

ÁRBOLES DE DECISIÓN
APRENDIZAJE SUPERVISADO

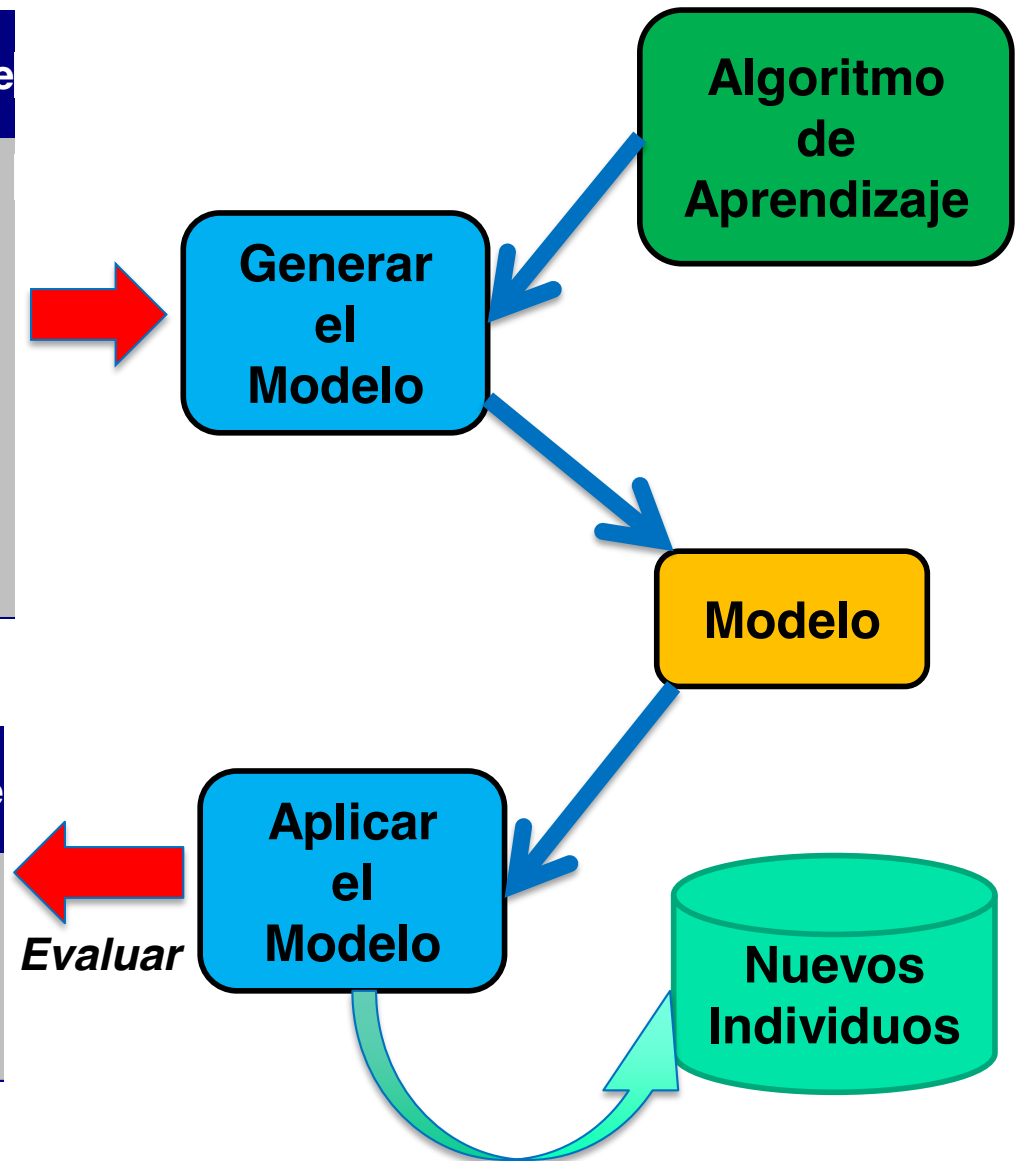
Modelo general de los métodos de Clasificación

<i>Id</i>	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

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

Tabla de Testing



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 $f: D \rightarrow C$ tal que cada t_i es asignada una clase C_j .
- $f: D \rightarrow C$ podría ser una Red Neuronal, un Árbol de Decisión, un modelo basado en Análisis Discriminante, o una Red Bayesiana.

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	CoficienteCre...	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

categorica

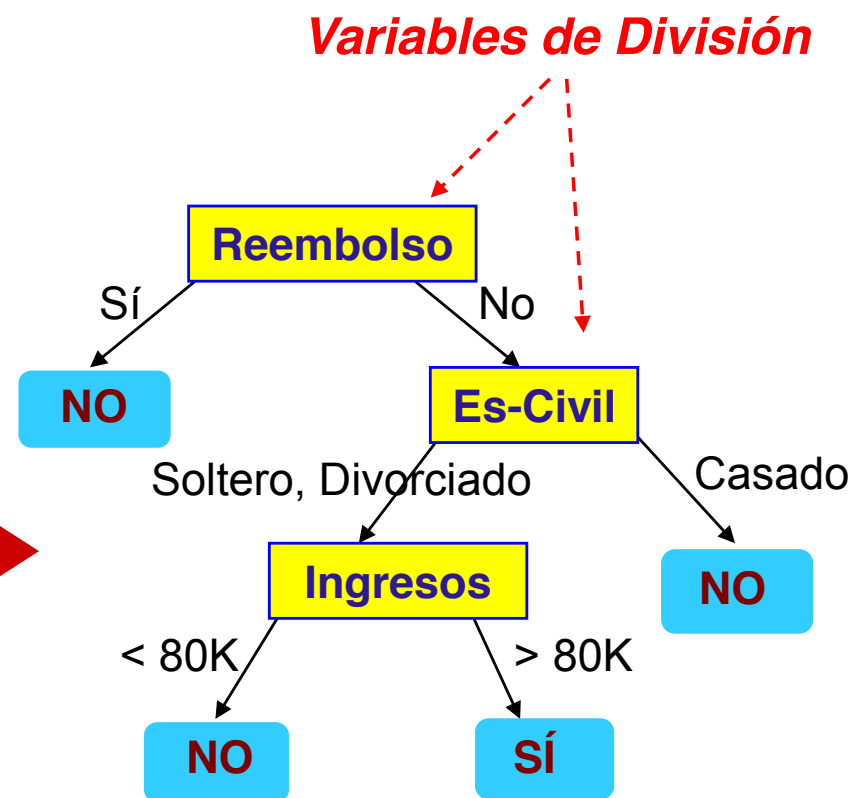
categorica

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	No	Casado	75K	No
10	No	Soltero	90K	Sí

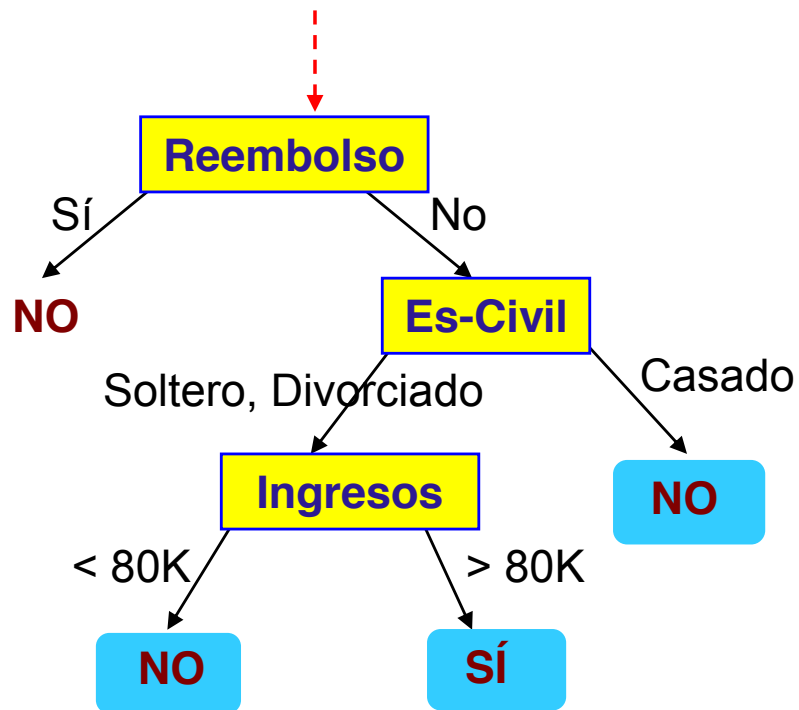
Tabla de Aprendizaje



Modelo: Árbol de Decisión

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

Inicia desde la raíz del árbol



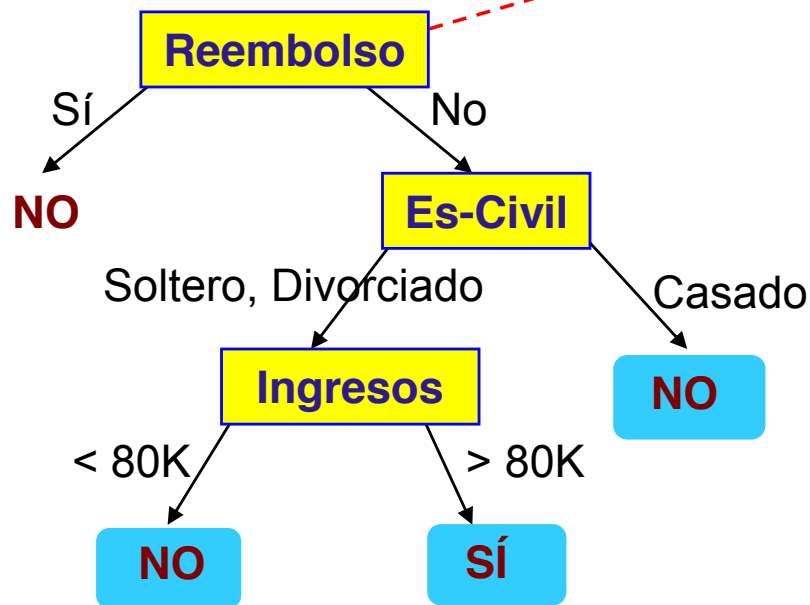
Datos de Prueba

Reembolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?

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

Datos de Prueba

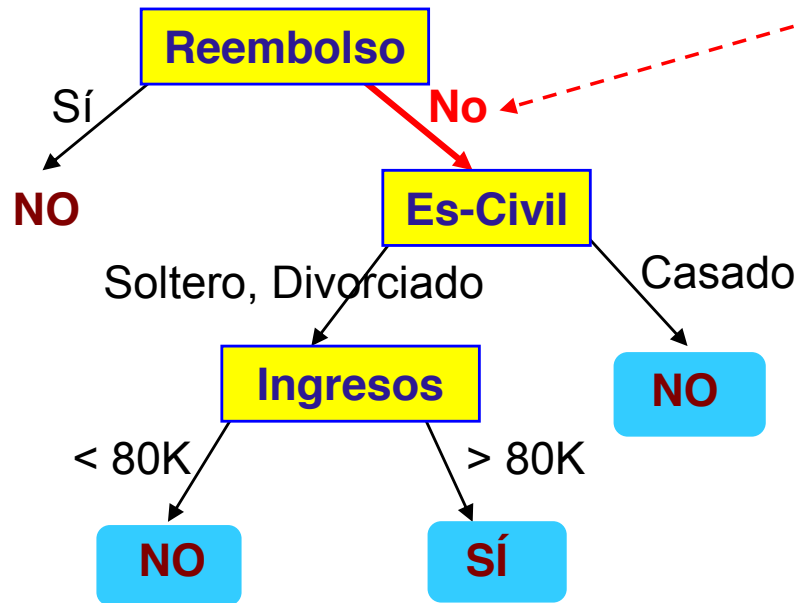
Reembolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?



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

Datos de Prueba

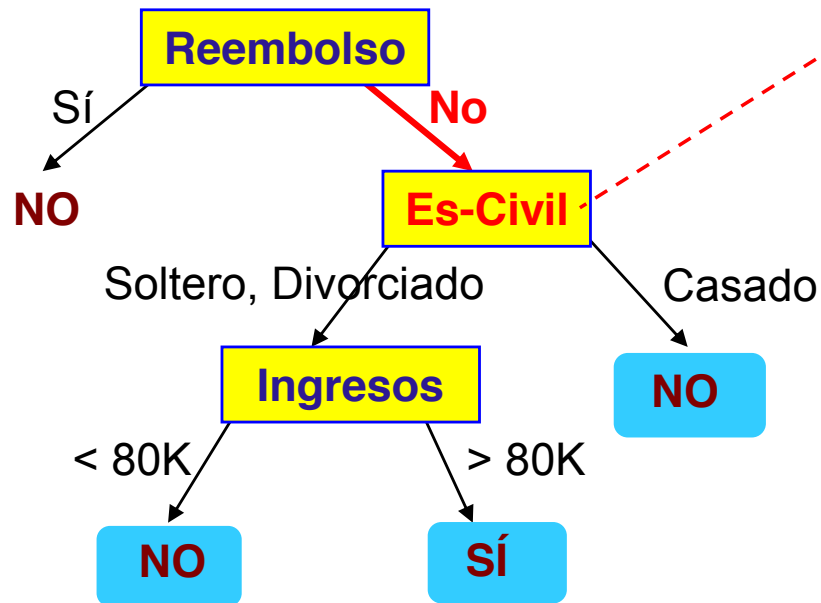
Reembolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?



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

Datos de Prueba

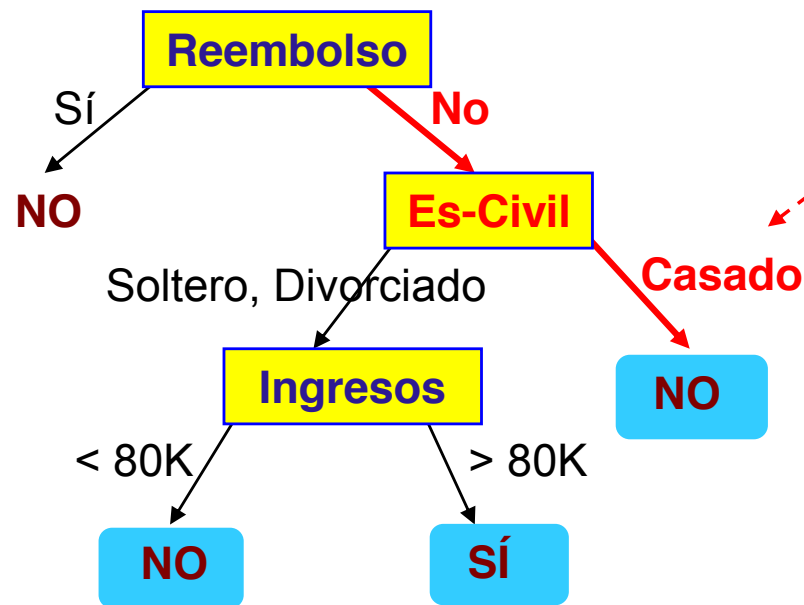
Reembolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?



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

Datos de Prueba

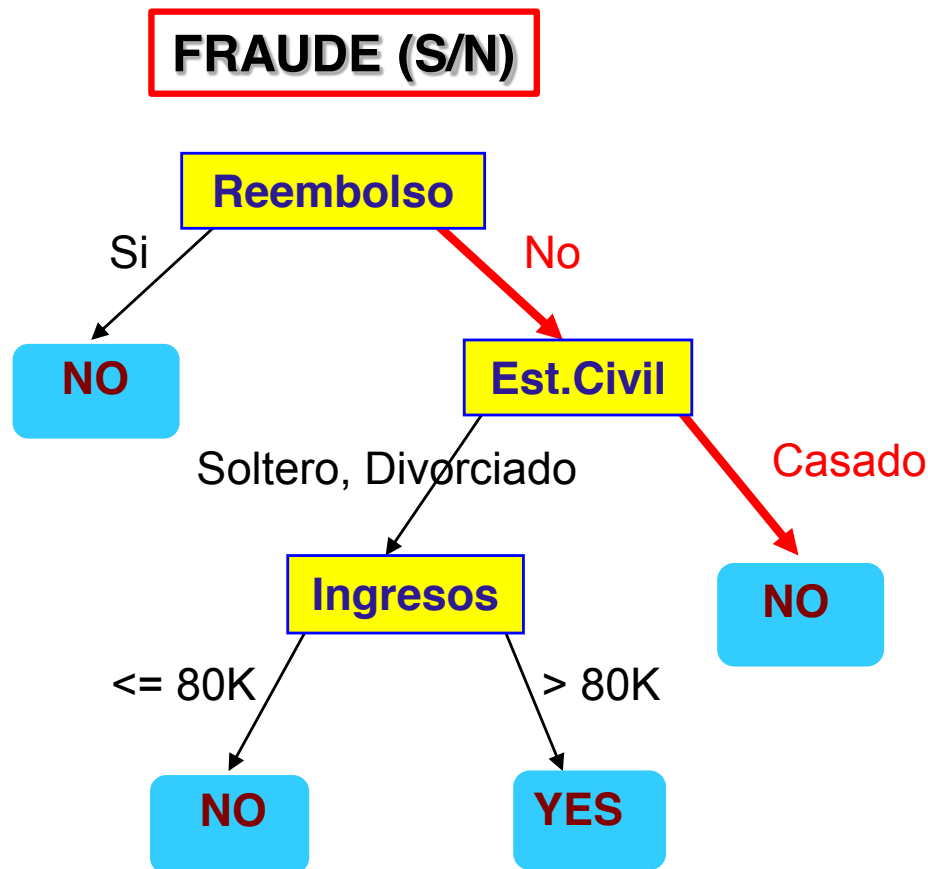
Reembolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?



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

Datos de Prueba

Reembolso	Estado Civil	Ingresos	Fraude
No	Casado	80K	?



Asigna "No" a la variable de Clase "Fraude"

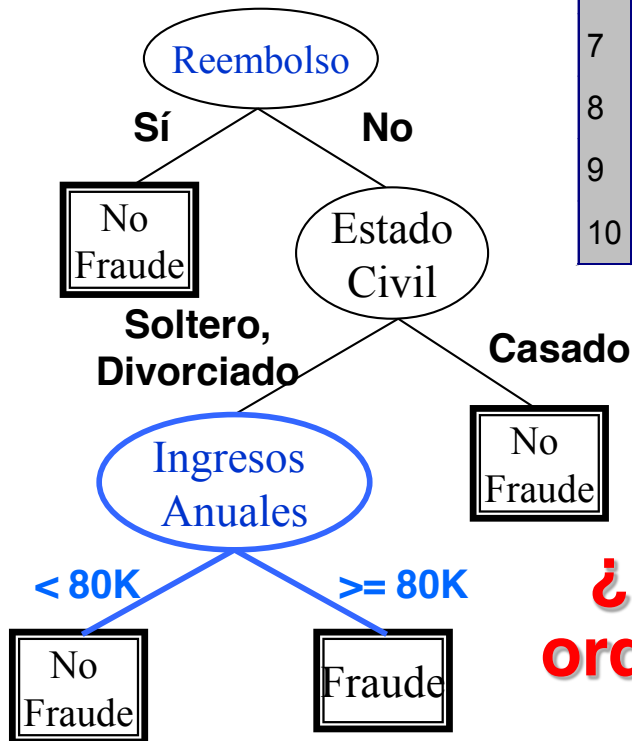
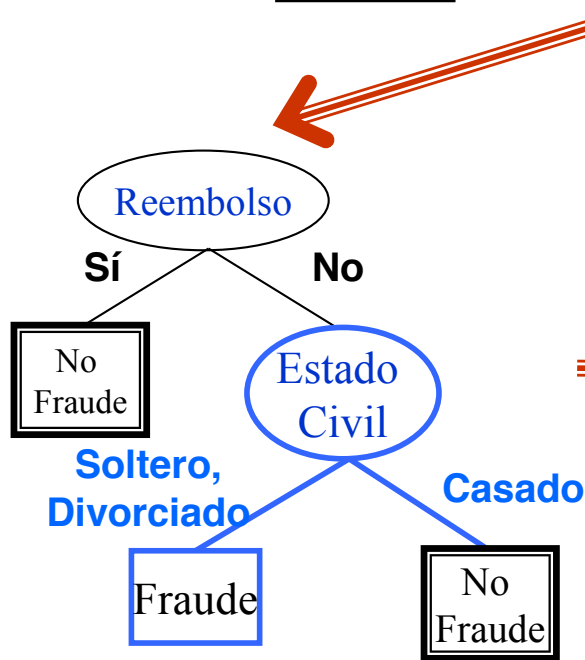
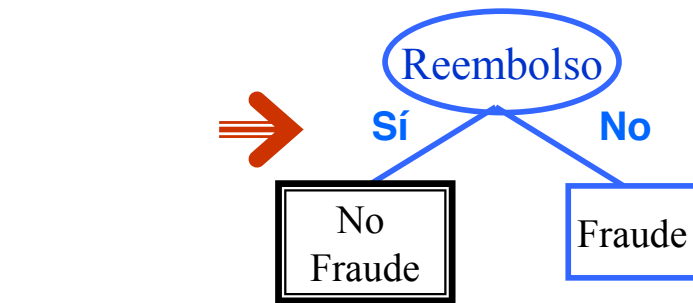
¿Cómo se generan los árboles de decisión?

- Muchos algoritmos usan una versión con un enfoque "top-down" o "dividir y conquistar" conocido como Algoritmo de Hunt.
- Sea D_t el conjunto de registros de entrenamiento en un nodo t dado.
- Sea $y_t = \{y_1, y_2, \dots, y_c\}$ 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	No			
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 enfoque computacional es eficiente por lo que es muy popular.

¿Cómo aplicar el algoritmo de Hunt?

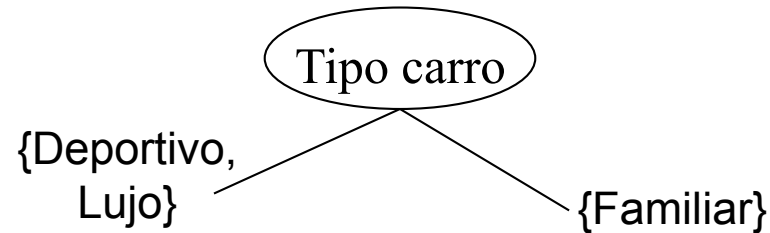
- ✓ Utilizando el enfoque de optimización aún se tienen que decidir tres cosas:
 1. ¿Cómo dividiremos las variables?
 2. ¿Qué variables (atributos) utilizar y en que orden? ¿Qué criterio utilizar para seleccionar la "mejor" división?
 3. ¿Cuándo dejar de dividir? Es decir, ¿Cuándo termina el algoritmo?

¿Cómo aplicar el algoritmo de Hunt?

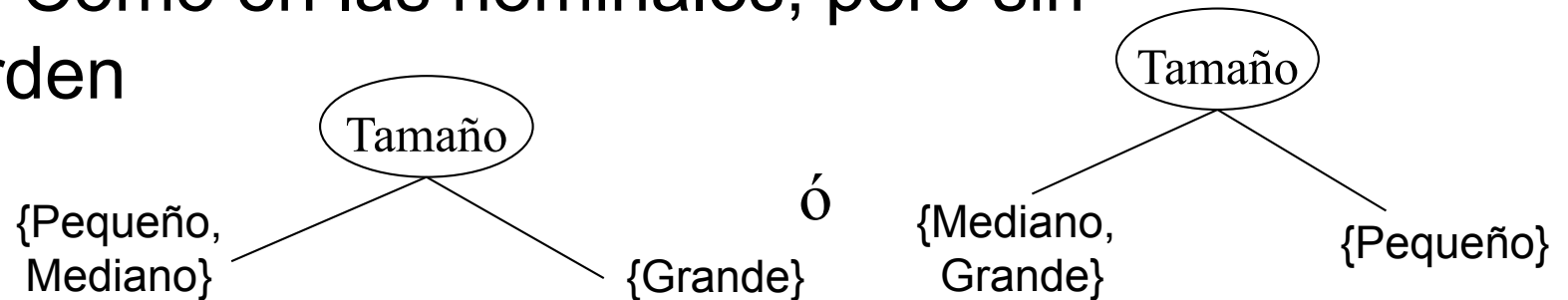
- Para la pregunta 1, se tendrán en cuenta sólo divisiones binarias tanto para predictores numéricos como para los categóricos, esto se explica más adelante (Método CART).
- Para la pregunta 2 se considerarán el *Error de Clasificación*, el *Índice de Gini* y la *Entropía*.
- La pregunta 3 tiene una 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):

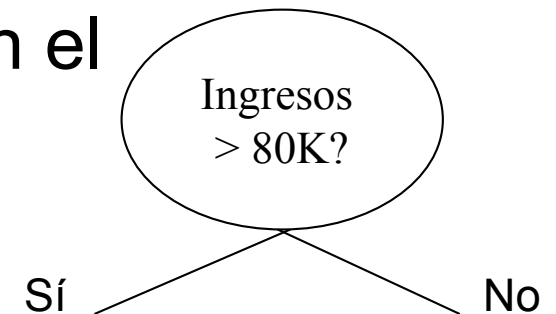
Nominales:



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 probabilidad de pertenecer a la clase “j” estando en el nodo t.
- Muchas veces simplemente se usa p_j

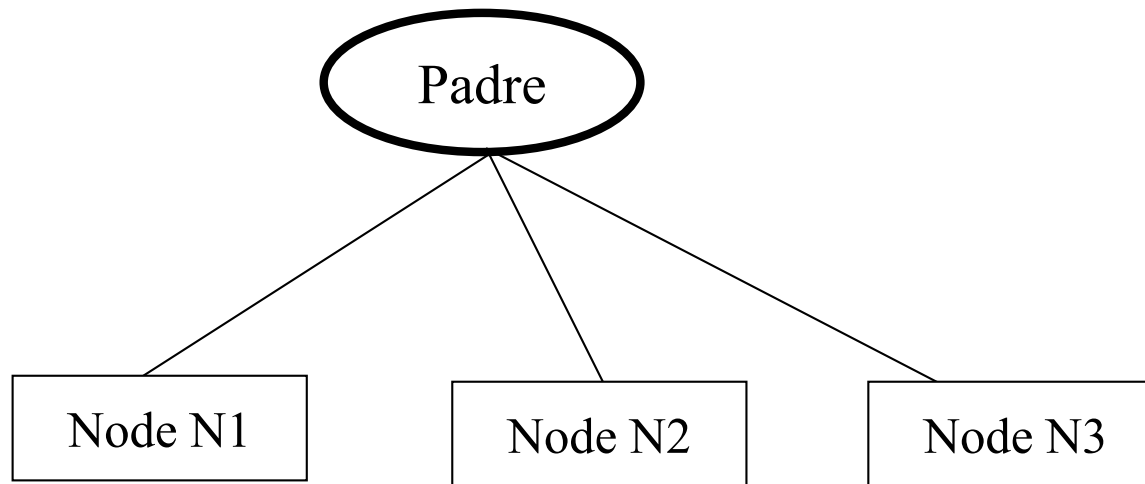
Pregunta #2: Se usarán el *Error de Clasificación*, el Índice de *Gini* y la *Entropía*

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

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

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

Ejemplo de cálculo de índices:



	N1	N2	N3
C1	0	1	2
C2	6	5	4

Ejemplo de cálculo de *Gini*

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

C1	0
C2	6

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

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

C1	1
C2	5

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

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

C1	2
C2	4

$$P(C1) = 2/6 \quad P(C2) = 4/6$$

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

Ejemplo de cálculo de la *Entropía*

C1	0
C2	6

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

$$\text{Entropía} = -0 \log 0 - 1 \log 1 = -0 - 0 = 0$$

C1	1
C2	5

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

$$\text{Entropía} = - (1/6) \log_2 (1/6) - (5/6) \log_2 (5/6) = 0.65$$

C1	2
C2	4

$$P(C1) = 2/6 \quad P(C2) = 4/6$$

$$\text{Entropía} = - (2/6) \log_2 (2/6) - (4/6) \log_2 (4/6) = 0.92$$

Ejemplo de cálculo del *Error de Clasificación*

C1	0
C2	6

$$\text{Error Clasificación} = 1 - \max[0/6, 6/6] = 0$$

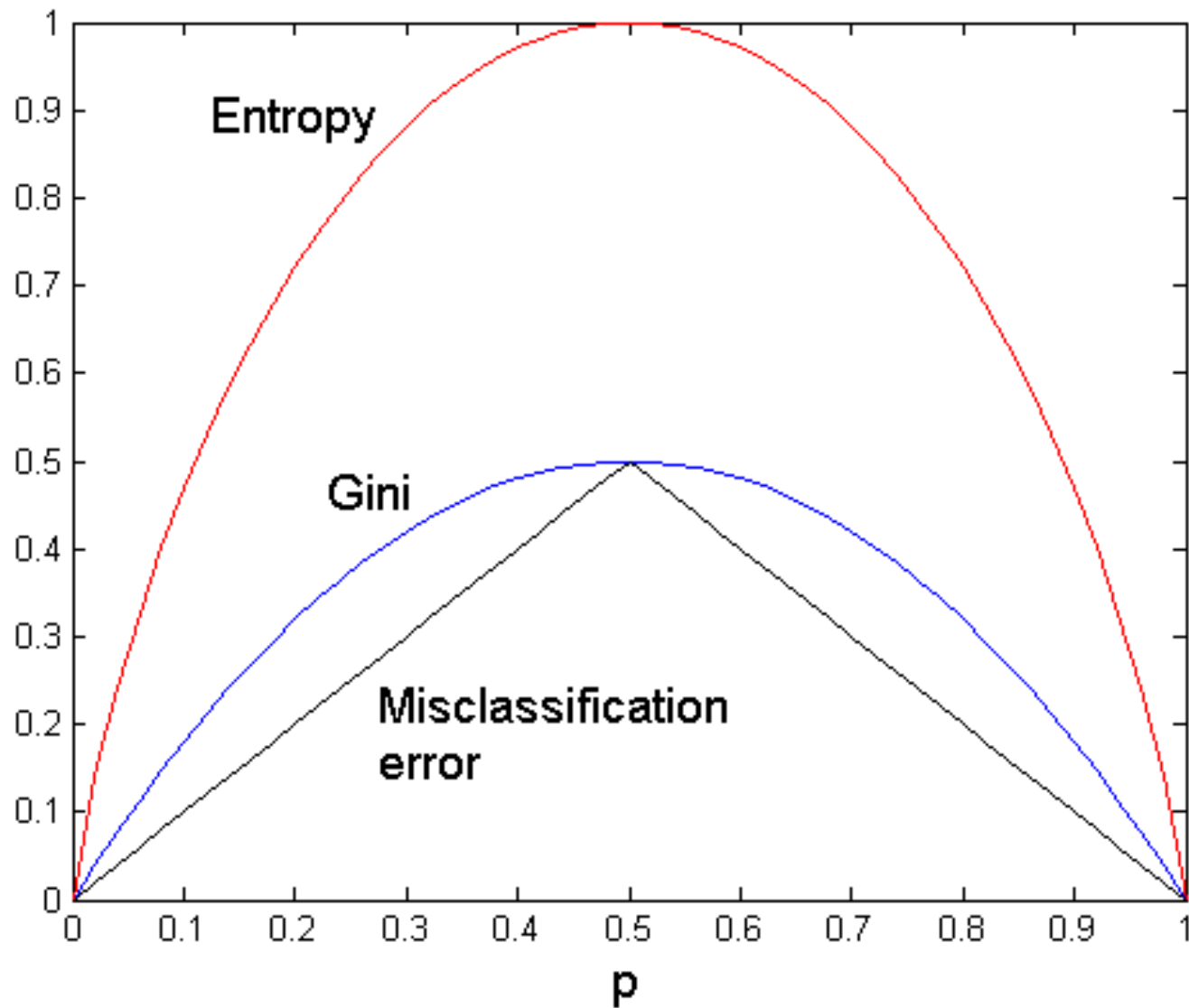
C1	1
C2	5

$$\text{Error Clasificación} = 1 - \max[1/6, 5/6] = 0,167$$

C1	2
C2	4

$$\text{Error Clasificación} = 1 - \max[2/6, 4/6] = 0,333$$

Comparación Gráfica



Gini Split

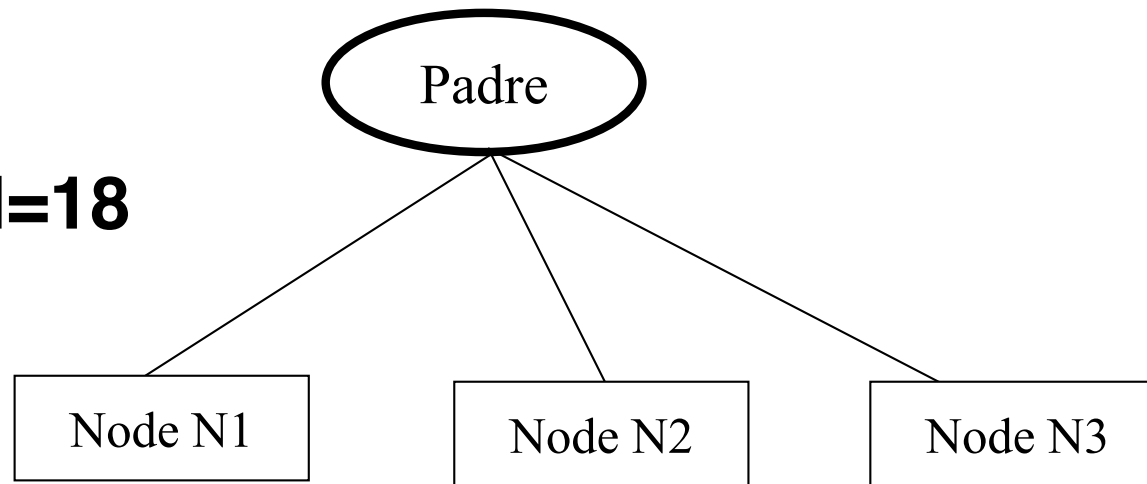
- ❖ Después de que el índice de Gini se calcula en cada nodo, el valor total del índice de Gini se calcula como el promedio ponderado del índice de Gini en cada nodo:

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

Ejemplo de cálculo $GINI_{Split}$

$I_{padre}=18$

Cardinalidad=18



	N1	N2	N3
C1	0	1	2
C2	6	5	4

Ejemplo de cálculo de $GINI_{split}$

C1	0
C2	6

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

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

C1	1
C2	5

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

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

C1	2
C2	4

$$P(C1) = 2/6 \quad P(C2) = 4/6$$

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

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

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:

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

Información Ganada $\rightarrow IG_{\text{Split}}$

- ✓ 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.

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

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

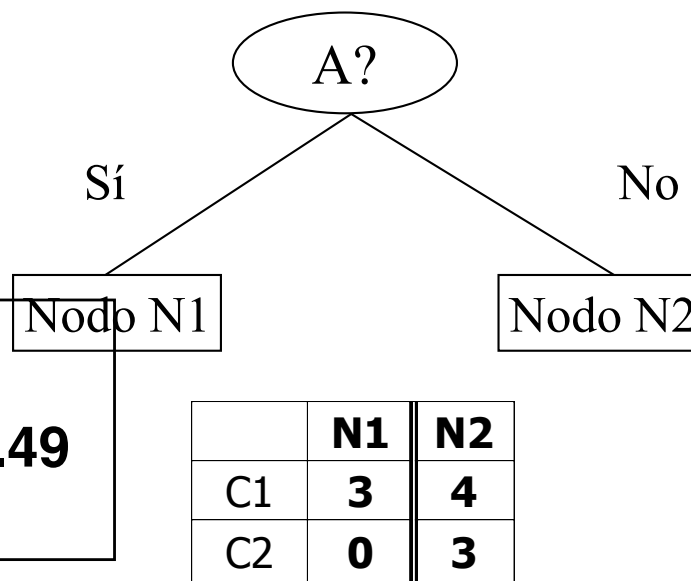
Ejemplo: Información Ganada $\rightarrow IG_{Split}$

	Padre
C1	7
C2	3
Gini = 0.42	

$$\begin{aligned} \text{Gini}(N1) &= 1 - (3/3)^2 - (0/3)^2 \\ &= 0 \end{aligned}$$

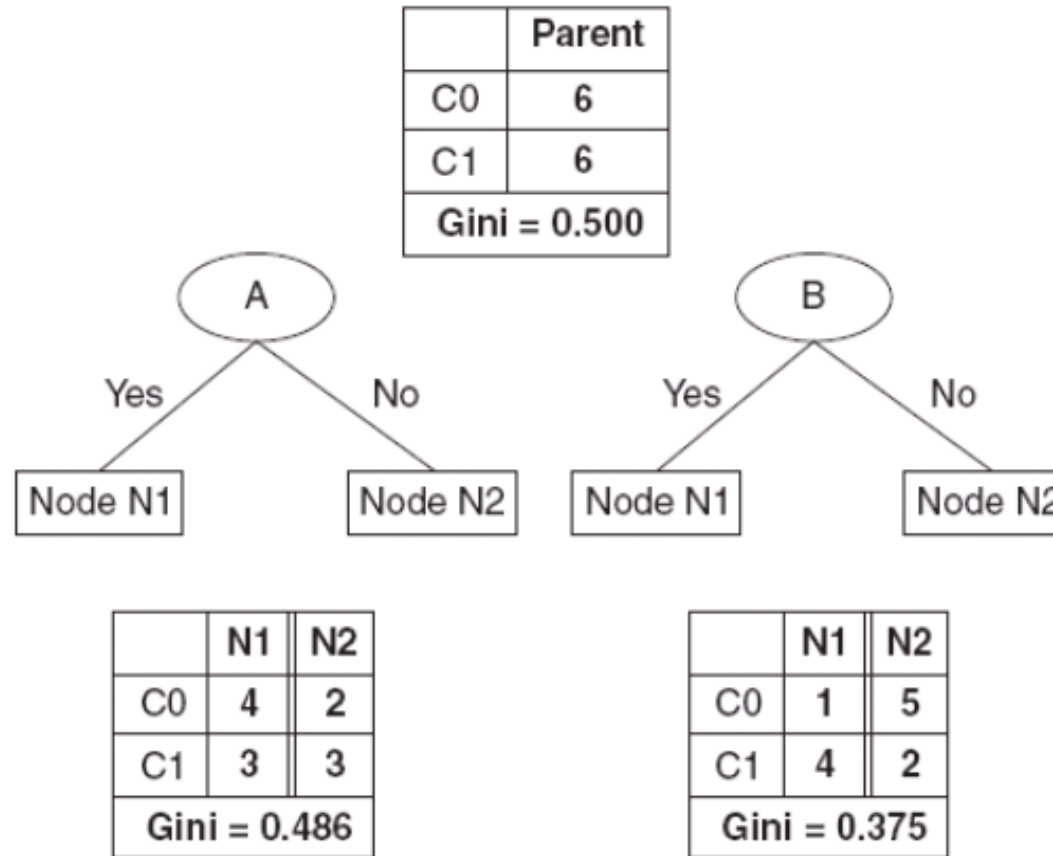
$$\begin{aligned} \text{Gini}(N2) &= 1 - (4/7)^2 - (3/7)^2 \\ &= 0.490 \end{aligned}$$

$$\begin{aligned} \text{GINI}_{split} &= \\ &= 3/10 * 0 + 7/10 * 0.49 \\ &= 0.343 \end{aligned}$$

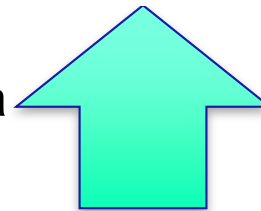


$$IG_{split} = I(padre) - \left(\sum_{i=1}^k \frac{n_i}{n} I(i) \right) = 0.42 - 0.343 = 0.077$$

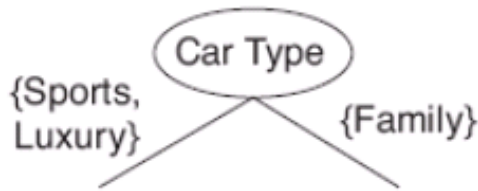
¿Cómo escoger la mejor división?



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

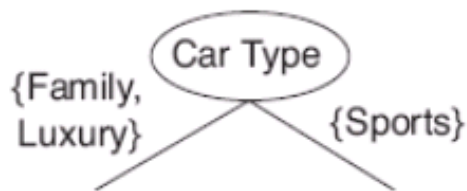


¿Cómo escoger la mejor división?

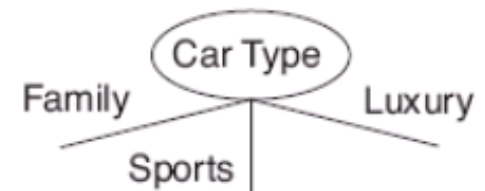


	Car Type	
	{Sports, Luxury}	{Family}
C0	9	1
C1	7	3
Gini	0.468	

(a) Binary split



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



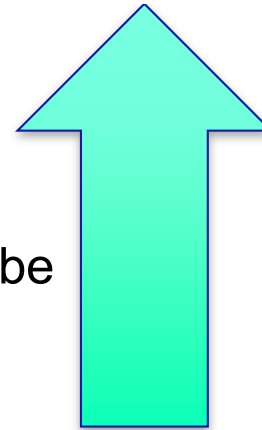
	Car Type		
	Family	Sports	Luxury
C0	1	8	1
C1	3	0	7
Gini	0.163		

(b) Multiway split

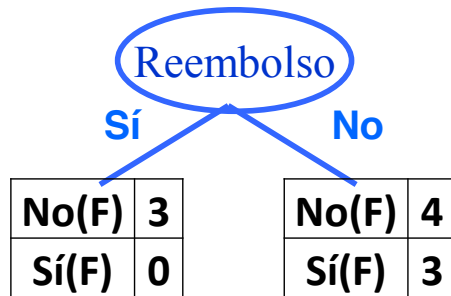
¿Cómo escoger la mejor división?

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			
		55		65		72		80		87		92		97		110		122		172		230	
Split Positions →		≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
	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		<u>0.300</u>		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}**



¿Porqué se escoge “Reembolso” como variable inicial?



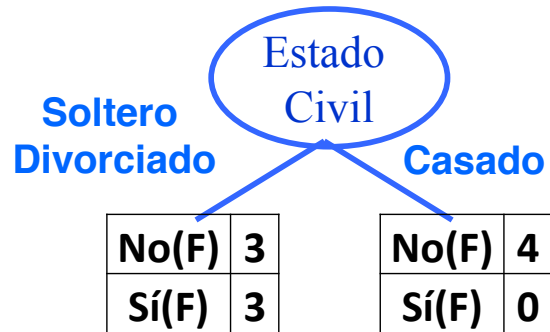
$$\text{GINI}(\text{No}) = 1 - (4/7)^2 - (3/7)^2 = 1 - 0.326 - 0.184 = 0.142$$

$$\text{GINI}(\text{Sí}) = 1 - (0/3)^2 - (3/3)^2 = 1 - 0 - 1 = 0$$

$$\text{GINI}_{\text{split}} = (3/10) * 0 + (7/10) * 0.142 = \mathbf{0.0994}$$

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	No	Casado	75K	No
10	No	Soltero	90K	Sí

¿Porqué se escoge “Reembolso” como variable inicial?



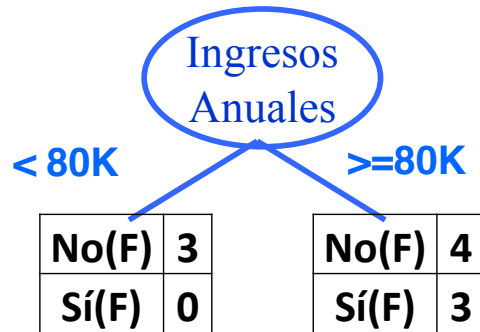
Id	Reembolso	Estado Civil	Ingresos Anuales	Fraude
1	Sí	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No
7	Sí	Divorciado	220K	No
8	No	Soltero	85K	Sí
9	No	Casado	75K	No
10	No	Soltero	90K	Sí

$$\text{GINI}(\text{Casado}) = 1 - (0/4)^2 - (4/4)^2 = 0$$

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

$$\text{GINI}_{\text{split}} = (6/10) * 0.5 + (4/10) * 0 = \mathbf{0.3}$$

¿Porqué se escoge “Reembolso” como variable inicial?



$$\text{GINI}(\geq 80K) = 1 - (3/7)^2 - (4/7)^2 = 0.49$$

$$\text{GINI}(< 80K) = 1 - (0/3)^2 - (3/3)^2 = 1 - 0 - 1 = 0$$

$$\text{GINI}_{\text{split}} = (3/10) * 0 + (7/10) * 0.49 = \mathbf{0.343}$$

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	No	Casado	75K	No
10	No	Soltero	90K	Sí

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)

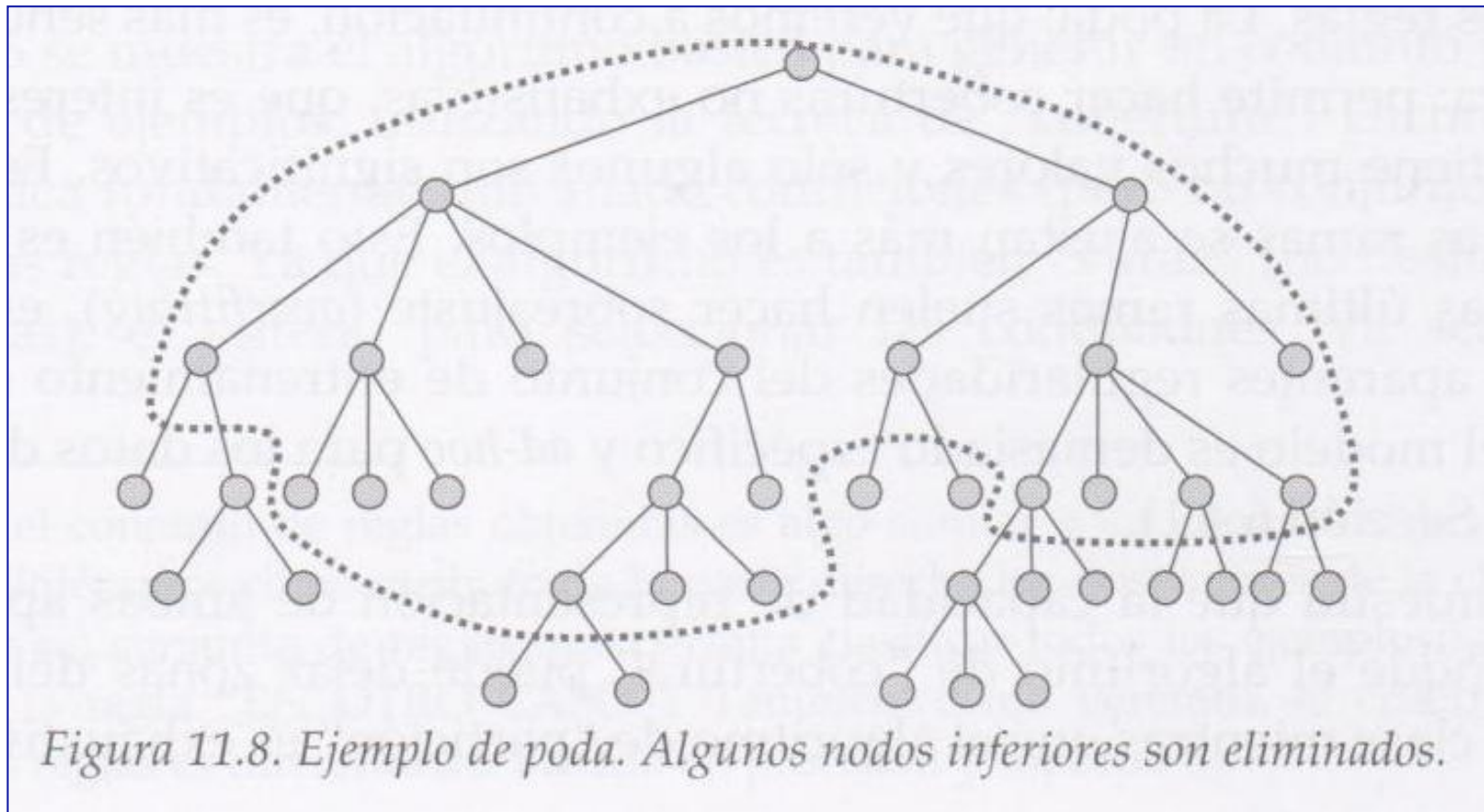
Algoritmo CART

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

1. Para $j= 1, 2, \dots, p$ 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 $d(1), d(2), \dots, d(p)$

FIN

Árbol podado y reestructurado



Árboles de Decisión en Rattle

> library(rattle)

Rattle: A free graphical interface for data mining with R.
Versión 2.6.21 Copyright (c) 2006-2012 Togaware Pty Ltd.
Escriba 'rattle()' para agitar, sacudir y rotar sus datos.

> rattle()

Ejemplo 1: IRIS.CSV

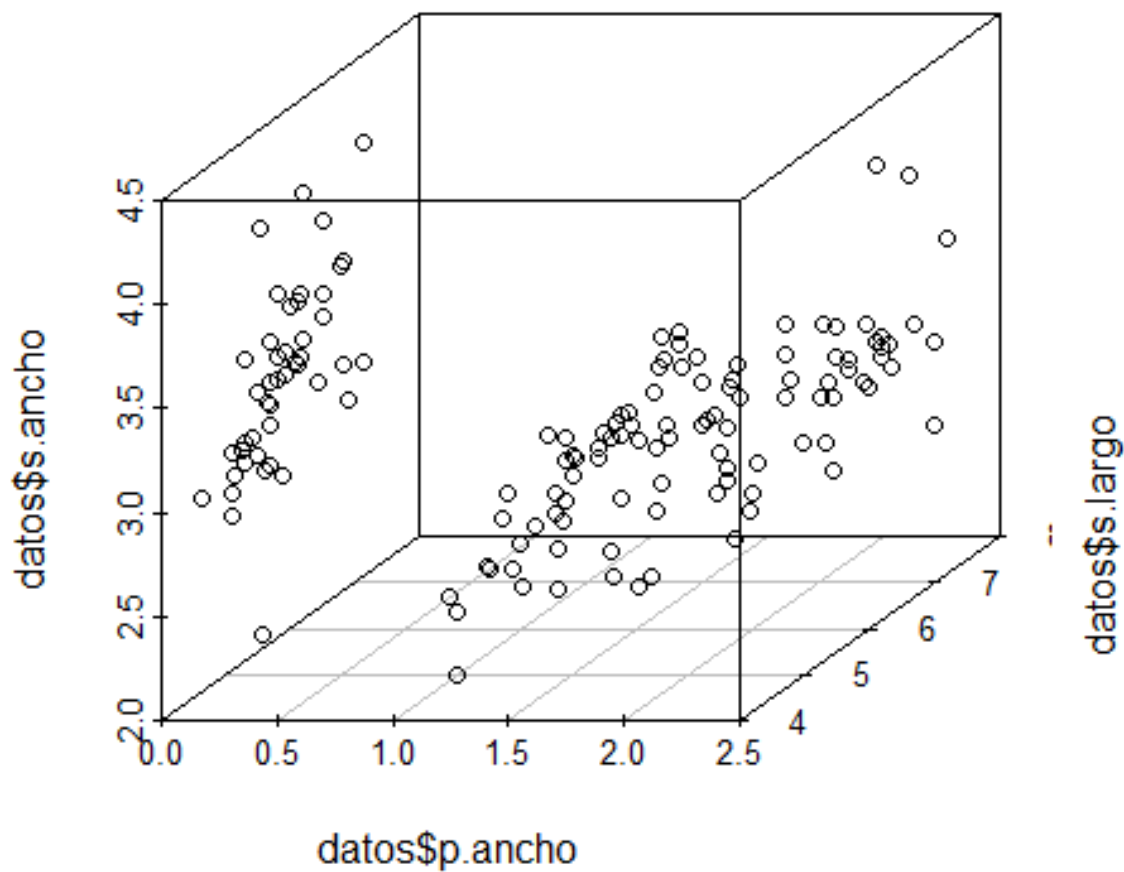
Ejemplo con la tabla de datos IRIS

IRIS Información de variables:

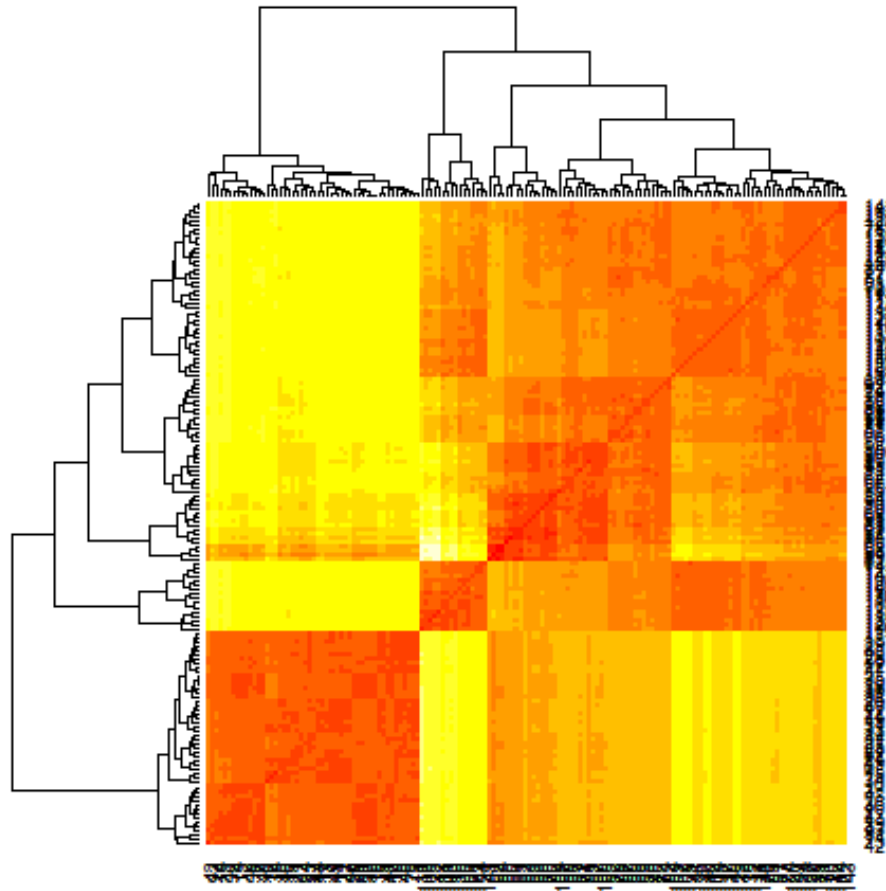
- 1.sepal largo en cm
- 2.sepal ancho en cm
- 3.petal largo en cm
- 4.petal ancho en cm
- 5.clase:
 - Iris Setosa
 - Iris Versicolor
 - Iris Virginica



	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

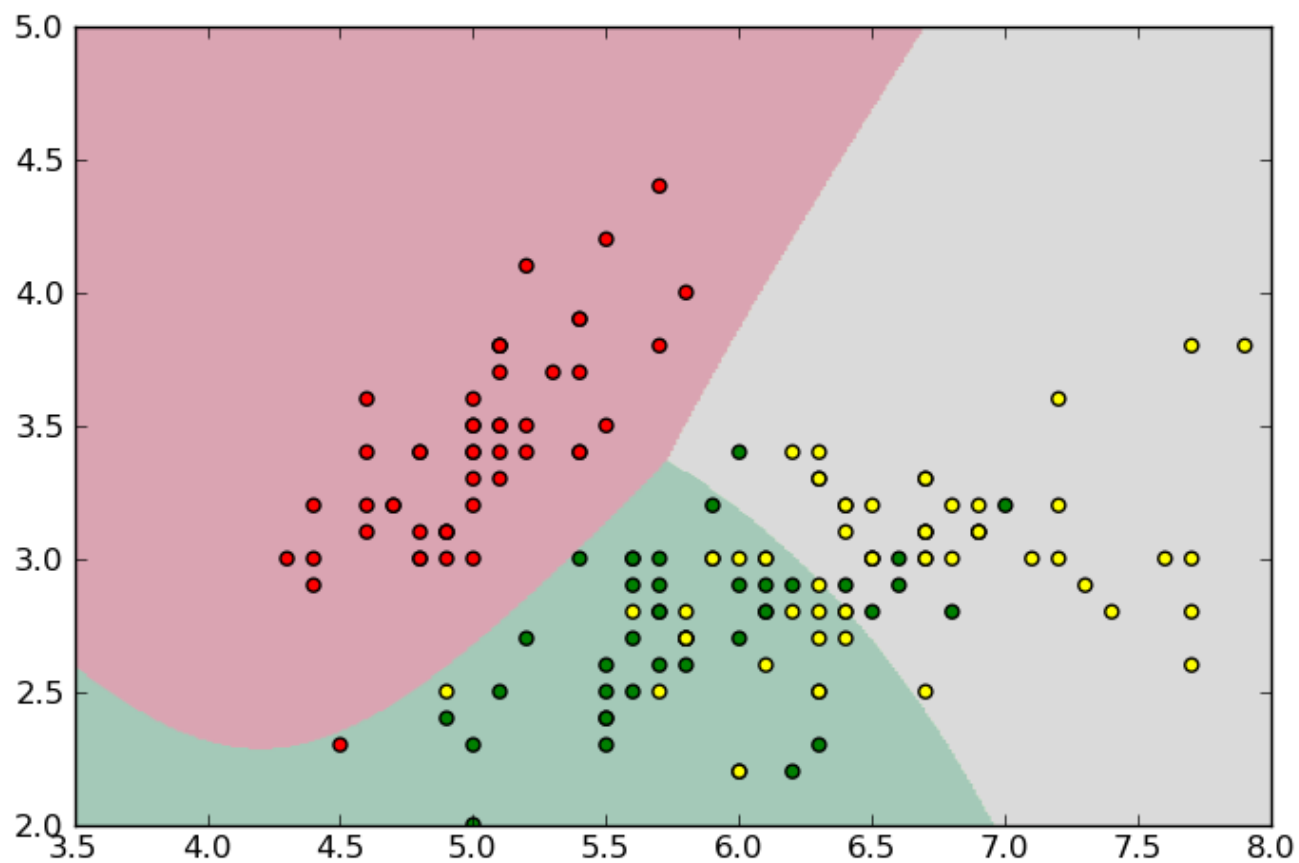


```
> library(scatterplot3d)
> scatterplot3d(datos$p.ancho,datos$s.largo,datos$s.ancho)
```



```
> library(rgl)
> D <- as.matrix(dist(datos[,1:4]))
> heatmap(D)
```

Ejemplo 1: iris.csv



Árboles de Decisión en Rattle

Minero de datos R - [Rattle (iris.csv)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Origen: ☒ Hoja de cálculo ☐ ARFF ☐ ODBC ☐ Conjunto de datos R ☐ Archivo de datos R ☐ L

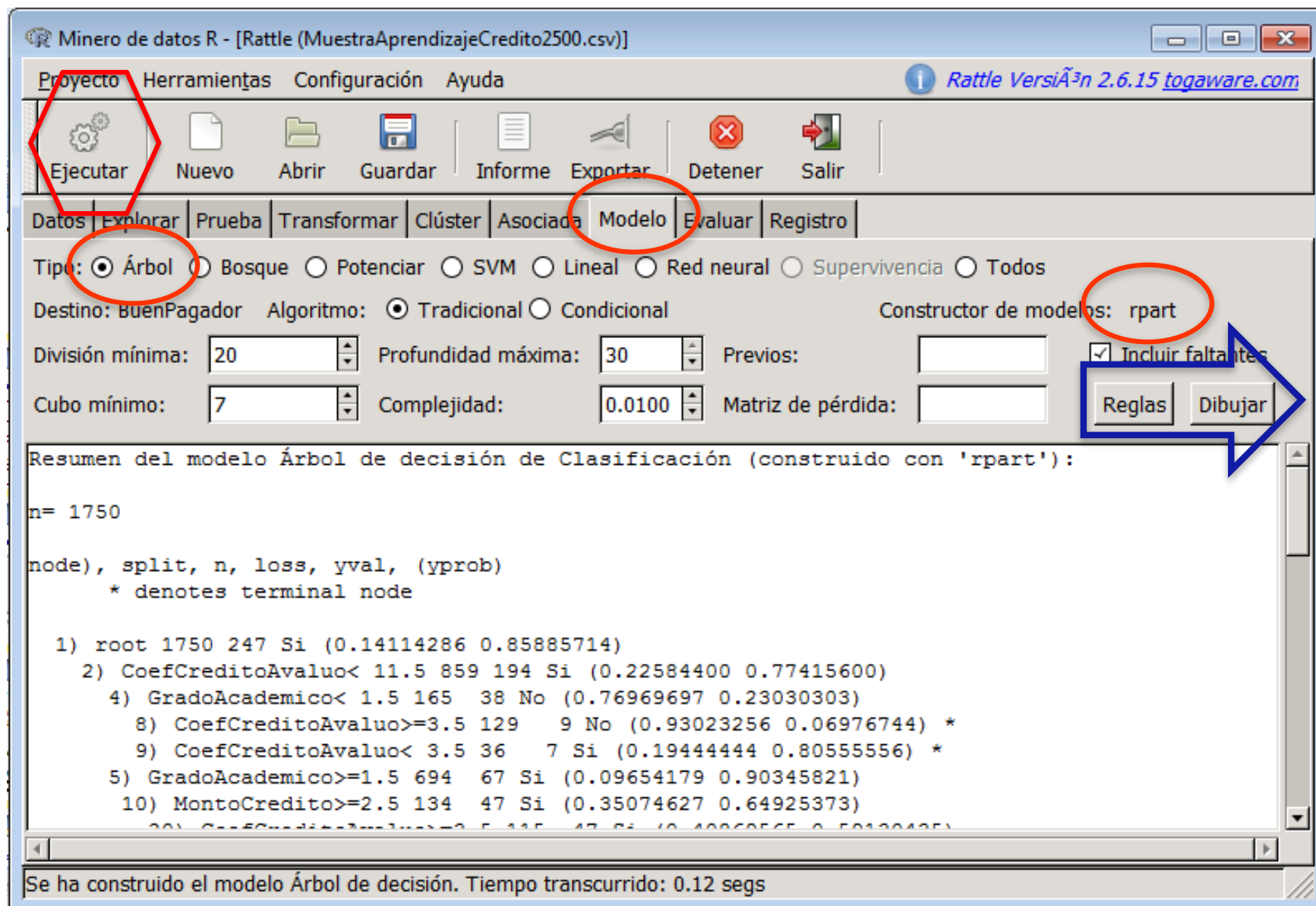
Archivo: Separador: ; Decimal: . ☒ Encabezado

☒ Partición 70/15/15 Semilla: 42 Ver Editar

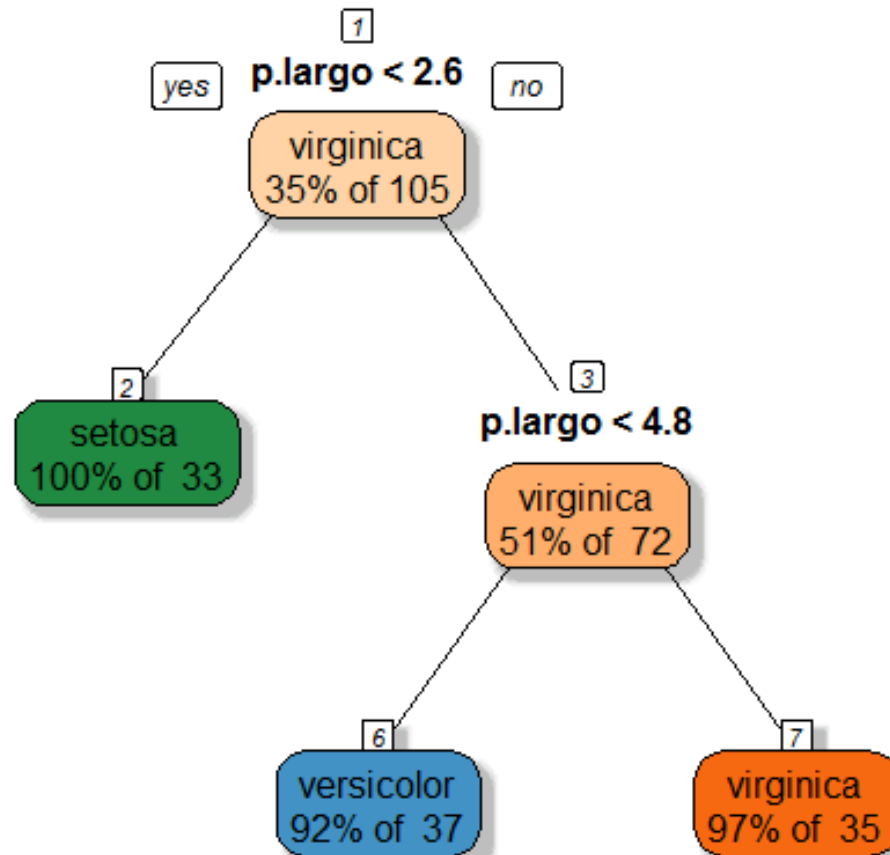
☒ Entrada ☐ Ignorar Calculadora de peso:

Tipo de datos de destino: ☒ Automática ☐ Categórica

No.	Variable	Tipo de datos	Entrada	Destino	Riesgo	Ident	Ignorar	Weight	Comentario
1	s.largo	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 35
2	s.ancho	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 23
3	p.largo	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 43
4	p.ancho	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 22
5	tipo	Categórica	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 3



Árbol de decisión iris.csv \$ tipo



Reglas en Rattle

Árbol como reglas:

Rule number: 2 [tipo=setosa cover=33 (31%) prob=1.00]
p.largo < 2.6

Rule number: 7 [tipo=virginica cover=35 (33%) prob=0.00]
p.largo ≥ 2.6
p.largo ≥ 4.85

Rule number: 6 [tipo=versicolor cover=37 (35%) prob=0.00]
p.largo ≥ 2.6
p.largo < 4.85

Matriz de confusión en Rattle (Matriz de Error)

Minero de datos R - [Rattle (iris.csv)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Tipo: ☒ Matriz de error ☐ Riesgo ☐ Curva de costo ☐ Hand ☐ Elevación ☐ ROC ☐ Precisión ☐ Sensibilidad

Modelo: ☒ Árbol ☐ Potenciar ☐ Bosque ☐ SVM ☐ Lineal ☐ Red neural ☐ Supervivencia ☐ KMeans ☐ HClust

Datos: ☐ Entrenamiento ☐ Convalidación ☒ Prueba ☐ Completo ☐ Ingresar ☐ Archivo CSV

Variable de riesgo: Informe: ☒ Clase ☐ Probabilidad Incluir: ☐

Matriz de error para el modelo Árbol de decisión en iris.csv [prueba] (cuentas):

	Predicho		
Real	setosa	versicolor	virginica
setosa	11	0	0
versicolor	0	6	2
virginica	0	0	4

Matriz de error para el modelo Árbol de decisión en iris.csv [prueba] (%):

	Predicho		
Real	setosa	versicolor	virginica
setosa	48	0	0
versicolor	0	26	9
virginica	0	0	17

Ejemplo 2:

Credit-Scoring

MuestraAprendizajeCredito2500.csv
MuestraTestCredito2500.csv

```
> setwd("C:/Users/Oldemar/Google Drive/Curso Minería 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 Rattle

Minero de datos R - [Rattle (MuestraAprendizajeCredito2500.csv)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos: Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Origen: ☒ Hoja de cálculo ☐ ARFF ☐ ODBC ☐ Conjunto de datos R ☐ Archivo de datos R ☐ Librería ☐ Co

Archivo: Separador: ; Decimal: . ☒ Encabezado

☒ Partición 70/15/15 Semilla: 42 Ver Editar

☒ Entrada ☐ Ignorar Calculadora de peso:

Tipo de datos de destino: ☒ Automática ☐ Categórica ☐ Numérica

No.	Variable	Tipo de datos	Entrada	Destino	Riesgo	Ident	Ignorar	Weight	Comentario
1	MontoCredito	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 4
2	IngresoNeto	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 2
3	CoefCreditoAvaluo	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 12
4	MontoCuota	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 4
5	GradoAcademico	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 2
6	BuenPagador	Categórica	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 2

Minero de datos R - [Rattle (MuestraAprendizajeCredito2500.csv)]

Proyecto Herramientas Configuración Ayuda

Rattle Versión 2.6.21 togawa

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada **Modelo** Evaluar Registro

Tipo: ☒ Árbol ☐ Bosque ☐ Potenciar ☐ SVM ☐ Lineal ☐ Red neural ☐ Supervivencia ☐ Todos

Destino: BuenPagador Algoritmo: ☒ Tradicional ☐ Condicional

Constructor de modelos: **rpart**

División mínima: 20 Profundidad máxima: 30 Previos:

Cubo mínimo: 7 Complejidad: 0.0100 Matriz de pérdida:

☐ Incluir valores

Reglas Dibujar

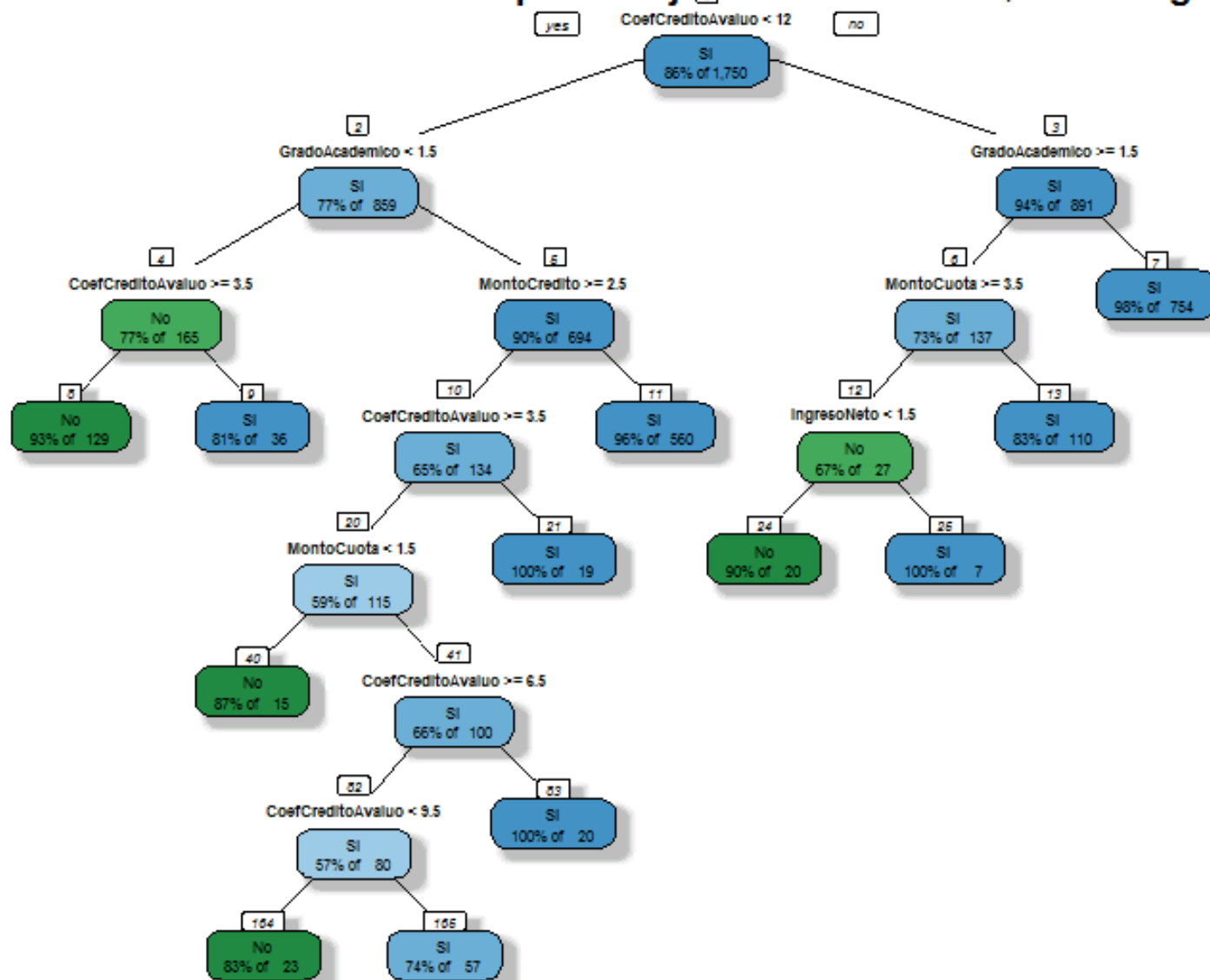
Resumen del modelo Árbol de decisión de Clasificación (construido con 'rpart'):

n= 1750

```
node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 1750 247 Si (0.14114286 0.85885714)
 2) CoefCreditoAvaluo< 11.5 859 194 Si (0.22584400 0.77415600)
    4) GradoAcademico< 1.5 165 38 No (0.76969697 0.23030303)
        8) CoefCreditoAvaluo>=3.5 129 9 No (0.93023256 0.06976744) *
        9) CoefCreditoAvaluo< 3.5 36 7 Si (0.19444444 0.80555556) *
    5) GradoAcademico>=1.5 694 67 Si (0.09654179 0.90345821)
        10) MontoCredito>=2.5 134 47 Si (0.35074627 0.64925373)
            20) CoefCreditoAvaluo>=3.5 115 47 Si (0.40869565 0.59130435)
                40) MontoCuota< 1.5 15 2 No (0.86666667 0.13333333) *
                41) MontoCuota>=1.5 100 34 Si (0.34000000 0.66000000)
```

Árbol de decisión MuestraAprendizajeCredito2500.csv \$ BuenPagador



Reglas en Rattle

Árbol como reglas:

Rule number: 25 [BuenPagador=Si cover=7 (0%) prob=1.00]

CoefCreditoAvaluo \geq 11.5

GradoAcademico \geq 1.5

MontoCuota \geq 3.5

IngresoNeto \geq 1.5

Rule number: 21 [BuenPagador=Si cover=19 (1%) prob=1.00]

CoefCreditoAvaluo $<$ 11.5

GradoAcademico \geq 1.5

MontoCredito \geq 2.5

CoefCreditoAvaluo $<$ 3.5

.....

Matriz de confusión en Rattle (Matriz de Error)

Minero de datos R - [Rattle (MuestraAprendizajeCredito2500.csv)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Tipo: ☒ Matriz de error ☐ Riesgo ☐ Curva de costo ☐ Hand ☐ Elevación ☐ ROC ☐ Precisión ☐ Sensibilidad ☐ C

Modelo: ☒ Árbol ☐ Potenciar ☐ Bosque ☐ SVM ☐ Lineal ☐ Red neural ☐ Supervivencia ☐ KMeans ☐ HClust

Datos: ☐ Entrenamiento ☐ Convalidación ☐ Prueba ☐ Completo ☐ Ingresar ☒ Archivo CSV ☐ Muestr... ☐

Variable de riesgo: Informe: ☒ Clase ☐ Probabilidad ☐ Incluir ☐ Id

Matriz de error para el modelo Árbol de decisión en MuestraTestCredito2500.csv (cuenta

	Predicho	
Real	No	Si
No	234	111
Si	22	2133

Matriz de error para el modelo Árbol de decisión en MuestraTestCredito2500.csv (%):

	Predicho	
Real	No	Si
No	9	4
Si	1	85

Error general: 0.0532

Curva ROC

- Una curva ROC compara la tasa de falsos positivos con la de verdaderos positivos.
- El área bajo la curva ROC = 0.8967

