Tema 5: Modelos predictivos

Minería de Datos

viu Universidad Internacional de Valencia

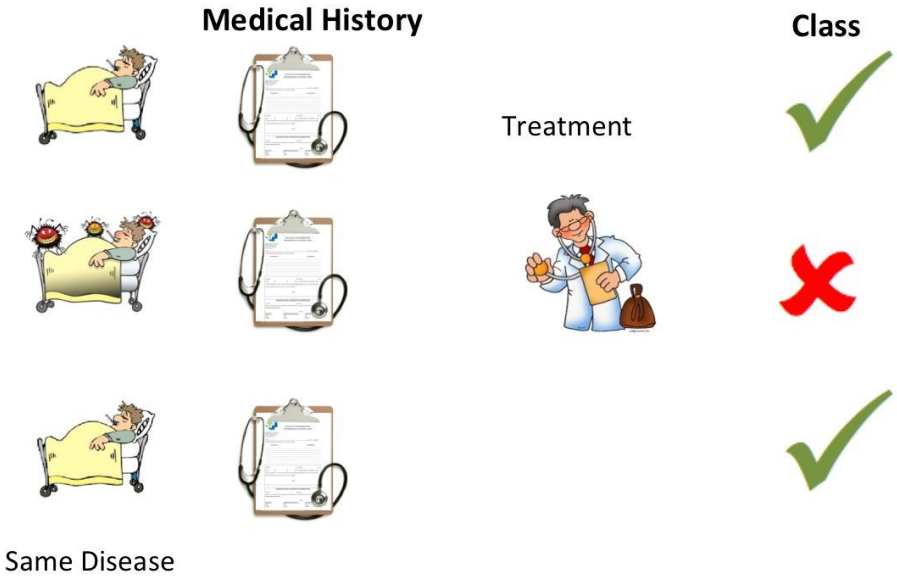
De:
Planeta Formación y Universidades

Índice

- El aprendizaje automático es un tipo de inteligencia artificial (IA) que proporciona a los ordenadores la capacidad de aprender sin ser programados explícitamente. El aprendizaje automático se centra en el desarrollo de programas informáticos capaces de aprender por sí mismos a crecer y cambiar cuando se exponen a nuevos datos.

- Los algoritmos de aprendizaje automático se describen como "supervisados" o "no supervisados".

- En los **algoritmos supervisados**, las **clases** están predeterminadas.
- Estas clases pueden concebirse como un conjunto finito, al que previamente ha llegado un ser humano.



- La tarea del algoritmo es buscar patrones y construir modelos matemáticos.
  - Decision Tree induction.
  - Naive Bayes
  - k-NN (K-nearest neighbour)
  - …

- A continuación, estos modelos se evalúan en función de su capacidad predictiva en relación con las medidas de varianza de los propios datos.

- **Modelos de clasificación:** Útil para predecir el valor de un **atributo categórico** (discreto o nominal).

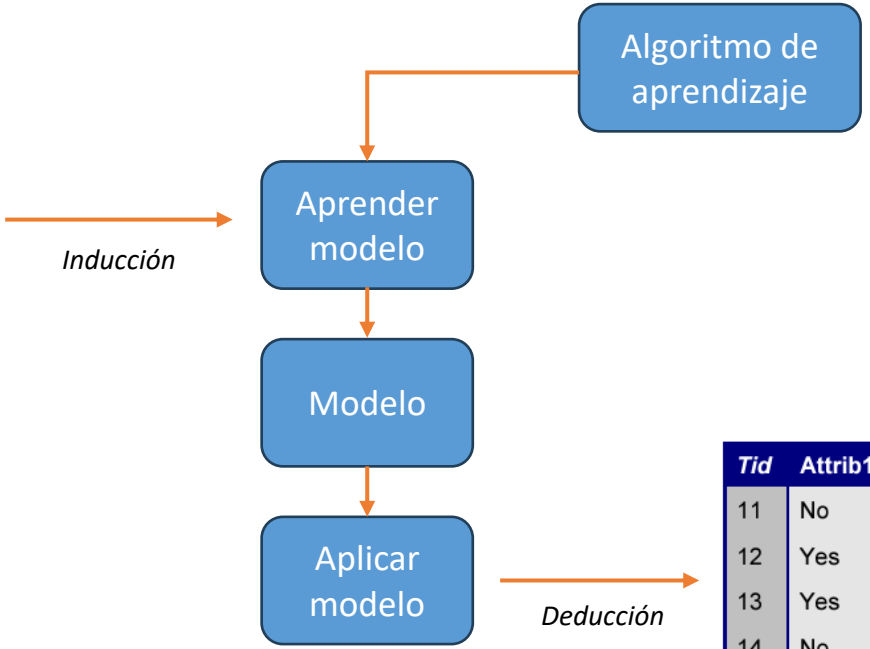
- **Modelos de predicción:** Útil para modelar funciones que contienen **valores continuos**.

- La tarea del algoritmo es buscar patrones y construir modelos matemáticos.
  - Decision Tree induction.
  - Naive Bayes
  - k-NN (K-nearest neighbour)
  - …

- A continuación, estos modelos se evalúan en función de su capacidad predictiva en relación con las medidas de varianza de los propios datos.

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Conjunto de entrenamiento (train)



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Conjunto test

- Los casos del conjunto de entrenamiento deben aparecer debidamente etiquetados con la **clase** a la que corresponden.

- **Clasificación:** Útil para predecir el valor de un atributo categórico (discreto o nominal).

- **Predicción:** Útil para modelar funciones que contengan valores continuos (predecir valores que son desconocidos).

- Una vez construido el modelo a partir del conjunto de entrenamiento, se usa dicho modelo para **clasificar los datos del conjunto de prueba**.

- Comparando los casos etiquetados del conjunto de prueba con el resultado de aplicar el modelo, se obtiene un **porcentaje de clasificación**.

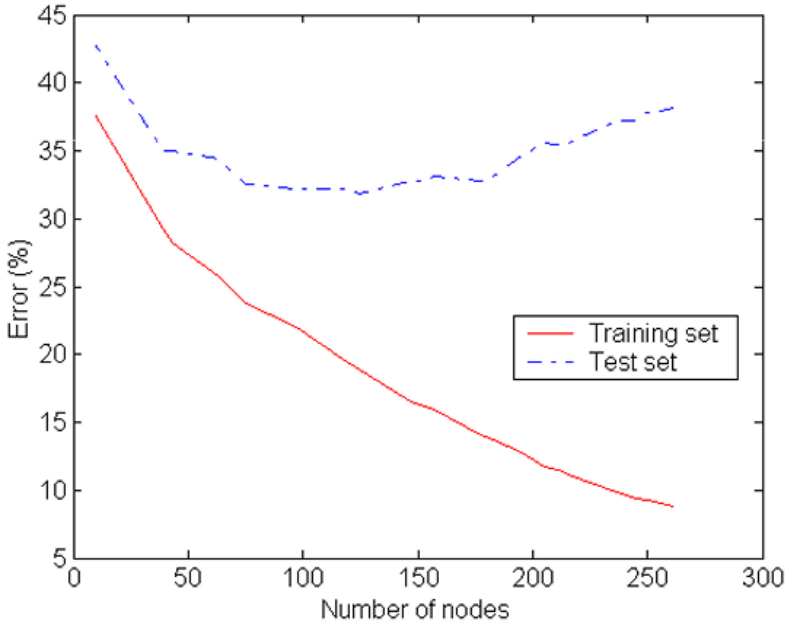
- Si la precisión del clasificador es aceptable, podremos utilizar el modelo para clasificar nuevos casos (de los que desconocemos realmente su clase).

- Cuanto mayor sea su complejidad, los modelos de clasificación tienden a ajustarse más al conjunto de entrenamiento utilizado en su construcción (**sobreaprendizaje**), lo que los hace menos útiles para clasificar nuevos datos.

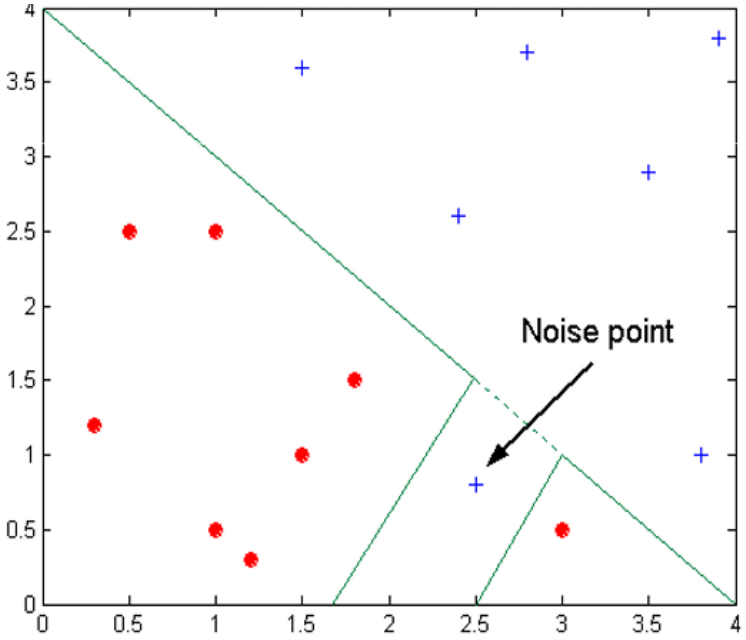
- En consecuencia, **el conjunto de prueba debe ser siempre independiente del conjunto de entrenamiento.**

- El error de clasificación en el conjunto de entrenamiento **NO** es un buen estimador de la precisión del clasificador.

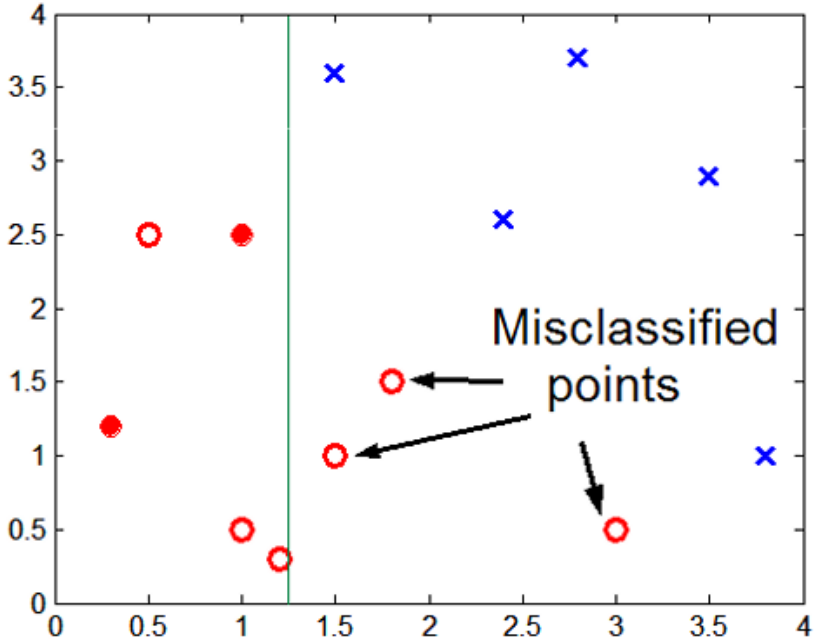
Sobreaprendizaje por la complejidad del clasificador utilizado



Sobreaprendizaje por la presencia de ruido en el dataset



Sobreaprendizaje por la escasez de muestras



Los modelos de clasificación más conocidos son:

- Árboles de decisión.
- SVMs (Support Vector Machines).
- Redes bayesianas.
- Reglas.
- Clasificadores basados en casos.
- Clasificadores paramétricos.
- Redes neuronales.

Los modelos de clasificación más conocidos son:

- **Árboles de decisión.**
- SVMs (Support Vector Machines).
- Redes bayesianas.
- **Reglas.**
- Clasificadores basados en casos.
- Clasificadores paramétricos.
- Redes neuronales.

- **Se debe contar con un conjunto de datos etiquetados previamente**. Es decir que, para cada ejemplo, debe conocerse la respuesta esperada.

- Una vez entrenado, es capaz de predecir el valor del atributo indicado previamente como la respuesta esperada (label).

Ejemplo:

Modelo para predecir a qué clase de flor de iris corresponde una flor dada en base a la información suministrada (ancho y largo del pétalo y del sépalo de la flor)

El conjunto de datos original se dividirá en dos partes:

Conjunto de datos de entrenamiento

Se utilizarán para construir el modelo. Como el aprendizaje es supervisado el método buscará ajustar su respuesta a lo indicado en estos ejemplos.

Conjunto de datos de testeo

Una vez construido el modelo será utilizado para medir su calidad.
Se espera que la respuesta del modelo coincida lo más posible con lo indicado en estos ejemplos.

- Es un modelo de predicción muy utilizado en Minería de Datos.
- Por su forma jerárquica, permite visualizar la organización de los atributos.
- Se construye a partir de la identificación sucesiva de los atributos más relevantes.

Predicción

- Recorriendo sus ramas se obtienen reglas que permiten tomar decisiones.
- Si todas las hojas se refieren al mismo atributo y es discreto es un árbol de clasificación.

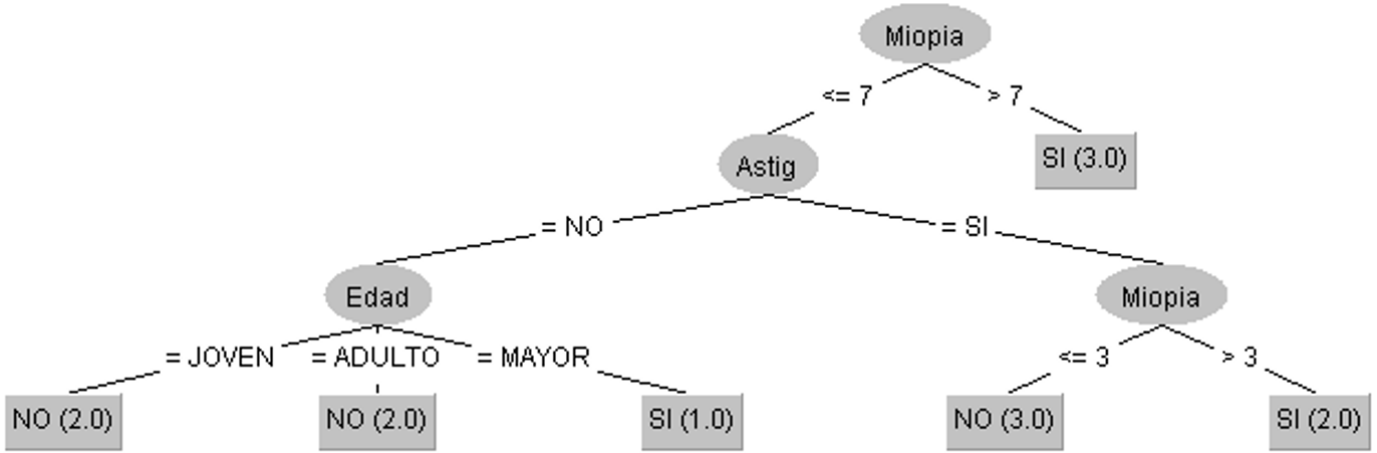
Descripción

- Su estructura jerárquica les permite mostrar cómo está organizada la información disponible.

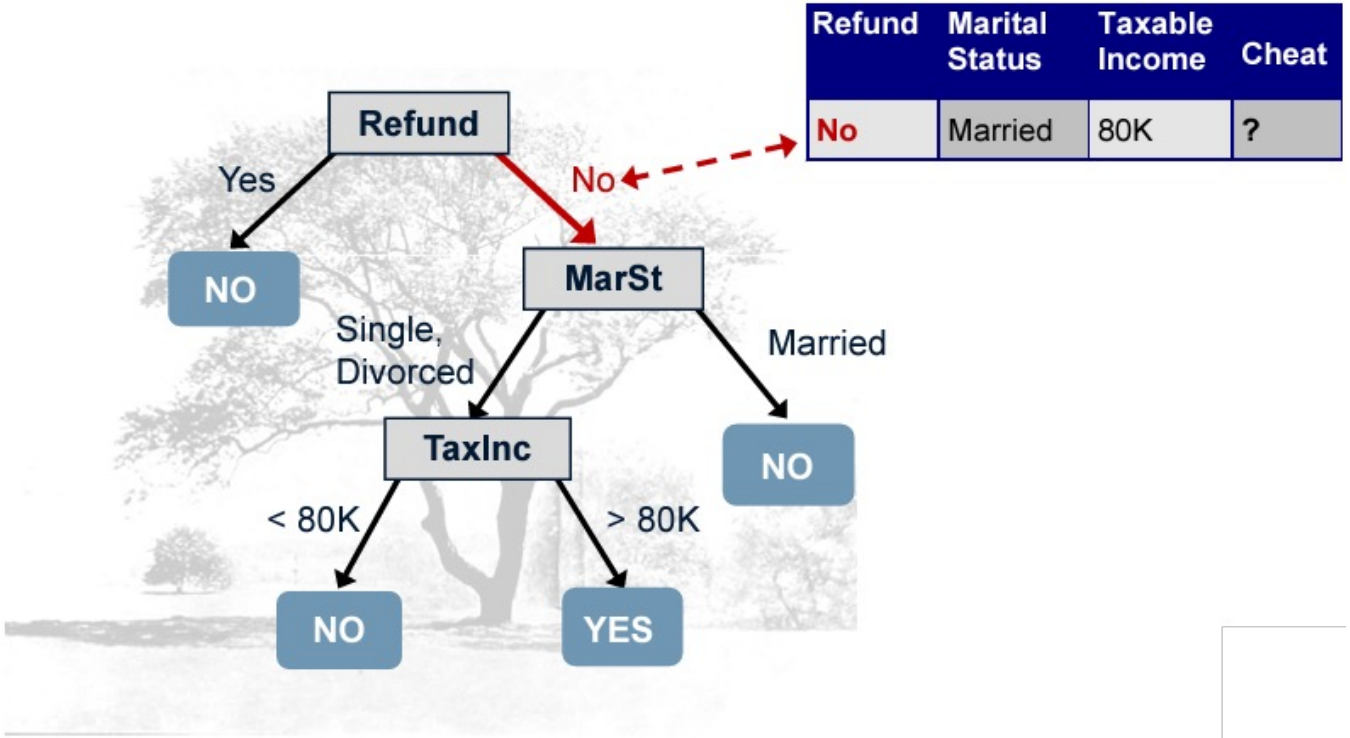
Suponiendo que se dispone de la siguiente información de pacientes tratados previamente por problemas visuales

- Edad
- Astigmatismo (si o no)
- Grado de miopía
- Recomendación de operarse (si o no)

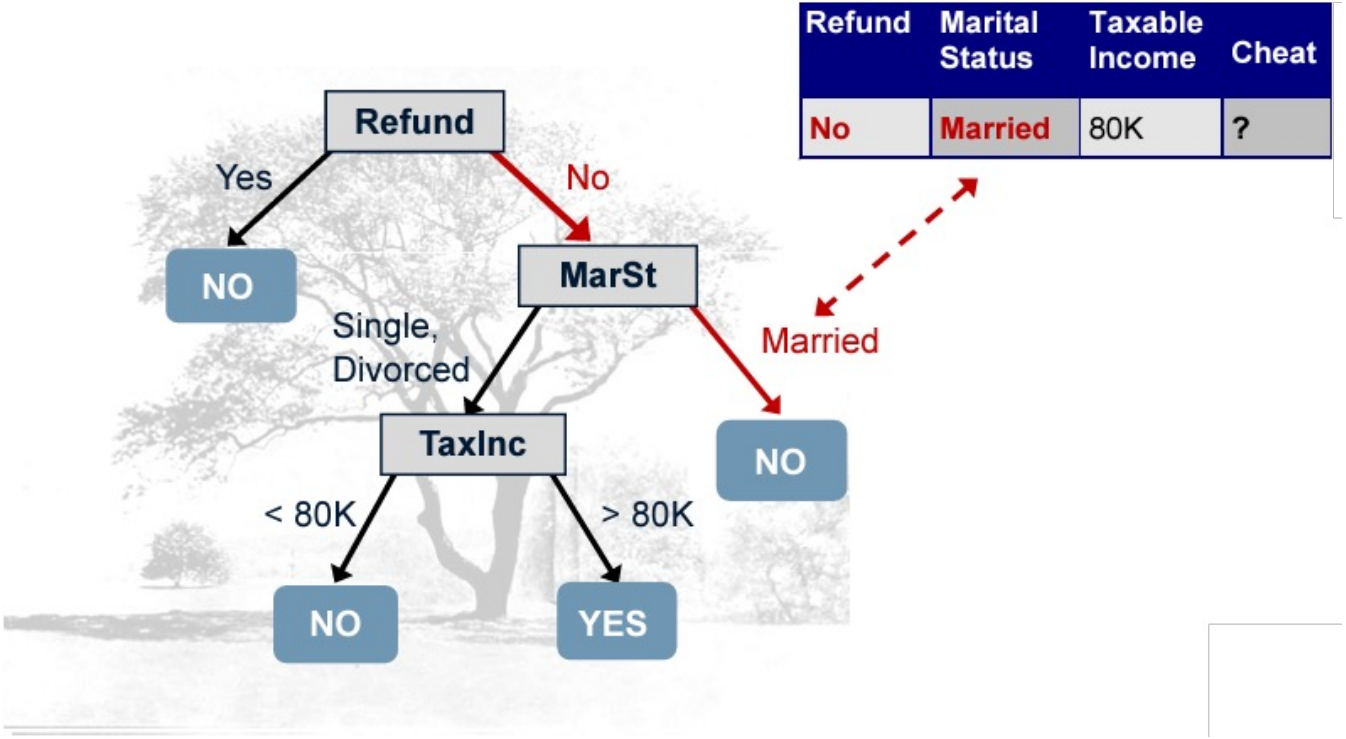
A partir de esta información puede obtenerse un modelo en forma de árbol que resuma el criterio seguido para recomendar si debe operarse o no.



Note que las opciones son excluyentes. Un atributo cualitativo no puede aparecer más de una vez en la misma rama pero un atributo cuantitativo sí puede

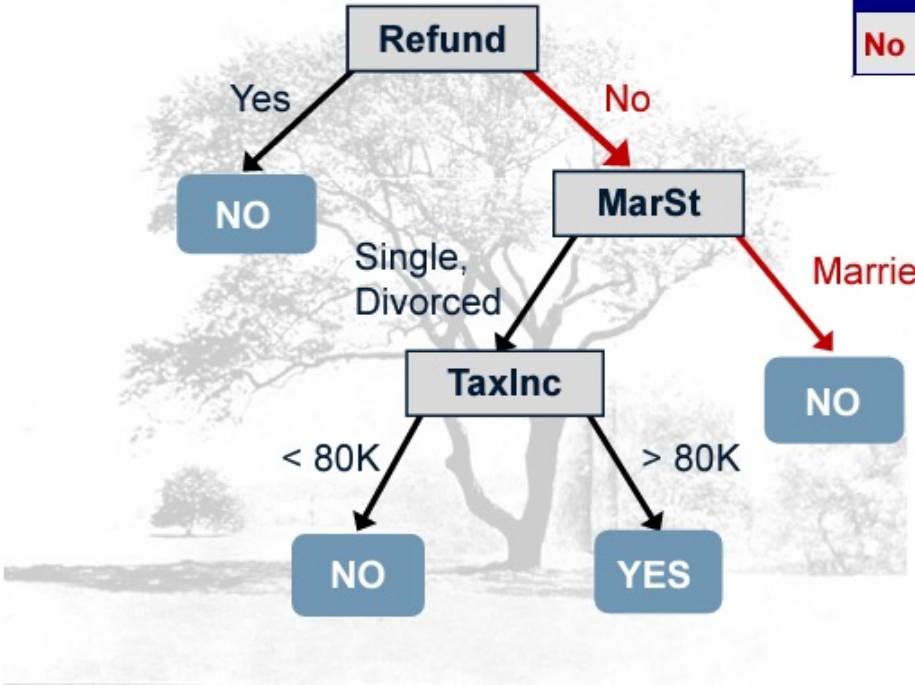


Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

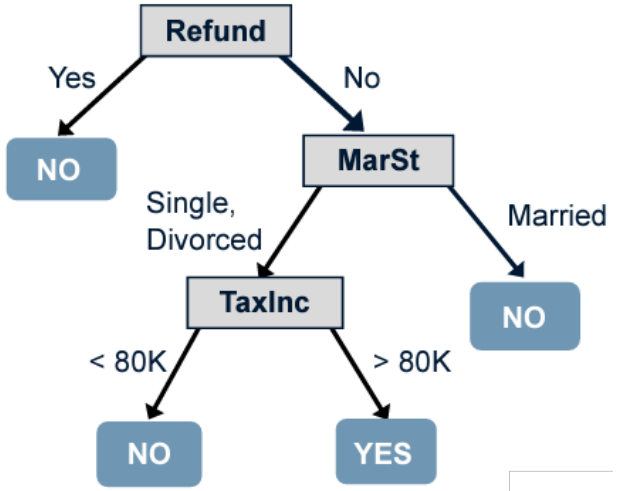
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	NO



Clase 'NO'

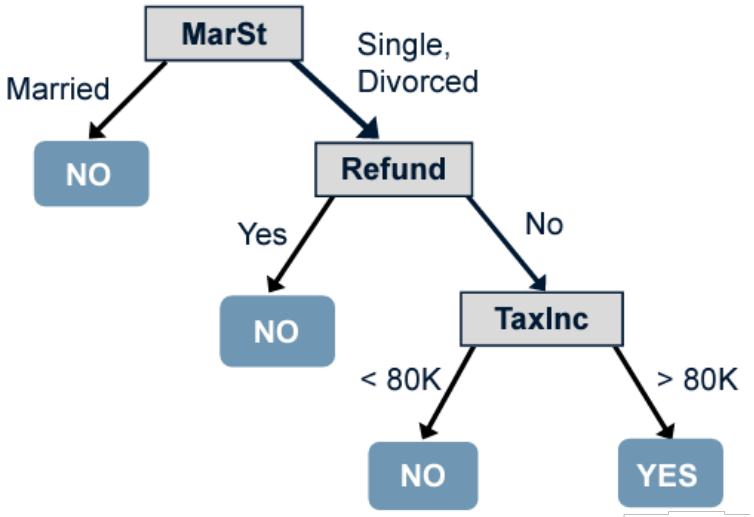
	categórico	categórico	continuo	clase
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Conjunto de entrenamiento



Modelo de clasificación: Árbol de decisión

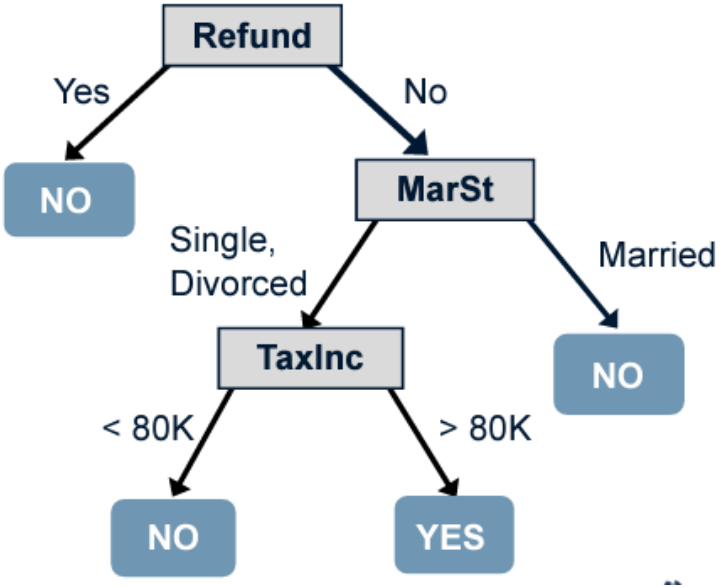
	categórico	categórico	continuo	clase
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

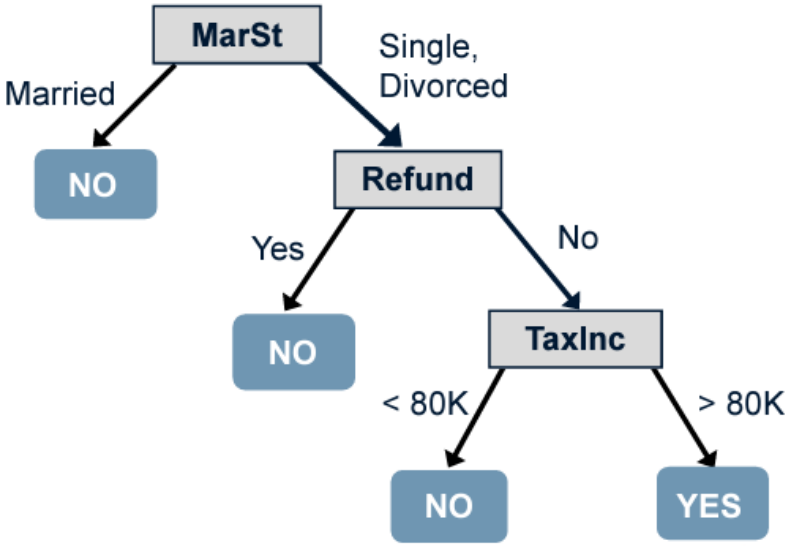


Conjunto de entrenamiento

Modelo de clasificación: Árbol de decisión

¿Qué árbol es mejor?





El árbol se construye de la forma **top-down recursive divide-and-conquer**.

Al comienzo, todos los ejemplos de entrenamiento están en el nodo raíz. Los ejemplos se particionan recursivamente basado en los atributos seleccionados.

Los atributos se seleccionan en base a una heurística o una medida estadística (p.ej., **ganancia de información**).

Condiciones para detener el particionamiento:

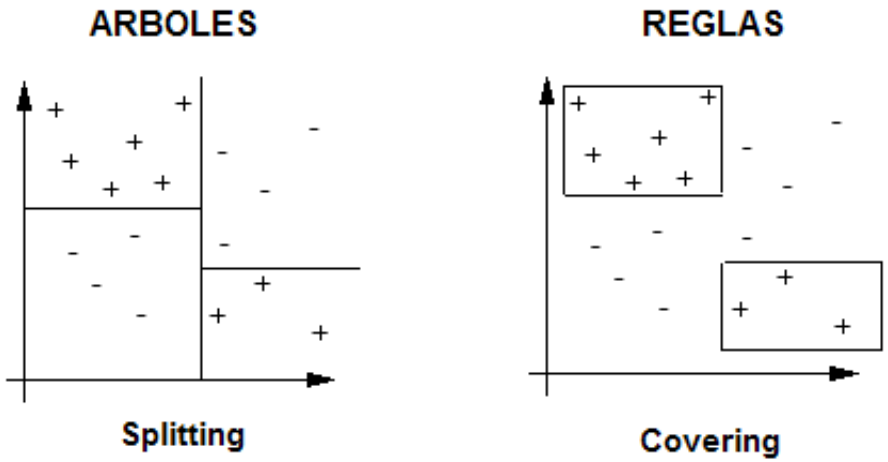
- Todas las muestras, para un nodo dado, corresponden a la misma clase.
- No hay atributos restantes para particionar. Se usa **voto mayoritario** para clasificar la hoja.
- No quedan más muestras (registros del conjunto de entrenamiento).

- A partir de la información disponible se busca obtener reglas de la forma:
- Si cuentas-Morosas > 0 entonces Devuelve-credito = no
- Si Cuentas-Morosas=0 Y [(Salario>2500) O (D-credito>10)]
  entonces Devuelve-credito= si

- En general las reglas son más compactas que los árboles. Especialmente si puede usarse una regla por defecto.
- Cada regla puede representar un concepto distinto. Esto permite agregar/quitar reglas fácilmente cosa que no es fácil de hacer en el árbol.
- Un mismo ejemplo puede ser cubierto por más de una regla. En el árbol, un mismo ejemplo sólo pertenece a una hoja del árbol.

La estrategia utilizada para aprender reglas, está basada en **covering**, esto es, encontrar condiciones de reglas (par atributo-valor) que cubra la mayor cantidad de ejemplos de una clase, y la menor del resto de las clases. Se considera el cubrir una sola clase.

La idea básica es añadir condiciones al antecedente de la regla que se está construyendo buscando maximizar la cobertura minimizando errores.



ZeroR
Es el más simple de todos. Clasifica todos los ejemplos como pertenecientes a la clase mayoritaria.

OneR
Clasifica en base a un único atributo.

PRISM
La construcción de las reglas busca caracterizar (cubrir) exactamente a los datos. A medida que se cubren los ejemplos, se eliminan de la entrada de datos. Este mecanismo de construcción lleva a obtener una lista de decisión pues el orden de ejecución de las reglas queda predeterminado.

PART
Al igual que PRISM genera una lista de reglas. A diferencia de los métodos convencionales, las reglas surgen de la mejor rama de un árbol construido parcialmente.

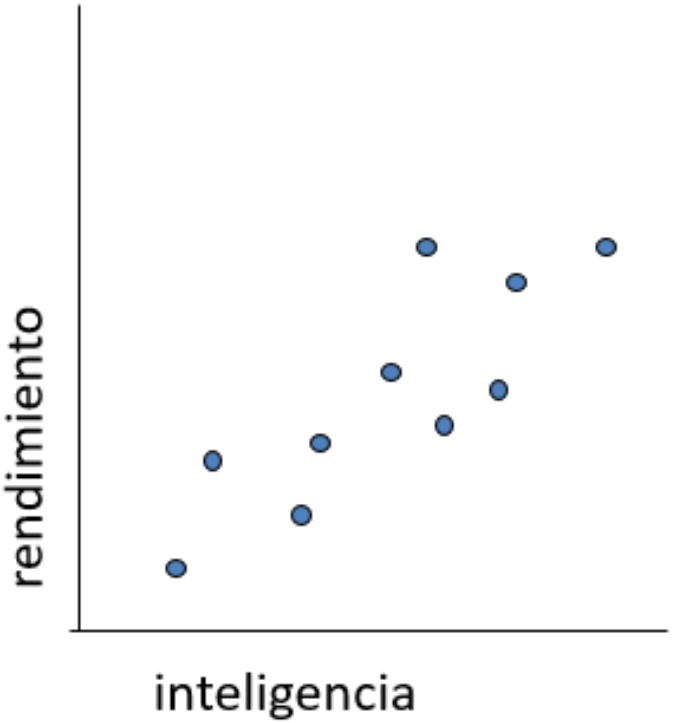
- Establecer una correlación entre dos variables se puede considerar como un primer paso para predecir una variable a partir de la otra o, incluso, otras en el caso de que se traten de regresiones múltiples.

- En definitiva, si sabemos que la variable X está muy correlacionada con Y, esto quiere decir que **podemos predecir Y a partir de X**, por lo tanto, **X sirve como predictor de Y**.

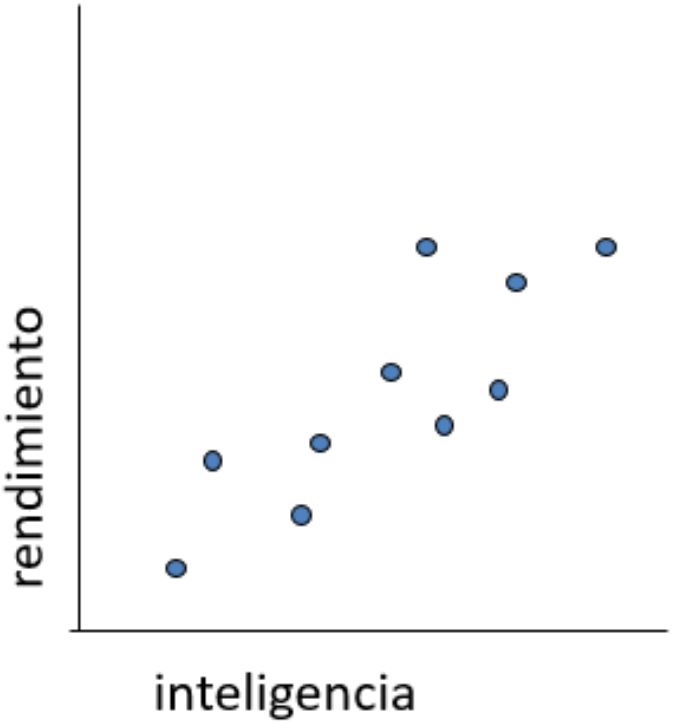
- La regresión permite modelar la relación entre una o más **variables independientes** (predictores) con una **variable dependiente** (variable de respuesta).

- **¿Qué similitudes tiene la clasificación con la regresión?** En que se construye un modelo a partir del conjunto de entrenamiento y que se utiliza el modelo para predecir el valor de una variable.

- **¿En qué se diferencia la clasificación de la regresión?** En que el modelo de regresión define una función continua.

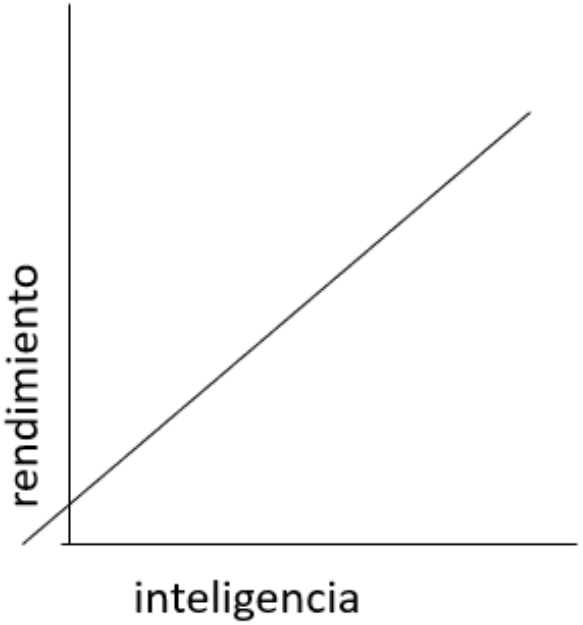


Cuando se trata de regresión con 2 variables, el objetivo consiste en **ajustar los puntos del diagrama de dispersión para las variables X e Y**.



Sin embargo, ¿cómo conseguimos pintar la "línea" más óptima en el que permita "unir" todos los puntos?

Para ello, necesitamos un criterio. Uno de los más empleados se denomina **criterio de mínimos cuadrados**.



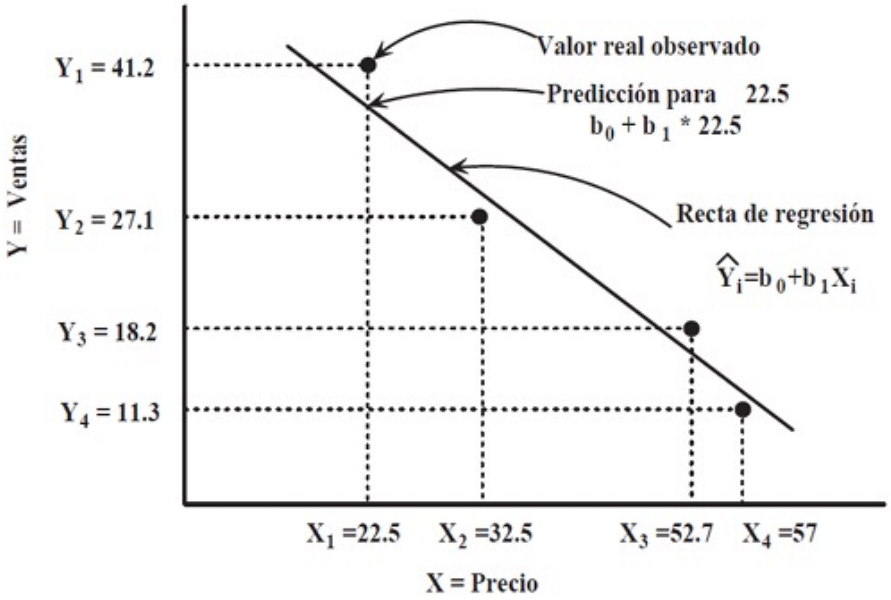
$$Y = a + bX$$

- A es el **desplazamiento** (es donde la recta corta el eje Y).
- B es la **pendiente**:
  - En el caso de las relaciones positivas, B será positivo.
  - en el caso de las relaciones negativas, B será negativo.
  - Si no hay relación, B será aproximadamente 0.

Los métodos de regresión más conocidos son:

- Regresión lineal.
- Regresión no lineal
- Árboles de regresión como, por ejemplo, CART.

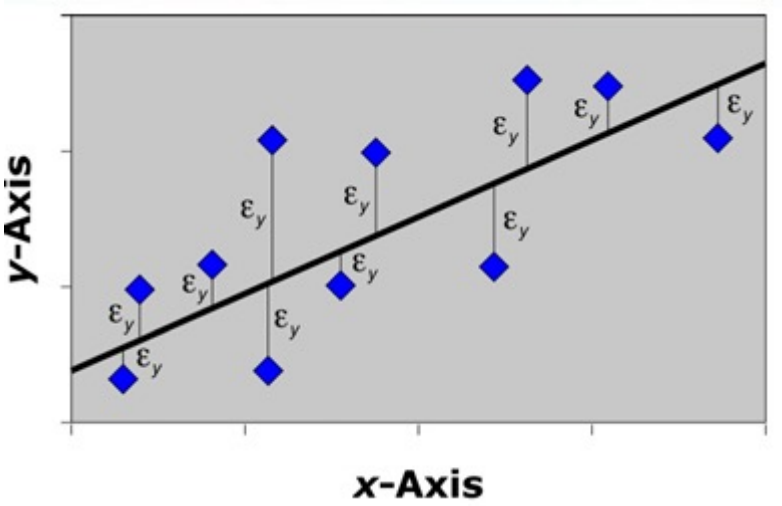
$$\hat{Y}_i = b_0 + b_1 X_i$$



b_0 (desplazamiento) y b_1 (pendiente) son los coeficientes de regresión.

El **método de los mínimos cuadrados** es utilizado para construir la recta que mejor se ajusta a los datos.

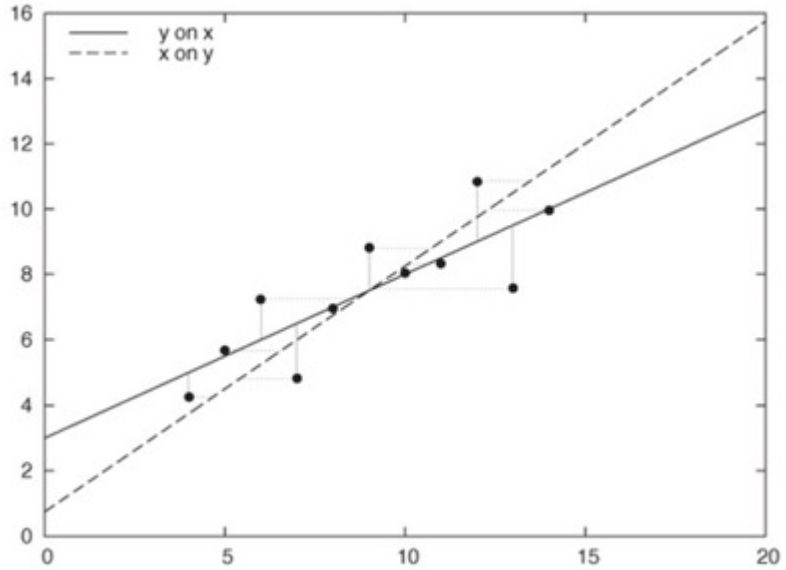
Se utiliza cuando tengamos tan sólo una variable independiente.



$$\sum_{i=1}^{n}\left(Y_i - Y_i^{'}\right)^2$$

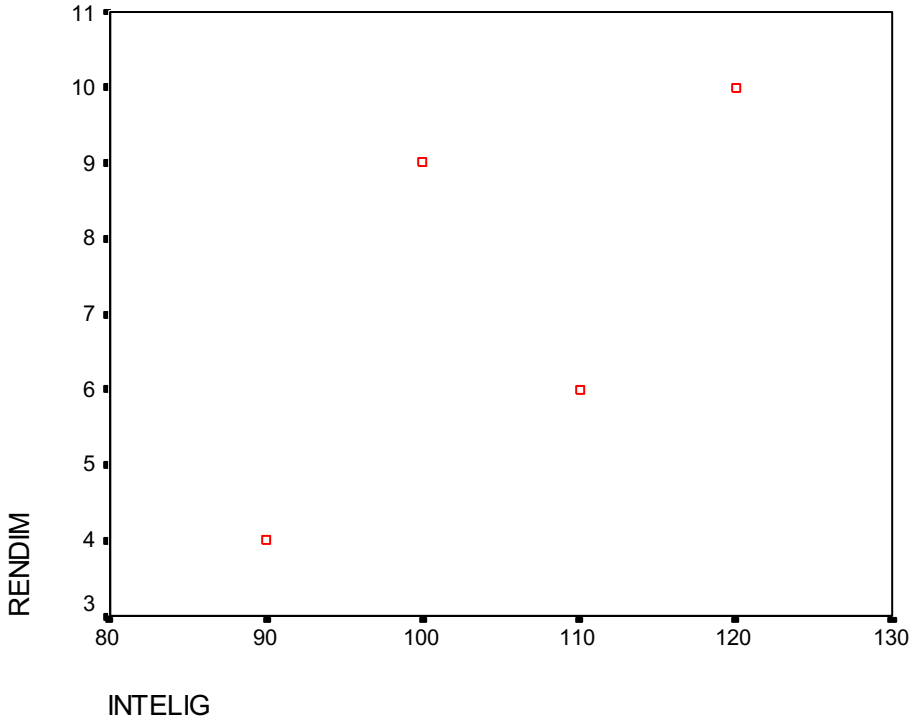
El **método de los mínimos cuadrados** minimiza la suma de los cuadrados de los residuos (ε_i).

En definitiva, se obtiene **las diferencias entre las predicciones y los valores observados.**

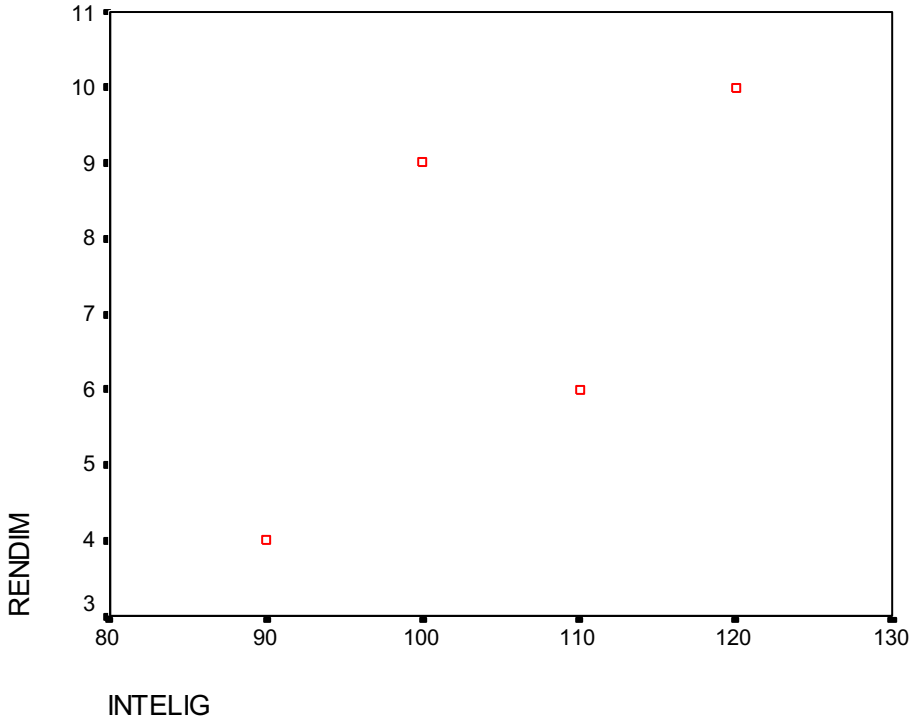


¡¡CUIDADO!!

Al utilizar la regresión lineal simple, la recta y=f(x) es distinta a la recta que se obtiene si x=f(y).



Intelig (X)	Rendimiento (Y)
120	10
100	9
90	4
110	5



$$Y = a + bX$$

¿Cómo calculamos la pendiente (b) y desplazamiento (a)?

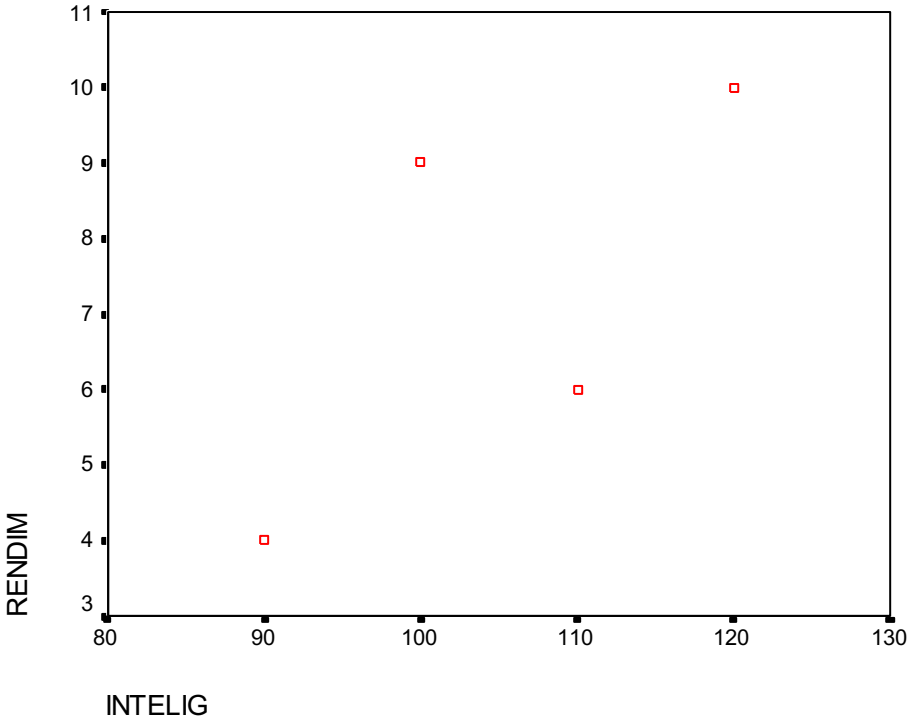
A partir de las puntuaciones de los sujetos: **directas, típicas y diferenciales.**

Puntuación directa: Se expresan los parámetros a y b de forma directa.

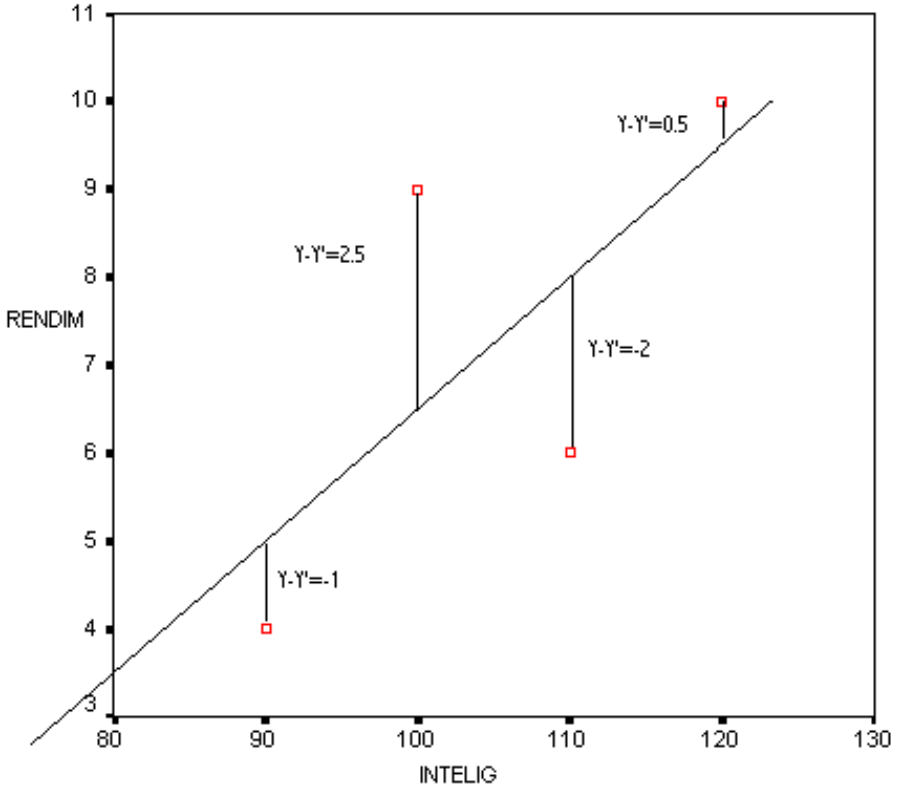
Puntuación diferencial: Tienen la peculiaridad de estimar la ecuación de regresión a partir de los valores de 0 en X e Y, es decir, que el desplazamiento (a) de dicha recta debe ser 0.

Puntuación típica: Tiene como ventaja expresar la relación entre X e Y en puntuaciones no dependientes de la escala en que se miden estas variables. En definitiva, resulta útil cuando las dos variables se encuentran en escalas de medición diferentes.

P. directas	P. diferenciales	P. típicas
$\hat{Y} = a + bX$	$(Y_i - \overline{Y}) = b(X_i - \overline{X}) + e_i$	$Z_{iY} = r_{XY}\, Z_{iX} + e_i$
$A = \overline{Y} - B\overline{X}$ $B = \dfrac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$	$a = 0$ $b = \dfrac{\sum_1^N xy}{\sum_1^N x}$	$r = \dfrac{\sum_1^N Z_x Z_y}{N}$ $z_{iX} = \dfrac{(X_i - \overline{X})}{S_X}$



¿Cómo obtenemos el valor de desplazamiento (a) y el valor de la pendiente (b) usando puntuaciones directas?



Observa que:

Cada unidad de INTELIG hace aumentar 0,15 el RENDIM.

Fórmulas:

Desplazamiento (a):

$$A = \bar{Y} - B\bar{X}$$

Pendiente (b):

$$B = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

	X	Y	XY	X2
suj1	120	10	1200	14400
suj2	100	9	900	10000
suj3	90	4	360	8100
suj4	110	6	660	12100
	4		SUMA	SUMA
			3120	44600
	PROMEDIO	PROMEDIO		
	105	7.25		
	N			
	4			

$$A = 7'25 - 0'15 \cdot 105 = -8'5$$

$$B = \frac{3120 - 4 \cdot 105 \cdot 7'25}{44600 - 4 \cdot 105^2} = 0'15$$

	X	Y	XY	X2
suj1	120	10	1200	14400
suj2	100	9	900	10000
suj3	90	4	360	8100
suj4	110	6	660	12100
	4		SUMA	SUMA
			3120	44600
	PROMEDIO	PROMEDIO		
	105	7.25		
	N			
	4			

Fórmula de la recta con puntuaciones directas:

Y’ = -8.5 + 0.15X

Resumen de la recta según puntuaciones

P. directas	P. diferenciales	P. típicas
$\hat{Y} = a + bX$	$(Y_i - \overline{Y}) = b(X_i - \overline{X}) + e_i$	$Z_{iY} = r_{XY}\, Z_{iX} + e_i$
Y'= -8.5 + 0.15X	y' = 0.15x	$z_{y'} = 0.703 z_x$

Resumen de coeficientes no estandarizados vs estandarizados (Pearson)

Ord. y pendiente (punt.directas)

Ord. y pendiente (punt.típicas)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-8.500	11.324		-.751	.531
	INTELIG	.150	.107	.703	1.399	.297

a. Variable dependiente: RENDIM

Resumen de coeficientes no estandarizados vs estandarizados (Pearson)

Ord. y pendiente (punt.directas)

Ord. y pendiente (punt.típicas)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-8.500	11.324		-.751	.531
	INTELIG	.150	.107	.703	1.399	.297

a. Variable dependiente: RENDIM

Pearson = 0.703

¿Con qué valor coincide el coeficiente de Pearson?

P. directas	P. diferenciales	P. típicas
$\hat{Y} = a + bX$	$(Y_i - \overline{Y}) = b(X_i - \overline{X}) + e_i$	$Z_{iY} = r_{XY} Z_{iX} + e_i$
Y'= -8.5 + 0.15X	y' = 0.15x	$z_{y'} = 0.703 z_x$

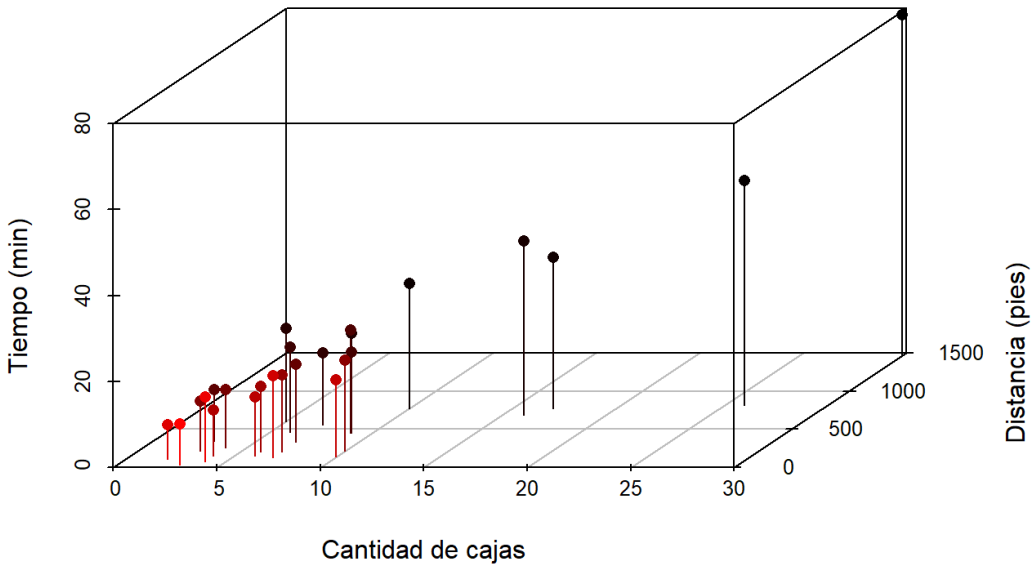
La pendiente en puntuaciones típicas = Correlación de Pearson

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.703[a]	.495	.242	2.398

a. Variables predictoras: (Constante), INTELIG

b. Variable dependiente: RENDIM



Se utiliza cuando tenemos varias variables independientes.

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots$$

Rendimiento (X)	Ansiedad (Y)	Neurotoxinas (Z)
9	3	5
3	12	15
6	8	8
2	9	7
7	7	6

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	11.288	2.221		5.082	.037
	ANSIED	-1.139	.510	-1.293	-2.233	.155
	NEUROT	.365	.421	.502	.868	.477

a. Variable dependiente: RENDIM

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.904[a]	.817	.634	1.744

a. Variables predictoras: (Constante), NEURO, ANSIE

Gracias

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades