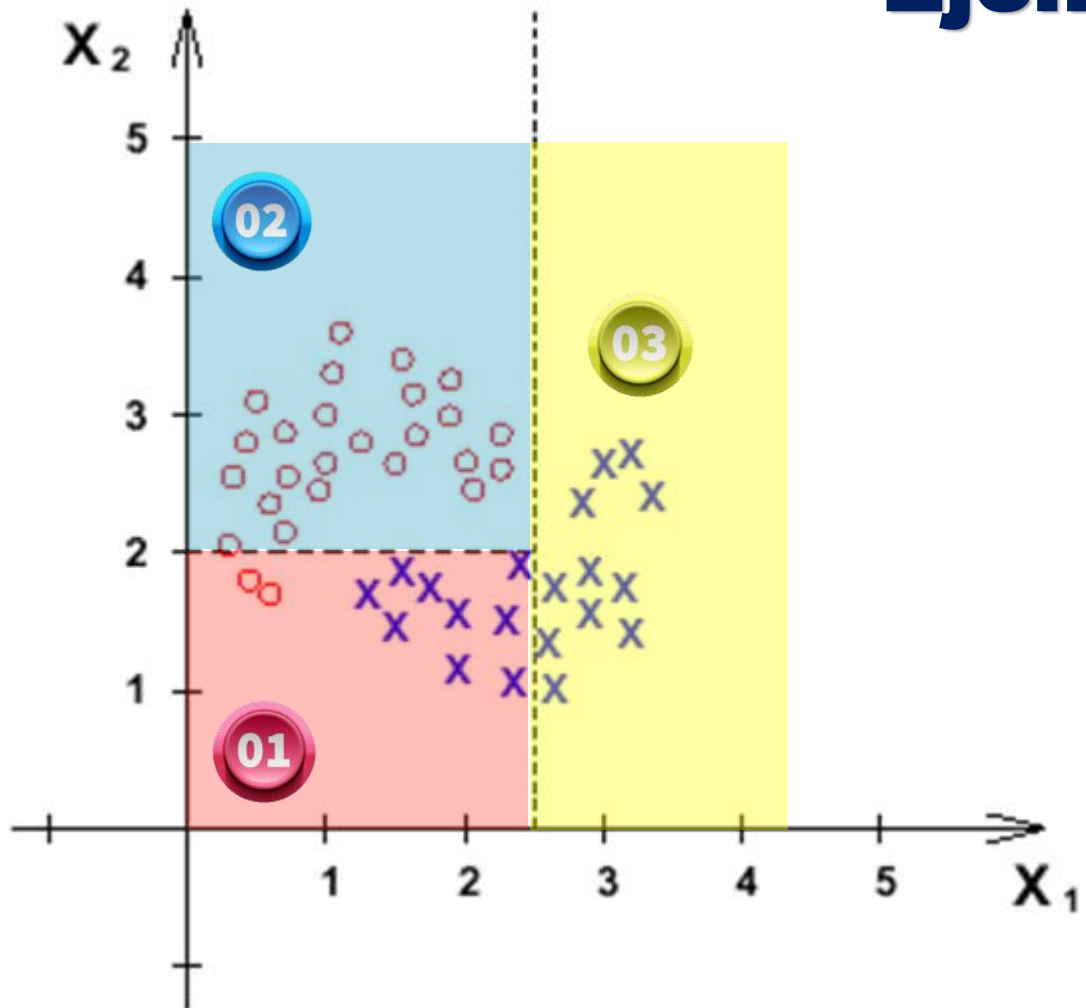
## Ejemplo CART



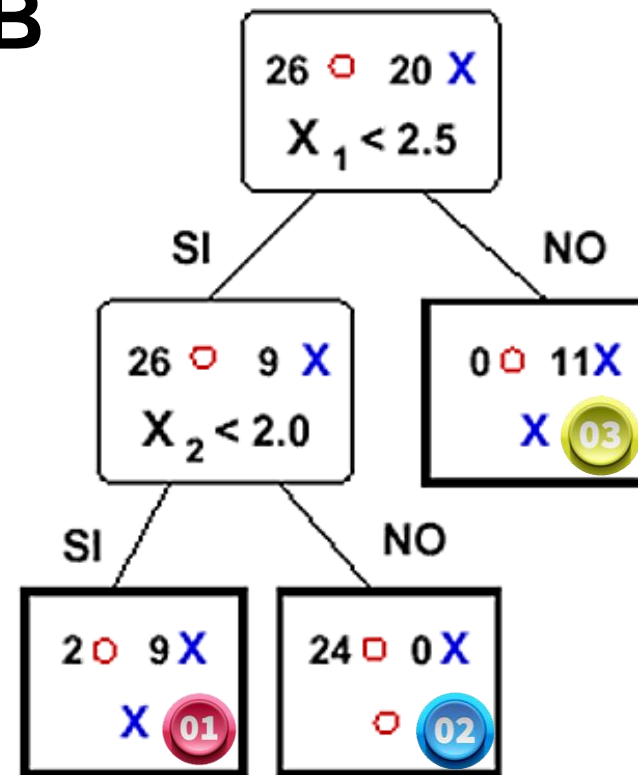
A



## Ejemplo CART

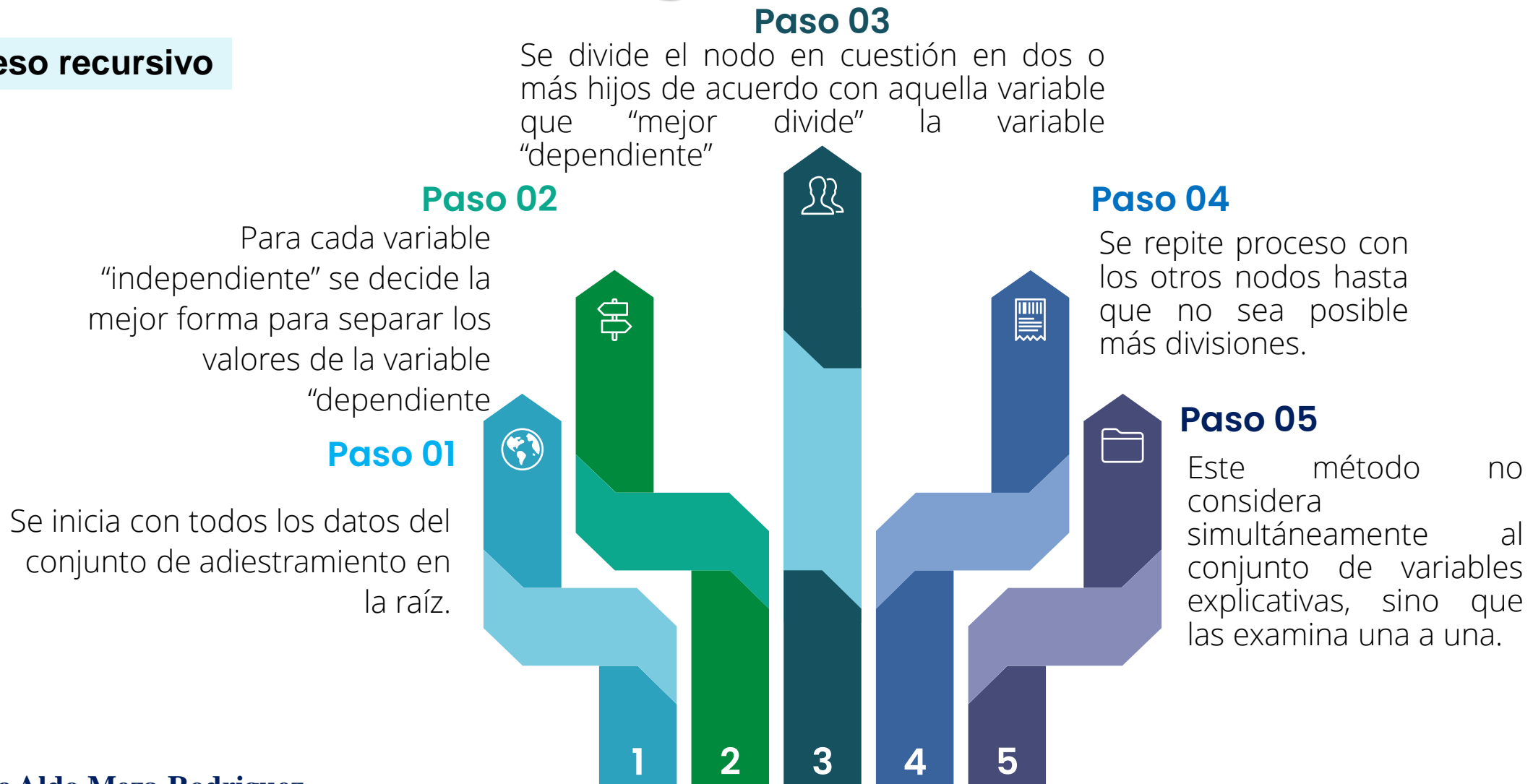


B



## Lógica del árbol

### Proceso recursivo



## **Puntos clave en la construcción de árbol de clasificación**

1. ¿De qué forma se hacen las particiones y se selecciona la mejor de entre las posibles en cada momento?
2. ¿Cual es el criterio para determinar que un nodo es homogéneo? ó ¿Cuando se debe declarar un nodo como terminal, o por el contrario, continuar su división?
3. ¿Cómo asignar una etiqueta a un nodo terminal?



## Formulación de las preguntas

### Variables categóricas

Si  $X_i$  es un atributo **categorico**, que toma valores en  $\{c_1, c_2, \dots, c_L\}$ , se incluyen las preguntas:

$$¿X_i = c ?$$

**Por ejemplo.** Si  $X_2$  toma valores en  $\{A, B, C\}$ ,  
 $¿X_2 = \{A\} ?$ ,  $¿X_2 = \{B\} ?$ ,  $¿X_2 = \{C\} ?$

### Variables continuas

Si  $X_i$  es un atributo **continuo**, se incluyen las preguntas:

$$¿X_i \leq v ?$$

donde  $v$  es valor real, teóricamente cualquiera. En **CART**,  $v$  es el punto medio de dos valores consecutivos de  $X_i$

**Por ejemplo.** Si  $X_1$  es real, con valores 0.1, 0.5, 1.0  
 $¿X_1 \leq (0.1 + 0.5)/2 ?$ ,  $¿X_1 \leq (0.5 + 1.0)/2 ?$

## PODA DEL ÁRBOL



En una primera fase se construye un árbol que contenga cientos de nodos.



En una segunda fase, el árbol es podado, eliminándose las ramas innecesarias hasta dar con el tamaño adecuado del árbol.



Este proceso compara simultáneamente todos los posibles subárboles resultado de podar en diferente grado el árbol original.