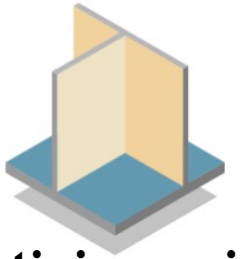
APRENDIZAJE SUPERVISADO

Anibal Sosa, PhD

UNIVERSIDAD ICESI

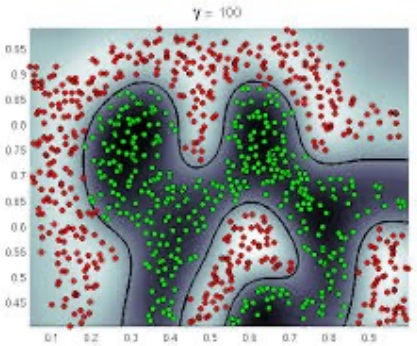
AGENDA



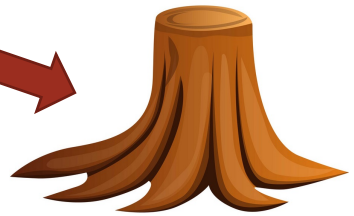
Particionamiento

Criterios de particionamiento

Modelos de ensamble



Aprendizaje supervisado



Decision Stump

Árbol de decisión

Poda

UNIVERSIDAD ICESI

TALLER: PROSPECCIÓN DE CLIENTES

Una compañía de seguros quiere contactar los mejores clientes potenciales de una base de datos que acaban de adquirir con **10.000** personas, para ofrecerles un plan. Cuentan con la información de campañas anteriores incluyendo diferentes características como edad, género y salarios, así como la indicación de si la oferta fue exitosa o no.

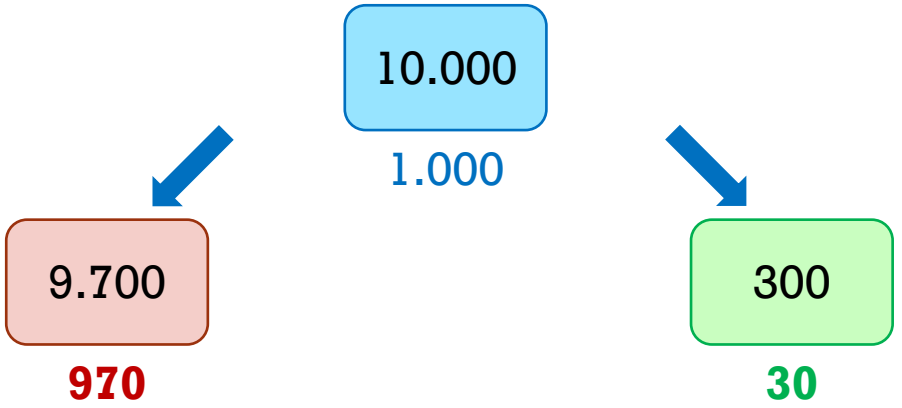
Teniendo en cuenta el costo del paquete de publicidad por correo, solo pueden contactar **1.000** clientes potenciales.

Sabemos de las campañas anteriores, que solo el **3%** de las personas contactadas adquirían el plan, pero esta tasa varía considerablemente si empezamos a considerar sub poblaciones con características particulares (edad, salario, etc).

UNIVERSIDAD ICESI

CASO DE ESTUDIO: PROSPECCIÓN DE CLIENTES

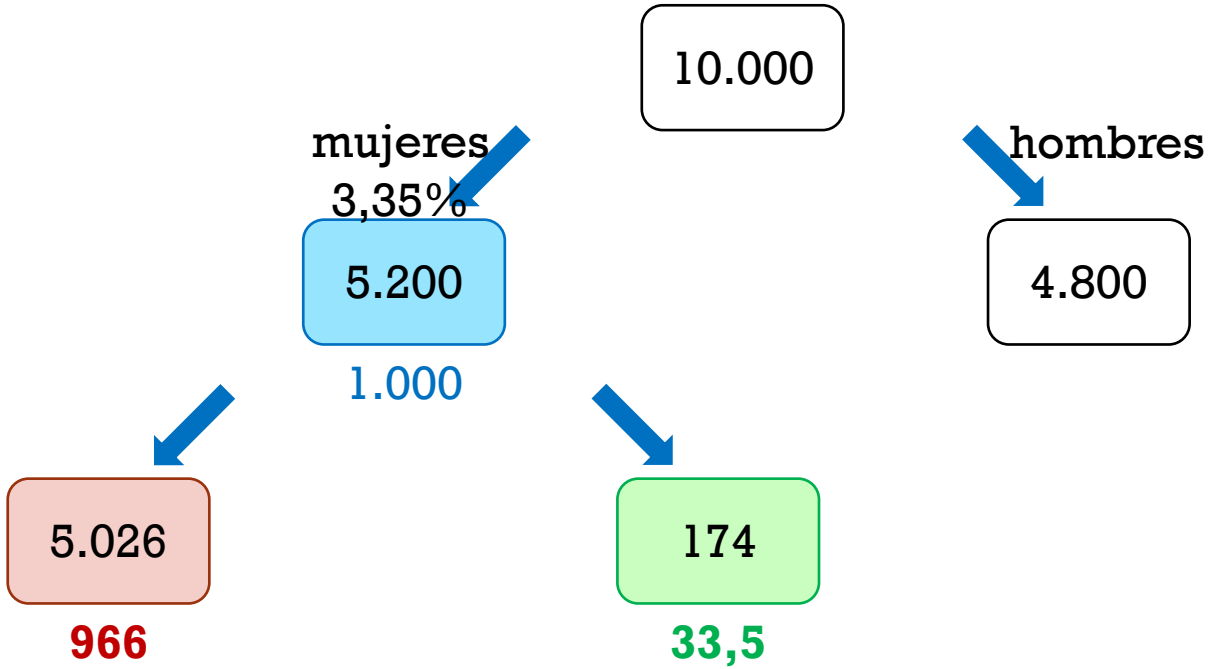
Escoja 10.000 clientes potenciales aleatoriamente



éxito: 3% (30)

CASO DE ESTUDIO: PROSPECCIÓN DE CLIENTES

Las mujeres son más propensas a comprar seguros (3,35%) y hay 5.200 mujeres en la BD

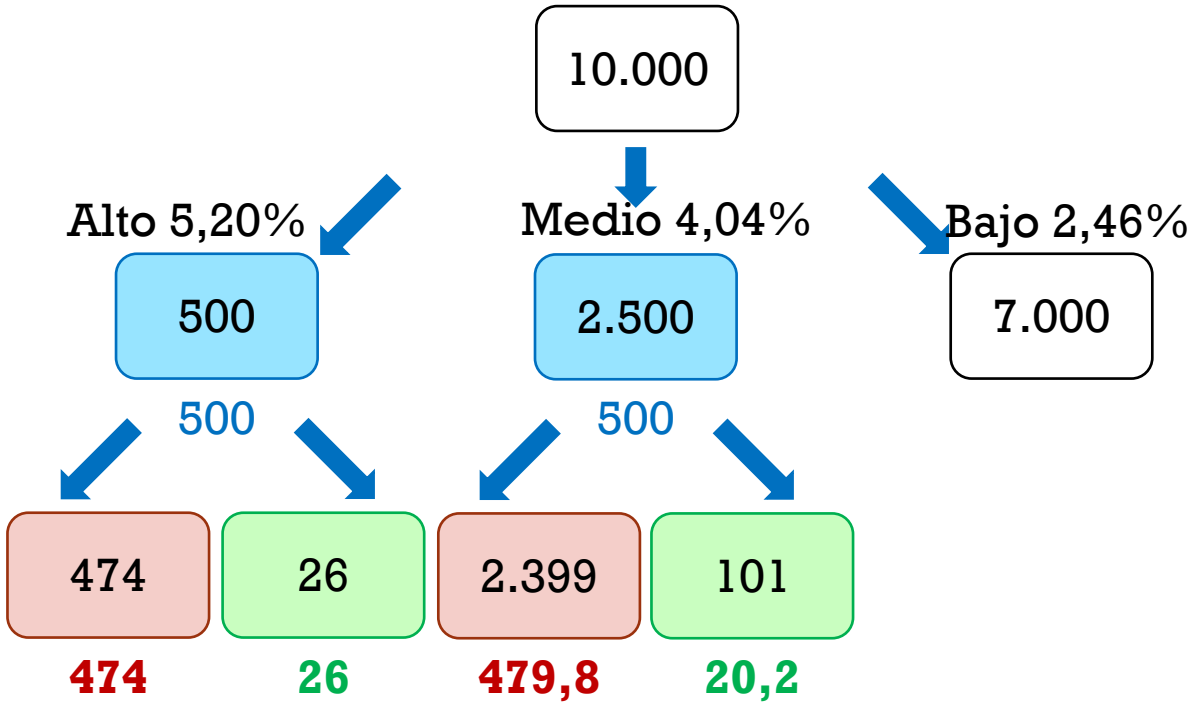


éxito: 3,35% (33,5)

CASO DE ESTUDIO: PROSPECCIÓN DE CLIENTES

Las tasas de éxito por tipo de salario son las siguientes:

- Alto (500 en el grupo): 5,20%
- Medio (2500 en el grupo): 4,04%
- Bajo (7000 en el grupo): 2,46%



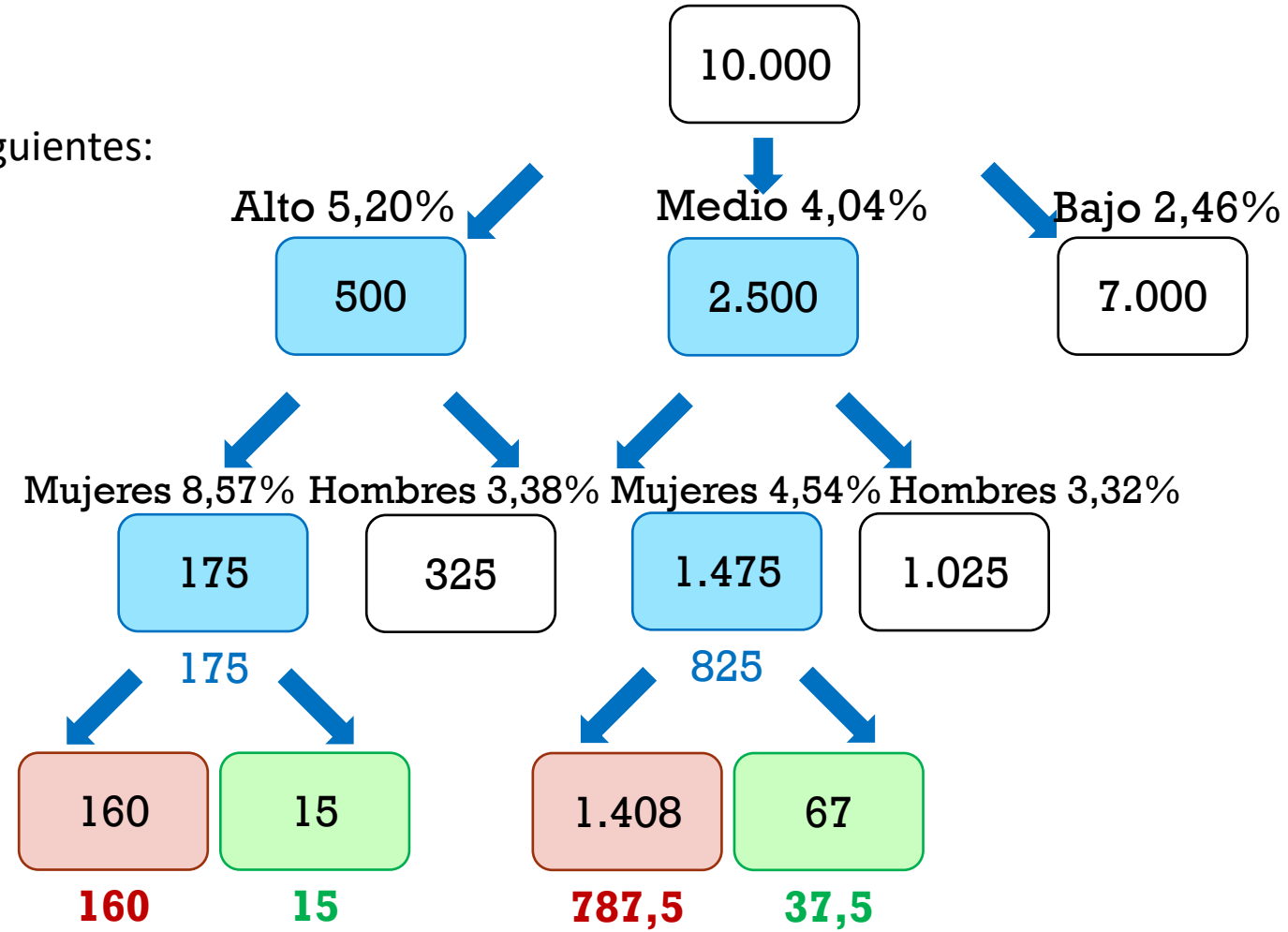
éxito: 4,62% (46,2)

CASO DE ESTUDIO: PROSPECCIÓN DE CLIENTES

Las tasas de éxito por tipo de salario son las siguientes:

- Alto (500 en el grupo): 5,20%
- Medio (2500 en el grupo): 4,04%
- Bajo (7000 en el grupo): 2,46%

- Mujeres con salario alto (175): 8,57%
- Hombre con salario alto (325): 3,38%
- Mujeres con salario medio (1475): 4,54%
- Hombres con salario medio (1025): 3,32%



¿Cómo hago esto de una manera más inteligente?

éxito: 5,25% (52,5)

UNIVERSIDAD ICESI

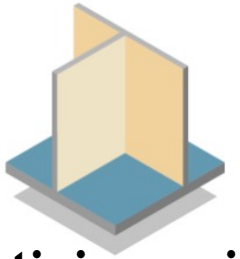
TALLER DE PARTICIONAMIENTO CATEGÓRICO: PREGUNTAS PARA ADIVINAR ENTIDADES

- Dinámica: Identificar una entidad u objeto con el menor número de preguntas posibles con respuestas binarias (árbol de preguntas con la menor profundidad).
  - → Particionamiento de atributos **binarios**
  - → Reconocimiento de la pertinencia de los atributos en la clasificación

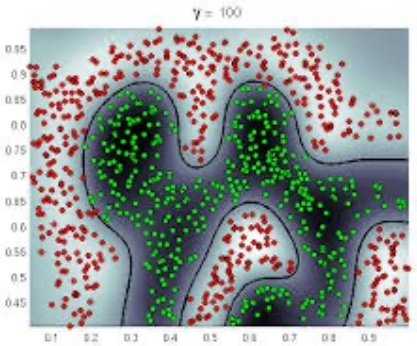
Everything, everywhere all at once	Python Data Science Handbook
Nicolas Maduro	Buga
Crónica de una muerte anunciada	Moneyball
Fito Paez	Michael Jackson
Ferrari (equipo de F1)	Millonarios
Lionel Messi	Ivan Cepeda
Real Madrid	Freakonomics
París	Freddie Mercury

UNIVERSIDAD ICESI

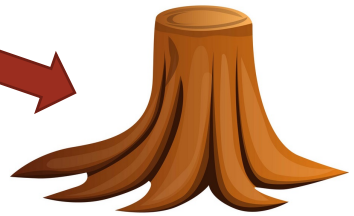
AGENDA



Particionamiento



Aprendizaje supervisado



Decision Stump

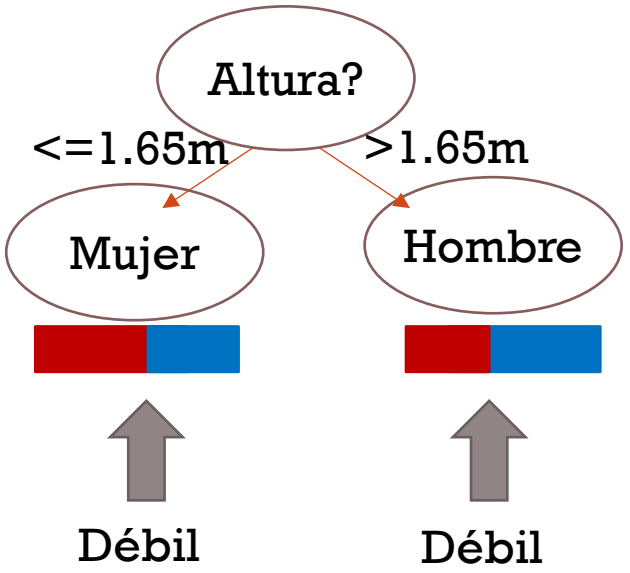
UNIVERSIDAD ICESI

DECISION STUMP

UNIVERSIDAD ICESI

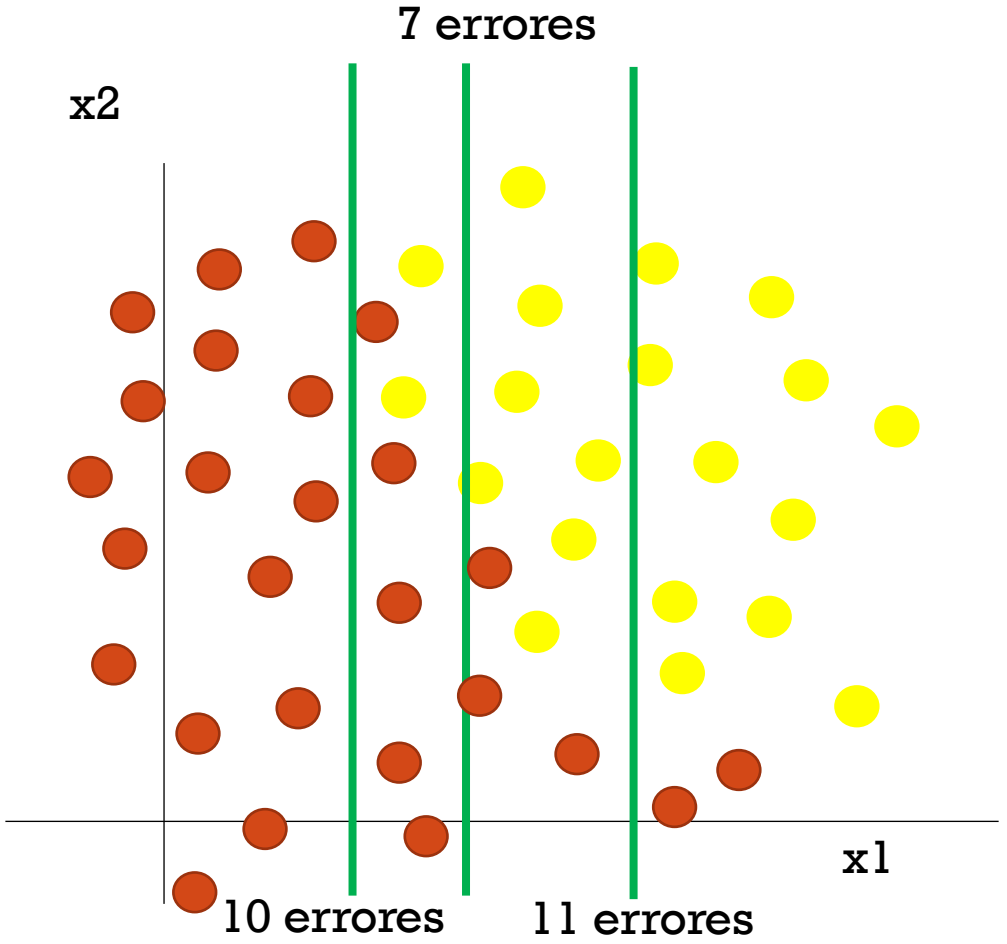
DECISION STUMP

- Busca el mejor particionamiento considerando una **sola variable** predictiva
- Árbol de decisión de un solo nivel
- Es un "**very weak learner**", que produce una sola regla de decisión. Por ejemplo:
  - Las personas que miden más de 1.65 metros son hombres, y las que no, son mujeres
- Muy utilizado en modelos de ensamble (sobre todo **Boosting**)



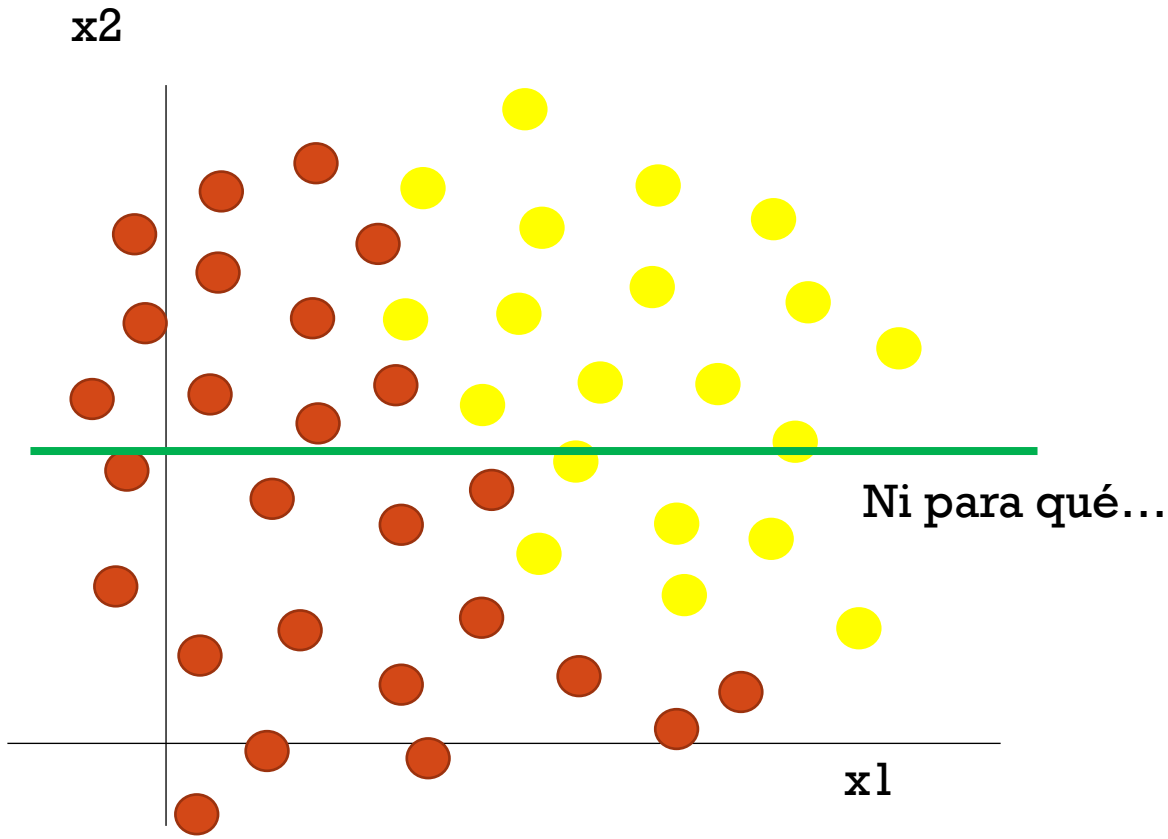
DECISION STUMP

- El particionamiento en las variables numéricas solo se puede realizar de manera perpendicular a los ejes
  - Se busca minimizar el error de clasificación
  - Es necesario buscar todos los particionamientos posibles en todas las variables predictivas
- ¿Cómo serían las reglas y el número de errores de los clasificadores siguientes?



DECISION STUMP

- El particionamiento en las variables numéricas solo se puede realizar de manera perpendicular a los ejes
  - Se busca minimizar el error de clasificación
  - Es necesario buscar todos los particionamientos posibles en todas las variables predictivas
- ¿Cómo serían las reglas y el número de errores de los clasificadores siguientes?



UNIVERSIDAD ICESI

DECISION STUMP

- Las variables numéricas deben ser discretizadas
- Hay varias maneras de realizar el análisis del mejor punto de corte, utilizando diferentes métricás:
  - Entropía condicional
  - Gini
  - CHAID
- Más adelante haremos un taller respecto estas métricas

¿Cuál particionamiento es mejor entre p1 y p2?

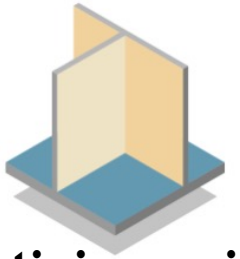
humidity	play (X)	p1	p2
54	yes	a	a
58	no	a	a
59	yes	a	a
60	yes	a	a
60	yes	a	a
62	yes	a	a
63	yes	b	a
80	yes	b	a
81	yes	b	a
89	no	b	b
90	no	b	b
90	no	b	b
90	no	b	b
92	yes	b	b

TALLER DE EVALUACIÓN DE UN MODELO DE CLASIFICACIÓN (ÁRBOLES)

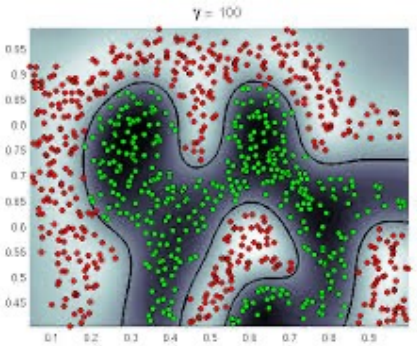
- DATASET: base de datos de 20000 clientes que han cancelado (churn) o no los servicios de una compañía. La idea es poder predecir en un futuro quiénes son los clientes más propensos a hacer churn, para poder desarrollar campañas que lo prevengan.
- Encontrar particionamientos que permitan mejorar la tasa de correctitud del baseline
- 06-02-EXCEL-ParticionamientoChurn-STUD.xlsx

UNIVERSIDAD ICESI

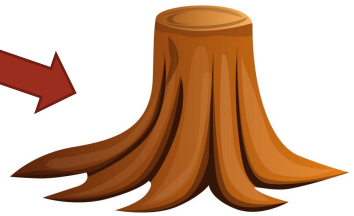
AGENDA



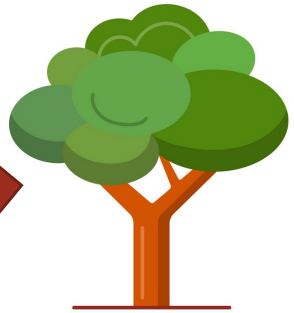
Particionamiento



Aprendizaje supervisado



Decision Stump



Árbol de decisión

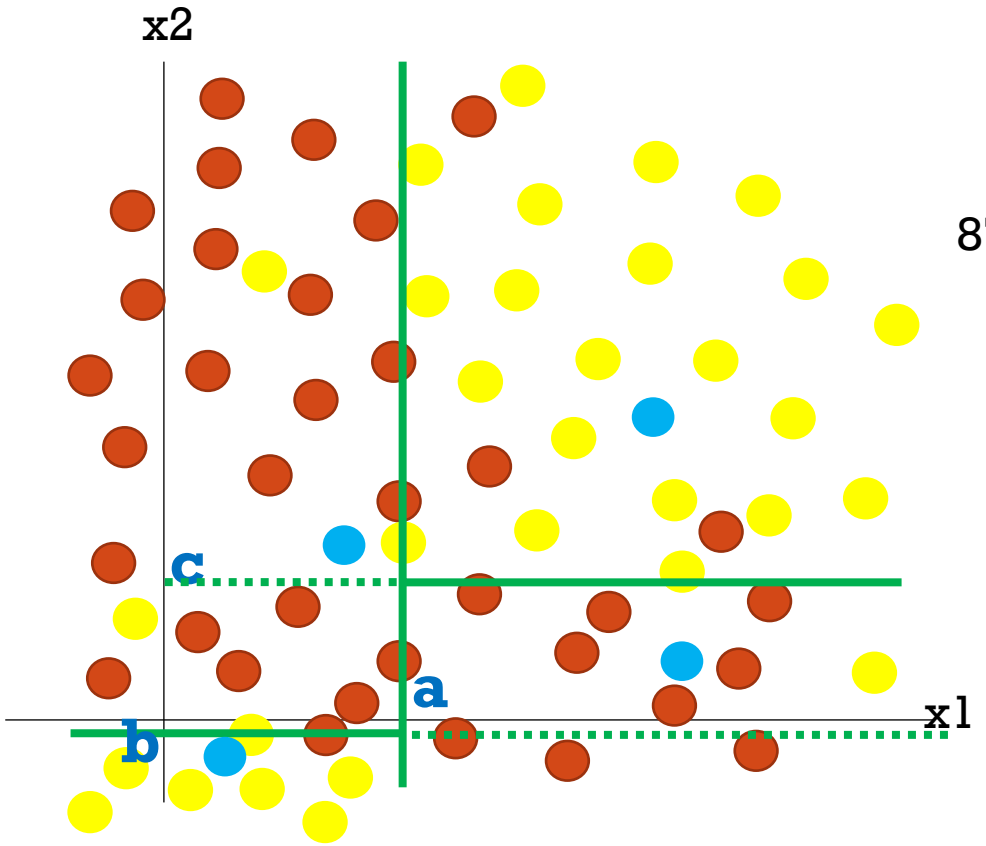
UNIVERSIDAD ICESI

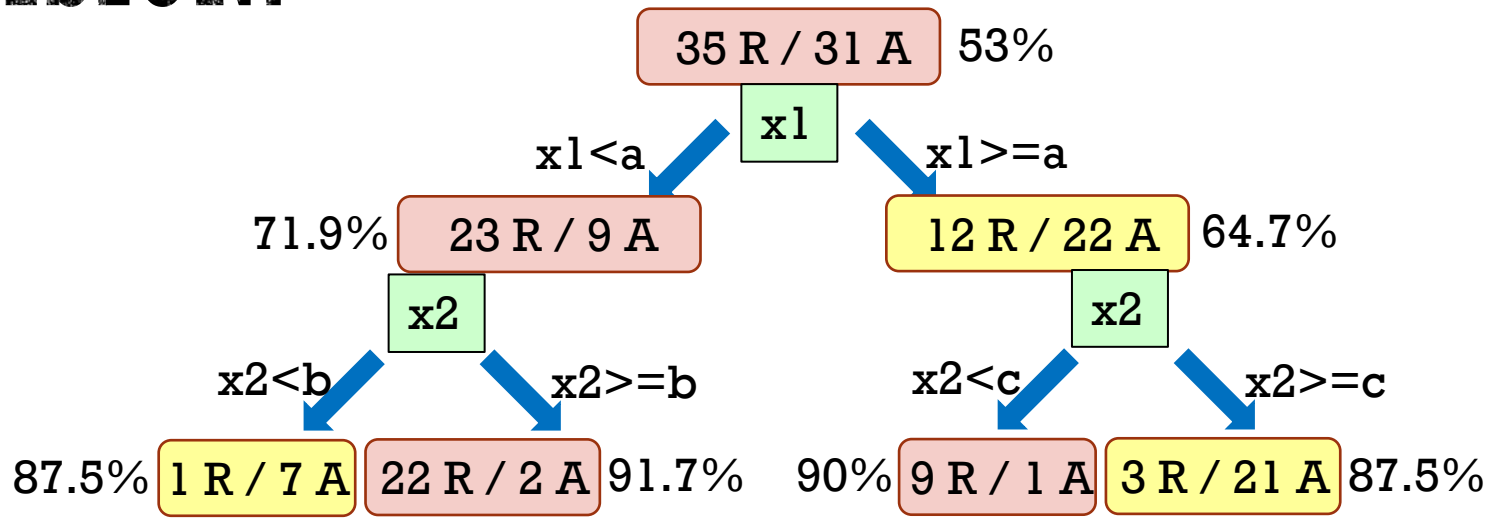
ÁRBOLES DE DECISIÓN: ALGORITMO

Dividir & conquistar: se divide de manera incremental el espacio en regiones no sobrelapadas, que constituyen los nodos del árbol:

- **Seleccionar un factor** que separe **óptimamente** los valores objetivo del nodo actual, crear una rama por cada valor minimizando una función de impureza del nodo en cuestión
- **Dividir** el conjunto de datos del nodo con respecto a los valores del factor seleccionado y crear los nodos correspondientes
- **Repetir recursivamente** hasta que
- todas las instancias de los nodos hoja sean de la misma clase
  - no existan más atributos por los cuales particionar
  - se llegue a un criterio de parada definido (pre-poda)

ÁRBOLES DE DECISIÓN: CLASIFICACIÓN



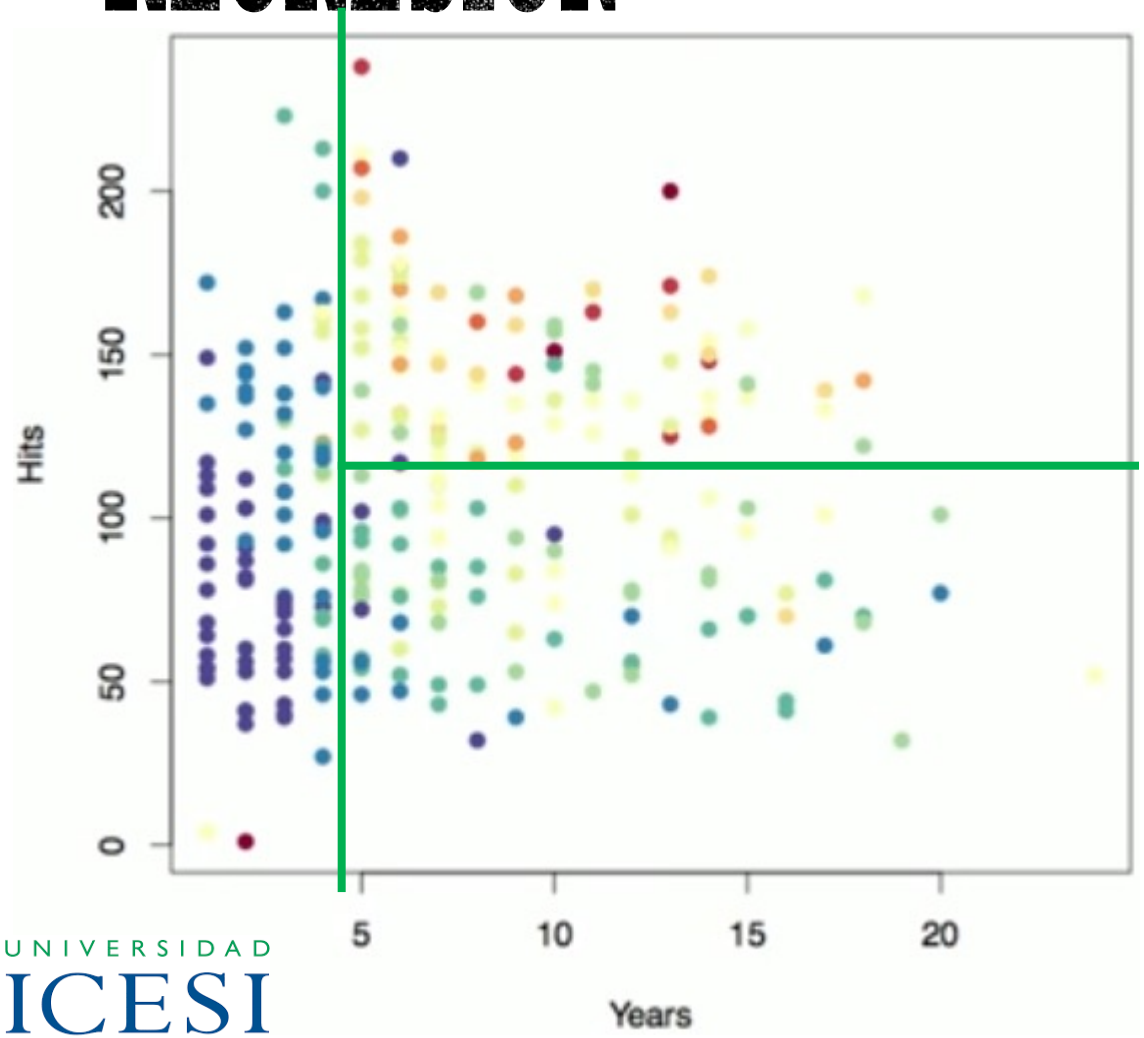


Paso	Accuracy
Raíz	35/66 = 53%
1era partición	45/66 = 68.2%
2a partición (rama izq.)	51/66 = 77.3%
3a partición (rama der.)	59/66 = 89.4%

Se minimiza localmente una función de costo que considera la **impureza** de los nodos terminales del árbol

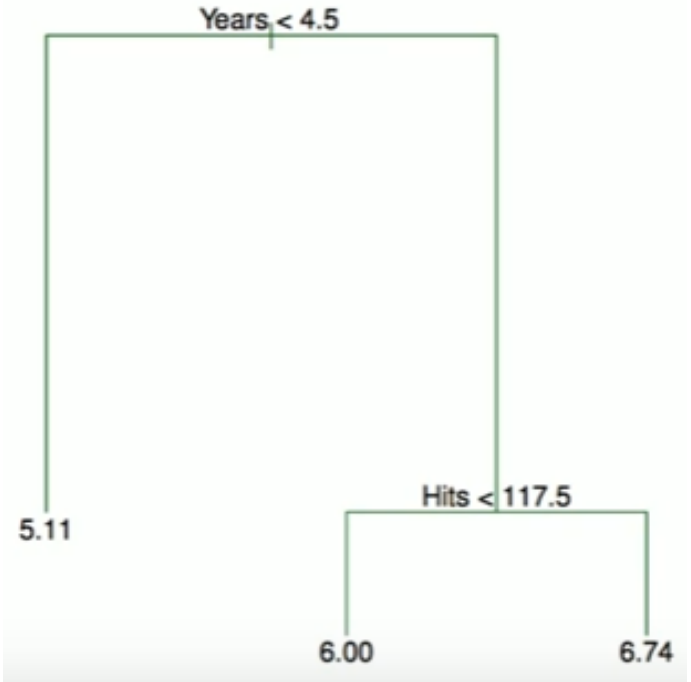
UNIVERSIDAD ICESI

ÁRBOLES DE DECISIÓN: REGRESIÓN



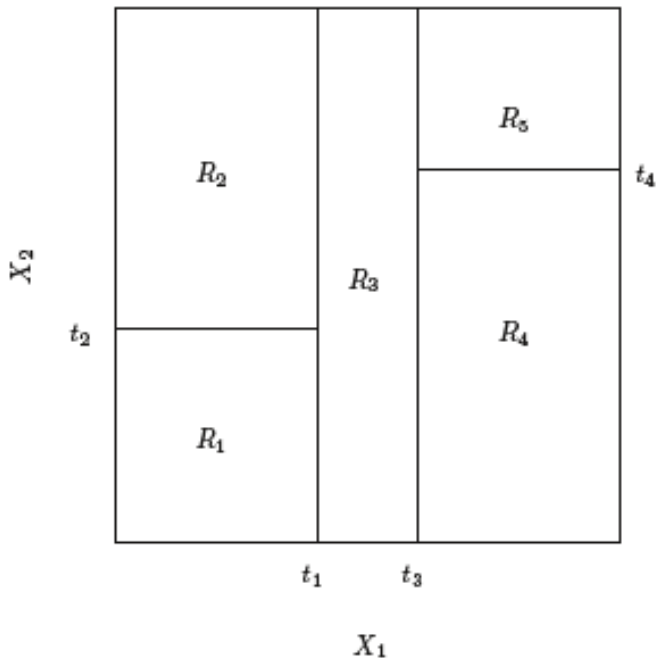
Evolución de lo salarios de beisbolistas (color) con respecto a años de experiencia (abscisa) y número de bateos exitosos (ordenada).

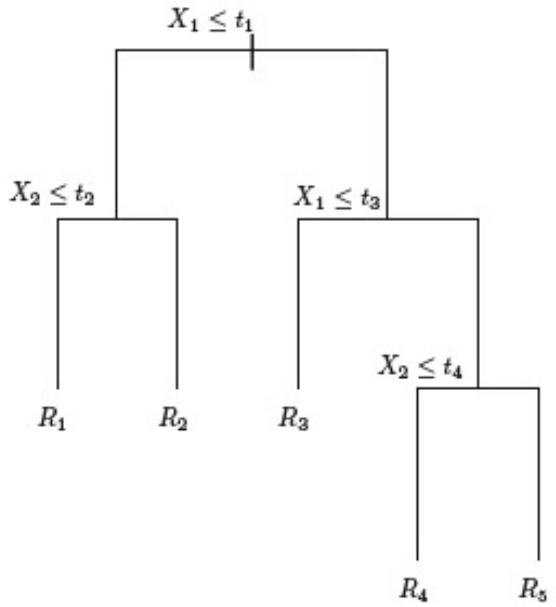
Se minimiza localmente $\sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - \hat{y}_{R_j})^2$

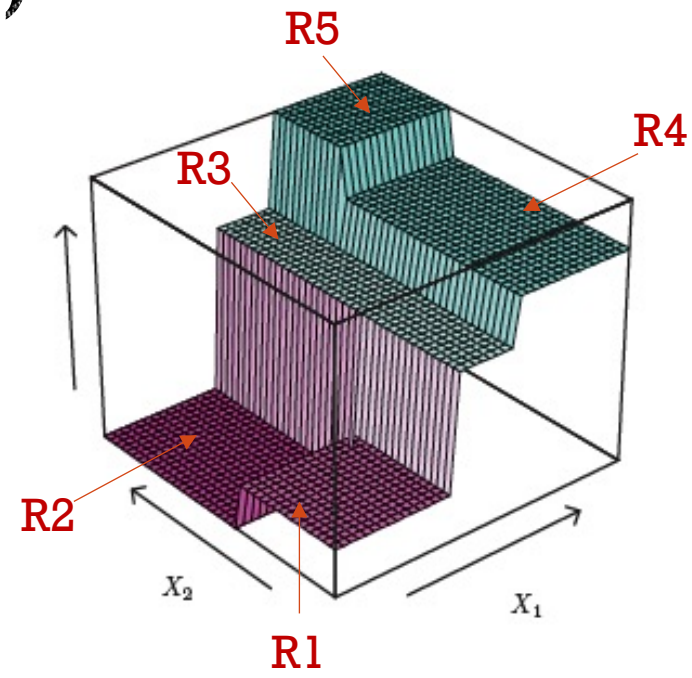


ISLR, 2013

ÁRBOLES DE DECISIÓN: REGRESIÓN (REPRESENTACIÓN)







ISLR, 2013

UNIVERSIDAD ICESI

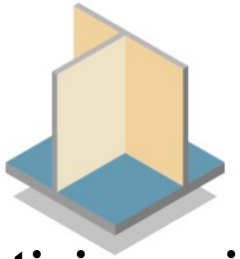
ÁRBOLES DE DECISIÓN

- **Aprendizaje inductivo**: generalización
- Algoritmo **greedy**: busca óptimos locales a cada etapa, que no son necesariamente los óptimos globales
- **Simple** de comprender, implementar y explotar
- Puede ser usado para **clasificación** y **regresión**
- Clasificador **no lineal** (considera interacciones entre los factores)
- Mejor **performance** en contextos no lineales
- Tamaño variable, **escalable** (BIG DATA)

ÁRBOLES DE DECISIÓN

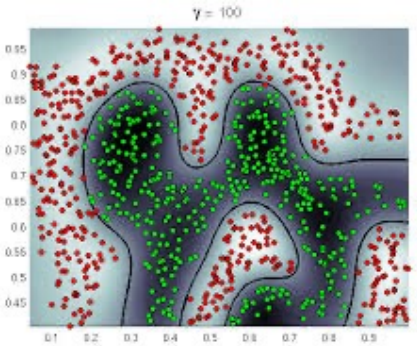
- Los datos deben ser **categóricos.** Variables continuas deben ser **discretizadas**.
- Un árbol de decisión se puede representar como un conjunto de **reglas** booleanas
- Una nueva instancia puede ser clasificada **siguiendo las ramas** del árbol
- Ideal para los casos en que un pequeño número de atributos provee una gran cantidad de la información
- Prueba diferentes atributos categóricos para aprender una clase, hace una selección automática de variables importantes.
- No se basa en ninguna noción de distancia, el modelo es **indiferente** a **escalas**

AGENDA

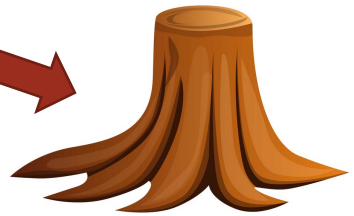


Particionamiento

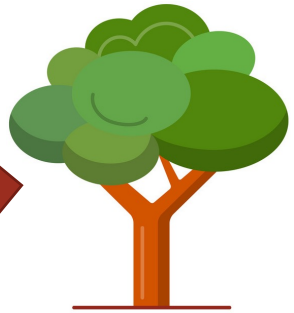
Criterios de particionamiento



Aprendizaje supervisado



Decision Stump



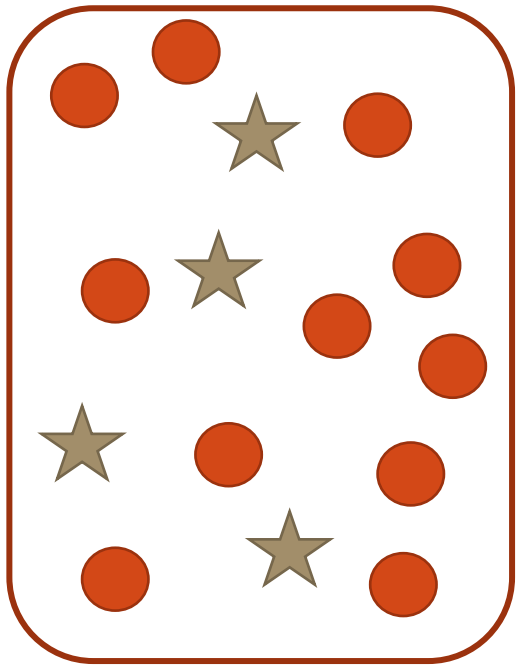
Árbol de decisión

UNIVERSIDAD ICESI

ÁRBOLES DE DECISIÓN

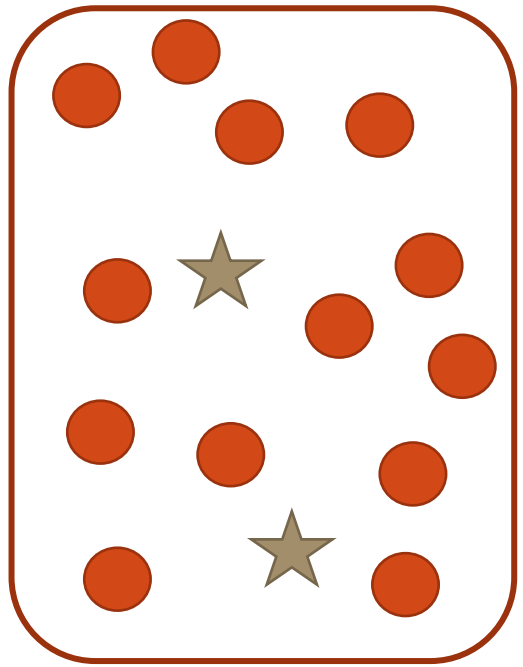
- **Existen diferentes criterios para determinar el mejor atributo de particionamiento en cada nodo → :**
  - **CART** (Classification and Regression Trees). Sólo particiones binarias, usando la métrica de impureza Gini para la clasificación y la reducción de varianza para la regresión
  - **ID3** (Iterative Dichotomizer). Basado en ganancia de información y entropía como criterio de división
  - **C4.5** Extensión de ID3, basado en la razón de ganancia de información. Considera atributos continuos y discretos, información faltante, diferentes costos de clasificación y poda
  - **CHAID** (Chi-squared Automatic Interaction Detector). Utiliza la métrica Chi cuadrado para la clasificación y pruebas F para la regresión
  - …

ÁRBOLES DE DECISIÓN: ID3



¿Cuál conjunto de datos presenta mayor desorden?

- →Entropia, como medida de desorden
- →Búsqueda de particiones cada vez mas puras
- →Reducción del desorden - Ganancia de información



UNIVERSIDAD ICESI

ÁRBOLES DE DECISIÓN: ID3

Utiliza métricas de la **teoría de información**

- Seleccionar el atributo que más reduce el desorden en la variable objetivo del dataset
  - Entropía:

$$H(Y) = -\sum_i p(Y = y_i) * log_2\big(p(Y = y_i)\big)$$

$$H(Y) = 0, si\ no\ hay\ erroers\ de\ clasificación$$

  - Ent.Cond.

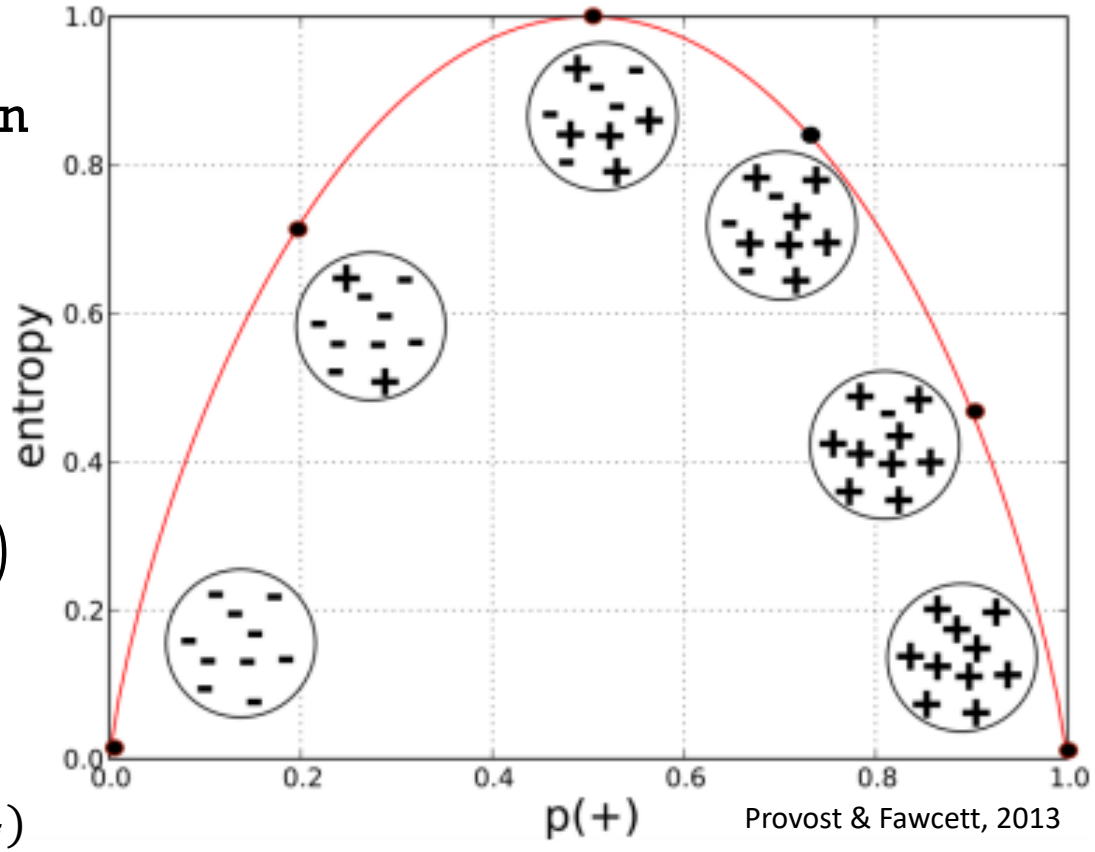
$$H\big(Y|X = x_j\big) = -\sum_i p\big(Y = y_i|X = x_j\big) * log_2\left(p\big(Y = y_i|X = x_j\big)\right)$$

  - Ent.Cond.Prom.

$$H(Y|X) = \sum_j p\big(X = x_j\big) * H(Y|X = x_j)$$

  - Ganancia de información

$$Gain\big(Y,\ X = x_j\big) = H(Y) - \sum_j p\big(X = x_j\big) * H(Y|X = x_j)$$



Provost & Fawcett, 2013

UNIVERSIDAD ICESI

TALLER: ÁRBOLES DE DECISIÓN ID3

	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Dataset de clima: 14 instancias, 4 variables independientes para predecir una clase con 2 categorías posibles

¿Cuál es el mejor atributo para particionar el dataset?

1. Calcular la entropía de la clase (“play”)

Play (Y)				H
p(Y=no)	35.7%	-p(Y=no) log p(Y=no)	0.53	
p(Y=yes)	64.3%	-p(Y=yes) log p(Y=yes)	0.41	0.940

2. Calcular la entropía condicional para cada atributo y su ganancia de información

Outlook								GAIN
p(sunny)	35.7%	p(yes|sunny)	40.0%	p(no|sunny)	60.0%	0.971		
p(overcast)	28.6%	p(yes|overcast)	100.0%	p(no|overcast)	0.0%	0.000	0.694	0.247
p(rainy)	35.7%	p(yes|rainy)	60.0%	p(no|rainy)	40.0%	0.971		

Temperature								GAIN
p(hot)	28.6%	p(yes|hot)	50.0%	p(no|hot)	50.0%	1.000		
p(mild)	42.9%	p(yes|mild)	66.7%	p(no|mild)	33.3%	0.918	0.911	0.029
p(cool)	28.6%	p(yes|cool)	75.0%	p(no|cool)	25.0%	0.811		

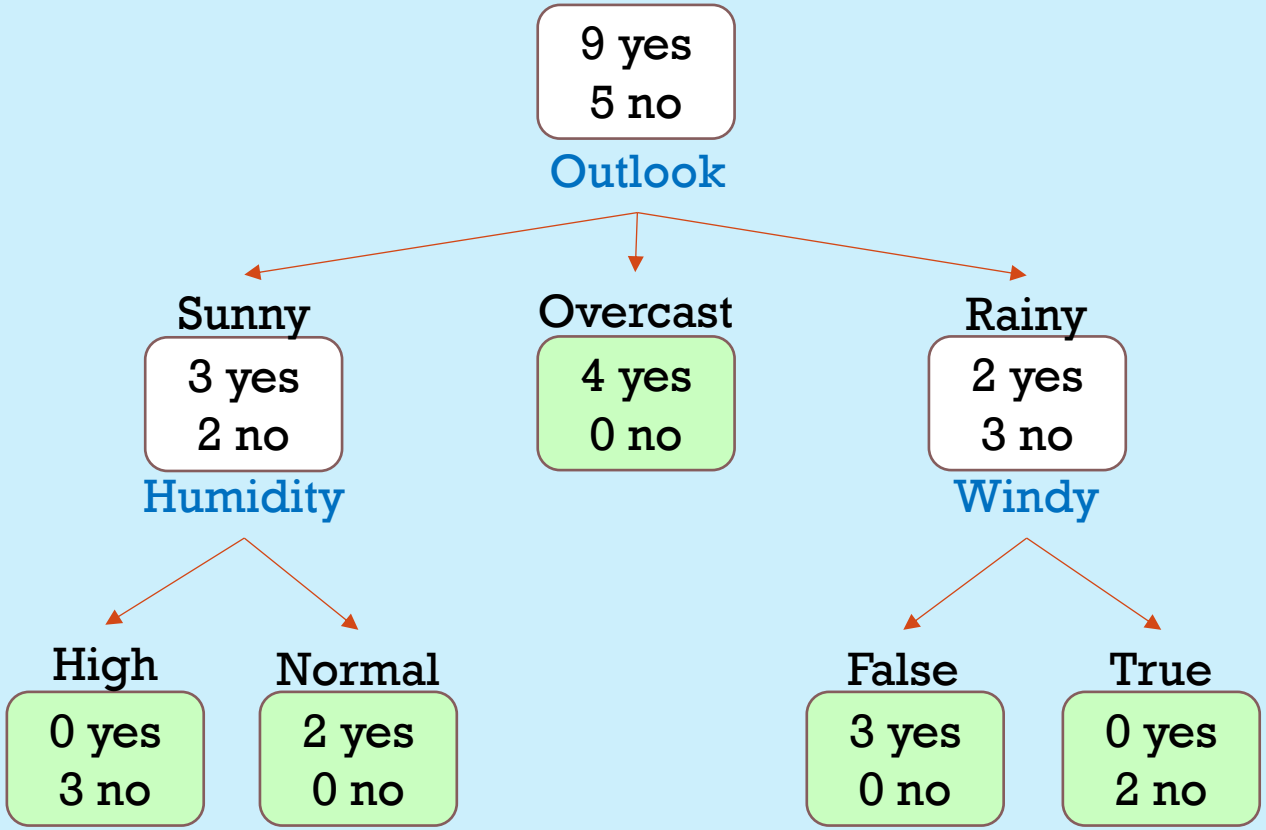
Humidity								GAIN
p(normal)	50.0%	p(yes|normal)	85.7%	p(no|normal)	14.3%	0.592	0.788	0.152
p(high)	50.0%	p(yes|high)	42.9%	p(no|high)	57.1%	0.985		

Windy								GAIN
p(FALSE)	57.1%	p(yes|W=FALSE)	75.0%	p(no|W=FALSE)	25.0%	0.811	0.892	0.048
p(TRUE)	42.9%	p(yes|W=TRUE)	50.0%	p(no|W=TRUE)	50.0%	1.000		

3. Particionar según el atributo con mayor ganancia de información
4. Parar si todas las hojas son puras o ya no hay mas atributos

UNIVERSIDAD ICESI

TALLER: ÁRBOLES DE DECISIÓN ID3



UNIVERSIDAD ICESI

ÁRBOLES DE DECISIÓN C4.5

Problemas con ID3 → c4.5:

- Problema con la ganancia de información como criterio de particionamiento:
  - preferencia por los atributos de mayor cardinalidad
  - c4.5 utiliza el ratio de ganancia, que maximiza la información generada por la partición

$$Gain(Play, Outlook) = 0.247$$

$$Entropia(Outlook) = -\left(\frac{5}{14} log_2 \frac{5}{14} + \frac{4}{14} log_2 \frac{4}{14} + \frac{5}{14} log_2 \frac{5}{14}\right) = 1.577$$

$$GainRatio(Play, Outlook) = \frac{0.247}{1.577} = 0.156$$

		Play		
		yes	no	Total
Outlook	Sunny	2	3	**5**
	Overcast	4	0	**4**
	Rainy	3	2	**5**
	Total	**9**	**5**	**14**

ÁRBOLES DE DECISIÓN C4.5

Extensión ID3 → c4.5:

- Manejo de datos faltantes (ID3 los ignora para hacer el modelo):
  - Influencia en la selección de variable de particionamiento:
    - →Multiplicación del information gain por la proporción de valores completos
    - →Consideración de valor adicional en el cálculo de la entropía de la variable de particionamiento
  - Consideración de registros con valores faltantes durante el entrenamiento
    - →Las instancias con valores faltantes del test se transmiten a todos los nodos, ponderadas con respecto a la proporción de instancias completas del nodo
  - Consideración de registros con valores faltantes durante la predicción
    - →Todas las ramas de la variable en cuestión son exploradas
    - →Se obtiene una predicción para cada sub-árbol, se agregan y se retorna la distribución de probabilidad de las clases posibles

ÁRBOLES DE DECISIÓN C4.5

Problemas con ID3 → c4.5:

- Poda como lucha para el overfitting
  - →Aprendizaje completo del árbol
  - →Reemplazo de subárbol por un nodo hoja, si se reduce el error de clasificación sobre un set de test
- Consideración de atributos numéricos
  - →Discretización en intervalos binarios, utilizando information gain

TALLER: DISCRETIZACIÓN

ID	outlook	temperature	humidity	windy	play (X)	p1	p2
5	rainy	cool	54	FALSE	yes	a	a
6	rainy	cool	58	TRUE	no	a	a
10	rainy	mild	59	FALSE	yes	a	a
7	overcast	cool	60	TRUE	yes	a	a
9	sunny	cool	60	FALSE	yes	a	a
11	sunny	mild	62	TRUE	yes	a	a
13	overcast	hot	63	FALSE	yes	b	a
3	overcast	hot	80	FALSE	yes	b	a
12	overcast	mild	81	TRUE	yes	b	a
2	sunny	hot	89	TRUE	no	b	b
14	rainy	mild	90	TRUE	no	b	b
1	sunny	hot	90	FALSE	no	b	b
8	sunny	mild	90	FALSE	no	b	b
4	rainy	mild	92	FALSE	yes	b	b

Ahora el atributo "humidity" es numérico.

¿Cómo encuentro el mejor particiona-miento de la variable numérica con respecto a la variable objetivo?

Considerar todas las particiones binarias posibles y escoger la que presente la mayor ganancia de información (en este taller solo consideramos 2). Para cada partición se deben seguir los siguientes pasos:

1. Calcular la entropía de clase del subconjunto actual

2. Calcular las entropías condicionales para cada partición, teniendo en cuenta todos condicionamientos posibles. Calcular la ganancia de información correspondiente.

3. Escoger la partición con la mayor ganancia de información

UNIVERSIDAD ICESI

ÁRBOLES DE DECISIÓN: CART

- Solo árboles con particionamientos **binarios**
- Manejo de datos faltantes: Utilización de varias variables de particionamiento **sustitutas** (surrogate) para suplantar la variable cuyo valor es faltante
- **Gini** como criterio de impureza para el particionamiento:
  - 0 pureza perfecto: todas las instancias de la misma clase
  - 0.5 impureza: distribución equitativa de las instancias entre ambas clases
- Algoritmo
  - Para cada atributo
    - Para cada posible split binario del atributo
      - Calcular el Gini para ambos subnodos

        $gini = \sum p * (1 - p) = 1 - \sum p^2,$ donde p es la probabilidad de cada clase.
      - Calcular el promedio ponderado del Gini de las particiones
    - Seleccionar el split binario con el menor promedio de Gini
  - Seleccionar el atributo con el menor promedio de Gini

ÁRBOLES DE DECISIÓN: CHAID

- Particiones en 2 o más subconjuntos
- Chi cuadrado como criterio de particionamiento: significancia estadística de las diferencias entre los nodos hijos y el nodo padre
- Algoritmo

  Para cada atributo

  1. Calcular el valor esperado para cada combinación con la variable objetivo
  2. Calcular el Chi cuadrado para cada nodo hijo

$$\chi^2 = \frac{(Observado - Esperado)^2}{Esperado}$$

  3. Escoger la variable con el mayor valor del chi cuadrado.

- Solo sirven para clasificación

PRIMER PARCIAL

Revisión de las preguntas:

ÁRBOLES DE DECISIÓN: CART

	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

n	yes	p	w	Split	Gini	Avg.Gini
5	2	40.0%	35.7%	sunny	0.480	0.394
9	7	77.8%	64.3%	(overcast | rainy)	0.346	

n	yes	p	w	Split	Gini	Avg.Gini
4	4	100.0%	28.6%	overcast	0.000	0.357
10	5	50.0%	71.4%	(sunny | rainy)	0.500	

n	yes	p	w	Split	Gini	Avg.Gini
5	3	60.0%	35.7%	rainy	0.480	0.457
9	6	66.7%	64.3%	(sunny | overcast)	0.444	

n	yes	p	w	Split	Gini	Avg.Gini
4	2	50.0%	28.6%	hot	0.500	0.443
10	7	70.0%	71.4%	(mild | cool)	0.420	

n	yes	p	w	Split	Gini	Avg.Gini
6	4	66.7%	42.9%	mild	0.444	0.458
8	5	62.5%	57.1%	(hot | cool)	0.469	

n	yes	p	w	Split	Gini	Avg.Gini
4	3	75.0%	28.6%	cool	0.375	0.450
10	6	60.0%	71.4%	(hot | mild)	0.480	

n	yes	p	w	Split	Gini	Avg.Gini
7	6	85.7%	50.0%	normal	0.245	0.439
7	3	42.9%	50.0%	high	0.633	

n	yes	p	w	Split	Gini	Avg.Gini
8	6	75.0%	57.1%	FALSE	0.375	0.429
6	3	50.0%	42.9%	TRUE	0.500	

UNIVERSIDAD ICESI

DECISION TREES: CHAID

	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Outlook	play		
Observed	yes	no	
sunny	2	3	5
overcast	4	0	4
rainy	3	2	5
	9	5	14

Expected			
sunny	3,2	1,8	5,0
overcast	2,6	1,4	4,0
rainy	3,2	1,8	5,0
	9,0	5,0	14,0

CHI2			
sunny	0,46	0,83	1,28
overcast	0,79	1,43	2,22
rainy	0,01	0,03	0,04
	1,27	2,28	3,55

Temperature	play		
Observed	yes	no	
hot	2	2	4
mild	4	2	6
cool	3	1	4
	9	5	14

Expected			
hot	2,6	1,4	4,0
mild	3,9	2,1	6,0
cool	2,6	1,4	4,0
	9,0	5,0	14,0

CHI2			
hot	0,13	0,23	0,36
mild	0,01	0,01	0,01
cool	0,07	0,13	0,20
	0,20	0,37	0,57

Humidity	play		
Observed	yes	no	
high	3	4	7
normal	6	1	7
	9	5	14

Expected			
hot	4,5	2,5	7,0
normal	4,5	2,5	7,0
	9,0	5,0	14,0

CHI2			
hot	0,50	0,90	1,40
normal	0,50	0,90	1,40
	1,00	1,80	2,80

Windy	play		
Observed	yes	no	
FALSE	6	2	8
TRUE	3	3	6
	9	5	14

Expected			
FALSE	5,1	2,9	8,0
TRUE	3,9	2,1	6,0
	9,0	5,0	14,0

CHI2			
hot	0,14	0,26	0,40
normal	0,19	0,34	0,53
	0,33	0,60	0,93

UNIVERSIDAD ICESI

REFERENCIAS

- *Introduction to Statistical Learning with Applications in R (ISLR), G. James, D. Witten, T. Hastie & R. Tibshirani, 2014*
- *Data Science for Business*, Foster Provost & Tom Fawcett, O'Reilly, 2013
- *Machine Learning,* Tom M. Mitchell, McGraw-Hill, 1997