Introducción al Aprendizaje Automático

Agenda del día (Lecture):

- ¿Qué es la inteligencia artificial?
- Historia y evolución de la IA
- Machine Learning
- Tipos de aprendizaje
- Recapitulación de la exploración de datos
- Scikit-learn
- Estandarización de datos
- ¿Qué es correlación?
- Coeficiente de Pearson (relaciones lineales)
- Regresión lineal

Agenda del día (Practice - Homework):

- Flujo de trabajo en Scikit-Learn
- Regresión lineal
  - Regresión lineal simple
  - "Derivación" para obtener m y b
  - Evaluación de modelos
  - Hipótesis estadísticas de la regresión lineal
  - Regresión lineal múltiple
  - Regresión lineal asociada a funciones polinómicas
  - ¿cómo pasar cualquier función a un polinomio? (serie de Taylor/Maclaurin)
- Práctica adicional (No existen code reviews en este apartado)

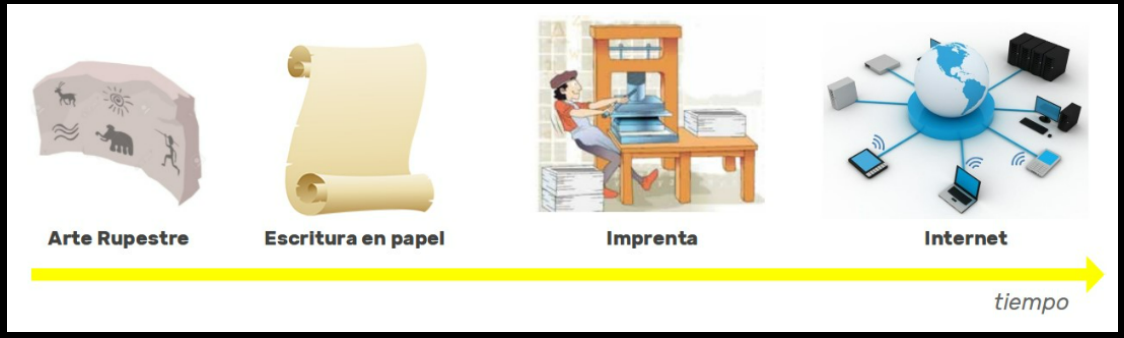
Objetivos de aprendizaje

- Comprender la historia y evolución de la Inteligencia Artificial (específicamente Machine Learning).
- Entender el concepto de Aprendizaje supervisado vs Aprendizaje no supervisado.
- Comprender el concepto de Correlación.
- Conocer el flujo de trabajo en ML.
- Identificar falencias en los Datos.
- Conocer qué transformaciones se pueden hacer sobre los Datos.
- Conocer Scikit-Learn.
- Comprender las ventajas de reescalar los Datos.
- Comprender el funcionamiento del algoritmo Regresión Lineal.

Inteligencia Artificial

Una tecnología es relevante en la medida en que altera un sistema productivo. Cada vez que una tecnología alteró un sistema productivo, tuvo consecuencias relevantes en el modelo social

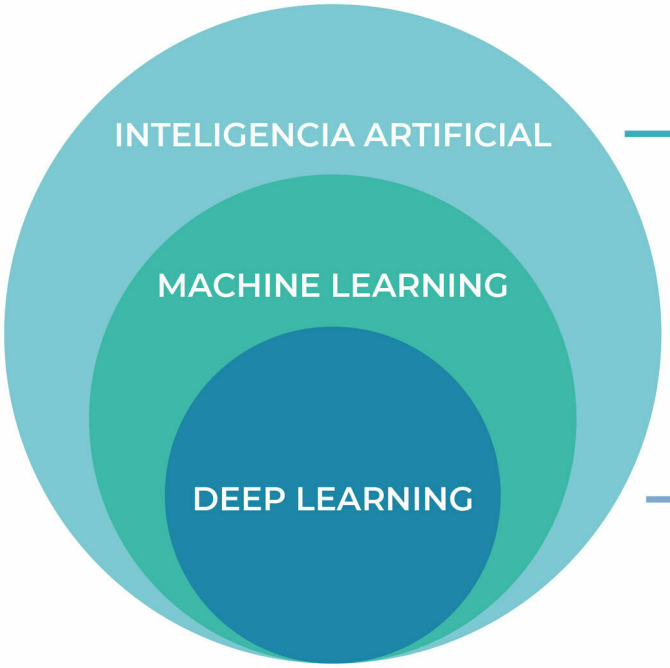
La tecnología también afecta a la transmisión del conocimiento



¿Qué es la Inteligencia Artificial?

Es una rama de las ciencias de la Computación que diseña y crea entidades con la capacidad de percibir datos de su entorno, analizarlos, asimilarlos y utilizarlos para conseguir un objetivo; de forma semejante a las capacidades humanas de cognición, y razonamiento.

Conceptos importantes

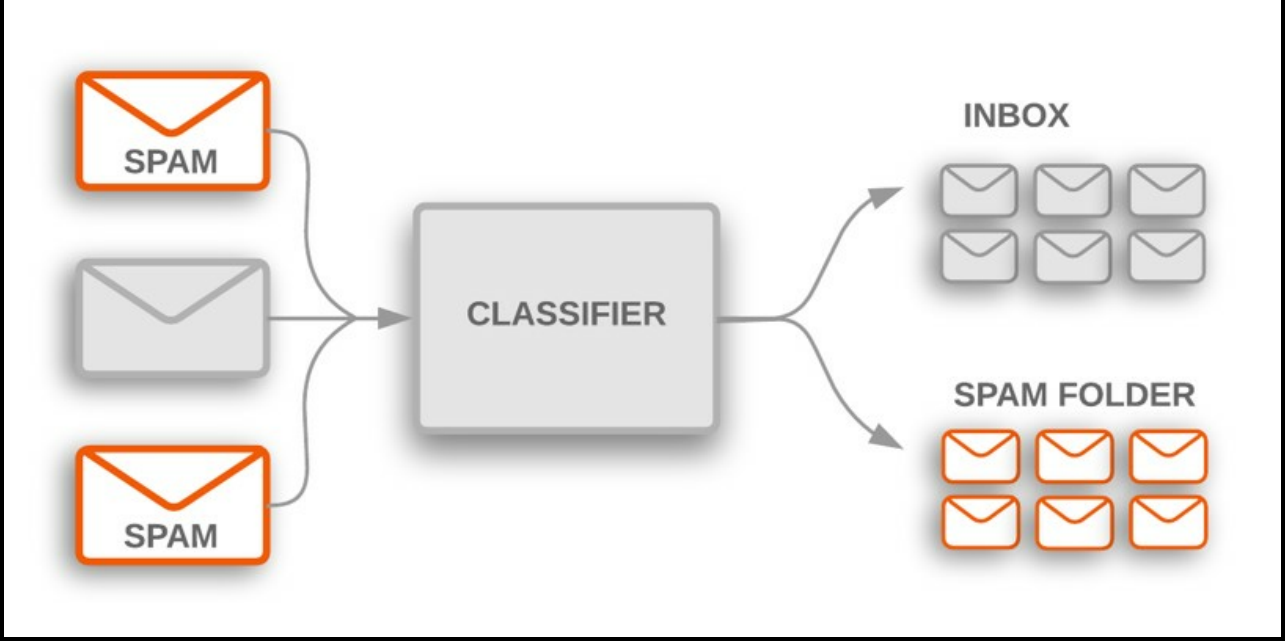


Cualquier técnica que permite a los ordenadores imitar el comportamiento humano.

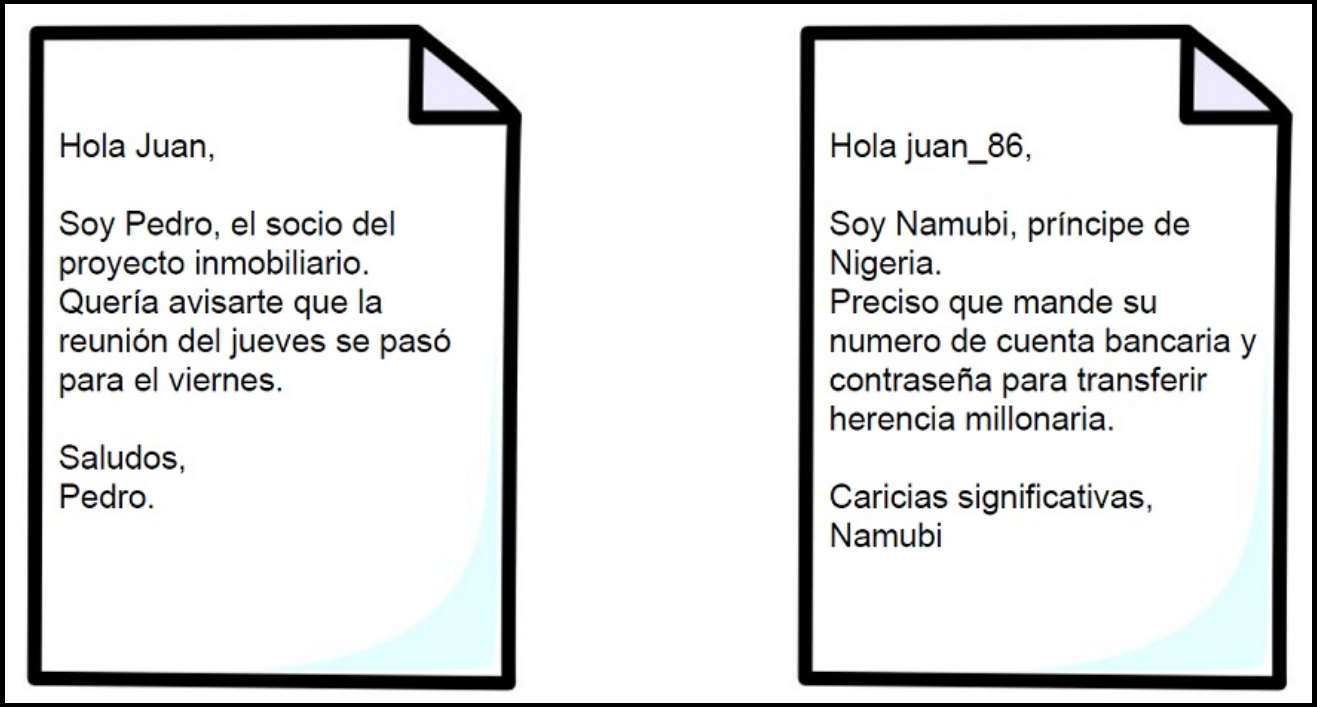
Subconjunto de técnicas de IA que utilizan métodos estadísticos para permitir que las máquinas mejoren con experiencias.

Subconjunto de ML que hace que el cálculo de de las redes neuronales multicapa sea factible.

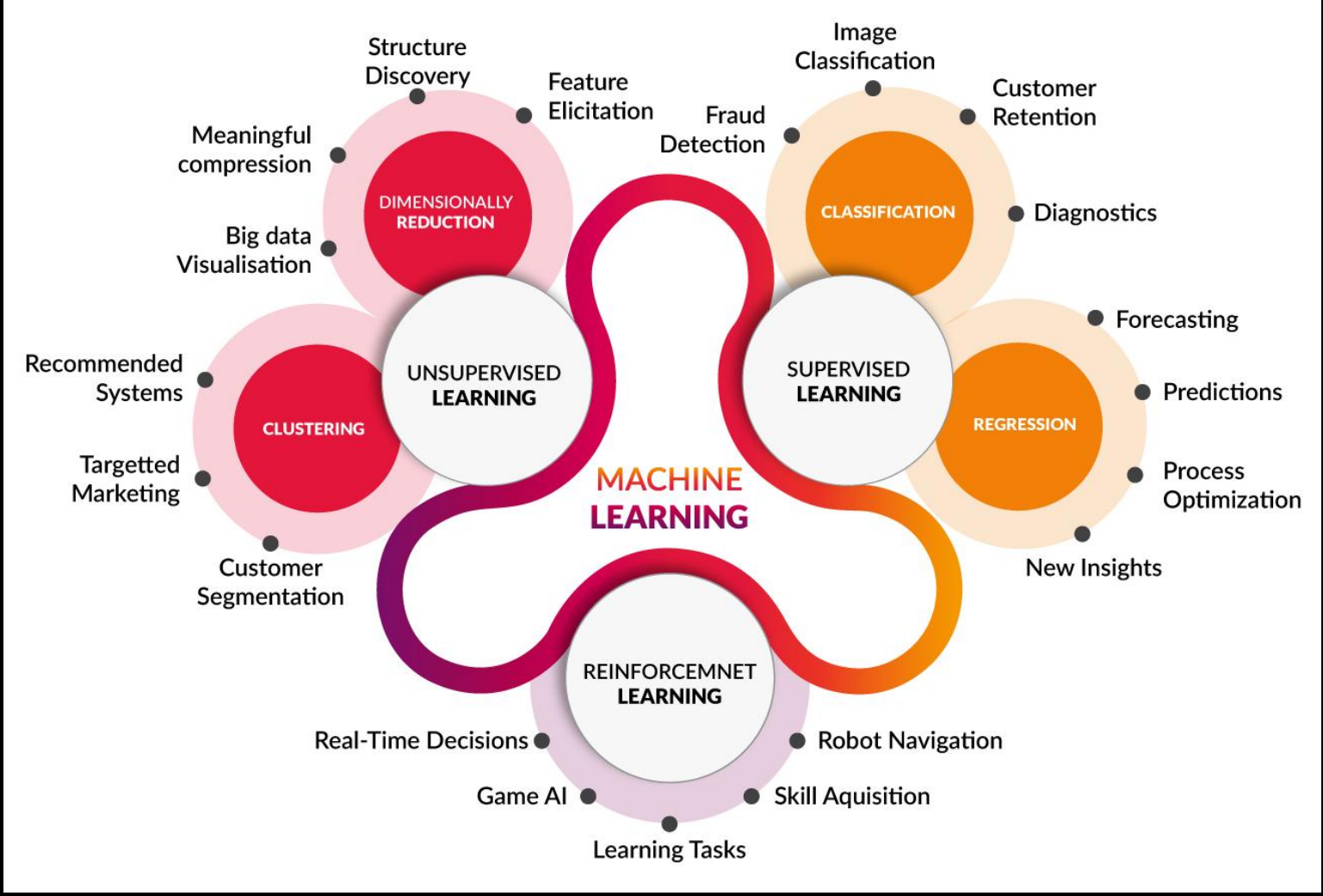
¿Cómo hacemos para que las computadoras aprendan de los datos?



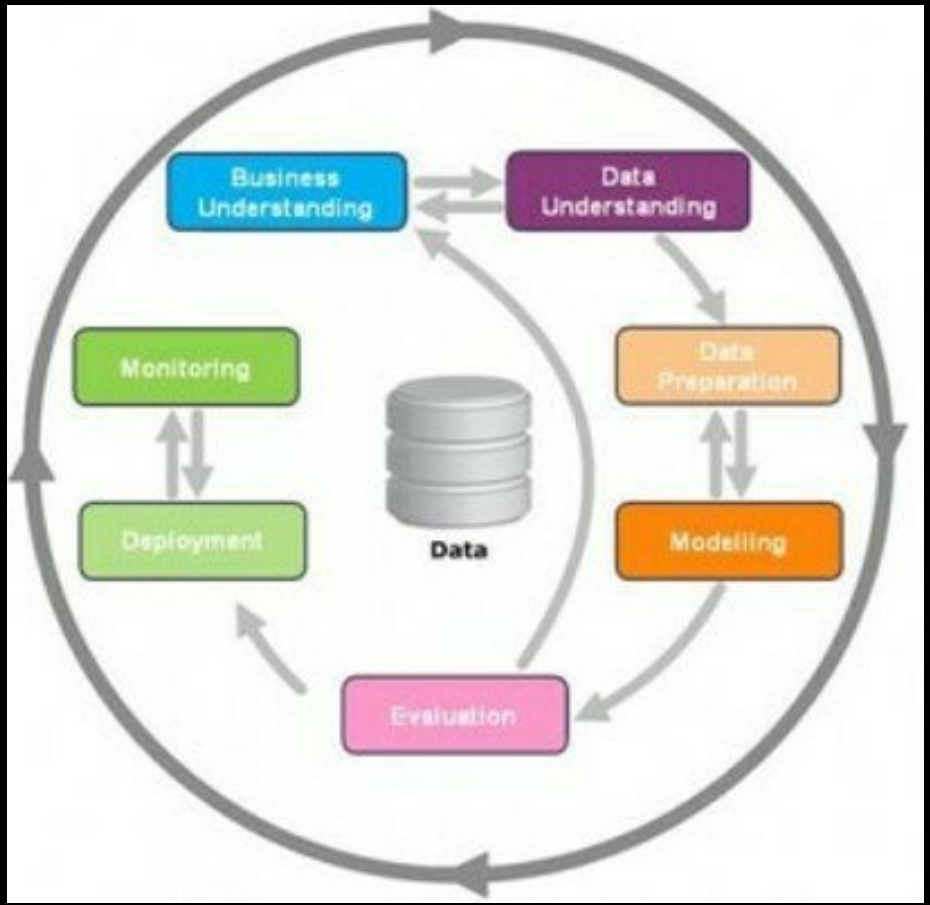
¿Cómo hacemos para que las computadoras aprendan de los datos?



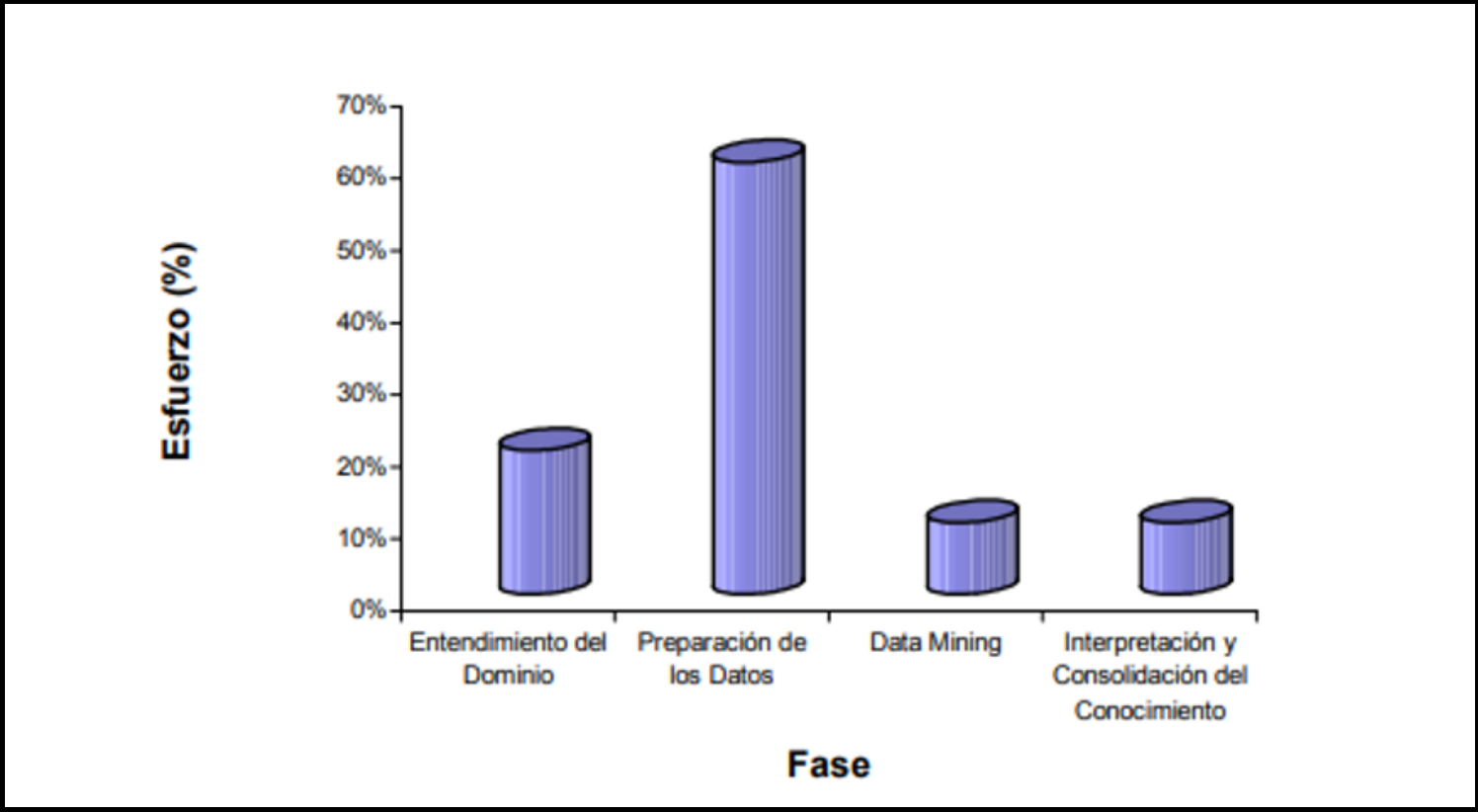
Machine Learning



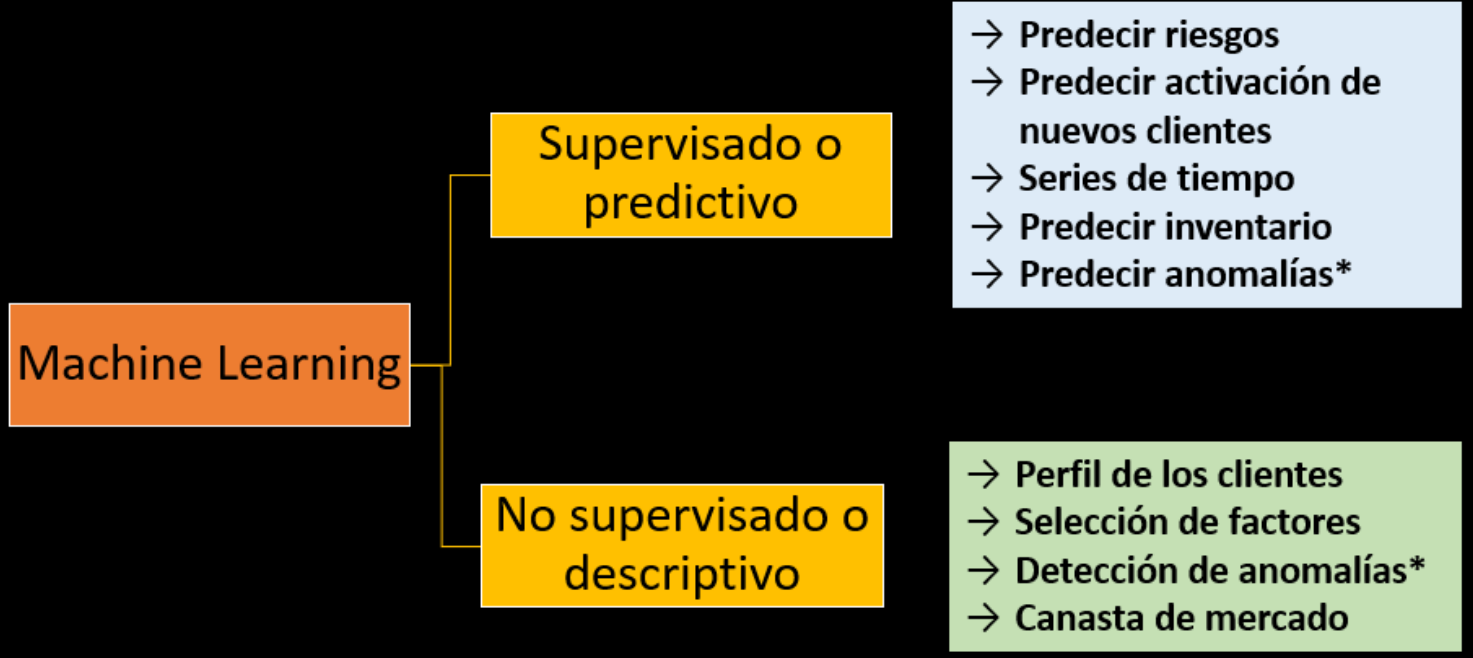
Metodología CRISP-DM



Fase vs Esfuerzo



Tipos de aprendizaje



Aprendizaje supervisado

Supervisado o predictivo

ID	ATR1	ATR2	ATR3	VARIABLE OBJETIVO
1				ALTO
2				BAJO
3				MEDIO
4				ALTO
n				MEDIO

Se conoce el **histórico** de la variable objetivo

ID	ATR1	ATR2	ATR3	VARIABLE OBJETIVO
n+1	mm	34	nn	

La variable que queremos predecir es una clase (categoría)

Clasificación

Aprendizaje supervisado

Supervisado o predictivo

ID	ATR1	ATR2	ATR3	VARIABLE OBJETIVO
1				0.2
2				0.9
3				0.5
4				0.1
n				0.4

Se conoce el histórico de la variable objetivo

ID	ATR1	ATR2	ATR3	VARIABLE OBJETIVO
n+1	mm	34	nn	

La variable que queremos predecir es una variable continua, como una probabilidad, edad, valor numérico

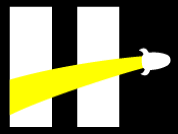
Regresión

Aprendizaje no supervisado

No supervisado o descriptivo

ID	ATR1	ATR2	ATR3
1			
2			
3			
4			
n			

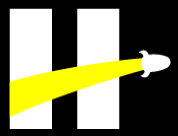
¿Cómo va a afectar la IA la vida de las personas?



Exploración de los Datos

Los datos con los que vamos a estar trabajando, son en definitiva la fuente del conocimiento necesario que debemos adquirir para poder resolver las preguntas que nos hacemos, entonces, es preciso conocer todas sus características, algunas de ellas son:

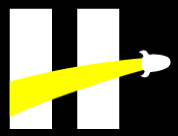
- Variabilidad.
- Estadística.
- Distribución.
- Rangos.



Falencias en los datos

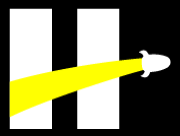
Como primera medida, antes de comenzar a realizar las tareas de análisis, vamos a encontrarnos con ciertas cuestiones que hacen a la calidad y fiabilidad del dato, y debemos resolverlas, entre ellas:

- Faltantes: ¿Qué hacer?
- Rangos de datos numéricos.
- Normalización.
- Errores: Su tratamiento.



Transformación de Datos

Es el proceso que más tiempo lleva en un flujo de Data Science y resulta muy importante no perder el objetivo de por qué lo hacemos. Por un lado, los modelos de Machine Learning que usemos, que van a "aprender" de nuestros datos, sólo entienden de números. La pregunta que queremos responder nos va a indicar cómo tenemos que trabajar con nuestro dataset.



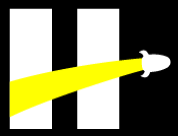
Tratamiento sobre Variables

Cualitativas Ordinales

- Tamaño de una prenda de ropa: XS, S, M, L, XL
- Tipo de Nafta por octanaje: 95, 98, más de 98

Podemos hacer una asignación a número enteros manteniendo el orden:

- S $\rightarrow$ 0
- M $\rightarrow$ 1
- L $\rightarrow$ 2



Tratamiento sobre Variables Cuantitativas Ordinales

- Este es uno de los tipos de **encoding** más comunes que vamos a tener que hacer. En el ejemplo, queremos llevar al género, los valores male y female, a 0 y 1. Lo importante es no perder cuál es cuál. Esto, en Pandas, lo podemos hacer con la función map(). Este tipo de encoding se denomina **Label_enconding**.

Tratamiento sobre Variables Cualitativas Nominales

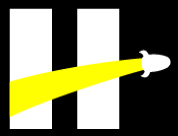
- Nacionalidad
- Tipo de Vino
- Especies de flores
- Color de auto

Tratamiento sobre Variables Cualitativas Nominales

Se llevan a variables **dummies** con **One-Hot Encoding**. La variable dummie será entonces aquella que tome valores 0 o 1, en función de la presencia o no de un atributo. Puede hacer que nuestro dataset crezca mucho.

Obs.	Ciudad
1	Rosario
2	Buenos Aires
3	Rosario
4	Mar del Plata
5	Córdoba

Obs.	D_BA	D_C	D_R
1	0	0	1
2	1	0	0
3	0	0	1
4	0	0	0
5	0	1	0



Tratamiento sobre Variables Cuantitativas

Son aquellas variables que se miden o se cuentan. Pueden ser discretas o continuas. Hay una relación de orden entre ellas. Se puede aplicar funciones de agregación.

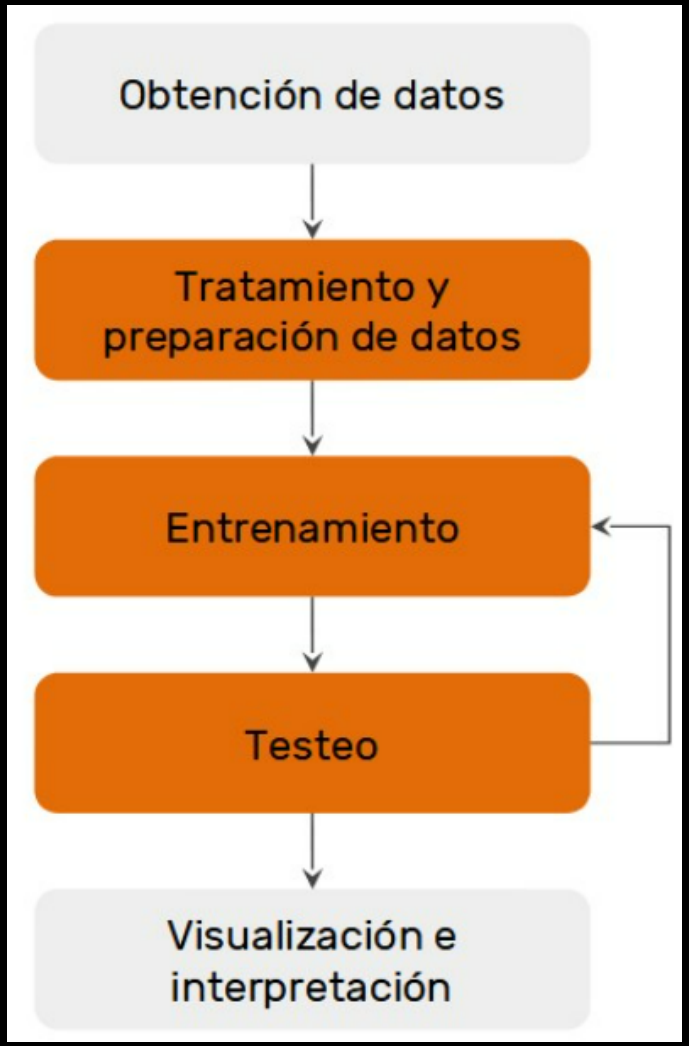
- Edad, Altura y Peso.
- Puntaje, precio de un vino.
- Valor de un pasaje.

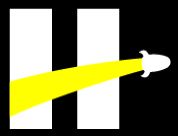
En general, ya vienen en un formato “cómodo” para trabajar, pero a veces queremos agruparlas según grupos o rangos, por ej.: agrupar edades en rangos (bebés, niños, adolescentes, adultos, ancianos), esto se denomina **Discretización y Binning**.

Scikit-Learn

Scikit-Learn es la librería base para Machine Learning en Python.

- Procesamiento de los datos
- Modelos de Clasificación y Regresión
- Métricas de Evaluación de algoritmos

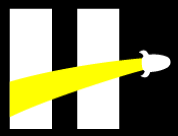




Scikit-Learn

Vamos a encontrar que Scikit-Learn trabaja con Clases e implementa de manera uniforme los atributos y métodos de sus objetos:

- Estimadores: Todos tienen el método fit()
- Predictores: Todos tienen el método predict()
- Transformadores: Todos tienen el método transform()
- Modelos: Todos tienen el método score()



Scikit-Learn

Las siguientes clases son las herramientas disponibles para procesar datos:

- SimpleImputer: Rellena valores faltantes.
- OneHotEncoder: Pasa de variables categóricas a dummies. Notar que con N instancias, son necesarias solo N-1 nuevas columnas.
- LabelEncoder: Pasa variables categóricas a valores numéricos.
- KBinsDiscretizer: Para discretización y binning, la principal diferencia con Pandas es que Scikit-Learn decide los límites de los bines de acuerdo a una estrategia que le pasemos de parámetro.
- SelectKBest: Selecciona atributos del dataset en base a diferentes criterios de evaluación. Puede servir como respaldo o referencia del análisis que se está realizando.

Reescalar los Datos

Muchos algoritmos funcionan mejor normalizando sus variables de entrada. Lo que en este caso significa comprimir o extender los valores de la variable para que estén en un rango definido. Sin embargo, una mala aplicación de la normalización o una elección descuidada del método de normalización puede arruinar los datos y, con ello, el análisis.

MinMax Scaler:

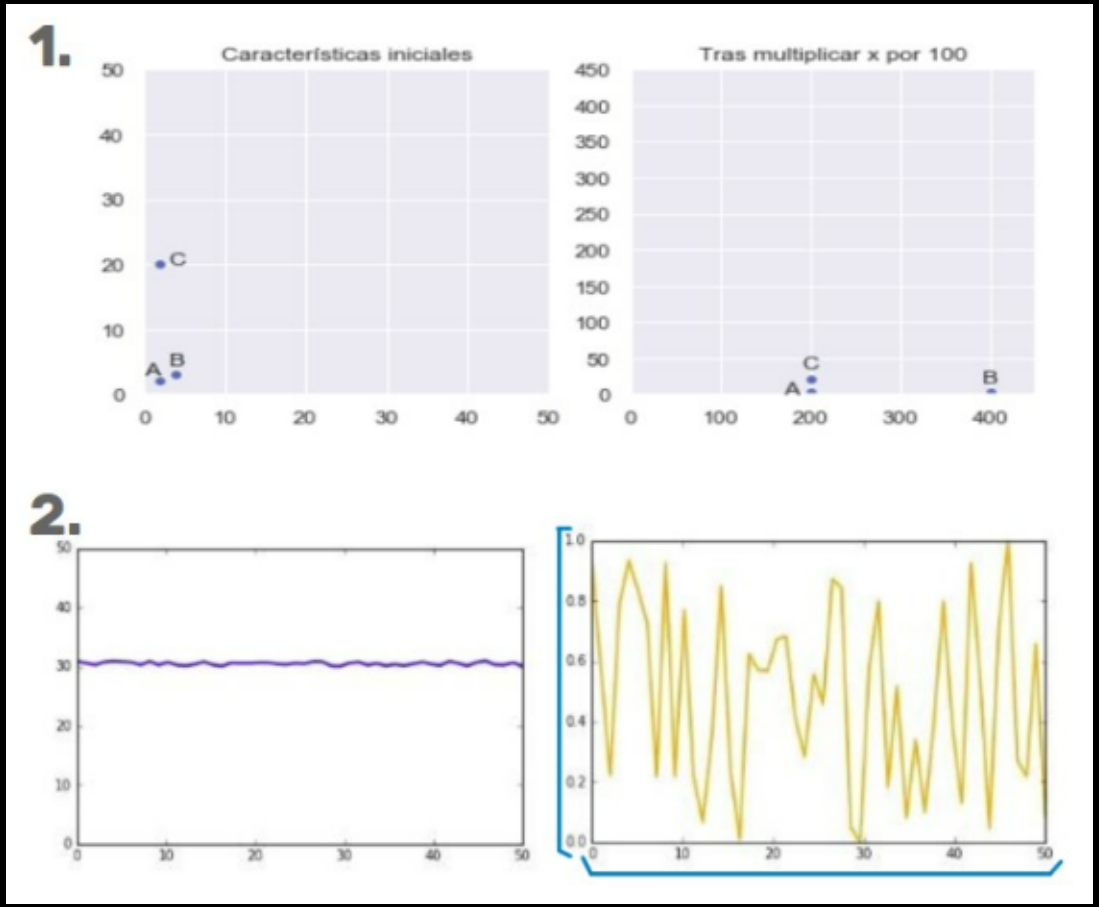
Las entradas se normalizan entre dos límites definidos

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Reescalar los Datos

Tener en cuenta que si se reescala un atributo, quizás sea conveniente reescalar otro, debido a que estamos rompiendo la proporcionalidad de los datos.

En 1 originalmente, A estaba más cerca de B, al multiplicar por 100, quedó más cerca de C. En 2 el ruido de la señal se hizo más notorio.

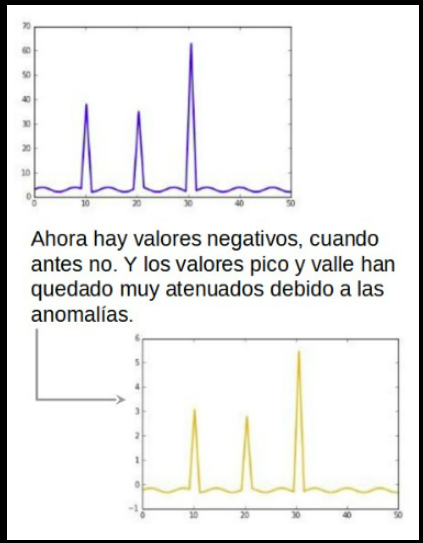


Reescalar los Datos

Standard Scaler:
A cada dato se le resta la media de la variable y se le divide por la desviación típica.

$$X_{normalized} = \frac{X - X_{mean}}{X_{stddev}}$$

Si bien puede resultar conveniente en datos que no tienen distribución de probabilidad Gaussiana o Normal debido a que se puede trabajar mejor bajo ese esquema, tanto la media como la desviación típica son muy sensibles a outliers.



Correlación entre Variables

Eventualmente vamos a querer conocer si existe una **variación conjunta** entre dos variables. Si este es el caso, podríamos ver que si una de las variables aumenta o disminuye su valor, que la otra también lo hace. La covarianza es una medida que intenta cuantificar esa relación:

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Correlación entre Variables

Con la covarianza también podemos determinar el coeficiente de relación o la recta de regresión, pero tiene el inconveniente de depender de la escala de los datos, motivo por el cual definimos la correlación, que es la covarianza dividida la desviación estándar de cada variable aleatoria obteniendo un valor que va de -1 a 1.

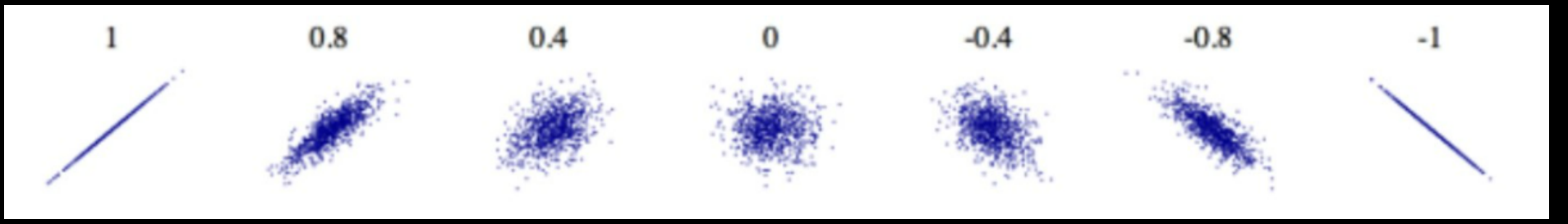
Donde:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Coeficiente

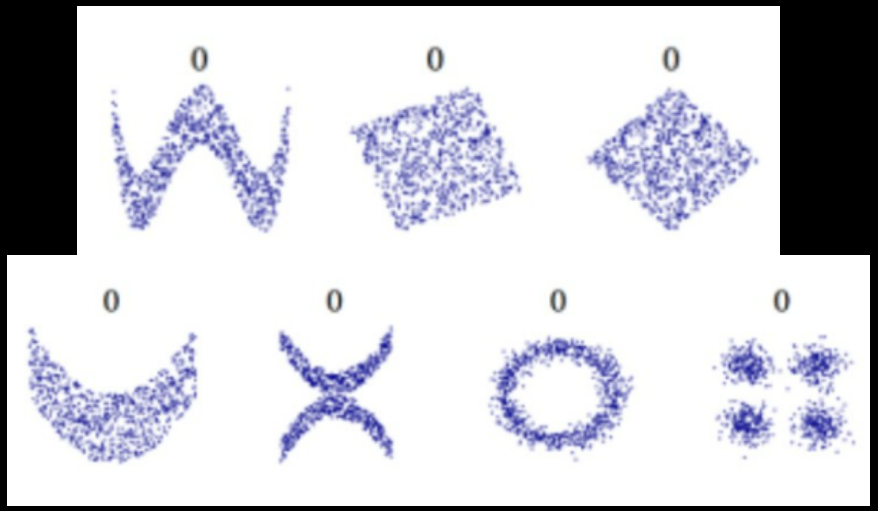
Se denomina **coeficiente de correlación lineal** o de **Pearson**, y es una cantidad adimensional.

- Correlación no implica causalidad.
- La correlación de Pearson es muy útil para encontrar correlaciones lineales.
- Si la relación entre las variables NO es lineal, existen otras correlaciones que pueden ser útiles: Spearman y Kendall.
- Coeficiente Negativo significa que son inversamente proporcionales entre sí con el valor del factor de coeficiente de correlación.
- Coeficiente Positivo significa que son directamente proporcionales entre sí, la media varía en la misma dirección con el factor del valor del coeficiente de correlación.
- Si el coeficiente de correlación es 0, significa que no existe una relación lineal entre las variables, sin embargo, podría existir otra relación funcional.



Correlación no Lineal

Si no hay relación lineal entre dos variables, entonces el coeficiente de Pearson será ciertamente 0. Sin embargo, si es 0, solo podemos decir que no existe una relación lineal, pero podría existir otra relación funcional:



Regresión Lineal

Consiste en predecir una respuesta numérica Y en base a atributos X1, X2, ..., Xp.

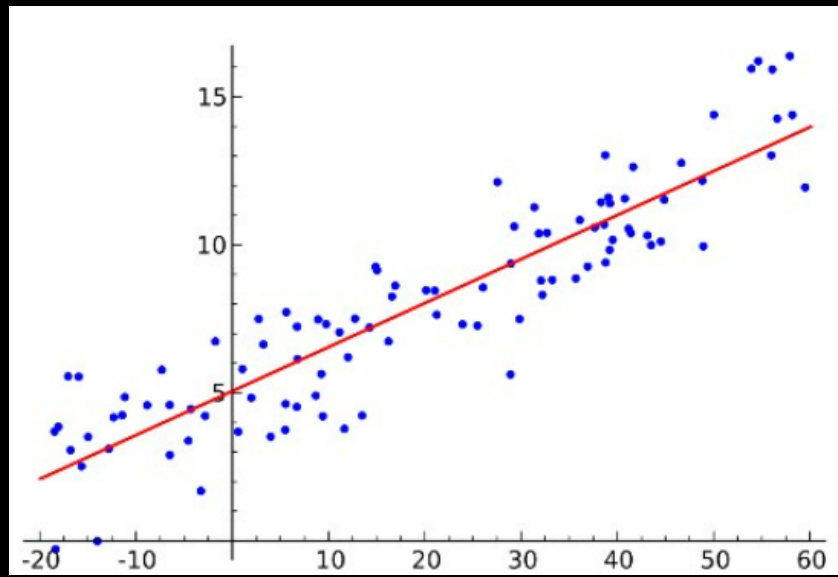
$$Y \approx f(X1, X2, ..., Xp)$$

Se busca Y = mX + b que mejor ajuste a los datos

m: pendiente

b: ordenada al origen

Se trata de aproximar los valores a una función lineal y aplicarla a los nuevos valores.

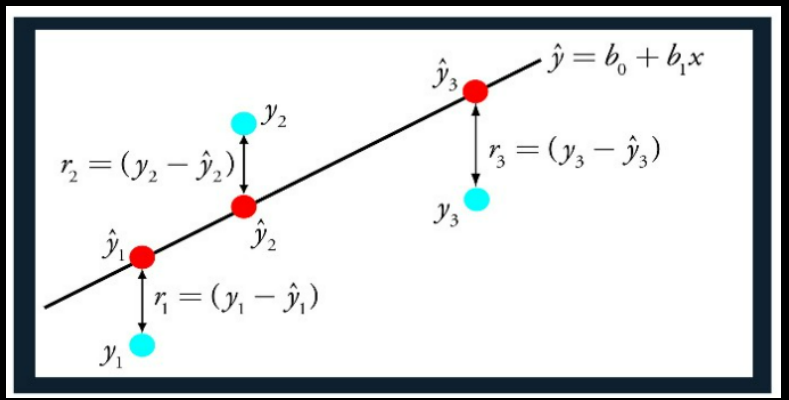


Medición de Error en Regresión

MAE (Error abstoluto medio):

- Se suman las distancias entre el valor en y real, y el predicho. Aunque esos errores tienen distinto signo. Si sumamos sin considerar eso, podría suceder que se cancelen.
- Sumando los valores absolutos, queda resuelto ese problema:
- Sin embargo ahora, a mayor cantidad de muestras el error se hace mayor.

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$



Medición de Error en Regresión

MSE (Error cuadrático medio)

- Se calcula el promedio de la suma de errores entre la predicción de un punto y su valor real elevado al cuadrado.
- Lo que se obtiene es error cuadrático medio, y la línea que lo minimice será la propuesta por la regresión.

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2$$