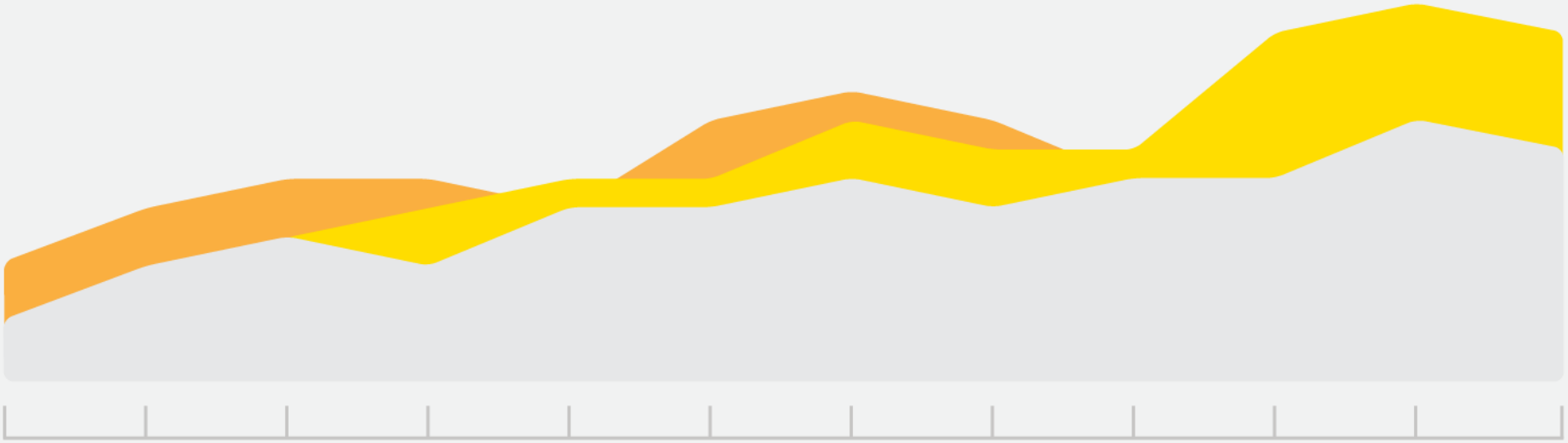
oldemar rodríguez
CONSULTOR en MINERIA DE DATOS

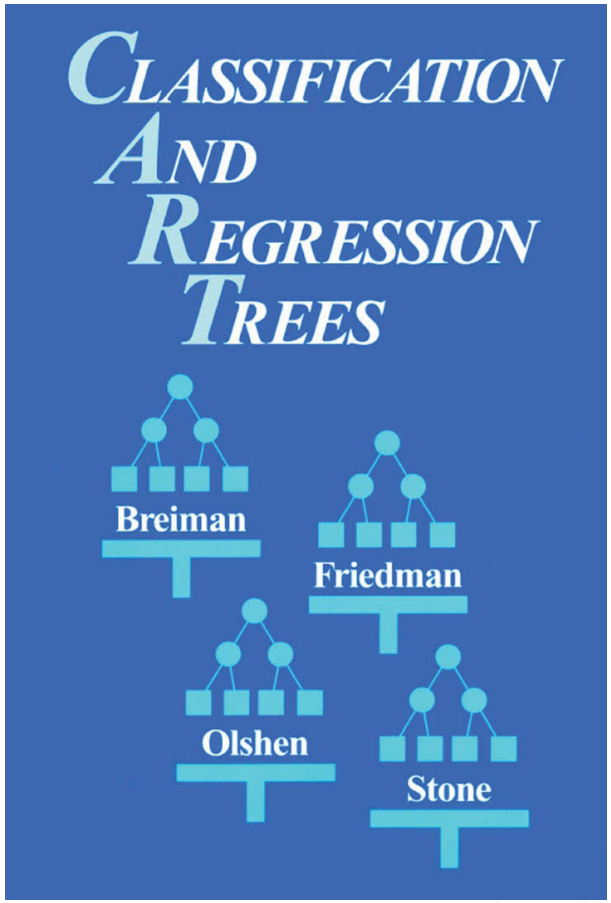
Aprendizaje Supervisado

Árboles de Decisión

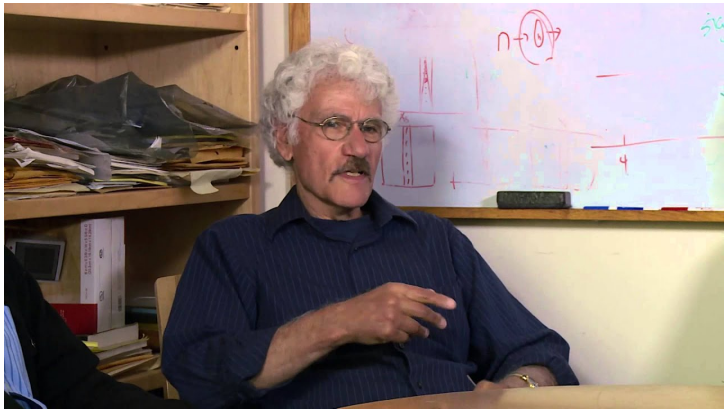


Método CART

Classification And Regression Trees



L. Brieman



J. Friedman

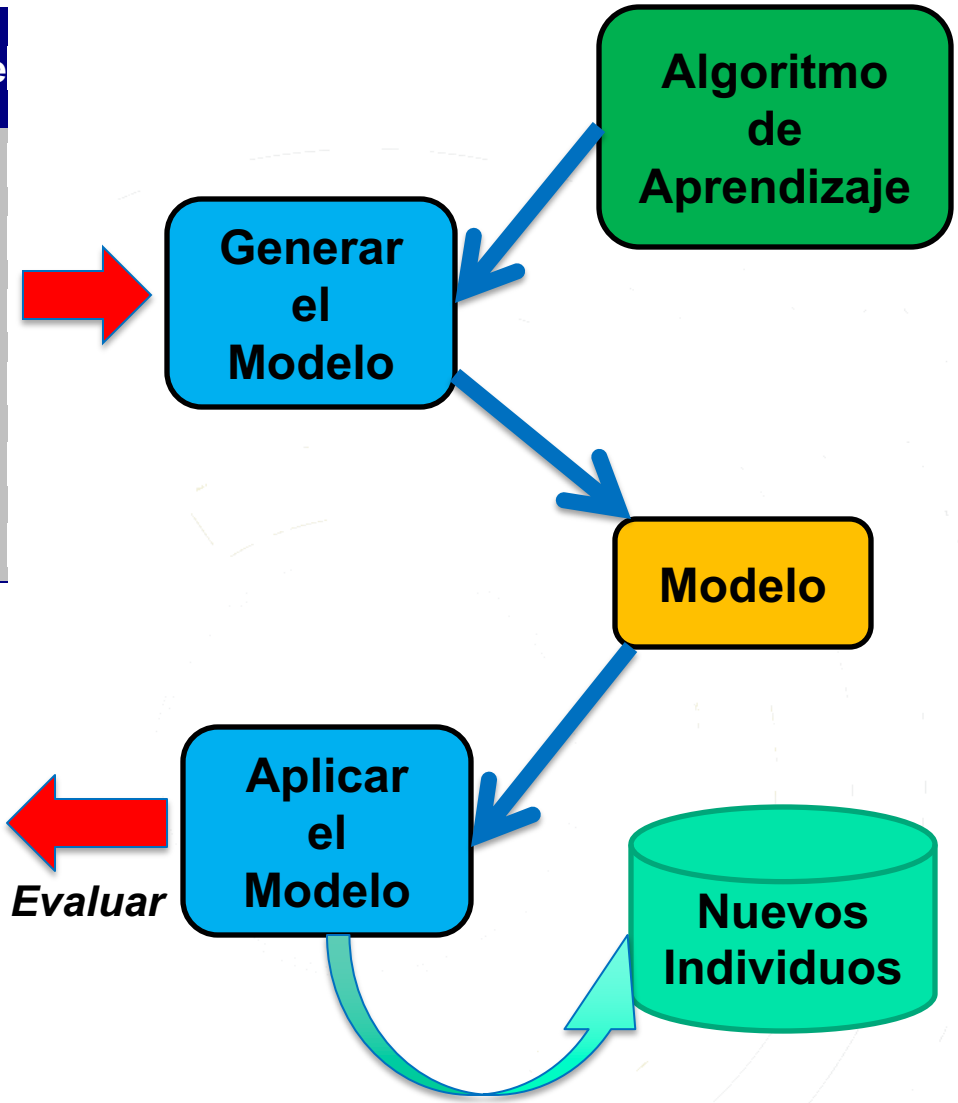
Modelo general de los métodos de Clasificación

Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
1	Sí	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No

Tabla de Aprendizaje

Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
7	No	Soltero	80K	No
8	Si	Casado	100K	No
9	No	Soltero	70K	No

Tabla de Testing



Clasificación: Definición

- Dada una colección de registros (conjunto de entrenamiento) cada registro contiene un conjunto de variables (atributos) denominado x, con un variable (atributo) adicional que es la clase denominada y.

- El objetivo de la ***clasificación*** es encontrar un modelo (una función) para predecir la clase a la que pertenecería cada registro, esta asignación una clase se debe hacer con la mayor precisión posible.

- Un conjunto de prueba (tabla de testing) se utiliza para determinar la precisión del modelo. Por lo general, el conjunto de datos dado se divide en dos conjuntos al azar de el de entrenamiento y el de prueba.

Definición de Clasificación

- Dada una base de datos $D = \{t_1, t_2, \dots, t_n\}$ de tuplas o registros (individuos) y un conjunto de clases $C = \{C_1, C_2, \dots, C_m\}$, el **problema de la clasificación** es encontrar una función $f: D \to C$ tal que cada t_i es asignada una clase C_j.

- $f: D \to C$ podría ser una Red Neuronal, un Árbol de Decisión, un modelo basado en Análisis Discriminante, o una Red Beyesiana.

Ejemplo: Créditos en un Banco

Tabla de Aprendizaje

Variable Discriminante

OLDEMARRR.DMEx...ditoViviendaPeq

	Id	MontoCredito	IngresoNeto	CoeficienteCre...	MontoCuota	GradoAcademico	BuenPagador
▶	1	2	4	3	1	4	1
	2	2	3	2	1	4	1
	3	4	1	1	4	2	2
	4	1	4	3	1	4	1
	5	3	3	1	3	2	2
	6	3	4	3	1	4	1
	7	4	2	1	3	2	2
	8	4	1	3	3	2	2
	9	3	4	3	1	3	1
	10	1	3	2	2	4	1
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Con la Tabla de Aprendizaje se entrena (aprende) el modelo matemático de predicción, es decir, a partir de esta tabla se calcula la función ***f*** de la definición anterior.

Ejemplo: Créditos en un Banco

Tabla de Testing

Variable Discriminante

OLDEMARRR.DME...iviendaPeqPRED | OLDEMARRR.DMEx...ditoViviendaPeq

	Id	MontoCredito	IngresoNeto	CoeficienteCre...	MontoCuota	GradoAcademico	BuenPagador
▶	11	3	3	3	3	1	2
	12	2	2	2	2	1	1
	13	2	2	3	2	1	1
	14	1	3	4	3	2	2
	15	1	2	4	2	1	1
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL

- Con la Tabla de Testing se valida el modelo matemático de predicción, es decir, se verifica que los resultados en individuos que no participaron en la construcción del modelo es bueno o aceptable.
- Algunas veces, sobre todo cuando hay pocos datos, se utiliza la Tabla de Aprendizaje también como de Tabla Testing.

Ejemplo: Créditos en un Banco

Nuevos Individuos

Variable Discriminante

OLDEMARRR.DMEx ...editoViviendaNI

	Id	MontoCredito	IngresoNeto	CoeficienteCre...	MontoCuota	GradoAcademico	BuenPagador
	100	4	4	2	2	3	?
	101	1	4	3	2	4	?
	102	3	2	3	4	2	?
▶*	NULL	NULL	NULL	NULL	NULL	NULL	NULL

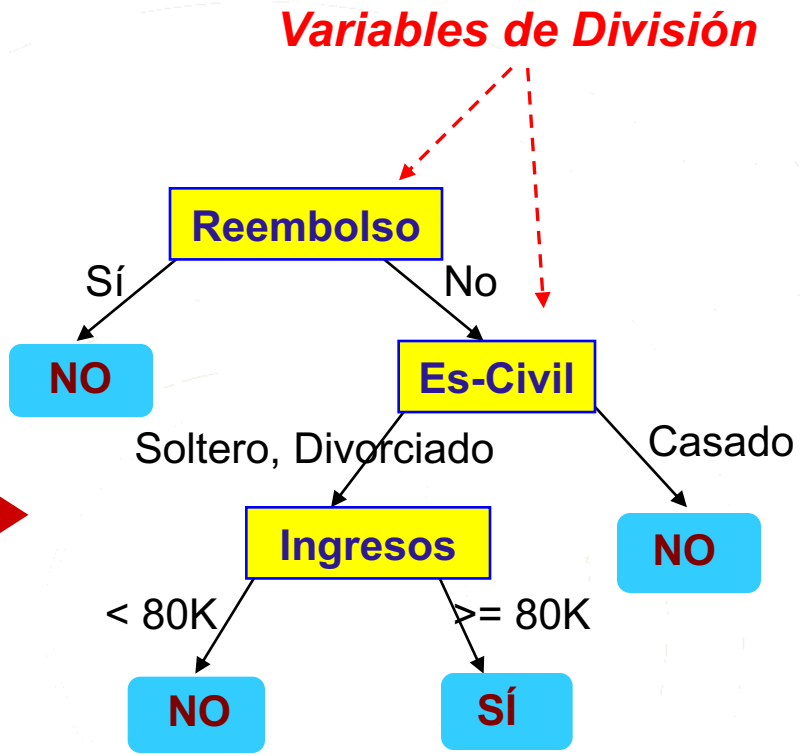
Con la Tabla de Nuevos Individuos se predice si estos serán o no buenos pagadores.

Un ejemplo de un árbol de decisión

categórica categórica continua clase

Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
1	Sí	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No
7	Sí	Divorciado	220K	No
8	No	Soltero	85K	Sí
9	Sí	Casado	75K	No
10	No	Soltero	90K	Sí

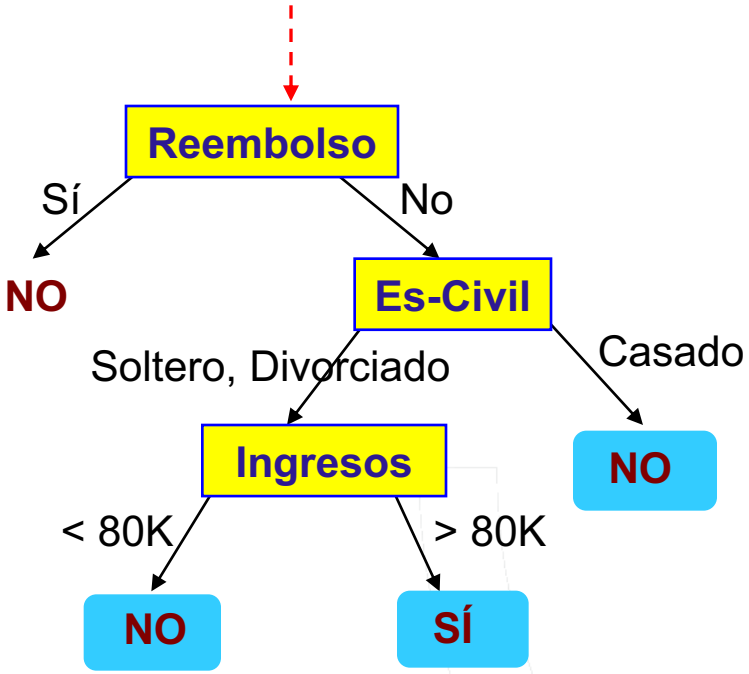
Tabla de Aprendizaje



Modelo: Árbol de Decisión

Aplicando el modelo de árbol para predecir la clase para una nueva observación

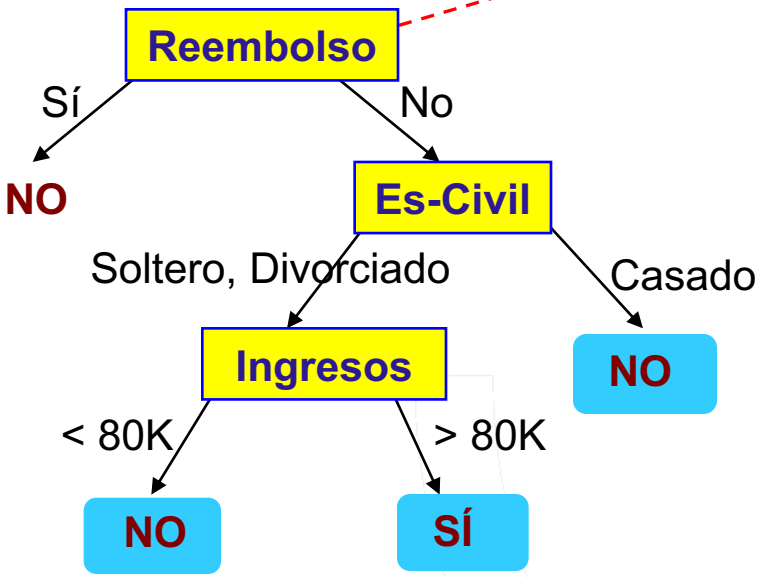
Inicia desde la raíz del árbol



Datos de Prueba

Reebolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?

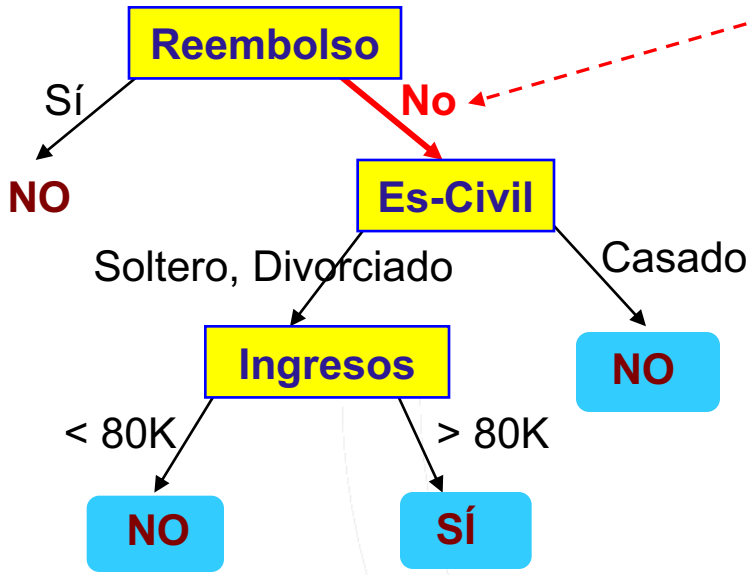
Aplicando el modelo de árbol para predecir la clase para una nueva observación



Datos de Prueba

Reebolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?

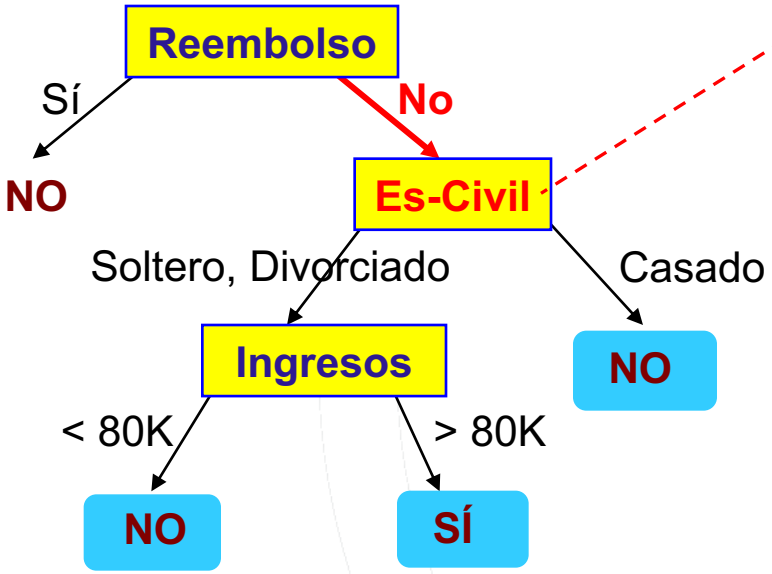
Aplicando el modelo de árbol para predecir la clase para una nueva observación



Datos de Prueba

Reebolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?

Aplicando el modelo de árbol para predecir la clase para una nueva observación



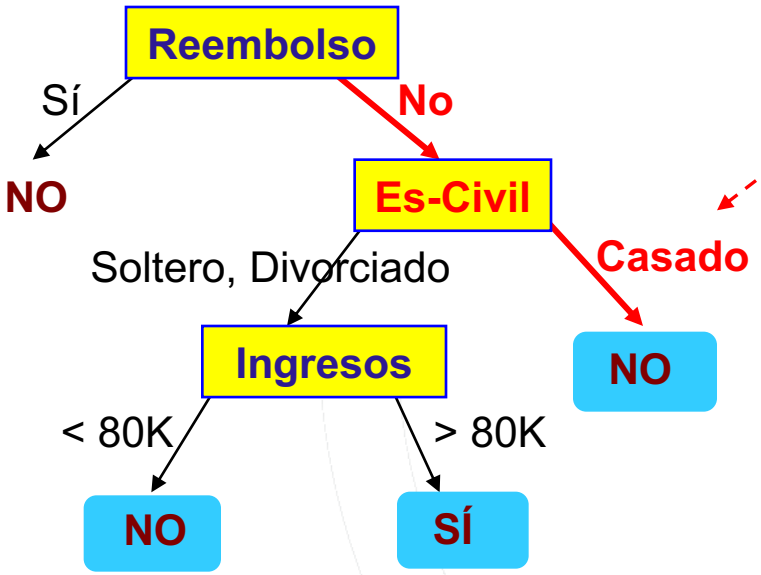
Datos de Prueba

Reebolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?

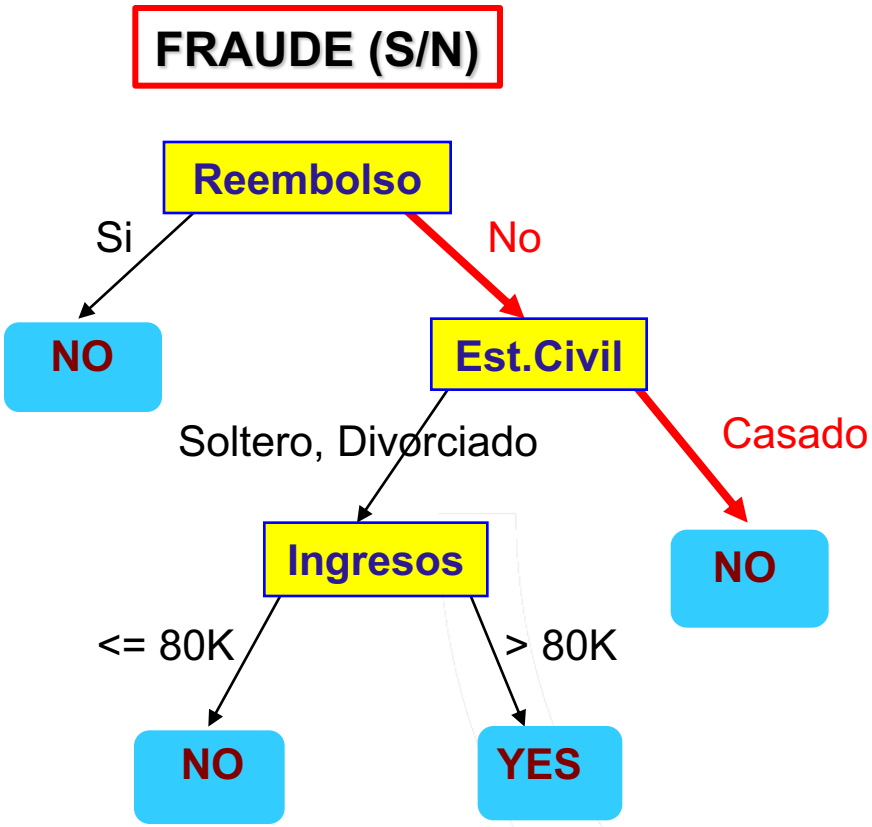
Aplicando el modelo de árbol para predecir la clase para una nueva observación

Datos de Prueba

Reebolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?



Aplicando el modelo de árbol para predecir la clase para una nueva observación



Datos de Prueba

Reebolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?

Asigna “No” a la variable de Clase “Fraude”

¿Cómo se generan los árboles de decisión?

- Muchos algoritmos usan una versión con un enfoque "top-down" o "dividir y conquista" conocido como Algoritmo de Hunt.

- Sea D_t el conjunto de registros de entrenamiento en un nodo t dado.

- Sea $y_t = \{y_1, y_2, \ldots, y_c\}$ el conjunto de etiquetas de las clases.

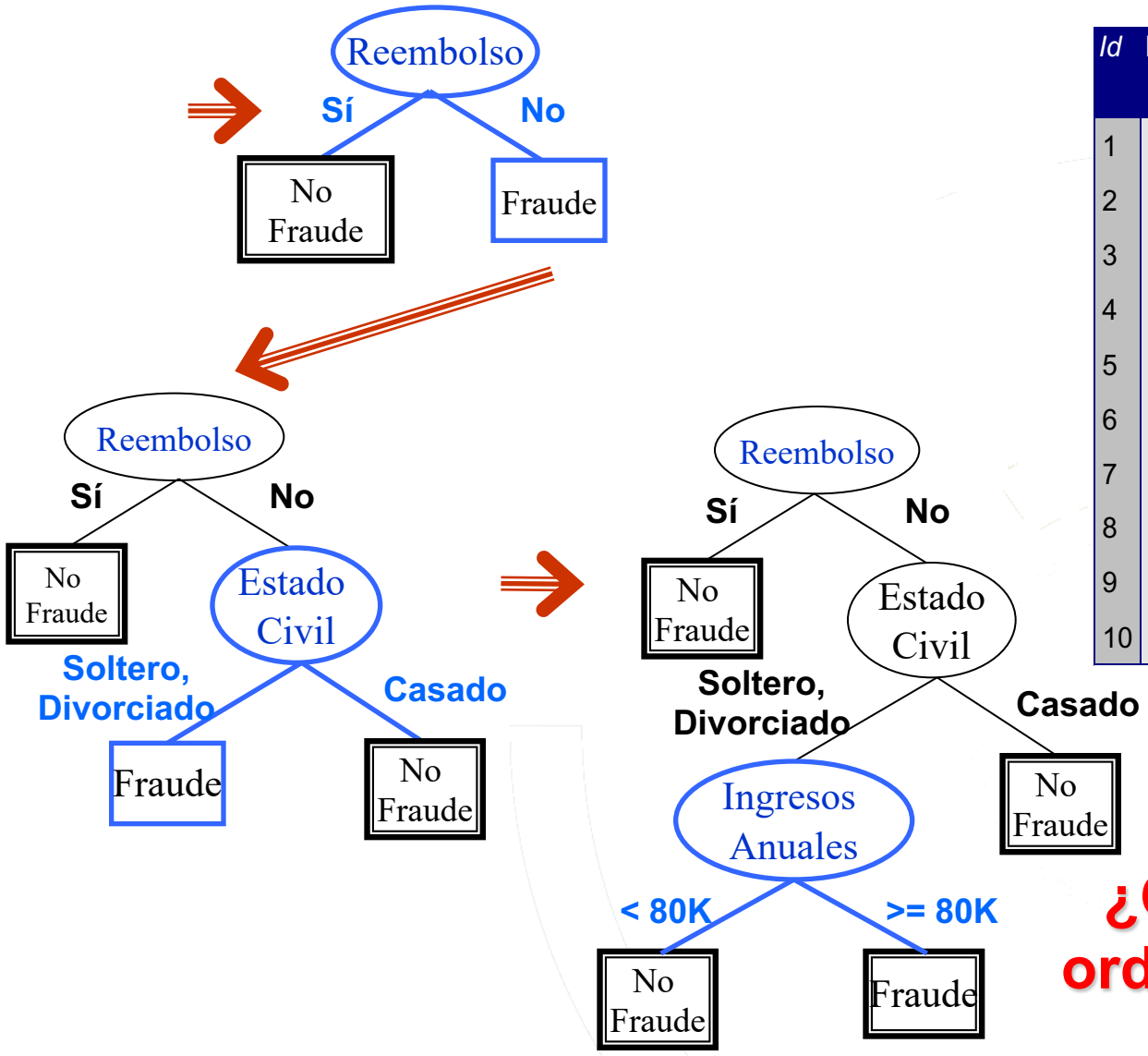
Algoritmo de Hunt:

- Si todos los registros D_t pertenecen a la misma clase y_t, entonces t es un nodo hoja que se etiqueta como y_t

- Si D_t contiene registros que pertenecen a más de una clase, se escoge una variable (atributo) para dividir los datos en subconjuntos más pequeños.

- Recursivamente se aplica el procedimiento a cada subconjunto.

Un ejemplo del algoritmo de Hunt



Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
1				
2	No			
3	No	Soltero	70K	No
4				
5	No	Divorciado	95K	Sí
6	No			
7				
8	No	Soltero	85K	Sí
9				
10	No	Soltero	90K	Sí

¿Cómo se escoge el orden de las variables?

¿Cómo aplicar el algoritmo de Hunt?

- Por lo general, se lleva a cabo de manera que la separación que se elige en cada etapa sea ***óptima*** de acuerdo con algún criterio.

- Sin embargo, puede no ser óptima al final del algoritmo (es decir no se encuentre un árbol óptimo como un todo). Aún así, este el enfoque computacional es eficiente por lo que es muy popular.

¿Cómo aplicar el algoritmo de Hunt?

- ✓ Utilizando el enfoque de optimización aún se tienen que decidir tres cosas:

1. ¿Cómo dividiremos las variables?

2. ¿Qué variables (atributos) utilizar y en que orden? ¿Qué criterio utilizar para seleccionar la "mejor" división?

3. ¿Cuándo dejar de dividir? Es decir, ¿Cuándo termina el algoritmo?

¿Cómo aplicar el algoritmo de Hunt?

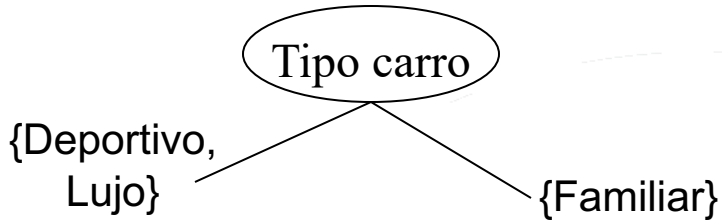
- Para la pregunta 1, se tendrán en cuenta sólo divisiones binarias tanto para predictores numéricos como para los categóricos, esto se explica más adelante (Método CART).

- Para la pregunta 2 se considerarán el *Error de Clasificación*, el *Índice de Gini* y la *Entropía.*

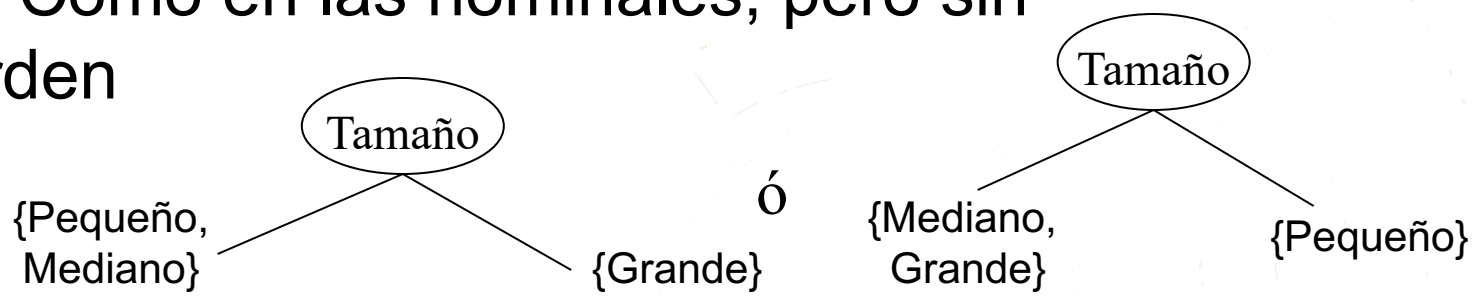
- La pregunta 3 tiene una respuesta difícil de dar porque implica la selección del modelo. Se debe tomar en cuenta qué tanto se quieren afinar las reglas generadas.

Pregunta #1: Solamente se usarán divisiones Binarias (Método CART):

Nominales:



Ordinales: Como en las nominales, pero sin violar el orden



Numéricas: Frecuentemente se divide en el punto medio

Pregunta #2: Se usarán los siguientes criterios de *IMPUREZA*: el *Error de Clasificación*, el *Índice de Gini* y la *Entropía*, para esto se define la siguiente probabilidad:

- ***p(j|t)*** *= La probalidad de pertenecer a la clase "j" estando en el nodo t.*
- *Muchas veces simplemente se usa* p_j

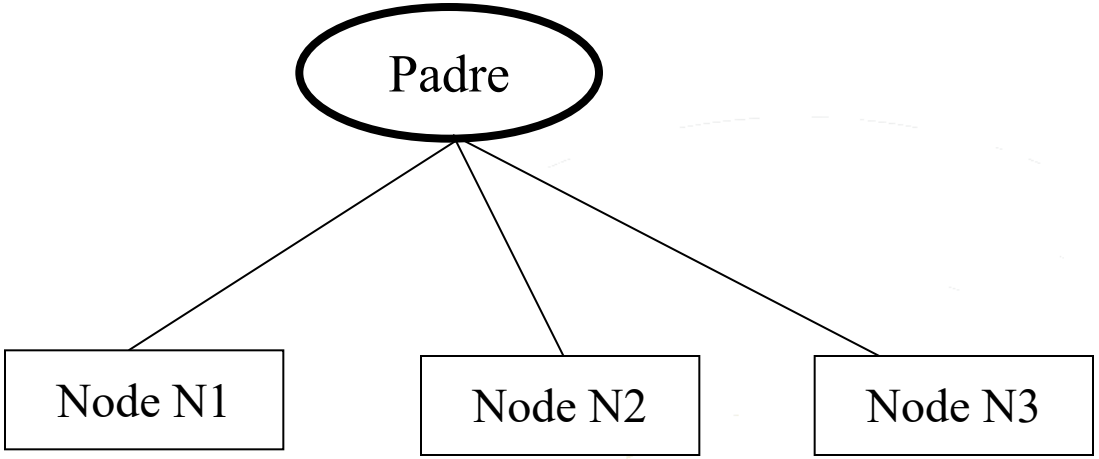
Pregunta #2: Se usarán el *Error de Clasificación*, el Índice de *Gini* y la *Entropía*

Error de clasificación: $$Error(t) = 1 - \max_j [\, p(j \mid t)]$$

Índice de Gini: $$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

Entropía: $$Entropía(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

Ejemplo de cálculo de índices:



	N1	N2	N3
C1	0	1	2
C2	6	5	4

Ejemplo de cálculo de *Gini*

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

C1	**0**
C2	**6**

P(C1|N1) = 0/6 = 0 P(C2|N1) = 6/6 = 1

Gini = 1 – P(C1|N1)² – P(C2|N1)² = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1|N2) = 1/6 P(C2|N2) = 5/6

Gini = 1 – (1/6)² – (5/6)² = 0.278

C1	**2**
C2	**4**

P(C1|N2) = 2/6 P(C2|N2) = 4/6

Gini = 1 – (2/6)² – (4/6)² = 0.444

Ejemplo de cálculo de la *Entropía*

C1	**0**
C2	**6**

P(C1|N1) = 0/6 = 0 **P(C2|N1) = 6/6 = 1**

Entropía = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1|N2) = 1/6 **P(C2|N2) = 5/6**

Entropía = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

C1	**2**
C2	**4**

P(C1|N3) = 2/6 **P(C2|N3) = 4/6**

Entropía = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Ejemplo de cálculo del *Error de Clasificación*

C1	**0**
C2	**6**

Error Clasificación = 1-max[0/6,6/6]= 0

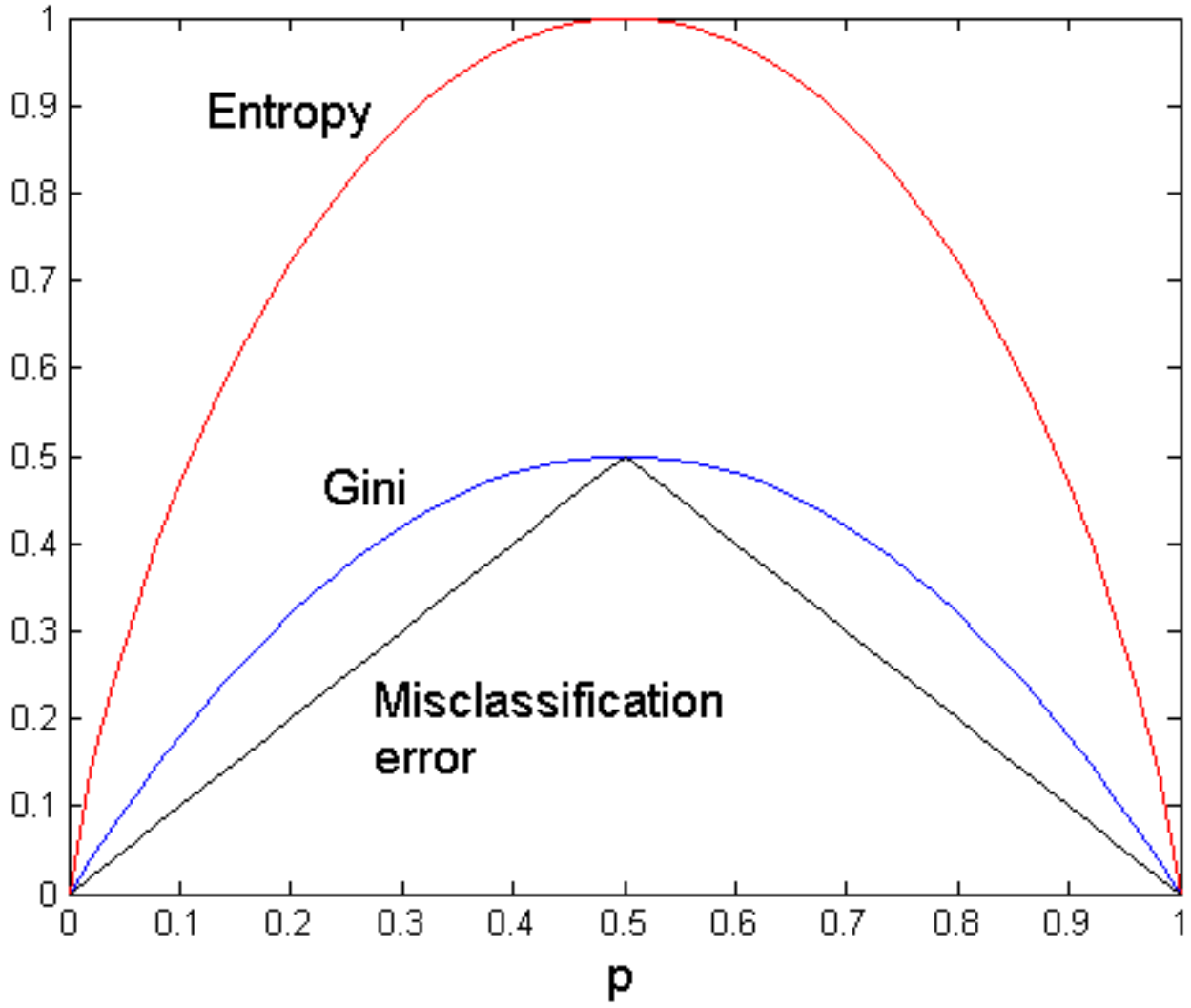
C1	**1**
C2	**5**

Error Clasificación = 1-max[1/6,5/6]= 0,167

C1	**2**
C2	**4**

Error Clasificación = 1-max[2/6,4/6]= 0,333

Comparación Gráfica



Gini Split

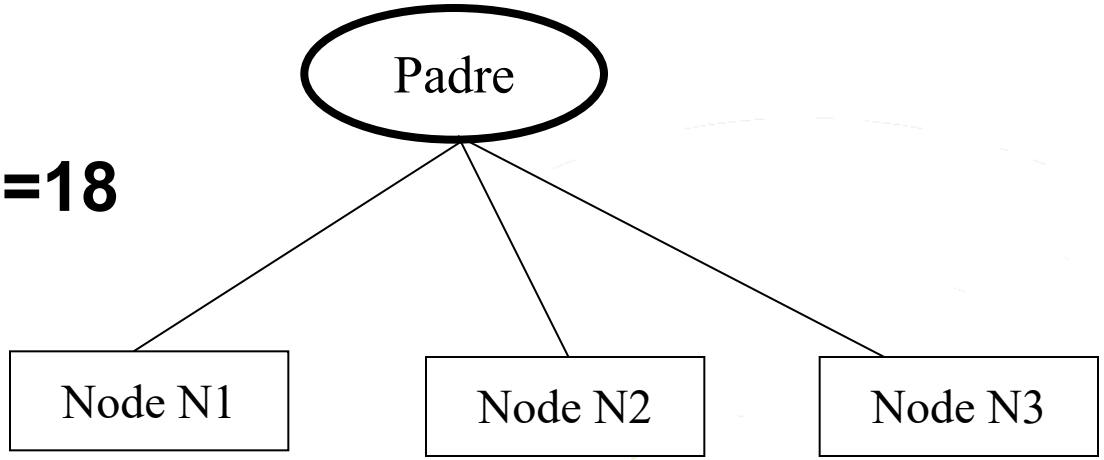
- Después de que el índice de Gini se calcula en cada nodo, el valor total del índice de Gini se calcula como el promedio ponderado del índice de Gini en cada nodo:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Ejemplo de cálculo $GINI_{Split}$

|padre|=18

Cardinalidad=18



	N1	N2	N3
C1	0	1	2
C2	6	5	4

Ejemplo de cálculo de $GINI_{split}$

C1	**0**
C2	**6**

P(C1|N1) = 0/6 = 0 P(C2|N1) = 6/6 = 1

Gini = 1 – $P(C1)^2$ – $P(C2)^2$ = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1|N2) = 1/6 P(C2|N2) = 5/6

Gini = 1 – $(1/6)^2$ – $(5/6)^2$ = 0.278

C1	**2**
C2	**4**

P(C1|N3) = 2/6 P(C2|N3) = 4/6

Gini = 1 – $(2/6)^2$ – $(4/6)^2$ = 0.444

$GINI_{split}$ = (6/18)*0+(6/18)*0.278+(6/18)*0.444 = 0.24

En este caso todos los pesos son iguales porque todas las clases tienen 6 elementos

Entropía Split

- Después de que el índice de Entropía se calcula en cada nodo, el valor total del índice de Entropía se calcula como el promedio ponderado del índice de Entropía en cada nodo:

$$Entropia_{split} = \sum_{i=1}^{k} \frac{n_i}{n} Entropia(i)$$

Error de Clasificación Split

- Después de que el Error de Clasificación (EC) se calcula en cada nodo, el valor total del índice del EC se calcula como el promedio ponderado del índice EC en cada nodo:

$$EC_{split} = \sum_{i=1}^{k} \frac{n_i}{n} EC(i)$$

Ejemplo con error de clasificación

Ejemplo con error de clasificación				
	N1	N2	N3	N4
C1	0	4	2	1
C2	6	4	3	5

- **N1 es un nodo completamente puro**
- **N2 es un nodo completamente impuro**
- **N3 es un nodo bastante impuro**
- **N4 es un nodo bastante puro**

Cálculo para el nodo N1

Cálculos para N1	
C1	0
C2	6

Probabilidades:

- **P(C1|N1)=0/6=0**
- **P(C2|N1)=6/6=1**

Luego el error de clasificación es 1-max(p(j|t)):

EC(N1)=1-max(0,1)=0

Cálculo para el nodo N2

Cálculos para N2	
C1	4
C2	4

Probabilidades:

- **P(C1|N2)=4/8=0,5**
- **P(C2|N2)=4/8=0,5**

Luego el error de clasificación es 1-max(p(j|t)):

EC(N2)=1-max(0,5,0,5)=0,5

Cálculo para el nodo N3

Cálculos para N2	
C1	2
C2	3

Probabilidades:

- **P(C1|N3)=2/5=0,4**
- **P(C2|N3)=3/5=0,6**

Luego el error de clasificación es 1-max(p(j|t)):

EC(N3)=1-max(0,4,0,6)=0,4

Cálculo para el nodo N4

Cálculos para N2	
C1	1
C2	5

Probabilidades:

- **P(C1|N4)=1/6=0,1667**
- **P(C2|N4)=5/6=0,8333**

Luego el error de clasificación es 1-max(p(j|t)):

EC(N3)=1-max(0,1667,0,8333)=0,1667

Error de clasificación Split

Datos previos

Total de individuos 25.

Individuos de N1=6 , EC(N1)=0.

Individuos de N2=8 , EC(N2)=0,5.

Individuos de N3=5 , EC(N3)=0,4.

Individuos de N4=6 , EC(N4)=0,1667.

ECS= (6/25)*0+(8/25)*0,5+(5/25)*0,4+(6/25)*0,1667=0,280008

Información Ganada → IG_{Split}

- ✓ Cada vez que se va a hacer una nueva división en el árbol (split the tree) se debe comparar el grado de impureza del nodo padre respecto al grado de impureza de los nodos hijos.
- ✓ Esto se calcula con el índice de Información Ganada (IG), que es la resta de la impureza del nodo padre menos el promedio ponderado de las impurezas de los nodos hijos.
- ✓ La idea en IG_{Split} sea máximo y esto se logra si el promedio ponderado de las impurezas de los nodos hijos es mínimo.

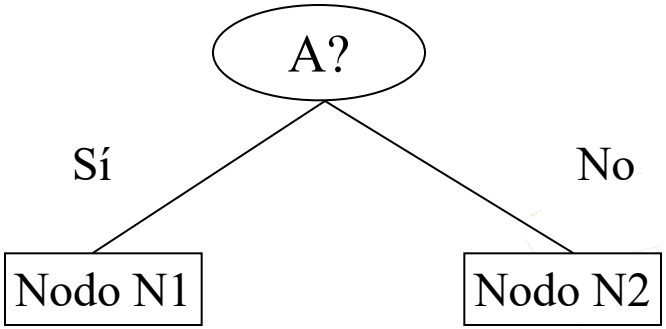
$$\Delta = IG_{split} = I(padre) - \left(\sum_{i=1}^{k} \frac{n_i}{n} I(i) \right)$$

- Donde I es el índice de GINI, la Entropía o el Error de Clasificación.

Ejemplo: Información Ganada → IG_{Split}

	Padre
C1	7
C2	3
Gini = 0.42	

Gini(N1)
= 1 – (3/3)² – (0/3)²
= 0



Gini(N2)
= 1 – (4/7)² – (3/7)²
= 0.490

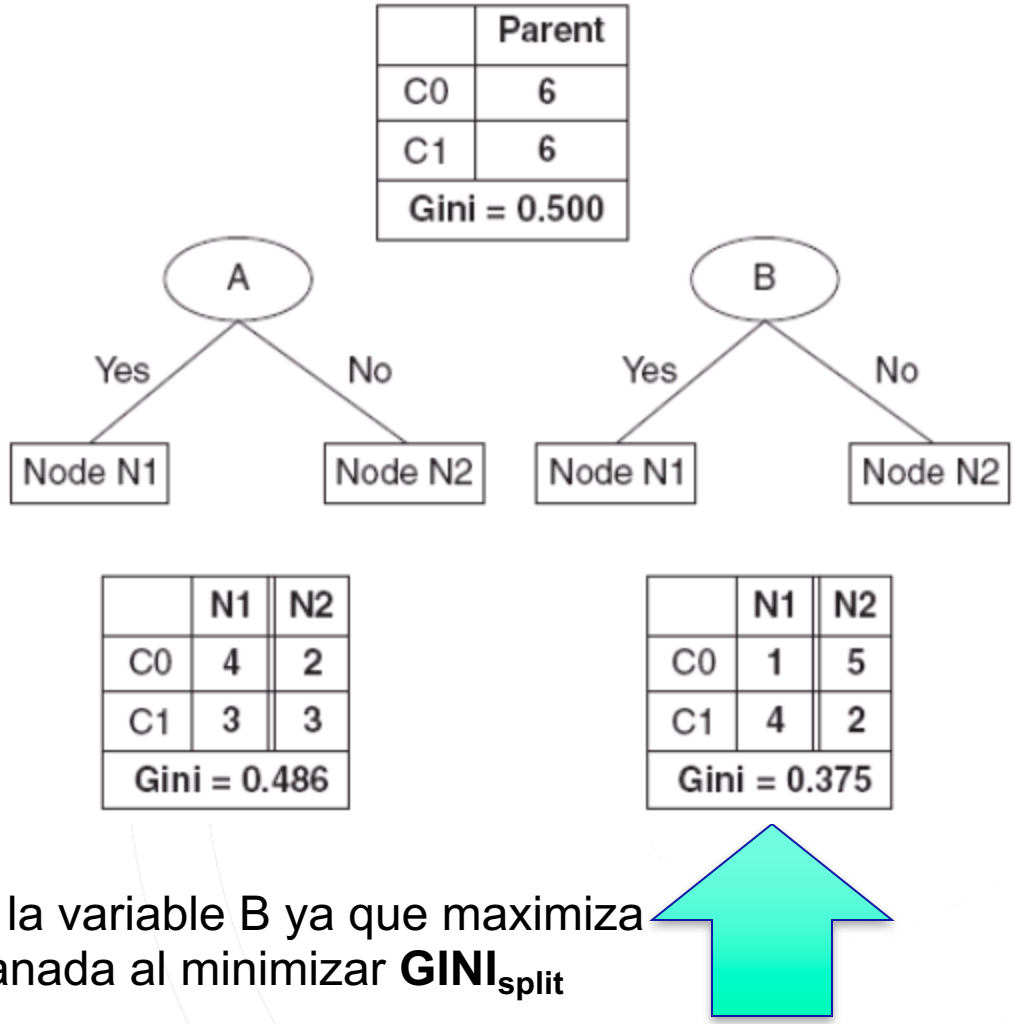
$GINI_{split}$ =

= 3/10 * 0 + 7/10 * 0.49
= 0.343

	N1	N2
C1	3	4
C2	0	3

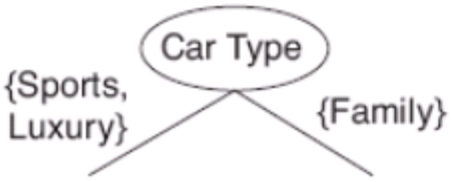
$$IG_{split} = I(padre) - \left(\sum_{i=1}^{k} \frac{n_i}{n} I(i) \right) = 0.42 - 0.343 = 0.077$$

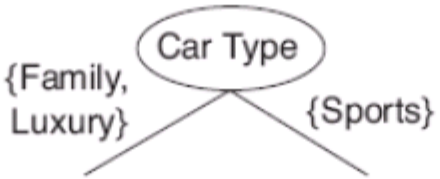
¿Cómo escoger la mejor división?

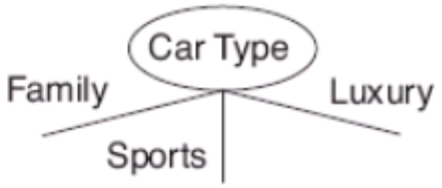


Se debe escoger la variable B ya que maximiza la Información Ganada al minimizar **$GINI_{split}$**

¿Cómo escoger la mejor división?







	Car Type	
	{Sports, Luxury}	{Family}
C0	9	1
C1	7	3
Gini	0.468	

	Car Type	
	{Sports}	{Family, Luxury}
C0	8	2
C1	0	10
Gini	0.167	

(a) Binary split

	Car Type		
	Family	Sports	Luxury
C0	1	8	1
C1	3	0	7
Gini	0.163		

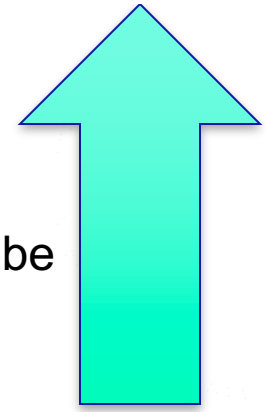
(b) Multiway split

Pero si se tiene solamente división binaria se escoge esta división ya que maximiza la Información Ganada al minimizar **$GINI_{split}$**

En caso de tener división múltiple se escoge esta división ya que maximiza la Información Ganada al minimizar **$GINI_{split}$**

¿Cómo escoger la mejor división?

Class	No	No	No	Yes	Yes	Yes	No	No	No	No
	Annual Income									
Sorted Values →	60	70	75	85	90	95	100	120	125	220
Split Positions →	55	65	72	80	87	92	97	110	122	172	230
	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >
Yes	0 3	0 3	0 3	0 3	1 2	2 1	3 0	3 0	3 0	3 0	3 0
No	0 7	1 6	2 5	3 4	3 4	3 4	3 4	4 3	5 2	6 1	7 0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	*0.300*	0.343	0.375	0.400	0.420



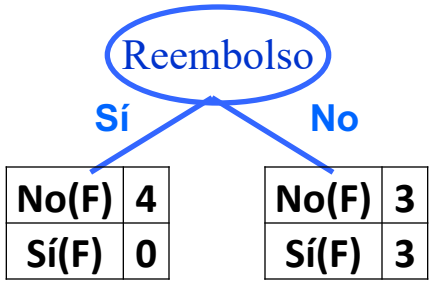
La variable “Annual Income” se debe dividir en “97” ya que maximiza la Información Ganada al minimizar **$GINI_{split}$**

¿Por qué se escoge "Reembolso" como variable inicial?

R/ Tiene el menor $GINI_{split}$

(mayor información ganada)
Empatado con Estado Civil,

pero aparece de primero



Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
1	Sí	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No
7	Sí	Divorciado	220K	No
8	No	Soltero	85K	Sí
9	Sí	Casado	75K	No
10	No	Soltero	90K	Sí

$GINI(No) = 1-(3/6)^2-(3/6)^2 = 1-0.25-0.25 = 0.5$

$GINI(Sí) = 1-(0/4)^2-(4/4)^2 = 1-0-1 = 0$

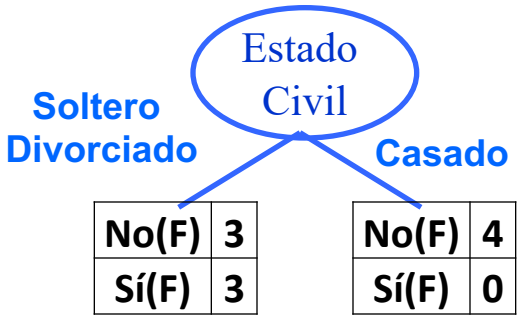
$GINI_{split} = (4/10)*0+(6/10)* 0.5$ **= 0.3**

¿Por qué se escoge "Reembolso" como variable inicial?

R/ Tiene el menor $GINI_{split}$

(mayor información ganada)
Empatado con Estado Civil,

pero aparece de primero



Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
1	Sí	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No
7	Sí	Divorciado	220K	No
8	No	Soltero	85K	Sí
9	Sí	Casado	75K	No
10	No	Soltero	90K	Sí

$GINI(Casado) = 1-(0/4)^2-(4/4)^2 = 0$

$GINI(Soltero/Divorciado) = 1-(3/6)^2-(3/6)^2 = 1-0.25-0.25 = 0.5$

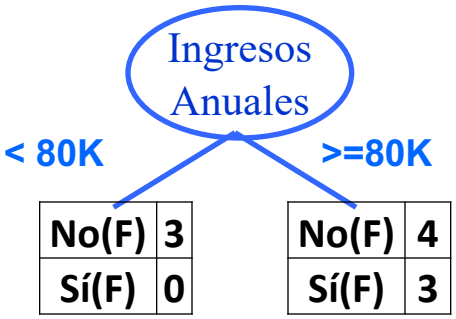
$GINI_{split} = (6/10)*0.5+(4/10)*0$ **= 0.3**

¿Por qué se escoge "Reembolso" como variable inicial?

R/ Tiene el menor $GINI_{split}$

(mayor información ganada)
Empatado con Estado Civil,

pero aparece de primero



Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
1	Sí	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No
7	Sí	Divorciado	220K	No
8	No	Soltero	85K	Sí
9	Sí	Casado	75K	No
10	No	Soltero	90K	Sí

$GINI(>=80) = 1-(4/7)^2-(3/7)^2 = 1-0.326530612-0.183673469 = 0.489795918$

$GINI(<80) = 1-(0/3)^2-(3/3)^2 = 1-0-1 = 0$

$GINI_{split} = (3/10)*0+(7/10)* 0.489795918$ **$= 0.342857143$**

Pregunta #3 ¿Cuándo dejar de dividir?

- Esta es una difícil ya que implica sutil selección de modelos.
- Una idea sería controlar el *Error de Clasificación* (o el *Índice de Gini* o la *Entropía*) en el conjunto de datos de prueba de manera que se detendrá cuando el índice selecciona comience a aumentar.
- La "Poda" (pruning) es la técnica más popular. Usada en el Método CART propuesto por Breiman, Friedman, Olshen, and Stone, 1984, (CART=Classification And Regression Trees)

Algoritmo CART

Para cada nodo ***v*** del Árbol hacer los pasos 1 y 2

1. Para *j= 1,2,…,p* calcular: (*p*=número de variables)
   - Todas las divisiones binarias correspondientes a la variable discriminante ***Y***
   - La división binaria óptima *d(j)* correspondiente a la variable ***Y***, es decir la división binaria maximiza el descenso de la impureza
2. Recursivamente calcular la mejor división binaria para *d(1), d(2), …,d(p)*

FIN

Árbol podado y reestructurado

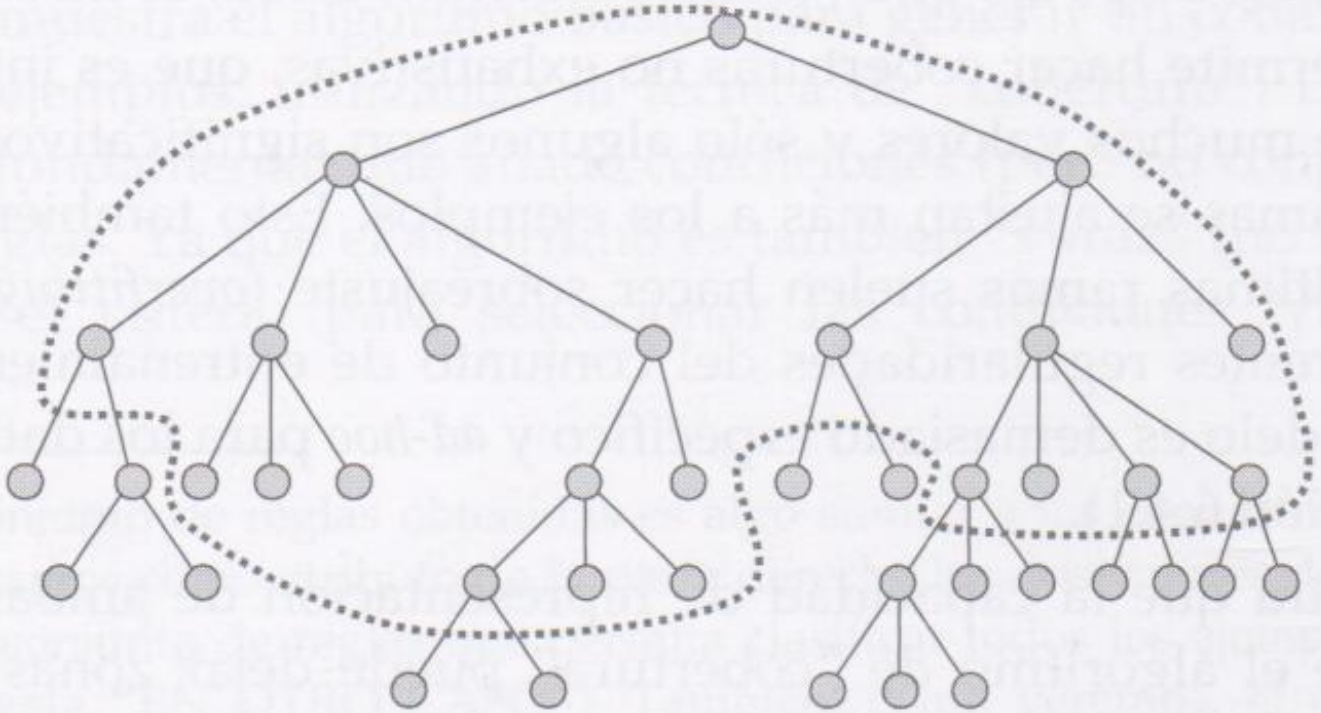
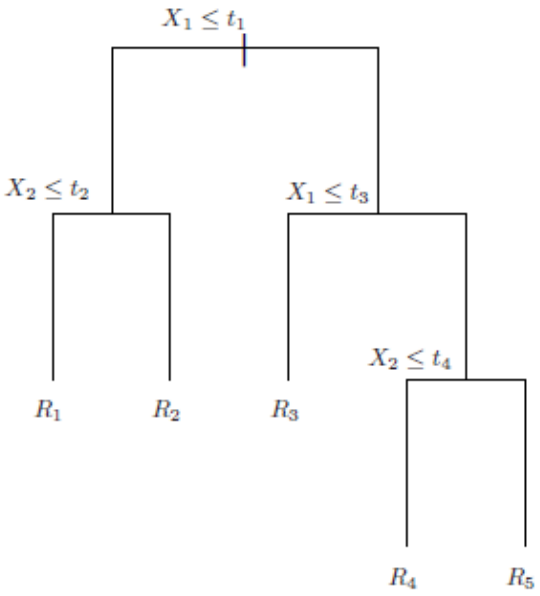
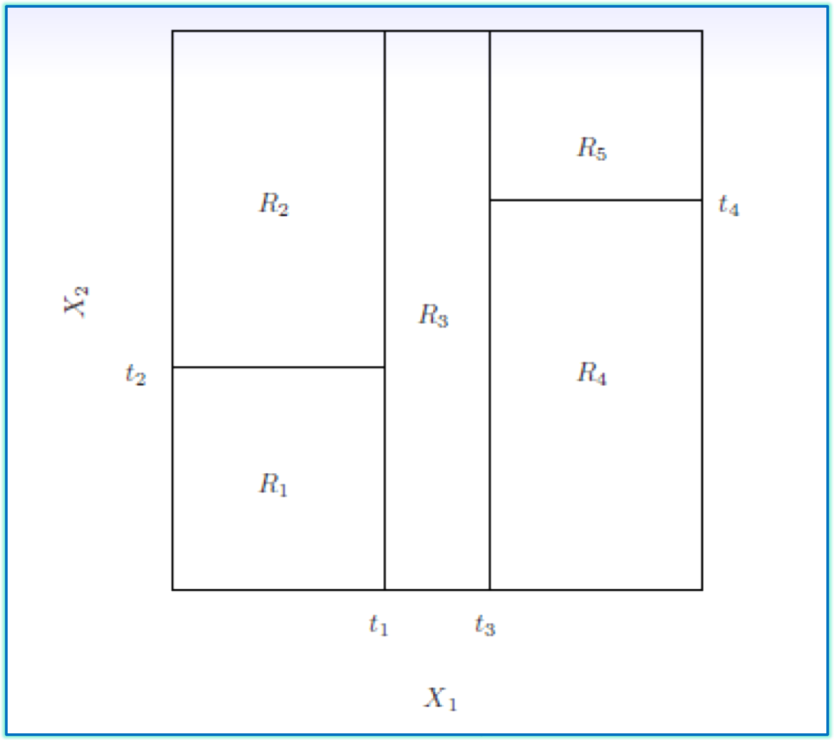
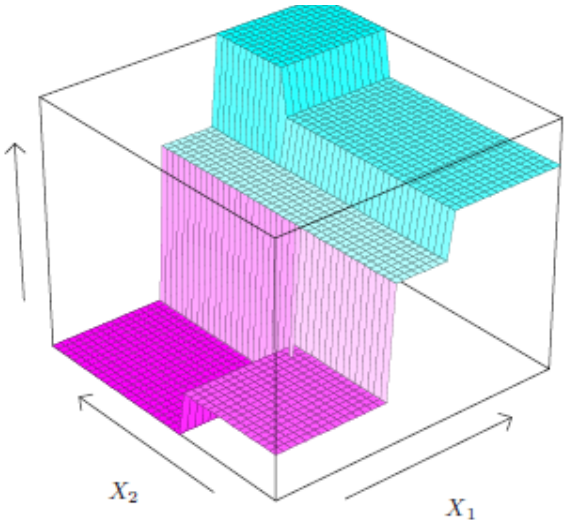


Figura 11.8. Ejemplo de poda. Algunos nodos inferiores son eliminados.



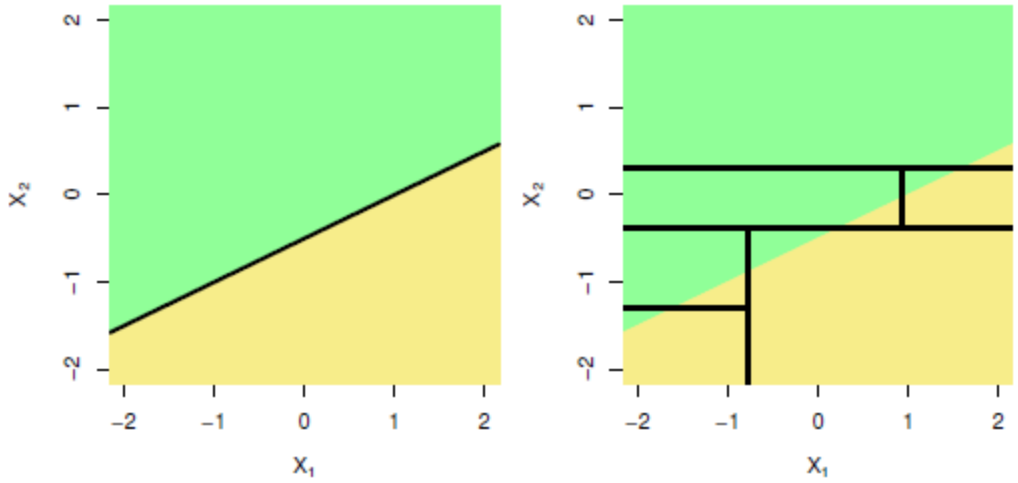




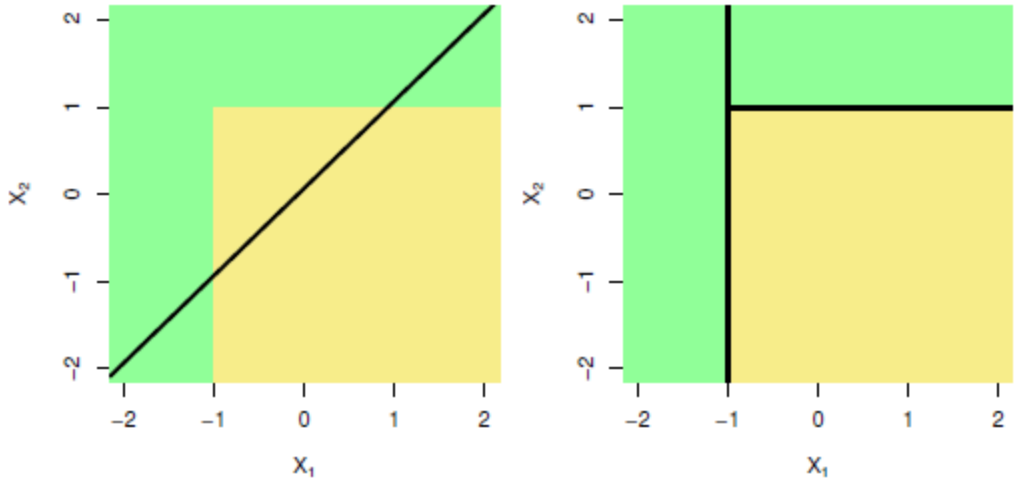
El ***Diagrama de Voronoi*** define las fronteras de la clasificación

Árboles versus Modelos Lineales

Mejor Modelo Lineal



Mejor un Árbol o un Bosque Aleatorio



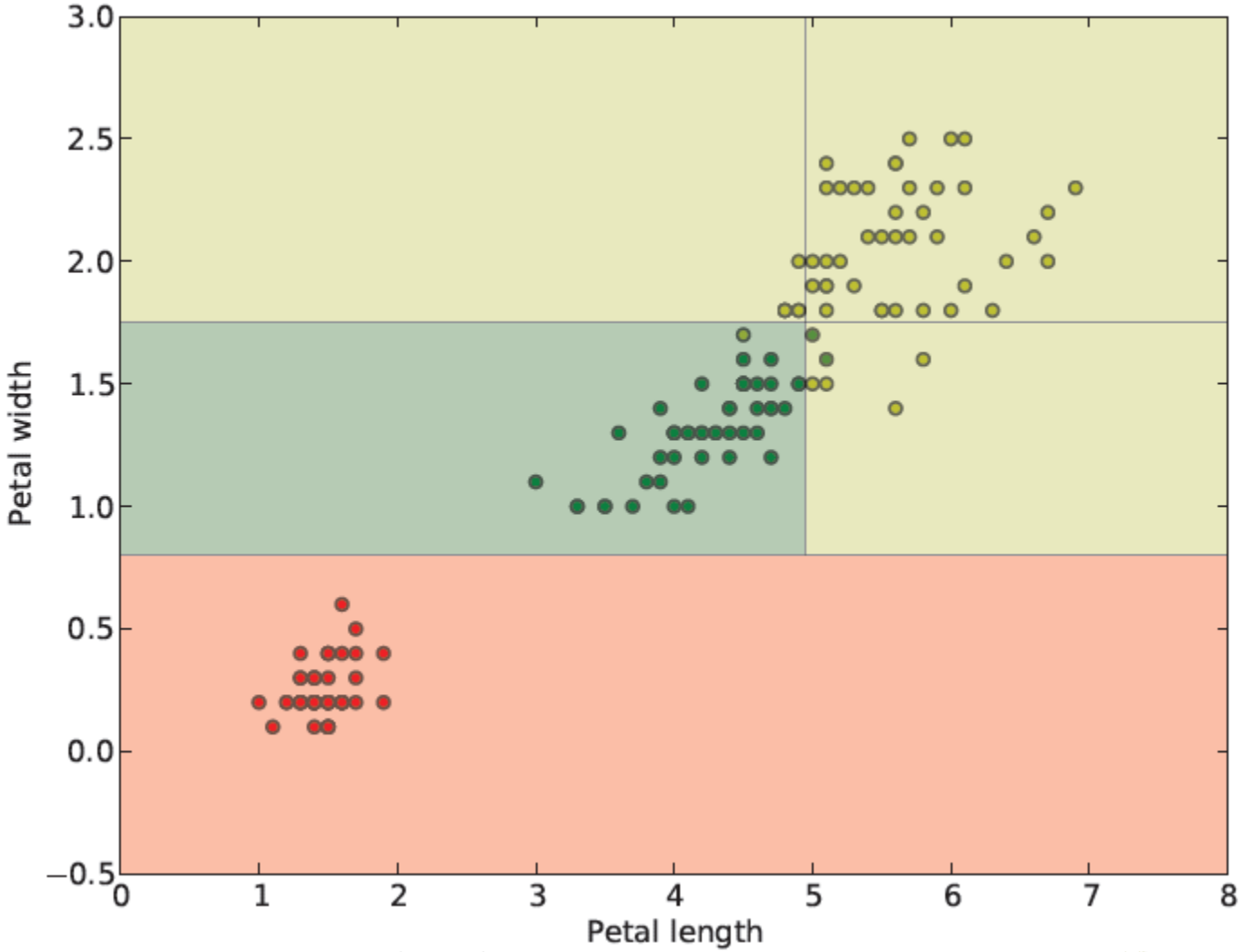
Ejemplo 1: IRIS.CSV

Ejemplo con la tabla de datos IRIS
IRIS Información de variables:
1.sepal largo en cm
2.sepal ancho en cm
3.petal largo en cm
4.petal ancho en cm
5.clase:

- Iris Setosa
- Iris Versicolor
- Iris Virginica

	A	B	C	D	E
1	s.largo	s.ancho	p.largo	p.ancho	tipo
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3.0	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5.0	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5.0	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3.0	1.4	0.1	setosa
15	4.3	3.0	1.1	0.1	setosa
16	5.8	4.0	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1.0	0.2	setosa
25	....	...	...	...	....

Ejemplo 1: iris.csv



Ejemplo 2: Credit-Scoring

MuestraAprendizajeCredito2500.csv
MuestraTestCredito2500.csv

```
> setwd("C:/Users/Oldemar/Google Drive/Curso Mineria Datos II - Optativo/Datos")
> taprendizaje<-read.csv("MuestraAprendizajeCredito2500.csv",sep = ";",header=T)
> taprendizaje
```

	MontoCredito	IngresoNeto	CoefCreditoAvaluo	MontoCuota	GradoAcademico	BuenPagador
1	1	1	1	1	1	Si
2	3	1	1	1	1	Si
3	2	1	1	1	1	Si
4	1	2	1	1	1	Si
5	1	1	1	1	1	Si
6	2	1	1	1	1	Si
7	4	1	1	1	1	Si
8	1	2	1	1	1	Si
9	1	2	1	1	1	Si
10	3	2	1	1	1	Si
11	1	1	1	1	1	Si
12	1	2	1	1	1	Si
13	3	1	1	1	1	Si
14	3	1	1	1	1	Si
15	2	1	1	1	1	Si
16	3	1	1	1	1	Si
17	3	1	1	1	1	Si

Descripción de Variables

MontoCredito

1=Muy Bajo
2=Bajo
3=Medio
4=Alto

MontoCuota

1=Muy Bajo
2=Bajo
3=Medio
4=Alto

IngresoNeto

1=Muy Bajo
2=Bajo
3=Medio
4=Alto

GradoAcademico

1=Bachiller
2=Licenciatura
3=Maestría
4=Doctorado

CoeficienteCreditoAvaluo

1=Muy Bajo
2=Bajo
3=Medio
4=Alto

BuenPagador

1=NO
2=Si

Árboles de Decisión en R

Package 'rpart'

June 27, 2012

Priority recommended

Version 3.1-54

Date 2012-06-27

DateNote March 2002 version of rpart

Author Terry M Therneau and Beth Atkinson <atkinson@mayo.edu>. R port by Brian Ripley. Note that maintainers are not available to give advice on using a package they did not author.

Maintainer Brian Ripley <ripley@stats.ox.ac.uk>

Description Recursive partitioning and regression trees

Title Recursive Partitioning

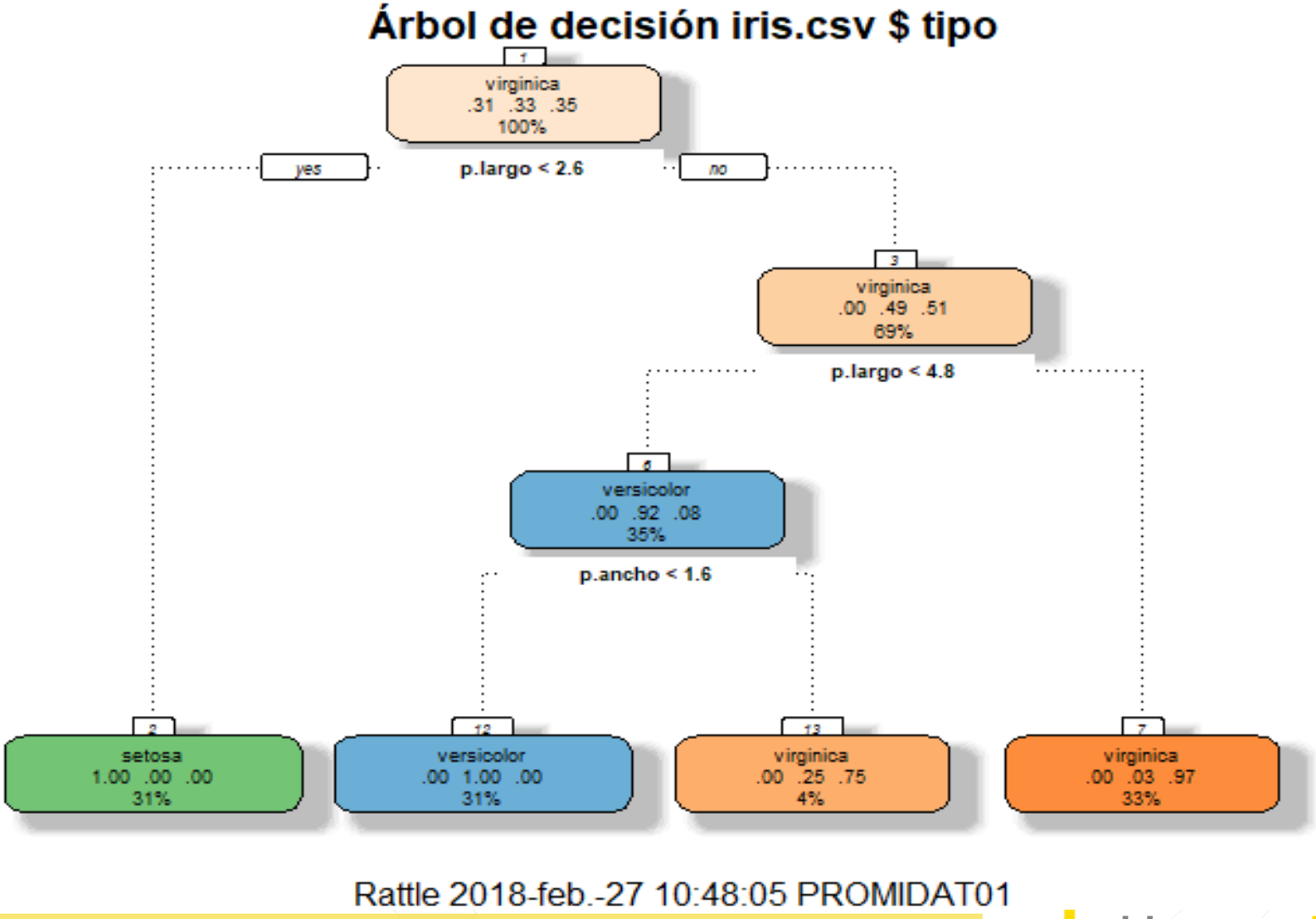
Depends R (>= 2.14.0), graphics, stats, grDevices

El paquete “rpart” utiliza el algoritmo CART + Pruning (poda)

Instalando y usando el paquete “rpart”:

- `install.packages('rpart',dependencies=TRUE)`
- `library(rpart)`

Reglas

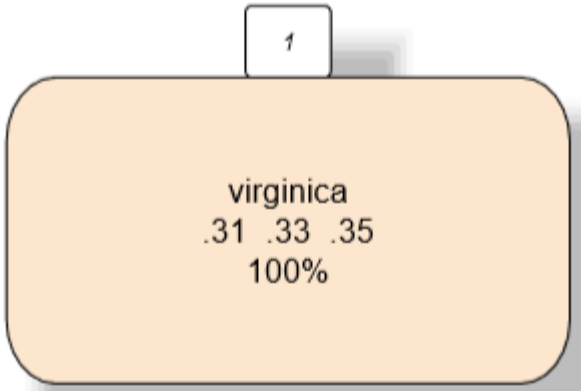


Reglas

Primero veamos como interpretar el árbol de decisión que nos genera rattle con una división mínima de 10 y todos los demás parámetros por defecto:

El primer nodo es el nodo raíz en el cual se encuentran el 100% de los datos, esto es a lo que se le llama el cover.

Dado que tomamos el 70% de los datos para la tabla de training entonces contamos con 105 individuos de los 150 totales, de los cuales el 31% (33 datos) son setosas, el 33% (35 datos) son versicolor y el 35% (37 datos) virginica. Esto es a lo que llamamos la probabilidad de acierto, que suman entre todas 100% aproximadamente.

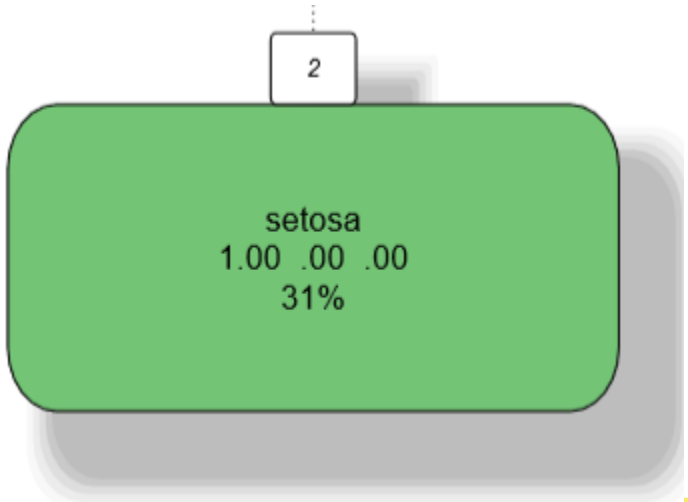


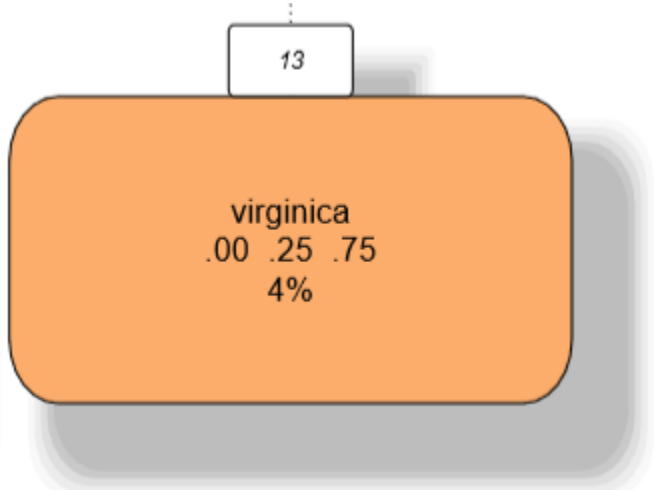
Reglas

Ahora pasemos a la regla número 2; en esta se tiene un cover que representa un 31% del total de los datos, de los cuales el 100% se clasifican como setosas, el 0% como versicolor y el 0% como virginica.

La regla número 13 por otro lado tiene un 4% de los datos de los cuales la probabilidad de que sean setosa es del 0%, versicolor del 25% y virginica del 75%.

El orden de cual probabilidad corresponde a que categoría se puede ver en la matriz de confusión o en los niveles de la variable.



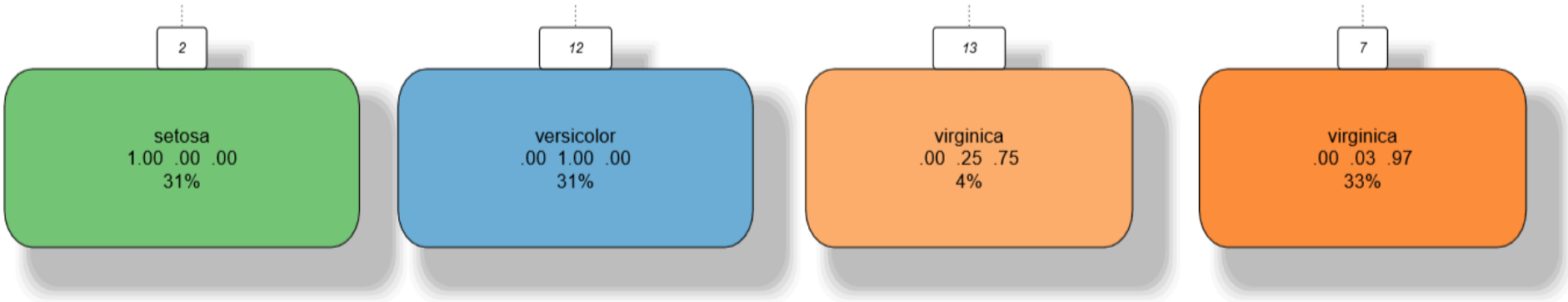


¿Cuál es la mejor *Regla*?

Decimos que una regla es buena si tiene una probabilidad de acierto alta junto con un cover alto, por lo tanto una regla es mejor que otra en cuanto su cover y probabilidades sean más altas que las de la otra regla.

Además, en nuestro caso solo consideramos los nodos terminales para la elección de las mejores reglas, por ejemplo no tomamos en cuenta los nodos 1-3-6.

Para nuestro caso particular podemos ver que la regla 2-12-7 son mejores que la 13, pues esta última tiene un cover muy bajo.



Gracias....

oldemar rodríguez

CONSULTOR en M1NƐR14 DƐ D4T0S