



Universidad Nacional Autónoma de México
Facultad de Ingeniería

Clasificación

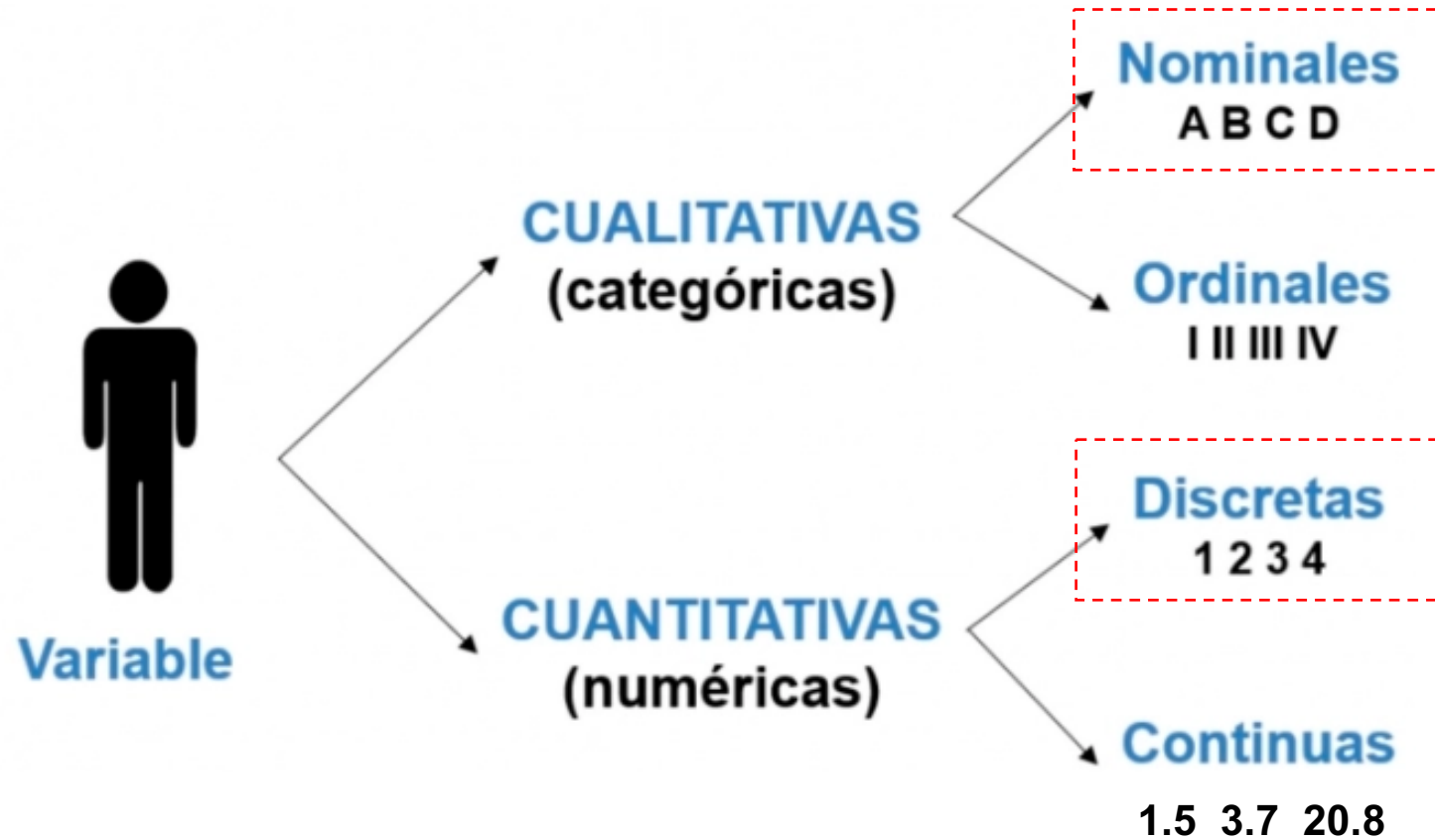
Regresión Logística

Guillermo Molero-Castillo

guillermo.molero@ingenieria.unam.edu

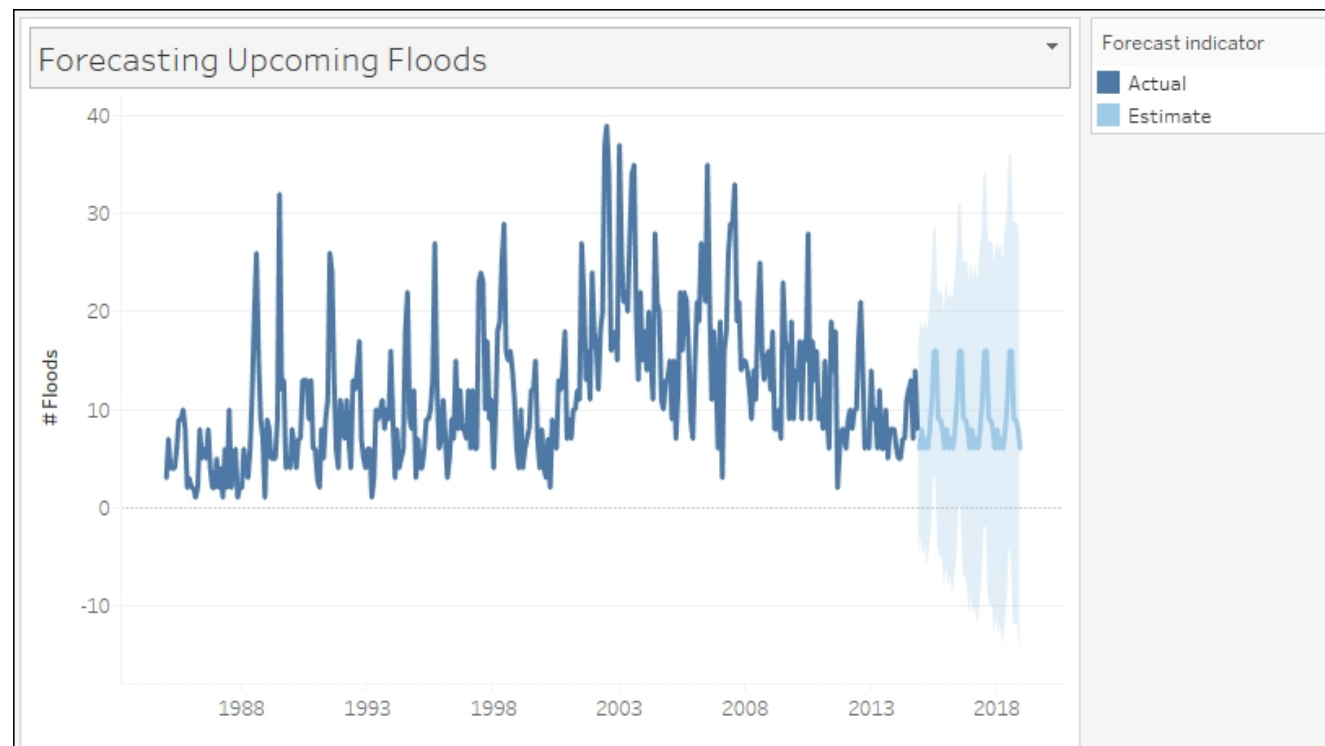
Enero, 2021

Tipos de variables



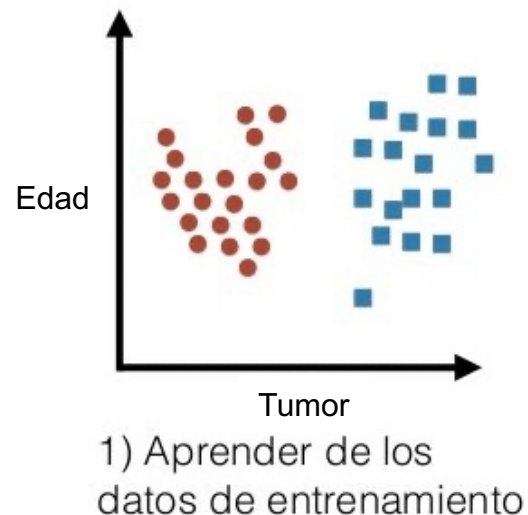
Pronóstico

- Modela funciones de valor continuo, es decir, predice valores desconocidos o faltantes.



Clasificación

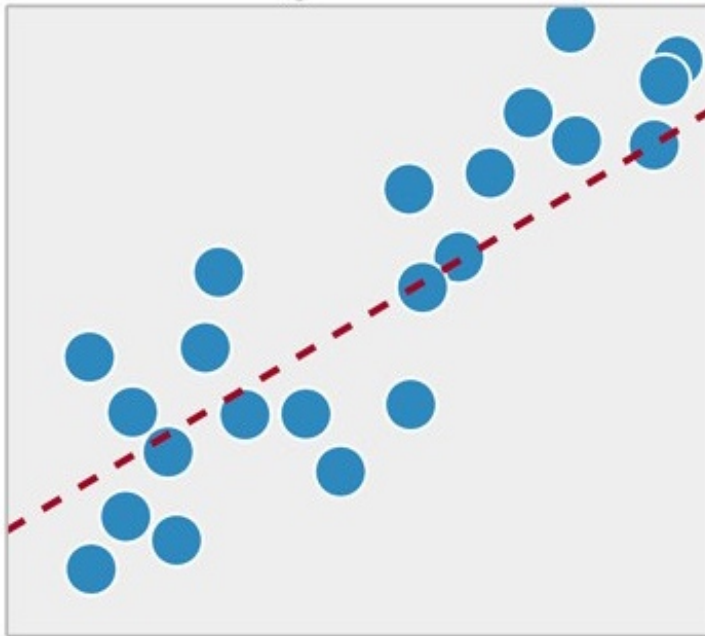
- Predice etiquetas de una o más clases de tipo discretas (0, 1, 2) o nominales (A, B, C; o positivo, negativo; y otros).
- Para esta clasificación se construye un modelo a través de un conjunto de entrenamiento (*training*).
- Se evalúa el modelo con un conjunto de prueba, que es independiente del entrenamiento. De lo contrario, se produce un sobre-ajuste (ajuste excesivo).



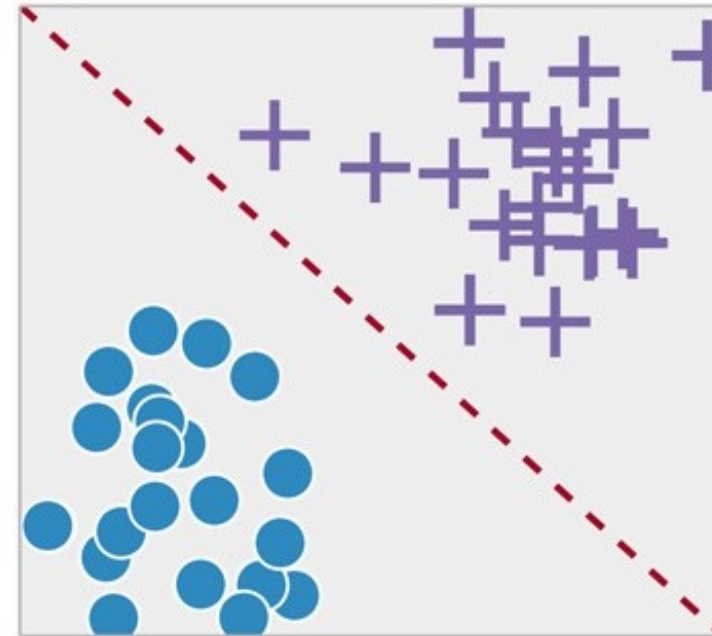
Contexto

En ambos casos, **pronóstico y clasificación**, si la precisión es aceptable, se utiliza el modelo para pronosticar o clasificar nuevos datos, cuyos valores o etiquetas no se conocen.

Pronóstico



Clasificación



Algoritmos

Sin duda los modelos de predicción son importantes, ayudan a automatizar actividades. Sin embargo, solo dice lo que sucederá, pero no lo que se debería hacer.

- Linear regression / Logistic regression
- Support vector machines
- Bayesian methods
- Nearest Neighbor (kNN)
- Artificial Neural Networks
- Decision trees
- ...

Los diferentes algoritmos tienen diferentes fortalezas y debilidades.
Se debe seleccionar el enfoque de predicción que sea adecuado para el problema.

Regresión logística

Regresión Logística

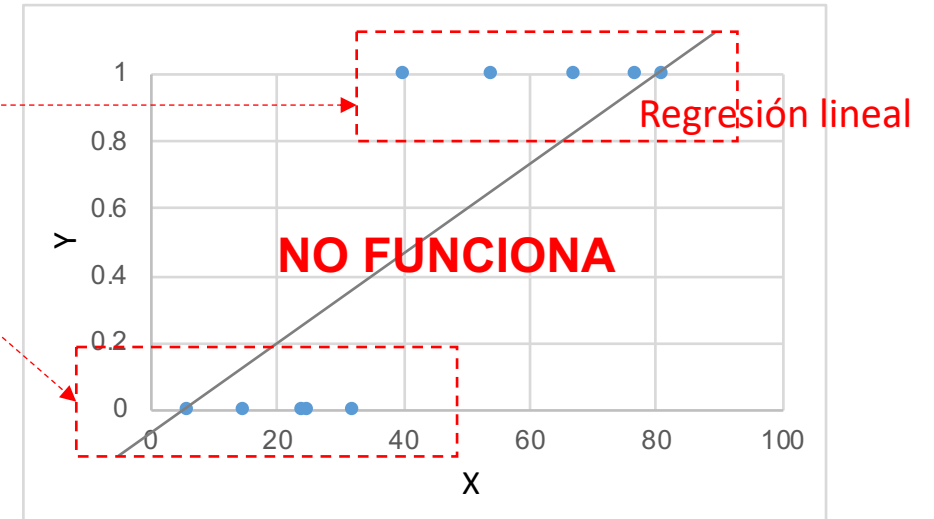
La regresión logística es otro tipo de algoritmo de aprendizaje supervisado cuyo objetivo es **predecir valores binarios** (0 o 1). Este algoritmo consiste en una transformación a la regresión lineal.

La transformación se debe a que una regresión lineal **no funciona** para **predecir una variable binaria**.

Datos		
Nro	X	Y
1	22	0
2	54	1
3	67	1
4	6	0
5	77	1
6	32	0
7	81	1
8	4	0
9	5	0
10	40	1

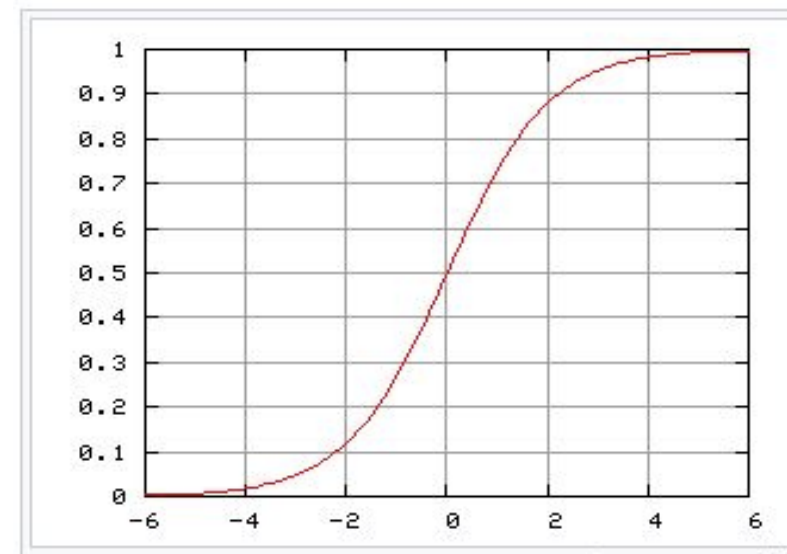
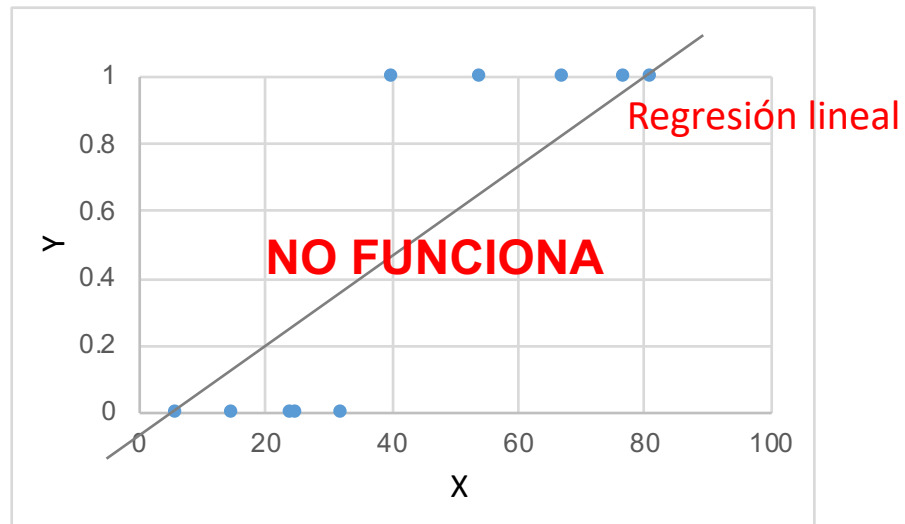
Clasificación: $Y = \{0, 1\}$

$$Y = a + b_i X_i + u$$



Regresión Logística

- Se utiliza la misma estructura que la regresión lineal, pero se **transforma** la variable respuesta (0 o 1) en una probabilidad.
- Para esta transformación se utiliza la función logística (conocida también como sigmoide).

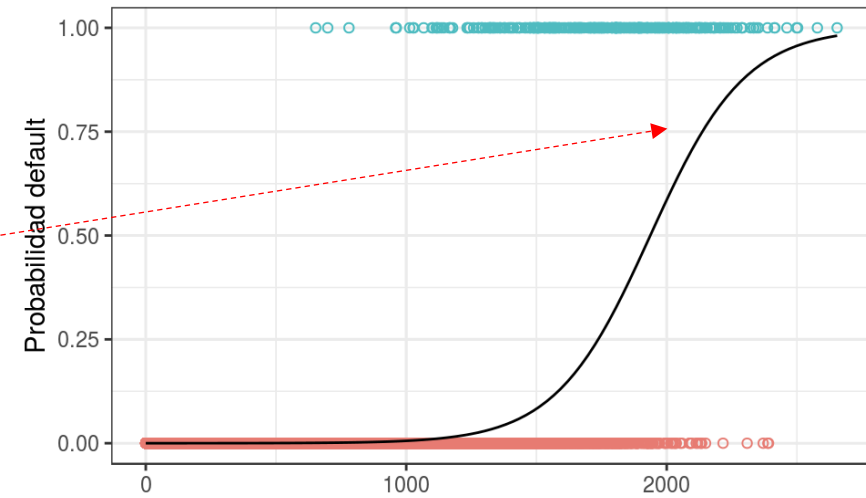


Regresión Logística

Función logística

$$\text{Función logística} = \frac{1}{1+e^{-x}} = \frac{1}{1+e^{-(a+bX)}}$$

Curva logística normalizada



e es conocido como el **número de Euler** por Leonhard Euler.

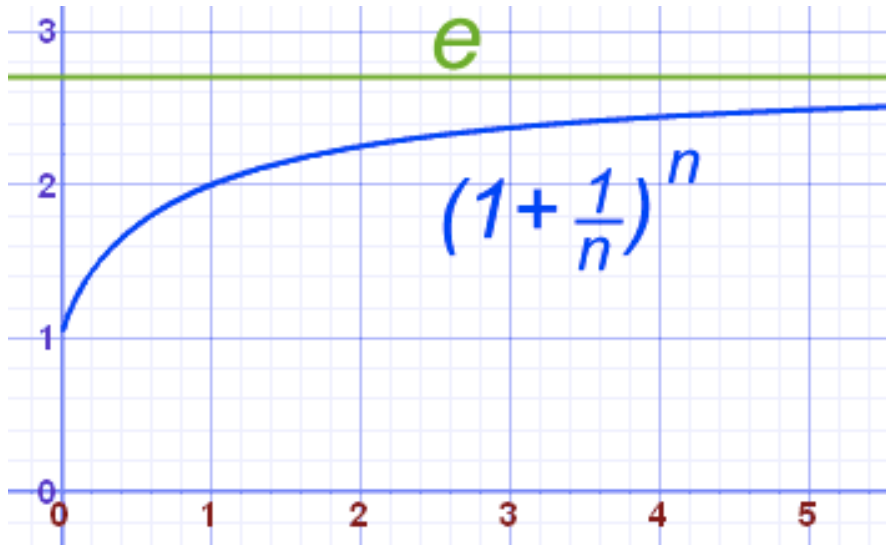
e es una constante matemática, que es la base del logaritmo natural (inventado por John Napier).

e es aproximadamente [2.718281828](#)

Regresión Logística

Número de Euler

- Euler definió una función exponencial a través de una función inversa: $(1 + 1/n)^n$
- El propósito fue tener una función, con diversas aplicaciones, para el cálculo del área cubierta por una hipérbola, el interés compuesto continuo y otros usos actuales.



n	$(1 + 1/n)^n$
1	2.00000
2	2.25000
5	2.48832
10	2.59374
100	2.70481
1000	2.71692
10000	2.71815
100000	2.71827

Regresión Logística

Funcionamiento

Paso 1


Se transforma **Y** en el logaritmo de la probabilidad de Y, esto es: $\ln\left(\frac{p}{1-p}\right)$

Datos		
Nro	X	Y
1	25	0
2	54	1
3	67	1
4	6	0
5	77	1
6	32	0
7	81	1
8	24	0
9	15	0
10	40	1



$$Y = a + b_i X_i + u$$

A esta transformación también se conoce como *razón de probabilidad de ser verdadero* (**Odds Ratio**).


$$Probabilidad = \ln\left(\frac{p}{1-p}\right) = a + b_i X_i$$

Regresión Logística

Funcionamiento

Paso 2

Se calcula la regresión lineal para predecir el logaritmo: $\ln\left(\frac{P}{1-P}\right) = a + b_i X_i$

Paso 3

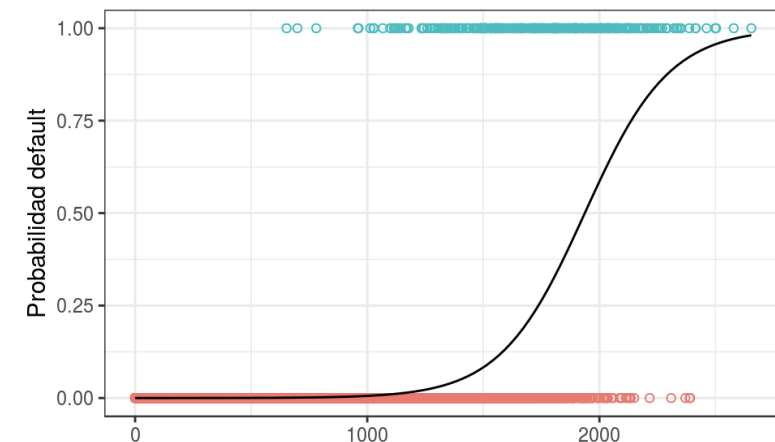
Se transforma el resultado de la regresión lineal en la probabilidad final.

$$\frac{1}{1+e^{-x}} = \frac{1}{1+e^{-(a+b_i X_i)}}$$



Donde $e = 2.718281828$

- Si la probabilidad es superior a **0.5** se asigna **1**.
- Si es menor a **0.5** se asigna **0**.



Regresión Logística

En resumen

Datos		
Nro	X	Y
1	25	0
2	54	1
3	67	1
4	6	0
5	77	1
6	32	0
7	81	1
8	24	0
9	15	0
10	40	1

→ ~~$Y = a + bX + u$~~

Paso 1

- Se transforma **Y** en el logaritmo de la probabilidad de **Y**

$\ln\left(\frac{P}{1-P}\right) = a + bX$

Paso 2

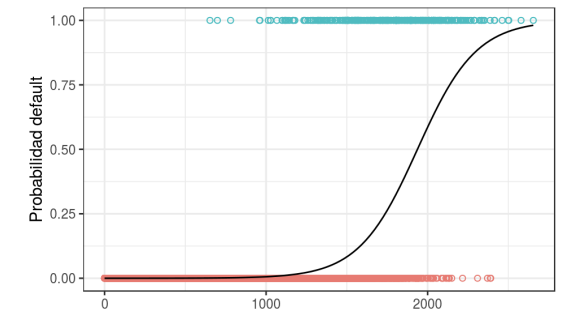
Se calcula la regresión lineal.

Paso 3

$\frac{1}{1 + e^{-(a+bX)}}$

Se transforma el resultado en una probabilidad final.

Regresión logística



Ejemplo

Ejemplo

Sean dos variables:

Datos		
Nro	X	Y
1	22	0
2	54	1
3	67	1
4	6	0
5	77	1
6	32	0
7	81	1
8	4	0
9	5	0
10	40	1
11		
12		
13		
14		
15		
38.80		0.50

$$\ln\left(\frac{p}{1-p}\right) = a + b_i X_i$$

Intercepto
Punto de corte

Pendiente
Coeficiente

Variable independiente
Variable predictora

Ejemplo

Solución:

- 1
- 2
- 3

Datos		
Nro	X	Y
1	22	0
2	54	1
3	67	1
4	6	0
5	77	1
6	32	0
7	81	1
8	4	0
9	5	0
10	40	1
11		
12		
13		
14		
15		
38.80		0.50

x (varianza)	x*Y	x ²
-16.80	0	282.24
15.20	15	231.04
28.20	28	795.24
-32.80	0	1075.84
38.20	38	1459.24
-6.80	0	46.24
42.20	42	1780.84
-34.80	0	1211.04
-33.80	0	1142.44
1.20	1	1.44
125		8026

Ejemplo

Solución:

Datos		
Nro	X	Y
1	22	0
2	54	1
3	67	1
4	6	0
5	77	1
6	32	0
7	81	1
8	4	0
9	5	0
10	40	1
11		
12		
13		
14		
15		
	38.80	0.50

x (varianza)	x*Y	x ²
-16.80	0	282.24
15.20	15	231.04
28.20	28	795.24
-32.80	0	1075.84
38.20	38	1459.24
-6.80	0	46.24
42.20	42	1780.84
-34.80	0	1211.04
-33.80	0	1142.44
1.20	1	1.44
	125	8026

4 Pendiente (b)

$$b = \frac{\sum(x * Y)}{\sum x^2}$$

$$b = 125 / 8026$$

$$b = 0.016$$

Ejemplo

Solución:

Datos		
Nro	X	Y
1	22	0
2	54	1
3	67	1
4	6	0
5	77	1
6	32	0
7	81	1
8	4	0
9	5	0
10	40	1
11		
12		
13		
14		
15		
	38.80	0.50

1	2	3
x (varianza)	x*Y	x ²
-16.80	0	282.24
15.20	15	231.04
28.20	28	795.24
-32.80	0	1075.84
38.20	38	1459.24
-6.80	0	46.24
42.20	42	1780.84
-34.80	0	1211.04
-33.80	0	1142.44
1.20	1	1.44
	125	8026

4

Pendiente (b)

$$b = 125 / 8026$$

$$b = 0.016$$

5

Intercepto (a)

$$Y = a + bX$$

$$a = \bar{y} - b\bar{x}$$

$$a = 0.5 - (0.016 * 38.8)$$

$$a = -0.104$$

Ejemplo

Solución:

Nro	Datos	
	X	Y
1	22	0
2	54	1
3	67	1
4	6	0
5	77	1
6	32	0
7	81	1
8	4	0
9	5	0
10	40	1
11		
12		
13		
14		
15		
	38.80	0.50

x (varianza)	x*Y	x ²	Ŷ (pronóstico)
-16.80	0	282.24	0.25
15.20	15	231.04	0.76
28.20	28	795.24	0.97
-32.80	0	1075.84	-0.01
38.20	38	1459.24	1.13
-6.80	0	46.24	0.41
42.20	42	1780.84	1.19
-34.80	0	1211.04	-0.04
-33.80	0	1142.44	-0.02
1.20	1	1.44	0.54
125		8026	

6 Pronóstico: \hat{Y}

$$Y_i = a + bX_i$$

$$\hat{Y}_1 = -0.104 + 0.016(22)$$

$$\hat{Y}_1 = -0.104 + 0.352$$

$$\hat{Y}_1 = 0.25$$

$$\hat{Y}_2 = -0.104 + 0.016(54)$$

$$\hat{Y}_2 = -0.104 + 0.864$$

$$\hat{Y}_2 = 0.76$$

$$\hat{Y}_2 = -0.104 + 0.016(67)$$

$$\hat{Y}_2 = -0.104 + 1.072$$

$$\hat{Y}_2 = 0.97$$

□ □ □

Ejemplo

Solución:

Datos							
Nro	X	Y	x (varianza)	x*Y	x ²	Ŷ (pronóstico)	Prob (ln)
1	22	0	-16.80	0	282.24	0.25	0.56
2	54	1	15.20	15	231.04	0.76	0.68
3	67	1	28.20	28	795.24	0.97	0.73
4	6	0	-32.80	0	1075.84	-0.01	0.50
5	77	1	38.20	38	1459.24	1.13	0.76
6	32	0	-6.80	0	46.24	0.41	0.60
7	81	1	42.20	42	1780.84	1.19	0.77
8	4	0	-34.80	0	1211.04	-0.04	0.49
9	5	0	-33.80	0	1142.44	-0.02	0.50
10	40	1	1.20	1	1.44	0.54	0.63
11							
12							
13							
14							
15							
38.80 0.50			125		8026		

7

$$\ln\left(\frac{p}{1-p}\right) = a + b_i X_i$$

$$\frac{1}{1 + e^{-(a+bX)}}$$

$$\ln_1 = 1/(1+e^{-(0.25)})$$

$$\ln_1 = 1/(1+ 2.718281828^{-0.25})$$

$$\ln_1 = 0.5621$$

$$\ln_2 = 1/(1+ 2.718281828^{-0.76})$$

$$\ln_2 = 0.6813$$

$$\ln_3 = 1/(1+ 2.718281828^{-0.97})$$

$$\ln_3 = 0.7251$$

...

Ejemplo

Solución:

Datos								
Nro	X	Y	x (varianza)	x*Y	x ²	Ŷ (pronóstico)	Prob (ln)	Clase
1	22	0	-16.80	0	282.24	0.24	0.56	1
2	54	1	15.20	15	231.04	0.74	0.68	1
3	67	1	28.20	28	795.24	0.94	0.73	1
4	6	0	-32.80	0	1075.84	-0.01	0.50	0
5	77	1	38.20	38	1459.24	1.09	0.76	1
6	32	0	-6.80	0	46.24	0.39	0.60	1
7	81	1	42.20	42	1780.84	1.16	0.77	1
8	4	0	-34.80	0	1211.04	-0.04	0.49	0
9	5	0	-33.80	0	1142.44	-0.03	0.50	0
10	40	1	1.20	1	1.44	0.52	0.63	1
11								
12								
13								
14								
15								
38.80 0.50			125		8026			

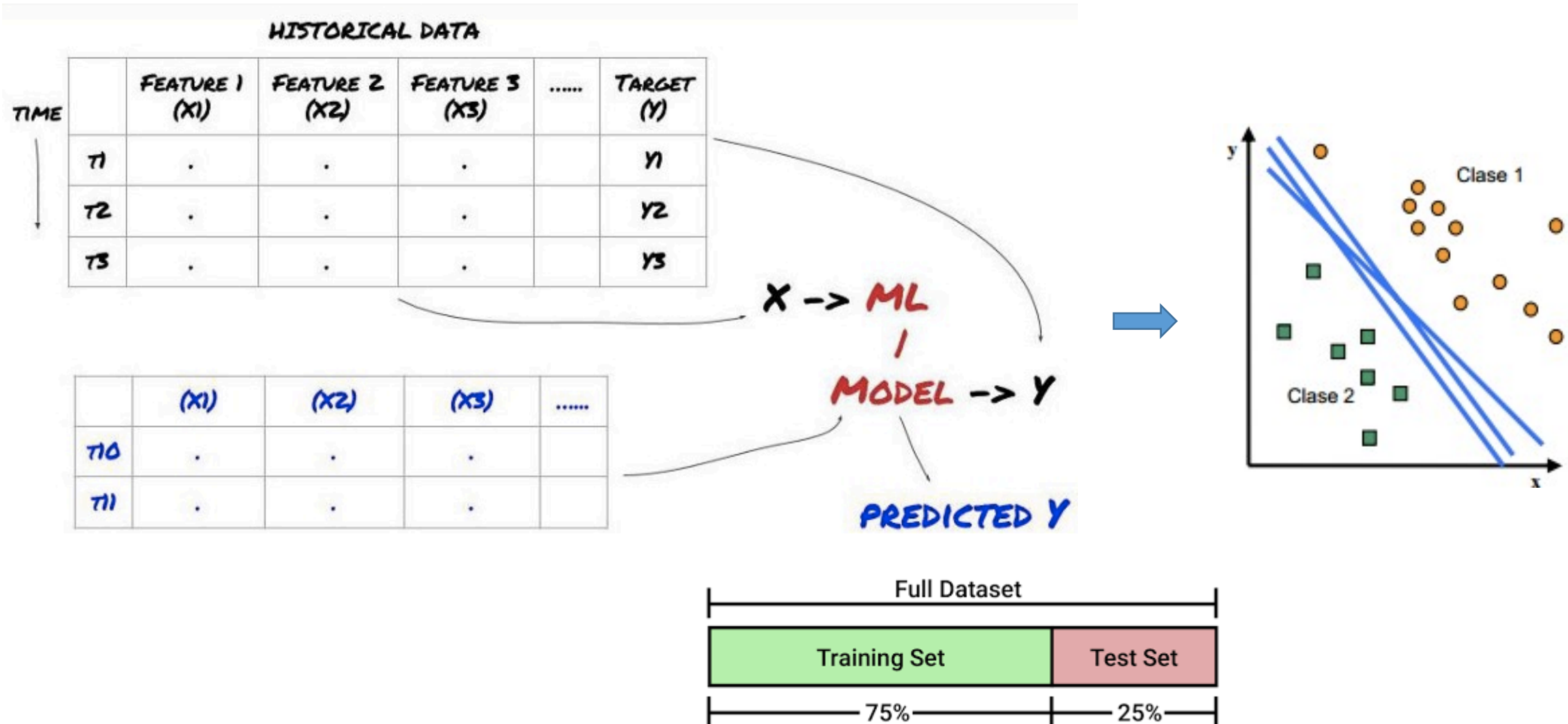
8 Clase

- Si la probabilidad es superior a 0.5 se asigna 1.
- Si es menor o igual a 0.5 se asigna 0.

Validación de la clasificación

1. Matriz de clasificación

Contexto



Matriz de clasificación

- Una matriz de clasificación, conocida también como matriz de confusión, se utiliza para evaluar una clasificación binaria.
- En la variable clase el conjunto de entrenamiento toma dos valores posibles: 0 o 1; positivo o negativo; falso o verdadero.
- Los valores positivos y negativos que se predicen correctamente se conocen como **verdaderos positivos (VP)** y **verdaderos negativos (VN)**, respectivamente.
- Mientras que los valores clasificados incorrectamente se denominan **falsos positivos (FP)** y **falsos negativos (FN)**.

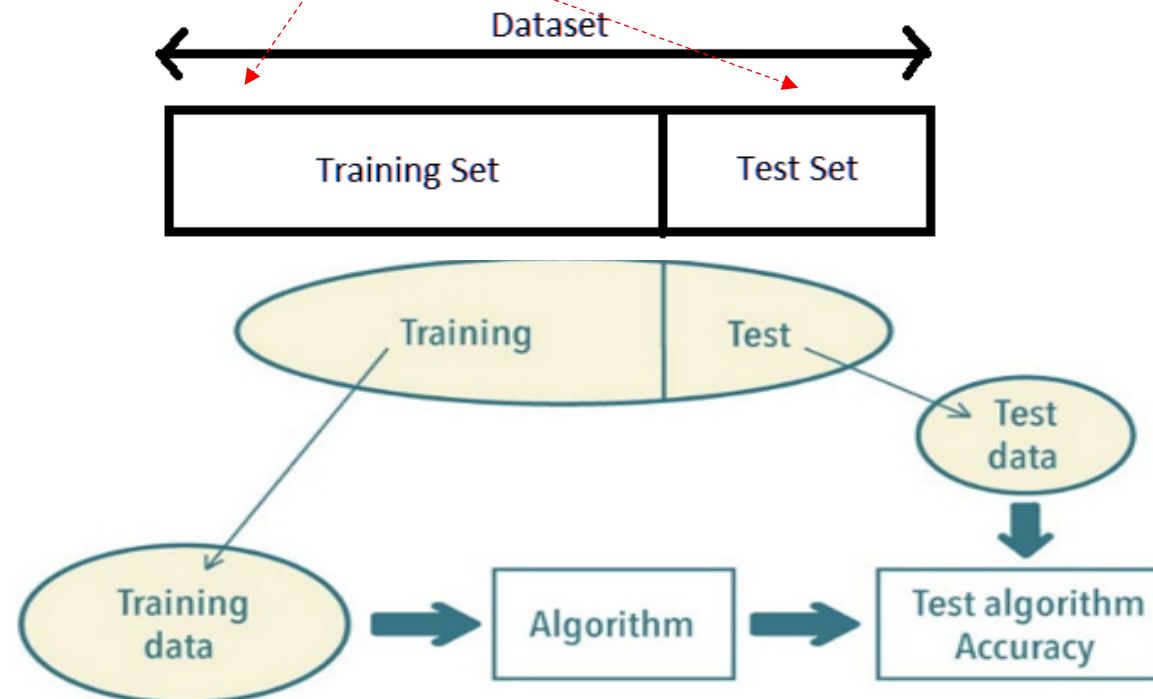
		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Matriz de clasificación

Criterio de división

Para utilizar este método de evaluación del modelo se necesita dividir los datos en:

- a) Datos de entrenamiento (*Training*): 80, 75, o 70%
- b) Datos de prueba (*Test*): 20, 25, o 30%



Matriz de clasificación

Procedimiento

- 1) Se evalúan todos los elementos y se determina si la **predicción (clase)** coincide con los **valores reales (Y)**.
- 2) Se cuentan todos los elementos y se muestran los totales obtenidos en la matriz.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Matriz de clasificación

Mediciones

- | | |
|------------------|---|
| 1) Exactitud | (Accuracy) |
| 2) Tasa de error | (Misclassification Rate) |
| 3) Precisión | (Precision) |
| 4) Sensibilidad | (Recall, Sensitivity, True Positive Rate) |
| 5) Especificidad | (Specificity, True Negative Rate) |

Matriz de clasificación

1) **Exactitud (Accuracy).** Es el porcentaje de datos clasificados correctamente.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

$$Exactitud = \frac{VP + VN}{Total} = \frac{VP + VN}{VP + VN + FP + FN}$$

Matriz de clasificación

2) **Precisión (Precision).** Es el porcentaje de clasificación positiva.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

$$Precisión = \frac{VP}{Total\ clasificados\ positivos} = \frac{VP}{VP + FP}$$

Matriz de clasificación

3) Tasa de error (Misclassification Rate). Porcentaje de datos clasificados incorrectamente.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

$$Tasa\ de\ error = \frac{FP + FN}{Total} = \frac{FP + FN}{VP + VN + FP + FN}$$

Matriz de clasificación

4) **Sensibilidad (True Positive Rate)**. Es el porcentaje de clasificación del total positivos.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

$$\text{Sensibilidad} = \frac{VP}{\text{Total positivos}} = \frac{VP}{VP + FN}$$

Matriz de clasificación

5) **Especificidad (True Negative Rate).** Es el porcentaje de clasificación del total negativos.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

$$\text{Especificidad} = \frac{VN}{\text{Total negativos}} = \frac{VN}{VN + FP}$$

Matriz de clasificación

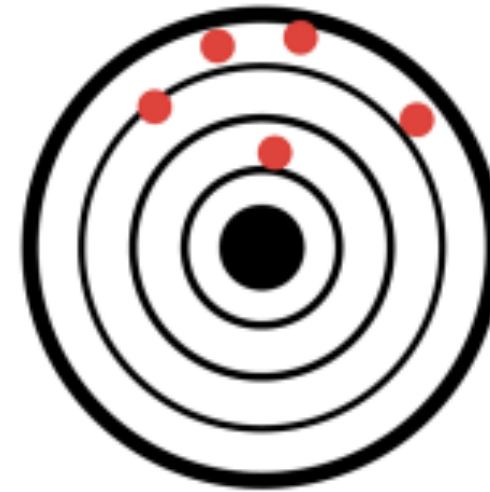
Eficiencia y precisión



Baja eficiencia pero
buena precisión



Buena eficiencia y
buena precisión



Baja eficiencia y
baja precisión

La **exactitud** simboliza el grado de conformidad, mientras que la **precisión** indica el grado de reproducibilidad.

Matriz de clasificación

Ejemplo 1

n = 2000

		Predicción	
		Positivo	Negativo
Observado (Real)	Positivo	VP 1100	FN 100
	Negativo	FP 60	VN 740

$$\text{Exactitud} = \frac{VP+VN}{VP+VN+FP+FN} = \frac{1100+740}{1100+740+60+100} = \frac{1840}{2000} = \mathbf{0.92}$$

$$\text{Precisión} = \frac{VP}{VP+FP} = \frac{1100}{1100+60} = \frac{1100}{1160} = \mathbf{0.95}$$

$$\text{Tasa de error} = \frac{FP+FN}{VP+VN+FP+FN} = \frac{60+100}{1100+740+60+100} = \frac{160}{2000} = \mathbf{0.08}$$

$$\text{Sensibilidad} = \frac{VP}{VP+FN} = \frac{1100}{1100+100} = \frac{1100}{1200} = \mathbf{0.916}$$

$$\text{Especificidad} = \frac{VN}{VN+FP} = \frac{740}{740+60} = \frac{740}{800} = \mathbf{0.925}$$

Matriz de clasificación

Ejemplo 2

n = 10000

		Predicción	
		Positivo	Negativo
Observado (Real)	Positivo	VP 3200	FN 340
	Negativo	FP 240	VN 6220

$$\text{Exactitud} = \frac{VP+VN}{VP+VN+FP+FN} = \frac{3200+6220}{3200+6220+340+240} = \frac{9420}{10000} = \mathbf{0.942}$$

$$\text{Precisión} = \frac{VP}{VP+FP} = \frac{3200}{3200+240} = \frac{3200}{3440} = \mathbf{0.93}$$

$$\text{Tasa de error} = \frac{FP+FN}{VP+VN+FP+FN} = \frac{240+340}{3200+6220+340+240} = \frac{580}{10000} = \mathbf{0.058}$$

$$\text{Sensibilidad} = \frac{VP}{VP+FN} = \frac{3200}{3200+340} = \frac{3200}{3540} = \mathbf{0.90}$$

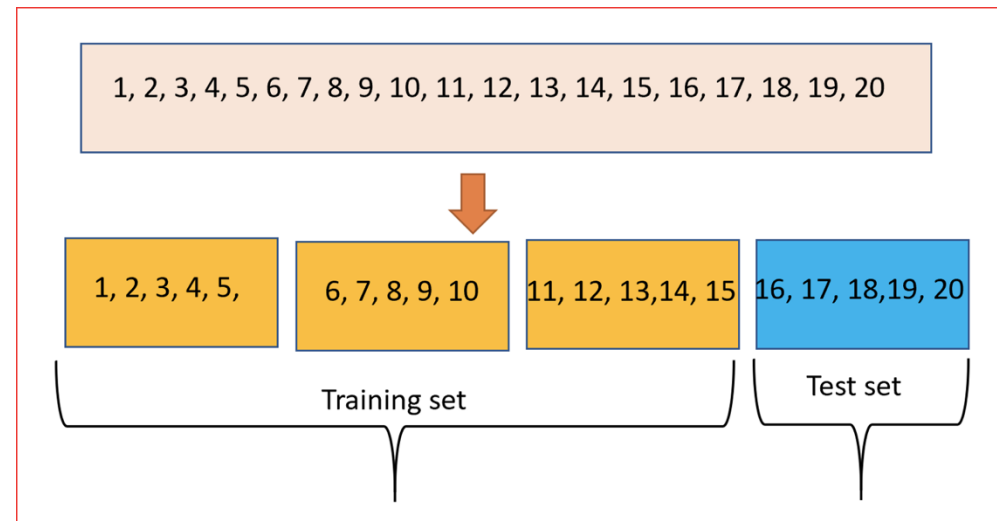
$$\text{Especificidad} = \frac{VN}{VN+FP} = \frac{6220}{6220+240} = \frac{6220}{6460} = \mathbf{0.96}$$

2. Validación cruzada

Validación cruzada

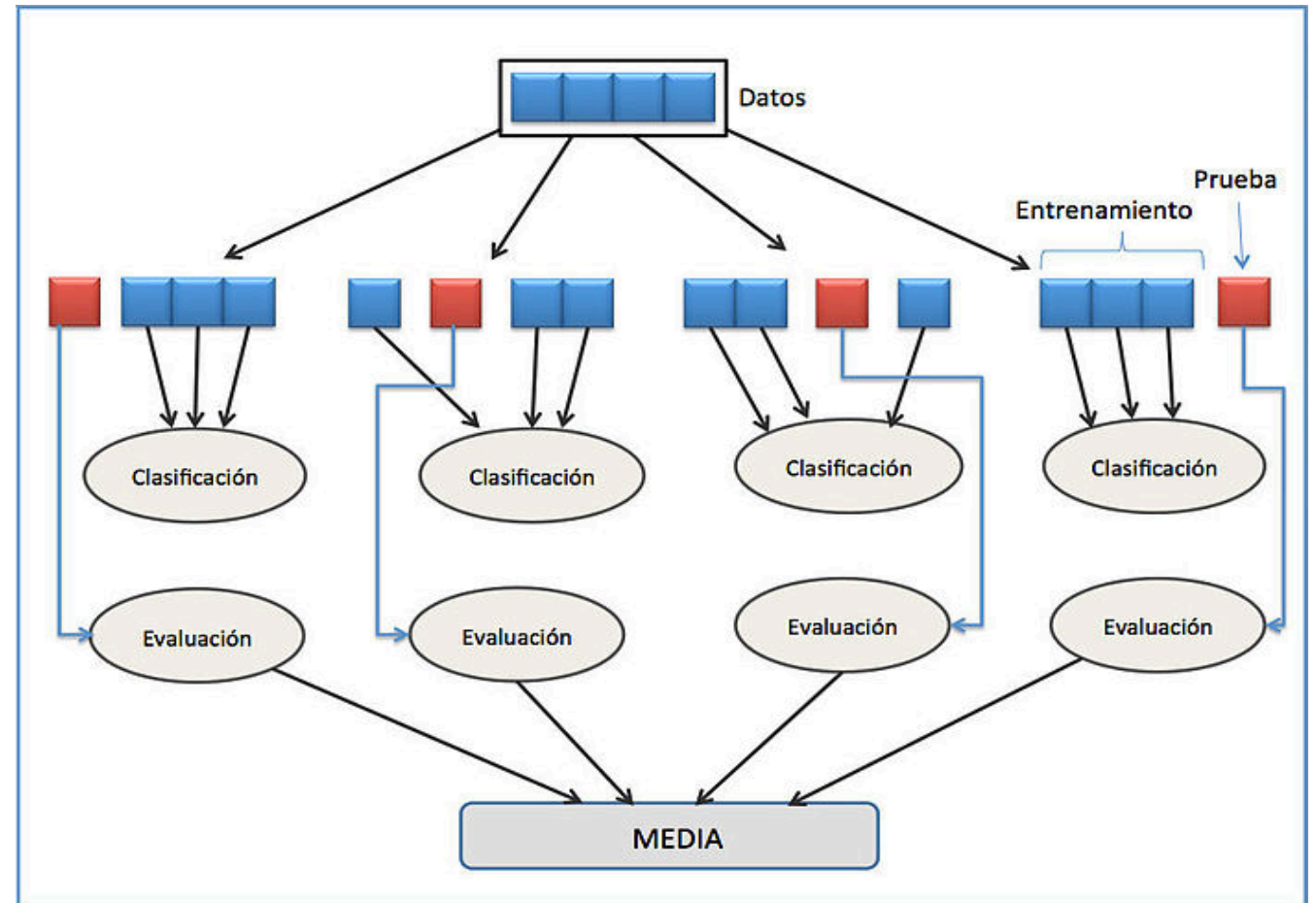
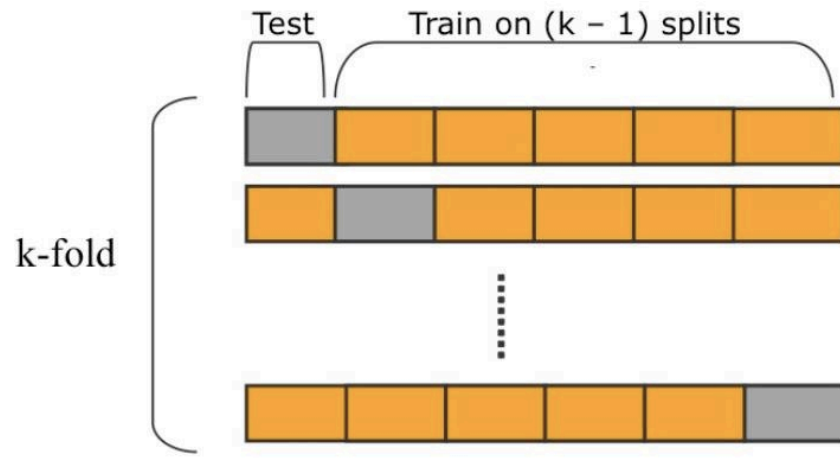
Consiste en dividir los datos en dos conjuntos: **entrenamiento** (*training*) y **prueba** (*test*).

- La clasificación se ajusta a un conjunto de datos de entrenamiento.
- Posteriormente, se calcula los valores de salida con los datos de prueba (**valores que no se han analizado antes**).
- La evaluación depende de la **división** entre los datos de entrenamiento y de prueba.



Validación cruzada

Para esta división se emplea el concepto de '**Cross validation**' de 'K' iteraciones.



Validación cruzada

Ejemplo

Se desea predecir la desafiliación de clientes y se requiere saber la eficiencia de la predicción. Una forma de lograr esto es mediante la **validación cruzada**.

	Plan_Internacional	Min_En_Dia	Min_Internacionales	Reclamos	Llamadas_Internacionales	Desafiliado	
TEST	no	265.1	10	1	3	no	no yes no
	no	129.1	12.7	4	6	yes	
	no	123	3	2	8	no	
Entrenamiento	no	116.9	7	0	10	yes	
	no	119.1	2.9	2	12	no	
	no	187.7	9.1	0	5	no	
	no	128.8	11.2	1	2	no	
	no	156.6	12.3	3	5	no	
	no	332.9	5.4	4	9	yes	

Iteración 1

$\frac{3}{3} = 100\%$ eficiencia en predicción

Validación cruzada

Ejemplo

	Plan_Internacional	Min_En_Dia	Min_Internacionales	Reclamos	Llamadas_Internacionales	Desafiliado
Entren...	no	265.1	10	1	3	no
	no	129.1	12.7	4	6	yes
	no	123	3	2	8	no
TEST	no	116.9	7	0	10	yes
	no	119.1	2.9	2	12	no
	no	187.7	9.1	0	5	no
Entren...	no	128.8	11.2	1	2	no
	no	156.6	12.3	3	5	no
	no	332.9	5.4	4	9	yes

Iteración 2

$\frac{1}{3} = 33\%$ eficiencia en predicción

Validación cruzada

Ejemplo

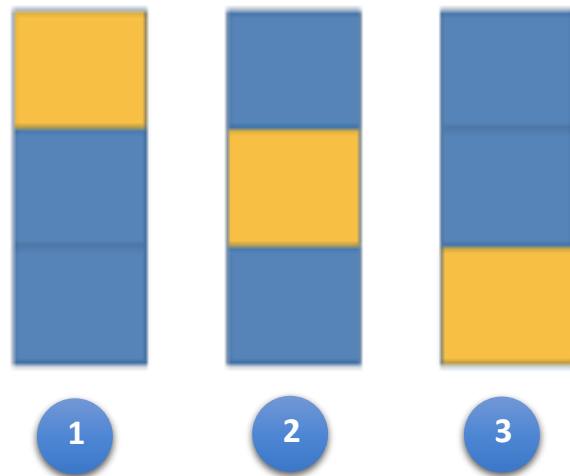
	Plan_Internacional	Min_En_Dia	Min_Internacionales	Reclamos	Llamadas_Internacionales	Desafiliado
Entrenamiento	no	265.1	10	1	3	no
	no	129.1	12.7	4	6	yes
	no	123	3	2	8	no
	no	116.9	7	0	10	yes
	no	119.1	2.9	2	12	no
	no	187.7	9.1	0	5	no
TEST	no	128.8	11.2	1	2	no
	no	156.6	12.3	3	5	no
	no	332.9	5.4	4	9	yes

Iteración 3

$\frac{2}{3} = 66\%$ eficiencia en predicción

Validación cruzada

Ejemplo



Iteración 1: $3/3 = 100\%$

Iteración 2: $1/3 = 33\%$

Iteración 3: $2/3 = 66\%$

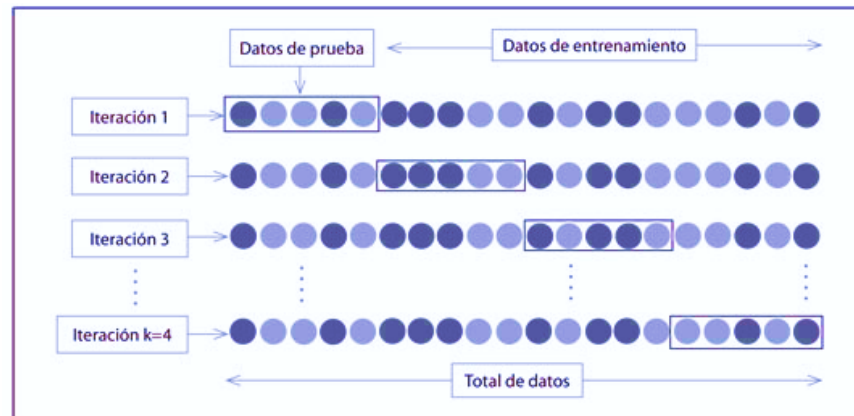
Promedio: 66.3%

Por lo tanto, la eficiencia de la predicción es de 66.3%

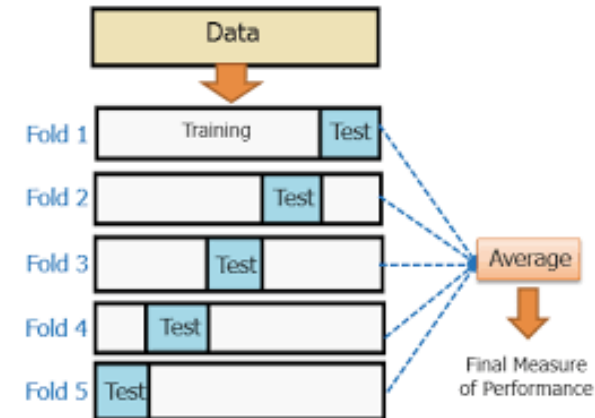
Validación cruzada

Divisiones comunes

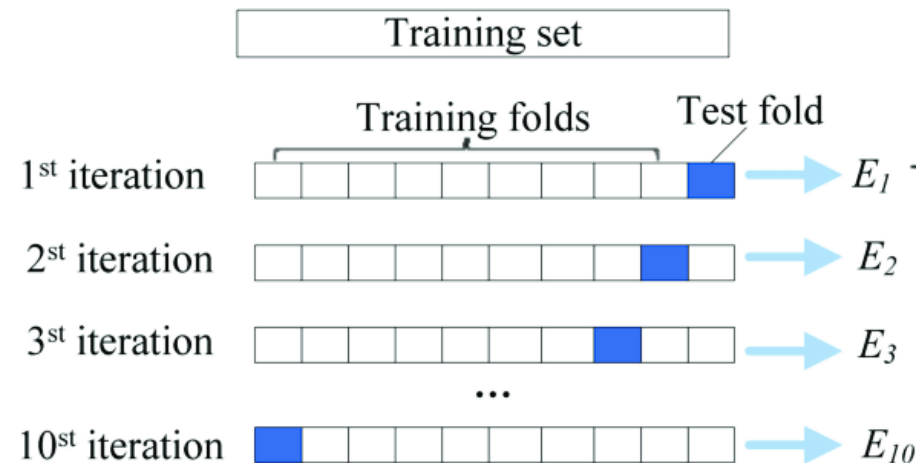
K = 4



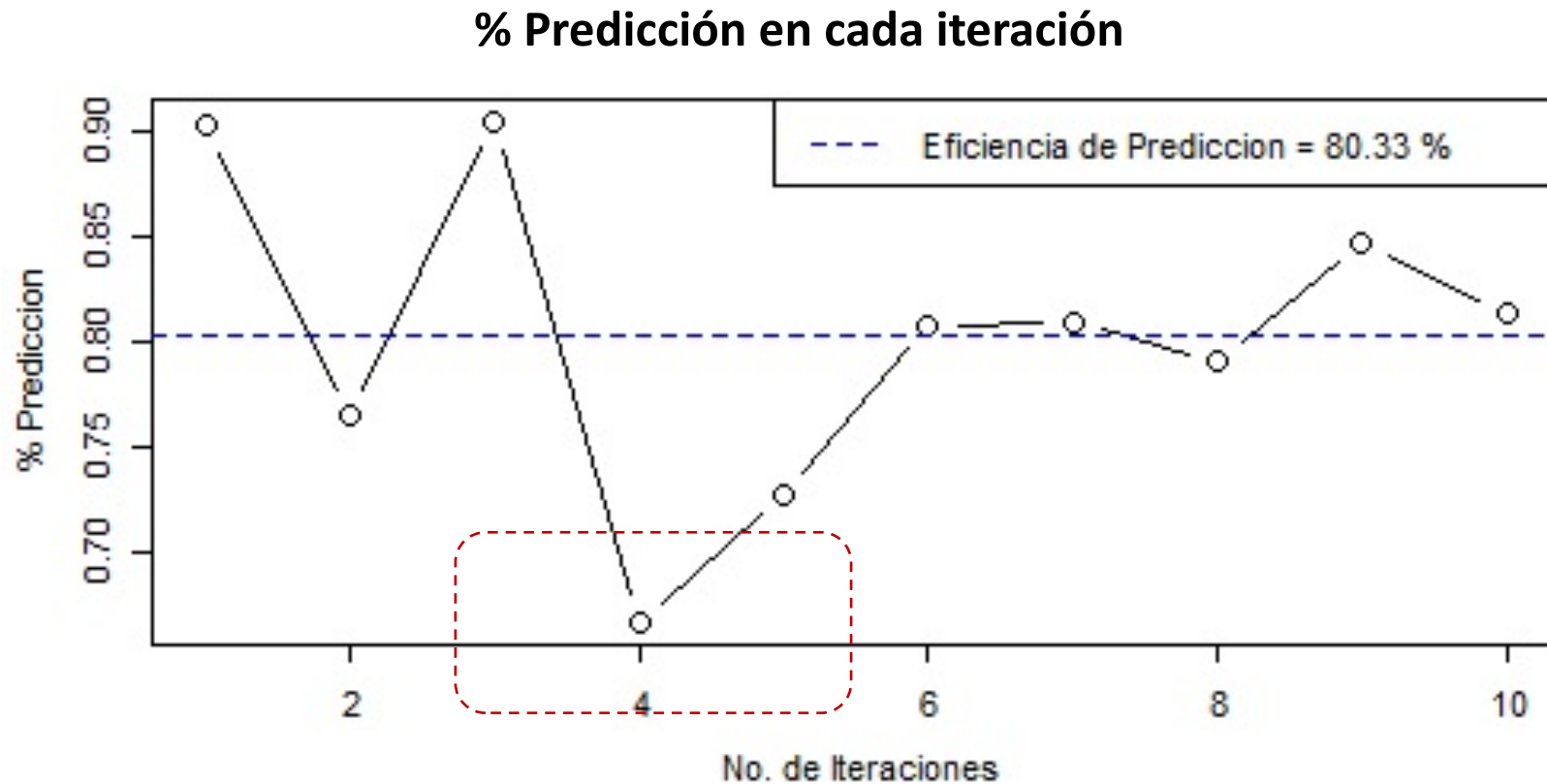
K = 5



K = 10



Validación cruzada



El % de predicción de cada iteración puede graficarse a partir del promedio de la eficiencia general del modelo predictivo.

Lo que se busca

Caso de estudio

- Registros de imágenes digitalizadas de **569** pacientes
- Variables independientes 10
- Variable dependiente 1 (**Diagnóstico**)
- Omitir la variable **Identificador**.

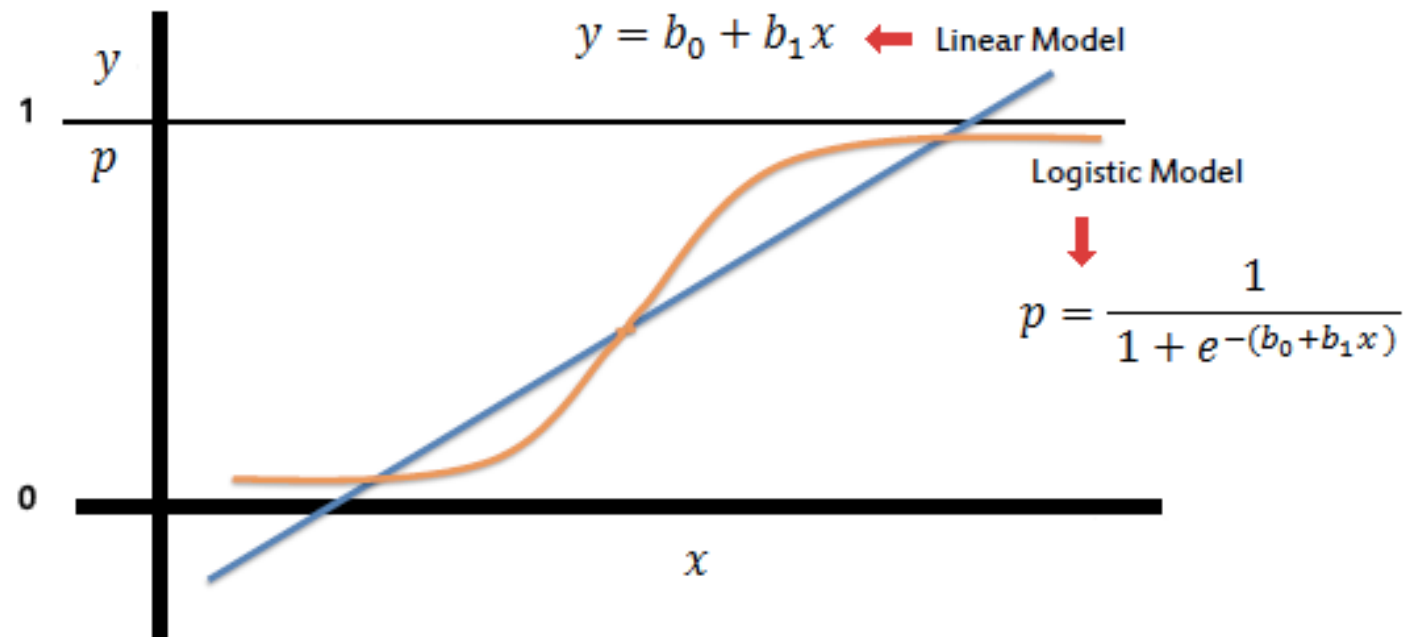
1	Identificador	Diagnosis	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity	Concave points	Symmetry	Fractal dimension
2	P-842302	0	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
3	P-842517	0	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
4	P-84300903	0	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999
5	P-84348301	0	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744
6	P-84358402	0	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883
7	P-843786	0	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613
8	P-844359	0	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742
9	P-84458202	0	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451
10	P-844981	0	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389
11	P-84501001	0	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243
12	P-845636	0	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697

0 = Maligno; 1 = Benigno


De las 569 observaciones, 357 son tumores benignos y 212 malignos

Regresión logística

- Diagnóstico (variable dependiente)



Sistemas de inferencia basadas en modelos de predicción

ID	<input type="text"/>	
Radius	<input type="text"/>	Compactness <input type="text"/>
Textura	<input type="text"/>	Concavity <input type="text"/>
Perimeter	<input type="text"/>	Concave_points <input type="text"/>
Area	<input type="text"/>	Symmetry <input type="text"/>
Smoothness	<input type="text"/>	Fractal_dimension <input type="text"/>
Diagnosis		<input type="text"/> 

Sistemas de inferencia basadas en modelos de predicción

ID	<input type="text"/>		
Area	<input type="text"/>	Concavity	<input type="text"/>
Texture	<input type="text"/>	Symmetry	<input type="text"/>
Compactness	<input type="text"/>	Fractal_dimension	<input type="text"/>
Diagnosis		<input type="text"/>	