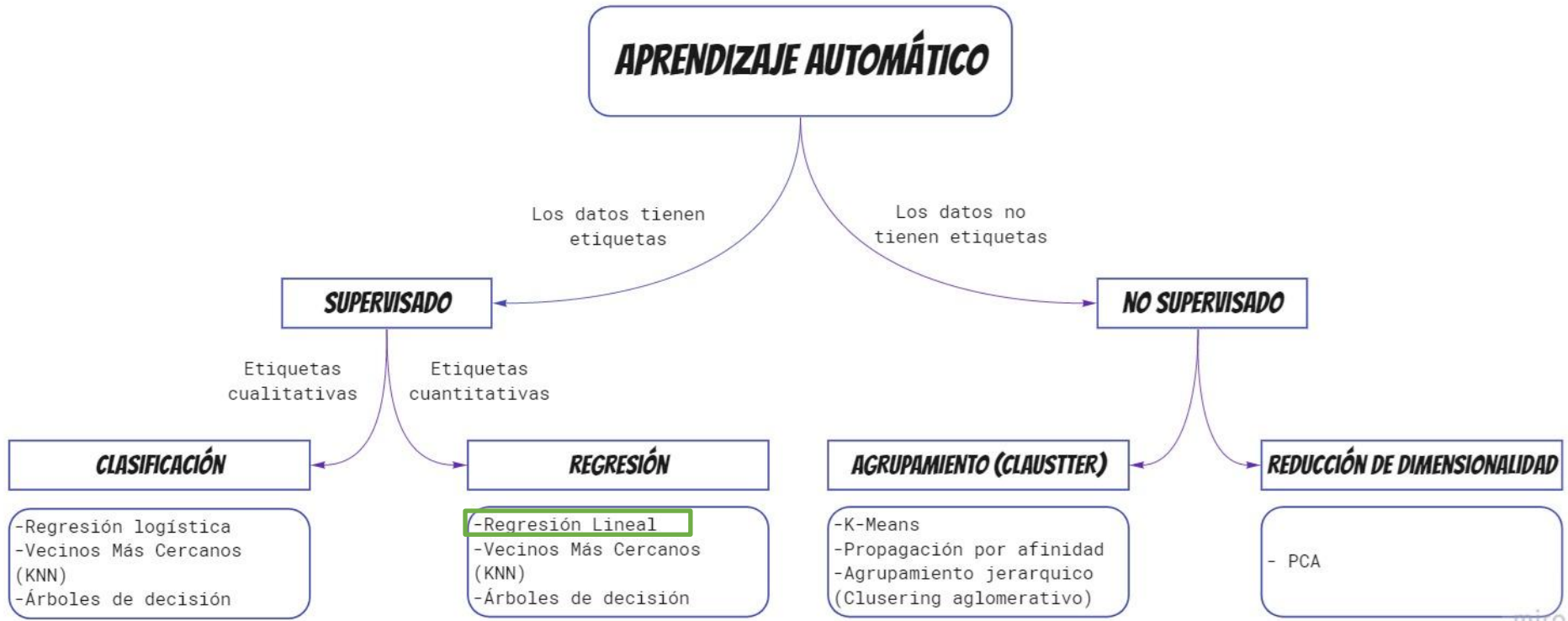


# Ciencia de Datos

- Módulo 2

Regresión lineal y métricas

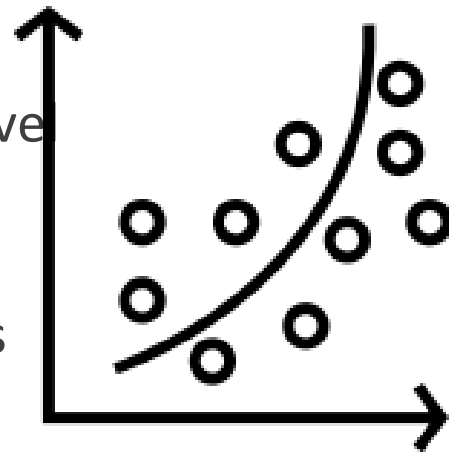




## Regresiones

Cuando nos referimos a un modelo de regresión, son modelos que tratan de predecir variables **numéricas** como precios, cantidades, pesos, etc. Se basan en estimar el valor de la variable a predecir en función de las otras variables observadas.

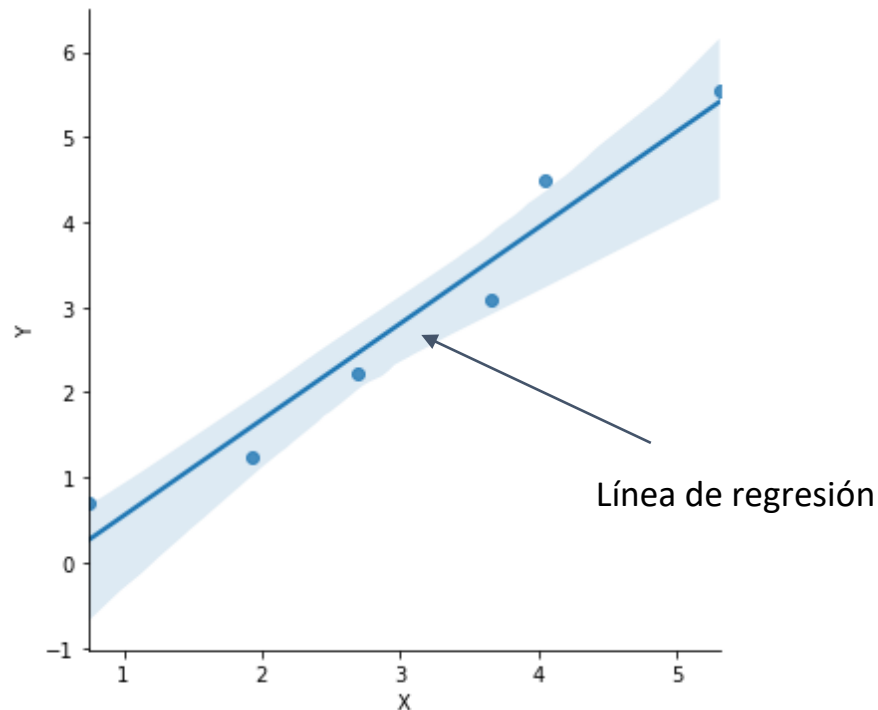
Por ejemplo, ver la probabilidad de lluvia dado el nivel de humedad y la velocidad del viento. Teniendo esta relación podemos saber cual es la probabilidad de lluvia en cualquier momento en el que tengamos el nivel de humedad y la velocidad del viento.



# Regresiones - Regresión lineal

Ejemplo:

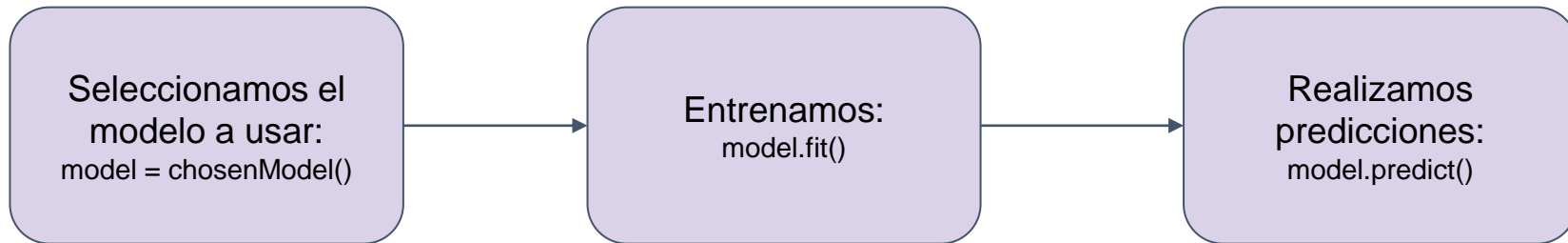
X	Y
0.75	0.70
1.93	1.23
2.69	2.22
3.66	3.09
4.05	4.5
5.31	5.54



## Regresión lineal

En Python usaremos la librería Scikit-Learn para aplicar los modelos de ML a nuestros datos.

Esta librería de código abierto nos permite entrenar y realizar predicciones en forma secuencial.



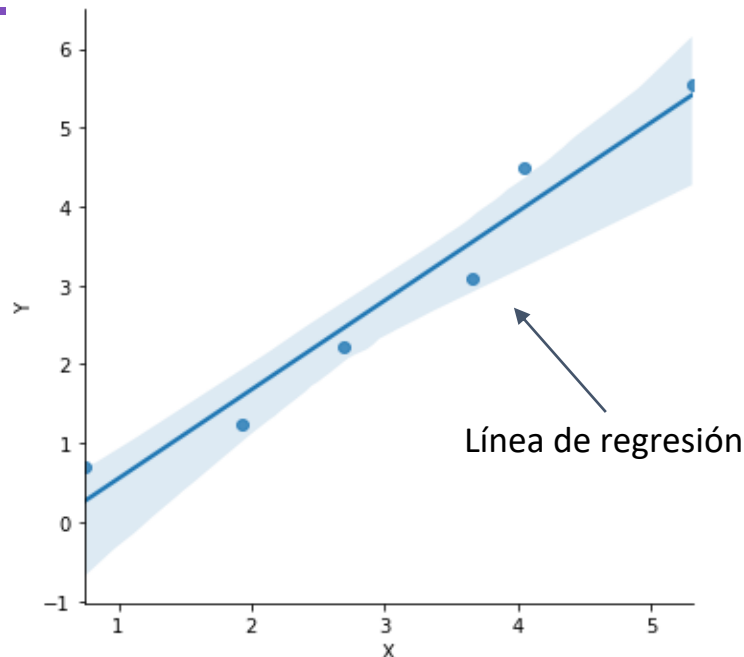
# Regresiones - Regresión lineal

Fórmula:

$$Y = \beta_0 + \beta_1 \cdot x$$

Intercepto  
(punto origen en eje Y)

Pendiente de línea



X	Y
0.75	0.70
1.93	1.23
2.69	2.22
3.66	3.09
4.05	4.5
5.31	5.54

[https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression\\_es.html](https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_es.html)



## Métricas de evaluación

Entrenamos el modelo con el método `.fit` y los datos de entrenamiento

Luego utilizamos el método `.predict` con los datos de testeo (`X_test`) para crear un objeto con las predicciones realizadas por el modelo. Se comparará los valores predichos por el modelo con los valores reales (`y_test`)

**Regresión:** Mide la diferencia entre la predicción y el valor real



Fuente: <https://www.edx.org/es/course/machine-learning-with-python-from-linear-models-to>



## Métricas de evaluación - Regresión

**MAE:** Media del Error Absoluto (mean\_absolute\_error) es el promedio de la diferencia absoluta entre el valor observado y los valores predichos. El error absoluto medio o MAE es un puntaje lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio.

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

**MSE:** La Media del Error cuadrático (mean\_squared\_error) es la media de los errores al cuadrado

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$





## Métricas de evaluación - Regresión

**RMSE:** La raíz del error cuadrático medio  $\text{np.sqrt}(\text{mean\_squared\_error})$  raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor pronosticado. Indica el ajuste absoluto del modelo a los datos, cuán cerca están los puntos de datos observados de los valores predichos del modelo. El error cuadrático medio o RMSE es una medida absoluta de ajuste

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

**R<sup>2</sup>** (r2\_score) es la proporción de la varianza total explicada por el modelo. La mejora en el modelo de regresión da como resultado aumentos proporcionales en R-cuadrado. De 0 a 1 siendo 1 el mejor.

$$R^2 = 1 - \frac{\sum (y_j - \hat{y}_j)^2}{\sum (y_j - \bar{y}_j)^2}$$

