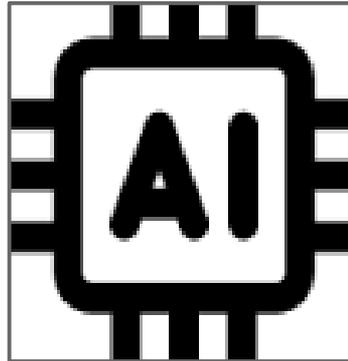


Introducción a la Inteligencia Artificial

Modelos de Regresión



En esta Presentación

1. Introducción a los modelos de Regresión

- Presentación

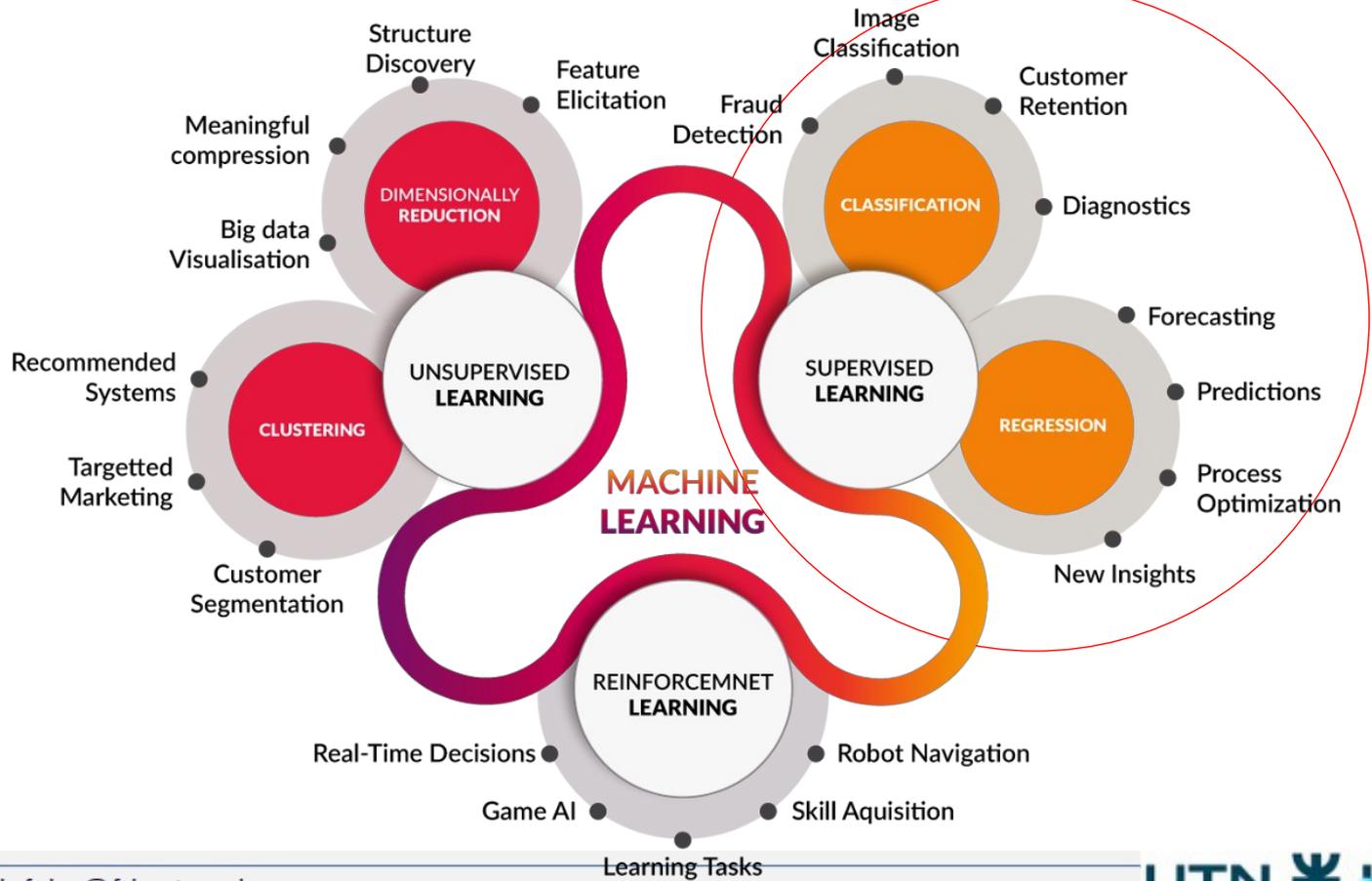
2. El aprendizaje supervisado

- Enfoque y abordaje

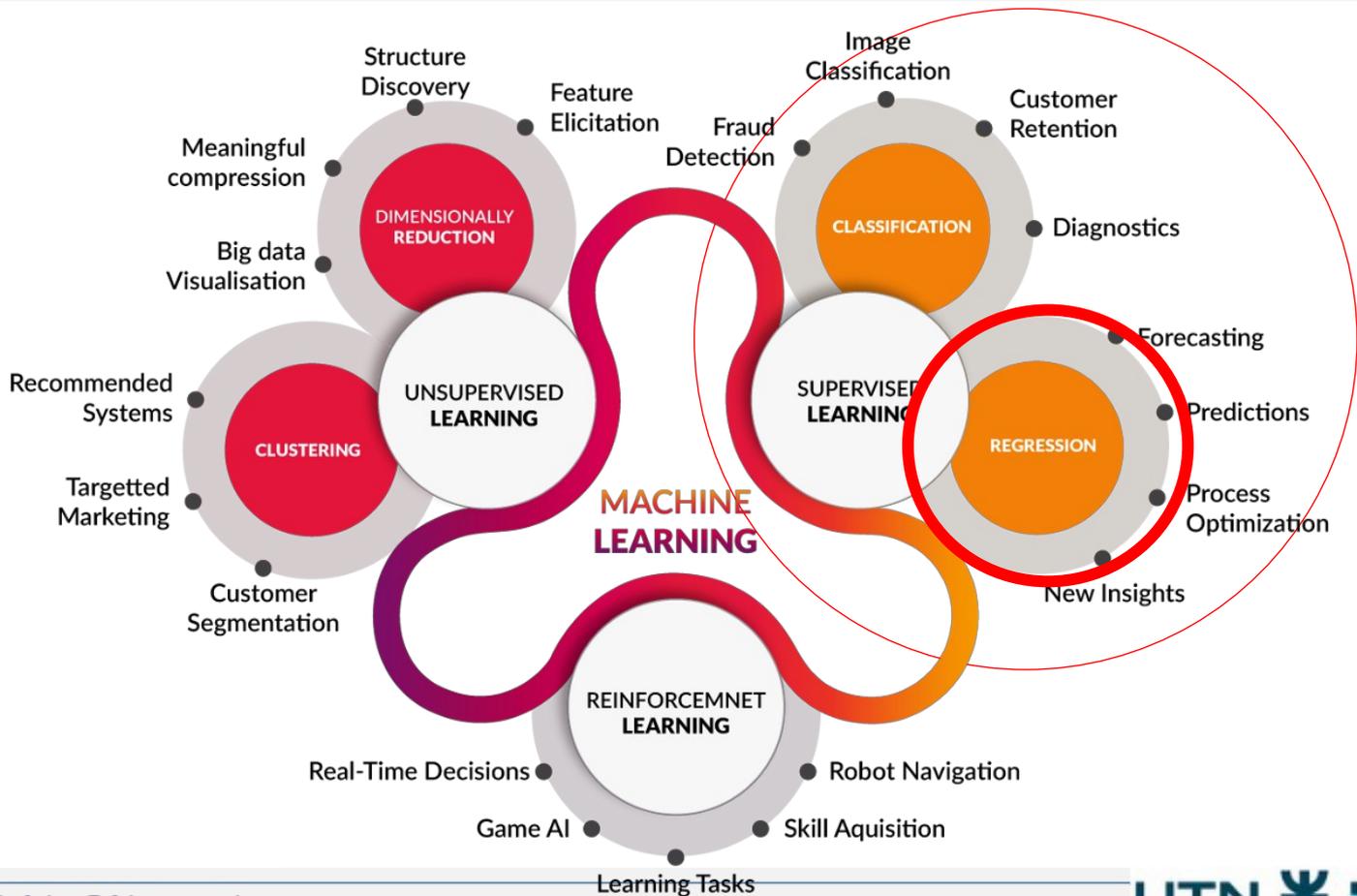
3. Algoritmos de Regresión

- Tipos de regresión
- Casos de uso

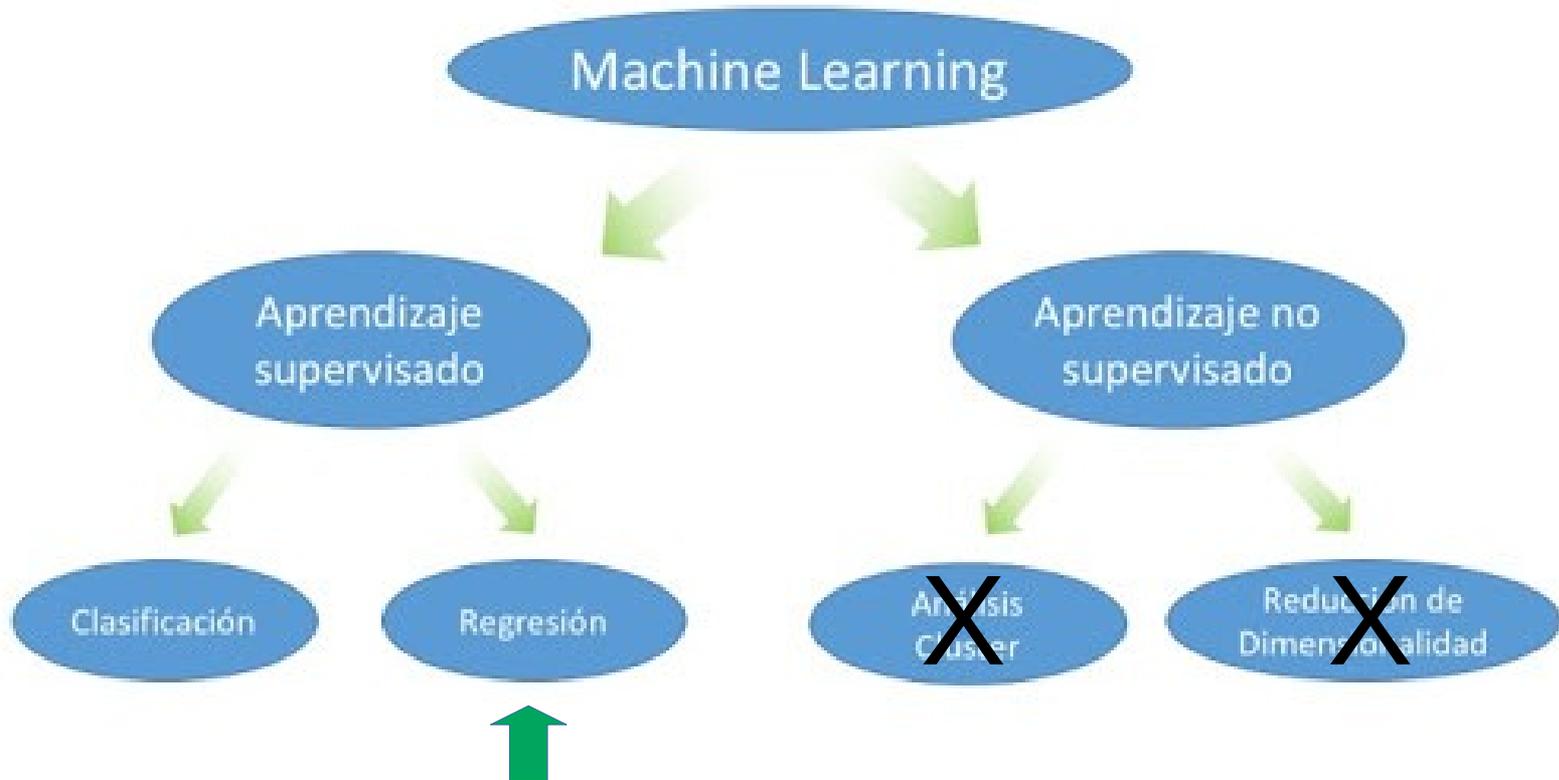
Introducción a los Modelos de Regresión



Introducción a los Modelos de Regresión



Introducción a los Modelos de Regresión



Enfoque y abordaje

En Aprendizaje Supervisado, los datos tienen etiquetas de clase.

Estos modelos se utilizan cuando existe un conjunto de datos de los cuales se conoce la clase a la que pertenecen.

Fuente: Ronchetti, F. Curso de Aprendizaje Automático, UNLP, 2018

Enfoque y abordaje

1. Aprendizaje: Se necesita definir la variable objetivo y un conjunto de datos etiquetados. Se lo separa en entrenamiento y prueba.

- **Entrenamiento:** Se entrena un modelo que aprende cómo llegar a los resultados etiquetados.
- **Prueba:** Con el modelo se hacen predicciones, y se lo evalúa comparándolas con el resultado.

2. Predicción: Se usa al modelo generado para predecir resultados a partir de datos no etiquetados.

Enfoque y abordaje

Aprendizaje

Dataset Etiquetado

x_1	x_2	...	y
1	5	2	Y_1
3	2	6	Y_2
2	6	1	Y_3
8	3	2	Y_4
5	7	4	Y_n

Train

x_1	x_2	...	y
1	5	2	Y_1
3	2	6	Y_2
2	6	1	Y_3
5	7	4	Y_n

Test

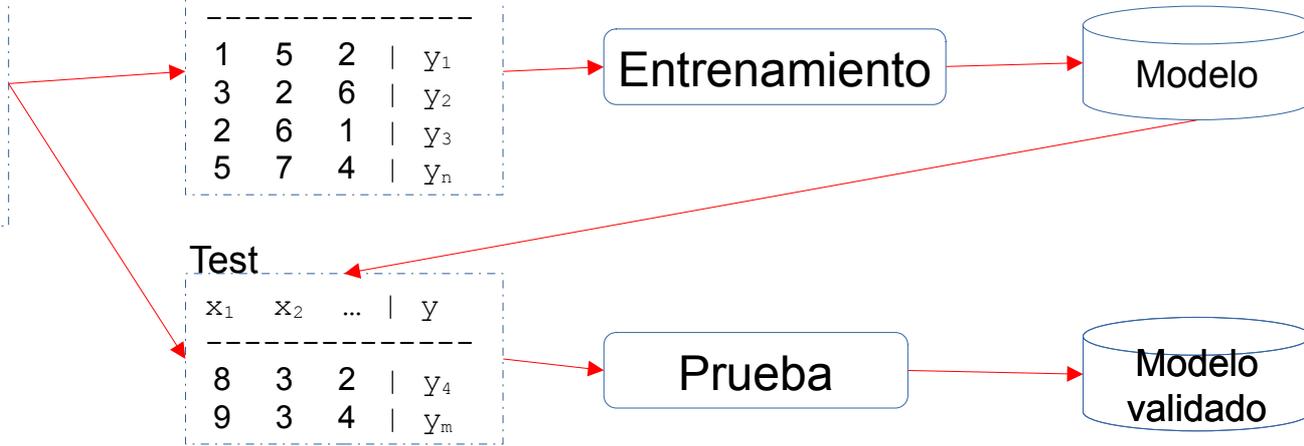
x_1	x_2	...	y
8	3	2	Y_4
9	3	4	Y_m

Entrenamiento

Modelo

Prueba

Modelo validado



Enfoque y abordaje

Predicción

x_1	x_2	...		y
1	3	2		?
4	7	3		?

Dataset No Etiquetado

Predicción

x_1	x_2	...		y
1	3	2		y_1
4	7	3		y_2

Dataset Etiquetado



Enfoque y abordaje

Aprendizaje

Dataset Etiquetado

x_1	x_2	...		y
1	5	2		y_1
3	2	6		y_2
2	6	1		y_3
8	3	2		y_4
5	7	4		y_n

Train

x_1	x_2	...		y
1	5	2		y_1
3	2	6		y_2
2	6	1		y_3
5	7	4		y_n

Entrenamiento

Modelo

Test

x_1	x_2	...		y
8	3	2		y_4
9	3	4		y_m

Prueba

Modelo validado

Predicción

x_1	x_2	...		y
1	3	2		?
4	7	3		?

Predicción

x_1	x_2	...		y
1	3	2		y_1
4	7	3		y_2

Regresión: Presentación

Por su sencillez, para realizar predicciones, los modelos de Regresión son los que más rápidamente nos pueden entregar un resultado con alta calidad y muy buena aproximación.

La mayoría de los problemas pueden tratarse con modelos de Regresión.

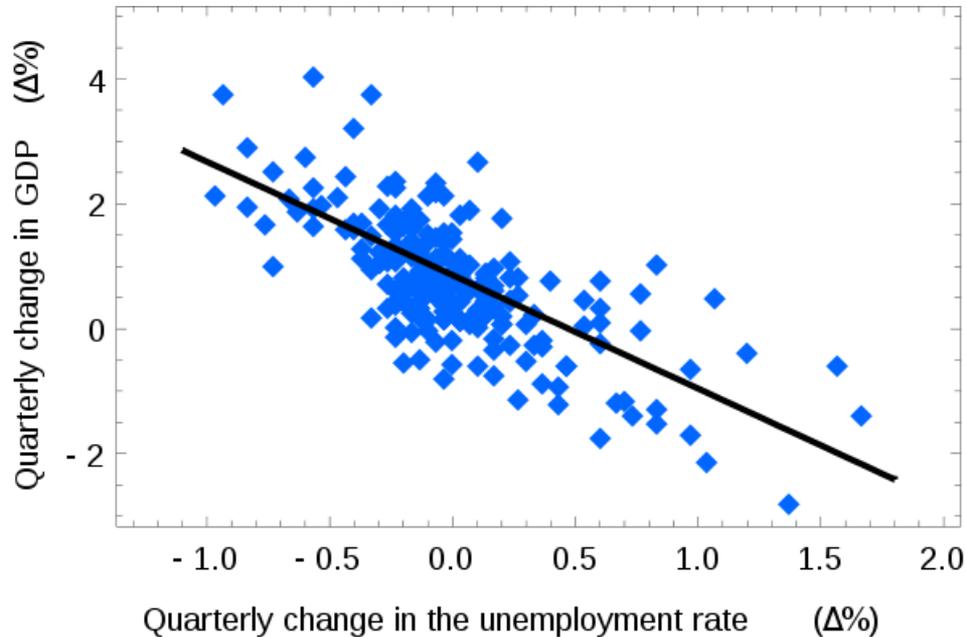
El modelo entrega un indicador: el R^2 o MSE que indica en promedio, cuán cerca o lejos de los datos reales está el modelo generado.

Coeficiente de Determinación R^2

R^2 es un estadístico que determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo.

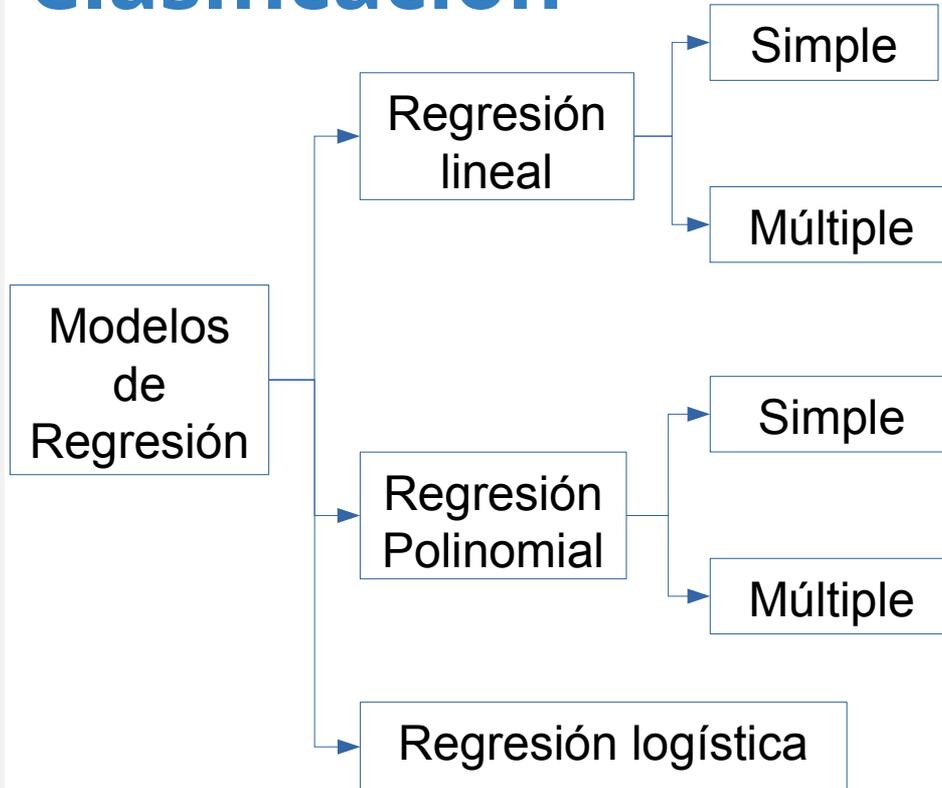
Mientras los puntos no disten mucho de la línea de la regresión, el coeficiente de determinación adoptará valores altos.

Coeficiente de Determinación R^2

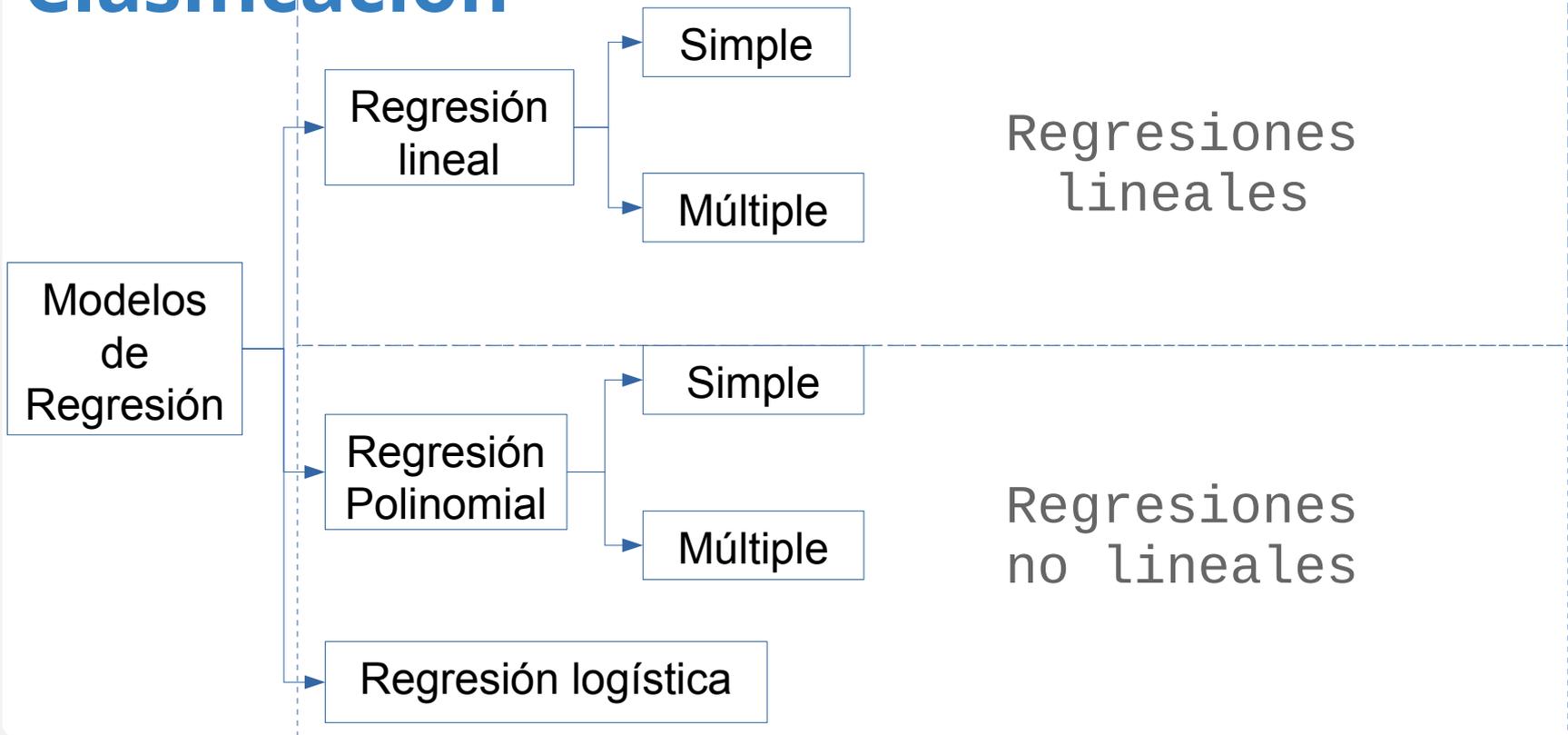


Fuente: https://es.wikipedia.org/wiki/Coeficiente_de_determinaci%C3%B3n

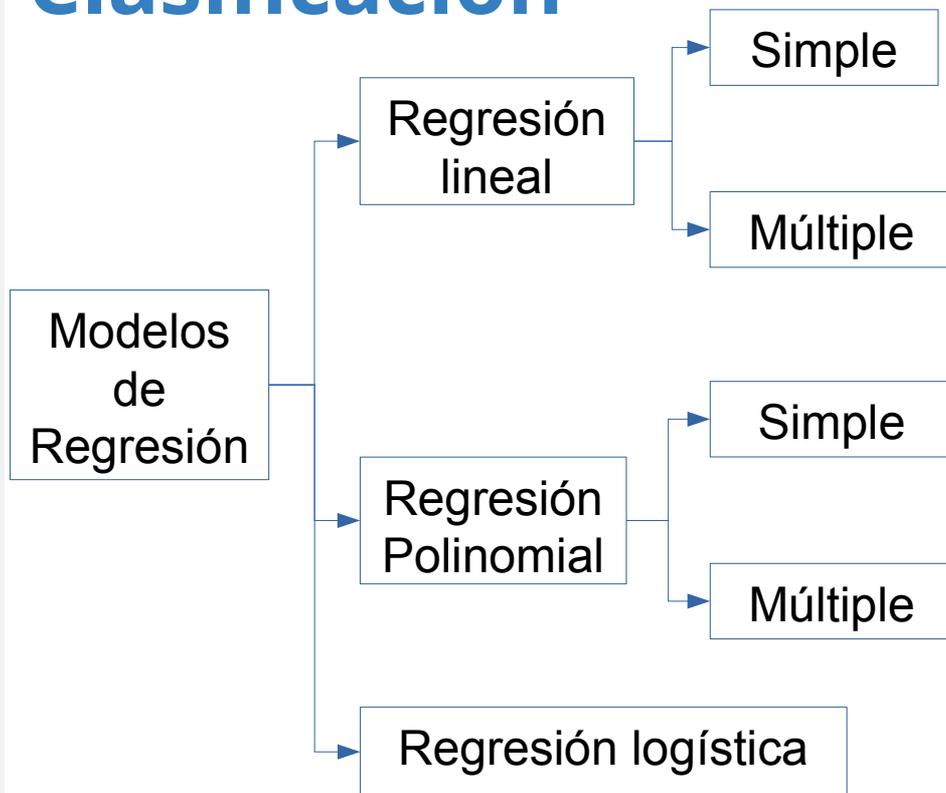
Clasificación



Clasificación



Clasificación



Una variable predictora. Relación lineal. Salida numérica. Predicciones rápidas.

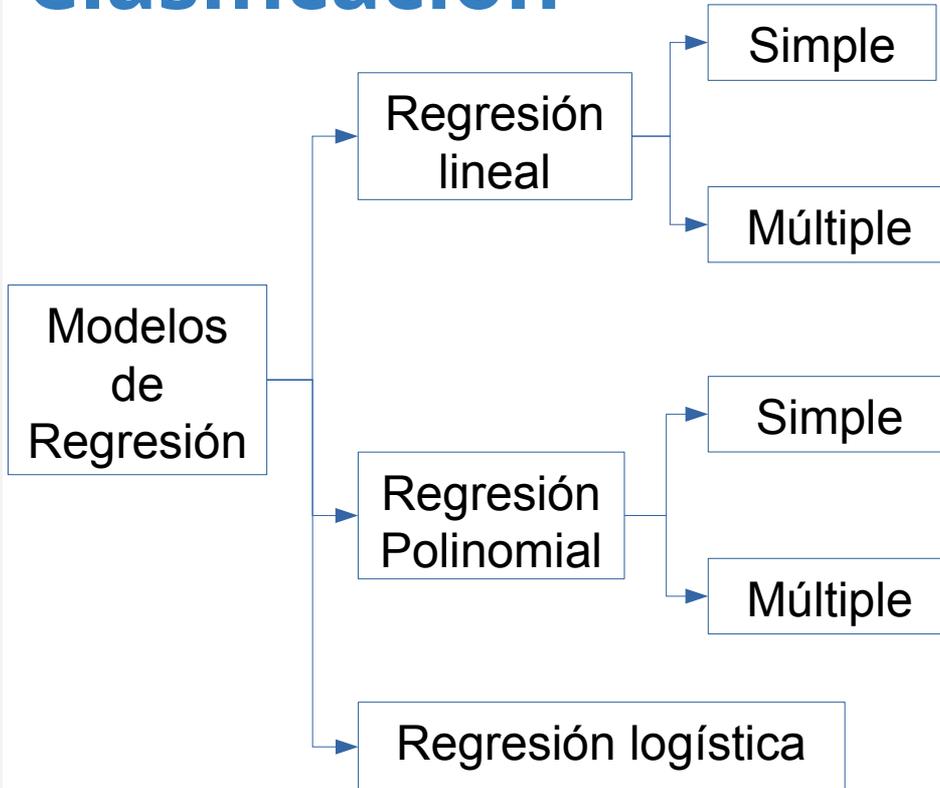
Varias variables predictoras. Relación Lineal. Salida numérica. Se usa en Predicciones de valores.

Una variable predictora. Relación polinómica. Salida numérica. Predicción de cotizaciones.

Varias variables predictoras. Relación polinómica. Salida numérica. Predicción de cotizaciones.

Varias variables predictoras. Salida binaria. Predicción de clases.

Clasificación



Predicciones de valores

Predicciones de clase

Regresión Lineal Simple

Forma de la expresión:

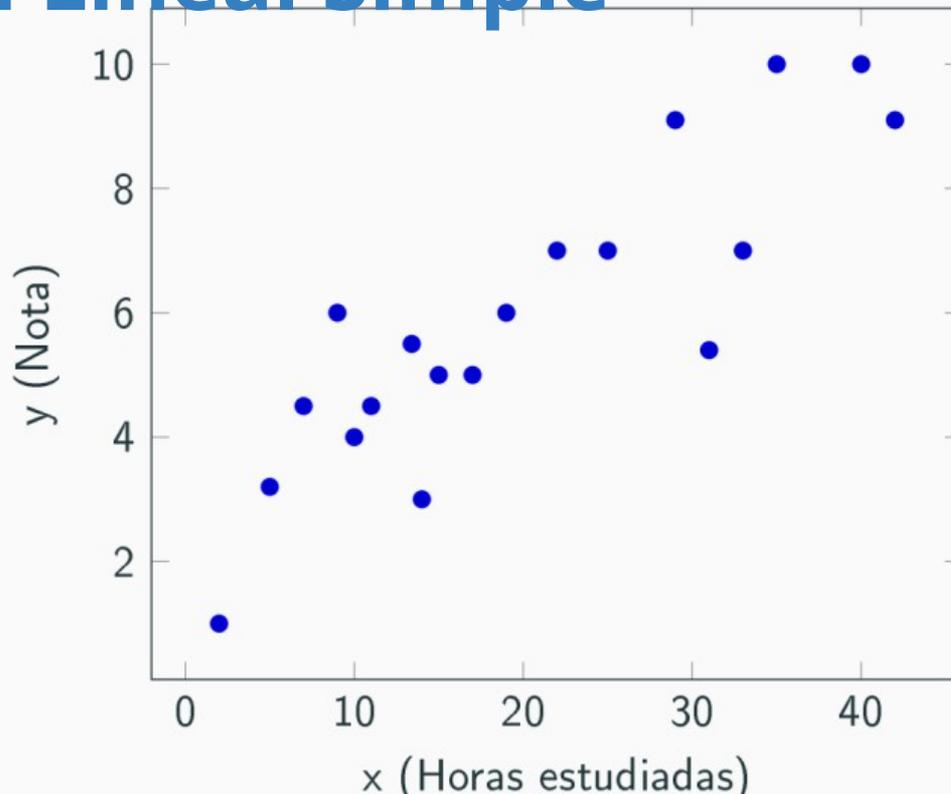
$$y = m \cdot x + b$$

Ejemplo:

Se dispone de una tabla en la que se tienen datos de las horas que un alumno le dedicó al estudio de una materia y la nota obtenida.

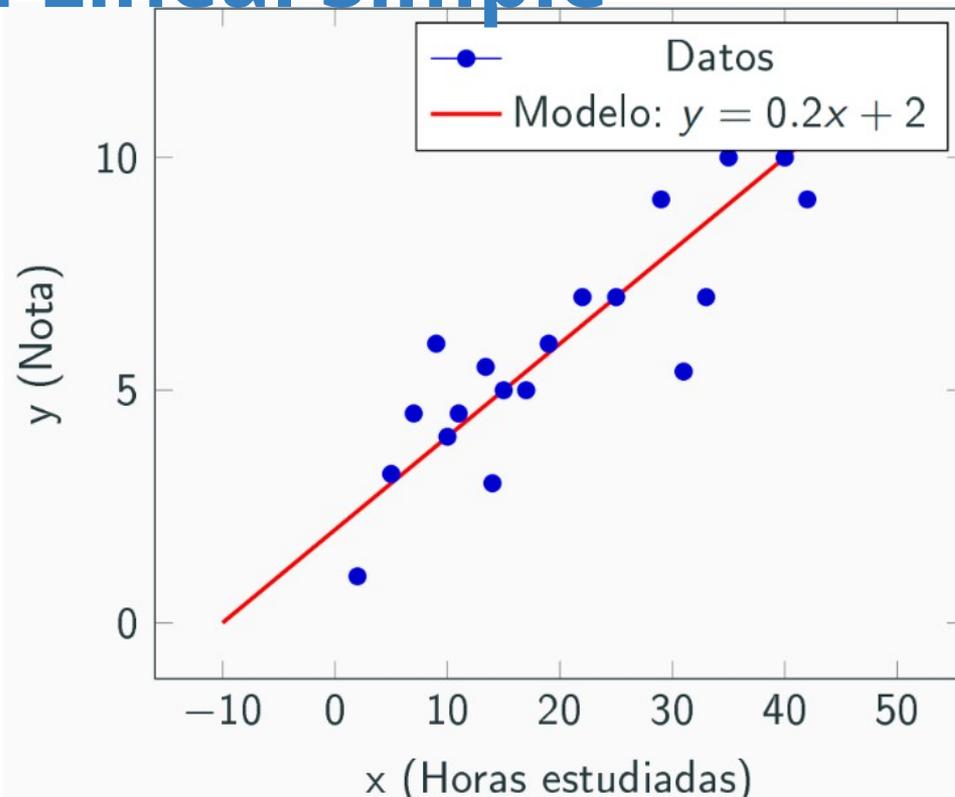
Se pretende predecir la nota que un alumno sacará en el próximo examen sabiendo las horas dedicadas al estudio.

Regresión Lineal Simple



estudio	nota
2	1
5	3.2
7	4.5
9	6
10	4
11	4.5
13.4	5.5
14	3
15	5
.....	
<i>n</i> datos o ejemplos, 2 columnas	

Regresión Lineal Simple



Regresión Lineal Simple

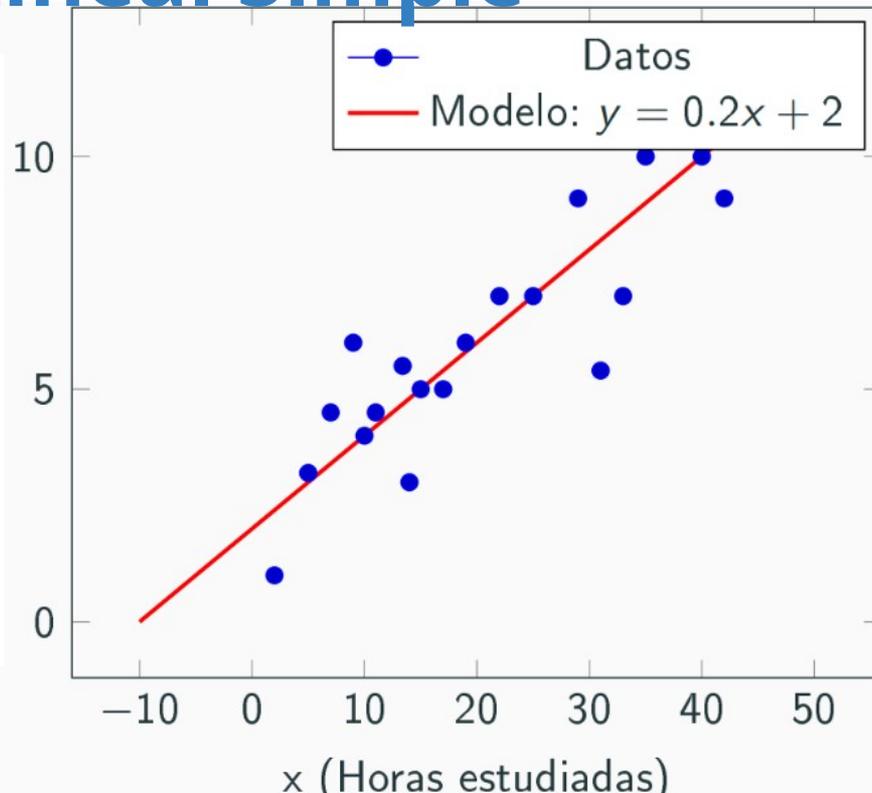
Pregunta:

¿Qué nota obtendrá un alumno que estudió 25hs?

Respuesta:

$$y(25) = 0,2 \cdot 25 + 2$$

$$y(25) = 7$$



Regresión Lineal Múltiple

Forma de la expresión:

$$y = C_1 \cdot X_1 + C_2 \cdot X_2 + \dots + C_n \cdot X_n$$

Ejemplo:

Disponemos de una tabla en la que se tienen datos de pesos de mujeres (pesos_mujeres.csv) y de su altura y edad.

Se pretende predecir el peso a partir de la altura y edad.

Regresión Lineal Múltiple

Ejemplo:

El conjunto de datos:

	altura_cm	edad_años	peso_kg
0	115	6	25
1	118	7	24
2	125	8	30
3	140	10	35
4	140	70	55

Regresión Lineal Múltiple

Ejemplo:

Entrenamiento: Evaluación

$$R^2 = 0.7076$$

Fórmula:

$$\text{Peso} = 0.342 \text{ edad} + 0.615 \text{ altura} - 49.53$$

Regresión Lineal Múltiple

Ejemplo:

Predicción: Ejemplo

* Ingrese altura_cm: 152

* Ingrese edad_años: 41

==> peso_kg_pred= 58,01839328156421

Regresión Polinomial Simple

Principales características:

- Se usan para poder predecir valores a partir de una variable predictora.
- La relación entre las variables no es lineal
- La salida es numérica.

Regresión Polinomial Simple

Fundamento:

Las funciones lineales son rápidas y sencillas, pero no aproximan muy bien para la mayoría de los datasets.

Funcionamiento:

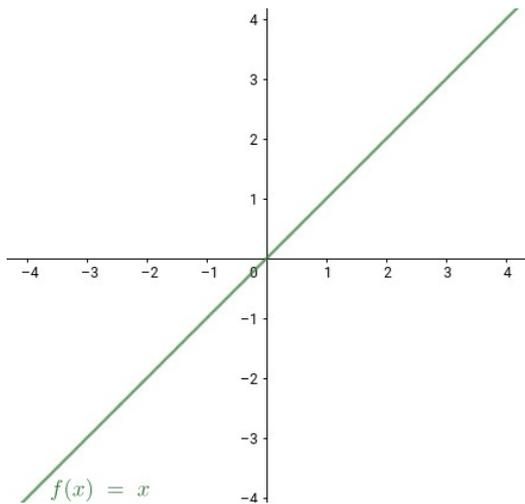
Se tiene una variable explicativa o predictora

Se trata de hallar una función polinómica de diferente grado $P(x)$ que aproxime mejor al conjunto de datos.

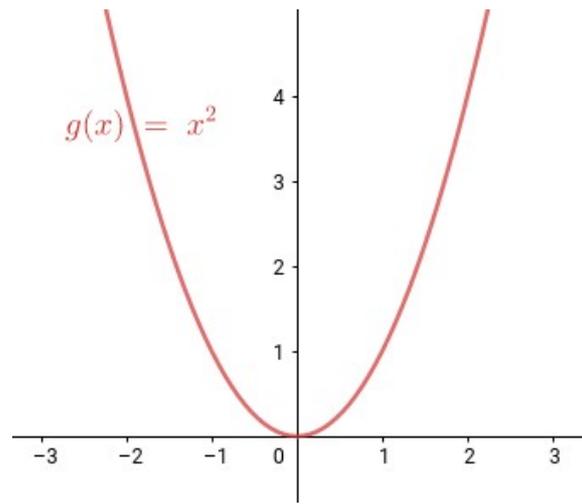
Regresión Polinomial Simple

Forma de la función polinomial simple:

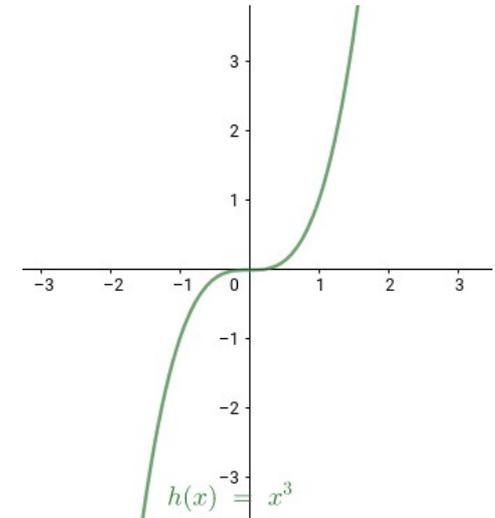
$$P(x) = c_0 + c_1 \cdot x + c_2 \cdot x^2 + \dots + c_n \cdot x^n$$



Grado 1



Grado 2



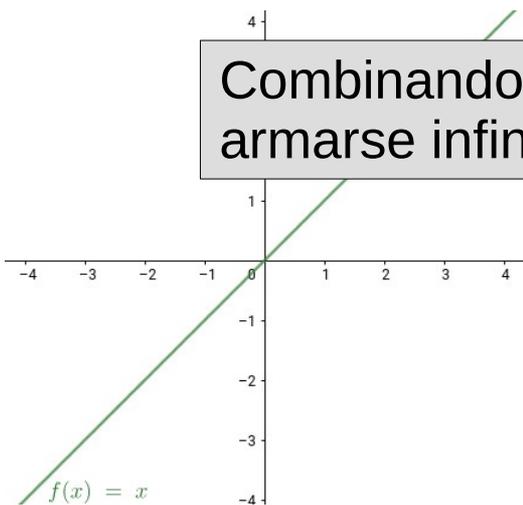
Grado 3

Regresión Polinomial Simple

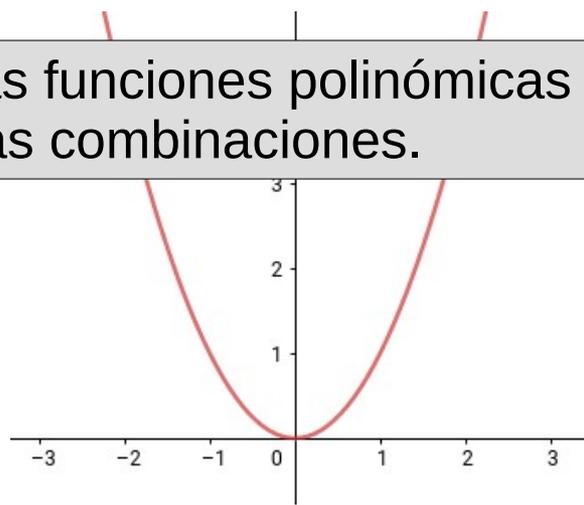
Forma de la función polinomial simple:

$$P(x) = c_0 + c_1 \cdot x + c_2 \cdot x^2 + \dots + c_n \cdot x^n$$

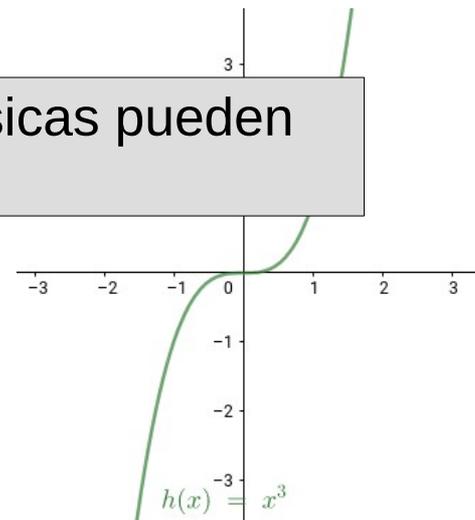
Combinando las funciones polinómicas básicas pueden armarse infinitas combinaciones.



Grado 1



Grado 2



Grado 3

Regresión Polinomial Simple

Ejemplo:

Salario según nivel

¿Con qué función puede
Modelarse mejor a este conjunto
de datos?

¿Puede estimarse el salario de
un nivel intermedio ?

	nivel	salario
0	1	45000
1	2	50000
2	3	60000
3	4	80000
4	5	110000
5	6	150000
6	7	200000
7	8	300000
8	9	500000
9	10	1000000

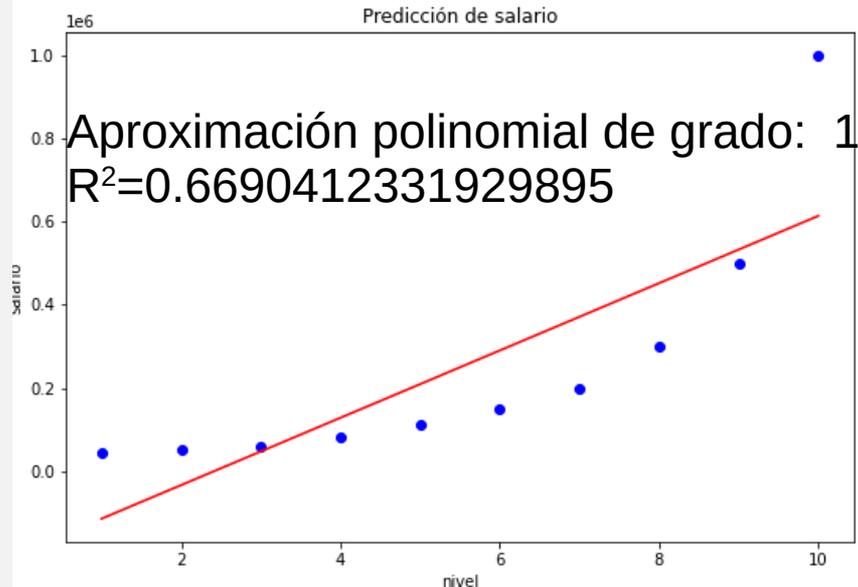
Dataset Salarios.csv

Regresión Polinomial Simple

¿Con qué función puede modelarse mejor a este conjunto de datos?

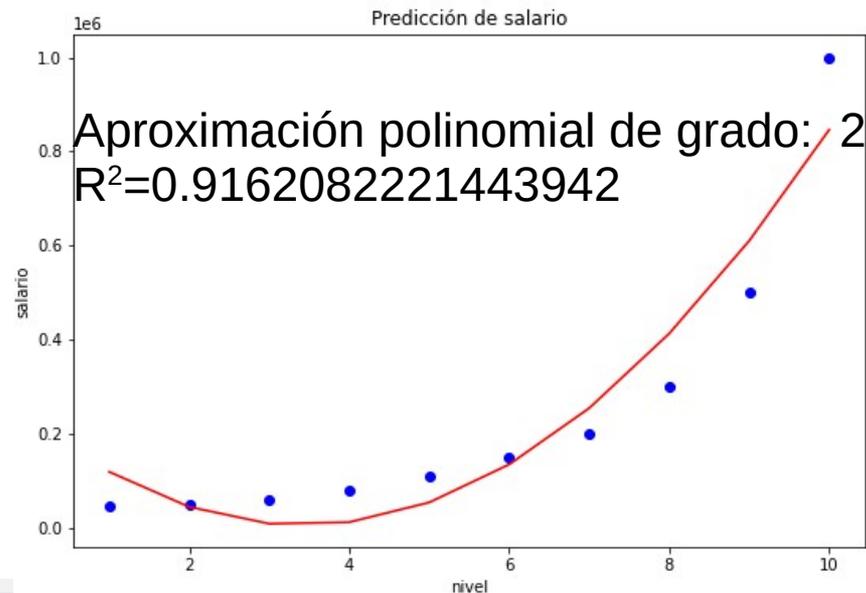
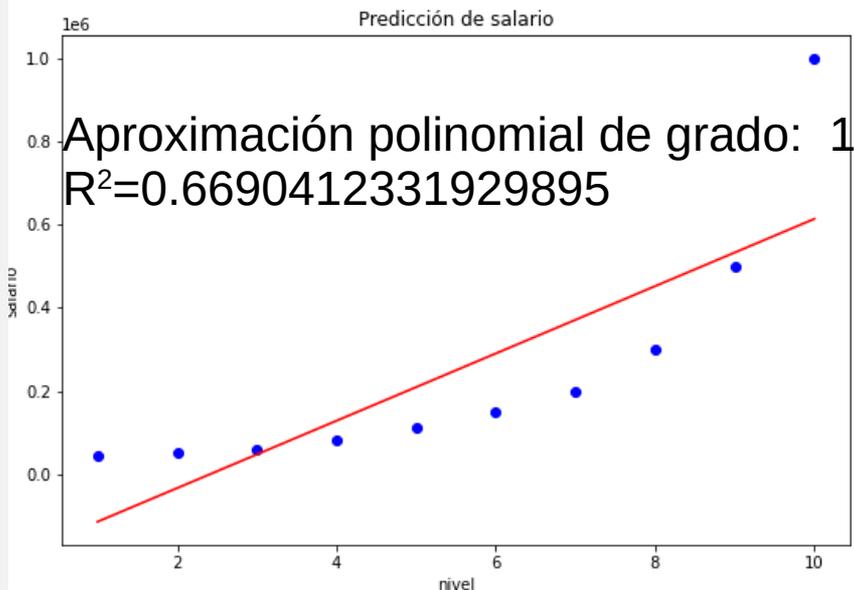
Regresión Polinomial Simple

¿Con qué función puede modelarse mejor a este conjunto de datos?



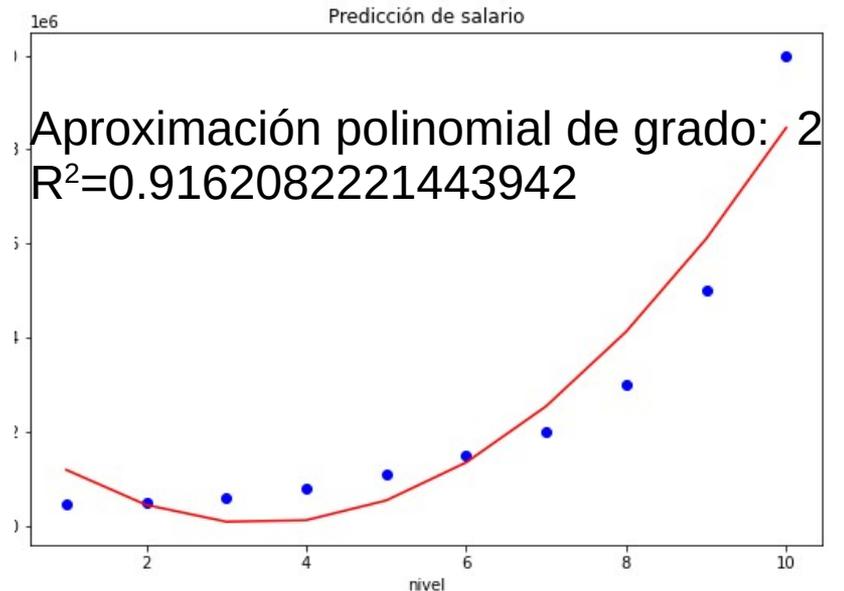
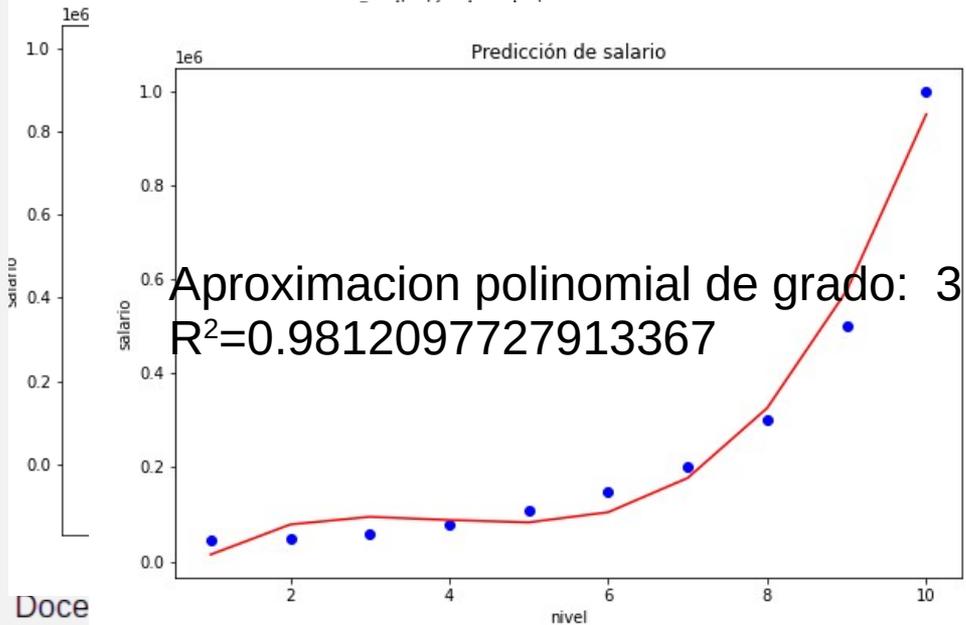
Regresión Polinomial Simple

¿Con qué función puede modelarse mejor a este conjunto de datos?



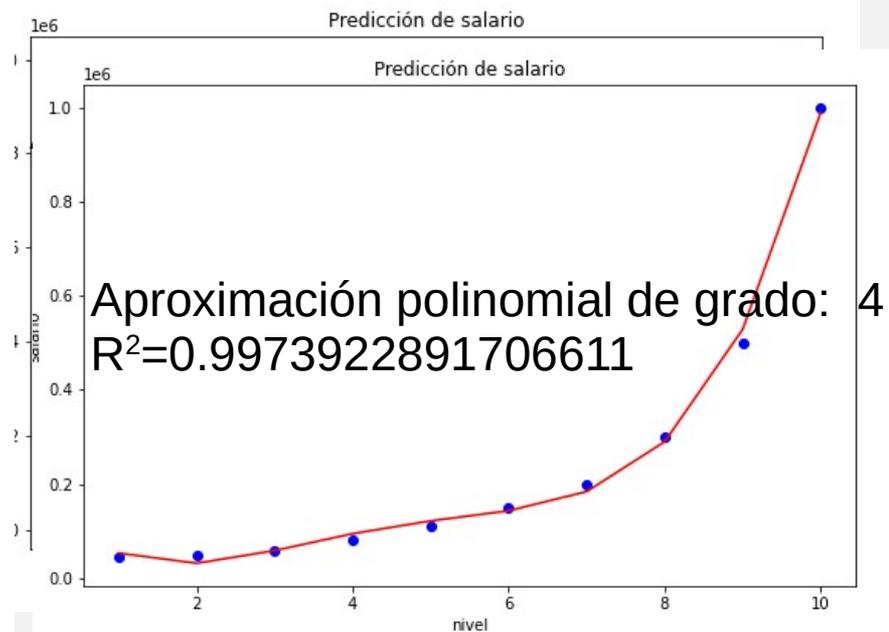
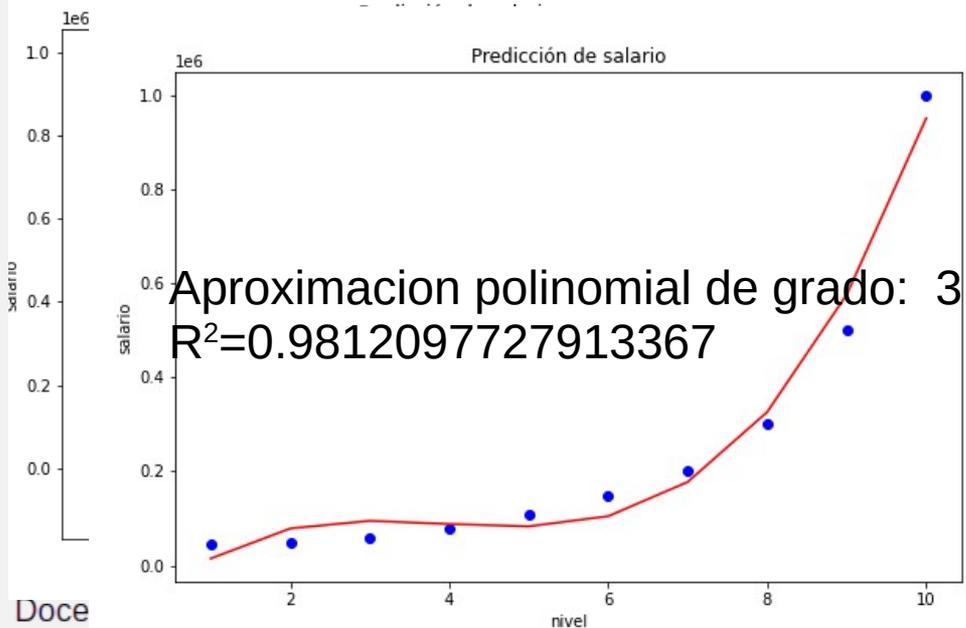
Regresión Polinomial Simple

¿Con qué función puede modelarse mejor a este conjunto de datos?



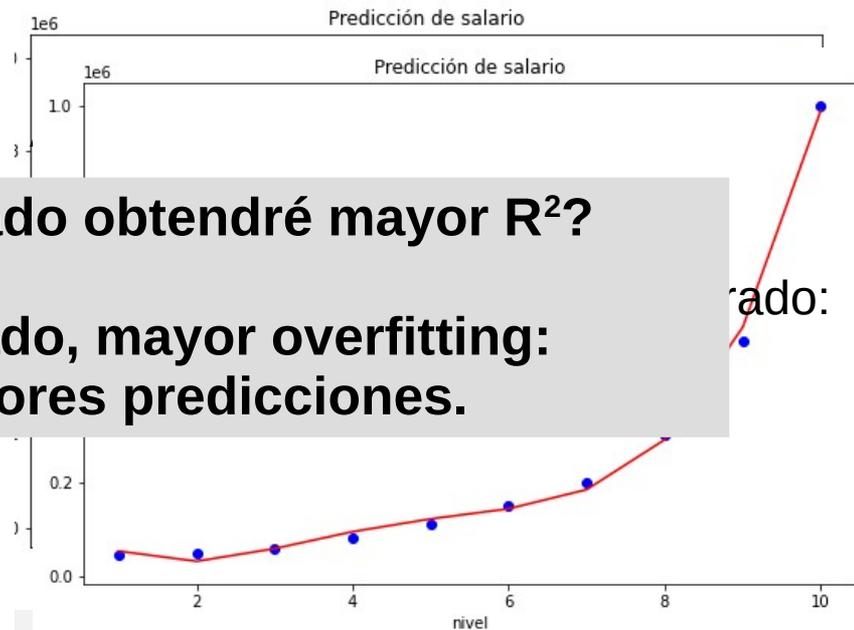
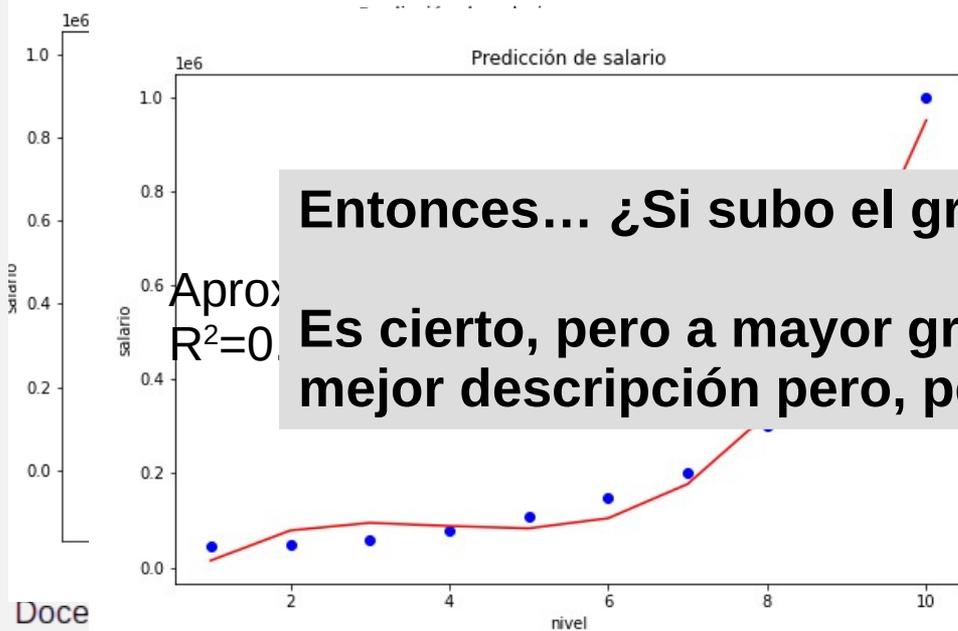
Regresión Polinomial Simple

¿Con qué función puede modelarse mejor a este conjunto de datos?



Regresión Polinomial Simple

¿Con qué función puede modelarse mejor a este conjunto de datos?



Entonces... ¿Si subo el grado obtendré mayor R^2 ?

Es cierto, pero a mayor grado, mayor overfitting: mejor descripción pero, peores predicciones.

Aprox
 $R^2=0.9$

Regresión Polinomial Múltiple

Funcionamiento:

Se tiene “n” variables explicativa o predictora ($n > 1$)

Se trata de hallar una función polinómica de diverso grado $F(x_1, x_2, \dots, x_n)$ que aproxime mejor al conjunto de datos.

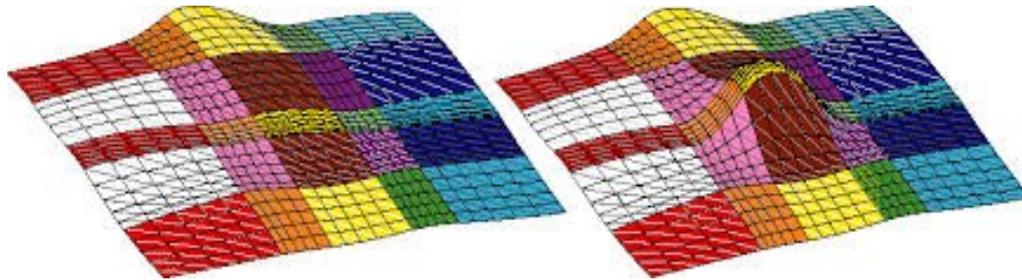
Uso:

Predicción de valores, a partir de varias variables numéricas.

Regresión Polinomial Múltiple

Forma de una expresión polinomial múltiple:

$$F(x, y) = a_0 + b_0 \cdot x_0 + b_1 \cdot x_1 + \dots + b_n \cdot x_n + c_0 \cdot y_0 + c_1 \cdot y_1 + \dots + c_n \cdot y_n$$



Fuente (imagen): <https://dcain.etsin.upm.es/~leonardo/tema5.htm>

Regresión Logística

Supongamos que tenemos información sobre el tamaño de varios tumores y si finalmente fueron benignos (0), o malignos (1).

Queremos crear un modelo que sea capaz de discriminar entre ambas clases: clasificación binaria.

Regresión Logística

Supongamos que tenemos información sobre el tamaño de varios tumores y si finalmente fueron benignos (0), o malignos (1).

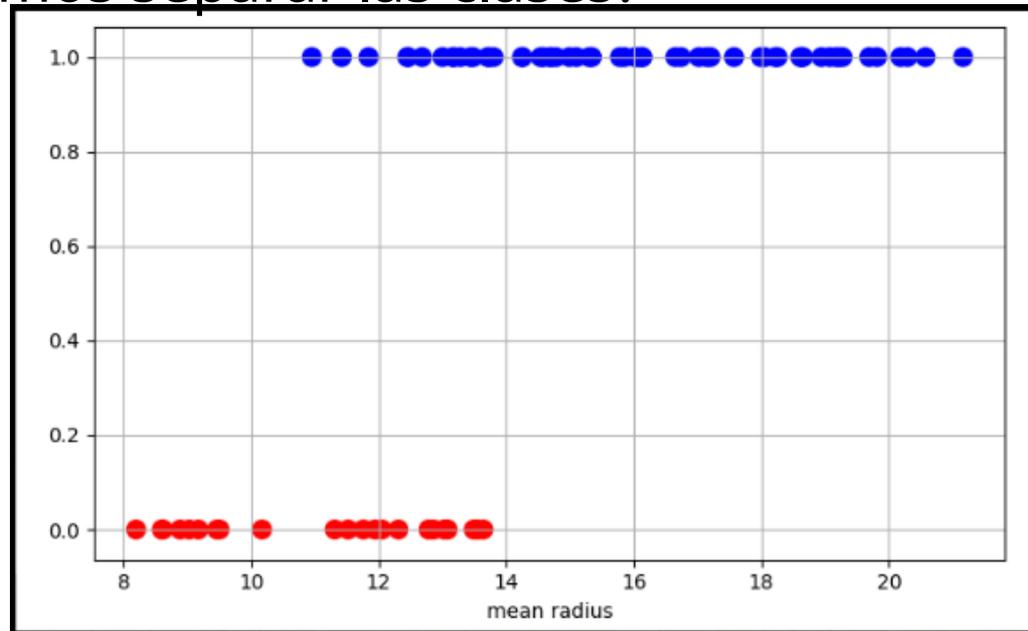
Queremos crear un modelo que sea capaz de discriminar entre ambas clases: clasificación binaria.

Es decir, dada cierta información del caso, se desea obtener un modelo que prediga si el tumor será benigno o maligno.

Fuente: Ronchetti, Franco. Curso de Aprendizaje Automático, UNLP, 2018

Regresión Logística

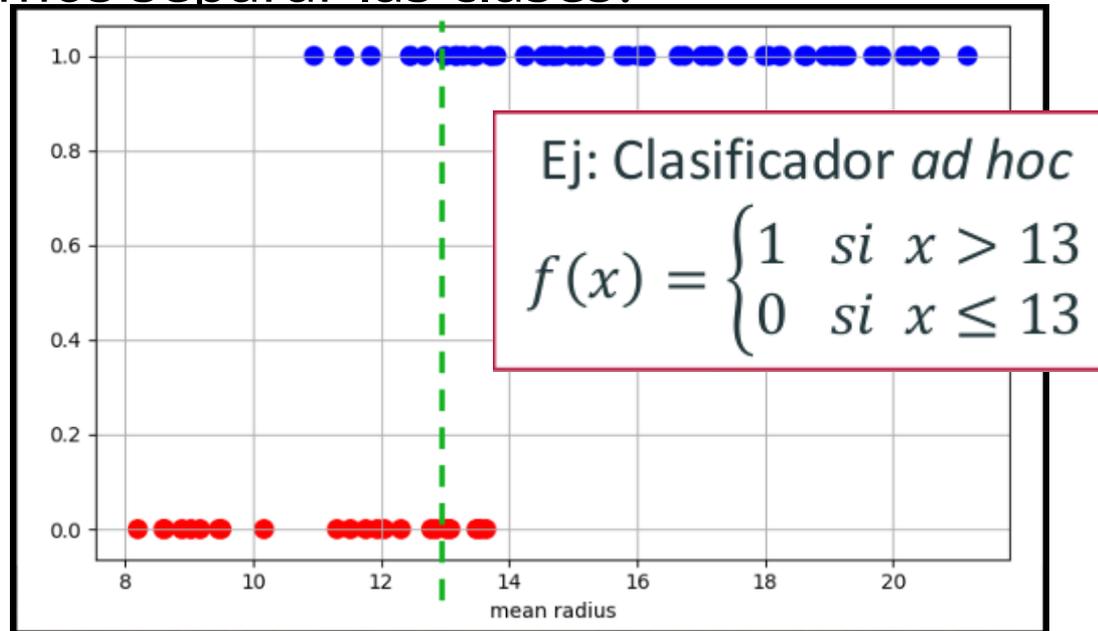
¿Cómo podemos separar las clases?



Fuente: Ronchetti, Franco. Curso de Aprendizaje Automático, UNLP, 2018

Regresión Logística

¿Cómo podemos separar las clases?



Fuente: Ronchetti, Franco. Curso de Aprendizaje Automático, UNLP, 2018

Regresión Logística

Un modo de solucionarlo es utilizando una función Sigmoidal, que tiene un rango de salida entre 0 y 1. Además, la función de Error es diferenciable.

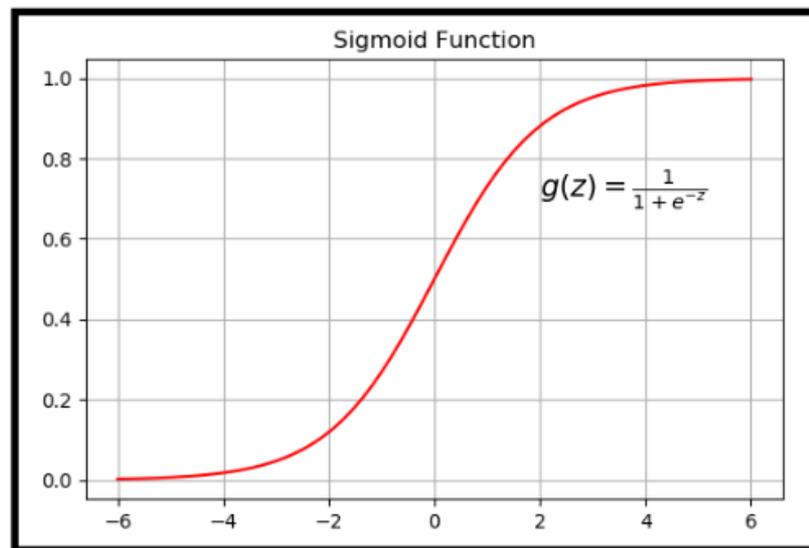
$$g(z) = \frac{1}{1+e^{-z}}$$

$$g(f(x)) = \frac{1}{1+e^{-wx-b}}$$

$g(z)$ tiene dos asíntotas horizontales:

- Tiende a 1 cuando x tiende a $+\infty$
- Tiende a 0 cuando x tiende a $-\infty$

Fuente: Ronchetti, Franco. Curso de Aprendizaje Automático, UNLP, 2018



Regresión Lineal y Polinomial

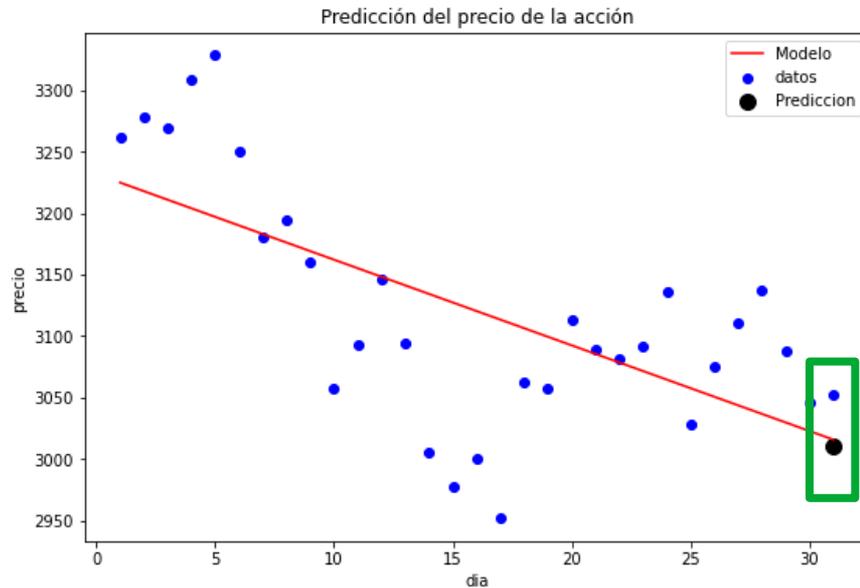
- Se usan para poder predecir valores numéricos partir de variables predictoras numéricas.
- Es posible obtener una expresión (fórmula) en forma de polinomio que permite describir el fenómeno.
- Indicador: R^2

Regresión Lineal y Polinomial

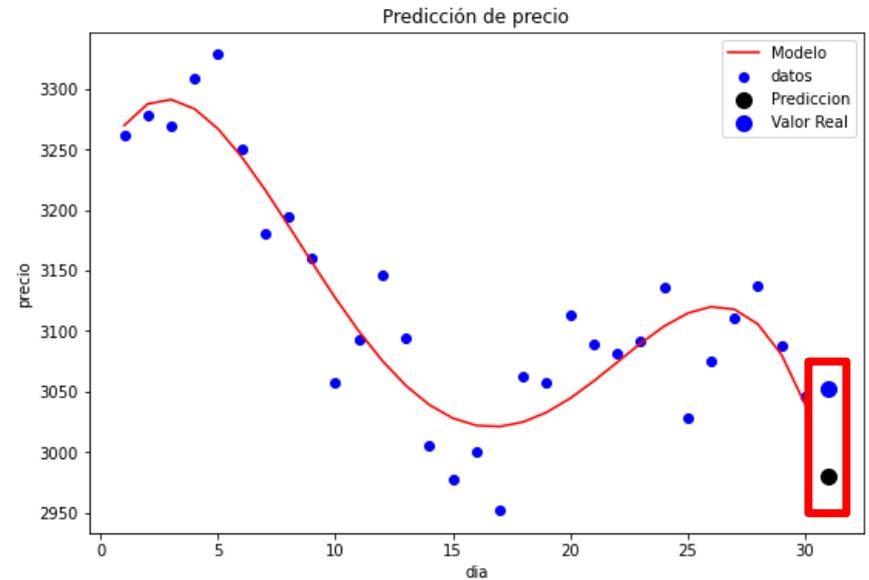
- La representación geométrica de las regresiones lineales corresponde a rectas, planos o hiper-planos: geometría lineal. Se obtienen valores de R^2 no muy altos, pero son mejores predicciones para tendencias.
- La representación de las regresiones polinomiales corresponde a líneas o superficies curvas: geometría no lineal. Se obtienen valores de R^2 mucho más altos, pero con peores predicciones en tendencias gracias al sobre-ajuste u overfitting. Sin embargo las predicciones para valores intermedios son excelentes!

Regresión Lineal y Polinomial

Predicción de tendencias



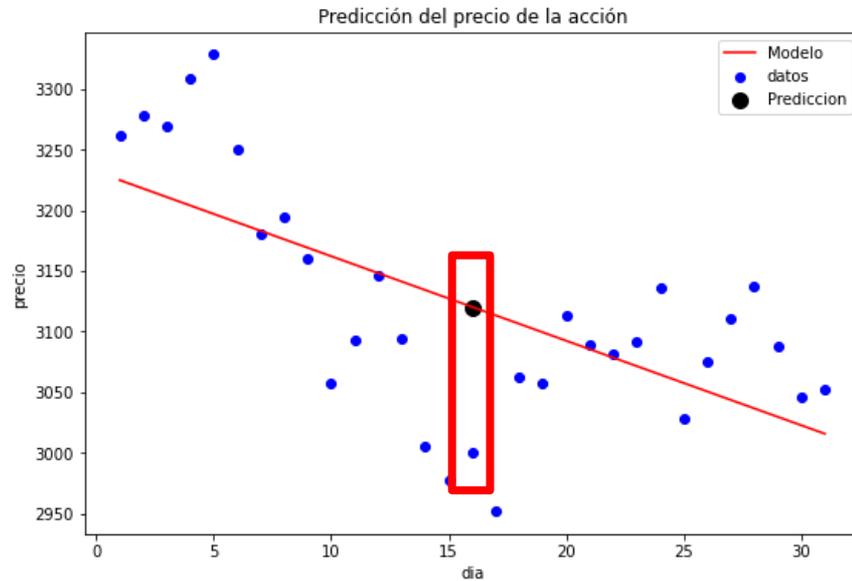
Predicción con modelo Lineal ($R^2=0,41$)



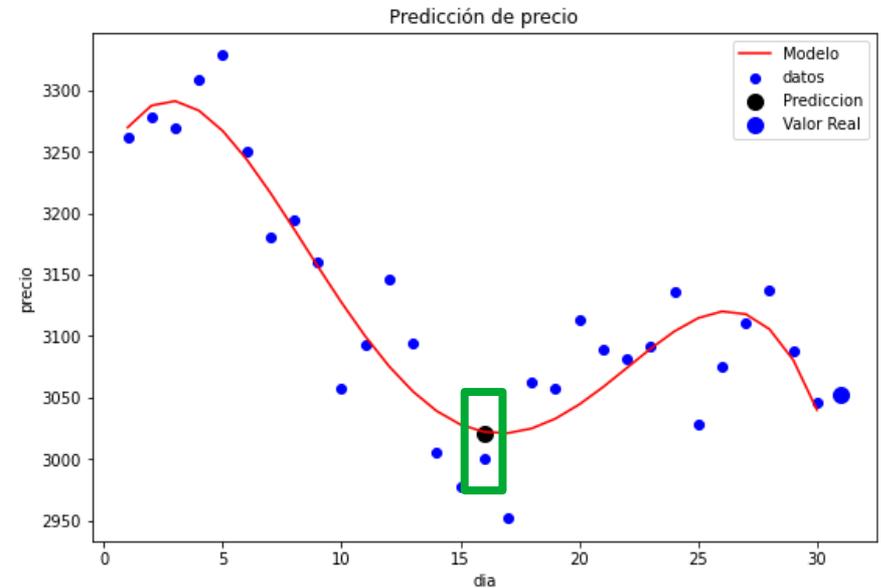
Predicción con modelo polinomial ($R^2=0,83$)

Regresión Lineal y Polinomial

Predicción de valores intermedios



Predicción con modelo Lineal ($R^2=0,41$)



Predicción con modelo polinomial ($R^2=0,83$)

Regresión Logística

- La regresión logística puede usarse para tratar de correlacionar la probabilidad (valor entre 0 y 1) de una variable cualitativa binaria o clase (que puede tomar los valores reales "0" y "1") con variables escalares x_i .
- Para evaluar estos modelos se utiliza la "Matriz de confusión".

https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica

Regresión Logística

Matriz de Confusion

```
[[ 9 18]
 [ 3 42]]
```

Reporte de las Clasificaciones

	precision	recall	f1-score	support
0	0.75	0.33	0.46	27
1	0.70	0.93	0.80	45
accuracy			0.71	72
macro avg	0.72	0.63	0.63	72
weighted avg	0.72	0.71	0.67	72

Conclusiones:

Se presentaron varios algoritmos de regresión distintos: algunos para una variable predictora (simples) y otros para varias variables predictoras (múltiples).

Anticipandonos a los Clasificadores Binarios, se presentó al regresor logístico.

Se presentaron las bondades de unos y otros.

Se presentaron casos de uso.