dcc
CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

# Curso DM
# "Regresiones"

Primavera 2023

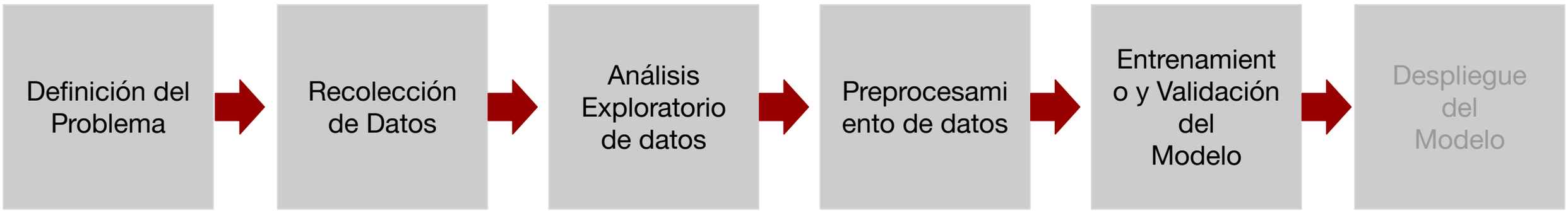
Basado en las slides de Felipe Bravo

# Metodología para el tratamiento de los datos

*Múltiples fuentes de datos:*
*Temperatura, presión, lluvias, nubosidad, relieve, tendencias del cambio climático, etc…*

*Limpieza y preparación de los datos. Trabajo con datos faltantes, selección/generación de características relevantes*

*Poner el modelo en producción*




*Heladas en el valle central provocan pérdidas en la producción de frutales.*

***Solución:*** *Predicción de Heladas*

*Exploración usando estadísticas de resumen (frecuencias, medidas de tendencia central y dispersión) y visualización.*

*Creación, evaluación y ajuste de modelos de Machine Learning*

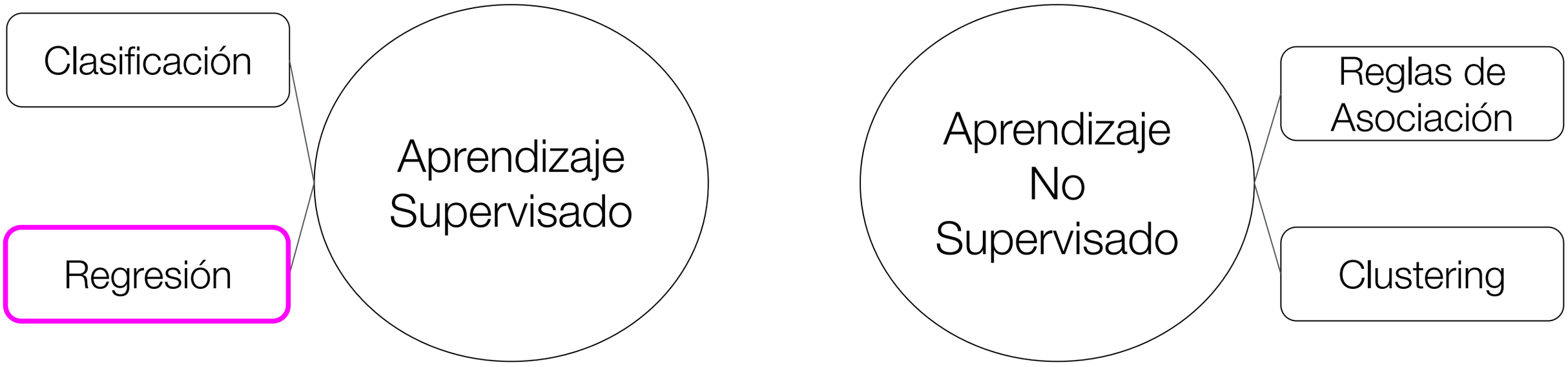
***Retroalimentación:*** *Evaluar el rendimiento del modelo desplegado en los nuevos datos. Reentrenar en caso que el modelo quede desactualizado.*

Fuente: Repositorio Github Pablo Badilla. https://github.com/pbadillatorrealba/MDS7202

# Machine Learning

- Machine Learning: Aprendizaje Automático, o Aprendizaje de Máquinas
- Estudio, diseño y desarrollo de algoritmos que permiten a los computadores aprender sin ser explícitamente programados (Arthur Samuel, 1959).

**Tipos de aprendizaje y sus tareas:**



# Sobre la Regresión

- Técnica utilizada para modelar la relación entre una variable dependiente (objetivo) y una o más variables independientes (predictoras o explicativas).
- Se utiliza para predecir los valores numéricos de un conjunto de datos determinado.
- Viene del área de Machine Learning
- Método de “aprendizaje supervisado”

# Sobre la Regresión

La regresión y la clasificación (ambas tareas predictivas) son bastante similares entre sí, con la diferencia que:

- En la clasificación las salidas son discretas, en la regresión son valores continuos.

**Regression Data**

| $X_1$ | $X_2$ | $X_3$ | $X_p$ | Y |
|---|---|---|---|---|
| | | | | 5.2 |
| | | | | 1.3 |
| | | | | 23.0 |
| | | | | 7.4 |

Numeric Target

**Classification Data**

| $X_1$ | $X_2$ | $X_3$ | $X_p$ | Y |
|---|---|---|---|---|
| | | | | cat |
| | | | | dog |
| | | | | cat |
| | | | | cat |

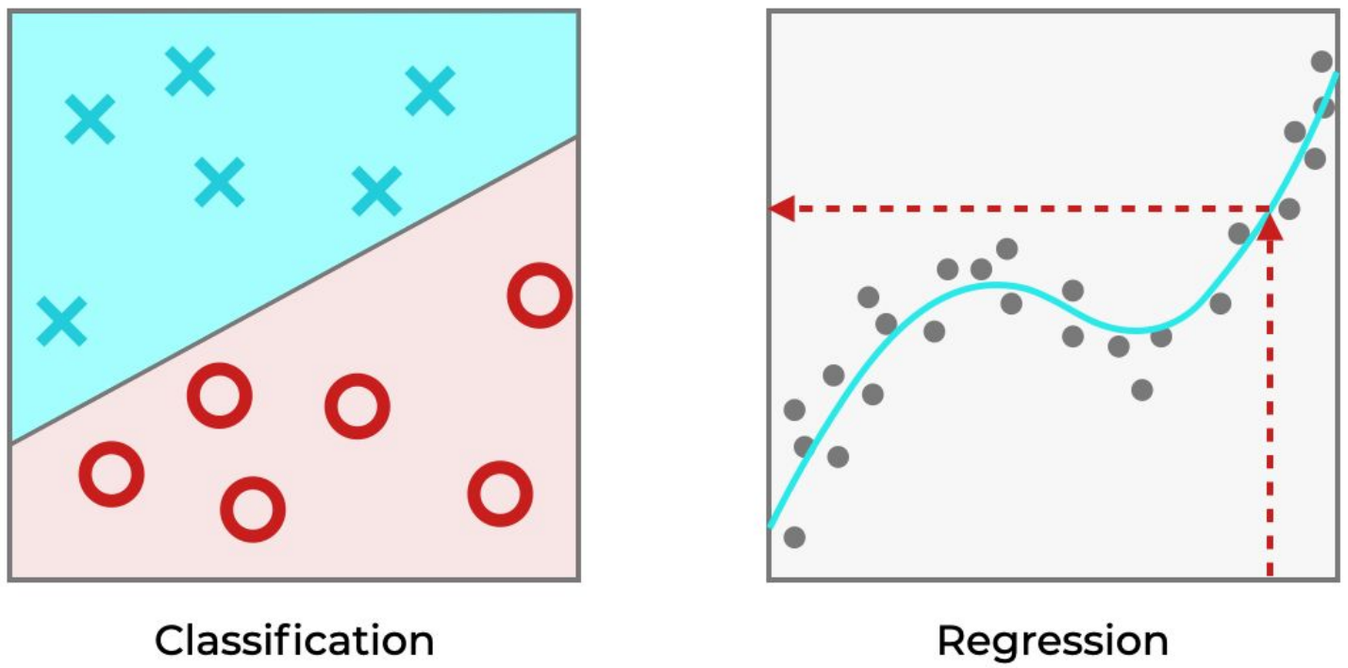
Categorical "Labels"

Fuente: https://www.sharpsightlabs.com/blog/regression-vs-classification/

# Sobre la Regresión

La regresión y la clasificación (ambas tareas predictivas) son bastante similares entre sí, con la diferencia que:

- En la clasificación las salidas son discretas, en la regresión son valores continuos.




Fuente: https://www.sharpsightlabs.com/blog/regression-vs-classification/

# Sobre la Regresión

Ejemplos de aplicación:

- Modelado medioambiental.
- Predicción del consumo energético.
- Predicción o pronóstico financiero.
- Proyección de ventas.
- Estimación de la demanda de productos para la gestión del stock.

# Sobre la Regresión

## Ejemplo: Predicción del precio de una propiedad

| habitaciones_desde | bannos_desde | metros_construidos_desde | metros_terreno | precio_desde |
|---|---|---|---|---|
| 4.0 | 3.0 | 105.0 | 120.0 | 3.051520e+08 |
| 3.0 | 3.0 | 95.0 | 113.0 | 1.928006e+08 |
| 4.0 | 3.0 | 120.0 | 137.0 | 2.663145e+08 |
| 2.0 | 2.0 | 72.0 | 86.0 | 1.886394e+08 |
| 4.0 | 3.0 | 120.0 | 130.0 | 2.912815e+08 |
| ... | ... | ... | ... | ... |

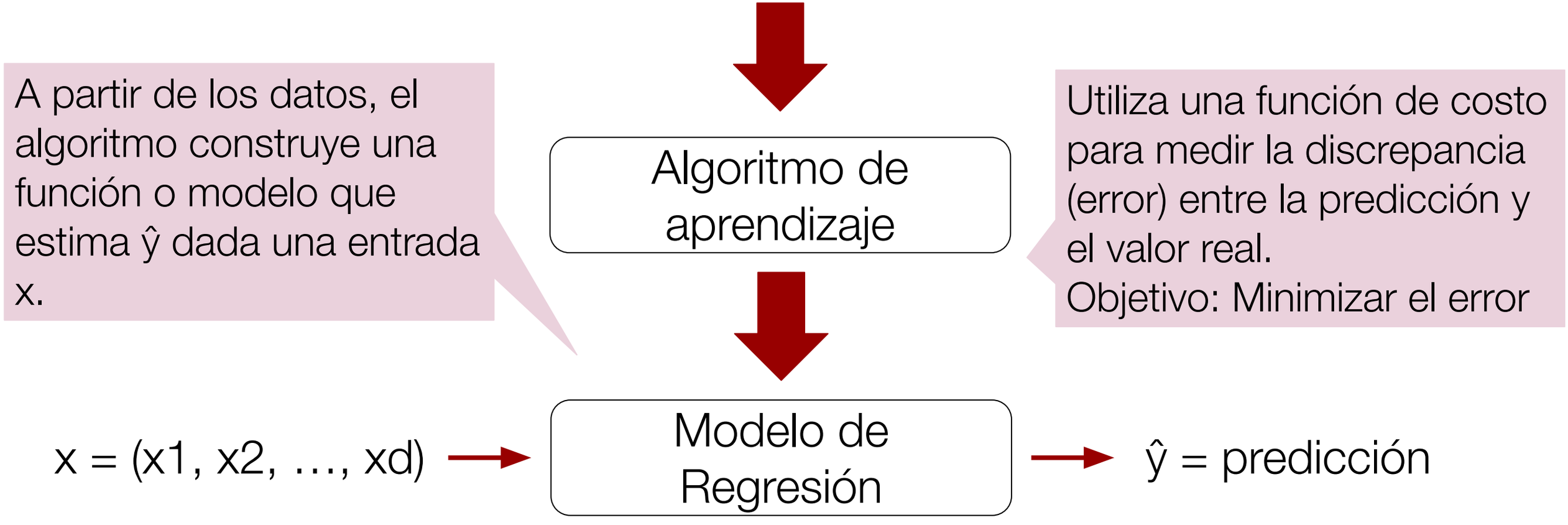
x = Variables independientes (o explicativas)

y = Variable dependiente

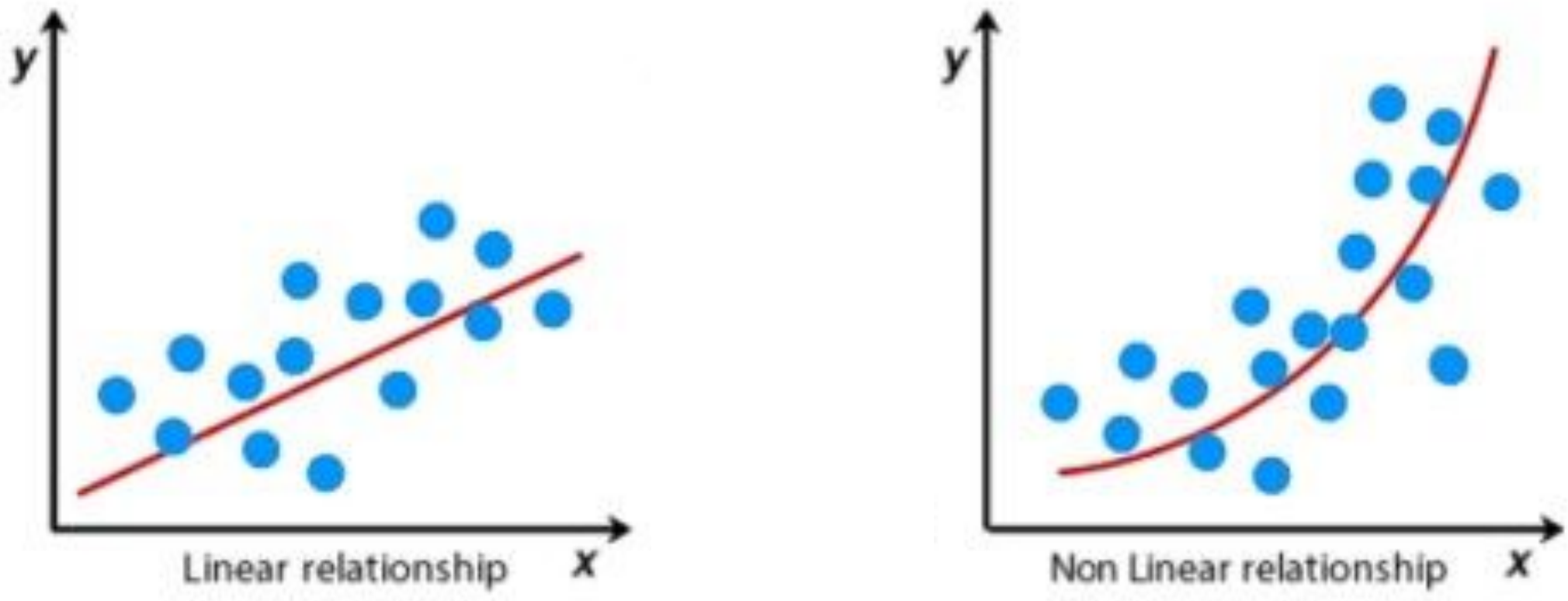
¿Qué hace un algoritmo de Machine Learning para resolver la tarea de regresión?

# Sobre la Regresión

| x1 | x2 | ... | xd | y |
|---|---|---|---|---|
| | | | | |
| | | | | |
| … | … | … | … | … |



# Modelo lineal vs no lineal




Fuente: https://www.analyticsvidhya.com/

# Modelos de Regresión

- Un modelo de regresión se usa para modelar la relación de una variable dependiente **y** numérica con n variables independientes $x_1$, $x_2$, ..., $x_n$ .
- A grandes rasgos queremos conocer el valor **esperado** de **y** a partir los valores de **x**:

$$\mathbb{E}(y|x_1, x_2, \ldots, x_n)$$

Usamos estos modelos cuando creemos que la variable de respuesta **y** puede ser modelada por otras variables independientes también conocidas como **covariables o atributos**.

# Modelos de Regresión

- Para realizar este tipo de análisis necesitamos un dataset formado por **m** observaciones que incluyan tanto a la variable de respuesta como a cada uno de los atributos.

- Nos referimos al proceso de **ajustar una función de regresión** al proceso en que a partir de los datos inferimos una función de hipótesis **h** que nos permite predecir valores de **y** desconocidos usando los valores de los atributos.

# Modelos de Regresión

- A este proceso de ajustar una función a partir de los datos se le llama en machine learning como **entrenamiento**.
- Se entiende que las funciones **aprenden** a partir de los datos.
- Como necesitamos observaciones donde el valor de **y** sea conocido para aprender la función, se le llama a este tipo de técnicas como técnicas de **aprendizaje supervisado**.
- Cuando **y** es una variable categórica hablamos de un problema de **clasificación**.

# Tipos de modelos


## Simple

Relación entre la variable objetivo y una variable independiente

| metros_terreno | precio_desde |
|---|---|
| 120.0 | 3.051520e+08 |
| 113.0 | 1.928006e+08 |
| 137.0 | 2.663145e+08 |
| 86.0 | 1.886394e+08 |
| 130.0 | 2.912815e+08 |
| ... | ... |

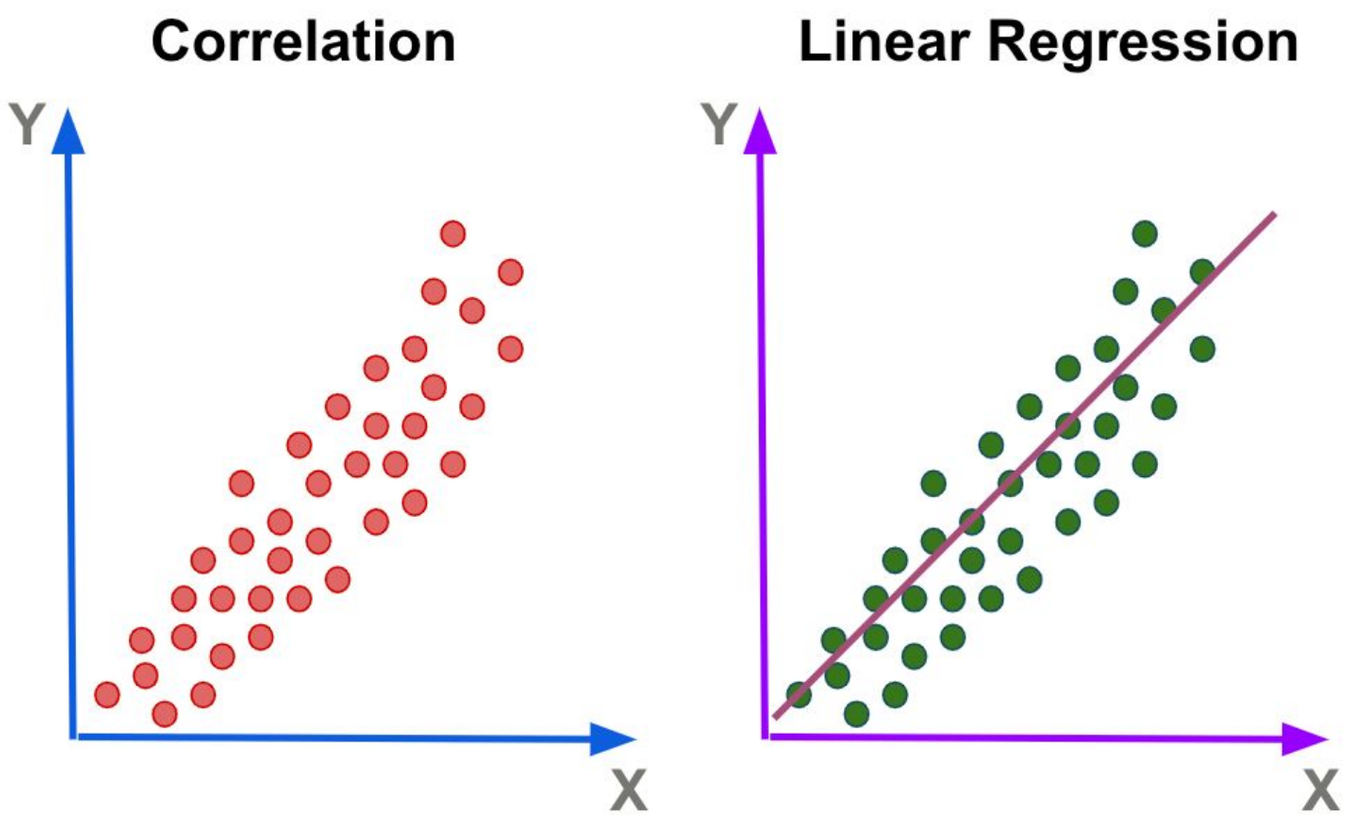
## Múltiple

Relación entre la variable objetivo y dos o más variables independientes

| bannos_desde | metros_terreno | precio_desde |
|---|---|---|
| 3.0 | 120.0 | 3.051520e+08 |
| 3.0 | 113.0 | 1.928006e+08 |
| 3.0 | 137.0 | 2.663145e+08 |
| 2.0 | 86.0 | 1.886394e+08 |
| 3.0 | 130.0 | 2.912815e+08 |
| ... | ... | ... |

# Regresión Lineal Simple

Cuantifica el impacto de un cambio en una variable sobre otra, es decir, cuantifica la relación. La correlación sólo indica si existe una asociación fuerte, moderada o débil.




Es una forma de ajustar una línea recta óptima a los datos; **permite cuantificar relaciones y hacer predicciones.**

Fuente: https://www.lennysnewsletter.com/

# Regresión Lineal Simple

Se tiene una única variable independiente **x** para modelar la variable dependiente **y**. Se asume la siguiente relación lineal entre la variables:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \forall i$$

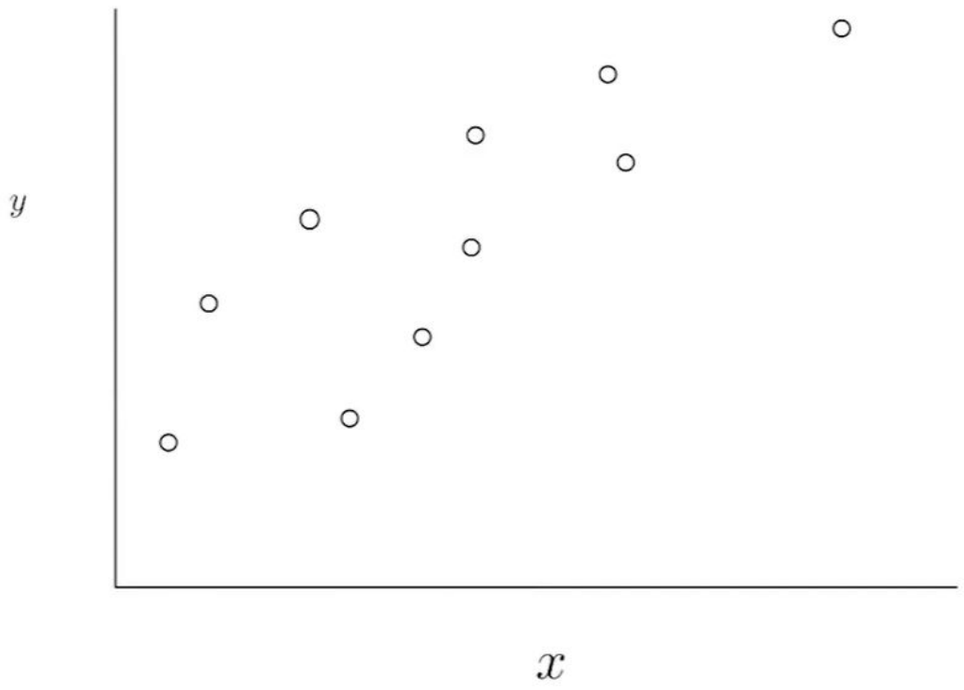
- El parámetro $\beta_0$ representa el intercepto de la recta (el valor de **y** cuando **x** vale cero).
- El parámetro $\beta_1$ es la pendiente y representa el cambio de **y** cuando variamos el valor de **x**. Entre mayor sea la magnitud de este parámetro mayor será la relación lineal entre las variables.
- Los valores $\varepsilon_i$ corresponden a los errores asociados al modelo.

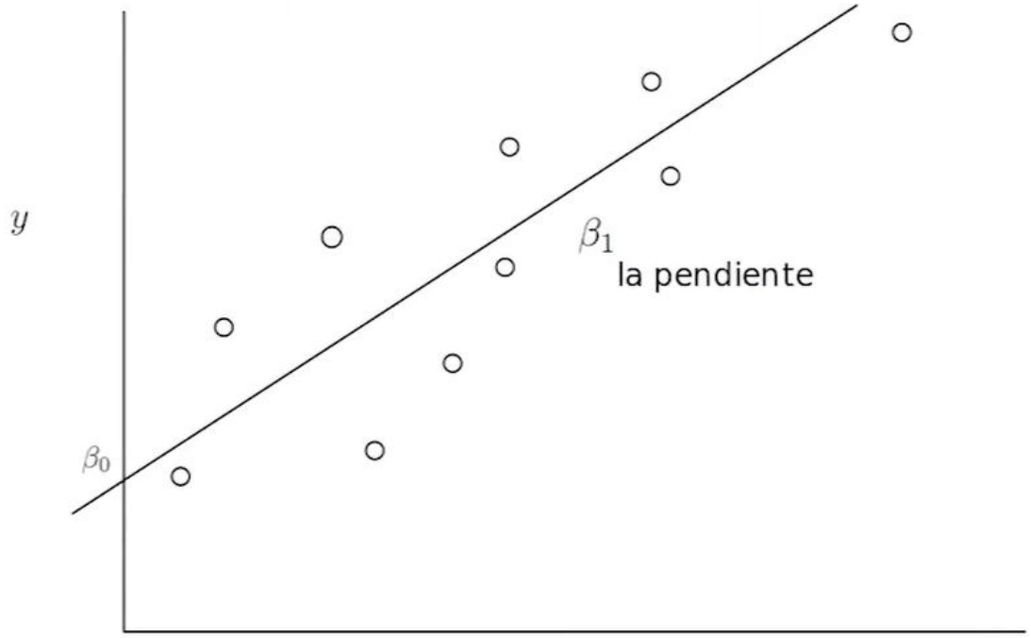
# Regresión Lineal Simple

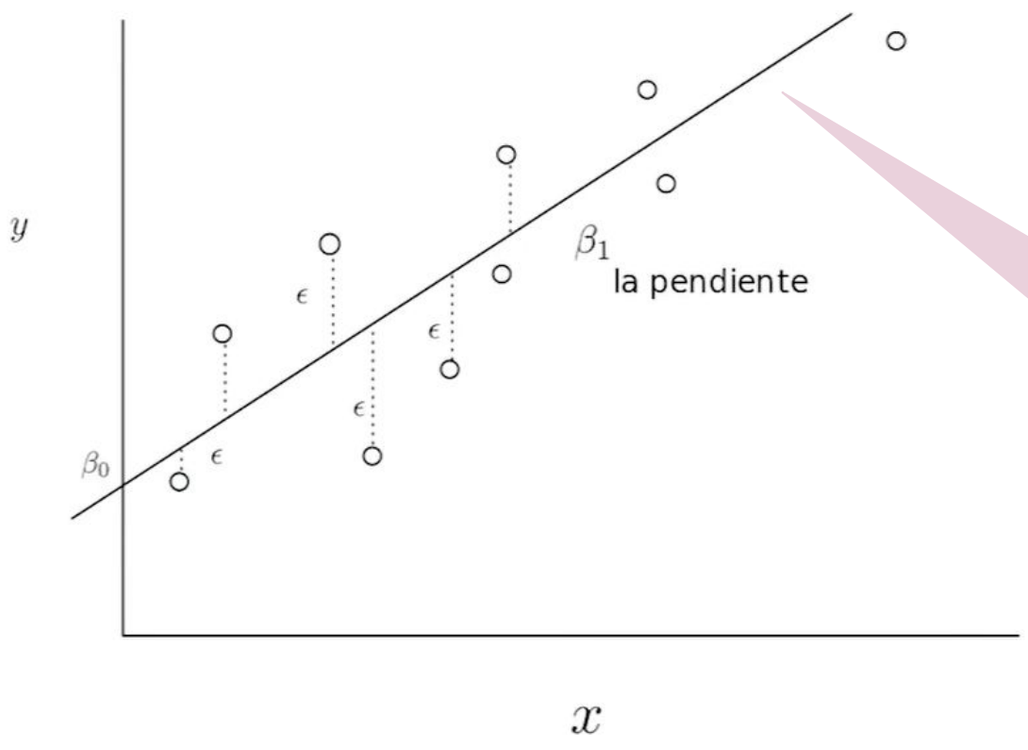
- Tenemos que encontrar una función lineal o recta $\mathbf{h_\beta}$ que nos permita encontrar una estimación de **y**, $\hat{\mathbf{y}}$ para cualquier valor de **x** con el mínimo error esperado.

$$h(x) = \beta_0 + \beta_1 x$$

# Regresión Lineal Simple




¿Cómo encontrar la recta óptima?

# Mínimos Cuadrados

- El método de mínimos cuadrados ordinarios se usa para estimar $\beta_0$ y $\beta_1$ minimizando la suma de los errores cuadráticos (SSE) de los datos observados.
- Supongamos que tenemos **m** observaciones de **y** y de **x**, calculamos la suma de los errores cuadráticos (SSE) o E de error de la siguiente forma:

$$E = \sum_{i=1}^{m} (y_i - h(x_i))^2 = \sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (1)$$

# Mínimos Cuadrados

- Para encontrar los parámetros que minimizan el error calculamos las derivadas parciales de SSE respecto a $\beta_0$ y $\beta_1$ . Luego igualamos las derivadas a cero y resolvemos la ecuación para despejar los parámetros.

$$\frac{\partial E}{\partial \beta_0} = -2 \sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{2}$$

$$\frac{\partial E}{\partial \beta_1} = -2 \sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \tag{3}$$

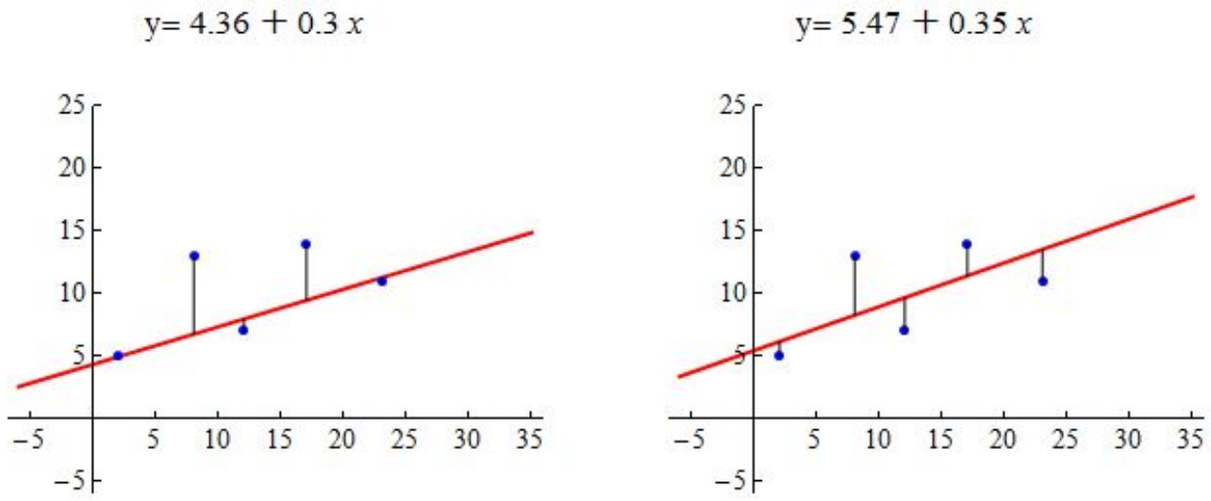
# Mínimos Cuadrados (2)

- Del sistema de ecuaciones anterior se obtienen las soluciones normales:

$$\hat{\beta}_1 = \frac{\sum_i^m (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^m (x_i - \overline{x})^2} \tag{4}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{5}$$

- El modelo ajustado representa la recta de mínimo error cuadrático.



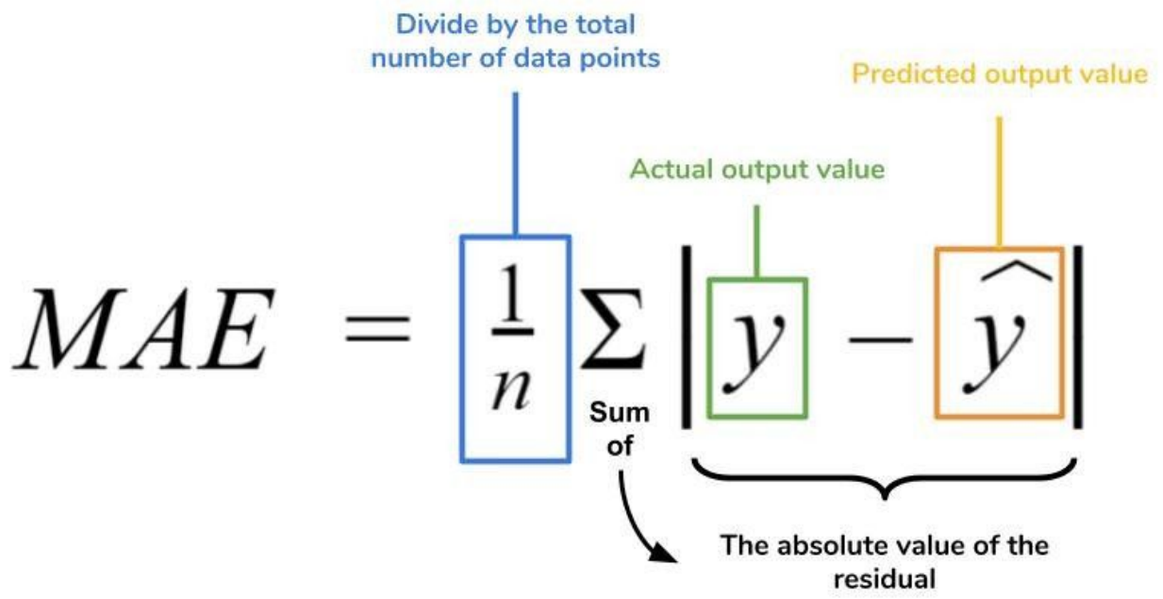

¿Cómo evaluar mi modelo de regresión?

Fuente: https://courses.lumenlearning.com/

# Validación de la capacidad de generalización

El enfoque clásico de Minería de Datos mide la capacidad del modelo para **predecir correctamente** la salida para **nuevos** datos:

1. Separar dataset en **train** y **test.**
2. Entrenar el modelo en la partición **train**
3. Evaluar el modelo en la partición **test.** Por ejemplo, calculando el Error Absoluto Medio (MAE) o el Error Cuadrático Medio (RMSE) (predicción vs valor real).




Fuente: https://www.aionlinecourse.com/

# Validación de la bondad de ajuste: Coeficiente de Determinación $R^2$

- Una vez ajustado nuestro modelo lineal debemos evaluar la calidad del modelo.
- Una medida muy común es el coeficiente de determinación $R^2$. Para calcularlo debo calcular otros errores distintos a los errores cuadráticos SSE.
- Se define a la suma cuadrática total (**SST**) como el error predictivo cuando usamos la media y para predecir la variable de respuesta y (es muy similar a la varianza de la variable):

$$\mathrm{SST} = \sum_{i}^{m} (y_i - \bar{y})^2$$

# Coeficiente de Determinación $R^2$

- Luego tenemos a la suma de los cuadrados explicada por el modelo (SSM) que nos indica la variabilidad de los valores predichos por el modelo respecto a la media.

$$\text{SSM} = \sum_{i}^{m} (\hat{y}_i - \bar{y})^2$$

- Se define el coeficiente de determinación para un modelo lineal R2 como:

$$R^2 = \frac{\text{SSM}}{\text{SST}} = \frac{\sum_i^m (\hat{y}_i - \bar{y})^2}{\sum_i^m (y_i - \bar{y})^2}$$

El coeficiente adquiere valores entre 0 a 1, mientras más cercano a 1, mayor será la calidad del modelo.

El valor de R2 es equivalente a la correlación lineal (Pearsons) entre **y** e **ŷ** al cuadrado.

$$R^2 = \text{cor}(y, \hat{y})^2$$

# Consideraciones

- Introducir valores para X fuera del rango (visto en los datos) se consideraría extrapolación y puede ser muy arriesgado.
- No es recomendable utilizar un modelo para predecir fuera del rango de los datos observados sobre los que se construyó el modelo.

# Regresión Lineal Múltiple

- Supongamos que tenemos n variables independientes: $x_1$, $x_2$, ..., $x_n$.
- Intuitivamente, estas variables en conjunto podrían explicar de mejor manera la variabilidad de la variable de respuesta **y** que un modelo simple.
- Se define un modelo lineal multivariado de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_n x_{i,n} + \epsilon_i \quad \forall i \in \{1, m\}$$

# Regresión Lineal Múltiple

- En el modelo multivariado se extienden todas las propiedades del modelo lineal simple.
- Se puede representar el problema de manera matricial:

$$Y = X\beta + \epsilon$$

- Donde **Y** es un vector de m×1 de variables de respuesta:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

# Regresión Lineal Múltiple

- X es una matriz de m×(n + 1) con las variables explicativas. Tenemos m observaciones de las n variables. La primera columna es constante igual a 1, ($x_{i,0} = 1 \ \forall i$) para incluir la variable de intercepto $\beta_0$ de manera limpia.

$$X = \begin{pmatrix} x_{1,0} & x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,0} & x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m,0} & x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix}$$

- Luego, β es un vector de parámetros de (n + 1) × 1

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

# Regresión Lineal Múltiple

- Finalmente, ε es un vector con los errores del modelo de dimensiones $m \times 1$.

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}$$

- Usando la notación matricial, podemos ver que la suma de los errores cuadráticos (SSE) se puede expresar como:

$$\text{SSE} = (Y - X\beta)^T (Y - X\beta)$$

- Minimizando esta expresión derivando el error en función de β e igualando a cero se llega a las ecuaciones normales:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

# Supuestos del Modelo Lineal

Cada vez que ajustamos un modelo lineal estamos asumiendo implícitamente ciertos supuestos sobre los datos.

| | | |
|---|---|---|
| 01 | Linealidad | La variable de respuesta se relaciona linealmente con los atributos. |
| 02 | Normalidad | Los errores tienen distribución normal de media cero: $\varepsilon_i \sim N(0, \sigma^2)$ |
| 03 | Homocedasticidad | Los errores tienen varianza constante (mismo valor de $\sigma^2$). |
| 04 | Independencia | Los errores son independientes entre sí. |

# Interpretación Probabilística

- Considerando los supuestos anteriores podemos ver que la densidad de probabilidad (PDF) de los errores ε se definen por una normal de media cero y varianza constante:

$$\text{PDF}(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$$

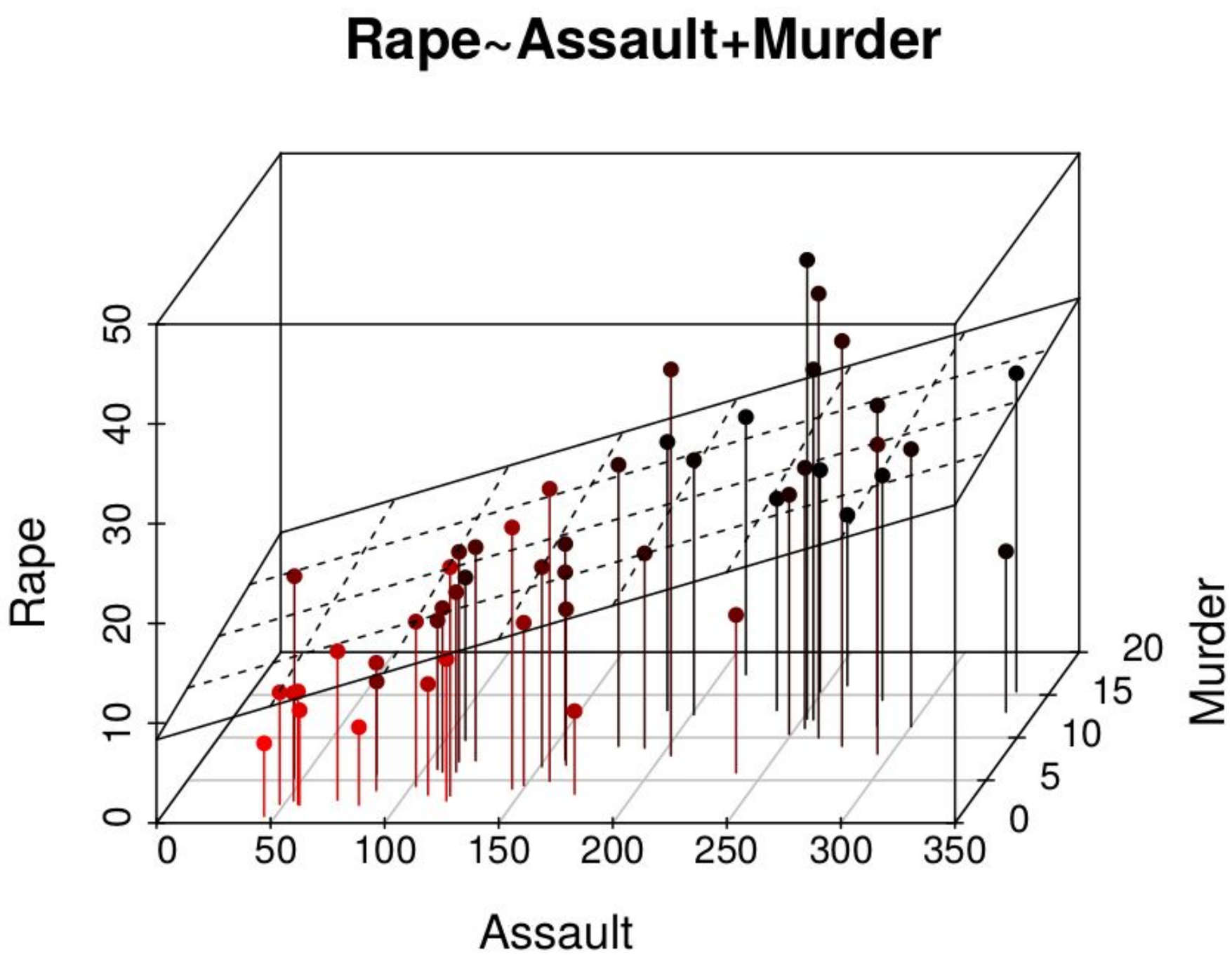
- Esto implica que:

$$\text{PDF}(y_i|x_i;\beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - h_\beta(x_i))^2}{2\sigma^2}\right)$$

# Interpretación Probabilística

- Lo que implica que la distribución de **y** dada los valores de **x** y parametrizada por β sigue una distribución normal.
- Luego si uno estima los parámetros de β usando una técnica de estimación llamada **máxima verosimilitud** llega a los mismos resultados que haciendo una estimación por mínimos cuadrados.
- Esto nos dice que cuando estimamos los parámetros del modelo usando mínimos cuadrados estamos realizando las mismas hipótesis probabilísticas mencionadas anteriormente.

# Ejemplo


Rape~Assault+Murder
Rape
0
10
20
30
40
50
0
50
100
150
200
250
300
350
Assault
Murder
0
5
10
15
20

# Machine Learning en Python


Python es considerado uno de los lenguajes de programación más populares que provee bibliotecas potentes para Data Science y ML. https://www.python.org/downloads/

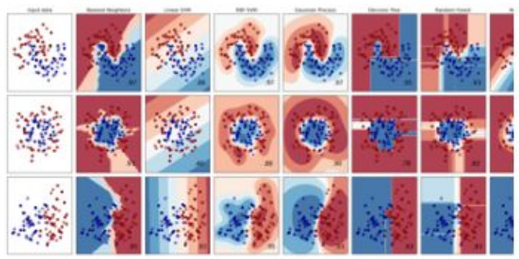

Scikit-learn es una biblioteca de aprendizaje automático de código abierto que permite el aprendizaje supervisado y no supervisado. https://scikit-learn.org/stable/index.html

# Machine Learning en Python

## Classification

Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.
**Algorithms:** Gradient boosting, nearest neighbors, random forest, logistic regression, and more...

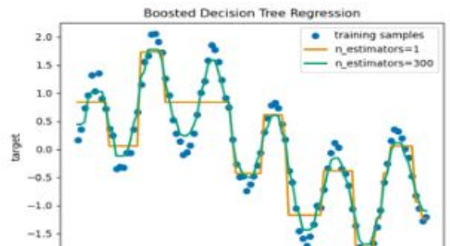


Examples

## Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.
**Algorithms:** Gradient boosting, nearest neighbors, random forest, ridge, and more...

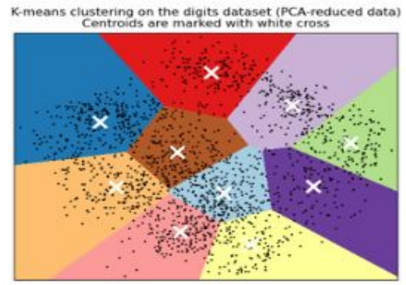



Examples

## Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes
**Algorithms:** k-Means, HDBSCAN, hierarchical clustering, and more...

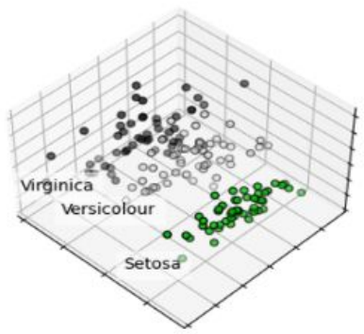



Examples

## Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency
**Algorithms:** PCA, feature selection, non-negative matrix factorization, and more...

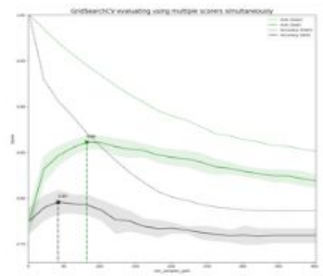



Examples

## Model selection

Comparing, validating and choosing parameters and models.

**Applications:** Improved accuracy via parameter tuning
**Algorithms:** grid search, cross validation, metrics, and more...

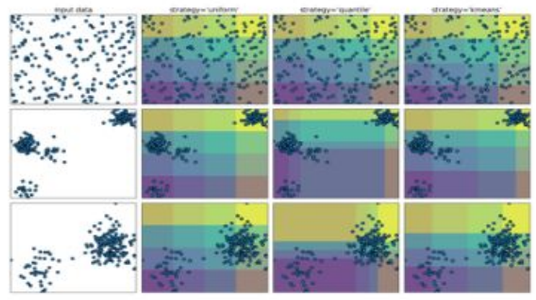


Examples

## Preprocessing

Feature extraction and normalization.

**Applications:** Transforming input data such as text for use with machine learning algorithms.
**Algorithms:** preprocessing, feature extraction, and more...



Examples

https://scikit-learn.org/

# Machine Learning en Python


Herramientas para:

- Preprocesamiento de datos
- Selección de modelos
- Ajuste de modelos
- Evaluación de modelos
- ...

dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl

/ DCCUCHILE