

Tipos de problemas

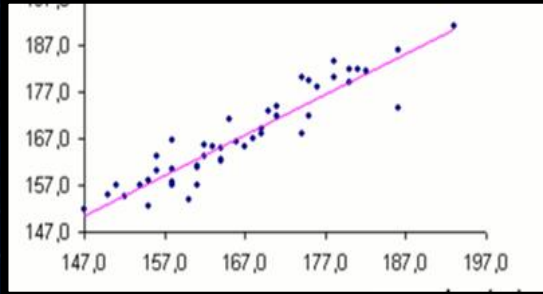
The background of the slide is a photograph of a clear night sky. The Milky Way galaxy is visible as a bright, hazy band of light stretching across the upper half of the frame. Numerous individual stars are scattered throughout the sky. In the lower portion of the image, the dark, silhouetted branches of evergreen trees are visible against the starry background.

Aprendizaje Automático. Tipos de problemas.

Tipo de problema

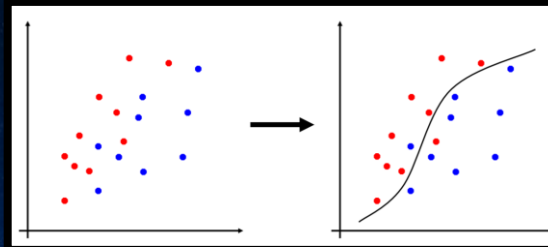
Regresión

Ej: dadas las características de una persona, cuál debería ser su peso?

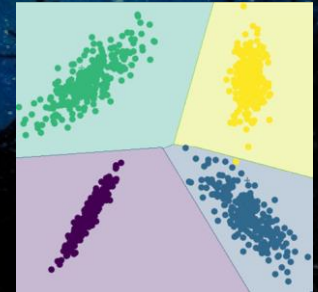


Clasificación

Clasificación
Binaria



Clasificación
multiclase



Datos

Para hacer Machine Learning necesitamos DATOS

	A	B	C	D	E
1	sepal.length	sepal.width	petal.length	petal.width	class
2	5,10	3,50	1,40	0,20	Iris-setosa
3	4,90	3,00	1,40	0,20	Iris-setosa
4	4,70	3,20	1,30	0,20	Iris-setosa
5	4,60	3,20	1,40	0,20	Iris-setosa
6	5,30	3,70	1,50	0,20	Iris-setosa
7	5,00	3,30	1,40	0,20	Iris-setosa
8	7,00	3,20	4,70	1,40	Iris-versicolor
9	6,40	3,20	4,50	1,50	Iris-versicolor
10	6,90	3,10	4,90	1,50	Iris-versicolor
11	5,50	2,30	4,00	1,30	Iris-versicolor
12	7,60	3,00	6,60	2,10	Iris-virginica
13	4,90	2,50	4,50	1,70	Iris-virginica
14	7,30	2,90	6,30	1,80	Iris-virginica
15	6,70	2,50	5,80	1,80	Iris-virginica
16	7,20	3,60	6,10	2,50	Iris-virginica
17	6,50	3,20	5,10	2,00	Iris-virginica
18	6,40	2,70	5,30	1,90	Iris-virginica
19	6,80	3,00	5,50	2,10	Iris-virginica

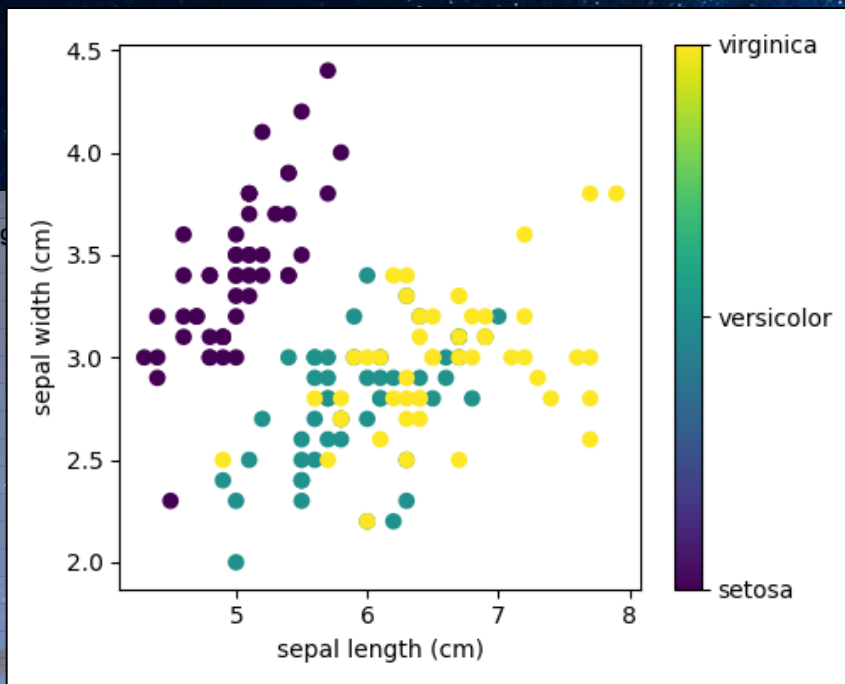


Un dataset típico

- Generalmente almacenado como archivo CSV o XLS.
- Cada fila es un registro (un ejemplo)
- Cada columna es un atributo (variable o feature)
- Dos tipos de datos: numéricos y nominales.
- Siempre necesitamos datos numéricos.
- Aprendizaje supervisado.

Visualización de los Datos

¿Cómo podemos visualizar el dataset?



- Al tener dos atributos podemos hacer un gráfico 2D.
- ¿Qué pasaría si tenemos 3 o más atributos?

	A
1	sepal length
2	5,10
3	4,90
4	4,70
5	4,60
6	5,30
7	5,00
8	7,00
9	6,40
10	6,90
11	5,50
12	7,60
13	4,90
14	7,30
15	6,70
16	7,20
17	6,50
18	6,40
19	6,80

2,70	5,30	1,90	Iris-virginica
3,00	5,50	2,10	Iris-virginica

Problemas de Regresión

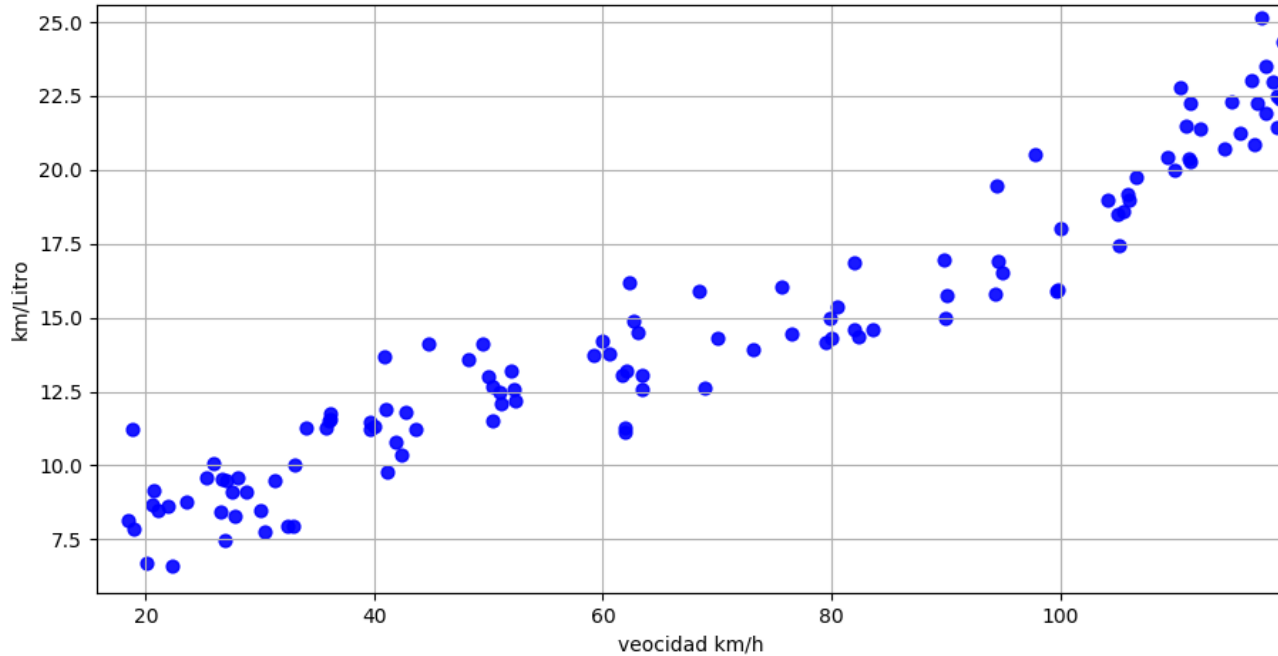
Supongamos el siguiente problema:

Queremos predecir cuál es el rendimiento (km/L) que tiene un auto en base a la velocidad a la que circula.

velocidad km/h	km/Litro
20	8.00
75	15.20
22	8.60
10	18.00
51	12.50
52	13.20
60	14.20

Problemas de Regresión

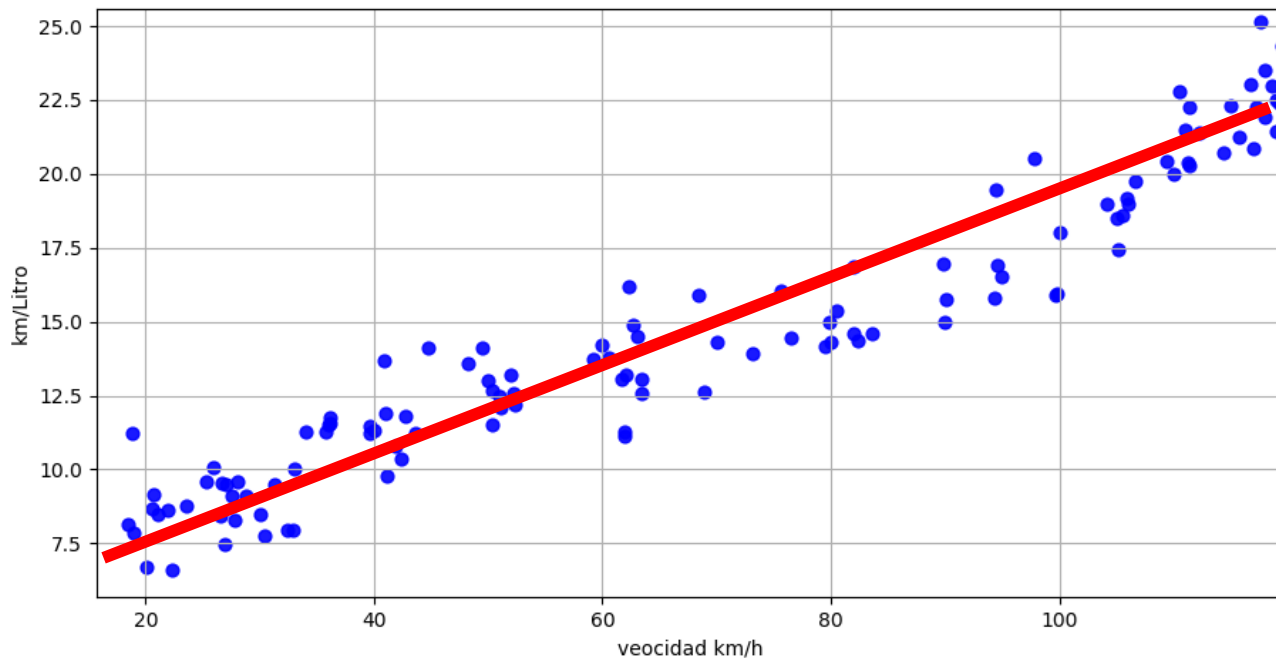
¿Cómo podemos generar un modelo que resuelva este problema?



velocidad km/h	km/Litro
20	8.00
75	15.20
22	8.60
10	18.00
51	12.50
52	13.20
60	14.20

Problemas de Regresión

¿Cómo podemos generar un modelo que resuelva este problema?



Podríamos generar una ecuación lineal.

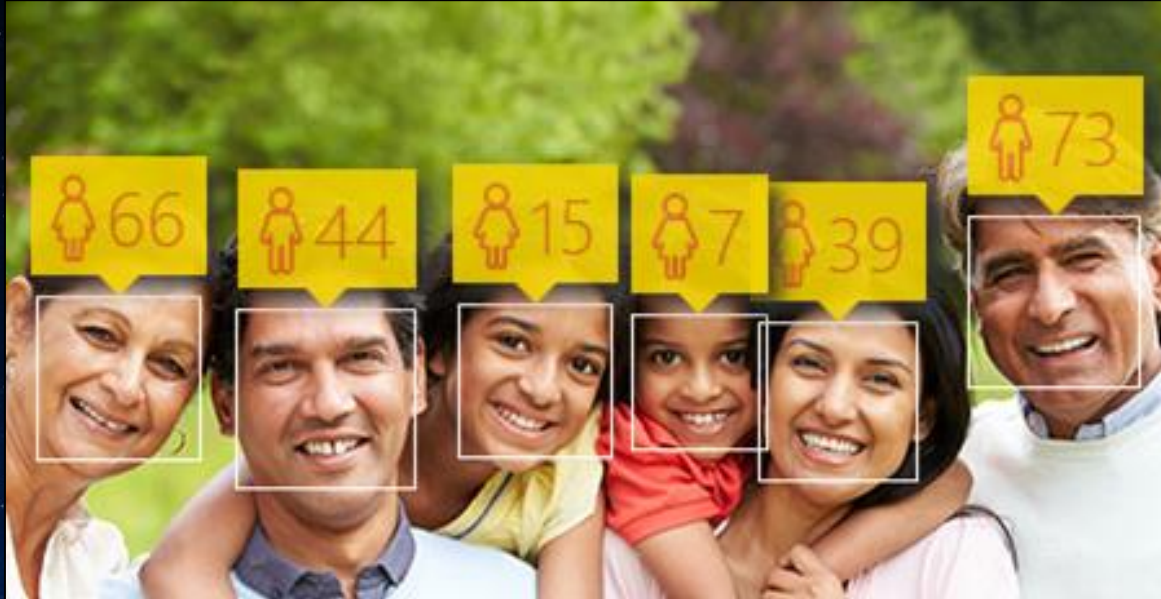
El modelo va a tener un error (ya que muy pocos elementos pasan por la recta), pero es simple y funciona “bien”.

$$y = 0.2X + 7$$

Problemas de Regresión

En el caso de imágenes

Input:
una imagen



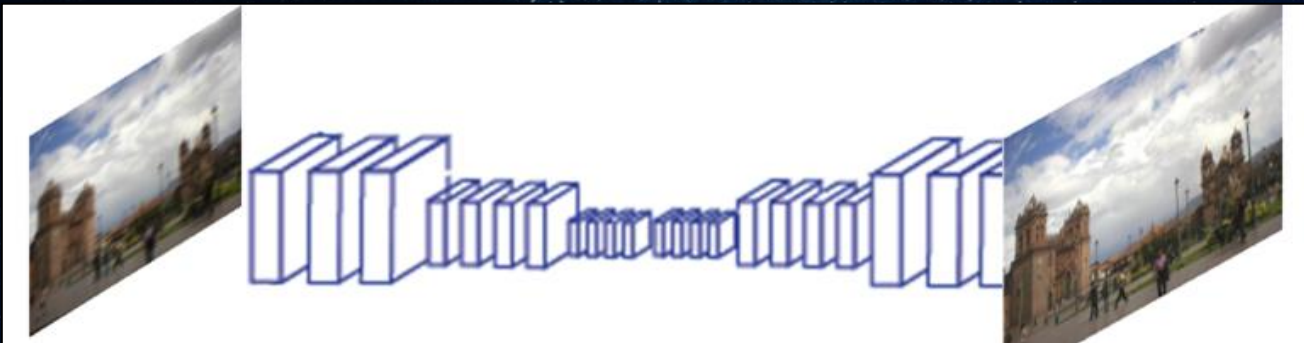
Output:
Un valor real

Ejemplo: estimar la edad de una persona.

Problemas de Regresión

En el caso de imágenes

Input:
una imagen

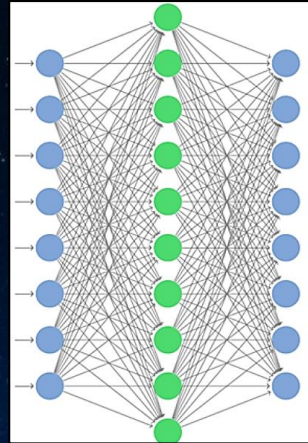


Output:
una imagen



Problemas de Clasificación

Clasificación binaria



Modelo

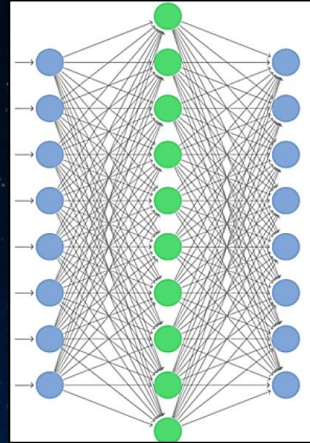


Perro

No Perro

Problemas de Clasificación

Clasificación Multiclase



Modelo



Gato

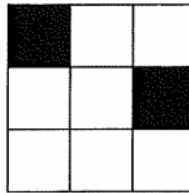
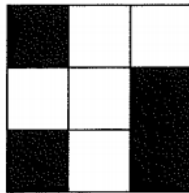
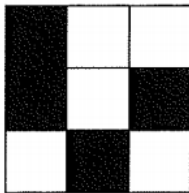
Perro

Ave

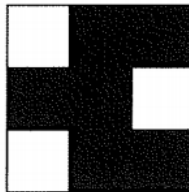
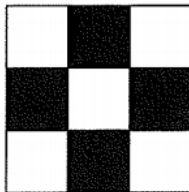
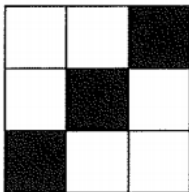
Rana

Ciervo

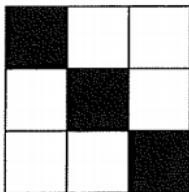
Problemas de Clasificación



$$f = -1$$



$$f = +1$$



$$f = ?$$

En el Aprendizaje Automático, el modelo aprenderá en base a los datos. Si los datos no son representativos, el modelo no generalizará bien a nueva información.

Problemas de Clasificación

Supongamos el siguiente problema:

Queremos clasificar dos especies de flores según el largo y ancho de sus pétalos. Tenemos la siguiente información. X1= Largo, X2= Ancho.

x1	x2	Clase
0,7	0,2	0
2	0,5	0
3	1	0
3	2	0
0,5	3	1
1	2	1
1,5	3,5	1
3	3,5	1

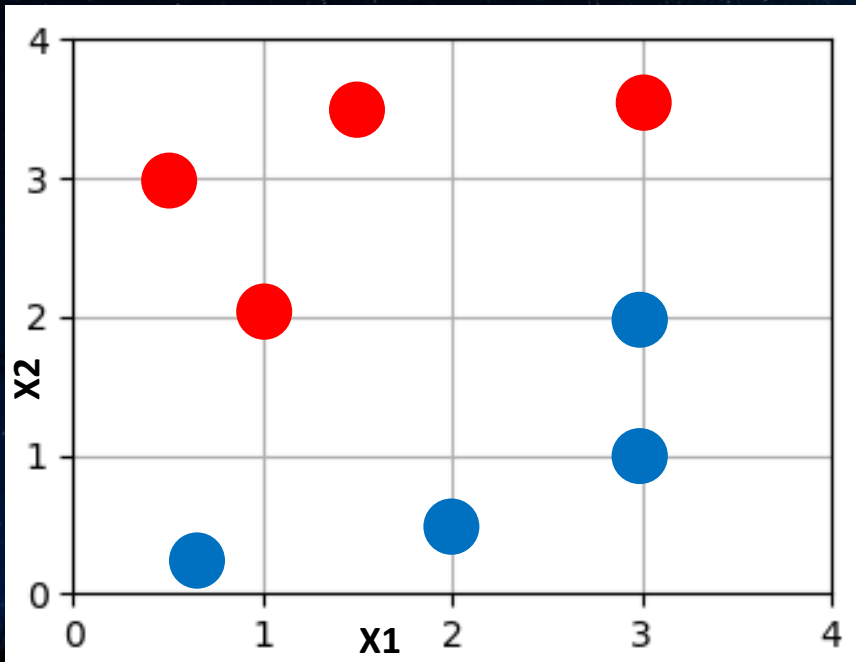
A veces, la “clase”
estará en formato
nominal (texto).

Ej. “flor_tipo_1” y
“flor_tipo_2”

Problema de **clasificación binaria**, que
podemos afrontar con **aprendizaje
supervisado**.

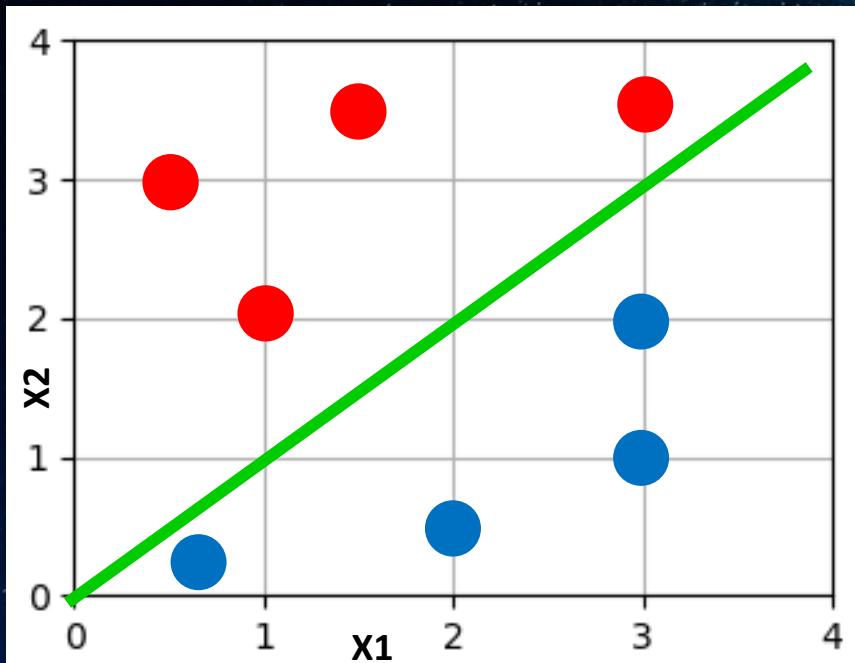
Problemas de Clasificación

Visualización en 2D



x_1	x_2	Clase
0,7	0,2	0
2	0,5	0
3	1	0
3	2	0
0,5	3	1
1	2	1
1,5	3,5	1
3	3,5	1

Problemas de Clasificación

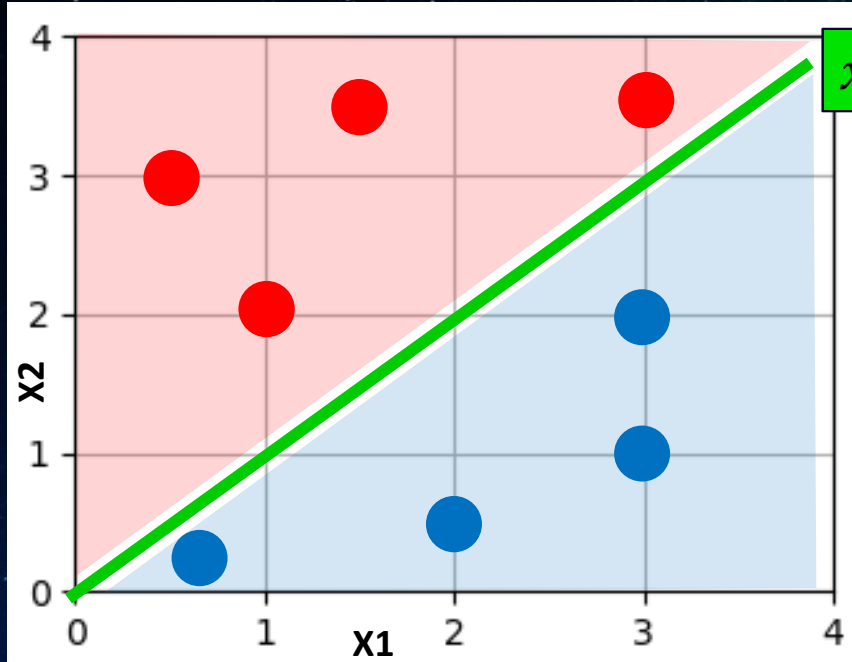


El modelo más simple que podemos pensar: **Lineal**

Ecuación de decisión:
 $x_1 = x_2$

$$g(x) = \begin{cases} 0 & \text{si } x_1 - x_2 > 0 \\ 1 & \text{si } x_1 - x_2 \leq 0 \end{cases}$$

Problemas de Clasificación

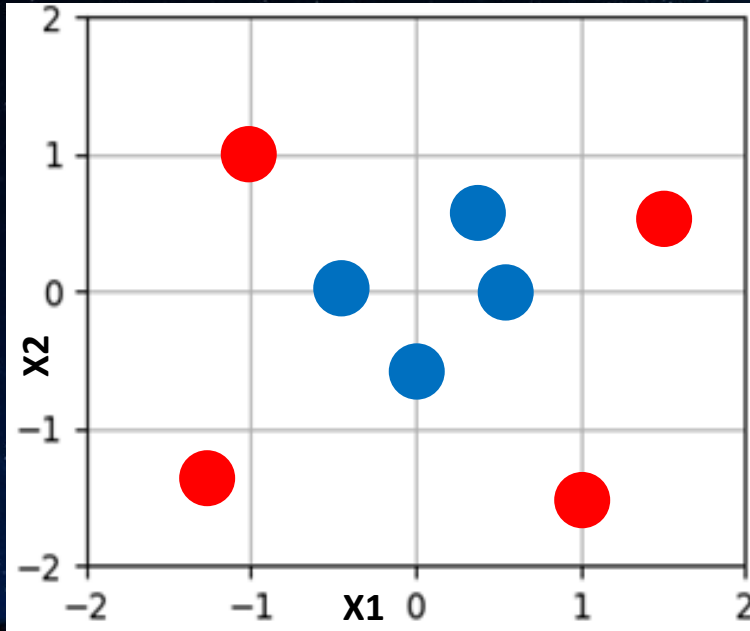


$$x_1 - x_2 = 0$$

**Frontera de
decisión**

Problemas de Clasificación

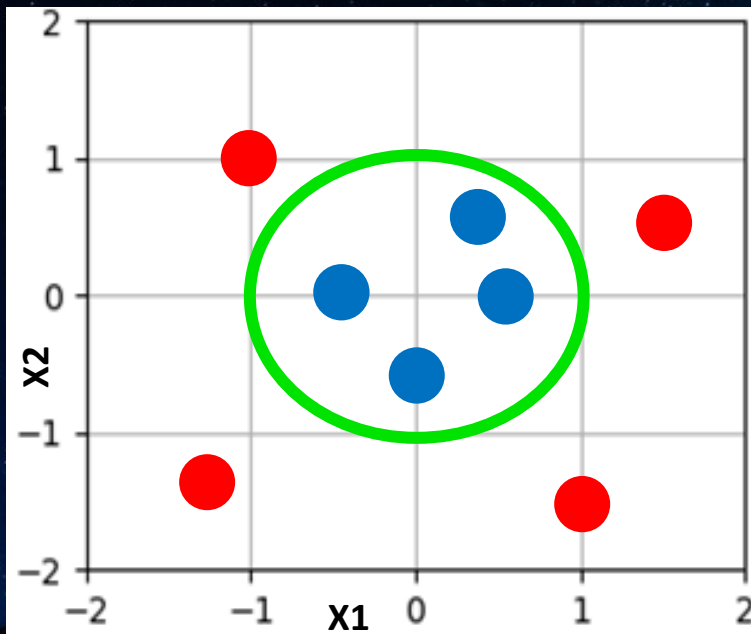
¿Y ahora?



x_1	x_2	Clase
0	-0,5	0
0,5	0	0
-0,5	0	0
0,3	0,5	0
-1,3	-1,3	1
1	-1,5	1
-1	1	1
1,5	0,5	1

Problemas de Clasificación

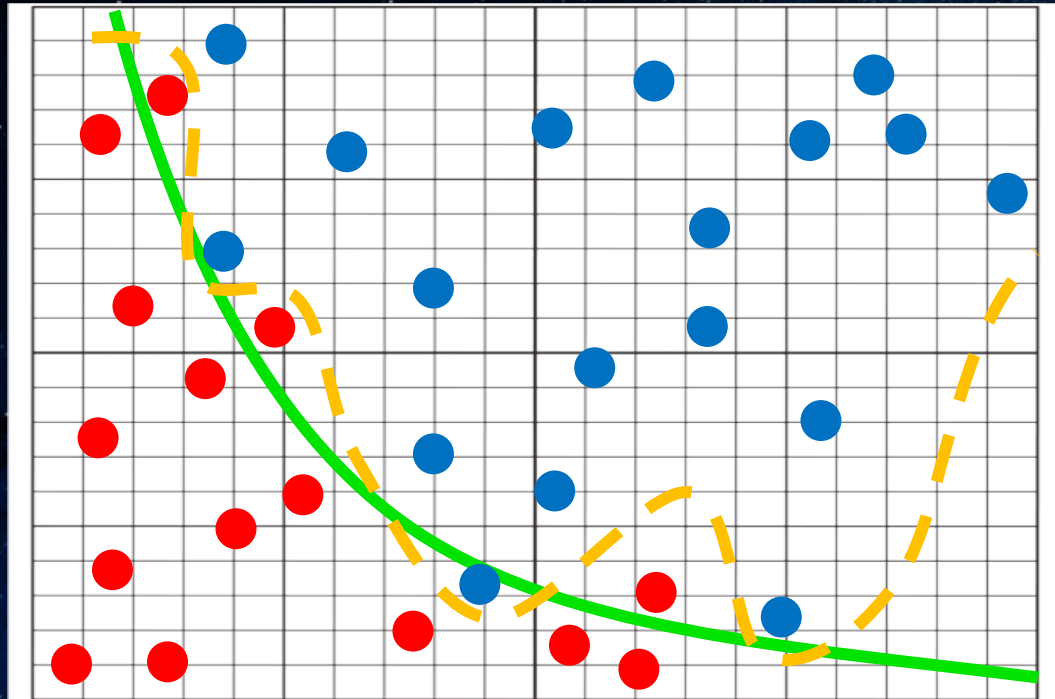
Un modelo lineal ya no nos sirve.





$$y = 1 \quad \text{si} \quad x_1^2 + x_2^2 - 1 > 0$$

Generalización de un modelo

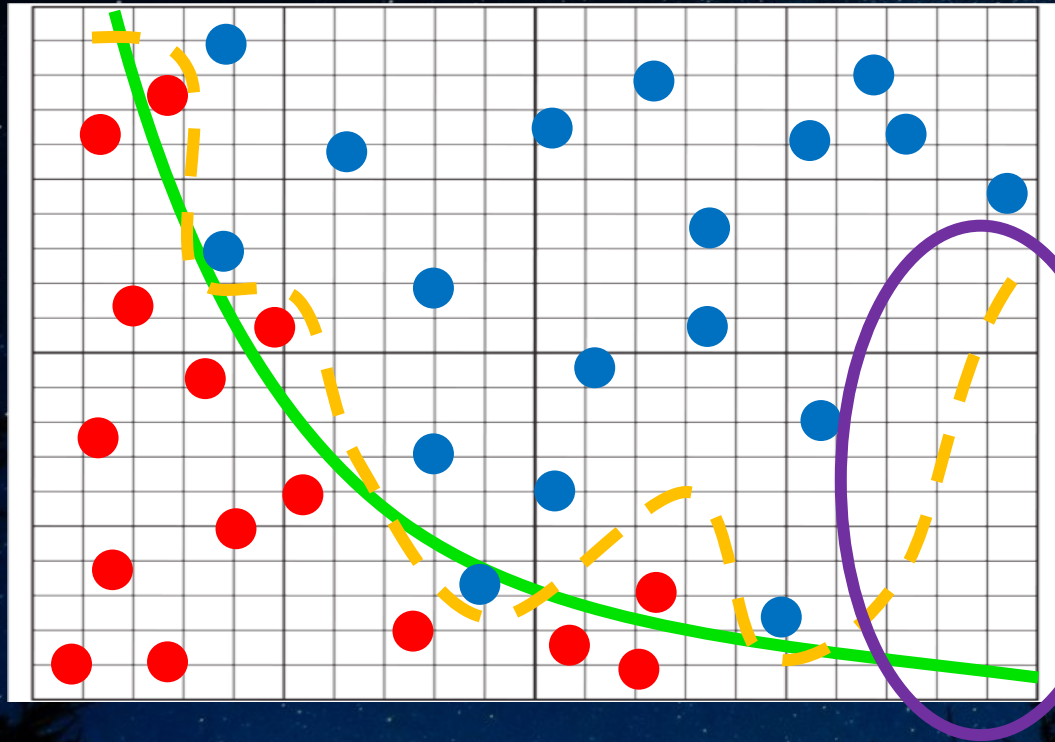
¿Qué curva es mejor?



Generalmente, un modelo más simple lo vamos a considerar mejor, ya que generaliza de forma más efectiva, aunque no sea perfecto.

 Polinomio simple
 Polinomio complejo

Generalización de un modelo



¿Qué ocurre en esta área?

Si no hay datos, esa zona no es representativa para un modelo de Aprendizaje Automático. Pero podría ser una zona del espacio a considerar en futuros datos (desconocidos por el modelo).

Ej: una flor con tamaño de pétalo muy grande.