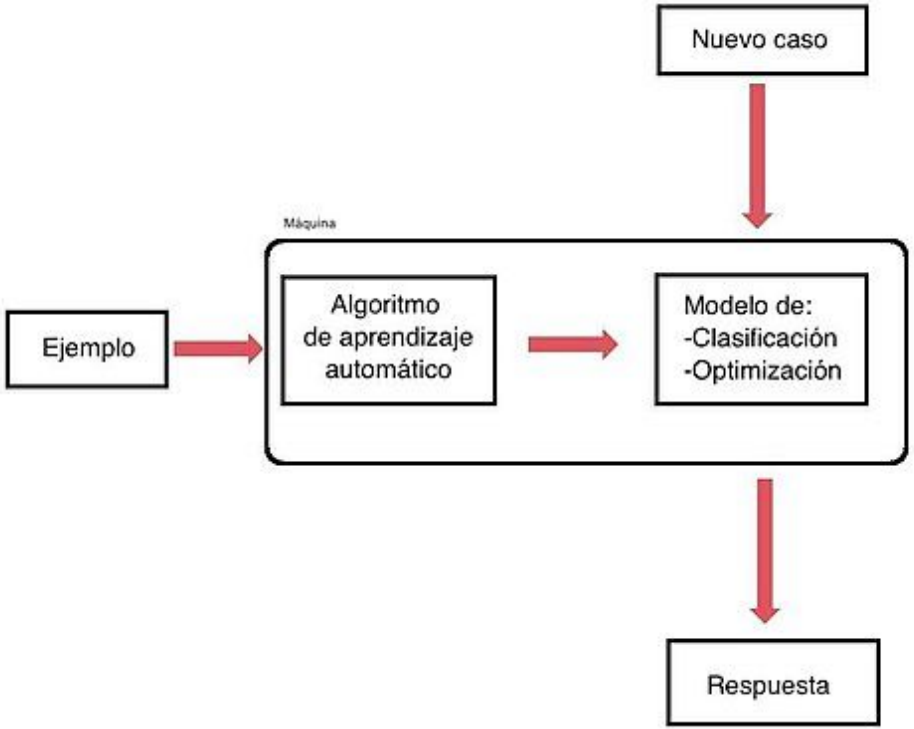
Resolución de Problemas de Regresión Mediante Machine Learning y Caret

Eduardo Illueca
Fernández
Marzo 2022

f SéNeCa (+)
Agencia de Ciencia y Tecnología
Región de Murcia
UNIVERSIDAD DE MURCIA
hop ubiquitous

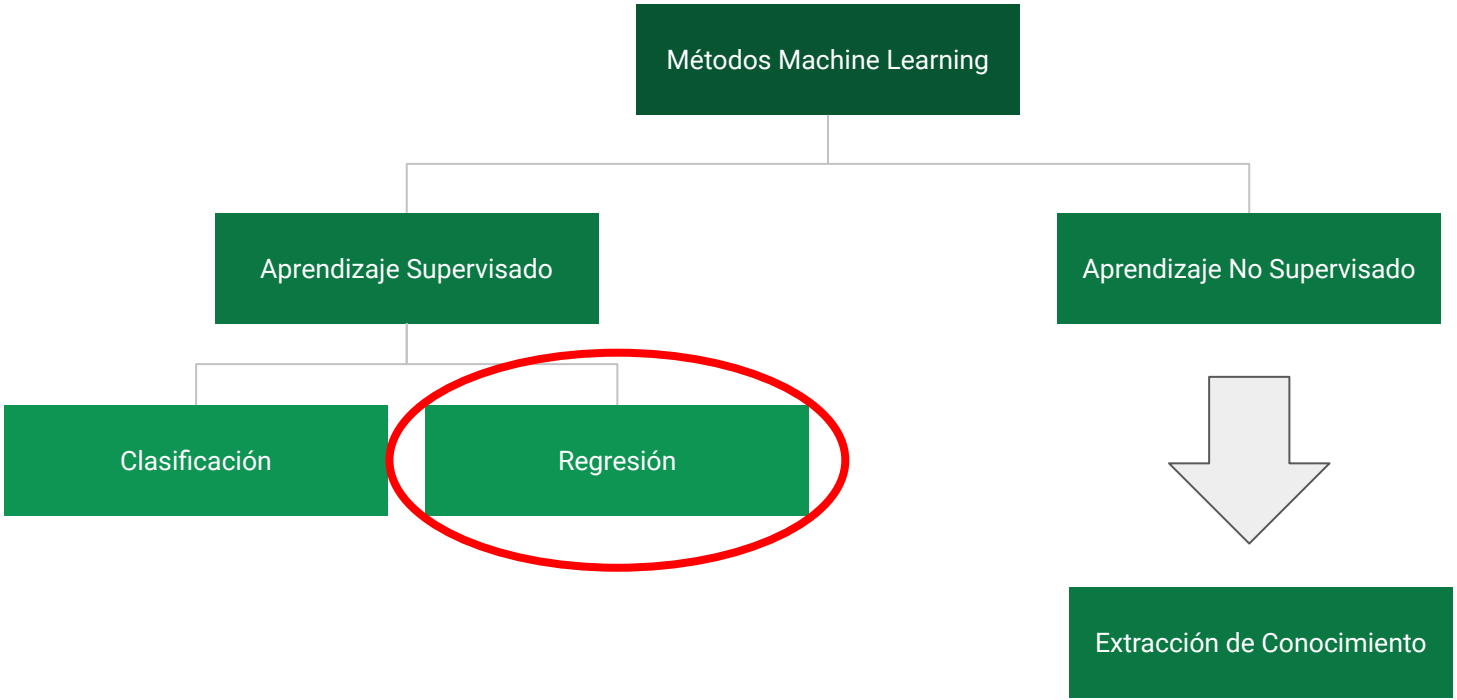
INTRODUCCIÓN

¿Qué es *machine learning*?



MÉTODOS MACHINE LEARNING

What are the process that relates mobility emissions with observed air quality?



APRENDIZAJE SUPERVISADO

El aprendizaje supervisado consiste en encontrar un modelo que aproxime la función que relaciona un conjunto de *inputs* (X) con una variable objetivo u *outputs* (y), minimizando una función de coste que cuantifica la bondad del modelo para la predicción de la variable y.

Matemáticamente

$$f(X) = y$$

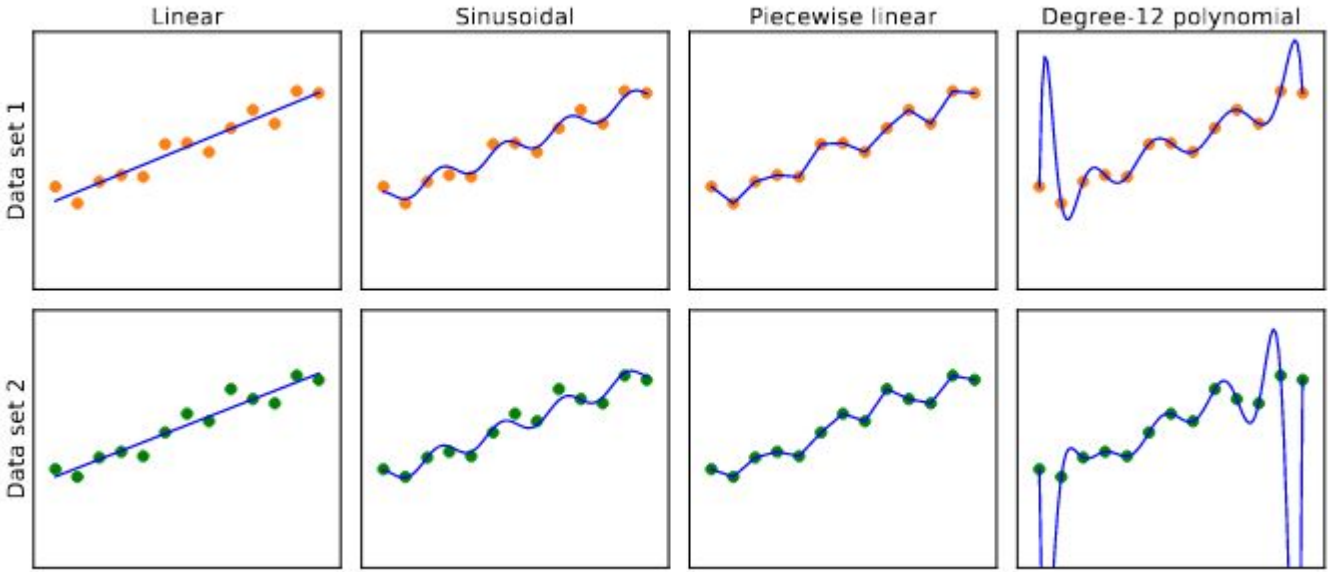
que minimiza

$$C(y_{true}, y_{predicted})$$

Es supervisado ya que necesita conocer el valor real de la variable objetivo en el aprendizaje. Es lo que se denomina *datos etiquetados*

PROBLEMAS DE REGRESIÓN

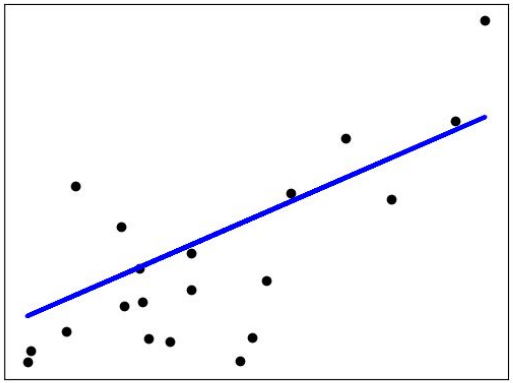
y es una variable continua



BREVE REPASO HISTÓRICO

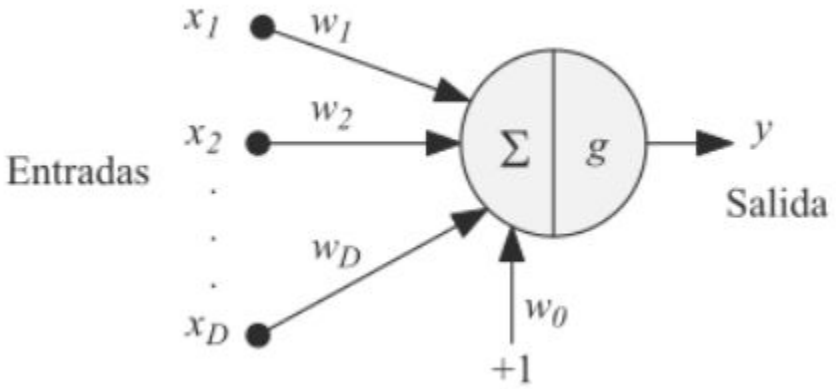
1805. Primer método de ajuste por mínimos cuadrados por Legrenge, extendido por Gauss

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$



BREVE REPASO HISTÓRICO

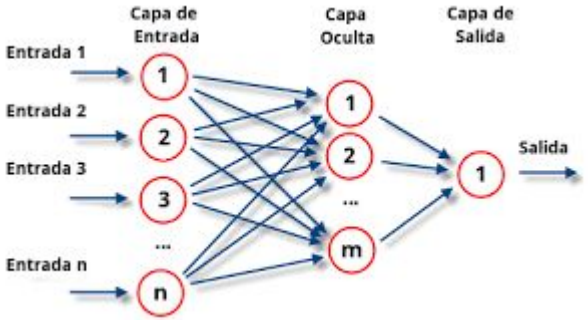
1942. McCuloth and Philips proponen el primer modelo matemático para representar el funcionamiento de una neurona.



$$a = \sum_{i=1}^{D} w_i x_i + w_0 \quad \Rightarrow \quad a = g(\sum_{i=1}^{D} w_i x_i + w_0) = g(\sum_{i=0}^{D} w_i x_i)$$

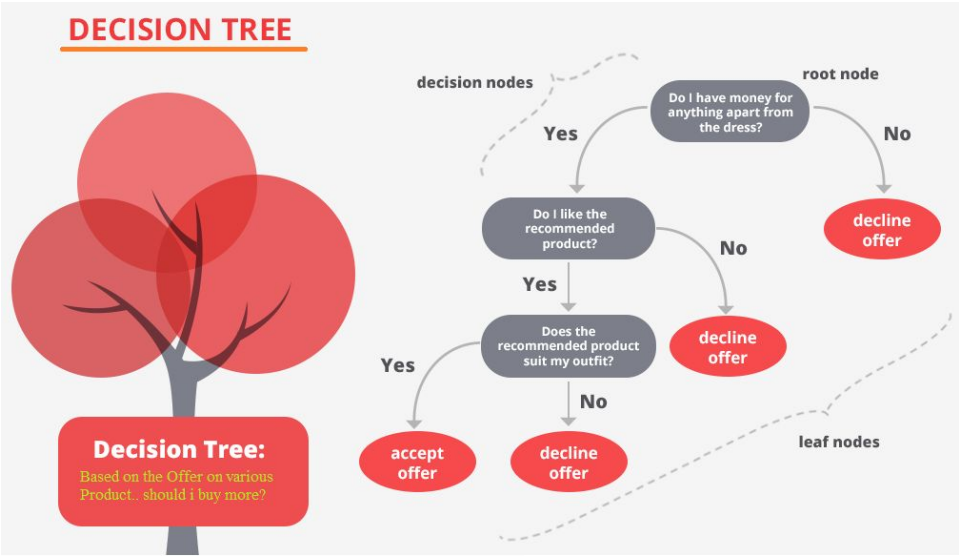
BREVE REPASO HISTÓRICO

1958. McCuloth and Philips proponen el primer modelo matemático para representar el funcionamiento de una neurona.

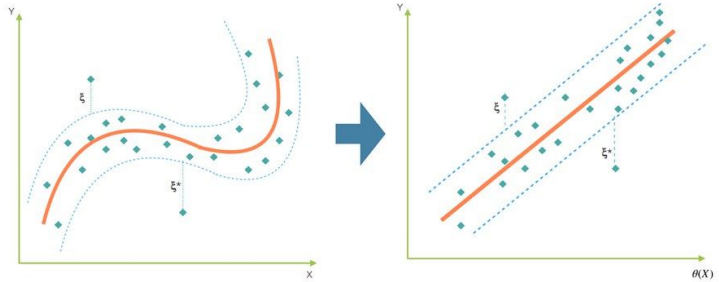


1985. Publicación del algoritmo de *backpropagation* por Rumelhart et al.

OTROS MODELOS



Random Forest



EJERCICIO PRÁCTICO

Se ha medido la concentración media del último año de diferentes metales pesados y compuestos tóxicos en agua en 10000 localidades distintas a lo largo de Europa, en ng/m3. A continuación, se obtiene el incremento de la mortalidad en el último año en dicha localidad.

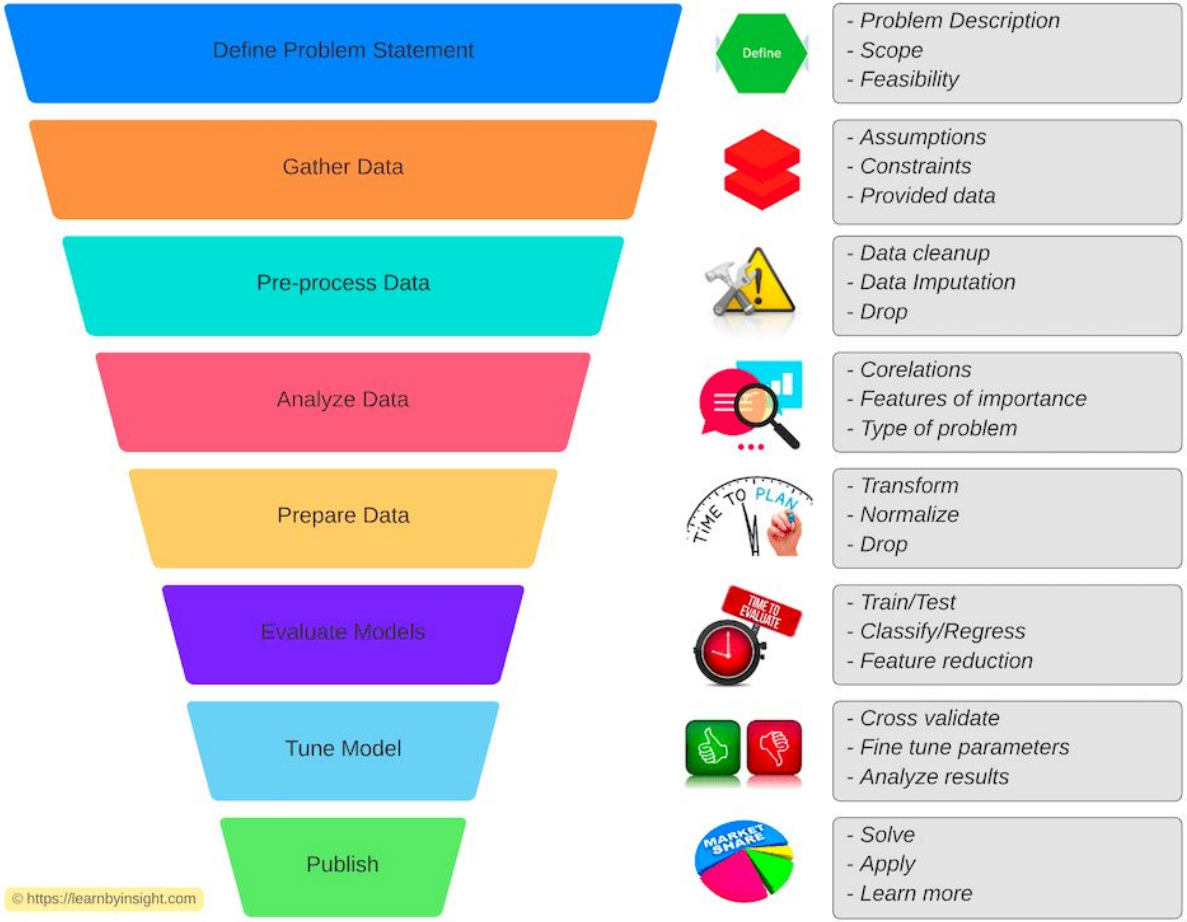
	Fe <dbl>	Ca <dbl>	Cu <dbl>	Hg <dbl>	As <dbl>	Be <dbl>	MortalityIncrement <dbl>
1	6.610855	2.820871	228.6156	21.37445	25.13515	0.9538274	1.1910277
2	9.245143	1.343969	103.4423	20.10891	17.66526	0.9135707	1.8166323
3	7.474468	4.497077	250.9428	33.27214	14.50601	0.9333067	2.9279593
4	9.876095	4.653075	246.6358	44.43551	28.89707	0.9436449	1.9594415
5	8.166730	3.197520	275.6806	37.32236	20.40677	0.9153085	2.4136668
6	6.961055	3.739619	163.7042	30.36394	20.14564	0.9472573	2.0216588
7	6.782077	2.686917	228.7978	33.14800	16.06977	1.0807524	2.5404211
8	6.602718	2.924007	240.1185	41.56409	19.28896	0.9485341	2.4860230
9	7.247903	1.401626	233.1337	33.71688	23.26361	0.9489442	1.9748916
10	5.670633	1.488838	270.3272	30.27260	25.91615	0.9697708	1.5089318

1-10 of 1,000 rows

Previous 1 2 3 4 5 6 ... 100 Next

¿Existe una relación entre la concentración de metales y el incremento de la mortalidad? ¿Es posible generar un modelo que prediga el incremento de la mortalidad en función de la concentración de metales pesados?

OTROS MODELOS



Este *workflow* es flexible, y la mejor forma de conocerlo es a través de la experiencia

Cada PROBLEMA es ÚNICO y debe ser tratado como tal!

BIBLIOGRAFÍA

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.".

Méndez, J. P., & Morales, R. M. (2008). Inteligencia artificial. *Técnicas, métodos y aplicaciones*

Legendre, A. M. (1806). *Nouvelles méthodes pour la détermination des orbites des comètes; par AM Legendre...* chez Firmin Didot, libraire pour lew mathematiques, la marine, l'architecture, et les editions stereotypes, rue de Thionville.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science.