Intuición sobre Regresión Logística

Luis R. Furlán Julio 2021

Conocimiento de fondo

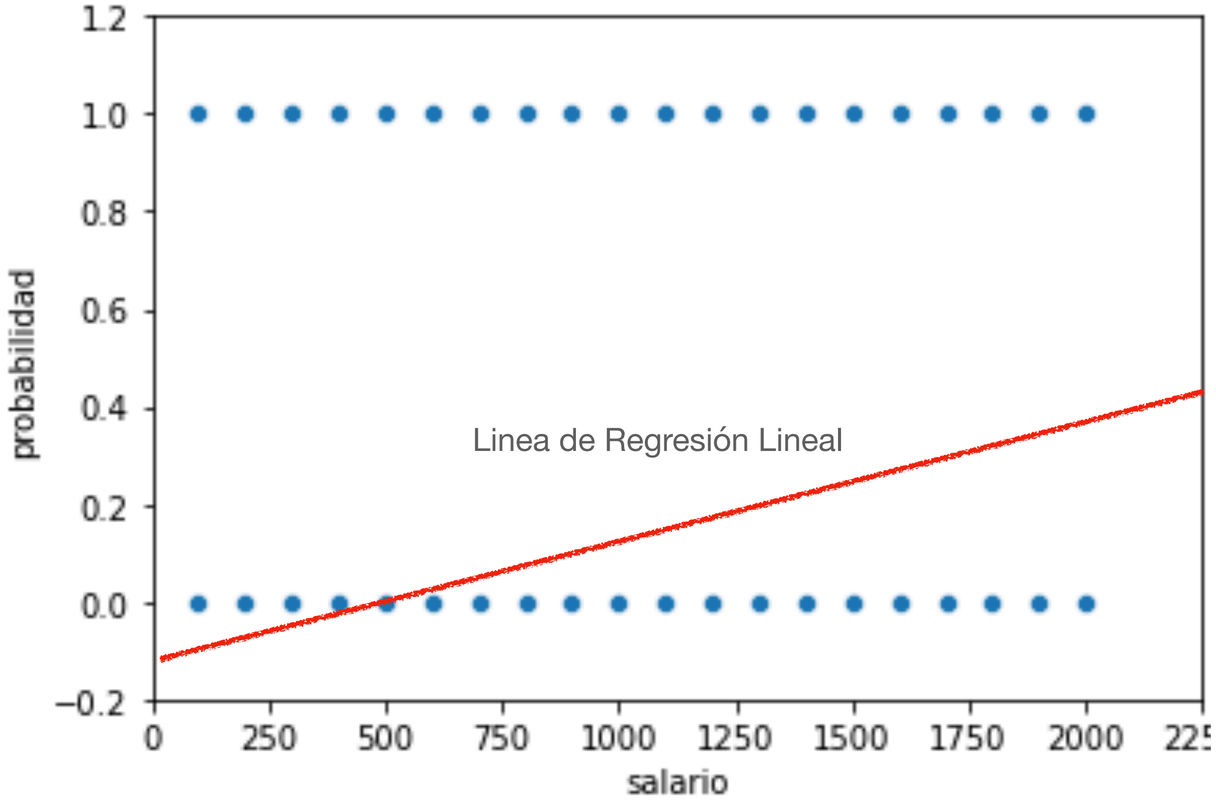
- Queremos saber sobre la Regresión Logística como un método de clasificación
- Algunos ejemplos de problemas de clasificación
  - eMails “Spam” vrs “Ham”
  - Default (dejar de pagar) en préstamos (si/no)
  - Diagnóstico de enfermedades
- Los anteriores son ejemplos de Clasificación Binaria

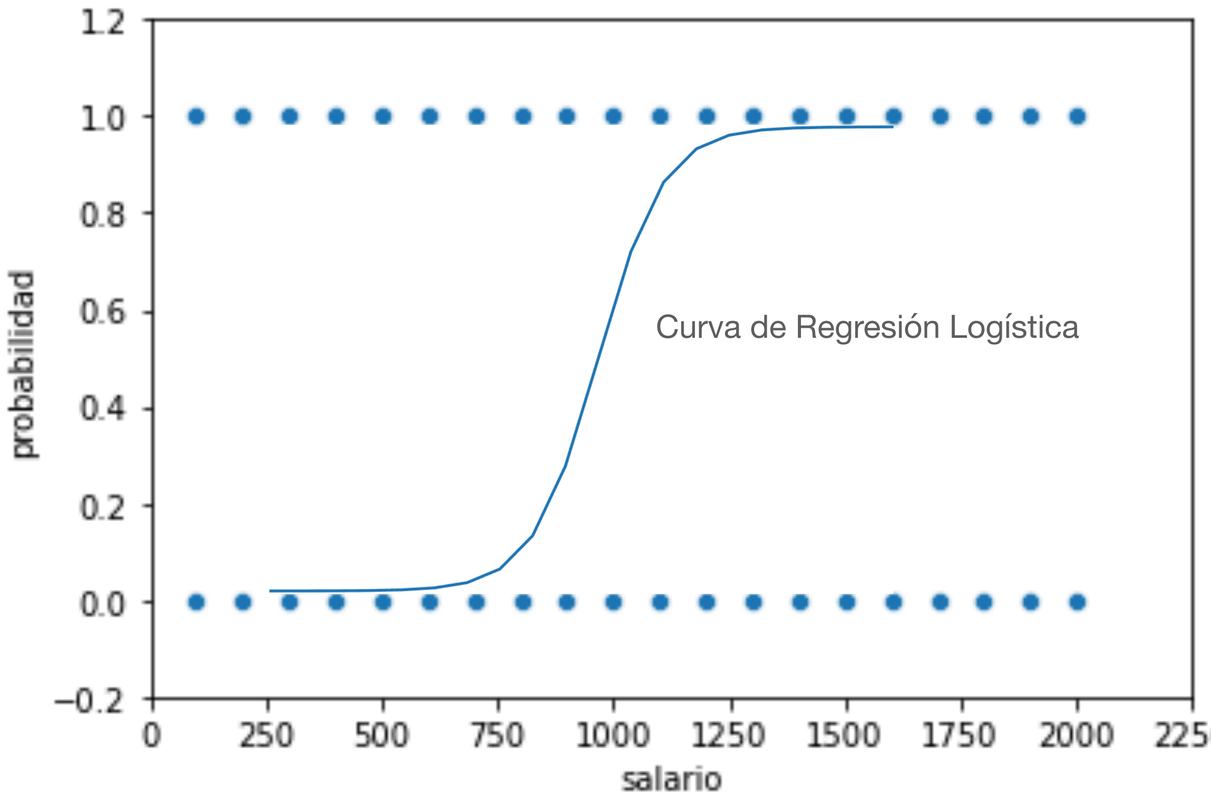
Conocimiento de fondo

- Hasta ahora solo hemos visto problemas de regresión en los que tratamos de predecir un valor continuo (ej. el precio de una casa).
- La regresión logística nos permite resolver problemas de clasificación en los que tratamos de predecir categorías discretas.
- La convención para la clasificación binaria es tener dos clases, 0 y 1

Conocimiento de fondo

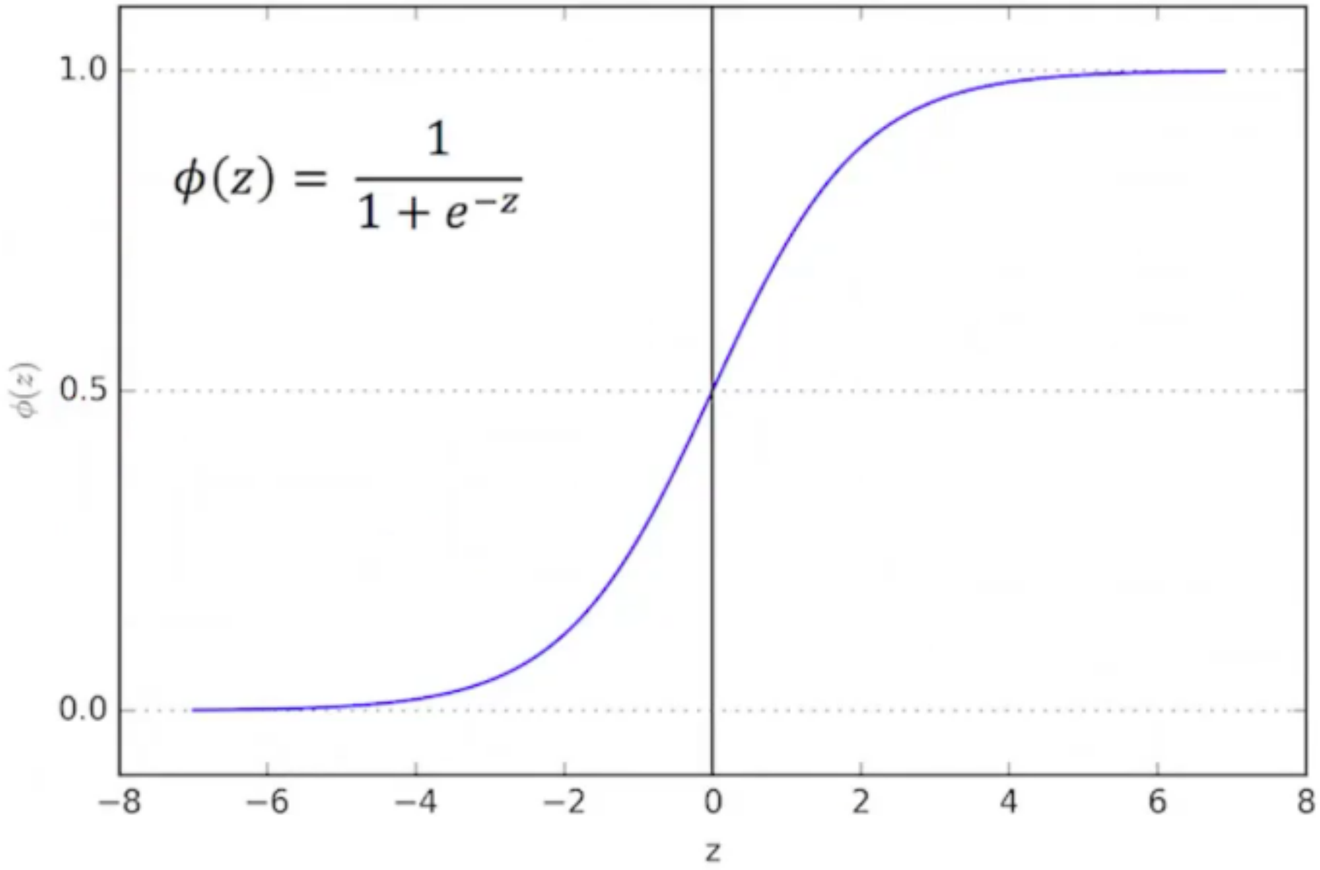
- No podemos utilizar un modelo de regresión lineal en grupos binarios. Simplemente no se ajusta





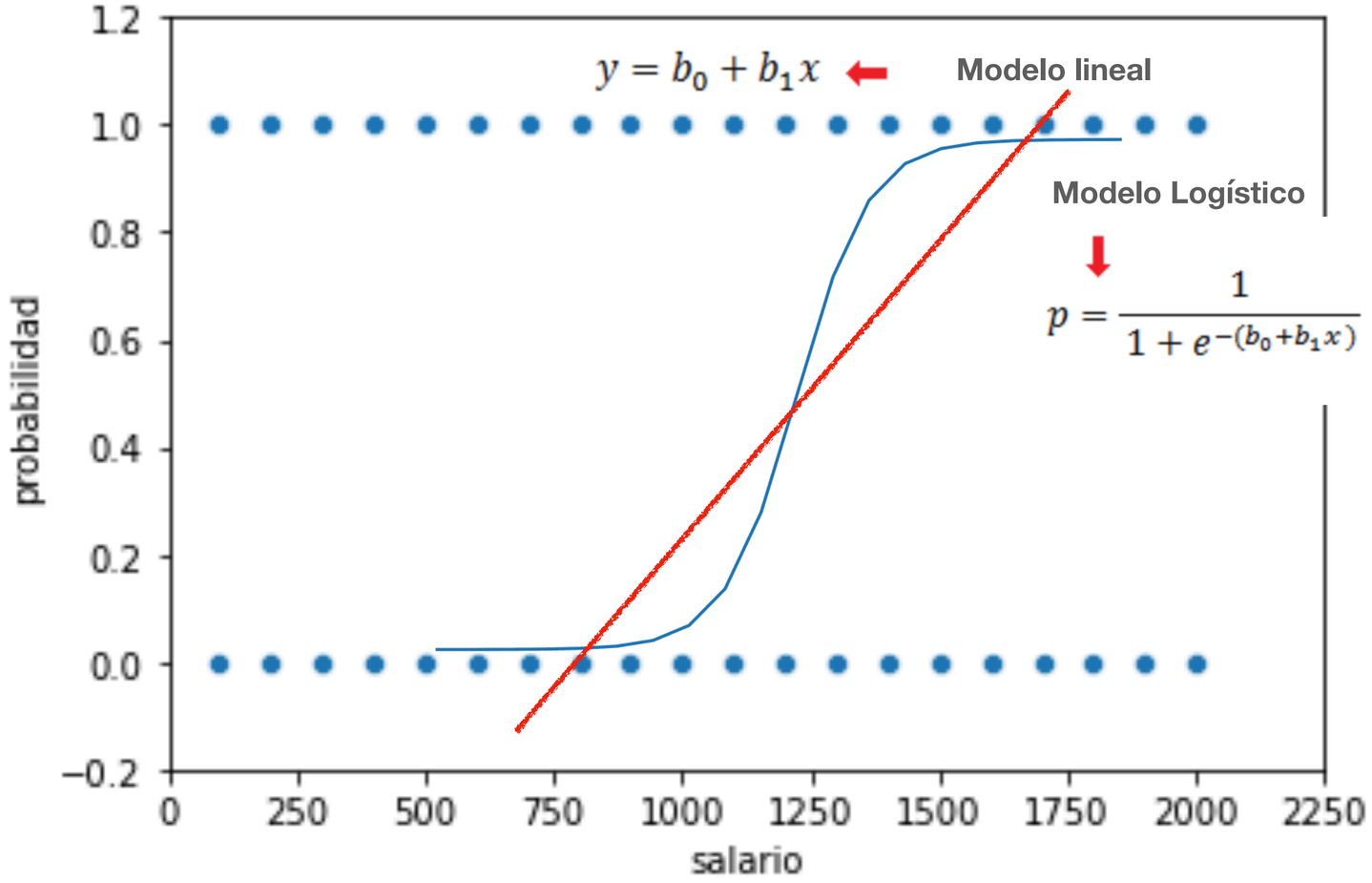
Función Sigmoide

- La función Sigmoide (Logística) recibe cualquier valor y retorna valores entre 0 y 1



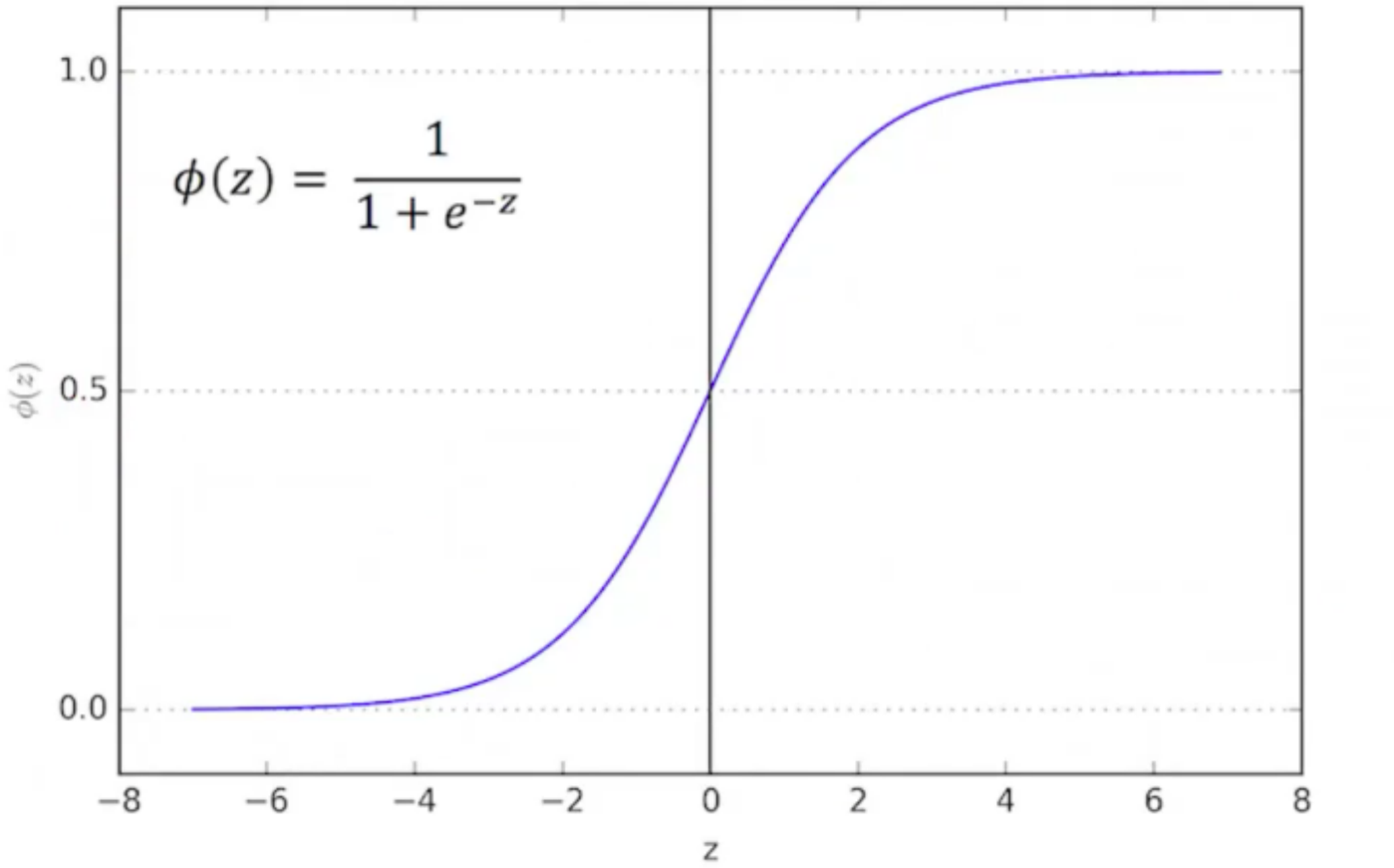
Función Sigmoide

- Esto quiere decir que podemos tomar nuestra solución de Regresión Lineal y colocarla en la función Sigmoide



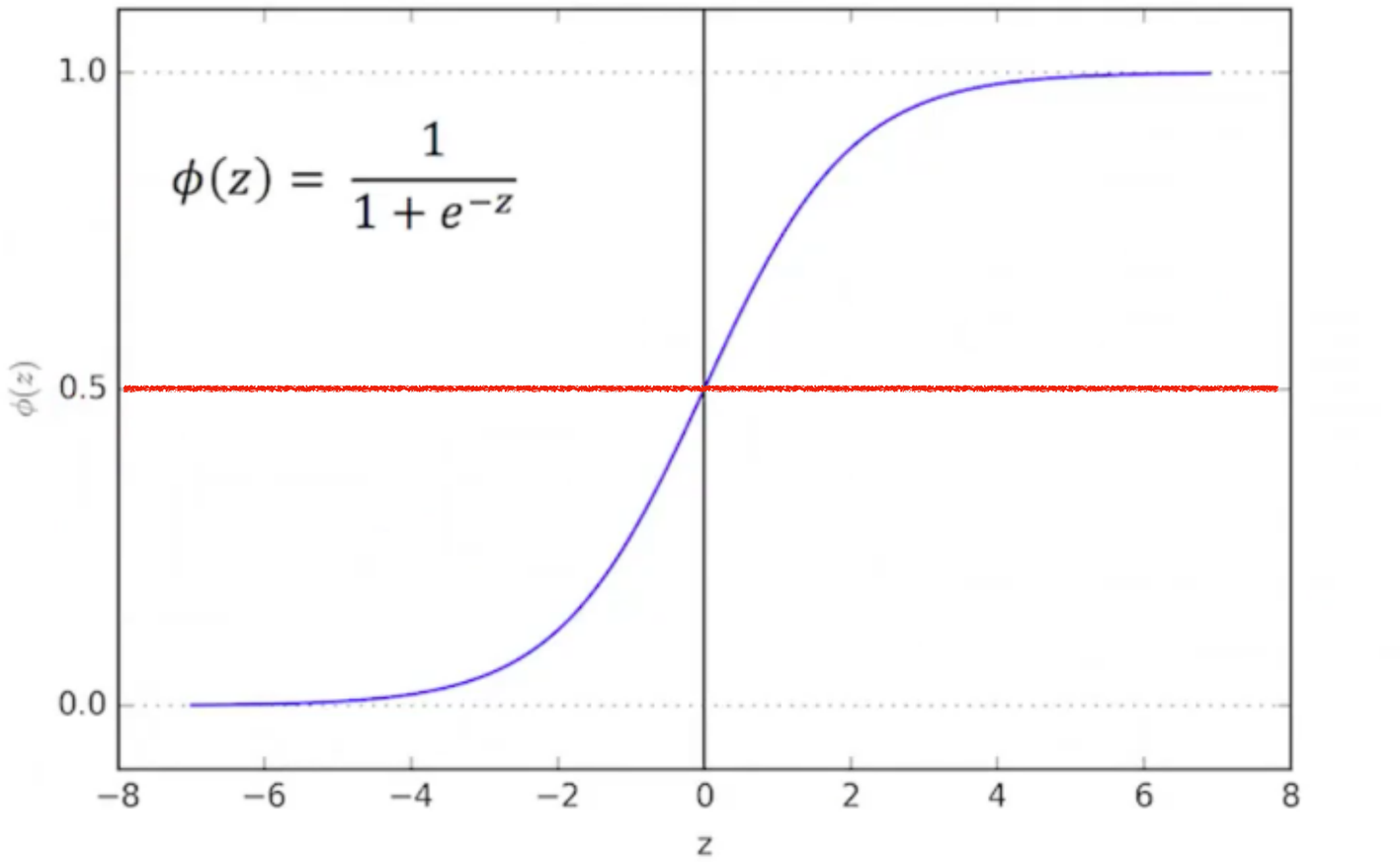
Función Sigmoide

- Esto resulta en una probabilidad de 0 a 1 de pertenecer a la clase 1



Función Sigmoide

- Podemos fijar un punto de corte en 0.5, y definir que cualquier cosa debajo resulta en la clase 0, cualquier cosa arriba es clase 1



Evaluación del modelo

- Luego de entrenar un modelo de regresión logística con datos de entrenamiento, se debe evaluar el modelo con datos de prueba.
- Podemos utilizar una matriz de confusión para evaluar modelos de clasificación

N = 165	Predijo: NO	Predijo: SI
Real: NO	50	10
Real: SI	5	100

Ejemplo: Prueba de presencia de enfermedad

NO = prueba negativa = Falso = 0
SI = prueba positiva = Verdadero = 1

Matriz de confusión

N = 165	Predijo: NO	Predijo: SI
Real: NO	TN = 50	FP = 10
Real: SI	FN = 5	TP = 100

Terminología Básica:

True Positives = TP
True Negatives = TN
False Positive = FP (Error Tipo I)
False Negatives = FN (Error Tipo II)

Matriz de confusión

N = 165	Predijo: NO	Predijo: SI	
Real: NO	TN = 50	FP = 10	60
Real: SI	FN = 5	TP = 100	105
	55	110	

Exactitud:

En general, ¿qué tan frecuente es correcto?

(TP + TN) / total = 150/165 = 0.91

Matriz de confusión

N = 165	Predijo: NO	Predijo: SI	
Real: NO	TN = 50	FP = 10	60
Real: SI	FN = 5	TP = 100	105
	55	110	

Tasa de Error (Misclassification Rate):

En general, ¿con qué frecuencia está errado?

(FP + FN) / total = 15/165 = 0.09

Matriz de confusión

N = 165	Predijo: NO	Predijo: SI	
Real: NO	TN = 50	FP = 10	60
Real: SI	FN = 5	TP = 100	105
	55	110	

Precisión:

Habilidad de no etiquetar una muestra como positiva, cuando es negativa

TP / (TP + FP) = 100/110 = 0.91

Matriz de confusión

N = 165	Predijo: NO	Predijo: SI	
Real: NO	TN = 50	FP = 10	60
Real: SI	FN = 5	TP = 100	105
	55	110	

Reconocimiento (Recall):

Habilidad del clasificador para encontrar todos los casos positivos

TP / (TP + FN) = 100/105 = 0.95

Matriz de confusión

N = 165	Predijo: NO	Predijo: SI	
Real: NO	TN = 50	FP = 10	60
Real: SI	FN = 5	TP = 100	105
	55	110	

Punteo F-beta:

Puede interpretarse como una media harmónica ponderada de la precisión y el reconocimiento (recall).

F-1 = (precisión + recall) / 2 (con beta = 1)

El punteo F-beta llega a su mejor valor cuando se acerca a 1 y el peor cuando se acerca a 0.

El punteo F-beta le da mayor peso a recall por un facto de beta. Un beta = 1 quiere decir que ambos factores tienen la misma importancia.

Matriz de confusión

N = 165	Predijo: NO	Predijo: SI	
Real: NO	TN = 50	FP = 10	60
Real: SI	FN = 5	TP = 100	105
	55	110	

Soporte (Support):

El número de ocurrencias de cada clase:

NO: TN + FP = 60
SI: FN + TP = 105

Tipos de error

Error de Tipo I
Falso Positivo

Error de Tipo II
Falso Negativo

Ejemplo con Python

Exploraremos un ejemplo de Regresión Logística utilizando el famoso conjunto de datos sobre el Titanic para tratar de predecir si un pasajero sobrevivió o no, en base a las características (features) del pasajero.