

# Decision Trees and ROC

## Seminar

Jesús D. Hernández Londoño<sup>1</sup> & Ramineh Lopez Yasdani <sup>2</sup>

<sup>1</sup>Universidad de Puerto Rico. Departamento de Matemáticas.  
Recinto Mayagüez.

<sup>2</sup>Universidad de Puerto Rico. Departamento de Matemáticas.  
Recinto Mayagüez.

1 de Septiembre de 2020



# Contenido

- 1 Historia
- 2 Definición
- 3 Construcción de un Decision Trees
- 4 Podando un Decision Trees
- 5 Ventajas y Desventajas
- 6 Métrica
- 7 Referencias



# Contenido

- 1 Historia
- 2 Definición
- 3 Construcción de un Decision Trees
- 4 Podando un Decision Trees
- 5 Ventajas y Desventajas
- 6 Métrica
- 7 Referencias



- Originó en las Ciencias Sociales con el trabajo de investigación de Sonquist y Morgan (1964).
- Sonquist, Baker y Morgan (1971) extendió a problemas de clasificación.
- En estadística, Kass (1980), algoritmo recursivo de árbol no binario, CHAID (Chi-square automatic interaction detection).
- Reiman, Friedman, Olshen y Stone (1984), CART (Classification and regression trees), para la construcción de árboles.



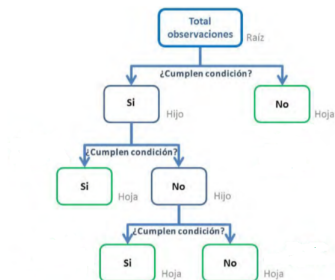
# Contenido

- 1 Historia
- 2 Definición**
- 3 Construcción de un Decision Trees
- 4 Podando un Decision Trees
- 5 Ventajas y Desventajas
- 6 Métrica
- 7 Referencias



## Definición

- Un modelo de Machine Learning para problemas de clasificación y de regresión.
- Es un gráfico que ilustra reglas de decisión partiendo de un nodo raíz que contiene todas las observaciones.



# Contenido

- 1 Historia
- 2 Definición
- 3 Construcción de un Decision Trees**
- 4 Podando un Decision Trees
- 5 Ventajas y Desventajas
- 6 Métrica
- 7 Referencias



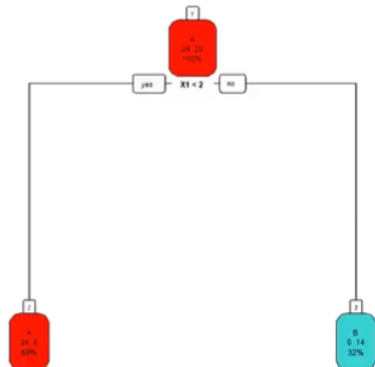
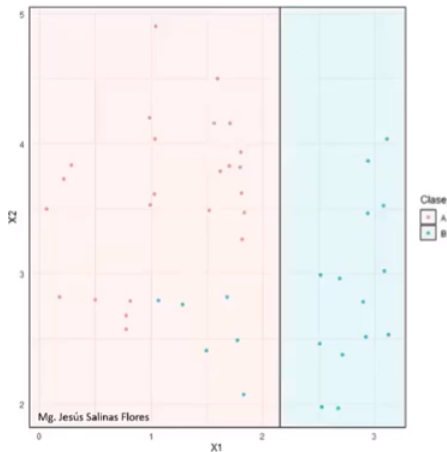
# Construcción de un Decision Trees

- 1 Divide el espacio muestral de las características de los predictores en un subconjunto de hiperrectángulos. En las predictoras numéricas toma los puntos medios de cada observación y en las categóricas los agrupa por el mismo factor (ejemplo: R, T).
- 2 Calcula la suma de los cuadrados del error o medidas de impureza.
- 3 Elige la predictora y característica que minimice el paso anterior.

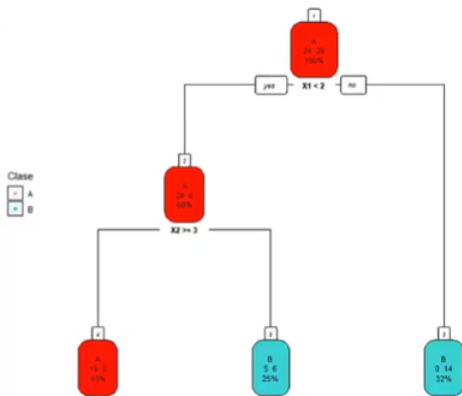
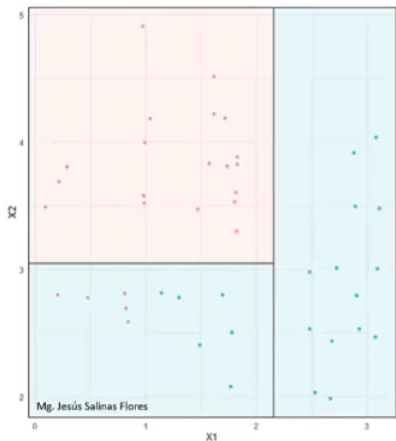




# Construcción de un Decision Trees



# Construcción de un Decision Trees



# Contenido

- 1 Historia
- 2 Definición
- 3 Construcción de un Decision Trees
- 4 Podando un Decision Trees**
- 5 Ventajas y Desventajas
- 6 Métrica
- 7 Referencias



# Podando un Decision Trees

Para elegir un árbol optimo necesitamos minimizar la medida de mérito del árbol  $T$  o medida de costo-complejidad.

$$R_{\alpha}(T) = Resub(T) + \alpha|T|$$

$Resub(T)$ , Tasa de error de clasificación.

$\alpha \geq 0$ , Complexity parameter.

$|T|$ , Número de nodos de  $T$ .

Nota:

- Cuando  $\alpha = 0$  se obtiene el arbol más grande.
- Este proceso nos ayudará a evitar el sobreajuste.



# Contenido

- 1 Historia
- 2 Definición
- 3 Construcción de un Decision Trees
- 4 Podando un Decision Trees
- 5 Ventajas y Desventajas**
- 6 Métrica
- 7 Referencias



- 1 Puede aplicarse a cualquier tipo de variables, cuantitativas o cualitativas.
- 2 No tiene problemas para trabajar con datos missing.
- 3 Es resistente a la presencia de valores atípicos.
- 4 Es un clasificador no paramétrico, esto significa que no requiere suposiciones.



- ❶ Es bastante sensible a pequeñas perturbaciones en los datos, dado que sigue gran parte de la tendencia de los datos. (overfitting).
- ❷ No es fácil elegir el árbol óptimo.
- ❸ Requiere gran número de datos y buena variabilidad para que funcione correctamente.
- ❹ Inestabilidad: un pequeño cambio en los datos puede modificar ampliamente la estructura del árbol.



# Contenido

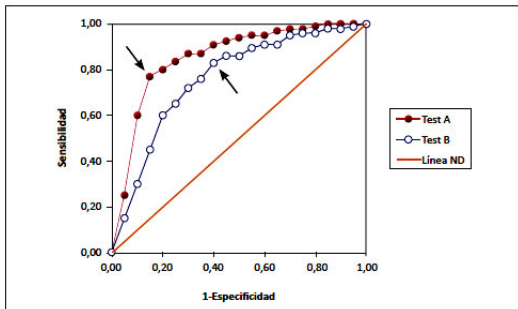
- 1 Historia
- 2 Definición
- 3 Construcción de un Decision Trees
- 4 Podando un Decision Trees
- 5 Ventajas y Desventajas
- 6 Métrica**
- 7 Referencias



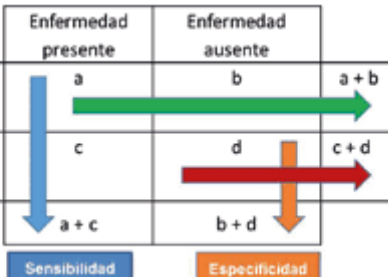


# ROC

La curva ROC es el punto de corte en el que se alcanza la sensibilidad y especificidad más alta. Evalúa la capacidad del modelo para clasificar correctamente. Un valor de 1 significa que el modelo es perfecto, un valor de 0.5 indica que el modelo no es útil.



	Enfermedad presente	Enfermedad ausente	
Prueba positiva	a	b	a + b
Prueba negativa	c	d	c + d
	a + c	b + d	



**Sensibilidad** (blue box)      **Especificidad** (orange box)

**VPP** (green box)      **VPN** (red box)



El Chinese Mini Mental Status Test (CMMS) es una prueba de 114 reactivos que pretende identificar personas con Alzheimer y demencia senil en China.

CMMS score	No Demencia	Demencia
0-5	0	2
6-10	0	1
11-15	3	4
16-20	9	5
21-25	16	3
26-30	18	1
<b>total</b>	<b>46</b>	<b>16</b>

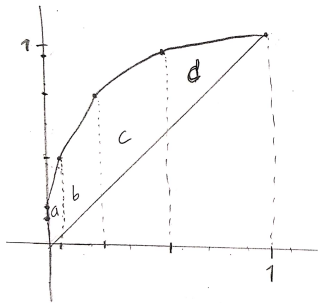


	$x$	$y$
Cohorte	1 - especificidad	Sensibilidad
$\leq 5$	$1 - 1 = 0$	$12/16 = 0.125$
$\leq 10$	$1 - 1 = 0$	$3/16 = 0.1875$
$\leq 15$	$1 - \frac{43}{46} = \frac{3}{46} = 0.065$	$7/16 = 0.4375$
$\leq 20$	$1 - \frac{34}{46} = \frac{6}{23} = 0.26$	$12/16 = 0.75$
$\leq 25$	$1 - \frac{18}{46} = \frac{14}{23} = 0.60$	$15/16 = 0.93$
$\leq 30$	$1 - 0 = 1$	1



# AUC

Es el área bajo la curva, Esta refleja que tan bueno es la prueba o modelo para discriminar pacientes con y sin la enfermedad.



$AUC=0.86$

Nota: Esta métrica se usa para datos desbalanceados.



# Contenido

- 1 Historia
- 2 Definición
- 3 Construcción de un Decision Trees
- 4 Podando un Decision Trees
- 5 Ventajas y Desventajas
- 6 Métrica
- 7 Referencias**



 Tibshirani R., Hastie T., Witten D., James G.

*An Introduction to Statistical Learning with applications in R.*

 Acuña E.

*Data Mining and Machine Learning.*

Decision Trees

 Árboles de decisión y Random Forest.

<https://bookdown.org/content/2031/>.

Orellana J.

